Advanced Gene Mapping Course

November 7-11, 2022
The Rockefeller University
New York, NY

Lectures

Table of Contents

Lectures given by: [1]Heather Cordell, [2]Suzanne Leal, [3]Gao Wang, [4]Judy Matuk, [5]Andrew DeWan, [6]Jurg Ott, and [7]Shamil Sunyaev

Genome-wide association studies (GWAS) - Part 1

Heather J. Cordell

Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University, UK

heather.cordell@ncl.ac.uk

Genome-wide association studies (GWAS)

- Popular (and highly successful) approach over past $\sim$ 15 years
- Enabled by advances in high-throughput (microarray-based) genotyping technologies
- Idea is to measure the genotype at a set of single nucleotide polymorphisms (SNPs) across the genome, in a large set of unrelated individuals
  - Cases and controls
  - Or population cohort measured for relevant quantitative phenotypes (height, weight, blood pressure etc)
  - Or related individuals (family data) – but need to analyse differently

Genome-wide association studies (GWAS)

Two individuals

Person 1	ACCTGTGTGCCCAATGGCGTCCCATACTATCGG
	ACCTGTGCGCCCAATGGCGTCCCATACTATCGG
Person 2	ACCTGTGCGCCCAGTGGCGTCCCATACTATCGG
	ACCTGTGCGCCCAGTGGCGTCCCATAGTATCGG

- Test each SNP for association/correlation with disease or quantitative phenotype

Association testing: case/control studies

- Collect sample of affected individuals (cases) and unaffected individuals (controls)
  - Or a else a sample of random "population" controls
    - Most of whom will not have the disease of interest
- Examine the association (correlation) between alleles present at a genetic locus and presence/absence of disease
  - By comparing the distribution of genotypes in affected individuals with that seen in controls

Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2|2	500 $(= a)$	200 $(= b)$
1|2	1100 $(= c)$	820 $(= d)$
1|1	400 $(= e)$	980 $(= f)$
Total	2000	2000

Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2|2	500 $(= a)$	200 $(= b)$
1|2	1100 $(= c)$	820 $(= d)$
1|1	400 $(= e)$	980 $(= f)$
Total	2000	2000

- Test for association (correlation) between genotype and presence/ absence of disease using standard χ^2 test for independence on 2 df

Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2|2	500 $(= a)$	200 $(= b)$
1|2	1100 $(= c)$	820 $(= d)$
1|1	400 $(= e)$	980 $(= f)$
Total	2000	2000

- Test for association (correlation) between genotype and presence/ absence of disease using standard χ^2 test for independence on 2 df
  - Defined as $\sum_{i=1,6} \frac{(O_i - E_i)^2}{E_i}$ where O_i and E_i are observed and expected counts (calculated from the row and column totals) respectively
  - Generates a p value indicating how significant the association/ correlation appears to be

Case/control studies

- Each person can have one of 3 possible genotypes at a diallelic genetic locus

Genotype	Cases	Controls
2|2	500 $(= a)$	200 $(= b)$
1|2	1100 $(= c)$	820 $(= d)$
1|1	400 $(= e)$	980 $(= f)$
Total	2000	2000

- Test for association (correlation) between genotype and presence/ absence of disease using standard χ^2 test for independence on 2 df
  - Defined as $\sum_{i=1,6} \frac{(O_i - E_i)^2}{E_i}$ where O_i and E_i are observed and expected counts (calculated from the row and column totals) respectively
  - Generates a p value indicating how significant the association/ correlation appears to be
- Two odds ratios can be estimated
  - $\text{OR}\,(2|2 : 1|1) = \frac{af}{be}$
  - $\text{OR}\,(1|2 : 1|1) = \frac{cf}{de}$

Odds ratios

- Odds of disease are defined as P(diseased)/P(not diseased)
  - Odds ratio OR$(2|2:1|1)$ repesents the factor by which your odds of disease must be multiplied, if you have genotype $2|2$ as opposed to $1|1$
    - i.e. the 'effect' of genotype $2|2$

Odds ratios

- Odds of disease are defined as P(diseased)/P(not diseased)
  - Odds ratio OR$(2|2:1|1)$ repesents the factor by which your odds of disease must be multiplied, if you have genotype $2|2$ as opposed to $1|1$
    - i.e. the 'effect' of genotype $2|2$
- Similarly, we can define the OR for $1|2$ vs $1|1$
  - As the factor by which your odds of disease must be multiplied, if you have genotype $1|2$ as opposed to $1|1$
    - i.e. the 'effect' of genotype $1|2$

Odds ratios

- Odds of disease are defined as P(diseased)/P(not diseased)
  - Odds ratio OR$(2|2:1|1)$ repesents the factor by which your odds of disease must be multiplied, if you have genotype $2|2$ as opposed to $1|1$
    - i.e. the 'effect' of genotype $2|2$
- Similarly, we can define the OR for $1|2$ vs $1|1$
  - As the factor by which your odds of disease must be multiplied, if you have genotype $1|2$ as opposed to $1|1$
    - i.e. the 'effect' of genotype $1|2$
- ORs are closely related (often $\approx$) genotype relative risks
  - The factor by which your probability of disease must be multiplied, if you have genotype $1|2$ as opposed to $1|1$ (say)
- If your genotype has no effect on your probability (and therefore on your odds) of disease, then the ORs=1.
  - So the association test can be thought of as a test of the null hypothess that the ORs=1

Genotype relative risks

- If a disease is reasonably rare, the odds ratio approximates the genotype relative risk (GRR, RR)

Genotype	Penetrance	GRR	Odds	OR
1/1	0.01	1.0	$0.01/0.99 = 0.0101$	1.00
1/2	0.02	2.0	$0.02/0.98 = 0.0204$	2.02
2/2	0.05	5.0	$0.05/0.95 = 0.0526$	5.21

- If your genotype has no effect on your probability (and therefore your RR) of disease, then both the ORs and the GRRs=1.

Dominant/recessive effects

Dominant:

Genotype	Cases	Controls	Total
2|2 and 1|2	500+1100	200+820	700+1920
1|1	400	980	1380
Total	2000	2000	4000

Recessive:

Genotype	Cases	Controls	Total
2|2	500	200	700
1|2 and 1|1	1100+400	820+980	1920+1380
Total	2000	2000	4000

- Can also rearrange table to examine effects of alleles (1 df tests):

Counting alleles

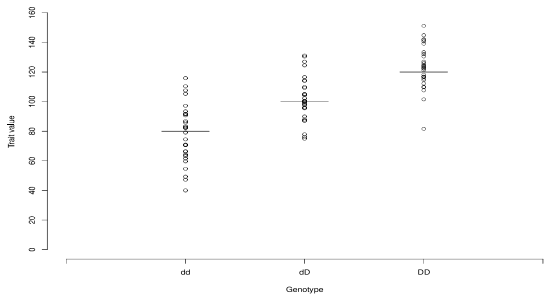
	Counts in	
Allele	Cases	Controls
2	2100 (=a)	1220 (=b)
1	1900 (=c)	2780 (=d)
Total	4000	4000

Allelic OR = ad/bc

- χ^2 test statistic on 1 df = $\sum_i (O_i - E_i)^2/E_i$ where O_i and E_i are the observed and expected values in cell i.
  - Assumes HWE under null and multiplicative allelic effects under alternative: considers chromosomes as independent units
  - Better approach: use counts in previous genotype table to perform a Cochran-Armitage trend test
  - Even better approach: use linear or logistic regression

Testing for association: quantitative traits

- Linear regression provides a natural test for quantitative traits
  - Testing the null hypothesis that the slope = 0



Logistic regression

- Used in case/control studies
  - Outcome is affected or unaffected
  - Model probability (and thus odds) of disease p as function of variable x coding for genotype:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x \quad \equiv c + mx$$

  - Use observed genotypes in cases and controls to estimate the values of regression coefficients β_0 and β_1
    - And to test whether $\beta_1 = 0$

Logistic regression

- Standard method used in standard epidemiological studies e.g. of risk factors such as smoking in lung cancer
- Main advantage is you can include more than one predictor in the regression equation e.g.

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where x_1, x_2, x_3 code for

- genotypes at 3 loci
- measured environmental covariates (e.g. age, sex, smoking etc),
- genetic principal component scores (to adjust for population substructure),
- interactions between loci etc. etc.

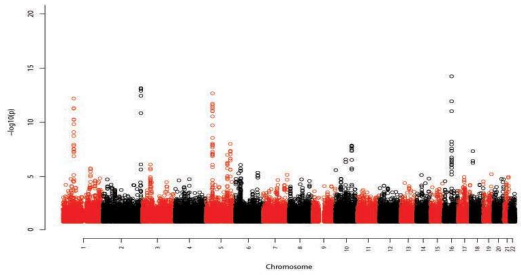
Testing for association

- All methods produce a test statistic and a p value at each SNP, indicating how significant the association/correlation observed appears to be
  - i.e. how likely it was to have occurred by chance
  - The threshold to declare 'genome-wide significance' is usually around $p = 5 \times 10^{-8}$
    - To account for multiple testing of many SNPs across the genome

Testing for association

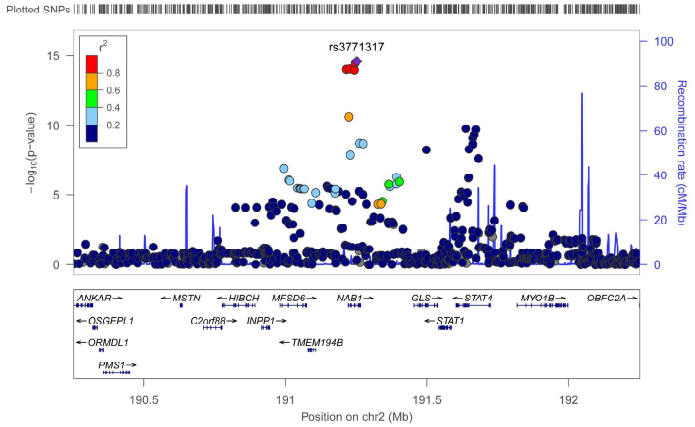
- All methods produce a test statistic and a p value at each SNP, indicating how significant the association/correlation observed appears to be
  - i.e. how likely it was to have occurred by chance
  - The threshold to declare 'genome-wide significance' is usually around $p = 5 \times 10^{-8}$
    - To account for multiple testing of many SNPs across the genome
- Alternative (Bayesian) methods produce a Bayes Factor
  - Indicates how likely the data is under the alternative hypothesis (of association between genotype and phenotype)
    - Compared to under the null hypothesis (of no association between genotype and phenotype)
  - Requires you to make some prior assumptions regarding the likely strength of associations (i.e. the value of the β's)
  - Choosing a sensible threshold (e.g. $\log_{10}$ BF> 4) requires you to make some prior assumptions regarding what proportion of SNPs in the genome are likely to be associated with the phenotype

Manhattan Plots



- At any location showing 'significant' association, we expect to see several SNPs in the same region showing association/correlation with phenotype
  - Due to the correlation or linkage disequilibrium (LD) between neighbouring SNPs

Close-up of hit region



Historical Perspective: Complement Factor H in AMD

- First (?) GWAS was by Klein et al. (2005) Science 308:385-389
- Typed 116,204 SNPs in 96 cases (with age-related macular degeneration, AMD) and 50 controls
  - Very small sample size – they were very lucky to find anything!
  - Luck was due to the fact the polymorphism has a very large effect (recessive OR=7.4)
- Klein et al. followed up on two SNPs passing threshold ($p < 4.8 \times 10^{-7}$)
  - Plus a third SNP that just failed to pass significance threshold, but lay in same region as first SNP

Complement Factor H in AMD

- Of the 3 SNPs followed up:
  - One appeared to be due to genotyping errors: significance disappeared on filling in some missing genotypes
  - First and third SNP lie in intron of Complement Factor H (*CFH*) gene
    - Lies in region previously implicated by family-based linkage studies
- Resequencing of the region identified a polymorphism of plausible functional effect
- Immunofluorescence experiments in the eyes of AMD patients supported the involvement of *CFH* in disease pathogenesis.

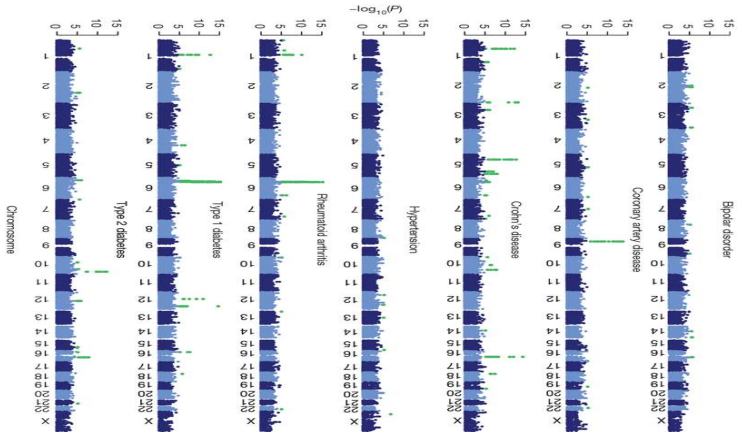
GWAS

- GWAS really got going in around 2007
  - See Visscher et al. (2012) AJHG 90:7-24 "Five Years of GWAS Discovery"
  - And Visscher et al. (2017) AJHG 101:5-22 "10 Years of GWAS Discovery: Biology, Function and Translation"
- 2007/2008 saw a slew of high-profile GWAS publications
  - Breast cancer (Easton et al. 2007)
  - Rheumatoid Arthritis (Plenge et al. 2007)
  - Type 1 and Type 2 diabetes (Todd et al. 2007; Zeggini et al. 2008)
- Arguably the most influential was the Wellcome Trust Case Control Consortium (WTCCC) study of 7 different diseases
  - http://www.wtccc.org.uk/

WTCCC

- Nature 447: 661-678 (2007)
- Considered 2000 cases for each of the following diseases:
  - Bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, type 2 diabetes
- Compared each disease cohort to common control panel
  - 3000 population-based controls
  - From 1958 birth cohort and National Blood Service
- Highly successful
  - WTCCC found 24 separate association signals
  - Including highly convincing signals in 5 out of the 7 diseases studied
  - All were replicated in subsequent independent follow-up studies

Manhattan plots for 7 diseases



Lessons from WTCCC (and others)

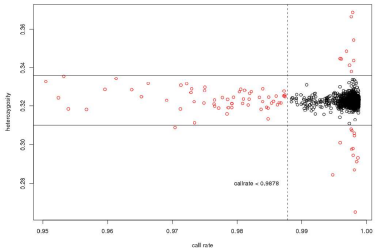
- Typically used rather standard statistical/epidemiological methods (χ^2 tests, t tests, logistic regression etc.)
- Success largely due to:
  - An appreciation of the importance of large sample size ($>$ 2000 cases, similar or greater number of controls)
  - Stringent quality control procedures for discarding low-quality SNPs and/or samples
  - Stringent significance thresholds ($p = 5 \times 10^{-8}$) to account for multiple testing and/or low prior prob of true effect
  - Importance of replication in an independent data set

Short break

Quality Control

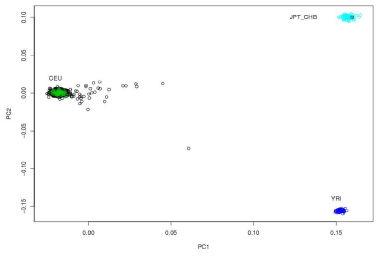
- Stringent QC checks are required for GWAS data
- Discard samples (people) deemed unreliable
  - Low genotype call rates, excess heterozygosity etc.
  - X chromosomal markers useful for checking gender
    - Males should 'appear' homozygous at all X markers
  - Genome-wide SNP data useful for checking relationships and ethnicity
- Discard data from SNPs deemed unreliable
  - On basis of genotype call rates, Mendelian misinheritances, Hardy-Weinberg disequilibrium
  - Exclude SNPs with low minor allele frequency (MAF)

QC: call rates and heterozygosity



- 61 sample exclusions (low call-rate); 23 exclusions (heterozygosity)
- SNP exclusions also made based on call-rates, MAF and Hardy-Weinburg equilibrium (HWE)

QC: ethnicity tests



- Multidimensional scaling (with 210 HapMap individuals) identifies 33 samples with non-Caucasian ancestry
- Similar multivariate methods can be used to model more subtle population differences between samples...

Multivariate Analysis

- Several related multivariate analysis techniques have been proposed for detecting population structure in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)

Multivariate Analysis

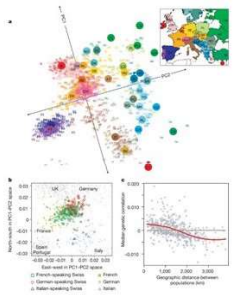
- Several related multivariate analysis techniques have been proposed for detecting population structure in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)
- If population differences can be detected (and adjusted for) in association analysis, this offers a way to deal with the problem of population stratification
  - Population sampled actually consists of several 'sub-populations' that do not really intermix
  - Can lead to spurious false positives (type 1 errors) in case/control studies

Multivariate Analysis

- Several related multivariate analysis techniques have been proposed for detecting population structure in genome-wide association studies
  - Principal components analysis (PCA)
  - Principal coordinates analysis (PCoA)
  - Multidimensional scaling (MDS)
- If population differences can be detected (and adjusted for) in association analysis, this offers a way to deal with the problem of population stratification
  - Population sampled actually consists of several 'sub-populations' that do not really intermix
  - Can lead to spurious false positives (type 1 errors) in case/control studies
- These techniques can also be used in quality control (QC) procedures, to check for (and discard) population outliers

Principal components analysis (PCA)

Genes mirror geography within Europe



J Novembre *et al.* (2008) *Nature* **456(7218):98-101**, doi:10.1038/nature07331

Principal Components Analysis

- Price et al. (2006) Nature Genetics 38:904-909; Patterson et al. (2006) PLoS Genetics 2(12):e190
  - Based on popn genetics ideas from Cavalli-Sforza (1978)
- Idea is to form a large matrix M of SNP counts (0,1,2) corresponding to the genotype at a L loci (=rows) for n individuals (=columns)

$$\mathrm{M} = \begin{pmatrix} g_{11} & g_{12} & \cdot & g_{1n} \\ g_{21} & g_{22} & \cdot & g_{2n} \\ g_{31} & g_{32} & \cdot & g_{3n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ g_{L1} & g_{L2} & \cdot & g_{Ln} \end{pmatrix}$$

Principal Components Analysis

- Subtract row means and normalise by function of row allele frequency $\sqrt{f_l(1-f_l)}$ to give matrix X

$$X = \begin{pmatrix} x_{11} & x_{12} & . & x_{1n} \\ x_{21} & x_{22} & . & x_{2n} \\ x_{31} & x_{32} & . & x_{3n} \\ . & . & . & . \\ . & . & . & . \\ x_{L1} & x_{L2} & . & x_{Ln} \end{pmatrix}$$

- This matrix will be used as starting point for PCA
  - In principal we could start with a different matrix – in particular not all PCA approaches would normalise by $\sqrt{f_l(1-f_l)}$

Multivariate Analysis

- Estimate covariance matrix $\Psi = X^T X$ between all pairs of individuals, with entries ψ_{ij} defined as the covariance (summing over SNPs) between column i and j of X
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)

Multivariate Analysis

- Estimate covariance matrix $\Psi = X^T X$ between all pairs of individuals, with entries ψ_{ij} defined as the covariance (summing over SNPs) between column i and j of X
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute the eigenvectors $\vec{v_j}$ and eigenvalues λ_j of matrix Ψ
    - Co-ordinate j of the kth eigenvector represents the ancestry of individual j along 'axis' k

Multivariate Analysis

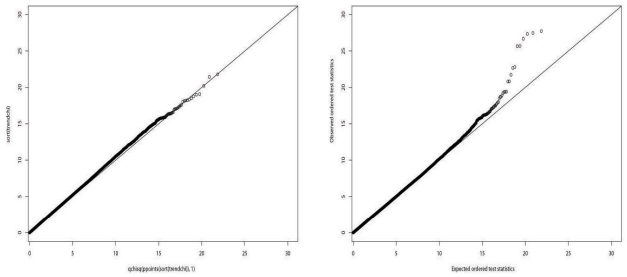
- Estimate covariance matrix $\Psi = X^T X$ between all pairs of individuals, with entries ψ_{ij} defined as the covariance (summing over SNPs) between column i and j of X
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute the eigenvectors $\vec{v_j}$ and eigenvalues λ_j of matrix Ψ
    - Co-ordinate j of the kth eigenvector represents the ancestry of individual j along 'axis' k
- For technical details, see McVean (2009) PLoS Genetics 5;10:e1000686

Multivariate Analysis

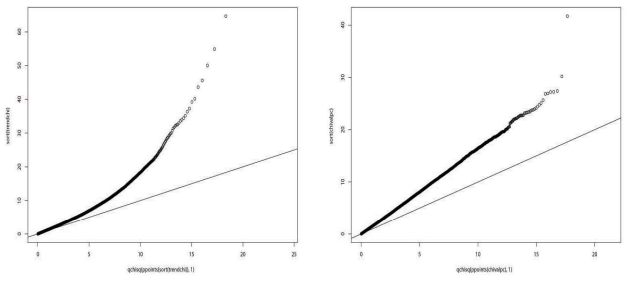
- Estimate covariance matrix $\Psi = X^T X$ between all pairs of individuals, with entries ψ_{ij} defined as the covariance (summing over SNPs) between column i and j of X
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute the eigenvectors $\vec{v}_j$ and eigenvalues λ_j of matrix Ψ
    - Co-ordinate j of the kth eigenvector represents the ancestry of individual j along 'axis' k
- For technical details, see McVean (2009) PLoS Genetics 5;10:e1000686
- Many genetics packages e.g. (PLINK) will allow you to calculate the top 10 (or more) PCs
  - Different geographic populations can often be well separated by just the first two or three PCs
    - Useful for outlier detection
  - For more subtle differences, you may need to calculate more PCs
    - And include them as covariates in the regression equation
    - Post-GWAS QC can determine whether you have included 'enough'

Post GWAS QC: Q-Q Plots (good)

- Plot ordered test statistics (y axis) against their expected values under the null hypothesis (x axis)



Q-Q Plots (bad)



Population stratification

- A QQ plot showing constant inflation (straight line with slope > 1) can indicate population stratification/population substructure
- Simple solution: Genomic Control (Devlin and Roeder 1999)
  - Use your observed test statistics to estimate the slope (=inflation factor λ)
  - Divide each test statistic by λ to get an adjusted (deflated) test statistic
- More complicated solution: use PCA/MDS or similar
- Even more complicated solution: use linear mixed models

Relatedness

- With genome-wide data, can also infer relationships based on average identity by descent (IBD) $\Psi = X^T X$ or identity by state (IBS)
  - Using 'thinned' subset of markers with high minor allele frequency (MAF) and in approximate linkage equilibrium
  - Simple relationships (PO, FS, MZ/duplicates) can identified with only a few hundred markers
  - More complicated relationships require 10,000-50,000 SNPs
- Various software packages, including PLINK, KING and TRUFFLE

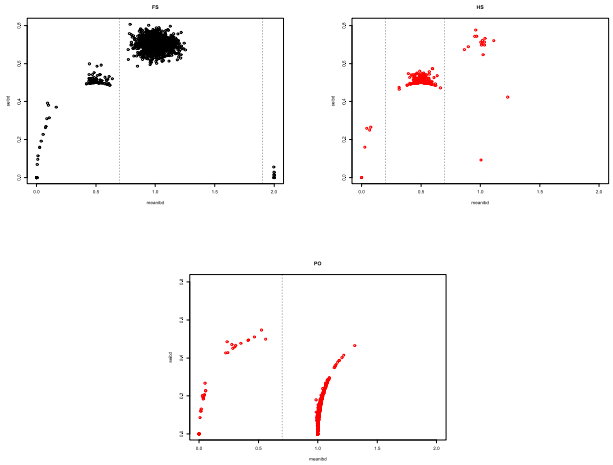
Expected IBD sharing

- Assuming no inbreeding, the IBD state probabilities are:

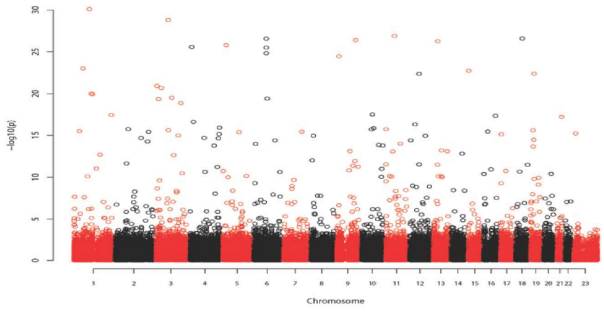
	Number of alleles shared IBD		
Relationship	2	1	0
MZ twins	1	0	0
Parent–Offspring	0	1	0
Full siblings	1/4	1/2	1/4
Half siblings	*0*	*1/2*	*1/2*
Grandchild–grandparent	*0*	*1/2*	*1/2*
Uncle/aunt–nephew/niece	*0*	*1/2*	*1/2*
First cousins	0	1/4	3/4
Second cousins	0	1/16	15/16
Double 1st cousins	1/16	6/16	9/16

- A useful visualisation tool is to plot SE(IBD) vs mean(IBD) (as estimated across the genome)
  - Or kinship coefficient ($\frac{1}{2}$P(IBD=2)+$\frac{1}{4}$P(IBD=1)) against P(IBD=0)

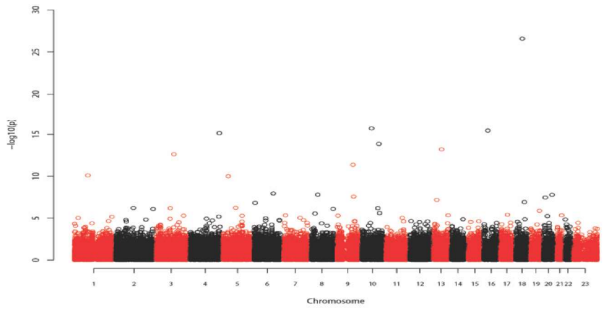
Full/half sibs and parent-offspring



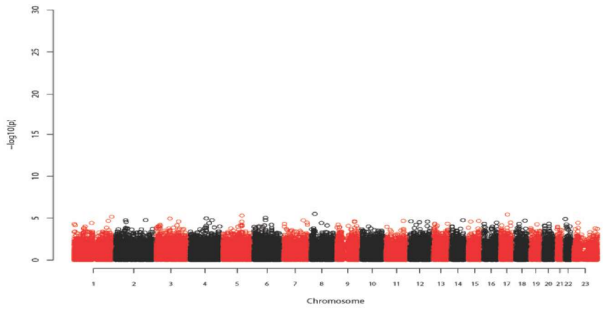
CHD GWAS results (low QC)



CHD GWAS results (better QC)



CHD GWAS results (final QC)



Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using meta-analytic techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies

Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using meta-analytic techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies
- Meta-analysis is often made easier by using *imputation*
  - Inferring (probabilistically) the genotypes at SNPs which have not actually been genotyped
    - On the basis of their known correlations with nearby SNPs that have been genotyped
    - Using a reference panel of people (e.g. 1000 Genomes) who have been genotyped at all SNPs

Genome-wide meta-analysis

- Puts together data (or results) from a number of different studies
  - Could analyse as one big study
  - But preferable to analyse using meta-analytic techniques
    - At each SNP construct an overall test based on the results (log ORs and standard errors) from the individual studies
- Meta-analysis is often made easier by using *imputation*
  - Inferring (probabilistically) the genotypes at SNPs which have not actually been genotyped
    - On the basis of their known correlations with nearby SNPs that have been genotyped
    - Using a reference panel of people (e.g. 1000 Genomes) who have been genotyped at all SNPs
- Enables meta-analysis of studies that used different genotyping platforms
  - By imputing to generate data at a common set of SNPs
    - Ideally while accounting for the imputation uncertainty in the downstream statistical analysis
    - In practice often don't bother - use post-imputation QC to remove poorly-imputed SNPS

Data Quality Control
NGS and Genotype Array Data

1

DNA Collection

- Blood samples
  - For unlimited supply of DNA
    - Transformed cell lines
      - Is expensive
    - Whole genome amplification
      - Allows for the creation of large amounts of DNA from initial small DNA sample
        - » Perform WGA on each sample three or more times and use pooled samples
      - Can experience lower call rates and higher genotyping error rates
      - Not recommend for whole genome sequencing or copy number variant (CNV) analysis
- Buccal Swabs
  - Small amounts of DNA
  - DNA not stable
- Saliva (Origene collection kit)

Measurement of DNA Concentrations

- Nanodrop
- Picogreen

2

Effect of Genotyping Error – Same Error Rates for Cases and Controls

- For family-based association studies - Trios
  - Can increase both type I and II error
- Population based studies
  - Increases type II error only

Quantitative Traits
If genotyping error is not correlated with trait values type II errors will be increased

3

Effects of Genotyping Error – Different Error Rates for Cases and Controls

- Cases and controls are sequenced/genotyped
  - At different times
  - Different institutions
  - Or one group, e.g., case or control, is predominately sequenced/genotyped in the same batch
- Can lead to different genotyping error rates in cases and controls
  - In this situation both type I and II error can be increased
- If sequencing/genotyping cases and controls
  - Randomize cases and controls so they are spread evenly across batches

Quantitative Traits
If genotyping error is correlated with trait values, it will also increase type I and II errors, e.g., individuals with elevated systolic blood pressure are genotyped in one batch and those with systolic blood pressure within the normotensive range in another batch

4

Genotype SNPs (~20-96) before Exome or Whole Genome Sequencing

- Genotype markers which can be used as DNA fingerprint
- Allows for Assessment of DNA quality
- Aids in determining the the genetic sex of study subjects
  - To aid in identification of potential sample swaps
- Detects cryptic duplicates
- For family data
  - Aids in determining close familial relationships
    - Non-paternity
    - Sample swaps
    - Cryptic relationships

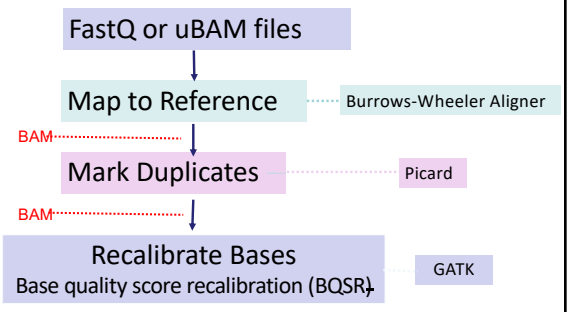
5

Detecting Genotyping Errors

- Duplicate samples genotyped using arrays to detect inconsistencies
  - Can use duplicate samples that are inconsistent to adjust clusters to improve allele calls
    - Will not detect systematic errors
- Usually generated only for genotype array data
  - Due to expense, duplicate samples are usually not generated for exome or whole genome sequencing studies

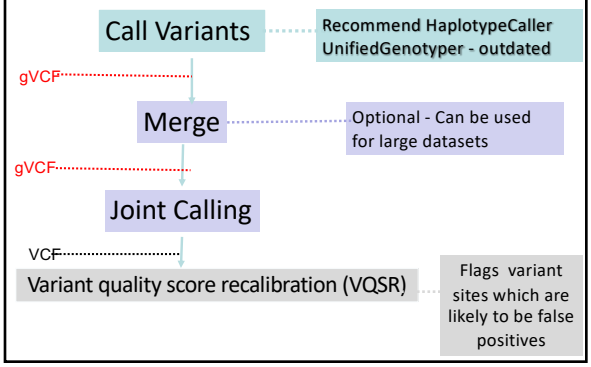
6

Variant Calling Pipeline -Step 1 Preprocessing



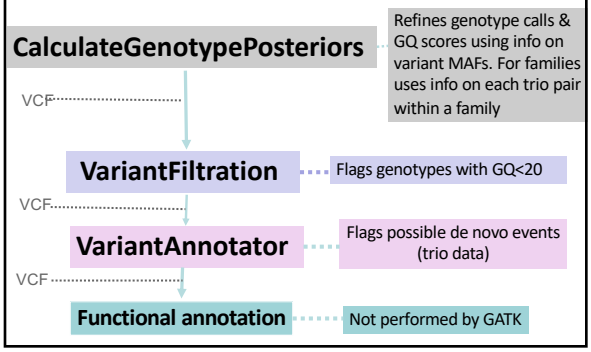
7

Variant Calling Pipeline-Step 2 Variant Discovery



8

Variant Calling Pipeline - Step 3 Call Set Refinement



9

A Short List of Additional Software to Detect Genetic Variation

- Exome data Copy Number Variation
  - CoNIFER (Copy Number Inference From Exome Reads)
    - Krumm et al. 2012
  - XHMM
    - Fromer et al. 2014
- WGS data structural variation
  - MetaSV
    - Mohiyuddin et al. 2015
  - LUMPY
    - Layer et al. 2014

10

Variant Calling

- BAM files are large and take considerable resources
  - Storage is expensive
  - One 30x whole genome is ~80-90 gigabytes
  - A small study of 1,000 samples will consume 80 terabytes of disk space
- The cost of cloud computing to call variants
  - (Souilmi et al. 2015)
  - $5 per exome
  - $50 per genome
    - For 1,000 samples
      - $5,000 exome
      - $50,000 genome

11

Working with gVCF Files

- Instead of obtaining VCF files
- Can obtain gVCF files to perform joint calling and complete the GATK pipeline
  - A whole genome gVCF
    - ~1 Gigabyte
      - 1/100th the size of a BAM file for one individual

12

Influences on Sequence Quality

- DNA quality
  - Age of sample
  - Extraction method
  - Source of sample
    - e.g., blood, skin punch, buccal
- Sequencing machines (read length)
- Median sequencing depth
- Alignment
- Variant calling method used
  - Single nucleotide variants and insertion/deletions
  - Structural variants

13

NGS Data Quality Control

- Extremely important to perform before data analysis
  - Poor data quality can increase type I and II errors
  - Due to inclusion of false positive variant sites or incorrect genotype calls
- Protocols for data QC are still in their infancy
  - No set protocols for QC
- QC is data specific
  - Dependent on read depth
  - Batch effects
  - Availability of duplicate samples
  - etc.

14

NGS Data Quality – Removal of Genotype Calls and Samples

- Sequence depth of coverage
  - DP_variant
    - High DP could be an indication of copy number variants
      - Which can introduce false positive variant calls
        - » Due to down sampling in GATK maximum DP is 250
  - DP_genotype
    - Concerned if depth is too low or too high
      - Low insufficient reads to call a variant site
    - Remove genotypes with low read depth, e.g., DP≤8
- Genotype quality (GQ) score
  - Removal of sites with low genotype quality core, e.g., GQ≤ 20

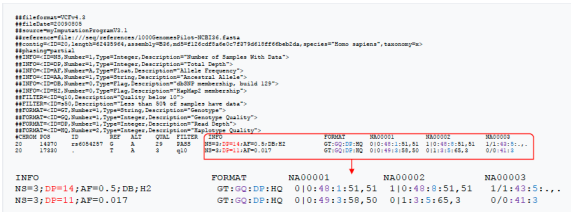
15

NGS Data Quality – Removal of Genotype Calls and Samples

- Sequence depth of coverage
  - DP_variant
    - High DP could be an indication of copy number variants
      - Which can introduce false positive variant calls
        - » Due to down sampling in GATK maximum DP is 250
  - DP_genotype
    - Concerned if depth is too low or too high
      - Low insufficient reads to call a variant site
    - Remove genotypes with low read depth, e.g., DP≤8
- Genotype quality (GQ) score
  - Removal genotypes with a low genotype quality core, e.g., GQ≤ 20

16

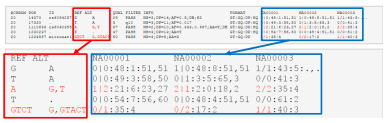
VCF Example



17

Variants with more than 2 Alleles

- Genetic analysis tools are usually developed to analyze variant sites that are diallelic
- Some sites may have >2 alleles
- The alleles at these sites need to be split
  - New loci are made each multi-allelic site each with only 2 alleles
    - bcftools
- Multiallelic sites can have higher error rates compared to diallelic sites



18

NGS Data Quality – Removal of Genotype Calls and Samples

- Removal of sites with missing data
  - e.g., missing > 10% of genotypes
- Removal of "novel" variant sites which only occur in one batch and the alternative allele is observed multiple times or the minor allele frequency (MAF) is high in overall sample
- Removal of sites that deviate from Hardy-Weinberg Equilibrium (HWE)
  - Must be performed by population, e.g., African American and European American
  - Related individuals should be removed from the sample before testing for deviations from HWE

19

NGS Data Quality Control

- GATK - Variant Quality Score Recalibration (VQSR)
  - Used to determine variant sites of bad quality
    - Variant site is a false positive call
- However even after this step
  - Concordance of duplicates (when available) and
  - and Ti/Tv ratios are often low
- Additional QC steps needs to be performed

20

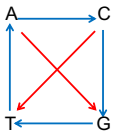
NGS Data Quality Control

- Values which are used for DP (genotype), GQ, and missing data cut offs are based upon
  - Concordance rates
    - If there are duplicate samples are available
  - Ti/Tv ratios
    - By individual
    - By batch
    - Entire data set
  - Amount of data removed
    - QC can remove substantial amounts of data which should be avoided
      - e.g., >15% of variant sites

21

Transition/Transversion (Ti/TV) Ratios

- Transition
  - Purine → Purine
  - Pyrimidine → Pyrimidine
- Transversion
  - Purine → Pyrimidine
  - Pyrimidine → Purine

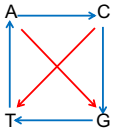


Transition
Transversion

22

Transition/Transversion (Ti/TV) Ratios

- Ti/Tv Ratios
  - Whole genome ~2.0
  - Exome novel ~2.7
  - Exome known ~3.5
- Ti/Tv ratios can be calculated by
  - Sample or
  - Dataset
- Ti/Tv ratios can be evaluated for subsets of data
  - e.g., by batch



Transition
Transversion

23

Sequence Data QC Overview

- Variant and genotype call level
  - Evaluation of batch effects
- Genotype call level – Removal of genotype calls
  - Low or high depth of coverage DP< 8
  - Low genotype quality score GQ< 20
- Removal of individual samples
  - >20% missing data
    - After taking the intersect of capture arrays
  - Samples without phenotype information

24

Sequence Data QC Overview

- Variant level – removal of variant sites
  - Low call rate
    - i.e., missing call rate > 10%
  - "Novel" variant sites observed ≥2 only in a single batch
  - Deviation from Hardy-Weinberg-Equilibrium
    - Population specific
    - Unrelated individuals
      - e.g., $p<5 \times 10^{-8}$, $p<5 \times 10^{-15}$

25

Data Clean – Assessing Sex Chromosomes

- When data is collected on study subjects they are asked about their gender/sex and not their genetic sex
  - Differences in gender/sex and genetic sex can be due to
    - Sample swaps
    - Study subjects who are not cisgender
- Some study subjects may have neither a XX nor XY karyotype
  - Turner syndrome X0
  - Klinefelter syndrome XXY

26

Data Clean – Assessing Chromosomal Sex

- Study subjects labeled as females with an excess of homozygous genotypes on the X chromosome can denote
  - That their genetic sex is male
  - Turner Syndrome

27

Data Clean – Assessing Chromosomal Sex

- Study subjects labeled as males with an excess of heterozygous SNPs* on the X chromosome can denote
  - That their genetic sex is female
  - Klinefelter syndrome
- Note: Individuals who are XY will also be heterozygous for markers in the pseudoautosomal regions
- Availability of Y chromosome data
  - Can greatly aid in determining genetic sex and if an individual has Turner or Klinefelter syndrome

*Both males and females have two alleles for each locus on the X chromosome in the datafile, although males are hemizygous

28

Data Clean – Assessing Sex Chromosomes

- Individuals whose labeled gender/sex does not match their genetic sex are removed from the analysis
- This observation may be due to a sample swap
  - When samples are swapped
    - Phenotype data will be incorrect
      - e.g., may be a case when labeled as a control

29

Checking for Duplicate and Related Individuals

- Duplicate samples are sometimes included in a study as part of quality control to detect inconsistencies
  - Will not detect systematic errors
  - Usually not included in exome and whole genome sequencing studies
  - Intentional duplicates can easily be removed before data quality control
- Cryptic duplicates (unintentional)
  - DNA sample aliquoted more than once
  - Individual ascertained more than once for a study
    - e.g. The same individual undergoes the same operation more than once and is ascertained each time
- Individuals who are related to each other may participate in the same study
  - Unknown to the investigator
  - Or be part of the study design

30

Duplicate and Related Individuals Need to be Identified

- For duplicate samples
  - Only one can be retained
- For related individuals
  - PCA is performed first with unrelated individuals and related individuals are then projected onto the PCs of unrelated individuals
  - Mixed-models need to be used to analyze the data if related individuals are included*
    - Case-Control
      - Generalized linear mixed models (GLMM)
    - Quantitative traits
      - Linear mixed models (LMM)
  - If not type I error rates can be increased

*If only a few related individuals in sample, may wish to remove them or use LMM/GLMM to control type I errors. Must use LMM/GLMM if related individuals are included in the dataset. If possible, opt for LMM/GLMM since it can help to control type I error due to other types of structure in the data, even when no closely related individuals are included in the analysis.

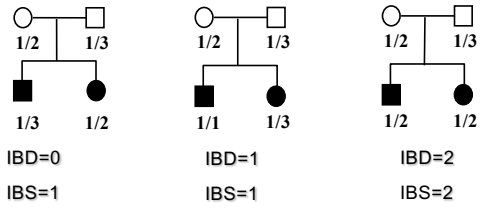
31

Identifying Duplicate and Related Individuals

- Duplicate and related individuals can be detected
  - By examining **Identity-by-State** (IBS) adjusted for allele frequencies (p-hat) between all pairs of individuals within a sample
  - Identify-by-descent (IBD) sharing can be estimated

32

Identity by Descent (IBD)/Identity-by-State (IBS)



33

IBD Sharing Estimated Pairwise for all Individuals in a Samples

- PLINK (Purcell et al. 2007)
- Uses sequence (or genotype array) data to check IBD
  - Prune markers to remove those in LD
    - e.g., r^2<0.1
- P-hat is calculated using the "population" allele frequency
- Used to approximates IBD sharing
- IBD is the number of alleles of alleles which are shared between a pair of individuals
  - Can either share 0, 1, and 2 alleles

34

Identifying Duplicate and Related Individuals

- Monozygote twins and duplicate samples will share 100% of their alleles IBD
  - IBD=2 is 1.0 (can be lower due to genotyping error)
- Siblings and child-parent pairs will share 50% of their alleles IBD
  - For parent-child IBD=1 is 1.0 (IBD=0 is 0 & IBD=2 is 0)
  - For sibs IBD=1 is ~0.50 (IBD=0 is ~0.25 & IBD=2 is ~0.25)
    - For more distantly related individuals the IBD measure will be lower

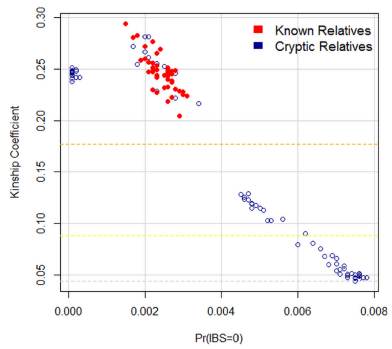
35

Identifying Duplicate and Related Individuals

- KING [Kinship-based INference for Gwas (*Manichaikul et al. 2010*)] can also be used to identify duplicate and related individuals
  - KING is more robust to population substructure and admixture
    - Prune markers for LD (e.g., r^2<0.1)
  - Provides kinship coefficients
    - Duplicate samples
      - Kinship coefficient equals 0.5
    - Siblings
      - Kinship coefficient equals 0.25

36

King Graphical Output



37

Multiple Individuals observed that are distantly "Related"

- If individuals in sample come from different populations
  - e.g., individuals from the same population within the sample will have inflated p-hat values due to incorrect allele frequencies
    - Incorrectly appear to be related to each other
- "Relatedness" amongst many individuals can also be observed when batches are combined if they have different error rates
  - Individuals from the same batch appear to be related
- DNA contamination can cause "relatedness" between multiple individuals

38

Principal Components Analysis (PCA) / Multidimensional Scaling (MDS)

- Can be used to identify outliers
- Population substructure
  - Individuals from different ancestry
    - e.g., African American samples included in samples of European Americans
- Batch effects
- Use a subset of markers which have been LD pruned
  - Only very low levels of LD between marker loci
    - e.g., $r^2<0.1$
  - MAF cutoff dependent on sample size
    - e.g MAF> 0.01
      - Can use lower MAF for large sample sizes

39

Principal Components Analysis (PCA) / Multidimensional Scaling (MDS)

- Unrelated individuals are used to generate PC plots
  - Related individuals are projected onto to the PC plots
- Plot 1st component vs. 2nd component
  - Additional PCs should also be plotted
    - e.g.. PCs 1-10
- Mahalanobis distance can be used to determine outliers
  - e.g., <1

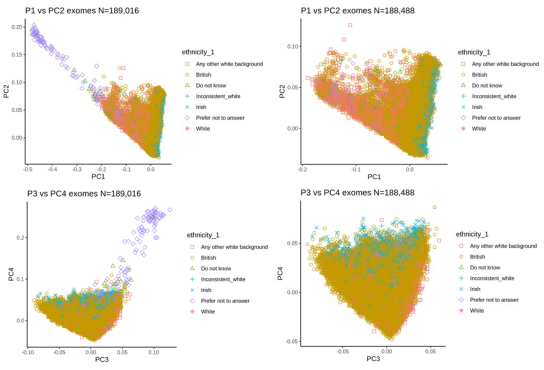
40

PCA/MDS Can be Used to Identify Outliers

- Individuals of different ancestry
  - e.g., African American samples included with European Americans samples
  - Can use samples from HapMap/1000 genomes to help to determine the ancestry for samples that are outliers
    - Should not include HapMap/1000 genomes samples when calculating components to control for population substructure/admixture
- Batch effects

41

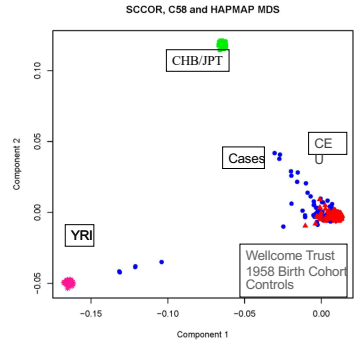
Principal Components Analysis Example



Exclusion of Outliers using Mahalanobis distance (0.997)

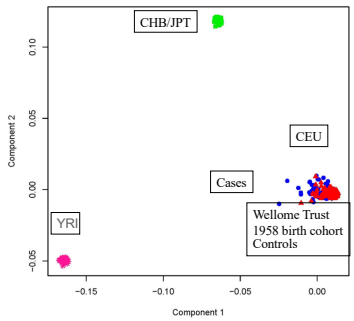
42

Detecting Outliers Using PCA and HapMap Sample



43

Detecting Outliers Using PCA and HapMap Sample



44

Detecting Genotyping Error – Examining HWE

- Testing for deviations from HWE not very powerful to detect genotyping errors
- The power to detect deviations from HWE dependent on:
  - Error rates
  - Underlying error model
    - Random
    - Heterozygous genotypes -> homozygous genotypes
    - Homozygous genotypes ->Heterozygous genotype
  - Minor allele frequencies (MAF)

45

Detecting Genotyping Error – Examining HWE

- Controls and Cases are evaluated separately
  - Deviation found only in cases can be due to an association
- Test for deviation from HWE only in samples of the same ancestry
  - Population substructure can introduce deviations from HWE
- Do not include related individuals when testing for deviations from HWE
  - Can cause deviations from HWE

46

Detecting Genotyping Error – Examining HWE

- What criterion is used to remove variants due to a deviation from HWE
  - GWAS studies have used 5.0 x 10^{-7} to 5.0 x 10^{-15}
- Quantitative Traits
  - Caution should be used removing markers which deviate from HWE may be due to an association
    - Remove markers with extreme deviations from HWE and Flag markers with less extreme deviations from HWE
- When performing imputation need to be more stringent in removing variants which deviate from HWE

47

Sequence Data QC Overview

- Remove variant sites that fail VQSR
- Remove genotypes with low DP, GQ scores, etc.
- Remove variant sites with large percent of missing data
- Remove samples with missing large percent of missing data
- Evaluate genetic sex of individuals based upon X and Y chromosomal data
  - Sample mix-ups
  - Individuals with Turner or Klinefelter Syndrome

48

Sequence Data QC Overview

- Evaluate samples for cryptically related individuals and duplicates
  - Use variants which have been pruned for LD
    - e.g., $r^2<0.1$
  - King or Plink algorithm
    - Always remove duplicate individuals
      - Retaining only one in the sample
    - If sample includes related samples use linear mix models (LMM)/Generalized LMM (GLMM) to control for relatedness
      - Best to perform even for data without related individuals
    - If only a few related individuals can retain only one individual of a relative group if not using LMM or GLMM

