2017 SISG Brisbane Module 10: Statistical & Quantitative Genetics of Disease

Lecture 5
Polygenic models of disease risk
Naomi Wray

Aims of Lecture 5

- 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

- Introduction to polygenic risk scores

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:

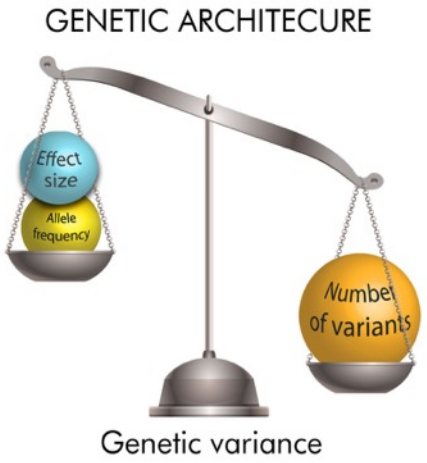
Many risk variants of small effect identified

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



Start simple
- Assume all risk alleles have the same frequency
- Assume all risk alleles have the same effect size
- Then risk is a reflection of the count of the number of risk alleles
- What is the shape of the relationship between count of risk alleles and probability of disease?

Count risk alleles

SNP 1:	Aa	1
SNP 2:	Aa	1
SNP 3:	AA	2
SNP 4:	aa	0
SNP 5:	aa	0

4 risk alleles =s

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

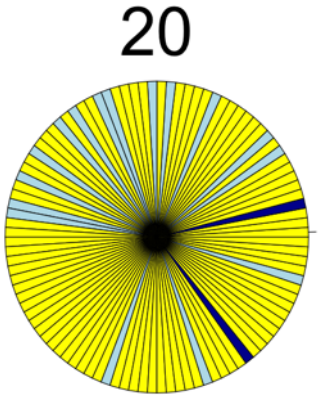
n loci 100

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

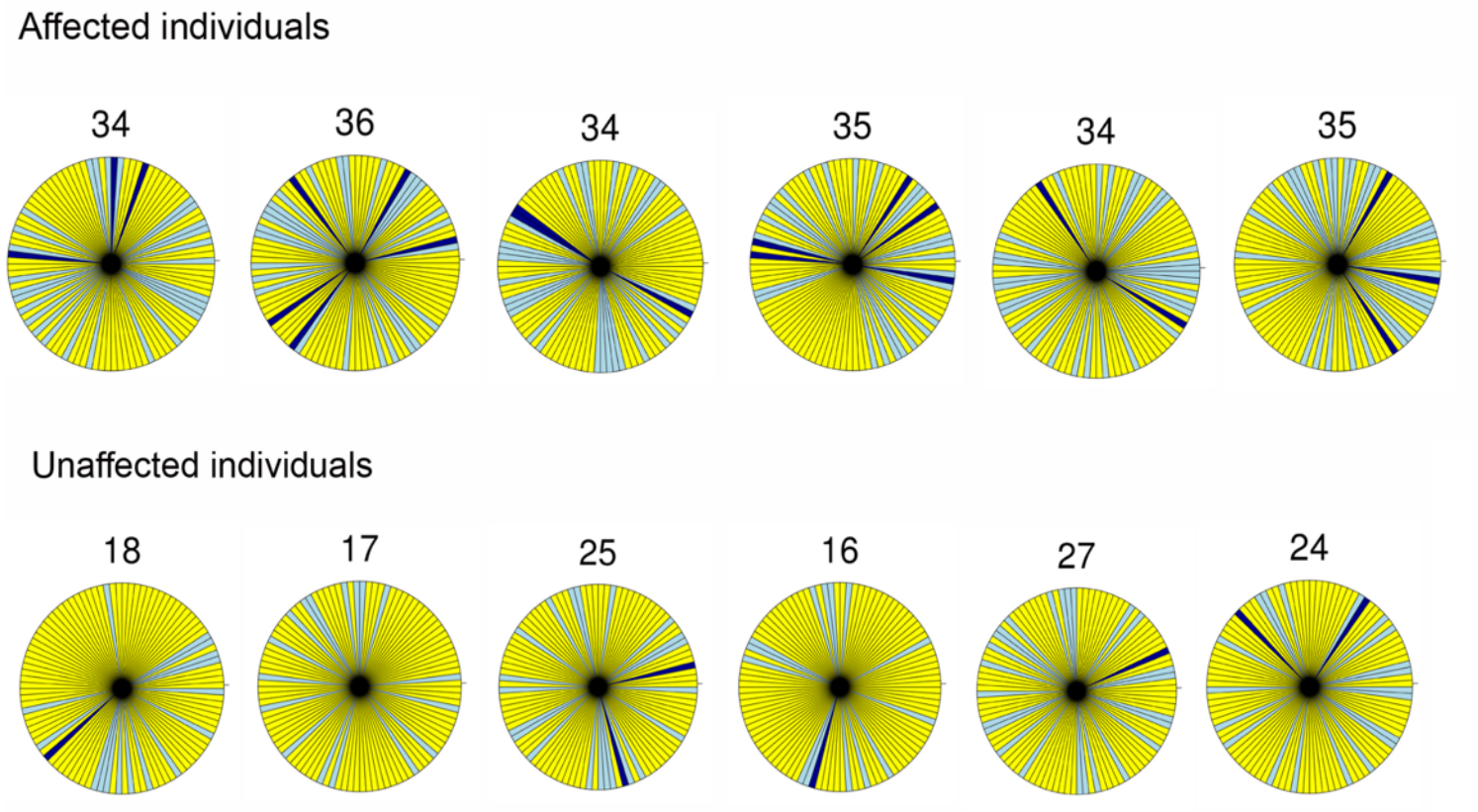
Visualising common complex genetic diseases
Polygenic genetic architecture

- 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

How to combine risk loci to explain disease

Additive on disease scale

Multiplicative on disease scale

Constrained multiplicative on disease scale

Multiplicative Odds on disease scale

Liability threshold model

Basic single locus risk model

Single locus disease model:

G = genotype; D=disease; K = overall disease risk in population

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

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

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

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

if $R_{Bb} = R$; $R_{BB} = R^2$ multiplicative on disease scale

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

$R_{BB} = 2^* R_{Bb}$ additive on the disease scale

Two-locus risk model

Single locus disease model:

G = genotype; D=disease; K = overall disease risk in population

P(G) P(D)

