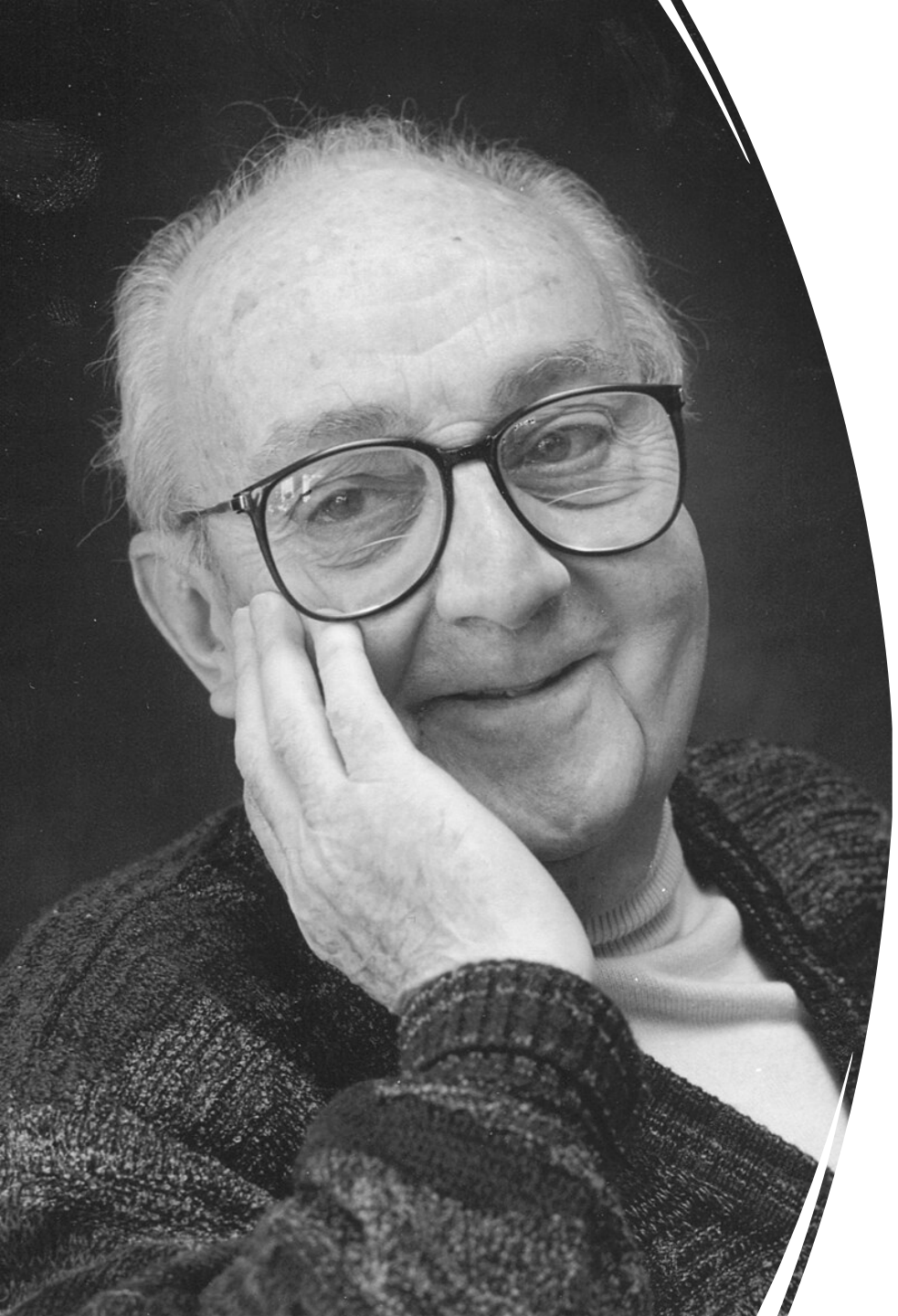
Introduction to GWAS

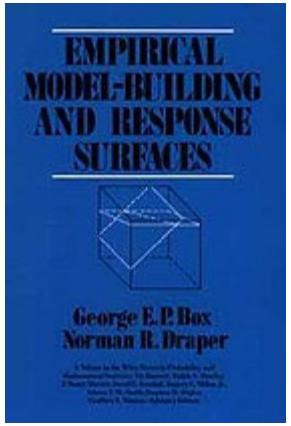
Sarah Medland



George Box

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

Box, G.E.P. and Draper, N.R. (1987) Empirical model-building and response surfaces. New York: Wiley. p424



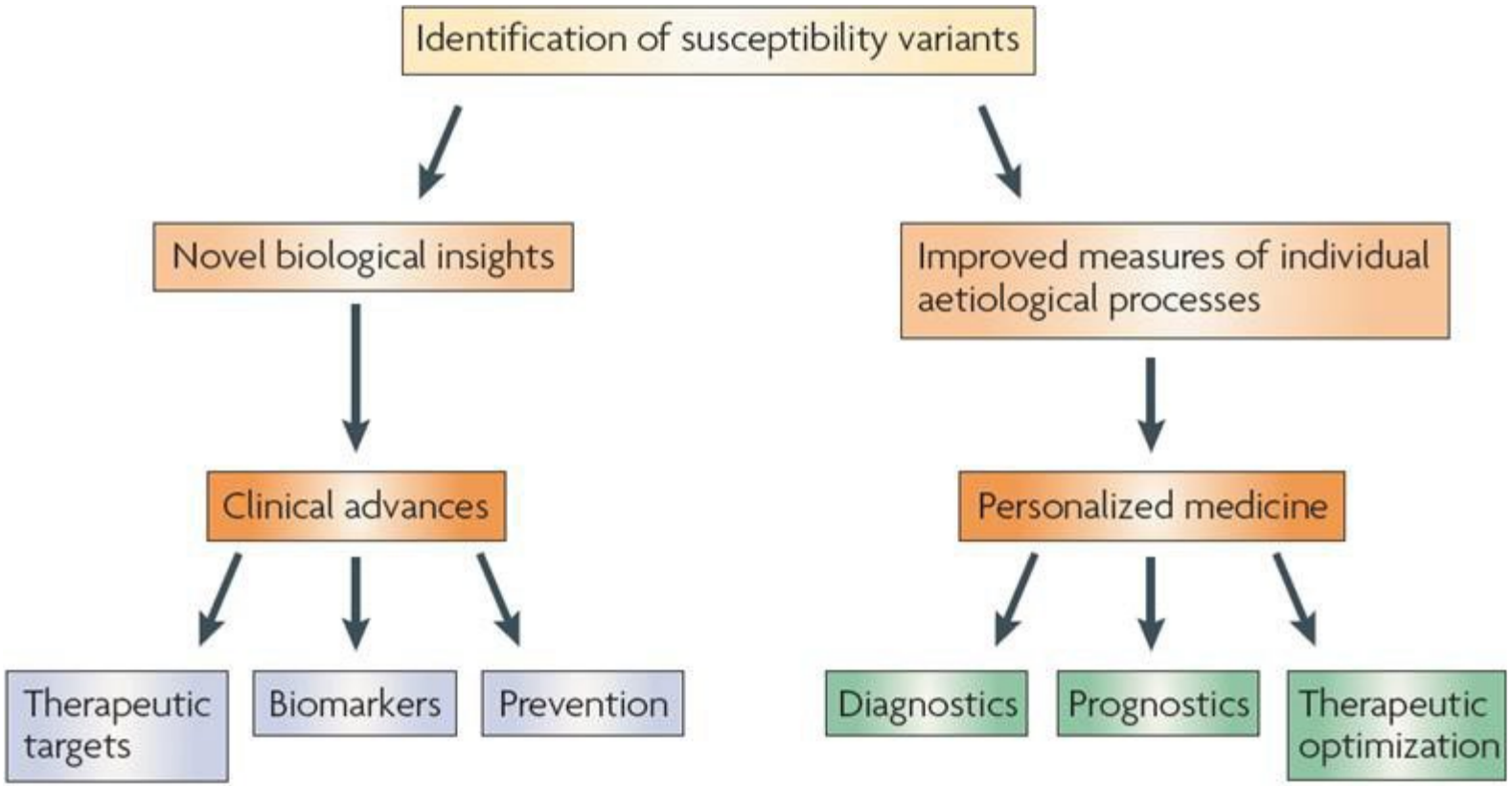

SAFETY REMINDER
USE THE RIGHT TOOL
FOR THE RIGHT JOB
IN THE RIGHT WAY

What is GWAS

- A hypothesis free study of genetic variation across the entire human genome
- Tests for genetic associations with continuous traits or with the presence / absence of disease
- With a focus on common loci
- Can detect with direct and indirect signals

Hirschhorn & Daly. *Nat Rev Genet* (2014)

Why do it?



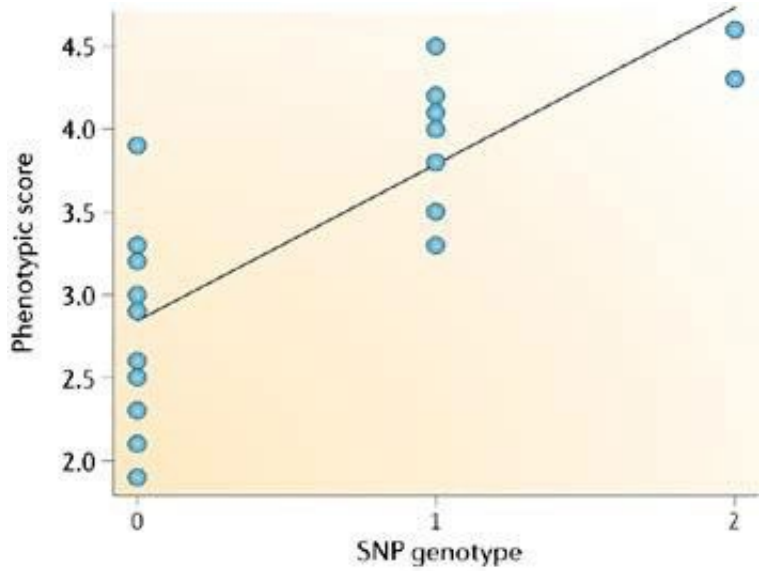
McCarthy et al. *Nat Rev Genet* (2008)

Association with unrelated individuals

Quantitative Trait

Linear Regression

$$\hat{Y} = \alpha + \beta X + \varepsilon$$



$\hat{Y}$ = score on phenotype

X = 0, 1 or 2 copies of allele ("G")

$\beta = 0$ no association

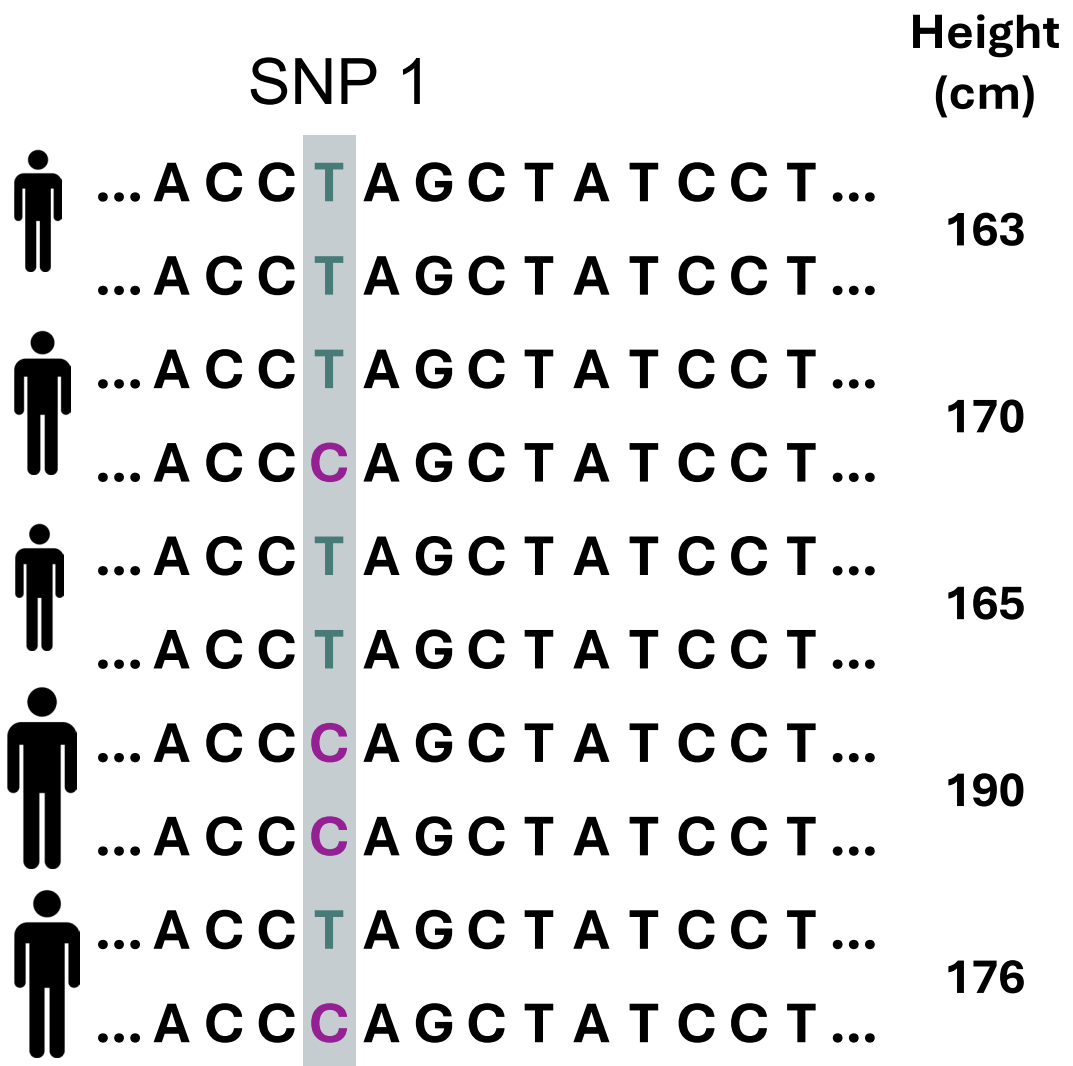
$\beta > 0$ G allele associated with higher score on trait

