Advanced Gene Mapping Course

May 22-26, 2023
The Rockefeller University
New York, NY

Lectures

Table of Contents

Lectures given by: [1]Heather Cordell, [2]Suzanne Leal, Wayne Patterson; [4]Gao Wang, [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
  - OR $(2|2 : 1|1) = \frac{af}{be}$
  - 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 hypothesis 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

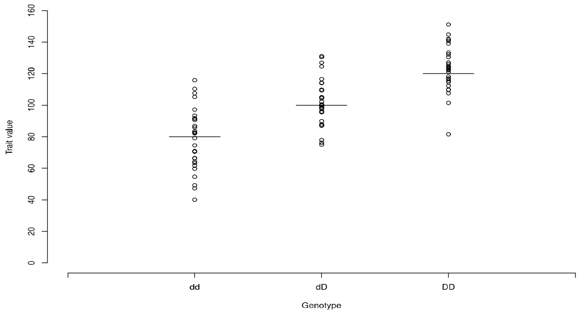
	Counts in	
Allele	Cases	Controls
2	2100 ($=a$)	1220 ($=b$)
1	1900 ($=c$)	2780 ($=d$)
Total	4000	4000

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.

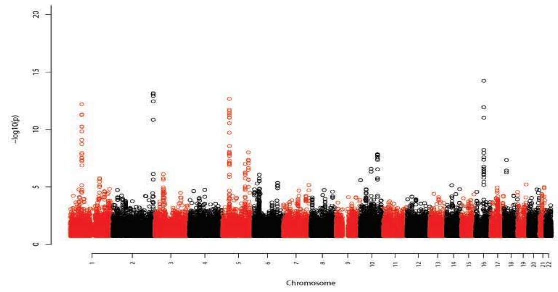
Testing for association

- 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

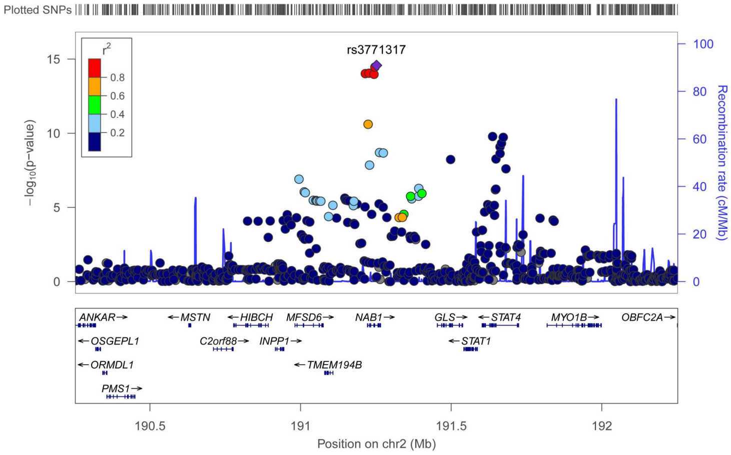
- All methods produce a test statistic and a p value at each SNP, indicating how significant the association/correlation observed appears to be
  - i.e. how likely it was to have occurred by chance
  - The threshold to declare 'genome-wide significance' is usually around $p = 5 \times 10^{-8}$
    -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

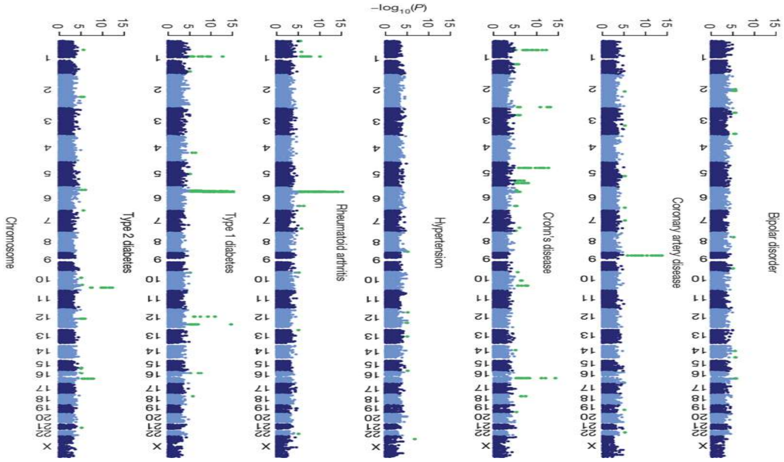
GWAS

- GWAS really got going in around 2007
  - Visscher et al. (2012) AJHG 90:7-24 "Five Years of GWAS Discovery"
  - Visscher et al. (2017) AJHG 101:5-22 "10 Years of GWAS Discovery: Biology, Function and Translation"
  - Abdellaoui et al. (2023) AJHG 110:179-194 "15 Years of GWAS Discovery: Realizing the promise"
- 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)

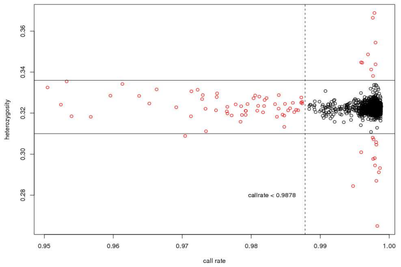
- Typically used rather standard statistical/epidemiological methods (χ^2 tests, t tests, logistic regression etc.)
- Success largely due to:
  - An appreciation of the importance of large sample size (> 2000 cases, similar or greater number of controls)
  - Stringent quality control procedures for discarding 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

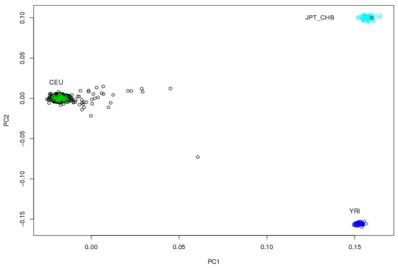
- Stringent QC checks are required for GWAS data
- Discard samples (people) deemed unreliable
  - Low genotype call rates, excess heterozygosity etc.
  - X chromosomal markers useful for checking gender
    - Males should 'appear' homozygous at all X markers
  - Genome-wide SNP data useful for checking 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
- MDS or similar multivariate methods can also 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

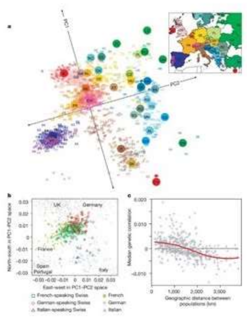
Multivariate Analysis

- 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) gross 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)

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

Principal Components Analysis

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

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

- 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 $\Psi = \mathrm{X}^T\mathrm{X}$ 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 $\Psi = \mathrm{X}^T\mathrm{X}$ 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

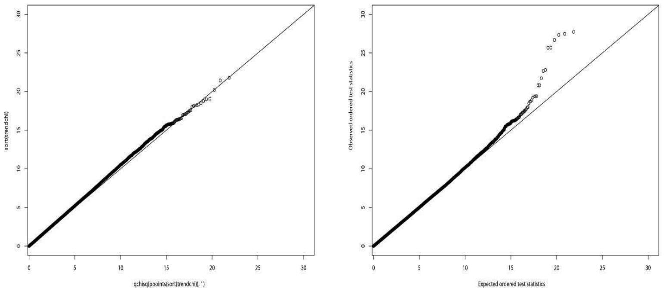
- Estimate covariance matrix $\Psi = \mathrm{X}^T\mathrm{X}$ between all pairs of individuals, with entries ψ_{ij} defined as the covariance (summing over SNPs) between column i and j of X
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute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

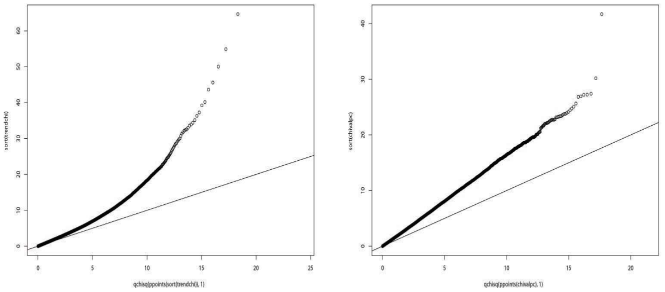
- Estimate covariance matrix $\Psi = \mathrm{X}^T\mathrm{X}$ between all pairs of individuals, with entries ψ_{ij} defined as the covariance (summing over SNPs) between column i and j of X
  - Represents average genome-wide identity by descent (IBD) (estimated from identity by state, IBS)
  - Compute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

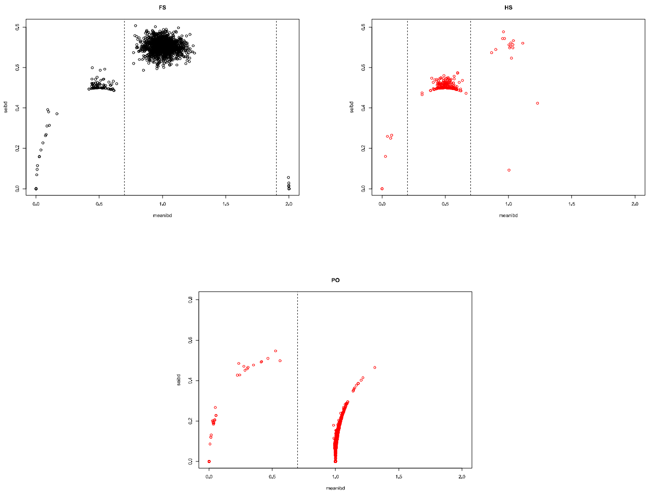
Expected IBD sharing

- Assuming no inbreeding, the IBD state probabilities are:

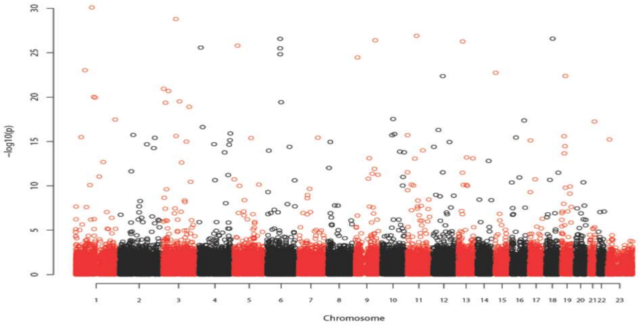
	Number of alleles shared IBD		
Relationship	2	1	0
MZ twins	1	0	0
Parent–Offspring	0	1	0
Full siblings	1/4	1/2	1/4
Half siblings	*0*	*1/2*	*1/2*
Grandchild–grandparent	*0*	*1/2*	*1/2*
Uncle/aunt–nephew/niece	*0*	*1/2*	*1/2*
First cousins	0	1/4	3/4
Second cousins	0	1/16	15/16
Double 1st cousins	1/16	6/16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

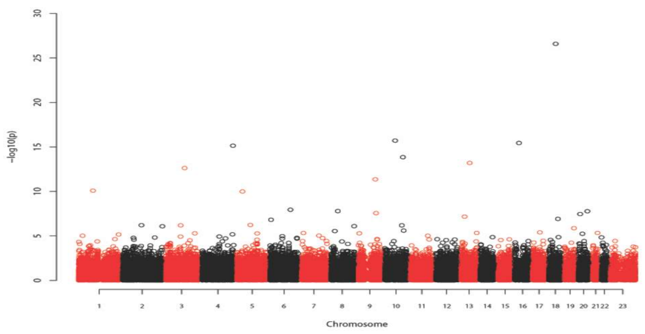
Full/half sibs and parent-offspring



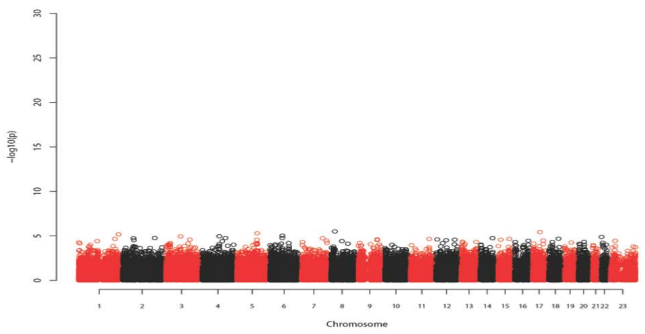
CHD GWAS results (low QC)



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

Suzanne M. Leal, Ph.D.
Sergievsky Family Professor of Neurological Sciences
Director of the Center for Statistical Genetics
Columbia University
sml3@Columbia.edu

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

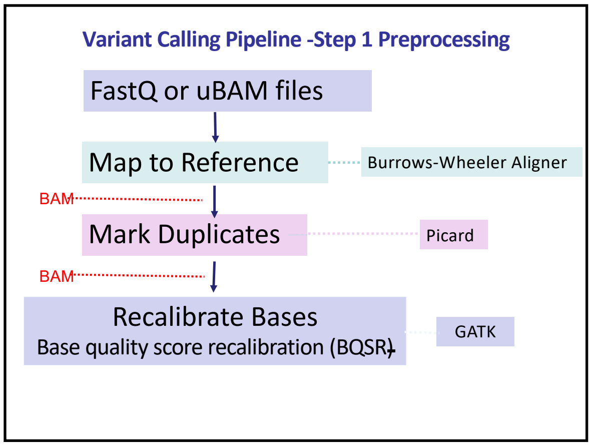
- Genotype markers which can be used as DNA fingerprint
- Allows for Assessment of DNA quality
- Aids in determining the the genetic sex of study subjects
  - To aid in identification of potential sample swaps
- Detects cryptic duplicates
- For family data
  - Aids in determining close familial relationships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

FastQ or uBAM files

Map to Reference — Burrows-Wheeler Aligner

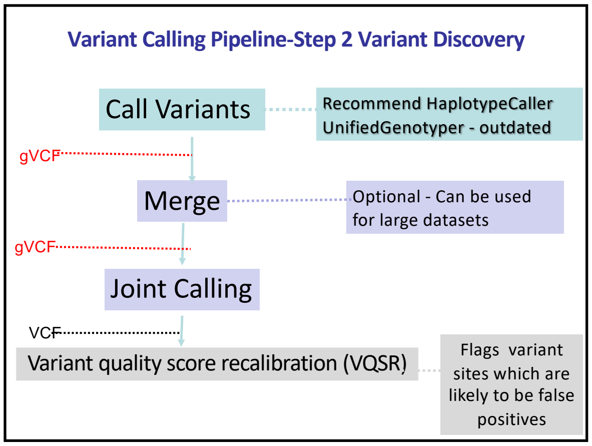
BAM

Mark Duplicates — Picard

BAM

Recalibrate Bases
Base quality score recalibration (BQSR) — GATK

7



Variant Calling Pipeline-Step 2 Variant Discovery

Call Variants — Recommend HaplotypeCaller
UnifiedGenotyper - outdated

gVCF

Merge — Optional - Can be used for large datasets

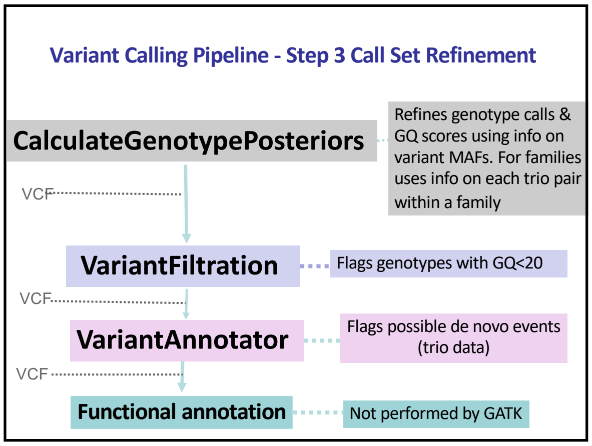
gVCF

Joint Calling

VCF

Variant quality score recalibration (VQSR) — Flags variant sites which are likely to be false positives

8



Variant Calling Pipeline - Step 3 Call Set Refinement

CalculateGenotypePosteriors — Refines genotype calls & GQ scores using info on variant MAFs. For families uses info on each trio pair within a family

VCF

VariantFiltration — Flags genotypes with GQ<20

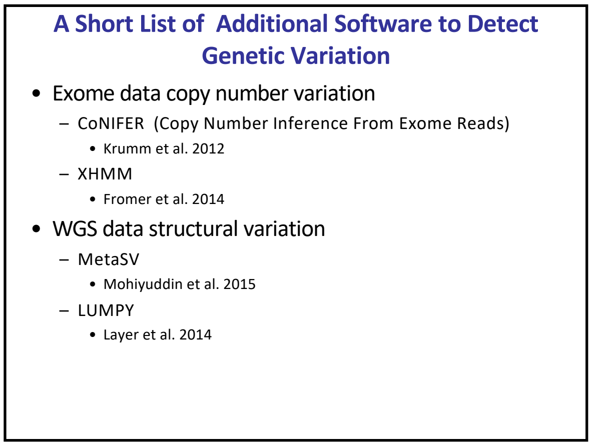
VCF

VariantAnnotator — Flags possible de novo events (trio data)

VCF

Functional annotation — Not performed by GATK

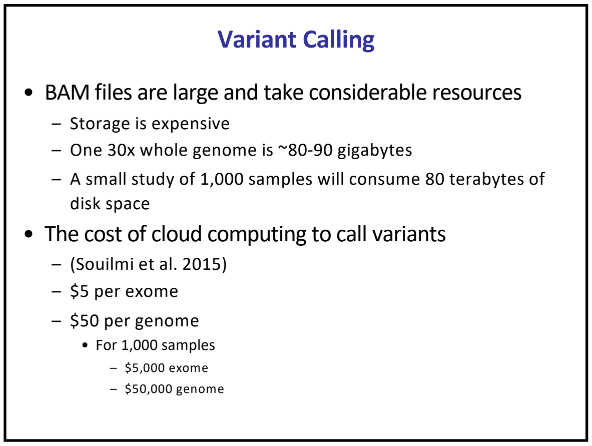
9



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

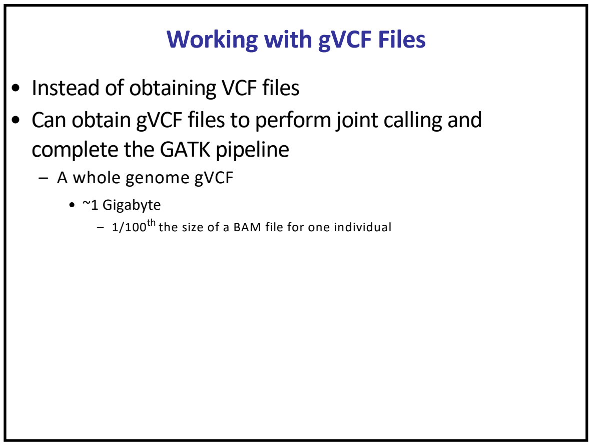
10



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

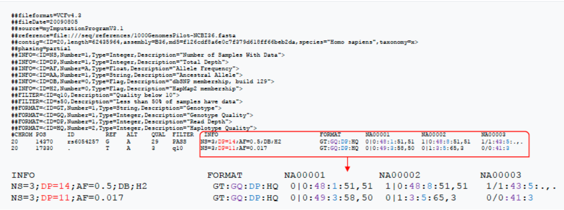
15

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



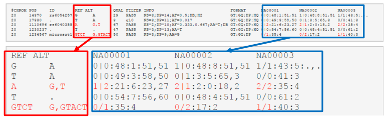
```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT   QUAL  FILTER  INFO                           FORMAT       NA00001         NA00002         NA00003
20     14370   rs6054257 G    A     29    PASS    NS=3;DP=14;AF=0.5;DB;H2        GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20     17330   .         T    A     3     q10     NS=3;DP=11;AF=0.017            GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3
```

```
INFO                         FORMAT        NA00001         NA00002          NA00003
NS=3;DP=14;AF=0.5;DB;H2       GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51   1/1:43:5:.,.
NS=3;DP=11;AF=0.017           GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3     0/0:41:3
```

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



```
#CHROM POS     ID        REF ALT      QUAL FILTER INFO                              FORMAT      NA00001         NA00002         NA00003
20     14370   rs6054257 G    A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ  |0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20     17330   .         T    A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ  |0:49:3:58,50  0|1:3:5:65,3    0/0:41:3
20     1110696 rs6040355 A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ  |2:21:6:23,27  2|1:2:0:18,2    2/2:35:4
20     1230237 .         T    .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ  |0:54:7:56,60  0|0:48:4:51,51  0/0:61:2
20     1234567 microsatl GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP     /1:35:4        0/2:17:2        1/1:40:3
```

```
REF  ALT      NA00001         NA00002         NA00003
G    A        0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
T    A        0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3
A    G,T      1|2:21:6:23,27  2|1:2:0:18,2    2/2:35:4
T    .        0|0:54:7:56,60  0|0:48:4:51,51  0/0:61:2
GTCT G,GTACT  0/1:35:4        0/2:17:2        1/1:40:3
```

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

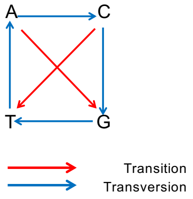
NGS Data Quality Control

- 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

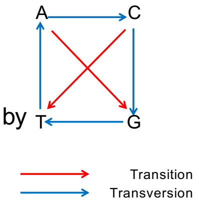
- Transition
  - Purine → Purine
  - Pyrimidine → Pyrimidine
- Transversion
  - Purine → Pyrimidine
  - Pyrimidine → Purine



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

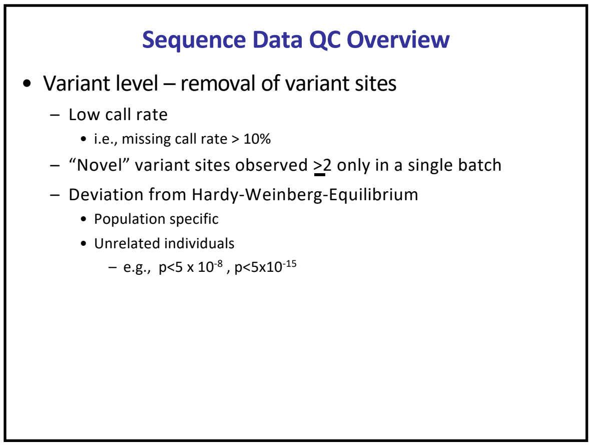


23

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

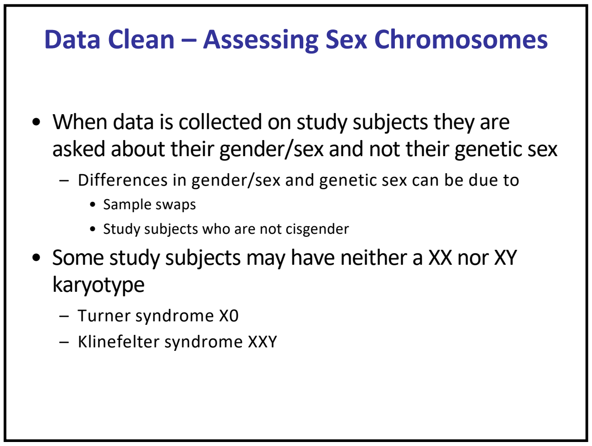
24



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., $p<5 \times 10^{-8}$, $p<5x10^{-15}$

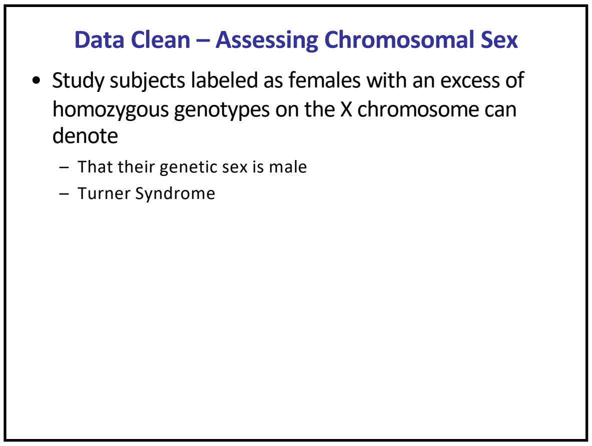
25



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

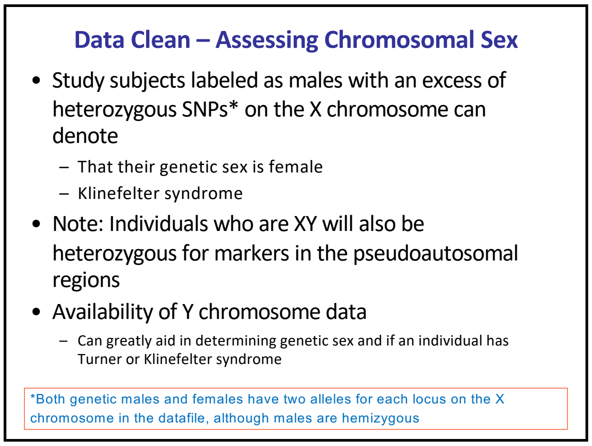
26



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

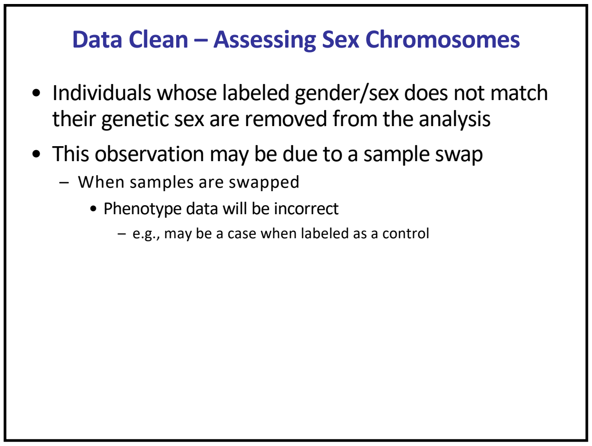


Data Clean – Assessing Chromosomal Sex

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

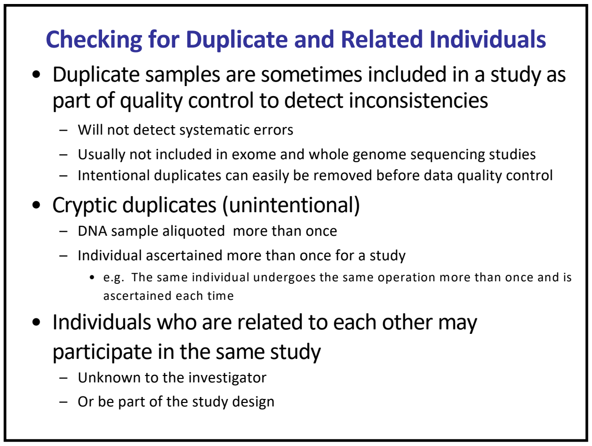
28



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.

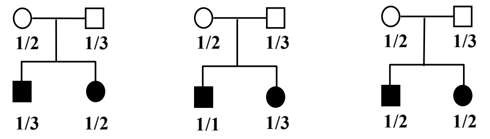
31

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)



IBD=0 IBS=1

IBD=1 IBS=1

IBD=2 IBS=2

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

UK Biobank Related Individuals > Kinship Coefficient 0.0625

White European

# of Relatives	# of relatives
1	86089
2	18491
3	3691
4	707
5	165
6	40
7	9
8	5
9	1
10	11
11	2
12	2
16	1
19	1
25	1
30	1
3985	1

African

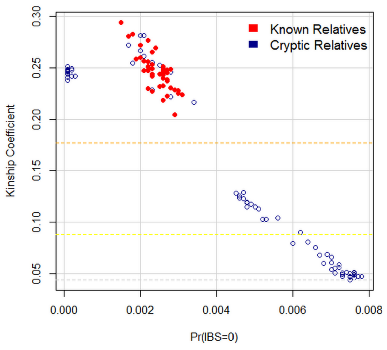
# of relatives	# of individuals
1	715
2	153
3	26
4	10
5	3
6	5
7	5
8	4
9	1
10	4
11	2
13	3
17	2
19	3
20	2
21	1
23	1
.	.
.	.
.	.
390	1
391	1
393	1
396	1

Asian

# of relatives	# of individuals
1	743
2	115
3	33
4	4
5	4

37

King Graphical Output



38

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

39

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

40

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

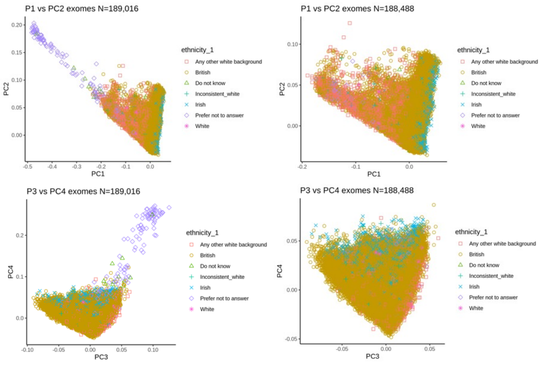
41

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

42

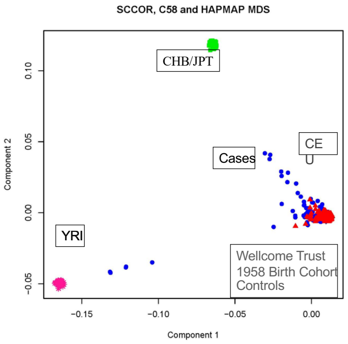
Principal Components Analysis Example



Exclusion of Outliers using Mahalanobis distance (0.997)

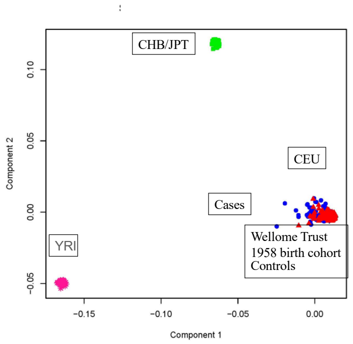
43

Detecting Outliers Using PCA and HapMap Sample



44

Detecting Outliers Using PCA and HapMap Sample



45

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)

46

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

47

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

48

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

49

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

50

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

51

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

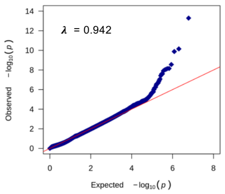
52

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

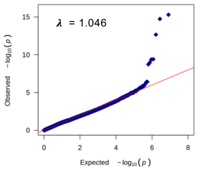
53

QQ Plot of Exome Wide P Values
UK Biobank 200K



Hearing aid users

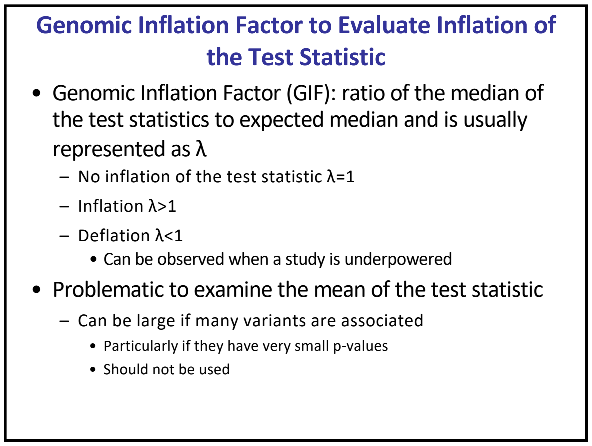
Case N= 6,436
Controls N= 96,601



Problem hearing
with background noise

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

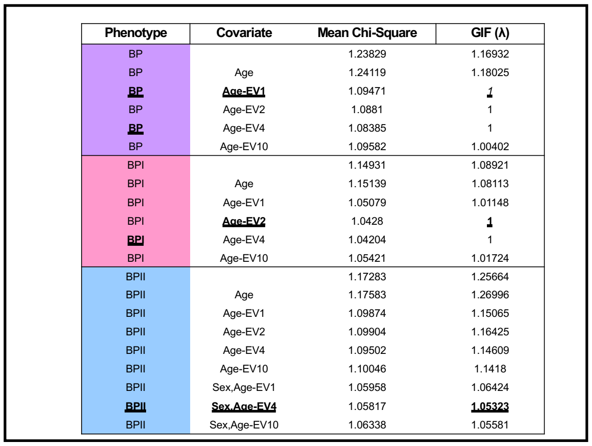
54



Genomic Inflation Factor to Evaluate Inflation of the Test Statistic

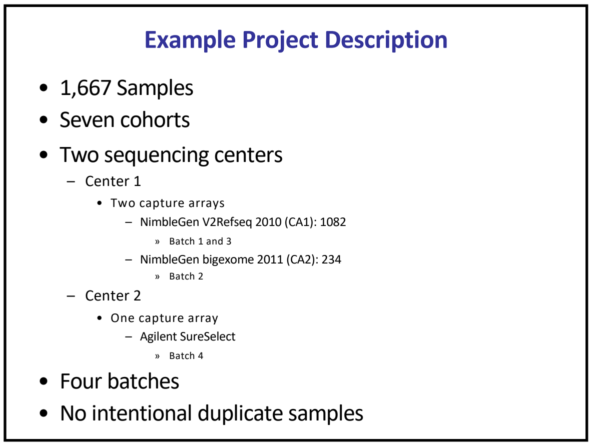
- Genomic Inflation Factor (GIF): ratio of the median of the test statistics to expected median and is usually represented as λ
  - No inflation of the test statistic λ=1
  - Inflation λ>1
  - Deflation λ<1
    - Can be observed when a study is underpowered
- Problematic to examine the mean of the test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

55



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

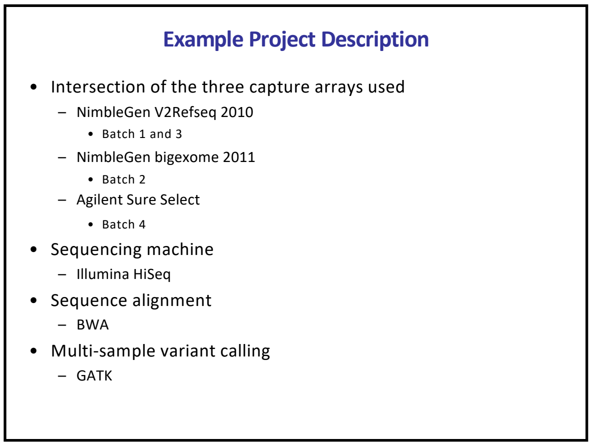
56



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

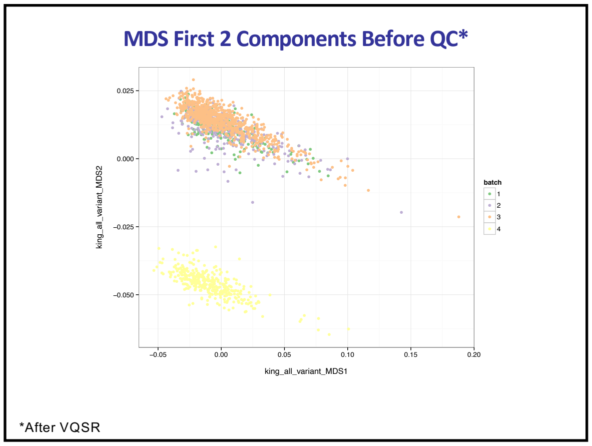
57



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

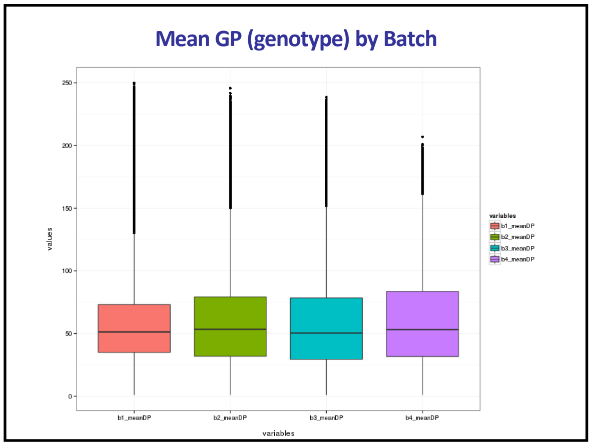
58



MDS First 2 Components Before QC*

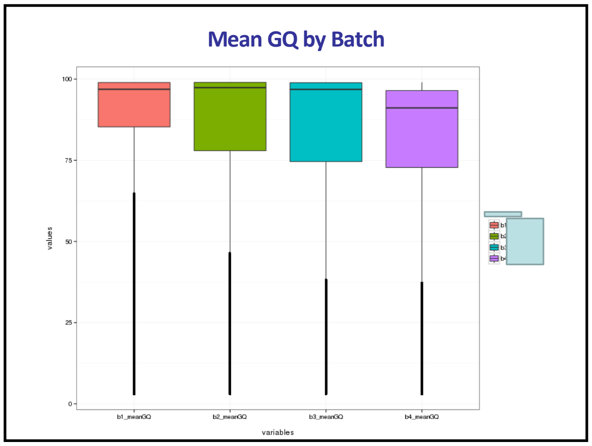
*After VQSR

59



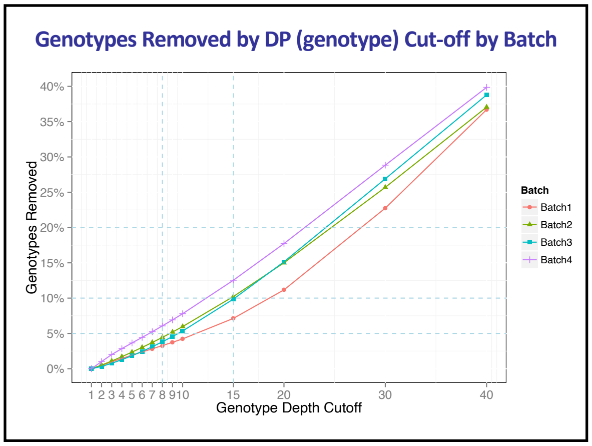
Mean GP (genotype) by Batch

60



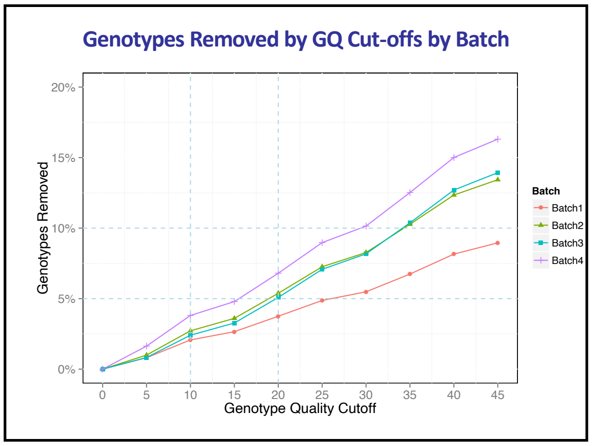
Mean GQ by Batch

61



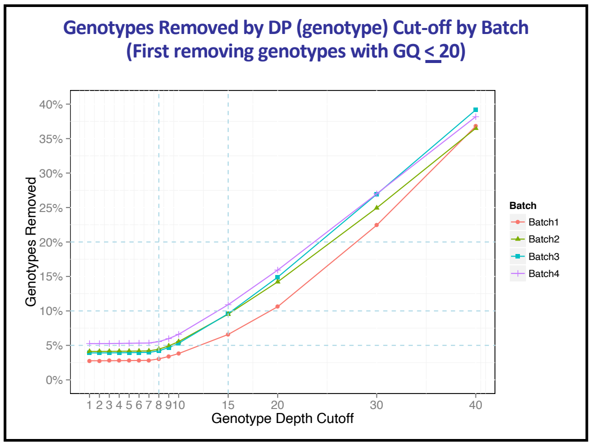
Genotypes Removed by DP (genotype) Cut-off by Batch

62



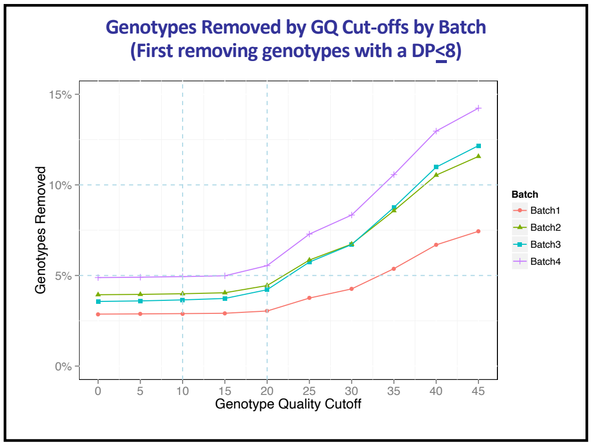
Genotypes Removed by GQ Cut-offs by Batch

63



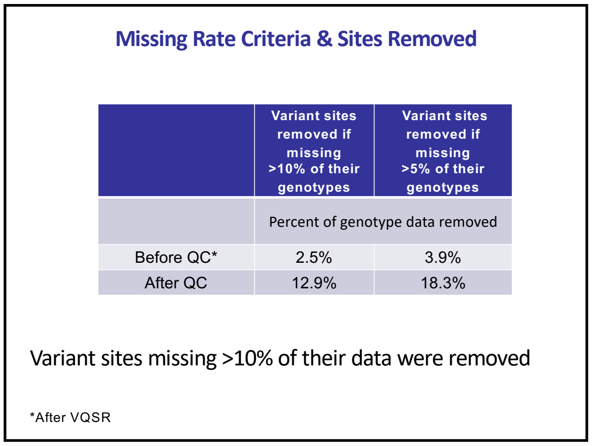
Genotypes Removed by DP (genotype) Cut-off by Batch (First removing genotypes with GQ < 20)

64



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

65



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

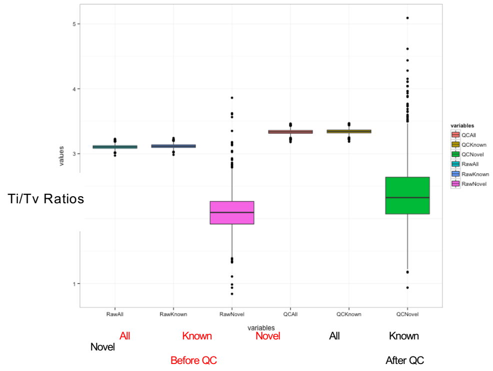
66

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≤5x10^{-8} N=4,414	3.41 ± 0.04	2.39 ± 0.54	3.40 ± 0.04

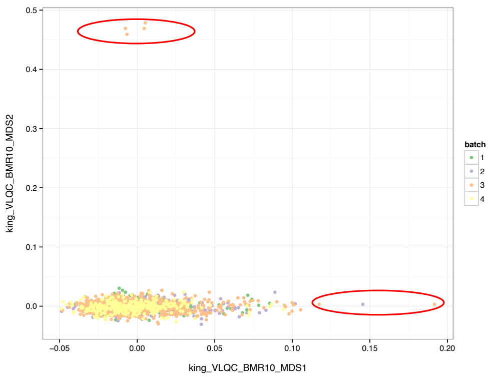
67

Ti/Tv Ratios by Individual Before and After QC



68

MDS First 2 Components After QC



69

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)

70

Convenience Controls

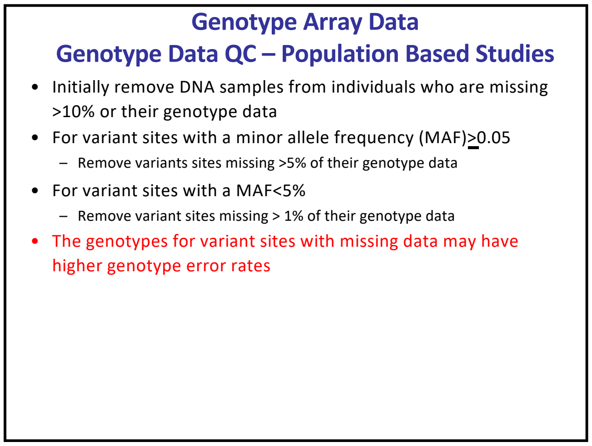
- Can reduce the cost of a study
- Genotype data
- Type I error can be increased
  - Ascertainment from different population
  - Differential genotyping error
    - Even if performed at the same facility
- Proper QC can reduce or remove biases

71

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

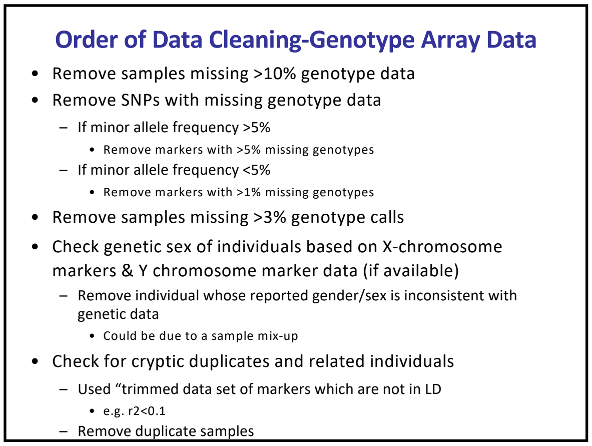
72



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

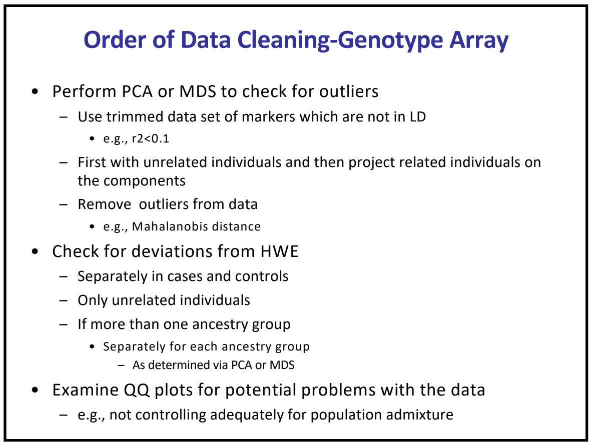
73



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

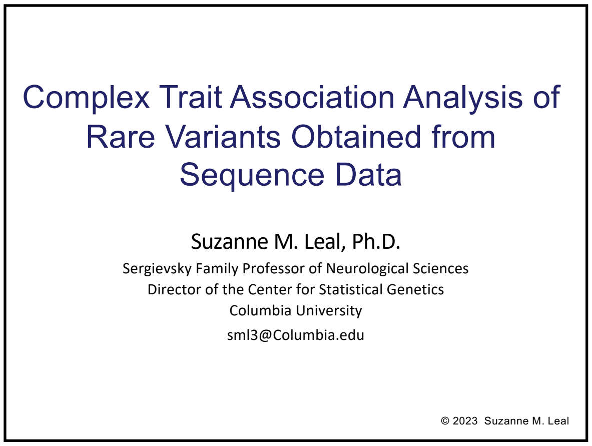
74



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

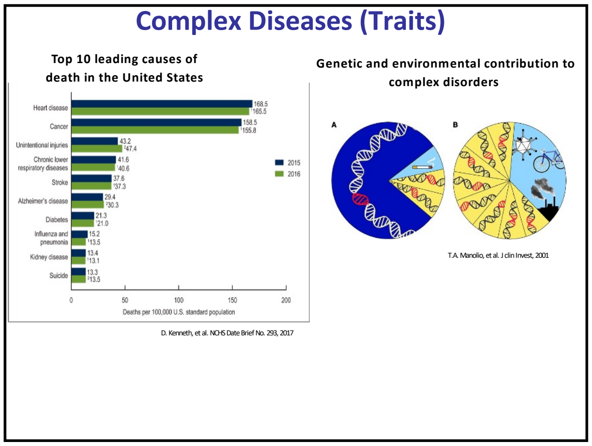
75



Complex Trait Association Analysis of Rare Variants Obtained from Sequence Data

Suzanne M. Leal, Ph.D.
Sergievsky Family Professor of Neurological Sciences
Director of the Center for Statistical Genetics
Columbia University
sml3@Columbia.edu

1



Complex Diseases (Traits)

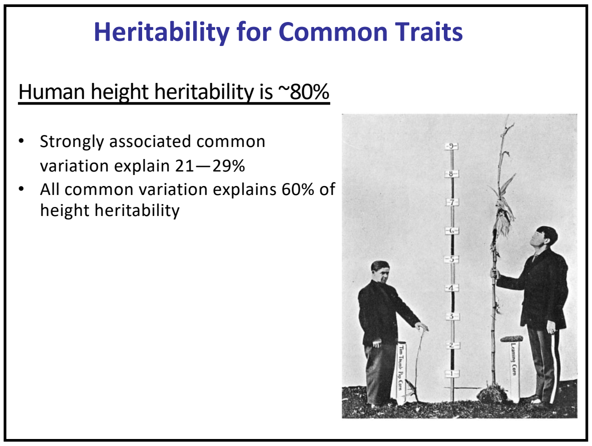
Top 10 leading causes of death in the United States

D. Kenneth, et al. NCHS Date Brief No. 293, 2017

Genetic and environmental contribution to complex disorders

T.A. Manolio, et al. J clin Invest, 2001

2

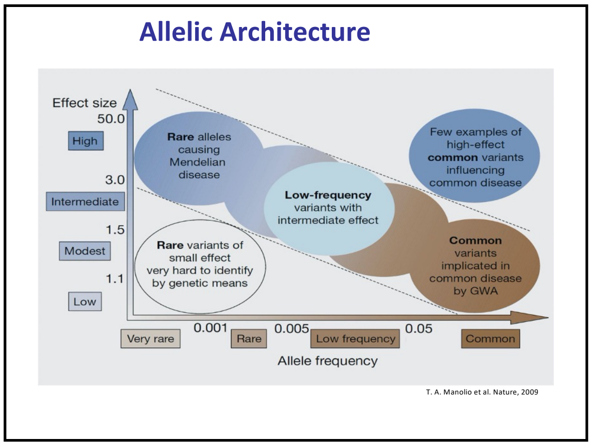


Heritability for Common Traits

Human height heritability is ~80%

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

3

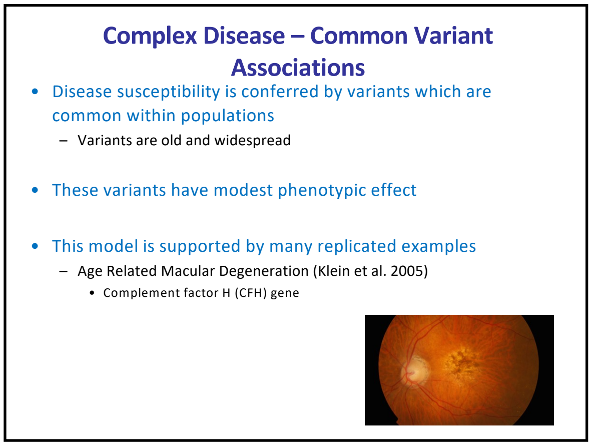


Allelic Architecture

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

T. A. Manolio et al. Nature, 2009

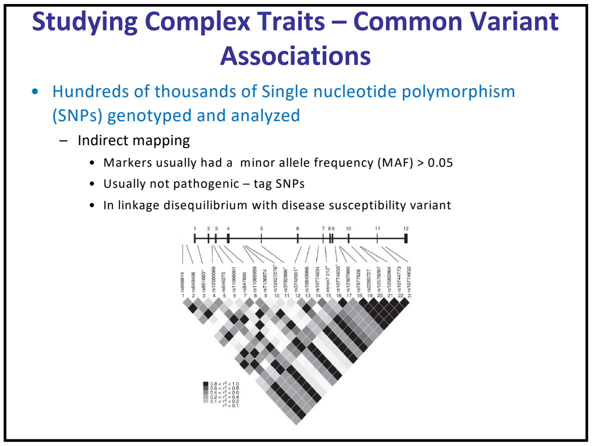
4



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 many replicated examples
  - Age Related Macular Degeneration (Klein et al. 2005)
    - Complement factor H (CFH) gene

5

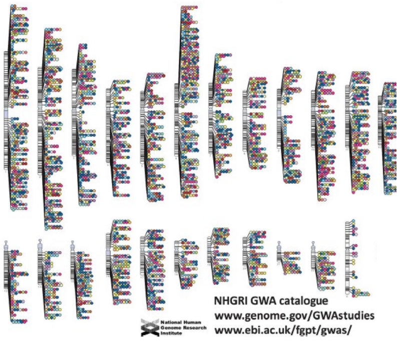


Studying Complex Traits – Common Variant Associations

- 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



NHGRI GWA catalogue
www.genome.gov/GWAstudies
www.ebi.ac.uk/fgpt/gwas/

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

7

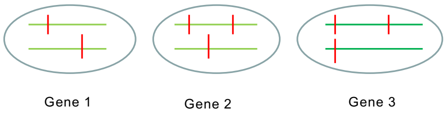
Complex Disease – Rare Variant Associations

- 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

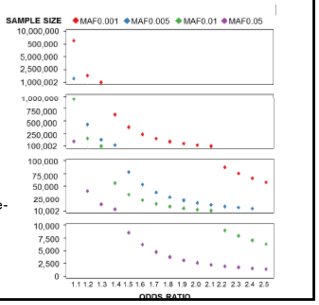
Example

α=5 x 10^{-8}*

Disease prevalence 5%

1-β =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

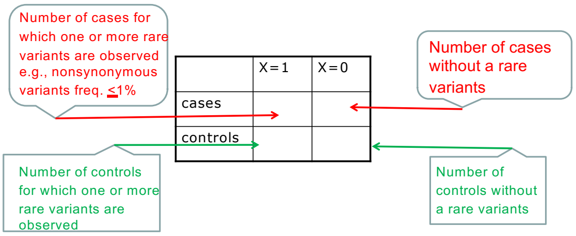
14

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%

Number of cases without a rare variants

	X=1	X=0
cases		
controls		

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

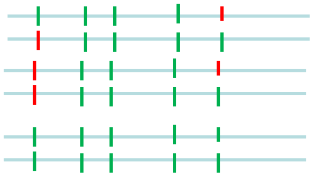
15

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

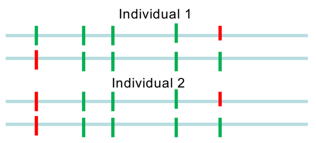
Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- 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

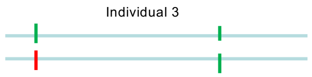
- Example of coding used in regression framework
  - Gene region with 5 variant sites – data available on all sites
  -



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

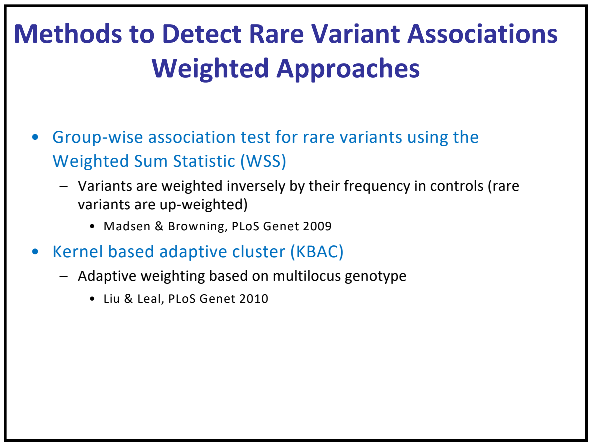


Coded 1/2 (0.5)

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

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

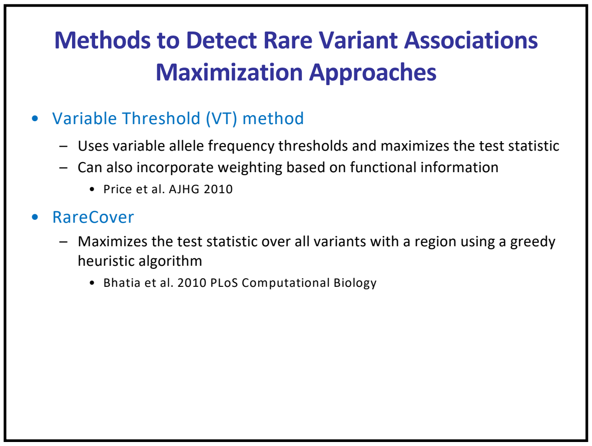
18



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

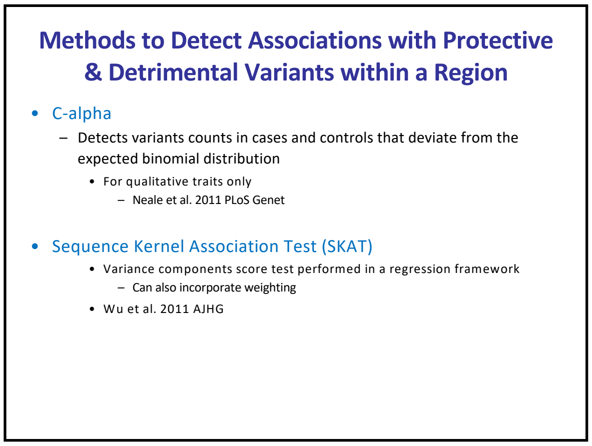
19



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

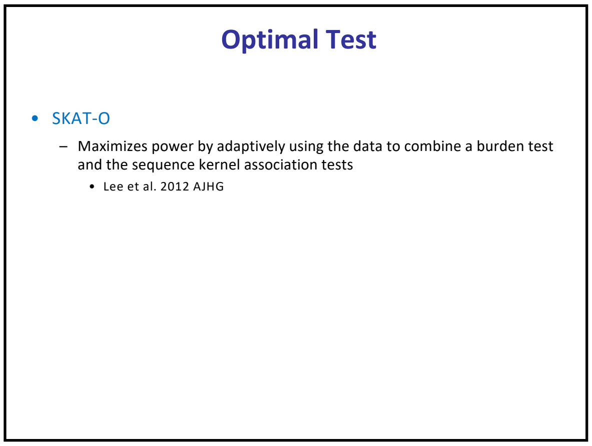
20



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

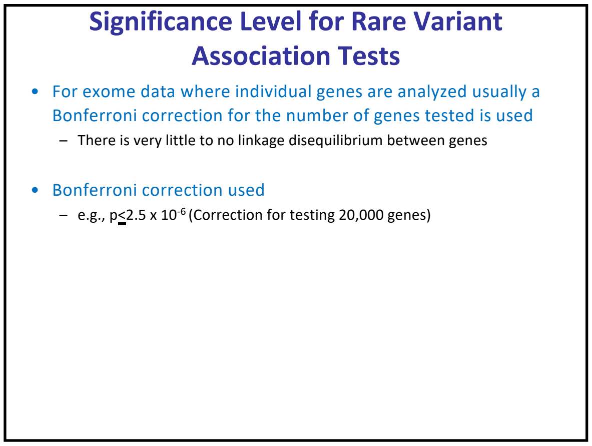
21



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

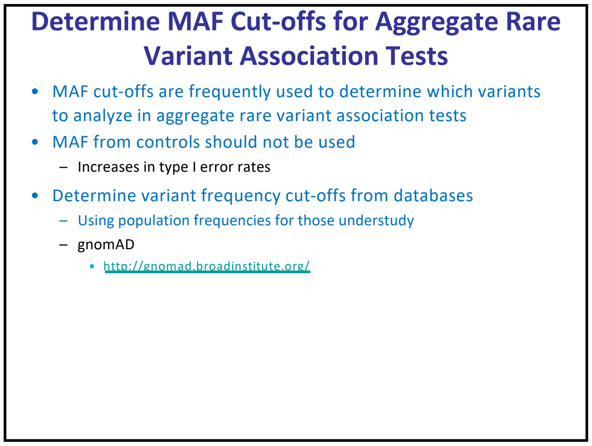
22



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

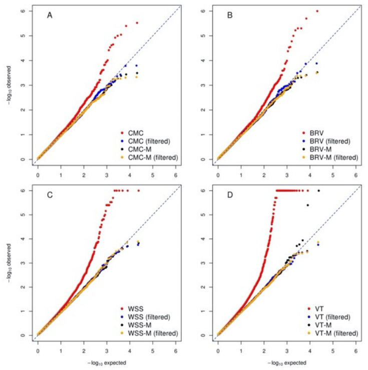
- Genotypes are no longer assigned 0 (1/1), 1 (1/2) or 2 (2/2)
  - Due to uncertainty
- Each genotype is assigned a probability
  - Probabilities sum to 1
- For example
  - Probability of 0 (1/1) genotype is 0.98 and 1 (1/2) genotype is 0.015
- The dosage can be estimated for this example as follows

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

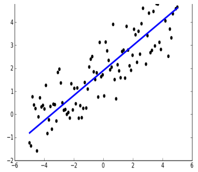
Results



27

Rare Variant Aggregate Methods

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

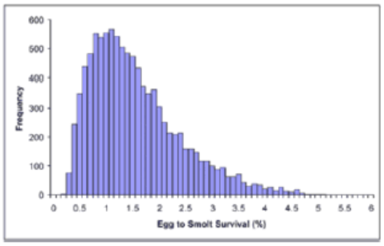


- Almost all rare variant aggregate methods have been extended to be implemented within a regression framework
- Some have also been implemented in a linear mixed model (LMM)/generalized LMM (GLMM)

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



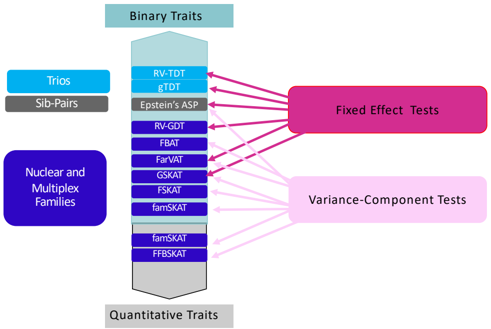
29

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 biobank sized datasets
- 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

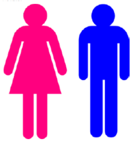
LMM and GLMM: Analysis of Related & Unrelated Individuals

- A few programs which can perform rare variant aggregate analysis
  - REGENIE - Burden test, SKAT, & SKAT-O
  - SMMAT - Burden, SKAT, & SKAT-O
  - rvtests (Zhan 2020) implements BOLT-LMM to perform burden association analysis

35

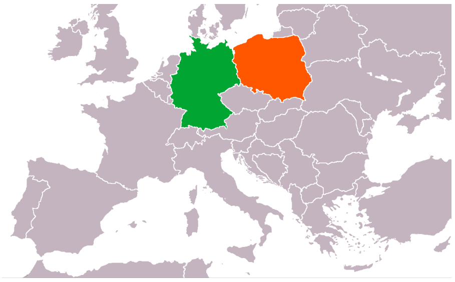
Rare Variant Association Analysis - Confounders

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



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

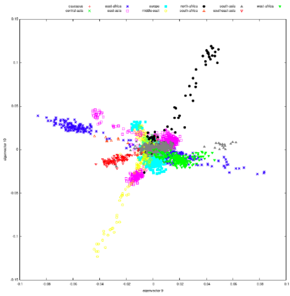
38

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
    - For admixed populations
    - Rare variation



40

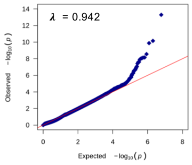
Rare Variant Aggregate Association Analysis

- 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

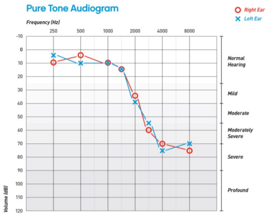
Part II
Example of a Rare Variant Association Study

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

43

Age-related Hearing Loss (ARHL) (aka Presbycusis)

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

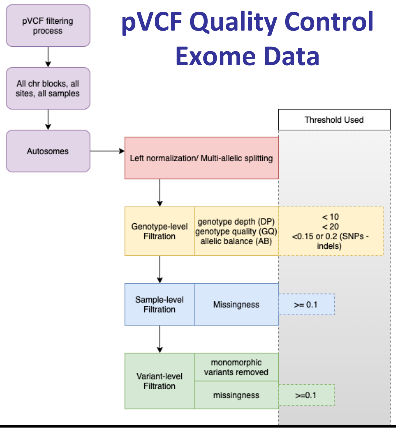
46

Genetic Data Analyzed

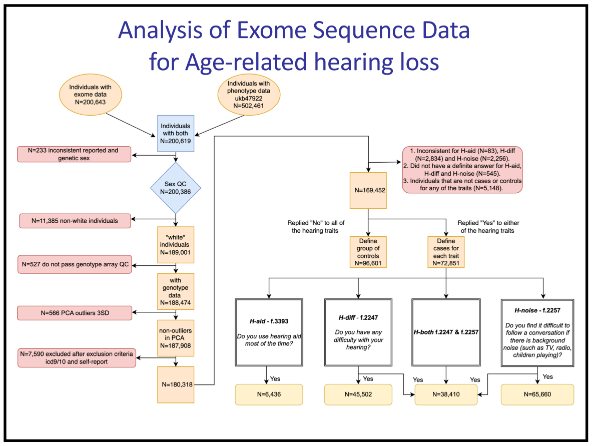
- Exome data
  - ~200,000 participants
- Imputed variant data (secondary replication sample for common variants)
  - ~300,000 participants
    - Did not have exome data at the time of the study

47

pVCF Quality Control Exome Data



48



Analysis of Exome Sequence Data for Age-related hearing loss

Individuals with exome data N=200,643

Individuals with phenotype data ukb47922 N=502,461

Individuals with both N=200,619

N=233 inconsistent reported and genetic sex

Sex QC N=200,386

N=11,385 non-white individuals

"white" individuals N=189,001

N=527 do not pass genotype array QC

with genotype data N=188,474

N=566 PCA outliers 3SD

non-outliers in PCA N=187,908

N=7,590 excluded after exclusion criteria icd9/10 and self-report

N=180,318

N=169,452

1. Inconsistent for H-aid (N=83), H-diff (N=2,834) and H-noise (N=2,256).
2. Did not have a definite answer for H-aid, H-diff and H-noise (N=545).
3. Individuals that are not cases or controls for any of the traits (N=5,148).

Replied "No" to all of the hearing traits

Define group of controls N=96,601

Replied "Yes" to either of the hearing traits

Define cases for each trait N=72,851

***H-aid* - f.3393**

Do you use hearing aid most of the time?

Yes

N=6,436

***H-diff* - f.2247**

Do you have any difficulty with your hearing?

