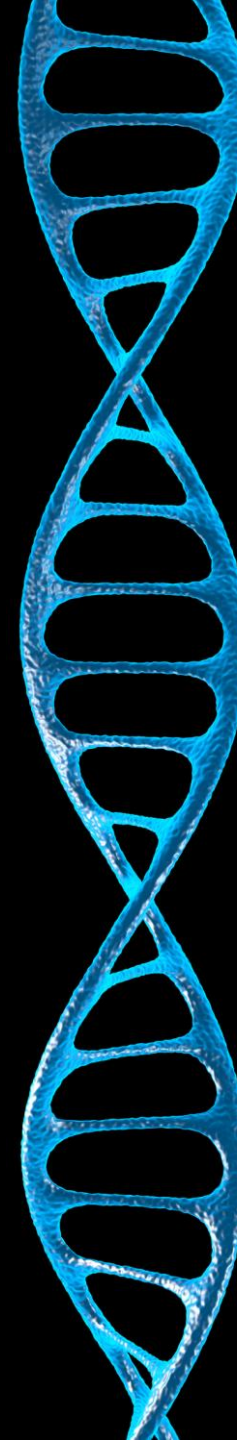


Statistical Genetics 101

DNA basics
Heritability
Gene discovery
Polygenic prediction

Aysu Okbay



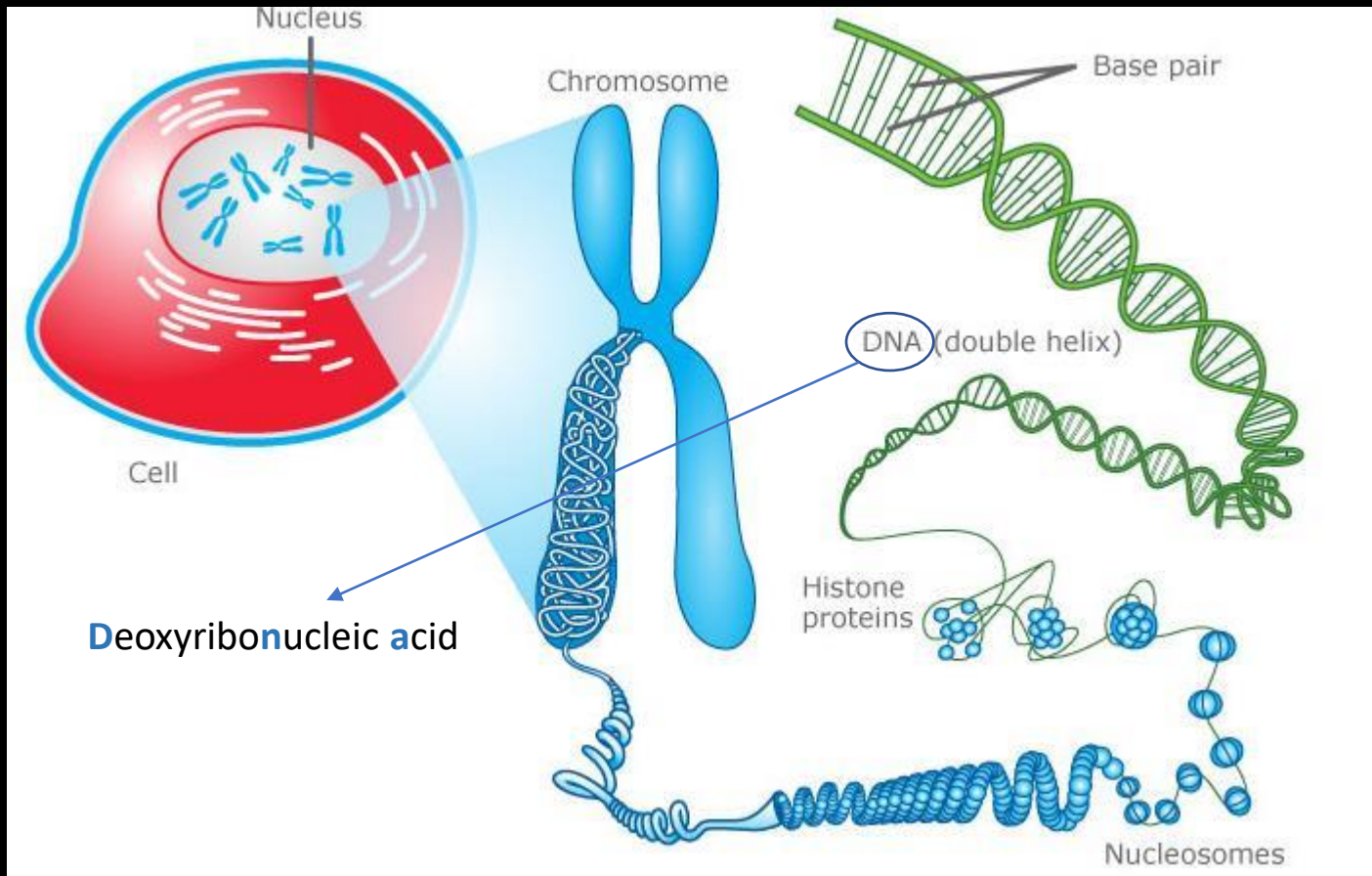
Part A - DNA Basics

Structure of DNA

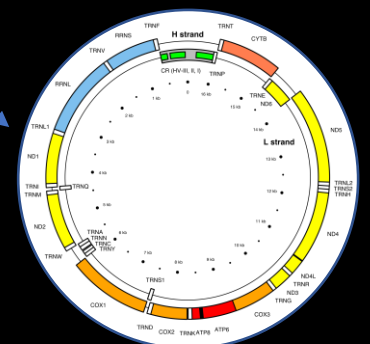
Genetic inheritance

Linkage disequilibrium

The Architecture of life



- Humans have ~40 trillion cells
- Each cell has a nucleus, containing DNA packed in
 - 23 pairs of chromosomes in somatic cells
 - 23 chromosomes in sex cells
- Some DNA exists in mitochondria in 1 small **circ**ular chromosome (maternally inherited)



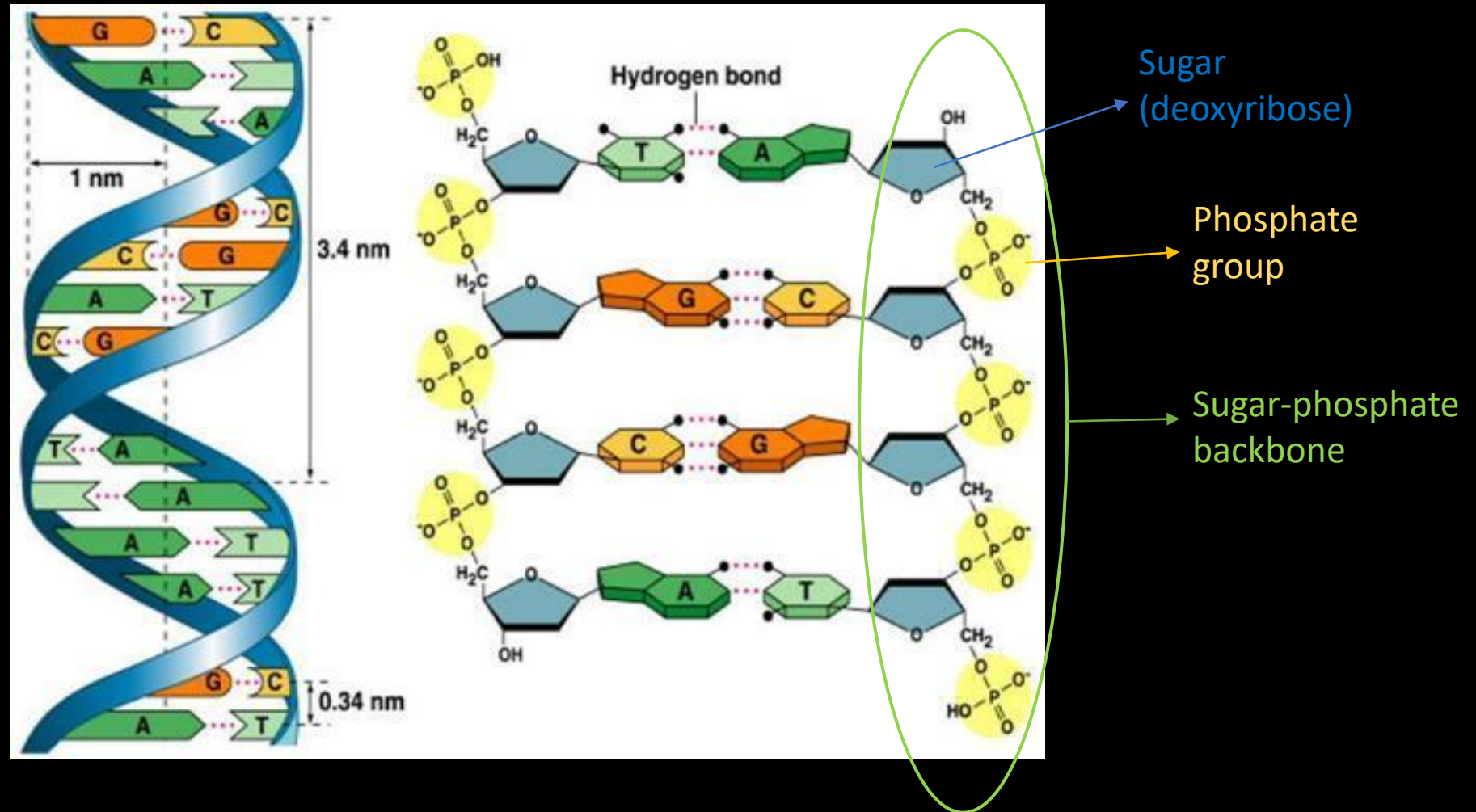
Basic Structure of DNA

The code of life is spelled with 4 “letters”:

- Adenine (A)
- Cytosine (C)
- Guanine (G)
- Thymine (T)

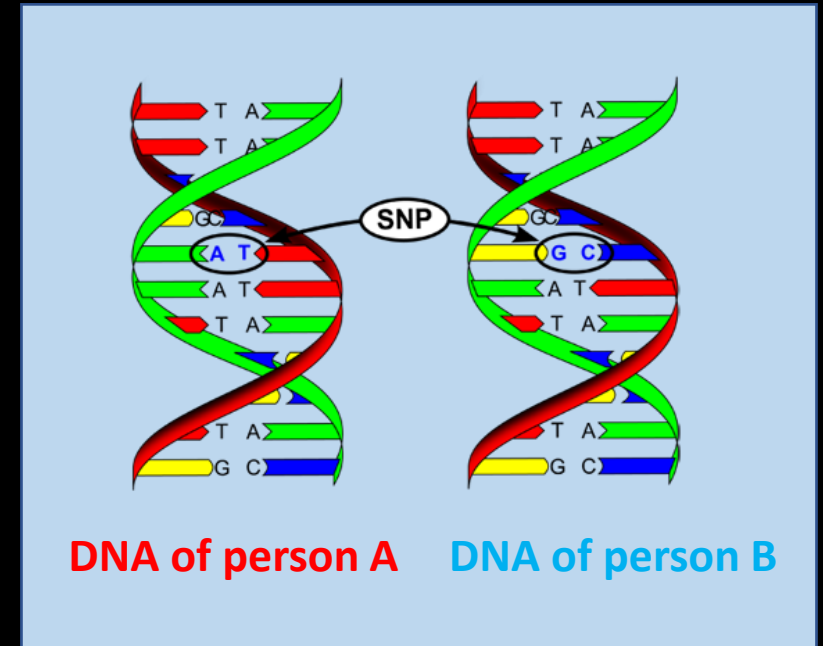
3.3 billion base pairs in the genome

- 1.8 meters long
- 3000 books of 500 pages each



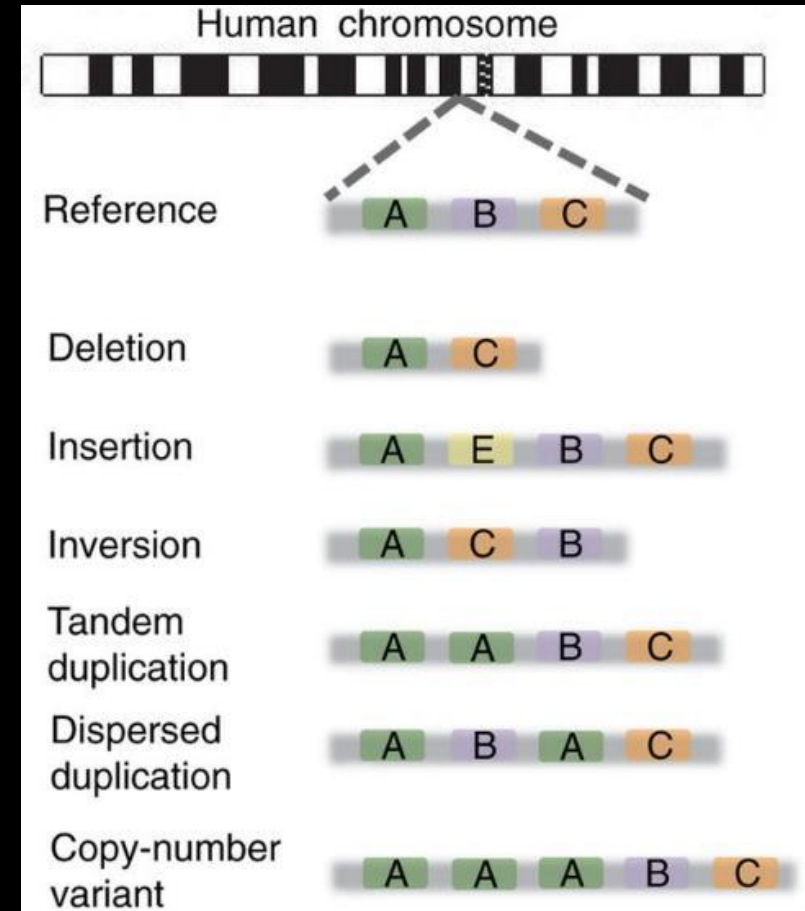
Genetic variants

- Any 2 humans share ~99.6% of DNA
 - 0.4% of genome varies between humans
- Single Nucleotide Polymorphism (**SNP**)
 - 1 base pair difference between individuals
 - Most common form of genetic variation
- Started as a “**point mutation**” in evolution
 - Became more frequent in population



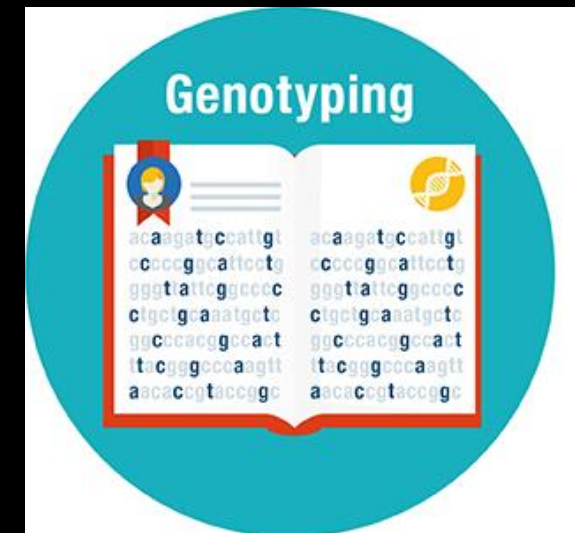
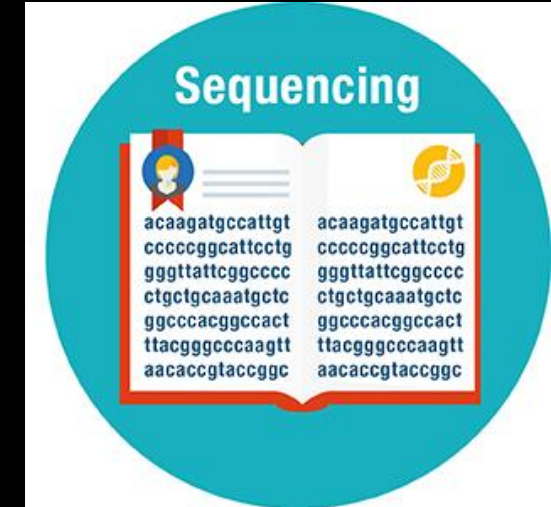
Structural variants

- Less common than SNPs
 - But they span more base-pairs in the genome than SNPs (1000 Genome consortium, *Nature* 2015)
- Common structural variants are correlated with SNPs

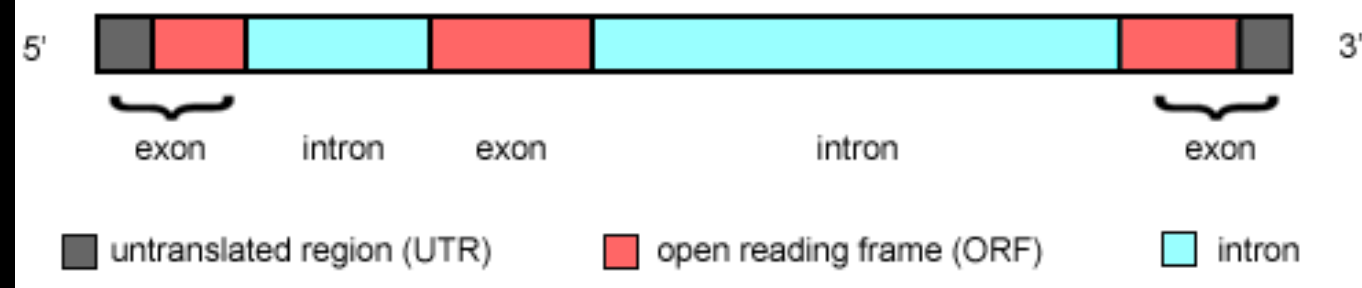


Measuring the genome

- Whole genome sequencing measures the full genome
 - 3.3 billion x 2 base pairs
 - Rare mutations
- **Genotyping** (“chip”/“array”) measures common genetic variants
 - 250,000 – 2 mln base pairs (x2)
 - Correlation structure (linkage disequilibrium) allows us to impute other SNPs



Genes

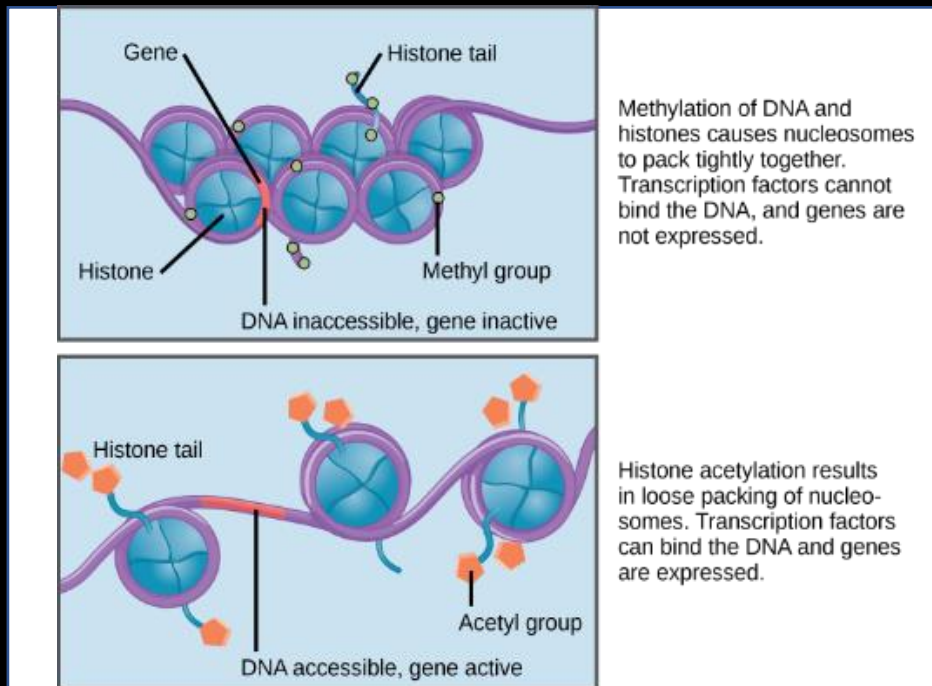


- A gene is a sequence of DNA that encodes one polypeptide (protein)
- Size: 1000 – 2mil bases
- Human genome
 - Only ~3% is genes
 - 21,000 protein-coding genes (a third expressed only in the brain)
- Genes make “functionally important” RNA molecule
 - RNA can be translated into amino acid sequence
 - ... or not: RNA can modify expression of other genes
- Traits & common diseases are mostly **not** the results of protein structure differences
 - But: differences in *abundance* of protein (gene expression)

Gene expression

DNA binds strongly to histones

- Chemical modification of histone “tightens” or “loosens” connection with DNA



- We can measure gene expression by measuring mRNA levels
 - RNA sequencing
 - Chips (microarrays)
- Expression level (how much mRNA there is) is highly variable across
 - time points
 - cell types
 - cell states (mediated by chromatin state, i.e. how tightly packed the DNA is, among other things)