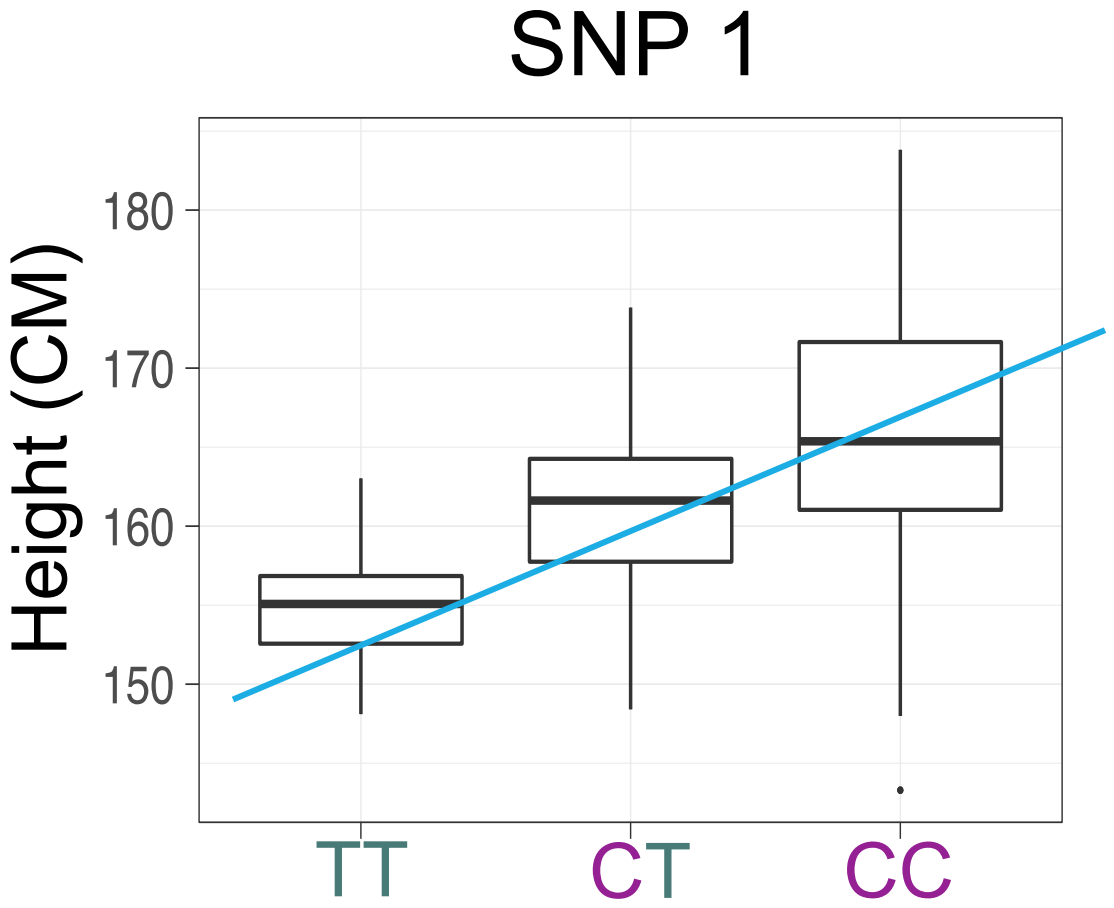
$\beta < 0$ G allele associated with lower score on trait

Balding. *Nat Rev Genet* (2006)

Genetic association tests

To identify genetic variants that are associated with a complex trait



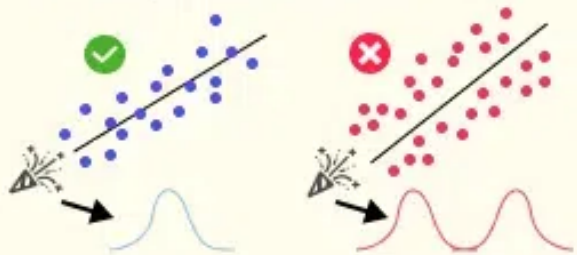


Assumptions of Linear Regression

1. Linearity
(Linear relationship between Y and each X)

2. Homoscedasticity
(Equal variance)

3. Multivariate Normality
(Normality of error distribution)

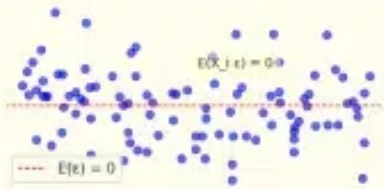


4. Independence
(of observations. Includes "no autocorrelation")

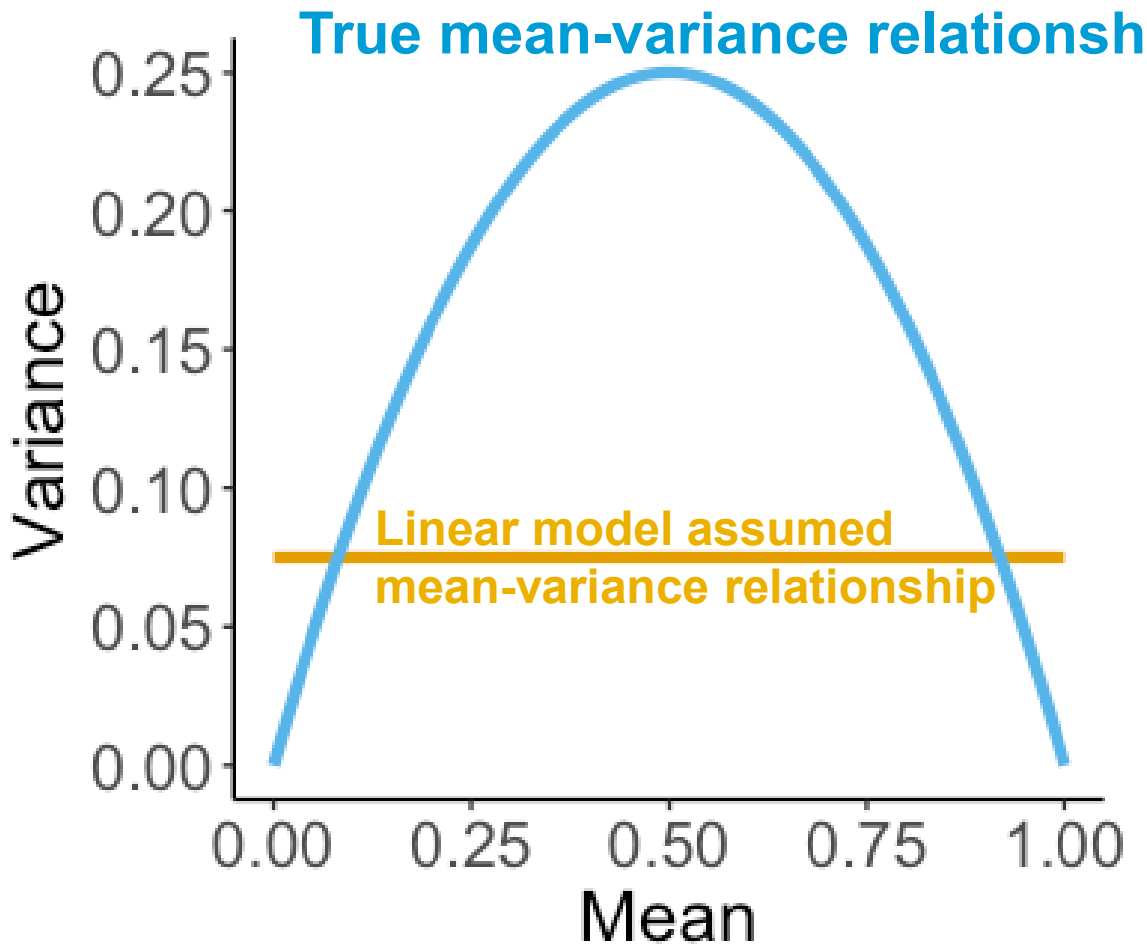
5. Lack of Multicollinearity
(Predictors are not correlated with each other)

$X_1 + X_2$ $X_1 \sim X_2$

6. Absence of endogeneity
(No correlation between predictors and errors.)

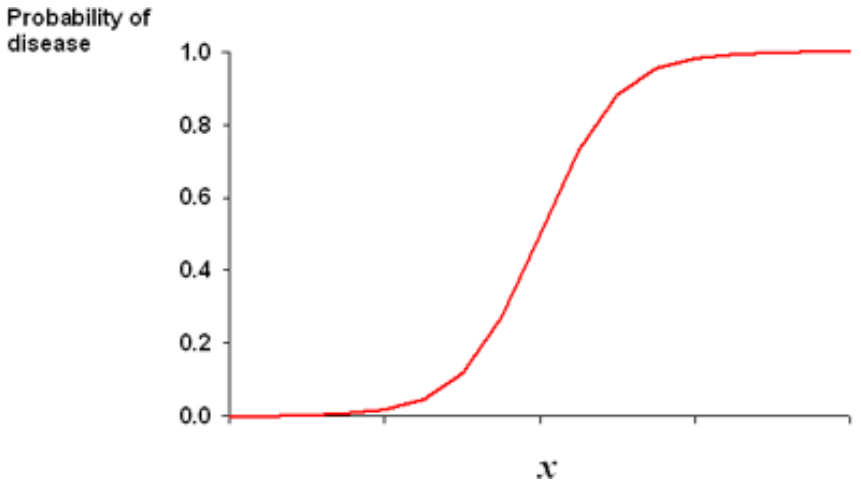


Using linear regression for binary phenotypes (coded as 0 and 1) can lead to inflated type I errors

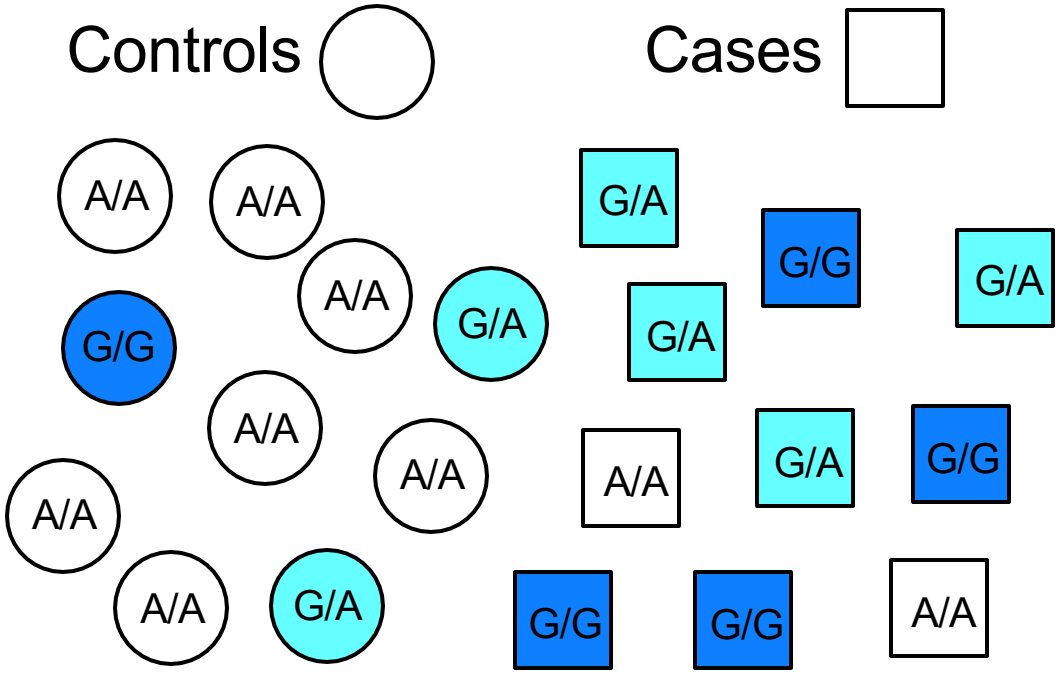


adapted from Chen, H., Wang, C., *et. al.* (2016)

Case-Control



Logistic Regression



$\ln(P/1-P) = \alpha + \beta X + \varepsilon$

β = difference in log odds for cases vs. controls

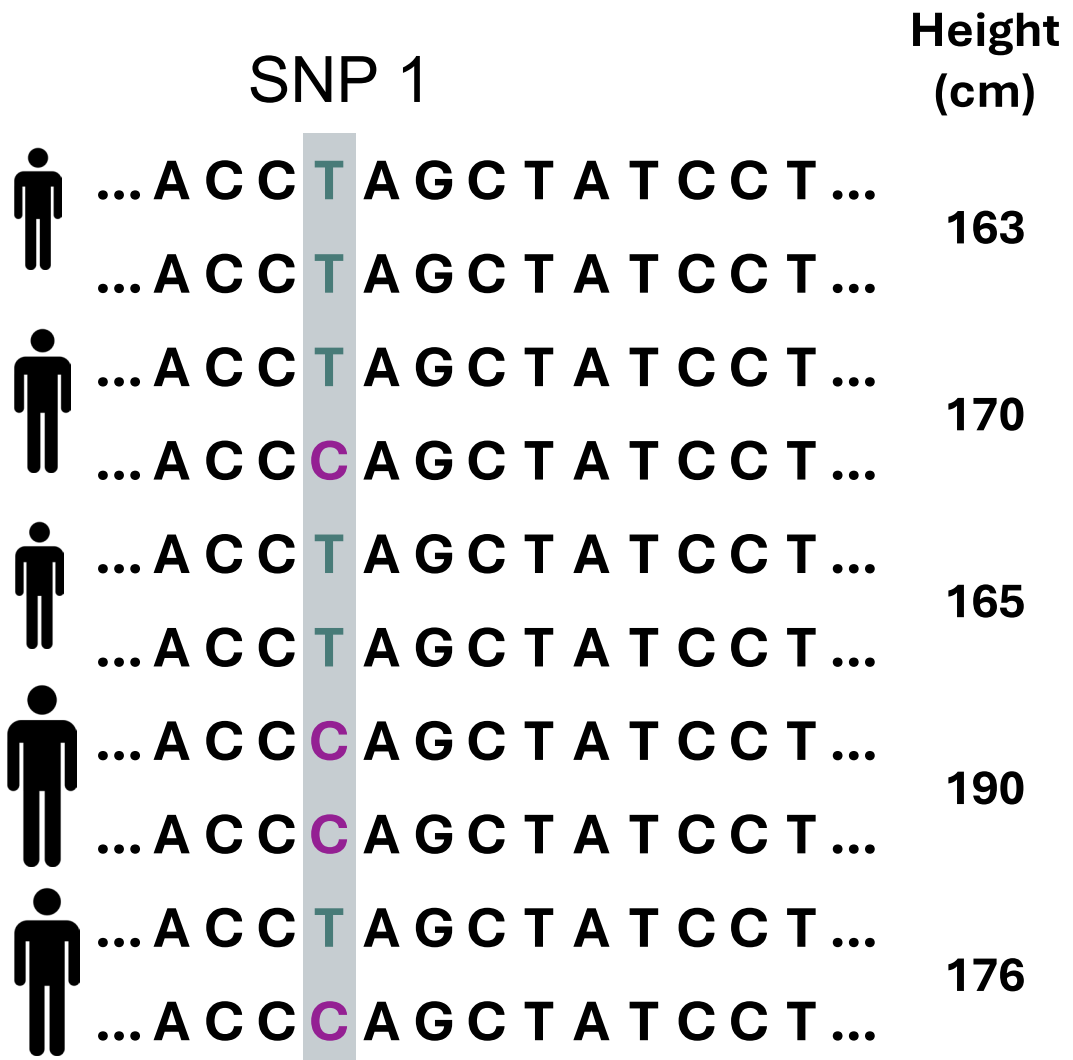
$e^{(\beta)}$ = difference in odds
= Odd Ratio (OR)