Yes

N=45,502

***H-both* f.2247 & f.2257**

Yes

Yes

N=38,410

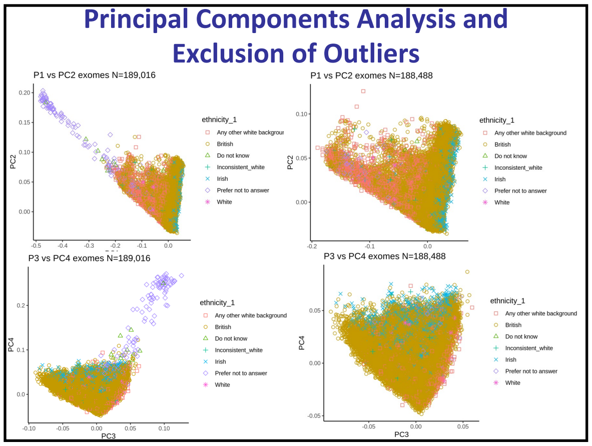
***H-noise* - f.2257**

Do you find it difficult to follow a conversation if there is background noise (such as TV, radio, children playing)?

Yes

N=65,660

49



Principal Components Analysis and Exclusion of Outliers

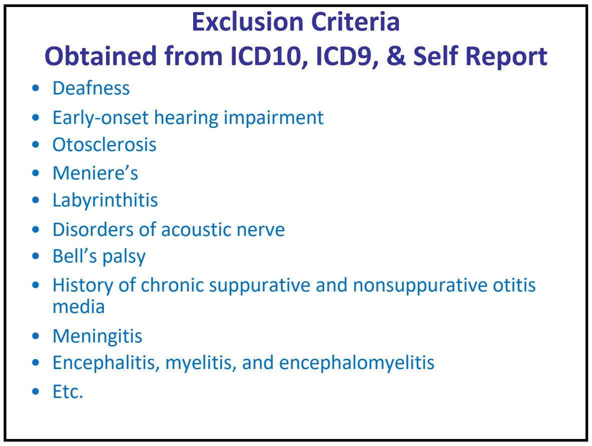
P1 vs PC2 exomes N=189,016

P1 vs PC2 exomes N=188,488

P3 vs PC4 exomes N=189,016

P3 vs PC4 exomes N=188,488

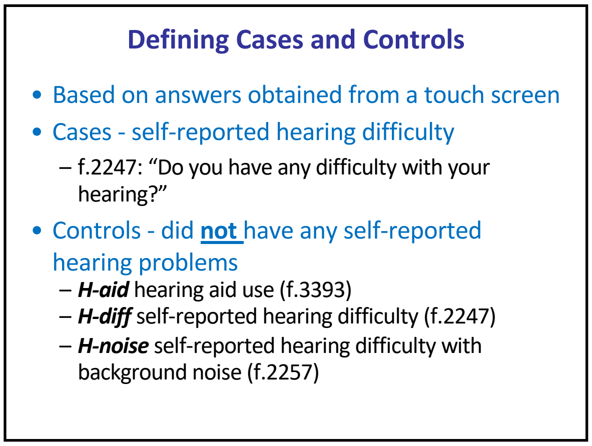
50



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.

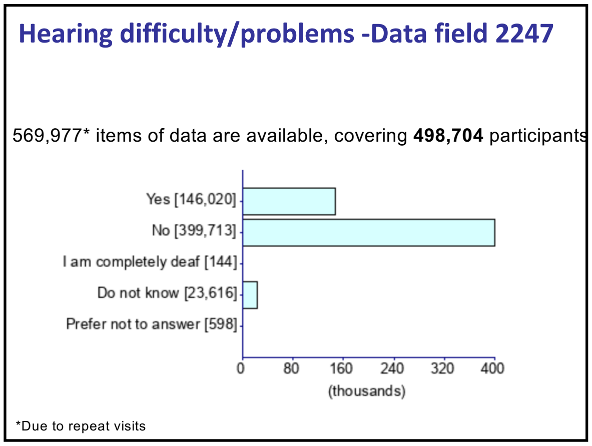
51



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)

52

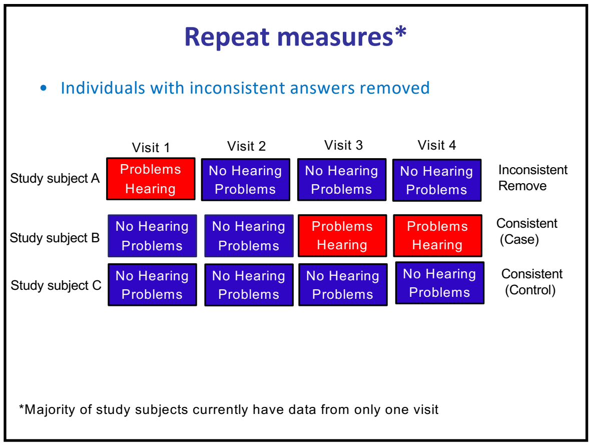


Hearing difficulty/problems -Data field 2247

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

*Due to repeat visits

53



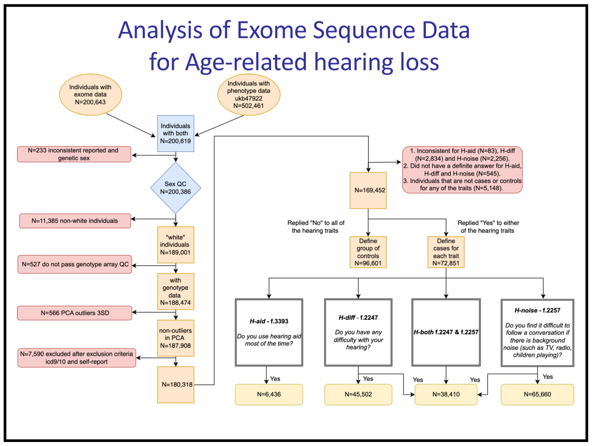
Repeat measures*

- 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

54



Analysis of Exome Sequence Data for Age-related hearing loss

Individuals with exome data N=200,643

Individuals with phenotype data ukb47922 N=502,461

Individuals with both N=200,619

N=233 inconsistent reported and genetic sex

Sex QC N=200,386

N=11,385 non-white individuals

"white" individuals N=189,001

N=527 do not pass genotype array QC

with genotype data N=188,474

N=566 PCA outliers 3SD

non-outliers in PCA N=187,908

N=7,590 excluded after exclusion criteria icd9/10 and self-report

N=180,318

N=169,452

1. Inconsistent for H-aid (N=83), H-diff (N=2,834) and H-noise (N=2,256).
2. Did not have a definite answer for H-aid, H-diff and H-noise (N=545).
3. Individuals that are not cases or controls for any of the traits (N=5,148).

Replied "No" to all of the hearing traits

Define group of controls N=96,601

Replied "Yes" to either of the hearing traits

Define cases for each trait N=72,851

***H-aid* - f.3393**
Do you use hearing aid most of the time?

***H-diff* - f.2247**
Do you have any difficulty with your hearing?

***H-both* f.2247 & f.2257**

***H-noise* - f.2257**
Do you find it difficult to follow a conversation if there is background noise (such as TV, radio, children playing)?

Yes

N=6,436

Yes

N=45,502

Yes

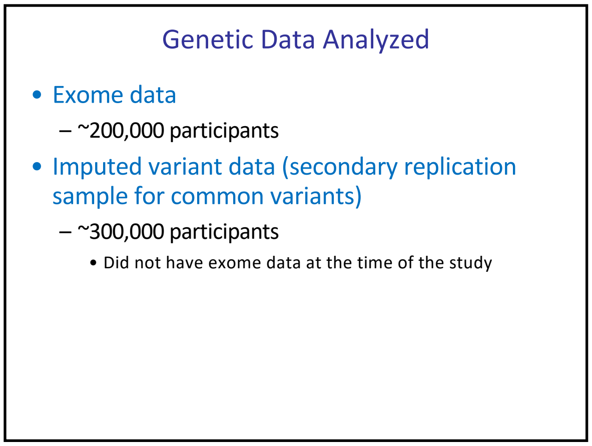
Yes

N=38,410

Yes

N=65,660

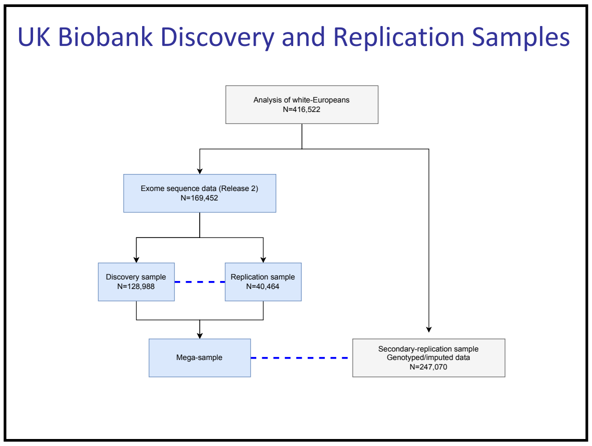
55



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

56



UK Biobank Discovery and Replication Samples

Analysis of white-Europeans N=416,522

Exome sequence data (Release 2) N=169,452

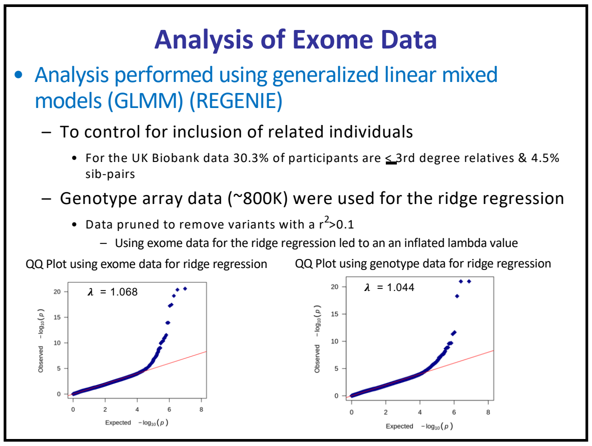
Discovery sample N=128,988

Replication sample N=40,464

Mega-sample

Secondary-replication sample Genotyped/imputed data N=247,070

57



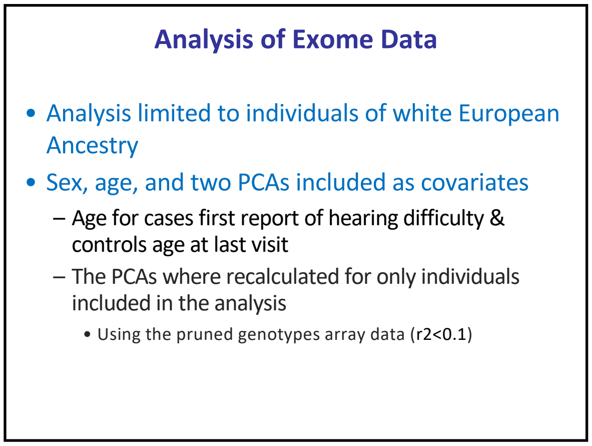
Analysis of Exome Data

- 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

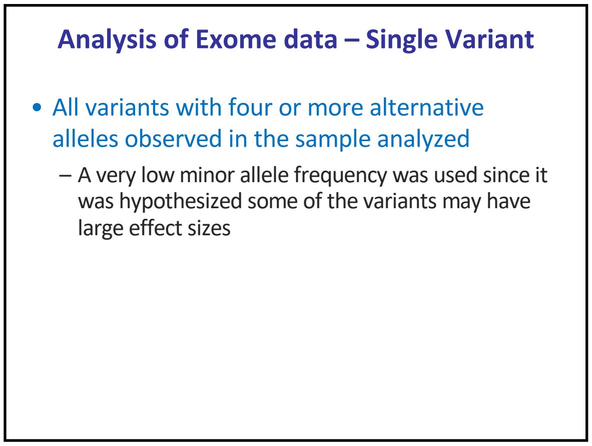
58



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)

59



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

60

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

61

Significance Levels

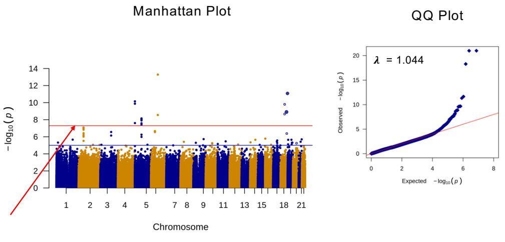
- Discovery sample
  - A genome-wide significance level was used to reject the null hypothesis of no association
    - p≤5.0x10^{-8}
- Replication sample
  - Permutation was used to obtain empirical p-values
    - Adjusting for the phenotypes and variants brought to replication
      - p≤0.05

For the replication it is not necessary to use a genome-wide significance level of 5 x 10^{-8} for single variant tests or 2.5 x 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

62

Hearing Difficulty - Data Field 2247


Manhattan Plot

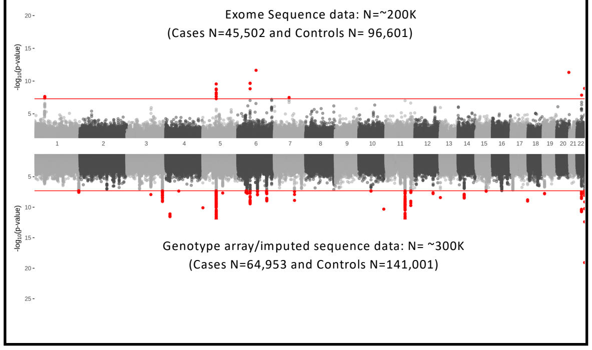
QQ Plot

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

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

63

Hudson Plot Discovery and Replication Hearing Difficulty Data Field 2247


Exome Sequence data: N=~200K
(Cases N=45,502 and Controls N= 96,601)

Genotype array/imputed sequence data: N= ~300K
(Cases N=64,953 and Controls N=141,001)

64

Analysis of the Discovery Sample & Replication Single Variant Analysis

Discovery sample single-variant associations analysis for age-related hearing loss traits

CHR	SNP	EA	EAF	Gene	H-aid			H-diff			H-noise			H-both		
					Beta(OR)	SE	P	Beta(OR)	SE	P	Beta(OR)	SE	P	Beta(OR)	SE	P
5	rs537688122	G	6.65x10^{-4}	*PDCD6*	1.99(7.3)	0.29	2.25x10^{-10}	1.32(3.7)	0.17	*1.12x10^{-15}*	1.04(2.8)	0.16	*5.50x10^{-11}*	1.27(3.6)	0.18	*1.02x10^{-12}*
5	rs549592074	C	5.58x10^{-4}	*PDCD6*	1.99(7.3)	0.32	1.95x10^{-8}	1.35(3.9)	0.18	*7.05x10^{-14}*	1.07(2.9)	0.18	*6.69x10^{-10}*	1.28(3.6)	0.19	*5.52x10^{-11}*
5	rs571370281	G	7.04x10^{-4}	*PDCD6*	1.92(6.8)	0.28	6.02x10^{-10}	1.33(3.8)	0.16	*1.14x10^{-16}*	1.03(2.8)	0.16	*2.26x10^{-11}*	1.29(3.6)	0.17	*9.66x10^{-14}*
6	rs1574430	C	6.09x10^{-1}	*SLC22A7*										0.06(1.1)	0.01	*2.77x10^{-8}*
6	rs2242416	G	6.09x10^{-1}	*CRIP3*										0.06(1.1)	0.01	*2.60x10^{-8}*
6	rs121912560	G	7.63x10^{-5}	*MYO6*	5.48(239.8)	1.12	1.79x10^{-10}	3.54 (34.5)	0.90	*3.41x10^{-9}*				3.73(41.7)	0.90	*3.76x10^{-10}*

Genome wide-significant variants ($p < 5x10^{-8}$) with hearing aid (H-aid), hearing difficulty (H-diff), hearing difficulty with background noise (H-noise) and the combined hearing trait (H-both) in the analysis of the discovery sample of White-European individuals from the UK Biobank. The p-values for replicated associations (empirical p-values <0.05 adjusting for variants and traits brought to replication) are shown in red; CHR - chromosome; EA - effect allele, EAF - effect allele frequency, OR - odds ratio, SE - standard error, P - p-value

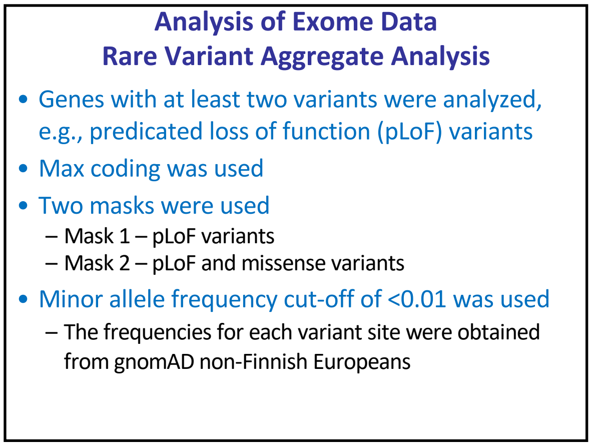
65

Mega analysis single variant associations analysis with age-related hearing impairment traits

CHR	SNP	EA	EAF	Gene	H-aid			H-diff			H-noise			H-both		
					Beta(OR)	SE	P	Beta(OR)	SE	P	Beta(OR)	SE	P	Beta(OR)	SE	P
1	rs11589562	C	0.424	*MAST2*				-0.05(0.95)	0.01	*2.25x10^{-8}*						
1	rs2275426	A	0.431	*MAST2*				-0.05(0.95)	0.01	*3.39x10^{-8}*						
1	rs1707336	G	0.435	*MAST2*				-0.05(0.95)	0.01	*3.63x10^{-8}*						
1	rs1707304	A	0.436	*PIK3R3*				-0.05(0.95)	0.01	*2.34x10^{-8}*				-0.05(0.95)	0.01	*3.30x10^{-8}*
5	rs537688122^{a}	G	7x10^{-4}	*PDCD6*	1.79(6.0)	0.25	7.06x10^{-11}	1.35(3.9)	0.14	1.04x10^{-21}	1.1(3.0)	0.14	4.96x10^{-16}	1.32(3.8)	0.15	1.11x10^{-18}
5	rs549592074^{a}	C	6x10^{-4}	*PDCD6*	1.70(5.5)	0.28	2.48x10^{-8}	1.37(3.9)	0.16	5.19x10^{-19}	1.08(3.0)	0.15	2.19x10^{-13}	1.32(3.8)	0.16	6.63x10^{-16}
5	rs571370281	G	7x10^{-4}	*PDCD6*	1.71(5.5)	0.24	1.34x10^{-10}	1.31(3.7)	0.14	1.00x10^{-21}	1.04(2.8)	0.14	1.83x10^{-15}	1.28(3.6)	0.15	8.00x10^{-19}
5	rs7714670	C	0.467	*ARHGEF28*	0.11(1.1)	0.02	*9.99x10^{-9}*	0.05(1.05)	0.01	*1.63x10^{-9}*				0.05(1.05)	0.01	*1.06x10^{-8}*
5	rs11949860	A	0.462	*ARHGEF28*	0.11(1.1)	0.02	*3.87x10^{-8}*	0.05(1.05)	0.01	*9.92x10^{-9}*						
5	rs35525194	G	0.471	*ARHGEF28*	0.11(1.1)	0.02	*7.03x10^{-9}*	0.05(1.05)	0.01	*2.19x10^{-9}*				0.05(1.05)	0.01	*1.21x10^{-8}*
5	rs6453022	A	0.501	*ARHGEF28*	0.11(1.1)	0.02	*7.30x10^{-9}*	0.05(1.05)	0.01	*2.75x10^{-10}*				0.06(1.06)	0.01	*4.13x10^{-10}*
5	rs7716253	C	0.524	*ARHGEF28*	0.11(1.1)	0.02	*8.82x10^{-9}*	0.05(1.05)	0.01	*6.29x10^{-9}*				0.05(1.05)	0.01	*2.19x10^{-8}*
5	rs2973549	A	0.478	*ARHGEF28*	0.11(1.1)	0.02	*1.23x10^{-8}*	0.05(1.05)	0.01	*2.22x10^{-8}*						
5	rs2973548	T	0.478	*ARHGEF28*	0.11(1.1)	0.02	*2.61x10^{-8}*	0.05(1.05)	0.01	*4.90x10^{-8}*						
6	rs146694394	T	0.005	*SYNJ2*										0.33(1.4)	0.06	*1.72x10^{-8}*
6	rs1574430	C	0.608	*SLC22A7*				0.05(1.05)	0.01	2.10x10^{-10}	0.05(1.05)	0.01	4.2x10-^{10}	0.06(1.06)	0.01	8.06x10^{-12}
6	rs2242416	G	0.606	*CRIP3*				0.05(1.05)	0.01	*2.25x10^{-10}*	0.05(1.05)	0.01	*3.8x10^{-10}*	0.06(1.06)	0.01	*8.13x10^{-12}*
6	rs2254303	A	0.606	*CRIP3*				0.05(1.05)	0.01	*1.49x10^{-9}*	0.05(1.05)	0.01	*1.90x10^{-9}*	0.06(1.06)	0.01	*3.88x10^{-11}*
6	rs765264064^{a}	C	6x10^{-5}	*FILIP1*	3.01(20.3)	0.48	2.81x10^{-9}									
6	rs121912560^{a}	G	0.005	*MYO6*	5.28(196.3)	0.98	5.15x10^{-14}	3.73(41.9)	0.87	2.26x10^{-12}	3.26(26.1)	0.86	1.09x10^{-9}	3.86(47.7)	0.88	8.72x10^{-13}
7	rs2286276	T	0.284	*TBL2*										0.05(1.05)	0.01	4.66x10^{-8}
7	rs61010704	G	0.283	*MLXIPL*				0.05(1.05)	0.01	3.16x10^{-8}				0.05(1.05)	0.01	2.72x10^{-8}
22	rs371997714	G	0.293	*BAIAP2L2*				0.05(1.05)	0.01	*1.40x10^{-8}*						
22	rs36062310	A	0.043	*KLHDC7B*				0.12(1.1)	0.02	*1.32x10^{-9}*				0.12(1.3)	0.02	*6.66x10^{-9}*

Genome-wide significant variants (p < 5x10-8) with hearing aid (H-aid), hearing difficulty (H-diff), hearing difficulty with background noise (H-noise) and the combined hearing trait (H-both) in the analysis of the mega-sample of White-European individuals from the UK Biobank. The p-values for replicated associations (empirical p-values <0.05 adjusting for variants and traits brought to replication) are shown in red. ^{a}Variant not found present in the replication sample; CHR -chromosome; EA - effect allele, EAF - effect allele frequency, OR – odds

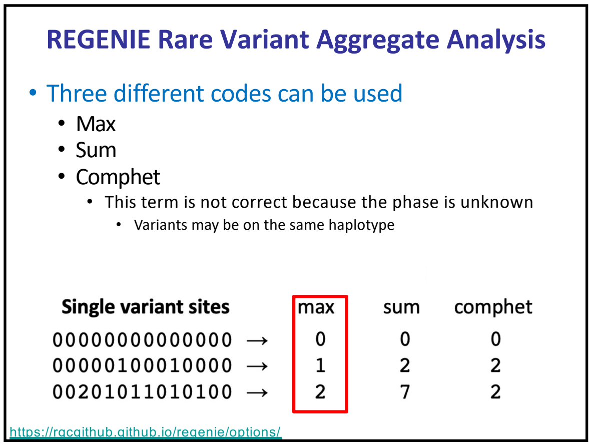
66



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

67



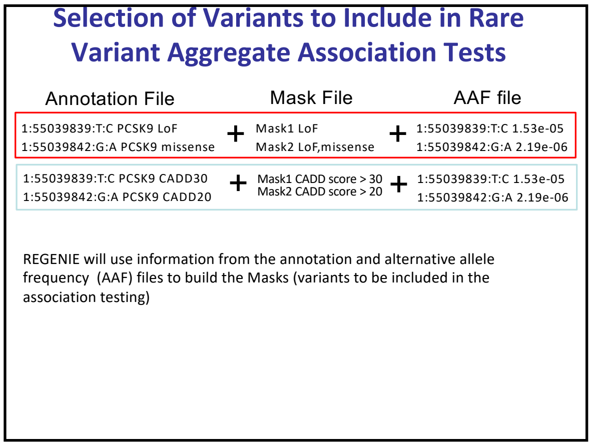
REGENIE Rare Variant Aggregate Analysis

- 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/

68

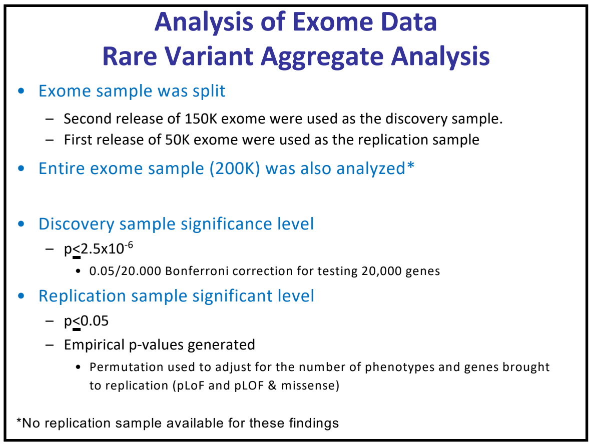


Selection of Variants to Include in Rare Variant Aggregate Association Tests

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)

69

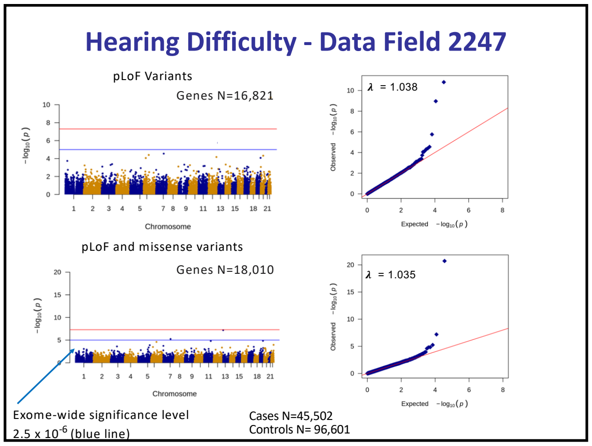


Analysis of Exome Data
Rare Variant Aggregate Analysis

- 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
  - $p \leq 2.5\text{x}10^{-6}$
    - 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

70



Hearing Difficulty - Data Field 2247

pLoF Variants

Genes N=16,821

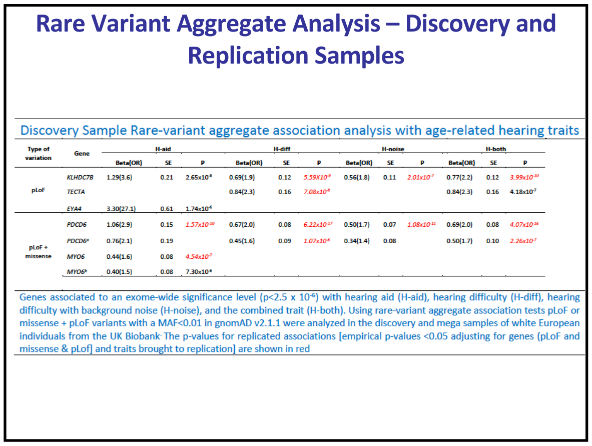
pLoF and missense variants

Genes N=18,010

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

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

71



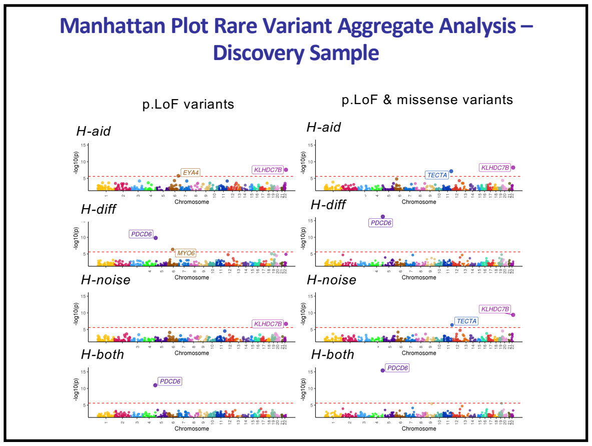
Rare Variant Aggregate Analysis – Discovery and Replication Samples

Discovery Sample Rare-variant aggregate association analysis with age-related hearing traits

Type of variation	Gene	H-aid			H-diff			H-noise			H-both		
		Beta(OR)	SE	P	Beta(OR)	SE	P	Beta(OR)	SE	P	Beta(OR)	SE	P
pLoF	*KLHDC7B*	1.29(3.6)	0.21	2.65x10^{-8}	0.69(1.9)	0.12	*5.59X10^{-9}*	0.56(1.8)	0.11	*2.01x10^{-7}*	0.77(2.2)	0.12	*3.99x10^{-10}*
	TECTA				0.84(2.3)	0.16	*7.08x10^{-8}*				0.84(2.3)	0.16	4.18x10^{-7}
	EYA4	3.30(27.1)	0.61	1.74x10^{-6}									
pLoF + missense	*PDCD6*	1.06(2.9)	0.15	*1.57x10^{-10}*	0.67(2.0)	0.08	*6.22x10^{-17}*	0.50(1.7)	0.07	*1.08x10^{-11}*	0.69(2.0)	0.08	*4.07x10^{-16}*
	PDCD6[a]	0.76(2.1)	0.19		0.45(1.6)	0.09	*1.07x10^{-6}*	0.34(1.4)	0.08		0.50(1.7)	0.10	*2.26x10^{-7}*
	MYO6	0.44(1.6)	0.08	*4.54x10^{-7}*									
	MYO6[b]	0.40(1.5)	0.08	7.30x10^{-6}									

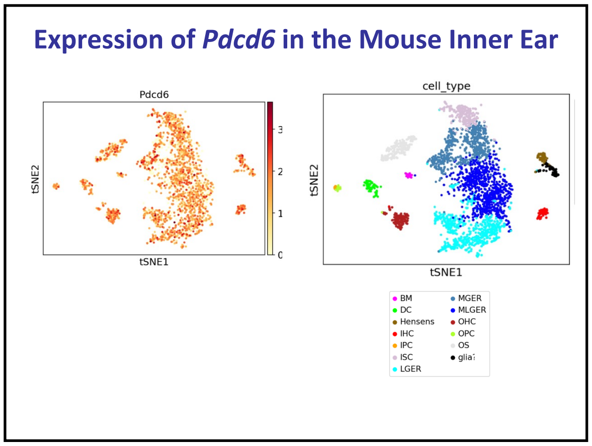
Genes associated to an exome-wide significance level ($p<2.5 \times 10^{-6}$) with hearing aid (H-aid), hearing difficulty (H-diff), hearing difficulty with background noise (H-noise), and the combined trait (H-both). Using rare-variant aggregate association tests pLoF or missense + pLoF variants with a MAF<0.01 in gnomAD v2.1.1 were analyzed in the discovery and mega samples of white European individuals from the UK Biobank. The p-values for replicated associations [empirical p-values <0.05 adjusting for genes (pLoF and missense & pLof] and traits brought to replication] are shown in red

72



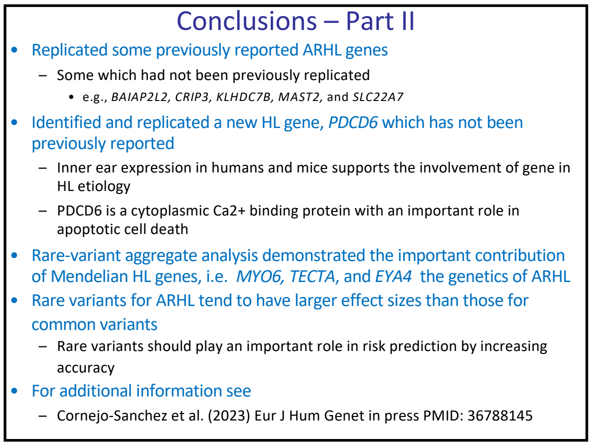
Manhattan Plot Rare Variant Aggregate Analysis – Discovery Sample

73



Expression of *Pdcd6* in the Mouse Inner Ear

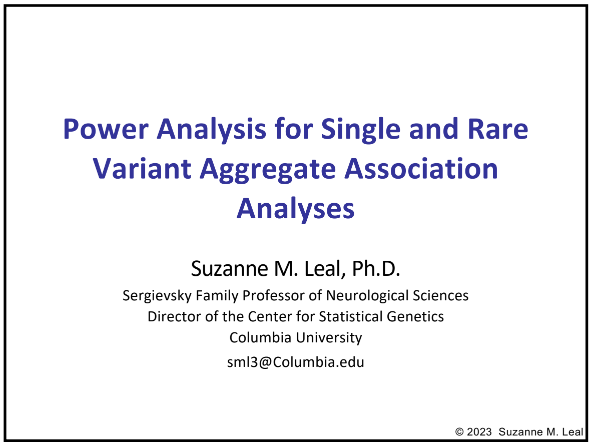
74



Conclusions – Part II

- 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, *PDCD6* which has not been previously reported
  - Inner ear expression in humans and mice supports the involvement of gene in HL etiology
  - PDCD6 is a cytoplasmic Ca2+ binding protein with an important role in apoptotic cell death
- Rare-variant aggregate analysis demonstrated the important contribution of Mendelian HL genes, i.e. *MYO6, TECTA*, and *EYA4* the genetics of ARHL
- 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
- For additional information see
  - Cornejo-Sanchez et al. (2023) Eur J Hum Genet in press PMID: 36788145

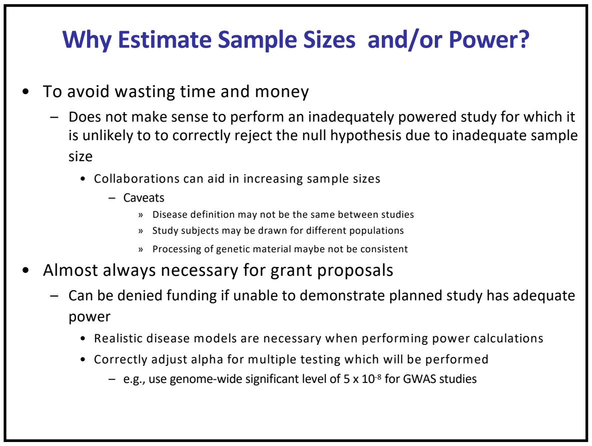
75



Power Analysis for Single and Rare Variant Aggregate Association Analyses

Suzanne M. Leal, Ph.D.
Sergievsky Family Professor of Neurological Sciences
Director of the Center for Statistical Genetics
Columbia University
sml3@Columbia.edu

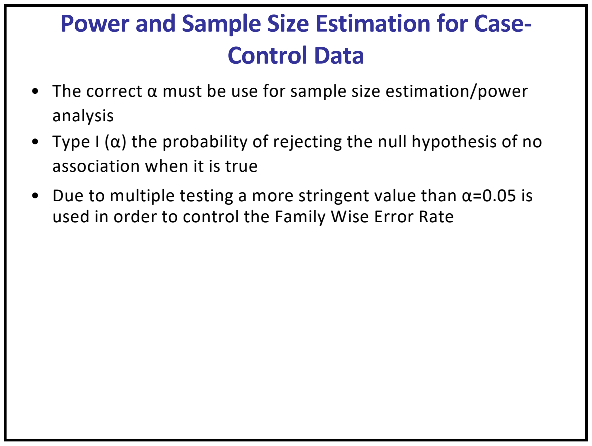
1



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 x 10^{-8} for GWAS studies

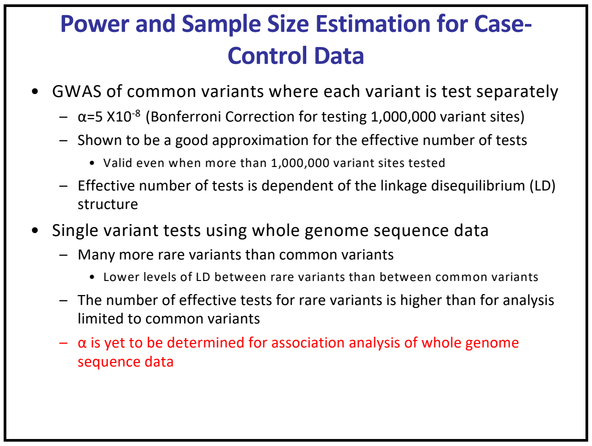
2



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

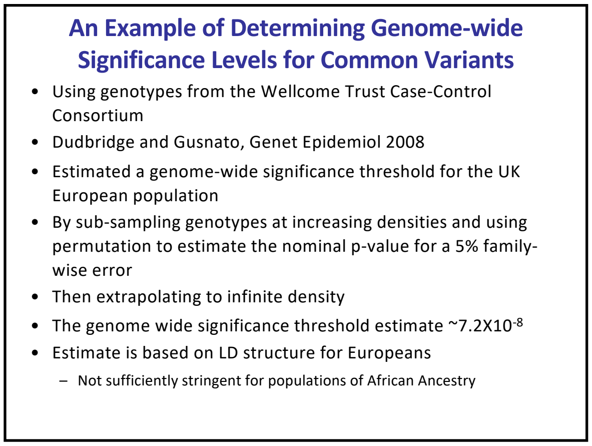
3



Power and Sample Size Estimation for Case-Control Data

- GWAS of common variants where each variant is test separately
  - α=5 X10^{-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

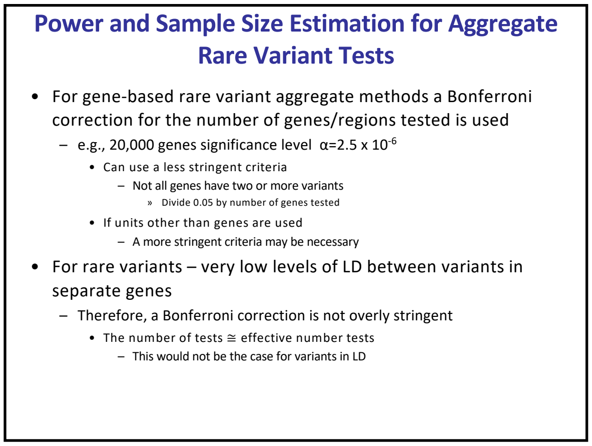
4



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 ~7.2X10^{-8}
- Estimate is based on LD structure for Europeans
  - Not sufficiently stringent for populations of African Ancestry

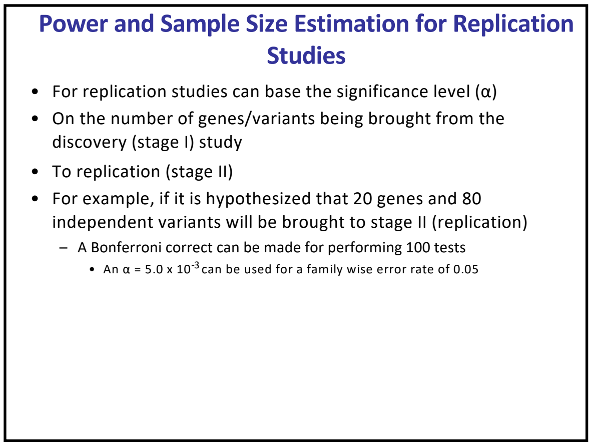
5



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 α=2.5 x 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

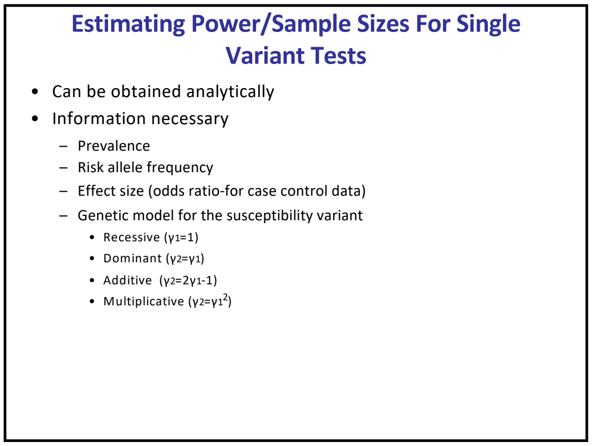
6



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

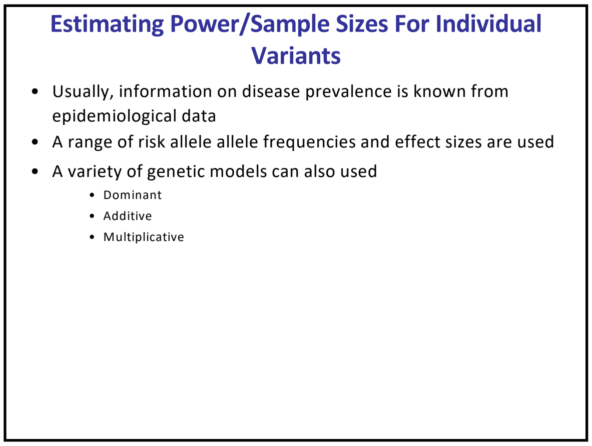
7



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 (γ_1=1)
    - Dominant ($\gamma_2=\gamma_1$)
    - Additive ($\gamma_2=2\gamma_1-1$)
    - Multiplicative ($\gamma_2=\gamma_1^2$)

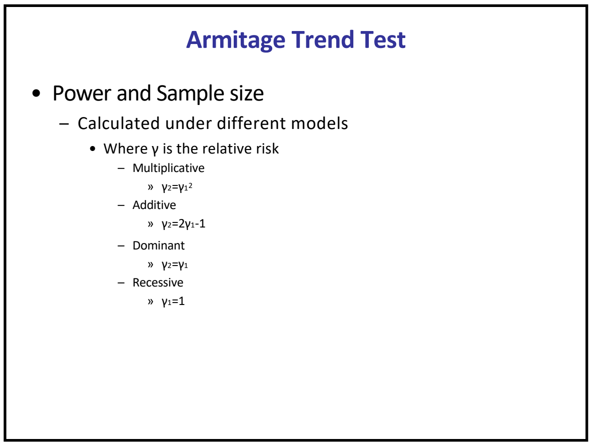
8



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

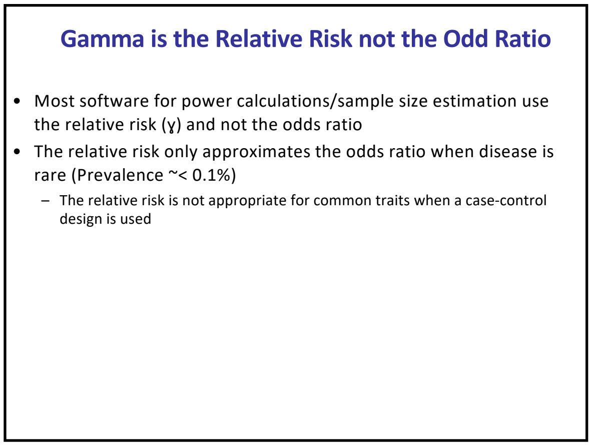
9



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
        - » $\gamma_1=1$

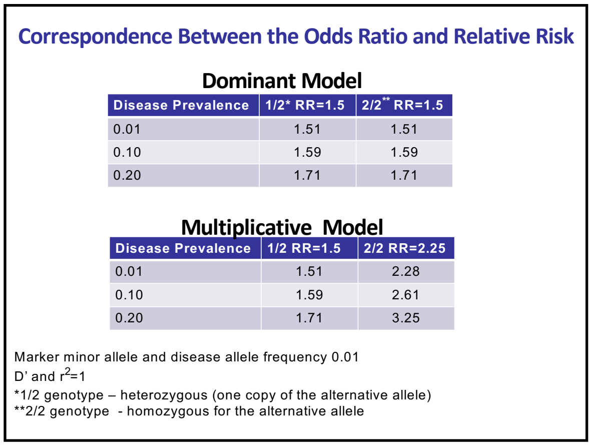
10



Gamma is the Relative Risk not the Odd Ratio

- Most software for power calculations/sample size estimation use the relative risk (ɣ) 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

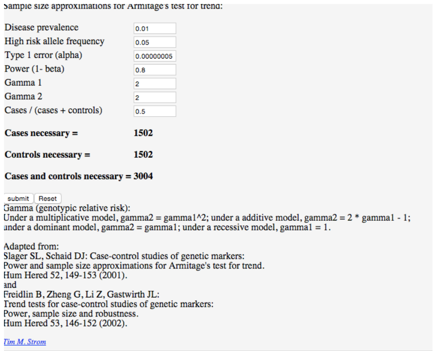
12

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



Sample size approximations for Armitage's test for trend:

Disease prevalence	0.01
High risk allele frequency	0.05
Type 1 error (alpha)	0.00000005
Power (1- beta)	0.8
Gamma 1	2
Gamma 2	2
Cases / (cases + controls)	0.5

Cases necessary = 1502

Controls necessary = 1502

Cases and controls necessary = 3004

Gamma (genotypic relative risk):
Under a multiplicative model, gamma2 = gamma1^2; under a additive model, gamma2 = 2 * gamma1 - 1;
under a dominant model, gamma2 = gamma1; under a recessive model, gamma1 = 1.

Adapted from:
Slager SL, Schaid DJ: Case-control studies of genetic markers:
Power and sample size approximations for Armitage's test for trend.
Hum Hered 52, 149-153 (2001).
and
Freidlin B, Zheng G, Li Z, Gastwirth JL:
Trend tests for case-control studies of genetic markers:
Power, sample size and robustness.
Hum Hered 53, 146-152 (2002).

Tim M. Strom

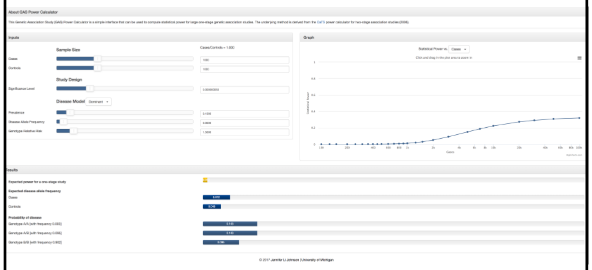
14

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



About GAS Power Calculator

This Genetic Association Study (GAS) Power Calculator is a simple interface that can be used to compute statistical power for large one-stage genetic association studies. The underlying method is derived from the CaTS power calculator for two-stage association studies (2006).

Inputs

Sample Size

Cases/Controls = 1.000

Cases	1000
Controls	1000

Study Design

Significance Level	0.00000005

Disease Model: Dominant

Prevalence	0.1000
Disease Allele Frequency	0.0500
Genotype Relative Risk	1.5000

Graph

Statistical Power vs. Cases

Results

Expected power for a one-stage study	0.0
Expected disease allele frequency	
Cases	0.072
Controls	0.048
Probability of disease	
Genotype A/A (with frequency 0.003)	0.143
Genotype A/B (with frequency 0.095)	0.143
Genotype B/B (with frequency 0.902)	0.095

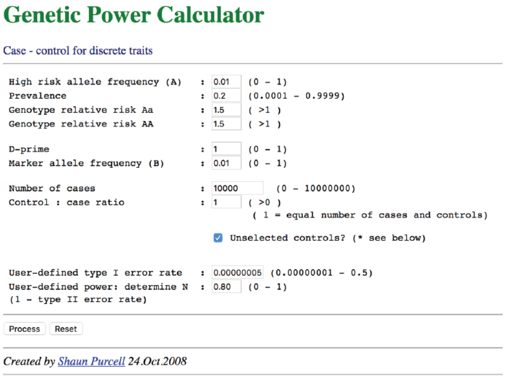
© 2017 Jennifer Li Johnson | University of Michigan

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



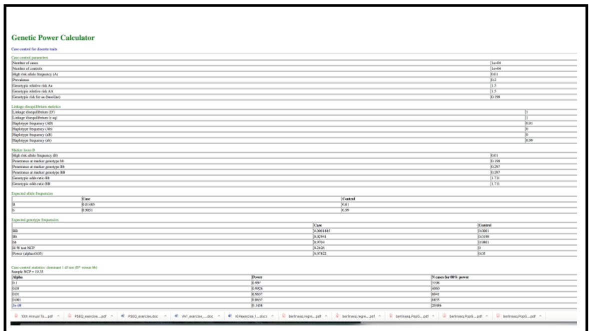
Genetic Power Calculator

Case - control for discrete traits

```
High risk allele frequency (A)   :  0.01   (0 - 1)
Prevalence                       :  0.2    (0.0001 - 0.9999)
Genotype relative risk Aa        :  1.5    ( >1 )
Genotype relative risk AA        :  1.5    ( >1 )

D-prime                          :  1      (0 - 1)
Marker allele frequency (B)      :  0.01   (0 - 1)

Number of cases                  :  10000      (0 - 10000000)
Control : case ratio             :  1      ( >0 )
                                     ( 1 = equal number of cases and controls)

                           ☑ Unselected controls? (* see below)

User-defined type I error rate   :  0.00000005  (0.00000001 - 0.5)
User-defined power: determine N  :  0.80   (0 - 1)
(1 - type II error rate)
```

Created by Shaun Purcell 24.Oct.2008

18



Genetic Power Calculator

Case-control for discrete traits

Case-control parameters

Number of cases	1e+04
Number of controls	1e+04
High risk allele frequency (A)	0.01
Prevalence	0.2
Genotypic relative risk Aa	1.5
Genotypic relative risk AA	1.5
Genotypic risk for aa (baseline)	0.198

Linkage disequilibrium statistics

Linkage disequilibrium (D')	1
Linkage disequilibrium (r-sq)	1
Haplotype frequency (AB)	0.01
Haplotype frequency (Ab)	0
Haplotype frequency (aB)	0
Haplotype frequency (ab)	0.99

Marker locus B

High risk allele frequency (B)	0.01
Penetrance at marker genotype bb	0.198
Penetrance at marker genotype Bb	0.297
Penetrance at marker genotype BB	0.297
Genotypic odds ratio Bb	1.711
Genotypic odds ratio BB	1.711

Expected allele frequencies

	Case	Control
B	0.01485	0.01
b	0.9851	0.99

Expected genotype frequencies

	Case	Control
BB	0.0001485	0.0001
Bb	0.02941	0.0198
bb	0.9704	0.9801
H-W test NCP	0.2426	0
Power (alpha=0.05)	0.07822	0.05

Case-control statistics: dominant 1 df test (B* versus bb)

Sample NCP = 19.33

Alpha	Power	N cases for 80% power
0.1	0.997	3198
0.05	0.9926	4060
0.01	0.9657	6041
0.001	0.8657	8833
5e-08	0.1458	20486

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., 2.5x10^{-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

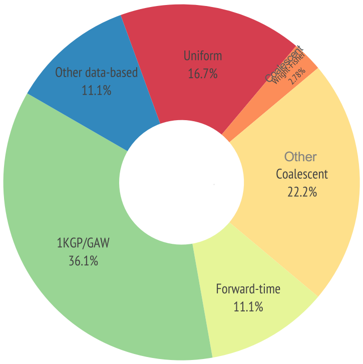
25

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

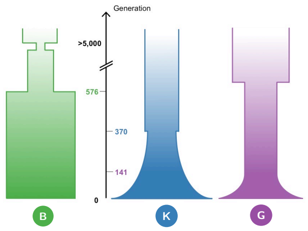
Simulation Methods



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” (1.0×10^{-5} – 1.8×10^{-2})
-“Nonsense, splice site and frameshift” ($>1.8 \times 10^{-2}$)

28

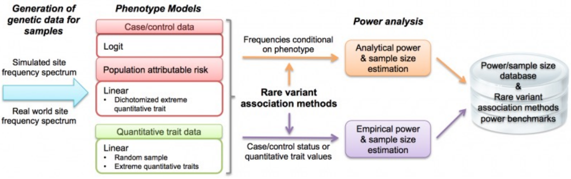
SKAT Power Calculator

- 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

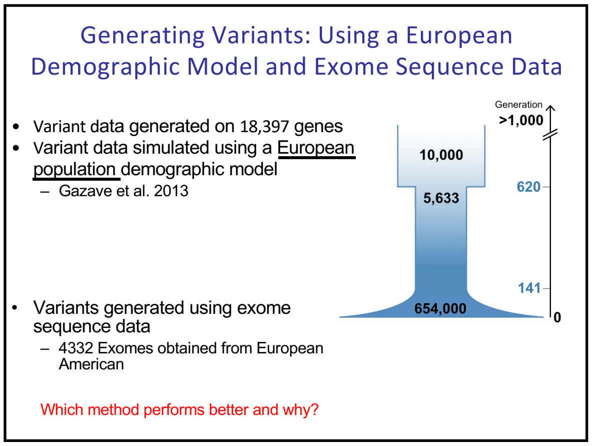
SEQPower

http://www.bioinformatics.org/spower/


Generation of genetic data for samples
Simulated site frequency spectrum
Real world site frequency spectrum
Phenotype Models
Case/control data
Logit
Population attributable risk
Linear
• Dichotomized extreme quantitative trait
Quantitative trait data
Linear
• Random sample
• Extreme quantitative traits
Frequencies conditional on phenotype
Rare variant association methods
Case/control status or quantitative trait values
Power analysis
Analytical power & sample size estimation
Empirical power & sample size estimation
Power/sample size database & Rare variant association methods power benchmarks

Wang et al. 2014 Bioinformatics

30

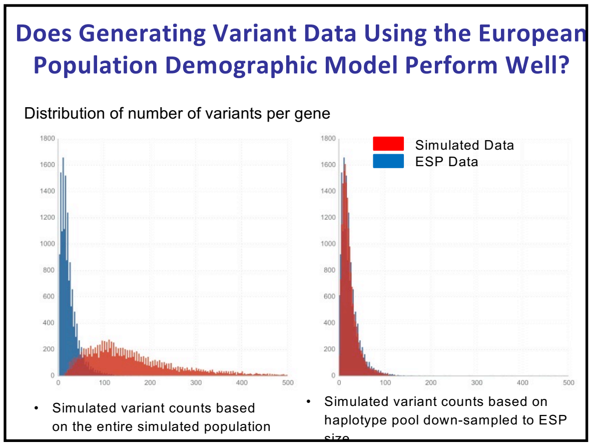


Generating Variants: Using a European Demographic Model and Exome Sequence Data

- 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

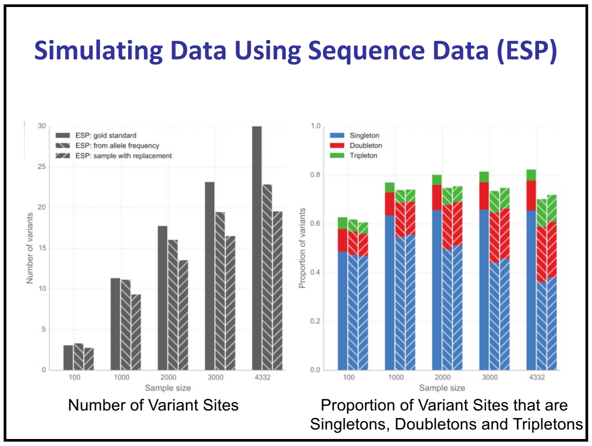


Does Generating Variant Data Using the European Population Demographic Model Perform Well?

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

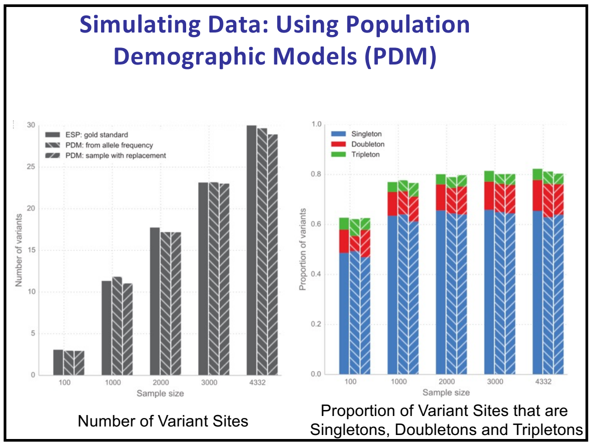


Simulating Data Using Sequence Data (ESP)

Number of Variant Sites

Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

33

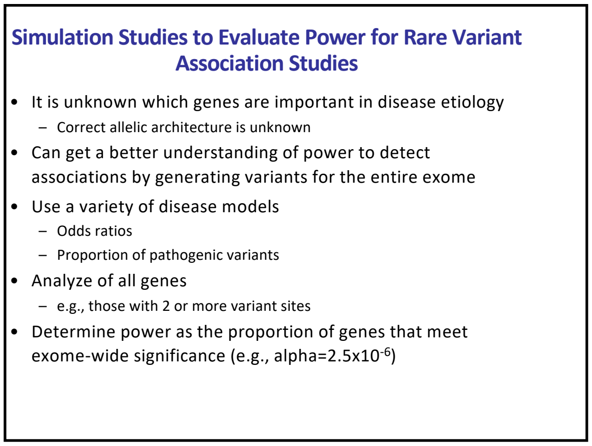


Simulating Data: Using Population Demographic Models (PDM)

Number of Variant Sites

Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

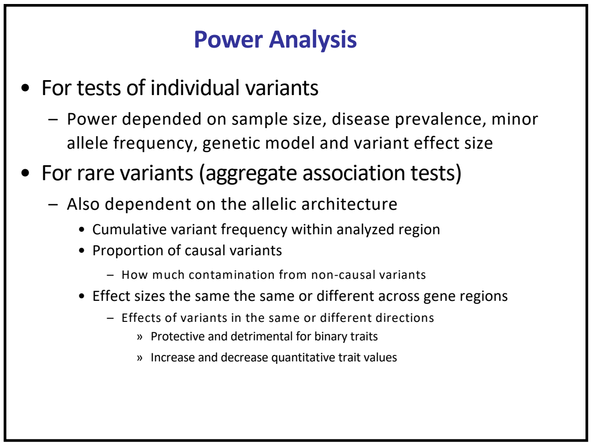
34



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=2.5x10^{-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

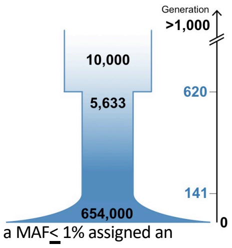
Power Analysis Rare Variants (Aggregate Association Tests)

- 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?

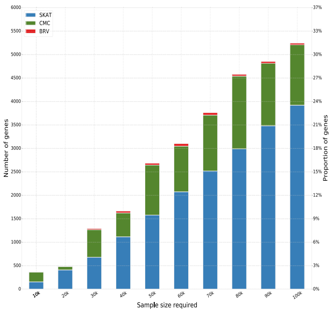
- Data generated on 18,397 genes
- Variant data simulated using a European population demographic model
  - Gazave et al. 2013



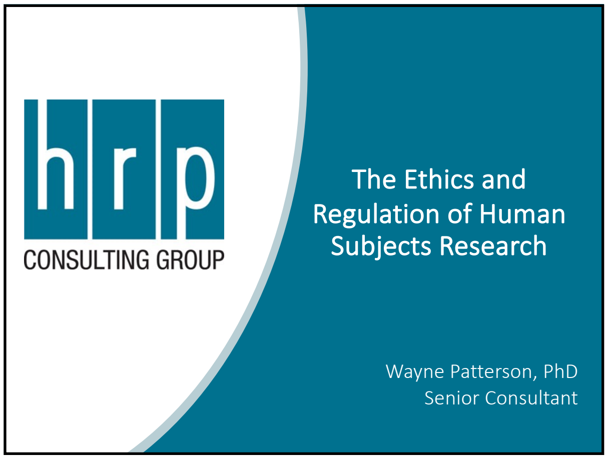
- Every missense, nonsense and splice with a MAF$\leq$ 1% assigned an odds ratio of 1.5
- Sample sizes to detect X number of genes determined for
  - $\alpha = 2.5 \times 10^{-6}$
  - power=0.8

38

Sample Sizes Necessary to Detect an Association (Case-Control Data)



39

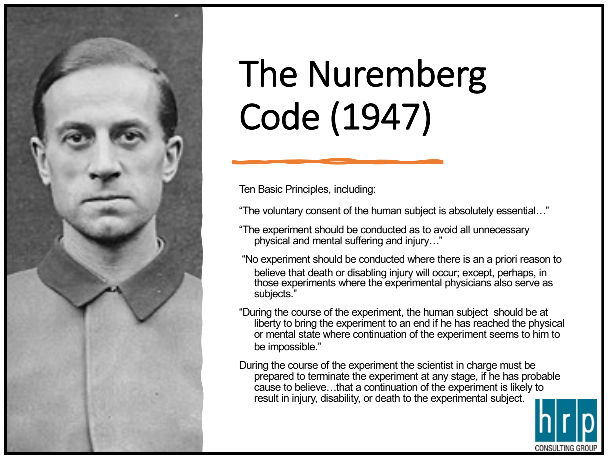


hrp
CONSULTING GROUP

The Ethics and Regulation of Human Subjects Research

Wayne Patterson, PhD
Senior 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…"

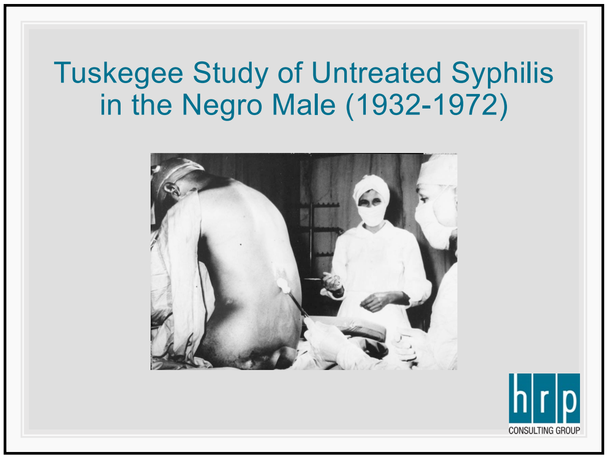
"No experiment should be conducted where there is an a priori reason to believe that death or disabling injury will occur; except, perhaps, in those experiments where the experimental physicians also serve as subjects."

"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.

hrp CONSULTING GROUP

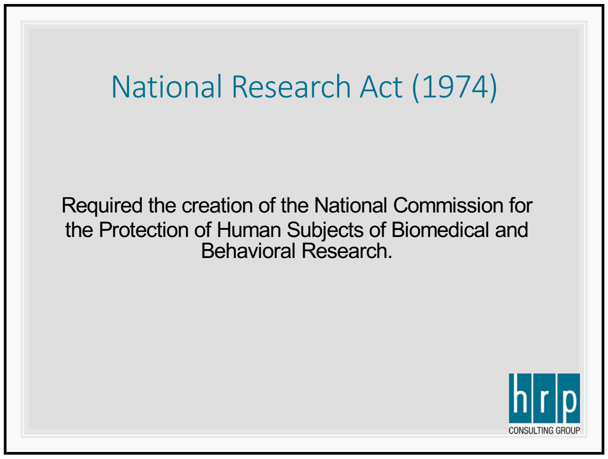
2



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

hrp CONSULTING GROUP

3

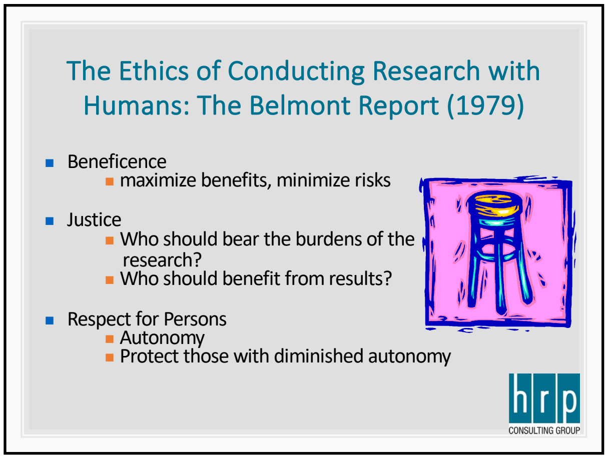


National Research Act (1974)

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

hrp CONSULTING GROUP

4

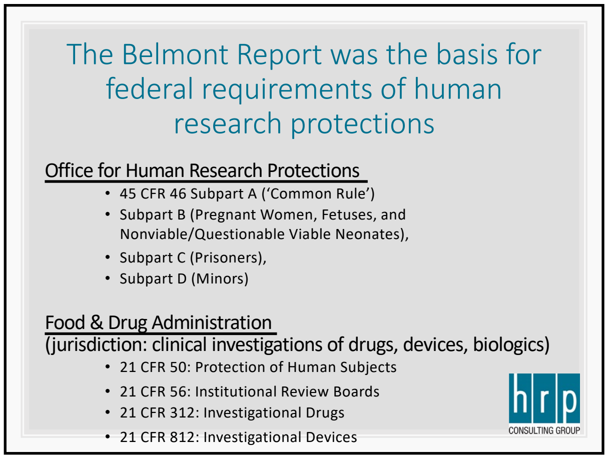


The Ethics of Conducting Research with Humans: The Belmont Report (1979)

- 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

hrp CONSULTING GROUP

5



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

Office for Human Research Protections

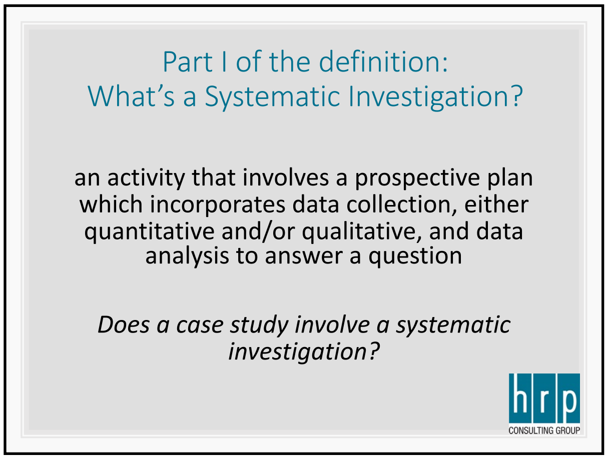
- 45 CFR 46 Subpart A ('Common Rule')
- Subpart B (Pregnant Women, Fetuses, and Nonviable/Questionable Viable Neonates),
- Subpart C (Prisoners),
- Subpart D (Minors)

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

hrp CONSULTING GROUP

6



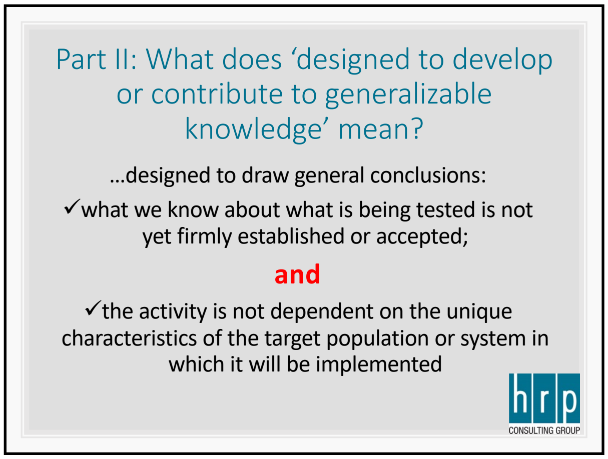
Part I of the definition: What's a Systematic Investigation?

