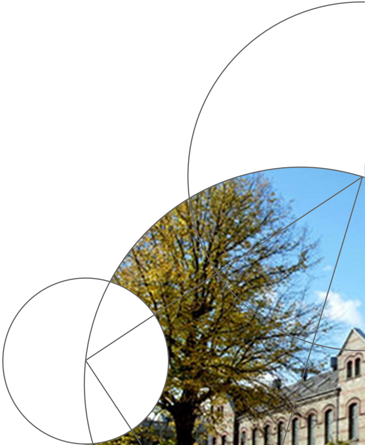
Faculty of Health Sciences

Statistical methods in bioinformatics

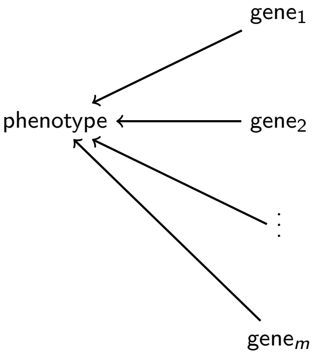
Genome-wide association studies

Claus Thorn Ekstrøm
Biostatistics,
University of Copenhagen
E-mail: ekstrom@sund.ku.dk

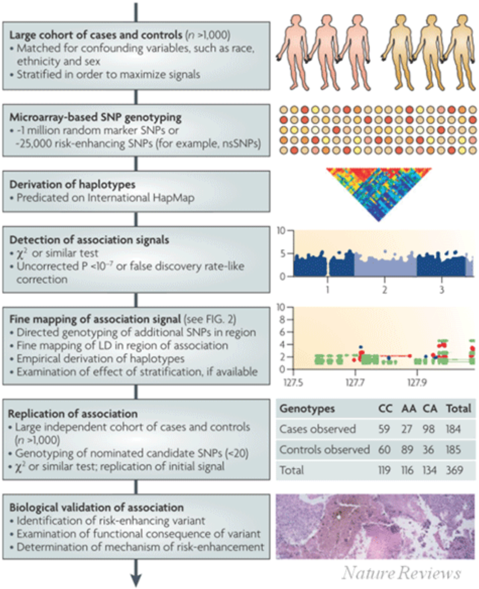


Overview

- Discover regions in the genome associated with disease/trait
- Look for variations in that occur more frequently in people with a particular disease than in people without the disease
- Consider thousands/millions of SNPs at the same time



GWAS overview — steps



Large cohort of cases and controls (*n* >1,000)

- Matched for confounding variables, such as race, ethnicity and sex
- Stratified in order to maximize signals

Microarray-based SNP genotyping

- -1 million random marker SNPs or -25,000 risk-enhancing SNPs (for example, nsSNPs)

Derivation of haplotypes

- Predicated on International HapMap

Detection of association signals

- χ^2 or similar test
- Uncorrected $P < 10^{-7}$ or false discovery rate-like correction

Fine mapping of association signal (see FIG. 2)

- Directed genotyping of additional SNPs in region
- Fine mapping of LD in region of association
- Empirical derivation of haplotypes
- Examination of effect of stratification, if available

Replication of association

- Large independent cohort of cases and controls (*n* >1,000)
- Genotyping of nominated candidate SNPs (<20)
- χ^2 or similar test; replication of initial signal

Biological validation of association

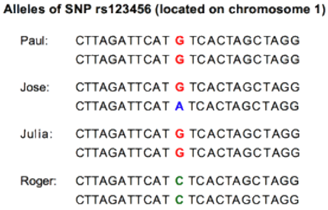
- Identification of risk-enhancing variant
- Examination of functional consequence of variant
- Determination of mechanism of risk-enhancement

Genotypes	CC	AA	CA	Total
Cases observed	59	27	98	184
Controls observed	60	89	36	185
Total	119	116	134	369

Nature Reviews

What is a SNP?

A single nucleotide polymorphism is a DNA sequence variation that occurs when a single nucleotide in the genome is altered (in a significant proportion of the population, say $>$ 1%).