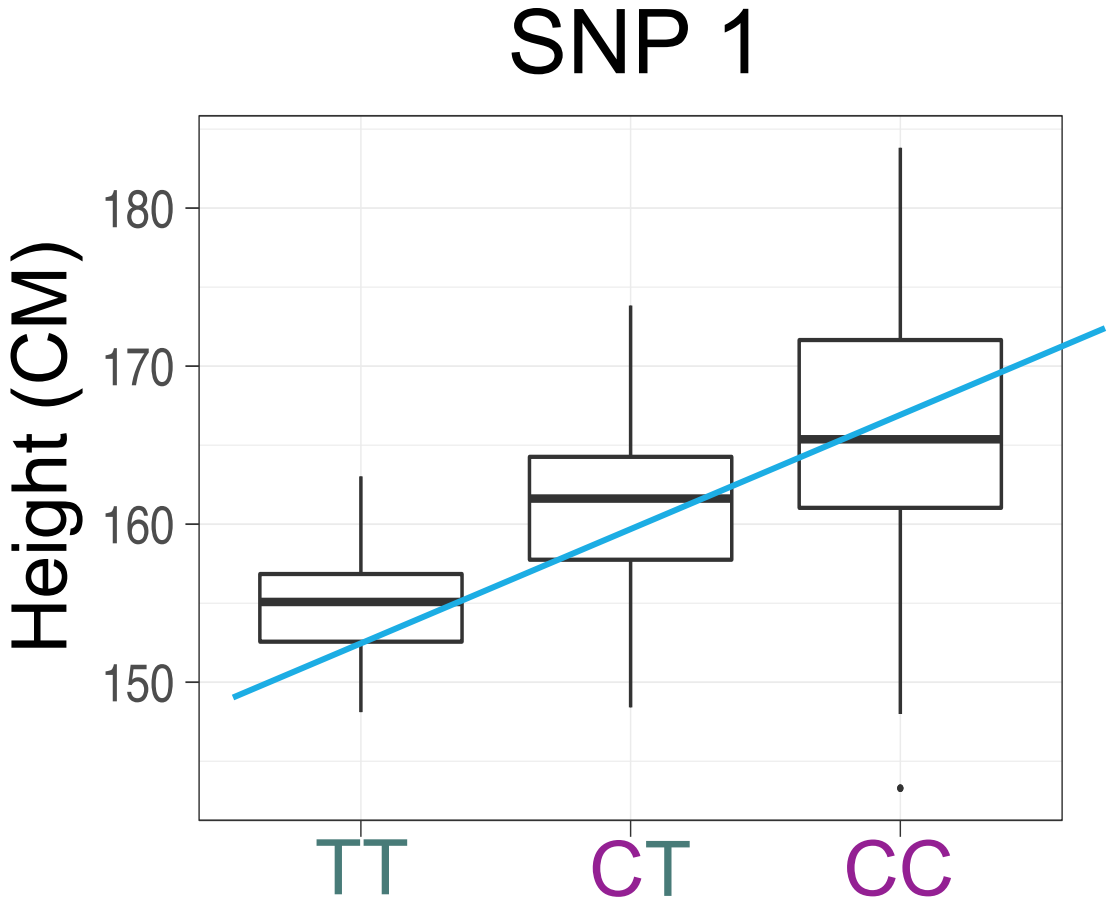
Allelic effect is an OR:
OR > 1 increased risk
OR < 1 decreased risk

The G allele is associated with disease

Genetic association tests

To identify genetic variants that are associated with a complex trait





Linear regression: $Y = X\alpha + G\beta + \epsilon$

G = 0, 1, or 2

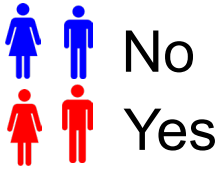
X: covariates, e.g. age, sex, ancestry, batch…

H_0: $\beta = 0$

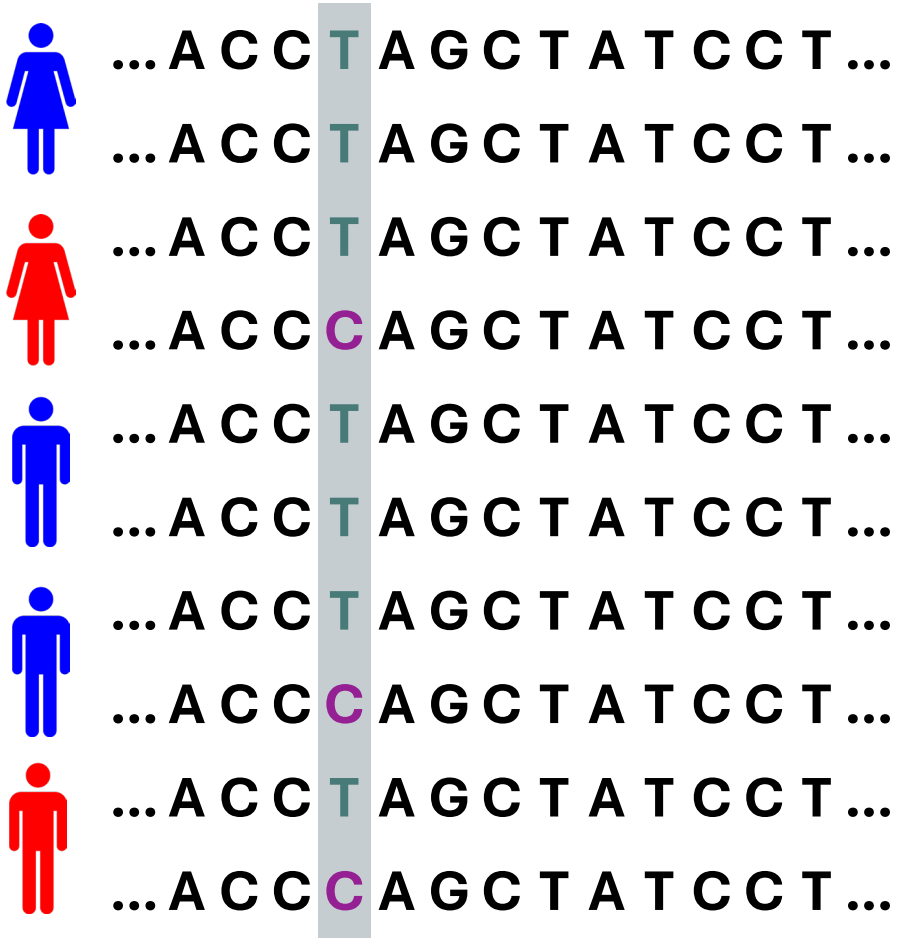
H_1: $\beta \neq 0$

Genetic association tests

To identify genetic variants that are associated with a complex disease/disorder

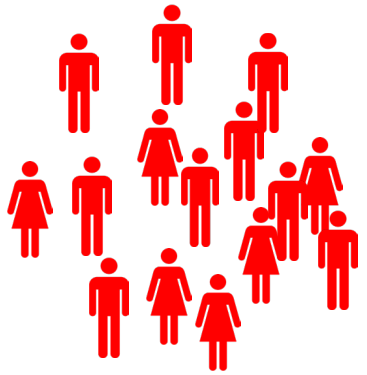


SNP 1

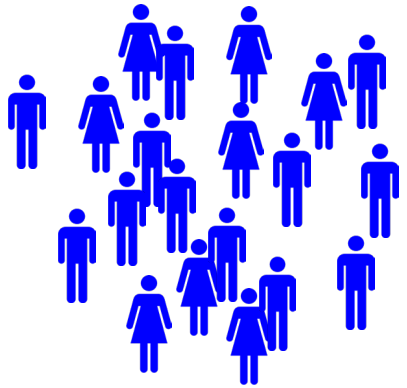


SNP 1

Frequency of C



VS.



Cases

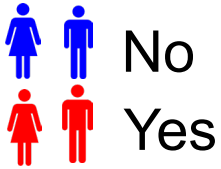
2/4 = 50%

Controls:

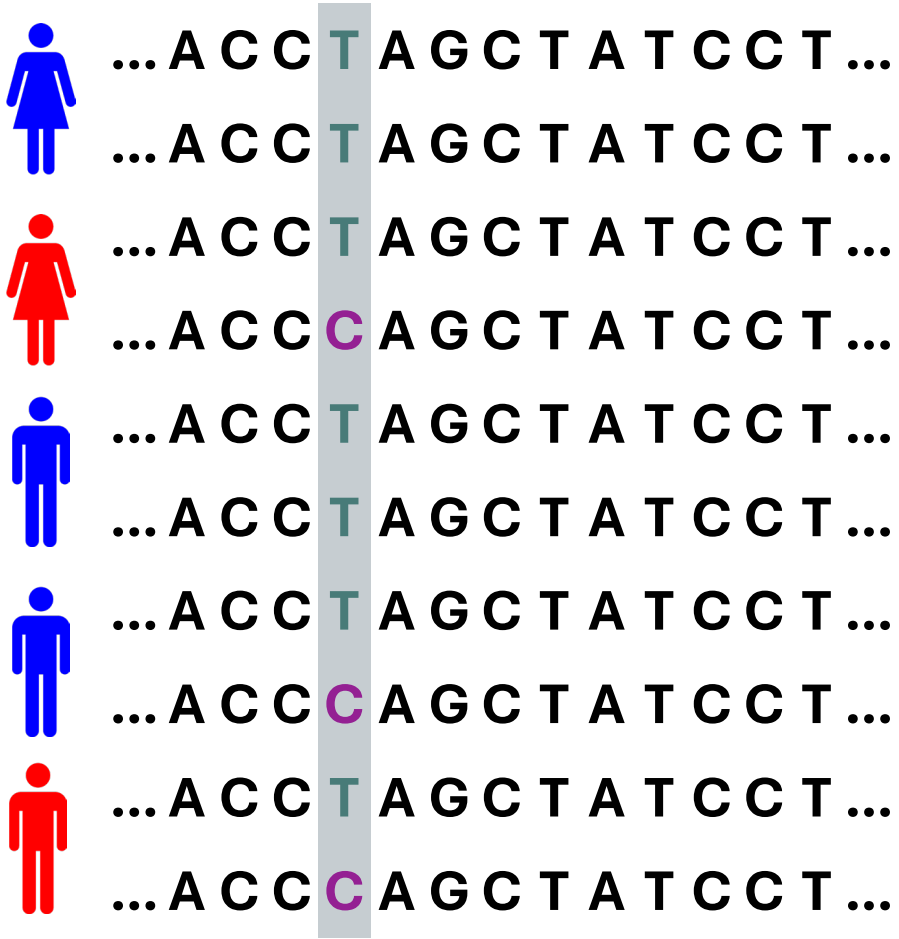
1/6 = 16.7%

Genetic association tests

To identify genetic variants that are associated with a complex disease/disorder

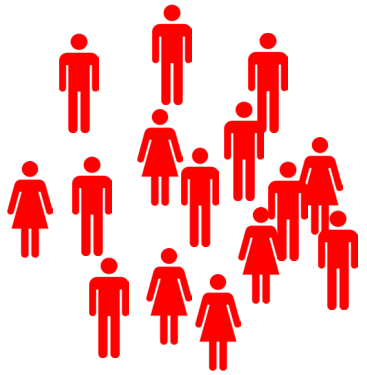


SNP 1

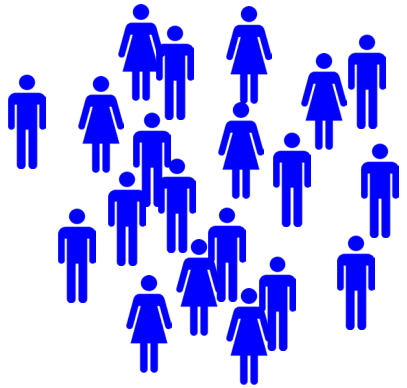


SNP 1

Frequency of C



VS.



Cases

2/4 = 50%

Controls:

1/6 = 16.7%

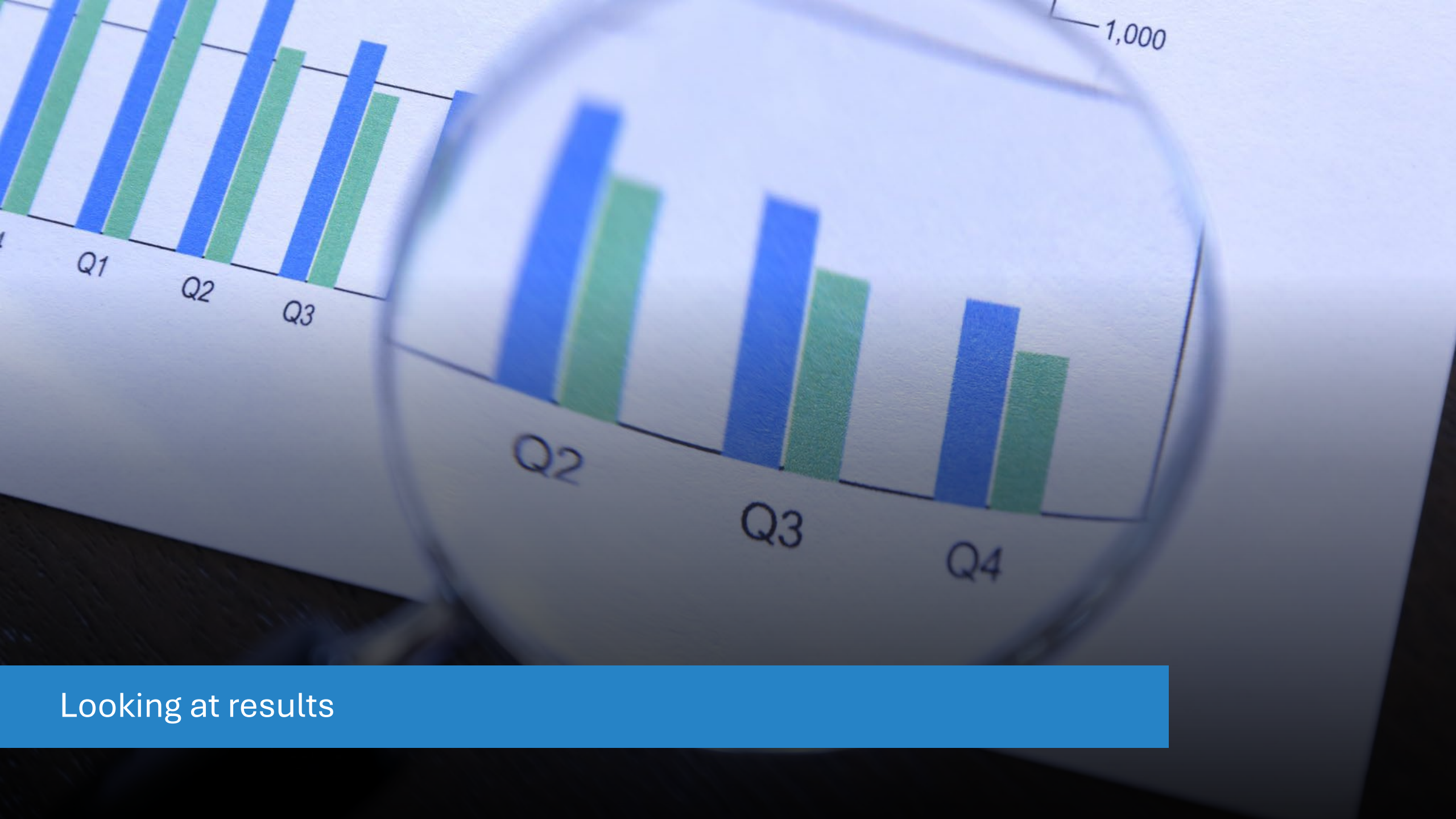
Logistic regression: $logit(\pi) = X\alpha + G\beta$

π: probability of being a case given X and G

G = 0, 1, or 2

H_0: $\beta = 0$

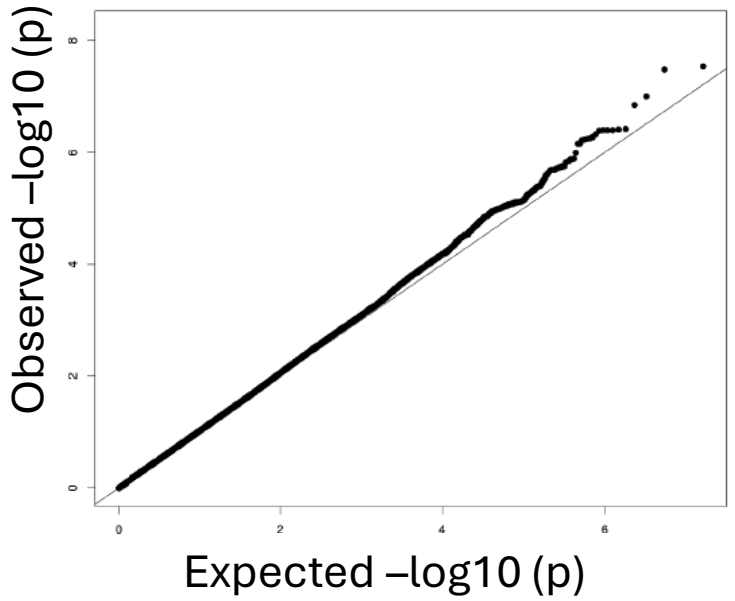
H_1: $\beta \neq 0$



Looking at results

QQ (quantile-quantile) plot

- Checks the overall distribution of test statistics or –log10 p-values of our results with the expectation under the null hypothesis of no association (the diagonal line shows where the points should fall under the null).
- Evaluates systematic bias and inflation (undetected sample duplications, unknown familial relationships, gross population stratification, problems in QC…).



