SISG 2014

SISG Module 19: Statistical & Quantitative Genetics of Disease

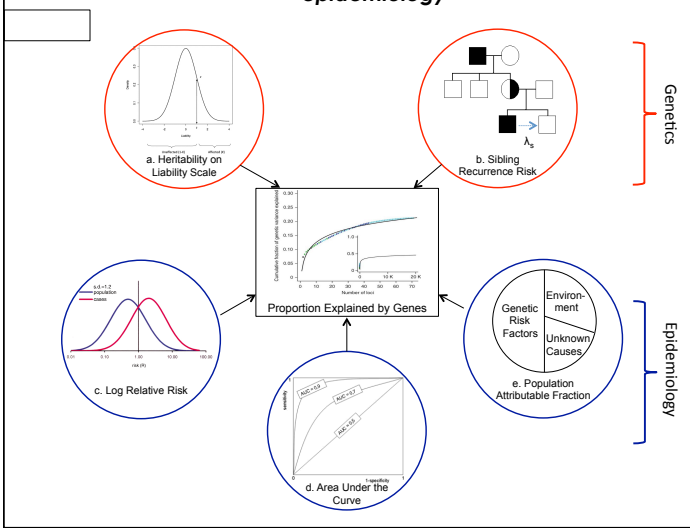
19th Summer Institute in Statistical Genetics

UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Module 19:
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
- To mix why with how

Course Outline

Monday morning (Naomi)
- Genetic epidemiology of disease
- Heritability of liability

Monday afternoon (John)
- Association analysis
- Rare variants

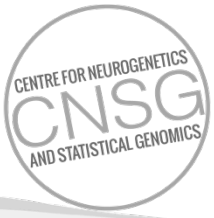
Tuesday morning (John)
- Measuring the contribution to risk of individual variants

Tuesday afternoon (Naomi)
- Polygenic models of disease
- Polygenic risk scoring

Wednesday morning
- Power (Naomi)
- Pleiotropy (John)

Module 19:
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

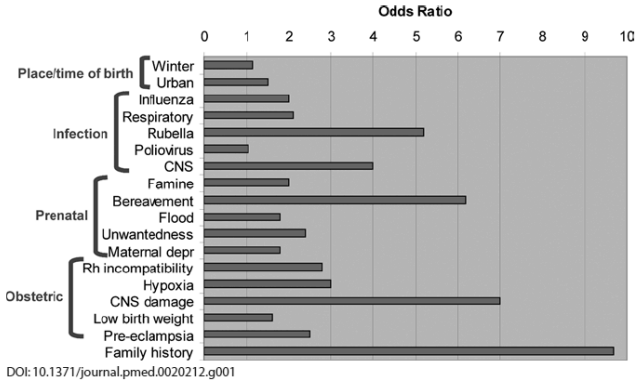
If an individual is affected > 60% will have no known family history

Bringing together genetic epidemiology and quantitative genetics

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

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)

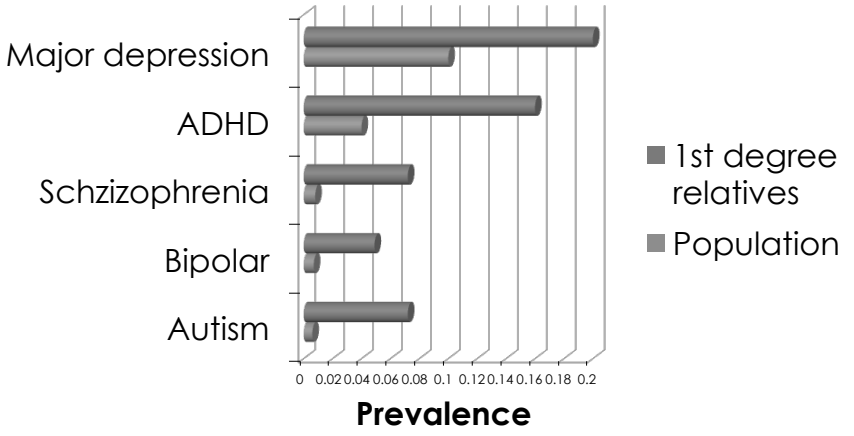
Sullivan, PLoS Med 05

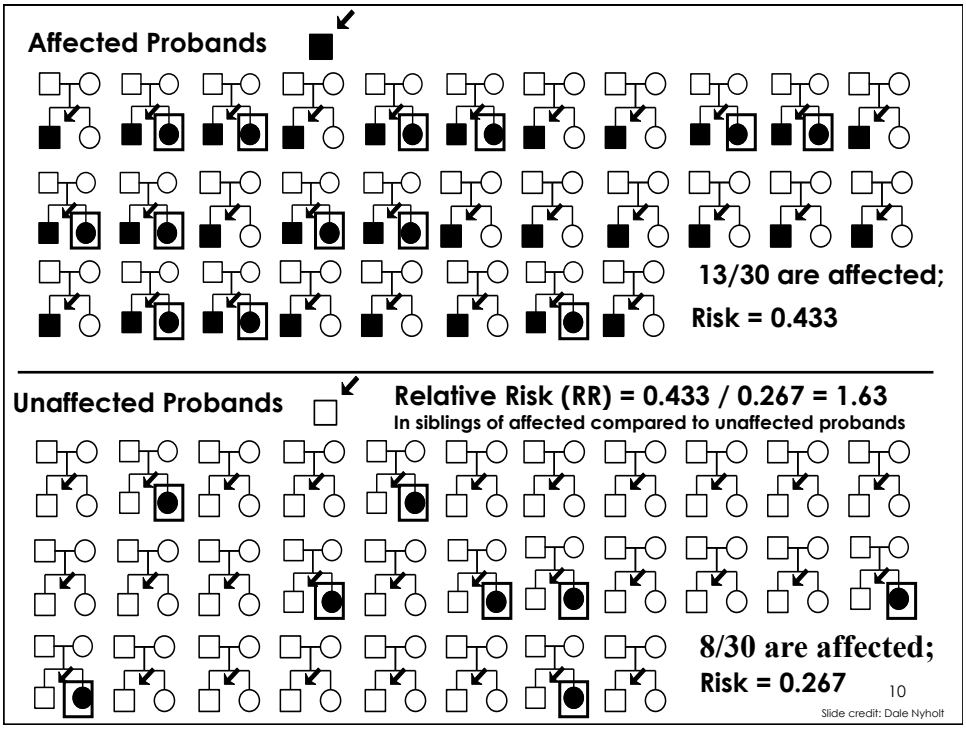
Complex genetic diseases

- 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





Slide credit: Dale Nyholt

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?

Relative risk to relatives (λ_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

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

K= "Prevalence" in Quantitative Genetics

Very specific meanings in epidemiology:

Prevalence – the proportion of people in a population who have the disease in a stated time frame

- Point prevalence – at point of assessment
- Period prevalence – at any time during the period of assessment
- Cumulative prevalence – had disease at any time in life
- Lifetime prevalence – had disease at any time in life

Incidence- the proportion of people who are newly diagnosed with a disease in a given time frame,

- Annual incidence– newly diagnosed within a 12-month period
- Lifetime incidence – diagnosed at any time in their life

Lifetime Morbid Risk- the proportion of a birth cohort that are diseased in their lifetime

The same

Disease Prevalence, Incidence and Lifetime Morbid Risk

Usually reported
Prevalence – point prevalence
Incidence- annual incidence
Lifetime Morbid Risk-

Schizophrenia:
Age of onset: 20's
Long mean life expectancy after diagnosis (albeit reduced)
Annual incidence: 2.5 per 10,000
Prevalence: 46 per 10,000
LMR: 72 per 10,000

Motor Neurone Disease:
Age of onset: 60's
Life expectancy after diagnosis: 2-5 years
Annual incidence: 0.3 per 10,000
Prevalence: 0.6 per 10,000
LMR: 25 per 10,000

How well is prevalence estimated? Sampling variance of risks

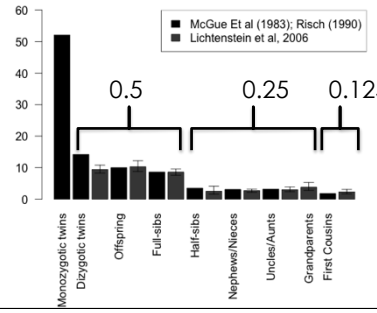
Either diseased or not diseased ~ Bernouilli (K)
Sampling variance $\frac{K(1-K)}{n}$
Standard error $\sqrt{(K(1-K)/n)}$

K = 0.01 n=10 s.e = 0.03
n=100 s.e.= 0.01
n= 1000 s.e. = 0.003

K = 0.10 n=10 s.e = 0.10
n=100 s.e.= 0.03
n= 1000 s.e. = 0.01

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 | X=1)

Probability that both X and Y =1: $E(XY) = K*K_R$
$Cov(X,Y) = E(XY) – E(X)*E(Y) = E(XY) – K^2$

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

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
Relationship between X and Y is linear (has to be only 2 values each)

$$Y = \mu_Y + b_{Y,X}(X-\mu_X) + \varepsilon = K + \frac{cov(Y,X)}{var(X)}(X-K) + \varepsilon = K + \frac{cov_R}{K(1-K)}(X-K) + \varepsilon$$

K_R proportion of relatives of affected individuals (i.e. X=1) who are affected

$$K_R = K + \frac{cov_R}{K(1-K)}(1-K) = K + \frac{cov_R}{K} \qquad cov_R = (K_R - K)K = (\lambda_R - 1)K^2$$

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

James (1971) relationship between K and K_R

$$cov_R = (\lambda_R - 1)K^2$$

cov_R is the covariance between phenotypes of relatives
= covariance between genetic values of relatives when we assume no covariance between environmental values

Aside 1: Heritability

Aside 2: Genetic covariances between family members

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

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

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

Aside 3: Sharing of dominance effects

When individuals share the same genotype by descent

- So parents and children do not share dominance effects (in absence of inbreeding)
- At any locus the probability that full-sibs share the same genotype is ¼
- Therefore across all loci they share ¼ of the dominance variance

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 $\quad \widehat{h_o^2} = \frac{(10-1)0.0085}{\frac{1}{2}(1-0.0085)} = 0.154$

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

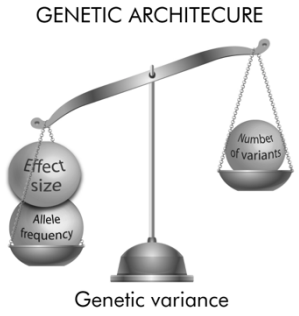
James (1971) Frequency in relatives for an all-or-non trait Ann Hum Genet 35 47

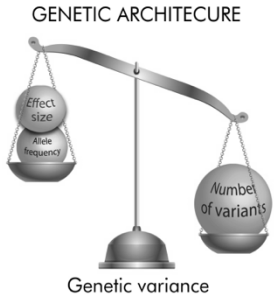
All we measure is recurrence risks to relatives what can we conclude about genetic architecture?

What do we mean by genetic architecture?

Genetic Architecture

- Number
- Frequency
- Effect size
- Interaction of genetic variants





All we measure is recurrence risks to relatives what can we conclude about genetic architecture?

Many familial diseases of a dichotomous nature appear to have a substantial genetic component, yet are not transmitted in a simple Mendelian fashion. For these disorders a great deal of effort has been expended in an attempt to characterize their transmission Suarez et al, 1978, AnnHG

e two models most often fitted to a body of (presence/absence) data are the major single locus model and a polygenic or multifactorial model. These two models represent opposite extremes on the continuum of genetic transmission and it is felt that if these models cannot be distinguished from each other there is little point in attempting to fit an intermediate oligogenic model (cf. Wilson, 1974). It has been shown that over a wide range of conditions it is often difficult to discriminate between these two models (Smith,
Suarez et al, 1976, AnnHG

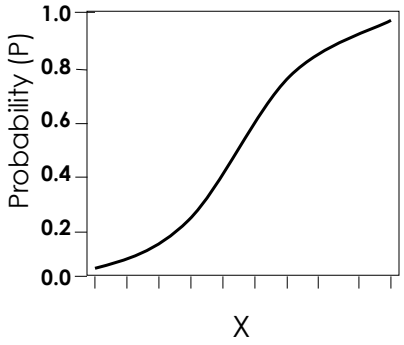
For schizophrenia (Risch 1990; Craddock et al, 1995)

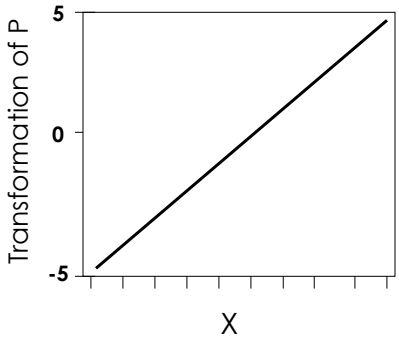
- recurrence risks not consistent with a single locus model
- need estimates from multiple relatives to make more detailed conclusions
- difficult to separate oligogenic from polygenic/multifactorial models

Suarez et al (1976) Limits of the general 2-allele single locus model with incomplete penetrance Ann HG
Risch (1990) Linkage Strategies for Genetically Complex Traits *AJHG*
Craddock et al (1995) The mathematical limits of multi-locus models AJHG

Additivity of effects depends on scale

In analysis of data we commonly make transformations between scales





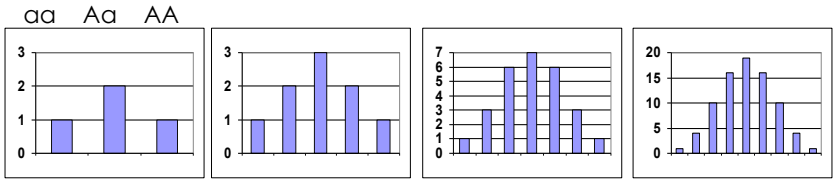
In trying to understand observed risks to relatives makes sense and is consistent with the laws of inheritance we seek a transformation of probability of disease

27

Complex diseases are multifactorial

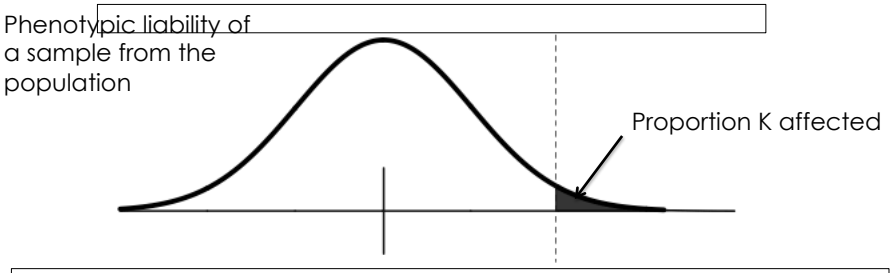
- Multiple genetic effects
- Multiple non-genetic effects including
  - Known environmental risk factors
  - Unknown environmental risk factors
  - Stochastic/chance events
- Things caused by many factors tend to be normally distributed

1 Locus	**2 Loci**	**3 Loci**	**4 Loci**
Individuals carry 0,1,2 risk alleles	Individuals carry 0-4 risk alleles	Individuals carry 0-6 risk alleles	Individuals carry 0-8 risk alleles



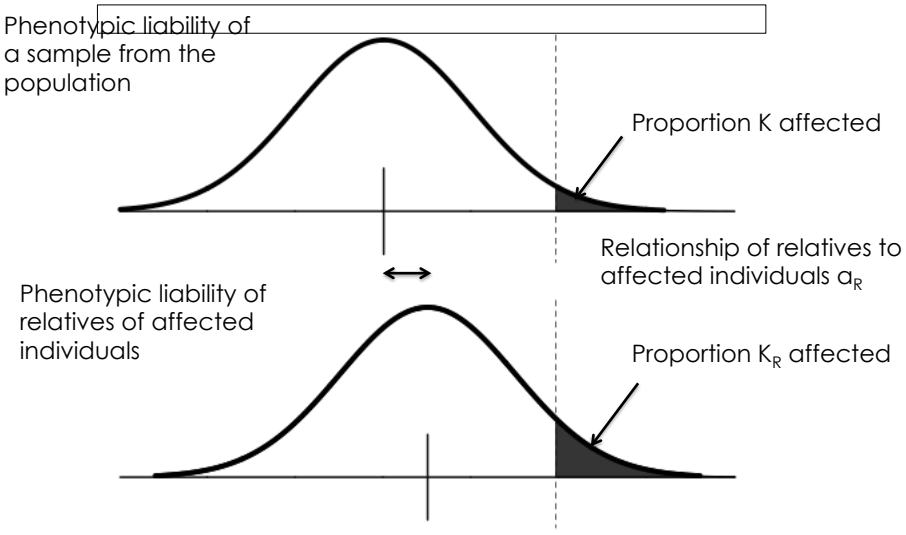
28

Liability threshold model



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)



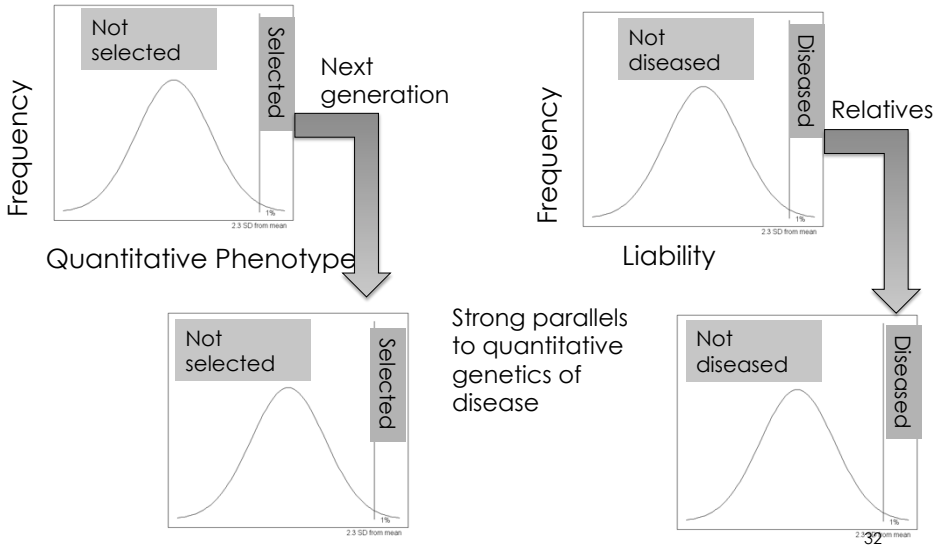
Using normal distribution theory what percentage of the variance in liability is attributale to genetic factors given K, K_R and a_R

A(n incomplete) history of the liability threshold model

- Pearson (1900) –used "tetrachoric correlation" to describe correlation between 2 dichotomous traits
- 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
- Robertson & Lerner (1949) provided the quantitative genetics maths for all-or-none traits – thinking about response to selection
- Dempster & Lerner (1950) with Appendix by Robertson – converts between heritability on the 0/1 and liability scales
- Crittenden (1961) & Falconer (1965) – provided the quantitative genetics maths – probably independently –for disease traits
- Smith (1970) – clarifies consistency of high heritability, low prevalence and high discordance of MZ twins
- Credit is generally given to Falconer (1965), but really it should be Robertson & Lerner (1949)

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
Robertson & Lerner (1949) The heritability of all-or-none traits: viability of poultry. Genetics
Dempster & Lerner (1950) Heritability of threshold characters. Genetics 35
Crittenden (1961) an interpretation o familial aggregation based on multiple genetic and environmental factors Ann NY Acad Sci 91 769
Falconer (1965) The inheritance of liability to certain diseases, estimated from incidences in relatives, Ann. Hum Genet. 29 51
Smith (1970) Heritabiity of liability and concordance in monozygous twins Ann Hum Genet 34: 85

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



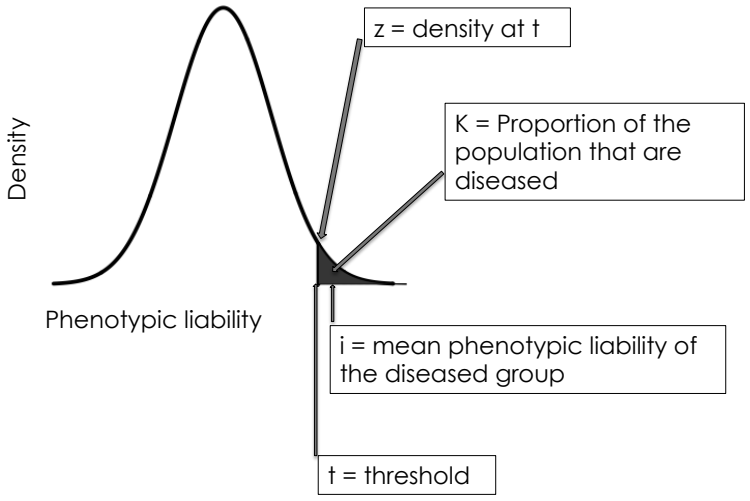
Heritage of the Edinburgh School

Douglas Falconer
Falconer, 1965 Annals of Human Genetics
The incidence of liability to certain diseases, estimated from incidence amongst relatives
Falconer & McKay text book

Charlie Smith
Smith, 1970 Annals of Human Genetics
Heritability of liability and concordance of monozygotic twins

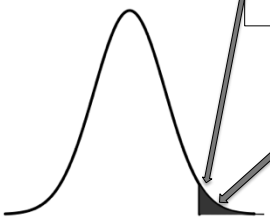
Alan Robertson
Appendix of Dempster and Lerner, Genetics,1950
Heritability of threshold characters

Definitions



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



Standard Deviation =1
$\sigma_p = 1$

Phenotypic liability

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)

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= Φ^{-1}(1-K) = qnorm(1-K)

Inverse standard normal distribution (probit) function

Mean of diseased group

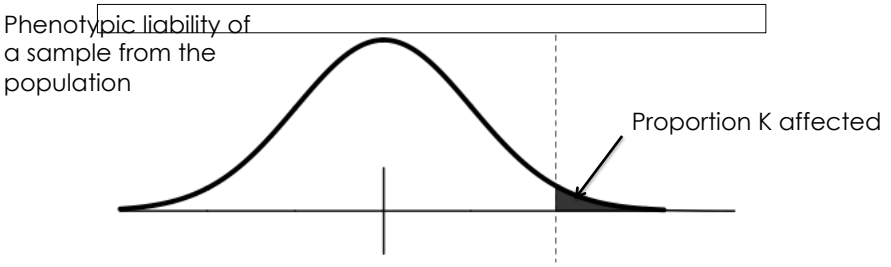
- Pearson & Lee (1908) On the generalized probable error in normal correlation. Biometrika
- Lee (1915) Table of Gaussian tail functions..Biometrika
- Fisher (1941) Properties and application of Hh functions. Introduction to mathematical tables
- Cohen (1949) On estimating the mean and standard deviation 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

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

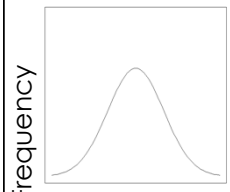
Lynch and Walsh equations 2.13 and 2.14; variance equation 2.15

Falconer (1965)



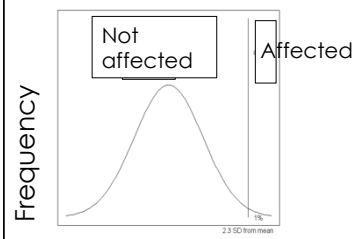
Assumption of normality
- Only appropriate for multifactorial disease
- i.e. more than a few genes but doesn't have to be highly polygenic
- Key – unimodal

Visualising heritability of the "disease" of loftiness



Short families

Tall families

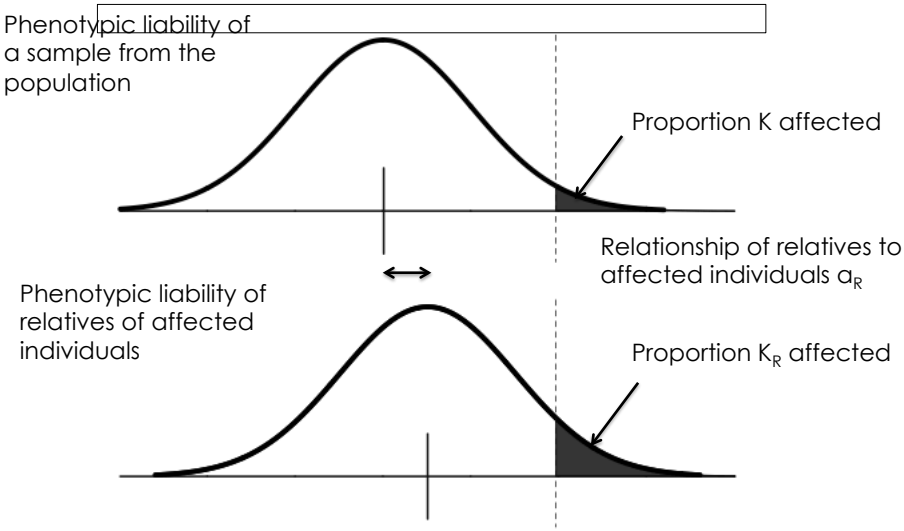


Liability of risk

Would be common to find more than one 'lofty' in a family

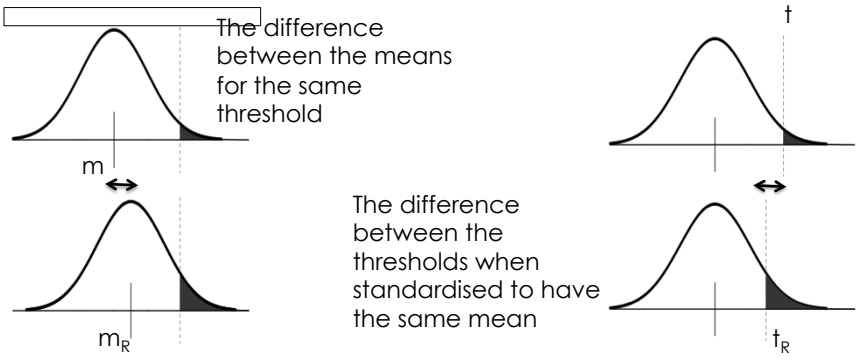
Heritability of height ~0.8
Heritability of schizophrenia ~0.8

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

Falconer (1965)

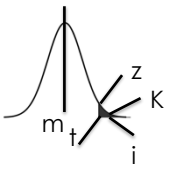


m_R-m = t-t_R

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 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

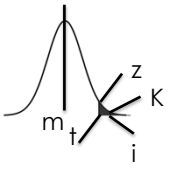
$Y = \mu_Y + b_{Y,X}(X-\mu_x) + \varepsilon$

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

Graph calculator

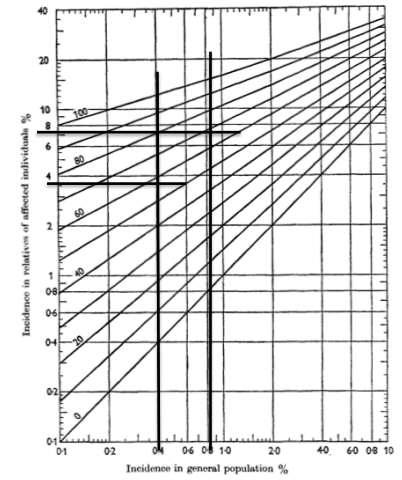


Fig. 4. Graph for estimating the heritability of liability from two observed incidence when the relatives are sibs, parents, or children. Explanation in text.

