Advanced Gene Mapping Course

January 27-31, 2020
The Rockefeller University
New York, NY

Lectures

Table of Contents

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 12 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 cases and controls
  - Or related individuals (family data) – but need to analyse differently

Two individuals

Person 1	ACCTGTGTGCCCAATGGCGTCCCATACTATCGG
	ACCTGTGCGCCCAATGGCGTCCCATACTATCGG
Person 2	ACCTGTGCGCCCAGTGGCGTCCCATACTATCGG
	ACCTGTGCGCCCAGTGGCGTCCCATAGTATCGG

- Test each SNP for association/correlation with disease 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 SNP (with alleles coded 1 and 2)

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
- Two odds ratios can be estimated
  - $\mathrm{OR}\,(2|2 : 1|1) = \frac{af}{be}$
  - $\mathrm{OR}\,(1|2 : 1|1) = \frac{cf}{de}$

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 $2|2$
- Similarly, we can define the OR for $1|2$ vs $1|1$
  - As the factor by which your odds of disease must be multiplied, if you have genotype $1|2$ as opposed to $1|1$
    - i.e. the 'effect' of genotype $1|2$
- ORs are closely related (often $\approx$) genotype relative risks
  - The factor by which your probability of disease must be multiplied, if you have genotype $1|2$ as opposed to $1|1$ (say)
- If your genotype has no effect on your probability (and therefore on your odds) of disease, then the ORs=1.
  - So the association test can be thought of as a test of the null hypothess that the ORs=1

Genotype relative risks

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

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

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

Dominant/recessive effects

Dominant:

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

Recessive:

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

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

Counting alleles

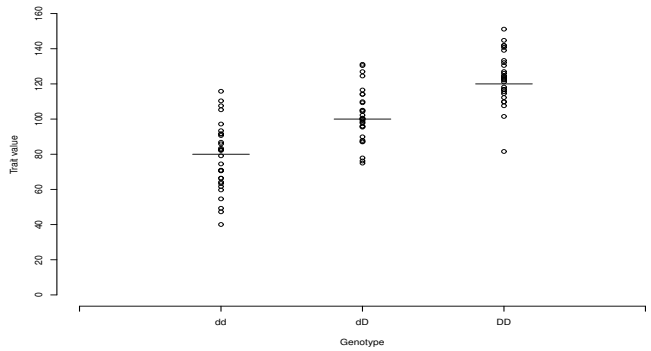
	Counts in	
Allele	Cases	Controls
2	2100 ($=a$)	1220 ($=b$)
1	1900 ($=c$)	2780 ($=d$)
Total	400	400

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.

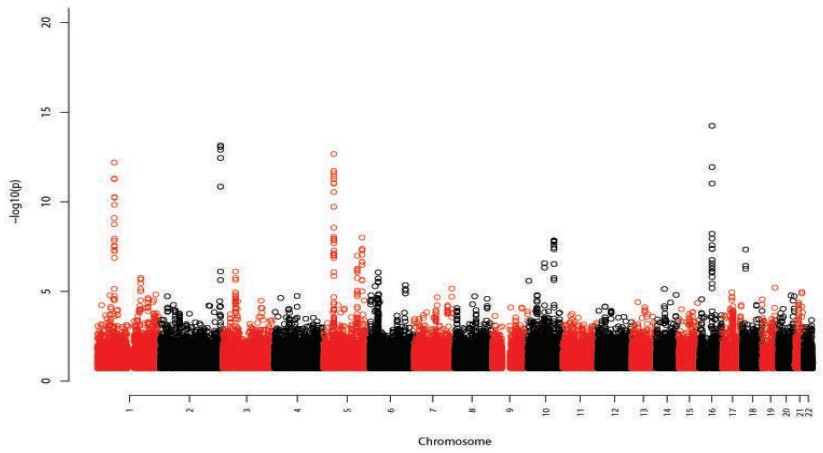
$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

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.

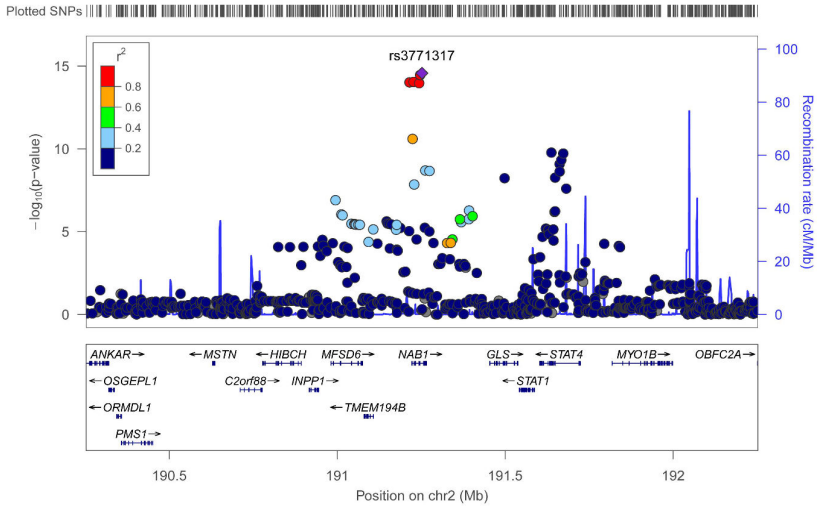
Testing for association

- 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
- 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

Manhattan Plots



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.

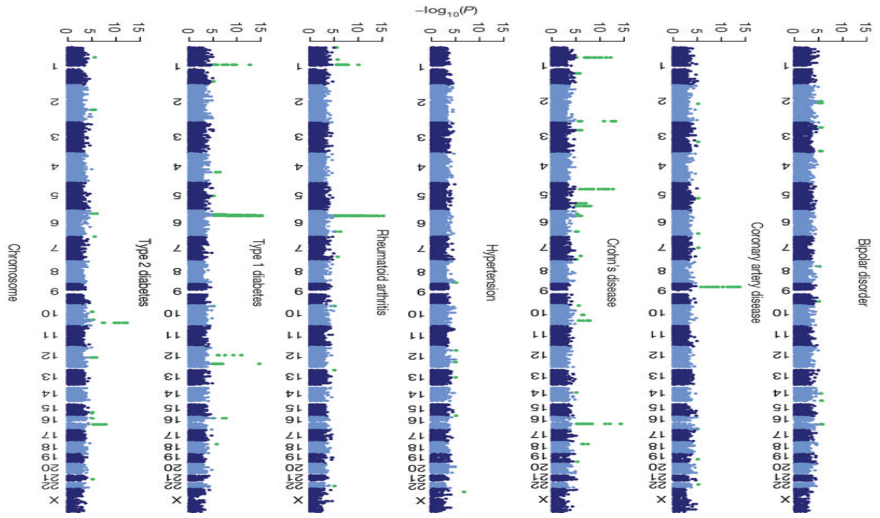
GWAS

- GWAS really got going about 12 or 13 years ago
  - See Visscher et al. (2012) AJHG 90:7-24 "Five Years of GWAS Discovery"
  - And Visscher et al. (2017) AJHG 101:5-22 "10 Years of GWAS Discovery: Biology, Function and Translation"
- 2007/2008 saw a slew of high-profile GWAS publications
  - Breast cancer (Easton et al. 2007)
  - Rheumatoid Arthritis (Plenge et al. 2007)
  - Type 1 and Type 2 diabetes (Todd et al. 2007; Zeggini et al. 2008)
- Arguably the most influential was the Wellcome Trust Case Control Consortium (WTCCC) study of 7 different diseases
  - http://www.wtccc.org.uk/

WTCCC

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

Manhattan plots for 7 diseases



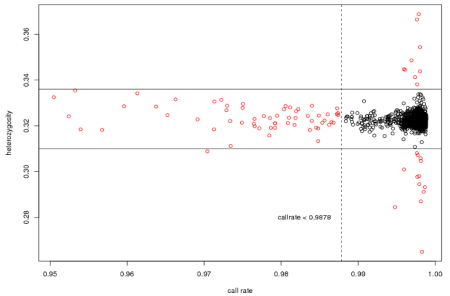
Lessons from WTCCC (and others)

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

Quality Control

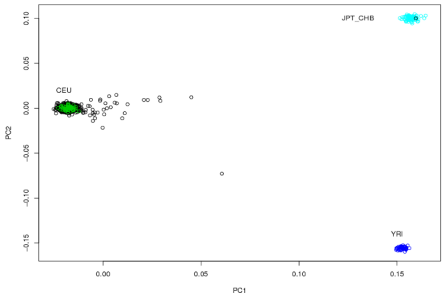
- Stringent QC checks are required for GWAS data
- Discard samples (people) deemed unreliable
  - Low genotype call rates, excess heterozygosity etc.
  - X chromosomal markers useful for checking gender
    - Males should 'appear' homozygous at all X markers
  - Genome-wide SNP data useful for checking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

QC: call rates and heterozygosity



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

QC: ethnicity tests



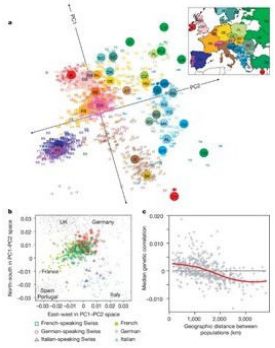
- Multidimensional scaling (with 210 HapMap individuals) identifies 33 samples with non-Caucasian ancestry
- Similar methods can be used to model more subtle population differences between samples

Multivariate Analysis

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

Principal components analysis (PCA)

Genes mirror geography within Europe



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

Principal Components Analysis

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

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

Principal Components Analysis

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

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

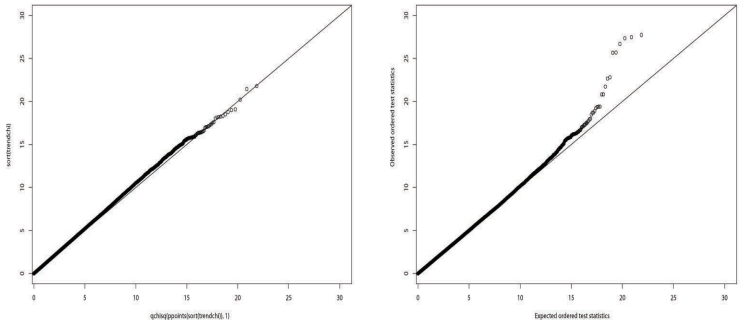
- This matrix will be used as starting point for PCA
  - In principal we could start with a different matrix – in particular not all PCA approaches would normalise by $\sqrt{f_l(1-f_l)}$

Multivariate Analysis

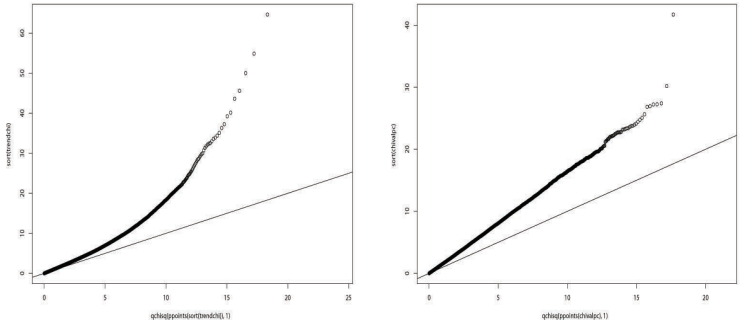
- Estimate covariance matrix $\Psi = \mathrm{X}^T\mathrm{X}$ between all pairs of individuals, with entries ψ_{ij} defined as the covariance (summing over SNPs) between column i and j of X
  - Represents average genome-wide IBD (estimated from IBS)
  - Compute the eigenvectors $\vec{v}_j$ and eigenvalues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

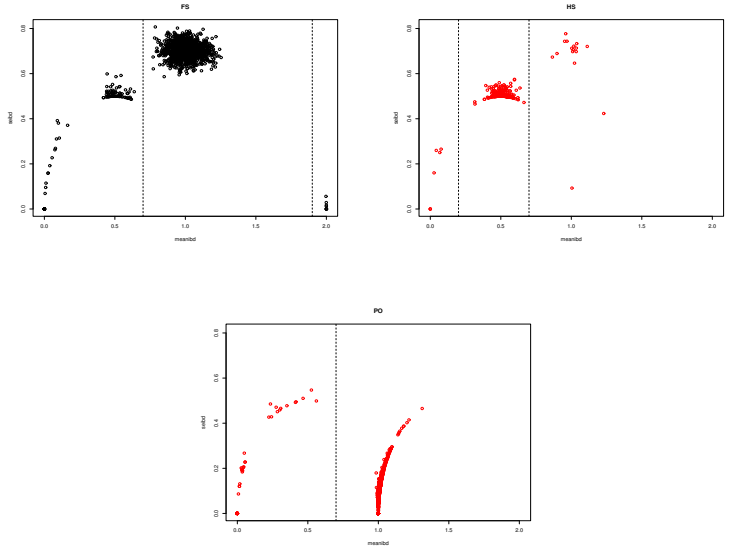
Expected IBD sharing

- Assuming no inbreeding, the IBD state probabilities are:

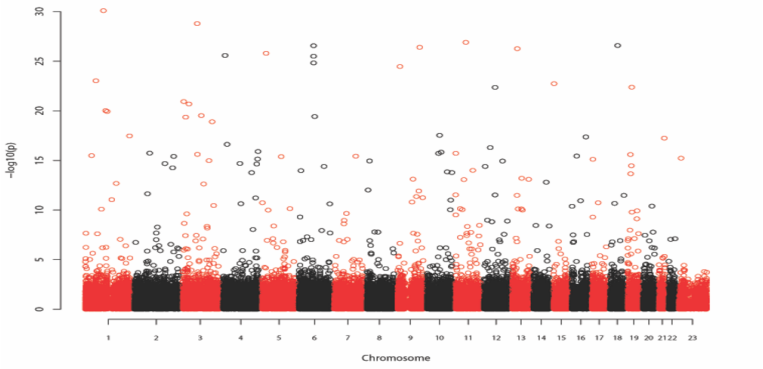
	Number of alleles shared IBD		
Relationship	2	1	0
MZ twins	1	0	0
Parent–Offspring	0	1	0
Full siblings	1/4	1/2	1/4
Half siblings	*0*	*1/2*	*1/2*
Grandchild–grandparent	*0*	*1/2*	*1/2*
Uncle/aunt–nephew/niece	*0*	*1/2*	*1/2*
First cousins	0	1/4	3/4
Second cousins	0	1/16	15/16
Double 1st cousins	1/16	6/16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

- A useful visualisation tool is to plot SE(IBD) vs mean(IBD)

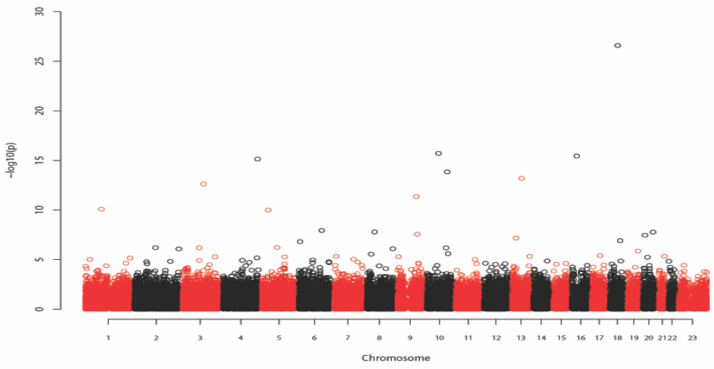
Full/half sibs and parent-offspring



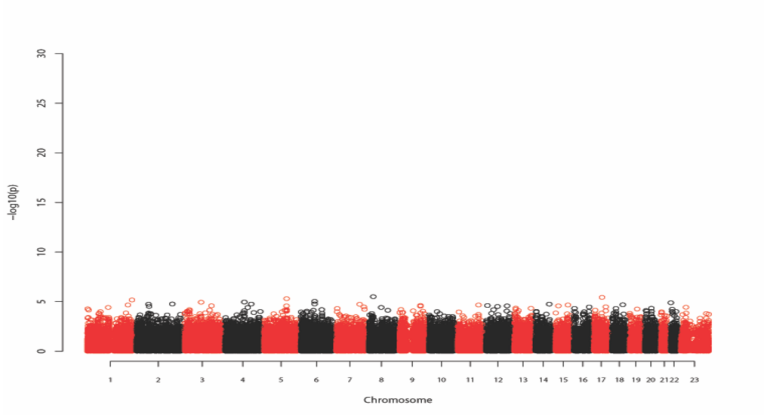
CHD GWAS results (low QC)



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
- 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

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 to use WGS samples for Copy Number Variant analysis
- Buccal Swabs
  - Small amounts of DNA
  - DNA not stable
- Saliva (Origene collection kit)

Measurement of DNA Concentrations

- Nanodrop
- Picogreen

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
- Determines the sex of the individual
  - To aid in identification of sample swaps
- Detects cryptic duplicates
- For family data
  - Aids in determining close familial relationships
    - Non-paternity
    - Sample swaps
    - Cryptic relationships

Detecting Genotyping Errors

- Duplicate samples genotyped to detect inconsistencies
  - Can use duplicate samples that are inconsistent to adjust clusters to improve allele calls
    - Will not detect systematic errors
- Usually not performed for exome and whole genome sequencing studies

Effects of Genotyping Error

- If there is no bias in genotyping error between cases and controls
  - Same rates of genotyping errors in case and control data
- For family based association studies - Trios
  - Can increase both type I and II error
- Population based studies
  - Increases type II error only
- Cases and controls are genotyped
  - At different times
  - Different institutions
  - Or one group, case or control, is predominately genotyped at one time/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 genotyping cases and controls
  - Randomize cases and controls so they are spread evenly across genotyping runs.

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

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 populations and errors
  - Can not adjust for confounders
    - e.g. sex, population substructure/admixture
- Don't perform an aggregate test using frequency information from gnomAD

Genotype Data QC – Population Based Studies

- Remove DNA samples from individuals who are missing >3% or their genotype data
  - May choose to use an even more stringent criteria
- Low DNA quality can lead to higher genotypes error rates at markers with available genotype data
- Lower call rates in individuals may be due to DNA contamination with another DNA sample
- To avoid markers with higher genotyping error rates
  - For markers with a minor allele frequency (MAF)$\geq$0.05
    - Remove markers missing >5% of their genotype data
  - For makers with a MAF<5%
    - Remove markers missing > 1% of their genotype data

Additional QC Family based studies

- Detect double recombination events over small distances
  - Can be an indication of genotyping error
    - Merlin
- Detect non-Mendelian errors of segregation
  - Can be due to genotyping errors*
  - If a larger number are observed
    - Could be due to incorrect specification of pedigree structure
      - e.g. non-paternity
    - Pedcheck

*Many genotyping errors will not be detected for single nucleotide variants (SNVs). The probability of detecting non-Mendelian segregation due genotyping errors decreases with increasing MAF

Data Clean – Accessing Sex

- Males with an excess of heterozygous SNPs on the X chromosome can denote
  - Males mislabeled as females
  - Males with Klinefelter syndrome
  - Note: Males will be heterozygous for markers in the pseudoautosomal regions
- Females with an excess of homozygous genotypes on the X chromosome can denote
  - Females mislabeled as males
  - Females with Turner Syndrome

Data Clean – Accessing Sex

- Males mislabeled as females and females mislabeled as males
- Can be observed due to sample mix-ups
- Samples for which the sex is incorrect
  - Should be removed from the analysis
- Probably not the person you think it is

Checking for Potential DNA Contamination

- DNA samples which have been contaminated by another DNA sample will have a larger proportion of their genotypes being heterozygous*
- If cross contamination of samples within the same study will observe “relatedness” amongst study subjects
- Can also be observed from principal components analysis (PCA)/multidimensional scaling (MDS)
  - Samples which are cross contaminated will cluster together and not cluster with other samples

*Higher levels of heterozygous markers will be observed in individuals of sub-Saharan African ancestry compared to those of European and Asian ancestry.

Checking for Duplicate and Related Individuals

- Duplicate samples are sometimes included in a study as part of quality control
  - These can easily 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

Identifying Duplicate and Related Individuals

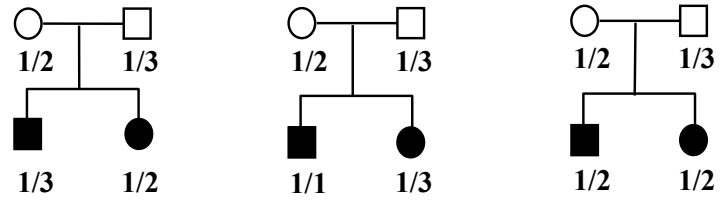
- Genotype data from one of the duplicates needs to be removed from the dataset
- Only one related individual should be retained in the data set
  - If related individuals remain in the data set mixed-models need to be used to analyze the data*
    - Case-Control
      - Generalized linear mixed models
    - Quantitative traits
      - Linear mixed models
  - If not type I error rates can be increased

*If only a few related individuals in sample, may wish to remove them or use mixed-models to control type I errors. Must use mixed models if many related individuals in dataset. Due to the construction of the generalized relationship matrix (GRM) can be problematic to apply for large samples sizes, e.g. UKbiobank.

Identifying Duplicate and Related Individuals

- Duplicate and related individuals can be detected by examining identify by state (IBS) adjusted for allele frequencies (p-hat) between all pairs of individuals within a sample
  - To estimate IBD sharing

IBD/IBS



IBD=0 IBS=1

IBD=1 IBS=1

IBD=2 IBS=2

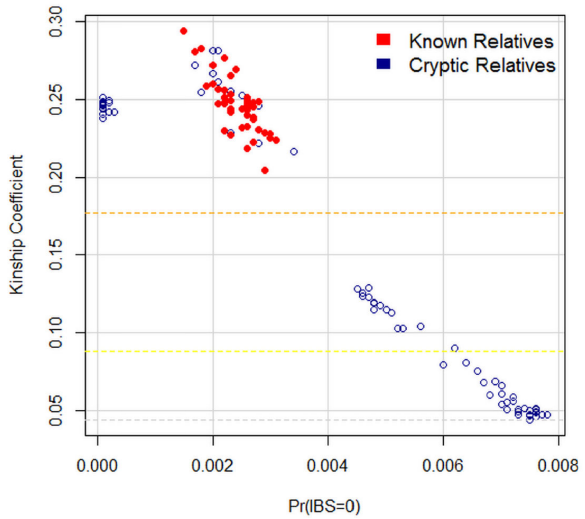
Identifying Duplicate and Related Individuals

- IBS is the number of alleles of alleles which are shared between a pair of individuals
  - Can either share 0, 1, and 2 alleles
- Duplicate individuals will have IBS measures of 1 or close to 1
  - If there is genotyping error could make IBS< 1
- IBS measures are adjusted for by allele frequencies p-hat
  - Approximates IBD sharing

Identifying Duplicate and Related Individuals

- Siblings and child-parent pairs will share 0.50 of their alleles IBD
  - For parent-child IBD=1 is ~1.0
  - For sibs IBD=1 is ~0.50
    - For more distantly related individuals the IBD measure will be lower
- Can use whole genome scan data to check IBD
  - Should "trim" markers so that they are not in LD
  - Caution should be used if using candidate genes are being studied since many markers will be in LD which can inflate the IBD measure
- PLINK or KING can be used to identify duplicate and related individuals

King Graphical Output



Observing Low Levels of IBD Sharing for Multiple Individuals

- P-hat is calculated using the "population" allele frequency
- If individuals in sample come from different populations
  - Individuals from the same population within the sample will have inflated p-hat values due to incorrect allele frequencies
    - Appear incorrectly 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

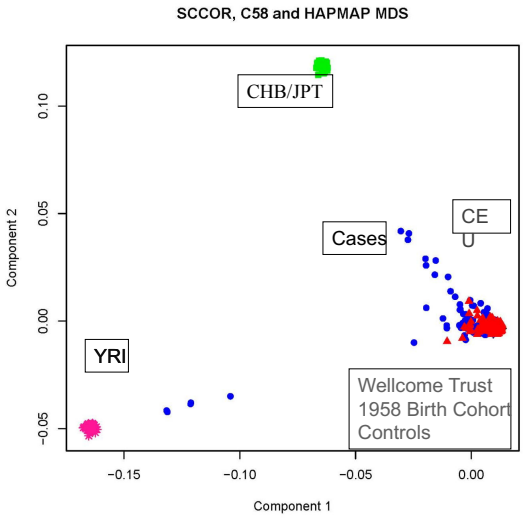
PCA/MDS

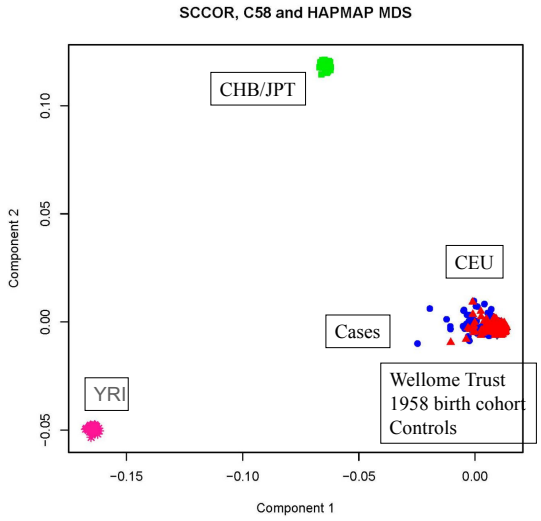
- Can be used to find outliers
- Individuals from different ethnic backgrounds
  - African American Samples included in samples of European Americans
- Use a subset of markers which have been LD pruned
  - So there is only very low levels of LD between marker loci
- Plot 1st component vs. 2nd component
  - Additional components should also be plotted to determine potential biases in the data

PCA/MDS

- Used to identify outliers
  - Individuals from different ancestries
    - e.g. African American samples included with European Americans samples
      - Can use samples from HapMap to help to determine what ethnic group outliers belong to.
        - » Should not include HapMap samples when calculating components to control for population substructure/admixture
  - Related individuals
  - Problems with genotype quality
    - Batch effects

Detecting Outliers Using PCA





Detecting Genotyping Error – Examining HWE

- Testing for deviations from Hardy-Weinberg Equilibrium (HWE) not very powerful to detect genotyping errors
- The power to detect deviations from HWE is dependent on
  - Error rates
  - Underlying Error Model
    - Random
    - Heterozygous genotypes -> homozygous genotypes
    - Homozygous genotypes ->Heterozygous genotype
  - MAFs

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 a sample of the same ancestry
  - Population substructure can introduce deviations from HWE
- Do not include related individuals when testing for deviations from HWE
  - Can cause deviation from HWE
- 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

Detecting Genotyping Error

- 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
- In the presents of genotyping error
  - Random error and equal allele frequencies
  - Power to detect deviation from HWE is α
- Pseudo SNPs lend themselves to readily be detected through testing for deviation from HWE

Testing for Deviations in HWE - α value

- For different studies a variety of α values are used
  - Need to consider that multiple testing is being performed
- WTCCC used a criterion $p<5 \times 10^{-7}$ to remove SNPs
- A large variety of criterion are used to reject the null hypothesis of HWE
  - A criterion of 1.0×10^{-4} is often used in published studies
- Deviations from HWE can be used to flag SNPs
  - e.g. Remove those SNPs with deviations from HWE with $p<5.0 \times 10^{-8}$
    - Note should be more conservative if performing quality control for imputation
  - SNPs can then be investigated in more detail later for reasons for the observation of deviation from HWE
    - If there are significant association results
- A significant result with a large deviations from HWE in cases and controls is probably due to genotyping error or a pseudoSNP

Deviation from HWE

- HWD coefficient (Weir 1996)

$$D = P_{11} - p^2$$

- For SNPs (2 allele system)
- The proportion observed for N genotypes G_{11}, G_{12} and G_{22}
  - Is P_{11}, P_{12} and P_{12} respectively
- p is the allele frequency which is estimated by

  $(2*G_{11} + G_{22})/2N$

Deviation from HWE

- Under HWE D=0
- Negative values of D indicate an excess of heterozygote genotypes
- Positive values of D indicate an excess of homozygote genotypes
- For a diallelic system
  - D can range from -0.25 to 0.25
- Markers not in HWE
  - May be due to population admixture
    - Excess of heterozygous genotypes (-D)
  - Copy number repeats (CNVs)
    - Either an excess of heterozygous (-D) or homozygous genotypes (+D)
- Heterozygous Advantage
  - Excess of heterozygous genotypes (-D)

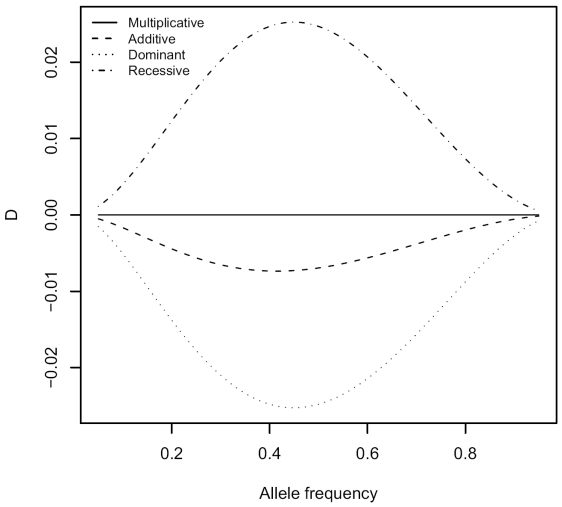
Detecting Genotyping Error

- Chance
  - Either an excess of heterozygous (-D) or homozygous genotypes (+D)
- Deviation from HWE due to genotyping error
  - Either an excess of heterozygous (-D) or homozygous genotypes (+D)
- Deviation from HWE due to pseudo SNPs
  - Excess of heterozygous (-D) genotypes
- Indication that deviation from HWE due to genotyping error
  - Higher genotype drop out rates for specific markers
- Pseudo SNPs
  - Primers map to multiple genomic regions

Deviation from HWE

- Genotype data from cases deviate from HWE when the SNP which is being tested is functional or in LD with a functional variant
  - Additive genetic model (D- excess heterozygous genotypes)
  - Dominant genetic model (D- excess heterozygous genotypes)
  - Recessive genetic model (D+ excess homozygous genotypes)
  - Multiplicative genetic model there is no deviation from HWE (D=0.0)
- Controls – unaffected individuals
  - Display an extremely small deviation from HWE
  - Deviation from HWE is higher with higher disease prevalence
- Only unascertained samples have no deviation from HWE at the functional locus

Deviation from HWE



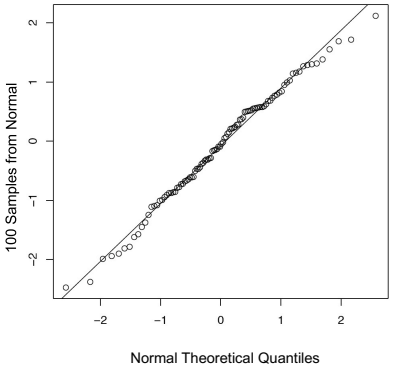
Odds Ratio
1.5

QQ plot

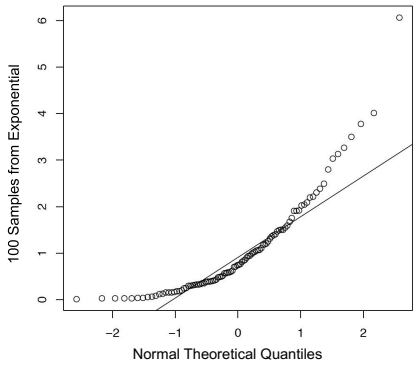
- QQ (quantile-quantile) plot is a graphical method for diagnosing differences between a random sample and a probability distribution
- Assume there are n points, ordered as x(1)<…<x(n). Then for i=1,…,n, we plot x(i) against the ith quantile (qi) of the specified distribution (e.g., normal).
- If the random sample is from the specified distribution, the QQ plot will be a straight line
- Let F() be the cumulative distribution function of a random variable, e.g., F(z)=P(x<z).
- The ordered n samples, x(1)<…<x(n), are treated as n empirical quantiles and the ith theoretical quantile, qi, corresponding to x(i) can be calculated as qi=F-1((i-0.5)/n)
- Plot qi on x axis and x(i) on y axis

QQ Plot Examples

Normal-Normal QQ plot



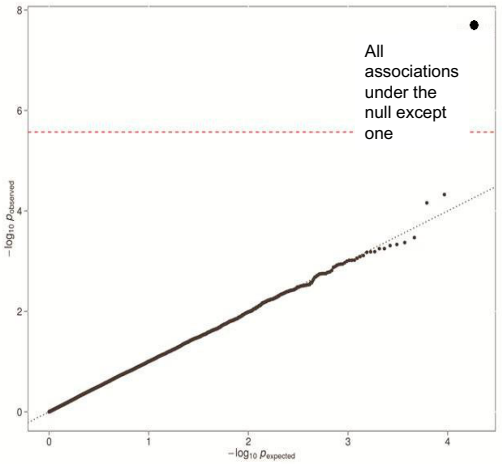
Exponential-Normal QQ plot



Genome Wide Association Diagnosis

- Thousands of markers 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 a intuitive way to visually detect biases

QQ Plot of Exome Wide P Values



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
- Problematic to examine the mean of the test statistic
  - If for a number of variants the null can be rejected
    - Particularly if they have very small p-values
      - The mean test statistic will be inflated

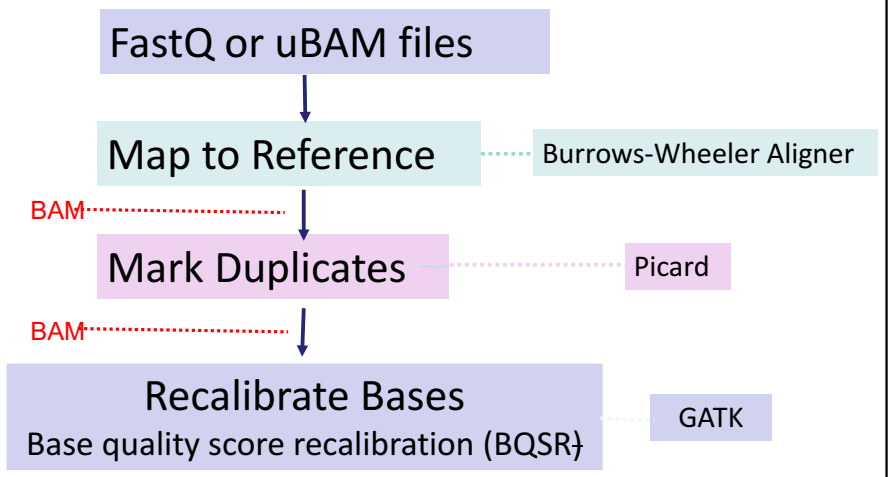
Phenotype	Covariate	Mean Chi-Square	GIF (λ)
BP		1.23829	1.16932
BP	Age	1.24119	1.18025
BP	**Age-EV1**	1.09471	*1*
BP	Age-EV2	1.0881	1
BP	Age-EV4	1.08385	1
BP	Age-EV10	1.09582	1.00402
BPI		1.14931	1.08921
BPI	Age	1.15139	1.08113
BPI	Age-EV1	1.05079	1.01148
BPI	**Age-EV2**	1.0428	**1**
BPI	Age-EV4	1.04204	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

Significance Results

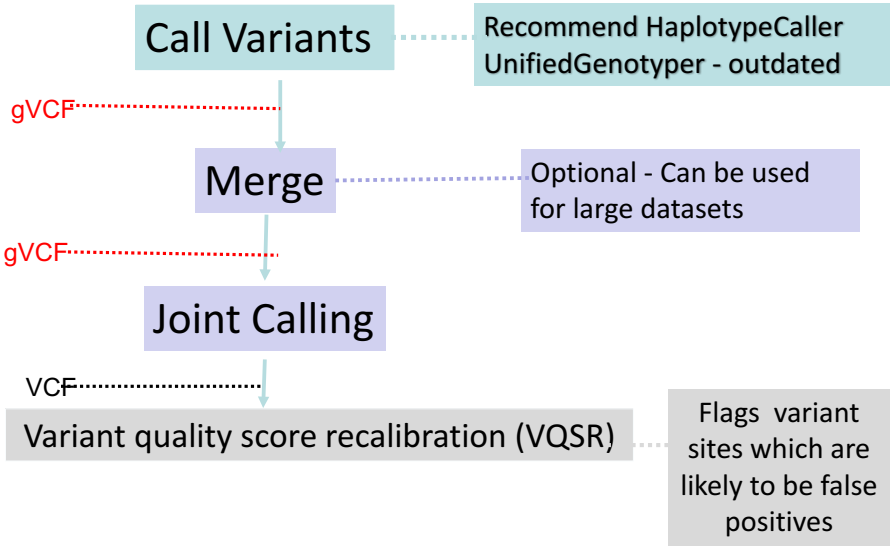
- For those SNPs with strong associations
- Cluster plots should be examined
- Poor clustering – overlap between clusters
  - Could be the reason for the strong association
- Cluster caller should be adjusted when possible and data reanalyzed

NGS Data Quality Control

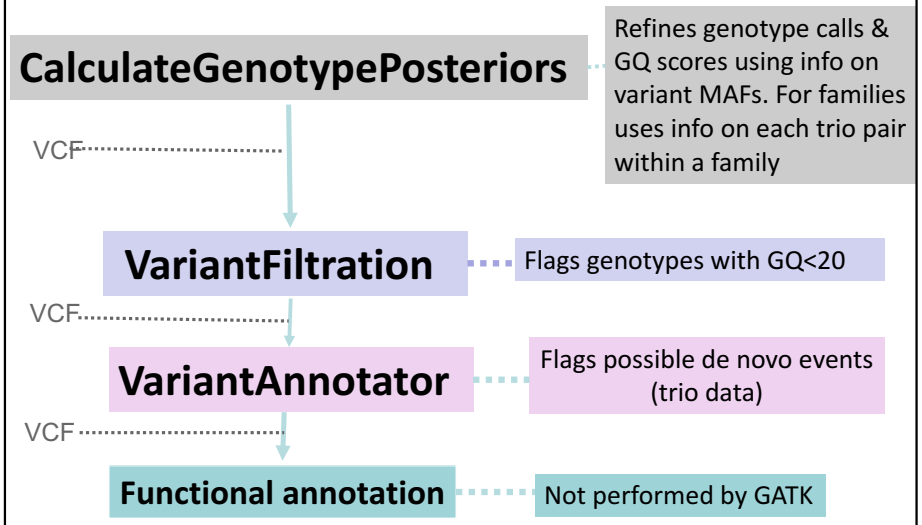
Variant Calling Pipeline -Step 1 Preprocessing



Variant Calling Pipeline-Step 2 Variant Discovery



Variant Calling Pipeline - Step 3 Call Set Refinement



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

Working with gVCF

- Instead of obtaining VCF files
- Can obtain gVCF files to perform joint calling and the rest of the GATK pipeline
  - A whole genome gVCF
    - ~1 Gigabyte
      - 1/100th the size of a BAM file for one individual
- Need additional information on how variants were called
  - e.g. HaplotypeCaller or UnifiedGenotyper
    - Not valid to use Unified Genotyper

Influences on Sequence Quality

- DNA quality
  - Age of sample
  - Extract method
  - Source of sample
- Sequencing machines (read length)
- Median sequencing depth
- Alignment
- Variant calling method used
  - SNVs and Indels

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
- Sequence quality can be influenced by
  - DNA quality
  - Sequencing machines (read length)
  - Sequencing depth
  - Alignment
  - Variant Calling
    - SNVs and Indels
- Protocols for data QC are still in their infancy
  - No set protocols for QC
- QC which has to be performed is data specific
  - Dependent on read depth
  - Batch effects
  - Availability of duplicate samples
  - etc

NGS Data Quality – Removal of Genotype Calls and Samples

- Sequence read genotype depth (GD)
  - Concerned if GD is too low or too high*
    - GD too low insufficient reads to call a variant site
    - GD too high can be an indication of copy number variants which can introduce false positive variant calls
      - *Due to down sampling in GATK maximum GD is 250
    - Remove genotypes with low read depth, e.g. GD<8
  - Genotype quality (GQ) score
    - Removal of sites with low genotype quality core, e.g. GQ< 20
- Remove individuals who are missing genotype calls/variant sites, e.g. > 10%
  - To remove individuals with bad quality data who can potentially have incorrect genotype calls
- If using different capture arrays use the intersect of the arrays

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 if the study consists of more than one ancestry group, e.g. African American and European American
  - Related individuals should also be removed from the sample before testing for deviations from HWE

NGS Data Quality Control

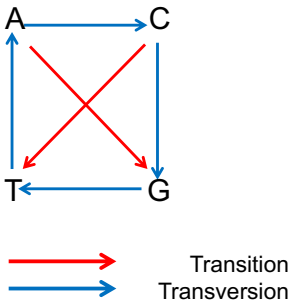
- Variant Quality Score Recalibration (VQSR) or
  - GATK
- Used to determine variant sites of bad quality
- However even after this step
  - Concordance of duplicates (when available) and
  - and Ti/Tv ratios are often low
- Additional QC steps needs to be performed

NGS Data Quality Control

- Values which are used for GD, GQ, and missing data cut offs are based upon
  - Concordance rates
    - if there duplicate samples are available
  - Ti/Tv ratios
    - For individuals
    - Entire sample
  - Removal of batch effects
    - As evaluated by multidimensional scaling (MDS) or
    - Principal components analysis (PCA)
  - Amount of data removed
    - QCl can remove substantial amounts of data which should be avoided
      - e.g. >15% of variant sites

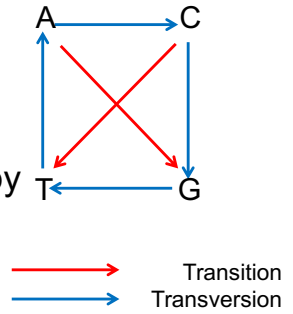
Transition/Transversion (Ti/TV) Ratios

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



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



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

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

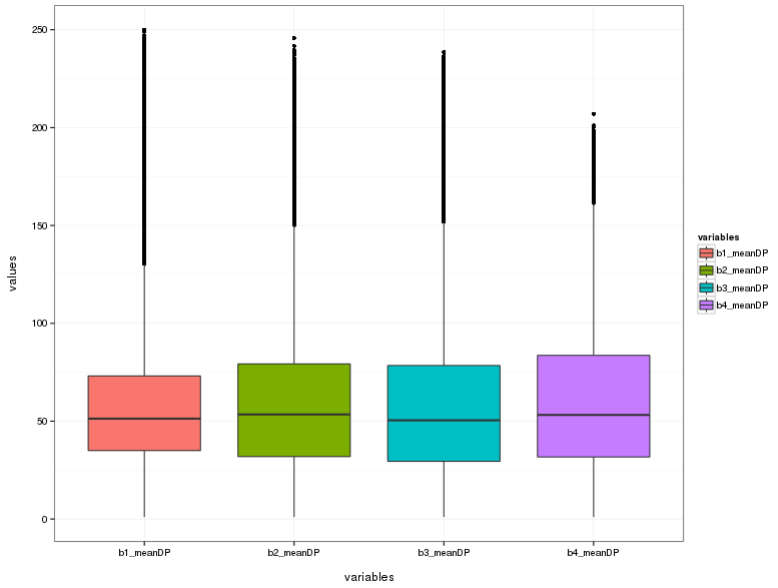
Sequence Data QC Overview

- Variant and genotype call level
  - Evaluation of batch effects
- Genotype call level – Removal of genotype calls
  - Low or high depth of coverage GD < 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
- 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
      - $p<5 \times 10^{-8}$

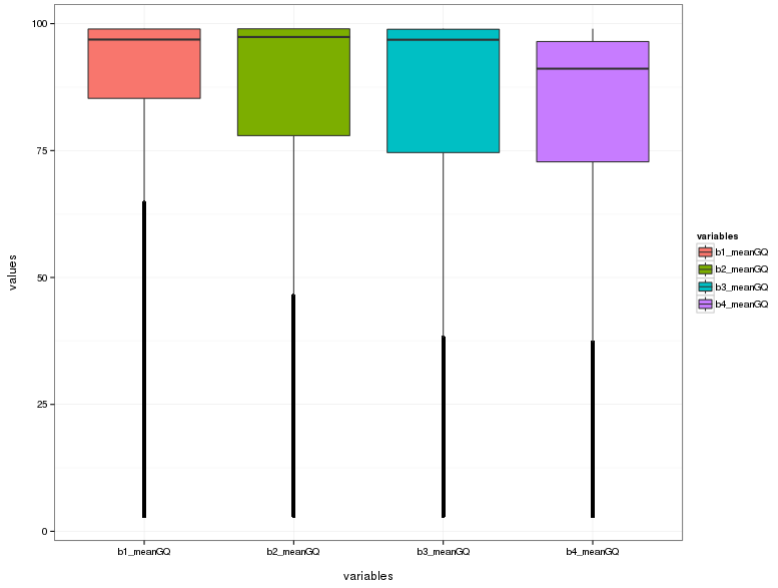
Sequence Data QC Overview

- Detection of sample outliers
  - Perform multidimensional scaling (MDS) to detect outliers
    - Due to population substructure/admixture and batch effects
    - Remove effects by
      - Additional QC
      - Removal of outliers and\or
      - Inclusion of MDS or PCA components in the association analysis
- Evaluate sex of individuals based upon X and Y chromosomal data
  - Sample mix-ups
  - Individuals with Turner or Klinefelter Syndrome
- Evaluate samples for cryptically related individuals and duplicates
  - King or Plink algorithm
    - Retain one duplicate of a pair
    - Retain only one individual of a relative group or control for relatedness in the analysis, i.e. mixed models
- Post Analysis - Quantile-Quantile (QQ)plots
  - To evaluate uncontrolled batch effects and population substructure/admixture

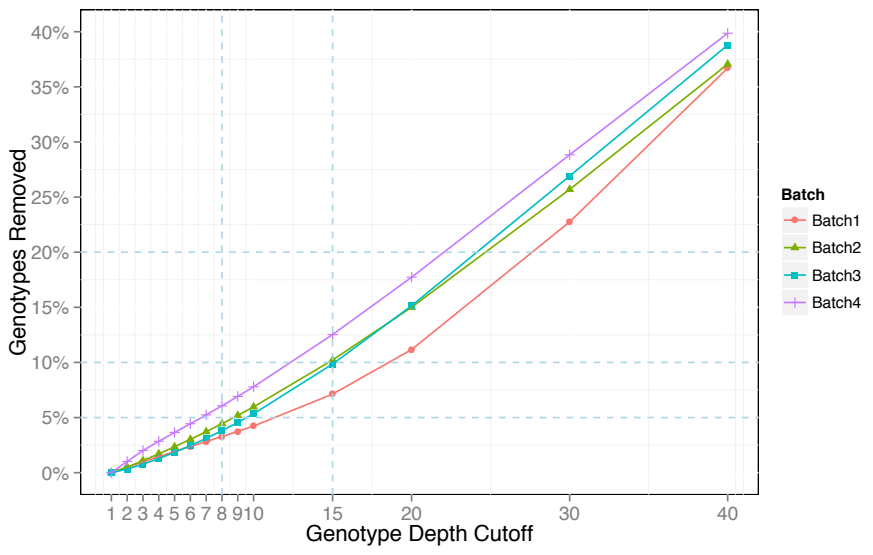
Mean DP by Batch



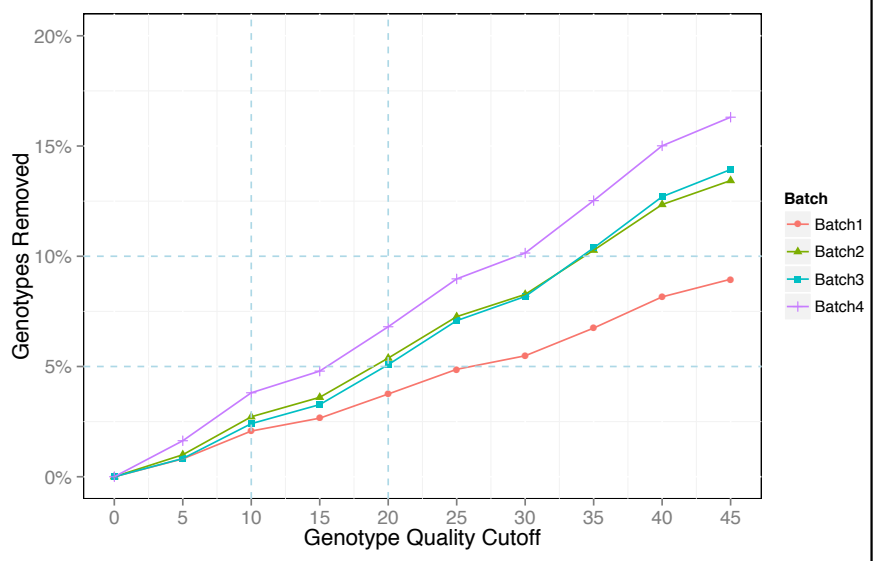
Mean GQ by Batch



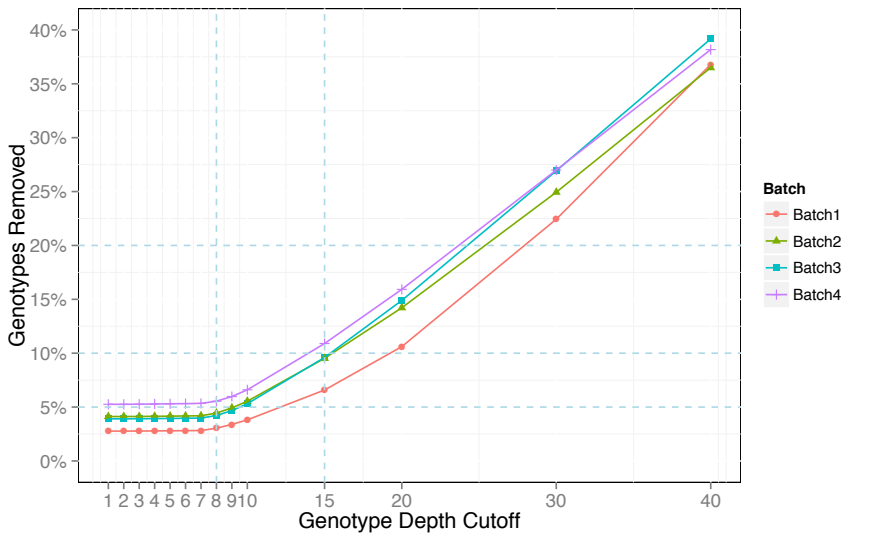
Genotypes Removed by GD Cut-off by Batch



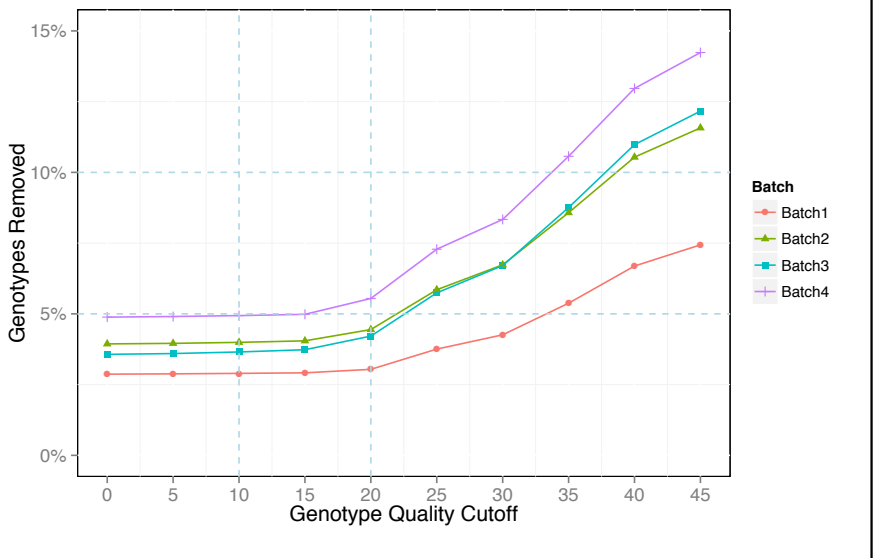
Genotypes Removed by GQ Cut-offs by Batch



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



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



Missing Rate Criteria & Sites Removed

	10%	5%
Before QC*	2.5%	3.9%
After QC	12.9%	18.3%

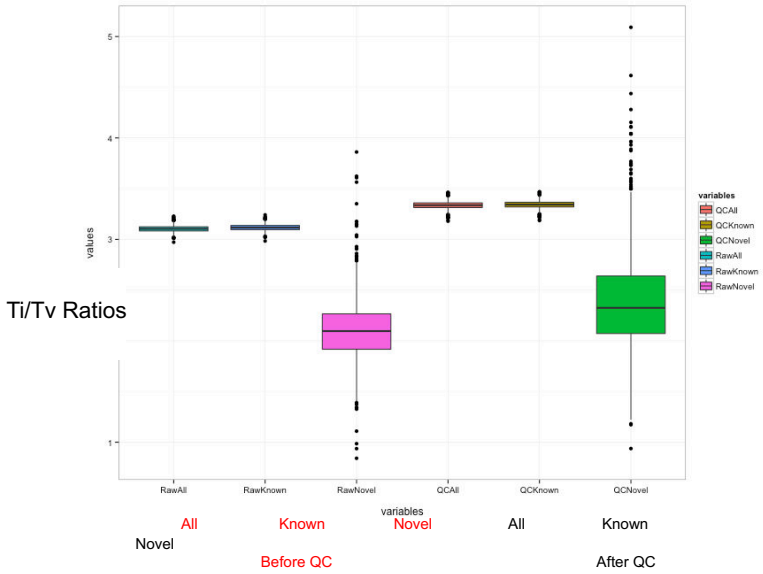
*After VQSR

Variant sites missing >10% of their data were removed

Ti/Tv Ratios during QC Process

	Known	Novel	All
Before VQSR	2.95 ± 0.05	1.18 ± 0.29	2.86 ± 0.07
Before QC	3.12 ± 0.03	2.01 ± 0.32	3.11 ± 0.03
Genotype QC GD<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

Ti/Tv Ratios by Individual Before and After QC



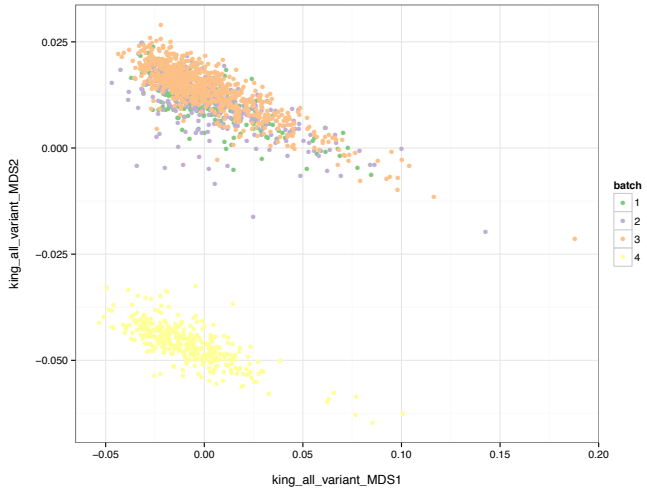
Detecting Outliers Using Multidimensional Scaling (MDS)

- Multidimensional Scaling (MDS) and principal components analysis (PCA) are frequently used to detect outliers
  - MDS & PCA can also be used to control for substructure in the analysis
- Outliers can be cause by
  - Population stratification
  - Population substructure
  - Batch Effects

Sequence Data QC

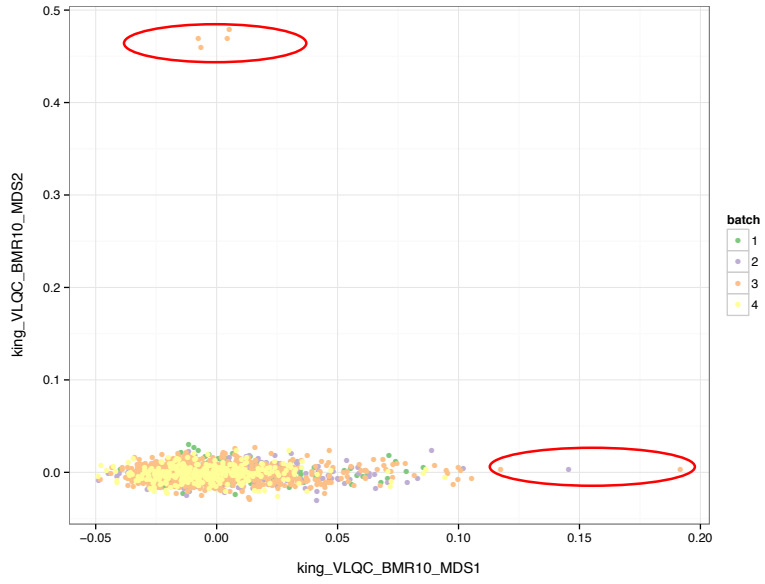
- Batch effects can sometimes be removed with additional QC
- Extreme outliers should be removed
- Additionally MDS or PCA components can be included in the analysis to control for population substructure\admixture and batch effects
  - Unless correlated with the outcome (phenotype)
- Batch effects (dummy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

MDS First 2 Components Before QC*



*After VQSR

MDS First 2 Components After QC



Order of Data Cleaning

- Remove samples missing >10% genotype data
- Remove SNPs with missing genotype data
  - If MAF >5%
    - Remove markers with >5% missing genotypes
  - If MAF <5%
    - Remove markers with >1% missing genotypes
- Remove samples missing >3% genotype calls
- Check for sex of individuals based on X-chromosome markers
  - Remove individual whose reported 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. $r^2 < 0.5$
  - Retain duplicate with best genotyping quality

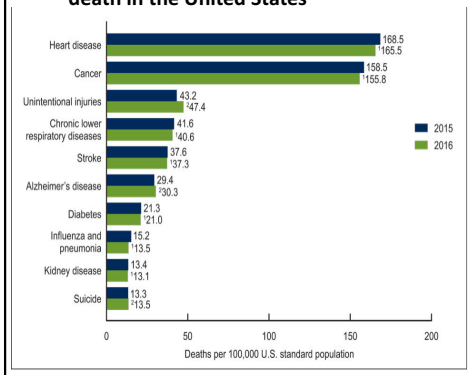
Order of Data Cleaning

- Perform principal components analysis to check for outliers
  - Use trimmed data set of markers which are not in LD
    - E.g. $r^2 < 0.5$
  - Remove outliers from data
- Check for deviations from Hardy Weinberg Equilibrium
  - Separately in cases and controls
  - If more than one ethnic group
    - Separately for each ethnic group
- Examine QQ plots for potential problems with the data
  - e.g. not controlling adequately for population admixture

Complex Trait Association Analysis of Rare Variants Obtained from Sequence Data: Population-Based Data

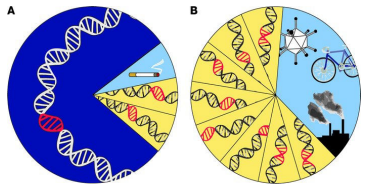
Complex Diseases (Traits)

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

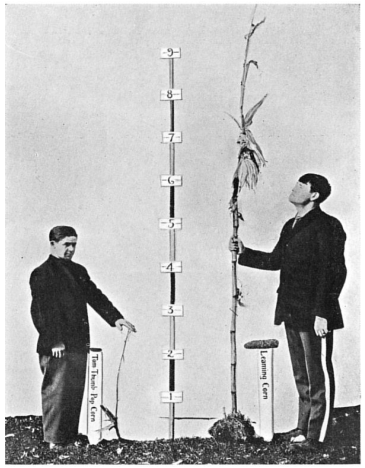


T.A. Manolio, et al. J clin Invest, 2001

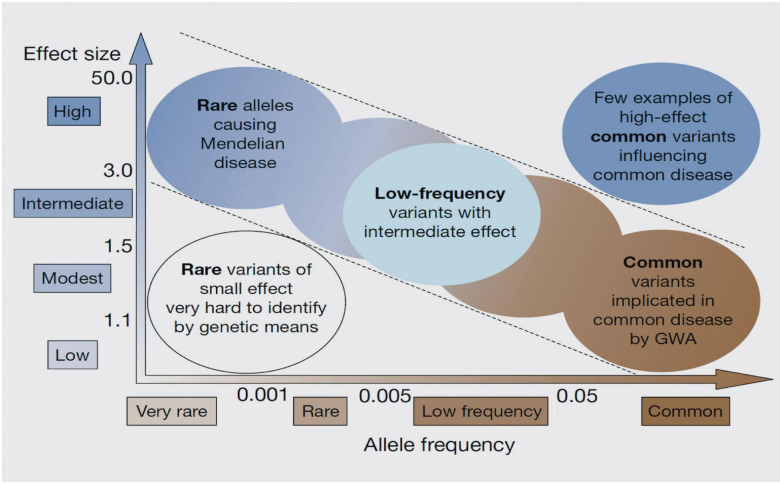
Heritability for Common Traits

Human height heritability is ~80%

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



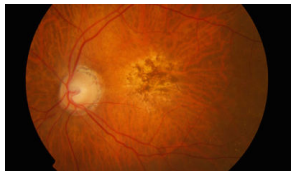
Allelic Architecture



T. A. Manolio et al. Nature, 2009

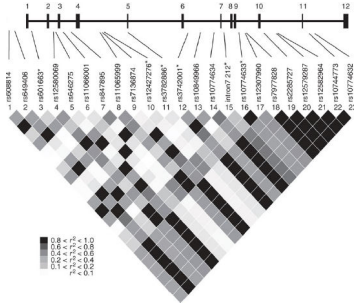
Complex Disease – Common Variant Associations

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

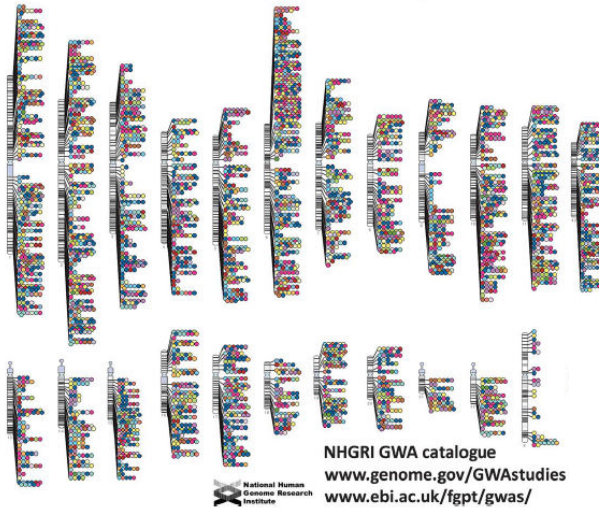


Studying Complex Traits – Common Variant Associations

- 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



Complex Trait – Common Variant Associations



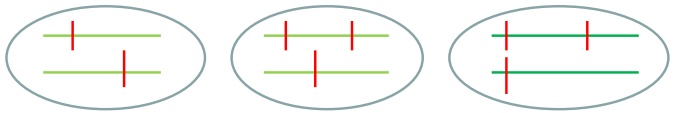
- Although highly successful in identifying thousands of complex trait loci
- Usually pathogenic susceptibility variant(s) not identified

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 >10 years ago
  - Dallas Heart Study
    - Small sample ~1,200 indiviuals
      - Multi-ethnic
      - Used “extreme” sampling
    - Plasma low density lipoprotein levels (Cohen et al. 2004)
      - NPC1L1

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



Gene 1 Gene 2 Gene 3

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.

