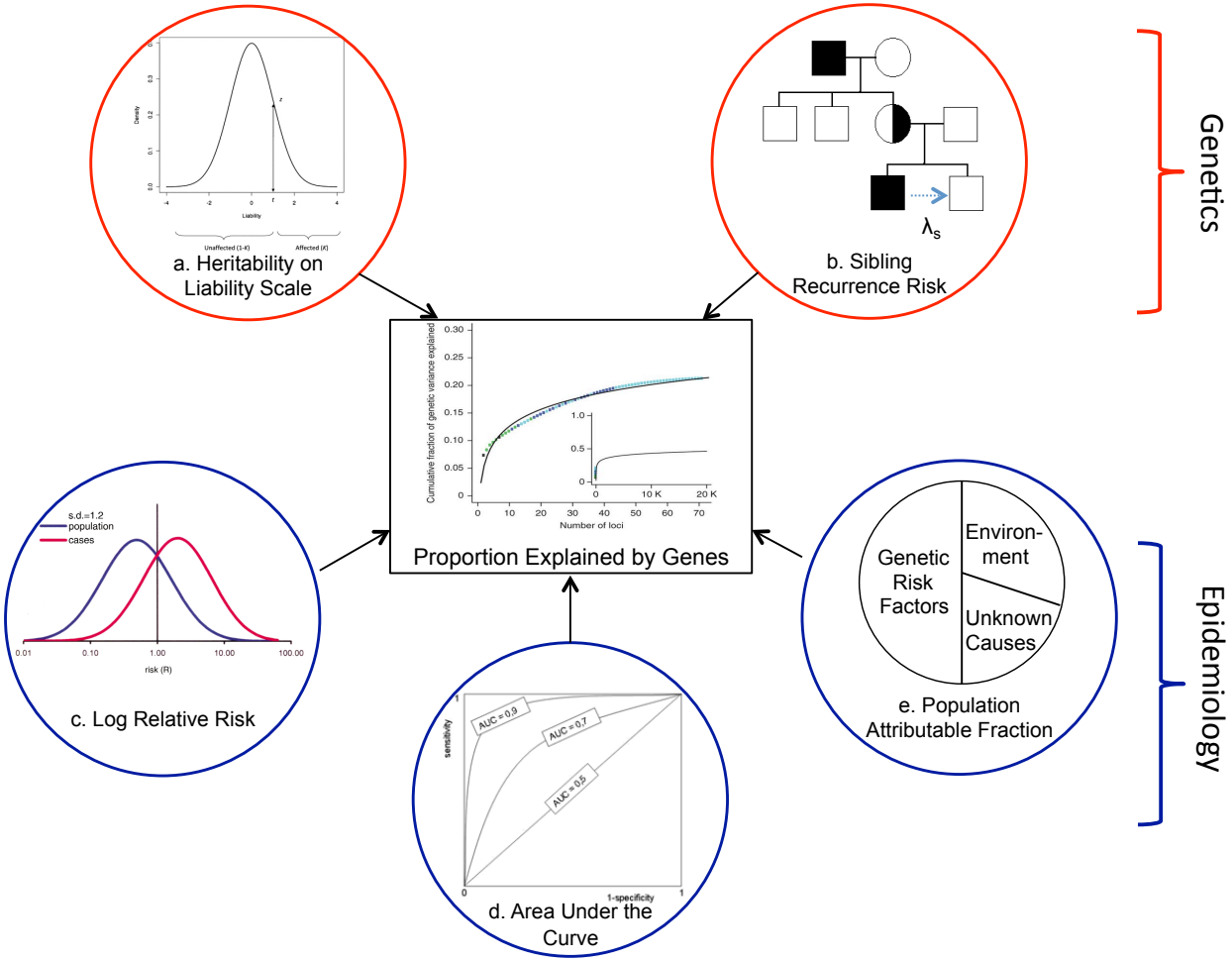
2016 Module 18:
Statistical & Quantitative Genetics of Disease

Converging fields of genetics, epidemiology & genetic epidemiology



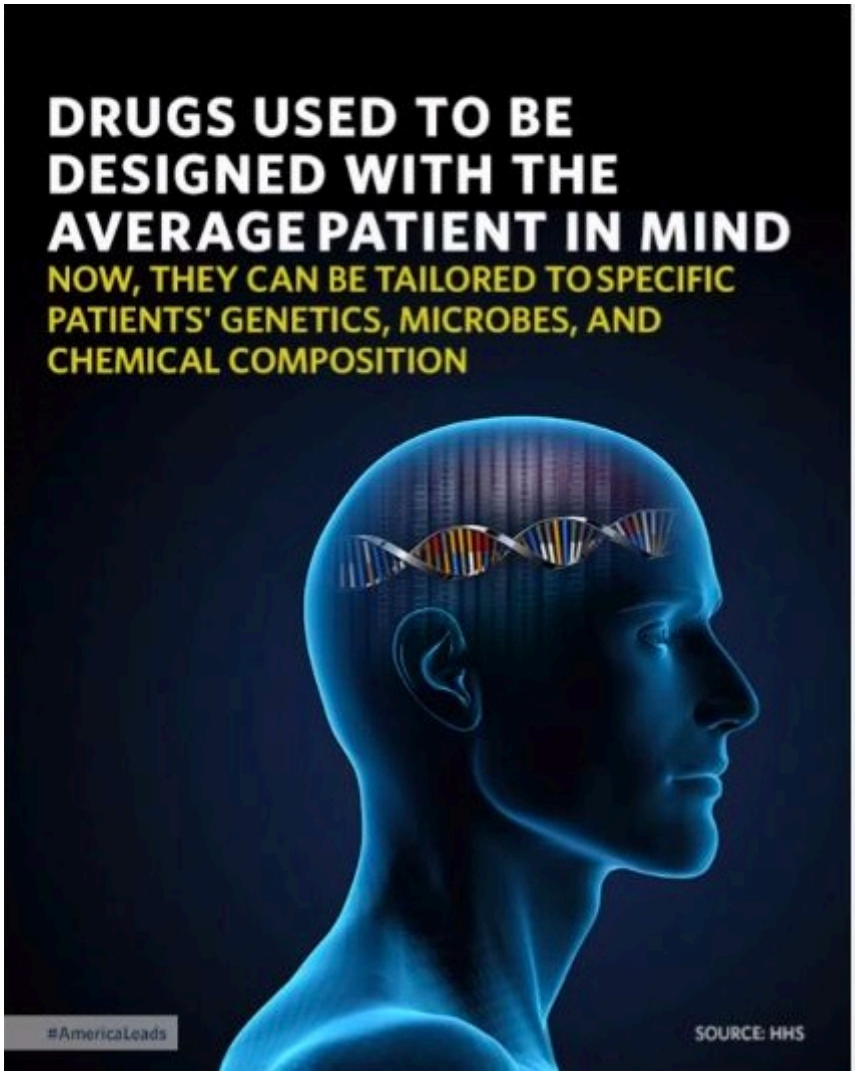
Naomi Wray
Psychiatric genetics

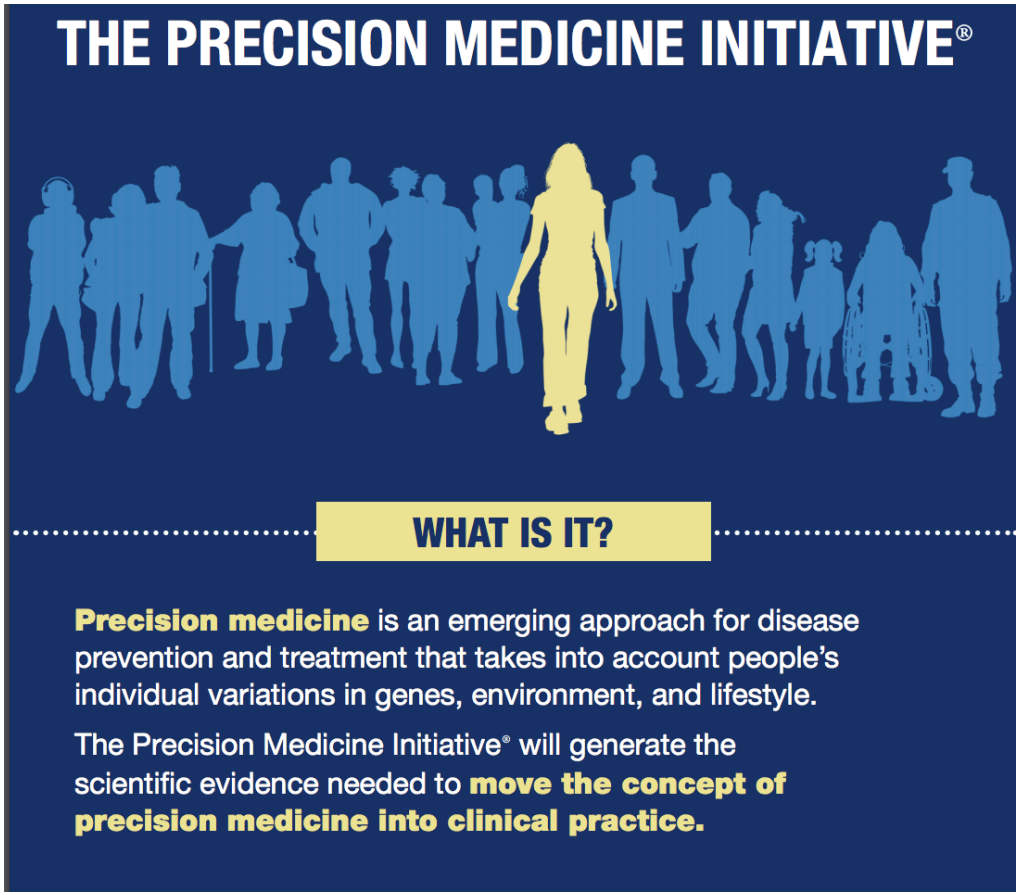
John Witte
Cancer genetics

Motivation for this module

- To unite the language of quantitative genetics (QG) and epidemiology
- Quantitative genetics of disease is often a tack on to QG of quantitative traits –here we make it the focus
- The new era of genomics bring QG of genetics of disease back into the foreground – a renewed relevance
- Understanding of prediction of disease risk in the precision medicine era

Precision Medicine Initiatives





http://syndication.nih.gov/multimedia/pmi/infographics/pmi-infographic.pdf

LONGER-TERM GOALS

Create a research cohort of **> 1 million American volunteers** who will share genetic data, biological samples, and diet/lifestyle information, all linked to their electronic health records if they choose.

Pioneer **a new model for doing science** that emphasizes **engaged participants, responsible data sharing, and privacy protection.**

Research based upon the cohort data will:

- Advance **pharmacogenomics,** the right drug for the right patient at the right dose
- Identify new targets for **treatment and prevention**
- Test whether **mobile devices** can encourage healthy behaviors
- Lay **scientific foundation** for precision medicine for **many diseases**

Course Outline

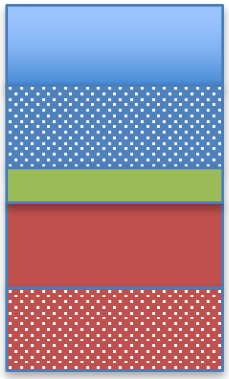
Wednesday afternoon
- Lecture 1 Genetic epidemiology of disease; Heritability of liability (Naomi)
- Lecture 2 Single locus disease analysis: design, logistic regression, covariates (John)

Thursday morning(John)
- Lecture 3: Single locus disease model; Power calculation for disease model (Naomi)
- Lecture 4: Interpreting measures of variation; multivariate models (John)

Thursday afternoon(Naomi)
- Lecture 5:Multi-locus disease model (Naomi)
- Lecture 6: Modeling interactions: gene-environment, epistasis (John)

Friday morning(John)
- Lecture 7: Risk Prediction – measures of accuracy for risk prediction of disease (Naomi)
- Lecture 8: Pleiotropy; LDscore (John)

Friday afternoon
- Lecture 9: Risk prediction – theory (Naomi)
- Lecture 10: Rare variants; Risk Prediction – application (John)



Naomi lecture — More quantitative genetics theory

practical

Coffee

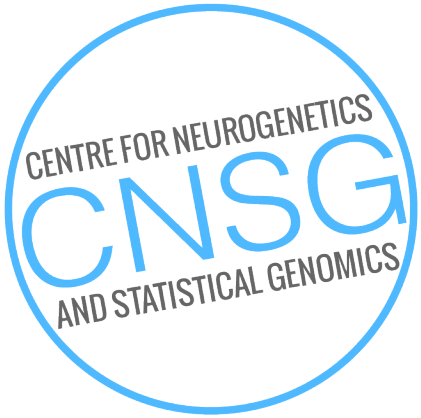
John lecture — More statistics/data analysis

practical

2016 Module 18:
Statistical & Quantitative Genetics of Disease

Lecture 1
Quantifying the genetic contribution to disease

Naomi Wray



Aims of Lecture 1

If a disease affects 1% of the population and has heritability 80%

We will show why these statements are consistent :

If an individual is affected ~8% of his/her siblings affected

If an MZ twin is affected ~50% of their co-twins are affected

If an individual is affected > 60% will have no known family history

Bringing together genetic epidemiology and quantitative genetics

- The key papers were published 40 and 70 years ago……

Structure

Disease data and risk to relatives

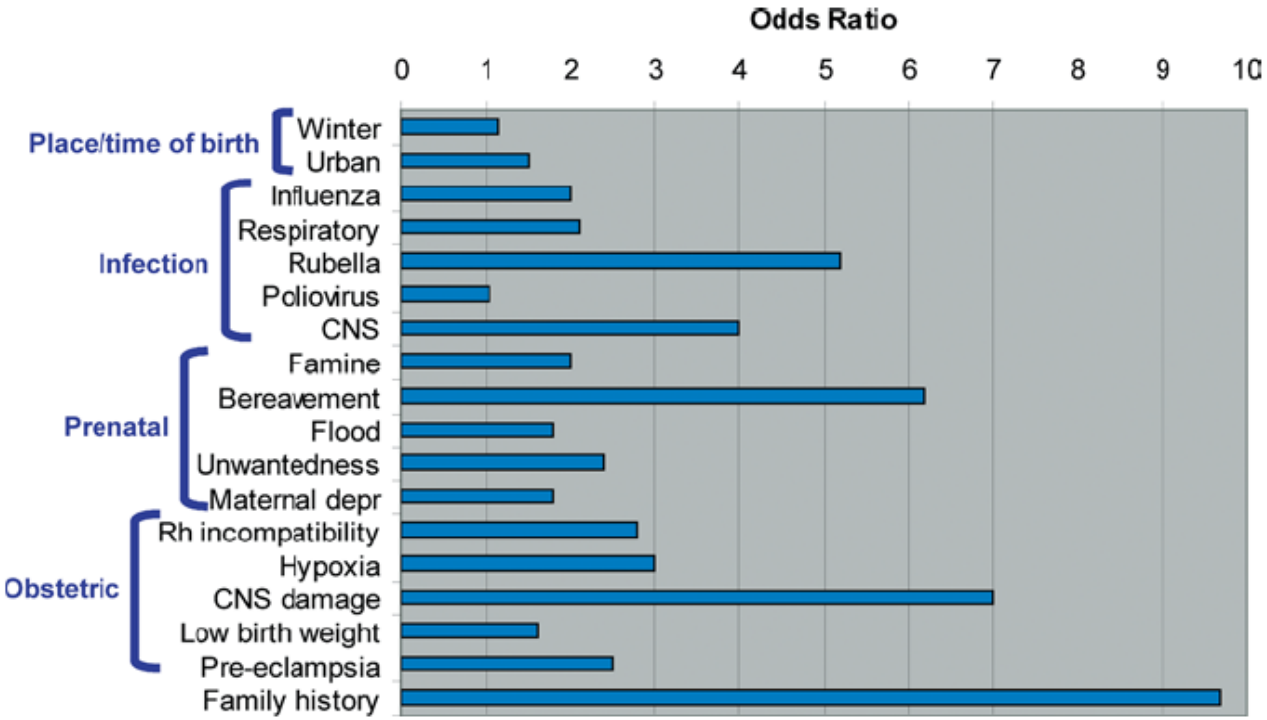
Aside: Quantitative Traits

Liability Threshold Model

Practical

Disease data and risk to relatives

Risk Factors for Schizophrenia



DOI: 10.1371/journal.pmed.0020212.g001

Figure 1. Comparison of a Selected Set of Relatively Well-Established Risk Factors for Schizophrenia, Focusing Mainly on Pre- and Antenatal Factors [6] (abbreviations: CNS, central nervous system; depr, depression; Rh, Rhesus)

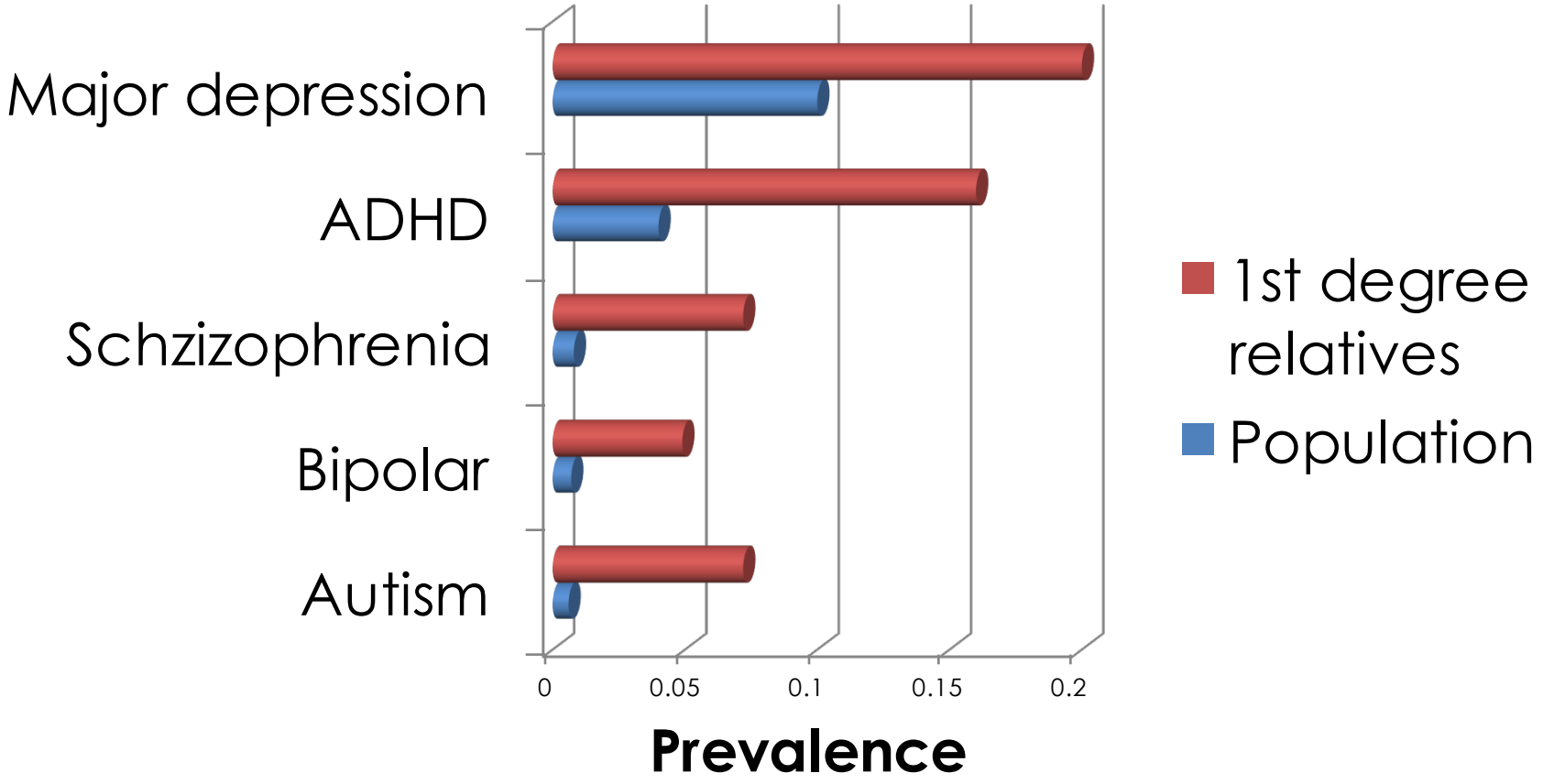
Sullivan, PLoS Med 05

Complex genetic diseases

- Unlike Mendelian disorders, there is no clear pattern of inheritance
- Tend to "run" in families
- Few large pedigrees of multiply affected individuals
- Most people have no known family history

What can we learn from genetic epidemiology about genetic architecture?

Evidence for a genetic contribution comes from risks to relatives



Relative risk to relatives
Recurrence risk to relatives

How much more likely are you to be diseased if your relative is affected compared to a person selected randomly from the population?

$$\text{Relative risk to relatives } (\lambda_R) = \frac{p(\text{affected} \mid \text{relative affected})}{p(\text{affected in population})} = \frac{K_R}{K}$$

How to estimate p(affected | relative affected) ?

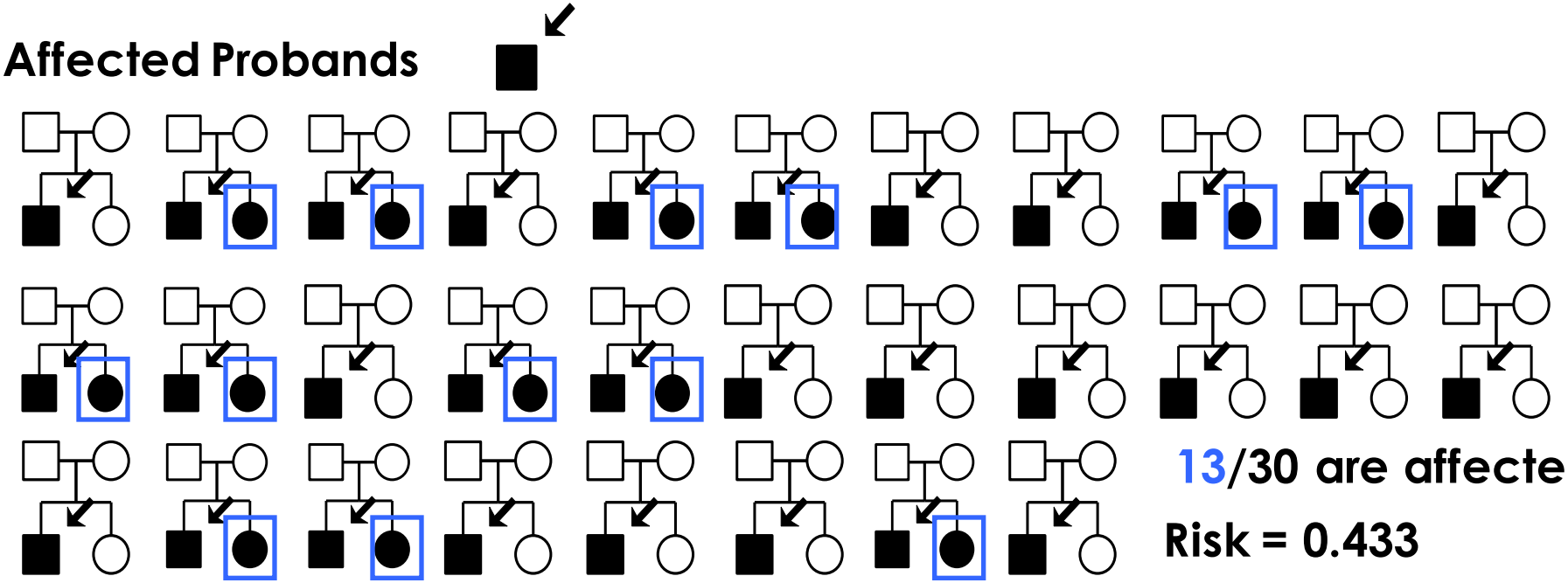
- Collect population samples – cases infrequent
- Collect samples of case families and assess family members

How to estimate p(affected in population) ?

- Census or national health statistics
  - Is definition of affected same in population sample as family sample
- Collect control families and assess family members

$$\text{If disease is not common} \quad \lambda_R = \frac{p(\text{sibling affected} \mid \text{case family})}{p(\text{sibling affected} \mid \text{control family})}$$

Affected Probands



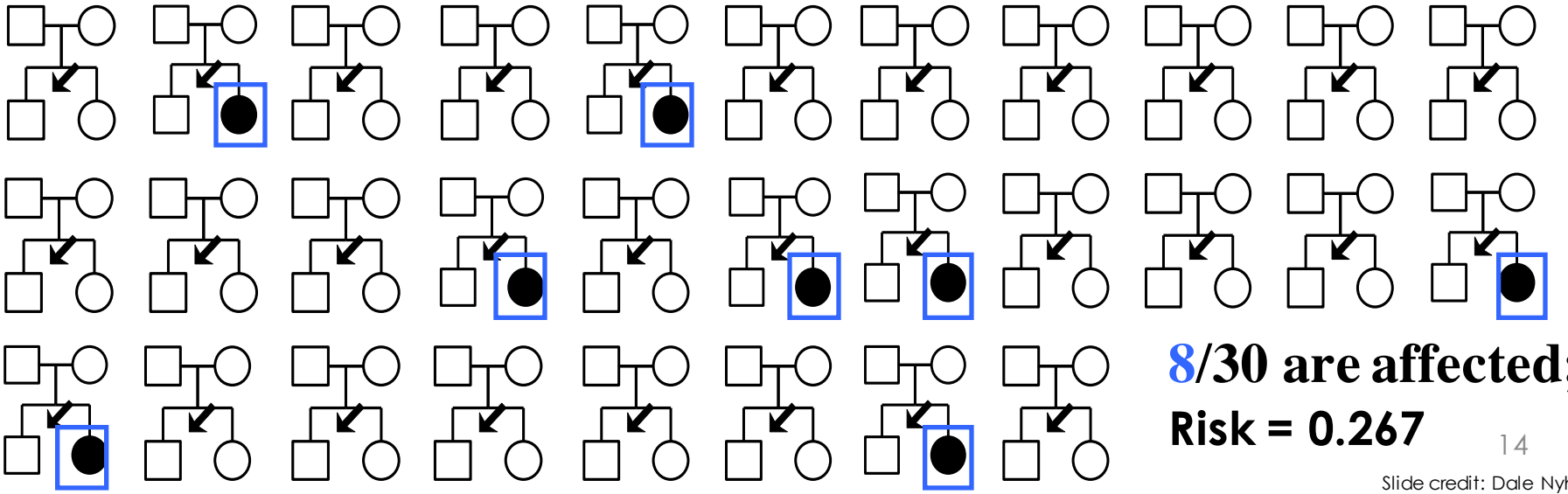
13/30 are affected;

Risk = 0.433

Unaffected Probands

Relative Risk (RR) = 0.433 / 0.267 = 1.63

In siblings of affected compared to unaffected probands



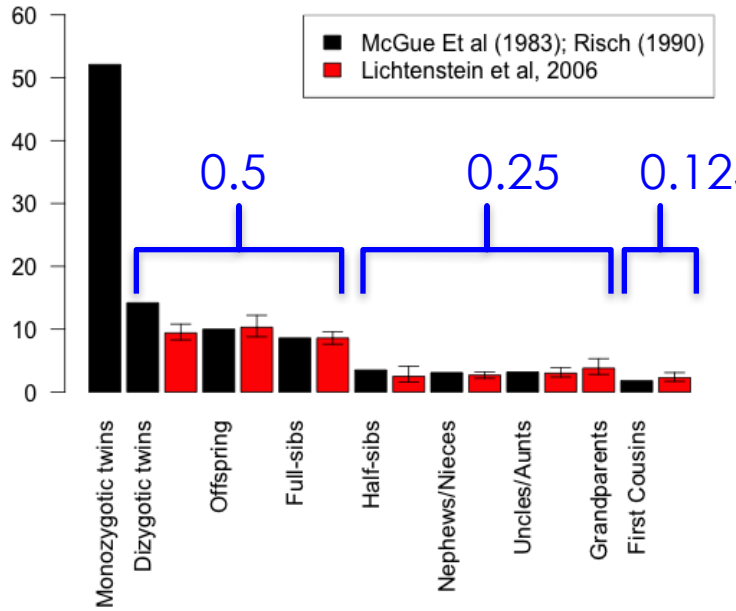
8/30 are affected;

Risk = 0.267

Slide credit: Dale Nyholt

Schizophrenia risks to relatives

Relatives	Coefficient of relationship	Risch McGue et al	Lichtenstein et al Estimate	95% CI
Monozygotic twins	1	**52.1**		
Dizygotic twins	½	**14.2**		
Parent	½		**9.4**	**8.3 - 10.8**
Offspring	½	**10.0**	**10.3**	**8.8 - 12.2**
Full-sibs	½	**8.6**	**8.6**	**7.6 - 9.6**
Half-sibs	¼	**3.5**	**2.5**	**1.6 - 4.1**
Nephews/Nieces	¼	**3.1**	**2.7**	**2.2 - 3.2**
Uncles/Aunts	¼	**3.2**	**3.0**	**2.4 - 3.9**
Grandparents	¼		**3.8**	**2.8 - 5.3**
First Cousins	1/8	**1.8**	**2.3**	**1.7 – 3.1**
Offspring of 2 affected parents	½ but ascertained		**89**	**19 - 672**



Baseline risk, K = 0.85% McGue et al
= 0.407% Lichtenstein et al

coefficient of relationship

Risch(1990) Linkage Strategies for Genetically Complex Traits AJHG
McGue et al (1983) Genetic Epidemiology 2: 99
Lichtenstein et al (2006) Recurrence risks for schizophrenia in a Swedish National Cohort.Psychological Medicine

James (1971) relationship between K and K_R

X = scores of disease yes/no for individuals

Y = scores of disease yes/no in relatives of X

K proportion of the population affected

$E(X) = E(Y) = K$

$K_R = E(Y \mid X=1)$

Probability that both X and Y =1: $E(XY) = K*K_R$

$Cov(X,Y) = E(XY) - E(X)*E(Y) = K*K_R - K^2$

So $Cov_R = Cov(X,Y) = K*K_R - K^2 = (K_R - K)K = (\lambda_R - 1)K^2$

James (1971) Frequency in relatives for an all-or-non trait Ann Hum Genet 35 47

Derivation from Risch (1990) Linkage strategies for genetically complex traits. I Multi-locus models. AJHG

Aside 1: Heritability

$$P = G + \varepsilon$$

P = phenotype
G = genetic factors
ε = residual, anything other than genetic, including environmental and stochastic factors

Broad sense heritability $$H^2 = \frac{\sigma_G^2}{\sigma_p^2}$$

- Parameters vs Estimates
- Often confused and confusing
- We can measure P but we cannot directly measure G or A.
- Estimate variance of G or A by using cohorts of individuals for whom we know the coefficient of relationship, but difficulties arise because of confounding

$$P = A + \varepsilon$$

P = phenotype
A = additive genetic factors
ε = residual, anything other than additive genetic, including environmental and stochastic factors

Narrow sense heritability $$h^2 = \frac{\sigma_A^2}{\sigma_p^2}$$

Aside 2: Covariances between relatives

$P = A + \varepsilon$

$V(P) = V(A) + V(\varepsilon)$, A and E uncorrelated

$P_{child} = A_{child} + \varepsilon = \frac{1}{2} A_{mum} + \frac{1}{2} A_{dad} + A_{seg} + \varepsilon$

$V(A_{child}) = \frac{1}{4} V(A_{mum}) + \frac{1}{4} V(A_{dad}) + V(A_{seg})$

$V(A) = \frac{1}{4} V(A) + \frac{1}{4} V(A) + V(A_{seg})$ so $V(A_{seg}) = \frac{1}{2} V(A)$

$Cov(P_{child},P_{dad}) = Cov(A_{child},A_{dad}) = Cov(\frac{1}{2} A_{mum} + \frac{1}{2} A_{dad} + A_{seg}, A_{dad}) = \frac{1}{2} V(A)$

$Cov(P_{child},P_{sib}) = Cov(A_{child},A_{sib}) =$
$Cov(\frac{1}{2} A_{mum} + \frac{1}{2} A_{dad} + A_{seg\text{-}ch}, \frac{1}{2} A_{mum} + \frac{1}{2} A_{dad} + A_{seg\text{-}sib})$
$= \frac{1}{4} V(A) + \frac{1}{4} V(A) = \frac{1}{2} V(A)$

General covariance between relatives

cov_R = covariance between relatives on the disease scale

$$cov_R = a_R V_{Ao} + u_R V_{Do} + a_R^2 V_{AAo} + a_R u_R V_{ADo} + \cdots$$

	V_A	V_D	V_{AA}	V_{AD}	V_{DD}
Offspring-parent	½	0	¼	0	0
Half-sib	¼	0	$^1/_{16}$	0	0
Full-sib	½	¼	¼	$^1/_8$	$^1/_{16}$
MZ twin	1	1	1	1	1
General	a_R	u_R	a_R^2	$a_R\, u_R$	u_R^2

$cov_R = (K_R - K)K = (\lambda_R - 1)K^2$ $\quad V_P = K(1-K)$ (from a few slides back!)

An estimate of narrow sense (additive) heritability on the disease scale is

$$\widehat{h_o^2} = \frac{(\lambda_R - 1)K^2}{a_R K(1-K)} = \frac{(\lambda_R - 1)K}{a_R(1-K)}$$

But covR contains non-additive genetic terms.
We don't know if non-additive genetic effects exist - What to do?

Estimate $\widehat{h_o^2}$ from different types of relatives to see if the estimates are consistent

James (1971) Frequency in relatives for an all-or-non trait Ann Hum Genet 35 47

James (1971) genetic variance on the disease scale

$$\widehat{h_o^2} = \frac{(\lambda_R - 1)K^2}{a_R K(1-K)} = \frac{(\lambda_R - 1)K}{a_R(1-K)}$$

K = 0.0085

λ_{OP}= 10 a_R= ½ $\quad \widehat{h_o^2} = \dfrac{(10-1)0.0085}{\frac{1}{2}(1-0.0085)} = 0.154$

λ_{HS} = 3 a_R= ¼ $\quad \widehat{h_o^2} = 0.069$

λ_{FS} = 8.6 a_R= ½ $\quad \widehat{h_o^2} = 0.130$

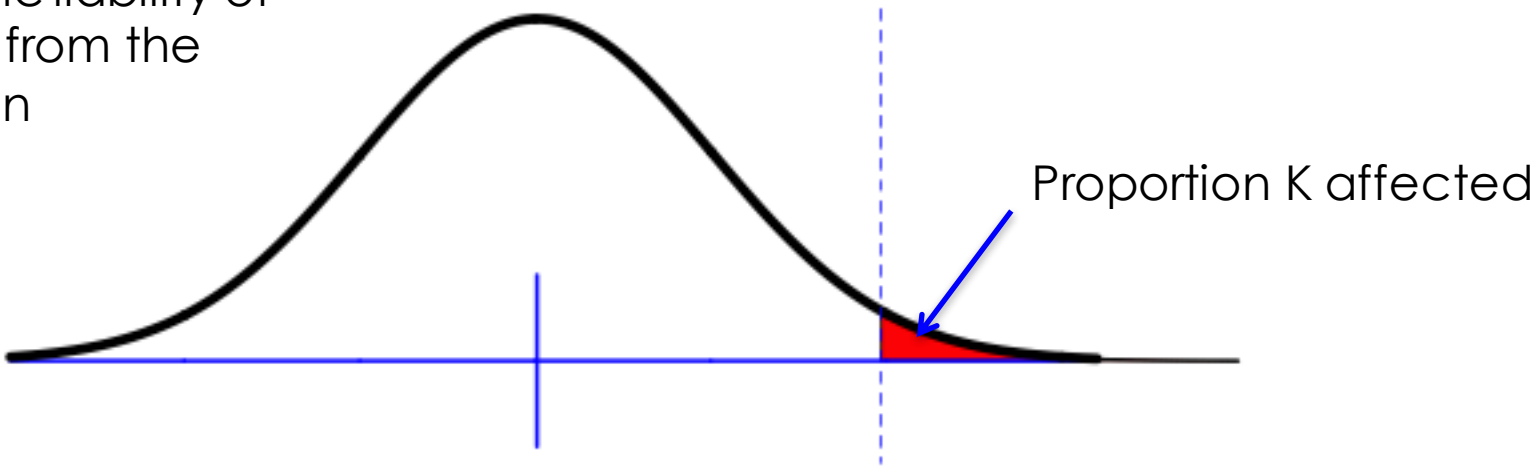
λ_{MZ}= 52 a_R= 1 $\quad \widehat{h_o^2} = 0.438$

The estimates of $\widehat{h_o^2}$ are very different (even if sampling variance is taken into account)

Implies that the estimates of $\widehat{h_o^2}$ are contaminated by non-additive variance on this scale of measurement

Liability threshold model

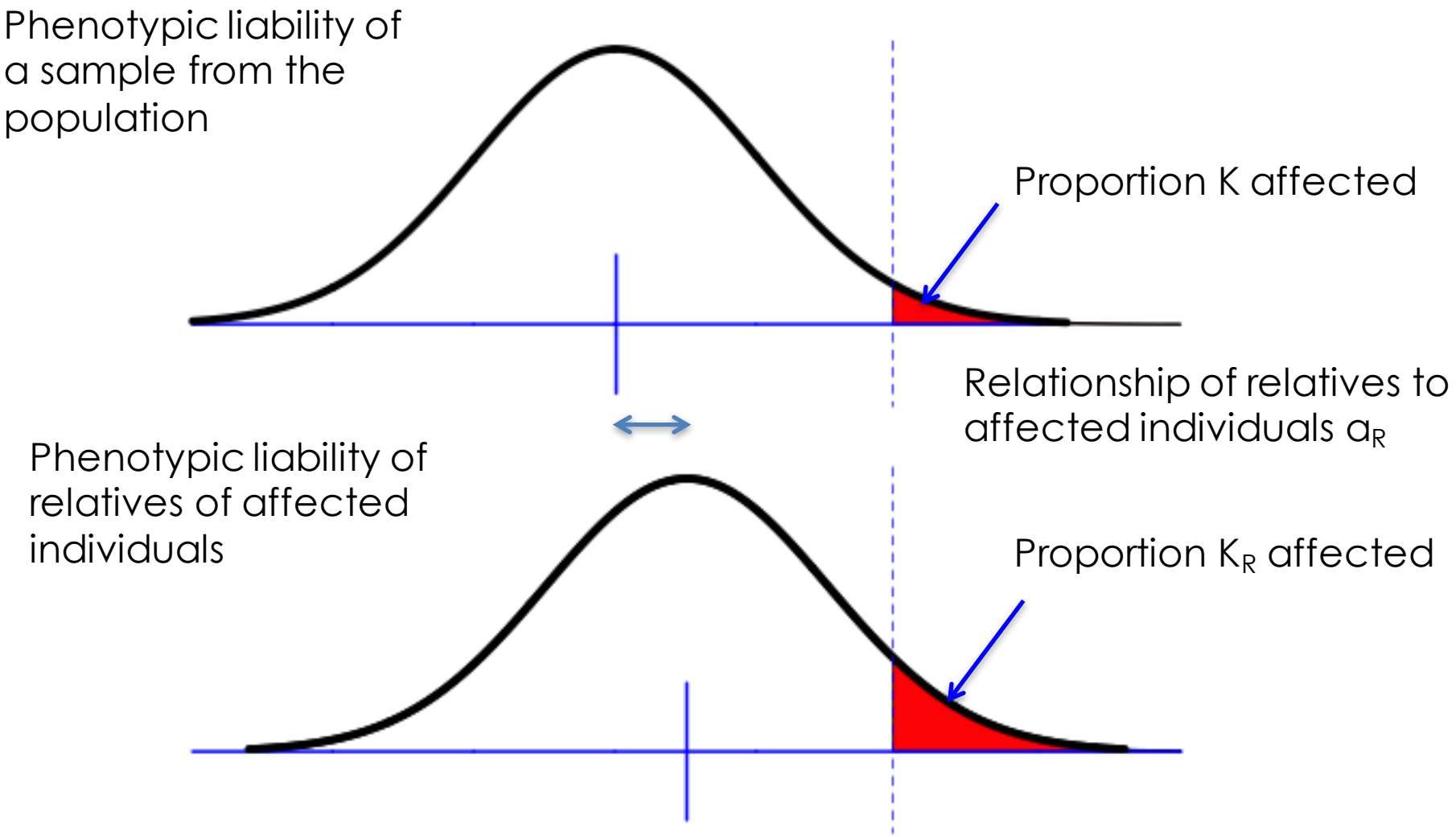
Phenotypic liability of a sample from the population



Assumption of normality

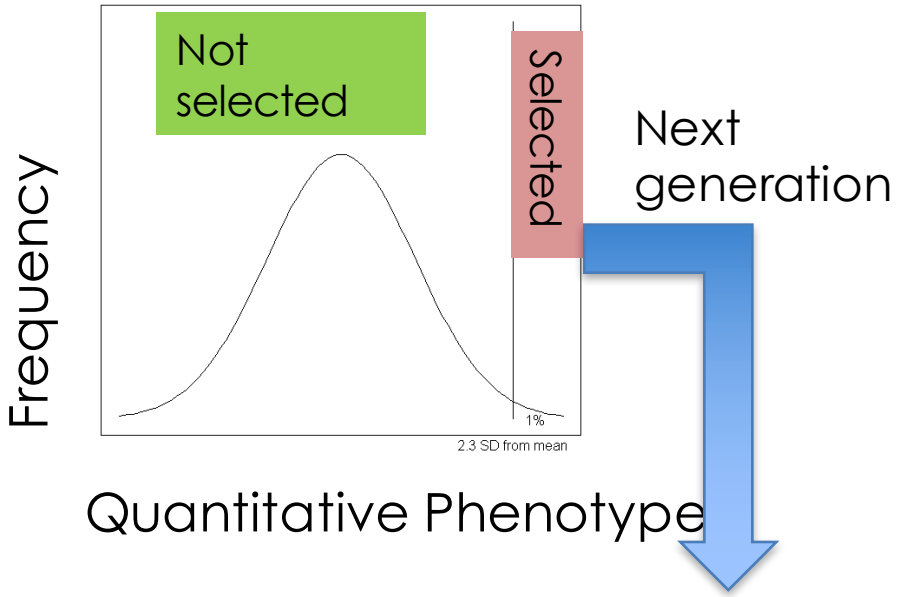
- Only appropriate for multifactorial disease
- i.e. more than a few genes but doesn't have to be highly polygenic
- Key – unimodal

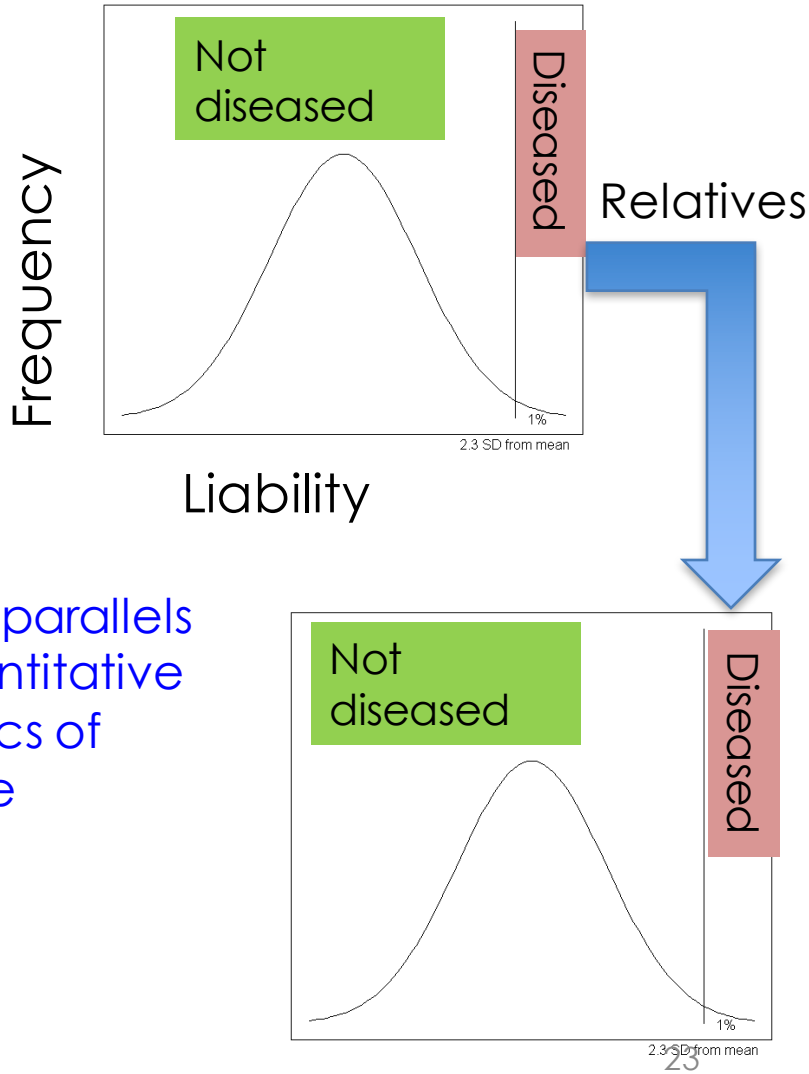
Falconer (1965)



Using normal distribution theory what percentage of the variance in liability is attributale to genetic factors given K, K_R and a_R

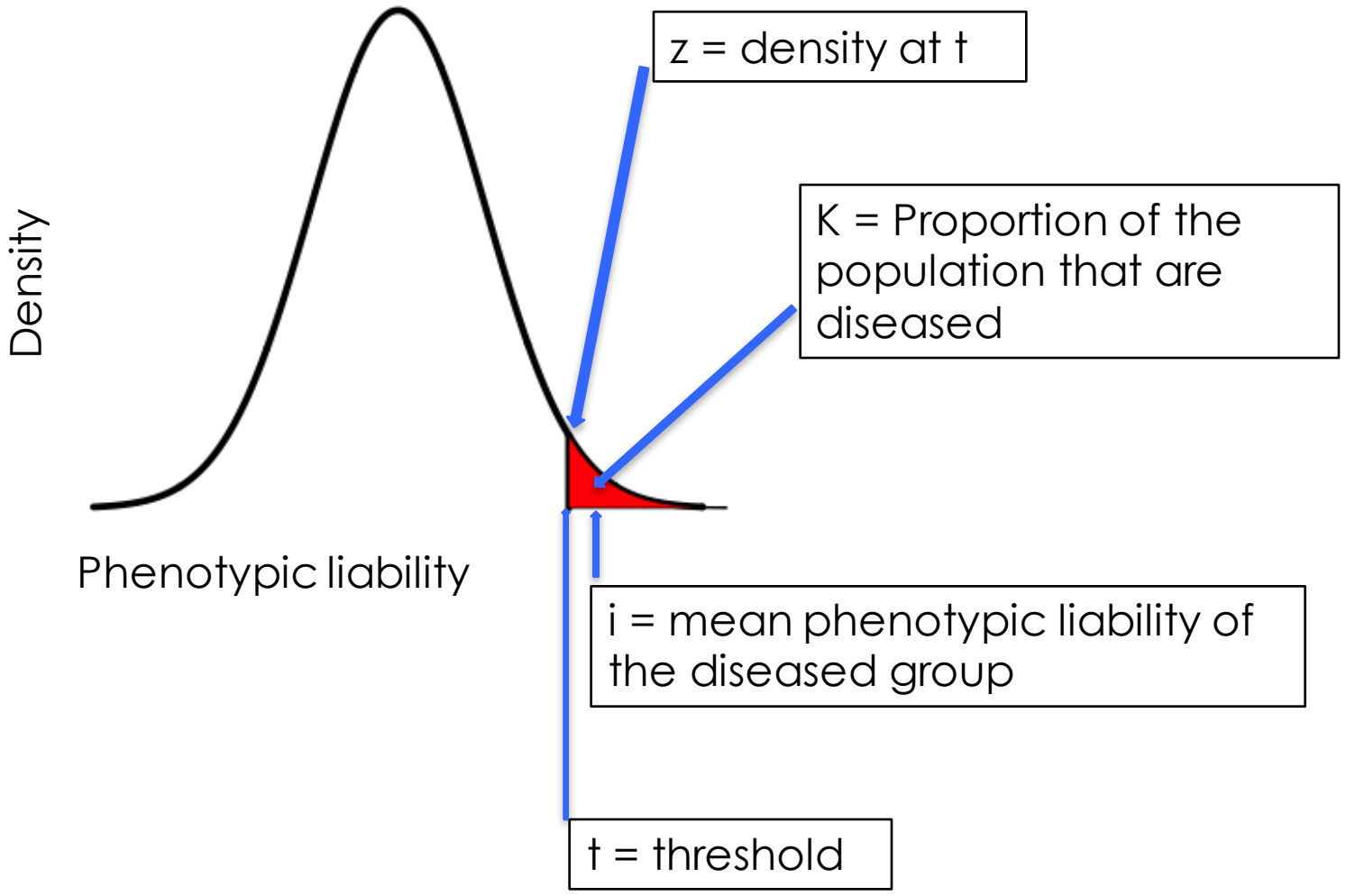
Prediction of response to selection and rates of inbreeding under directional selection





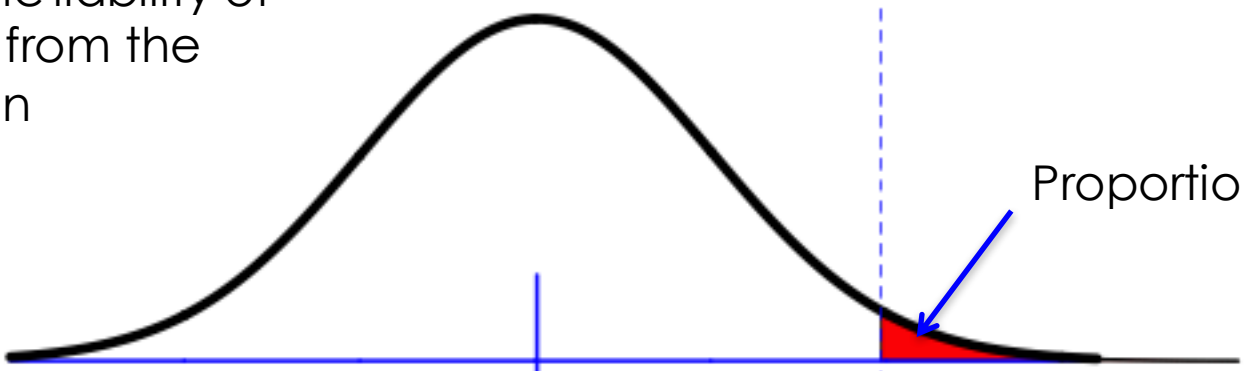
Strong parallels to quantitative genetics of disease

Definitions



How to get from observed risks to relatives to heritability?- Falconer (1965)

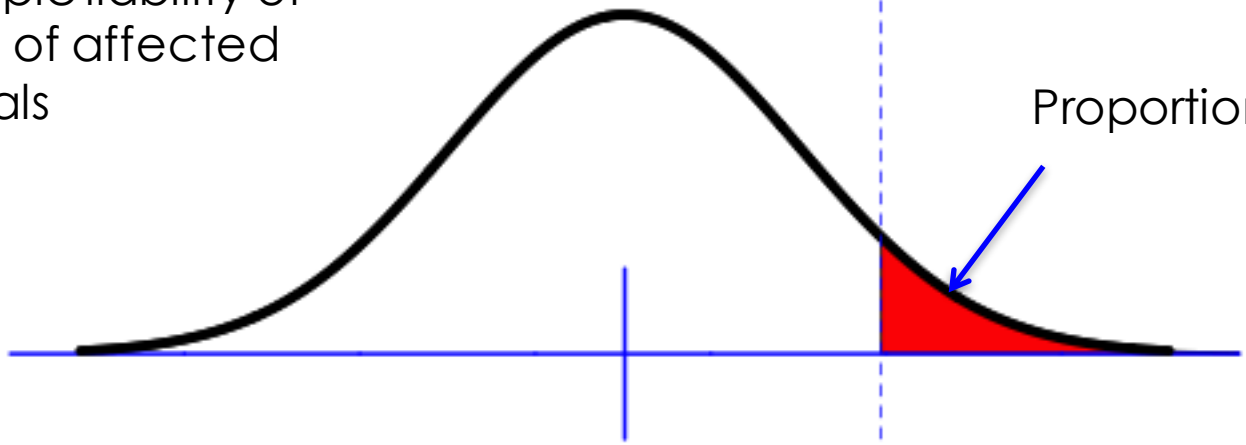
Phenotypic liability of a sample from the population



Proportion K affected

Relationship of relatives to affected individuals r

Phenotypic liability of relatives of affected individuals



Proportion K_R affected

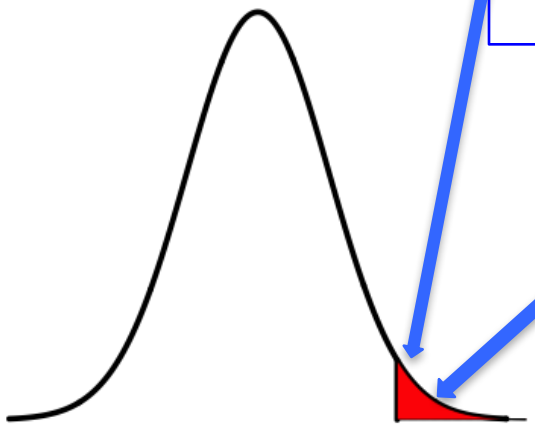
Using normal distribution theory what percentage of the variance in liability is attributale to genetic factors given K, K_R and r

Liability Threshold Model –truncated normal distribution theory

Φ(x) =cumulative density until liability x
standard normal distribution function
φ (x) = probability density at x
Phi

z = density at t

z = φ (t) $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ = dnorm(t)



K = Proportion of the population that are diseased

K= 1-Φ(t) = 1-pnorm(t)

Standard Deviation =1
$\sigma_p = 1$

i = mean phenotypic liability of the diseased group

i= z/K "selection intensity"

Variance in liability amongst the diseased individuals
= σ_p^2 (1-k), where k = i(i-t)

t = threshold

t= $\Phi^{-1}(1-K)$ = qnorm(1-K)

Inverse standard normal distribution (probit) function

Mean of diseased group

- Pearson & Lee (1908) On the generalized probable error in normal correlation. Biometrika
- Lee (1915) Table of Gaussian tail functions..Biometrika
- Fisher (1941) Properties and application of Hh functions. Introduction to mathematical tables
- Cohen (1949) On estimating the mean and standard deviation of truncated normal distributions Am Stat Association
- Cohen & Woodward (1953)Pearson-Lee-Fisher Functions of singly truncated normal distributions. Biometrics

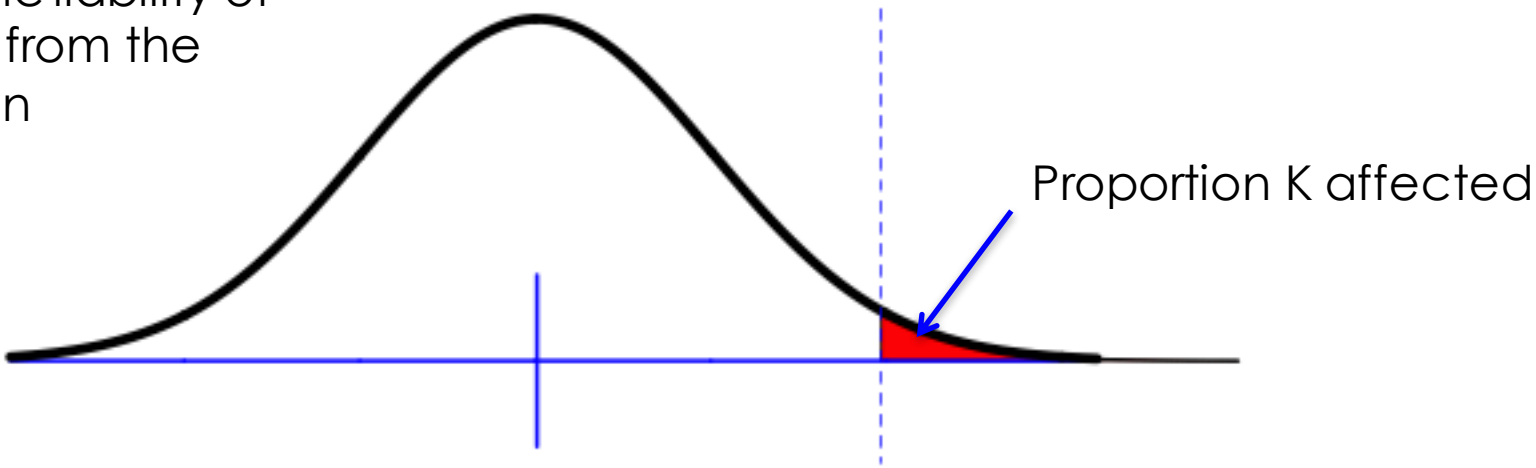
Mean (i): = sum(x * freq of x)
The phenotype frequencies must sum to 1, hence the denominator

$$i = \frac{\int_t^\infty x\phi(x)dx}{\int_t^\infty \phi(x)dx} = \frac{\int_t^\infty x\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx}{K} = \frac{\phi(t)}{K} = \frac{z}{K}$$

Lynch and Walsh equations 2.13 and 2.14; variance equation 2.15

Falconer (1965)

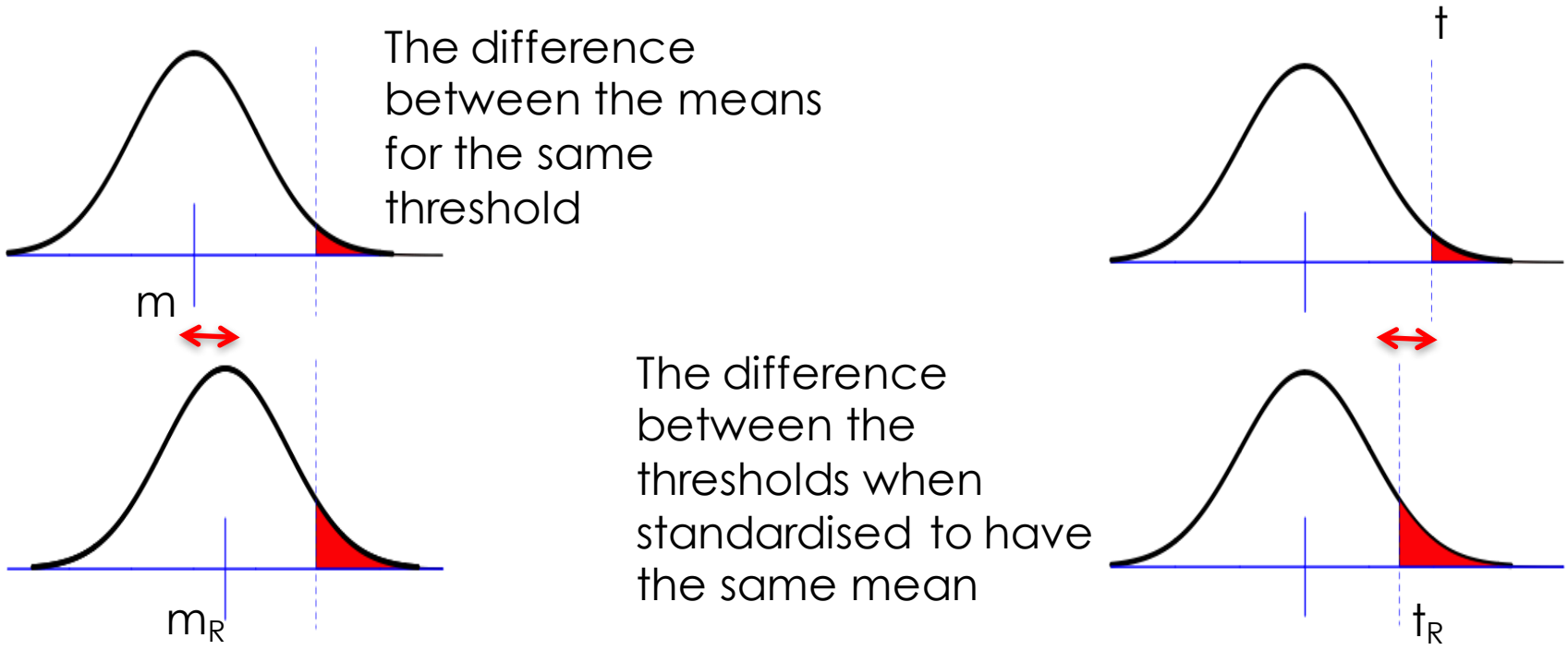
Phenotypic liability of a sample from the population



Assumption of normality

- Only appropriate for multifactorial disease
- i.e. more than a few genes but doesn't have to be highly polygenic
- Key – unimodal

Falconer (1965)



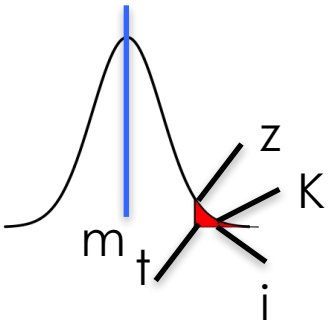
$$m_R - m = t - t_R$$

Given the difference in thresholds, and given known additive genetic relationship between relatives, what proportion of the total variance must be due to genetic factors

Falconer (1965) The inheritance of liability to certain diseases, estimated from incidences in relatives, Ann. Hum Genet. 29 51

Crittenden (1961) an interpretation o familial aggregation based on multiple genetic and environmental factors 29 Ann NY Acad Sci 91 769

Calculate heritability of liability using regression theory



X = phenotypic liability for individuals

Y = phenotypic liability for relatives of X

E(X) = E(Y) = m = 0

Relationship between X and Y is linear

$Y = \mu_Y + b_{Y.X}(X-\mu_x)+ \varepsilon$

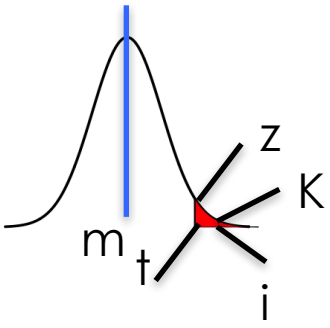
$= m + \frac{cov(A_R,A)}{Var(X)}(X-m) + \varepsilon$, since m = 0

$= \frac{a_R\sigma_a^2}{\sigma_p^2} X + \varepsilon = a_R h^2 X + \varepsilon$

Falconer (1965) The inheritance of liability to certain diseases, estimated from incidences in relatives, Ann. Hum Genet. 29 51

Crittenden (1961) an interpretation o familial aggregation based on multiple genetic and environmental factors Ann NY Acad Sci 91 769

Calculate heritability of liability using regression theory



X = phenotypic liability for individuals

Y = phenotypic liability for relatives of X

$Y = a_R h^2 X + \varepsilon$

For affected individuals X = i

Expected phenotypic liability of relatives of those affected

$E(Y \mid X>t) = m_R - m = t - t_R$

Substitute $t - t_R = a_R h^2 i$

Rearrange $h^2 = (t - t_R)/i a_R$

Falconer (1965) The inheritance of liability to certain diseases, estimated from incidences in relatives, Ann. Hum Genet. 29 51

Crittenden (1961) an interpretation o familial aggregation based on multiple genetic and environmental factors Ann NY Acad Sci 91 769

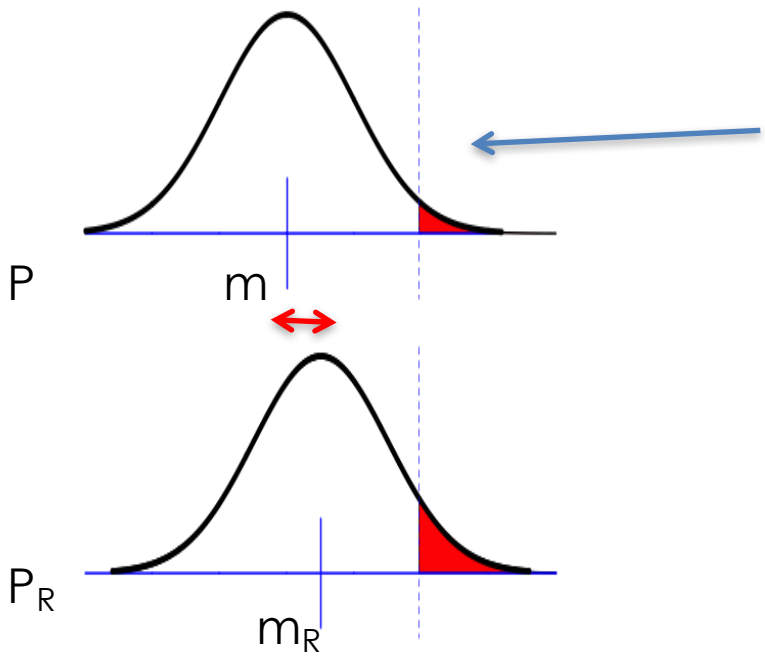
Assumptions made by Falconer (1965)

Assumption: Covariance between relatives reflects only shared additive genetic effects

Check: Use different types of relatives with different a_R and different u_R(dominance coefficient) and different shared environment to see consistency of estimates of h^2

Assumption: Phenotypic variance in relatives is unaffected by ascertainment on affected probands

Accounting for reduction in variance in relatives as a result of ascertainment on affected individuals



Variance in liability amongst the diseased individuals
= σ_p^2 (1-k), where k = i(i-t)

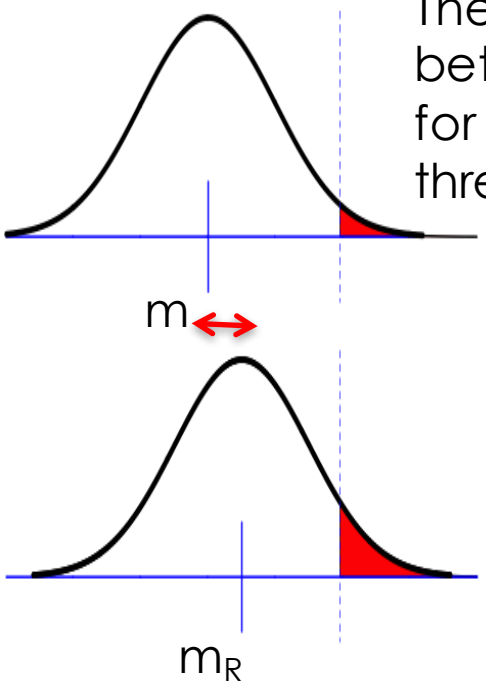
Variance in liability amongst relatives the diseased individuals

$V(P_R \mid P>t) = V(P_R) - kCov(P_R,P)^2$

$= 1 - k(a_R h^2)^2 = 1 - ka_R^2 h^4$

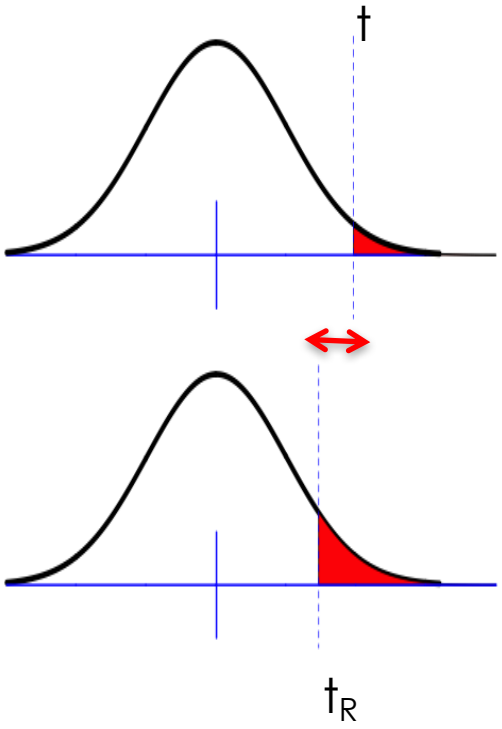
Reich, James, Morris (1972) The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. Ann Hum Gen 36: 163.