McGue et al K =0.85%
λ_{sib} = 8.6
K_R= 8.6*0.85 = 7.31%
h^2 ~70%

Lichtenstein et al K =0.407%
λ_{sib} = 8.6
K_R= 8.6*0.407 = 3.5%
h^2~60%

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

Assumptions made by Falconer (1965)

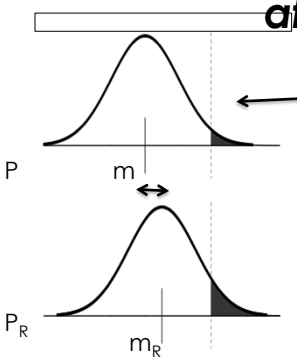
Assumption: Covariance between relatives reflects only shared additive genetic effects

Check: Use different types of relatives with different $a_R$and 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

…..we'll take a look by simulation

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

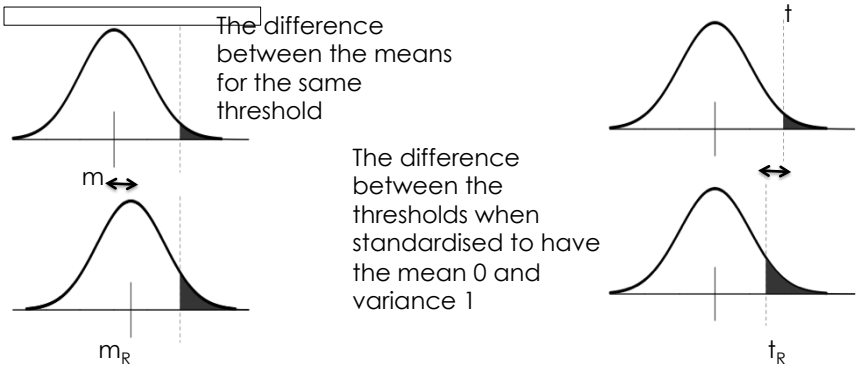


Variance in liability amongst the diseased individuals
= σ_p^2 (1-k), where k = i(i-t)

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

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

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

For affected individuals X = i
Expected phenotypic liability of relatives of those affected
$E(Y \mid X>t) = m_R - m = t - t_R\sqrt{1 - ka_R^2h^4}$

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

Sampling variance?

Variance of the estimate of a threshold based on the estimate of prevalence K

$$V(\hat{t}) = \left(\frac{dt}{dK}\right)^2 V(\hat{K}) = \left(\frac{-1}{z}\right)^2 V(\hat{K}) = \frac{1}{z^2}\frac{K(1-K)}{N}$$

$$h^2 = \frac{t - t_R}{i a_R}$$

If K has been estimated from a very large sample then we can ignore the sampling variance of t and i

$$V(\hat{h^2}) = \frac{1}{i^2 a_R^2} V(\widehat{t_R})$$

$$= \frac{1}{i^2 a_R^2 z_R^2}\frac{K_R(1-K_R)}{N_R} \quad = \frac{(1-K_R)}{a_R^2 i^2 i_R^2 K_R N_R} \quad = \frac{(1-K_R)}{a_R^2 i^2 i_R^2 N_{RA}}$$

Where N_{RA} is the number of affected relatives

Sampling variance?

$$h^2 = \frac{t - t_R}{i a_R}$$

If K also has sampling variance associated, then

$$V(\widehat{h^2}) = \left(\frac{1}{i a_R} - \widehat{h^2}(i - t)\right)^2 V(\hat{t}) + \left(\frac{1}{i a_R}\right)^2 V(\widehat{t_R})$$

$$V(\widehat{h^2}) = \left(\frac{1}{i a_R} - \widehat{h^2}(i - t)\right)^2 \frac{K(1-K)}{z^2 N} + \left(\frac{1}{i a_R}\right)^2 \frac{K_R(1-K_R)}{z_R^2 N_R}$$

$$s.e(\widehat{h^2}) = \frac{1}{i a_R}\sqrt{\left[\frac{(1-K)}{i^2 K N}\right](1 + a_R h^2 k)^2 + \frac{(1-K_R)}{i_R^2 N_R K_R}} \qquad k = i\,(i - t)$$

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

Assumptions in estimation of heritability of liability

Assumption: Covariance between relatives reflects only shared additive genetic effects

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

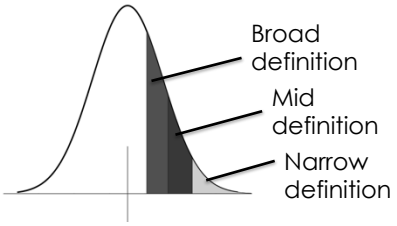
- In particular 1st degree relatives, MZ twins children of two affected parents

Low prevalence high heritability diseases will show more consistency between estimates of narrow sense heritability based on different types of family members on liability scale than on disease scale.

Empirical estimates suggest that non-additive genetic effects on the liability scale and shared family effects and are small

Multiple thresholds

Collect narrow and broader definitions of disease (eg psychiatric) in cohorts of families – see if risks are consistent with a single liability distribution.



Proband

	Sibs: Broad	Sibs: Mid	Sibs: Narrow
Broad	λ_{BB}	λ_{BM}	λ_{BN}
Mid	λ_{MB}	λ_{MM}	λ_{MN}
Narrow	λ_{NB}	λ_{NM}	λ_{NN}

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

Misunderstandings in MZ twins

"Estimates of heritability of liability to a number of diseases turned out to be quite high. This was a surprise to those not versed in quantitative genetics who tended to think that low concordances rates in MZ twins implied low heritability." Fraser (1976)

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

Smith (1970) MZ twins

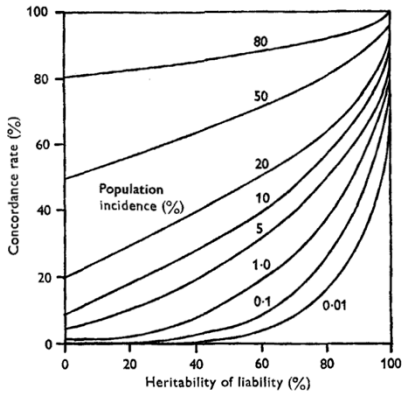


Fig. 2. Expected concordance rate in monozygotic (MZ) twins given the population incidence and the heritability of liability.

For rare diseases concordance is not expected to be high even when heritability is high

…..no need to invoke epigenetics

Smith (1970) Heritabiity of liability and concordance in monozygous twins Ann Hum Genet 34: 85

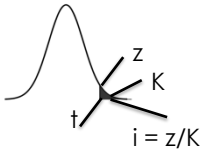
Relationship between heritabilities on disease and liability scales

Consider a linear regression of genetic values on the disease scale (A_{01}) on genetic values on the liability scale (A_L):

$A_{01} = \mu + bA_L$ $\quad b = \frac{cov(A_{01},A_L)}{var(A_L)}$

$Var(A_{01}) = b^2 Var(A_L) = \frac{cov(A_{01},A_L)^2}{var(A_L)}$ by differential calculus normal distribution theory….

$$h_{01}^2 = \frac{z^2 h_L^2}{K(1-K)} = \frac{i^2 K h_L^2}{(1-K)}$$



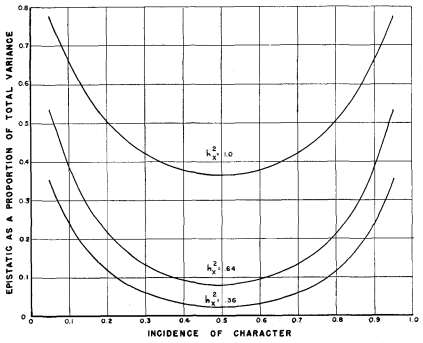
$$h_L^2 = \frac{(1-K)h_{01}^2}{i^2 K}$$

Robertson (1950) Appendix of Dempster & Lerner (1950) Heritability of threshold characters. Genetics 35

Relationship between heritability on the disease and liability scales

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

On the disease scale



Lines are heritability of liability

= Prevalence

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

Estimation of heritability on large samples

- Linear mixed model
  - Accounts for different relationships between individuals
  - Include fixed effects eg sex
  - Maximum likelihood estimate of the threshold that best fits the data
- The "information" essence comes from the risks to relatives

Lichtenstein et al (2009) Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet 373: 234.

Wray & Gottesman (2012) Using summary data from the Danish National Registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder . Frontiers in Genetics

Assumption – familiality is only caused by additive genetic factors

- Use different types of relatives to estimate heritability of liability
- Check for consistency between estimates

- Empirical estimates suggest shared family effects and non-additive genetic effects are small

Practical

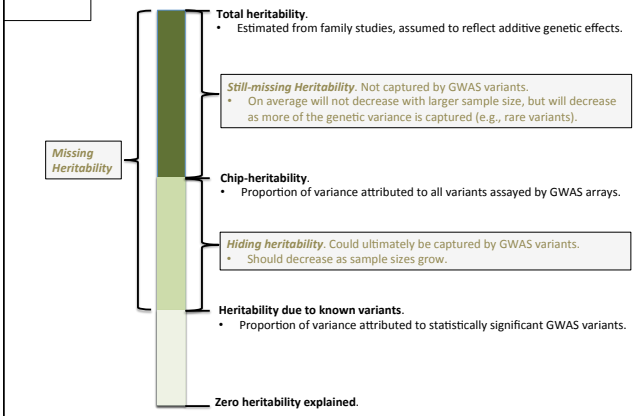
Uses simulation to give understanding to the theory.

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

Take home message 1



- Don't over-interpret
- Just a benchmark
- Often subject to large sampling variance

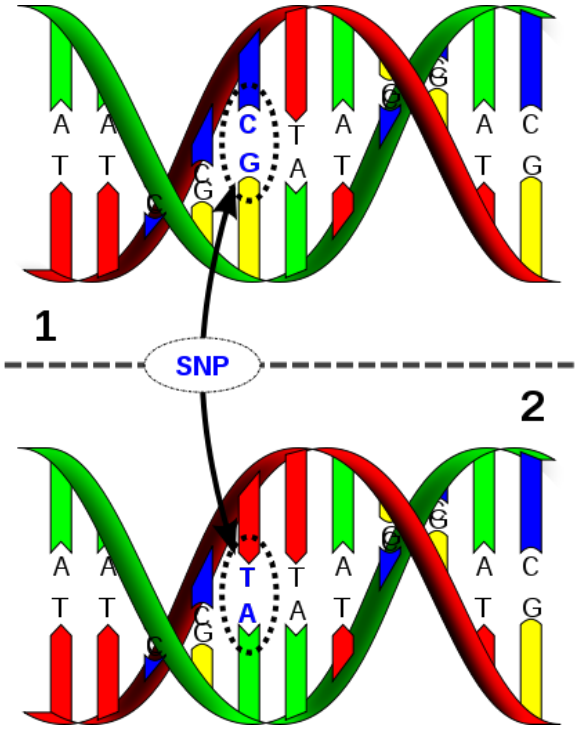
Module 19: Statistical and Quantitative Genetics of Disease

John Witte

Lecture #2

Now Assume we have DNA

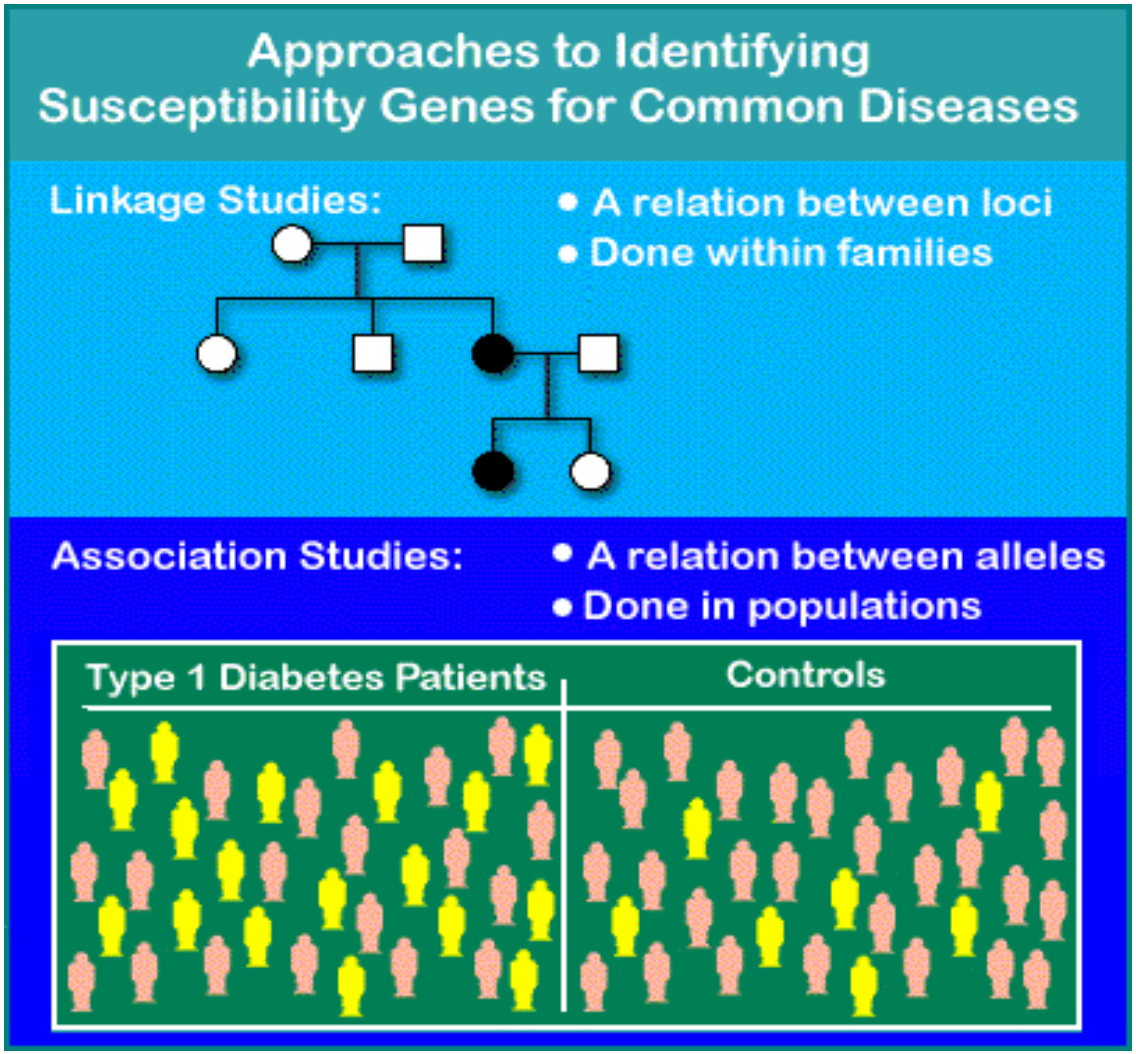
- Building on the previous lecture… what next?
- Estimate impact of genetic variants on disease.
- Who should we study?
- What analytic approaches?
- Once detect risk variants, how much variation do they explain.
- Can we make inferences about genetic architecture?



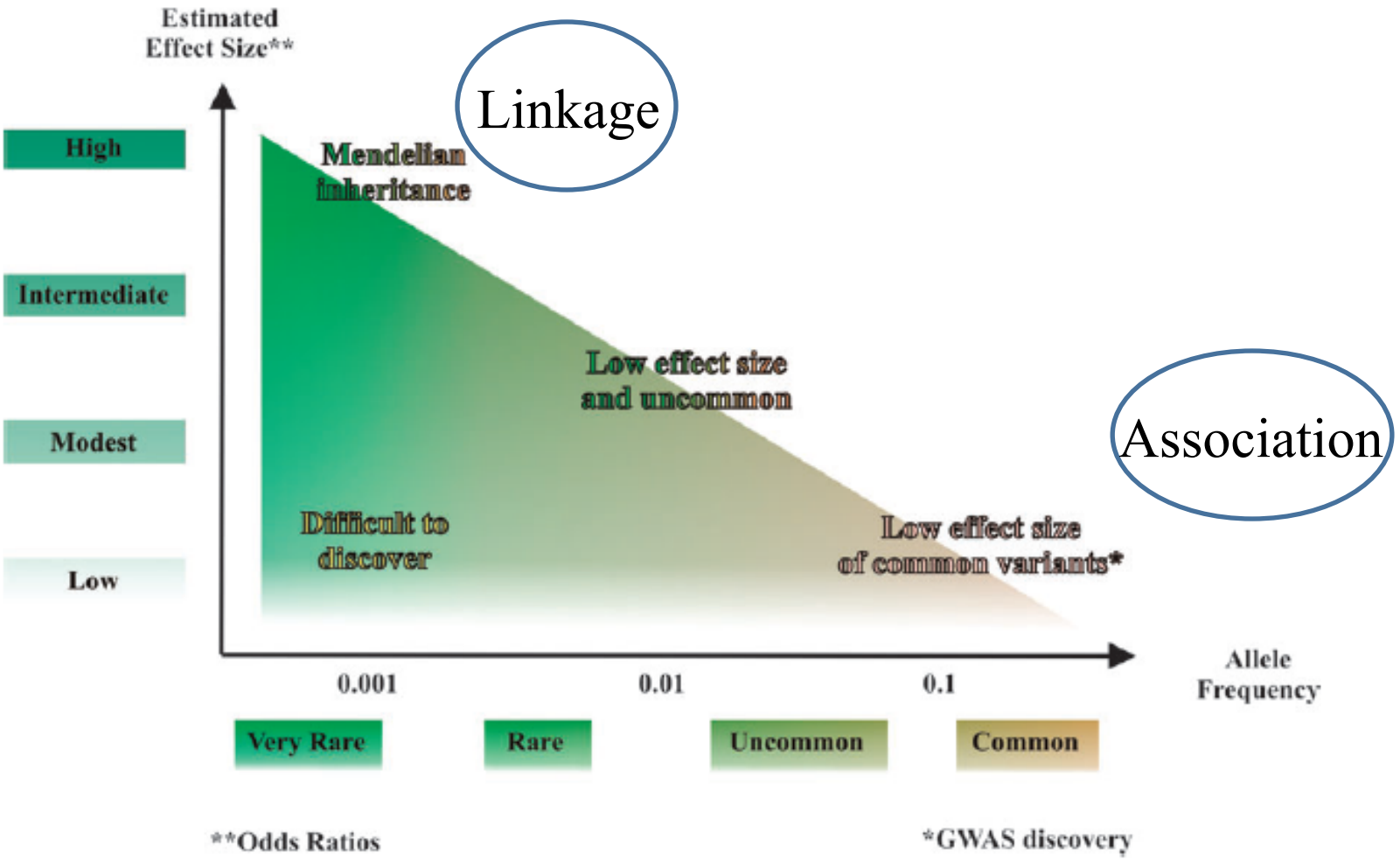
Outline

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

1. Linkage *versus* Association



Linkage *versus* Association



Chung et al. Carcinogenesis 31:111-120, 2010

Linkage *versus* Association

Exposure Frequency

Effect Size

Identity By Descent

Sample Size Required

Linkage			
Genotypic risk ratio (γ)	Frequency of disease allele A (p)	Probability of allele sharing (Y)	No. of families required (N)
4.0	0.01	0.520	4260
	0.10	0.597	185
	0.50	0.576	297
	0.80	0.529	2013
2.0	0.01	0.502	296,710
	0.10	0.518	5382
	0.50	0.526	2498
	0.80	0.512	11,917
1.5	0.01	0.501	4,620,807
	0.10	0.505	67,816
	0.50	0.510	17,997
	0.80	0.505	67,816

Comparison of linkage and association studies. Number of families needed for identification of a disease gene.

Risch and Merikangas, 1996 Science

Linkage *versus* Association

		Linkage			Association			
					Singletons		Sib pairs	
Genotypic risk ratio (γ)	Frequency of disease allele A (p)	Probability of allele sharing (Y)	No. of families required (N)	Probability of transmitting disease allele A P(tr-A)	Proportion of heterozygous parents (Het)	(N)	(Het)	(N)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

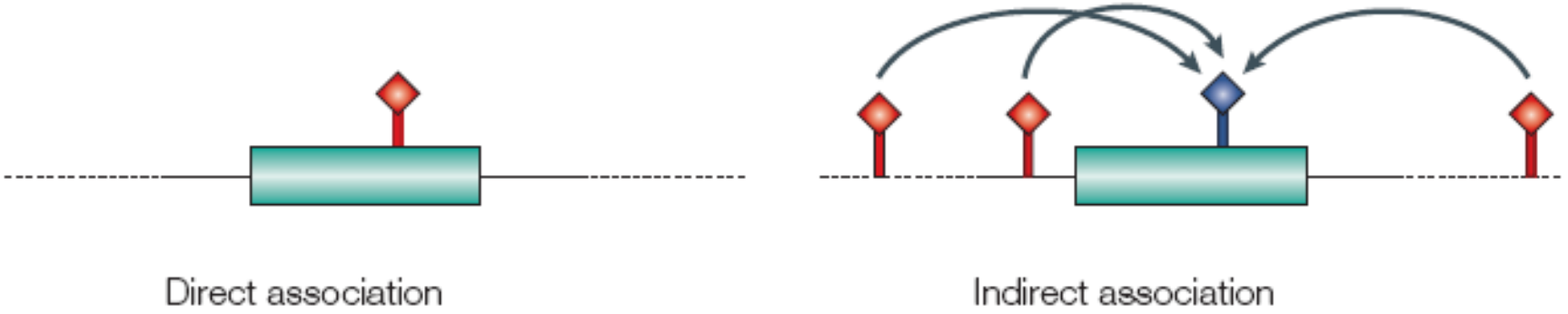
Comparison of linkage and association studies. Number of families needed for identification of a disease gene.

Risch and Merikangas, 1996 Science

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



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

Hirschhorn & Daly, Nat Rev Genet 2005

Assume two loci, A and B
Locus A has two alleles A and a
Locus B has two alleles B and b

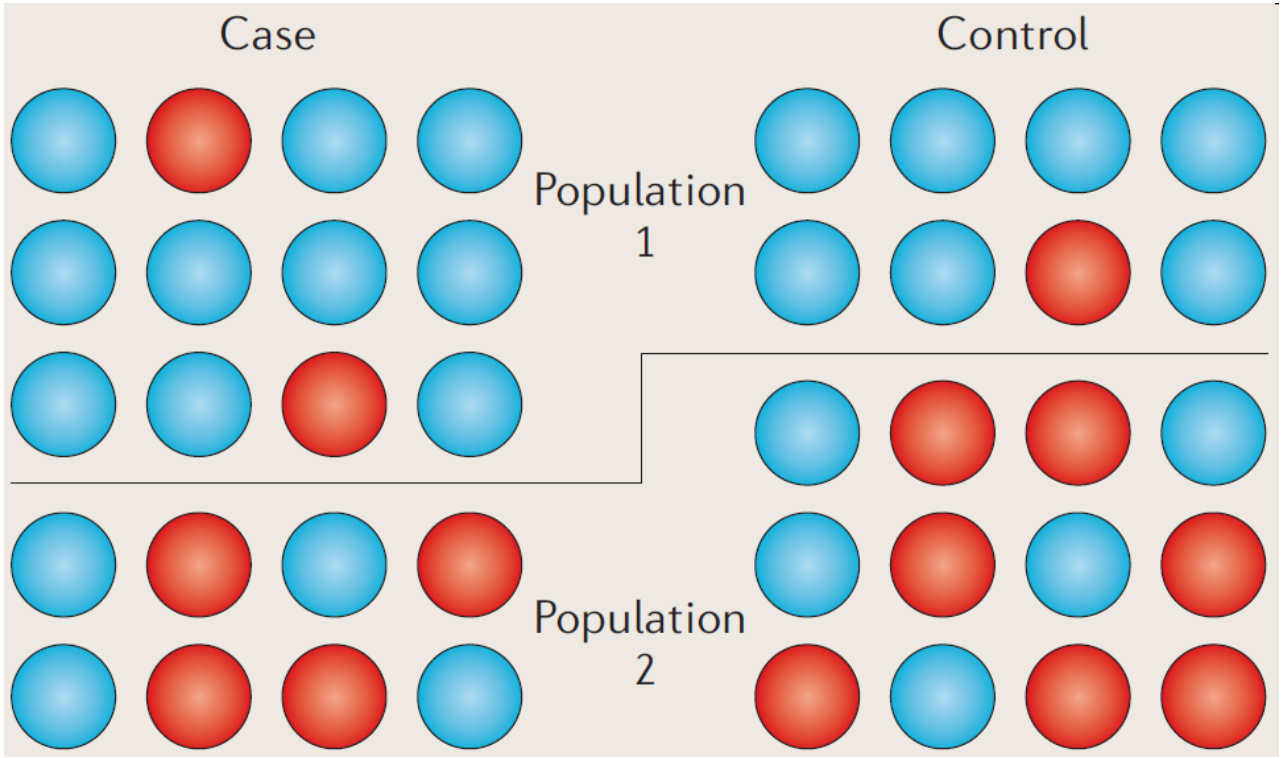
$D = P_{AB} - p_A p_B = P_{AB}P_{ab} - P_{Ab}P_{aB}$
$D' = D/\max(D)$
where
$\max(D) = \min(p_A p_b, p_a p_B)$ if $D>0$ or
$\min(p_A p_B, p_a p_b)$ if $D<0$
$r^2 = D^2/(p_A p_a p_B p_b)$.