Alleles of SNP rs123456 (located on chromosome 1)

Paul:	CTTAGATTCAT G TCACTAGCTAGG CTTAGATTCAT G TCACTAGCTAGG
Jose:	CTTAGATTCAT G TCACTAGCTAGG CTTAGATTCAT A TCACTAGCTAGG
Julia:	CTTAGATTCAT G TCACTAGCTAGG CTTAGATTCAT G TCACTAGCTAGG
Roger:	CTTAGATTCAT C TCACTAGCTAGG CTTAGATTCAT C TCACTAGCTAGG

- Human genome sequence is 99.9% identical in all people.
- Almost all common SNPs have only two alleles.
- Quite abundant — roughly 1/1000 bases
- A SNP close to a gene acts as marker for that gene.

SNP data

One of the two alleles will be the least frequent in the population. Its frequency is called minor allele frequency (MAF).

Each individual has 0, 1 or 2 copies of the rare allele.

id	bmi	g1	g2	g3	g4	g5	g6	g7
1	23	0	2	0	1	0	0	1
2	31	0	1	0	0	2	0	1
3	26	0	1	1	0	0	0	0
4	35	2	1	1	0	2	1	0

Each gene corresponds to the genotype at the particular SNP.

SNP data

Statistical model

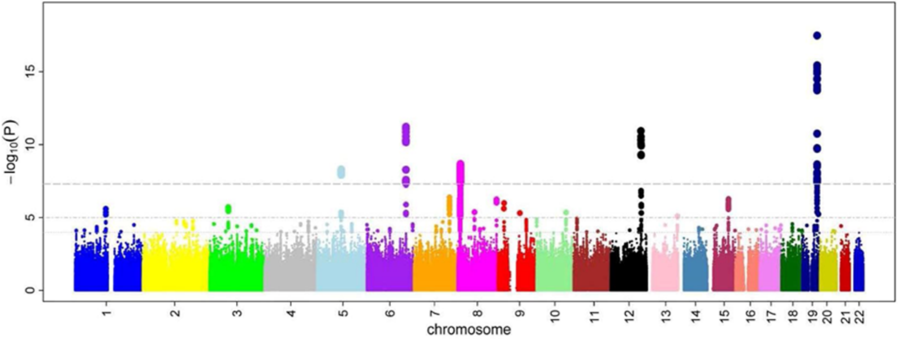
$$\text{bmi}_i = \mathbf{X}\beta + \varepsilon_i = \alpha + \sum_{j=1}^{m} \beta x_{im} + \varepsilon_i$$

- Problem with large number of predictors — we need the methods from this Monday.
- How do we handle more complex systems? Multiple genes at the same time?
- Missing data
- External information

Single marker analysis — standard approach

- Identify SNPs where one allele is significantly associated to the outcome (quantitative or binary).
- Identify chromosomal regions where one haplotype is significantly associated to the outcome (quantitative or binary).
- Run an analysis for each SNP! Will need the methods we introduced on Monday.

Manhattan plot



Standard approach — "millions" of t tests

```
> library("data.table")         # Get package
> geno <- fread("hapmap1.ped")  # Read file
> for(j in seq_along(geno)){    # Convert 0 to NA
+          set(geno, i=which(geno[[j]]==0), j=j, value=NA)
+ }
> pheno <- fread("qt.phe")      # Read phenotypes
> geno$V1 <- NULL ; geno$V2 <- NULL ; geno$V3 <- NULL ;
> geno$V4 <- NULL ; geno$V5 <- NULL ; geno$V6 <- NULL
> # Collapse pairs of alleles to genotype
> geno2 <- geno[, lapply(1:(ncol(.SD)/2),
+               function(x) sum(.SD[[2*x-1]], .SD[[2*x]])-2),
+               by = 1:nrow(geno),
+               .SDcols = grep('^V', names(geno), value = TRUE)]
> keep <- which(apply(geno2, 2, sd) > 0) # Remove those with no var
> geno2 <- geno2[,keep, with=FALSE]
> geno2$nrow <- NULL
```

Standard approach — "millions" of t tests

```
> library("MESS")
> mres <- mfastLmCpp(pheno$V3, as.matrix(geno2))
> pval <- 2*pt(-abs(mres[,3]), df=nrow(pheno)-2)
> head(sort(pval))

[1] 5.272596e-09 1.095443e-06 1.630357e-05 2.885730e-05 3.580539e-05
[6] 3.715969e-05

> head(names(geno2)[order(pval)])

[1] "V10602" "V81525" "V37137" "V12225" "V18546" "V53636"
```

Controlling for genomic controls

There can be a "shift" in p-values due to unfulfilled assumptions — they become too small.