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

Types of Aggregate Analyses

- Frequency cut offs used to determine which variants to include in the analysis
  - Rare Variants (e.g. <1% frequency)
  - Rare and low (1-5%) 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

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

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

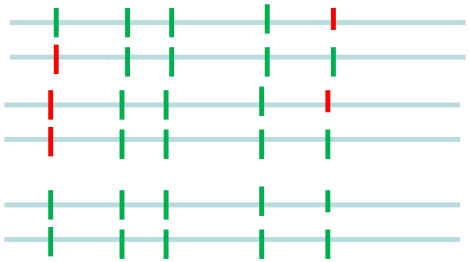
Can also use same coding in a regression framework

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

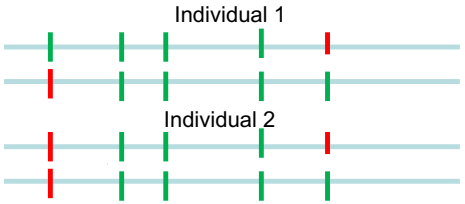
Red bars: Minor allele has been observed

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

GRANVIL

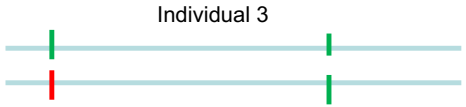
- Example of coding used in regression framework
  - Gene region with 5 variant sites – data available on all sites
  -



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

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

Methods to Detect Rare Variant Associations Maximization Approaches

- Variable Threshold (VT) method
  - Uses variable allele frequency thresholds and maximizes the test statistic
  - Also can 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

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

Optimal Test

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

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
- Often a Bonferroni correction for testing 20,000 genes is often used as the significance level cut-off
  - 2.5×10^{-6}

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
  - ExAC
    - http://exac.broadinstitute.org/
  - gnomAD
    - http://gnomad.broadinstitute.org/

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%
- Impute missing genotypes using observed allele frequencies
  - For the entire sample
    - Not based on case or control status
- Analyze imputed data using dosages

National Heart Lung and Blood Institute Exome Sequencing Project (NHLBI-ESP)

NHLBI-ESP Study Design

Samples Selected from Large Studies

HeartGO	LungGO	WHISP
N>60,000	N>19,000	N>142,000
e.g. Jackson Heart Study, Framingham	e.g. Lung Health Study, Acute Lung Injury	i.e. Women's Health Initiative

Sequencing Performed at the Broad & University of Washington

Read Mapping and Alignment -BWA

Multi-Sample Variant Calling (glfMultiples –UMAKE)

European Americans (EA) N=4420

~6700 Exomes

African Americans (AA) N=2312

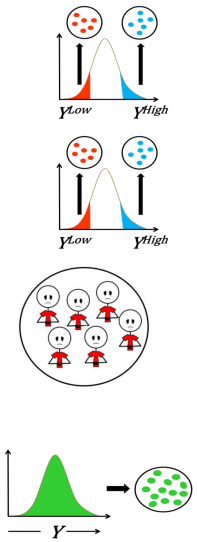
Selection for the 12 Primary Traits

Extreme quantitative trait values
- Low-density lipoprotein (N=657)
- Blood pressure (N=812)

Disease severity
- Asthma (N=190)
- Chronic obstructive pulmonary disease (N=623)

Disease endpoints
- Stroke (N=551)
- Early onset myocardial infarction (N=1007)

Deeply phenotyped individuals
- Randomly selected to be used as controls (N=964)



Extensive Secondary Phenotypic Data

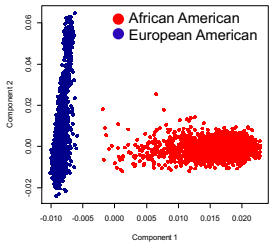
- C-reactive protein (N=3379)
- EKG measurements(N_{EKG-QT}= 3442)
- Fasting blood glucose (N=2470)
- Fibrinogen (N=2915)
- High-density lipoprotein (N=3770)
- Intima-media thickness (N=2079)
- Low-density lipoprotein (N=2685)
- Red blood cell count (N=1103)
- Systolic blood pressure (N=4423)
- Triglycerides (N=3728)
- Uric acid (N=2169)
- Waist-to-hip ratio (N=3853)
- White blood cell count (N=3792)
- von Willebrand factor (N=1587)

- 59 Secondary phenotypes*
  - 48 quantitative traits
  - 11 qualitative traits
- *Some traits are both primary and secondary
  - i.e. asthma, blood pressure, BMI, COPD, LDL, T2D

Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

Sex Check

Duplicates & Related Sample Check

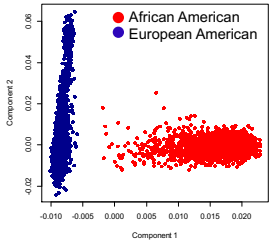
Designate EAs & AAs
Multidimensionality scaling (MDS)

Variant Site Removal
Deviation from HWE

Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

Sex Check

Duplicates & Related Sample Check

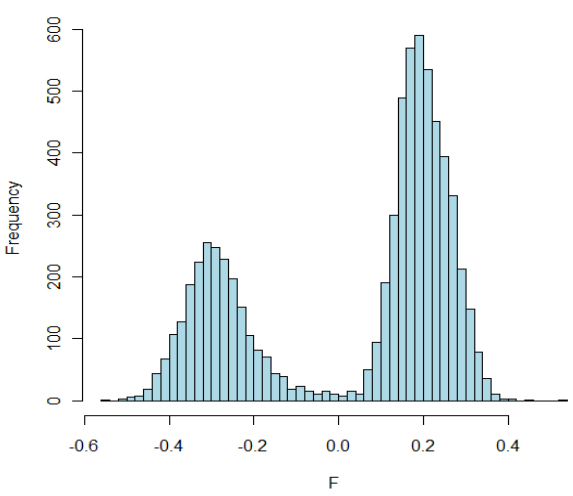
Designate EAs & AAs
Multidimensionality scaling (MDS)

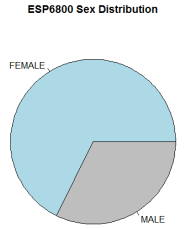
Variant Site Removal
Deviation from HWE

Support Vector Machine (SVM)

- A machine-learning algorithm, to separate likely true positive and false-positive variant sites.
- Uses VCF annotation related to quality of each SNV, including
  - Overall depth
  - Fraction of samples with coverage
  - Fraction of reference bases in heterozygous individuals (allele balance)
  - Inbreeding coefficient
  - In all 16 parameters were used
- Training set
  - False positives
    - SNVs that deviated significantly from expected values in three or more annotation categories
  - True positives
  - SNVs at HapMap polymorphic sites and Omni 2.5 array polymorphic sites in the 1000 Genomes project data
- The SVM classifier was used to identify all likely false positive sites
- Those variant sites which fail the support vector machine (SVM) (Likely false positive variant sites)
  - Are flagged and removed from further analysis

ESP6800 SEX CHECK





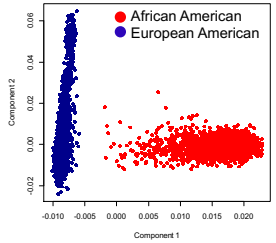
Sex	Report	Exome
Males	2608	2612
Females	4191	4211

*14 individuals inconsistent sex information

Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

Sex Check

Duplicates & Related Sample Check

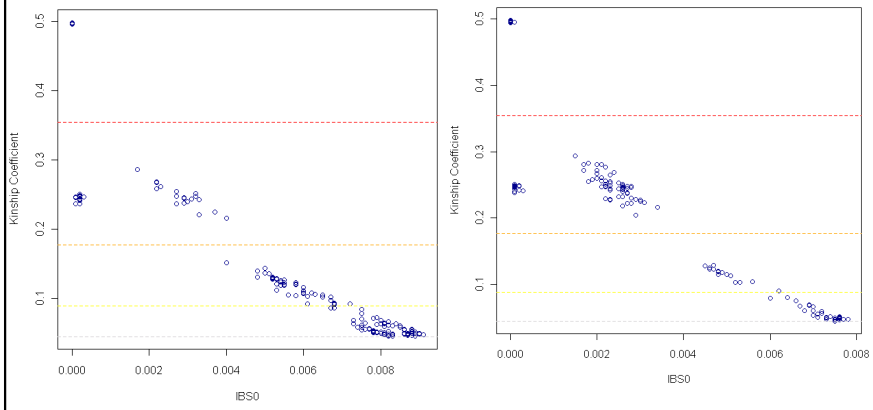
Designate EAs & AAs
Multidimensionality scaling (MDS)

Variant Site Removal
Deviation from HWE

KING Relatedness

African Americans

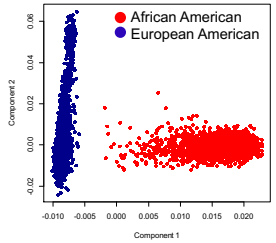
European Americans



Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

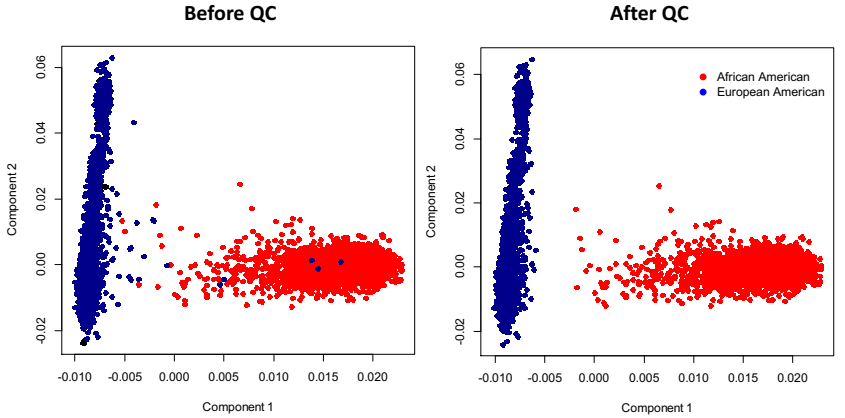
Sex Check

Duplicates & Related Sample Check

Designate EAs & AAs
Multidimensionality scaling (MDS)

Variant Site Removal
Deviation from HWE

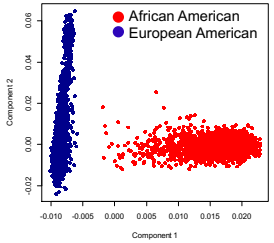
MDS Before and After Removal of Outliers



Data Quality Control & Association Analysis

Very different for rare variant sequence data than for common variants obtained from genotyping arrays

Analysis performed using Variant Association Tools
http://varianttools.sourceforge.net/VAT



Exome Data Quality Control

Variant Site Removal
Support Vector Machine

Variant Call Removal
Read Depth <10X

Variant Site Removal
>500X mean depth, Missing >10% genotypes

Sex Check

Duplicates & Related Sample Check

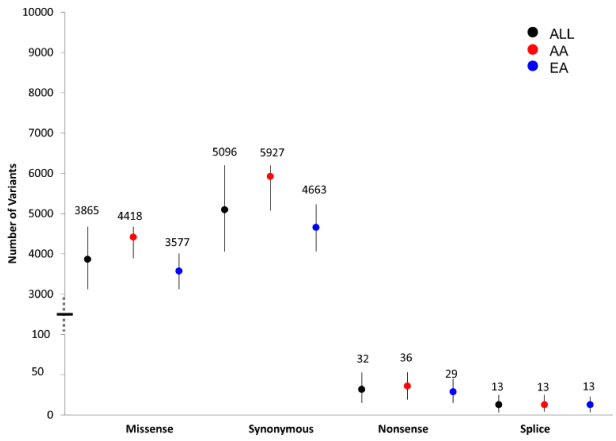
Designate EAs & AAs
Multidimensionality scaling (MDS)

Variant Site Removal
Deviation from HWE

Removal of Additional Variant Sites

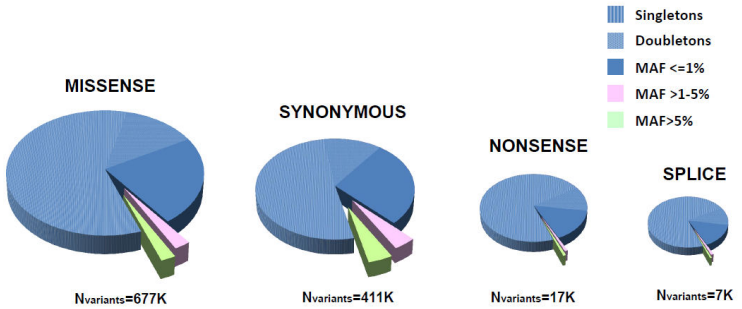
- Variant sites which deviate from HWE
  - Using a p-value $<1x10^{-7}$ criterion
    - Number of variant sites which deviate from HWE expectations:
      - EA: 2332
      - AA: 2663

Average Number of Variant Site Per Individual

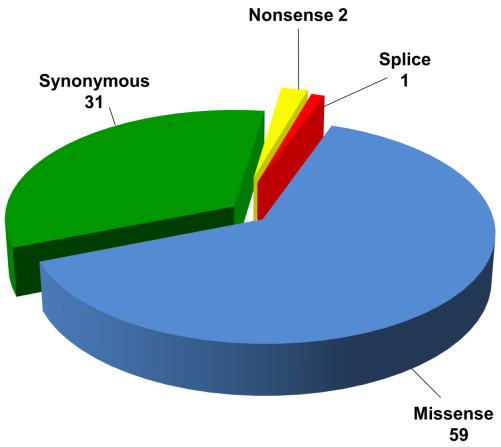


Intersect of 4 capture array targets - 16,206 genes

Most Variants are Rare



Average Number of Unique Variants per Individual



Analysis of Phenotypes and Exome Data

Extreme QTs
Dichotomized

Disease Traits
Case-control

QTs
Analyze QT values

Control for Population Substructure
Included population-specific C1 & C2

Selection of Covariates
Phenotype-specific model selection

Association Analysis

Single Variant Association Analysis
- All variant types included
- e.g. synonymous, missense, etc.

Rare Variant Aggregate Association Analysis
- CMC, SKAT (MAF ≤1%) and VT (MAF<5%)
- Variant types restricted within gene region
- i.e. missense, nonsense, spice site

Analysis of Phenotypes and Exome Data

Extreme QTs
Dichotomized

Disease Traits
Case-control

QTs
Analyze QT values

Control for Population Substructure
Included population-specific C1 & C2

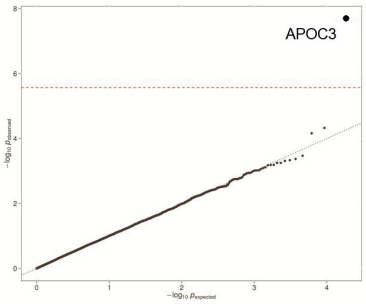
Selection of Covariates
Phenotype-specific model selection

Association Analysis

Single Variant Association Analysis
- All variant types included
- e.g. synonymous, missense, etc.

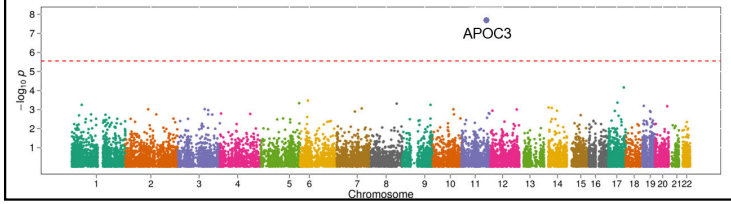
Rare Variant Aggregate Association Analysis
- CMC, SKAT (MAF ≤1%) and VT (MAF<5%)
- Variant types restricted within gene region
- i.e. missense, nonsense, spice site

Burden of Rare Variants *APOC3* Associated with Triglycerides Levels



APOC3 *	Beta	# Variant sites	P-value (CMC)
Meta Analysis			**3.0×10^{-8}**
EA (N=2074)	-0.598	7	3.9×10^{-6}
AA (N=1654)	-0.357	4	1.6×10^{-3}

*Rare variants reduce triglyceride levels
Triglyceride levels natural log transformed
Covariates: age, age^2, BMI & current smoker
3 variant sites overlap between EA & AA



Auer et al.
AJHG 2016

The Exome Chip

NHLBI-ESP the largest contributor of sequence data for the development of the exome chip

- ~240,000 missense, nonsense and splice site variants
- NHLBI-ESP findings are being followed up using the exome chip
- Novel findings are also being pursed
- More than 100,000 exome chips being genotyped and analyzed using samples from the ESP cohorts

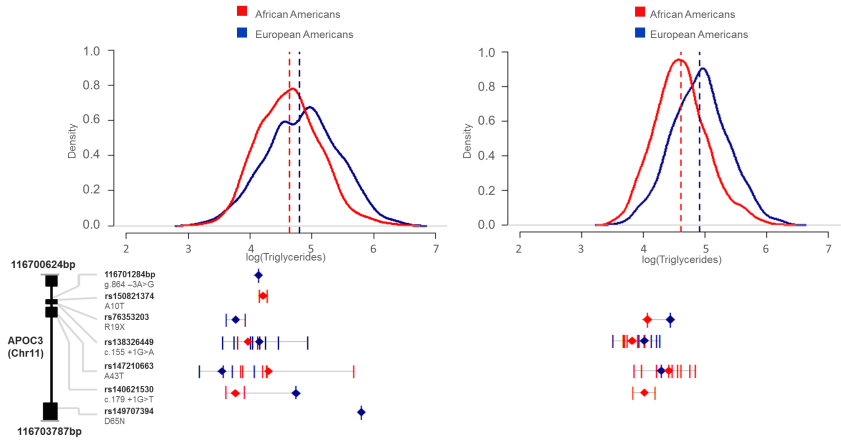
Replication with the Exome Chip
APOC3 Associated with Triglycerides Levels

APOC3 *	Sample Size	# Variant Sites	P-value
Meta Analysis	**8,069**		**1.7×10^{-18}**
Women's Health Initiative (WHI) Exome Chip			
Meta Analysis	4,341		9.4×10^{-12}
European Americans	2,301	3	1.3×10^{-6}
African Americans	2,041	4	1.6×10^{-6}
Exome Sequencing Project			
Meta Analysis	3,728		3.0×10^{-8}
European Americans	2,074	7	3.9×10^{-6}
African Americans	1,654	4	1.6×10^{-3}

*Reduces triglyceride levels
Triglyceride levels natural log transformed
Covariates: age, age^2, sex, BMI & current smoker

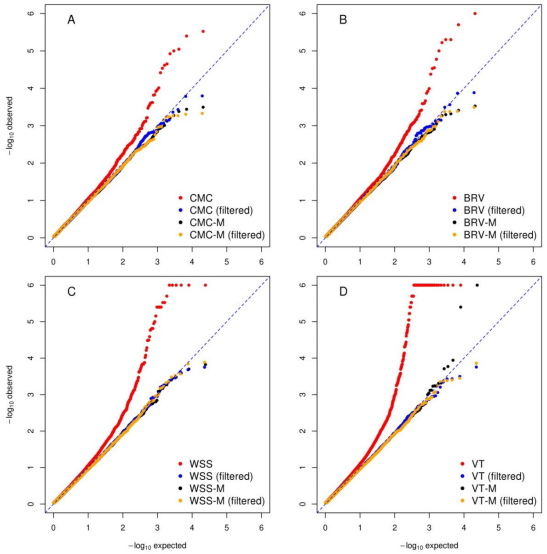
Triglyceride Levels for Carriers of APOC3 Rare Variants

Exome Sequencing Project | Women's Health Initiative



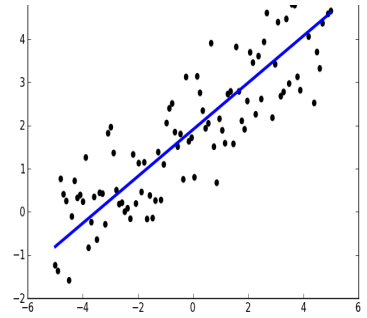
For additional information see Auer et al. 2016

Results



Rare Variant Aggregate Methods

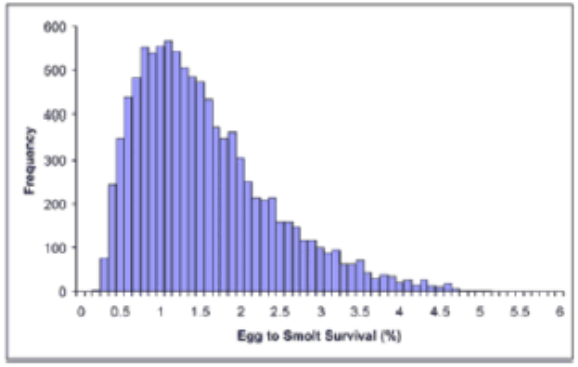
- Ideally should be performed in a regression framework
  - Logistic
  - Linear regression



- Almost all methods have been extended to be implemented within a regression framework

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

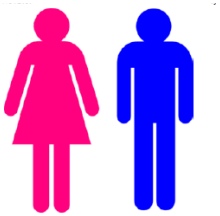


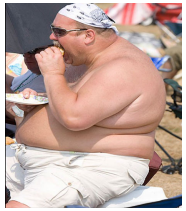
Analyzing Quantitative Variants

- 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
  - If their our confounders which need to be controlled
    - Residuals are generated were confounders have been adjusted

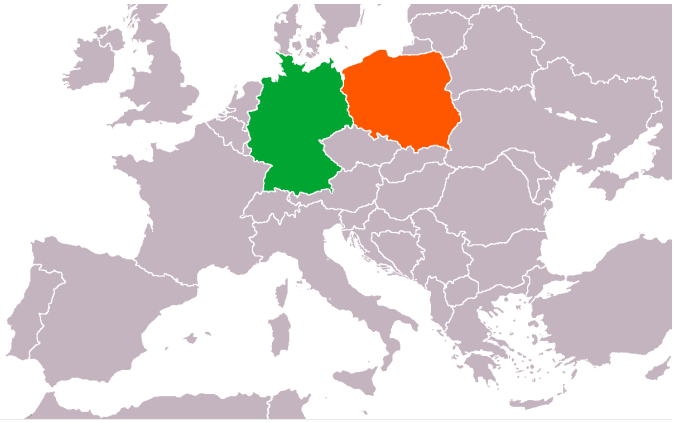
Rare Variant Aggregate Methods

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





Confounder -Population Substructure and Admixture



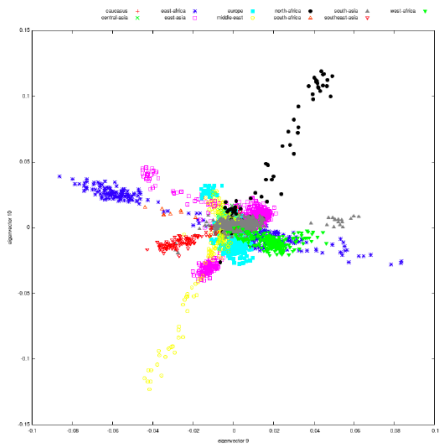
Rare Variant Aggregate Methods

- If proportion of cases and controls sampled from each populations 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

Example Pima Indians

Rare Variant Aggregate Methods

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



Rare Variant Aggregate Methods

- Best to obtain components to include in the regression model
  - using variants which are not in LD e.g. r^2<0.5 (prunned)
  - covering a wide range of the allelic frequency spectrum e.g. >0.1%
- Success of PCA\MDS in controlling for population substructure\admixture can be evaluated through lambda and examining Quantile-Quantile (QQ) plots

Linear Mixed Models and Generalized linear Mixed Models

- Linear mixed models and their extension Generalized linear mixed models for binary traits
  - Can offer better control of type I error than linear and logistic regression
    - When observations are not independent
      - Related or cryptically related individuals are included in the sample
      - Population structure
- Models both fixed and random effects

Genome-wide association studies (GWAS) - Part 2

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

Heather J. Cordell

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

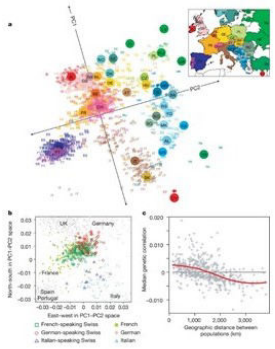
heather.cordell@ncl.ac.uk

Linear Mixed Models (LMMs)

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

Population stratification and relatedness

Genes mirror geography within Europe



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

Linear Mixed Models (LMMs)

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

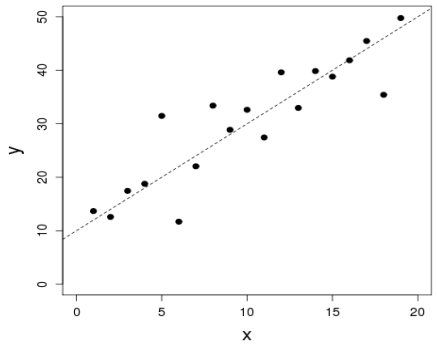
$$y = mx + c \qquad \text{or} \qquad y = \beta_0 + \beta_1 x$$

- This model may also be written

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

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

Linear Regression



Linear Mixed Models (LMMs)

- In linear regression we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Here β_0 and β_1 are fixed effects while ϵ_i is a random error
- In matrix notation we can write this model:

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

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

Linear Mixed Models (LMMs)

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

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

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

LMMs in genetics

- In genetics we generally work with two equivalent forms of LMM
- One is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$
  - The random effect u_l corresponds to a scaled additive effect of causal variant (locus) l
    - Assuming many (m) such causal variants all across the genome
  - The random effects u_l all have variance σ_u^2 and are uncorrelated with each other
    - So $\mathbf{u} = (u_1, u_2, \ldots u_m) \quad \sim N(0, \mathbf{I}\sigma_u^2)$
  - $\mathbf{Z}$ is a standardized genotype matrix i.e. z_{il} takes value

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

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

LMMs in genetics

- The other form is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon}$, or

$$y_i = \sum_k \beta_k x_{ik} + g_i + \epsilon_i$$

  - Where the β_k are k fixed effects (e.g. covariates or SNPs to be tested for association), and random effect $g_i = \sum_{l=1}^{m} z_{il} u_l$ is the total genetic effect in individual i, summed over all the causal loci
- Here, g_i is considered as a random effect operating in individual i
  - The vector of random effects $\mathbf{g}$ takes distribution $\mathbf{g} \sim N(0, \mathbf{G}\sigma_a^2)$
    - Where $\mathbf{G}$=$\mathbf{ZZ'}/m$ is the genetic relationship matrix (GRM) between individuals at the causal loci
    - $\sigma_a^2 = m\sigma_u^2$ is the total additive genetic variance
- For family data (close relatives), the expected values of the elements of $\mathbf{G}$ are equal to twice the kinship coefficients Φ_{ij} i.e. $\mathbf{G}$ is equal to twice the kinship matrix $\boldsymbol{\Phi}$
  - Models their relatedness at the causal loci (and elsewhere)

Use of LMMs in genetics

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

Testing for association using LMMs

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

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

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

Testing for association using LMMs

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

Estimating the genetic relationship matrix

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

Estimating the genetic relationship matrix

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

Software implementations

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

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

- Model can also be extended to bivariate traits (Korte et al. 2012, Nat Genet 44:1066-1071), implemented in MTMM/ASREML and DISSECT

Binary traits

- For binary traits, coding cases and controls as a 1/0 quantitative trait is not optimal
  - Though in practice it seems to work reasonably well
- LTMLM (Hayeck et al. 2015) and LEAP (Weissbrod et al. 2015) instead use an underlying *liability model* to improve power
  - Assuming known disease prevalence
- Chen et al. (2016) [AJHG 98:653-66] showed that high levels of population stratification can invalidate the analysis, when applied to a case/control sample
  - Resulting in a mixture of **inflated** and **deflated** test statistics
  - Developed GMMAT software to address this problem
  - CARAT software (Jiang et al. 2016, AJHG 98:243-55) also appears to address this problem effectively
- SAIGE software (Zhou et al. 2018, AJHG 50(9):1335-1341) implements a mixed model association test that deals with large case-control imbalance

Elucidating genetic architecture

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

Elucidating genetic architecture

- Basic idea is to use formulation

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

with $\mathbf{g} \sim N(0, \mathbf{A}\sigma^2_a)$ and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma^2_e)$ so $\mathbf{V} = \mathbf{A}\sigma^2_a + \mathbf{I}\sigma^2_e$

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

Partitioning variance

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

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

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

Other approaches

- Some recent work has focussed on estimating (a) heritability explained by sets of SNPs, and (b) genetic correlations across traits, using summary statistics only
  - Bulik-Sullivan et al. (2015) [Nat Genet 47:291-295]
  - Bulik-Sullivan et al. (2015) [Nat Genet 47:1236-1241]
    - Clever idea that allows the variance component parameters to be estimated via a simple regression on 'LD Scores'

Gene-gene (and gene-environment) interactions

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

Definition of (pairwise) interaction

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

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

Marginal effect of factor 1

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

Marginal effect of factor 2

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

Main effects of factors 1 and 2

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

Main effects and interaction term

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

Interaction

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

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

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

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

  - Trait value only differs from baseline if both factors present

Gene-gene interaction (epistasis)

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

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

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

Gene-gene interaction

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

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

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

	Locus H		
Locus G	2	1	0
2	$\beta_0 + 2\beta_G + 2\beta_H + 4\beta_{GH}$	$\beta_0 + 2\beta_G + \beta_H + 2\beta_{GH}$	$\beta_0 + 2\beta_G$
1	$\beta_0 + \beta_G + 2\beta_H + 2\beta_{GH}$	$\beta_0 + \beta_G + \beta_H + \beta_{GH}$	$\beta_0 + \beta_G$
0	$\beta_0 + 2\beta_H$	$\beta_0 + \beta_H$	β_0

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way
  - Good starting point for further investigation of their (joint) action

Gene-environment (G×E) interactions

- The same regression model

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

can be used to model interaction between a genetic factor G and an environmental factor H

  - With the environmental variable x_2 coded in binary fashion (e.g. smoking) or quantitatively (e.g. age)
- Focus of analysis is often risk estimation
  - Estimating genetic risks in particular environments
  - Estimating effect of environmental factor on particular genetic background
    - Important for treatment/screening strategies and public health interventions
- For G×G, focus of interest is more related to
  - Increasing power to detect an effect (by taking into account the effects of other genetic loci)
  - Modelling the biology, especially related to the joint action of the loci

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

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

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone
- Depending on circumstances, any of these tests may be a sensible option
- Most tests of interaction/joint action can be thought of as a version of one or other of these tests
  - Although different tests vary in their precise details
  - And their relationship to the logistic regression formulation not always clearly described

G×G versus G×E in the context of GWAS

- Typically GWAS measure thousands if not millions of genetic variants
  - But only a few (tens or at most 100s) of environmental factors
- Feasible to consider all G×E combinations
- All pairwise G×G combinations possible, but much more time consuming
  - And leads to greater multiplicity of tests
  - Also, why stop at 2-way interactions?
    - Could look at all 3 way, 4 way etc. combinations
    - Scale of problem quickly gets out of hand
    - Less obvious reason to do this for G×E...

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"
- Or have proposed filtering based on single-locus significance or other (biological or statistical) considerations
  - Reduces multiple testing burden, improves interpretability
- Or have proposed testing at the gene level rather than the SNP level
  - Ma et al. (2013) PLoS Genet 9(2): e1003321
    - Compared 4 different tests that combine P values from pairwise (SNP x SNP) interaction tests
    - Showed that the truncated tests did best
    - Presented an application only considering gene pairs known to exhibit protein-protein interactions

Case-only analysis

- Piergorsh et al. 1994; Yang et al. 1999; Weinberg and Umbach 2000
- Several authors have shown that, for binary predictor variables, a test of the interaction term β_{12} in the logistic regresssion model

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

  can be obtained by testing for correlation (association) between the genotypes at two separate loci, within the sample of cases
- Gains power from making assumption that genotypes (alleles) at the two loci are uncorrelated in the population
  - So only really suitable for unlinked or loosely linked loci (since closely linked loci are likely to be in LD)
- Alternatively contrast the genotype correlations in cases with those seen in controls (`--fast-epistasis` in PLINK)

Testing correlation between loci

- A similar idea is implemented in EPIBLASTER (Kam-Thong et al. 2011; EJHG 19:465-571)
- Wu et al. (2010) (PLoS Genet 6:e1001131) also proposed a similar approach – though needs adjustment to give correct type I error rates
- See also Joint Effects (JE) statistics (Ueki and Cordell 2012; PLoS Genetics 8(4):e1002625)
- All these methods test whether correlation exists (case-only) or is different in cases and controls (case/control), via testing a log OR for association between two loci
  - However, the log OR for association (λ) encapsulates a slightly different quantity between the different methods
- All implemented (along with standard logistic and linear regression) in CASSI
  - http://www.staff.ncl.ac.uk/richard.howey/cassi/

Empirical evidence for G×G interactions

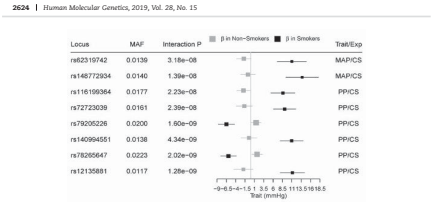
- Epistasis among *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* in multiple sclerosis (Lincoln et al. 2009 PNAS 106:7542-7547)
- *HLA-C* and *ERAP1* in psoriasis (Strange et al. 2010)
- *HLA-B27* and *ERAP1* in ankylosing spondylitis (Evans et al. 2011)
- *BANK1* and *BLK* in SLE (Castillejo-Lopez et al. 2012)
- Gusareva et al. (2014) found a reasonably convincing (partially replicating) interaction between SNPs on chromosome 6 (*KHDRBS2*) and 13 (*CRYL1*) in Alzheimer's disease
- Dai et al. (2016) [AJHG 99:352-365] identified 3 loci simultaneously interacting with established risk factors gastresophageal reflux, obesity and tobacco smoking, with respect to risk for Barrett's esophagus

Empirical evidence for G×G interactions

- Hemani et al. 2014 (Nature 508:249-253) found 501 instances of epistatic effects on gene expression, of which 30 could be replicated in two independent samples
  - Many SNPs are close together, could represent haplotype effects?
  - Or the effect of a single untyped variant?
  - See caveats in
    - Wood et al. (2014) Nature 514(7520):E3-5. PMID:25279928
    - Fish et al. (2016) Am J Hum Genet 99(4):817830. PMID:27640306

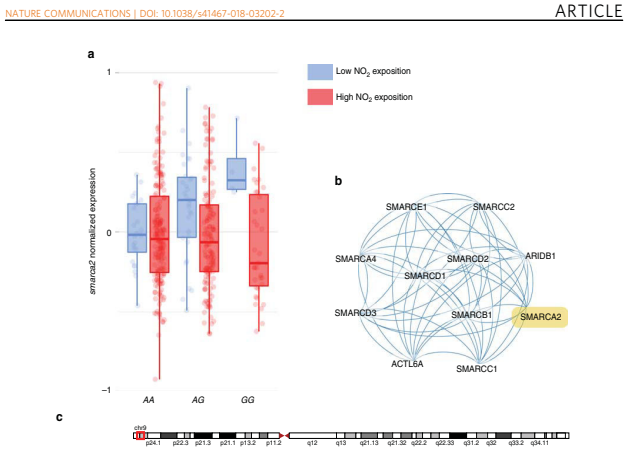
Empirical evidence for G×E interactions

- Myers et al. (2014) Hum Mol Genet 23(19): 5251-9 "Genome-wide Interaction Studies Reveal Sex-Specific Asthma Risk Alleles"
- Small et al. (2018) Nat Genet 50(4): 572-580 "Regulatory Variants at KLF14 Influence Type 2 Diabetes Risk via a Female-Specific Effect on Adipocyte Size and Body Composition"
- Sung et al. (2019) Hum Molec Genet 28(15): 2615-2633 "A multi-ancestry genome-wide study incorporating gene-smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure."



Empirical evidence for G×E interactions

- Favé et al. (2018) Nat Commun 9(1): 827 "Gene-by-environment Interactions in Urban Populations Modulate Risk Phenotypes"



Power Analysis for Single and Rare Variant Aggregate Association Analyses

Why Estimate Sample Sizes and/or Power?

- Not wasting your time and money
  - Carrying out a study for which you will never find a true association due to inadequate sample sizes
- Almost always necessary for grant proposals
  - Usually will be denied funding if cannot demonstrate planned study has adequate power

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

Power and Sample Size Estimation for Case-Control Data

- GWAS of common variants where each variant is test separately
  - $\alpha = 5 \times 10^{-8}$ (Bonferroni Correction for testing 1,000,000 variant sites)
  - Shown to be a good approximation for the effective number of tests
    - Valid even when more than 1,000,000 variant sites tested
  - Effective number of tests is dependent of the LD structure
- Analysis of individual variants for whole genome sequence data
  - More rare variants than common variants
    - Also have lower levels of LD than between common variants
  - The number of effective tests is higher than for analysis limited to common variants
  - α yet to be determined

Determining Genome-wide Significance Levels

- 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 the genotypes at increasing densities and using permutation to estimate the nominal p-value for 5% family-wise error
- Then extrapolating to infinite density
- The genome wide significance threshold was estimated to be ~$7.2X10^{-8}$
- Estimate is based on LD structure for Europeans
  - Not sufficiently stringent for populations of African Ancestry

Power and Sample Size for Aggregate Rare Variant Tests

- For gene based methods a Bonferroni correction for the number of genes/regions tested is used
  - e.g. 20,000 genes significance level $\alpha = 2.5 \times 10^{-6}$
    - Can use a less stringent criteria
      - Not all genes have two or more variants
        - » Divide 0.05 by number of genes tested
    - If units other than genes used may have to use a more stringent
- Little LD between variants in separate genes
  - Little to no correlation between tests
    - Bonferroni correction is not overly stringent

Power and Sample Size 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 is hypothesized that 20 genes and 80 independent variants will be brought to stage II
  - A Bonferroni correct can be made for performing 100 tests
    - An $\alpha = 5.0 \times 10^{-3}$ cab be used for a family wise error rate of 0.05

Estimating Power/Sample Sizes For Individual Variants

- Can be obtained analytically
- Information necessary
  - Prevalence
  - Risk allele frequency
  - Effect size (odds ratio-for case control data)
  - Genetic model for the susceptibility variant
    - Recessive ($\gamma_1=1$)
    - Dominant ($\gamma_2=\gamma_1$)
    - Additive ($\gamma_2=2\gamma_1-1$)
    - Multiplicative ($\gamma_2=\gamma_1^2$)

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 are also used
  - Dominant
  - Additive
  - Multiplicative

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
        - » $\gamma_1=1$

Gamma is the Relative Risk

- Many programs work with the relative risk (ɣ)
- Relative risk only approximates odds ratio when disease is rare
  - Not appropriate for common trait
- Example risk variant and marker allele frequency 0.01
  - D' and $r^2=1$

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

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

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

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

submit Reset

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.

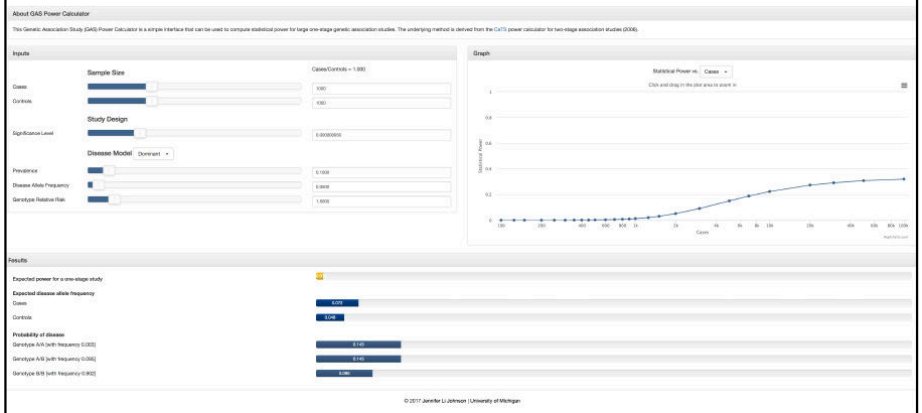
Adapted from:
Slager SL, Schaid DJ: Case-control studies of genetic markers:
Power and sample size approximations for Armitage's test for trend.
Hum Hered 52, 149-153 (2001).
and
Freidlin B, Zheng G, Li Z, Gastwirth JL:
Trend tests for case-control studies of genetic markers:
Power, sample size and robustness.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

Tim M. Strom

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 results

GAS Power Calculator



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 relative risk
  - Displays odds ratios

