Genome-Wide Association Studies

Live Session

International Statistical Genetics Workshop – 2026

Abdel Abdellaoui

a.abdellaoui@amsterdamumc.nl

Today’s plan

- 5 minutes to get to know people in your room (with ice breaker question)
- Live lecture (~20 minutes)
- 1 hour practical
- 10 minute break
- 1 hour practical
- ~20 minutes of questions

Ice breaker question of the day:

Tell each other which trait you would like to do a GWAS on and why

(you get bonus points if it’s a trait that hasn’t been subjected to a GWAS yet)

When did GWAS begin?

SCIENCE • VOL. 273 • 13 SEPTEMBER 1996

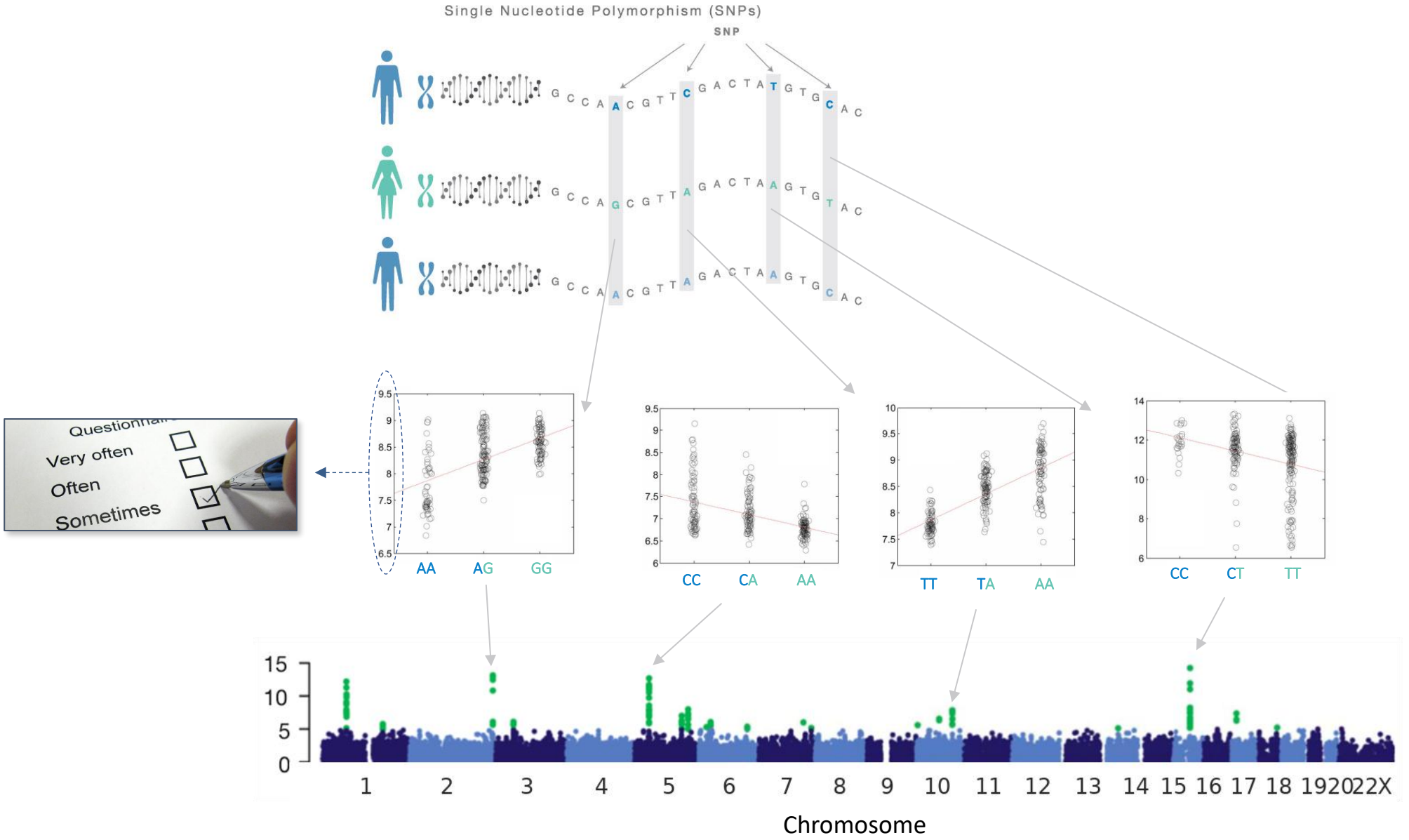
The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

Thus, the primary limitation of genome-wide association tests is not a statistical one but a technological one. A large number of genes (up to 100,000) and polymorphisms (preferentially ones that create alterations in derived proteins or their expression) must first be identified, and an extremely large number of such polymorphisms will need to be tested. Although testing such a large number of polymorphisms on several hundred, or even a thousand families, might currently seem implausible in scope, more efficient methods of screening a large number of polymorphisms (for example, sample pooling) may be possible.

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

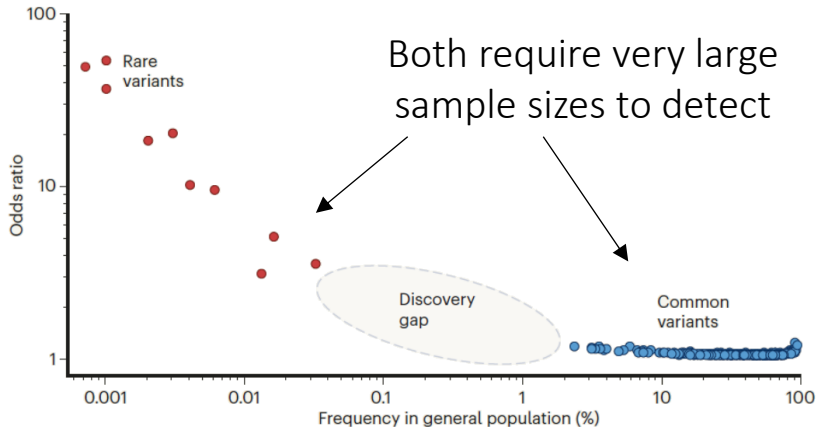
Comparison of linkage and association studies. Number of families needed for identification of a disease gene.