- Most markers should be unassociated
- Survey markers with a low prior probability of association with disease ("null markers")
- The inflation factor, λ, is the ratio of the *observed median value* of the χ^2-statistic for the null markers divided by the expected median value of the χ^2-statistic (approximately 0.456 for 1 df tests).
- If $\lambda > 1$ then downscale all subsequent statistics by λ.

```
> statistics <- mres[,3]^2 ; median(statistics)
[1] 0.5471164
> rescaled <- statistics/(median(statistics)/0.456)
> results <- 1-pchisq(rescaled, df=1)
> head(sort(results)) # p-values
[1] 3.274656e-09 1.697990e-06 3.075233e-05 5.555227e-05 6.93321[illegible]05
[6] 7.201973e-05
> head(names(geno2)[order(results)])
```

Standard approach — alternative?

```
> library(glmnet)
> res <- glmnet(as.matrix(geno2), pheno$V3)
> plot(res)
```

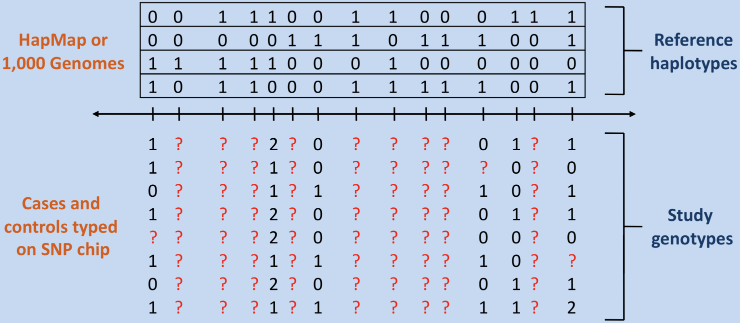
Problems with this approach?

Could we fix them?

Effects or p-values?

Digression — missing data

Typical imputation scenario



HapMap or 1,000 Genomes

Reference haplotypes

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

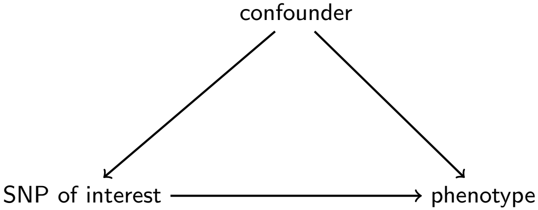
Cases and controls typed on SNP chip

Study genotypes

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

Another digression: association/causation

- Suppose that genotypes at a particular SNP are significantly associated with the outcome
- This may be because the SNP is associated with some other factor (a confounder), which is associated with outcome but is not in the same causal pathway.



confounder

SNP of interest

phenotype

Possible confounders

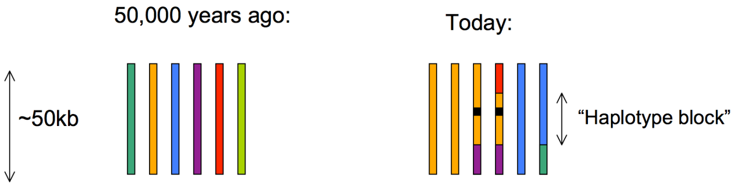
Possible confounders of genetic associations:

- Ethnic ancestry
- Genotyping batch, genotyping centre
- DNA quality

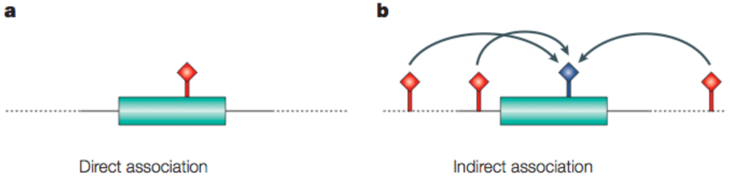
Need to control for possible confounders!
How can we do that? What if we haven't measured it?

