camh
Centre for Addiction and Mental Health

Krembil Centre for Neuroinformatics

Using big data, artificial intelligence and brain modelling to fundamentally change our understanding of mental illness.

SUMMER SCHOOL 2020

Day 1

Reproducible Neuroinformatics and Psychiatric Genetics

camh
Krembil Centre for Neuroinformatics

A quick how-to guide on Crowdcast

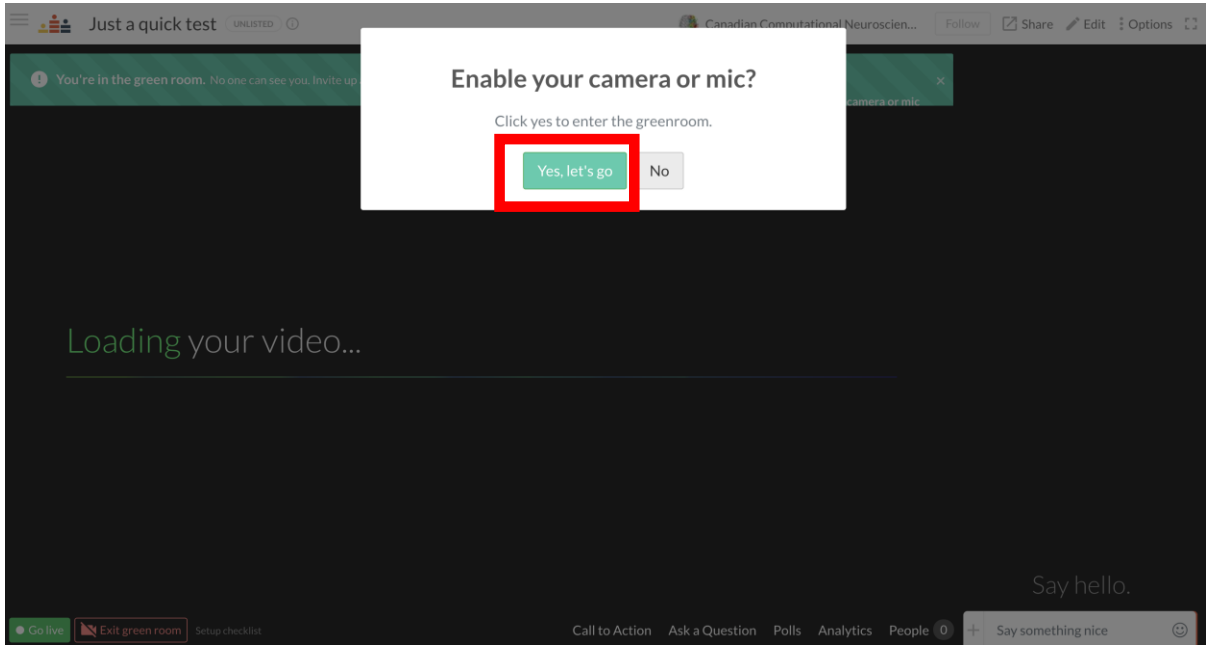
Ask a question, participate in the ongoing chat, come “on stage”, etc.

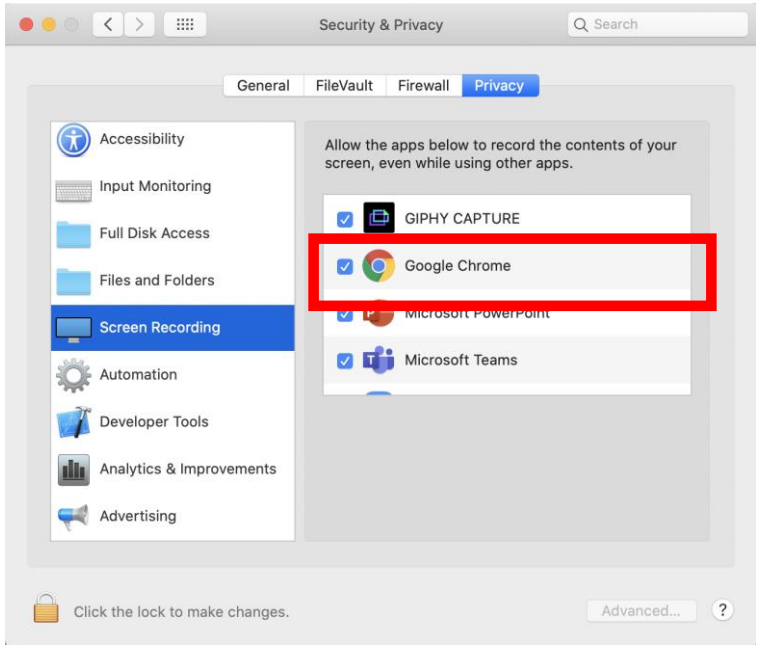
Thanks to Dr. Scott Rich

Crowdcast

camh
Krembil Centre for Neuroinformatics

If you want to “come on stage” with video/audio



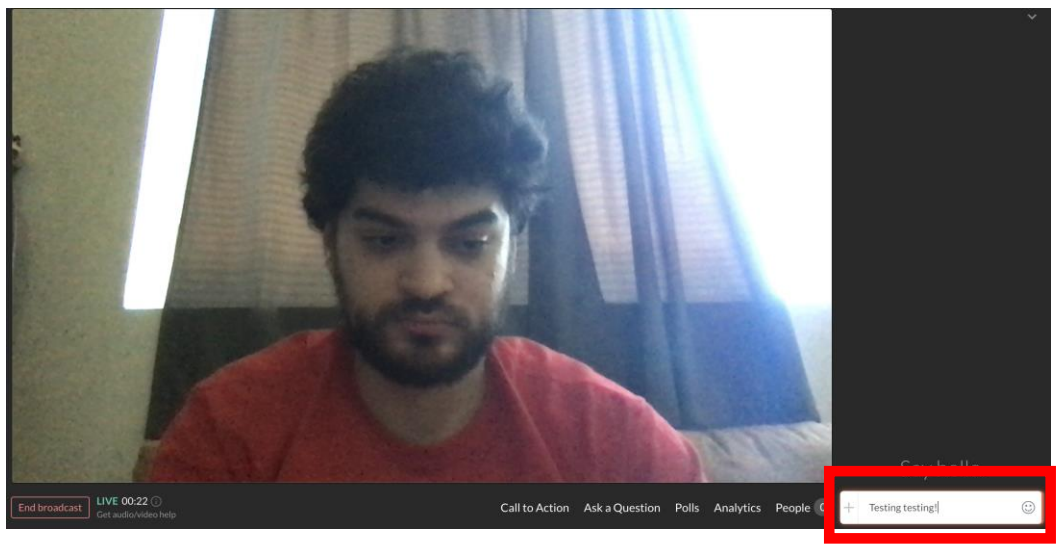


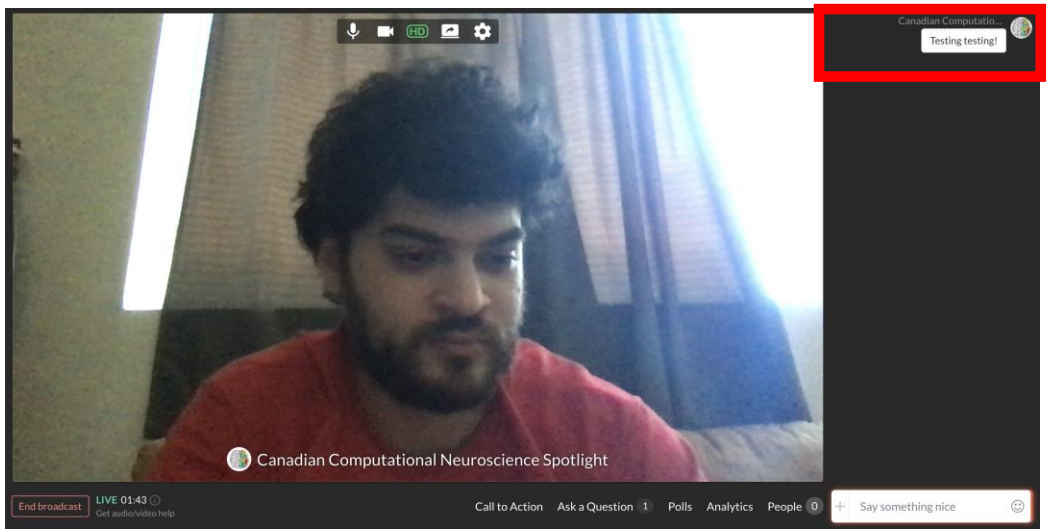
https://www.crowdcast.io/setup

Crowdcast

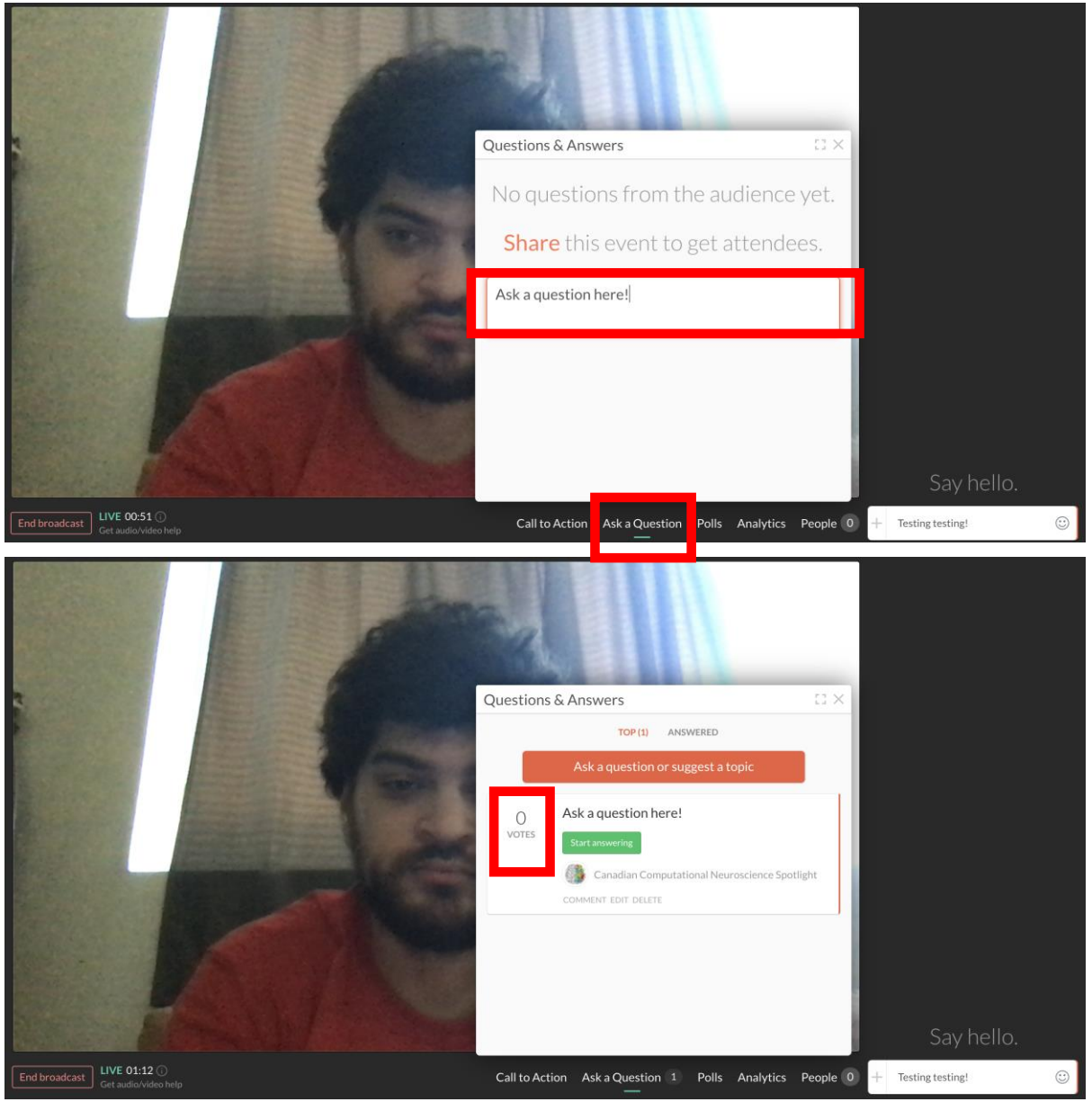
camh
Krembil Centre for Neuroinformatics

Use the “chat” function on the right-hand side of the screen both before and during presentations





Use the “Ask a Question” feature, and up-vote other questions you’d like answered!



Course Overview

Day 1: Reproducible Science and Psychiatric Genetics

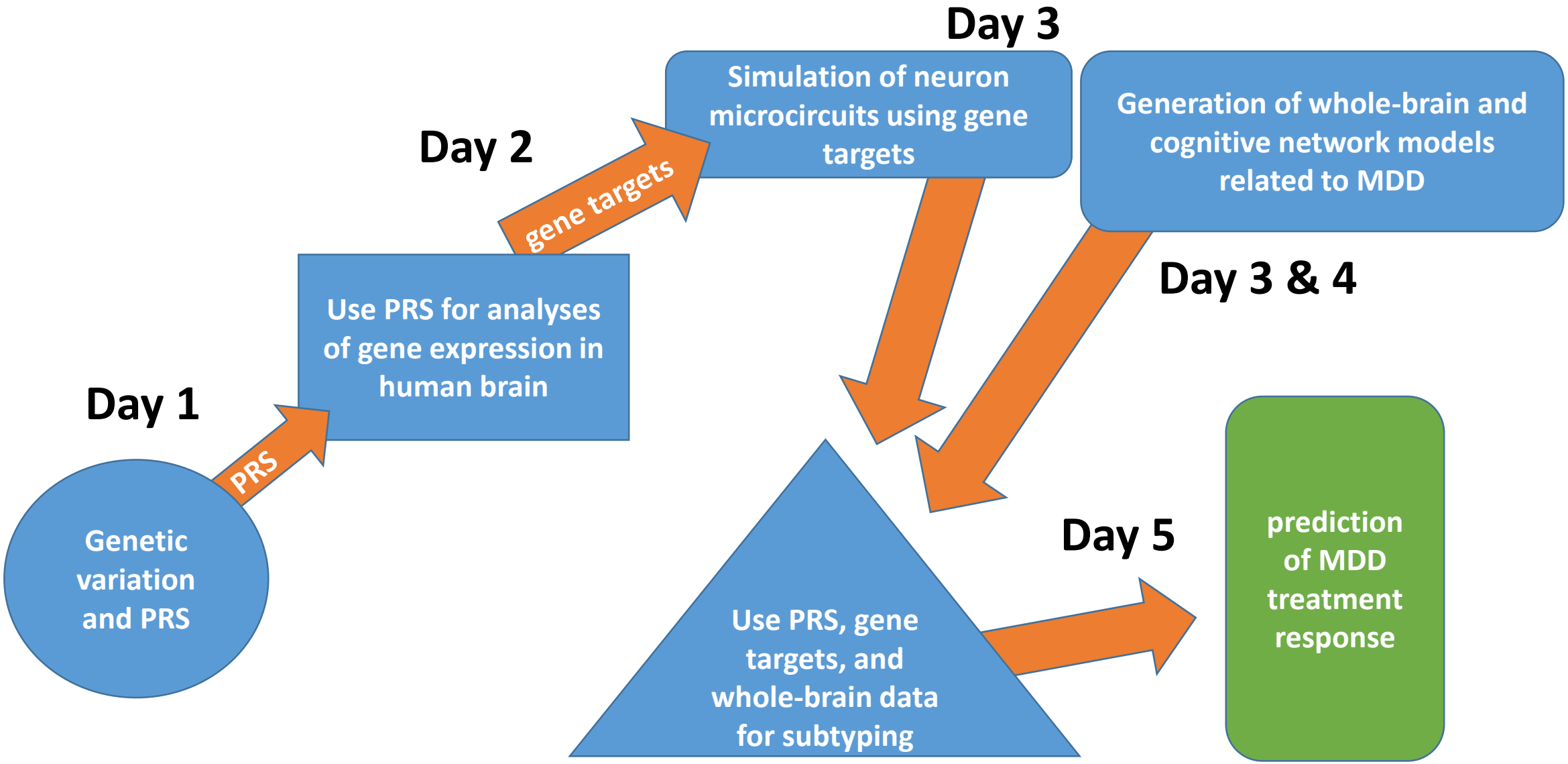
Day 2: Functional Genomics

Day 3: Cortical microcircuit and whole-brain Modelling

Day 4: Bayesian models of perception and learning

Day 5: Psychiatric epidemiology and Population-based subtyping

Graphical Overview



Ultimate Goal

Aim: To use multiple data types, including sociodemographics, genomics, neuroimaging, and multi-scale electrophysiology to predict whether a patient diagnosed with major depressive disorder (MDD) will respond to a particular intervention.

Hypothesis: biologically-informed subgroups will provide improved prognosis over existing known clinical MDD subtypes.

Day 1: Lead Instructors

camh
Krembil Centre for Neuroinformatics

Dr. Daniel Felsky

Independent Scientist and Head of Whole Person Modelling at KCNI
Assistant Professor in the Department of Psychiatry at the University of Toronto.
PhD in neuroimaging and genetics of Alzheimer's disease.
Postdocs in computational neuroimunology and bioinformatics at Broad Institute, Harvard Medical School & Columbia University.
Teaching Fellow for the Global Initiative for Neuropsychiatric Genetics Education in Research (USA, South Africa, Uganda, UK).

Dr. Erin Dickie

Project Scientist in KCNI and the Kimel Family Translational Imaging-Genetics Research Lab
Assistant Professor in the Department of Psychiatry at the University of Toronto.
PhD in Neuroscience, from McGill.
Postdocs in population neuroscience (Rotman Research Inst., Baycrest Toronto) and developmental neuroscience at SickKids in Toronto
Research focus in neuroinformatics for precision brain mapping.

Day 1: Teaching Assistants

camh
Krembil Centre for Neuroinformatics

Dr. Sajel Patel
Postdoctoral Fellow

Dr. Michael Wainberg
Postdoctoral Fellow

Day 1: Topics and Schedule

camh
Krembil Centre for Neuroinformatics

1. Introduction to the KCNI

2. Reproducible Neuroinformatics

3. Major Depressive Disorder **(Break 10:30-10:45am)**

4. Genetic Variation and Genotyping

5. Genome-wide Association Study (GWAS) and Linkage Disequilibrium (LD)

6. Polygenic Risk **(Lunch 12:15-1:00pm)**

Day 1: Learning Objectives

camh
Krembil Centre for Neuroinformatics

1. Understand the concepts and implementation of reproducible neuroinformatics

2. Develop insight into the clinical presentation and genetics of major depression

3. Understand basic concepts of single nucleotide variation in human genetics

4. Understand the steps involved in GWAS and the principle of LD

5. Understand the concept of polygenic risk and how scores are calculated

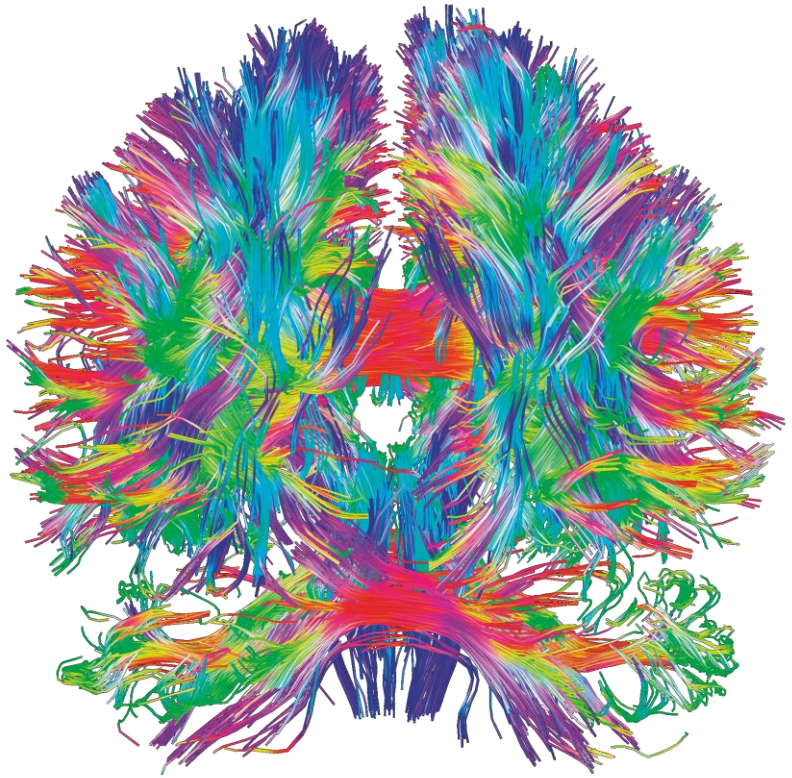
The Krembil Centre for Neuroinformatics

camh | Krembil Centre for Neuroinformatics

Multiscale Data
Integrating high quality clinical and behavioural data including genetics, imaging, and mobile

Artificial Intelligence
Defining, diagnosing and predicting mental illness and brain disorders

Multiscale Brain Modeling
Identifying cellular and synaptic mechanisms of brain function and dysfunction



Open Science
Open, reproducible, team science, FAIR data principles,
Global collaboration

Education
Disseminating knowledge,
Training the next generation of scientists

Innovation
Ethical, responsible incubator for mental health technologies

Principle Investigators

Computational Genomics
Shreejoy Tripathy

Brain Circuit Modeling
Etay Hay

Whole Brain Modeling
John Griffiths

Cognitive Network Modeling
Andreea Diaconescu

Integrative Neuroanatomy
Leon French

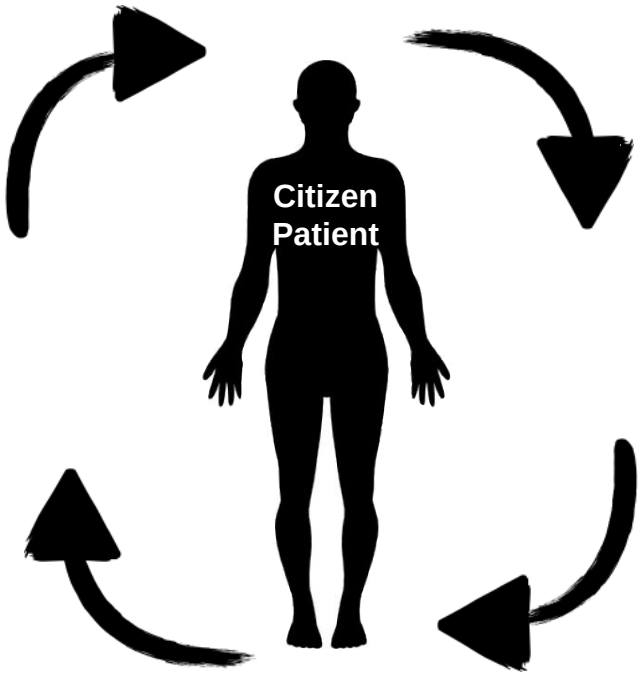
Whole Person Modeling
Daniel Felskey

camh | Krembil Centre for Neuroinformatics

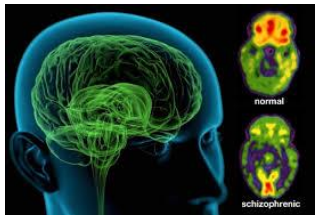
Krembil Centre for Neuroinformatics

Open Science, Team Science, Data-driven Mental Health

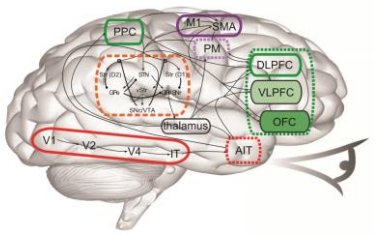
**Clinicians,
Health Professionals**



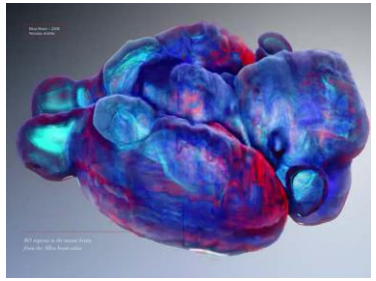
**Researchers
& Clinical
Scientists**

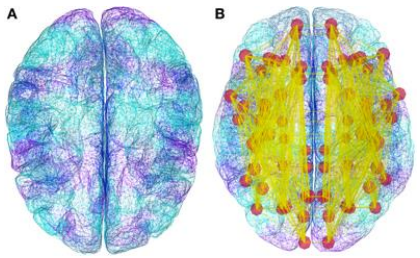


**Medical,
Information
Technology,
Wearables**

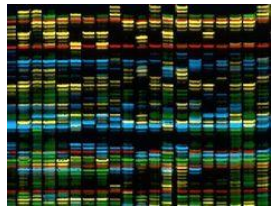


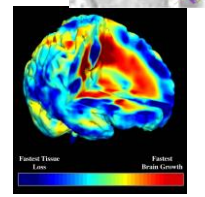
**Analysis, Brain Modeling,
Simulation, Classification,
Prediction, Machine Learning, AI**

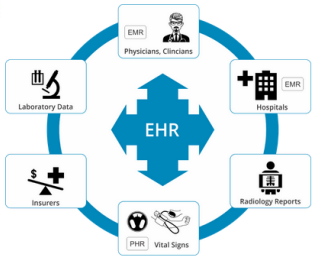




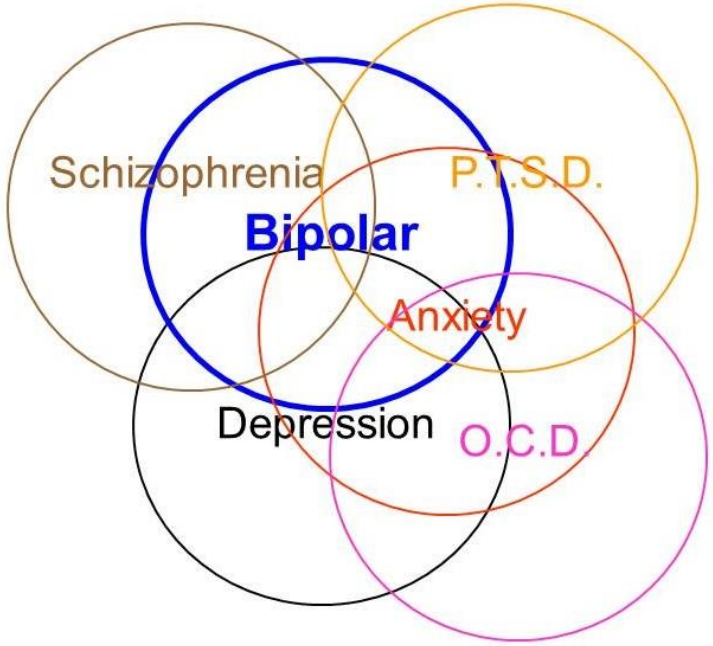
**Multiscale Data
Genomics, Brain Imaging,
Eectronic Health Records**