Part A - DNA Basics

Structure of DNA

Genetic inheritance

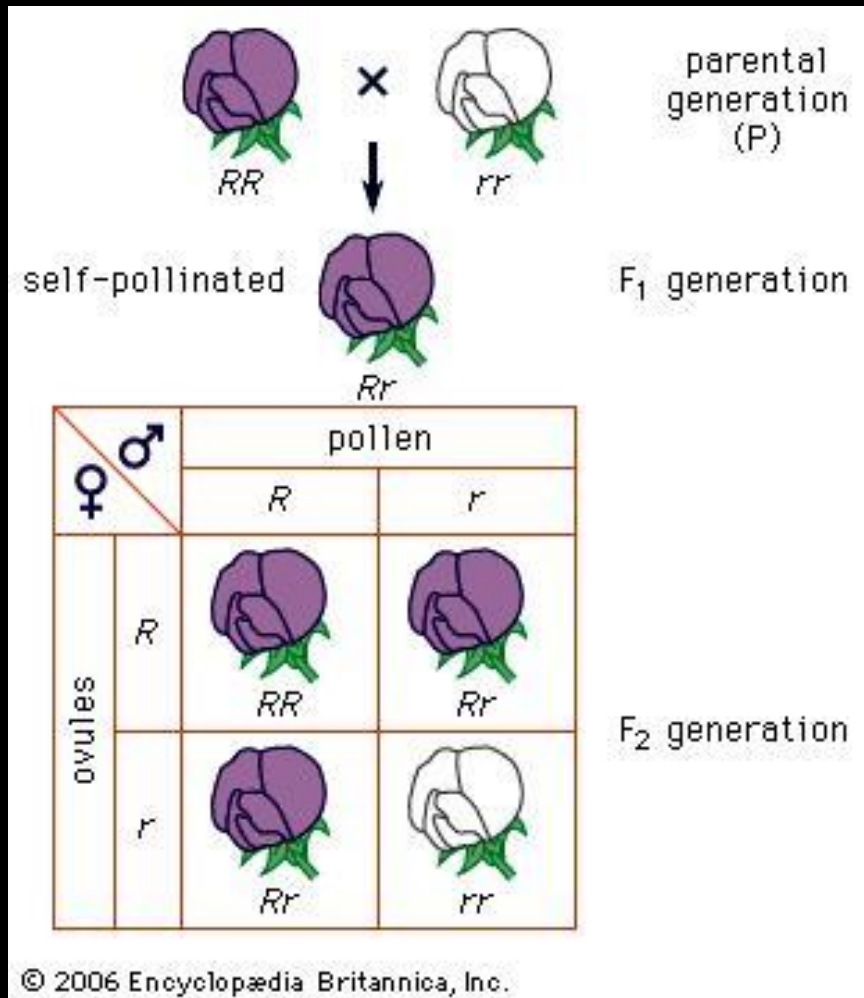
Linkage disequilibrium

Mendel's Laws



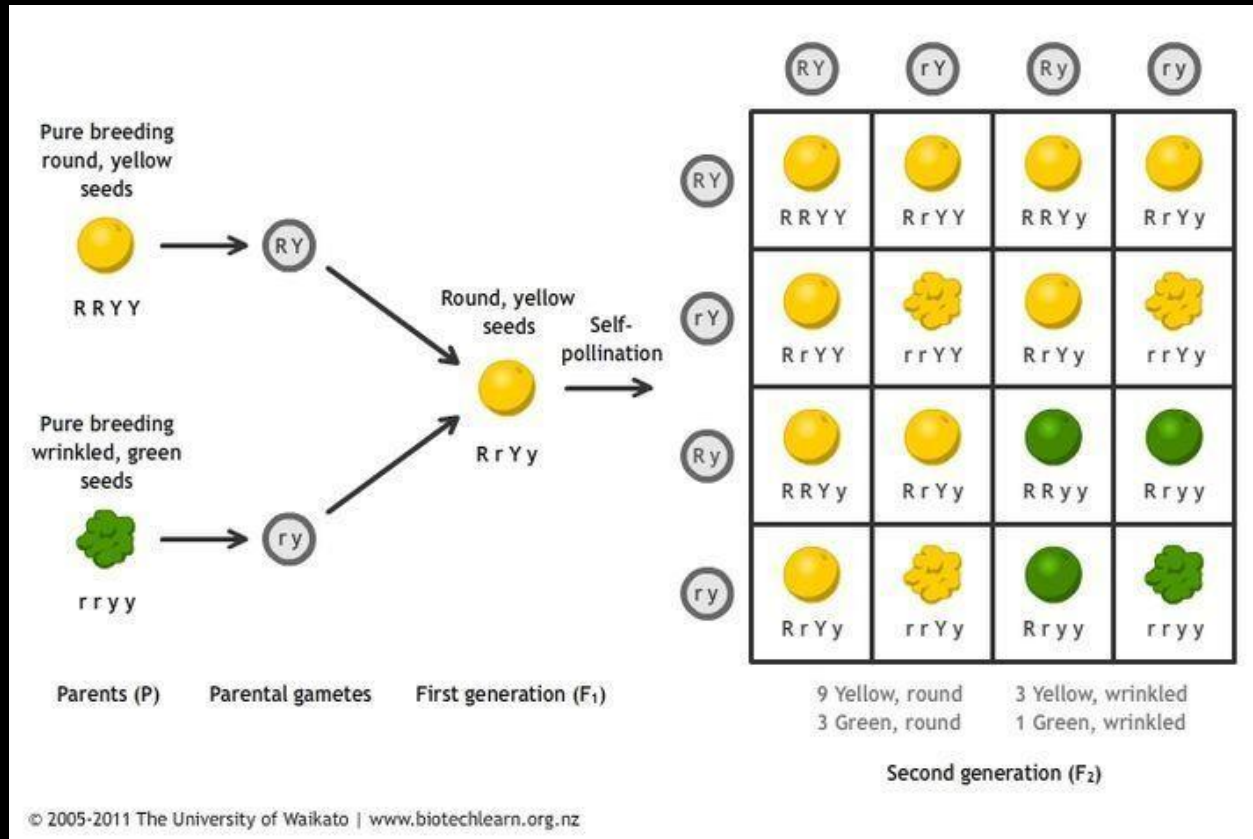
- Gregor Mendel (1822-1884)
 - Augustinian friar who studied inheritance in pea plants and discovered fundamental laws in genetics
- If these laws hold for a given trait, mode of inheritance & genotype probabilities can be derived from observing family characteristics

Law of dominance & Law of segregation



- **Law of dominance and uniformity:** One allele can dominate the expression of the other
 - **Dominant:** Only need 1 copy to have trait/disease
 - **Recessive:** Need 2 copies to have trait/disease
- **Law of segregation:**
 - A gene can exist in more than one form or allele.
 - Organisms inherit two alleles for each trait.
 - The two alleles, one from each parent, separate during gamete formation so that gamete carries only one.

Law of independent assortment



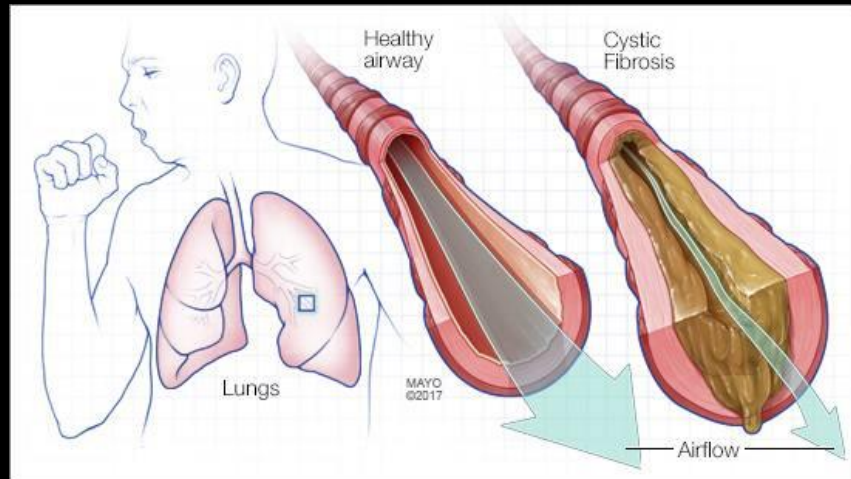
Genes for different traits segregate independently during the formation of gametes

→ does not hold if there is “linkage disequilibrium”!

Mendelian traits

Traits that are largely determined by **one** gene according to law of dominance

Severe diseases



Huntington, *BRCA*-linked breast cancer, cystic fibrosis, Rett's disease, sickle cell disease

Simple physical traits

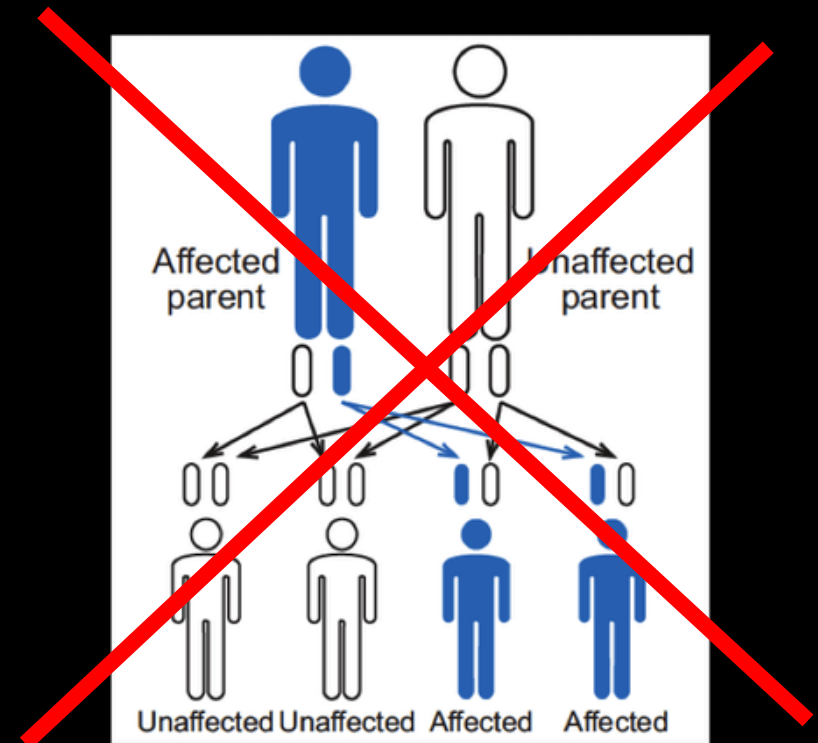
	AB	Ab	aB	ab
AB	AA BB	AA Bb	Aa BB	Aa Bb
Ab	AA Bb	AA bb	Aa Bb	Aa bb
aB	Aa BB	Aa Bb	aa BB	aa Bb
ab	Aa Bb	Aa bb	aa Bb	aa bb

Ear lobe attachment, eye color, cheek dimples, blood type

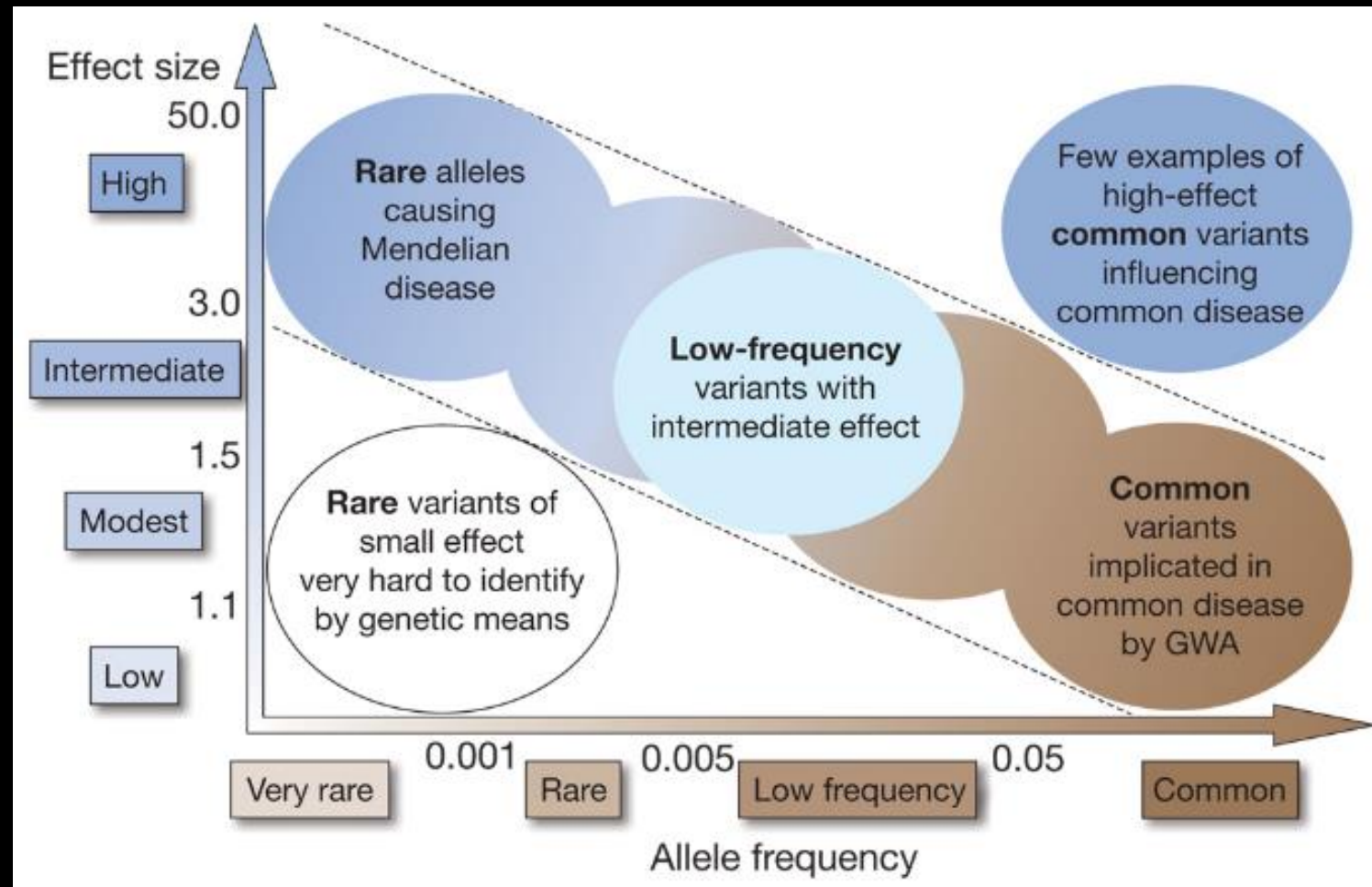
Polygenic traits

Most phenotypes do ***not*** follow simple Mendelian inheritance. Why?

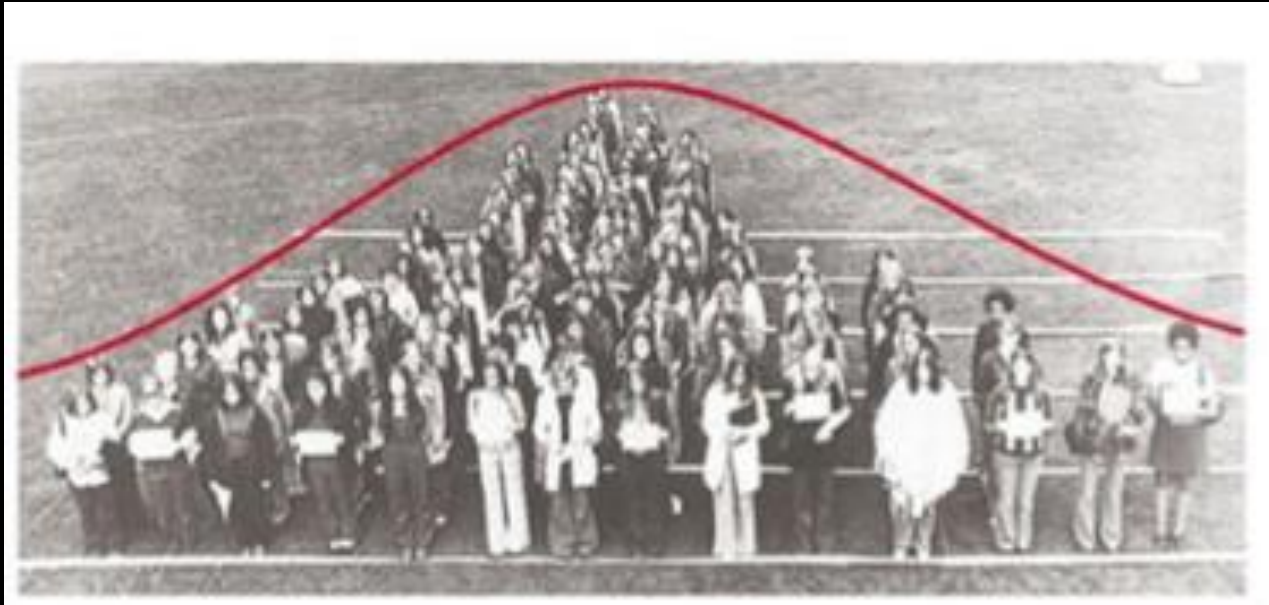
- Mutations with large effects are *usually* harmful
- Mutation/selection balance
 - Organisms with rare mutations don't tend to (live long enough to) procreate
 - But new *de novo* mutations in population keep disease prevalence up
- We call these “complex”, “polygenic”, or “quantitative” traits
 - Many genetic variants (1,000s) affect trait
 - Phenotypic resemblance increases with increasing degrees of genetic relatedness



Common disease, common variant hypothesis



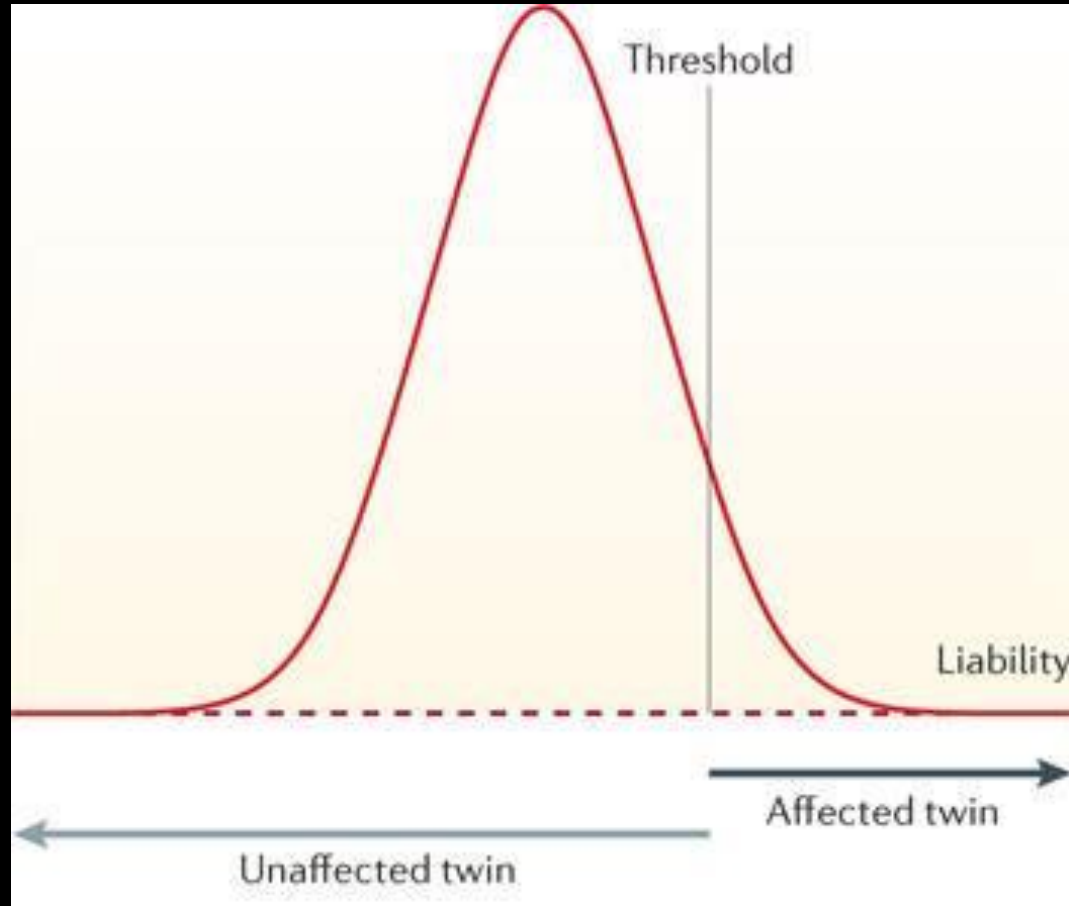
Polygenic inheritance – *Continuous traits*



Most genetic effects **additively** affect phenotype

Phenotype following polygenic inheritance will be **normally distributed** in a reasonably large sample

Polygenic inheritance - *Common diseases*



How do dichotomous diseases exist if many genes are involved? Two explanations:

- Liability-threshold model
 - Disease susceptibility as a continuous measure
 - Underlying additive genetic factors
- Dichotomy is artificial