Reich et al: heritability of liability



The difference between the means for the same threshold

The difference between the thresholds when standardised to have the mean 0 and variance 1



$$m_R - m = t - t_R \sqrt{1 - ka_R^2 h^4}$$

Reich, James, Morris (1972) The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. Ann Hum Gen 36: 163.

Reich et al: heritability of liability

X = phenotypic liability for individuals

Y = phenotypic liability for relatives of X

t

NB. Distribution of relatives may also be skewed – especially for MZ twins-Estimates could be biased upwards

$Y = a_R h^2 X + \varepsilon$

For affected individuals X = i

Expected phenotypic liability of relatives of those affected

$E(Y \mid X>t) = m_R\text{-}m = t - t_R\sqrt{1 - ka_R^2h^4}$

Substitute $t - t_R\sqrt{1 - ka_R^2h^4} = a_R h^2 i$

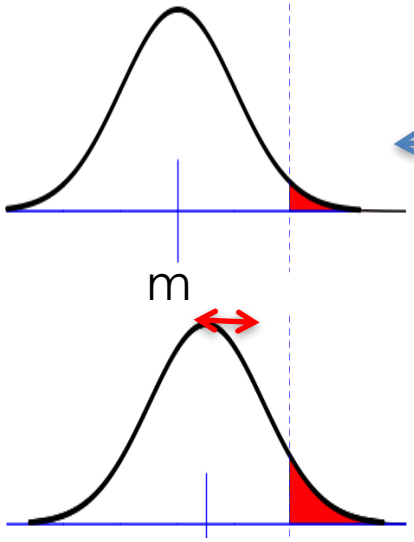
Rearrange

$$h^2 = \frac{t - t_R\sqrt{1 - (1 - t/i)(t^2 - t_R^2)}}{a_R(i + (i - t)t_R^2)}$$

Also useful – calculation of t_R when K and h^2 are known

$$t_R = \frac{t - a_R i h^2}{\sqrt{1 - a_R^2 h^4 k}}$$

Accounting for reduction in variance in relatives as a result of ascertainment on affected individuals



Variance in liability amongst the diseased individuals
= ((1-i(i-t)) = (1-k)

Variance in liability amongst relatives the diseased individuals = 1- $i(i-t)(a_R h^2)^2$

$$t_R = \frac{t - a_R i h^2}{\sqrt{1 - a_R^2 h^4 k}}$$

$$h^2 = \frac{t - t_R\sqrt{1 - (1 - t/i)(t^2 - t_R^2)}}{a_R(i + (i - t)t_R^2)}$$

```
> h2l=function(t,tR,i,aR){(t-tR*sqrt(1-(1-t/i)*(t^2-tR^2)))/(aR*(i+(i-t)*tR^2))} # heritability of liability with Reich et al correct
ion **use this one
> (h2l_est=h2l(t_est,t_dad,i_est,0.5))
[1] 0.7857835
> (h2l_est=h2l(t_est,t_MZ,i_est,1))
[1] 0.7985478
```

Reich, James, Morris (1972) The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. Ann Hum Gen 36: 163.

Practical

Uses simulation to give understanding to the theory.

How to calculate heritability of liability from risks to relatives.

Feel for sample size and sampling variation

Relationship between narrow sense heritability on disease and liability scales

Module 19: Statistical and Quantitative Genetics of Disease

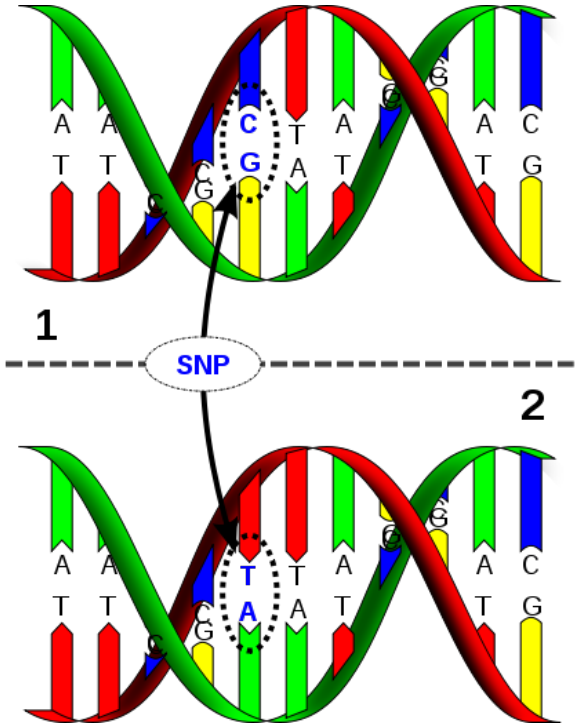
John Witte

Session #2:

Single locus analysis: design, analysis, logistic regression, covariates.

Now Assume We Can Collect DNA on Cases and Controls

- What study design should we use?
- What analytic approaches?
- Conventional: estimate impact of single genetic variants on disease.



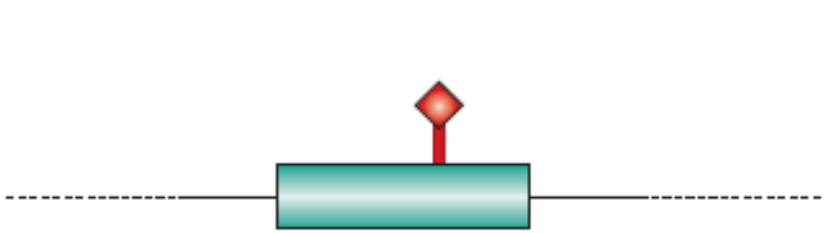
Outline

1. Association Approaches
2. Linkage Disequilibrium
3. Population Stratification / Study Design
4. Association Analysis
5. Odds ratios and relative risks
6. Logistic regression
7. Covariates

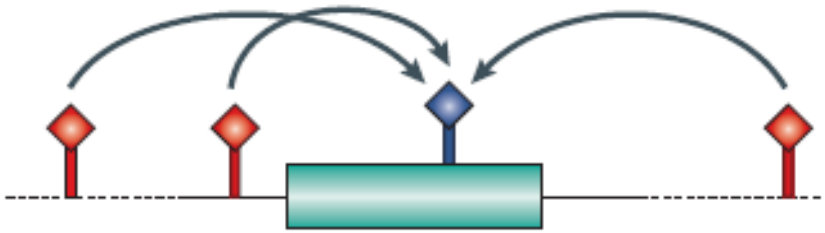
1. Association Study Approaches

- Direct vs. Indirect
- Candidate genes: hypotheses about biological mechanisms.
  - Functional
  - All common variants
  - Exome arrays
- All common variants in genome (GWAS)
- All variants in genes/genome (sequencing)
  - Expensive

2. Linkage Disequilibrium



Direct association



Indirect association

Hirschhorn & Daly, Nat Rev Genet 2005

The non-random association of alleles at two or more loci, that descend from single, ancestral chromosomes.

Assume two loci with alleles {A, a} and {B, b}

$D = P_{AB} - p_A p_B = P_{AB}P_{ab} - P_{Ab}P_{aB}$

$D' = D / \max(D)$

where

$\max(D) = \min(p_A p_b, p_a p_B)$ if $D>0$ or $\min(p_A p_B, p_a p_b)$ if $D<0$

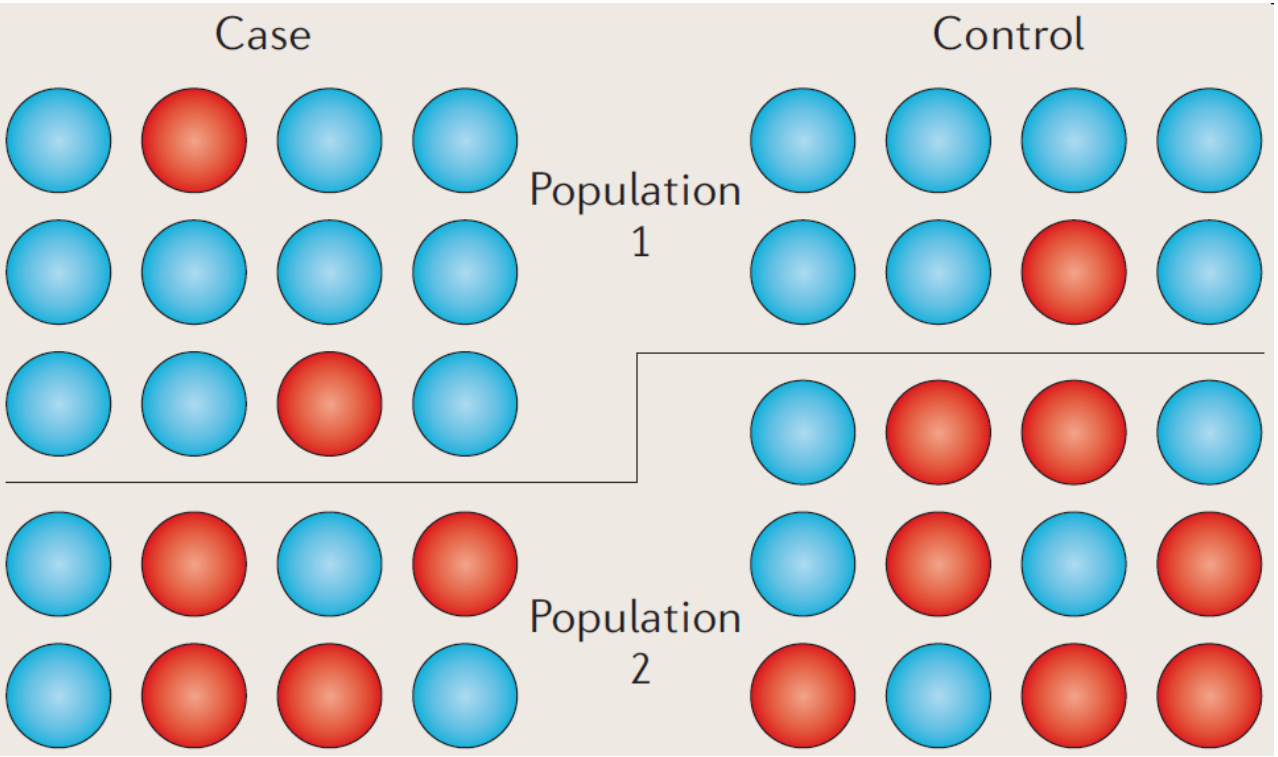
$r^2 = D^2/(p_A p_a p_B p_b)$.

3. Population Stratification & Study Design

- Key principle of association studies: select controls from the cases' source population.
- Those individuals who—if they were diseased—would become cases.
- Otherwise potential for bias (e.g., population stratification) and reduced efficiency.

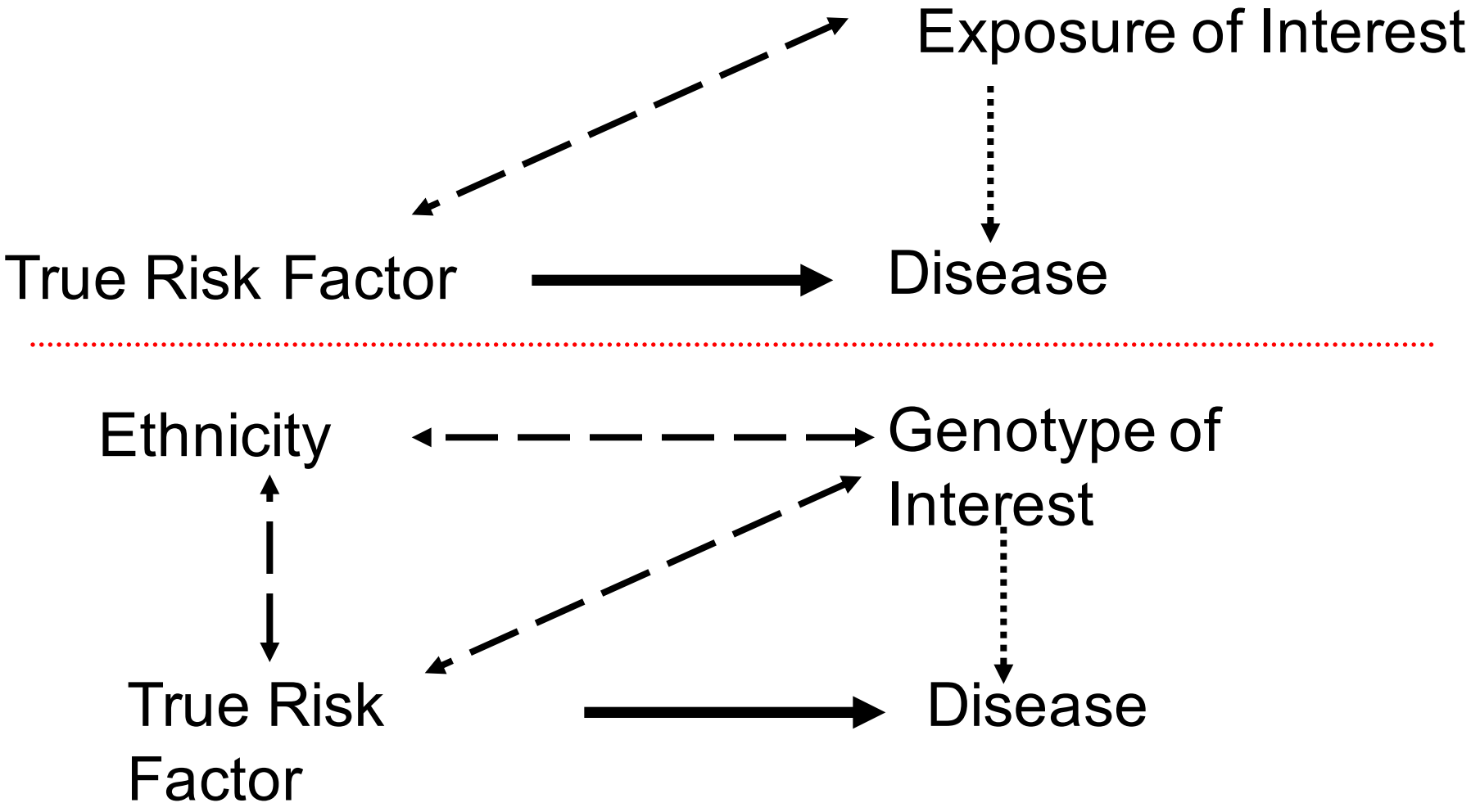
Population Stratification

- Two populations have different allele frequencies and background rates of disease.
- Can lead to biased association results.



Balding, Nature Reviews Genetics 2010

Population Stratification: Confounding



Wacholder, *JNCI*, 2000

Example

Study Population: 4,290 Pima and Papago Native Americans

Genetic Variant: Gm 3;5,13, 15 haplotype (Gm system of human immunoglobulin G)

Outcome: Type 2 diabetes

Question: Is the Gm 3; 5,13, 15 haplotype associated with Type 2 diabetes?

Knowler, *AJHG*, 1998

Population Stratification: Gm3;5,13,14 in admixed sample of Native Americans of the Pima and Papago tribes

Full heritage Native American population

	+	-
Gm3;5,13,14	~1%	~99%

NIDDM prevalence ~40%

Caucasian population

	+	-
Gm3;5,13,14	~66%	~34%

NIDDM prevalence ~15%

Gm3,5,13,14 haplotype	Cases	Controls
+	7.80%	29.00%
-	92.20%	71.00%

Unadjusted for ethnic background

OR = 0.27 (95% 0.18-0.40)

Different genotype frequency, different phenotype frequency

Population Stratification: Gm3;5,13,14 in admixed sample of Native Americans of the Pima and Papago tribes

Index of N Am heritage	Gm3;5,13,14 haplotype	% Diabetes
0	65.8%	18.5%
4	42.1%	28.5%
8	1.6%	39.2%

Gm3,5,13,14 haplotype	Cases	Controls
+	7.80%	29.00%
-	92.20%	71.00%

Adjusted for ethnic background

OR = 0.83 (95% 0.58-1.18)

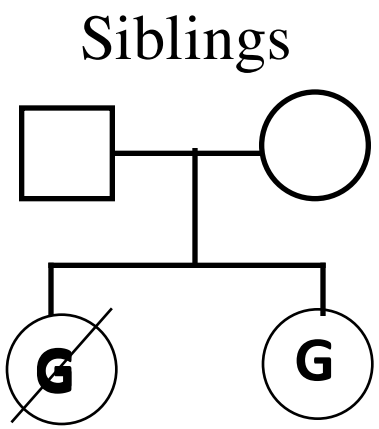
Previous result just picked out race/ethnicity!

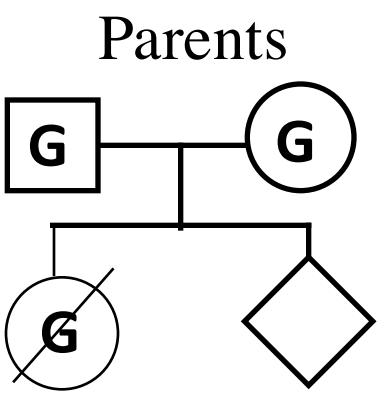
How can we address the potential bias due to population stratification?

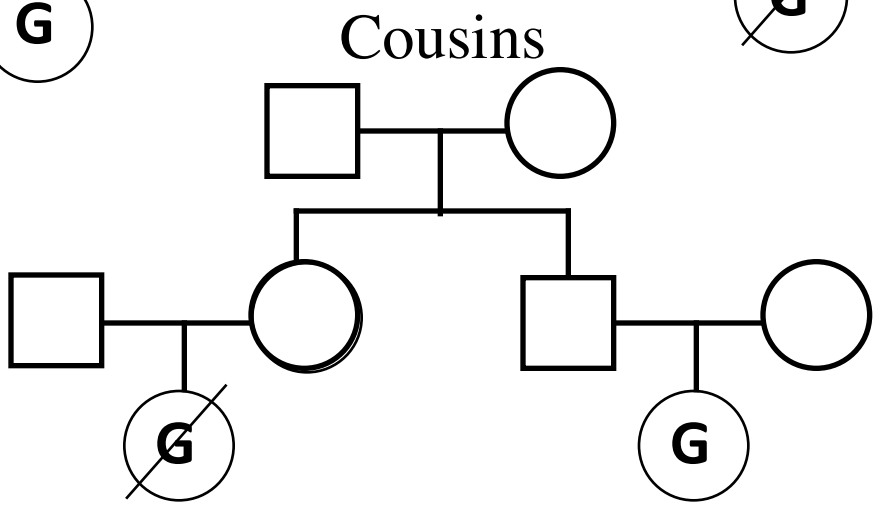
Addressing Population Stratification

- Match on self-reported ethnicity
  (Wacholder et al., / Thomas & Witte, CEBP 2002)
- Family-based studies
  (Witte et al., AJE 1999)
- Genomic control
  (Devlin and Roeder, Biometrics, 1999)
  Adjust test statistics for 'inflation' (bias) using empirical χ^2 distribution, comparing median observed to expected ($\chi^2_{new} = \chi^2_{old}/\diamond$).
- Principal Components
  (Price et al., Nat Genet 2006)
  Adjust regression for PCs as a proxy for genetic ancestry.

Family-Based Association Studies







Termed "**psuedo**"-controls (the hypothetical other mendelian transmissions)

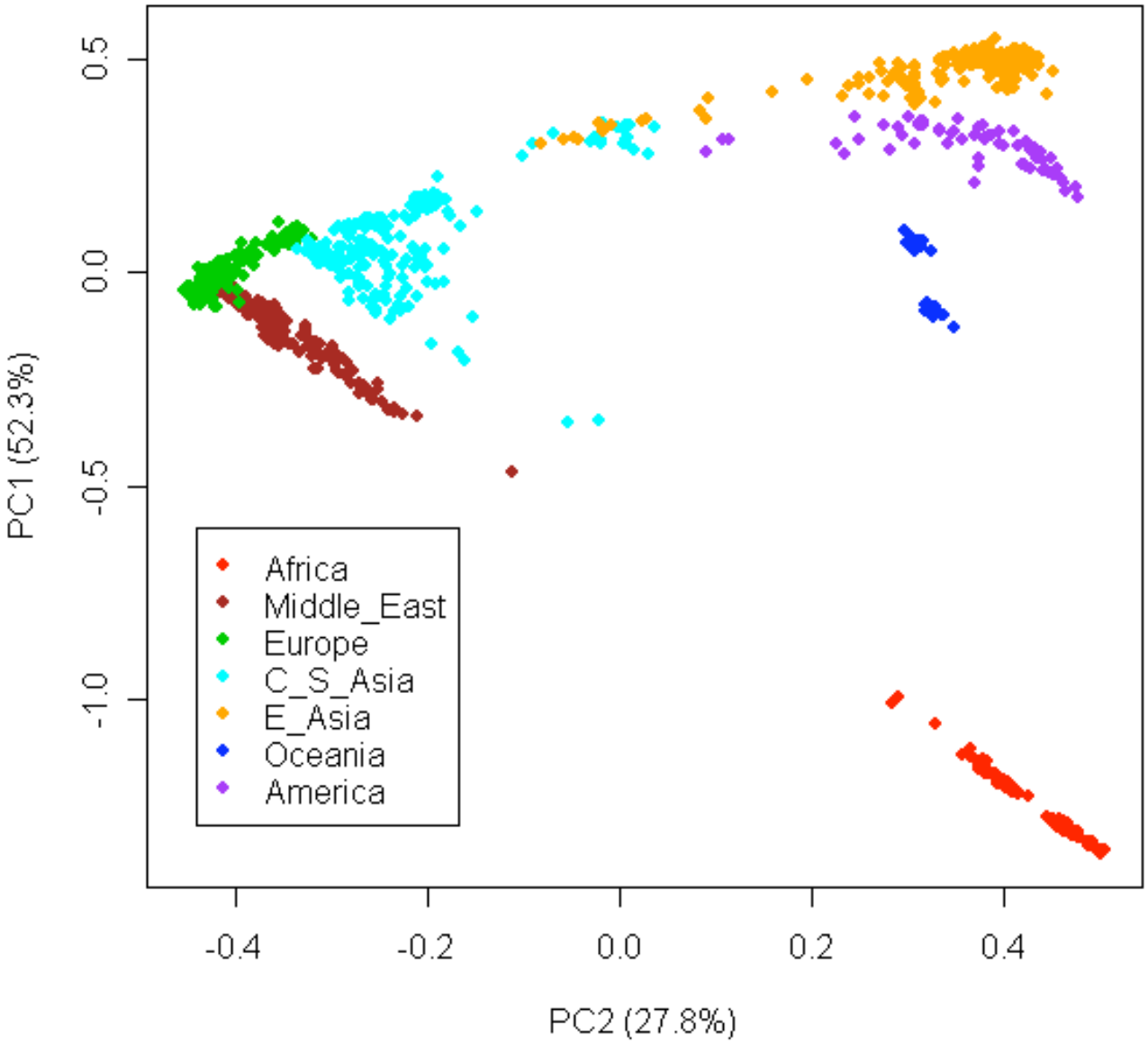
Comparison of Designs

- Family-based designs can:
  - Be less efficient than population-based designs.
  - Require more recruitment efforts

	Rare Recessive High Risk	Common Low Risk	Rare Dominant High Risk
Population-based	100%	100%	100%
Case-sibling	69%	51%	50%
Case-cousin	97%	88%	88%
TDT	231%	102%	101%

Witte et al. AJE 1999

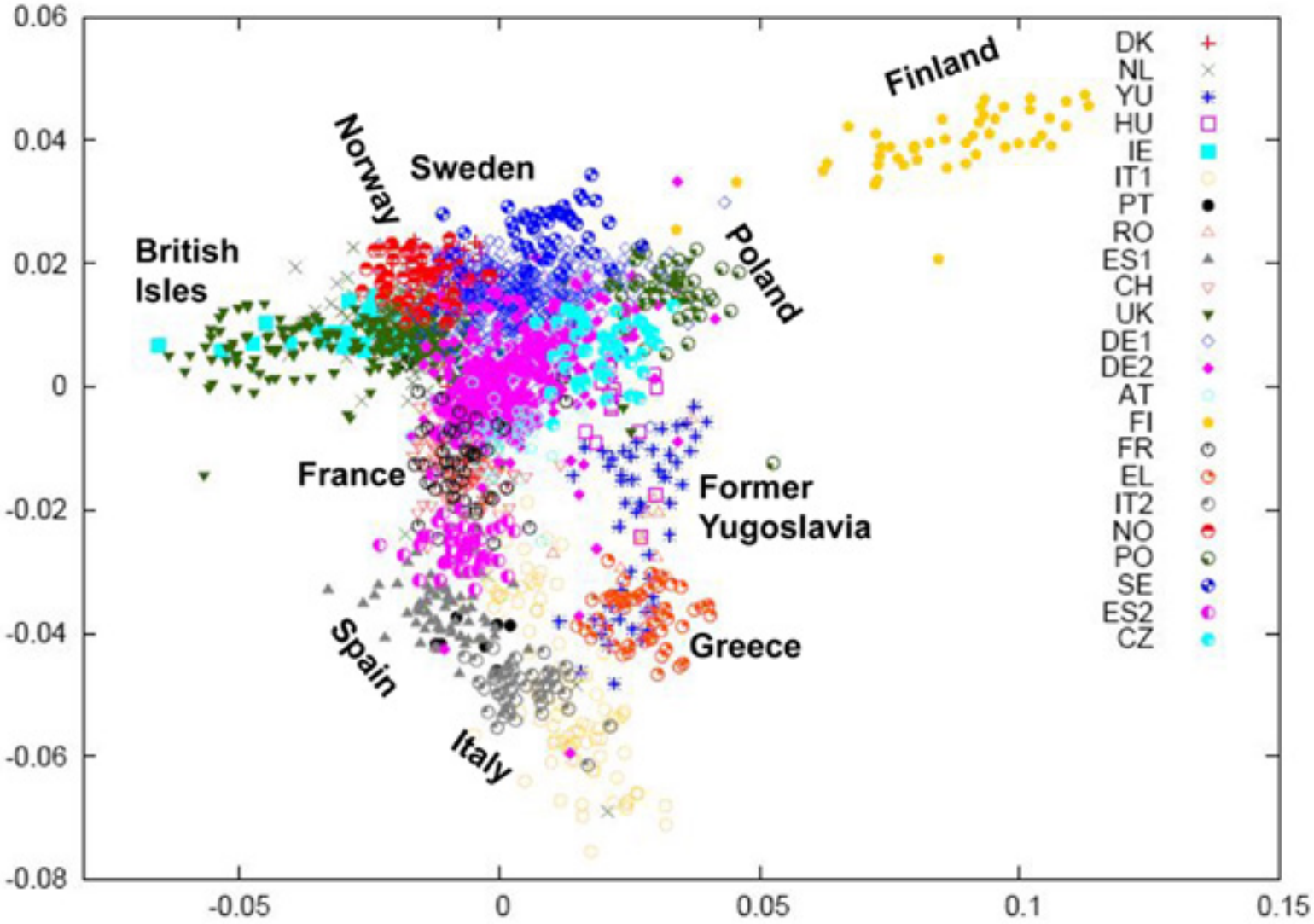
Adjusting for Principal Components



- Maximize variance between subjects using all SNPs.
- Clusters individuals from different populations.

Li et al., Science 2008

PCs Detect Fine Population Structure



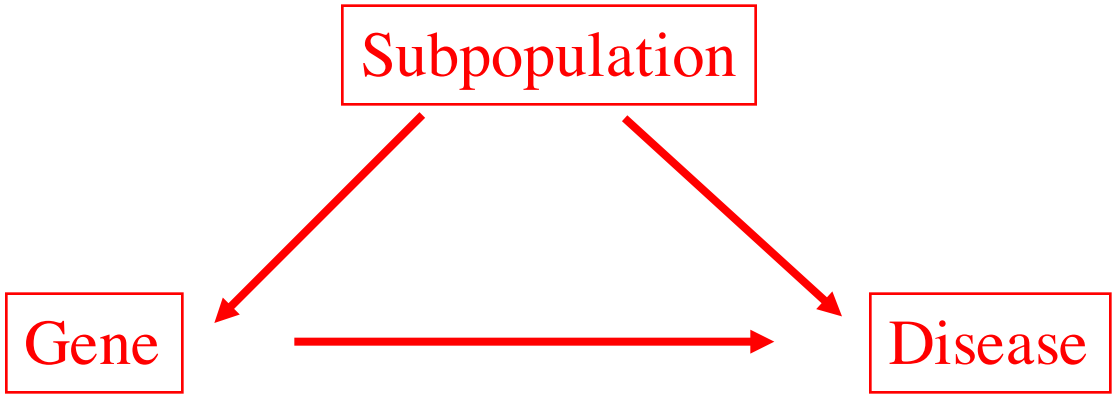
Razib, Current Biology 2008

Continuum of Assoc Study Designs

Population Stratification

Overmatching

(Bias……versus……..efficiency)



Sharing of genes & envt.

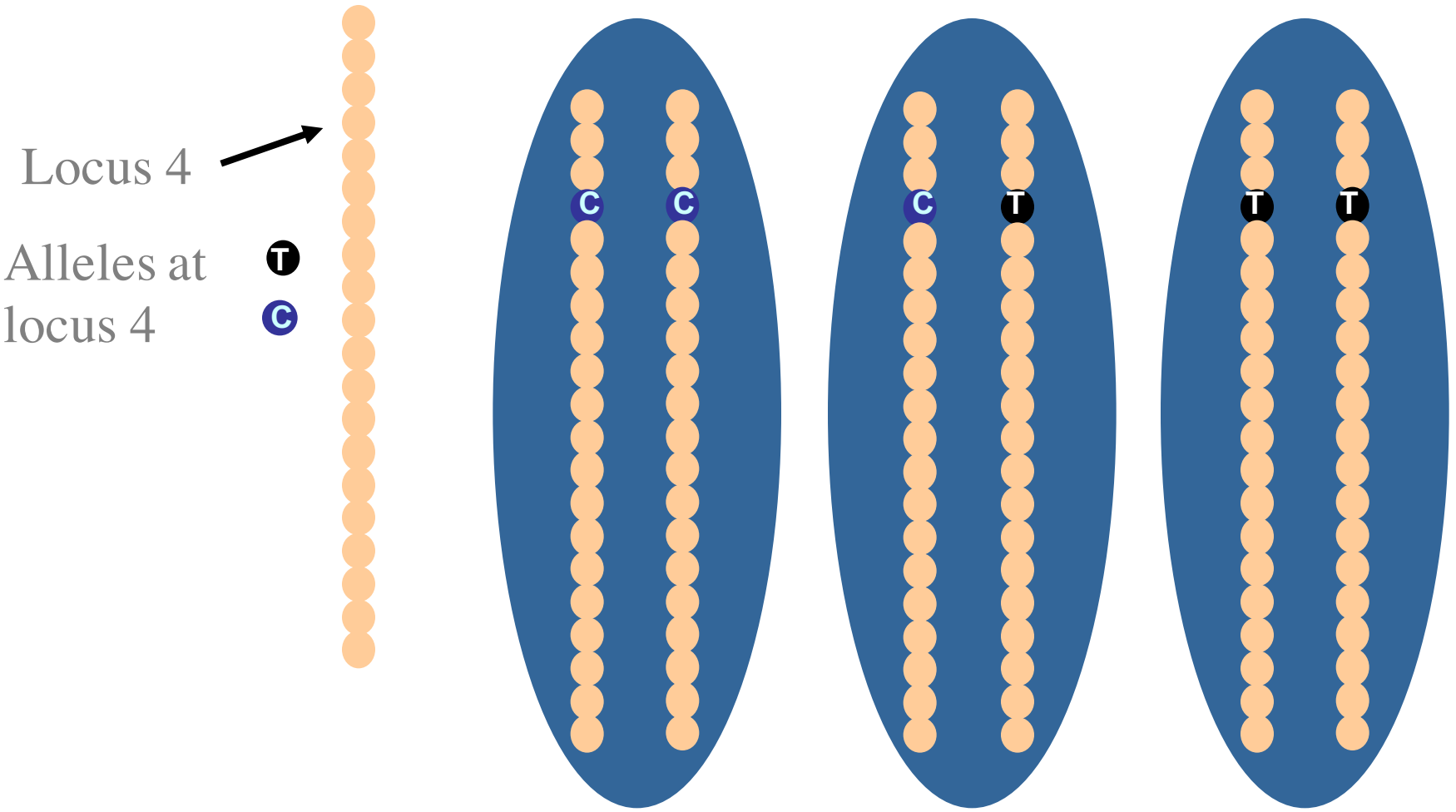
Efficiency

Also, recruitment issues

Outline

1. Association Approaches
2. Linkage Disequilibrium
3. Population Stratification / Study Design
4. Association Analysis
5. Odds ratios and relative risks
6. Logistic regression
7. Covariates

4. Association Analysis: Genotypes



Locus: chromosomal location that's polymorphic.
Alleles: different variants @ locus

- Each somatic cell is diploid (two copies of each autosome)
- Thus 3 genotypes at locus 4 (use only one strand, often forward): CC, CT, TT

Association Analysis

Genotype	Cases	Controls	OR
CC	A	D	AF/DC
CT	B	E	BF/EC
TT	C	F	1

Simple chi-square test comparing genotype frequencies (2 d.f.)
Called a co-dominant analysis

Testing for Association

Observed:					Expected	
Geno	Case	Control	Total	OR	Case	Control
CC	A	D	A+D=nCC	AF/DC	nCC*nCase/n	nCC*nCont/n
CT	B	E	B+E=nCT	AE/BD	nCT*nCase/n	nCT*nCont/n
TT	C	F	C+F=nTT	1	nTT*nCase/n	nTT*nCont/n
Total	A+B+C =nCase	D+E+F =nCont	A+B+C+D+E+F =n			

Sum (Observed – Expected)^2/Expected. Chi squared with 2 degrees of freedom.

Expected cell count = row_total * column_total / total

Testing for Association

Observed:					Expected	
Geno	Case	Control	Total	OR	Case	Control
CC	20	5	25	12	25*35/65=13.5	25*30/65=11.5
CT	10	10	20	3	20*35/65=10.8	20*30/65=9.2
TT	5	15	20	1	20*35/65=10.8	20*30/65=9.2
Total	35	30	65			
	=nCase	=nCont	=n			

Sum (Observed – Expected)^2/Expected

= (20-13.5)^2/13.5 + (10-10.8)^2/10.8 + (5-10.8)^2/10.8
+ (5-11.5)^2/11.5 + (10-9.2)^2/9.2 + (15-9.2)^2/9.2

= 13.7

P-value = 0.0011

Co-dominant model

Genetic Model

Genotype	OR
CC	R
CT	r
TT	1

ORs depend on genetic model

R = r = 1 not risk allele

R > r = 1 recessive

R = r > 1 dominant

$R = r^2 > 1$ log additive

(Assuming positive association)

Testing for Association

2 df Genotype		
Genotype	Case	Control
CC	A	D
CT	B	E
TT	C	F

~chi_sq(2df)

Recessive (G)		
	Case	Control
CC	A	D
CT or TT	B+C	E+F

Dominant (G)		
	Case	Control
CC or CT	A+B	D+E
TT	C	F

(1df)

What model should we use here?

Genotype	Case	Control
CC	20	5
CT	10	10
TT	5	15
	P=0.0011	

	Case	Control
CC	20	5
CT or TT	15	25
	P=0.0020	

	Case	Control
CC or CT	30	15
TT	5	15
	P=0.0045	

Genetic Model

If genetic model known:

- Collapse genotypes into 2x2 table, 1 d.f. test
- Or trend test for log additive
- Use logistic regression: coding; covariates, odds ratios

If genetic model unknown?

- Log-additive is default. Why?

- Could use all three models (dom, rec, log additive).
- Compare fit with the co-dominant (2d.f.) model (LR test).
- Can't use LR test to compare models since not nested.
- Model with best fit and smallest P is best?
- Use permutation test (MAX test).

5. Odds Ratios and Relative Risks

When does the OR estimate the RR?

1. When the disease is "rare"

	D	Ď
CC or CT	A_1	B_1
TT	A_0	B_0

$$RR = \frac{\frac{A_1}{(A_1 + B_1)}}{\frac{A_0}{(A_0 + B_0)}} = \frac{q+}{q-}$$

q+: Incidence in carriers (exposed)

q-: Incidence in non-carriers (non-exposed)

$$OR = \frac{\frac{A_1}{B_1}}{\frac{A_0}{B_0}} = \frac{\frac{q+}{(1-q+)}}{\frac{q-}{(1-q-)}} = \frac{q+}{q-} * \frac{1- \frac{A_0}{(A_0+B_0)}}{1- \frac{A_1}{(A_1+B_1)}}$$

Odds Ratios and Relative Risks

2. When exposure distribution among the controls is the same as the 'person-time' in the cases' source population.

	D	Ď
CC or CT	A_1	B_1
TT	A_0	B_0

$$\frac{I_1}{I_0} = RR$$

$$\frac{A_1}{T_1} = I_1 \qquad I_0 = \frac{A_0}{T_0}$$

$$\frac{B_1}{T_1} = \frac{B_0}{T_0} = r$$

Let:

T_1 = Amount of exposed person-time

I_1 = Incident rate of exposed

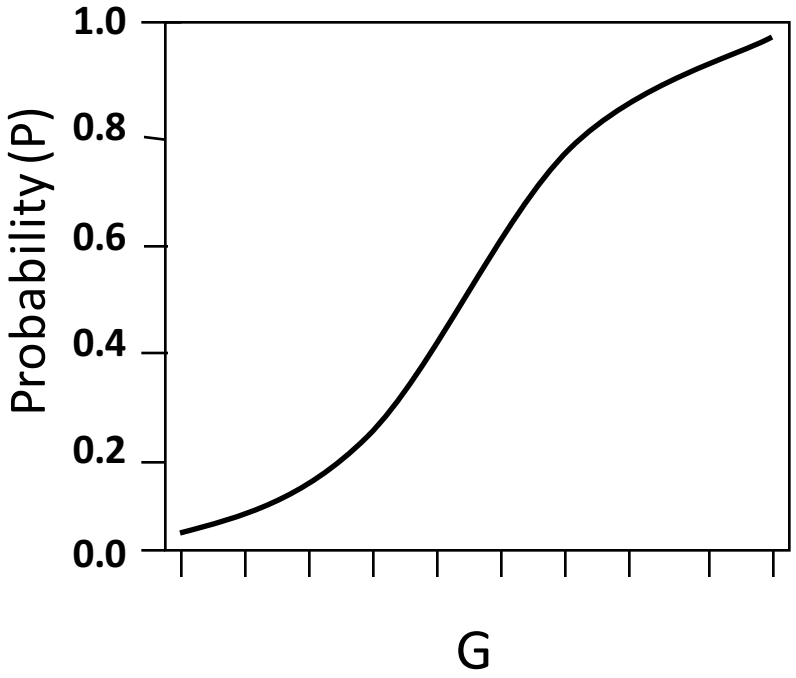
T_0 = Amount of unexposed person-time

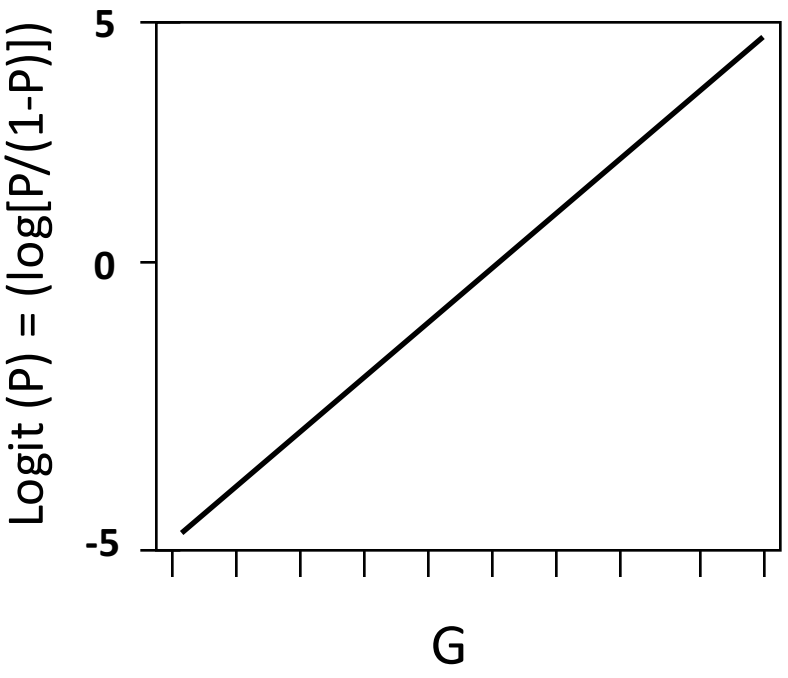
I_0 = Incident rate of unexposed

r = Sampling rate

$$OR = \frac{\frac{A_1}{B_1} * r = \frac{A_1}{T_1}}{\frac{A_0}{B_0} * r = \frac{A_0}{T_0}} = RR$$

6. Logistic Regression





$$P(D \mid G) = \frac{1}{1 + e^{-(\alpha + \beta G)}}$$

$$\log\left[\frac{P}{(1-P)}\right] = \alpha + \beta G$$

The log odds of disease increases linearly with G.

Interpretation of Coefficients

- The logistic regression coefficients: $\beta = \log(OR)$

- Assume G=1 (carrier), G=0 (non-carrier)

$$\log [P_1 /(1 - P_1)] = \alpha + \beta * 1$$

$$\log [P_0 /(1 - P_0)] = \alpha + \beta * 0$$

so

$$\log [P_1 /(1 - P_1)] - \log [P_0 /(1 - P_0)] = \beta$$

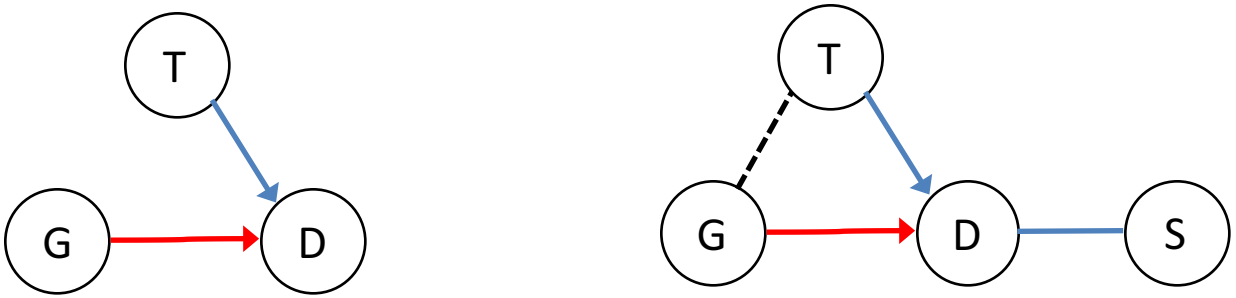
or

$$\log[P_1 /(1 - P_1) / (P_0 /(1 - P_0))] = \log(OR) = \beta$$

- The OR for the effect of G on disease risk is e^{β}
- For multiple variants, assumes joint effects are multiplicative.

7. Including Covariates in Regression

- Confounders: PCs for population stratification.
- Modifiers: Envt or Genetic interactions.
- Independent predictors?

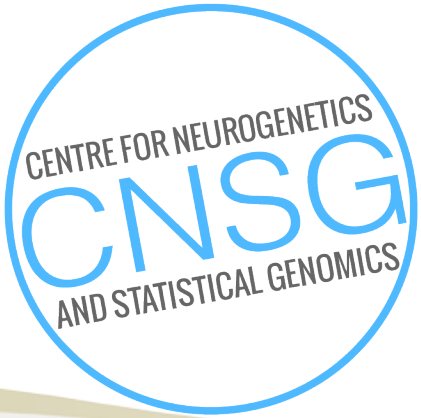


Zaitlen et al.; Mefford & Witte, PloS Genet, 2012

Module 18:
Statistical & Quantitative Genetics of Disease

Lecture 3
Polygenic models of disease risk

Naomi Wray



Aims of Lecture 3

Theory

- Single locus disease model
- Power calculations

Single locus disease model

Single locus disease model:
G = genotype; D=disease; K = overall disease risk in population

	P(G)
aa	$(1\text{-}p)^2$
Aa	$2p(1\text{-}p)$
AA	p^2

Single locus disease model

Single locus disease model:
G = genotype; D=disease; K = overall disease risk in population

	P(G)	P(D|G)
aa	$(1\text{-}p)^2$	f_0
Aa	$2p(1\text{-}p)$	f_0R
AA	p^2	f_0R^2

$$P(\text{Disease})=K=f_0(1\text{-}p)^2+f_0R2p(1\text{-}p)+f_0R^2 \quad =f_0(1+p(R\text{-}1))^2$$

$$f_0=K/(1+p(R\text{-}1))^2$$

Single locus disease model

Single locus disease model:
G = genotype; D=disease; K = overall disease risk in population

	P(G)	P(D|G)	P(D) =P(D|G)p(G)
aa	$(1\text{-}p)^2$	f_0	$(1\text{-}p)^2 f_0$
Aa	$2p(1\text{-}p)$	f_0R	$2p(1\text{-}p) f_0R$
AA	p^2	f_0R^2	$p^2 f_0R^2$
			Sum= K

$$P(\text{Disease})=K=f_0(1\text{-}p)^2 + f_0R2p(1\text{-}p) + f_0R^2 \quad = f_0(1+p(R\text{-}1))^2$$

$$f_0=K/(1+p(R\text{-}1))^2$$

Power of association test – case/control

Single locus disease model:
G = genotype; D=disease; K = overall disease risk in population

	P(G)	P(D|G)	P(D) =P(D|G)p(G)	P(G|D) =P(G)/P(D)
aa	$(1\text{-}p)^2$	f_0	$(1\text{-}p)^2 f_0$	$(1\text{-}p)^2 f_0/K$
Aa	$2p(1\text{-}p)$	f_0R	$2p(1\text{-}p) f_0R$	$2p(1\text{-}p) f_0R/K$
AA	p^2	f_0R^2	$p^2 f_0R^2$	$p^2 f_0R^2/K$
			Sum= K	

$$P(\text{Disease})=K=f_0(1\text{-}p)^2 + f_0R2p(1\text{-}p) + f_0R^2 \quad = f_0(1+p(R\text{-}1))^2$$

$$f_0=K/(1+p(R\text{-}1))^2$$

What is power?

When we set up a statistical test

- The null hypothesis is EITHER
  - true
  - false
- With the data available we EITHER
  - reject the null hypothesis
  - fail to reject the null hypothesis

	Null hypothesis is true	**Null hypothesis is false**
Reject the null hypothesis	Type I error False positive	Correct Outcome True positive
Fail to reject the null hypothesis	Correct Outcome True negative	Type II error False Negative

Power = probability of rejecting the null hypothesis when the null hypothesis is false

=1 –probability of failing to reject the null hypothesis when the null hypothesis is false

= 1- probability(Type II error)

Power depends on statistical test, effect size to be detected, sample size, acceptable level of Type I error

Non-centrality parameter depends on statistical test, effect size to be detected, sample size

Relative power of a GWAS for a quantitative trait compared to a disease trait

First step:

How to calculate power in an association study?

Genetic Power Calculator

S. Purcell & P. Sham, 2001-2009

This site provides automated power analysis for variance components (VC) quantitative trait locus (QTl

If you use this site, please reference the following Bioinformatics article:

```
Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator:
design of linkage and association genetic mapping studies of complex
traits. Bioinformatics, 19(1):149-150.
```

Modules

Genetic Power Calculator

Quantitative Case-Control

Total QTL variance	:		(0 - 1)
Dominance : additive QTL effects	:		(0 - 1)
QTL increaser allele frequency	:		(0 - 1)
Marker M1 allele frequency	:		(0 - 1)
Linkage disequilibrium (D-prime)	:		(0 - 1)
Number of cases	:		(>0)
Case lower threshold	:		
Case upper threshold	:		
Control:case ratio	:		(>0)
Controls lower threshold	:		
Controls upper threshold	:		
User-defined type I error rate	:	0.05	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	:	0.80	(0 - 1)

Send Clear Form

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A)	:		(0 - 1)
Prevalence	:		(0.0001 - 0.9999)
Genotype relative risk Aa	:		(>1)
Genotype relative risk AA	:		(>1)
D-prime	:		(0 - 1)
Marker allele frequency (B)	:		(0 - 1)
Number of cases	:		(0 - 10000000)
Control : case ratio	:		(>0) (1 = equal number of cases and controls)
Unselected controls? (* see below)			
User-defined type I error rate	:	0.05	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	:	0.80	(0 - 1)

Process Reset

Created by Shaun Purcell 24.Oct.2008

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A) : .2 (0 - 1)
Prevalence : .01 (0.0001 - 0.9999)
Genotype relative risk Aa : 1.2 (>1)
Genotype relative risk AA : 1.44 (>1)

D-prime : 1 (0 - 1)
Marker allele frequency (B) : .2 (0 - 1)

Number of cases : 5000 (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)

Case-control statistics: allelic 1 df test (B versus b)
Sample NCP = 28.59

Alpha	Power	N cases for 80% power
0.1	0.9999	1081
0.05	0.9996	1372
0.01	0.9972	2042
0.001	0.9802	2985
5e-08	0.4586	6924

Power of a case-control study

Power of a disease trait

p = frequency of risk allele in population

p_{case} = frequency of risk allele in cases

p_{cont} = frequency of risk allele in controls

v = proportion of a sample of N that are cases

$\bar{p}$ = mean allele frequency across cases and controls

= v p_{case} + (1-v) $p_{control}$

Yang et al (2009)

Power of association test – case/control

Single locus disease model:
G = genotype; D=disease; K = overall disease risk in population

	P(G)	P(D|G)	P(D) =P(D|G)p(G)	P(G|D) =P(G)/P(D)
aa	$(1-p)^2$	f_0	$(1-p)^2 f_0$	$(1-p)^2 f_0/K$
Aa	$2p(1-p)$	f_0R	$2p(1-p) f_0R$	$2p(1-p) f_0R/K$
AA	p^2	f_0R^2	$p^2 f_0R^2$	$p^2 f_0R^2/K$
			Sum= K	

$P(\text{Disease})=K = f_0(1-p)^2 + f_0R2p(1-p) + f_0R^2 \quad = f_0(1+p(R-1))^2$

$f_0 = K/(1+p(R-1))^2$

p_{case} = ½ P(Aa|D)+p(AA|D) Allele frequency in cases

$$= f_0 pR\big((1-p)+pR\big)/K = \frac{pR}{(1+p(R-1))}$$

Find allele frequency in controls in the same way

$$p_{cont} = \frac{p}{1-K}\left(1 - \frac{KR}{(1+p(R-1))}\right)$$

Power of a case-control study

Power of a disease trait

p = frequency of risk allele in population

p_{case} = frequency of risk allele in cases

p_{cont} = frequency of risk allele in controls

v = proportion of a sample of N that are cases

$\bar{p}$ = mean allele frequency across cases and controls

= v p_{case} + (1-v) $p_{control}$

Z-Test statistic of association = test of difference of two proportions =

$$\frac{p_{case} - p_{cont}}{s.e.(pooled\ sample\ p)} = \frac{p_{case} - p_{cont}}{s.e.(\bar{p})}$$

χ^2 non-centrality parameter $= NCP_{01} = \dfrac{(p_{case} - p_{cont})^2}{var(\bar{p})}$

$$var(\bar{p}) = 2\bar{p}(1-\bar{p})\left(\frac{1}{Nv} + \frac{1}{N(1-v)}\right)$$

Power of a case-control study

$$NCP_{01} = \frac{(p_{case} - p_{cont})^2}{var(\bar{p})}$$

a = significance level - acceptable level of type I error

$t = \Phi^{-1}\left(\frac{\alpha}{2}\right)$ Normal distribution threshold above which null hypothesis will be rejected

$$Power = \Phi\left(\sqrt{NCP_{01}} + t\right)$$

N=10000,v=0.5,p=0.2,R=1.2,K=0.01,a=5e-8,K=0.01, power = 0.46

Agrees with the genetic power calculator

Yang et al (2009) Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies. Genetic Epidemiology

Approximate variance explained by a locus

Regression of disease on jth SNP, $x_{[j]}$ = 0,1,2

$$y_{01} = K + b_{01}x_{[i]} + \varepsilon$$

When x[j]=0 $\widehat{y_{01}} = K$ = P(Disease | Genotype =aa)

When x[j]=1 $\widehat{y_{01}} = K + b_{01}$ = P(Disease | Genotype =Aa)

Relative Risk = R= P(Disease | Genotype =Aa)/P(Disease | Genotype =Aa)

$$= (K+b_{01})/K \quad \text{so} \quad b_{01} = K(R-1)$$

Variance attributable to the locus on the disease scale

$$\sigma^2_{A_{01[j]}} = h^2_{01[j]}K(1-K) = b^2_{01}var(x) = 2p(1-p)b^2_{01}$$

$$h^2_{01[j]} = 2p(1-p)b^2_{01}/K(1-K)$$

$$h^2_{L[j]} = \frac{(1-K)h^2_{01[i]}}{i^2K} = \frac{2p(1-p)b^2_{01}}{i^2K^2} = \frac{2p(1-p)(R-1)^2}{i^2}$$

Assumes a population sample not a case control sample

See Lecture 1: Dempster & Lerner (1950) Appendix by Alan Robertson. Heritability of threshold characters. Genetics 35

Power of a case-control association study expressed in terms of variance explained by the locus

χ^2 non-centrality parameter $= NCP_{01} = \dfrac{(p_{case} - p_{cont})^2}{var(\bar{p})}$

$$NCP_{01} = \frac{2\bar{p}(1-\bar{p})(R-1)^2 v(1-v)N}{(1-K)^2(1+p(R-1))^2}$$

If R is small then $(1+p(R-1))^2 \approx 1$ e.g., p=0.2, R=1.2, $(1+p(R-1))^2 = 1.08$

Variance explained by a locus $= h^2_{L[j]} \approx \dfrac{2p(1-p)(R-1)^2}{i^2}$

$$NCP_{01} \approx \frac{h^2_{L[j]} i^2 v(1-v)N}{(1-K)^2}$$

Yang et al (2009) Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies. Genetic Epidemiology

Power of a association study of a quantitative trait

χ^2 *non-centrality parameter* $= NCP_{QT} = \dfrac{N_{QT} h^2_{L[i]}}{1 - h^2_{L[i]}}$

When the variance explained is the same in c-c and for quantitative trait

$NCP_{01} \approx \dfrac{h^2_{L[j]} i^2 v(1-v) N_{01}}{(1-K)^2}$

$$\frac{NCP_{01}}{NCP_{QT}} \approx \frac{i^2 v(1-v) N_{01}}{(1-K)^2 N_{QT}}$$

Yang et al (2009) Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies. Genetic Epidemiology

Practical

- Power in case-control study design
  - Code of slides 3-6
  - Curve function for power in case-control study design

Module 19: Statistical and Quantitative Genetics of Disease: Interpreting measures of variation explained; Multivariate analysis

John Witte

Lecture #4

Outline

1. Measures of Variation Explained
2. Multivariate Analysis

1. Measures of Variation Explained

- Assume we've identified risk variants from single locus models.
- Once discovered, what next?
  - Search for more risk variants?
  - Focus on their biology?
  - Probably both!
- Depends on their overall impact on disease.
- Can assess with a number of measures
  - give values between 0 and 100%

Measures to Assess Impact

- Heritability explained
- Sibling recurrence risk explained
- Log RR: familial risk explained
- Area under the receiver-operating curve (AUC)
- Population attributable fraction (PAF)

Key questions:

- How do these measures compare?
- Do they provide similar info?
- Does genetic architecture of disease impact differences?