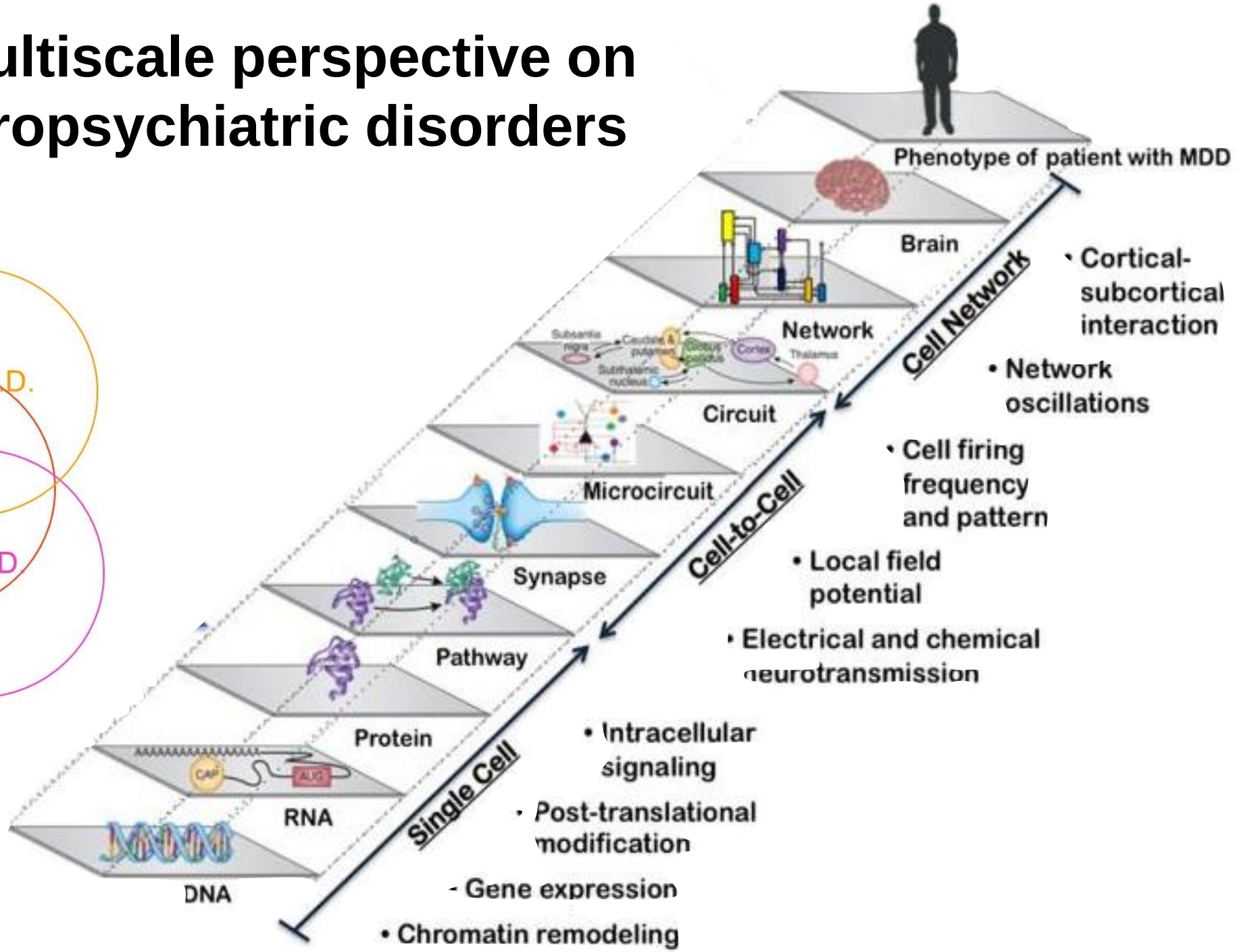




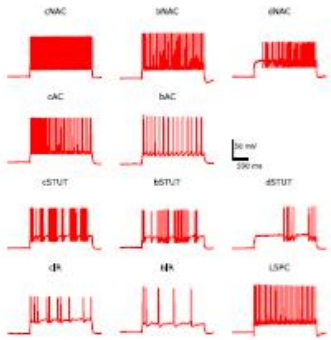


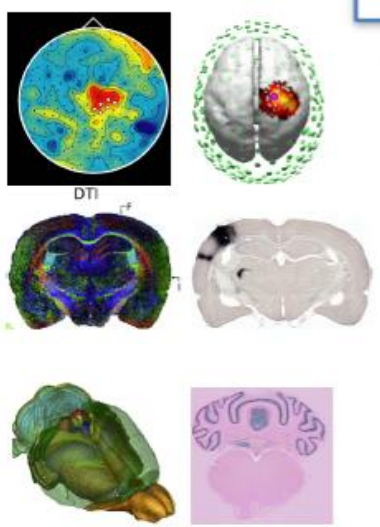
A multiscale perspective on neuropsychiatric disorders



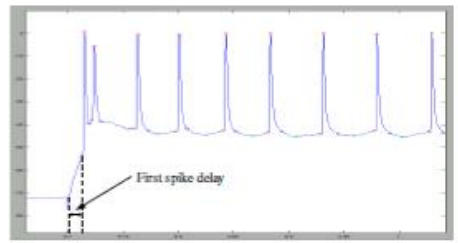


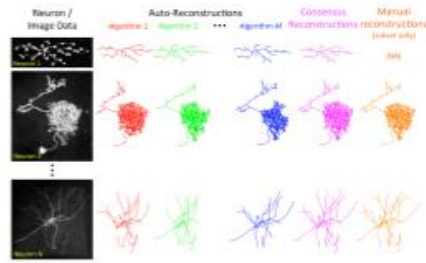
Data Repositories

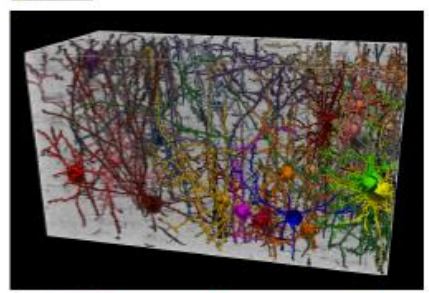


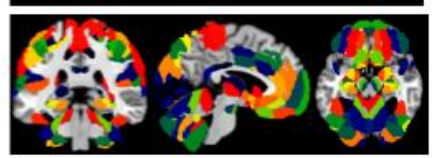


Feature Extraction

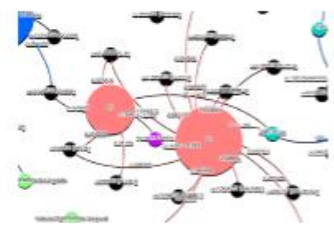




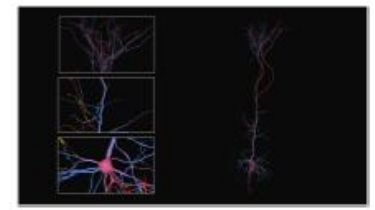


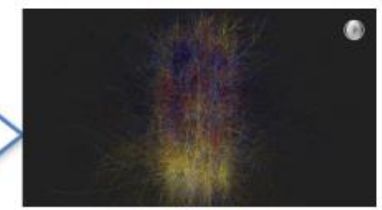


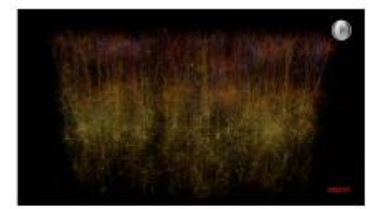
Knowledge Management and Search

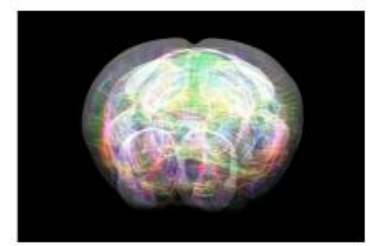


Data-driven Modelling and Simulation









BrainHealth Databank

Dr. Joanna Yu, Senior Project Manager

High Quality Longitudinal Data

Accumulation of high quality longitudinal data

Supporting Delivery of Evidence Based Care

Visualize patient trajectories
machine learning decision support

Integrating Research Measures with Care

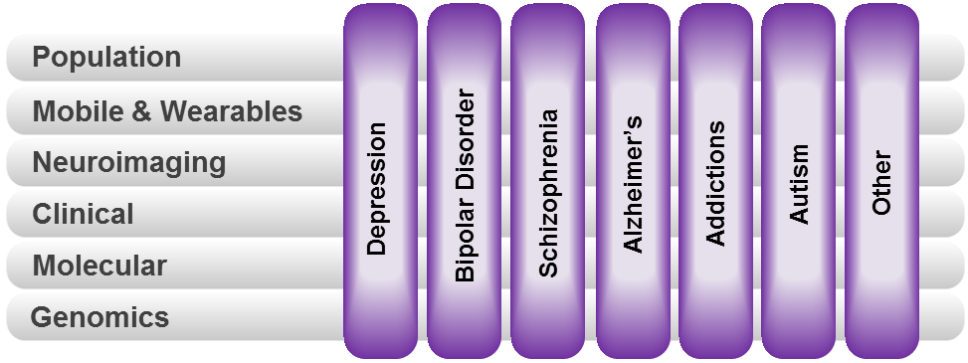
Biobank

- Proteomics
- Genomics

Mobile and Wearable Devices

- Sleep and activity
- Portable EEG

Open Data, Enabling Large Scale Analytics & Modelling



- Open, multiscale brain health atlas across the lifespan and disorders

WELCOME TO OUR CENTRE

Reproducible Neuroinformatics - Why

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

Buckheit and Donoho
WaveLab and Reproducible Research, 1995

Reproducible Neuroinformatics - Why

Your number one collaborator is yourself six months ago: And they don't answer emails

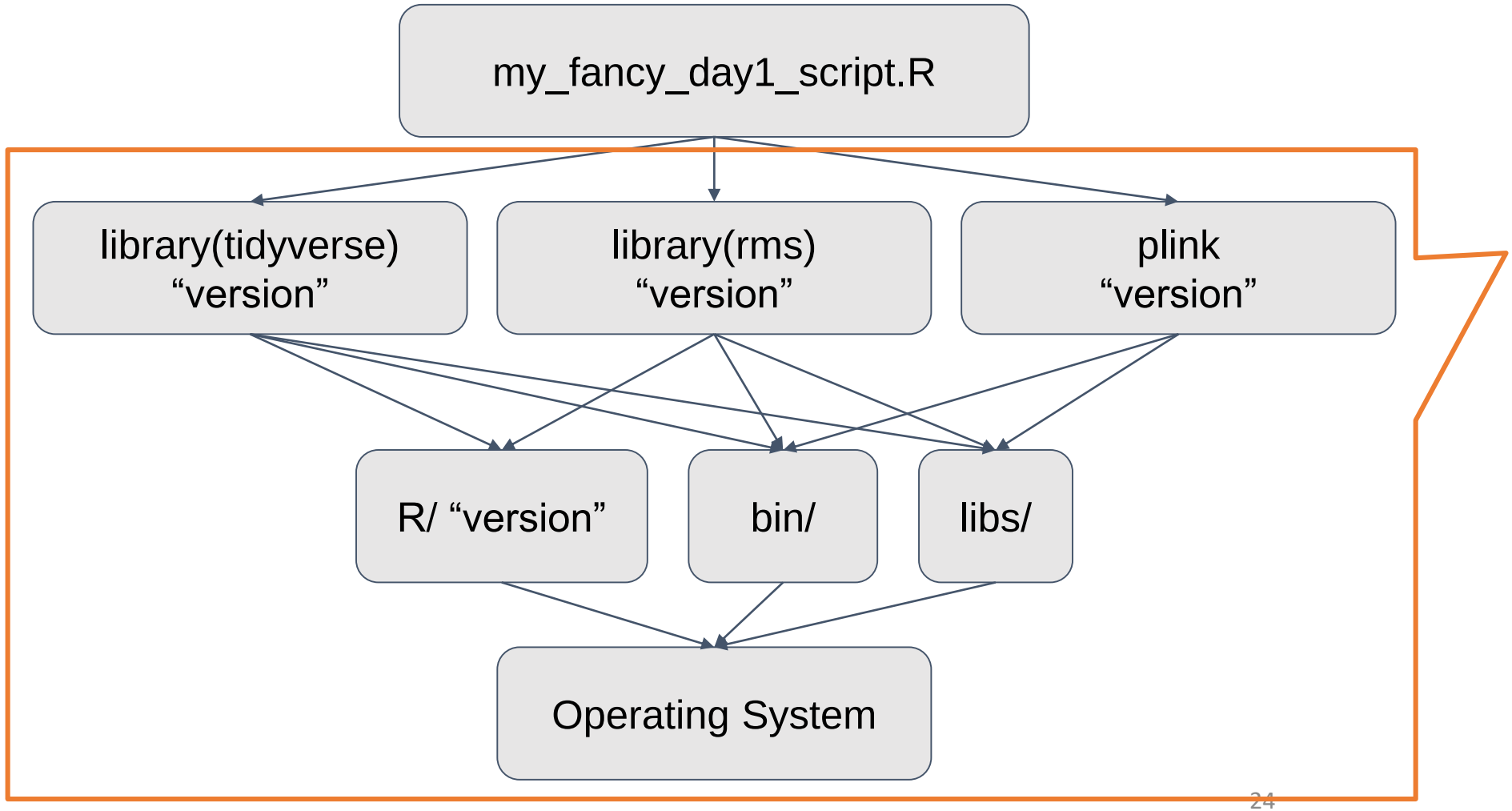
Reproducible Neuroinformatics - Why

camh
Krembil Centre for Neuroinformatics

You
colla
mor
ansv

six
on't

Reproducible Neuroinformatics - Why



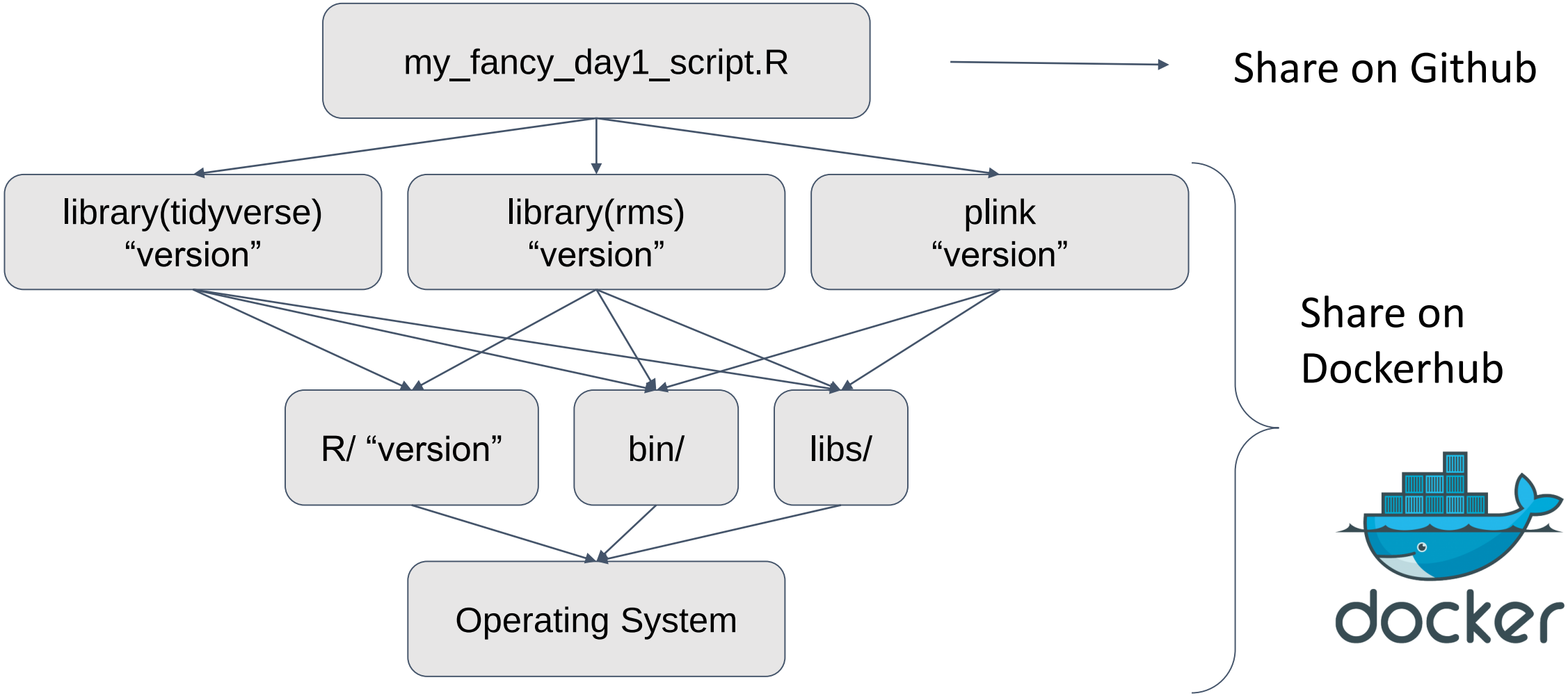
Every box here is something that could be installed differently (*or not at all*) by the next user

This will cause my_fancy_day1_script.R to:

1) crash/not work at all
2) produce unexpected/ different results
3) maybe still work?

Reproducible Neuroinformatics - Solution

camh
Krembil Centre for Neuroinformatics



Git and Github - What and Why?

GitHub

'git' is a tool for tracking what changes to a folder (usually a folder filled with code) when and by who..

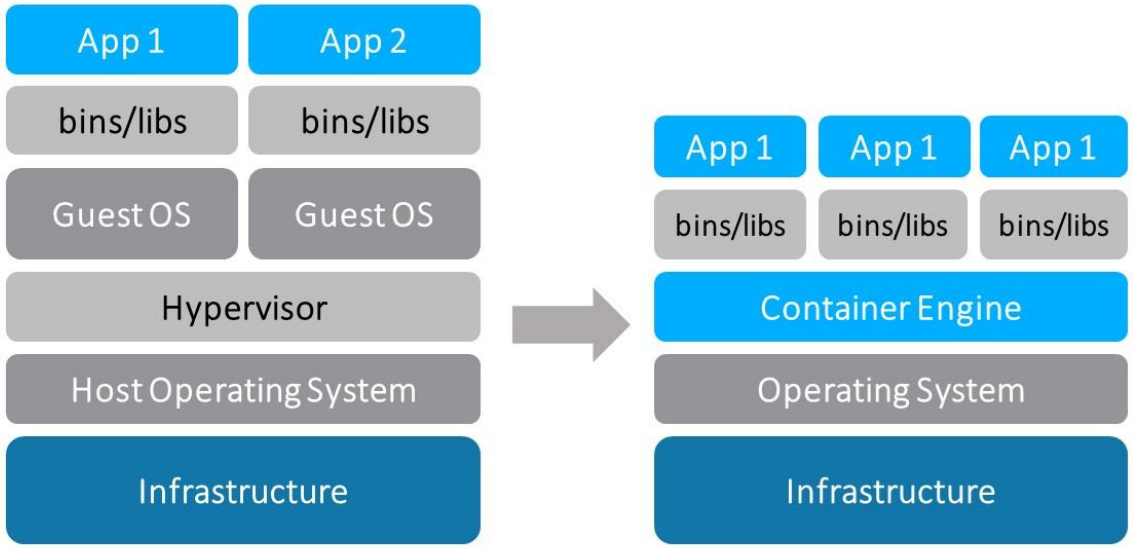
- Like MS Word's "track changes"
  - ...but for code
  - ..and on steroids..

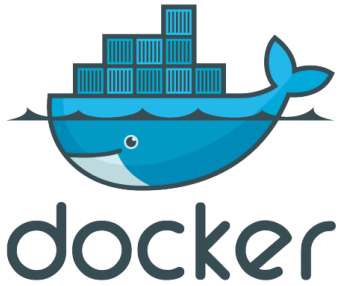
Github is a website where everyone shares there code with themselves, their teams and with the world.

If has a a lot of useful features for:

- working with teams
- reading other people's code
- integrating with other platforms
  - continuous integration (CI) to test for bugs
  - Dockerhub
- hosting documentation websites and wiki's
- releasing versions

Containers - what and why?





Docker - is a tool for sharing software + the dependencies

- the install instructions are stored script called "Dockerfile"
- it's like a virtual machine
  - without a display
  - that takes up a little less disk space
  - that can be installed in one line

Some Docker vocabulary

- **image**: your install of the software
- **container**: one instance of that software that is *usually* still running.

Dockerhub is a website that hosts docker images. So that anyone - anywhere is the world can run it!

Learning steps for today

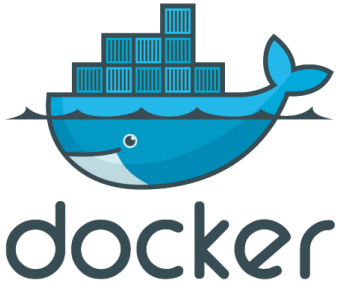
camh
Krembil Centre for Neuroinformatics

Today - altogether - we will try to:

1. **Install Docker Desktop** on our machines
2. Use `docker pull` to get the tutorial software
3. Use `docker run` to start up open up the RStudio instance on your machines
4. Use `git clone` to download the code for School.

Instructions are available at: https://github.com/edickie/kcni-school-envs

Step 1: Install Docker Desktop



Installing Docker *should* not be harder than installing any other program on your computer.[1]

Download link and install instructions at: https://www.docker.com/products/docker-desktop.

To check your install open up a terminal (in windows this is Powershell or WSL) and type:

```
docker run hello-world
```

If this fails! Fear not - we have a plan!

camh
Krembil Centre for Neuroinformatics

If you can't install Docker on your local computer (because you probably don't have enough administrative rights - or you don't have enough space on you home computer) We have a plan!

You can run the software on the SciNet teach cluster!

Instructions are available at: https://github.com/edickie/kcni-school-envs

Step 2: pull the tutorial software

camh
Krembil Centre for Neuroinformatics

In the same terminal window where you typed “docker pull hello-world” now type:

```
docker pull edickie/kcnischool-rstudio:latest
```

Then you should see lots of things happening! What is happening? - *docker is downloading ~ 3G of software for today’s lesson into an “image”*

Copy and paste this line from: https://github.com/edickie/kcni-school-envs

Step 3: run the kcnischool-rstudio image

camh
Krembil Centre for Neuroinformatics

In the same terminal we will now type the run command ***Before running this*** *- replace* <path/to/my/data> *with a folder on your computer where you want to store materials for this course.*

```
docker run --rm -it \
 -e DISABLE_AUTH=true \
 -p 127.0.0.1:8787:8787 \
 -v <path/to/your/data>:/home/rstudio/kcni-school-data \
 edickie/kcnischool-rstudio:latest
```

Step 3: run the kcnischool-rstudio image

camh
Krembil Centre for Neuroinformatics

In the same terminal we will now type the run command
Before running this *- replace* <path/to/my/data> *with a folder on your computer where you want to store materials for this course.*

"docker, please run my image"

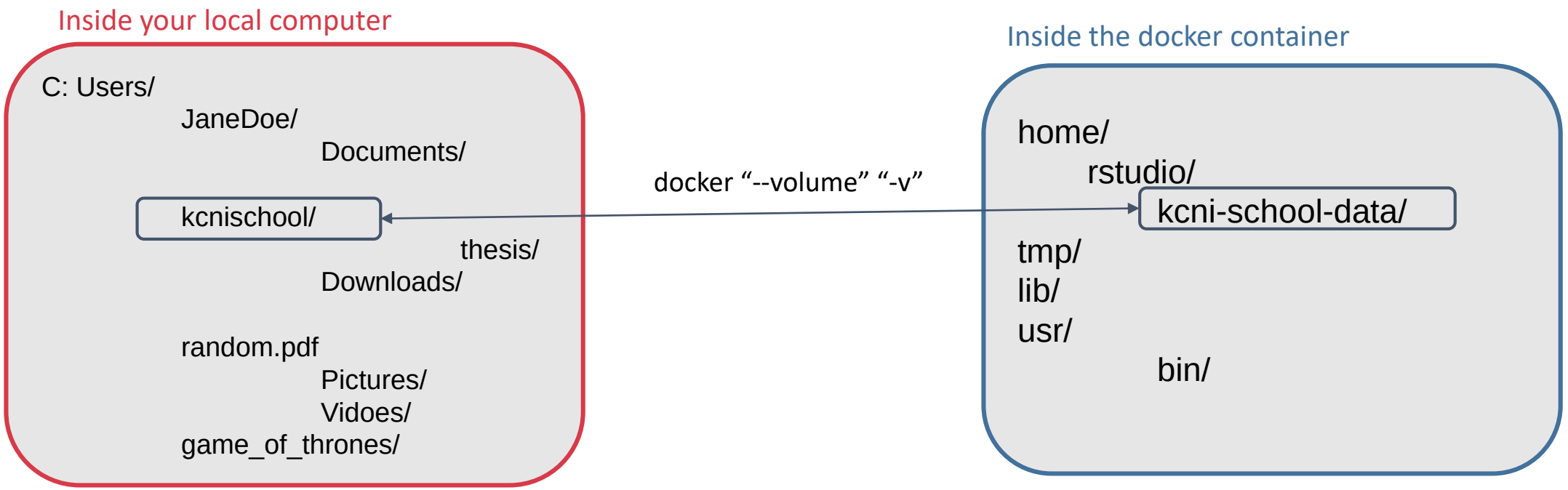
```
docker run --rm -it \
 -e DISABLE_AUTH=true \
 -p 127.0.0.1:8787:8787 \
 -v <path/to/your/data>:/home/rstudio/kcni-school-data \
 edickie/kcnischool-rstudio:latest
```

the docker image name

The concept of Docker “volume mounts”

camh
Krembil Centre for Neuroinformatics

```
docker run --rm -it \
 -e DISABLE_AUTH=true \
 -p 127.0.0.1:8787:8787 \
 -v C:\\Users\JaneDoe\Documents\kcnischool:/home/rstudio/kcni-school-data \
 edickie/kcnischool-rstudio:latest
```



Step 3c: check your browser

camh
Krembil Centre for Neuroinformatics

Copy and paste code from: https://github.com/edickie/kcni-school-envs

```
docker run --rm -it \
 -e DISABLE_AUTH=true \
 -p 127.0.0.1:8787:8787 \
 -v <path/to/your/data>:/home/rstudio/kcni-school-data \
 edickie/kcnischool-rstudio:latest
```

connect port 8787 inside the image to my computer

Open up your browser and go to: http://localhost:8787/

Step 4: Inside Rstudio (terminal)

camh
Krembil Centre for Neuroinformatics

Get the tutorial scripts by typing

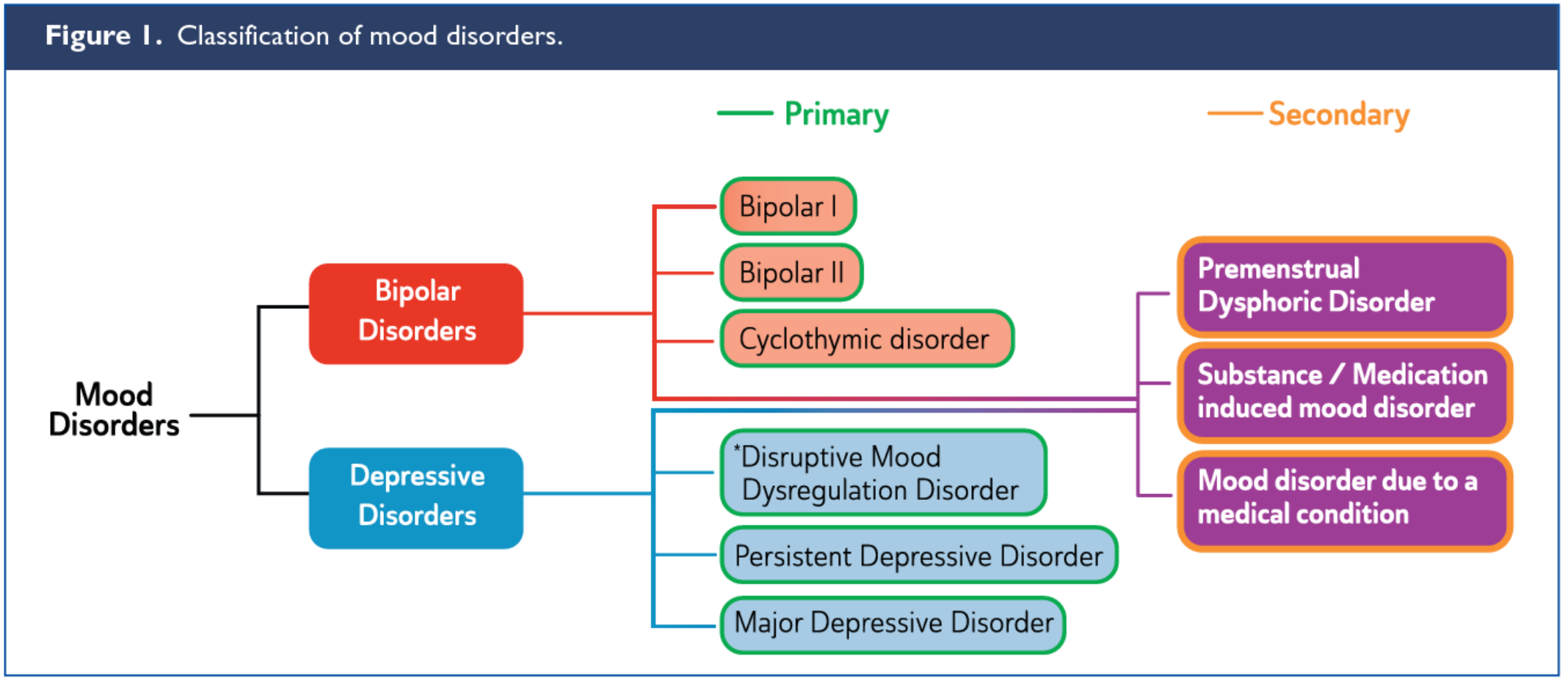
```
cd /home/rstudio/kcni-school-data

git clone --recurse-submodules \
https://github.com/krembilneuroinformatics/kcni-school-lessons.git
```