Multiple testing

$p < 5 \times 10^{-8}$

Genetic Epidemiology 32: 227–234 (2008)

Estimation of Significance Thresholds for Genomewide Association Scans

Frank Dudbridge* **and Arief Gusnanto**

MRC Biostatistics Unit, Institute for Public Health, Cambridge, United Kingdom

Genetic Epidemiology 32: 381–385 (2008)

Brief Report

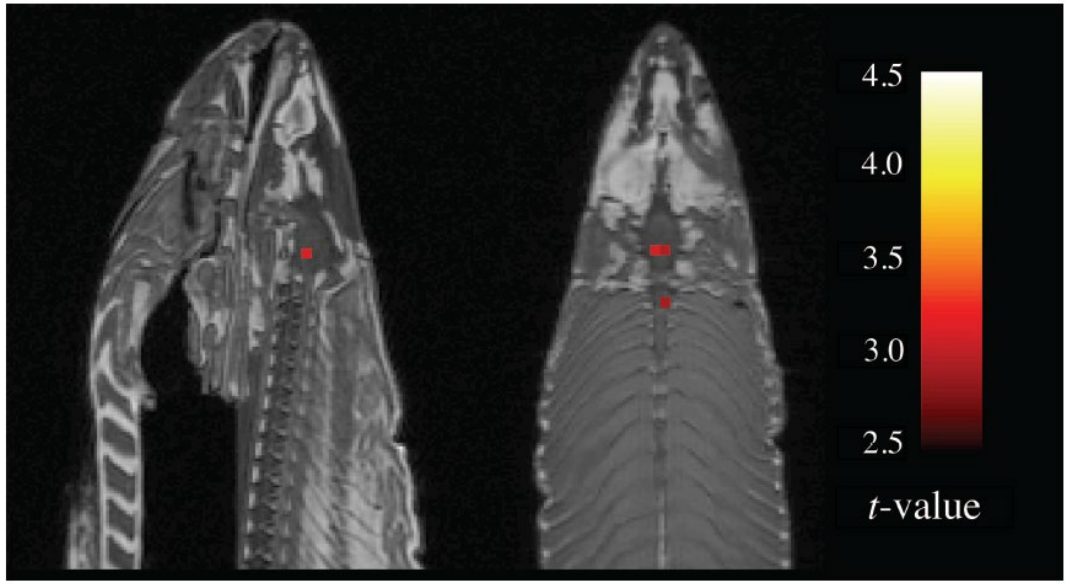
Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants

Itsik Pe'er,[1] Roman Yelensky,[2–4] David Altshuler,[2,3,5–7] and Mark J. Daly[2,5,8]*

Journal of Serendipitous and Unexpected Results

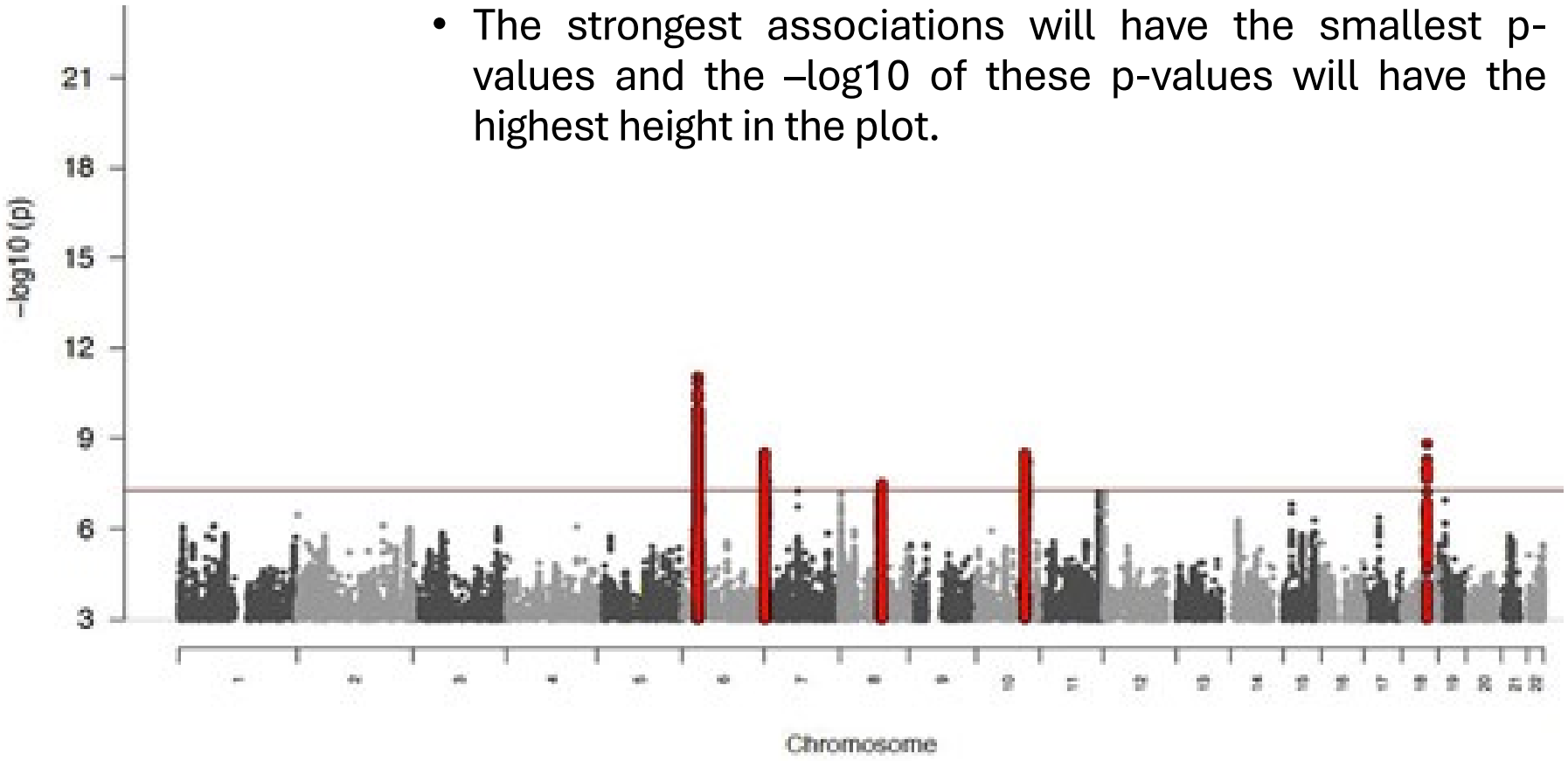
Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett[1*], Abigail A. Baird[2], Michael B. Miller[1] and George L. Wolford[3]

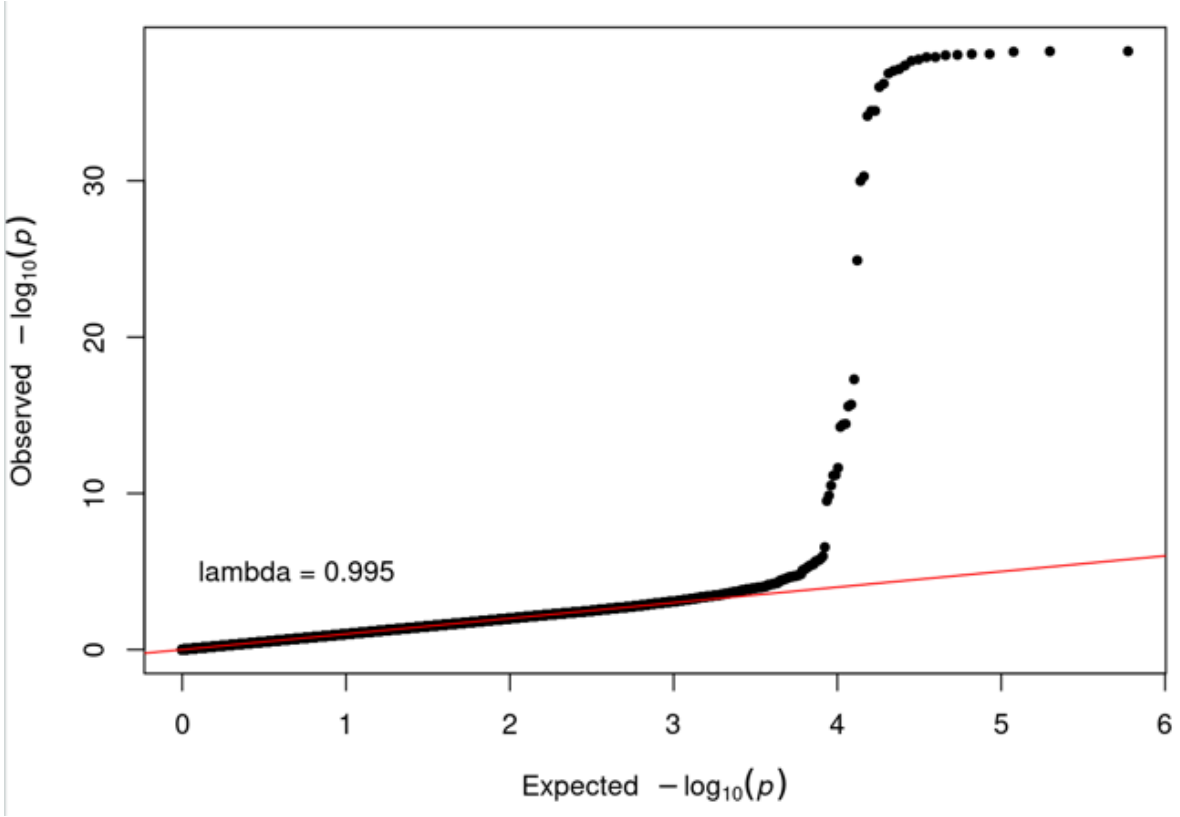


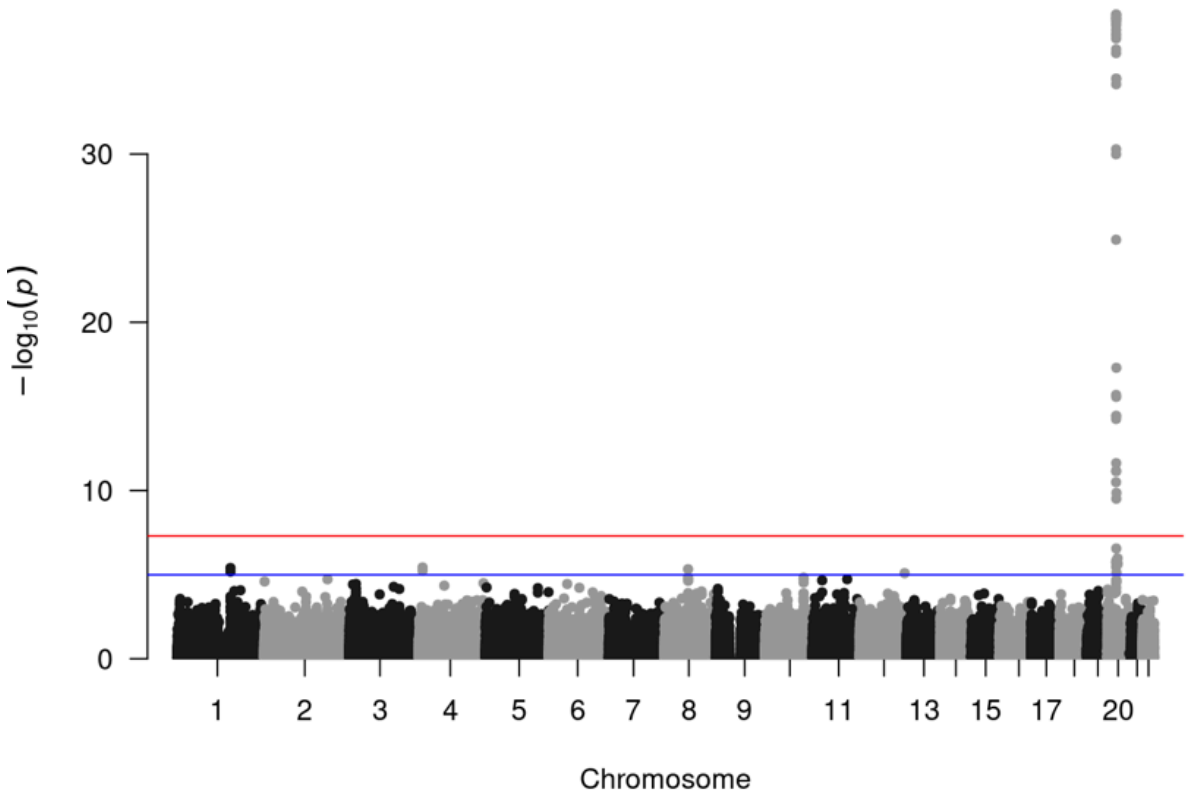
Manhattan plot

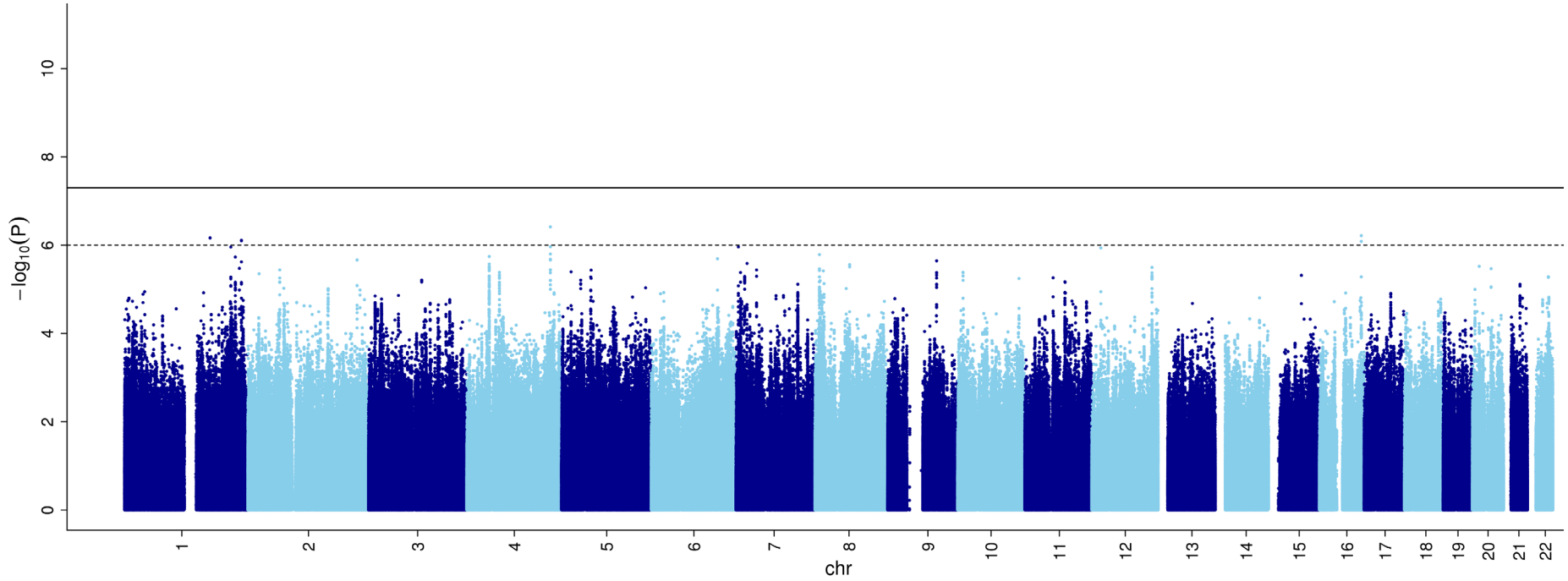
- Plots the -log10 of the association p-value for each SNP against the genomic coordinates.
- The strongest associations will have the smallest p-values and the –log10 of these p-values will have the highest height in the plot.