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)
```

Process Reset

Created by Shaun Purcell 24.Oct.2008

PAWE

- Power Association With Errors
  - Will give same results for case-control studies of discrete traits as Genetic Power Calculator when calculations are done without errors
- 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

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

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

Power Analysis for Rare Variant Aggregate Association Tests

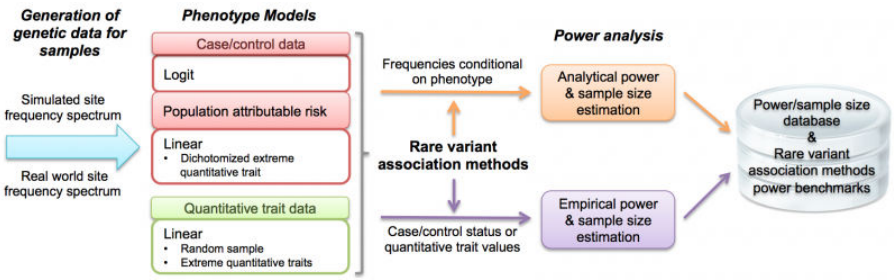
- Many unknown parameters must be modeled
  - Allelic architecture within a genetic region
    - Varied across genes and populations
  - Effects of variants within a region
    - Fixed or varied effect sizes of causal variants
    - Bidirectional effect of variants
    - Proportion of non-causal variants
- Power usually must be estimated empirically
- Simplified assumptions can be made to obtain analytical estimates
  - All variants have the same effect size
  - No non-causal variants

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

SEQPower

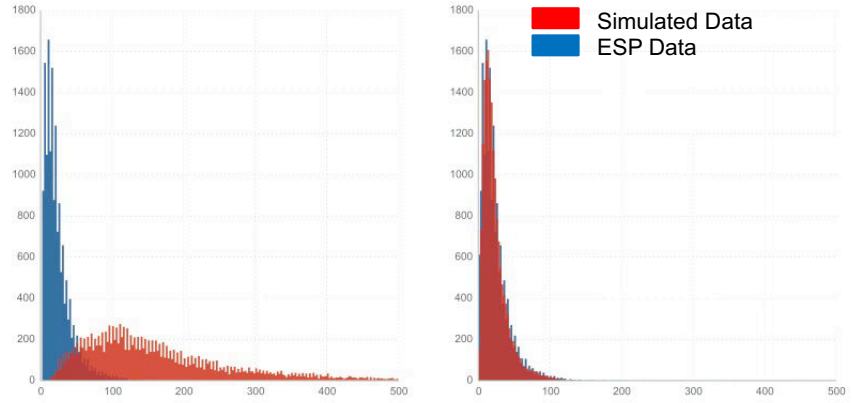
http://www.bioinformatics.org/spower/



Wang et al. 2014 Bioinformatics

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

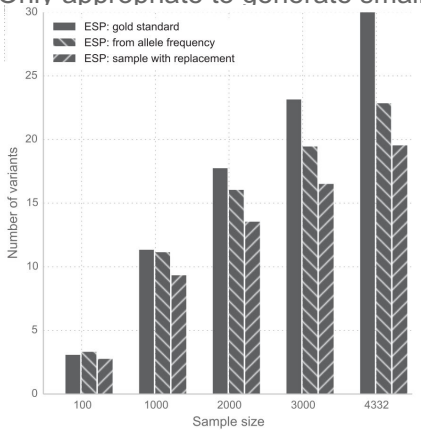
Distribution of number of variants per gene



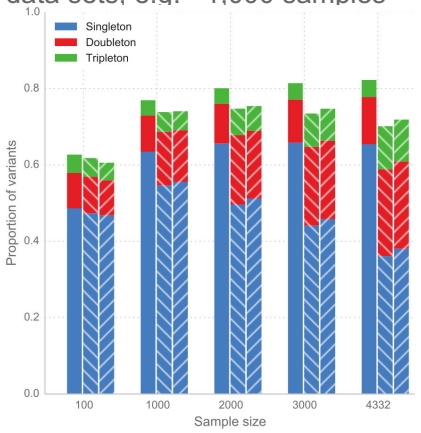
- Simulated variant counts based on the entire simulated population
- Simulated variant counts based on haplotype pool down-sampled to ESP size

Simulating Data Using Sequence Data (ESP)*

*Only appropriate to generate small data sets, e.g. <1,000 samples



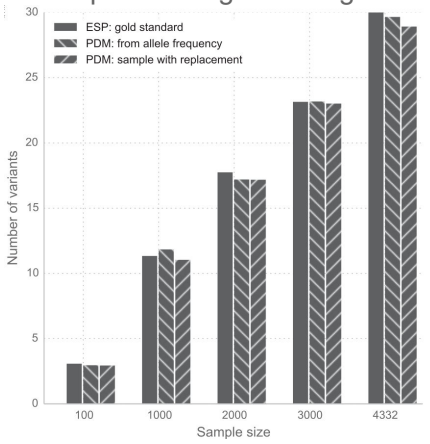
Number of Variant Sites



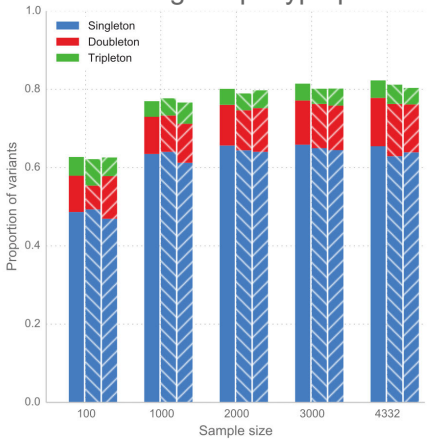
Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

Simulating Data: Using Population Demographic Models (PDM)*

*Resample or using MAF to generate data from large haplotype pools



Number of Variant Sites



Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

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 3 or more variant sites
- Determine power as the proportion of genes that meet exome-wide significance (alpha=2.5×10^{-6})

Power Analysis

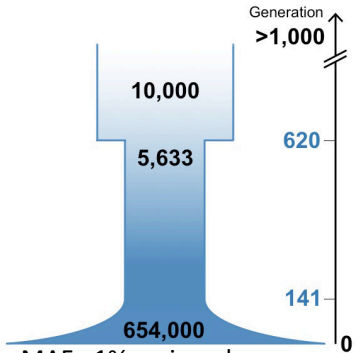
- For tests of individual variants
  - Power depended on sample size, disease prevalence, minor allele frequency, genetic model and variant effect size
- For rare variants (aggregate association tests)
  - Also dependent on the allelic architecture
    - Cumulative variant frequency within analyzed region
    - Proportion of causal variants
      - How much contamination by 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
        - » Increase and decrease quantitative trait values

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 some genes even with hundreds of thousands of samples power will still be low, while for others a few thousand samples may be sufficient

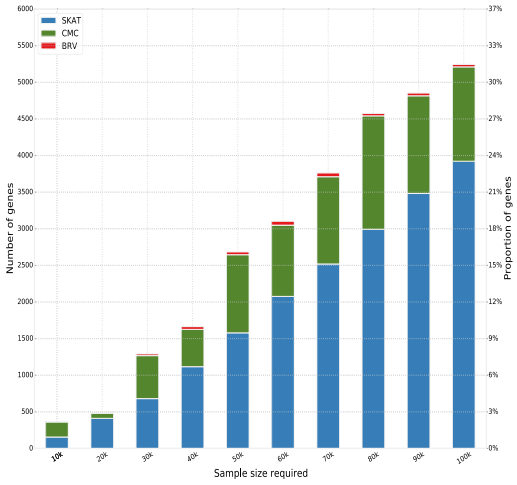
How Large of a Sample Size is Necessary to Detect Rare Variant Associations?

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



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

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

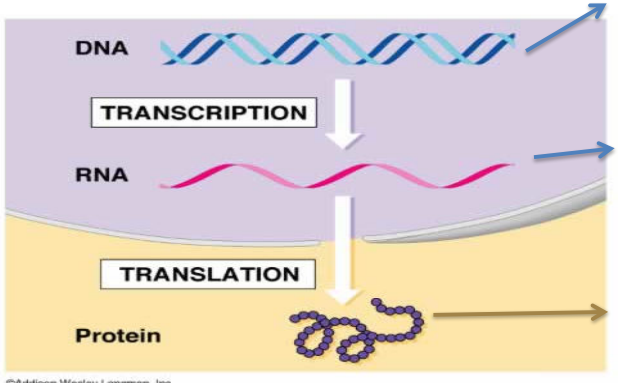


Integrative Approaches in Biobanks : Getting to Biological Mechanisms of Disease

Nancy J. Cox, PhD
Vanderbilt Genetics Institute
Vanderbilt University Medical Center

But First Some Review

Brief Review

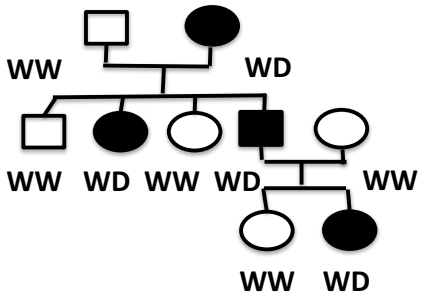


Full blueprint stored in architect's office

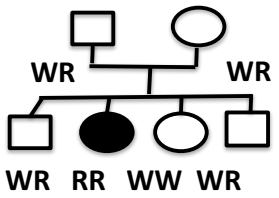
Working copy brought to construction zone

Fully assembled building

Rare Disease Transmitted in Families



Dominant



Recessive

Finding the Genetic Cause of a Rare Mendelian Disease

I am a cool dude.

Finding the Genetic Cause of a Rare Mendelian Disease

I am a coal dude.

Geneticists call this a “missense” mutation; still makes a protein, but the protein isn’t making as much sense.

Finding the Genetic Cause of a Rare Mendelian Disease

I am a cool rude.

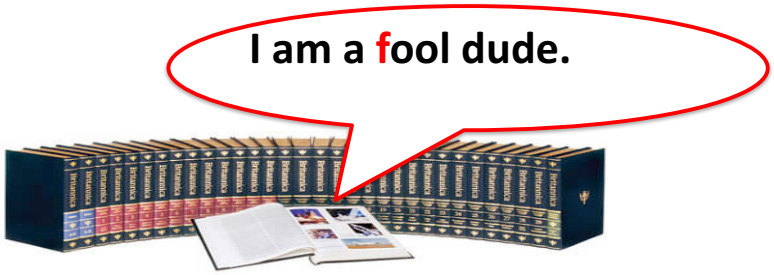
This makes even less sense; it is a more deleterious missense mutation.

Finding the Genetic Cause of a Rare Mendelian Disease

I am a cmol dude.

This is closer to a “nonsense” mutation; mutation creates a protein that can’t go to completion.

Finding the Genetic Cause of a Rare Mendelian Disease

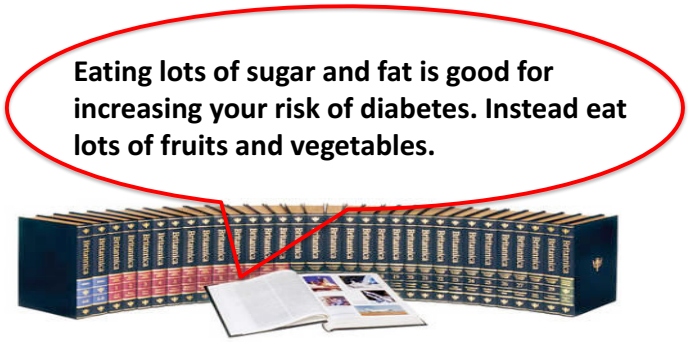


This might be considered a “gain-of-function” mutation. The protein is active – but is doing something entirely new.

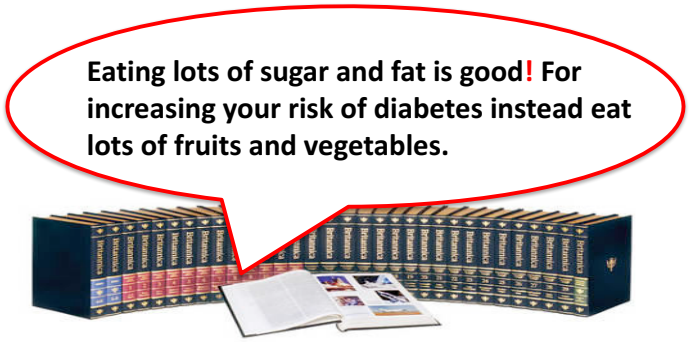
Common Diseases

- **Are familial, in that the risk of disease in a relative of an affected person is higher than the risk in the general population, but the transmission is not simple**
- **Many genetic and non-genetic risk factors contribute to common diseases**
- **Any disease that sends people to doctors or hospitals may have a genetic component**

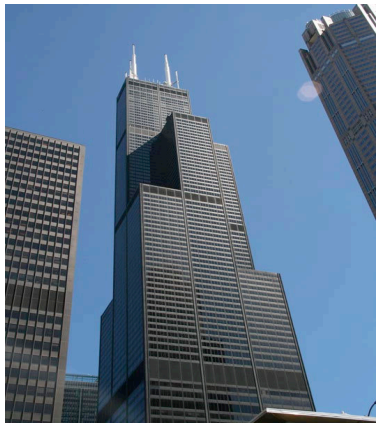
Finding Genetic Factors Contributing to a Common Disease…



Finding Genetic Factors Contributing to a Common Disease…



Genomic Discoveries for Rare Mendelian Disease



Genomic Discoveries for Common Disease

Genomic Discoveries for Common Disease

G C A C G G T T G T T

T C C C A G C T A G C

G G T T C G T A T C T

G C A C G G T T G T T

T T A A A C T T C C C

C C C T G G G C G T T

Picking Mendelian Cherries

Relating Variation to Phenotype

Genotyping Sequencing

Relating Variation to Phenotype

+

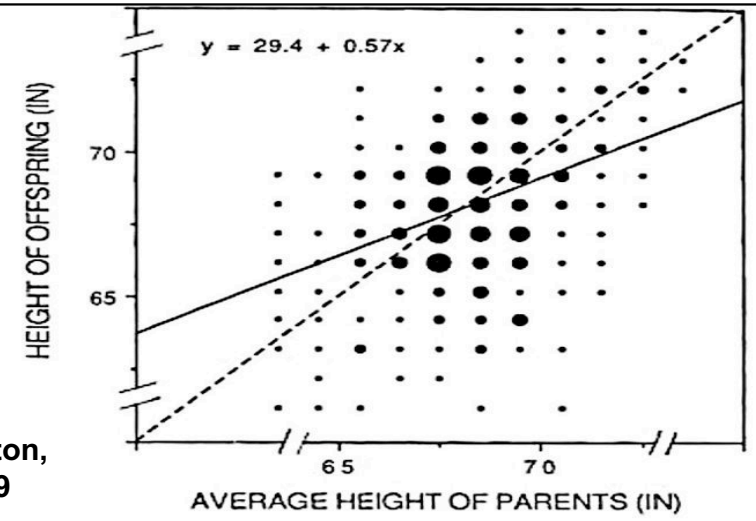
=

Common Variants

Rare Variants

Genome Interrogation

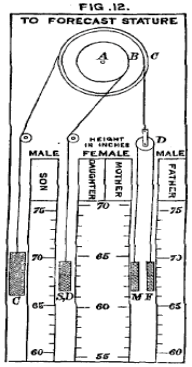
Galton, 1889



Galton, 1889

contrived more than one form of apparatus by which the probable stature of the children of known parents can be mechanically reckoned. Fig. 12 is a representation of one of them, that is worked with pulleys and weights. A, B, and C are three thin wheels with grooves round their edges. They are screwed together so as to form a single piece that turns easily on its axis. The weights M and F are attached to either end of a thread that passes over the movable pulley D. The pulley itself hangs from a thread which is wrapped two or three times round the groove of B and is then secured to the wheel. The weight SD hangs from a thread that is wrapped two or three times round the groove of A, and is then secured to the wheel. The diameter of A is to that of B as 2 to 3. Lastly, a thread is wrapped in the opposite direction round the wheel C, which may have any convenient diameter, and is

FIG. 12.
TO FORECAST STATURE



Estimating Heritability from SNPs

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]

ARTICLE

Estimating Missing Heritability for Disease from Genome-wide Association Studies

Sang Hong Lee,[1] Naomi R. Wray,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1,*]

Table 2 Comparison of results of different polygenic methods across diseases

			Caused by common GWAS SNPs		
				Polygenic modeling and Bayesian inference	
Disease	Prevalence (%)	Family based heritability[a]	LMM-based heritability (s.e.)	Total variance explained (50% CI)	*N* SNPs (50% CI)
Rheumatoid arthritis	1	0.53–0.68 (–0.13 MHC)[b]	0.32 (0.037)	0.18 (0.15–0.20) (+0.04 known non-MHC)[b]	2,231 (1,588–2,740)
Celiac disease	1	0.5–0.87 (–0.35 MHC)[b]	0.33 (0.042)	0.44 (0.40–0.47)	2,550 (1,907–3,061)
MI/CAD	6	0.3–0.63	0.41 (0.067)	0.48 (0.43–0.54)	1,766 (1,215–2,125)
T2D mellitus	8	0.26–0.69	0.51 (0.065)	0.49 (0.46–0.53)	2,919 (2,335–3,442)

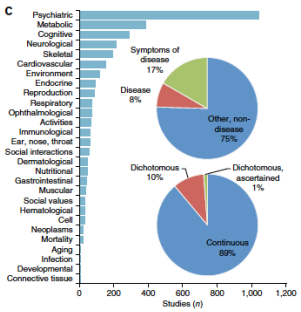
[a]Family based heritability estimates were taken from previous data for rheumatoid arthritis[27,28], celiac disease[18,30], MI/CAD[31,32] and T2D[33,34]. [b]We excluded some loci in certain analyses: although the family based heritability estimates are based on the whole genome, the extended MHC region was removed from the common GWAS SNP analyses for rheumatoid arthritis and celiac disease, and validated non-MHC loci were further removed from the polygenic modeling analysis of the rheumatoid arthritis GWAS data. 50% CI, 50% credible interval; s.e., standard error.

Stahl et al, Nat Gen

Meta-analysis of the heritability of human traits based on fifty years of twin studies

VOLUME 47 | NUMBER 7 | JULY 2015 NATURE GENETICS

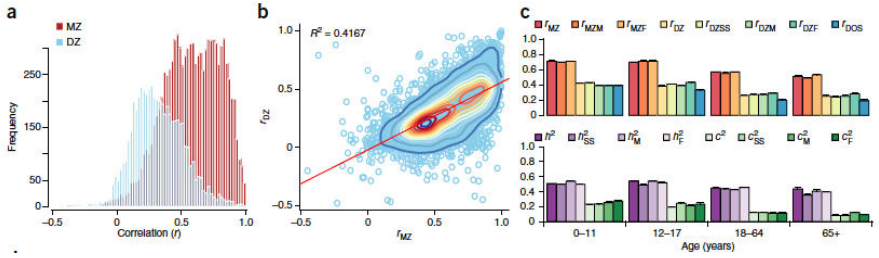
Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]



Meta-analysis of the heritability of human traits based on fifty years of twin studies

VOLUME 47 | NUMBER 7 | JULY 2015 NATURE GENETICS

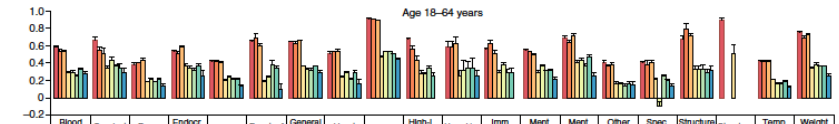
Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]



Meta-analysis of the heritability of human traits based on fifty years of twin studies

VOLUME 47 | NUMBER 7 | JULY 2015 NATURE GENETICS

Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]

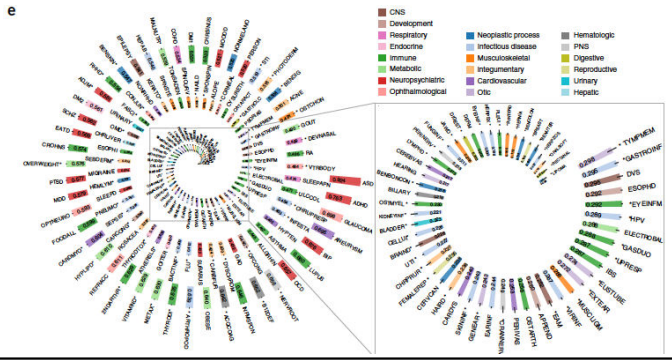


	Blood pressure funct.	Conduct dis.	Depr. episode	Endocr. gland funct.	Food	Funct. of brain	General metab. funct.	Heart funct.	Height	High-L. cognitive funct.	Hyperkin-etic dis.	Imm. system funct.	Ment. beh. dis. alc.	Ment. beh. dis. tob.	Other anxiety dis.	Spec. personal. dis.	Structure of the eyeball	Structure of mouth	Temp. pers. funct.	Weight maint. funct.
r_{MZ}	0.59	0.67	0.39	0.53	0.42	0.65	0.65	0.52	0.92	0.68	0.58	0.56	0.55	0.69	0.41	0.41	0.68	0.89	0.42	0.76
r_{MZM}	0.54	0.55	0.40	0.52	0.42	0.70	0.63	0.53	0.91	0.57	0.58	0.62	0.53	0.64	0.36	0.39	0.79		0.42	0.70
r_{MZF}	0.53	0.52	0.44	0.58	0.41	0.60	0.66	0.54	0.89	0.44	0.62	0.52	0.49	0.72	0.39	0.40	0.71		0.42	0.73
r_{DZ}	0.29	0.34	0.18	0.37	0.20	0.19	0.36	0.24	0.47	0.28	0.26	0.30	0.30	0.41	0.17	0.22	0.33	0.52	0.21	0.34
r_{DZSS}	0.30	0.43	0.21	0.34	0.24	0.24	0.33	0.29	0.53	0.27	0.31	0.38	0.36	0.44	0.16	−0.09	0.33		0.17	0.39
r_{DZM}	0.25	0.36	0.18	0.32	0.22	0.39	0.32	0.21	0.53	0.34	0.35	0.29	0.31	0.37	0.14	0.25	0.33		0.17	0.36
r_{DZF}	0.33	0.34	0.22	0.37	0.22	0.35	0.36	0.30	0.51	0.25	0.34	0.30	0.31	0.47	0.16	0.20	0.30		0.19	0.36
r_{DOS}	0.28	0.29	0.14	0.26	0.14	0.10	0.30	0.16	0.45		0.25		0.22	0.26	0.15	0.14	0.32		0.13	0.25

Classification of common human diseases derived from shared genetic and environmental determinants

VOLUME 49 | NUMBER 9 | SEPTEMBER 2017 NATURE GENETICS

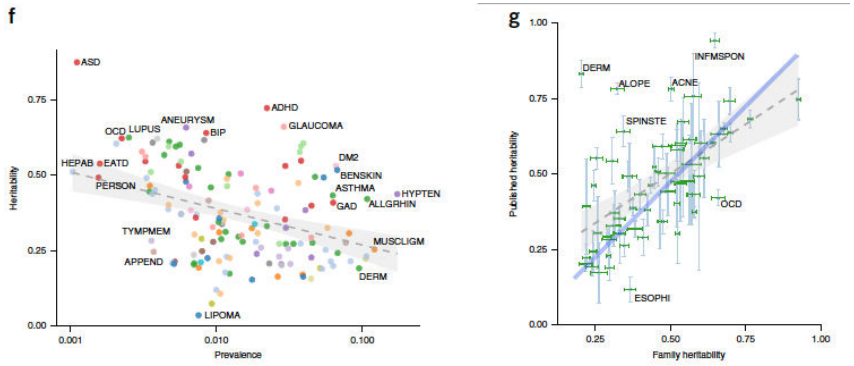
Kanix Wang[1,2], Hallie Gaitsch[2], Hoifung Poon[3], Nancy J Cox[4] & Andrey Rzhetsky[2,5]



Classification of common human diseases derived from shared genetic and environmental determinants

VOLUME 49 | NUMBER 9 | SEPTEMBER 2017 NATURE GENETICS

Kanix Wang[1,2], Hallie Gaitsch[2], Hoifung Poon[3], Nancy J Cox[4] & Andrey Rzhetsky[2,5]



Classification of common human diseases derived from shared genetic and environmental determinants

VOLUME 49 | NUMBER 9 | SEPTEMBER 2017 NATURE GENETICS

Kanix Wang[1,2], Hallie Gaitsch[2], Hoifung Poon[3], Nancy J Cox[4] & Andrey Rzhetsky[2,5]

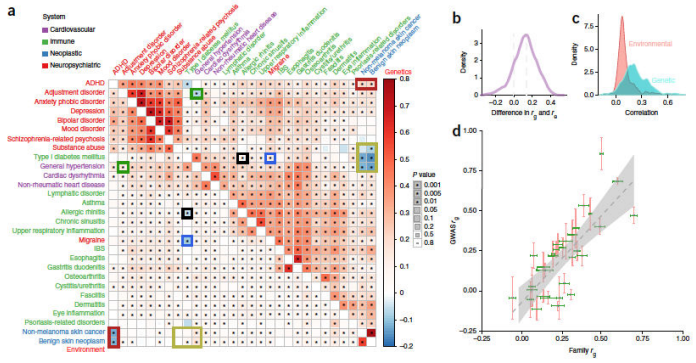
Table 1 Disease prevalence and heritability estimates for the 30 most prevalent diseases in our study

Disease	Prevalence	h^2	h^2 s.d.
Cardiac dysrhythmia	0.045	0.240	0.011
General hypertension	0.173	0.462	0.009
Esophageal disease	0.077	0.292	0.008
Functional digestive disorder	0.051	0.203	0.009
Type II diabetes mellitus	0.066	0.561	0.010
Allergic rhinitis	0.108	0.445	0.006
Asthma	0.063	0.457	0.008
Atopic contact dermatitis	0.095	0.202	0.006
Chronic sinusitis	0.047	0.523	0.008
Eye inflammation	0.045	0.292	0.009
Osteoarthritis	0.068	0.256	0.012
Cellulitis	0.061	0.226	0.007
Ear infection	0.106	0.244	0.007
Eye infection	0.053	0.200	0.009
Fungal infection	0.063	0.211	0.007
UTI	0.083	0.227	0.007
Viral warts HPV	0.038	0.289	0.009
Acne	0.036	0.501	0.010
Keratosis	0.058	0.344	0.015
General spondylosis spine disorder	0.081	0.325	0.008
Muscle ligament disorder	0.121	0.268	0.006
Synovium tendon bursa disorder	0.039	0.180	0.009
Benign colon neoplasm	0.039	0.173	0.019
Benign skin neoplasm	0.067	0.547	0.007
Non-melanoma skin cancer	0.054	0.520	0.008
Anxiety phobic disorder	0.063	0.432	0.007
Depression	0.038	0.579	0.006
Substance abuse	0.045	0.422	0.010
Breast disorder	0.044	0.166	0.010
Disease of the female reproductive organs	0.105	0.235	0.009

Classification of common human diseases derived from shared genetic and environmental determinants

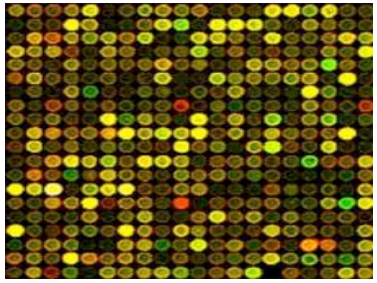
VOLUME 49 | NUMBER 9 | SEPTEMBER 2017 NATURE GENETICS

Kanix Wang[1,2], Hallie Gaitsch[2], Hoifung Poon[3], Nancy J Cox[4] & Andrey Rzhetsky[2,5]



Transcriptomics Review

Measuring Transcript Levels



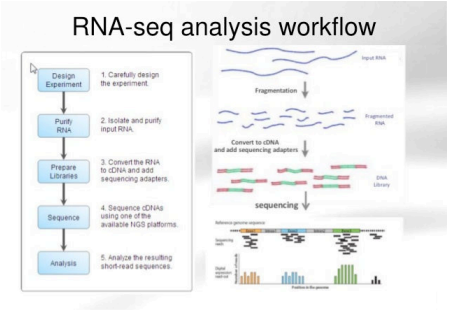
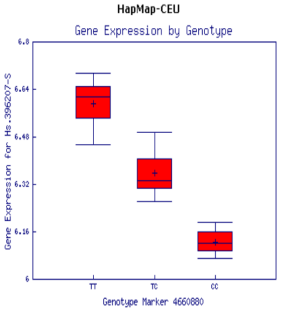
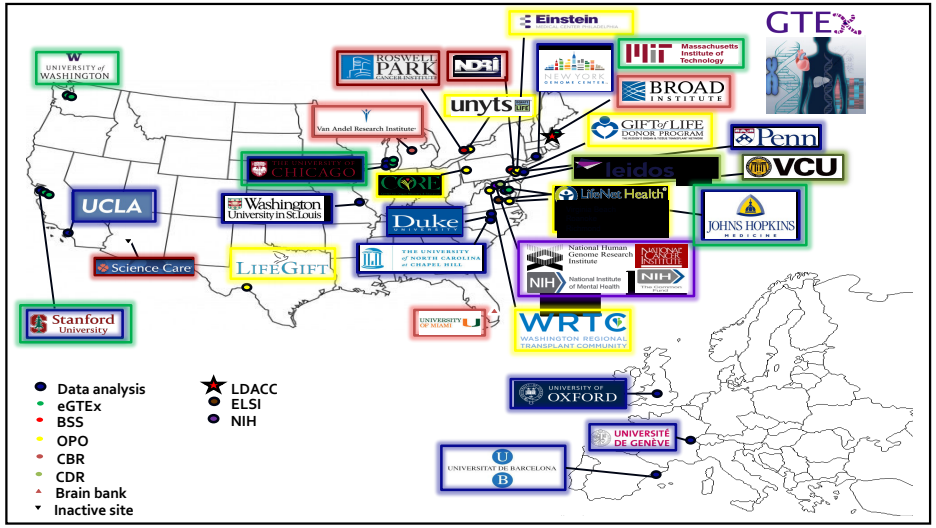


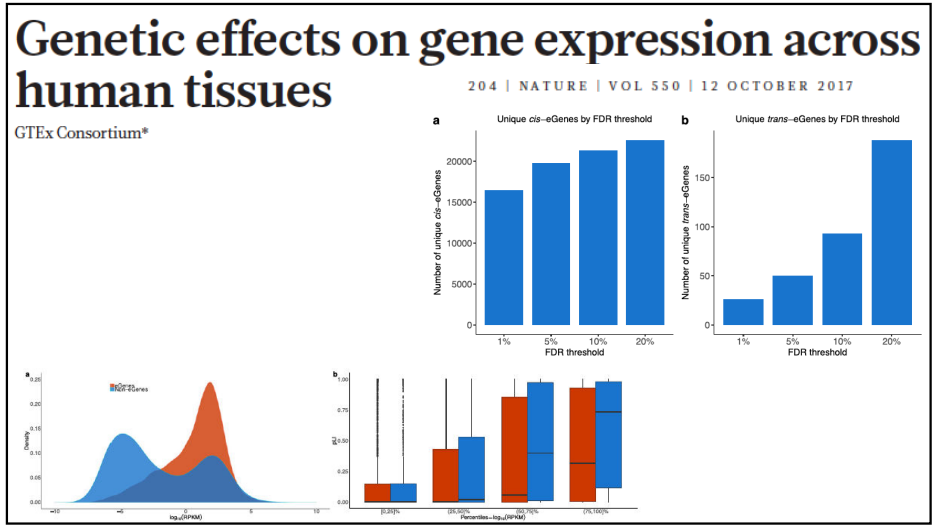
Illustration of eQTL
(An integration: genome x transcriptome)

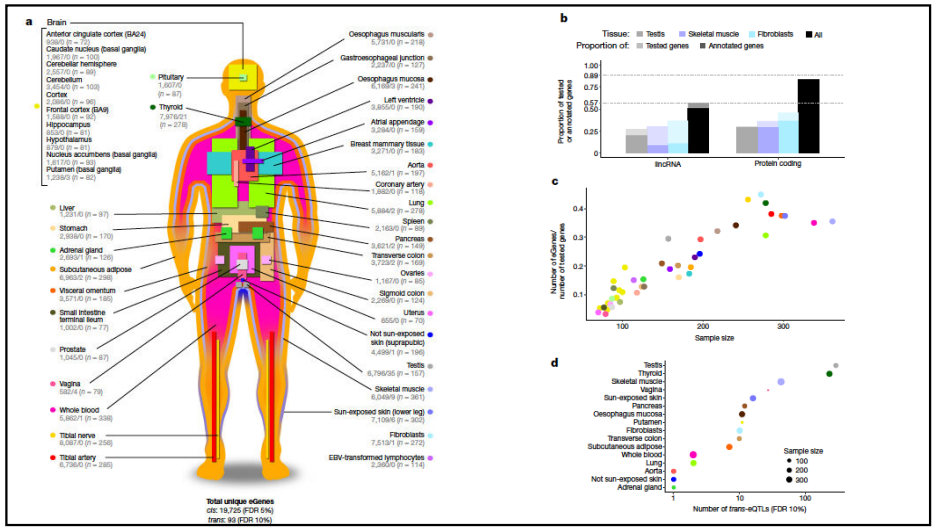
Gene Expression



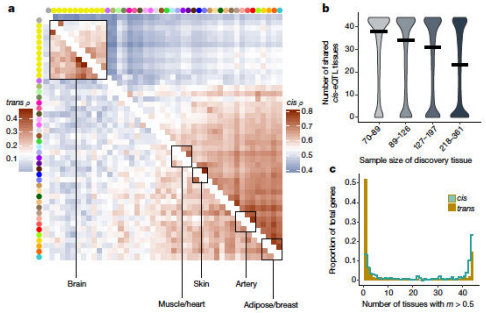
Genotype



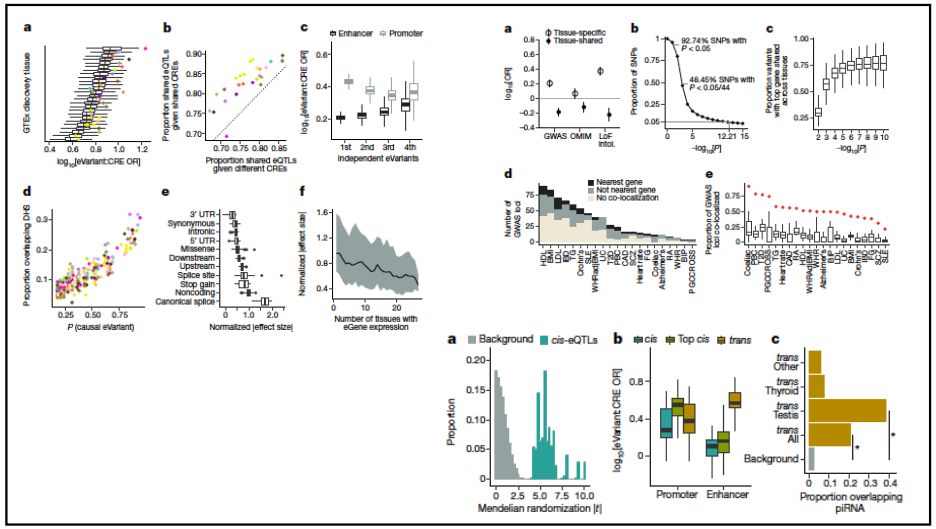


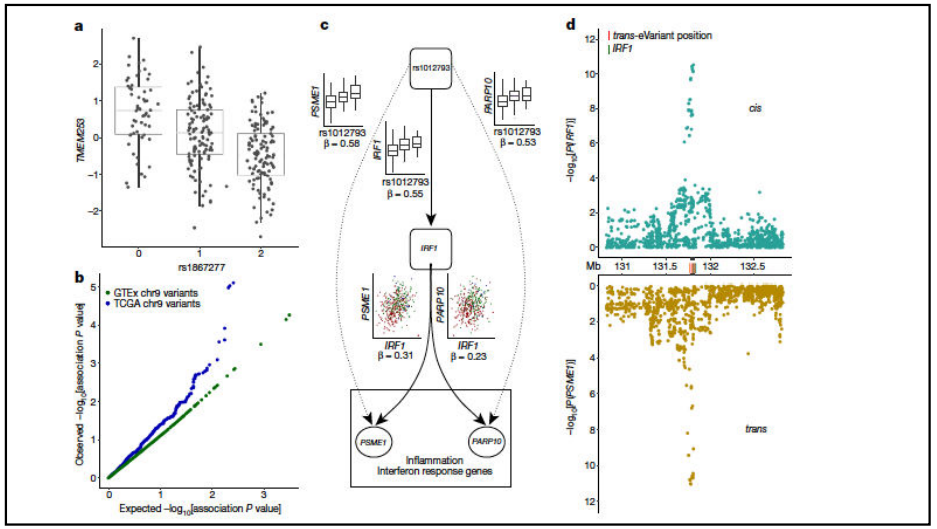


Tissue sharing of eQTL effects



Larger sample sizes allow detection of more tissue-specific eQTLs

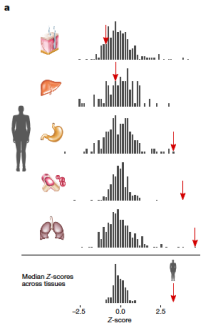




The impact of rare variation on gene expression across tissues

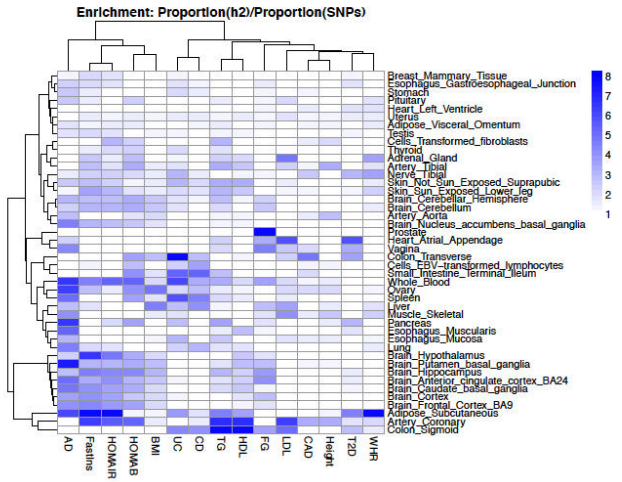
12 OCTOBER 2017 | VOL 550 | NATURE | 239

Xin Li[1]*, Yungil Kim[2]*, Emily K. Tsang[1,3]*, Joe R. Davis[1,4]*, Farhan N. Damani[2], Colby Chiang[5], Gaelen T. Hess[4], Zachary Zappala[1,4], Benjamin J. Strober[6], Alexandra J. Scott[5], Amy Li[4], Andrea Ganna[7,8,9], Michael C. Bassik[4], Jason D. Merker[1], GTEx Consortium†, Ira M. Hall[5,10,11], Alexis Battle[2]§ & Stephen B. Montgomery[1,4]§

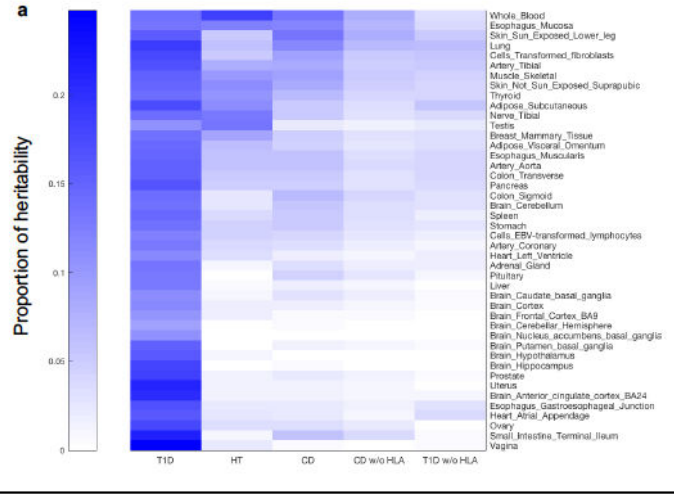


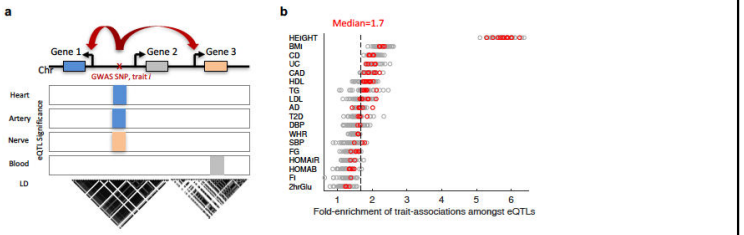
we observed that 58% of underexpression and 28% of overexpression outliers had rare variants near the relevant gene, compared to 8% for non-outliers (Fig. 3c).

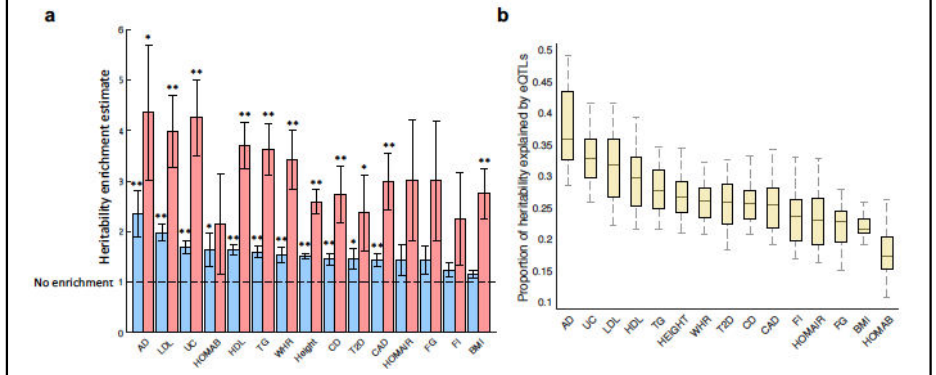
GWAS and Transcriptomics



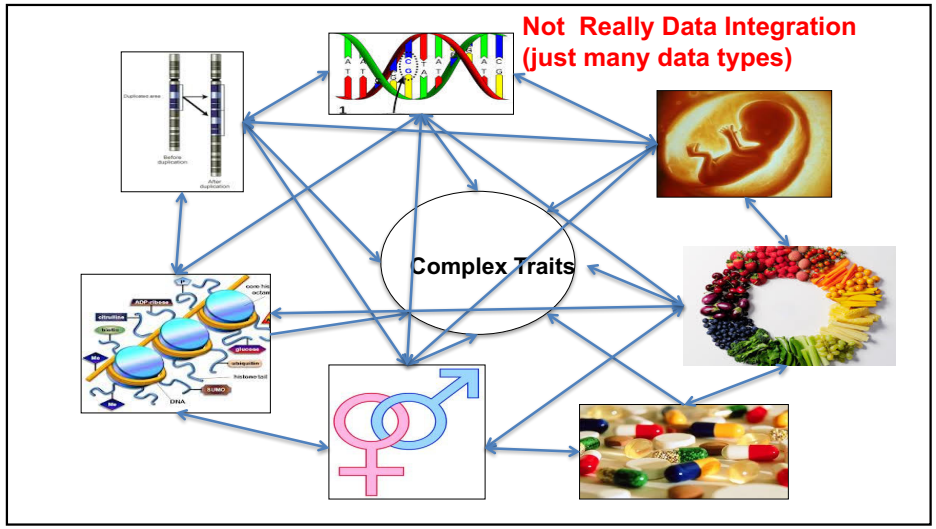
Fold enrichment (proportion of h^2 explained by eQTLs / proportion of total SNPs that were tissue-specific eQTLs)



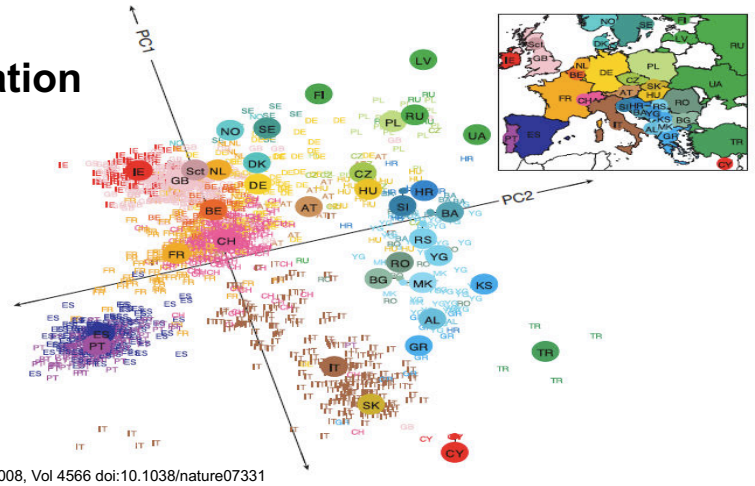




What is Data Integration?



Data Integration



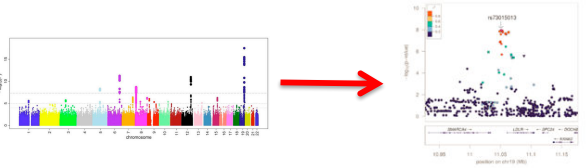
Novembre et al, 2008, Vol 4566 doi:10.1038/nature07331

Data Integration: Phenome X -OMICS

What can we do uniquely well in biobank data?

How can biobank strengths improve understanding of mechanisms of disease?

What Biobanks Do Uniquely Well?

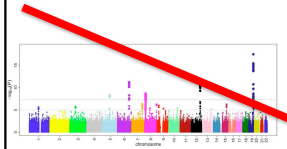


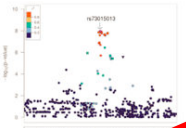
CAUSAL GENE

Relevant tissues, cell types,

MECHANISM

What Biobanks Do Uniquely Well?





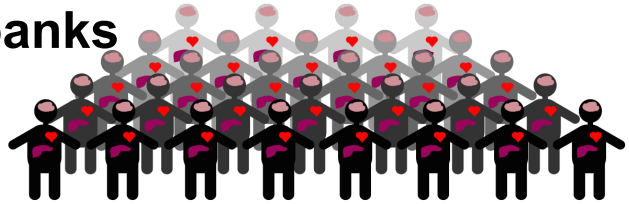
CAUSAL GENE

NOT!

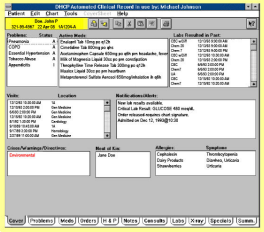
Relevant tissues, cell types,

MECHANISM

Biobanks



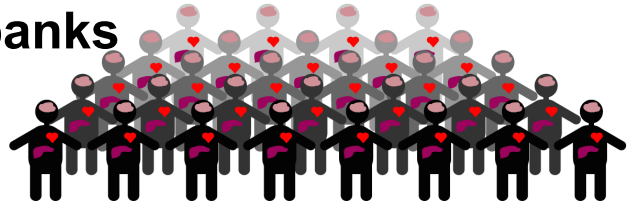
EHR



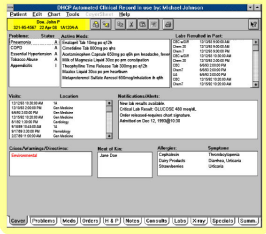
-OMICS

- Genome variation
- Transcriptomes
- Metabolomes
- Methylomes

Biobanks



EHR

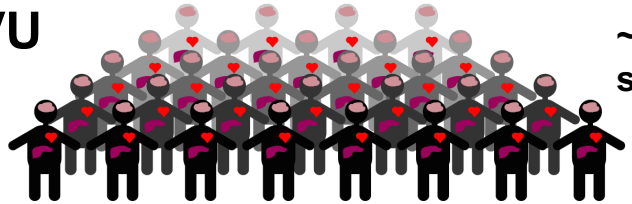


-OMICS

- Genome variation
- Transcriptomes
- Metabolomes
- Methylomes

Measure
Impute

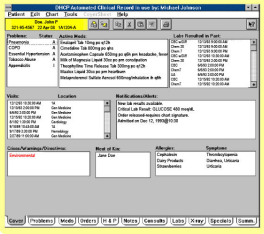
BioVU



~250,000 samples

~10-15 yrs of EHR

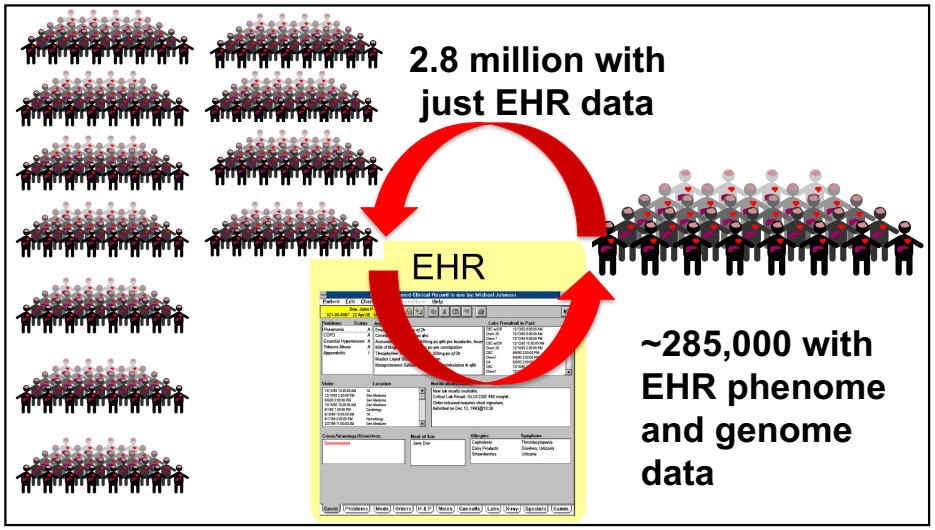
EHR



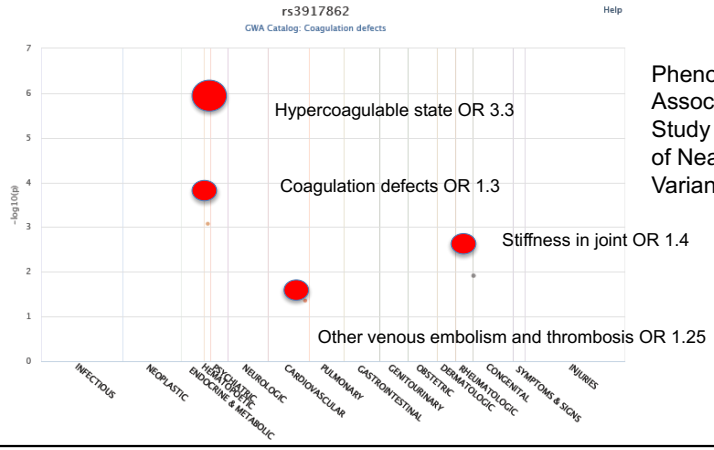
-OMICS

- Genome variation
- Transcriptomes
- Metabolomes

120K
Imputed



What can we do uniquely well in BioVU?



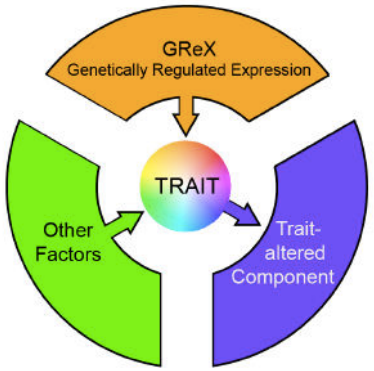
Phenome-wide Association Study (PheWAS) of Neanderthal Variants

GWAS: What variants are associated with this phenotype?

PheWAS: What phenome is associated with this variant?

PrediXcan (Gamazon et al., 2015, Nat Genet)

https://github.com/hakyimlab/PrediXcan

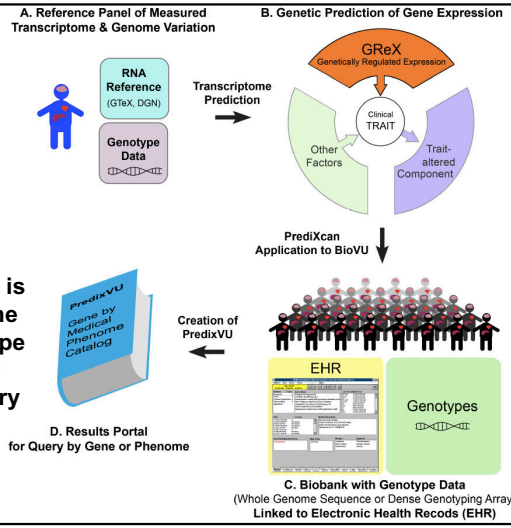


Reference panel: GTEx

$$T = \sum_k \omega_k X_k + \varepsilon$$

GReX – Genetically Regulated Expression

GTEx: WGS and RNA-Seq in 44 tissues from ~450 (>950) subjects



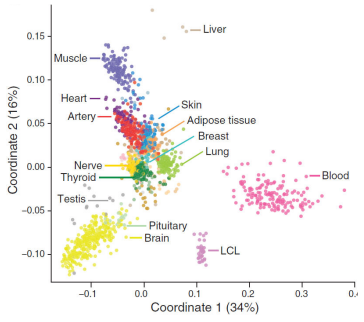
High quality prediction performance in 18,483 genes and lncRNAs

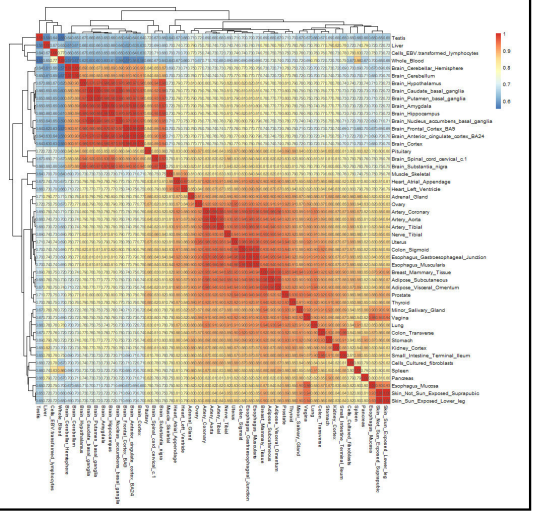
PredixVU catalog is a discovery engine for gene-phenotype relationships and iteration to primary mechanism of disease

BioVU: >250K subjects, linked to EHR going back on average 10-15 years; ~50K with genome interrogation, >120K by 2018

Comparison	PrediXcan	UTMOST	XT-SCAN
			Eric Gamazon
Cross-tissue	No	Yes	Yes
Prediction Model	Elastic Net	Sparse Group LASSO	Weighted Elastic Net
Incorporate regulatory elements	No	No	Yes (DHS)
Incorporate expression similarity	No	No	Yes
			Dan Zhou

Gene expression similarity



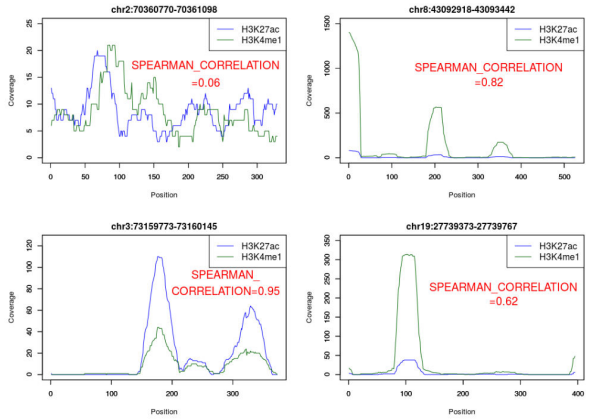


DHS profile similarity

1. For each gene, Compare the DHS profile (around TSS) similarity for each tissue-tissue pair.

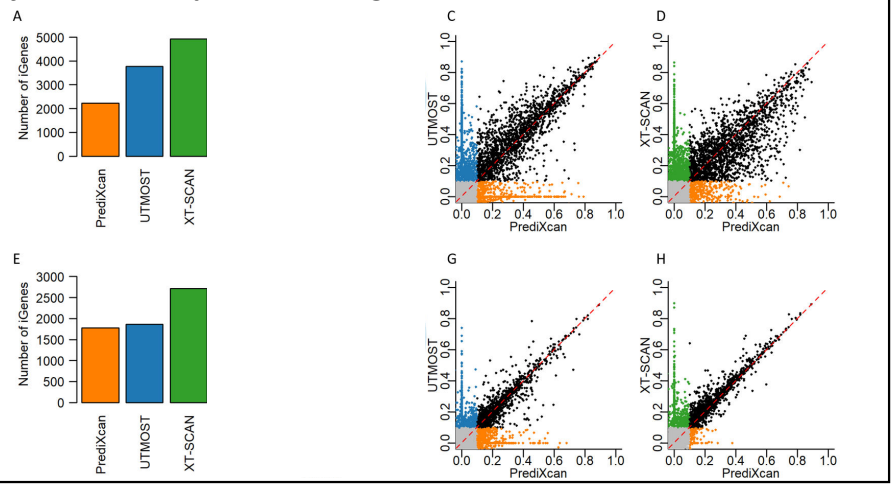
2. Quantile normalization for the similarity across genome.

Not enough data for histone modifications, even for H3K27ac, H3K4me3, etc.



Deschenes A, (2018). similaRpeak: Metrics to estimate a level of similarity between two ChIP-Seq profiles.

Performance comparison among PrediXcan, UTMOST, and XT-SCAN



What Discoveries Do Biobanks Uniquely Enable?

Tennessee Biobanks Special Recipe:

Genes (with known function)

Phenome!

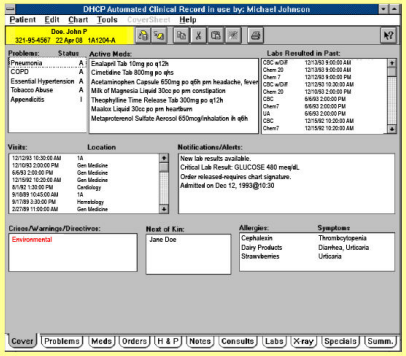
Gene-based PheWAS

What does this gene do?

What does the natural variation in the expression of this gene associate with across the medical phenome?

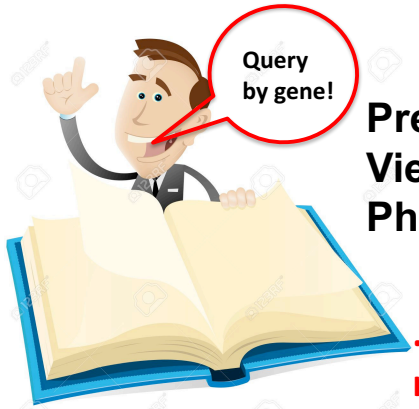
The Full Medical Phenome!

EHR



- Characterize larger-scale consequences of genes
- More/better ways to mechanism
- PheRS larger, richer targets for gene discovery

Genome X Transcriptome X EHR



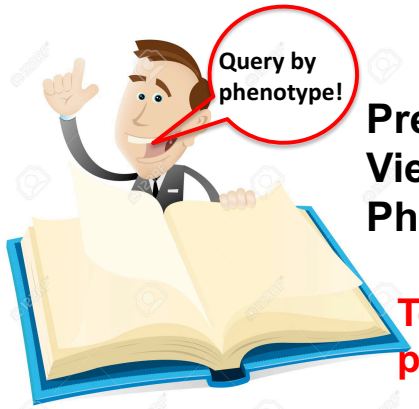
PredixVU: A Catalog for Viewing Gene-Based Phenome-Wide Association

… or gene set, pathway, network …

ZNF577

Trait	R^2	OR per unit SD	P-value	Cases	Controls
Viral hepatitis C	0.37	0.84	6.21E-07	808	20904
Viral hepatitis	0.37	0.86	4.38E-06	937	20904
Malignant neoplasm of liver, primary	0.37	0.81	1.61E-04	267	22080
Cancer of liver and intrahepatic bile duct	0.37	0.82	2.43E-04	318	22080
Precordial pain	0.37	1.22	4.34E-04	374	12283

Genome X Transcriptome X EHR



PredixVU: A Catalog for Viewing Gene-Based Phenome-Wide Association

Test pleiotropy, investigate phenome relationships, …

Viral Hepatitis C

Gene	R^2	OR per unit SD	P-value	Cases	Controls
ZNF577	0.37	8.44E-01	6.21E-07	808	20904
ZNF649	0.12	8.48E-01	1.85E-06	808	20904
SPAG1	0.22	8.63E-01	2.98E-05	808	20904
KLRC1	0.42	8.57E-01	4.31E-05	808	20904
CST2	0.14	8.64E-01	4.42E-05	808	20904

Intestinal Infection

Gene	R^2	OR per unit SD	P-value	Cases	Controls
NDUFA4	0.03	1.16	1.83E-09	1608	24187
C10orf120	0.006	1.13	4.92E-08	1608	24187
TJP1	0.066	1.12	5.27E-06	1608	24187
YEATS2	0.012	0.89	6.33E-06	1608	24187
DGAT2	0.12	1.12	1.11E-05	1608	24187
GNA12	0.23	1.12	1.32E-05	1608	24187
ZNF525	0.13	0.89	1.96E-05	1608	24187
ALLC	0.15	0.89	2.08E-05	1608	24187
HDGFRP2	0.005	0.90	3.45E-05	1608	24187

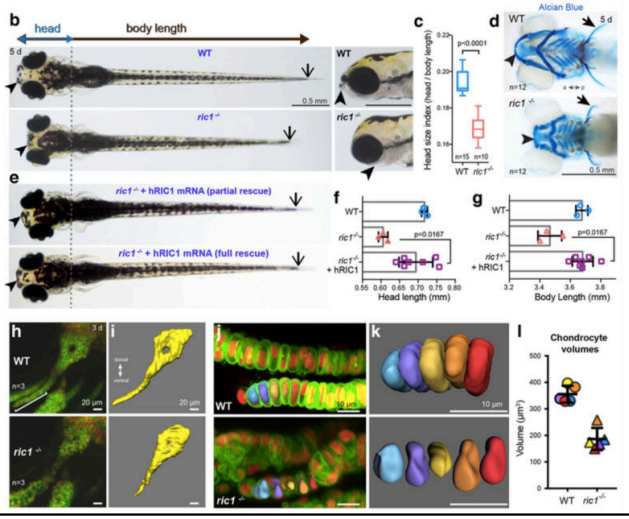
A Fish(ing) Story

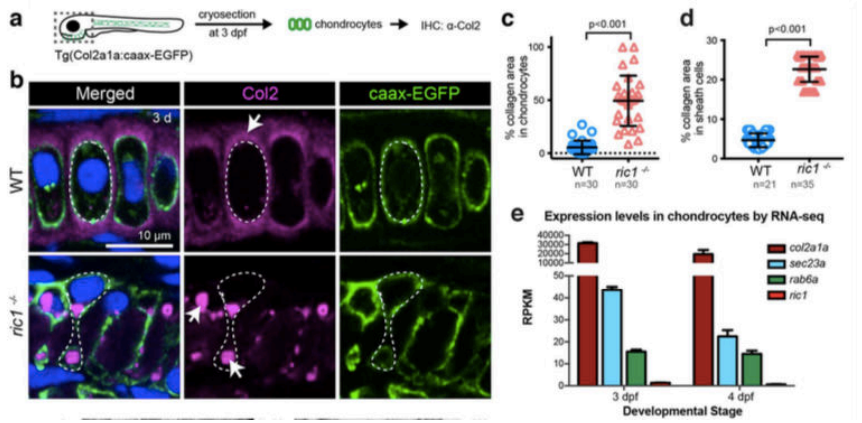
Tennesee Biobanks Special Recipe:

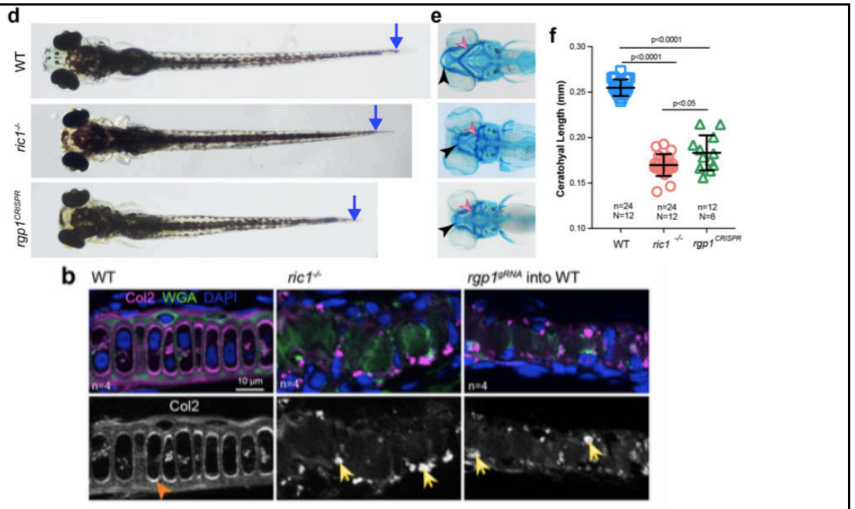
RIC1

Mechanism: Moves fibrillar collagens out of cells

Phenome?



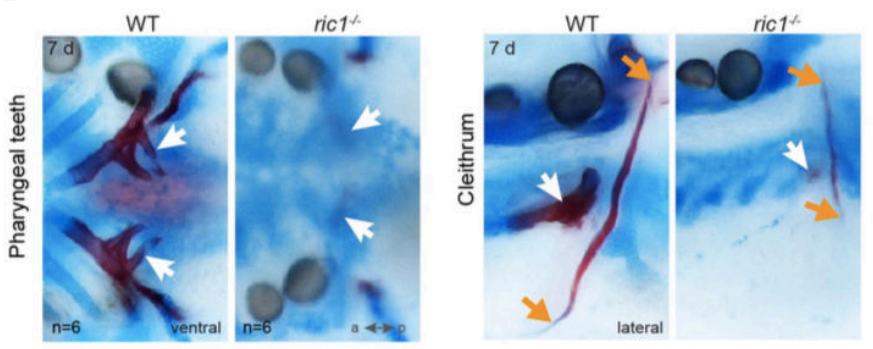




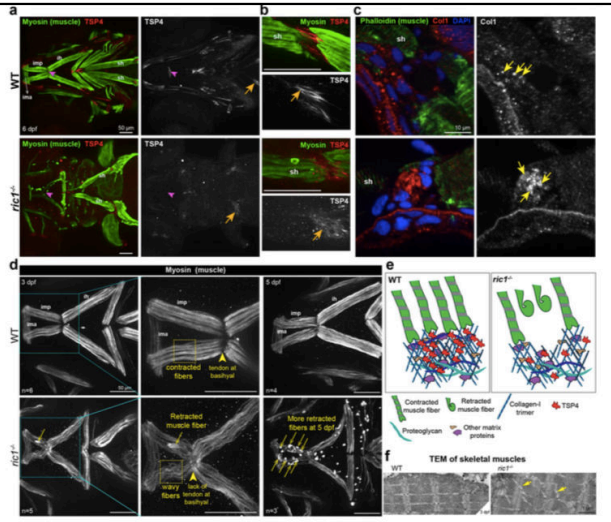
BioVU Phenome for *RIC1/RGP1*

- Bone, fracture, and connective tissue disorders 
- Tooth development and eruption 
- Gait disorders 
- Asthma, heart valve defects and replacement
- Esophageal, gut and digestive disorders 
- Strabismus, amblyopia 
- Neurological and neuropsychiatric disorders 

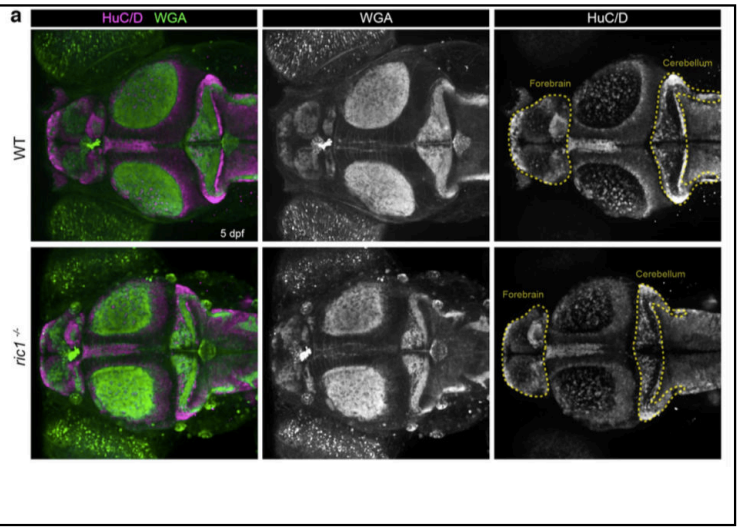
Results of BioVU studies prompted zebrafish studies of tooth development…

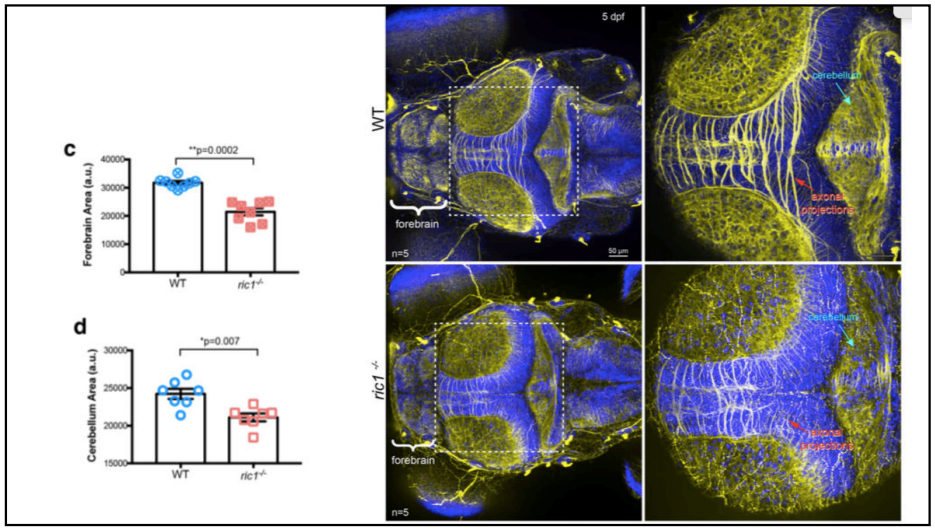


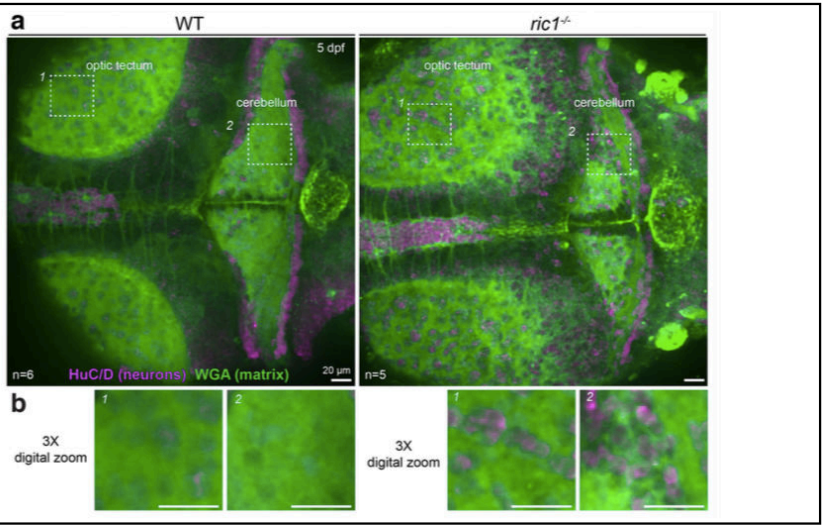
Muscle attachment studies in small (eye) and large muscles…



And brain studies

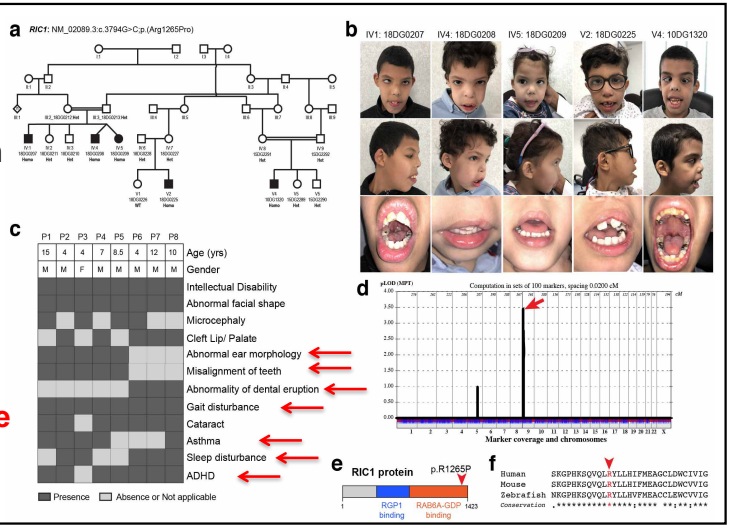


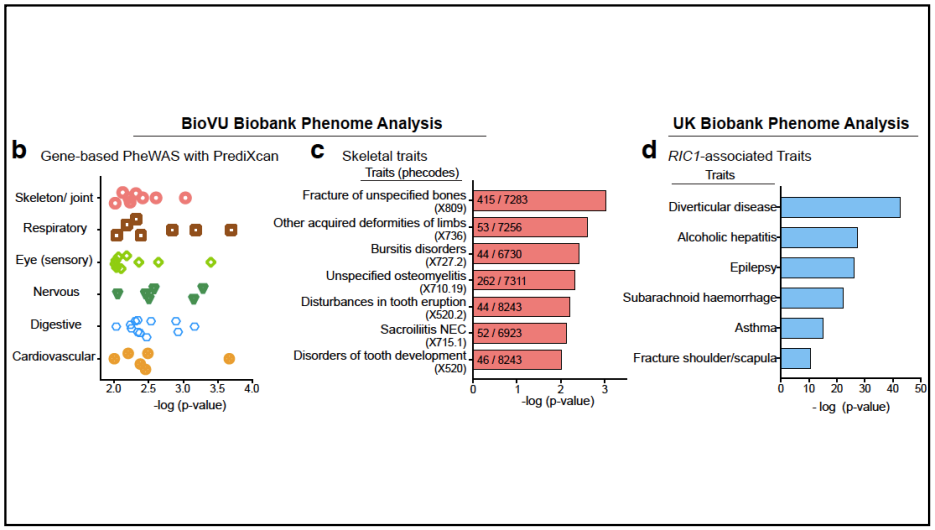




BioVU results used to guide the re-evaluation of patient phenomes

From Moms: Gut Phenome





Pleiotropy:

Failure of basement membrane: digestive disorders, vitamin / mineral deficiencies, malabsorption

RIC1, RGP1 **fail to traffic fibrillar collagens – the most abundant extracellular matrix proteins in vertebrates, providing tissues and organs with form, stability and connectivity**

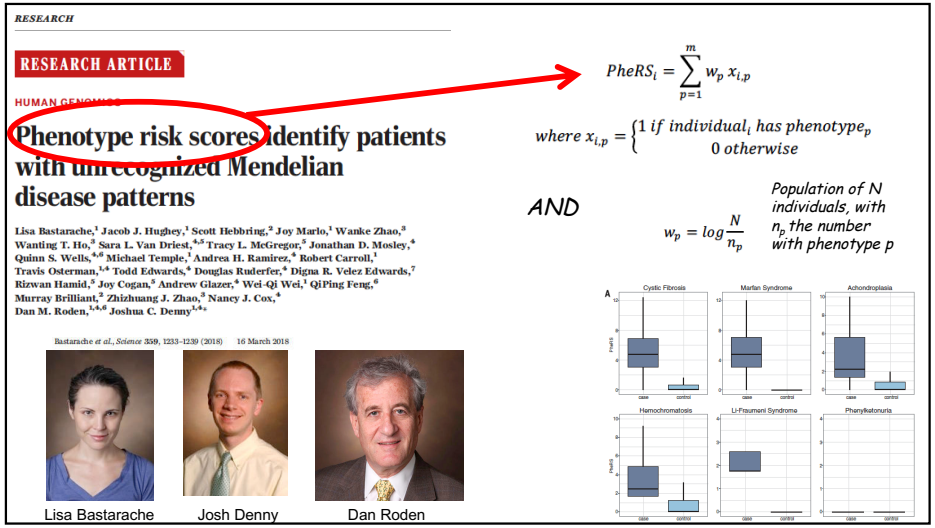
Inadequate ECM, failure to produce cartilage: asthma, organ prolapse, connective tissue disorders, valve defects, fracture, tooth defects

Failure of muscle attachment: strabismus, amblyopia, gait disorders

Inadequate brain ECM: Epilepsy, pervasive developmental delay, intellectual disability, ADHD, alcohol related disease, restless legs

Phenome Fishing to Find More Genes in the *RIC1/RGP1* Space

Use a Phenome Risk Score (PheRS) as the phenotype (rather than an individual disease code)



Genes with Significant Association to *RIC1/RGP1* PheRS

COL14A1, COL3A1, COL5A2, COL9A3 (Duh!)

Exchanges bound GDP for free GTP – a key part of the *RIC1/RGP1* machine for collagen transport

FAM124A, FGD6, EPC2, EPPK1, ADAMTSL4, ATP5H, EDEM2, CPT1A

ER-associated degradation of misfolded glycoproteins

Attaches to fibrillin and fibrillar collagens to stabilize extracellular matrix

Creating a Poly-gene Score for ECM Strength

- Directions of effects consistent across the genes functioning to strengthen ECM
- Sum predicted expression across gene set to probe phenome associations to ECM strength

Pleiotropy is ubiquitous.

Continuum from Mendelian to Complex

Continuum from LOF to deleterious to ↓expression

Mendelian Genes

The 10% we see as Mendelian disease

The 90% under the surface – influence on common disease

Genes for Cardiac Congenital Anomalies

RESEARCH

RESEARCH ARTICLE

HUMAN GENOMICS

Phenotype risk scores identify patients with unrecognized Mendelian disease patterns

Lisa Bastarache,[1] Jacob J. Hughey,[1] Scott Hebbring,[2] Joy Marlo,[1] Wanke Zhao,[3] Wanting T. Ho,[3] Sara L. Van Driest,[4,5] Tracy L. McGregor,[5] Jonathan D. Mosley,[4] Quinn S. Wells,[4,6] Michael Temple,[1] Andrea H. Ramirez,[4] Robert Carroll,[1] Travis Osterman,[1,4] Todd Edwards,[4] Douglas Ruderfer,[4] Digna R. Velez Edwards,[7] Rizwan Hamid,[5] Joy Cogan,[5] Andrew Glazer,[4] Wei-Qi Wei,[1] QiPing Feng,[6] Murray Brilliant,[2] Zhizhuang J. Zhao,[3] Nancy J. Cox,[4] Dan M. Roden,[1,4,6] Joshua C. Denny[1,4]*

Bastarache *et al.*, *Science* 359, 1233–1239 (2018) 16 March 2018

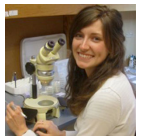
$$PheRS_i = \sum_{p=1}^{m} w_p \, x_{i,p}$$

$$where\ x_{i,p} = \begin{cases} 1\ if\ individual_i\ has\ phenotype_p \\ 0\ otherwise \end{cases}$$

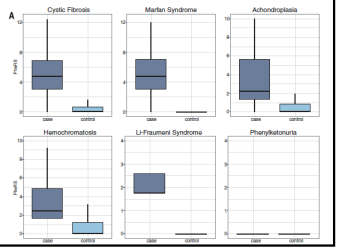
AND

$$w_p = log\frac{N}{n_p}$$

Population of N individuals, with n_p the number with phenotype p



Tyne Miller



VACTERL: at least 3 of…
vertebral defects, anal atresia, cardiac defects, tracheo-esophageal fistula, renal anomalies, and limb abnormalities

GReX Associated with VACTERL PheRS

Gene	P	Tissue
SNX25	8.39E-18	ArteryAorta
UBE2W	8.39E-18	Brain_Cerebellum
RPS17	8.39E-18	Brain_NucleusAccumbensBasalGanglia
MYBL1	8.39E-18	Liver
HLA_DQB1	8.39E-18	Vagina
PCSK4	8.39E-18	WholeBlood
HLA_DQA2	4.90E-16	Spleen
HLA_DQA1	1.27E-13	Vagina
DHODH	2.27E-11	Prostate
ACTA2	3.60E-09	SkinNOTSunExposed
POLQ	1.56E-07	ArteryAorta
VSX1	1.56E-07	Brain_Cerebellum
ZFYVE1	1.56E-07	Brain_NucleusAccumbensBasalGanglia
LILRB4	1.56E-07	WholeBlood
ACSL1	2.24E-07	Adrenal

GREX associated with VACTERL PheRS

Gene	P	Tissue
SNX25	8.39E-18	ArteryAorta
UBE2W	8.39E-18	Brain_Cerebellum
RPS17	8.39E-18	Brain_NucleusAccumbensBasalGanglia
MYBL1	8.39E-18	Liver
HLA_DQB1	8.39E-18	Vagina
PCSK4	8.39E-18	WholeBlood
HLA_DQA2	4.90E-16	Spleen
HLA_DQA1	1.27E-13	Vagina
DHODH	2.27E-11	Prostate
ACTA2	3.60E-09	SkinNOTSunExposed
POLQ	1.56E-07	ArteryAorta
VSX1	1.56E-07	Brain_Cerebellum
ZFYVE1	1.56E-07	Brain_NucleusAccumbensBasalGanglia
LILRB4	1.56E-07	WholeBlood
ACSL1	2.24E-07	Adrenal

J Biol Chem. 2016 Feb 5;291(6):3030-42. doi: 10.1074/jbc.M115.676601. Epub 2015 Nov 24.

Loss of the Ubiquitin-conjugating Enzyme UBE2W Results in Susceptibility to Early Postnatal Lethality and Defects in Skin, Immune, and Male Reproductive Systems.

Wang B[1], Mettlach SA[2], Vincent M[3], Huber AK[2], Basrur V[4], Mangelberger D[5], Zeng L[2], Elenitoba-Johnson K[4], Miller RA[2], Irani DN[2], Dlugosz AA[5], Schnell S[3], Scaglione KM[3], Paulson HL[6].

Author information

Abstract

UBE2W ubiquitinates N termini of proteins rather than internal lysine residues, showing a preference for substrates with intrinsically disordered N termini. The in vivo functions of this intriguing E2, however, remain unknown. We generated Ube2w germ line KO mice that proved to be susceptible to early postnatal lethality without obvious developmental abnormalities. Although the basis of early death is uncertain, several organ systems manifest changes in Ube2w KO mice. Newborn Ube2w KO mice often show altered epidermal maturation with reduced expression of differentiation markers. Mirroring higher UBE2W expression levels in testis and thymus, Ube2w KO mice showed a disproportionate decrease in weight of these two organs (~50%), suggesting a functional role for UBE2W in the immune and male reproductive systems. Indeed, Ube2w KO mice displayed sustained neutrophilia accompanied by increased G-CSF signaling and testicular vacuolation associated with decreased fertility. Proteomic analysis of a vulnerable organ, presymptomatic testis, showed a preferential accumulation of disordered proteins in the absence of UBE2W, consistent with the view that UBE2W preferentially targets disordered polypeptides. These mice further allowed us to establish that UBE2W is ubiquitously expressed as a single isoform localized to the cytoplasm and that the absence of UBE2W does not alter cell viability in response to various stressors. Our results establish that UBE2W is an important, albeit not essential, protein for early postnatal survival and normal functioning of multiple organ systems.