Major Depressive Disorder (MDD)

Mood Disorder Classification

camh
Krembil Centre for Neuroinformatics



(Malhi et al., 2015, Aust. N. Z. J. Psychiatry)

Mood Disorder Classification

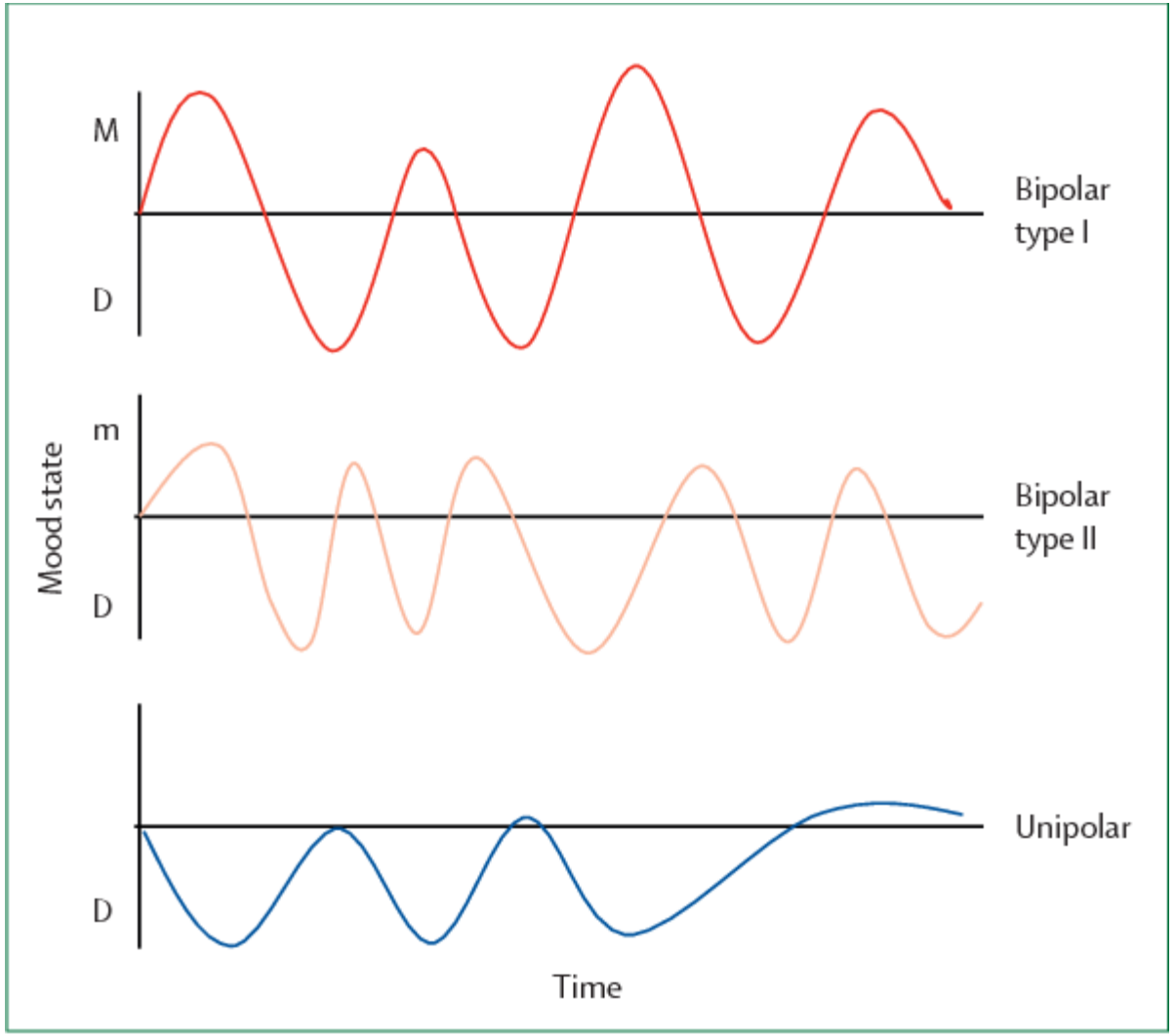
camh
Krembil Centre for Neuroinformatics

Classified among mood disorders

"Major Depression"
"Unipolar Depression"
"Major Depressive Disorder"

Depressive subtypes/considerations:

- Dysthymia (mild, persistent depression)
- Seasonal Affective Disorder
- Atypical Depression
- Psychotic Depression
- Postpartum / Antepartum
- Premenstrual Depression
- Situational Depression



(Phillips and Kupfer, 2013, Lancet)

Major Depressive Disorder (MDD)

camh
Krembil Centre for Neuroinformatics

Diagnostic and Statistical Manual – 5 (DSM-5) Criteria:

- **Five or more** consistently within 2 week period, emerging from previous functioning:
  1. Depressed mood* (essential for diagnosis)
  2. Diminished pleasure/interest in all or most activities* (essential for diagnosis)
  3. Weight loss, decreased appetite
  4. Sleep disturbance
  5. Psychomotor agitation or retardation
  6. Fatigue
  7. Feelings of worthlessness or guilt
  8. Diminished ability to think or concentrate
  9. Recurrent thoughts of death; suicidal ideation

- symptoms **cause distress or impaired functioning** in life
- not attributable to **substance use** or other health condition
- not better explained by co-morbid **psychotic illness**
- No history of **manic or hypomanic episode**

(Otte et al., 2016, Nat Rev.)

Manic and Depressive Classification

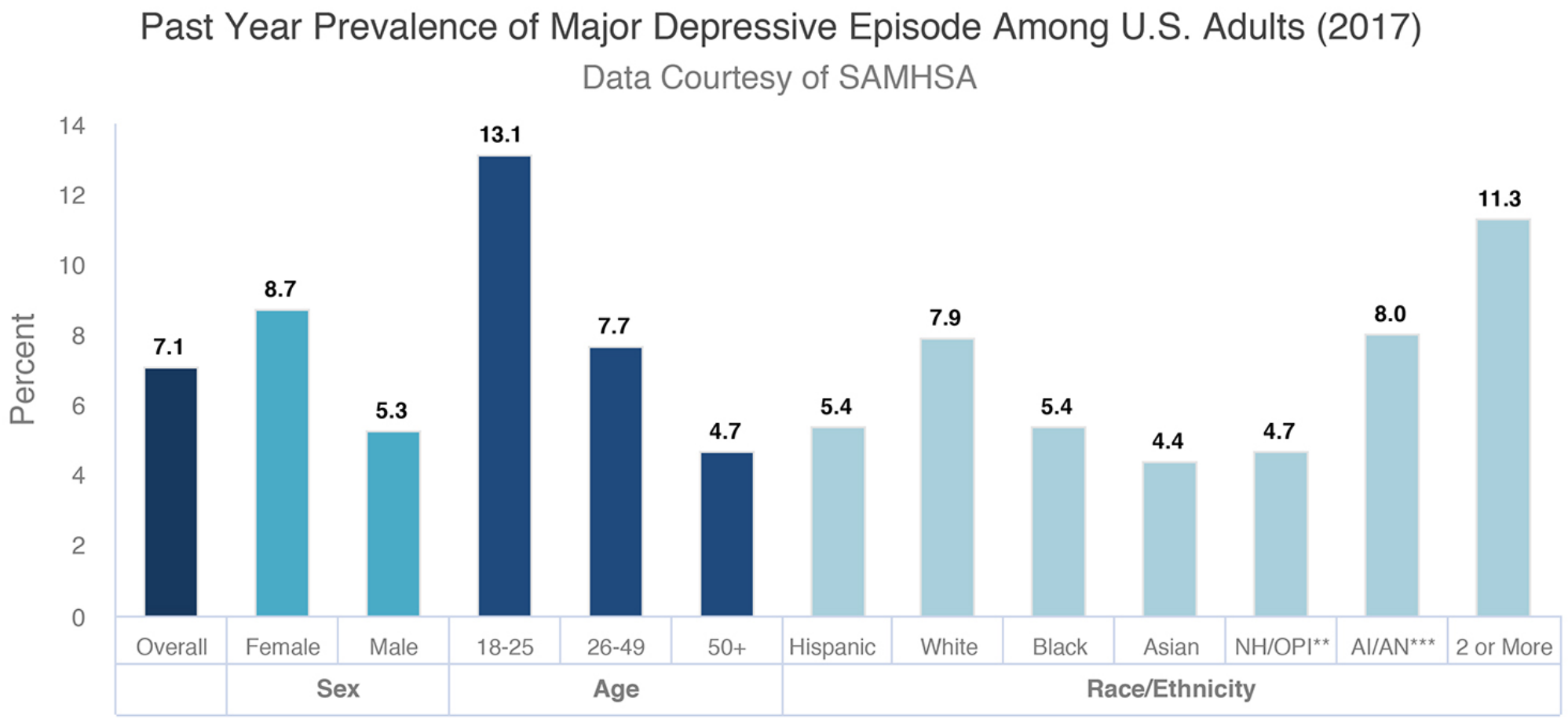
camh
Krembil Centre for Neuroinformatics

Core symptoms	Elevated or irritable mood		Elevated or irritable mood + depressed mood or loss of interest		Depressed mood or loss of interest
Manic	≥3 or ≥4		≥3		
Depressive			≥5		≥5
DSM-IV	Manic		Mixed		Depressive
DSM-5	Manic	Manic with mixed features	← →	Depressive with mixed features	Depressive
Core symptoms	Elevated or irritable mood + energy or activity	Elevated or irritable mood + energy or activity		Depressed mood or loss of interest	Depressed mood or loss of interest
Manic	≥3 or ≥4	≥3 or 4		≥3	
Depressive		≥3		≥5	≥5

(Calabrese et al., 2017, Am J Psych)

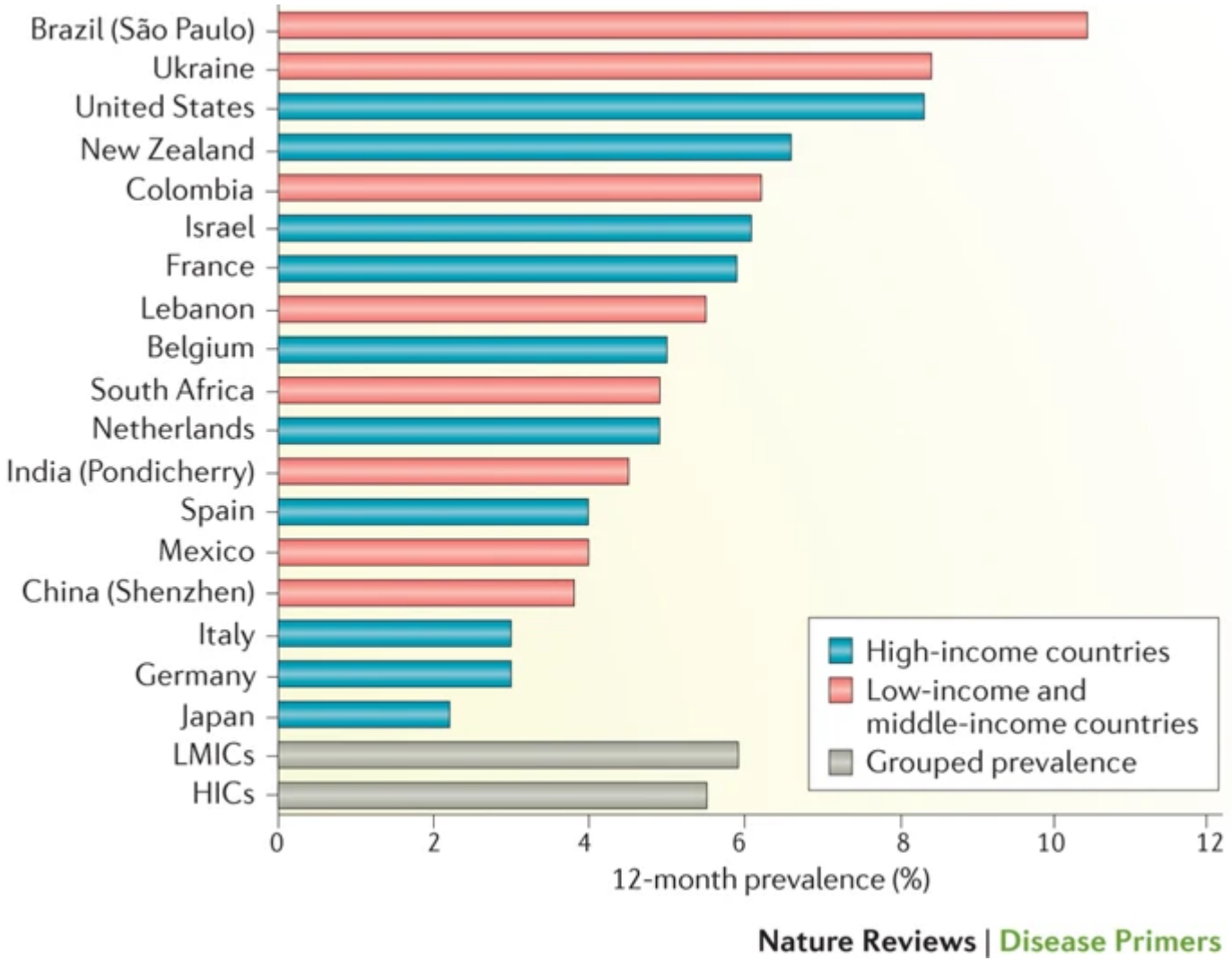
Major Depressive Disorder (MDD)

camh
Krembil Centre for Neuroinformatics



(Otte et al., 2016, Nat Rev.)

Major Depressive Disorder (MDD)



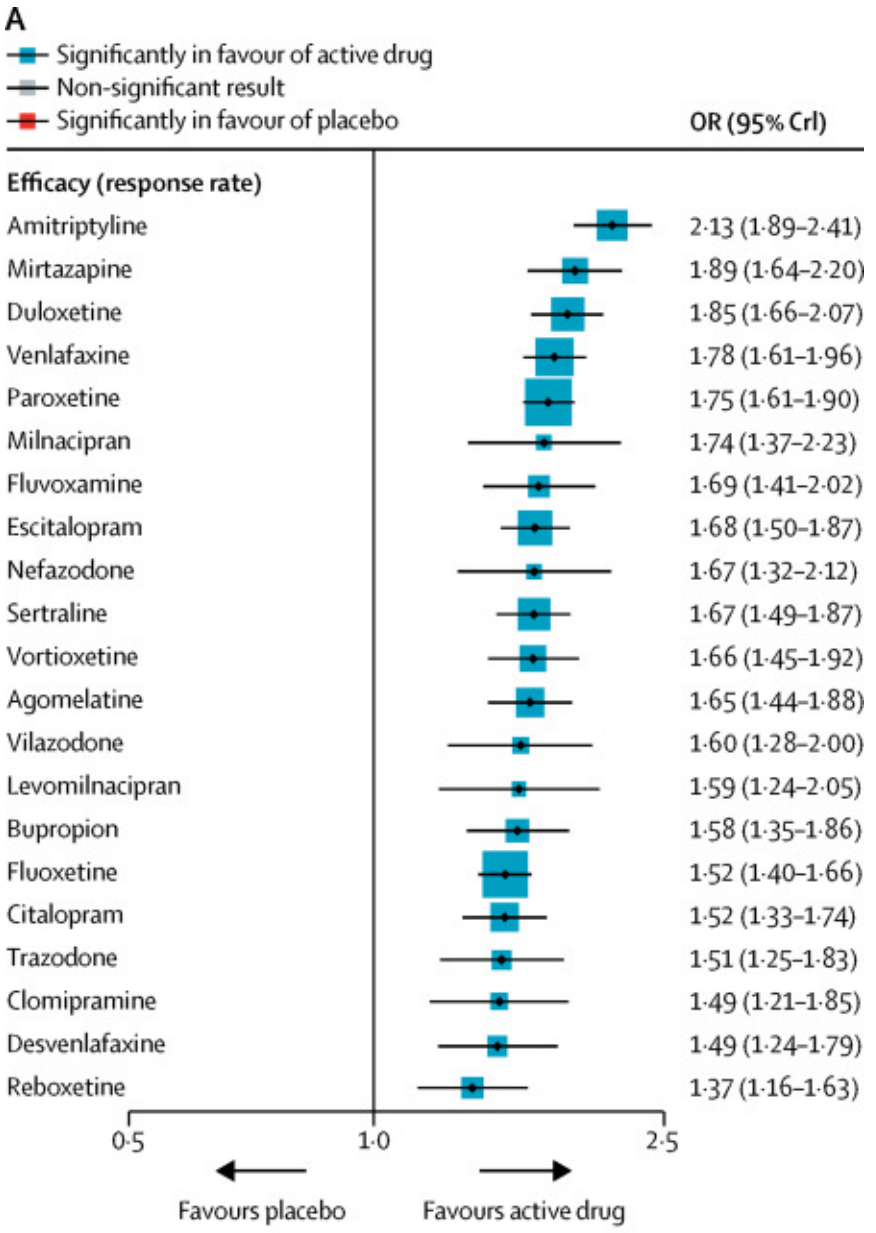
(Otte et al., 2016, Nat Rev.)

Major Depressive Disorder (MDD)

camh
Krembil Centre for Neuroinformatics

The good news...

522 double-blind studies
n=116 477



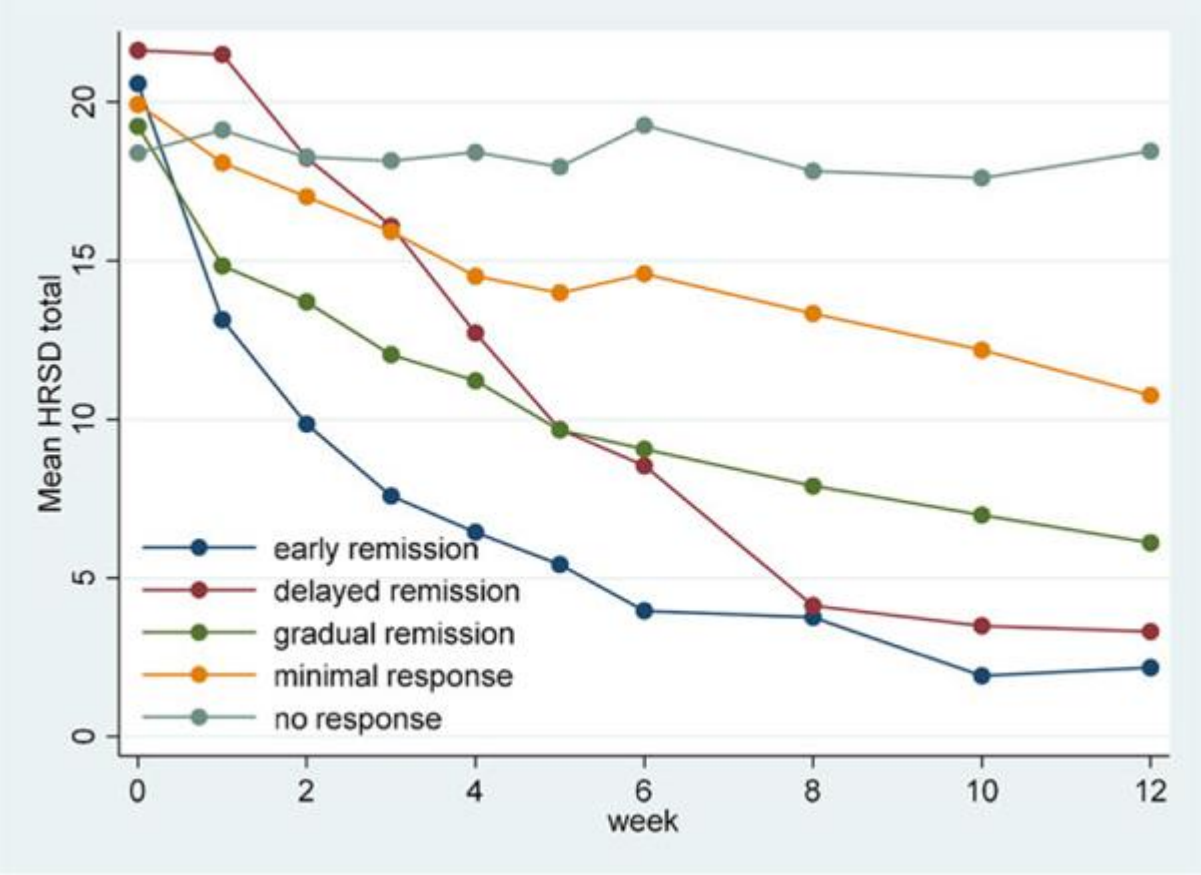
(Cipriani et al., 2018, Lancet)

Major Depressive Disorder (MDD)

camh
Krembil Centre for Neuroinformatics

The not-so-good news...

(The Predictors of Remission in Depression to Individual and Combined Treatments [PReDICT] study)



Baseline controlled (bc-ΔHRSD) class	N	Total (%)	Baseline HRSD total
Early remission	43	12.5	20.6 (3.6)
Delayed remission	24	7.0	21.6 (4.3)
Gradual remission	111	32.3	19.2 (3.5)
Minimal response	138	40.1	19.9 (4.0)
No response	28	8.1	18.4 (3.0)
Total			

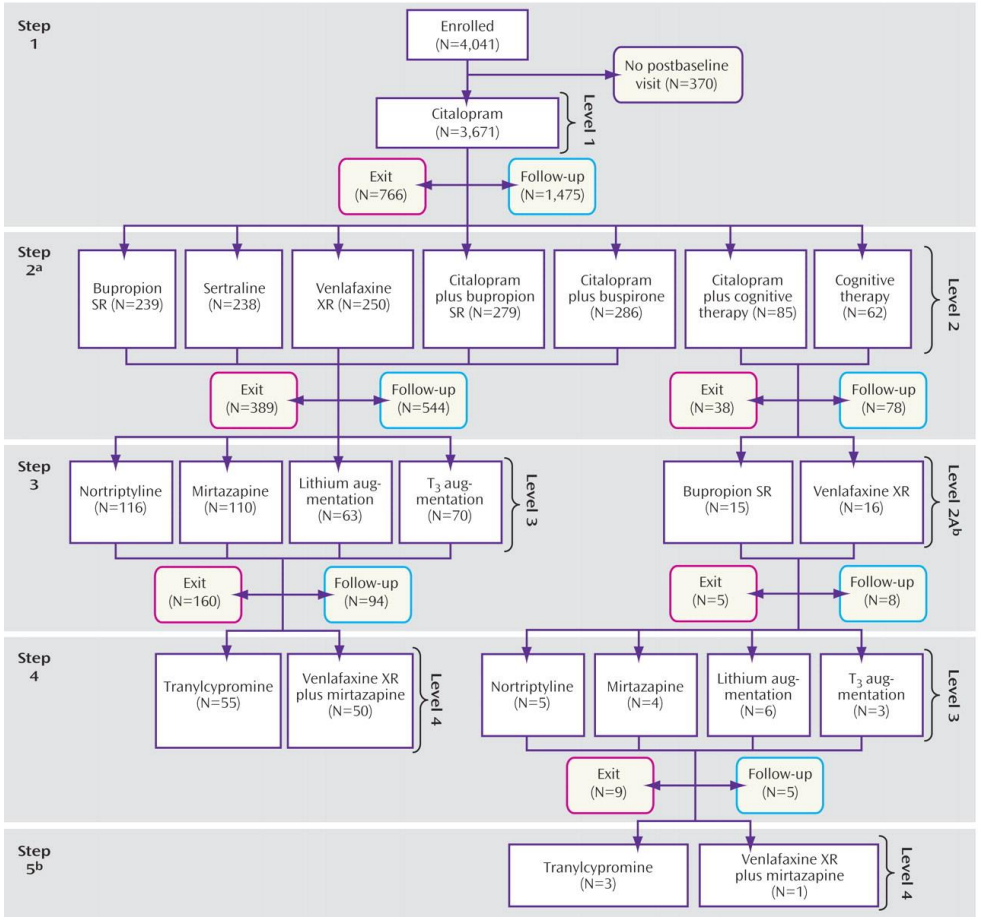
(Kelly et al., 2018, Depress Anxiety)

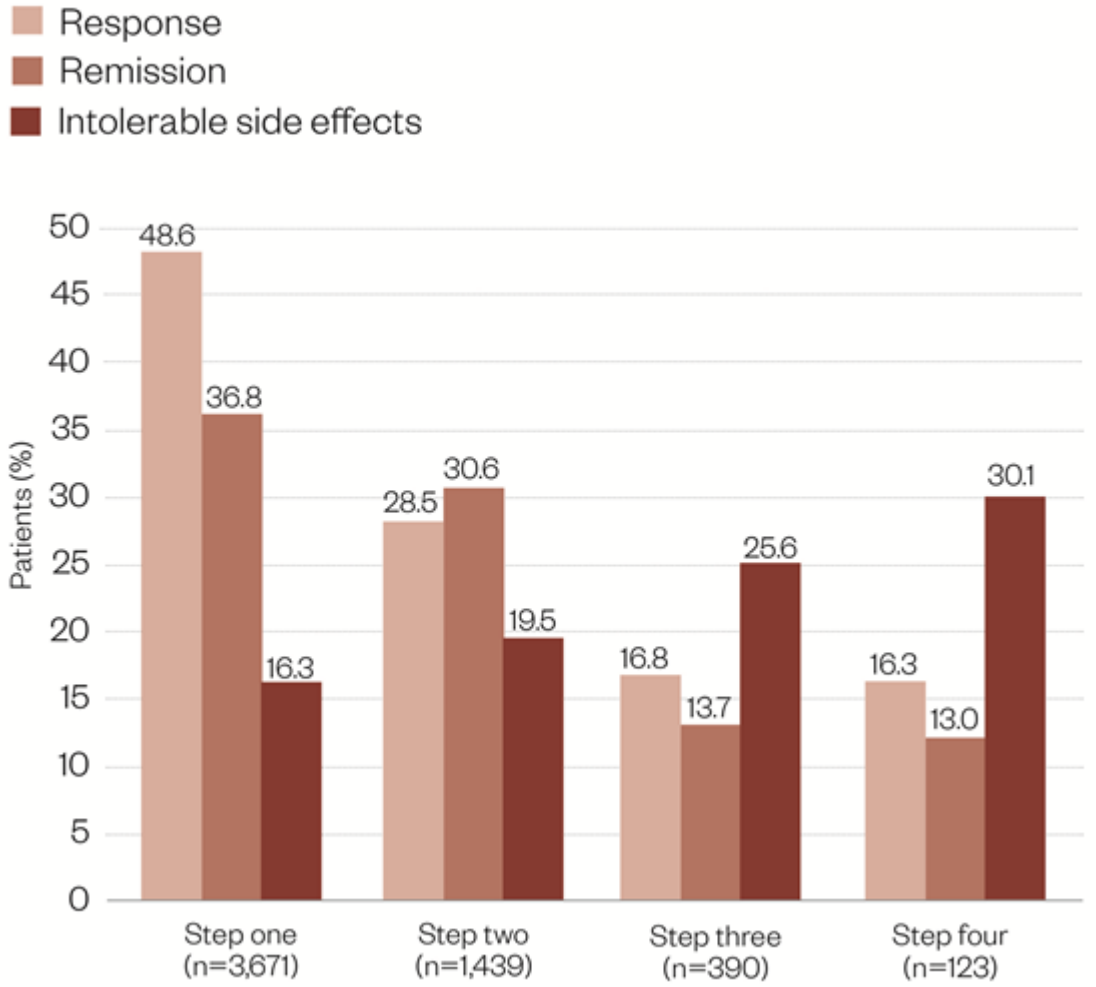
Major Depressive Disorder (MDD)

camh
Krembil Centre for Neuroinformatics

The not-so-good news...