Helpful confounding

Linkage disequilibrium (LD) is the non-independence of alleles at nearby markers in a population because of a lack of recombinations between the markers



Direct and indirect association testing



Functional SNP is genotyped and an association is found

Functional SNP (blue) is not genotyped, but a number of other SNPs (red), in LD with the functional SNP, are genotyped, and an association is found for these SNPs

Digression — Fisher's approach

Statistical model

$$\mathrm{bmi}_i = \mathbf{X}\beta + \varepsilon_i = \alpha + \sum_{j=1}^{m} \beta x_{im} + \varepsilon_i$$

Fisher assumed that there was a large number of genes each with small effect. Thus, the combined effect of the genes can be well approximated by a Gaussian distribution (law of large number idea).

In general we expect additivity of effects but that may not be correct.
What assumptions have we made so far?

The common disease — common variant hypothesis

The "Common Disease, Common Variant (CDCV)" hypothesis argues that genetic variations with appreciable frequency in the population at large, but relatively low "penetrance" are the major contributors to genetic susceptibility to common diseases.

- GWASs have identified thousands of common variants.
- The infinitesimal model is standard genetic theory.
- Common variants collectively capture most of the genetic variance in GWASs

But:

- The missing heritability has not been accounted for.
- The quantitative trait locus (QTL) paradox: QTLs that are consistently detected in pedigrees and in experimental crosses are not observed in outbred populations.
- Very few common variants have been functionally validated

The common disease — rare variant hypothesis

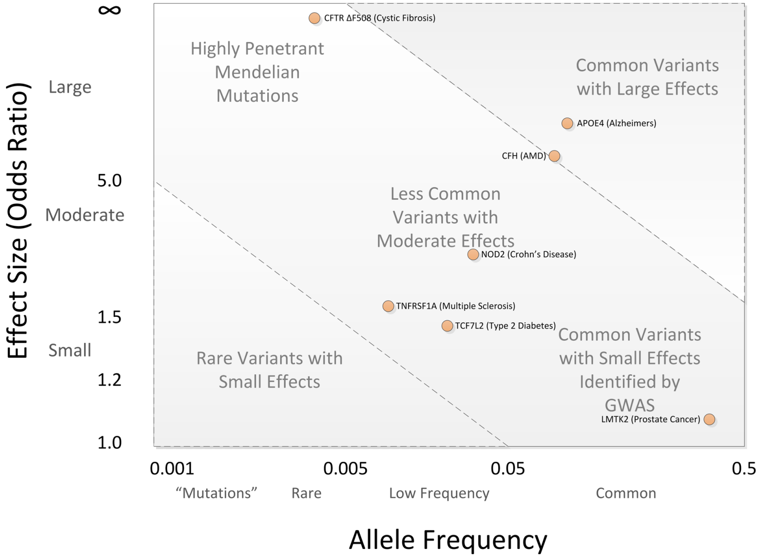
The "Common Disease, Rare Variant (CDRV)" hypothesis, on the contrary, argues that multiple rare DNA sequence variations, each with relatively high penetrance, are the major contributors to genetic susceptibility to common diseases.

- Evolutionary theory predicts that disease alleles should be rare
- Many rare familial disorders are due to rare alleles of large effect
- Empirical population genetic data show that deleterious variants are rare

But:

- Simulation of the allele frequency distribution of data from genome-wide association studies (GWASs) is not consistent with rare variant explanations.
- Genome-wide associations are consistent across populations

What can we hope to find?



Which markers can we hope to find?

Recall the t test statistic for a single explanatory variable (the kth):

$$T = \frac{\hat{\beta}_k - 0}{\mathrm{SE}(\hat{\beta}_k)} = \frac{\hat{\beta}_k - 0}{s/SS_x} = \frac{(\hat{\beta}_k) \cdot SS_x}{s},$$

where SS_x is $(n-1) \cdot Var(X_k)$.