49

Sequence Data QC Overview

- Detection of sample outliers
  - Perform principal components analysis (PCA) or multidimensional scaling (MDS) to detect outliers
    - Use variants pruned for LD
      - e.g. $r^2<0.1$
    - Use unrelated individuals and then project related individuals onto the PCs
- Due to population substructure/admixture and batch effects
- Remove effects by
  - Additional QC
  - Removal of outliers (can be determined by Mahalanobis distance) and\or
  - Inclusion of MDS or PCA components in the association analysis

50

Sequence Data QC Overview

- Remove/flag variant sites that deviate from HWE in controls
  - HWE should be only be tested in unrelated individuals from the same population
- Post Analysis - Quantile-Quantile (QQ) plots
  - To evaluate uncontrolled batch effects and population substructure/admixture

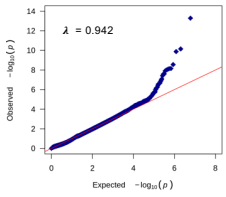
51

QQ Plots - Genome Wide Association Diagnosis

- Thousands of variants/genes are tested simultaneously
- The p-values of neutral markers follow the uniform distribution
- If there are systematic biases, e.g., population substructure, genotyping errors, there will be a deviation from the uniform distribution
- QQ plots offers an intuitive way to visually detect biases
- Observed p-values are ordered from largest to smallest and their $-\log_{10}(p)$ values are plotted on the y axis and the expected $-\log_{10}(p)$ values under the null (uniform distribution) on the x axis

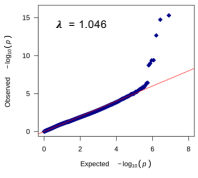
52

QQ Plot of Exome Wide P Values UK Biobank 200K



Hearing aid users

Case N= 6,436
Controls N= 96,601



Problem hearing with background noise

Cases N=65,660
Controls N= 96,601

53

Genomic Inflation Factor to Evaluate Inflation of the Test Statistic

- Genomic Inflation Factor (GIF): ratio of the median of the test statistics to expected median and is usually represented as λ
  - No inflation of the test statistic λ=1
  - Inflation λ>1
  - Deflation λ<1
    - Can be observed when a study is underpowered
- Problematic to examine the mean of the test statistic
  - Can be large if many variants are associated
    - Particularly if they have very small p-values
    - Should not be used

54

Phenotype	Covariate	Mean Chi-Square	GIF (λ)
BP		1.23829	1.16932
BP	Age	1.24119	1.18025
BP	**Age-EV1**	1.09471	**1**
BP	Age-EV2	1.0881	1
BP	Age-EV4	1.08385	1
BP	Age-EV10	1.09582	1.00402
BPI		1.14931	1.08921
BPI	Age	1.15139	1.08113
BPI	Age-EV1	1.05079	1.01148
BPI	**Age-EV2**	1.0428	**1**
BPI	Age-EV4	1.04204	1
BPI	Age-EV10	1.05421	1.01724
BPII		1.17283	1.25664
BPII	Age	1.17583	1.26996
BPII	Age-EV1	1.09874	1.15065
BPII	Age-EV2	1.09904	1.16425
BPII	Age-EV4	1.09502	1.14609
BPII	Age-EV10	1.10046	1.1418
BPII	Sex,Age-EV1	1.05958	1.06424
BPII	**Sex,Age-EV4**	1.05817	**1.05323**
BPII	Sex,Age-EV10	1.06338	1.05581

55

Example -Project Description

- 1,667 Samples
- Seven cohorts
- Two sequencing centers
  - Center 1
    - Two capture arrays
      - NimbleGen V2Refseq 2010 (CA1): 1082
        - » Batch 1 and 3
      - NimbleGen bigexome 2011 (CA2): 234
        - » Batch 2
  - Center 2
    - One capture array
      - Agilent SureSelect
        - » Batch 4
- Four batches
- No intentional duplicate samples

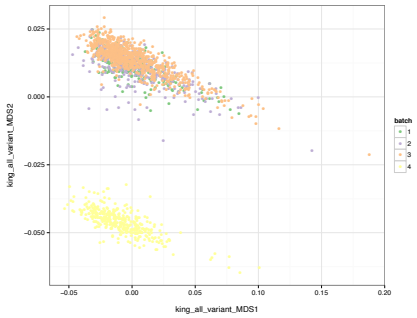
56

Example Project Description

- Intersection of the three capture arrays used
  - NimbleGen V2Refseq 2010
    - Batch 1 and 3
  - NimbleGen bigexome 2011
    - Batch 2
  - Agilent Sure Select
    - Batch 4
- Sequencing machine
  - Illumina HiSeq
- Sequence alignment
  - BWA
- Multi-sample variant calling
  - GATK

57

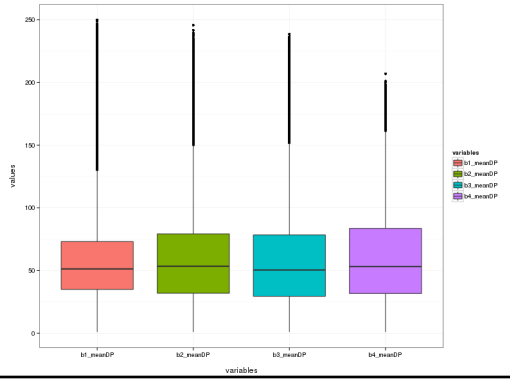
MDS First 2 Components Before QC*



*After VQSR

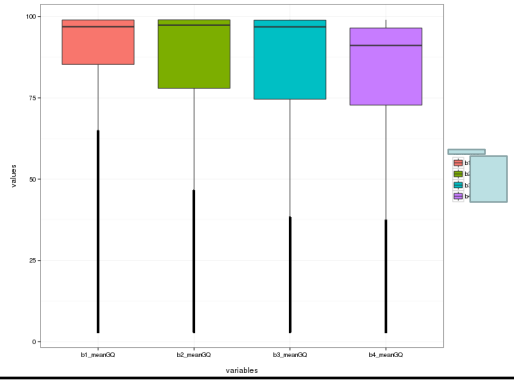
58

Mean GP (genotype) by Batch



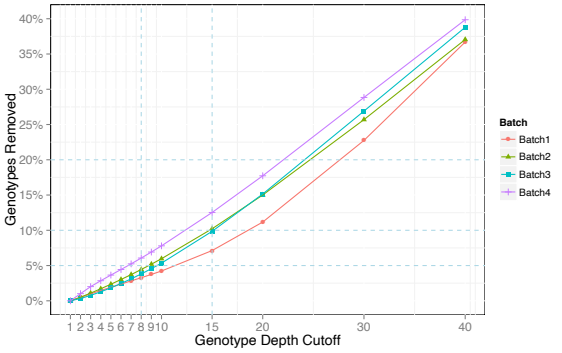
59

Mean GQ by Batch



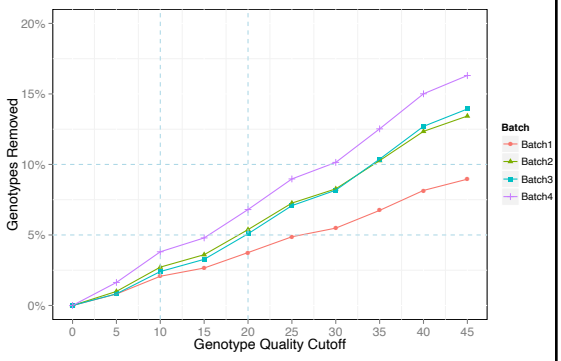
60

Genotypes Removed by DP (genotype) Cut-off by Batch



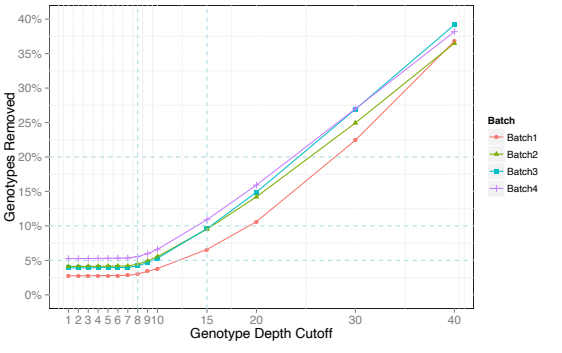
61

Genotypes Removed by GQ Cut-offs by Batch



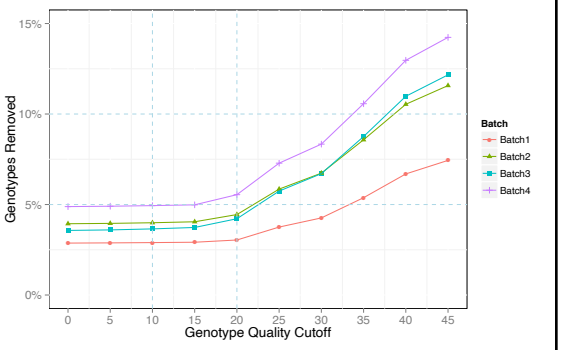
62

Genotypes Removed by DP (genotype) Cut-off by Batch (First removing genotypes with GQ ≤ 20)



63

Genotypes Removed by GQ Cut-offs by Batch (First removing genotypes with a DP≤8)



64

Missing Rate Criteria & Sites Removed

	Variant sites removed if missing >10% of their genotypes	Variant sites removed if missing >5% of their genotypes
	Percent of genotype data removed	
Before QC*	2.5%	3.9%
After QC	12.9%	18.3%

Variant sites missing >10% of their data were removed

*After VQSR

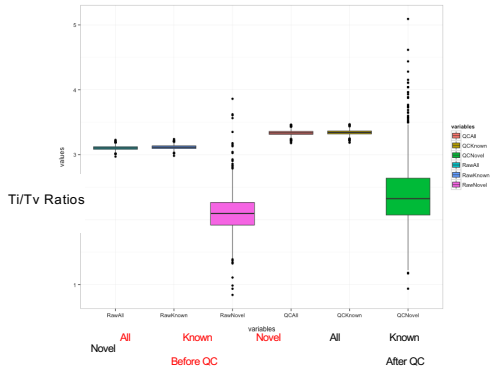
65

Ti/Tv Ratios during QC Process

	Known	Novel	All
Before VQSR	2.95 ± 0.05	1.18 ± 0.29	2.86 ± 0.07
Before additional QC	3.12 ± 0.03	2.01 ± 0.32	3.11 ± 0.03
Genotype QC DP≤8, GQ≤20	3.18 ± 0.04	2.10 ±0.32	3.16 ± 0.03
Remove sites missing >10% genotypes	3.39 ± 0.04	2.42 ± 0.52	3.39 ± 0.04
Remove batch specific novel sites ≥2 N=17,835	3.39 ± 0.04	2.41 ± 0.53	3.39 ± 0.04
Remove sites deviating from HWE $p \leq 5 \times 10^{-8}$ N=4,414	3.41 ± 0.04	2.39 ± 0.54	3.40 ± 0.04

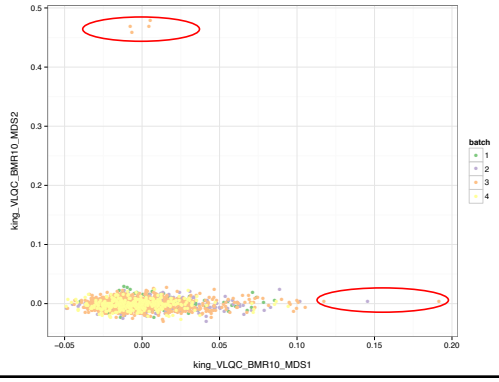
66

Ti/Tv Ratios by Individual Before and After QC



67

MDS First 2 Components After QC



68

Sequence Data QC

- Batch effects can sometimes be removed with additional QC
- Extreme outliers should be removed
- Additionally, MDS\PCA components can be included in the analysis to control for population substructure\admixture and batch effects
  - Unless correlated with the outcome (phenotype)
  - The MDS or PCA components should be recalculated after QC only including those samples included in the analysis
- Batch (dummy coding) may be included as a covariate in the analysis
  - Unless correlated with the outcome (phenotype)

69

Convenience Controls

- Can reduce the cost of a study
- Genotype data
- Type I error can be increased
  - Ascertainment from different population
  - Differential genotyping error
    - Even if performed at the same facility
- Proper QC can reduce or remove biases

70

Convenience Controls–Sequence Data

- Obtain BAM files and recall cases and control together
  - Can still have differential errors between cases and controls
  - Check variant frequency by variant types in cases and control
    - Synonymous variants should have the same frequencies
    - Would not expect large differences in numbers of variants between cases and controls
- For single variants can compare difference in frequencies with gnomAD but is problematic
  - Differences in frequencies can be due to differences in ancestry and/or sequencing errors
  - Cannot adjust for confounders
    - e.g., sex, population substructure/admixture
- Don't perform an aggregate test using frequency information obtained from databases, e.g., gnomAD, TOPMed Bravo

71

Genotype Array Data
Genotype Data QC – Population Based Studies

- Initially remove DNA samples from individuals who are missing >10% or their genotype data
- For variant sites with a minor allele frequency (MAF)≥0.05
  - Remove variants sites missing >5% of their genotype data
- For variant sites with a MAF<5%
  - Remove variant sites missing > 1% of their genotype data
- The genotypes for variant sites with missing data may have higher genotype error rates

72

Order of Data Cleaning-Genotype Array Data

- Remove samples missing >10% genotype data
- Remove SNPs with missing genotype data
  - If minor allele frequency >5%
    - Remove markers with >5% missing genotypes
  - If minor allele frequency <5%
    - Remove markers with >1% missing genotypes
- Remove samples missing >3% genotype calls
- Check genetic sex of individuals based on X-chromosome markers & Y chromosome marker data (if available)
  - Remove individual whose reported gender/sex is inconsistent with genetic data
    - Could be due to a sample mix-up
- Check for cryptic duplicates and related individuals
  - Used "trimmed data set of markers which are not in LD
    - e.g. r2<0.1
  - Remove duplicate samples

73

Order of Data Cleaning-Genotype Array

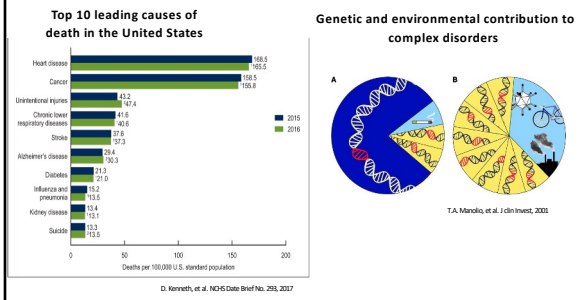
- Perform PCA or MDS to check for outliers
  - Use trimmed data set of markers which are not in LD
    - e.g., r2<0.1
  - First with unrelated individuals and then project related individuals on the components
  - Remove outliers from data
    - e.g., Mahalanobis distance
- Check for deviations from HWE
  - Separately in cases and controls
  - Only unrelated individuals
  - If more than one ancestry group
    - Separately for each ancestry group
      - As determined via PCA or MDS
- Examine QQ plots for potential problems with the data
  - e.g., not controlling adequately for population admixture

74

Complex Trait Association Analysis of Rare Variants Obtained from Sequence Data

1

Complex Diseases (Traits)

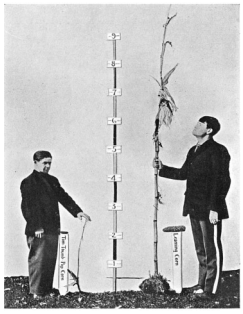


2

Heritability for Common Traits

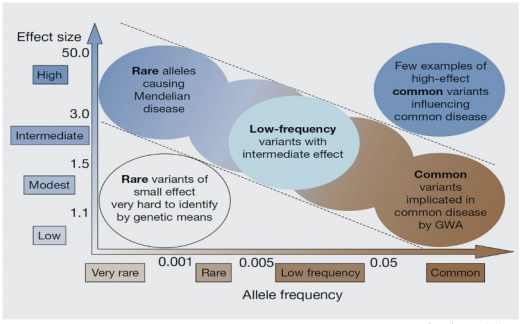
Human height heritability is ~80%

- Strongly associated common variation explain 21—29%
- All common variation explains 60% of height heritability



3

Allelic Architecture



4

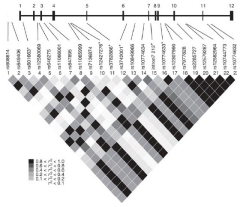
Complex Disease – Common Variant Associations

- Disease susceptibility is conferred by variants which are common within populations
  - Variants are old and widespread
- These variants have modest phenotypic effect
- This model is supported by a large number of replicated examples
  - Age Related Macular Degeneration (Klein et al. 2005)
    - Complement factor H (CFH) gene

5

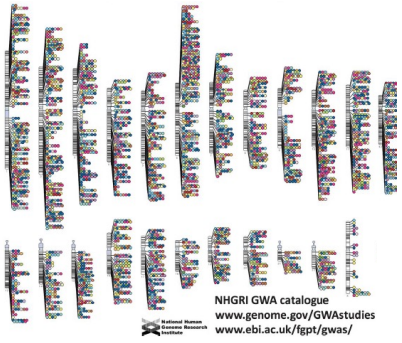
Studying Complex Traits – Common Variant Associations

- Hundreds of thousands of Single nucleotide polymorphism (SNPs) genotyped and analyzed
  - Indirect mapping
    - Markers usually had a minor allele frequency (MAF) > 0.05
    - Usually not pathogenic – tag SNPs
    - In linkage disequilibrium with disease susceptibility variant



6

Complex Trait – Common Variant Associations



- Although highly successful in identifying thousands of complex trait loci
- Usually pathogenic susceptibility variant(s) not identified

7

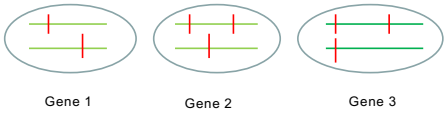
Complex Disease – Rare Variant Associations

- Complex traits are the result of multiple rare variants
  - Although first thought to large effects, there effect sizes are usually small
- Although these variants are rare, e.g., MAF<0.005
  - Collectively they may be quite common
- Direct tests of this hypothesis where first reported >15 years ago
  - Dallas Heart Study
    - Small sample ~1,200 individuals
      - Multi-ancestry
      - Used "extreme" sampling
    - Plasma low density lipoprotein levels (Cohen et al. 2004)
      - NPC1L1

8

Rationale for Rare Variant Aggregate Association Tests

- Testing individual variants with low effect sizes and minor allele frequencies (MAFs)
  - Underpowered to detect associations
- Testing variants in aggregate increases MAFs
  - Improving the power to detect associations



9

Caveats - Aggregate Rare Variant Association Tests

- Misclassification of variants can reduce power
  - Inclusion of non-causal variants
  - Exclusion of causal variants
- Analysis is limited to
  - Genes
  - Genes within pathways
- Analysis outside of exonic regions is problematic
  - Unlikely a sliding window approach will work
    - Size of window unknown and will differ across the genome
  - A better understanding of functionality outside the coding regions is necessary
    - Predicted functional regions, enhancer regions, transcription factors, DNase I hypersensitivity sites, etc.

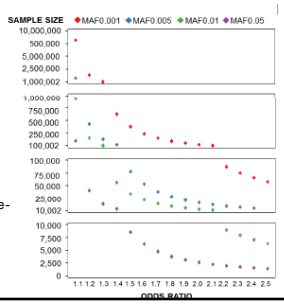
10

Analysis of Rare Variants

- For biobank sized datasets higher frequency rare variants, e.g., 0.5% can be analyzed individually
  - Using same same methods implemented for common variants

Example
$\alpha = 5 \times 10^{-8}$*
Disease prevalence 5%
$1-\beta = 0.80$

*Note: a more stringent significance criterion may be necessary for genome-wide sequence data. Due to a larger number of effective tests compared to analysis of common variant GWAS panels



11

A Few Rare Variant Association Tests

- Combined Multivariate Collapsing (CMC)
  - Li and Leal AJHG 2008
- Burden of Rare Variants (BRV)
  - Auer, Wang, Leal Genet Epidemiol 2013
- Weighted Sum Statistic (WSS)
  - Madsen and Browning PloS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Liu and Leal PloS Genet 2010
- Variable Threshold (VT)
  - Price et al. AJHG 2010

Fixed Effect Tests

- Sequence Kernel Association Test (SKAT)
  - Wu et al. AJHG 2011

Random Effect Test

- SKAT-0
  - Lee et al. AJHG 2012

Optimal test

12

Types of Aggregate Analyses

- Frequency cut offs used to determine which variants to include in the analysis
  - Rare Variants (e.g., MAF<0.05% frequency)
  - Rare and low (MAF=0.05-5%) frequency variants
- Maximization approaches
- Tests developed to detection associations when variants effects are bidirectional
  - e.g., protective and detrimental
- Incorporate weights based upon annotation
  - Frequency
    - e.g., gnomAD
  - Functionality
    - CADD c-scores

13

Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Combined multivariate & collapsing (CMC)
  - Li & Leal, AJHG 2008
- Collapsing scheme which can be used in the regression framework
  - Can use various criteria to determine which variants to collapse into subgroups
    - Variant frequency
    - Predicted functionality

14

CMC

- Define covariate Xj for individual j as

$$X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

- Compute Fisher exact test for 2x2 table

Number of cases for which one or more rare variants are observed e.g., nonsynonymous variants freq. ≤1%

	X=1	X=0
cases		
controls		

Number of cases without a rare variants

Number of controls for which one or more rare variants are observed

Number of controls without a rare variants

Can also use same coding in a regression framework

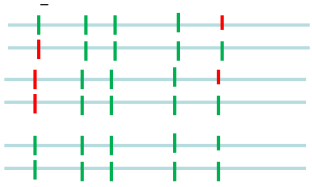
15

CMC

- Example of coding used in regression framework:
  - Binary coding

$$X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

  - Gene region with 5 variant sites
  -



Individual	Coding
1	1
2	1
3	0

Rare Variant Sites

Green bars: Major allele is observed in the study subject

Red bars: Minor allele has been observed

16

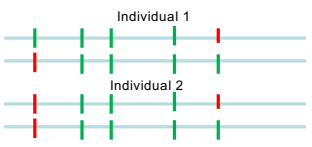
Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Gene-or Region-based Analysis of Variants of Intermediate and Low frequency (GRANVIL)
  - Aggregate number of rare variants used as regressors in a linear regression model
  - Can be extended to case-control studies
    - Morris & Zeggini 2010 Genet. Epidemiol
  - Test also referred to as MZ

17

GRANVIL

- Example of coding used in regression framework
  - Gene region with 5 variant sites – data available on all sites
  -



Coded 2/5 (0.4)

Coded 2/5 (0.4) Note same coding for heterozygous and homozygous genotypes

- Missing data for three of the five variant sites

Coded 1/2 (0.5)

Burden Rare Variant (BRV) extension (Auer et al. 2013 Genet Epidemiol)

Individual 1: Coded 2
Individual 2: Coded 3
Individual 3: Coded 1

18

Methods to Detect Rare Variant Associations Weighted Approaches

- Group-wise association test for rare variants using the Weighted Sum Statistic (WSS)
  - Variants are weighted inversely by their frequency in controls (rare variants are up-weighted)
    - Madsen & Browning, PLoS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Adaptive weighting based on multilocus genotype
    - Liu & Leal, PLoS Genet 2010

19

Methods to Detect Rare Variant Associations Maximization Approaches

- Variable Threshold (VT) method
  - Uses variable allele frequency thresholds and maximizes the test statistic
  - Can also incorporate weighting based on functional information
    - Price et al. AJHG 2010
- RareCover
  - Maximizes the test statistic over all variants with a region using a greedy heuristic algorithm
    - Bhatia et al. 2010 PLoS Computational Biology

20

Methods to Detect Associations with Protective & Detrimental Variants within a Region

- C-alpha
  - Detects variants counts in cases and controls that deviate from the expected binomial distribution
    - For qualitative traits only
      - Neale et al. 2011 PLoS Genet
- Sequence Kernel Association Test (SKAT)
  - Variance components score test performed in a regression framework
    - Can also incorporate weighting
  - Wu et al. 2011 AJHG

21

Optimal Test

- SKAT-O
  - Maximizes power by adaptively using the data to combine a burden test and the sequence kernel association tests
    - Lee et al. 2012 AJHG

22

Significance Level for Rare Variant Association Tests

- For exome data where individual genes are analyzed usually a Bonferroni correction for the number of genes tested is used
  - There is very little to no linkage disequilibrium between genes
- Bonferroni correction used
  - e.g., $p \leq 2.5 \times 10^{-6}$ (Correction for testing 20,000 genes)

23

Determine MAF Cut-offs for Aggregate Rare Variant Association Tests

- MAF cut-offs are frequently used to determine which variants to analyze in aggregate rare variant association tests
- MAF from controls should not be used
  - Increases in type I error rates
- Determine variant frequency cut-offs from databases
  - Using population frequencies for those understudy
  - gnomAD
    - http://gnomad.broadinstitute.org/

24

Problem of Missing Genotypes for Aggregate Rare Variant Association Tests

- Same frequency of missing variant calls in cases and controls
  - Decrease in power
- More variant calls missing for either cases or controls
  - Increase in Type I error
  - Decrease in power
- Remove variant sites which are missing genotypes, e.g., >10%
- Can impute missing genotypes using observed allele frequencies
  - For the entire sample
    - Not based on case or control status
- Analyze imputed data using dosages

25

Dosages

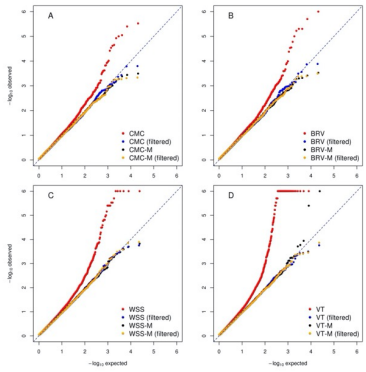
- Genotypes are no longer assigned 0 (1/1), 1 (1/2) or 2 (2/2)
  - Due to uncertainty
- Each genotype is assigned a probability
  - Probabilities sum to 1
- For example
  - Probability of 0 (1/1) genotype is 0.98 and 1 (1/2) genotype is 0.015
- The dosage can be estimated for this example as follows

0 x 0.98 = 0
1 x 0.015 = 0.015
2 x 0.005 = 0.01
Dosage = 0.025

- Instead of using the most likely genotype the dosage is used

26

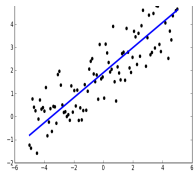
Results



27

Rare Variant Aggregate Methods

- Ideally should be performed in a regression framework to adjust for covariates
  - Logistic
  - Linear regression

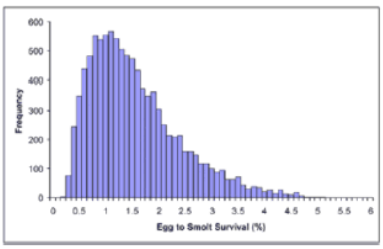


- Almost all rare variant aggregate methods have been extended to be implemented within a regression framework
- Some have also been implemented in a linear mixed model (LMM)/generalized LMM
  - Usually limited to fix effect tests

28

Analyzing Quantitative Variants

- Most rare variant aggregate analysis methods can be performed on quantitative traits
- If phenotype data includes outliers or deviates from normality
  - Can increase type I errors



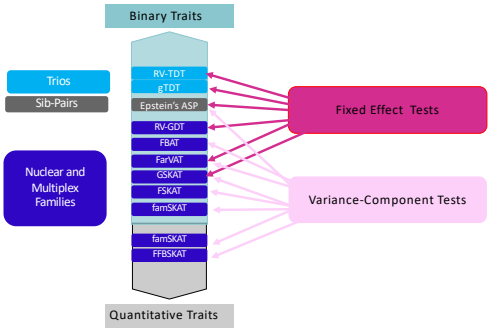
29

Analyzing Quantitative Variants

- For data that deviates from normality
  - Quantile-quantile normalization
- For data that includes outliers
  - Winsorize
- Don't winsorize and then normalize
- Instead of analyzing quantitative trait values
  - Residual can be generated
    - Adjusting for confounders

30

Family-based Methods for Rare Variant Aggregate Association Analysis



31

Linear Mixed Model (LMM) & generalized LMM (GLMM) Analysis of Related & Unrelated Individuals

- LMM is an extension of the linear model to allow for both fixed & random effects and also allows for non-independence of samples
  - Early implementations calculated the kinship matrix Φ on the basis of known relationships
  - Amin et al. (2007) proposed to estimate kinships based on genome-wide variant data
    - The generalized relationship matrix (GRM) can be estimated for all individuals using for example identical-by-descent (IBD) sharing
- Extended to binary (case-control) traits - GLMM

32

LMM and GLMM: Analysis of Related & Unrelated Individuals

- Can be applied to analyze families, cryptically related, & unrelated individuals
  - e.g., UK Biobank
    - 500K study subjects of which 30.3% are ≤3rd degree relatives & 4.5% sib-pairs
- More recent implementation for large scale data using a variety of methods
  - BOLT-LMM (Loh et al. 2015)
  - FastGWA (Jiang et al. 2019)
  - SAIGE (Zhao et al. 2015)*
  - REGENIE (Mbatchou et al. 2020) *
  - SMMAT (Chen et al. 2019)**
- *Can be used to analyze data where case to control ratio is very unbalanced
  - e.g., 20 cases for every control
- **Cannot be used for UK Biobank Scale data

33

LMM and GLMM: Analysis of Related & Unrelated Individuals

- To allow for use with very biobank sized data
- REGENIE does not use the GRM
  - It uses whole genome regression, i.e., the ridge regression
    - In essence, it includes all the SNVs as covariates in the null model
      - Performed by blocks to avoid having to load the entire genome in memory
        - » Using different effect size differences per block
- This large-scale approximation may not control type I error for individuals that are closely related
  - e.g., when only families are being analyzed
  - Can use for example SMMAT
    - Which uses the GRM

34

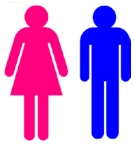
LMM and GLMM: Analysis of Related & Unrelated Individuals

- A few programs which can perform rare variant aggregate analysis
  - REGENIE - Burden test
  - SMMAT - Burden, SKAT, & SKAT-O tests
  - rvtests (Zhan 2020) implements BOLT-LMM to perform burden association analysis
- An alternative for rare variant aggregate analysis
  - Recode variants within gene regions and then analyze

35

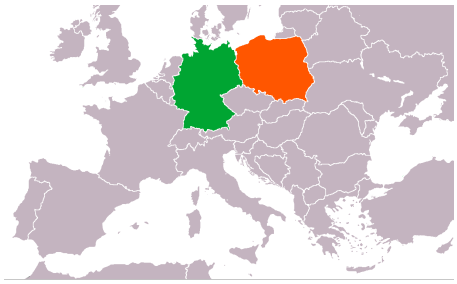
Rare Variant Aggregate Methods

- Can control for covariates in the analysis which are potential confounders
  - Age
  - Sex
  - Batch
  - Body Mass Index (BMI)
  - Smoking pack years



36

Confounder -Population Substructure and Admixture



37

Population Substructure and Admixture

- If proportion of cases and controls sampled from each population is different
  - Can occur due to
    - Disease frequency is different between populations
    - Sloppy sampling
- Population substructure\admixture can cause detection of differences in variant frequencies within a gene which is due to sampling and not disease status
  - False positive findings can be increased

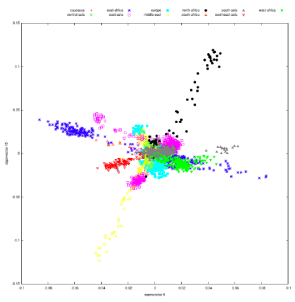
38

Example River People

39

Population Substructure and Admixture

- Currently PCA or MDS are use to control for population substructure\admixture
  - Controls on the global level
  - May not be sufficient in particular for admixed populations



40

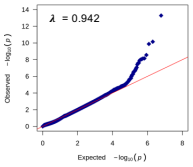
Rare Variant Aggregate Association Analysis

- When analyzing different populations, e.g.,
  - Africans
  - Europeans
- When analyzing data from different source
  - Analyze each group separately
- Meta-analysis can be used to combine the results from each group

41

Rare Variant Aggregate Methods

- Best to obtain principal components to include in the regression model (including LMM and GLMM)
  - using variants which are not in LD e.g., r^2<0.1 (pruned)
  - covering a wide range of the allelic frequency spectrum e.g., >0.1%
  - Evaluate how many components need to be included
    - Don't include a fix number of components
      - e.g., 5 or 10 components
- Success of PCA\MDS in controlling for population substructure\admixture can be evaluated through lambda and examining Quantile-Quantile (QQ) plots



42

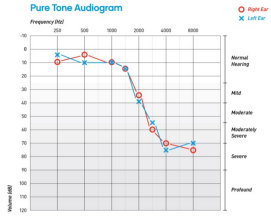
Part II
Example of a Rare Variant Association Study

Analysis of UK Biobank Exome Data to Study the Etiology of Late-onset Hearing Impairment

43

Age-related Hearing Impairment (ARHI) (aka Presbycusis)

- ARHI can impact quality of life and daily functioning
- ARHI is one of the most common adult conditions
  - In the USA
    - ARHI affects 50% of individuals >75 years of age
    - It is estimated that 30-40 million will be affected with significant ARHI by 2030



44

Goals of the Study

- Using data from the UK Biobank to detect associations between self-reported measures of ARHL and genetic variants
  - **H-aid** self-reported hearing aid use (f.3393: "Do you use a hearing aid most of the time?")
  - **H-diff** self-reported hearing difficulty (f.2247: "Do you have any difficulty with your hearing?")
  - **H-noise** self-reported hearing difficulty with background noise (f.2257: "Do you find it difficult to follow a conversation if there is background noise e.g., TV, radio, children playing)?
  - **H-both** individuals with both H-diff and H-noise
- With an emphasis of understanding the role that rare variation plays in ARHL
  - Current analysis - exome sequence data

45

UK Biobank

- 500,000 individuals randomly sampled
  - Aged 40-69 at time of enrollment
    - To be followed for at least 20 years
    - Predominantly white Europeans
      - Also includes South Asians and individuals of African Ancestry and smaller number of individuals of a few other ancestries
- Extensive phenotype data
  - Qualitative and quantitative traits
    - ICD-10 and ICD-9 codes
    - Self reports
    - Cognitive test
    - Brain MRIs
    - NMR-metabolomics data
- Genetic Data
  - Genotype and imputed data
  - Exome sequence data
  - Whole genome sequence data
    - 200K currently available
    - Remining sample - Quarter 1 2023
  - Telomere length data

*Data showcase can be used to examine phenotypes and sample sizes available

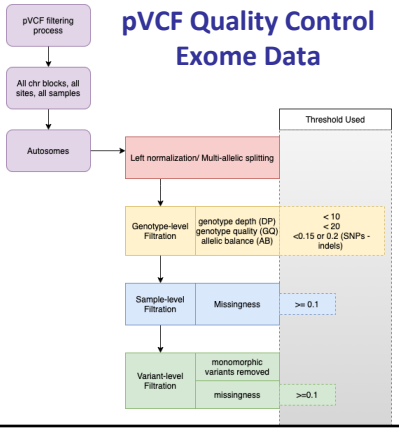
46

Genetic Data Analyzed

- Exome data
  - ~200,000 participants
- Imputed variant data (secondary replication sample for common variants)
  - ~300,000 participants
    - Did not have exome data at the time of the study

47

pVCF Quality Control Exome Data



48

Filtering Strategy

N=200,619
Individuals with exomes and phenotype data

↓

N=200,386
Individuals consistent for sex

↓

N=189,001
Individuals self-report to be of European ancestry

↓

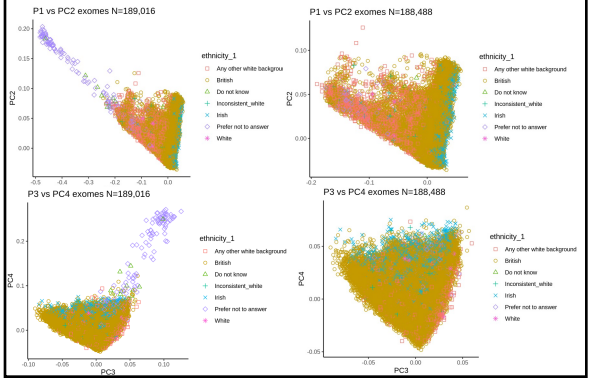
N=187,908
After removal of outliers based on PCA analysis – Mahalanobis distance (≤0.997)

↓

N=180,318
After applying phenotype exclusion criteria (ICD10, ICD9, & self-report)

49

Principal Components Analysis and Exclusion of Outliers



50

Exclusion Criteria
Obtained from ICD10, ICD9, & Self Report

- Deafness
- Early-onset hearing impairment
- Otosclerosis
- Meniere's
- Labyrinthitis
- Disorders of acoustic nerve
- Bell's palsy
- History of chronic suppurative and nonsuppurative otitis media
- Meningitis
- Encephalitis, myelitis, and encephalomyelitis
- Etc.

51

Filtering Strategy

N=200,619
Individuals with exomes and phenotype data

↓

N=200,386
Individuals consistent for sex

↓

N=189,001
Individuals self-report to be of European ancestry

↓

N=187,908
After removal of outliers based on PCA analysis – Mahalanobis distance (≤0.997)

↓

N=180,318
After applying phenotype exclusion criteria (ICD10, ICD9, & self-report)

52

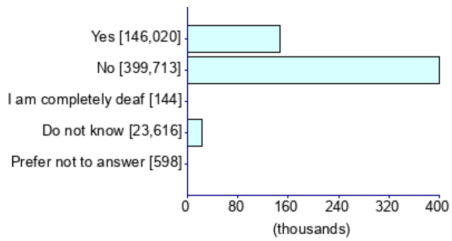
Defining Cases and Controls

- Based on answers obtained from a touch screen
- Cases - self-reported hearing difficulty
  - f.2247: "Do you have any difficulty with your hearing?"
- Controls - did **not** have any self-reported hearing problems
  - ***H-aid*** hearing aid use (f.3393)
  - ***H-diff*** self-reported hearing difficulty (f.2247)
  - ***H-noise*** self-reported hearing difficulty with background noise (f.2257)

53

Hearing difficulty/problems -Data field 2247

569,977* items of data are available, covering **498,704** participants



*Due to repeat visits

54

Repeat measures*

- Individuals with inconsistent answers removed

	Visit 1	Visit 2	Visit 3	Visit 4	
Study subject A	Problems Hearing	No Hearing Problems	No Hearing Problems	No Hearing Problems	Inconsistent Remove
Study subject B	No Hearing Problems	No Hearing Problems	Problems Hearing	Problems Hearing	Consistent (Case)
Study subject C	No Hearing Problems	No Hearing Problems	No Hearing Problems	No Hearing Problems	Consistent (Control)

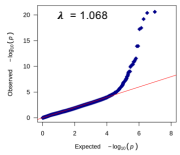
*Majority of study subjects currently have data from only one visit

55

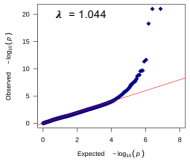
Analysis of Exome Data

- Analysis performed using generalized linear mixed models (GLMM) (REGENIE)
  - To control for inclusion of related individuals
    - For the UK Biobank data 30.3% of participants are ≤3rd degree relatives & 4.5% sib-pairs
  - Genotype array data (~800K) were used for the ridge regression
    - Data pruned to remove variants with a r^2>0.1
      - Using exome data for the ridge regression led to an an inflated lambda value

QQ Plot using exome data for ridge regression



QQ Plot using genotype data for ridge regression



56

Analysis of Exome Data

- Analysis limited to individuals of white European Ancestry
- Sex, age, and two PCAs included as covariates
  - Age for cases first report of hearing difficulty & controls age at last visit
  - The PCAs where recalculated for only individuals included in the analysis
    - Using the pruned genotypes array data (r2<0.1)

57

Analysis of Exome data – Single Variant

- All variants with four or more alternative alleles observed in the sample analyzed
  - A very low minor allele frequency was used since it was hypothesized some of the variants may have large effect sizes

58

Analysis of Exome data – Single Variant

- Discovery sample
  - Second release of 150K exome
- Replication sample
  - First release of 50K exomes
- Entire exome sample (200K)
- Secondary Replication Sample*
  - To replicate findings from the entire exome sample
  - Genotype and Imputed data (Haplotype Reference Consortium Panel)
    - 300K individuals who were not included in the exome data
      - Imputed variants with an INFO score > 0.3 were analyzed

*Only used for replication of common variants

59

Significance Levels

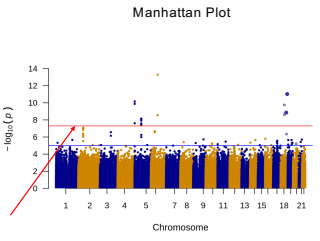
- Discovery sample
  - A genome-wide significance level was used to reject the null hypothesis of no association
    - $p \leq 5.0 \times 10^{-8}$
- Replication sample
  - Permutation was used to obtain empirical p-values
    - Adjusting for the phenotypes and variants brought to replication
      - $p \leq 0.05$

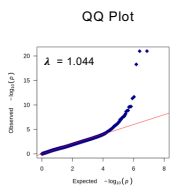
For the replication it is not necessary to use a genome-wide significance level of 5×10^{-8} for single variant tests or 2.5×10^{-6} for gene-based rare variant aggregate analysis. Significance level is adjusted for the number of variants/genes tested in the replication sample

- Bonferroni correction
- Estimate empirical p-values

60

Hearing Difficulty - Data Field 2247



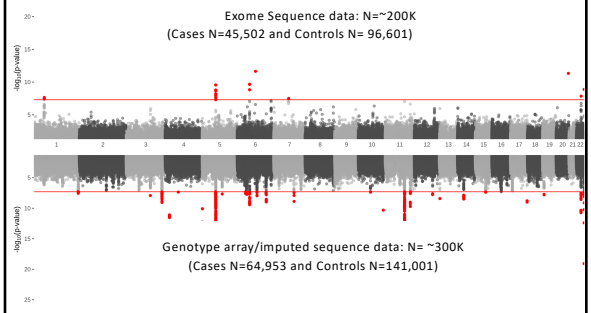


Genome-wide significance level 5 x 10^{-8} (red line)

Cases N=45,502
Controls N= 96,601

61

Hudson Plot Discovery and Replication Hearing Difficulty Data Field 2247



62

Analysis of Exome Data Rare Variant Aggregate Analysis

- Genes with at least two variants were analyzed, e.g., predicated loss of function (pLoF) variants
- Max coding was used
- Two masks were used
  - Mask 1 – pLoF variants
  - Mask 2 – pLoF and missense variants
- Minor allele frequency cut-off of <0.01 was used
  - The frequencies for each variant site were obtained from gnomAD non-Finnish Europeans

63

REGENIE Rare Variant Aggregate Analysis

- Three different codes can be used
  - Max
  - Sum
  - Comphet
    - This term is not correct because the phase is unknown
      - Variants may be on the same haplotype

Single variant sites	max	sum	comphet
00000000000000 →	0	0	0
00000100010000 →	1	2	2
00201011010100 →	2	7	2

https://rgcgithub.github.io/regenie/options/

64

Selection of Variants to Include in Rare Variant Aggregate Association Tests

Annotation File		Mask File		AAF file
1:55039839:T:C PCSK9 LoF 1:55039842:G:A PCSK9 missense	+	Mask1 LoF Mask2 LoF,missense	+	1:55039839:T:C 1.53e-05 1:55039842:G:A 2.19e-06
1:55039839:T:C PCSK9 CADD30 1:55039842:G:A PCSK9 CADD20	+	Mask1 CADD score > 30 Mask2 CADD score > 20	+	1:55039839:T:C 1.53e-05 1:55039842:G:A 2.19e-06

REGENIE will use information from the annotation and alternative allele frequency (AAF) files to build the Masks (variants to be included in the association testing)

65

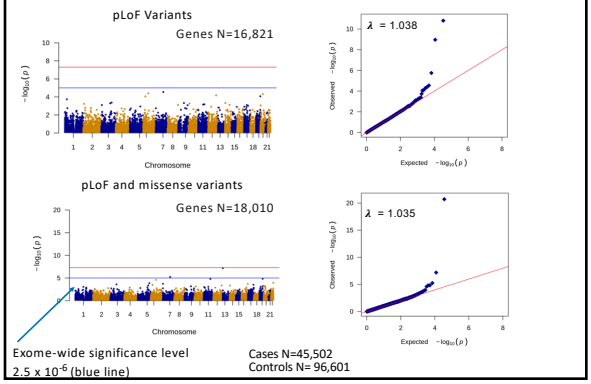
Analysis of Exome Data Rare Variant Aggregate Analysis

- Exome sample was split
  - Second release of 150K exome were used as the discovery sample.
  - First release of 50K exome were used as the replication sample
- Entire exome sample (200K) was also analyzed*
- Discovery sample significance level
  - $p \leq 2.5 \times 10^{-6}$
    - 0.05/20.000 Bonferroni correction for testing 20,000 genes
- Replication sample significant level
  - $p \leq 0.05$
  - Empirical p-values generated
    - Permutation used to adjust for the number of phenotypes and genes brought to replication (pLoF and pLOF & missense)

*No replication sample available for these findings

66

Hearing Difficulty - Data Field 2247

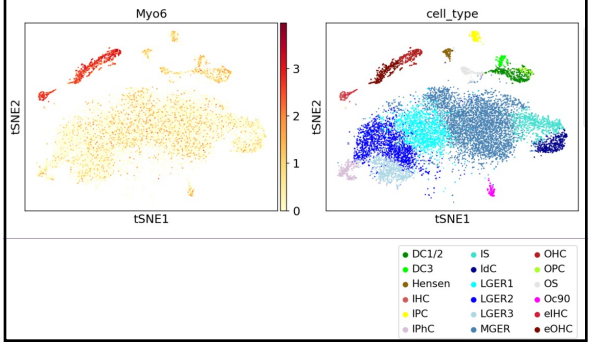


Exome-wide significance level
2.5×10^{-6} (blue line)

Cases N=45,502
Controls N= 96,601

67

Expression - gEAR



68

Results

- Replicated some previously reported ARHL genes
  - Some which had not been previously replicated
    - e.g., *BAIAP2L2, CRIP3, KLHDC7B, MAST2,* and *SLC22A7*
- Identified and replicated a new HL gene which has not been previously reported
  - Inner ear expression in humans and mice supports the involvement of gene in HL etiology
- Rare-variant aggregate analysis demonstrated the important contribution of Mendelian HL genes, i.e. *MYO6, TECTA*, and *EYA4* the genetics of ARHL

69

Results

- Rare variants for ARHL tend to have larger effect sizes than those for common variants
  - Rare variants should play an important role in risk prediction by increasing accuracy
- Although most of the studies findings were replicated in independent samples of white Europeans
  - Additional studies are necessary to elucidate whether these variants/genes play a role in the genetic etiology of ARHL in other populations

70

Linkage disequilibrium in genetic association studies

Gao Wang, Ph.D.

Advanced Gene Mapping Course, November 2022

The Gertrude H. Sergievsky Center and Department of Neurology
Columbia University Vagelos College of Physicians and Surgeons

Genetic association studies

Identify genetic variants **associated** with **complex traits**

- Association does not imply causality
- Disease, quantitative traits, molecular phenotypes

in order to

- Understand biological mechanism
- Identify potential drug targets
- Identify individuals with high disease risk

Sources of association signals

Causal association — meaningful

- Tested genetic variations influence traits directly

Linkage disequilibrium (LD) — useful

- Tested genetic variations associated with other nearby variations that influence traits
- Meaningful or misleading, in different contexts

Population stratification — misleading

- Tested genetic variations is unrelated to traits, but is associated due to sampling differences
- eg, minor allele frequency, disease prevalence

Sources of association signals: analysis tools

Causal association — meaningful

- Fine-mapping, colocalization, Mendelian randomization

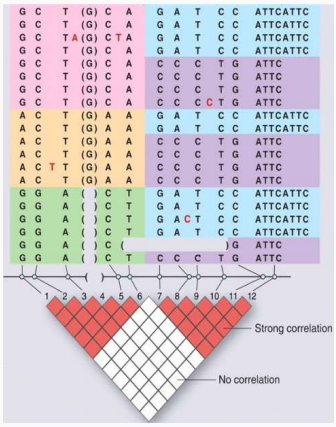
Linkage disequilibrium (LD) — useful

- Meaningful: LD scores regression, polygenic risk scores (PRS), transcriptome-wide association studies (TWAS)
- Misleading: fine-mapping, LD pruning / clumping

Population stratification — misleading

- Principle component analysis, linear (mixed) models

LD in human genome is pervasive

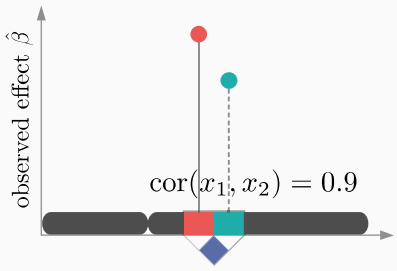


Altshuler *et al.* (2008)

Impact of LD on GWAS analysis

Oligogenic: trait influenced by a few genetic variants

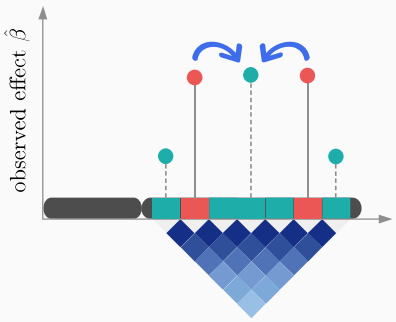
- Misleading: difficult to identify causal variants
- Useful: 'tag SNPs' in array based GWAS design



Impact of LD on GWAS analysis

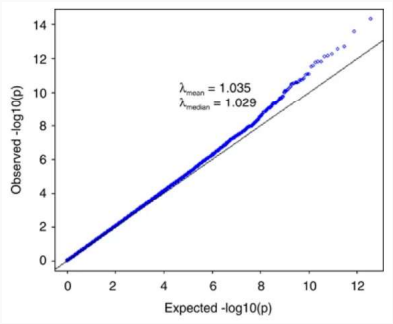
Polygenic: trait influenced by numerous genetic variants

- Misleading: stronger association due to more LD 'friends'
- Useful: whole-genome prediction with sparse models



A second thought on genomic inflation

Population stratification? Or, polygenic inheritance + LD?