(NIMH Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study)





(Rush et al., 2006, Am J Psych.)

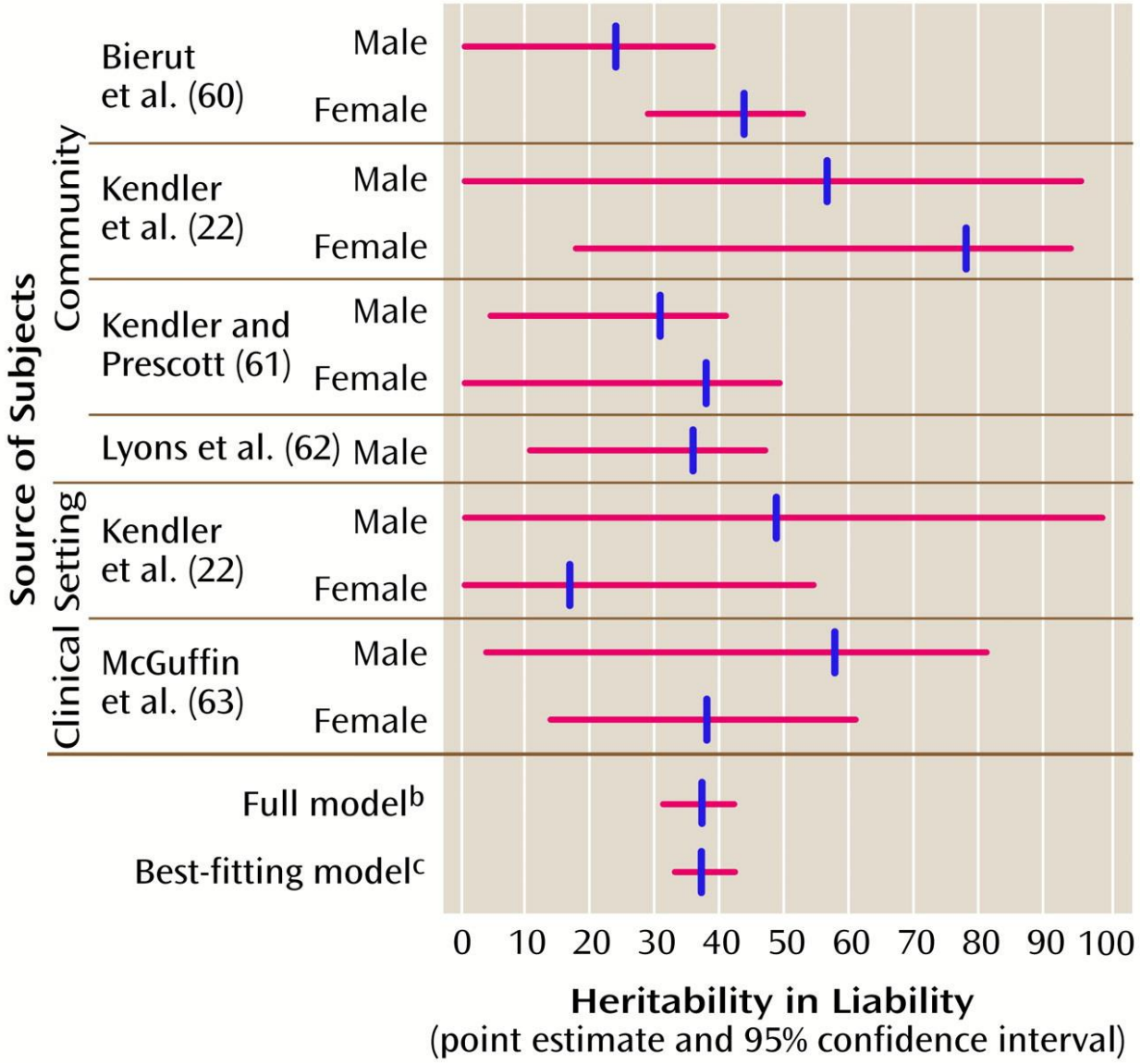
Genetics of MDD - Heritability

camh
Krembil Centre for Neuroinformatics

Family and twin-based meta-analysis

~40% heritability

The component of disease risk that is **attributable to genetic factors**, as opposed to environment (nature vs. nurture).

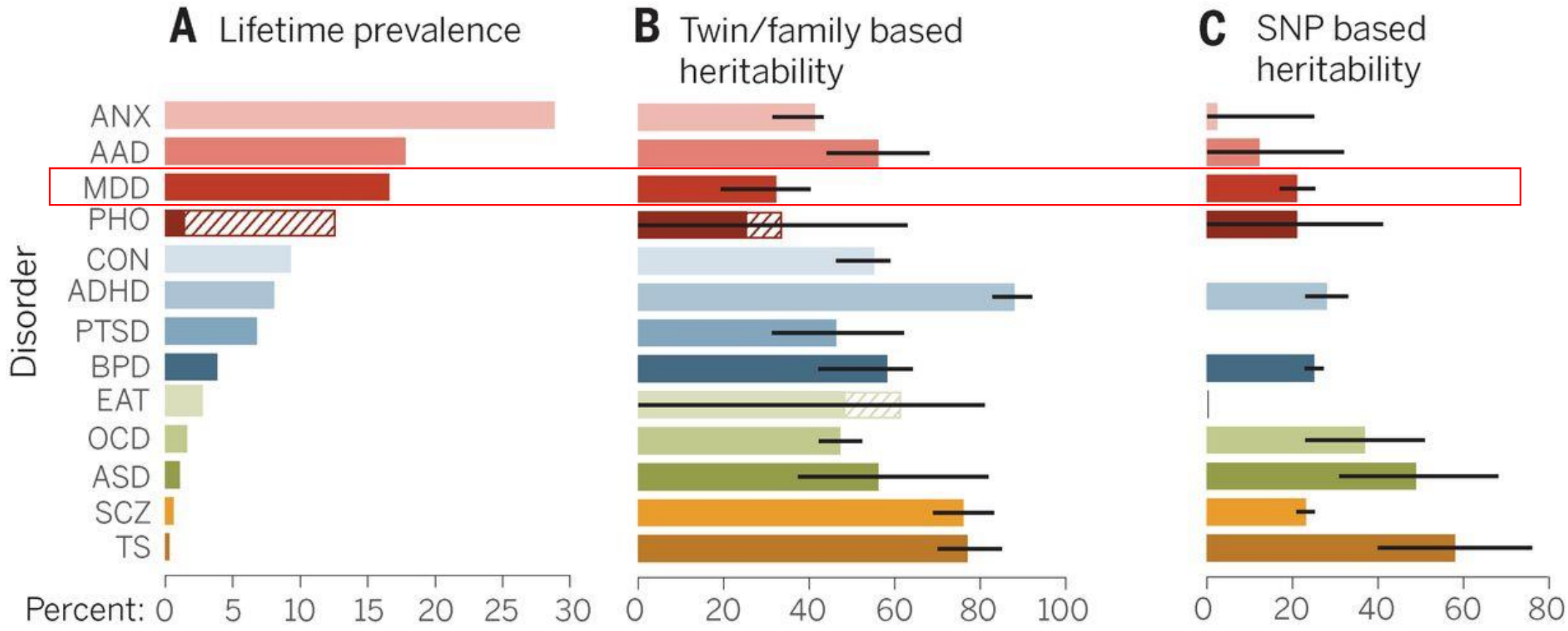


(Sullivan et al., 2000, Am J Psych.)

Genetics of MDD - Heritability

camh
Krembil Centre for Neuroinformatics

ANX = any anxiety disorder; **AAD** = alcohol abuse disorder; **ASD** = autism spectrum disorder; **BPD** = bipolar disorder; **MDD** = major depressive disorder; **PHO** = any phobia; **CON** = conduct disorders; **OCD** = obsessive compulsive disorder; **PTSD** = post traumatic stress disorder; **EAT** = eating disorders; **TS** =Tourette syndrome



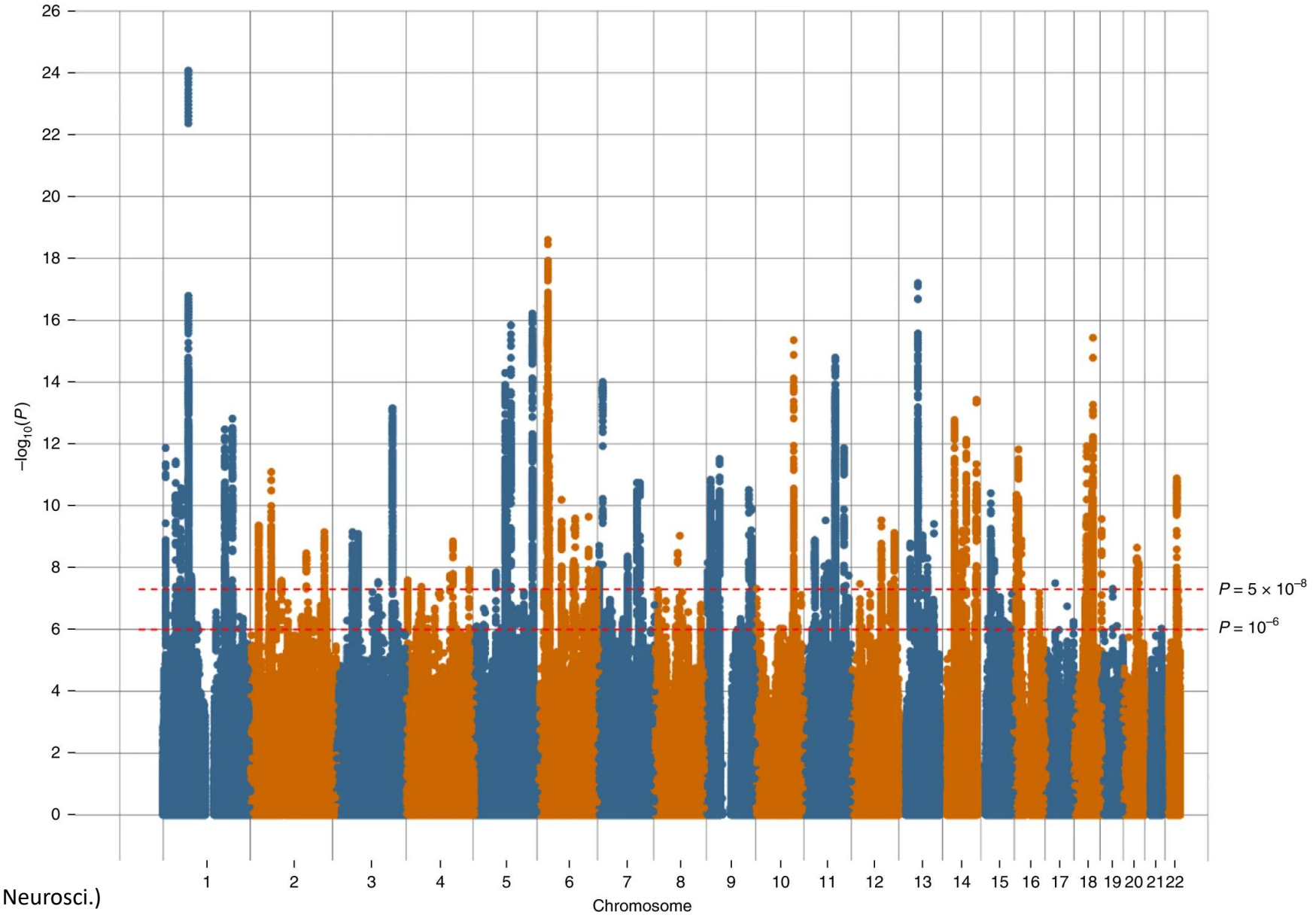
(Geschwind and Flint, 2015)

Genetics of MDD

camh
Krembil Centre for Neuroinformatics

n=807,553
meta-analysis

h_{snp}=0.089



(Howard et al., 2019, Nat Neurosci.)

Genetics of MDD

camh
Krembil Centre for Neuroinformatics

"Missing Heritability"

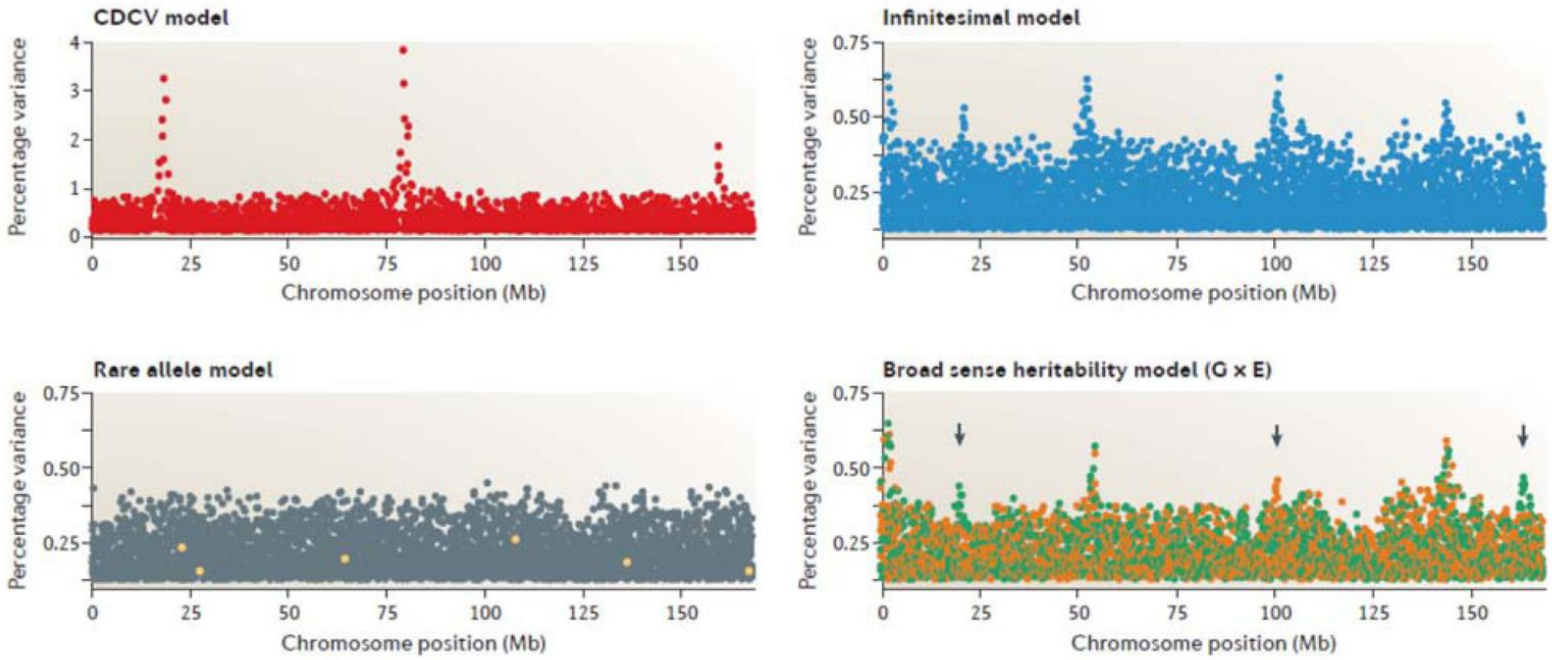
Measurement error

Phenotypic Heterogeneity

Genetic models:
Common-disease common-variant

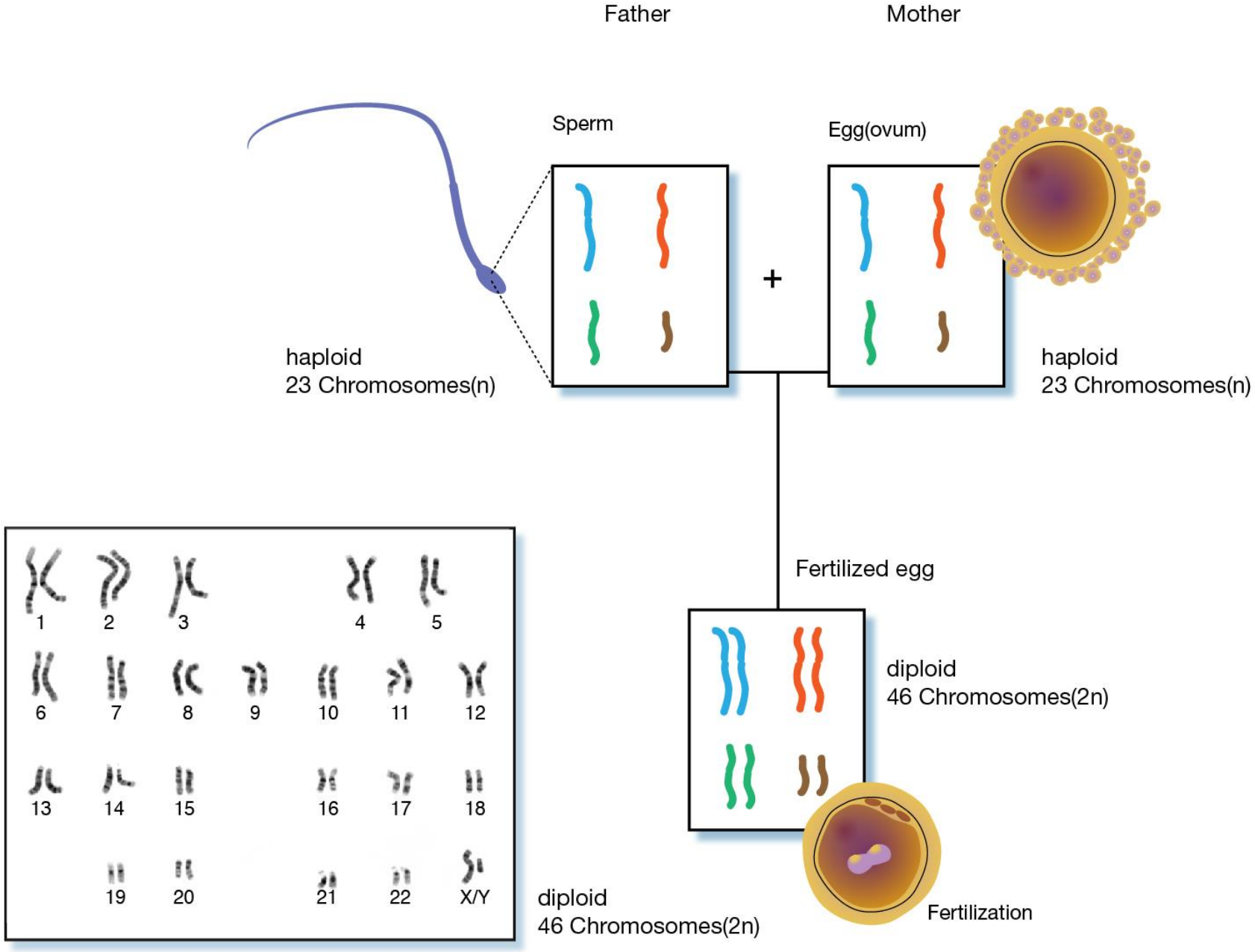
Infinitesimal model (Fisher 1918)
Rare allele model
Broad sense heritability model
Omnigenic model (Boyle 2017)

Structural variation (CNVs, Indels)

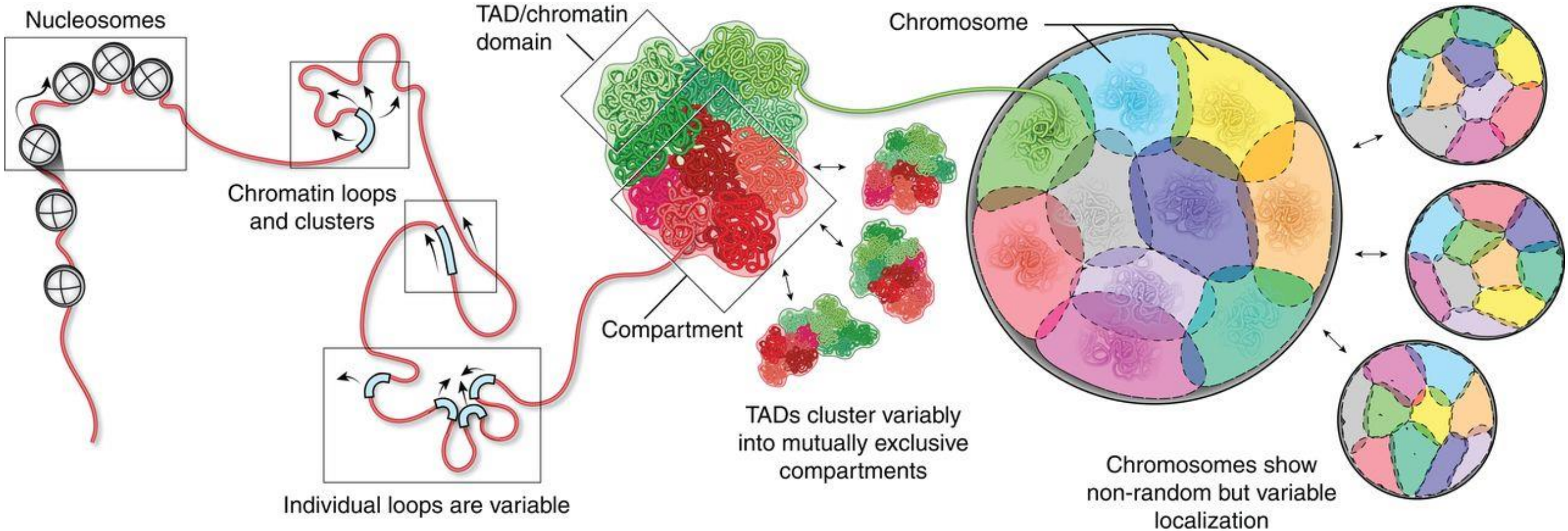


(Gibson, 2015, Nat. Rev. Genet)

Genetic Variation and Genotyping

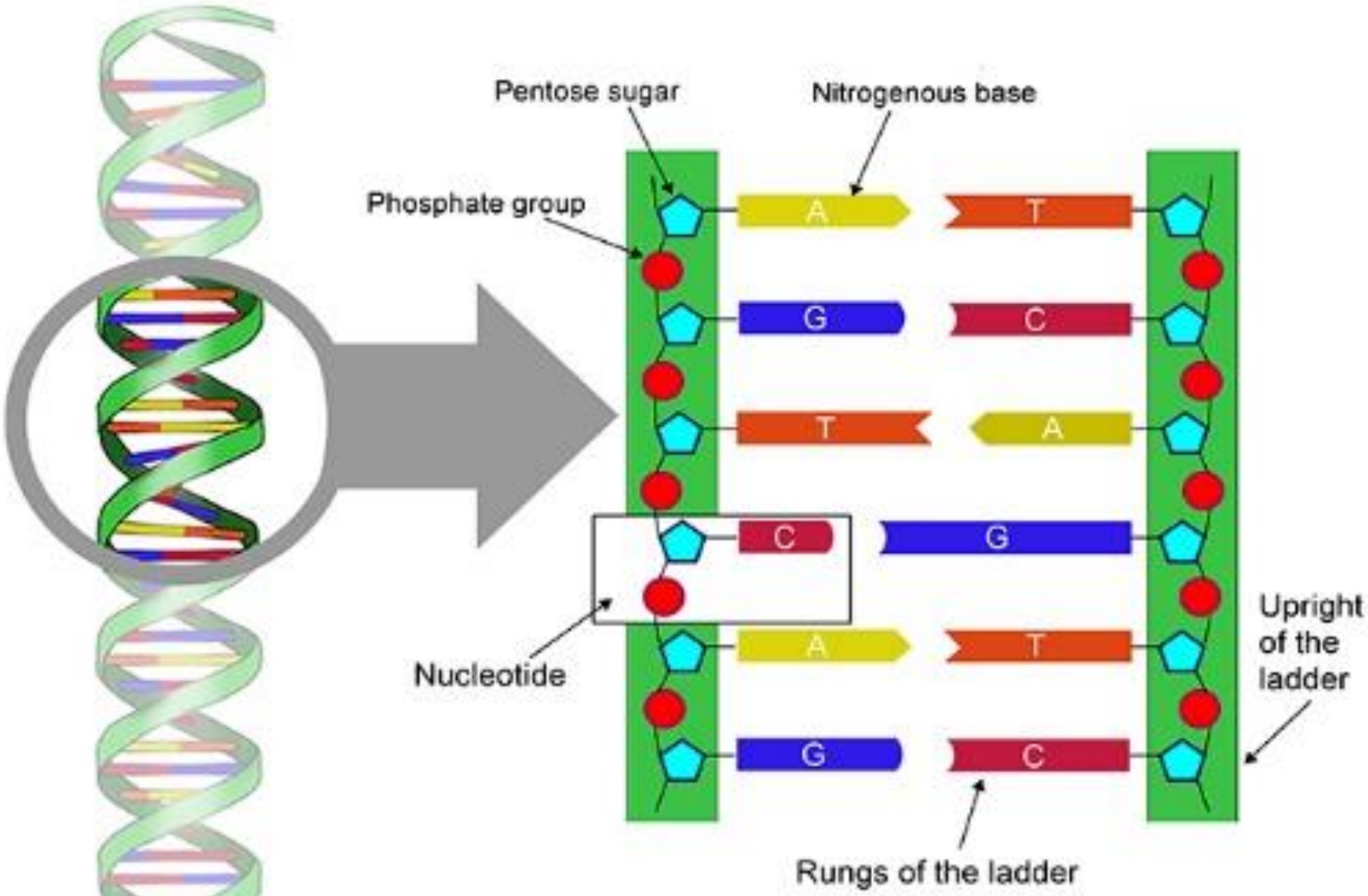


(National Human Genome Research Institute, NIH)



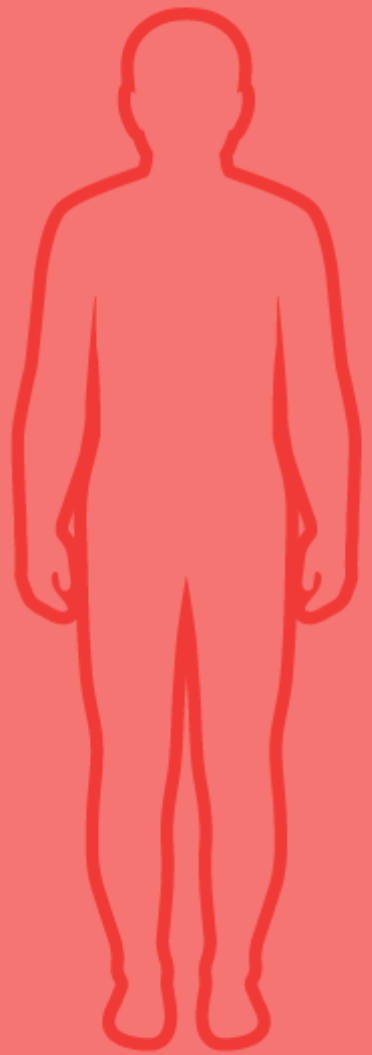
TAD: topologically associating domain

(Finn and Misteli, 2019, Science)


Pentose sugar
Nitrogenous base
Phosphate group
A
T
G
C
T
A
C
G
Nucleotide
A
T
G
C
Upright of the ladder
Rungs of the ladder

The genetic similarity between a **human** and a **human** is...