Single Nucleotide Polymorphism (SNPs)
SNP
Very often
Often
Sometimes
AA
AG
GG
CC
CA
AA
TT
TA
AA
CC
CT
TT
Chromosome

Some general lessons learned about complex traits from GWASs

Some general lessons learned about complex traits from GWASs

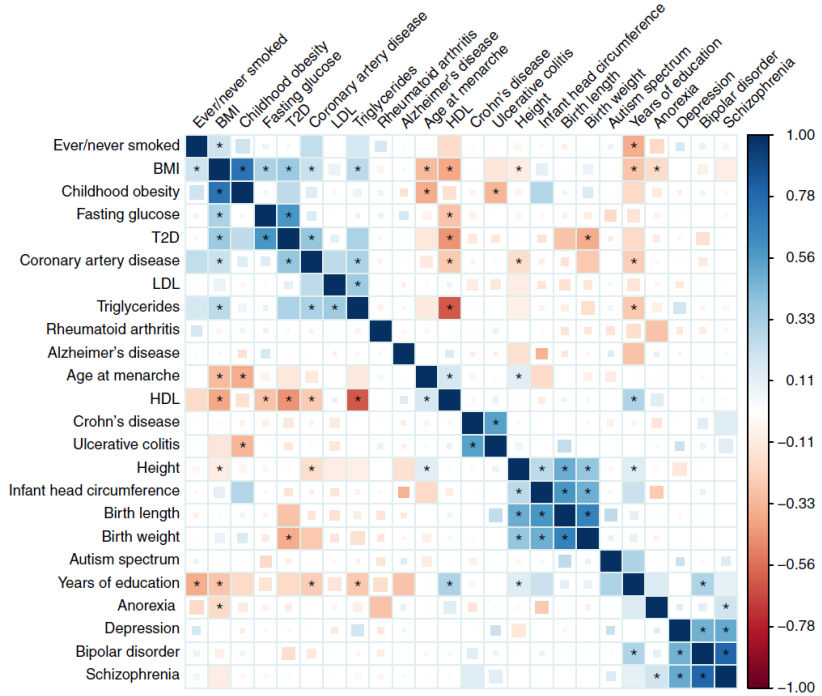
Influenced by many common variants with small effects and rare variants with larger effects



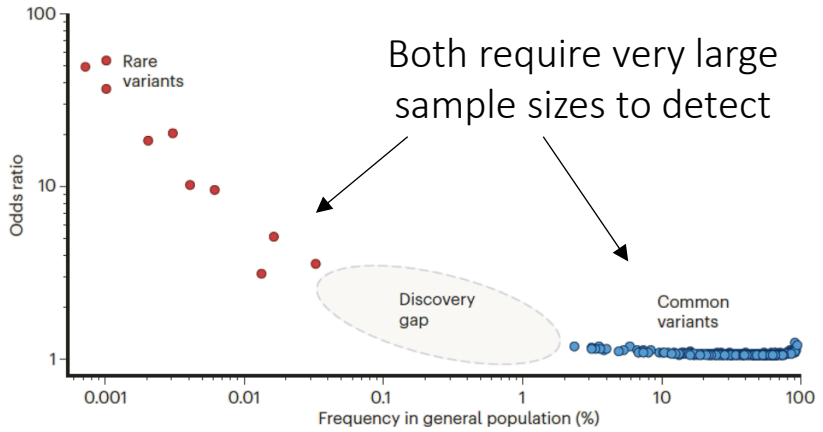
Negative selection keeps large-effect variants rare

Some general lessons learned about complex traits from GWASs

Complex traits share much of their genetic effects with other complex traits



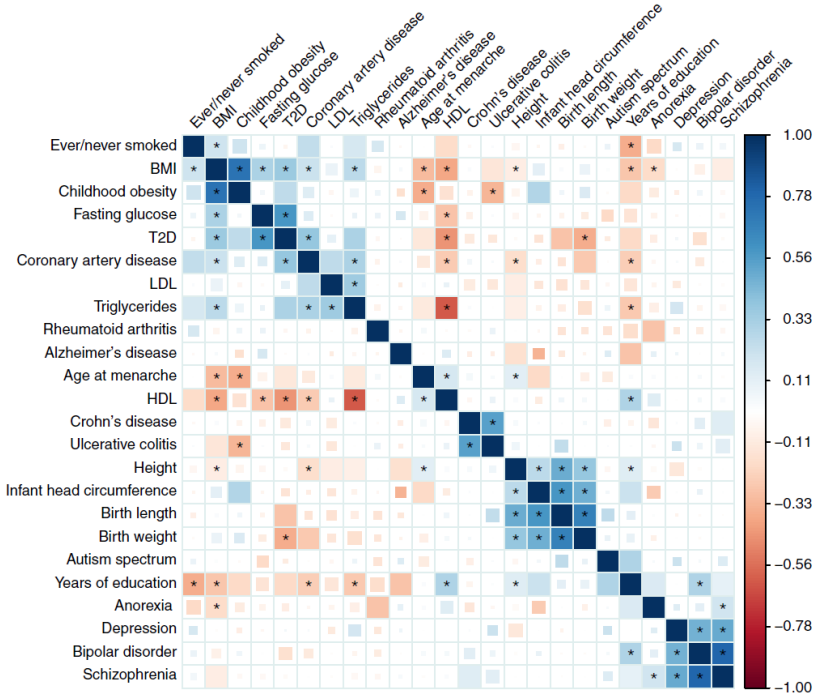
Influenced by many common variants with small effects and rare variants with larger effects



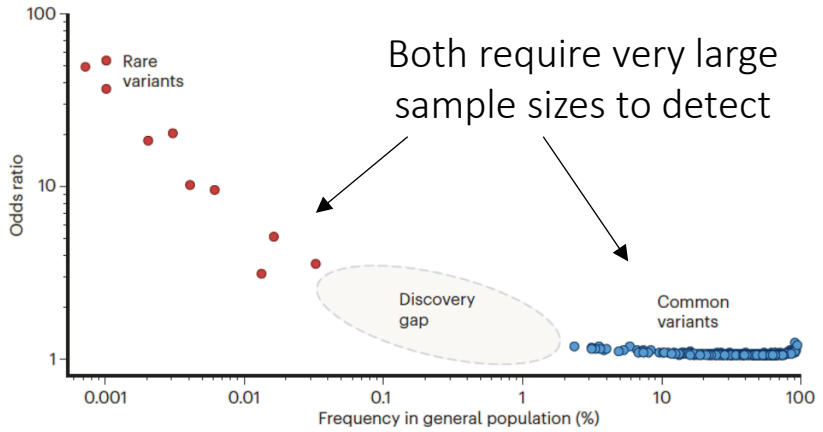
Negative selection keeps large-effect variants rare