Suggested reading: Yang et al (2011) EJHG

LD score regression (LDSC)

LD score regression model without population stratification

Chi-square GWAS statistic of variant j

Sample size

Narrow sense heritibility

LD score of variant j

Total number of variants

$$\mathrm{E}[\chi_j^2] = 1 + \frac{Nh_g^2}{M} l_j$$

$$l_j = \sum_{k \neq j} r_{jk}^2$$

LD score: sum of squared Pearson's correlation coefficient between SNP j and other (neighboring) SNPs

LD score regression (LDSC)

Separating h_g^2 and population stratification

Population stratification factor

$$\mathrm{E}[\chi_j^2] = N\alpha + 1 + \frac{Nh_g^2}{M} l_j$$

Regression intercept

Regression slope

LD score of variant j

A more powerful and accurate correction factor for GWAS summary statistics compared to genomic control approach.

- Bulik-Sullivan et al (2015) Nature Genetics — the LDSC regression paper
- Zhu and Stephens (2017) AoAS — a neat, alternative LDSC regression model derivation in supplemental material

LDSC application: heritability estimation

Narrow sense heritibility

- Proportion of phenotypic variation explained by additive genetic factors

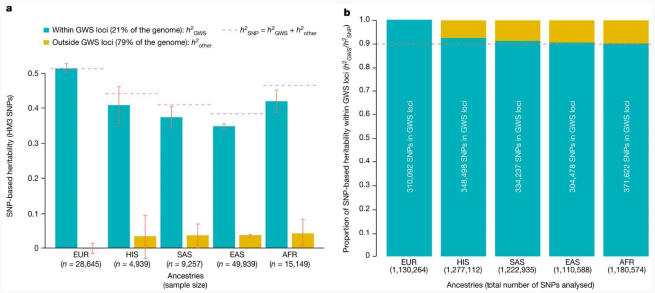
Estimation strategy

- Pedigree design: genetic covariance and IBD sharing
- Population design: linear mixed models

Population design, summary statistics

- LDSC to estimate SNP-based heritability
- Stratified LDSC (S-LDSC) to partition heritability by functional annotations

Variance of height explained in GWAS



Yengo *et al.* (2022) Nature

Fine-mapping with summary statistics: current methods and practical considerations

Gao Wang, Ph.D.

Advanced Gene Mapping Course, November 2022

The Gertrude H. Sergievsky Center and Department of Neurology
Columbia University Vagelos College of Physicians and Surgeons

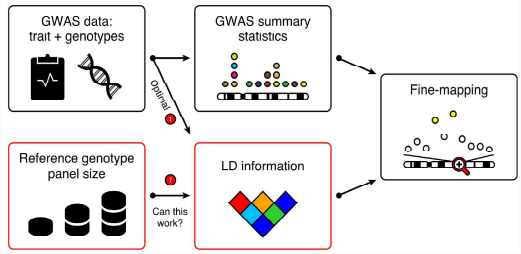


Figure: Benner *et al.* (2017) Am. J. Hum. Genet.

Association analysis summary statistics

z-scores from univariate association studies:

$$\hat{z}_j := \hat{\beta}_j / s_j,$$

where

$$\hat{\beta}_j := (\boldsymbol{x}_j^\intercal \boldsymbol{x})^{-1} \boldsymbol{x}^\intercal \boldsymbol{y} \quad s_j := \sqrt{\hat{\sigma}_j^2 (\boldsymbol{x}_j^\intercal \boldsymbol{x})^{-1}}$$

- **Sufficient** statistics: $\boldsymbol{x}^\intercal \boldsymbol{x}, \boldsymbol{x}^\intercal \boldsymbol{y}, \hat{\sigma}_j^2$
- **"Summary"** statistics:
  - z-scores: $\hat{\boldsymbol{z}}$
  - Genotypic correlation: $\hat{\boldsymbol{R}}$

Reasons to work with summary statistics

Advantage over full data (genotypes and phenotypes):

- Easier to obtain and share with others
- Convenient to use: QC and data wrestling barely needed
- Computationally suitable for large-sample fine-mapping
  - $\mathcal{O}(p^2)$ (summary statistics) $\ll \mathcal{O}(np)$ (full data)
  - when sample size $n \gg$ variants in fine-mapped region p

Suggested reading: Pasaniuc and Price (2017) Nat. Rev. Genet.

Regression with Summary Statistics (RSS)

$$\hat{z} \sim N(\hat{R}z, \hat{R})$$

Assumptions:

1. Heritability of any single SNP is small
2. $\hat{R}$ is sample genotypic correlation for **the same study**
3. Genotypes used to computed z and $\hat{R}$ are accurate

Properties of per SNP z scores

- z-score for a SNP depends on effects of both itself and other correlated SNPs:

$$\mathrm{E}(\hat{z}_j|\hat{R}) = \sum_{i=1}^{p} r_{ij} z_j.$$

GWAS marginal effects are biased due to LD!

- z-scores are correlated,

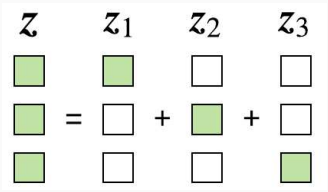
$$\mathrm{Cor}(\hat{z}_j, \hat{z}_k) = r_{jk}, \forall j, k$$

- Recall the previously discussed connection with LDSC

Fine-mapping via RSS model

"Single effect": z_l's

$$\hat{z} \sim N(\hat{R}z, \hat{R})$$
$$z = \sum_{l=1}^{L} z_l$$
$$z_l = \gamma_l z_l$$
$$z_l \sim N(0, \omega_l^2)$$
$$\gamma_l \sim \mathrm{Mult}(1, \pi)$$



Suggested reading:

Zou et al (2022) PLoS Genet.

Alternative models of GWAS summary statistics

The $\hat{z}$ model:

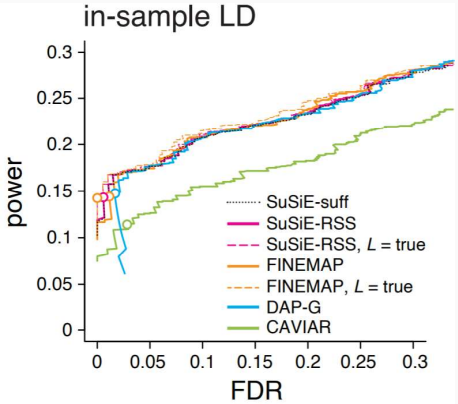
$$\hat{z} \sim N(\hat{R}z, \hat{R})$$

The $\hat{b}, \hat{s}$ model:

$$\hat{b}|\hat{\mathbf{s}} \sim N(\hat{\mathbf{S}}\hat{R}\hat{\mathbf{S}}^{-1}b, \hat{\mathbf{S}}\hat{R}\hat{\mathbf{S}})$$

- Both models can be easily written as SuSiE regression
  - $\hat{z}$ model: lower MAF variants have larger effects
  - $\hat{b}, \hat{s}$ model: effect sizes are the same regardless of MAF
- Fine-mapping using $\hat{z}$ model: CAVIAR, FINEMAP
- Fine-mapping using $\hat{b}, \hat{s}$ model: DAP-G

Summary statistics methods comparison



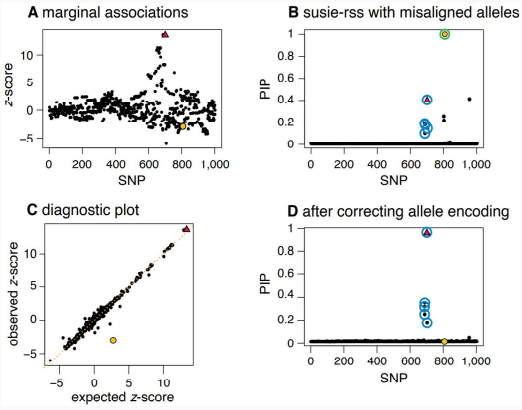
Zou *et al.* (2022) PLoS Genet.

Impact of allele flips

What is allele flip?

- Different allele encoding between GWAS and LD reference
- *e.g.* AA=0, AC=1, CC=2 in GWAS; AA=2, AC=1, CC=0 in LD reference genotype
- A challenging problem coupled with strand flip, when merging sequence data from different platforms

Impact of allele flips

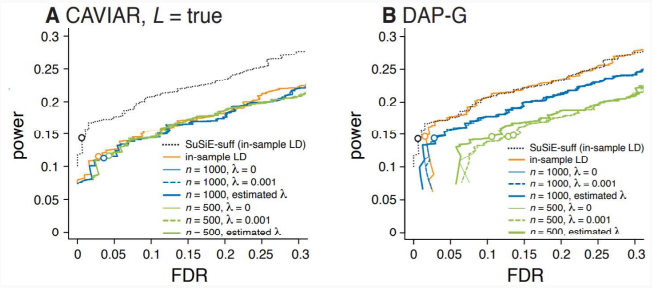


Zou *et al.* (2022) PLoS Genet.

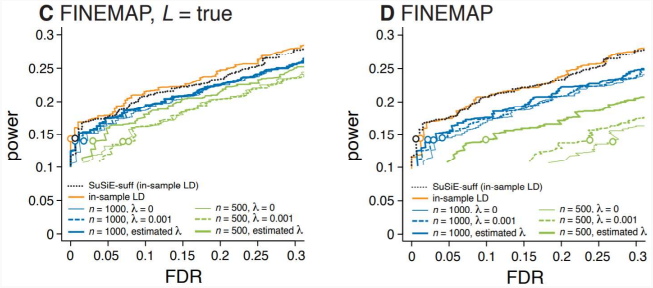
Addressing the allele flip challenge

- `susieR::susie_rss()` function implements a diagnosis
- `bigsnpr::snp_match()` function implements a basic allele matching for two sets of summary statistics
- Other resources
  - Allele flip illustration: `https://statgen.us/lab-wiki/compbio_tutorial/allele_qc`
  - A powerful, multi-set data merger (by Yin Huang): `https://cumc.github.io/xqtl-pipeline/pipeline/misc/summary_stats_merger.html`

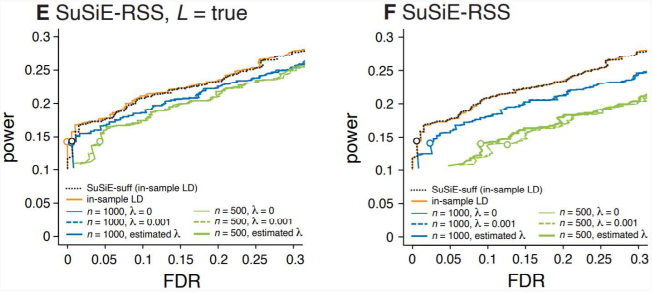
Impact of mis-matched LD reference: PIP



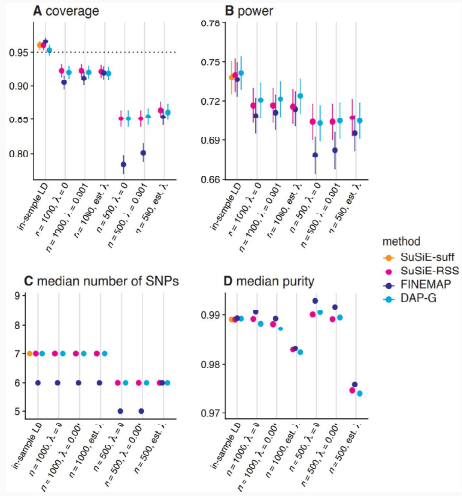
Impact of mis-matched LD reference: PIP



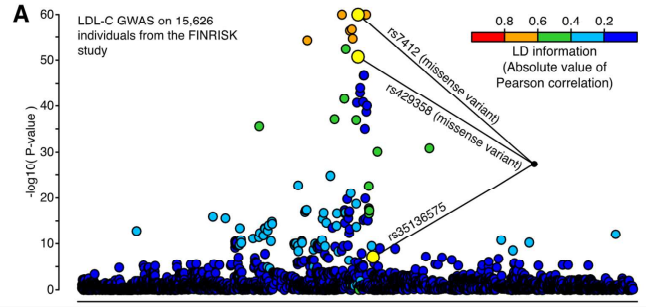
Impact of mis-matched LD reference: PIP



Impact of mis-matched LD reference: credible sets

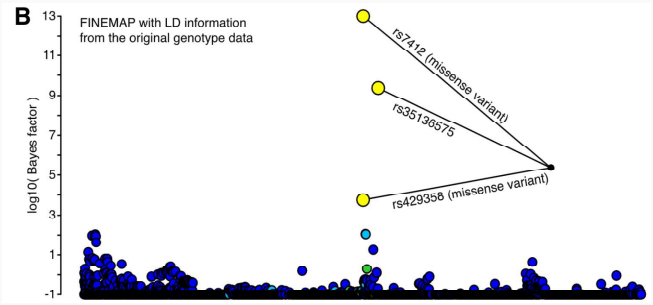


Impact of mis-matched LD reference: real data



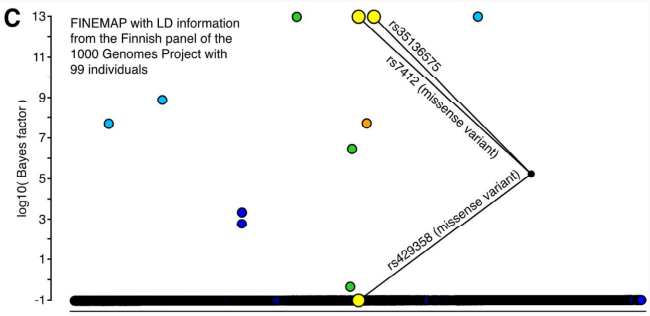
Benner *et al.* (2017) Am. J. Hum. Genet.

Impact of mis-matched LD reference: real data



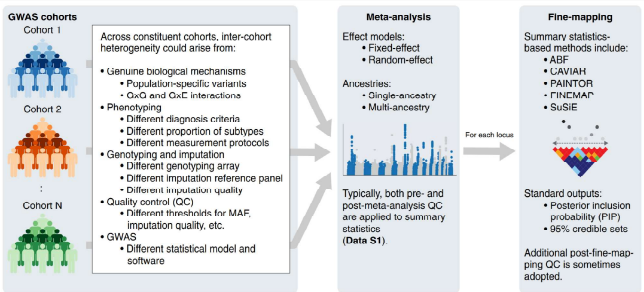
Benner *et al.* (2017) Am. J. Hum. Genet.

Impact of mis-matched LD reference: real data



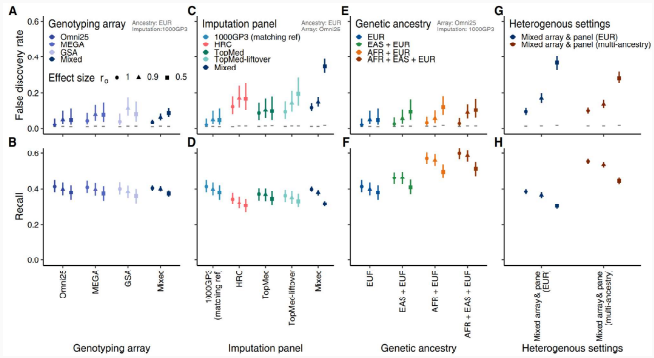
Benner *et al.* (2017) Am. J. Hum. Genet.

Fine-mapping in meta-analysis: overview



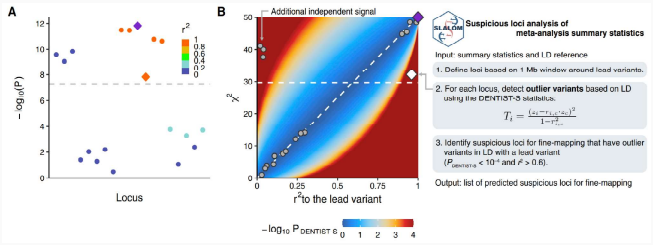
Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: key factors



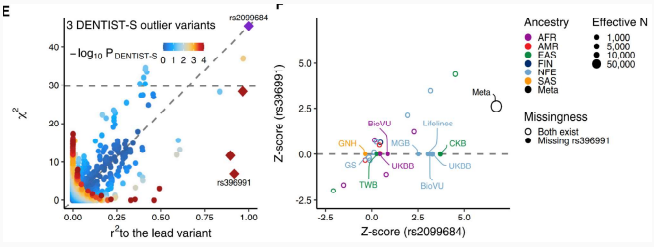
Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: diagnosis



Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: diagnosis



Kanai *et al.* (2022) Cell Genomics

Covariate adjustment in LD reference

Consider two GWAS regression analysis:

1. Evaluate SNP effect in Trait $\sim$ SNP+Age+Sex+PCs
2. Fit model Trait $\sim$ Age+Sex+PCs, compute residual of Trait (remove covariates), and evaluate SNP effect in model Residual_Trait $\sim$ SNP

Are these two analysis equivalent?

Covariate adjustment in LD reference

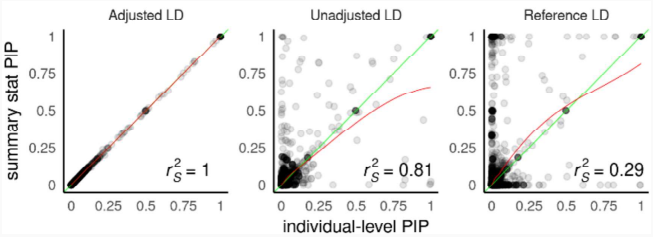
Consider two GWAS regression analysis:

1. Evaluate SNP effect in Trait $\sim$ SNP+Age+Sex+PCs
2. Fit model Trait $\sim$ Age+Sex+PCs, compute residual of Trait (remove covariates), and evaluate SNP effect in model Residual_Trait $\sim$ SNP

They are not equivalent because covariates should also be removed from SNP data: Residual_Trait $\sim$ Residual_SNP

Covariate adjustment in LD reference

Covariates should be removed from genotype before computing LD reference for fine-mapping



Quick *et al.* (2020) biorxiv

Fine-mapping with summary statistics: current methods and practical considerations

Gao Wang, Ph.D.

Advanced Gene Mapping Course, November 2022

The Gertrude H. Sergievsky Center and Department of Neurology
Columbia University Vagelos College of Physicians and Surgeons

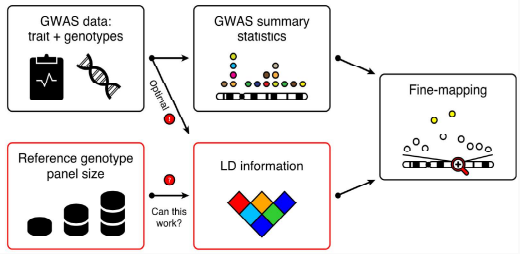


Figure: Benner *et al.* (2017) Am. J. Hum. Genet.

Association analysis summary statistics

z-scores from univariate association studies:

$$\hat{z}_j := \hat{\beta}_j / s_j,$$

where

$$\hat{\beta}_j := (\boldsymbol{x}_j^\intercal \boldsymbol{x})^{-1} \boldsymbol{x}^\intercal \boldsymbol{y} \quad s_j := \sqrt{\hat{\sigma}_j^2 (\boldsymbol{x}_j^\intercal \boldsymbol{x})^{-1}}$$

- **Sufficient** statistics: $\boldsymbol{x}^\intercal \boldsymbol{x}, \boldsymbol{x}^\intercal \boldsymbol{y}, \hat{\sigma}_j^2$
- **"Summary"** statistics:
  - z-scores: $\hat{\boldsymbol{z}}$
  - Genotypic correlation: $\hat{\boldsymbol{R}}$

Reasons to work with summary statistics

Advantage over full data (genotypes and phenotypes):

- Easier to obtain and share with others
- Convenient to use: QC and data wrestling barely needed
- Computationally suitable for large-sample fine-mapping
  - $\mathcal{O}(p^2)$ (summary statistics) $\ll \mathcal{O}(np)$ (full data)
  - when sample size $n \gg$ variants in fine-mapped region p

Suggested reading: Pasaniuc and Price (2017) Nat. Rev. Genet.

Regression with Summary Statistics (RSS)

$$\hat{z} \sim N(\hat{R}z, \hat{R})$$

Assumptions:

1. Heritability of any single SNP is small
2. $\hat{R}$ is sample genotypic correlation for **the same study**
3. Genotypes used to computed z and $\hat{R}$ are accurate

Properties of per SNP z scores

- z-score for a SNP depends on effects of both itself and other correlated SNPs:

$$\mathrm{E}(\hat{z}_j|\hat{R}) = \sum_{i=1}^{p} r_{ij} z_j.$$

GWAS marginal effects are biased due to LD!

- z-scores are correlated,

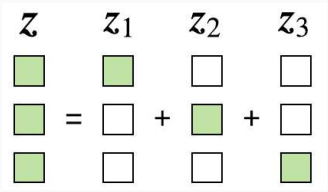
$$\mathrm{Cor}(\hat{z}_j, \hat{z}_k) = r_{jk}, \forall j, k$$

- Recall the previously discussed connection with LDSC

Fine-mapping via RSS model

"Single effect": z_l's

$$\hat{z} \sim N(\hat{R}z, \hat{R})$$
$$z = \sum_{l=1}^{L} z_l$$
$$z_l = \gamma_l z_l$$
$$z_l \sim N(0, \omega_l^2)$$
$$\gamma_l \sim \mathrm{Mult}(1, \pi)$$



Suggested reading:

Zou et al (2022) PLoS Genet.

Alternative models of GWAS summary statistics

The $\hat{z}$ model:

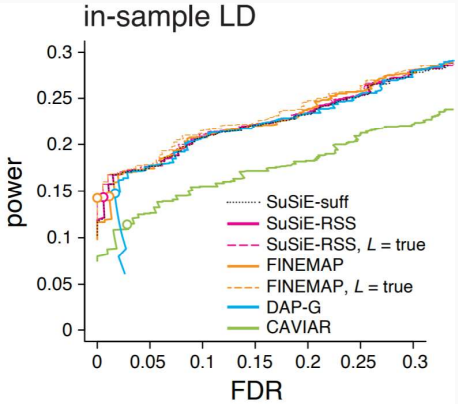
$$\hat{z} \sim N(\hat{R}z, \hat{R})$$

The $\hat{b}, \hat{s}$ model:

$$\hat{b}|\hat{\mathbf{s}} \sim N(\hat{\mathbf{S}}\hat{R}\hat{\mathbf{S}}^{-1}b, \hat{\mathbf{S}}\hat{R}\hat{\mathbf{S}})$$

- Both models can be easily written as SuSiE regression
  - $\hat{z}$ model: lower MAF variants have larger effects
  - $\hat{b}, \hat{s}$ model: effect sizes are the same regardless of MAF
- Fine-mapping using $\hat{z}$ model: CAVIAR, FINEMAP
- Fine-mapping using $\hat{b}, \hat{s}$ model: DAP-G

Summary statistics methods comparison



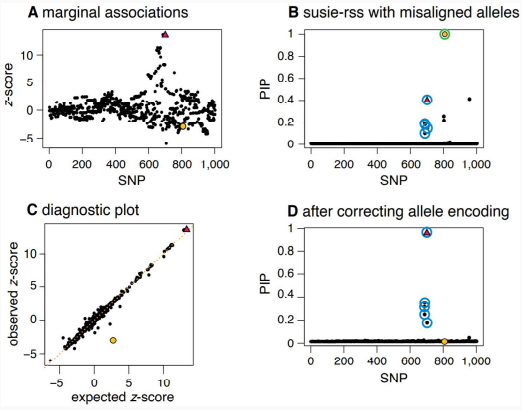
Zou *et al.* (2022) PLoS Genet.

Impact of allele flips

What is allele flip?

- Different allele encoding between GWAS and LD reference
- *e.g.* AA=0, AC=1, CC=2 in GWAS; AA=2, AC=1, CC=0 in LD reference genotype
- A challenging problem coupled with strand flip, when merging sequence data from different platforms

Impact of allele flips

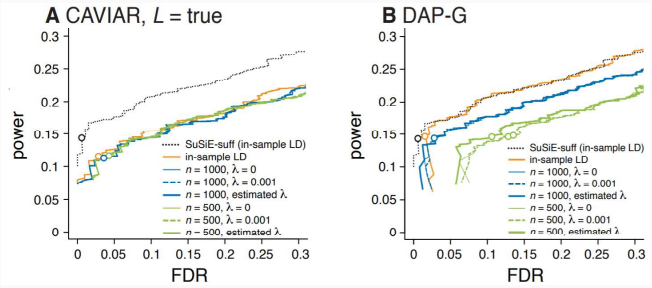


Zou *et al.* (2022) PLoS Genet.

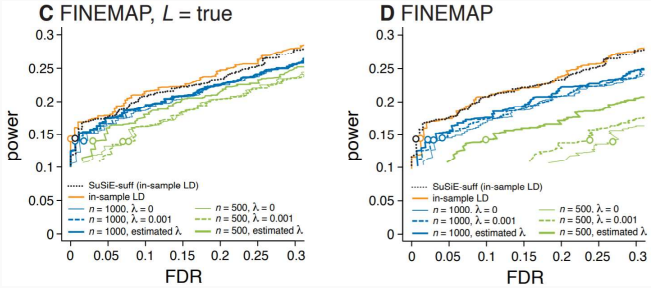
Addressing the allele flip challenge

- `susieR::susie_rss()` function implements a diagnosis
- `bigsnpr::snp_match()` function implements a basic allele matching for two sets of summary statistics
- Other resources
  - Allele flip illustration: `https://statgen.us/lab-wiki/compbio_tutorial/allele_qc`
  - A powerful, multi-set data merger (by Yin Huang): `https://cumc.github.io/xqtl-pipeline/pipeline/misc/summary_stats_merger.html`

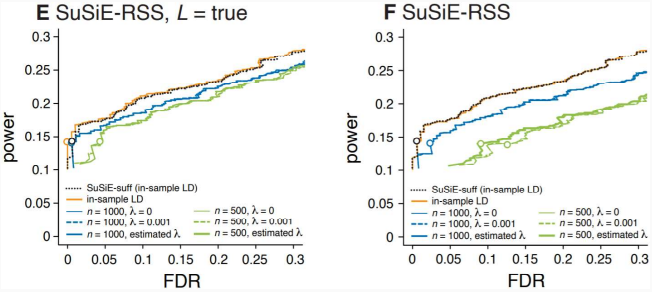
Impact of mis-matched LD reference: PIP



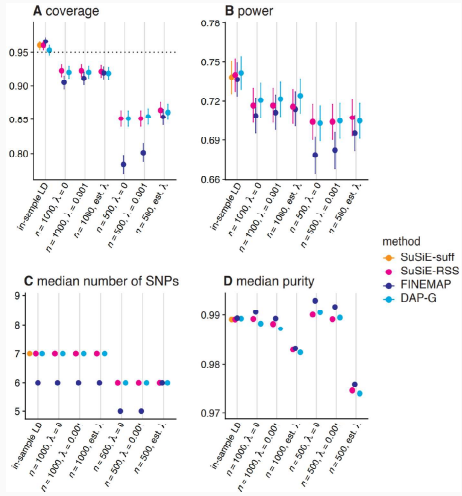
Impact of mis-matched LD reference: PIP



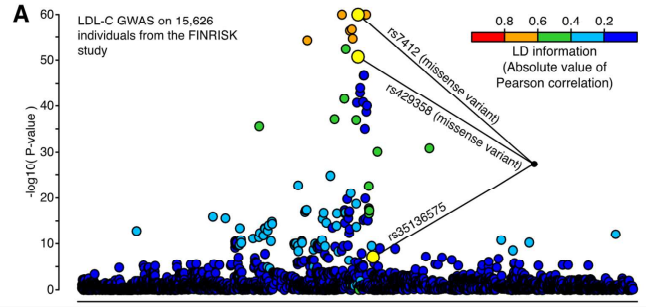
Impact of mis-matched LD reference: PIP



Impact of mis-matched LD reference: credible sets

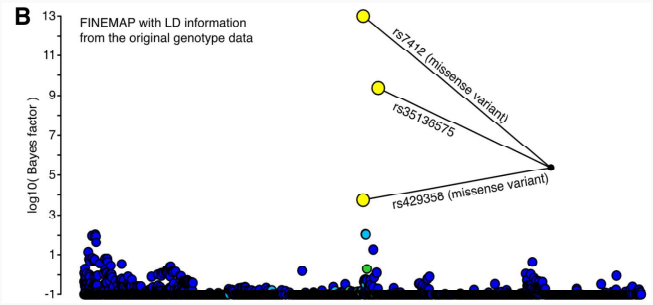


Impact of mis-matched LD reference: real data



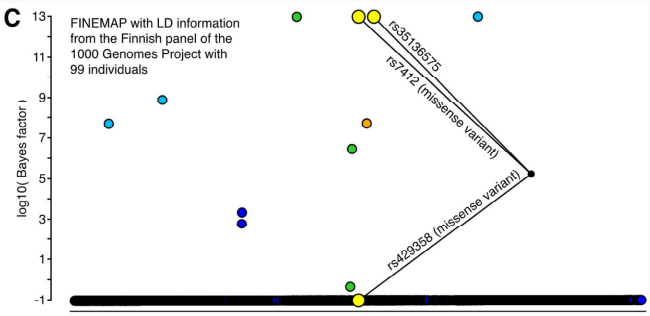
Benner *et al.* (2017) Am. J. Hum. Genet.

Impact of mis-matched LD reference: real data



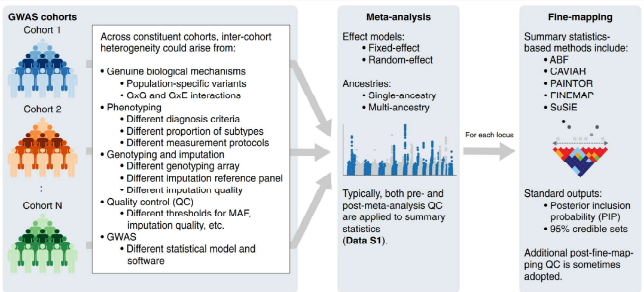
Benner *et al.* (2017) Am. J. Hum. Genet.

Impact of mis-matched LD reference: real data



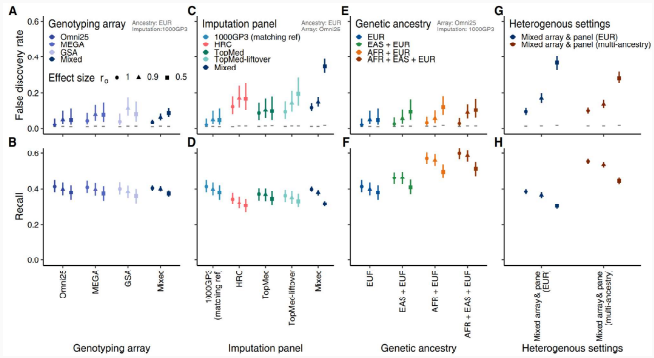
Benner *et al.* (2017) Am. J. Hum. Genet.

Fine-mapping in meta-analysis: overview



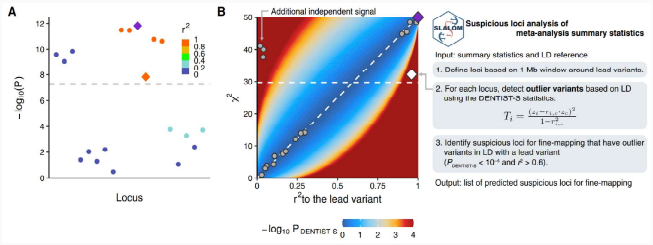
Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: key factors



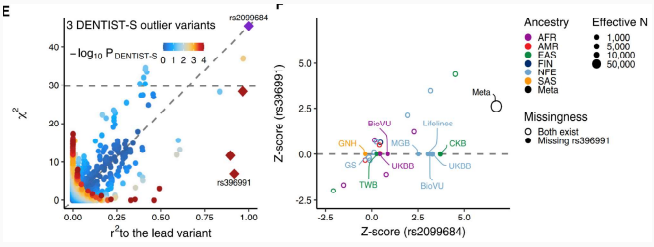
Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: diagnosis



Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: diagnosis



Kanai *et al.* (2022) Cell Genomics

Covariate adjustment in LD reference

Consider two GWAS regression analysis:

1. Evaluate SNP effect in Trait $\sim$ SNP+Age+Sex+PCs
2. Fit model Trait $\sim$ Age+Sex+PCs, compute residual of Trait (remove covariates), and evaluate SNP effect in model Residual_Trait $\sim$ SNP

Are these two analysis equivalent?

Covariate adjustment in LD reference

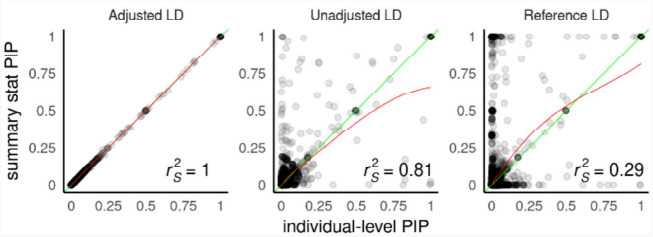
Consider two GWAS regression analysis:

1. Evaluate SNP effect in Trait $\sim$ SNP+Age+Sex+PCs
2. Fit model Trait $\sim$ Age+Sex+PCs, compute residual of Trait (remove covariates), and evaluate SNP effect in model Residual_Trait $\sim$ SNP

They are not equivalent because covariates should also be removed from SNP data: Residual_Trait $\sim$ Residual_SNP

Covariate adjustment in LD reference

Covariates should be removed from genotype before computing LD reference for fine-mapping



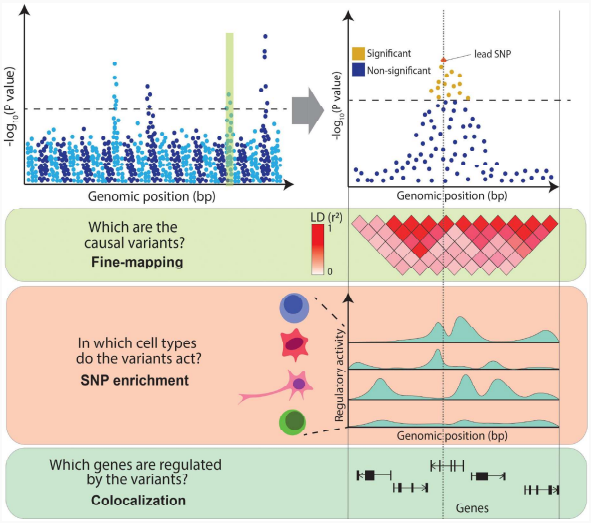
Quick *et al.* (2020) biorxiv

Integrating GWAS with functional annotations

Gao Wang, Ph.D.

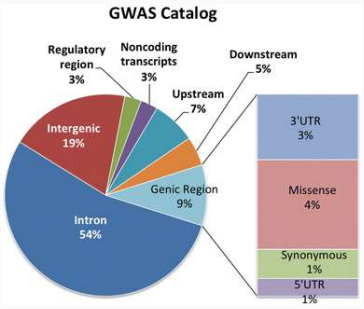
Advanced Gene Mapping Course, November 2022

The Gertrude H. Sergievsky Center and Department of Neurology
Columbia University Vagelos College of Physicians and Surgeons



GWAS variants catelog by functional annotations

Most GWAS variants are non-coding



Lee *et al.* (2018) *Human Genetics*

Functional enrichment in fine-mapped variants

Signals concentrated in tissue / cell specific functional area

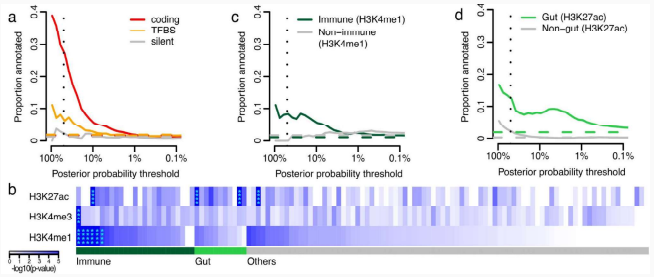


Figure: Huang *et al.* (2017) Nature

Functional annotation in aggregated rare variant association analysis

Functional annotation filters in aggregated tests

Aggregated tests are sensitive to (mis-)classification of functional variants. Different sets can be evaluated in practice:

- Loss of function: start-loss, stop-gain, splice sites
- Damaging missense: start-loss, stop-gain, splice sites, nonsynonymous with REVEL score > 0.5
  - Ioannidis et al (2016) AJHG
- All: start-loss, stop-gain, splice sites, nonsynonymous

Annotations integrated to aggregated tests

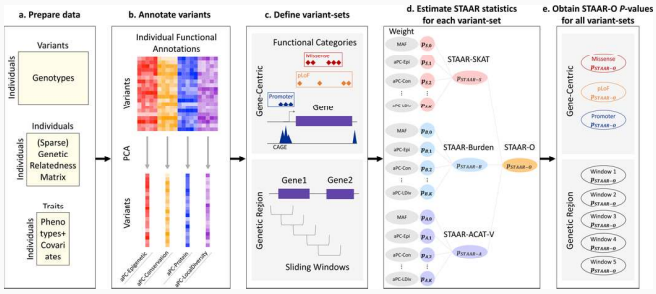


Figure: Li *et al.* (2020) Nature Genetics

Also see Li *et al.* (2022) Nature Methods

Annotations integrated to aggregated tests

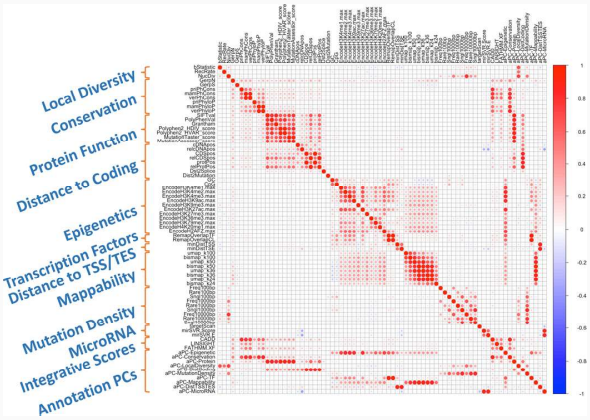


Figure: Li *et al.* (2020) Nature Genetics

Functional annotation in common variant association analysis

A polygenic model: stratified LD score regression

Chi-square GWAS statistic of variant j

$$\mathrm{E}[\chi_j^2] = 1 + \frac{Nh_g^2}{M} l_j$$

Sample size

Narrow sense heritibility

LD score of variant j

Total number of variants

$$l_j = \sum_{k \neq j} r_{jk}^2$$

LD score: sum of squared Pearson's correlation coefficient between SNP j and other (neighboring) SNPs

A polygenic model: stratified LD score regression

Chi-square GWAS statistic of variant j

$$\mathrm{E}[\chi_j^2] = 1 + \frac{Nh_g^2}{M} l_j$$

Sample size

Narrow sense heritibility

LD score of variant j

Total number of variants

$$l_j = \sum_{k \neq j} r_{jk}^2$$

LD score: sum of squared Pearson's correlation coefficient between SNP j and other (neighboring) SNPs

- Perform LDSC restricted to a functional category
- **Enrichment:** The proportion of SNP-heritability in the category divided by the proportion of SNPs

Cell-type enrichment in GWAS traits via S-LDSC

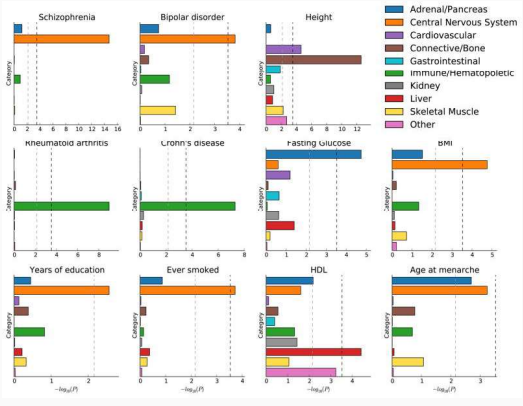


Figure: Finucane *et al.* (2015) Nature Genetics

Integration approaches

- Integrate directly as range based binary annotations
  - Finucane et al (2015) Nature Genetics — Stratified LDSC paper
- Possibility to work with variant specific continuous annotations
  - Gazal et al (2017) Nature Genetics
- Compute variant level annotations from epigenomic feature ranges
  - Deep Learning methods
  - Zhou et al (2015) Nature Genetics, Zhou et al (2018) Nature Genetics

A sparse model (a somewhat oligogenic view)

Generalized linear model for SNP effects given K annotations

$$\beta_j = (1 - \pi_j)\delta_0 + \pi_j g(\Theta)$$

$$\pi_j := \Pr(\gamma_j = 1 | \boldsymbol{\alpha}, \boldsymbol{d})$$

$$\log\left[\frac{\pi_j}{1-\pi_j}\right] = \alpha_0 + \sum_{k=1}^{K} \alpha_k d_{kj}$$

α are **log fold enrichment** of functional genomic features

- Suggested reading: Wen (2016) AoAS

Enrichment of DNase I in GTEx eQTLs

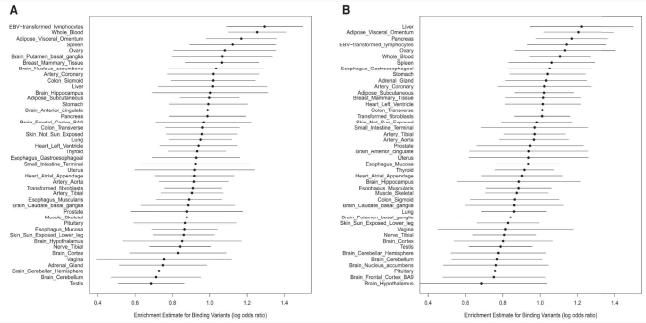
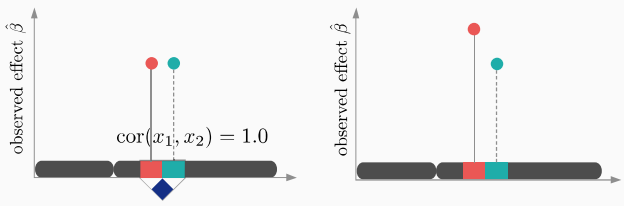


Figure: Wen *et al.* (2016) AJHG

Integrative fine-mapping with functional annotations

Annotations improves fine-mapping resolution



Integrating functional information prioritizes the left SNP.

Recall the toy example

Probability of association assuming **one effect variable**,

$$\frac{\mathrm{LR}_1}{\mathrm{LR}_1 + \mathrm{LR}_2} = 0.87 \quad \frac{\mathrm{LR}_2}{\mathrm{LR}_1 + \mathrm{LR}_2} = 0.13$$

Recall the toy example

Probability of association assuming **one effect variable**,

$$\frac{\mathrm{LR}_1}{\mathrm{LR}_1 + \mathrm{LR}_2} = 0.87 \quad \frac{\mathrm{LR}_2}{\mathrm{LR}_1 + \mathrm{LR}_2} = 0.13$$

What if we determine *a priori* that SNP 1 is **twice as important** as SNP 2?

$$\frac{2 \times \mathrm{LR}_1}{2 \times \mathrm{LR}_1 + \mathrm{LR}_2} = 0.93 \quad \frac{\mathrm{LR}_2}{2 \times \mathrm{LR}_1 + \mathrm{LR}_2} = 0.07$$

Fine-mapping with functional annotations

Recall the BVSR model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \\ \boldsymbol{e} &\sim N(0, \sigma^2 I_n) \\ \gamma_j &\sim \text{Bernoulli}(\boldsymbol{\pi}) \\ \boldsymbol{b}_\gamma | \boldsymbol{\gamma} &\sim g(\cdot) \\ \boldsymbol{b}_{-\gamma} | \boldsymbol{\gamma} &\sim \delta_0 \end{aligned}$$

Key idea: π, prior inclusion probability, can be modelled by **enrichment** of functional annotations

Genome-wide approach with S-LDSC

- A single locus may not have enough power to leverage annotation enrichment
- Genome-wide evaluation of thousands of annotations can increase power of fine-mapping
  - Lead to new loci to discover
- Functional enrichment can be done under the same framework
  - Prioritize genomic features / tissues / cell-types
- **Enrichment coefficient may be transferrable cross population**
  - Weissbrod *et al.* (2021) medrxiv

Functionally informed fine-mapping in UK Biobank

In analyses of 49 UK Biobank traits, PolyFun + SuSiE identified >32% more fine-mapped variant–trait pairs compared to using SuSiE alone.

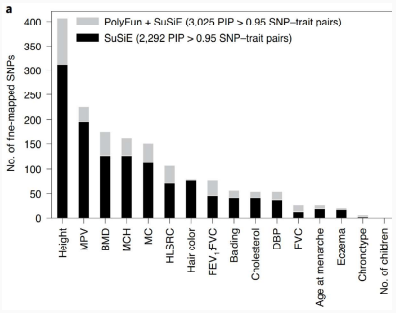


Figure: Weissbrod *et al.* (2020) Nat. Genet.

Example: *SuSiE* with functional informed prior

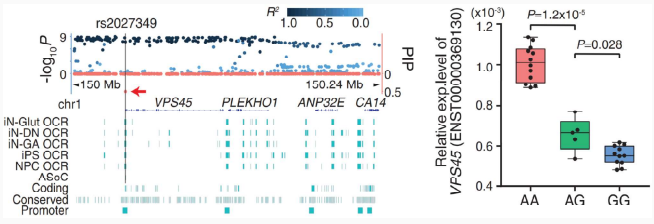


Figure: Zhang *et al.* (2020) Science

Caution: disease specific enrichment

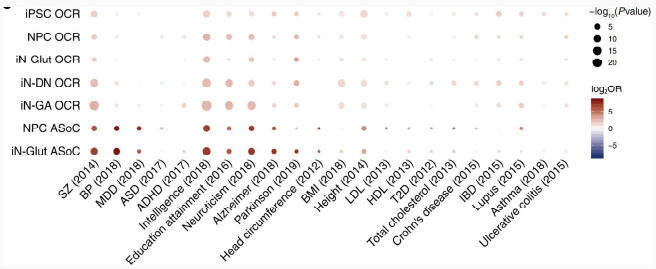


Figure: Zhang *et al.* (2020) Science

Complex phenotype prediction and transcriptome-wide association studies

Gao Wang, Ph.D.

Advanced Gene Mapping Course, November 2022

The Gertrude H. Sergievsky Center and Department of Neurology
Columbia University Vagelos College of Physicians and Surgeons

1. Rationale and assumptions
2. Univariate TWAS methods (credits: Haky Im @ UChicago)
3. Multivariate TWAS methods
4. Connections between TWAS and fine-mapping, colocalization and Mendelian Randomization

Rationale and assumptions

Motivation: eQTLs are enriched in GWAS signals

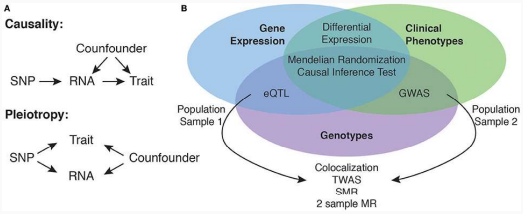


Figure: Heinig (2018) Front. Cardiovasc. Med.

Transcriptome-wide association study (TWAS)

Contributions of multiple genetic variants to complex traits through their impact on molecular phenotypes

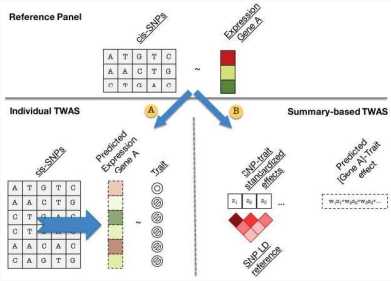


Figure: Gusev *et al.* (2016) Nat. Genet.

TWAS challenge: association vs causality

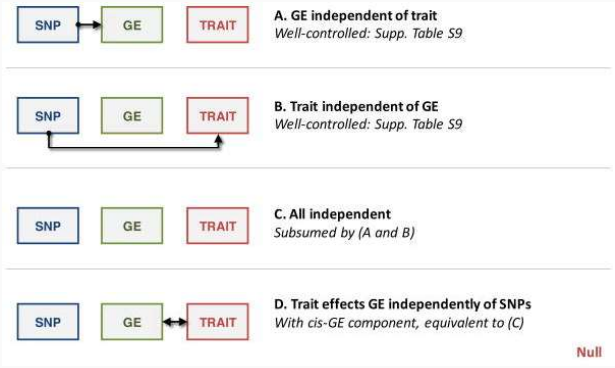


Figure: Gusev *et al.* (2016) Nat. Genet.