99.9%



SOURCE: National Human Genome Research Institute

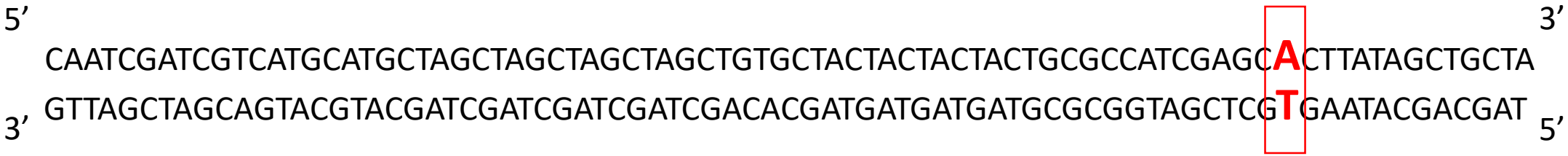
BUSINESS INSIDER

5’ 3’

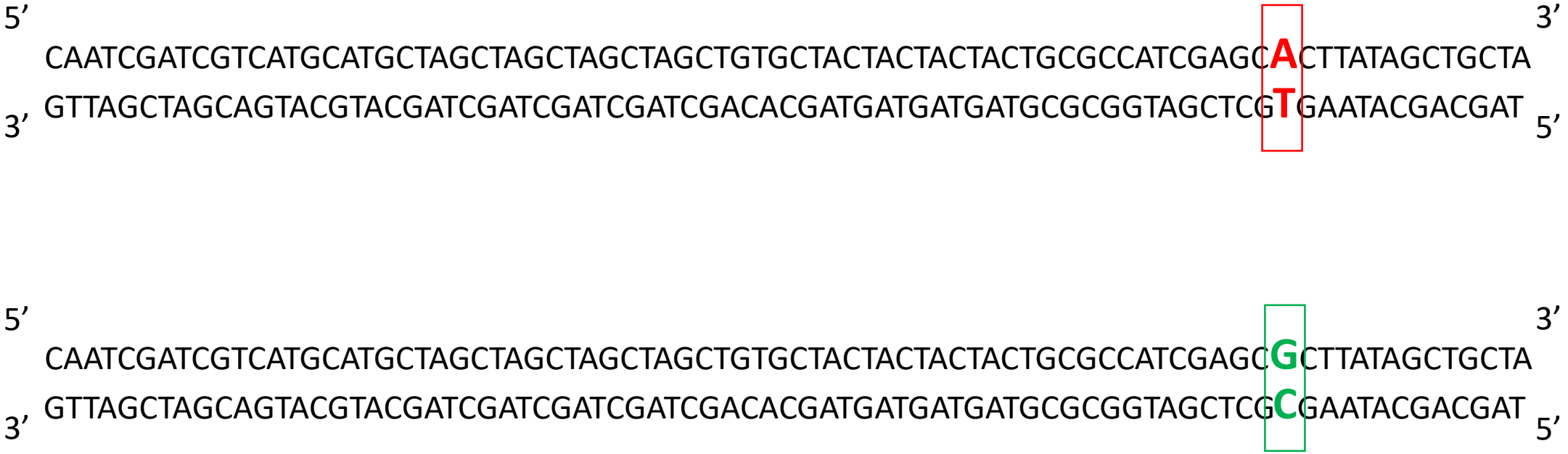
CAATCGATCGTCATGCATGCTAGCTAGCTAGCTAGCTGTGCTACTACTACTACTGCGCCATCGAGCACTTATAGCTGCTA

3’ GTTAGCTAGCAGTACGTACGATCGATCGATCGATCGACACGATGATGATGATGCGCGGTAGCTCGTGAATACGACGAT 5’

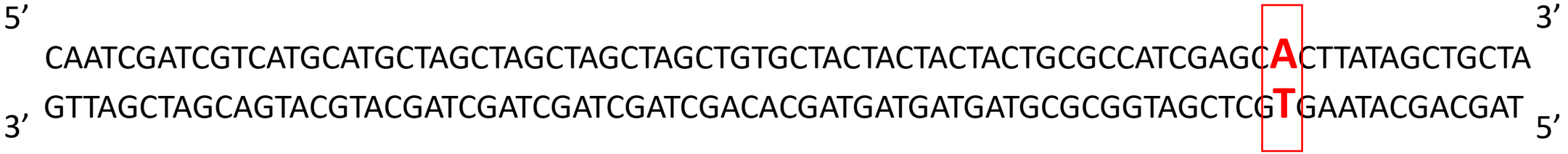
SNP rs123 [A/G]

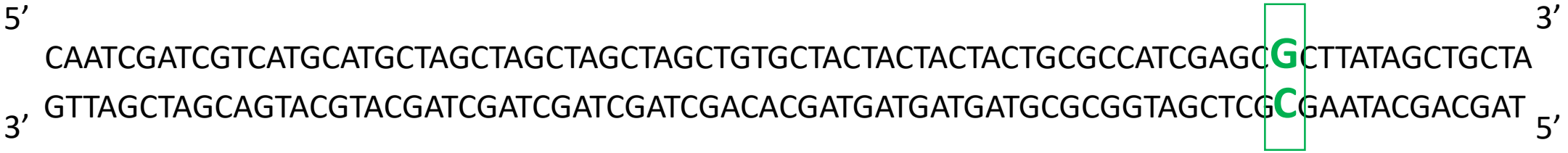

5’
3’
CAATCGATCGTCATGCATGCTAGCTAGCTAGCTAGCTGTGCTACTACTACTACTGCGCCATCGAGCACTTATAGCTGCTA
3’
GTTAGCTAGCAGTACGTACGATCGATCGATCGATCGACACGATGATGATGATGCGCGGTAGCTCGTGAATACGACGAT
5’

SNP rs123 [A/G]


5’
3’
CAATCGATCGTCATGCATGCTAGCTAGCTAGCTAGCTGTGCTACTACTACTACTGCGCCATCGAGCACTTATAGCTGCTA
3’
GTTAGCTAGCAGTACGTACGATCGATCGATCGATCGACACGATGATGATGATGCGCGGTAGCTCGTGAATACGACGAT
5’
5’
3’
CAATCGATCGTCATGCATGCTAGCTAGCTAGCTAGCTGTGCTACTACTACTACTGCGCCATCGAGCGCTTATAGCTGCTA
3’
GTTAGCTAGCAGTACGTACGATCGATCGATCGATCGACACGATGATGATGATGCGCGGTAGCTCGCGAATACGACGAT
5’

SNP rs123 [A/G]

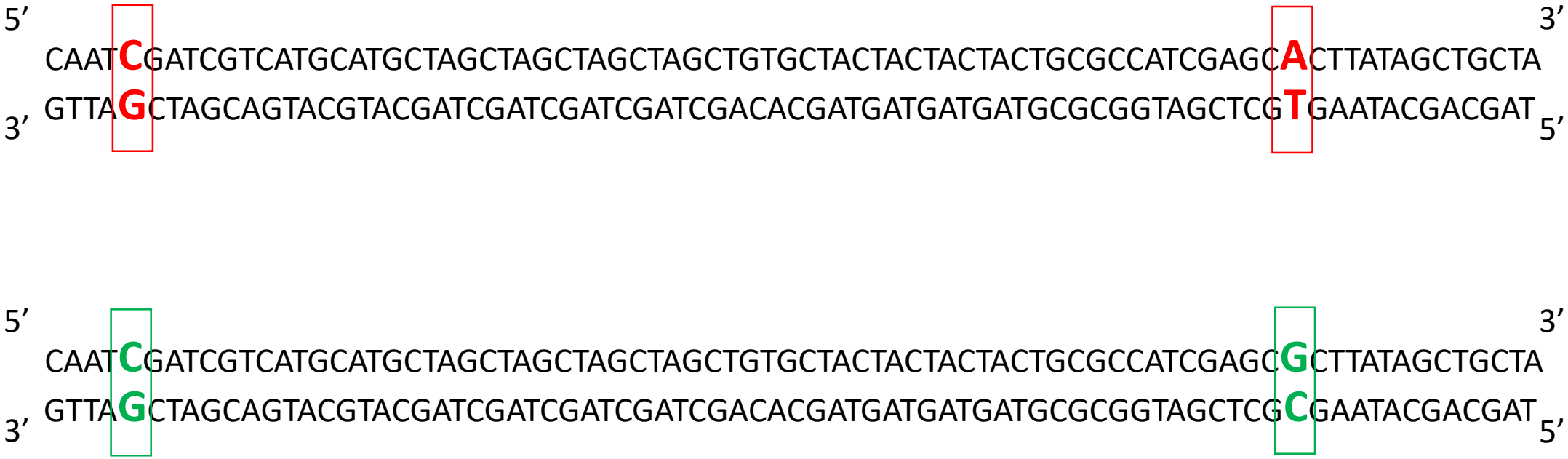




Individual is A/G genotype for SNP rs123 (**heterozygous**)

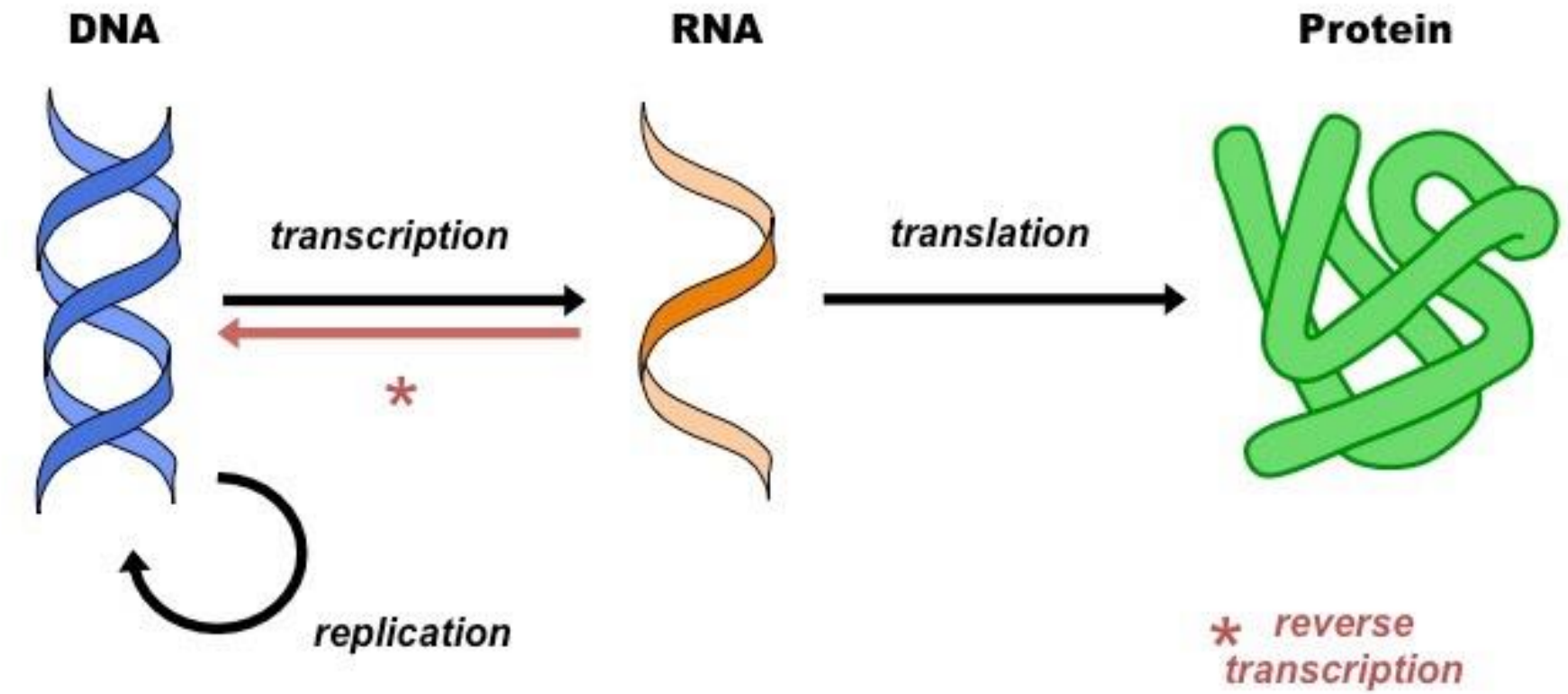
SNP rs456 [C/A]

SNP rs123 [A/G]

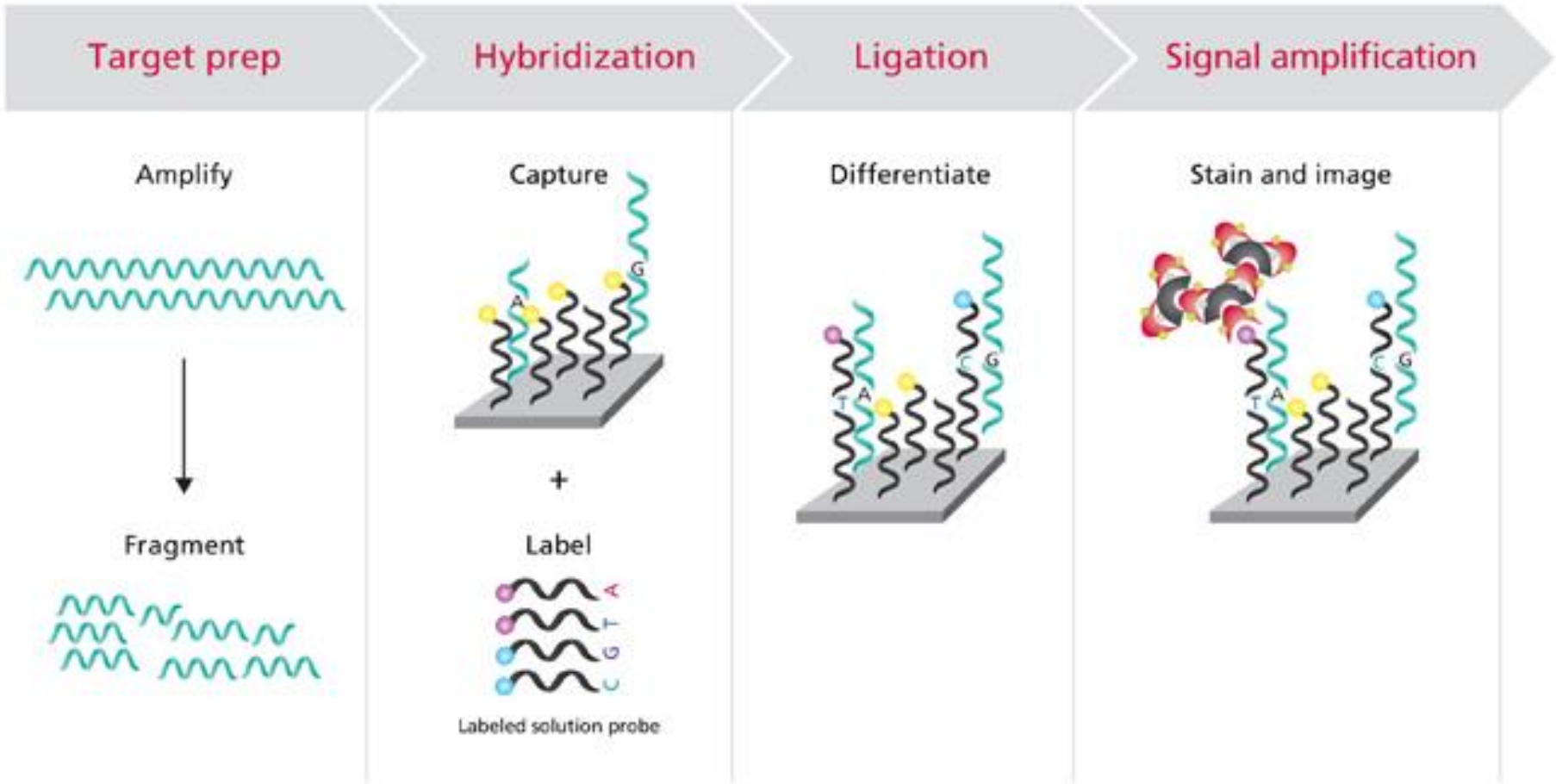


Individual is A/G genotype for SNP rs123 (**heterozygous**)
Individual is C/C genotype for SNP rs456 (**homozygous**)

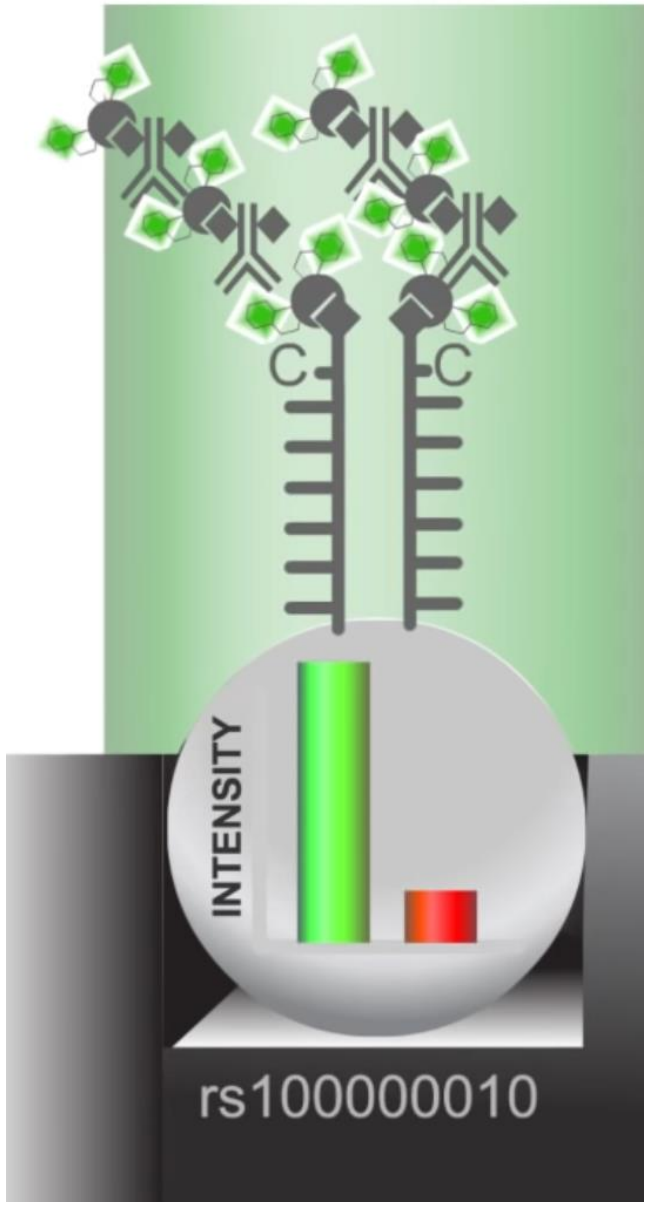
The Central Dogma



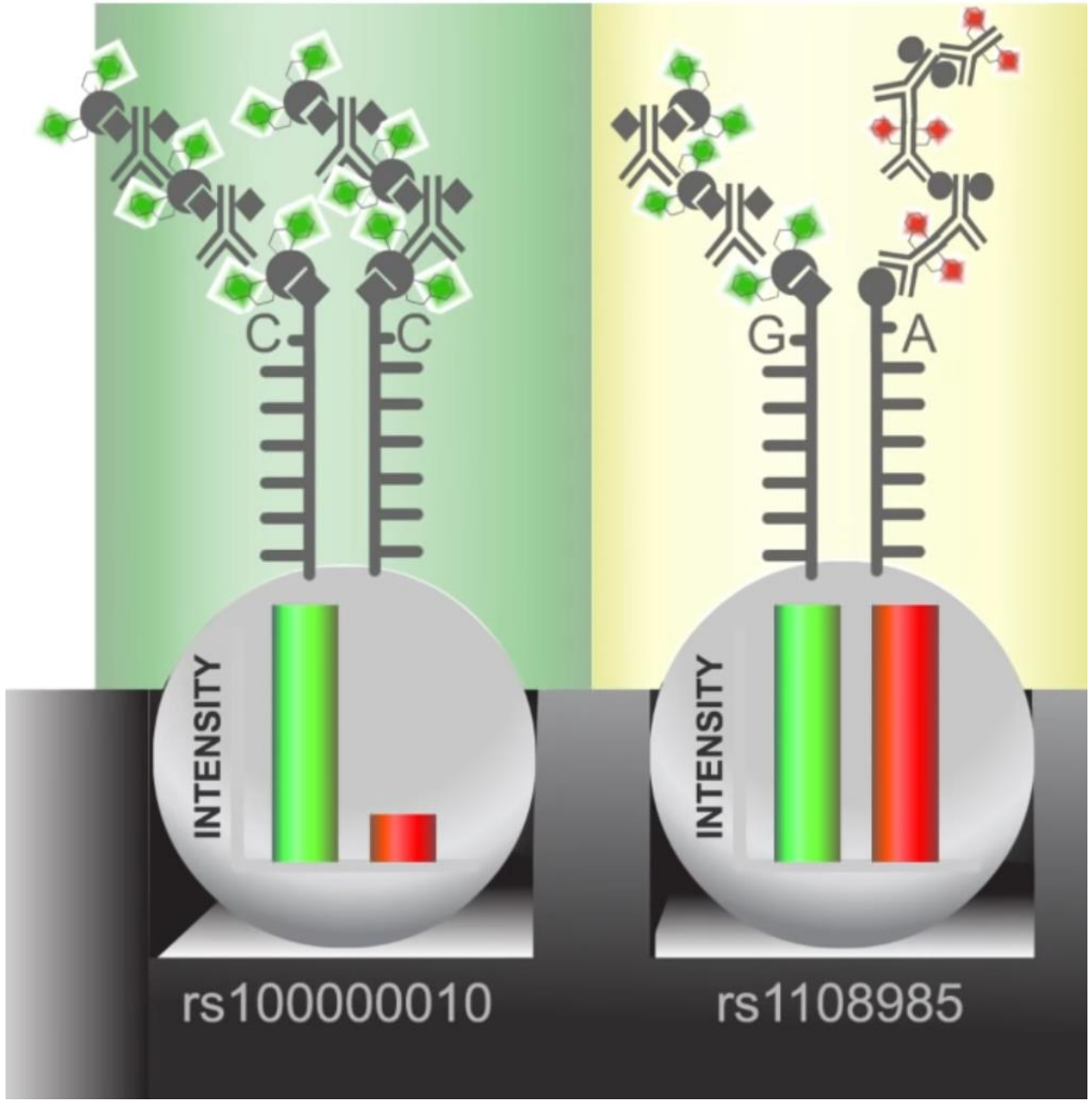
(BioNinja, 2020)


Target prep
Hybridization
Ligation
Signal amplification
Amplify
Fragment
Capture
+
Label
Labeled solution probe
Differentiate
Stain and image

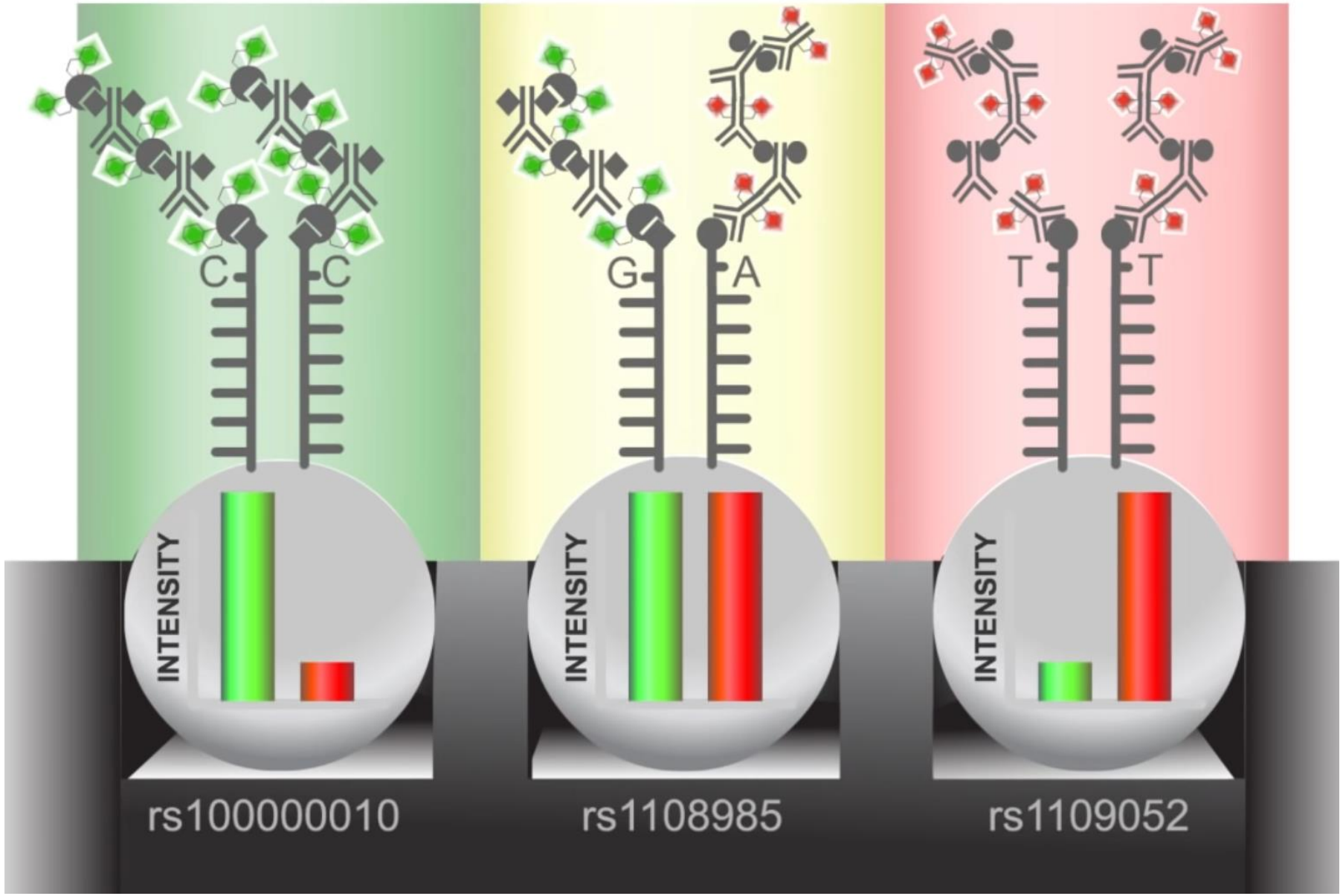
(Affymetrix)



(Illumina)