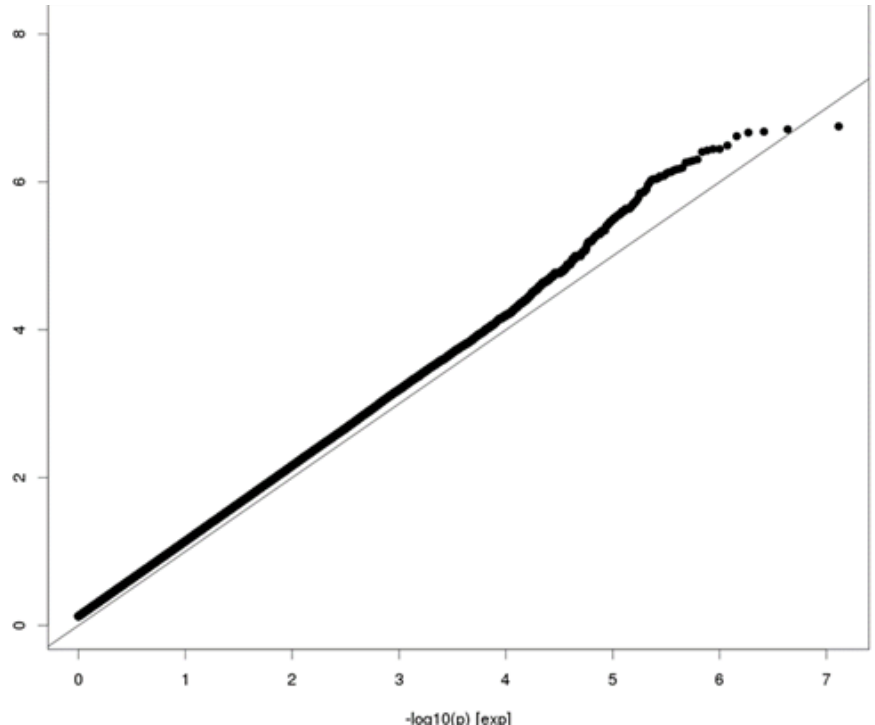


What are these telling us?

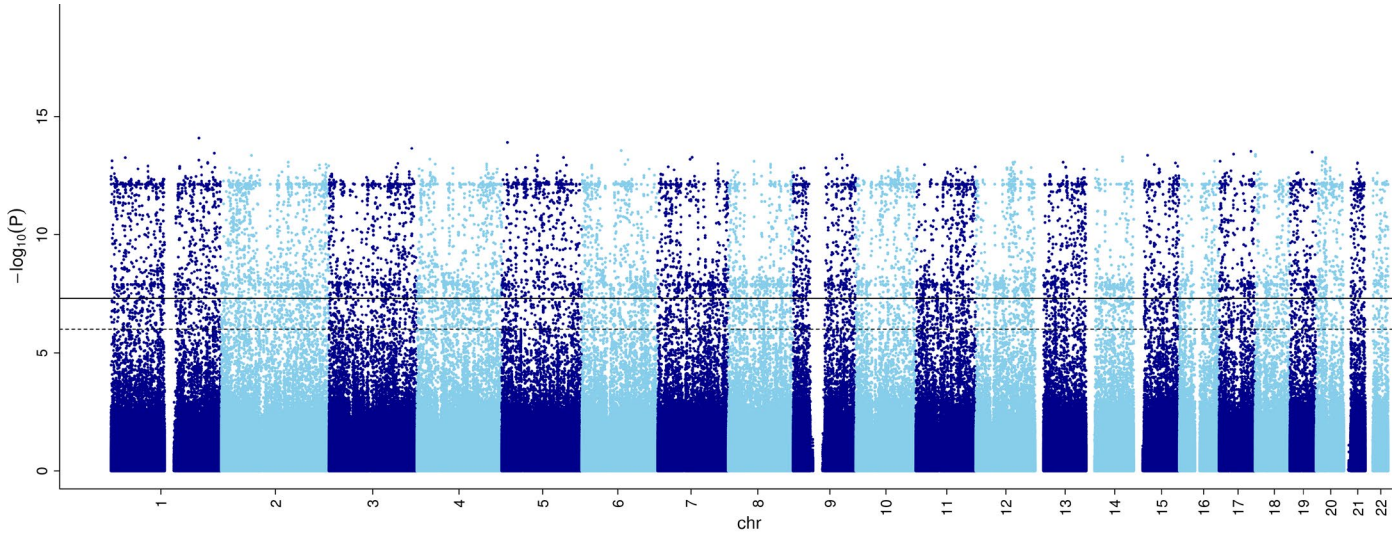


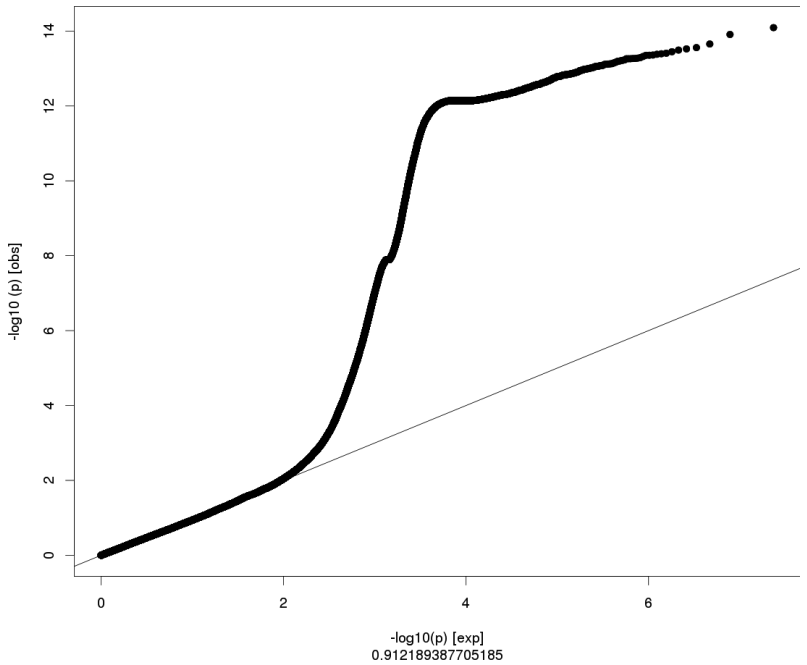



−log10(P)
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
chr


0
2
4
6
8
0
1
2
3
4
5
6
7
-log10(p) [exp]

And these? What are these telling us?



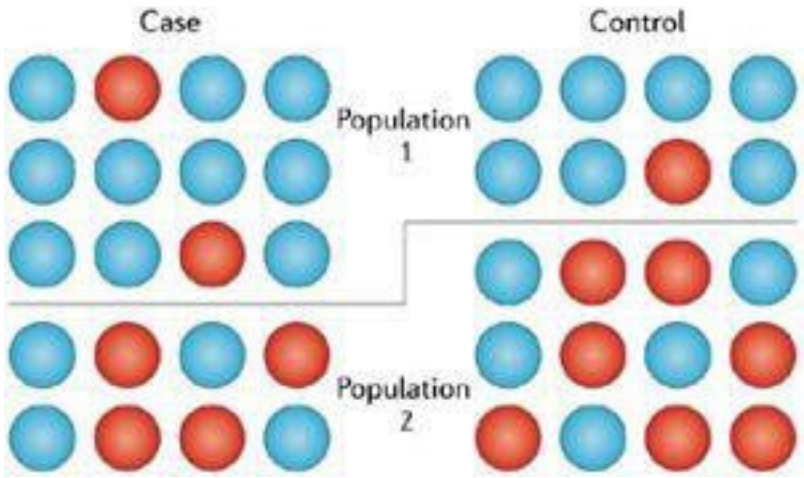


There be dragons…

Confounders

- Population Stratification

Mean trait or case frequency differences between populations



Alleles with frequency differences between populations

False positive / negative associations

Balding. *Nat Rev Genet* (2006)

Analyzing X Chromosome

More to come later (Medland)

- Often overlooked, but important to analyze, too
- Can analyze it, you just have to do it a little differently than the autosomes
- Imputation can be done and servers understand the different chromosomes and regions on them
- Dosage differences between sexes, dosage compensation and X inactivation are all important features
- X inactivation varies for different tissues

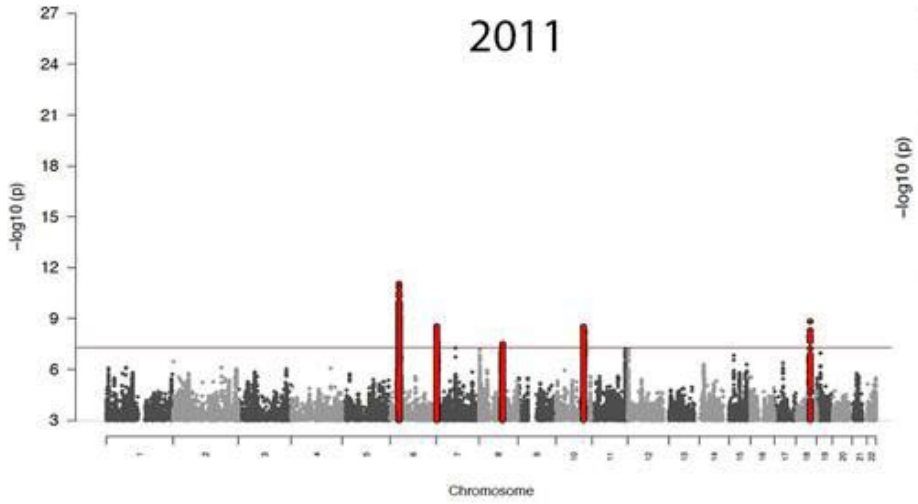
Briefings in Bioinformatics, 2022, **23(5)**, 1–9
https://doi.org/10.1093/bib/bbac287
Review

A systematic review of analytical methods used in genetic association analysis of the X-chromosome

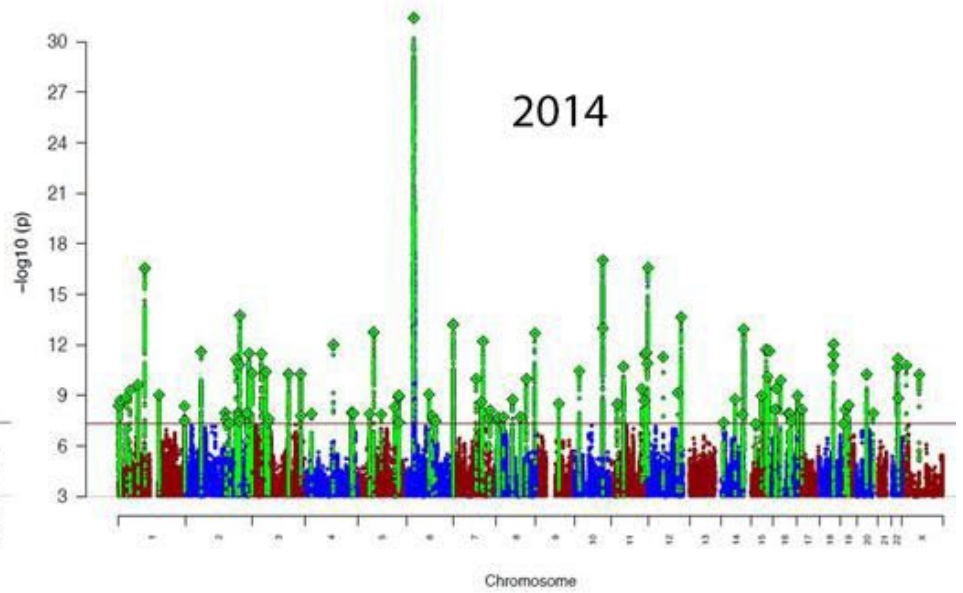
Nick Keur, Isis Ricaño-Ponce, Vinod Kumar and Vasiliki Matzaraki
Corresponding author. Vasiliki Matzaraki, E-mail: Vasiliki.Matzaraki@radboudumc.nl

Sample Size & Power

Schizophrenia Working Group of the Psychiatric Genomics Consortium.