Different Messages?

- Results in contrasting and confusing use of these measures.
- Example,
  - for Crohn's disease variants in *NOD2* reported to explain:
    - 1-2% of heritability
    - ~5% of familial risk
    - 18% of the PAF

Heritability Explained

Measures	Genotype[a]		
	bb	Bb	BB
General notation			
Population frequency[b]	$(1-p)^2$	$2p(1-p)$	p^2
Genotype risk[c]	w_{bb}	w_{Bb}	w_{BB}
Mean genotype risk (M)[d]	$(1-p)^2\ w_{bb}$	$2p(1-p)\ w_{Bb}$	$p^2\ w_{BB}$
Variance of genotype risk (V)[d]	$(1-p)^2\ (w_{bb} - M)^2$	$2p(1-p)\ (w_{Bb} - M)^2$	$p^2\ (w_{BB} - M)^2$
Scale-specific genotype risks			
Observed risk[e]	k_{bb}	$k_{bb}\ RR_{Bb}$	$k_{bb}\ RR_{BB}$
Relative risk	*1*	RR_{Bb}	RR_{BB}
Log relative risk	0	$\log(RR_{Bb})$	$\log(RR_{BB})$
Liability threshold[f]	$-\Phi^{-1}(1- k_{bb})$	$-\Phi^{-1}\ (1- k_{bb}\ RR_{Bb})$	$-\Phi^{-1}\ (1- k_{bb}\ RR_{BB})$
Quantitative genetics notation			
Genotype risk	-a	$d = w_{Bb}-(w_{bb}+w_{BB})/2$	$a = w_{BB} - (w_{bb}+w_{BB})/2$
Deviations from the mean[g]			
Total	$-a-M = -2p(a+(1-p)d)$	$d-M = a((1-p)-p)+d(1-2p(1-p))$	$a-M = 2(1-p)(a-pd)$
Additive[h]	$-2p\alpha$	$((1-p)-p)\alpha$	$2(1-p)\alpha$
Dominance	$-2p^2d$	$2p(1-p)d$	$2(1-p)^2d$

Heritability Explained

Heritability: $h^2_{L[i]} = V_{AL[i]} / V_{PL[i]} = V_{AL[i]} / (V_{GL[i]} + 1)$

where

V*L[i] = additive (*=A), phenotype (*=P), genetic (*=G) variance.

$V_A = (1-p)^2 4p^2\alpha^2 + 2p(1-p)((1-p)-p)^2\alpha^2 + p^2 4(1-p)^2\alpha^2$

$= 2p(1-p)\alpha^2$

$\alpha = a+d((1-p)-p)$ (ave effect of replacing a b allele by a B allele).

$V_D = (1-p)^2 4p^4 d^2 + 2p(1-p)4p^2(1-p)^2 d^2 + p^2 4(1-p)^4 d^2$

$= (2p(1-p)d)^2$

$V_G = V_A + V_D$ (Applied to liability risk genotypic values.)

Heritability explained: $h^2_{L[i]} / h^2_L$

Across multiple variants: $h^2_{L[i]} / h^2_L$ (Falconer & Mackay 1996)

Heritability Approximation

If we can assume small RR and a multiplicative model ($RR_{Bb}^2 = RR_{BB}$).

Then, $h^2_{Lapprox[\,i]} = 2p(1-p)(RR_{Bb}-1)^2/x^2$

where

x = the mean liability of cases, approximated as z/K

z is the height of the standard normal distribution at the threshold T that truncates the proportion K, $T= \Phi^{-1}(1-K)$

Heritability explained: $h^2_{Lapprox[i]} / h^2_L$

Stahl et al., Nat Genet 2012

Sibling Recurrence Risk Explained

- Proportion of the total sibling risk explained by the risk variants (observed scale).
- Siblings share $V_{AO}/2 + V_{DO}/4$ of risk.

$$\lambda_{S[i]} = 1 + \frac{V_{AO[i]}/2 + V_{DO}[i]/4}{K^2}$$

$V_{AO[i]} = k^2_{bb} 2*p(1-p)(p*(RR_{BB}-RR_{Bb})+(1-p)*(RR_{Bb}-1))^2$

$V_{DO[i]} = k^2_{bb} p^2(1-p)^2(RR_{BB}+1-2*RR_{Bb})^2$

Sibling risk explained: $\log(\lambda_{S[i]}) / \log(\lambda_S)$

Across multiple variants: $\sum \log(\lambda_{s[i]}) / \log(\lambda_s)$

Log RR: Familial Risk Explained

- More epidemiologic approach.
- Genetic variance attributable to the ith locus on the log risk scale:

$$V_{Glog[i]} = (1-p)^2\mathrm{M}^2 + 2p(1-p)(\log(RR_{Bb}) - \mathrm{M})^2 + p^2(\log(RR_{BB}) - \mathrm{M})^2$$

where M is the mean value of log relative risk,

M= $2p(1\text{-}p)\log(RR_{Bb}) + p^2\log(RR_{BB})$.

$$V_{Glog[i]} = 2p(1-p)\log(RR_{Bb})^2$$

- Multiple alleles, log-risk ~N with var=$2\log(\lambda_S)$
- Variation explained: $V_{Glog[i]}/\ 2\log(\lambda_S)$
- Across multiple variants $\sum_i V_{Glog[i]}/\ 2\log(\lambda_S)$

Pharoah et al., Nat Genet 2002

Area Under the Curve

$$AUC_{L[i]} = \Phi\left(\frac{(x-v)h^2_{L[i]}}{\sqrt{h^2_{L[i]}(1-h^2_{L[i]}x(x-T)+1-h^2_{L[i]}v(v-T))}}\right)$$

where

x = mean liability among cases

v = -x * K(1-K)

T= population threshold (determined from the disease prevalence K)

- Proportion explained: divide risk variant AUC by the maximum attainable AUC for a genetic risk predictor.

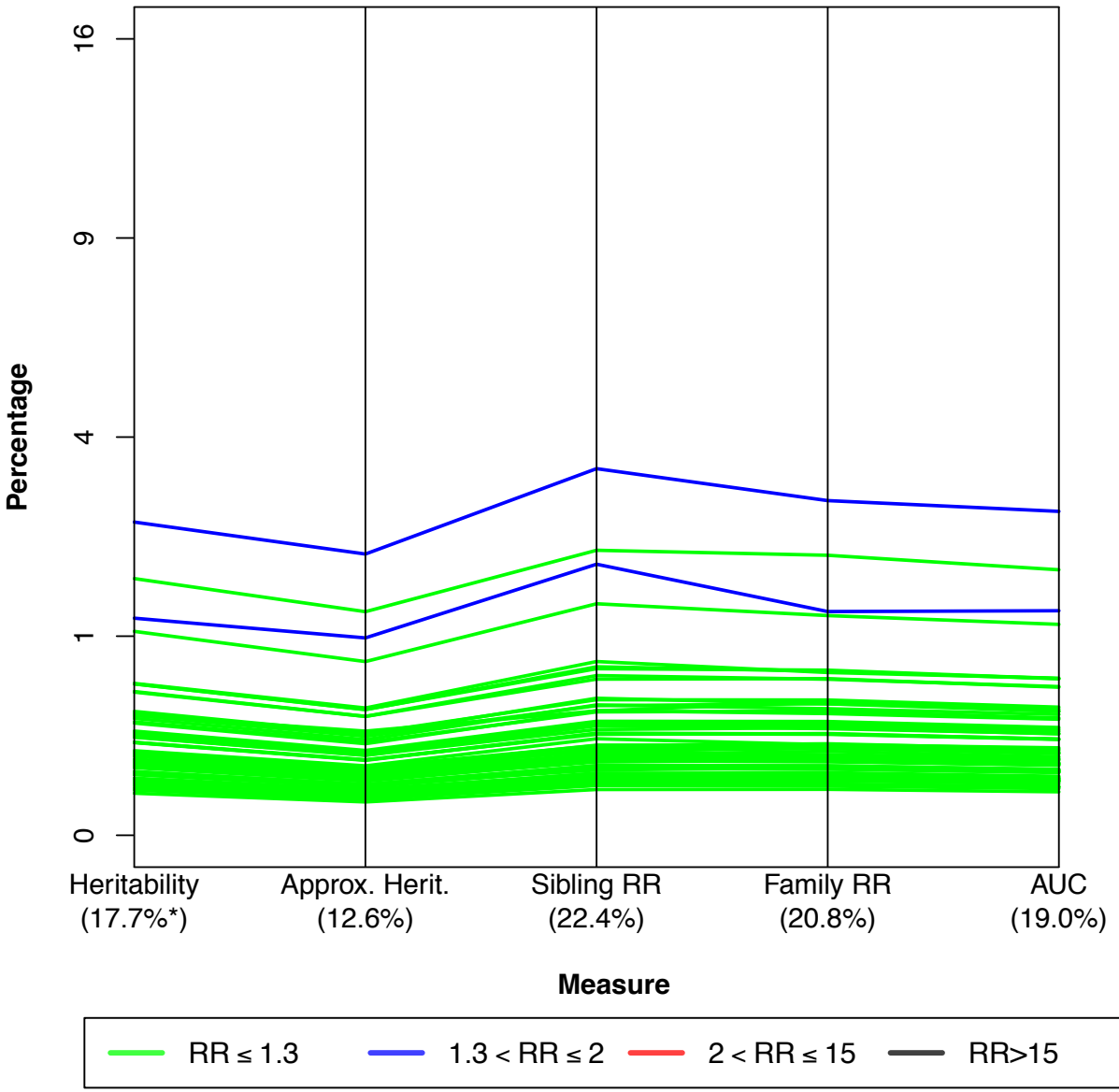
  $[(AUC_{L[i]}-0.5) / (AUC_{Max}-0.5)]^2$

Wray et al., Plos Genet 2010

Application

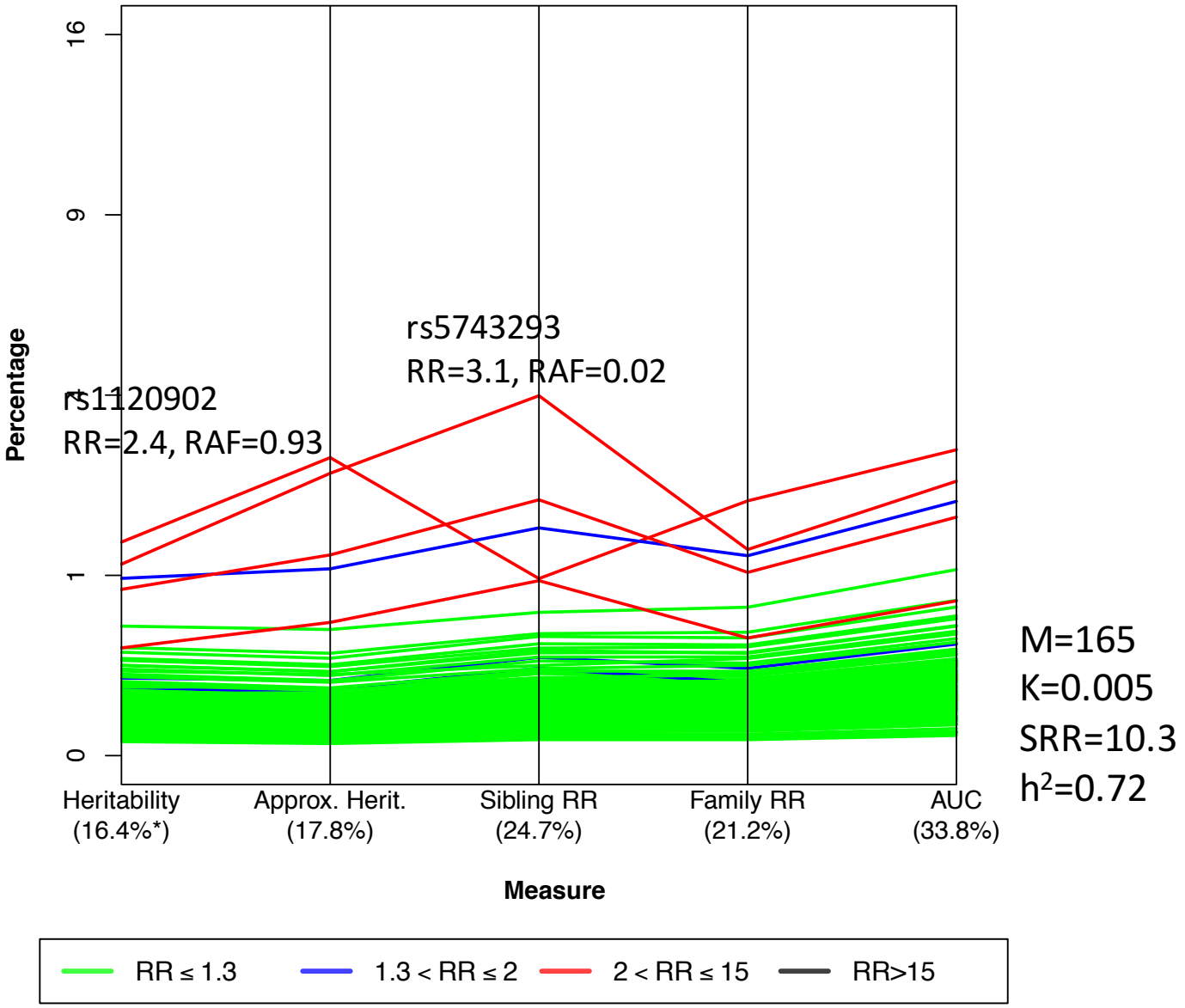
- Explore how these measures can imply different impacts of genetic variants on disease.
- Calculate them across studies of:
  a) Breast cancer
  b) Crohn's disease
  c) Rheumatoid arthritis
  d) Schizophrenia

Results: Breast Cancer



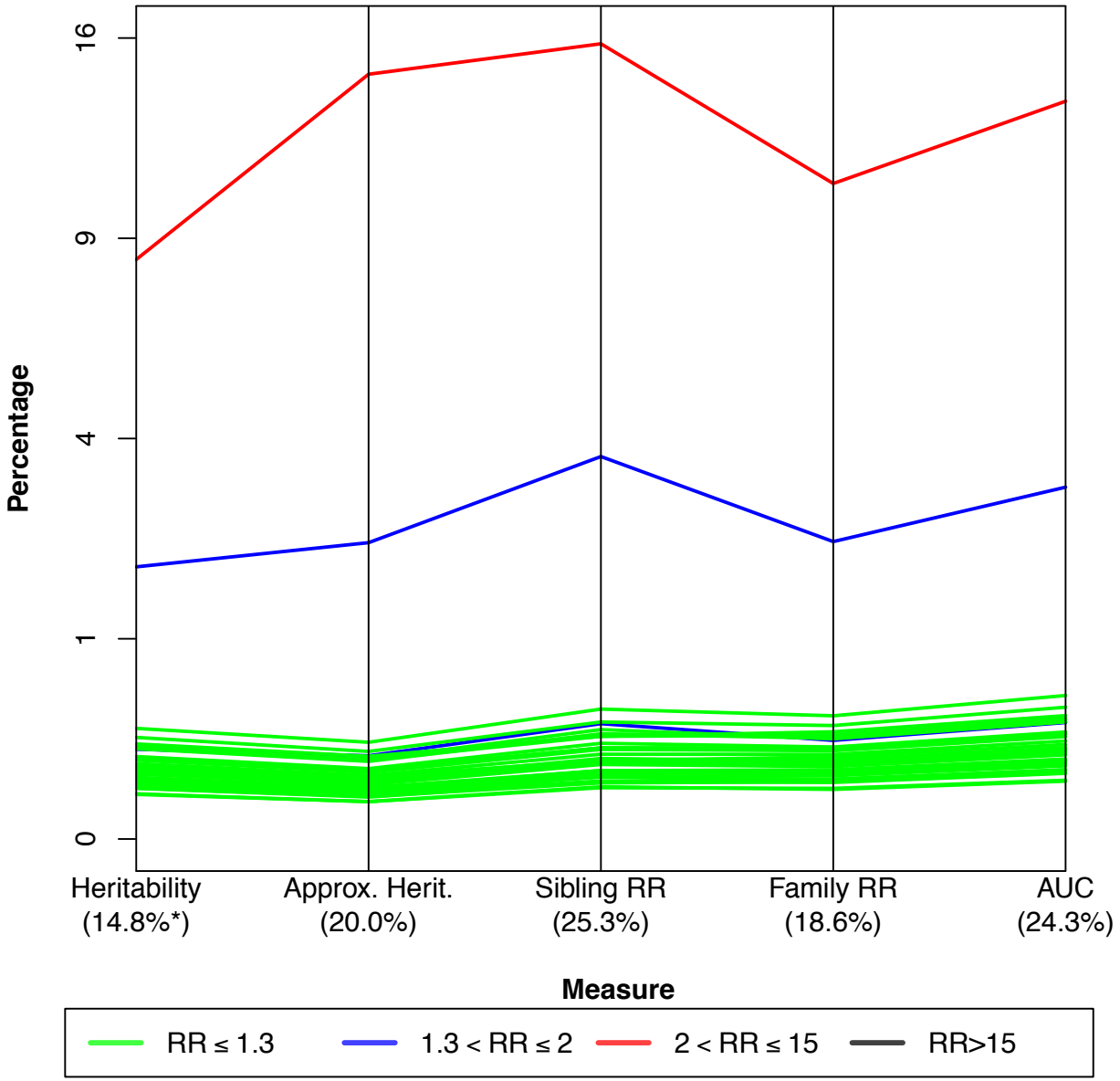
M=65
K=0.12
SRR=2
h^2=0.6

Results: Crohn's Disease



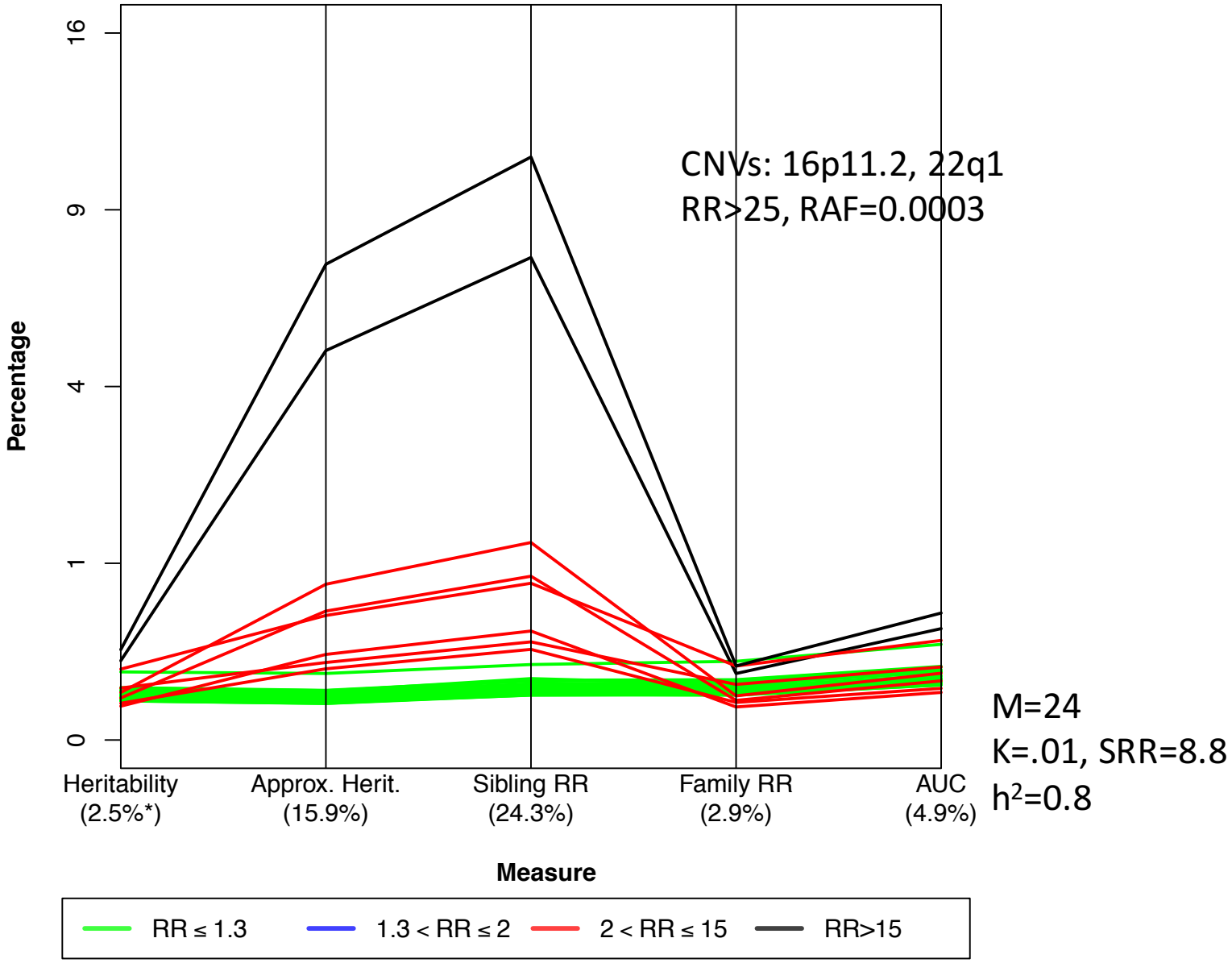
Results: RA

s6910071, HLA-DRB1E
RR=2.88, RAF=0.22



M=36
K=.01, SRR=6
h^2=0.63

Results: Schizophrenia



What goes into Denominator?

- All measures considered here require specification of a denominator.
- The apparent impact of genetic variants can hinge on the baseline or overall risks.
- Undertake probabilistic sensitivity analyses to explore how results vary across risks.
- Final results in terms of benchmarking, not exact estimates.

Population Attributable Fraction

- Proportion by which disease reduced in a population if exposure to a risk factor(s) was reduced or removed.

$$PAF = \frac{K - k_{bb}}{K} = 1 - \frac{k_{bb}}{K}$$

$$PAF = \frac{2p(1-p)(RR_{Bb} - 1) + p^2(RR_{BB} - 1)}{1 + 2p(1-p)(RR_{Bb} - 1) + p^2(RR_{BB} - 1)}$$

- For multiple variants:

$$PAF_{Total} = 1 - \prod_i (1 - PAF_i)$$

Example of PAF

Nature Genetics **32**, 581 - 583 (2002)
Published online: 4 November 2002 | doi:10.1038/ng1021

RNASEL Arg462Gln variant is implicated in up to 13% of prostate cancer cases

Graham Casey[1], Phillippa J. Neville[1], Sarah J. Plummer[1], Ying Xiang[1], Lisa M. Krumroy[1], Eric A. Klein[2], William J. Catalona[3], Nina Nupponen[4], John D. Carpten[4], Jeffrey M. Trent[4], Robert H. Silverman[1] & John S. Witte[5]

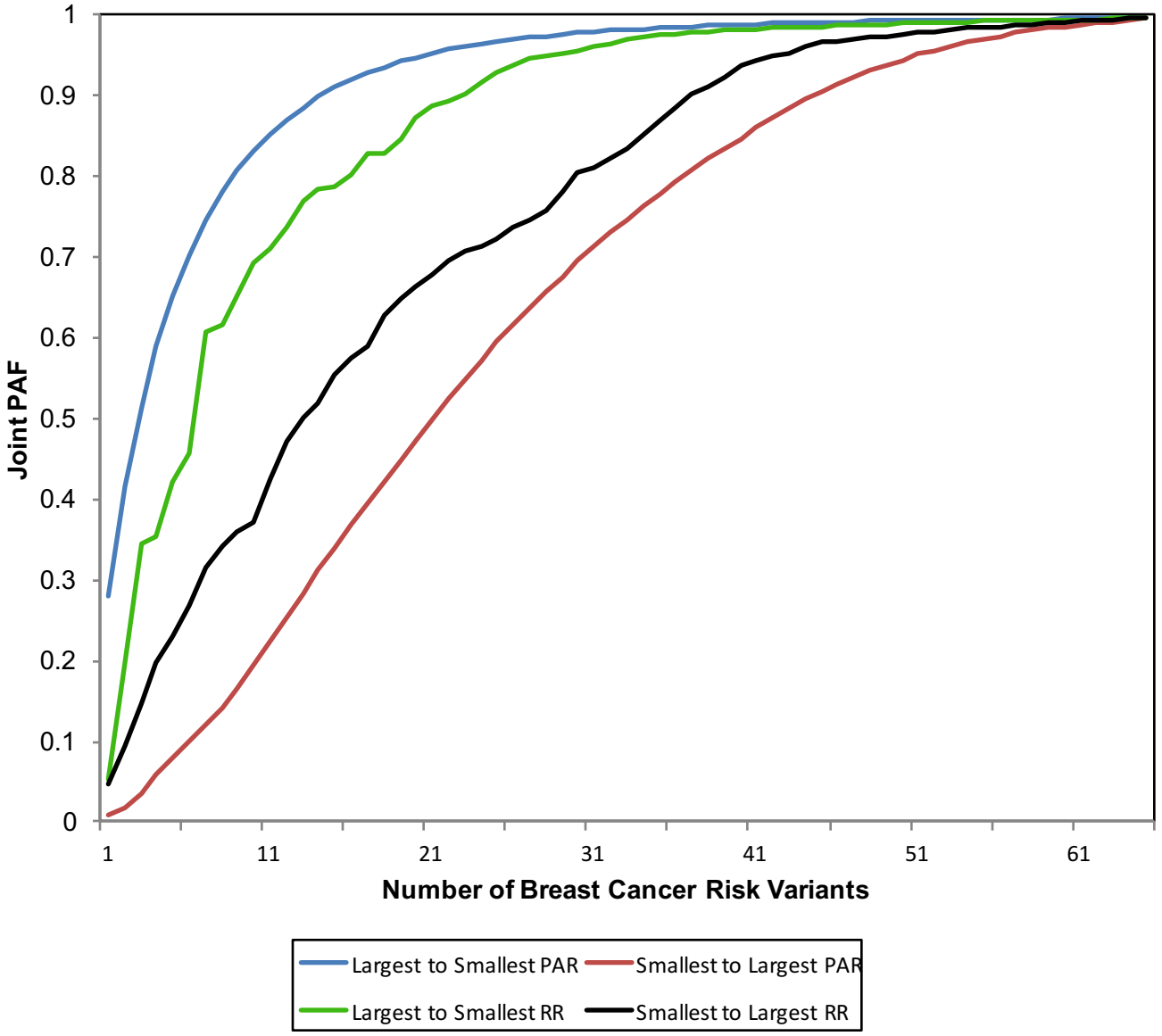
Population Attributable Fraction

- ~ Order of magnitude larger than other measures.
- As RAF > 0.50, PAF only measure that increases.
- When RR and RAF get large, single variant PAF approaches 100%.
- Examples:
  - Breast cancer variant (rs10771399, RR=1.2, RAF = 0.90) PAF=28%
  - Schizophrenia rare variant (CNV at 16p11.2, RR=26, RAF = 0.0003) PAF =1.4%
  - Combined PAF > 90% (=100% with ½ Crohn's variants)

Computational Anomaly in PAF

- Apparent impact of each additional risk variant depends on which variants have already been incorporated.
- E.g., assume two genetic variants for a disease:
  - each with individual PAF=0.50
  - combined PAF = 0.75 (=1-(1-0.5)^2).
- Remove 1 variant disease by ½.
- Remove 2nd disease by ½ in remaining popln. Or by ¼ in original population.

PAF curve Depends on SNP Order



Another Issue with PAF…

- Combined PAF not analogous to that obtained by removing an environmental exposure (smoking).
- As the number of known risk loci continues to increase, essentially everyone in the population will carry a number of risk alleles.
- Then any preventative treatment directed at countering the risk loci would have to be applied to the entire population, which seems very unrealistic.

Take Home…

- For common and rare variants of varying penetrance, use heritability explained or the proportion of genetic risk on a log-scale.
- Avoid approximation to the heritability and sibling relative risk because they break down for rare, high-penetrance variants (vastly inflated estimates).
- Issues with AUC, and PAF has a number of undesirable properties.

Outline

1. Measures of Variation Explained
2. Multivariate Analysis

2. Multivariate Analysis

- Single Locus Analysis

$$\text{logit}(P(D \mid G)) = \beta_0 + G_l \beta_l, \quad l = 1, \ldots, m$$

- Multiple Loci

$$\text{logit}(P(D \mid G)) = \beta_0 + G_1 \beta_1 + \ldots + G_m \beta_m$$

Hierarchical Model

$$\text{logit}(P(D \mid G)) = \beta_0 + G_1\beta_1 + \ldots + G_m\beta_m$$

$$\underline{\beta} = \mathbf{Z}\underline{\alpha} + \underline{\delta}$$

$$\underline{\delta} \sim MVN(0, \mathbf{T}), \mathbf{T} = \underline{\tau}^2\mathbf{I}$$

Efron & Morris, 1974
Witte, 1996
Conti and Witte, 2003
Chen and Witte, 2007

Posterior Estimates

- Weighted to reflect precision of ML and prior estimates

$$\underline{\tilde{\beta}} = \mathbf{BZ}\underline{\tilde{\alpha}} + (\mathbf{I} - \mathbf{B})\underline{\hat{\beta}}$$

$$\text{where } \underline{\tilde{\alpha}} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}\underline{\hat{\beta}}$$

$$\text{and } \mathbf{B} = \frac{\mathbf{V}}{\mathbf{V} + \mathbf{T}}, \mathbf{W} = (\mathbf{V} + \mathbf{T})^{-1}$$

Incorporating Additional Info?

- Part of a known pathway?
- Within linkage \ association regions?
- Potentially functional?
- Degree of conservation?
- Tagging other SNPs?
- Copy number polymorphism?

Z matrix

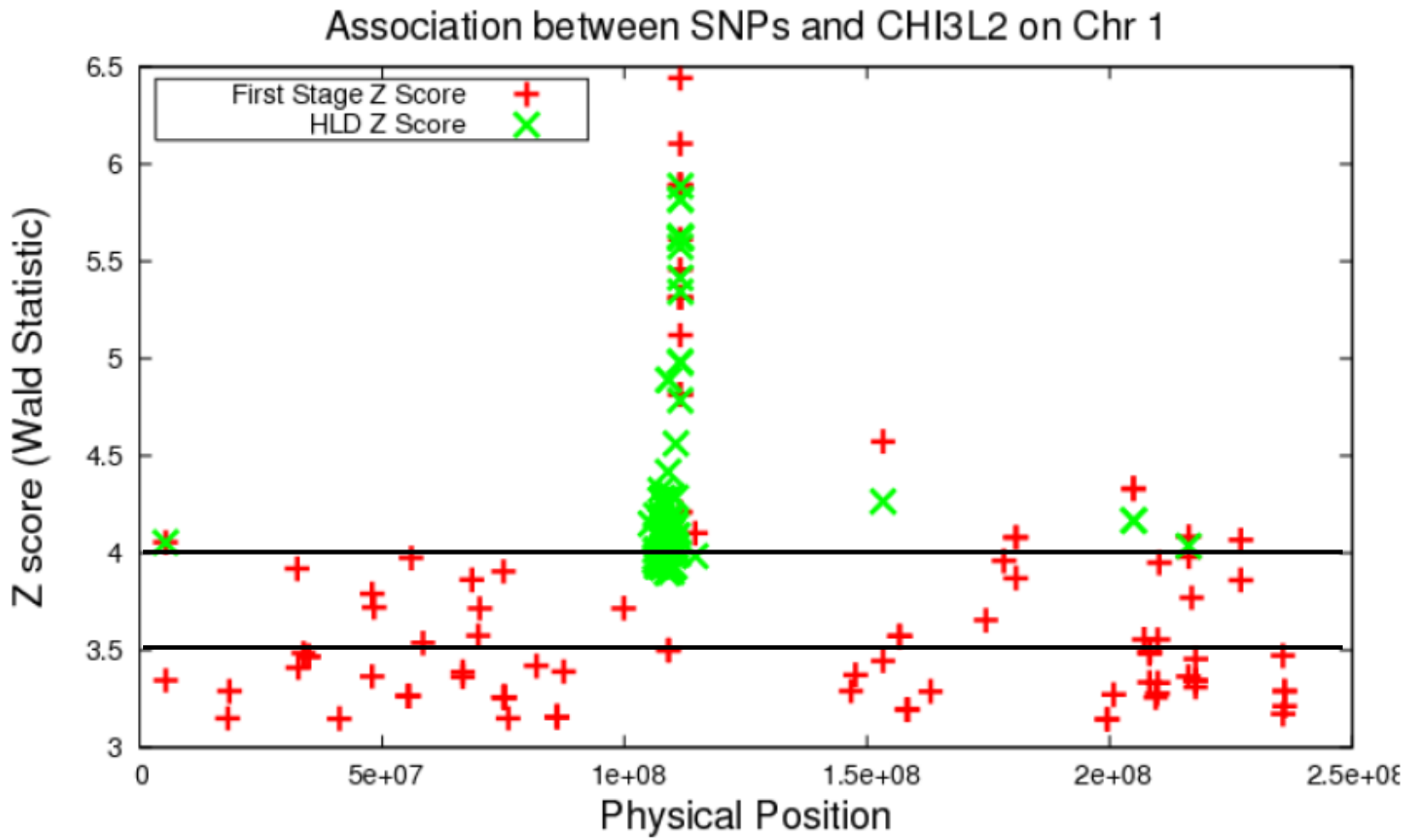
			Functional Category					LD sum columns						
SNP	**connectivity**	**conservation**	**mRNA UTR**	**ns coding**	**intron**	**locus**	**syn coding**	**mRNA UTR**	**ns coding**	**intron**	**locus**	**syn coding**	**conservation**	**linkage**
1	206	21	0	1	0	0	0	1	0	1	0	0	42	4.4
2	4	32	1	0	0	0	0	0	1	1	0	0	31	5.5
3	15	10	0	0	1	0	0	1	1	0	0	0	53	4.3
4	56	15	0	0	0	1	0	0	0	0	0	0	0	3
5	108	14	0	0	1	0	1	0	0	0	0	0	0	2
6	340	9	0	0	0	1	0	0	0	0	0	0	0	2
7	356	31	1	0	0	0	0	0	0	2	1	1	84	2

HM Example: SNPs and Expression

- Previous result:
  - Linkage to chromosome 1, and association between SNP in chitinase 3-like 2 (*CHI3L2*) promoter and *CHI3L2's* expression level
- Genoptypes:
  - Affy 500K data, unrelated CEPH individuals
- Prior information:
  - Linkage region (& LOD scores)
  - Functionality
  - Conservation scores
  - Number of SNPs tagged

Cheung et al., Nature 2005; 437:1365-1369.

HM Example Results

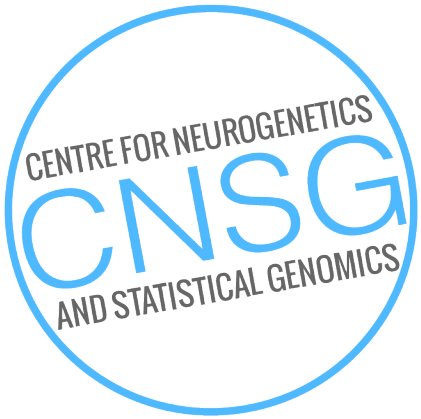


Chen & Witte, AJHG 2007

2016 Module 18:
Statistical & Quantitative Genetics of Disease

Lecture 5
Polygenic models of disease risk

Naomi Wray



Aims of Lecture 5

Theory

- To consider polygenic models of genetic risk
- To demonstrate that many polygenic models are consistent with empirical data and that they can be considered equivalent
- To understand the conclusion that the liability threshold model is the model of choice
- To understand the criticisms and controversy of the liability threshold model

Genetic models of disease

Mendelian disease:

- Individuals that possess the mutation get the disease.
- Dominant e.g Huntington's or recessive e.g. Cystic fibrosis

Mendelian disease with variable penetrance.

- Only those with the mutation get the disease
- Not everyone with the mutation gets the disease.
- E.g. C9orf72 in Motor Neurone Disease

Compound heterozygote disease.

- Like recessive Mendelian but individuals carry two different rare mutations in the same gene.

Two-hit diseases

- Hypothesized, but examples?

Oligogenic diseases –caused by presence of several genetic risk variants

Polygenic diseases – caused by multiple genetic risk variants

Multifactorial diseases- caused by multiple genetic risk variants and other risk factors

Common complex genetic diseases are likely to be polygenic multifactorial

Evidence:

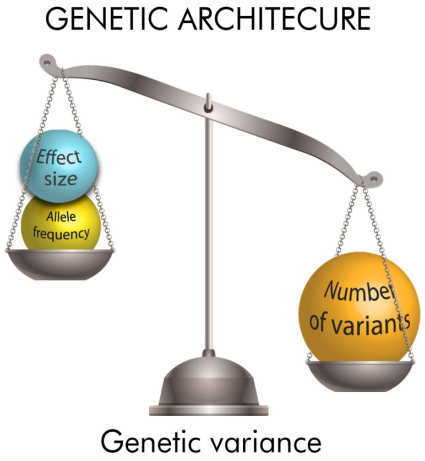
Many risk variants of small effect identified

Implications:

- We all carry risk alleles
- Each affected person may carry a unique portfolio
- Polygenic model can accommodate some people having few loci of larger effect and others having many loci of small effect
- The more loci involved, to be consistent with low prevalence, the probability of disease has to increase steeply with the number of loci.
- The more loci involved, the more likely they have a pleiotropic effect, which would be consistent with them being common in the population
- The more loci involved implies that we are highly robust to perturbations – but this breaks down when the burden of risk factors become too great

Modeling polygenic genetic risk

- "Easiest" to understand by thinking of individual risk loci and how they act together to cause disease
  - The frequency of the risk alleles
    - Drawn from a distribution
    - All the same
  - The effect size of the risk alleles
    - Drawn from a distribution
    - All the same – relative risk associated
  - Interaction between risk loci
    - Complex
    - All act in the same way



Basic Model

p = freq of risk allele 0.1
1-p = freq of non-risk allele

Assume Hardy- Weinberg equilibrium in the population
Genotype frequencies
P(bb) = $(1-p)^2$
P(Bb) = $2p(1-p)$
P(BB) = p^2

Relative risk associated with one risk allele R

n loci 100

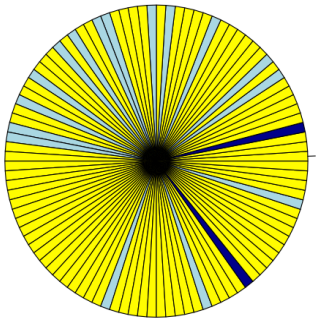
Theoretical minimum number of risk loci : 0 0
Theoretical maximum number of risk loci possible: 2n 200

Mean number of risk loci: 2np 20
Variance in number of risk loci: 2np(1-p) 18
Range in number of loci expected 2np +/- (3.5)$\sqrt{2np(1-p)}$ 5 - 36

Visualising common complex genetic diseases
Polygenic genetic architecture

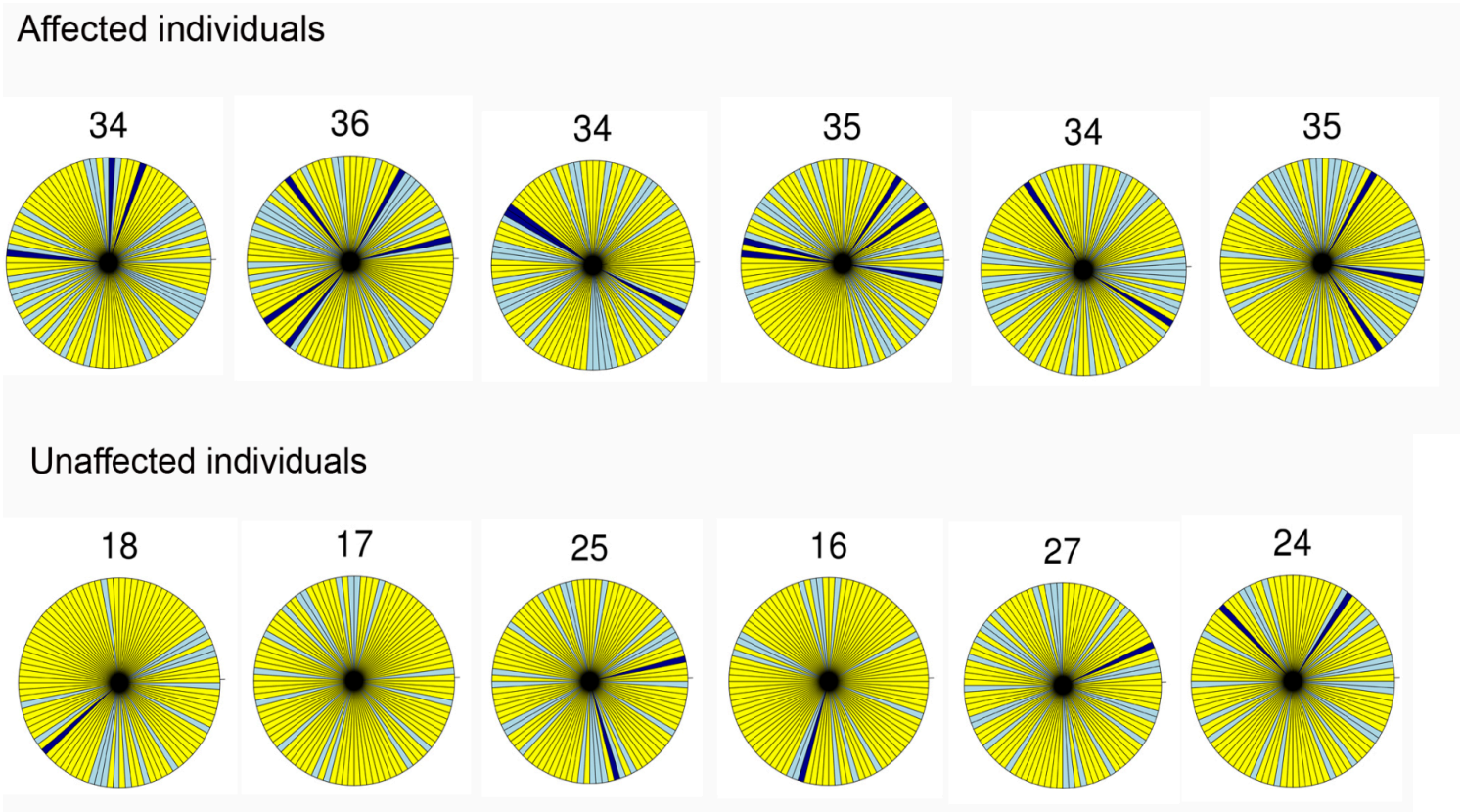
- Imagine a disorder underpinned by
  - 100 loci : 2 alleles at each locus
  - Each risk allele has frequency 0.1

20



0 risk alleles = yellow
1 risk allele = light blue
2 risk alleles = dark blue
Average person a person carries 2 alleles * 100 loci *0.1 = 20 risk alleles
Everybody carries some risk alleles
Range in population ~5-36 (mean +/- 3.5 sd)
Polygenic burden : top 1% carry > 33 risk alleles

Visualising variation between individuals for common complex genetic diseases



Not all affected individuals carry the risk allele at any particular locus
Unaffected individuals carry multiple risk loci
Consequences of risk alleles depend on the genetic and environmental background

How to combine risk loci to explain disease

- Additive on disease scale
- Multiplicative on disease scale
- Constrained multiplicative on disease scale
- Multiplicative Odds on disease scale
- Liability threshold model

Basic genetic risk model

Single locus disease model:
G = genotype; D=disease; K = overall disease risk in population

	P(G)	P(D|G)	P(D)=P(D|G)p(G)	P(G|D)=P(G)/P(D)
aa	$(1-p)^2$	f_0	$(1-p)^2 f_0$	$(1-p)^2 f_0/K$
Aa	$2p(1-p)$	$f_0 R_{Bb}$	$2p(1-p) f_0 R_{Bb}$	$2p(1-p) f_0 R_{Bb}/K$
AA	p^2	$f_0 R_{BB}$	$p^2 f_0 R_{BB}$	$p^2 f_0 R_{BB}/K$
			Sum= K	

$$P(\text{Disease})=K=f_0(1-p)^2 + f_0 R_{Bb} 2p(1-p) + f_0 R_{BB} p^2 \quad = f_0(1+p(R-1))^2$$

$$= f_0((1-p)^2 + R_{Bb} 2p(1-p) + R_{BB}\, p^2)$$

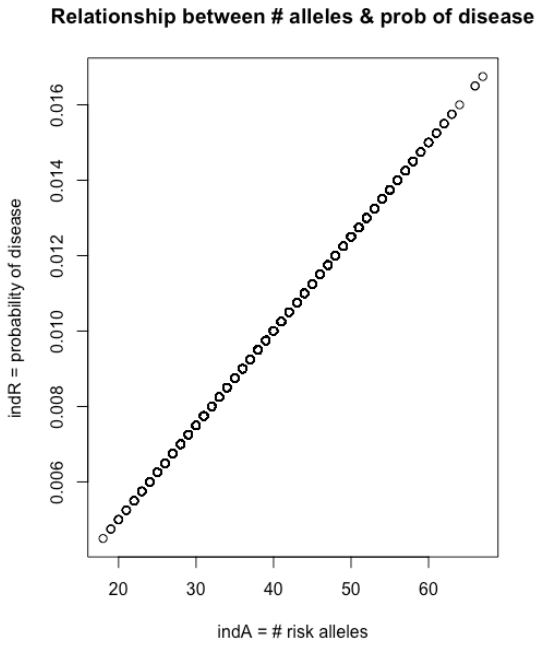
$$f_0 = K/((1-p)^2 + R_{Bb} 2p(1-p) + R_{BB}\, p^2)$$

if $R_{Bb} = R$; $R_{BB} = R^2$

$$f_0 = K/(1+p(R-1))^2$$

Additive on the disease scale

Probability of disease increases additively/linearly with the number of loci (x) carried.



$P(D|x = s) = b*R*s$

Constraint

$$\sum_{x=0}^{2n} P(D|x)P(x) = K$$

$E(P(D|x)) = E(b*R*x) = b*R*E(x) = b*R*2np = K$

So $b = K/2npR$

Looking at the additive model

Base

N = 1e5 # number of families

n = 100 # number of loci

R = 1.1 # relative risk of each risk allele

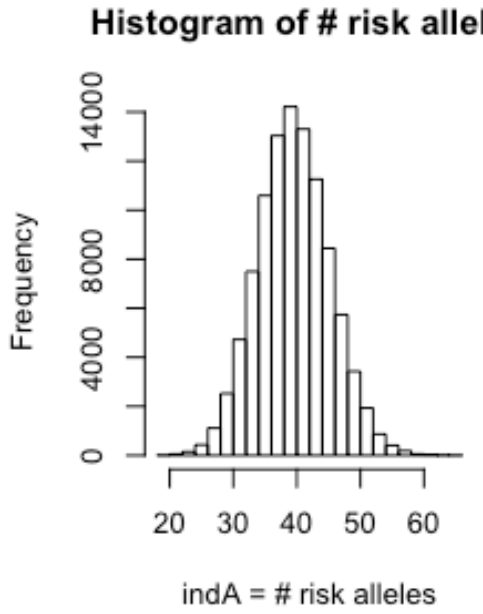
p = 0.2 # allele frequency of each risk allele

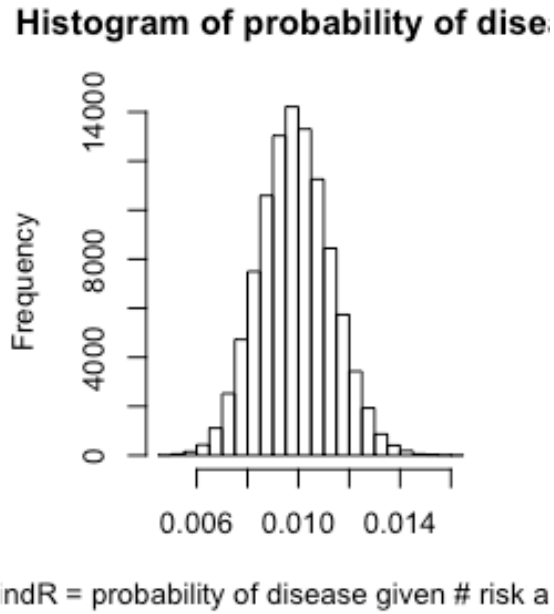
K = 0.01 # probability of disease

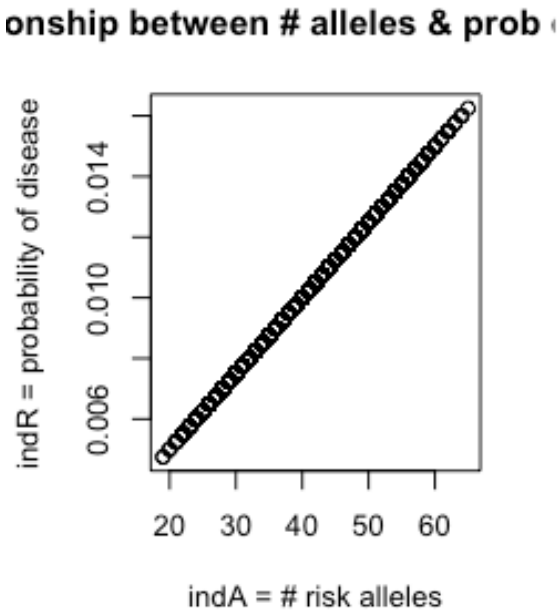
Follow up:

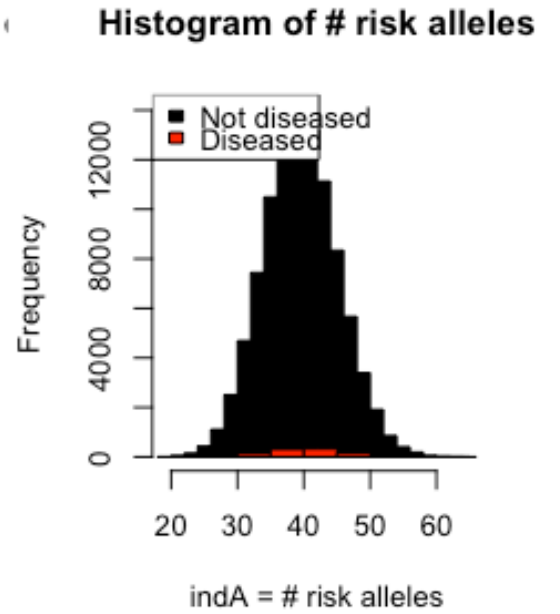
Base, R=1.5, p=0.5, K =0.1

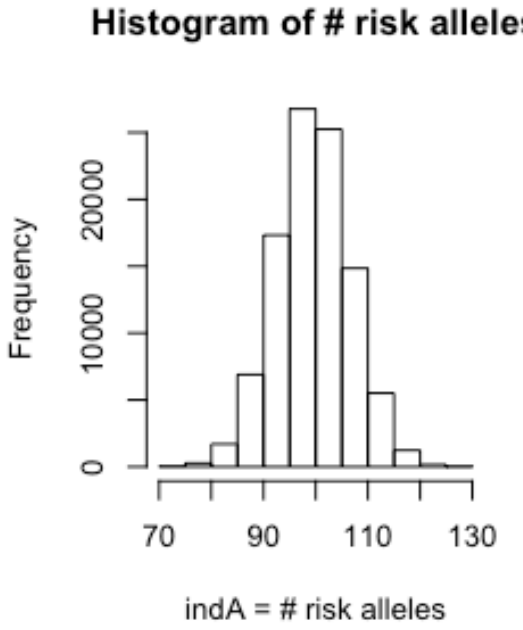
Look at maximum probability of disease and consider whether this model will generate an increased risk in relatives

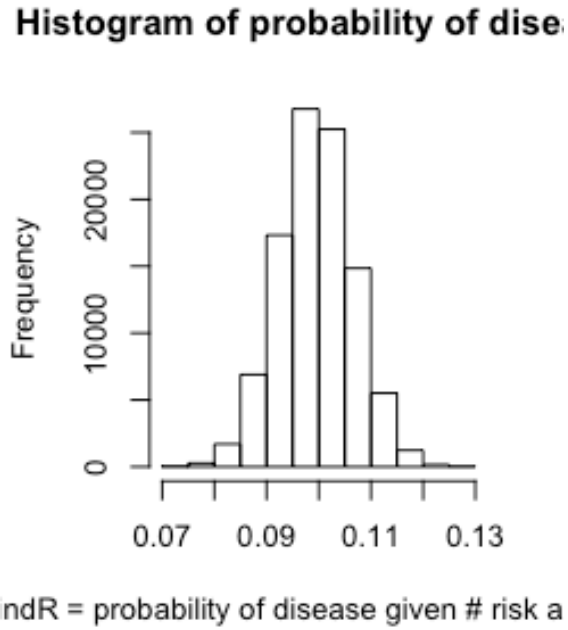

Histogram of # risk alleles
Frequency
0
4000
8000
14000
20
30
40
50
60
indA = # risk alleles

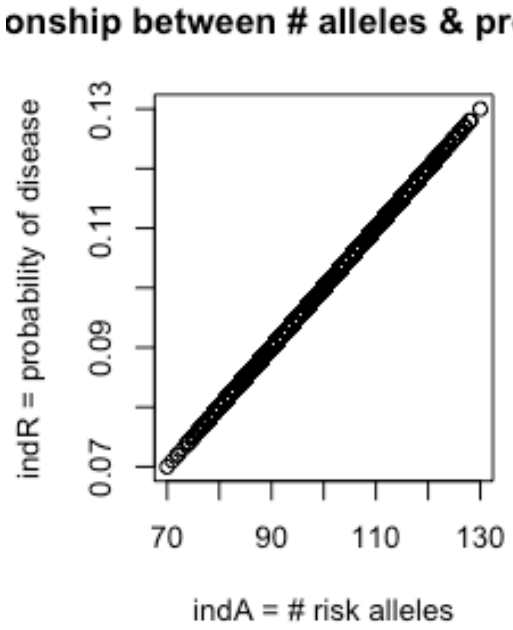

Histogram of probability of dise
Frequency
0
4000
8000
14000
0.006
0.010
0.014
ndR = probability of disease given # risk a

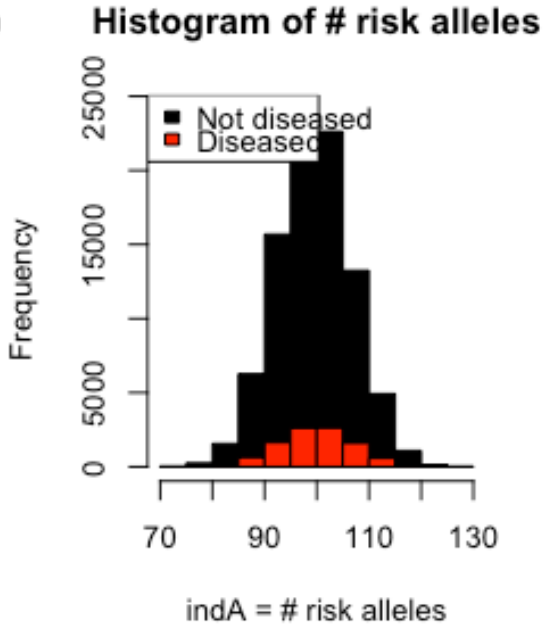

onship between # alleles & prob
indR = probability of disease
0.006
0.010
0.014
20
30
40
50
60
indA = # risk alleles


Histogram of # risk alleles
Not diseased
Diseased
Frequency
0
4000
8000
12000
20
30
40
50
60
indA = # risk alleles


Histogram of # risk alleles
Frequency
0
10000
20000
70
90
110
130
indA = # risk alleles


Histogram of probability of dise
Frequency
0
10000
20000
0.07
0.09
0.11
0.13
ndR = probability of disease given # risk a


onship between # alleles & prob
indR = probability of disease
0.07
0.09
0.11
0.13
70
90
110
130
indA = # risk alleles


Histogram of # risk alleles
Not diseased
Diseased
Frequency
0
5000
15000
25000
70
90
110
130
indA = # risk alleles

Additive model

- Mathematically tractable
- To achieve additivity of risk loci and correct disease prevalence, does not give high probability of disease with large number of risk loci
- Not consistent with high heritability
- Not consistent with observed risks to relatives

- Can "fudge" the additive model by saying
  - $P(D|x < n1) = 0$
  - $P(D|n1<x<n2)$ = additive with x
  - $P(D|x> n2) = 1$

  Is non-linear with x

  Not mathematically tractable

Multiplicative on the disease scale

Probability of disease increases multiplicatively with the number of risk loci (x)

$P(D|x = s) = f_0 R^s$

Multiplicative on the risk scale

When s =0, $P(D|x = 0) = f_0$

Constraint

$$\sum_{x=0}^{2n} P(D|x)P(x) = K$$

Binomial expansion

$E(P(D|x)) = E(f_0 R^s) = f_0 (pR + (1-p))^{2n} = f_0 (1 + p(R-1)p)^{2n} = K$

$f_0 = K/(1 + p(R-1)p)^{2n}$

Additive on the log risk scale

$\text{Log}(P(D|x=s)) = s \log(f_0 R)$

Looking at the multiplicative model

Base

N = 1e5 # number of families
n = 100 # number of loci
R = 1.1 # relative risk of each risk allele
p = 0.2 # allele frequency of each risk allele
K = 0.01 # probability of disease

Follow up:

Base, K=0.1

Base K = 0.1, R =1.2

Look at maximum probability of disease and consider whether this model will generate an increased risk in relatives

Add fix

Multiplicative model

- Mathematically tractable
- High probability of disease with large number of risk loci so consistent with high heritability and can be consistent with observed risks to relatives

BUT

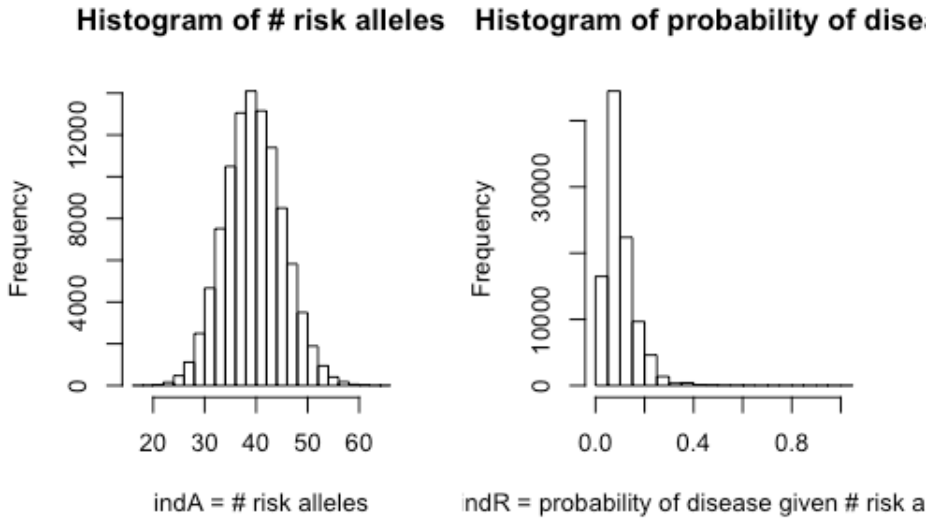
- Probability of disease for an individual can be > 1

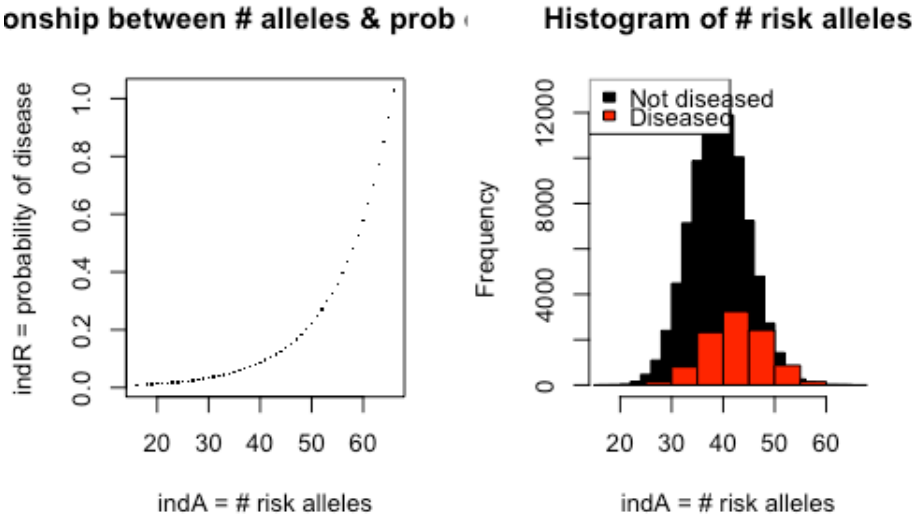
IF constrain so that max probability of disease is 1

THEN

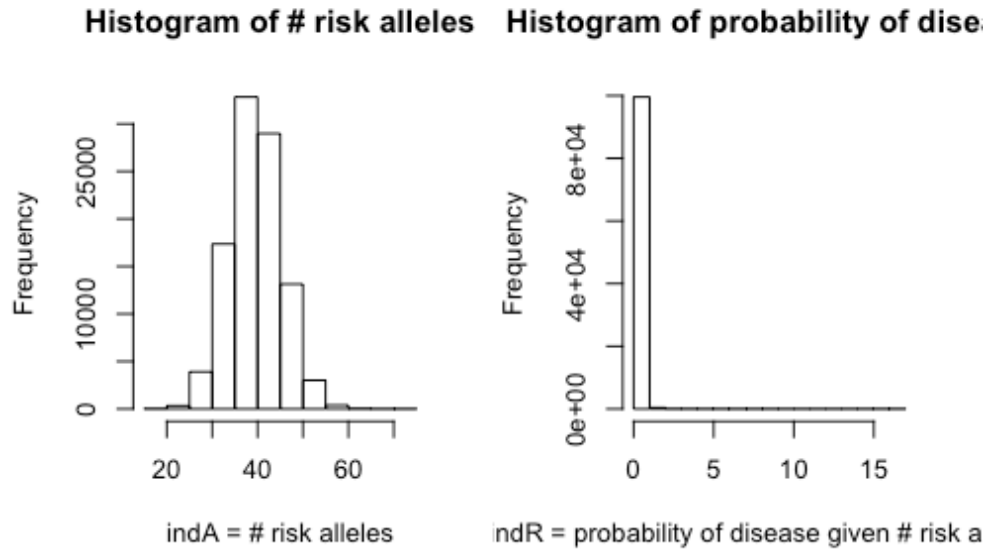
- E(P(D|x)) is no longer K
- Need to fudge to retain this property
- Loses mathematical tractability

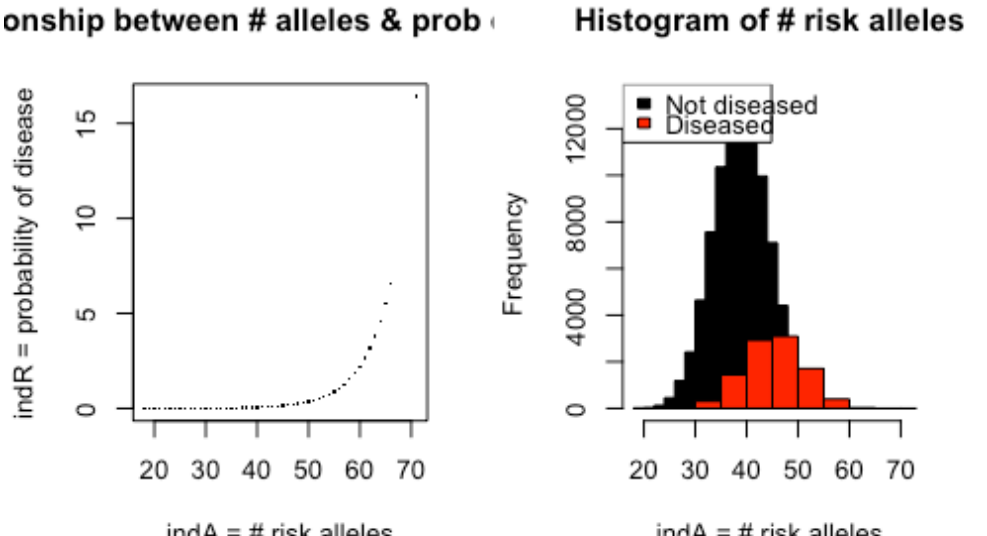
K=0.1, p=0.2, R=1.1


Histogram of # risk alleles
Frequency
0
4000
8000
12000
20
30
40
50
60
indA = # risk alleles
Histogram of probability of dise
Frequency
0
10000
30000
0.0
0.4
0.8
ndR = probability of disease given # risk a


onship between # alleles & prob
indR = probability of disease
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
indA = # risk alleles
Histogram of # risk alleles
Not diseased
Diseased
Frequency
0
4000
8000
12000
20
30
40
50
60
indA = # risk alleles

K=0.1, p=0.2, R=1.2


Histogram of # risk alleles
Frequency
0
10000
25000
20
40
60
indA = # risk alleles
Histogram of probability of dise
Frequency
0e+00
4e+04
8e+04
0
5
10
15
ndR = probability of disease given # risk a


onship between # alleles & prob
indR = probability of disease
0
5
10
15
20
30
40
50
60
70
indA = # risk alleles
Histogram of # risk alleles
Not diseased
Diseased
Frequency
0
4000
8000
12000
20
30
40
50
60
70
indA = # risk alleles

Epidemiology risk model

$Odds(Disease) = P(Disease)/(1-P(Disease))$

$Odds(Disease \mid x = s) = Odds(Disease \mid x = 0)\gamma^x = C\gamma^x$

s = number of risk loci carried by an individuals

γ = odds ratio for each risk locus

$P(Disease \mid x = s) = C\gamma^s/(1-C\gamma^s)$

Good: probability of disease does not exceed 1
Bad: mathematically intractable

Janssen et al (2006) Predictive testing for complex diseases using multiple genes: Fact or fiction? Genet Med 8 395
Lu & Elston (2008) Using the optimal ROC to design a predictive test, exemplified with Type 2 Diabetes AJHG 82

Epidemiology risk modelling

- R = risk = probability of disease
- logR = y ~N($\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$)
- R ~ LogNormal($\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$) = LN($\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$)
- $\boldsymbol{\mu}$ is arbitrary but Pharaoh set as $\boldsymbol{\mu} = -\boldsymbol{\sigma}^2/2$, but can also be calculated from disease prevalence *K*

$$\sigma^2 = \log(\lambda_{MZ}) = 2\log(\lambda_{sib})$$

$$\mu = logK - \sigma^2/2$$

Pharaoh et al (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nature Genetics
Sieh et al (2014) The role of genome sequencing in personalised breast cancer prevention. Cancer Epi Biom & Prev

Epidemiology risk model

$$E[R] = K = \int e^{\mu+\sigma x}\phi(x)dx \; = \cdots = \; e^{\mu+\sigma^2/2}$$

Two relatives, with risk of disease R_1R_2

$$R_1 = \; e^{\mu+\sigma z_1}$$
$$R_2 = \; e^{\mu+\rho\sigma z_1+\sqrt{(1-\rho^2)}\sigma z_2}$$

Probability that both are affected R_1R_2

$$E[R_1R_2] = \int R_1\phi(z_1)R_2\phi(z_2)dz_1 \; dz_2 = \cdots = e^{2\mu+\sigma^2(1+\rho)}$$

Recurrence risk = $\lambda_{relative} = \frac{E[R_1R_2]}{K^2} = e^{\rho\sigma^2}$

$$\lambda_{MZ} = \; e^{\sigma^2}$$

$$\lambda_{sib} = \; e^{0.5\sigma^2}$$

$$\sigma^2 = \log(\lambda_{MZ}) = \; 2\log\,(\lambda_{sib})$$

$$\mu = logK - \sigma^2/2$$

See thesis from Luke Jostins ftp://ftp.sanger.ac.uk/pub/resources/theses/lj4/thesis.pdf chapter 2 – contains typos

Liability threshold model

Doesn't parameterise in terms of number of risk loci

Only parameterises in terms of

- prevalence of disease and heritability of liability

OR

- prevalence of disease and risk to relatives

i.e.