TWAS challenge: association vs causality

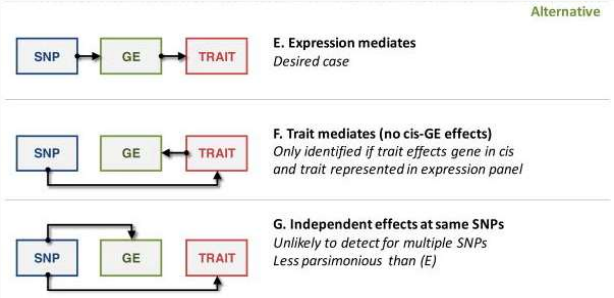


Figure: Gusev *et al.* (2016) Nat. Genet.

TWAS challenge: technical considerations

Ideal TWAS setup

- Homogenous population
- Tissue and cell-type specific
- Training data-set is large and complete ($N > 200$)

But in reality

- Cross population TWAS aplications
- Multiple tissue and cell-types
- Availability of individual level data vs summary statistics

TWAS methods overview

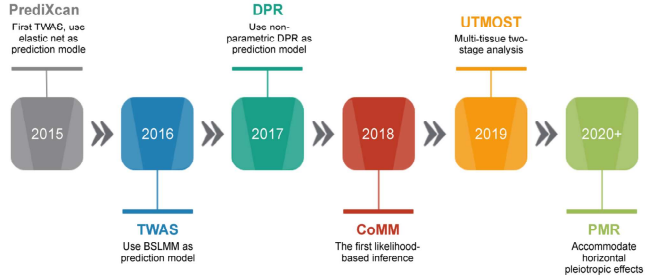


Figure: Zhu and Zhou *et al.* (2020) Quantitative Biology

Univariate TWAS methods (credits: Haky Im @ UChicago)

Univariate TWAS methods overview

$$Y = \sum_{k=1}^{M} \beta_k X_k + \epsilon$$

Univariate Regression → GWAS

$$\|Y - X_k\beta_k\|_2$$

Penalized regression → Ridge, LASSO, Elastic Net

$$\|Y - \sum_k X_k\beta_k\|_2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta_2\|_2$$

These methods can also be used for Polygenic Risk Score (PRS) calculations

Simple regression method

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\mathrm{GWAS}} X_k$$

Univariate Regression → GWAS

Ridge regression / BLUP

REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

AJHG 2011

Penalized regression

Ridge

$$Y - \sum_{k=1}^{M} \hat{\beta}_k^{\text{Ridge}} X_k$$

$$\|Y - \sum_k X_k \beta_k\|_2 + \qquad \lambda_2 \|\beta_2\|_2$$

11

Other penalized regression

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie
Stanford University, USA

Penalized regression

LASSO

Elastic Net

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\text{E-N}} X_k$$

$$\|Y - \sum_k X_k \beta_k\|_2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta_2\|_2$$

12

Bayesian variable selection regression

OPEN ACCESS Freely available online PLOS | GENETICS

Polygenic Modeling with Bayesian Sparse Linear Mixed Models

Xiang Zhou[1]*, Peter Carbonetto[1], Matthew Stephens[1,2]*

$$Y = \sum_{k=1}^{M} \beta_k^L X_k + \sum_{k=1}^{M} \beta_k^S X_k + \epsilon$$

$$\beta_k^L \sim N(0, \sigma_L^2)$$

$$\beta_k^S \sim N(0, \sigma_S^2)$$

MultiBLUP: improved SNP-based prediction for complex traits

Doug Speed and David J Balding

Genome Res. published online June 24, 2014
Access the most recent version at doi:10.1101/gr.169375.113

13

Choice of methods: cross validation

TWAS / FUSION

Functional Summary-based Imputation

New! RWAS (Grishin et al.) models for TCGA ATAC-seq

New! CONTENT (Thompson et al.) context-specific models for single-cell and bulk expression

New! GTEx v8 models

FUSION is a suite of tools for performing transcriptome-wide and regulome-wide association studies (TWAS and RWAS). FUSION builds predictive models of the genetic component of a functional/molecular phenotype and predicts and tests that component for association with disease using GWAS summary statistics. **The goal is to identify associations between a GWAS phenotype and a functional phenotype that was only measured in reference data.** We provide precomputed predictive models from multiple studies to facilitate this analysis.

Please cite the following manuscript for TWAS methods:

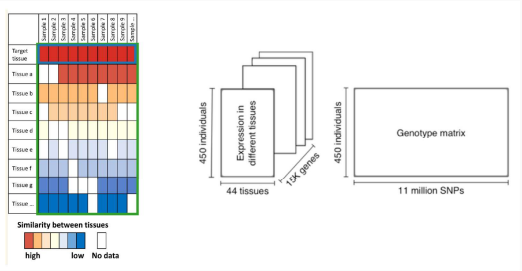
Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies" 2016 *Nature Genetics*

14

Multivariate TWAS methods

Multivariate TWAS methods overview

Leverage similarity between molecular phenotypes



- UTMOST, Yu *et al.* (2019) Nature Genetics
- MR-JTI, Zhou *et al.* (2020) Nature Genetics

15

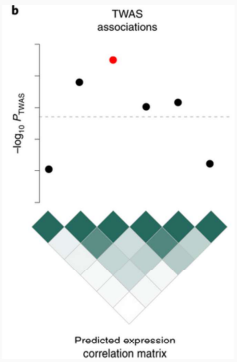
Multivariate TWAS hands-on exercise

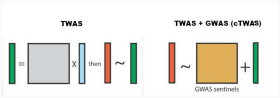
```
statgen-setup launch --tutorial twas
```

16

Connections between TWAS and fine-mapping, colocalization and Mendelian Randomization

TWAS and fine-mapping: variable selection






Article | Published: 29 March 2019

Probabilistic fine-mapping of transcriptome-wide association studies

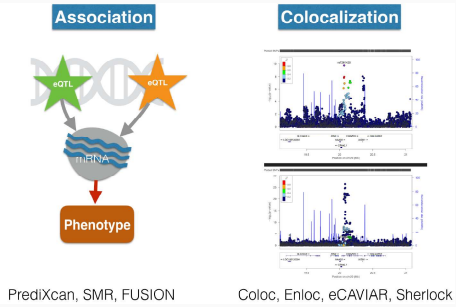
Nicholas Mancuso, Malika K. Freund, Ruth Johnson, Huwenbo Shi, Gleb Kichaev, Alexander Gusev & Bogdan Pasaniuc

Nature Genetics **51**, 675–682 (2019) | Cite this article

10k Accesses | 115 Citations | 89 Altmetric | Metrics

17

TWAS and colocalization: pleiotropy



PrediXcan, SMR, FUSION Coloc, Enloc, eCAVIAR, Sherlock

- Image credit: Haky Im @ UChicago
- Locus level colocalization combined with TWAS: Hukku *et al.* (2022) Am. J. Hum. Genet.

18

TWAS and Mendelian randomization

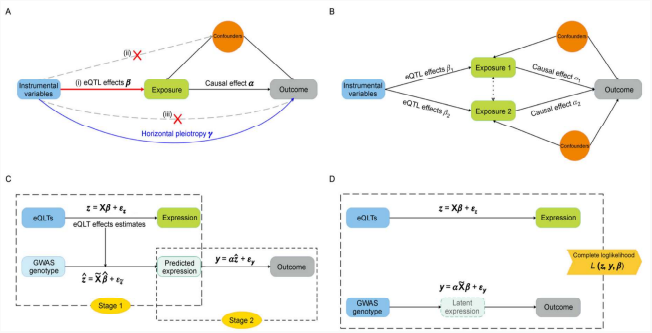


Figure: Zhu and Zhou (2022) Quantitative Biology

19

Multivariate analysis in genetic association studies

Gao Wang, Ph.D.

Advanced Gene Mapping Course, November 2022

The Gertrude H. Sergievsky Center and Department of Neurology
Columbia University Vagelos College of Physicians and Surgeons

Motivation

Beyond per trait per variant association studies

Statistical fine-mapping (multiple regressors)

- Identify non-zero effect variables by accounting for LD

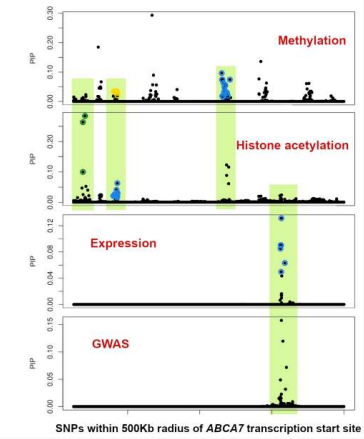
Meta-analysis (multiple responses)

- Integrate information across multiple conditions / studies

"Causal" variants across multiple conditions?

- Cross-population fine-mapping; colocalization; pleiotropy; mediation; ...

The problem



4

The problem

For a genetic variable analyzed in two conditions:

$P($**"causal" in trait 1 & 2** | association data for 1 & 2$)$

5

The problem

For a genetic variable analyzed in two conditions:

$P($**"causal" in trait 1 & 2** | association data for 1 & 2$)$

Denote data as D_1 and D_2, and use indicator variables γ_1, γ_2 for variable having effects in 1 and 2, and hyperparameters Θ:

$$P(\gamma_1 = 1, \gamma_2 = 1 | D_1, D_2, \Theta)$$

5

Multivariate relationships?

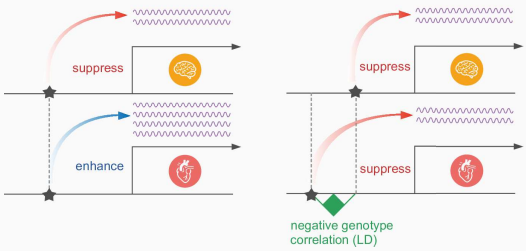


Figure: Pleiotropy or Linkage?

6

Meta-analysis review

Fixed effect and random effects models

Different assumptions on **effects across studies**

- Fixed effect model: all studies *share a common effect size*
- Random effects model: effect sizes are random variables *from an underlying distribution*

Fixed effect (FE) model

Let $\hat{\beta}_i$ be the observed effect size of study i, $1 \leq i \leq k$, and s_i^2 its variance. The true effect size is β. The observed effect is modelled as

$$\hat{\beta}_i \sim N(\beta, s_i^2),$$

with likelihood function

$$L(\beta) = P(\hat{\boldsymbol{\beta}}|\beta) = \prod_i^k P(\hat{\beta}_i|\beta) \propto \prod_i^k \exp\left[-\sum_i^k \frac{(\hat{\beta}_i - \beta)^2}{2s_i^2}\right].$$

Fixed effect (FE) model

Let $\hat{\beta}_i$ be the observed effect size of study i, $1 \leq i \leq k$, and s_i^2 its variance. The true effect size is β. The observed effect is modelled as

$$\hat{\beta}_i \sim N(\beta, s_i^2),$$

with likelihood function

$$L(\beta) = P(\hat{\boldsymbol{\beta}}|\beta) = \prod_i^k P(\hat{\beta}_i|\beta) \propto \prod_i^k \exp\left[-\sum_i^k \frac{(\hat{\beta}_i - \beta)^2}{2s_i^2}\right].$$

Let $w_i = 1/s_i^2$ be the weight of study i. The MLE of summary effect is

$$\hat{\beta} = \frac{\sum_i^k w_i \hat{\beta}_i}{\sum_i^k w_i}$$ **Inverse variance weighting**

Random effects (RE) model

Let $\hat{\beta}_i$ be the observed effect size of study i, $1 \leq i \leq k$, and s_i^2 its variance. Let β_i be the true effect size of study i. The observed effect is modelled as

$$\hat{\beta}_i|\beta_i \sim N(\beta_i, s_i^2), \quad \beta_i \sim N(\beta, \sigma^2)$$

with likelihood function

$$P(\hat{\boldsymbol{\beta}}|\beta, \sigma^2) \propto \prod_i^k \frac{1}{s_i^2 + \sigma^2} \exp\left[-\sum_i^k \frac{(\hat{\beta}_i - \beta)^2}{2(s_i^2 + \sigma^2)}\right].$$

Random effects (RE) model

Let $\hat{\beta}_i$ be the observed effect size of study i, $1 \leq i \leq k$, and s_i^2 its variance. Let β_i be the true effect size of study i. The observed effect is modelled as

$$\hat{\beta}_i|\beta_i \sim N(\beta_i, s_i^2), \quad \beta_i \sim N(\beta, \sigma^2)$$

with likelihood function

$$P(\hat{\boldsymbol{\beta}}|\beta, \sigma^2) \propto \prod_i^k \frac{1}{s_i^2 + \sigma^2} \exp\left[-\sum_i^k \frac{(\hat{\beta}_i - \beta)^2}{2(s_i^2 + \sigma^2)}\right].$$

RE has weight $w_i^* = 1/(s_i^2 + \sigma^2)$; summary effect $\hat{\beta}$ can be similarly computed as FE, replacing w_i with w_i^*. σ^2 can be estimated (*e.g.* , MLE).

Meta-analysis: a multivariate regression prospective

Multivariate model(s) for effect sizes

Consider a parametric model on **effect sizes** across studies,

$$B_j|\gamma = 1 \sim MVN(0, U)$$

Consider 2 studies, *e.g.* height GWAS in Europeans and Africans.

Fixed-effect model multivariate analysis

Effect sizes are exactly the same between two studies,

$$U_{\text{fixed}} = \sigma_0^2 \times \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

11

Random effects model multivariate analysis

Effect sizes are different between two studies, but are from the same distribution,

$$U_{\text{random}} = \sigma_0^2 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

12

Other multivariate models

$$U_{\text{partially shared}} = \sigma_0^2 \times \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where $|\rho| \leq 1$. This contains the two meta-analysis models as special cases!

13

Other flexible multivariate models

More generally,

$$U = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix}$$

- Pro: more generic than U_{fixed} and U_{random}
- Con: 3 parameters to deal with, compared to one σ_0^2

14

Analogy to popular multivariate models (some necessary but, not sufficient)

- Colocalization correlation matrix:

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

- Condition specific correlation matrix:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

15

Analogy to popular multivariate models (some necessary, but not sufficient)

- Mediation:

$$U_{\text{mediation}} = \sigma_0^2 \times \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & \rho_2 \end{bmatrix}$$

- Genotype $\rightarrow$ Trait 1 $\rightarrow$ Trait 2.
- Effect on trait 2 should be smaller than that on trait 1.

16

Variant colocalization: variable selection in meta-analysis

The problem

For a genetic variable analyzed in GWAS and eQTL studies:

$$P(\gamma_g = 1, \gamma_e = 1 | D_g, D_e, \Theta)$$

17

Colocalization method: *coloc*

coloc [Giambartolomei *et al.* (2014) PLoS Genet.]

- On X: "one causal" assumption
- On Y: the null + 4 combinations given "one causal"
  1. In 1 but not 2
  2. In 2 but not 1
  3. In 1 and 2 but not the same variable
  4. In 1 and 2 and the same variable (colocalization)
  5. No association in both data 1 and 2

Colocalization method: *eCAVIAR*

eCAVIAR [Hormozdiari *et al.* (2016) Am. J. Hum. Genet.]

- On X: multiple effect variables
- On Y: each effect variable can be
  1. In 1 but not 2
  2. In 2 but not 1
  3. In both 1 and 2
  4. No association in both data 1 and 2

eCAVIAR effects assumption

Effect sizes are independent,

$$U = \begin{bmatrix} \sigma_g^2 & 0 \\ 0 & \sigma_e^2 \end{bmatrix}$$

Colocalization method: *enloc*

enloc [Wen *et al.* (2017) PLoS Genet.]

- Key difference: cross-condition effects **not** independent
- **eQTL signals are enriched in GWAS**

Colocalization method: *enloc*

enloc [Wen *et al.* (2017) PLoS Genet.]

- Key difference: cross-condition effects **not** independent
- **eQTL signals are enriched in GWAS**

But how?

- Recall **fine-mapping with functional annotation** for j

$$\log\left[\frac{\pi}{1-\pi}\right] = \alpha_0 + \alpha\gamma_e$$

and in this context

$$\pi := P(\gamma_g = 1 | \gamma_e = 1)$$

enloc two step procedure

1. Obtain $P(\gamma_g = 1)$ and $P(\gamma_e = 1)$ using fine-mapping
2. Fit the enrichment model via **multiple imputation**

Connections among colocalization methods

- *eCAVIAR* is a special case of *enloc* with $\alpha = 0$.
- *coloc* is a special case of "one causal" fine-mapping based *enloc* with fixed, high**(!)** α value by default.
- Recent *coloc* extension: *coloc* version 5, aka *SuSiE-coloc* removed the "one causal" assumption.
  - Wallace (2021) PLoS Genetics
  - `https://chr1swallace.github.io/coloc/`

Connections among colocalization methods

- *eCAVIAR* is a special case of *enloc* with $\alpha = 0$.
- *coloc* is a special case of "one causal" fine-mapping based *enloc* with fixed, high**(!)** α value by default.
- Recent *coloc* extension: *coloc* version 5, aka *SuSiE-coloc* removed the "one causal" assumption.
  - Wallace (2021) PLoS Genetics
  - `https://chr1swallace.github.io/coloc/`

Summary: **pattern** and **scale** of effect size correlations, represented as different **prior** models.

Practical considerations

- Choice of prior
  - Best to **estimate enrichment α from data**
  - $\alpha \in [0, 5]$ suggested by $> 4,000$ GWAS + GTEx data
- LD reference mismatch: underestimate α, thus power loss

Hukku *et al.* (2021) Am. J. Hum. Genet.

Multivariate adaptive shrinkage and fine-mapping

More phenotypes, more complications

Figure: Plausible patterns of sharing

Major challenges

- **For a given variant**: the less assumption made on multivariate effects, the more parameters to estimate.
  - FE and RE models are restrictive but easy to fit.
- **Different variants**: may fit in different multivariate effect models

A naive mixture model

"FE and RE are equally likely for any variant":

$$U_{mixed} = 0.5 \times \begin{bmatrix} \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 \end{bmatrix} + 0.5 \times \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix}$$

Prior allows for possibility of both; data will determine where posterior lands.

A data-adaptive mixture model

Instead of making assumptions, can we **learn from data**:

- What are the latent structures for multivariate effects?
- How often does each structure appear?

and use these to construct the mixture model?

Patterns of sharing: factor analysis

Decomposing effect estimates, $\widehat{B} = LF + E$

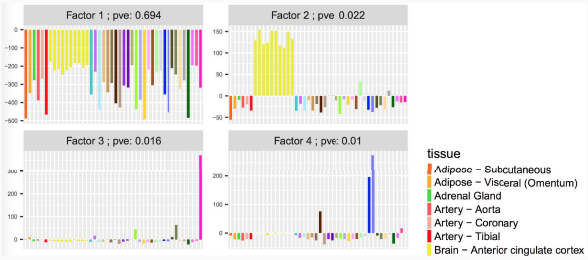


Figure: Sparse factor analysis of GTEx data

Incorporating all possible patterns

Multivariate effects of a variant follows the k-th pattern with probability π_k:

$$U_{mixed} = \pi_1 \times \begin{bmatrix} 2.4 & 0.3 \\ 0.3 & 1.5 \end{bmatrix} + \pi_2 \times \begin{bmatrix} 1.6 & 0.001 \\ 0.001 & 0.02 \end{bmatrix} + \pi_3 \times \cdots$$

This is the Multivariate Adaptive Shrinkage Prior.

- Step 1: estimated π_k via EM algorithm using data across genome.
- Step 2: apply this prior to each variant in association mapping.

Multivariate effect size sharing in eQTLs

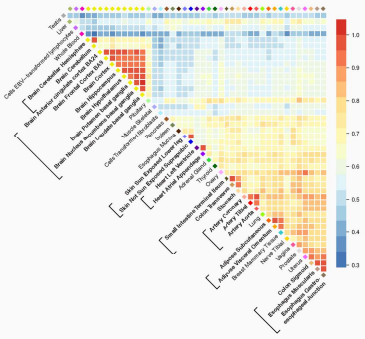


Figure: Quantitative characterization of eQTL effects heterogeneity in GTEx

Application to multivariate fine-mapping

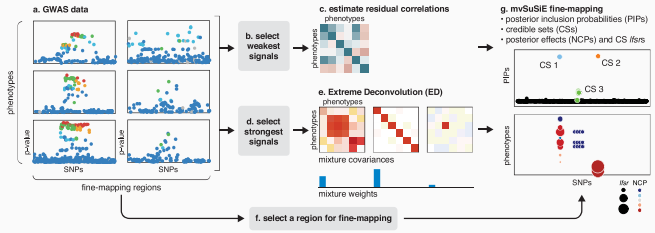


Figure: mvSuSiE fine-mapping with adaptive shrinkage model

The Ethics and Regulation of Human Subjects Research

Judy Matuk, M.S.
Independent Consultant

1

The Nuremberg Code (1947)

Ten Basic Principles, including:

"The voluntary consent of the human subject is absolutely essential…"

"The experiment should be conducted as to avoid all unnecessary physical and mental suffering and injury…"

"No experiment should be conducted where there is an a priori reason to believe that death or disabling injury will occur; except, perhaps, in those experiments where the experimental physicians also serve as subjects."

"During the course of the experiment, the human subject should be at liberty to bring the experiment to an end if he has reached the physical or mental state where continuation of the experiment seems to him to be impossible."

During the course of the experiment the scientist in charge must be prepared to terminate the experiment at any stage, if he has probable cause to believe…that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject.

2

Tuskegee Study of Untreated Syphilis in the Negro Male (1932-1972)

3

National Research Act (1974)

Required the creation of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

4

The Ethics of Conducting Research with Humans: The Belmont Report (1979)

- Beneficence
  - maximize benefits, minimize risks
- Justice
  - Who should bear the burdens of the research?
  - Who should benefit from results?
- Respect for Persons
  - Autonomy
  - Protect those with diminished autonomy

5

The Belmont Report was the basis for federal requirements of human research protections

Office for Human Research Protections

- 45 CFR 46 Subpart A ('Common Rule')
- Subpart B (Pregnant Women, Fetuses, and Nonviable/Questionable Viable Neonates),
- Subpart C (Prisoners),
- Subpart D (Minors)

Food & Drug Administration
(jurisdiction: clinical investigations of drugs, devices, biologics)

- 21 CFR 50: Protection of Human Subjects
- 21 CFR 56: Institutional Review Boards
- 21 CFR 312: Investigational Drugs
- 21 CFR 812: Investigational Devices

6

What is the Common Rule?

It is **the** Federal Policy for the Protection of Human Subjects

Originally promulgated in 1991, with no significant changes, until 1/21/19!

Rockefeller's Federal Wide Assurance (FWA) certifies compliance with this federal policy (for human research conducted or supported by Common Rule agencies…)

7

What's so Common about the Common Rule?

✓19 federal agencies follow the new Common Rule, e.g.,

- DHHS, including NIH (45 CFR 46, Subpart A)*
- DoD (32 CFR 219)
- NSF (45 CFR 690)
- DoEnergy (10 CFR 745)
- Department of VA (38 CFR 16)
- DoEducation (34 CFR 97)

*FDA is within DHHS, but also has its own regulations

*DoJ has not signed on yet

8

First Question: Is your activity "human subjects research" (HSR)?

9

Specifically:

1. Is it HSR according to the Common Rule?
2. Is it HSR according to FDA?

(could be both!)

10

Start with the Common Rule

First assess:

Does the activity involve Research?

11

Common Rule Definition of Research:

"…**a systematic investigation**, including research development, testing and evaluation, **designed to develop or contribute to generalized knowledge**…"

(Both parts of the definition must be met)

12

Part I of the definition: What's a Systematic Investigation?

an activity that involves a prospective plan which incorporates data collection, either quantitative and/or qualitative, and data analysis to answer a question

Does a case study involve a systematic investigation?

13

Part II: What does 'designed to develop or contribute to generalizable knowledge' mean?

…designed to draw general conclusions:

✓what we know about what is being tested is not yet firmly established or accepted;

and

✓the activity is not dependent on the unique characteristics of the target population or system in which it will be implemented

14

An activity is not likely to be generalizable if the intent is:

The evaluation or improvement of a process, practice, or program at the site where the activity is being conducted

Results only to be applied to populations, or inform practice within the target population or within the site where the activity is being conducted

Implementation and evaluation of an evidence-based practice, process, or program (is it functioning as intended within the site where the activity is being conducted or with the local target population

15

If the activity IS research: Does the research involve human subjects, according to the Common Rule?

A living individual about whom an investigator conducting research:

(i) Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or

(ii) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens.

16

Once you determine if the activity is or is not human subjects research according to the Common Rule…

You still need to assess if the activity is human subjects research according to FDA

17

FDA Decisions

Does the activity evaluate an FDA-regulated test article (i.e., drug, biologic, device)?

Does the activity involve Human Subjects?

An individual who is, or becomes, a participant in research, either as a recipient of the test article or as a control. A subject may be either a healthy human or a patient. *Also included in the FDA human subject definition: The use of a biological specimen –even if de-identified-from an individual used to test an investigational device*

Does the activity involve research (clinical investigation)?

Any experiment that involves a test article and one or more human subjects…

18

If the activity IS human subjects research, next question: Is it exempt from the federal regulations? *

*this does not mean exempt from institutional review!

19

Focus on: Exemption #4

Secondary research* for which consent is not required

***Secondary research only!** (i.e., re-using identifiable information and/or identifiable biospecimens that were, or will be, are collected for another reason, e.g., clinical or research)

20

Exemption #4: Secondary research uses of identifiable private information or identifiable biospecimens can be exempt under this category, if at least one of the following criteria is met:

21

Exemption 4(ii)

Identifiable private information…is **recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained** directly or through identifiers linked to the subject, the investigator does not contact the subjects, and the investigator will not re-identify subjects;

22

Exemption 4 (iii)

"The research involves only information collection and analysis involving the investigator's use of identifiable health information when that use is regulated under 45 CFR parts 160 AND 164, subparts A and E [HIPAA], for the purposes of "health care operations" or "research" as those terms are defined at 45 CFR 164.501 or "public health activities and purposes" as described under 45 CFR 164.512(b)"

23

What are the ethical standards that should be considered for all exempt studies?

Criteria	Yes	No	NA
The research holds out no more than minimal risk to participants	☐	☐	
Selection of participants is equitable	☐	☐	
If there is recording of identifiable information, there are adequate provisions to maintain the confidentiality of the data	☐	☐	☐
If there are interactions with participants, there are adequate provisions to protect the privacy interests of participants	☐	☐	☐
If there are interactions with participants, the consent process or information provided to potential subjects includes the following: ☐ N/A – there are no interactions and no other need for consent			
That the activity involves research	☐	☐	☐
A description of the procedures	☐	☐	☐
For Category 3 research that involves subject deception: A statement that subjects will be unaware of or misled regarding the nature or purposes of the research	☐	☐	☐
That participation is voluntary	☐	☐	☐
Name and contact information for the researcher	☐	☐	☐

24

If the activity IS human subjects research, but does not qualify for exemption, **it is** HSR that is not exempt, i.e., it is subject to federal regulations governing human research protection…

…including review by a federally mandated Institutional Review Board (IRB)

25

Two Types of Non-Exempt Review

1. Expedited Review
2. Full Board Review

26

For a non-exempt study to qualify for **Expedited (not full IRB Board) Review…**

…The research must be all of the following:

- no greater than minimal risk
- not involve prisoners (per OHRP guidance)
- not be classified
- not involve identifiable data that would place subjects at risk of criminal or civil liability or be damaging to the subjects financial standing, employability, insurability, reputation, or be stigmatizing. If it could, reasonable protections must be in place so that risks related to invasion of privacy and breach of confidentiality are no greater than minimal, **and**
- Fit into one or more of these categories: https://www.hhs.gov/ohrp/regulations-and-policy/guidance/categories-of-research-expedited-review-procedure-1998/index.html

27

If the nonexempt research doesn't qualify for expedited review, it must be reviewed at a convened IRB meeting.

28

Whether expedited or full board, a study must meet federally-defined criteria in order to be approved

i.e.,

"The .111 Criteria"

29

§ 46.111 Criteria for IRB approval of research.

(a) In order to approve research covered by this policy the IRB shall determine that all of the following requirements are satisfied:

30

1. Risks to subjects are minimized:

(i) By using procedures which are consistent with **sound research design** and which do not unnecessarily expose subjects to risk, and

(ii) Whenever appropriate, by using procedures already being performed on the subjects for diagnostic or treatment purposes

31

2. Risks to subjects are reasonable in relation to anticipated benefits, if any, to subjects, and the importance of the knowledge that may reasonably be expected to result

32

3. Selection of Subjects is Equitable

Consider:

- The setting in which the research will be conducted
- Who is included, who is excluded? Does it make scientific sense? Ethical sense?
- If applicable: Are children in a study involving a test article that hasn't first been tested in adults? Pregnant women before non-pregnant women?
- Costs or compensation that may impact 'fairness'
- Screening and recruitment?
- What about non-English speakers?

33

4. Informed consent will be sought from each prospective subject or the subject's legally authorized representative, in accordance with, and to the extent required by, §46.116

If not:

Are **ALL** the criteria for waiving informed consent or for altering/excluding specific elements of informed consent met?

34

5. Informed consent will be appropriately documented or appropriately waived in accordance with §46.117

If not:

Does the research meet one of the allowable criteria to waive documentation?

35

6. When appropriate, the research plan makes adequate provision for monitoring the data collected to ensure the safety of subjects

- **What** data will be monitored for safety purposes? **When? How?**
- **Who** will be responsible for evaluating safety data? Is a DSMB needed?
- Stopping Rules?
- Communication plan of findings to investigators and IRBs (from the IRB of Record or Sponsor)

36

7.When appropriate, there are adequate provisions to protect the privacy of subjects…

Consider:

- Settings where recruitment, consent, and research procedures and interactions will occur
- Provisions to ensure privacy for each of the above
- Provisions to ensure privacy when contacting or soliciting information from subjects

37

…and to protect the confidentiality of subject data

General:

- How will the data/biospecimens be stored?
- If identifiers will be removed or replaced, is there a possibility that such information/biospecimens could be re-identified?
- Will the data/biospecimens be shared/transmitted/ transferred to a third party or otherwise disclosed or released? How?
- Is there a potential risk of harm to individuals if the data/biospecimens are lost, stolen, compromised, or otherwise used in a way contrary to the parameters of the study?
- Plans for data retention and destruction?

38

A closer look at data security: minimize the risk of disclosure or breach of data

- Obtaining the data
  - What is the sensitivity of the data? Are all the data points that will be accessed or gathered for the research necessary to achieve the objectives of the research?
- Recording the data
  - What (if any) identifiers, including codes, will be recorded for the research?
- Storing the data
  - Where will paper research records, including signed consent forms, be stored? How will paper records be kept secure and restricted to authorized project personnel?
  - Where will the electronic research data be study be stored (University-provided database application like REDCap, IT file server, etc.)?
  - If there a key that links code numbers to identifiers, that list should be kept separate from the coded data, including copies of signed informed consent forms. Additionally, access to that list/key must be restricted authorized research personnel.

39

Data security, continued

- Transporting or transmitting the data
  - If any research data will be collected on a mobile device, such as an electronic tablet, cell phone, or wireless activity tracker, details are needed regarding the physical security of the device, electronic security, and how the transfer of data from device to research storage location will be securely accomplished.
  - If any research data will be directly entered/sent by subjects over the internet or via email, will a University-provided database application (like REDCap) be used, or is there an encrypted tunnel to the site/application?
- Access to the data
  - How will the investigators ensure only approved research personnel have access to the stored research data? Password-protected files, role-based security, etc.?
- Sharing of the data
  - Will data be transferred or disclosed to or from the University? Is a contract or data transfer agreement necessary? What (if any) identifiers will be included? How will the data be securely transferred or disclosed (University-approved secure file transfer, etc.)?

40

And (111.b) When some or all of the subjects are likely to be vulnerable to coercion or undue influence, such as children, prisoners, individuals with impaired decision-making capacity, or economically or educationally disadvantaged persons, additional safeguards have been included in the study to protect the rights and welfare of these subjects.

(set aside issues with children, pregnant women/fetuses, prisoners, regulations for which are codified in the Common Rule subparts---more on that in a moment)

- What are some considerations when determining if additional safeguards are necessary and sufficient?
  - Examples:
    - For economically disadvantaged…is there payment? What is the amount? schedule?
    - For educationally disadvantaged…is the consent process particularly simplified? Should there be a witness to the consent process?

41

That's it for the .111 criteria… but that's not all!

Pregnant Women?
Subpart B of 45 CFR 46

Prisoners?
Subpart C of 45 CFR 46

Children?
Subpart D of 45 CFR 46

Department of Education (ED)?
Family Educational Rights and Privacy Act (FERPA) (34 CFR 99)
and the Protection of Pupil Rights Amendment (PPRA) (34 CFR 98)
See resources provided by ED when developing your research protocol

Investigational Drugs, biologics, devices?
FDA regulations at 21 CFR 50, 21 CFR 56, 21 CFR 312, 21 CFR 812

HIPAA?
45 CFR Part 160 and Subparts A and E of Part 164

42

43

Yale

From cross-phenotype associations to pleiotropy in human genetic studies

Andrew DeWan, PhD, MPH
Associate Professor of Epidemiology
Director, Yale Center for Perinatal, Pediatric and Environmental Epidemiology
Yale School of Public Health

Work done in collaboration with Yasmmyn Salinas, PhD, MPH, Assistant Professor of Epidemiology
Yale School of Public Health

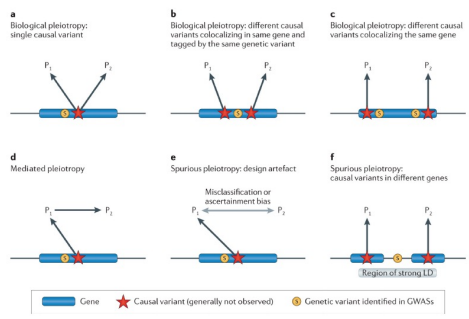
Yale SCHOOL OF PUBLIC HEALTH

1

Pleiotropy

- Phenomenon in which a genetic locus affects more than one trait or disease
- Molecular level
  - Single gene with multiple physiological function
  - Two domains of a single gene product with different functions and affecting multiple phenotypes
  - Gene product with a single function that affects multiple phenotypes acting in multiple tissues
- Statistical level
  - A locus displaying cross-phenotype associations is often considered pleiotropic
  - Can be at the variant, gene or region level

2

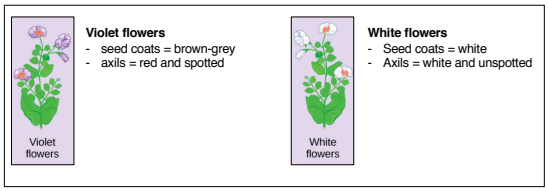


Solovieff et al. Nat Rev Genet. 2013 July ; 14(7): 483–495. doi:10.1038/nrg3461.

3

Early example of "pleiotropy"

Gregor Mendel documented one of the earliest examples of pleiotropy in his pea plant experiments



Mendel, J. G., 1866 Experiments in plant hybridization. Verhandlungen des naturforschenden Vereines in Brunn 4: 3–47 (in German).

4

Examples in humans

- Marfan syndrome
  - FBN1 (fibrillin-1)
  - thinness, joint hypermobility, limb elongation, lens dislocation, and increased susceptibility to heart disease.
- Holt-Oram syndrome,
  - TBX5 (transcription factor)
  - cardiac and limb defects
- Nijmegen breakage syndrome
  - NBS1 (DNA damage repair protein)
  - microcephaly, immunodeficiency, and cancer predisposition

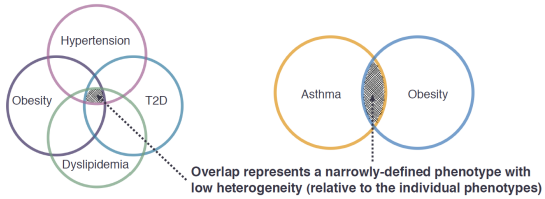
5

Pleiotropy and complex disease comorbidity

- Examples of correlated (comorbid) disease
  - Obesity, hypertension, dyslipidemia, type 2 diabetes (metabolic disorder)
  - Depression, anxiety, personality disorders (psychiatric disorder)
  - Asthma, obesity (pro-inflammatory conditions)
- Why do certain disease occur together
  - Causality
  - Shared environmental risk factors
  - Shared genetic risk factors

6

Pleiotropy and complex disease comorbidity



7

Pleiotropy and complex disease comorbidity

- Pleiotropy-informed analyses consider multiple phenotypes together and take into account the correlation between the phenotypes
  - Analyzing multiple correlated phenotype (e.g. comorbid diseases) is equivalent to analyzing a single narrowly-defined phenotype with low heterogeneity

8

9

Pleiotropy and complex disease comorbidity

- Detecting shared genetics and/or molecular pathways between comorbid diseases can help us understand exactly how the etiology of the diseases overlap
- Etiologic overlaps:
  - provide opportunities for novel interventions that prevent or treat the comorbidity, rather than preventing/treating each disease separately
  - facilitate drug repurposing (that is, known drugs targeting a pleiotropic locus may be repurposed to treat other diseases controlled by that locus, precluding the need for the development and testing of a brand-new drug)

9

Abundant Pleiotropy in Human Complex Diseases and Traits

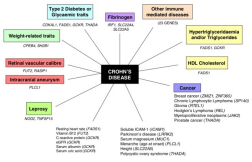
Shanya Sivakumaran,[1,6] Felix Agakov,[1,2,6] Evropi Theodoratou,[1,6] James G. Prendergast,[3] Lina Zgaga,[1,4] Teri Manolio,[5] Igor Rudan,[1] Paul McKeigue,[1] James F. Wilson,[1] and Harry Campbell[1,*]

The American Journal of Human Genetics *89*, 607–618, November 11, 2011

Table 6. Extent of Pleiotropy in Different Disease Classes

	Genes			SNPs		
Disease Class	Pleiotropic (%)	Nonpleiotropic (%)	p Value[a]	Pleiotropic (%)	Nonpleiotropic (%)	p Value[a]
All (comparison group)	233 (16.9)	1147 (83.1)	–	77 (4.6)	1610 (95.4)	–
Immune-mediated phenotypes	106 (37.7)	175 (62.3)	<0.0001	31 (8.3)	343 (91.7)	0.0066
Cancer	49 (34.8)	92 (65.2)	<0.0001	8 (4.8)	158 (95.2)	0.8456
Metabolic syndrome	79 (28.5)	198 (71.5)	<0.0001	30 (8.4)	327 (91.6)	0.0056

[a] Fisher's exact test p value.



10

11

Pleiotropy in gene mapping

- Mapping a single genotype to multiple phenotypes has the potential to uncover novel links between traits or diseases
- It can also offer insights into the mechanistic underpinnings of known comorbidities
- It can increase power to detect novel associations with one or more phenotypes

11

12

A practitioners' guide for studying pleiotropy in genetic epi studies

Am J Epidemiol. 2017 Aug 11. doi: 10.1093/aje/kwx296. [Epub ahead of print]

Statistical Analysis of Multiple Phenotypes in Genetic Epidemiological Studies:From Cross-Phenotype Associations to Pleiotropy.

Salinas YD, Wang Z, DeWan AT.

Abstract

In the context of genetics, pleiotropy refers to the phenomenon in which a single genetic locus affects more than one trait or disease. Genetic epidemiological studies have identified loci associated with multiple phenotypes, and these cross-phenotype associations are often incorrectly interpreted as examples of pleiotropy. Pleiotropy is only one possible explanation for cross-phenotype associations. Cross-phenotype associations may also arise due to issues related to study design, confounder bias, or non-genetic causal links between the phenotypes under analysis. Therefore, it is necessary to dissect cross-phenotype associations carefully to uncover true pleiotropic loci. In this review, we describe statistical methods that can be used to identify robust statistical evidence of pleiotropy. First, we provide an overview of univariate and multivariate methods for discovery of cross-phenotype associations and highlight important considerations for choosing among available methods. Then, we describe how to dissect cross-phenotype associations by using mediation analysis. Pleiotropic loci provide insights into the mechanistic underpinnings of disease comorbidity, and may serve as novel targets for interventions that simultaneously treat multiple diseases. Discerning between different types of cross-phenotype associations is necessary to realize the public health potential of pleiotropic loci.

KEYWORDS: genetic epidemiology; mediation analysis; pleiotropy

12

13

Guidelines for generating robust statistical evidence of pleiotropy

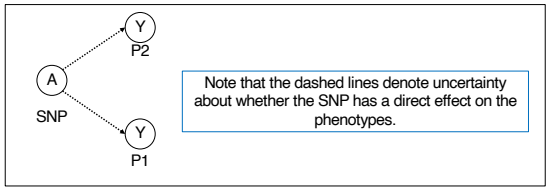
Discover CP associations ⇨ Dissect CP associations ⇨ **Classify** them as examples of biological, mediated, or spurious pleiotropy

13

14

Cross-phenotype (CP) associations

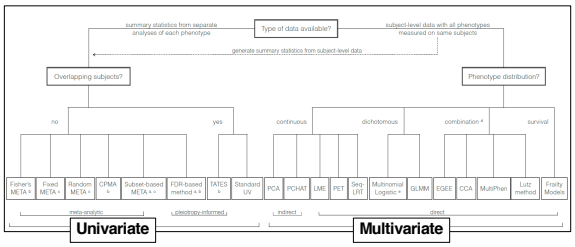
Statistical associations between a **single genetic locus** – a single gene or a single variant within a gene – and **multiple phenotypes**



14

15

Analytic options for discovery of CP associations



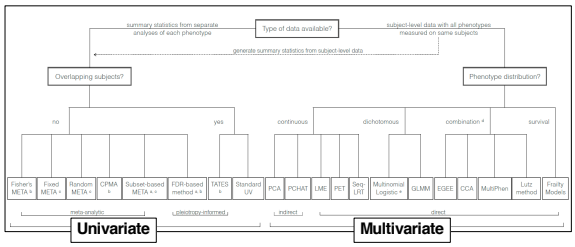
Key distinction:

- Univariate methods examine the association between a given SNP and each trait *separately*
- Multivariate methods examine the association between a given SNP and each trait by modeling the traits *jointly*

15

16

Analytic options for discovery of CP associations



Choice between univariate and multivariate approaches depends on:

- Types of data available on our phenotypes of interest
  - Summary statistics vs. individual-level data?
  - Are the phenotypes measured on the same subjects?
- Distribution of the phenotypes (e.g., quantitative or disease trait)

16

17

Univariate methods are by far the most commonly used to detect CP associations

- Univariate methods include (but are not limited to) the methods you've discussed in class so far:
  - allelic Chi-Square test
  - genotypic Chi-Square test
  - regression-based methods
- The overall approach is to:
  - obtain univariate association p-values for each phenotype
  - declare CP associations at genetic loci that are statistically significantly associated with each phenotype

17

18

Hypothetical example: Discovery of CP associations for hypertension and heart disease by using logistic regression

Step 1. Fit two univariate regression models within PLINK

$$E[hypertension] = \beta_0 + \boldsymbol{\beta_1} * SNP$$

$$E[heart\ disease] = \beta_0 + \boldsymbol{\beta_1} * SNP$$

Word of caution: The univariate tests of association should be marginal tests (conducted irrespectively of the second phenotype) NOT conditional tests (conducted on a subset defined based on absence/presence of the second phenotype). In this example, what that means is that the regression for hypertension should be fit on all subjects *irrespectively* of their heart disease status; and the regression for heart disease should be fit on all subjects *irrespectively* of their hypertension status. More on this later!

evidence to declare a CP association at this SNP.

18

Hypothetical example: Discovery of CP associations for hypertension and heart disease by using logistic regression

Step 1. Fit two univariate regression models within PLINK

$$E[hypertension] = \beta_0 + \boldsymbol{\beta_1} * SNP$$

$$E[heart\ disease] = \beta_0 + \boldsymbol{\beta_1} * SNP$$

Step 2. For a given SNP, examine p-values for $\boldsymbol{\beta_1}$ from each model.

- P-value for $\boldsymbol{\beta_1}$ in hypertension model = 1.03×10^{-12}
- P-value for $\boldsymbol{\beta_1}$ in heart disease model = 6.02×10^{-9}

Step 3. Declare CP associations at a given SNP, if the p-values for $\boldsymbol{\beta_1}$ in each model surpass the study significance threshold.

- Assuming the standard GWAS significance threshold (alpha=5×10^{-8}), there is a statistically significant association with both hypertension and heart disease at this particular SNP. Therefore, we have sufficient statistical evidence to declare a CP association at this SNP.

19

Using multivariate methods to increase the power to detect cross-phenotype associations

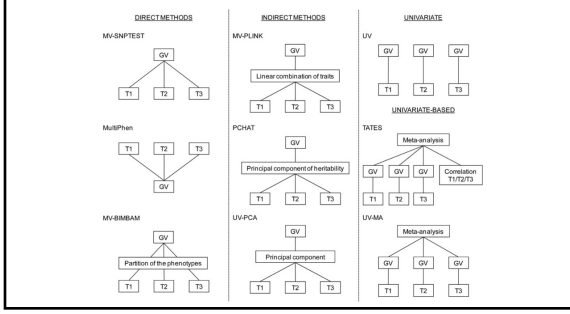
20

A Comparison of Multivariate Genome-Wide Association Methods

Tessel E. Galesloot[1], Kristel van Steen[2,3], Lambertus A. L. M. Kiemeney[1,4], Luc L. Janss[5]*, Sita H. Vermeulen[1,4]*

1 Department for Health Evidence, Radboud university medical center, Nijmegen, The Netherlands, 2 Systems and Modeling Unit, Montefiore Institute, University of Liège, Liège, Belgium, 3 Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, 4 Department of Urology, Radboud university medical center, Nijmegen, The Netherlands, 5 Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

PLOS ONE | www.plosone.org 1 April 2014 | Volume 9 | Issue 4 | e95923



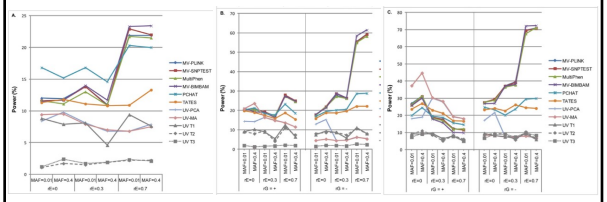
21

Table 1. Simulation scenarios.

# traits associated with QTL	Heritability (h^2_j)	Effect size (a_j)	rG	rE	MAF (q)
1	$h^2_1 = 0.1\%$, $h^2_2 = h^2_3 = 0$	$a_1 > 0$, $a_2 = a_3 = 0$	0	$3 \times 0/3 \times 0.3/3 \times 0.7$	0.01/0.4
2	$h^2_1 = h^2_2 = 0.1\%$, $h^2_3 = 0$	$a_1 = a_2$, $a_3 = 0$	+	$3 \times 0/3 \times 0.3/3 \times 0.7$	0.01/0.4
	$h^2_1 = h^2_3 = 0.1\%$, $h^2_2 = 0$	$-a_1 = a_2$, $a_3 = 0$	−	$3 \times 0/3 \times 0.3/3 \times 0.7$	0.01/0.4
3	$h^2_1 = h^2_2 = h^2_3 = 0.1\%$	$a_1 = a_2 = a_3$	+	$3 \times 0/3 \times 0.3/3 \times 0.7$	0.01/0.4
	$h^2_1 = h^2_2 = h^2_3 = 0.1\%$	$-a_1 = a_2 = a_3$	−	$3 \times 0/3 \times 0.3/3 \times 0.7$	0.01/0.4

MAF indicates minor allele frequency; j, trait; QTL, quantitative trait locus; rE, residual correlation; rG, genetic correlation.

doi:10.1371/journal.pone.0095923.t001



22

A comparison of univariate and multivariate GWAS methods for analysis of multiple dichotomous phenotypes

Yasmmyn D. Salinas[1], Andrew T. DeWan[1], and Zuoheng Wang[2]

[1]/Department of Chronic Disease Epidemiology; [2]/Department of Biostatistics, Yale School of Public Health, Yale University, 60 College St, New Haven, Connecticut, USA

Genet. Epidemiol. 41 (7), 689-689

23

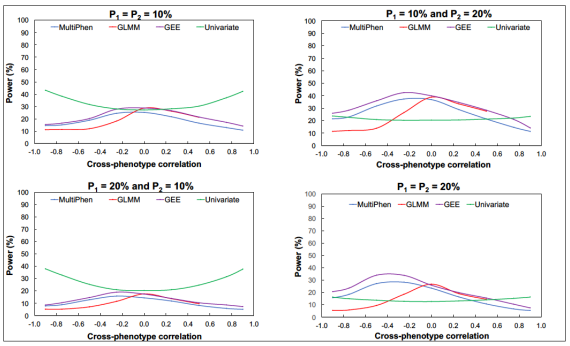
Simulation scenarios

# traits associated	h_j^2	$r_{Y1,Y2}$	P_j
1	h_1^2=0.1%,h_2^2=0%	[-0.9,0.9]	P1 = P2 = 10%
			P1 = P2 = 20%
			P1 = 10%, P2 = 20%
			P1 = 20%, P2 = 10%
2	$h_1^2 = h_2^2$= 0.1%	[-0.9,0.9]	P1 = P2 = 10%
			P1 = P2 = 20%
			P1 = 10%, P2 = 20%
			P1 = 20%, P2 = 10%
2	h_1^2 = 0.1%,h_2^2 = 0.05%	[-0.9,0.9]	P1 = P2 = 10%
			P1 = P2 = 20%
			P1 = 10%, P2 = 20%
			P1 = 20%, P2 = 10%

24

PLEIOTROPY PRESENT
equal effect sizes

Figure 2. Power when both phenotypes are associated with the SNP ($h_1^2 = h_2^2 = 0.1\%$) [a]

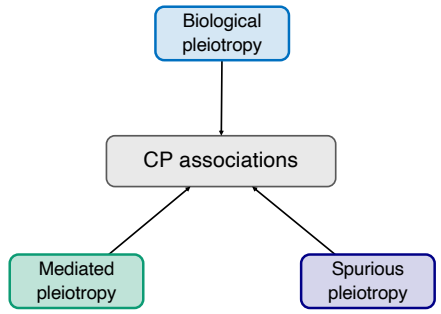


[a] Results for GLMMs are shown for $r_{Y1,Y2} \leq 0.5$ only, since the models experienced convergence issues for $r_{Y1,Y2} > 0.5$.