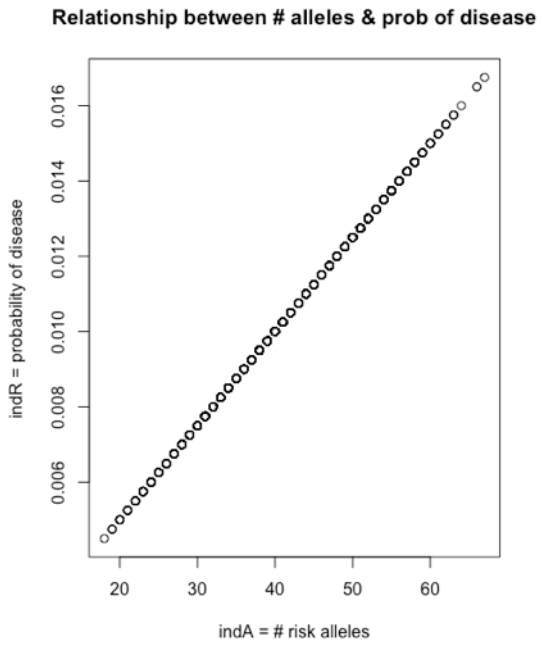
# risk alleles	1st locus/2nd locus	Frequency	Risk Additive	Risk Multiplicative
0	0/0	$(1\text{-}p)^4$	f_0	f_0
1	1/0 or 0/1	$2p(1\text{-}p)(1\text{-}p)^2 + 2p(1\text{-}p)(1\text{-}p)^2$	f_0R	f_0R
2	2/0 or 0/2 or 1/1	$p^2 + p^2 + 4p^2(1\text{-}p)^2$	$2f_0R$	f_0R^2
3	2/1 or 1/2	$p^2 2p(1\text{-}p) + p^2 2p(1\text{-}p)$	$3f_0R$	f_0R^3
4	2/2	p^4	$4f_0R$	f_0R^4

Sum P(D) = K

f_0 = probability of disease with no risk alleles. This baseline probability differs between the models

Additive on the disease scale

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



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

Constraint

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

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

So b = K/2npR

Looking at the additive model

Base

N = 1e5 # number of families 100,000

n = 100 # number of loci

R = 1.1 # relative risk of each risk allele

p = 0.2 # allele frequency of each risk allele

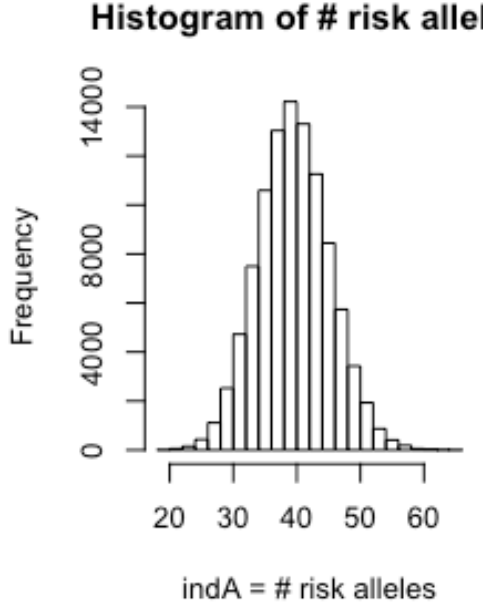
K = 0.01 # probability of disease

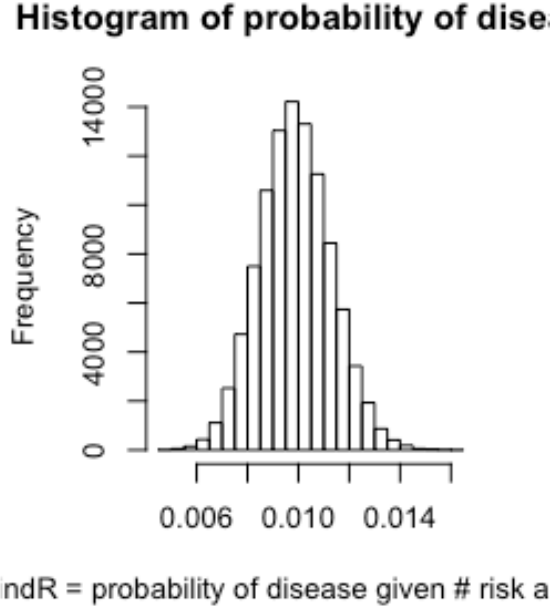
Follow up:

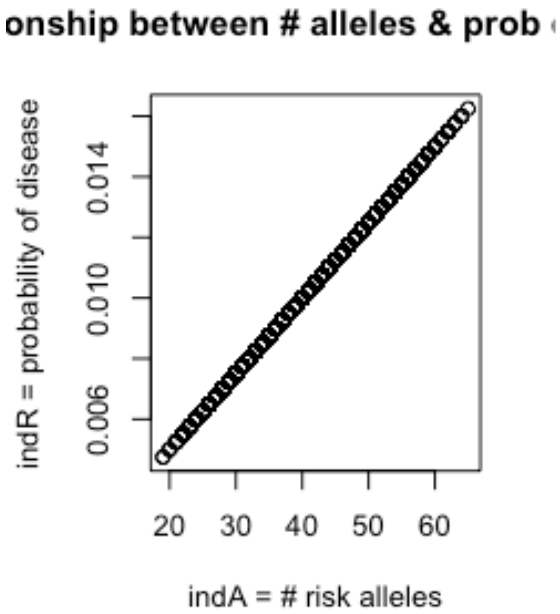
Base, R=1.5, p=0.5, K =0.1

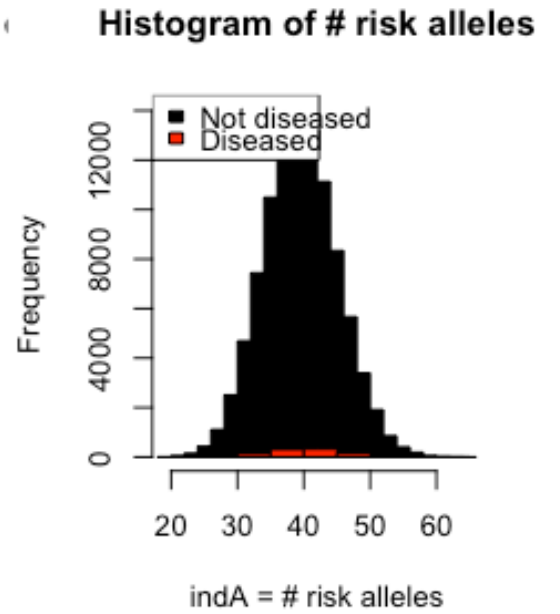
Look at maximum probability of disease and consider whether this model will generate an increased risk in relatives