9,394 cases
12,462 controls



36,989 cases
113,075 controls

Power Calculation Tools

Consider: Effect size, Sample size, Prevalence, MAF

Purcell, Cherny, & Sham. *Bioinformatics,* 2003

http://zzz.bwh.harvard.edu/gpc/

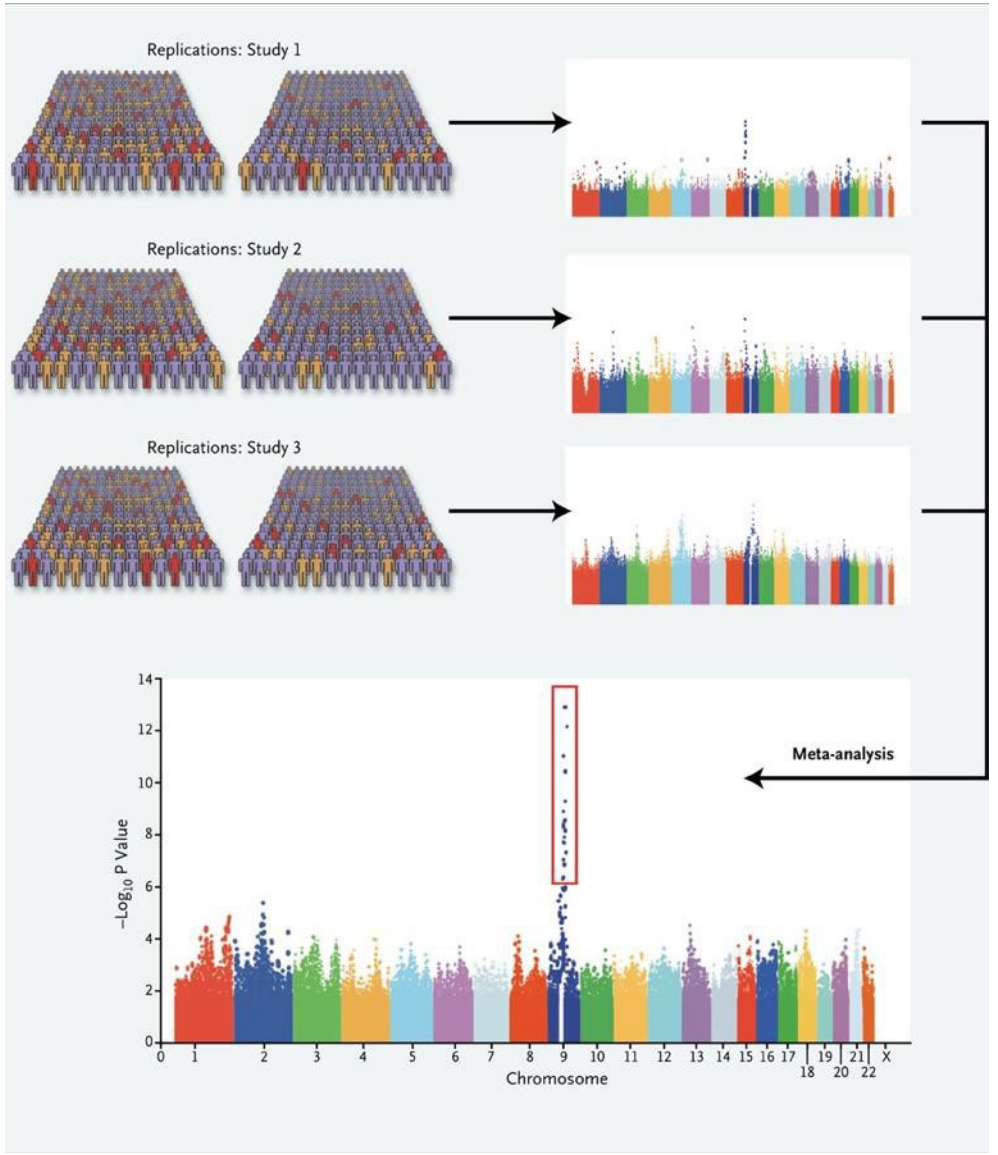
Johnson & Abecasis. *bioRxiv*, 2017

https://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html

Replication

1. Significance
2. Size
3. Direction

Meta-analysis has mostly replaced replication…
More on this this afternoon



Manolio. *N Engl J Med*, 2010

There are many tools for GWAS

- The most appropriate method depends on the structure of your data: stratification, relatedness Sample size Etc…
- We’re going to use plink2 for today’s practical: fast, simple command line program that is a general workhorse software for managing data and running analyses.

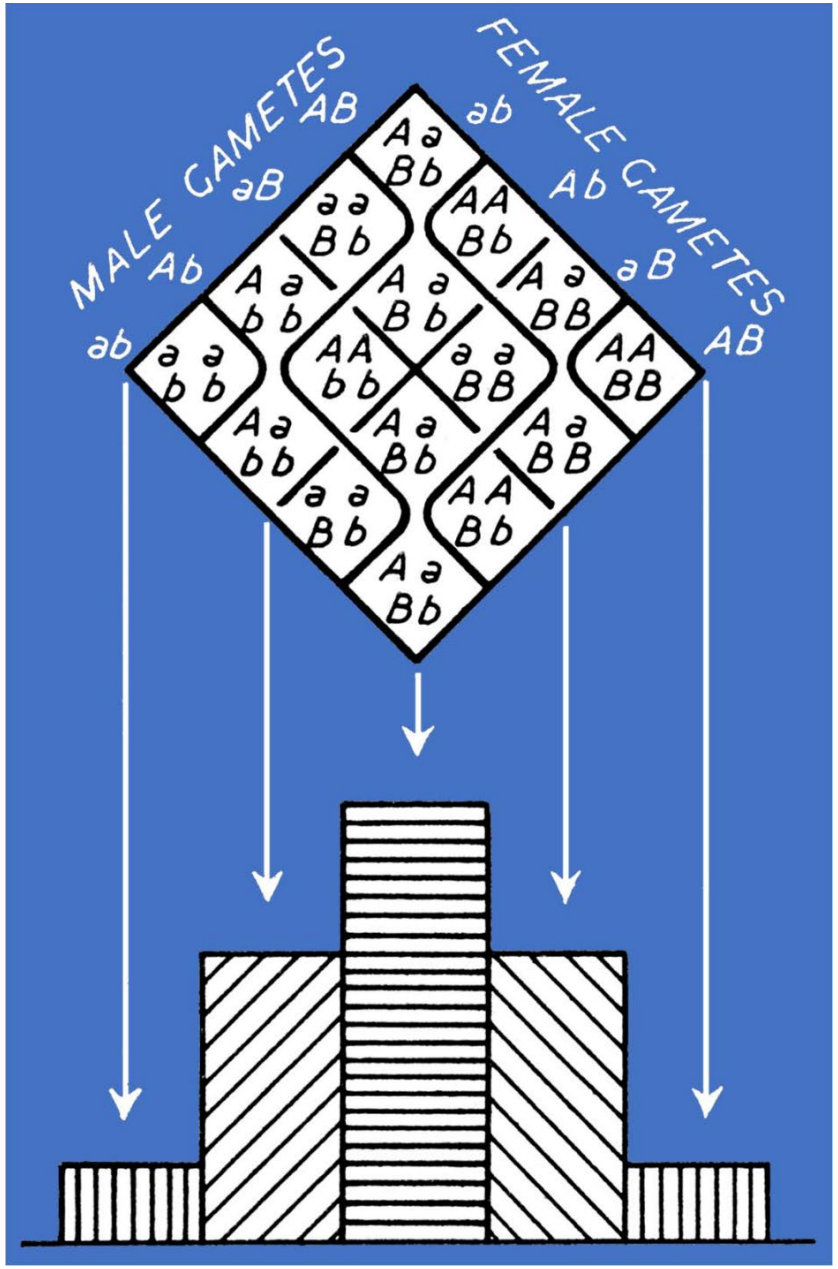
  - Original plink ped/map format
  - Binary plink format (plink1.9): bed/bim/fam
    - Fast, very useful
  - Plink2 format: pgen/psam/pvar
    - Very fast
    - Saves space compared to others
    - Can include dosage, phase, INFO (similar to VCF format)

Today’s practical is on qualtrics:

https://qimr.az1.qualtrics.com/jfe/form/SV_8IWskkB41ezAMlg

Link is in Qualtrics.txt

Association with family data

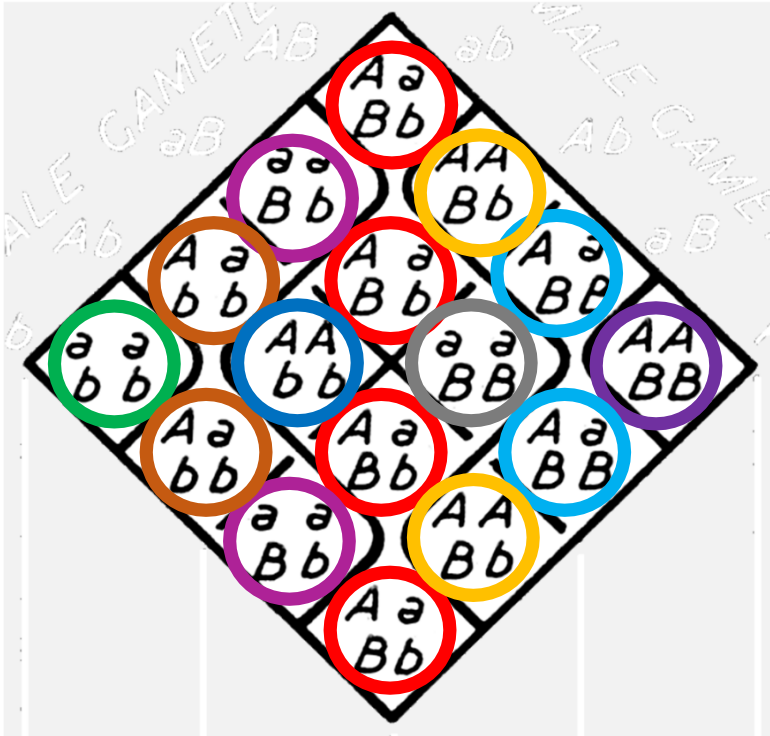


Punnett square

Two genes A and B. Parents are both heterozygotes (AaBb).

Their offspring may have different genotypes.

K Mather, Biometrical Genetics, Dover Publ, 1949



In the population traits of e.g. ab/ab individuals differ from the phenotypes of AB/AB individuals.

Do we see the same differences if these two individuals are siblings?

I.e., is variation within families equal to variation between families?

If yes: “true” genetic association

If no: ? (confounding)

Lindon Eaves
(e.g. Inferring the Causes of Human Variation, 1977)

The genetic and environmental variation is partitioned into within and between family components.

G1=*within* - family genetic component

G2=*between* - family genetic component

E1=*within*-family environment ("E")

E2=*between*-family environment ("C")

In the absence of GE interaction or GE correlation total variance is partitioned into: $\sigma^2 t = \sigma^2 w + \sigma^2 b$, and familial resemblance is: $ICC = \sigma^2 b / (\sigma^2 w + \sigma^2 b)$

AJHG

Volume 64, Issue 1, January 1999, Pages 259-267

Combined Linkage and Association Sib-Pair Analysis for Quantitative Traits

D.W. Fulker [1 2], S.S. Cherny [1 2], P.C. Sham [2], J.K. Hewitt [1]

Summary

An extension to current maximum-likelihood variance-components procedures for mapping quantitative-trait loci in sib pairs that allows a simultaneous test of allelic association is proposed. The method involves modeling of the allelic means for a test of association, with simultaneous modeling of the sib-pair covariance structure for a test of linkage. By partitioning of the mean effect of a locus into between- and within-sibship components, the method controls for spurious associations due to population stratification and admixture. The power and efficacy of the method are illustrated through simulation of various models of both real and spurious association.

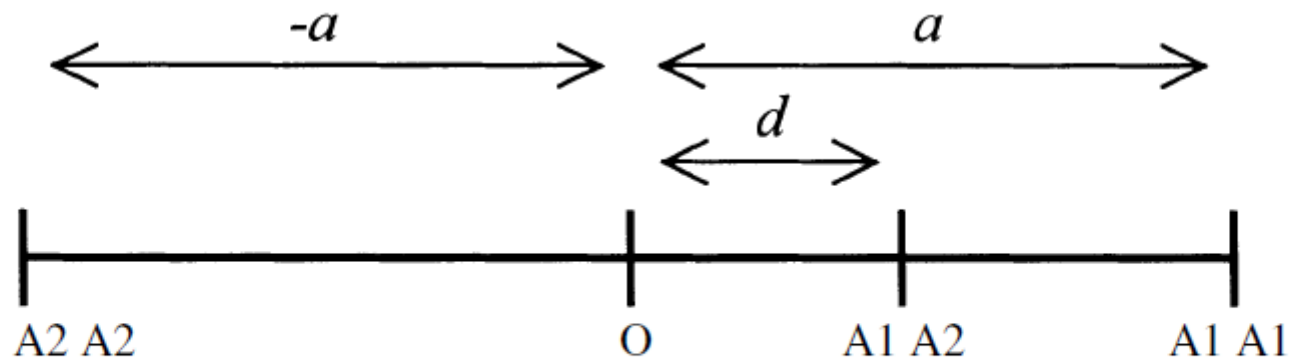


Figure 1

Graphical illustration of the genotypic values for a diallelic locus.

Table 1

Summary of Genotypic Values, Frequencies, and Dominance Deviation for Three Genotypes A1A1, A1A2, and A2A2

Genotype	A1A1	A1A2	A2A2
Genotypic value	a	d	$-a$
Frequency	p^2	$2pq$	q^2
Frequency x value	$a\,p^2$	$2dpq$	$-a\,q^2$
Deviation from the population mean	$2q(a - dp)$	$a(q - p) + d(1 - 2pq)$	$-2p(a + dq)$
Dominance deviation	$-2q^2d$	$2dpq$	$-2p^2d$

Table I. Expected Sib-Pair Means and Differences and Their Frequencies for a Single Additive Two-Allele Locus

Genotype		Additive effects				
Sib 1	Sib 2	Sib 1	Sib 2	Mean	Difference/2	Frequency
A_1A_1	A_1A_1	a	a	a	0	$p^4 + p^3q + (p^2q^2/4)$
A_1A_1	A_1A_2	a	0	$a/2$	$a/2$	$p^3q + (p^2q^2/2)$
A_1A_1	A_2A_2	a	$-a$	0	a	$p^2q^2/4$
A_1A_2	A_1A_1	0	a	$a/2$	$-a/2$	$p^3q + (p^2q^2/2)$
A_1A_2	A_1A_2	0	0	0	0	$p^3q + 3p^2q^2 + pq^3$
A_1A_2	A_2A_2	0	$-a$	$-a/2$	$a/2$	$(p^2q^2/2) + pq^3$
A_2A_2	A_1A_1	$-a$	a	0	$-a$	$p^2q^2/4$
A_2A_2	A_1A_2	$-a$	0	$-a/2$	$-a/2$	$(p^2q^2/2) + pq^3$
A_2A_2	A_2A_2	$-a$	$-a$	$-a$	0	$(p^2q^2/4) + pq^2 + q^4$