Some general lessons learned about complex traits from GWASs

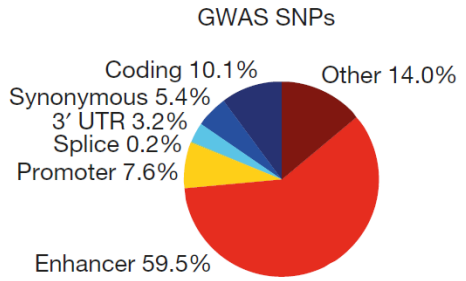
Complex traits share much of their genetic effects with other complex traits



Influenced by many common variants with small effects and rare variants with larger effects



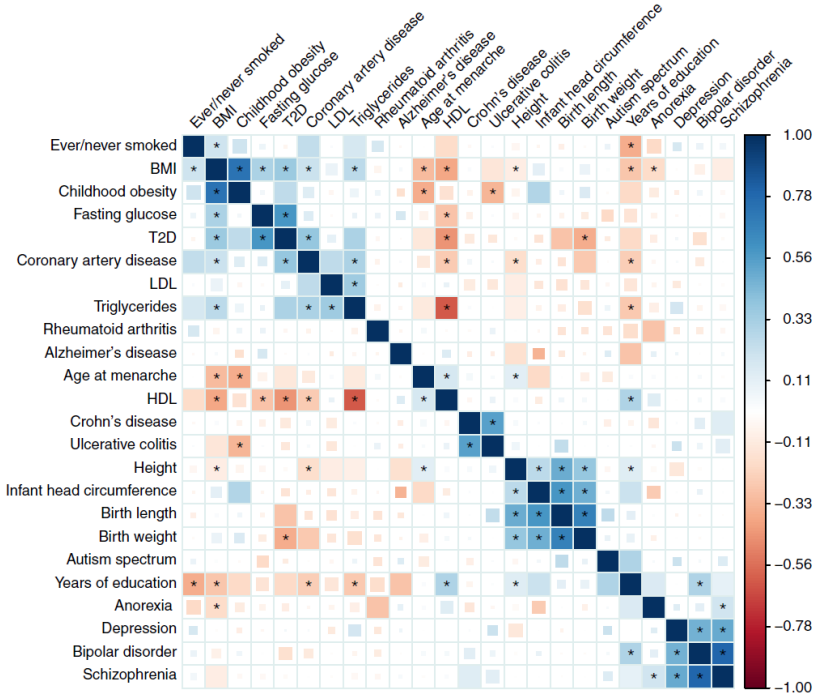
Negative selection keeps large-effect variants rare



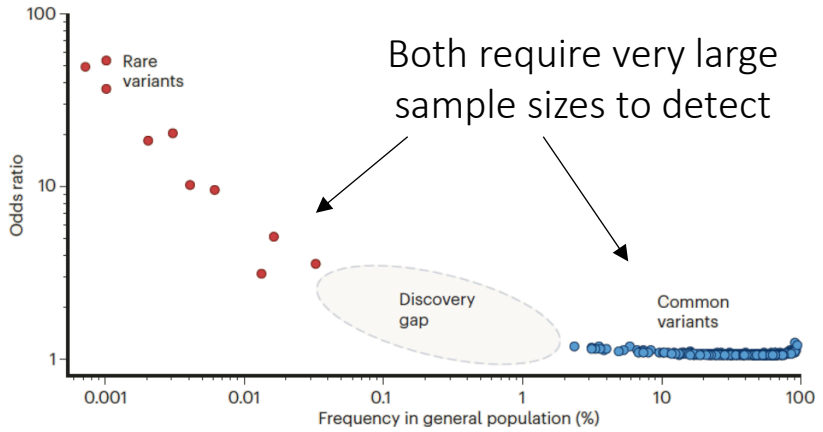
Most GWAS hits fall in regulatory regions, only few in coding regions

Some general lessons learned about complex traits from GWASs

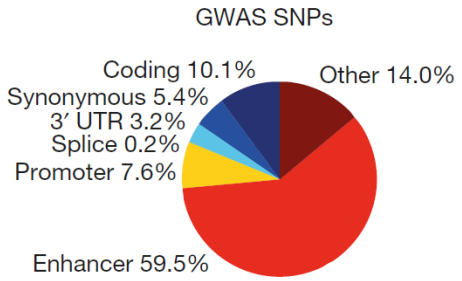
Complex traits share much of their genetic effects with other complex traits



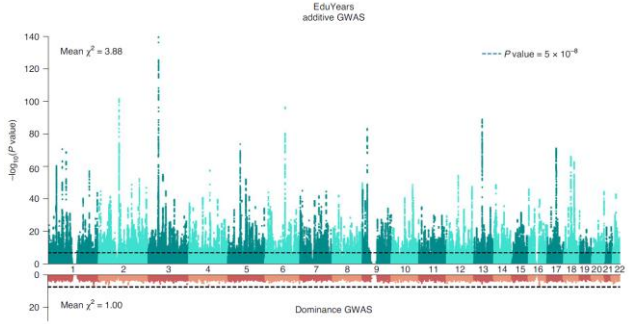
Influenced by many common variants with small effects and rare variants with larger effects



Negative selection keeps large-effect variants rare



Most GWAS hits fall in regulatory regions, only few in coding regions



Genetic effects on complex traits are overwhelmingly additive

How do we separate signal (= causal genetic effects) from noise (e.g., confounding)?

Looks like signal, but isn't:

- **Multiple testing**: millions of tests, many $p < 0.05$ by chance
- **Technical artefacts**: genotyping errors, batch effects
- **Population stratification**: Genetic differences between ancestries look like trait associations
- **Cryptic relatedness**: Hidden family structure inflates test statistics