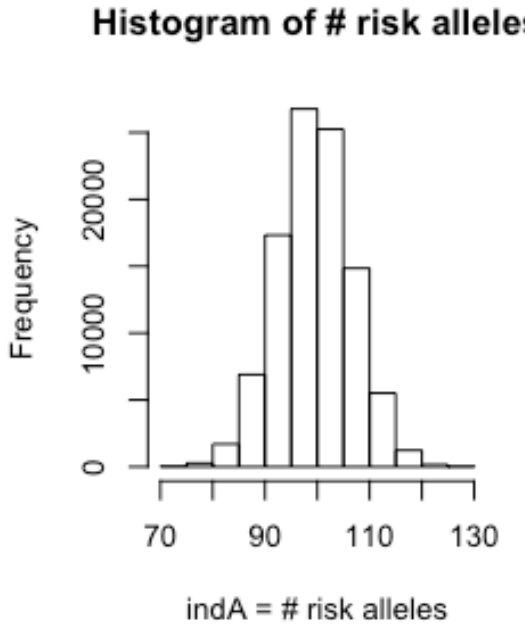
SWITCH TO R

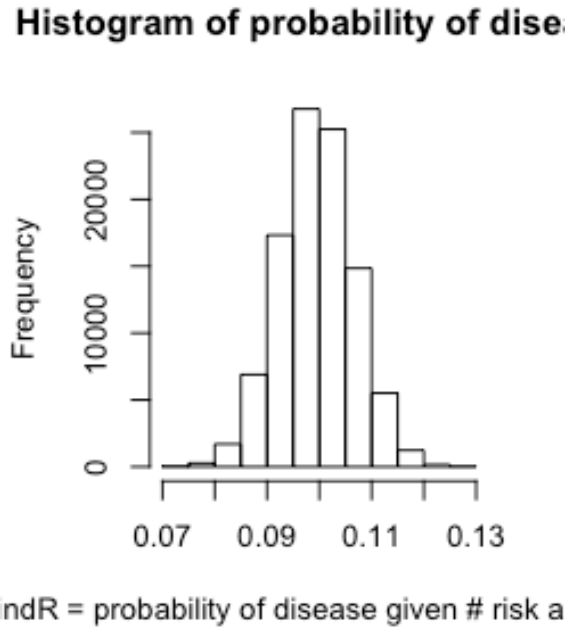

Histogram of # risk alleles
Frequency
0
4000
8000
14000
20
30
40
50
60
indA = # risk alleles

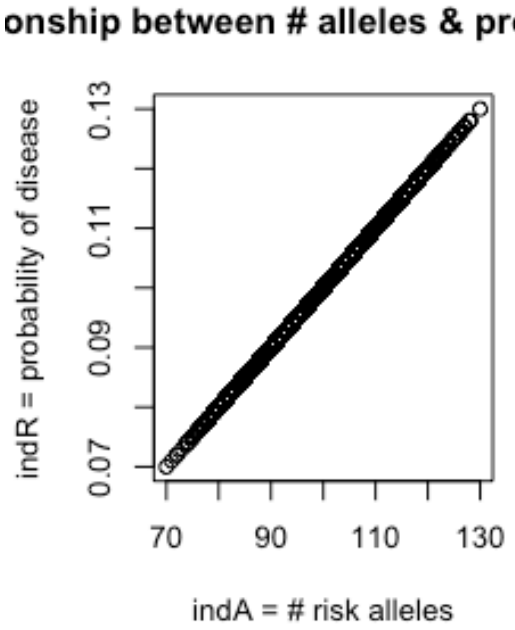

Histogram of probability of dise
Frequency
0
4000
8000
14000
0.006
0.010
0.014
ndR = probability of disease given # risk a

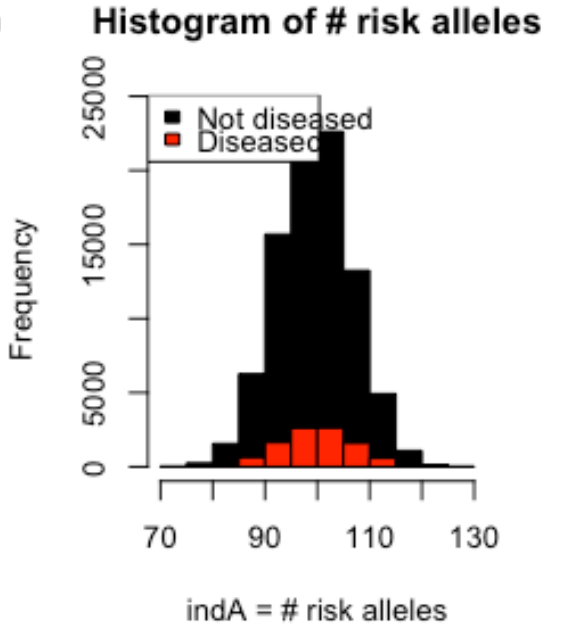

onship between # alleles & prob
indR = probability of disease
0.006
0.010
0.014
20
30
40
50
60
indA = # risk alleles


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


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


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


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


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

Additive model

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

- Can "fudge" the additive model by saying
  - $P(D \mid x < n1) = 0$
  - $P(D \mid n1<x<n2)$ = 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

Base

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

Follow up:

Base, K=0.1

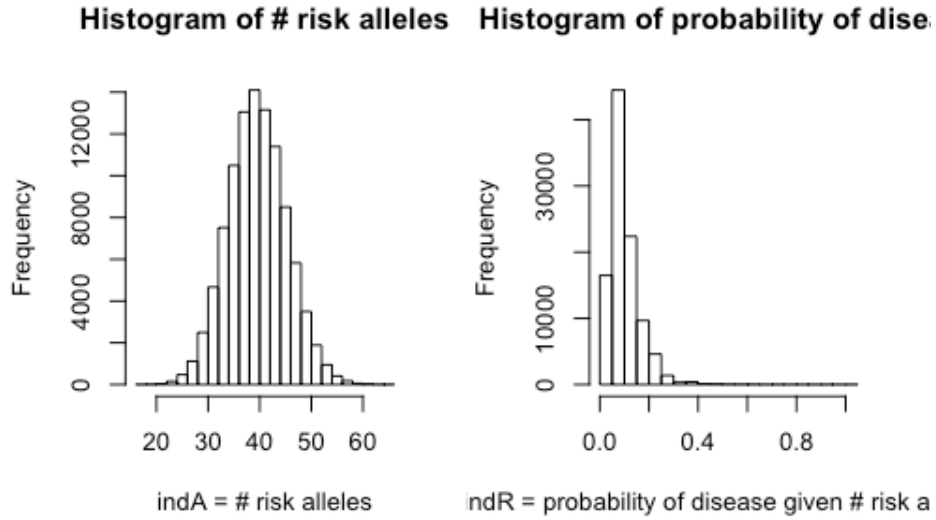
Base K = 0.1, R =1.2

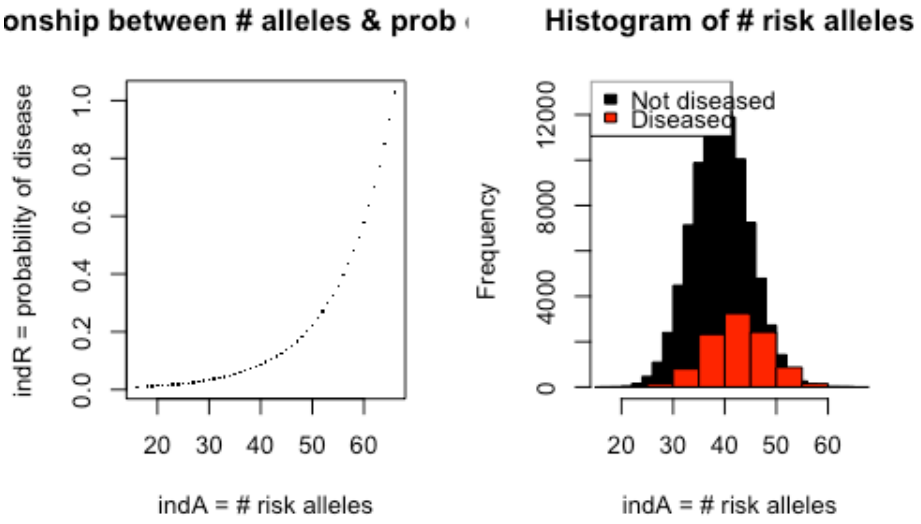
Look at maximum probability of disease and consider whether this model will generate an increased risk in relatives