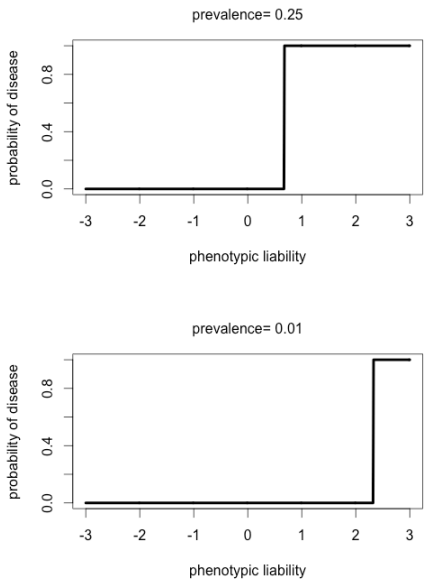
- In terms of total variance explained which could cover a range of genetic architectures

Variance explained by a locus
depends on frequency (p) and effect size(a) : $2p(1-p)a^2$

Variance explained is the same for
p= 0.1, a=0.1 as for p= 0.5, a=0.06

- BUT is the liability threshold model realistic?

Controversy – the abrupt threshold is not biological



"Contrary to the argument regarding the conservatism of the multifactorial threshold model for describing the inheritance of congenital malformations, little biological insight has resulted from the series of tautological, albeit grandiose, mathematical assumptions currently comprising the basis for this hypothesis." Melnick & Shields

The theoretical foundation of genome-wide association studies

GWAS are founded on the polygenic model of disease liability, which itself arises from an assertion of breathtaking audacity by the godfather of quantitative genetics, DS Falconer. In an attempt to demonstrate the relevance of quantitative genetics to the study of human disease, Falconer, based on work of others before him (for example, [24]), came up with a nifty solution [25]. Even though disease states are typically all-or-nothing, and even though the actual risk of disease is clearly very discontinuously distributed in the population (being dramatically higher in relatives of affected people, for example), he claimed that it was reasonable to assume that there was something called the underlying liability to the disorder that was actually continuously distributed.

Mitchell (2012) What is complex about complex disorders Genome Biol 12: 237
Edwards(1969) Familial predisposition in man, Br Med Bull
Melnick & Shields (1976) Allelic restriction: a biologic alternative to multifactorial threshold model. The Lancet
Many references to the criticism in papers of the time eg Smith (1970)

Is the abrupt threshold non-biological?

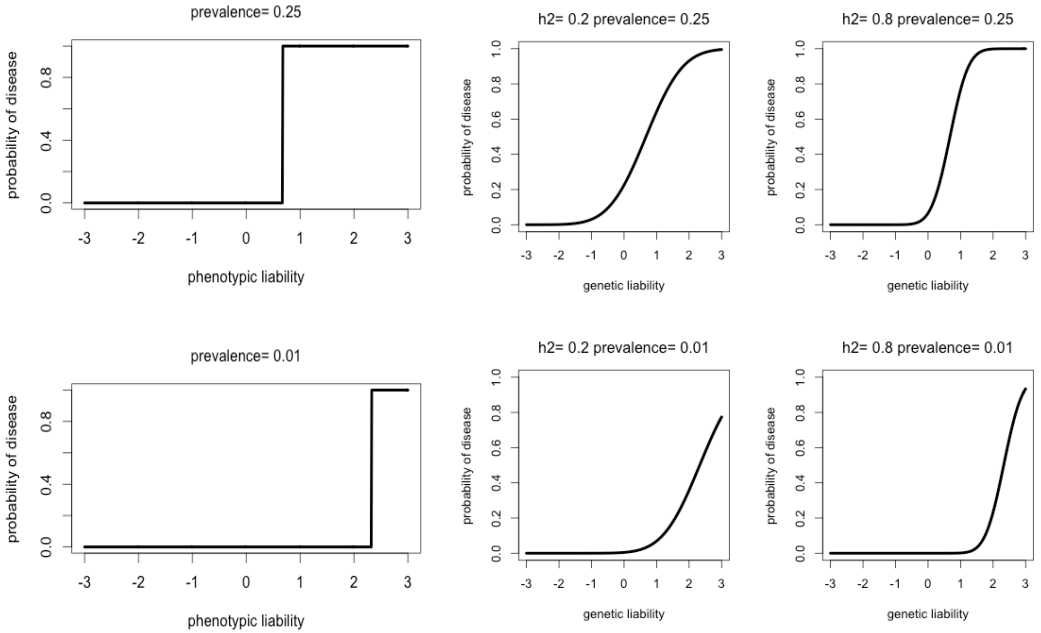
- People are classed as diseased or not disease, any error in this classifcation, contributes of a heritability of < 1.

- Wright(1934) showed that 3 vs 4 toes in guinea pigs "cannot correspond to alternate phases of a single factor (=gene)" and used crosses to show several factors ("> 3") underly a physiological threshold

- Fraser (1976) Detailed explanation of the biology consistent with a multifactorial threshold model for cleft palate

Fraser(1976) The multifactorial/Threshold concept –uses and misuses Teratology
Wright (1934) An analysis of variability in number of digits in an inbred strain of guineapig. Genetics 19 506
Wright (1934) The results of crosses between inbred strains of guinea pigs, differing in the number of. Genetics 19 537

No need to invoke abrupt threshold of phenotypic liability – instead use Probability of risk of disease under liability threshold model



"The abrupt threshold is thus conceptual rather than real and may be avoided by redefining the variance and risk function." Smith 1970

$$P(Disease|genetic\ liability = x) = \Phi\left(\frac{x - t}{\sqrt{\sigma_e^2}}\right) = \Phi\left(\frac{x - t}{\sqrt{1 - h^2}}\right)$$

Probit model

Two parameters: disease prevalence and heritability

Probit model can be parameterised in terms of number of risk loci

Curnow (1972) The multifactorial model for the inheritance of liabilty to disease and its implications for risk to relatives. Biometrics

Curnow & Smith (1975) Multifactorial models for familial diseases in man. J Royal Stat Soc A 138

Controversy – many models fit empirical data

"One cause of scepticism of the liability threshold model was the realization that the empirical data would also fit other models (Morton, '67; Smith, '71), such as a major gene combined with polygenic and environmental variation (Morton and MacLean, '74,a single locus with two alleles, each with incomplete penetrance (Reich et al., '72, or a heterogeneous mixture of cases determined either by a major locus with incomplete dominance and reduced penetrance or by environmental factors (Chung et al., '74, or various combinations of these (Elston and Stewart, '73; Lange and Elston, '75).

This is because the extreme tail of the distribution (which is all one can usually see when diseases are uncommon) are not good indicators of the shape of the main body of the distribution."

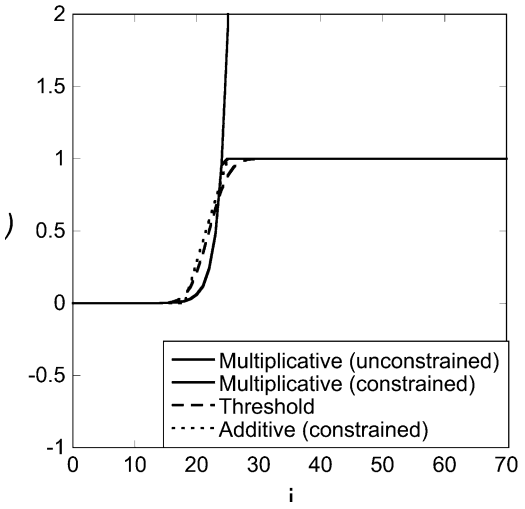
Need risk to disease from relatives of different types of relatives to start to distinguish between models
Not easy to collect, large sampling variances

Fraser(1976) The multifactorial/Threshold concept –uses and misuses Teratology

Exchangeable models of disease

- For diseases 0.5%-2%
- High heritability
- Requires there be a large variance in risk among individuals. Consequently risk considered as a function of the number of causative alleles has to be steeply increasing.

Probability of disease



Number of risk alleles(x)

Multiplicative model – standard model used but allows probability of disease to be >1.
$P(Disease)=P(Disease \mid x=0)R^x$
Constrained multiplicative model – constrain the multiplicative model to have a maximum probability of 1

"Additive" model
P(Disease)=b+xR, b=-18/7 set P(Disease)<0 to 0 and P(Disease)>1to 1

Slatkin (2008)Exchangeable models of complex inherited diseases Genetics

Which polygenic model to use?

The liability threshold model is the model of choice because

- It is the simplest parameterization that fits the observable data
- It is mathematically tractable
- It makes least assumptions about genetic architecture

"Most models are wrong some models are useful"

Practical 7

1. Additive risk model.
   a. Run code
   b. Change parameters
2. Multiplicative risk model.
   a. Run code
   b. Change parameters
3. Logistic risk model.
   a. Run code
   b. Change parameters
4. Liability threshold model
   a. Run code
   b. Change parameters

Module 19: Statistical and Quantitative Genetics of Disease: Gene-Environment Interaction

John Witte

Lecture #6: Gene-Environment Interactions

Overview

1. Conventional approaches
2. Case-only GxE
3. Empirical-Bayes case-only / case-control
4. Two-step approaches
5. Gene-sets / pathways
6. Other…

Gauderman et al., submitted 2016

Gene-Environment Interactions

- Difference in the magnitude or direction of effect of an environmental exposure on disease risk in people with different genotypes (or vice-versa).
- Effect modification
- Important because it may:
  - Identify populations with environmental exposures at increased risk.
  - Increase power and/or statistical accuracy.
  - Clarify biological mechanisms of disease risk.
  - Explain some of the missing heritability.

1. Conventional Analysis

- Assume case-control data.

  $\text{Logit}(\Pr(D=1|G,C) = \alpha_0 + \beta_G G + \beta_C C$

  – D = binary trait or disease outcome

  – G = genetic variant (e.g., SNP coded 0, 1, 2)

  – C = set of potential confounders

$\exp(\beta_G)$ = 'marginal effect' of G on D

  – averaging (or marginalizing) over the environmental (E) exposure-specific effects of G.

- E may or may not be included in **C**

Conventional GxE Model

$$\text{Logit}(P(D=1|G,E,C))=\alpha_0+ \beta_G G+ \beta_E E+ \beta_{GxE} GxE+\beta_C C$$

$\exp(\beta_G)$ = main effect of G on D (G=1, E=0)

$\exp(\beta_E)$ = main effect of E on D (G=0, E=1)

$\exp(\beta_{GxE})\ \exp(\beta_G)\ \exp(\beta_E)$ = overall effect (G=1, E=1)

- For a cohort study, use a log-linear model to estimate relative risks or a proportional hazards model to estimate hazard rate ratios if time-to disease data are available.
- For a quantitative outcome, use linear regression.

Interaction Scale

Multiplicative

Departure from multiplicative effects implies odds-ratios associated with one risk-factor varies by the level of the other risk-factor and vice-versa.

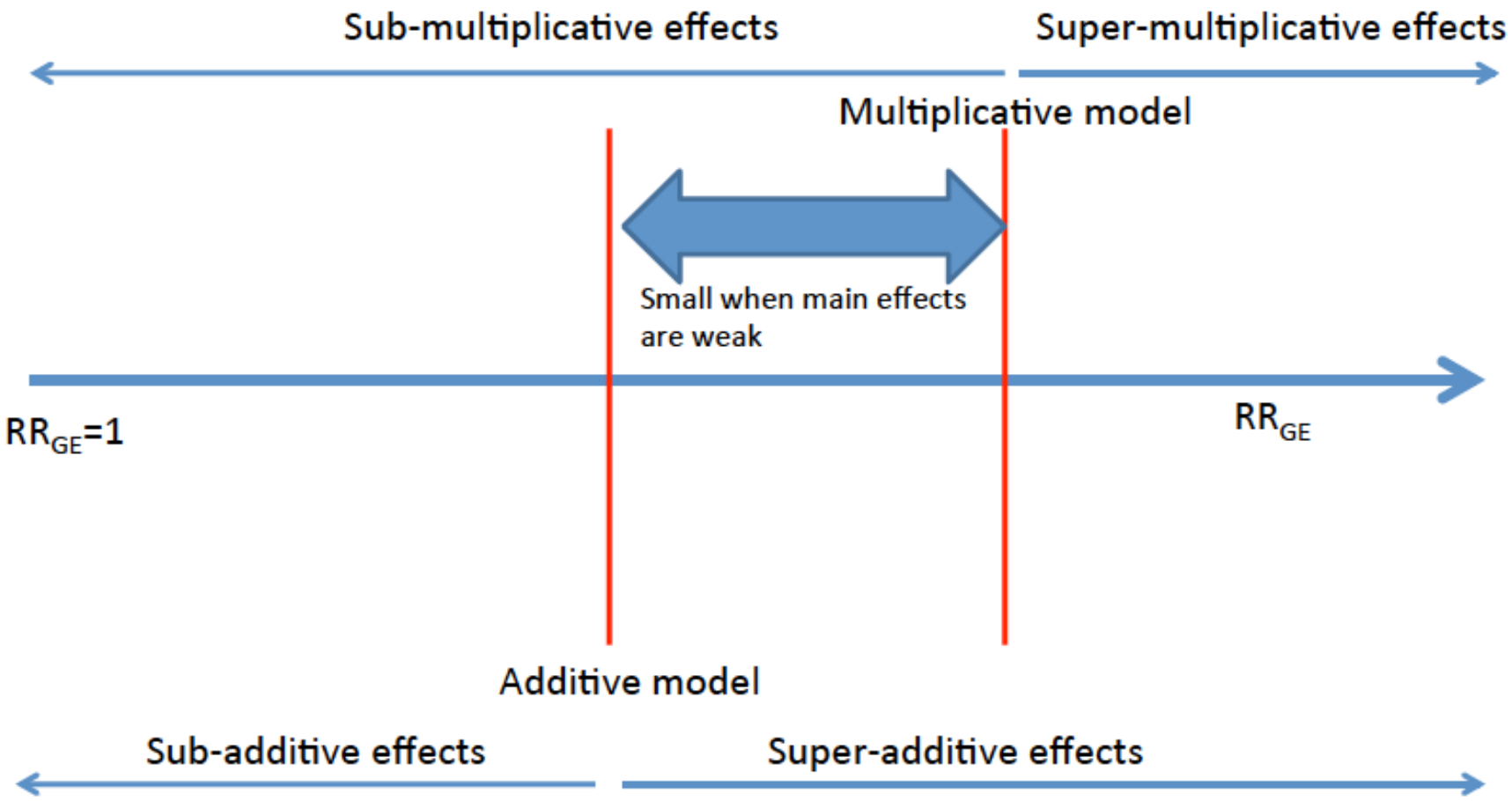
$$GxE_{Multp} = \exp(\beta_{GxE})\exp(\beta_G)\exp(\beta_E) / (\exp(\beta_G)\exp(\beta_E))$$
$$= \exp(\beta_{GxE}) = \text{interaction effect}$$

Additive

Departure from additivity implies that absolute risk-reduction associated with removal of one risk-factor depends on the levels of another and vice-versa.

$$GxE_{ADD} = \exp(\beta_{GxE})\exp(\beta_G)\exp(\beta_E) - \exp(\beta_G) - \exp(\beta_E) + 1$$

Joint effects for two risk factors



Factor V Leiden Mutations, Oral Contraceptive Use, and Venous Thrombosis

Strata	Cases	Controls
G+E+	25	2
G+E-	10	4
G-E+	84	63
G-E-	36	100
Total	155	169

OR

G+E+: 34.7

G+E-: 6.9

G-E+: 3.7

G-E-: Reference

$OR_{Interaction\ (mult)}$
$= OR_{G+E+} / OR_{G+E-}\ OR_{G-E+}$
$= 34.7 / 6.9 \times 3.7$
$= 1.4$

$OR_{Interaction\ (add)}$
$= OR_{G+E+} - OR_{G+E-} - OR_{G-E+} + 1$
$= 34.7 - 6.9 - 3.7 + 1$
$= 25.1$

Vandenbroucke et al., The Lancet 1994

Testing for Multiplicative GxE Interactions

$$\text{Logit}(P(D=1|G,E,C))=\alpha_0+ \beta_G G+ \beta_E E+ \beta_{GxE} GxE+\beta_C C$$

- 1 df test. H_0: $\beta_{GxE} = 0$.
- 2 df test. Joint null H_0: $\beta_G= \beta_{GxE} = 0$.
- 2 df often more powerful than 1 df test.

Controlling Confounding

- When testing GxE interaction, need to consider inclusion of confounders C in the model, but also G x C and E x C interactions.
- GxE interaction effects can themselves be confounded by other interactions.
- Potential Confounders: PCAs, etc.

Keller MC. *Biol Psychiatry* 2014;75(1):18-24.

Why so few GxE Interactions detected?

- Limited power.
- Challenges measuring E (both for discovery and replication).
- Model misspecification.

- A number of approaches can increase power.

2. G-E Interaction: Case-Only

Strata	Cases	Controls
G+E+	a	b
G+E-	c	d
G-E+	e	f
G-E-	g	h

Odds Ratio (OR)
ah / bg
ch / dg
eh / fg
1

$$OR_{Interaction} = OR_{G+E+} / OR_{G+E-}\ OR_{G-E+}$$
$$= ah/bg / (ch/dg)\ (eh/fg)$$
$$= (ag/ce) / (bh/df)$$
$$= ag/ce \quad \text{if no G-E assoc in controls } (bh/df = 1).$$

Piegorsch et al., 1994

Case-Only Model

$$\text{Logit}(P(G=g|E,D=1)) = \gamma_0 + \gamma_{GxE}E$$

$\exp(\gamma_{GxE})$ = GxE interaction effect

- H_0: $\gamma_{GxE} = 0$.
- Wald test asymptotically equivalent to H_0: $\beta_{GxE} = 0$ (assuming log-additive coding for g, 0,1,2).

If G-E are associated in source population, then can give high false positive rate.

Overview

1. Conventional approaches
2. Case-only GxE
3. Empirical-Bayes case-only / case-control
4. Two-step approaches
5. Gene-sets / pathways
6. Other?

3. Empirical-Bayes GxE Test

- Case-only more efficient than case-control, but can give biased results (e.g., if G-E assumption violated).
- Use EB hybrid model to combine case-control and case-only approaches (bias versus efficiency trade-off).

$$\beta_{EB} = K(\beta_{GxE}) + (1-K)\gamma_{GxE}$$

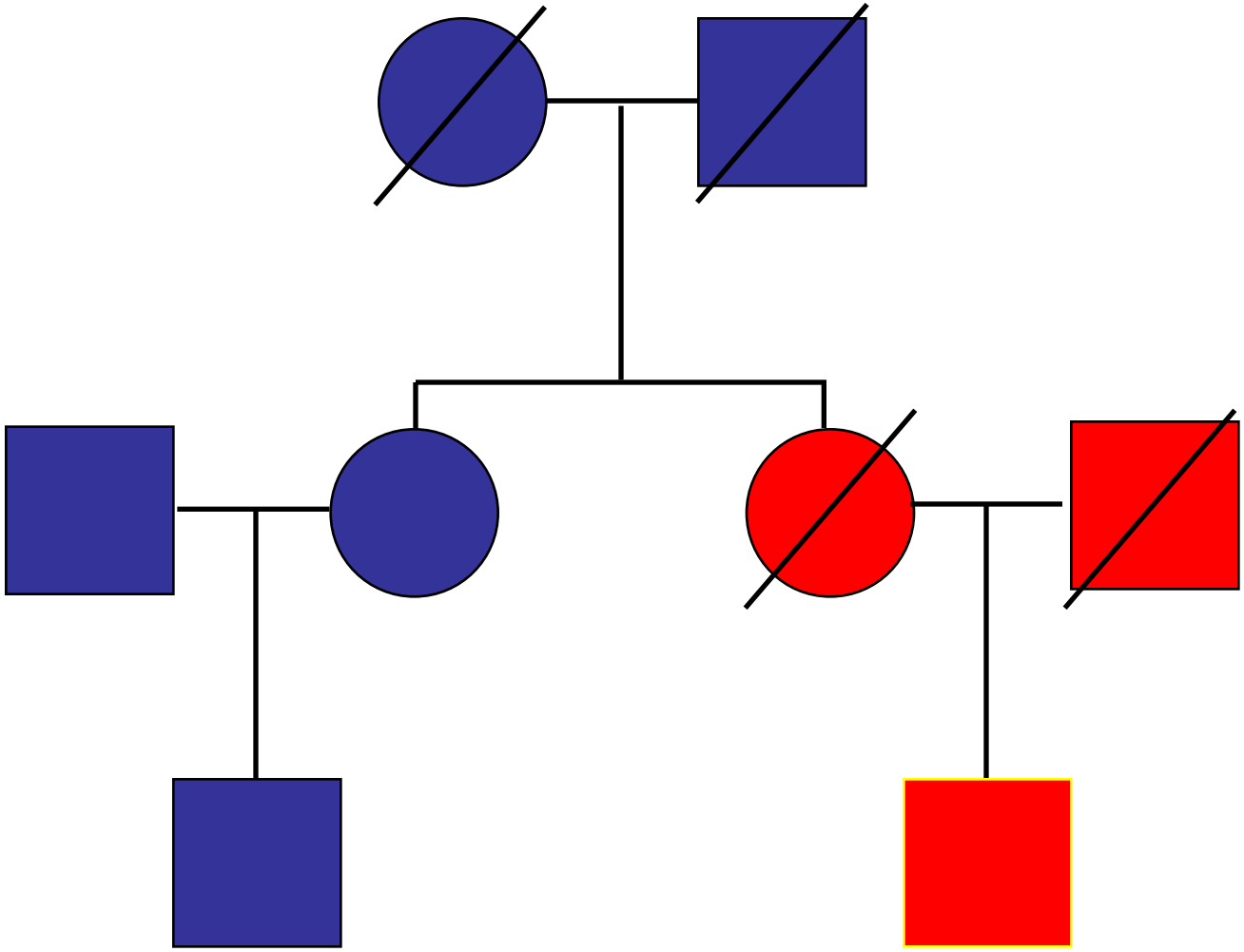
where $K=\theta_{GE}^2/(\sigma_{GxE}^2+\theta_{GE}^2)$

θ_{GE} = G-E association

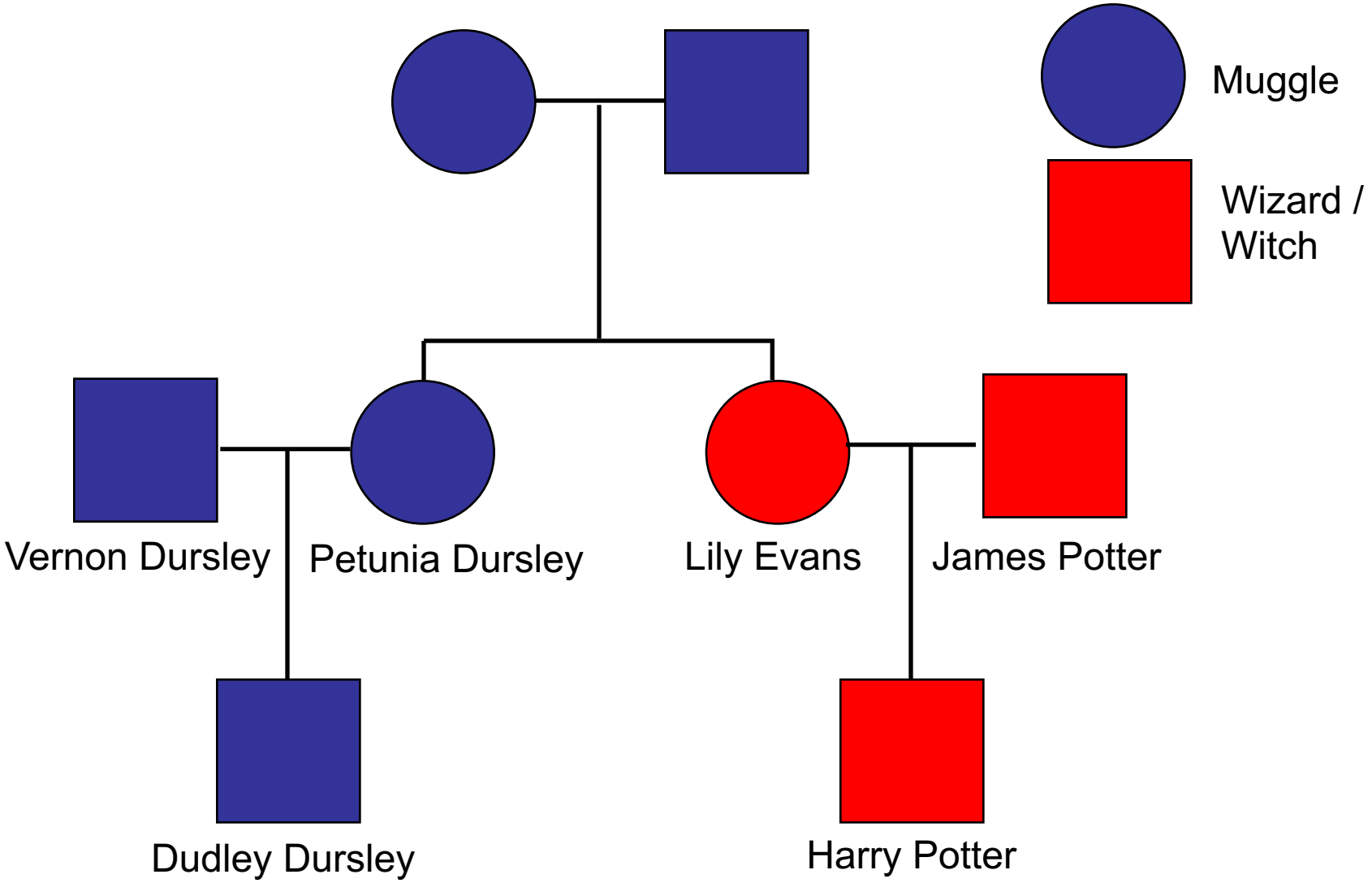
- If $\theta_{GE} \neq 0$ or if σ_{GxE}^2 is small, larger weight assigned to β_{GxE}.
- If $\theta_{GE} = 0$ (G-E independence), $\gamma_{GxE} \cong \beta_{GxE}$, use γ_{GxE} (more efficient).

- H_0: $\beta_{EB} = 0$. More power than case-control, helps control type I error from case-only.

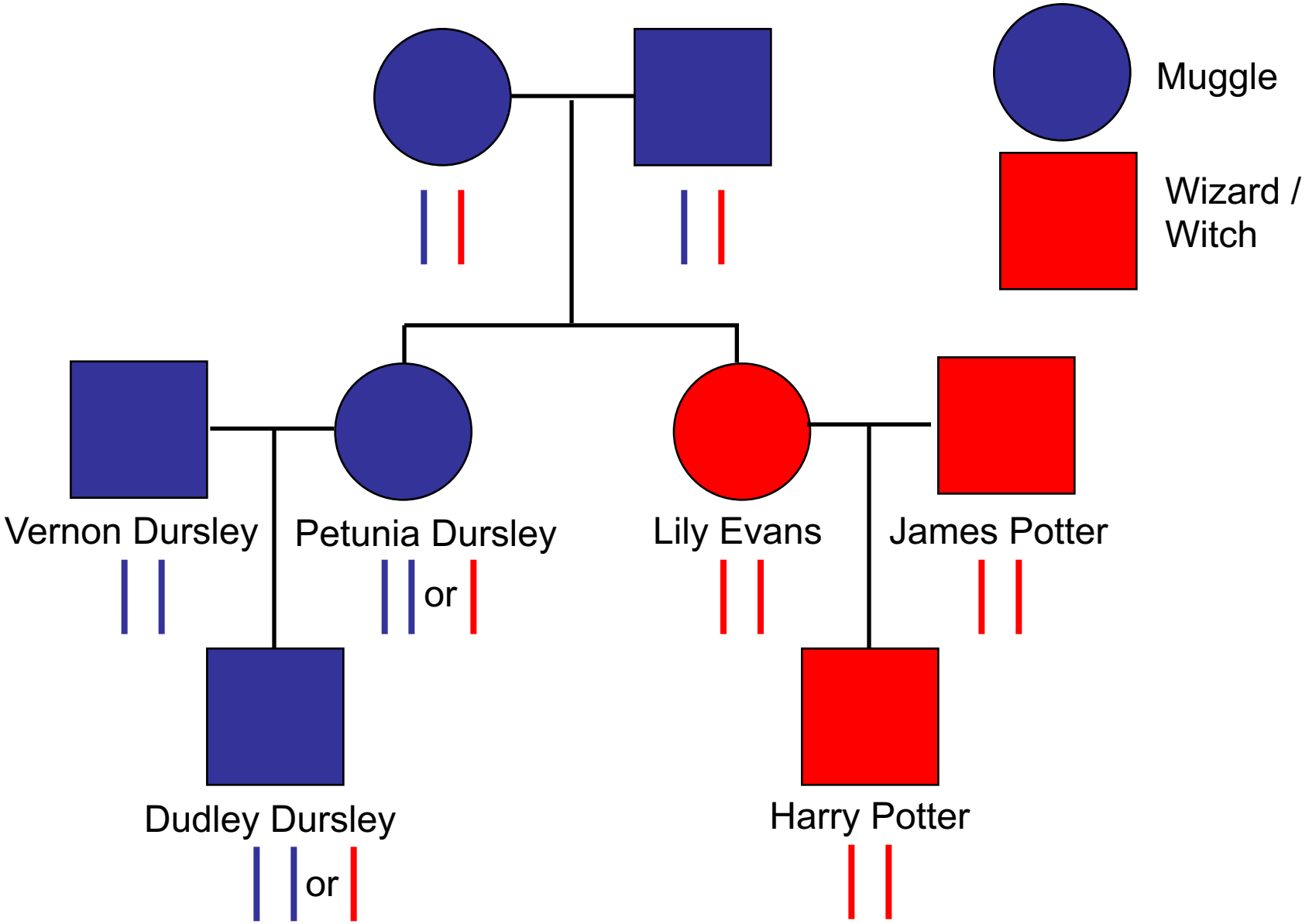
Mukherjee and Chatterjee, 2008



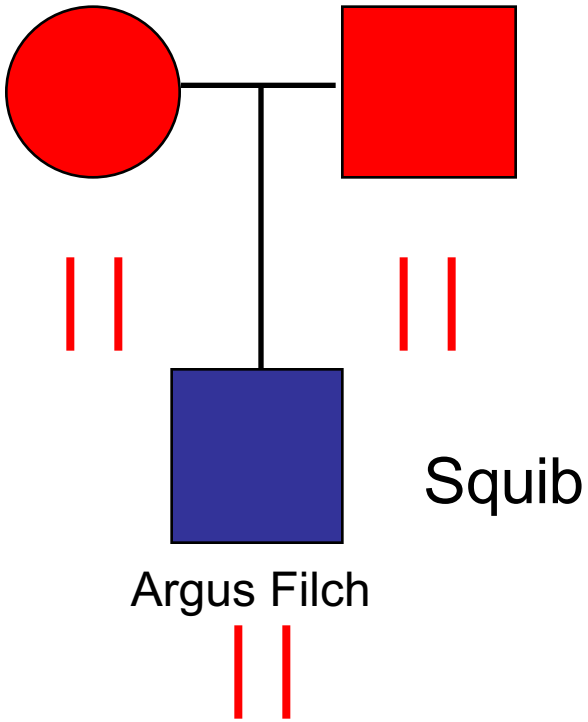
Harry Potter's Pedigree



Harry Potter's Pedigree



What About Filch?



4. Two-Step GxE Tests

- Step 1 screen: For each SNP, compute screening test statistic T_1 and corresponding p-value p_1.
- Step 2 test: Prioritize SNPs based on p_1, and conduct GxE interaction test T_2 with corresponding p-value p_2.
- Key requirement: T_1 and T_2 are independent.

Kooperberg and LeBlanc, 2008; Murcray et al., 2009; 2011; Hsu et al., 2012

Two Step GxE: Case-Control Data

Step 1:

- Test for marginal D-G association

  Logit(Pr(D=1 | G) = $l_0 + l_G G$, and/or

- Test for E-G association

  Logit(Pr(G | E) = $d_0 + d_E E$

Step 2:

- Test for GxE interaction, only using SNPs passing Step 1 threshold (fewer comparisons).
- Can use an E Bayes procedure here.
- Additional info from Step 1 increases power by up to 50% over conventional approach.

Hybrid 2-Step Approach

- Step 1: test DG and EG.
- Retain SNPs that pass at least one of these tests.

- Step 2: Apply case-control analysis and test GxE, correcting for the number of SNPs retained from step 1.

Murcray et al., 2011;

Cocktail Method

- Step 1: If p<threshold for EG, assign SNP that p.
- Else, assign SNP from DG (marginal) analysis.

- Step 2:
  - If p from DG, then test for GxE using case-only model.
  - If p from EG, then test GxE using case-control analysis.
  - Use weighted hypothesis testing.

Hsu et al., 2012

EDGxE Approach

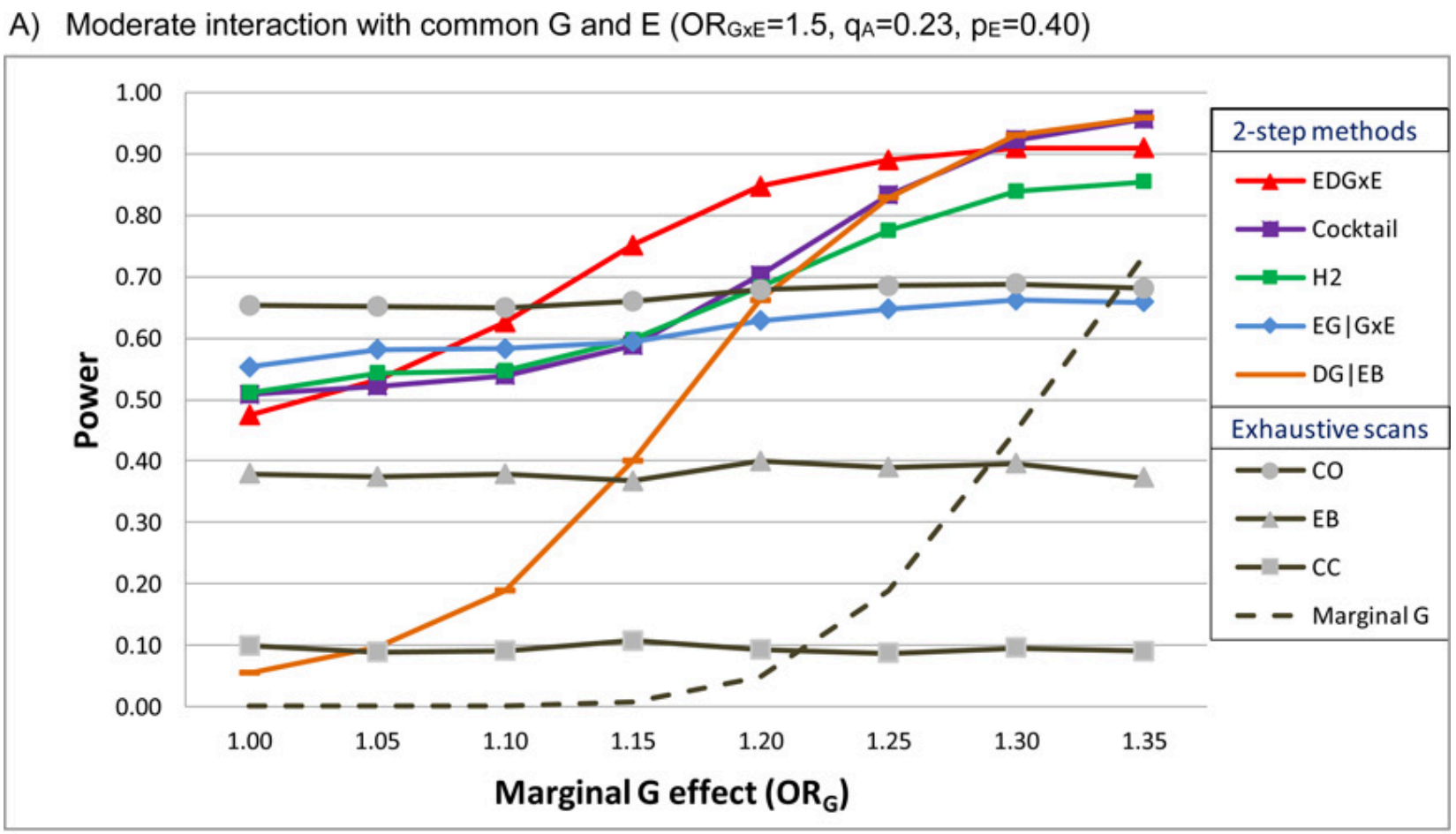
- Step 1: combines the DG and EG tests into single 2 df test.
- Step 2: weighted hypothesis testing of case-control analysis.

Gauderman et al., 2013

Power Gains for Two-step

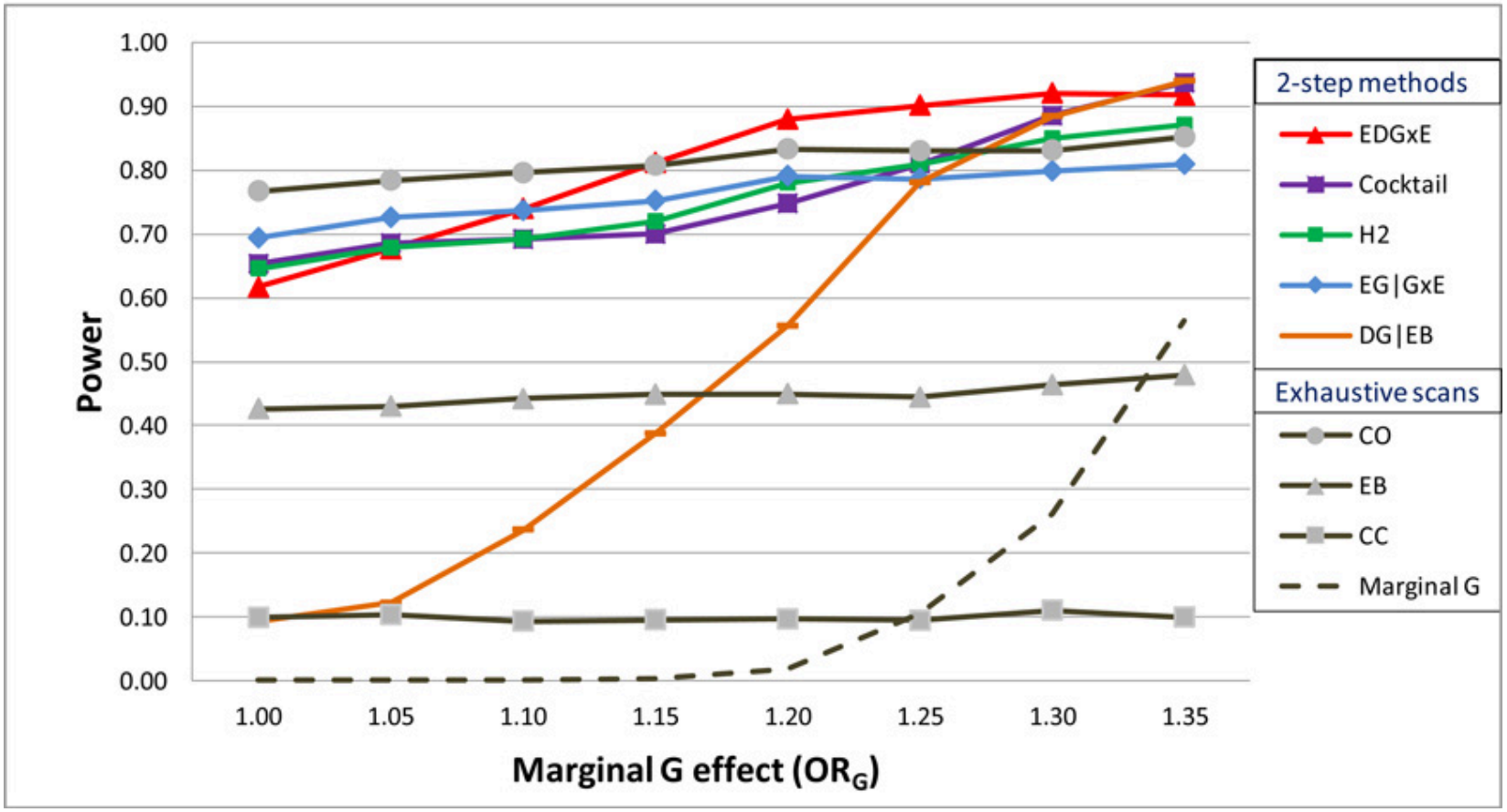
- Assume G has MAF = 0.3, and for E 30% exposed
- $\exp(b_G) = \exp(b_E) = 1.0$, $\exp(b_{GxE}) = 1.5$.
- For a GWAS, 80% power (alpha = 0.05)
- For conventional GxE model, N=10,060.
- Two-step approaches:
  - D-G screening, N=6,630
  - E-G screening, N=4,472
  - EDGE screening, N=3,994.

Comparison of GxE Tests



Gauderman et al., 2013

B) Strong interaction with less common G and E (OR_{GxE}=2.0, q_A=0.14, p_E=0.10)



Gauderman et al., 2013

Step 2: Weighted hypothesis testing

- Partition SNPs into groups, where higher ranked SNPs have less stringent alpha level.
- B most significant SNPs in step 1 tested in step 2 at significance level (α/2)/B, next 2B at (α/4)/2B, next 4B at (α/8)/4B, etc.
- Maintains overall GWAS alpha level, but uses larger alpha level for most promising interactions.

Ionita-Laza et al. 2007

Overview

1. Conventional approaches
2. Case-only GxE
3. Empirical-Bayes case-only / case-control
4. Two-step approaches
5. Gene-sets / rare variants
6. Other?

GxE for Gene Sets / Rare Variants

- Burden and variance components tests.
- Combination of burden and variance component GxE tests.
- Can incorporate GxE term into kernel.

6. Other…

- GxE interactions using Summary Stats
- Analyses stratified by E.
- Then test for differences in G main effects.

- Note: same methods can be applied to GxG interactions.

Epistasis: Gene-Gene Interactions

- Similar issues as with gene-environment interaction (e.g., multiplicative vs additive scale)
- $P(Y=1|g_1,g_2)=\beta_0 + \beta_1 X(g_1) + \beta_2 X(g_2) + \beta_{12} X(g_1)\, X(g_2)$
- Usually test when g_1 is from one gene, and g_2 from another gene (e.g., take GWAS hits)
- Feasible to do all pairwise: plink: --fast-epistasis
  - "4.5 billion two-locus tests generated from a 100K data set took just over 24 hours to run" (http://pngu.mgh.harvard.edu/~purcell/plink/)

Example: GWAS of Psoriasis

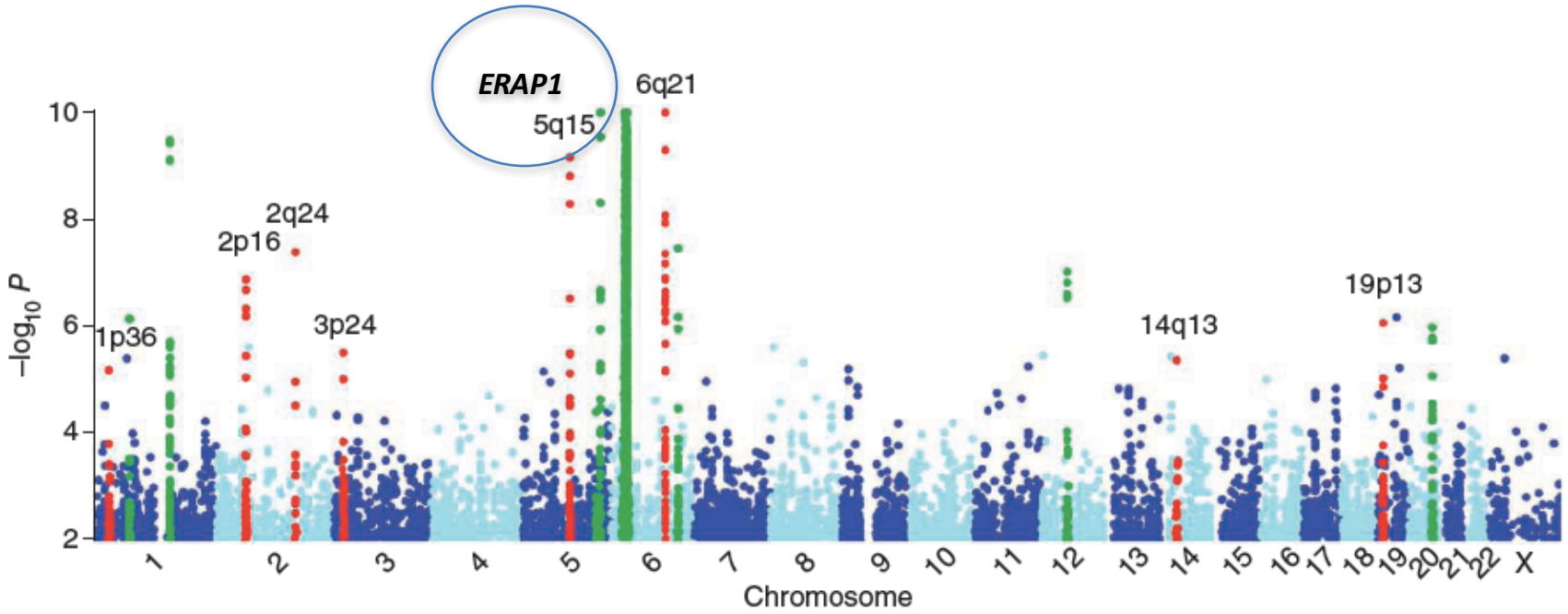


Figure 1 Plot of genome-wide association results. Genome-wide association results from 523,067 SNPs on chromosomes 1–22 and 12,408 SNPs on the X chromosome using the additive model in SNPTEST. The $-\log_{10} P$ values are thresholded at 10^{-10}. Regions in red are described in **Table 2**. Regions which have been shown previously to be associated with psoriasis and which replicated in this study are highlighted in green, as described in **Table 1**.

Take the hits, and follow up on gene-gene interaction test -- (nextslide)-->

Strange et al. Nature Genetics 2010

Example of Gene-Gene Interaction

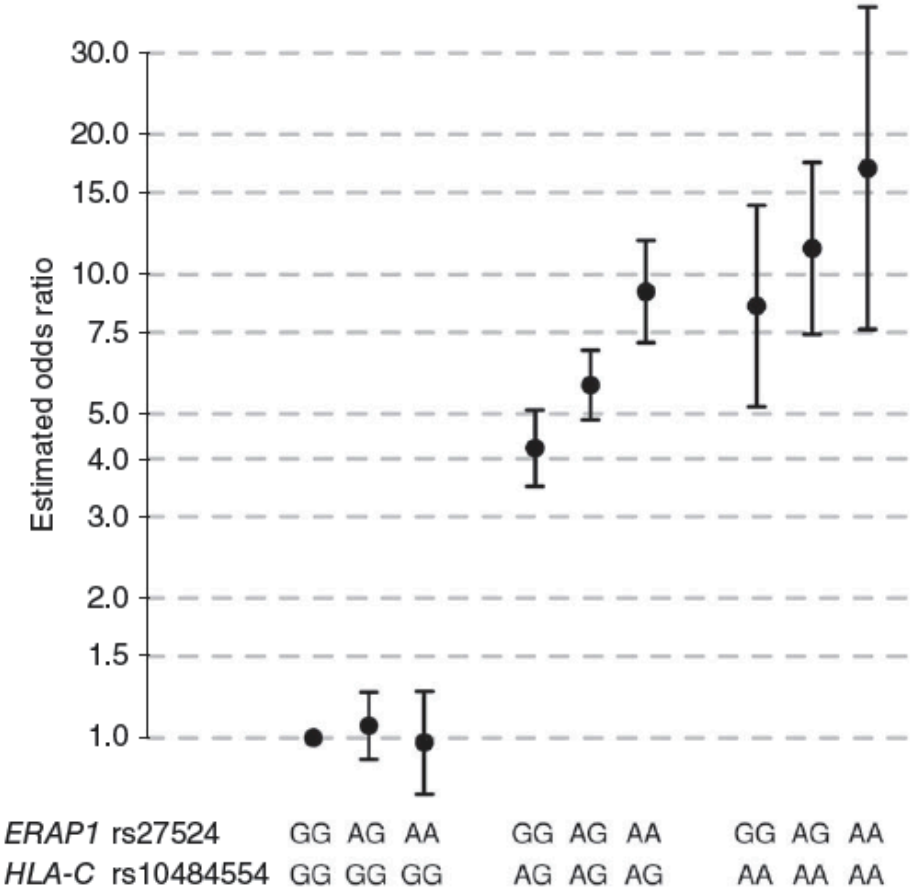


Figure 3 Statistical interaction between *ERAP1* and *HLA-C* genotypes.