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?

hrp CONSULTING GROUP

13



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

…designed to draw general conclusions:

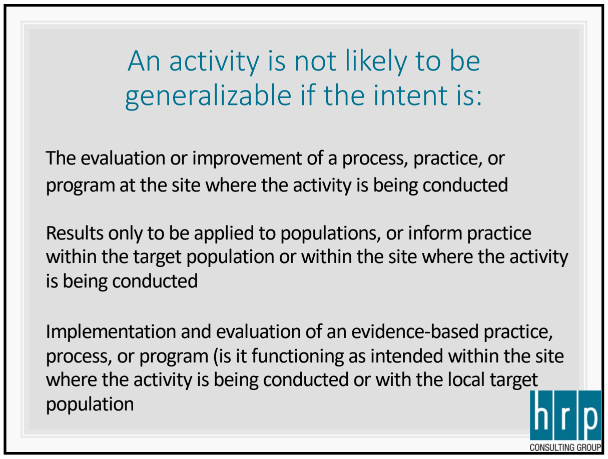
✓what we know about what is being tested is not yet firmly established or accepted;

and

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

hrp CONSULTING GROUP

14



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

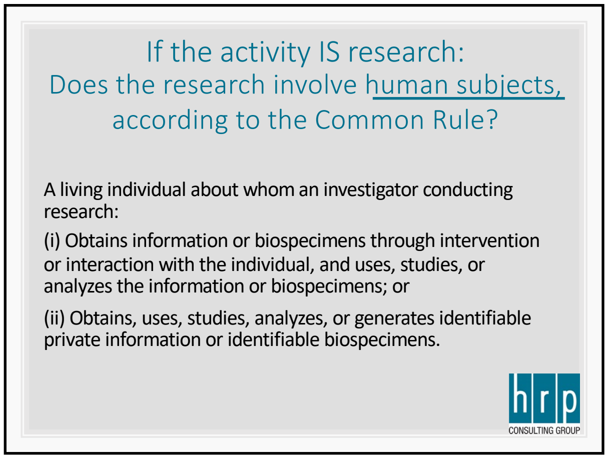
The evaluation or improvement of a process, practice, or program at the site where the activity is being conducted

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

hrp CONSULTING GROUP

15



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

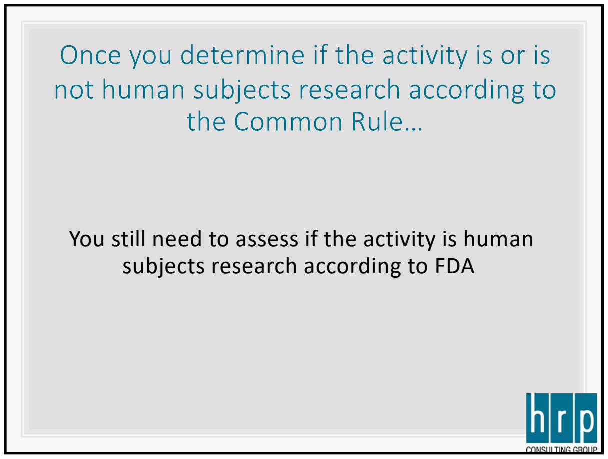
A living individual about whom an investigator conducting research:

(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.

hrp CONSULTING GROUP

16

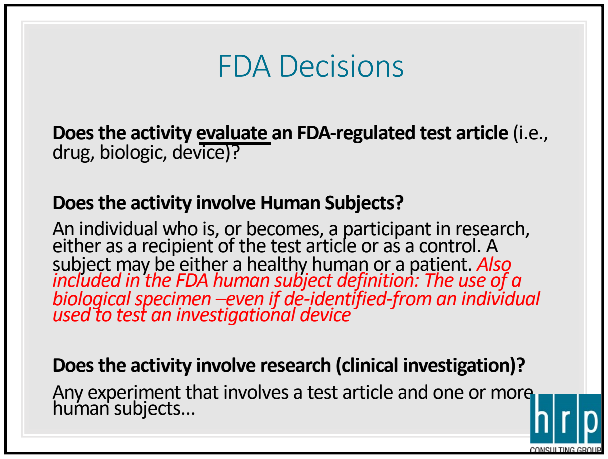


Once you determine if the activity is or is not human subjects research according to the Common Rule…

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

hrp CONSULTING GROUP

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…

hrp CONSULTING GROUP

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, 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;

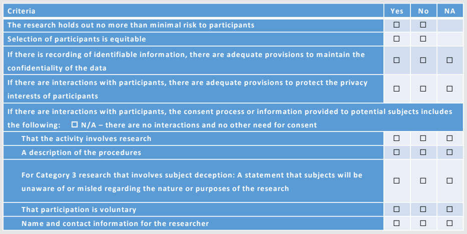
22

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)"

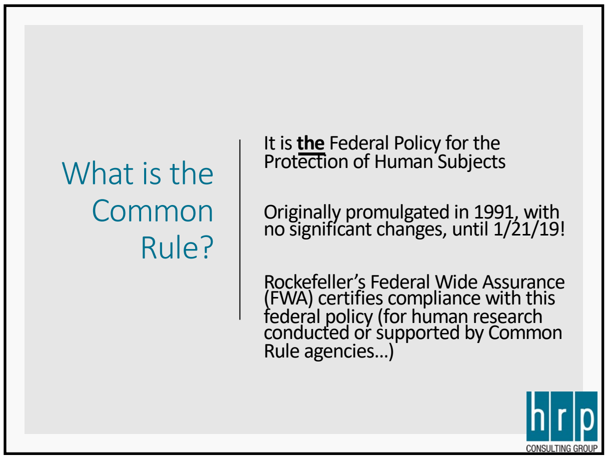
23

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



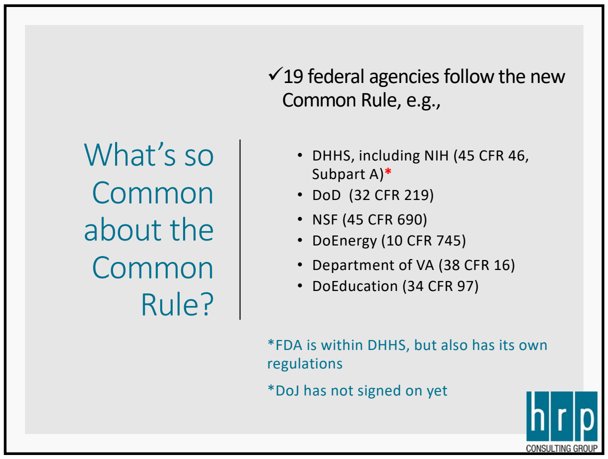
What is the Common Rule?

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?

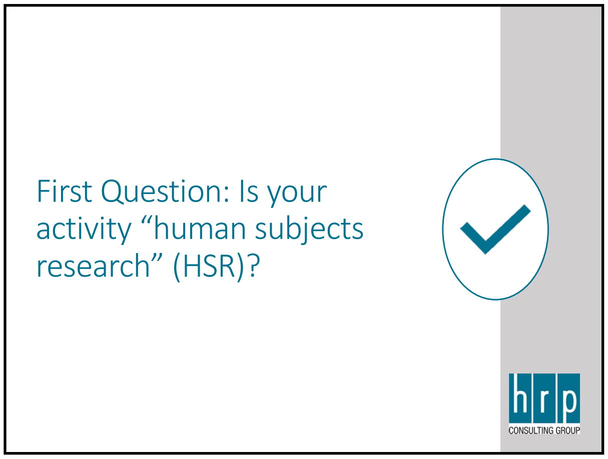
✓19 federal agencies follow the new Common Rule, e.g.,

- 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

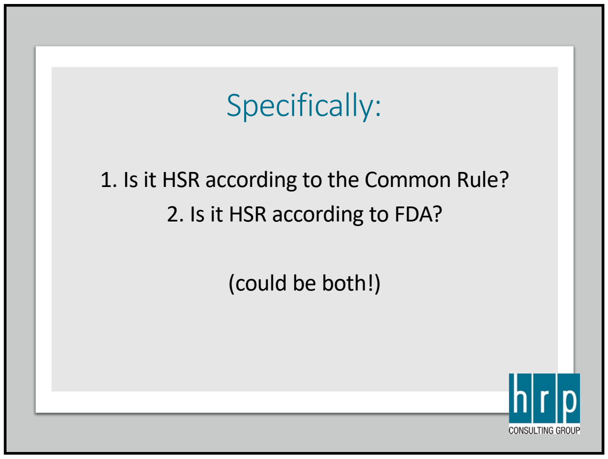
*DoJ has not signed on yet

8



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

9

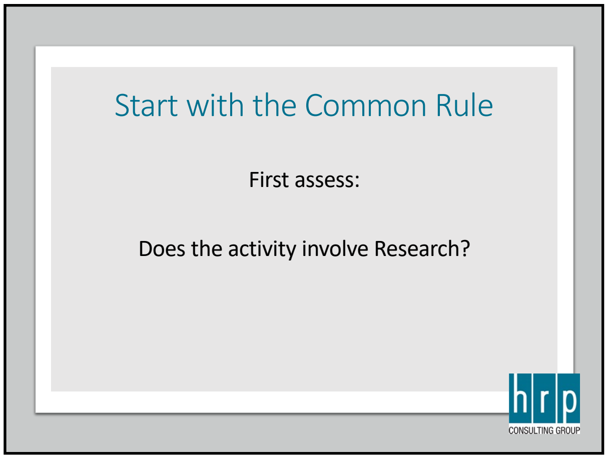


Specifically:

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

(could be both!)

10

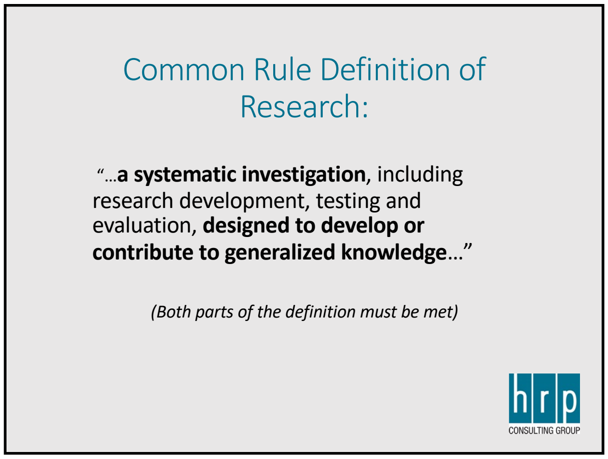


Start with the Common Rule

First assess:

Does the activity involve Research?

11

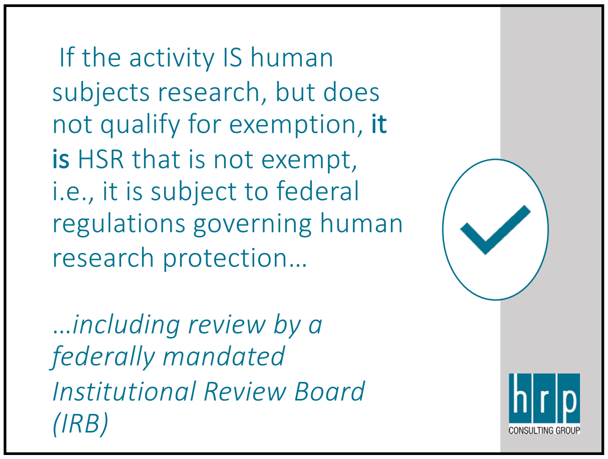


Common Rule Definition of Research:

"…**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)

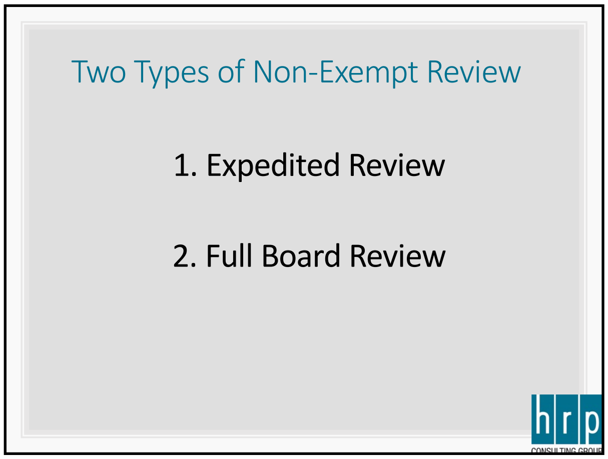
12



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)

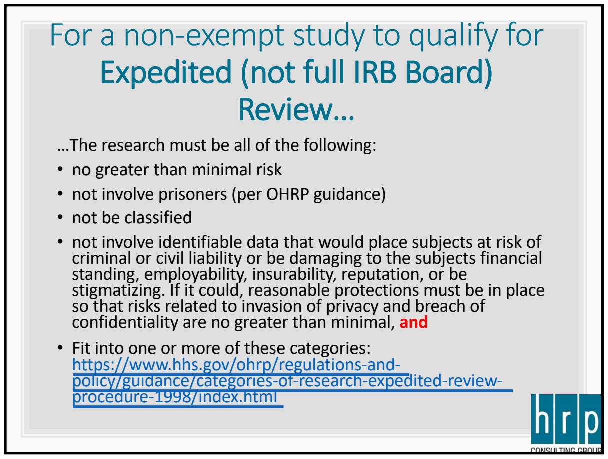
25



Two Types of Non-Exempt Review

1. Expedited Review
2. Full Board Review

26

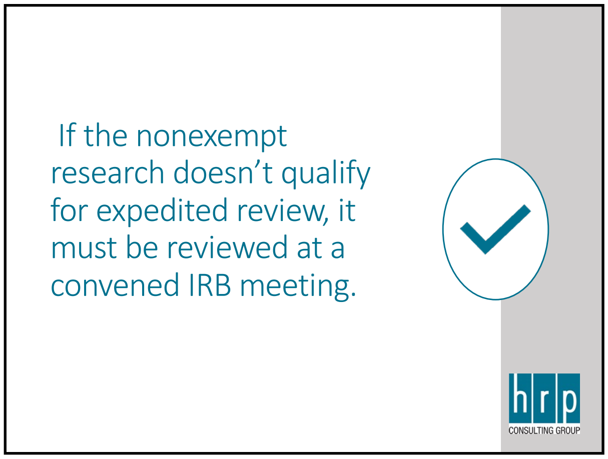


For a non-exempt study to qualify for **Expedited (not full IRB Board) Review…**

…The research must be all of the following:

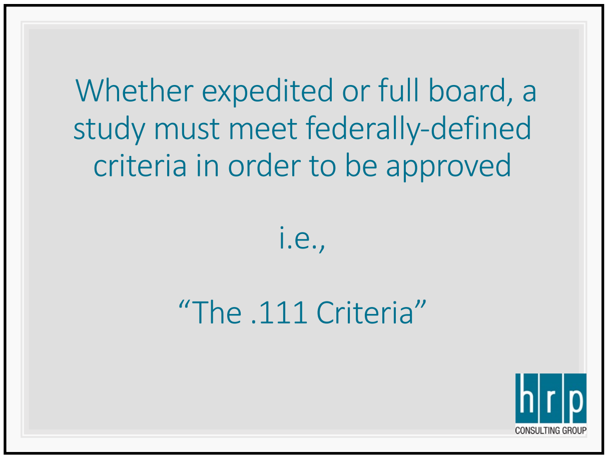
- no greater than minimal risk
- not involve prisoners (per OHRP guidance)
- not be classified
- not involve identifiable data that would place subjects at risk of criminal or civil liability or be damaging to the subjects financial standing, employability, insurability, reputation, or be stigmatizing. 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

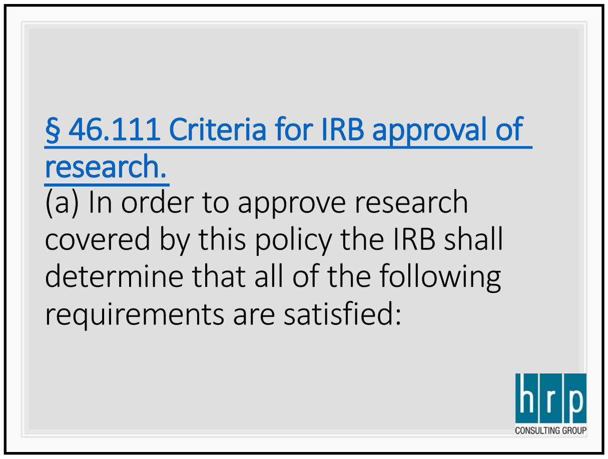


Whether expedited or full board, a study must meet federally-defined criteria in order to be approved

i.e.,

"The .111 Criteria"

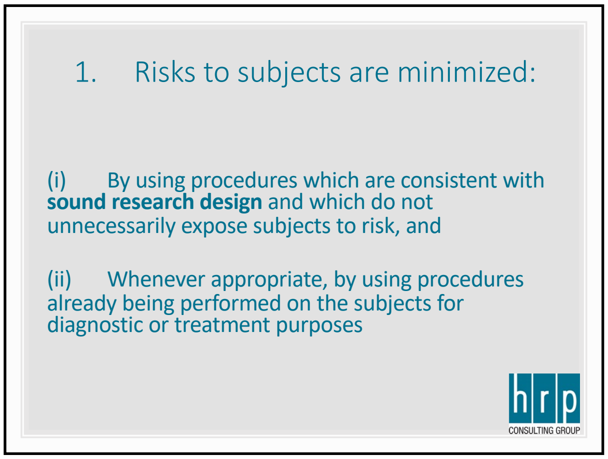
29



§ 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

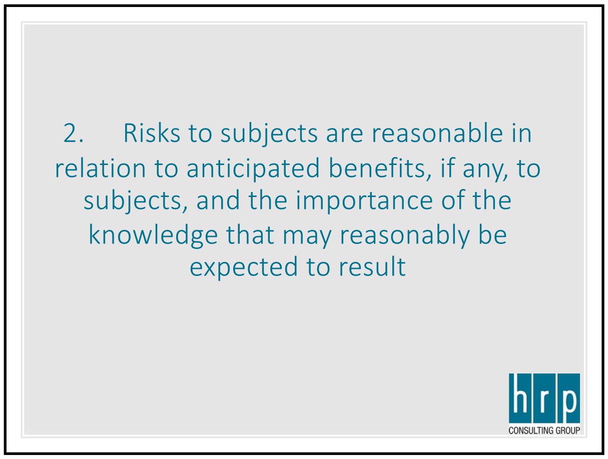


1. Risks to subjects are minimized:

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

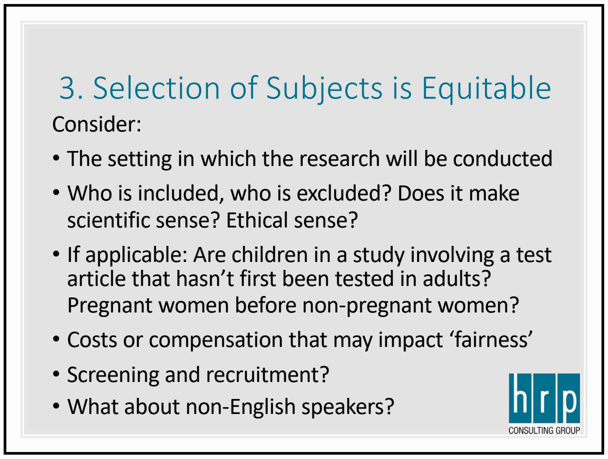
(ii) Whenever appropriate, by using procedures already being performed on the subjects for diagnostic or treatment purposes

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

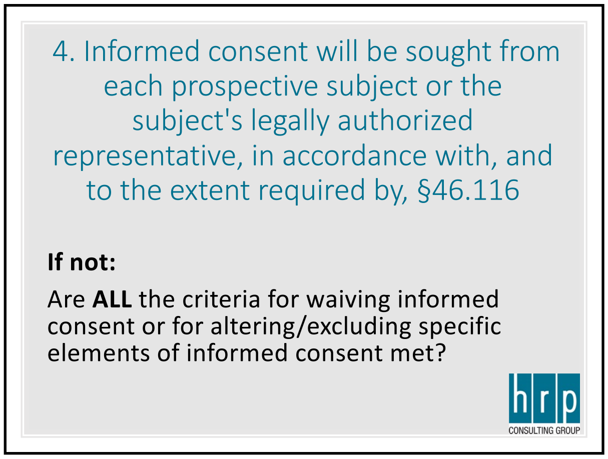


3. Selection of Subjects is Equitable

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

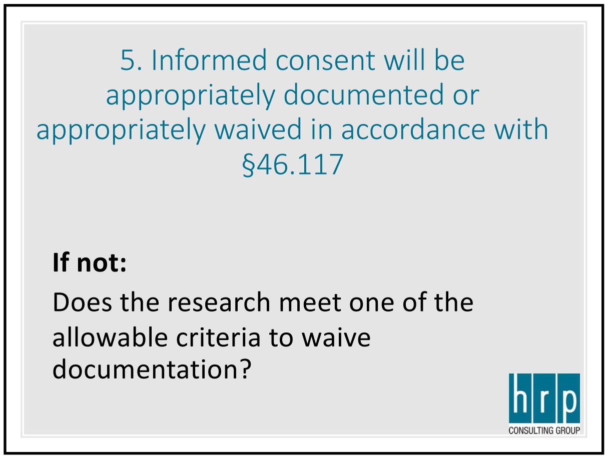


4. Informed consent will be sought from each prospective subject or the subject's legally authorized representative, in accordance with, and to the extent required by, §46.116

If not:

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

34

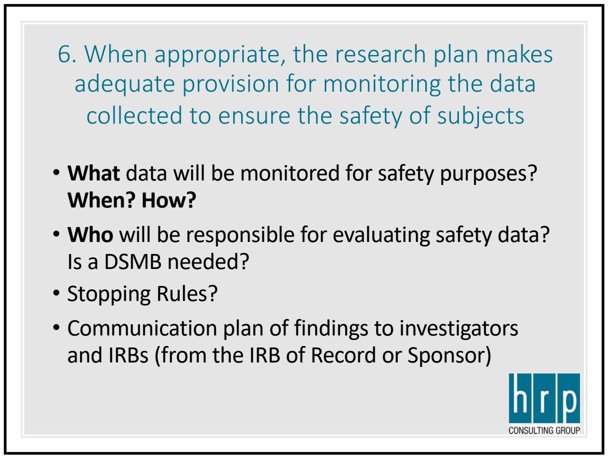


5. Informed consent will be appropriately documented or appropriately waived in accordance with §46.117

If not:

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

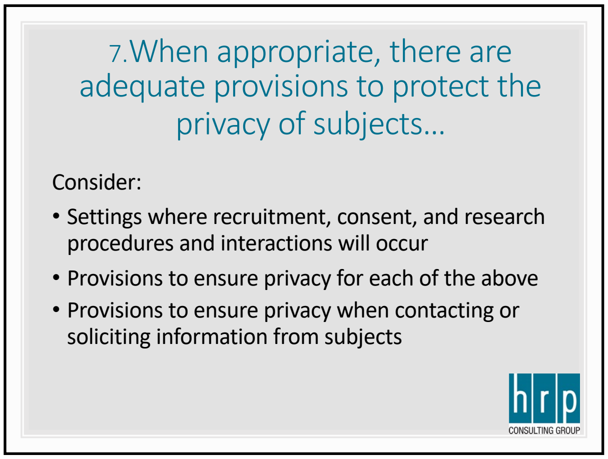
35



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

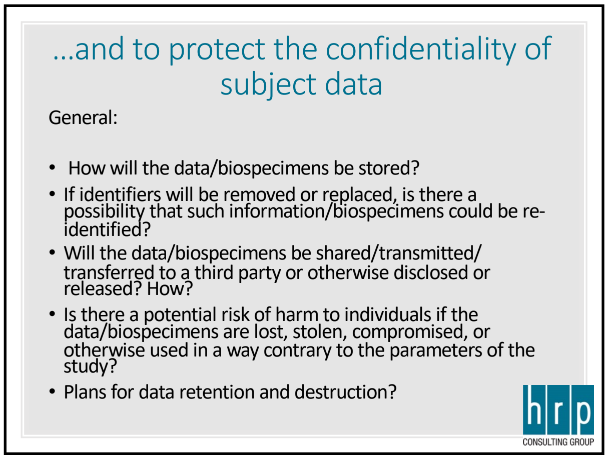


7.When appropriate, there are adequate provisions to protect the privacy of subjects…

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

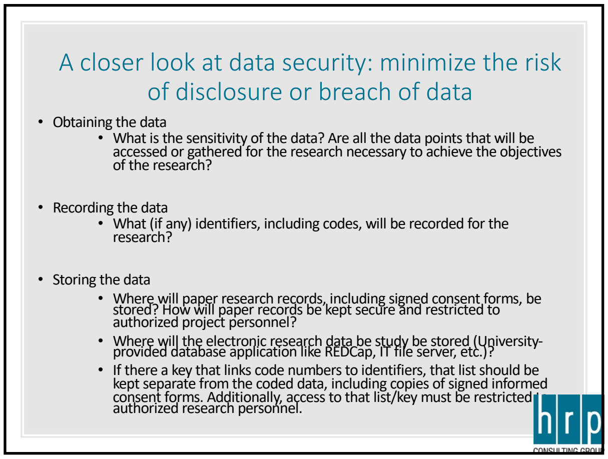


…and to protect the confidentiality of subject data

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?

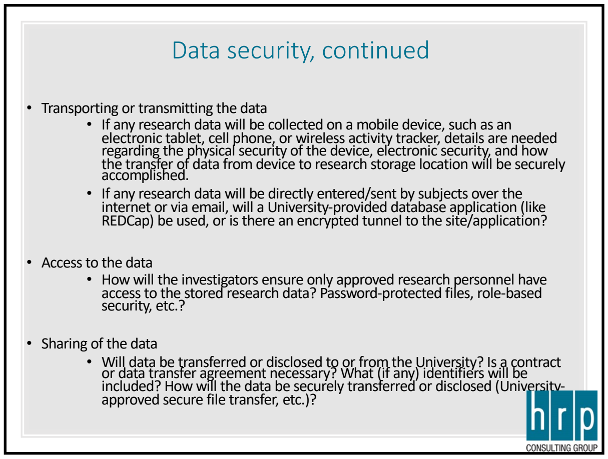
38



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.

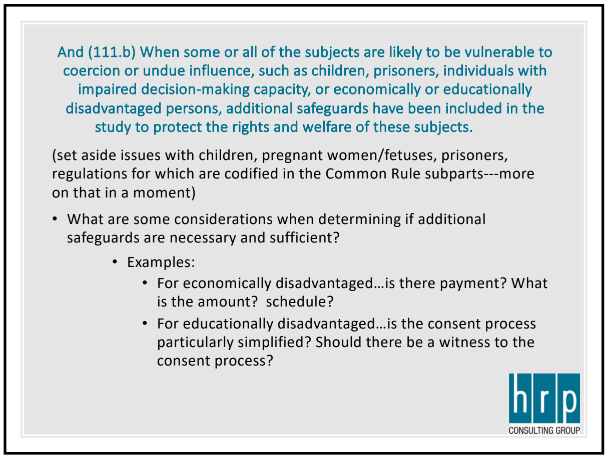
39



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

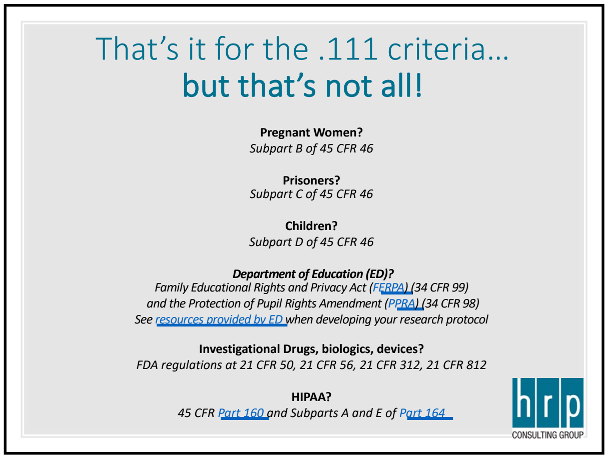


And (111.b) When some or all of the subjects are likely to be vulnerable to coercion or undue influence, such as children, prisoners, individuals with impaired decision-making capacity, or economically or educationally disadvantaged persons, additional safeguards have been included in the study to protect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

hrp
CONSULTING GROUP

43

Linkage disequilibrium in genetic association studies

Gao Wang, Ph.D.

Advanced Gene Mapping Course, May 2023

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

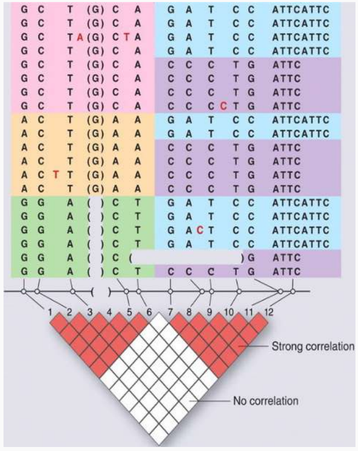
Linkage disequilibrium (LD) — useful

- 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

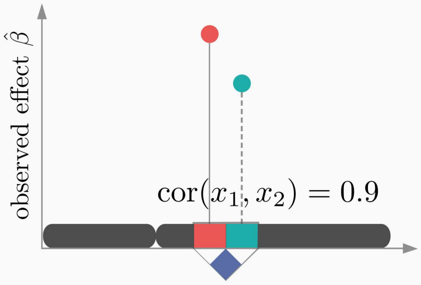


Altshuler *et al.* (2008)

Impact of LD on GWAS analysis

Oligogenic: trait influenced by a few genetic variants

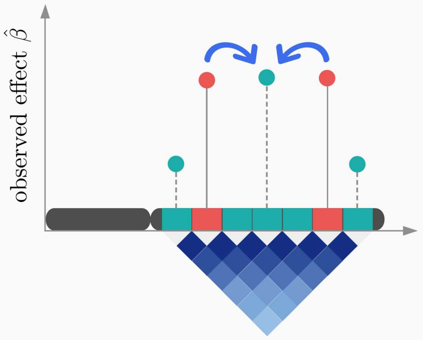
- Misleading: difficult to identify causal variants
- Useful: 'tag SNPs' in array based GWAS design



Impact of LD on GWAS analysis

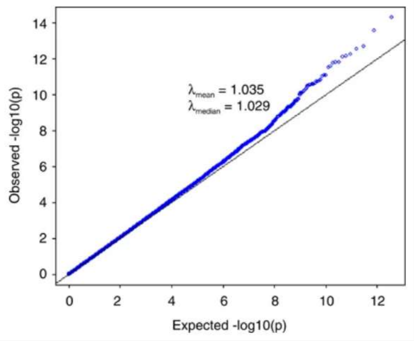
Polygenic: trait influenced by numerous genetic variants

- 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

$$\mathsf{E}[\chi_j^2] = 1 + \frac{N h_g^2}{M} l_j$$

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

LD score regression (LDSC)

Separating h_g^2 and population stratification

Population stratification factor

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

LD score of variant j

Regression intercept

Regression slope

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

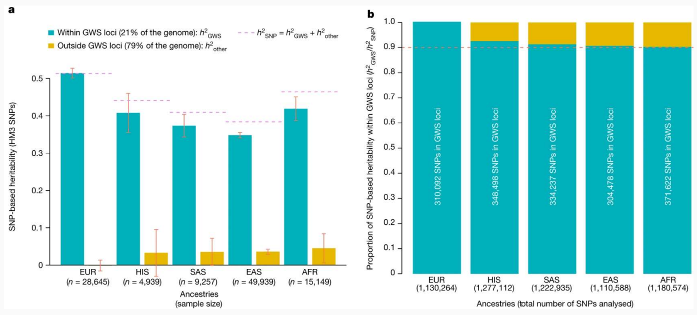
Estimation strategy

- 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

Statistical fine-mapping in genetic association studies

Gao Wang, Ph.D.

Advanced Gene Mapping Course, May 2023

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

1. Fine-mapping: background and challenge
2. A (naively) simple approach to fine-mapping
3. Probabilistic fine-mapping: Bayesian Variable Selection
4. A simple Bayesian variable selection with applications to fine-mapping
5. Other variable selection problems in genetics

Fine-mapping: background and challenge

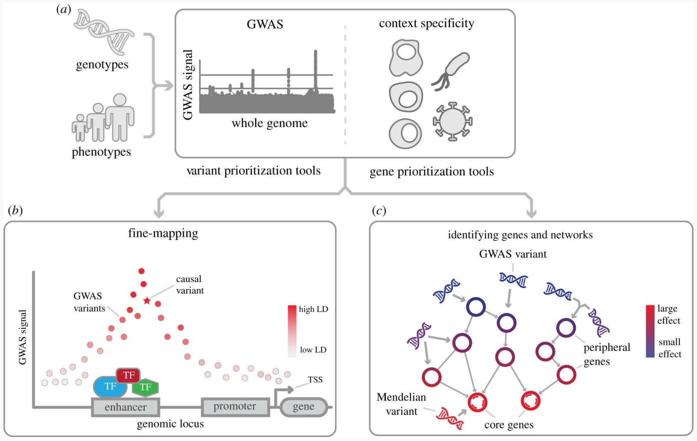
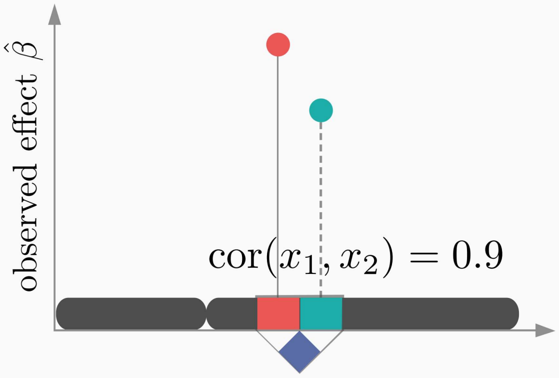


Figure: Broekema *et al.* (2020) Open Biol.

Correlated variables in association studies

Due to a phenomenon called **linkage disequilibrium** (LD)



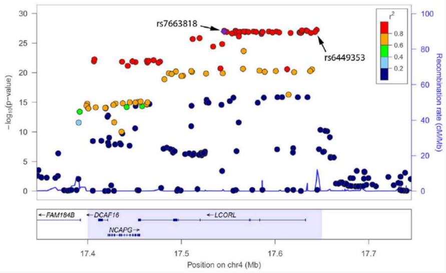


Figure: N'Diaye *et al.* (2011) PLoS Genet.

Objectives

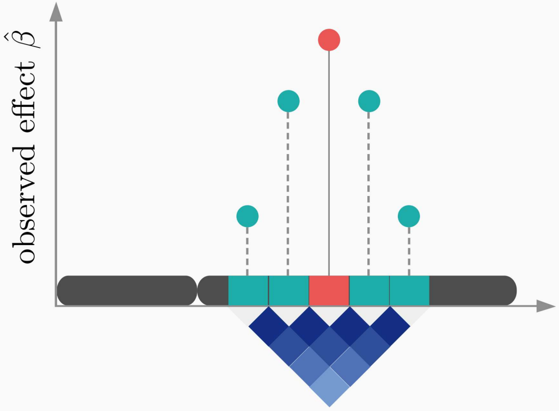
Statistical fine-mapping **aids in** the identification of causal variants, in order to

- interpret association signals (pinpoint to genes)
- understand biological function of a variant
- elucidate genetic architecture of complex and molecular phenotypes

6

Identify non-zero effect ("causal") variables

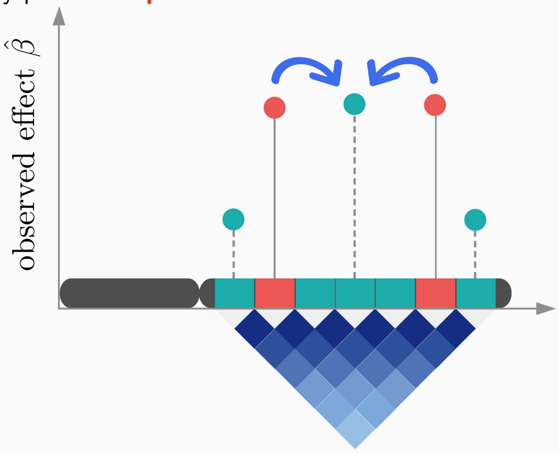
Simply pick the **top** association in an LD block? Maybe?



7

Identify non-zero effect ("causal") variables

Simply pick the **top** association in an LD block? ... or not!



7

Architecture: sparse effects, polygenic background

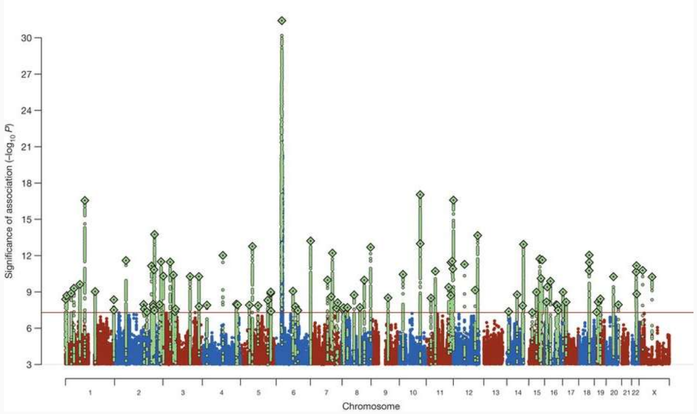


Figure: O'Donovan *et al.* (2014) Nature

8

Challenge: large-sample computational challenge

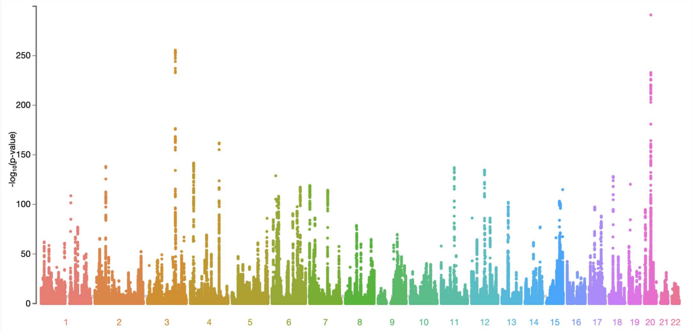


Figure: UK Biobank height GWAS, `http://nealelab.is/uk-biobank`

9

A (naively) simple approach to fine-mapping

“One causal SNP” assumption

Effect variable (red) correlated with non-effect variable (green)

	case		control		p-value
	A1	A2	A1	A2	
SNP 1	1200	800	1000	1000	2.1×10^{-10}
SNP 2	1191	809	1000	1000	1.3×10^{-9}

“One causal SNP” assumption

	case		control		p-value
	A1	A2	A1	A2	
SNP 1	1200	800	1000	1000	2.1×10^{-10}
SNP 2	1191	809	1000	1000	1.3×10^{-9}

Compute likelihood ratios (LR) H_1 vs H_0,

$$\text{LR}_1 = 6.15 \times 10^8 \quad \text{LR}_2 = 0.94 \times 10^8$$

“One causal SNP” assumption

	case		control		p-value
	A1	A2	A1	A2	
SNP 1	1200	800	1000	1000	2.1×10^{-10}
SNP 2	1191	809	1000	1000	1.3×10^{-9}

Compute likelihood ratios (LR) H_1 vs H_0,

$$\text{LR}_1 = 6.15 \times 10^8 \quad \text{LR}_2 = 0.94 \times 10^8$$

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

$$\frac{\text{LR}_1}{\text{LR}_1 + \text{LR}_2} = 0.87 \qquad \frac{\text{LR}_2}{\text{LR}_1 + \text{LR}_2} = 0.13$$

Per variable contingency table analysis, R code

```r
# returns likelihood ratio of H_1 vs H_0
get_2x2_lr = function(tbl) {
    tbl = as.table(matrix(tbl, 2,2,
      dimnames=list(status=c('case','control'),
      genotype=c('minor_allele','major_allele'))))
    test = MASS::loglm(~status+genotype,data=tbl)
    return(exp(test$lrt / 2))
}
lr1 = get_2x2_lr(c(1200,800,1000,1000))
lr2 = get_2x2_lr(c(1190,809,1000,1000))
```

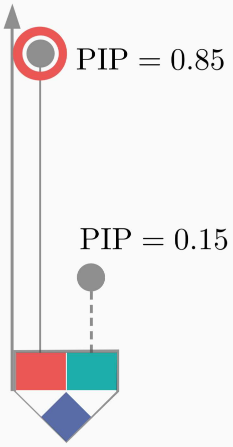
A “single effect” Bayesian variable selection

Use Bayes Factor, and compute **posterior inclusion probability**

case		control		p-value
A1	A2	A1	A2	
1200	800	1000	1000	2.1×10^{-10}
1191	809	1000	1000	1.3×10^{-9}

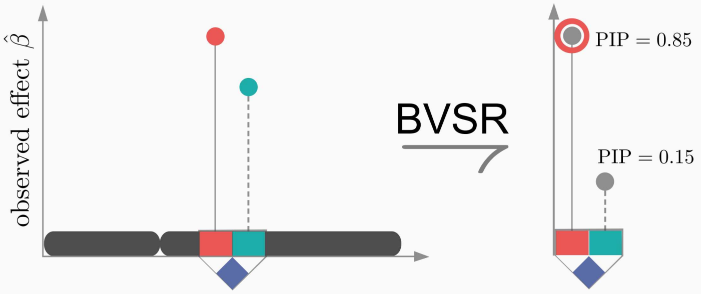
$$\text{PIP}_1 = \frac{\text{BF}_1}{\text{BF}_1 + \text{BF}_2} = 0.85$$

$$\text{PIP}_2 = \frac{\text{BF}_2}{\text{BF}_1 + \text{BF}_2} = 0.15$$



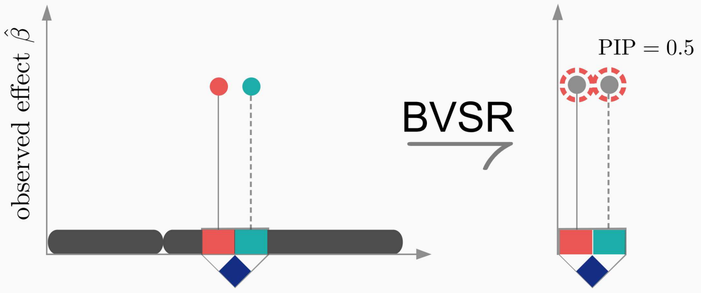
Bayesian variable selection: PIP

Computes **Posterior Inclusion Probability** (PIP)



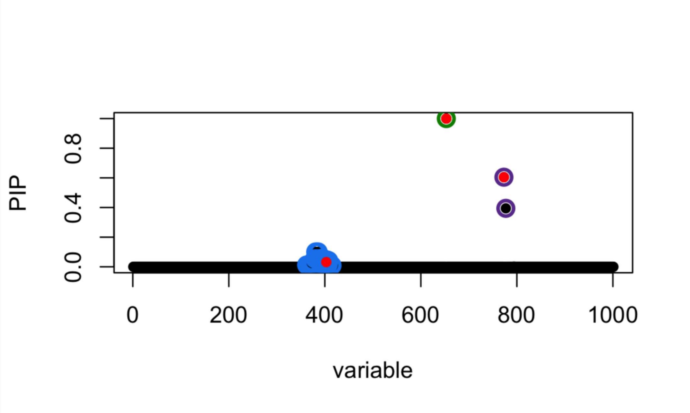
Bayesian variable selection: PIP

Computes **Posterior Inclusion Probability** (PIP)



Bayesian variable selection: Credible Sets

'Clusters' of signals to account for correlations between variables (eg LD)



Bayesian variable selection: Credible Sets

- **95% credible set** S: $\Pr(effect\ variable\ in\ \mathrm{S}) \geq 95\%$
- *e.g.* , "Single effect" model:

$$\sum_{j \in S} PIP_{(j)} \geq 95\%$$

  where $PIP_{(j)}$'s are in descending order.
- Formal definition: Wang *et al.* (2020) J. R. Stat. Soc. B

Multiple effects: step-wise search

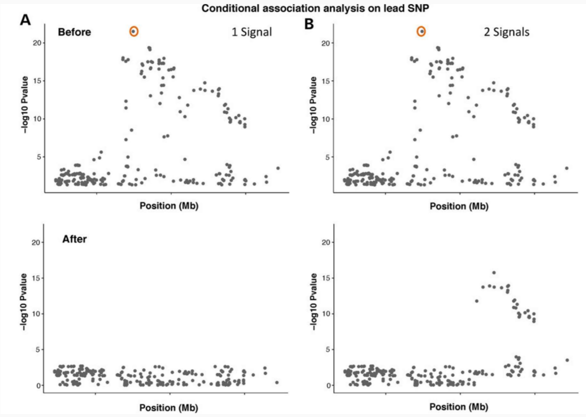


Figure: Spain and Barrett (2015) Hum. Mol. Genet.

A simple frequentist conditional analysis

Forward selection algorithm

1. For each SNP fit a simple linear regression model
2. Select the SNP j that has the largest model likelihood
3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \boldsymbol{X}_j \hat{b}_j$, and repeat

A simple frequentist conditional analysis

Forward selection algorithm

1. For each SNP fit a simple linear regression model
2. Select the SNP j that has the largest model likelihood
3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \boldsymbol{X}_j \hat{b}_j$, and repeat

A greedy algorithm to choose the "best" SNPs, but is incapable of capturing multiple SNPs in LD

To quantify uncertainty

Bayesian forward selection algorithm

1. For each SNP j, fit a simple Bayesian linear regression model to get Bayes Factor BF_j
2. Form weight for each SNP, $w_j \propto \text{BF}_j$
3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \sum_j w_j \boldsymbol{X}_j \hat{b}_j$, and repeat

To quantify uncertainty

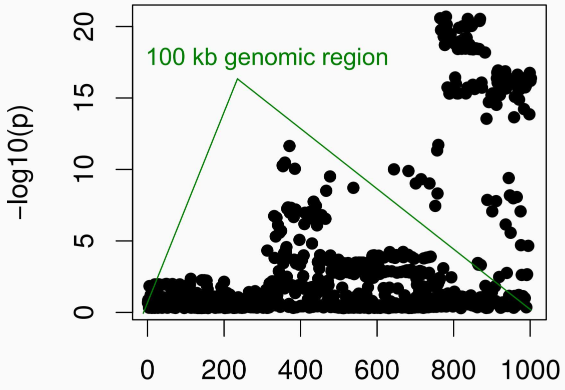
Bayesian forward selection algorithm

1. For each SNP j, fit a simple Bayesian linear regression model to get Bayes Factor BF_j
2. Form weight for each SNP, $w_j \propto \text{BF}_j$
3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \sum_j w_j \boldsymbol{X}_j \hat{b}_j$, and repeat

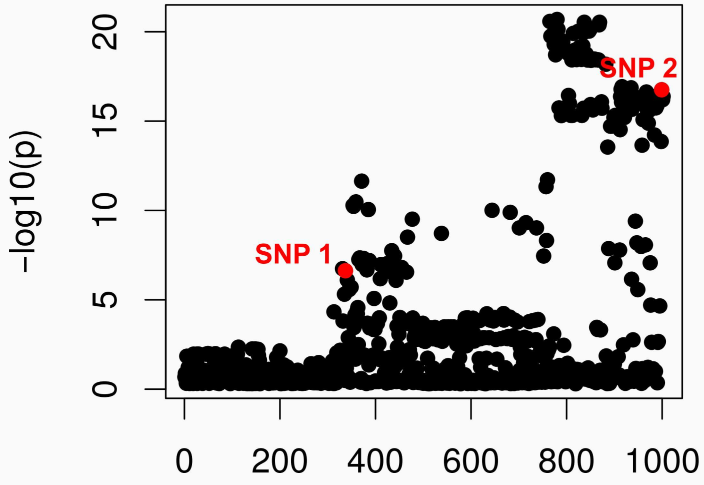
What if a "bad decision" is made early on?

A motivating example

data available as `data(susieR::N2finemapping)`



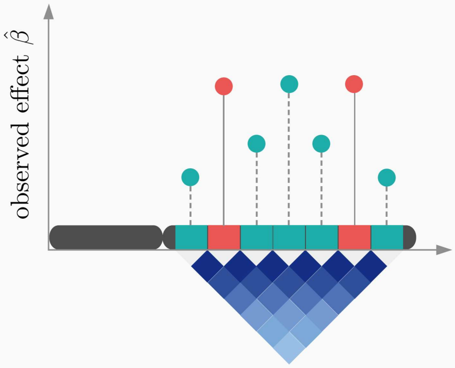
A motivating example



Probabilistic fine-mapping:
Bayesian Variable Selection

Detecting multiple effect variables

Intuition: A model involving the two effect variables should fit the data better than that involving the top variable.



Bayesian Variable Selection Regression (BVSR)

Fine-mapping is a particular multiple regression problem:

$$\boldsymbol{y}_{n\times1} = \boldsymbol{X}_{n\times p}\boldsymbol{b}_{p\times1} + \boldsymbol{e}_{n\times1}$$

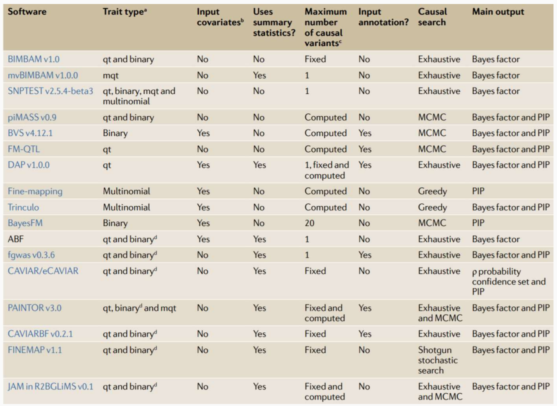
- $\boldsymbol{b}$ is sparse: most of its elements are 0's
- Columns of $\boldsymbol{X}$ are very correlated

Why BVSR?

- Other sparse variable selection regression may not work
  - designed to minimize prediction errors, *e.g.* LASSO

Why BVSR?

- Other sparse variable selection regression may not work
  - designed to minimize prediction errors, *e.g.* LASSO
- Bayesian variable selection regression (BVSR)
  - can evaluate significance of effect variables
  - can quantify **uncertainty** in variables selected



Software	Trait type[a]	Input covariates[b]	Uses summary statistics?	Maximum number of causal variants[c]	Input annotation?	Causal search	Main output
BIMBAM v1.0	qt and binary	No	No	Fixed	No	Exhaustive	Bayes factor
mvBIMBAM v1.0.0	mqt	No	Yes	1	No	Exhaustive	Bayes factor
SNPTEST v2.5.4-beta3	qt, binary, mqt and multinomial	No	No	1	No	Exhaustive	Bayes factor
piMASS v0.9	qt and binary	No	No	Computed	No	MCMC	Bayes factor and PIP
BVS v4.12.1	Binary	Yes	No	Computed	Yes	MCMC	Bayes factor and PIP
FM-QTL	qt	No	No	Computed	Yes	MCMC	Bayes factor and PIP
DAP v1.0.0	qt	Yes	Yes	1, fixed and computed	Yes	Exhaustive	Bayes factor and PIP
Fine-mapping	Multinomial	Yes	No	Computed	No	Greedy	PIP
Trinculo	Multinomial	Yes	No	Computed	No	Greedy	Bayes factor and PIP
BayesFM	Binary	Yes	No	20	No	MCMC	PIP
ABF	qt and binary[d]	Yes	Yes	1	No	Exhaustive	Bayes factor
fgwas v0.3.6	qt and binary[d]	No	Yes	1	Yes	Exhaustive	Bayes factor and PIP
CAVIAR/eCAVIAR	qt and binary[d]	No	Yes	Fixed	No	Exhaustive	ρ probability confidence set and PIP
PAINTOR v3.0	qt, binary[d] and mqt	No	Yes	Fixed and computed	Yes	Exhaustive and MCMC	Bayes factor and PIP
CAVIARBF v0.2.1	qt and binary[d]	No	Yes	Fixed	Yes	Exhaustive	Bayes factor and PIP
FINEMAP v1.1	qt and binary[d]	No	Yes	Fixed	No	Shotgun stochastic search	Bayes factor and PIP
JAM in R2BGLiMS v0.1	qt and binary[d]	No	Yes	Fixed and computed	No	Exhaustive and MCMC	Bayes factor and PIP

Figure: Schaid *et al.* (2018) Nat. Rev. Genet.

BVSR model

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

γ: model configurations; $\boldsymbol{\pi}$: prior inclusion probability.

BVSR results

Assess **combinations** of variables

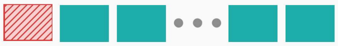
SNPs	1	2	3	4	5	⋯	Probability
model configurations	1	0	1	0	0	⋯	0.25
	1	0	0	1	0	⋯	0.25
	0	1	1	0	0	⋯	0.25
	0	1	0	1	0	⋯	0.25

- $\text{PIP}_j := Pr(z_j \text{ is non-zero})$

$$\text{PIP} = (0.5, 0.5, 0.5, 0.5, 0, \cdots)$$

Assessing multi-effects configurations

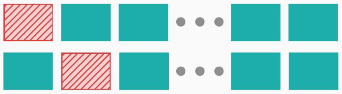
$L = 1$



$\Leftarrow \Pr(\mathcal{M}_1)$

Assessing multi-effects configurations

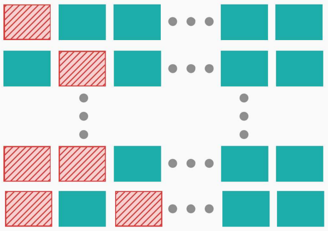
$L = 1$



$\Leftarrow \Pr(\mathcal{M}_1)$
$\Leftarrow \Pr(\mathcal{M}_2)$

Assessing multi-effects configurations

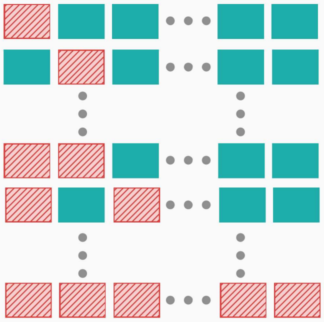
$L = 2$



$\Leftarrow \Pr(\mathcal{M}_1)$
$\Leftarrow \Pr(\mathcal{M}_2)$
$\Leftarrow \Pr(\mathcal{M}_J)$
$\Leftarrow \Pr(\mathcal{M}_{J+1})$

Assessing multi-effects configurations

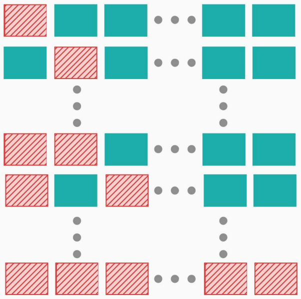
$L = P$



$\Leftarrow \Pr(\mathcal{M}_1)$
$\Leftarrow \Pr(\mathcal{M}_2)$
$\Leftarrow \Pr(\mathcal{M}_J)$
$\Leftarrow \Pr(\mathcal{M}_{J+1})$
$\Leftarrow \Pr(\mathcal{M}_P)$

Assessing multi-effects configurations

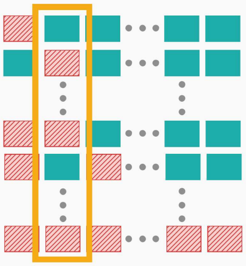
$L = P$



$\Rightarrow \mathrm{PIP}(\mathcal{M}_1)$
$\Rightarrow \mathrm{PIP}(\mathcal{M}_2)$
$\Rightarrow \mathrm{PIP}(\mathcal{M}_J)$
$\Rightarrow \mathrm{PIP}(\mathcal{M}_{J+1})$
$\Rightarrow \mathrm{PIP}(\mathcal{M}_P)$

Assessing multi-effects configurations

Marginal associations

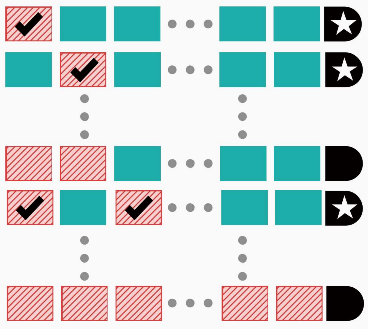


$\mathrm{PIP}(\mathcal{M}_1)$
$\mathrm{PIP}(\mathcal{M}_2)$
$\mathrm{PIP}(\mathcal{M}_J)$
$\mathrm{PIP}(\mathcal{M}_{J+1})$
$\mathrm{PIP}(\mathcal{M}_P)$

$$\mathrm{PIP}_2 = \mathrm{PIP}(\mathcal{M}_2) + \mathrm{PIP}(\mathcal{M}_J) + \mathrm{PIP}(\mathcal{M}_P)$$

Assessing multi-effects configurations

The 95% (smallest) Credible Set



$\alpha_1 = 0.70$

$\alpha_2 = 0.15$

$\alpha_3 = 0.02$

$\alpha_4 = 0.10$

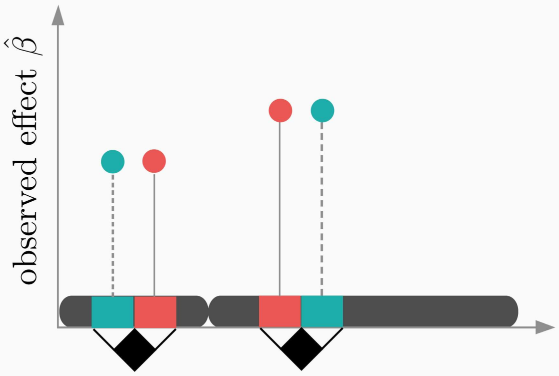
$\alpha_5 = 0.00$

BVSR inference: posterior methods

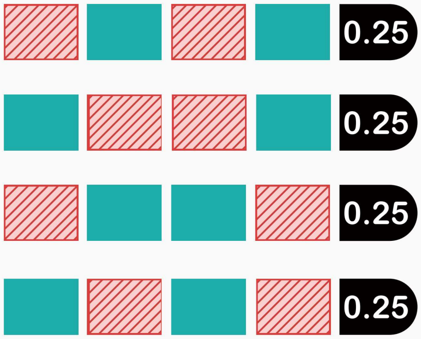
BVSR is computationally challenging!

- MCMC: BIMBAM, Guan & Stephens (2011)
- Enumeration: CAVIAR, Hormozdiari *et al.* (2014)
- Schochastic search: FINEMAP, Benner *et al.* (2016)
- Deterministic approximation: DAP-G, Wen *et al.* (2016)

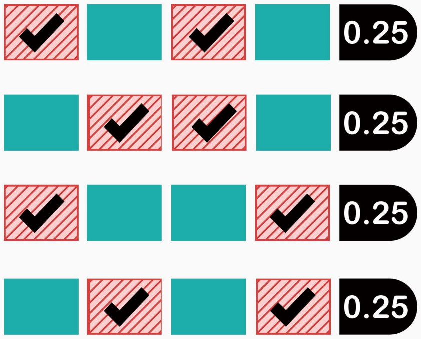
Summarizing BVSR results



Summarizing BVSR results



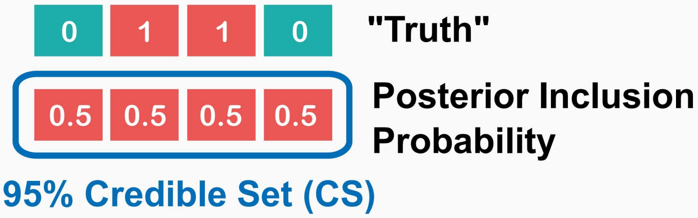
Summarizing BVSR results



Summarizing BVSR results

0	1	1	0	"Truth"
0.5	0.5	0.5	0.5	Posterior Inclusion Probability

Summarizing BVSR results


0 1 1 0 "Truth"

0.5 0.5 0.5 0.5 Posterior Inclusion Probability

95% Credible Set (CS)

Limitation of BVSR inference

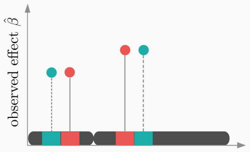
0.5 0.5 0.5 0.5

- There are 2 signals expected $(0.5 + 0.5 + 0.5 + 0.5)$
- But **which two?** Any two?
- 95% certainty that **all** effect variables are captured?

Limitation of BVSR inference

0.5 0.5 0.5 0.5

- There are 2 signals expected $(0.5 + 0.5 + 0.5 + 0.5)$
- But **which two?** Any two?
- 95% certainty that **all** effect variables are captured?
- We need to quantify this better!



Quantifying uncertainty in variable selection

Consider a sparse regression example

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{x}_j b_j + \boldsymbol{e} \quad \boldsymbol{e} \sim N(0, \sigma^2 I_n), \tag{1}$$

where $\boldsymbol{x}_1 = \boldsymbol{x}_2, \boldsymbol{x}_3 = \boldsymbol{x}_4$, $b_1 \neq 0, b_4 \neq 0, b_{j \notin \{1,4\}} = 0$.

Quantifying uncertainty in variable selection

Consider a sparse regression example

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{x}_j b_j + \boldsymbol{e} \quad \boldsymbol{e} \sim N(0, \sigma^2 I_n), \tag{1}$$

where $\boldsymbol{x}_1 = \boldsymbol{x}_2, \boldsymbol{x}_3 = \boldsymbol{x}_4$, $b_1 \neq 0, b_4 \neq 0, b_{j \notin \{1,4\}} = 0$.

We are interested in making the following statement,

$(b_1 \neq 0$ **or** $b_2 \neq 0)$ **AND** $(b_3 \neq 0$ **or** $b_4 \neq 0)$.

Quantifying uncertainty in variable selection

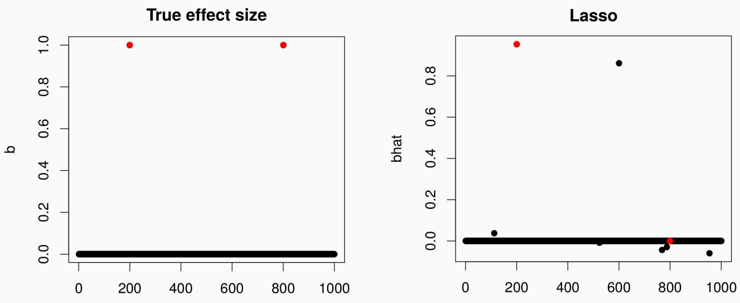
We are interested in making the following statement,

$(b_1 \neq 0$ **or** $b_2 \neq 0)$ **AND** $(b_3 \neq 0$ **or** $b_4 \neq 0)$.

1. There are two independent variables with non-zero effect
2. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (and $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$) are too similar to distinguish
3. yet they can be prioritized relative to each other

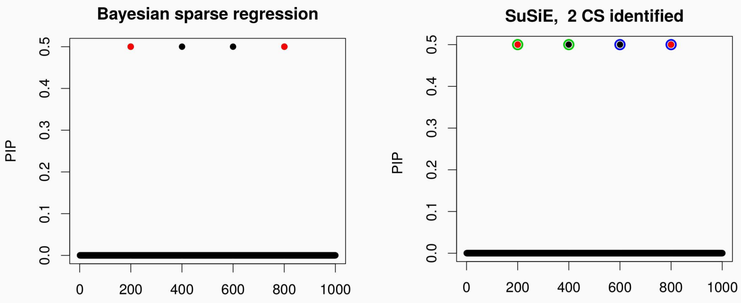
Quantifying uncertainty in variable selection

$$b_1 \neq 0 \text{ or } b_2 \neq 0, \text{ and } b_3 \neq 0 \text{ or } b_4 \neq 0.$$



Quantifying uncertainty in variable selection

$$b_1 \neq 0 \text{ or } b_2 \neq 0, \text{ and } b_3 \neq 0 \text{ or } b_4 \neq 0.$$

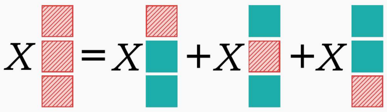


A simple Bayesian variable selection with applications to fine-mapping

The Sum of Single Effects model (*SuSiE*)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$$

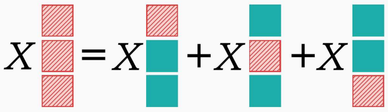


Wang *et al.* (2020) J. R. Stat. Soc. B

The Sum of Single Effects model (*SuSiE*)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$$



A variational approximation to posterior under SuSiE

$$q(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L) = \prod_l q_l(\boldsymbol{b}_l)$$

- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L$ are treated as **independent** *a posteriori*.
- **Do not** assume q_l factorizes over the elements of $\boldsymbol{b}_l$.

A fast Bayesian variable selection algorithm

Iterative Bayesian forward selection algorithm (IBSS)

- For each iteration t
  1. For each SNP j fit $\boldsymbol{y} = \boldsymbol{X}_j b_j^{(t)} + \boldsymbol{e}$ get $\text{BF}_j^{(t)}$
  2. Form weight for each SNP $w_j^{(t)} \propto \text{BF}_j^{(t)}$
  3. Form residuals $\boldsymbol{y}' := \boldsymbol{y} - \sum_j w_j^{(t)} \boldsymbol{X}_j \hat{b}_j^{(t)}$ and repeat
- Until converge

Coordinate ascent algorithm; convergence based on evidence lower bound (ELBO)

SuSiE model, formal notation

"single effect": $\boldsymbol{b}_l$'s

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(0, \sigma^2 I_n)$$
$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$$
$$\boldsymbol{b}_l = \gamma_l \beta_l$$
$$\gamma_l \sim \text{Mult}(1, \boldsymbol{\pi})$$
$$\beta_l \sim N(0, \sigma_{0_l}^2)$$
$$\sigma_{0_l}^2 \geq 0$$

A mean-field approximation

$$q(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L) = \prod_l q_l(\boldsymbol{b}_l)$$

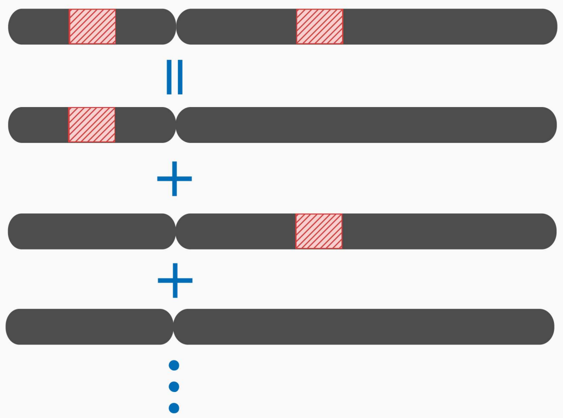
- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L$ are treated as **independent** *a posteriori*.
- **Do not** assume q_l factorizes over the elements of $\boldsymbol{b}_l$.