Add fix

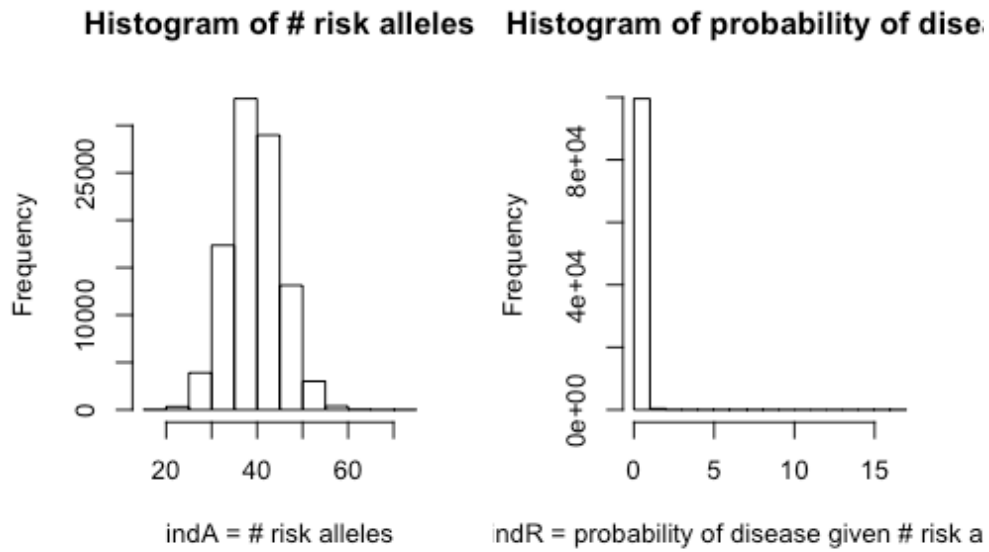
SWITCH to R

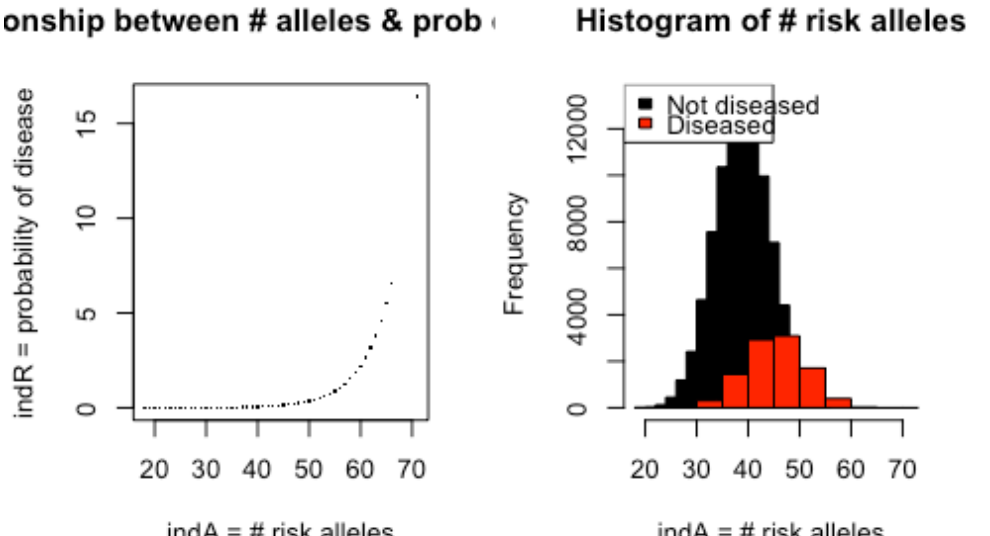
K=0.1, p=0.2, R=1.1


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


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

K=0.1, p=0.2, R=1.2


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


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

Multiplicative model

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

BUT

- Probability of disease for an individual can be > 1

IF constrain so that max probability of disease is 1

THEN

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

Epidemiology risk model

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

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

s = number of risk loci carried by an individuals

γ = odds ratio for each risk locus

P(Disease | x = s)= Cγ^s/(1-Cγ^s)

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

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

Epidemiology risk modelling

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

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

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

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

Epidemiology risk model

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

Two relatives, with risk of disease R_1R_2

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

Probability that both are affected R_1R_2

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

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

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

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

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

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

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

Liability threshold model

Doesn't directly 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 – so can be indirectly parameterised in terms of number of risk loci

Variance explained by a locus

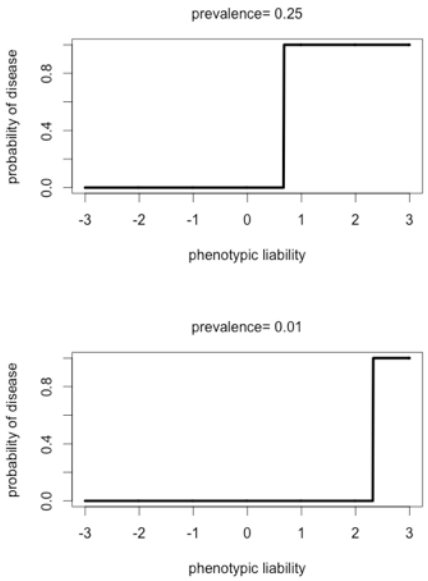
depends on frequency (p) and effect size(a) : $2p(1-p)a^2$

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?