Strange et al. Nature Genetics 2010

Endnote on Interactions

Challenges	Old Approach	Solutions/New Approach
Interaction can be dependent on scale	Only multiplicative scale considered	Consider evaluating interaction on both additive and multiplicative scales
SNP-based analyses can lack power	Single step analysis subject to multiple comparisons burden due to large number of SNPs considered at once	Conduct more efficient 2-step tests
	Single variant approach agnostic to biological information	Conduct gene-based/set-based tests
	Individual studies report results independently	Conduct meta-analysis across studies/cohorts
	Only homogenous populations considered, typically of European decent	Consider admixture analysis, if appropriate
Exposure measurement can be inconsistent and imperfect	Individual studies independently determine method of exposure measurement	Work towards common core of exposures and definitions
	Employ easiest measurement method for largest study sample possible	Prioritize improving precision of measurements
Software is not available to conduct efficient GxE	Individual analysts tweak existing software to generate limited GxE results	Implement new software designed for high-volume GxE analyses using novel

Gauderman et al., submitted 2016

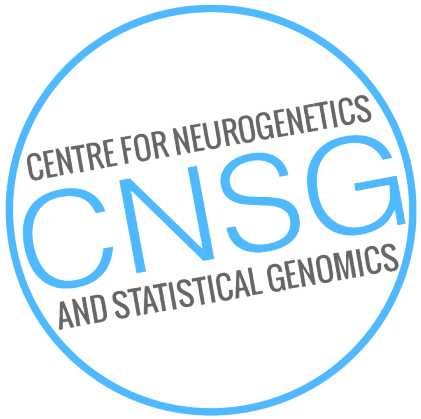
GxE Software

Program	GxE	Case-only	EB	2-Step	Additive models
PLINK	X				
GxEScan	X	X		X	
CGEN	X	X	X		X

2016 Module 18:
Statistical & Quantitative Genetics of Disease

Lecture 7
Risk profile scores

Naomi Wray

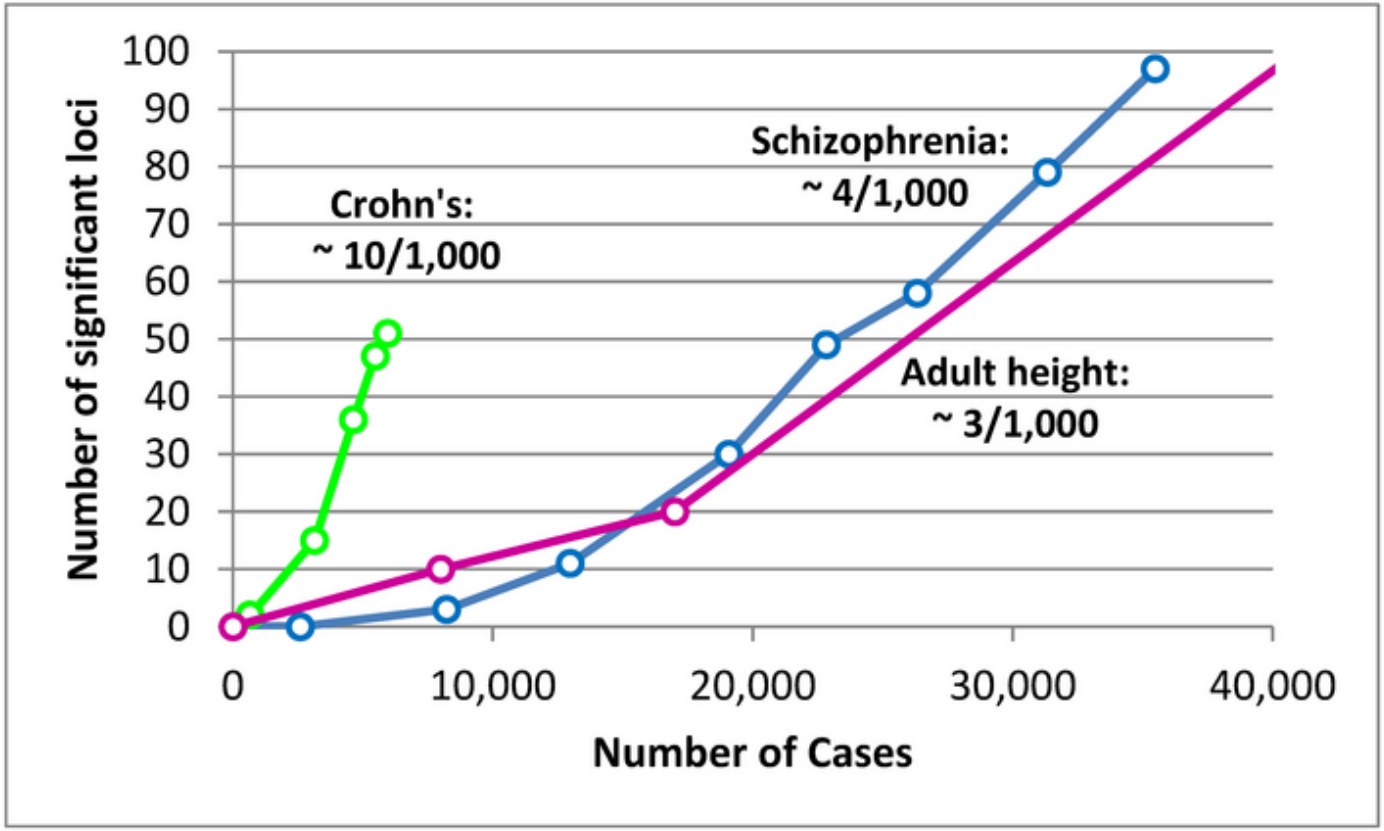


Aims of Lecture 7

1. Calculation of Risk Profile Scores
2. Examples of Use of Risk Profile Scores
3. Statistics to evaluate risk profile scores
    a. Nagelkerke's R^2
    b. AUC
    c. Decile Odds Ratio
    d. Variance explained on liability scale
    e. Risk stratification

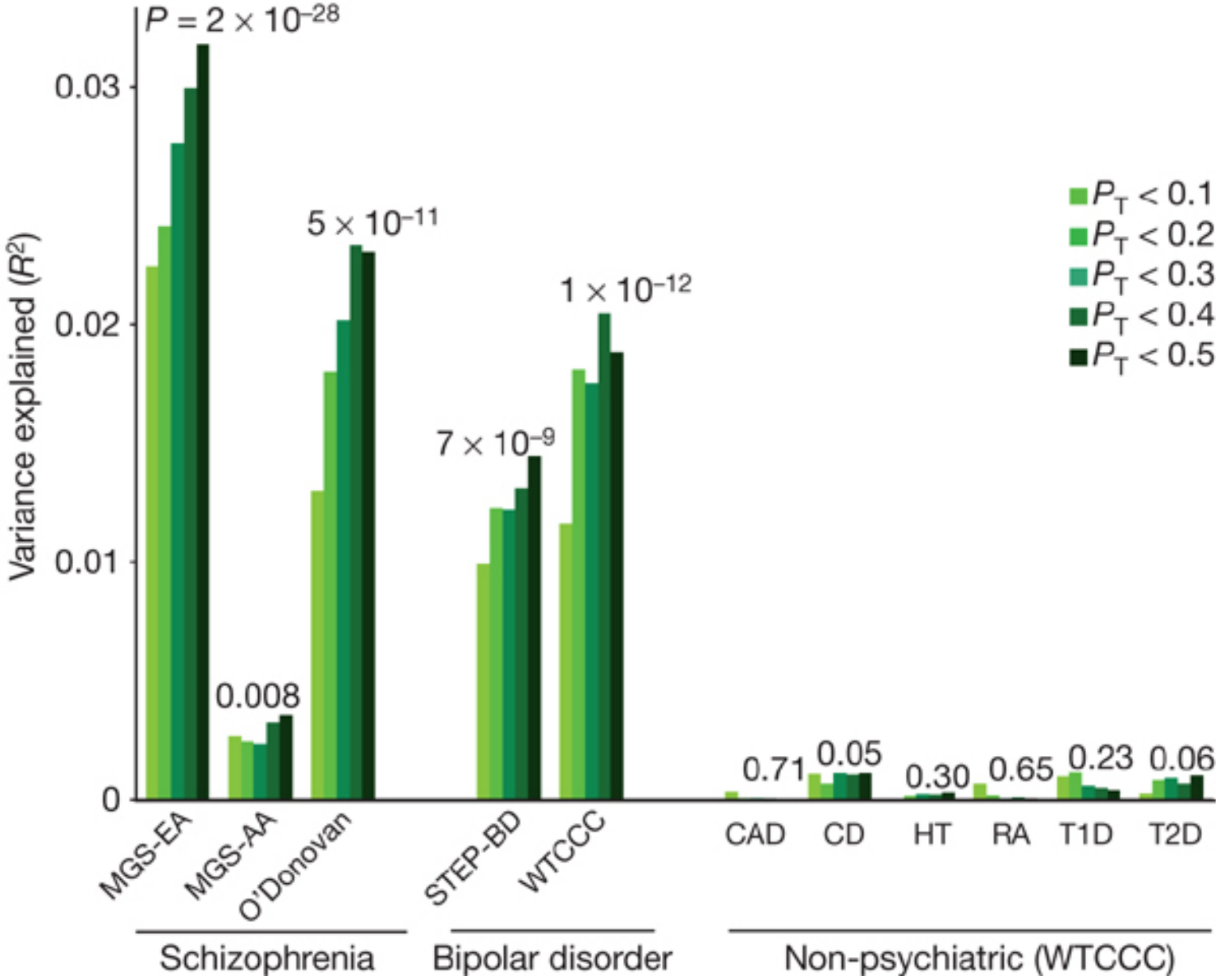
Polygenic risk profile

Evidence for a polygenic contribution to disease



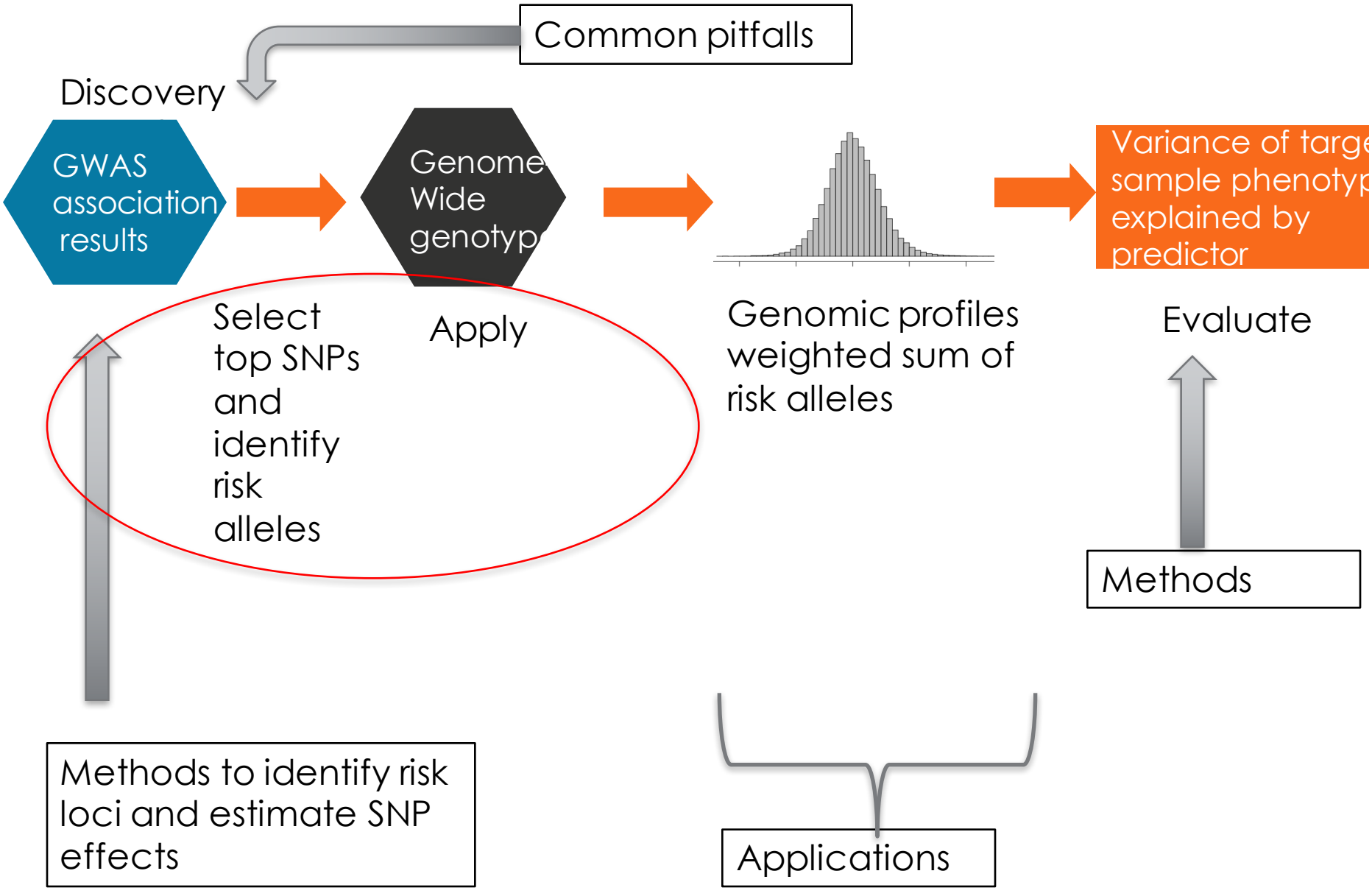
Levinson et al (2014) Genetic studies of major depressive disorder. Why are there no GWAS findings and what can we do about it? Biological Psychiatry

Risk Profile Scoring

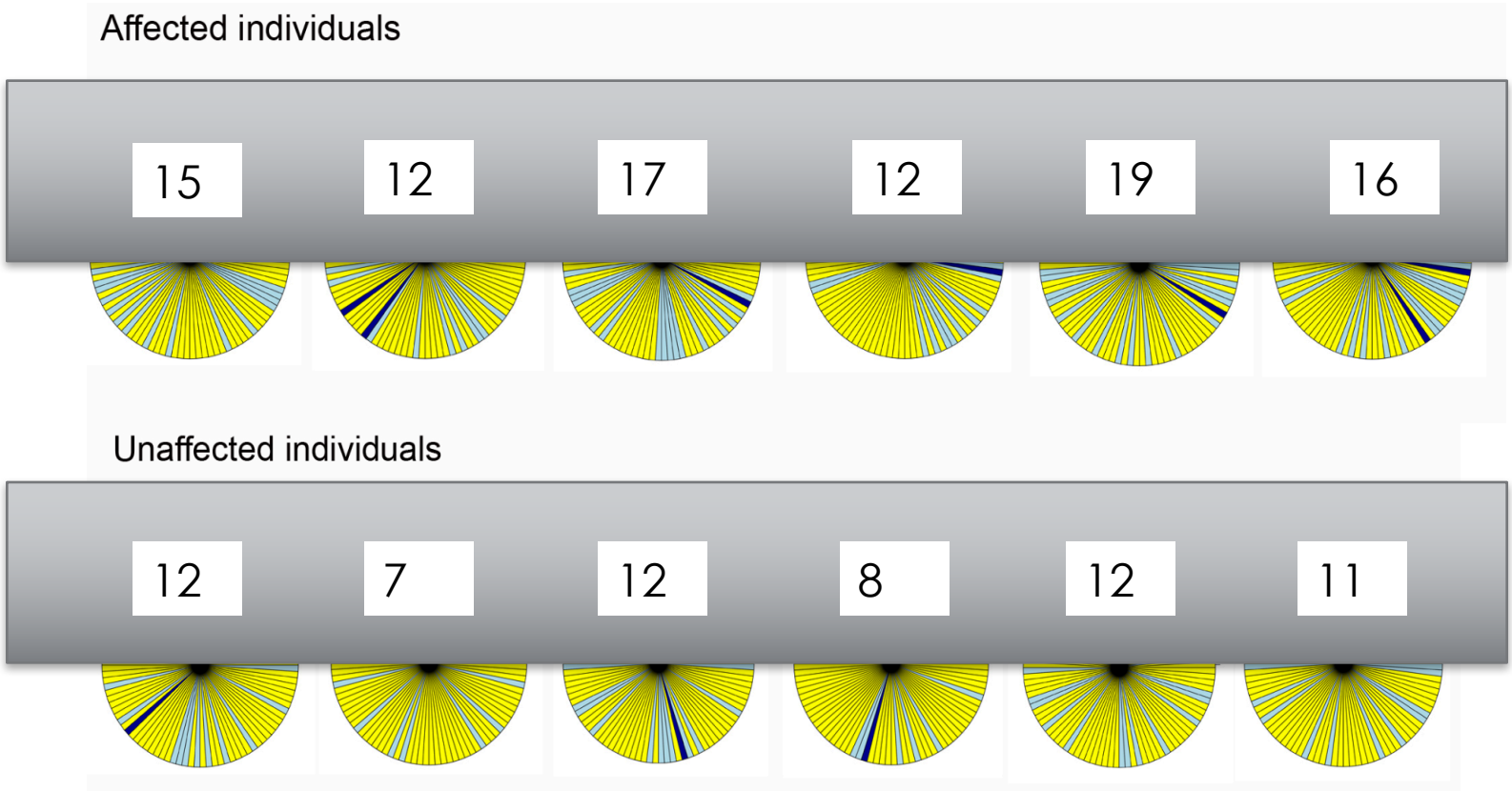


Purcell / ISC et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder *Nature* 2009

SNP profiling schematic



Visualising variation between individuals for common complex genetic diseases



- Not all affected individuals carry the risk allele at any particular locus
- Unaffected individuals carry multiple risk loci
- Consequences of risk alleles depend on the genetic and environmental background

Steps 1 - 3 in polygenic risk scoring

1. Identify Discovery sample with genome-wide association analysis summary statistics

1. Identify Target sample with genome-wide genotypes.
   - The Target sample should not include individuals closely related to those in the Discovery sample. Results can be inflated if there is overlap between samples.

2. Determine the list of SNPs in common between Discovery and Target samples

See: Wray et al (2013) Pitfalls of predicting complex traits from SNPs. Nature Reviews Genetics

Steps 4-7 in polygenic risk scoring as currently commonly applied

4. Construct a clumped SNP list: association p-value informed removal of correlated SNPs,
   - e.g. LD threshold of $r^2 < 0.2$ across 500 kb.
   - e.g.,in the program PLINK: −clump-p1 1–clump-p2 1–clump-r2 0.2–clump-kb 500

5. Limit SNP list to those with association p-value less than a defined threshold
   - often several thresholds are considered, i.e., <0.00001, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3 etc.
   –

6. Generate genomic profile scores in the target sample: e.g., sum of risk alleles weighted by Discovery sample log(odds ratio).
   - e.g., in PLINK: –score

7. Evaluate

Purcell / ISC et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder *Nature* 2009

Polygenic Modeling

Calculate polygenic risk score for individual j

$$Score_j = \frac{\sum_{i=1}^{m} \ln(OR_i) \times SNP_{ij}}{m}$$

where

- ln(*ORi*) = effect size or 'score' for SNP_i from 'discovery' sample
- SNP_{ij} = # of alleles (0,1,2) for SNP_i, person j in 'target' sample.
- m = number of SNPs considered in test set

Purcell / ISC et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder *Nature* 2009

Consider step 4

4. Construct a clumped SNP list: association p-value informed removal of correlated SNPs,
   - e.g. LD threshold of $r^2 < 0.2$ across 500 kb.
   - e.g.,in the program PLINK: –clump-p1 1–clump-p2 1–clump-r2 0.2–clump-kb 500

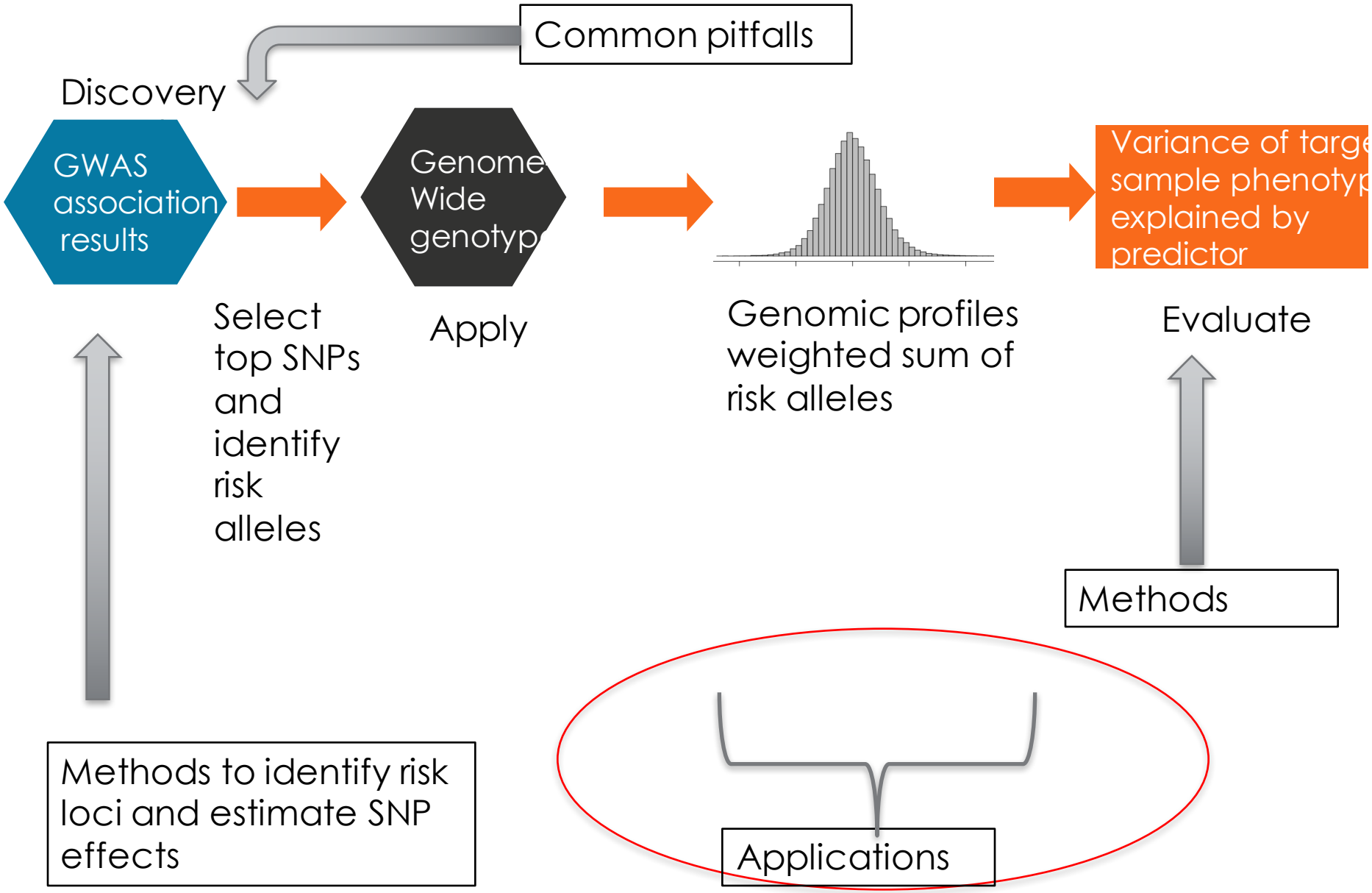
This step can be improved upon to make it less arbitary

Purcell / ISC et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder *Nature* 2009

Step 7 in polygenic risk scoring

7. Evaluate efficacy of score predictor.
   - Regression analysis:
   - y= phenotype, x = profile score.
   - Compare variance explained from the full model (with x) compared to a reduced model (covariates only).
   - Check the sign of the regression coefficient to determine if the relationship between y and x is in the expected direction.

SNP profiling schematic



Applications of polygenic Risk Profile Scoring

Discovery & Target samples could be:

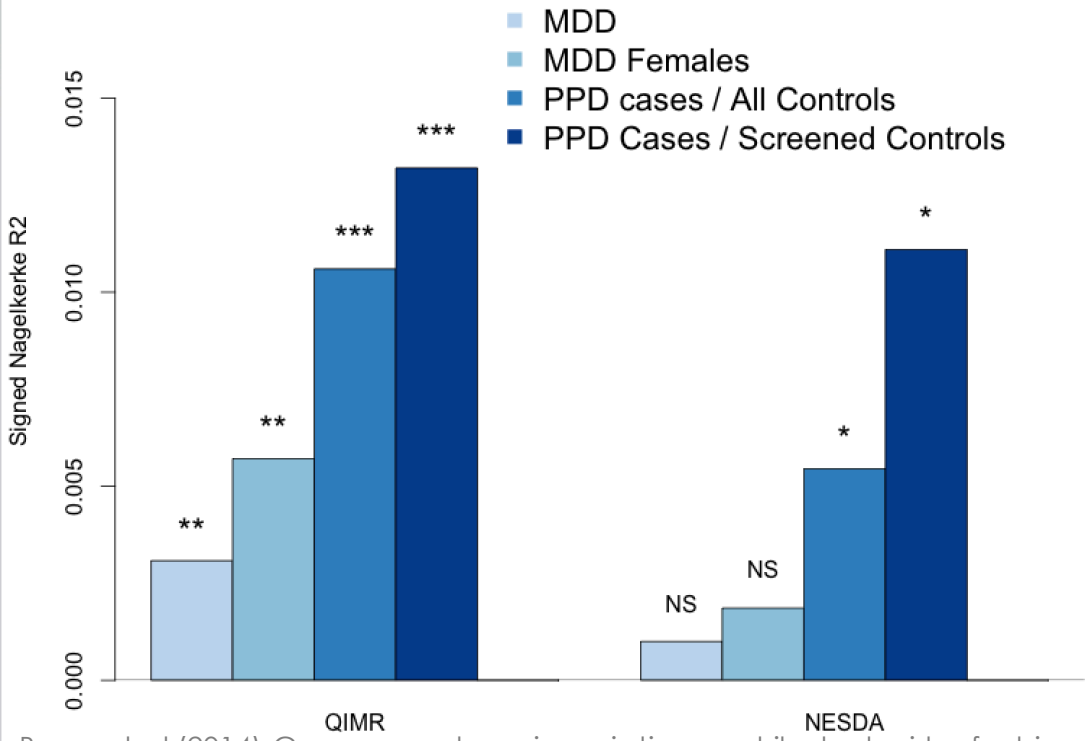
A. Same Disorder - demonstrates polygenicity even in absence of genome-wide significant SNP associations
B. Different disorders - demonstrates genetic overlap between disorders

A. Target samples are disorder subtypes
   - investigates genetic genetic heterogeneity
   - think carefully about how the heterogeneity is represented in the Discovery sample if Target and Discovery are the same disease

Example Disorder Sub-types. Discovery: PGC-BPD Target: Postnatal depression in MDD

Postnatal depression – a more homogeneous subtype of depression?

Female only
Same bio-social stressor



Enda Byrne
Carillo-Roa
Samantha Meltzer-Brody
Nick Martin
Brenda Penninx

Tania

NB. Null result in the ALSPAC community sample measured for PND but not MDD

Byrne et al (2014) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Archives of Women's Health. In press

Applications of polygenic Risk Profile Scoring

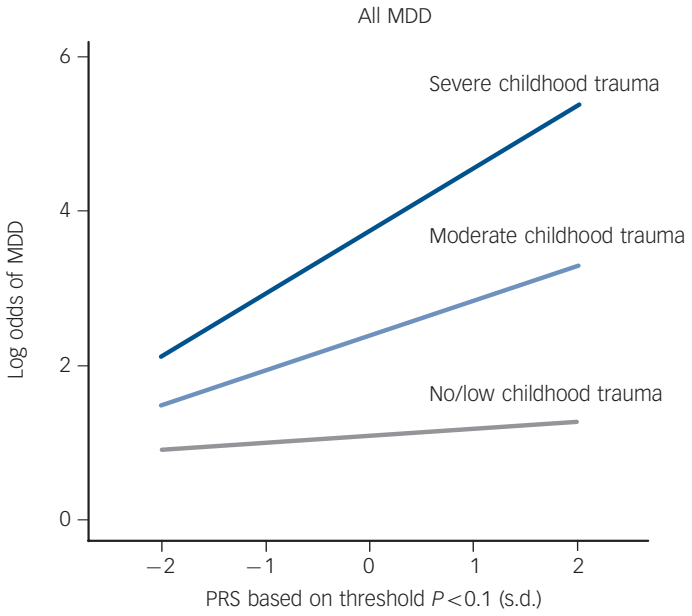
Discovery & Target samples could be:

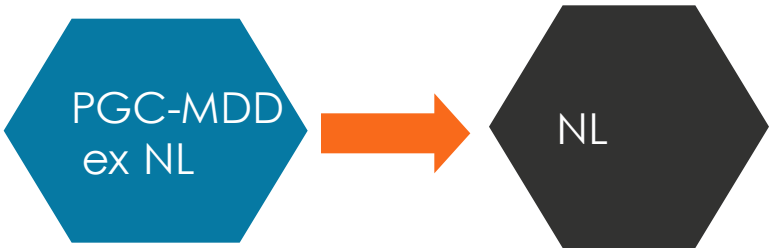
A. Same Disorder - demonstrates polygenicity even in absence of genome-wide significant SNP associations
B. Different disorders - demonstrates genetic overlap between disorders

A. Target samples are disorder subtypes
   - investigates genetic genetic heterogeneity
   - think carefully about how the heterogeneity is represented in the Discovery sample if Target and Discovery are the same disease

D. Target samples have the same disease as the discovery sample and have environmental risk factors recorded
   - investigate GxE
   - think carefully about how the environmental risk factor is represented in the Discovery sample

Application of Polygenic Risk Profiling Scores to investigate GxE, e.g., depression and childhood trauma





Peyrot et al (2014) Effect of polygenic risk scores on depression in childhood trauma Biol Psychiatry

Applications of polygenic Risk Profile Scoring

Discovery & Target samples could be:

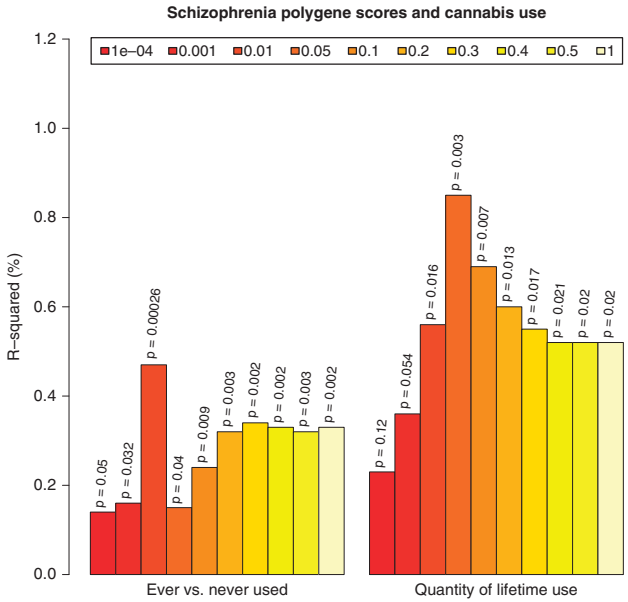
A. Same Disorder - demonstrates polygenicity even in absence of genome-wide significant SNP associations
B. Different disorders - demonstrates genetic overlap between disorders

A. Target samples are disorder subtypes
   - investigates genetic genetic heterogeneity
   - think carefully about how the heterogeneity is represented in the Discovery sample if Target and Discovery are the same disease

D. Target samples have the same disease as the discovery sample and have environmental risk factors recorded
   - investigate GxE
   - think carefully about how the environmental risk factor is represented in the Discovery sample

E. Target samples are recorded for an environmental risk factor
   - insight into GxE

Example: E in target sample
Discovery: schizophrenia
Target: Cannabis use



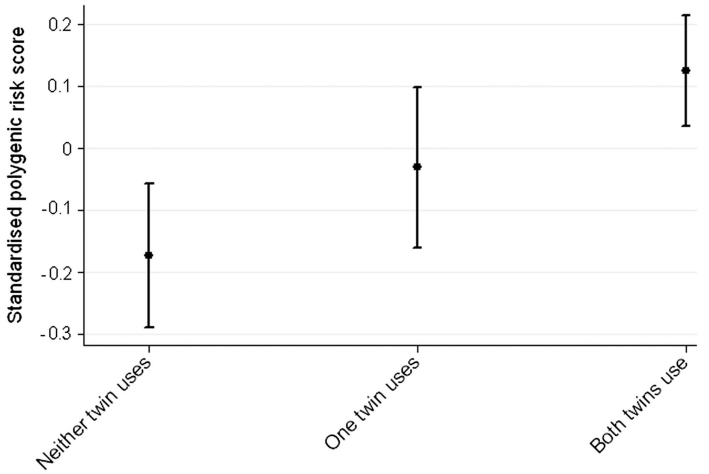
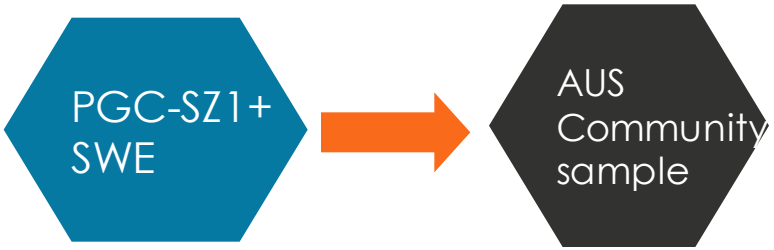


Figure 2. Mean standardized polygenic risk scores for pairs of twins when neither ($n = 272$), one ($n = 273$) or both twins ($n = 445$) had reported use of cannabis. An ordinal regression reported a significant association ($P = 0.001$).



Power et al (2014) Effect of polygenic risk scores on depression in childhood trauma Mol Psychiatry

Factors affecting accuracy of risk prediction

Genetic architecture of the trait - unknown

Sample size of discovery sample – maximise

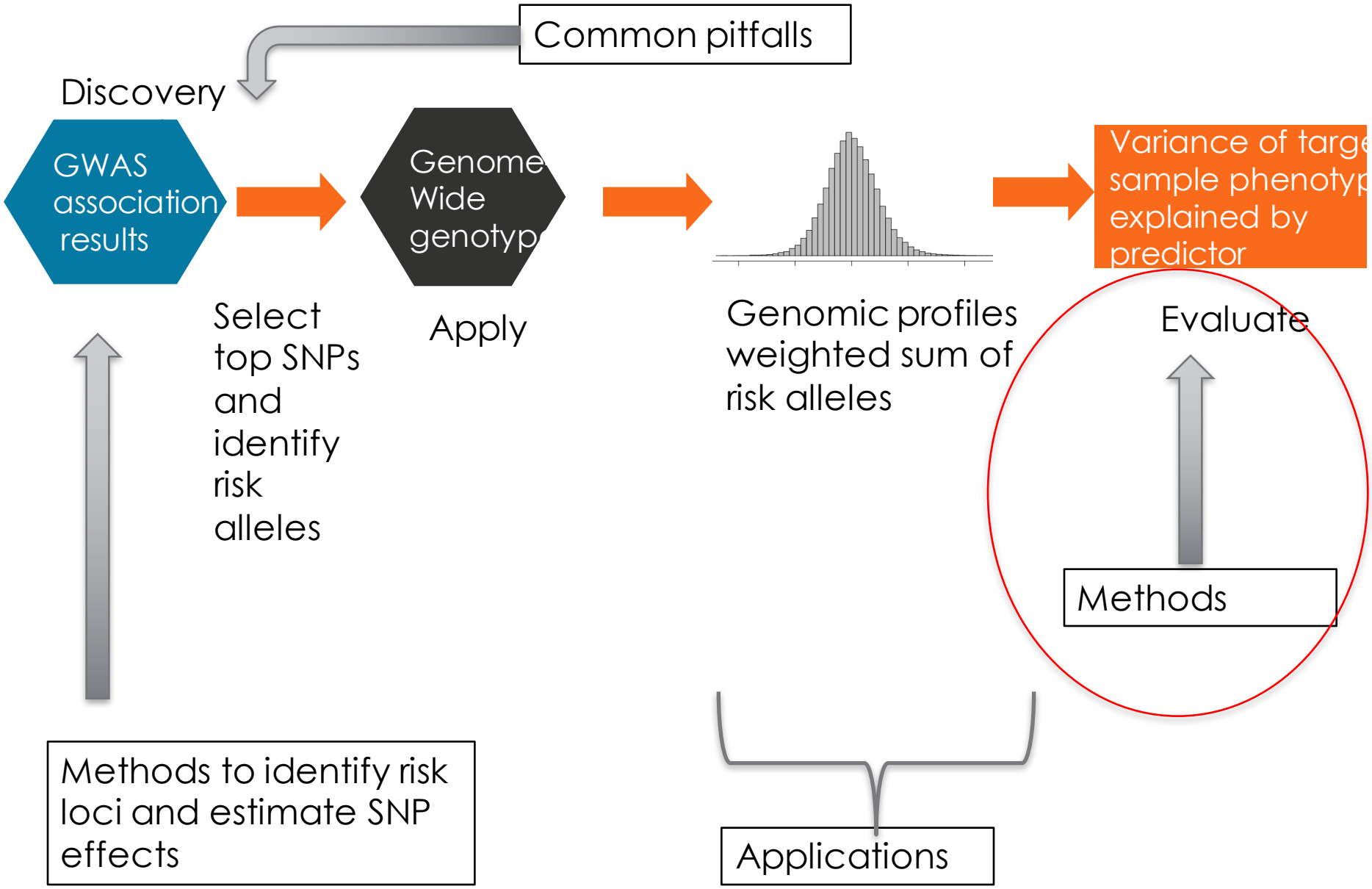
Sample size of target sample – be sufficiently large (once achieved not so much gained by increasing further)

Variance explained by genetic factors

Dudbridge (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genetics
Wray et al (2014) Polygenic methods and their application to psychiatric traits. Journal of Child Psychology & Psychiatry (in press)

Evaluating Polygenic Risk Scores

SNP profiling schematic



Statistics to evaluate polygenic risk scoring 1.

1. Nagelkerke's R^2
   - Pseudo-R^2 statistic for logistic regression

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

Cox & Snell R^2

$$= 1 - \exp\left(\frac{2}{N}\right)(LogLikelihood\ (Reduced\ model) - LogLikelihood(Full\ model))$$

Full model: y ~ covariates + score
Reduced model: y ~ covariates
N: sample size

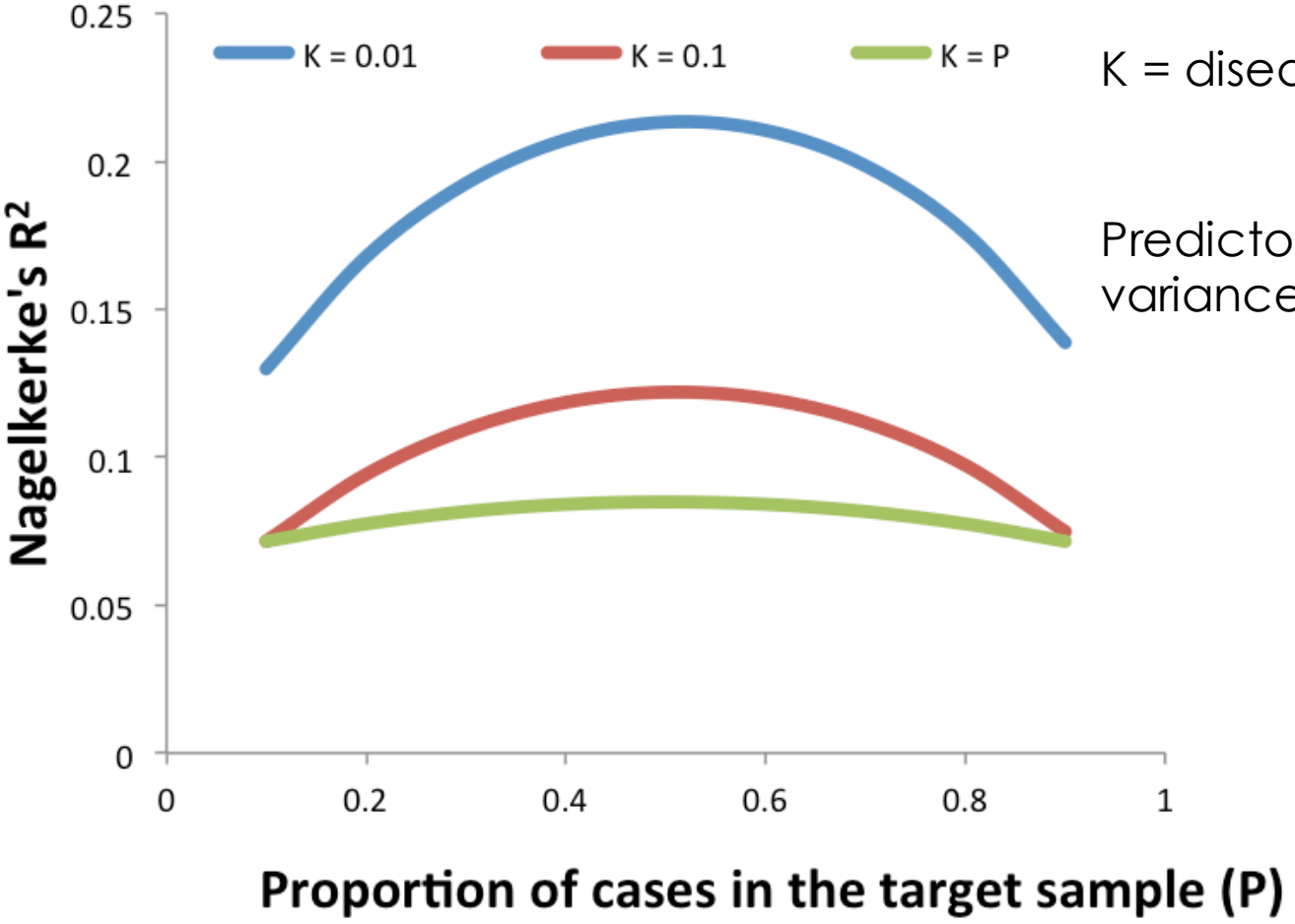
Logistic, y= case/control = 1/0

This definition gives R2 for a quantitative trait.
For a binary trait in logistic regression, C&S R2 has maximum

$$= 1 - \exp\left(\frac{2}{N}\right)(LogLikelihood\ (Reduced\ model))$$

Nagelkerke's R^2 divides Cox & Snell R^2 by its maximum to give an R^2 with usual properties of between 0 and 1.

Problem with Nagelkerke's R^2



K = disease prevalence

Predictor explains 7% of variance in liability

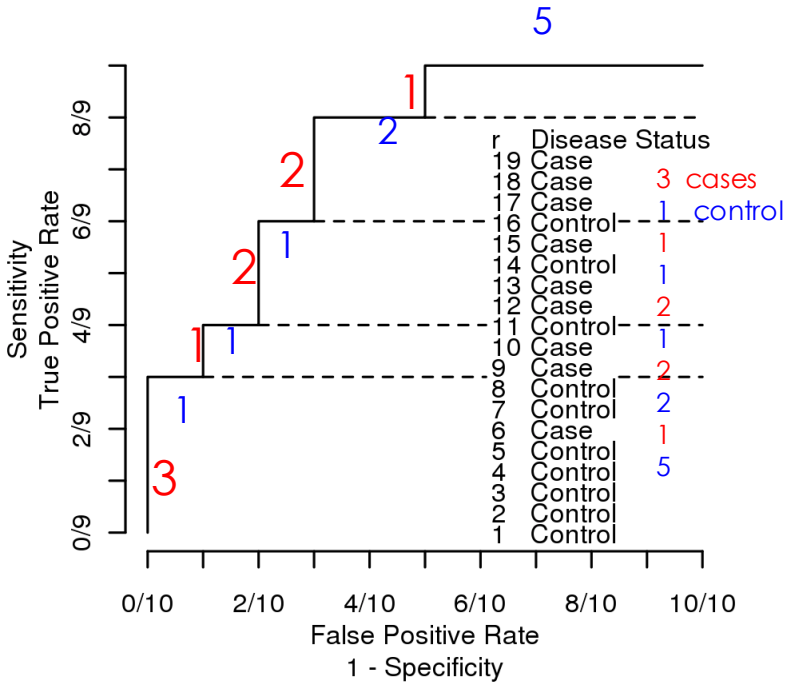
Statistics to evaluate polygenic risk scoring 2.

2. Area Under Receiver Operator Characteristic Curve

- Well established measure of validity of tests for classifier diseased vs non-diseased individuals
- Nice property – independent to proportion of cases and controls in sample
- Range 0.5 to 1
- 0.5 the score has no predictive value
- Probability that a randomly selected case has a score higher than a randomly selected control

Visualising AUC

- Rank individuals on score
- Start at origin on graph
- Work through list of ranked individuals
- Move one unit along y-axis if next individual is a case
- Move one unit along x-axis is next individual is a control6



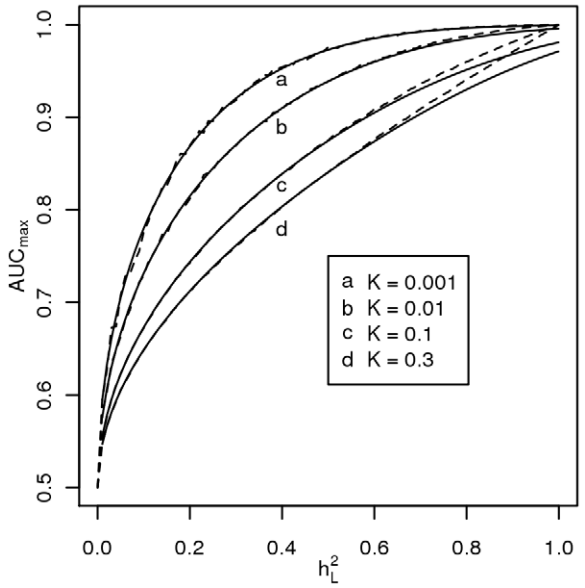
Problem with AUC

Well recognised as a measure of clinical validity

A measure of how well genomic profile predicts yes/no phenotype

But hides the fact that is should be judged as a measure of analytic validity

A measure of how well genomic profile predicts genotype



The maximum AUC achievable depends on the heritability of the disease

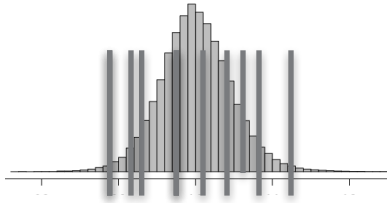
Many useful properties

Problem is genetic interpretation

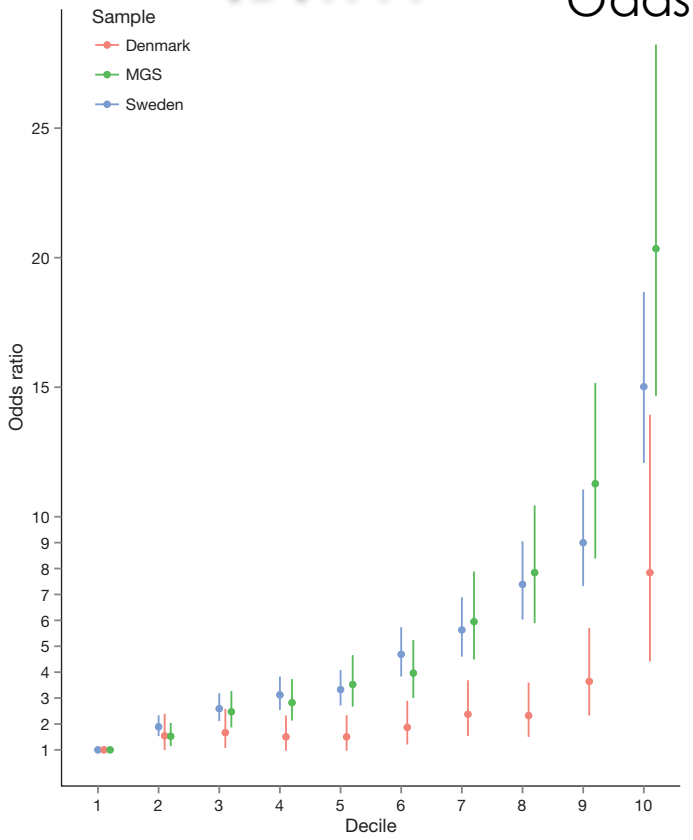
Wray et al (2010) The genetic interpretation of area under the receiver operator characteristic curve in genomic profiling. PLoS Genetics

Statistics to evaluate polygenic risk scoring 3.

3. Odds Ratio



Cut distribution into deciles
Each decile will include both cases and controls
Odds of being a case in each decile
Odds ratio for each decile compared to the 1st decile

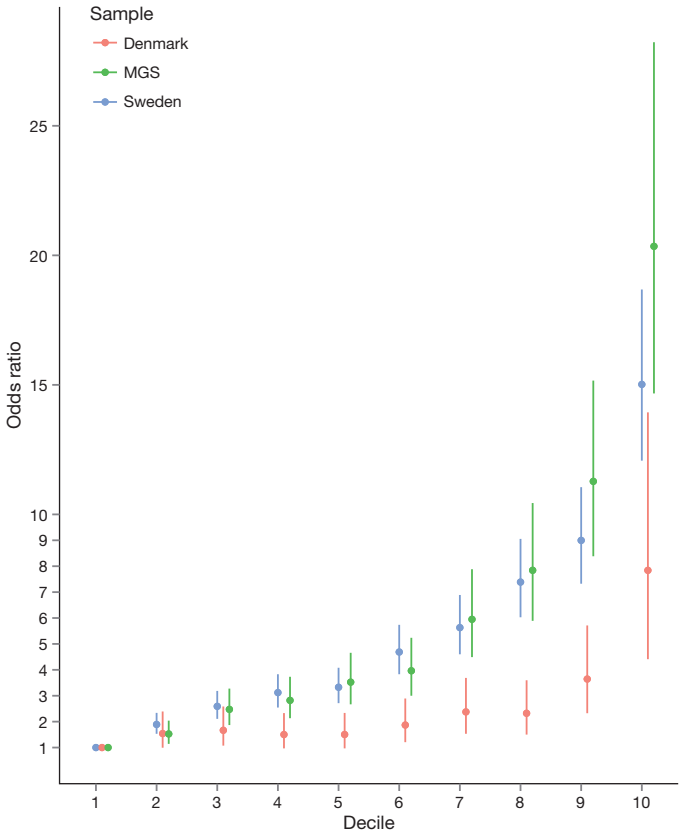


- Good visualisation
- Shows that there could be utility in using high vs low profile risk scores
- But remember case-control samples are 50% cases
- Would look less impressive if a population sample

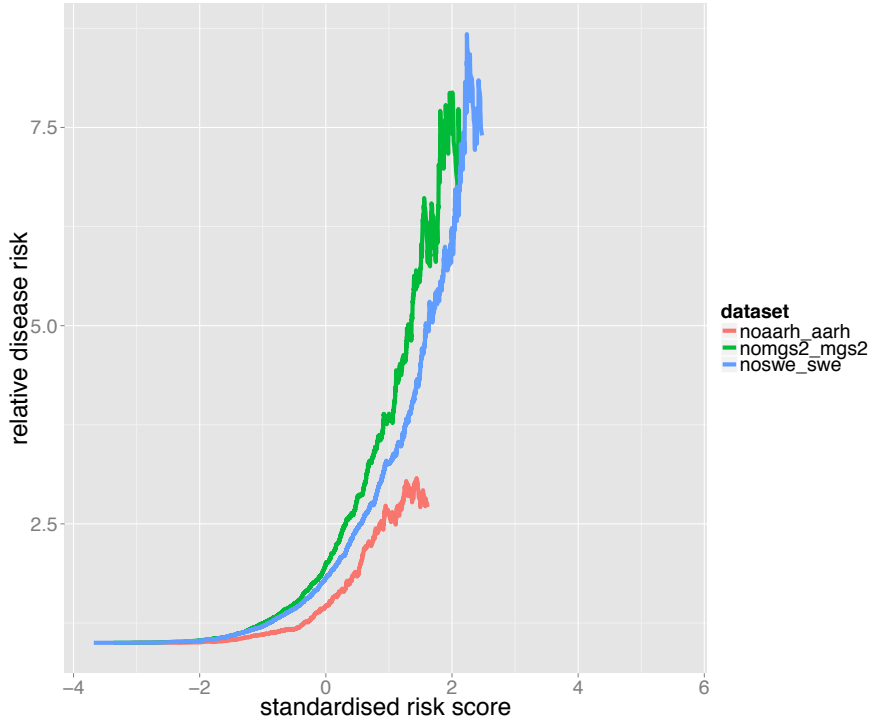
PGC-SCZ 2014 108 loci Nature

Statistics to evaluate polygenic risk scoring 3.

In case control samples



Same data scaled to population risk



Statistics to evaluate polygenic risk scoring 4.

3. R^2 on liability scale

Linear model

Full model: y ~covariates + score y = case/control = 1/0

Reduced model: y~ covariates

Calculate R^2 attributable to score

If target sample is a population sample i.e. prevalence of cases in sample = prevalence of cases in controls

Then R^2 is a measure of the proportion of variance in case-control status attributable to the genomic risk profile score

= heritability attributable to genomic profile score $h^2_{GRPS-01}$ on the disease scale

Convert to liability scale (see lecture 1)

$$h^2_{GRPS} = \frac{h^2_{GRPS-01} K(1-K)}{z^2}$$

Lee et al (2012) A better coefficient of determination for genetic profile analysis. Genetic Epidemiology

Statistics to evaluate polygenic risk scoring 4 cont.

3. R^2 on liability scale cont.

   If target sample is a case-control sample

   i.e. prevalence of cases in sample >> prevalence of cases in controls

   Then R^2 is a measure of the proportion of variance in case-control status attributable to the genomic risk profile score

   = heritability attributable to genomic profile score on the case-control scale

$$h^2_{GRS-CC}$$

Convert to the liability scale

$$h^2_{GRS} = \frac{h^2_{GRS-CC} C}{1 + h^2_{GRS-CC} C}$$

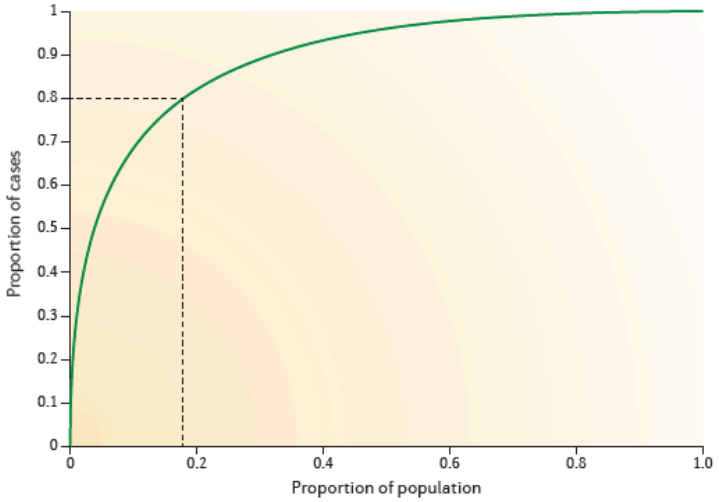
Where C is:

$$C = \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)}$$

h^2_{GRS} is on the same scale as heritability estimated from family studies and GREML SNP-chip heritability

Lee et al (2012) A better coefficient of determination for genetic profile analysis. Genetic Epidemiology

Statistics to evaluate polygenic risk scoring
5.Stratification & health economics



Population risk of 1%

80% of cases in
top 18% of genetic risk

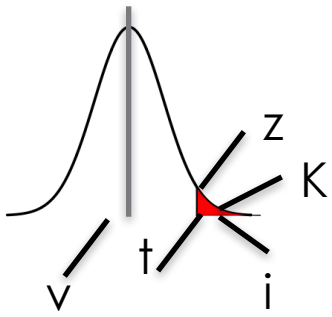
For every 1,000 people treated with intervention could "save" 10
Treat only 18% = 180 and "save" 8 (4%)

Number of people treated to save 1 reduced from 100 to 22.5

Polychronakos & Li NRG (2011) Understanding Type I Diabetes through genetics. Nat Rev Genetics

Area Under ROC

Phenotypic liability variance is 1



Variance explained by genetic predictor r^2 (has max h^2)

Mean phenotypic liability of cases = i = z/K

Mean genetic liability of controls = v = -iK/(1-K)

Mean liability of cases explained by predictor = ir^2

Mean genetic liability of controls = vr^2

Variance of genetic predictor in cases = $r^2(1-r^2i(i-t))$

Variance of genetic predictor in controls = $r^2(1-r^2v(v-t))$

Using normal distribution theory can work out

Proportion of cases captured when x% of population screened

Proportion of population that needs to be screened in order to capture 80% of the cases

Improvement between predictors

Difference in AUC

Net reclassification index

The NRI, as originally proposed, seeks to quantify whether a new marker provides clinically relevant improvements in prediction. In the definition of "net reclassification indices," the risk prediction model with established predictors is called the "old" model. The model that adds the new marker is the "new" model. "Events" are cases—persons who have or will have the disease or outcome in the absence of intervention. "Nonevents" are controls. The formula defining the NRI is[4]

$$\mathrm{NRI} = P(\mathrm{up}|\mathrm{event}) - P(\mathrm{down}|\mathrm{event}) + P(\mathrm{down}|\mathrm{nonevent}) - P(\mathrm{up}|\mathrm{nonevent}). \quad (1)$$

$$\mathrm{NRI}_{\mathrm{e}} = P(\mathrm{up}|\mathrm{event}) - P(\mathrm{down}|\mathrm{event})$$

$$\mathrm{NRI}_{\mathrm{ne}} = P(\mathrm{down}|\mathrm{nonevent}) - P(\mathrm{up}|\mathrm{nonevent})$$

Topic of debate
Needs more research

Kerr et al (2014) NRI for evaluating risk prediction indices.

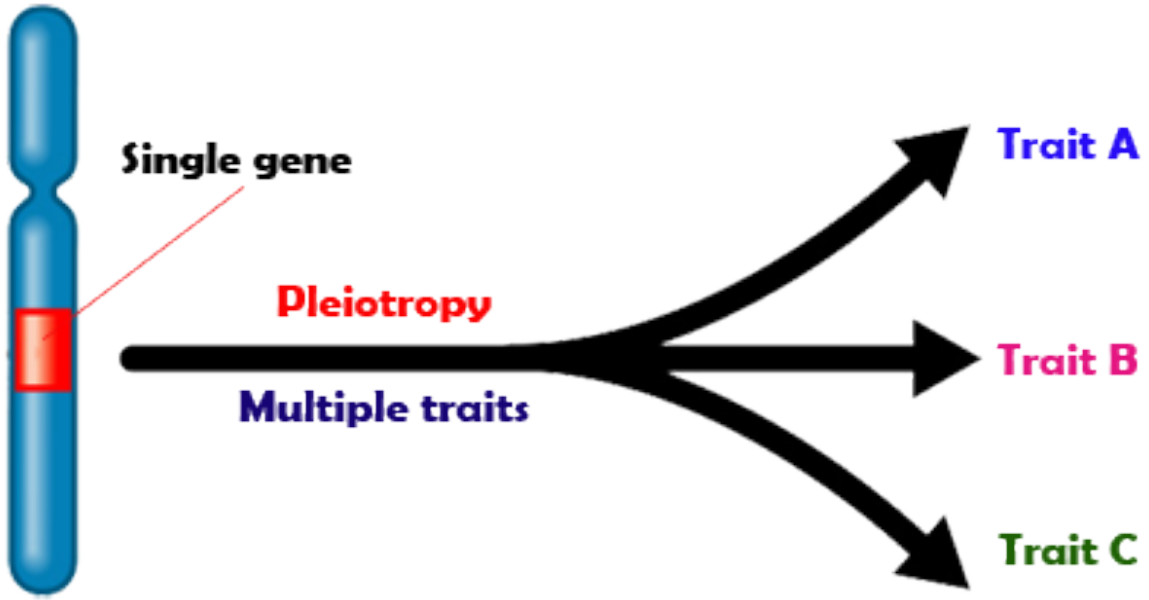
Module 18: Statistical and Quantitative Genetics of Disease: Pleiotropy / Co-heritability

John Witte

Lecture 8

Pleiotropy

- From Greek: Pleio (many) and tropic (affecting).

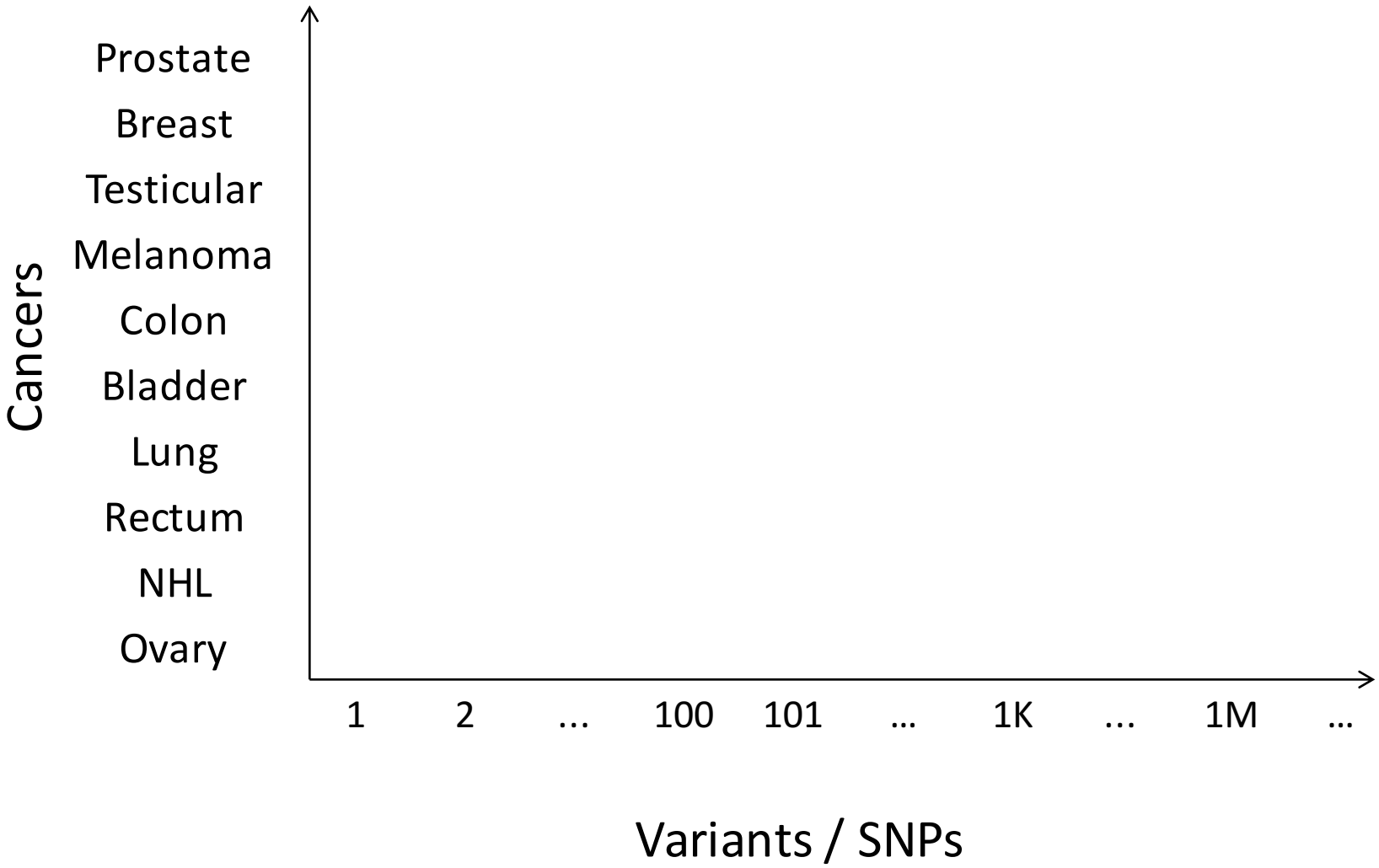


- One gene, multiple traits.