IBSS algorithm, formal notation

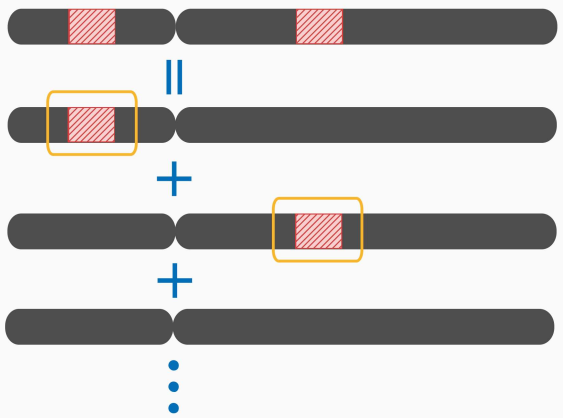
Algorithm Iterative Bayesian forward selection

```
Require: data $\boldsymbol{y}$ and variable matrix $\boldsymbol{X}$.
Require: Single Effect Regression: $\text{SER}(\boldsymbol{y}, \boldsymbol{X}) \rightarrow (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \sigma_1^2)$
Initialize $\boldsymbol{\alpha}_l, \boldsymbol{\mu}_l, \bar{\boldsymbol{b}}_l$ for $l = 1, \ldots, L$.
repeat
    for $l$ in $1, \ldots, L$ do
        $\boldsymbol{r}_l \leftarrow \boldsymbol{y} - \sum_{l' \neq l} \boldsymbol{X}\bar{\boldsymbol{b}}_{l'}$
        $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_l, \sigma_l^2) \leftarrow \text{SER}(\boldsymbol{r}_l, \boldsymbol{X})$
        $\bar{\boldsymbol{b}}_l \leftarrow \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_l$
until converged
return $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_L$.
```

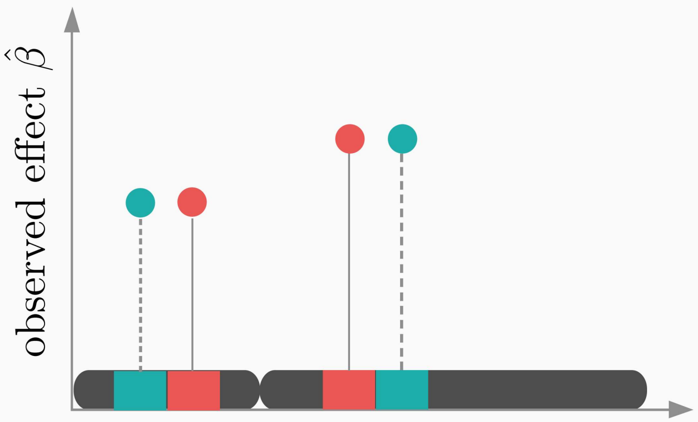
SuSiE model yields single-effect CS



SuSiE model yields single-effect CS



IBSS algorithm illustration



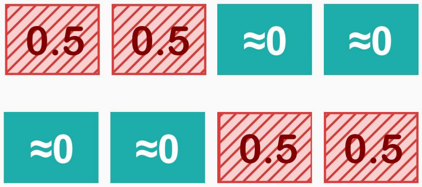
IBSS algorithm illustration

1. At random (zero) initialization, fit single effect model on $\boldsymbol{y}$

0.5	0.5	≈0	≈0

IBSS algorithm illustration

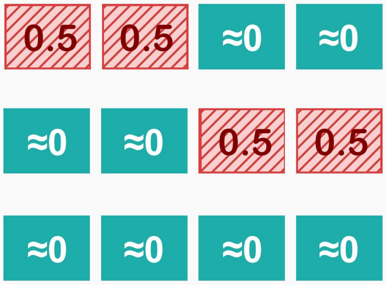
2. Compute residual r_2 using fitted model, and do it again



0.5	0.5	≈0	≈0
≈0	≈0	0.5	0.5

IBSS algorithm illustration

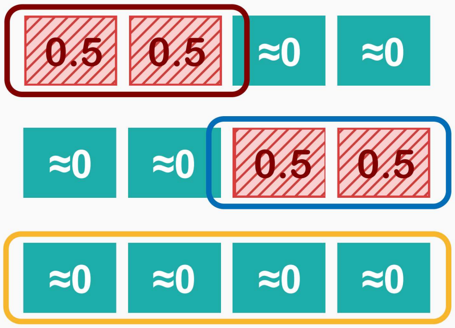
3. Compute residual r_3 using fitted model, and do it again



0.5	0.5	≈0	≈0
≈0	≈0	0.5	0.5
≈0	≈0	≈0	≈0

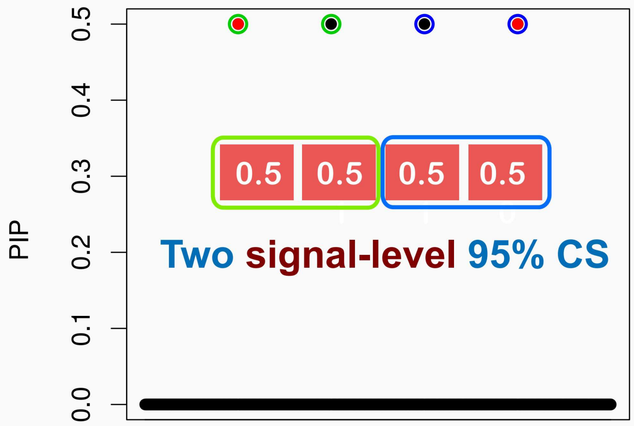
IBSS algorithm illustration

4. Iterate until converge; compute **single-effect credible sets**



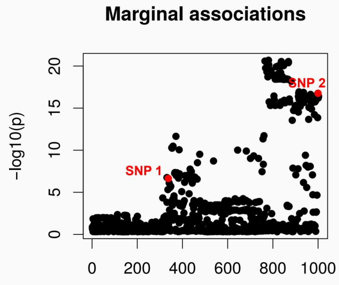
0.5	0.5	≈0	≈0
≈0	≈0	0.5	0.5
≈0	≈0	≈0	≈0

SuSiE CS illustration

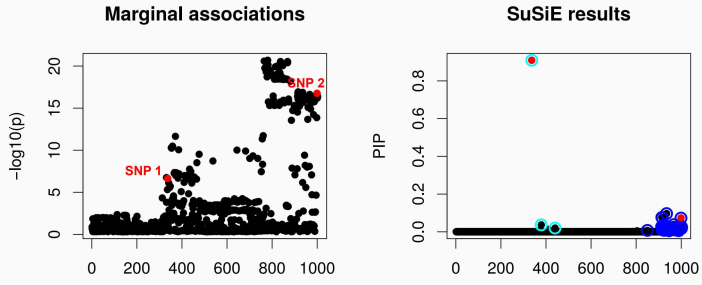


Two signal-level 95% CS

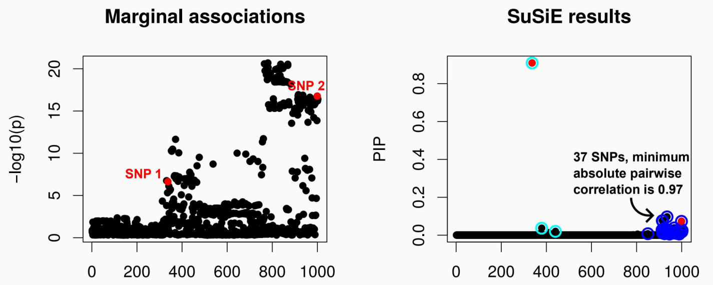
Real-world example illustrated



Real-world example illustrated

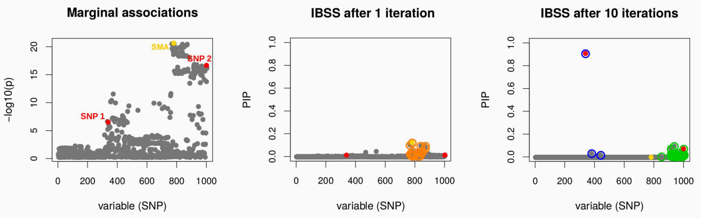


Real-world example illustrated

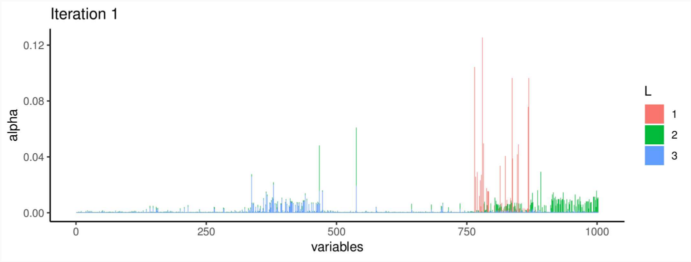


37 SNPs, minimum absolute pairwise correlation is 0.97

Real-world example illustrated



The IBSS algorithm iterations breakdown



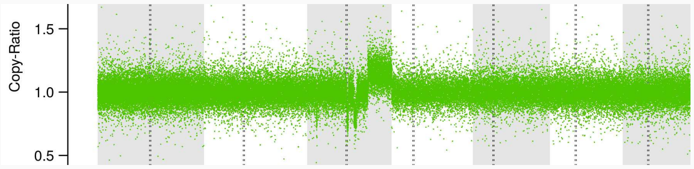
Other variable selection problems in genetics

Similar model, different problems

- X is gene expression, y is tissue / cell type
- X is pathway, y is gene-set
- X is functional annotation, y is GWAS effect size
- **X is "step matrix", y is spatially-structured variable**

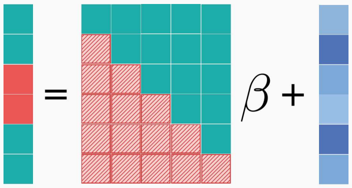
The "changepoint" problem

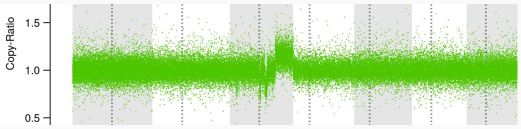
Data is piecewise constant, *e.g.* copy number variation



The “changepoint” problem

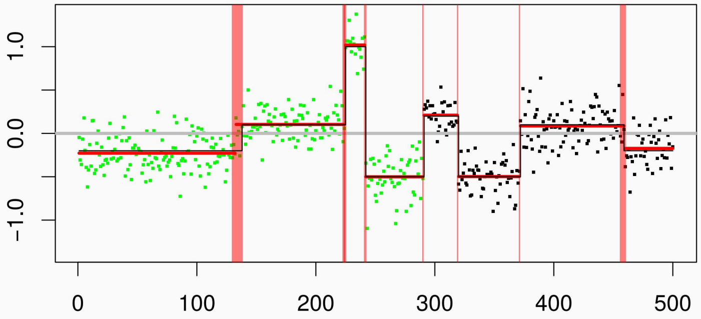
Can be modelled as linear combination of step functions





Example: simulated DNA copy number variation

SuSiE vs Circular Binary Segmentation Olshen *et al.* (2004) *Biostatistics*



Notice the benefit of quantifying uncertainty in this example

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

Gao Wang, Ph.D.

Advanced Gene Mapping Course, May 2023

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

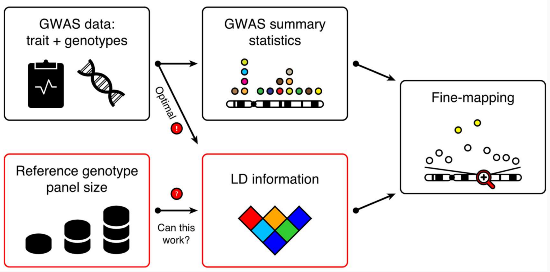


Figure: Benner *et al.* (2017) Am. J. Hum. Genet.

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)

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

Assumptions:

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

Properties of per SNP z scores

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

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

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

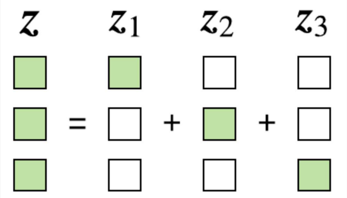
$$\hat{z} \sim N(\hat{R}z, \hat{R})$$

$$z = \sum_{l=1}^{L} z_l$$

$$z_l = \gamma_l z_l$$

$$z_l \sim N(0, \omega_l^2)$$

$$\gamma_l \sim \text{Mult}(1, \pi)$$



Suggested reading:

Zou et al (2022) PLoS Genet.

$\hat{\beta}$ and SE($\hat{\beta}$) based models

The $\hat{z}$ model:

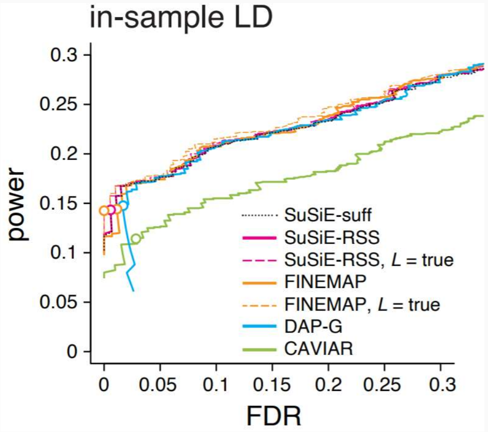
$$\hat{z} \sim N(\hat{R}z, \hat{R})$$

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

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

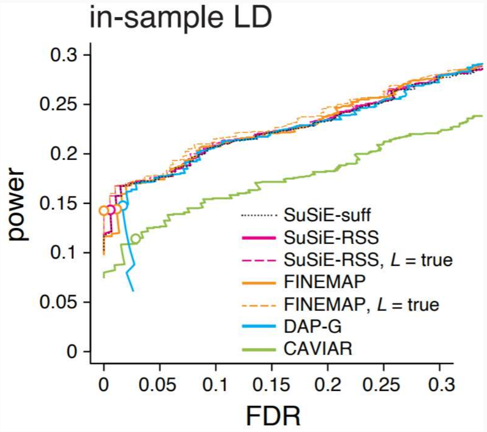
- Both models can be easily written as SuSiE regression
  - $\hat{z}$ model: lower MAF variants have larger effects
  - $\hat{b}, \hat{s}$ model: effect sizes are the same regardless of MAF
- $\hat{b}, \hat{s}$ model takes sample size into consideration
  - No longer have to assume small effect per SNP
- $\hat{z}$ model: CAVIAR, FINEMAP (2016)
- $\hat{b}, \hat{s}$ model: FINEMAP (2018), SuSiE_RSS

Summary statistics methods comparison



Zou *et al.* (2022) PLoS Genet

Summary statistics methods comparison



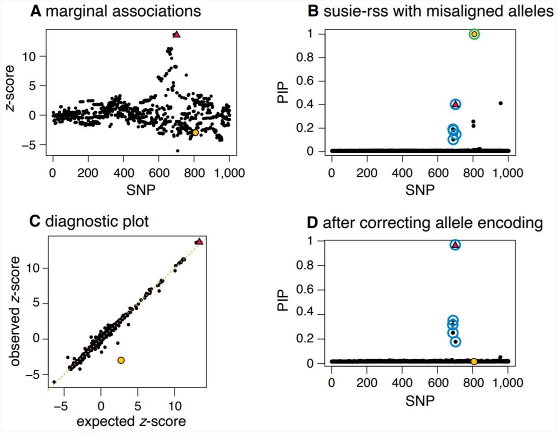
Zou *et al.* (2022) PLoS Genet.

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

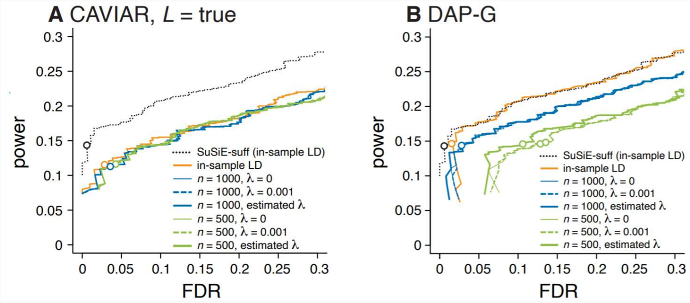


Zou *et al.* (2022) PLoS Genet.

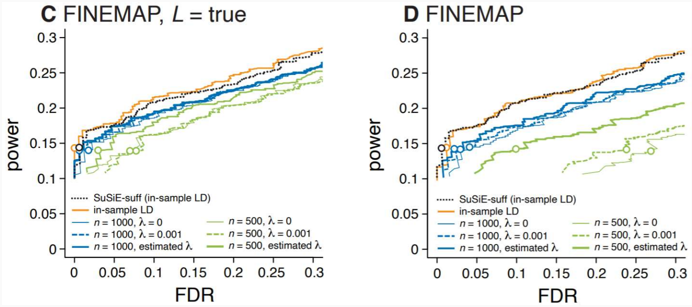
Addressing the allele flip challenge

- `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`

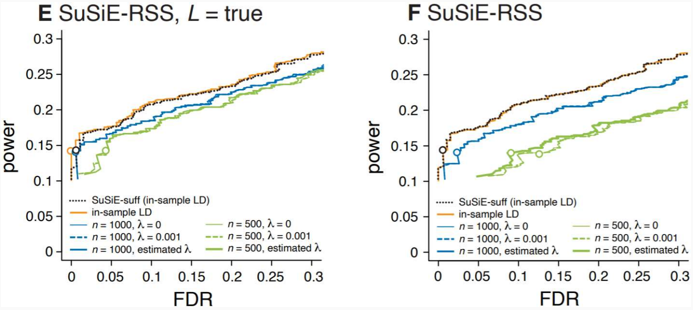
Impact of mis-matched LD reference: PIP



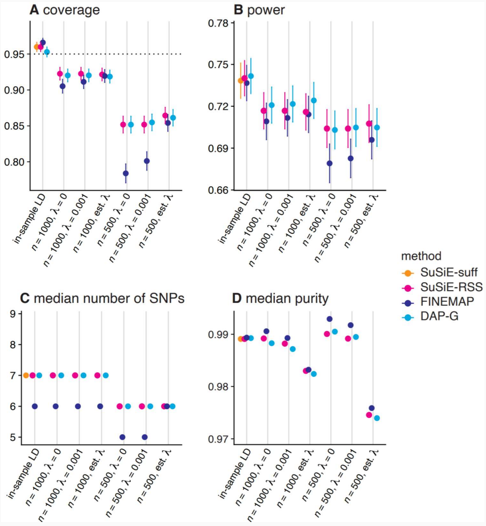
Impact of mis-matched LD reference: PIP



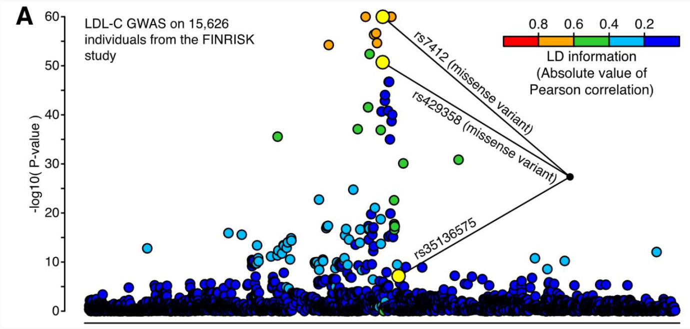
Impact of mis-matched LD reference: PIP



Impact of mis-matched LD reference: credible sets

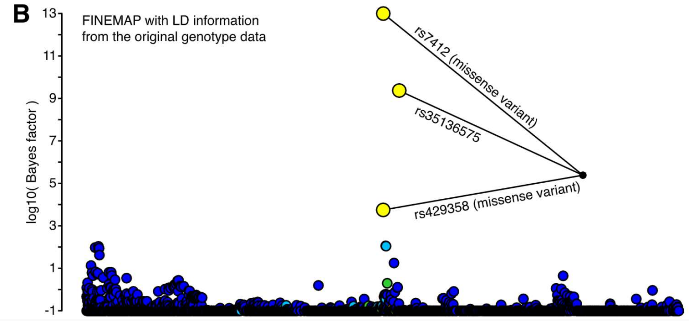


Impact of mis-matched LD reference: real data



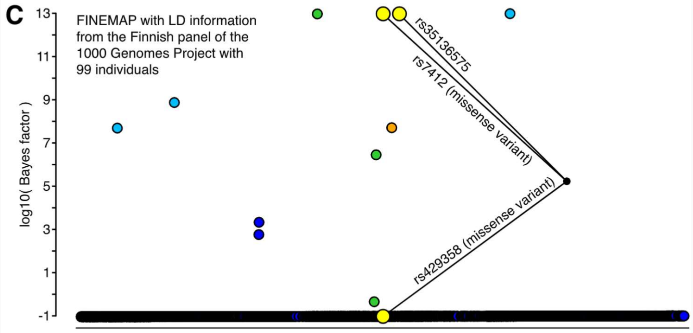
Benner *et al.* (2017) Am. J. Hum. Genet.

Impact of mis-matched LD reference: real data



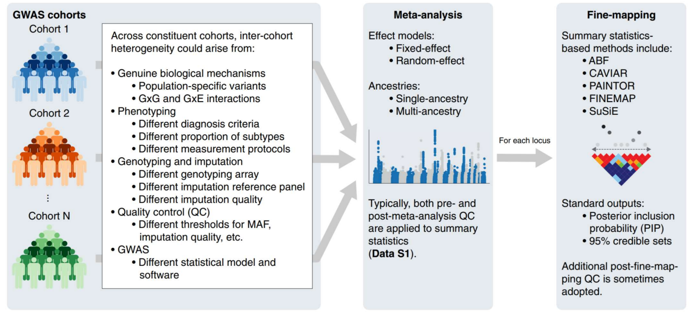
Benner *et al.* (2017) Am. J. Hum. Genet.

Impact of mis-matched LD reference: real data



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

Fine-mapping in meta-analysis: overview



GWAS cohorts

Cohort 1

Cohort 2

⋮

Cohort N

Across constituent cohorts, inter-cohort heterogeneity could arise from:

- Genuine biological mechanisms
  - Population-specific variants
  - GxG and GxE interactions
- Phenotyping
  - Different diagnosis criteria
  - Different proportion of subtypes
  - Different measurement protocols
- Genotyping and imputation
  - Different genotyping array
  - Different imputation reference panel
  - Different imputation quality
- Quality control (QC)
  - Different thresholds for MAF, imputation quality, etc.
- GWAS
  - Different statistical model and software

Meta-analysis

Effect models:
- Fixed-effect
- Random-effect

Ancestries:
- Single-ancestry
- Multi-ancestry

Typically, both pre- and post-meta-analysis QC are applied to summary statistics (**Data S1**).

For each locus

Fine-mapping

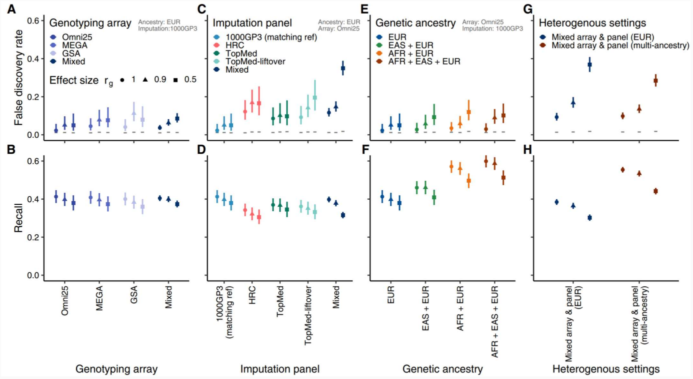
Summary statistics-based methods include:
- ABF
- CAVIAR
- PAINTOR
- FINEMAP
- SuSiE

Standard outputs:
- Posterior inclusion probability (PIP)
- 95% credible sets

Additional post-fine-mapping QC is sometimes adopted.

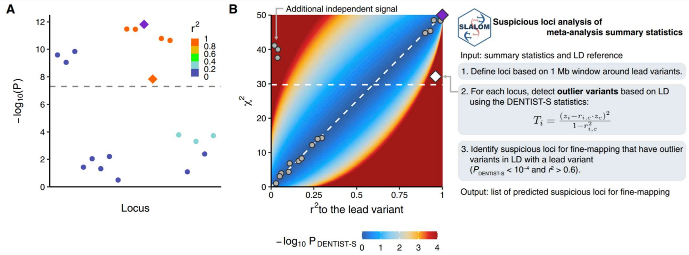
Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: key factors



Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: diagnosis



Suspicious loci analysis of meta-analysis summary statistics

Input: summary statistics and LD reference

1. Define loci based on 1 Mb window around lead variants.
2. For each locus, detect **outlier variants** based on LD using the DENTIST-S statistics:

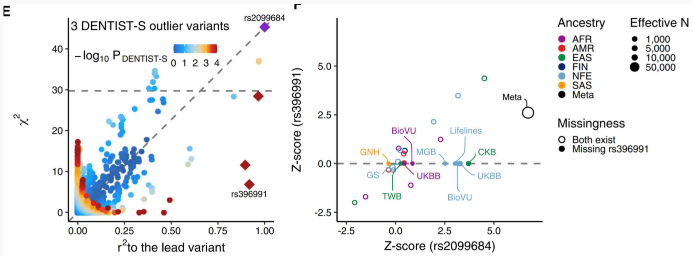
$$T_i = \frac{(z_i - r_{i,c} \cdot z_c)^2}{1 - r_{i,c}^2}$$

3. Identify suspicious loci for fine-mapping that have outlier variants in LD with a lead variant ($P_{\text{DENTIST-S}} < 10^{-4}$ and $r^2 > 0.6$).

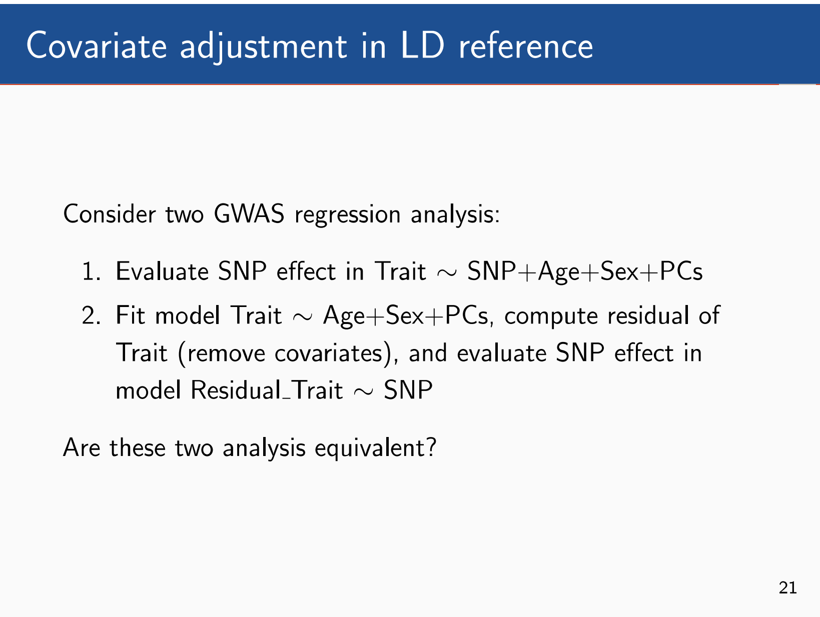
Output: list of predicted suspicious loci for fine-mapping

Chen *et al.* (2021) Nat. Comm. (DENTIST)
Kanai *et al.* (2022) Cell Genomics

Fine-mapping in meta-analysis: diagnosis



Kanai *et al.* (2022) Cell Genomics

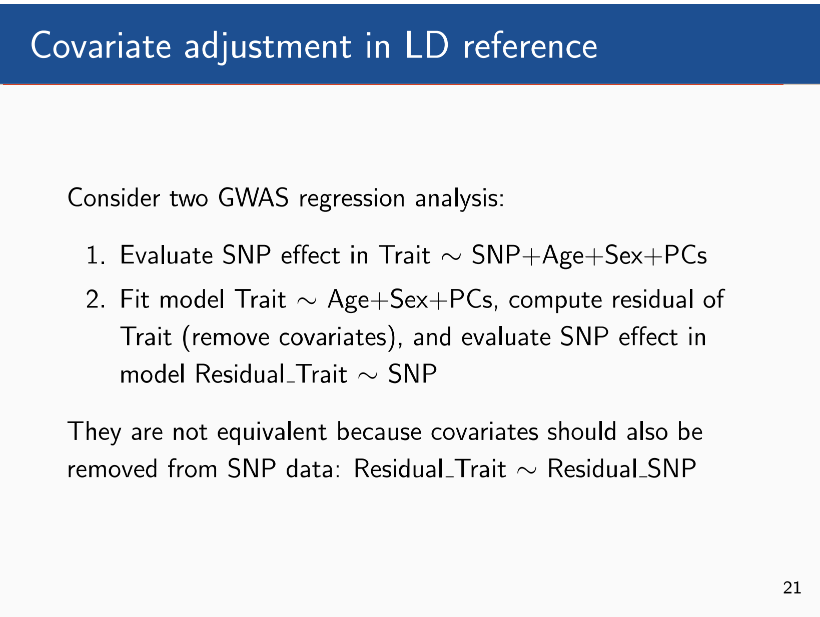


Covariate adjustment in LD reference

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?

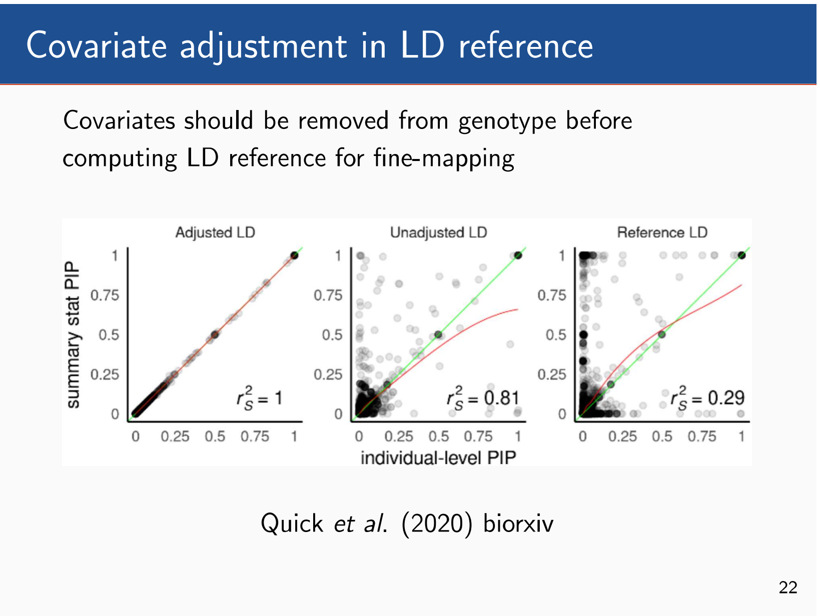


Covariate adjustment in LD reference

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

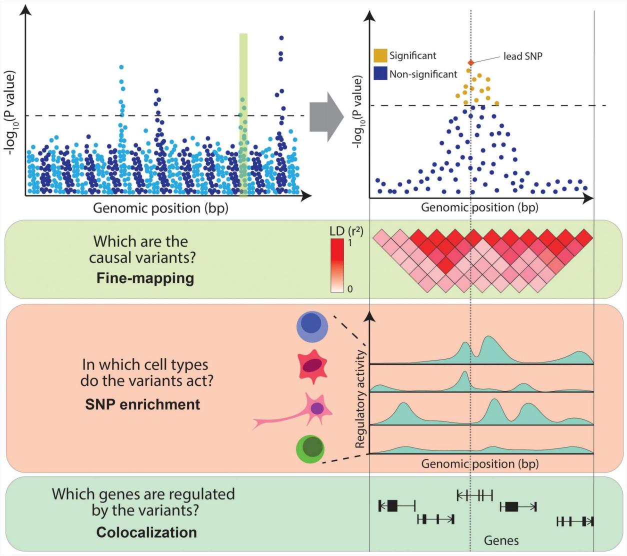
Quick *et al.* (2020) biorxiv

Integrating GWAS with functional annotations

Gao Wang, Ph.D.

Advanced Gene Mapping Course, May 2023

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



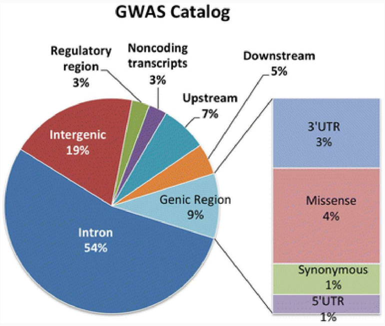
Which are the causal variants?
Fine-mapping

In which cell types do the variants act?
SNP enrichment

Which genes are regulated by the variants?
Colocalization

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

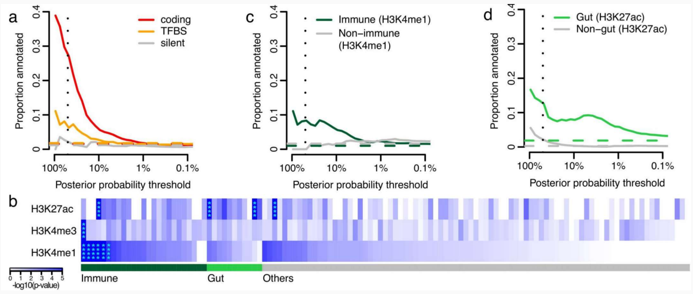


Figure: Huang *et al.* (2017) Nature

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

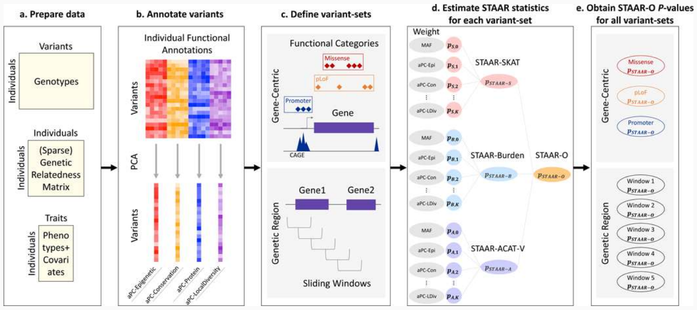


Figure: Li *et al.* (2020) Nature Genetics

Also see Li *et al.* (2019) AJHG; Li *et al.* (2022) Nature Methods

Annotations integrated to aggregated tests

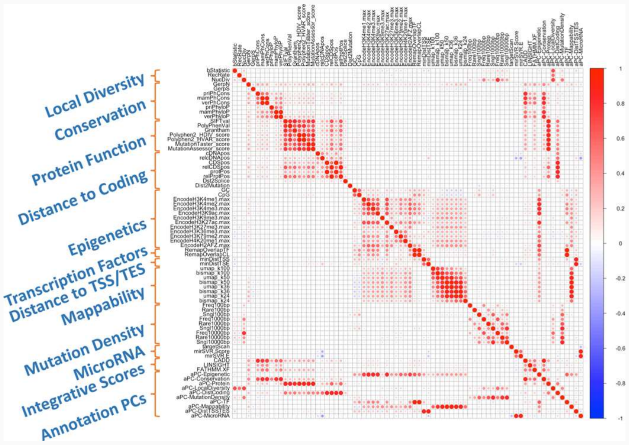


Figure: Li *et al.* (2020) Nature Genetics

Functional annotation in common variant association analysis

A polygenic model: stratified LD score regression

Chi-square GWAS statistic of variant j

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

A polygenic model: stratified LD score regression

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

- 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

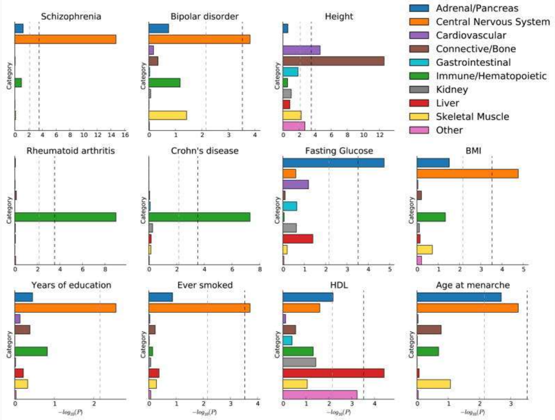


Figure: Finucane *et al.* (2015) Nature Genetics

Integration approaches

- Integrate directly as range based binary annotations
  - Finucane et al (2015) Nature Genetics — Stratified LDSC paper
- Extension: variant specific continuous annotations
  - Gazal et al (2017) Nature Genetics
- Tissue specific variant level annotations independent of GWAS results
  - 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

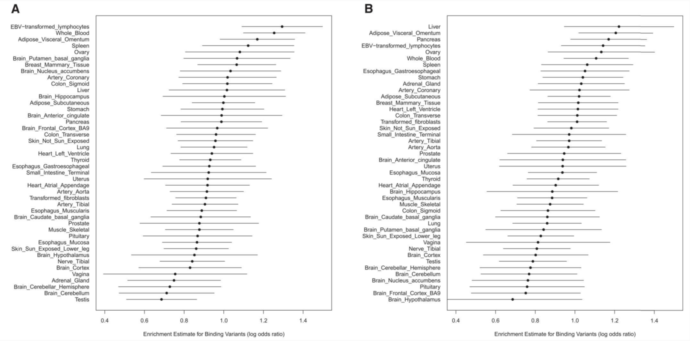
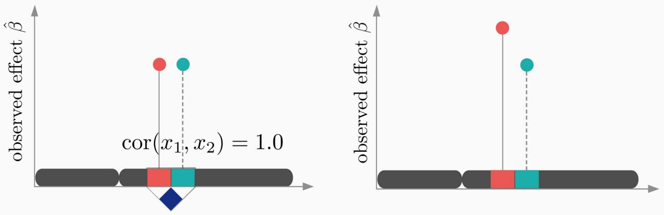


Figure: Wen *et al.* (2016) AJHG

Integrative fine-mapping with functional annotations

Annotations improves fine-mapping resolution


observed effect $\hat{\beta}$
$\mathrm{cor}(x_1, x_2) = 1.0$
observed effect $\hat{\beta}$

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**,

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

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

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

Fine-mapping with functional annotations

Recall the BVSR model

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

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.

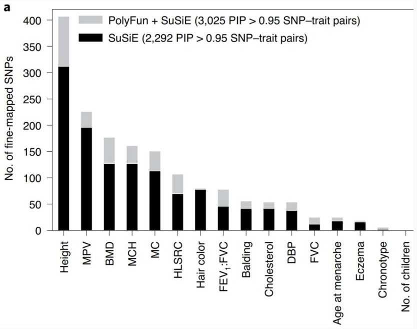


Figure: Weissbrod *et al.* (2020) Nat. Genet.

Example: *SuSiE* with functional informed prior

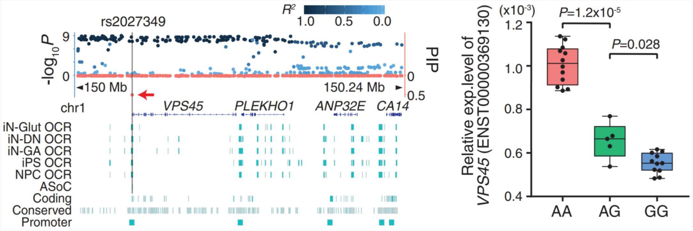


Figure: Zhang *et al.* (2020) Science

Caution: disease specific enrichment

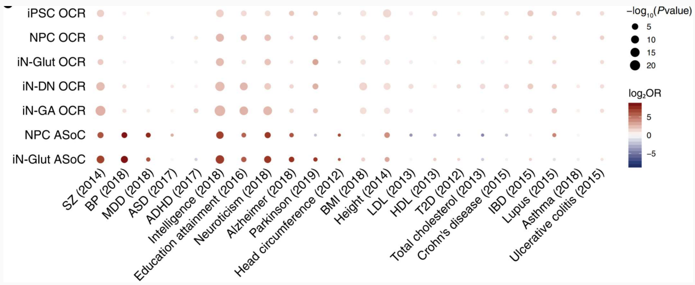


Figure: Zhang *et al.* (2020) Science

Complex phenotype prediction and transcriptome-wide association studies

Gao Wang, Ph.D.

Advanced Gene Mapping Course, May 2023

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

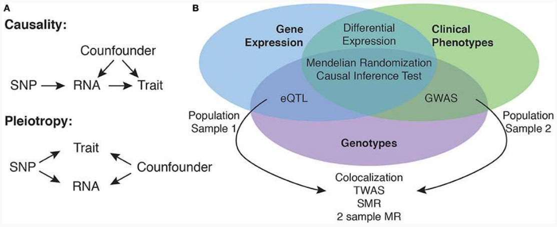


Figure: Heinig (2018) Front. Cardiovasc. Med.

Transcriptome-wide association study (TWAS)

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

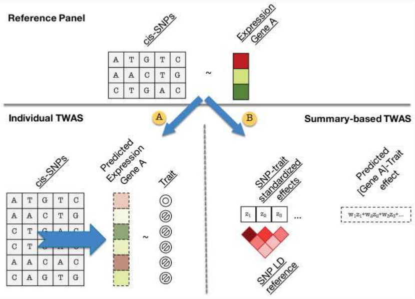
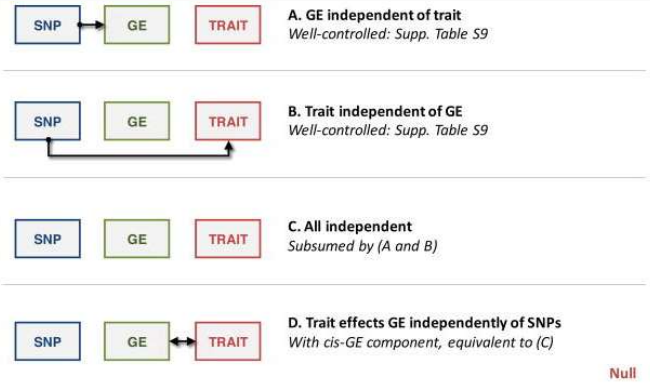


Figure: Gusev *et al.* (2016) Nat. Genet.

TWAS challenge: association vs causality



A. GE independent of trait
Well-controlled: Supp. Table S9

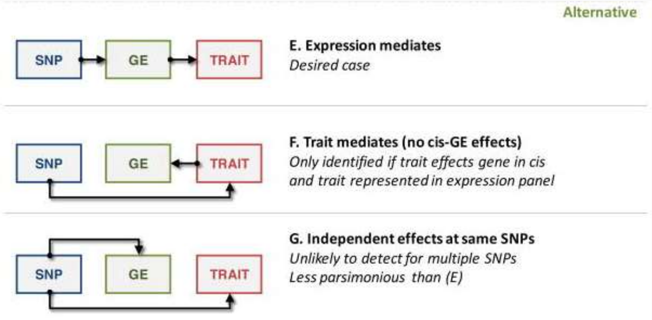
B. Trait independent of GE
Well-controlled: Supp. Table S9

C. All independent
Subsumed by (A and B)

D. Trait effects GE independently of SNPs
With cis-GE component, equivalent to (C)

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

TWAS challenge: association vs causality



E. Expression mediates
Desired case

F. Trait mediates (no cis-GE effects)
Only identified if trait effects gene in cis and trait represented in expression panel

G. Independent effects at same SNPs
Unlikely to detect for multiple SNPs
Less parsimonious than (E)

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

TWAS challenge: technical considerations

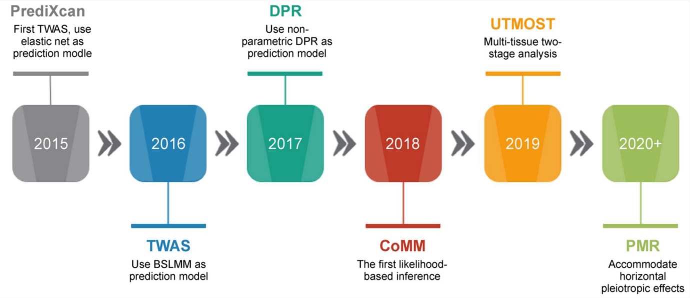
Ideal TWAS setup

- 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



PrediXcan First TWAS, use elastic net as prediction modle

DPR Use non-parametric DPR as prediction model

UTMOST Multi-tissue two-stage analysis

TWAS Use BSLMM as prediction model

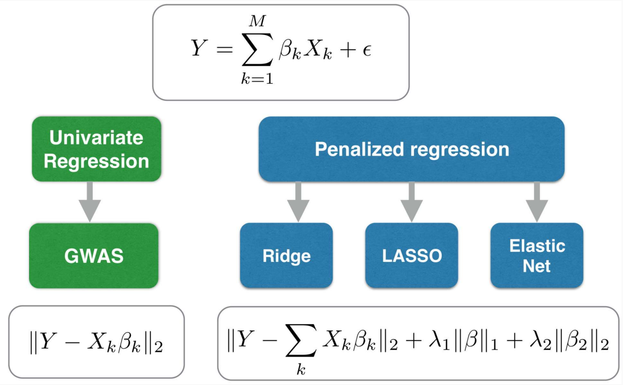
CoMM The first likelihood-based inference

PMR Accommodate horizontal pleiotropic effects

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

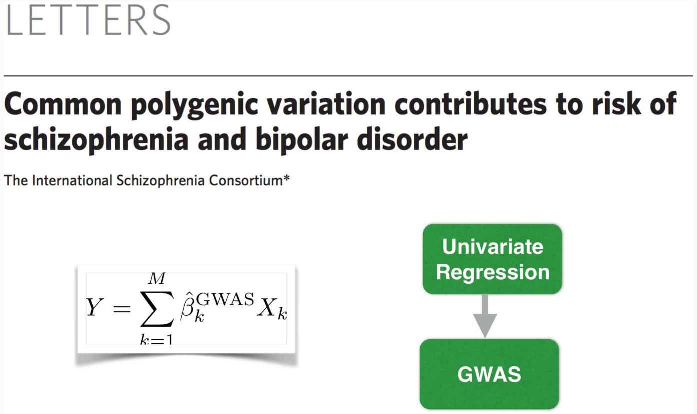
$$\|Y - X_k\beta_k\|_2$$

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

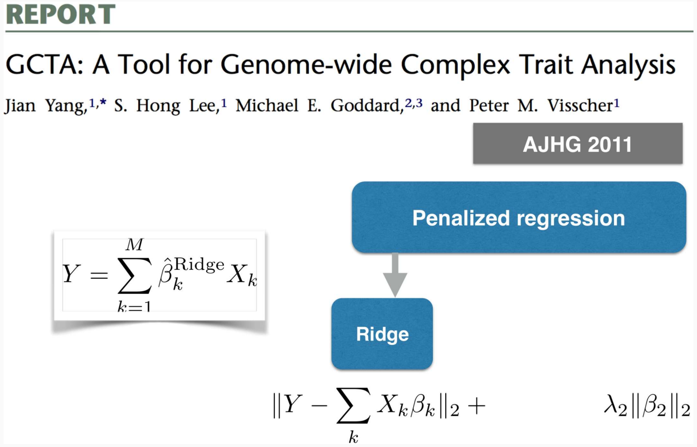
Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

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

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

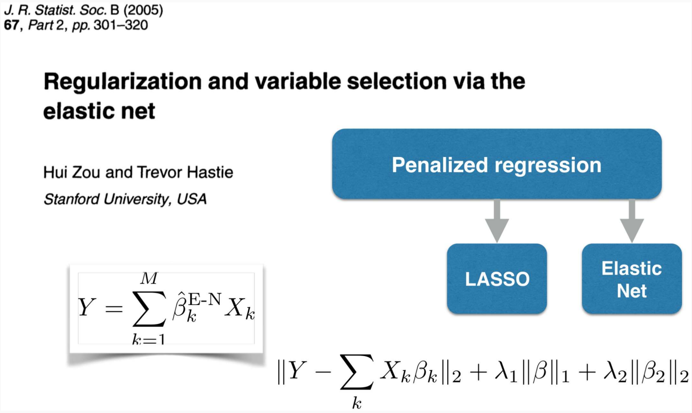
Penalized regression

$$Y = \sum_{k=1}^{M} \hat\beta_k^{\text{Ridge}} X_k$$

Ridge

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

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

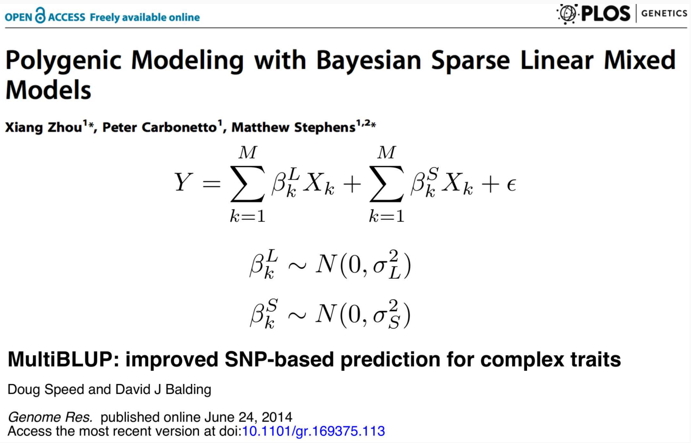
LASSO

Elastic Net

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

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

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*]

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

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

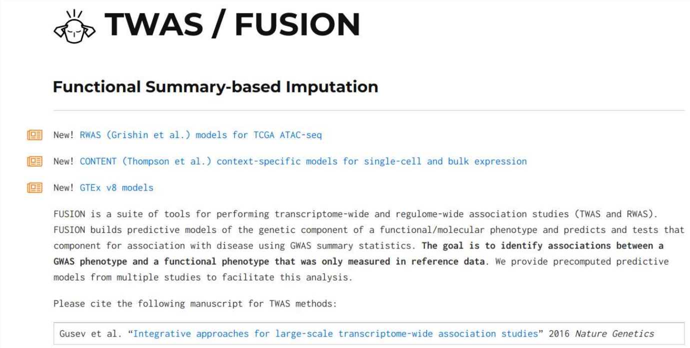
$$\beta_k^S \sim N(0, \sigma_S^2)$$

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

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

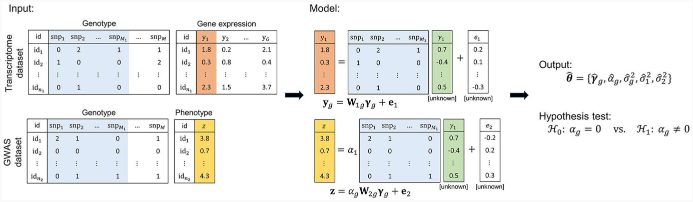
New! GTEx v8 models

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*

Likelihood based approach



Input:

Transcriptome dataset

Genotype

id	snp_1	snp_2	...	snp_{M_1}	...	snp_M
id_1	0	2		1		1
id_2	1	0		0		2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
id_{n_1}	0	1		0		0

Gene expression

id	y_1	y_2	...	y_G
id_1	1.8	0.2		2.1
id_2	0.3	0.8		0.4
⋮	⋮	⋮	⋮	⋮
id_{n_1}	2.3	1.5		3.7

GWAS dataset

Genotype

id	snp_1	snp_2	...	snp_{M_1}	...	snp_M
id_1	2	1		0		1
id_2	0	0		0		0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
id_{n_2}	0	1		1		1

Phenotype

id	z
id_1	3.8
id_2	0.7
⋮	⋮
id_{n_2}	4.3

Model:

$$\begin{bmatrix} y_1 \\ 1.8 \\ 0.3 \\ \vdots \\ 2.3 \end{bmatrix} = \begin{bmatrix} snp_1 & snp_2 & \cdots & snp_{M_1} \\ 0 & 2 & & 1 \\ 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & & 0 \end{bmatrix} \underset{\text{[unknown]}}{\begin{bmatrix} \gamma_1 \\ 0.7 \\ -0.4 \\ \vdots \\ 0.5 \end{bmatrix}} + \underset{\text{[unknown]}}{\begin{bmatrix} e_1 \\ 0.2 \\ 0.1 \\ \vdots \\ -0.3 \end{bmatrix}}$$

$$\mathbf{y}_g = \mathbf{W}_{1g}\boldsymbol{\gamma}_g + \mathbf{e}_1$$

$$\begin{bmatrix} z \\ 3.8 \\ 0.7 \\ \vdots \\ 4.3 \end{bmatrix} = \alpha_1 \begin{bmatrix} snp_1 & snp_2 & \cdots & snp_{M_1} \\ 2 & 1 & & 0 \\ 0 & 0 & & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & & 1 \end{bmatrix} \underset{\text{[unknown]}}{\begin{bmatrix} \gamma_1 \\ 0.7 \\ -0.4 \\ \vdots \\ 0.5 \end{bmatrix}} + \underset{\text{[unknown]}}{\begin{bmatrix} e_2 \\ -0.2 \\ 0.2 \\ \vdots \\ 0.3 \end{bmatrix}}$$

$$\mathbf{z} = \alpha_g \mathbf{W}_{2g}\boldsymbol{\gamma}_g + \mathbf{e}_2$$

Output:
$\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\gamma}}_g, \hat\alpha_g, \hat\sigma_g^2, \hat\sigma_1^2, \hat\sigma_2^2\}$

Hypothesis test:
$\mathcal{H}_0\colon \alpha_g = 0$ vs. $\mathcal{H}_1\colon \alpha_g \neq 0$

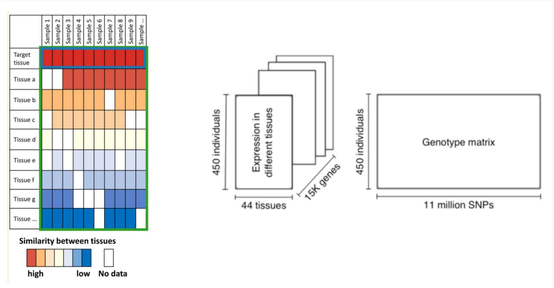
Figure: CoMM, Yeung *et al.* (2019)

Also see Yuan *et al.* (2022) likelihood based Mendelian Randomization

Multivariate TWAS methods

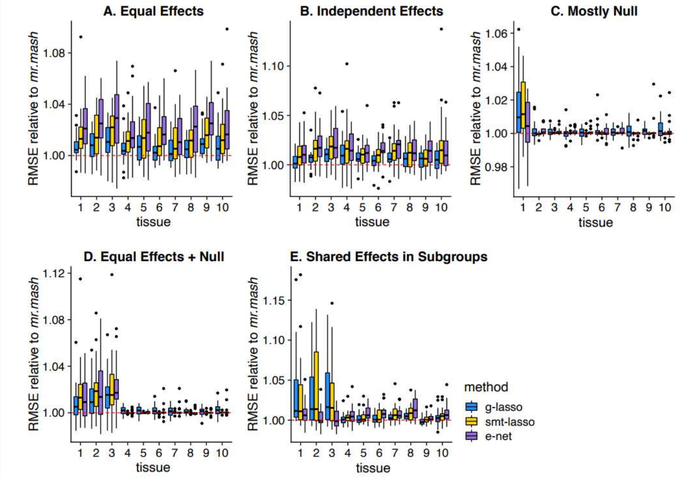
Multivariate TWAS methods overview

Leverage similarity between molecular phenotypes



- UTMOST, Yu *et al.* (2019) Nature Genetics
- MR-JTI, Zhou *et al.* (2020) Nature Genetics
- mr.mash, Morgante *et al.* (2023) PLoS Genetic (to appear)

Multivariate TWAS method: mr.mash

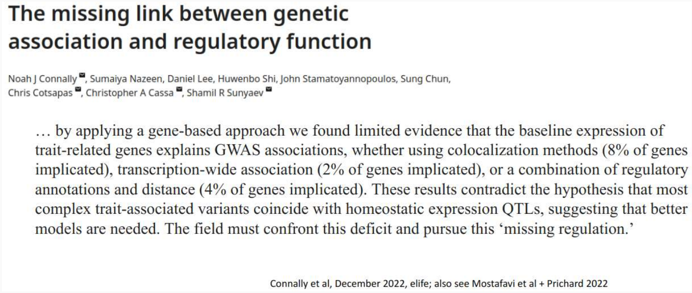


Multivariate TWAS hands-on exercise

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

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

Missing regulation in eQTL and GWAS



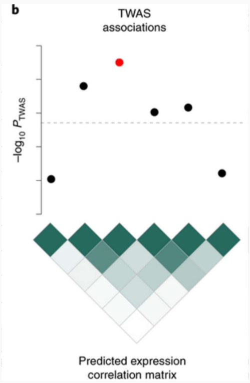
The missing link between genetic association and regulatory function

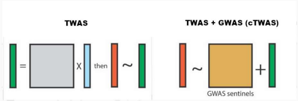
Noah J Connally ✉, Sumaiya Nazeen, Daniel Lee, Huwenbo Shi, John Stamatoyannopoulos, Sung Chun, Chris Cotsapas ✉, Christopher A Cassa ✉, Shamil R Sunyaev ✉

> … by applying a gene-based approach we found limited evidence that the baseline expression of trait-related genes explains GWAS associations, whether using colocalization methods (8% of genes implicated), transcription-wide association (2% of genes implicated), or a combination of regulatory annotations and distance (4% of genes implicated). These results contradict the hypothesis that most complex trait-associated variants coincide with homeostatic expression QTLs, suggesting that better models are needed. The field must confront this deficit and pursue this 'missing regulation.'

Connally et al, December 2022, elife; also see Mostafavi et al + Prichard 2022

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

TWAS and fine-mapping: variable selection

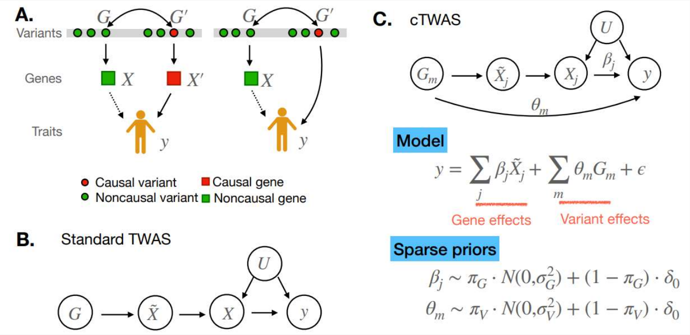

A.
Variants
G
G'
G
G'
Genes
X
X'
X
Traits
y
y
● Causal variant
🟥 Causal gene
● Noncausal variant
🟩 Noncausal gene
B.
Standard TWAS
U
G
$\tilde{X}$
X
y
C.
cTWAS
U
G_m
$\tilde{X}_j$
X_j
β_j
y
θ_m
Model
$y = \underbrace{\sum_j \beta_j \tilde{X}_j}_{\text{Gene effects}} + \underbrace{\sum_m \theta_m G_m}_{\text{Variant effects}} + \epsilon$
Sparse priors
$\beta_j \sim \pi_G \cdot N(0,\sigma_G^2) + (1-\pi_G)\cdot\delta_0$
$\theta_m \sim \pi_V \cdot N(0,\sigma_V^2) + (1-\pi_V)\cdot\delta_0$

Figure: Zhao *et al.* (2022) biorxiv

TWAS and colocalization: pleiotropy

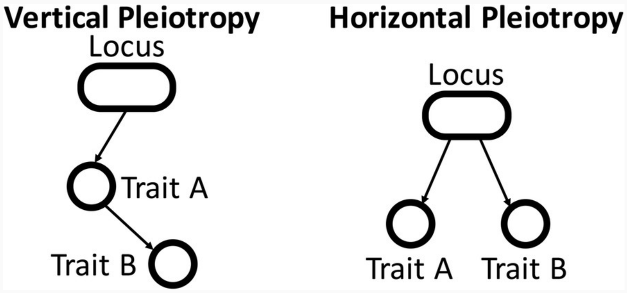
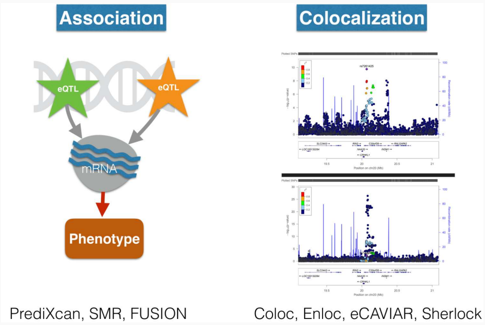


Figure: Jordan *et al.* (2019) Genome Biology

TWAS and colocalization: pleiotropy



- Image credit: Haky Im @ UChicago
- "Locus level" colocalization: Hukku *et al.* (2022) AJHG; Okamoto *et al.* (2023) AJHG.

TWAS and colocalization: statistical framework

$$\boldsymbol{M} = \mu_M \mathbf{1} + \boldsymbol{G}\boldsymbol{\beta}_E + \mathrm{e}_M, \mathrm{e}_M \sim \mathrm{N}\left(\boldsymbol{0}, \sigma_M^2 \boldsymbol{I}\right)$$
$$\boldsymbol{Y} = \mu_Y \mathbf{1} + \gamma \boldsymbol{M} + \boldsymbol{G}\boldsymbol{\beta}_Y + \mathrm{e}_Y, \mathrm{e}_Y \sim \mathrm{N}\left(\boldsymbol{0}, \sigma_Y^2 \boldsymbol{I}\right)$$

- "locus level", $Pr(\gamma \neq 0|\text{Data}) \propto Pr(\gamma \neq 0)Pr(\text{Data})$
- $\Pr(\gamma \neq 0) = Pr(coloc) \times Pr(twas)$
- Data: z-score from TWAS.
- Key idea: Test $\gamma = 0$, not to estimate γ which is Mendelian Randomization.

TWAS and Mendelian randomization

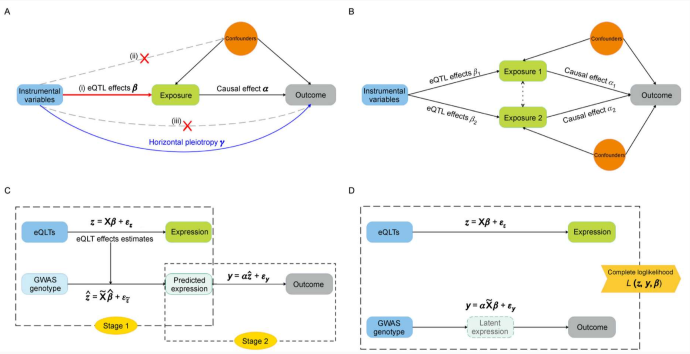


Figure: Zhu and Zhou (2022) Quantitative Biology

TWAS can be viewed as two-sample MR — using various IV selection methods.

Multivariate analysis in genetic association studies

Gao Wang, Ph.D.

Advanced Gene Mapping Course, May 2023

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

1. Motivation
2. Meta-analysis review
3. Meta-analysis: a multivariate regression prospective
4. Variant colocalization: variable selection in meta-analysis
5. Multivariate adaptive shrinkage and fine-mapping

Motivation

Beyond per trait per variant association studies

Statistical fine-mapping (multiple regressors)

- Identify non-zero effect variables by accounting for LD

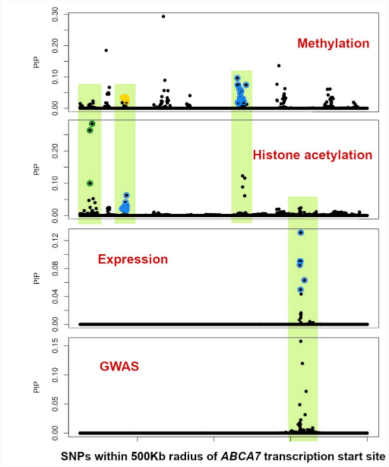
Meta-analysis (multiple responses)

- Integrate information across multiple conditions / studies

"Causal" variants across multiple conditions?

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

The problem



The problem

For a genetic variable analyzed in two conditions:

$$P(\textbf{“causal” in trait 1 \& 2} \mid \text{association data for 1 \& 2})$$

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:

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

Multivariate relationships?

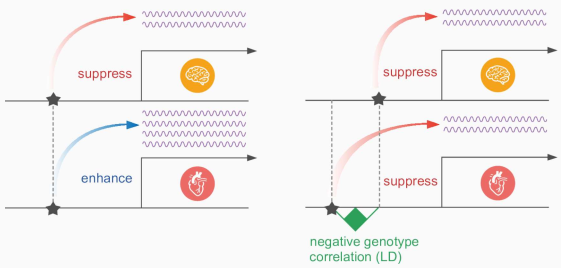


Figure: Pleiotropy or Linkage?

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

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

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

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

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

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

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

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

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,

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

Random effects model multivariate analysis

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

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

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!

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

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:

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

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 → Trait 1 → Trait 2.
- Effect on trait 2 should be smaller than that on trait 1.

Variant colocalization: variable selection in meta-analysis

The problem

For a genetic variable analyzed in GWAS and eQTL studies:

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

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 between 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 between 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/`

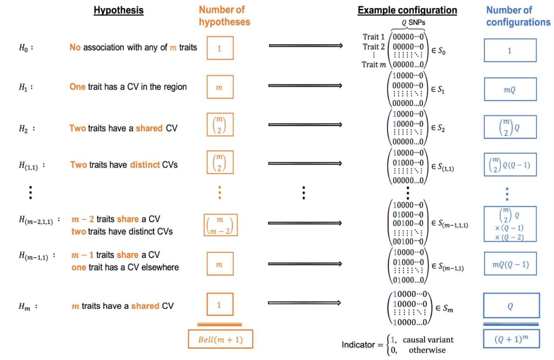
Summary: **pattern** and **scale** of effect size correlations, represented as different **prior** models.

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.