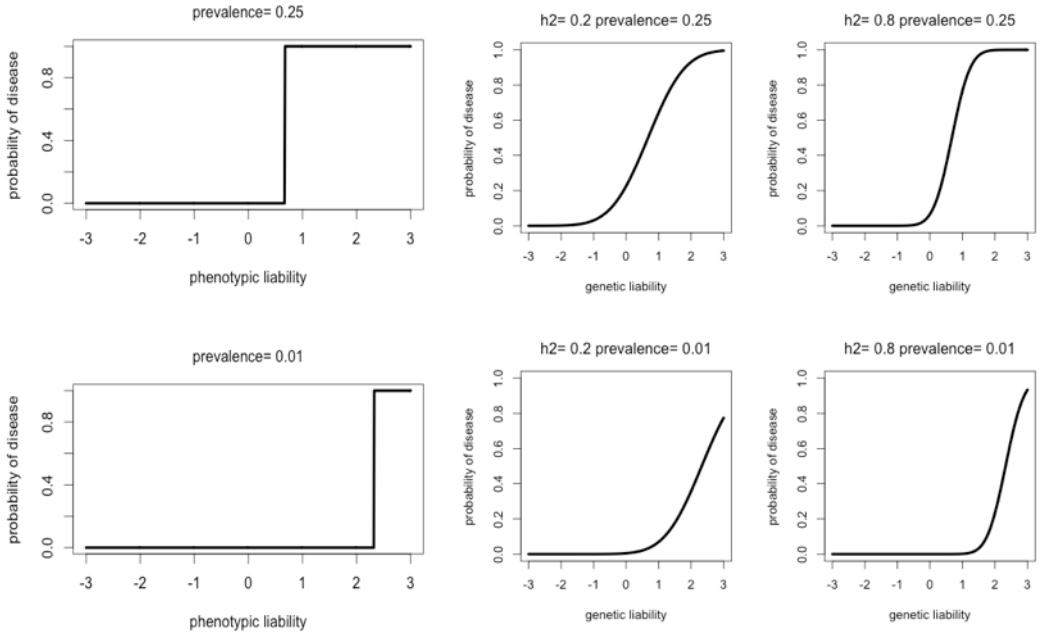
- People are classed as diseased or not disease, any error in this classification, contributes of a heritability of < 1.

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

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

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

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



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

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

Probit model

Two parameters: disease prevalence and heritability

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

Curnow (1972) The multifactorial model for the inheritance of liabilty to disease and its implications for risk to relatives. Biometrics
Curnow & Smith (1975) Multifactorial models for familial diseases in man. J Royal Stat Soc A 138

Controversy – many models fit empirical data

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

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

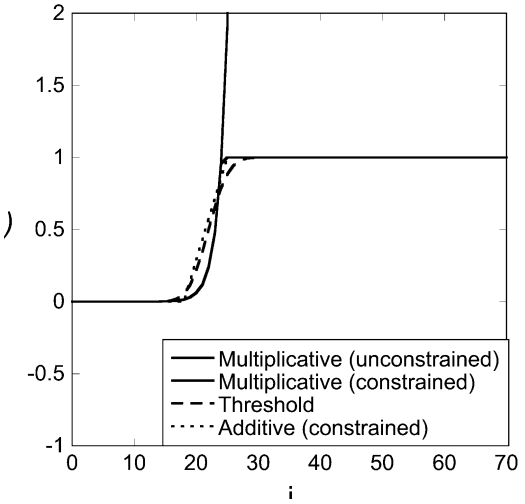
Need risk to disease from relatives of different types of relatives to start to distinguish between models
Not easy to collect, large sampling variances

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

Exchangeable models of disease

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

Probability of disease



Number of risk alleles(x)

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

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

Which polygenic model to use?

The liability threshold model is the model of choice because

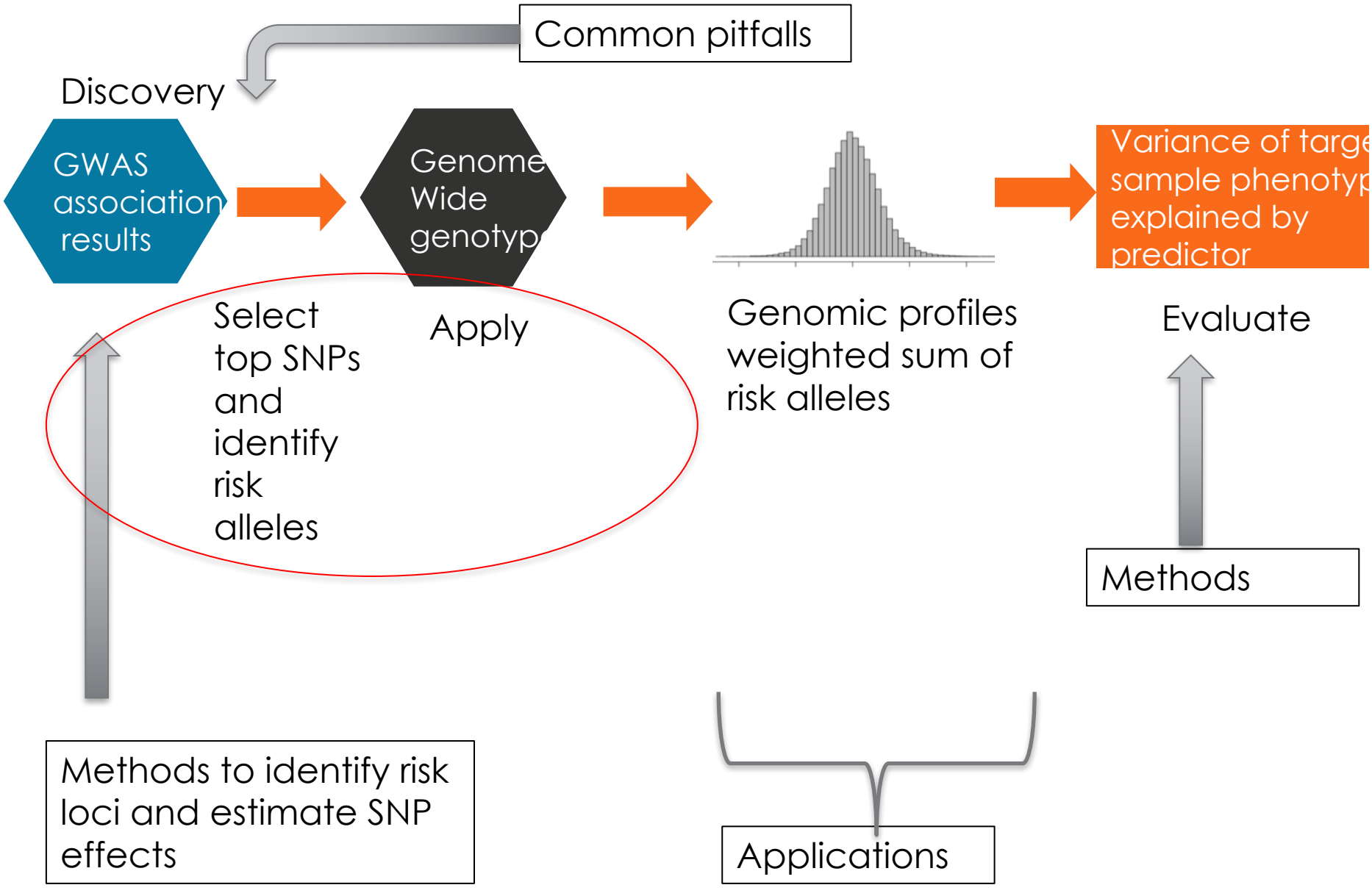
- It is the simplest parameterization that fits the observable data
- It is mathematically tractable
- It makes least assumptions about genetic architecture