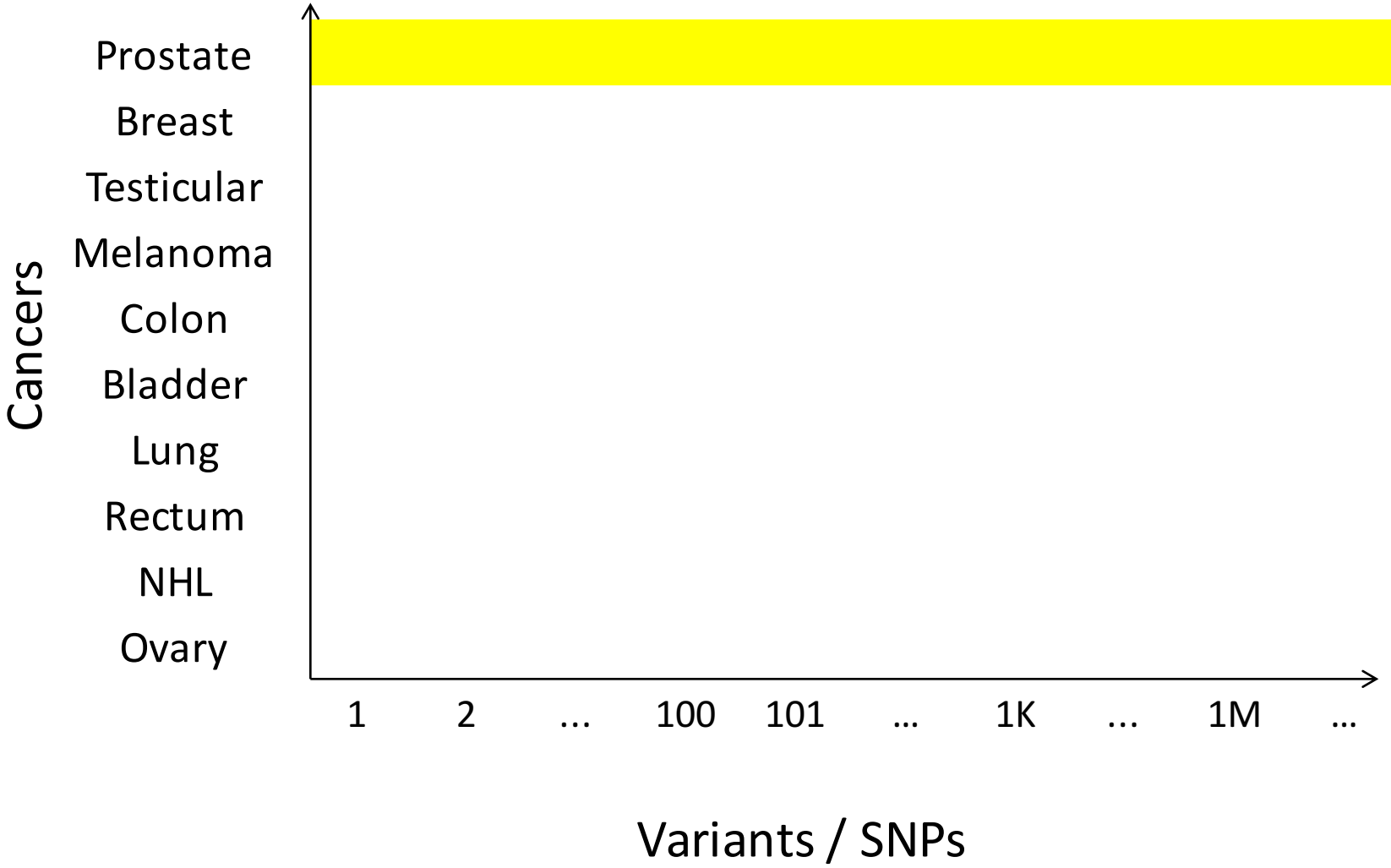
Assessing Pleiotropy

1. Pleiotropy ‘look-ups’
2. Meta-analysis (ASSET)
3. Multiphenotype
4. Multilevel pleiotropy
5. Polygenic risk scores
6. Co-heritability

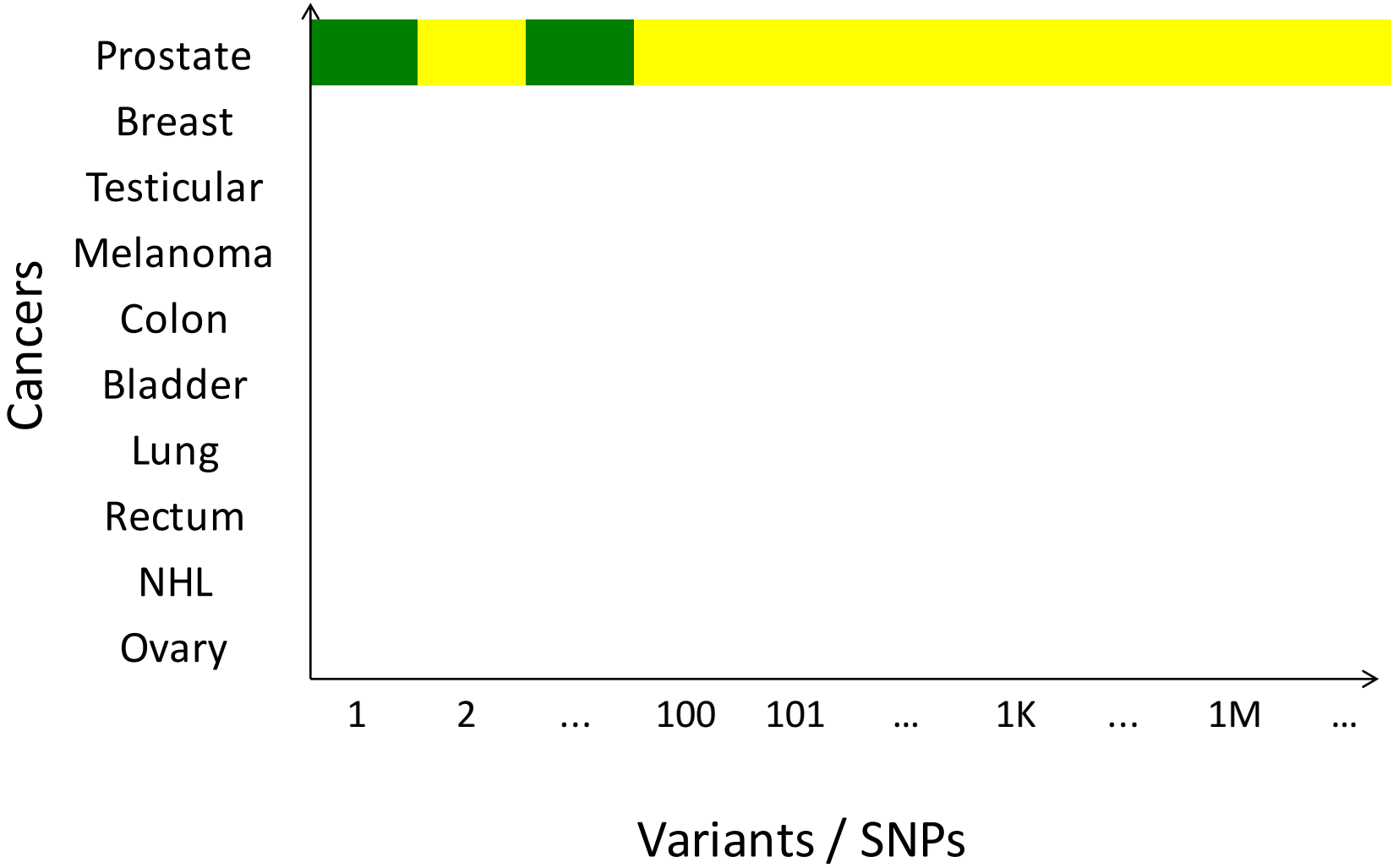
1. Pleiotropy “Look-ups”



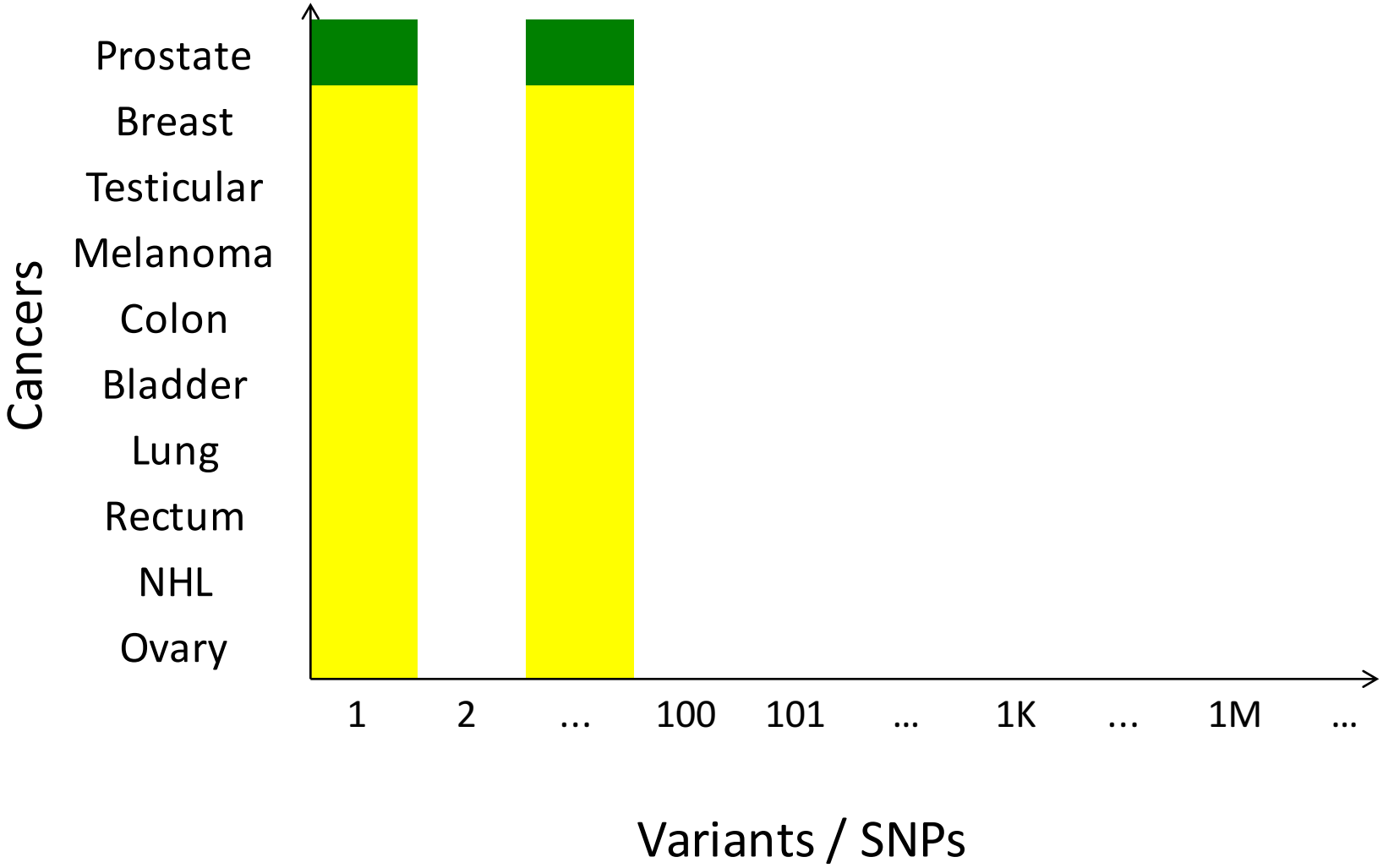
Pleiotropy “Look-ups”



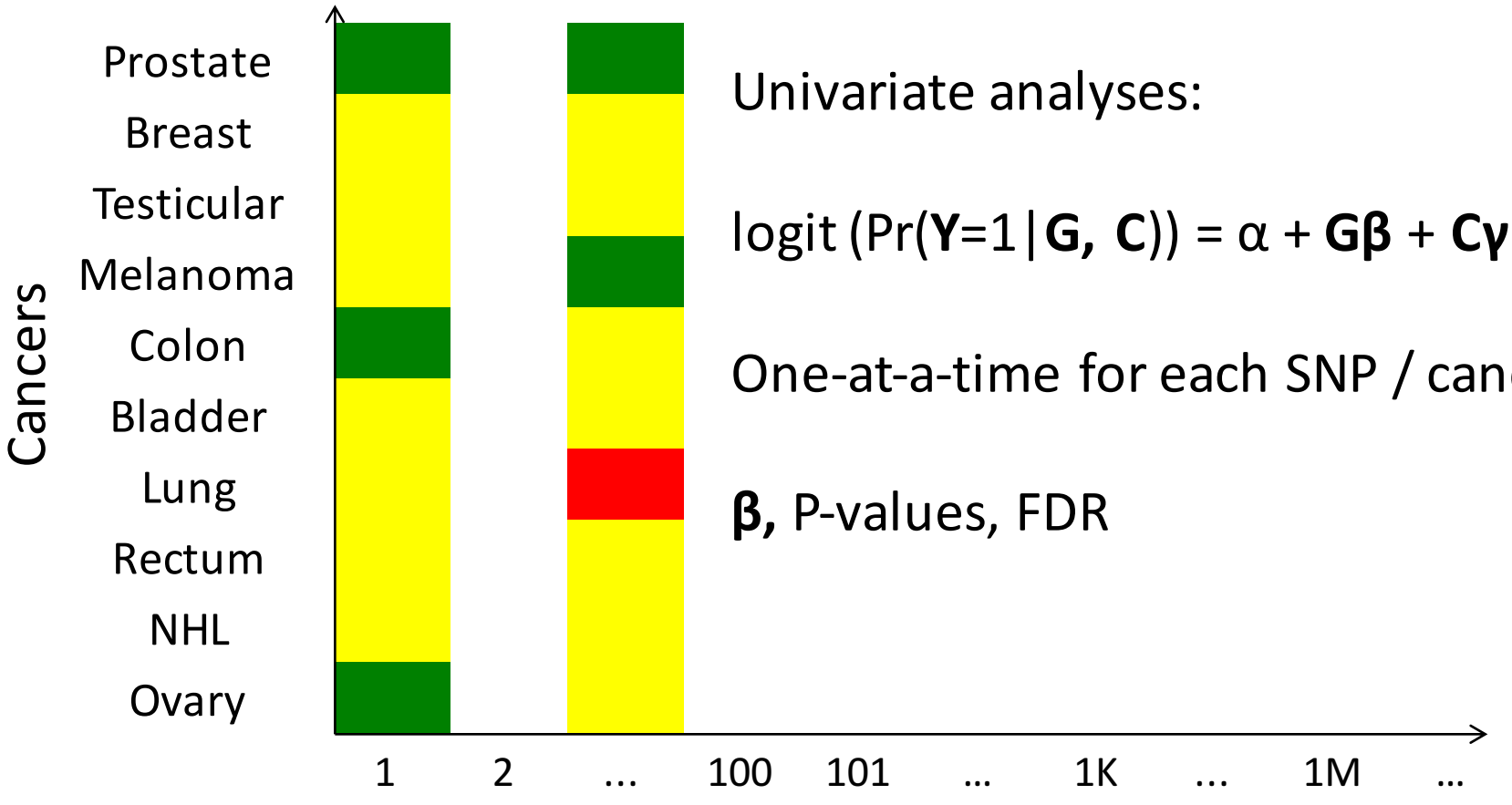
Pleiotropy “Look-ups”



Pleiotropy “Look-ups”



Pleiotropy “Look-ups”



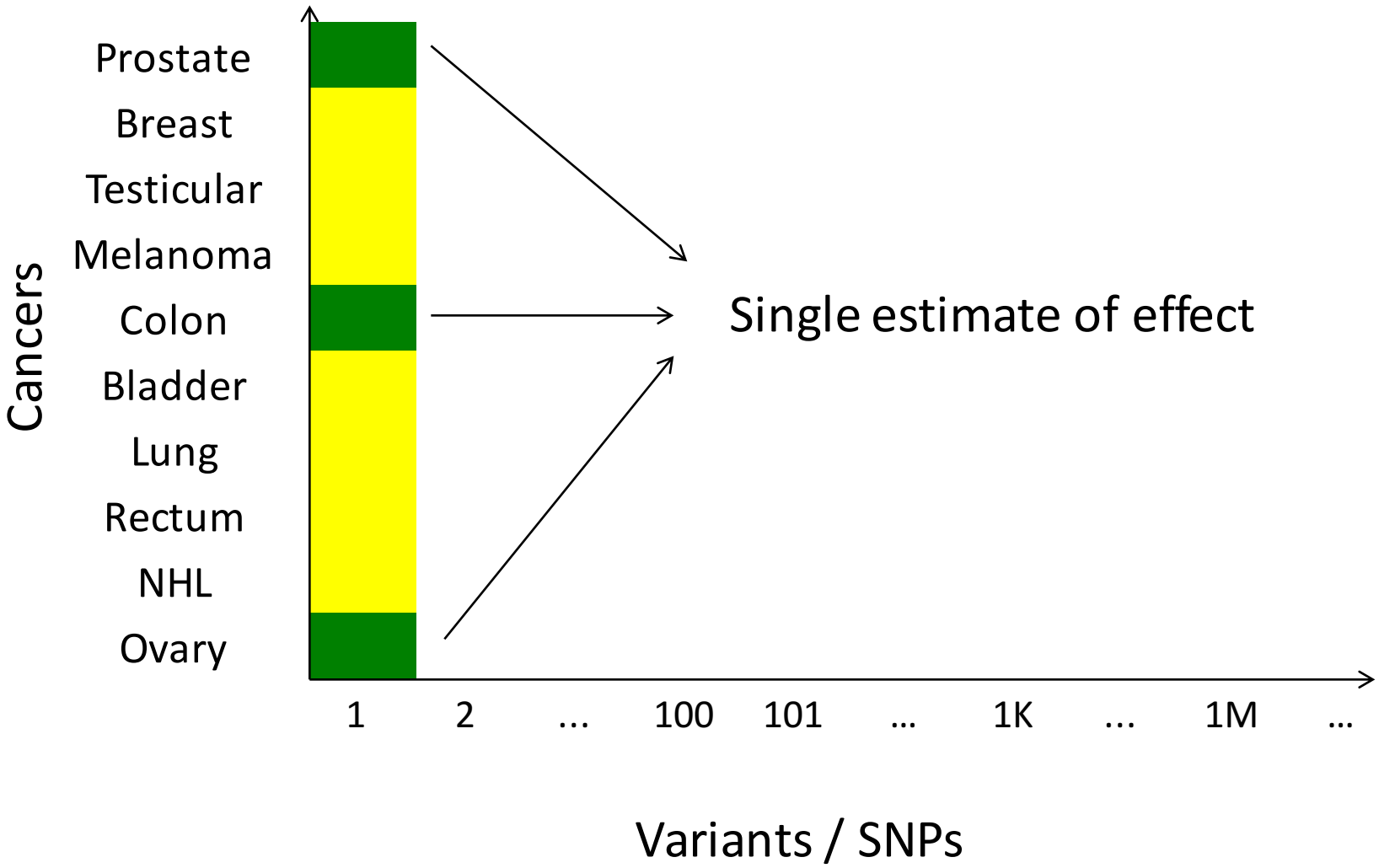
Univariate analyses:

logit (Pr(**Y**=1|**G, C**)) = α + **Gβ** + **Cγ**

One-at-a-time for each SNP / cancer

β, P-values, FDR

2. Meta-Analysis Approach



ASSET

Standard fixed-effects

$$Z_{meta} = \sum_{k=1}^{K} \sqrt{\pi_k} Z_k \quad \text{Where} \quad Z_k = \beta_k / se(\beta_k)$$

$$\pi_k = n_k / \sum_{k=1}^{K} n_k$$

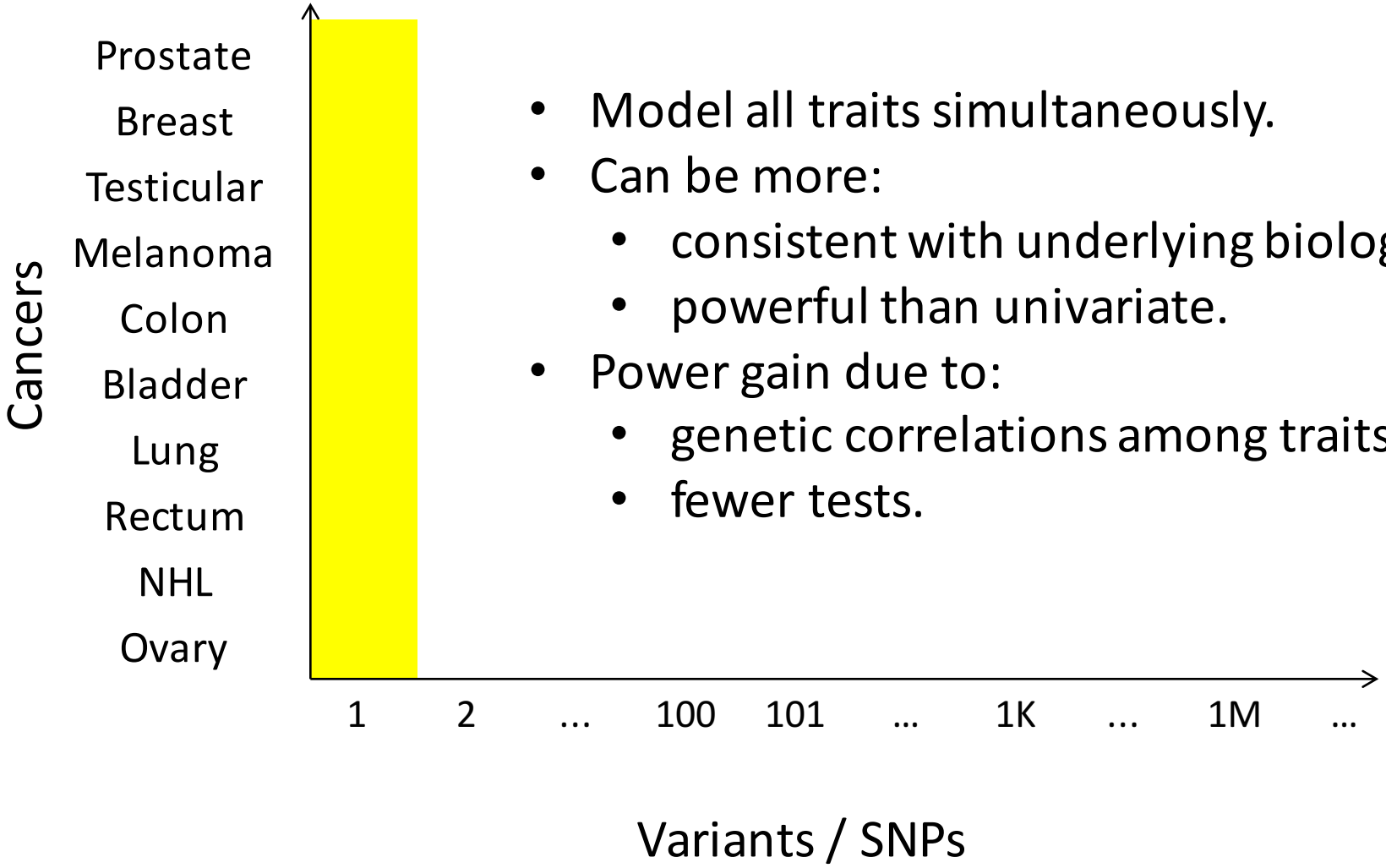
Subset-based

$$Z_{\max-meta} = \max_{s \in S} \left| Z(s) \right|$$

Where $$Z(s) = \sum_{k \in S} \sqrt{\pi_k(s)} Z_k$$

Bhattacharjee *et al. AJHG,* 2012

3. Multiphenotype Approach



- Model all traits simultaneously.
- Can be more:
  - consistent with underlying biology;
  - powerful than univariate.
- Power gain due to:
  - genetic correlations among traits;
  - fewer tests.

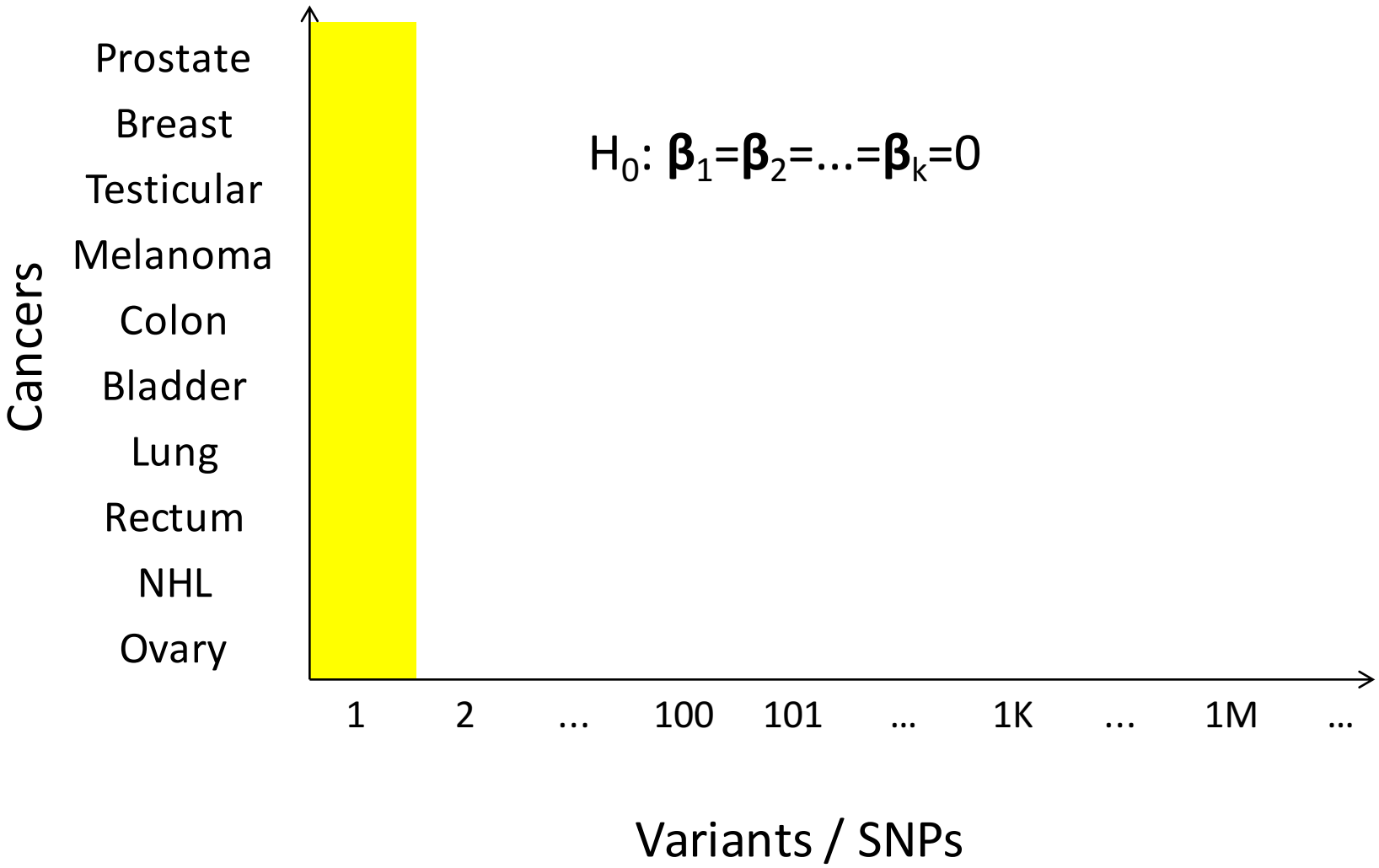
Multinomial Regression

$$\text{logit}\left(\Pr(\mathbf{Y_i}=1|\mathbf{G}, \mathbf{C})\right) = \alpha_i + \mathbf{G_i}\boldsymbol{\beta}_i + \mathbf{C}\boldsymbol{\gamma}_i$$

Y is multivariate with dimension = # traits

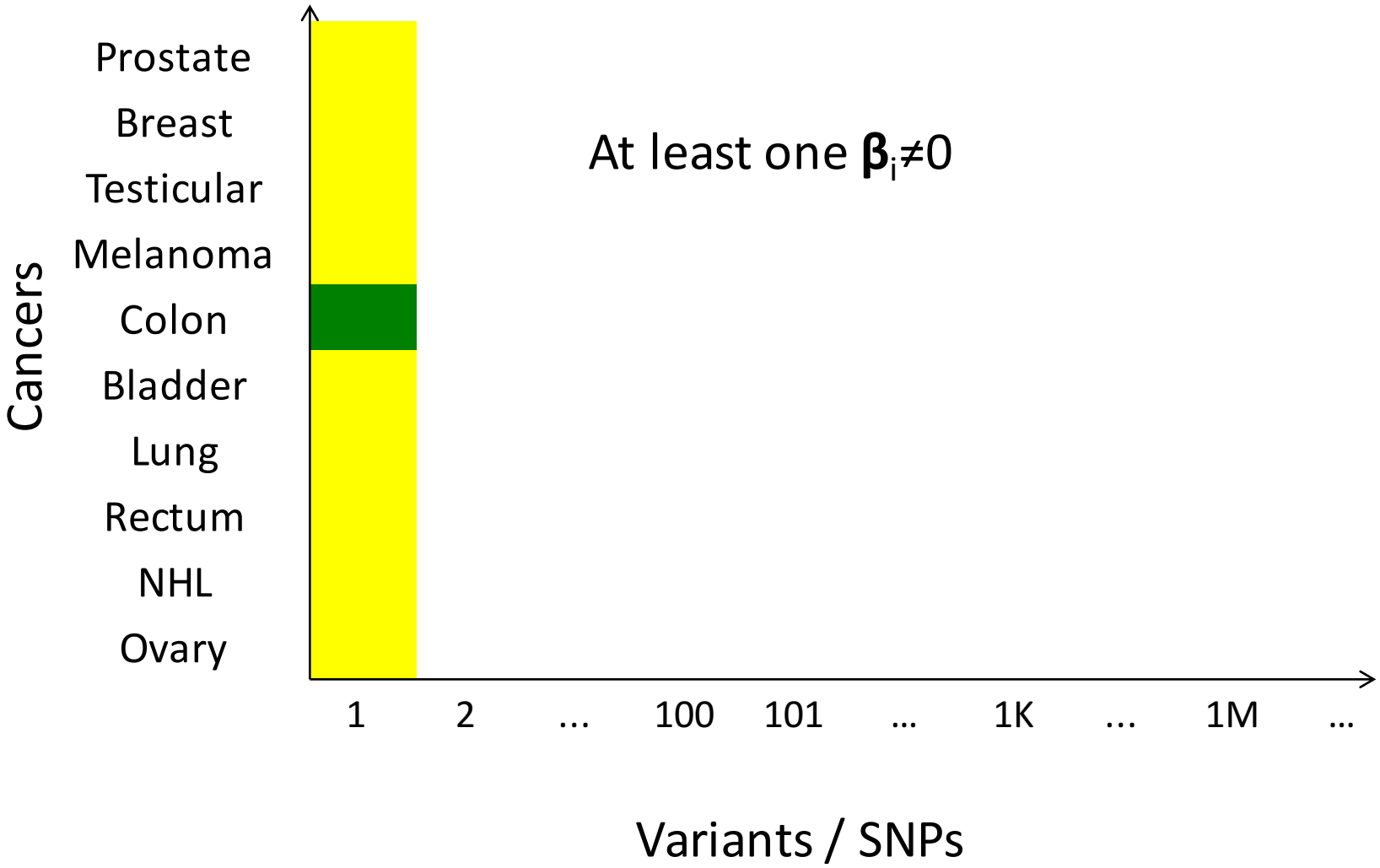
Test different pleiotropic models by specifying assumptions about the $\boldsymbol{\beta}_i$.

Null Model

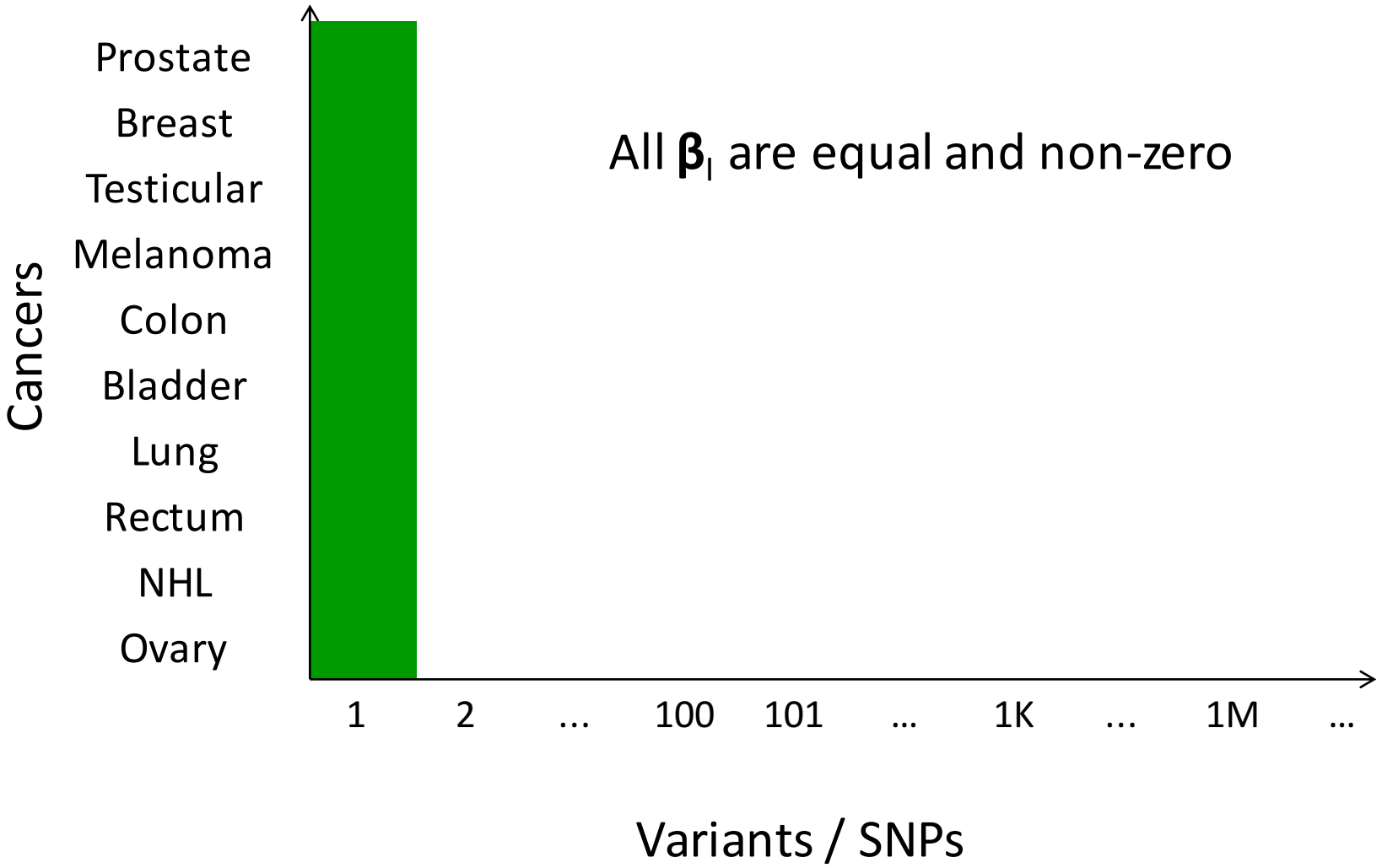


H_0: $\beta_1=\beta_2=...=\beta_k=0$

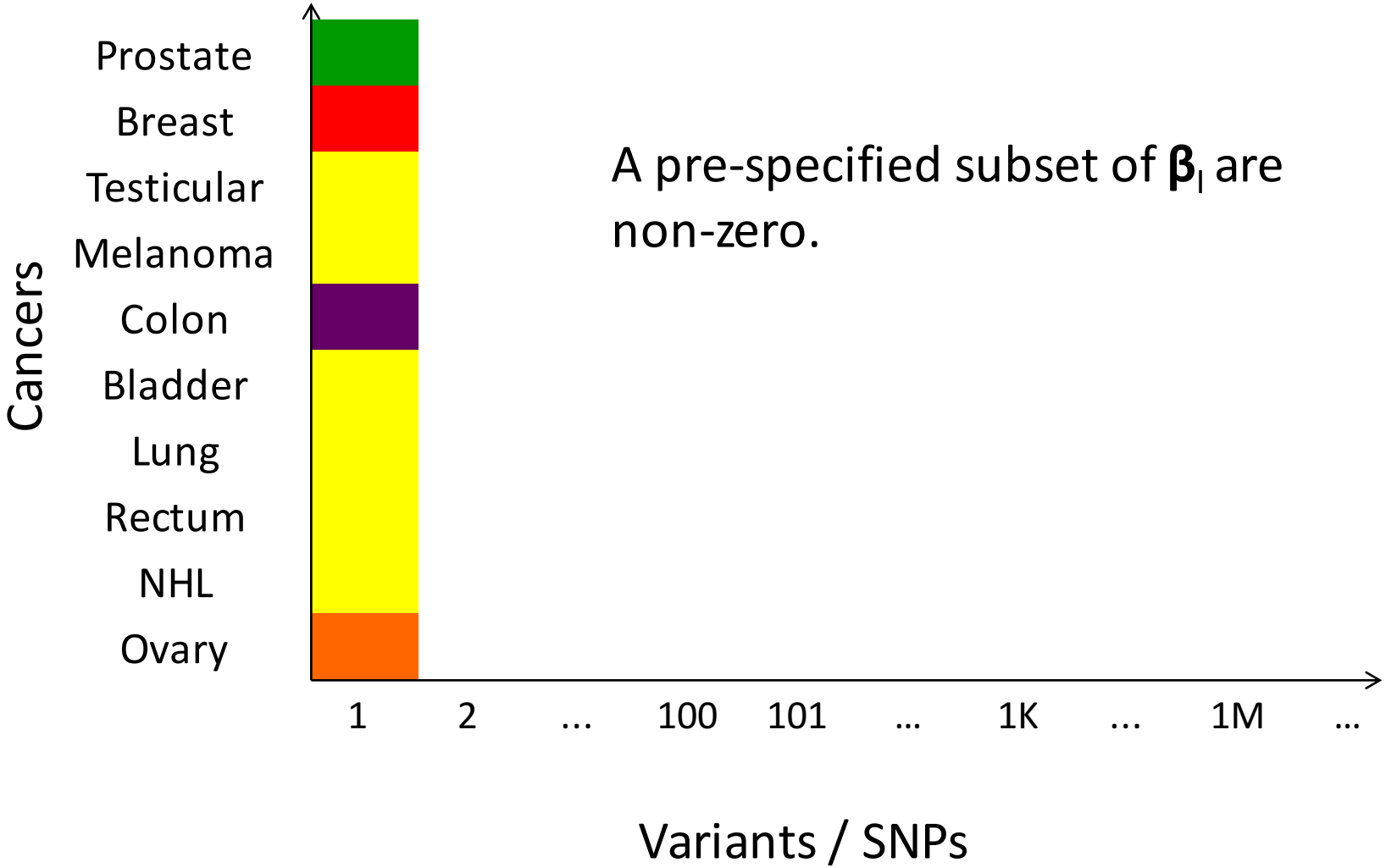
Alternative Hypotheses



Alternative Hypotheses

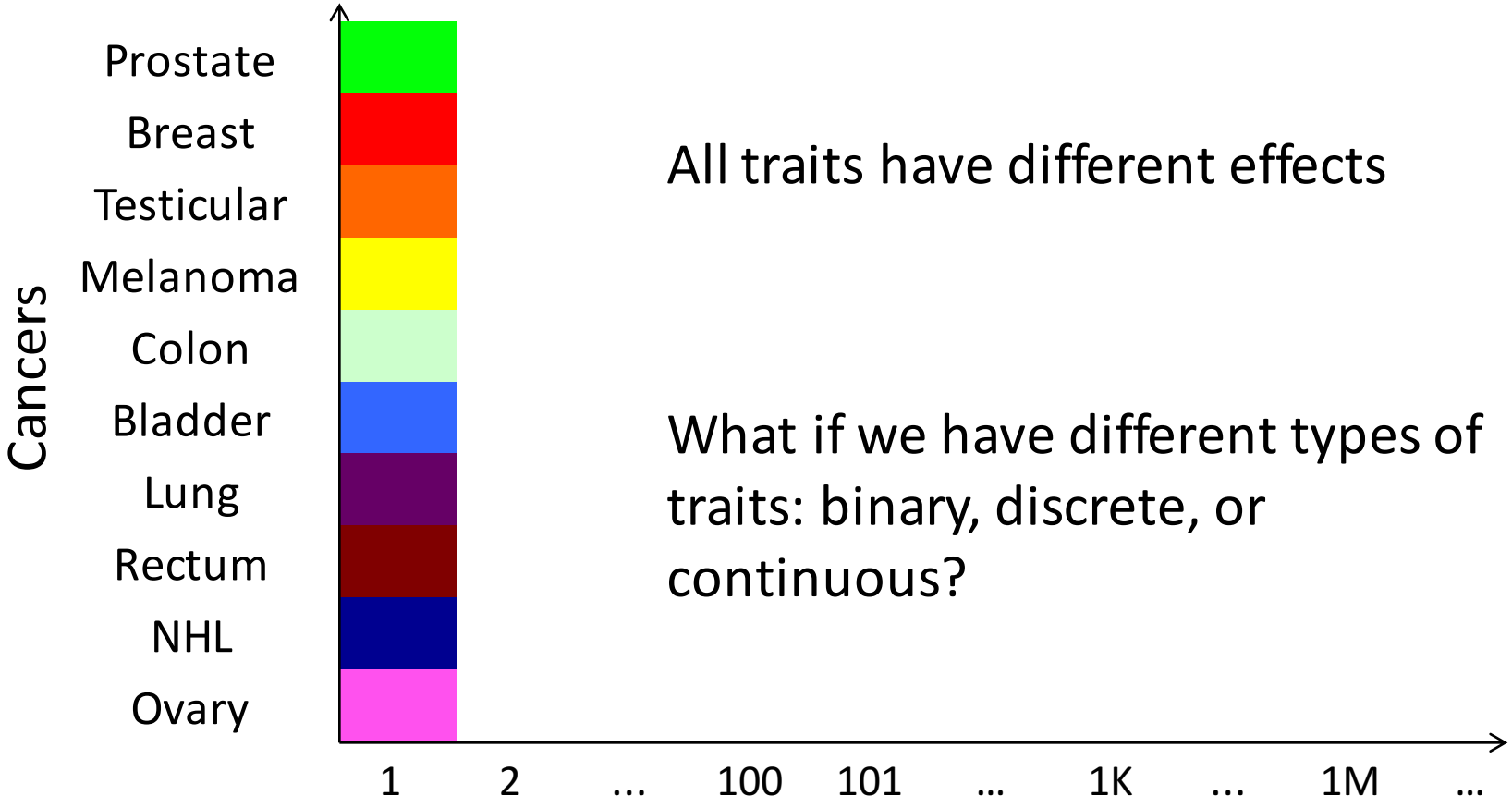


Alternative Hypotheses



A pre-specified subset of $\boldsymbol{\beta}_l$ are non-zero.

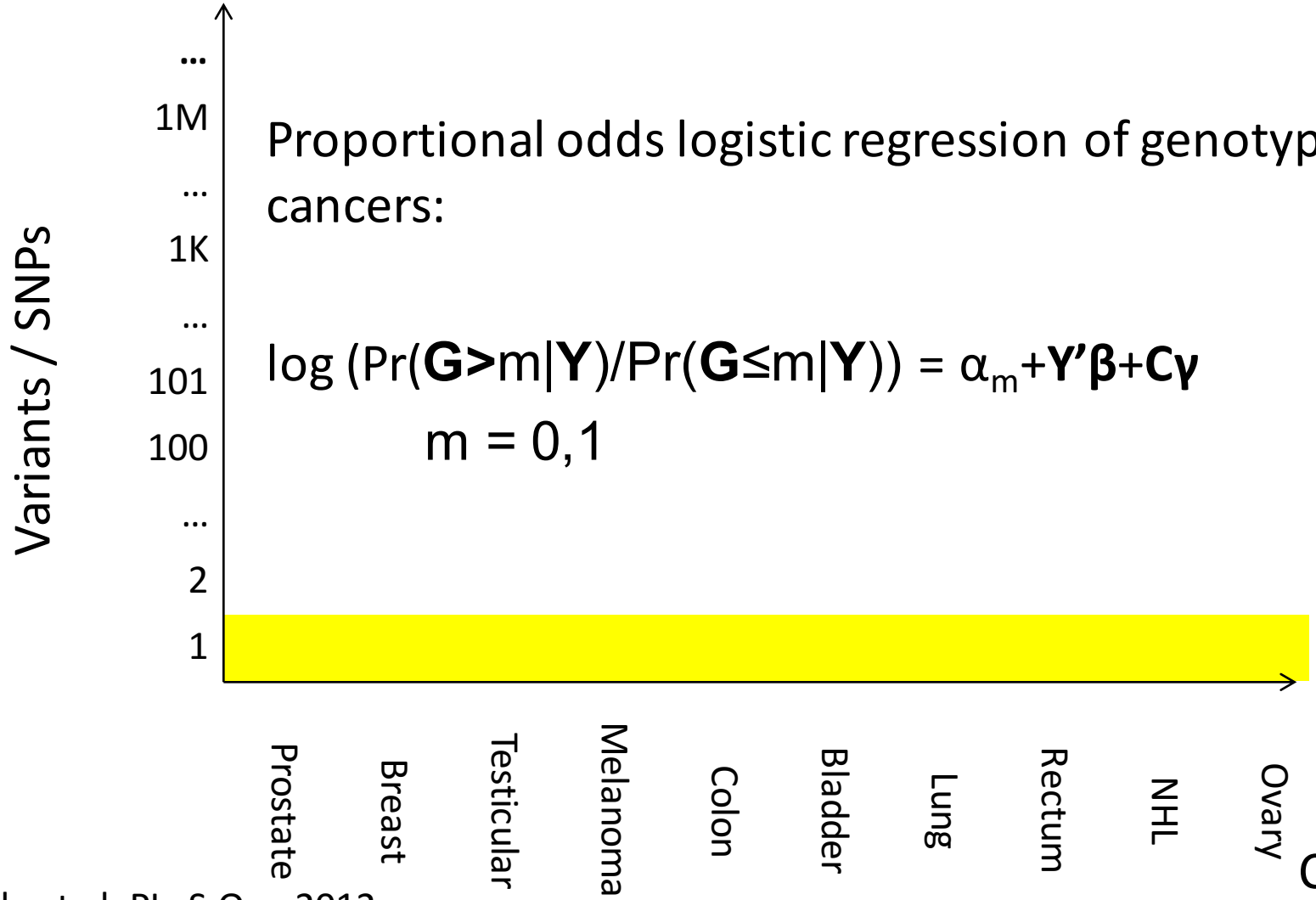
Alternative Hypotheses



All traits have different effects

What if we have different types of traits: binary, discrete, or continuous?

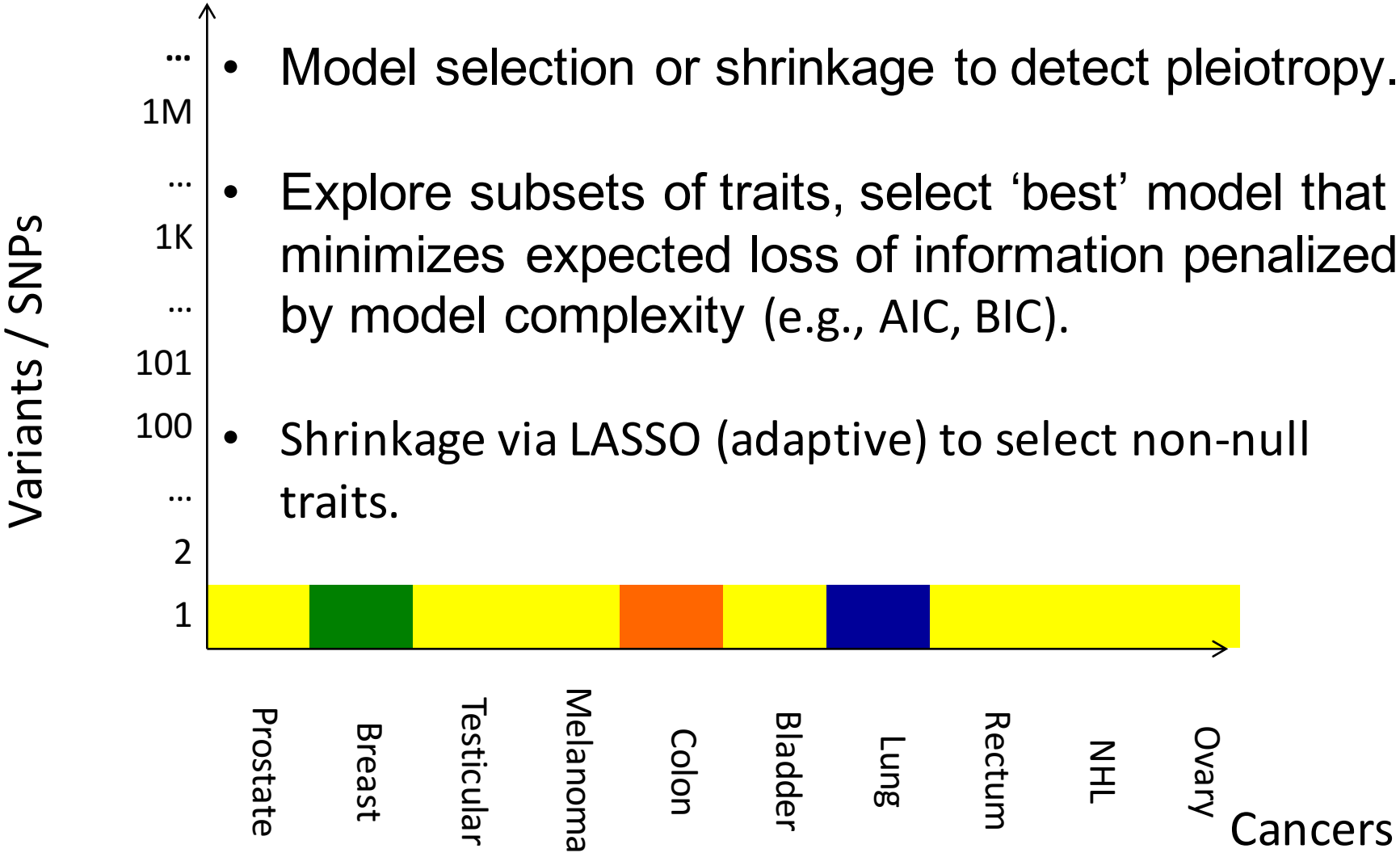
MultiPhen: 'Inverse Regression'



O'Reilly et al.,PLoS One 2012

MultiPhen: 'Inverse Regression'

- Model selection or shrinkage to detect pleiotropy.
- Explore subsets of traits, select 'best' model that minimizes expected loss of information penalized by model complexity (e.g., AIC, BIC).
- Shrinkage via LASSO (adaptive) to select non-null traits.



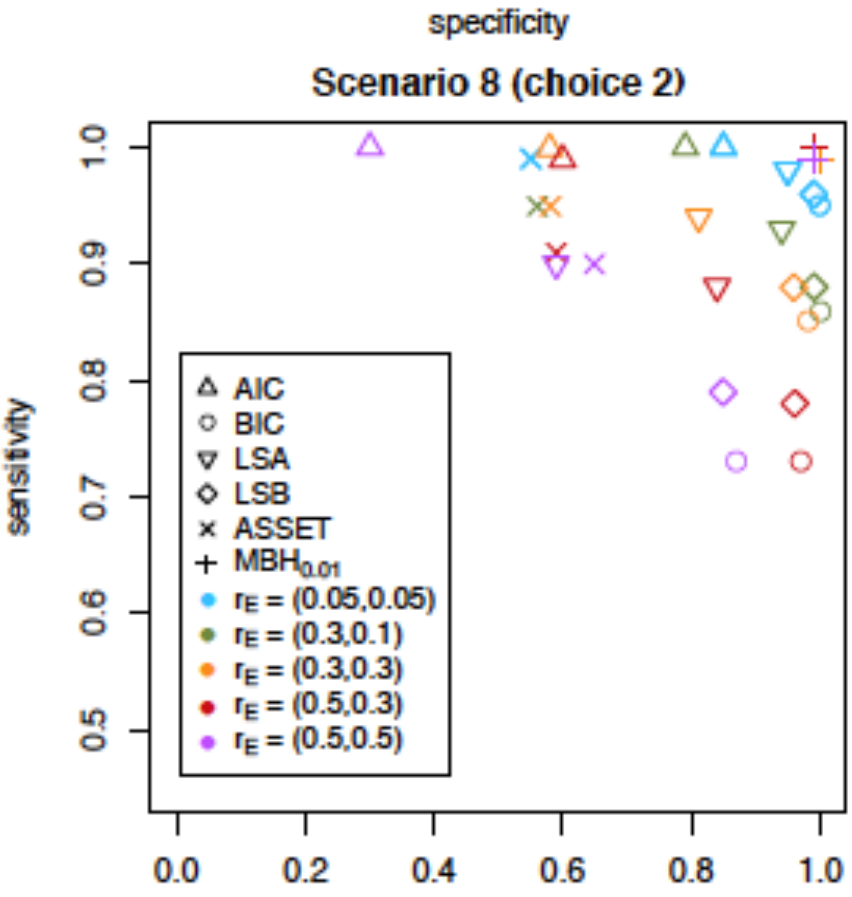
Comparison of FDR, ASSET & MultiPhen

- Traits simulated under additive model (Galesloot et al. 2014)
- Single causal variant with influence on a subset of traits.

- Number of traits: 4 - 20
- Residual correlation between traits: 0.05 - 0.3
- Heritability of trait due to variant: 0 - 0.4%
- LD between variant and typed SNP: 0.80, 0.95
- MAF at variant and SNP: 0.1, 0.2
- Number of individuals: 10K - 30K

Majumdar, Haldar, Witte, Genetic Epi 2016.

Results



Overall pleiotropy:

MultiPhen > ASSET (power)

- Except when all traits associated & in same direction.

Increases with increasing correlation.

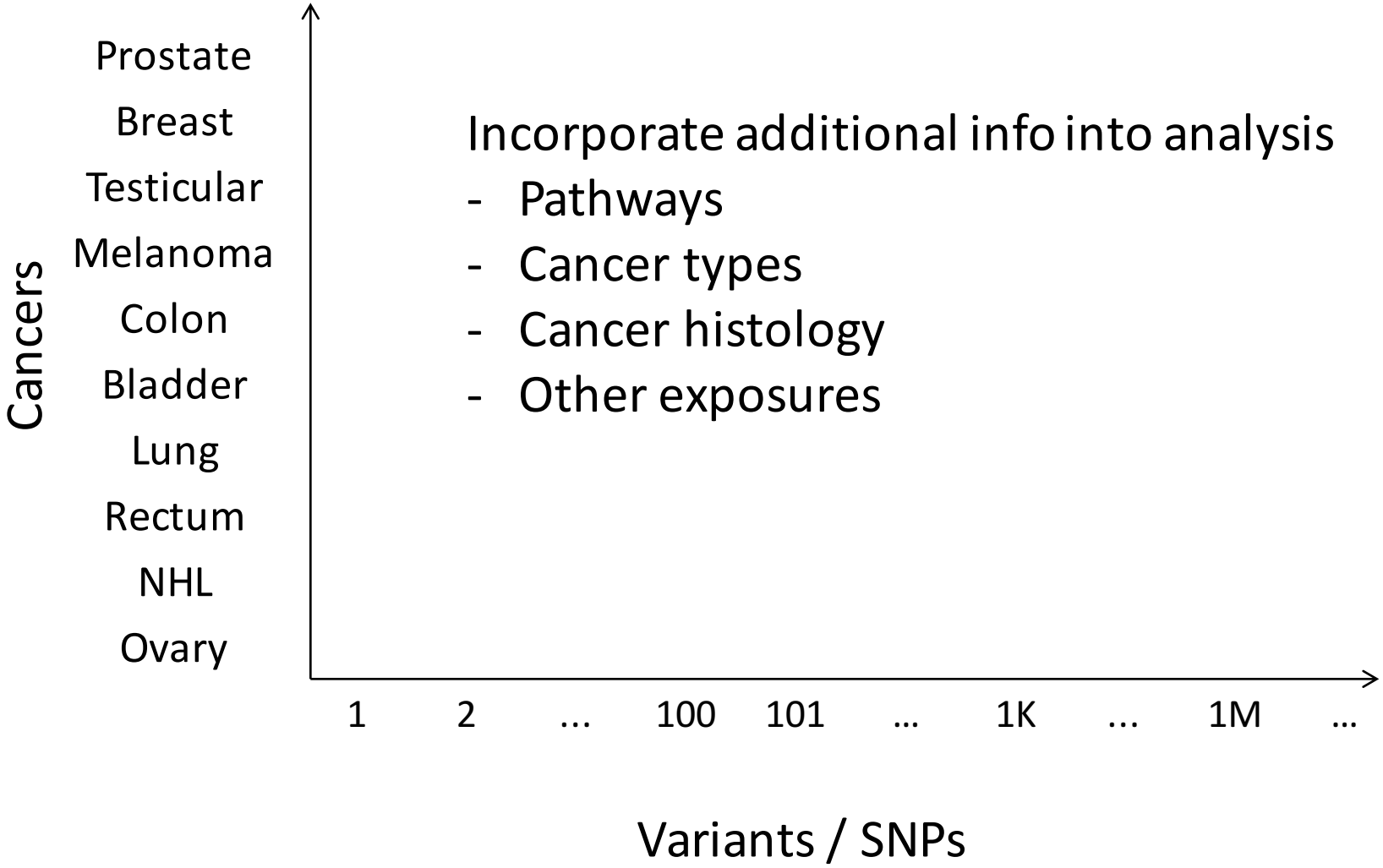
Traits underlying pleiotropy:

FDR > MultiPhen > ASSET (sens /spec)

- Except FDR = MultiPhen when weak correlation
- MultiPhen = ASSET when strong correlation.

Majumdar, Haldar, Witte, Genetic Epi 2016.