Andrologia. 2019 Feb;51(1):e13171. doi: 10.1111/and.13171. Epub 2018 Oct 15.

Paternal factors and embryonic development: Role in recurrent pregnancy loss.

Dhawan V[1], Kumar M[1], Deka D[2], Malhotra N[2], Singh N[2], Dadhwal V[2], Dada R[1].

Author information

Abstract

The events occurring at the maternal-foetal interface define a successful pregnancy but the current paradigm has shifted towards assessing the contribution of spermatozoa for embryogenesis. Spermatozoa with defective DNA integrity may fertilise the oocyte but affect subsequent embryonic development. The present case-control study was conducted in male partners of couples experiencing recurrent pregnancy loss (RPL) to assess the gene expression of spermatozoal FOXG1, SOX3, OGG1, PARP1, RPS6, RBM9, RPS17 and RPL29. This was correlated with reactive oxygen species (ROS) levels and DNA Fragmentation Index (DFI). Semen samples were obtained from 60 cases and 30 fertile controls. Gene expression was done by qPCR analysis, and relative quantification was calculated by the $2^{-\Delta\Delta Ct}$ method. Chemiluminescence and the sperm chromatin structure assay were used to measure the ROS and DFI levels respectively. FOXG1, OGG1, RPS6 and RBM9 were seen to be upregulated, while SOX3 and PARP1 were downregulated. Relative expression of SOX3, OGG1, RPS6 and RPS17 showed a significant difference between patients and controls ($p < 0.05$). RPL patients were seen to have high ROS (>27.8; $p = 0.001$) and DFI (>30.7; $p < 0.0001$) with respect to controls. Sperm transcript dysregulation and oxidative DNA damage can be "carried over" after implantation, thus affecting embryogenesis and health of the future progeny.

GReX Associated with VACTERL PheRS

Gene	P	Tissue
SNX25	8.39E-18	ArteryAorta
UBE2W	8.39E-18	Brain_Cerebellum
RPS17	8.39E-18	Brain_NucleusAccumbensBasalGanglia
MYBL1	8.39E-18	Liver
HLA_DQB1	8.39E-18	Vagina
PCSK4	8.39E-18	WholeBlood
HLA_DQA2	4.90E-16	Spleen
HLA_DQA1	1.27E-13	Vagina
DHODH	2.27E-11	Prostate
ACTA2	3.60E-09	SkinNOTSunExposed
POLQ	1.56E-07	ArteryAorta
VSX1	1.56E-07	Brain_Cerebellum
ZFYVE1	1.56E-07	Brain_NucleusAccumbensBasalGanglia
LILRB4	1.56E-07	WholeBlood
ACSL1	2.24E-07	Adrenal

Mol Genet Metab. 2016 Sep;119(1-2):83-90. doi: 10.1016/j.ymgme.2016.06.008. Epub 2016 Jun 14.

Elevated plasma dihydroorotate in Miller syndrome: Biochemical, diagnostic and clinical implications, and treatment with uridine.

Duley JA[1], Henman MG[2], Carpenter KH[3], Bamshad MJ[4], Marshall GA[2], Ooi CY[5], Wilcken B[7], Pinner JR[8].

Author information

Abstract

BACKGROUND: Miller syndrome (post-axial acrofacial dysostosis) arises from gene mutations for the mitochondrial enzyme dihydroorotate dehydrogenase (DHODH). Nonetheless, despite demonstrated loss of enzyme activity dihydroorotate (DHO) has not been shown to accumulate, but paradoxically urine orotate has been reported to be raised, confusing the metabolic diagnosis.

METHODS: We analysed plasma and urine from a 4-year-old male Miller syndrome patient. DHODH mutations were determined by PCR and Sanger sequencing. Analysis of DHO and orotic acid (OA) in urine, plasma and blood-spot cards was performed using liquid chromatography-tandem mass spectrometry. In vitro stability of DHO in distilled water and control urine was assessed for up to 60h. The patient received a 3-month trial of oral uridine for behavioural problems.

RESULTS: The patient had early liver complications that are atypical of Miller syndrome. DHODH genotyping demonstrated compound-heterozygosity for frameshift and missense mutations. DHO was grossly raised in urine and plasma, and was detectable in dried spots of blood and plasma. OA was raised in urine but undetectable in plasma. DHO did not spontaneously degrade to OA. Uridine therapy did not appear to resolve behavioural problems during treatment, but it lowered plasma DHO.

CONCLUSION: This case with grossly raised plasma DHO represents the first biochemical confirmation of functional DHODH deficiency. DHO was also easily detectable in dried plasma and blood spots. We concluded that DHO oxidation to OA must occur enzymatically solved the biochemical conundrum in previous reports of Miller syndrome patients, and opened the ening.

Front Immunol. 2017 Aug 24;8:1013. doi: 10.3389/fimmu.2017.01013. eCollection 2017.

Decidual Macrophage Functional Polarization during Abnormal Pregnancy due to *Toxoplasma gondii*: Role for LILRB4.

Li Z[1], Zhao M[2], Li T[1], Zheng J[3], Liu X[1], Jiang Y[1], Zhang H[1], Hu X[1].

Author information

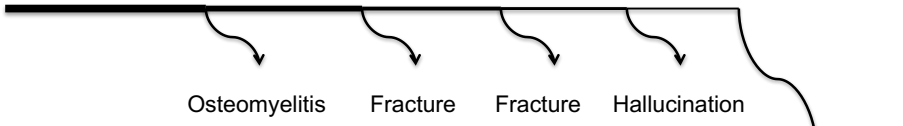
Abstract

During gestation, *Toxoplasma gondii* infection produces a series of complications including stillbirths, abortions, and congenital malformations. The inhibitory receptor, LILRB4, which is mainly expressed by professional antigen-presenting cells (especially macrophages and dendritic cells) may play an important immune-regulatory role at the maternal-fetal interface. To assess the role of LILRB4 during *T. gondii* infection, $LILRB4^{+/+}$ and *T. gondii* infected pregnant mouse models were established. Further, human primary-decidual macrophages were treated with anti-LILRB4 neutralizing antibody and then infected with *T. gondii*. These *in vivo* and *in vitro* models were used to explore the role of LILRB4 in *T. gondii*-mediated abnormal pregnancy outcomes. The results showed that abnormal pregnancy outcomes were more prevalent in $LILRB4^{+/+}$ infected pregnant mice than in wild-type infected pregnant mice. In subsequent experiments, expression levels of LILRB4, M1, and M2 membrane-functional molecules, arginine metabolic enzymes, and related cytokines were assessed in uninfected, infected, LILRB4-neutralized infected, and $LILRB4^{+/+}$ infected models. The results demonstrated *T. gondii* infection to downregulate LILRB4 on decidual macrophages, which strengthened M1 activation functions and weakened M2 tolerance functions by changing M1 and M2 membrane molecule expression, synthesis of arginine metabolic enzymes, and cytokine secretion profiles. These changes contributed to abnormal pregnancy outcomes. The results of this study provide not only a deeper understanding of the immune mechanisms operational during abnormal pregnancy, induced by *T. gondii* infection, but also identify potential avenues for therapeutic and preventive treatment of congenital toxoplasmosis.

G X E

Using pleiotropy to aid discovery…

No gene exists simply to cause human disease…



Xue Zhong, PhD

Phenome Associations to GReX of *PSEN1*

Trait	OR per unit SD	p-value	Cases	Controls
Unspecified osteomyelitis	8.38E-01	4.03E-05	524	19865
Fracture of upper limb	8.95E-01	6.03E-05	1355	21096
Acute osteomyelitis	7.79E-01	7.06E-05	231	19865
Fracture of humerus	8.41E-01	7.18E-05	504	21096
Bacterial enteritis	1.15E+00	7.67E-05	948	
Intestinal infection due to C. difficile	1.15E+00	1.10E-04	894	
Heart failure with reduced EF	9.15E-01	1.10E-04	2262	
Osteomyelitis	8.70E-01	1.41E-04	735	
Osteomyelitis, periostitis, and other	8.80E-01	2.14E-04	835	
Congenital anomalies of limbs	6.75E-01	2.34E-04	76	

Fractures: patella, radius, ulna, vertebral column
Brain: intracranial hemorrhage, subdural hemorrhage, cerebral aneurism, Alzheimers

TREM2 GReX Associations

Nasu-Hakola Disease

Alzheimers and other dementias

Trait	OR per unit SD	p-value	Cases	Controls
Chronic osteomyelitis	8.18E-01	1.40E-05	344	19865
Acute osteomyelitis	7.93E-01	1.97E-05	231	19865
Cerebral laceration and contusion	7.34E-01	5.55E-05	Fractures: patella, radius, ulna, vertebral column Brain: Delerium, psychogenic disorder, dementia with cerebral degenerations	
Unspecified osteomyelitis	8.59E-01	1.17E-04		
Osteomyelitis	8.79E-01	1.51E-04		
Diverticulitis	8.47E-01	4.98E-04		
Hallucinations	7.44E-01	7.18E-04		

Osteoclasts and microglia share a developmental lineage…

- Substantial overlap in genes expressed, and the genetic architecture of their gene regulation
- Additional of the genes implicated in Alzheimers show similar patterns of phenome association

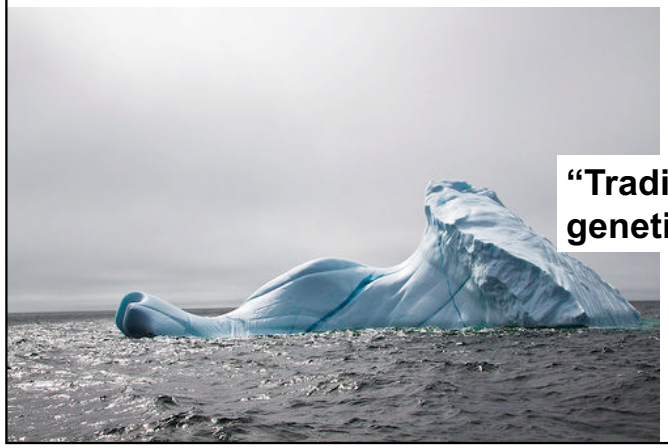
In 2.8M Subjects with EHR…

- Age at diagnosis for osteomyelitis is 15-20 years earlier than diagnosis of dementia
- Subjects with osteomyelitis diagnosed before 60 have 4-6 X increased risk of dementia diagnosis after age 70

In Silico Drug Trials

- Already conducting in silico drug trials: Do patients taking a drug for one indication have altered risk or age at onset for another disease (e.g. AD)? YES
- *Pleiotropic phenome* discovered through genetics creates a larger target for discovery and validation for *in silico* and later clinical trials

Translation Today



Translation Tomorrow

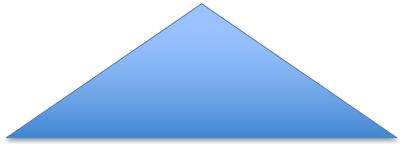
Improving Commonly Used Biomarkers

Genetics → Biomarkers ← Environment

Biomarkers

Metabolome proxies of coding variants

LDL



Cystatin C

New –Omics Biomarkers

Discovery of New Biomarker Relationships

Ranges and Thresholds

- Some biomarker measurements are adjusted for sex, age, race/ethnicity
- It is not about race/ethnicity! It is about DNA variation with frequency differences by historical geographic ancestry

Predictive Modeling

- Done extensively in learning healthcare systems now
- Where can we add genetics to the modeling and get improvement?
- For the most expensive conditions, even marginal improvements in prediction yield better care, lower costs

The IOTA Concept:

When we already have the genome information, even an iota of improvement over current practice (or existing predictive models) will have value for the most expensive and common conditions

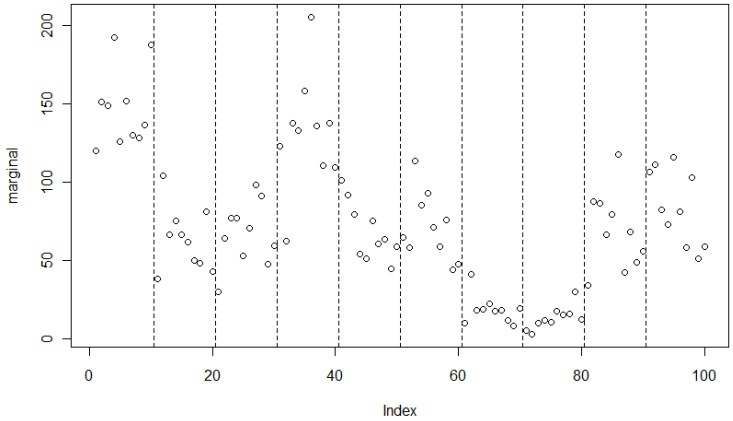
Statistical fine-mapping of GWAS signals

Bogdan Pasaniuc
Associate Professor
Computational Medicine
UCLA
@bpasaniuc
pasaniuc@ucla.edu

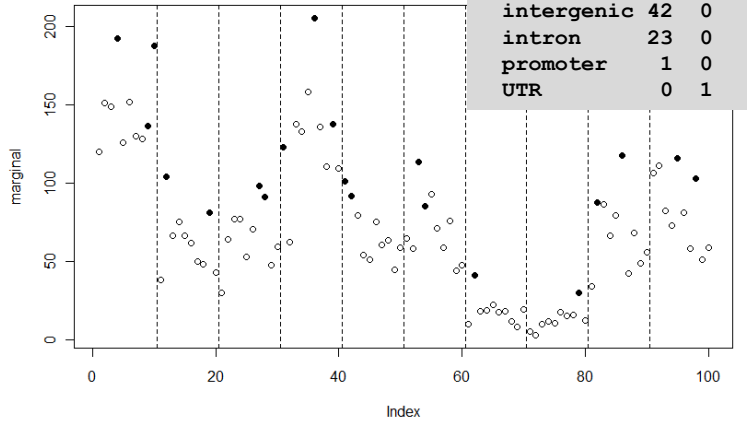
Advanced gene mapping course, January 2020

Motivating example:

- You've conducted a GWAS using in ~100,000 individuals.
- 10 regions are GWAS significant.
- You have resources to conduct follow-up studies of 10 SNPs.
- How do you chose the SNPs for functional follow-up?



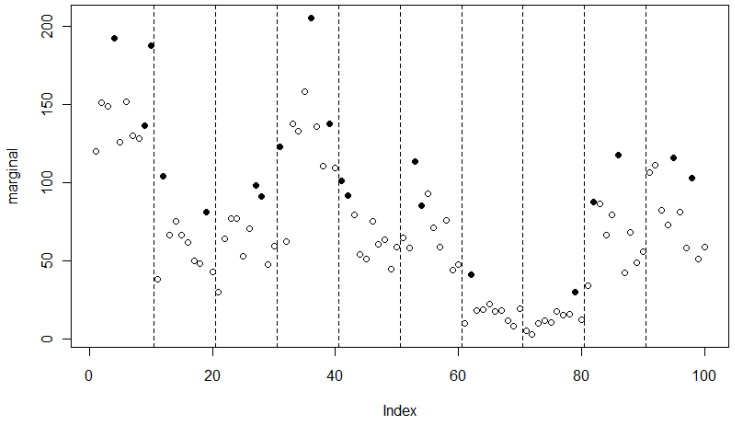
How would you choose the ten SNPs to follow up? How would you use the genetic association data? What other information could you use? How would you combine the association results with this additional information?



	truth	
	0	1
dhs	14	19
intergenic	42	0
intron	23	0
promoter	1	0
UTR	0	1

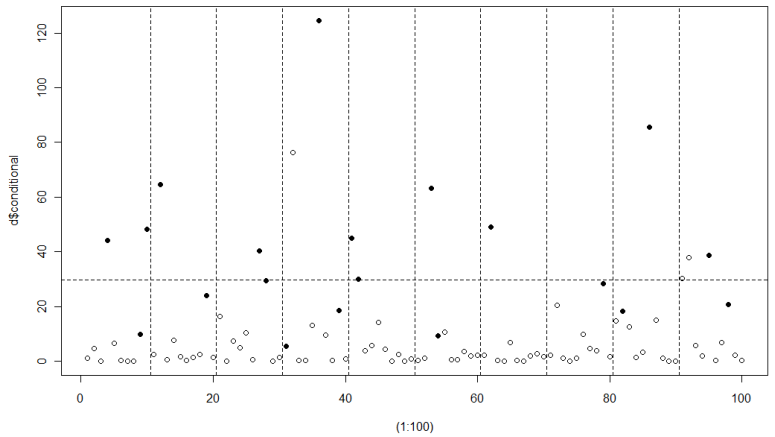
Filled in points = true causal SNPs

Global ranking emphasizes regions with high effect/LD



Filled in points = true causal SNPs

Conditional



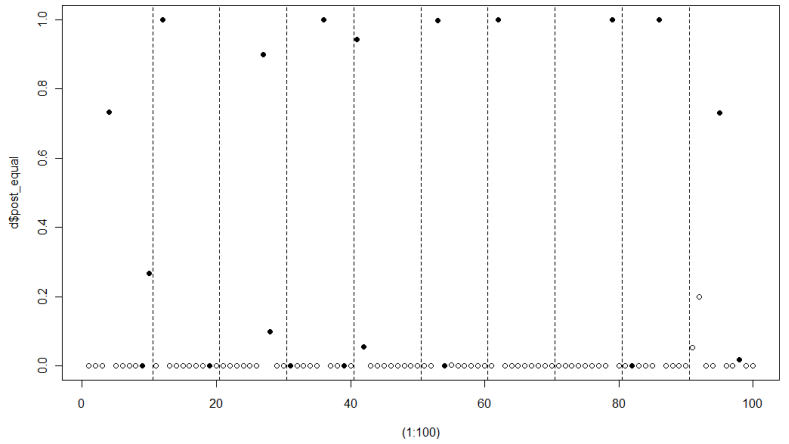
Stepwise conditional analysis

Does this analysis match the goal?

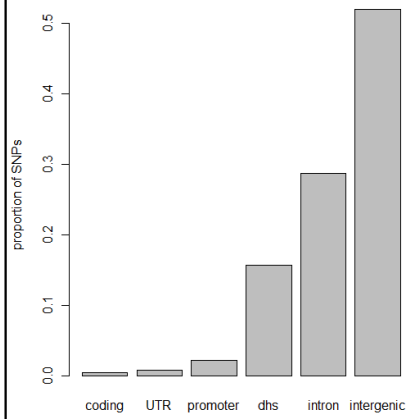
Maybe not: the more “LD friends” a causal variant has, the more likely the stepwise approach will pick one of its “friends”

This is particularly concerning when power is modest (unlike in our well-powered example)

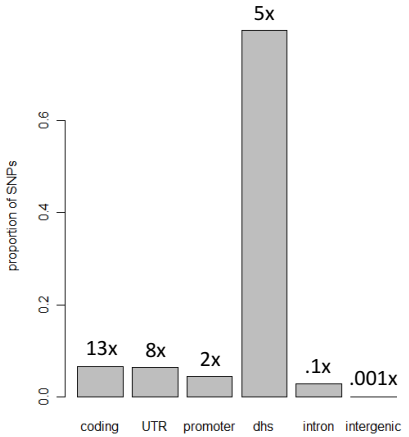
Posterior: 1 causal



Proportion of 1,000 Genomes SNPs in different categories



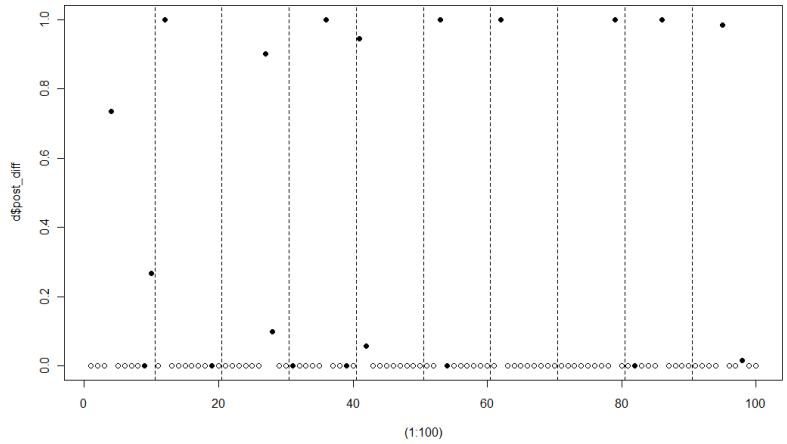
Proportion of 1,000 Genomes SNPs that are causally associated with complex traits



Gusev et al. Regulatory variants explain much more heritability than coding variants across 11 common diseases. AJHG 2014

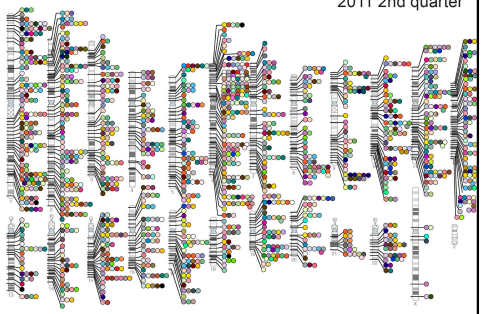
Multiple lines of evidence and several different types of analysis suggest that certain functional categories are enriched for SNPs associated with complex traits. We might infer that causal SNPs are similarly enriched.

Posterior: 1 causal with different weights



Thousands of loci found by GWAS

- GWAS ➔ >1,000 variants associated to various diseases
- Variants at any locus are highly correlated (LD)
- GWAS-Variants are **not causal**

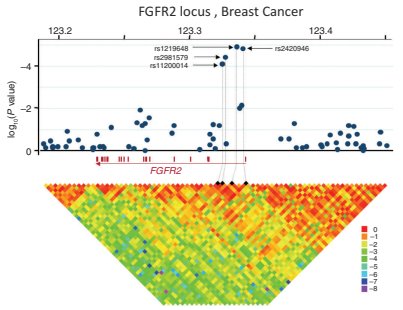


NHGRI GWA Catalog, www.genome.gov/GWAStudies

How do we identify causal variants at GWAS loci?

Fine mapping studies

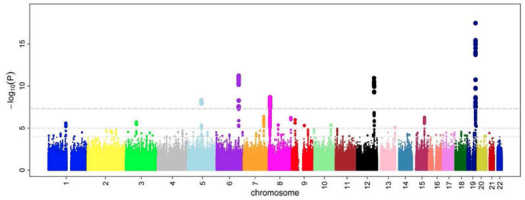
- **Goal:** find causal variants at GWAS associated loci
- Sample size required for differentiating SNPs in LD[Udler et al GenEpi 2010]:
  - $r^2(0.3)$ -> N=3,600 (maf=0.3, R=1.2)
  - $r^2(0.8)$ -> N=12,500 (maf=0.3, R=1.2)



[Hunter et al NG 2007]

Large sample sizes required for fine mapping!

An optimization approach to fine-mapping



- Many GWAS associated loci
  - Height (>180), BC(~30), PC(~30)…
- 2-step approach:
  1. Select set of SNPs for functional validation from considered loci
  2. Test set of variants in functional assays
- **Goal:** find as many causals within fixed budget

Methods for fine-mapping

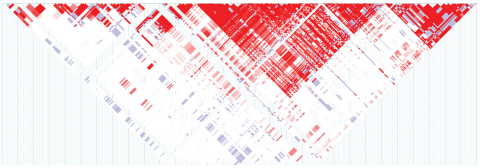
- GOAL: prioritize variants for functional assays
- Prioritization approaches:
  - Marginal association statistics
    - Ignores LD ➔ suboptimal performance
  - Conditioning approaches
    - No clear strategy for selecting variants to condition on
    - Stopping condition?
    - Does not correctly model LD for prioritizing causal variants
  - Probabilistic approach
    - Estimates probability of each variant to be causal
    - Can integrate ENCODE genomic features as priors

Statistical model ➔ causal SNP probabilities

(statistical fine-mapping, see Schaid et al NRG 2018)

Pervasive Linkage Disequilibrium (LD) in the human genome

- Neighboring SNPs are inherited together on haplotypes
- Block-like correlation structure across the human genome



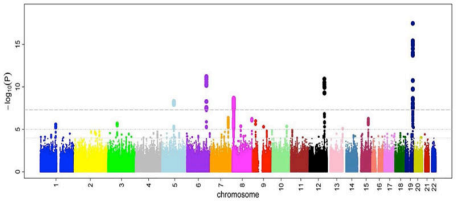
Barret et al. *Bioinformatics* 2005

Only need to measure a subset of the markers for effective GWAS!

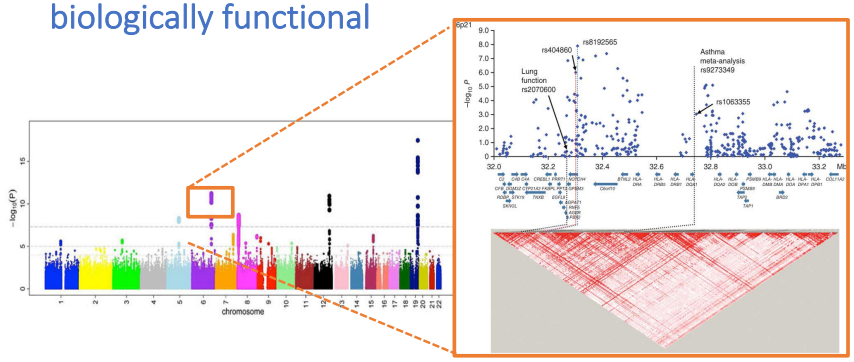
- Neighboring SNPs are inherited together on haplotypes
- Block-like correlation structure across the human genome
- Fill in the rest through imputation

Marchini et al. *Nat Rev Genet* 2010

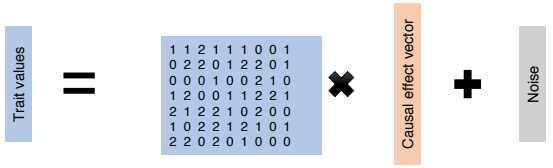
Groups of neighboring variants will display significant associations due to LD



Fine mapping aims to figure which variants are responsible for the observed association → biologically functional



Basic linear model for trait



$$y_{[n,1]} = X_{[n,m]} * \beta_{[m,1]} + \epsilon_{[n,1]}$$

n: individuals ***m***: SNPs ***X***: standardized genotype matrix var(***y***)=1

GWAS marginal effect is biased due to LD

GWAS effect size at SNP i:

$$\widehat{\boldsymbol{\beta}_{\boldsymbol{GWAS,i}}} = \frac{1}{n}X_i y = \frac{1}{n}X_i([X_1 \cdots X_p]\beta + \epsilon)$$
$$= \left[\frac{1}{n}X_i X_1 \cdots \frac{1}{n}X_i X_p\right]\beta + \frac{1}{n}X_i\epsilon$$
$$= \boldsymbol{\Sigma_{j=1}^{p}\, r_{ij}\beta_j + \frac{1}{n}X_i\epsilon}$$

Or in matrix notation:

$$\widehat{\boldsymbol{\beta}_{\boldsymbol{GWAS}}} = \boldsymbol{MVN}(\boldsymbol{V\beta}, \frac{\boldsymbol{V(1-h_g^2)}}{\boldsymbol{n}})$$

Am. J. Hum. Genet. 69:1–14, 2001

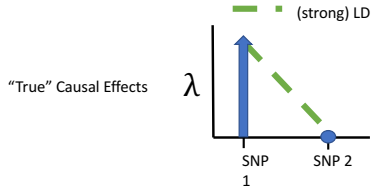
REVIEW ARTICLE
Linkage Disequilibrium in Humans: Models and Data

Jonathan K. Pritchard[1] and Molly Przeworski[2]
[1]Department of Statistics, University of Oxford, Oxford

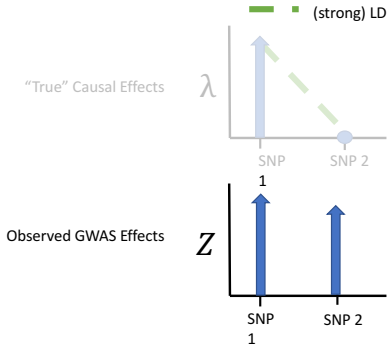
In this review, we describe recent empirical and theoretical work on the extent of linkage disequilibrium (LD) in the human genome, comparing the predictions of simple population-genetic models to available data. Several studies report significant LD over distances longer than those predicted by standard models, whereas some data from short, intergenic regions show less LD than would be expected. The apparent discrepancies between theory and data present a challenge—both to modelers and to human geneticists—to identify which important features are missing from our understanding of the biological processes that give rise to LD. Salient features may include demographic complications such as recent admixture, as well as genetic factors such as local variation in recombination rates, gene conversion, and the potential segregation of inversions. We also outline some implications that the emerging patterns of LD have for association-mapping strategies. In particular, we discuss what marker densities might be necessary for genomewide association scans.

Pritchard&Przeworski AJHG 2001, Shi et al, AJHG 2016, ...

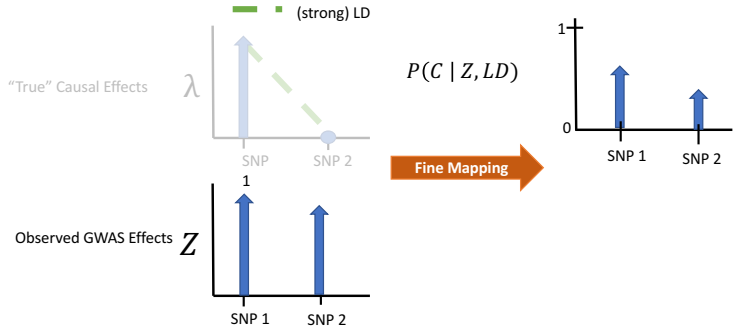
LD induces correlations between causal and non-causal SNPs



LD induces correlations between causal and non-causal SNPs



Given the correlation structure and association strength, quantify the probability that SNPs explain the signal



CAVIAR, CAVIARBF, FINEMAP, PAINTOR, Susie, etc…
(nicely reviewed in Schaid et al NRG 2018)

Approximate estimation of posterior probabilities (1-causal assumption)

- Assumptions
  - one causal SNP per locus
  - causal variant is typed

$$P(c \mid \bar{s}) = \frac{P(\bar{s} \mid c)P(c)}{P(c)}$$

- Marginal association statistics are sufficient to estimate posterior [Maller et al Nat Gen 2012]
  - Can be extended to frequentist approach
  - Statistics at nearby SNPs are independent of phenotype conditional on causal variant
- "Robust to misspecifications"
- Can estimate confidence sets

Allowing for multiple causals improves accuracy

- Causal status vector ➔

$$C = \{c_1, c_2, \cdots c_m\}$$

- Use Bayes to compute probability of each vector

$$P(C|S) = \frac{P(S|C)P(C)}{P(S)}$$

- Model for association statistics

Causal status (Unknown)

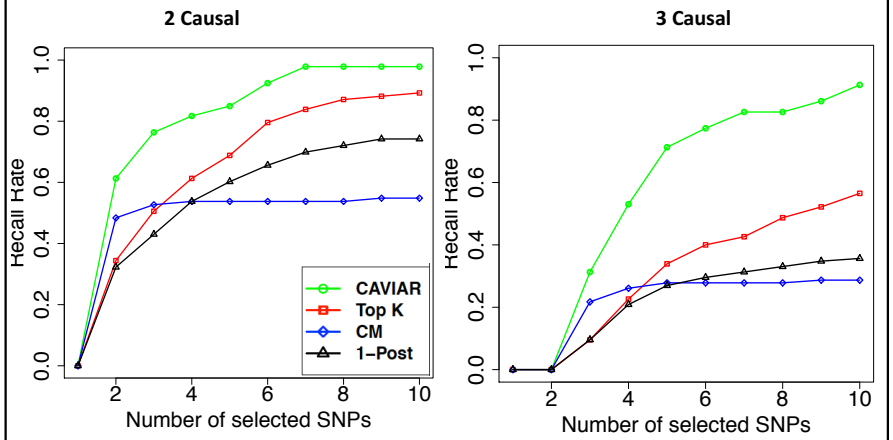
$$S \sim MVN(\lambda_c \Sigma C, \Sigma)$$

Z-score (Known)

NCP of causal SNP (Known)

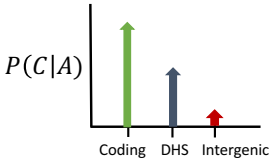
LD pattern (Known)

Allowing for multiple causals improves accuracy

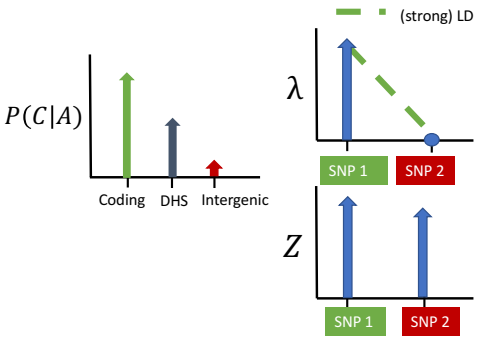


[Hormozdiari et al. Genetics 2014]

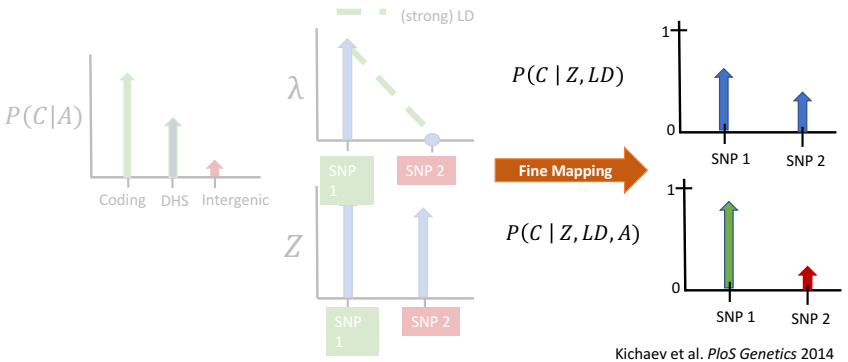
Probabilistic models provide a principled way to incorporate prior biological knowledge



Probabilistic models provide a principled way to incorporate prior biological knowledge



Probabilistic models provide a principled way to incorporate prior biological knowledge

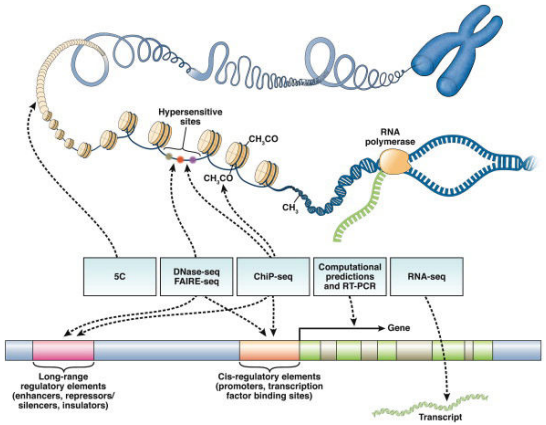


Kichaev et al. *PloS Genetics* 2014

(1) Where to get external functional information?
(e.g., classes of SNPs more likely to have function)

(2) How to quantify which classes of SNPs more useful for our trait of interest?
(e.g., regulatory in tissue X vs tissue Y)

(1) ENCODE/ROADMAP provides a functional map of the human genome (or use your own data)



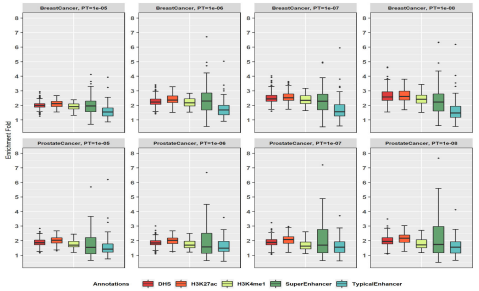
"Here, we assign biochemical functions for 80% of the genome"
—ENCODE Consortium 2012 Nature

(2) Functional enrichment

- Question:
  - Is GWAS signal concentrated in particular ”functional” areas of the genome?
- Functional enrichment = GWAS Signal / proportion of genome covered by functional annotation

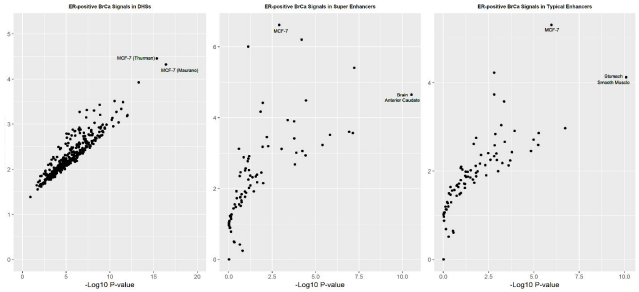
(2) Functional enrichment quantification
Option 1: count of biofeatures with GWAS signal

- Question:
  - Is GWAS signal concentrated in particular ”functional” areas of the genome?
- Functional enrichment = GWAS Signal / proportion of genome covered by functional annotation
- Example from Breast/Prostate Cancer (Chen..Linstroem Hum Genet 2019)

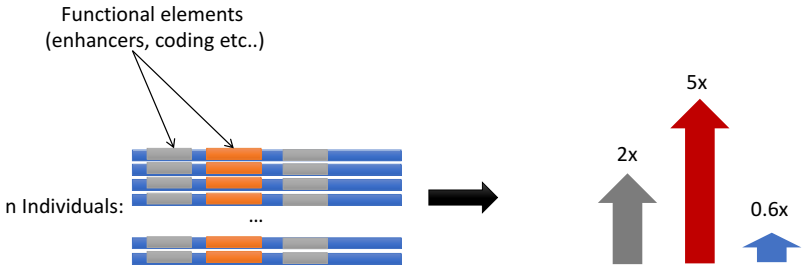


(2) Functional enrichment quantification
Option 1: count of biofeatures with GWAS signal

- Question:
  - Is GWAS signal concentrated in particular ”functional” areas of the genome?
- Functional enrichment = GWAS Signal / proportion of genome covered by functional annotation
- Example from Breast/Prostate Cancer (Chen..Linstroem Hum Genet 2019)

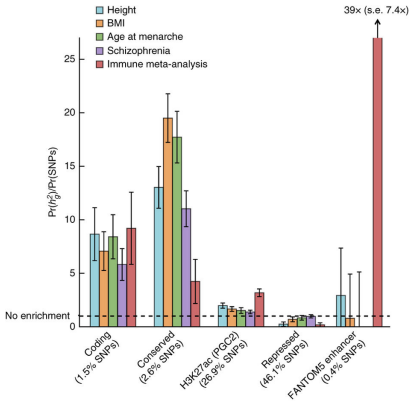


(2) Functional enrichment quantification
Option 2: SNP-heritability enrichment



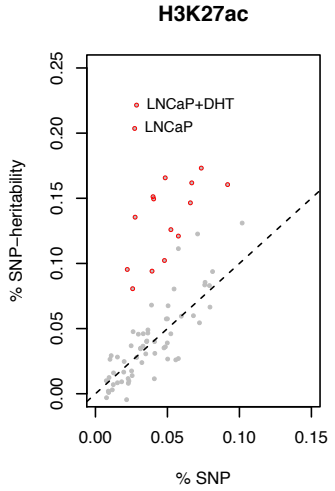
Can be estimated directly from summary GWAS (e.g., Finucane et al Nat Gen 2014)

Example of functional enrichment



- Finucane et al Nat Genet 2016

Top H3k27ac marks enriched in Prostate Cancer



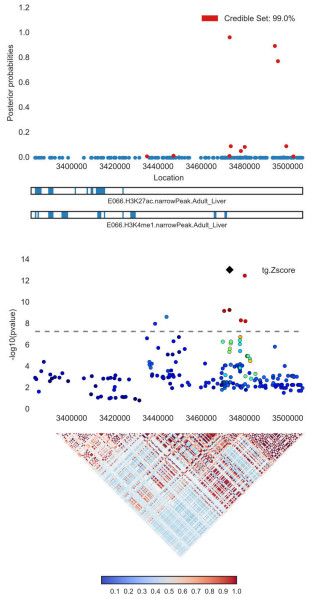
[Gusev et al NatComm 2016]

How do we put all together

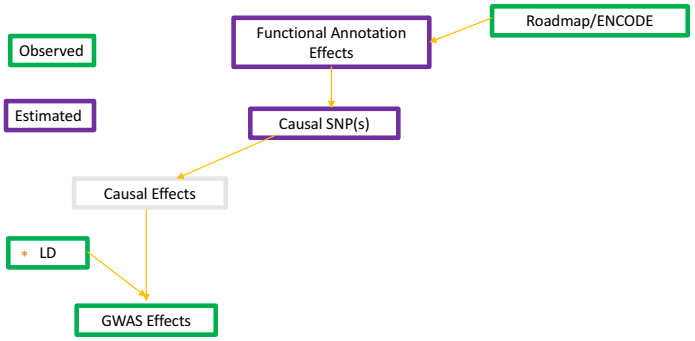
STATISTICAL FINE-MAPING

OUTPUT
- Probability of each SNP to be causal
- (optional: functional enrichment)

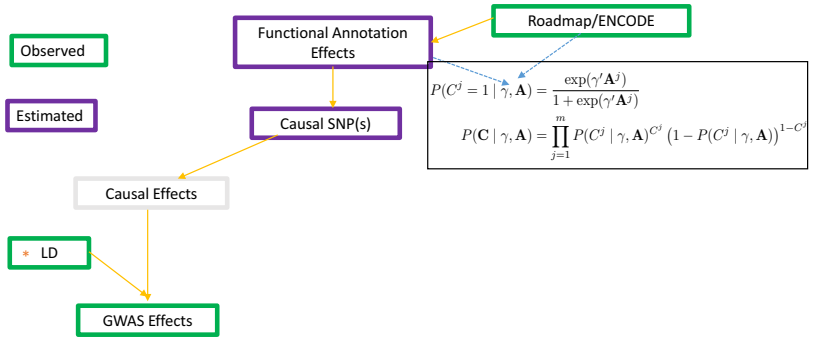
INPUT:
- Functional classes of SNPs (e.g., RoaMap/ENCODE)
- GWAS output (p-values for all SNPs)
- LD patterns (correlation structure among SNPs)



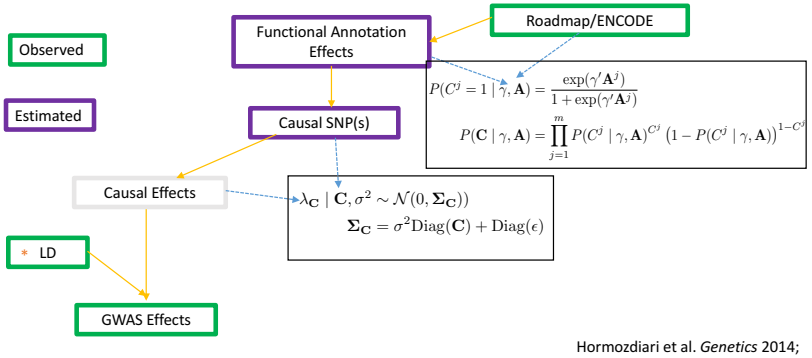
Combining this into a Bayesian Hierarchical Model: Big Picture Schematic



Combining this into a Bayesian Hierarchical Model: Big Picture Schematic

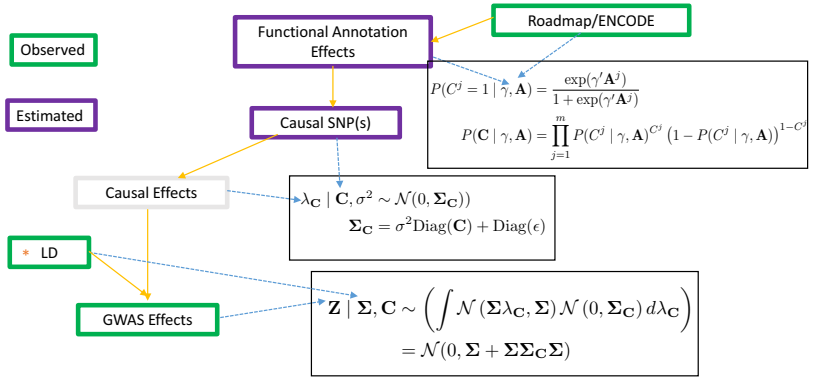


Combining this into a Bayesian Hierarchical Model: Big Picture Schematic

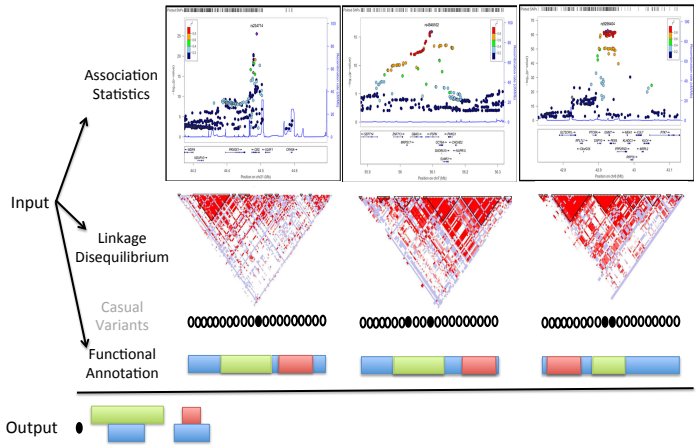


Hormozdiari et al. *Genetics* 2014;
Chen et al. *Genetics* 2015

Combining this into a Bayesian Hierarchical Model: Big Picture Schematic

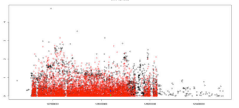


Model: visual representation



PAINTOR; RIVIERA; CAVIARBF;FINEMAP etc…

Association statistics (z-scores)

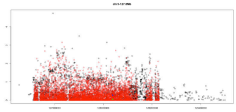


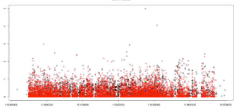
LD structure

Functional annotation

Locus 1

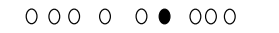
Association statistics (z-scores)



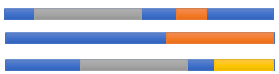


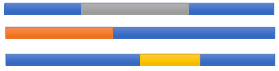
LD structure

Causal/non-causal SNPs



Functional annotation

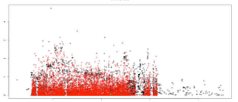




Locus 1 Locus 2

Statistical Model (N-SNPs, M functional classes)

Z: N-size vector of observed Association statistics



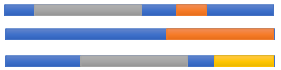
$$P(Z_j|C_j;\lambda_j) = \mathcal{N}(Z_j;\Sigma_j(\lambda_j \circ C_j),\Sigma_j) \quad \text{(p.d.f of multivariant normal)}$$

C: N-size vector 0/1 indicating causal status

$$P(C_j;\gamma) = \prod_i P(c_{ij};\gamma)$$

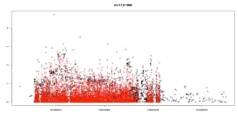
$$P(c_{ij};\gamma) = \left(\frac{1}{1+\exp(\gamma^T A_{ij})}\right)^{c_{ij}} \left(\frac{1}{1+\exp(-\gamma^T A_{ij})}\right)^{1-c_{ij}}$$

A: NxM matrix of 0/1 indicating annotation membership for every SNP



Statistical Model (N-SNPs, M functional classes)

Z: N-size vector of observed Association statistics



λ_j: effect size of SNP j

$$P(Z_j|C_j;\lambda_j) = \mathcal{N}(Z_j;\Sigma_j(\lambda_j \circ C_j),\Sigma_j) \quad \text{(p.d.f of multivariant normal)}$$

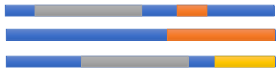
Σ_i: LD at locus i

C: N-size vector 0/1 indicating causal status

$$P(C_j;\gamma) = \prod_i P(c_{ij};\gamma)$$

$$P(c_{ij};\gamma) = \left(\frac{1}{1+\exp(\gamma^T A_{ij})}\right)^{c_{ij}} \left(\frac{1}{1+\exp(-\gamma^T A_{ij})}\right)^{1-c_{ij}}$$

A: NxM matrix of 0/1 indicating annotation membership for every SNP



γ: effect of annotation on probability of SNP to be causal

Getting output from statistical model: Bayes

$$P(\mathbf{C} \mid \mathbf{Z}, \mathbf{\Sigma}, \mathbf{A}, \gamma) = \frac{\mathcal{N}(0, \mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{\Sigma_C}\mathbf{\Sigma}) P(\mathbf{C} \mid \mathbf{A}, \gamma)}{\sum_{C \in \mathcal{C}} \mathcal{N}(0, \mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{\Sigma_C}\mathbf{\Sigma}) P(\mathbf{C} \mid \mathbf{A}, \gamma)}$$

- Causal vector formulation gives the flexibility to model **multiple causal variants** at any risk locus
- Iterative procedure to estimate annotation effects across all fine mapping loci using Maximum Likelihood and EM.

Practical considerations

Fine-mapping credible sets are miscalibrated when assuming a single causal variant.

Causal Assump.	Method	Functional Annotations	Causals Identified	90% Credible Set Size
Single	Maller et al.	None	64.2%	265.0
Multiple	CAVIAR	None	91.9%	510.3

90% credible set = set of SNPs that consume 90% of the total posterior probability mass

Maller et al. *Nat Gen* 2012;
Hormozdiari et al. *Genetics* 2014;

Leveraging functional enrichment improves fine-mapping resolution

Causal Assump.	Method	Functional Annotations	Causals Identified	90% Credible Set Size
Single	Maller et al.	None	64.2%	265.0
Multiple	CAVIAR	None	91.9%	510.3
Multiple	**PAINTOR**	**Included**	**91.2%**	**393.7**

up to ~30% improvement

90% credible set = set of SNPs that consume 90% of the total posterior probability mass

Maller et al. *Nat Gen* 2012; Hormozdiari et al. *Genetics* 2014; Kichaev et al. *Plos Genetics* 2014

Statistical fine-mapping identifies bona-fide causal variants

LETTERS

A thrifty variant in *CREBRF* strongly influences body mass index in Samoans

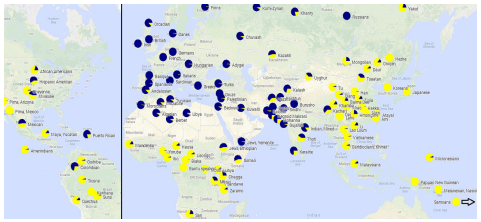
Ryan L Minster[1,13], Nicola L Hawley[2,13], Chi-Ting Su[1,12,13], Guangyun Sun[3,13], Erin E Kershaw[4], Hong Cheng[3], Olive D Buhule[5,12], Jerome Lin[3], Muagututi'a Sefuiva Reupena[6], Satupa'itea Viali[7], John Tuitele[8], Take Naseri[9], Zsolt Urban[1,14], Ranjan Deka[3,14], Daniel E Weeks[1,5,14] & Stephen T McGarvey[10,11,14]

*"Bayesian fine-mapping with PAINTOR strongly supported following up the missense variant. The two variants in the region with the highest posterior probability (PP) of being causal were **rs373863828** (PP = 0.80) and rs150207780 (PP = 0.22); when Encyclopedia of DNA Elements (ENCODE) functional annotation was included, these probabilities increased to 0.92 and 0.34, respectively"*