25

Problem: CP associations need not be indicative of pleiotropy

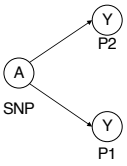
26



27

Biological pleiotropy

Independent associations between a genetic locus (A) and multiple phenotypic outcomes (Y)

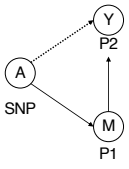


The SNP has a direct effect on each phenotype. (Note that direct or causal effects are depicted with solid lines).

28

Mediated pleiotropy

Association between a genetic locus (A) and an intermediate phenotype (M) that causes a second phenotypic outcome (Y)

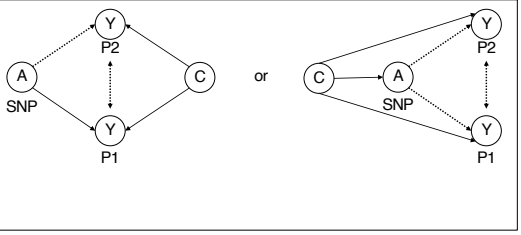


A non-genetic causal link between M and Y induces an association between A and Y, even in the absence of a direct effect of A on Y.

29

Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes



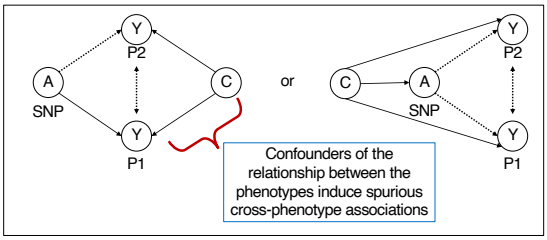
*Linkage disequilibrium is the non-random co-segregation of alleles.

30

31

Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes



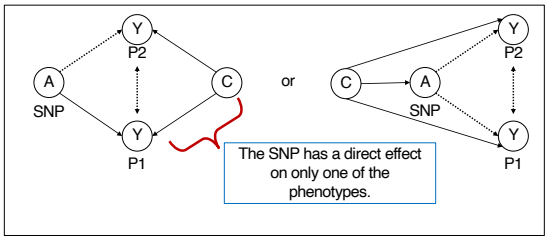
*Linkage disequilibrium is the non-random co-segregation of alleles.

31

32

Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes



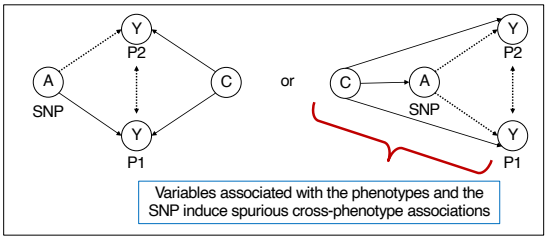
*Linkage disequilibrium is the non-random co-segregation of alleles.

32

33

Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes



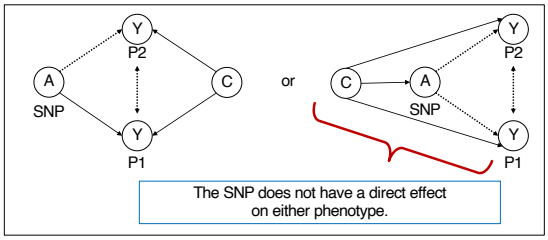
*Linkage disequilibrium is the non-random co-segregation of alleles.

33

34

Spurious pleiotropy

Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes

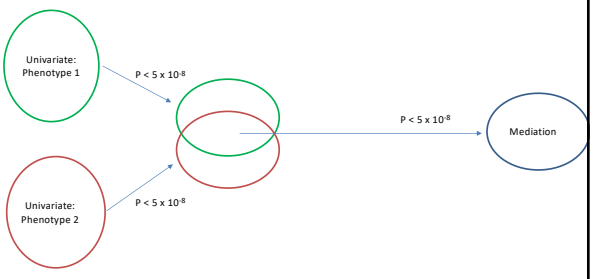


*Linkage disequilibrium is the non-random co-segregation of alleles.

34

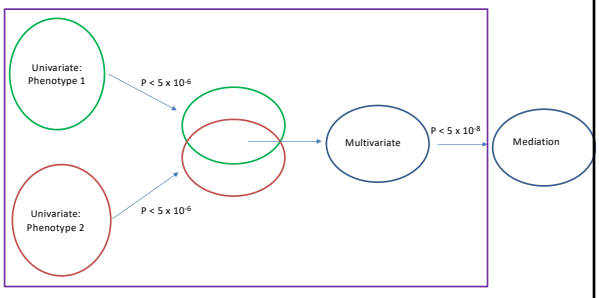
35

Pleiotropy exercise (Parts 1 and 2)



36

Pleiotropy exercise (Parts 1 and 2)



37

Guidelines for generating robust statistical evidence of pleiotropy

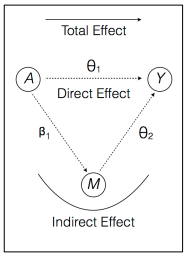
Discover CP associations → **Dissect** CP associations → Classify them as examples of biological, mediated, or spurious pleiotropy

37

38

Mediation analysis provides a tool for dissecting CP associations

- Mediation analysis decomposes the **total effect** of the SNP (*A*) on a phenotypic outcome (*Y*) into:
  - **Direct effect:** effect of *A* on *Y* that occurs independently of an intermediate phenotype (*M*)
  - **Indirect effect:** effect of *A* on *Y* that occurs through the intermediate phenotype *M*

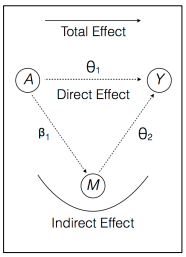


38

39

Mediation analysis: Data requirements

- All phenotypes must be measured on the same subjects
- Temporality must be ascertained
  - The occurrence of the intermediate variable *M* must precede that of the phenotypic outcome variable *Y*

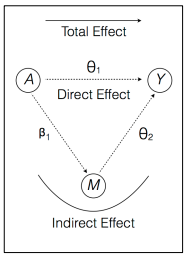


39

40

Mediation analysis: Assumptions

- There must be no unmeasured:
  - confounders of the total effect
  - confounders of the relationship between SNP *A* and the mediator *M*
  - confounders of the relationship between mediator *M* and phenotypic outcome *Y*



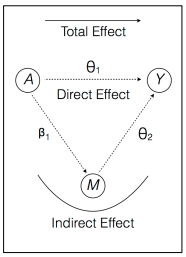
40

41

Mediation analysis: Assumptions

Typically met in genetic epi studies!

- There must be no unmeasured:
  - confounders of the total effect
  - confounders of the relationship between SNP *A* and the mediator *M*
  - confounders of the relationship between mediator *M* and phenotypic outcome *Y*

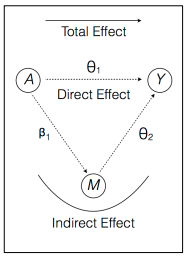


41

42

Mediation analysis: Assumptions

- There must be no unmeasured:
  - confounders of the total effect
  - confounders of the relationship between SNP *A* and the mediator *M*
  - confounders of the relationship between mediator *M* and phenotypic outcome *Y*



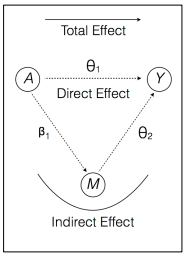
Requires adjustment for known confounders to prevent bias
(Note: this effectively restricts the use of mediation analyses to datasets in which data on such variables have been collected)

42

Mediation analysis: Regression-based approach

- Requires fitting two regression models, one for mediator *M* and one for phenotypic outcome *Y:*
  - $E[M\,|a,c] = \beta_0 + \boldsymbol{\beta_1} a + \beta_2' c$
  - $E[Y\,|a,m,c\,] = \theta_0 + \boldsymbol{\theta_1} a + \boldsymbol{\theta_2} m + \theta_4' c$

Assesses the effect of *A* on *M*, while controlling for measured confounders *(C)*

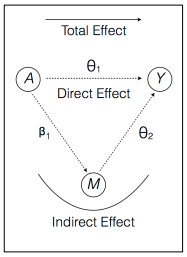


43

Mediation analysis: Regression-based approach

- Requires fitting two regression models, one for mediator *M* and one for phenotypic outcome *Y:*
  - $E[M\,|a,c] = \beta_0 + \boldsymbol{\beta_1} a + \beta_2' c$
  - $E[Y\,|a,m,c\,] = \theta_0 + \boldsymbol{\theta_1} a + \boldsymbol{\theta_2} m + \theta_4' c$

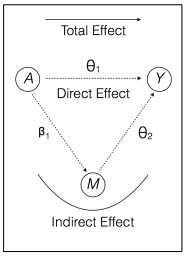
Assesses the effect of *A* on *Y*, while controlling for both *M* and *C*



44

Mediation analysis: Regression-based approach

- Requires fitting two regression models, one for mediator *M* and one for phenotypic outcome *Y:*
  - $E[M\,|a,c] = \beta_0 + \boldsymbol{\beta_1} a + \beta_2' c$
  - $E[Y\,|a,m,c\,] = \theta_0 + \boldsymbol{\theta_1} a + \boldsymbol{\theta_2} m + \theta_4' c$
- The parameter estimates from these models (**namely $\boldsymbol{\beta_1}$, $\boldsymbol{\theta_1}$, and $\boldsymbol{\theta_2}$**) are used to estimate the direct and indirect effects



45

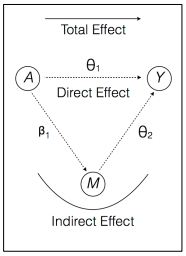
Guidelines for generating robust statistical evidence of pleiotropy

Discover CP associations → **Dissect** CP associations → **Classify** them as examples of biological, mediated, or spurious pleiotropy

46

Mediation analysis: Interpretation

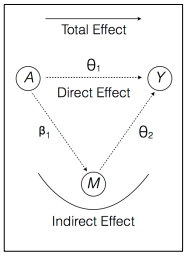
- **Mediated pleiotropy**
  - Complete mediation: SNP *A* is associated with mediator *M* and the total effect of *A* on phenotypic outcome *Y* is equal to its indirect effect (i.e., the direct effect is equal to 0).
  - Incomplete mediation: SNP *A* is associated with mediator *M* and *A* has both direct and indirect effects on phenotypic outcome *Y* (i.e., the total effect is equal to the sum of the direct and indirect effects)
- **Biological pleiotropy**
  - SNP A is associated with mediator M, and the total effect of SNP A on phenotypic outcome Y is equal to its direct effect (i.e., the indirect effect is equal to 0)



47

Mediation analysis: Interpretation

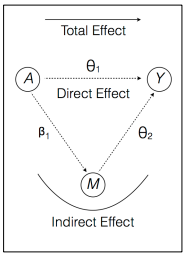
- **Mediated pleiotropy**
  - Complete mediation: SNP *A* is associated with mediator *M* and the total effect of *A* on phenotypic outcome *Y* is equal to its indirect effect (i.e., the direct effect is equal to 0).
- **Biological pleiotropy**
  - SNP A is associated with mediator M, and the total effect of SNP A on phenotypic outcome Y is equal to its direct effect (i.e., the indirect effect is equal to 0)
  - Incomplete mediation: SNP *A* is associated with mediator *M* and *A* has both direct and indirect effects on phenotypic outcome *Y* (i.e., the total effect is equal to the sum of the direct and indirect effects)



48

Mediation analysis: Interpretation

- **Spurious pleiotropy**
  - SNP A is not associated with mediator M after controlling for measured confounders



49

mediation R package

```
> med.fit<-glm(W1~rs1_2, data=combined, family=binomial("logit"))
> out.fit<-glm(W2~W1+rs1_2, data=combined, family=binomial("logit"))
> med.out<-mediate(med.fit,out.fit, treat="rs1_2", mediator="W1", boot=TRUE, boot.ci.type="bca", sims=1000)
> summary(med.out)

Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the BCa Method
```

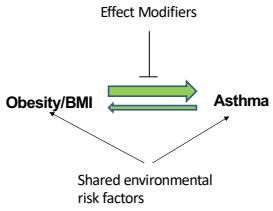
	Estimate	95% CI Lower	95% CI Upper	p-value
ACME (control)	0.02152	0.01823	0.03	<2e-16 ***
ACME (treated)	0.02199	0.01868	0.03	<2e-16 ***
ADE (control)	0.00723	0.00415	0.01	<2e-16 ***
ADE (treated)	0.00771	0.00443	0.01	<2e-16 ***
Total Effect	0.02922	0.02461	0.03	<2e-16 ***
Prop. Mediated (control)	0.73634	0.65429	0.84	<2e-16 ***
Prop. Mediated (treated)	0.75247	0.67272	0.85	<2e-16 ***
ACME (average)	0.02175	0.01847	0.03	<2e-16 ***
ADE (average)	0.00747	0.00426	0.01	<2e-16 ***
Prop. Mediated (average)	0.74441	0.66254	0.84	<2e-16 ***

50

Empirical searches for pleiotropic loci for asthma and obesity

51

Asthma-obesity comorbidity



Ford ES. The epidemiology of obesity and asthma. J Allergy Clin Immunol. 2005;115(5):897-909; quiz 10.
Stukus DR. Obesity and asthma: The chicken or the egg? J Allergy Clin Immunol. 2014.
Kim SH, Sutherland ER, Gelfand EW. Is there a link between obesity and asthma? Allergy Asthma Immunol Res. 2014;6(3):189-95.
Egan KB, Ettinger AS, DeWan AT, Holford TR, Holmen TL, Bracken MB. Longitudinal associations between asthma and general and abdominal weight status among Norwegian adolescents and young adults: the HUNT Study. Pediatric obesity. 2014.

52

Am J Hum Genet. 2009 Jul;85(1):87-96. doi: 10.1016/j.ajhg.2009.06.011. Epub 2009 Jul 2.

PRKCA: a positional candidate gene for body mass index and asthma.

Murphy A[1], Tantisira KG, Soto-Quirós ME, Avila L, Klanderman BJ, Lake S, Weiss ST, Celedón JC.

Study design

- Two phases:
  - genome-wide linkage analysis of BMI
  - follow-up family-based candidate-gene association study of BMI and asthma
- Strategy for candidate-gene study:
  - Authors focused on a single gene (*PRKCA*) within the BMI linkage peak because:
    - animal models suggest role of PRKCA in obesity; and
    - published association studies of other genes within the linkage peak had found no association with BMI.

53

Study population

- Costa Rica study
  - N = 415 asthmatic children + parents
- Childhood Asthma Management Program
  - N = 493 non-Hispanic White asthmatic children + parents

Note that ALL children in both study populations are asthmatic

54

55

Phenotype definitions

- Body mass index (BMI)
  - calculated from objective measures of height and weight
- Asthma
  - physician-diagnosed asthma + one of the following:
    - 2 respiratory symptoms or asthma attacks in prior year
    - increased airway responsiveness or bronchodilator response

55

56

Statistical methods

- Univariate family-based association tests (FBATs) were used to test *PRKCA* SNPs for association with BMI and asthma *separately*
  - Note: The FBAT statistic takes into account the phenotype of the **offspring only**
- Significance threshold used by study authors: $\alpha = 9.5 \times 10^{-5}$

56

57

Results for BMI

Table 3. Evidence for Association of *PRKCA* with BMI in Costa Rica and CAMP

Marker	Location (BP)[a]	Minor Allele	Allele Frequency: CR	Allele Frequency: CAMP	Number of Informative Families[b] (number of offspring with 0/1 recoded genotype): CR	Number of Informative Families[b] (number of offspring with 0/1 recoded genotype): CAMP	Effect Size[c]: CR	Effect Size[c]: CAMP	CR p Value[d,e]	CAMP Replication p Value[d,e] (two-sided)	Joint p Value[f] (CR, CAMP two-sided)
rs228883	61874457	T	0.27	0.33	91 (67/24)	110 (80/39)	2.45	1.60	+0.0011	+0.0038 (+0.0076)	$5.6 \times 10^{-5**}$ (1.0×10^{-4})
rs1005651	61868473	C	0.26	0.33	83 (60/23)	113 (83/39)	2.27	1.60	+0.0019	+0.0039 (+0.0077)	$9.5 \times 10^{-5**}$ (1.8×10^{-4})
rs228875	61924337	A	0.29	0.35	101 (70/31)	129 (92/46)	1.71	1.22	+0.0109	+0.0182 (+0.0364)	0.0019 (0.0035)
rs2244497	61931405	C	0.31	0.36	120 (86/34)	136 (98/47)	1.69	1.21	+0.0160	+0.0171 (+0.0341)	0.0025 (0.0046)

Two BMI-associated variants

57

58

Results for asthma

Table 4. Evidence for Association of *PRKCA* with Asthma in Costa Rica and CAMP

Marker	Location (BP)[a]	Minor Allele	Allele Frequency: CR	Allele Frequency: CAMP	Number of Informative Families[b] (number of offspring with 0/1 recoded genotype): CR	Number of Informative Families[b] (number of offspring with 0/1 recoded genotype): CAMP	Costa Rica p Value[c,d]	CAMP Replication p Value[c,d] (two-sided)	Joint p Value[e] (CR, CAMP two-sided)
rs732191	61779673	G	0.46	0.35	168 (117/51)	141 113/43	−0.0194	−0.0214 (−0.0428)	0.0036 (0.0067)
rs9895580	61789701	C	0.47	0.35	168 (117/51)	141 114/43	−0.0171	−0.0160 (−0.0320)	0.0025 (0.0047)
rs4411531	61793662	A	0.29	0.12	88 (70/18)	25 (24/1)	−0.0058	−0.0058 (−0.0117)	0.0004 (0.0007)
rs8080771	61824330	G	0.46	0.35	164 (116/48)	108 (90/29)	−0.0161	−0.0070 (−0.0140)	0.0011 (0.0021)
rs11652956	61839798	G	0.29	0.12	83 (65/18)	23 (22/1)	−0.0101	−0.0111 (−0.0222)	0.0011 (0.0021)
rs7221968	61848731	C	0.27	0.11	79 (63/16)	18 (17/1)	−0.0122	−0.0216 (−0.0432)	0.0024 (0.0045)
rs7405806	61862056	A	0.49	0.31	164 (109/55)	90 (77/20)	−0.0309	−0.0009 (−0.0018)	0.0003 (0.0006)
rs11079657	61862528	A	0.38	0.23	129 (94/35)	60 (56/8)	−0.0092	−0.0002 (−0.0004)	$2.6 \times 10^{-5**}$ ($5.0 \times 10^{-5**}$)

One asthma-associated variant

58

59

Conclusions

- **Authors' conclusion: *PRKCA* displays pleiotropy for asthma and BMI (pleiotropy at gene level)**
  - Two variants (rs228883 and rs1005651) displayed statistically significant associations with body mass index
  - A different variant (rs11079657) displayed a statistically significant association with asthma.

59

60

Conclusions

- **Our conclusion: *PRKCA* is associated with asthma and with BMI among asthmatics (no true CP association!)**
  - There is insufficient evidence to declare a CP association at *PRKCA* because the test of association with BMI was not a marginal test
    - FBAT test for BMI only took into account the phenotype of the offspring – which were ALL asthmatic
  - Thus, it remains to be seen whether the association with BMI is also present among non-asthmatics subjects
  - Without that information, we would not be able to assess whether asthma is a **mediator** or a **moderator** of the relationship between *PRKCA* and BMI.

60

61

A GWAS study of pleiotropy

Discovery and Mediation Analysis of Cross-Phenotype Associations Between Asthma and Body Mass Index in 12q13.2

Yasmmyn D. Salinas*, Zuoheng Wang, and Andrew T. DeWan

* Correspondence to Dr. Yasmmyn D. Salinas, Department of Chronic Disease Epidemiology, Yale School of Public Health, 60 College Street, New Haven, CT 06520 (e-mail: yasmmyn.salinas@yale.edu).

Am J Epidemiol. 2021;190(1):85–94

61

62

Study design

- Two parts:
  - Genome-wide search for cross-phenotype associations with asthma and body mass index
  - Follow-up mediation analysis to dissect genome-wide significant CP associations

62

63

Study population

- N = 305,945 White, British subjects from the UK Biobank (a population-based prospective cohort study of > 500,000 subjects, aged 40-69 years at baseline)

biobank uk
Improving the health of future generations

63

64

Phenotype definitions

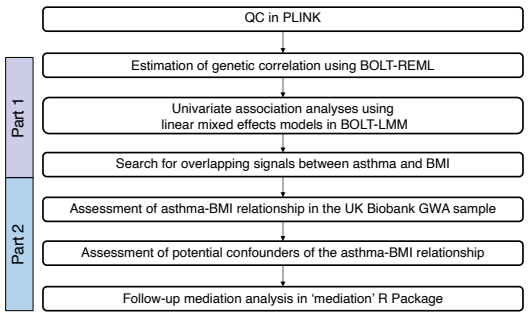
- BMI at baseline (kg/m^2):
  - calculated based on height and weight measurements collected by trained UK Biobank staff at the recruitment sites
- Asthma diagnosed prior to baseline (yes/no):
  - ascertained via the question “Has a doctor ever told you that you had asthma?”
  - Note: In mediation analyses, two subgroups were created based on age-at-diagnosis

biobank uk
Improving the health of future generations

64

65

Statistical Methods



65

66

Overlap in GWA signals

Association with BMI among the 1,457 SNPs with genome-wide significant p-values for asthma

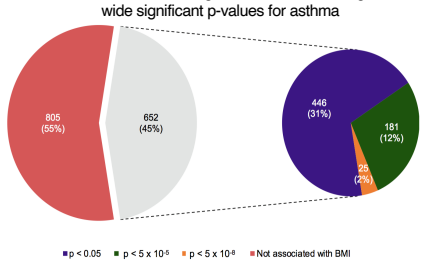


Figure 1. Overlap in GWA signals between asthma and BMI. Results for asthma are for the analysis of all asthmatic subjects (35,373 asthmatics vs. 270,572 non-asthmatics). Results for BMI are for the quantitative BMI analysis (n=305,945). Both analyses are sex- and age-adjusted. The threshold for genome-wide significance was alpha=5x10^{-8}.

66

Overlap in GWA signals

Association with asthma among the 1,699 SNPs with genome-wide significant p-values for BMI

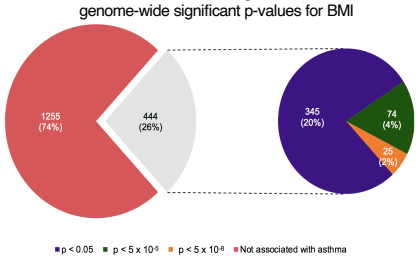
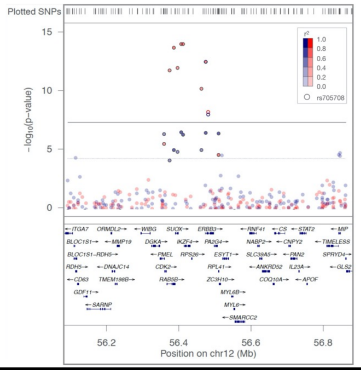


Figure 1. Overlap in GWA signals between asthma and BMI. Results for asthma are for the analysis of all asthmatic subjects (35,373 asthmatics vs. 270,572 non-asthmatics). Results for BMI are for the quantitative BMI analysis (n=305,945). Both analyses are sex- and age-adjusted. The threshold for genome-wide significance was alpha=5x10^{-8}.

67

Regional plot around rs705708 for BMI (blue) and asthma (red)



68

Cross-phenotype associations in 12q13.2

Table 2. Cross-phenotype associations in 12q13.2 [a]

					Asthma		BMI[c]	
SNP	Gene	BP	Effect/reference allele	EAF	OR (95% CI)	P[c]	beta (95% CI)	P[d]
rs2069408	*CDK2*	56,364,321	G/A	0.3388	1.04 (1.02, 1.06)	3.30×10^{-6}	-0.06 (-0.08, -0.04)	5.40×10^{-7}
rs1873914	*RAB3*	56,379,427	C/G	0.4237	1.06 (1.04, 1.08)	2.40×10^{-12}	-0.05 (-0.07, -0.02)	7.90×10^{-5}
rs705702 [b]	*SUOX*	56,390,636	G/A	0.3376	1.07 (1.05, 1.09)	3.10×10^{-14}	-0.05 (-0.08, -0.03)	1.10×10^{-5}
rs10876864 [b]	*SUOX*	56,401,085	G/A	0.4279	1.06 (1.04, 1.08)	1.50×10^{-12}	-0.05 (-0.07, -0.03)	1.60×10^{-5}
rs1701704	*IKZF4*	56,412,487	G/T	0.3433	1.07 (1.05, 1.09)	1.50×10^{-14}	-0.06 (-0.09, -0.04)	3.70×10^{-7}
rs2456973	*IKZF4*	56,416,928	C/A	0.3432	1.07 (1.05, 1.09)	1.50×10^{-14}	-0.06 (-0.08, -0.04)	6.00×10^{-7}
rs11171739 [b]	*ERBB3*	56,470,625	C/T	0.4337	1.06 (1.04, 1.07)	8.80×10^{-11}	-0.05 (-0.07, -0.03)	1.10×10^{-5}
rs2292239	*ERBB3*	56,482,180	T/G	0.3470	1.07 (1.05, 1.08)	4.50×10^{-13}	-0.06 (-0.08, -0.04)	4.20×10^{-7}
rs705708	*ERBB3*	56,488,913	A/G	0.4712	1.05 (1.03, 1.07)	7.20×10^{-9}	-0.06 (-0.09, -0.04)	1.30×10^{-6}
rs11171747 [b]	*ESYT1*	56518408	T/G	0.6180	1.04 (1.02, 1.05)	2.90×10^{-5}	-0.06 (-0.08, -0.04)	4.50×10^{-7}

Abbreviations: BP = base-pair ; BMI = body mass index; CI = confidence interval; EAF = effect allele frequency; OR = odds ratio; SNP = single-nucleotide polymorphism

a. Results shown for SNPs with $p < 5x10^{-8}$ for asthma and $p < 0.05$ for BMI.
b. For intergenic SNPs, the nearest gene is listed, with priority given to genes directly downstream of variant.
c. P-value from BOLT-LMM, derived using the standard "infinitesimal" mixed model.
d. P-value from BOLT-LMM, derived using the Gaussian mixture model.

69

Decomposing the effect of rs705708 on BMI via mediation analysis

70

Among childhood asthmatics (n=4,817) and common set of non-asthmatics (n=181,304)

total effect = -0.0656

direct effect = -0.0655

rs705708 → BMI

\+ asthma

varies by sex

indirect effect = -0.0001*

Population Average

Adult asthmatics (n=16,801) and common set of non-asthmatics (n=181,304)

total effect = -0.0560

direct effect = -0.0582

rs705708 → BMI

\+ asthma \+

indirect effect = 0.0022

Population Average

Note: Effect estimates shown are adjusted for common determinants of asthma and BMI: age, sex, breast-feeding status, exposure to maternal smoking, and smoking status at asthma diagnosis (adult analyses only). Unless otherwise noted by an asterisk(*), all paths are significant at the 0.05 level.

71

Conclusions

- rs705708 has a positive direct effect on asthma
  - Stronger in magnitude for childhood asthma
- rs705708 has a negative direct effect on BMI
  - Consistent in magnitude and direction in analyses including childhood vs. adult asthmatics
- This suggests that locus 12q13.2, tagged by rs705708, has pleiotropic effects on asthma and BMI.

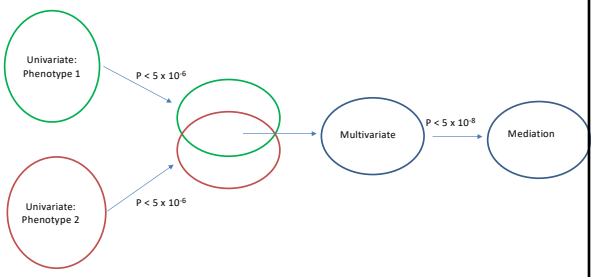
72

Conclusions

- 12q13.2 is multigenic and our CP associations span genes *CDK2, RAB5*, *SUOX*, *IZK4*, *RPS26*, *ERBB3*, and *ESYT1*.
  - rs705708 is the top regional BMI signal and resides in *ERBB3*.
  - The top regional asthma signal, rs2456973, resides in *IZKF4*.
  - While rs705708 and rs2456973 could be in LD with the same causative variant in either *ERBB3* or *IKZF4* or another gene in 12q13.2, it is also possible that each variant could tag a distinct, trait-specific causative variant in different genes.
- Therefore, locus 12q13.2 displays pleiotropic effects on asthma and BMI, but this may not be an example of pleiotropy at the gene level (biological pleiotropy).

73

Pleiotropy exercise (Part 3)



74

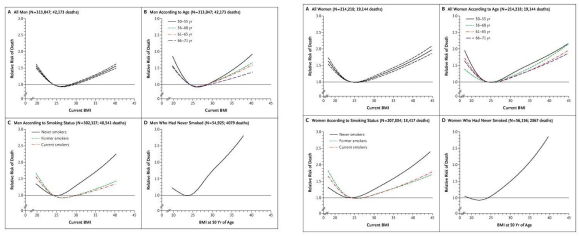
Yale

Mendelian randomization: An Introduction

Andrew DeWan, PhD, MPH
Associate Professor of Epidemiology
Director, Yale Center for Perinatal, Pediatric and Environmental Epidemiology
Yale School of Public Health

Yale SCHOOL OF PUBLIC HEALTH

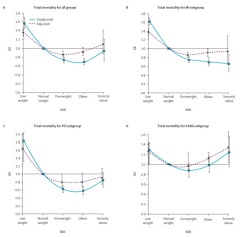
1



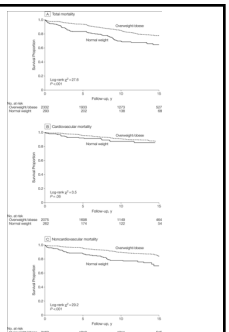
Adams et al. (2006) Overweight, Obesity and Mortality in a Large Prospective Cohort of Persons 50 to 71 Years Old. N Engl J Med 355:763-778

2

The "Obesity Paradox"



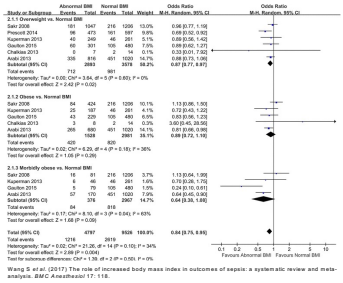
Romero-Corral A et al. (2006) Association of bodyweight with total mortality and with cardiovascular events in coronary artery disease: a systematic review of cohort studies. The Lancet 368:666-676



Carnethon M et al. (2012) Association of Weight Status With Mortality in Adults With Incident Diabetes. JAMA 308:581-590.

3

BMI and Bloodstream Infection (BSI)/Sepsis Mortality



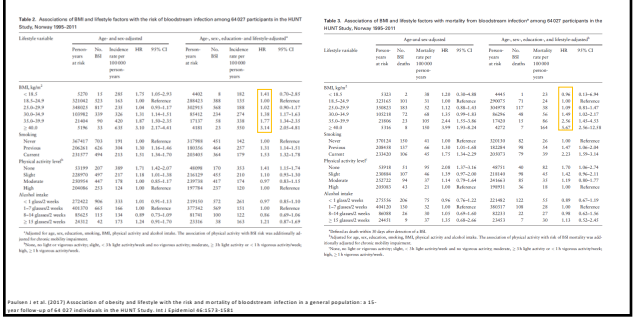
Wang S et al. (2017) The role of increased body mass index in outcomes of sepsis: a systematic review and meta-analysis. *BMC Anesthesiol* 17: 118.

4

Areas of Concern (BMI/BSI as an example)

- Selection Bias: If obesity is associated with BSI risk, non-obese patients may have other characteristics that cause their BSI that in turn are more strongly associated with mortality
- Reverse Causation: if measured BMI is affected by BSI
- Confounding: if factors such as chronic diseases and smoking habits that affect both BMI and BSI mortality are not adequately adjusted

5



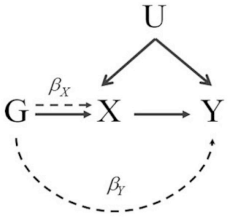
Paulsen J et al. (2017) Association of obesity and lifestyle with the risk and mortality of bloodstream infection in a general population: a 15-year follow-up of 64 027 individuals in the HUNT Study. Int J Epidemiol 46:1573-1581

6

Mendelian randomization

- Mimic randomized trial using genetic data as instruments for exposures
- Leverages information on genetic variants that segregate randomly at conception
- If an association between the instrument and outcome is detected, a causal relationship for this association is strengthened

7

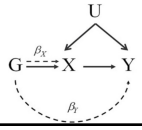


Dimou NL and Tsilidis KK. (2018) A primer in Mendelian Randomization Methodology with a Focus on Utilizing Published Summary Association Data. Methods Mol Biol. 2018;1793: 211-230

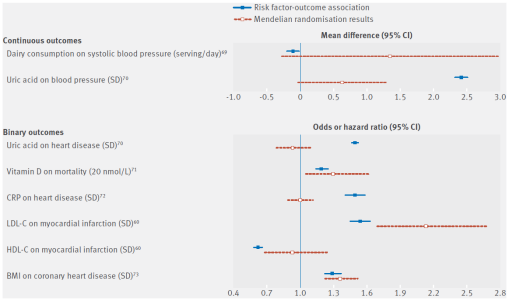
8

MR Assumptions

- The genetic instrument (G) is associated with the exposure (X)
- The genetic instrument is not associated with any confounder (U) of the exposure-outcome association
- The genetic instrument is conditionally independent of the outcome (Y) given the exposure and confounders



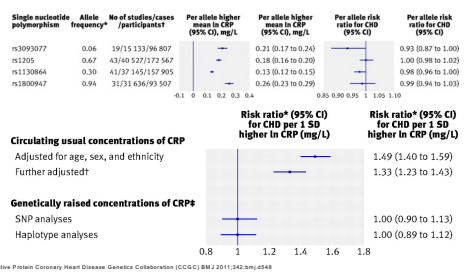
9



Davies et al. (2018) Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601

10

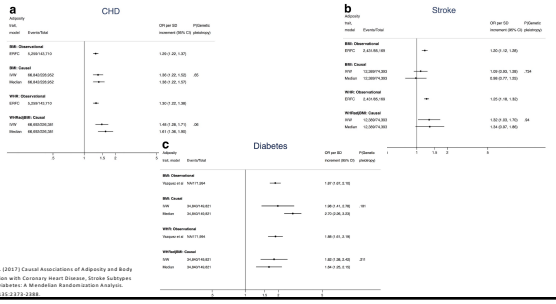
CRP and Heart Disease



C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC) BMJ 2011;342:bmj.d548

11

BMI and CHDIStrokelType 2 Diabetes



Dale CE et al. (2017) Causal Associations of Adiposity and Body Fat Distribution with Coronary Heart Disease, Stroke Subtypes and Type 2 Diabetes: A Mendelian Randomization Analysis. Circulation. 135:2373-2388.

12

One-sample vs. two-sample designs

One-sample

- Genotype(s), risk factor and outcome all measured in the same set of study subjects
- Individual level data must be available

Two-sample

- Genotype(s) and risk factor measured in one set of study subjects and genotype(s) and outcome measured in a separate set of study subjects
- Can use summary statistics or individual level data

13

One-sample vs. two-sample designs

Assumption/Issue	One-sample	Two-sample
Instrument variable related to risk factor	Weak instrument biases towards the confounded regression result	Weak instrument biases towards the null
Confounders	Can (and should) check this for measured confounders	Not often possible when using summary statistics
Pleiotropy	Multiple methods to explore this issue (including MR-Egger)	Multiple methods to explore this issue (including MR-Egger) and may be more powerful with large consortium datasets since methods tend to be statistically inefficient
Subgroup analyses	Possible if large sample sizes and data on relevant risk factors are available	Only possible if individual level data are available
Bias from adjustments made in GWAS	N/A as all adjustments made in the same set of subjects	Summary data may or may not have been adjusted

Adapted from: Lawlor DA (2016) Commentary: Two-sample Mendelian randomisation: opportunities and challenges. Int J Epi 45: 908-915.

14

Selecting genetic variants for an instrument

- Single or multiple variants
- Current recommendation is to select variant(s) that are significantly associated with the exposure at the genome-wide level
- Want a strong genetic instrument to avoid weak instrument bias
  - A single variant or variants with modest effects in small samples are likely to have low power and can suffer from bias
- If selecting multiple variants these should not be in LD and assumes negligible gene-gene interaction among variants

15

Instrument strength

- Measured using the F statistic in the regression of the IV on the exposure

$$F = \frac{N-K-1}{K} * \frac{R^2}{1-R^2}$$

R^2: proportion of the variance of the exposure explained by IV

N: sample size

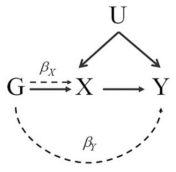
K: number of genetic variants

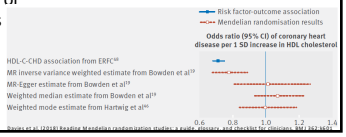
General Rule: F < 10 is an indication of a weak instrument

16

Pleiotropy

- Assumption that the IV is not associated with Y independently from X
- Presence of pleiotropy can bias the causal estimate
- Sensitivity analyses such as MR-Egger can be used to test whether or not the pleiotropy assumption has been violated





17

Testing MR: Wald Ratio

- Simple ratio of the effects of the instrument variable on the outcome over the instrument variable on the exposure
- Can be implemented in both one and two sample designs
  - One sample can use either a single variant or a GRS
  - Two sample design that uses multiple variants requires a method for combining Wald Ratios

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{ZY}}{\hat{\beta}_{ZX}}$$

18

Testing MR: 2 stage least squares (2SLS)

- Single continuous instrument (GRS)
- Only for one sample method
- Assumes a linear relationship between exposure and outcome

- Regress X on G
- Calculate genetically predicted values of X
- Regress Y on genetically predicted values of X
- Fix the standard errors (e.g. sandwich estimator)

19

Testing MR: Inverse variant weighted

- One or two sample designs
- Tends to give more reliable results in the presence of heterogeneity and when using large number of instruments
- Fixed (assumes no heterogeneity across SNP) or random effects meta-analysis

For each variant calculate the Wald ratio:

$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{Y}_j}$$

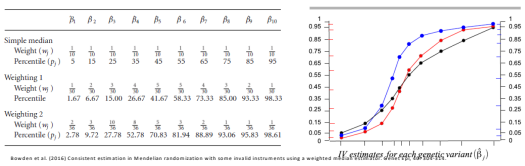
Combine into an overall estimate using a formula from meta-analysis literature:

$$\hat{\beta}_{IVW} = \frac{\Sigma_j \hat{Y}_j^2 \sigma_{Yj}^{-2} \hat{\beta}_j}{\Sigma_j \hat{Y}_j^2 \sigma_{Yj}^{-2}}$$

20

Testing MR: Weighted Median

- Calculate the Wald ratio for each instrument
- Select the median value according to the weighted method



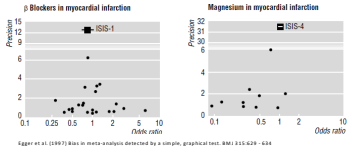
Bowden et al. (2016) Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Int J Epi, 44: 512-525

- Valid estimate when more than half of the genetic variants satisfy the IV assumptions
- No single IV contributes more than 50% of the weight

21

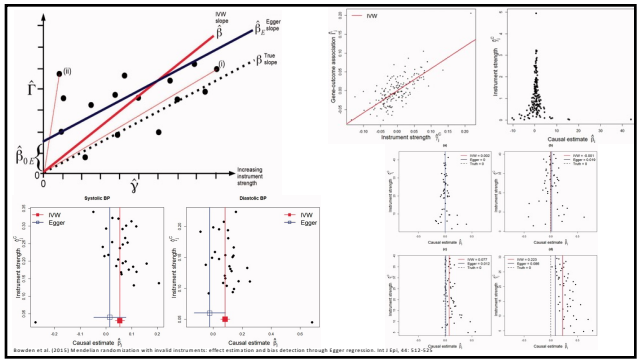
Testing MR: MR-Egger

- Provide a valid causal estimate in the presence of some violations of the MR assumptions (mainly pleiotropy)
- MR consisting of a single study with multiple IVs is analogous to a meta-analysis
- Bias resulting from pleiotropy is analogous to small study bias in meta-analysis
  - Small studies with less precise estimates tend to report larger estimates than big studies with more precise estimates
- Regress the standard normal deviate (odds ratio divided by its se) on the estimate's precision (inverse of the se)
  - Without bias, intercept = 0, and in the presence of bias the intercept is a measure of asymmetry



Egger et al. (1997) Bias in meta-analysis detected by a simple, graphical test. BMJ 315:629 - 634

22



Bowden et al. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epi, 44: 512-525

23

Databases and software

Table 3 | Databases of genome-wide association study results

Data source	Description	Number of traits	Integrated with statistics package?
MR-Base	A curated database of genome-wide association study results with integrated R package for MR[23]	Over 1000	Yes
PhenoScanner	A curated database of genome-wide association study results with integrated R package for MR[37]	Over 500	Yes
GWAS catalog	Searchable database of genome-wide association study results[38]	Over 24 000	No

Davies et al. (2018) Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601

24

Body mass index and risk of dying from a bloodstream infection: A Mendelian randomization study

Tormod Rogne[1,2,3]*, Erik Solligård[1,3], Stephen Burgess[4,5], Ben M. Brumpton[6,7,8], Julie Paulsen[9], Hallie C. Prescott[10,11], Randi M. Mohus[1,3], Lise T. Gustad[1,12], Arne Mehl[12], Bjørn O. Asvold[6,13], Andrew T. DeWan[1,2‡], Jan K. Damås[1,14,15‡]

PLOS Medicine | https://doi.org/10.1371/journal.pmed.1003413 November 16, 2020

Assess the causal association between BMI and risk of and mortality from BSI by overcoming the limitations of previous observational studies by conducting an MR study in a general population of approximately 56,000 participants in Norway with 23 years of follow-up

25

Study Population

- The Trondelag Health Study (HUNT) is a series of cross-sectional surveys carried out in Nord-Trondelag County, Norway
- 130,000 inhabitants who are representative of the general Norwegian population in terms of morbidity, mortality, sources of income and age distribution
- Based on HUNT2 survey conducted in 1995-1997 with 65,236 participants, 55,908 of whom had complete data for the analysis

26

Table 1. Background characteristics.

Characteristic	Total population (*n* = 55,908)	BSI incidence (*n* = 2,547)	BSI death (*n* = 451)
Age (years)[§]	48.3 (36.5–62.3)	63.6 (52.9–71.4)	67.3 (57.1–74.5)
Male sex*	26,324 (47.1)	1,345 (52.8)	263 (58.3)
BMI (kg/m^2)^	26.3 (4.1)	27.7 (4.5)	27.9 (4.8)
Median follow-up time (years)[§]	21.1 (17.1–21.8)	13.8 (8.4–18.3)	13.3 (7.7–17.9)
Self-reported cancer*	1,955 (3.7)	144 (6.2)	24 (5.9)
Smoking*			
Never	23,594 (43.0)	876 (35.2)	156 (35.6)
Previous	15,133 (27.6)	893 (35.8)	164 (37.4)
Current	16,117 (29.4)	723 (29.0)	118 (26.9)
Physical activity*			
None	3,821 (7.6)	243 (11.9)	54 (15.4)
Slight	15,662 (31.0)	714 (34.9)	117 (33.3)
Moderate	17,167 (34.0)	693 (33.9)	116 (33.1)
High	13,810 (27.4)	397 (19.4)	64 (18.2)
Education*			
≤9 years	19,033 (35.7)	1,305 (55.8)	240 (58.8)
10–12 years	23,468 (44.0)	762 (32.6)	125 (30.6)
≥13 years	10,832 (20.3)	274 (11.7)	43 (10.5)

BMI, body mass index; BSI, bloodstream infection. Data are presented as
^mean (standard deviation)
[§]median (25th–75th percentiles), or
* *n* (%). BSI incidence is based on first occurrence; otherwise, last occurrence is used. Education defined as follows: ≤9 years ("primary school 7–10 years, continuation school, folk high school"), 10–12 years ("high school, intermediate school, vocational school, 1–2 years high school" and "university qualifying examination, junior college, A levels"), and ≥13 years ("university or other post-secondary education, less than 4 years" and "university/college 4 years or more"). Activity defined as follows: none ("no light or vigorous activity"), slight ("<3 h light activity/week and no vigorous activity"), moderate ("≥3 h light activity/week or <1 h vigorous activity/week"), or high ("≥1 h vigorous activity/week").

27

Outcome

- Linked to all prospectively recorded blood cultures at the two community hospitals in the catchment area (Levanger and Namsos Hospitals) as well as St. Olav's Hospital in Trondheim (tertiary referral center)
- Data on blood cultures were available from January 1, 1995 through the end of 2017
- Date of death and emigration out of Nord-Trondelag County were obtained from the Norwegian population registry
- BSI was defined as a positive blood culture of pathogenic bacteria
- BSI mortality was defined as death within 30 days of BSI diagnosis

28

Genetic Instrument

- Based on a BMI meta-analysis of ~700,000 individuals (Yengo L et al. (2018) Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. Hum. Mol. Genet., 27, 3641–3649.)
- 939 of 941 SNPs identified as associated with BMI ($p<5x10^{-8}$, two SNPs did not pass imputation quality control)
- Genetic risk score (GRS) was calculated for BMI using the --score command in PLINK (version 1.9) and weighted based on the effect estimates from the meta-analysis
- GRS (939 variants) explained 4.2% of the variation in BMI in the population (F-statistic = 2,461)

29

Analysis Methods

- Fractional polynomial model (suggestion of a nonlinear relationship between BMI and BSI)
- 2-stage least squares (with sandwich estimator) for analyses assuming a linear relationship between exposure and outcome
- Sensitivity analyses
  - MR Egger (random effects)
  - INW
  - Weighted median
  - 2-sample (using Yengo et al. for SNP-exposure associations)

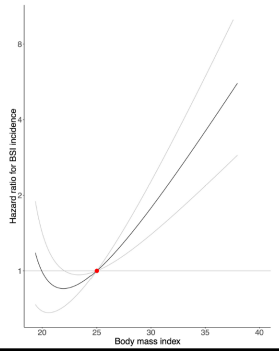
30

Table 1. Background characteristics.

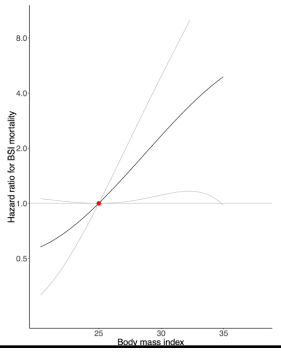
Characteristic	Total population (n = 55,908)	BSI incidence (n = 2,547)	BSI death (n = 451)
Age (years)[§]	48.3 (36.5–62.3)	65.6 (52.9–71.4)	67.3 (57.1–74.5)
Male sex*	26,324 (47.1)	1,345 (52.8)	265 (58.3)
BMI (kg/m²)[‡]	26.3 (4.1)	27.7 (4.5)	27.9 (4.8)
Median follow-up time (years)[§]	21.1 (17.1–21.8)	13.8 (8.4–18.3)	13.3 (7.7–17.9)
Self-reported cancer*	1,955 (3.7)	144 (6.2)	24 (5.9)
Smoking*			
Never	23,594 (43.0)	876 (35.2)	156 (35.6)
Previous	15,133 (27.6)	893 (35.8)	164 (37.4)
Current	16,117 (29.4)	723 (29.0)	118 (26.9)
Physical activity*			
None	3,821 (7.6)	243 (11.9)	54 (15.4)
Slight	15,662 (31.0)	714 (34.9)	117 (33.3)
Moderate	17,167 (34.0)	693 (33.9)	116 (33.1)
High	13,810 (27.4)	397 (19.4)	64 (18.2)
Education*			
≤9 years	19,033 (35.7)	1,305 (55.8)	240 (58.8)
10–12 years	23,468 (44.0)	762 (32.6)	125 (30.6)
≥13 years	10,832 (20.3)	274 (11.7)	43 (10.5)

BMI, body mass index; BSI, bloodstream infection. Data are presented as
^mean (standard deviation)
§median (25th–75th percentiles), or
*n (%). BSI incidence is based on first occurrence; otherwise, last occurrence is used. Education defined as follows: ≤9 years ("primary school 7–10 years, continuation school, folk high school"), 10–12 years ("high school, intermediate school, vocational school, 1–2 years high school" and "university qualifying examination, junior college, A levels"), and ≥13 years ("university or other post-secondary education, less than 4 years" and "university/college 4 years or more"). Activity defined as follows: none ("no light or vigorous activity"), slight (" <3 h light activity/week and no vigorous activity"), moderate (" ≥3 h light activity/week or <1 h vigorous activity/week"), or high (" ≥1 h vigorous activity/week").