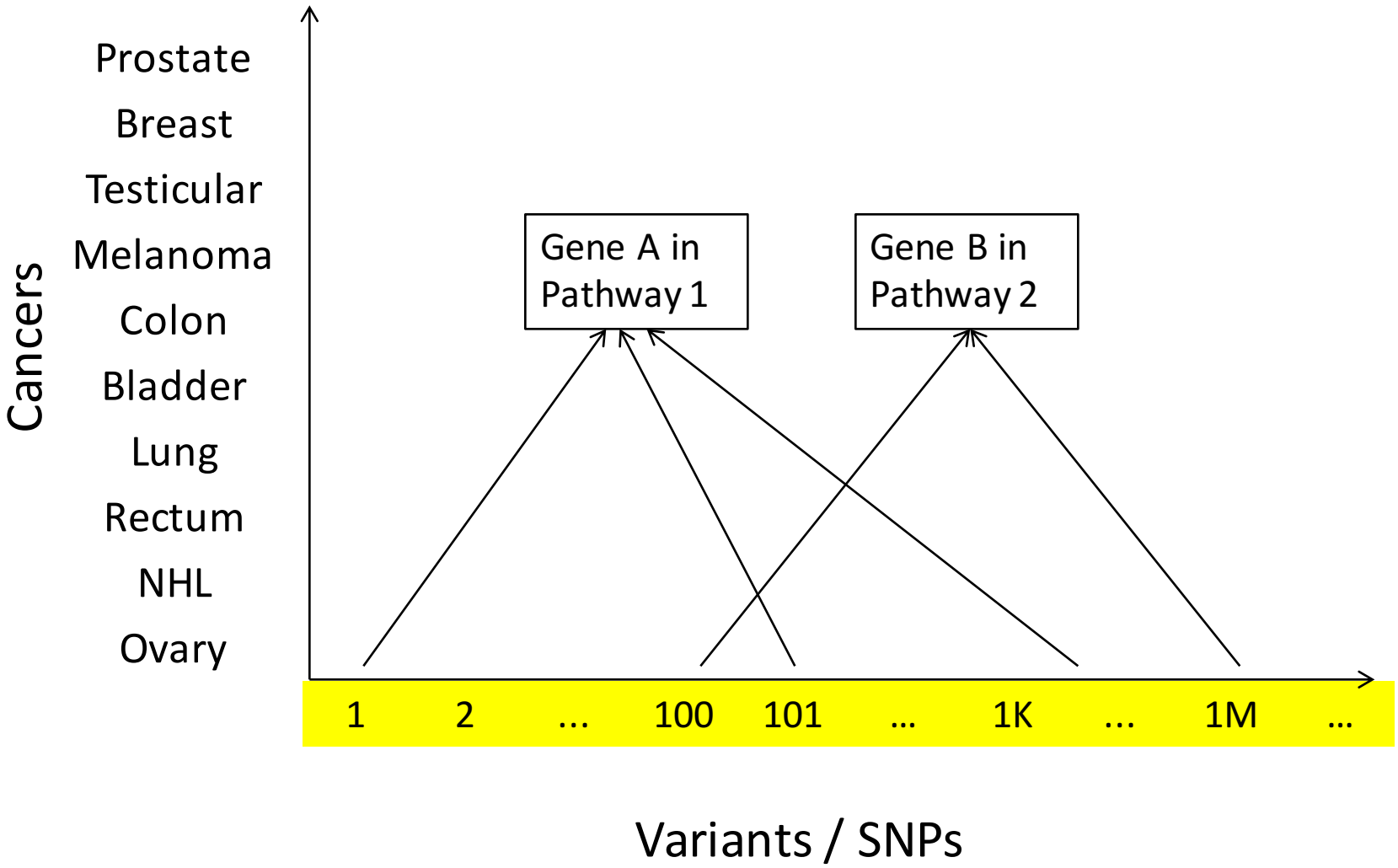
4. Multilevel Pleiotropy



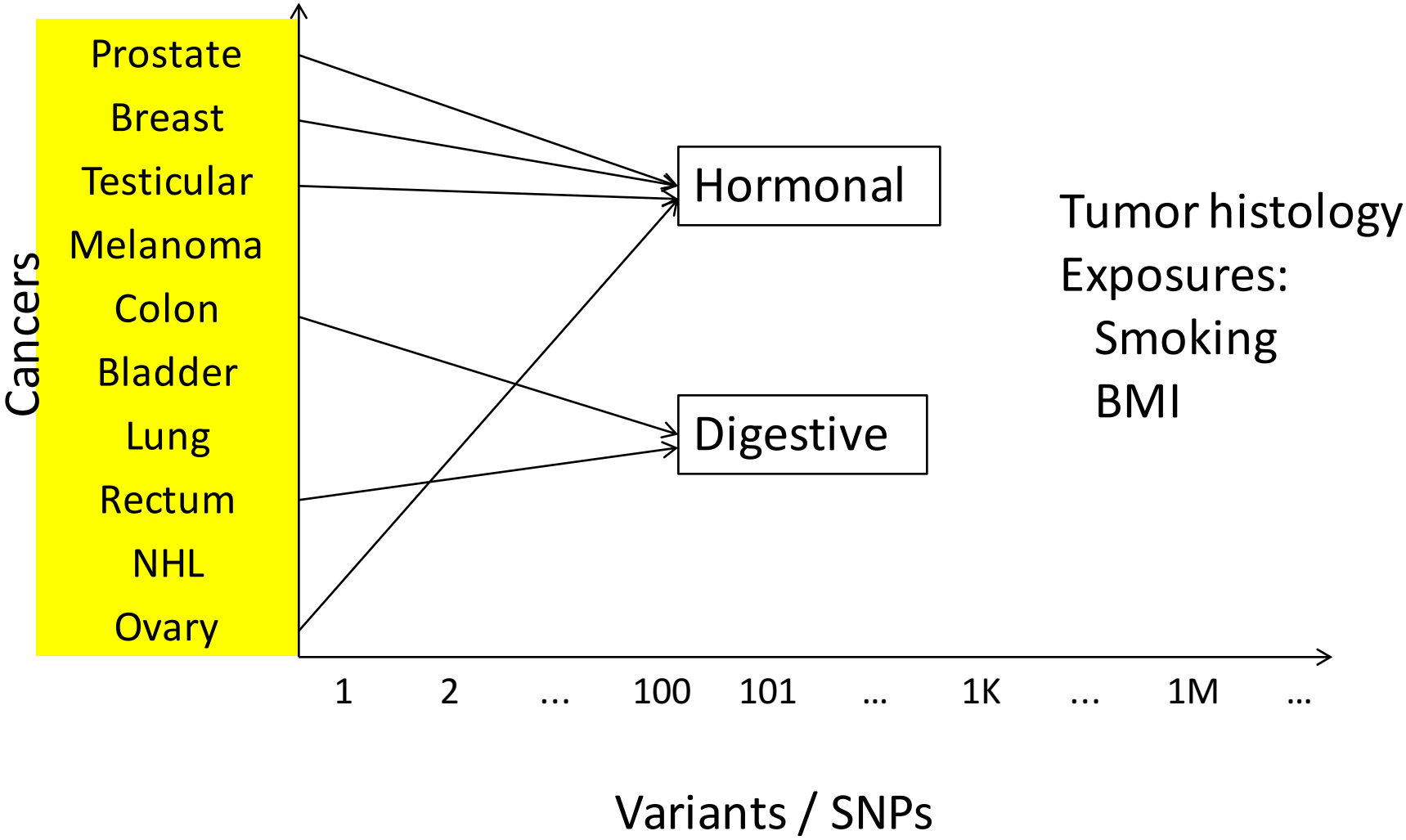
Incorporate additional info into analysis

- Pathways
- Cancer types
- Cancer histology
- Other exposures

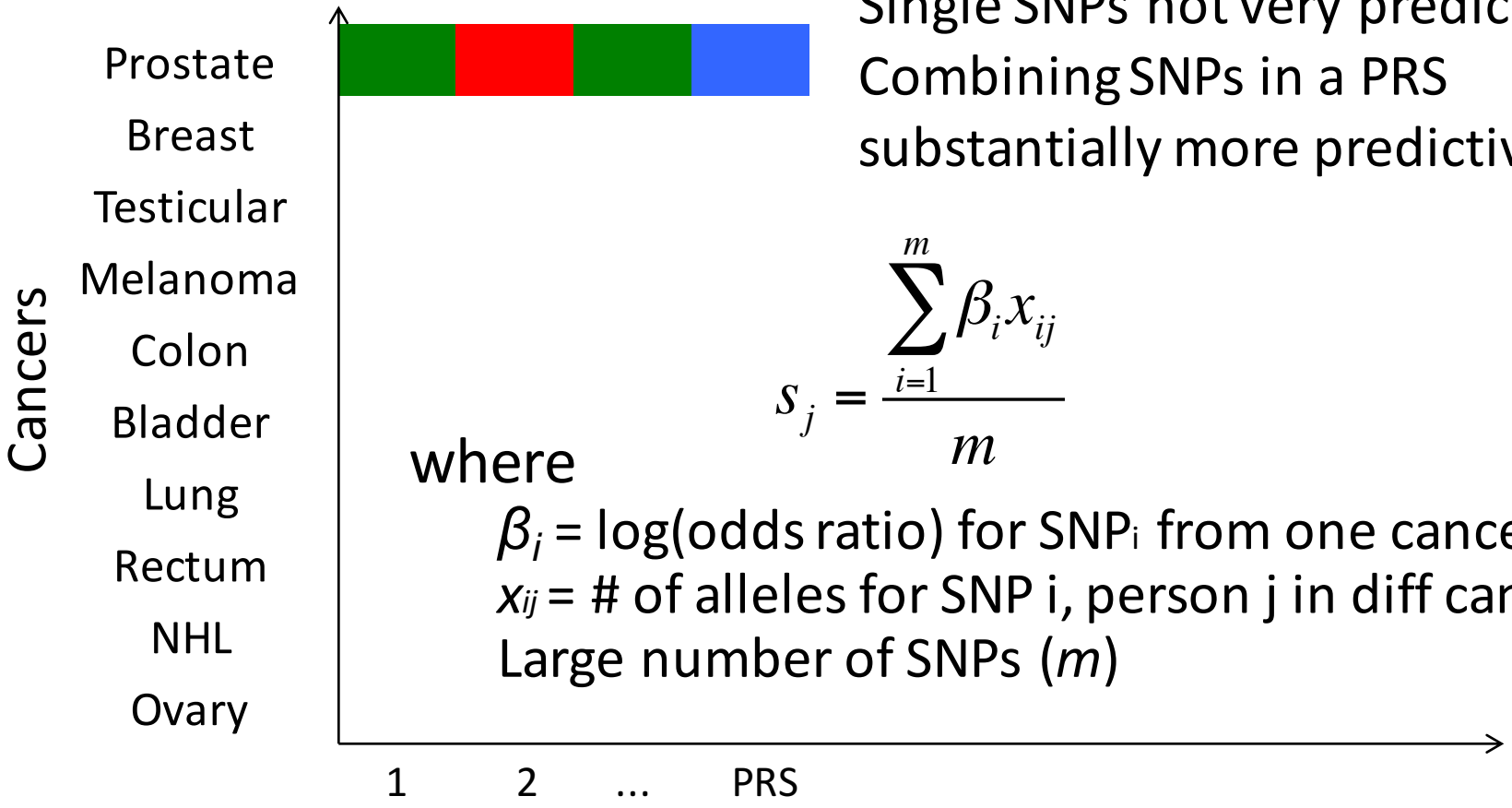
Gene / Pathway Priors



Leverage Individual-level Data



5. Polygenic Risk Scores (PRS)



Single SNPs not very predictive.
Combining SNPs in a PRS substantially more predictive.

$$s_j = \frac{\sum_{i=1}^{m} \beta_i x_{ij}}{m}$$

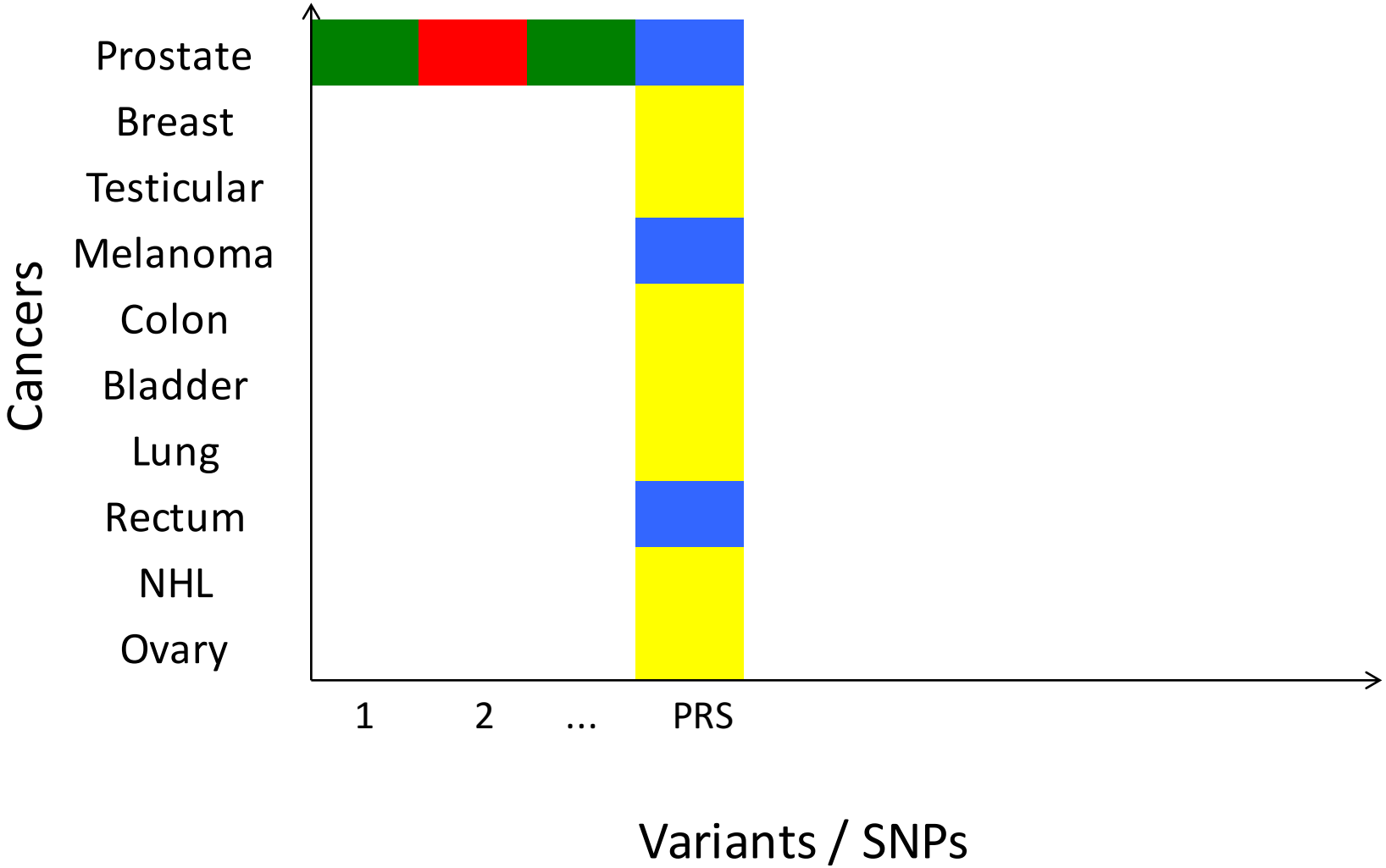
where

β_i = log(odds ratio) for SNP_i from one cancer.

x_{ij} = # of alleles for SNP i, person j in diff cancer.

Large number of SNPs (m)

Polygenic Risk Score Pleiotropy



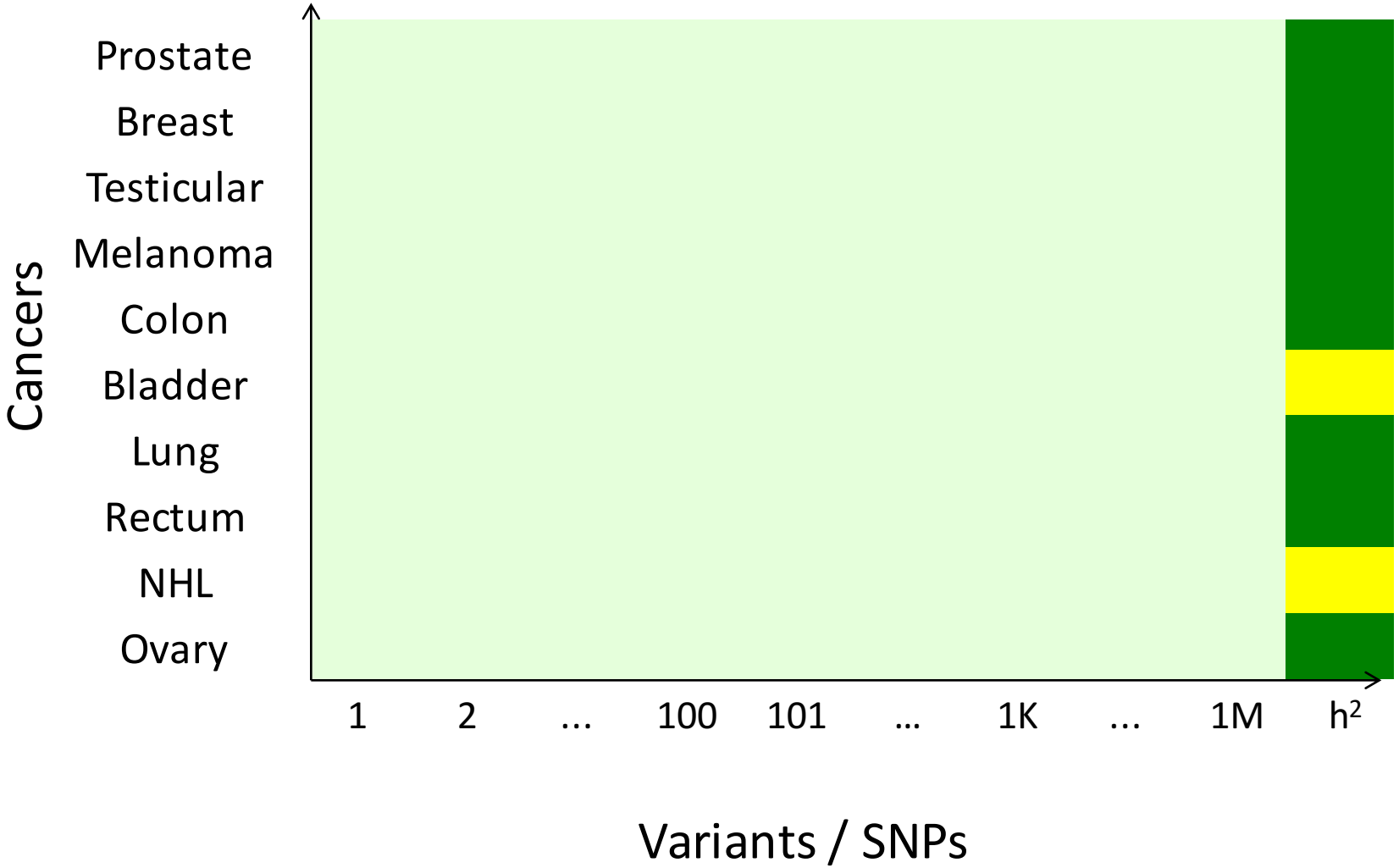
PanCancer PRS in UK Biobank

Target Cancer	Risk Score Profile: Bladder	Breast	Colorectal	Endometrial	Esophagus/Gastric	Lung	Prostate	Testicular
Bladder	-	0.008	0.20	0.27	0.17	0.09	0.67	0.03
Colon	0.18	0.81	-	0.74	0.59	0.97	0.01	0.01
Kidney	0.002	0.20	0.46	0.34	0.23	0.07	0.81	0.23
Lung	0.39	0.66	0.46	0.54	0.03	-	0.31	0.51
Melanoma	0.32	0.99	0.40	0.04	0.39	0.99	0.54	0.70
Prostate	0.21	0.63	0.003	0.00001	0.27	0.47	-	0.02
Rectum	0.72	0.89	-	0.79	0.01	0.01	0.94	0.61
Testicular	0.02	0.23	0.75	0.0002	0.57	0.21	0.10	-

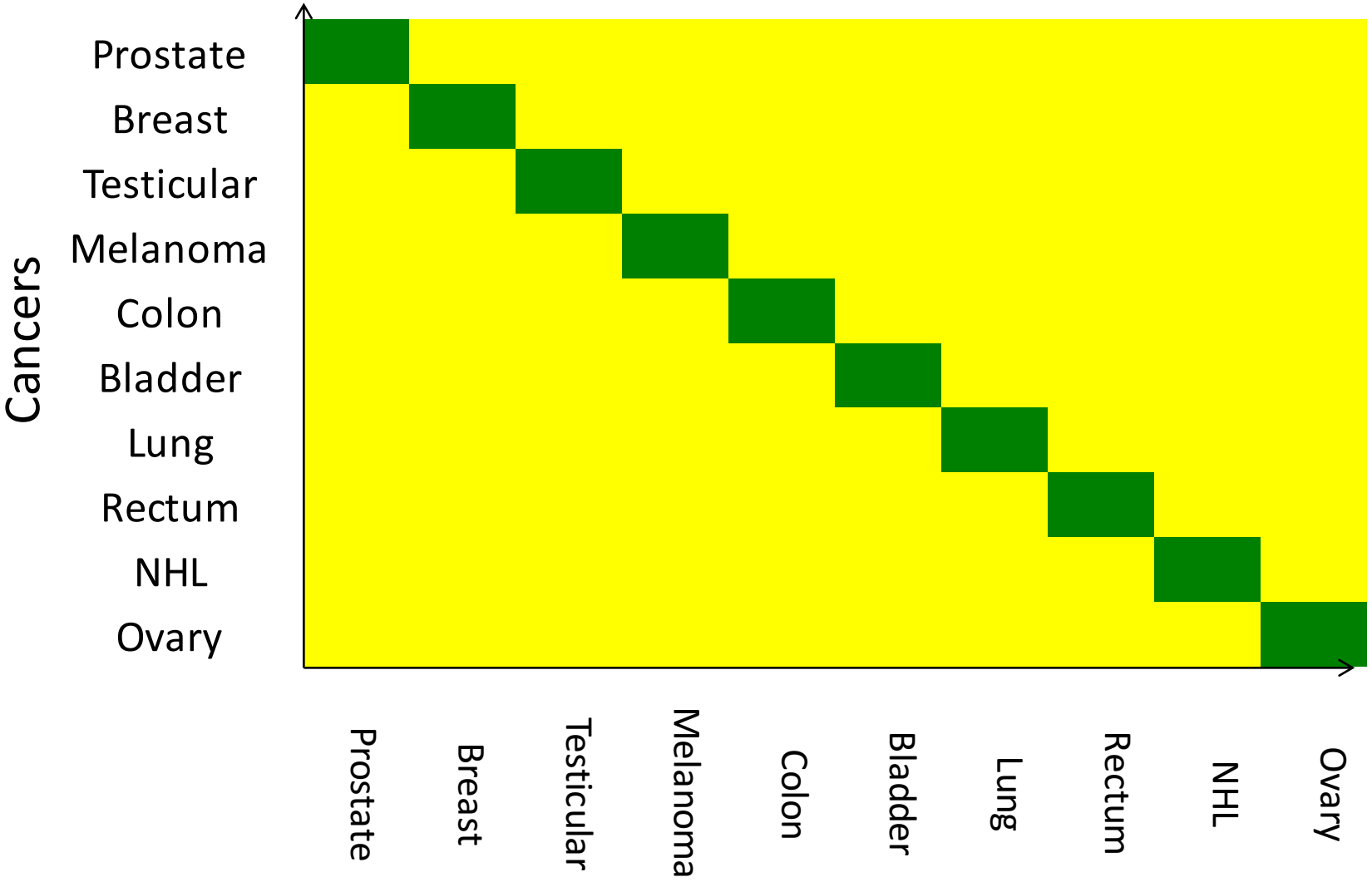
Negative Association

Positive Association

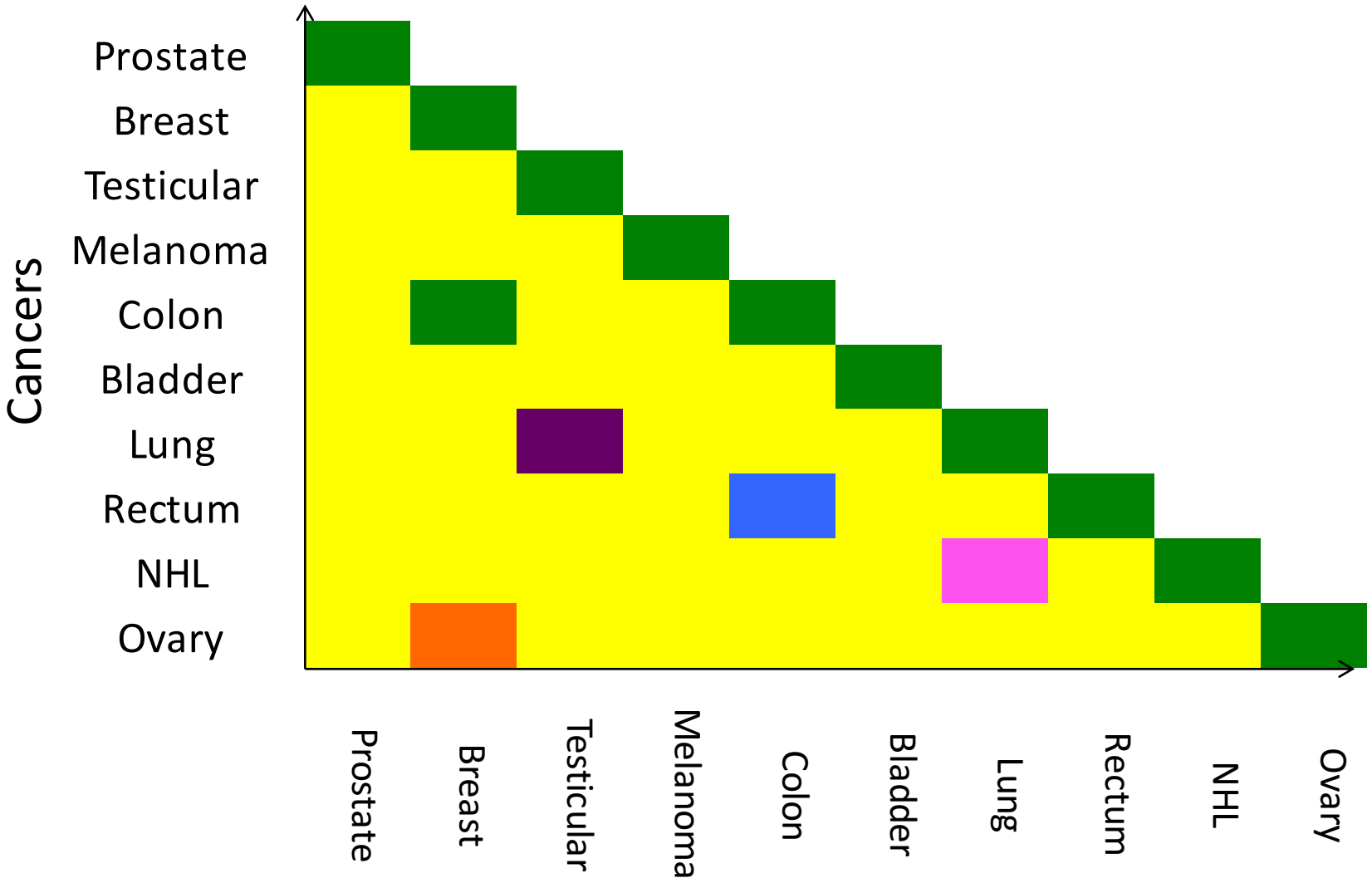
6. Co-heritability



Co-heritability

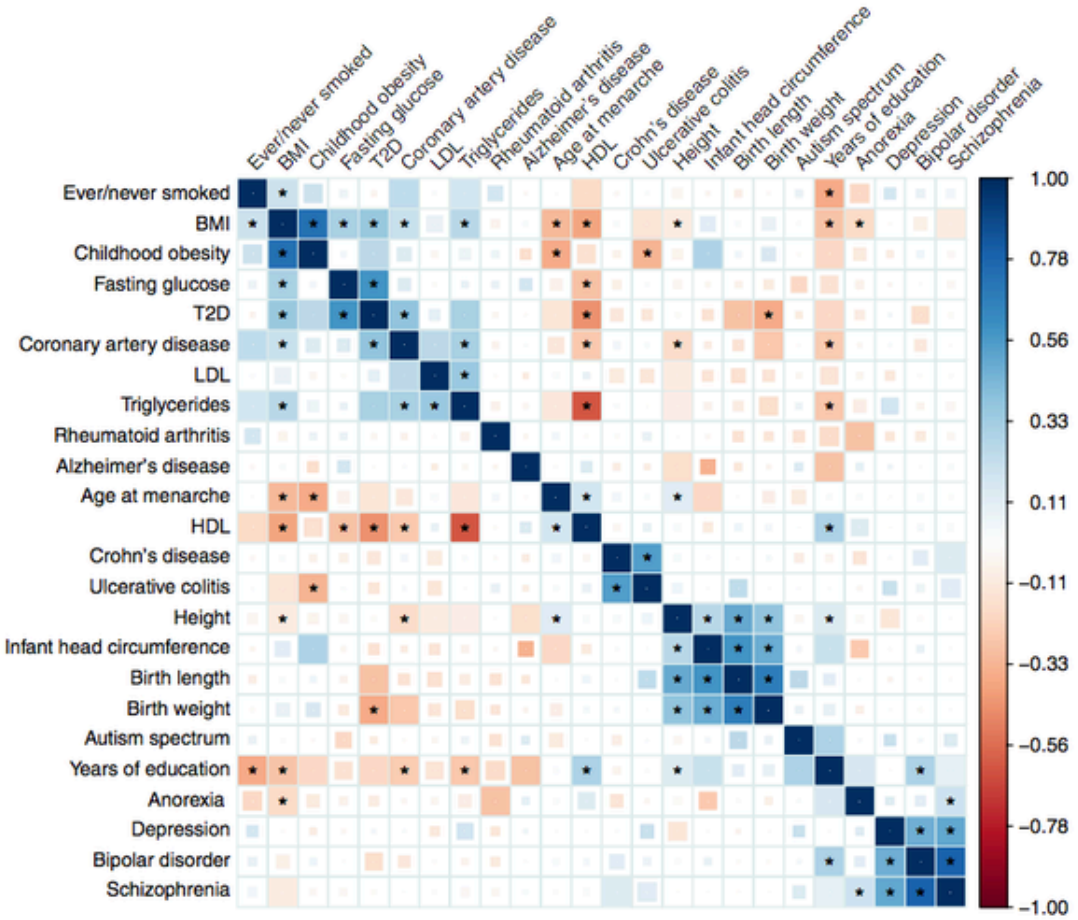


Co-heritability



Co-heritability with Summary Statistics

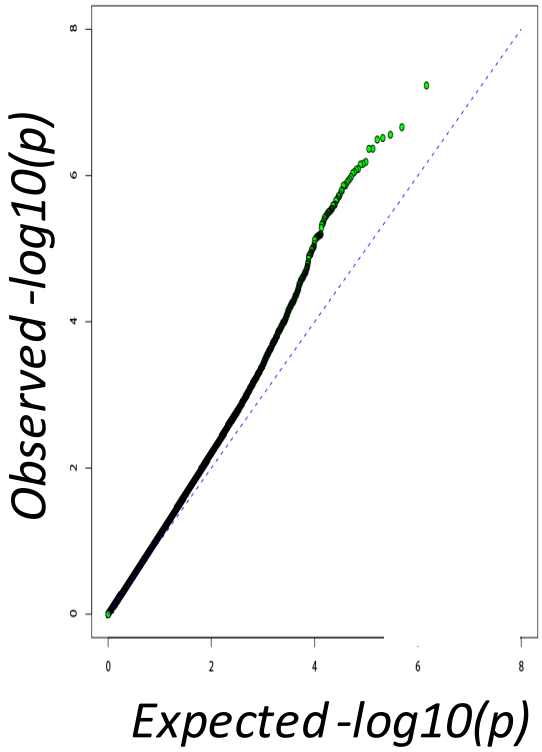
Cross-trait LD Score Regression



Bulik-Sullivan, et al. NG 2015b

LD Score:
Distribution of Associated SNPs

QQ-Plot



If a proportion of SNPs associated:
observed = expected
(median test statistic)

If observed > expected:
genomic inflation

Due to population stratification?

Yang et al (2011). EJHG

Genomic Inflation Expected under Polygenic Inheritance

Under null hypothesis:

- Mean test statistic (λ_{mean}) = median test statistic (λ_{median})

Under polygenic inheritance (no population stratification):

$\lambda_{mean} > \lambda_{median}$

λ_{mean} reflects SNP heritability h_g^2

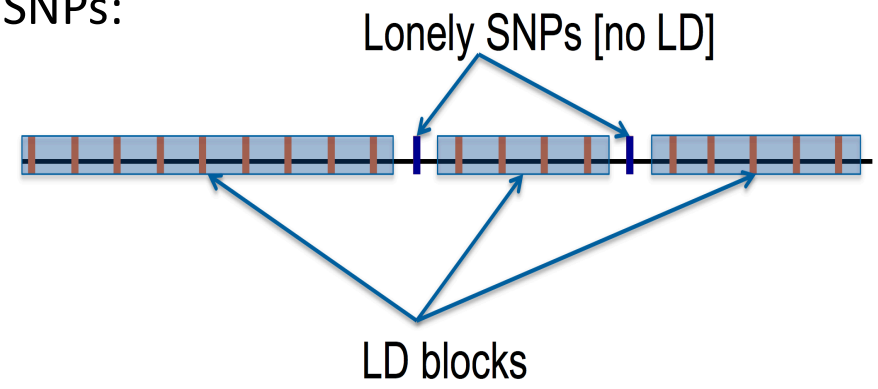
λ_{median} reflects # causal variants contributing to h_g^2

Controlling for genomic inflation may remove both pop strat and real effects. How to tell them apart?

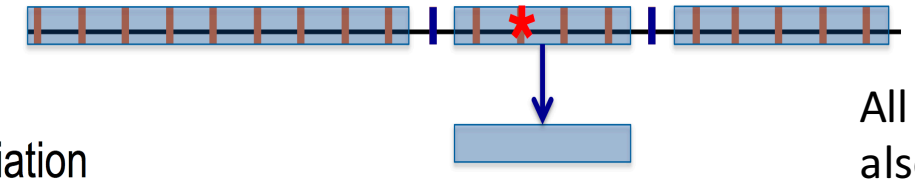
Yang et al (2011). EJHG

Impact of LD on Association

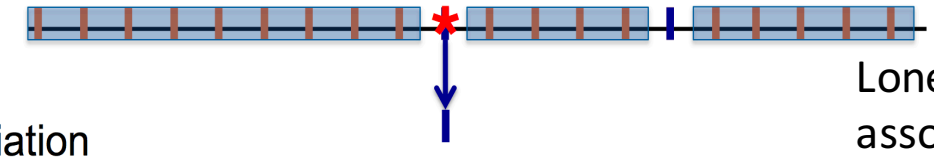
LD among SNPs:



Consider causal SNPs



All in LD with causal SNP also associated



Lonely SNPs only associated if causal

- More tagging of SNPs, more likely to tag a causal variant.
- If all SNPs equally likely associated given LD status, expect more association for SNPs with more LD 'friends'.
- This is a reasonable assumption under a polygenic genetic architecture.

Slide: Ben Neale

Expected Value of Summary Stats

$$\lambda_{mean} = E\left[\chi_j^2\right] = 1 + \frac{Nh_g^2}{M} l_j$$

- N: Sample size
- h_g^2: SNP heritability
- l_j: LD Score of SNP j: amount of genetic variation tagged by j.
- M: Number of SNPs

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

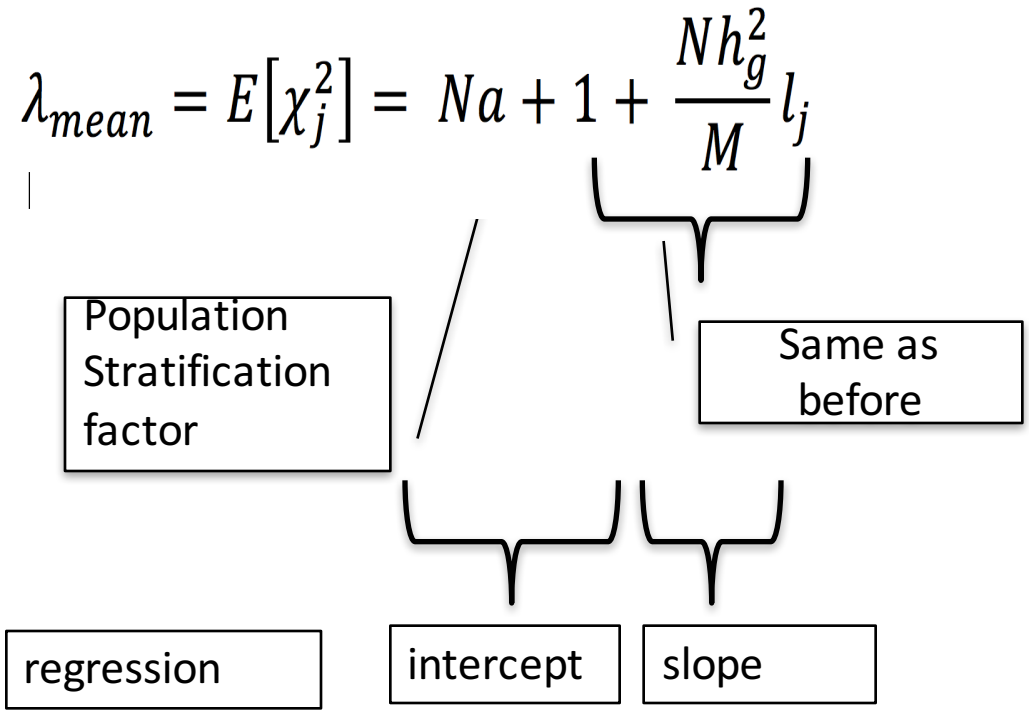
LD Score:
r^2 LD between SNP j and neighboring SNPS

But can't separate out population stratification here.

Bulik-Sullivan, et al. NG 2015

Expected Value of Summary Stats

Separating h_g^2 and population stratification

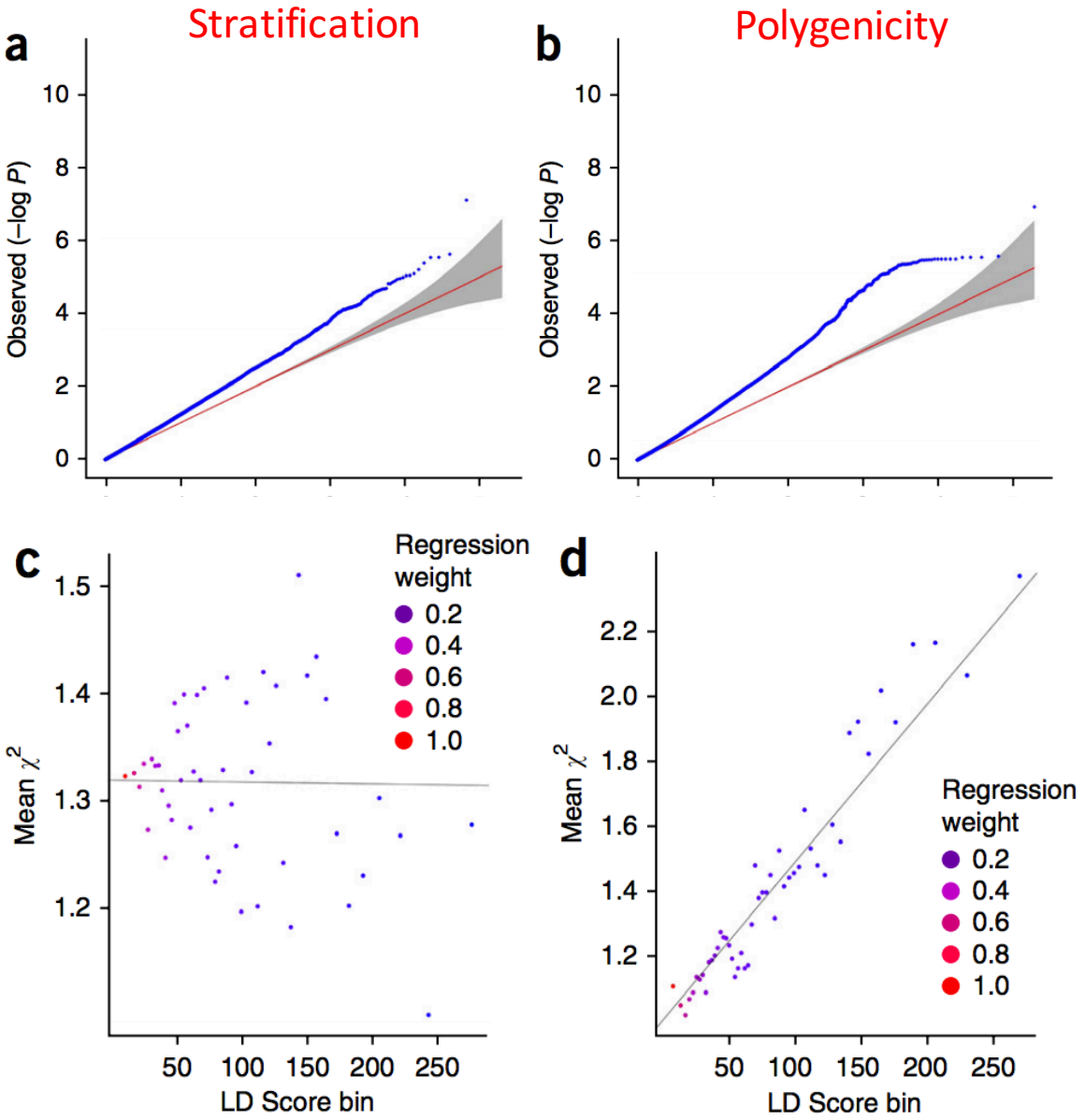


Bulik-Sullivan, et al. NG 2015

Polygenicity vs Population Stratification

GWAS QQ Plot

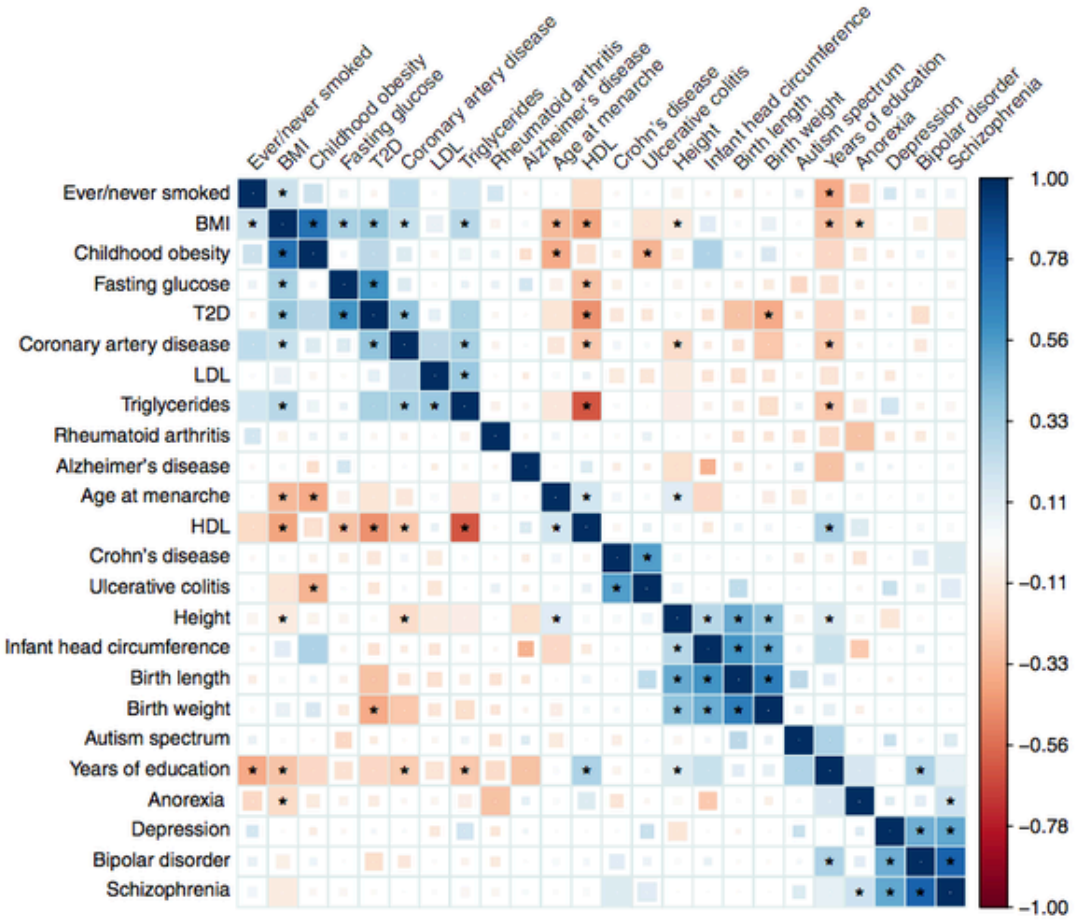
Regression of association X^2 statistic on LD score



Bulik-Sullivan, et al. NG 2015

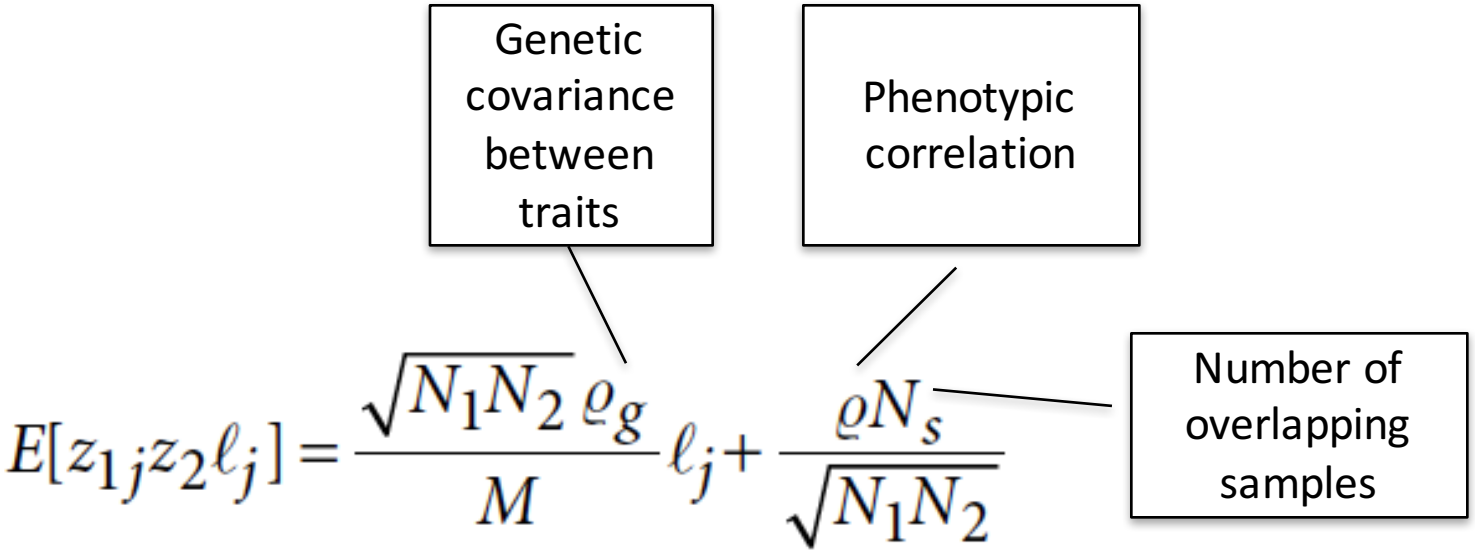
Co-heritability with Summary Statistics

Cross-trait LD Score Regression



Bulik-Sullivan, et al. NG 2015b

Cross-trait LD-score Regression



- Use LD Score to estimate the genetic correlations between diseases with summary statistics

Bulik-Sullivan, et al. NG 2015

2016 Module 18:
Statistical & Quantitative Genetics of Disease

Lecture 9
Tying up some ends on risk prediction

Naomi Wray

Aims of Lecture 9

1. Variance explained by genetic factors
2. Factors affecting accuracy of risk prediction
3. Pitfalls of Risk Prediction

Variance explained by genetic factors

Definition of heritabilities

Proportion of variance attributable to genetic factors

From family data and phenotypic records h^2
Proportion of variance attributable to genetic factors accounting for all genetic variants across the frequency spectrum (Lecture 1: Wray)

From genome-wide significant SNP h^2-i
Proportion of variance attributable to a single variant (Lecture 2: Witte)

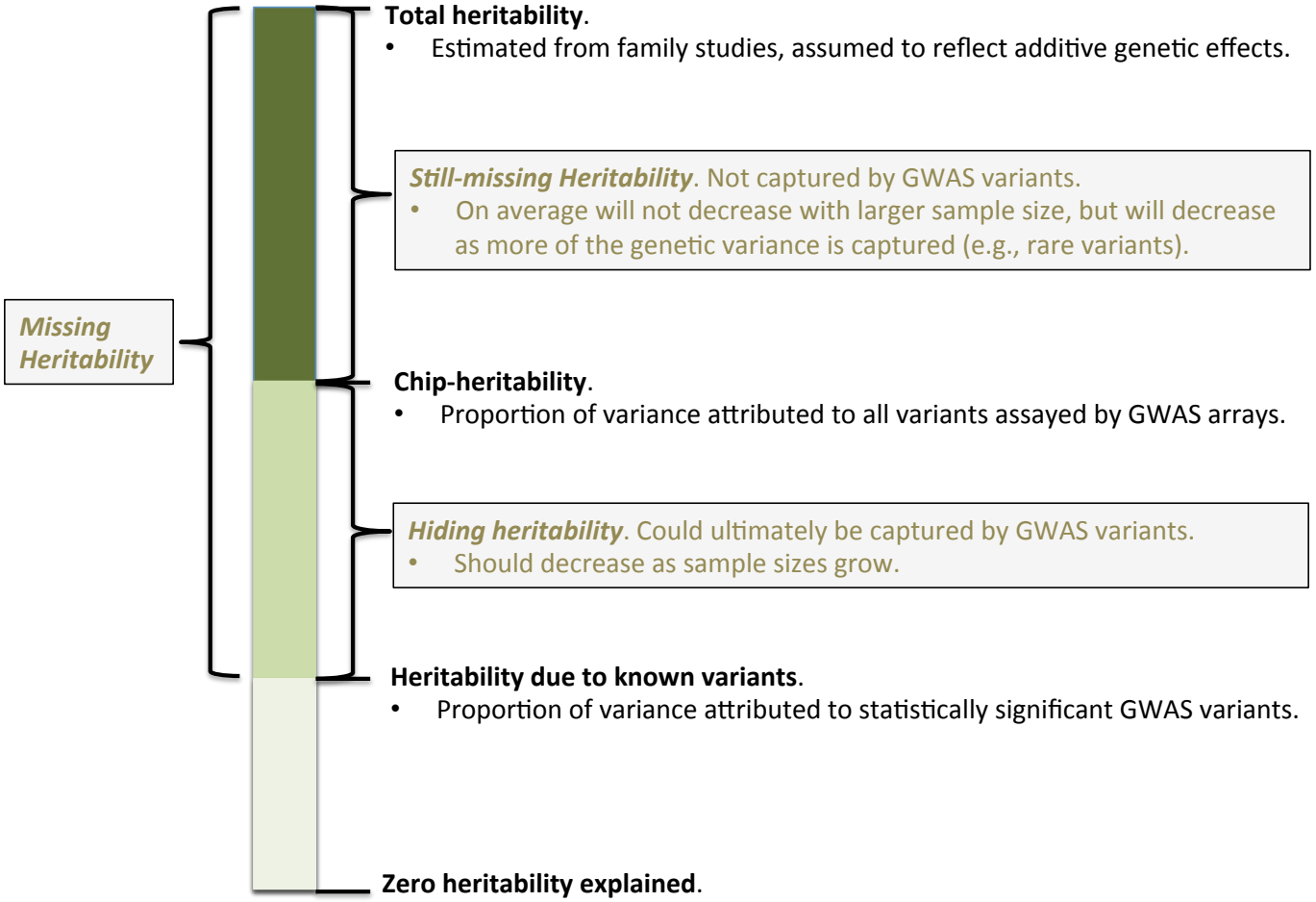
From genome-wide significant SNP h^2-GWS
Proportion of variance attributable to genome-wide significant SNPs

From genome-wide significant SNP h^2-profile score
Proportion of variance attributable to a set SNPs (Lecture 7: Wray)

From all SNPs h^2-SNP or h^2-chip or h^2-g
Proportion of variance attributable to common SNPs on SNP chips (GREML; LDScore; Lecture 8: Witte)

The heritabilities

Figure 4



Witte, Visscher, Wray (2014) The contributions of genetic variants to disease depends on the ruler. Nature Genetics

Variance explained by sets of SNPs e.g., all GWS SNPs

If independent loci then simply sum up estimates from individual SNPs

If not independent need to use set based tests

- Set-based test (--sbat)
  - Using GWAS summary data
  - Similar as PLINK --set or VEGAS but more accurate and faster
  - Working on more powerful improvements
- Can be used for discovery
  - e.g. gene-based tests (genome-wide)
- Can be used to test prior hypotheses
  - e.g. do all known together explain more variation than expected by chance?

http://www.cnsgenomics.com/software/

h^2-SNP or h^2-chip or h^2g

Purpose:

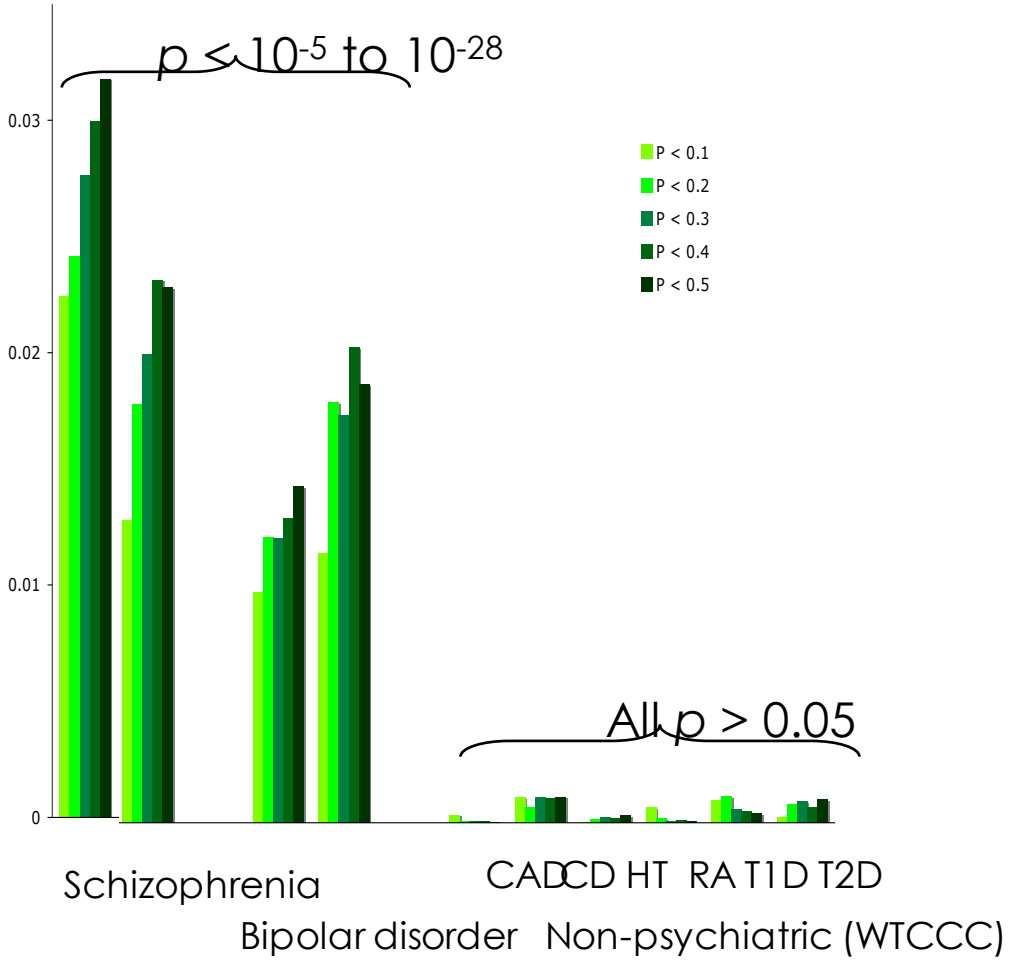
- *Detects the signal contributed from variants that are not genome-wide significant*
- *Detects if cases are more similar to other cases genome-wide than they are to controls without specifying at which loci there are more similar*
- *Gives an indication of what could be detected as GWS as sample size increases*
- *Uses data available from currently available data to inform on future experimental design*

h^2-SNP or h^2-chip or h^2g

Five ways to measure:

- Compare empirical results to simulations – Purcell et al (2009) Nature
- GREML – linear mixed models- GREML Yang et al (2010), Lee et al (2011)
  (Visscher & Goddard SISG Unit)
- Haseman-Elston or PCGC Haseman & Elston (1972), Chen (2014) Frontiers Gen, Golan et al (2014) PNAS
- Transformation of polygenic risk scores – Dudbridge (2013)
- LDscore –GWAS summary statistics - Yang et al (2011) EJHG, Bulik-Sullivan et al (2015) Nat Gen

h^2-SNP by simulation to explain these results



Purcell et al (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature

Simulations

- Simulations considered a wide range of genetic architectures
  - Number risk loci, effect size
  - Which genetic architectures generated the same pattern of results we observed with the real data
- Results showed
  - Most genetic architectures were not consistent with empirical result
  - But many architectures were consisted with the empirical results
  - All consistent models pointed to h^2SNP= ~0.3 for schizophrenia

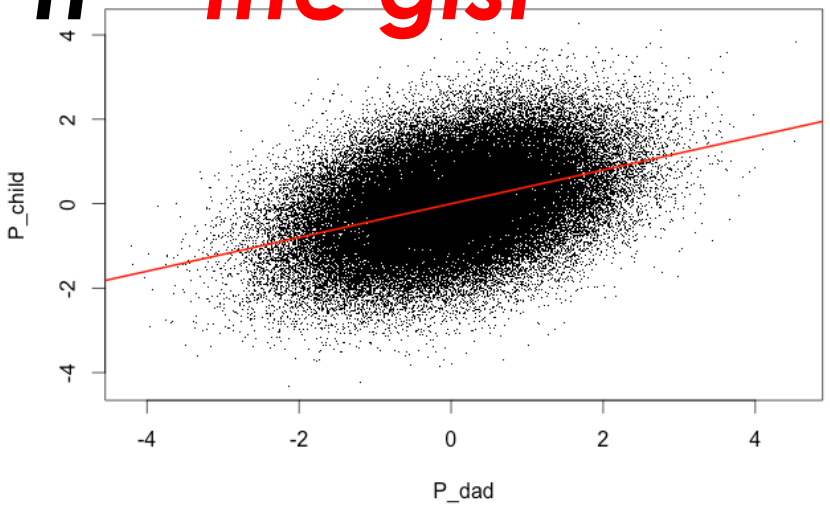
- Results not consistent with
  - Many extremely rare (MAF < 1/10,000) variant

Purcell et al (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature

Direct estimation of h^2-SNP parallels estimation of h^2– the gist

Coefficient of relationship of parent and offspring = ½

Estimate of heritability
= 2*correlation of offspring and parent



Coefficient of relationship of grand-parent and grand-offspring = ¼

Estimate of heritability
= 4*correlation of grandoffspring and grandparent

Coefficient of relationship of individual and distant relative= 1/r

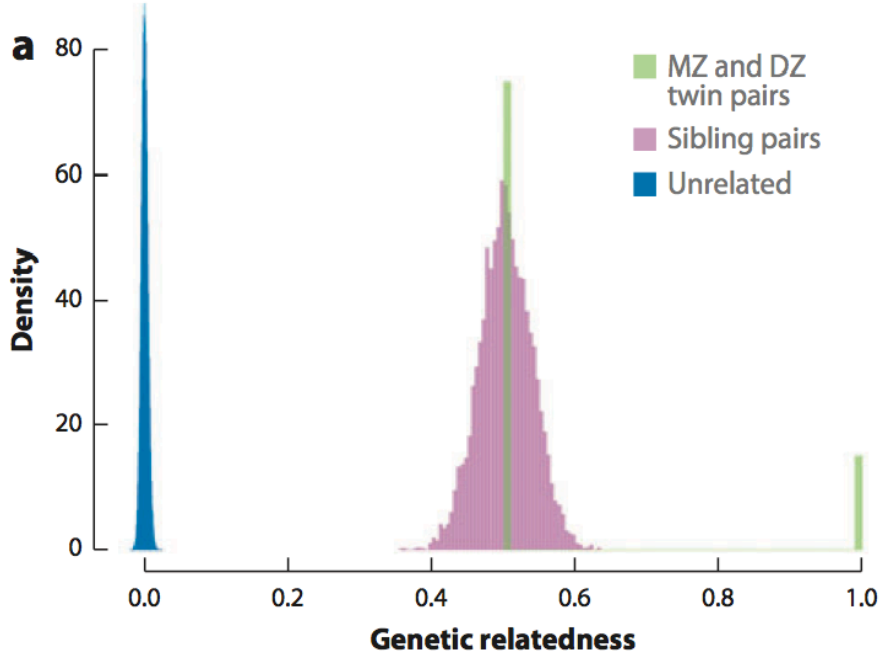
Estimate of heritability
= r*correlation between distant relatives on individuals

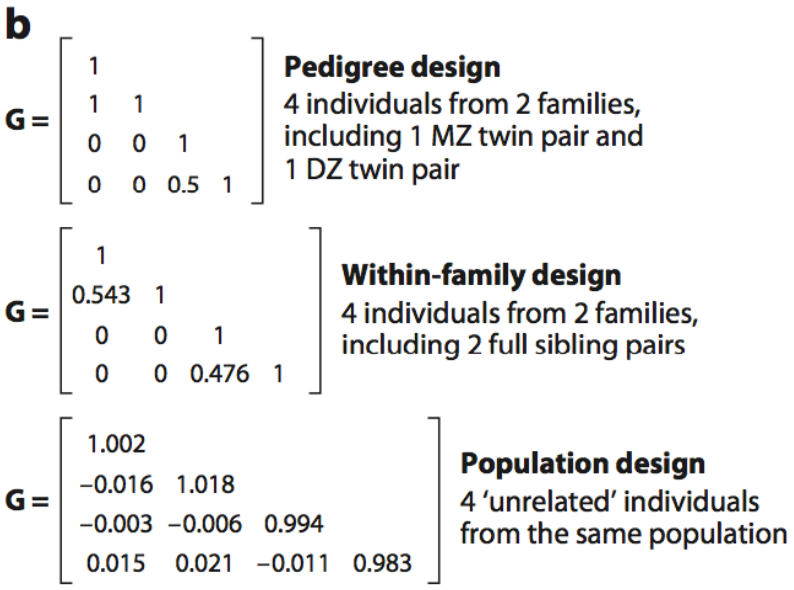
In real analysis twiddles and caveats

- But the gist of where the information is coming from

Yang et al (2010) Nature

Visualizing where the information is coming from





Vinkhuyzen et al (2013) Estimation and Partition of Heritability in Human Populations Using Whole-Genome Analysis Methods Annual reviews of genetics.