"Most models are wrong some models are useful"

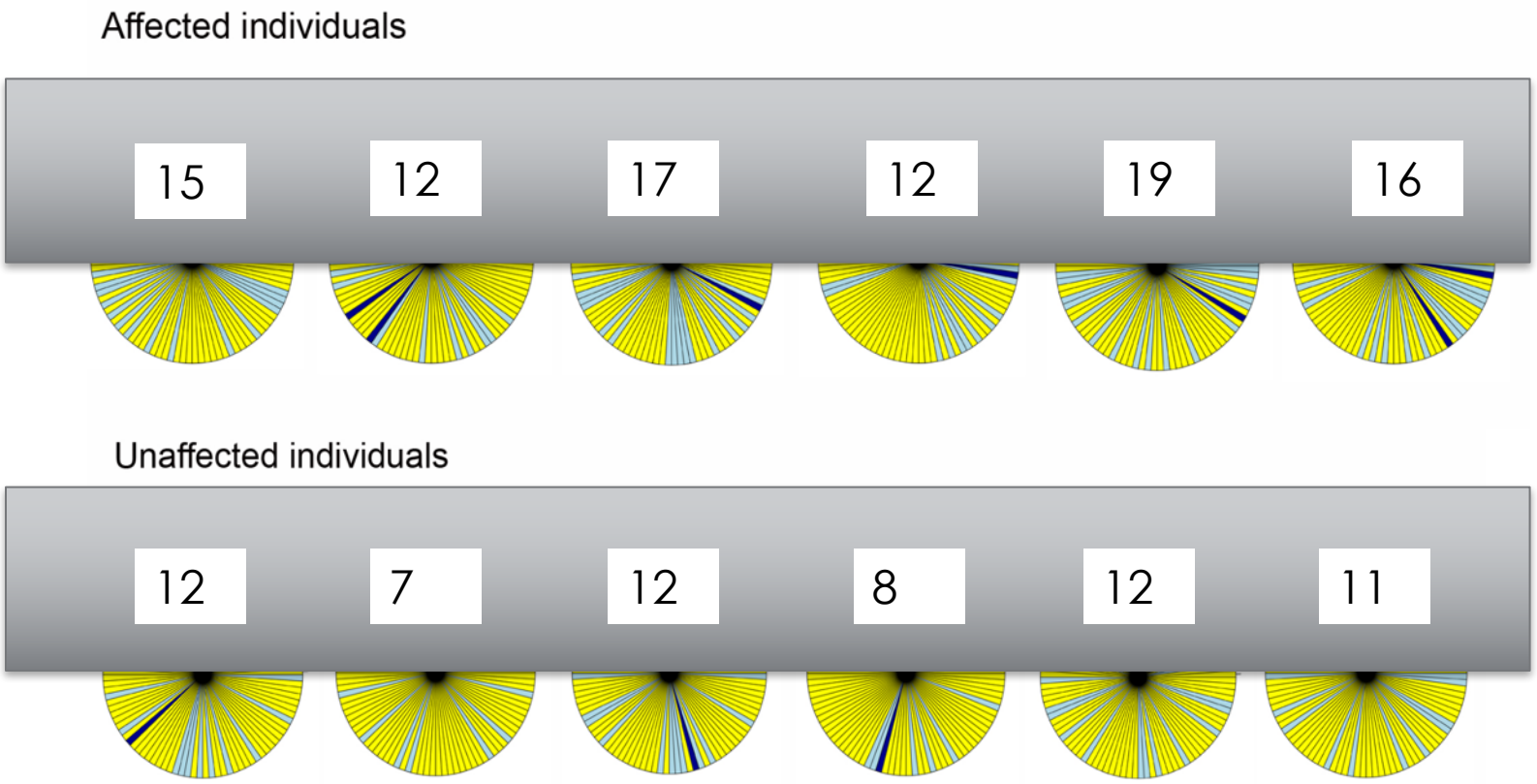
From theory to data

Profile risk scoring

SNP profiling schematic



Visualising variation between individuals for common complex genetic diseases



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

Acronyms, synonyms

- PRS
  - Polygenic risk score
- GPRS
  - Genomic profile risk score
- PGS
  - Polygenic score
- GRS
  - Genetic risk score
- Gene score
- Genetic score
- Genotypic score
- Allele score
- Profile score
- Linear predictor

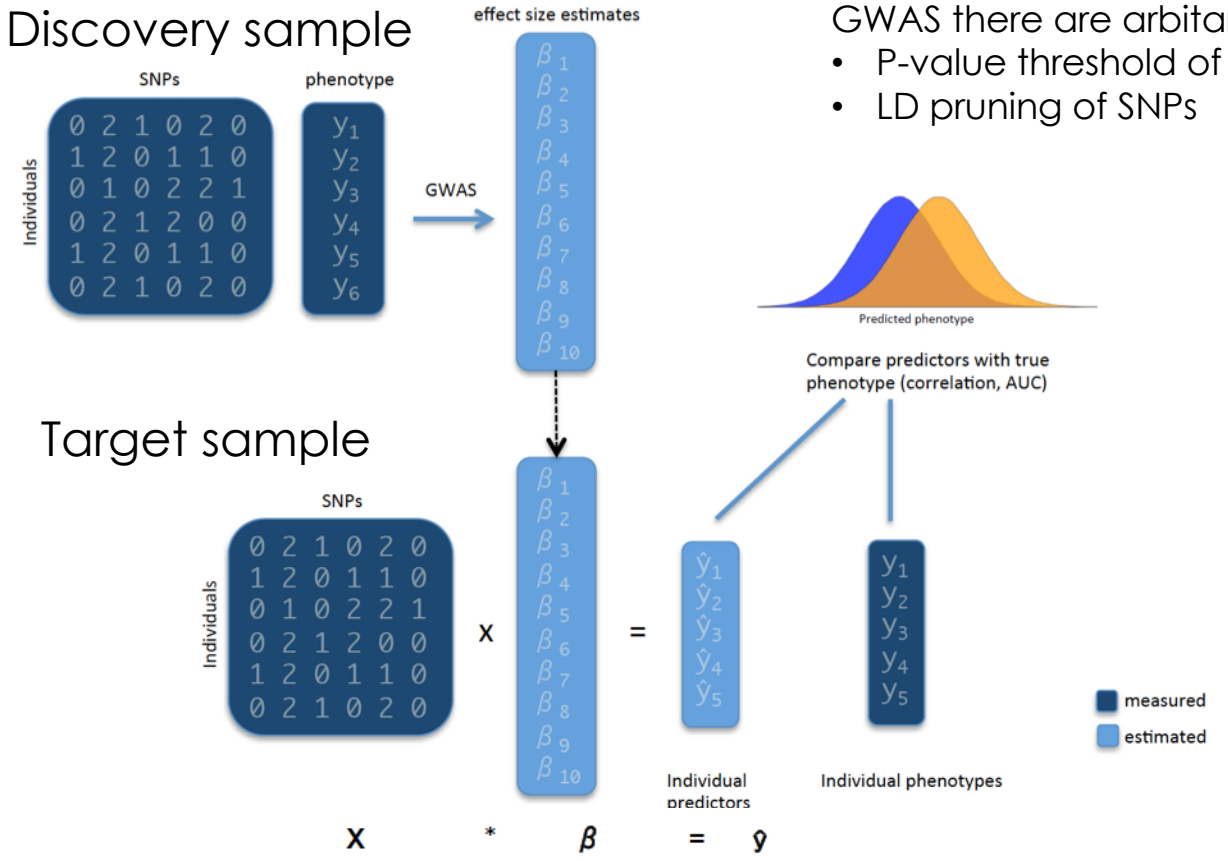
- Discovery = Training
- Target = Replication = Test