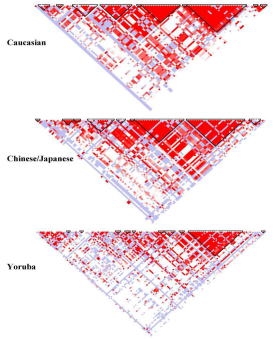
What if we have multiple ancestries? Can we still perform statistical fine-mapping?

Divergent population histories give rise to unique genetic backgrounds

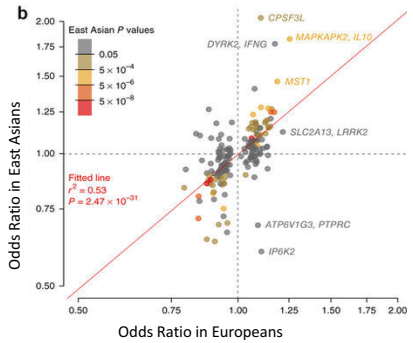
Allele Frequencies



LD Patterns



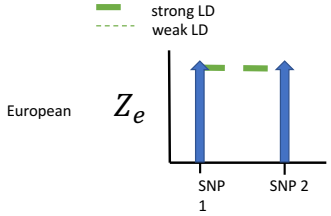
GWAS loci tend to replicate in non-European populations → shared causal variants



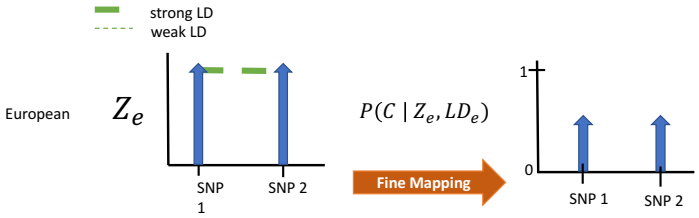
Liu et al. *Nat Genet* 2015

see also: Marigorta et al. *Plos Genet*. 2013
Zaitlen et al. *Am J Hum Genet.* 2011, PAGE consortium Nature 2019

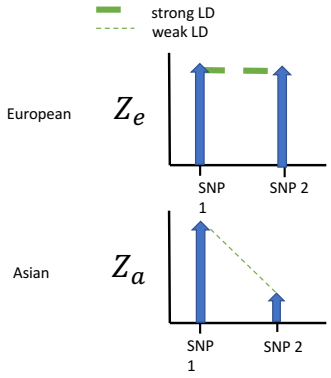
Leveraging genetic diversity to improve fine-mapping



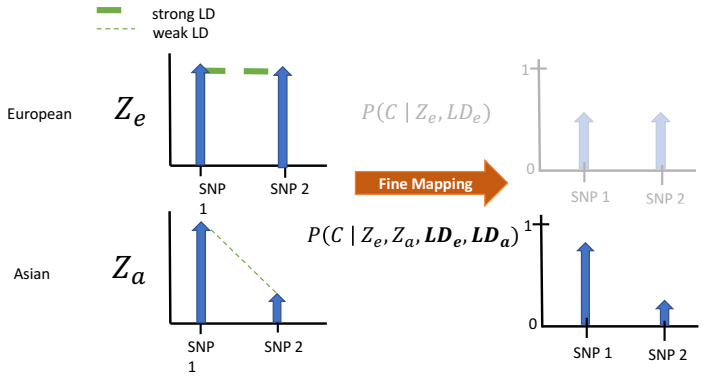
Leveraging genetic diversity to improve fine-mapping



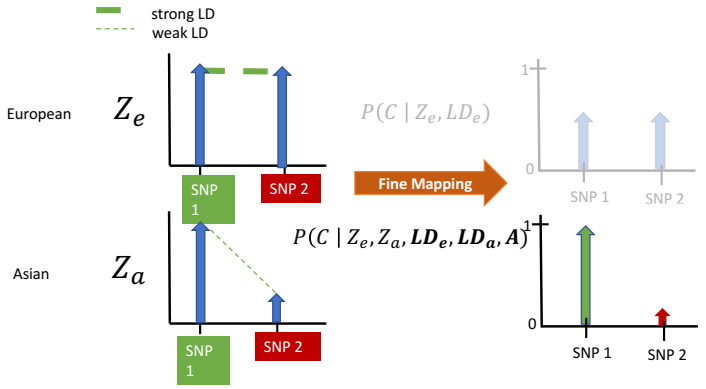
Leveraging genetic diversity to improve fine-mapping



Leveraging genetic diversity to improve fine-mapping



Leveraging genetic diversity and functional annotations to improve fine-mapping



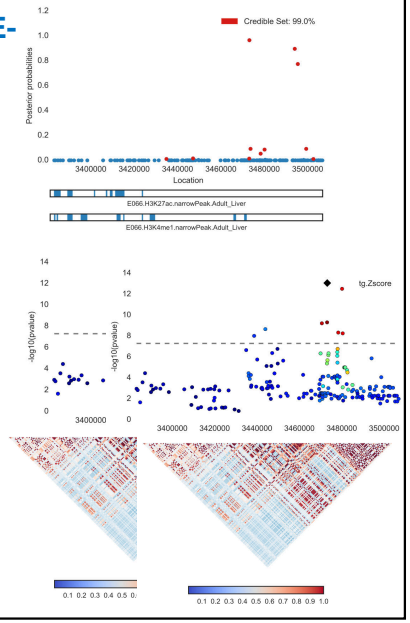
MULTI-ETHNIC STATISTICAL FINE-MAPING

OUTPUT

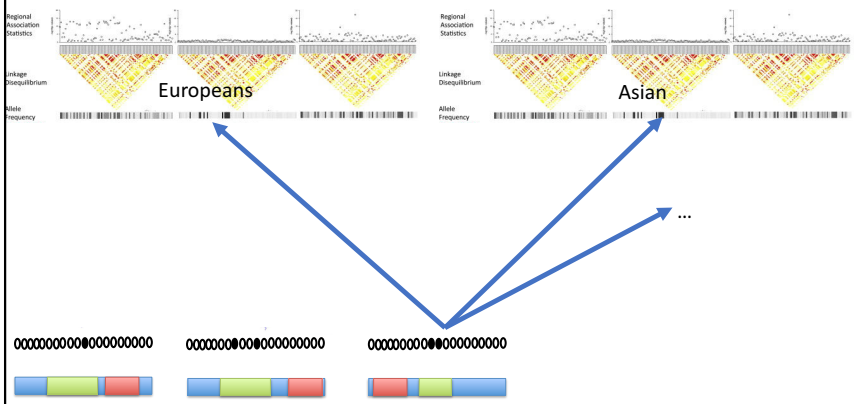
- Probability of each SNP to be causal
- (optional: functional enrichment)

INPUT:

- Functional classes of SNPs (e.g., RoaMap/ENCODE)
- GWAS output for each ancestry (p-values for all SNPs)
- LD patterns for each ancestry (correlation structure among SNPs)



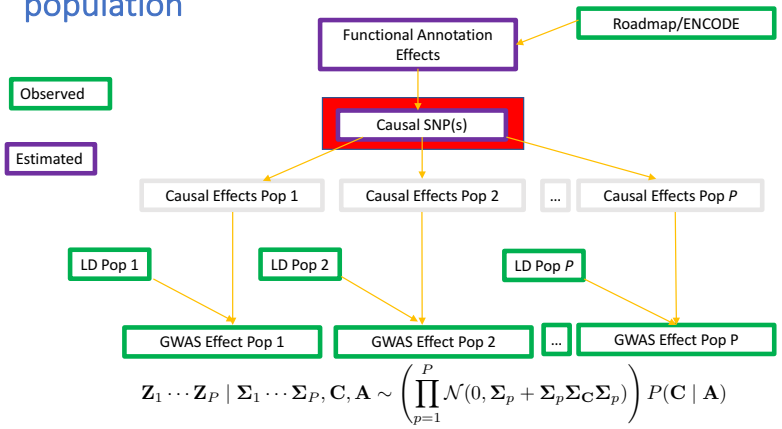
Integrating functional annotation data in trans-ethnic fine-mapping



Main assumption: same causal variants with causal probability from functional data

[Kichaev et al AJHG 2015; Morris Gen Epi 2011; Liu et al AJHG 2016; etc...]

Generalizing integrative fine-mapping from a single population



$$\mathbf{Z}_1 \cdots \mathbf{Z}_P \mid \mathbf{\Sigma}_1 \cdots \mathbf{\Sigma}_P, \mathbf{C}, \mathbf{A} \sim \left(\prod_{p=1}^{P} \mathcal{N}(0, \mathbf{\Sigma}_p + \mathbf{\Sigma}_p \mathbf{\Sigma}_\mathbf{C} \mathbf{\Sigma}_p) \right) P(\mathbf{C} \mid \mathbf{A})$$

Trans-ethnic fine-mapping reduces the size of the credible causal sets

LETTER

doi:10.1038/nature12873

Genetics of rheumatoid arthritis contributes to biology and drug discovery

A list of authors and their affiliations appears at the end of the paper

N Europeans ≈ 68K
N Asians ≈ 36K

Final Annotations selected by model:

DHS (Skin Keratinocytes, Th2, and B-lymphocytes), Immune Enhancers, Exonic regions

Average Size of 90% credible set

	Data	No Annotations	With Annotations
Single Population	Asian	35.2	31.9
	European	32.0	28.7
Multi Population	Meta analysis	28.5	25.0
	Asian, European	24.0	21.7

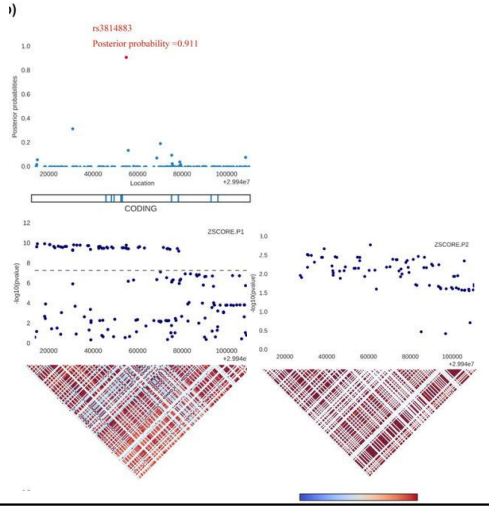
+32%

Example of multi-ethnic fine-mapping Europeans+East Asians

Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia

Zhiqiang Li, Jianhua Chen [...] Yongyong Shi

Nature Genetics 49, 1576–1583 (2017) | Download Citation

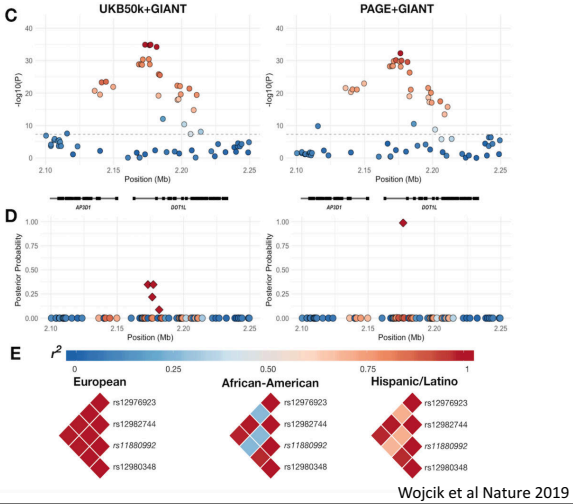


Example of multi-ethnic fine-mapping

Genetic analyses of diverse populations improves discovery for complex traits

Genevieve L. Wojcik, Mariaelisa Graff, [...] Christopher S. Carlson

Nature 570, 514–518 (2019) | Download Citation



Wojcik et al Nature 2019

Big picture

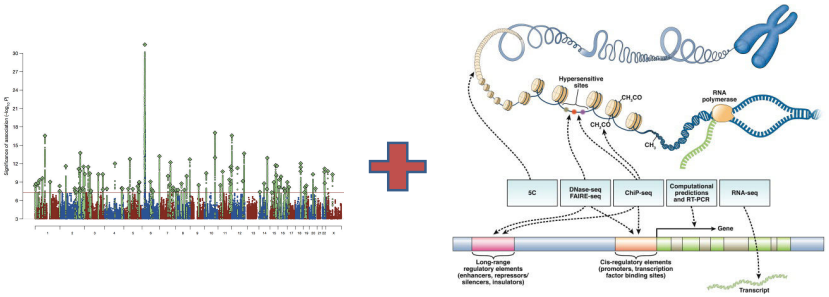
GWAS Marginal Summary Statistics + Linkage Disequilibrium Reference Panel + Functional Annotation

LD Score Regression (or many others) → (1) Find functional annotations of interest

PAINTOR/Susie (or many others) → (2) Estimate SNP Probabilities of Causality

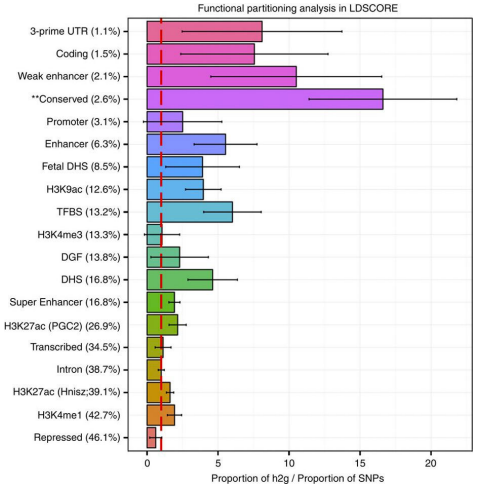
Questions

RECAP



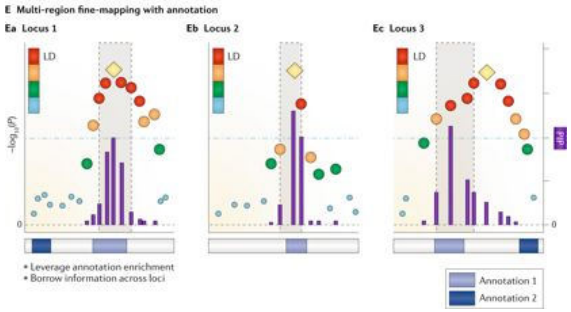
- Functional enrichment (LDSCORE regression, Finucane et al NG 2016)

Identify functional elements of interest



(hippocampal volume) Hibar et al Nat Comm 2017

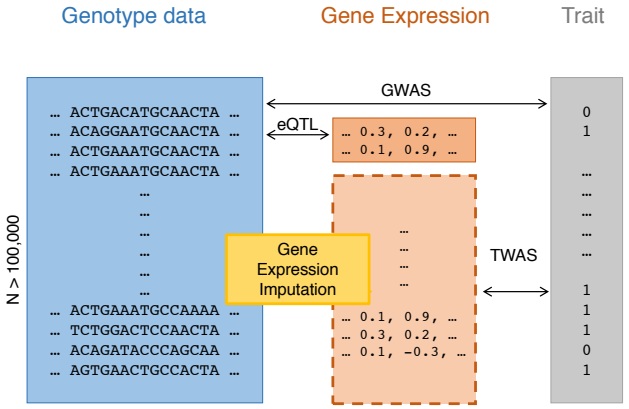
Prioritize genetic variants likely to be biologically causal



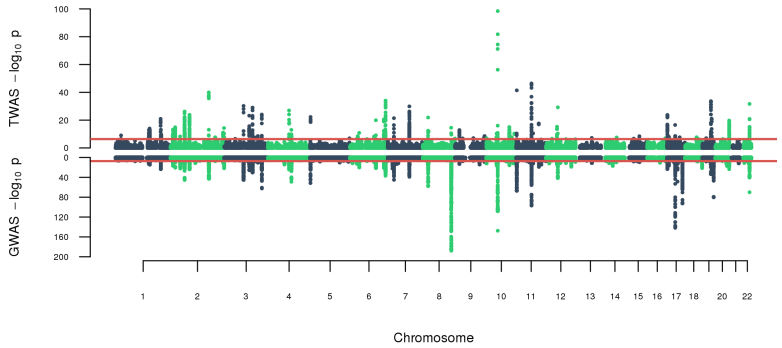
Schaid et al Nat Rev Genet 2018

Questions

Statistical fine-mapping at gene-level in TWAS/PrediXscan/etc…



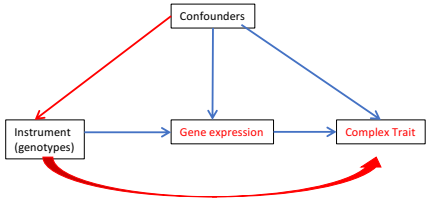
TWAS example



Mancuso et al. Nat Comm 2018

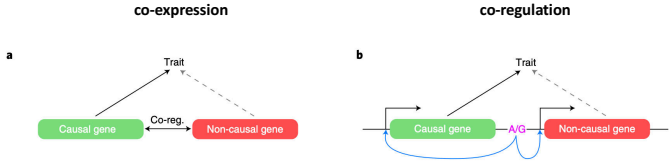
Are all TWAS significant genes causal?

TWAS = Mendelian Randomization under strong assumptions



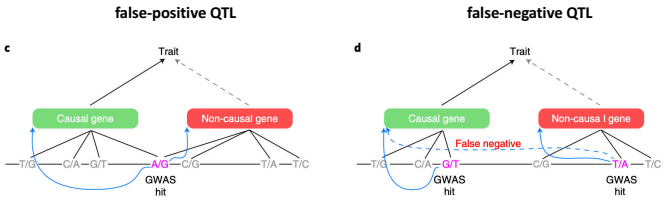
- Instrument must be associated with the exposure
- **Instrument must not have an association with outcome, except through the exposure**
- Instrument is not related to measured or unmeasured confounders

non-causal TWAS associations due to co-regulation



Wainberg et al. 2019 *Nat Genet*

non-causal TWAS associations due to tagging



Wainberg et al. 2019 *Nat Genet*

TWAS interpretation

- Two sample Mendelian Randomization test
  - Estimate **mediating** effect of gene expression **under very strong assumptions!**
  - Zhu et al. Nat Gen 2016; Barfield et al Gen Epi 2018, …
- Test of association (genetic covariance) expression and trait
  - **Similarity** between trait / GE at local genetics
  - Gusev et al. Nat Gen 2016; Mancuso et al. AJHG 2017, Wainberg et al Nat Gen 2019; Mancuso et al Nat Genet 2019

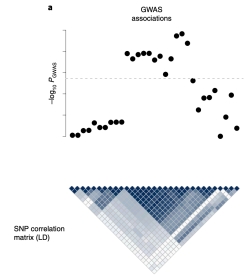
Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg[1], Nasa Sinnott-Armstrong[2], Nicholas Mancuso[3], Alvaro N. Barbeira[4], David A. Knowles[5,6], David Golan[2], Raili Ermel[7], Arno Ruusalepp[7,8], Thomas Quertermous[9], Ke Hao[10], Johan L. M. Björkegren[8,10,11,12], Hae Kyung Im[4], Bogdan Pasaniuc[3,13,14], Manuel A. Rivas[15] and Anshul Kundaje[1,2]

But we really want causal genes…

Probabilistic fine-mapping of TWAS

GWAS: Fine-mapping multiple associated SNPs in LD at a locus

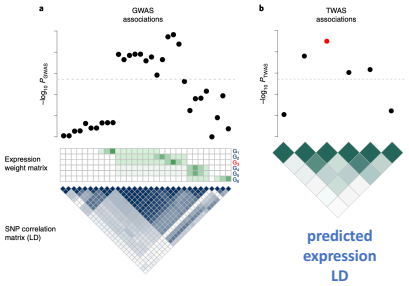


Mancuso et al. 2019 *Nat Genet*

Probabilistic fine-mapping of TWAS

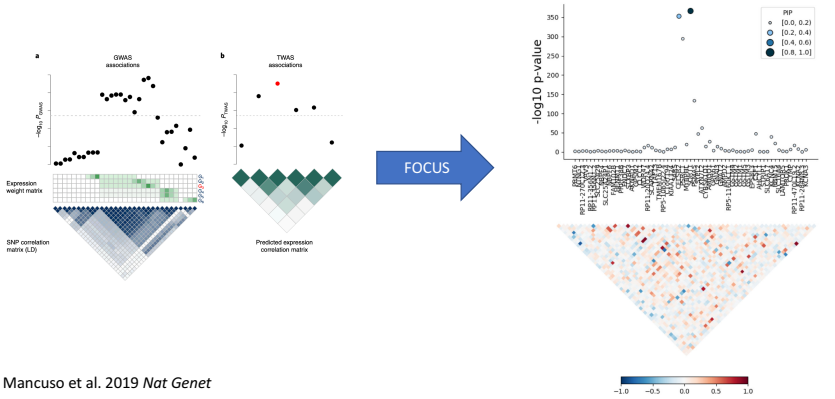
GWAS: Fine-mapping multiple associated SNPs in LD at a locus

TWAS: Fine-mapping multiple gene models ➔ posterior probability for each gene ➔ **credible sets** (FOCUS)



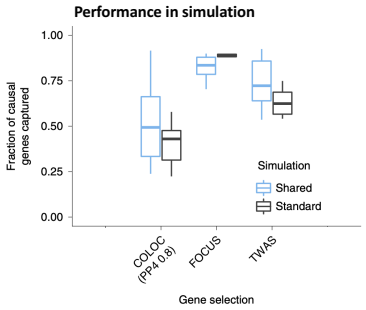
Mancuso et al. 2019 *Nat Genet*

Example of TWAS fine-mapping



Mancuso et al. 2019 *Nat Genet*

FOCUS accurately identifies credible genes



Performance in real data

Table 1 | Summary of gene-based fine-mapping in lipid GWAS risk regions

Lipid trait	GWAS risk regions	GWAS risk regions with TWAS-significant genes	TWAS genes at risk regions	Genes in 90%-credible sets
HDL	43	18	64	30
LDL	36	20	56	40
Total cholesterol	51	24	73	53
Triglycerides	30	13	33	25
Overall	160	75	226	148
Unique	89	46	146	100

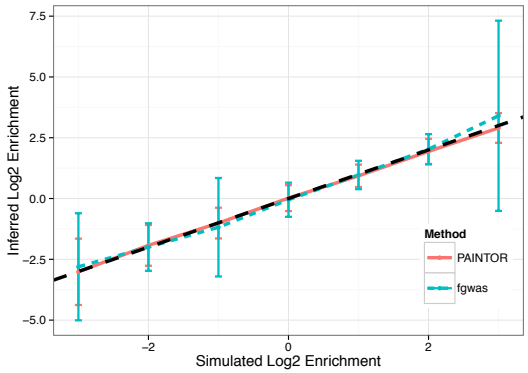
A GWAS risk region is defined to be an LD block defined by LDetect[41] harboring at least one genome-wide-significant SNP ($P < 5 \times 10^{-8}$) reported in ref. [16]. A TWAS gene is a gene whose predicted expression reaches transcriptome-wide significance of $P < 0.05/15,277$. Overall results are presented as total counts across traits with unique results discarding repeated elements.

1.5 genes on average in the 95% credible set

Mancuso et al. 2019 *Nat Genet*

Questions

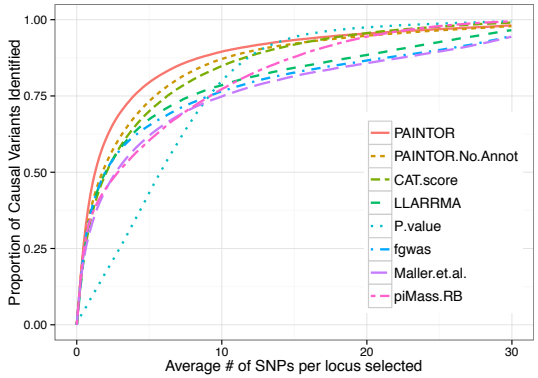
Unbiased estimation of enrichment of causal variants in simulations



Hapgen simulations
100 loci 10Kb
Europeans 100G
h2=0.25
N=5,000

[Kichaev et al. Plos Genetics 2014]

Functional data improves fine-mapping accuracy

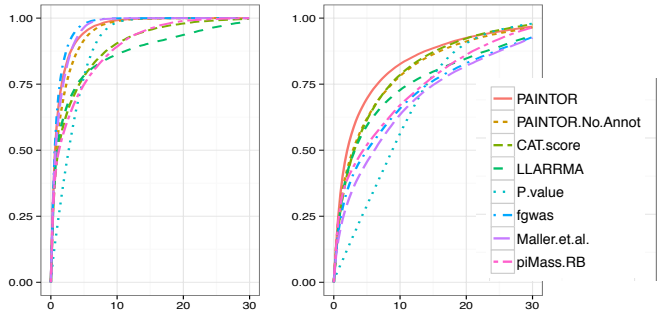


[Kichaev et al. Plos Genetics 2014]

Functional data improves fine-mapping accuracy

Loci with 1 causal

Loci with >1 causal



[Kichaev et al. Plos Genetics 2014]

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)
Probabilities (single causal per locus) (Maller et al NG'12)	2.7 (172.4)	25.0 (1616.1)

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)
Probabilities (single causal per locus) (Maller et al NG'12)	2.7 (172.4)	25.0 (1616.1)
PAINTOR (multiple causals)	1.7 (108.9)	11.4 (734.4)

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)
Probabilities (single causal per locus) (Maller et al NG'12)	2.7 (172.4)	25.0 (1616.1)
PAINTOR (multiple causals)	1.7 (108.9)	11.4 (734.4)
PAINTOR (multiple causals; with ENCODE)	1.2 (78.7)	9.7 (625.6)

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

How many variants to functionally test to find causal variants?

Method (assumptions)	50% of all causals	90% of all causals
p-value	5.7 (369.5)	12.3 (796.9)
Probabilities (single causal per locus) (Maller et al NG'12)	2.7 (172.4)	25.0 (1616.1)
PAINTOR (multiple causals)	1.7 (108.9)	11.4 (734.4)
PAINTOR (multiple causals; with ENCODE)	1.2 (78.7)	9.7 (625.6)
PAINTOR (multiple causals; true ENCODE prior)*	*1.2 (77.9)*	*9.4 (610.6)*

Simulations:
100 loci 10Kb
Europeans 100G
h2=0.25
N=10,000

Assuming 1 causal per locus yields miss-calibrated causal sets

Causal Set	Method	Annotations	# Causals	# SNPs
90%	Maller et al.	-	63.2	325.5
95%	Maller et al.	-	68.8	409.8
99%	Maller et al.	-	76.7	579.4

[Kichaev et al. Plos Genetics 2014]

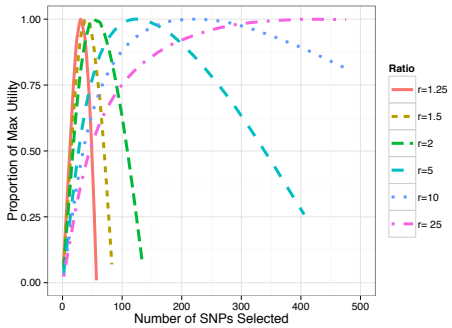
Assuming 1 causal per locus yields miss-calibrated causal sets

Causal Set	Method	Annotations	# Causals	# SNPs
90%	Maller et al.	-	63.2	325.5
	PAINTOR	-	90.1	569.4
	PAINTOR	+	89.9	461.0
95%	Maller et al.	-	68.8	409.8
	PAINTOR	-	95.8	741.4
	PAINTOR	+	95.6	628.8
99%	Maller et al.	-	76.7	579.4
	PAINTOR	-	101.5	1118.9
	PAINTOR	+	101.6	1001.8

[Kichaev et al. Plos Genetics 2014]

How many SNPs to follow-up?

- Depends on ratio of benefit of finding a causal variant to cost of testing a variant (r=B/C)
  - r=10 ➔the benefit of finding a causal outweighs 10 times the cost of testing 1 SNP

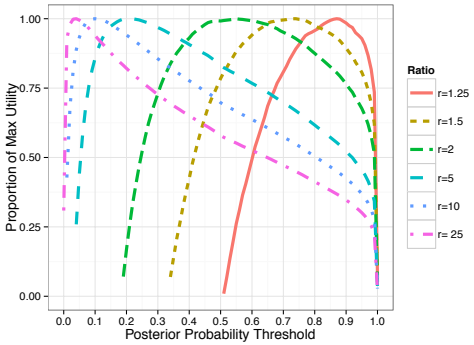


r=10, ~3.8 SNPs per locus to identify ~72.6% of all causals
r=100, ~13.1 SNPs per locus to identify ~96.2% of all causals

[Kichaev et al. Plos Genetics 2014]

How many SNPs to follow-up?

- Thresholding on posterior probability of causality gives a principled way of maximizing utility.



[Kichaev et al. Plos Genetics 2014]

Other considerations

- Locus size (function of LD):
  - Increasing locus size increases performance of PAINTOR vs existing methods
    - 10Kb: 27.4 vs 11.4 variants to follow-up to find 90% of causals
    - 50Kb: 110.7 vs 24.1 variants to follow-up to find 90% of causals
- Causal not in the data:
  - Median distance in Kb increases by ~6%
    - 21.6 vs 22.0 median Kb to the top 10 SNPs
    - 1.6 vs 4.0 minimum Kb to the top 10 SNPs
- Sample size
  - 19.0, 12.5, 10.8 variants per locus to find 90% of all causals for 2.5k, 5k and 10k samples

Yale

From cross-phenotype associations to pleiotropy in human genetic studies

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

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

Yale SCHOOL OF PUBLIC HEALTH

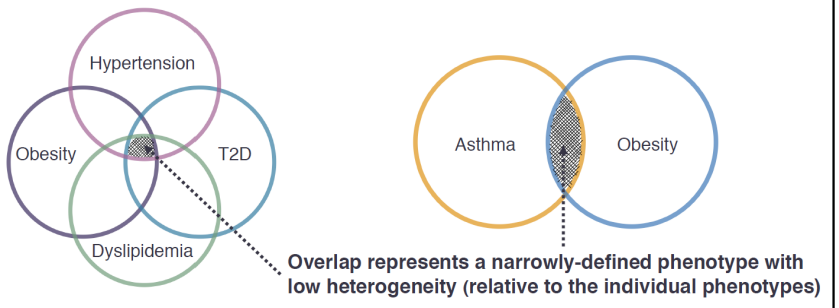
Pleiotropy

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

Pleiotropy and 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

Pleiotropy and disease comorbidity



Pleiotropy and 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

Pleiotropy and 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)

Expression of autism spectrum and schizophrenia in patients with a 22q11.2 deletion

Jacob A.S. Vorstman [a,*], Elemi J. Breetvelt [a], Kirstin I. Thode [b], Eva W.C. Chow [c,d], Anne S. Bassett [c,d]

[a] Rudolf Magnus Institute of Neuroscience, Department of Psychiatry, University Medical Center Utrecht, Utrecht, The Netherlands
[b] Department of Psychiatry, Malcolm Grow Medical Center, Joint Base Andrews, Andrews AFB, MD, USA
[c] Clinical Genetics Research Program, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[d] Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada

Schizophrenia Research 143 (2013) 55–59

Table 1
Autistic symptoms and probable ASD during childhood, assessed in 77 adults with 22q11.2DS, comparing those with and without schizophrenia.

	Total (n=77)	22q11.2DS-SZ (n=36)	22q11.2DS-Co (n=41)	p				
Mean SRS T-score (95% CI)	73.1 (69.7–76.6)	72.4 (67.3–78.0)	73.7 (68.9–78.6)	0.36[a]				
Subjects categorized as probable ASD	n (%)	n (%)	n (%)	p	Uncorrected OR (95% CI)	p	Corrected OR (95% CI)	p
SRS T score cut-off 60	59 (76.6%)	27 (75.0%)	32 (78.0%)	0.75[b]	0.84 (0.29–2.43)	0.75	0.36 (0.085–1.52)	0.17[c]
SCQ cut-off 15	13 (16.9%)	3 (8.3%)	10 (24.4%)	0.06[b]	0.28 (0.07–1.12)	0.07	0.20 (0.03–1.46)	0.11[d]
SCQ cut-off 12	27 (35.1%)	13 (36.1%)	14 (34.1%)	0.86[b]	1.10 (0.44–2.78)	0.86	1.53 (0.49–5.27)	0.49[d]

CI= confidence interval.
[a] Mann–Whitney–Wilcoxon.
[b] Chi-square.
[c] Corrected for age and IQ.
[d] Corrected for gender, age, IQ.

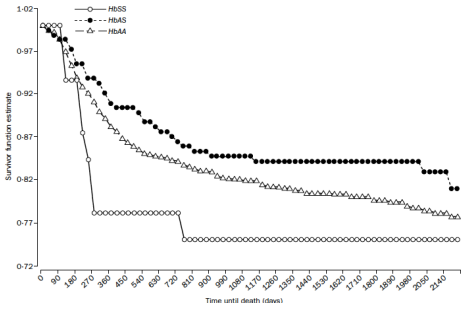
A B S T R A C T

Background: Copy number variants (CNVs) associated with neuropsychiatric disorders are increasingly being identified. While the initial reports were relatively specific, i.e. implicating vulnerability for a particular neuropsychiatric disorder, subsequent studies suggested that most of these CNVs can increase the risk for more than one neuropsychiatric disorder. Possibly, the different neuropsychiatric phenotypes associated with a single genetic variant are really distinct phenomena, indicating pleiotropy. Alternatively, seemingly different disorders could represent the same phenotype observed at different developmental stages or the same underlying pathogenesis with different phenotypic expressions.
Aims: To examine the relation between autism and schizophrenia in patients sharing the same CNV.
Method: We interviewed parents of 78 adult patients with the 22q11.2 deletion (22q11.2DS) to examine if autistic symptoms during childhood were associated with psychosis in adulthood. We used Chi-square, T-tests and logistic regression while entering cognitive level, gender and age as covariates.
Results: The subgroup of 22q11.2DS patients with probable ASD during childhood did not show an increased risk for psychosis in adulthood. The average SRS scores were highly similar between those with and those without schizophrenia.
Conclusions: ASD and schizophrenia associated with 22q11.2DS should be regarded as two unrelated, distinct phenotypic manifestations, consistent with true neuropsychiatric pleiotropy. 22q11.2DS can serve as a model to examine the mechanisms associated with neuropsychiatric pleiotropy associated with other CNVs.

Protective effects of the sickle cell gene against malaria morbidity and mortality

Michael Aidoo, Dianne J Terlouw, Margarette S Kolczak, Peter D McElroy, Feiko O ter Kuile, Simon Kariuki, Bernard L Nahlen, Altaf A Lal, Venkatachalam Udhayakumar

Lancet 2002; **359:** 1311–12



	Crude incidence/1000 person-months			Adjusted relative risk (95% CI)			
	HbAA	*HbAS*	*HbSS*	*HbAS vs HbAA*	p	*HbSS vs HbAA*	p
Severe malaria anaemia episodes	4·0	2·0	1·5	0·40 (0·30–0·60)	0·0001	0·29 (0·1–0·9)	0·04
All severe anaemia episodes	8·8	6·8	7·5	0·61 (0·46–0·80)	0·0006	0·63 (0·35–1·2)	0·15
(Hb <6 g/dL plus any parasitaemia)							
High density parasitaemia episodes	20	17·3	15·8	0·73 (0·65–0·84)	0·0001	0·52 (0·36–0·74)	0·0002

Hb=haemoglobin. *Hb*SS was associated with lower parasite incidence than HbAA haemoglobin levels and parasitaemia were determined using routine monthly finger-prick blood samples and samples collected any time the children were reported ill. All data points collected monthly for the entire time children participated in the study were used in data analyses unless indicated otherwise. Only birthweight among the various covariates considered (same as for survival analysis) was controlled for in the final model.

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

A practitioners' guide for studying pleiotropy in genetic epi studies

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

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

Salinas YD, Wang Z, DeWan AT.

Abstract

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

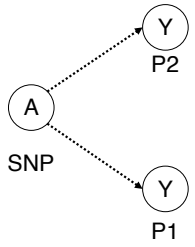
KEYWORDS: genetic epidemiology; mediation analysis; pleiotropy

Guidelines for generating robust statistical evidence of pleiotropy

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

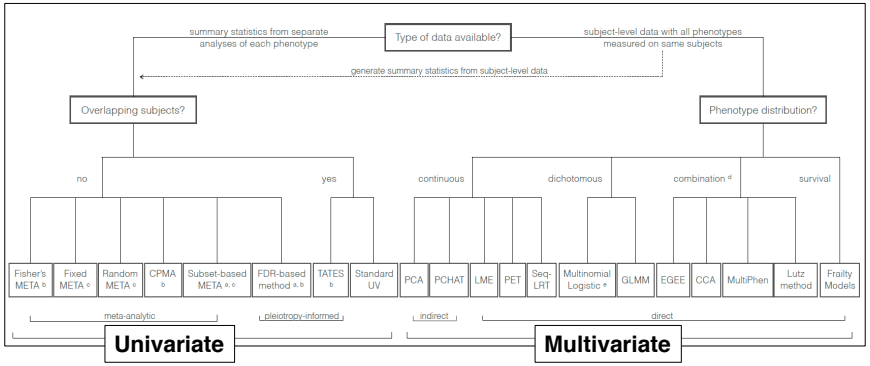
Cross-phenotype (CP) associations

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.

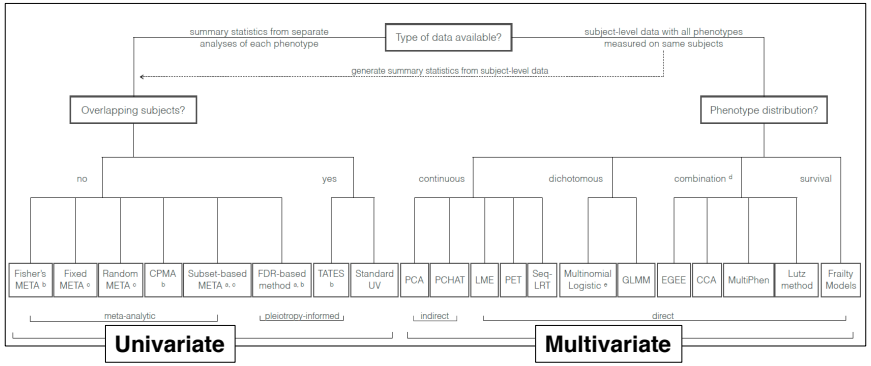
Analytic options for discovery of CP associations



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*

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)

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

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!

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 <u>each</u> model.

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

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

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

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

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

Statistical power of multi-trait methods

- For *quantitative* trait methods, it has been shown that:
  - Multivariate analyses achieve greater power than univariate analyses both in the presence **(Allison 1998)** and absence of cross-trait genetic correlation or pleiotropy **(Galesloot 2014)**
- Therefore, joint analysis of *quantitative* phenotypes has the potential to enhance the statistical power of genetic studies.

Statistical power of multi-trait methods

- With this potential for greater statistical power, multivariate methods could contribute to the investigation of the 'missing heritability' of complex diseases.
- However, it is unknown whether the trends observed for *quantitative* traits also hold for methods that can analyze multiple *disease* (case-control) phenotypes.
- Understanding the performance of these methods is essential to their successful application to real data.

Objective

- To evaluate the relative statistical power of methods for analysis of two disease (case/control) phenotypes in the presence and absence of pleiotropy using simulated genotype and phenotype data.

Data Simulation

- Genotypes were simulated for a bi-allelic SNP with **MAF = 0.20** by sampling two alleles independently from a binomial distribution.
- Genotypes (coded as 0/1/2) are the sum of the two alleles.

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%

Methods evaluated

1. **Standard univariate approach**
   - **models fitted**
     $\text{logit}[E(Y_{i1})] = \beta_0 + \beta_1 X_i$
     $\text{logit}[E(Y_{i2})] = \beta_0 + \beta_1 X_i$
   - **p-value extracted**
     - the minimum of the two univariate p-values
2. **Reversed ordinal logistic regression (MultiPhen)**
   - **model fitted**
     $\text{logit}[E(X_i \leq c)] = \alpha_c + \beta_1 Y_{i1} + \beta_2 Y_{i2}$, for c = 1, 2, or 3 genotype categories
   - **p-value extracted**
     - the p-value for a Likelihood Ratio Test for model fit, evaluating the null hypothesis that $\beta_1 = \beta_2 = 0$
3. **Generalized estimating equations (GEEs)**
   - **model fitted**
     $\text{logit}[E(Y_{ij})] = \beta_0 + \beta_1 j_i + \beta_2 X_i + \beta_{12} X_i j_i$
   - **p-value extracted**
     - the p-value for the test of the hypothesis that $\beta_2 = \beta_{12} = 0$
4. **Generalized linear mixed models (GLMMs)**
   - **model fitted**
     $\text{logit}[E(Y_{ij})] = \beta_0 + \beta_1 j_i + \beta_2 X_i + \beta_{12} X_i j_i + b_{ij}$
   - **p-value extracted**
     - the p-value for the test of the hypothesis that $\beta_2 = \beta_{12} = 0$

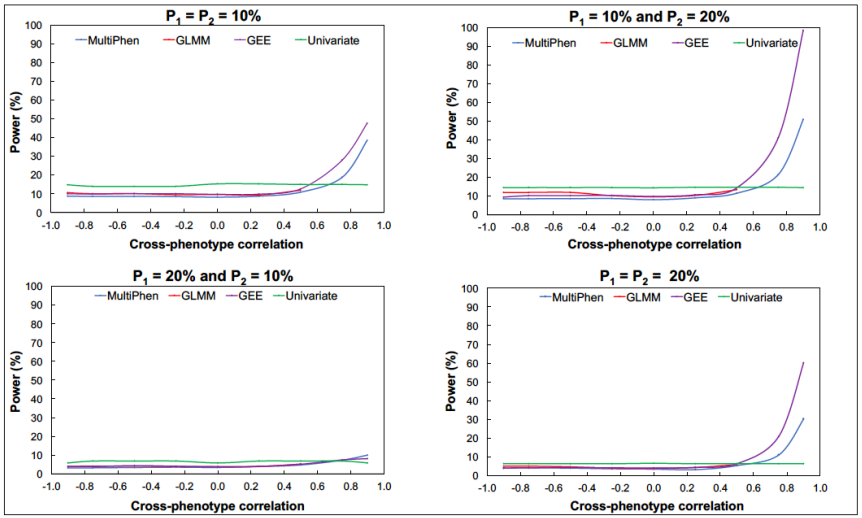
* Y_{ij} represent the case/control status of the i^{th} subject, measured for phenotypes j = 1 or 2; X_i are the individual genotypes; b_{ij} are the random effects correlated within the i^{th} subject; and j_i is an indicator variable for the phenotypes (coded as 0/1).

Power

- We defined power as the percentage of the 10,000 replicates for which the extracted p-value for a given scenario was smaller than a genome-wide significance level of 5×10^{-8} .

PLEIOTROPY ABSENT

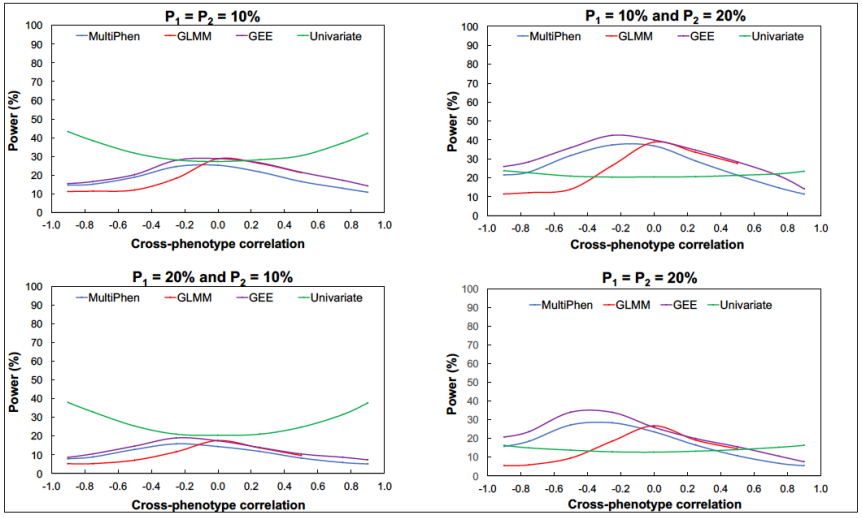
Figure 1. Power when one phenotype (Y_1) is associated with the SNP ($h_1^2 = 0.1\%$; $h_2^2 = 0\%$) [a]



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

PLEIOTROPY PRESENT
equal effect sizes

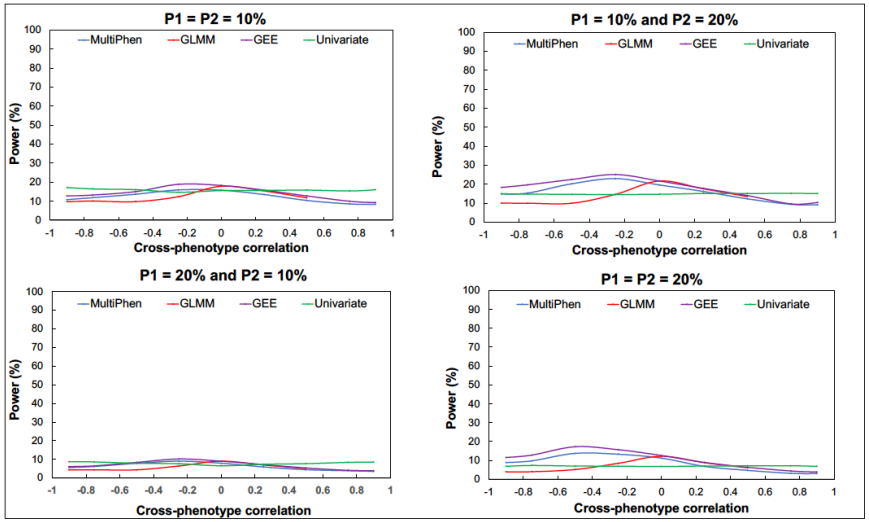
Figure 2. Power when both phenotypes are associated with the SNP ($h_1^2 = h_2^2 = 0.1\%$) [a]



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

PLEIOTROPY PRESENT
unequal effect sizes

Figure 3. Power when both phenotypes are associated with the SNP (h_1^2 = 0.1%, h_2^2=0.05%) [a]



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

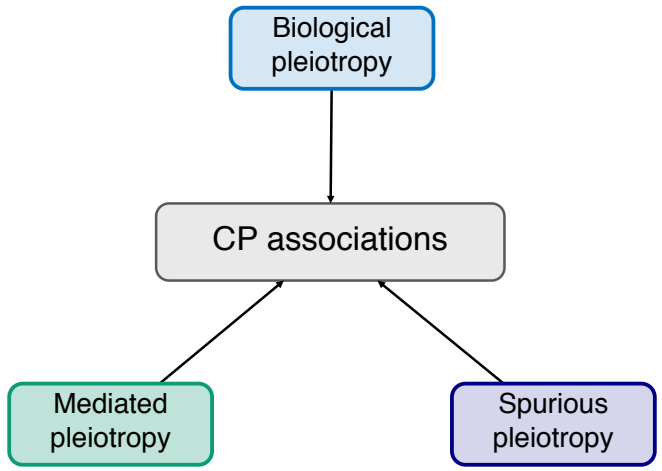
Conclusions

- The performance of the univariate approach appeared to complement that of multivariate methods, with notable patterns:
  - **in the absence of pleiotropy**, multivariate methods had better performance **for $r_{Y1,Y2} > 0.5$** while univariate methods had better performance **for $r_{Y1,Y2} < 0.5$**
  - **in the presence of pleiotropy (positive genetic correlation)**, the multivariate approach lost power **for $r_{Y1,Y2} > 0$**, while the univariate approach gained power across this range of values
- Thus, to improve GWAS discovery, it may be beneficial to use univariate and multivariate approaches in parallel.

31

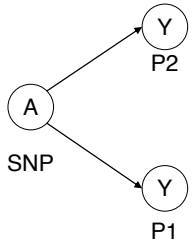
Problem: CP associations need not be indicative of pleiotropy

32



Biological pleiotropy

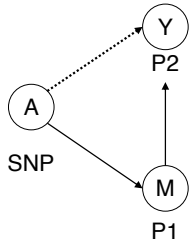
Independent associations between a genetic locus (A) and multiple phenotypic outcomes (Y)



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

Mediated pleiotropy

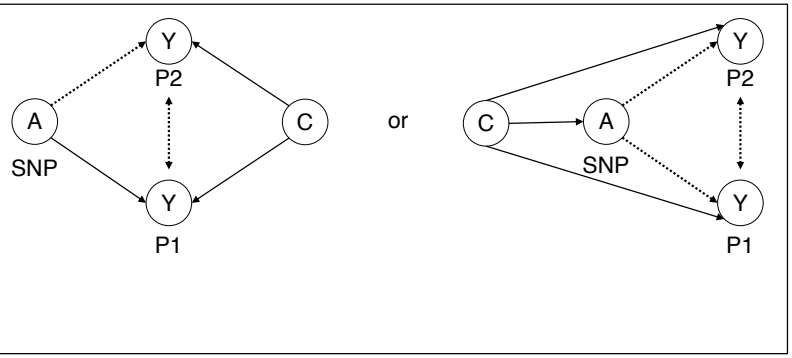
Association between a genetic locus (A) and an intermediate phenotype (M) that causes a second phenotypic outcome (Y)



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.

Spurious pleiotropy

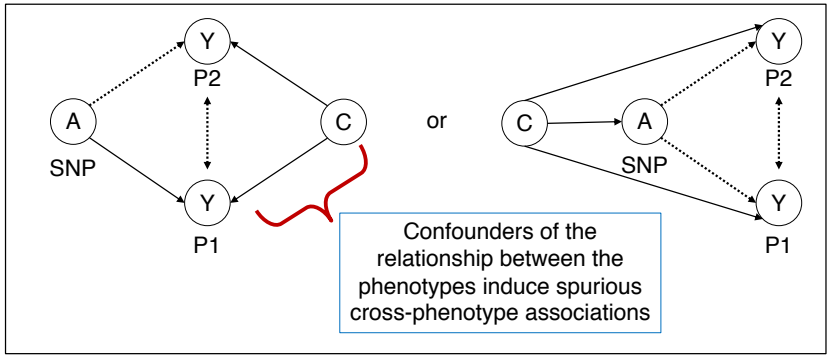
Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes



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

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

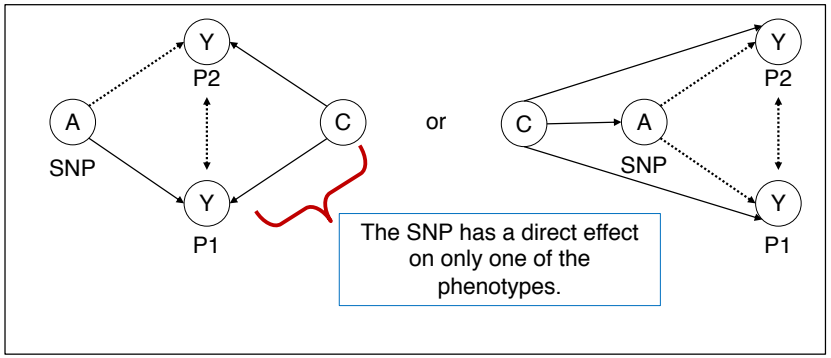


Confounders of the relationship between the phenotypes induce spurious cross-phenotype associations

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

Spurious pleiotropy

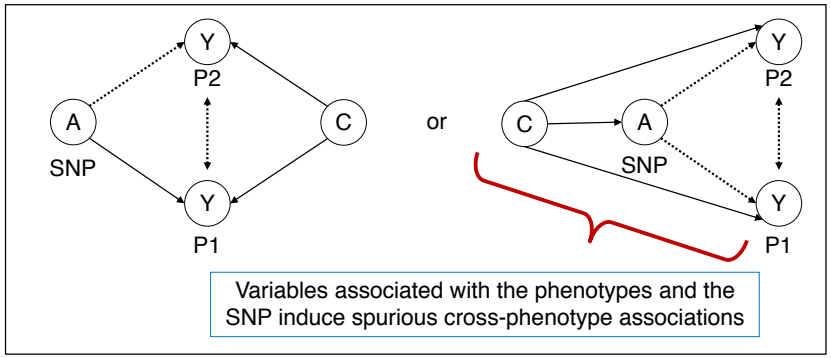
Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes



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

Spurious pleiotropy

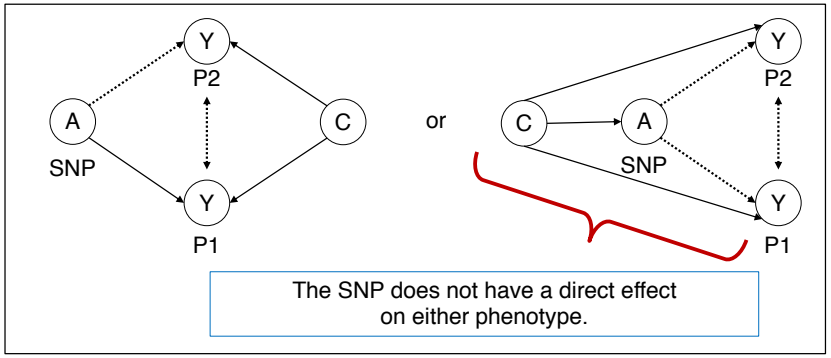
Artifactual associations with multiple phenotypes due to issues related to study design, confounding, or associations with markers in strong linkage disequilibrium* with multiple causal variants in different genes



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

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.

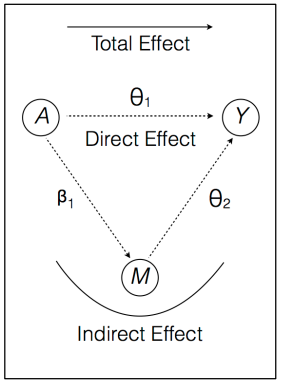
Guidelines for generating robust statistical evidence of pleiotropy

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

41

Mediation analysis provides a tool for dissecting CP associations

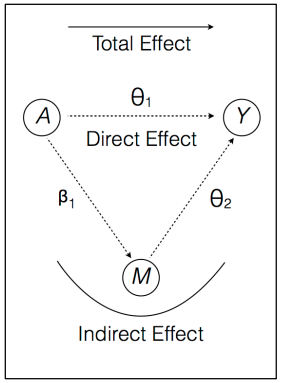
- Mediation analysis decomposes the **total effect** of the SNP (*A)* on a phenotypic outcome (*Y*) into:
  - **Direct effect:** effect of *A* on *Y* that occurs independently of an intermediate phenotype (*M)*
  - **Indirect effect:** effect of *A* on *Y* that occurs through the intermediate phenotype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



42

Mediation analysis: Data requirements

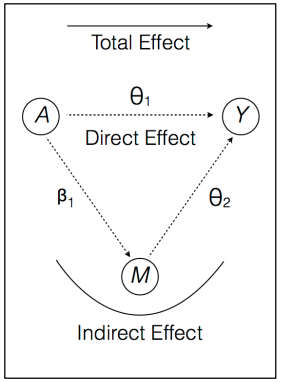
- All phenotypes must be measured on the same subjects
- Temporality must be ascertained
  - The occurrence of the intermediate variable *M* must precede that of the phenotypic outcome variable *Y*



43

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*

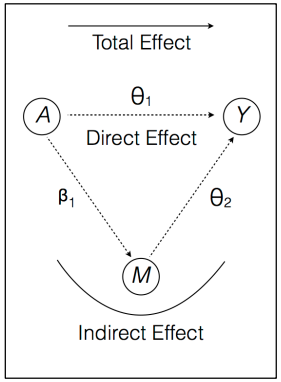


44

Mediation analysis: Assumptions

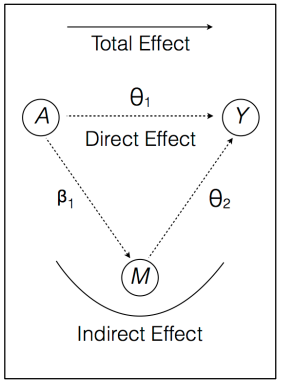
Typically met in genetic epi studies!

- 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*



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*

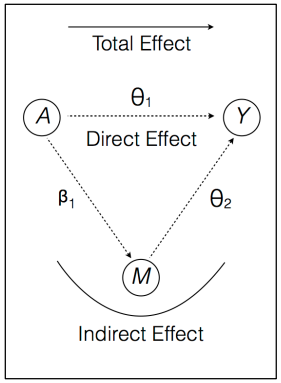


Requires adjustment for known confounders to prevent bias
(Note: this effectively restricts the use of mediation analyses to datasets in which data on such variables have been collected)

Mediation analysis: Regression-based approach

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

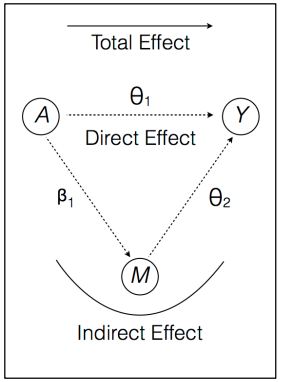
Assesses the effect of *A* on *M*, while controlling for measured confounders *(C)*



Mediation analysis: Regression-based approach

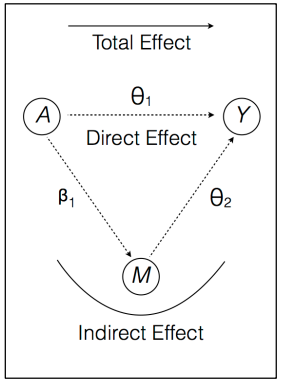
- Requires fitting two regression models, one for mediator *M* and one for phenotypic outcome *Y:*
  - $E[M\,|a,c] = \beta_0 + \boldsymbol{\beta_1} a + \beta_2' c$
  - $E[Y\,|a,m,c\,] = \theta_0 + \boldsymbol{\theta_1} a + \boldsymbol{\theta_2} m + \theta_4' c$

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



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

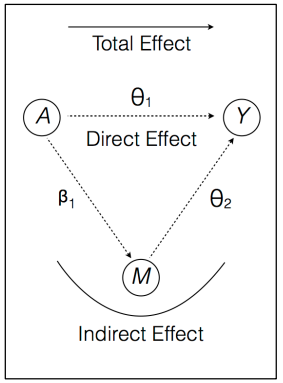


Guidelines for generating robust statistical evidence of pleiotropy

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

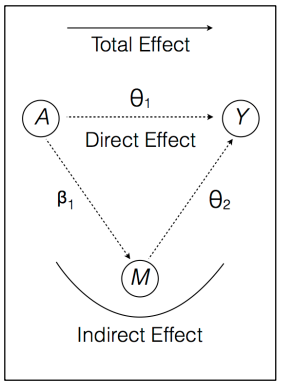
Mediation analysis: Interpretation

- **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)



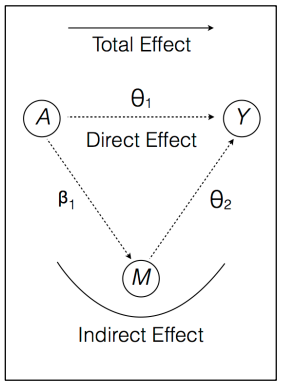
Mediation analysis: Interpretation

- **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)



Mediation analysis: Interpretation

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



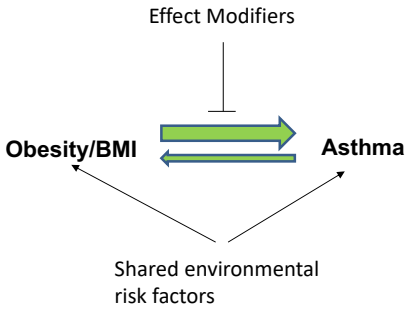
mediation R package