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 (FP and FN).



Balding, Nature Reviews Genetics 2010

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 ($\chi^2_{new} = \chi^2_{old}/\lambda$).
- Principal Components
  (Price et al., Nat Genet 2006)
  Adjust regression for PCs as a proxy for genetic ancestry.
- Mixed model (Yang et al. Nat Genet 2010)

Point/Counterpoint

Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations?[1]

Duncan C. Thomas [2] and John S. Witte

Department of Preventive Medicine, University of Southern California, Los Angeles, California 90033-9987 [D. C. T.], and Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106-4945 [J. S. W.]

easily arise in recently admixed populations for example, will often not be replicated in different populations but are nevertheless "interesting" as an indication that a causal gene may be in the general region. [In fact, studying admixed populations can benefit linkage disequilibrium mapping (8).] Although

Point/Counterpoint

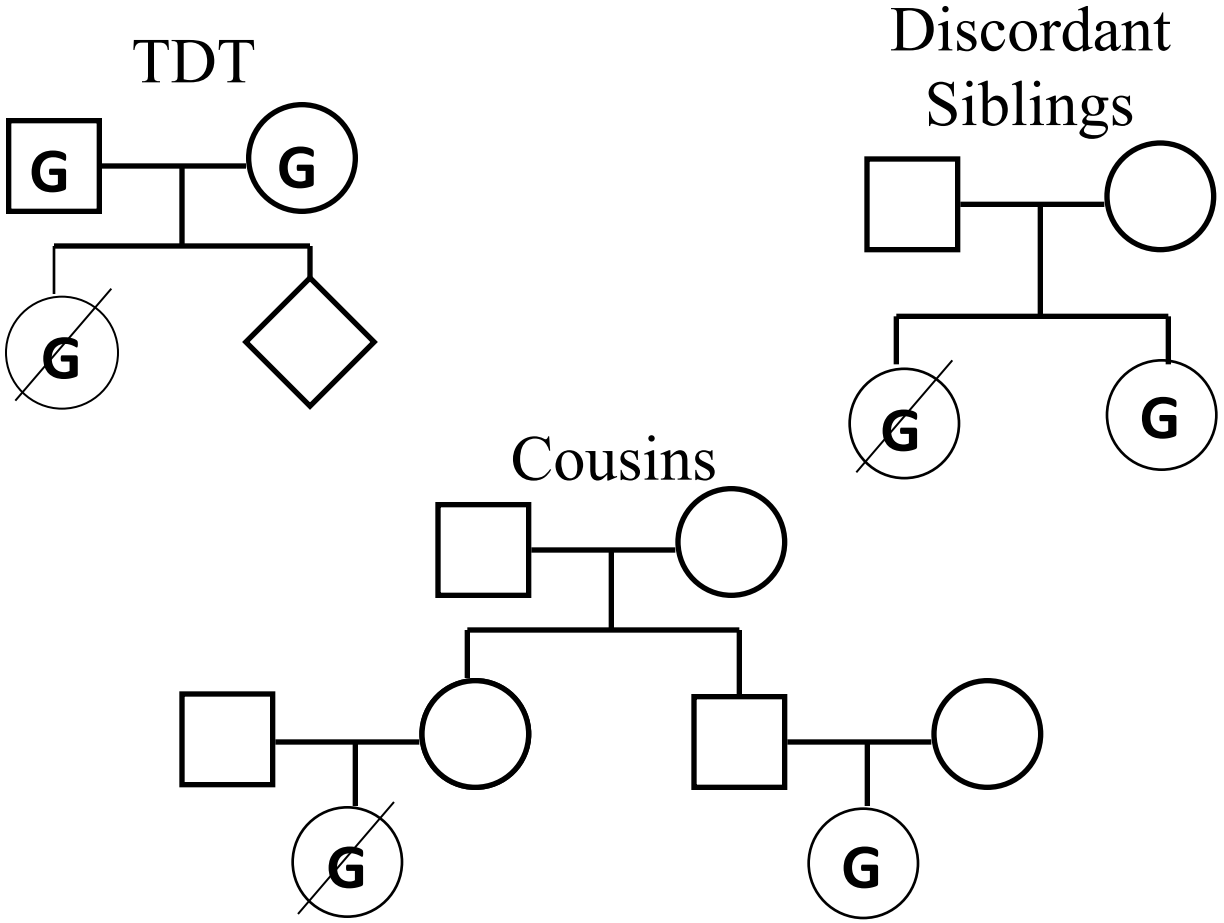
Counterpoint: Bias from Population Stratification Is Not a Major Threat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer

Sholom Wacholder,[1] Nathaniel Rothman, and Neil Caporaso

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20854

uals at similar risk. Similarly, a gradient in risk of disease by socioeconomic status does not itself cause disease but may be a reflection of similarity in lifestyle or access to preventive health care. Whether the consideration is ethnicity or socioeconomic status, controlling for the factors that explain

Family-Based Studies



Transmission Disequilibrium Test (TDT)

- Based on transmissions from parents to affected offspring.
- Compare observed to expected transmissions.
- Need parents' and offspring genotypes.

Transmitted parental allele		Non-transmitted parental allele: A	Non-transmitted parental allele: a
	A	w	x
	a	y	z

- $(x-y)^2/(x+y) \sim \chi_1^2$; McNemar's test.
- No information in w or z (homozygous parents).
- Tests whether there is an excess of the A allele in the affected offspring than expected based on Mendel's laws.

Transmission Disequilibrium Test (TDT)

- Example: 94 families
  - 78 parents transmit allele A
  - 46 transmit allele a

Transmitted parental allele		Non-transmitted parental allele: A	Non-transmitted parental allele: a
	A	.	78
	a	46	.

- χ_1^2: $(78-46)^2/(78+46)=8.26$, p-value=0.004
- Note: for rare variants the observed transmission probability from parents to offspring may be less than the expected value, leading to conservative tests.

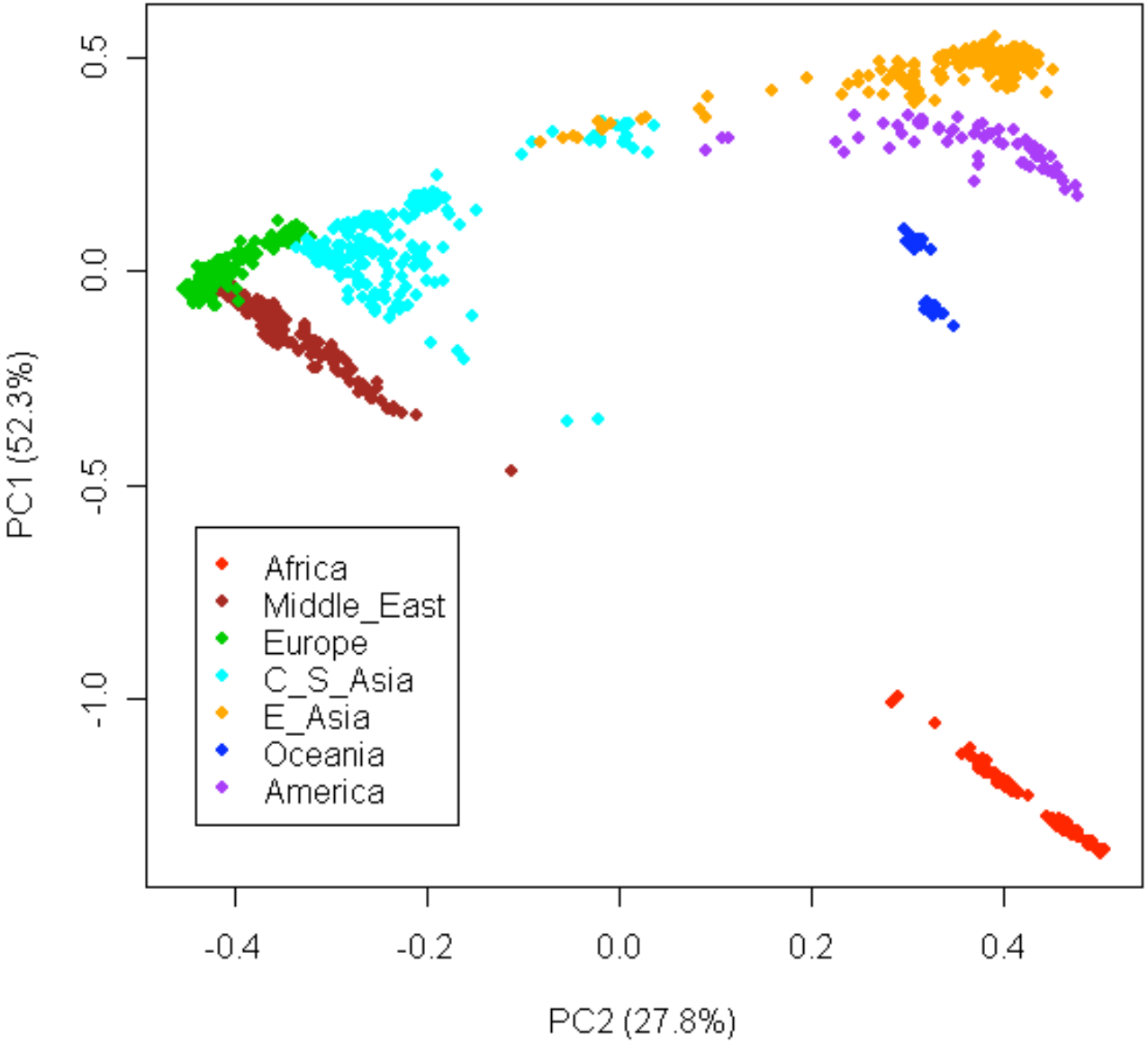
Comparison of Designs

- 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

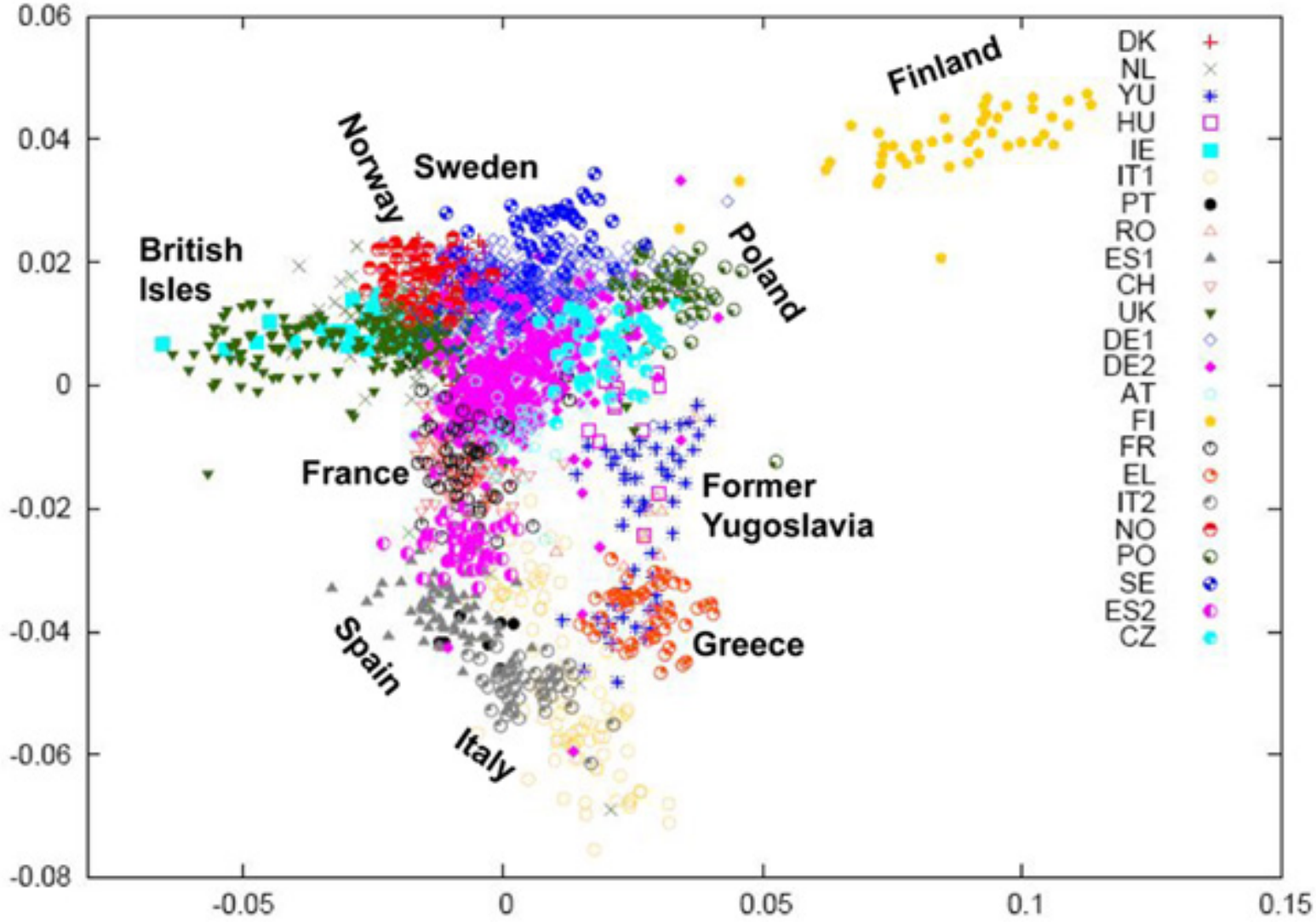
Adjusting for Principal Components



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

Li et al., Science 2008

PCs Detect Fine Population Structure



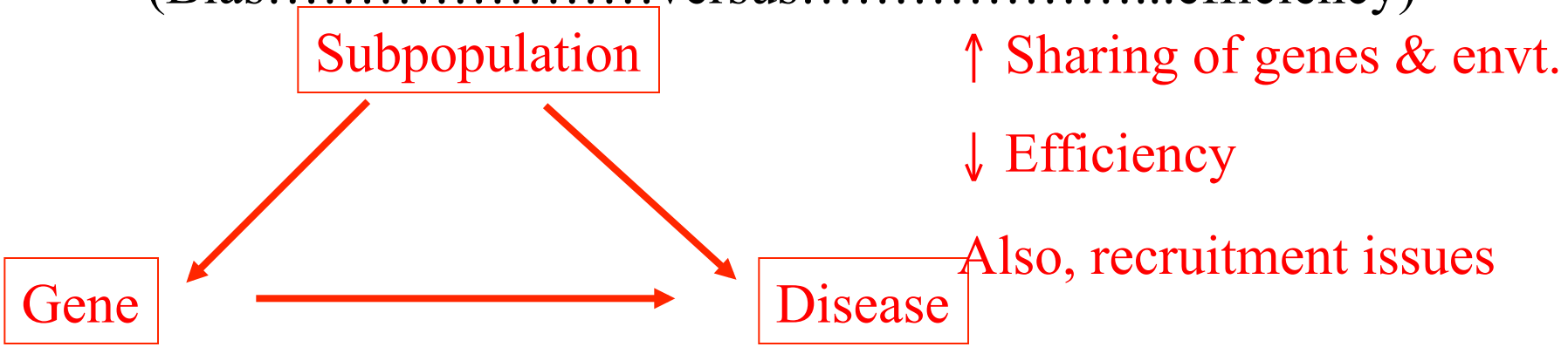
Razib, Current Biology 2008

Continuum of Assoc Study Designs

Population Stratification

Overmatching

(Bias……versus……..efficiency)



↑ Sharing of genes & envt.

↓ Efficiency

Also, recruitment issues

Outline

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

4. Association Analysis

Observed:

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

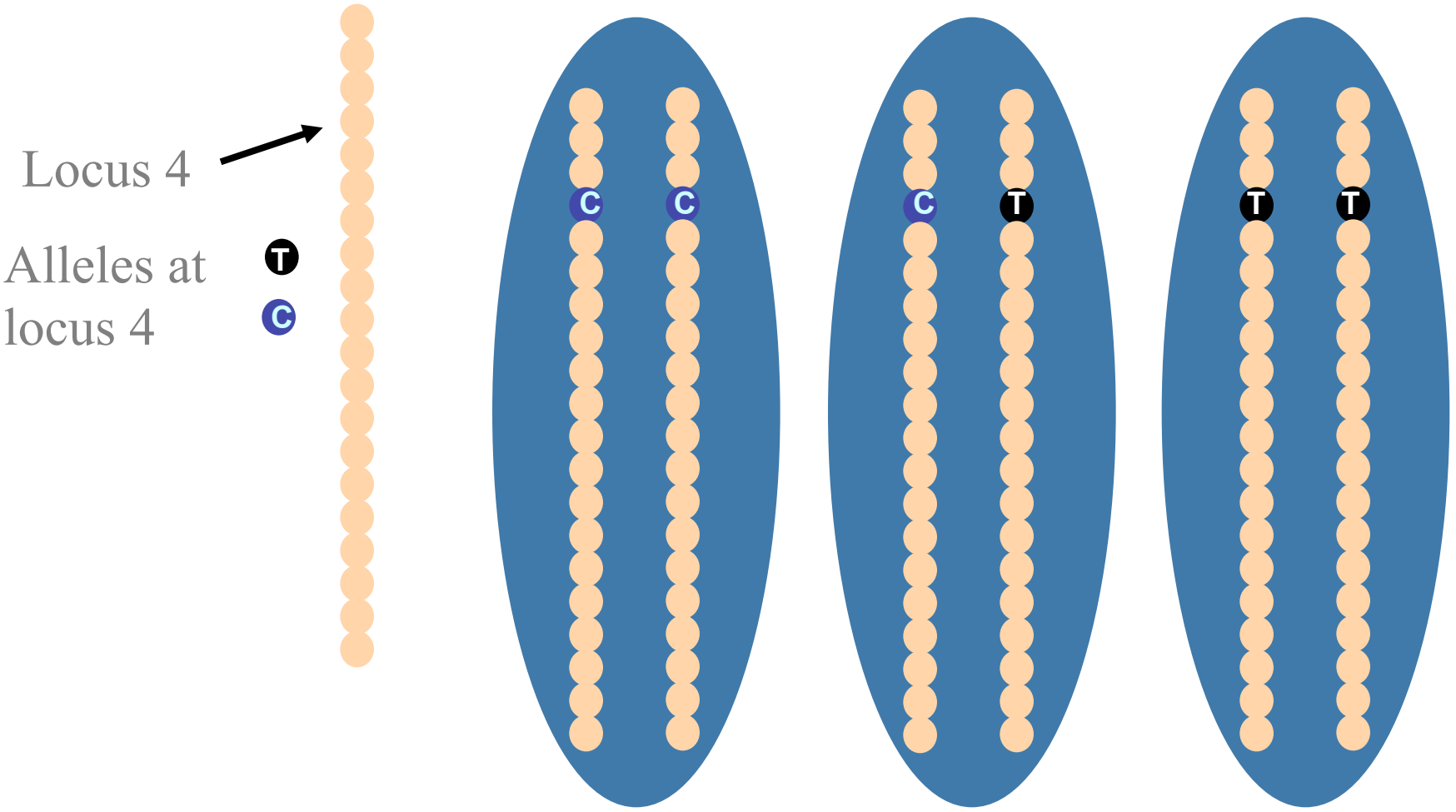
Expected

Case	Control
nCC*nCase/n	nCC*nCont/n
nCT*nCase/n	nCT*nCont/n
nTT*nCase/n	nTT*nCont/n

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

Expected cell count = row_total * column_total / total

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

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)

Association Analysis

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						

What model should we use here?

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

Testing Association

If genetic model known:

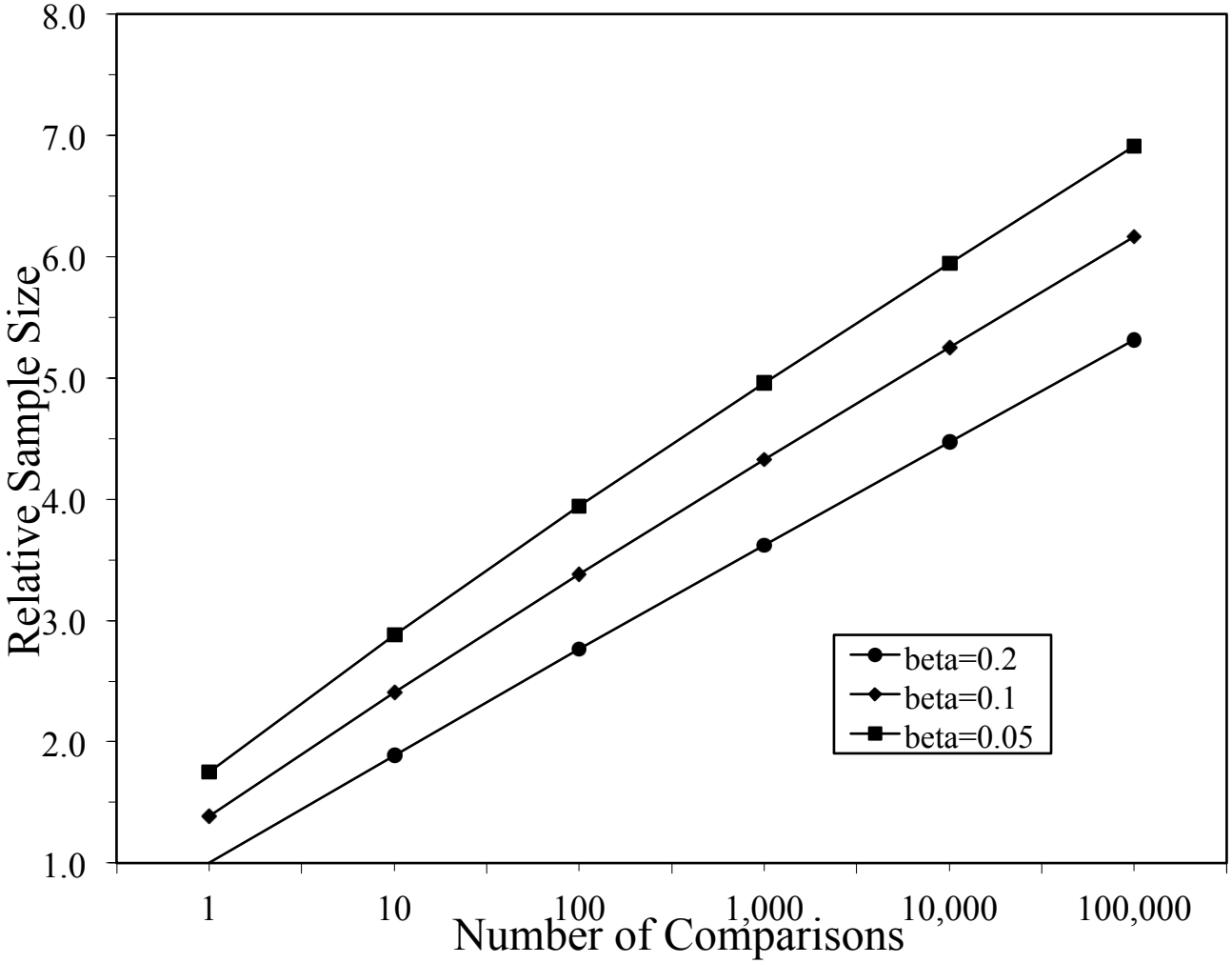
- Collapse genotypes into 2x2 table, 1 d.f. test
- Or trend test for log additive
- Use logistic regression: coding; covariates, odds ratios

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).