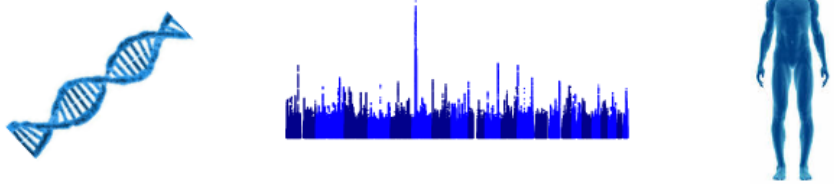
Slide credit: Frank Dudbridge

Polygenic risk predictons turns SNP efects into predicted genetic risk for an individual

When SNP effects are estimated by standard GWAS there are arbitary decisions

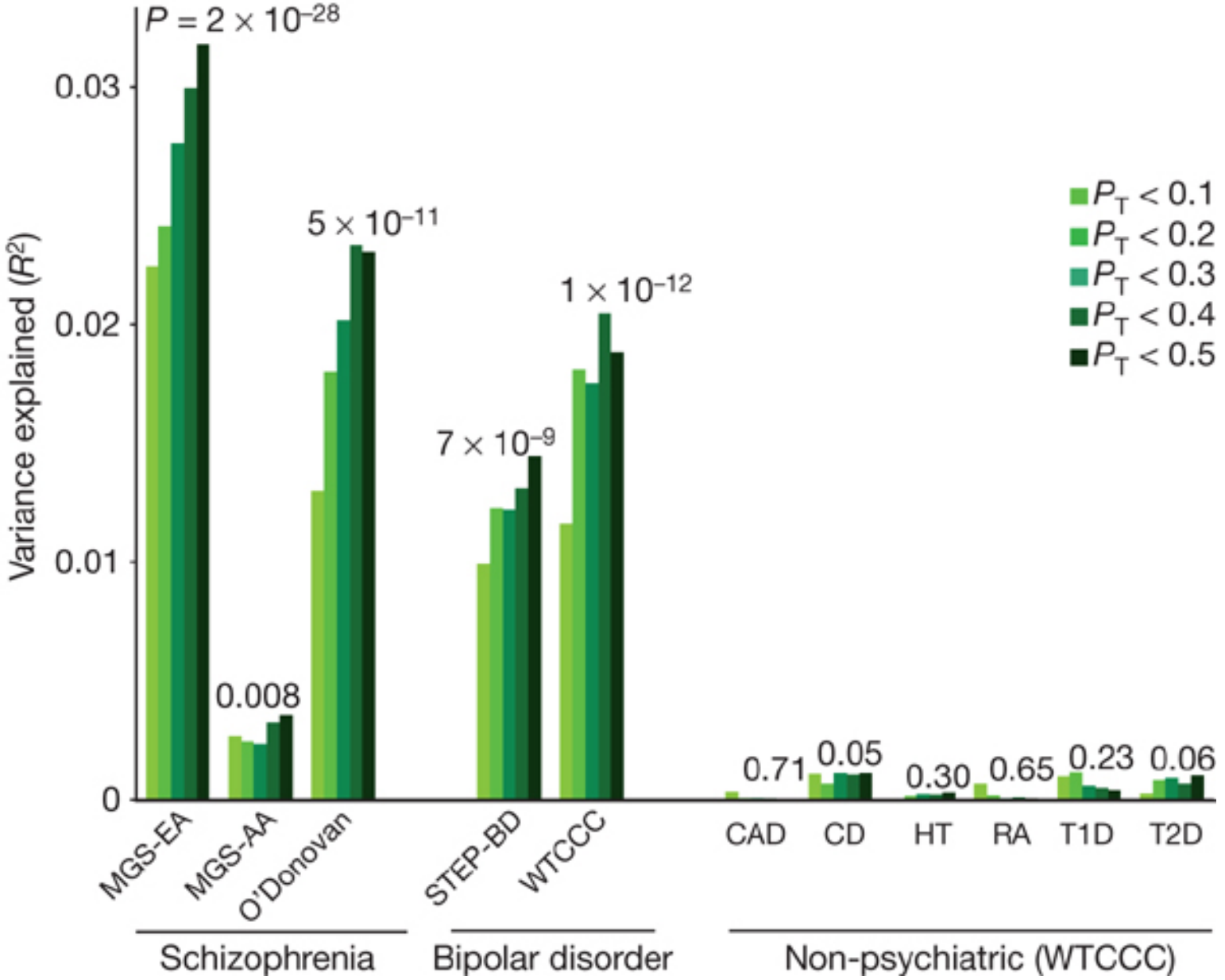
- P-value threshold of SNPs
- LD pruning of SNPs





Slide credit: Robert Maier

First Application of Risk Profile Scoring



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

PRS ("profile scoring") in PLINK

THE JOURNAL OF CHILD PSYCHOLOGY AND PSYCHIATRY

Research Review: Polygenic methods and their application to psychiatric traits

Naomi R. Wray,[1] Sang Hong Lee,[1] Divya Mehta,[1] Anna A.E. Vinkhuyzen,[1] Frank Dudbridge,[2] and Christel M. Middeldorp[3,4]

- Reduce the SNPs to a set in approximate LD
  - Best done using "clumping": keep the most associated SNP, remove those in LD with it, then keep the next most associated remaining SNP, etc
  - Recommend LD threshold of $r^2<0.1$

  plink --clump-p1 1 --clump-p2 1 --clump-r2 0.1 --clump-kb 500
- Estimate effect sizes of these SNPs

  plink --assoc

PRS ("profile scoring") in PLINK

- Create a file listing SNPs and their effect sizes
  - Take logs of odds ratios

  awk
- Create a file listing p-value thresholds to select SNPs into PRS
- Generate PRS for subjects in the target sample

  plink --score --q-score-range
- Regress target phenotype against PRS

  glm(y~prs, family=binomial)