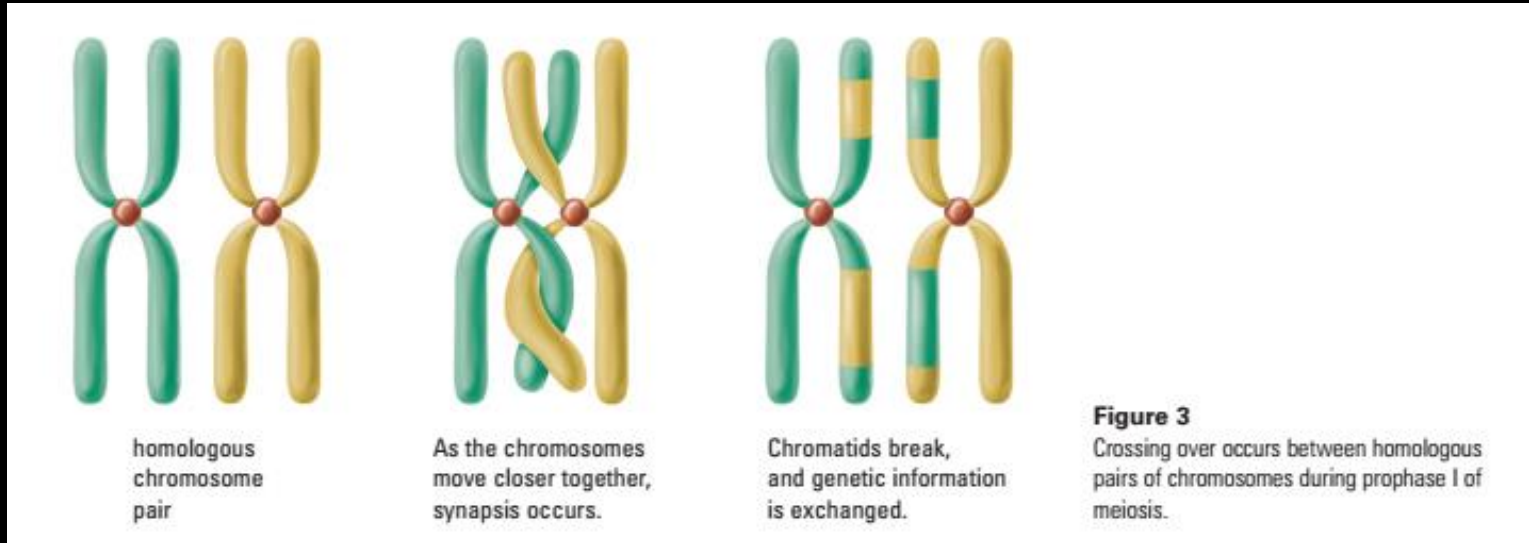
Part A - DNA Basics

Structure of DNA

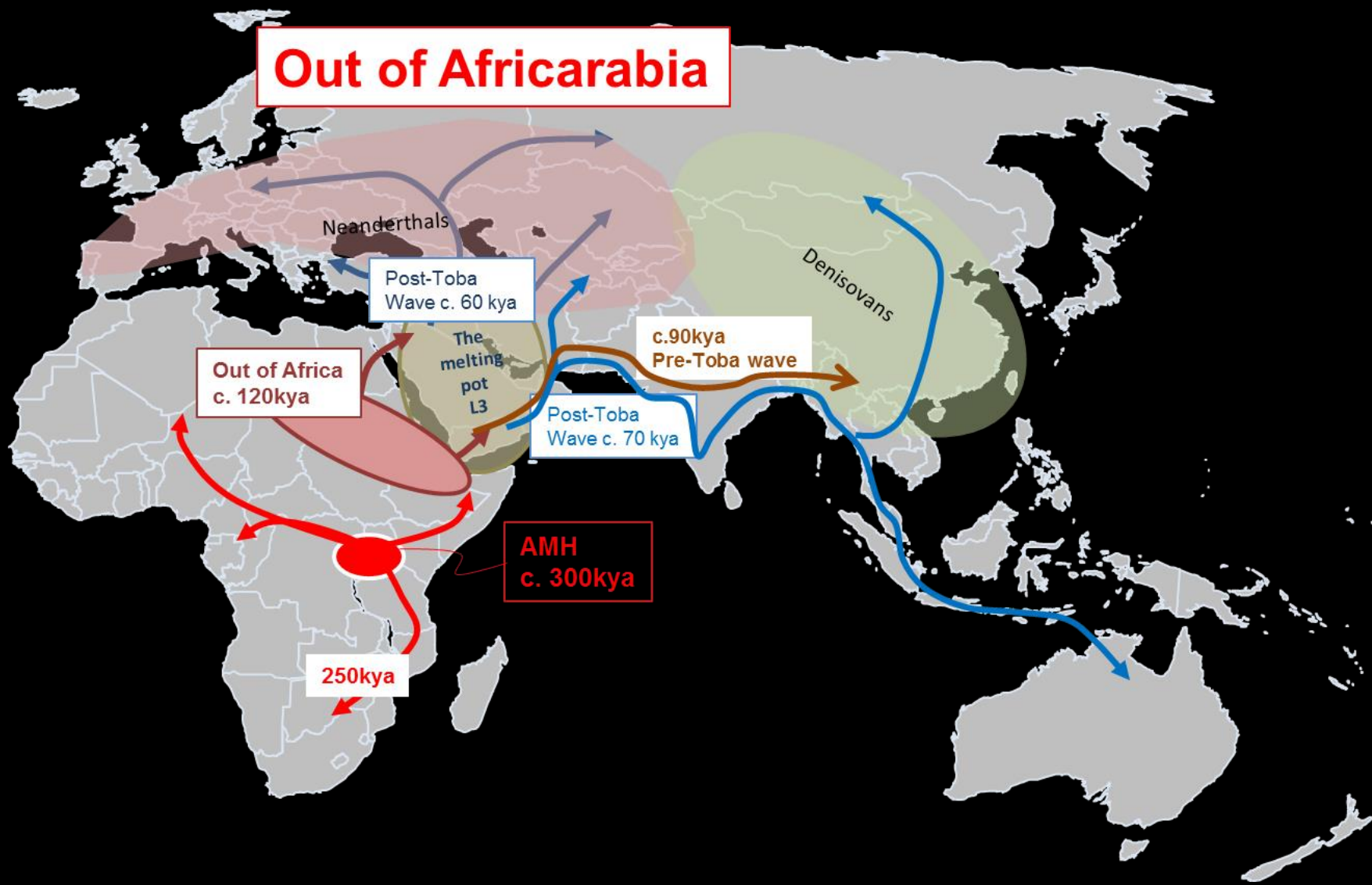
Genetic inheritance

Linkage disequilibrium

Recombination during meiosis



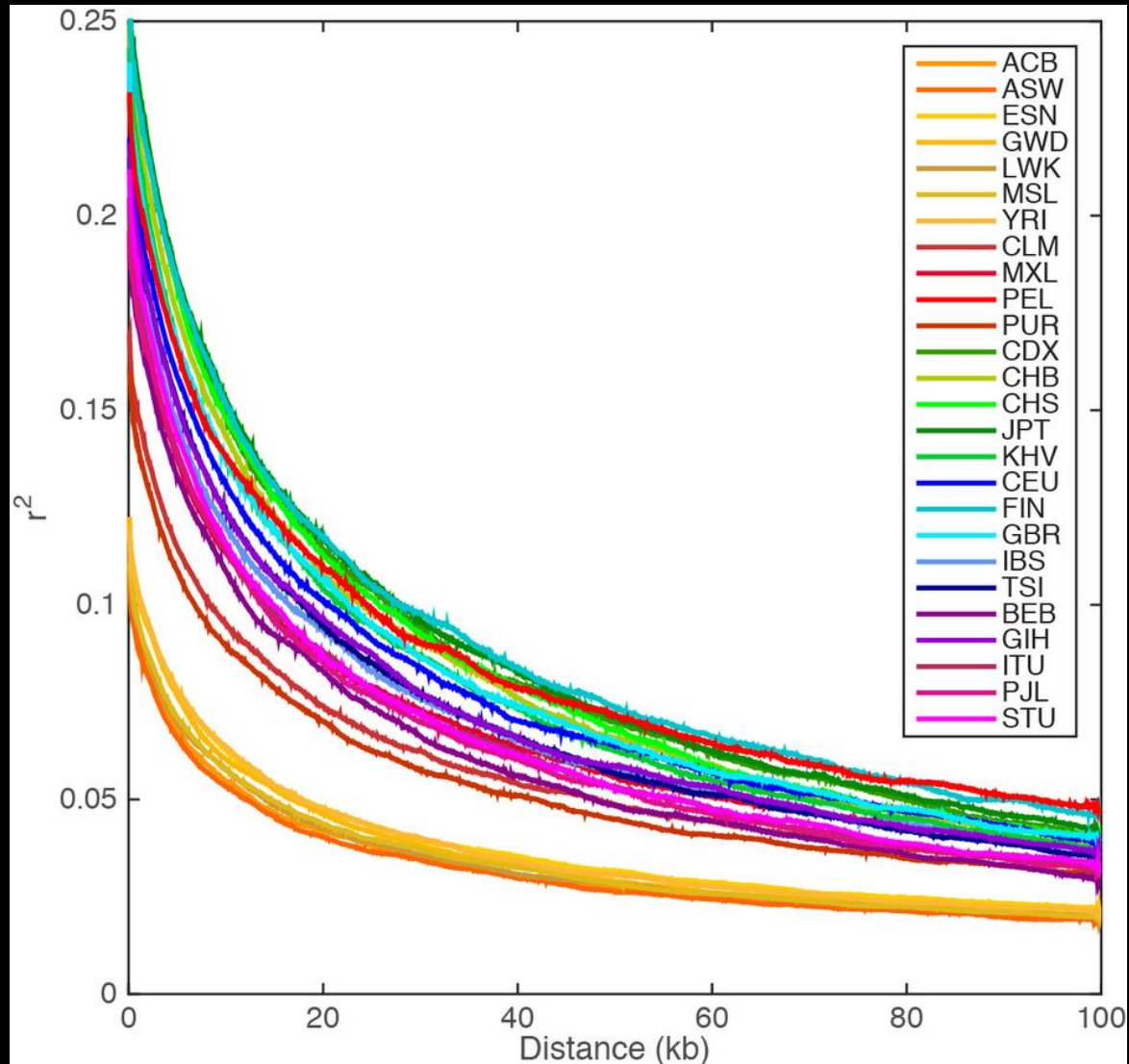
- Happens in ovaries/testes during meiosis I
 - Each chromosome duplicates to form a sister chromatid
 - Recombination across homologous chromosomes occurs (on average one time for each chromosome)
 - Then each chromatid is transmitted to a different gamete in meiosis II.
- THE reason sexual reproduction produces such variable offspring (2^{23} possible combinations even without recombination)
 - Probability of recombination is a function of distance



- LD structure varies based on population, specifically,
- The size and age of the population can shorten haplotype blocks
 - Recombination “hot spots” and patterns can vary by population

European genome contains ~1 mln independent haplotype blocks

LD as a function of physical distance



- Decay is fastest for African, slowest for East Asian populations
- Each common variant has over 15-20 good tagging variants ($R^2 > 0.8$) in non-African populations, but only about 8 in African populations
- For lower MAF, differences are less marked

1000 Genomes Project (2015, *Nature*)
Extended Data Figure 10

Part B - Heritability

What is heritability?

Estimating heritability with twin studies

Estimating heritability with molecular genetic data

Meaning of heritability

The proportion of observed differences in a trait among individuals of a population that is due to genetic differences among these individuals

- Heritability estimates
 - are not informative about the molecular genetic architecture of a trait
 - can vary across environments / populations
 - are population parameters and have no direct translation for individuals
- Traits can be heritable without being hardwired
- Heritable means “pre-wired” (flexible and subject to change) rather than “hard-wired” (fixed and immutable)
- A trait that does not vary in a population may be *inherited* (e.g. having two legs), but it is not *heritable*

True/False?

- If h^2 is high, differences between groups are due to genetic differences.
 - Heritability of obtaining a College degree
 - “Given the environmentalists’ extraordinary ingenuity in explaining away evidence that genes are important in the first place, their failure to argue that genes may exert their impact largely through the environment is puzzling.” (Jencks, 1980, p. 730)
- Hans Eysenck once remarked upon learning that income was moderately heritable that the British Commission for the Distribution of Income might as well “pack up”.
 - “...if it were shown that a large proportion of the variance in eyesight were due to genetic causes, then the Royal Commission on the Distribution of Eyeglasses might as well pack up.” (Goldberger 1979, p. 337)
 - Heritability may be *induced* or *reduced* by specific features of the environment
 - Even if h^2 is very high, this does not mean policy making is irrelevant or may not have profound effects on the outcomes
- If h^2 is high, there can still be rapid changes in the mean of the trait over time.
 - Height & nutrition in Europe 1900 – 2000
- High h^2 means there are a few genes with large effects.
- Low h^2 means there are no genes with large effects.

Narrow vs broad-sense heritability

Narrow-sense heritability:

- Accounts for additive genetic effects only

Broad-sense heritability:

- Includes additive effects
- Includes dominance
- Includes epistasis (interactions of alleles at different loci, e.g. gene*gene)

- Most models estimate narrow-sense heritability, including all classic twin studies and GREML.

Part B - Heritability

What is heritability?

Estimating heritability with twin studies

Estimating heritability using molecular genetic data

Estimating heritability – twin studies

- **Standard case:** DZ and MZ twins reared together
- **Assumptions:**
 - Genetic correlation for MZ twins is 1
 - Genetic correlation for DZ twins is 0.5
 - MZ and DZ twins share their environment to the same extent (or all the additional similarity in the environments of MZ twins is due to their genetic similarity)
 - Random mating (no assortative mating on the trait in question)
 - Genetic effects are additive
- Heritability is approximately twice the difference between MZ and DZ correlations of the trait:

$$\hat{h}^2 = 2 \times (r_{MZ} - r_{DZ})$$

Estimating heritability – twin studies

More formally:

Variance components

$$\text{Var}(y) = \text{Var}(g) + \text{Var}(c) + \text{Var}(e)$$

y – phenotype, g – genotype, c – common environment, e – unique environment

$$\text{Var}(g) = \text{Var}(a) \longrightarrow \text{Assuming genetic effects are additive}$$

a – additive

Narrow-sense heritability $h^2 = \text{Var}(a) / \text{Var}(y)$

• Correlation between MZ twins is $r_{MZ} = h^2 + c^2$

• Correlation between DZ twins is $r_{DZ} = \frac{h^2}{2} + c^2$

shared additive genetic component

shared environment component

Estimating heritability – twin studies

- Subtracting second equation from first:

$$r_{MZ} - r_{DZ} = h^2 - \frac{h^2}{2} + c^2 - c^2 = \frac{h^2}{2}$$

$$h^2 = 2(r_{MZ} - r_{DZ})$$

- Shared environmental effects:

$$c^2 = r_{MZ} - h^2$$

- Non-shared environment:

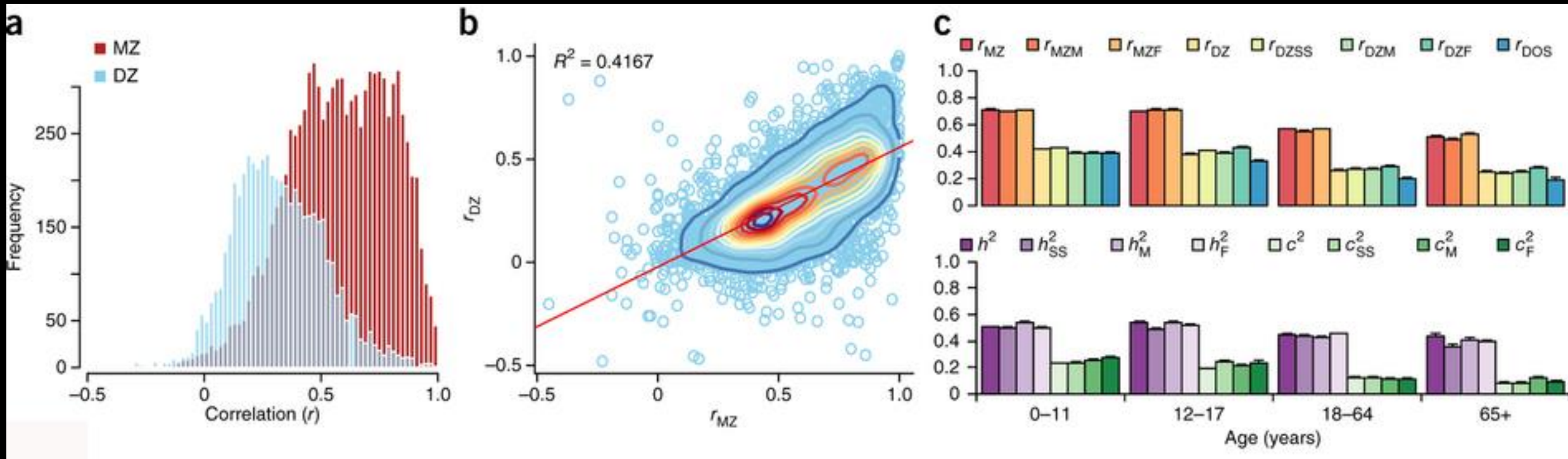
$$1 = h^2 + c^2 + e^2$$

$$1 = [2(r_{MZ} - r_{DZ})] + [r_{MZ} - 2(r_{MZ} - r_{DZ})] + e^2$$

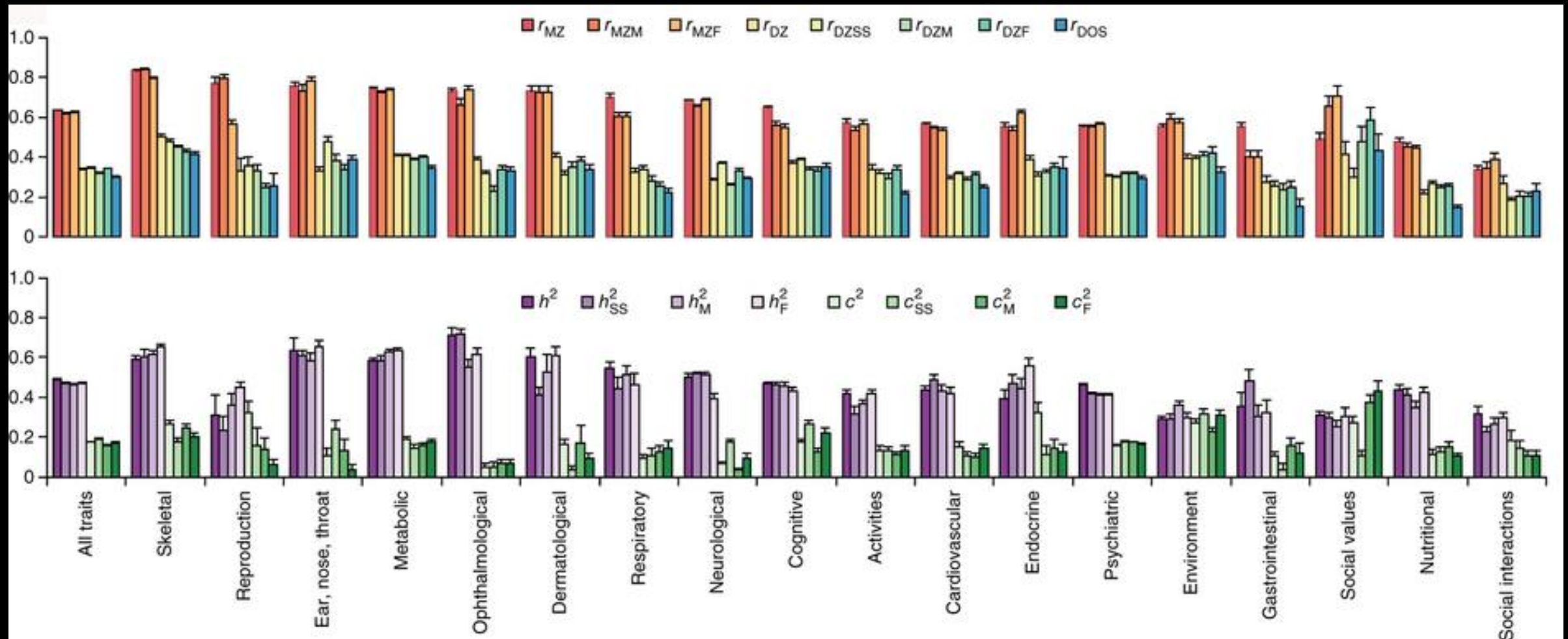
$$1 = r_{MZ} + e^2$$

$$e^2 = 1 - r_{MZ}$$

Estimating heritability – twin studies



Estimating heritability – twin studies



Part B - Heritability

What is heritability?

Estimating heritability with twin studies

Estimating heritability using molecular genetic data

Estimating heritability – GREML

- Heritability can also be estimated from molecular genetic data in population samples
 - Yang et al. (2010), *Nature Genetics*, doi:10.1038/ng.608
 - **GREML**: Genomic-relatedness-matrix **R**estricted **M**aximum **L**ikelihood
- **Assumptions:**
 - Individuals are comprehensively and accurately genotyped
 - No relationship between genetic similarity and shared environment (i.e. exclude closely related individuals)
 - Genetic effects are additive and infinitesimal
 - Effect sizes are inversely proportional to MAF and independent from LD-score of causal variants

Estimating heritability – GREML

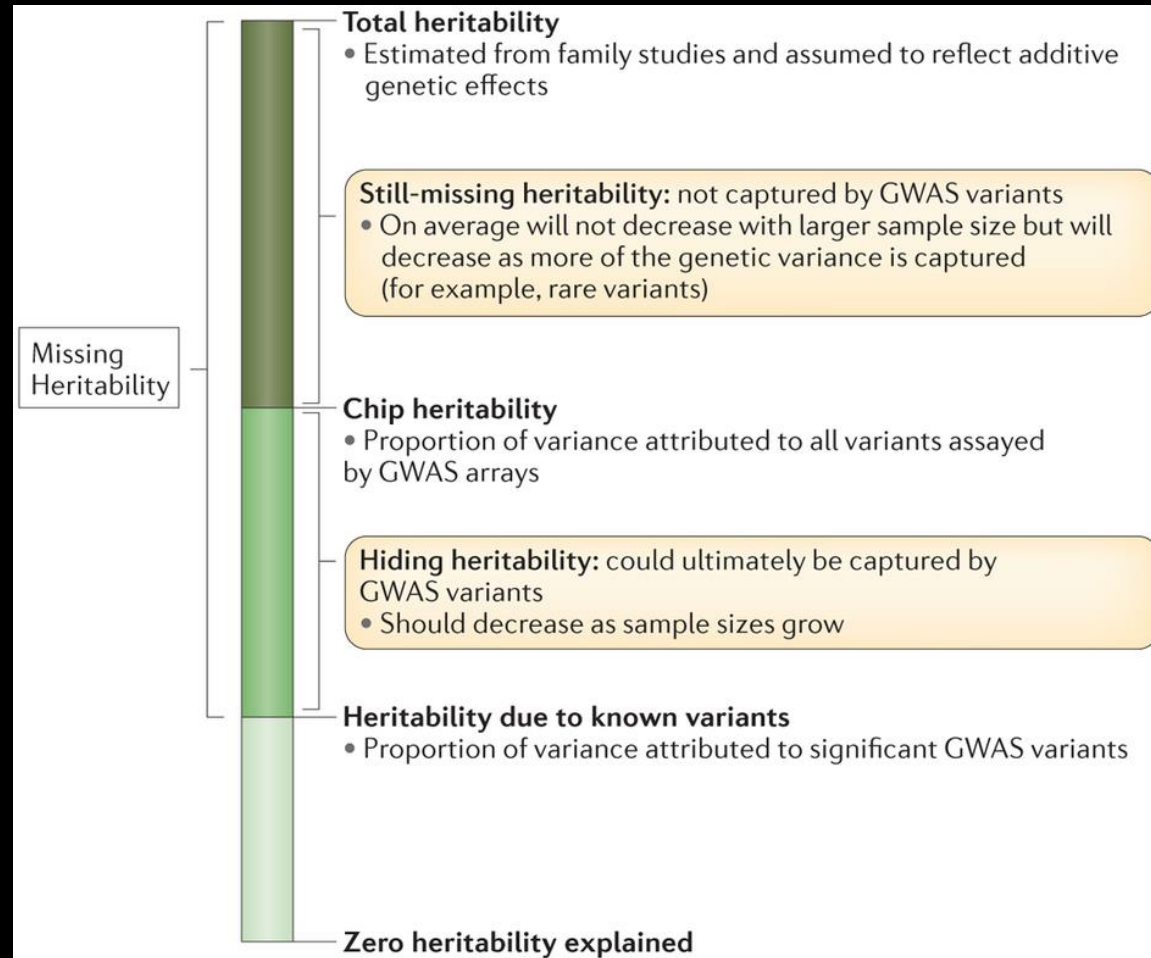
- Steps:

- Estimate pairwise genetic relatedness among individuals
- Exclude individuals who are more similar genetically than expected by chance
 - i.e., genetic similarity > 0.025 (\sim second cousins or closer)
- Examine whether individuals who are more closely related have more similar phenotypes

- Interpretation:

- Resulting estimate can be interpreted as proportion of variance accounted for by linear effects of the genotyped markers
 - A.k.a. “narrow-sense SNP-based heritability” (h^2_{SNP})
- Defines the upper bound of the predictive accuracy that a polygenic score constructed from those markers could have

“Missing heritability”



Source: Witte et al., 2014. The contribution of genetic variants to disease depends on the ruler. *Nature Reviews Genetics* 15, 765-776.

Part C- Genetic discovery

Candidate gene studies

GWAS

Imputation

Meta-analysis

QC of GWAS summary statistics

Life after GWAS

Genetic discovery

- Suppose you want to analyze the genetic influence on some outcome y_i

$$y_i = \mu + \sum_{j=1}^J \beta_j x_{ij} + \mathbf{Z}_i \boldsymbol{\gamma} + \varepsilon_i$$

μ : the mean value of y in the population

β_j : the effect of SNP j

x_{ij} : genotype of individual i at SNP j

\mathbf{Z}_i : vector of covariates for individual i

$\boldsymbol{\gamma}$: vector of covariate effects

ε_i : effect of exogenous residual factors

- If $\beta_j \neq 0$, we call SNP j “causal”

Candidate gene studies – 1

- Choose x_{ij} based on theory or prior biological insights
- Set significance threshold $\alpha = .05 / J$
- **Advantages:**
 - Keep J low
 - Eminently reasonable, and has worked when hypotheses are direct (e.g., APOE and Alzheimer's)
- **Problems:**
 - Theory on biological mechanisms for behavior is often weak
 - >14,000 genes expressed in the brain
 - We don't know yet what all of them are doing
 - But it's easy to post-rationalize empirical results
 - Difficult to control for population stratification

Candidate gene studies – 2

- **Problems (cont'd):**

- Typical candidate gene studies assume (implicitly) large effect sizes of genetic variants
 - This assumption is false for genetically complex (non-Mendelian) traits
 - As a result of wrong assumptions about plausible effect sizes, many candidate gene studies were underpowered (small N)

Also, in the age of cheap genome-wide data, it's difficult to justify to look at only a few of them.