Multi-trait colocalization



Hypothesis		Number of hypotheses	Example configuration	Number of configurations
H_0 :	No association with any of m traits	1	Q SNPs Trait 1, Trait 2, ⋮, Trait m: $\begin{pmatrix} 00000\cdots0 \\ 00000\cdots0 \\ \vdots \\ 00000\ldots0 \end{pmatrix} \in S_0$	1
H_1 :	One trait has a CV in the region	m	$\begin{pmatrix} 10000\cdots0 \\ 00000\cdots0 \\ \vdots \\ 00000\ldots0 \end{pmatrix} \in S_1$	mQ
H_2 :	Two traits have a shared CV	$\binom{m}{2}$	$\begin{pmatrix} 10000\cdots0 \\ 10000\cdots0 \\ \vdots \\ 00000\ldots0 \end{pmatrix} \in S_2$	$\binom{m}{2}Q$
$H_{(1,1)}$:	Two traits have distinct CVs	$\binom{m}{2}$	$\begin{pmatrix} 10000\cdots0 \\ 01000\cdots0 \\ \vdots \\ 00000\ldots0 \end{pmatrix} \in S_{(1,1)}$	$\binom{m}{2}Q(Q-1)$
⋮		⋮	⋮	⋮
$H_{(m-2,1,1)}$:	$m-2$ traits share a CV two traits have distinct CVs	$\binom{m}{m-2}$	$\begin{pmatrix} 10000\cdots0 \\ 01000\cdots0 \\ 00100\cdots0 \\ \vdots \\ 00100\cdots0 \end{pmatrix} \in S_{(m-1,1,1)}$	$\binom{m}{2}Q \times (Q-1) \times (Q-2)$
$H_{(m-1,1)}$:	$m-1$ traits share a CV one trait has a CV elsewhere	m	$\begin{pmatrix} 10000\cdots0 \\ 01000\cdots0 \\ \vdots \\ 01000\ldots0 \end{pmatrix} \in S_{(m-1,1)}$	$mQ(Q-1)$
H_m :	m traits have a shared CV	1	$\begin{pmatrix} 10000\cdots0 \\ 10000\cdots0 \\ \vdots \\ 10000\ldots0 \end{pmatrix} \in S_m$	Q
		$Bell(m+1)$	Indicator $= \begin{cases} 1, & \text{causal variant} \\ 0, & \text{otherwise} \end{cases}$	$(Q+1)^m$

Figure: `HyPrColoc`, Foley *et al.* (2021) Nat. Comm.

Assuming a single causal variant in the loci.

Multivariate adaptive shrinkage and fine-mapping

More phenotypes, more complications

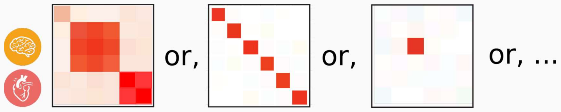


Figure: Plausible patterns of sharing

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":

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

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, $\widehat{B} = LF + E$

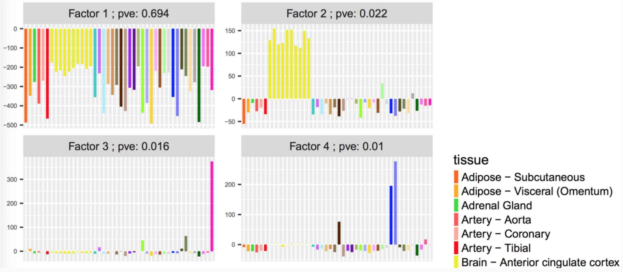


Figure: Sparse factor analysis of GTEx data

Incorporating all possible patterns

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

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

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

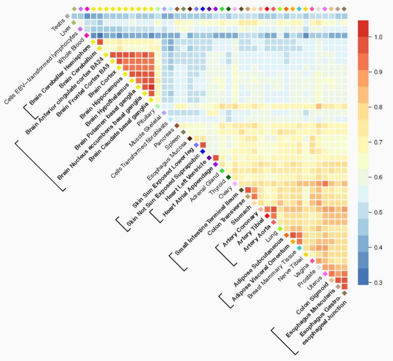


Figure: Quantitative characterization of eQTL effects heterogeneity in GTEx

Application to multivariate fine-mapping

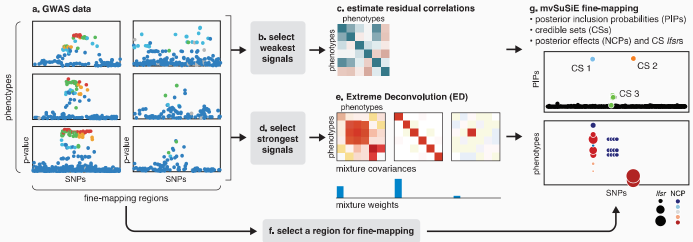

a. GWAS data
phenotypes
p-value
SNPs
p-value
SNPs
fine-mapping regions
b. select weakest signals
d. select strongest signals
c. estimate residual correlations
phenotypes
phenotypes
e. Extreme Deconvolution (ED)
phenotypes
phenotypes
mixture covariances
mixture weights
g. mvSuSiE fine-mapping
• posterior inclusion probabilities (PIPs)
• credible sets (CSs)
• posterior effects (NCPs) and CS lfsrs
PIPs
CS 1
CS 2
CS 3
phenotypes
SNPs
lfsr
NCP
f. select a region for fine-mapping

Figure: mvSuSiE fine-mapping with adaptive shrinkage model

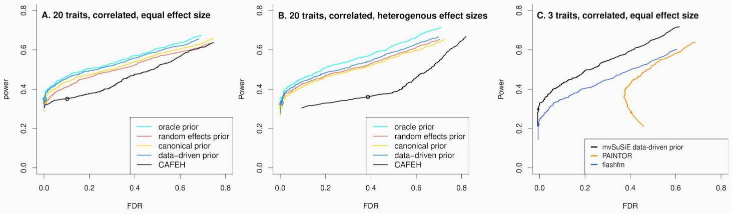
Zou *et al.* (2023) biorxiv

Multi-trait fine-mapping methods & challenges

	mvSuSiE	CAFEH	PAINTOR	MTHESS	BayesSUR	flashfm	msCaviar	HyPrColoc	moloc
>5 traits integrated	🟩	🟩	▨	▨	▨	▨	▨	🟩	▨
>10 traits integrated	🟩	🟩						🟩	
Multiple causal signals	🟩	🟩	▨	🟩	🟩	▨	▨		
Individual level data	🟩	🟩		🟩	🟩				
Summary statistics	🟩	🟩	🟩			🟩	🟩	🟩	🟩
Missing data	▨					🟩			
Trait specific LD							🟩		
Correlated effects	🟩			🟩	🟩	🟩	🟩		
Trait specific effects	🟩	🟩				🟩		🟩	🟩
Arbitrary heterogeneous effects	🟩								
Arbitrary multi-trait colocalization		🟩				🟩		🟩	🟩
Correlated traits	🟩			🟩	🟩	🟩		🟩	
Partial sample overlap	🟩					🟩		🟩	
Functional annotation	▨	▨	🟩			▨			
Trait specific functional annotation									
Genome-wide scalability	🟩	🟩						🟩	

Reference: CAFEH: Arvanitis et al (2022); PAINTOR: Kichaev et al (2017); MTHESS: Lewin et al (2016); BayesSUR: Zhao et al (2021); flashfm: Hernández et al (2021); msCaviar: LaPierre et al (2021); HyPrColoc: Foley et al (2021); moloc: Giambartolomei et al (2018).

Comparison to other methods



GWAS application: 16 blood traits in UK Biobank

Analysis overview

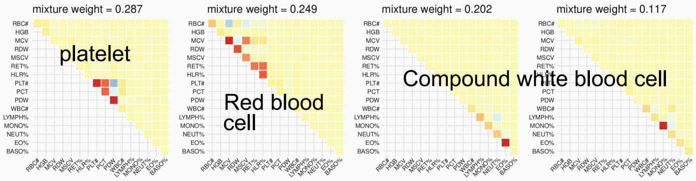
- Sample size 248,980; 975 candidate regions fine-mapped
- Average #SNPs per region 4,776; maximum 36,605

GWAS application: 16 blood traits in UK Biobank

Analysis overview

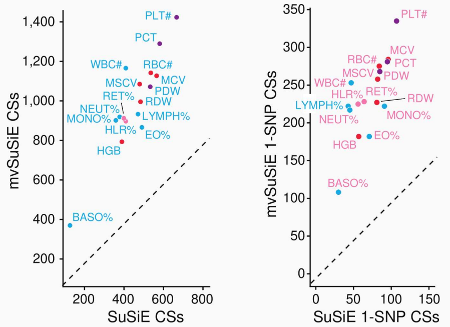
- Sample size 248,980; 975 candidate regions fine-mapped
- Average #SNPs per region 4,776; maximum 36,605

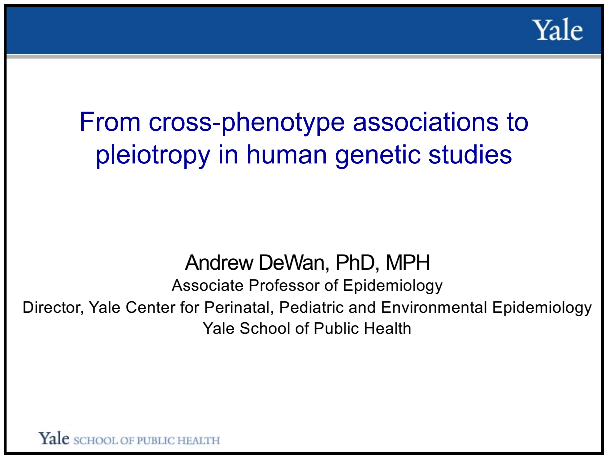
Top patterns of effect size sharing inferred from data:



GWAS application: 16 blood traits in UK Biobank

Many more signals identified compared to fine-mapping per each trait





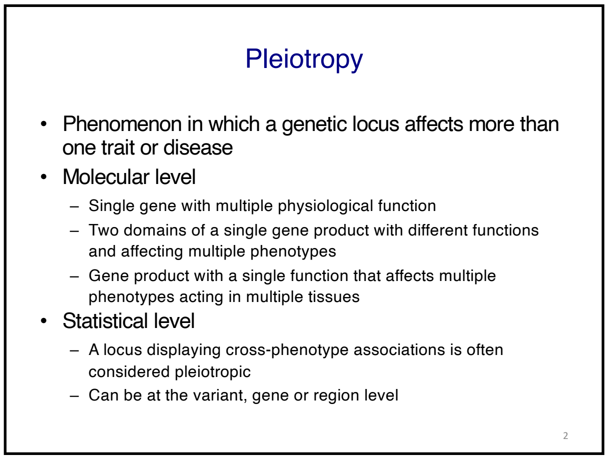
Yale

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

Yale SCHOOL OF PUBLIC HEALTH

1

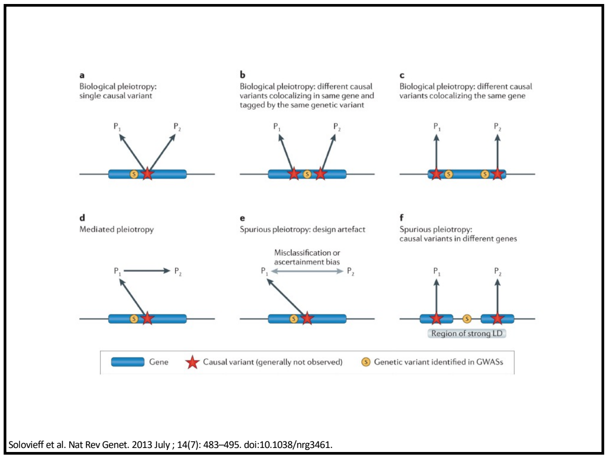


Pleiotropy

- 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

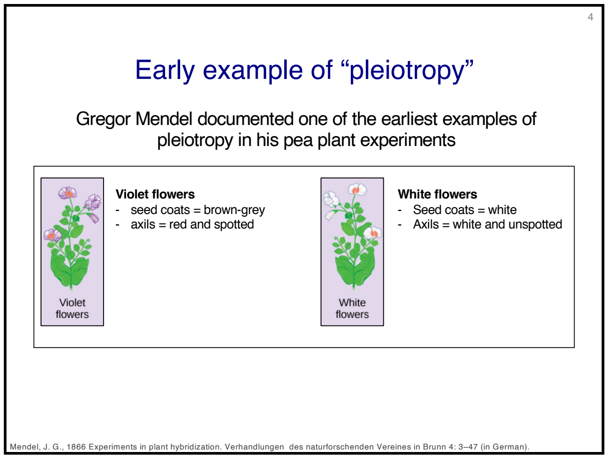
2



a
Biological pleiotropy: single causal variant
b
Biological pleiotropy: different causal variants colocalizing in same gene and tagged by the same genetic variant
c
Biological pleiotropy: different causal variants colocalizing the same gene
d
Mediated pleiotropy
e
Spurious pleiotropy: design artefact
Misclassification or ascertainment bias
f
Spurious pleiotropy: causal variants in different genes
Region of strong LD
P1
P2
S
Gene
Causal variant (generally not observed)
Genetic variant identified in GWASs

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

3



4

Early example of “pleiotropy”

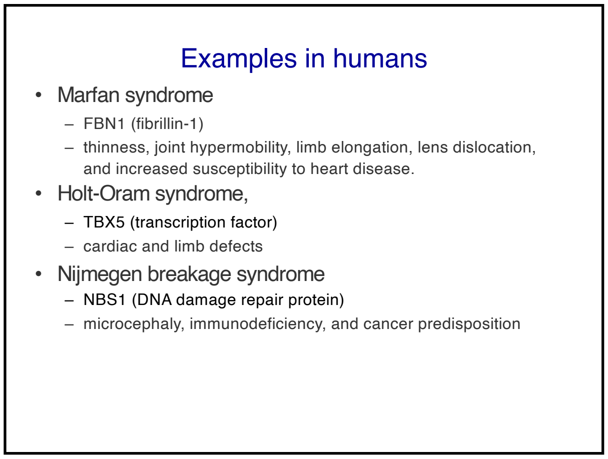
Gregor Mendel documented one of the earliest examples of pleiotropy in his pea plant experiments

Violet flowers
- seed coats = brown-grey
- axils = red and spotted

White flowers
- Seed coats = white
- Axils = white and unspotted

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

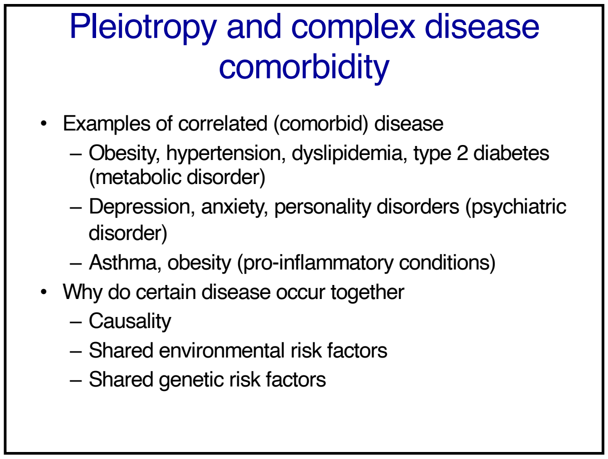
4



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

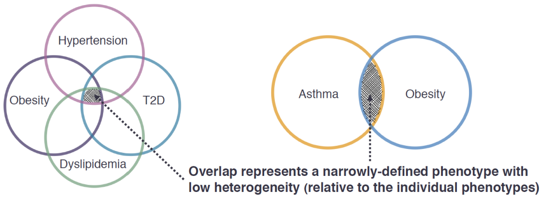


Pleiotropy and complex disease comorbidity

- 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



Overlap represents a narrowly-defined phenotype with low heterogeneity (relative to the individual phenotypes)

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

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

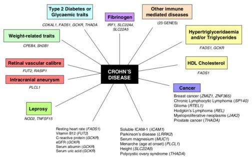
Shanya Sivakumaran,[1,6] Felix Agakov,[1,2,6] Evropi Theodoratou,[1,6] James G. Prendergast,[3] Lina Zgaga,[1,4] Teri Manolio,[5] Igor Rudan,[1] Paul McKeigue,[1] James F. Wilson,[1] and Harry Campbell[1,*]

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.



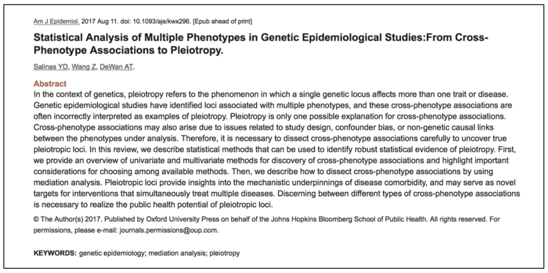
10

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

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.

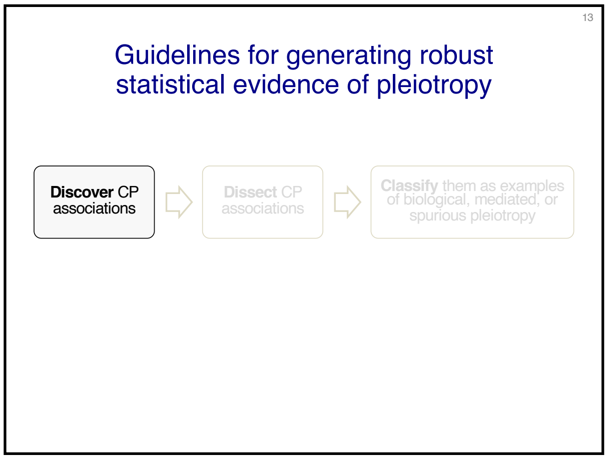
Abstract

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.

© The Author(s) 2017. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

KEYWORDS: genetic epidemiology; mediation analysis; pleiotropy

12

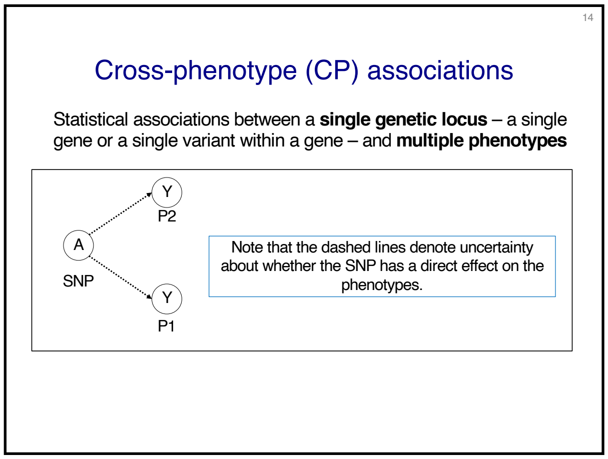


13

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



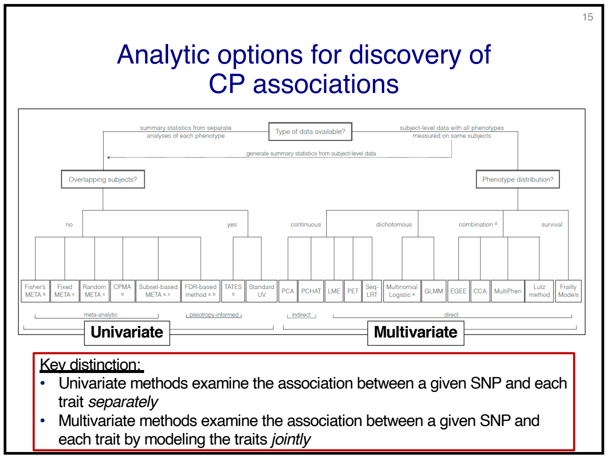
14

Cross-phenotype (CP) associations

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

Note that the dashed lines denote uncertainty about whether the SNP has a direct effect on the phenotypes.

14



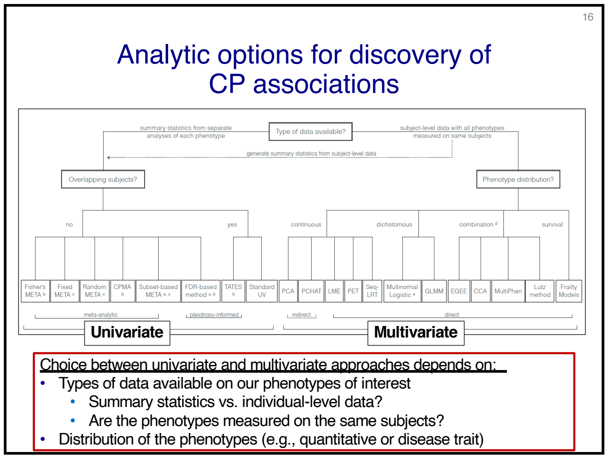
15

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



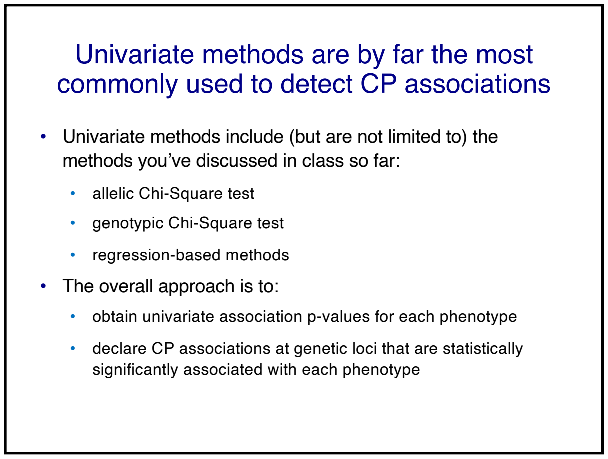
16

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)

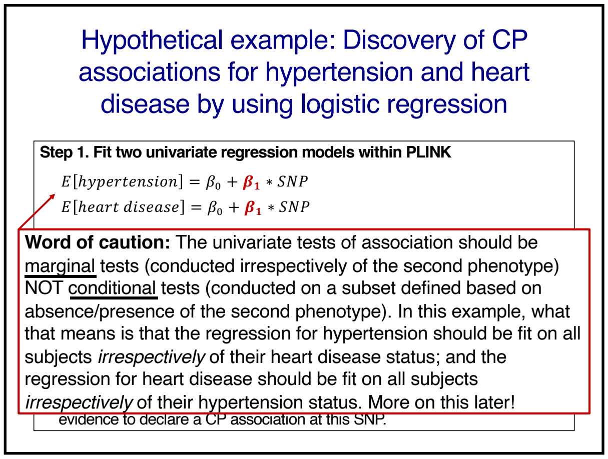
16



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



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

Step 1. Fit two univariate regression models within PLINK

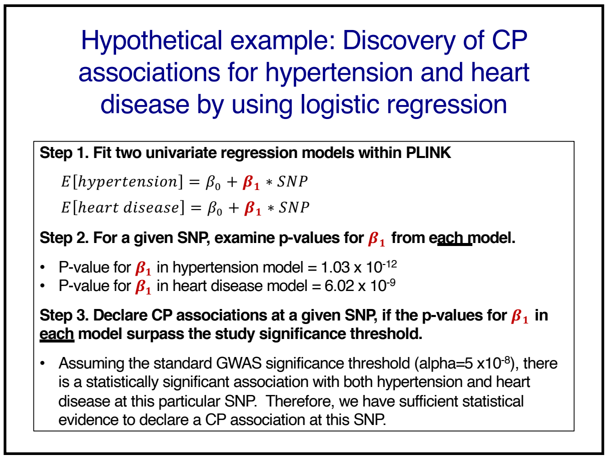
$E[hypertension] = \beta_0 + \boldsymbol{\beta_1} * SNP$

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

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

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

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

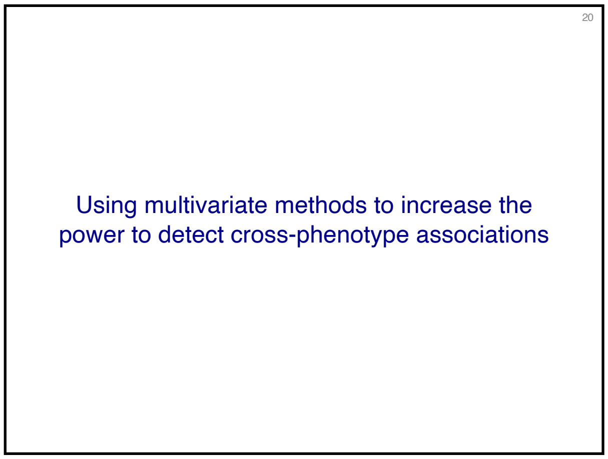
Step 2. For a given SNP, examine p-values for $\boldsymbol{\beta_1}$ from each model.

- P-value for $\boldsymbol{\beta_1}$ in hypertension model = 1.03 x 10^{-12}
- P-value for $\boldsymbol{\beta_1}$ in heart disease model = 6.02 x 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 x10^{-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.

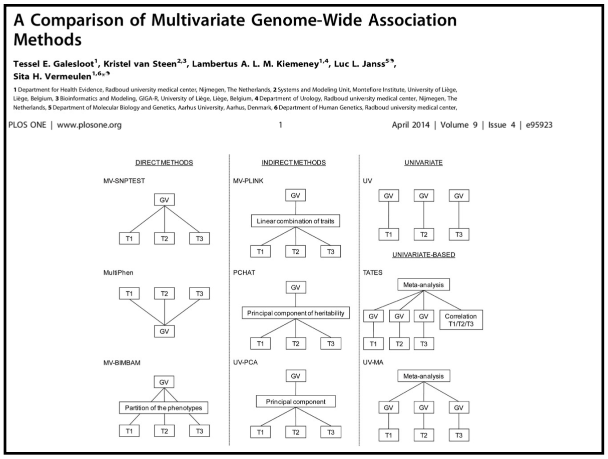
19



20

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[1,6]*

[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, [6] Department of Human Genetics, Radboud university medical center,

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

21

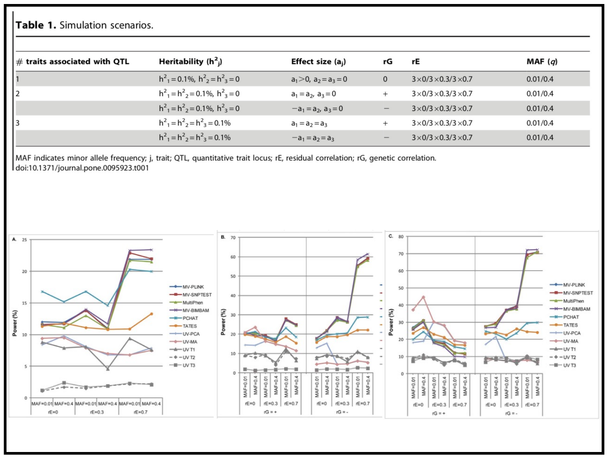
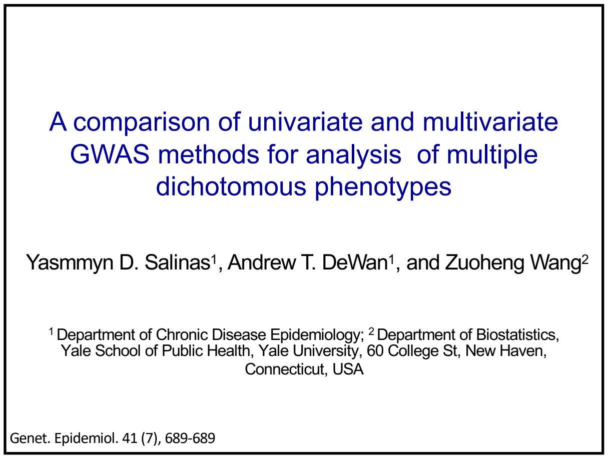


Table 1. Simulation scenarios.

# 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	3×0/3×0.3/3×0.7	0.01/0.4
2	$h^2_1 = h^2_2 = 0.1\%$, $h^2_3 = 0$	$a_1 = a_2$, $a_3 = 0$	+	3×0/3×0.3/3×0.7	0.01/0.4
	$h^2_1 = h^2_2 = 0.1\%$, $h^2_3 = 0$	$-a_1 = a_2$, $a_3 = 0$	−	3×0/3×0.3/3×0.7	0.01/0.4
3	$h^2_1 = h^2_2 = h^2_3 = 0.1\%$	$a_1 = a_2 = a_3$	+	3×0/3×0.3/3×0.7	0.01/0.4
	$h^2_1 = h^2_2 = h^2_3 = 0.1\%$	$-a_1 = a_2 = a_3$	−	3×0/3×0.3/3×0.7	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



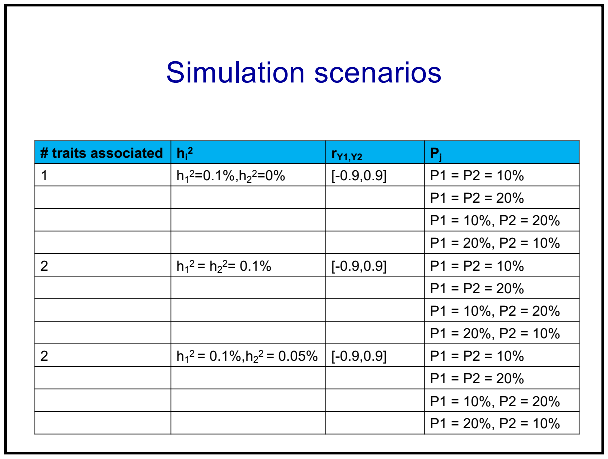
A comparison of univariate and multivariate GWAS methods for analysis of multiple dichotomous phenotypes

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



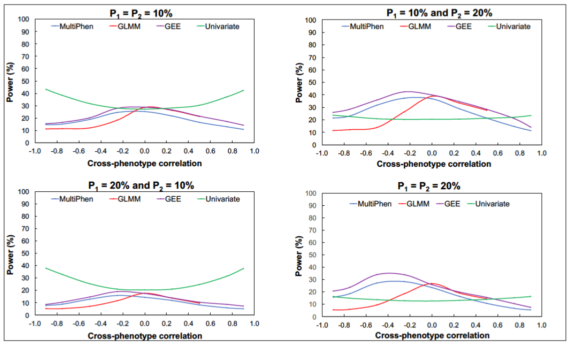
Simulation scenarios

# traits associated	h_i^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]

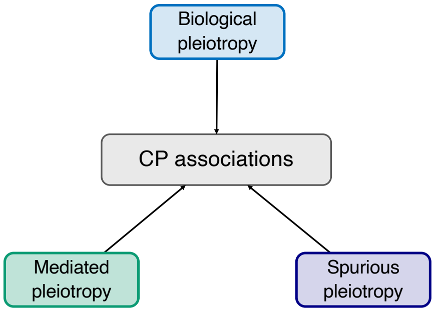


[a] Results for GLMMs are shown for $r_{Y1,Y2} \leq 0.5$ only, since the models experienced convergence issues for $r_{Y1,Y2} > 0.5$.

25

Problem: CP associations need not be indicative of pleiotropy

26



Biological pleiotropy

CP associations

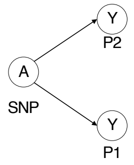
Mediated pleiotropy

Spurious pleiotropy

27

Biological pleiotropy

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

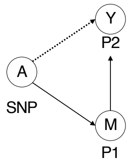


The SNP has a direct effect on each phenotype. (Note that direct or causal effects are depicted with solid lines).

28

Mediated pleiotropy

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

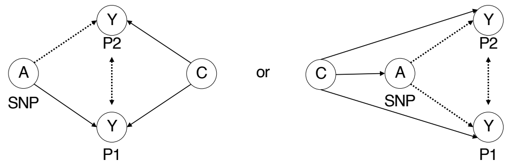


A non-genetic causal link between M and Y induces an association between A and Y, even in the absence of a direct effect of A on Y.

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

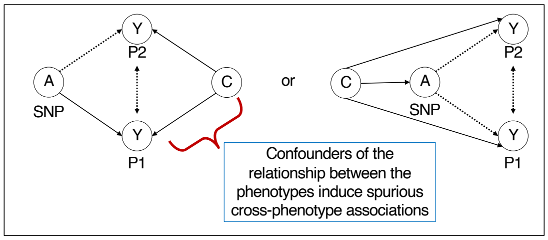


*Linkage disequilibrium is the non-random co-segregation of alleles.

30

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



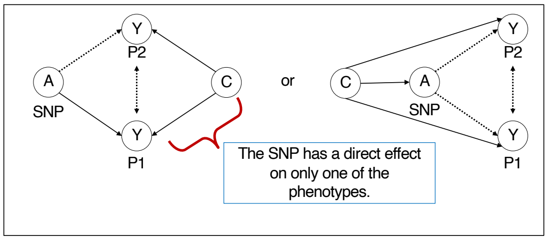
Confounders of the relationship between the phenotypes induce spurious cross-phenotype associations

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

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



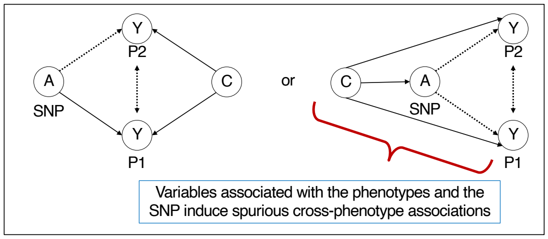
The SNP has a direct effect on only one of the phenotypes.

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

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



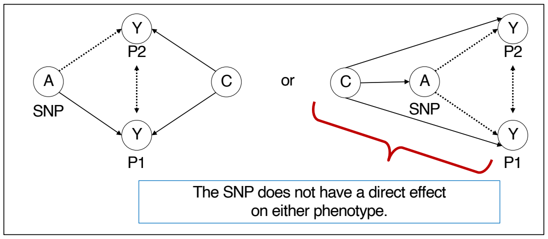
Variables associated with the phenotypes and the SNP induce spurious cross-phenotype associations

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

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

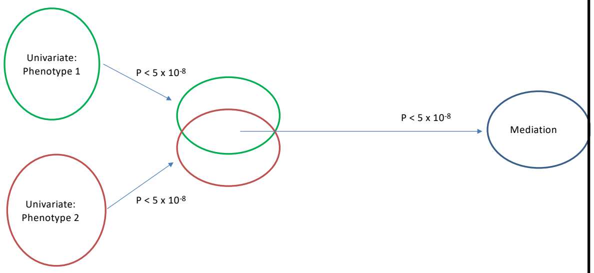


The SNP does not have a direct effect on either phenotype.

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

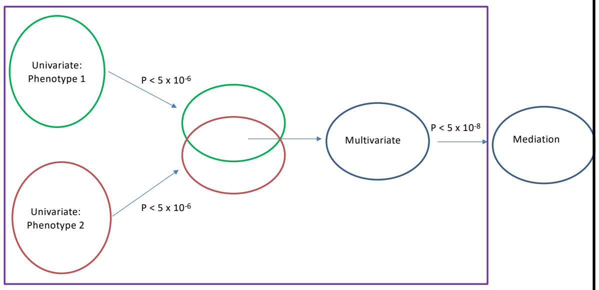
34

Pleiotropy exercise (Parts 1 and 2)



35

Pleiotropy exercise (Parts 1 and 2)



36

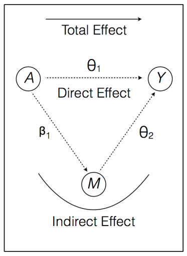
Guidelines for generating robust statistical evidence of pleiotropy

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

37

Mediation analysis provides a tool for dissecting CP associations

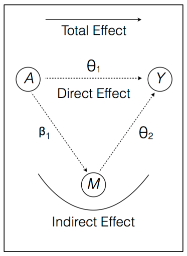
- Mediation analysis decomposes the **total effect** of the SNP (*A)* on a phenotypic outcome (*Y*) into:
  - **Direct effect:** effect of *A* on *Y* that occurs independently of an intermediate phenotype (*M)*
  - **Indirect effect:** effect of *A* on *Y* that occurs through the intermediate phenotype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



38

Mediation analysis: Data requirements

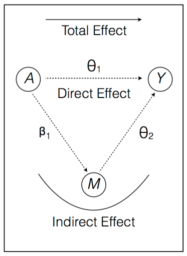
- All phenotypes must be measured on the same subjects
- Temporality must be ascertained
  - The occurrence of the intermediate variable *M* must precede that of the phenotypic outcome variable *Y*



39

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*

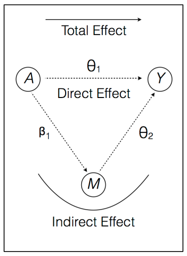


40

Mediation analysis: Assumptions

Typically met in genetic epi studies!

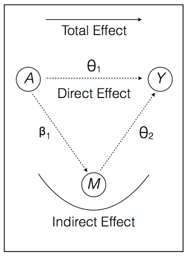
- There must be no unmeasured:
  - confounders of the total effect
  - confounders of the relationship between SNP *A* and the mediator *M*
  - confounders of the relationship between mediator *M* and phenotypic outcome *Y*



41

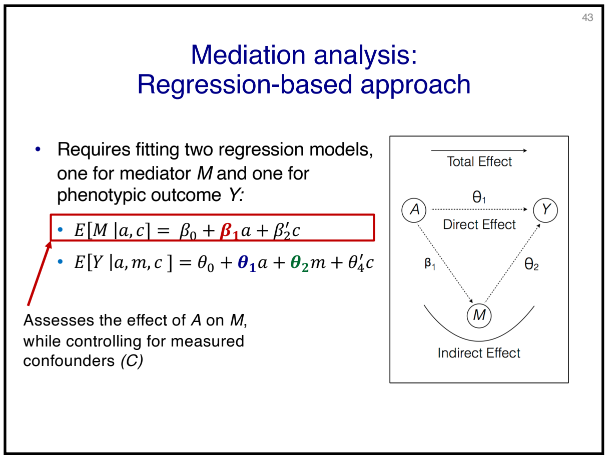
Mediation analysis: Assumptions

- 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



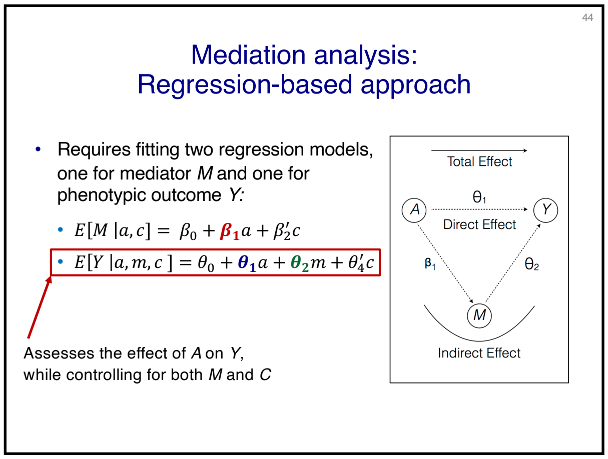
43

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 *M*, while controlling for measured confounders *(C)*

43



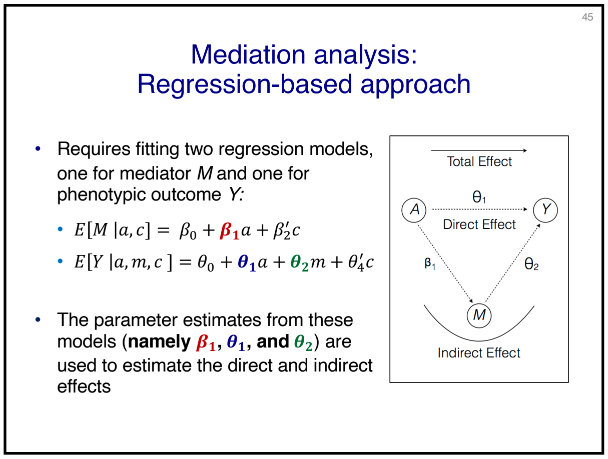
44

Mediation analysis: Regression-based approach

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

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

44

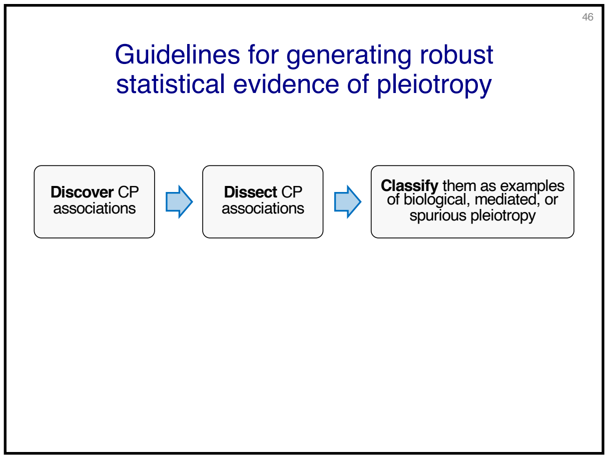


45

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

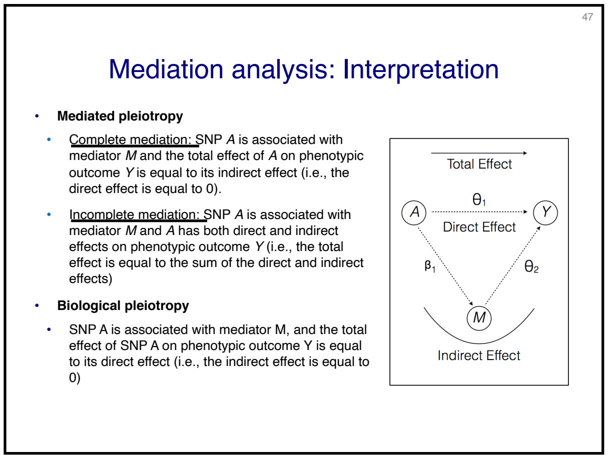


46

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

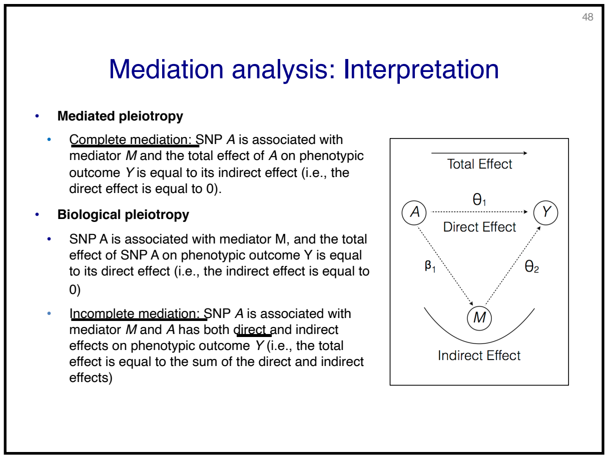


47

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).
  - 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

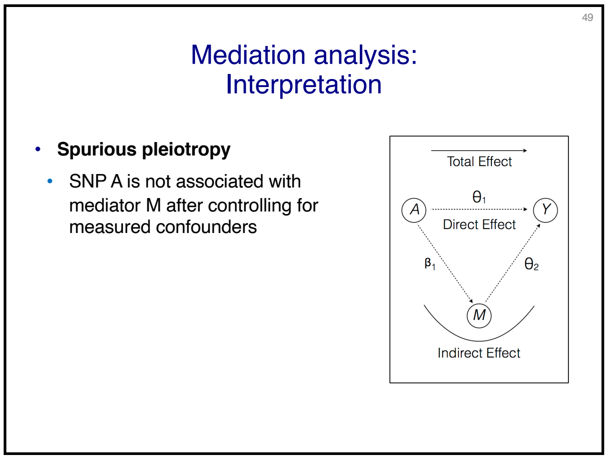


48

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)