PRSice

- R package to simplify calculations of PRS
- http://prsice.info

Bioinformatics, 31(9), 2015, 1466–1468

PRSice: Polygenic Risk Score software

Jack Euesden*, Cathryn M. Lewis and Paul F. O'Reilly*

MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

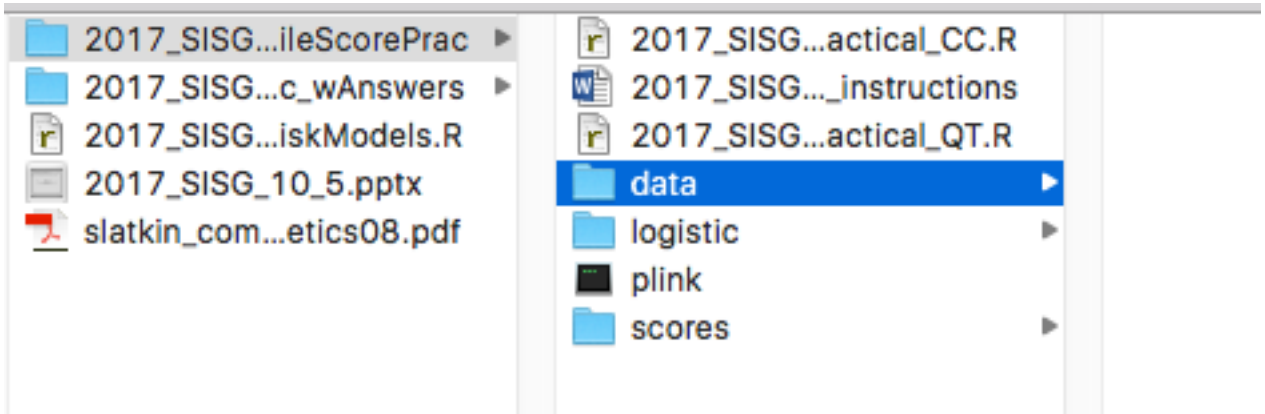
Practical –Risk Models

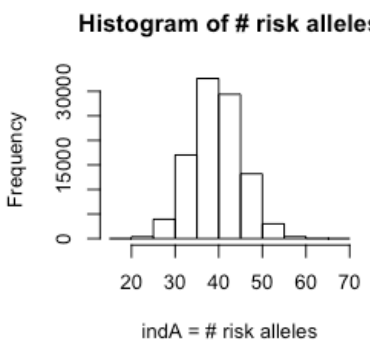
1. Additive risk model.
   a. Run code
   b. Change parameters
2. Multiplicative risk model.
   a. Run code
   b. Change parameters
3. Logistic risk model.
   a. Run code
   b. Change parameters
4. Liability threshold model
   a. Run code
   b. Change parameters

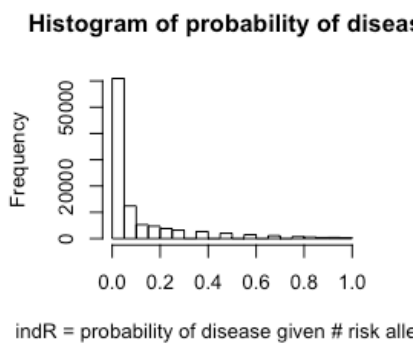
Practical – Polygenic risk scoring

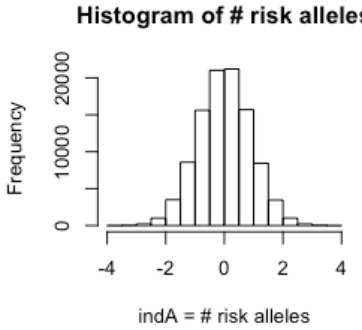
- Demonstrate impact of overlap between Discovery & Target sample
- Learn to simulate SNP data
- Explore impact of
  - N sample size
  - M markers
  - h^2 true variance explained by SNPs

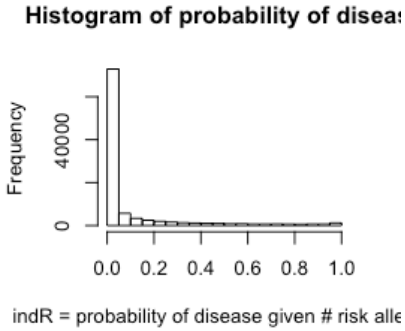
on variance explained by predictor

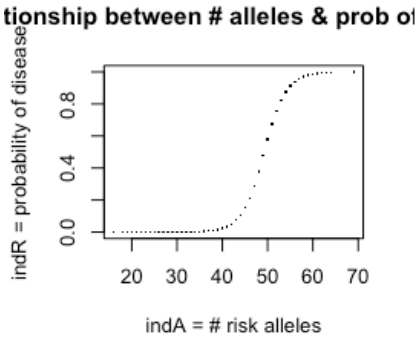


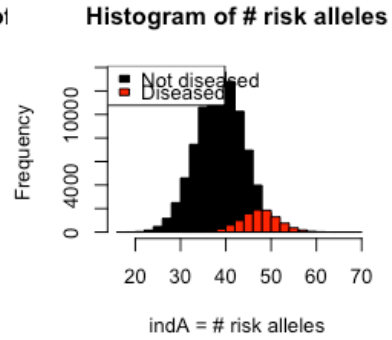

Histogram of # risk alleles
Frequency
0
15000
30000
20
30
40
50
60
70
indA = # risk alleles

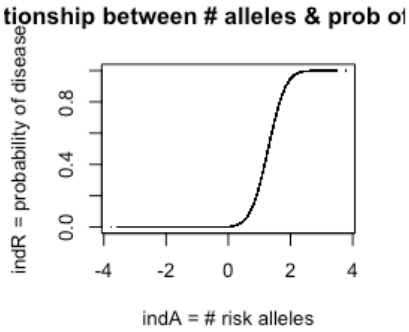

Histogram of probability of disea
Frequency
0
20000
50000
0.0
0.2
0.4
0.6
0.8
1.0
indR = probability of disease given # risk alle

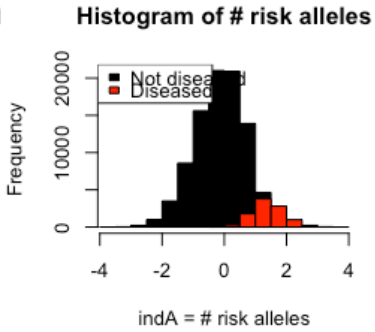

Histogram of # risk alleles
Frequency
0
10000
20000
-4
-2
0
2
4
indA = # risk alleles


Histogram of probability of disea
Frequency
0
40000
0.0
0.2
0.4
0.6
0.8
1.0
indR = probability of disease given # risk alle


tionship between # alleles & prob of
indR = probability of disease
0.0
0.4
0.8
20
30
40
50
60
70
indA = # risk alleles


Histogram of # risk alleles
Not diseased
Diseased
Frequency
0
4000
10000
20
30
40
50
60
70
indA = # risk alleles


tionship between # alleles & prob of
indR = probability of disease
0.0
0.4
0.8
-4
-2
0
2
4
indA = # risk alleles


Histogram of # risk alleles
Not disea
Diseased
Frequency
0
10000
20000
-4
-2
0
2
4
indA = # risk alleles