Article | Open access | Published: 09 May 2022

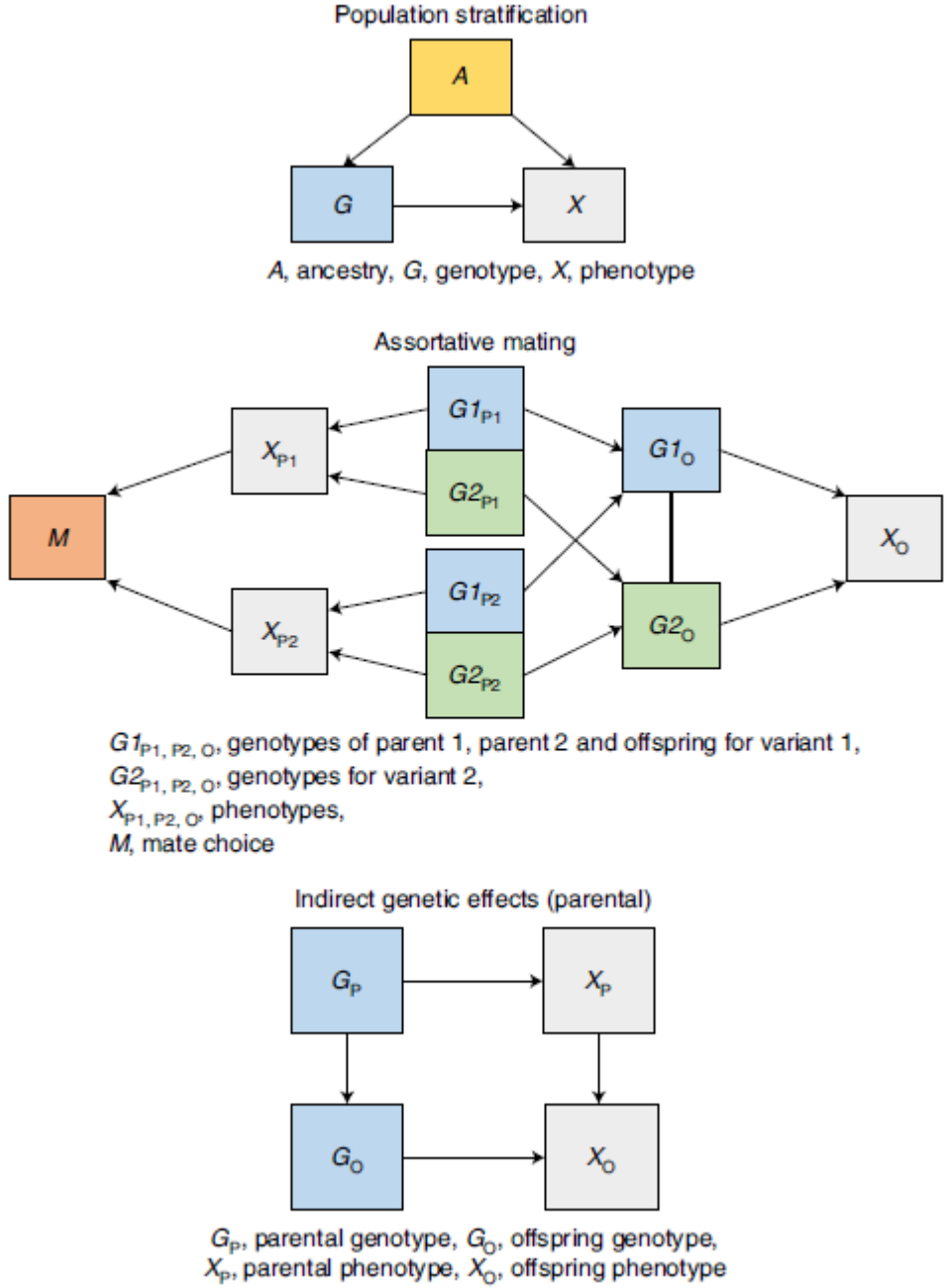
Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects

Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A. Lind, Teemu Palviainen, Matthijs D. van der Zee, Rosa Cheesman, Massimo Mangino, Yunzhang Wang, Shuai Li, Lucija Klaric, Scott M. Ratliff, Lawrence F. Bielak, Marianne Nygaard, Alexandros Giannelis, Emily A. Willoughby, Chandra A. Reynolds, Jared V. Balbona, Ole A. Andreassen, Social Science Genetic Association Consortium, Within Family Consortium, ... Neil M. Davies

+ Show authors

Nature Genetics **54**, 581–592 (2022)

Fig. 1 | Demographic and indirect genetic effects. Population stratification: population stratification is defined as the distortion of associations between a genotype and a phenotype when ancestry *A* influences both genotype *G* (via differences in allele frequencies) and the phenotype *X*. Principal components and linear mixed model methods control for ancestry but they may not completely control for fine-scale population structure. Assortative mating: assortative mating is a phenomenon where individuals select a partner based on phenotypic (dis)similarities. For example, tall individuals may prefer a tall partner. Assortative mating can induce correlations between causes of an assorted phenotype in subsequent generations. If a phenotype *X* is influenced by two independent genetic variants *G1* and *G2* then assortment on *X* (represented by effects of *X* on mate choice *M*) will induce positive correlations between *G1* in parent 1 and *G2* in parent 2 and vice versa. Parental transmission will then induce correlations between otherwise independent *G1* and *G2* in offspring. These correlations can distort genetic association estimates. Indirect genetic effects: indirect genetic effects are effects of relative genotypes (via relative phenotypes and the shared environment) on the index individual's phenotype. These indirect effects influence population GWAS estimates because relative genotypes are also associated with genotypes of the index individual. Indirect genetic effects of parents on offspring are of most interest because they are likely to be the largest. However, indirect genetic effects of siblings or more distal relatives are also possible.



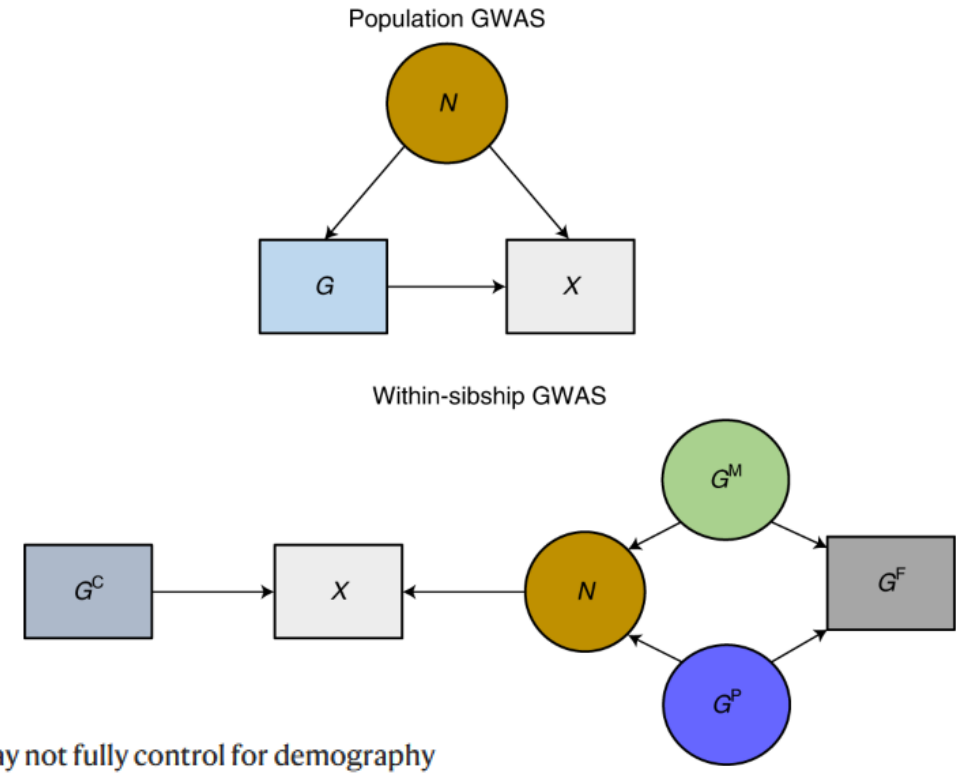
Article | Open access | Published: 09 May 2022

Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects

Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A. Lind, Teemu Palviainen, Matthijs D. van der Zee, Rosa Cheesman, Massimo Mangino, Yunzhang Wang, Shuai Li, Lucija Klaric, Scott M. Ratliff, Lawrence F. Bielak, Marianne Nygaard, Alexandros Giannelis, Emily A. Willoughby, Chandra A. Reynolds, Jared V. Balbona, Ole A. Andreassen, Social Science Genetic Association Consortium, Within Family Consortium, … Neil M. Davies

+ Show authors

Nature Genetics **54**, 581–592 (2022) | Cite this article



As outlined in Fig. 1, estimates from population GWAS may not fully control for demography (population stratification and assortative mating) and may also capture indirect genetic effects of relatives. For simplicity we use *N* to represent all sources of associations between *G* and *X* that do not relate to direct effects of *G*. Circles indicate unmeasured variables and squares indicate measured variables. If parental genotypes are known, *G* can be separated into nonrandom (determined by parental genotypes) and random (relating to segregation at meiosis) components. Within-sibship GWAS include the mean genotype across a sibship (G^F) (a proxy for the mean of the paternal and maternal genotypes $G^{P,M}$) as a covariate to capture associations between *G* and *X* relating to parents. The within-sibship estimate is defined as the effect of the random component: that is, the association between family-mean-centered genotype G^C (that is, $G - G^F$) and *X*. Demography and indirect genetic effects of parents (*N*) will be captured by G^F. The association between G^C and *X* will not be influenced by these sources of association but could be affected by indirect effects of the siblings themselves, which are not controlled for.

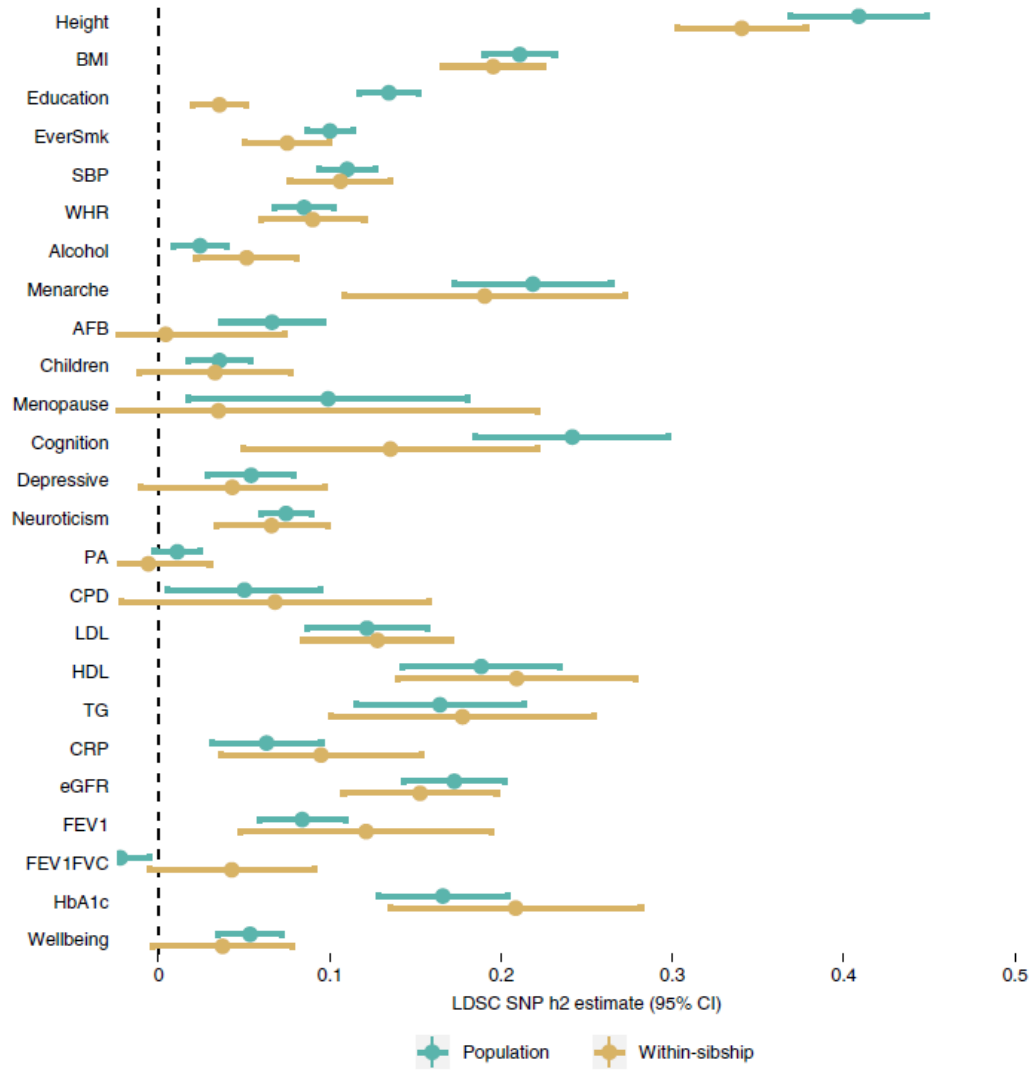
Article | Open access | Published: 09 May 2022

Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects

Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A. Lind, Teemu Palviainen, Matthijs D. van der Zee, Rosa Cheesman, Massimo Mangino, Yunzhang Wang, Shuai Li, Lucija Klaric, Scott M. Ratliff, Lawrence F. Bielak, Marianne Nygaard, Alexandros Giannelis, Emily A. Willoughby, Chandra A. Reynolds, Jared V. Balbona, Ole A. Andreassen, Social Science Genetic Association Consortium, Within Family Consortium, ... Neil M. Davies

+ Show authors

Nature Genetics **54**, 581–592 (2022) | Cite this article



Family-GWAS reveals effects of environment and mating on genetic associations

Tammy Tan, Hariharan Jayashankar, Junming Guan, Seyed Moeen Nehzati, Mahdi Mir, Michael Bennett, Esben Agerbo, Rafael Ahlskog, Ville Pinto de Andrade Anapaz, Bjørn Olav Åsvold, Stefania Benonisdottir, Laxmi Bhatta, Dorret I. Boomsma, Ben Brumpton, Archie Campbell, Christopher F. Chabris, Rosa Cheesman, Zhengming Chen, China Kadoorie Biobank Collaborative Group, Eco de Geus, Erik A. Ehli, Abdelrahman G. Elnahas, Estonian Biobank Research Team, Finngen, Andrea Ganna, Alexandros Giannelis, Liisa Hakaste, Ailin Falkmo Hansen, Alexandra Havdahl, Caroline Hayward, Jouke-Jan Hottenga, Mikkel Aagaard Houmark, Kristian Hveem, Jaakko Kaprio, Arnulf Langhammer, Antti Latvala, James J. Lee, Mikko Lehtovirta, Liming Li, LifeLines Cohort Study, Kuang Lin, Richard Karlsson Linnér, Stefano Lombardi, Nicholas G. Martin, Matt McGue, Sarah E. Medland, Andres Metspalu, Brittany L. Mitchell, Guiyan Ni, Ilja M. Nolte, Matthew T. Oetjens, Sven Oskarsson, Teemu Palviainen, Rashmi B. Prasad, Anu Reigo, Kadri Reis, Julia Sidorenko, Karri Silventoinen, Harold Snieder, Tiinamaija Tuomi, Bjarni J. Vilhjálmsson, Robin G. Walters, Emily A. Willoughby, Bendik S. Winsvold, Eivind Ystrom, Jonathan Flint, Loic Yengo, Peter M. Visscher, Augustine Kong, Elliot M. Tucker-Drob, Richard Border, David Cesarini, Patrick Turley, Aysu Okbay, Daniel J. Benjamin, Alexander Strudwick Young

doi: https://doi.org/10.1101/2024.10.01.24314703

Genome-wide association studies (GWAS) have discovered thousands of replicable genetic associations, guiding drug target discovery and powering genetic prediction of human phenotypes and diseases. However, genetic associations can be affected by gene-environment correlations and non-random mating, which can lead to biased inferences in downstream analyses. Family-based GWAS (FGWAS) uses the natural experiment of random assignment of genotype within families to separate out the contribution of direct genetic effects (DGEs) — causal effects of alleles in an individual on an individual — from other factors contributing to genetic associations. Here, we report results from an FGWAS meta-analysis of 34 phenotypes from 17 cohorts. We found evidence that factors uncorrelated with DGEs make substantial contributions to genetic associations for 27 phenotypes, with population stratification confounding — a form of gene-environment correlation — likely the major cause. By estimating SNP heritability and genetic correlations using DGEs, we found evidence that assortative mating has led to overestimation of SNP heritability for 5 phenotypes and overestimation of the degree of shared genetic effects (pleiotropy) between 22 pairs of phenotypes. Polygenic predictors constructed from DGEs are particularly useful for studying natural selection, assortative mating, and indirect genetic effects (effects of relatives' genes mediated through the family environment). We validate our meta-analysis results by predicting phenotypes in hold-out samples using polygenic predictors constructed from DGEs, achieving statistically significant out-of-sample prediction for 24 phenotypes with little attenuation of predictive power within-families. We provide FGWAS summary statistics for 34 phenotypes that can be used for downstream analyses. Our study provides both a template for performing FGWAS and an argument for its value for debiasing inferences and understanding the impact of environment and mating patterns.