Power for Detecting Association



Witte et al., Stat Med 2000

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

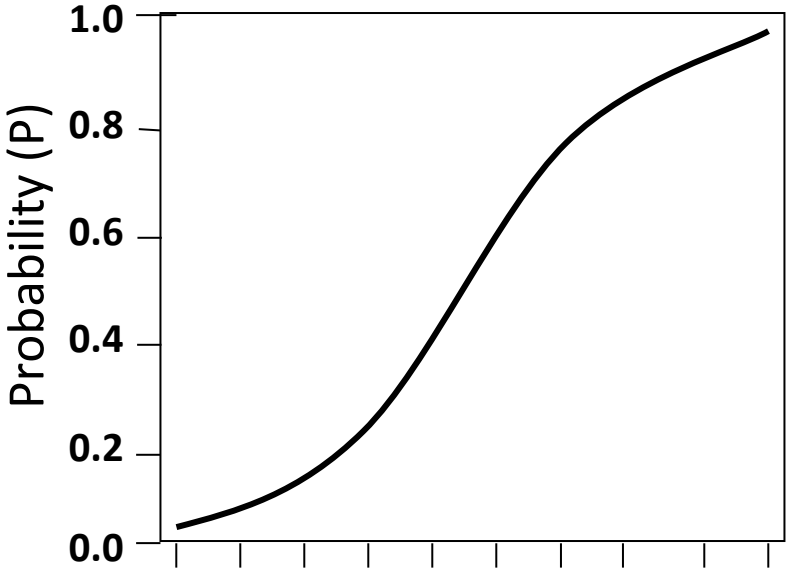
T_0 = Amount of unexposed person-time

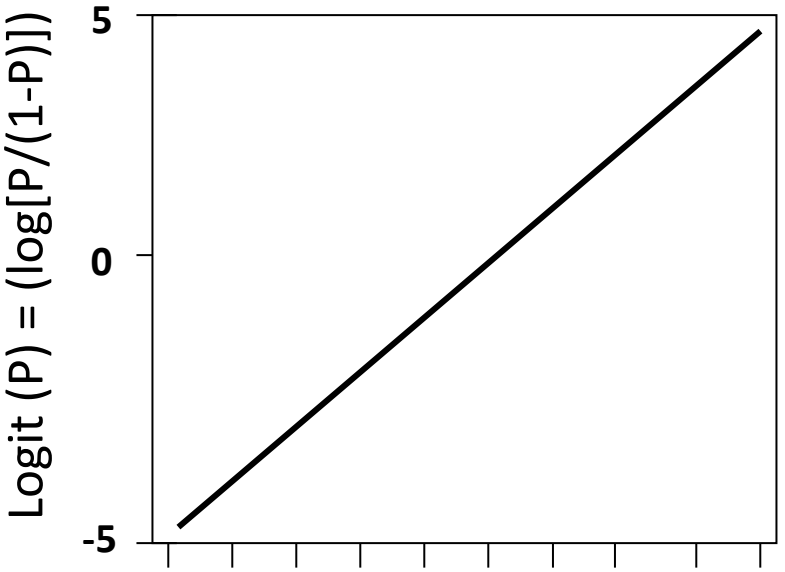
I_0 = Incident rate of unexposed

r = Sampling rate

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

6. Logistic Regression





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

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

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

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

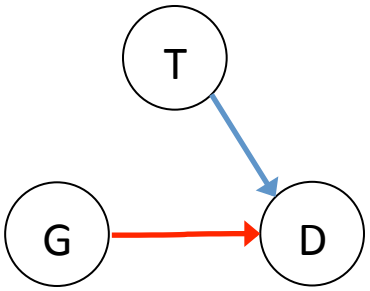
Or

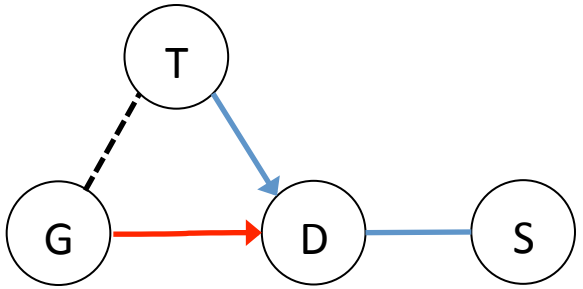
$$\log[P_1 /(1 – P_1) / (P_0 /(1 – P_0))] = \log (OR) = \beta$$

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

Including Covariates in Regression

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





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

Outline

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

7. Rare Variants (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

Goals:

- Identify new genes.
- Improve risk prediction.
- Help explain missing heritability.

Rare Variants & Rare Monogenic Traits

Paper	Phenotype	N (Affected)	Comparison
Exomes			
Ng (1)	FSS*	4 (unrelated)	HapMap exomes, dbSNP
Ng (2)	Miller	4 (2 sibs, 2 unrelated)	HapMap exomes, dbSNP
Genomes			
Roach (3)	Miller S, PCD	2 (sibs=Ng (2))	2 parents
Lupski (4)	CMT**	4 (sibs) (1 sequenced)	4 unaffected sibs, parents

*Freeman-Sheldon syndrome

**Charcot-Marie-Tooth disease

(1) Nat 2009; (2) Nat Genet 2009; (3) Science 2010 (4) NEJM 2010

Single Variant Analysis

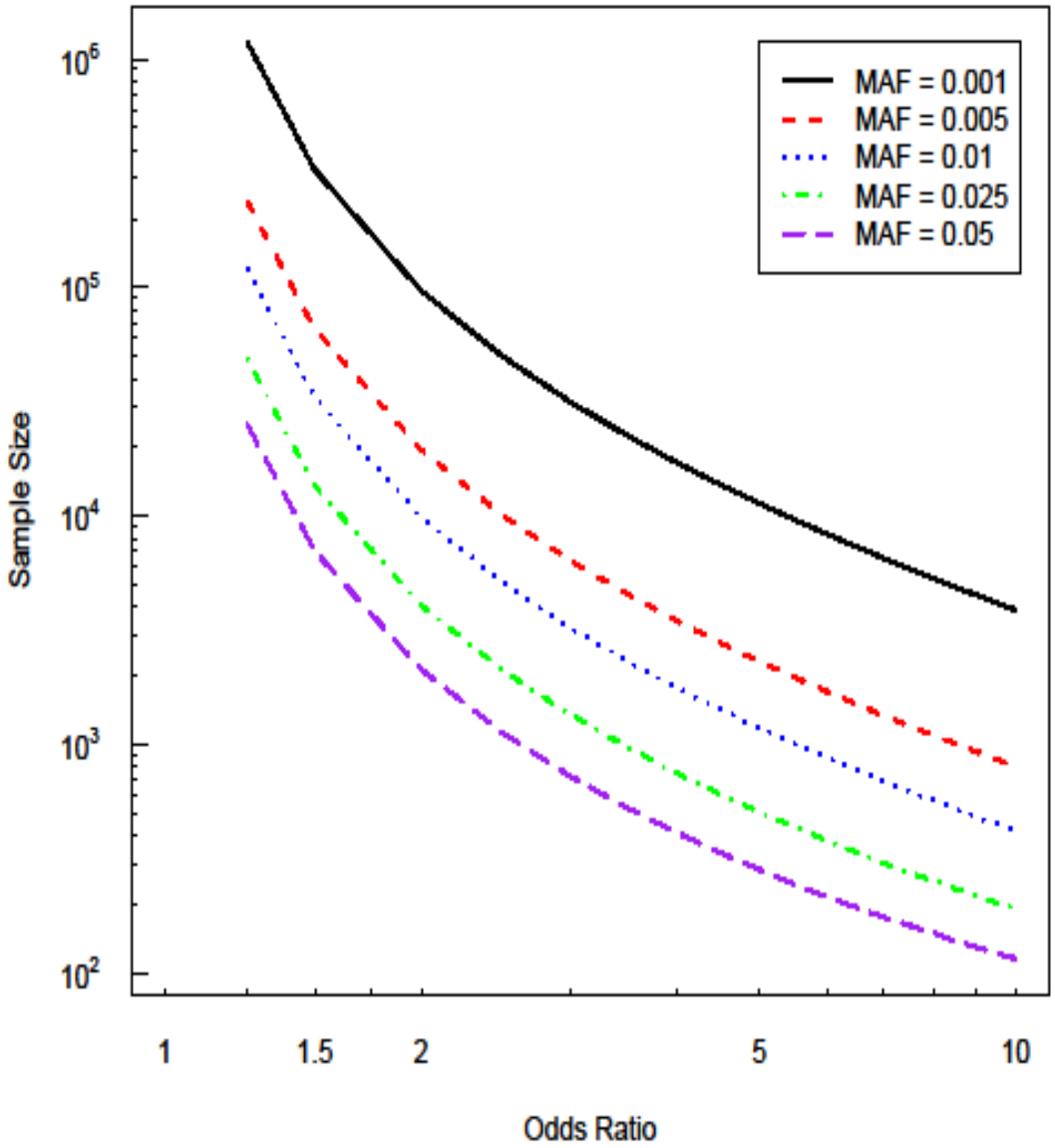
- Low power unless sample size is very large.

IFIH1 rs35744605 Genotype	Controls	T1D Cases	OR (p-value)
GG	9621	8109	1.0
GT	131	76	0.69 (9×10^{-3})
TT	0	0	-

MAF = (76 + 131) / [76 + 131 + 2*(9621 + 8109)] = 0.0058

Nejentsev…Todd. Science 2009;324:387.

Sample Size for Rare / Less Common Variants



Analysis of Rare Variants

Focus on a set of k variants

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

- Also difficult to model due to sparsity.
- Instead up-weight analyses for most likely causal variants.

Rare Variant Tests

- 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 individual's 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)

$$i_k = I_{\{p_k < 0.01\}} f_k \qquad f_k = I_{\{k \in \text{Nonsynonymous}\}}$$

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 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 ("step-up") (Hoffmann, Marini & Witte, 2010)

Module 19: Statistical and Quantitative Genetics of Disease

John Witte

Lecture #3

Outline

1. Hierarchical Modeling
2. Interactions
3. Measures of Variation Explained

1. Hierarchical Modeling

- Approaches have existed for > 50 years.
- The following have hierarchical components:
  - Variance components
  - Empirical-Bayes
  - Penalized Regression

World Cup Goal Keeper Saves

- A measure of goalies' effectiveness is their proportion of shots saved.
- p_i = # of saves / total number of shots on goal.
- When only the first one or two world cup matches have been played, how can we best estimate their true p_i?
- This is analogous to estimation:
  - use a sample of data to try and estimate true parameters.

Conventional Estimators

Decide between:

1) each goalie's current save proportion,

$\hat{p}_i$ = # saves / # shots, i = 1, ... k;

2) the average save proportion across all goalies,

$$\bar{p} = \frac{\sum_{i=1}^{k} \hat{p}_i}{k}$$

1) might be unstable due to the limited number of observations (i.e., small sample size).
2) suggests no differences among the goalies.

Hierarchical Estimation

- Compromises between these extremes, pulling the individual estimates toward the group mean (relative to standard deviation).

- Assume *a priori* that they are exchangeable (i.e., before seeing the data, we can't distinguish among them):

  $p_i \sim N(\mu, \tau^2)$ (Prior)

- The hierarchical estimator for p_i is given by

$$p_i^{HM} = [\sigma_i^2 / (\sigma_i^2 + \tau_i^2)]\mu + [\tau_i^2 / (\tau_i^2 + \sigma_i^2)]\hat{p}_i$$

  where σ_i^2 = var(p_i).

Hierarchical Estimates

- Now $$\frac{\sigma_i^2}{\tau_i^2 + \sigma_i^2} = E\left[\frac{(k-3)\hat{\sigma}_i^2}{\sum_{i=1}^{k}(\hat{p}_i - \bar{p})^2}\right] \qquad \mu = E(\bar{p}) = \frac{\sum_{i=1}^{k}\hat{p}_i}{k}$$

- where $$\hat{\sigma}_i^2 = \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$$

- and n_i = # saves (assumes saves ~ binomial).

- Hierarchical estimates of the save proportion for goalie i

$$\tilde{p}_i^{HM} = \left[\frac{(k-3)\hat{\sigma}_i^2}{\sum_k(\hat{p}_i - \bar{p})^2}\right]\bar{p} + \left[1 - \frac{(k-3)\hat{\sigma}_i^2}{\sum_k(\hat{p}_i - \bar{p})^2}\right]\hat{p}_i$$

Comparison of Results

- Variance ~ 1/3 for hierarchical estimates in comparison with conventional ML estimates.

- In fact, $$E\left[\sum_{i=1}^{k} (p_i - \hat{p}_i)^2\right] > E\left[\sum_{i=1}^{k} (p_i - \tilde{p}_i^{HM})^2\right]$$

when k >= 4.

Back to Genetics…

- Conventional Analytics Approach

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

- 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

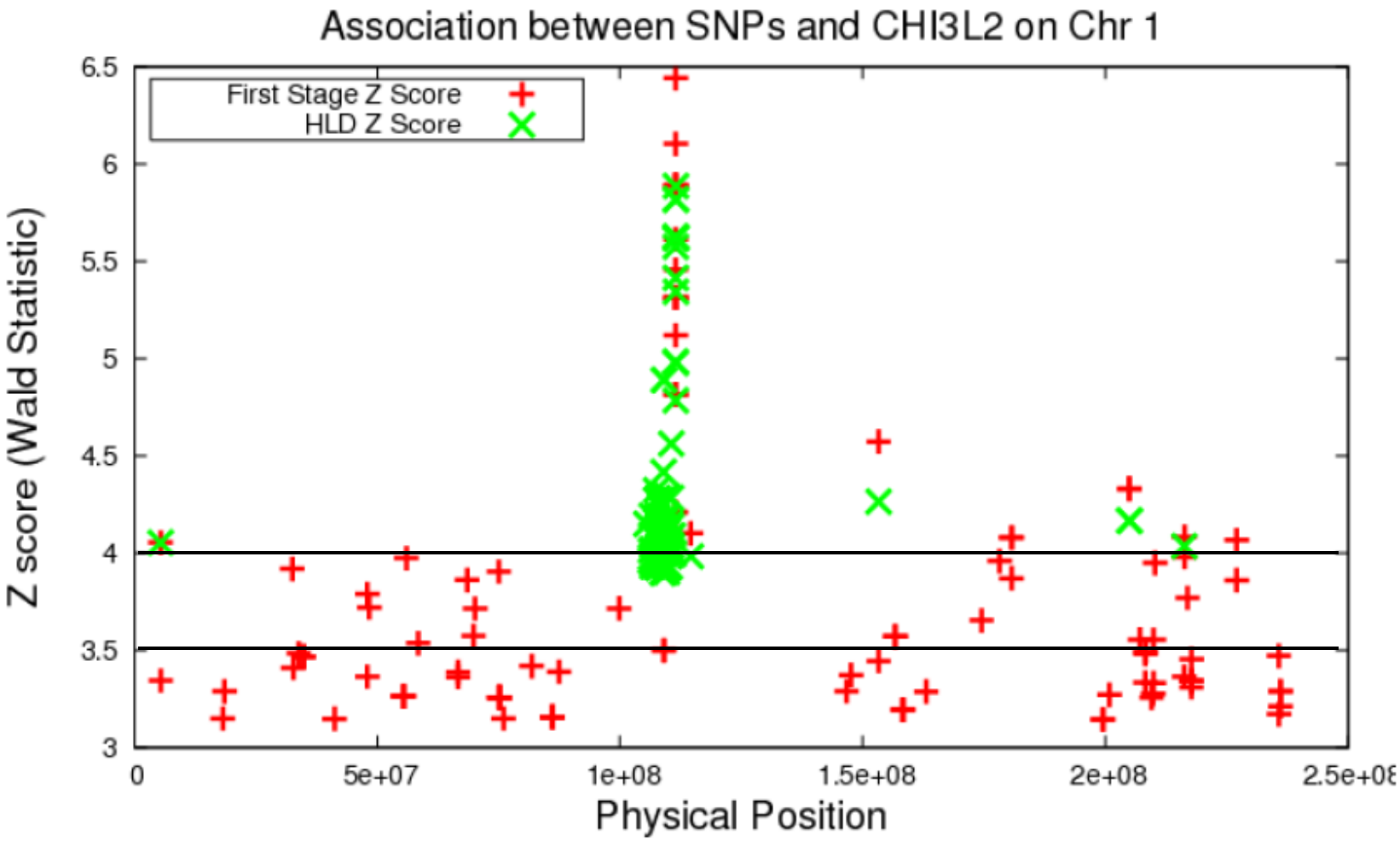
			Functional Category					LD sum columns						
SNP	**connectivity**	**conservation**	**mRNA UTR**	**ns coding**	**intron**	**locus**	**syn coding**	**mRNA UTR**	**ns coding**	**intron**	**locus**	**syn coding**	**conservation**	**linkage**
1	206	21	0	1	0	0	0	1	0	1	0	0	42	4.4
2	4	32	1	0	0	0	0	0	1	1	0	0	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

2. 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.

Multiplicative vs. Additive Scales

	High-risk genotype present		High-risk genotype absent	
Disease outcome	Exposure present	Exposure absent	Exposure present	Exposure absent
Affected	a	b	e	f
Unaffected	c	d	g	h
Disease risk estimate	$r_{11} = a/(a+c)$	$r_{01} = b/(b+d)$	$r_{10} = e/(e+g)$	$r_{00} = f/(f+h)$
Measure of exposure effect within genotype strata:				
Multiplicative scale	$RR_{G+} = r_{11}/r_{01}$		$RR_{G-} = r_{10}/r_{00}$	
Additive scale	$RD_{G+} = r_{11}-r_{01}$		$RD_{G-} = r_{10}-r_{00}$	
RR with common referent	$RR_{G+E+} = r_{11}/r_{00}$	$RR_{G+E-} = r_{01}/r_{00}$	$RR_{G-E+} = r_{10}/r_{00}$	$RR_{G-E-} = r_{00}/r_{00} =$ 1 (reference)

$RR_{G+}=RR_{G-}$ or $RR_{G+E+} = RR_{G+E-} \times RR_{G-E+}$ no interaction on multiplicative scale

$RD_{G+}=RD_{G-}$ or $RR_{G+E+} = RR_{G+E-} + RR_{G-E+} - 1$ no interaction on additive scale

Austin, 2014

Multiplicative vs Additive Interactions

Measurement scale and interaction effect	Cohort study
Multiplicative scale	
No interaction	$RR_{G+E+} = RR_{G+E-} \times RR_{G-E+}$
Synergistic interaction	$RR_{G+E+} > RR_{G+E-} \times RR_{G-E+}$
Antagonistic interaction	$RR_{G+E+} < RR_{G+E-} \times RR_{G-E+}$
Additive scale	
No interaction	$RR_{G+E+} = RR_{G+E-} + RR_{G-E+} - 1$
Synergistic interaction	$RR_{G+E+} > RR_{G+E-} + RR_{G-E+} - 1$
Antagonistic interaction	$RR_{G+E+} < RR_{G+E-} + RR_{G-E+} - 1$

[a]Formulas for the ORs are approximations based on the approximat

If $RR_{G+E-} \neq 1$ and $RR_{G-E+} \neq 1$, then no interaction on one scale implies interaction on the other!

So different conclusions depending on the scale assumed (multiplicative vs. additive). Austin, 2014

Example: 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

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

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

Vanderbroucke et al., The Lancet 1994

Case-Only GxE Studies

- Search for gene-gene or gene-environment interactions.
- OR among cases only estimates departure from multiplicative joint effect of G and E.
- How does this work?
- Assumptions?