Replication attempts of candidate gene studies

Behav Genet (2012) 42:1–2
DOI 10.1007/s10519-011-9504-z

BRIEF COMMUNICATION

Editorial Policy on Candidate Gene Association and Candidate Gene-by-Environment Interaction Studies of Complex Traits

John K. Hewitt

“The literature on candidate gene associations is full of reports that have not stood up to rigorous replication... As a result, the psychiatric and behavior genetics literature has become confusing and it now seems likely that many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge.”

The literature on candidate gene associations is full of reports that have not stood up to rigorous replication. This is the case both for straightforward main effects and for candidate gene-by-environment interactions (Duncan and Keller 2011). As a result, the psychiatric and behavior genetics literature has become confusing and it now seems likely that many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge. The reasons for this are complex, but include the likelihood that effect sizes of individual polymorphisms are small, that studies have therefore been underpowered, and that multiple hypotheses and methods of analysis have been explored; these conditions will result in an unacceptably high proportion of false findings (Ioannidis 2005).

studies of complex traits, especially when reporting complex interaction effects based on novel phenotypes and groupings. Of course, we understand that this has not been done routinely—sometimes it is not practical—and so authors have preferred to publish the initial paper without such replication. We also recognize that there are historical examples where early failures to replicate were themselves misleading because of heterogeneity or poor methodology.

However, for a candidate gene or candidate gene-by-environment interaction study of a complex trait to be considered for publication in *Behavior Genetics* it should usually have one or more of the following characteristics:

- It is a rigorously conducted, adequately powered, direct replication study of a previously reported result; for well-conducted replication studies, there is no editorial

Part C- Genetic discovery

Candidate gene studies

GWAS

Imputation

Meta-analysis

QC of GWAS summary statistics

Life after GWAS

Genome-wide association studies (GWAS)

Hypothesis-free scan of all J SNPs

For **SNP** $j = 1, \dots, M$:

1. Estimate $y_i = \mu + \beta_j x_{ij} + \mathbf{Z}_i \gamma + \varepsilon_i$
using ordinary least squares (OLS) or something more sophisticated
2. Store the relevant ‘stuff’, typically including at least:
 - $\hat{\beta}_j$: estimated effect of **SNP** j
 - $\text{SE}(\hat{\beta}_j)$: the standard error (SE) of the estimated SNP effect
 - P_j : the two-sided p -value for the t -statistic for the null hypothesis that $\beta_j = 0$

→ All done ‘for you’ using command-line tools such as PLINK

Multiple hypothesis testing – 1/2

GWAS: scans many SNPs for association with Y

- we are testing many independent hypotheses!

Conventional significance threshold $\alpha = 5\%$?

- if null hypothesis (H_0 : given SNP has no effect) is true?
 - 5% chance to falsely reject in each test
 - for one million independent, non-associated SNPs:
 - we expect $0.05 \times 10^6 = 50,000$ false positives!... Unacceptable!
 - and chance of observing no false positive is basically zero



Simple solution? Bonferroni correction!

- Set $\alpha^* = \alpha / (\# \text{ independent tests})$ and use that as significance level

Multiple hypothesis testing – 2/2

What does a Bonferroni correction accomplish?

- it sets a very high bar: but that's good!
- chance of finding one (or more) false positives? $\approx \alpha$
- chance of finding no false positive at all? $\approx 1 - \alpha$

For GWAS using common SNPs for Europeans:

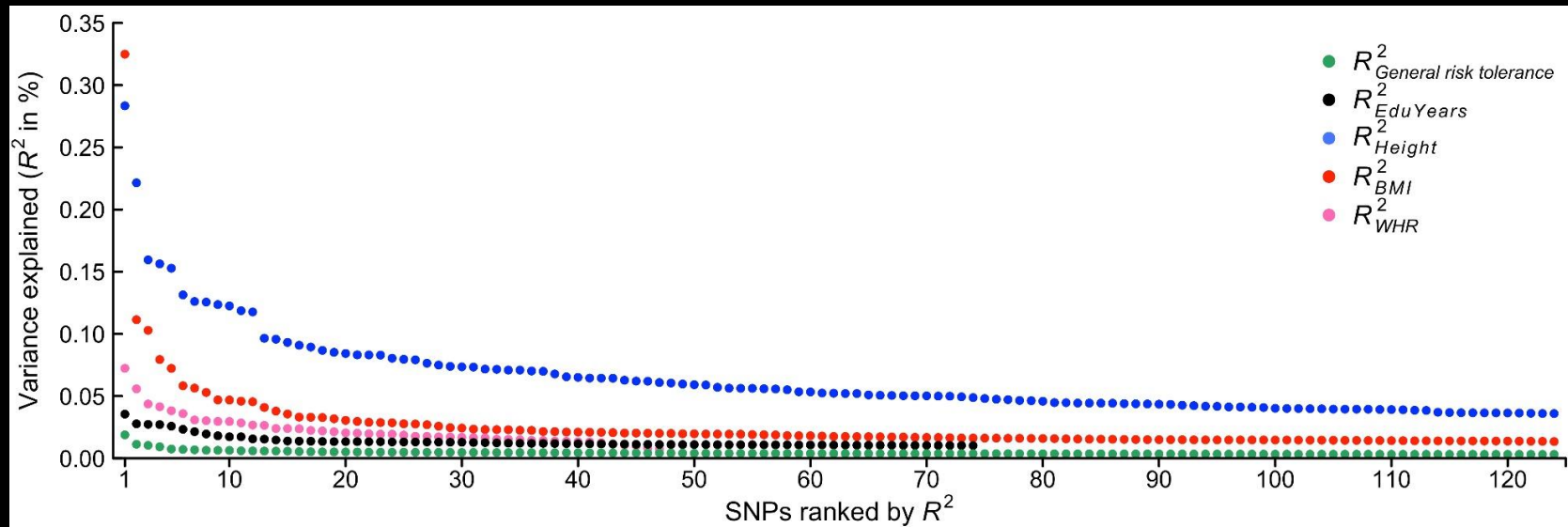
- # independent tests ≈ 1 million
- So $\alpha = 0.05$ yields $\alpha^* = 5 \times 10^{-8} = 0.000005\%$

Only if $\hat{\beta}_j$ has P -value below α^* do we say

- “SNP j is *genome-wide significant*”

Challenges

- Many traits influenced by ‘thousands’ of SNPs with small effects

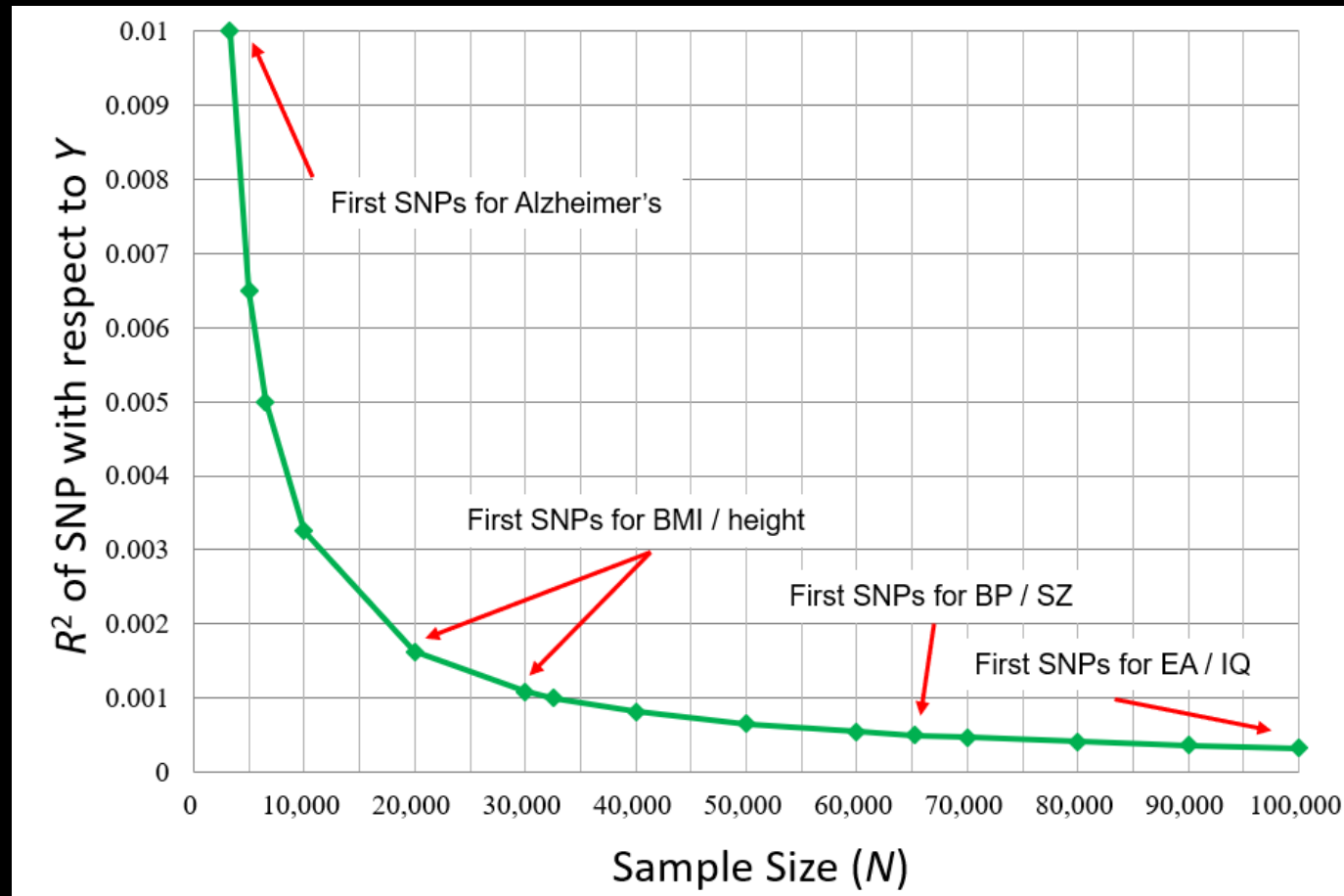


Effect sizes genome-wide significant SNPs in terms of R^2 w.r.t. various traits. SNP with lowest p -value for each approximately independent locus is displayed. Effect sizes are subject to the statistical winner's curse (i.e. true effects are likely even smaller). Source: Linnér et al. 2018,

<https://www.biorxiv.org/content/early/2018/02/08/261081>

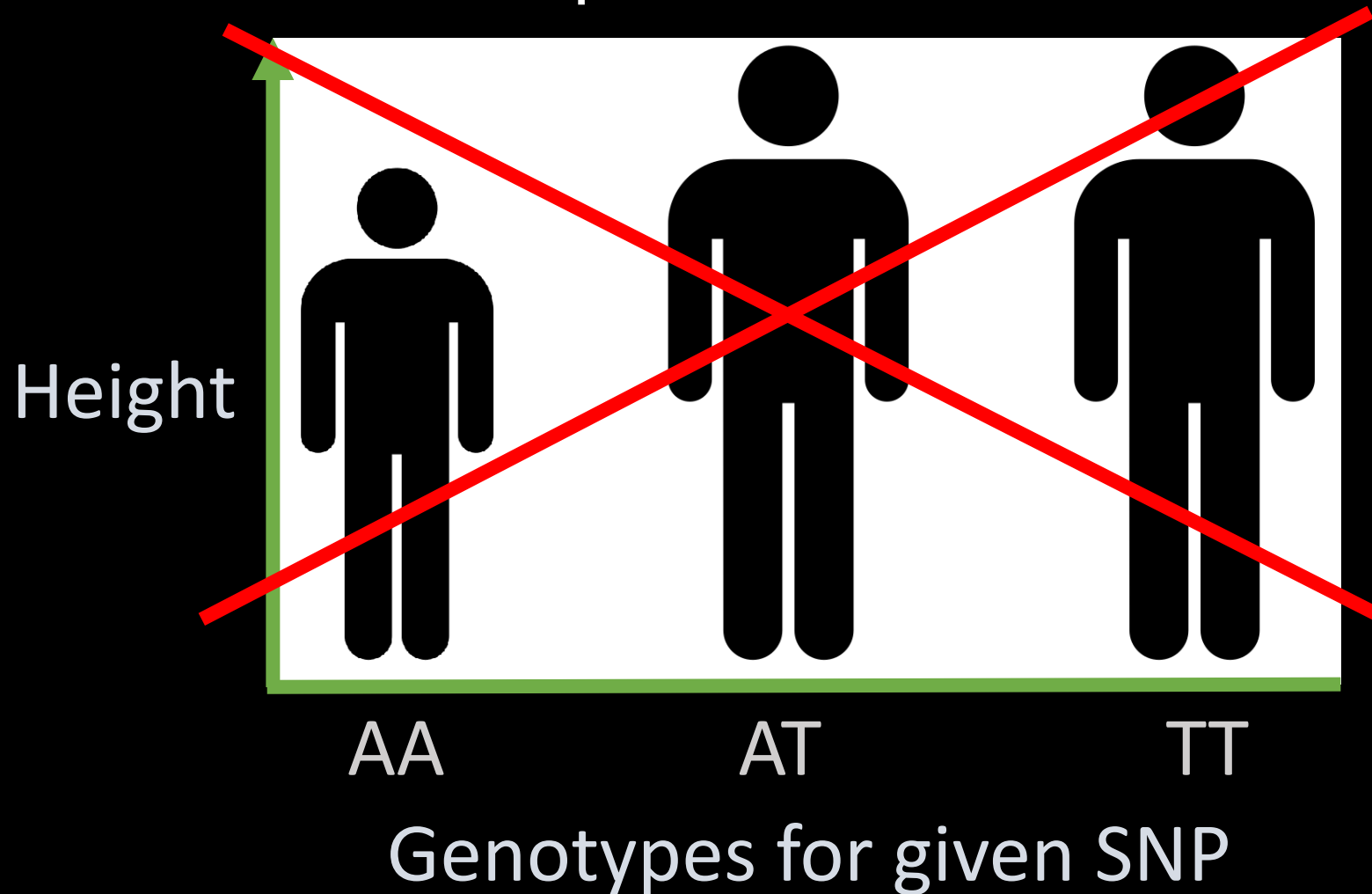
- Very large sample sizes are needed!

Putting power in context



Required R^2 (y-axis) between SNP and Y to have 50% power for given N (x-axis) and $\alpha = 5 \times 10^{-8}$

GWAS Assumptions – Linearity



Little evidence for pervasive dominance effects for complex traits.

So linear SNP effects are reasonable starting point!

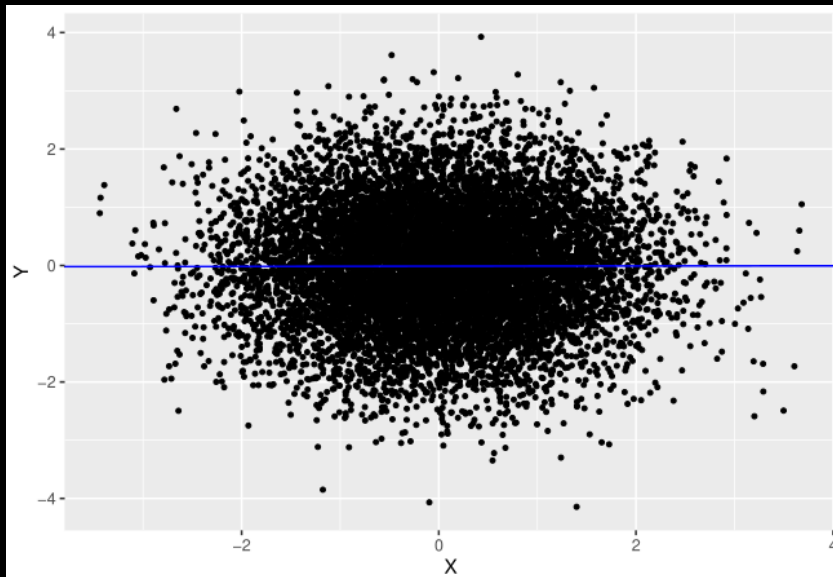
(GWAS including dominance effects is possible, but beyond this lecture)

GWAS Assumptions – Random sampling

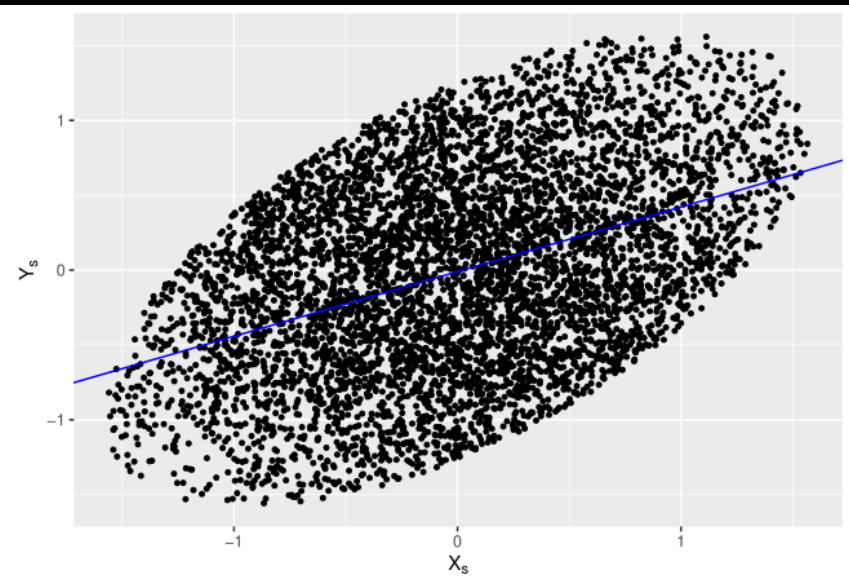
1. Non-representative samples (e.g. case–control cohorts)

- $\hat{\beta}_j$ can reflect spurious associations

Some simulation with random sampling:



same simulation, taking non-random subsample:



GWAS Assumptions – Random sampling

2. Observations are not independent

- Family samples or cryptic relatedness in the sample
- $\hat{\beta}_j$ not necessarily biased, but...
- $SE(\hat{\beta}_j)$ tends to be too low:
 - you have fewer ‘independent’ pieces of info than OLS thinks, increasing number of false positives!
- Solutions:
 - Exclude related individuals
 - Better: Use models that adjust for relatedness (e.g. mixed linear models)

GWAS Assumptions – Zero conditional mean

Error term not correlated with SNP (and covariates)

- Very intricate problem!
- Various things can go wrong

First, the obvious bit:

1. 'omitted-SNP bias'

- SNP j is correlated to nearby SNP h
 - because the two are in close proximity: linkage disequilibrium (LD)
- SNP h has causal effect on Y
- Yet, while estimating β_j we do not control for SNP h
- So (roughly speaking) $E[\hat{\beta}_j] \approx \beta_j + \rho_{jh}\beta_h + \text{other terms}$
where ρ_{jh} denotes the correlation between the two SNPs

GWAS Assumptions – Zero conditional mean

- $E[\hat{\beta}_j] = \beta_j + \rho_{jh}\beta_h + \text{other terms}$: estimate picks up effects ALL correlated SNPs
- GWAS needs and embraces this!
 - Ideally: fit all SNPs jointly, but... $M \gg N \rightarrow$ massive collinearity
 - So 'naïve' approach of GWAS: 1 regression per SNP
 - And embrace the fact that the estimated SNP effect 'picks up' on...
 - effects of many nearby variants that the SNP is correlated with!
- Embrace it how?
 - By acknowledging that a significant SNP only is a pointer to a region!
 - Applying tools that utilize expected correlational patterns between $\hat{\beta}_j$'s

In short: a problem that isn't really a problem, rather feature of GWAS results

GWAS Assumptions – Zero conditional mean

2. Other omitted variables, e.g. some types of gene-environment correlation

- *population stratification*

- A systematic difference in allele frequencies between (sub)populations due to different ancestry.
- Can cause false positives if the trait values also differ between the (sub)populations.

- *genetic nurture*

- My genes are correlated to parental genes
- Parental genes partially shaped my environment
- Environment partially shaped my outcomes
- Ideally: control for parental genotypes, or similar strategies, involving family data

Population stratification example: Chopstick gene

Sample 1: Americans $\chi^2 = 0, p = 1$			
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2: Chinese $\chi^2 = 0, p = 1$			
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

And also a clear difference in the proportion of cases and controls

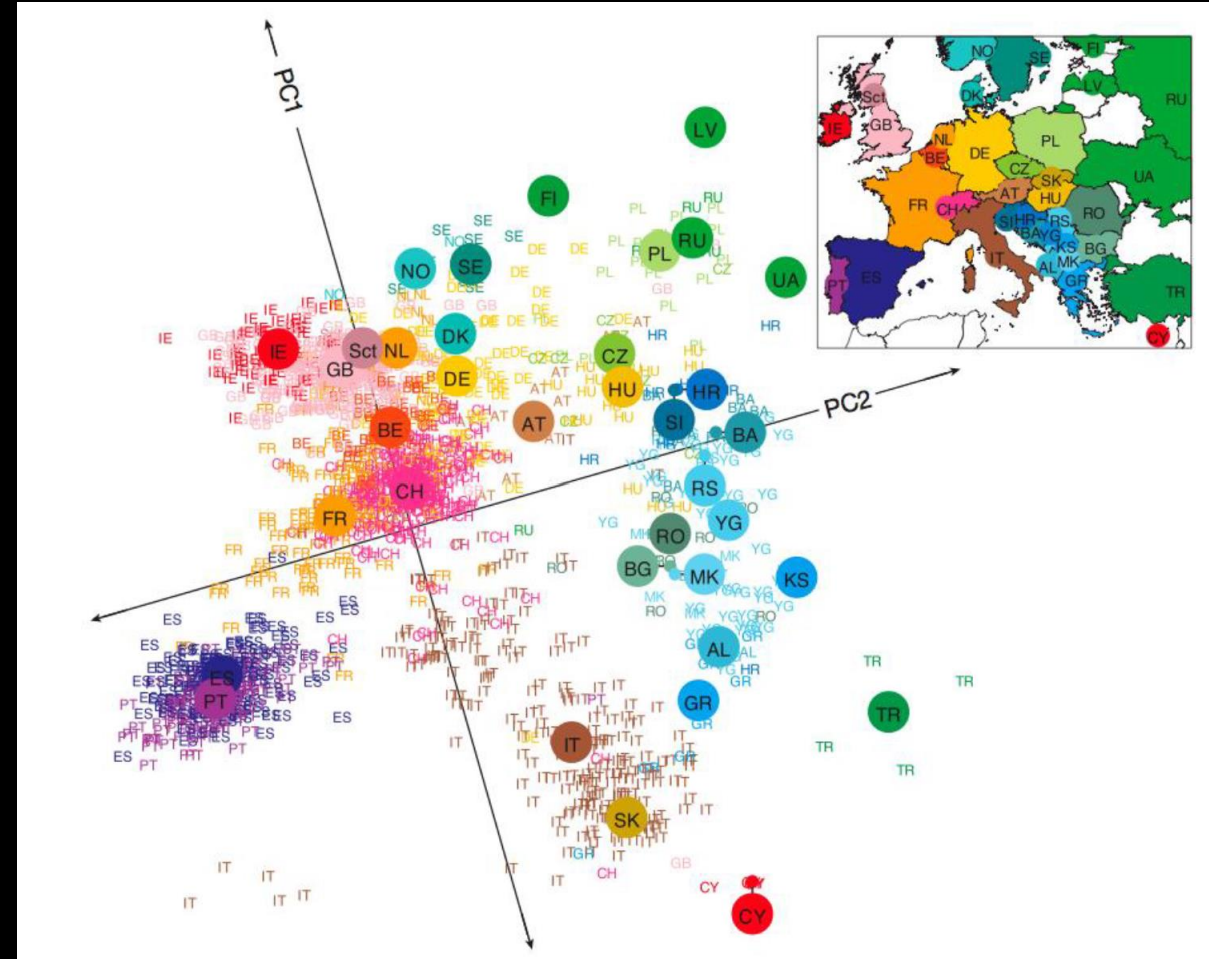
There is a clear allele frequency difference between Americans and Chinese

Sample 3: Americans + Chinese $\chi^2 = 34.2, p = 4.9 \times 10^{-9}$			
	Yes	No	Total
Allele 1	640	340	980
Allele 2	400	100	500
Total	1040	440	1480

How to deal with pop strat?