Discussed in Lecture 2

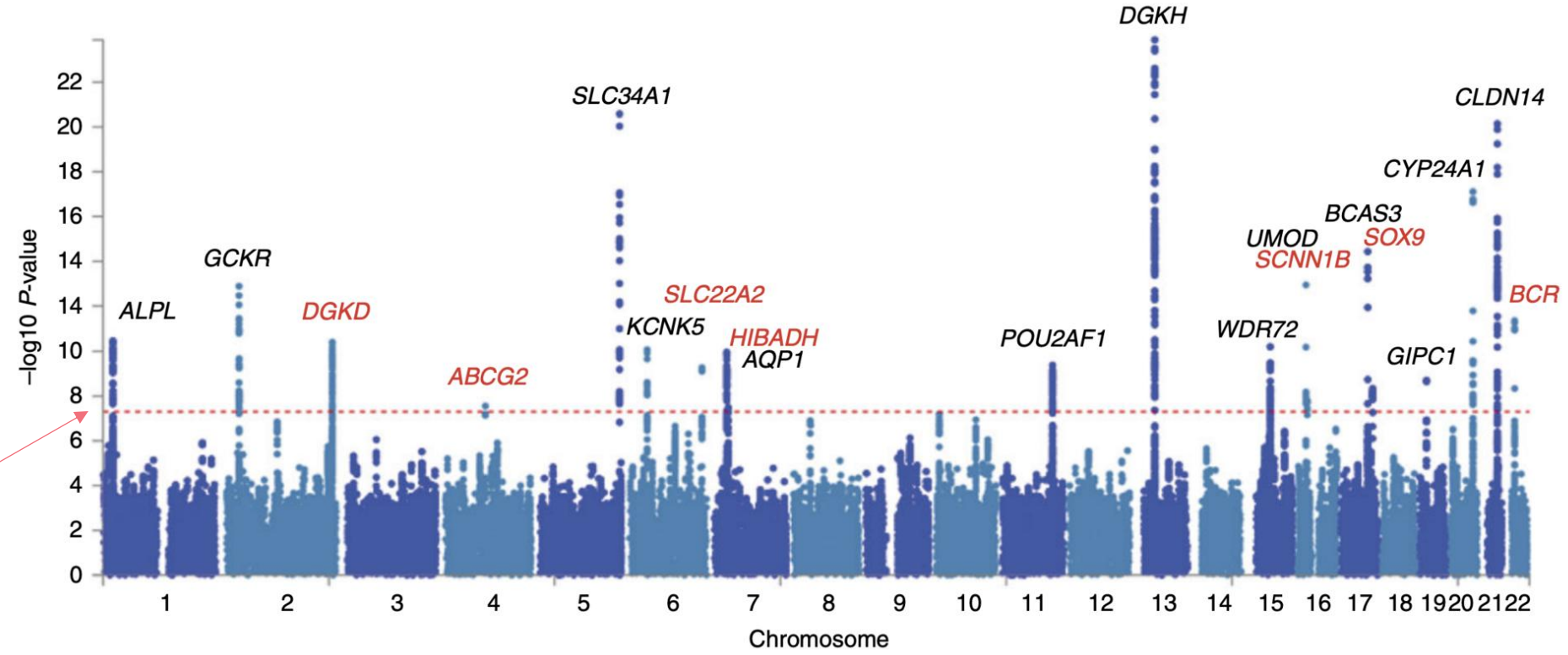
Real signal, but inflated or misunderstood:

- **Gene-environment correlation**: Real genetic effects amplified by correlated environment
- **Assortative mating**: Creates genome-wide correlations that inflate polygenic signals

Discussed in Lecture 4

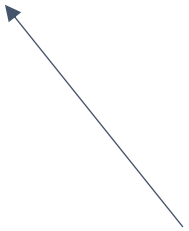
Multiple testing correction

Bonferroni corrected significance threshold



How to deal with technical artefacts?

- Exclude SNPs with a low Minor Allele Frequency (MAF) (< .01 or < .05)
- Exclude SNPs & individuals with excessive missingness (> 1% or 5% respectively)
- Exclude SNPs that deviate grossly from HWE ($p < 10^{-6}$)
- Exclude individuals with extreme values for F (too homozygous or too heterozygous)
- Exclude SNPs based on quality score or clustering metrics (availability depends on genotype platform used)
- Exclude SNPs with Mendelian errors (if child-parent data are available)



Good study design and good QC

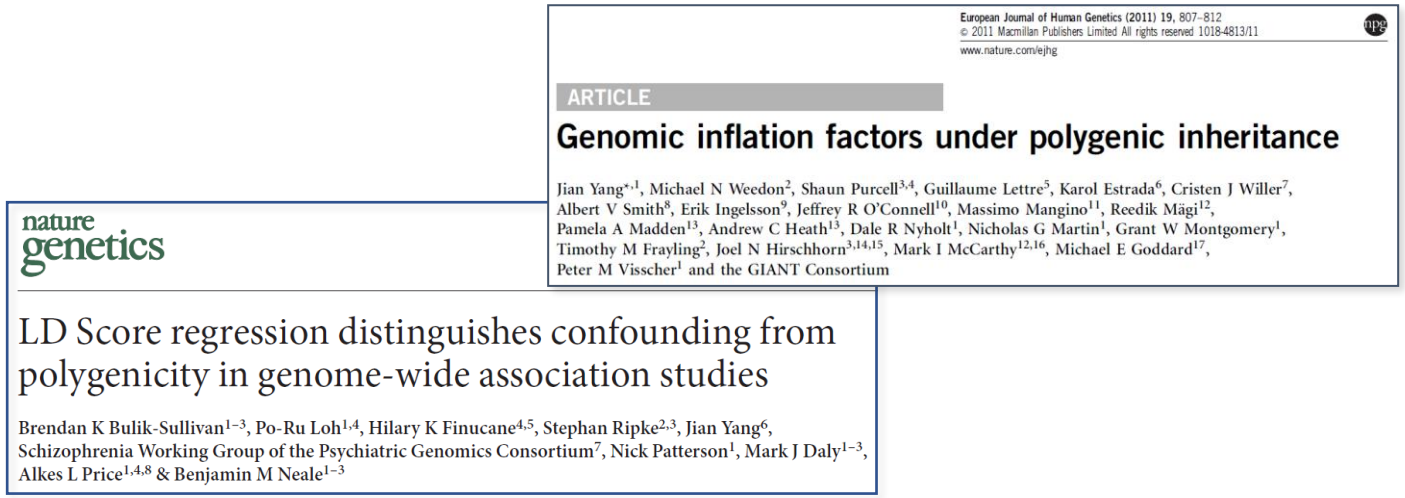
Dealing with population stratification & cryptic relatedness

- Genomic Control (GC)
- Principal Component Analysis (does not deal with cryptic relatedness)
- Within Family Association
- Linear Mixed Modeling

Dealing with population stratification & cryptic relatedness

A bit outdated (unless inflation factor from LDSC regression is used)

- Genomic Control (GC)
- Principal Component Analysis (does not deal with cryptic relatedness)
- Within Family Association
- Linear Mixed Modeling