E+

	G+	G-
Case	A_{11}	A_{10}
Control	B_{11}	B_{10}

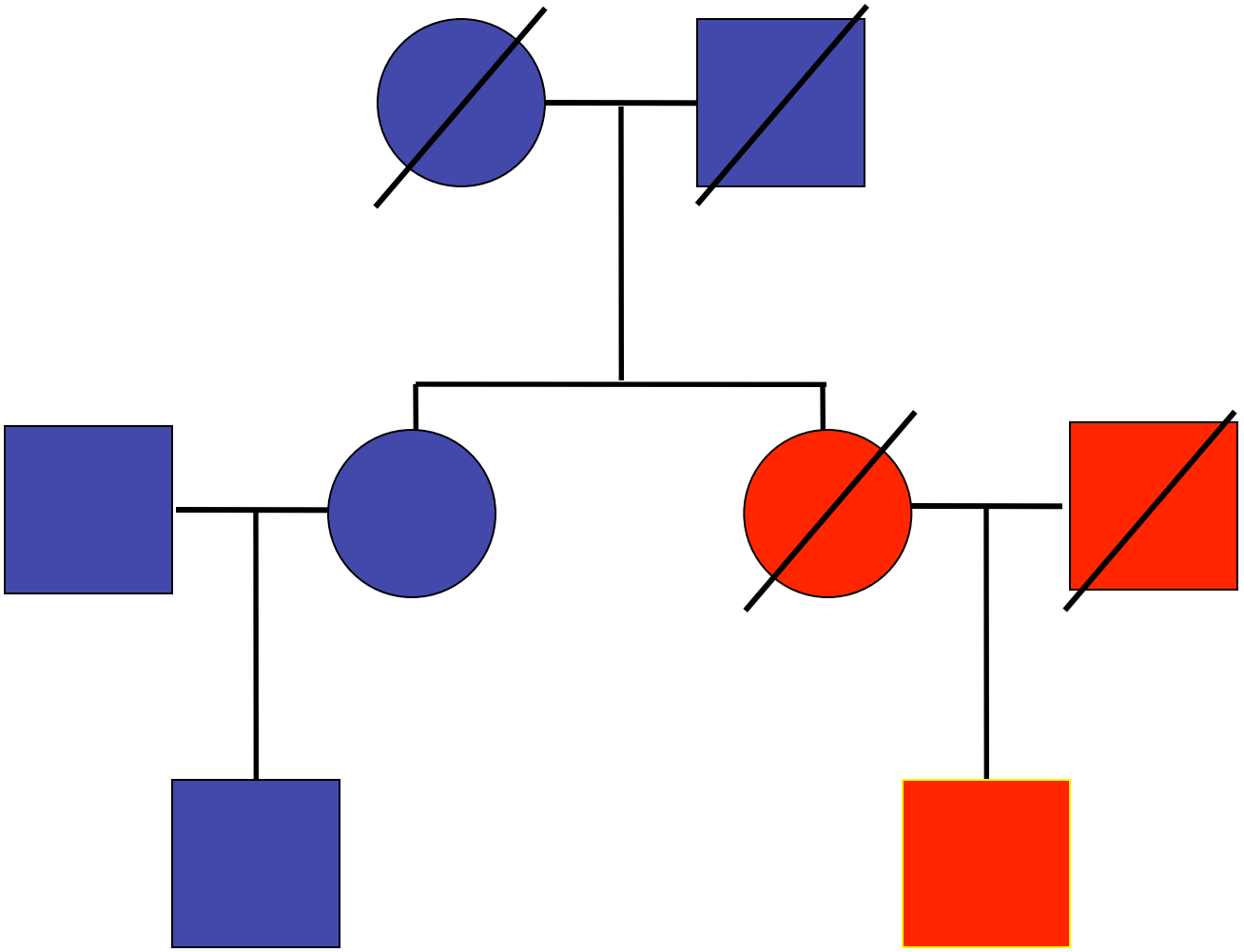
E-

G+	G-
A_{01}	A_{00}
B_{01}	B_{00}

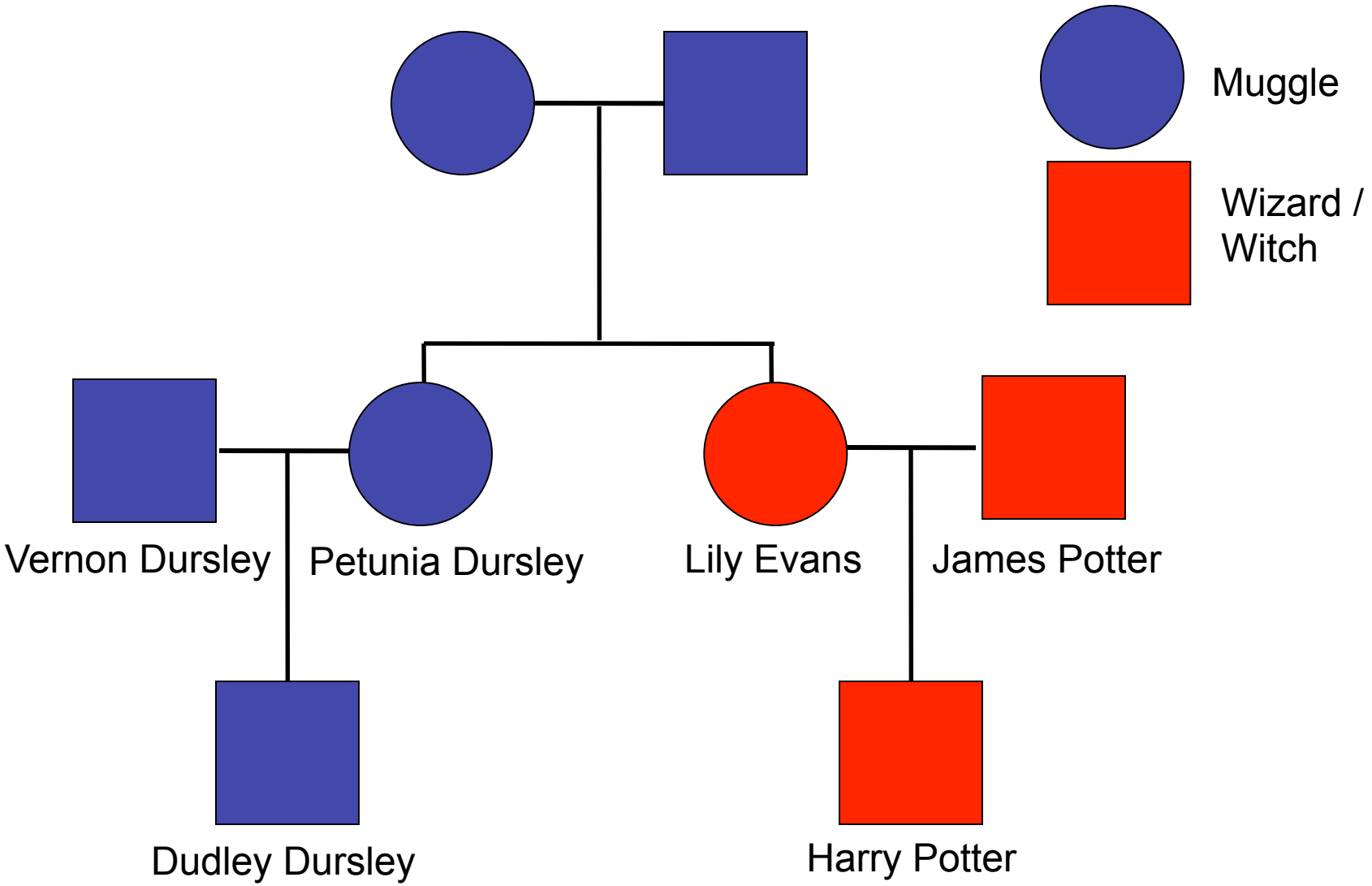
Outline

1. Hierarchical Modeling
2. Interactions
3. Measures of Variation Explained

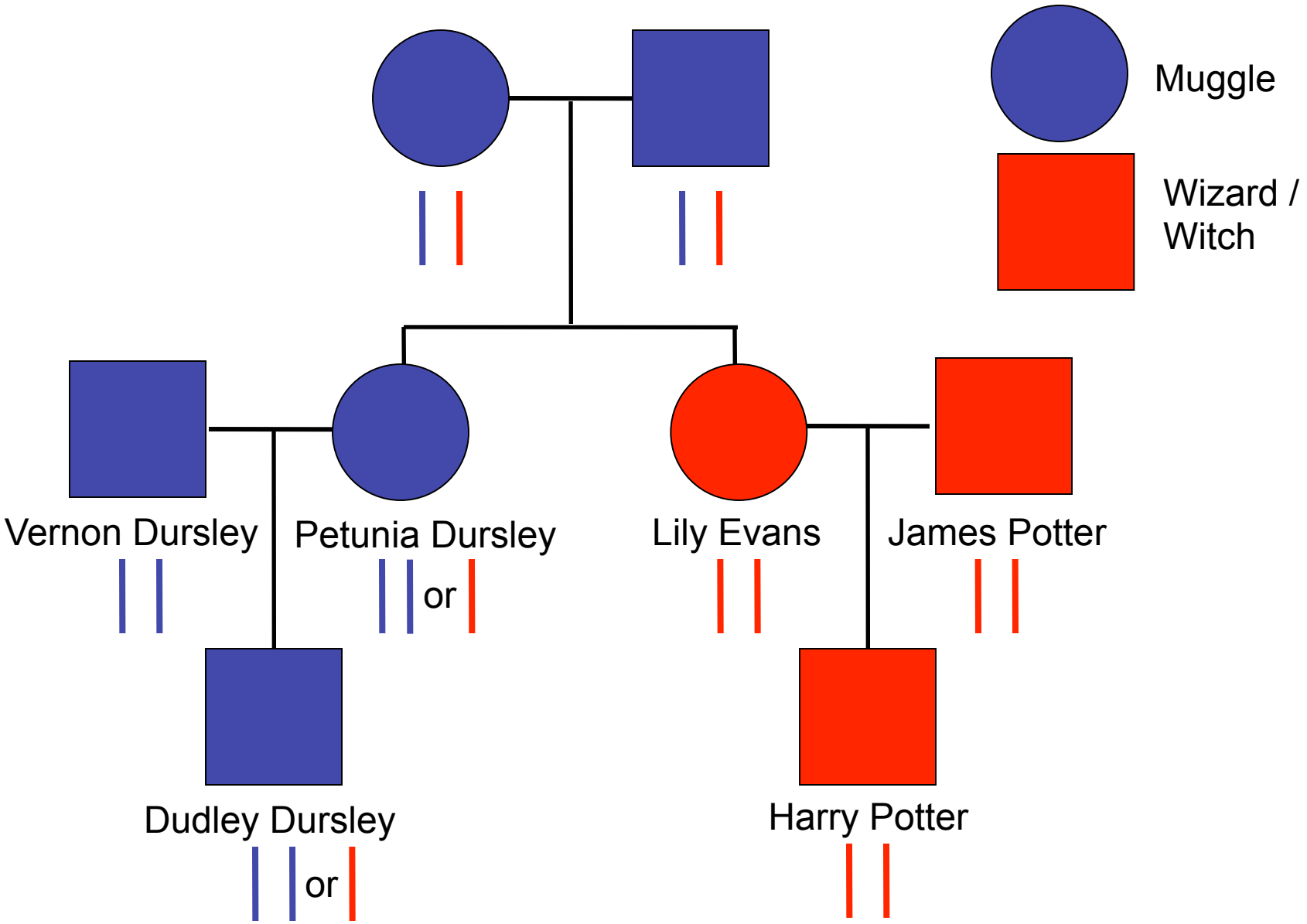
3. Measures of Variation Explained



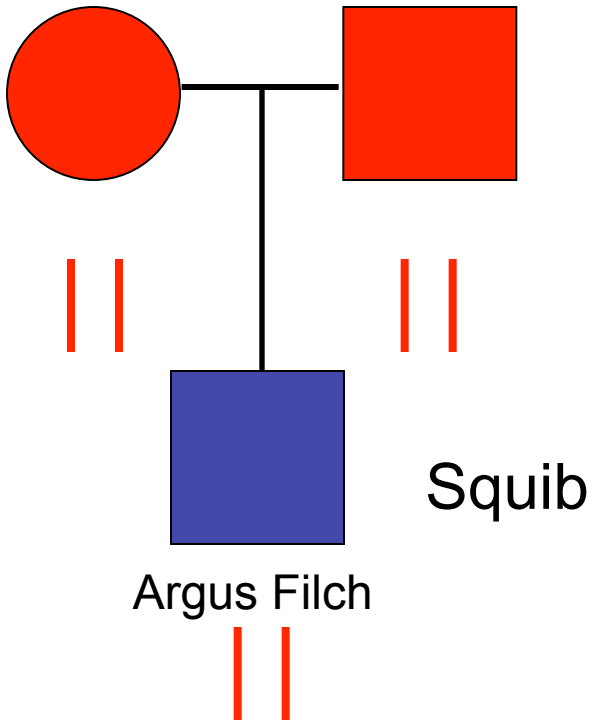
Harry Potter's Pedigree



Harry Potter's Pedigree



What About Filch?



Measures Assessing Impact of Genetic Variants on Disease

- Assume we've identified many risk variants.
- 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

Genetics *and* Epidemiology

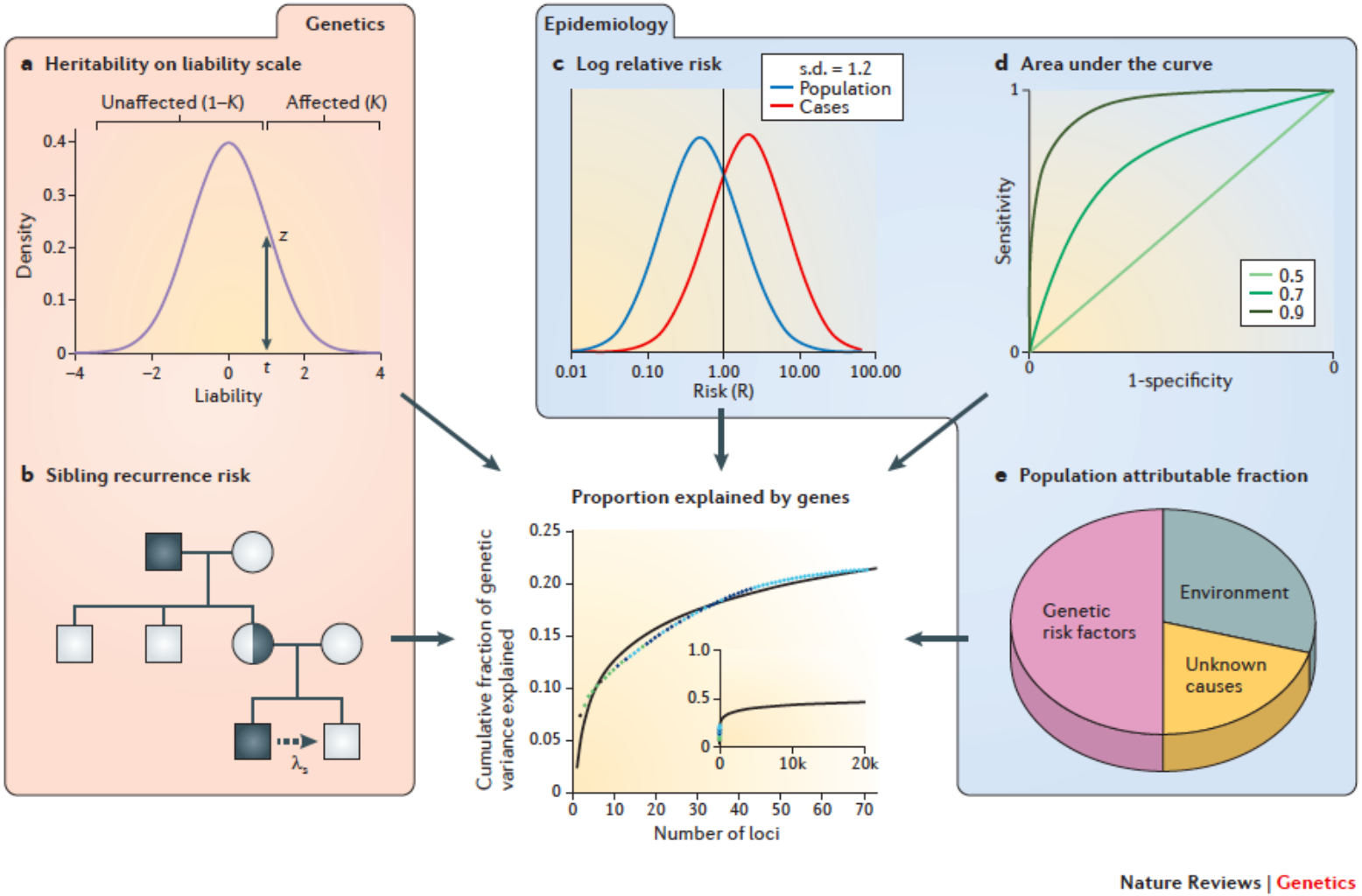
- Measures developed with different goals and within traditionally disparate fields:
  - Quantitative geneticists: interested in genetic basis underlying trait variability (liability threshold models).
  - Epidemiologists: estimate associations (log-risk models).

Genetics *and* Epidemiology

	Genetics	Epidemiology
Origins	Biological science of inheritance Experimental/breeding	Methodological sciences / public health Observational/analytic studies
Focus	Mechanisms of inheritance, often on rare diseases	Etiologic factors, primarily on common diseases
Tools	Family studies Laboratory methods Statistical models Population models	Observational studies Case-control studies Cohort designs Clinical trials
Goals	Understand mechanisms of inheritance	Understand distribution etiology, and progression of diseases
Overlap	Diseases of interest Population studies Ultimate goal of disease prevention	

Austin, 2014

Measures of Genetic Variation



Witte, Visscher, Wray; to appear.

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.

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

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

α = 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: $\sum_i h^2_{L[i]} / h^2_L$ (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.

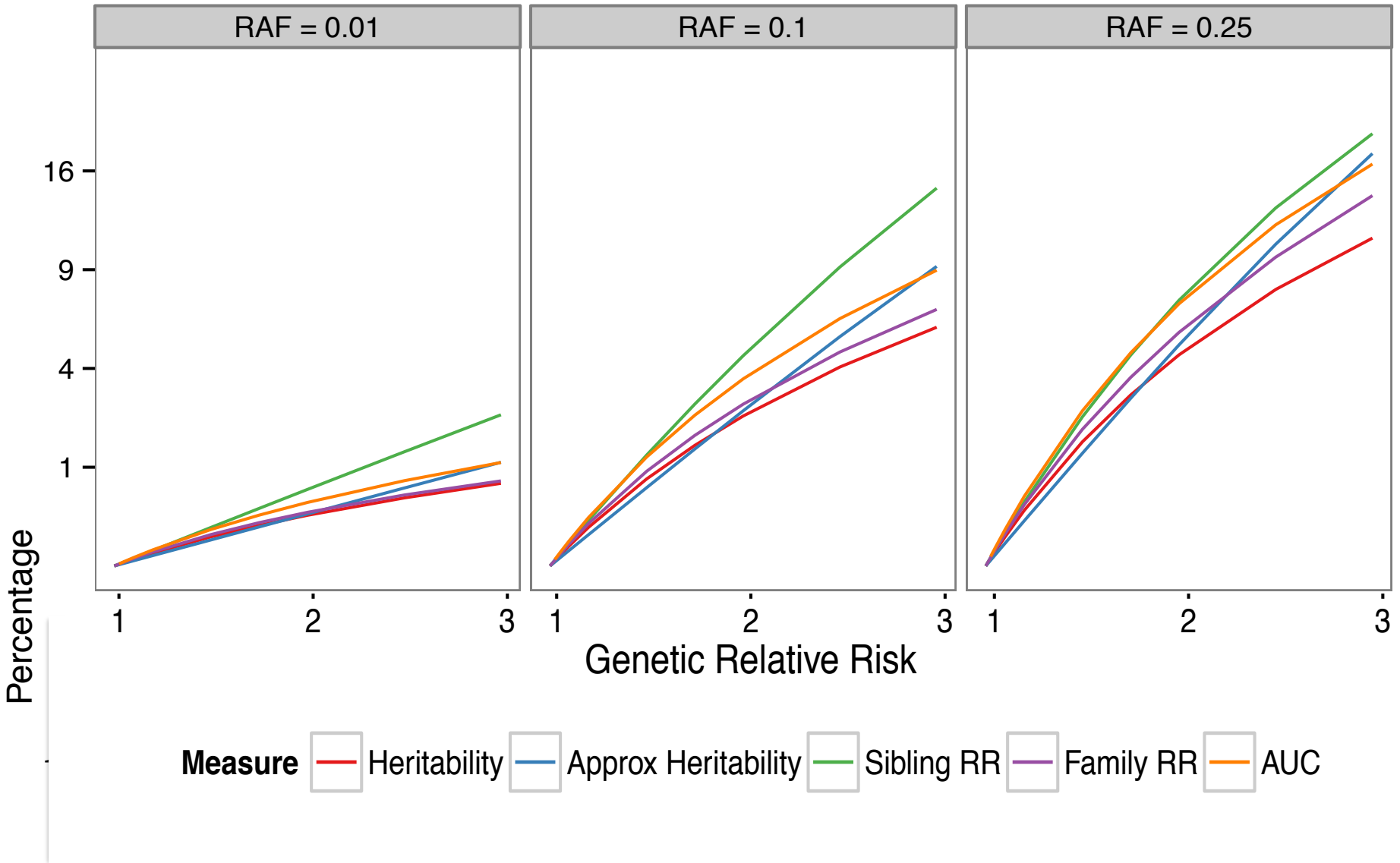
$[(AUC_{L[i]}-0.5) / (AUC_{Max}-0.5)]^2$

Wray et al., Plos Genet 2010

Comparison of Measures

- Look across a range of risk allele frequencies (RAF) and a range of genetic relative risks (RR).
- Assume popln risk (K) = 0.01, and sib RR=5.
- i.e., heritability = 55%
- Multiplicative effect of risk variant on disease (observed scale).
- Single risk variant: any differences substantially increase as # variants increase.

Results


RAF = 0.01
RAF = 0.1
RAF = 0.25
16
9
4
1
Percentage
1
2
3
Genetic Relative Risk
Measure
Heritability
Approx Heritability
Sibling RR
Family RR
AUC

Application

- Further explore how these measures can imply different impacts of genetic variants on disease.
- Calculate them across studies of:

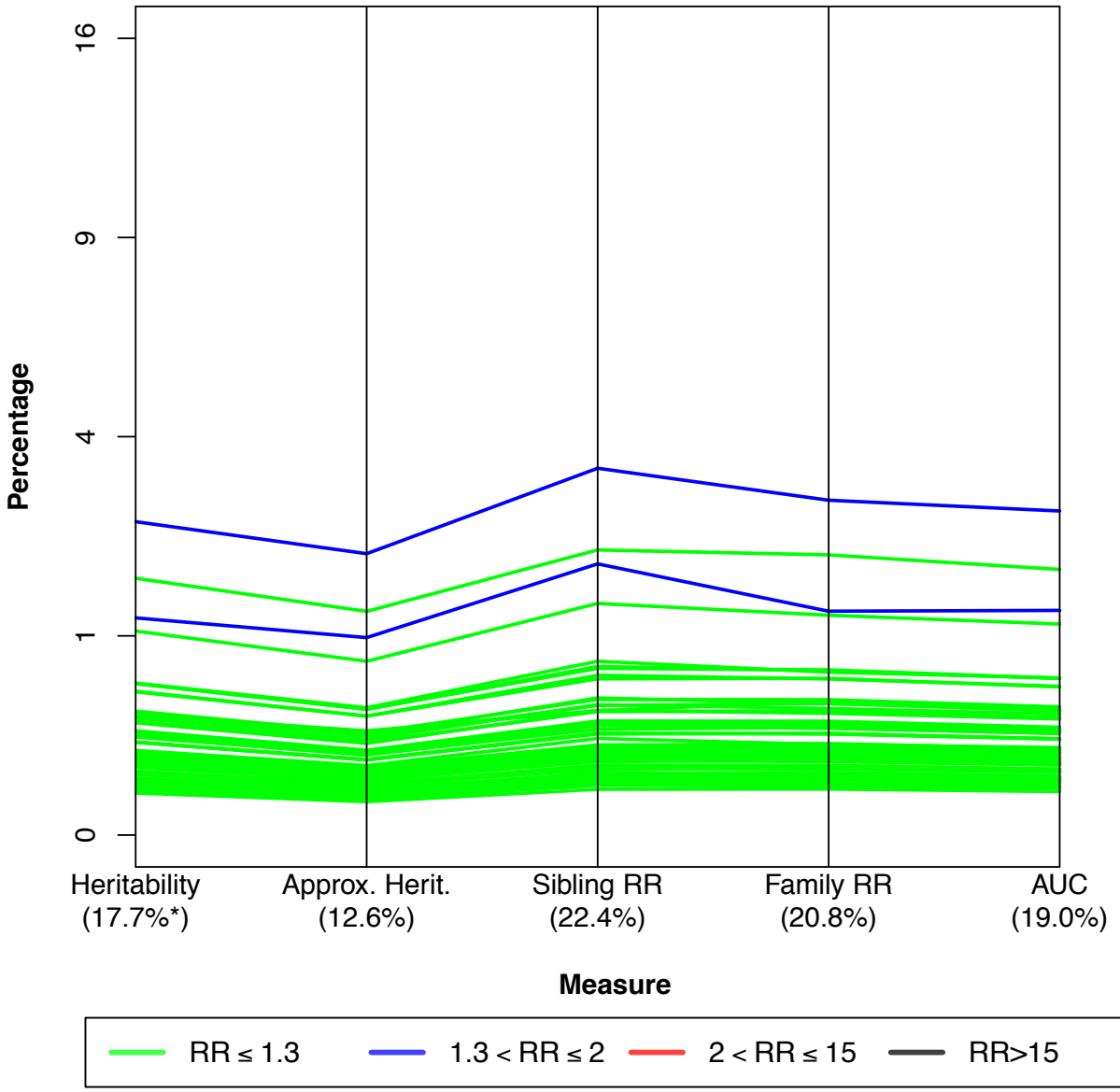
  a) Breast cancer

  b) Crohn's disease

  c) Rheumatoid arthritis

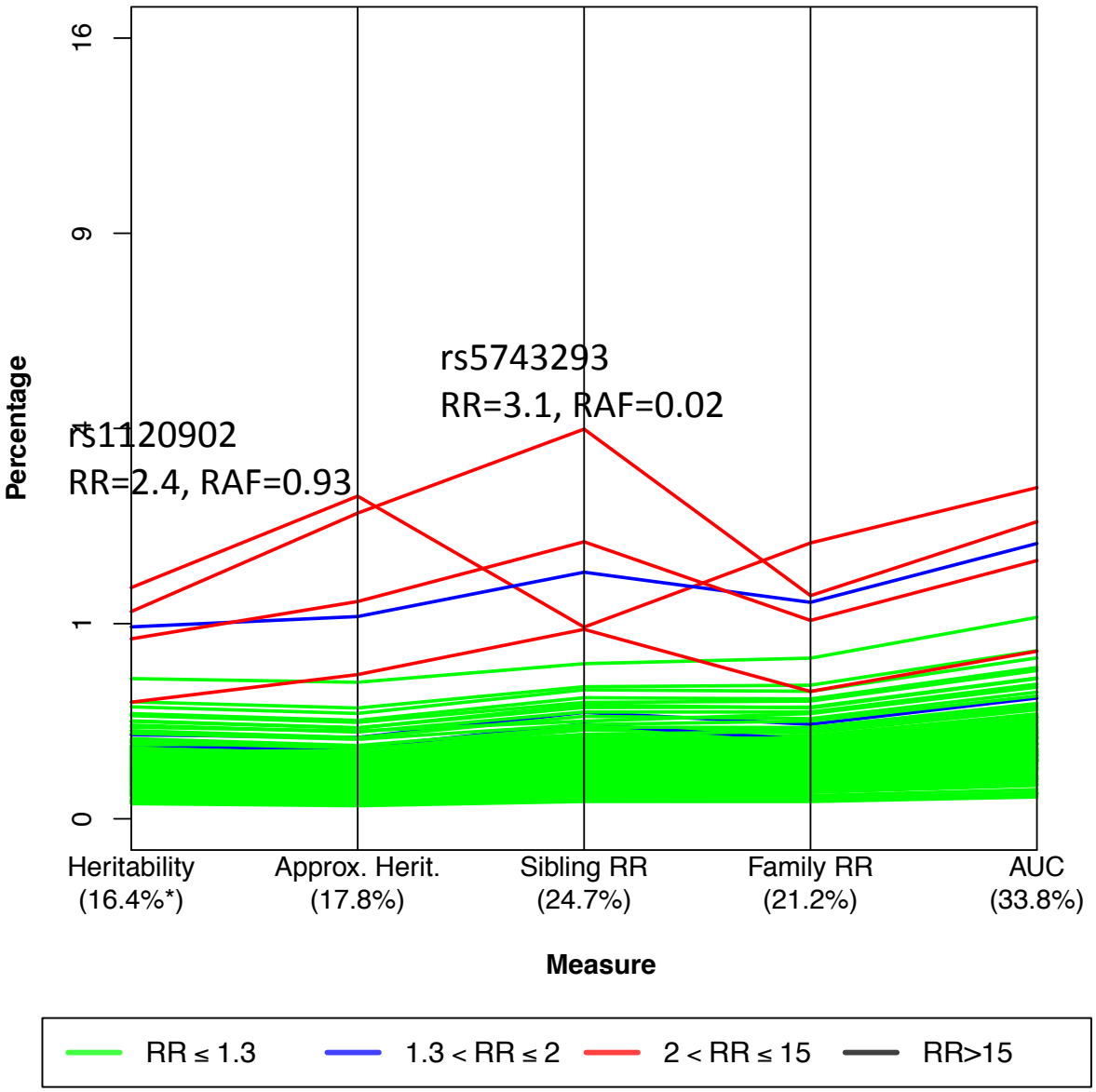
  d) Schizophrenia

Results: Breast Cancer



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

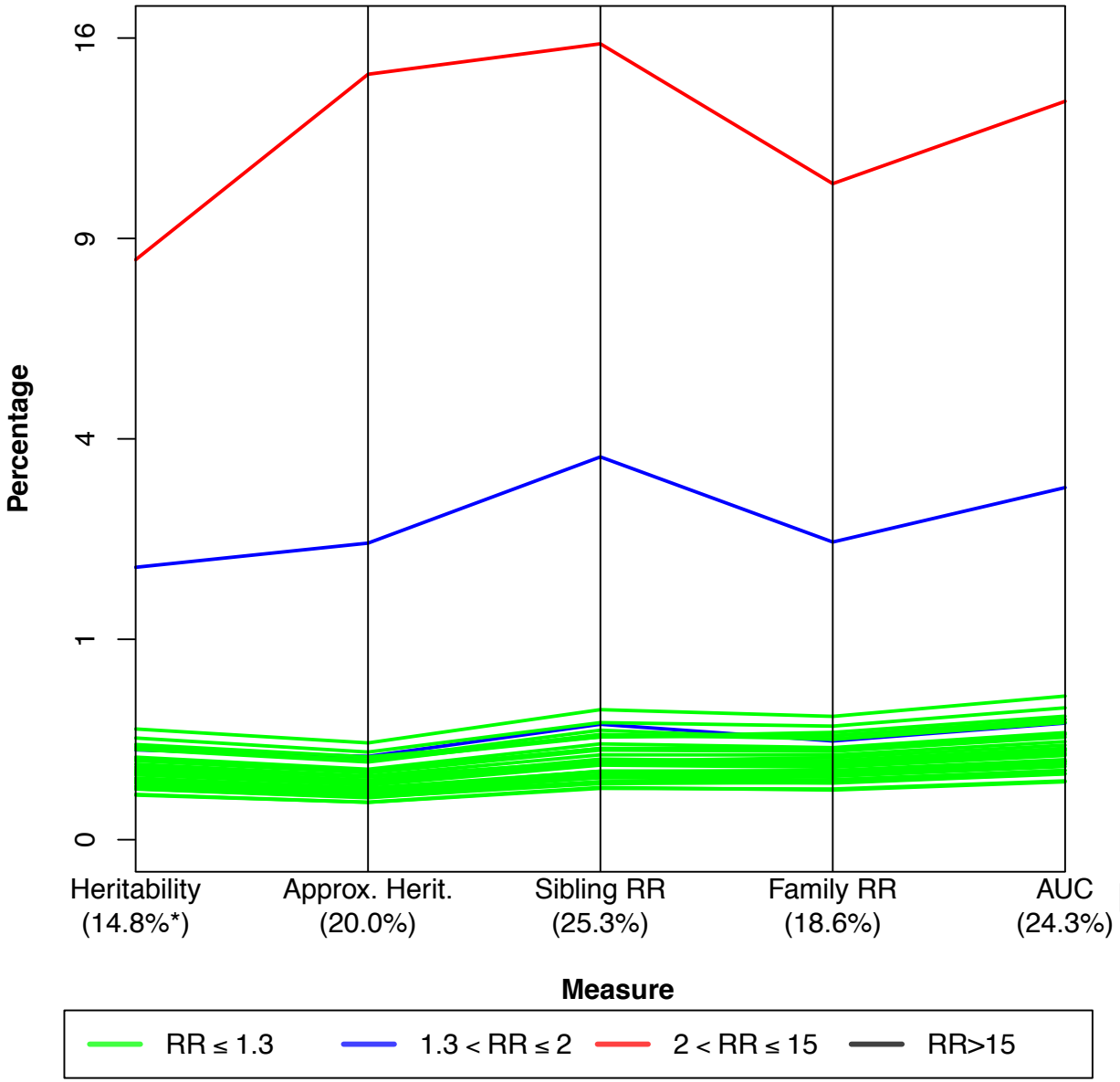
Results: Crohn's Disease



M=165
K=0.005, SRR=10.3
h^2=0.72

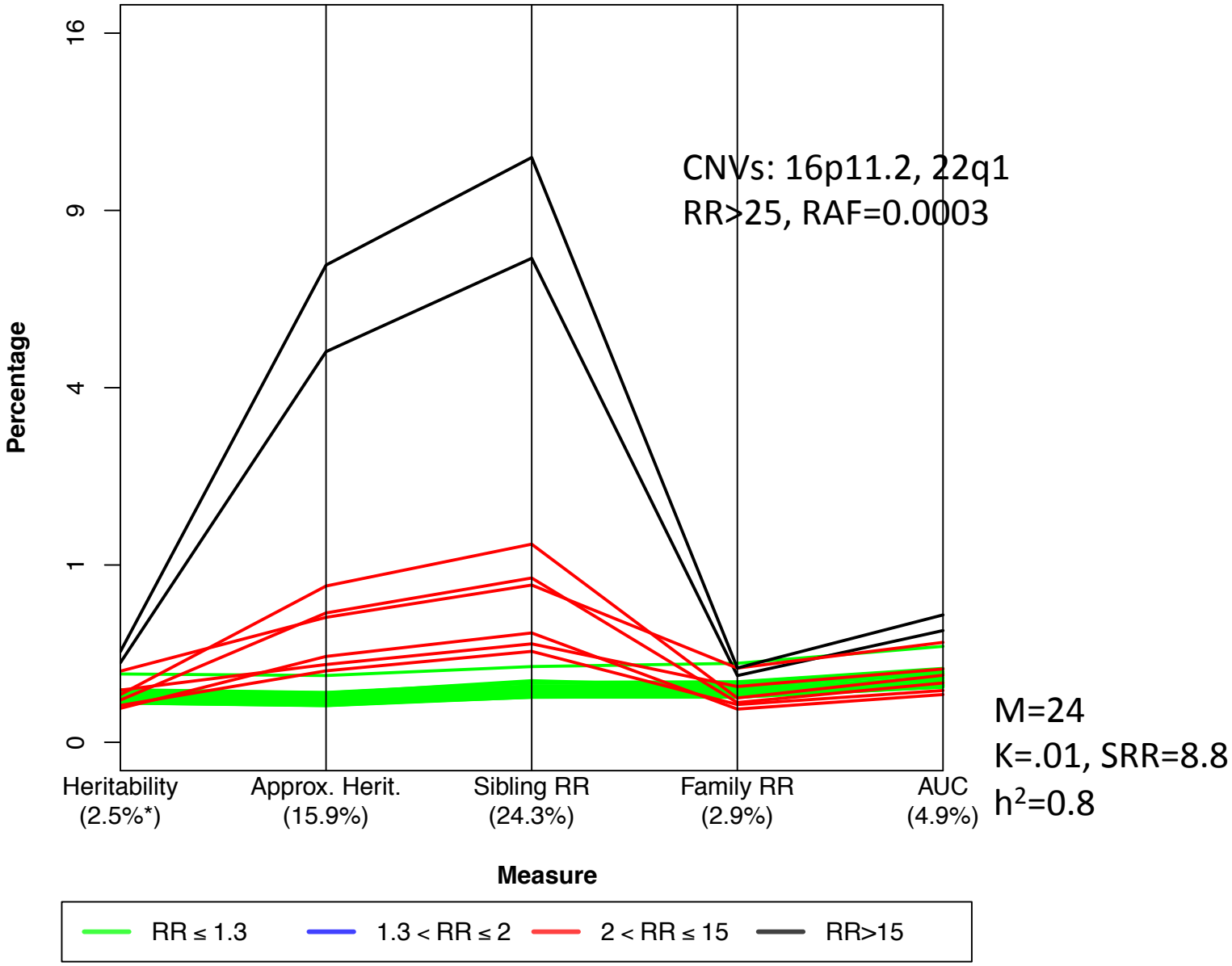
Results: RA

rs6910071, HLA-DRB1E
RR=2.88, RAF=0.22



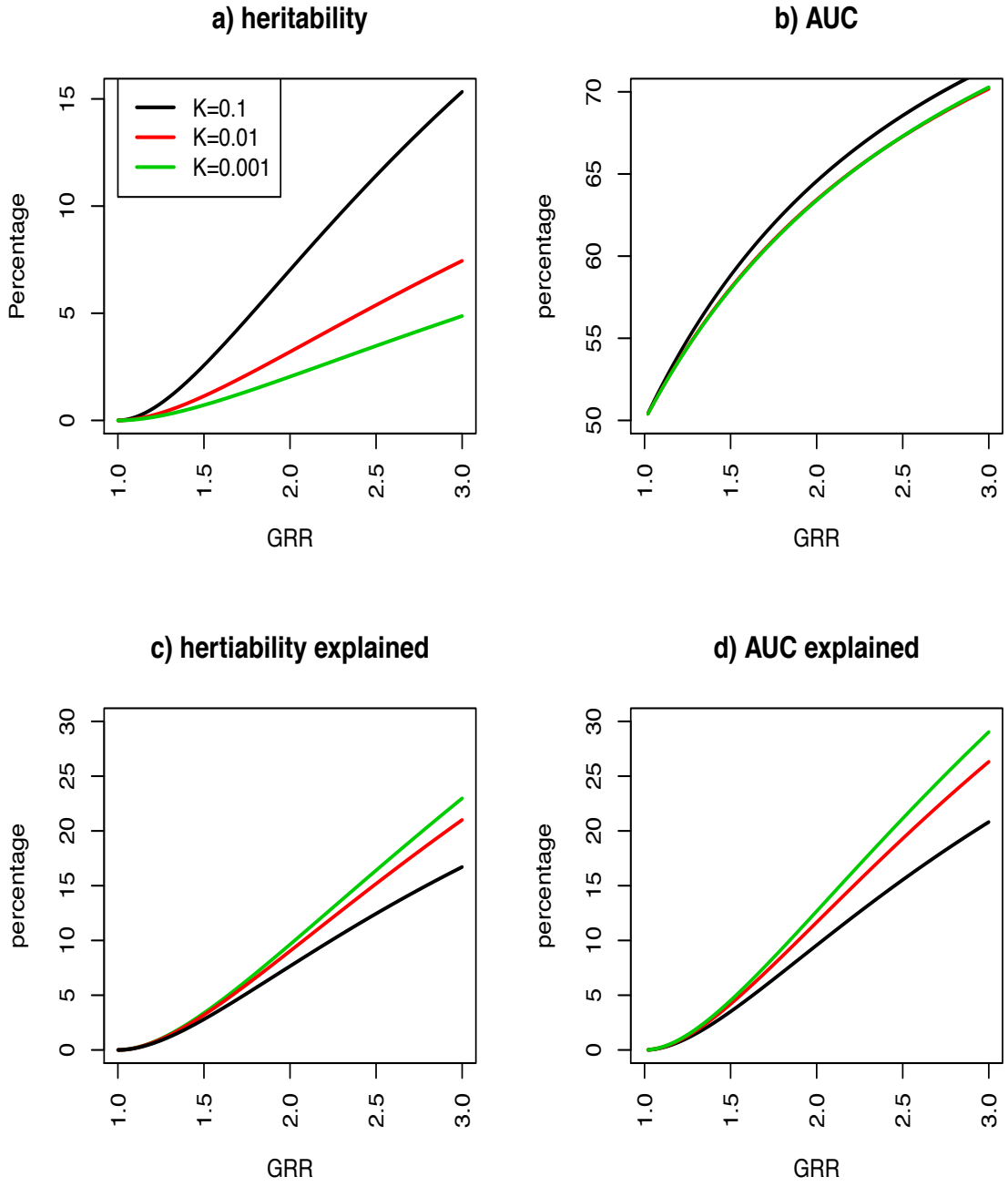
M=36
K=.01, SRR=6
h^2=0.63

Results: Schizophrenia



Impact of K

- The heritability and AUC depend on the baseline disease risk (K).
- The proportion of heritability and AUC explained is lower with increasing K.



RAF=0.5, SRR=3

h^2=0.21, 0.35, 0.92

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]

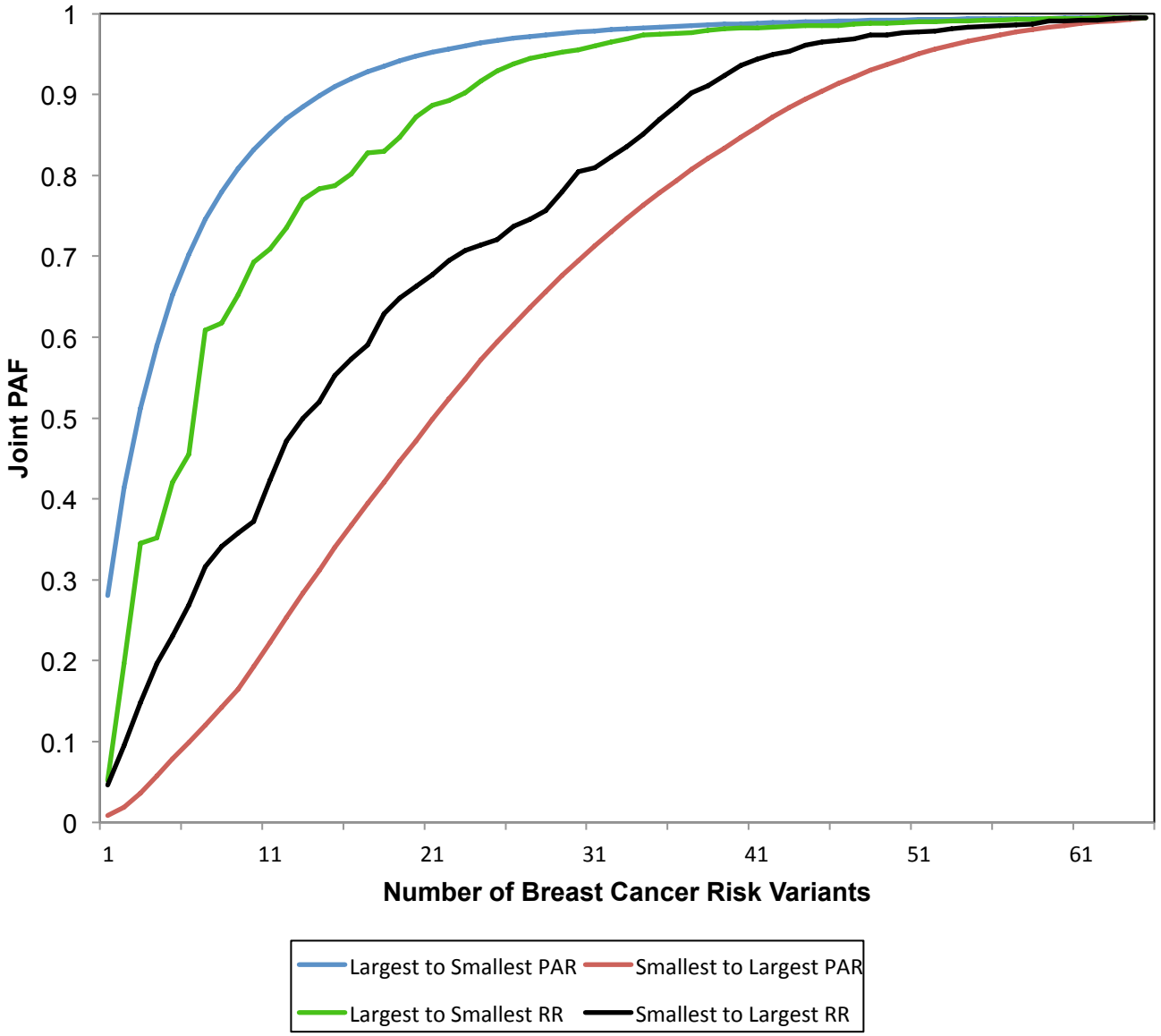
Population Attributable Fraction

- ~ 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 (e.g. a pharmacologic intervention) 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.

Module 19:
Statistical & Quantitative Genetics of Disease

Lecture 4
Polygenic models of disease risk
Naomi Wray

Aims of Lecture 4

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

Data

- Polygenic risk scoring

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:

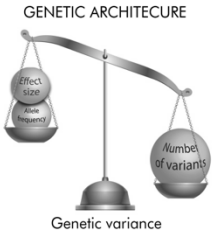
Many risk variants of small effect identified

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

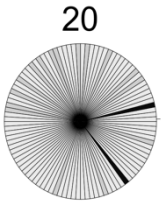
n loci 100

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

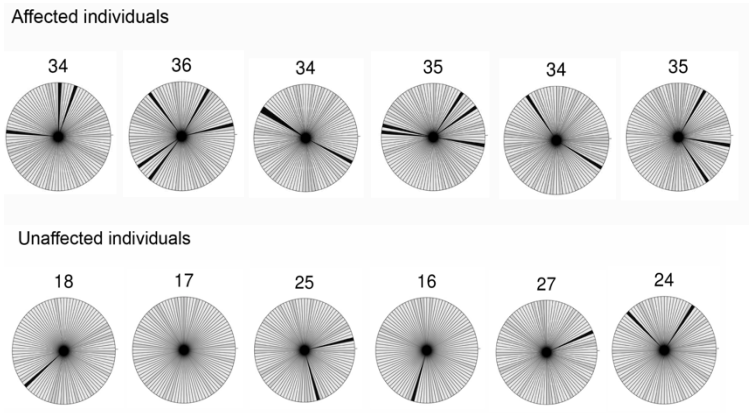
Visualising common complex genetic diseases
Polygenic genetic architecture

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



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

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

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
  - $P(D \mid x < n1) = 0$
  - $P(D \mid n1 < x < n2)$ = additive with x
  - $P(D \mid x > n2) = 1$

  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)

$P(D \mid x = s) = f_0 R^s$ Multiplicative on the risk scale

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

Constraint

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

Binomial expansion

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

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

Additive on the log risk scale

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

Looking at the multiplicative model

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 x
- Need to fudge to retain this property
- Loses mathematical tractability

Epidemiology risk model

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

Odds(Disease | x =s) = Odds(Disease | x =0) γ^x = C γ^x

γ = odds ratio for each risk locus

P(Disease | x = s)= C γ^s/(1-C γ^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

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

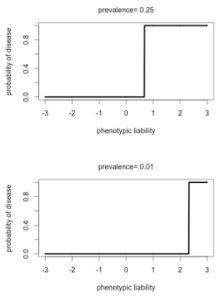
i.e.
- In terms of total variance explained which could cover a range of genetic architectures

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.

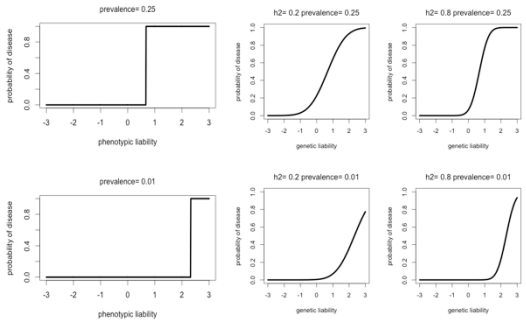
Mitchell (2012) What is complex about complex disorders Genome Biol 12: 237
Edwards(1969) Familial predisposition in man, Br Med Bull
Melnick & Shields (1976) Allelic restriction: a biologic alternative to multifactorial threshold model. The Lancet
Many references to the criticism in papers of the time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

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. "

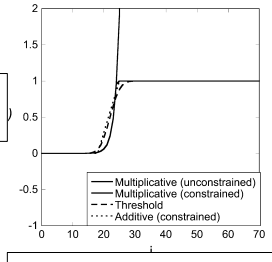
Need risk to disease from relatives of different types of relatives to start to distinguish between models
Not easy to collect, large sampling variances

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.
$P(Disease)=P(Disease|x=0)R^x$
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"

Genetic counseling – can make predictions about difficult to measure scenarios

TABLE 1

RECURRENCE RISKS (× 1,000) IN SIBSHIPS

POPULATION FREQUENCY (%)	HERITABILITY (%)	No. NORMAL SIBS	No. AFFECTED PARENTS: 0 — No. Affected Sibs: 0	1	2	No. AFFECTED PARENTS: 1 — No. Affected Sibs: 0	1	2	No. AFFECTED PARENTS: 2 — No. Affected Sibs: 0	1	2
10.0	100*	0	055*	165	244	283	409	493	731	751	769
		1	049	149	230	234	351	433	678	695	712
		2	043	136	210	198	306	387	641	657	672
	80	0	064	165	252	235	347	432	557	607	649
		1	057	148	228	200	302	382	493	543	585
		2	052	134	208	175	267	342	444	493	535
	50	0	080	151	220	178	257	326	335	401	457
		1	074	139	203	160	233	297	301	363	417
		2	068	129	188	147	213	273	275	332	383
	20	0	093	123	154	129	161	193	174	207	240
		1	089	119	148	124	155	186	167	199	230
		2	087	115	143	120	149	179	160	191	221
1.0	100*	0	007	073	144	112	240	338	633	649	666
		1	007	067	135	096	209	301	606	617	630
		2	006	062	127	084	185	271	588	597	606
	80	0	008	065	142	083	185	278	409	466	516
		1	008	060	130	074	164	248	370	423	470
		2	008	055	120	067	148	224	338	389	434
	50	0	009	039	084	043	093	151	146	206	263
		1	009	037	079	040	087	141	135	191	245
		2	009	035	075	038	082	132	127	178	229
0.1	100*	0	001	038	108	049	156	257	620	629	640
		1	001	035	100	044	137	230	604	612	620
		2	001	033	093	040	123	207	592	599	605
	80	0	001	025	082	029	098	179	317	374	424
		1	001	024	076	027	089	163	290	345	392
		2	001	022	071	025	082	149	268	320	366
	50	0	001	010	032	010	034	069	066	109	153
		1	001	010	031	010	032	066	063	104	145
		2	001	009	030	010	031	063	060	099	139

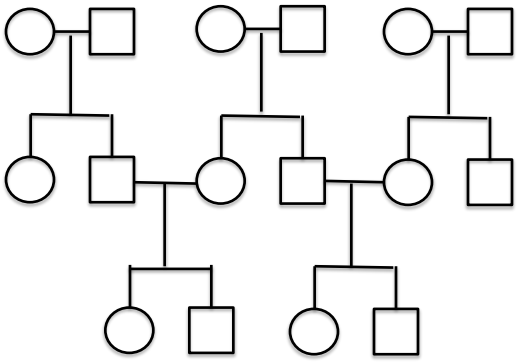
* Evaluated for $h^2 = 99\%$.

Smith (1971) Recurrence risks for multifactorial inheritance Ann Hum Genet 34: 85

Under a totally polygenic genetic architecture

- How often do cases show as familial ?
- How often do cases show as sporadic ?

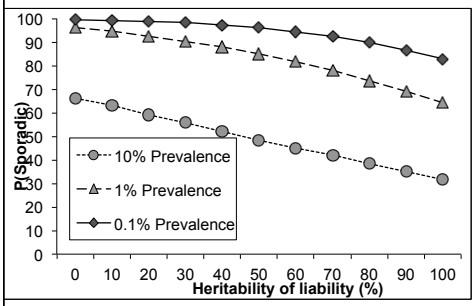
Simulate families under a polygenic model



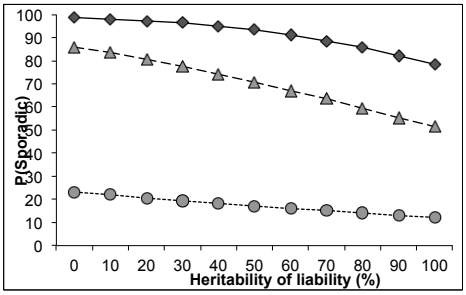
Two key parameters:
Proportion of population affected
Proportion of variance of disease attributable to genetic factors = heritability

Simulation
- Idealised
- True disease status of all family member is known

Probability of having no 1st degree relative affected



Probability of having no 1st, 2nd or 3rd degree relative affected

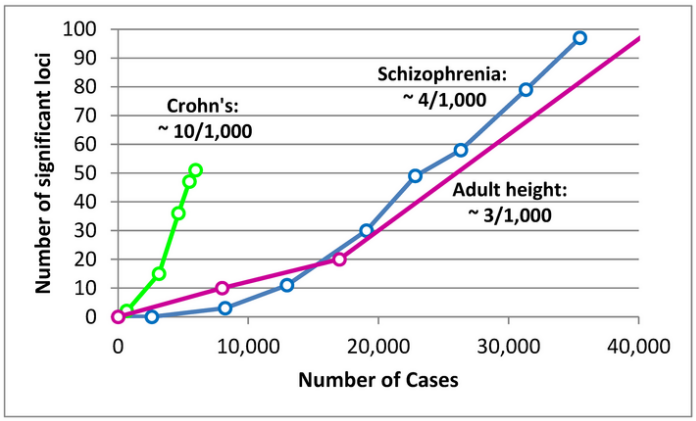


Under a polygenic model most cases appear sporadic

Yang et al (2009) Sporadic cases are the norm for complex disease. European J Human Genetics

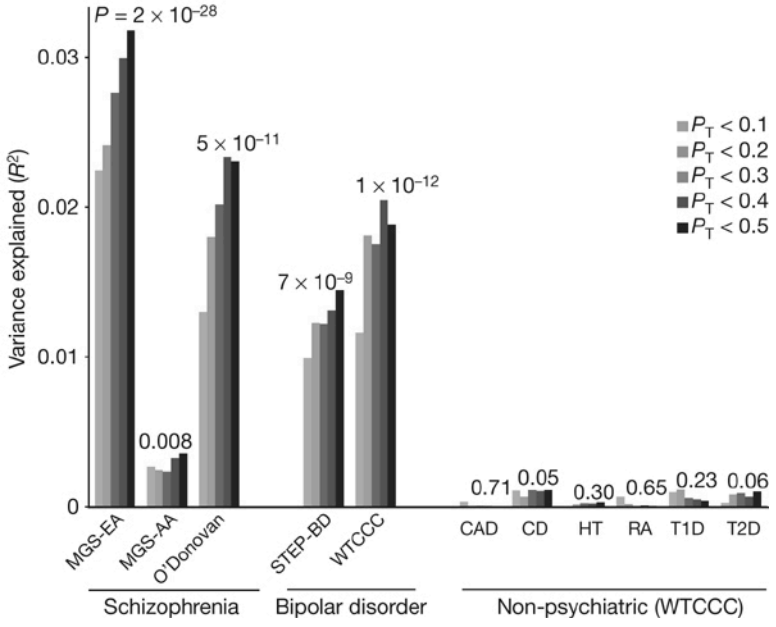
Polygenic risk profile

Evidence for a polygenic contribution to disease



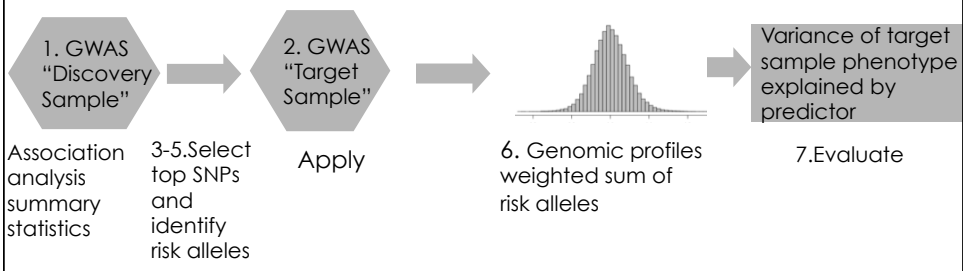
Levinson et al (2014) Genetic studies of major depressive disorder. Why are there no GWAS findings and what can we do about it? Biological Psychiatry (submitted)

Risk Profile Scoring



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

Polygenic Risk Profile Scoring



Wray et al (2014) Polygenic methods and their application to psychiatric traits. Journal of Child Psychology & Psychiatry (in press)

Steps 1 - 3 in polygenic risk scoring

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

2. 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.