Start by including genetically homogenous samples into the GWAS!

- Control for genetic principal components in GWAS
- Genomic control (GC):
 - Estimates the factor with which the test statistics are inflated due to the population structure and/or cryptic relatedness
 - Divide SEs by square-root of the factor
- Mixed linear modeling
- Within family association



Part C - Genetic discovery

Candidate gene studies

GWAS

Imputation

Meta-analysis

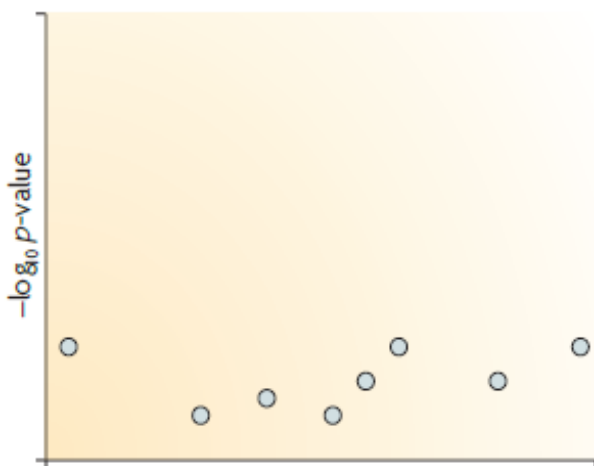
QC of GWAS summary statistics

Life after GWAS

Imputation

- The correlation structure of SNPs can be exploited to save costs
 - Instead of measuring all SNPs, a carefully selected set of SNPs from each haplotype can be chosen and genotyped
 - The remaining SNPs of the haplotype can be imputed with high accuracy using information about the correlation structure
- Typically use publicly available reference datasets, such as haplotypes from major sequencing projects
 - [1000 Genomes](#) ($N=2,504$, ≈ 84.7 mil SNPs, 3.6 mil short indels, 60,000 structural variants, final release of phase 3 Oct 2014)
 - [Haplotype Reference Consortium](#) – HRC (based on $N=38,821$)
 - [TopMed](#) (97,256 reference samples, ≈ 308 mio genetic variants)

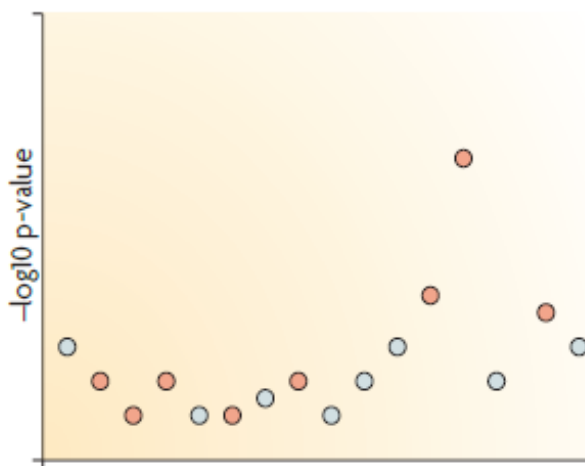
b Testing association at typed SNPs may not lead to a clear signal



d Reference set of haplotypes, for example, HapMap

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

f Testing association at imputed SNPs may boost the signal



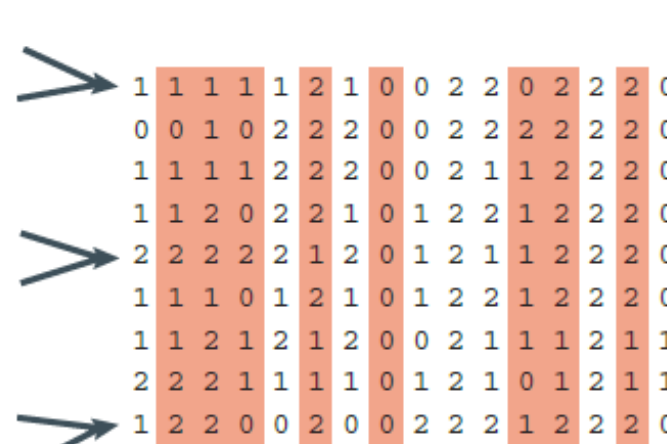
a Genotype data with missing data at untyped SNPs (grey question marks)

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

c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel



e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

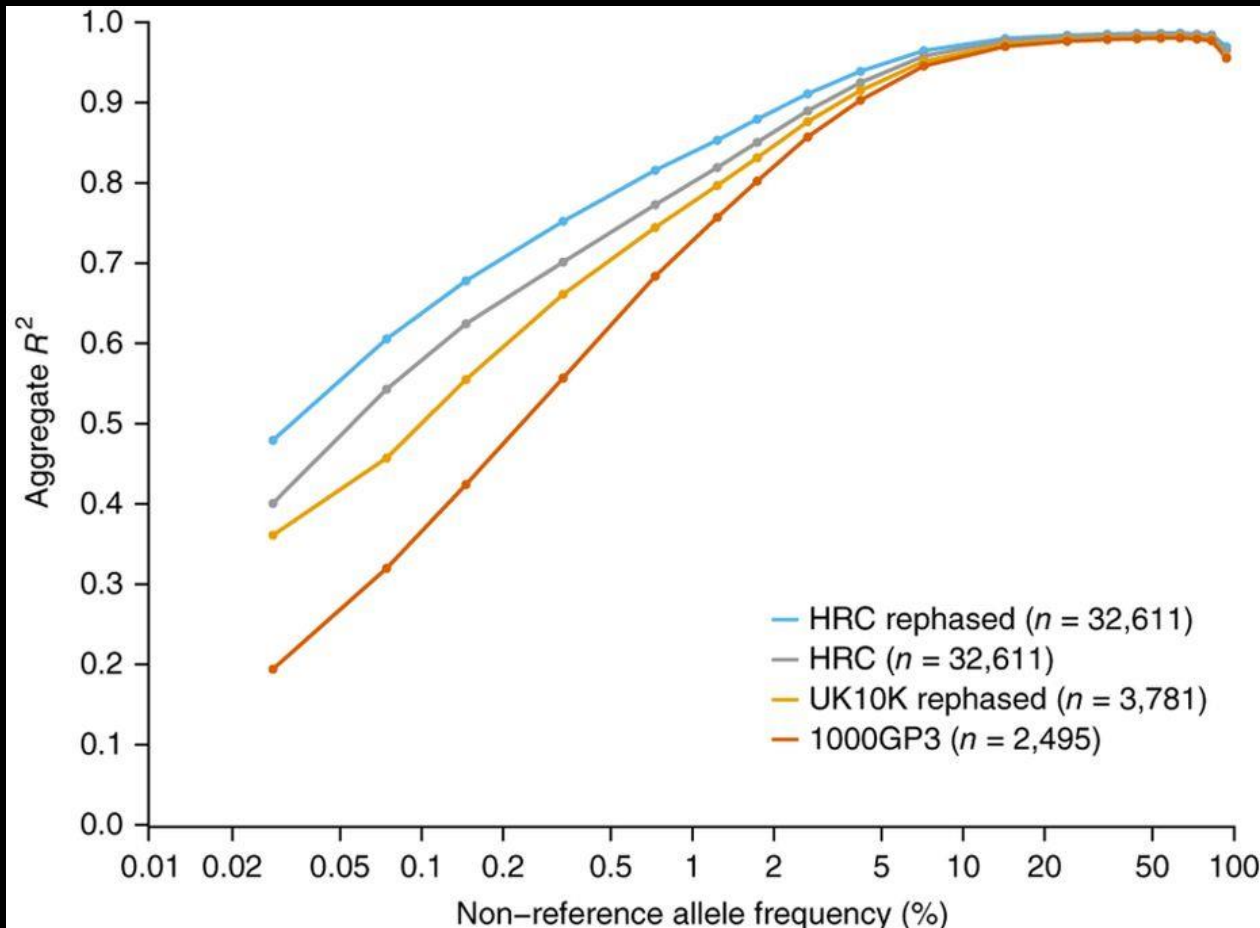


Source:
Marchini &
Howie (2010).
*Nature Review
Genetics*. DOI:
10.1038/nrg27
96

Advantages of imputation

- Advantages of imputation:
 - Cheaper than direct genotyping
 - Possibility to pool results from cohorts that were genotyped on different platforms
 - Higher power than GWAS on directly genotyped data
 - Up to 10% gain in power (Marchini & Howie 2010)
 - Fine-mapping of regions around associated SNPs
 - Correction of genotyping errors
 - If directly genotyped SNP call conflicts with other directly genotyped SNPs of the same person

Imputation quality across ancestries / MAF spectrum



- All reference panels do well for common SNPs (MAF > 5%).
- Imputation accuracy of rare SNPs (MAF < 1%) substantially improved with HRC
 - mainly for Europeans because HRC consists primarily of Europeans

Part C - Genetic discovery

Candidate gene studies

GWAS

Imputation






Meta-analysis

QC of GWAS summary statistics



Life after GWAS

Back to GWAS challenges

- Very large sample sizes are needed
 - Pooling data from many different datasets (“cohorts”)
 - Meta-analysis
 - Mega-analysis
 - Proxy-phenotypes that are available in $N > 100,000$
 - Multivariate analysis (e.g. MTAG, Genomic SEM)

RESEARCH ARTICLE | BIOLOGICAL SCIENCES |     

Common genetic variants associated with cognitive performance identified using the proxy-phenotype method

Cornelius A. Rietveld, Tõnu Esko, Gail Davies, , and Philipp D. Koellinger 

Edited by Michael S. Gazzaniga, University of California, Santa Barbara, CA, and approved August 14, 2014 (received for review March 12, 2014)

September 8, 2014





111 (38) 13790-13794

<https://doi.org/10.1073/pnas.1404623111>

PNAS

Article | Published: 01 January 2018


Multi-trait analysis of genome-wide association summary statistics using MTAG

Patrick Turley , Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A. Furlotte, 23andMe Research Team, Social Science Genetic Association Consortium, Patrik Magnusson, Sven Oskarsson, Magnus Johannesson, Peter M. Visscher, David Laibson, David Cesarini , Benjamin M. Neale  & Daniel J. Benjamin 

Nature Genetics **50**, 229–237 (2018) | [Cite this article](#)

Article | Published: 08 April 2019

Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits

Andrew D. Grotzinger , Mijke Rhemtulla, Ronald de Vlaming, Stuart J. Ritchie, Travis T. Mallard, W. David Hill, Hill F. Ip, Riccardo E. Marioni, Andrew M. McIntosh, Ian J. Deary, Philipp D. Koellinger, K. Paige Harden, Michel G. Nivard & Elliot M. Tucker-Drob

Nature Human Behaviour **3**, 513–525 (2019) | [Cite this article](#)

Pooling datasets

- Mega-analyses

- Individual-level data are uploaded to a common server and centrally analyzed
 - More options for analyzing the data
 - Analyses can be done quicker
 - Genotype and phenotype data can be checked and adjusted directly
 - But often not possible in practice (e.g. IRB, legal issues, privacy)

- Meta-analyses

- Summary statistics from GWAS are uploaded to a common server and meta-analyzed
 - Practically feasible (protecting individual-level data)
 - Local analysts know their data better than central analysts
 - Slow
 - Limited range of analyses that can be conducted

Estimated β_i and their SE from these analyses are asymptotically identical

No efficiency gain from using individual-level data!

Lin & Zeng (2010 *Genetic Epidemiology*, DOI:10.1002/gepi.20435)

Meta-analysis workflow

1. Write analysis-plan and post it online
2. Invite cohorts
 - Descriptive statistics
 - Collaboration agreement
3. Cohorts conduct GWAS according to analysis-plan
4. Cohorts upload results to a secure server
5. Meta-analysts conduct Quality Control (QC)
 - Follow-up on problems, missing information
6. Data freeze
7. Meta-analysts conduct meta-analysis
 - Cross-checking of results

Meta-analysis weighting scheme - 1

	Sample-size weighting	Inverse-variance weighting
Inputs	N_i – sample size for study i P_i – p -value for study i Δ_i – direction of effect in study i	β_i – effect size estimate in i se_i – standard error in i
Intermediate statistics	$Z_i = \Phi^{-1} \left(1 - \frac{P_i}{2} \right) (\Delta_i)$ $w_i = \sqrt{N_i}$	$w_i = 1/se_i^2$ $se = \sqrt{1 / \sum_i w_i}$ $\beta = \sum_i \beta_i w_i / \sum_i w_i$
Overall Z-score	$Z = \left(\sum_i Z_i w_i \right) / \sqrt{\sum_i w_i^2}$	$Z = \beta / se$
Overall P -value	$P = 2\Phi(- Z)$	

Meta-analysis weighting scheme - 2

Sample-size weighting:

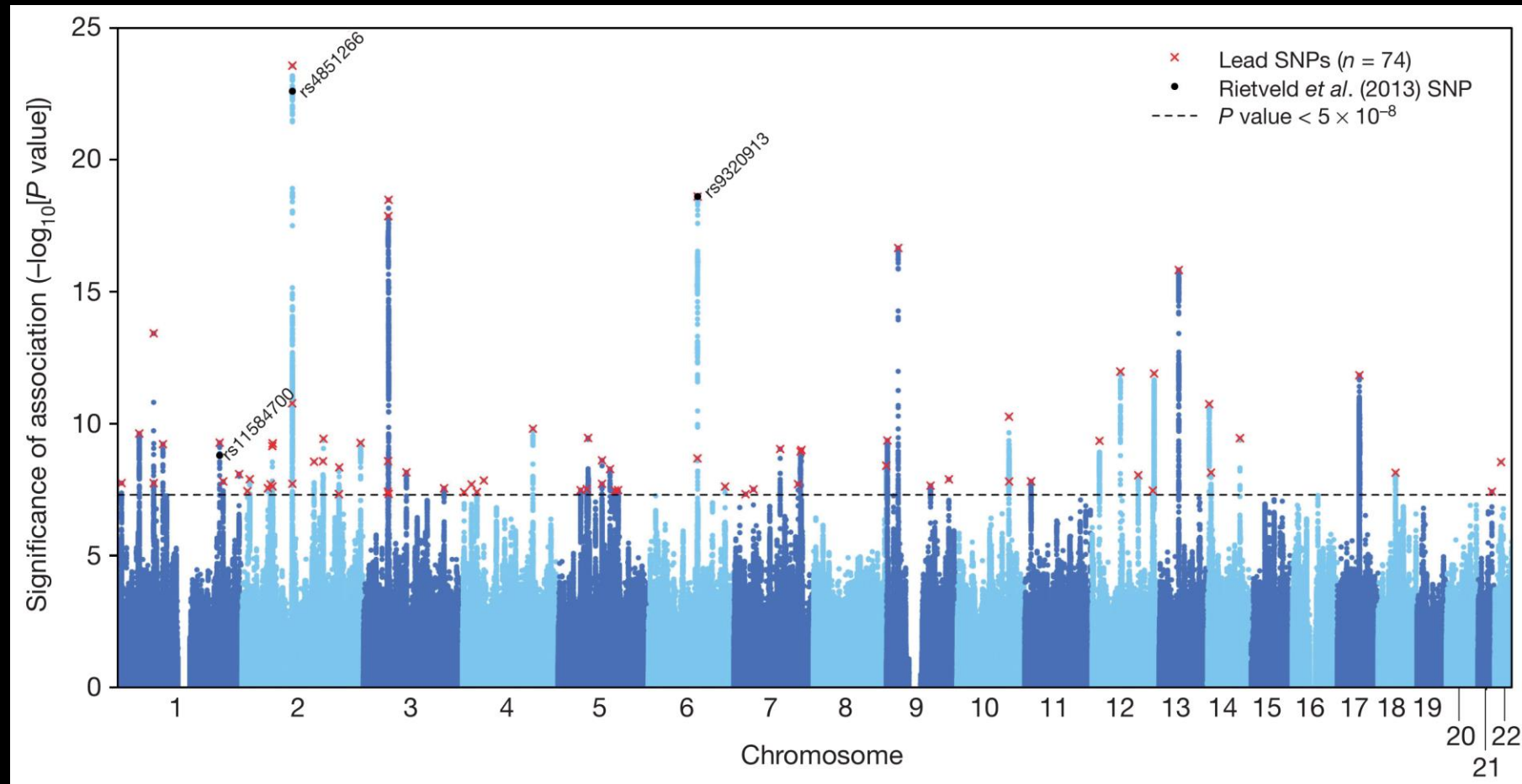
- Can be used when effect sizes have been estimated on different scales
 - height in feet and height in meters
 - different coding of the same outcome categories across cohorts

Inverse variance weighting:

- More precise estimates get higher weight.
- Better to use if trait was measured on the same scale, but with varying degrees of accuracy across samples (to avoid imprecise estimates from some large cohorts getting too much weight)
- Slightly more powerful in finite samples

Asymptotically, the two approaches are equivalent when the trait distribution is identical across samples.

Manhattan plot



Manhattan plot from work by Okbay *et al.* (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, **533**, 539–542.

Obtaining independent signals

How to go from genome-wide significant SNPs to “independent loci”?

- **Clumping** algorithm (implemented in Plink) :
 1. Take the SNP with lowest P -value (lead SNP).
 2. Check the correlation between the lead SNP and all SNPs within a window (e.g. 500kb, 1mb).
 3. Assign the SNPs with a correlation greater than your pre-specified threshold (e.g. 0.1) to the first clump (“locus”).
 4. Take the next most significant SNP, repeat steps 1-3 until no genome-wide significant SNPs remain.

Part C - Genetic discovery

Candidate gene studies vs. GWAS

Imputation

Meta-analysis

QC of GWAS summary statistics

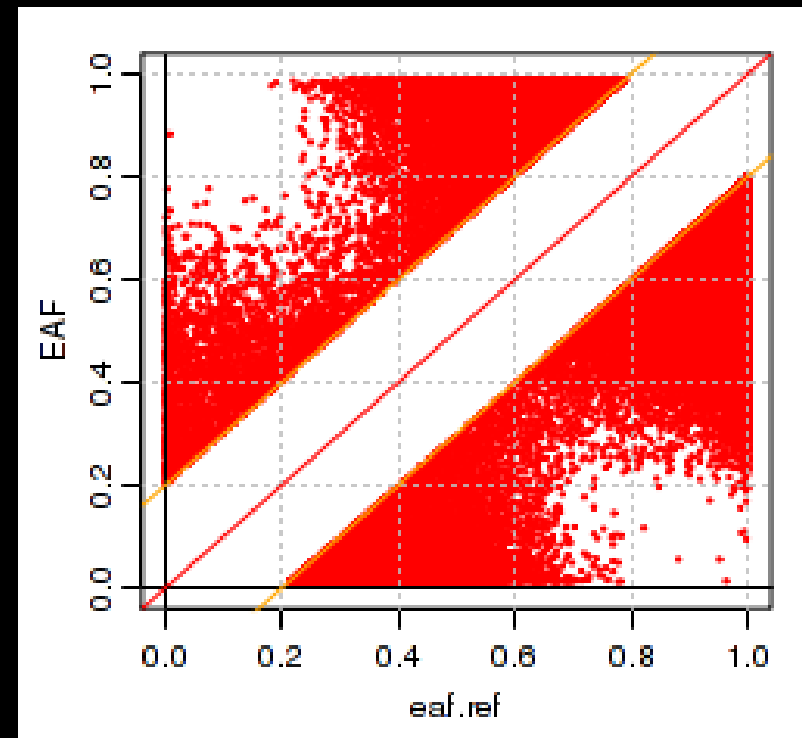
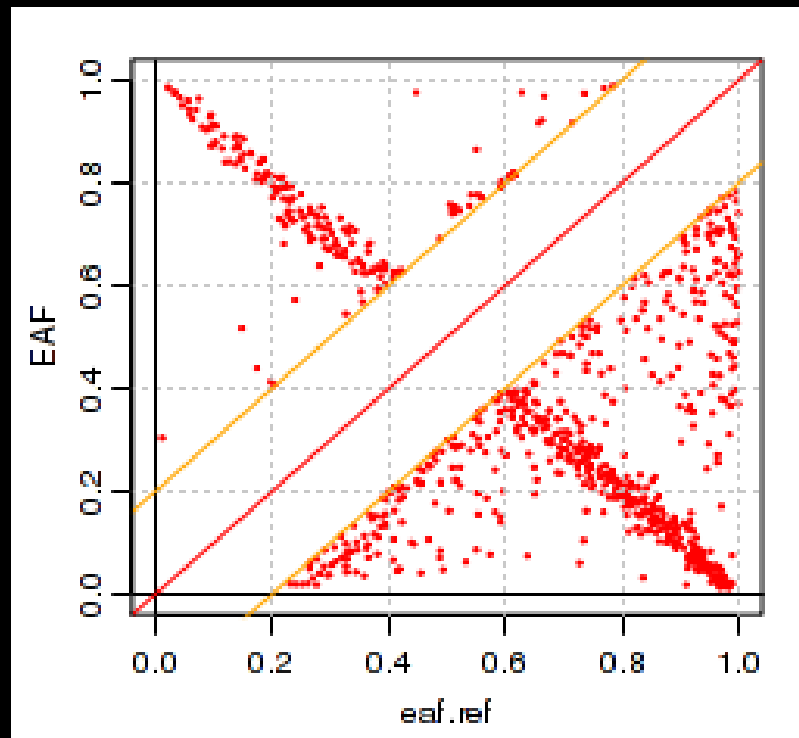
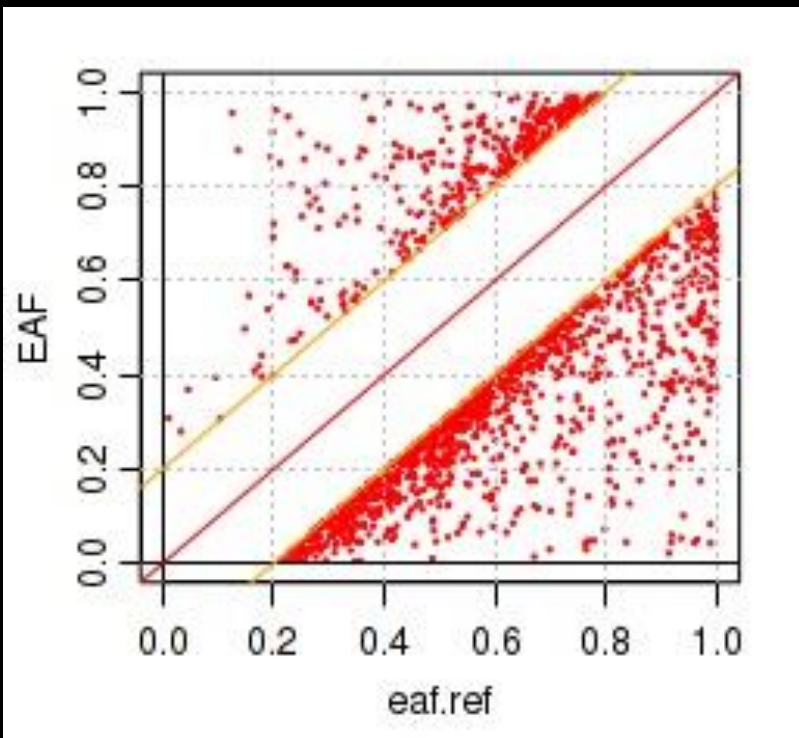
Life after GWAS

QC - What can go wrong?