Note that variance of a single marker, X, with MAF p is (assuming Hardy-Weinberg equilibrium):

Number of alleles	0	1	2
Frequency	$(1-p)^2$	$2 \cdot p \cdot (1-p)$	p^2

The variation of X is $2 \cdot p \cdot (1-p)$.
Smaller $p \Rightarrow$ smaller variance $\Rightarrow$ harder to detect an effect.

GWAS geared towards detection of “common” genes.

Heritability

Observed phenotype Y is given as

$$Y = \underbrace{A + D}_{\text{Genes}} + \underbrace{C + E}_{\text{Environment}} \tag{1}$$

such that

$$Cov(Y) = \sigma_a^2 2\Phi + \sigma_d^2 \Delta_7 + \sigma_c^2 J + \sigma_e^2 I.$$

Partition the variance into 4 sources:

$$\text{Var}(Y) = \sigma_a^2 + \sigma_d^2 + \sigma_c^2 + \sigma_e^2. \tag{2}$$

Divide the ACDE model (2) by the total phenotypic variance, Var(Y):

$$1 = \underbrace{h^2 + d^2}_{H^2} + c^2 + e^2 \tag{3}$$

Heritability estimation requirements

Classical approach: need families

Twins: compare correlation between MZ twins to correlations between same-sex DZ twins.

Families: look at contribution based on familial/genetic relationship to the total variance.

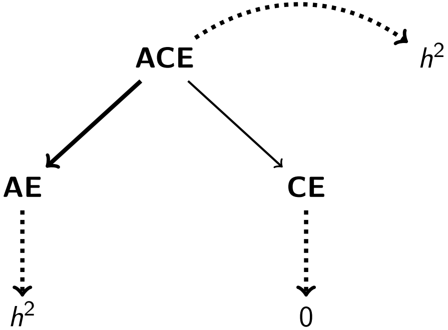
Example — heritability

	n	$h^2 \pm \text{SE}$
% body fat	296	0.47 ± 0.13
Waist-to-hip ratio	301	0.38 ± 0.13
BMI	305	0.46 ± 0.14
Serum leptin	269	0.25 ± 0.12
Birth length	249	0.44 ± 0.12
Birth weight	249	0.89 ± 0.09

The case of the missing hertability...