European Journal of Human Genetics (2011) 19, 807–812

www.nature.com/ejhg

ARTICLE

Genomic inflation factors under polygenic inheritance

Jian Yang*[,1], Michael N Weedon[2], Shaun Purcell[3,4], Guillaume Lettre[5], Karol Estrada[6], Cristen J Willer[7], Albert V Smith[8], Erik Ingelsson[9], Jeffrey R O'Connell[10], Massimo Mangino[11], Reedik Mägi[12], Pamela A Madden[13], Andrew C Heath[13], Dale R Nyholt[1], Nicholas G Martin[1], Grant W Montgomery[1], Timothy M Frayling[2], Joel N Hirschhorn[3,14,15], Mark I McCarthy[12,16], Michael E Goddard[17], Peter M Visscher[1] and the GIANT Consortium

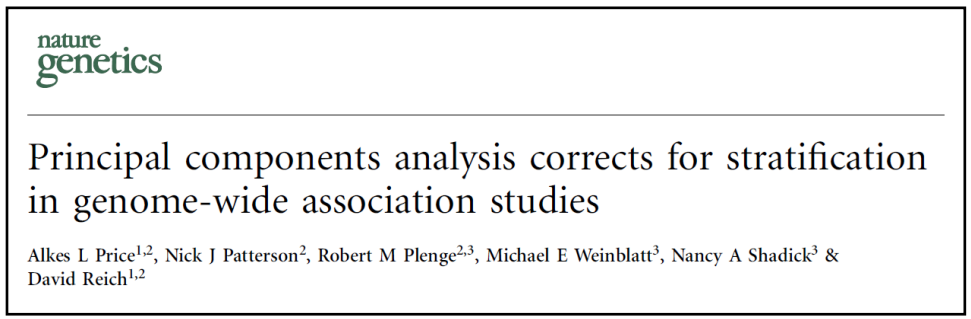
nature genetics

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan[1–3], Po-Ru Loh[1,4], Hilary K Finucane[4,5], Stephan Ripke[2,3], Jian Yang[6], Schizophrenia Working Group of the Psychiatric Genomics Consortium[7], Nick Patterson[1], Mark J Daly[1–3], Alkes L Price[1,4,8] & Benjamin M Neale[1–3]

Dealing with population stratification & cryptic relatedness

- Genomic Control (GC)
- Principal Component Analysis (does not deal with cryptic relatedness)
- Within Family Association
- Linear Mixed Modeling

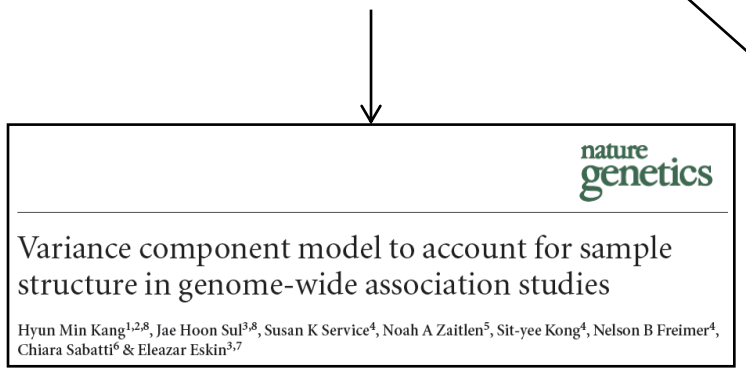


Dealing with population stratification & cryptic relatedness

- Genomic Control (GC)
- Principal Component Analysis (does not deal with cryptic relatedness)
- Within Family Association
- Linear Mixed Modeling

Dealing with population stratification & cryptic relatedness

- Genomic Control (GC)
- Principal Component Analysis (does not deal with cryptic relatedness)
- Within Family Association
- Linear Mixed Modeling → **Controlling for Genetic Relatedness Matrix (GRM)**


nature genetics

Variance component model to account for sample structure in genome-wide association studies

Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[4], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]


nature genetics

Advantages and pitfalls in the application of mixed-model association methods

Jian Yang[1,2,8], Noah A Zaitlen[3,8], Michael E Goddard[4,9], Peter M Visscher[1,2,9] & Alkes L Price[5–7,9]

The linear mixed model (LMM)

$$y = X\beta + g + \varepsilon$$

phenotype vector

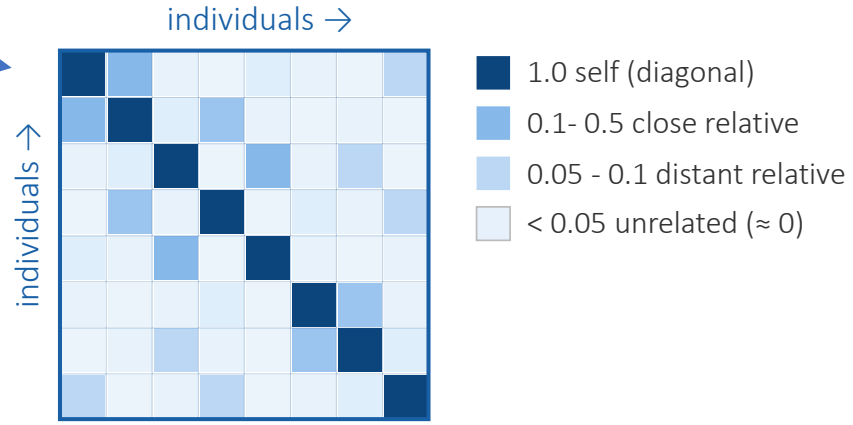
fixed effects
(SNP + covariates)
β_{SNP} = what we are interested in

~ N(0, σ^2 GRM)

genetic random effect, captures relatedness between all individuals

~ N(0, σ^2 I)

residual error
(independent)