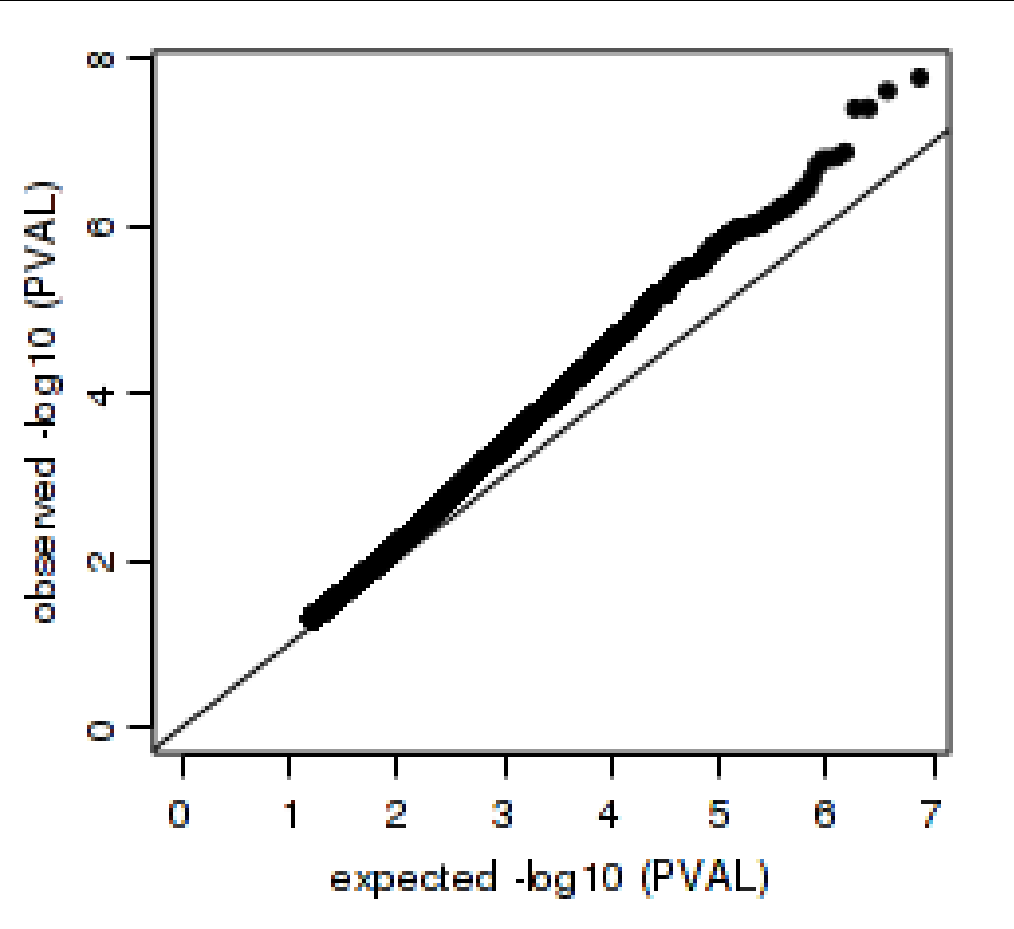
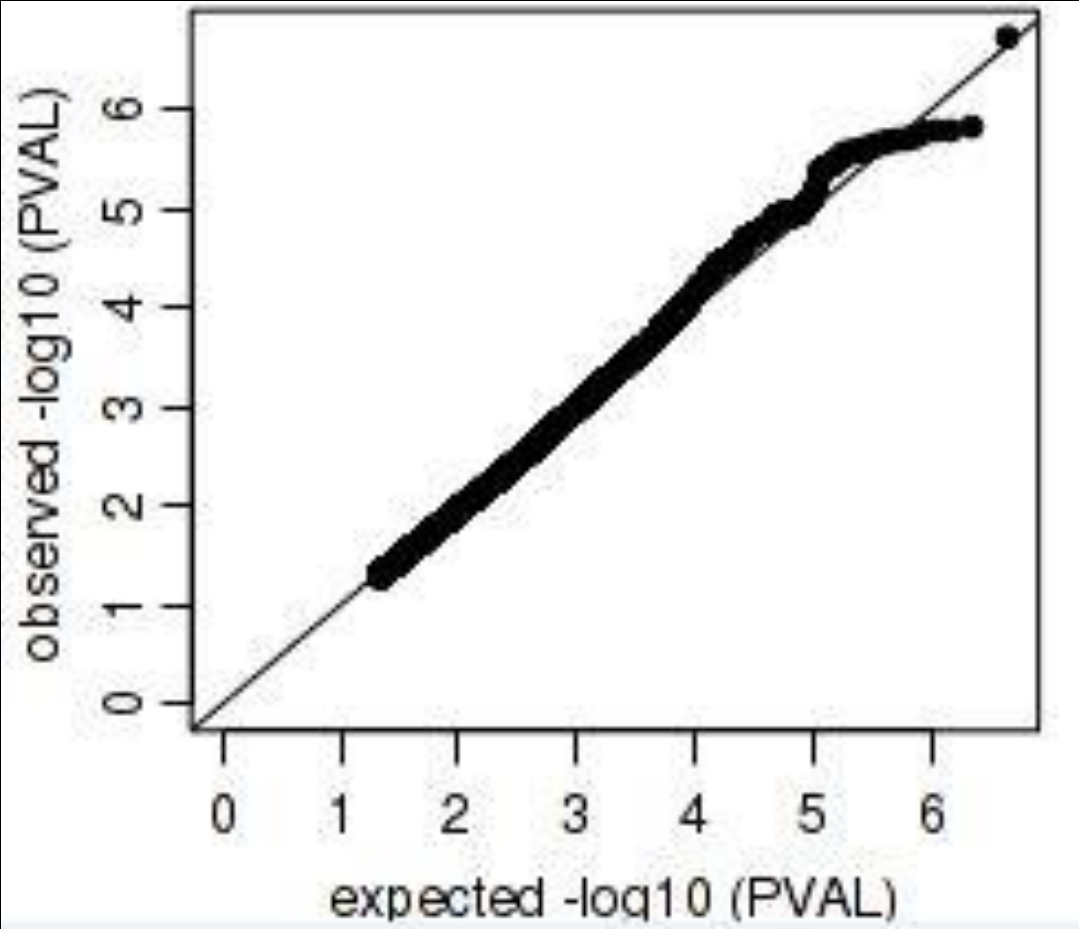
Anything you can think of! Examples:

1. Incorrect coding of the outcome, e.g. reverse-coding, incorrect missing value coding
2. Incorrect model specification, e.g. omission of necessary covariates, no proper control for population stratification
3. Problems with genotype data, e.g. flipped SNP alleles, errors in imputation
4. Unreliable SNPs, e.g. low imputation quality, low minor allele frequency.

Allele frequency plot



QQ-plot



Part C - Genetic discovery

Candidate gene studies

GWAS

Imputation

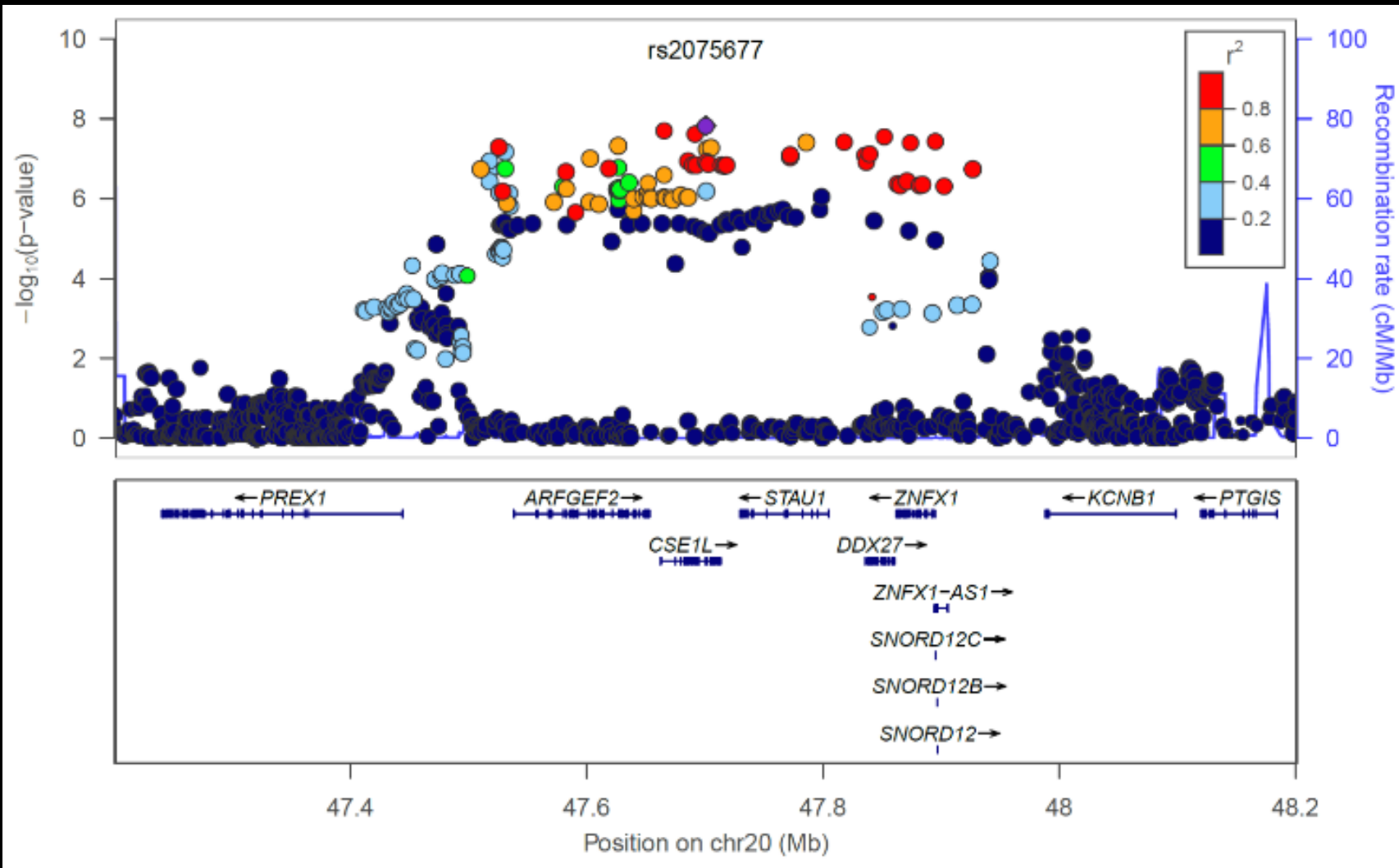
Meta-analysis

QC of GWAS summary statistics

Life after GWAS

What to do with GWAS results?

- Biological annotation
 - Which are the causal variants?
 - Which are the causal genes?
 - Which biological pathways are involved?
 - Which are the tissues of action?
- Heritability, genetic correlations with other traits
- Mendelian randomization
- Multi-trait analyses: Genomic SEM, MTAG
- Construct polygenic indices



Part D – Polygenic Prediction

Predictive power of polygenic indices

Constructing polygenic indices

Applications

Limitations & pitfalls

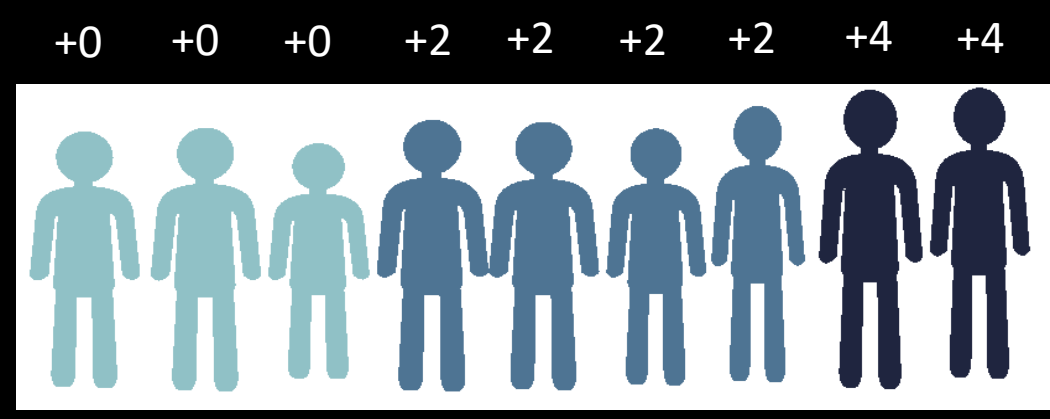
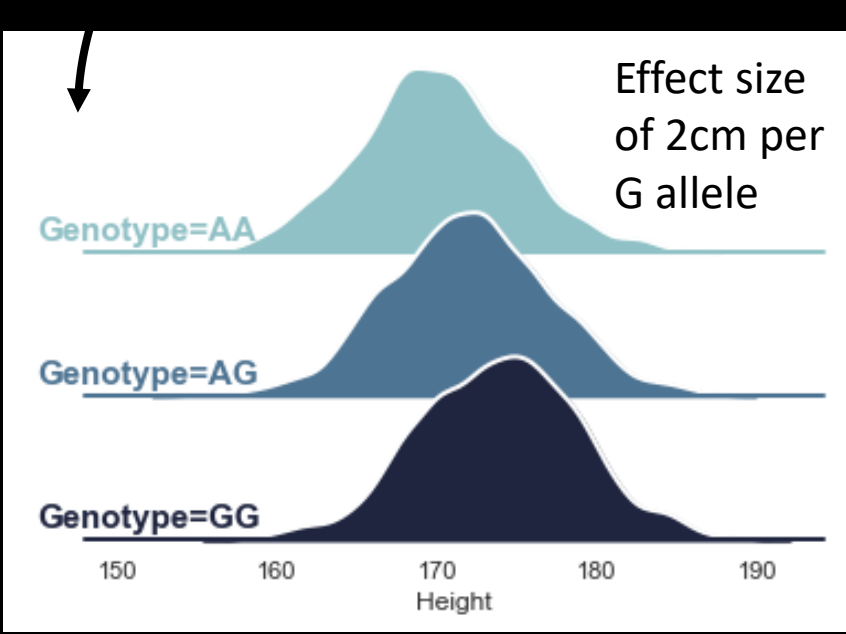
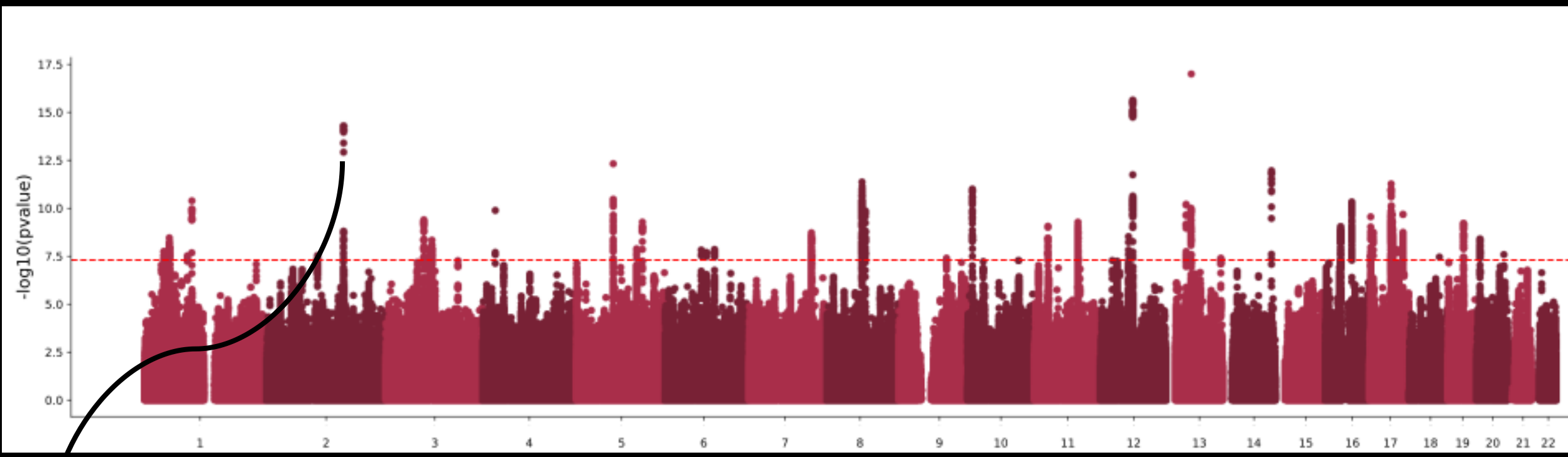
Polygenic score
(PGS)

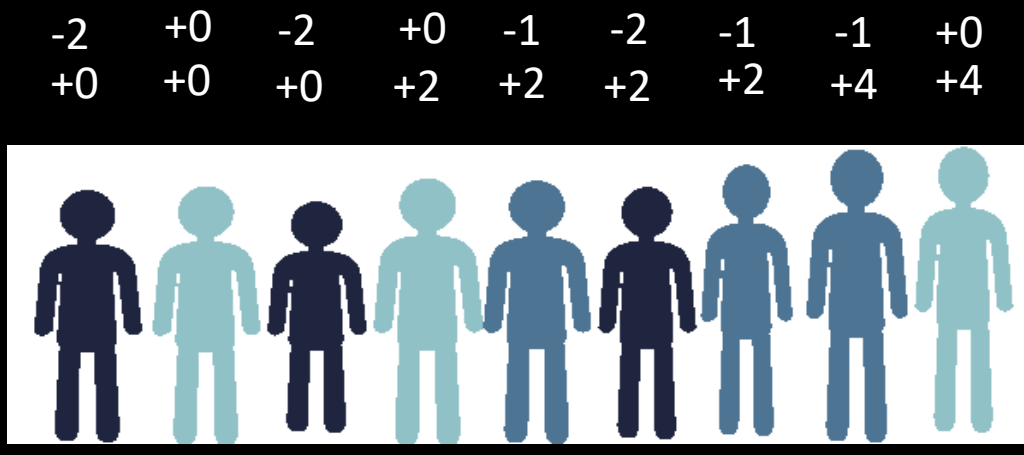
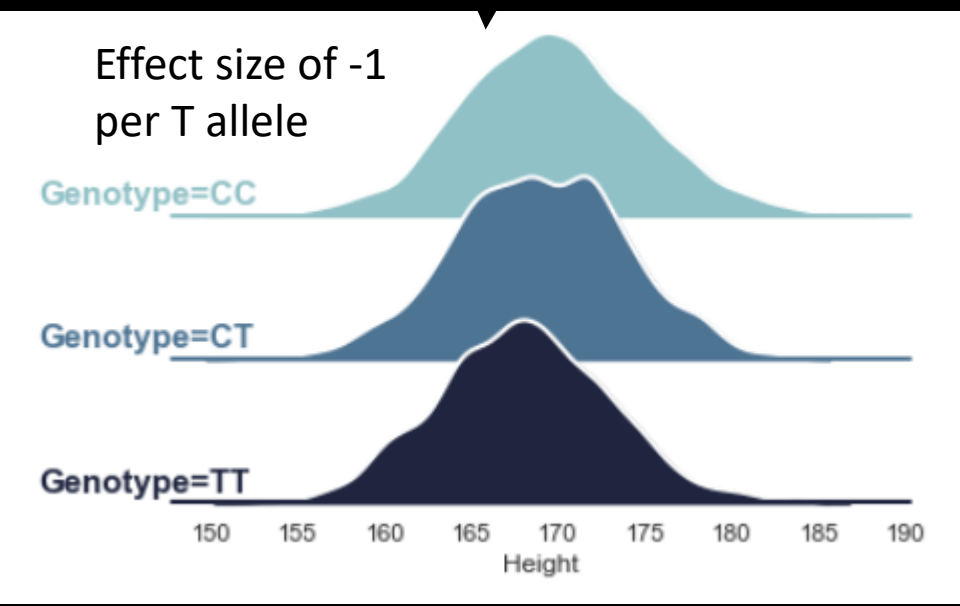
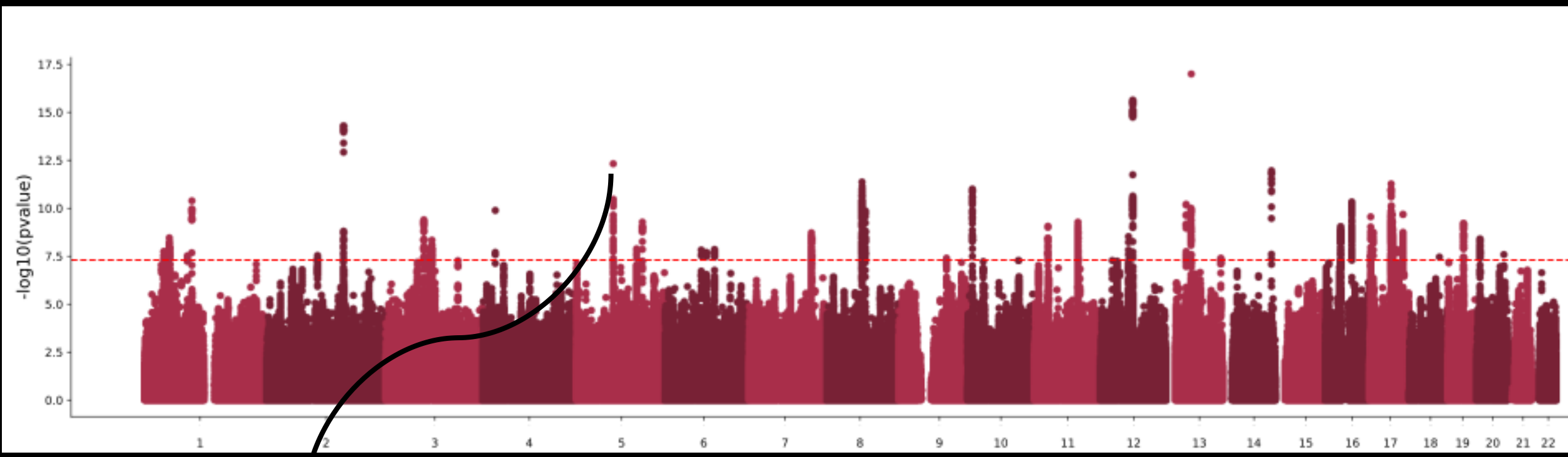
Polygenic index
(PGI)

Polygenic risk
score (PRS)

Genetic risk
score (GRS)

Genome-wide
score (GWS)





What is a polygenic index?