GREML: h^2-SNP

- Uses individuals who are unrelated in the classical sense
- Coefficient of relationship < 2nd cousins

How can we get an accurate estimate when coefficients of relationships are so tiny?

Because based on a very large number of pairwise relationships

Sample of 10,000 has ~$10{,}000^2/2$ = 50M

h^2 vs h^2-SNP

h^2
Proportion of variance attributable to all genetic variants – across whole frequency spectrum

Could be contaminated by non-additive genetic variance

Could be contaminated by environmental factors shared by close family members

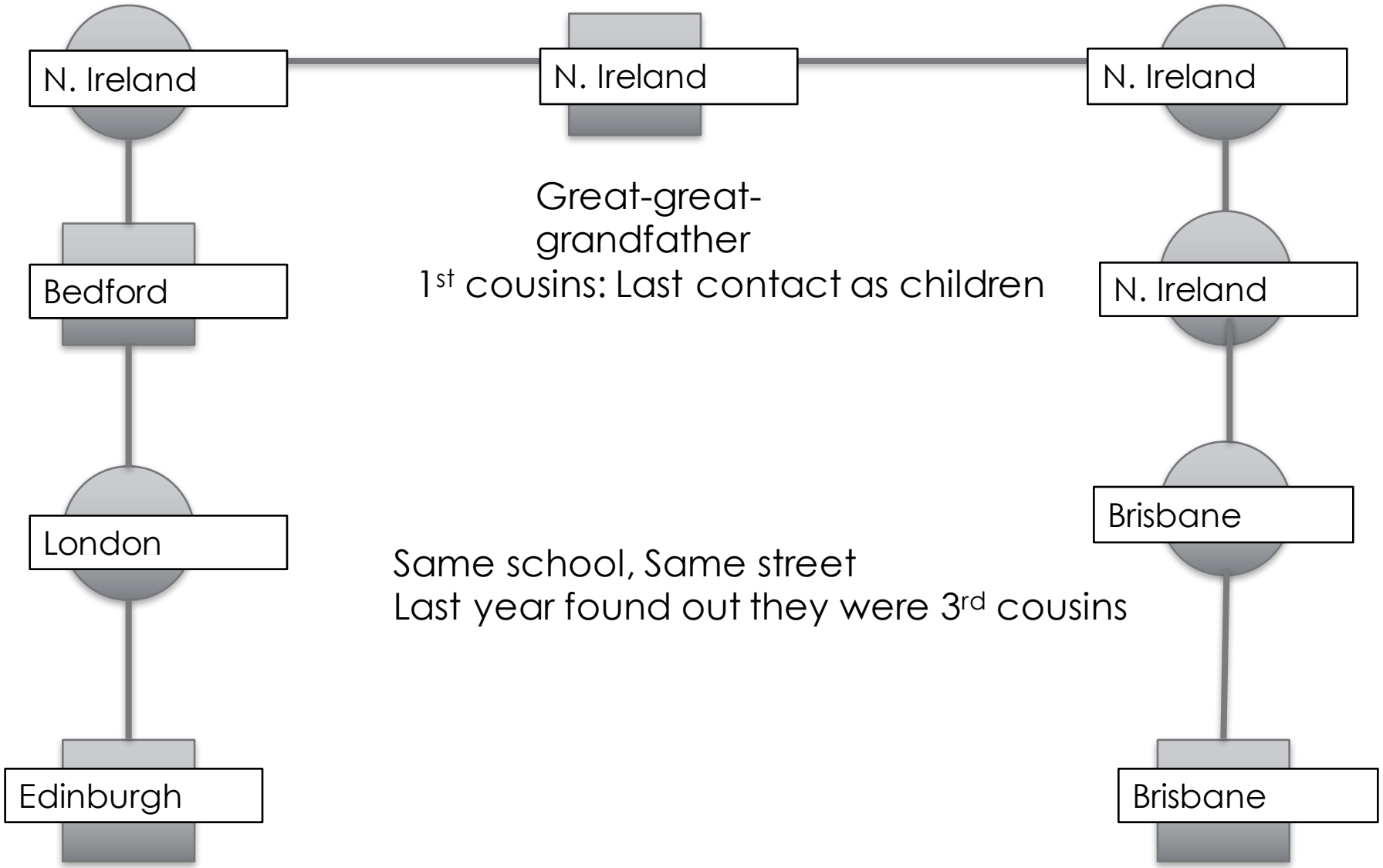
h^2-SNP
Proportion of variance attributable to common genetic variants

Non-additive genetic effects shared by distant relatives are tiny – expect no contamination

Distant relatives unlikely to share environmental effects

Do you know your 3rd cousins??

N=1 anecdote, skepticism check



Comparison of GREML and LDscore

Table 1. Estimates of genetic correlation between men and women from the bivariate GCTA-GREML analysis using individual-level data for five anthropometric traits.

Trait	Sample size (men versus women)	h^2_g (Men)		h^2_g (Women)		r_g		
		Est.	SE	Est.	SE	Est.	SE	$P(r_g = 1)$
Height	19 095 versus 24 504	0.447	0.018	0.431	0.015	1.022	0.031	0.483
BMI	19 016 versus 24 350	0.236	0.019	0.226	0.015	1.011	0.064	0.859
WCadjBMI	13 158 versus 15 874	0.167	0.026	0.174	0.022	0.774	0.119	0.057
HIPadjBMI	13 119 versus 15 854	0.231	0.026	0.185	0.022	0.855	0.101	0.149
WHRadjBMI	13 115 versus 15 846	0.159	0.026	0.182	0.022	0.607	0.112	4.4×10^{-4}

h^2_g = proportion of phenotypic variance explained by all SNPs used in the analysis. $P(r_g = 1)$: Wald's test P-value against $r_g = 1$.

Table 2. Estimates of genetic correlation between men and women from the LDSC regression analysis using summary data for five anthropometric traits.

Trait	Sample size (men versus women)	h^2_g (Men)		h^2_g (Women)		r_g		
		Est.	SE	Est.	Est.	Est.	SE	$P(r_g = 1)$
Height	60 505 versus 73 073	0.274	0.018	0.261	0.018	0.957	0.023	0.063
BMI	58 599 versus 67 935	0.167	0.012	0.186	0.010	0.879	0.035	5.9×10^{-4}
WCadjBMI	38 361 versus 42 727	0.143	0.014	0.110	0.013	0.780	0.071	1.9×10^{-3}
HIPadjBMI	32 920 versus 40 712	0.162	0.018	0.136	0.015	1.000	0.083	0.999
WHRadjBMI	34 594 versus 47 463	0.102	0.016	0.093	0.017	0.770	0.108	0.033

h^2_g = proportion of phenotypic variance explained by all SNPs used in the analysis. HIPadjBMI, BMI-adjusted hip circumference; WCadjBMI, BMI-adjusted waist circumference; WHRadjBMI, BMI-adjusted waist–hip ratio. The samples size shown in this table is the median of the per-SNP sample sizes reported in the summary data. $P(r_g = 1)$: Wald's test P-value against $r_g = 1$.

Yang et al (2015) Genome-wide heterogeneity between sexes and populations for human height and bod mass index. Hum Mol Gen

Comparison of GREML and LDscore

Table 1. Estimates of genetic correlation between men and women from the bivariate GCTA-GREML analysis using individual-level data for five anthropometric traits.

Trait	Sample size (men versus women)	h^2_g (Men)		h^2_g (Women)		r_g		
		Est.	SE	Est.	SE	Est.	SE	$P(r_g = 1)$
Height	19 095 versus 24 504	0.447	0.018	0.431	0.015	1.022	0.031	0.483
BMI	19 016 versus 24 350	0.236	0.019	0.226	0.015	1.011	0.064	0.859
WCadjBMI	13 158 versus 15 874	0.167	0.026	0.174	0.022	0.774	0.119	0.057
HIPadjBMI	13 119 versus 15 854	0.231	0.026	0.185	0.022	0.855	0.101	0.149
WHRadjBMI	13 115 versus 15 846	0.159	0.026	0.182	0.022	0.607	0.112	4.4×10^{-4}

h^2_g = proportion of phenotypic variance explained by all SNPs used in the analysis. $P(r_g = 1)$: Wald's test P-value against $r_g = 1$.

Supplementary Table 1 Estimates of genetic correlation between men and women from the bivariate LDSC regression analysis in the combined GWAS data for five anthropometric traits.

	Trait	Sample size (men vs. women)	h^2_g (Men)		h^2_g (Women)		r_g		
			Est.	SE	Est.	SE	Est.	SE	$P(r_g = 1)$
No adjustment	Height	19095 vs. 24504	0.412	0.042	0.374	0.039	1.006	0.051	0.905
	BMI	19016 vs. 24350	0.196	0.026	0.187	0.026	1.198	0.113	0.079
	WCadjBMI	13158 vs. 15874	0.158	0.037	0.118	0.030	0.986	0.226	0.952
	HIPadjBMI	13119 vs. 15854	0.192	0.046	0.109	0.031	1.000	0.219	0.999
	WHRadjBMI	13115 vs. 15846	0.112	0.038	0.139	0.033	0.593	0.224	0.069
Genomic Control	Height	19095 vs. 24504	0.367	0.038	0.326	0.034	1.006	0.051	0.907
	BMI	19016 vs. 24350	0.180	0.024	0.172	0.024	1.198	0.113	0.079
	WCadjBMI	13158 vs. 15874	0.153	0.036	0.114	0.029	0.986	0.226	0.952
	HIPadjBMI	13119 vs. 15854	0.184	0.044	0.104	0.029	1.000	0.219	0.999
	WHRadjBMI	13115 vs. 15846	0.108	0.037	0.134	0.032	0.593	0.224	0.069

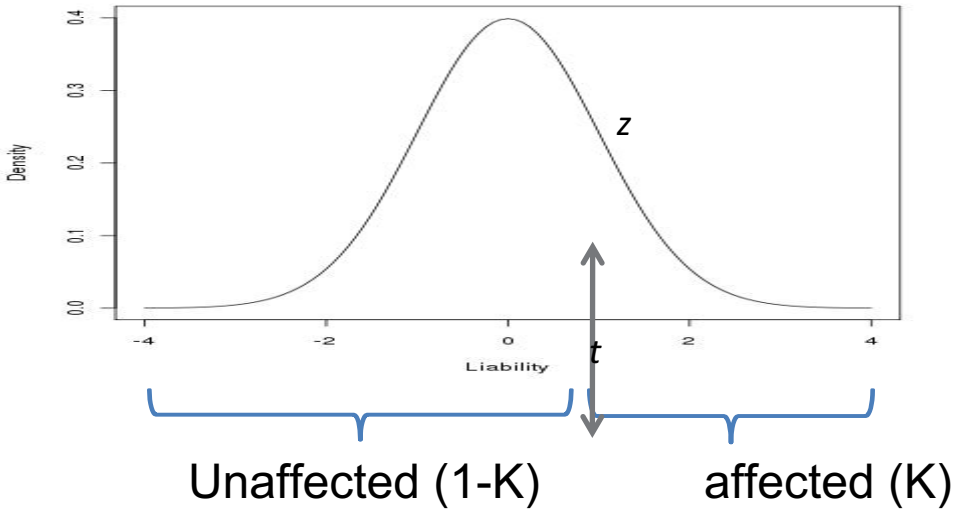
h^2_g = proportion of phenotypic variance explained by all SNPs used in the analysis. $P(r_g = 1)$: Wald test p-value against $r_g = 1$.

GREML: h^2-SNP for disease

- Observations are on disease scale but heritability is most interpretable on the liability scale
  - Use linear regression
  - Estimate on observed scale
  - Transform to Liability scale via Robertson Transformation
- Case-control samples are ascertained
  - Up date transformation
- Differences between case and control samples may reflect artefacts
  - Very stringent QC

Lee et al (2011) Estimating missing heritability for Disease from GWAS AJHG
Lee et al (2013) Estimation and partition of polygenic variation captured by common SNPs for AD, MS & Endo, HMG

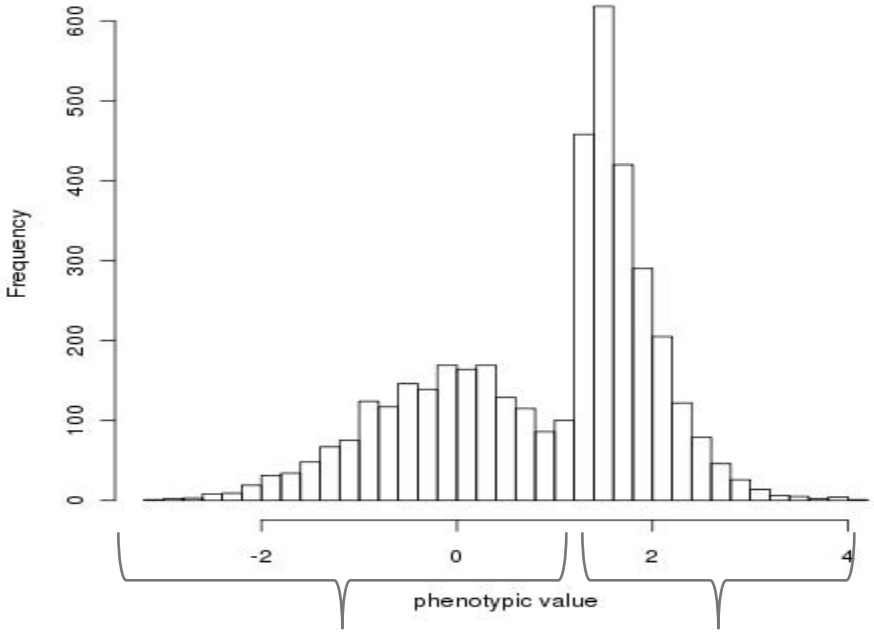
Ascertainment in case-control studies



$$h_l^2 = h_o^2 \frac{K(1-K)}{z^2}$$

Robertson (1950)
Appendix of Dempster and Lerner (1950)
See Lecture 1

$\widehat{h^2_{o_{cc}}}$ · Estimate of proportion of variance explained
· by SNP between cases and controls



$$h_l^2 = \widehat{h^2_{o_{cc}}} \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)}$$

Lee et al (2011)AJHG
Zhou & Stephens (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models PLoSG Text S3
Golan et al (2014) Measuring missing heritability: Inferring the contribution of common variants PNAS

Golan et al (2014) Measuring missing heritability: Inferring the contribution of common variants PNAS

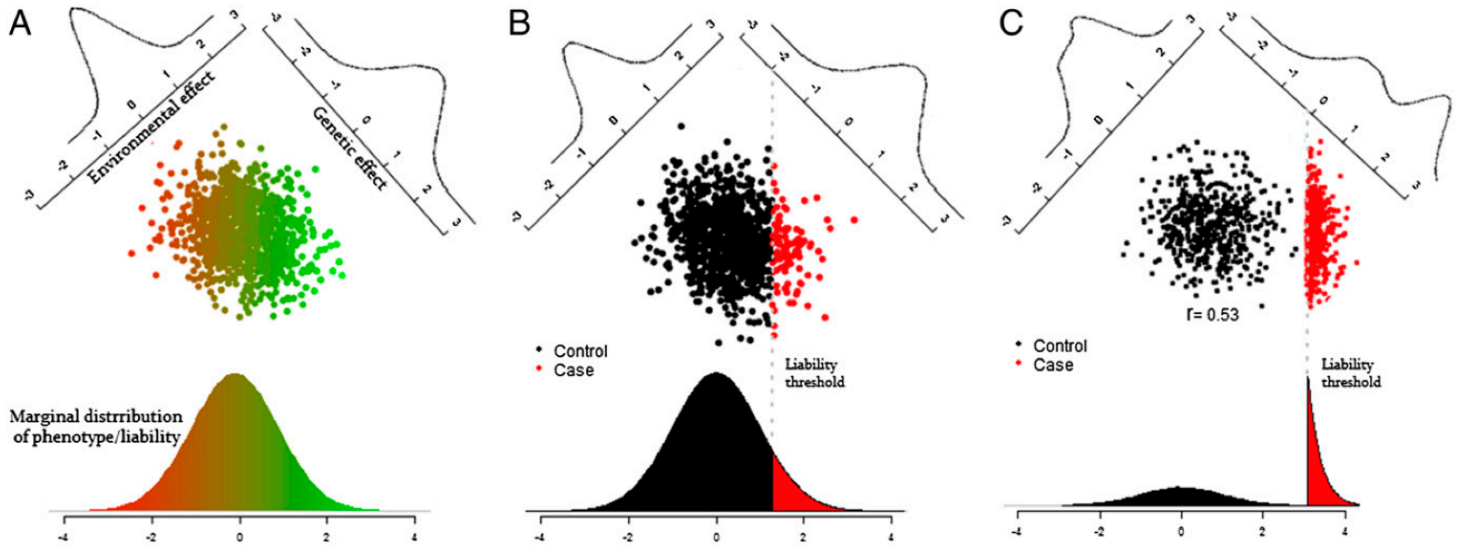


Fig. 1. Distributions of genetic effects, environmental effects, phenotypes, and liabilities in three study designs. In each of *A*, *B*, and *C*, a phenotype is assumed to depend on the sum of a genetic effect and an environmental effect. The scatterplot shows the joint distribution of the genetic and environmental effects, the upper left shows the marginal distributions of the environmental effect, the upper right shows the marginal distributions of the genetic effect, and the lower portion shows the marginal distribution of the phenotype. (*A*) Quantitative phenotype in a random sample of the population. (*B*) Disease phenotype in a random sample of the population. (*C*) Disease trait in a balanced case–control study. Disease phenotypes were simulated under a liability threshold model with disease prevalence of 10% (*B*) and 0.1% (*C*), with red points indicating affected individuals (liability above the threshold) and black points indicating unaffected individuals (liability below the threshold). In *C*, the marginal distributions of the genetic and environmental effects no longer are normally distributed, and there is an induced positive correlation between the genetic and environmental effects ($r = 0.53$).

Non-normality of liability

Case-control sampling induces GxE correlation

Solution use Haseman-Elston regression regression of phenotype correlation between each pair of individuals and genetic relationship between each pair of individuals

(see also Chen et al, 2013 Estimating heritability of complex traits from GWAS using IBS-based Haseman-Elston regression. Front Genet)

Golan et al (2014) Measuring missing heritability: Inferring the contribution of common variants PNAS

Denote by Z_{ij} the product of the standardized phenotypes:

$$Z_{ij} = \frac{(y_i - P)(y_j - P)}{P(1-P)}.$$

The variable Z_{ij} can obtain three values:

$$Z_{ij} = \begin{cases} \frac{1-P}{P} & y_i = y_j = 1 \\ -1 & y_i \neq y_j \\ \frac{P}{1-P} & y_i = y_j = 0 \end{cases}.$$

Regress Z_{ij} on A_{ij} (coefficient of relationship estimated from SNPs)

PCGC regression (phenotype correlation-genotype correlation regression) – general form of Haseman-Elston correction for fixed effects

Their simulation shows substantial underestimation of SNP-heritability from GREML applied to disease traits

NB their simulation strategies exacerbate differences that we see in real data

In the past we have used H-E in-house as a check, that all is well with GREML. Usually we see little difference in estimates, but standard errors smaller with GREML.

As sample sizes increase the induced GxE correlation will become more of a problem. See revision when posted of
Loh et al Nature Genetics 2015
Includes updated faster version of PCGC

Golan et al (2014) Measuring missing heritability: Inferring the contribution of common variants PNAS

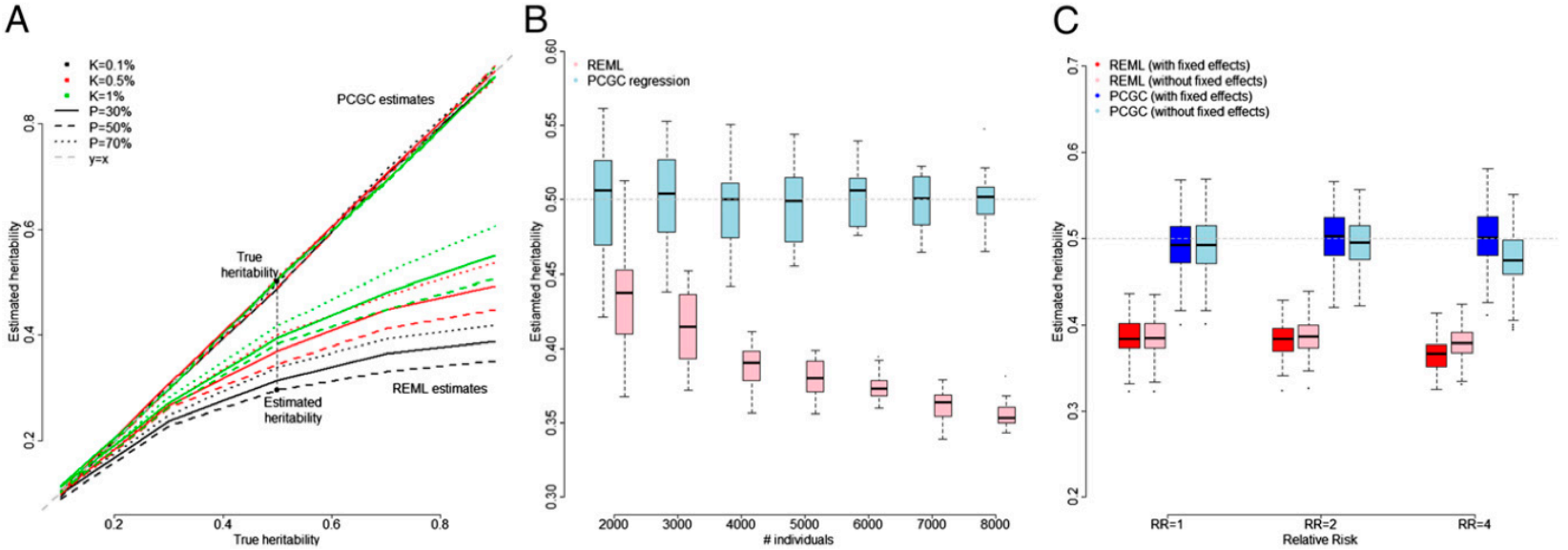


Fig. 2. Comparison of REML and PCGC regression. (*A*) REML yields biased estimates for case–control studies of diseases, whereas PCGC regression yields unbiased estimates. We simulated case–control studies for nine combinations of K (prevalence) and P (proportion of cases among overall samples), and for five values of h^2 (0.1, 0.3, 0.5, 0.7, and 0.9). For each combination of parameters, we show the average of 10 heritability estimates obtained by applying the REML method of Lee et al. (10) and PCGC regression to our simulated case–control data. REML produced biased estimates, whereas PCGC regression produced unbiased estimates for all scenarios. The bias of REML estimates increases as both the true heritability and overrepresentation of cases increase. To demonstrate the severity of the bias, consider the scenario of a disease with prevalence of 0.1% in a balanced case–control study (values typical for Crohn's disease or MS). When the true heritability is 50%, the estimated heritability would be 30% on average, as indicated by the black dots. (*B*) Heritability estimates for case–control studies with increasing sample size. Simulated case–control studies are as previously described, with the prevalence of the disease, the proportion of cases, and the heritability fixed at 1%, 30%, and 50%, respectively. The size of simulated studies ranged from 2,000 to 8,000. The bias of heritability estimates from REML increases with study size, whereas those from PCGC regression estimates remain unbiased. (*C*) Heritability estimation in the presence of fixed effects. We simulated case–control studies with an additional "sex" covariate, which either has no effect on the disease or increases the relative risk (RR) by twofold or fourfold. The prevalence of the disease in the population was 0.5%, the heritability was set to 50%, and the numbers of cases and controls were equal. Applying REML with or without accounting for the additional covariate resulted in underestimation of the heritability. Moreover, inclusion of the covariate as a fixed effect resulted in even lower estimates of heritability when the effect of the covariate on the phenotype was considerable. By contrast, PCGC regression correctly accounted for the presence of the covariate.

Summary

- Heritability from family records is not expected to be the same as SNP-h^2
- Five methods to estimate SNP-h^2
  - GCTA GREML = gold standard
  - Methods based on summary statistics are the best starting place
- We are not interested in decimal place accuracy missing heritability
- We should be interested in order of magnitude missing heritability
  - What can we learn from the data available now to inform on future experimental design

Risk Prediction ..again

Factors affecting accuracy of risk prediction

Genetic architecture of the trait – unknown

- Number, frequency, effect size
- How well marker effects are correlated with causal variants (LD)

Sample size of discovery sample – maximise

– how well marker effects are estimated

Sample size of target sample – be sufficiently large (once achieved not so much gained by increasing further)

- Precision of estimation of R^2

Dudbridge (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genetics
Wray et al (2014) Polygenic methods and their application to psychiatric traits. Journal of Child Psychology & Psychiatry (in press)

Single GWAS- how to split into discovery and target?

Split based on independently collected samples

What is the optimum split?

Equal sample sizes of discovery and target gives maximum power to detect association between discovery and target (Dudbridge).

But with large samples power achieves 1, so value of increasing target sample is redundant.

Rule of thumb.

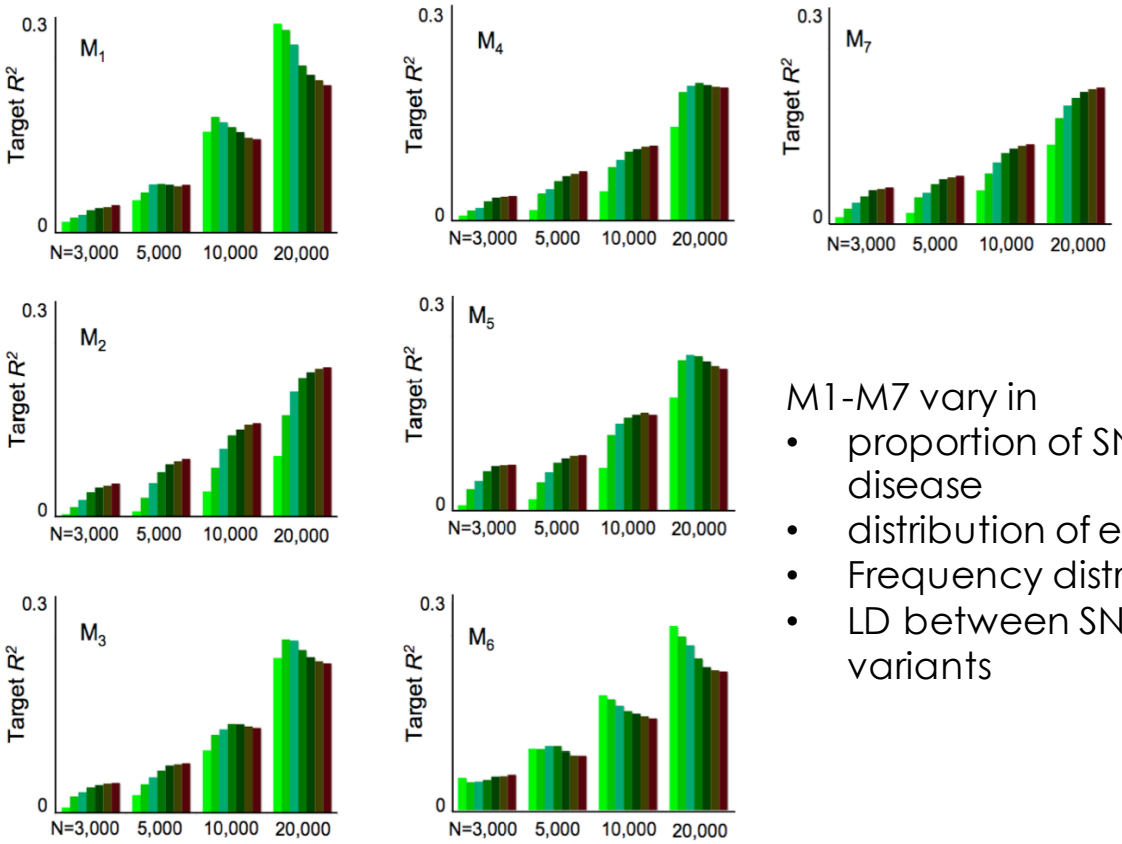
Split sample equally into discovery and target until target has ~2000 cases + 2000 controls, then add additional samples to discovery.

Then with larger sample sizes the accuracy of the estimation of SNP effects is increased and the accuracy of the GRS for an individual increases

Dudbridge (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genetics
Wray et al (2014) Polygenic methods and their application to psychiatric traits. Journal of Child Psychology & Psychiatry

Simulation study demonstrating the impact of sample size and genetic architecture on profile scoring

Figure S8: *Impact of increasing sample size on score analysis.*

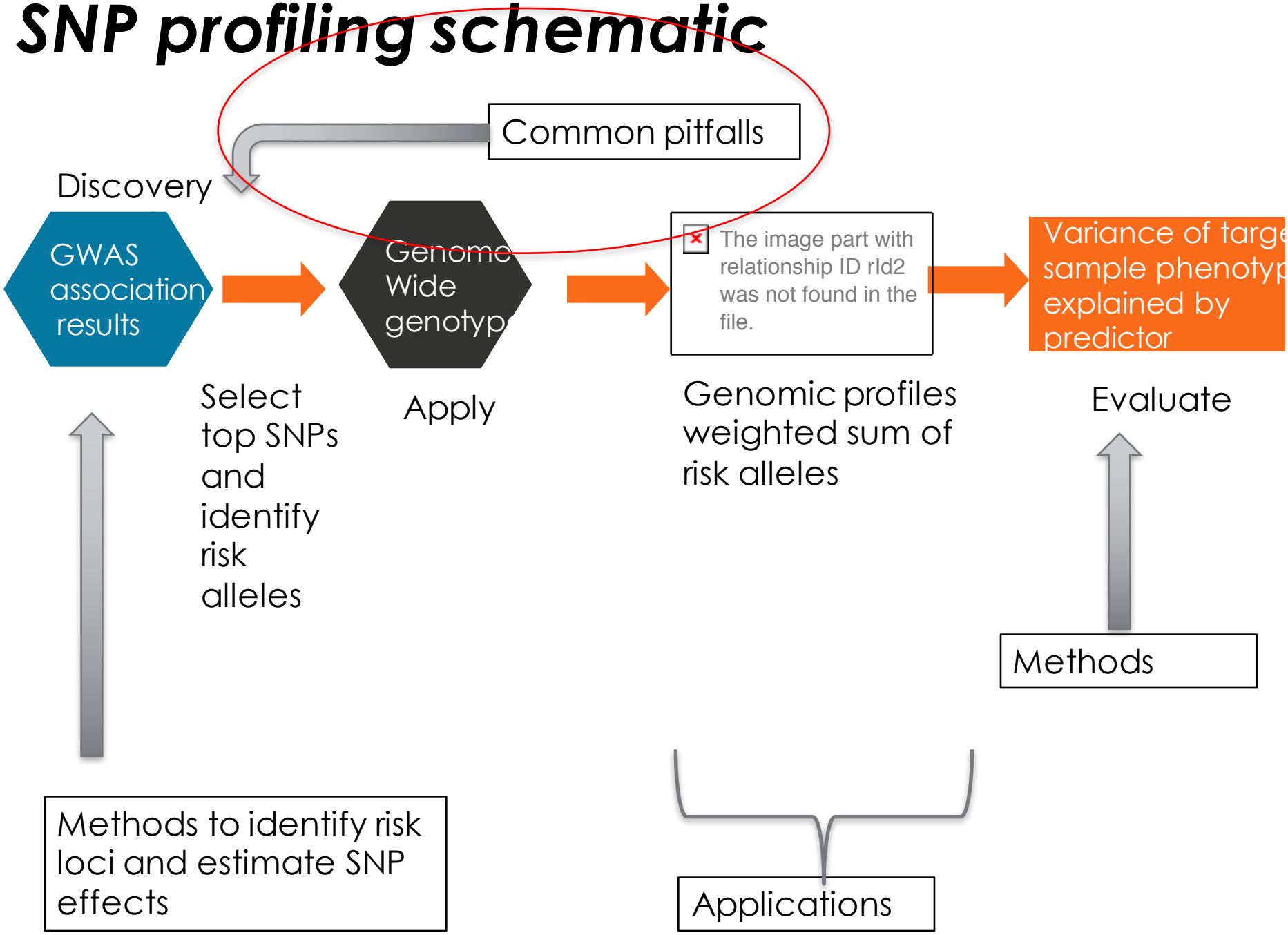


M1-M7 vary in

- proportion of SNPs associated in disease
- distribution of effect sizes
- Frequency distribution
- LD between SNPs and causal variants

Purcell et al (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature

SNP profiling schematic



Pitfall 1: No target (=validation) sample

- Report R^2 or AUC from discovery sample only
- Small n large p problem
- Even under null can get high R^2 within discovery sample when p >> n

Wray, Yang, Hayes, Price, Goddard, Visscher (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genetics

Pitfall 2: Overlapping Discovery & Target Sample

- Overlapping discovery & target samples
- Greater similarity between discovery & target samples than discovery & true validation samples
  - E.g. cross-validation samples
  - Not a pitfall, as such, but to be aware

Wray, Yang, Hayes, Price, Goddard, Visscher (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genetics

Pitfall 3: Less obvious non-independence

- Cross-validation but select associated SNPs from total sample

- Select SNPs in discovery sample, for those SNPs re-estimate effects in the target sample

Wray, Yang, Hayes, Price, Goddard, Visscher (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genetics

Selection bias

Select *m* 'best' markers out of *M* in total

'Prediction' in same sample

$E(R^2) >> m/N$

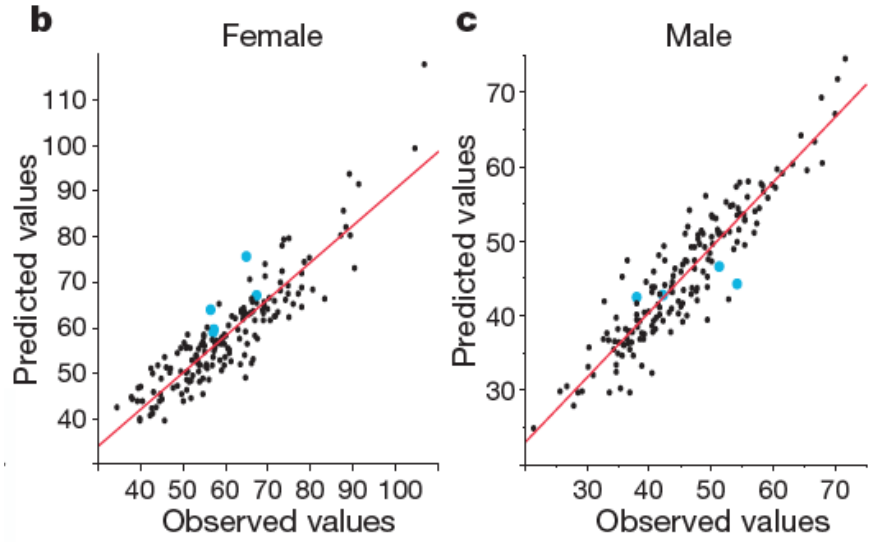
→ Lots of variation explained by chance

ARTICLE

doi:10.1038/nature10811

The *Drosophila melanogaster* Genetic Reference Panel

~15 best markers selected from 2.5 million markers



Practical

Module 18: Statistical and Quantitative Genetics of Disease: Rare Variants and Prediction

John Witte

Lecture 10

Rare Variants

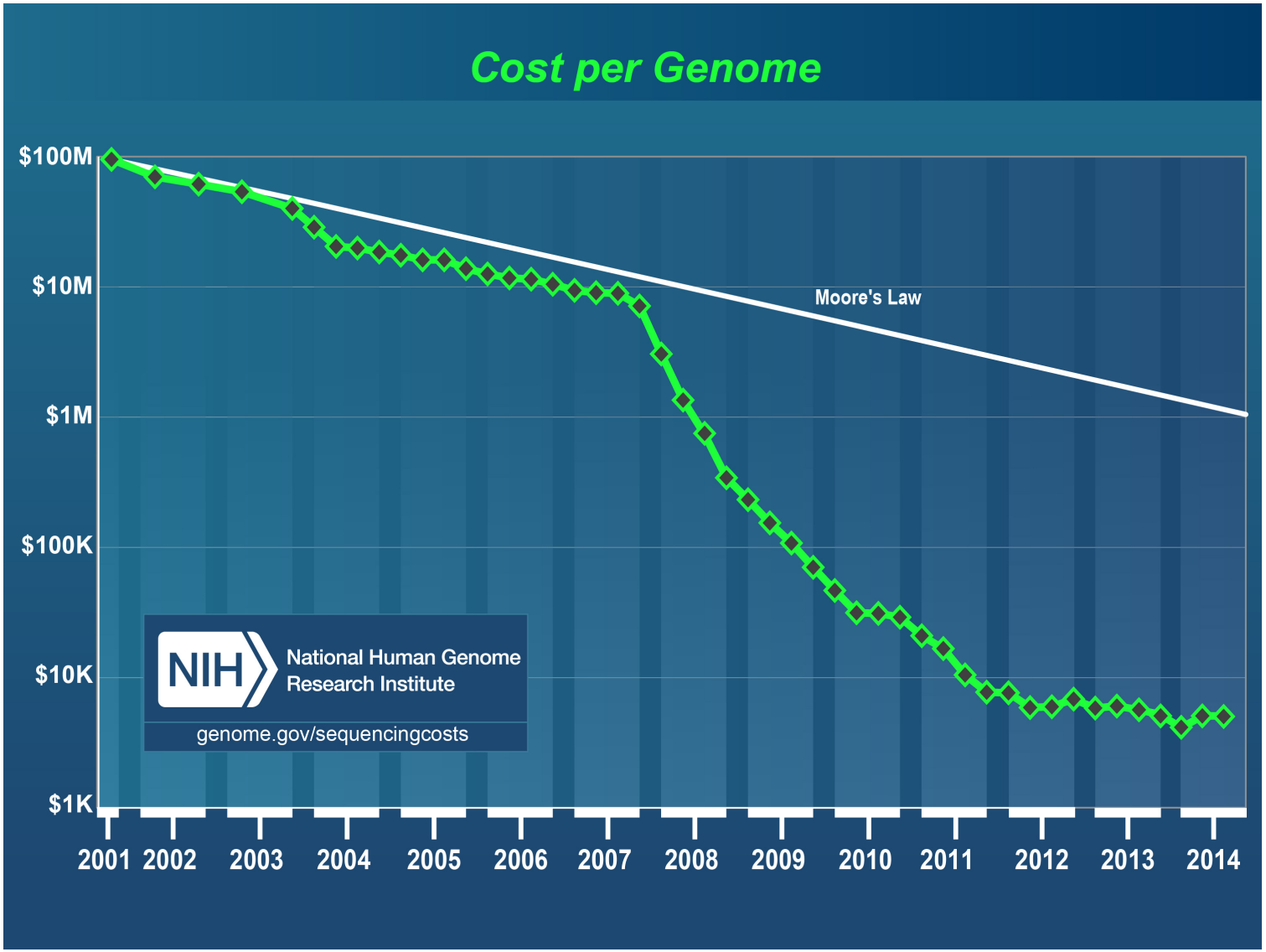
- “Common”: MAF > 0.05
- “Less common”: 0.05>MAF>0.01
- “Rare”: 0.01<MAF

- SNP: MAF>0.01 (Single Nucleotide Polymorphism)
- SNV: MAF<0.01 (Single Nucleotide Variant)

Rare Variants

- Previous GWAS focused on chips designed for MAF > 0.05 (most powered for MAF > 0.10)
- Exome arrays
- Sequencing (de novo)

Sequencing Costs have Fallen



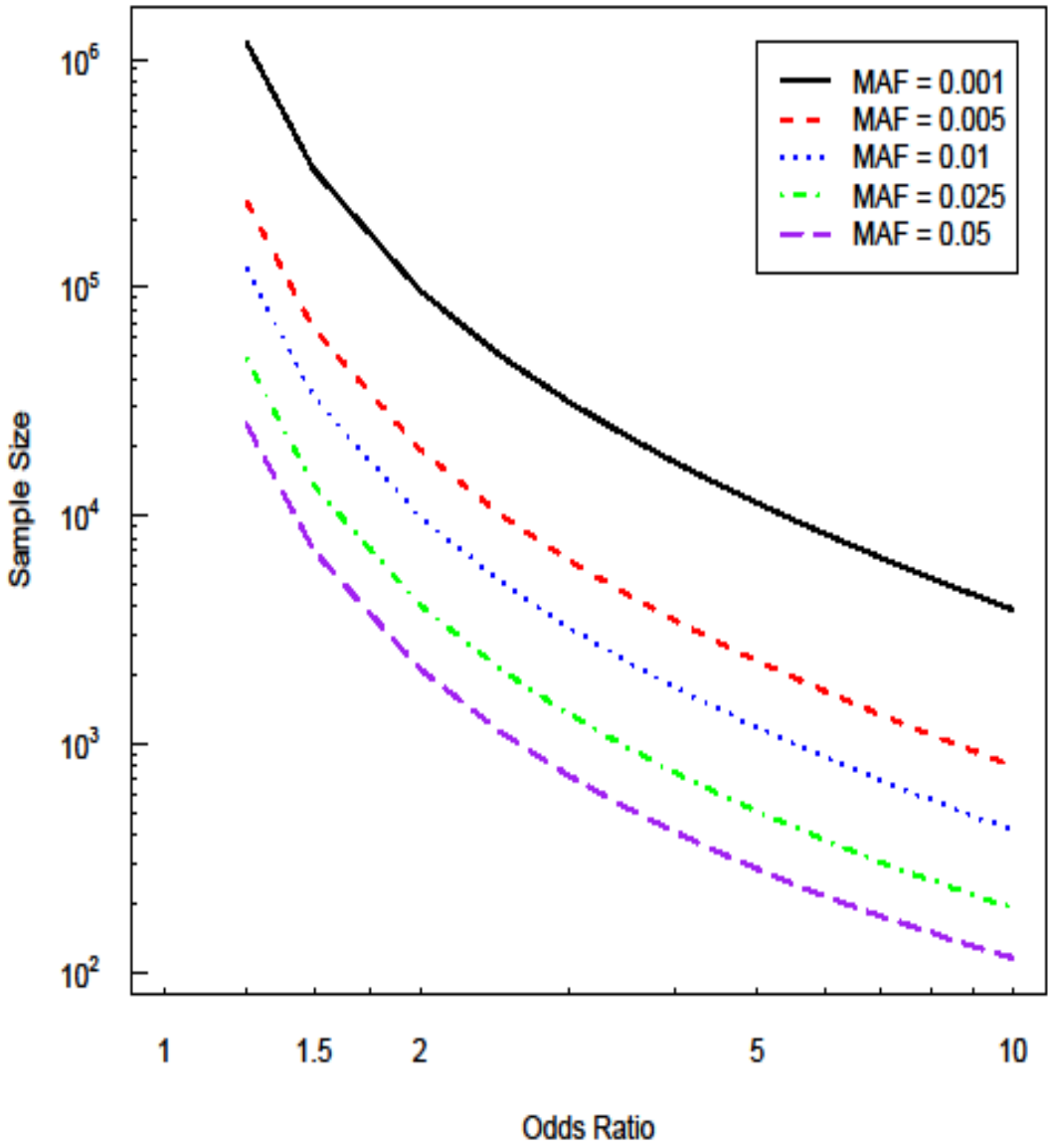
Analysis of Rare Variants

Focus on a set of k variants

$$g(Y_i) = \alpha_0 + \sum_k \beta_k X_{ik},$$

- Difficult to model due to sparsity.
- Limited power.

Sample Size for Rare Variants



Rare Variant Tests

- ‘Up-weight’ analyses for most likely causal variants.
- Burden tests (CAST, Collapsing, WSS).
- Variance component (dispersion) tests (SKAT, SKAT-O, C-alpha).
- Burden tests more powerful when a large percentage of rare variants are causal and have the same sign (direction of association).
- Variance component more powerful when there is a mixture of risk and protective variants, and most rare variants are not causal.

Burden Tests for Rare Variants

$$g(Y_i) = \alpha_0 + \gamma \left[\sum_k w_k X_{ik} \right].$$

Where w_k defines similarities among the variants for their aggregation / modeling

Estimate the effect of a weighted summary 'score' across each individuals' rare variants on outcome.

Key Aspect: Specifying w_k

$$w_k = a_k \times s_k \times i_k$$

where

a_k inverse variance weighting, controls' MAF

s_k direction of association; positive / negative

i_k Indicators for whether to aggregate

- Overall MAF
  - Hard cutpoint (e.g., MAF < 0.01)
- Functional information
  - Non-synonymous
  - Deleterious (SIFT)

Example: Cohort Allelic Sums Test (CAST)

Aggregate rare variants within three genes

a_k = 1

s_k = 1

i_k = 1 if rare, nonsynonymous

ABCa1, APOA1, or LCAT	>95% HDL	<5% HDL	OR (p-value)
No ns variants	125	107	1.0
ns variants	3	21	8.1 ($1x10^{-4}$)

Cohen et al., Science 2004;305:869.
Morgenthaler Mut Res 2007;615:28.

Difficult to determine best weighting / aggregation scheme *a priori*

Most approaches make strong assumptions about exchangeability and combination of rare variants for analysis.

Empirical ‘Step-Up’ Approach

- Data driven aggregation of rare variants
- Consider multiple possible groupings
- Select the “best” grouping (e.g., min P)
- Correct by permutation

- Possible groupings defined by:
  - MAF weighting / cutoffs
  - Positive or negative associations
  - Nonsynonomous
  - Deleterious (SIFT)
- All possible subsets, or those contributing most to signal

Hoffmann, Marini & Witte, 2010

Variance Components Approach

- SNP-set (Sequence) Kernel Association Test (SKAT) (Wu et al., AJHG 2011).
- Uses flexible weight kernels, which reflect different assumptions underlying the rare variant tests.
- For example, that rarer variants have larger effect sizes.

Test Stats for SKAT vs. Burden

$$\text{SKAT}: \quad Q_{\rho=0} = \sum_{j=1}^{m} w_j^2 \left[\sum_{i=1}^{n} \left(Y_i - \widehat{\mu_{i,0}} \right) X_{ij} \right]^2 \text{ and}$$

$$\text{Burden}: \quad Q_{\rho=1} = \left[\sum_{j=1}^{m} w_j \sum_{i=1}^{n} \left(Y_i - \widehat{\mu_{i,0}} \right) X_{ij} \right]^2 .$$

Prediction: Ozzy Osbourne!?

cientists to map Ozzy Osbourne's
enetic code to find out how he survived
o much substance abuse

NICK KLOPSIS
ILY NEWS WRITER

nday, June 14th 2010, 1:39 PM

Sayles/AP

zzy Osbourne has had many issues relating to drug and alcohol abuse, so scientists are apping out his genetic code to find out how his body can take it.

u can't kill rock and roll, but it's not ually this hard to kill a rocker.

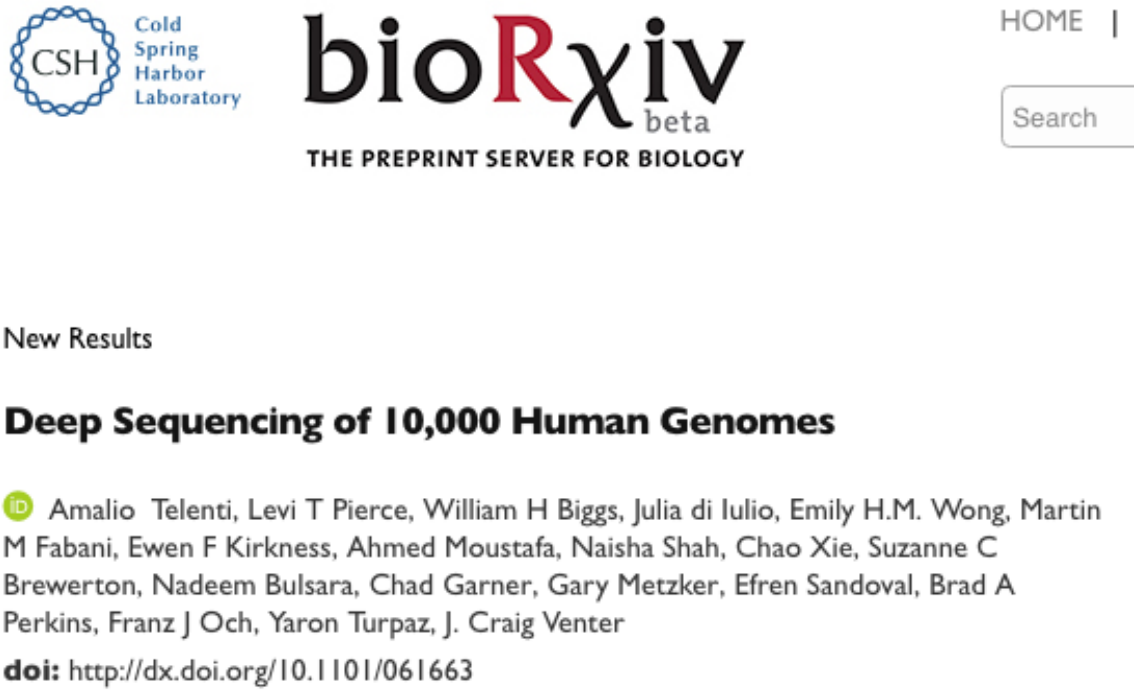
RELATED NEWS

Increased risk of:

- Alcohol and cocaine dependence.
- Hallucinations while on marijuana.

Slow to metabolise coffee.

More Info will Improve Prediction

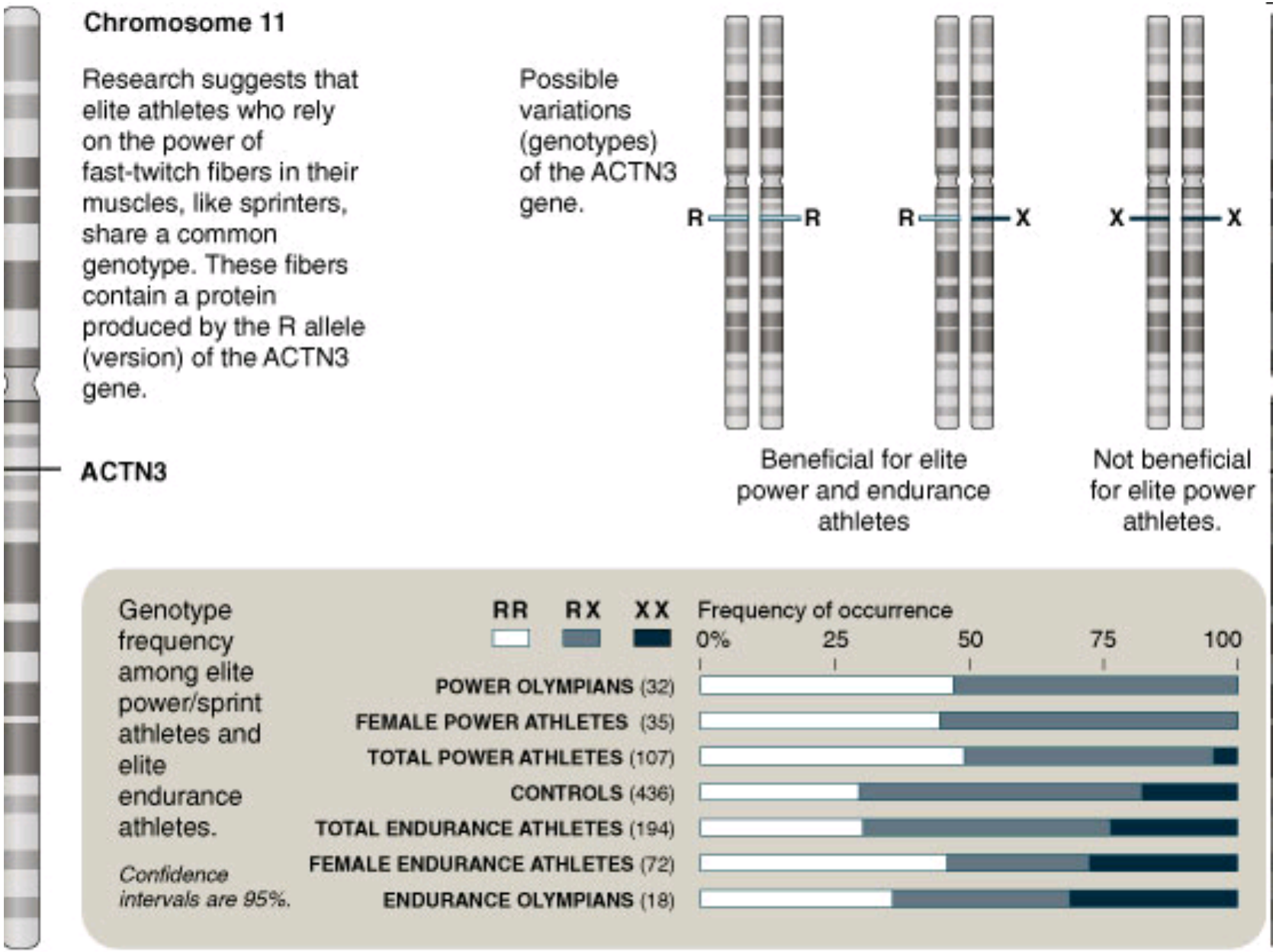


Precision Medicine Initiative Cohort

Key Features	Framingham Heart Study	Precision Medicine Initiative Cohort Program
Year Started	1948	2016
Number of Individuals	5,209*	1,000, 000
Age	30-62	All
Ancestry	>95% European	Diverse, cross-section of Americans
Medical data obtained	Every 2 yrs at office visit	Real world, real time, via mobile devices, Web
Focus	Heart disease	All medical conditions, health
Data return to participants	No	Yes
Data available for research community	No	Yes

* Initial cohort

Genetic Prediction



Sources: Stephen M. Roth, Ph.D., University of Maryland; American Journal of Human Genetics

NY Times, 11/30/08

AACR Meeting Abstra...

Home Log In

HOME PRODUCTS ABOUT US COMPANY RESEARCH SHOPPING CART

ATLAS SPORTS GENETICS

ATHLETIC TALENT LABORATORY ANALYSIS SYSTEM

Finding any great Olympic champion normally takes years to determine.

What if we knew a part of the answer when we were born?

See How...

The New York Times - Born to Run? Little Ones Get Test for Sports Gene

GIST™ Sports Profile

Genetic Testing for Speed/Power and Endurance Events

Home :: Privacy Policy :: Conditions Of Use :: Disclaimer :: Links

ATLAS SPORTS GENETICS

ATHLETIC TALENT LABORATORY ANALYSIS SYSTEM

larger image

ATLAS First

Recommended for ages 1 and up.
Description: Our *Atlas First* product is geared specifically at the youngest of athletes. Doing any type of performance based sport talent identification testing is very difficult below age 6 due to developmental levels of motor skills, strength and eye-hand coordination. *Atlas First* looks at only genetic markers, specifically the presence of ACTN3. Studies have found that individuals having the variant in both copies of their ACTN3 gene may have a natural predisposition to endurance events, one copy of their ACTN3 gene may be equally suited to for both endurance and sports/power event, neither copy of their ACTN3 gene may have a natural predisposition to sprint/power events. Knowing this information may be helpful, not in eliminating choices for sport activities but adding exposure to a host of team or individual sport events that may come easier to a young athlete.

The test is one of tool of many that can help children realize their athletic potential.

Other Products available through Atlas First

- Height monitoring charts
- Weight monitoring charts
- BMI charts
- Height Prediction Calculator (not genetic)

$149.00

My 23andMe Results for *ACTN3*

me
My Health and Traits
Browse Raw Data
My Profile

family & friends
Compare Genes
Family Inheritance

my ancestors
Maternal Line
Paternal Line
Ancestry Painting
Global Similarity

23andWe
Introduction
My Surveys (15)
Featured Research

community
23andMe Community

account
Genome Sharing
Inbox
Settings

health and traits

Traits | Research Reports (72)

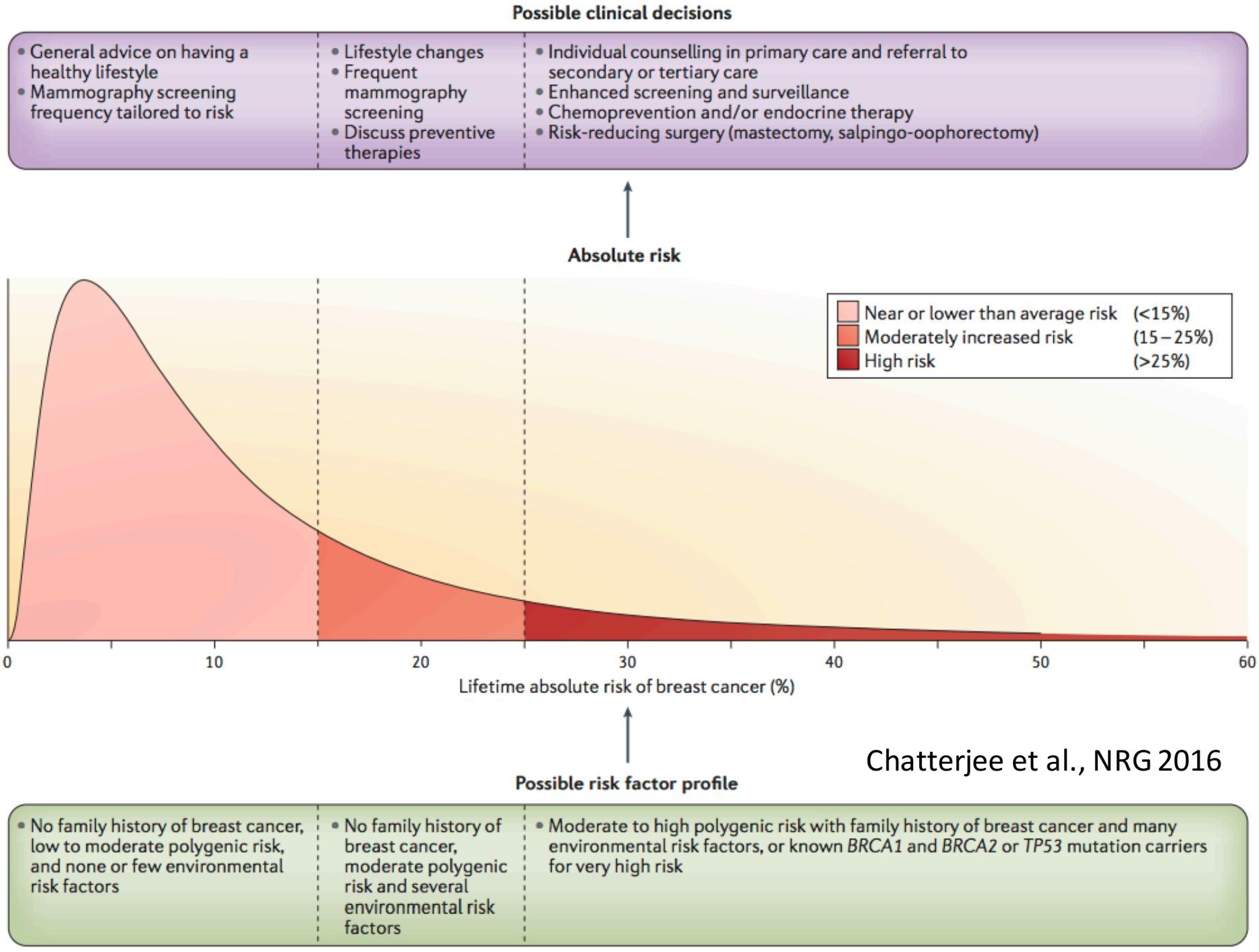
Show data for: John Witte

<< Return to All Clinical Reports | Disease Risks | Carrier Status | Traits | Recently Updated

Name ▲	Outcome	Last Updated
Alcohol Flush Reaction	Does Not Flush	Dec 19, 2007
Bitter Taste Perception	Can Taste	Nov 19, 2007
Earwax Type	Wet	Nov 19, 2007
Eye Color	Likely Blue	Mar 25, 2008
Lactose Intolerance	Likely Tolerant	Nov 19, 2007
Malaria Resistance (Duffy Antigen)	Not Resistant	Feb 28, 2008
Muscle Performance	Unlikely Sprinter	Nov 19, 2007
Non-ABO Blood Groups	See Report	Mar 25, 2008
Norovirus Resistance	Resistant	Jul 23, 2008
Resistance to HIV/AIDS	Not Resistant	Jan 27, 2008

The genotyping services of 23andMe are performed in LabCorp's CLIA-registered laboratory. The results presented here have not been cleared or approved by the FDA but have been analytically validated according to CLIA standards.

Talk with the Community



Chatterjee et al., NRG 2016