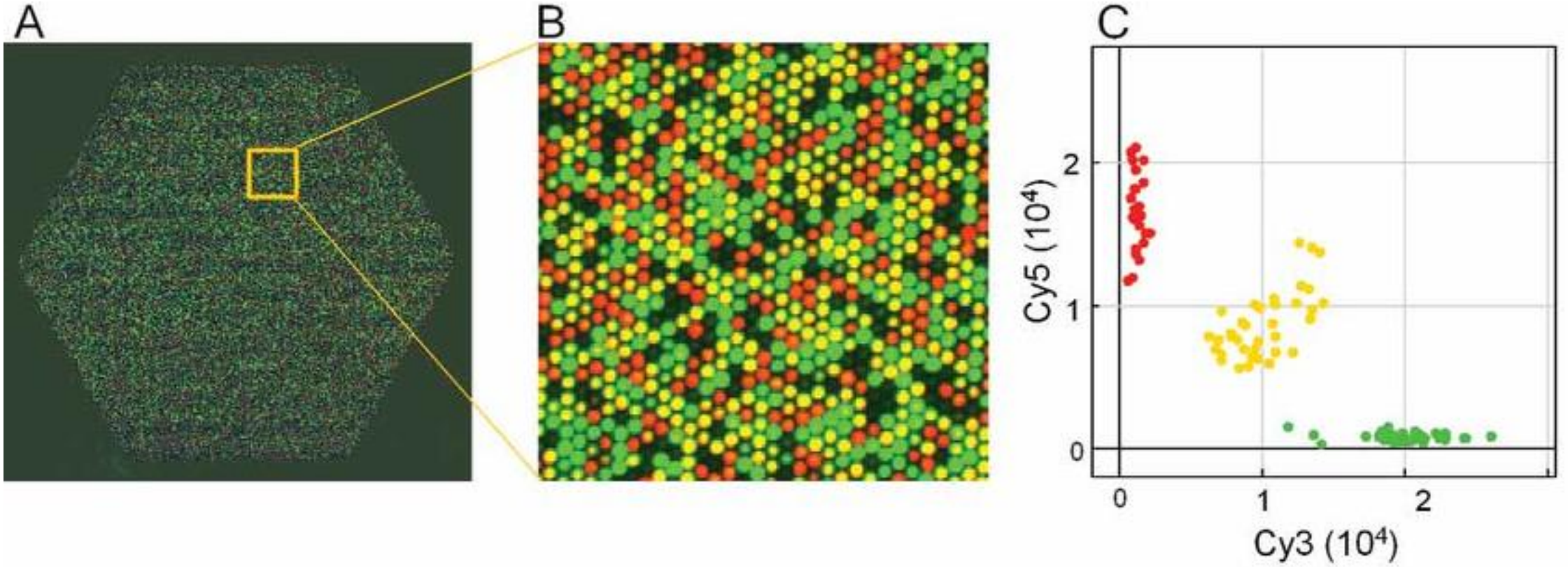


(Illumina)

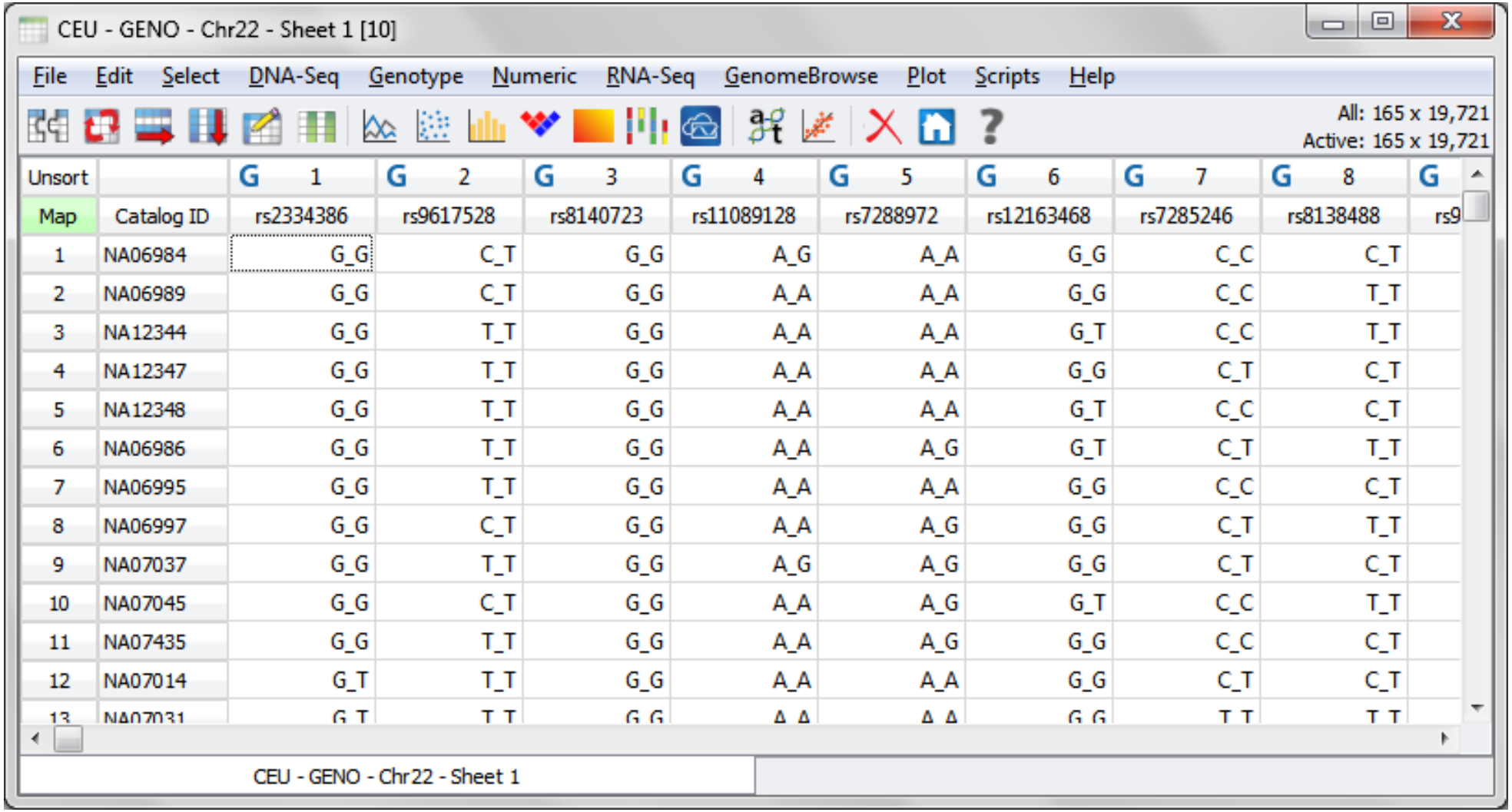

C
C
G
A
T
T
INTENSITY
INTENSITY
INTENSITY
rs100000010
rs1108985
rs1109052

(Illumina)

Chip-Based Genotyping



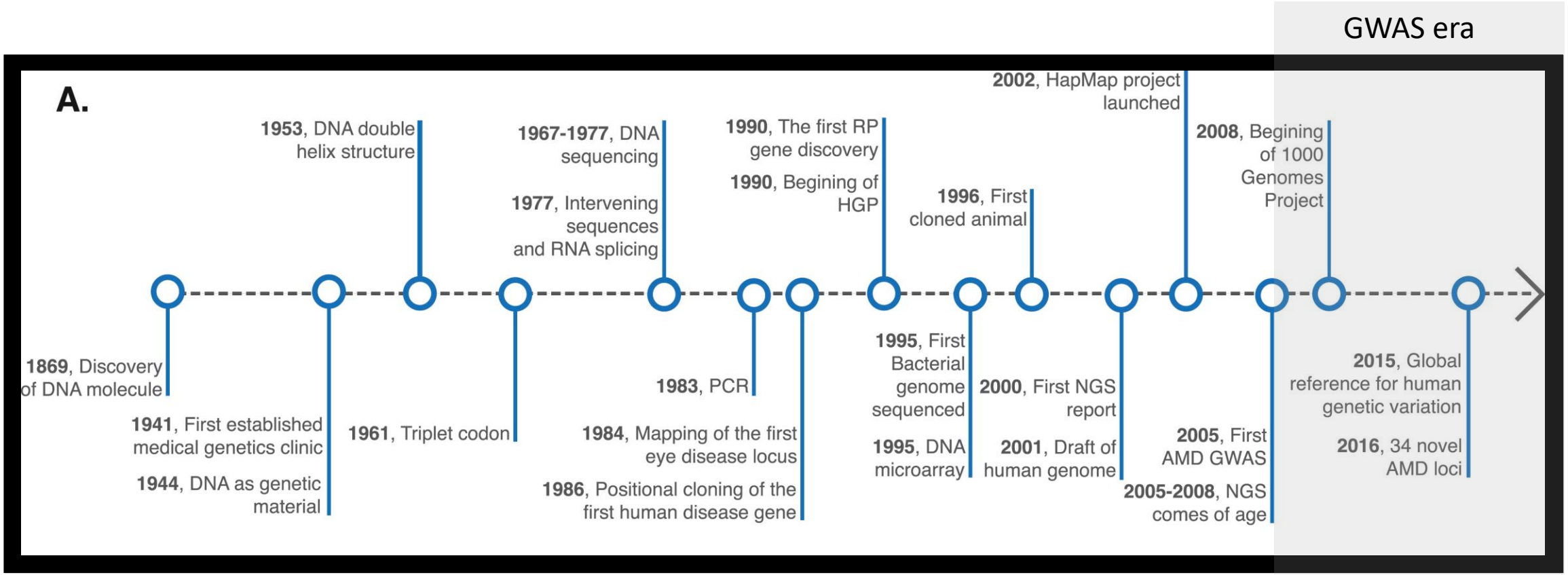
(Fan et al., 2003)


CEU - GENO - Chr22 - Sheet 1 [10]
File Edit Select DNA-Seq Genotype Numeric RNA-Seq GenomeBrowse Plot Scripts Help
All: 165 x 19,721
Active: 165 x 19,721
Unsort
Map
Catalog ID
G 1
G 2
G 3
G 4
G 5
G 6
G 7
G 8
G
rs2334386
rs9617528
rs8140723
rs11089128
rs7288972
rs12163468
rs7285246
rs8138488
rs9
1 NA06984 G_G C_T G_G A_G A_A G_G C_C C_T
2 NA06989 G_G C_T G_G A_A A_A G_G C_C T_T
3 NA12344 G_G T_T G_G A_A A_A G_T C_C T_T
4 NA12347 G_G T_T G_G A_A A_A G_G C_T C_T
5 NA12348 G_G T_T G_G A_A A_A G_T C_C C_T
6 NA06986 G_G T_T G_G A_A A_G G_T C_T T_T
7 NA06995 G_G T_T G_G A_A A_A G_G C_C C_T
8 NA06997 G_G C_T G_G A_A A_G G_G C_T T_T
9 NA07037 G_G T_T G_G A_G A_G G_G C_T C_T
10 NA07045 G_G C_T G_G A_A A_G G_T C_C T_T
11 NA07435 G_G T_T G_G A_A A_G G_G C_C C_T
12 NA07014 G_T T_T G_G A_A A_A G_G C_T C_T
CEU - GENO - Chr22 - Sheet 1

Genome-wide Association Study (GWAS)

GWAS: a Timeline

camh
Krembil Centre for Neuroinformatics



(Chaitankar et al., 2016)

Basic Types of GWAS

- Single-stage:
  - Discovery GWAS; one sample, one set of tests
- Multi-stage:
  - Discovery phase
    - usually largest sample, full agnostic genome-wide scan
  - Replication phase(s)
    - usually smaller, non-overlapping samples following up on significant and suggestive loci only (to reduce multiple comparisons)
  - Meta-analysis
    - Results from discovery phase and replication phase(s) are pooled and analyzed together. This is distinct from “mega-analyses”, where individual-level genotype and phenotype data are first pooled then analyzed.

Anatomy of GWAS

camh
Krembil Centre for Neuroinformatics

Genotype arrays:

- Affymetrix
- Illumina

400K – 2.5M SNPs

Subjects:

- sex discordance
- missingness and heterozygosity
- duplication
- cryptic relatedness
- ancestry

Variants:

- missingness
- Hardy-Wienberg Equilibrium
- minor allele frequency

Strand alignment to reference:

- HRC v1.1
- 1000G phase 3

Phasing

Imputation

Imputation quality score

Variant QC as in QC #1

Dosage or best guess?

Linear or Logistic

Need subject phenotypes

Technical and biological covariates

PLINK 1.90b

p-value correction

- $p<5.0x10^{-8}$
- sFDR

LD-based SNP clumping

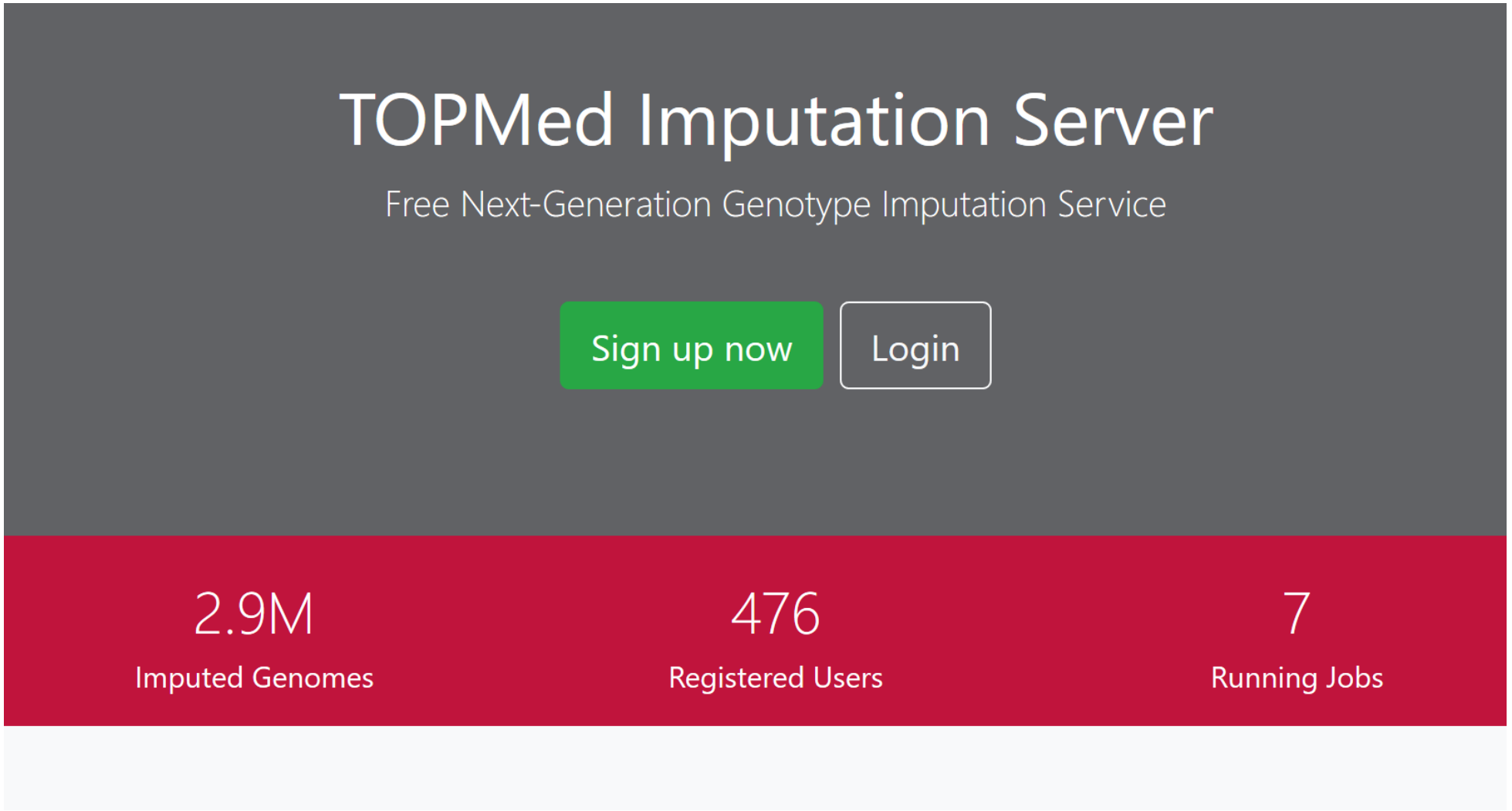
Q-Q plot

Manhattan plot

LocusZoom (regional) plot

(Recommend: *Anderson et al., 2010, Nature Protocols*)

Don’t panic…



https://imputation.biodatacatalyst.nhlbi.nih.gov/

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X_1$$

- Remember that a GWAS is just a collection of millions of individual regression tests, performed independently of each other.

- So what exactly *IS* regression and, more importantly, what do the numbers that we get out of a GWAS really mean?

Simple Linear Regression

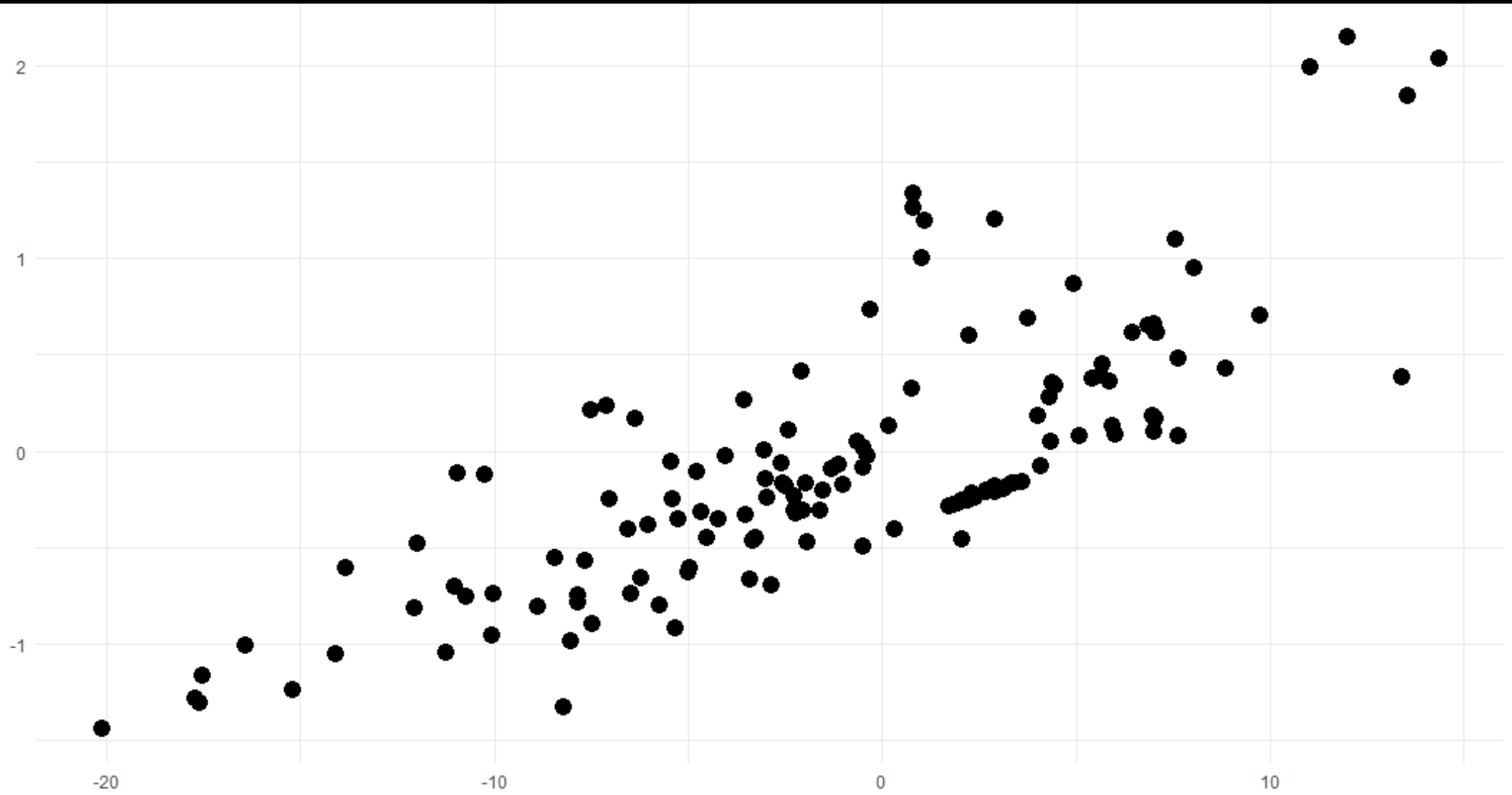
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

- Two continuous variables:
  - outcome (dependent) Y, and exposure (independent) X
  - Think of it as "Y is dependent on X"
  - R notation: Y ~ X

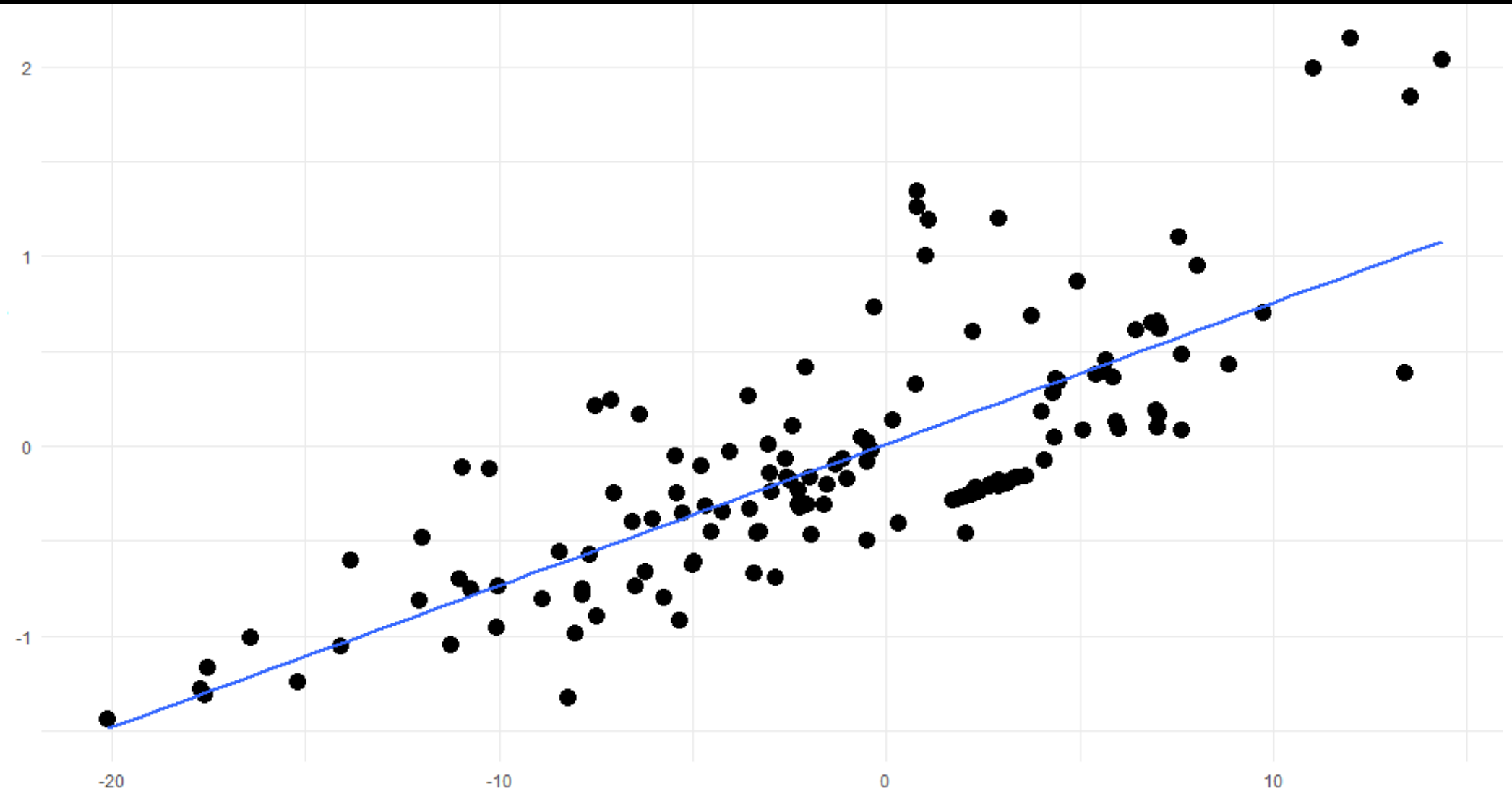
- The question is: ***how much*** is Y dependent on X, and, based on the variability in our data, can we say that the association is ***statistically significant***?
- Remember that statistical significance depends on the probability of observing our association in a random sample of a given size (n) if our two variables are actually not related to each other in the population.