- An index that linearly aggregates the estimated effects of individual SNPs on the trait of interest.
- Can be considered a measure of an individual's genetic propensity towards a trait.
- Defined as a **weighted sum of a persons genotypes at K loci**.
- Start with additive model using measured SNPs:

$$y_i = A_{SNP,i}(x_i) + \epsilon_{i,SNP} = \sum_{j=1}^K \beta_j x_{ij} + \epsilon_{i,SNP}$$

↓
additive SNP factor

What is a polygenic index?

Additive SNP factor:

$$A_{SNP,i}(x_i) \equiv \sum_{j=1}^K \beta_j x_{ij}$$

True effect size of
SNP j

PGI:

$$\hat{A}_{SNP,i}(x_i) \equiv \sum_{j=1}^K \hat{\beta}_j x_{ij}$$

Estimated effect size of
SNP j

$$\hat{\beta}_j = \beta_j + u_j \Rightarrow \hat{A}_{SNP,i} = \sum_{j=1}^K (\beta_j + u_j) x_{ij} = A_{SNP,i} + U_i \text{ where } U_i = \sum_{j=1}^K u_j x_{ij}$$

If u is mean-zero estimation
error uncorrelated with β_j

U is mean-zero
measurement error

$$E(\hat{A}_i | A_i) = A_i$$

Predictive power of a polygenic index

If we regress y on \hat{A}_{SNP} we get an OLS coefficient of

$$\begin{aligned}
 b &= \frac{Cov(\hat{A}_{SNP}, y)}{Var(\hat{A}_{SNP})} \\
 &= \frac{Cov(A_{SNP} + U_i, A_{SNP} + \epsilon_{SNP})}{Var(A_{SNP} + U)} \\
 &= \frac{Var(A_{SNP})}{Var(A_{SNP}) + Var(U)}
 \end{aligned}$$

And the expected predictive power is:

$$\begin{aligned}
 R^2 &= \frac{b^2 Var(\hat{A}_{SNP})}{Var(y)} \\
 &= \left(\frac{Var(A_{SNP})}{Var(A_{SNP}) + Var(U)} \right)^2 \frac{Var(\hat{A}_{SNP})}{Var(y)} \\
 &\quad \vdots \\
 &\approx \frac{h_{SNP}^2}{h_{SNP}^2 + \frac{M_e}{N}}
 \end{aligned}$$

Sometimes called the Daetwyler formula (Daetwyler et al. 2008)

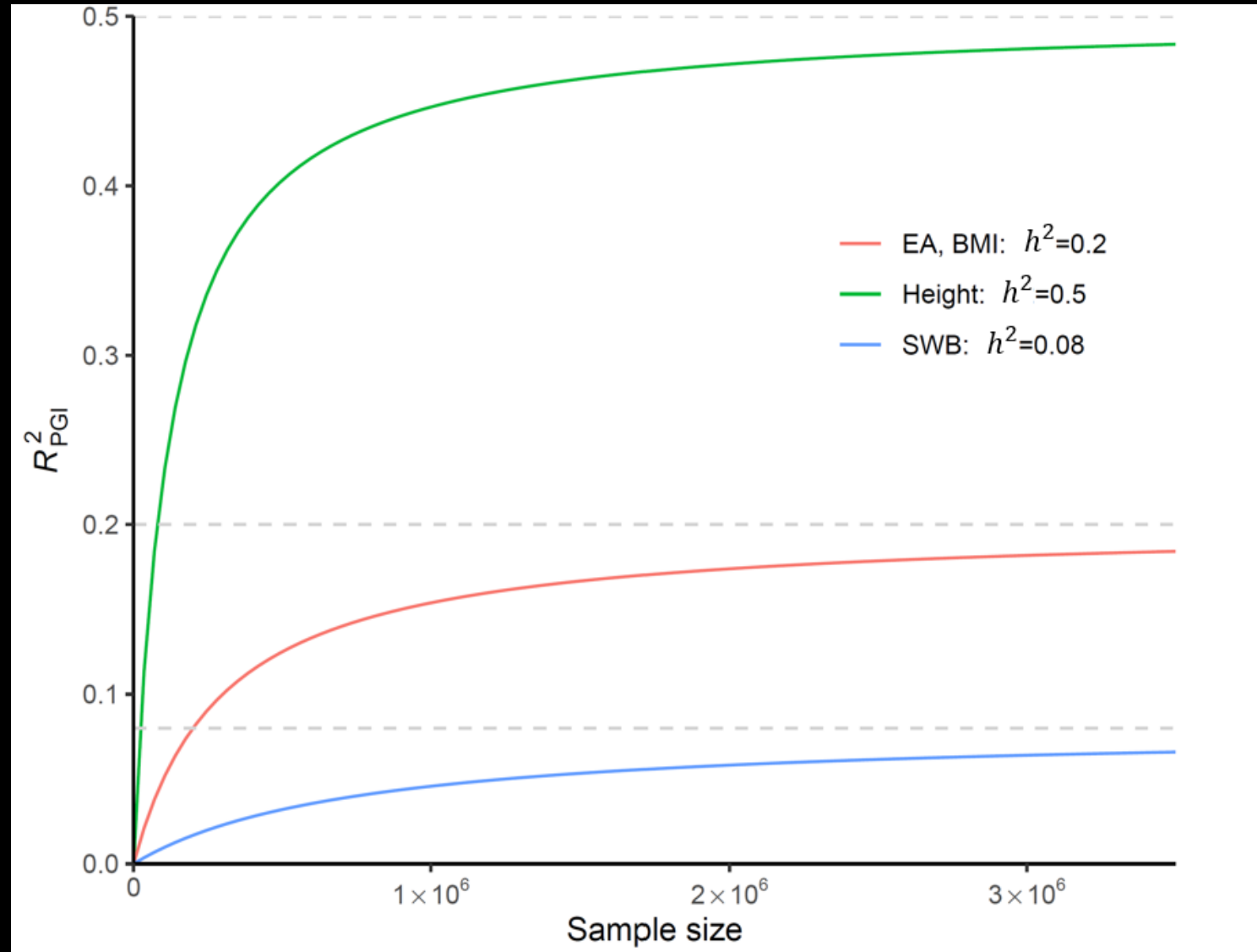
Effective number of SNPs in the PGI, estimated to be between 50k-70k in genome-wide data for EUR ancestry (Wray et al. 2013)

OLS:

$$y_i = a + bx_i + \epsilon_i$$

$$b = \frac{Cov(x,y)}{Var(x)}, R^2 = \frac{b^2 Var(x)}{Var(y)}$$

Theoretical projections for $R_{P_{GI}}^2$



Predictive power and heterogeneity

What if we are predicting into a cohort where the genetic architecture is not the same as the GWAS sample?

y, A_{SNP} : phenotype and additive SNP factor in the training (GWAS) sample

y^*, A_{SNP}^* : phenotype and additive SNP factor in the validation sample

$$A_{SNP,i}^* \neq A_{SNP,i} \rightarrow h_{SNP}^{2*} \equiv \frac{Var(A_{SNP,i}^*)}{Var(y_i^*)} \neq h_{SNP}^2$$

Define the genetic correlation to be

$$r_g = Corr(A_{SNP,i}^*, A_{SNP,i})$$

The expected predictive power

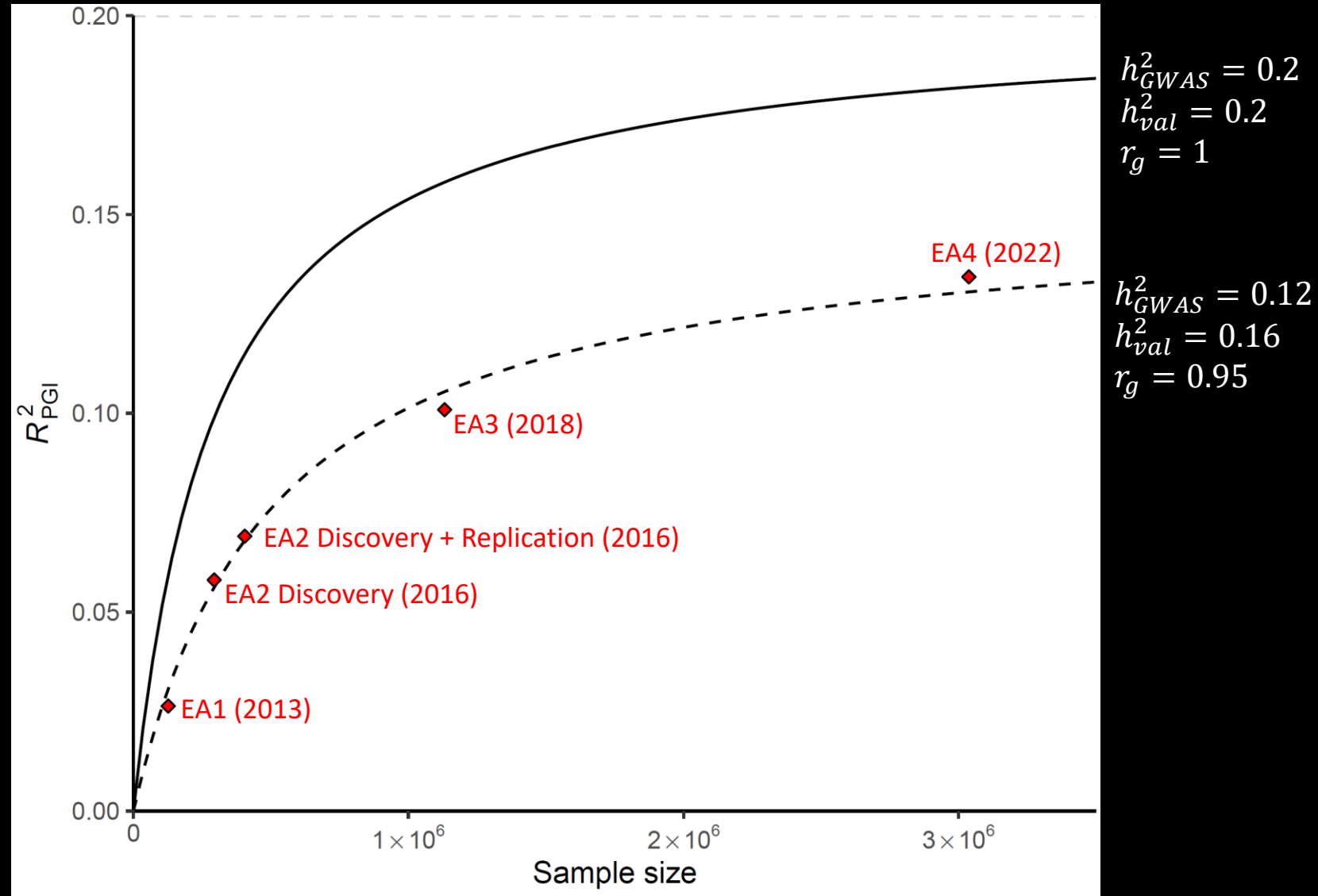
$$E(R^2) \approx \frac{h_{SNP}^2}{h_{SNP}^2 + \frac{M_e}{N}}$$

now becomes

$$E(R^2) \approx \frac{r_g h_{SNP}^2 h_{SNP}^{2*}}{h_{SNP}^2 + M_e/N}$$

(De Vlaming et al. 2016)

Theoretical projections for R_{PGI}^2 vs Observed R_{PGI}^2



Part D – Polygenic Prediction

Predictive power of polygenic indices

Constructing polygenic indices

Applications

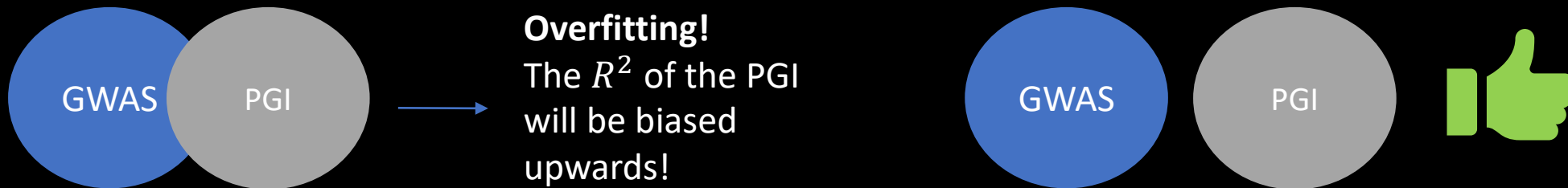
Limitations & pitfalls

Constructing polygenic indices

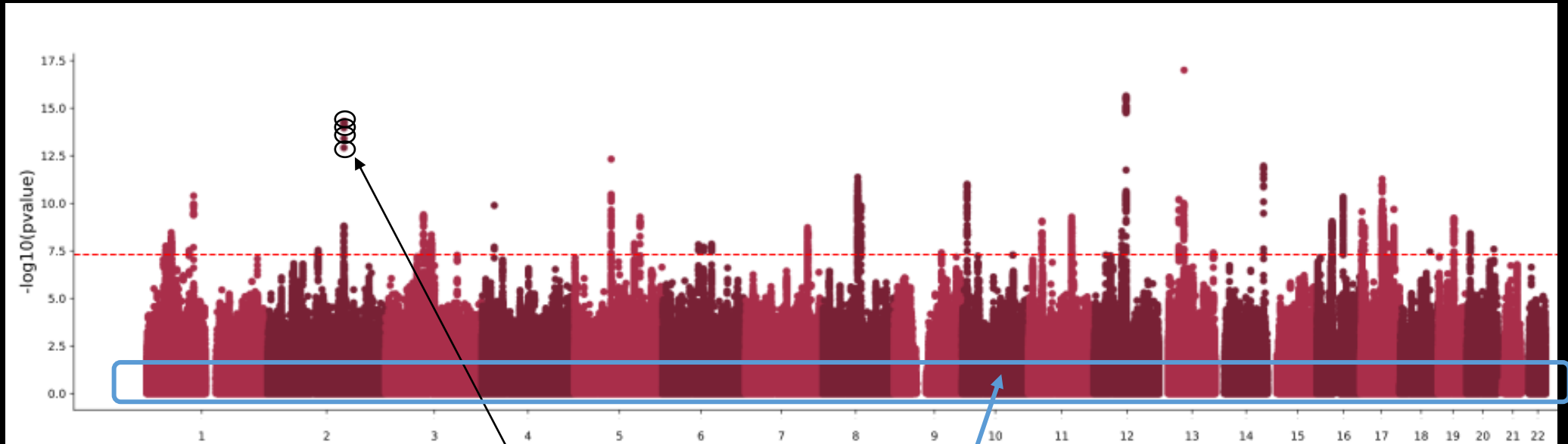
What is needed?

- Individual-level genotype data from a prediction sample.
- Weights: GWAS summary statistics from a discovery sample

Caution: The prediction sample should not overlap with the discovery sample!



Weights



GWAS results give us $\hat{\beta}_j^{GWAS}$, not β_j . Two issues to consider when constructing $\sum_{j=1}^K \hat{\beta}_j^{GWAS} x_{ij}$:

1. For some SNPs, $\hat{\beta}_j^{GWAS}$ may be a very noisy estimate of β_j and/or β_j may be close to 0, so adding those SNPs will add more noise than signal
2. If we include all SNPs, we will overweight (“double-count”) SNPs with high LD scores

Two solutions

Clumping and thresholding

Include only the most strongly associated SNP from each LD block (Purcell et al., 2009)

Weights: Set equal to GWAS coefficients.

Loci: Selected by

1. using a **clumping** algorithm that ensures the included markers are all approximately independent of each other
2. omitting SNPs whose P value for association with the phenotype is above a certain **threshold**

$$\sum_{j=1}^K \hat{\beta}_j^{GWAS} x_{ij}$$

Bayesian approaches

Include all SNPs but adjust the effect sizes for LD

Weights: Set to GWAS coefficients **adjusted for LD** → approximate results from a theoretical multiple regression of the phenotype on all SNPs

Loci: Include **all SNPs**, no LD-based pruning

Examples: LDpred (Vilhjalmsson et al. 2015, Prive et al. 2020), PRS-CS (Ge et al. 2019), SBayesR (Lloyd-Jones et al. 2019)

Practical considerations - (C+T)

P-value cutoff: Depends on

- the polygenicity of the trait
 - For highly polygenic traits, reasonable to expect prediction R^2 to increase when more SNPs are included
- the sample size of the discovery GWAS
 - smaller the GWAS sample, the larger the P -values → imposing a very strict P -value threshold may drop too many SNPs in a small GWAS.

Clumping parameters

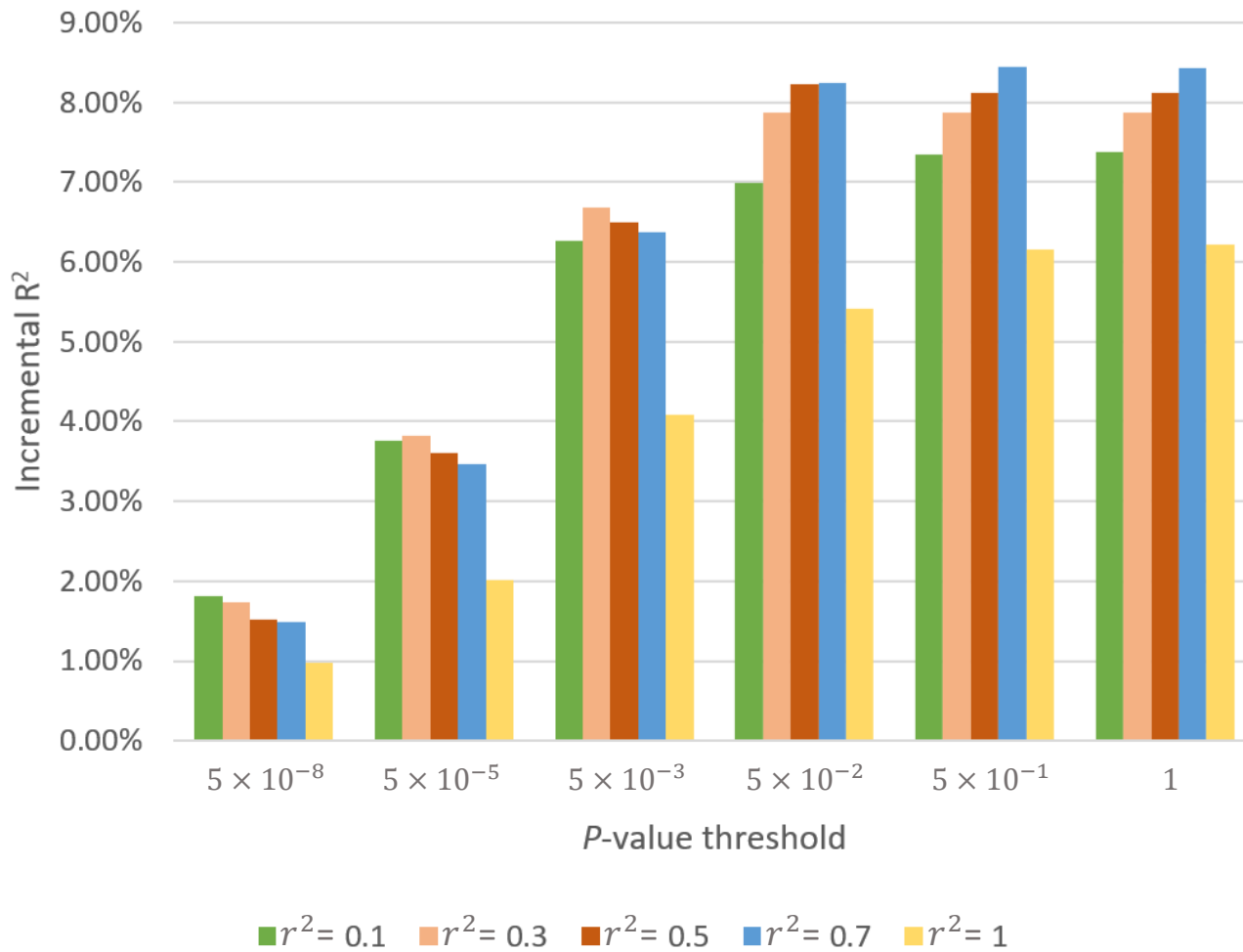
- r^2 threshold: Do not want to double-count, but also do not want to lose signal
- LD-window:
 - If too large, then errors in LD estimates can lead to apparent LD between unlinked loci.
 - If too small, there is risk of not accounting for LD between linked loci.

Imputed or genotyped SNPs?

Depends on

- genotyping chip coverage
- quality of imputed SNPs

Predictive power of C+T PGS with different clumping r^2 and P -value thresholds



- **Cohort:** Health and Retirement Study
- **Phenotype:** Educational attainment

Practical considerations - (Bayesian approaches)

Uses as weights

$$E(\beta_j | \hat{\beta}_j^{GWAS}, D) \xrightarrow{\text{LD matrix}}$$

By Bayes's rule,

$$f(\beta_j | \hat{\beta}_j^{GWAS}, D) = \frac{f(\hat{\beta}_j^{GWAS} | \beta, D) f(\beta_j | D)}{f(\hat{\beta}_j^{GWAS} | D)}$$

Shrinkage depends on the prior!