```
> 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 ***

Empirical searches for pleiotropic loci for asthma and obesity

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.

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.

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

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

Statistical methods

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

Results for BMI

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

Marker	Location (BP)[a]	Minor Allele	Allele Frequency CR	Allele Frequency CAMP	Number of Informative Families[b] (number of offspring with 0/1 recoded genotype) CR	Number of Informative Families[b] (number of offspring with 0/1 recoded genotype) CAMP	Effect Size[c] CR	Effect Size[c] CAMP	CR p Value[d,e]	CAMP Replication p Value[d,e] (two-sided)	Joint p Value[f] (CR, CAMP two-sided)
rs228883	61874457	T	0.27	0.33	91 (67/24)	110 (80/39)	2.45	1.60	+0.0011	+0.0038 (+0.0076)	5.6×10^{-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

Results for asthma

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

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

One asthma-associated variant

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.

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.

A GWAS study: Salinas et al. (In Press)

Discovery and mediation analysis of cross-phenotype associations with asthma and body mass index in 12q13.2

Salinas YD, Wang Z, and DeWan AT

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

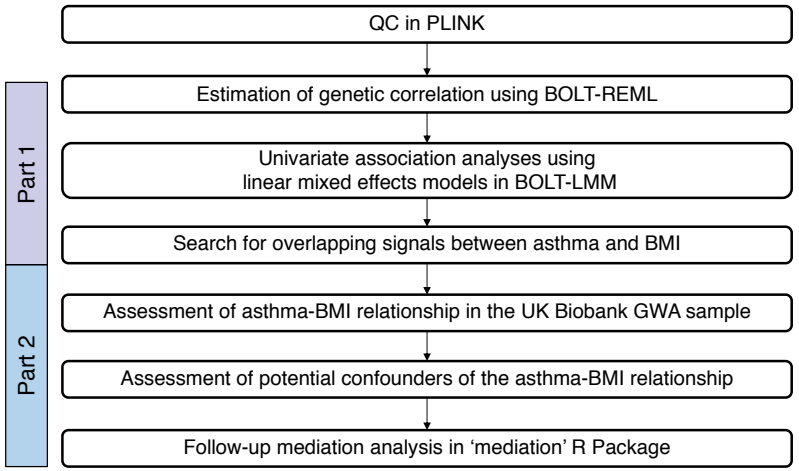
Study population

- 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)

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

Statistical Methods



Overlap in GWA signals

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

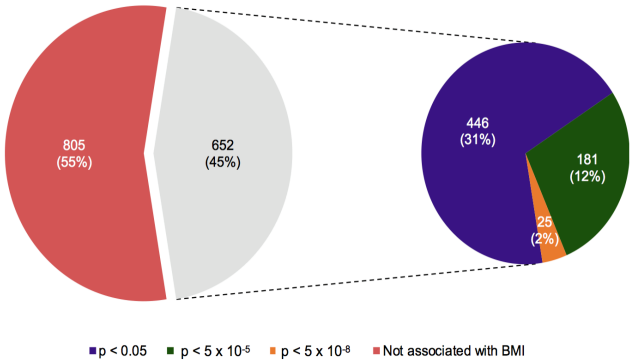


Figure 1. Overlap in GWA signals between asthma and BMI. Results for asthma are for the analysis of all asthmatic subjects (35,373 asthmatics vs. 270,572 non-asthmatics). Results for BMI are for the quantitative BMI analysis (n=305,945). Both analyses are sex- and age-adjusted. The threshold for genome-wide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

Overlap in GWA signals

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

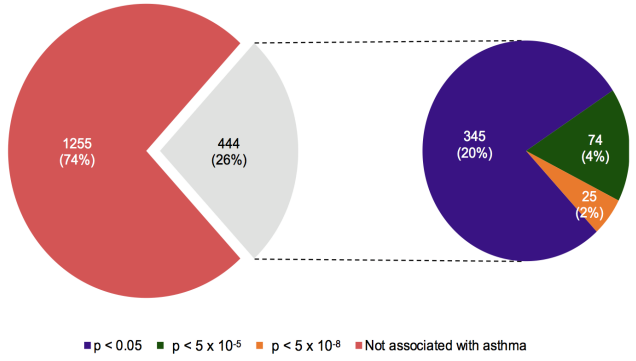
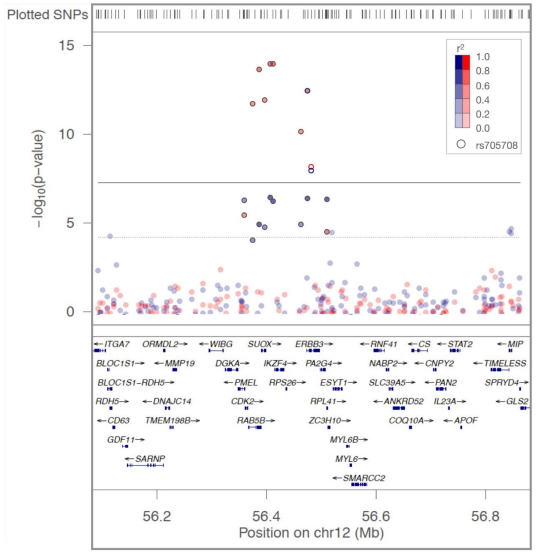


Figure 1. Overlap in GWA signals between asthma and BMI. Results for asthma are for the analysis of all asthmatic subjects (35,373 asthmatics vs. 270,572 non-asthmatics). Results for BMI are for the quantitative BMI analysis (n=305,945). Both analyses are sex- and age-adjusted. The threshold for genome-wide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

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



Cross-phenotype associations in 12q13.2

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

					Asthma		BMI[c]	
SNP	Gene	BP	Effect/reference allele	EAF	OR (95% CI)	P[c]	beta (95% CI)	P[d]
rs2069408	*CDK2*	56,364,321	G/A	0.3388	1.04 (1.02, 1.06)	3.30×10^{-6}	-0.06 (-0.08, -0.04)	5.40×10^{-7}
rs1873914	*RAB5*	56,379,427	C/G	0.4237	1.06 (1.04, 1.08)	2.40×10^{-12}	-0.05 (-0.07, -0.02)	7.90×10^{-5}
rs705702 [b]	*SUOX*	56,390,636	G/A	0.3376	1.07 (1.05, 1.09)	3.10×10^{-14}	-0.05 (-0.08, -0.03)	1.10×10^{-5}
rs10876864 [b]	*SUOX*	56,401,085	G/A	0.4279	1.06 (1.04, 1.08)	1.50×10^{-12}	-0.05 (-0.07, -0.03)	1.60×10^{-5}
rs1701704	*IKZF4*	56,412,487	G/T	0.3433	1.07 (1.05, 1.09)	1.50×10^{-14}	-0.06 (-0.09, -0.04)	3.70×10^{-7}
rs2456973	*IKZF4*	56,416,928	C/A	0.3432	1.07 (1.05, 1.09)	1.50×10^{-14}	-0.06 (-0.08, -0.04)	6.00×10^{-7}
rs11171739 [b]	*ERBB3*	56,470,625	C/T	0.4337	1.06 (1.04, 1.07)	8.80×10^{-11}	-0.05 (-0.07, -0.03)	1.10×10^{-5}
rs2292239	*ERBB3*	56,482,180	T/G	0.3470	1.07 (1.05, 1.08)	4.50×10^{-13}	-0.06 (-0.08, -0.04)	4.20×10^{-7}
rs705708	*ERBB3*	56,488,913	A/G	0.4712	1.05 (1.03, 1.07)	7.20×10^{-9}	-0.06 (-0.09, -0.04)	1.30×10^{-8}
rs11171747 [b]	*ESYT1*	56518408	T/G	0.6180	1.04 (1.02, 1.05)	2.90×10^{-5}	-0.06 (-0.08, -0.04)	4.50×10^{-7}

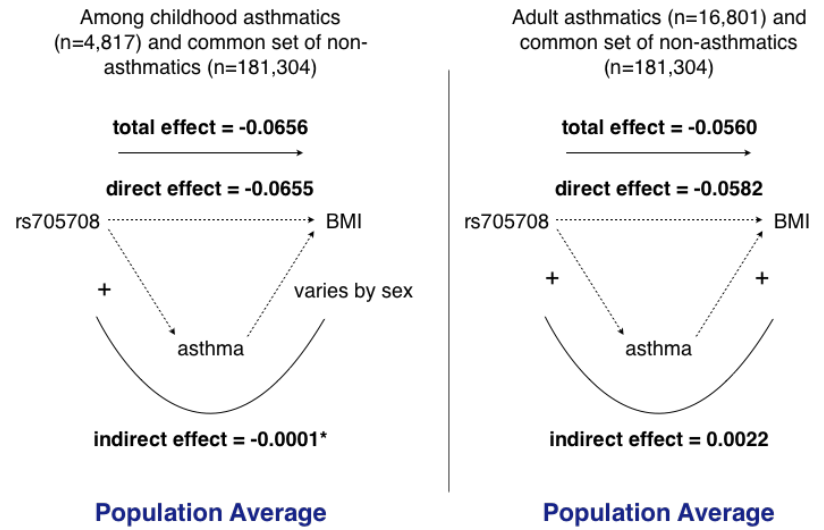
Abbreviations: BP = base-pair ; BMI = body mass index; CI = confidence interval; EAF = effect allele frequency; OR = odds ratio; SNP = single-nucleotide polymorphism

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

73

Decomposing the effect of rs705708 on BMI via mediation analysis

74



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.

75

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.

76

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).

Intro to population genetics

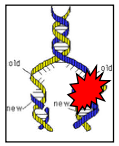
Shamil Sunyaev

Department of Biomedical Informatics
Harvard Medical School

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

Broad Institute of M.I.T. and Harvard

Forces responsible for genetic change

Mutation  μ

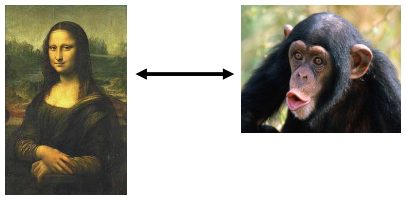
Selection  s

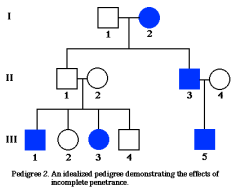
Drift  N_e

Population structure  F_{ST}

Mutations

Mutation rate in humans and flies





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

1.8x10^{-8} (Kondrashov)

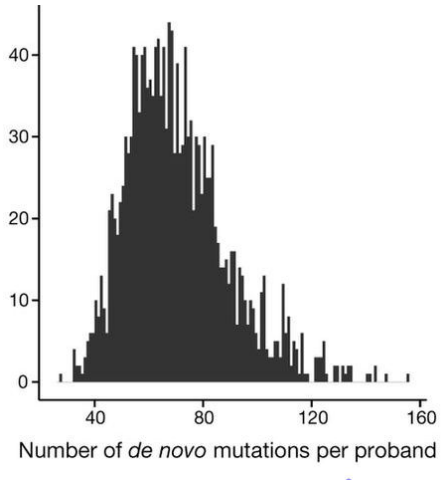
NGS estimates *~1.2X10*$^{-8}$ per nt changes genome

~70 per nt changes genome

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

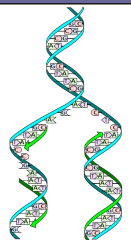
repeat extensions/contractions (10^{-5}*)*

Number of de novo mutations per individual

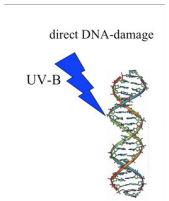


Jonsson et al., *Nature* 2017

Mutation rate is variable along the genome

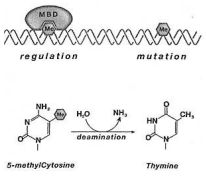


Replication fidelity



DNA damage

DNA repair



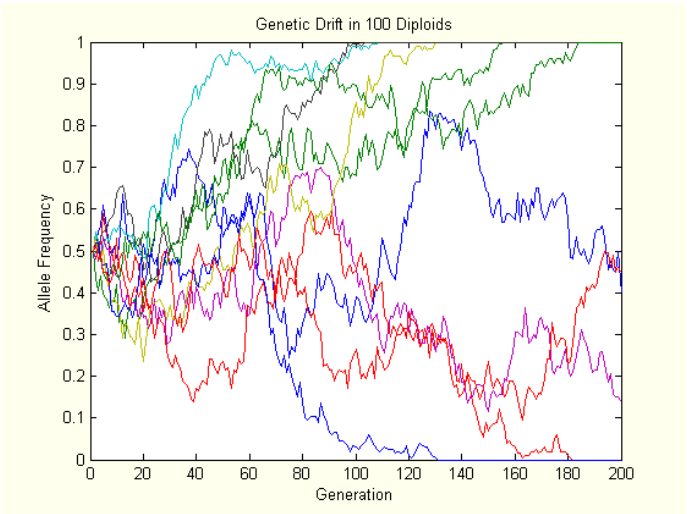
CpG deamination

Regional variation of mutation rate

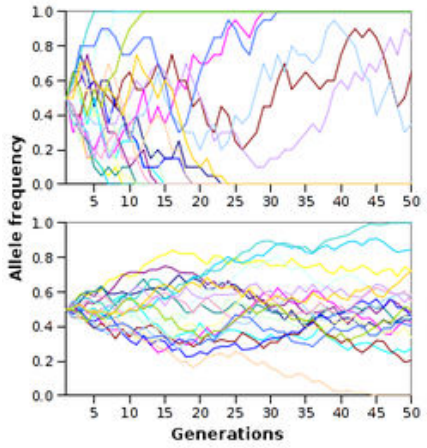
Context dependence of mutation rate

Genetic drift

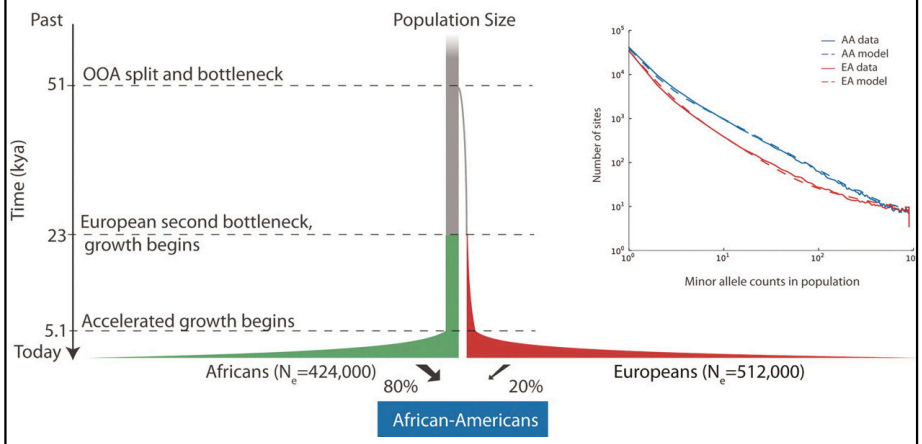
Drift is a random change of allele frequencies



Drift depends on population size



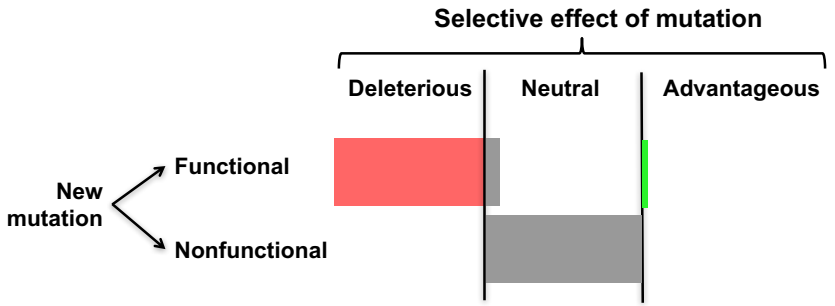
Demographic history



Tennessen et al. *Science* 2012

Selection

Most functional mutations are deleterious



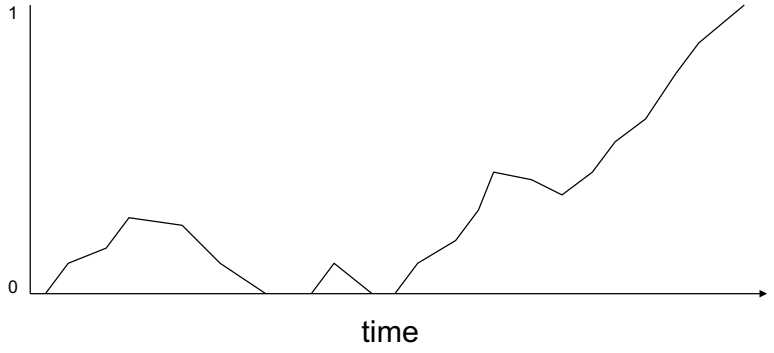
Selection indicates functional mutations, whether or not the tested trait is under selection

13

Methods of mathematical population genetics

Dynamic of allelic substitution

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



Diffusion approximation

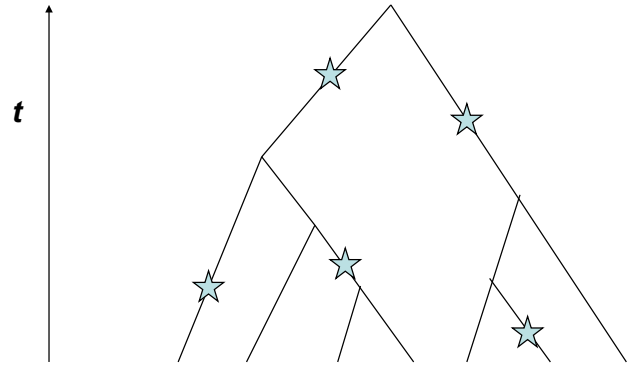
Random walk that does not jump long distances can be approximated by a diffusion process

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

Coalescent theory

Instead of modeling a population, we can model our sample

Time goes backwards !



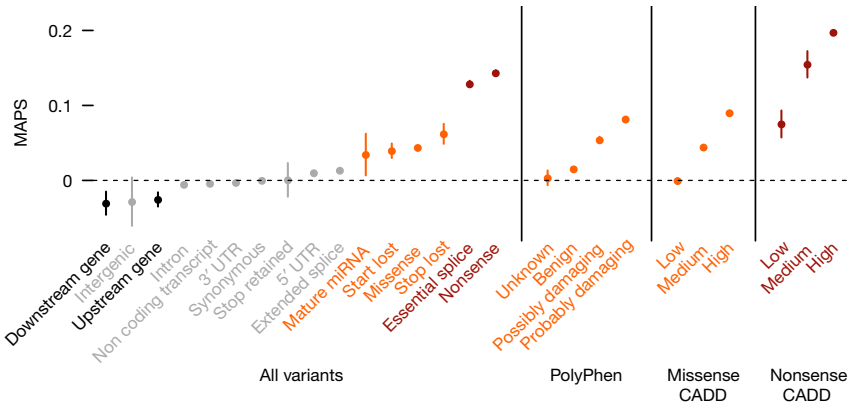
Natural selection in protein coding regions

Signatures of purifying selection

Reduced variation

Excess of rare alleles

Diversity and allele frequency



Am J Hum Genet 26:669–673, 1974

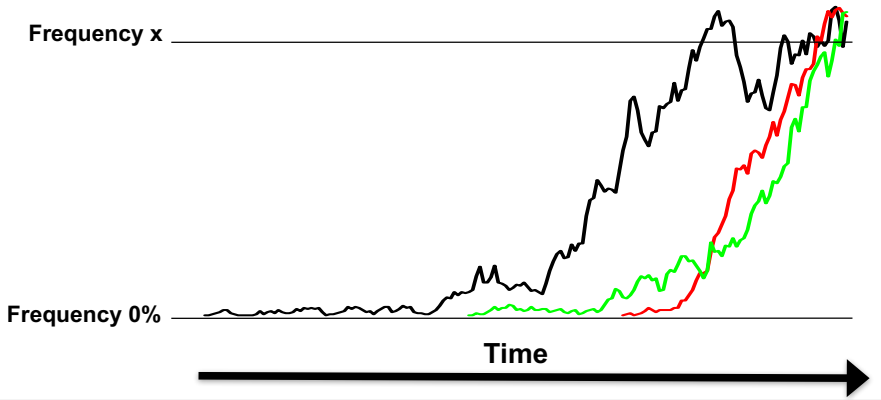
MAHLON V. R. FREEMAN, M.D.

The Age of a Rare Mutant Gene in a Large Population

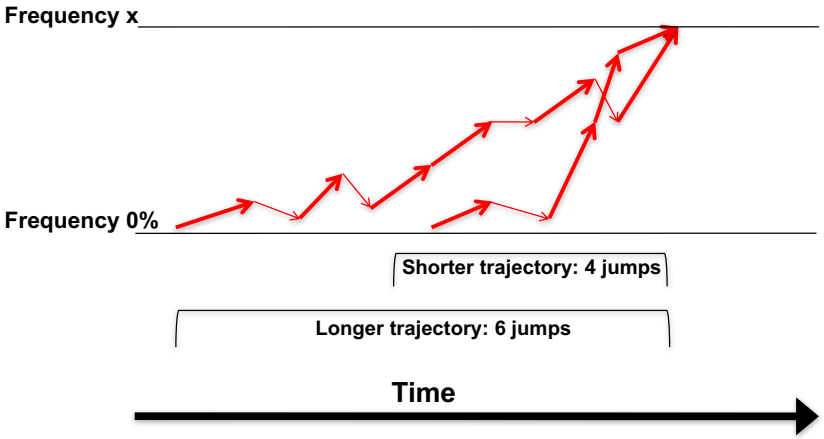
TAKEO MARUYAMA[1]

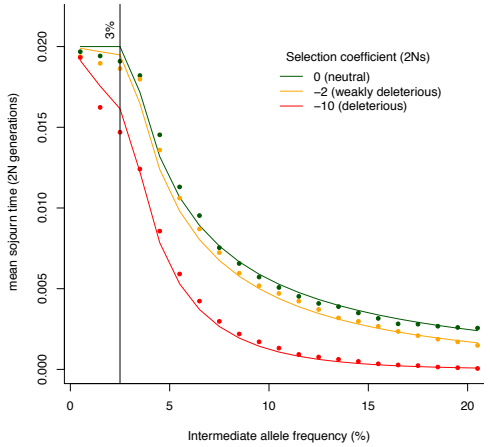
At a given frequency deleterious and advantageous alleles are younger than neutral

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



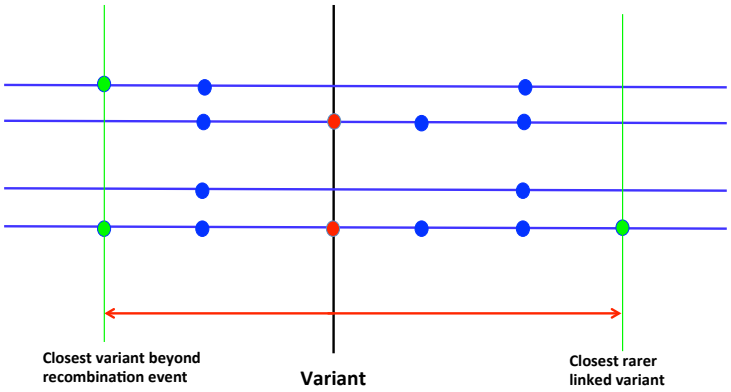
Intuition: shorter trajectories require fewer lucky jumps





Kiezun et al. *PLOS Genetics* 2013

Neighborhood clock (fuzzy clock)



Selection inference using frequencies of individual SNPs

Change in allele frequency **=**

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

Of the order of 10^{-8}

Demographic history

Population structure

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

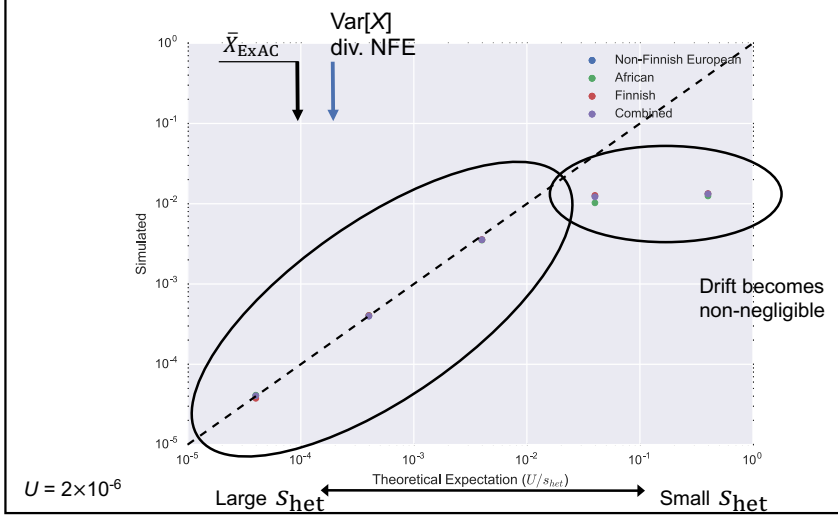
Selection inference using combined frequency of PTVs

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 $\lambda = U/hs$

Simulations



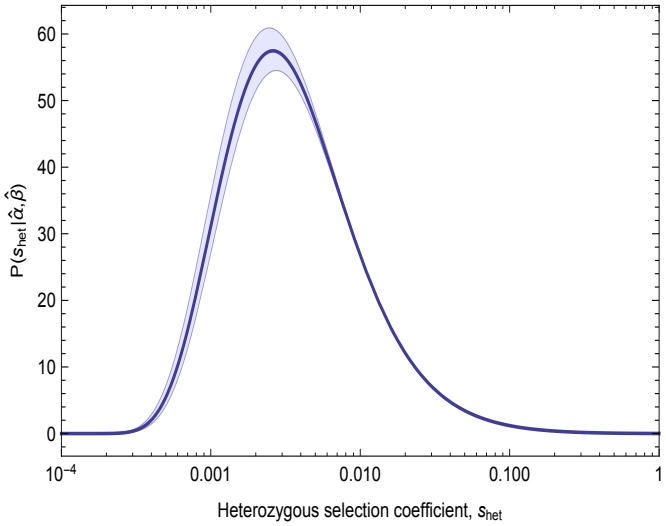
The model

PTV counts in each gene are Poisson distributed but we lack sufficient data to estimate selection coefficients

We can treat selection coefficients as random variables with a distribution to be estimated

$$P(n|\alpha,\beta;\nu) = \int P(n|s_{\text{het}};\nu)\, P(s_{\text{het}};\alpha,\beta)\, \mathrm{d}s_{\text{het}}$$

Distribution of selection coefficients



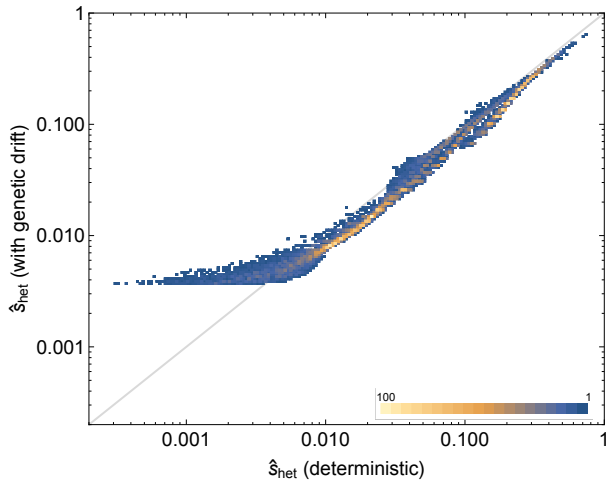
Cassa, Weghorn, Balick, Jordan et al. *Nature Genetics* 2017

Estimates for each gene

The estimated distribution over selection coefficients can be now used as a prior, and per gene estimates from posteriors

$$P\left(s_{\text{het},i}|n_i;\nu_i\right) = \frac{P\left(n_i|s_{\text{het},i};\nu_i\right)P\left(s_{\text{het},i};\hat{\alpha}_t,\hat{\beta}_t\right)}{\int P(n_i|s;\nu_i)P\left(s;\hat{\alpha}_t,\hat{\beta}_t\right)\,\mathrm{d}s}$$

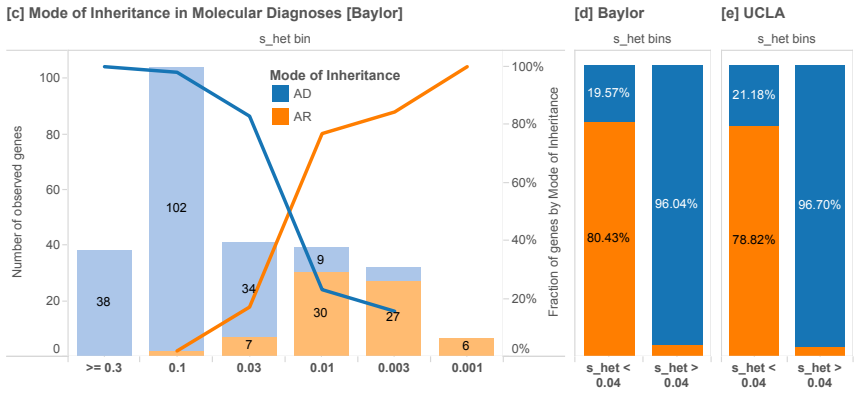
What happens if we incorporate drift?



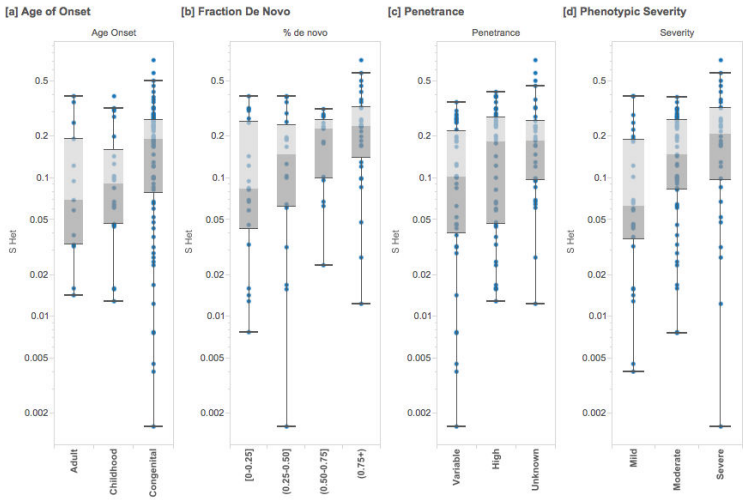
What happens if we incorporate drift?

1) The approach fails if selection is weak
2) The approach fails if mutational target is small
3) These considerations are important for regional constraint scores
4) Overall, the approach is non-informative in case of recessivity

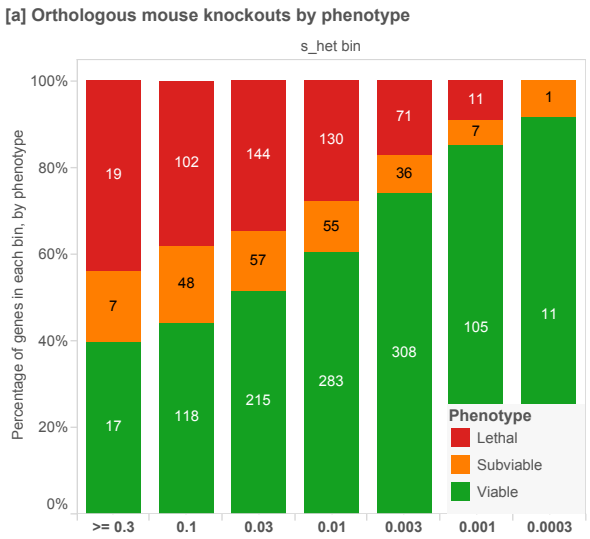
AD and AR Mendelian genes



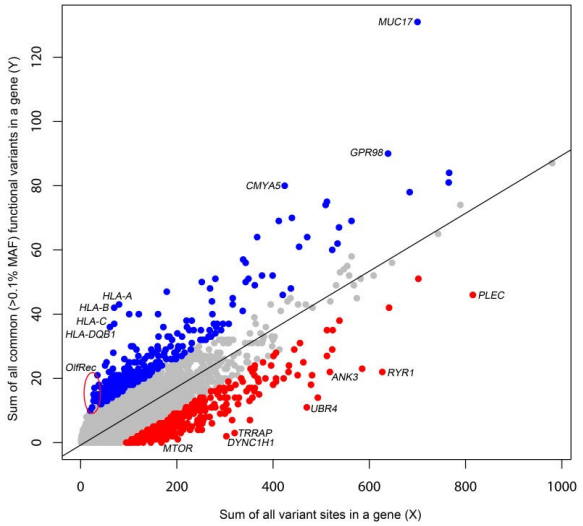
Age of onset, penetrance and severity



Concordance with mouse knockout data



RVIS



Petrovski et al. *PLOS Genetics* 2013

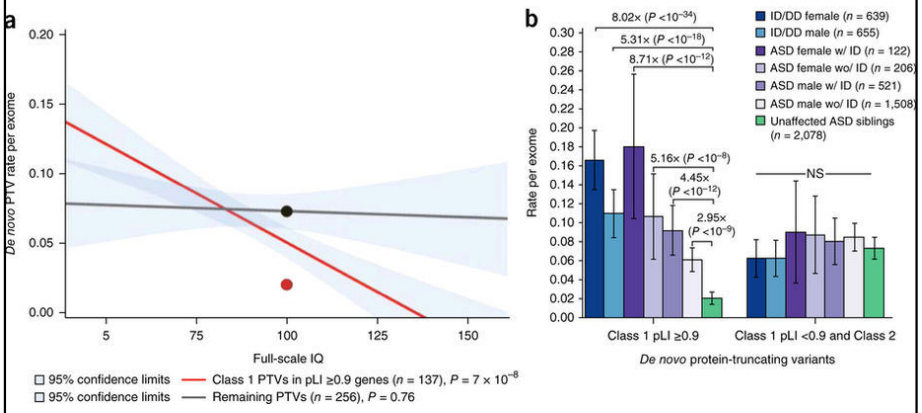
pLI

$$PTV_i \mid Z_i = c \sim Pois(N\lambda_c)$$

$$p(Z_i = c \mid \pi_c, PTV_i) = \frac{Pois(PTV_i \mid N\lambda_c)\pi_c}{\sum_c Pois(PTV_i \mid N\lambda_c)\pi_c},$$

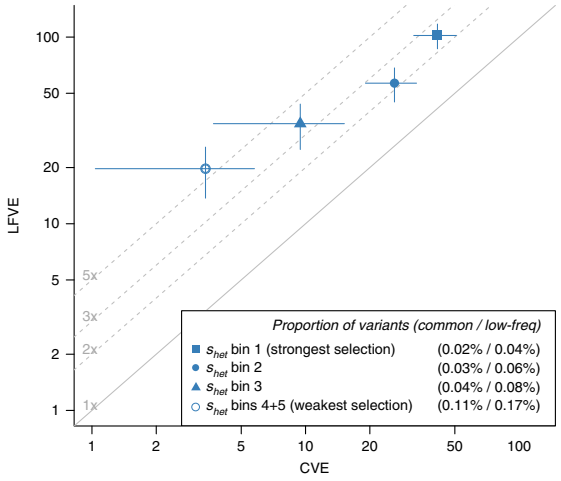
Lek et al. *Nature* 2016

De novo mutations in ASD



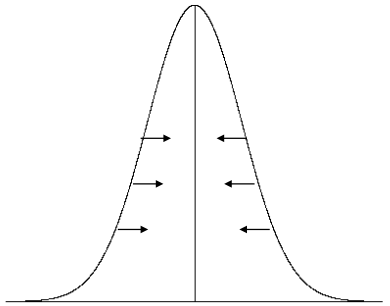
Kosmicki et al. *Nature Genetics* 2017

Heritability enrichment



Gazal et al. *Nature Genetics* 2018

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



Stabilizing selection

Selection may be related or unrelated to the trait

Technically, non-neutral genetic variation should not exist!

Forces to maintain variation:

Selection

Mutation

Why does a common genetic disease exist?

From evolutionary perspective common genetic disease should not exist: natural selection should remove disease-causing alleles from the population

Theory 1: MEDICALLY detrimental polymorphisms are not EVOLUTIONARY deleterious

- **Disease late onset** **(after the reproductive age)**
- **Changed environment and lifestyle (Selection direction reversal)**
- **Compensatory positive effect**

Balancing selection

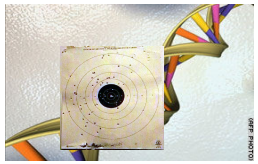
Frequency dependent selection

Antagonistic pleiotropy (Trade Off)

Examples: *APOE* (Alzheimer's disease)**, *AGT*** (Hypertension)**, *CYP3A*** (Hypertension)

Mutation/selection balance

Theory 2: Common diseases are due to multiple deleterious alleles in mutation-selection balance



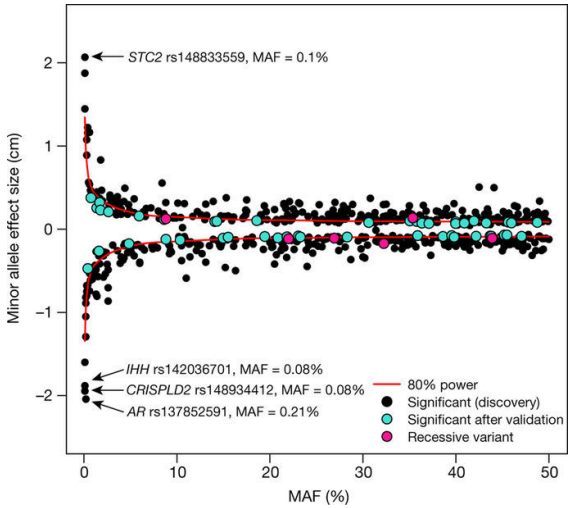
- **Weak selection**
- **High mutation rate**

CURRENT ESTIMATE:

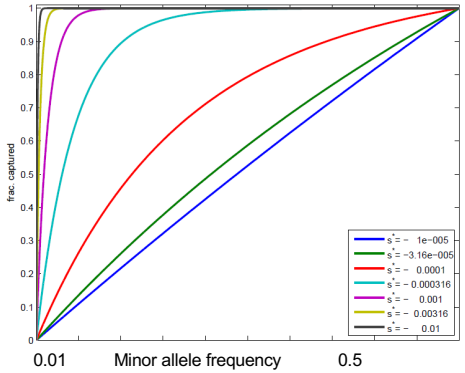
~70 new mutations per genome

~1 new coding mutation per genome

Rare coding alleles have larger effect sizes



Heritability by allele frequency

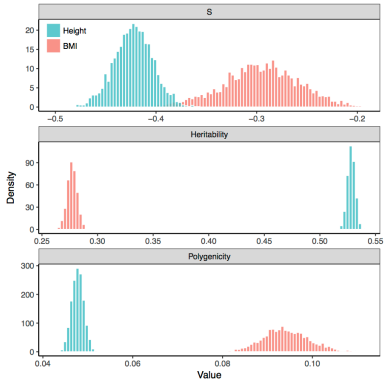


Effective population size: N=10,000

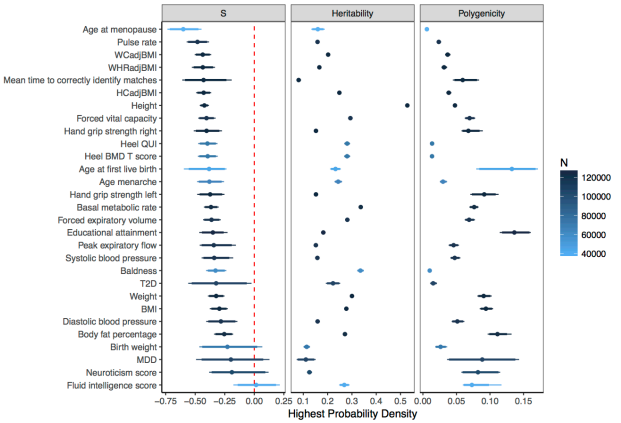
Evidence in favor of the highly polygenic model

$$\beta_j \sim N\left(0, [2p_j(1-p_j)]^S \sigma_\beta^2\right)\pi + \phi(1-\pi)$$

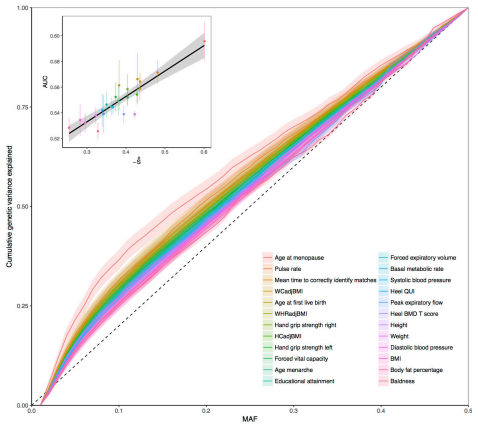
Evidence in favor of the highly polygenic model



Evidence in favor of the highly polygenic model



Evidence in favor of the highly polygenic model



Evolution, maintenance and allelic architecture of complex traits

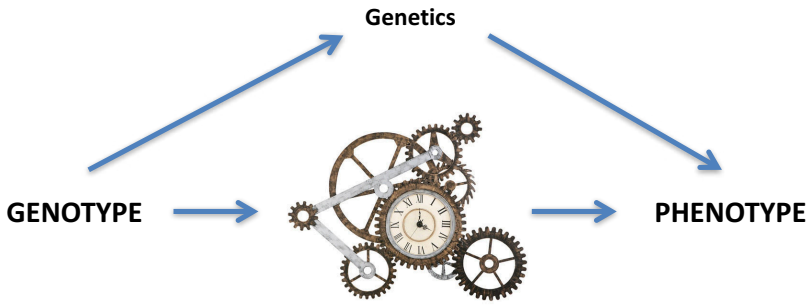
Shamil Sunyaev

Department of Biomedical Informatics
Harvard Medical School

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

Broad Institute of M.I.T. and Harvard

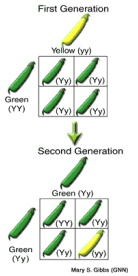
Why are we doing genetics?



The role of statistics

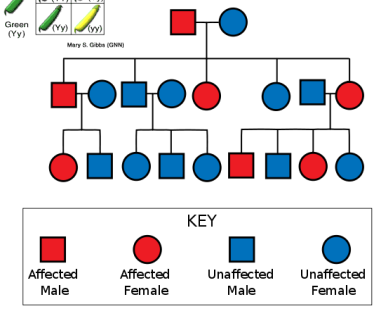
- Genetics for statistics is what physics is for mathematics
- Genetics is a leading motivation for development of new basic statistics
- Statistics is the main formal instrument (although not the only one)

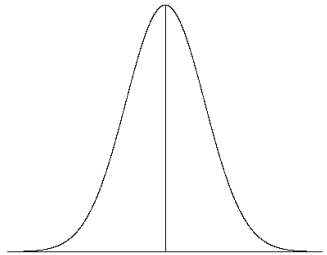
Simple and Complex Phenotypes



Simple phenotype

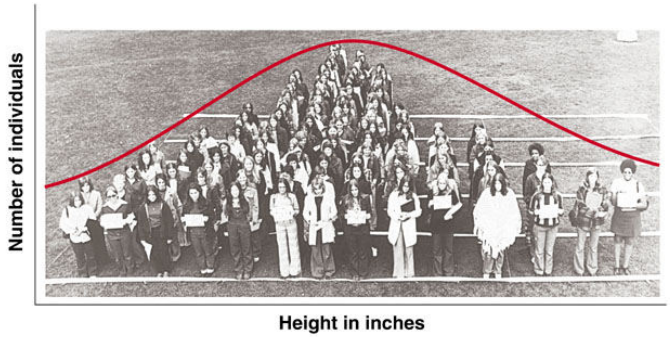
Complex phenotype



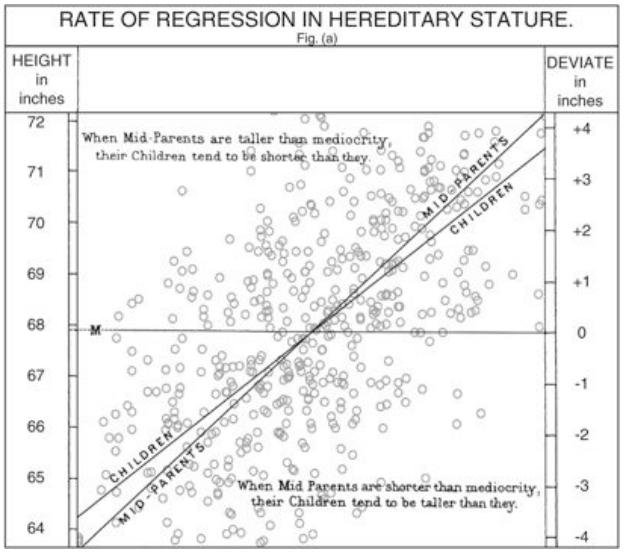


Complex traits are heritable but not in Mendelian fashion

Tobin/Dusheck, Asking About Life, 2/e
Figure 16.6



Complex traits are heritable but not in Mendelian fashion



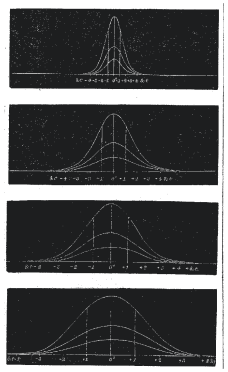
Infinitesimal model

NATURE [*April* 5, 1877

lies with
r towards
ie hinder
s are fully
ics on the
extremi-

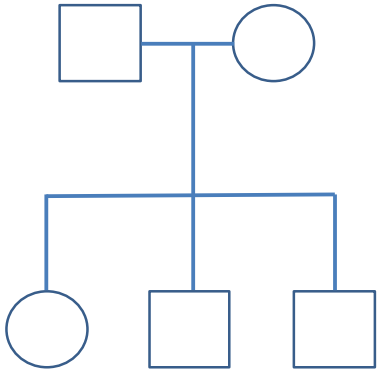
TYPICAL LAWS OF HEREDITY[1]

WE are far too apt to regard common events as matters of course, and to accept many things as obvious truths which are not obvious truths at all, but present problems of much interest. The problem to which I am about to direct attention is one of these.



7

Infinitesimal model: multivariate normal distribution in pedigrees



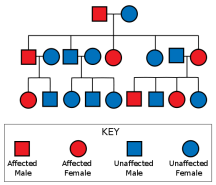
The pedigree defines the covariance matrix

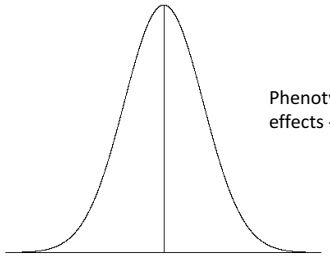
XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

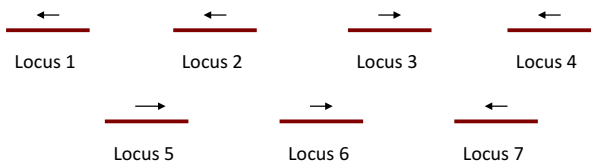
Quantitative Trait Loci (QTLs)

Inheritance at each locus is Mendelian. Loci are independent

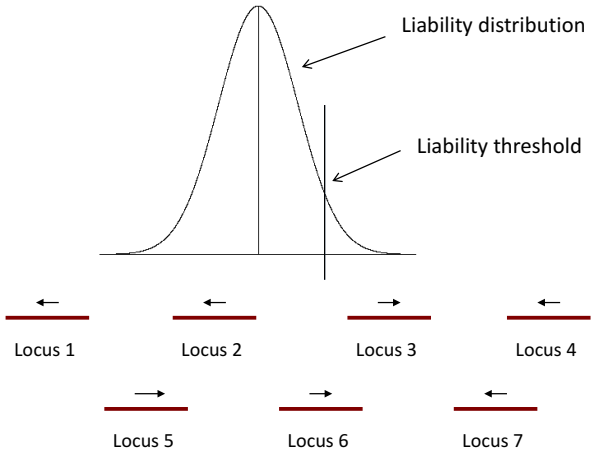




Phenotype is additive over locus effects -> normal distribution



Dichotomous complex traits such as disease



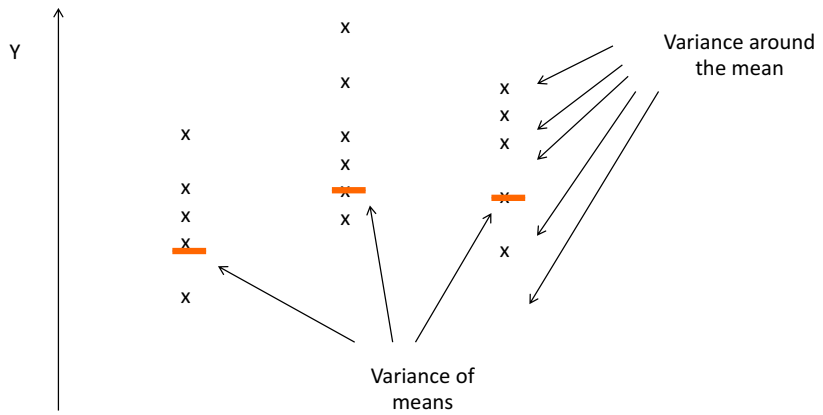
Population variation is fully described by variance

$$V = V_G + V_E$$

Genetic contribution

Everything else

Variance decomposition



Components of genetic variance

$$V_G = V_A + V_D + V_I + V_M$$

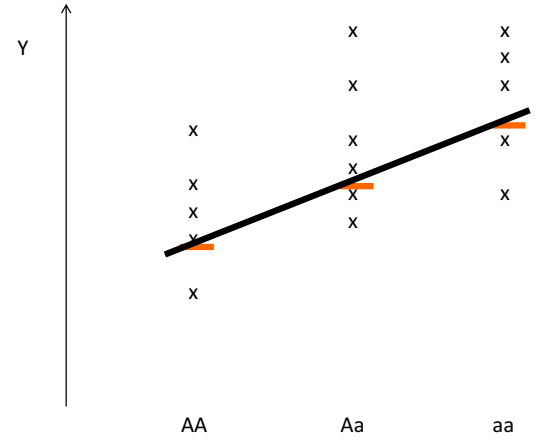
Main (additive) effects

Dominant effects

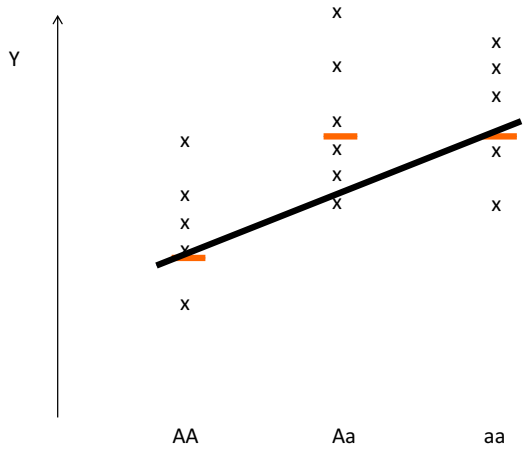
Genetic interactions

New mutations

Regression



Regression



Additive variance

Additive variance V_A is variance explained by the model

$$Y_j = \sum_i \beta_i X_{ij} + \varepsilon$$

$$V_A = 2\sum_i \beta_i^2 x_i (1 - x_i)$$

Variance components due to dominance and epistasis

Dominance variance V_D is variance explained by the residuals of the model additive over loci

Epistatic variance V_I is genetic variance that is not captured by the model additive over loci (presumably due to interactions)

Additive by additive pairwise epistasis

$$Y_j = \sum_i \beta_i X_{ij} + \sum_{lk} \beta_{lk} X_{lj} X_{kj} + \varepsilon$$

Other variance components

Epistasis can be additive by dominant and dominant by dominant

Epistasis can be due to higher order interactions

Mutational variance V_M – additional variance due to *de novo* mutations

Heritability

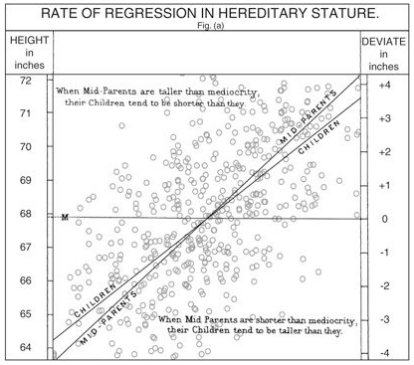
Broad sense

$$H^2 = \frac{V_G}{V}$$

Narrow sense

$$h^2 = \frac{V_A}{V}$$

Estimating heritability

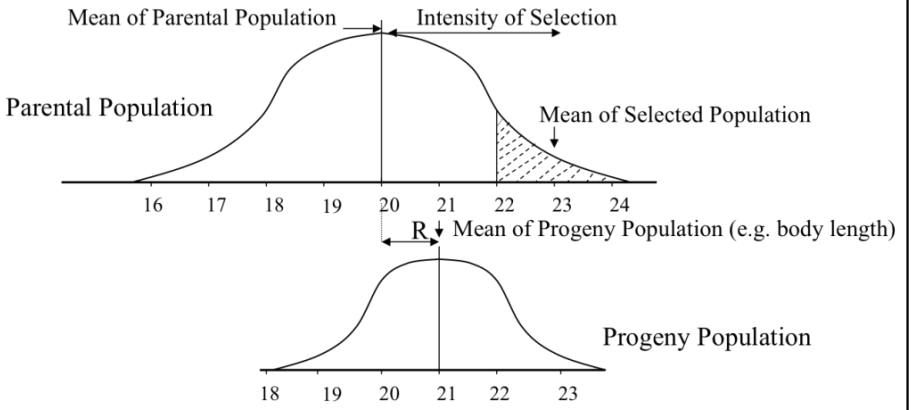


$$Cov(MP,O) = \frac{1}{2}V_A + \frac{1}{4}V_I$$

Narrow sense heritability

$$h^2 = \frac{V_A}{V} \approx \frac{Cov(MP,O)}{V(MP)}$$

Breeder's equation

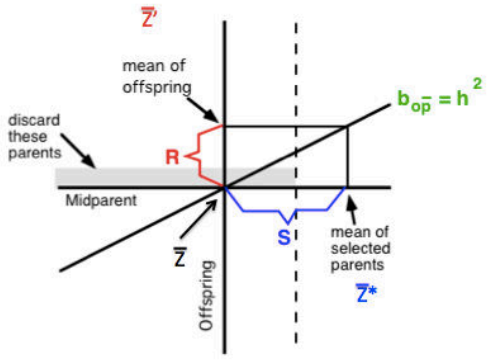


$$R = h^2 S$$

Breeder's equation

Response to Selection = heritability * Selection Differential

R = h^2 S



$$h^2 = \frac{V_A}{V_P}$$

With genotypic information in hand

Regress phenotype on genotype

$$Y_j = \sum_i \beta_i X_{ij} + \varepsilon$$

Additive variance

$$V_A = 2\sum_i \beta_i^2 x_i (1 - x_i)$$

Narrow sense heritability

$$h^2 = \frac{V_A}{V}$$

In the Ideal World

Regress phenotype on genotype

$$Y_j = \sum_i \beta_i X_{ij} + \varepsilon$$

Identify significant and reproducible associations.
Estimate effect sizes. Estimate additive variance.

$$\hat{V}_A = 2 \sum_i^{known} \hat{\beta}_i^2 \, x_i (1 - x_i)$$

Reality: missing heritability

$$\hat{h}^2 = \frac{\hat{V}_A}{V} << \frac{Cov(MP,O)}{V(MP)}$$

Current GWAS explain a minor fraction of heritability

The case of the missing heritability

Height – 10%, Blood lipids – 12%

Likely reasons for missing heritability

1. *Common variants of weak effect*
2. *Rare variants of larger effect*
3. *Epistatic interactions*

$$Cov(MP,O) = \frac{1}{2}V_A + \frac{1}{4}V_I$$

Questions about allelic architecture

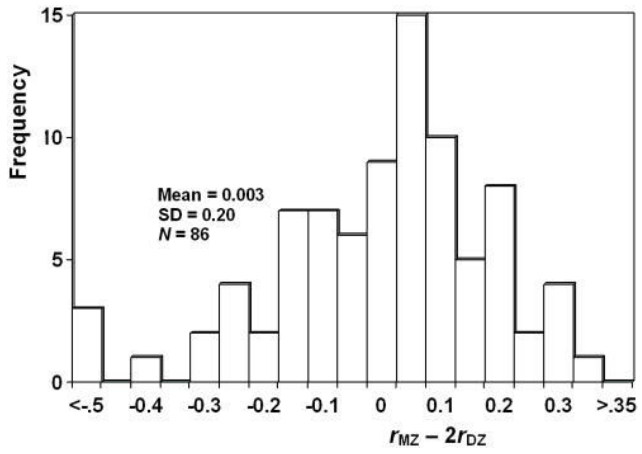
- How many loci are involved?
- Is variation underlying the trait rare or common?
- What is the distribution of effect sizes of variants involved in the trait?
- What is the role of epistasis and dominance?