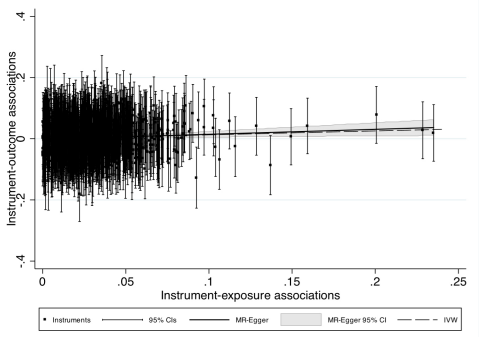
31



32



33



34

S5 Table. Mendelian randomization sensitivity analyses of linear association between body mass index and bloodstream infection mortality in the general population

	HR/OR	Lower	Upper	P-value	Intercept	Lower	Upper	P-value
One-sample								
MR-Egger, random effects	1.18	1.04	1.33	0.011	1.00	0.99	1.00	0.476
IVW, random effects	1.13	1.05	1.23	0.002	-	-	-	-
Median estimator, weighted	1.13	0.99	1.30	0.081	-	-	-	-
Two-sample								
MR-Egger, random effects	1.98	0.95	4.18	0.070	1.00	0.99	1.01	0.877
IVW, random effects	1.89	1.33	2.67	<0.001				
Median estimator, weighted	2.09	1.10	3.97	0.025				

HR, hazard ratio; IVW, inverse-variance weighted; OR, odds ratio. Assuming a linear relationship between body mass index and bloodstream infection mortality in the general population using the same 959 single nucleotide polymorphisms (SNP) as used to create the genetic risk score. Two-sample analyses use SNP-exposure associations from Yengo et al [ref 2 in Supplementary text], and SNP-outcome associations from HUNT. The F of the SNP-exposure associations were 54% in the one-sample MR-Egger regression, and 92% in the two-sample MR-Egger regression. Effect estimates reported as HR for one unit increase of body mass index in one-sample analyses and as OR for one standard deviation increase of body mass index in two-sample analyses.

35

STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies

Authors (in alphabetical order):

George Davey Smith, Neil M Davies, Niki Dimou, Matthias Egger, Valentina Gallo, Robert Golub, Julian PT Higgins, Claudia Langenberg, Elizabeth W Loder, J Brent Richards, Rebecca C Richmond, Veronika W Skrivankova, Sonja A Swanson, Nicholas J Timpson, Anne Tybjaerg-Hansen, Tyler J VanderWeele, Benjamin AR Woolf, James Yarmolinsky

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27857v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

36

INTRODUCTION

2. Background

Explain the scientific background and rationale for the reported study. Is causality between exposure and outcome plausible? Justify why MR is a helpful method to address the study question.

METHODS

4. Study design and data sources

Present key elements of study design early in the paper. Consider including a table listing sources of data for all phases of the study. For each data source contributing to the analysis, describe the following:

a) Describe the study design and the underlying population from which it was drawn. Describe also the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection, if available.

b) Give the eligibility criteria, and the sources and methods of selection of participants.

c) Explain how the analyzed sample size was arrived at.

d) Describe measurement, quality and selection of genetic variants.

e) For each exposure, outcome and other relevant variables, describe methods of assessment and, in the case of diseases, the diagnostic criteria used.

f) Provide details of ethics committee approval and participant informed consent, if relevant.

5. Assumptions

Explicitly state assumptions for the main analysis (e.g. relevance, exclusion, independence, homogeneity) as well assumptions for any additional or sensitivity analysis.

37

6. Statistical methods: main analysis

Describe statistical methods and statistics used.

a) Describe how quantitative variables were handled in the analyses (i.e., scale, units, model).

b) Describe the process for identifying genetic variants and weights to be included in the analyses (i.e, independence and model). Consider a flow diagram.

c) Describe the MR estimator, e.g. two-stage least squares, Wald ratio, and related statistics. Detail the included covariates and, in case of two-sample MR, whether the same covariate set was used for adjustment in the two samples.

d) Explain how missing data were addressed.

e) If applicable, say how multiple testing was dealt with.

7. Assessment of assumptions

Describe any methods used to assess the assumptions or justify their validity.

8. Sensitivity analyses

Describe any sensitivity analyses or additional analyses performed.

RESULTS

10. Descriptive data

d) For two-sample Mendelian randomization:

i. Provide information on the similarity of the genetic variant-exposure associations between the exposure and outcome samples.

ii. Provide information on extent of sample overlap between the exposure and outcome data sources.

38

11. Main results

a) Report the associations between genetic variant and exposure, and between genetic variant and outcome, preferably on an interpretable scale (e.g. comparing 25th and 75th percentile of allele count or genetic risk score, if individual-level data available).

b) Report causal effect estimate between exposure and outcome, and the measures of uncertainty from the MR analysis. Use an intuitive scale, such as odds ratio, or relative risk, per standard deviation difference.

c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time-period.

d) Consider any plots to visualize results (e.g. forest plot, scatterplot of associations between genetic variants and outcome versus between genetic variants and exposure).

12. Assessment of assumptions

a) Assess the validity of the assumptions.

b) Report any additional statistics (e.g., assessments of heterogeneity, such as I^2, Q statistic).

13. Sensitivity and additional analyses

a) Use sensitivity analyses to assess the robustness of the main results to violations of the assumptions.

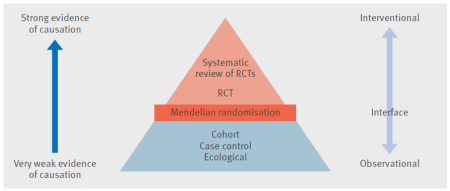
b) Report results from other sensitivity analyses (e.g., replication study with different dataset, analyses of subgroups, validation of instrument(s), simulations, etc.).

c) Report any assessment of direction of causality (e.g., bidirectional MR).

d) When relevant, report and compare with estimates from non-MR analyses.

e) Consider any additional plots to visualize results (e.g., leave-one-out analyses).

39



Davies et al. (2018) Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601

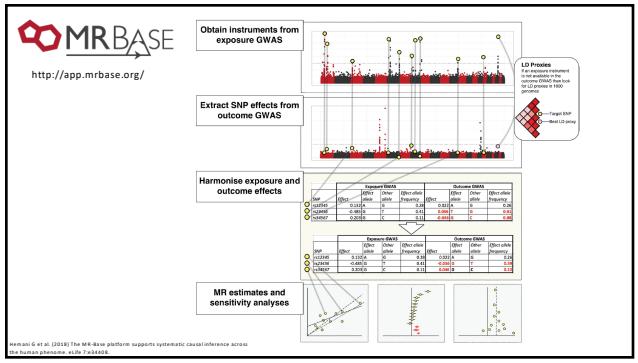
40

Limitations

- Exposure randomization may not be truly random
- Unmeasured or residual confounding (population stratification, parental genotype associated with outcome)
- Weak instrument bias resulting from measurement error for the exposure of interest
- Adaptation to the exposure
- Inconsistent results and selective publication

Mukamal KJ et al. (2020) Genetic instrument variable analysis: time to call Mendelian randomization what it is. The example of alcohol and cardiovascular disease. Eur J of Epidemiol 35:93-97.

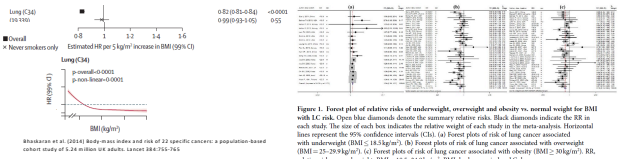
41



Hemani G et al. (2018) The MR-Base platform supports systematic causal inference across the human phenome. eLife 7:e34408.

42

BMI and Lung Cancer



Khaskaran et al. (2014) Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. Lancet 384:755–765

Figure 1. Forest plot of relative risks of underweight, overweight and obesity vs. normal weight for BMI with LC risk. Open blue diamonds denote the summary relative risks. Black diamonds indicate the RR in each study. The size of each box indicates the relative weight of each study in the meta-analysis. Horizontal lines represent the 95% confidence intervals (CIs). (a) Forest plots of risk of lung cancer associated with underweight (BMI ≤ 18.5kg/m²). (b) Forest plots of risk of lung cancer associated with overweight (BMI = 25–29.9kg/m²). (c) Forest plots of risk of lung cancer associated with obesity (BMI ≥ 30kg/m²). RR, relative risk; normal weight, BMI = 18.5–24.9kg/m²; BMI, body mass index; LC, lung cancer; m, men; w, women; mw, men and women.

Duan et al. (2015) Body mass index and risk of lung cancer: Systematic review and dose-response meta-analysis. Sci Rep 5:14988

43

Newcastle University

wellcome

Genome-wide association studies (GWAS) - Part 2

More advanced topics: Linear Mixed Models and G×G or G×E interactions

Heather J. Cordell

Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University, UK

heather.cordell@ncl.ac.uk

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories

Linear Mixed Models (LMMs)

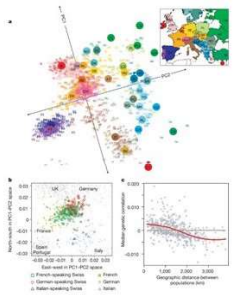
- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories
  - Investigating genetic correlations between different traits

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories
  - Investigating genetic correlations between different traits
  - Predicting trait values in a new individual

Population stratification and relatedness

Genes mirror geography within Europe



J Novembre *et al.* (2008) *Nature* **456(7218):98-101**, doi:10.1038/nature07331

Linear Mixed Models (LMMs)

- A linear mixed model is a statistical model in which the dependent variable is a linear function of both fixed and random independent variables
  - Known respectively as fixed and random effects
  - Fixed effects are considered 'fixed' at their measured values
  - Random effects are considered to be sampled from a distribution

Linear Mixed Models (LMMs)

- A linear mixed model is a statistical model in which the dependent variable is a linear function of both fixed and random independent variables
  - Known respectively as fixed and random effects
  - Fixed effects are considered 'fixed' at their measured values
  - Random effects are considered to be sampled from a distribution
- Recall the usual linear regression model

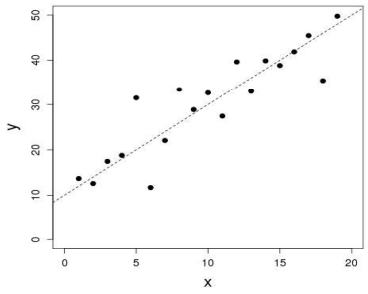
$$y = mx + c \qquad \text{or} \qquad y = \beta_0 + \beta_1 x$$

- This model may also be written

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  - y_i refers to the trait value of person i
  - x_i refers to the measured value of person i's predictor variable
  - ϵ_i refers to the displacement from the regression line
    - i.e. the discrepency between the observed and the predicted y value

Linear Regression



Linear Mixed Models (LMMs)

- In linear regression we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Here β_0 and β_1 are fixed effects while ϵ_i is a random error
  - x_i is the 'loading' of the fixed effect that someone has (based on their genotype)
- In matrix notation we can write this model:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

Linear Mixed Models (LMMs)

- In linear regression we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Here β_0 and β_1 are fixed effects while ϵ_i is a random error
  - x_i is the 'loading' of the fixed effect that someone has (based on their genotype)
- In matrix notation we can write this model:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- A LMM takes the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$
  - where $\mathbf{u}$ corresponds to a vector of random effects
    - with loadings specified in $\mathbf{Z}$

Linear Mixed Models (LMMs)

- E.g. suppose 2 fixed effects β_1 and β_2, and 3 random effects (plus n random errors)
- Then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$ corresponds to:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ . & . \\ . & . \\ x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ . & . & . \\ . & . & . \\ z_{n1} & z_{n2} & z_{n3} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + u_1 z_{i1} + u_2 z_{i2} + u_3 z_{i3} + \epsilon_i$

LMMs in genetics

- In genetics we generally work with two equivalent forms of LMM
- One is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$
  - The random effect u_l corresponds to a scaled additive effect of causal variant (locus) l
    - Assuming many (m) such causal variants all across the genome
  - $\mathbf{Z}$ is a standardized genotype matrix i.e. z_{il} takes value

$$\left(\frac{-2f_l}{\sqrt{2f_l(1-f_l)}}, \frac{(1-2f_l)}{\sqrt{2f_l(1-f_l)}}, \frac{2(1-f_l)}{\sqrt{2f_l(1-f_l)}} \right)$$

    if individual i has genotype (qq, Qq, QQ)
    - where f_l is the frequency of allele Q at locus l

LMMs in genetics

- The other form is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon}$
  - Where $g_i = \sum_{l=1}^{m} z_{il} u_l$ is the total genetic effect in individual i, summed over all the causal loci
- In this form, g_i **can be considered as a random effect operating in individual** i
  - The vector of random effects $\mathbf{g}$ takes distribution $\mathbf{g} \sim N(0, \mathbf{G}\sigma_a^2)$
    - Where $\mathbf{G}$ is the genetic relationship matrix (GRM) between individuals – i.e. their IBD sharing at the causal loci
    - $\sigma_a^2 = m\sigma_u^2$ is the total additive genetic variance
    - $\mathbf{G}$=$\mathbf{ZZ'}/m$

LMMs in genetics

- The other form is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon}$
  - Where $g_i = \sum_{l=1}^{m} z_{il} u_l$ is the total genetic effect in individual i, summed over all the causal loci
- In this form, g_i **can be considered as a random effect operating in individual** i
  - The vector of random effects $\mathbf{g}$ takes distribution $\mathbf{g} \sim N(0, \mathbf{G}\sigma_a^2)$
    - Where $\mathbf{G}$ is the genetic relationship matrix (GRM) between individuals – i.e. their IBD sharing at the causal loci
    - $\sigma_a^2 = m\sigma_u^2$ is the total additive genetic variance
    - $\mathbf{G}$=$\mathbf{ZZ'}/m$
- For family data (close relatives), the expected values of the elements of $\mathbf{G}$ equal the expected IBD sharing
  - i.e. twice the kinship coefficients
  - Thus $\mathbf{G}$ is just equal to twice the kinship matrix
    - Models their expected relatedness at the causal loci (and elsewhere)

Use of LMMs in genetics

- The formulation $\mathbf{y} = \mathbf{X}\beta + \mathbf{g} + \epsilon$ is known as the **Animal Model** and has been used extensively in plant and animal breeding
  - Mostly to predict the *breeding values* g_i in order to inform breeding strategies
    - E.g. to increase milk yield, meat production etc. etc.
  - Similar approaches could be used for *prediction* of trait values given genotype data
- In the mid 1990s it became popular in human genetics as the backbone of **variance components linkage analysis**
- Now commonly used in association analysis (GWAS)
  - To correct for relatedness, when testing for association

Testing for association using LMMs

- Idea is to test a fixed SNP effect β_1
  - While including a random effect γ_i that models relatedness
- Fit regression model: $y_i = \beta_0 + \beta_1 x_i + \gamma_i$
  - y is the trait value
  - x is a variable coding for genotype at the test SNP (e.g. an allele count, coded 0, 1, 2 for genotypes 1/1, 1/2, 2/2)
  - $\gamma_i = g_i + \epsilon_i$

Testing for association using LMMs

- Idea is to test a fixed SNP effect β_1
  - While including a random effect γ_i that models relatedness
- Fit regression model: $y_i = \beta_0 + \beta_1 x_i + \gamma_i$
  - y is the trait value
  - x is a variable coding for genotype at the test SNP (e.g. an allele count, coded 0, 1, 2 for genotypes 1/1, 1/2, 2/2)
  - $\gamma_i = g_i + \epsilon_i$
  - We assume $\gamma \sim MVN(0, \mathbf{V})$ where variance/covariance matrix $\mathbf{V}$ follows standard variance components model
    - Variance/covariance matrix structured as:

$$V_{ij} = \sigma_a^2 + \sigma_e^2 \quad (i = j)$$
$$V_{ij} = 2\Phi_{ij}\sigma_a^2 \quad (i \neq j)$$

    - σ_a^2, σ_e^2 represent the additive polygenic variance (due to all loci) and the environmental (=error) variance, respectively

Testing for association using LMMs

- LMMs were first (?) applied in human genetics by Boerwinkle et al. (1986) and Abney et al. (2002)
- Chen and Abecasis (2007) implemented them via the "FAmily based Score Test Approximation" (FASTA) in the MERLIN software package
  - Closely related to earlier QTDT method (Abecasis et al. 2000a;b) which implements a slightly more general/complex model
  - FASTA was also implemented in GenABEL, along with a similar test called GRAMMAR (Aulchenko et al. 2007)

Estimating the genetic relationship matrix

- These early implementations calculated the kinship matrix $\mathbf{\Phi}$ on the basis of known (theoretical) kinships constructed from known pedigree relationships
- Amin et al. (2007) proposed instead *estimating* the kinships based on genome-wide SNP data
  - Ideally we want to use $\mathbf{G}=\mathbf{ZZ}'/m$, the genetic relationship matrix (GRM) between individuals at the causal loci
  - Since we don't know the causal loci, we approximate $\mathbf{G}$ by $\mathbf{A}$, the overall GRM between individuals
    - Various different ways to estimate this, usually based on scaled (by allele frequency) matrix of *identity-by-state* (IBS) sharing

Estimating the genetic relationship matrix

- Once you move to estimating the GRM, you are no longer limited to using family data
- Kang et al. (2010) and Zhang et al. (2010) suggested applying the approach to apparently unrelated individuals
  - As a way of accounting for population substructure/stratification
  - Also proposed applying to binary traits (case/control coded 1/0)
  - Implemented in EMMAX and TASSEL software, respectively

Estimating the genetic relationship matrix

- Once you move to estimating the GRM, you are no longer limited to using family data
- Kang et al. (2010) and Zhang et al. (2010) suggested applying the approach to apparently unrelated individuals
  - As a way of accounting for population substructure/stratification
  - Also proposed applying to binary traits (case/control coded 1/0)
  - Implemented in EMMAX and TASSEL software, respectively
- Subsequently a number of other publications/software packages have implemented essentially the same model
  - FaST-LMM (Lippert et al. 2011)
  - GEMMA (Zhou and Stephens 2012)
  - GenABEL (GRAMMAR-Gamma) (Svishcheva et al. 2012)
  - MMM (Pirinen et al. 2013)
  - MENDEL (Zhou et al. 2014)
  - RAREMETALWORKER
  - GCTA
  - DISSECT

Software implementations

- Main difference between them is the precise computational tricks used to speed up the calculations
  - And the convenience/ease of use
    - See comparison in Eu-Ahsunthornwattana et al. (2014) PLoS Genetics 10(7):e1004445

Software implementations

- Main difference between them is the precise computational tricks used to speed up the calculations
  - And the convenience/ease of use
    - See comparison in Eu-Ahsunthornwattana et al. (2014) PLoS Genetics 10(7):e1004445
- BOLT-LMM (Loh et al. 2016) uses a slightly different approach, based on a Bayesian implementation of LMM formulation 1:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$$

  - One of the first mixed model packages that worked for really large-scale (e.g. UK Biobank) datasets
  - Now potentially (?) superseded by fastGWA module in GCTA
  - And by REGENIE (https://doi.org/10.1038/s41588-021-00870-7), which uses a slightly different formulation based on analysing the residuals following a whole-genome blockwise ridge regression
    - Again based on LMM formulation 1: $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$

Binary traits

- For binary traits, coding cases and controls as a 1/0 quantitative trait is not optimal
  - Though in practice it seems to work reasonably well
- LTMLM (Hayeck et al. 2015) and LEAP (Weissbrod et al. 2015) instead use an underlying *liability model* to improve power
  - Assuming known disease prevalence

Binary traits

- For binary traits, coding cases and controls as a 1/0 quantitative trait is not optimal
  - Though in practice it seems to work reasonably well
- LTMLM (Hayeck et al. 2015) and LEAP (Weissbrod et al. 2015) instead use an underlying *liability model* to improve power
  - Assuming known disease prevalence
- Chen et al. (2016) showed that high levels of population stratification can invalidate the analysis, when applied to a case/control sample
  - Resulting in a mixture of **inflated** and **deflated** test statistics
  - Developed GMMAT software to address this problem
  - See also CARAT software (Jiang et al. 2016, AJHG 98:243-55)

Binary traits

- SAIGE software (Zhou et al. 2018, AJHG 50(9):1335-1341) implements a mixed model test that deals with large case-control imbalance, as you might see (for example) in UK Biobank
- REGENIE also implements this same saddle point approximation (SPA) test
  - Along with an approximate Firth penalized likelihood-ratio test

Elucidating genetic architecture

- Seminal paper by Yang et al. (2010) [Nat Genet 42(7):565-9]
- Showed that by framing the relationship between height and genetic factors as an LMM, 45% of variance could be explained by considering 294,831 SNPs simultaneously
  - So-called 'SNP heritability' or 'chip heritability'
  - Demonstrated that modelling effects at all genotyped SNPs explained the 'known' heritability ($\approx 80\%$) much better than just the top SNPs from GWAS
- Moreover, if you estimate effects of additional SNPs in LD with the genotyped SNPS, the variance explained goes up to 84% (s.e. 16%), consistent with 'known' value
- Subsequently many papers have shown similar results for a variety of complex traits

Elucidating genetic architecture

- Basic idea is to use formulation

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{g} + \epsilon$$

with $\mathbf{g} \sim N(0, \mathbf{A}\sigma_a^2)$ and $\epsilon \sim N(0, \mathbf{I}\sigma_e^2)$ so $\mathbf{V} = \mathbf{A}\sigma_a^2 + \mathbf{I}\sigma_e^2$

  - $\mathbf{A}$ is the GRM between individuals, estimated using all genotyped SNPs
  - σ_a^2 and σ_e^2 estimated using REML (or MLE)
  - Thus we can estimate heritability accounted for by the genotyped SNPs as $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$
- Implemented in several software packages including GCTA and DISSECT
  - ALBI software (Schweiger et al. 2016, AJHG 98:1181-1192) can then be used to construct accurate confidence intervals for the heritability

Partitioning variance

- The same formulation can be used to partition the variance explained by different subsets of SNPs
  - Yang et al. (2010) partitioned variance onto each of the 22 autosomes using formulation

$$\mathbf{y} = \mathbf{X}\beta + \sum_{c=1}^{22} \mathbf{g_c} + \epsilon \qquad \text{with } \mathbf{V} = \sum_{c=1}^{22} \mathbf{A_c}\sigma_c^2 + \mathbf{I}\sigma_e^2,$$

    where $\mathbf{g_c}$ is a vector of effects attributed to the cth chromosome, and $\mathbf{A_c}$ is the GRM estimated from SNPs on the cth chromosome
    - Slight adjustment is needed for estimating variance explained by SNPs on chromosome X
- Similar partitioning can be used to examine subsets of SNPs defined in other ways e.g. according to MAF or functional annotation

Other approaches

- Some recent work has focussed on achieving similar ends
  - i.e. estimating
    - heritability explained by sets of SNPs
    - genetic correlations across traits

  using summary statistics only
  - Bulik-Sullivan et al. (2015) [Nat Genet 47:291-295]
  - Bulik-Sullivan et al. (2015) [Nat Genet 47:1236-1241]
    - Clever idea that allows the variance component parameters to be estimated via a simple regression on 'LD Scores'

Short break

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198
- Effects operating through interactions may not be visible unless you stratify by or take account of the interacting genetic (or environmental) factors
  - By modelling interactions, we hope to increase our power to detect loci with weak marginal effects

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198
- Effects operating through interactions may not be visible unless you stratify by or take account of the interacting genetic (or environmental) factors
  - By modelling interactions, we hope to increase our power to detect loci with weak marginal effects
- Phenomenon of biological interest?
  - Identifying genes that interact to cause disease could help us understand the mechanisms and pathways in disease progression

Definition of (pairwise) interaction

- Statistical interaction most easily described in terms a of (logistic) regression framework
  - Supppose x_1 and x_2 are binary factors whose presence/absence (coded 1/0) may be associated with a disease outcome
  - Logistic regression models their effect on the log odds of disease as:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

Marginal effect of factor 1

$$\log \frac{p}{1-p} = \beta_0 + \beta_2 x_2$$

Marginal effect of factor 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Main effects of factors 1 and 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Main effects and interaction term

  - For quantitative traits, use linear regression (replace $\log \frac{p}{1-p}$ with y)
  - For modelling as an LMM, add in a random effect γ

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value
      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'
    - The 'effect' of factor 2 is different in the presence/absence of factor 1

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value
      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'
    - The 'effect' of factor 2 is different in the presence/absence of factor 1
- Suppose no main effects ($\beta_1 = \beta_2 = 0$)

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_{12}$	β_0
0	β_0	β_0

  - Trait value only differs from baseline if both factors present

Gene-gene interaction (epistasis)

- However SNPs are not binary, but rather take 3 levels according to the number of copies (0,1,2) of the susceptibility allele possessed
- Most general 'saturated' (9 parameter) genotype model allows all 9 penetrances to take different values
  - Via modelling log odds in terms of:
    - A baseline effect (β_0)
    - Main effects of locus G (β_{G_1}, β_{G_2})
    - Main effects of locus H (β_{H_1}, β_{H_2})
    - 4 interaction terms

	Locus H		
Locus G	2	1	0
2	$\beta_0+\beta_{G_2}+\beta_{H_2}+\beta_{22}$	$\beta_0+\beta_{G_2}+\beta_{H_1}+\beta_{21}$	$\beta_0+\beta_{G_2}$
1	$\beta_0+\beta_{G_1}+\beta_{H_2}+\beta_{12}$	$\beta_0+\beta_{G_1}+\beta_{H_1}+\beta_{11}$	$\beta_0+\beta_{G_1}$
0	$\beta_0+\beta_{H_2}$	$\beta_0+\beta_{H_1}$	β_0

  - Corresponds in statistical analysis packages to coding x_1, x_2 (0,1,2) as a "factor"

Gene-gene interaction

- Alternatively we can assume additive effects of each allele at each locus:
  - Corresponds to fitting

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

  with x_1, x_2 coded (0,1,2)

	Locus H		
Locus G	2	1	0
2	$\beta_0 + 2\beta_G + 2\beta_H + 4\beta_{GH}$	$\beta_0 + 2\beta_G + \beta_H + 2\beta_{GH}$	$\beta_0 + 2\beta_G$
1	$\beta_0 + \beta_G + 2\beta_H + 2\beta_{GH}$	$\beta_0 + \beta_G + \beta_H + \beta_{GH}$	$\beta_0 + \beta_G$
0	$\beta_0 + 2\beta_H$	$\beta_0 + \beta_H$	β_0

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way
  - Good starting point for further investigation of their (joint) action

Gene-environment (G×E) interactions

- The same regression model

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

can be used to model interaction between a genetic factor G and an environmental factor H

  - With the environmental variable x_2 coded in binary fashion (e.g. smoking) or quantitatively (e.g. age)

Gene-environment (G×E) interactions

- The same regression model

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

can be used to model interaction between a genetic factor G and an environmental factor H

  - With the environmental variable x_2 coded in binary fashion (e.g. smoking) or quantitatively (e.g. age)

- Focus of analysis is often risk estimation
  - Estimating genetic risks in particular environments
  - Estimating effect of environmental factor on particular genetic background
    - Important for treatment/screening strategies and public health interventions
- For G×G, focus of interest is more related to
  - Increasing power to detect an effect (by taking into account the effects of other genetic loci)
  - Modelling the biology, especially related to the joint action of the loci

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone
- Depending on circumstances, any of these tests may be a sensible option

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone
- Depending on circumstances, any of these tests may be a sensible option
- Most tests of interaction/joint action can be thought of as a version of one or other of these tests
  - Although different tests vary in their precise details
  - And their relationship to the logistic regression formulation not always clearly described
    - See Howey and Cordell (2017) https://pubmed.ncbi.nlm.nih.gov/28852712/

G×G versus G×E in the context of GWAS

- Typically GWAS measure thousands if not millions of genetic variants
  - But only a few (tens or at most 100s) of environmental factors
- Feasible to consider all G×E combinations
- All pairwise G×G combinations possible, but much more time consuming
  - And leads to greater multiplicity of tests
  - Also, why stop at 2-way interactions?
    - Could look at all 3 way, 4 way etc. combinations
    - Scale of problem quickly gets out of hand
    - Less obvious reason to do this for G×E...

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"

- Or have proposed filtering based on single-locus significance or other (biological or statistical) considerations
  - Reduces multiple testing burden, improves interpretability

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"
- Or have proposed filtering based on single-locus significance or other (biological or statistical) considerations
  - Reduces multiple testing burden, improves interpretability

- Or have proposed testing at the gene level rather than the SNP level
  - Ma et al. (2013) PLoS Genet 9(2): e1003321
    - Compared 4 different tests that combine P values from pairwise (SNP x SNP) interaction tests
    - Showed that the truncated tests did best
    - Presented an application only considering gene pairs known to exhibit protein-protein interactions

Case-only analysis

- Piergorsh et al. 1994; Yang et al. 1999; Weinberg and Umbach 2000
- Several authors have shown that, for binary predictor variables, a test of the interaction term β_{12} in the logistic regresssion model

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  can be obtained by testing for correlation (association) between the genotypes at two separate loci, within the sample of cases
- Gains power from making assumption that genotypes (alleles) at the two loci are uncorrelated in the population
  - So only really suitable for unlinked or loosely linked loci (since closely linked loci are likely to be in LD)
- Alternatively contrast the genotype correlations in cases with those seen in controls (`--fast-epistasis` in PLINK)

Testing correlation between loci

- A similar idea is implemented in EPIBLASTER (Kam-Thong et al. 2011; EJHG 19:465-571)
- Wu et al. (2010) (PLoS Genet 6:e1001131) also proposed a similar approach – though needs adjustment to give correct type I error rates
- See also Joint Effects (JE) statistics (Ueki and Cordell 2012; PLoS Genetics 8(4):e1002625)
- All these methods test whether correlation exists (case-only) or is different in cases and controls (case/control)
  - Via testing a log OR for association between two loci
  - However, the log OR for association (λ) encapsulates a slightly different quantity between the different methods
- All implemented (along with standard logistic and linear regression) in CASSI
  - http://www.staff.ncl.ac.uk/richard.howey/cassi/

Empirical evidence for G×G interactions

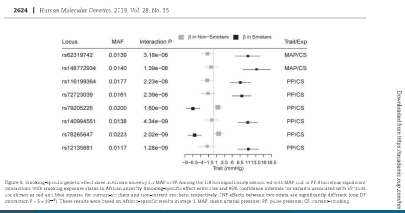
- Epistasis among *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* in multiple sclerosis (Lincoln et al. 2009 PNAS 106:7542-7547)
- *HLA-C* and *ERAP1* in psoriasis (Strange et al. 2010)
- *HLA-B27* and *ERAP1* in ankylosing spondylitis (Evans et al. 2011)
- *BANK1* and *BLK* in SLE (Castillejo-Lopez et al. 2012)
- Gusareva et al. (2014) found a reasonably convincing (partially replicating) interaction between SNPs on chromosome 6 (*KHDRBS2*) and 13 (*CRYL1*) in Alzheimer's disease
- Dai et al. (2016) [AJHG 99:352-365] identified 3 loci simultaneously interacting with established risk factors gastresophageal reflux, obesity and tobacco smoking, with respect to risk for Barrett's esophagus

Empirical evidence for G×G interactions

- Hemani et al. 2014 (Nature 508:249-253) found 501 instances of epistatic effects on gene expression, of which 30 could be replicated in two independent samples
  - Many SNPs are close together, could represent haplotype effects?
  - Or the effect of a single untyped variant?
  - See caveats in
    - Wood et al. (2014) Nature 514(7520):E3-5. PMID:25279928
    - Fish et al. (2016) Am J Hum Genet 99(4):817830. PMID:27640306
- The Hemani et al. paper was subsequently retracted (https://www.nature.com/articles/s41586-021-03766-y)

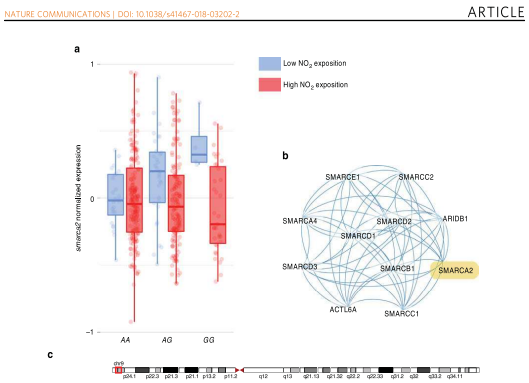
Empirical evidence for G×E interactions

- Myers et al. (2014) Hum Mol Genet 23(19): 5251-9 "Genome-wide Interaction Studies Reveal Sex-Specific Asthma Risk Alleles"
- Small et al. (2018) Nat Genet 50(4): 572-580 "Regulatory Variants at KLF14 Influence Type 2 Diabetes Risk via a Female-Specific Effect on Adipocyte Size and Body Composition"
- Sung et al. (2019) Hum Molec Genet 28(15): 2615-2633 "A multi-ancestry genome-wide study incorporating gene-smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure."



Empirical evidence for G×E interactions

- Favé et al. (2018) Nat Commun 9(1): 827 "Gene-by-environment Interactions in Urban Populations Modulate Risk Phenotypes"



Power Analysis for Single and Rare Variant Aggregate Association Analyses

1

Why Estimate Sample Sizes and/or Power?

- To avoid wasting time and money
  - Does not make sense to perform an inadequately powered study for which it is unlikely to to correctly reject the null hypothesis due to inadequate sample size
    - Collaborations can aid in increasing sample sizes
      - Caveats
        - » Disease definition may not be the same between studies
        - » Study subjects may be drawn for different populations
        - » Processing of genetic material maybe not be consistent
- Almost always necessary for grant proposals
  - Can be denied funding if unable to demonstrate planned study has adequate power
    - Realistic disease models are necessary when performing power calculations
    - Correctly adjust alpha for multiple testing which will be performed
      - e.g., use genome-wide significant level of 5×10^{-8} for GWAS studies

2

Power and Sample Size Estimation for Case-Control Data

- The correct α must be use for sample size estimation/power analysis
- Type I (α) the probability of rejecting the null hypothesis of no association when it is true
- Due to multiple testing a more stringent value than α=0.05 is used in order to control the Family Wise Error Rate

3

Power and Sample Size Estimation for Case-Control Data

- GWAS of common variants where each variant is test separately
  - $\alpha=5 \times 10^{-8}$ (Bonferroni Correction for testing 1,000,000 variant sites)
  - Shown to be a good approximation for the effective number of tests
    - Valid even when more than 1,000,000 variant sites tested
  - Effective number of tests is dependent of the linkage disequilibrium (LD) structure
- Single variant tests using whole genome sequence data
  - Many more rare variants than common variants
    - Lower levels of LD between rare variants than between common variants
  - The number of effective tests for rare variants is higher than for analysis limited to common variants
  - α is yet to be determined for association analysis of whole genome sequence data

4

An Example of Determining Genome-wide Significance Levels for Common Variants

- Using genotypes from the Wellcome Trust Case-Control Consortium
- Dudbridge and Gusnato, Genet Epidemiol 2008
- Estimated a genome-wide significance threshold for the UK European population
- By sub-sampling genotypes at increasing densities and using permutation to estimate the nominal p-value for a 5% family-wise error
- Then extrapolating to infinite density
- The genome wide significance threshold estimate $\sim 7.2 \times 10^{-8}$
- Estimate is based on LD structure for Europeans
  - Not sufficiently stringent for populations of African Ancestry

5

Power and Sample Size Estimation for Aggregate Rare Variant Tests

- For gene-based rare variant aggregate methods a Bonferroni correction for the number of genes/regions tested is used
  - e.g., 20,000 genes significance level $\alpha=2.5 \times 10^{-6}$
    - Can use a less stringent criteria
      - Not all genes have two or more variants
        - » Divide 0.05 by number of genes tested
    - If units other than genes are used
      - A more stringent criteria may be necessary
- For rare variants – very low levels of LD between variants in separate genes
  - Therefore, a Bonferroni correction is not overly stringent
    - The number of tests ≅ effective number tests
      - This would not be the case for variants in LD

6

Power and Sample Size Estimation for Replication Studies

- For replication studies can base the significance level (α)
- On the number of genes/variants being brought from the discovery (stage I) study
- To replication (stage II)
- For example, if it is hypothesized that 20 genes and 80 independent variants will be brought to stage II (replication)
  - A Bonferroni correct can be made for performing 100 tests
    - An $\alpha = 5.0 \times 10^{-3}$ can be used for a family wise error rate of 0.05

7

Estimating Power/Sample Sizes For Single Variant Tests

- Can be obtained analytically
- Information necessary
  - Prevalence
  - Risk allele frequency
  - Effect size (odds ratio-for case control data)
  - Genetic model for the susceptibility variant
    - Recessive ($\gamma_1=1$)
    - Dominant ($\gamma_2=\gamma_1$)
    - Additive ($\gamma_2=2\gamma_1-1$)
    - Multiplicative ($\gamma_2=\gamma_1^2$)

8

Estimating Power/Sample Sizes For Individual Variants

- Usually, information on disease prevalence is known from epidemiological data
- A range of risk allele allele frequencies and effect sizes are used
- A variety of genetic models can also used
  - Dominant
  - Additive
  - Multiplicative

9

Armitage Trend Test

- Power and Sample size
  - Calculated under different models
    - Where γ is the relative risk
      - Multiplicative
        - » $\gamma_2=\gamma_1^2$
      - Additive
        - » $\gamma_2=2\gamma_1-1$
      - Dominant
        - » $\gamma_2=\gamma_1$
      - Recessive
        - » $\gamma_1=1$

10

Gamma is the Relative Risk not the Odd Ratio

- Most software for power calculations/sample size estimation use the relative risk (γ) and not the odds ratio
- The relative risk only approximates the odds ratio when disease is rare (Prevalence ~< 0.1%)
  - The relative risk is not appropriate for common traits when a case-control design is used

11

Correspondence Between the Odds Ratio and Relative Risk

Dominant Model

Disease Prevalence	1/2* RR=1.5	2/2** RR=1.5
0.01	1.51	1.51
0.10	1.59	1.59
0.20	1.71	1.71

Multiplicative Model

Disease Prevalence	1/2 RR=1.5	2/2 RR=2.25
0.01	1.51	2.28
0.10	1.59	2.61
0.20	1.71	3.25

Marker minor allele and disease allele frequency 0.01
D' and $r^2=1$
*1/2 genotype – heterozygous (one copy of the alternative allele)
**2/2 genotype - homozygous for the alternative allele

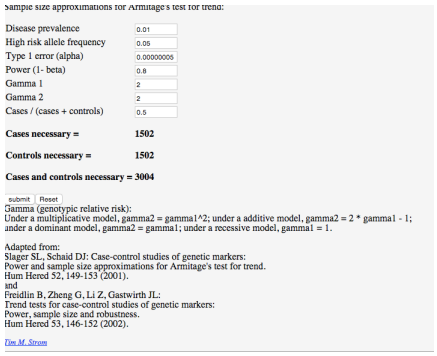
12

Armitage Trend Test - Power Calculations

- Information need
  - Population prevalence
  - Genetic Model
  - Risk allele frequency
- Tools
  - http://ihg.gsf.de/cgi-bin/hw/power2.pl
  - Reference Slager and Schaid 2001

13

Armitage Test for Trend



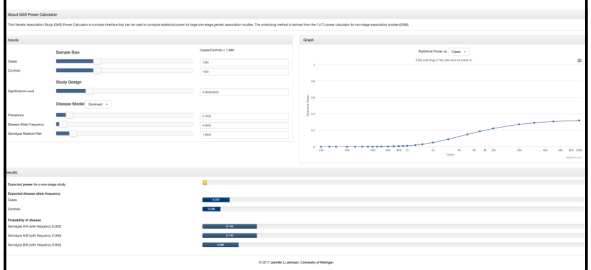
14

Genetic Association Study (GAS) Power Calculator

- http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html
- A one-stage study power calculator
  - Which was derived from CaTs
    - Which is to perform two-stage genome wide association studies
      - Skol et al. 2006
- Cochran Armitage Trend Test
- Displays graphs of the results

15

GAS Power Calculator



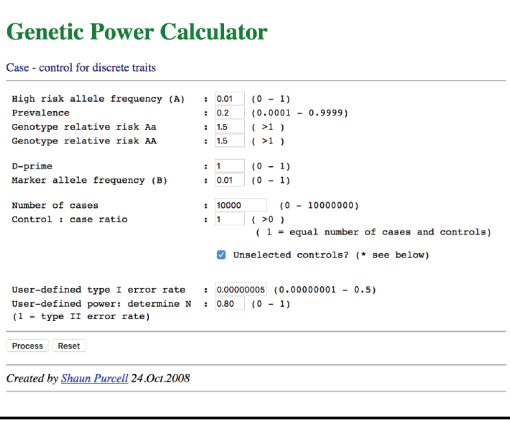
16

Genetic Power Calculator

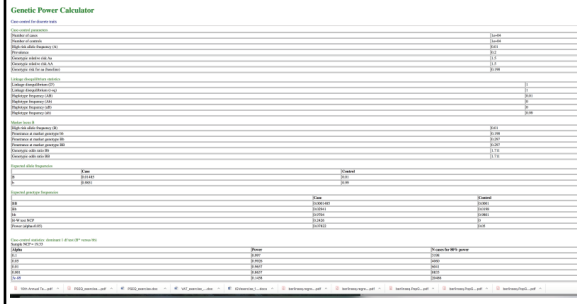
- http://zzz.bwh.harvard.edu/gpc/
- S Purcell & P Sham
- Uses the methods described in Sham PC et al. (2000) Am J Hum Genet 66:1616-1630
  - VC QTL linkage for sibships
  - VC QTL association for sibships
  - VC QTL linkage for sibships conditional on the trait
  - TDT for discrete traits
  - Case-Control for discrete traits
  - TDT for quantitative traits
  - Case-Control quantitative traits
- Although input is the relative risk
  - Displays odds ratios

17

Genetic Power Calculator



18



19

Power Association With Errors (PAWE)

- http://compgen.rutgers.edu/pawe/
- Implements the linear trend test
- Four different error models can be used
  - See online documentation for complete explanation
- Can either perform:
  - Power calculations for a fixed sample size
  - Sample size calculations for a fixed power
- The genotype frequencies can be generated either using a:
  - Genetic model free method or
  - Genetic model-based method

20

Quanto

- Provides sample size and power calculations for
- Genetic and environmental main effects
- Interactions
  - Gene x gene
  - Gene x environment
- Sample & power calculations can be carried for:
  - Case-control
    - Unmatched
    - Matched
  - Case-sibling
  - Case-parent (trios)
    - Quantitative
    - Qualitative
  - Independent sample of individuals
    - Quantitative traits
      - Assumption sampled from a random population
- Can only be run under windows
  - https://pphs.usc.edu/download-quanto/

21

Linkage Disequilibrium (LD)

- Power will be reduced if causal variant is not in perfect LD (r^2=1) with the tag SNP
- Can adjust sample size when r^2 <1 to increase power to the same level as when r^2=1
- Can estimate sample size when $r^2 \neq 1$
  - $N/r^2=N'$
  - Valid only for multiplicative model
  - (Pritchard and Przeworski, 2001)
- Power calculation almost always assume that r^2=1
- For whole genome sequence data this should be the case since usually the causal variant would be included in the data

22

Power Analysis for Rare Variant Aggregate Association Tests

- Many unknown parameters must be modeled
  - Allelic architecture within a genetic region
    - Varied across genes and populations
  - Effects of variants within a region
    - Fixed or varied effect sizes of causal variants
    - Bidirectional effect of variants
    - Proportion of non-causal variants
- Power estimated empirically
- Simplified assumptions can be made to obtain analytical estimates
  - All variants have the same effect size
  - No non-causal variants within a region that is analyzed in aggregate

23

Simplistic Analytical Power Calculation for Rare-variant Aggregate Association Analysis

- Assumption
  - All rare variants are causal and have the same effect size
- Although usual not be correct
  - Provides a gestalt of the power for a given samples or sample size for a given power
- Use aggregate of allele frequencies
  - For example, assume a cumulative allele frequency of 0.025
  - Use an exome-wide significant level e.g., 2.5×10^{-6}
- Provide disease prevalence and penetrance model
- Perform calculations in the same manner as was described for single variants

24

Empirical Power Calculations

- A variety of methods can be used to generate variant data to empirically estimate power
- Variant data is generated
  - Based upon a penetrance model samples of cases and controls are generated
  - Or a quantitative trait is generated based upon the genetic variance
- Multiple replicates are generated and analyzed
  - To determine the power

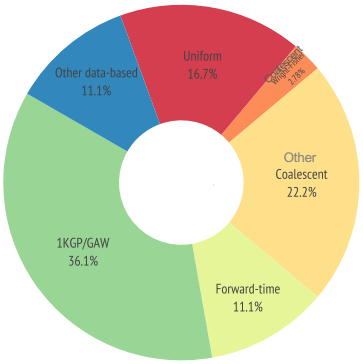
25

Empirical Power Calculations

- Examples
  - 5,000 replicates are generated each with 20,000 cases and 20,000 controls
    - The power is the proportion of replicates with p-value less than the specified threshold, e.g., $5x10^{-8}$
  - For rare-variant aggregate tests all autosomal genes are generated and those genes with more than two rare variants (e.g., predicted loss of function) are analyzed
    - The power is the proportion of genes that were tested with p-value which is below a specified threshold, e.g., $2.5x10^{-6}$

26

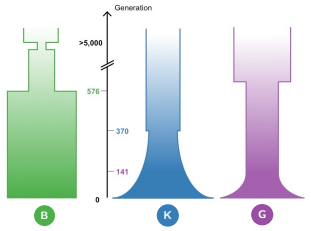
Simulation Methods



Note: Not all methods give a realistic distribution of variants & in particular for rare variants

27

Generating Exome Sequence Data Sets
Forward-time Simulation

Data	Haplotype Counts	Demographics
Boyko	105,814*	
Kyrukov	1,800,000*	
Gazave	1,308,000*	

*Selection coefficients used to define "variant type"
-"Missense" ($1.0 \times 10^{-8} - 1.8 \times 10^{-2}$)
-"Nonsense, splice site and frameshift" (>1.8×10^{-2})

28

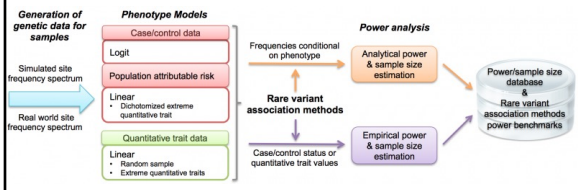
SKAT Power Calculator

- R Library
- Provides a haplotype matrix
  - 10,000 haplotypes over 200kb region
  - Simulated using a calibrated coalescent model (cosi)
  - Mimicking linkage disequilibrium structure of European ancestry
  - User can also provide haplotype data
- Power and sample size calculations for binary and quantitative traits
- User specify proportion of variants that increase or lower risk

29

SEQPower

http://www.bioinformatics.org/spower/

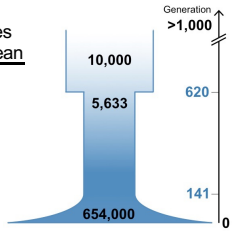


Wang et al. 2014 Bioinformatics

30

Generating Variants: Using a European Demographic Model and Exome Sequence Data

- Variant data generated on 18,397 genes
- Variant data simulated using a European population demographic model
  - Gazave et al. 2013



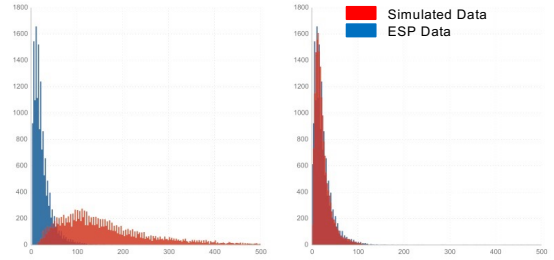
- Variants generated using exome sequence data
  - 4332 Exomes obtained from European American

Which method performs better and why?

31

Does Generating Variant Data Using the European Population Demographic Model Perform Well?