The case of the missing heritability

ბიგ მაკი
Coca-Cola
Coca-Cola


ACE
h^2
AE
CE
h^2
0

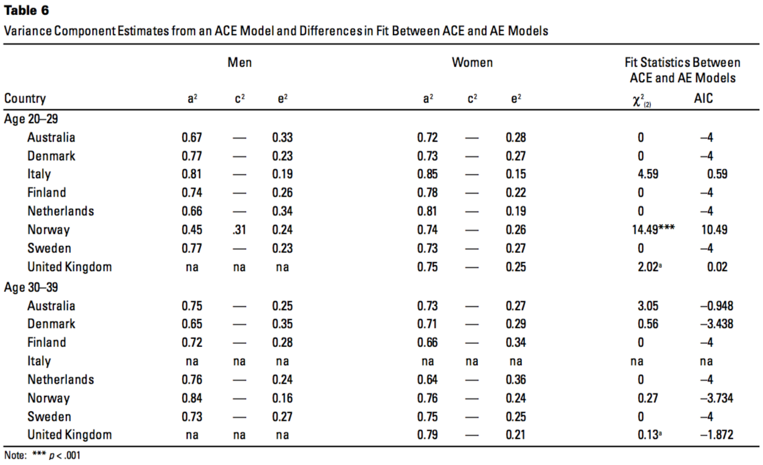


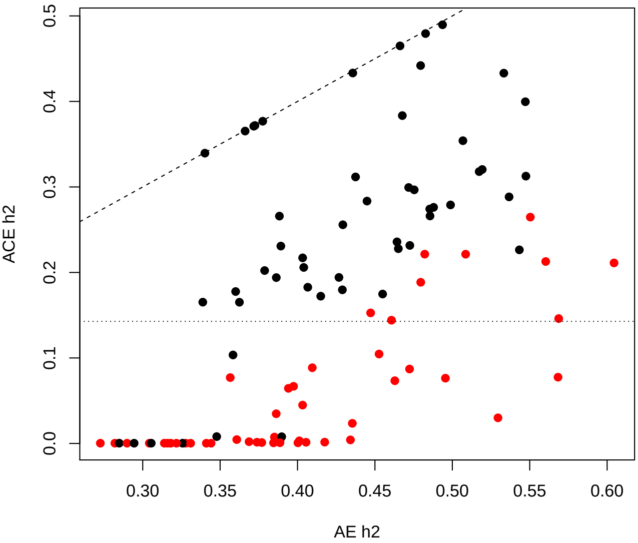
Table 6

Variance Component Estimates from an ACE Model and Differences in Fit Between ACE and AE Models

	Men			Women			Fit Statistics Between ACE and AE Models	
Country	a^2	c^2	e^2	a^2	c^2	e^2	$\chi^2_{(2)}$	AIC
Age 20–29								
Australia	0.67	—	0.33	0.72	—	0.28	0	–4
Denmark	0.77	—	0.23	0.73	—	0.27	0	–4
Italy	0.81	—	0.19	0.85	—	0.15	4.59	0.59
Finland	0.74	—	0.26	0.78	—	0.22	0	–4
Netherlands	0.66	—	0.34	0.81	—	0.19	0	–4
Norway	0.45	.31	0.24	0.74	—	0.26	14.49***	10.49
Sweden	0.77	—	0.23	0.73	—	0.27	0	–4
United Kingdom	na	na	na	0.75	—	0.25	2.02[a]	0.02
Age 30–39								
Australia	0.75	—	0.25	0.73	—	0.27	3.05	–0.948
Denmark	0.65	—	0.35	0.71	—	0.29	0.56	–3.438
Finland	0.72	—	0.28	0.66	—	0.34	0	–4
Italy	na	na	na	na	na	na	na	na
Netherlands	0.76	—	0.24	0.64	—	0.36	0	–4
Norway	0.84	—	0.16	0.76	—	0.24	0.27	–3.734
Sweden	0.73	—	0.27	0.75	—	0.25	0	–4
United Kingdom	na	na	na	0.79	—	0.21	0.13[a]	–1.872

Note: *** $p < .001$

$h^2 = 0.15, c^2 = 0.25$



Approximating “heritability” from unrelated individiuals

So what can we do if we want to estimate heritability from unrelated individuals?

Approximating “heritability” from unrelated individiuals

So what can we do if we want to estimate heritability from unrelated individuals?

Need an estimate of the genetic correlation among individuals in order to estimate the heritability. However, they are assumed unrelated.

Approximating "heritability" from unrelated individiuals

So what can we do if we want to estimate heritability from unrelated individuals?

Need an estimate of the genetic correlation among individuals in order to estimate the heritability. However, they are assumed unrelated.

Alternative: estimate the proportion of the variance explained by common SNPs in total (through LD with causal variants)

“Heritability” from unrelated individiuals

Statistical model:

$$Y = \mathbf{X}\beta + \mathbf{W}u + \varepsilon$$

where $\mathbf{W}$ is the $n \times m$ matrix of standardized genotypes:

$$w_{ij} = \frac{(x_{ij} - 2p_j)}{\sqrt{2 \cdot p_j \cdot (1 - p_j)}}$$

Note that the variance of Y is

$$\text{Var}(Y) = \mathbf{W}\mathbf{W}^t \sigma_u^2 + \sigma_e^2 I$$

Note that the size of the diagonals of $\mathbf{W}\mathbf{W}^t$ increases with the number of markers, m. So we standardize it:

$$\text{Var}(Y) = \frac{\mathbf{W}\mathbf{W}^t}{m} \sigma_g^2 + \sigma_e^2 I$$

Estimate of proportion of variance explained: $\frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$