GxG interactions

Why is epistasic variance commonly disregarded?

- In human genetics, epistatic interactions between common variants have not been observed.
- In a model with two (or several) loci, contribution of epistatic variance is relatively small.
- Long term response to selection in model organisms seems to contradict the importance of epistasis.

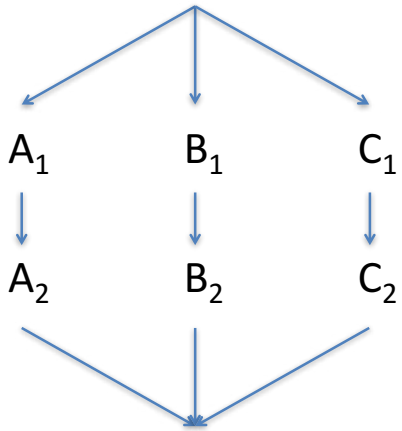
Any evidence for or against epistasis?



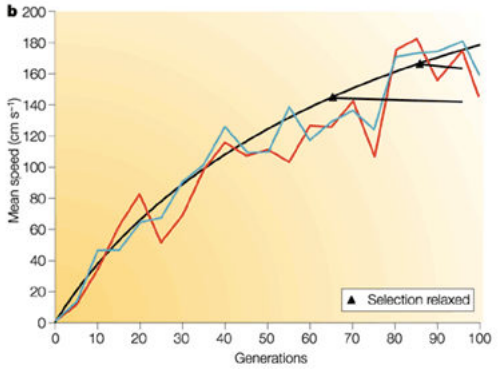
Why is epistasic variance might be of importance?

- A non-linear model involving many loci would generate a large epistatic variance.
- Interactions would be statistically undetectable.
- The model would not generate significant deviations from the observations.
- As an example, we may consider a model with multiple pathways involved.

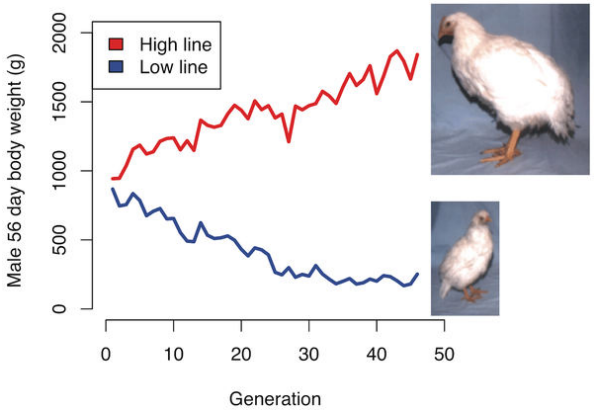
Multiple pathway model



Evidence in favor of the highly polygenic model



Evidence in favor of the highly polygenic model



Evidence in favor of the highly polygenic model

ANALYSIS

nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

	AA	Aa	aa
Genotypes	0	1	2
Normalized genotypes	$\frac{0-E(X)}{\sqrt{Var(X)}}$	$\frac{1-E(X)}{\sqrt{Var(X)}}$	$\frac{2-E(X)}{\sqrt{Var(X)}}$
Normalized genotypes	$\frac{-2q}{\sqrt{2pq}}$	$\frac{p-q}{\sqrt{2pq}}$	$\frac{2p}{\sqrt{2pq}}$

If SNP1 is causal and SNP2 is not, the apparent association of SNP2 is:

$$\hat{\beta}_2 = \beta_1 \cdot r_{12}$$

In non-normalized genotypes

$$\hat{\beta}_2 = \beta_1 \cdot r_{12} \cdot \sqrt{\frac{Var(X_1)}{Var(X_2)}}$$

X_{ij} – Normalized genotype of individual *i* at SNP *j*

In the matrix form:

$$\bar{y} = X\bar{\beta} + \varepsilon$$

Two important matrices:

$$LD = \frac{1}{M} X^T X$$

$$GRM = \frac{1}{N} XX^T$$

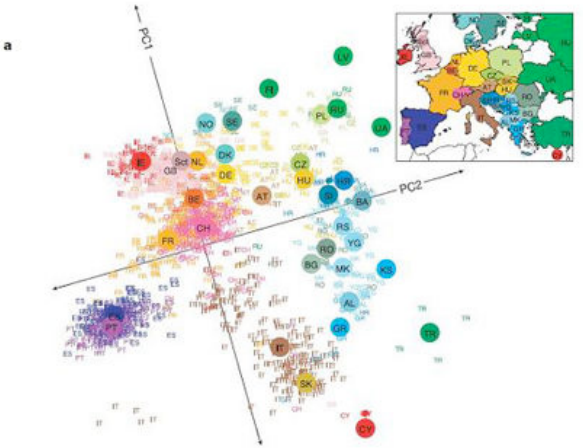
Principle component analysis (PCA)

$$GRM = \frac{1}{N} XX^T$$

Principle component are eigenvectors

First principle component corresponds to the largest eigenvalue

Europe



Linear Mixed Models (LMM)

- We can model effects of individual variants as random effects distributed as $N(0,\sigma^2)$.
- Random effect model is a model with error terms drawn from a multivariate normal distribution.
- In the infinitesimal model, co-variance matrix can be approximated using IBS (not IBD).

Linear Mixed Model (LMM)

Our model

$$Y_i = \sum_j \beta_j X_{ij} + \varepsilon$$

We have to fit markers individually

$$Y_i = \beta_1 X_1 + \sum_{j=2} \beta_j X_{ij} + \varepsilon \sim \beta_1 X_1 + \varepsilon'$$

For each SNP we can fit the model

$$Y_i = \beta X_i + u_i + \varepsilon$$

$$\varepsilon \sim N\left(0, I\sigma^2\right) \qquad u \sim MVN\left(0, GRM\right)$$

Remember from the Galton plot

Parent and offspring share 50% of DNA (IBD)

$$Cov(P,O) = \frac{1}{2} V_A$$

More generally, if fraction of the genome IBD is ***r***

$$Cov(A,B) = \frac{1}{r} V_A$$

If we assume that genetic effects are random

We assume that all SNPs have effects on the trait drawn from a normal distribution

$$Y_i = \mu_i + u_i + \varepsilon$$

$$Cov(u_i, u_k) = \frac{1}{N} \sigma^2 \sum_j X_{ij} X_{ik}$$

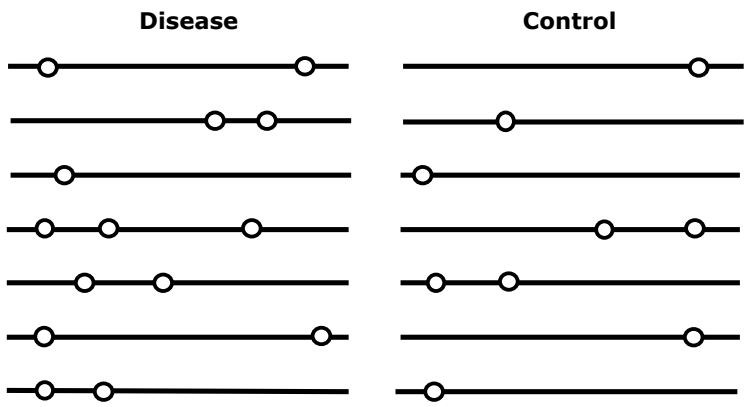
$$u \sim MVN\left(0, \sigma^2 GRM\right)$$

Challenges of the polygenic model

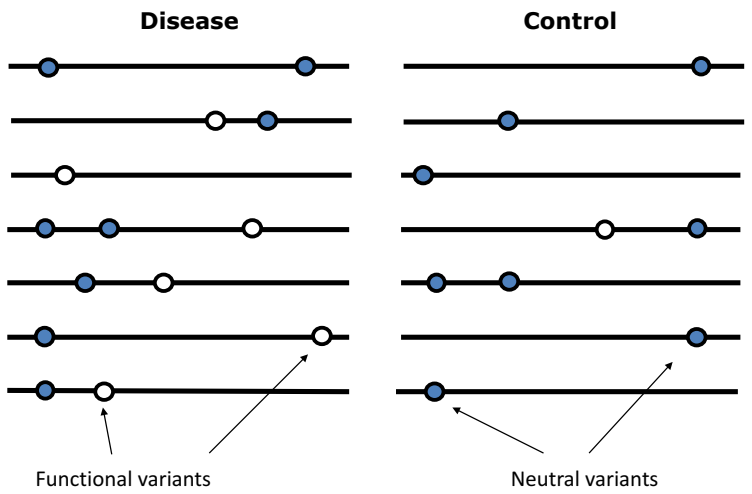
1) Need for a very large target size
2) Natural selection is expected to rapidly eliminate variants and reduce allele frequency of remaining variants
3) Variants must be either very rare or of very small effect sizes

Rare variants

This is a direct association!



This is a direct association!



Hyperlipidemia in Coronary Heart Disease

II. GENETIC ANALYSIS OF LIPID LEVELS IN 176 FAMILIES AND DELINEATION OF A NEW INHERITED DISORDER, COMBINED HYPERLIPIDEMIA

Joseph L. Goldstein, Helmut G. Schrott, William R. Hazzard, Edwin L. Bierman, and Arno G. Motulsky with the technical assistance of Ellen D. Campbell and Mary Jo Levinski

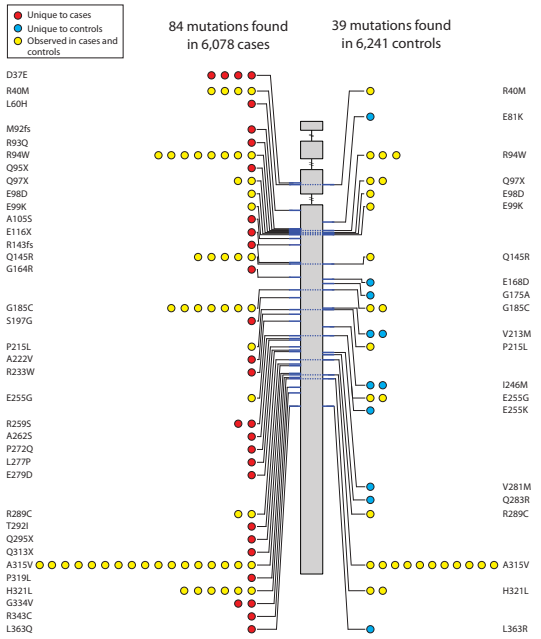
From the Departments of Medicine (Division of Medical Genetics, University Hospital, and Division of Metabolism and Gerontology, Veterans Administration Hospital) and Genetics, University of Washington, Seattle, Washington 98195

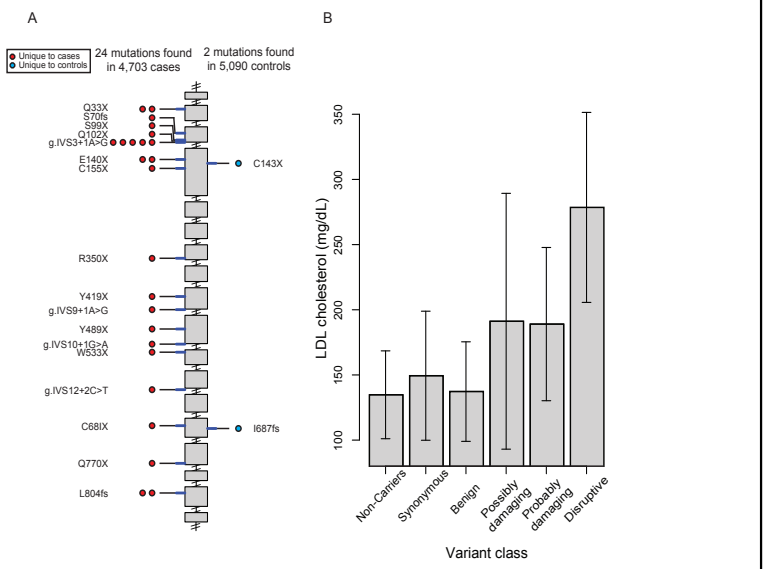
TABLE XII

Frequency of Hyperlipidemia

Disorder	Survivors of myocardial infarction: < Age 60 (a)	≥ Age 60 (b)	Ratio a/b	General population*
	%	%		%
A. Monogenic hyperlipidemia				
Familial hypercholesterolemia	4.1	0.7	5.9	~0.1–0.2
Familial hypertriglyceridemia	5.2	2.7	1.9	~0.2–0.3
Combined hyperlipidemia	11.3	4.1	2.8	~0.3–0.5
Total	20.6	7.5		~0.6–1.0
B. Polygenic Hypercholesterolemia	5.5	5.5	1.0	—
C. Sporadic Hypertriglyceridemia	5.8	6.9	0.8	—

Goldstein et al, *JCI*, 52:1544, 1973





Hyperlipidemia in Coronary Heart Disease

II. GENETIC ANALYSIS OF LIPID LEVELS IN 176 FAMILIES AND DELINEATION OF A NEW INHERITED DISORDER, COMBINED HYPERLIPIDEMIA

Joseph L. Goldstein, Helmut G. Schrott, William R. Hazzard, Edwin L. Bierman, and Arno G. Motulsky with the technical assistance of Ellen D. Campbell and Mary Jo Levinski

From the Departments of Medicine (Division of Medical Genetics, University Hospital, and Division of Metabolism and Gerontology, Veterans Administration Hospital) and Genetics, University of Washington, Seattle, Washington 98195

TABLE XII

Frequency of Hyperlipidemia

Disorder	Survivors of myocardial infarction: < Age 60 (a)	≥ Age 60 (b)	Ratio a/b	General population*
	%	%		%
A. Monogenic hyperlipidemia				
Familial hypercholesterolemia	4.1	0.7	5.9	~0.1–0.2
Familial hypertriglyceridemia	5.2	2.7	1.9	~0.2–0.3
Combined hyperlipidemia	11.3	4.1	2.8	~0.3–0.5
Total	20.6	7.5		~0.6–1.0
B. Polygenic Hypercholesterolemia	5.5	5.5	1.0	—
C. Sporadic Hypertriglyceridemia	5.8	6.9	0.8	—

Goldstein et al, *JCI*, 52:1544, 1973

Non-Parametric Polygenic Risk Prediction

Shamil Sunyaev

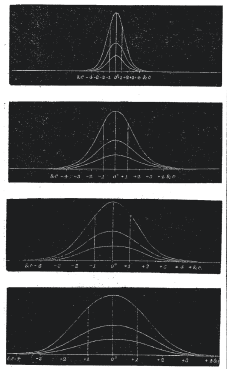
Department of Biomedical Informatics
Harvard Medical School

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

lies with
r towards
ıe hinder
; are fully
ics on the
· extremi-

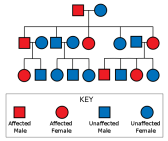
TYPICAL LAWS OF HEREDITY[1]

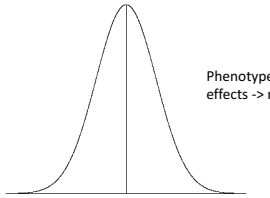
WE are far too apt to regard common events as matters of course, and to accept many things as obvious truths which are not obvious truths at all, but present problems of much interest. The problem to which I am about to direct attention is one of these.



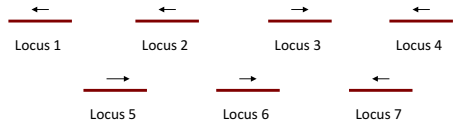
Quantitative Trait Loci (QTLs)

Inheritance at each locus is Mendelian. Loci are independent

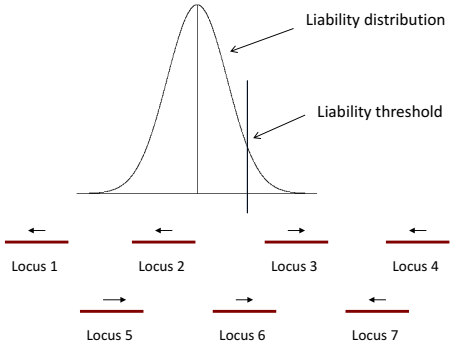




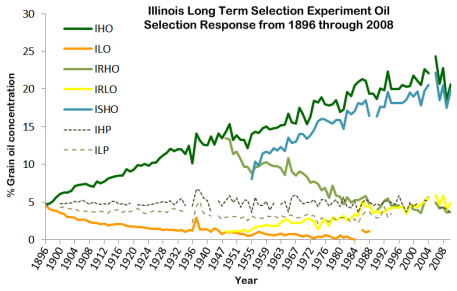
Phenotype is additive over locus effects -> normal distribution



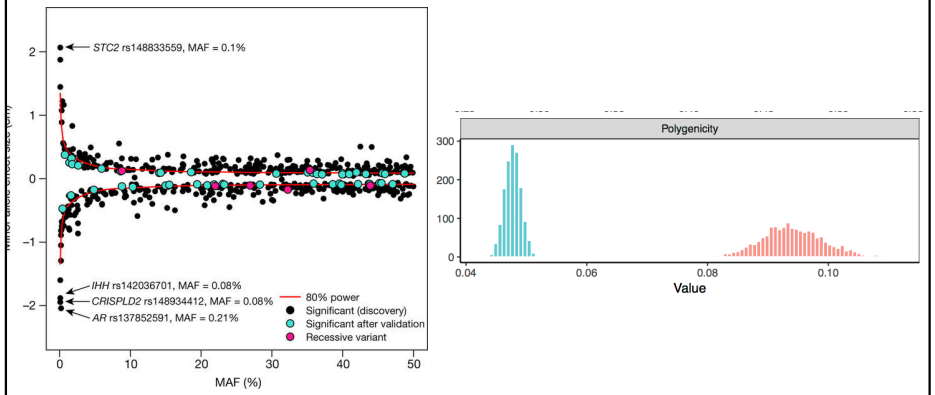
Binary traits such as diseases



Early evidence of high polygenicity of complex traits



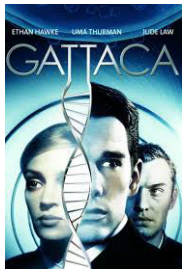
Evidence in favor of the highly polygenic model



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.

Genetic risk prediction



Genotype of an individual (Common SNPs) → Life-time risk of genetic disorders (Common complex genetic disorders)

Effect sizes of individual variants are very small

- Genotype at a single locus carries very little information about phenotype.
- I does not mean that one cannot predict phenotype from genotype.
- Accuracy (r^2) of an ideal genetic predictor equals heritability.

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)

LDL levels and risk of disease

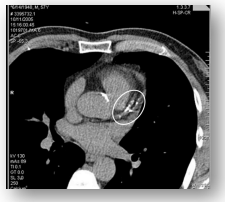
Annals of Internal Medicine | ARTICLE

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

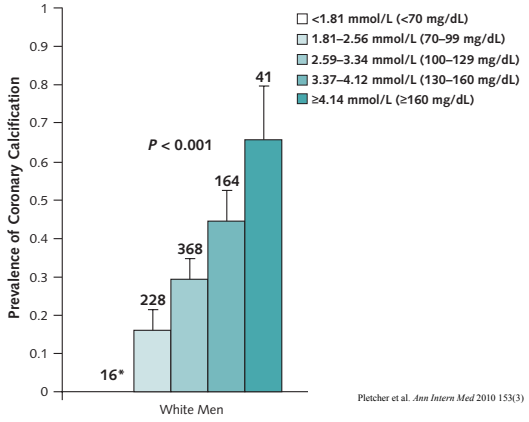
~3500 subjects < 35 years old

15-20 years



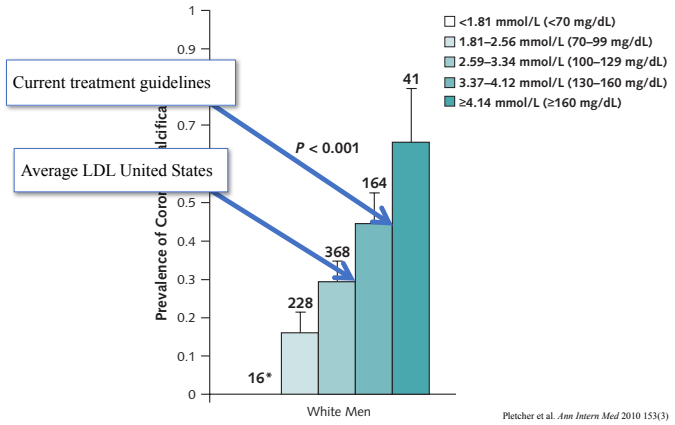
Piers *et al. BMC Cardiovascular Disorders* 2008 **8**:38

LDL levels and risk of disease



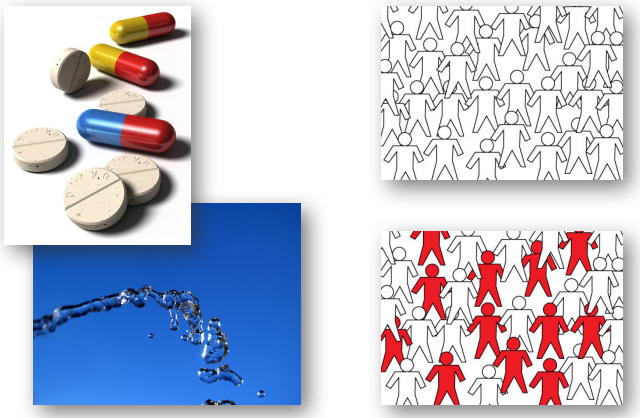
Pletcher et al. *Ann Intern Med* 2010 153(3)

LDL levels and risk of disease



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

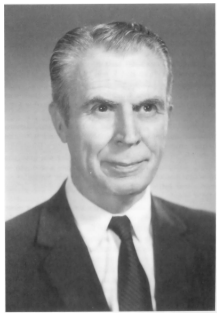
Selecting populations for treatment



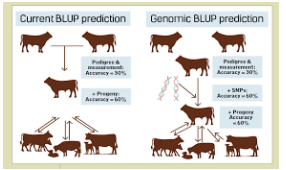
Why estimate genetic risk?

- An estimate of the long-term risk at birth
- Genetic risk can be combined together 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

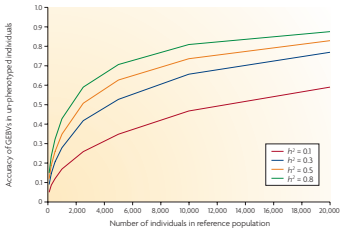
BLUP – Best Linear Unbiased Predictor



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



Accuracy of polygenic prediction in cattle



Poor transferability between breeds!

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

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

Methods that work with summary statistics

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

20

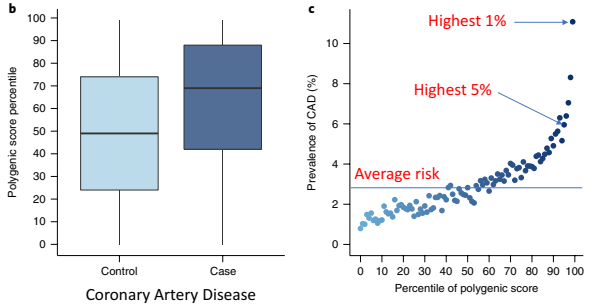
LDPred – a Bayesian method using summary statistics

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

Vilhjalmsson et al. 2015

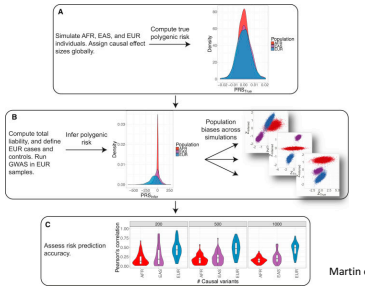
Also, check *BayesR*

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



Khera et al. 2018

With some caveats



Martin et al., *AJHG* 2017

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

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

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

Predicted genetic risk

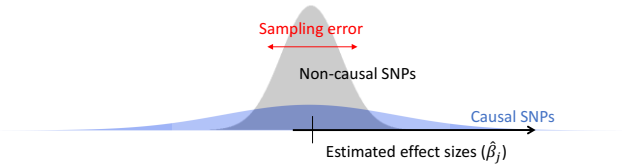
Estimated effect size

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

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

Predicted genetic risk

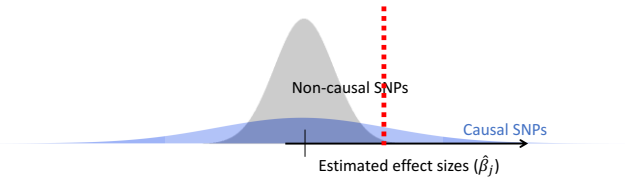
Estimated effect size



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

P-value Thresholding

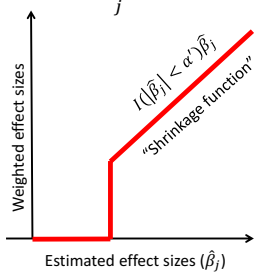
$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$



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

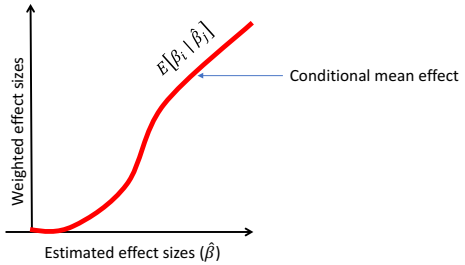
P-value Thresholding

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



The optimal polygenic score can be constructed with "conditional mean effects"

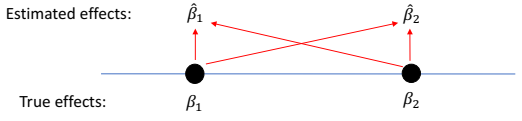
$$\hat{y}_i = \sum_j E[\beta_j \mid \hat{\beta}_j] x_{ij}$$

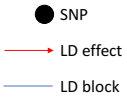


Goddard et al. 2009

Accounting for LD in summary data is a major challenge

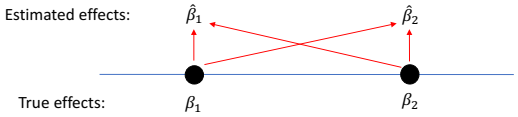
- Correlation between **apparent true genetic effects**



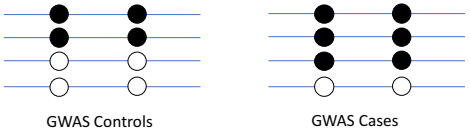


Accounting for LD in summary data is a major challenge

- Correlation between **apparent true genetic effects**



- Correlation between **sampling errors**



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

Our approach ("**N**on-**P**arametric **S**hrinkage" or NPS)

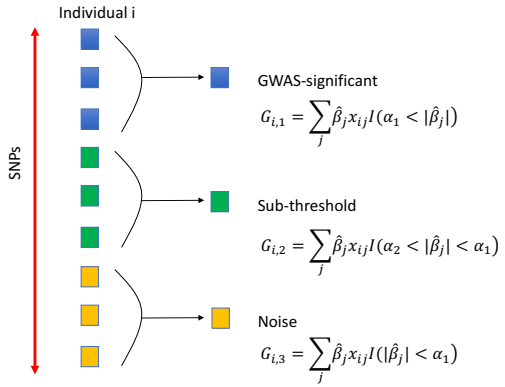
- No explicit specification of genetic architecture prior, thus "*non-parametric*"
- Learn conditional mean effects directly from training data

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

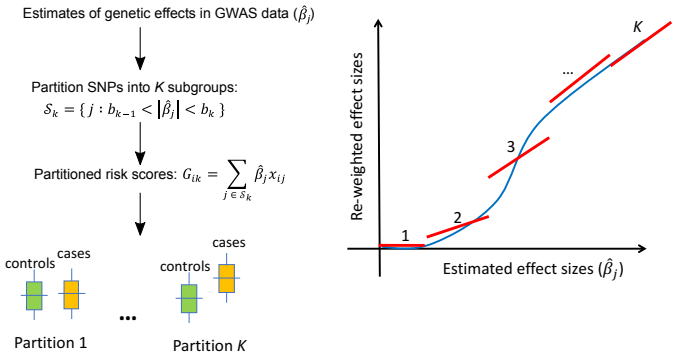
- Fully account for correlation in summary statistics

2. How to deal with LD

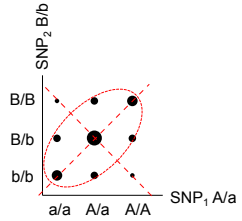
Partitioned risk scores



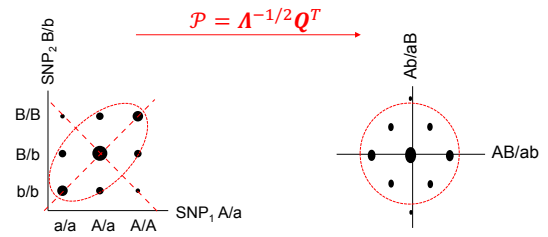
Piecewise linear interpolation on shrinkage curve



How to deal with LD?



Decorrelating linear projection $\mathcal{P}$

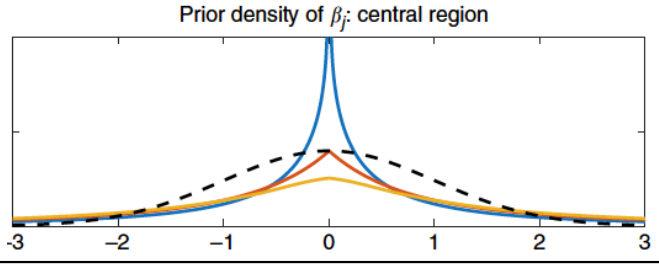


$\boldsymbol{\Sigma}$ is a local LD matrix and $\boldsymbol{\Sigma} = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^T$ by eigenvalue decomposition

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q}\,\boldsymbol{\Lambda}^{-1}\,\boldsymbol{Q}^T = (\boldsymbol{Q}\,\boldsymbol{\Lambda}^{-1/2})(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^T)$$

Other shrinkage methods: PRS-CS

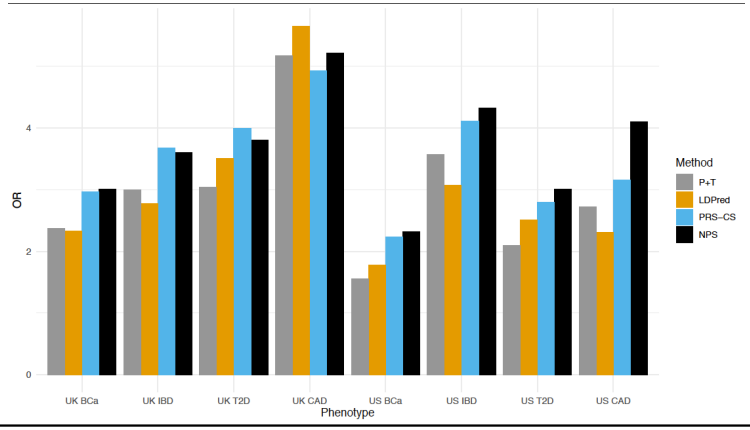
$$\beta_j \sim \mathrm{N}\left(0, \frac{\sigma^2}{N}\phi\psi_j\right), \qquad \psi_j \sim g,$$



Other shrinkage methods: PRS-CS

Lassosum – extension of *LASSO*

Accuracy of the 5% tail



Summary on the method

- NPS accounts for the correlation of sampling errors in GWAS summary statistics.
- NPS provides an extensible framework to estimate the shrinkage curve from training data.
- NPS is best-suited to take advantage of the high density of markers and imputation accuracy in latest GWAS datasets.

The preprint is available in BioRxiv:
Chun et al. "*Non-parametric polygenic risk prediction using partitioned GWAS summary statistics.*"

Software is available at: https://github.com/sgchun/nps

Is an extreme presentation with a family history Mendelian?

- It is often assumed that an extreme phenotypic presentation is due to a large effect Mendelian mutation.
- Apparently Mendelian family history is assumed to support a highly penetrant Mendelian mutation.
- Could these cases be polygenic (or, at least, not monogenic)?

42

Annotating gene sequence variation

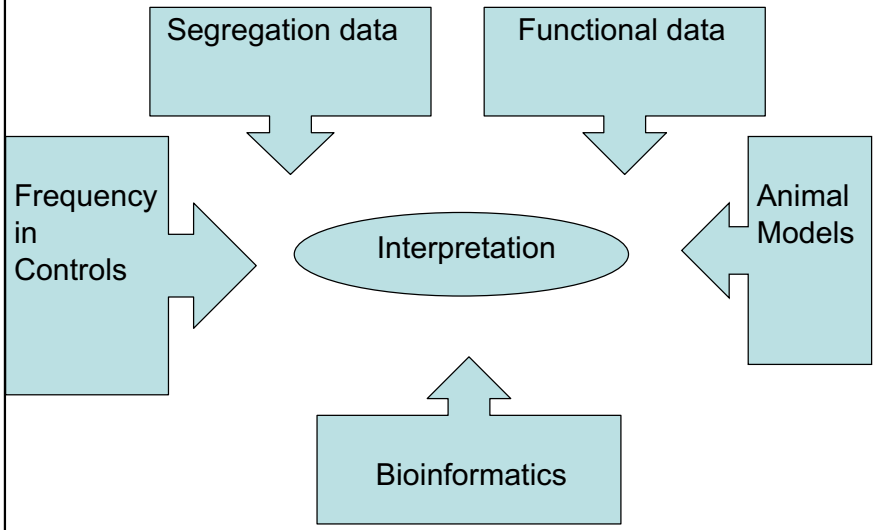
Shamil Sunyaev

Department of Biomedical Informatics
Harvard Medical School

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

Broad Institute of M.I.T. and Harvard

Identifying functionally significant causal variants in



Map variants on genomic annotation

Watch for multiple transcripts!

Watch for conflicting annotations!

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

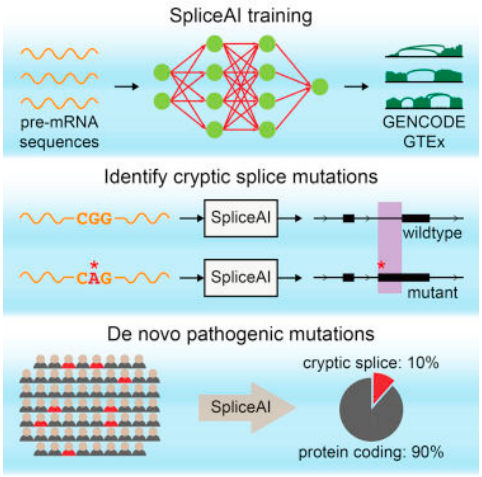
Variants involved in splicing

Variants in canonic splicing sites

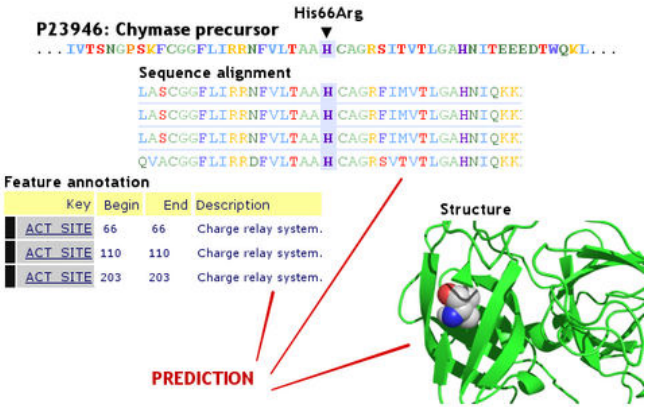
Variants in exonic or intronic splicing enhancers

Gain of splicing variants

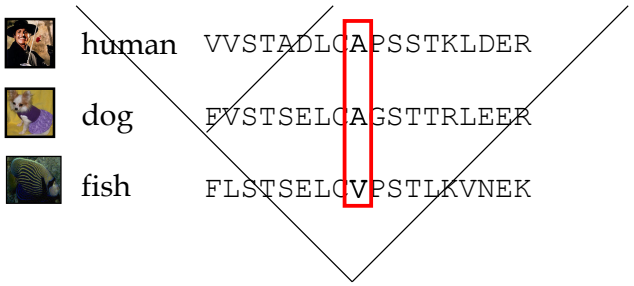
SpliceAI



Missense variants: computational predictions



Does the mutation fit the pattern of past evolution?



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

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

Continuous time Markov model

GLY → *VAL* → *ALA* → *GLY* → *ALA*

Continuous time Markov model

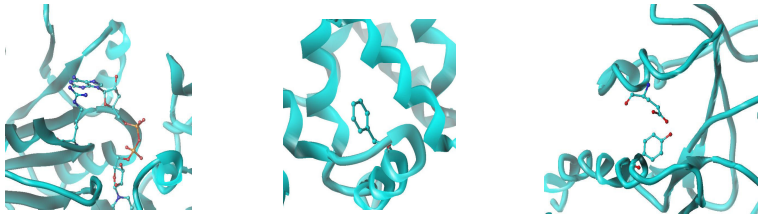
P – matrix of transition probabilities

$$P(t) = e^{Qt}$$

π – stationary distribution

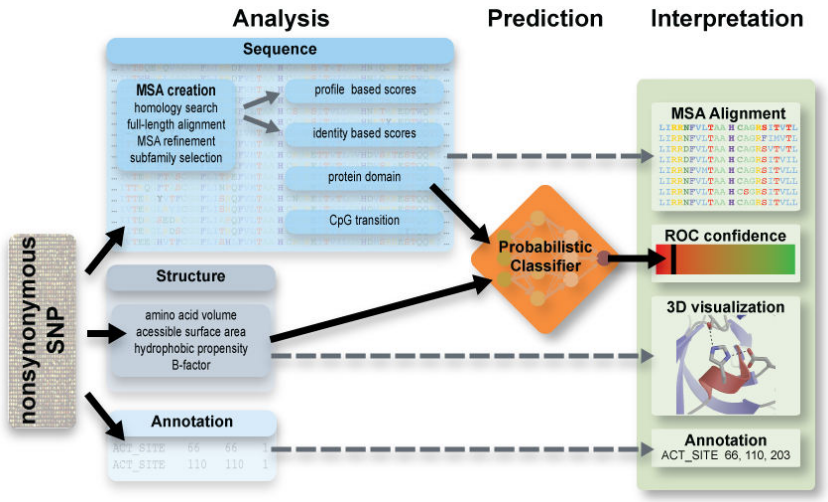
$$Q\bar{\pi} = 0$$

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.**

PolyPhen2



www.genetics.bwh.harvard.edu/pph2 Adzhubei, et al. Nature Methods 2010

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

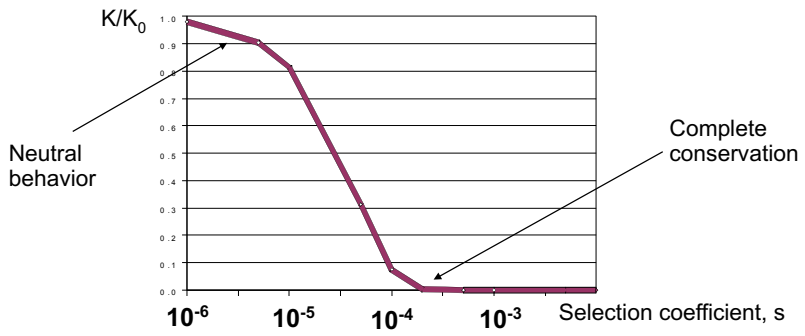
Conservation can be due to very weak selection

Every new mutation eventually will be either fixed or lost

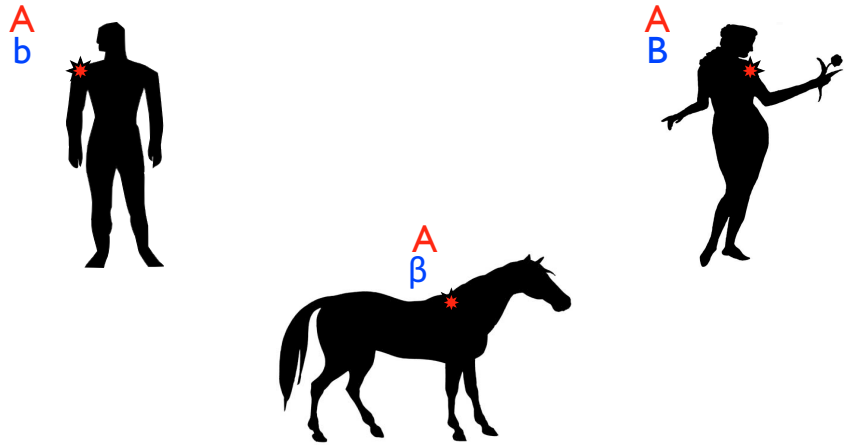
$$K = K_0 2 N_e \frac{(1-e^{-2s})}{(1-e^{-4N_e s})}$$

s – selection coefficient
N_e - effective population size

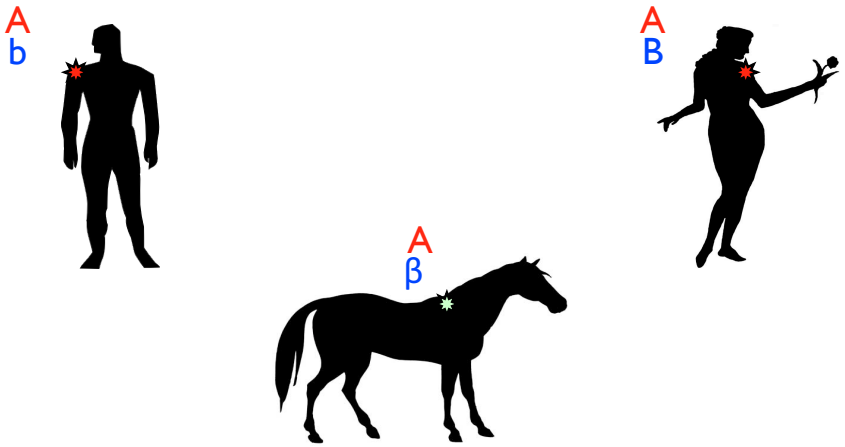
For humans estimated to be ~ 10 000



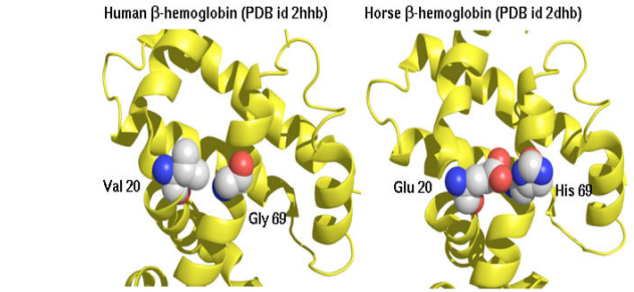
Constant evolutionary landscape



Epistatic interactions



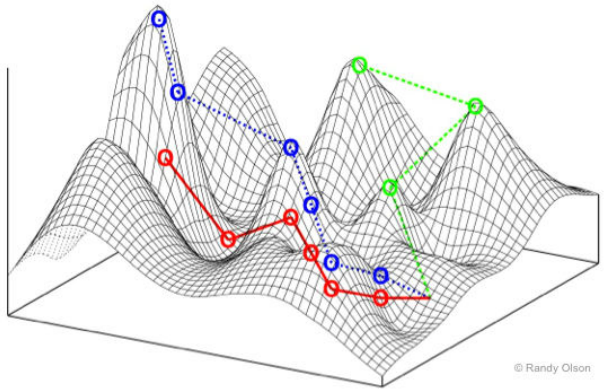
Compensatory mutations



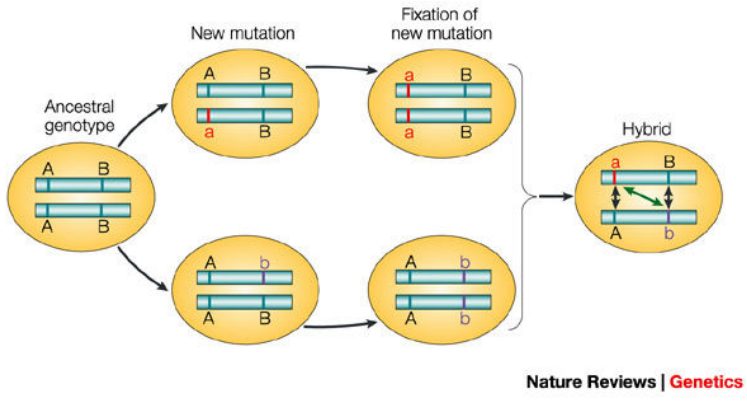
The phenomenon of compensatory mutations in different fields

- Biochemistry – protein stability, allosteric effects
- Genetics – incomplete penetrance
- Evolutionary biology – speciation, epistatic models of evolution

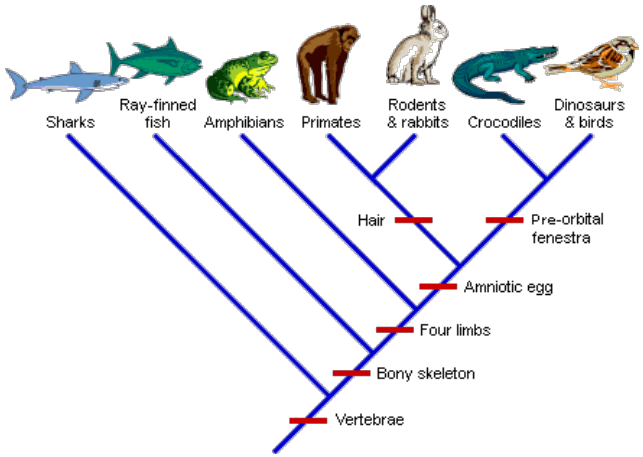
Ridges on the fitness landscape



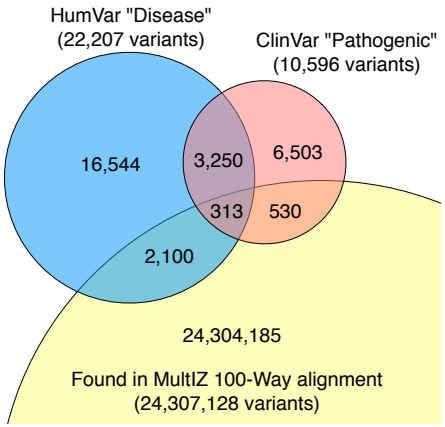
Dobzhansky-Muller incompatibility



Looking at vertebrate species



Many human disease mutations are found in vertebrates



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

How complex genetic suppression can be?

LETTER

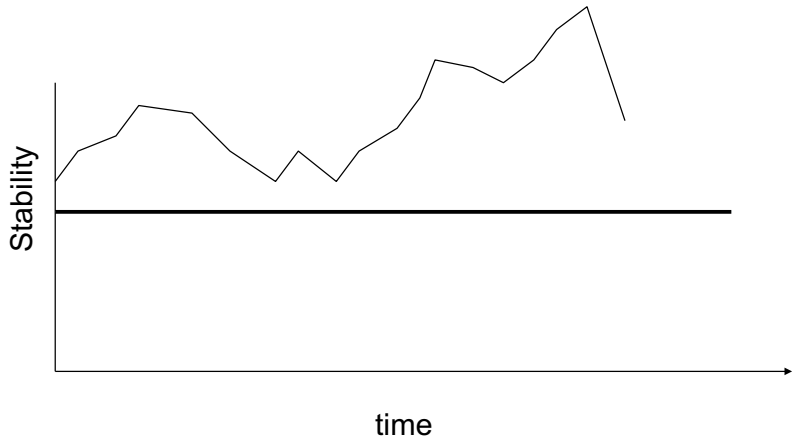
doi:10.1038/nature12678

Genetic incompatibilities are widespread within species

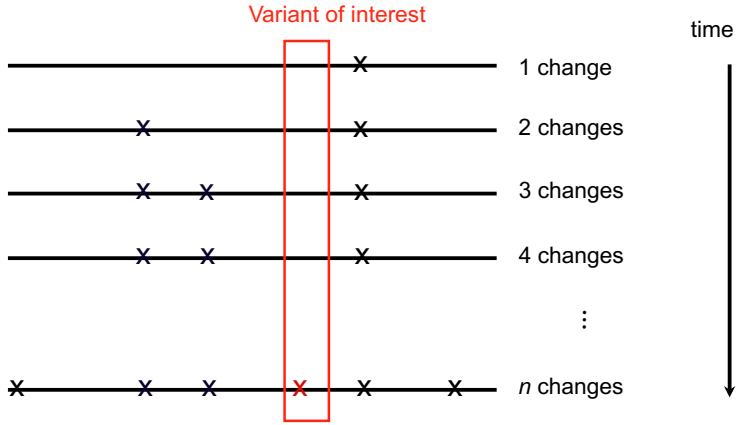
Russell B. Corbett-Detig[1], Jun Zhou[1], Andrew G. Clark[2,3], Daniel L. Hartl[1] & Julien F. Ayroles[1,2,4]

Numerous Dobzhansky-Muller incompatibilities in fly population

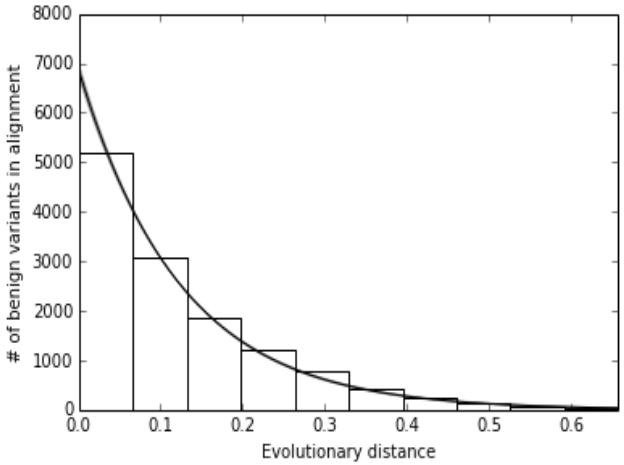
How complex genetic suppression can be?