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES
Cold Spring Harbor Laboratory BMJ Yale

Family-GWAS reveals effects of environment and mating on genetic associations

Tammy Tan, Hariharan Jayashankar, Junming Guan, Seyed Moeen Nehzati, Mahdi Mir, Michael Bennett, Esben Agerbo, Rafael Ahlskog, Ville Pinto de Andrade Anapaz, Bjørn Olav Åsvold, Stefania Benonisdottir, Laxmi Bhatta, Dorret I. Boomsma, Ben Brumpton, Archie Campbell, Christopher F. Chabris, Rosa Cheesman, Zhengming Chen, China Kadoorie Biobank Collaborative Group, Eco de Geus, Erik A. Ehli, Abdelrahman G. Elnahas, Estonian Biobank Research Team, Finngen, Andrea Ganna, Alexandros Giannelis, Liisa Hakaste, Ailin Falkmo Hansen, Alexandra Havdahl, Caroline Hayward, Jouke-Jan Hottenga, Mikkel Aagaard Houmark, Kristian Hveem, Jaakko Kaprio, Arnulf Langhammer, Antti Latvala, James J. Lee, Mikko Lehtovirta, Liming Li, LifeLines Cohort Study, Kuang Lin, Richard Karlsson Linnér, Stefano Lombardi, Nicholas G. Martin, Matt McGue, Sarah E. Medland, Andres Metspalu, Brittany L. Mitchell, Guiyan Ni, Ilja M. Nolte, Matthew T. Oetjens, Sven Oskarsson, Teemu Palviainen, Rashmi B. Prasad, Anu Reigo, Kadri Reis, Julia Sidorenko, Karri Silventoinen, Harold Snieder, Tiinamaija Tuomi, Bjarni J. Vilhjálmsson, Robin G. Walters, Emily A. Willoughby, Bendik S. Winsvold, Eivind Ystrom, Jonathan Flint, Loic Yengo, Peter M. Visscher, Augustine Kong, Elliot M. Tucker-Drob, Richard Border, David Cesarini, Patrick Turley, Aysu Okbay, Daniel J. Benjamin, Alexander Strudwick Young

doi: https://doi.org/10.1101/2024.10.01.24314703

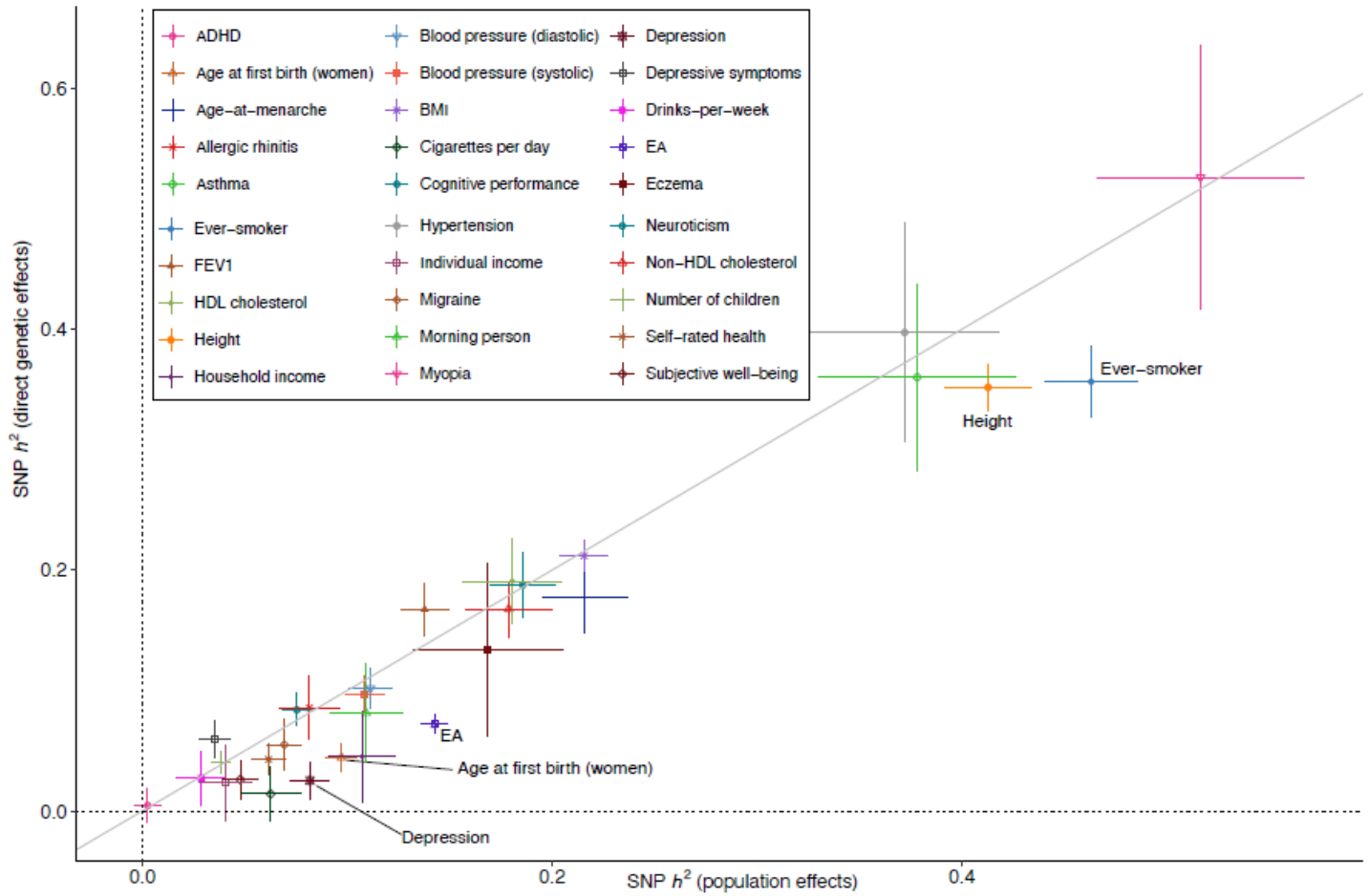
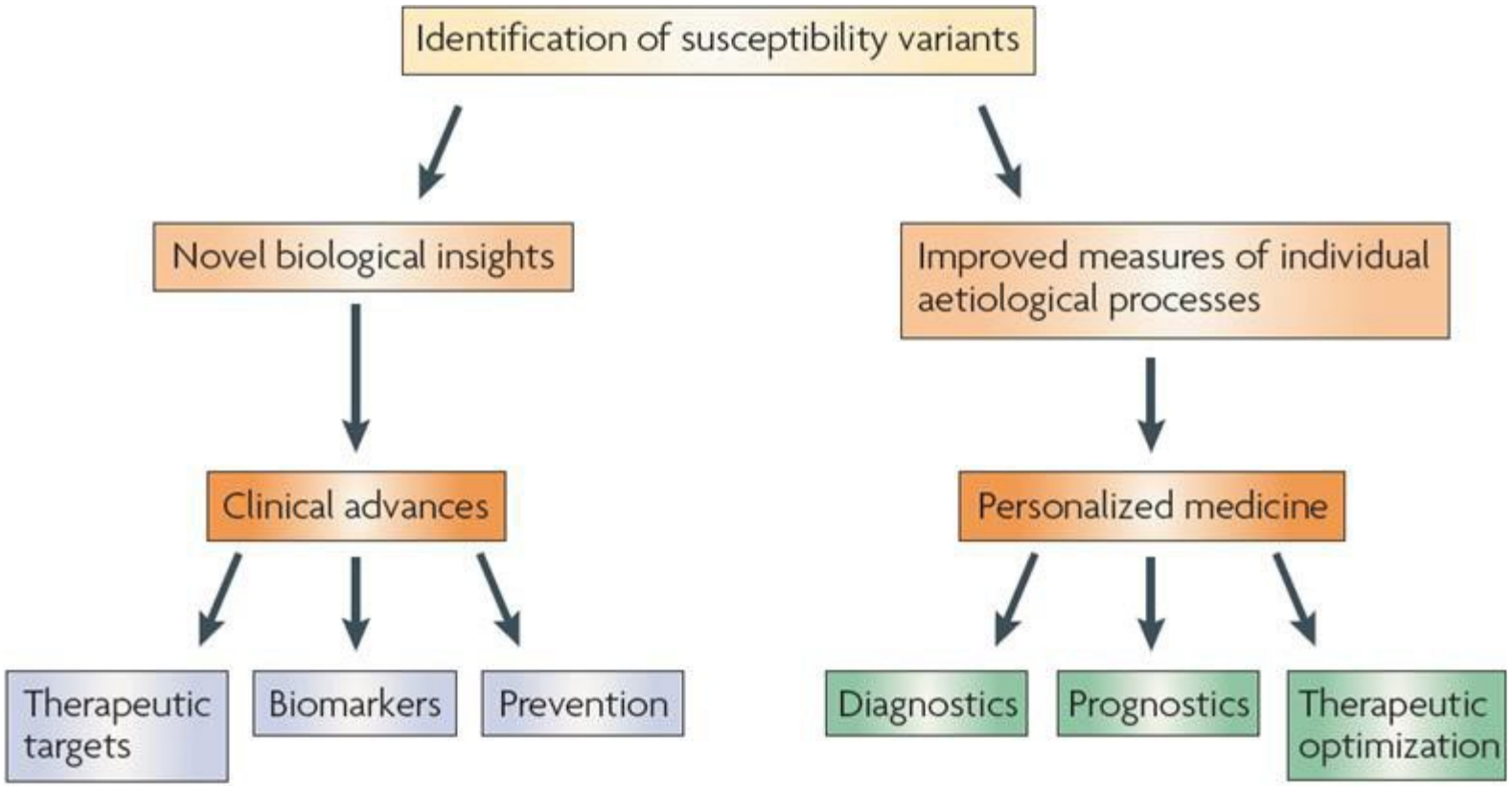
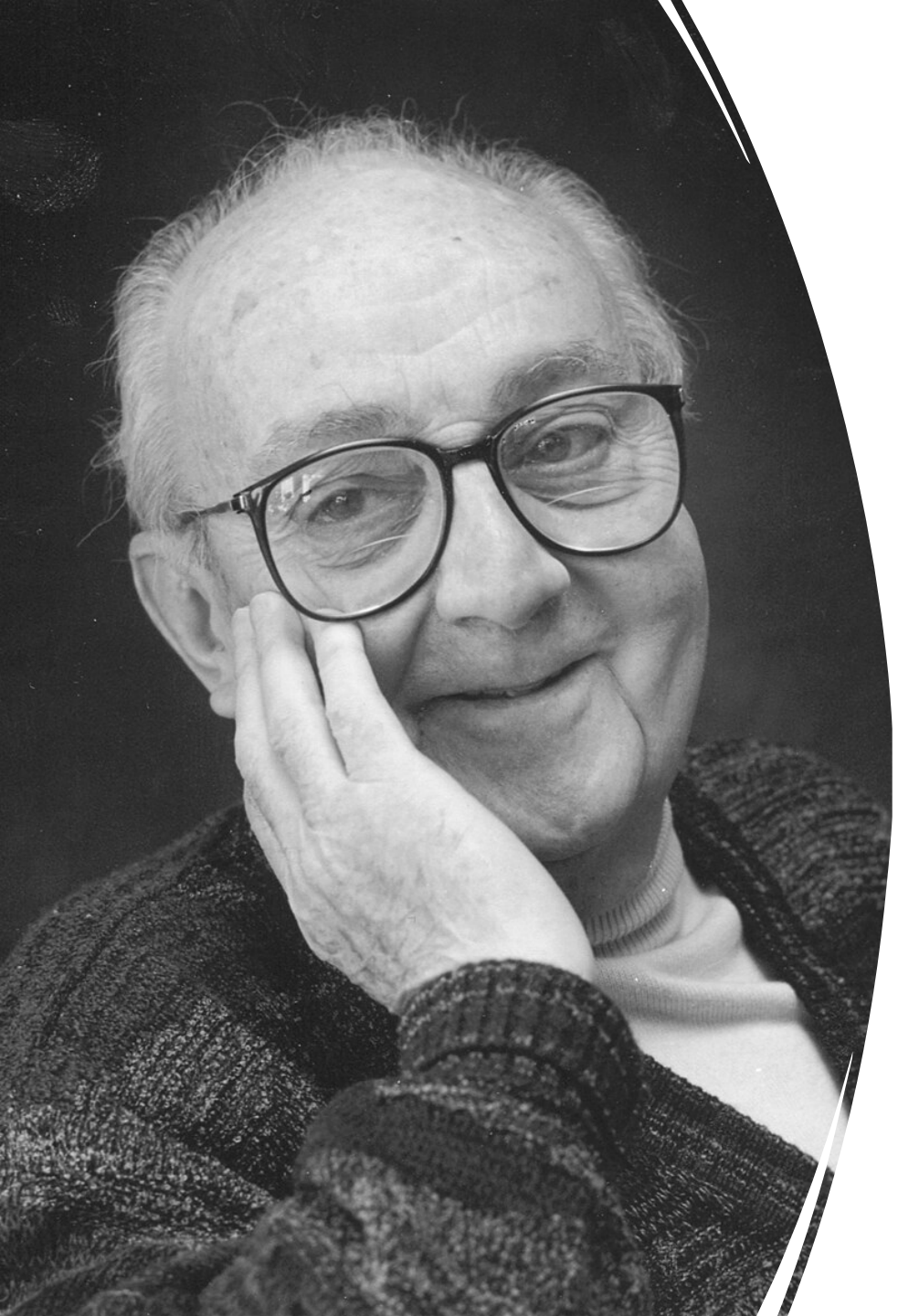


Figure 3. Comparison of SNP heritability estimates from direct genetic effects and population effects. *The x-axis is the SNP heritability estimate from applying LDSC[33] to genome-wide summary statistics on population effects. The y-axis is the SNP heritability estimate from applying LDSC to direct genetic effects (DGEs) (Methods). Vertical and horizontal error bars give the 95% confidence intervals. The diagonal line is the identity. We label the phenotypes with statistically detectable differences (FDR< 0.05, two-sided test): Age at first birth (women); EA, educational attainment; Ever-smoker, whether an individual has ever smoked; Depression; and Height.*

Why do it?



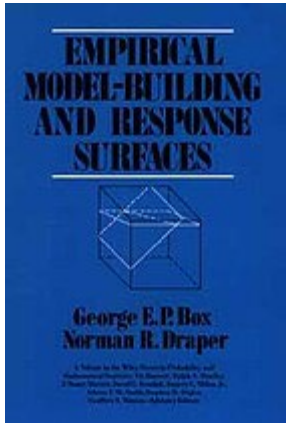
McCarthy et al. *Nat Rev Genet* (2008)



George Box

“Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”

Box, G.E.P. and Draper, N.R. (1987) Empirical model-building and response surfaces. New York: Wiley. p424




Questions?