Simple Linear Regression

camh
Krembil Centre for Neuroinformatics

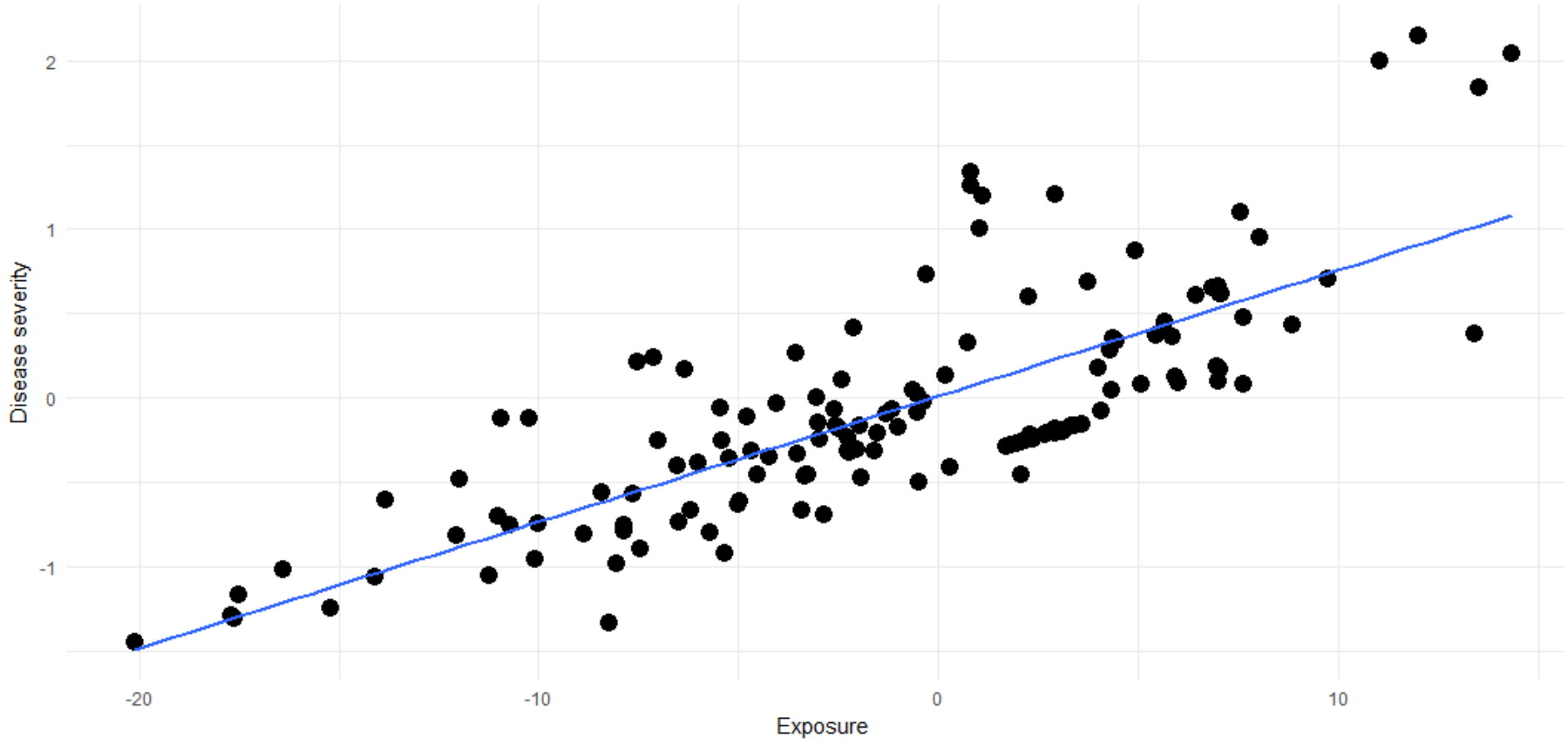


Simple Linear Regression



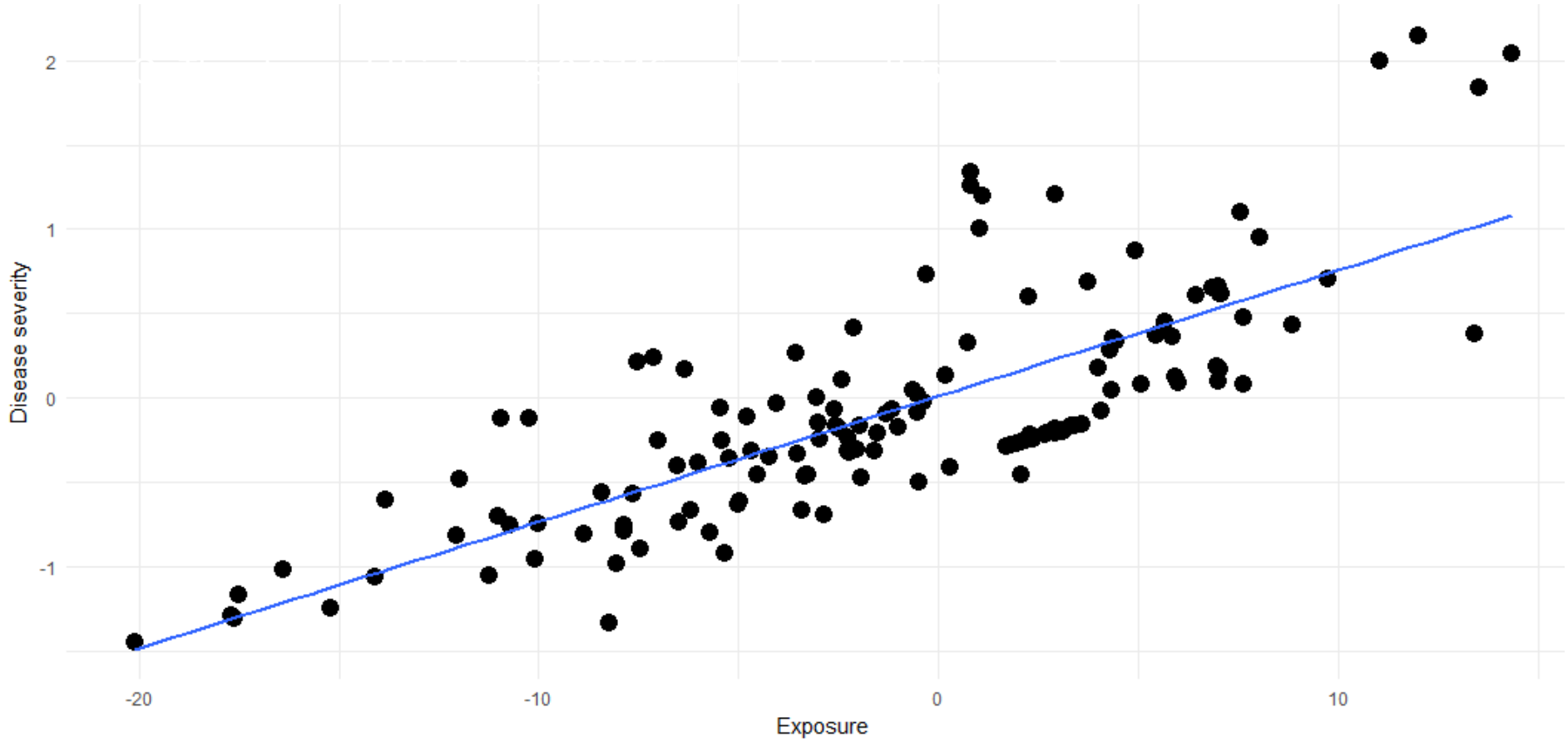
Simple Linear Regression

$$\hat{Y} = \beta_0 + \beta_1 X_1$$



Simple Linear Regression

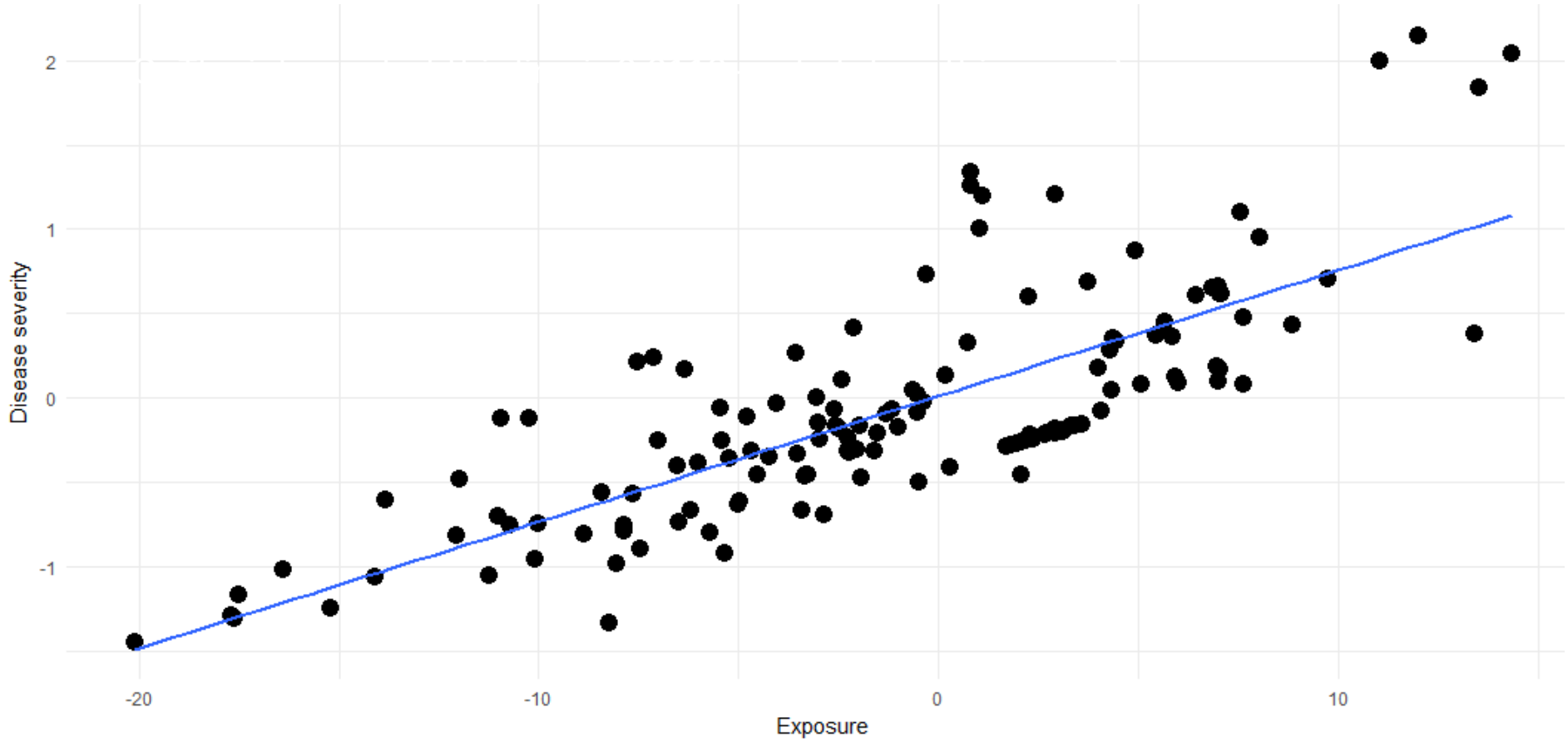
$$\hat{Y} = \beta_0 + \beta_1 X_1$$



A: For every change of +1 in our exposure (X) variable, there is a change of +0.0746 in our outcome (Y) variable. In this case, Y is “Disease Severity”

Simple Linear Regression

$$\hat{Y} = \beta_0 + \beta_1 X_1$$



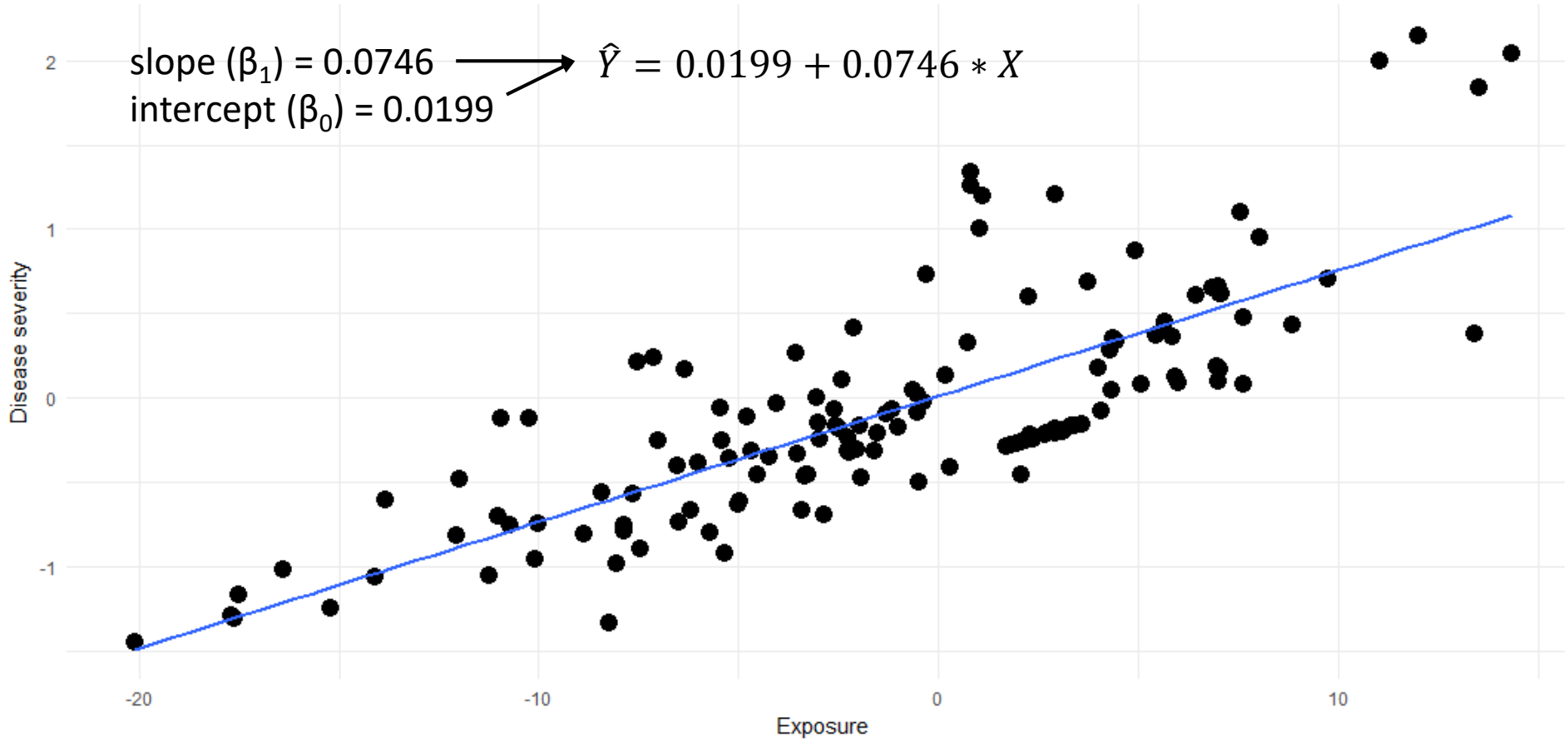
A: When X is equal to zero, Y is equal to 0.0199

Simple Linear Regression

$$\hat{Y} = \beta_0 + \beta_1 X_1$$

slope (β_1) = 0.0746 → $\hat{Y} = 0.0199 + 0.0746 * X$

intercept (β_0) = 0.0199



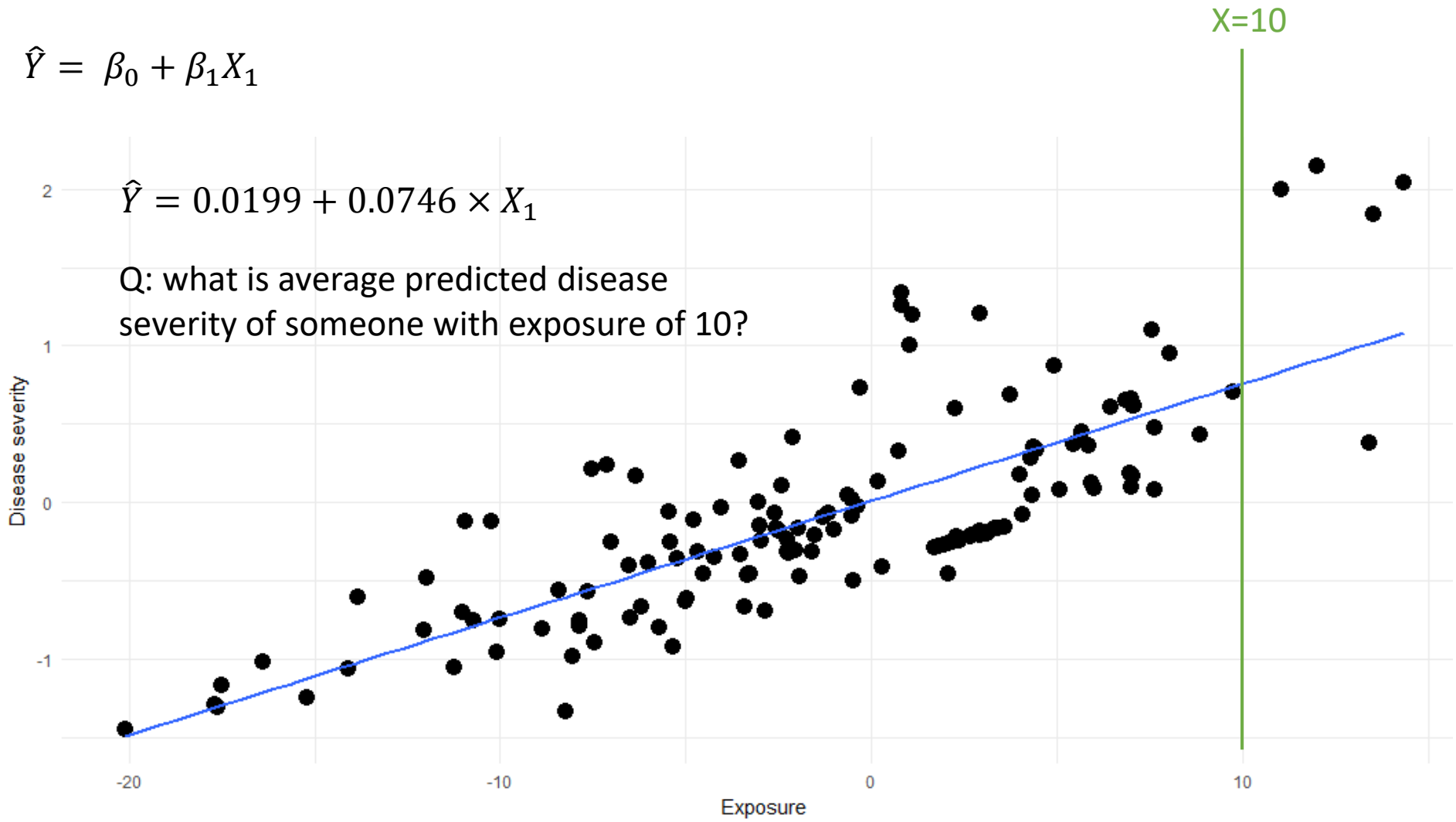
Simple Linear Regression

camh
Krembil Centre for Neuroinformatics

$$\hat{Y} = \beta_0 + \beta_1 X_1$$

$$\hat{Y} = 0.0199 + 0.0746 \times X_1$$

Q: what is average predicted disease severity of someone with exposure of 10?



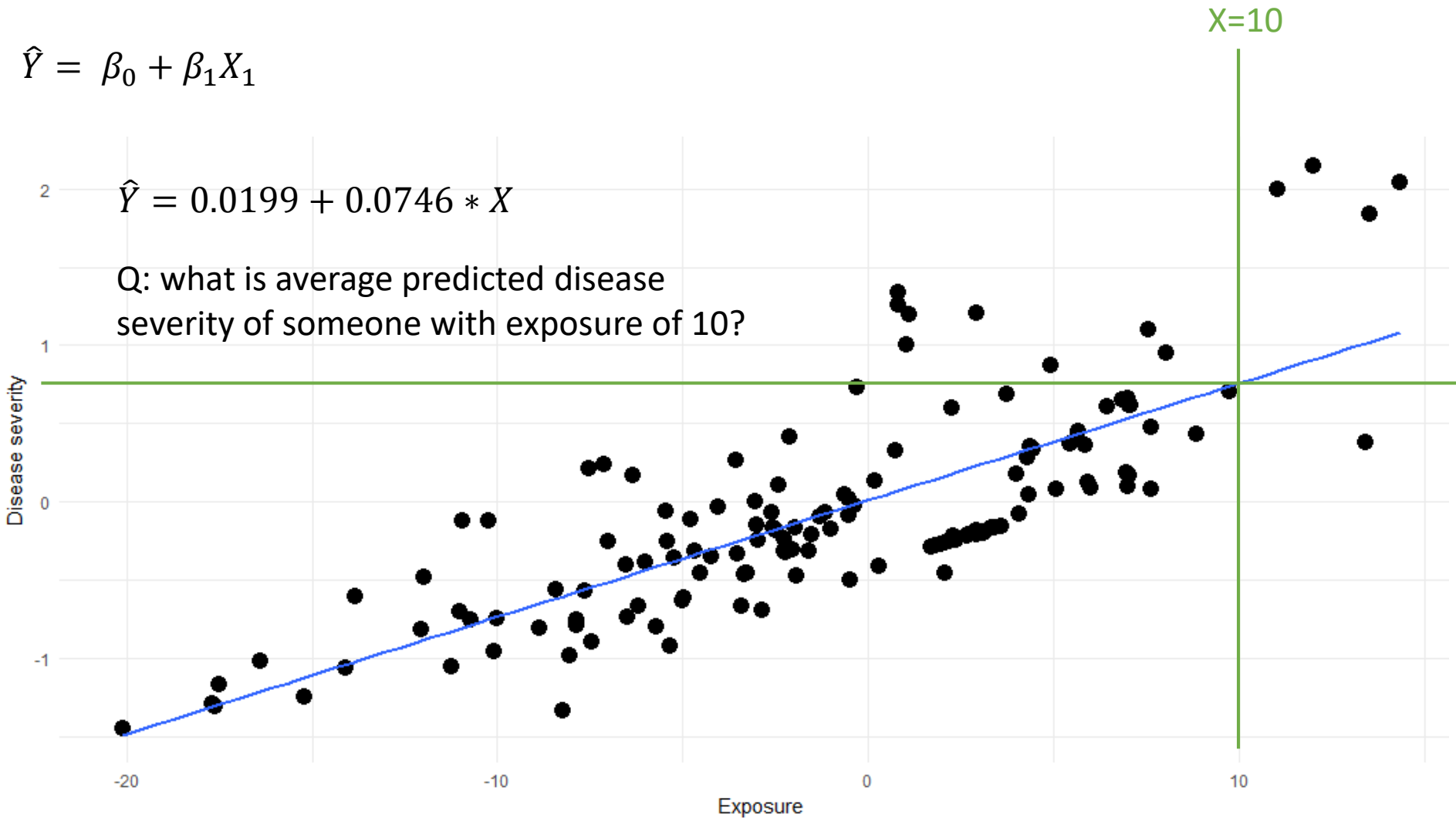
Simple Linear Regression

camh
Krembil Centre for Neuroinformatics

$$\hat{Y} = \beta_0 + \beta_1 X_1$$

$$\hat{Y} = 0.0199 + 0.0746 * X$$

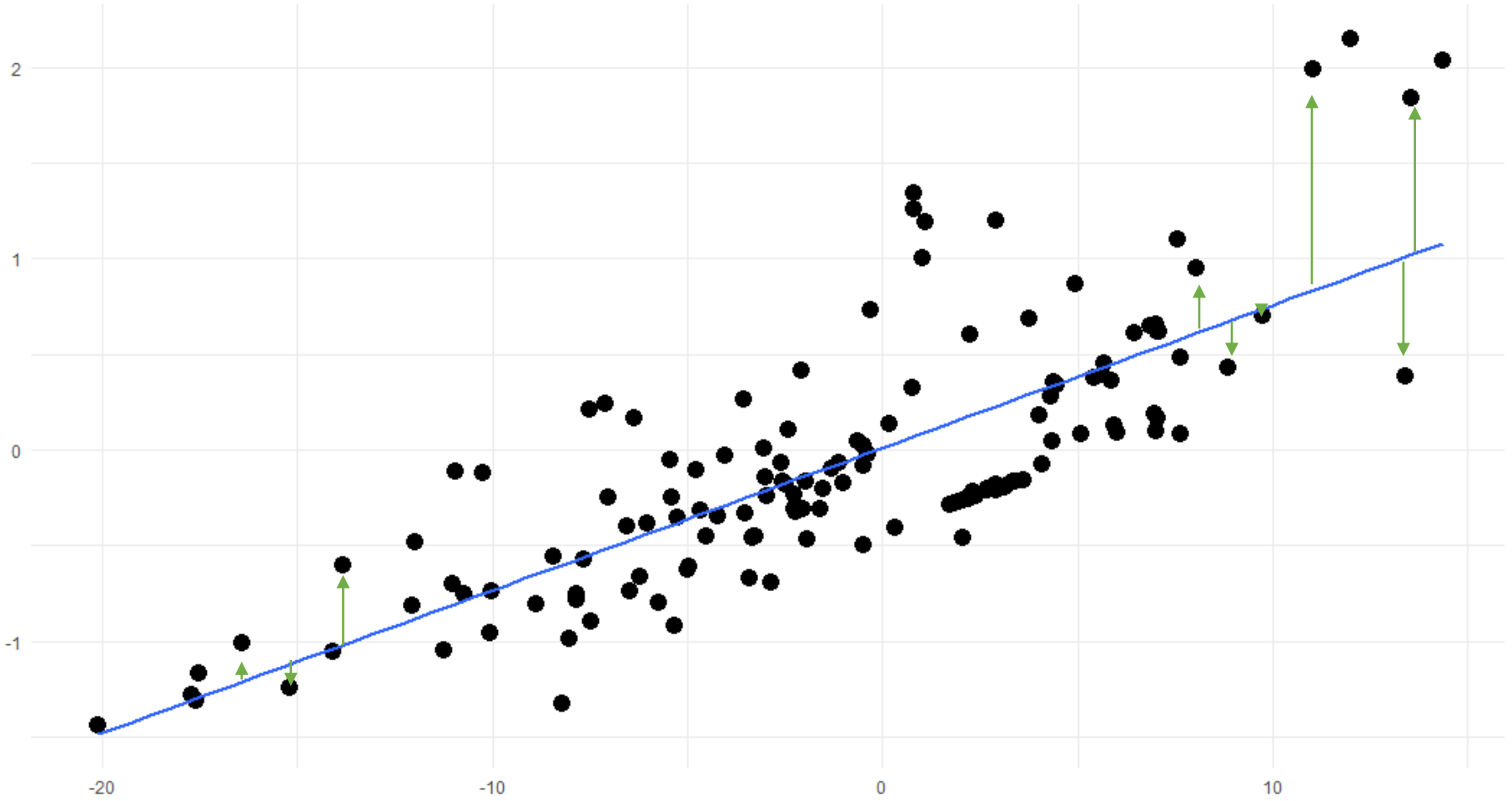
Q: what is average predicted disease severity of someone with exposure of 10?



A:

$$\hat{Y} = 0.0199 + 0.0746 * X$$
$$\hat{Y} = 0.0199 + 0.0746 * 10$$
$$\hat{Y} = 0.0199 + 0.746$$
$$\hat{Y} = 0.766$$

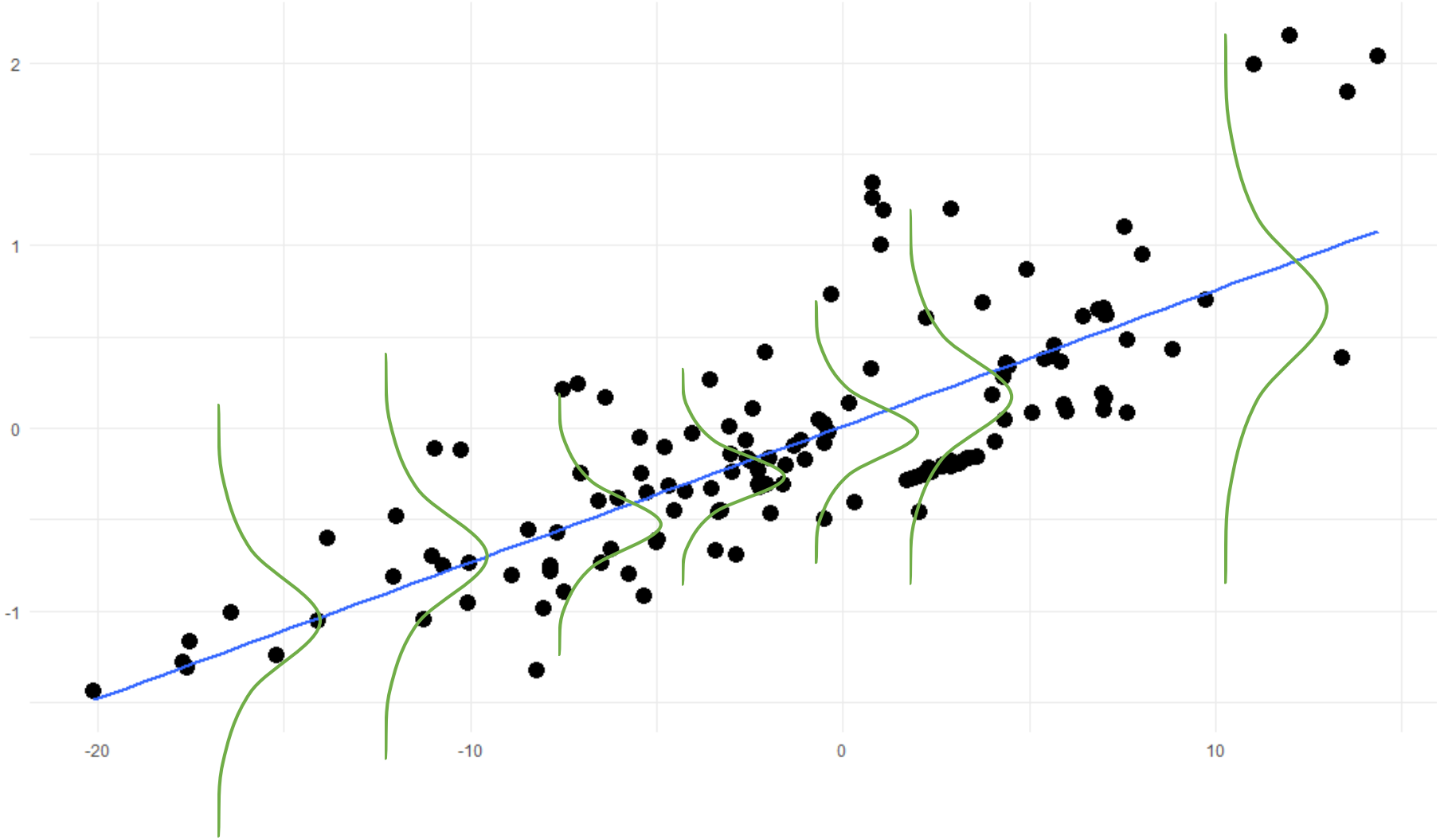
Variability and Confidence



At each point along X, there is some Y distribution around the line of best fit, which linear regression assumes to be *normal.*