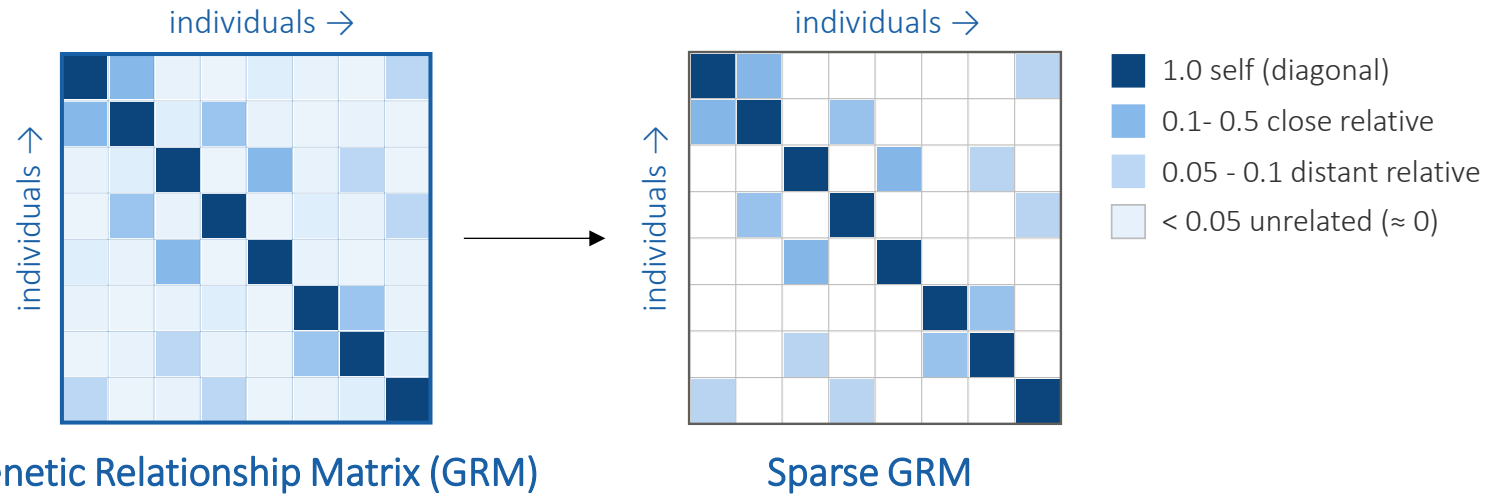


Genetic Relationship Matrix (GRM)

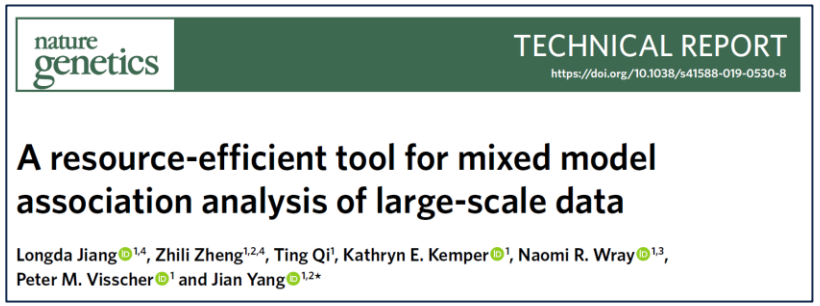
Each cell = relatedness between two individuals estimated from genome-wide SNP data

fastGWA

- Uses a sparse GRM, which only retains relatedness estimates above a threshold (~0.05)
- Part of GCTA



2019



nature genetics ARTICLES
https://doi.org/10.1038/s41588-022-01062-7
Check for updates
OPEN

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

Standard GWAS

requires: unrelated individuals

$y_i = \beta_l g_{il} + x_i'\alpha + \varepsilon_i$

β_l population effect

g_{il} focal individual's genotype

$x_i'\alpha$ covariates (age, sex, PCs)

β conflates direct effects with gene-environment correlation (rGE), assortative mating (AM), and population stratification

Sib-GWAS

requires: sibling pairs

$y_i = \beta_l G^C_{il} + G^F_{f(i)l} + x_i'\alpha + \varepsilon_i$

β_l direct genetic effect

$G^C_{il} = G_{il} - G^F_{f(i)l}$ (within-family deviation)

$G^F_{f(i)l}$ = mean sibling genotype

Including the family genetic background as a covariate isolates the random within-family variation due to Mendelian segregation, which captures more of the direct effect

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

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

Tammy Tan[1], Hariharan Jayashankar[2], Junming Guan[2], Seyed Moeen Nehzati[2], Mahdi Mir[2], Michael Bennett[1], Esben Agerbo[3], Rafael Ahlskog[4], Ville Pinto de Andrade Anapaz[5], Stefania Benonisdottir[6], Dorret I. Boomsma[7], Archie Campbell[8,9], Christopher F. Chabris[10], Zhengming Chen[11], China Kadoorie Biobank Collaborative Group, Eco de Geus[12,13], Erik A. Ehli[14], Abdelrahman G. Elnahas[15], Estonian Biobank Research Team[16], Finngen[17], Andrea Ganna[5,18], Alexandros Giannelis[19], Liisa Hakaste[5,20], Caroline Hayward[21], Jouke-Jan Hottenga[14,22], Mikkel Aagaard Houmark[23], Jaakko Kaprio[5], Antti Latvala[24], James J. Lee[19], Mikko Lehtovirta[5], Liming Li[25,26,27], LifeLines Cohort Study[28], Kuang Lin[11], Richard Karlsson Linnér[29,30,31], Stefano Lombardi[5,32,33], Nicholas G. Martin[34], Matt McGue[19], Sarah E. Medland[34], Andres Metspalu[15,16], Brittany L. Mitchell[34], Guiyan Ni[35], Ilja M. Nolte[36], Matthew T. Oetjens[30], Sven Oskarsson[4], Teemu Palviainen[5], Rashmi B. Prasad[5,37], Anu Reigo[15], Kadri Reis[15], Julia Sidorenko[35], Karri Silventoinen[38], Harold Snieder[36], Tinamaija Tuomi[5,20,37,39], Bjarni J. Vilhjálmsson[40,41,42], Robin G. Walters[21], Emily A. Willoughby[19], Jonathan Flint[43], Loic Yengo[35], Peter M. Visscher[21,35], Augustine Kong[44], Elliot M. Tucker-Drob[45], Richard Border[46,47], David Cesarini[1,48,49], Patrick Turley[50,51], Aysu Okbay[52], Daniel J. Benjamin[1,2,53], Alexander Strudwick Young[1,2,53]

Family-based GWAS (FGWAS)

requires: trios/sib pairs
(parental G can be imputed)

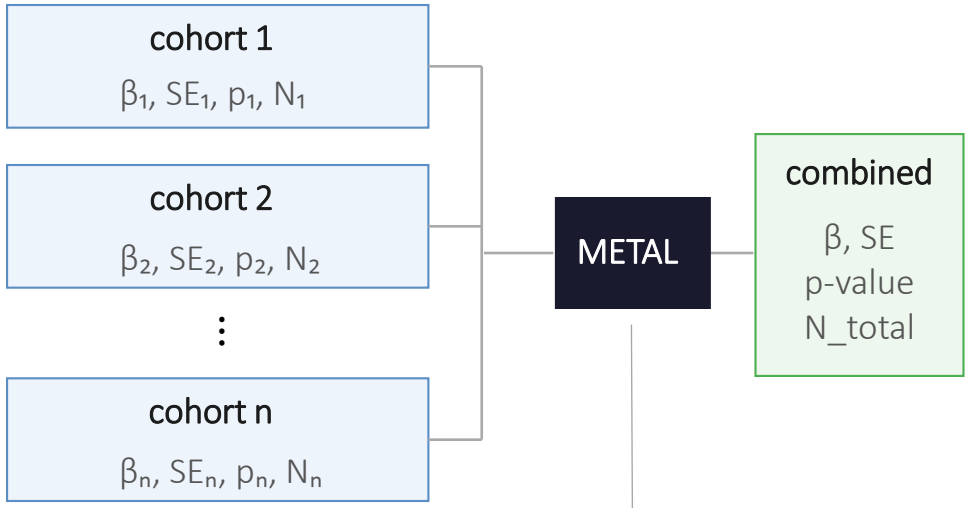
$y_i = \beta_l g_{il} + g_{p(i)l} + g_{m(i)l} + x_i'\alpha + \varepsilon_i$