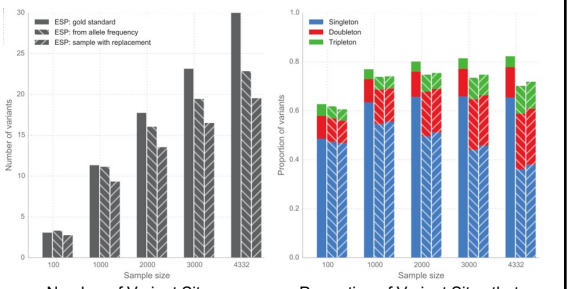
Distribution of number of variants per gene



- Simulated variant counts based on the entire simulated population
- Simulated variant counts based on haplotype pool down-sampled to ESP size

32

Simulating Data Using Sequence Data (ESP)

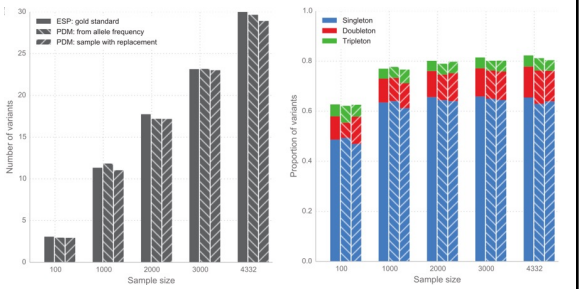


Number of Variant Sites

Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

33

Simulating Data: Using Population Demographic Models (PDM)



Number of Variant Sites

Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

34

Simulation Studies to Evaluate Power for Rare Variant Association Studies

- It is unknown which genes are important in disease etiology
  - Correct allelic architecture is unknown
- Can get a better understanding of power to detect associations by generating variants for the entire exome
- Use a variety of disease models
  - Odds ratios
  - Proportion of pathogenic variants
- Analyze of all genes
  - e.g., those with 2 or more variant sites
- Determine power as the proportion of genes that meet exome-wide significance (e.g., alpha=2.5×10^{-6})

35

Power Analysis

- For tests of individual variants
  - Power depended on sample size, disease prevalence, minor allele frequency, genetic model and variant effect size
- For rare variants (aggregate association tests)
  - Also dependent on the allelic architecture
    - Cumulative variant frequency within analyzed region
    - Proportion of causal variants
      - How much contamination from non-causal variants
    - Effect sizes the same the same or different across gene regions
      - Effects of variants in the same or different directions
        - » Protective and detrimental for binary traits
        - » Increase and decrease quantitative trait values

36

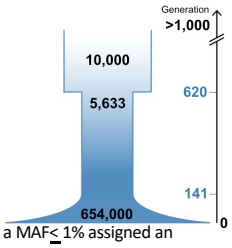
Power Analysis Rare Variants (Aggregate Association Tests)

- Power will not only vary between traits greatly
- The power to detect an association will also vary drastically between genes for the same complex trait
  - For some causal genes even with hundreds of thousands of samples power will be low
  - While for other causal genes a few thousand samples may be sufficient

37

How Large of a Sample Size is Necessary to Detect Rare Variant Associations?

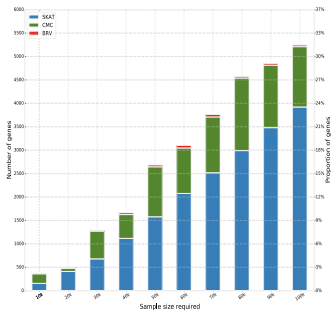
- Data generated on 18,397 genes
- Variant data simulated using a European population demographic model
  - Gazave et al. 2013



- Every missense, nonsense and splice with a MAF< 1% assigned an odds ratio of 1.5
- Sample sizes to detect X number of genes determined for
  - $\alpha = 2.5 \times 10^{-6}$
  - power=0.8

38

Sample Sizes Necessary to Detect an Association (Case-Control Data)



39

Genotype Pattern Mining For Digenic Traits

Advanced Gene Mapping Course, November 2022

Jurg Ott, Ph.D., Professor Emeritus
Rockefeller University, New York
https://lab.rockefeller.edu/ott/
ott@rockefeller.edu
PH +1 646 321 1013

Jurg Ott, PhD

Professor Emeritus and Director
Laboratory of Statistical Genetics
Rockefeller University
New York, NY 10065

https://lab.rockefeller.edu/ott/
EM: ott@rockefeller.edu
PH: +1 646 321 1013

Research Interests

Development of analysis methods for genetic data, genetic linkage and association analysis

Implementation in computer programs, dissemination on website

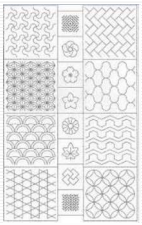
Collaboration with researchers world-wide on their data

Recent publications: [1-5] (#1 now freely available from https://www.jurgott.org/linkage/LinkageHandbook.pdf)

1. Terwilliger DJ, Ott J. Handbook of human genetic linkage. Johns Hopkins University Press. 1994.
2. Imai-Okazaki A, Li Y, Horpaopan S, Riazalhosseini Y, Garshasbi M, Mosse YP, et al. Heterozygosity mapping for human dominant trait variants. Hum Mutat. 2019 Apr 24;40(7):996-1004.
3. Horpaopan S, Fann CSJ, Lathrop M, Ott J. Shared genomic segment analysis with equivalence testing. Genet Epidemiol. 2020 Oct;44(7):741-47.
4. Okazaki A, Yamazaki S, Inoue I, Ott J. Population genetics: past, present, and future. Human genetics. 2020:1-10.
5. Okazaki A, Horpaopan S, Zhang Q, Randesi M, Ott J. Genotype pattern mining for pairs of interacting variants underlying digenic traits. Genes. 2021;12(8):1160.

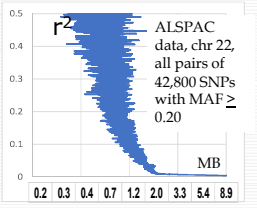
Topics

- Science develops independently in different fields:
  - Human gene mapping
  - Frequent Pattern Mining
- Case-control association analysis
  - Main effects in genetic association studies
  - Interaction effects in case-control data
- Mining consumer databases
  - The *Apriori* algorithm
  - Newer algorithms: *eclat, fpgrowth*
  - Analysis of AMD dataset



Main association effects

- Consider two DNA variants with minor alleles *A* and *T*. Even when on the same chromosome, the frequency of *A-T* chromosomes (haplotypes) is the product of allele frequencies, P(*A*-*T*) = P(*A*) P(*T*) → linkage equilibrium. Variants very close together: P(*A*-*T*) ≠ P(*A*) P(*T*) → LD, linkage disequilibrium.
- Disease variant vs marker variant: Different genotype frequencies in cases and controls → genetic association.
- **Recessive traits**: Variants close to disease tend to be homozygous (homozygosity mapping; Lander & Botstein, *Science* 1987;**236**:1567-70).
- **Dominant traits**: Variants close to disease tend to be heterozygous (Imai-Okazaki et al, *Hum Mutat* 2019;**40**:996-1004): P(het) > 1 – *f*, P(het, popul.) = 2*f*(1 – *f*), *f* = MAF



	AA	AG	GG
affected	0.18	0.70	0.12
unaffected	0.04	0.32	0.64
OR	~5	~5	~0.1

Multiple Hits ... Digenic Diseases

Ming & Muenke (2002) *Am J Hum Genet* **71**, 1017 (review)
Schaffer A (2013) *J Med Genet* 50, 641-52 (review)

Effect and Phenotype	Gene 1		Gene 2	
	Mutation	Phenotype	Mutation	Phenotype
Synergistic:				
RP	*ROM1*$^{+/G80insG}$	Normal	*RDS*$^{+/L185P}$	Normal
RP	*ROM1*$^{+/L114insG}$	Normal	*RDS*$^{+/L185P}$	Normal
Bardet-Biedl	*BBS2*$^{Y240X/Q359X}$	Normal	*BBS6*$^{+/Q147X}$	Normal
Deafness	*GJB2*$^{+/35delG}$	Normal	*GJB6*$^{+/-}$	Normal
Deafness	*GJB2*$^{+/167delT}$	Normal	*GJB6*$^{+/-}$	Normal
Hirschsprung	*RET*$^{+/I647T}$	Normal	*EDNRB*$^{+/S305N}$	Normal
Severe insulin resistance	*PPARG*$^{+/A353delAAAT}$	Normal	*PPP1R3A*$^{+/C1984delAG}$	Normal
Modifier:				
Juvenile-onset glaucoma	*MYOC*$^{+/G399V}$	Adult-onset glaucoma	*CYP1B1*$^{+/R368H}$	Normal
Usher 1	*USH3*$^{mut/mut}$	Usher 3	*MYO7A*$^{+/delG\ (exon\ 25)}$	Normal
Congenital nonlethal JEB	*COL17A1*$^{R1226X/L855X}$	Juvenile JEB	*LAMB3*$^{+/R635X}$	Normal
More severe ADPKD	*PKD1*$^{+/mut}$	Less severe ADPKD	*PKD2*$^{+/2152delA}$	Less severe ADPKD
More severe hearing loss	*DFNA1*	Mild hearing loss	*DFNA2*	Mild hearing loss
WS2/OA	*MITF*$^{+/I944delA}$	?WS2	*TYR*$^{+/R402Q}$	Normal
More severe WS2/OA	*MITF*$^{+/I944delA}$	?WS2	*TYR*$^{R402Q/R402Q}$	Normal

How to analyze interaction effects?

- Hyperlipidemia data: 5 relevant genes, ~200 variants in each gene, look for interactions in each pair of variants. Work with LR chi-square!

CASES	Variant 1				CONTROLS	Variant 1		
Var 2	GG	GT	TT		Var 2	GG	GT	TT
AA	...	...	...		AA	...	...	...
AC	...	...	...		AC	...	...	...
CC	...	...	...		CC	...	...	...

Data	chi-sq	df
cases	3.3591	4
controls	3.6658	4
both	1.4255	4
heterogeneity	5.5994	4

$\chi^2_{Heterogeneity} = \chi^2_{Cases} + \chi^2_{Controls} - \chi^2_{both}$

Var 1 ->	GG			GT			TT		
Var 2 ->	AA	AC	CC	AA	AC	CC	AA	AC	CC
cases	...	...	...	...	...	...	...	...	...
controls	...	...	...	...	...	...	...	...	...

Source	chi-sq	df
Var 1 main	0.4196	2
Var 2 main	48.1979	2
Interaction	5.5994	4
Total table	54.2169	8

$\chi^2_{Interaction} = \chi^2_{Total} - \chi^2_{Var1} - \chi^2_{Var2}$

- More sophisticated analysis by logistic regression (Cordell, *Nat Rev Genet* 2009;**10**:392-404).

Different Levels of Genetic Interactions

Okazaki & Ott (2022) *Trends in Genetics* 38 (10):1013-1018

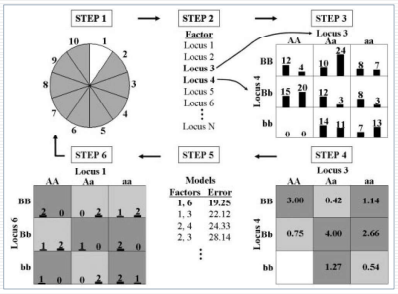
1. Traditionally, disease association has been carried out on the level of alleles or **genotypes**. The total number of pairs can be prohibitively large. While this level of analysis generally requires the most effort, it also entails the highest level of precision in the sense that disease-causing elements can be directly traced down to nucleotides.
2. Working with pairs of **variants** provides some economy of computational effort but may 'dilute' a signal from a single genotype pair when all nine genotype pairs in a pair of variants are analyzed jointly.
3. Finally, focusing on pairs of **genes** represents the most economical approach but is also the most imprecise among the three strategies. Also, focusing on genes disregards susceptibility elements outside of genes. Distant-acting transcriptional enhancers have been known for over 10 years to affect susceptibility to human disease and noncoding RNAs have been shown to be associated with many diseases, for example, cardiac hypertrophy.

Finding disease-associated pairs of genotypes

1. Multifactor Dimensionality Reduction (MDR)
   Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions ... Genet Epidemiol 2003;24:150–157
2. Exhaustive evaluation of all pairs of genotypes at all pairs of variants
3. Applying off-the-shelf pattern search algorithms
   Chee C-H, Jaafar J, Aziz IA, Hasan MH, Yeoh W. Algorithms for frequent itemset mining: a literature review. Artificial Intelligence Review. 2019;52(4):2603-21
4. Construction of Bayesian network
   Guo Y, Zhong Z, et al. Epi-GTBN: An approach of epistasis mining based on genetic Tabu algorithm and Bayesian network. BMC Bioinform 2019;20:444
5. Sophisticated computational approaches
   Titarenko SS, Titarenko VN, Aivaliotis G, Palczewski J. Fast implementation of pattern mining algorithms ... . Journal of Big Data. 2019; 13(6):37

1. MDR

- Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions ... Genet Epidemiol 2003;24:150–157
- Classify each of the 9 cells as high risk or low risk.
- Evaluate prediction error (case vs. control) by cross-validation.
- Find model that maximizes cross-validation consistency and minimizes prediction error.



2. Exhaustive search for interacting SNPs

- "Discovering Genetic Factors for psoriasis through exhaustively searching for significant second order SNP-SNP interactions"
- Kwan-Yeung Lee, Kwong-Sak Leung, Nelson L. S. Tang & Man-Hon Wong. *Sci Rep* 2018;**8**:15186
- Abstract: To deal with the enormous search space, our search algorithm is accelerated with eight **biological plausible interaction** patterns and a pre-computed look-up table. After our search, we have discovered several **SNPs having a stronger association to psoriasis when they are in combination with another SNP**...

All pairs of SNPs

Ueki & Cordell (2012) *PLoS Genet* 8(4): e1002625

- Interaction statistic, χ^2 (1 df). Implemented in *plink* with option `--fast-epistatisis joint-effects`
- Applied to schizophrenia data: 2,164 males, 853,934 SNPs
- Trend genotype test (*plink*), permutation testing with 10,000 replicates: 5 SNPs with $p < 0.05$ by permutation and Bonferroni.
- Interaction tests for all pairs of SNPs, disregarding the 5 SNPs significant in trend test.
- Much stronger results even though p_{Bon} obtained as 853,929 × p.
- General result?

p_{Bon} <	#SNPs	same chr.
0.01	259	41
0.05	452	63

3. Frequent Pattern Mining

- Thirty years ago, supermarkets started collecting huge amounts of consumer data at their cashiers. Consumer habits – if someone buys bread, how likely will they also buy milk and wine?
- **Apriori algorithm** (Agrawal et al, *ACM SIGMOD Conference on Management of Data* 1993; 207-216): Efficient search for frequent sets of items ("itemsets") purchased by one consumer ("transaction"). Development of **association rules**, that is, conditional probabilities P(Y | X), with Y and X being items or itemsets.
- Research published in conference proceedings, rarely in traditional journals.
- In the absence of strong main effects, we need to directly search for **genotype patterns** (at two or more variants) with different frequencies in cases and controls, without consulting main effects.
- Zhang Q, Long Q, Ott J. **AprioriGWAS**, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. PLoS Comput Biol. 2014 Jun;10(6):e1003627
- Other implementations of search algorithms, e.g. *fpgrowth* (https://borgelt.net/software.html). Huge memory demands: Using Linux desktop with 512 GB of memory.

4. Bayesian networks

- Guo Y et al. Epi-GTBN: An approach of epistasis mining based on ... Bayesian network. BMC Bioinform. 2019;20:444
- Like many other approaches, Epi-GTBN employs a Bayesian network, that is, a probabilistic model to represent actions and interactions among variants and phenotypes.
- Authors analyzed a well-known dataset on age-related macular degeneration (AMD), which has been investigated by various other researchers. For analysis by Epi-GTBN, to reduce the computational burden, only the 1,039 SNPs with **smallest p-values** ($p < 0.01$) out of the original 103,611 SNPs were retained.
- Results were comparable to those obtained elsewhere.
- Focusing on variants with strong main effects is fallacious! Frequencies of genotype patterns depend on main and interaction effects: Strong main effects are likely to lead to strong (significant) genotype patterns.

5. Newer Algorithms

1) http://www.philippe-fournier-viger.com/spmf/
2) Titarenko SS et al. Fast implementation of pattern mining algorithms ... . Journal of Big Data. 2019; 13(6):37

- 1) Superb documentation, freely downloadable. Large memory requirements although not as large as for *fpgrowth* (Borgelt).
- 2) Apriori principle vs. evaluating all genotype pairs.
- Schizophrenia data: 853,934 SNPs vs 344,831,940,768 pairs (diff. chrom)
- For a given pattern (pair) of genotypes, X, its relation with phenotype Y is specified by a 2 x 2 table.
- Support, $s = a + c$
- Confidence, $c = a/(a + c)$

Phenotype	X present	X absent
Y = 2, cases	*a*	*b*
Y = 1, controls	*c*	*d*

- Current computer approaches: Work with bitwise operations (1 word = 4 bytes = 32 bits) and with multiple threads in a single machine.

AMD data: Genotype pattern analysis

Klein et al (2005) Science 308 (5720:385-389

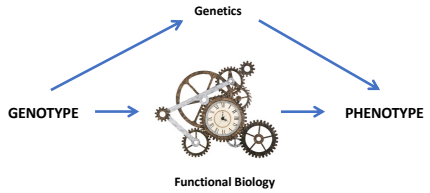
- Search for patterns (genotype pairs) with minimum support of 40. Perform 1000 random permutations for *p*-value estimation (corrected for multiple testing).
- Find m = 18,044,794 patterns.

supp	conf	chisq	pPerm	OR	ch1	ch2
40	90.0	47.2604	0.015	18.5	3	4
40	90.0	47.2604	0.015	18.5	3	4
53	81.1	42.5124	0.098	9.5	1	5
56	78.6	39.481	0.315	8.1	6	7
56	78.6	39.481	0.315	8.1	6	7

- Two patterns are significant, $p = 0.015$, compared with the best $p = 0.60$ in single-variant analysis by trend test.
- Expect many **more significant genotype patterns than single-variant results**.
- Different ways of establishing significance: Bonferroni correction; FDR depends on large number m of "null" results.
- Compare confidence of 90% with 43% of cases ("null" confidence) in data.

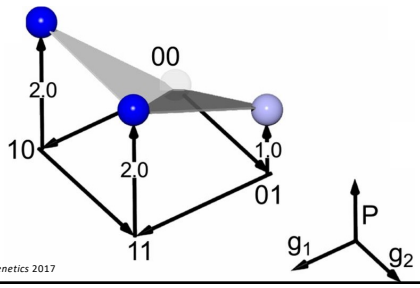
alpha	Perm	Bonf	FDR-BY
0.001	0	2	2
0.01	0	9	3
0.02	2	11	11
0.03	2	16	13
0.04	2	18	61
0.05	2	19	11905

Why are we doing genetics?



1

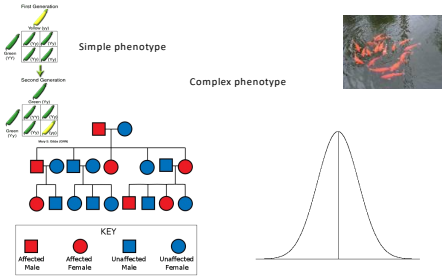
Genotype-Phenotype Map



Sailer & Harms, *Genetics* 2017

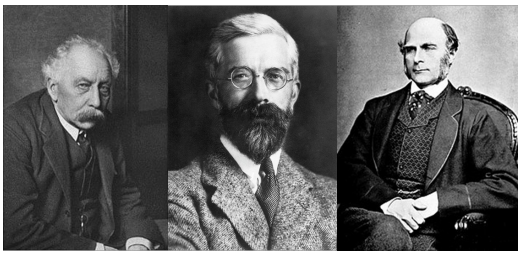
2

Simple and complex phenotypes



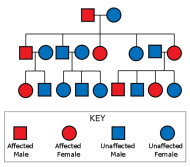
3

The Battle between Mendelians and Biometricians

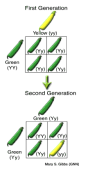


4

Mendel phenotypes segregate in pedigrees



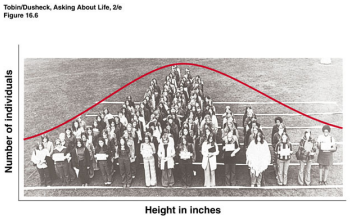
Dominant



Recessive

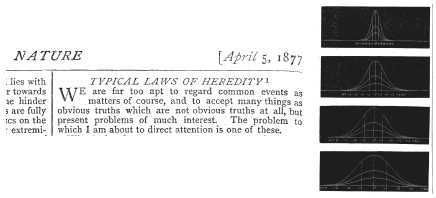
5

Statistical properties of quantitative traits



6

Complex traits are heritable but not in Mendelian fashion



NATURE [April 5, 1877

lies with
r towards
ne hinder
s are fully
cs on the
extremi-

TYPICAL LAWS OF HEREDITY [1]

WE are far too apt to regard common events as matters of course, and to accept many things as obvious truths which are not obvious truths at all, but present problems of much interest. The problem to which I am about to direct attention is one of these.

7

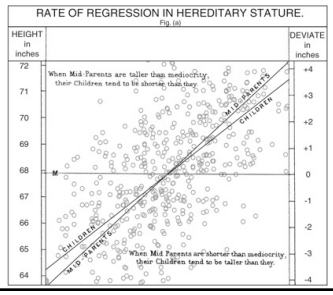
7

Observations

- Progeny of smaller plants are, on average, smaller
- Progeny of larger plants are, on average, larger
- ALL HAVE THE SAME VARIANCE!

8

Complex traits are heritable but not in Mendelian fashion

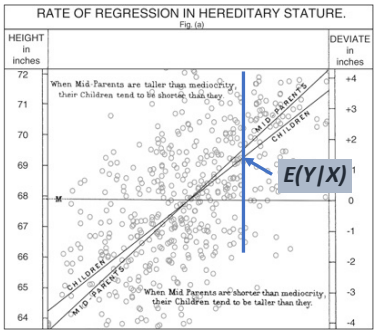


9

Introduce properties of the plot

- Children of tall parents are taller than average, and children of short parents are shorter than average.
- Children of tall parents are not as tall as their parents and children are not as short as their parents (regression to mediocrity).

10

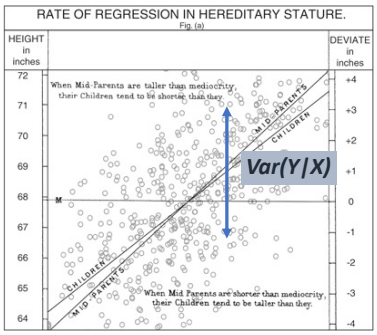


11

Properties of the Galton plot

- ***E(Y|X)*** – conditional expectation (regression)
- Regression is linear: $E(Y|X) = b_1X$
- $E(X|Y) = b_2Y$ (both regressions are linear)

12



13

Properties of the Galton plot

- ***Var(Y|X)*** is independent of ***X*** !

14

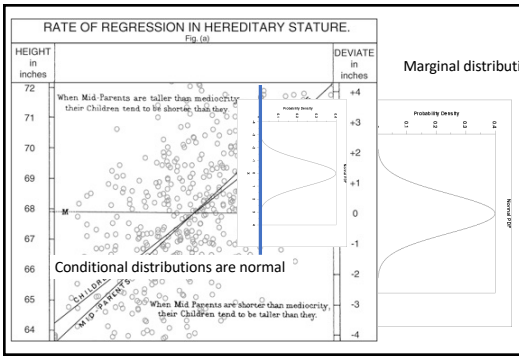
This distribution is bi-variate normal !

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

$\boldsymbol{\mu}$ is mean

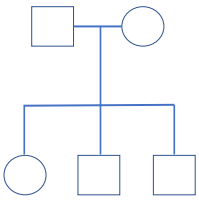
$\boldsymbol{\Sigma}$ is covariance matrix

15



16

Infinitesimal model: multivariate normal distribution in pedigrees



$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

The pedigree defines the covariance matrix !

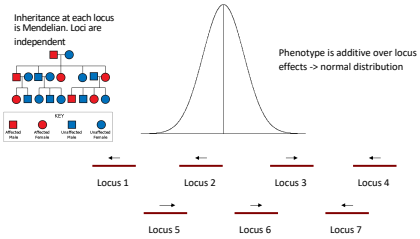
17

XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

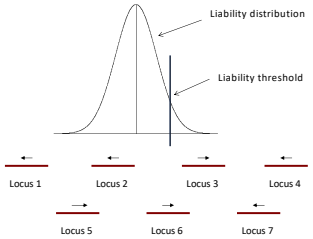
18

Quantitative Trait Loci (QTLs)



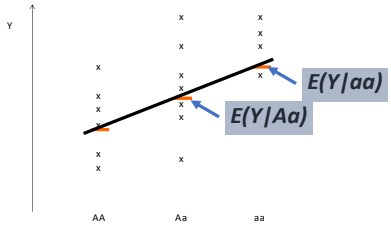
19

Binary (dichotomous) traits



20

Allelic effects – change in phenotypic mean per one allele ***a***



21

The model more formally

X – genotype represented as dosage of allele ***a***

	AA	***Aa***	***aa***
X	0	1	2

X_{ij} – genotype of individual ***i*** at locus ***j***

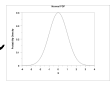
Y_i –phenotype of individual ***i*** (we will assume ***E(Y)=0***)

β_j –allelic effect at locus ***j***

22

The model more formally

$$Y_i = \underbrace{\sum_j \beta_j X_{ij}}_{\text{Genetic factors}} + \underbrace{\varepsilon}_{\text{Non-genetic factors}} \qquad \varepsilon \sim$$



23

The model more formally

$$\bar{y} = X\bar{\beta} + \varepsilon$$

24

This model explains

- Normal distribution in the population (why?)
- Bi-variate normal distribution on the Galton plot (why?)
- Regression to mediocrity (why?)

25

Questions about allelic architecture

- How much of the phenotypic variation is due to genes?
- How many loci are involved?
- Are alleles involved in the phenotype common or rare in the population?
- What about dominance?
- Do alleles interact?
- Do alleles have different effects in different environment?
- And BTW, what is biology behind the story?

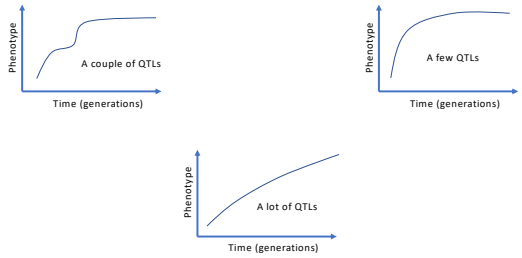
26

We start from the effect of selection on quantitative traits

- Historically, observations on selection gave first clue about polygenicity of quantitative traits
- We will look at the same questions from a different perspective a little later

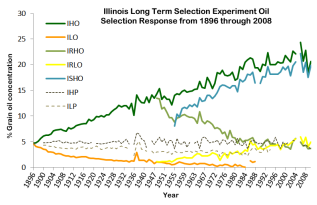
27

Selection experiment



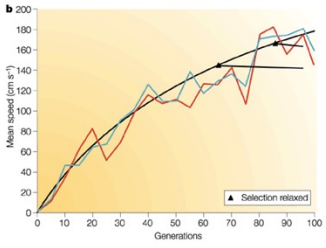
28

Evidence in favor of the highly polygenic model



29

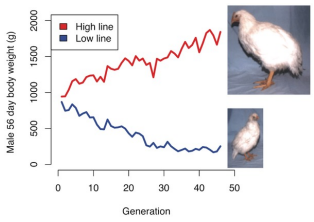
Evidence in favor of the highly polygenic model



Weber, *Genetics* 1996; Barton & Keightley, *Nature Reviews Genetics* 2002

30

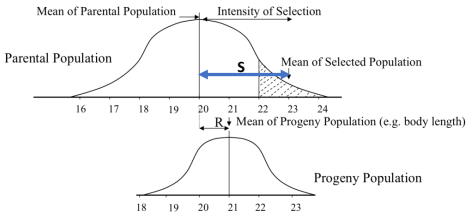
Evidence in favor of the highly polygenic model



Johansson et al., *PLOS Genetics* 2010

31

Breeder's equation



32

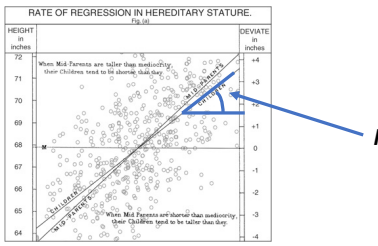
Breeder's equation

$$R = h^2 S$$

- Response to selection is proportional to selection differential
- Coefficient of proportionality is called "heritability".
- Long-term linear response to selection means that heritability stays constant for a long time.

33

Similarly, heritability is the slope on the Galton's plot

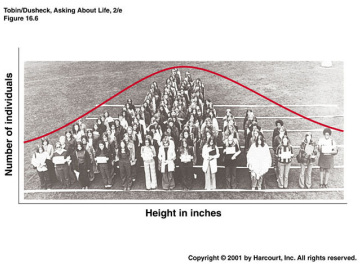


h^2

$$h^2 = \frac{Cov(x,y)}{Var(x)}$$

34

Everything is about variance



35

Law of total variance

Normal distribution has only two parameters.

If we are interested in the population variation, we are only interested in variance.

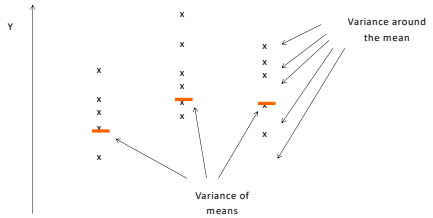
How can we explain reasons for variation in phenotypes?

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$$

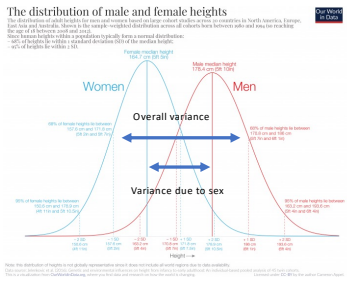
$$Fraction\ of\ variance\ due\ to\ factor\ X: \frac{Var(Y|X)}{Var(Y)}$$

36

Variance decomposition

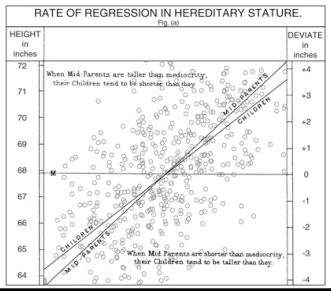


37



38

Variance due to parental height



39

Population variation is fully described by variance

$$V = V_G + V_E$$

Genetic contribution

Everything else

40

Components of genetic variance

$$V_G = V_A + V_D + V_I + V_M$$

Main (additive) effects

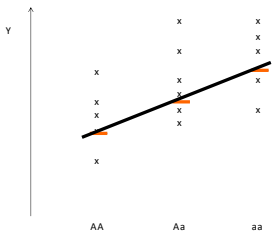
Dominant effects

Genetic interactions

New mutations

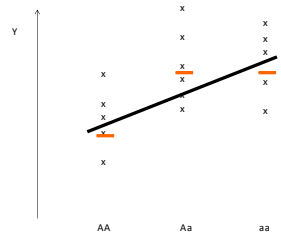
41

Regression



42

Regression



43

Additive model

	AA	*Aa*	*aa*
E(Y)	0	β	2β

44

Additive variance

Additive variance V_A is genetic variance explained by the additive model

$$Y_i = \sum \beta_j X_{ij} + \varepsilon$$

$$Var(Y) = Var\left(\sum \beta_j X_{ij}\right) + Var(\varepsilon)$$

$$Var(Y) = \sum Var(\beta_j X_{ij}) + \overline{\varepsilon^2}$$

$$Var(Y) = \sum \beta_j^2 Var(X_{ij}) + \overline{\varepsilon^2}$$

$$Var(Y) = \sum 2\beta_j^2 p_j q_j + \overline{\varepsilon^2}$$

45

Additive variance

$$Var(Y) = \sum 2\beta_j^2 p_j q_j + \overline{\varepsilon^2}$$

Additive genetic variance (variance due to independent effects of alleles)

"Environmental" variance

$$V_A = \sum 2\beta_j^2 p_j q_j$$

46

Heritability

Broad sense

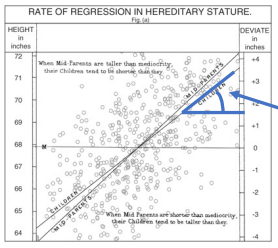
$$H^2 = \frac{V_G}{V}$$

Narrow sense

$$h^2 = \frac{V_A}{V}$$

47

Now, we can bring together epidemiological (or experimental) observations and molecular data



h²

$$h^2 = \frac{\sum 2\beta_j^2 p_j q_j}{V}$$

48

Just to recap

$$V_G = V_A + V_D + V_I + V_M$$

Main (additive) effects

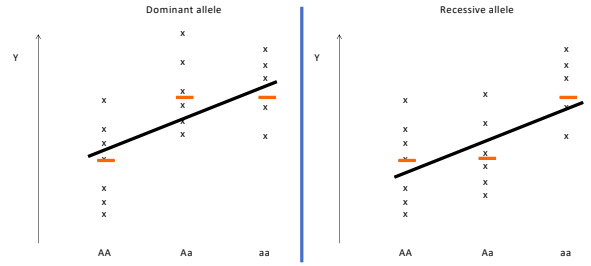
Genetic interactions

Dominant effects

New mutations

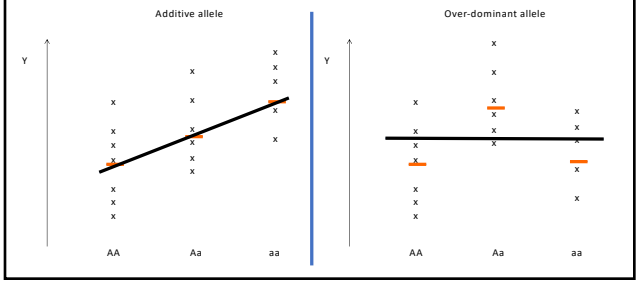
49

Dominance and dominance variance



50

Dominance and dominance variance



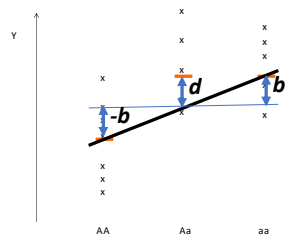
51

Dominance variance

- Dominance variance – variance unexplained by the additive model but explained by introducing dominance effects
- V_D is zero for the additive model
- V_D is usually large for the over-dominant model
- V_D is surprisingly smaller than V_A for the dominant (or recessive) model

Note that human medical genetics frequently uses a different definition of dominance

52



53

Variance component due to epistasis (genetic interactions)

Epistatic variance V_I is genetic variance that is not captured by the additive model but explained by interactions between alleles

Additive by additive pairwise epistasis

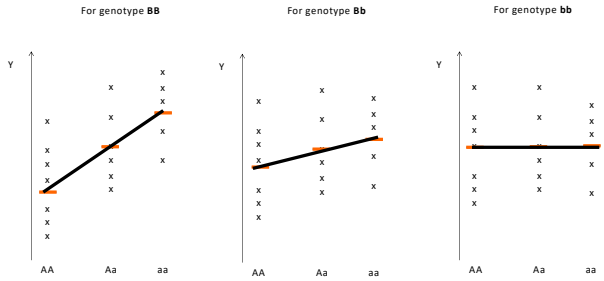
$$Y_i = \sum_j \beta_j X_{ij} + \sum_{ik} \beta_{ik} X_{il} X_{ik} + \varepsilon$$

54

Epistasis

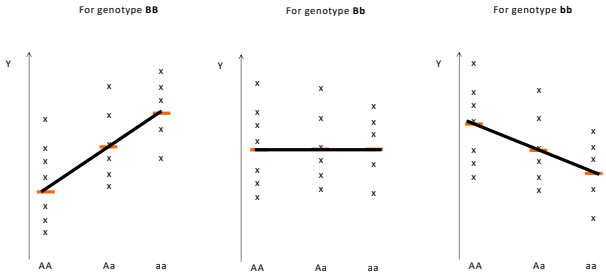
- Dependency of allelic effects on genetic background
- You can think of it as of interaction of allelic effects (usually denoted ***GxG***)
- For example, consider two loci with alleles ***A*** / ***a*** and ***B***/ ***b***
- Epistasis means that the effect of allele a is not the same for individuals with genotypes ***BB***, ***Bb*** and ***bb***
- In other words, effects of alleles ***a*** and ***b*** are not independent

55



56

Sign epistasis



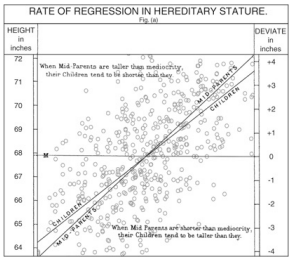
57

Epistatic variance

- For two loci, V_I is usually small even in the presence of epistatic interactions (physiological epistasis).
- Frequencies of all four haplotypes have to be substantial for V_I to play a role
- V_I can be large in case of sign epistasis. Sign epistasis is not biologically infeasible
- For multiple alleles, number of interacting terms is very large, so theoretically total V_I can be large even if individual pairwise terms are small

58

Estimating heritability



$$Cov(MP,O) = \frac{1}{2}V_A + \frac{1}{4}V_I$$

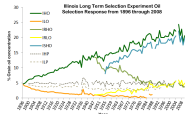
Children inherit half of parental alleles but only quarter of interactions!

59

More generally, for relatives ***A*** and ***B*** of degree ***r***

$$Cov(Y_A, Y_B) = \frac{1}{r}V_A + \frac{1}{r^2}V_I$$

Assuming no dominance



The long-term selection response suggests that the role of V_I is limited

A recent study of social networks says that humans are not the exception

60

Other variance components

Epistasis can be additive by dominant and dominant by dominant

Epistasis can be due to higher order interactions

Mutational variance $\boldsymbol{V_M}$ – additional variance due to *de novo* mutations

61

The case of the missing heritability

For GWAS significant loci

$$\frac{Cov(MP,O)}{V(MP)} \gg \frac{\sum 2\beta_j^2 p_j q_j}{V}$$

62

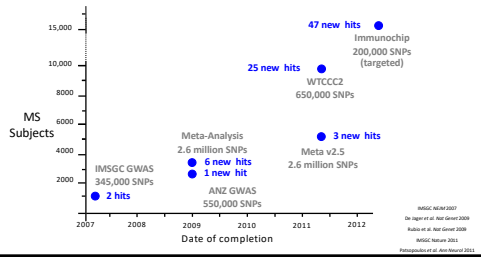
Likely reasons for missing heritability

1. *Common variants of weak effect*
2. *Rare variants of larger effect*
3. *Epistatic interactions*

$$Cov(MP,O) = \frac{1}{2}V_A + \frac{1}{4}V_I$$

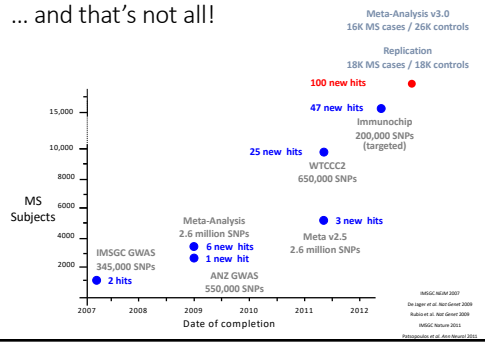
63

Multiple sclerosis GWAS



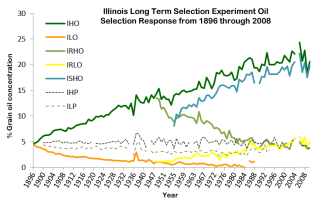
64

… and that's not all!



65

Evidence in favor of the highly polygenic model



66

Evidence in favor of the highly polygenic model

67

	AA	Aa	aa
Genotypes	0	1	2
Normalized genotypes	$\frac{0-E(X)}{\sqrt{Var(X)}}$	$\frac{1-E(X)}{\sqrt{Var(X)}}$	$\frac{2-E(X)}{\sqrt{Var(X)}}$
Normalized genotypes	$\frac{-2q}{\sqrt{2pq}}$	$\frac{p-q}{\sqrt{2pq}}$	$\frac{2p}{\sqrt{2pq}}$

In these notations, $V_A = \sum \beta^2$

68

$$Y_i = \sum_j \beta_j X_{ij} + \varepsilon$$

X_{ij} – Normalized genotype of individual i at SNP j

In the matrix form:

$$\bar{y} = X\bar{\beta} + \varepsilon$$

$$GRM = \frac{1}{N} XX^T$$

Now, assume that allelic effects are random

$$\beta \sim N(0, \sigma^2) \qquad V_A = M \cdot \sigma^2$$

69

If we assume that genetic effects are random

We assume that all SNPs have effects on the trait drawn from a normal distribution

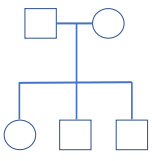
$$Y_i = \mu_i + u_i + \varepsilon$$

$$Cov(u_i, u_k) = \frac{1}{N}\sigma^2 \sum_j X_{ij} X_{ik}$$

$$u \sim MVN\left(0, \sigma^2 GRM\right)$$

70

Another way to think about it: extending the infinitesimal model



$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

The pedigree defines the covariance matrix !

71

For the whole population

- Assume that the distribution of the trait is multivariate normal.
- Covariance matrix is given by the observed covariance of genotypes (rather than pedigree, IBS instead or IBD)
- This covariance matrix is scaled by additive genetic variance

72

This model explains a large fraction (although not all) of heritability

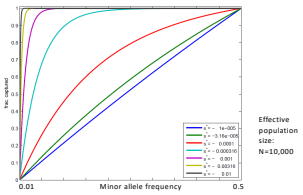
73

Challenges of the polygenic model

- The model literally assumes that all of the genome is important for the trait
- What can be the biological effect of so many loci?
- Natural selection is expected to rapidly eliminate variants and reduce allele frequency of remaining variants
- Variants must be either very rare or of very small effect sizes

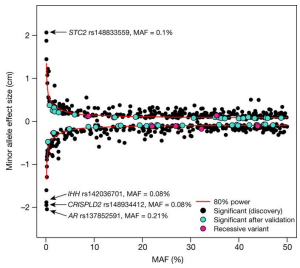
74

Heritability by allele frequency



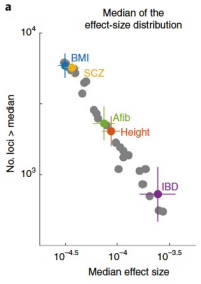
75

Rare coding alleles have larger effect sizes



76

Effect size distributions



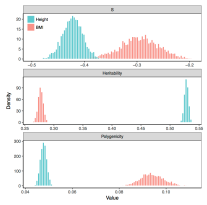
77

Evidence in favor of the highly polygenic model

$$\beta_j \sim N\left(0, [2p_j(1-p_j)]^S \sigma_\beta^2\right)\pi + \phi(1-\pi)$$

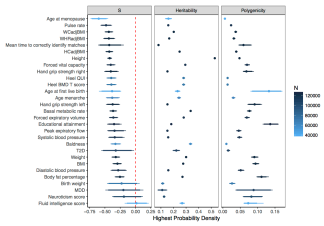
78

Evidence in favor of the highly polygenic model



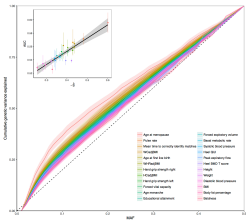
79

Evidence in favor of the highly polygenic model



80

Evidence in favor of the highly polygenic model



81

GxG interactions

82

Variance component due to epistasis (genetic interactions)

Epistatic variance V_I is genetic variance that is not captured by the additive model but explained by interactions between alleles

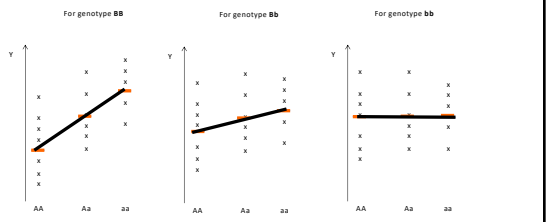
$$Y_i = \sum_j \beta_j X_{ij} + \sum_{ik} \beta_{ik} X_{il} X_{ik} + \varepsilon$$

83

Epistasis

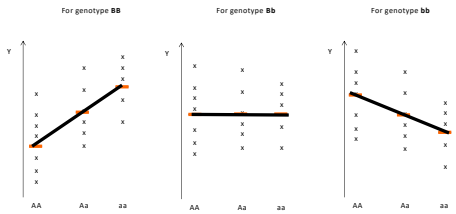
- Dependency of allelic effects on genetic background
- You can think of it as of interaction of allelic effects (usually denoted ***GxG***)
- For example, consider two loci with alleles ***A*** / ***a*** and ***B***/ ***b***
- Epistasis means that the effect of allele a is not the same for individuals with genotypes ***BB***, ***Bb*** and ***bb***
- In other words, effects of alleles ***a*** and ***b*** are not independent

84



85

Sign epistasis



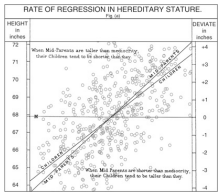
86

Epistatic variance

- For two loci, V_I is usually small even in the presence of epistatic interactions (physiological epistasis).
- Frequencies of all four haplotypes have to be substantial for V_I to play a role
- V_I can be large in case of sign epistasis. Sign epistasis is not biologically infeasible
- For multiple alleles, number of interacting terms is very large, so theoretically total V_I can be large even if individual pairwise terms are small

87

Estimating heritability



$$Cov(MP,O) = \frac{1}{2}V_A + \frac{1}{4}V_I$$

Children inherit half of parental alleles but only quarter of interactions!

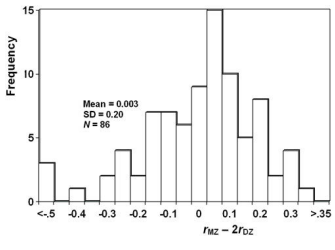
88

Why is epistatic variance commonly disregarded?

- In human genetics, epistatic interactions between common variants have not been observed.
- In a model with two (or several) loci, contribution of epistatic variance is relatively small.
- Long term response to selection in model organisms seems to contradict the importance of epistasis.

89

Evidence against epistasis



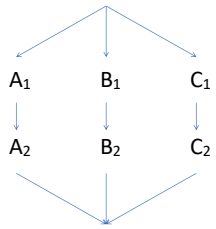
90

Why might epistatic variance be of importance?

- A non-linear model involving many loci would generate a large epistatic variance.
- Interactions would be statistically undetectable.
- The model would not generate significant deviations from the observations.
- As an example, we may consider a model with multiple pathways involved.

91

Multiple pathway model



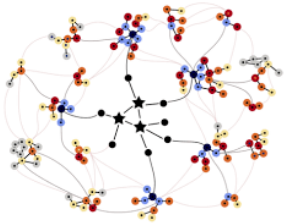
92

Genotype-phenotype map

- The map seems to be highly dimensional
- It is surprisingly linear, as far as we know at the time

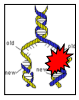
93

What is biology behind this?