3. 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 4-6 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

What would happen if we did step 4 before step 3

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

- $\ln(OR_i)$ = 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.

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

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

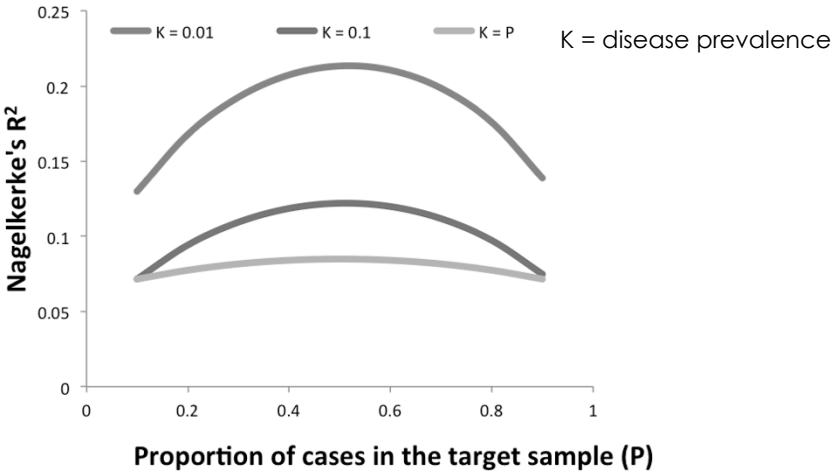
   Full model: y ~ covariates + score Logistic, y= case/control = 1/0
   Reduced model: y ~ covariates
   N: sample size

   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

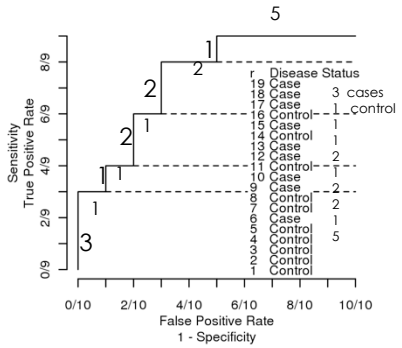
Statistics to evaluate polygenic risk scoring 2.

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

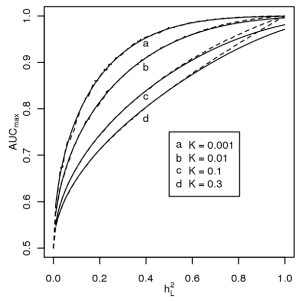
- Rank individuals on score
- Start at origin on graph
- Work through list of ranked individuals
- Move one unit along y-axis if next individual is a case
- Move one unit along x-axis is next individual is a control6



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

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. 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 3 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

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

   Convert to the liability scale

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

   Where C is:

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

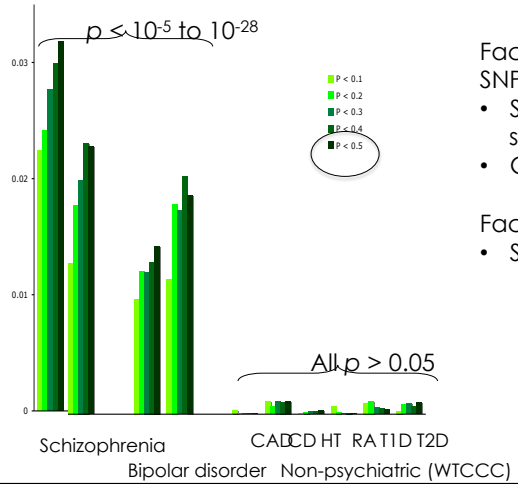
   h^2_{GRPS} is on the same scale as heritability estimated from family studies and GREML SNP-chip heritability

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

Practical

- Risk profile scoring

Factors affecting the proportion of SNPs that maximise efficacy



Factors affecting proportion of SNPs that give maximum R^2

- Sample size of discovery sample
- Genetic architecture

Factors that affect p-value of R^2

- Sample size of target sample

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

Single GWAS- how to split into discovery and target?

Split based on independently collected samples

What is the optimum split?

The larger the discovery sample the better effect sizes are estimated

The larger the target sample the R^2 is estimated with more precision

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

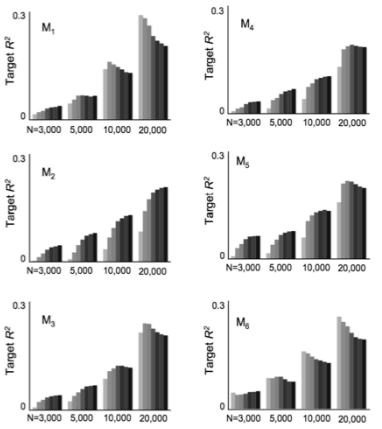
But with large samples power achieves 1, so value of increasing target sample is redundant.

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 GRPS 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 (in press)

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

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

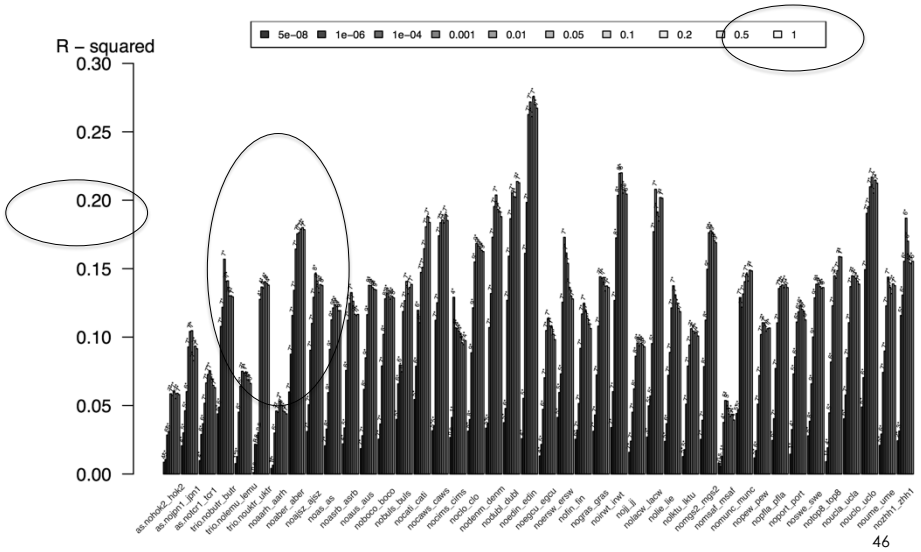


M1-M7 vary in

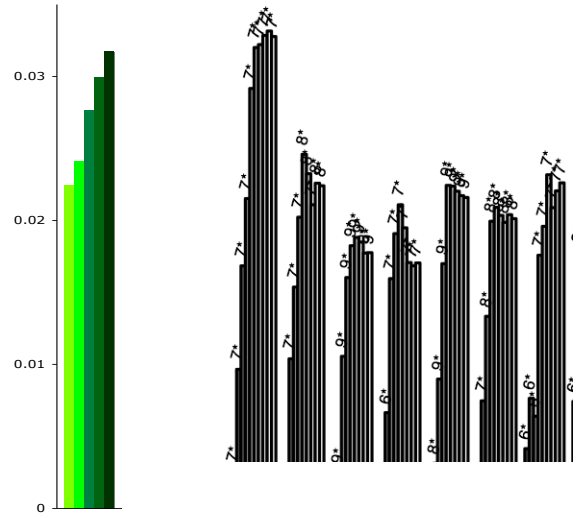
- proportion of SNPs associated in disease
- distribution of effect sizes
- Frequency distribution
- LD between SNPs and causal variants

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

PGC-SCZ Wave 2- leave one sampling out profile scoring



Impact of sample size on p-value threshold that maximises R^2



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

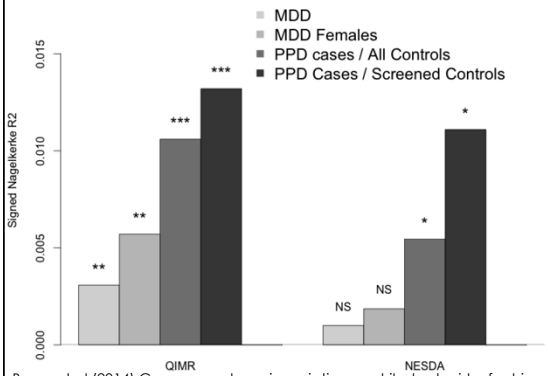
B. Different disorders - demonstrates genetic overlap between disorders

C. 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

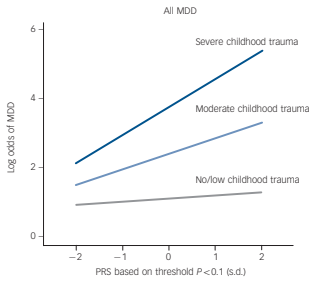
Discovery & Target samples could be:

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

C. 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

Example: GxE
Major Depressive Disorder and childhood trauma



Discovery sample: PGC-MDD ex target sample
Interpretation depends on proportion of those with childhood trauma in discovery sample = unknown

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

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

C. 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

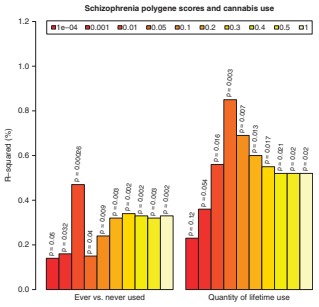


Figure 1. Results of polygenic risk scores for schizophrenia predicting variance explained (R^2) in cannabis use as both a binary trait of ever versus never, and as a quantitative trait of lifetime use within only users. Polygenic scores were created using different cutoffs for the inclusion of risk variants for schizophrenia, ranging from $P = 0.0001$ to 1.0.

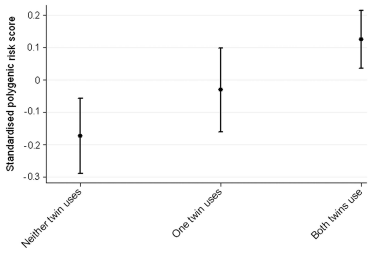


Figure 2. Mean standardized polygenic risk scores for pairs of twins when neither ($n = 272$), one ($n = 273$) or both twins ($n = 445$) had reported use of cannabis. An ordinal regression reported a significant association ($P = 0.001$).

Relationship between GPRS and GREML

- Estimates of SNP-heritability using GREML e.g. software GCTA see SISG module 23
- Purcell et al (2009) used simulation to determine what proportion of variance in liability to SCZ attributed to SNPs was consistent with observed R2
- Simulated many genetic architectures in terms on number of risk loci, effect size distribution, risk allele frequency distribution.
- What GPRS R^2 was observed for the real discovery and target samples
- Most architectures considered were not consistent with the observed results
- Many architectures were consistent with observed results
- For all consistent architectures the total proportion of variance explained by SNPs was 0.34

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

Relationship between GPRS and GREML

International Schizophrenia Consortium (ISC) Discovery sample:
3322 cases, 3587 controls
Molecular Genetics of Schizophrenia (MGS) Target sample:
2687 cases, 2656 controls
Nagelkerke's R^2 of 0.032 p-value of $2x10^{-28}$
Based on SNPs with p-value threshold 0.5 out of 74062 LD-pruned SNPs.

All simulation genetic architectures that were consistent with the empirical results pointed to a h^{2-} *SNP* of 0.34.

Application of GREML to the ISC data generated a direct estimate of h^{2-} *SNP* = 0.33 (95% CI 0.24-0.42) Supplementary Table 2 in Lee et al (2012).

Reducing to h^{2-} *SNP* = 0.27 (95% CI 0.21-0.33) after stringent QC, designed to reduce the chances that the reported estimate is inflated by artefacts such as population stratification.

These results demonstrate the relationship between GPRS and GREML via simulation.

Lee et al (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by SNPs. Nature Genetics

Relationship between GPRS and GREML

Dudbridge (2013) uses theory – uses liability threshold model to estimate the total variance attributable to SNPs that would generate the observed R^2

- Input
  - Discovery sample size and proportion of cases
  - Target sample size and proportion of cases
  - Disease prevalence
  - P-value of R^2 estimated from GPRS in target sample
  - Total number of SNPs
  - Proportion of SNPs in predictor

Relationship between GPRS and GREML

International Schizophrenia Consortium (ISC) Discovery sample:
3322 cases, 3587 controls
Molecular Genetics of Schizophrenia (MGS) Target sample:
2687 cases, 2656 controls
Nagelkerke's R^2 of 0.032 p-value of $2x10^{-28}$
Based on SNPs with p-value threshold 0.5 out of 74062 LD-pruned SNPs.
Prevalence of schizophrenia: 1%

Dudbridge R code Polygenescore
https://sites.google.com/site/fdudbridge/software/

"estimateVg2FromP" (p=2e-28, n1=3322+3587, nsnp=74062, n2=2687+2656, vg1=0, corr=1, plower=0, pupper=0.5, weighted=T, binary=T, prevalence1=0.01, prevalence2=.01, sampling1=3322/(3322+3587), sampling2=2687/(2687+2656), lambdaS1=NA, lambdaS2=NA, nullfraction=0, shrinkage=F, logrisk=F) generates an estimate of h^2- *SNP* = 0.287 (95% CI 0.236 - 0.337).

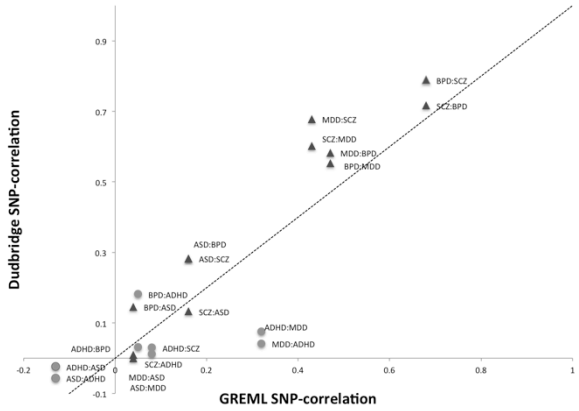
Relationship between GPRS and GREML

International Schizophrenia Consortium (ISC) Discovery sample:
3322 cases, 3587 controls
Molecular Genetics of Schizophrenia (MGS) Target sample:
2687 cases, 2656 controls
Nagelkerke's R^2 of 0.032 p-value of $2x10^{-28}$
Based on SNPs with p-value threshold 0.5 out of 74062 LD-pruned SNPs.
Prevalence of schizophrenia: 1%

Dudbridge R code Polygenescore
https://sites.google.com/site/fdudbridge/software/

"estimateVg2FromP" (p=2e-28, n1=3322+3587, nsnp=74062, n2=2687+2656, vg1=0, corr=1, plower=0, pupper=0.5, weighted=T, binary=T, prevalence1=0.01, prevalence2=.01, sampling1=3322/(3322+3587), sampling2=2687/(2687+2656), lambdaS1=NA, lambdaS2=NA, nullfraction=0, shrinkage=F, logrisk=F) generates an estimate of h^2- *SNP* = 0.287 (95% CI 0.236 - 0.337).

Relationship between GPRS and GREML



The Dudbridge correlation estimates are calculated using univariate GREML estimates of SNP heritability

GPRS results: Cross-Disorder Group of the PGC (2013) Lancet

GREML SNP-correlation: Cross-Disorder Group of the PGC (2013) Nature Genetics (more ADHD data)

Wray et al (2014) Polygenic methods and their application to psychiatric traits. Journal of Child Psychology & Psychiatry (in press)

Use of GPRS in experimental design

Calculate GPRS in your sample

Select high and low GPRS scorers for high cost phenotype study

Take home messages

Polygenic models of disease risk can be parameterised in different ways but they all have to have similar features to be consistent with observed risks to relatives.

Since the polygenic models are "exchangeable" the liability threshold model is the model of choice because it depends on the fewest parameters and so is the most general – and it is easy to work with.

Emerging evidence from GWAS is that common complex diseases are polygenic.

Polygenic risk prediction has utility even without understanding underlying biology

Module 19:
Statistical & Quantitative Genetics of Disease

Lecture 5
Power of case-control association studies
Naomi Wray

Aims of Lecture 5

- To understand power calculations of case-control association analysis.

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

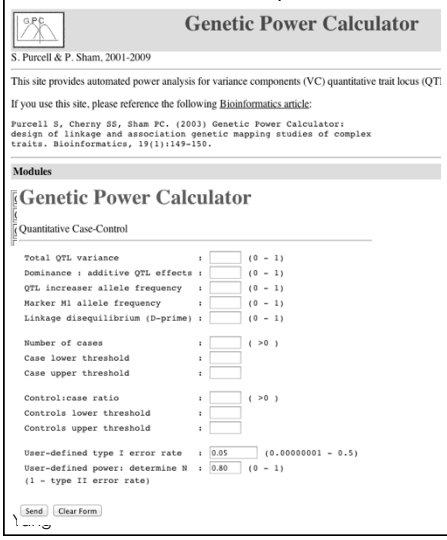
= 1- probability(Type II error)

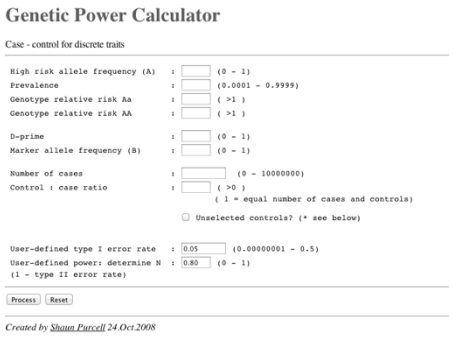
Power depends on statistical test, effect size to be detected, sample size, acceptable level of Type I error
Non-centrality parameter depends on statistical test, effect size to be detected, sample size

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

First step:

How to calculate power in an association study?





Power of association test – case/control

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

	P(G)
aa	$(1-p)^2$
Aa	$2p(1-p)$
AA	p^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	$(1-p)^2$	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

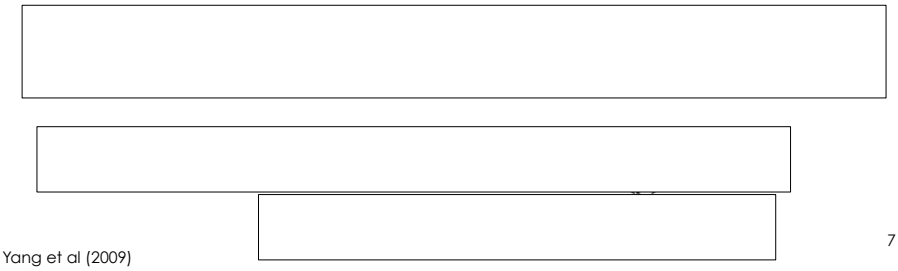
p_{case} = frequency of risk allele in cases

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) $p_{control}$



Yang et al (2009)

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

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

Yang

Power of a case-control study

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

α = *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, α =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)

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

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

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

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

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

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 $= NCP_{QT} = \frac{N_{QT} h^2_{L[i]}}{1 - h^2_{L[i]}}$

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

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

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

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

Practical

- Compare power for a common disease vs a rare disease

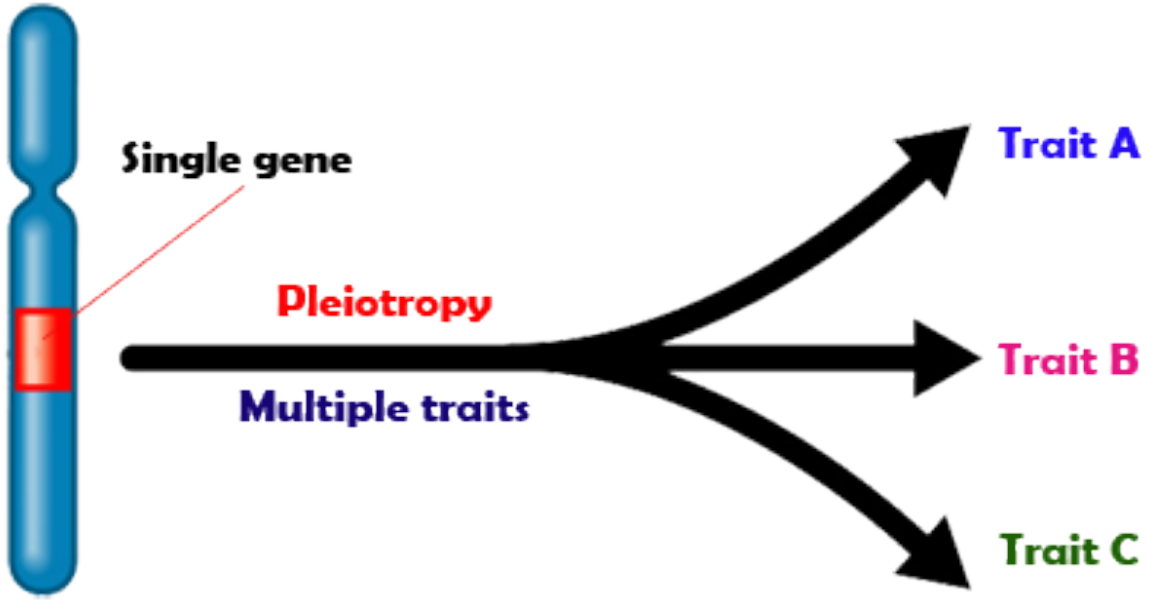
Module 19: Statistical and Quantitative Genetics of Disease

John Witte

Lecture 6

Pleiotropy

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



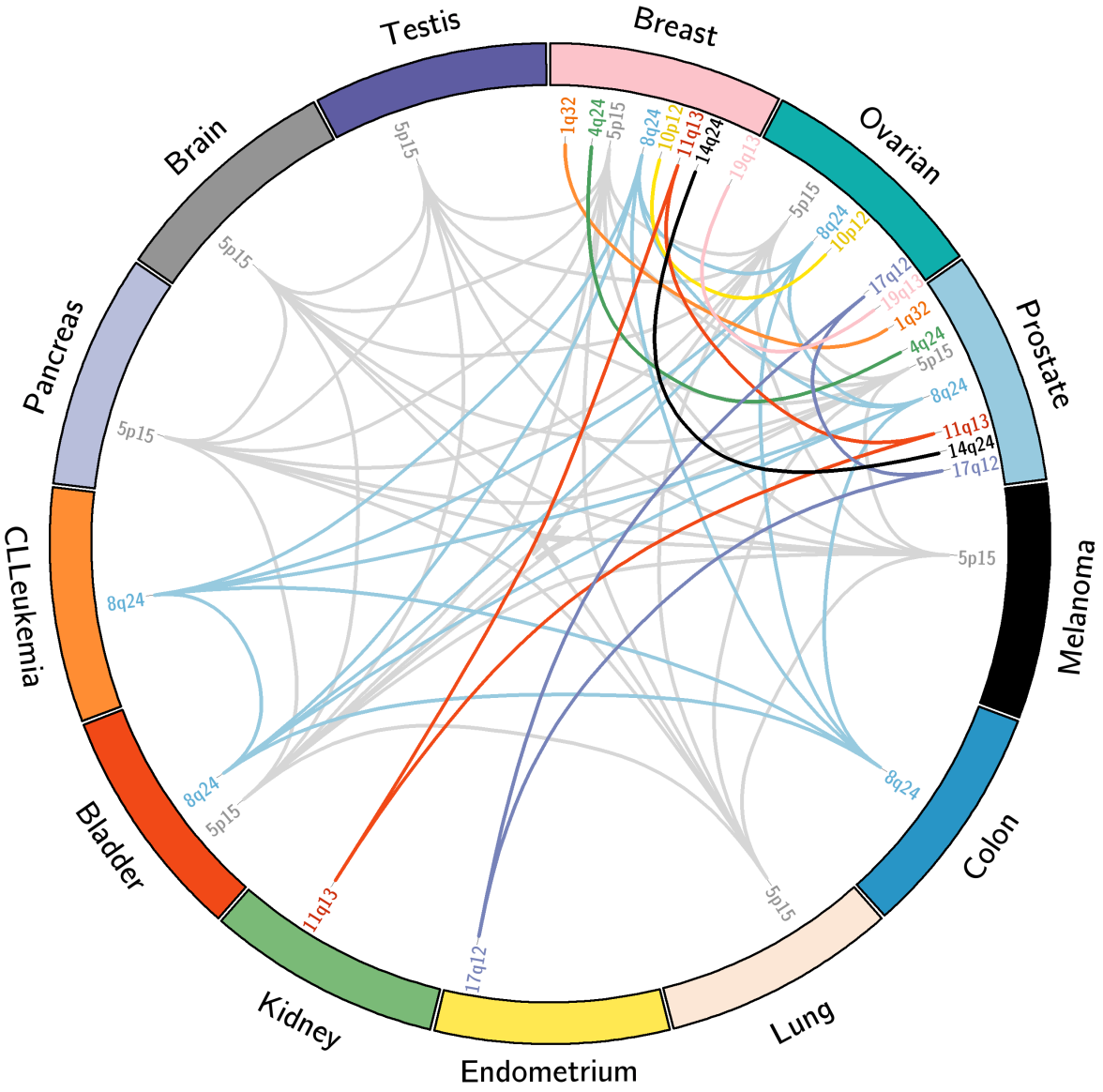
- One gene, multiple traits.

Examples of Pleiotropy

- Mendel's Pea Plants: those with colored seed coats also had colored flowers and colored leaf axils.
- Waardenburg syndrome: mutations in *PAX3* gene lead to hearing loss, different colored eyes, white forelock of hair.

- Antagonistic pleiotropy: *p53* gene stops damaged cells from reproducing (protects against cancer, but increases 'aging').

Pleiotropy



Sakoda, Jorgenson, Witte Nat Genet 2013

Evaluating Pleiotropy

- Conventional approach: calculate associations between genetic variant / mutation and multiple traits in a univariate manner.