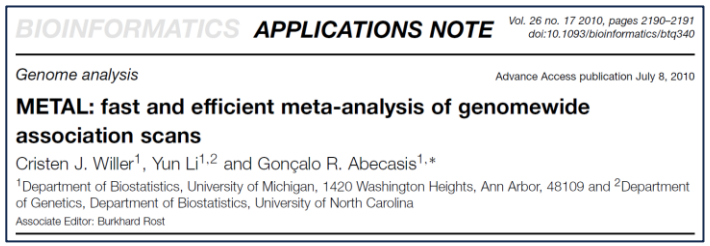
β_l direct genetic effect

$g_{p(i)l}$ paternal genotype (non-transmitted)

$g_{m(i)l}$ maternal genotype (non-transmitted)

GWAS meta-analysis





BIOINFORMATICS **APPLICATIONS NOTE** Vol. 26 no. 17 2010, pages 2190–2191
doi:10.1093/bioinformatics/btq340

Genome analysis Advance Access publication July 8, 2010

METAL: fast and efficient meta-analysis of genomewide association scans

Cristen J. Willer[1], Yun Li[1,2] and Gonçalo R. Abecasis[1,*]

[1]Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, 48109 and [2]Department of Genetics, Department of Biostatistics, University of North Carolina

Associate Editor: Burkhard Rost

In the **practical**, you will carry out a GWAS on two simulated datasets:

1. Perform basic **quality control (QC)** on genotype data using **PLINK**
2. Compute and plot **principal components (PCs)** to visualize population structure
3. Build a **genetic relationship matrix (GRM)** and run a **GWAS** using **GCTA**
4. Visualize your results with **Manhattan plots** and **QQ-plots** in **R**
5. Conduct a **meta-analysis** of two GWAS results using **METAL**

Practical

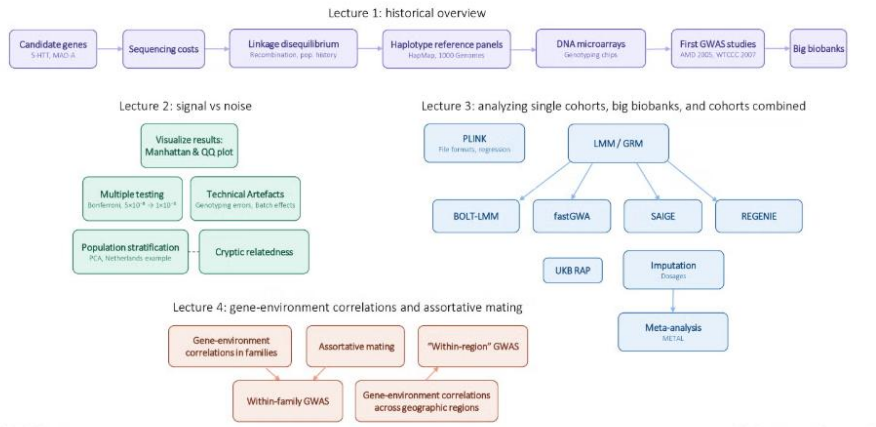
https://www.colorado.edu/ibg/workshop-2026/syllabus/genome-wide-association-studies#accordion-e8b5645a54520682bd7ba7abad61647bd-1

University of Colorado **Boulder**

Institute for Behavioral Genetics

About IBG | People | Community | Join IBG | Resources | Research | ISG Workshop

ISG 2026 — Day 4: Genome-Wide Association Studies (GWAS)



Lectures

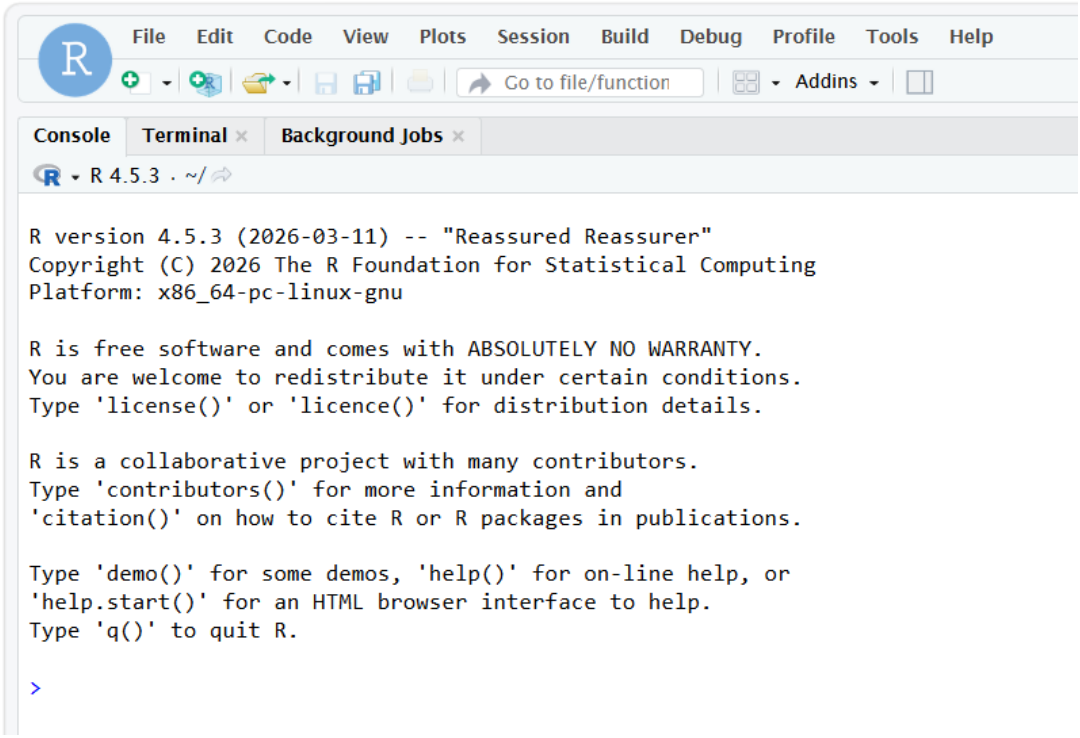
GWAS lecturs (86 minutes)

Practicals

This practical is guided through Qualtrics.