94

Forces responsible for genetic change

Mutation  μ

Selection  s

Drift  N_e

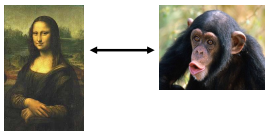
Population structure  F_{ST}

1

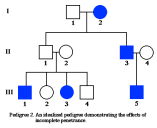
Mutations

2

Mutation rate in humans and flies



2.5x10^{-8} (Nachman & Crowell)



1.8x10^{-8} (Kondrashov)

NGS estimates *~1.2X10^{-8}* per nt changes genome

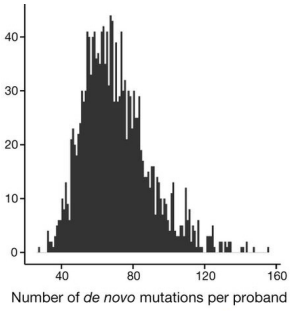
~70 per nt changes genome

Other events: indels (10^{-9})

repeat extensions/contractions (10^{-5})

3

Number of de novo mutations per individual

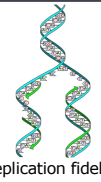


Number of *de novo* mutations per proband

Jonsson et al., *Nature* 2017

4

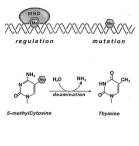
Mutation rate is variable along the genome



Replication fidelity

DNA damage

DNA repair



CpG deamination

Regional variation of mutation rate

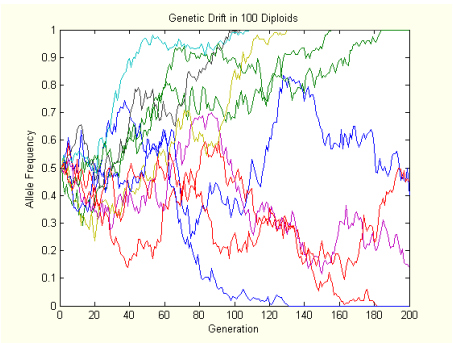
Context dependence of mutation rate

5

Genetic drift

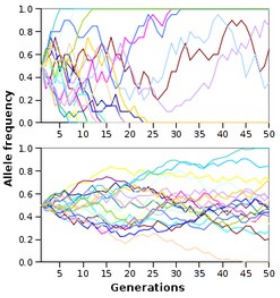
6

Drift is a random change of allele frequencies



7

Drift depends on population size



8

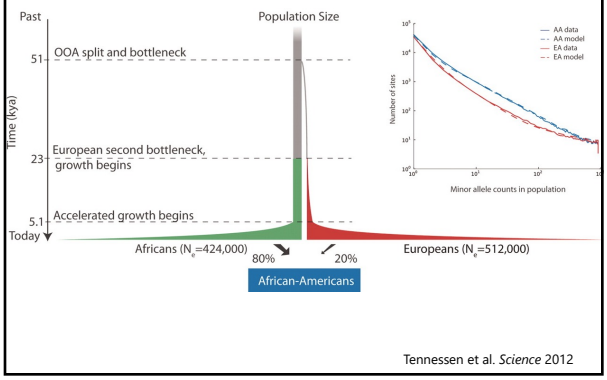
Effective population size

- In an idealized model, the intensity of genetic drift depends on population size (mean squared change in allele frequency is proportional to 1/Ne)
- In more realistic situations, effective population size (Ne) is a parameter characterizing intensity of drift

9

Demographic history

10

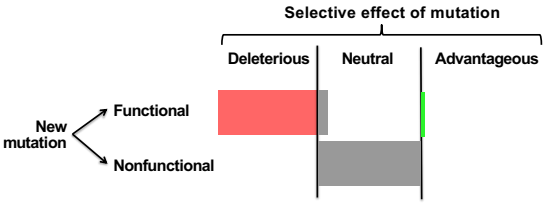


11

Selection

12

Most functional mutations are deleterious



Selection indicates functional mutations, whether or not the tested trait is under selection

13

Selection coefficient

- Selection coefficient (***s***) is the expected relative loss of fitness due to the sequence variant
- Variants with selection coefficients less than ~1/Ne are insensitive to selection. This is the drift barrier

14

Basic facts about human genetic variation

- Nucleotide diversity (density of nucleotide differences between two randomly chosen chromosomes) is about 0.001
- Most common SNPs are very old (~300-400K years old)
- Protein coding regions are showing clear signs of selection (reduced diversity and excess of rare alleles)

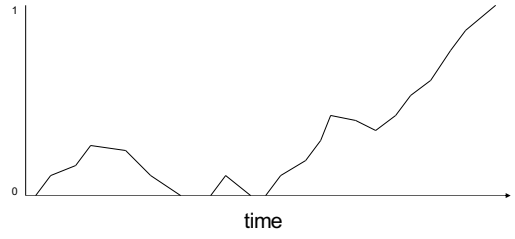
15

Methods of mathematical population genetics

16

Dynamic of allelic substitution

Mathematically, allele frequency change in a population follows a one-dimensional random walk



17

Diffusion approximation

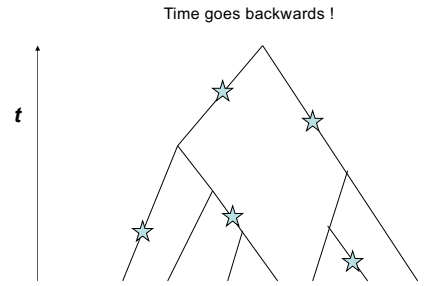
Random walk that does not jump long distances can be approximated by a diffusion process

$$\frac{\partial \phi(x,p,t)}{\partial t} = -\frac{\partial M\phi(x,p,t)}{\partial x} + \frac{1}{2}\frac{\partial^2 V\phi(x,p,t)}{\partial x^2}$$

18

Coalescent theory

Instead of modeling a population, we can model our sample



19

Natural selection in protein coding regions

20

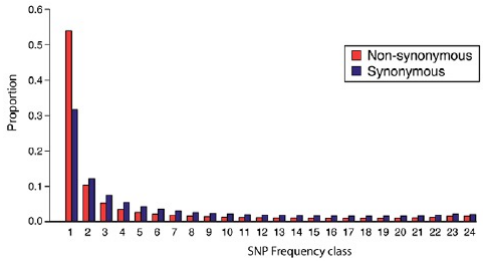
Signatures of purifying selection

Reduced variation

Excess of rare alleles

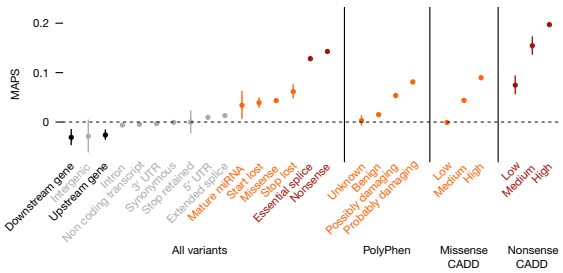
21

Site Frequency Spectrum



22

Diversity and allele frequency



23

Am J Hum Genet 26:669–673, 1974

MAHLON V. B. FREEMAN, M.D.

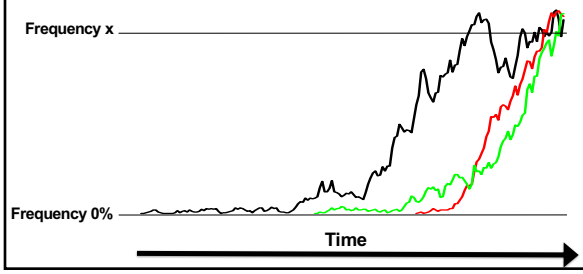
The Age of a Rare Mutant Gene in a Large Population

TAKEO MARUYAMA[1]

24

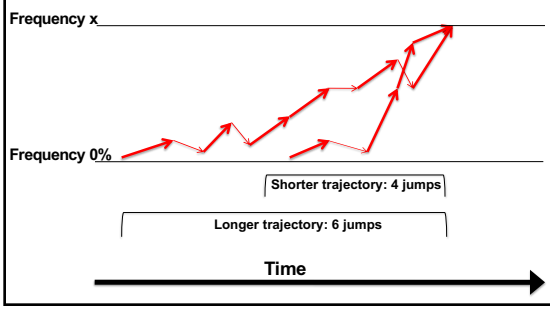
At a given frequency deleterious and advantageous alleles are younger than neutral

Maruyama effect (1974): at any frequency **advantageous** , or **deleterious** alleles are younger than **neutral** alleles

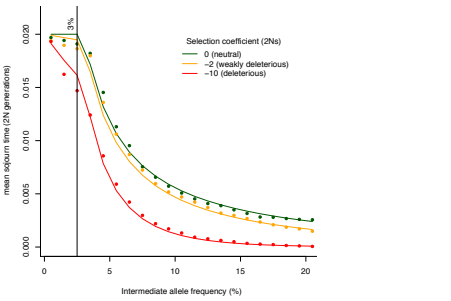


25

Intuition: shorter trajectories require fewer lucky jumps



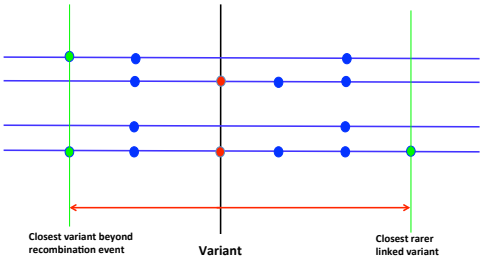
26



Kiezun et al. *PLOS Genetics* 2013

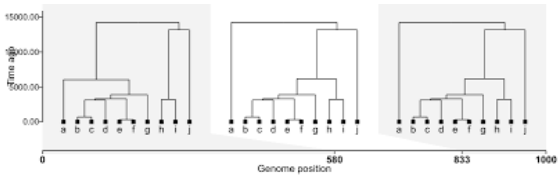
27

Neighborhood clock (fuzzy clock)



28

Tree sequences



29

Selection inference using frequencies of individual SNPs

Change in allele frequency **=**

= ~~***Mutation***~~ **+** ***Selection*** **+** ***Drift***

Of the order of 10^{-8}

Demographic history

Population structure

30

Focusing on rare deleterious PTVs

PTV – protein truncating variant (a.k.a. nonsense)

Combine all PTVs per gene – we assume that they have identical effects

Consider each gene as a bi-allelic locus – PTV / no PTV

31

Selection inference using combined frequency of PTVs

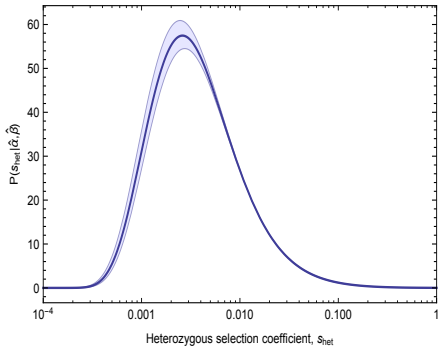
Change in allele frequency **=**

= ***Mutation*** **+** ***Selection*** **+** ~~***Drift***~~

Assuming string selection and a very large population, combined frequency of rare deleterious PTVs is expected to be Poisson distributed with $\lambda=U/hs$

32

Distribution of selection coefficients



Cassa, Weghorn, Balick, Jordan et al. *Nature Genetics*

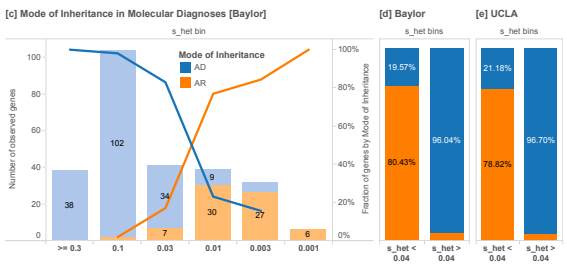
33

What happens if we incorporate drift?

1) The approach fails if selection is weak
2) The approach fails if mutational target is small
3) These considerations are important for regional constraint scores
4) Overall, the approach is non-informative in case of recessivity

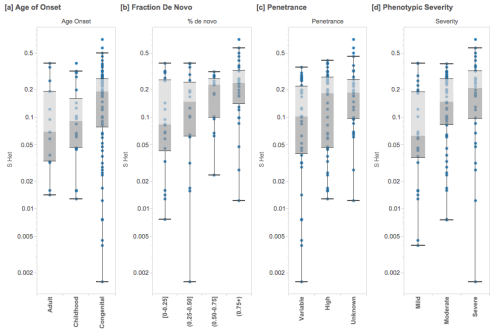
34

AD and AR Mendelian genes



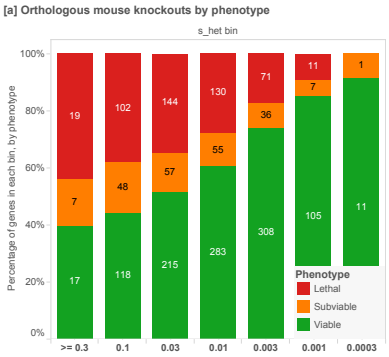
35

Age of onset, penetrance and severity



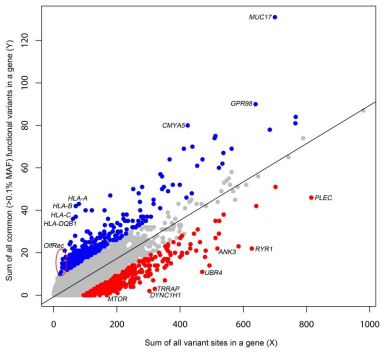
36

Concordance with mouse knockout data



37

RVIS



Petrovski et al. *PLOS Genetics* 2013

38

pLI

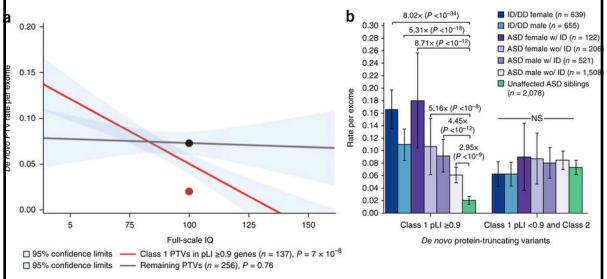
$$PTV_i \mid Z_i = c \sim Pois(N\lambda_c)$$

$$p(Z_i = c \mid \pi_c, PTV_i) = \frac{Pois(PTV_i \mid N\lambda_c)\pi_c}{\sum_c Pois(PTV_i \mid N\lambda_c)\pi_c},$$

Lek et al. *Nature* 2016

39

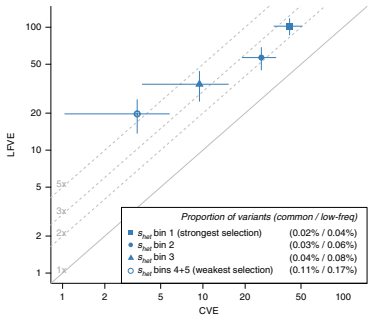
De novo mutations in ASD



Kosmicki et al. *Nature Genetics* 2017

40

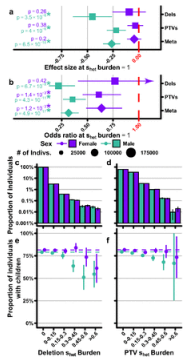
Heritability enrichment



Gazal et al. *Nature Genetics* 2018

41

Surprisingly, the estimates of historic selection seem relevant now



Gardner et al. *Nature* 2022

42

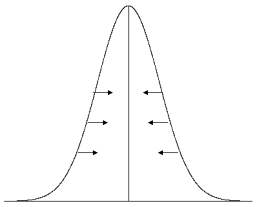
Evidence from modern populations

The overall mutation burden correlates with risk of mental illness

The overall mutation burden correlates with measures of mortality and morbidity

43

Stabilizing selection is the most common type of selection on a quantitative trait



Stabilizing selection

Selection may be related or unrelated to the trait

44

Technically, non-neutral genetic variation should not exist!

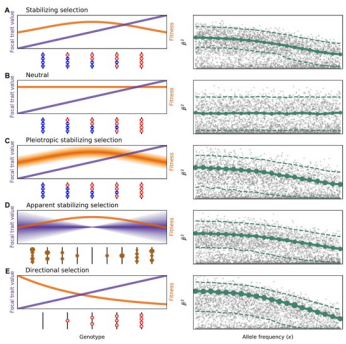
Forces to maintain variation:

Selection

Mutation

45

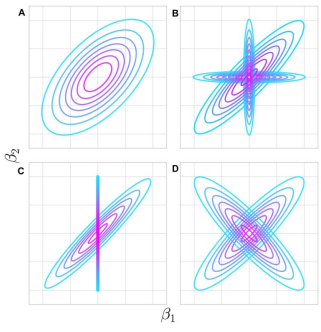
Possible theoretical models



Koch & Sunyaev *Front. Genet.* 2021

46

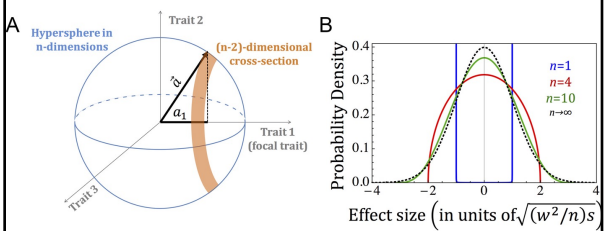
Shades of pleiotropy



Koch & Sunyaev *Front. Genet.* 2021

47

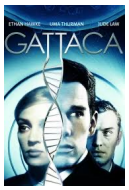
A highly pleiotropic model



Simons et al., *PLOS Biology* 2018

48

Genetic risk prediction



Genotype of an individual (Common SNPs) → Life-time risk of genetic disorders (Common complex genetic disorders)

1

Effect sizes of individual variants are very small

- Genotype at a single locus carries very little information about phenotype.
- It does not mean that one cannot predict phenotype from genotype.
- Accuracy (r^2) of an ideal genetic predictor equals heritability.

2

Measuring risk of myocardial infarction

Coronary Risk Prediction in Adults (The Framingham Heart Study)

PETER W.F. WILSON, MD, WILLIAM P. CASTELLI, MD, and WILLIAM B. KANNEL, MD

The Framingham Heart Study, an ongoing prospective study of adult men and women, has shown that certain risk factors can be used to predict the development of coronary artery disease. These factors include age, gender, total cholesterol level, high-density lipoprotein cholesterol level, systolic blood pressure, cigarette smoking, glucose intolerance and cardiac enlargement (left ventricular hypertrophy on electrocardiogram or enlarged heart on chest x-ray). Calculators and computers can be easily programmed using a multivariate logistic function that allows calculation of the conditional probability of cardiovascular events. These determinations, based on experience with 5,209 men and women participating in the Framingham study, estimate coronary artery disease risk over variable periods of follow-up. Modeled incidence rates range from <1% to >80% over an arbitrarily selected 8-year interval; however, they are typically <10%, and rarely exceed 45% in men and 25% in women.

(Am J Cardiol 1987;59:91G–94G)

3

LDL levels and risk of disease

Annals of Internal Medicine | ARTICLE

Nonoptimal Lipids Commonly Present in Young Adults and Coronary Calcium Later in Life: The CARDIA (Coronary Artery Risk Development in Young Adults) Study

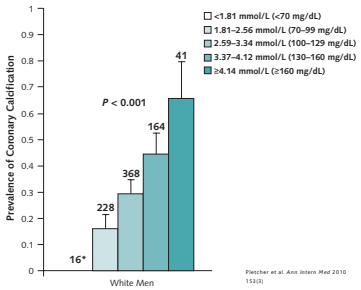
~3500 subjects < 35 years old

15-20 years

Pletcher et al. BMC Cardiovascular Disorders 2008

4

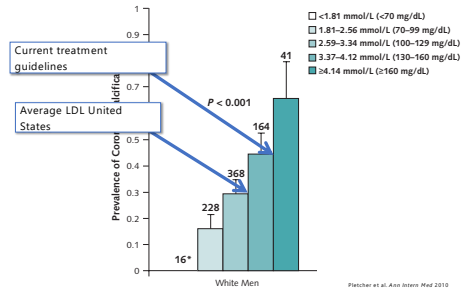
LDL levels and risk of disease



Pletcher et al. Ann Intern Med 2010 153(3)

5

LDL levels and risk of disease



Pletcher et al. Ann Intern Med 2010

6

Selecting populations for treatment

7

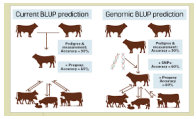
Why estimate genetic risk?

- An estimate of the long-term risk at birth
- Genetic risk can be combined with biomarkers and clinical features
- Genetics explains about 50% of risk. One cannot predict risk any better than that but 50% is a non-trivial proportion of risk

8

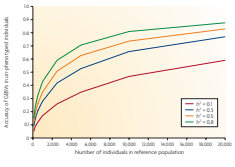
BLUP – Best Linear Unbiased Predictor

- Infinitesimal model
- Genetic effects are random
- Predict the expected genetic effect



9

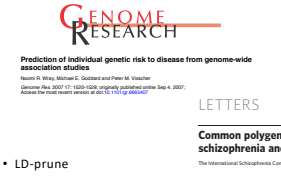
Accuracy of polygenic prediction in cattle



Poor transferability between breeds!

10

Applications in humans



- LD-prune
- Exclude SNPs of very small effect

11

Extensions of BLUP – multiple variance scales and binary phenotypes

MultiBLUP:	Speed and Balding. *Genome Research* 2014
Bayesian analysis:	MacLeod et al. *Genetics* 2014
BSLMM:	Zhou et al. *PLOS Genetics* 2013
GeRSI:	Golan and Rossett. *AJHG* 2014

12

Methods that work with summary statistics

- Summary statistics are easily available
- Most methods require a separate small individual level dataset to tune parameters

13

13

LDPred – a Bayesian method using summary statistics

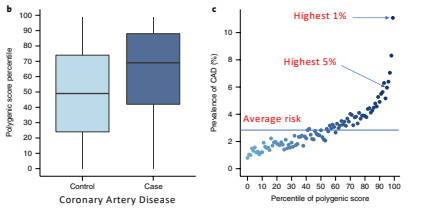
$$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) \text{with probability } p \\ 0 \text{ with probability } (1-p), \end{cases}$$

Vilhjalmsson et al. 2015

Also, check *BayesR*

14

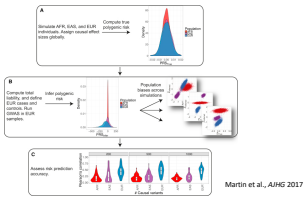
Extreme tails in the distributions of genetic risk scores are highly predictive



Khera et al. 2018

15

With some caveats



Martin et al., *AJHG* 2017

16

Linear models for genetic risk prediction

$$y_i = \sum_j \beta_j x_{ij}$$

Genetic risk of individual i

Genotype of SNP j and individual i

Effect size of SNP j

17

"Polygenic scores" can leverage summary statistics from a large GWAS study

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$

Predicted genetic risk

Estimated effect size

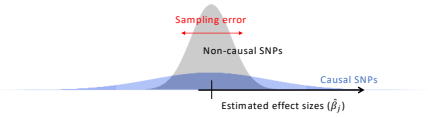
18

"Polygenic scores" can leverage summary statistics from a large GWAS study

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$

Predicted genetic risk

Estimated effect size

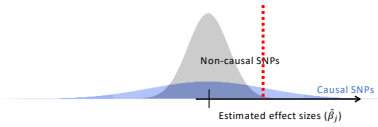


19

"Polygenic scores" can leverage summary statistics from a large GWAS study

P-value Thresholding

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$

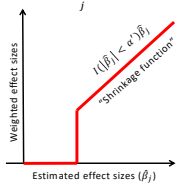


20

P-value thresholding can be reformulated as "shrinking" estimated effect sizes

P-value Thresholding

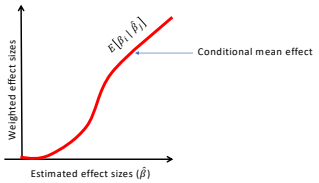
$$\hat{y}_i = \sum_j I(|\hat{\beta}_j| < \alpha')\hat{\beta}_j x_{ij}$$



21

The optimal polygenic score can be constructed with "conditional mean effects"

$$\hat{y}_i = \sum_j E[\beta_j \mid \hat{\beta}_j] x_{ij}$$

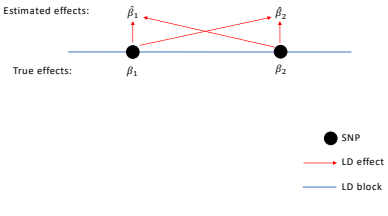


Goddard et al. 2009

22

Accounting for LD in summary data is a major challenge

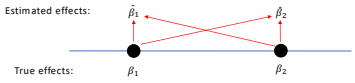
- Correlation between **apparent true genetic effects**



23

Accounting for LD in summary data is a major challenge

- Correlation between **apparent true genetic effects**



- Correlation between **sampling errors**

24

Our approach ("**N**on-**P**arametric **S**hrinkage" or NPS)

- No explicit specification of genetic architecture prior, thus "*non-parametric*"
- Learn conditional mean effects directly from training data
- Fully account for correlation in summary statistics

25

Our approach ("**N**on-**P**arametric **S**hrinkage" or NPS)

- No explicit specification of genetic architecture prior, thus "*non-parametric*"
- Learn conditional mean effects directly from training data

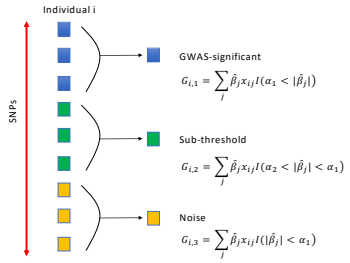
1. How to estimate $E[\beta_j \mid \hat{\beta}_j]$ without a Bayesian prior on $\boldsymbol{\beta}$

- Fully account for correlation in summary statistics

2. How to deal with LD

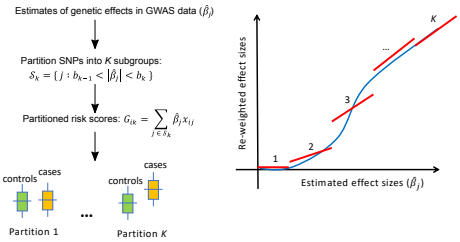
26

Partitioned risk scores



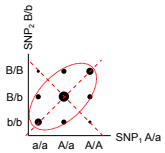
27

Piecewise linear interpolation on shrinkage curve



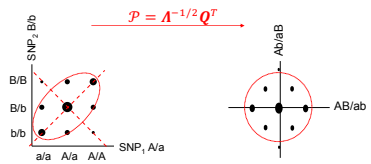
28

How to deal with LD?



29

Decorrelating linear projection $\mathcal{P}$



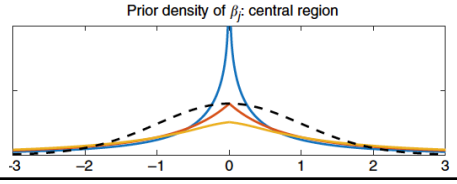
$\boldsymbol{\Sigma}$ is a local LD matrix and $\boldsymbol{\Sigma} = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^T$ by eigenvalue decomposition

$$\boldsymbol{\Sigma^{-1}} = \boldsymbol{Q}\,\boldsymbol{\Lambda^{-1}}\,\boldsymbol{Q}^T = (\boldsymbol{Q}\,\boldsymbol{\Lambda^{-1/2}})(\boldsymbol{\Lambda^{-1/2}}\boldsymbol{Q}^T)$$

30

Other shrinkage methods: PRS-CS

$$\beta_j \sim \mathrm{N}\left(0, \frac{\sigma^2}{N}\phi\psi_j\right), \qquad \psi_j \sim g,$$

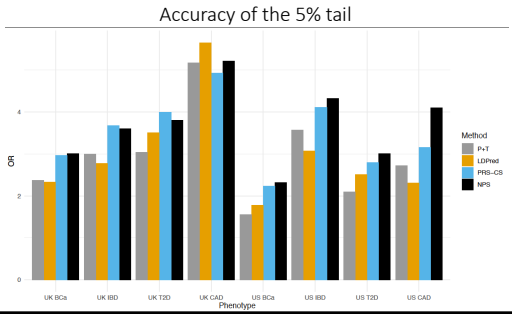


31

Other shrinkage methods: PRS-CS

Lassosum – extension of *LASSO*

32



33

Summary

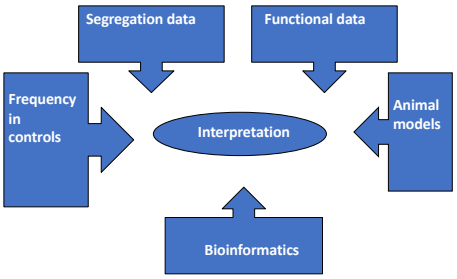
- NPS accounts for the correlation of sampling errors in GWAS summary statistics.
- NPS provides an extensible framework to estimate the shrinkage curve from training data.
- NPS is best-suited to take advantage of the high density of markers and imputation accuracy in latest GWAS datasets.

The preprint is available in BioRxiv:
Chun et al. AJHG 2020 *"Non-parametric polygenic risk prediction via partitioned GWAS summary statistics."*
Software is available at: https://github.com/sgchun/nps

34

Identifying functionally significant variants

1



2

Map variants onto genomic annotation

Watch for multiple transcripts!

Watch for conflicting annotations!

3

Nonsense variants

One of most significant types of variants usually leading to the complete loss of function.

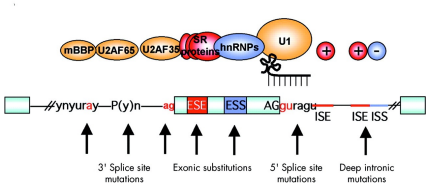
Nonsense variants are enriched in sequencing artifacts

Important considerations: i) location along the gene, ii) does the variant cause NMD? iii) is the variant in a commonly skipped exon?

Tool: LOFTEE

4

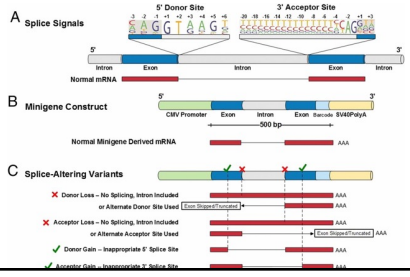
Variants involved in splicing



1) Variants in canonic splice sites
2) Variants in exonic or intronic splicing enhancers
3) Gain of splicing variants

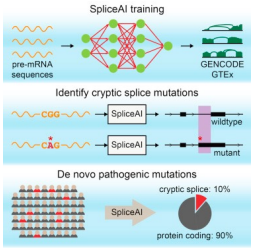
5

Experimental Methods: Minigene Assay and Massively Parallel Splicing Assay (MPSA)



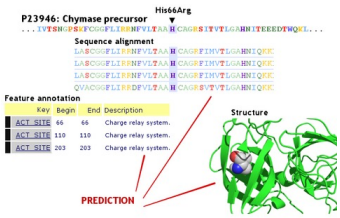
6

Computational Predictions: SpliceAI, MMSplice and other methods



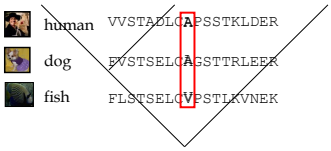
7

Missense variants: computational predictions



8

Does the mutation fit the pattern of past evolution?



Statistical issues:
-sequences are related by phylogeny
-generally, we have too few sequences

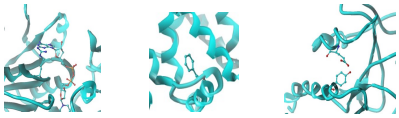
9

Does the mutation fit the pattern of past evolution?

- We assume a constant fitness landscape: what is good for fish is good for human!
- We can estimate whether the mutation fits the pattern of amino acid changes.
- We can also estimate rate of evolution at the amino acid site

10

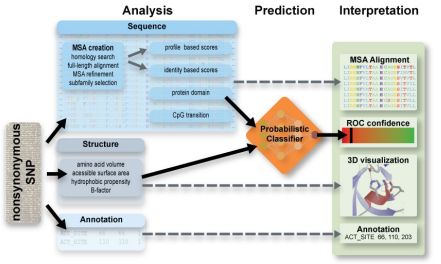
Protein structure view



- **Most of pathogenic mutations are important for stability (good news?).**
- **$\Delta\Delta G$ is difficult to estimate.**
- **Unfolded protein response pathway has to be taken into account.**
- **Heuristic structural parameters help but less than comparative genomics.**

11

PolyPhen2



www.genetics.bwh.harvard.edu/pph2 Adzhubei, et al. Nature Methods 2010

12

Weakly deleterious mutations

- Multiple independent lines of evidence suggest abundance of weakly deleterious alleles in humans
- Weakly deleterious variants may occur in highly conserved positions
- Weakly deleterious alleles probably contribute to complex phenotypes but not to simple Mendelian phenotypes

13

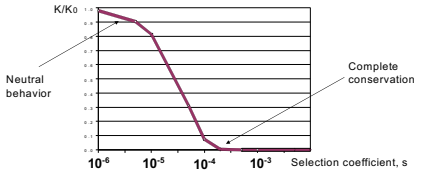
Conservation can be due to very weak selection!

Every new mutation eventually will be either fixed or lost

$$K = K_0 2N_e \frac{(1-e^{-2s})}{(1-e^{-4N_e s})}$$

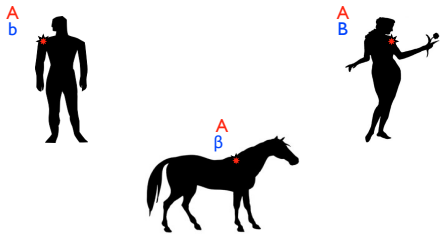
s – selection coefficient
N_e - effective population size

For humans estimated to be ~ 10 000



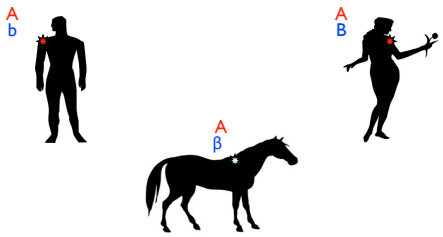
14

Constant fitness landscape



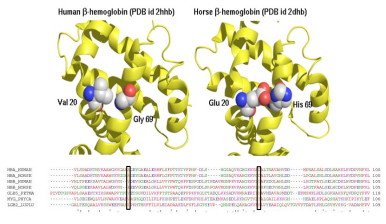
15

Epistatic interactions



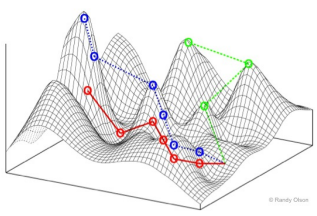
16

Compensatory mutations



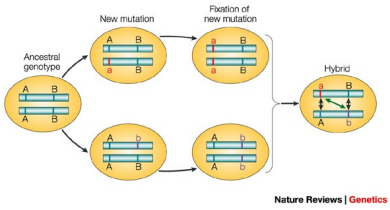
17

Ridges on the fitness landscape



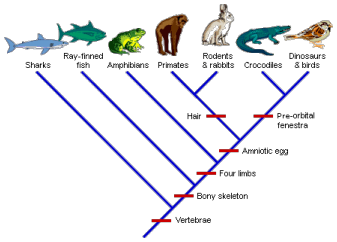
18

Dobzhansky-Muller incompatibility



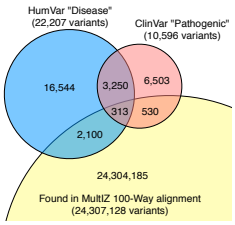
19

Looking at vertebrate species



20

Many human pathogenic mutations are found in vertebrates

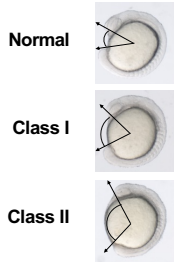


5.5-6.5% of presumably pathogenic human mutations are detected in mammals

21

Zebrafish model

- Model of Bardet-Biedl Syndrome (obesity, renal failure, vision loss)
- Caused by defects in primary cilium
- Embryonic convergence / extension phenotype in zebrafish
- Easily scorable phenotype



Images: Phoebe

22

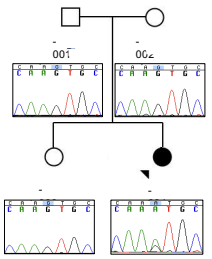
Testing double mutants

No injection		Human gene with disease mutant	
Knockdown		Double mutant (no suppression)	
Rescue with human gene		Double mutant (full suppression)	

Images: Phoebe

23

A newly identified gene



Clinical features
Global developmental delay
microcephaly
feeding issues
failure to thrive
abnormal muscle tone
low immunoglobulins
frequent respiratory infections

Clinical testing
normal female microarray
metabolic testing – negative
extensive genetic testing – negative

BTG2 De novo	*TTN* Compound het
NOS2 De novo	*LAMA1* Compound het

Stephan Frangakis

24

The mutation is a reversal to the mammalian ancestral state

BTG2	R80	L128	Q140	V141	L142
H. sapiens	R	L	Q	V	L
P. troglodytes	•	•	•	•	•
G. gorilla	•	•	•	•	•
M. musculus	K	V	•	M	M
R. norvegicus	K	V	•	M	M
H. glaber	•	V	•	M	M
S. domesticus	K	V	•	M	M
B. primigenius	K	V	•	M	M
E. ferus caballus	K	V	•	M	M
F. catus	K	V	•	M	M
C. lupus familiaris	K	V	•	M	M
D. novemcinctus	K	V	•	M	M
G. gallus	K	P	•	M	M

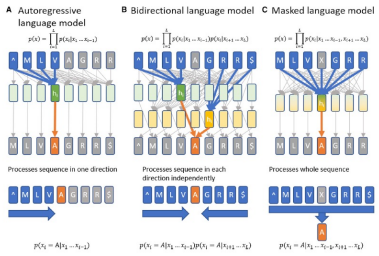
25

New methods directions

- Machine learning techniques have the potential to solve the epistasis problem
- Measures of population level constraint have the potential to solve the problem of distinguishing between strongly and weakly deleterious mutations.

26

Language models



Bepler & Berger, *Cell Systems*, 2021

27

Methods

PolyPhen2

SIFT

LRT

MutationTaster

FatHMM

SNPs3D

DeepSequence

Note, that state of the art is far from perfect

28

New developments

EVE – addresses the issue of epistatic interactions

VARITY – addresses the issue of slightly deleterious mutations

29

Umbrella methods

Condel

REVEL

VEST

CADD

M-CAP

30

Incorporating regional constraint

CCR

M-CAP

PrimateAI

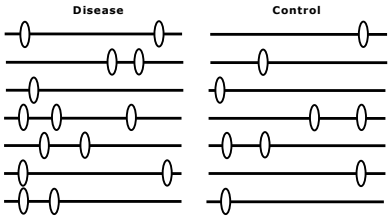
31

Applications

- Mendelian genetics – obvious
- Rare variant studies in search of drug targets

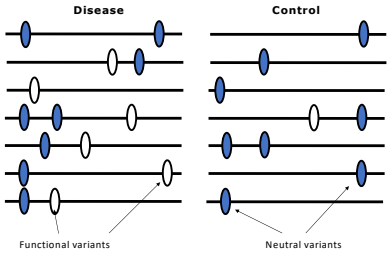
32

Rare variant collapsing study



33

Rare variant collapsing study



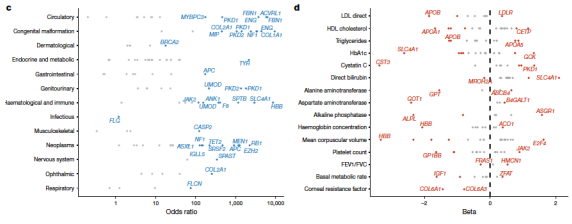
34

Predicting functional consequences increases power

- Inclusion of neutral variants reduces power of the test
- Combining variants with vastly different effect sizes reduces power of the test
- Most groups limit the tests to nonsense, splicing and missense variants that are predicted functional
- Assigning quantitative weights is probably a better approach, but nobody uses it in practice

35

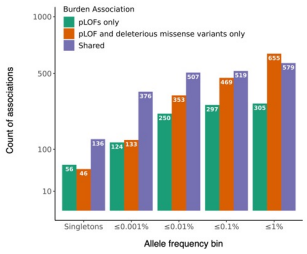
UK Biobank results (Wang et al.)



Variant grouping: nonsense, splicing, missense predicted by REVEL and MTR

36

UK Biobank results (Backman et al.)



Deleterious missense variants:

SIFT
PolyPhen2
LRT
Mutation Taster

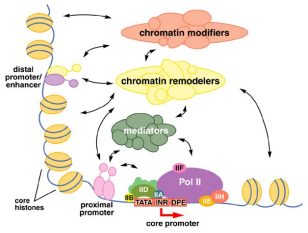
37

Non-coding variants

38

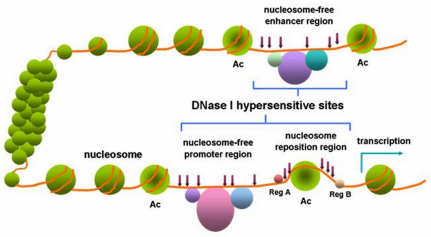
Regulatory variants

- Regulation: variants in promoters, enhancers, silencers, insulators



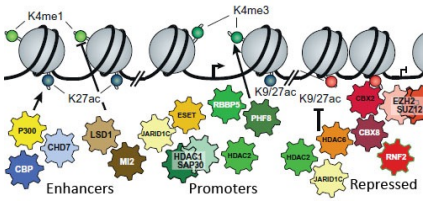
39

Chromatin accessibility



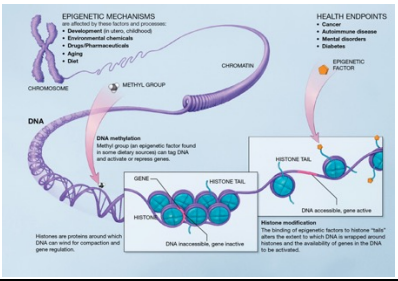
40

Chromatin modifications



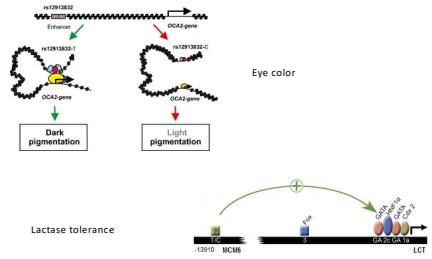
41

Epigenomics



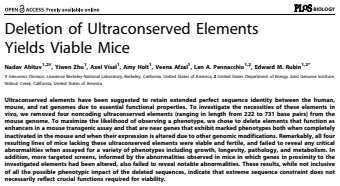
42

Non-disease alleles of large effect



43

Ultraconserved elements



OPEN ACCESS Freely available online

PLoS BIOLOGY

Deletion of Ultraconserved Elements Yields Viable Mice

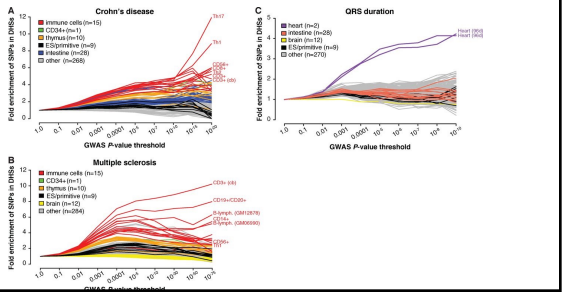
Nadav Ahituv[1,2*], Yiwen Zhu[1], Axel Visel[3], Amy Holt[1], Veena Afzal[3], Len A. Pennacchio[1,2], Edward M. Rubin[1,2*]

1 Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 2 United States Department of Energy Joint Genome Institute, Walnut Creek, California, United States of America

Ultraconserved elements have been suggested to retain extended perfect sequence identity between the human, mouse, and rat genomes due to essential functional properties. To investigate the necessities of these elements in vivo, we removed four noncoding ultraconserved elements (ranging in length from 222 to 731 base pairs) from the mouse genome. To maximize the likelihood of observing a phenotype, we chose to delete elements that function as enhancers in a mouse transgenic assay and that are near genes that exhibit marked phenotypes both when completely inactivated in the mouse and when their expression is altered due to other genomic modifications. Remarkably, all four resulting lines of mice lacking these ultraconserved elements were viable and fertile, and failed to reveal any critical abnormalities when assayed for a variety of phenotypes including growth, longevity, pathology, and metabolism. In addition, more targeted screens, informed by the abnormalities observed in mice in which genes in proximity to the investigated elements had been altered, also failed to reveal notable abnormalities. These results, while not inclusive of all the possible phenotypic impact of the deleted sequences, indicate that extreme sequence constraint does not necessarily reflect crucial functions required for viability.

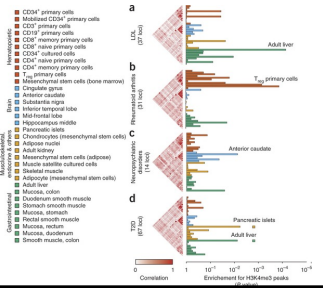
44

Enrichment of GWAS signals in regulatory elements



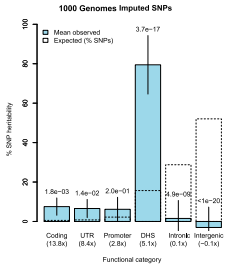
45

Enrichment of GWAS signals in regulatory elements



46

Partitioning heritability



47

Translating GWAS findings into mechanistic models

GWAS peak

↓

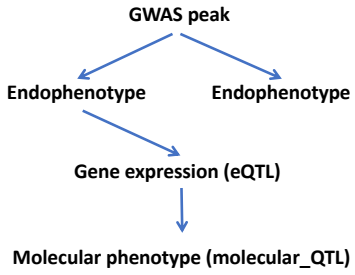
Controlled model system

↓

Biochemistry

48

Human Genetics all the way



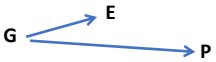
49

Causality

Mediation

G → E → P

Independent effects



Reverse causation

G → P → E

50

Co-localization

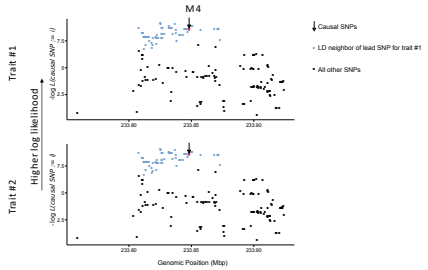
Same causal variant

G → E → P

Distinct variants

51

Co-localization problem



52

Methods

Coloc

eCAVIAR

JLIM

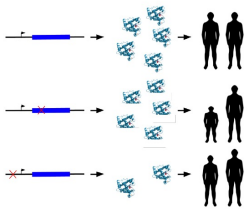
53

Genetic variants differ between Mendelian and complex traits

- **Complex trait variants**
  - Small effect size
  - Extremely large number of loci
  - Mostly non-coding (regulatory)
- **Mendelian & somatic cancer variants**
  - Large effect sizes
  - Small number of loci
  - Mostly coding
  - Are in "putatively causative" genes

54

The basic model



https://commons.wikimedia.org/wiki/File:Myoglobin.png
https://commons.wikimedia.org/wiki/File:Human_outline_generic.svg

By now we know that most complex trait loci never harbor mutations of large effect

55

Hypothesis

- Most genes involved in Mendelian components of complex traits are also causative for cognate common forms.
- Variants involved in common forms alter regulatory sequence of these genes.
- This in turn induces changes in gene expression; regulatory variants are *eQTLs*.

56

Genes and phenotypes

(for complex traits, GWAS is restricted to non-coding variants)

Mend. trait	GWAS trait	Tissue
Breast cancer	Breast cancer	breast mammary tissue
Crohn disease	Crohn's disease	small intestine terminal ileum colon sigmoid colon transverse
Dyslipidemia Hyperlipidemia Tangier's disease	HDL	liver adipose whole blood
Dwarfism	Height	skeletal muscle
Blood pressure	Blood pressure	heart atrial appendage kidney heart left ventricle
Dyslipidemia Hyperlipidemia	LDL	liver adipose tissue whole blood
Monogenic diabetes	Type II diabetes	pancreas skeletal muscle adipose whole blood
Ulcerative colitis	Ulcerative colitis	small intestine terminal ileum colon sigmoid colon transverse

Overall, 139 genes

89 (64%) fall under a GWAS peak of a cognate complex trait

Examples include:

LDL Receptor under a GWAS peak for LDL Cholesterol

Estrogen receptor under a GWAS peak for breast cancer

These genes are highly likely to mediate the effects of regulatory variants

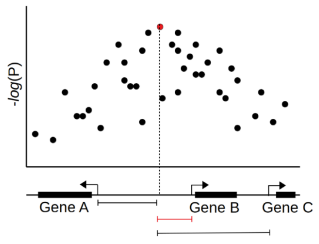
57

Statistical methods to locate the causative gene under GWAS peak

- Closest gene to peak
- Colocalization methods
  - JLIM
  - Coloc
  - eCAVIAR
- Transcriptome-wide association
  - FUSION
- Chromatin marks
  - Fine-mapping using SuSiE
  - Locate fine-mapped variants under chromatin modification peaks

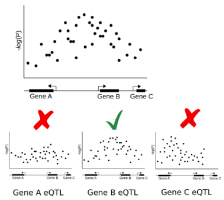
58

Distance of fine-mapped SNPs (by SuSiE) to the closest gene



59

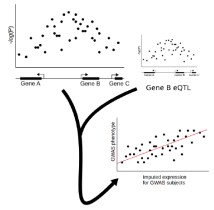
Colocalization of GWAS and eQTLs



Methods effectively compare the shape of two peaks.
Colocalization often returns multiple hits per locus.

60

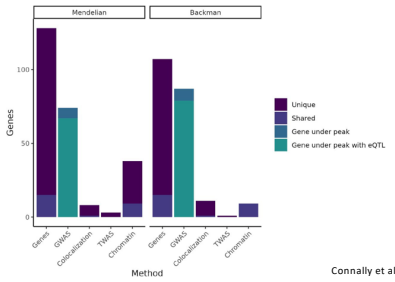
Transcriptome-wide association (TWAS)



TWAS often returns multiple hits per locus.

61

Results



Connally et al., *medRxiv*, 2021

62

Results

- Colocalization results are almost random
- TWAS results are almost random
- There is a significant enrichment of fine-mapped SNPs in regulatory elements marked by modified chromatin, but still in a minority of loci
- A simple strategy to link peaks to genes using the chromatin signal (ABD) does not link most of the genes successfully

63

Possible reasons for the negative result

- Perhaps expression of putatively causative genes
  - Matters only in a certain environmental context
  - Matters only at a specific developmental stage
  - Matters only as a stimulus response

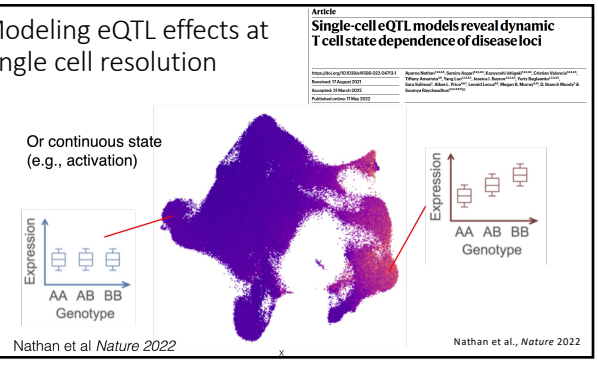
64

I find it highly surprising that

- A context independent large change in expression of LDLR due to a nonsense mutation leads to a large phenotypic change
- A smaller change in expression does not affect LDL levels, while non-coding effect on LDLR does

65

Modeling eQTL effects at single cell resolution



Nathan et al *Nature 2022*

Nathan et al., *Nature* 2022

66