Combined marker analysis

Until now: a series of single marker analyses. Possibly penalized regression

Combined marker analysis

Until now: a series of single marker analyses. Possibly penalized regression

Use external / additional information

- Multiple markers in the same region provide more information/better LD with the causal gene
- Multiple markers in the same biological pathway/set provide more information
- Gene-set enrichment analysis

Haplotypes

We observe unphased SNP genotypes

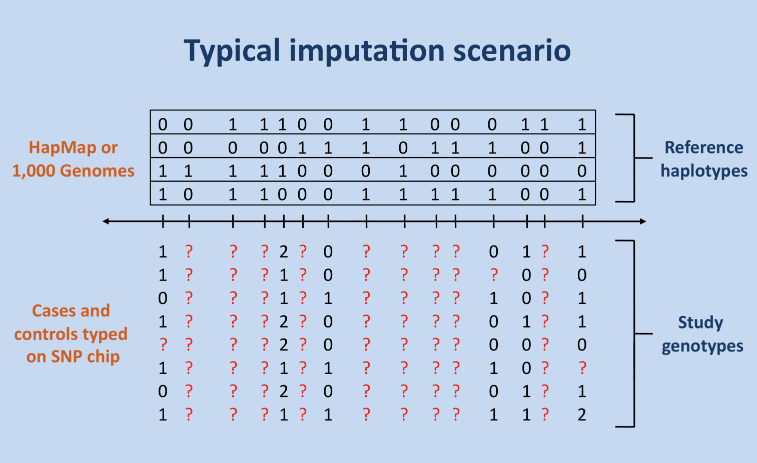
0 1 1 2

We would like to estimate the original haplotypes

0 1 1 1	0 1 0 1
0 0 0 1	0 0 1 1

How do we do that?

Haplotypes



Typical imputation scenario

HapMap or 1,000 Genomes

Reference haplotypes

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

Cases and controls typed on SNP chip

Study genotypes

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

Typical haplotypes can be estimated from, e.g., families.

Association of rare variants

Recall that GWAS is useful for common variants. However, rare variant might be just as important.

Example: variant with frequency of 0.001, disease with prevalence 10%. A case-control study with a two-fold increase effect in risk requires. 33k cases and controls! A 3-fold increase requires 11k cases and controls!

Association of rare variants

Recall that GWAS is useful for common variants. However, rare variant might be just as important.

Example: variant with frequency of 0.001, disease with prevalence 10%. A case-control study with a two-fold increase effect in risk requires. 33k cases and controls! A 3-fold increase requires 11k cases and controls!

- The burden test.
- The Sequence Kernel Association Test (SKAT) approach.

Association of rare variants

The burden test:

- Instead of testing rare variants individually, group variants likely to have similar function
- Score presence or absence of rare variants per individual. Use rare variant score to predict trait values
- If all variants are causal, leads to large increase in power
- In practice, success depends on:
  - Number of associated variants,
  - Number of neutral variants diluting signals

The burden test

- Instead of testing rare variants individually, group variants likely to have similar function
- Score presence or absence of rare variants per individual. Use rare variant score to predict trait values
- If all variants are causal, leads to large increase in power
- In practice, success depends on:
  - Number of associated variants,
  - Number of neutral variants diluting signals

The burden test

- Instead of testing rare variants individually, group variants likely to have similar function
- Score presence or absence of rare variants per individual. Use rare variant score to predict trait values
- If all variants are causal, leads to large increase in power
- In practice, success depends on:
  - Number of associated variants,
  - Number of neutral variants diluting signals

What do we do in practice to test the burden of rare variants:

- Count # rare variants within each gene
- Associate # variants with phenotype

Burden tests implicitly assume that all the rare variants in a region are causal and affect the phenotype in the same direction with similar magnitudes,

Using polygenic risk scores

What if we are not interested in genetics *per se* but just want to control for it?
Assumes a set $\mathcal{S}$ which are the SNPs of interest *and* their corresponding regression coefficients, $\hat{\beta}$s.