- For binary traits:

$$\text{logit}\,(\text{Prob}(\mathbf{y}=1 \mid \mathbf{x}, \mathbf{C})) = \alpha + \mathbf{x}\beta + \mathbf{C}\gamma.$$

where

y is a vector of disease status

x is a vector of genotypes (e.g., 0, 1, 2)

C is a matrix of covariates (e.g., ancestry principal components)

β and γ are the corresponding regression coefficients.

Univariate Pleiotropy

ARTICLE

Pleiotropic Associations of Risk Variants Identified for Other Cancers With Lung Cancer Risk: The PAGE and TRICL Consortia

Colon

ORIGINAL ARTICLE

Pleiotropic effects of genetic risk variants for other cancers on colorectal cancer risk: PAGE, GECCO and CCFR consortia

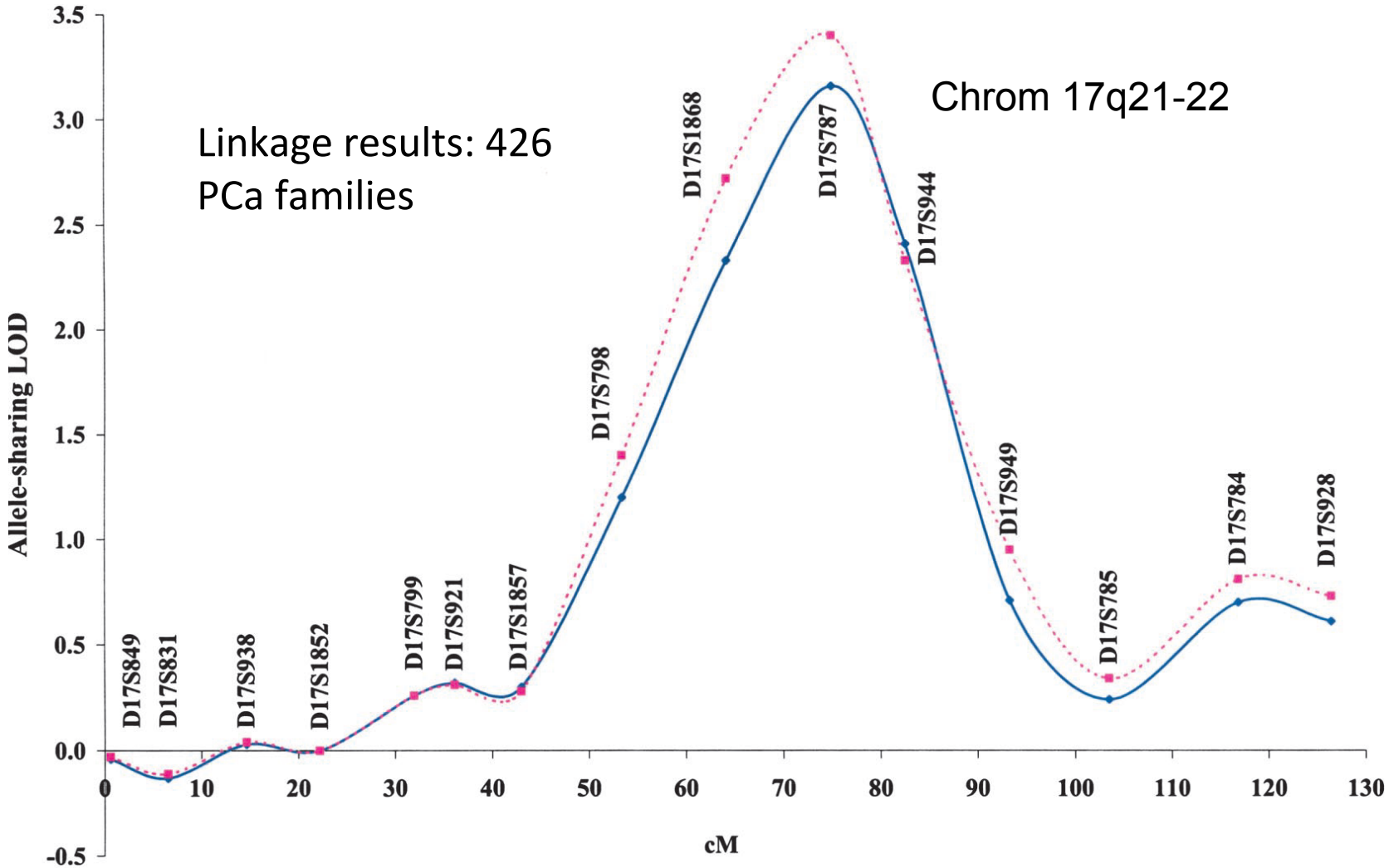
Carcinogenesis vol.00 no.00 p.1 of 6, 2014
doi:10.1093/carcin/bgu107
Advance Access publication May 15, 2014

Cross-cancer pleiotropic analysis of endometrial cancer: PAGE and E2C2 consortia

PheWAS

Phenome-wide association studies (PheWAS) analyze many phenotypes compared to a single genetic variant (or other attribute). This method was originally described using electronic medical record (EMR) data from EMR-linked in the Vanderbilt DNA biobank, BioVU, but can also be applied to other richly phenotyped sets.

Example: *HOXB13* and Cancer



Gillanders et al., JNCI 2004

HOXB13 & Cancer in Kaiser Cohort

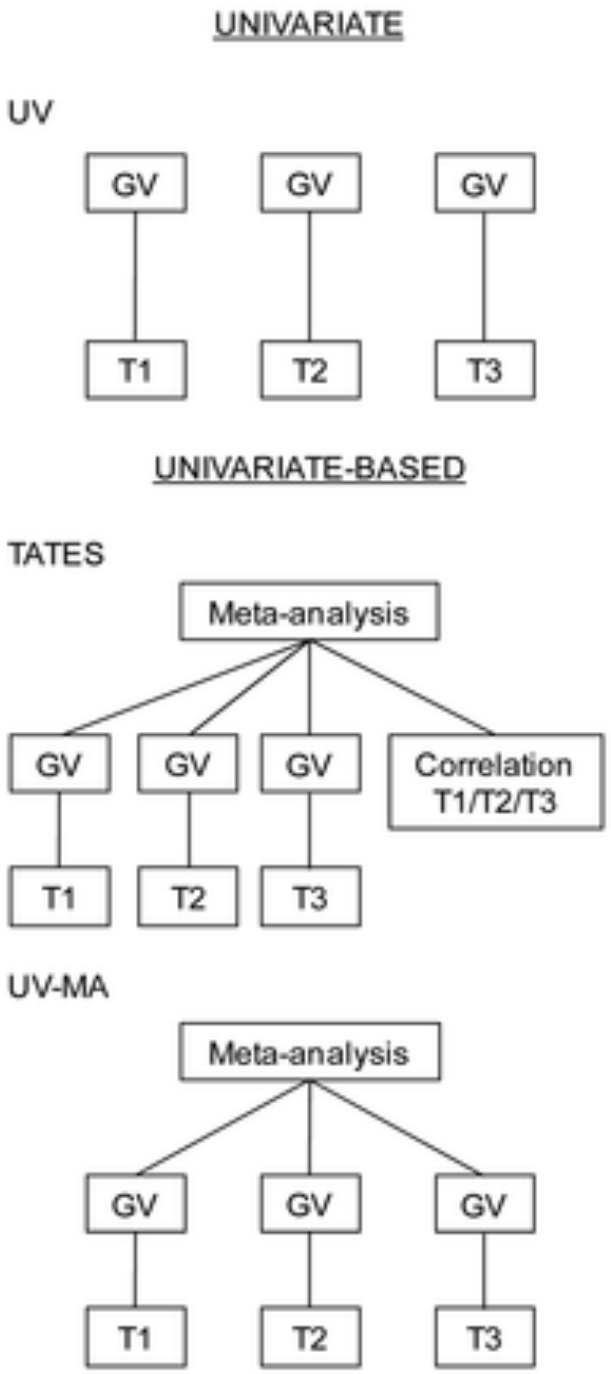
- GWAS on >100K individuals (not G84E).
- Imputed G84E into cohort (IMPUTE2).
- Reference panel:
  - 1KGP data: 2 carriers.
  - PCa study: 22 carriers, 71 non-carriers (Witte 2013)
- Focus on 15 cancers in Caucasians (>70K).

Prostate Cancer	Controls	OR	95% C.I.	P
1.86%	0.74%	2.62	(1.97, 3.48)	3.6×10^{-11}

Results for other Cancers

Cancer[a]	# Case carriers	# Cases	Frequency	Odds Ratio[b] (95% CI)	P-value[b]
Any Cancer[d]	165	13143	1.25%	1.60 (1.32, 1.93)	1.7×10^{-6}
Any (minus prostate)[e]	[illegible]	[illegible]	[illegible]	[illegible].62)	0.04
Breast	38	3183	1.19%	1.48 (1.04, 2.10)	0.029
Non-Hodgkin's Lymphoma	10	846	1.18%	1.81 (0.98, 3.36)	0.058
Kidney	5	303	1.65%	2.32 (0.94, 5.76)	0.069
Bladder	5	335	1.49%	2.05 (0.81, 5.22)	0.13
Melanoma	15	1301	1.15%	1.50 (0.89, 2.55)	0.13
Endometrium	6	578	1.04%	1.44 (0.64, 3.24)	0.38
Pancreas	2	149	1.34%	1.49 (0.28, 7.83)	0.64
Colon	11	1119	0.98%	1.42 (0.77, 2.60)	0.26
Lymphocytic Leukemia	1	218	0.46%	0.53 (0.05, 5.24)	0.59
Thyroid	3	217	1.38%	1.35 (0.35, 5.22)	0.67
Ovary	2	198	1.01%	1.32 (0.32, 5.49)	0.70
Multiple Myeloma	1	135	0.74%	1.26 (0.20, 7.92)	0.80
Lung	5	667	0.75%	1.10 (0.44, 2.72)	0.84
Oral	2	283	0.71%	0.98 (0.23, 4.10)	0.98

Can we do better than this?



van der Sluis et al.
PLoS Genet 2013

Galesloot et al.
PLoS One 2014

Meta-Analysis

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

Asset: Bhattacharjee *et al. AJHG,* 2012

Results for other Cancers

Cancer[a]	# Case carriers	# Cases	Frequency	Odds Ratio[b] (95% CI)	P-value[b]
Any Cancer[d]	165	13143	1.25%	1.60 (1.32, 1.93)	$1.7x10^{-6}$
Any (minus prostate)[e]	91	9167	0.99%	1.28 (1.01, 1.62)	0.04
Breast				10)	0.029
Non-Hodgkin's Lymph				36)	0.058
Kidney	5	303	1.65%	2.32 (0.94, 5.76)	0.069
Bladder	5	335	1.49%	2.05 (0.81, 5.22)	0.13
Melanoma	15	1301	1.15%	1.50 (0.89, 2.55)	0.13
Endometrium	6	578	1.04%	1.44 (0.64, 3.24)	0.38
Pancreas	2	149	1.34%	1.49 (0.28, 7.83)	0.64
Colon	11	1119	0.98%	1.42 (0.77, 2.60)	0.26
Lymphocytic Leukemia	1	218	0.46%	0.53 (0.05, 5.24)	0.59
Thyroid	3	217	1.38%	1.35 (0.35, 5.22)	0.67
Ovary	2	198	1.01%	1.32 (0.32, 5.49)	0.70
Multiple Myeloma	1	135	0.74%	1.26 (0.20, 7.92)	0.80
Lung	5	667	0.75%	1.10 (0.44, 2.72)	0.84
Oral	2	283	0.71%	0.98 (0.23, 4.10)	0.98

Can we do better than this?

Recall that this is a cohort!

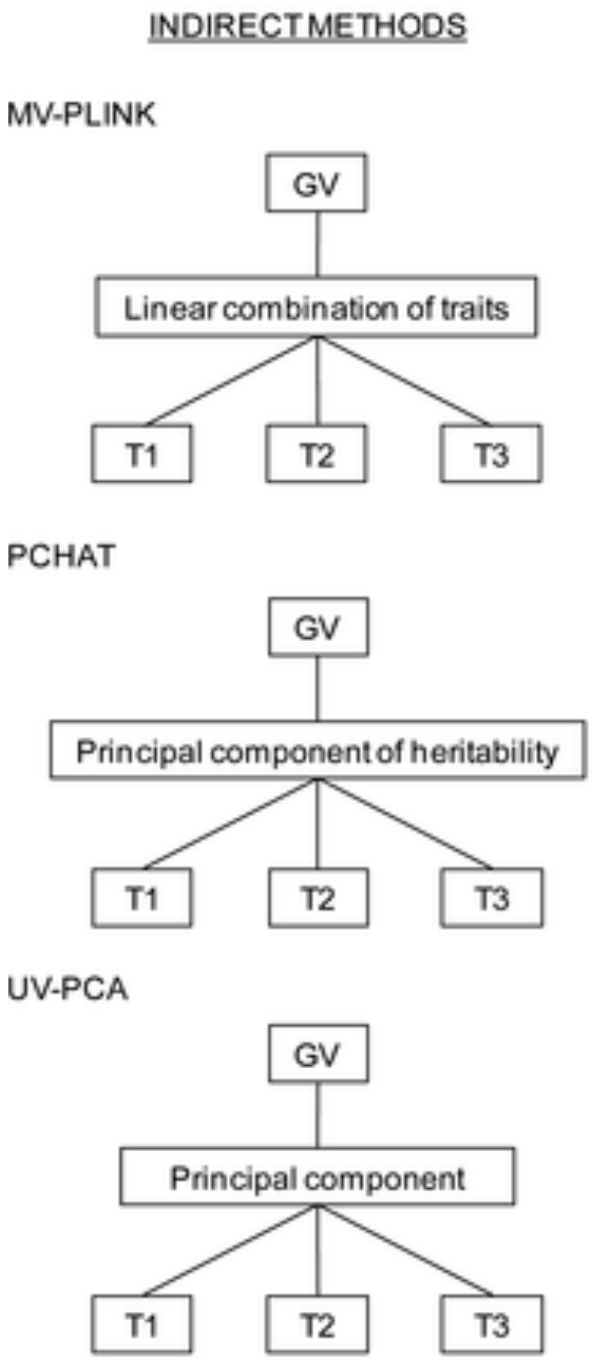
OR = 1.50
p = 0.042

Hoffmann…Witte, Submitted

Multivariate Approach

- Can be more:
  - consistent with underlying biology;
  - powerful than univariate.
- Power gain due to:
  - genetic correlations among traits;
  - fewer tests.
- (Univariate often ignore multiple comparisons.)

Lui et al. Genet Epi 2009
Galesloot et al. PLoS One 2014



MQFAM: Canonical Correlation Analysis / MANOVA
Ferreira and Purcell, Bioinformatics 2009

Reduce trait dimensions to optimal linear combinations

Heritability perspective.
Klei et al. Genet Epi 2008

What if we want to retain original trait form?

Galesloot et al.
PLoS One 2014

Evaluating Pleiotropy: multinomial model

- Multinomial logistic regression

$$\text{logit}\,(\text{Prob}(\mathbf{y_i}=1 \mid \mathbf{x}, \mathbf{C})) = \alpha_i + \mathbf{x_i}\beta_i + \mathbf{C}\gamma_\mathbf{i}.$$

y is now multivariate, dimension = # of traits

β_i are cancer-specific regression coefficients

- Test different models, specify assumptions about β_l.
- Null model: H_0: $\beta_1=\beta_2=...=\beta_k=0$.
- General pleiotropy model: genetic variant **x** has a different effect on each trait.

Evaluating Pleiotropy: multinomial model

- First, LRT comparing general and null models.
- Second, for each variant with $p<5x10^{-5}$, fit:
  - Basic, assumes that $\beta_1=\beta_2=...=\beta_k$ (all equal).
  - Subset, assumes particular trait subsets have the same variant associations (most likely pleiotropic).
  - Best fit w/ Bayesian Information Criterion (BIC).
- Report only best fitting model and the p-value from the LRT of the general to null models.
- Single model framework controls the Type I error rate at an experimentwise alpha-level

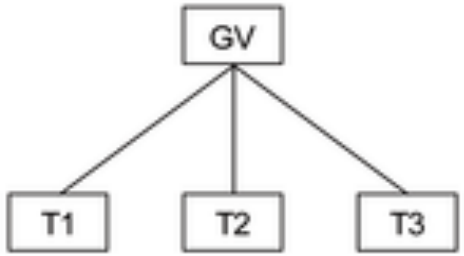
Lee et al., Hum Hered 2011; 72:10-20.

Complication

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

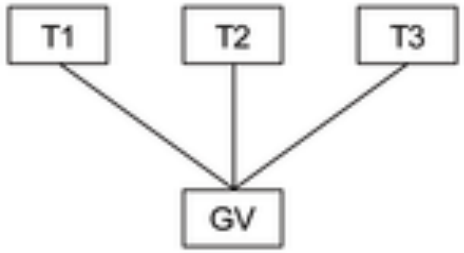
DIRECT METHODS

MV-SNPTEST



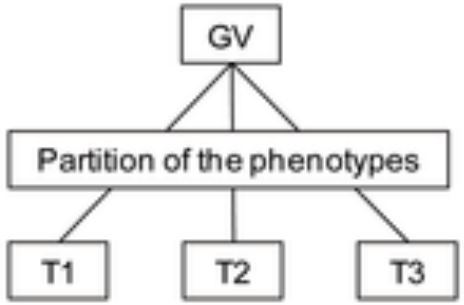
Marchini et al. Nat Genet 2007

MultiPhen



O'Reilly et al. PLoS One 2012

MV-BIMBAM



Stephens PLoS One 2013

Galesloot et al. PLoS One 2014

MultiPhen: ‘Reverse Genetics’

- Ordinal regression of genotype on traits

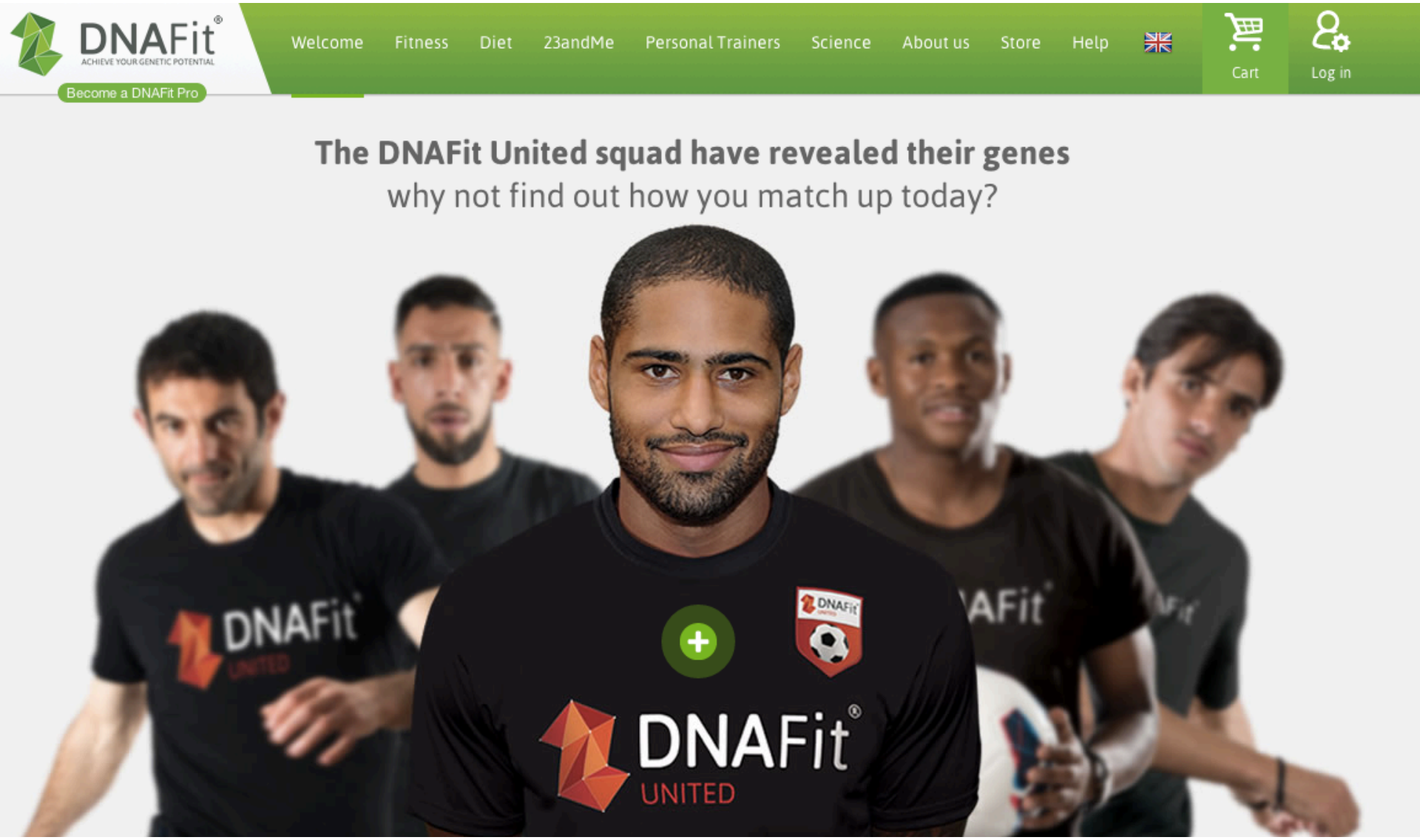
$$\log\left(\Pr(\mathbf{G}>m\,|\,\mathbf{Y})/\Pr(\mathbf{G}\leq m\,|\,\mathbf{Y})\right) = \alpha_m + \mathbf{Y}'\beta + \mathbf{C}\gamma$$

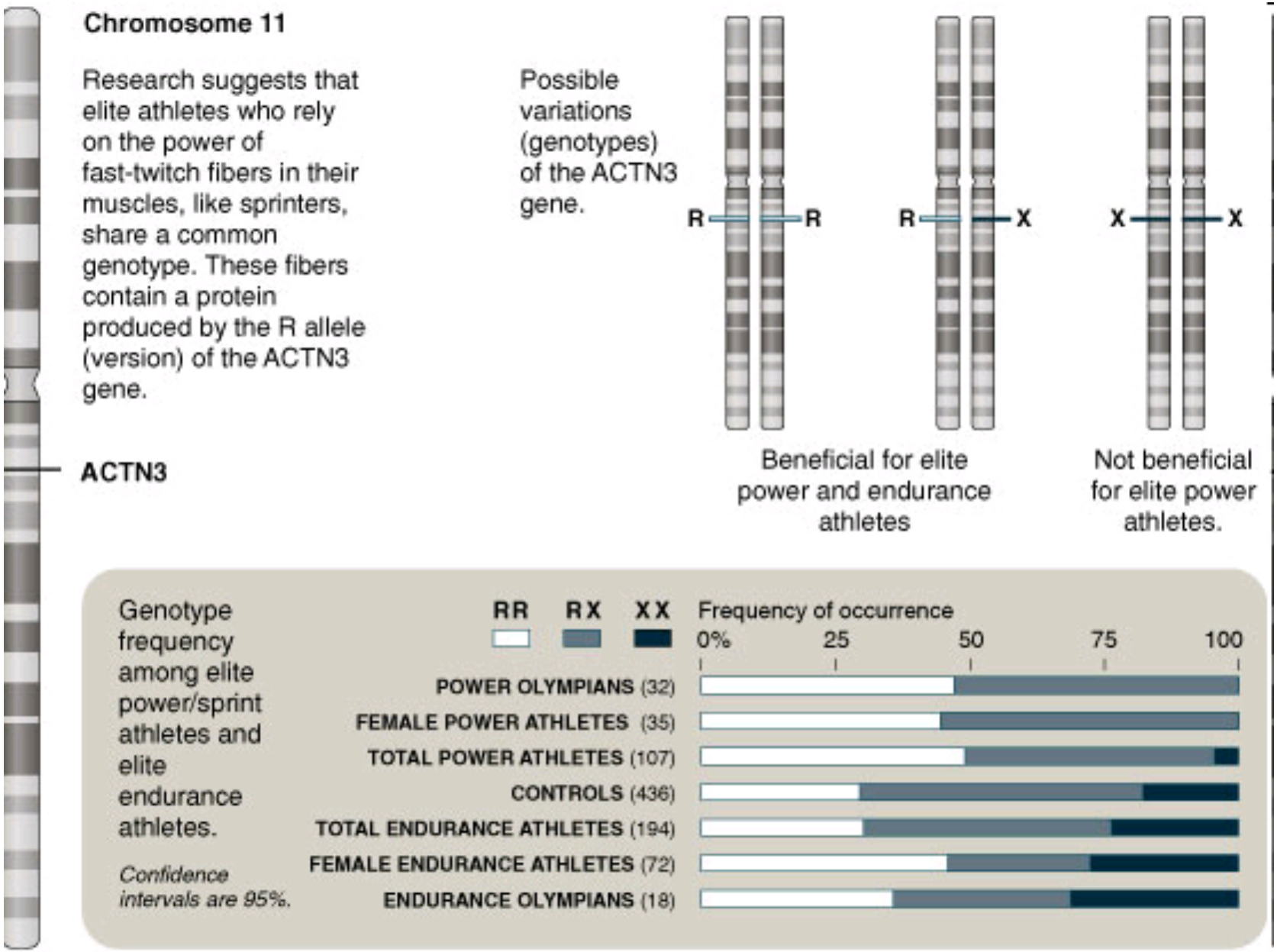
$$m = 0, 1$$

What now?

- Model testing / selection.

Genetic Prediction





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

NY Times, 11/30/08

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

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