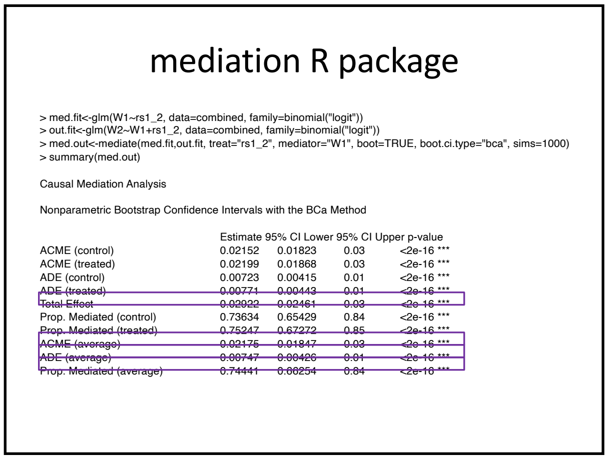
48



Mediation analysis: Interpretation

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

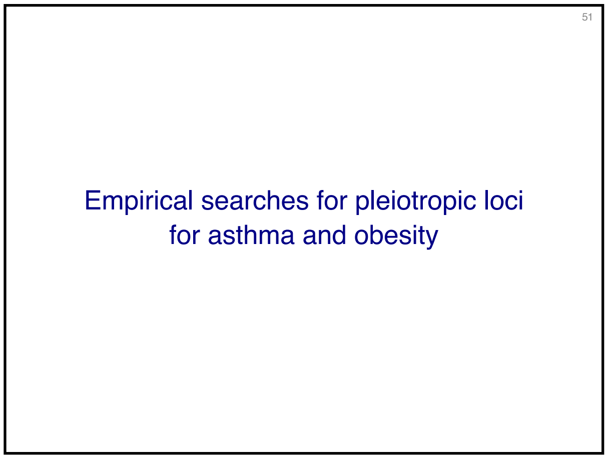
49



mediation R package

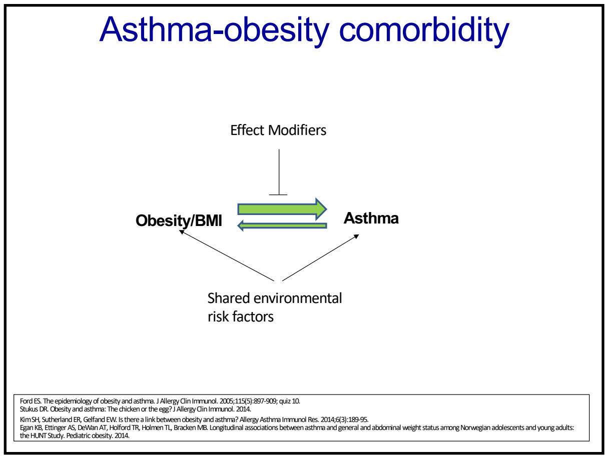
```
> med.fit<-glm(W1~rs1_2, data=combined, family=binomial("logit"))
> out.fit<-glm(W2~W1+rs1_2, data=combined, family=binomial("logit"))
> med.out<-mediate(med.fit,out.fit, treat="rs1_2", mediator="W1", boot=TRUE, boot.ci.type="bca", sims=1000)
> summary(med.out)

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

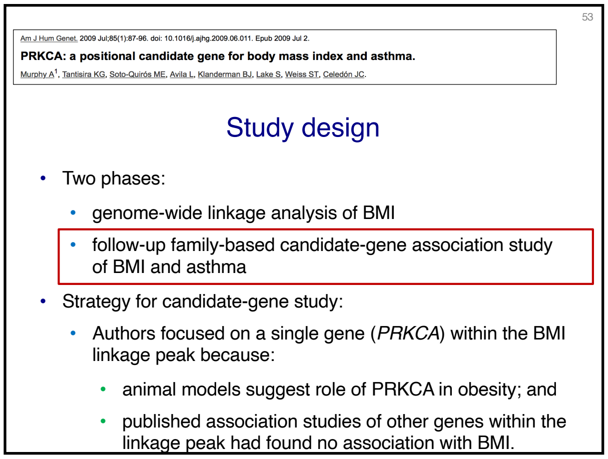
51



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.

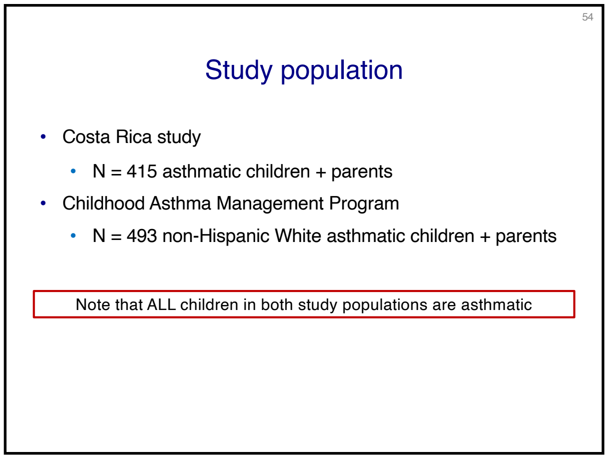
PRKCA: a positional candidate gene for body mass index and asthma.

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

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

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: α = 9.5 x 10^{-5}

56

Results for BMI

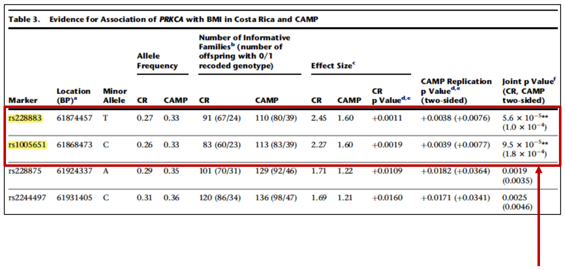


Table 3. Evidence for Association of *PRKCA* with BMI in Costa Rica and CAMP

			Allele Frequency		Number of Informative Families^{b} (number of offspring with 0/1 recoded genotype)		Effect Size^{c}				
Marker	Location (BP)^{a}	Minor Allele	CR	CAMP	CR	CAMP	CR	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)	5.6×10^{-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)	9.5×10^{-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

Results for asthma

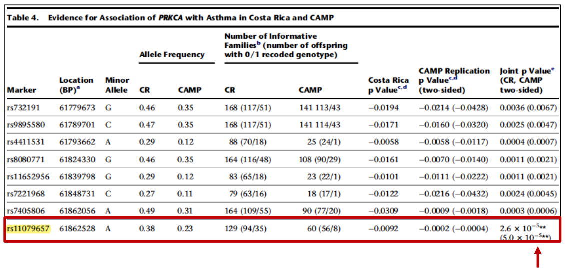


Table 4. Evidence for Association of *PRKCA* with Asthma in Costa Rica and CAMP

			Allele Frequency		Number of Informative Families^{b} (number of offspring with 0/1 recoded genotype)				
Marker	Location (BP)^{a}	Minor Allele	CR	CAMP	CR	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)	2.6×10^{-5}** (5.0×10^{-5}**)

One asthma-associated variant

58

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

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)

63

64

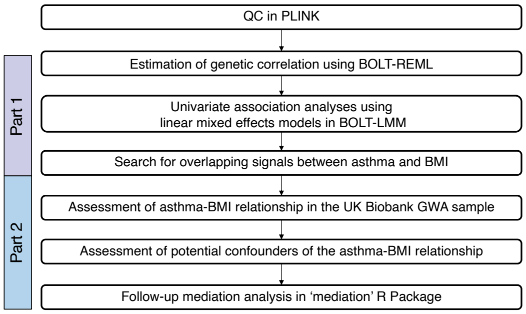
Phenotype definitions

- 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

64

65

Statistical Methods



QC in PLINK

Estimation of genetic correlation using BOLT-REML

Univariate association analyses using linear mixed effects models in BOLT-LMM

Search for overlapping signals between asthma and BMI

Assessment of asthma-BMI relationship in the UK Biobank GWA sample

Assessment of potential confounders of the asthma-BMI relationship

Follow-up mediation analysis in ‘mediation’ R Package

65

66

Overlap in GWA signals

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

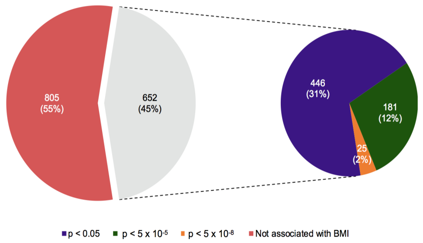
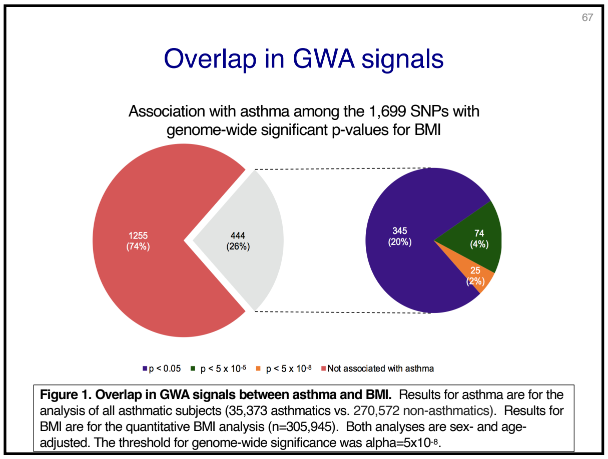


Figure 1. Overlap in GWA signals between asthma and BMI. Results for asthma are for the analysis of all asthmatic subjects (35,373 asthmatics vs. 270,572 non-asthmatics). Results for BMI are for the quantitative BMI analysis (n=305,945). Both analyses are sex- and age-adjusted. The threshold for genome-wide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

66

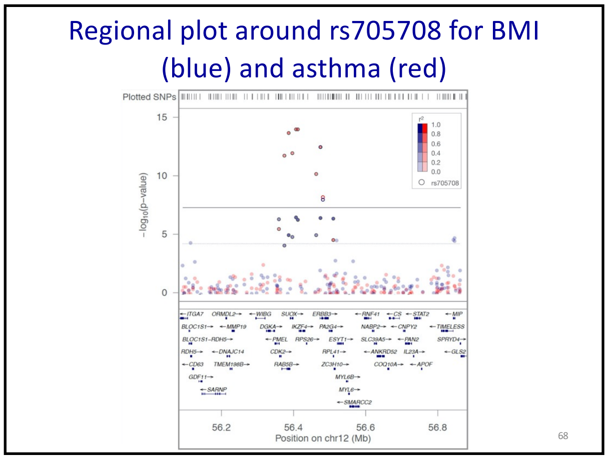


Overlap in GWA signals

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

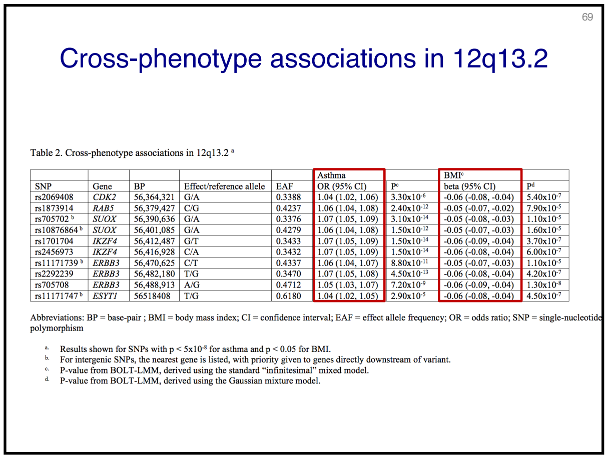
Figure 1. Overlap in GWA signals between asthma and BMI. Results for asthma are for the analysis of all asthmatic subjects (35,373 asthmatics vs. 270,572 non-asthmatics). Results for BMI are for the quantitative BMI analysis (n=305,945). Both analyses are sex- and age-adjusted. The threshold for genome-wide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

67



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

68



Cross-phenotype associations in 12q13.2

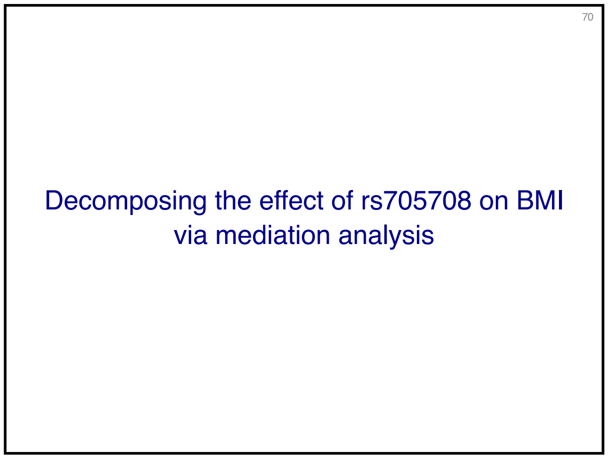
Table 2. Cross-phenotype associations in 12q13.2 [a]

					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)	$3.30x10^{-6}$	-0.06 (-0.08, -0.04)	$5.40x10^{-7}$
rs1873914	*RAB5*	56,379,427	C/G	0.4237	1.06 (1.04, 1.08)	$2.40x10^{-12}$	-0.05 (-0.07, -0.02)	$7.90x10^{-5}$
rs705702 [b]	*SUOX*	56,390,636	G/A	0.3376	1.07 (1.05, 1.09)	$3.10x10^{-14}$	-0.05 (-0.08, -0.03)	$1.10x10^{-5}$
rs10876864 [b]	*SUOX*	56,401,085	G/A	0.4279	1.06 (1.04, 1.08)	$1.50x10^{-12}$	-0.05 (-0.07, -0.03)	$1.60x10^{-5}$
rs1701704	*IKZF4*	56,412,487	G/T	0.3433	1.07 (1.05, 1.09)	$1.50x10^{-14}$	-0.06 (-0.09, -0.04)	$3.70x10^{-7}$
rs2456973	*IKZF4*	56,416,928	C/A	0.3432	1.07 (1.05, 1.09)	$1.50x10^{-14}$	-0.06 (-0.08, -0.04)	$6.00x10^{-7}$
rs11171739 [b]	*ERBB3*	56,470,625	C/T	0.4337	1.06 (1.04, 1.07)	$8.80x10^{-11}$	-0.05 (-0.07, -0.03)	$1.10x10^{-5}$
rs2292239	*ERBB3*	56,482,180	T/G	0.3470	1.07 (1.05, 1.08)	$4.50x10^{-13}$	-0.06 (-0.08, -0.04)	$4.20x10^{-7}$
rs705708	*ERBB3*	56,488,913	A/G	0.4712	1.05 (1.03, 1.07)	$7.20x10^{-9}$	-0.06 (-0.09, -0.04)	$1.30x10^{-8}$
rs11171747 [b]	*ESYT1*	56518408	T/G	0.6180	1.04 (1.02, 1.05)	$2.90x10^{-5}$	-0.06 (-0.08, -0.04)	$4.50x10^{-7}$

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

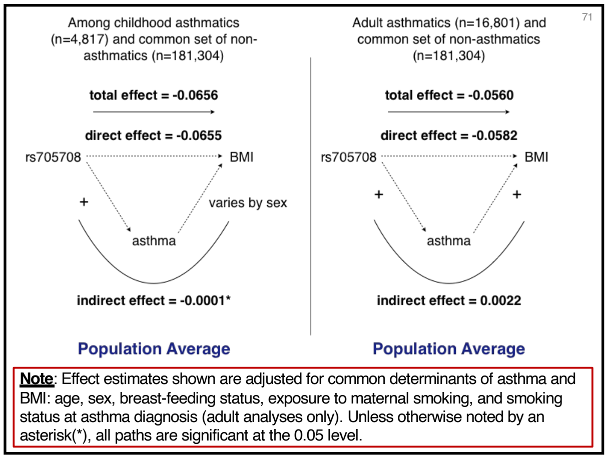
[a] Results shown for SNPs with $p < 5x10^{-8}$ for asthma and $p < 0.05$ for BMI.
[b] For intergenic SNPs, the nearest gene is listed, with priority given to genes directly downstream of variant.
[c] P-value from BOLT-LMM, derived using the standard "infinitesimal" mixed model.
[d] P-value from BOLT-LMM,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

+

varies by sex

asthma

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

+

+

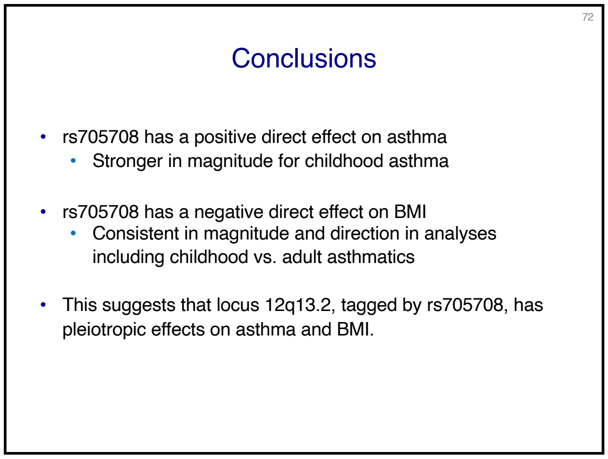
asthma

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.

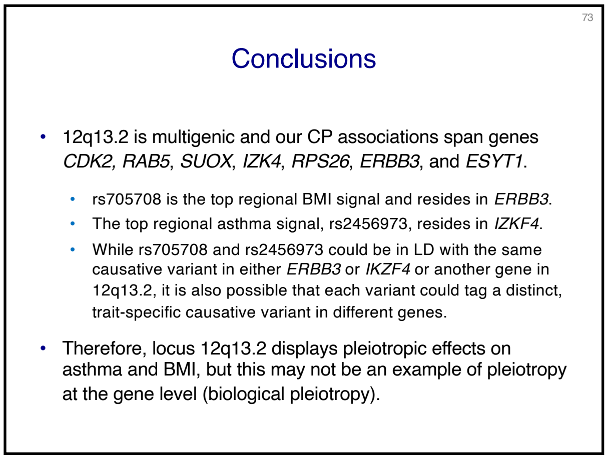
71



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.

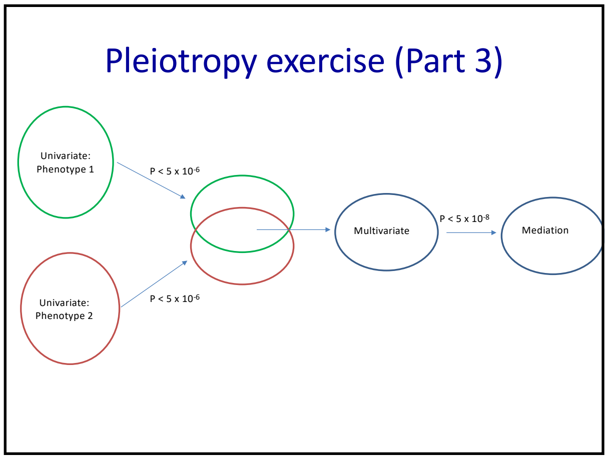
72



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)

Univariate: Phenotype 1

$P < 5 \times 10^{-6}$

Univariate: Phenotype 2

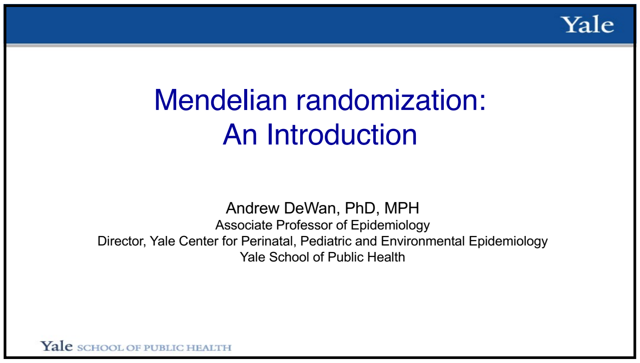
$P < 5 \times 10^{-6}$

Multivariate

$P < 5 \times 10^{-8}$

Mediation

74



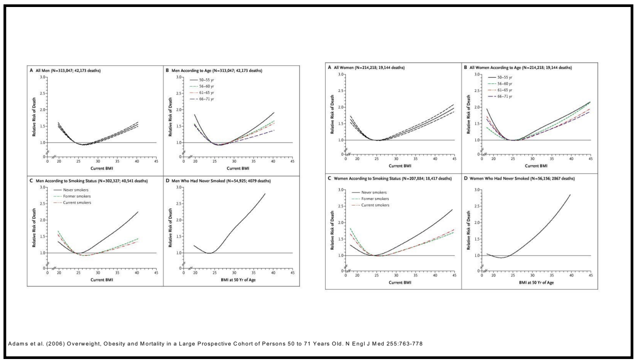
Yale

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

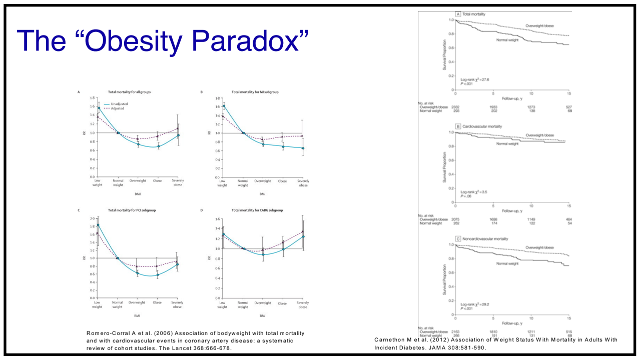
Yale SCHOOL OF PUBLIC HEALTH

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 255:763-778

2

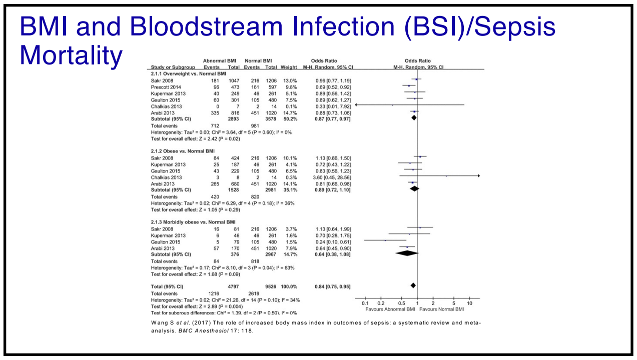


The "Obesity Paradox"

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 368:666-678.

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

Study or Subgroup	Abnormal BMI Events	Abnormal BMI Total	Normal BMI Events	Normal BMI Total	Weight	Odds Ratio M-H, Random, 95% CI
2.1.1 Overweight vs. Normal BMI						
Sakr 2008	181	1047	216	1206	13.0%	0.96 [0.77, 1.19]
Prescott 2014	96	473	161	597	9.8%	0.69 [0.52, 0.92]
Kuperman 2013	40	249	46	261	5.1%	0.89 [0.56, 1.42]
Gaulton 2015	60	301	105	480	7.5%	0.89 [0.62, 1.27]
Chalkias 2013	0	7	2	14	0.1%	0.33 [0.01, 7.92]
Arabi 2013	335	816	451	1020	14.7%	0.88 [0.73, 1.06]
Subtotal (95% CI)		**2893**		**3578**	**50.2%**	**0.87 [0.77, 0.97]**
Total events	712		981			
Heterogeneity: Tau² = 0.00; Chi² = 3.64, df = 5 (P = 0.60); I² = 0%						
Test for overall effect: Z = 2.42 (P = 0.02)						
2.1.2 Obese vs. Normal BMI						
Sakr 2008	84	424	216	1206	10.1%	1.13 [0.86, 1.50]
Kuperman 2013	25	187	46	261	4.1%	0.72 [0.43, 1.22]
Gaulton 2015	43	229	105	480	6.5%	0.83 [0.56, 1.23]
Chalkias 2013	3	8	2	14	0.3%	3.60 [0.45, 28.56]
Arabi 2013	265	680	451	1020	14.1%	0.81 [0.66, 0.98]
Subtotal (95% CI)		**1528**		**2981**	**35.1%**	**0.89 [0.72, 1.10]**
Total events	420		820			
Heterogeneity: Tau² = 0.02; Chi² = 6.29, df = 4 (P = 0.18); I² = 36%						
Test for overall effect: Z = 1.05 (P = 0.29)						
2.1.3 Morbidly obese vs. Normal BMI						
Sakr 2008	16	81	216	1206	3.7%	1.13 [0.64, 1.99]
Kuperman 2013	6	46	46	261	1.6%	0.70 [0.28, 1.75]
Gaulton 2015	5	79	105	480	1.5%	0.24 [0.10, 0.61]
Arabi 2013	57	170	451	1020	7.9%	0.64 [0.45, 0.90]
Subtotal (95% CI)		**376**		**2967**	**14.7%**	**0.64 [0.38, 1.08]**
Total events	84		818			
Heterogeneity: Tau² = 0.17; Chi² = 8.10, df = 3 (P = 0.04); I² = 63%						
Test for overall effect: Z = 1.68 (P = 0.09)						
Total (95% CI)		**4797**		**9526**	**100.0%**	**0.84 [0.75, 0.95]**
Total events	1216		2619			
Heterogeneity: Tau² = 0.02; Chi² = 21.26, df = 14 (P = 0.10); I² = 34%						
Test for overall effect: Z = 2.89 (P = 0.004)						
Test for subgroup differences: Chi² = 1.39, df = 2 (P = 0.50), I² = 0%						

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

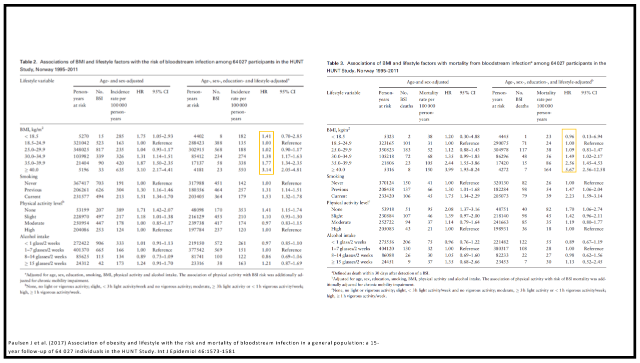


Table 2. Associations of BMI and lifestyle factors with the risk of bloodstream infection among 64 027 participants in the HUNT Study, Norway 1995–2011

Lifestyle variable	Age- and sex-adjusted					Age-, sex-, education- and lifestyle-adjusted[a]				
	Person-years at risk	No. BSI	Incidence rate per 100 000 person-years	HR	95% CI	Person-years at risk	No. BSI	Incidence rate per 100 000 person-years	HR	95% CI
BMI, kg/m²										
< 18.5	5270	15	285	1.75	1.05–2.93	4402	8	182	1.41	0.70–2.85
18.5–24.9	321042	523	163	1.00	Reference	288423	388	135	1.00	Reference
25.0–29.9	348025	817	235	1.04	0.93–1.17	302915	568	188	1.02	0.90–1.17
30.0–34.9	103982	339	326	1.31	1.14–1.51	85412	234	274	1.38	1.17–1.63
35.0–39.9	21404	90	420	1.87	1.50–2.35	17137	58	338	1.77	1.34–2.35
≥ 40.0	5196	33	635	3.10	2.17–4.41	4181	23	550	3.14	2.05–4.81
Smoking										
Never	367417	703	191	1.00	Reference	317988	451	142	1.00	Reference
Previous	206261	626	304	1.30	1.16–1.46	180356	464	257	1.31	1.14–1.51
Current	231577	494	213	1.51	1.34–1.70	203405	364	179	1.53	1.32–1.78
Physical activity level[b]										
None	53199	207	389	1.71	1.42–2.07	48098	170	353	1.41	1.15–1.74
Slight	228970	497	217	1.18	1.01–1.38	216129	455	210	1.10	0.93–1.30
Moderate	250954	447	178	1.00	0.85–1.17	239738	417	174	0.97	0.83–1.15
High	204086	253	124	1.00	Reference	197784	237	120	1.00	Reference
Alcohol intake										
< 1 glass/2 weeks	272422	906	333	1.01	0.91–1.13	219150	572	261	0.97	0.85–1.10
1–7 glasses/2 weeks	401370	665	166	1.00	Reference	377542	569	151	1.00	Reference
8–14 glasses/2 weeks	85625	115	134	0.89	0.73–1.09	81741	100	122	0.86	0.69–1.06
≥ 15 glasses/2 weeks	24312	42	173	1.24	0.91–1.70	23316	38	163	1.21	0.87–1.69

[a] Adjusted for age, sex, education, smoking, BMI, physical activity and alcohol intake. The association of physical activity with BSI risk was additionally adjusted for chronic mobility impairment.

[b] None, no light or vigorous activity; slight, < 3 h light activity/week and no vigorous activity; moderate, ≥ 3 h light activity or < 1 h vigorous activity/week; high, ≥ 1 h vigorous activity/week.

Table 3. Associations of BMI and lifestyle factors with mortality from bloodstream infection[a] among 64 027 participants in the HUNT Study, Norway 1995–2011

Lifestyle variable	Age-and sex-adjusted					Age-, sex-, education-, and lifestyle-adjusted[b]				
	Person-years at risk	No. BSI deaths	Mortality rate per 100 000 person-years	HR	95% CI	Person-years at risk	No. BSI deaths	Mortality rate per 100 000 person-years	HR	95% CI
BMI, kg/m²										
< 18.5	5323	2	38	1.20	0.30–4.88	4445	1	23	0.96	0.13–6.94
18.5–24.9	323165	101	31	1.00	Reference	290075	71	24	1.00	Reference
25.0–29.9	350823	183	52	1.12	0.88–1.43	304978	117	38	1.09	0.81–1.47
30.0–34.9	105218	72	68	1.35	0.99–1.83	86296	48	56	1.49	1.02–2.17
35.0–39.9	21806	23	105	2.44	1.55–3.86	17420	15	86	2.56	1.45–4.53
≥ 40.0	5316	8	150	3.99	1.93–8.24	4272	7	164	5.67	2.56–12.58
Smoking										
Never	370124	150	41	1.00	Reference	320130	82	26	1.00	Reference
Previous	208458	137	66	1.30	1.01–1.68	182284	98	54	1.47	1.06–2.04
Current	233420	106	45	1.75	1.34–2.29	205073	79	39	2.23	1.59–3.14
Physical activity level[c]										
None	53918	51	95	2.08	1.37–3.16	48751	40	82	1.70	1.06–2.74
Slight	230884	107	46	1.39	0.97–2.00	218140	98	45	1.42	0.96–2.11
Moderate	252722	94	37	1.14	0.79–1.64	241663	85	35	1.19	0.80–1.77
High	205083	43	21	1.00	Reference	198931	36	18	1.00	Reference
Alcohol intake										
< 1 glass/2 weeks	275556	206	75	0.96	0.76–1.22	221482	122	55	0.89	0.67–1.19
1–7 glasses/2 weeks	404120	130	32	1.00	Reference	380317	108	28	1.00	Reference
8–14 glasses/2 weeks	86088	26	30	1.05	0.69–1.60	82233	22	27	0.98	0.62–1.56
≥ 15 glasses/2 weeks	24451	9	37	1.35	0.68–2.66	23453	7	30	1.13	0.52–2.45

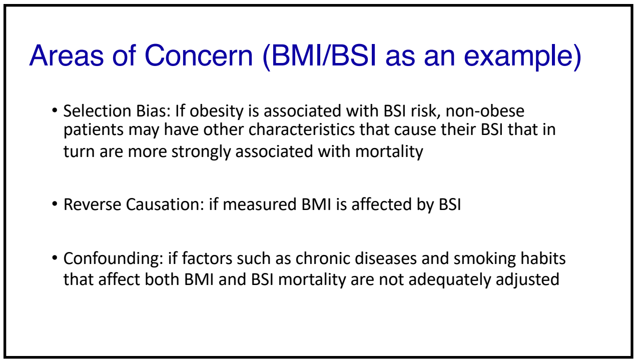
[a] Defined as death within 30 days after detection of a BSI.

[b] Adjusted for age, sex, education, smoking, BMI, physical activity and alcohol intake. The association of physical activity with risk of BSI mortality was additionally adjusted for chronic mobility impairment.

[c] None, no light or vigorous activity; slight, < 3h light activity/week and no vigorous activity; moderate, ≥ 3 h light activity or < 1 h vigorous activity/week; high, ≥ 1 h vigorous activity/week.

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

5



Areas of Concern (BMI/BSI as an example)

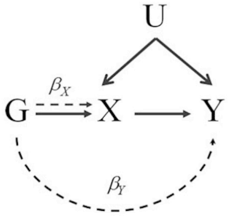
- Selection Bias: If obesity is associated with BSI risk, non-obese patients may have other characteristics that cause their BSI that in turn are more strongly associated with mortality
- Reverse Causation: if measured BMI is affected by BSI
- Confounding: if factors such as chronic diseases and smoking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

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

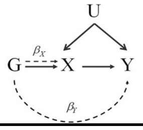


Dimou NL and Tsilidis KK. (2018) A primer in Mendelian Randomization Methodology with a Focus on Utilizing Published Summary Association Data. Methods Mol Biol. 2018;1793: 211-230

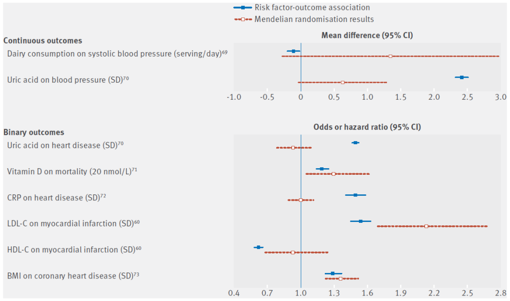
8

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



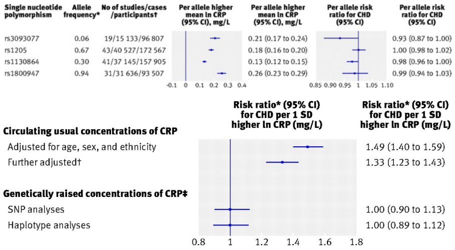
9



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

10

CRP and Heart Disease



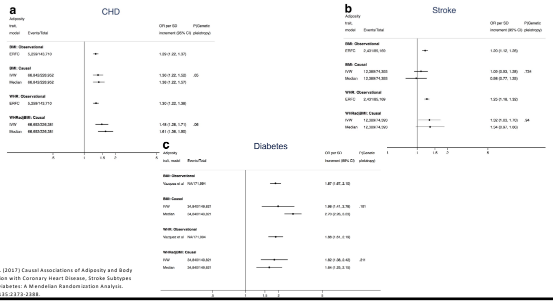
Single nucleotide polymorphism	Allele frequency*	No of studies/cases /participants†	Per allele higher mean ln CRP (95% CI), mg/L	Per allele risk ratio for CHD (95% CI)
rs3093077	0.06	19/15 133/96 807	0.21 (0.17 to 0.24)	0.93 (0.87 to 1.00)
rs1205	0.67	43/40 527/172 567	0.18 (0.16 to 0.20)	1.00 (0.98 to 1.02)
rs1130864	0.30	41/37 145/157 905	0.13 (0.12 to 0.15)	0.98 (0.96 to 1.00)
rs1800947	0.94	31/31 636/93 507	0.26 (0.23 to 0.29)	0.99 (0.94 to 1.03)

	Risk ratio* (95% CI) for CHD per 1 SD higher ln CRP (mg/L)
Circulating usual concentrations of CRP	
Adjusted for age, sex, and ethnicity	1.49 (1.40 to 1.59)
Further adjusted†	1.33 (1.23 to 1.43)
Genetically raised concentrations of CRP‡	
SNP analyses	1.00 (0.90 to 1.13)
Haplotype analyses	1.00 (0.89 to 1.12)

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

11

BMI and CHDIStrokeIType 2 Diabetes



a CHD

Adiposity trait, model	Events/Total	OR per SD increment (95% CI)	P(Genetic pleiotropy)
BMI: Observational			
ERFC	5,259/143,710	1.29 (1.22, 1.37)	
BMI: Causal			
IVW	66,842/228,952	1.36 (1.22, 1.52)	.65
Median	66,842/228,952	1.38 (1.22, 1.57)	
WHR: Observational			
ERFC	5,259/143,710	1.30 (1.22, 1.38)	
WHRadjBMI: Causal			
IVW	66,692/226,381	1.48 (1.28, 1.71)	.06
Median	66,692/226,381	1.61 (1.36, 1.90)	

b Stroke

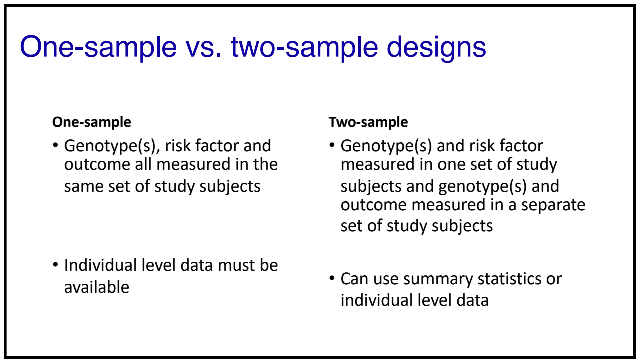
Adiposity trait, model	Events/Total	OR per SD increment (95% CI)	P(Genetic pleiotropy)
BMI: Observational			
ERFC	2,431/85,169	1.20 (1.12, 1.28)	
BMI: Causal			
IVW	12,389/74,393	1.09 (0.93, 1.28)	.734
Median	12,389/74,393	0.98 (0.77, 1.25)	
WHR: Observational			
ERFC	2,431/85,169	1.25 (1.18, 1.32)	
WHRadjBMI: Causal			
IVW	12,389/74,393	1.32 (1.03, 1.70)	.94
Median	12,389/74,393	1.34 (0.97, 1.86)	

c Diabetes

Adiposity trait, model	Events/Total	OR per SD increment (95% CI)	P(Genetic pleiotropy)
BMI: Observational			
Vazquez et al	NA/171,994	1.87 (1.67, 2.10)	
BMI: Causal			
IVW	34,840/149,821	1.98 (1.41, 2.78)	.101
Median	34,840/149,821	2.70 (2.26, 3.23)	
WHR: Observational			
Vazquez et al	NA/171,994	1.88 (1.61, 2.19)	
WHRadjBMI: Causal			
IVW	34,840/149,821	1.82 (1.38, 2.42)	.211
Median	34,840/149,821	1.64 (1.25, 2.15)	

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

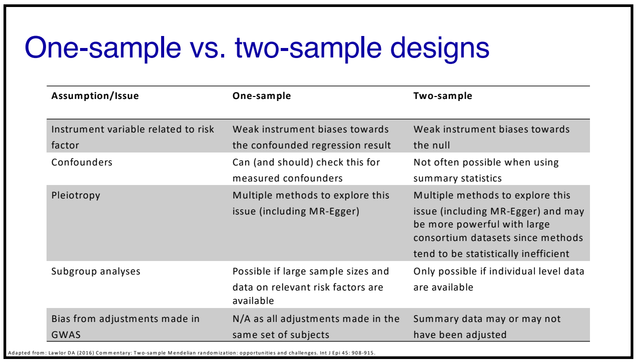
One-sample

- 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

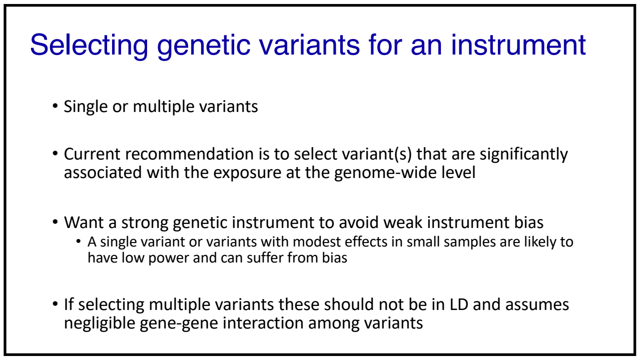


One-sample vs. two-sample designs

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 randomization: opportunities and challenges. Int J Epi 45: 908-915.

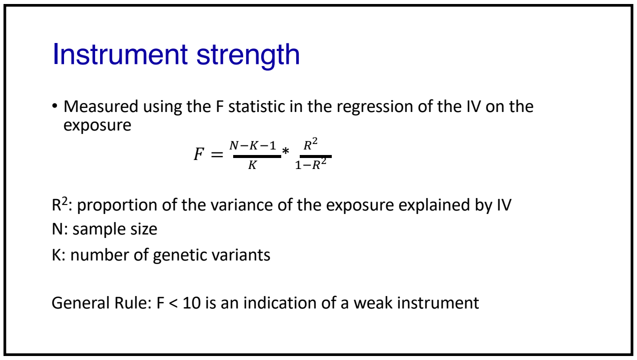
14



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

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

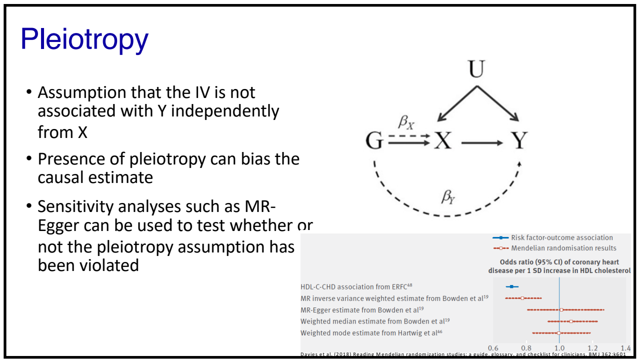
R^2: proportion of the variance of the exposure explained by IV

N: sample size

K: number of genetic variants

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

16



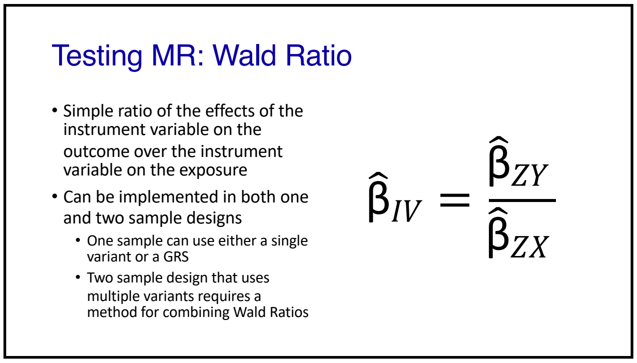
Pleiotropy

- 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

U
G
X
Y
β_X
β_Y
Risk factor-outcome association
Mendelian randomisation results
Odds ratio (95% CI) of coronary heart disease per 1 SD increase in HDL cholesterol
HDL-C-CHD association from ERFC[48]
MR inverse variance weighted estimate from Bowden et al[19]
MR-Egger estimate from Bowden et al[19]
Weighted median estimate from Bowden et al[19]
Weighted mode estimate from Hartwig et al[46]
0.6
0.8
1.0
1.2
1.4

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

17

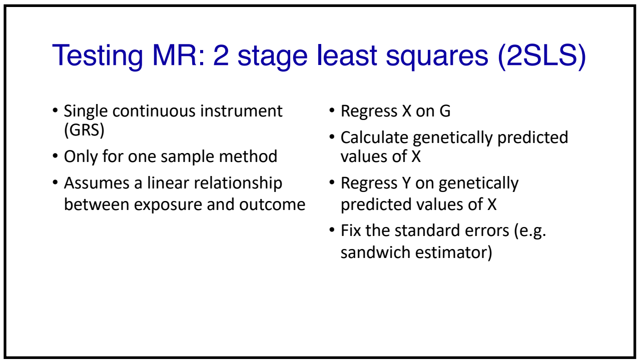


Testing MR: Wald Ratio

- 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

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

18

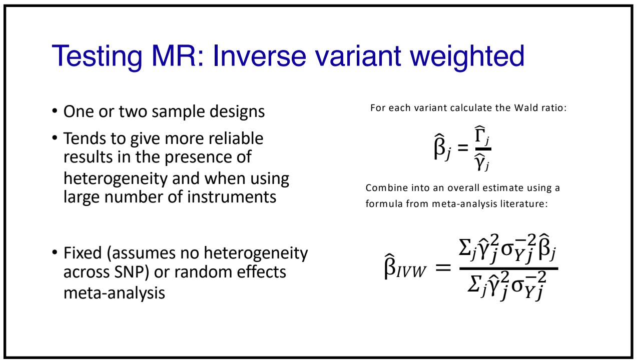


Testing MR: 2 stage least squares (2SLS)

- 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

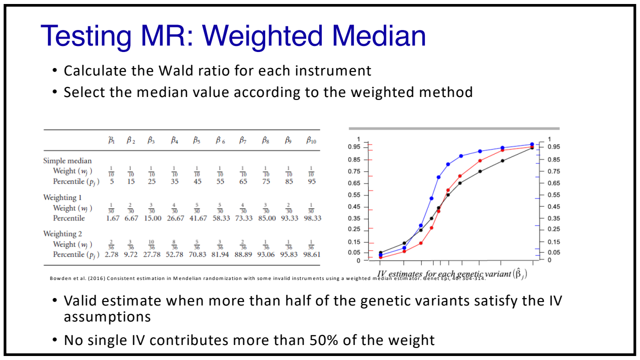
For each variant calculate the Wald ratio:

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

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

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

20



Testing MR: Weighted Median

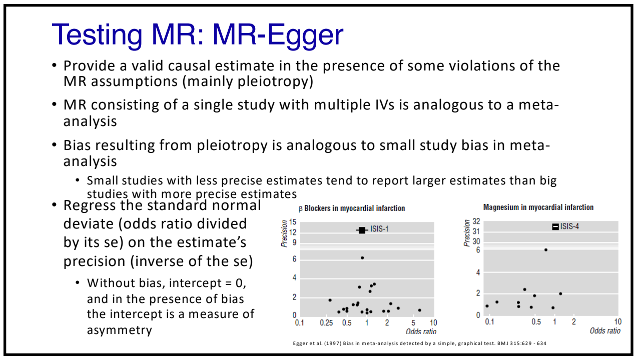
- Calculate the Wald ratio for each instrument
- Select the median value according to the weighted method

	$\hat{\beta}_1$	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Simple median										
Weight (w_j)	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Percentile (p_j)	5	15	25	35	45	55	65	75	85	95
Weighting 1										
Weight (w_j)	$\frac{1}{30}$	$\frac{2}{30}$	$\frac{3}{30}$	$\frac{4}{30}$	$\frac{5}{30}$	$\frac{5}{30}$	$\frac{4}{30}$	$\frac{3}{30}$	$\frac{2}{30}$	$\frac{1}{30}$
Percentile	1.67	6.67	15.00	26.67	41.67	58.33	73.33	85.00	93.33	98.33
Weighting 2										
Weight (w_j)	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
Percentile (p_j)	2.78	9.72	27.78	52.78	70.83	81.94	88.89	93.06	95.83	98.61

Bowden et al. (2016) Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epi, 40: 304-314.

- 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

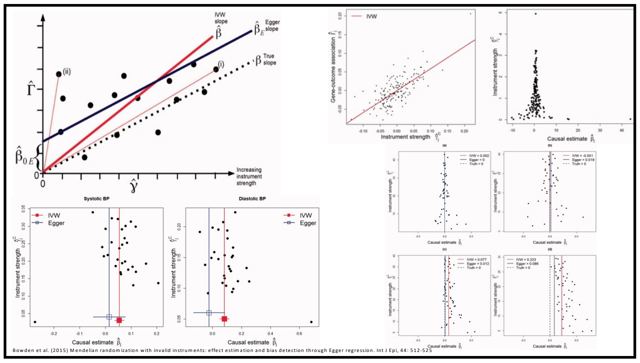


Testing MR: MR-Egger

- 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

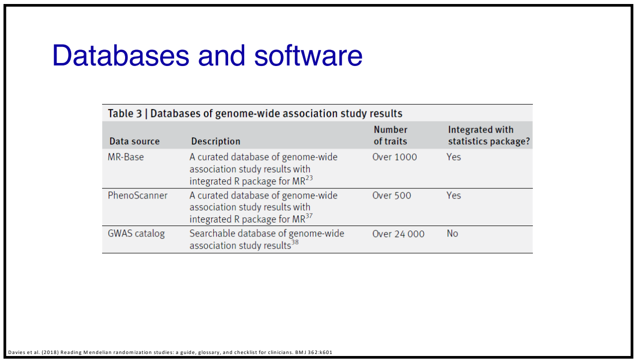
Egger et al. (1997) Bias in meta-analysis detected by a simple, graphical test. BMJ 315:629 - 634

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 randomization 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. Åsvold[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

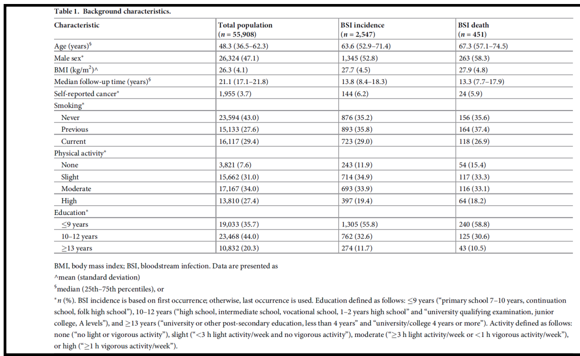


Table 1. Background characteristics.

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

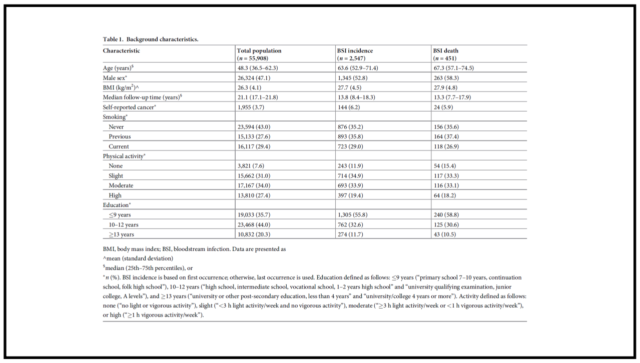


Table 1. Background characteristics.

Characteristic	Total population (n = 55,908)	BSI incidence (n = 2,547)	BSI death (n = 451)
Age (years)$^{\$}$	48.3 (36.5–62.3)	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)

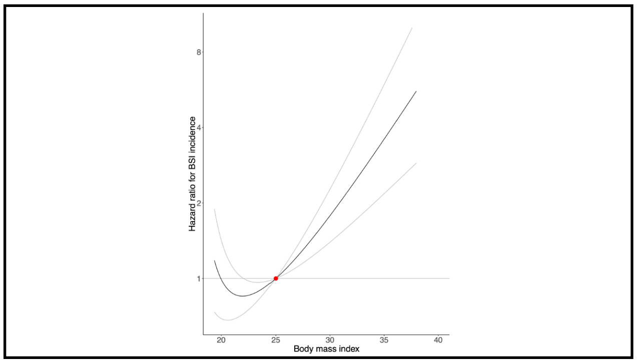
BMI, body mass index; BSI, bloodstream infection. Data are presented as

^mean (standard deviation)

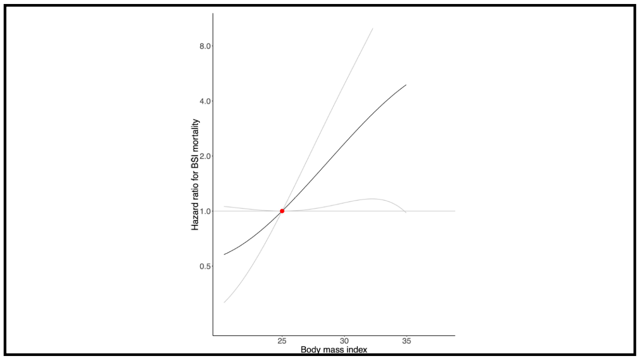
$^{\$}$median (25th–75th percentiles), or

*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").

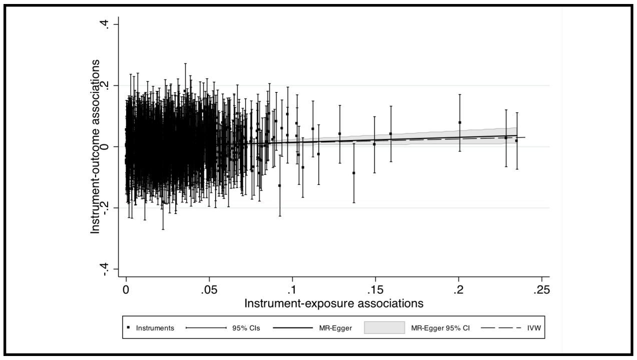
31



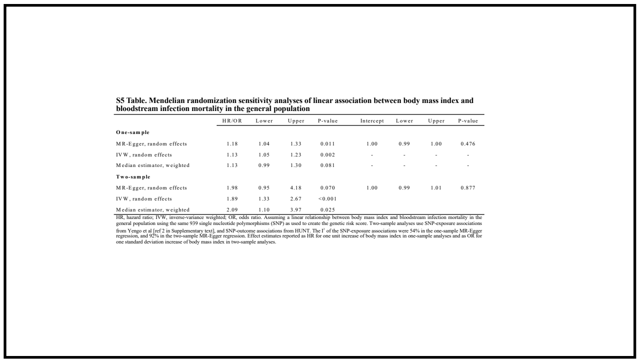
32



33



34

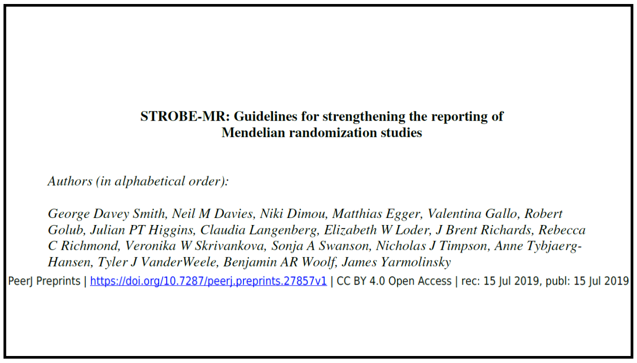


S5 Table. Mendelian randomization sensitivity analyses of linear association between body mass index and bloodstream infection mortality in the general population

	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 939 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 I^2 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

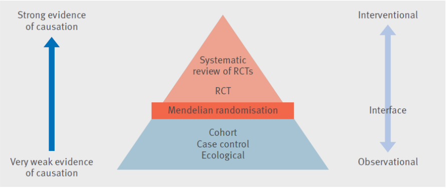


STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies

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

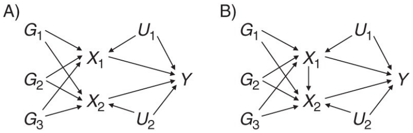
36



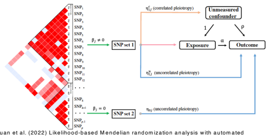
Davies et al. (2018) Reading Mendelian randomization studies: a guide, glossary, and checklist for clinicians. BMJ 362:k601

37

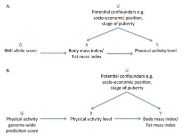
Some Advanced MR analysis approaches



Burgess and Thompson (2015) Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. Am J Epidemiol, 181(4): 251–260.



Yuan et al. (2022) Likelihood-based Mendelian randomization analysis with automated instrument selection and horizontal pleiotropic modeling. Sci. Adv. 8, eabl5744 (2022)

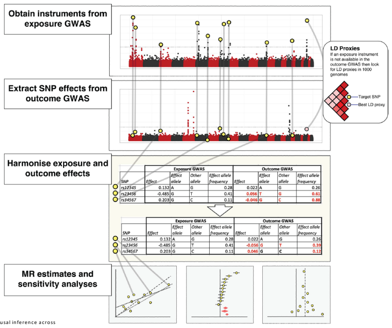

A.
U
Potential confounders e.g. socio-economic position, stage of puberty
G
BMI allelic score
X
Body mass index/ Fat mass index
Y
Physical activity level
B.
U
Potential confounders e.g. socio-economic position, stage of puberty
G
Physical activity genome-wide prediction score
X
Physical activity level
Y
Body mass index/ Fat mass index

Richmond et al. (2014) Assessing Causality in the Association between Child Adiposity and Physical Activity Levels: A Mendelian Randomization Analysis. PLoS Med 11(3): e1001618.

38

MRBASE

http://app.mrbase.org/


Obtain instruments from exposure GWAS
Extract SNP effects from outcome GWAS
LD Proxies
If an exposure instrument is not available in the outcome GWAS then look for LD proxies in 1000 genomes
Target SNP
Best LD proxy
Harmonise exposure and outcome effects
MR estimates and sensitivity analyses

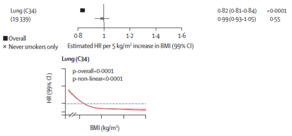
	Exposure GWAS				Outcome GWAS			
SNP	Effect	Effect allele	Other allele	Effect allele frequency	Effect	Effect allele	Other allele	Effect allele frequency
rs12345	0.132	A	G	0.28	0.022	A	G	0.26
rs23456	-0.485	G	T	0.41	0.056	T	G	0.61
rs34567	0.203	G	C	0.11	-0.046	G	C	0.88

	Exposure GWAS				Outcome GWAS			
SNP	Effect	Effect allele	Other allele	Effect allele frequency	Effect	Effect allele	Other allele	Effect allele frequency
rs12345	0.132	A	G	0.28	0.022	A	G	0.26
rs23456	-0.485	G	T	0.41	-0.056	G	T	0.39
rs34567	0.203	G	C	0.11	0.046	G	C	0.12

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

39

BMI and Lung Cancer



Bhaskaran 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

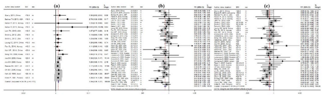
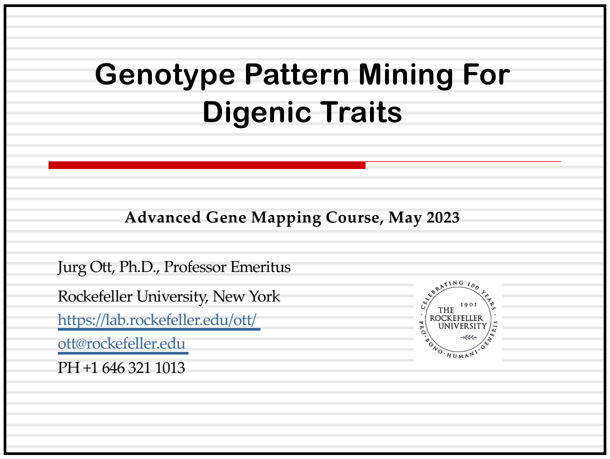


Figure 1. Forest plot of relative risks of underweight, overweight and obesity vs. normal weight for BMI with LC risk. Open blue diamonds denote the summary relative risks. Black diamonds indicate the RR in each study. The size of each box indicates the relative weight of each study in the meta-analysis. Horizontal lines represent the 95% confidence intervals (CIs). (a) Forest plots of risk of lung cancer associated with underweight (BMI ≤ 18.5 kg/m^2). (b) Forest plots of risk of lung cancer associated with overweight (BMI = 25–29.9 kg/m^2). (c) Forest plots of risk of lung cancer associated with obesity (BMI ≥ 30 kg/m^2). RR, relative risk; normal weight, BMI = 18.5–24.9 kg/m^2; 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 5:16938

40



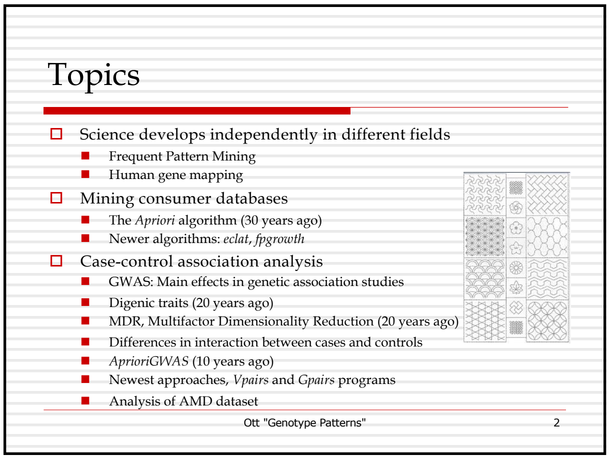
Genotype Pattern Mining For Digenic Traits

Advanced Gene Mapping Course, May 2023

Jurg Ott, Ph.D., Professor Emeritus

Rockefeller University, New York
https://lab.rockefeller.edu/ott/
ott@rockefeller.edu
PH +1 646 321 1013

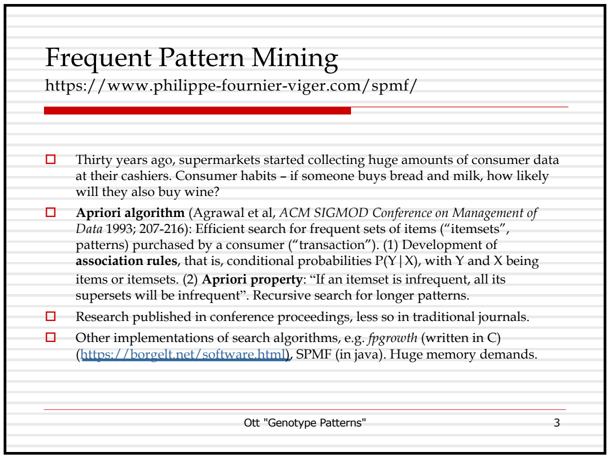
1



Topics

- □ Science develops independently in different fields
  - ■ Frequent Pattern Mining
  - ■ Human gene mapping
- □ Mining consumer databases
  - ■ The *Apriori* algorithm (30 years ago)
  - ■ Newer algorithms: *eclat*, *fpgrowth*
- □ Case-control association analysis
  - ■ GWAS: Main effects in genetic association studies
  - ■ Digenic traits (20 years ago)
  - ■ MDR, Multifactor Dimensionality Reduction (20 years ago)
  - ■ Differences in interaction between cases and controls
  - ■ *AprioriGWAS* (10 years ago)
  - ■ Newest approaches, *Vpairs* and *Gpairs* programs
  - ■ Analysis of AMD dataset

2

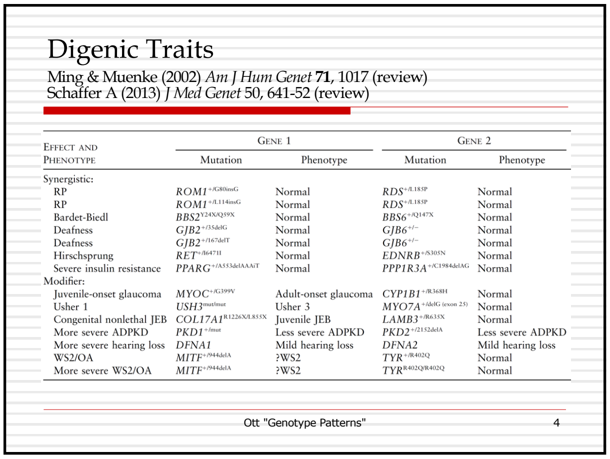


Frequent Pattern Mining

https://www.philippe-fournier-viger.com/spmf/

- □ Thirty years ago, supermarkets started collecting huge amounts of consumer data at their cashiers. Consumer habits – if someone buys bread and milk, how likely will they also buy wine?
- □ **Apriori algorithm** (Agrawal et al, *ACM SIGMOD Conference on Management of Data* 1993; 207-216): Efficient search for frequent sets of items ("itemsets", patterns) purchased by a consumer ("transaction"). (1) Development of **association rules**, that is, conditional probabilities P(Y|X), with Y and X being items or itemsets. (2) **Apriori property**: "If an itemset is infrequent, all its supersets will be infrequent". Recursive search for longer patterns.
- □ Research published in conference proceedings, less so in traditional journals.
- □ Other implementations of search algorithms, e.g. *fpgrowth* (written in C) (https://borgelt.net/software.html), SPMF (in java). Huge memory demands.

3

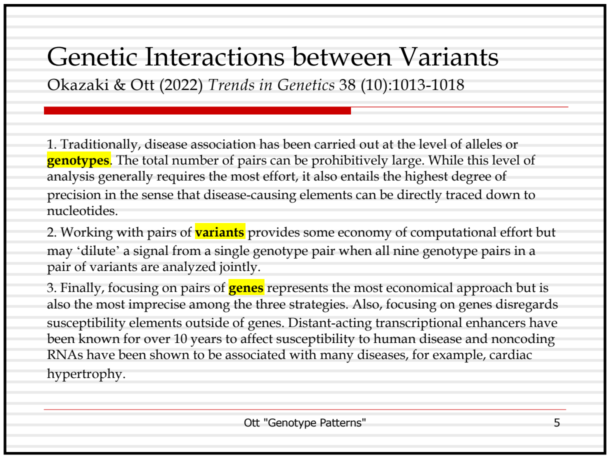


Digenic Traits

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*$^{Y24X/Q59X}$	Normal	*BBS6*$^{+/Q147X}$	Normal
Deafness	*GJB2*$^{+/35delG}$	Normal	*GJB6*$^{+/-}$	Normal
Deafness	*GJB2*$^{+/167delT}$	Normal	*GJB6*$^{+/-}$	Normal
Hirschsprung	*RET*$^{+/I647I}$	Normal	*EDNRB*$^{+/S305N}$	Normal
Severe insulin resistance	*PPARG*$^{+/A553delAAAiT}$	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*$^{+/944delA}$	?WS2	*TYR*$^{+/R402Q}$	Normal
More severe WS2/OA	*MITF*$^{+/944delA}$	?WS2	*TYR*$^{R402Q/R402Q}$	Normal

4



Genetic Interactions between Variants

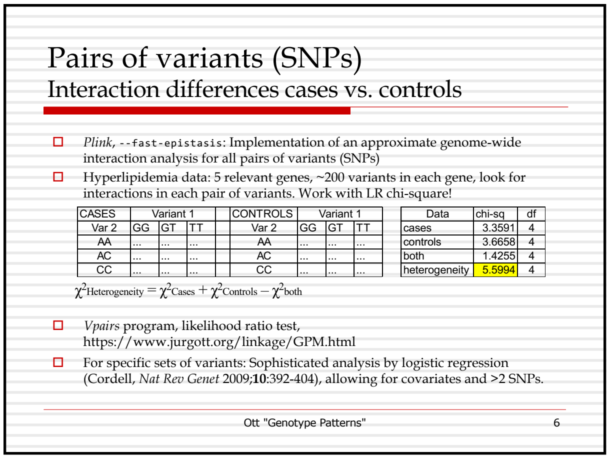
Okazaki & Ott (2022) *Trends in Genetics* 38 (10):1013-1018

1. Traditionally, disease association has been carried out at 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 degree 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.

5



Pairs of variants (SNPs)
Interaction differences cases vs. controls

- □ *Plink*, `--fast-epistasis`: Implementation of an approximate genome-wide interaction analysis for all pairs of variants (SNPs)
- □ Hyperlipidemia data: 5 relevant genes, ~200 variants in each gene, look for interactions in each pair of variants. Work with LR chi-square!

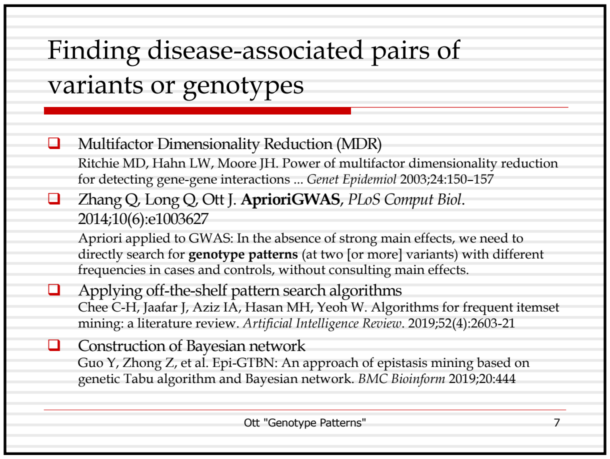
CASES	Variant 1				CONTROLS	Variant 1		
Var 2	GG	GT	TT		Var 2	GG	GT	TT
AA	...	...	...		AA	...	...	...
AC	...	...	...		AC	...	...	...
CC	...	...	...		CC	...	...	...

Data	chi-sq	df
cases	3.3591	4
controls	3.6658	4
both	1.4255	4
heterogeneity	5.5994	4

$$\chi^2_{\text{Heterogeneity}} = \chi^2_{\text{Cases}} + \chi^2_{\text{Controls}} - \chi^2_{\text{both}}$$

- □ *Vpairs* program, likelihood ratio test, https://www.jurgott.org/linkage/GPM.html
- □ For specific sets of variants: Sophisticated analysis by logistic regression (Cordell, *Nat Rev Genet* 2009;**10**:392-404), allowing for covariates and >2 SNPs.

6

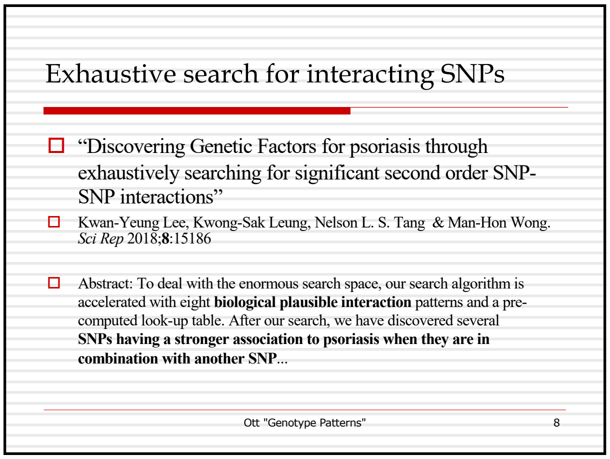


Finding disease-associated pairs of variants or genotypes

- ❑ 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
- ❑ Zhang Q, Long Q, Ott J. **AprioriGWAS**, *PLoS Comput Biol*. 2014;10(6):e1003627
  Apriori applied to GWAS: 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.
- ❑ 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
- ❑ 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

Ott "Genotype Patterns" 7

7

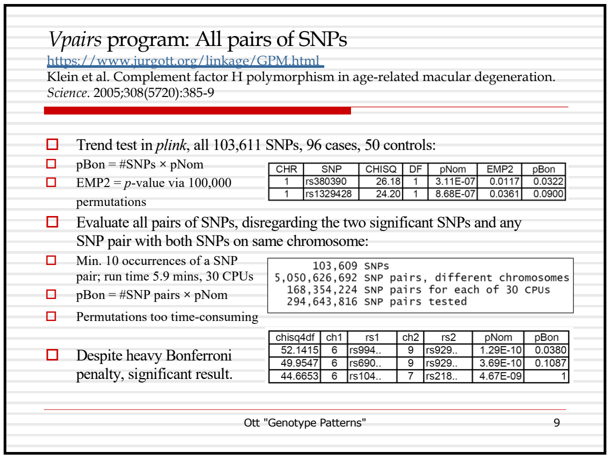


Exhaustive search for interacting SNPs

- □ "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**...

Ott "Genotype Patterns" 8

8



Vpairs program: All pairs of SNPs

https://www.jurgott.org/linkage/GPM.html

Klein et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308(5720):385-9

- □ Trend test in *plink*, all 103,611 SNPs, 96 cases, 50 controls:
- □ pBon = #SNPs × pNom
- □ EMP2 = *p*-value via 100,000 permutations

CHR	SNP	CHISQ	DF	pNom	EMP2	pBon
1	rs380390	26.18	1	3.11E-07	0.0117	0.0322
1	rs1329428	24.20	1	8.68E-07	0.0361	0.0900

- □ Evaluate all pairs of SNPs, disregarding the two significant SNPs and any SNP pair with both SNPs on same chromosome:
- □ Min. 10 occurrences of a SNP pair; run time 5.9 mins, 30 CPUs
- □ pBon = #SNP pairs × pNom
- □ Permutations too time-consuming

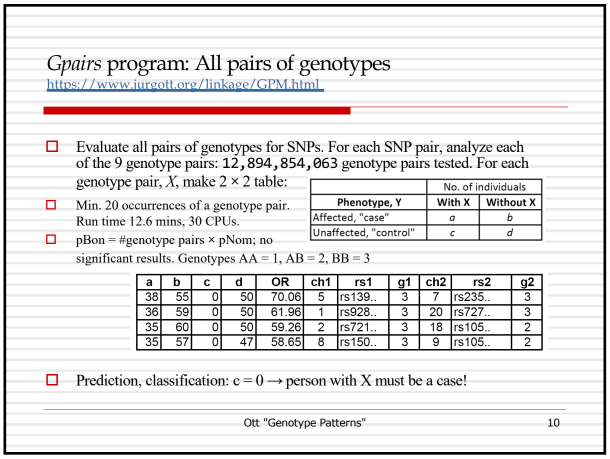
```
      103,609 SNPs
5,050,626,692 SNP pairs, different chromosomes
  168,354,224 SNP pairs for each of 30 CPUs
  294,643,816 SNP pairs tested
```

- □ Despite heavy Bonferroni penalty, significant result.

chisq4df	ch1	rs1	ch2	rs2	pNom	pBon
52.1415	6	rs994..	9	rs929..	1.29E-10	0.0380
49.9547	6	rs690..	9	rs929..	3.69E-10	0.1087
44.6653	6	rs104..	7	rs218..	4.67E-09	1

Ott "Genotype Patterns" 9

9



Gpairs program: All pairs of genotypes

https://www.jurgott.org/linkage/GPM.html

- □ Evaluate all pairs of genotypes for SNPs. For each SNP pair, analyze each of the 9 genotype pairs: `12,894,854,063` genotype pairs tested. For each genotype pair, *X*, make 2 × 2 table:

	No. of individuals	
Phenotype, Y	**With X**	**Without X**
Affected, "case"	*a*	*b*
Unaffected, "control"	*c*	*d*

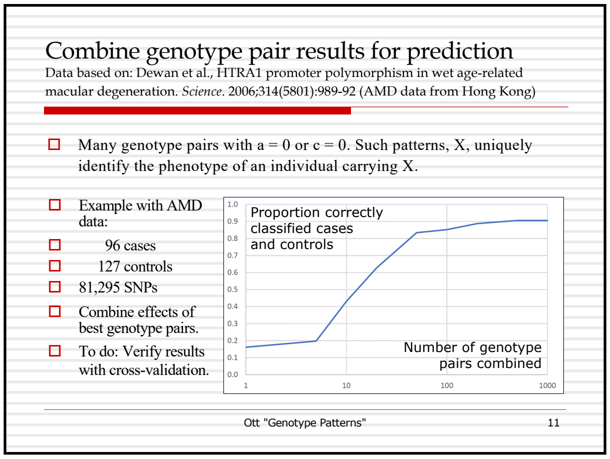
- □ Min. 20 occurrences of a genotype pair. Run time 12.6 mins, 30 CPUs.
- □ pBon = #genotype pairs × pNom; no significant results. Genotypes AA = 1, AB = 2, BB = 3

a	b	c	d	OR	ch1	rs1	g1	ch2	rs2	g2
38	55	0	50	70.06	5	rs139..	3	7	rs235..	3
36	59	0	50	61.96	1	rs928..	3	20	rs727..	3
35	60	0	50	59.26	2	rs721..	3	18	rs105..	2
35	57	0	47	58.65	8	rs150..	3	9	rs105..	2

- □ Prediction, classification: c = 0 → person with X must be a case!

Ott "Genotype Patterns" 10

10



Combine genotype pair results for prediction

Data based on: Dewan et al., HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*. 2006;314(5801):989-92 (AMD data from Hong Kong)

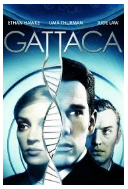
- □ Many genotype pairs with a = 0 or c = 0. Such patterns, X, uniquely identify the phenotype of an individual carrying X.
- □ Example with AMD data:
- □ 96 cases
- □ 127 controls
- □ 81,295 SNPs
- □ Combine effects of best genotype pairs.
- □ To do: Verify results with cross-validation.

Proportion correctly classified cases and controls
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
10
100
1000
Number of genotype pairs combined

Ott "Genotype Patterns" 11

11

Genetic risk prediction



Genotype of an individual → Life-time risk of genetic disorders

(Common SNPs) (Common complex genetic disorders)

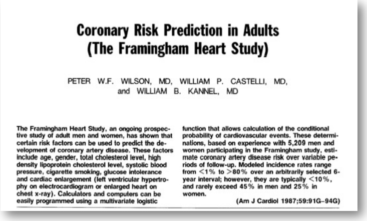
1

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)

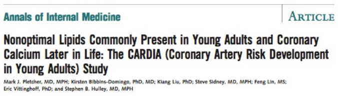
PETER W.F. WILSON, MD, WILLIAM P. CASTELLI, MD, and WILLIAM B. KANNEL, MD

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 6-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



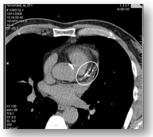
Annals of Internal Medicine | ARTICLE

Nonoptimal Lipids Commonly Present in Young Adults and Coronary Calcium Later in Life: The CARDIA (Coronary Artery Risk Development in Young Adults) Study

Mark J. Pletcher, MD, MPH; Kirsten Bibbins-Domingo, PhD, MD; Kiang Liu, PhD; Steve Sidney, MD, MPH; Feng Lin, MS; Eric Vittinghoff, PhD; and Stephen B. Hulley, MD, MPH

~3500 subjects < 35 years old

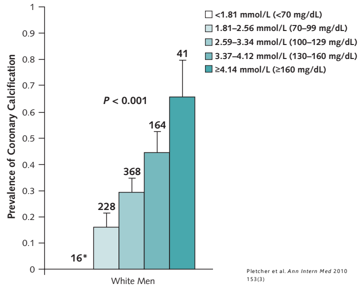
15-20 years →



Piers *et al. BMC Cardiovascular Disorders* 2008 **8**:38

4

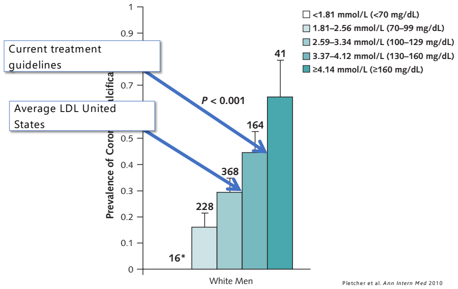
LDL levels and risk of disease



Pletcher et al. *Ann Intern Med* 2010 153(3)

5

LDL levels and risk of disease



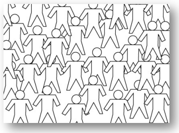
Current treatment guidelines

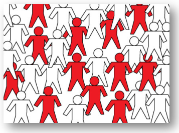
Average LDL United States

Pletcher et al. *Ann Intern Med* 2010

6

Selecting populations for treatment





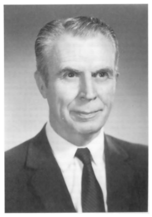
7

Why estimate genetic risk?

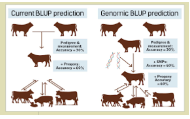
- An estimate of the long-term risk at birth
- Genetic risk can be combined with biomarkers and clinical features
- Genetics explains about 50% of risk. One cannot predict risk any better than that but 50% is a non-trivial proportion of risk

8

BLUP – Best Linear Unbiased Predictor

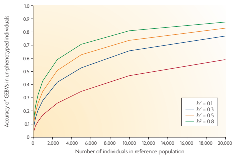


- Infinitesimal model
- Genetic effects are random
- Predict the expected genetic effect


Current BLUP prediction
Pedigree & measurement: Accuracy = 30%
+ Progeny: Accuracy = 60%
Genomic BLUP prediction
Pedigree & measurement: Accuracy = 30%
+ SNPs: Accuracy = 60%
+ Progeny Accuracy = 60%

9

Accuracy of polygenic prediction in cattle



Poor transferability between breeds!

10

Applications in humans

GENOME RESEARCH

Prediction of individual genetic risk to disease from genome-wide association studies

Naomi R. Wray, Michael E. Goddard and Peter M. Visscher

Genome Res. 2007 17: 1520-1528; originally published online Sep 4, 2007;
Access the most recent version at doi:10.1101/gr.6665407

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

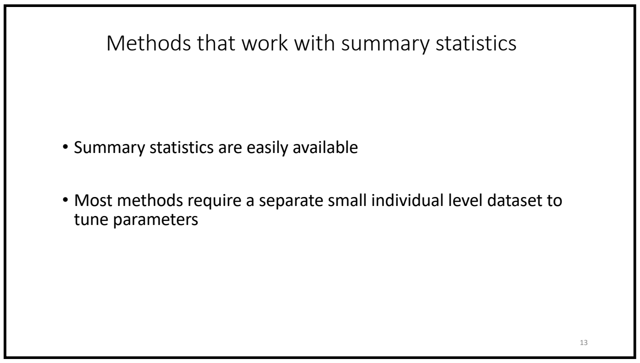
- LD-prune
- Exclude SNPs of very small effect

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

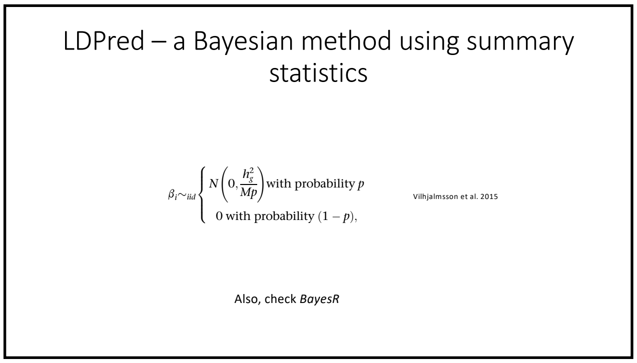


Methods that work with summary statistics

- Summary statistics are easily available
- Most methods require a separate small individual level dataset to tune parameters

13

13



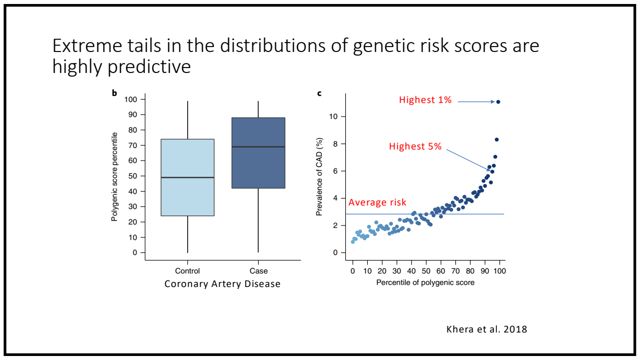
LDPred – a Bayesian method using summary statistics

$$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) \text{with probability } p \\ 0 \text{ with probability } (1-p), \end{cases}$$

Vilhjalmsson et al. 2015

Also, check *BayesR*

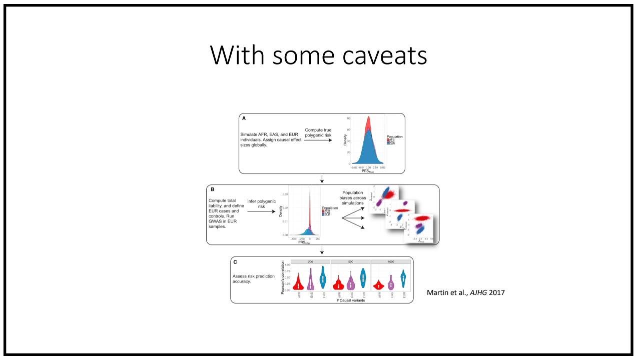
14



Extreme tails in the distributions of genetic risk scores are highly predictive

Khera et al. 2018

15



With some caveats

A

Simulate AFR, EAS, and EUR individuals. Assign causal effect sizes globally.

Compute true polygenic risk

B

Compute total liability, and define EUR cases and controls. Run GWAS in EUR samples.

Infer polygenic risk

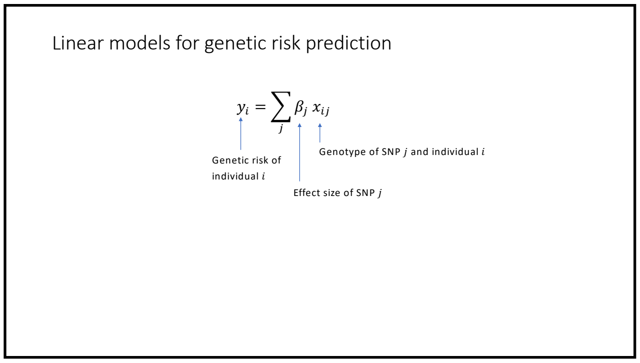
Population biases across simulations

C

Assess risk prediction accuracy.

Martin et al., *AJHG* 2017

16



Linear models for genetic risk prediction

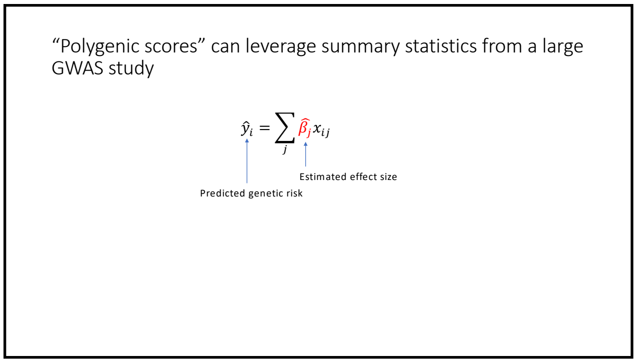
$$y_i = \sum_j \beta_j \, x_{ij}$$

Genetic risk of individual i

Effect size of SNP j

Genotype of SNP j and individual i

17



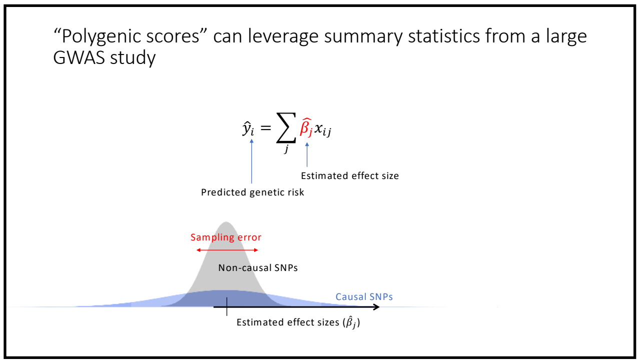
"Polygenic scores" can leverage summary statistics from a large GWAS study

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$

Predicted genetic risk

Estimated effect size

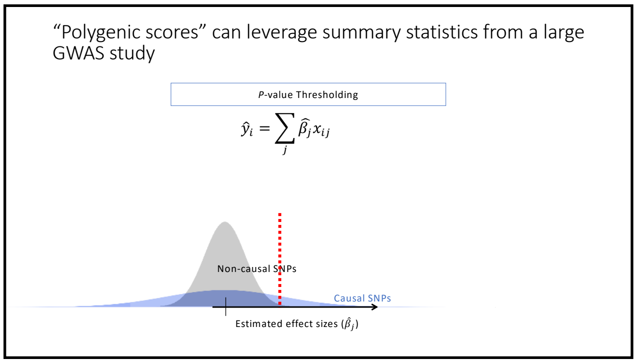
18



“Polygenic scores” can leverage summary statistics from a large GWAS study

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$

19

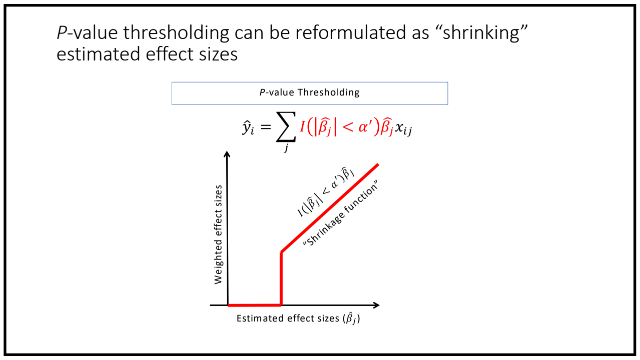


“Polygenic scores” can leverage summary statistics from a large GWAS study

P-value Thresholding

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$

20

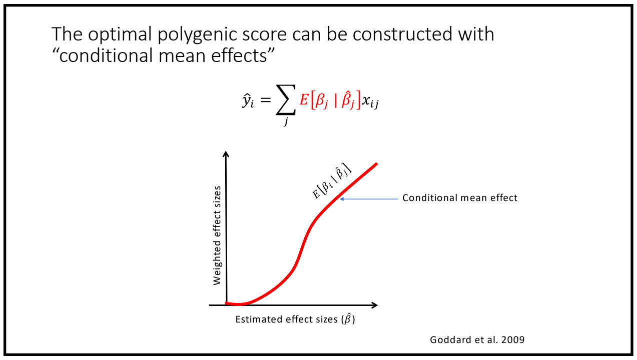


P-value thresholding can be reformulated as “shrinking” estimated effect sizes

P-value Thresholding

$$\hat{y}_i = \sum_j I(|\hat{\beta}_j| < \alpha')\hat{\beta}_j x_{ij}$$

21

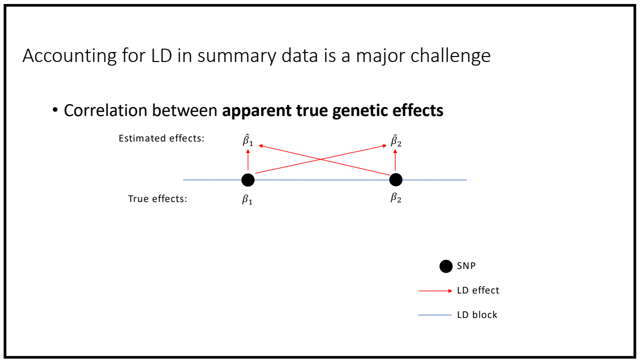


The optimal polygenic score can be constructed with “conditional mean effects”

$$\hat{y}_i = \sum_j E[\beta_j \mid \hat{\beta}_j] x_{ij}$$

Goddard et al. 2009

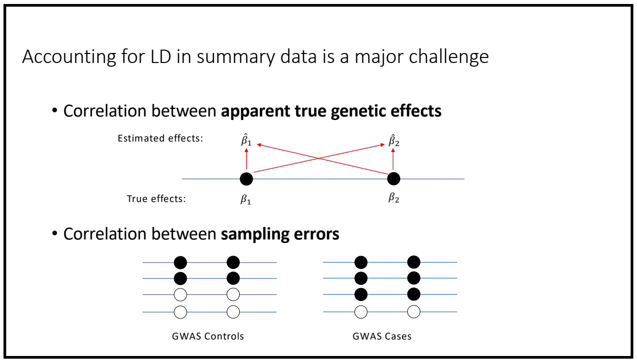
22



Accounting for LD in summary data is a major challenge

- Correlation between **apparent true genetic effects**

23

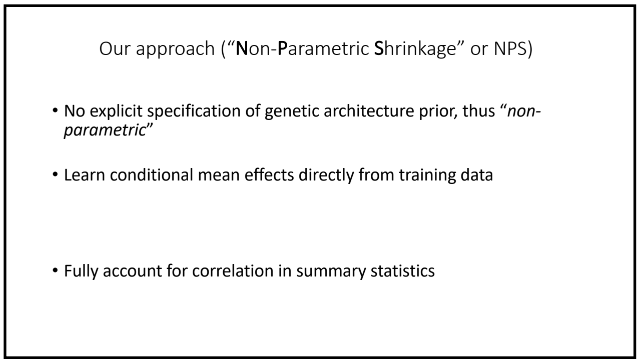


Accounting for LD in summary data is a major challenge

- Correlation between **apparent true genetic effects**

- Correlation between **sampling errors**

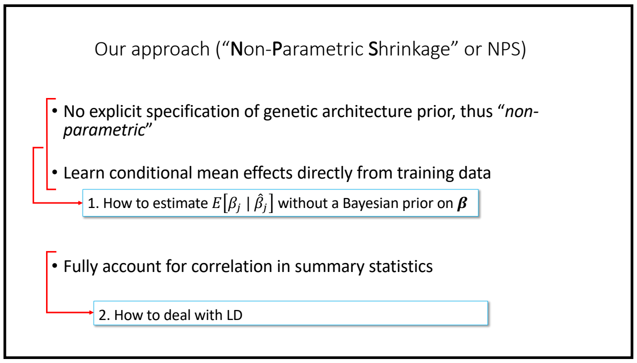
24



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)

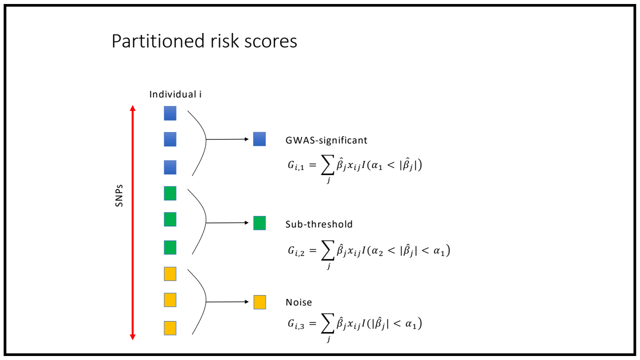
- No explicit specification of genetic architecture prior, thus “*non-parametric*”
- Learn conditional mean effects directly from training data

1. How to estimate $E\left[\beta_j \mid \hat{\beta}_j\right]$ without a Bayesian prior on $\boldsymbol{\beta}$

- Fully account for correlation in summary statistics

2. How to deal with LD

26



Partitioned risk scores

GWAS-significant

$$G_{i,1} = \sum_j \hat{\beta}_j x_{ij} I(\alpha_1 < |\hat{\beta}_j|)$$

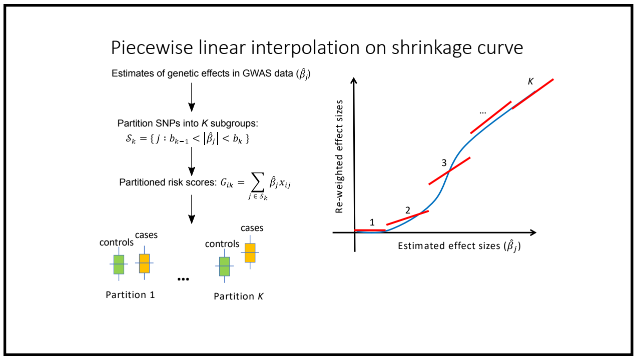
Sub-threshold

$$G_{i,2} = \sum_j \hat{\beta}_j x_{ij} I(\alpha_2 < |\hat{\beta}_j| < \alpha_1)$$

Noise

$$G_{i,3} = \sum_j \hat{\beta}_j x_{ij} I(|\hat{\beta}_j| < \alpha_1)$$

27



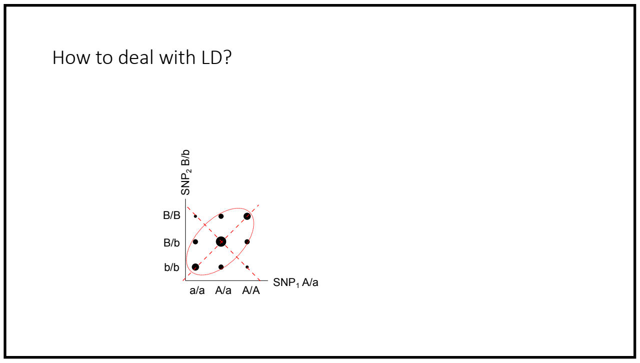
Piecewise linear interpolation on shrinkage curve

Estimates of genetic effects in GWAS data ($\hat{\beta}_j$)

Partition SNPs into K subgroups:
$\mathcal{S}_k = \{\, j : b_{k-1} < |\hat{\beta}_j| < b_k \,\}$

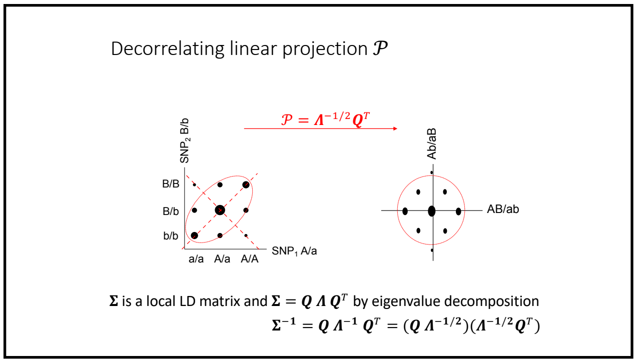
Partitioned risk scores: $G_{ik} = \sum_{j \in \mathcal{S}_k} \hat{\beta}_j x_{ij}$

28



How to deal with LD?

29



Decorrelating linear projection $\mathcal{P}$

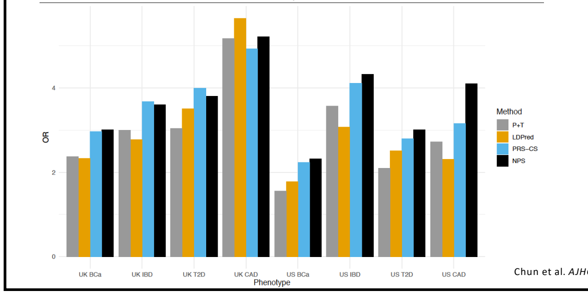
$$\mathcal{P} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^T$$

$\boldsymbol{\Sigma}$ 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

Accuracy of the 5% tail



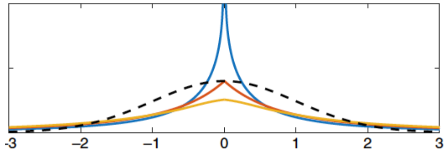
Chun et al. *AJHG* 2020

31

Other shrinkage methods: PRS-CS

$$\beta_j \sim \mathrm{N}\left(0, \frac{\sigma^2}{N}\phi\psi_j\right), \qquad \psi_j \sim g,$$

Prior density of β_j: central region



32

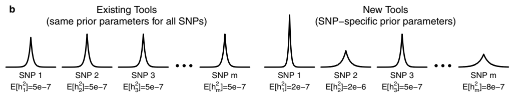
BayesR

$$\beta_j|\pi, \sigma_\beta^2 = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0, \gamma_2\sigma_\beta^2) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C\sigma_\beta^2) & \text{with probability } 1 - \sum_{c=1}^{C-1} \pi_c, \end{cases}$$

Lassosum – extension of *LASSO*

33

LDAK-Bolt-Predict



34

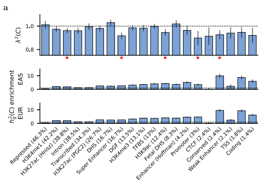
What makes PRS non-transferrable?

- Differences in allele frequencies between populations
- Differences in LD between populations
- Differences in effect sizes (although likely a minor contribution)

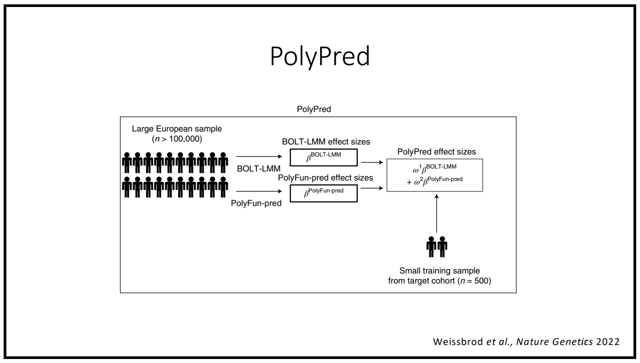
35

Slight differences in genetic effects between populations

Genetic correlations between populations are close but not equal to 1.
They are not uniformly distributed along the genome.



36



PolyPred

PolyPred

Large European sample
(n > 100,000)

BOLT-LMM

PolyFun-pred

BOLT-LMM effect sizes

$\beta^{\text{BOLT-LMM}}$

PolyFun-pred effect sizes

$\beta^{\text{PolyFun-pred}}$

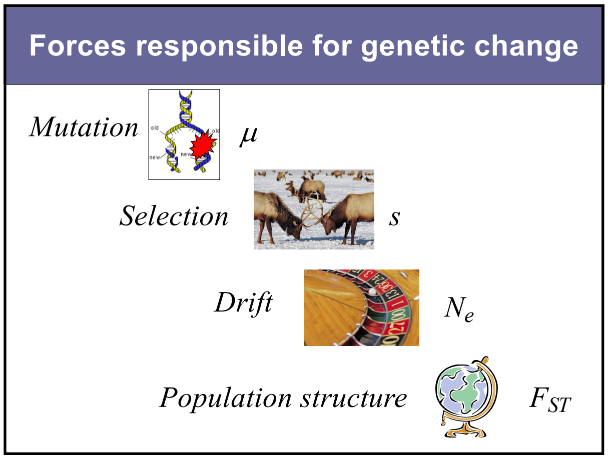
PolyPred effect sizes

$\omega^1\beta^{\text{BOLT-LMM}}$
$+ \omega^2\beta^{\text{PolyFun-pred}}$

Small training sample
from target cohort (n = 500)

Weissbrod *et al., Nature Genetics* 2022

37



Forces responsible for genetic change

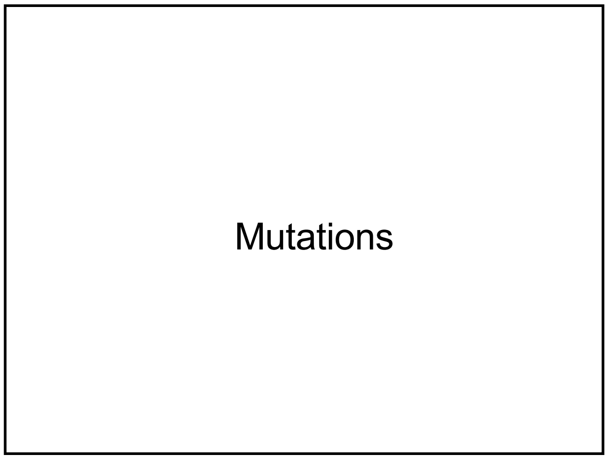
Mutation μ

Selection s

Drift N_e

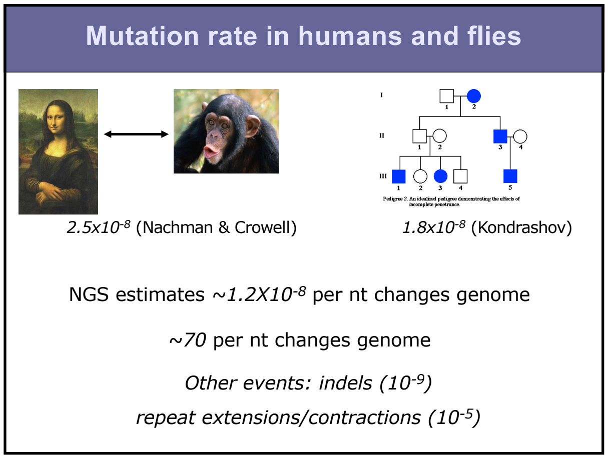
Population structure F_{ST}

1



Mutations

2



Mutation rate in humans and flies

2.5x10^{-8} (Nachman & Crowell)

1.8x10^{-8} (Kondrashov)

I
II
III
Pedigree 2. An idealized pedigree demonstrating the effects of incomplete penetrance.

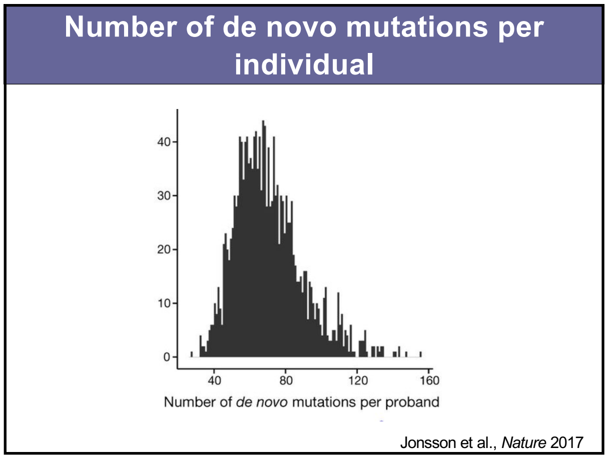
NGS estimates *~1.2X10^{-8}* per nt changes genome

~70 per nt changes genome

Other events: indels (10^{-9})

repeat extensions/contractions (10^{-5})

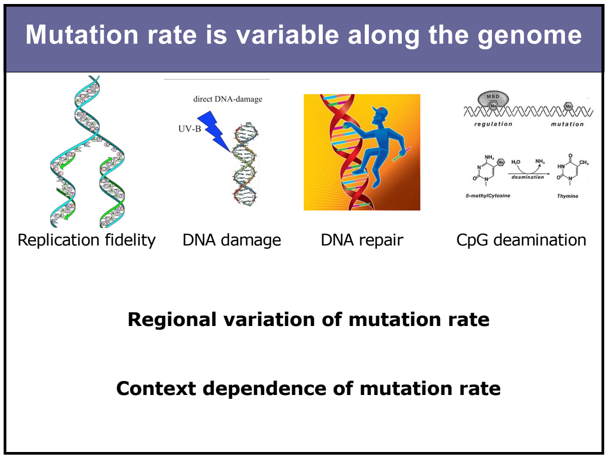
3



Number of de novo mutations per individual

Jonsson et al., *Nature* 2017

4



Mutation rate is variable along the genome

Replication fidelity

DNA damage

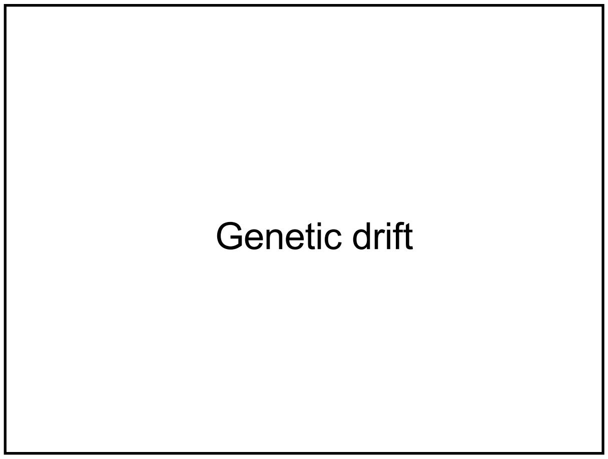
DNA repair

CpG deamination

Regional variation of mutation rate

Context dependence of mutation rate

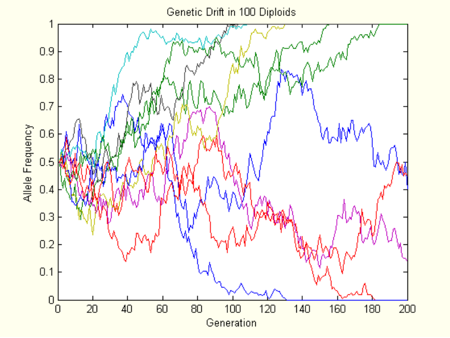
5



Genetic drift

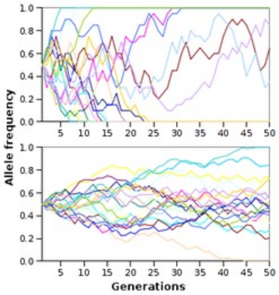
6

Drift is a random change of allele frequencies



7

Drift depends on population size



8

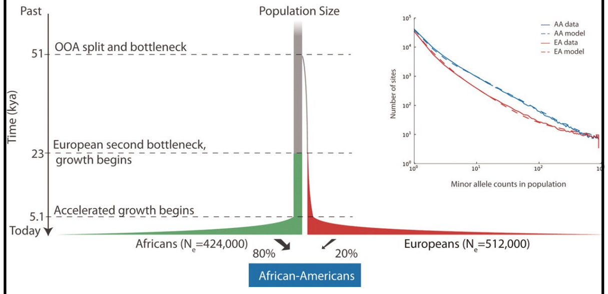
Effective population size

- 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



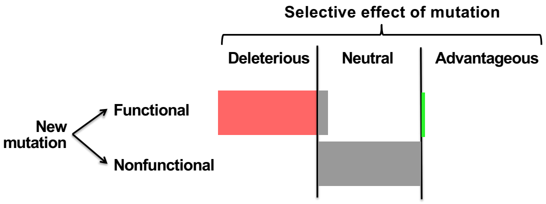
Tennessen et al. *Science* 2012

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

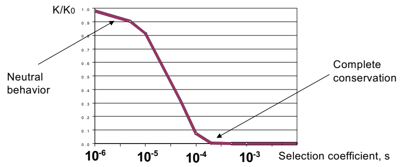
Conservation can be due to very weak selection!

Every new mutation eventually will be either fixed or lost

$$K = K_0 2 N_e \frac{(1-e^{-2s})}{(1-e^{-4N_e s})}$$

s – selection coefficient
N_e - effective population size

For humans estimated to be ~ 10 000



15

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)

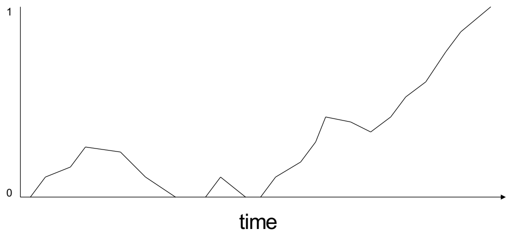
16

Methods of mathematical population genetics

17

Dynamic of allelic substitution

Mathematically, allele frequency change in a population follows a one-dimensional random walk



18

Diffusion approximation

Random walk that does not jump long distances can be approximated by a diffusion process

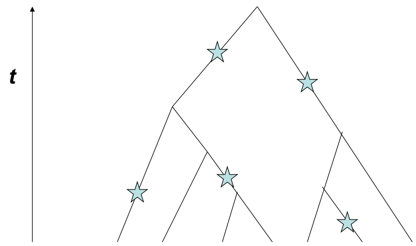
$$\frac{\partial \phi(x,p,t)}{\partial t} = -\frac{\partial M\phi(x,p,t)}{\partial x} + \frac{1}{2}\frac{\partial^2 V\phi(x,p,t)}{\partial x^2}$$

19

Coalescent theory

Instead of modeling a population, we can model our sample

Time goes backwards !



20

Signatures of purifying selection

Reduced variation

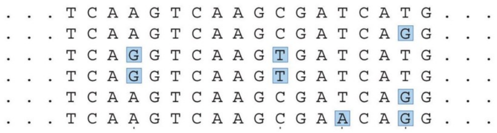
Excess of rare alleles

21

Commonly used summary statistics to characterize variation

22

Number of segregating sites