- Create a polygenic risk score

$$\mathrm{PRS}_i = \sum_{s \in \mathcal{S}} \hat{\beta}_s x_{is}$$

- Use the PRS as a predictor in a regression model

What are the problems with this approach? What are the advantages?

The sequence kernel association test

SKAT aggregates individual variant score test statistics with weights.

1. Consider a region, e.g., a gene, that has m variants.
2. The statistical model is

$$Y = \mathbf{X}\beta + \varepsilon$$

   Assume that the regression coefficients are random variables each with variance τw_j^2.
3. A test for no effect, $H_0 : \beta = 0$ corresponds to testing the variance scaling parameter $H_0 : \tau = 0$.
4. In practice the test statistic for H_0 becomes a weighted sum of the individual SNP score statistics.

The SKAT test allows for differing effect sizes and directions of the variants in the region.

Gene-set enrichment analysis

Idea: GSEA tests for enrichment of some pre-specified group S among N background genes.
How: the ensemble of genes in each gene set using a **metric** for each gene. Increases statistical power for “difference” between two outcomes.
If the set contains important genes then the genes in the set should cluster around the top of the list of important genes.

Any set of genes that makes sense to consider together.
The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways or coexpression in previous experiments.

Gene-set enrichment analysis

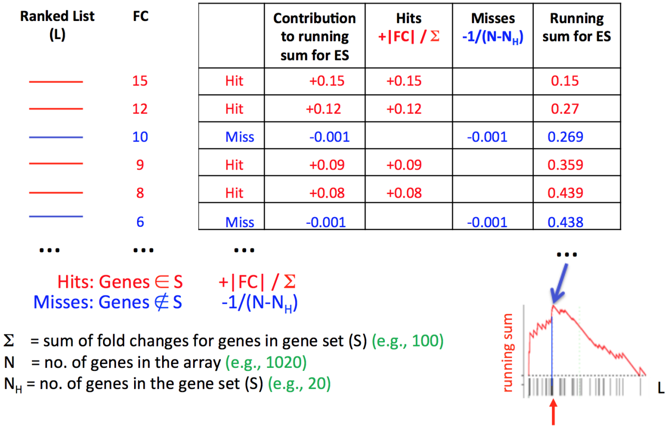
Start with ranked list (L) of genes that are in ($+$) or not in (-) a gene set (S) based on e.g., fold change, correlation, differences.

Evaluate the fraction of genes in S ($+$) weighted by their correlation and the fraction of genes not in S (-) present up to a given position i in L

Three key elements:

1. Calculate an enrichment score (ES) that reflects the degree to which a set S is overrepresented at the extremes
2. Estimate the statistical significance (the p-value) of the ES by using an empirical phenotype-based permutation test procedure.
3. Adjustment for Multiple Hypothesis Testing.

Gene-set enrichment analysis



Ranked List (L)	FC		Contribution to running sum for ES	Hits $+\lvert FC\rvert / \Sigma$	Misses $-1/(N-N_H)$	Running sum for ES
—	15	Hit	+0.15	+0.15		0.15
—	12	Hit	+0.12	+0.12		0.27
—	10	Miss	-0.001		-0.001	0.269
—	9	Hit	+0.09	+0.09		0.359
—	8	Hit	+0.08	+0.08		0.439
—	6	Miss	-0.001		-0.001	0.438
…	…	…				…

Hits: Genes $\in$ S $+\lvert FC\rvert / \Sigma$

Misses: Genes $\notin$ S $-1/(N-N_H)$

Σ = sum of fold changes for genes in gene set (S) (e.g., 100)

N = no. of genes in the array (e.g., 1020)

N_H = no. of genes in the gene set (S) (e.g., 20)

Enrichment score = value of maximum deviation from 0 of the running sum.

Gene-set enrichment analysis — testing

Uses gene set permutation.

- For gene set S, each permutation π is the random selection of s genes from all genes in the genome which have a probe on the platform.
- If $ES(S) > 0$, the resulting empirical p value for S is the fraction of the $ES_{\pi}(S)$ values that equal or exceed the actual enrichment score $ES(S)$.