PRS-CS: "Continuous shrinkage"

$$(\beta_j | D) \sim N(0, \phi \psi_j)$$

$$\psi_j \sim N(a, \delta_j)$$

$$\delta_j \sim N(b, 1)$$

Parameters a and b determine how aggressively to shrink small estimates and how much you don't shrink large ones

LDpred2: Gaussian or Spike-and-Slab

$$(\beta_j | D) \sim \begin{cases} N(0, \tau^2), & \text{with probability } \pi \\ 0 & \text{with probability } 1 - \pi \end{cases}$$

π can be estimated from data, sparsity allowed (if $\bar{\pi}_j < \pi$, b_j set to 0), $\tau^2 = h^2 / M\pi$

SBayesR: flexible finite mixture of normal distributions, sparsity allowed

$$(\beta_j | D) \sim \begin{cases} 0, & \text{with probability } \pi_1 \\ N(0, \gamma_2 \sigma_b^2), & \text{with probability } \pi_2 \\ \dots & \\ N(0, \gamma_c \sigma_b^2) & \text{with probability } 1 - \sum_{c=1}^{c-1} \pi_c \end{cases}$$

Practical considerations - (Bayesian approaches)

Reference genotype data to calculate LD matrix should be

- large enough
- representative of the GWAS sample
- cleaned
 - sample-level filters: related individuals, ancestry outliers, individuals with low genotyping rate
 - SNP-level filters: low SNP call rate, MAF, HWE P-value (genotyped SNPs), imputation accuracy (imputed SNPs)

Which method is better?

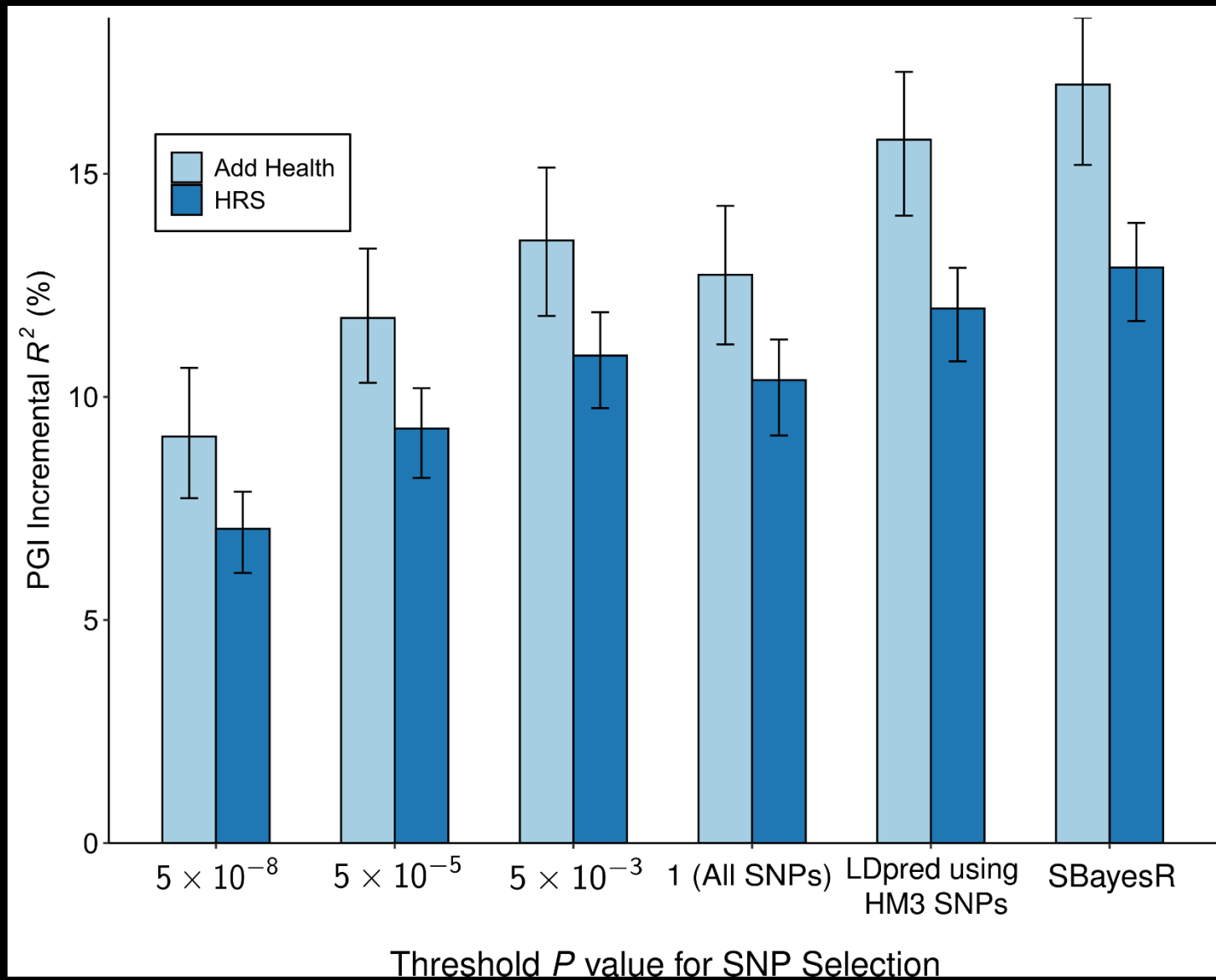
Clumping and thresholding

Faster and easier, but too black & white

- If clumping r^2 or P -value cutoffs too strict, it drops potentially causal SNPs.
- If clumping r^2 and P -value cutoffs too relaxed, there is a lot of double-counting and noise

Bayesian approaches

- utilize information from all SNPs by adjusting SNP weights for LD, but
 - if the reference panel is not a good match for the population from which summary statistics were obtained, prediction accuracy might be compromised
 - the assumed prior distribution might not accurately model the true genetic architecture



Part D – Polygenic Prediction

Predictive power of polygenic indices

Constructing polygenic indices

Applications

Limitations & pitfalls

Applications

Major advantage of PGI over specific genetic variants: can have much greater predictive power

e.g., if $R_{PGI}^2 = 0.07$, then 80% to detect its effect in a sample of size ~ 110 individuals. If $R_{PGI}^2 = 0.09$, then ~ 85 individuals.

→ Can study PGI in datasets containing high quality measures of outcomes, mediators, and covariates.



Identify correlates of genetic factors

e.g. Educational attainment PGI predicts early speech acquisition and is mediated by cognitive ability (Belsky et al., 2016).



Identify causal effects of genetic factors

Sibling data and family fixed effects → causal effect of PGI



Study treatment effect heterogeneity by genotype

e.g. Increase of compulsory schooling age in U.K. reduces BMI only among those with a high-BMI PGI (Barcellos, Carvalho, and Turley 2016)



Use as control variable

To control for confounding genetic factors or to increase statistical power for estimating the effect of a randomized treatment. If incremental R_{PGI}^2 is 15%, then power increase is equivalent to 17% increase in sample size (Rietveld, 2013)



Use for balance tests of randomization

PGIs should be identically distributed in treatment and control groups (Davies et al. 2016, Barcellos, Carvalho, and Turley 2016)



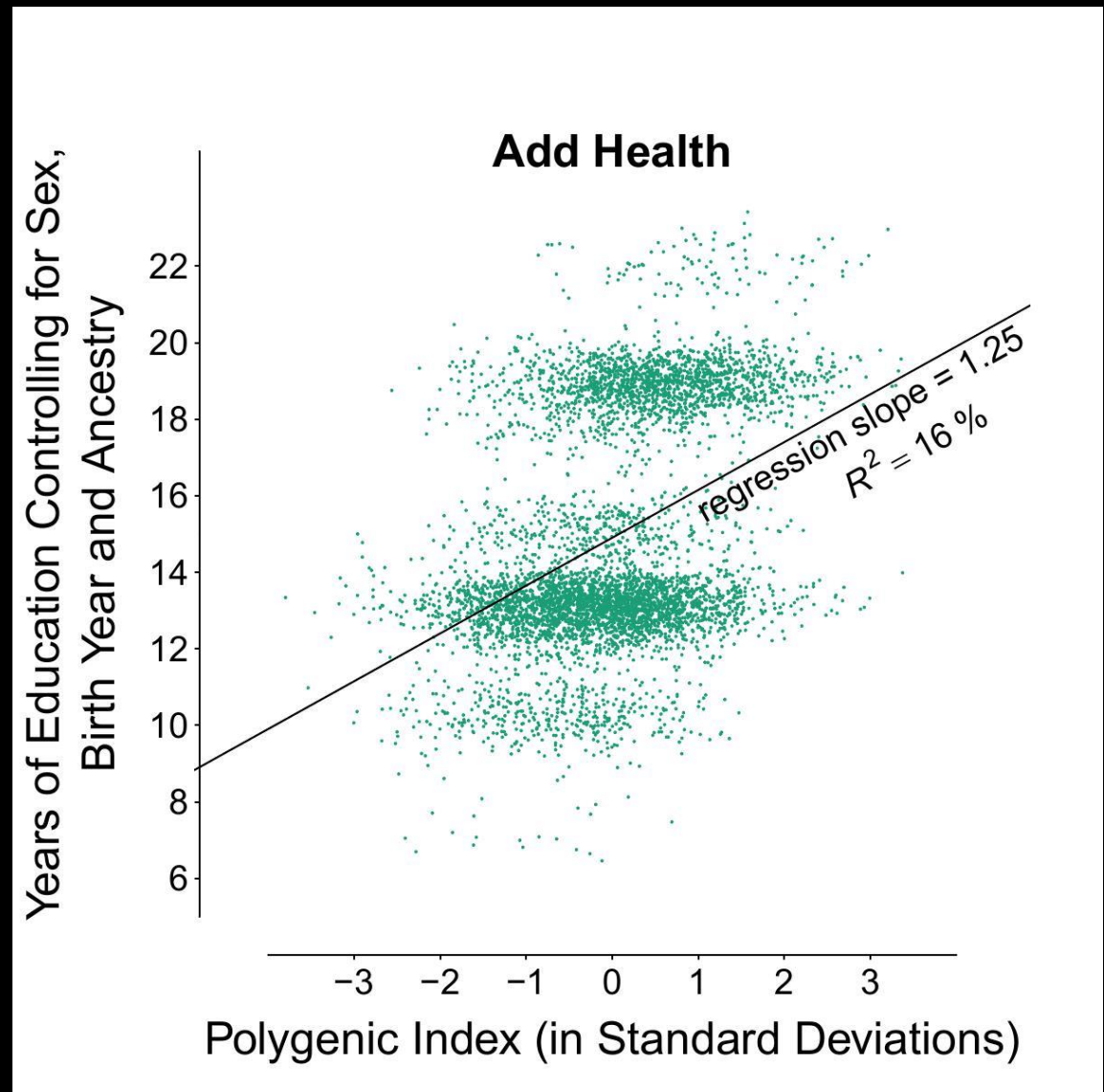
Identify at-risk individuals



Personalized treatment

⋮

Individual-level prediction is not accurate enough for most complex phenotypes!



Source: Okbay et al. (2022)

Part D – Polygenic Prediction

Predictive power of polygenic indices

Constructing polygenic indices

Applications

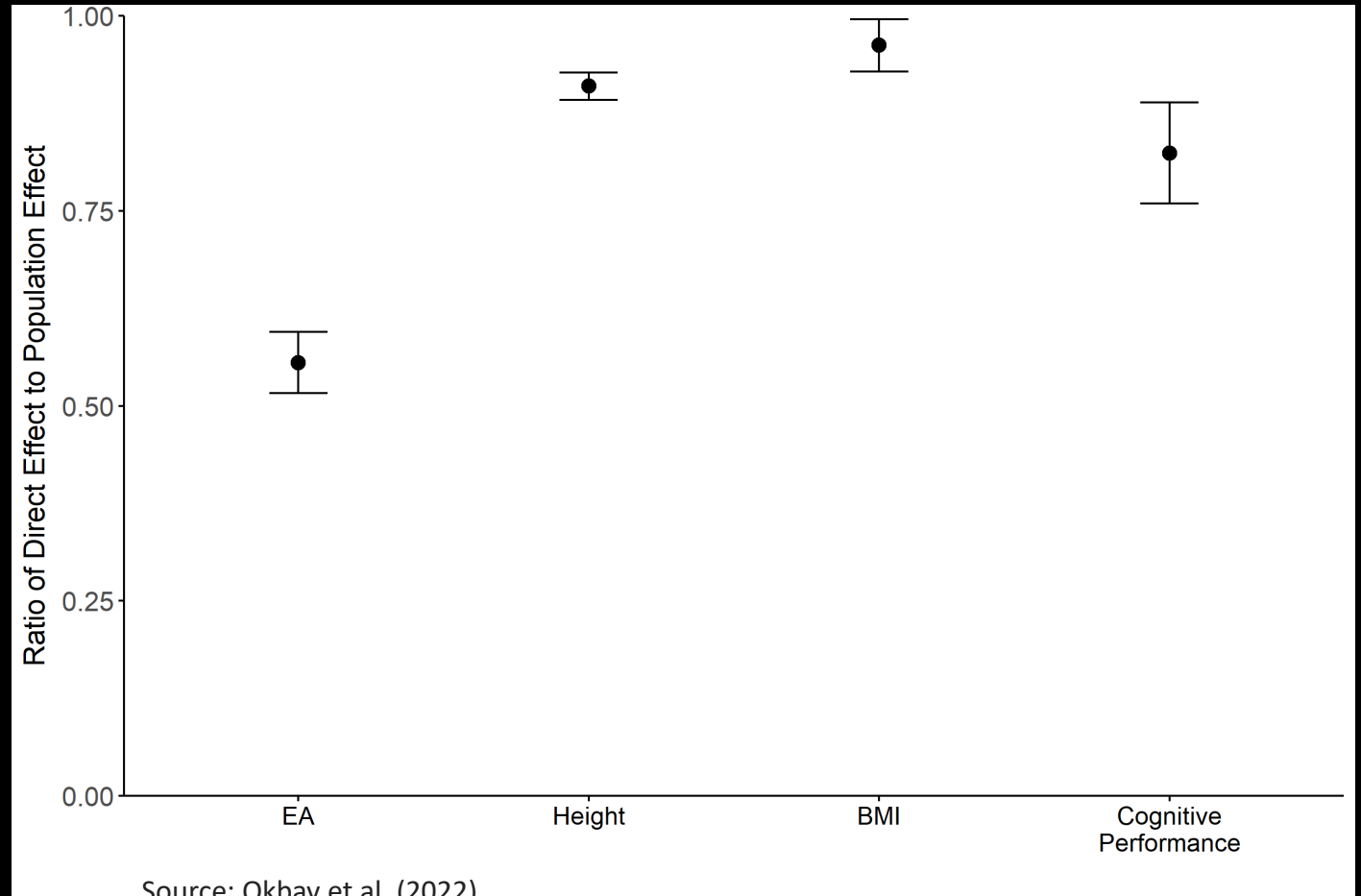
Limitations & pitfalls

Limitations and pitfalls - I

Mechanisms are poorly understood.

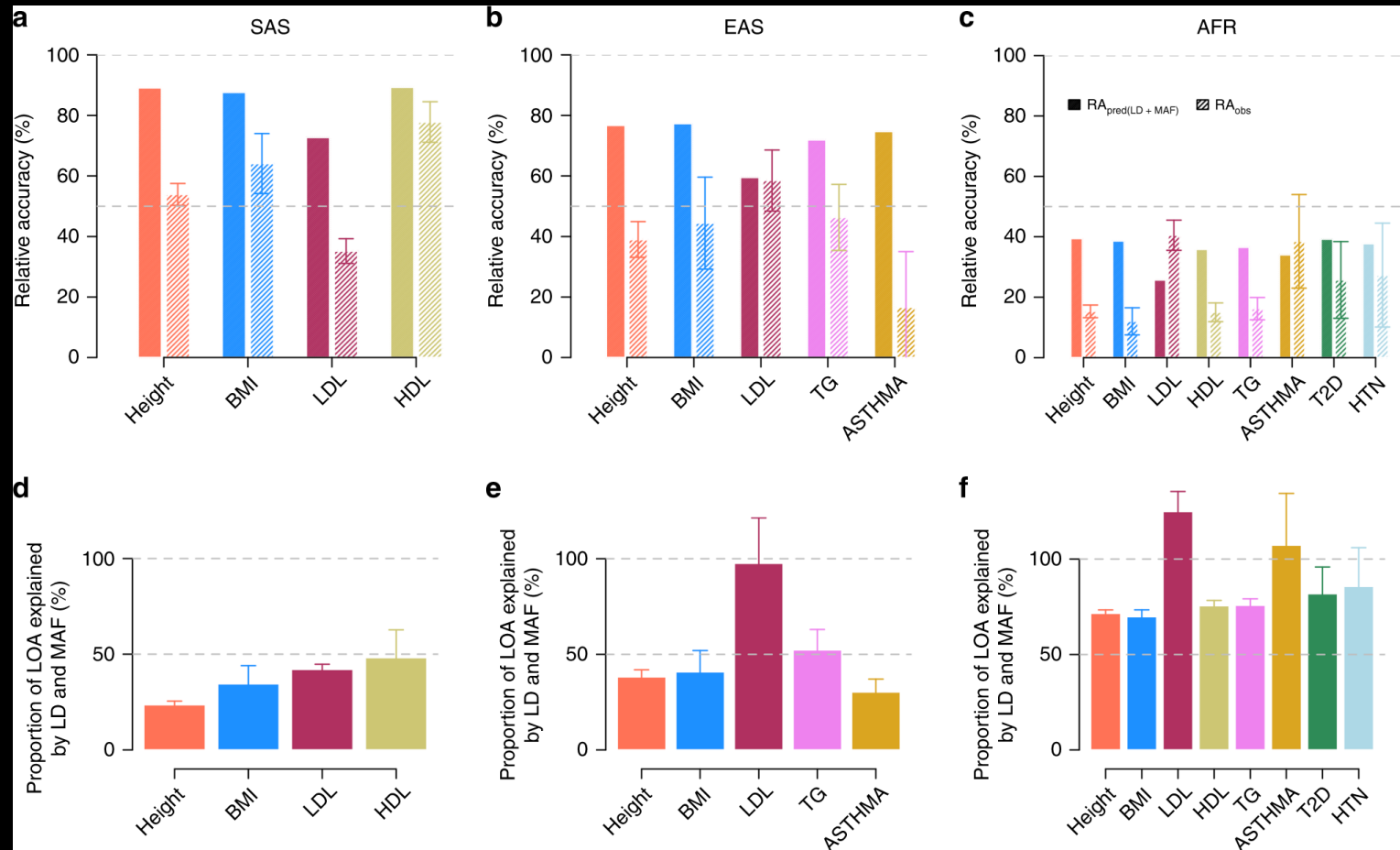
- Including many genetic variants
 - increases predictive power
 - requires including genetic variants with unknown function

→ makes it hard to specify what is captured by PGI.



Limitations and pitfalls - II

- Current polygenic indices far less predictive in non-European-descent samples.
 - For example, for the EA4 PGI:
 - $R^2 \approx 17\%$ for European-ancestry individuals in Add Health, 13% in HRS.
 - $R^2 \approx 2.3\%$ for African-ancestry individuals in Add Health, 1.3% in HRS.
- Relative accuracies of 15% and 11%



Source: Wang et al. (2020)

Limitations and pitfalls - III

Two sources of population stratification

- In the discovery phase
 - leads to bias in the GWAS estimates, so the PGI may give more weight to SNPs that just correspond to ancestry
- In the prediction phase
 - If the prediction sample is stratified, this can lead to bias in our PGI-based analyses even if SNP-weights are unbiased
- Interaction of bias in both phases
 - The combination of these two interact so group differences are strongly exaggerated

→ Important to control for PCs in prediction analyses!

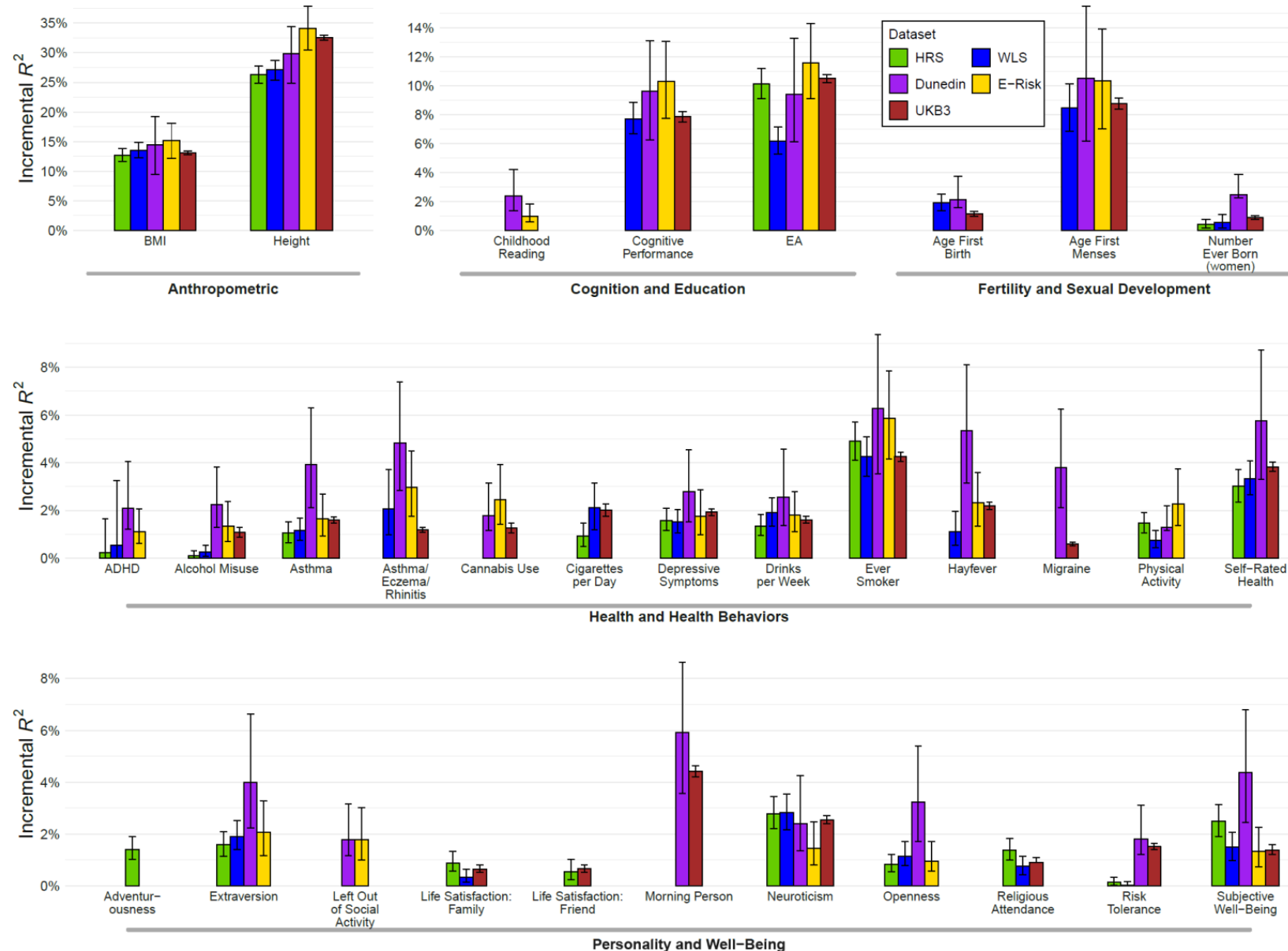
PGI Repository

v1.0

- 47 phenotypes
- 11 cohorts

v2.0 (coming soon)

- 7 new cohorts, 20 new phenotypes
- Parental PGIs



QUESTIONS?