```
. . . T C A A G T C A A G C G A T C A T G . . .
. . . T C A A G T C A A G C G A T C A G G . . .
. . . T C A G G T C A A G T G A T C A T G . . .
. . . T C A G G T C A A G T G A T C A T G . . .
. . . T C A A G T C A A G C G A T C A G G . . .
. . . T C A A G T C A A G C G A A C A G G . . .
```

k – number of sites variable in the sample

density of segregating sites is also frequently used

k is dominated by rare alleles

k strongly depend on sample size

23

Nucleotide diversity

$$\pi = \frac{2}{n(n-1)}\sum d_{ij}$$ ***d*$_{ij}$** - number of nucleotide differenced between sequences ***i*** and ***j***

$$\pi = \frac{n}{(n-1)}\sum 2p_k(1-p_k)$$ ***p*$_k$** – allele frequency at site ***k***

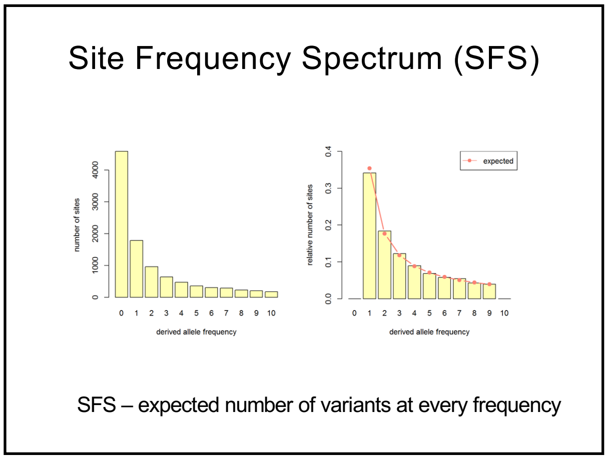
π – the average density of nucleotide differences between two sequences

π – per nucleotide heterozygosity

π is dominated by common alleles

π is independent of sample size

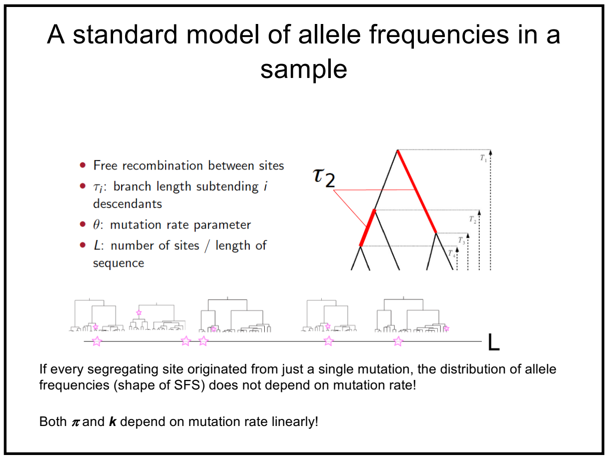
24



Site Frequency Spectrum (SFS)

SFS – expected number of variants at every frequency

25



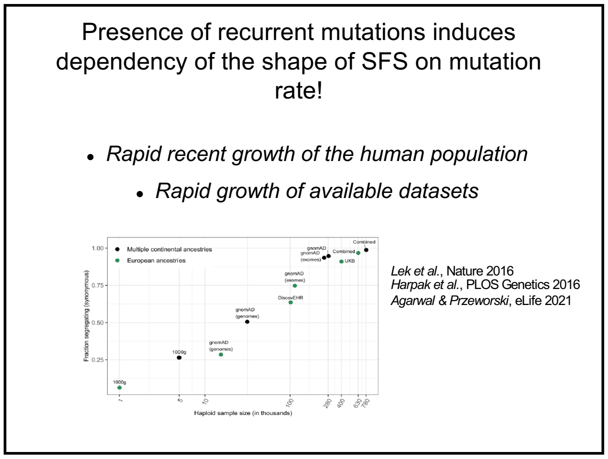
A standard model of allele frequencies in a sample

- Free recombination between sites
- τ_i: branch length subtending i descendants
- θ: mutation rate parameter
- L: number of sites / length of sequence

If every segregating site originated from just a single mutation, the distribution of allele frequencies (shape of SFS) does not depend on mutation rate!

Both ***π*** and ***k*** depend on mutation rate linearly!

26

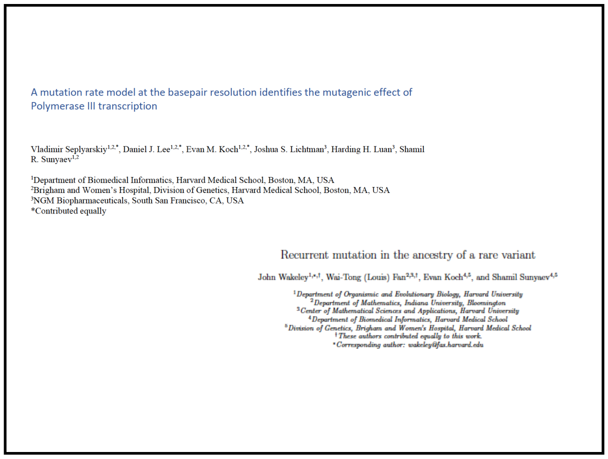


Presence of recurrent mutations induces dependency of the shape of SFS on mutation rate!

- *Rapid recent growth of the human population*
- *Rapid growth of available datasets*

Lek et al., Nature 2016
Harpak et al., PLOS Genetics 2016
Agarwal & Przeworski, eLife 2021

27



A mutation rate model at the basepair resolution identifies the mutagenic effect of Polymerase III transcription

Vladimir Seplyarskiy[1,2,*], Daniel J. Lee[1,2,*], Evan M. Koch[1,2,*], Joshua S. Lichtman[3], Harding H. Luan[3], Shamil R. Sunyaev[1,2]

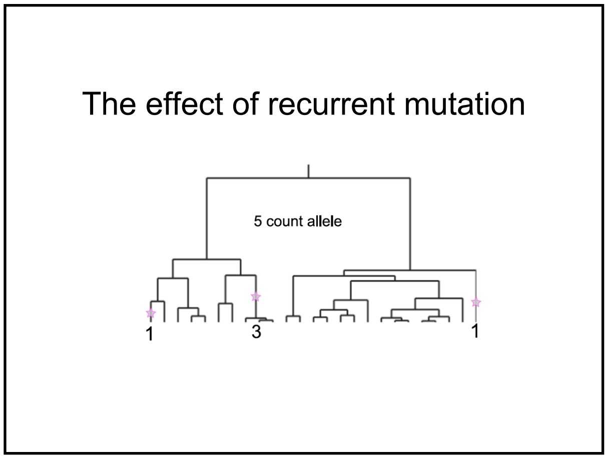
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Brigham and Women's Hospital, Division of Genetics, Harvard Medical School, Boston, MA, USA
[3]NGM Biopharmaceuticals, South San Francisco, CA, USA
*Contributed equally

Recurrent mutation in the ancestry of a rare variant

John Wakeley[1,*,†], Wai-Tong (Louis) Fan[2,3,†], Evan Koch[4,5], and Shamil Sunyaev[4,5]

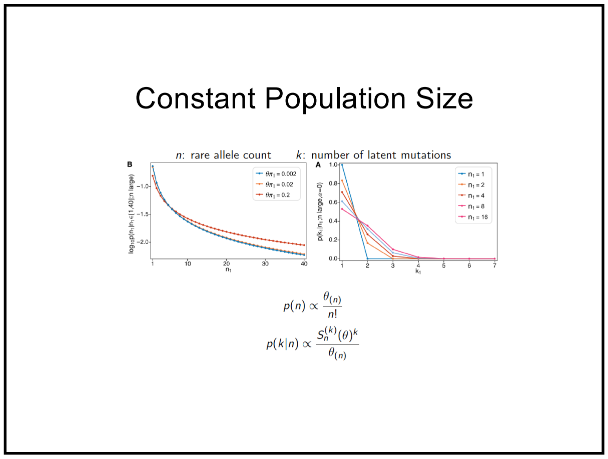
[1]*Department of Organismic and Evolutionary Biology, Harvard University*
[2]*Department of Mathematics, Indiana University, Bloomington*
[3]*Center of Mathematical Sciences and Applications, Harvard University*
[4]*Department of Biomedical Informatics, Harvard Medical School*
[5]*Division of Genetics, Brigham and Women's Hospital, Harvard Medical School*
†*These authors contributed equally to this work.*
**Corresponding author: wakeley@fas.harvard.edu*

28



The effect of recurrent mutation

29



Constant Population Size

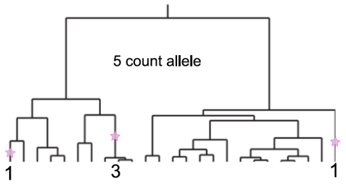
n: rare allele count k: number of latent mutations

$$p(n) \propto \frac{\theta_{(n)}}{n!}$$

$$p(k|n) \propto \frac{S_n^{(k)}(\theta)^k}{\theta_{(n)}}$$

30

More generally, we can sum over latent mutations



Desai & Plotkin., Genetics 2008

31

$$p(n) = \sum_{k=1}^{n} p(n|k)p(k)$$

- sum over recurrence

$$p(n|k) = \sum_{(i_1,\ldots,i_k)} \sum_{m=1}^{k} \frac{\mathrm{E}[\tau_i]}{\mathrm{E}[T_{total}]}$$

- sum over partitions, e.g. (n=5,k=3): 1+1+3, 2+2+1

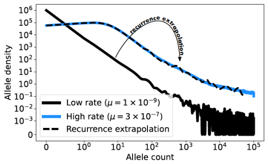
$$k \sim \mathrm{Poisson}(\theta \mathrm{E}[T_{total}])$$

- latent mutations

- $p(n)$: allele frequency distribution
- $p(k|n)$: recurrence distribution
- τ_i: total branch length with i descendants
- T_{total}: total size of the genealogy
- μ: mutation rate per generation
- θ: scaled mutation rate

32

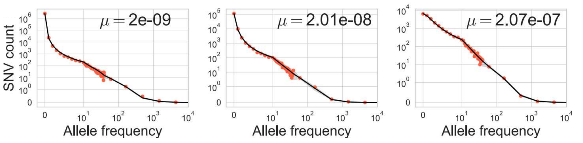
Predict SFS for high mutation rate sites from low mutation rate sites



Estimate $\mathrm{E}[\tau_i]$ by assuming no recurrent mutations at low-rate sites.

33

This works very well on real data



34

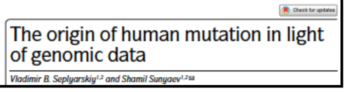
In order to measure selection, we need a good handle on mutation rate!

RESEARCH

HUMAN GENETICS

Population sequencing data reveal a compendium of mutational processes in the human germ line

Vladimir B. Seplyarskiy[12]†, Ruslan A. Soldatov[2]†, Evan Koch[12], Ryan J. McGinty[12], Jakob M. Goldmann[3], Ryan D. Hernandez[45], Kathleen Barnes[6], Adolfo Correa[789], Esteban G. Burchard[510], Patrick T. Ellinor[11], Stephen T. McGarvey[121314], Braxton D. Mitchell[151617], Ramachandran S. Vasan[1819], Susan Redline[2021], Edwin Silverman[22], Scott T. Weiss[202122], Donna K. Arnett[23], John Blangero[2425], Eric Boerwinkle[2627], Jiang He[2829], Courtney Montgomery[30], D. C. Rao[31], Jerome I. Rotter[32], Kent D. Taylor[32], Jennifer A. Brody[33], Yii-Der Ida Chen[34], Lisa de las Fuentes[3135], Chii-Min Hwu[36], Stephen S. Rich[37], Ani W. Manichaikul[37], Josyf C. Mychaleckyj[37], Nicholette D. Palmer[38], Jennifer A. Smith[3940], Sharon L. R. Kardia[40], Patricia A. Peyser[40], Lawrence F. Bielak[40], Timothy D. O'Connor[414243], Leslie S. Emery[44], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium‡, TOPMed Population Genetics Working Group, Christian Gilissen[3], Peter V. Kharchenko[2], Shamil Sunyaev[12]*



Check for updates

The origin of human mutation in light of genomic data

Vladimir B. Seplyarskiy[1,2] and Shamil Sunyaev[1,2]

35

Features of mutation rate variation

Direction of transcription and replication (DNA repair recruitment)

Regional variation associated with replication timing

Methylation rate (CpG transitions)

Enzymatic demelythation rate (CpG transversions)

Regions mutagenic in arrested oocytes

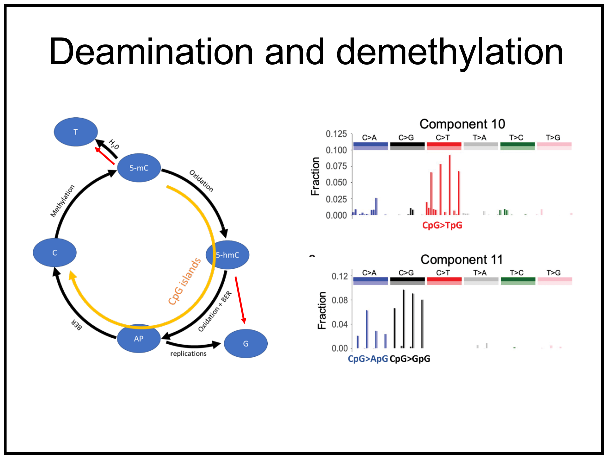
Mutagenic LINE elements

Sequence context

Transcription by RNA polymerase III

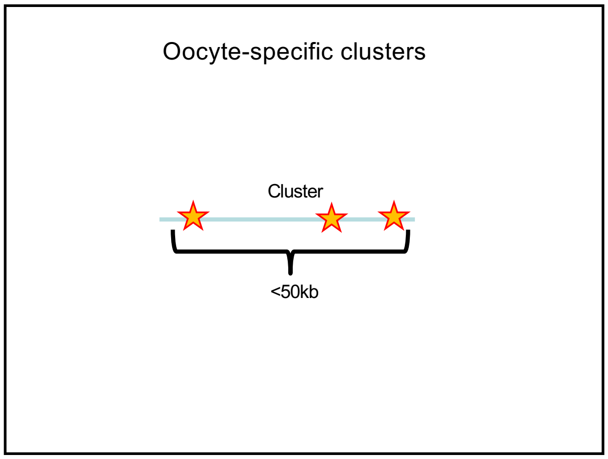
Transcription factor binding in testis

36



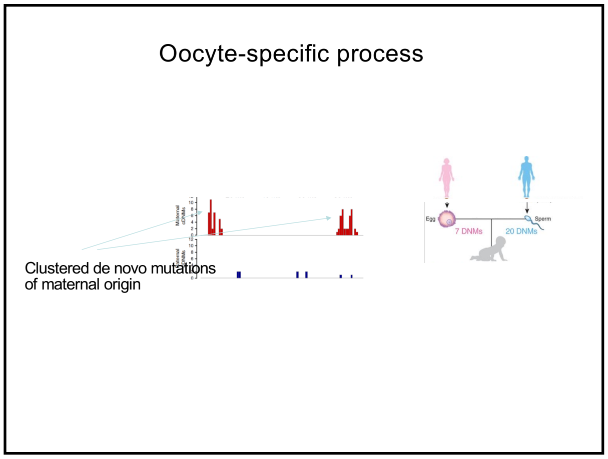
Deamination and demethylation

37



Oocyte-specific clusters

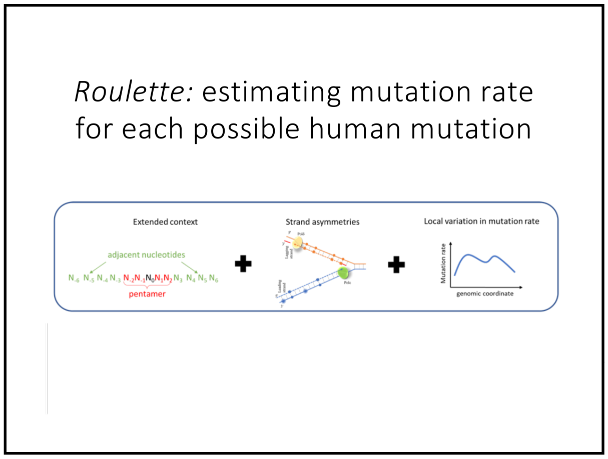
38



Oocyte-specific process

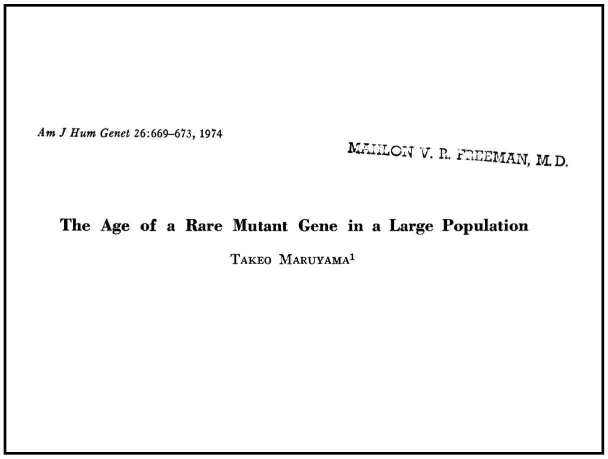
Clustered de novo mutations
of maternal origin

39



Roulette: estimating mutation rate for each possible human mutation

40



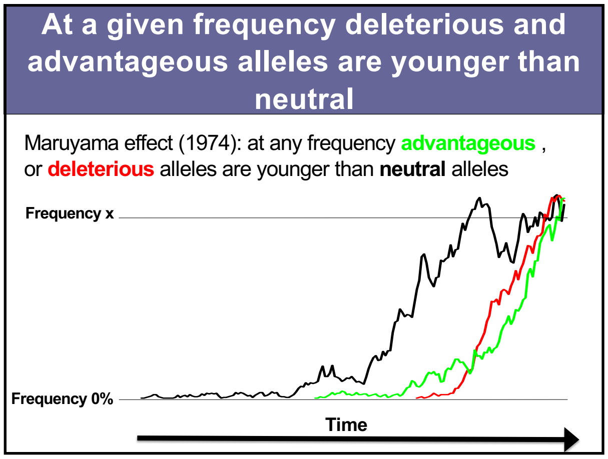
Am J Hum Genet 26:669–673, 1974

MAHLON V. R. FREEMAN, M.D.

The Age of a Rare Mutant Gene in a Large Population

TAKEO MARUYAMA[1]

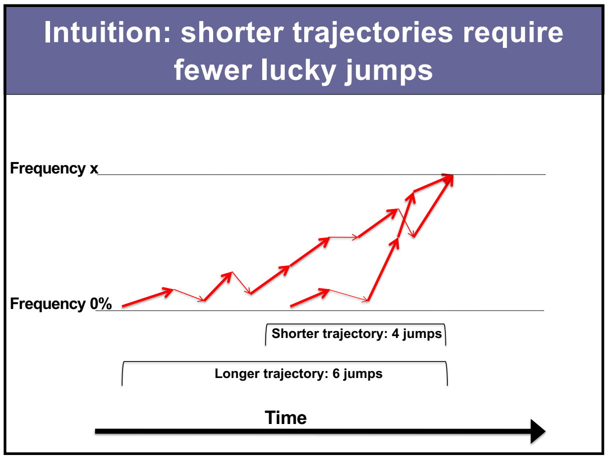
41



At a given frequency deleterious and advantageous alleles are younger than neutral

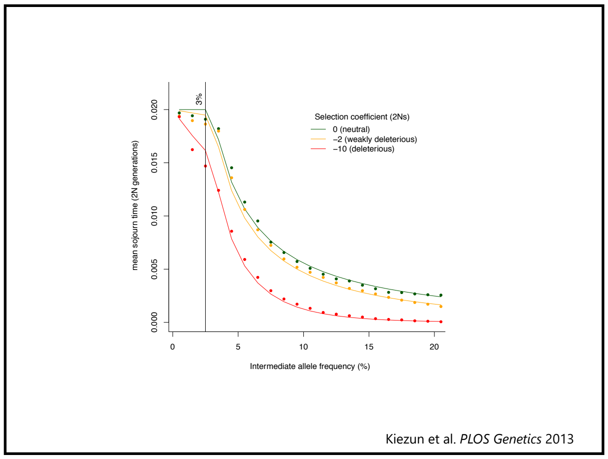
Maruyama effect (1974): at any frequency **advantageous** , or **deleterious** alleles are younger than **neutral** alleles

42



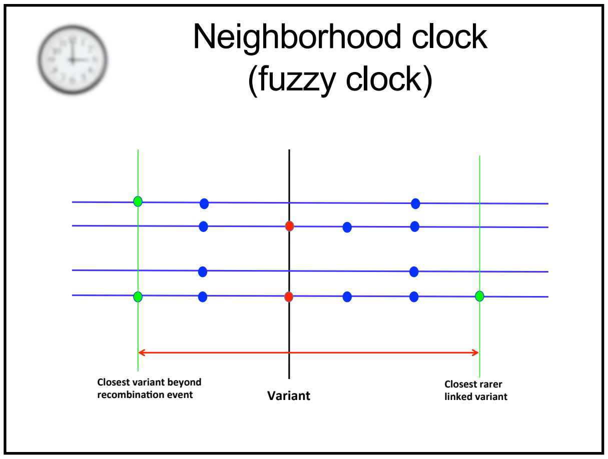
Intuition: shorter trajectories require fewer lucky jumps

43



Kiezun et al. *PLOS Genetics* 2013

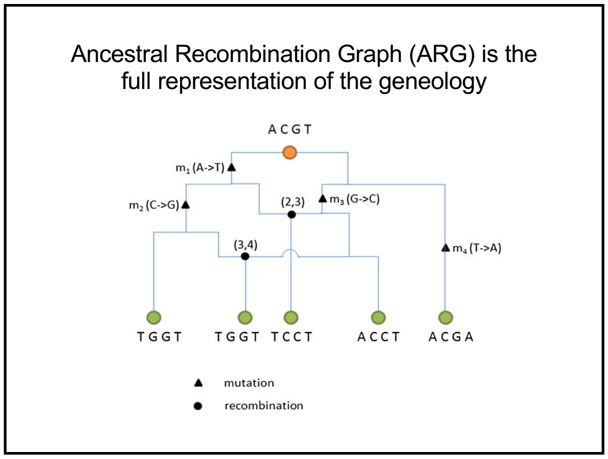
44



Neighborhood clock (fuzzy clock)

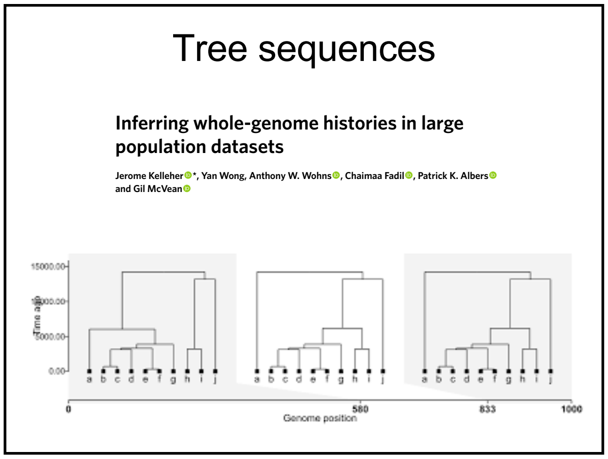
Closest variant beyond recombination event
Variant
Closest rarer linked variant

45



Ancestral Recombination Graph (ARG) is the full representation of the geneology

46

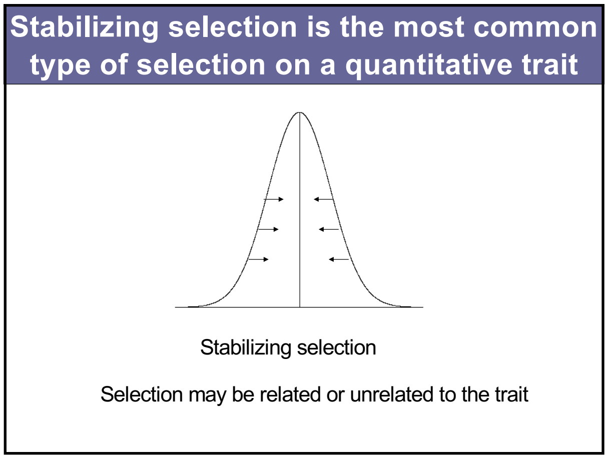


Tree sequences

Inferring whole-genome histories in large population datasets

Jerome Kelleher*, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers and Gil McVean

47



Stabilizing selection is the most common type of selection on a quantitative trait

Stabilizing selection

Selection may be related or unrelated to the trait

48

Technically, non-neutral genetic variation should not exist!

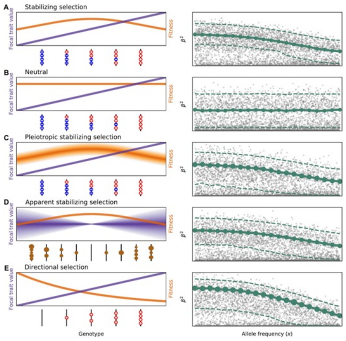
Forces to maintain variation:

Selection

Mutation

49

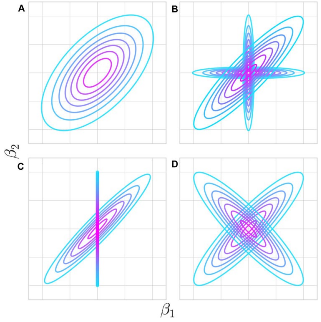
Possible theoretical models



Koch & Sunyaev *Front. Genet.* 2021

50

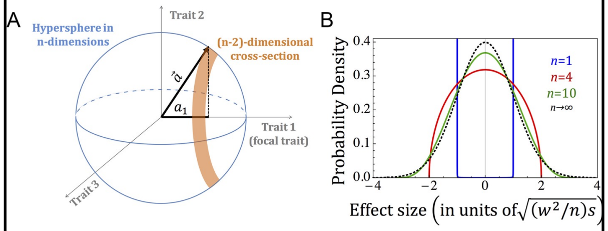
Shades of pleiotropy



Koch & Sunyaev *Front. Genet.* 2021

51

A highly pleiotropic model



Simons et al., *PLOS Biology* 2018

52

Functional annotation of genes and variants

1

Map variants onto genomic annotation

Watch for multiple transcripts!

Watch for conflicting annotations!

2

Nonsense variants

One of most significant types of variants usually leading to the complete loss of function.

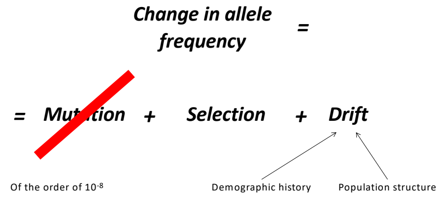
Nonsense variants are enriched in sequencing artifacts

Important considerations: i) location along the gene, ii) does the variant cause NMD? iii) is the variant in a commonly skipped exon?

Tool: LOFTEE

3

Selection inference from frequency of individual SNVs



Change in allele frequency **=**

= ***~~Mutation~~*** **+** ***Selection*** **+** ***Drift***

Of the order of 10^{-8}

Demographic history

Population structure

4

Focusing on rare deleterious PTVs

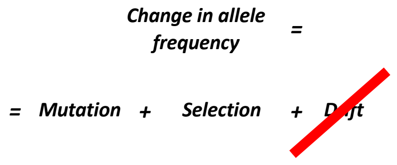
PTV – protein truncating variant (a.k.a. nonsense)

Combine all PTVs per gene – we assume that they have identical effects

Consider each gene as a bi-allelic locus – PTV / no PTV

5

Selection inference using combined frequency of PTVs

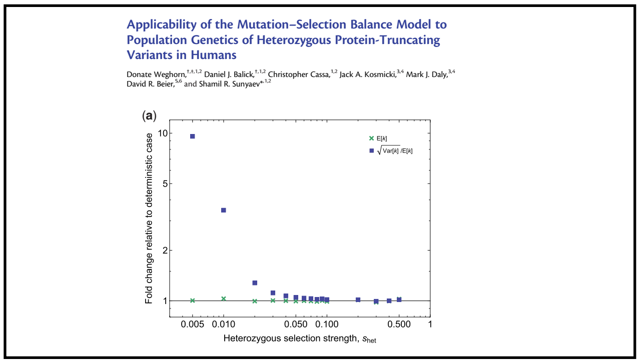


Change in allele frequency **=**

= ***Mutation*** **+** ***Selection*** **+** ***~~Drift~~***

Assuming string selection and a very large population, combined frequency of rare deleterious PTVs is expected to be Poisson distributed with $\lambda = U/hs$

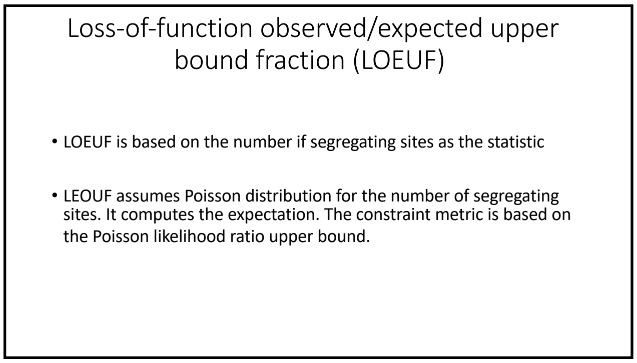
6



Applicability of the Mutation–Selection Balance Model to Population Genetics of Heterozygous Protein-Truncating Variants in Humans

Donate Weghorn,[†,‡,1,2] Daniel J. Balick,[†,1,2] Christopher Cassa,[1,2] Jack A. Kosmicki,[3,4] Mark J. Daly,[3,4] David R. Beier,[5,6] and Shamil R. Sunyaev*,[1,2]

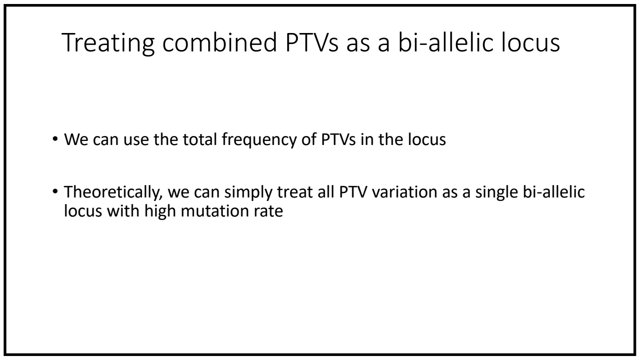
7



Loss-of-function observed/expected upper bound fraction (LOEUF)

- LOEUF is based on the number if segregating sites as the statistic
- LEOUF assumes Poisson distribution for the number of segregating sites. It computes the expectation. The constraint metric is based on the Poisson likelihood ratio upper bound.

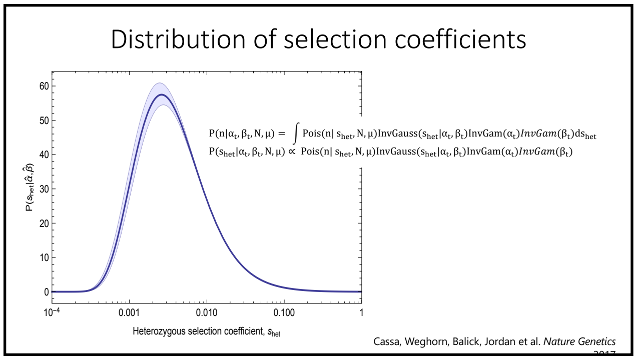
8



Treating combined PTVs as a bi-allelic locus

- We can use the total frequency of PTVs in the locus
- Theoretically, we can simply treat all PTV variation as a single bi-allelic locus with high mutation rate

9



Distribution of selection coefficients

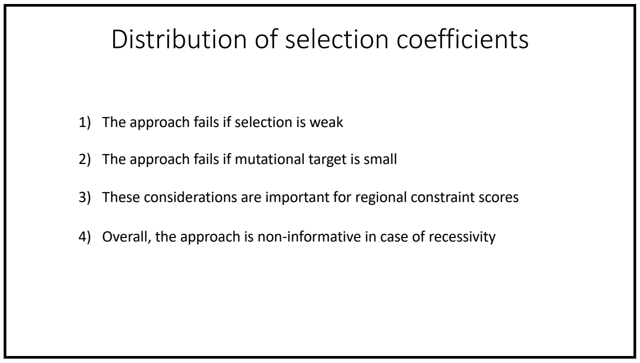
$$\mathrm{P}(\mathrm{n}|\alpha_t, \beta_t, \mathrm{N}, \mu) = \int \mathrm{Pois}(\mathrm{n}|\, s_{het}, \mathrm{N}, \mu)\mathrm{InvGauss}(s_{het}|\alpha_t, \beta_t)\mathrm{InvGam}(\alpha_t)\mathit{InvGam}(\beta_t)ds_{het}$$

$$\mathrm{P}(s_{het}|\alpha_t, \beta_t, \mathrm{N}, \mu) \propto \mathrm{Pois}(\mathrm{n}|\, s_{het}, \mathrm{N}, \mu)\mathrm{InvGauss}(s_{het}|\alpha_t, \beta_t)\mathrm{InvGam}(\alpha_t)\mathit{InvGam}(\beta_t)$$

$P(s_{het}|\hat{\alpha}, \hat{\beta})$
60
50
40
30
20
10
0
10^{-4}
0.001
0.010
0.100
1
Heterozygous selection coefficient, s_{het}

Cassa, Weghorn, Balick, Jordan et al. *Nature Genetics* 2017

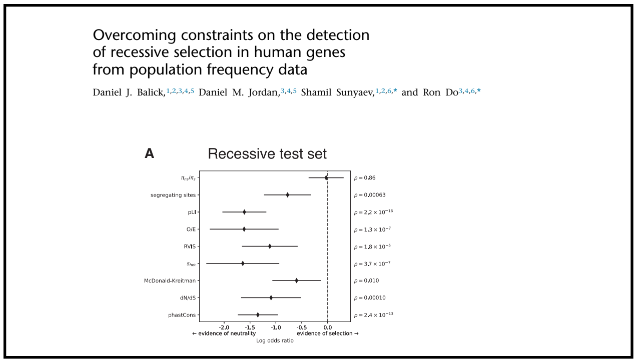
10



Distribution of selection coefficients

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

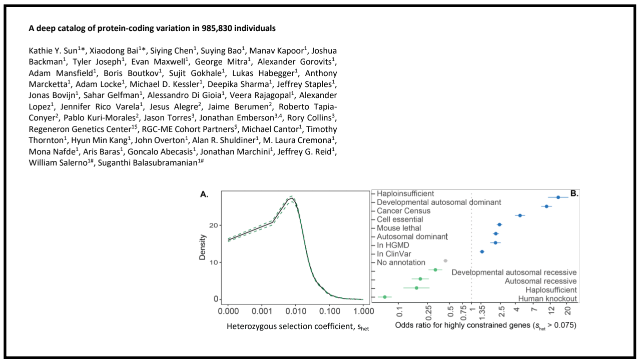
11



Overcoming constraints on the detection of recessive selection in human genes from population frequency data

Daniel J. Balick,[1,2,3,4,5] Daniel M. Jordan,[3,4,5] Shamil Sunyaev,[1,2,6,*] and Ron Do[3,4,6,*]

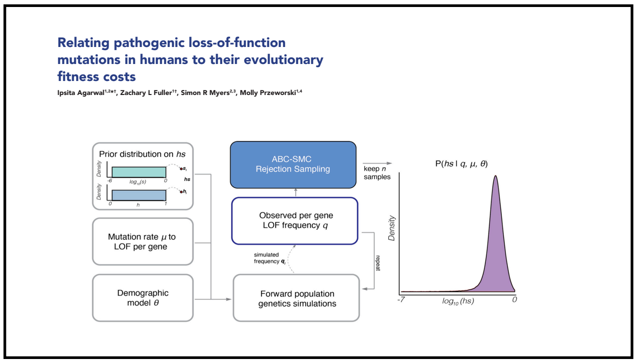
12



A deep catalog of protein-coding variation in 985,830 individuals

Kathie Y. Sun[1]*, Xiaodong Bai[1]*, Siying Chen[1], Suying Bao[1], Manav Kapoor[1], Joshua Backman[1], Tyler Joseph[1], Evan Maxwell[1], George Mitra[1], Alexander Gorovits[1], Adam Mansfield[1], Boris Boutkov[1], Sujit Gokhale[1], Lukas Habegger[1], Anthony Marcketta[1], Adam Locke[1], Michael D. Kessler[1], Deepika Sharma[1], Jeffrey Staples[1], Jonas Bovijn[1], Sahar Gelfman[1], Alessandro Di Gioia[1], Veera Rajagopal[1], Alexander Lopez[1], Jennifer Rico Varela[1], Jesus Alegre[2], Jaime Berumen[2], Roberto Tapia-Conyer[2], Pablo Kuri-Morales[2], Jason Torres[3], Jonathan Emberson[3,4], Rory Collins[3], Regeneron Genetics Center[1$], RGC-ME Cohort Partners[$], Michael Cantor[1], Timothy Thornton[1], Hyun Min Kang[1], John Overton[1], Alan R. Shuldiner[1], M. Laura Cremona[1], Mona Nafde[1], Aris Baras[1], Goncalo Abecasis[1], Jonathan Marchini[1], Jeffrey G. Reid[1], William Salerno[1#], Suganthi Balasubramanian[1#]

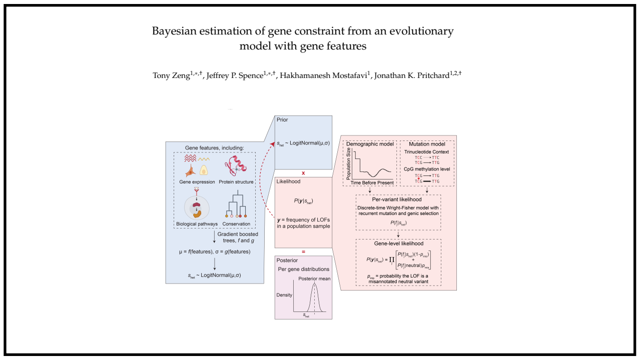
13



Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs

Ipsita Agarwal[1,2*†], Zachary L Fuller[1†], Simon R Myers[2,3], Molly Przeworski[1,4]

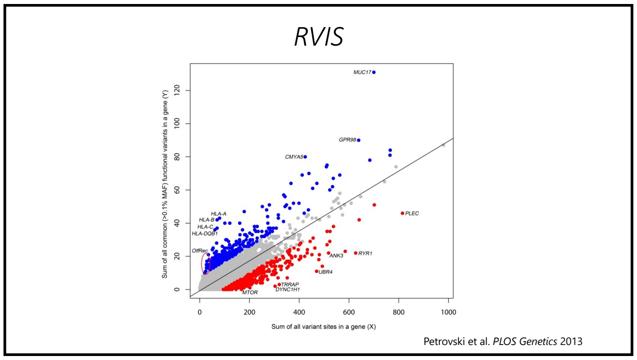
14



Bayesian estimation of gene constraint from an evolutionary model with gene features

Tony Zeng[1,*,†], Jeffrey P. Spence[1,*,†], Hakhamanesh Mostafavi[1], Jonathan K. Pritchard[1,2,†]

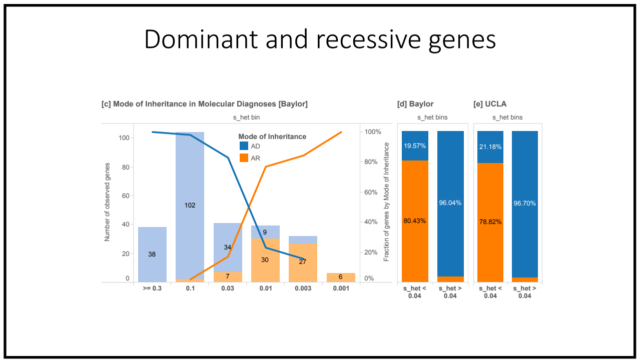
15



RVIS

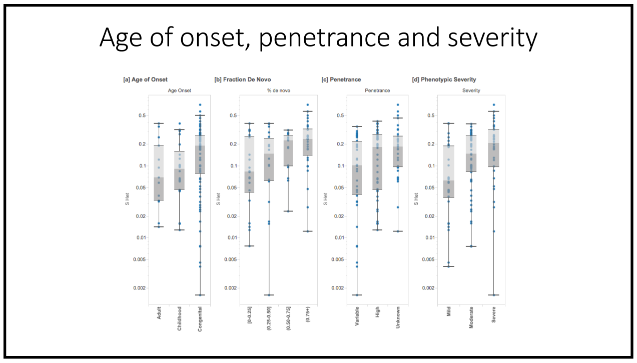
Petrovski et al. *PLOS Genetics* 2013

16



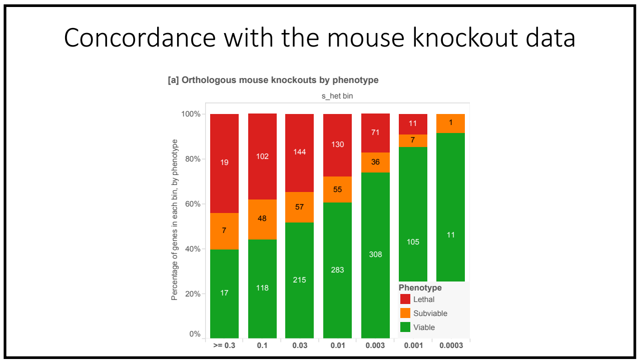
Dominant and recessive genes

17



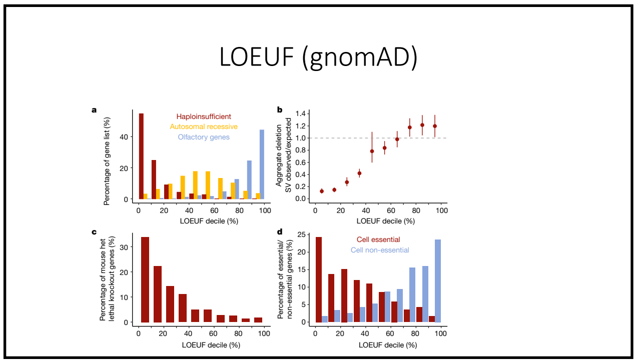
Age of onset, penetrance and severity

18



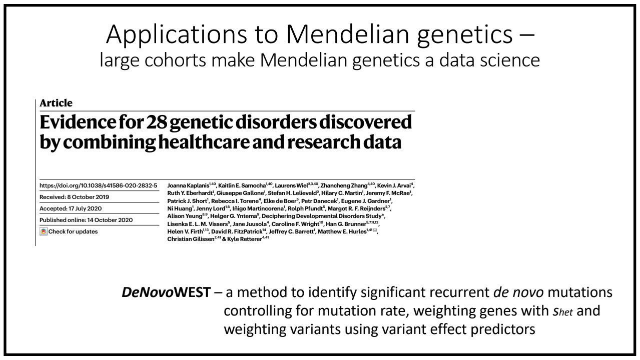
Concordance with the mouse knockout data

19



LOEUF (gnomAD)

20



Applications to Mendelian genetics –

large cohorts make Mendelian genetics a data science

Article

Evidence for 28 genetic disorders discovered by combining healthcare and research data

https://doi.org/10.1038/s41586-020-2832-5

Received: 8 October 2019

Accepted: 17 July 2020

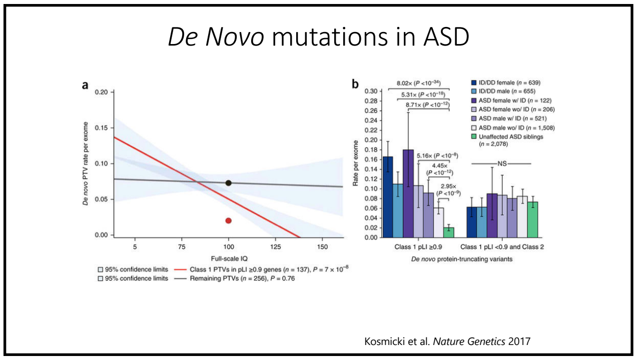
Published online: 14 October 2020

Check for updates

Joanna Kaplanis[1,40], Kaitlin E. Samocha[1,40], Laurens Wiel[2,3,40], Zhancheng Zhang[4,40], Kevin J. Arvai[4], Ruth Y. Eberhardt[1], Giuseppe Gallone[1], Stefan H. Lelieveld[2], Hilary C. Martin[1], Jeremy F. McRae[1], Patrick J. Short[1], Rebecca I. Torene[4], Elke de Boer[5], Petr Danecek[1], Eugene J. Gardner[1], Ni Huang[1], Jenny Lord[1,6], Iñigo Martincorena[1], Rolph Pfundt[5], Margot R. F. Reijnders[2,7], Alison Yeung[8,9], Helger G. Yntema[5], Deciphering Developmental Disorders Study*, Lisenka E. L. M. Vissers[5], Jane Juusola[4], Caroline F. Wright[10], Han G. Brunner[5,7,11,12], Helen V. Firth[1,13], David R. FitzPatrick[14], Jeffrey C. Barrett[1], Matthew E. Hurles[1,41] ✉, Christian Gilissen[2,41] & Kyle Retterer[4,41]

***DeNovo*WEST** – a method to identify significant recurrent *de novo* mutations controlling for mutation rate, weighting genes with s_{het} and weighting variants using variant effect predictors

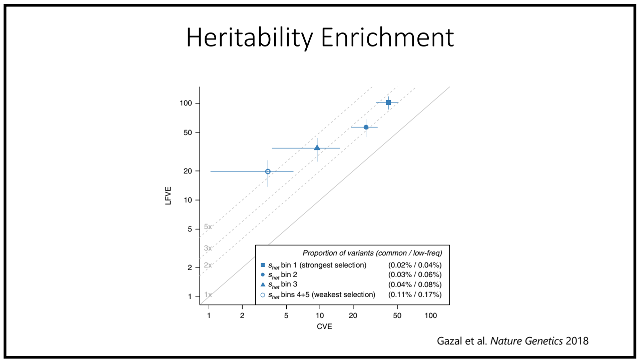
21



De Novo mutations in ASD

Kosmicki et al. *Nature Genetics* 2017

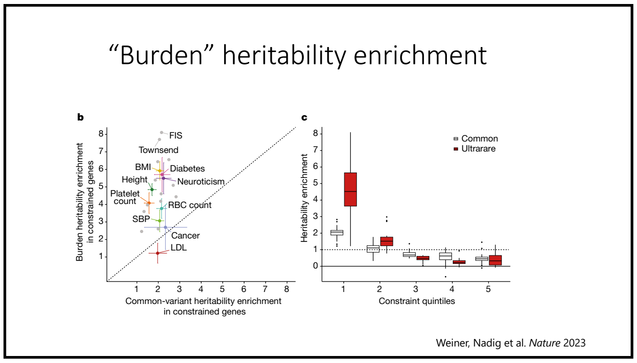
22



Heritability Enrichment

Gazal et al. *Nature Genetics* 2018

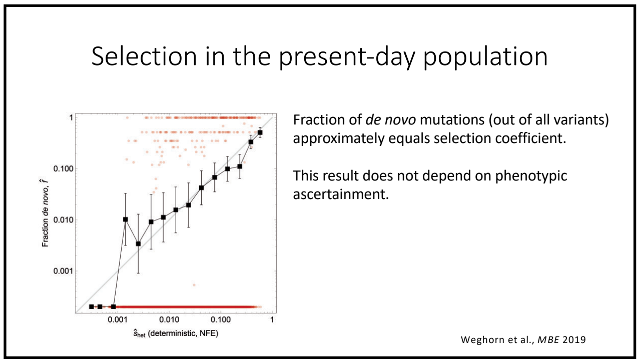
23



"Burden" heritability enrichment

Weiner, Nadig et al. *Nature* 2023

24



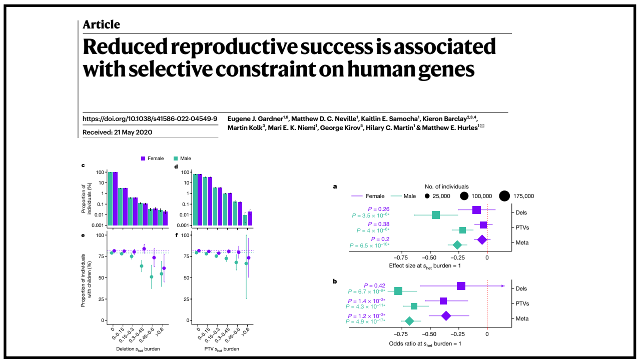
Selection in the present-day population

Fraction of *de novo* mutations (out of all variants) approximately equals selection coefficient.

This result does not depend on phenotypic ascertainment.

Weghorn et al., *MBE* 2019

25



Article

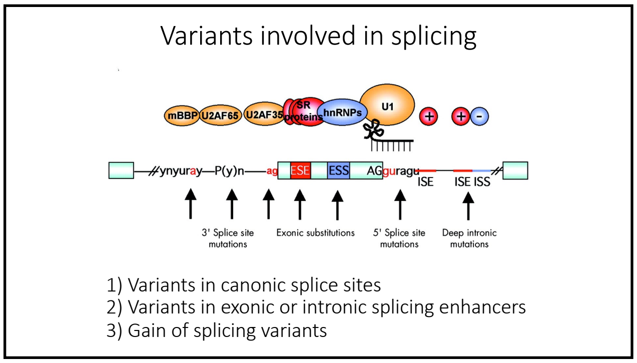
Reduced reproductive success is associated with selective constraint on human genes

https://doi.org/10.1038/s41586-022-04549-9

Received: 21 May 2020

Eugene J. Gardner[1,6], Matthew D. C. Neville[1], Kaitlin E. Samocha[1], Kieron Barclay[2,3,4], Martin Kolk[3], Mari E. K. Niemi[1], George Kirov[5], Hilary C. Martin[1] & Matthew E. Hurles[1]✉

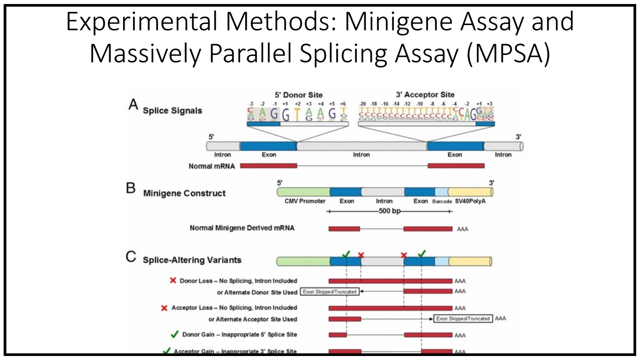
1



Variants involved in splicing

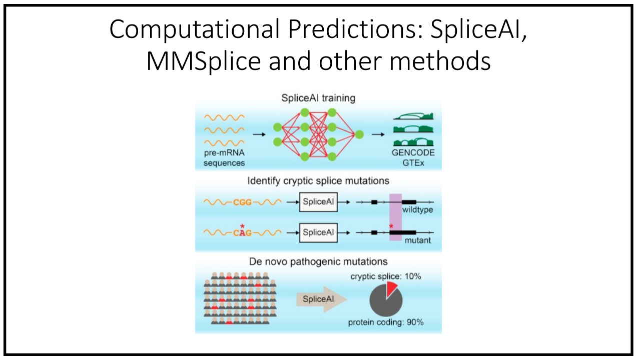
1) Variants in canonic splice sites
2) Variants in exonic or intronic splicing enhancers
3) Gain of splicing variants

2



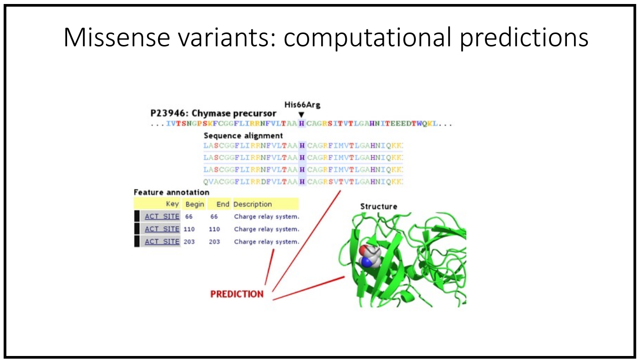
Experimental Methods: Minigene Assay and Massively Parallel Splicing Assay (MPSA)

3



Computational Predictions: SpliceAI, MMSplice and other methods

4



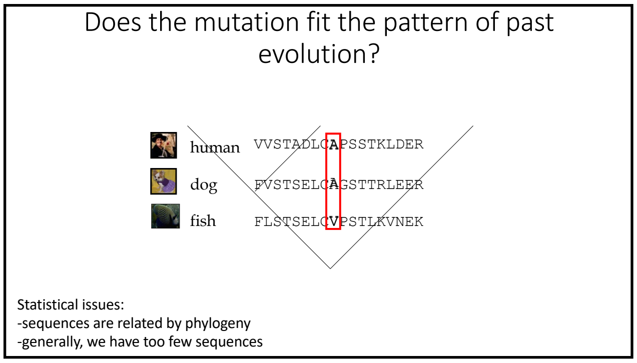
Missense variants: computational predictions

His66Arg
P23946: Chymase precursor
...IVTSNGPSKFCGGFLIRRNFVLTAAHCAGRSITVTLGAHNITEEEDTWQKL...
Sequence alignment
LASCGGFLIRRNFVLTAAHCAGRFIMVTLGAHNIQKK
LASCGGFLIRRNFVLTAAHCAGRFIMVTLGAHNIQKK
LASCGGFLIRRNFVLTAAHCAGRFIMVTLGAHNIQKK
QVACGGFLIRRDFVLTAAHCAGRSVTVTLGAHNIQKK
Structure
PREDICTION

Feature annotation

Key	Begin	End	Description
ACT_SITE	66	66	Charge relay system.
ACT_SITE	110	110	Charge relay system.
ACT_SITE	203	203	Charge relay system.

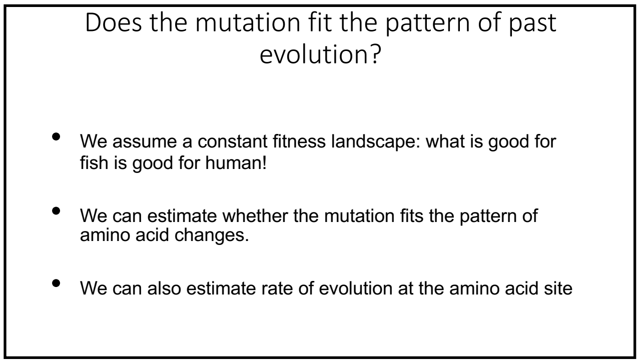
5



Does the mutation fit the pattern of past evolution?

Statistical issues:
-sequences are related by phylogeny
-generally, we have too few sequences

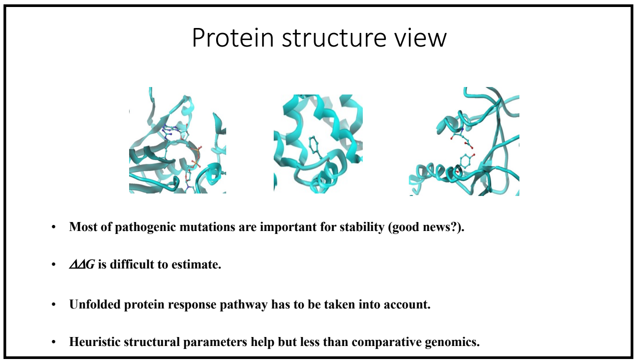
6



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

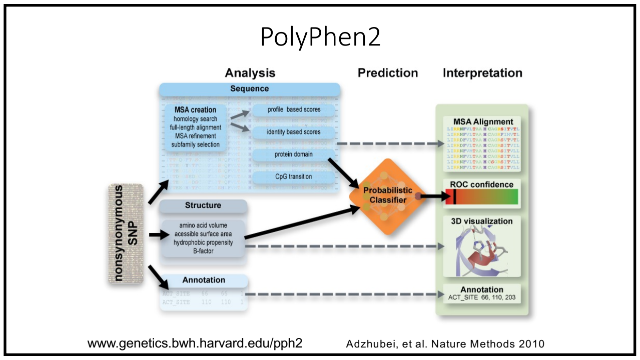
7



Protein structure view

- **Most of pathogenic mutations are important for stability (good news?).**