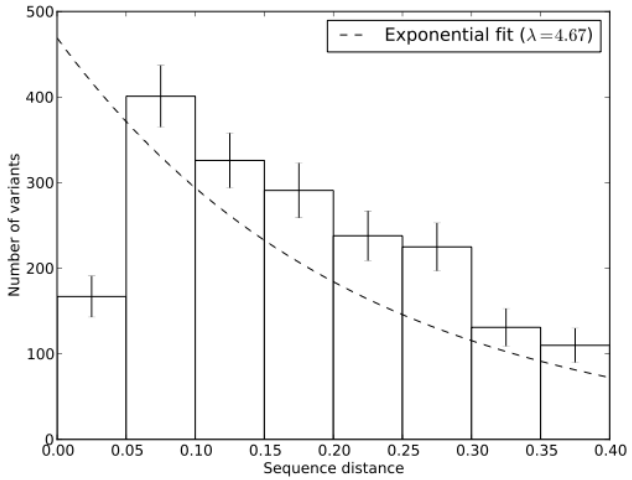
Model: accumulation of neutral variants



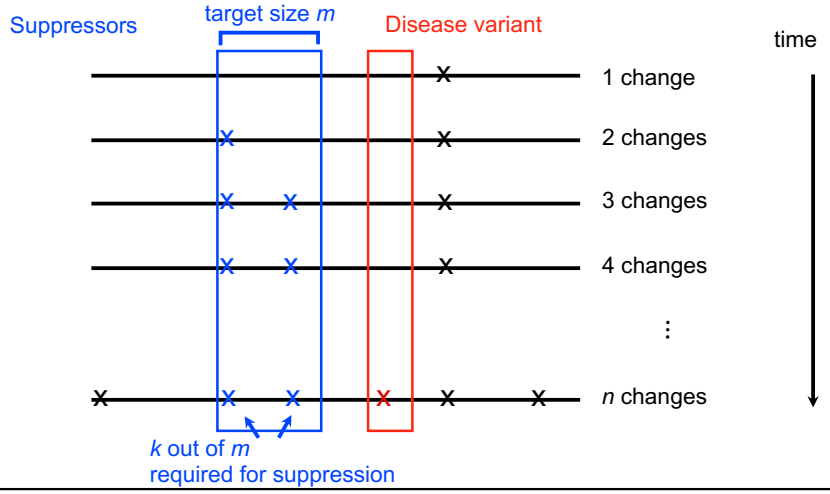
Exponential fit



Disease variants do not fit the Poisson expectation



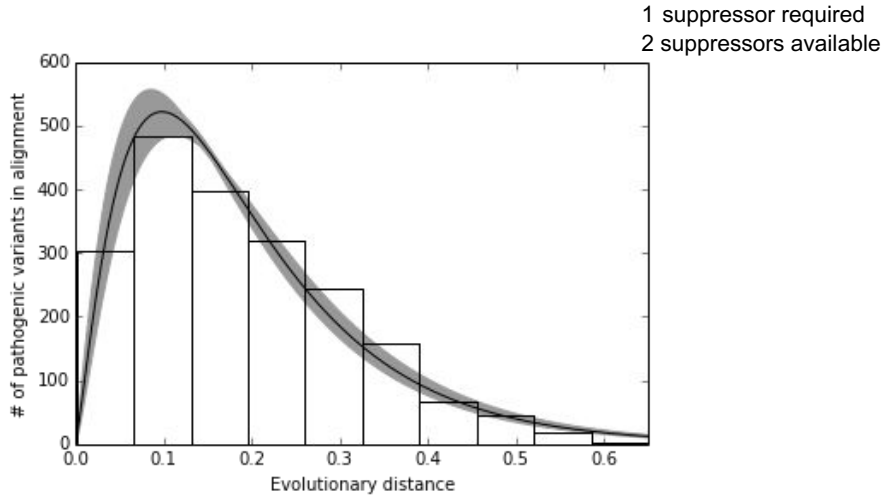
Model: accumulation of disease variants



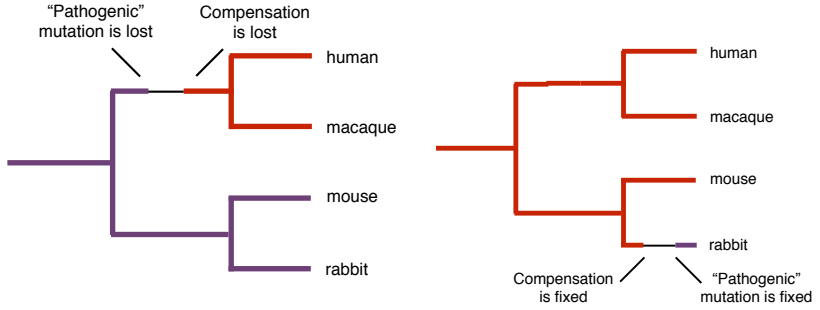
Model: Erlang distribution

$$L(k,t,\lambda) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!}$$

Model: fit for target size and number of compensatory changes



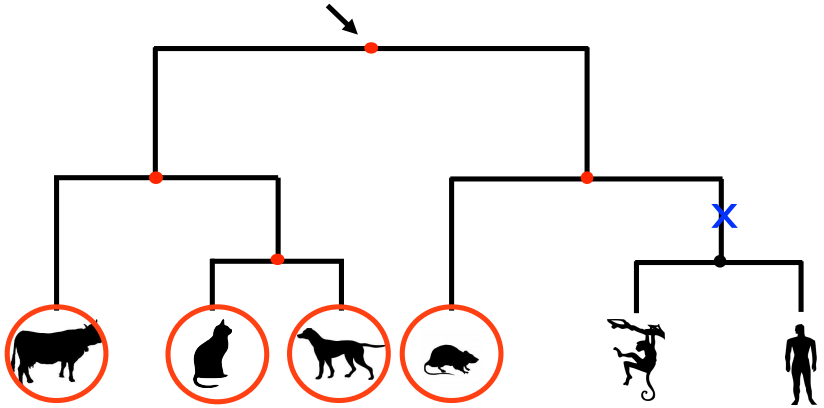
Evolutionary model of *cis*-complementation



Model Summary

- ~5-6% of human disease mutations have potential suppressors (i.e. are present in another mammalian species)
- In most cases, one large-effect suppressor is sufficient, out of only 1-2 available
- These values allow simple experiment to identify suppressors

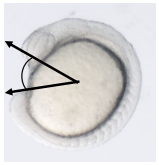
Predicting and testing suppressors



Zebrafish model

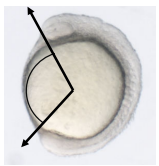
- Model of Bardet-Biedl Syndrome (obesity, renal failure, vision loss)
- Caused by defects in primary cilium
- Embryonic convergence / extension phenotype in zebrafish
- Easily scorable phenotype

Normal



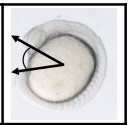
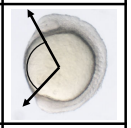
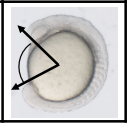
Class I

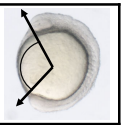
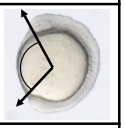
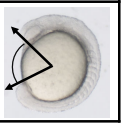
Class II



Images: Phoebe

Experiment interpretation

No injection	
Knockdown	
Rescue with human gene	

Human gene with disease mutant	
Double mutant (no suppression)	
Double mutant (full suppression)	

Images: Phoebe

Bardet-Biedl syndrome – BBS4 N165H

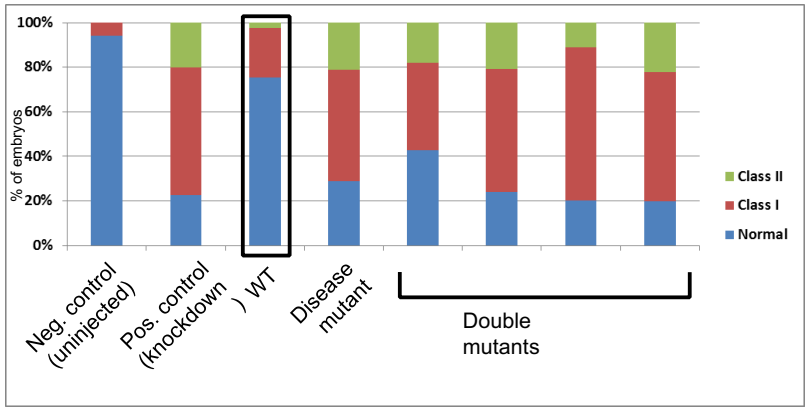
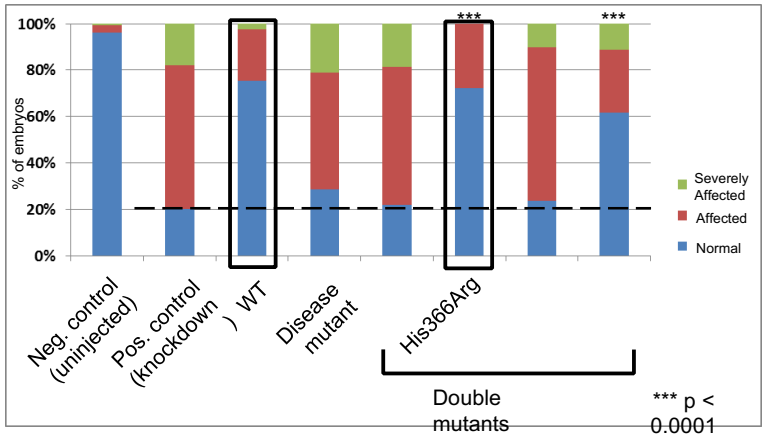


Figure: Stephan Frangakis

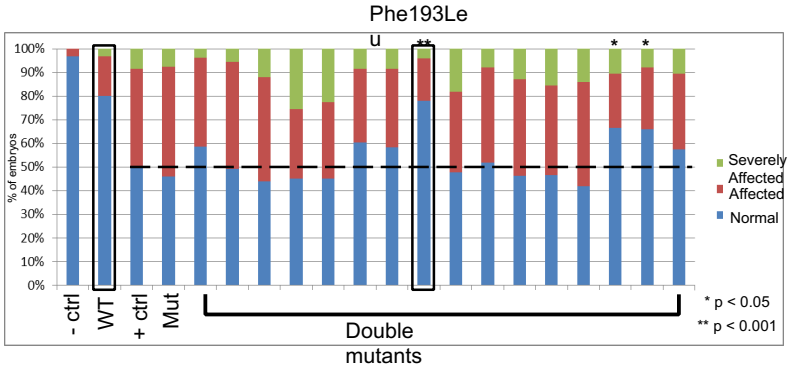
Bardet-Biedl syndrome – BBS4 N165H



Out of 9 candidates (7 shown here), 1 complete rescue

Figure: Stephan Frangakis

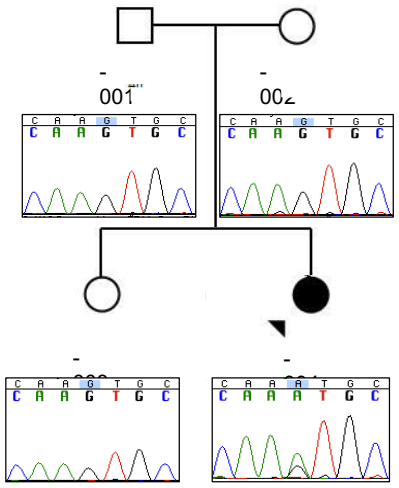
Bardet-Biedl syndrome – RPGRIP1L R937L



Out of 32 candidates, 1 complete rescue

Figure: Stephan Frangakis

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

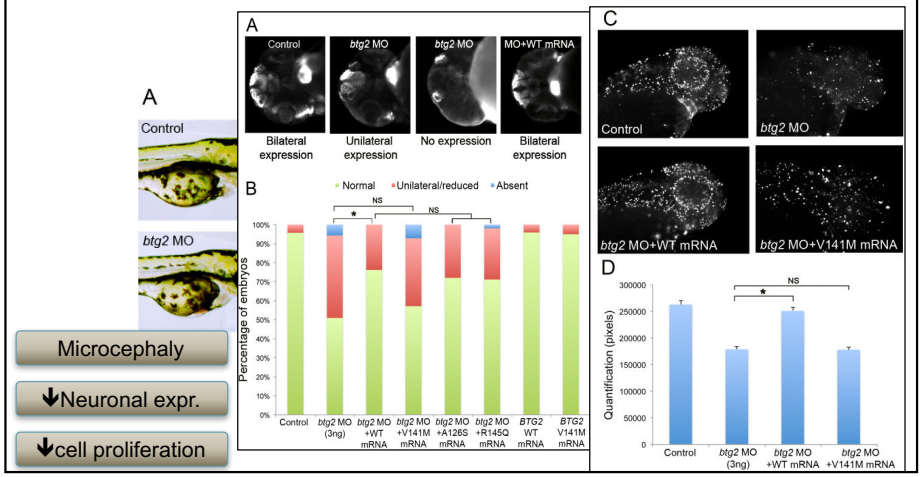
TTN Compound het

NOS2 De novo

LAMA1 Compound het

Stephan Frangakis

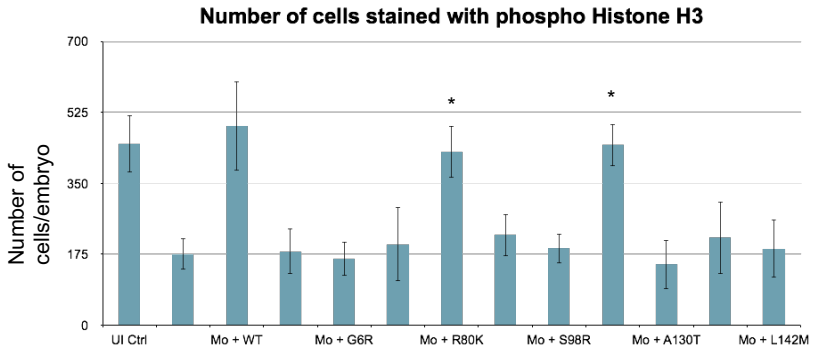
BTG2 is the disease driver



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

BTG2 has two compensatory mutations



*P<0.01 vs V141M rescue alone

Injection

Stephan Frangakis

Methods

PolyPhen2

SIFT

LRT

MutationTaster

FatHMM

SNPs3D

DeepSequence

Umbrella methods

Condel

REVEL

CADD

M-CAP

Incorporating regional constraint

CCR

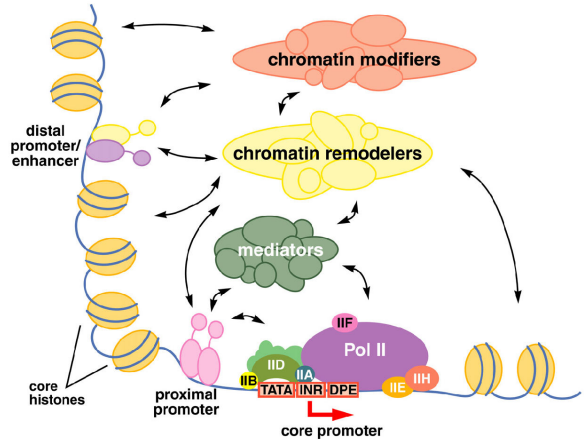
M-CAP

PrimateAI

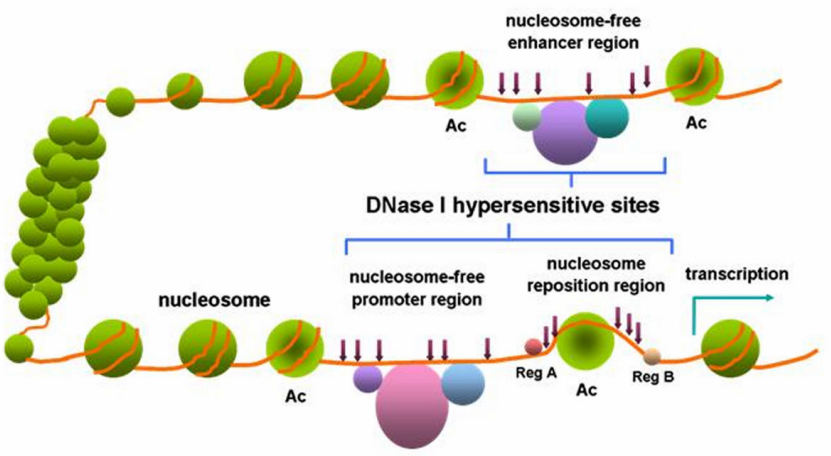
Non-coding variants

Regulatory variants

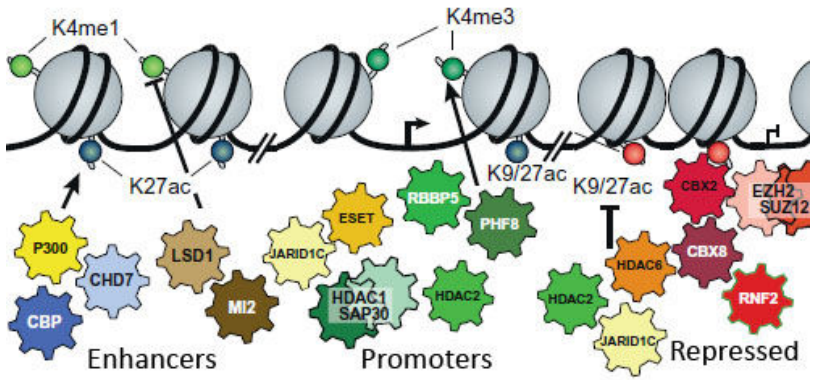
- Regulation: variants in promoters, enhancers, silencers, insulators



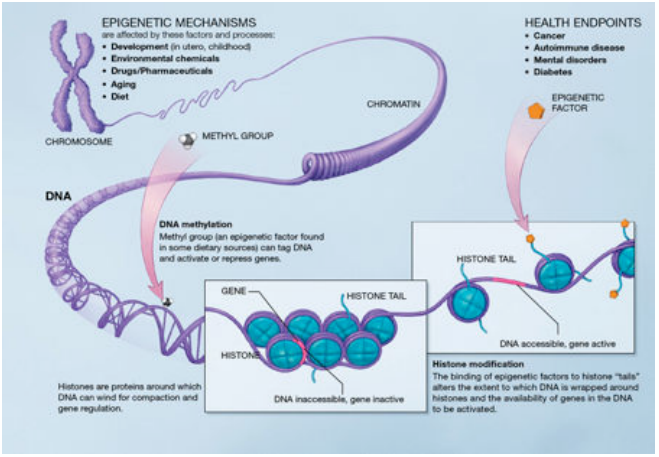
Chromatin accessibiliy



Chromatin modification



Epigenomics



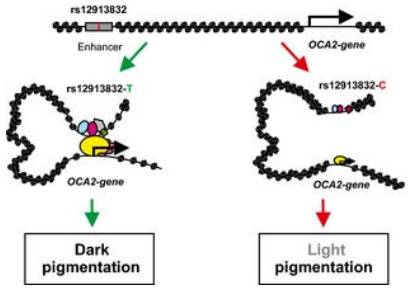
Why do we think that non-coding variation is of importance?

Regions and individual nucleotides conserved along phylogeny show signals of purifying selection in humans

Epigenetic studies report many well-localized regulatory marks

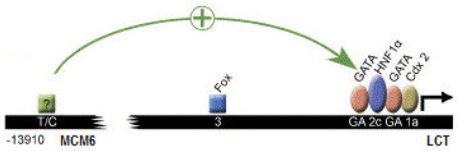
GWAS signals are predominantly located in non-coding regions

Non-disease alleles of large effect



Eye color

Lactase tolerance



Ultraconserved elements

OPEN ACCESS Freely available online PLOS BIOLOGY

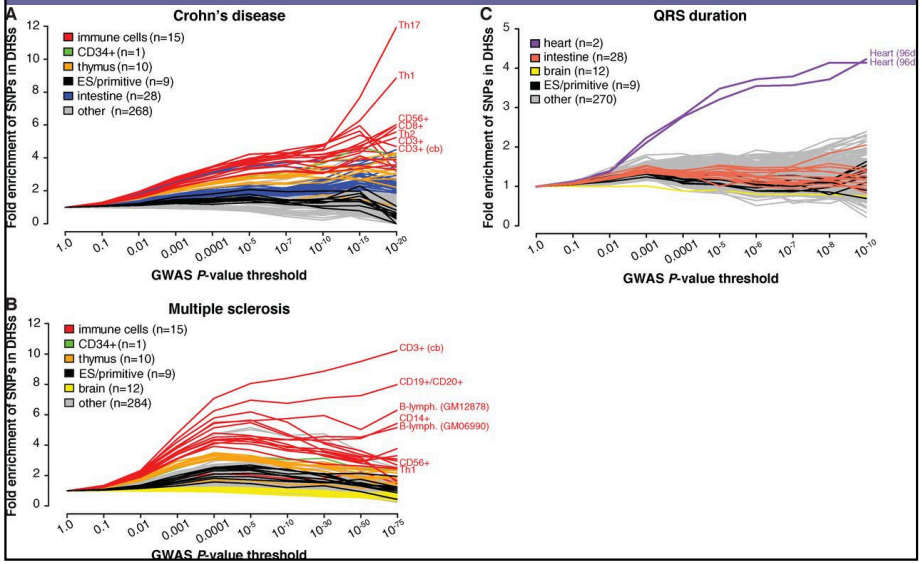
Deletion of Ultraconserved Elements Yields Viable Mice

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*]

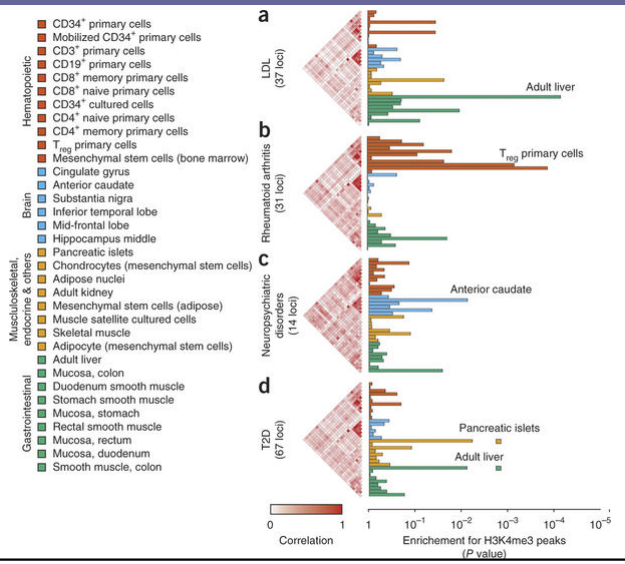
1 Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 2 United States Department of Energy Joint Genome Institute, Walnut Creek, California, United States of America

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.

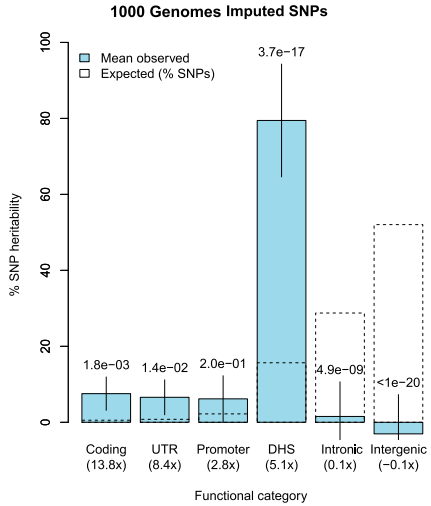
Enrichment of GWAS signals in putative regulatory elements



Enrichment of GWAS signals in putative regulatory elements



Partitioning heritability



GWAS SNPs co-localize with eQTLs

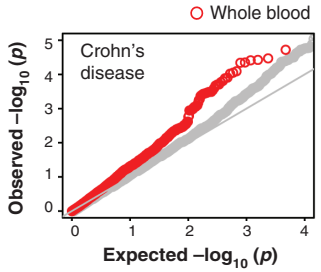
OPEN ACCESS Freely available online PLoS GENETICS

Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS

Dan L. Nicolae[1,2,3], Eric Gamazon[1], Wei Zhang[1], Shiwei Duan[1¤], M. Eileen Dolan[1,2], Nancy J. Cox[1,2*]

The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans

The GTEx Consortium*†



Translating GWAS findings into mechanistic models

GWAS peak

↓

Controlled model system

↓

Biochemistry

Human Genetics All the Way

GWAS peak

Endophenotype **Endophenotype**

Gene expression (eQTL)

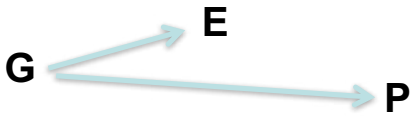
↓

Molecular phenotype (molecular_QTL)

Causality

Mediation

Independent effects

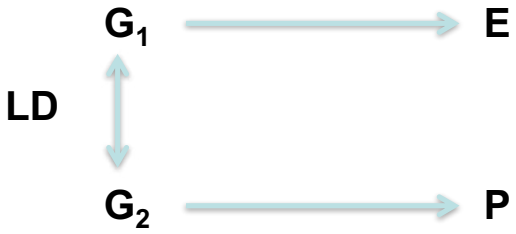


Reverse causation

Co-localization

Same causal variant

Distinct variants



Methods

Coloc

eCAVIAR

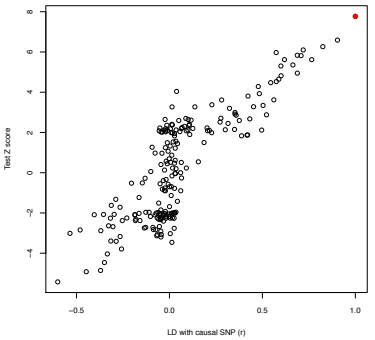
JLIM

PAINTOR's fine mapping model

- If a SNP is causal, then r^2 should predict association of other SNPs in the area:

  $$Z_2 \sim N(r_{12}\lambda_1, 1)$$

- Correlation between test statistics *Z* are approximated by MVN given local pairwise LD structure.

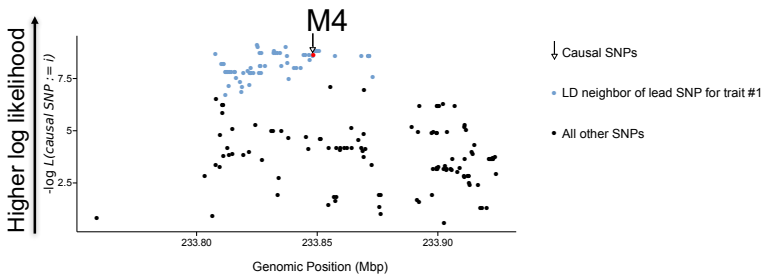


$$L(\mathbf{Z}|C = \{m\}; \lambda_m) = N(\mathbf{Z}; \Sigma(\lambda \circ C), \Sigma)$$
$$\propto e^{-\frac{1}{2}(\mathbf{Z}^T \Sigma^{-1} \mathbf{Z} - z_m^2)}$$

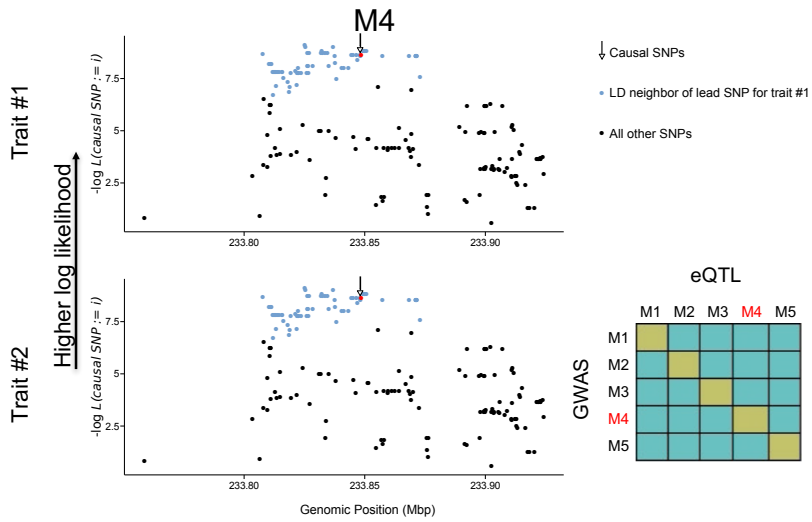
Parameters
λ: standardized effect size
Z: association statistic
C: indicator of causality
m: SNP considered

Kichaev *et al*. PLoS Genet. 2014; Chen *et al.* Genetics. 2015

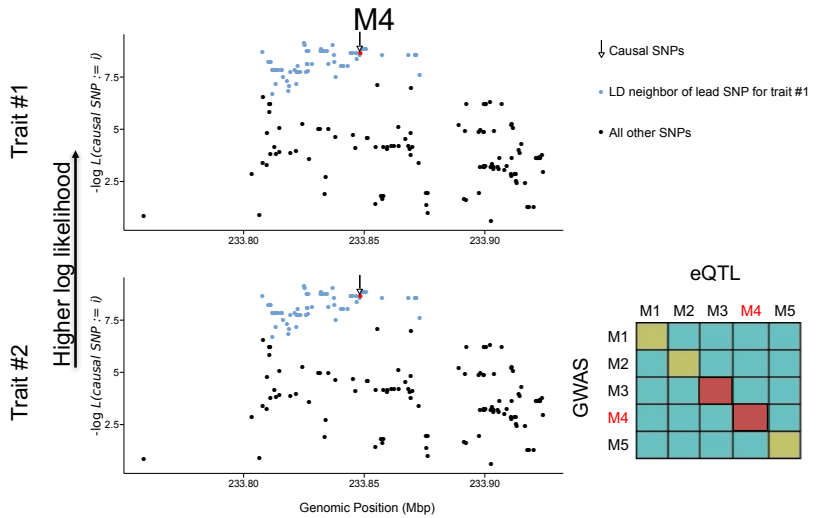
Likelihood of causal SNP



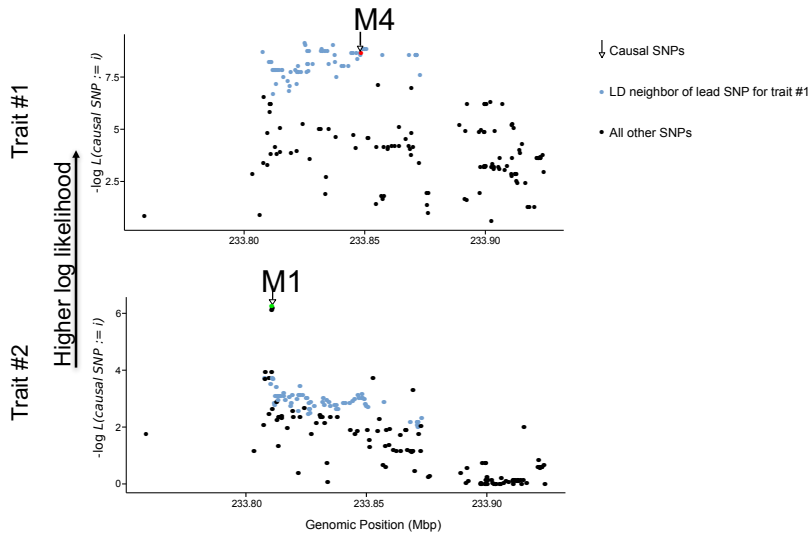
Joint likelihood: same causal variant



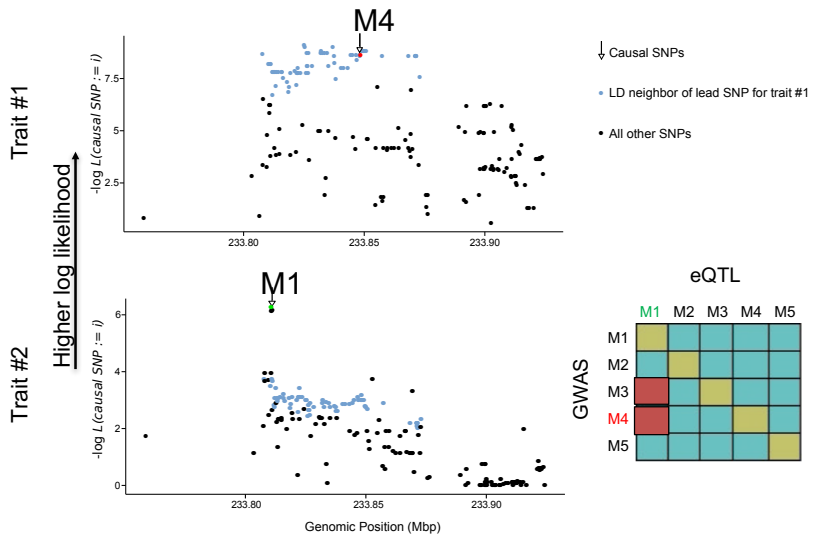
Joint likelihood: same causal variant



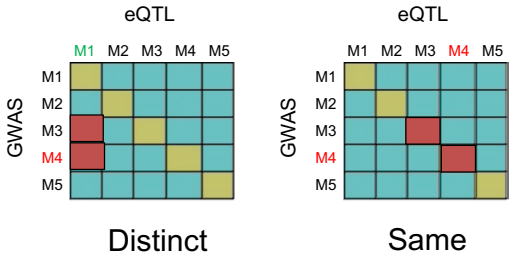
Joint likelihood: distinct variants



Joint likelihood: distinct variants



Joint likelihood test



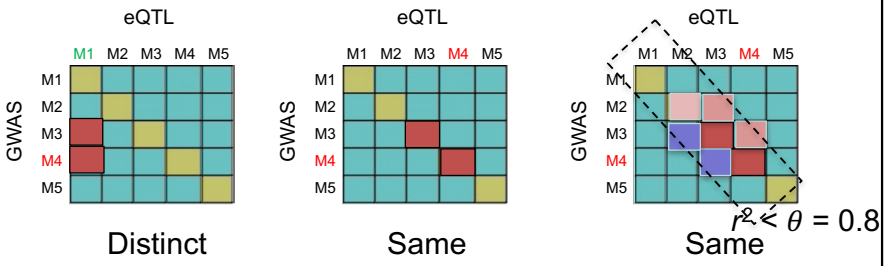
$$\Lambda = \sum_{r^2_{i,m^*>\theta}} L_1(i) \cdot \log \frac{L_1(\boldsymbol{i})L_2(\boldsymbol{i})}{\max_{r^2_{i,j}<\theta} L_1(\boldsymbol{i})L_2(\boldsymbol{j})}$$

$$= \sum_{r^2_{i,m^*>\theta}} e^{-\frac{1}{2}(z_i^2 - z_{m^*}^2)} \cdot (w_i^2 - \max_{r^2_{i,j}<\theta} w_j^2)$$

Parameters
z: disease association statistic
w: eQTL association statistic
m^*: lead disease-associated SN
θ: r^2 resolution limit

P-values can be estimated by permuting eQTLs

Joint likelihood test



$$\Lambda = \sum_{r^2_{i,m^*>\theta}} L_1(i) \cdot \log \frac{L_1(\boldsymbol{i})L_2(\boldsymbol{i})}{\max_{r^2_{i,j}<\theta} L_1(\boldsymbol{i})L_2(\boldsymbol{j})}$$

$$= \sum_{r^2_{i,m^*>\theta}} e^{-\frac{1}{2}(z_i^2 - z_{m^*}^2)} \cdot (w_i^2 - \max_{r^2_{i,j}<\theta} w_j^2)$$

Parameters
z: disease association statistic
w: eQTL association statistic
m^*: lead disease-associated SN
θ: r^2 resolution limit

P-values can be estimated by permuting eQTLs

Real data: autoimmune/inflammatory diseases

- Highly successful GWAS
- ImmunoChip: custom fine-mapping array
- Free availability of summary statistics on ImmunoBase
- Accessibility of disease relevant cell types and eQTLs data

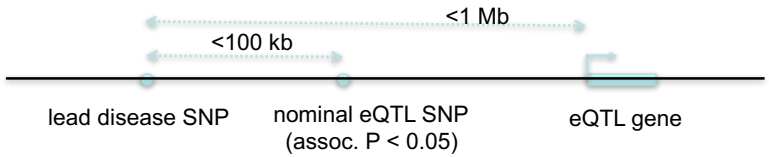
Disease	Densely genotyped [a]
MS	59
IBD	69
Crohn	19
UC	10
T1D	47
RA	34
CEL	34
Overall	272

* Excluding conditional hits
** Defined by ImmunoChip's densely genotyped fine-mapping intervals. Excluding MHC

Disease	Densely genotyped [a]	Number of loci: eQTL present [b] CD4+	CD14+	LCL	Total
MS	59				
IBD	69				
Crohn	19				
UC	10				
T1D	47				
RA	34				
CEL	34				
Overall	272				

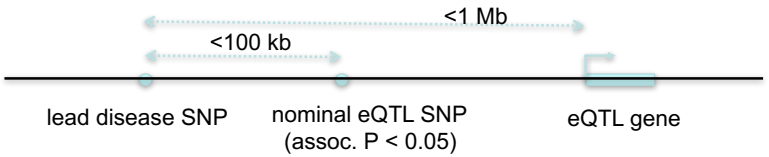
**** CD4/CD14+ (n=211/213) from Raj et al. Science 2014;
LCL (n=278) from Lappalainen et al. Nature 2013

Disease	Densely genotyped [a]	Number of loci: eQTL present [b] CD4+	CD14+	LCL	Total
MS	59				
IBD	69				
Crohn	19				
UC	10				
T1D	47				
RA	34				
CEL	34				
Overall	272				



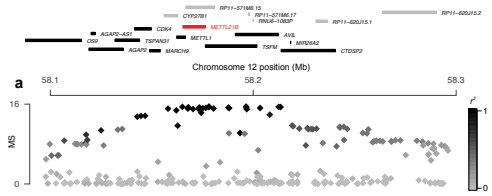
**** CD4/CD14+ (n=211/213) from Raj et al. Science 2014;
LCL (n=278) from Lappalainen et al. Nature 2013

Disease	Densely genotyped [a]	Number of loci: eQTL present [b] CD4+	CD14+	LCL	Total
MS	59	54	55	55	56
IBD	69	69	69	68	69
Crohn	19	18	18	18	18
UC	10	10	9	10	10
T1D	47	39	40	36	40
RA	34	34	34	34	34
CEL	34	34	34	34	34
Overall	272	258	259	255	261

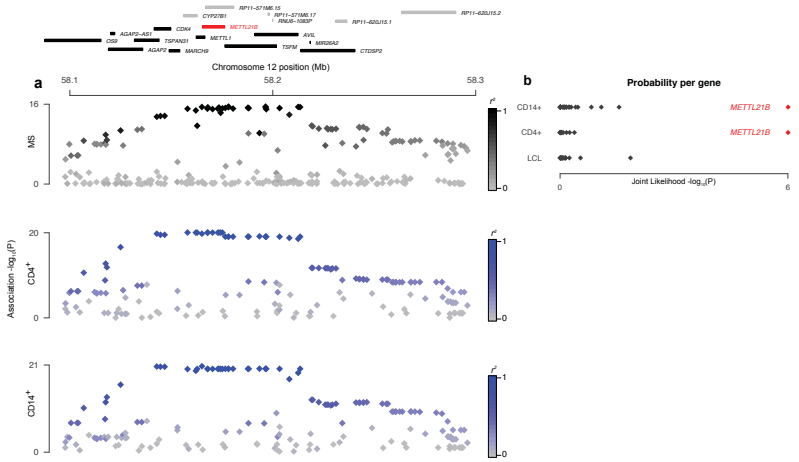


**** CD4/CD14+ (n=211/213) from Raj et al. Science 2014;
LCL (n=278) from Lappalainen et al. Nature 2013

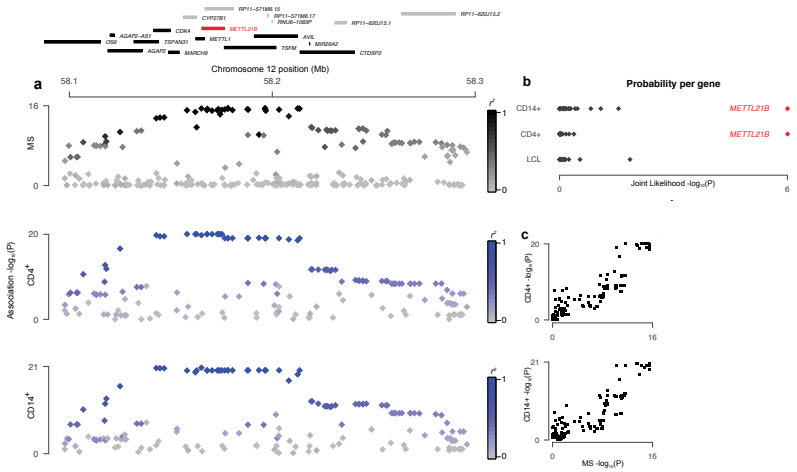
METTL21B eQTLs in CD4⁺ and CD14⁺ are consistent with association to MS



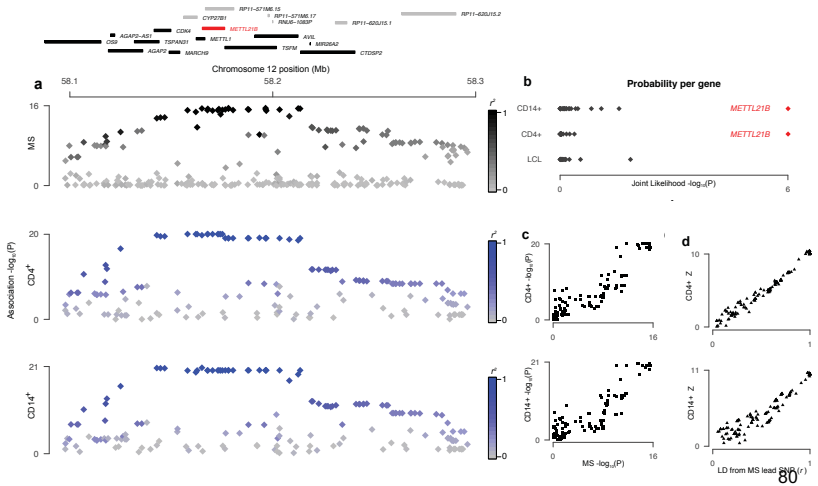
METTL21B eQTLs in CD4⁺ and CD14⁺ are consistent with association to MS



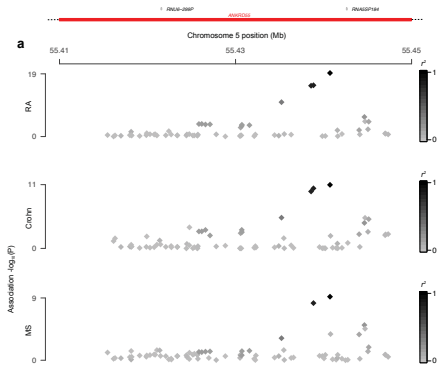
METTL21B eQTLs in CD4⁺ and CD14⁺ are consistent with association to MS



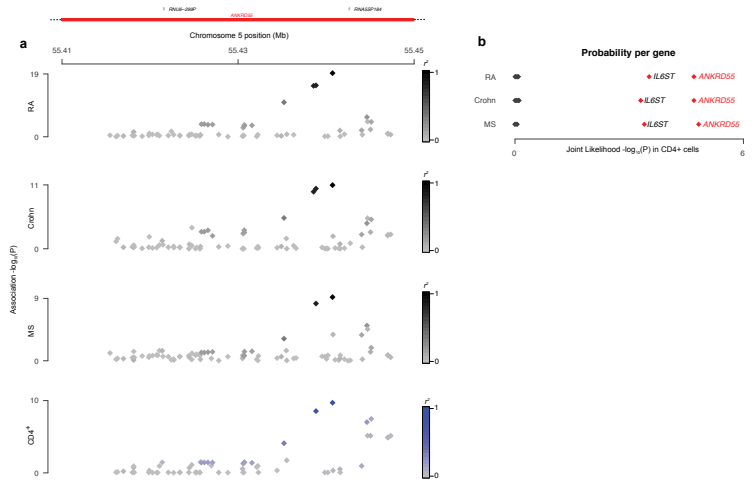
METTL21B eQTLs in CD4⁺ and CD14⁺ are consistent with association to MS



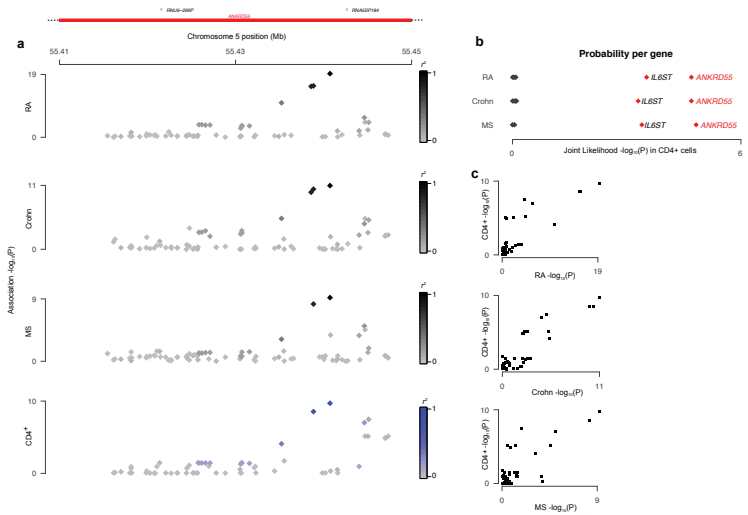
A GWAS peak shared across RA, Crohn, and MS is consistent with *ANKRD55* eQTL in T cells



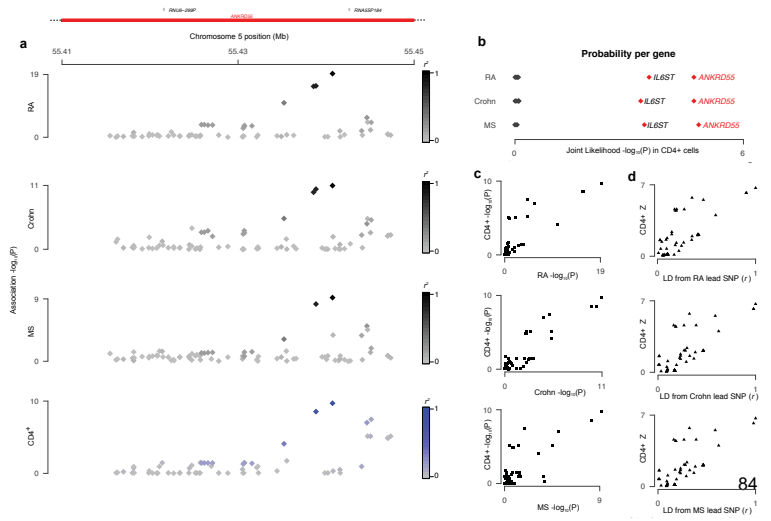
ANKRD55 eQTL in CD4+ is consistent with association with MS, Crohn, and RA



ANKRD55 eQTL in CD4+ is consistent with association with MS, Crohn, and RA



ANKRD55 eQTL in CD4+ is consistent with association with MS, Crohn, and RA

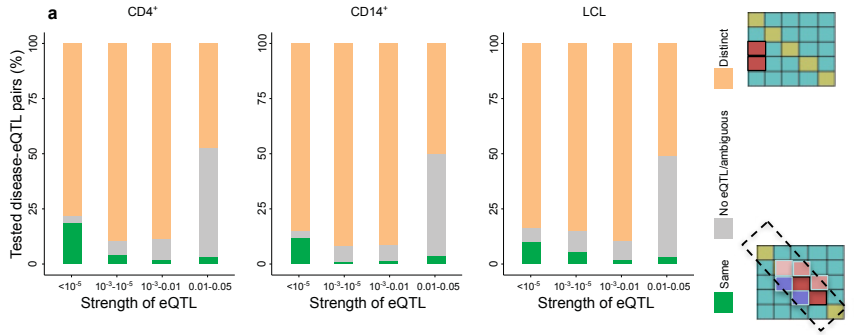


15% of disease loci have eQTL driven by the same variant (FDR 5%)

Disease	Number of loci							
	Densely genotyped [a]	eQTL present [b]				Driven by same effect [c]			
		CD4+	CD14+	LCL	Total	CD4+	CD14+	LCL	Total
MS	59	54	55	55	56	8	3	6	12
IBD	69	69	69	68	69	6	9	1	12
Crohn	19	18	18	18	18	2	1	0	3
UC	10	10	9	10	10	2	1	3	4
T1D	47	39	40	36	40	2	0	0	2
RA	34	34	34	34	34	2	0	1	3
CEL	34	34	34	34	34	3	2	0	5
Overall	272	258	259	255	261	25	16	11	41

* 75% of hits pass Bonferroni threshold as well.

~75% of tests disease eQTL pairs are driven by distinct variants



n=3,256 n=3,360 n=2,653

Summary on eQTLs

- ~15% of GWAS loci were mapped to eQTL genes.
- ~25% of GWAS loci are driven by the eQTLs of same effect.
- JLIM software

Methods

GWAVA (supervised)

CADD (predicts loss of genetic variation)

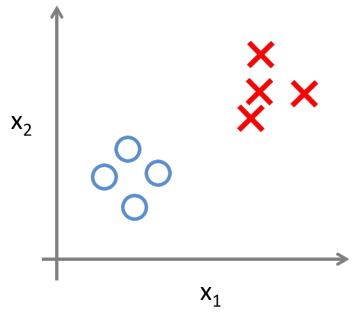
INSIGHT / LINSIGHT (population genetics)

Eigen (eigenvector in the annotation space)

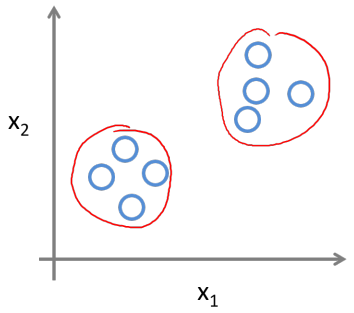
PINES (phenotype-specific)

Prediction Methods

Supervised Learning

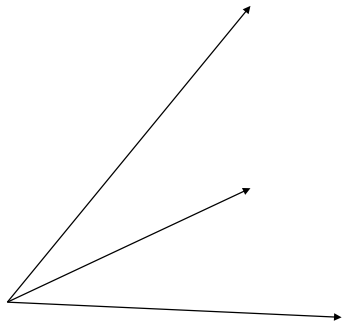


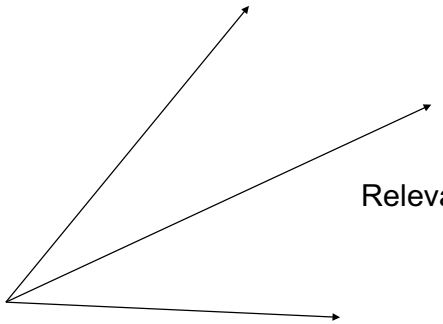
Unsupervised Learning

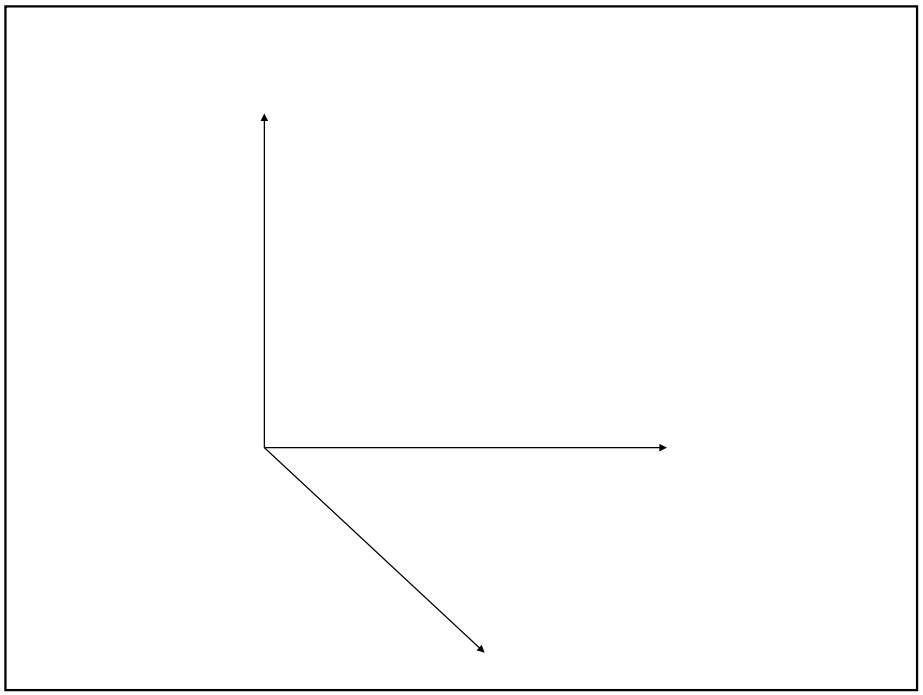


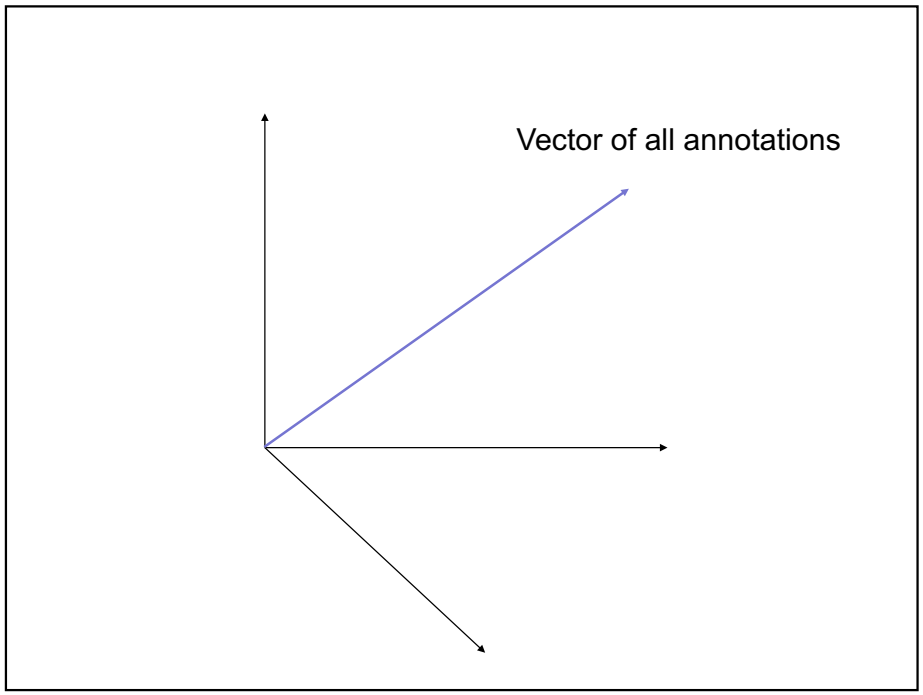
PINES

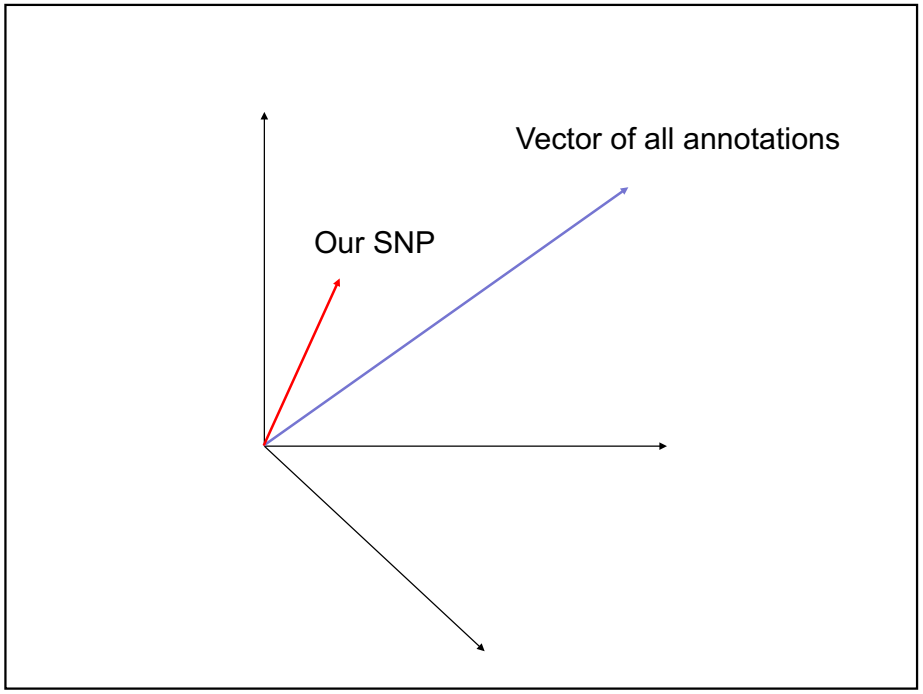
- Take all annotations and create an uncorrelated space
- Upweight axes corresponding to relevant cell/tissue types
- The score is based on the angle with the direction of the maximal possible annotation

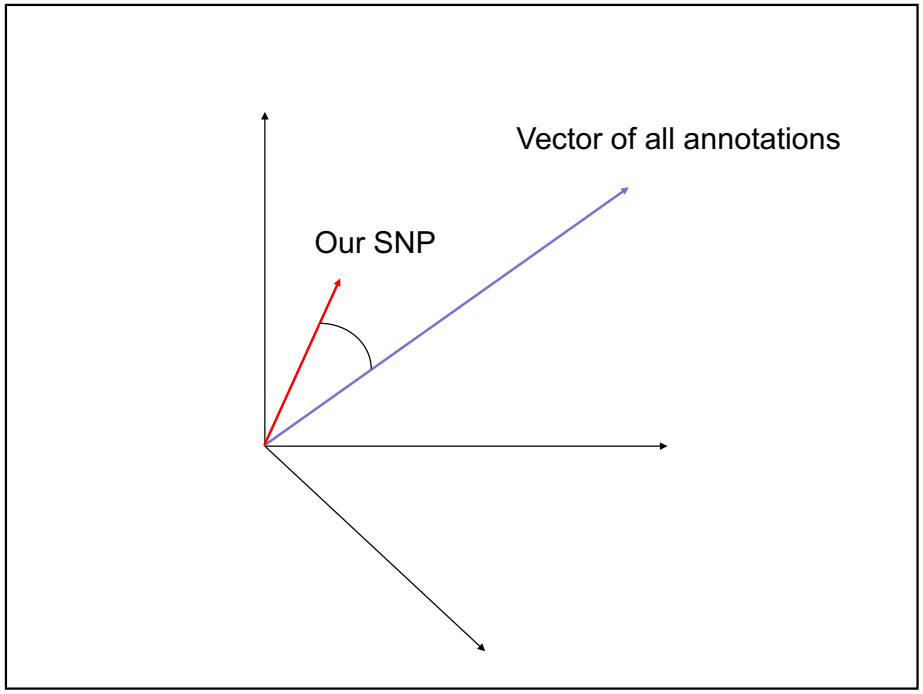




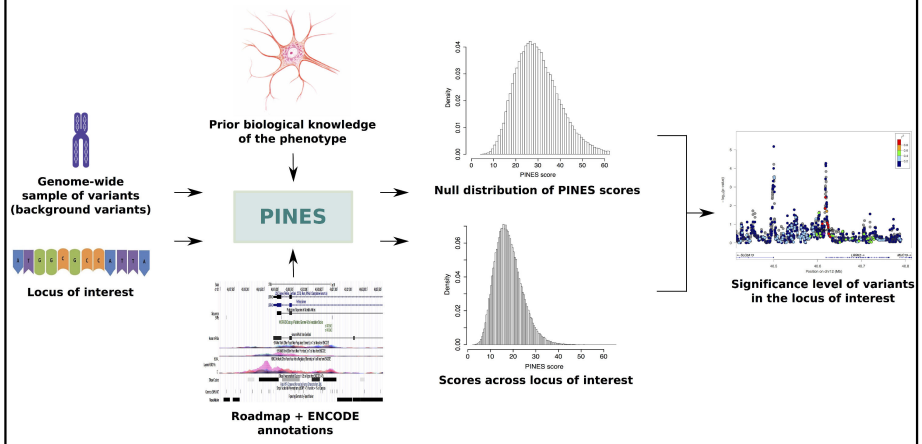


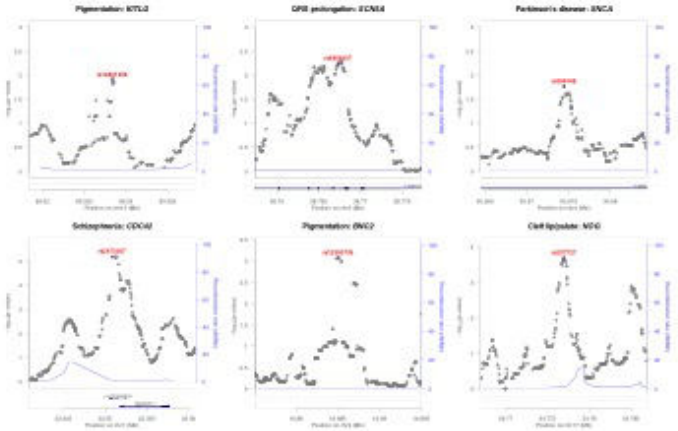

Vector of all annotations


Vector of all annotations
Our SNP


Vector of all annotations
Our SNP

PINES





PINES: http://genetics.bwh.harvard.edu/pines/