Click here

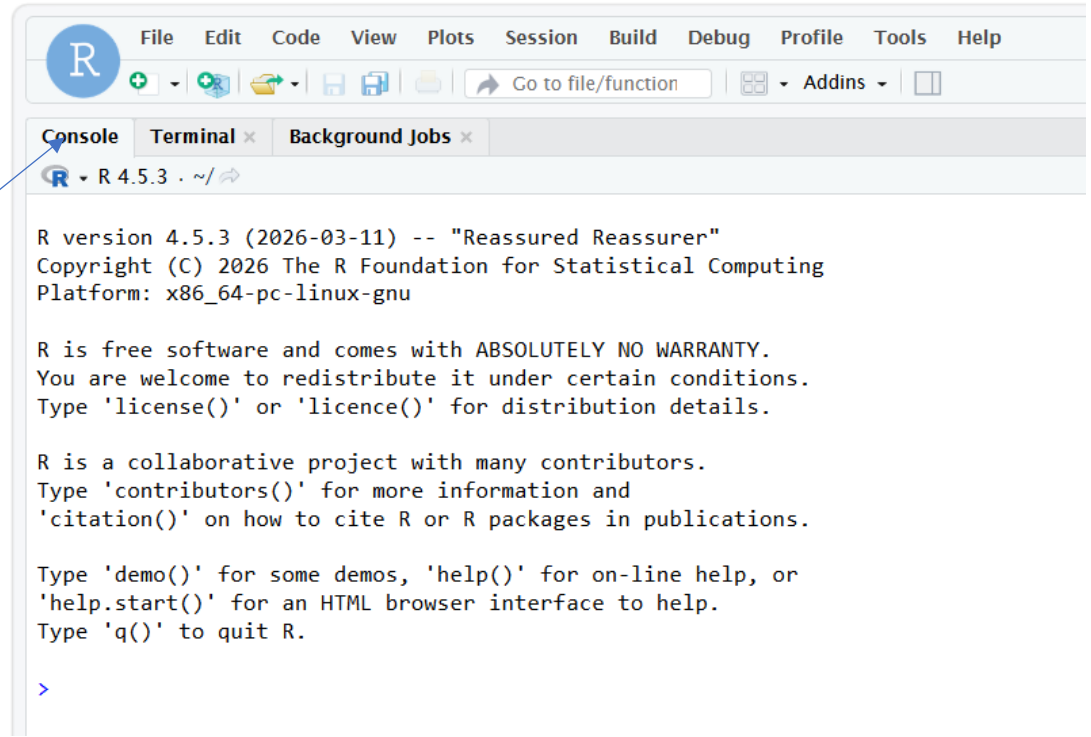
Practical


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Console
Terminal
Background Jobs
R 4.5.3 · ~/
R version 4.5.3 (2026-03-11) -- "Reassured Reassurer"
Copyright (C) 2026 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>

Practical

R

```
setwd("YOUR_WORKING_DIRECTORY")
```



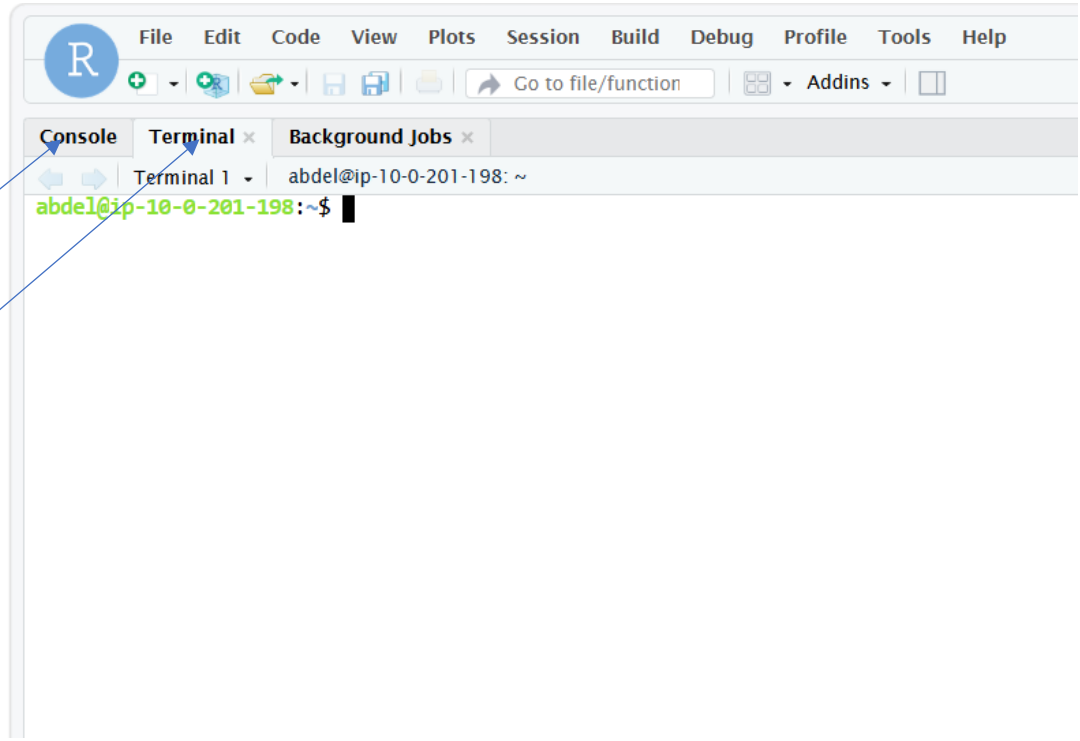
Practical

R

```
setwd("YOUR_WORKING_DIRECTORY")
```

TERMINAL

```
mkdir GWAS
cd GWAS
cp -r /home/abdel/2026/gwas_practical/* .
```



If you get stuck, you can find outputfiles as well as answers to the practical here:

`/home/abdel/2026/gwas_practical/answers/`

(i.e., the folder “answers” in your own working folder after you copied the files)

It contains all output files and these two files:

- **`ISG_2026_GWAS_Practical_Answers.txt`**: contains the answers to the questions in Qualtrics
- **`ISG_2026_GWAS_Practical_Commands.txt`**: contains all the commands (which are also in Qualtrics)