- ***ΔΔG* is difficult to estimate.**
- **Unfolded protein response pathway has to be taken into account.**
- **Heuristic structural parameters help but less than comparative genomics.**

8



PolyPhen2

Analysis

Prediction

Interpretation

Sequence

MSA creation
homology search
full-length alignment
MSA refinement
subfamily selection

profile based scores

identity based scores

protein domain

CpG transition

nonsynonymous SNP

Structure

amino acid volume
acessible surface area
hydrophobic propensity
B-factor

Annotation

ACT_SITE 66 66 1
ACT_SITE 110 110 1

Probabilistic Classifier

MSA Alignment

LIRRNFVLTAA H CAGRSITVTL
LIRRNFVLTAA H CAGRFIMVTL
LIRRDFVLTAA H CAGRSVTVTL
LIRRDFVLTAA H CAGRSITVIL
LIRRNFVMTAA H CAGRSITVLL
LIRRNFVLTAA H CAGRSITVLL
LIRRNFVLTAA H CSGRSITVLL
LIRRNFVLTAA H CAGRSITVLL

ROC confidence

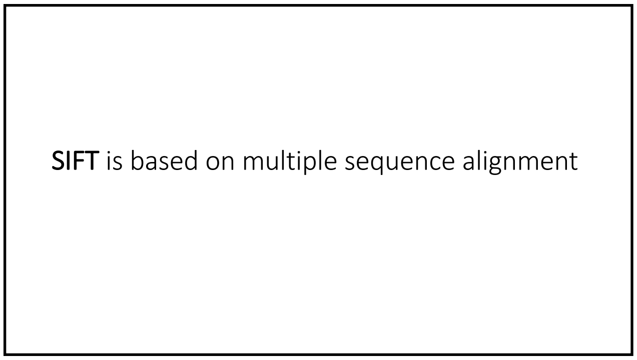
3D visualization

Annotation
ACT_SITE 66, 110, 203

www.genetics.bwh.harvard.edu/pph2

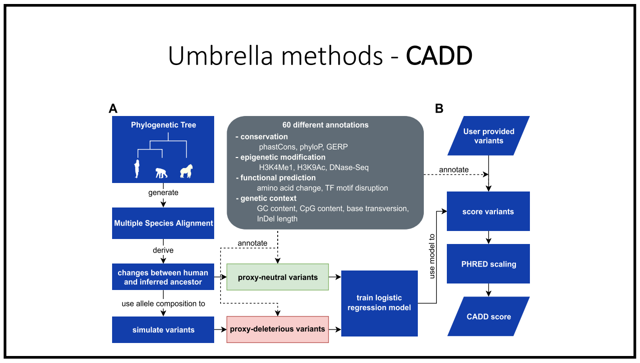
Adzhubei, et al. Nature Methods 2010

9



SIFT is based on multiple sequence alignment

10



Umbrella methods - **CADD**

A

Phylogenetic Tree

generate

Multiple Species Alignment

derive

changes between human and inferred ancestor

use allele composition to

simulate variants

60 different annotations
- conservation
  phastCons, phyloP, GERP
- epigenetic modification
  H3K4Me1, H3K9Ac, DNase-Seq
- functional prediction
  amino acid change, TF motif disruption
- genetic context
  GC content, CpG content, base transversion, InDel length

annotate

proxy-neutral variants

proxy-deleterious variants

train logistic regression model

use model to

B

User provided variants

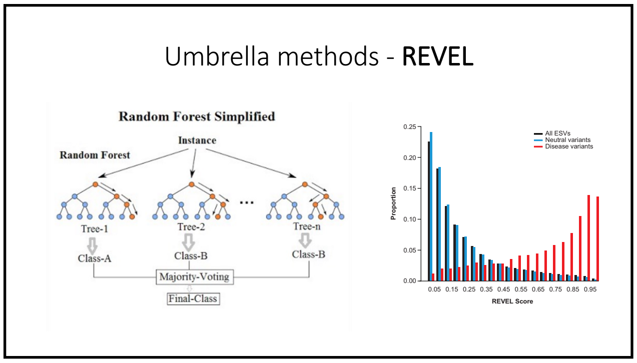
annotate

score variants

PHRED scaling

CADD score

11



Umbrella methods - **REVEL**

Random Forest Simplified

Instance

Random Forest

Tree-1

Tree-2

...

Tree-n

Class-A

Class-B

Class-B

Majority-Voting

Final-Class

12

Umbrella methods

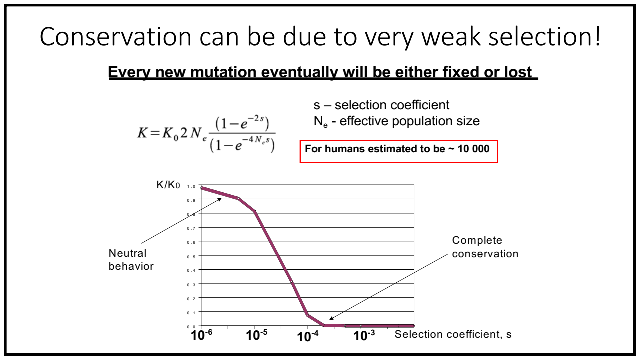
- **VEST4** – also an umbrella method using Random Forest
- **VARITY** – a new method using Gradient Boosting and focusing on de novo mutations and ultra rare variants

13

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

14



Conservation can be due to very weak selection!

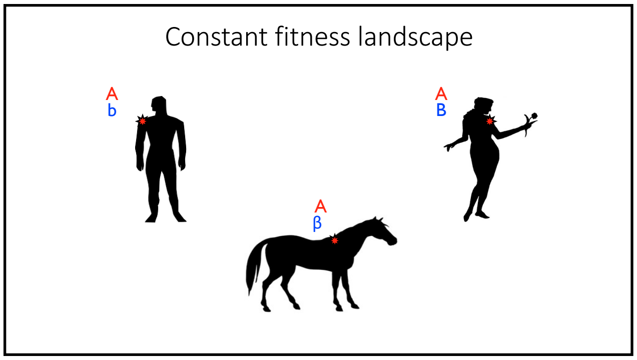
Every new mutation eventually will be either fixed or lost

$$K = K_0 2 N_e \frac{(1-e^{-2s})}{(1-e^{-4N_e s})}$$

s – selection coefficient
N_e - effective population size

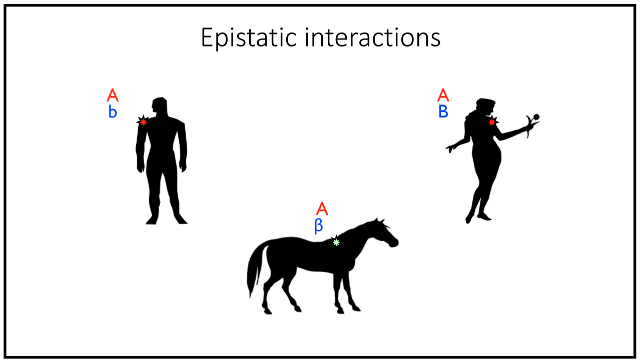
For humans estimated to be ~ 10 000

15



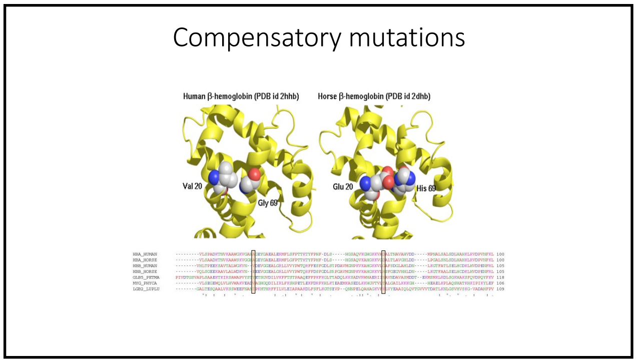
Constant fitness landscape

16



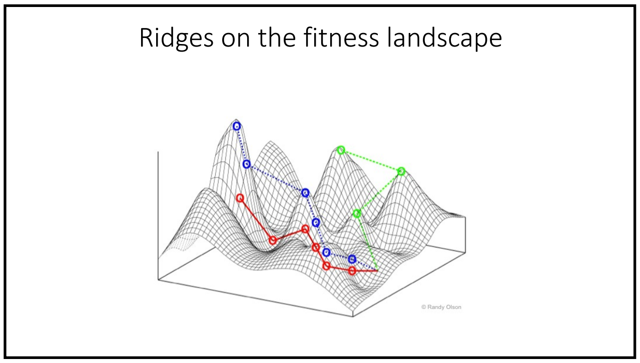
Epistatic interactions

17



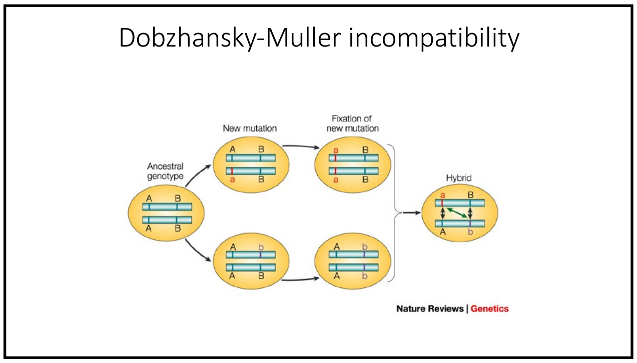
Compensatory mutations

18



Ridges on the fitness landscape

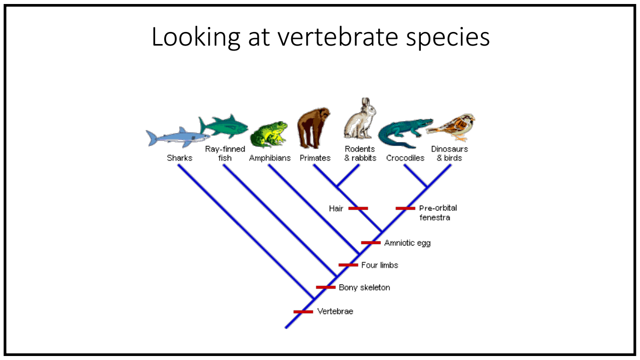
19



Dobzhansky-Muller incompatibility

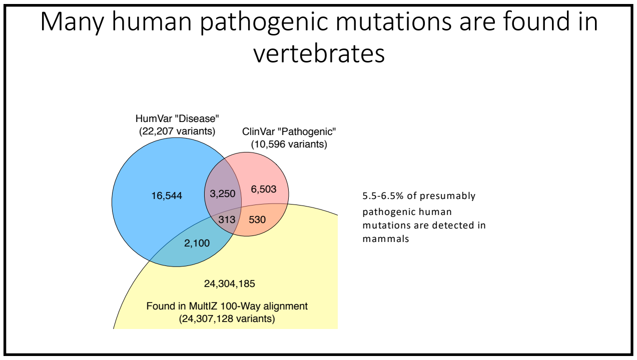
Nature Reviews | Genetics

20



Looking at vertebrate species

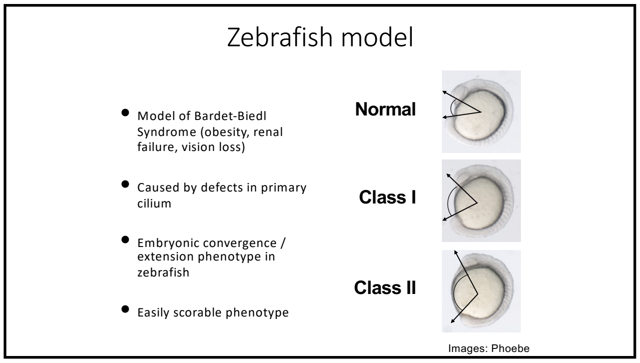
21



Many human pathogenic mutations are found in vertebrates

5.5-6.5% of presumably pathogenic human mutations are detected in mammals

22

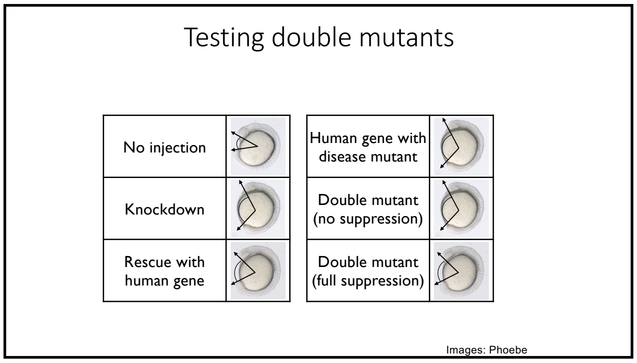


Zebrafish model

- 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

23

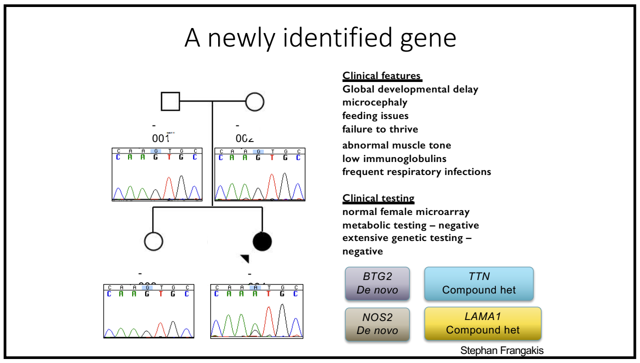


Testing double mutants

No injection		Human gene with disease mutant	
Knockdown		Double mutant (no suppression)	
Rescue with human gene		Double mutant (full suppression)	

Images: Phoebe

24



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

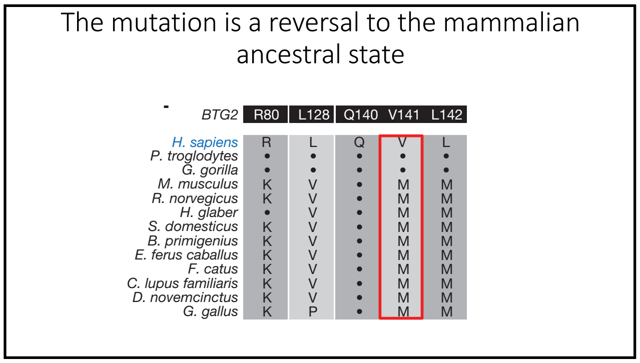
TTN
Compound het

NOS2
De novo

LAMA1
Compound het

Stephan Frangakis

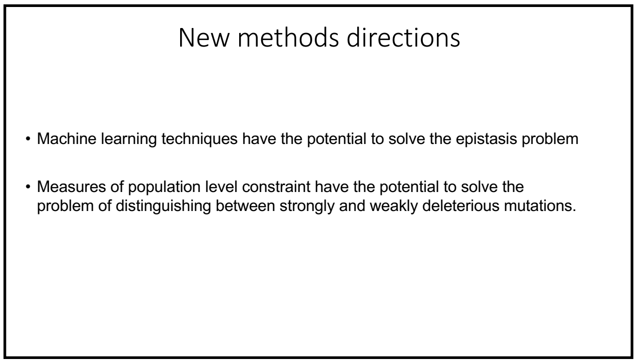
25



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

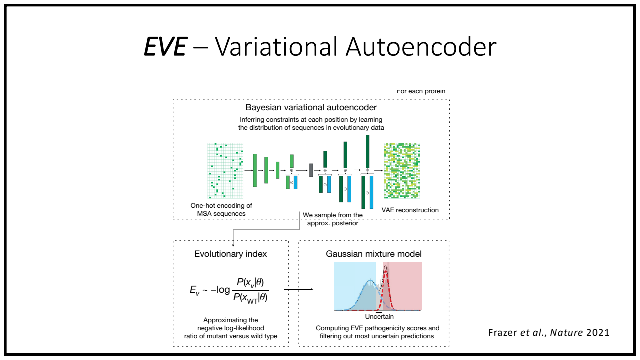
1



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.

2

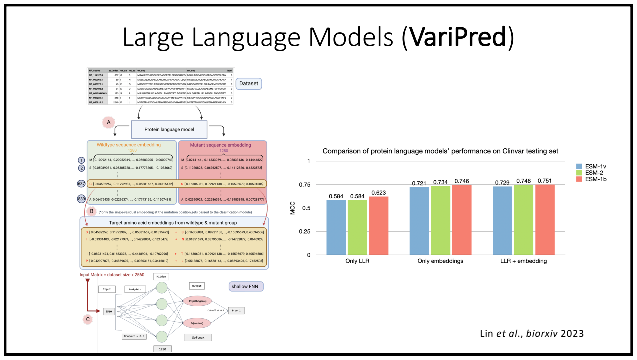


EVE – Variational Autoencoder

For each protein
Bayesian variational autoencoder
Inferring constraints at each position by learning the distribution of sequences in evolutionary data
One-hot encoding of MSA sequences
We sample from the approx. posterior
VAE reconstruction
Evolutionary index
$E_v \sim -\log \frac{P(x_v|\theta)}{P(x_{WT}|\theta)}$
Approximating the negative log-likelihood ratio of mutant versus wild type
Gaussian mixture model
Uncertain
Computing EVE pathogenicity scores and filtering out most uncertain predictions

Frazer *et al.*, *Nature* 2021

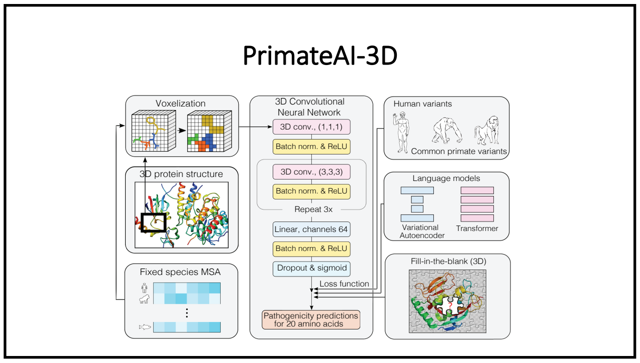
3



Large Language Models (**VariPred**)

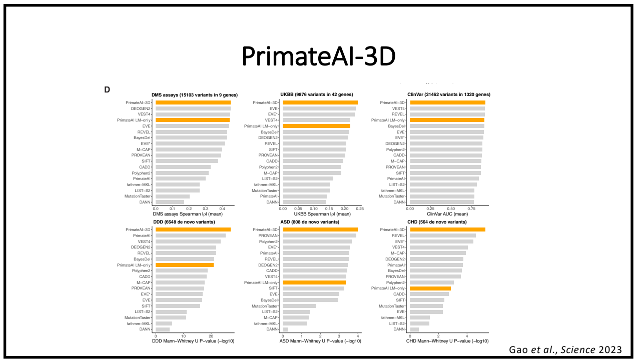
Lin *et al.*, *biorxiv* 2023

4



PrimateAI-3D

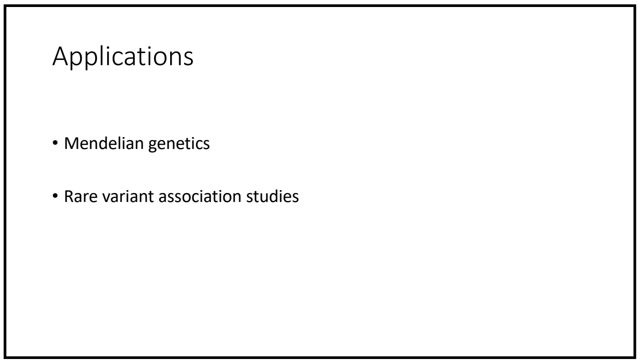
5



PrimateAI-3D

Gao *et al.*, *Science* 2023

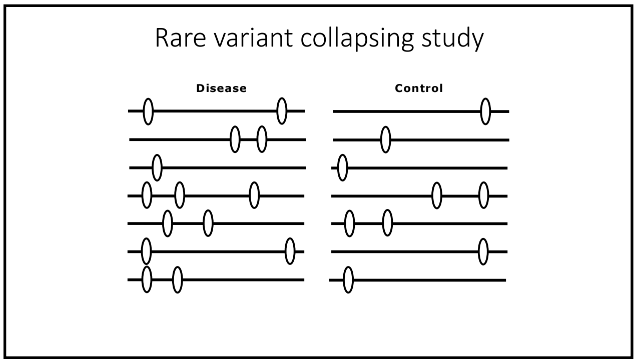
6



Applications

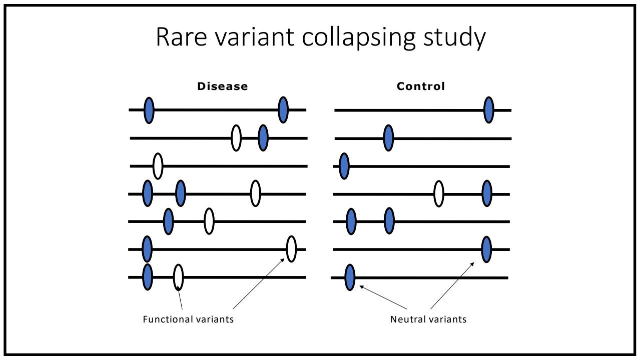
- Mendelian genetics
- Rare variant association studies

7



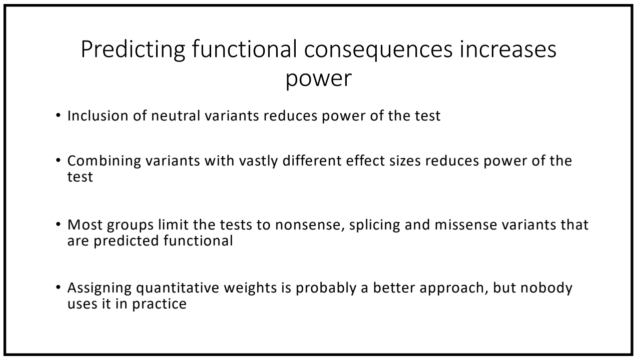
Rare variant collapsing study

8



Rare variant collapsing study

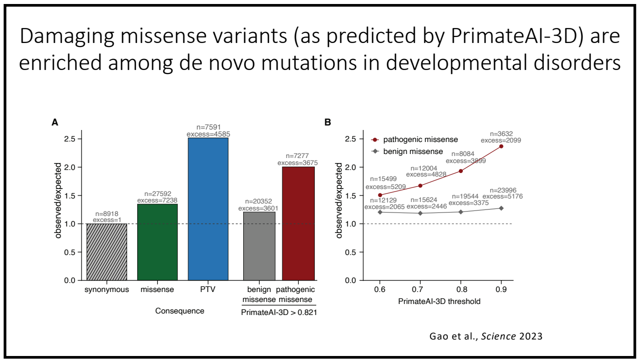
9



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

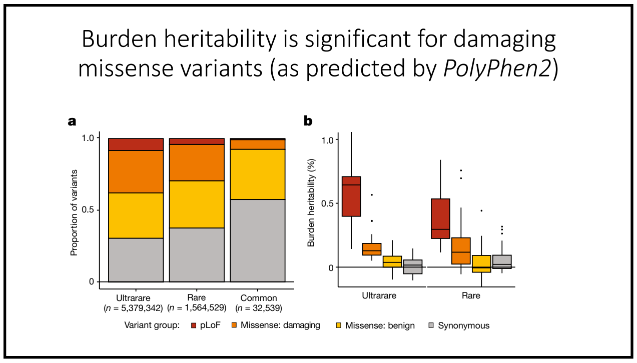
10



Damaging missense variants (as predicted by PrimateAI-3D) are enriched among de novo mutations in developmental disorders

Gao et al., *Science* 2023

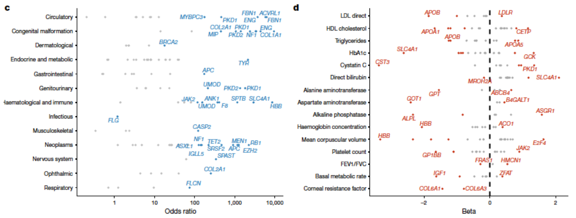
11



Burden heritability is significant for damaging missense variants (as predicted by *PolyPhen2*)

12

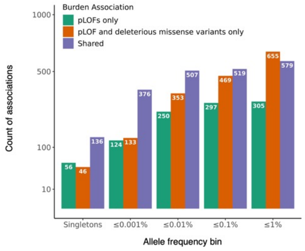
UK Biobank results (Wang et al.)



Variant grouping: nonsense, splicing, missense predicted by REVEL and MTR

13

UK Biobank results (Backman et al.)

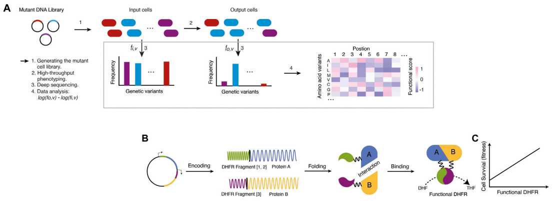


Deleterious missense variants:

SIFT
PolyPhen2
LRT
Mutation Taster

14

Experimental technologies – deep mutational scanning (DMS)

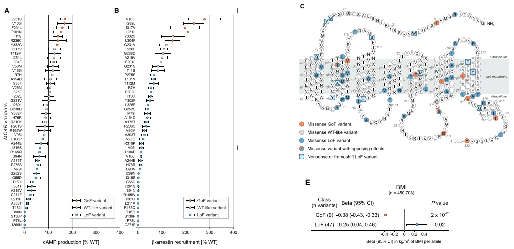


→ 1. Generating the mutant cell library.
2. High-throughput phenotyping.
3. Deep sequencing.
4. Data analysis: *log(fo,v) - log(fi,v)*

Wei & Li, *Frontiers in Genetics* 2023

15

MC4R example



Lotta *et al.*, *Cell* 2023

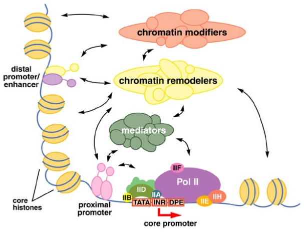
1

Non-coding variants

2

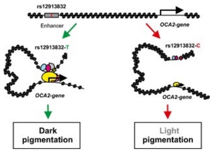
Regulatory variants

- Regulation: variants in promoters, enhancers, silencers, insulators



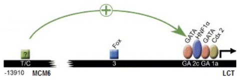
3

Non-disease alleles of large effect



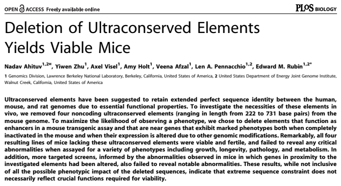
Eye color

Lactase tolerance



4

Ultraconserved elements



OPEN ACCESS Freely available online PLOS BIOLOGY

Deletion of Ultraconserved Elements Yields Viable Mice

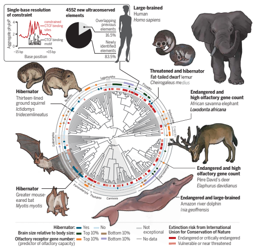
Nadav Ahituv[1,2¤], Yiwen Zhu[1], Axel Visel[1], Amy Holt[1], Veena Afzal[1], Len A. Pennacchio[1,2], Edward M. Rubin[1,2*]

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.

5

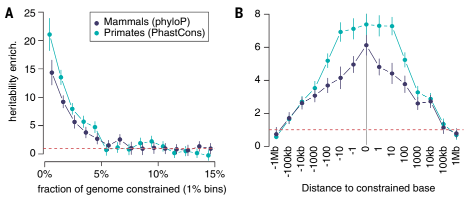
Zoonomia conservation


Single-base resolution of constraint
Aggregate phyloP
4
2
0
−15 bp
+15 bp
Base position
constrained CTCF binding sites
CTCF binding motif
4552 new ultraconserved elements
Overlapping previous elements
16.5%
Newly identified elements
83.5%
Large-brained
Human
Homo sapiens
Threatened and hibernator
Fat-tailed dwarf lemur
Cheirogaleus medius
Endangered and high olfactory gene count
African savanna elephant
Loxodonta africana
Hibernator
Thirteen-lined ground squirrel
Ictidomys tridecemlineatus
Endangered and high olfactory gene count
Père David's deer
Elaphurus davidianus
Endangered and large-brained
Amazon river dolphin
Inia geoffrensis
Hibernator
Greater mouse-eared bat
Myotis myotis
Lagomorpha
Primates
Rodentia
Eulipotyphla
Chiroptera
Perissodactyla
Pholidota
Carnivora
Cetartiodactyla
Afrotheria
Xenarthra
Hibernator: Yes No
Brain size relative to body size: Top 10% Bottom 10%
Olfactory receptor gene number: Top 10% Bottom 10%
(predictor of olfactory capacity)
Not exceptional
No data
Extinction risk from International Union for Conservation of Nature
Endangered or critically endangered
Vulnerable or near threatened

Christmas, Kaplow *et al.*, *Science* 2023

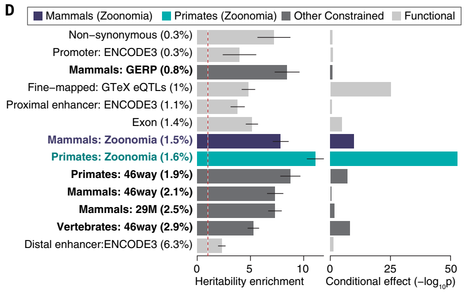
6

Heritability enrichment



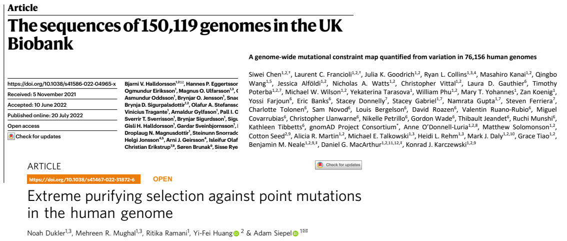
Sullivan, Meadows *et al.*, *Science* 2023

7



8

Population constraint in non-coding regions



Article

The sequences of 150,119 genomes in the UK Biobank

https://doi.org/10.1038/s41586-022-04965-x

Received: 5 November 2021

Accepted: 10 June 2022

Published online: 20 July 2022

Open access

Check for updates

Bjarni V. Halldorsson[1,2✉], Hannes P. Eggertssor
Ogmundur Eiriksson[1], Magnus O. Ulfarsson[1,3], (
Asmundur Oddsson[1], Brynjar O. Jensson[1], Snae
Brynja D. Sigurpalsdottir[1,2], Olafur A. Stefanssc
Vinicius Tragante[1], Arnaldur Gylfason[1], Pall I. O
Sverrir T. Sverrisson[1], Brynjar Sigurdsson[1], Sigu
Gisli H. Halldorsson[1], Gardar Sveinbjornsson[1], I
Droplaug N. Magnusdottir[1], Steinunn Snorradc
Helgi Jonsson[4,5], Arni J. Geirsson[4], Isleifur Olaf
Christian Erikstrup[7,8], Søren Brunak[9], Sisse Rye

A genome-wide mutational constraint map quantified from variation in 76,156 human genomes

Siwei Chen[1,2,†], Laurent C. Francioli[1,2,†], Julia K. Goodrich[1,2], Ryan L. Collins[1,3,4], Masahiro Kanai[1,2], Qingbo Wang[1,5], Jessica Alföldi[1,2], Nicholas A. Watts[1,2], Christopher Vittal[1,2], Laura D. Gauthier[6], Timothy Poterba[1,2,7], Michael W. Wilson[1,2], Yekaterina Tarasova[1], William Phu[1,2], Mary T. Yohannes[1], Zan Koenig[1], Yossi Farjoun[6], Eric Banks[6], Stacey Donnelly[7], Stacey Gabriel[1,7], Namrata Gupta[1,7], Steven Ferriera[7], Charlotte Tolonen[6], Sam Novod[6], Louis Bergelson[6], David Roazen[6], Valentin Ruano-Rubio[6], Miguel Covarrubias[6], Christopher Llanwarne[6], Nikelle Petrillo[6], Gordon Wade[6], Thibault Jeandet[6], Ruchi Munshi[6], Kathleen Tibbetts[6], gnomAD Project Consortium[*], Anne O'Donnell-Luria[1,2,8], Matthew Solomonson[1,2], Cotton Seed[2,9], Alicia R. Martin[1,2], Michael E. Talkowski[1,3], Heidi L. Rehm[1,3], Mark J. Daly[1,2,10], Grace Tiao[1,2], Benjamin M. Neale[1,2,9,‡], Daniel G. MacArthur[1,2,11,12,‡], Konrad J. Karczewski[1,2,9]

Check for updates

ARTICLE

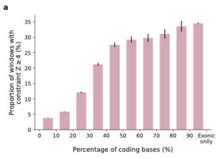
https://doi.org/10.1038/s41467-022-31872-6 OPEN

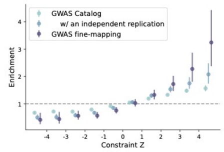
Extreme purifying selection against point mutations in the human genome

Noah Dukler[1,3], Mehreen R. Mughal[1,3], Ritika Ramani[1], Yi-Fei Huang[2] & Adam Siepel[1✉]

9

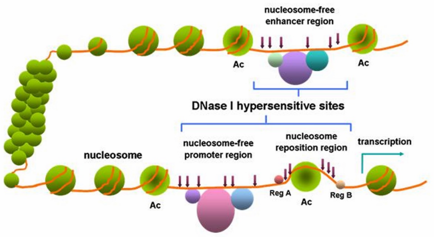
Population constraint in non-coding regions





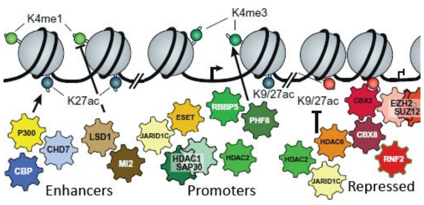
10

Chromatin accessibility


nucleosome-free enhancer region
Ac
Ac
DNase I hypersensitive sites
nucleosome
nucleosome-free promoter region
nucleosome reposition region
transcription
Reg A
Reg B
Ac
Ac

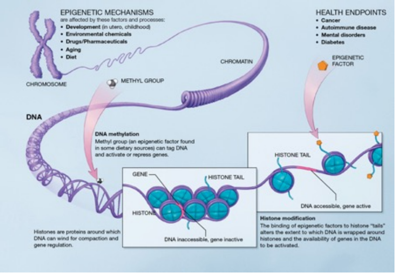
11

Chromatin modifications



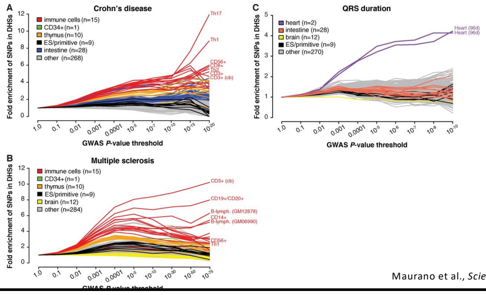
12

Epigenomics


EPIGENETIC MECHANISMS
are affected by these factors and processes:
• Development (in utero, childhood)
• Environmental chemicals
• Drugs/Pharmaceuticals
• Aging
• Diet
HEALTH ENDPOINTS
• Cancer
• Autoimmune disease
• Mental disorders
• Diabetes
CHROMOSOME
CHROMATIN
METHYL GROUP
EPIGENETIC FACTOR
DNA
DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.
GENE
HISTONE TAIL
HISTONE
HISTONE TAIL
DNA accessible, gene active
DNA inaccessible, gene inactive
Histones are proteins around which DNA can wind for compaction and gene regulation.
Histone modification
The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

13

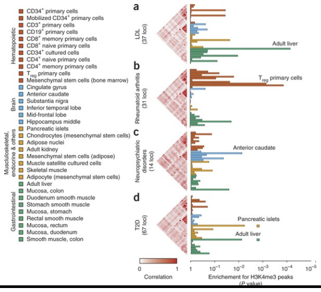
Enrichment of GWAS signals in regulatory elements



Maurano et al., *Science*, 2012

14

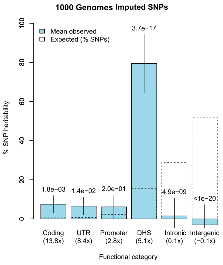
Enrichment of GWAS signals in regulatory elements



Trynka et al., *Nature Genetics*, 2014

15

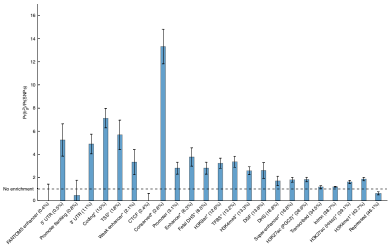
Partitioning heritability



Gusev & Price, *AJHG*, 2014

16

Heritability partitioning across annotations



Finucane et al., *Nature Genetics*, 2015

17

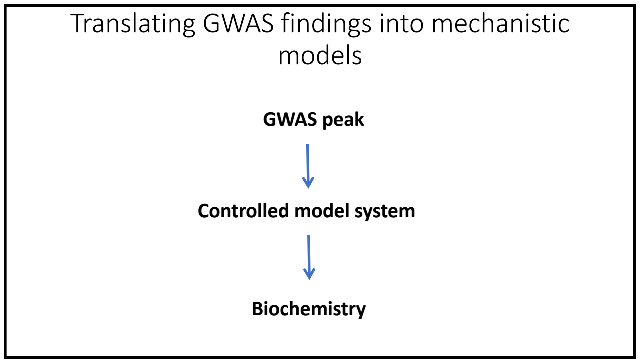
Application – function informed fine-mapping

Functionally informed fine-mapping and polygenic localization of complex trait heritability

Omer Weissbrod[1], Farhad Hormozdiari[1], Christian Benner[2], Ran Cui[3], Jacob Ulirsch[3,4], Steven Gazal[1], Armin P. Schoech[1], Bryce van de Geijn[1], Yakir Reshef[1], Carla Márquez-Luna[5], Luke O'Connor[3], Matti Pirinen[2,6,7], Hilary K. Finucane[3,8] and Alkes L. Price[1,3]

- Estimate heritability enrichment and convert the estimates into prior probabilities
- Use these prior in fine-mapping (with SuSiE or FINEMAP)

18



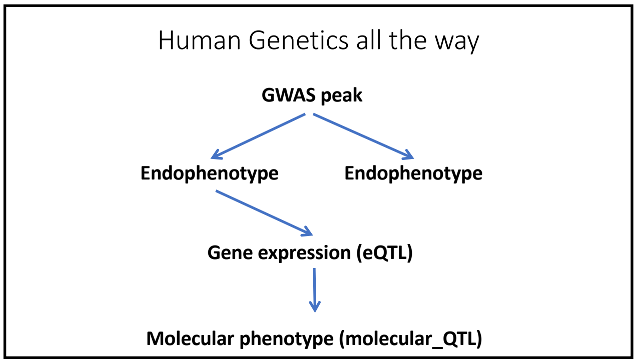
Translating GWAS findings into mechanistic models

GWAS peak

Controlled model system

Biochemistry

19



Human Genetics all the way

GWAS peak

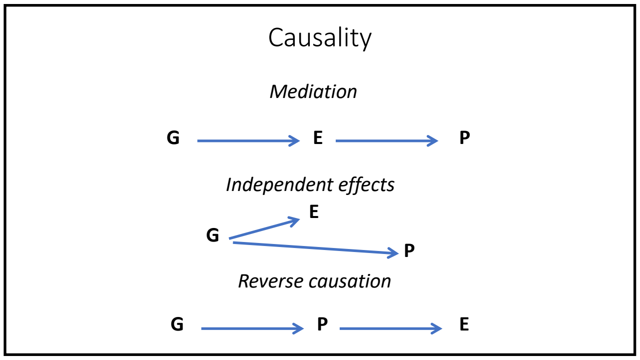
Endophenotype

Endophenotype

Gene expression (eQTL)

Molecular phenotype (molecular_QTL)

20



Causality

Mediation

$$\begin{tikzcd} \mathbf{G} \arrow[r] & \mathbf{E} \arrow[r] & \mathbf{P} \end{tikzcd}$$

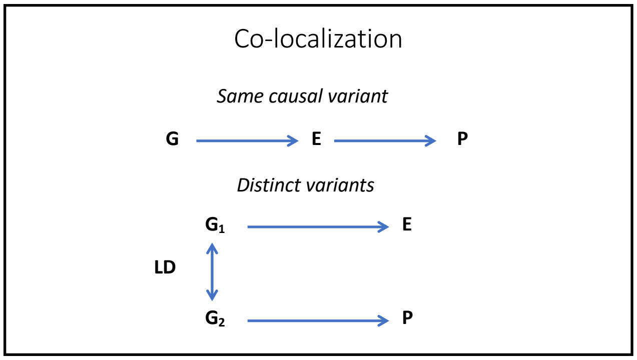
Independent effects

$$\begin{tikzcd} & \mathbf{E} \\ \mathbf{G} \arrow[ur] \arrow[dr] & \\ & \mathbf{P} \end{tikzcd}$$

Reverse causation

$$\begin{tikzcd} \mathbf{G} \arrow[r] & \mathbf{P} \arrow[r] & \mathbf{E} \end{tikzcd}$$

21



Co-localization

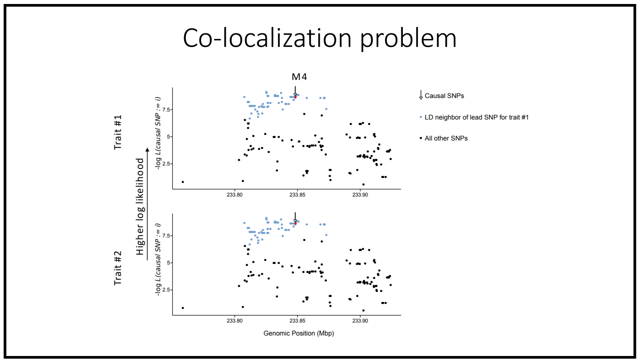
Same causal variant

$$\begin{tikzcd} \mathbf{G} \arrow[r] & \mathbf{E} \arrow[r] & \mathbf{P} \end{tikzcd}$$

Distinct variants

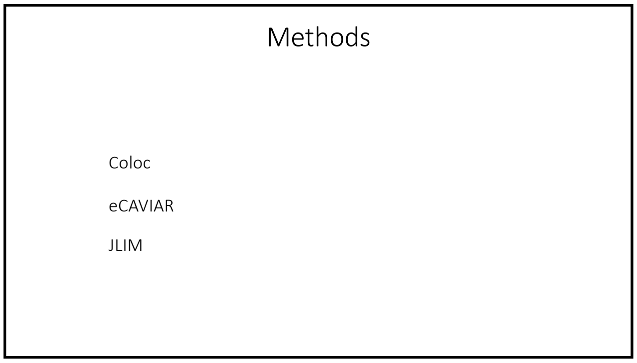
$$\begin{tikzcd} \mathbf{G_1} \arrow[r] \arrow[d, leftrightarrow, "\mathbf{LD}"'] & \mathbf{E} \\ \mathbf{G_2} \arrow[r] & \mathbf{P} \end{tikzcd}$$

22



Co-localization problem

23



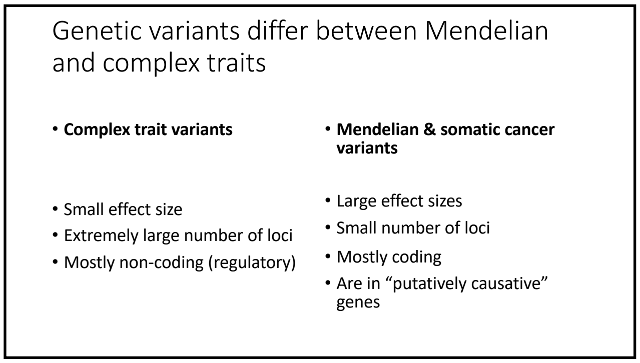
Methods

Coloc

eCAVIAR

JLIM

24



Genetic variants differ between Mendelian and complex traits

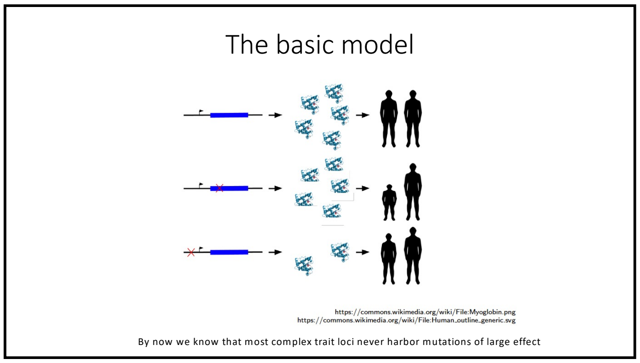
- **Complex trait variants**

- 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

25

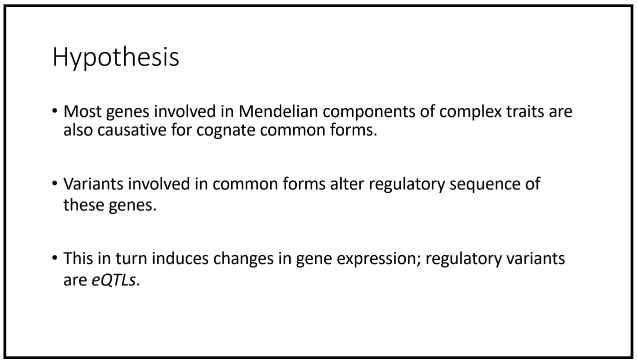


The basic model

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

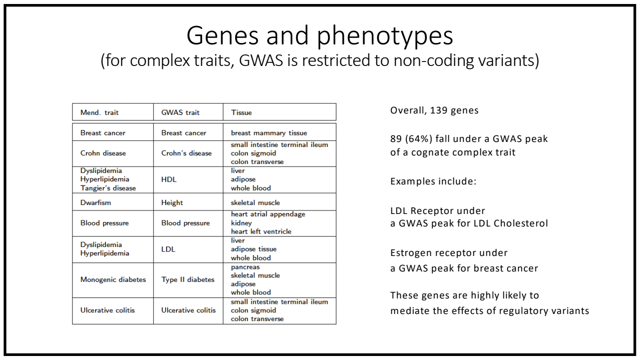
1



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*.

2



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

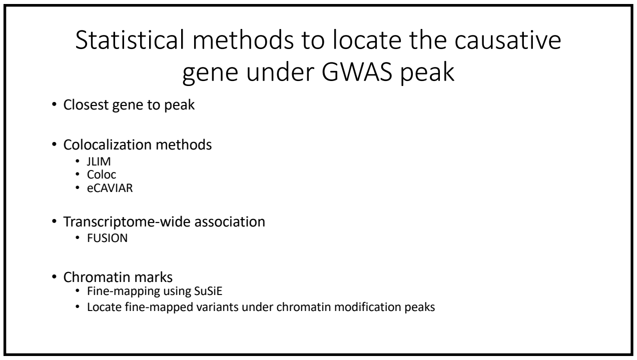
Examples include:

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

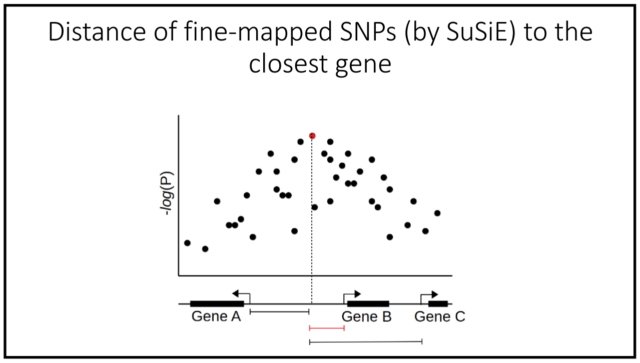
3



Statistical methods to locate the causative gene under GWAS peak

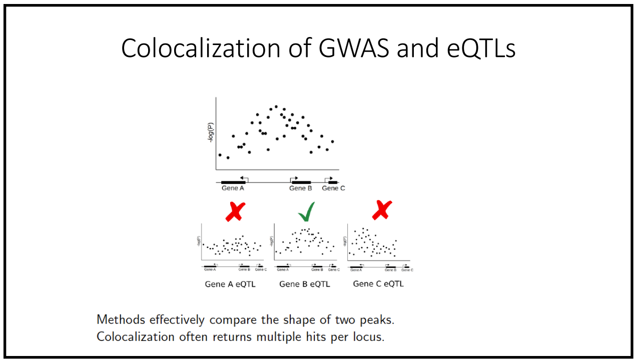
- Closest gene to peak
- Colocalization methods
  - JLIM
  - Coloc
  - eCAVIAR
- Transcriptome-wide association
  - FUSION
- Chromatin marks
  - Fine-mapping using SuSiE
  - Locate fine-mapped variants under chromatin modification peaks

4



Distance of fine-mapped SNPs (by SuSiE) to the closest gene

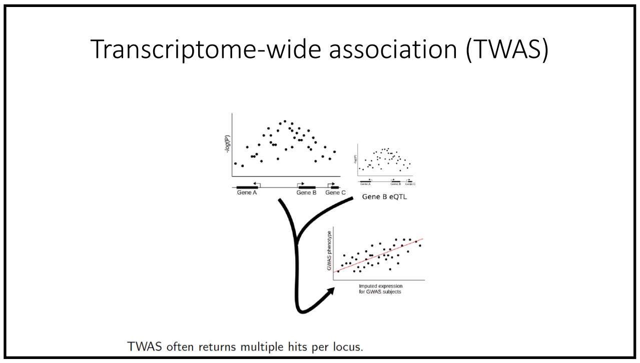
5



Colocalization of GWAS and eQTLs

Methods effectively compare the shape of two peaks.
Colocalization often returns multiple hits per locus.

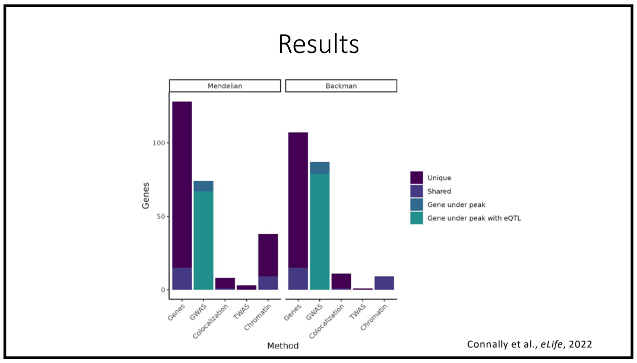
6



Transcriptome-wide association (TWAS)

TWAS often returns multiple hits per locus.

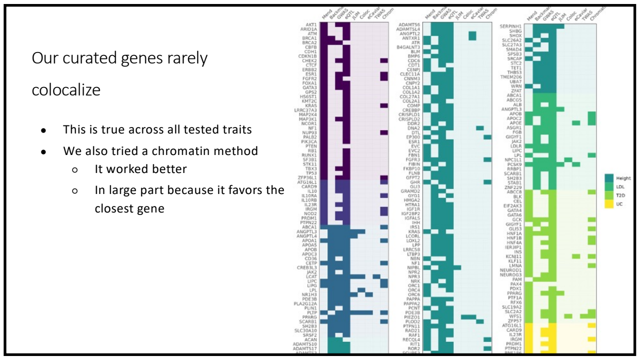
7



Results

Connally et al., *eLife*, 2022

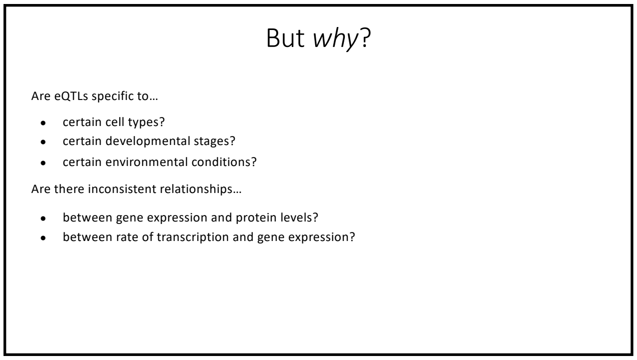
8



Our curated genes rarely colocalize

- This is true across all tested traits
- We also tried a chromatin method
  - It worked better
  - In large part because it favors the closest gene

9



But *why*?

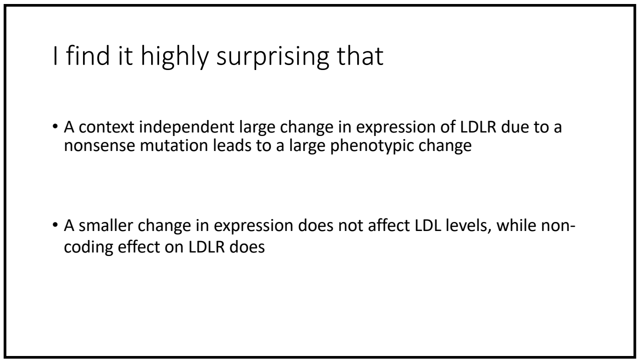
Are eQTLs specific to…

- certain cell types?
- certain developmental stages?
- certain environmental conditions?

Are there inconsistent relationships…

- between gene expression and protein levels?
- between rate of transcription and gene expression?

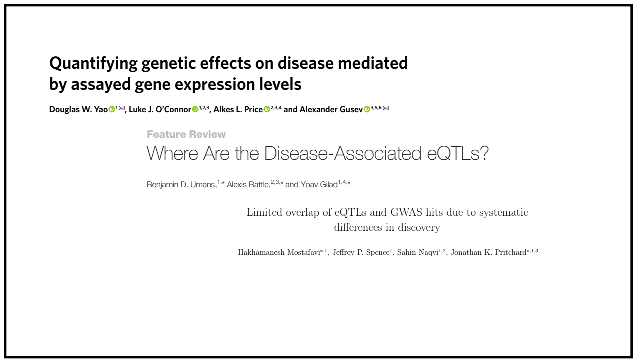
10



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

11



Quantifying genetic effects on disease mediated by assayed gene expression levels

Douglas W. Yao[1], Luke J. O'Connor[1,2,3], Alkes L. Price[2,3,4] and Alexander Gusev[3,5,6]

Feature Review

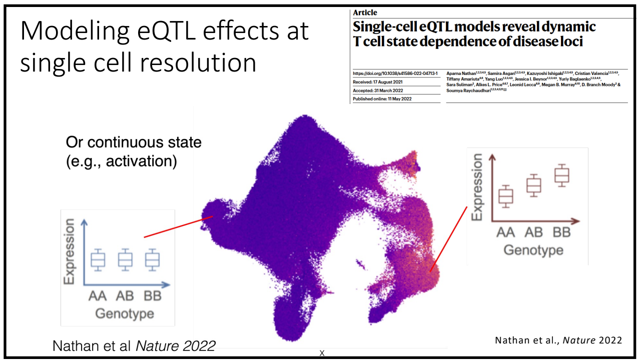
Where Are the Disease-Associated eQTLs?

Benjamin D. Umans,[1,*] Alexis Battle,[2,3,*] and Yoav Gilad[1,4,*]

Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery

Hakhamanesh Mostafavi[*,1], Jeffrey P. Spence[1], Sahin Naqvi[1,2], Jonathan K. Pritchard[*,1,3]

12



Modeling eQTL effects at single cell resolution

Article

Single-cell eQTL models reveal dynamic T cell state dependence of disease loci

https://doi.org/10.1038/s41586-022-04713-1

Received: 17 August 2021

Accepted: 31 March 2022

Published online: 11 May 2022

Aparna Nathan[1,2,3,4,5], Samira Asgari[1,2,3,4,5], Kazuyoshi Ishigaki[1,2,3,4,5], Cristian Valencia[1,2,3,4,5], Tiffany Amariuta[4,6], Yang Luo[1,2,3,4,5], Jessica I. Beynor[1,2,3,4,5], Yuriy Baglaenko[1,2,3,4,5], Sara Suliman[2], Alkes L. Price[4,6,7], Leonid Lecca[8,9], Megan B. Murray[9,10], D. Branch Moody[2] & Soumya Raychaudhuri[1,2,3,4,5,11] ✉

Or continuous state (e.g., activation)

Nathan et al *Nature 2022*

Nathan et al., *Nature* 2022

13