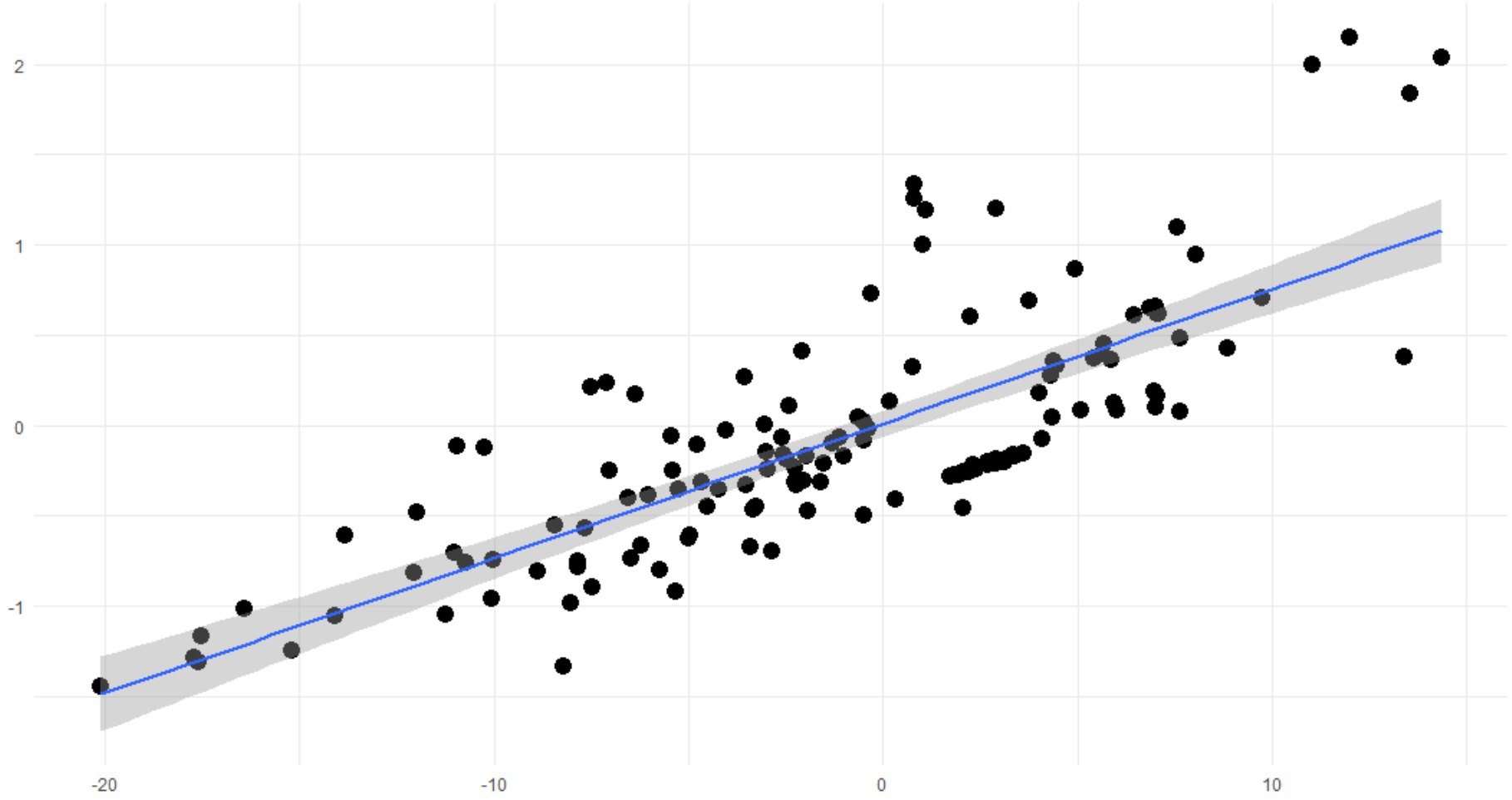
The distances between the observed points and the regression line are called “residuals”

Variability and Confidence



At each point along X, there is some Y distribution around the line of best fit, which linear regression assumes to be *normal*

Variability and Confidence

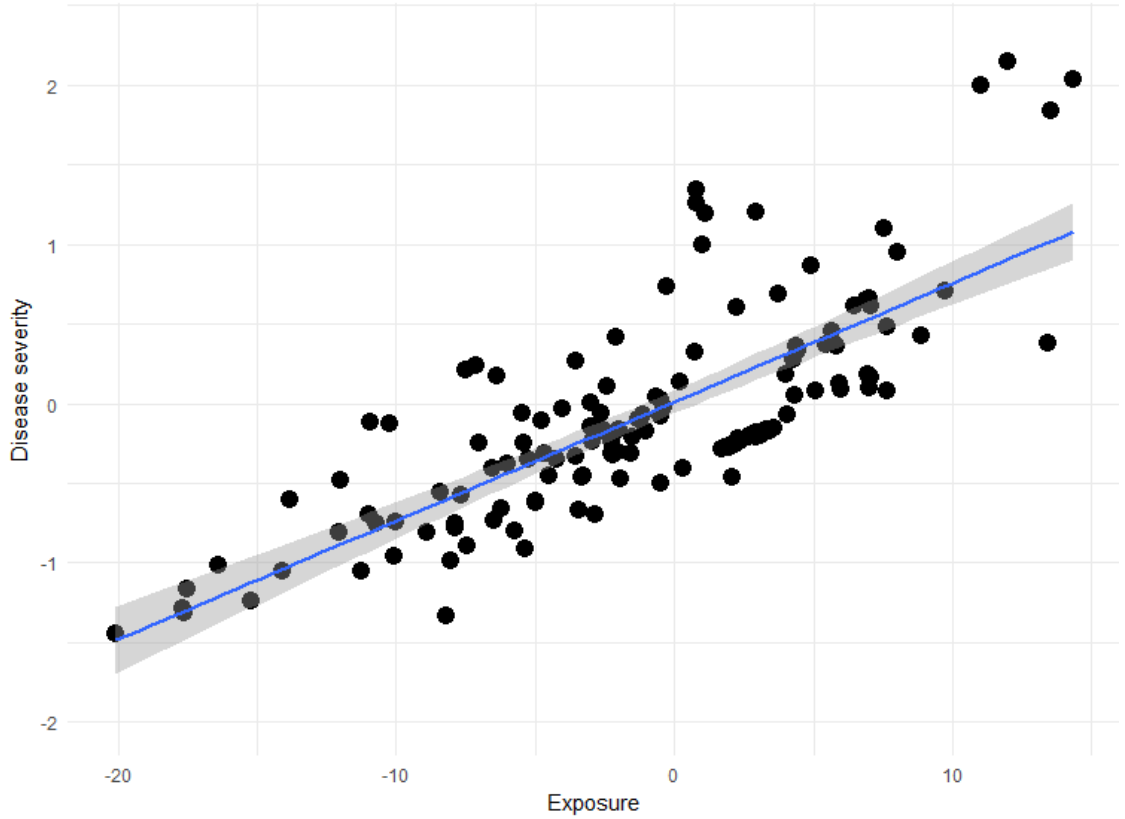


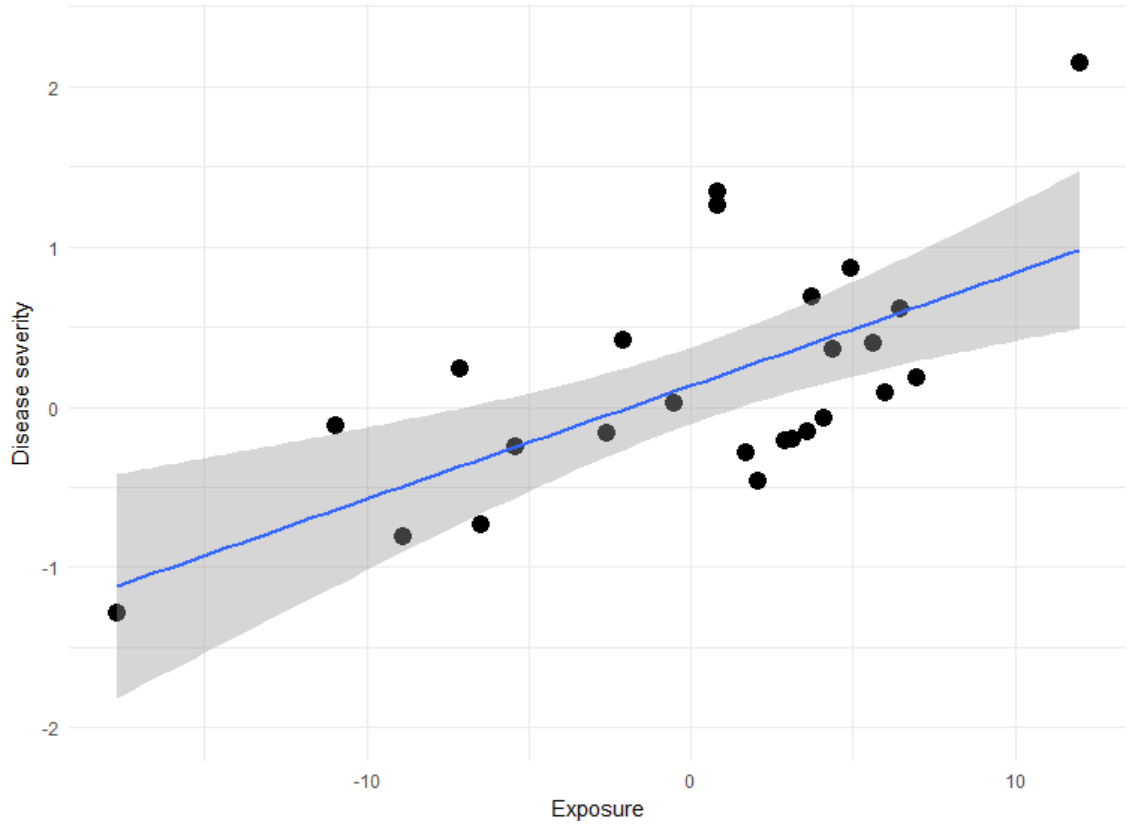
Since we have this margin of error around our regression line, we can draw confidence intervals around it.

So our slope and intercept (beta coefficients) each have a standard error associated with them.

Variability and Confidence

More data (n) and less variance (s^2) = ⬆ confidence (CI), ⬇ standard errors (SE)





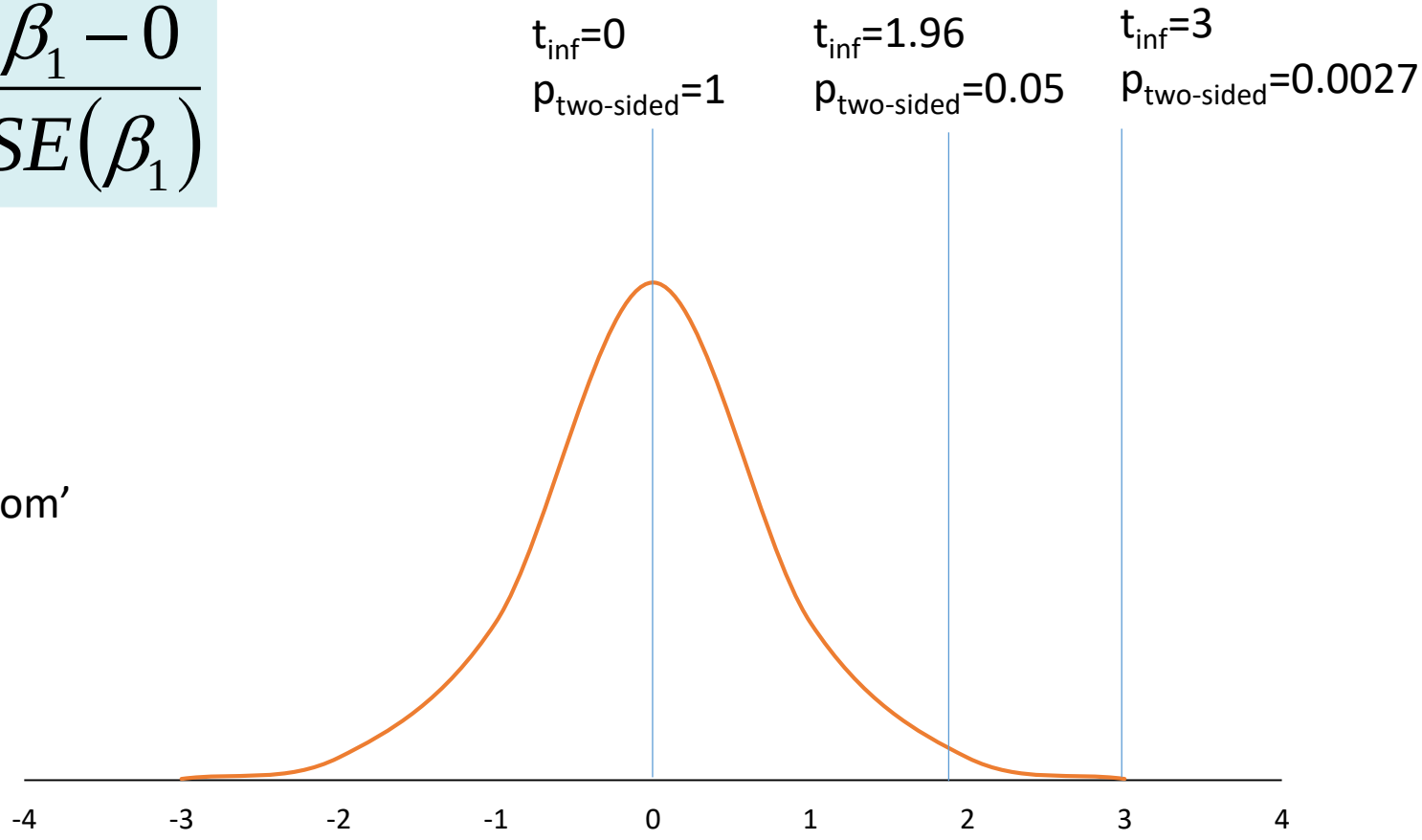
Putting it Together: Hypothesis Testing

Null: $\beta_1 = 0$
Alt: $\beta_1 \neq 0$

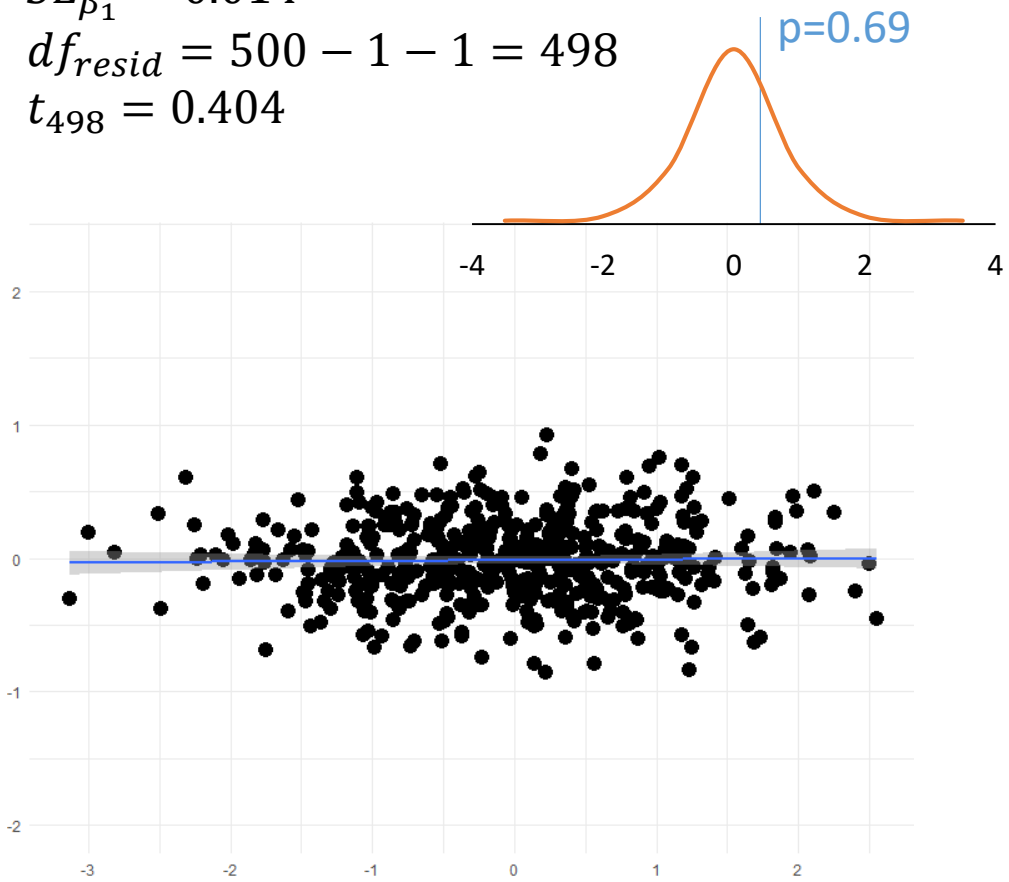
$$t = \frac{\beta_1 - 0}{SE(\beta_1)}$$

t distribution with some
'residual degrees of freedom'

$$df = n - k - 1$$

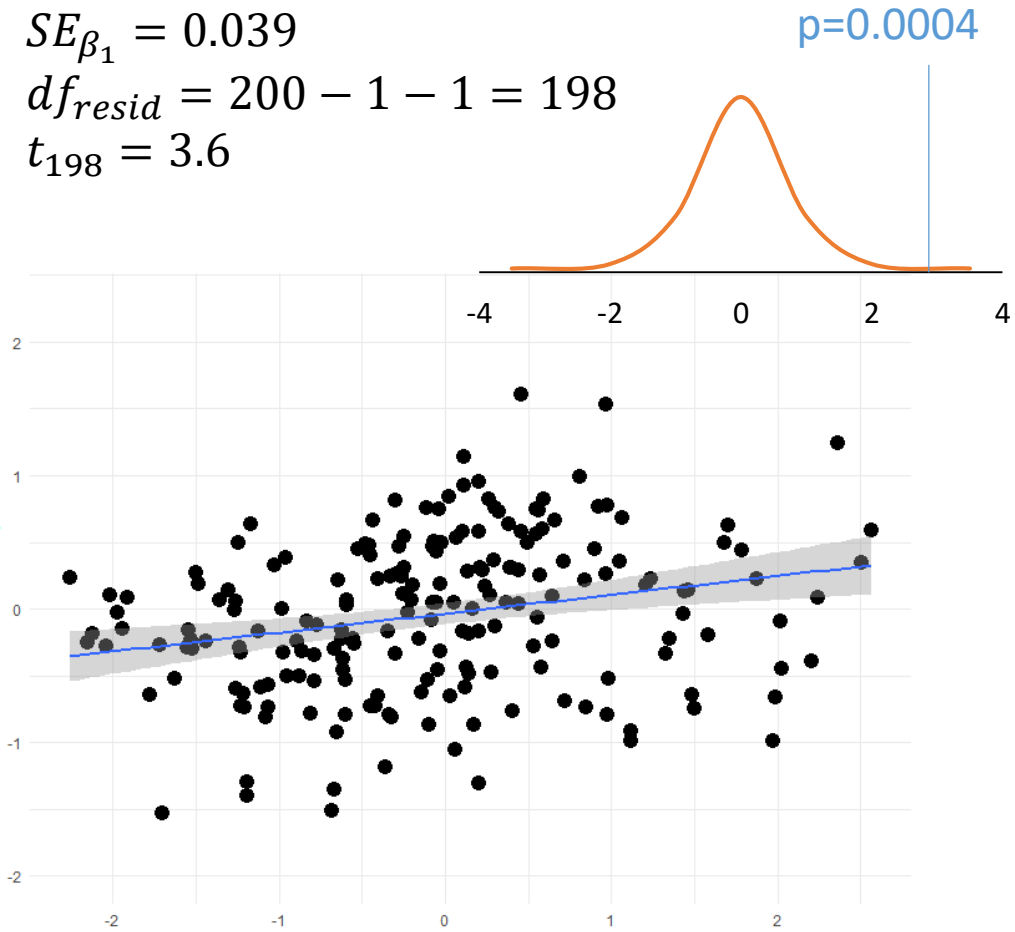


$\hat{Y} = -0.015 + 0.0056 \times X_1$
$n = 500$
$SE_{\beta_1} = 0.014$
$df_{resid} = 500 - 1 - 1 = 498$
$t_{498} = 0.404$



We **do not** reject the null that β_1=0 since p>0.05, and therefore find no relationship between Y and X_1

$\hat{Y} = -0.035 + 0.142 \times X_1$
$n = 200$
$SE_{\beta_1} = 0.039$
$df_{resid} = 200 - 1 - 1 = 198$
$t_{198} = 3.6$

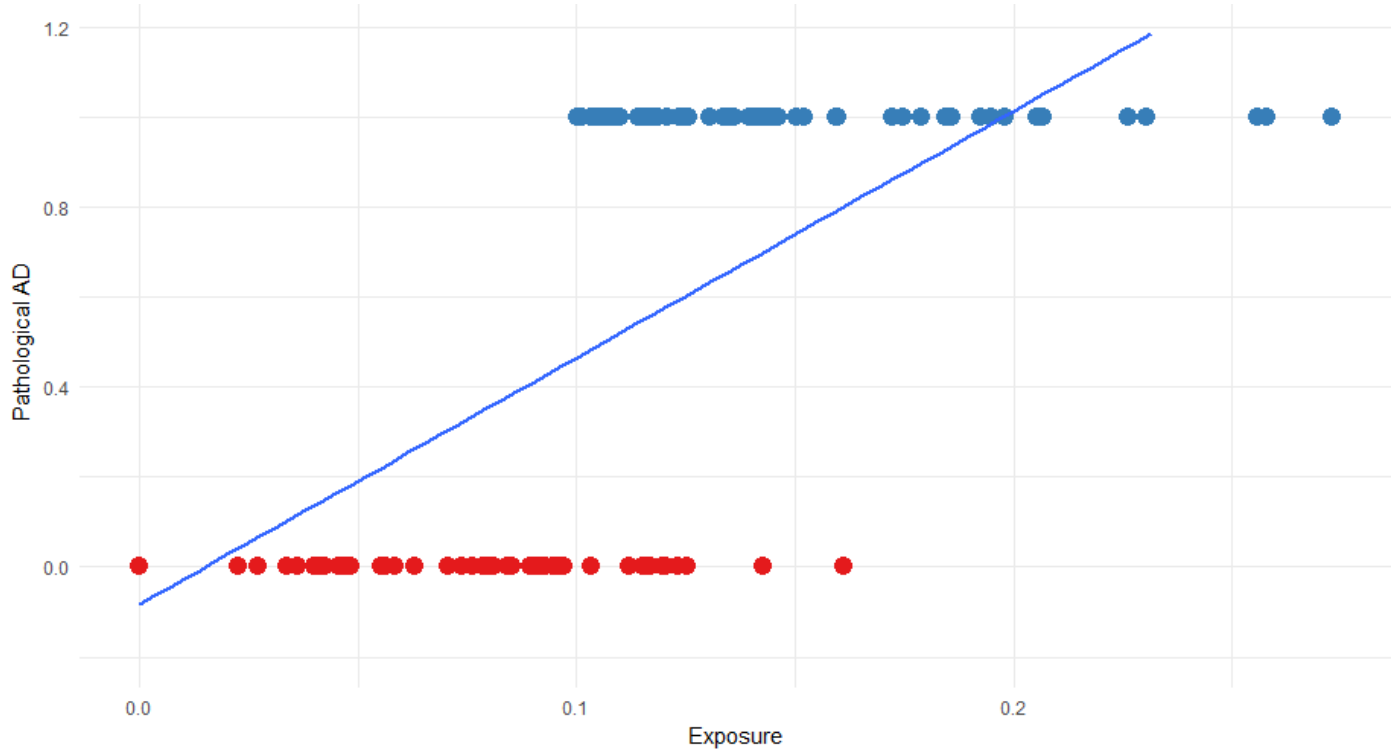


We **do** reject the null that β_1=0 since p<0.05, and find that Y is related to X_1 such that for every 1 unit increase in X_1 there is on average a 0.142 increase in Y

Logistic Regression

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1$$

- Q: How do you model a binary variable with a linear function?

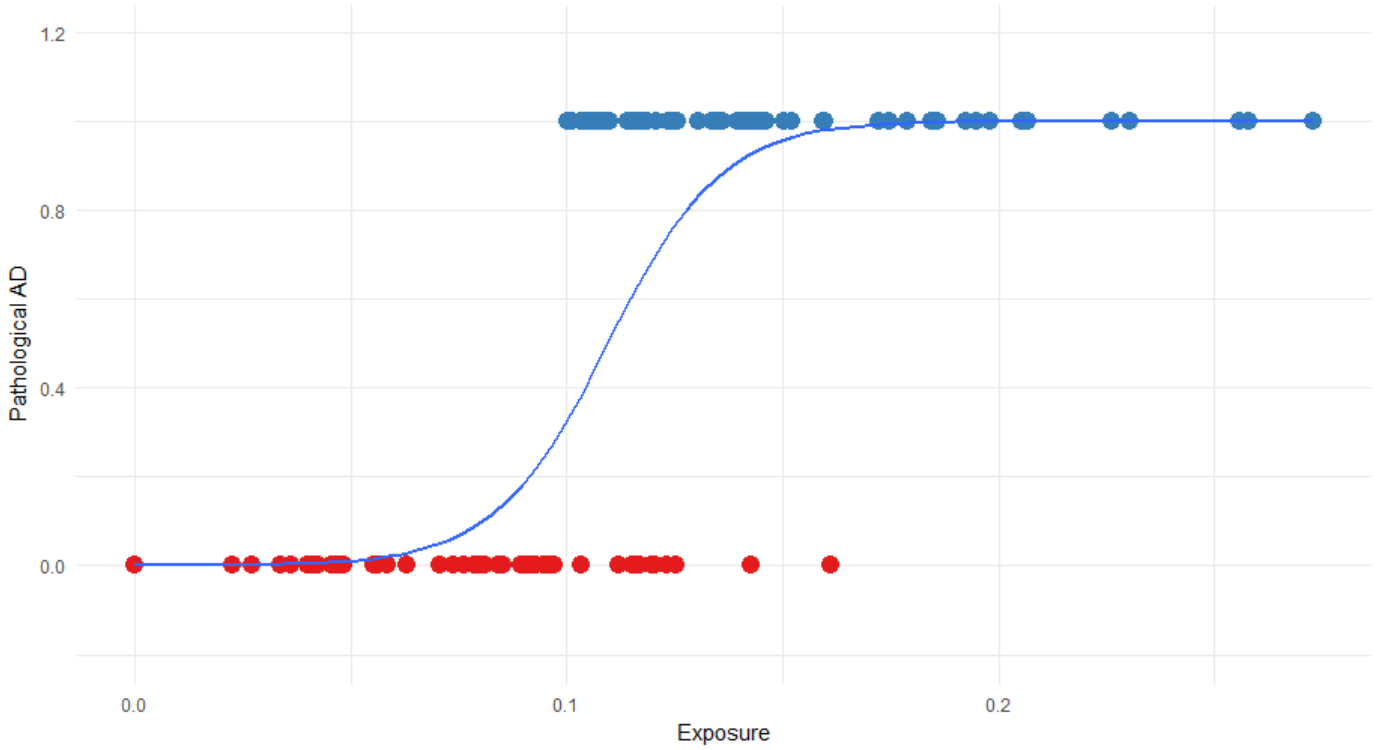


Remember:
Binary variables are considered on a scale of probability [0-1]

Logistic Regression

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1$$

- A: Transformation of the Y variable – in this case with a "logit"



This transformation crucially keeps the estimated values of Y between 0 and 1.

The effect coefficients are interpreted differently since:

A) Scale is no longer linear
B) The outcome is a probability

Logistic Regression

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1$$

- We convert beta coefficients from a logistic model into odds ratios by applying the exponential function (e is a constant, the base of the natural logarithm ~ 2.718):

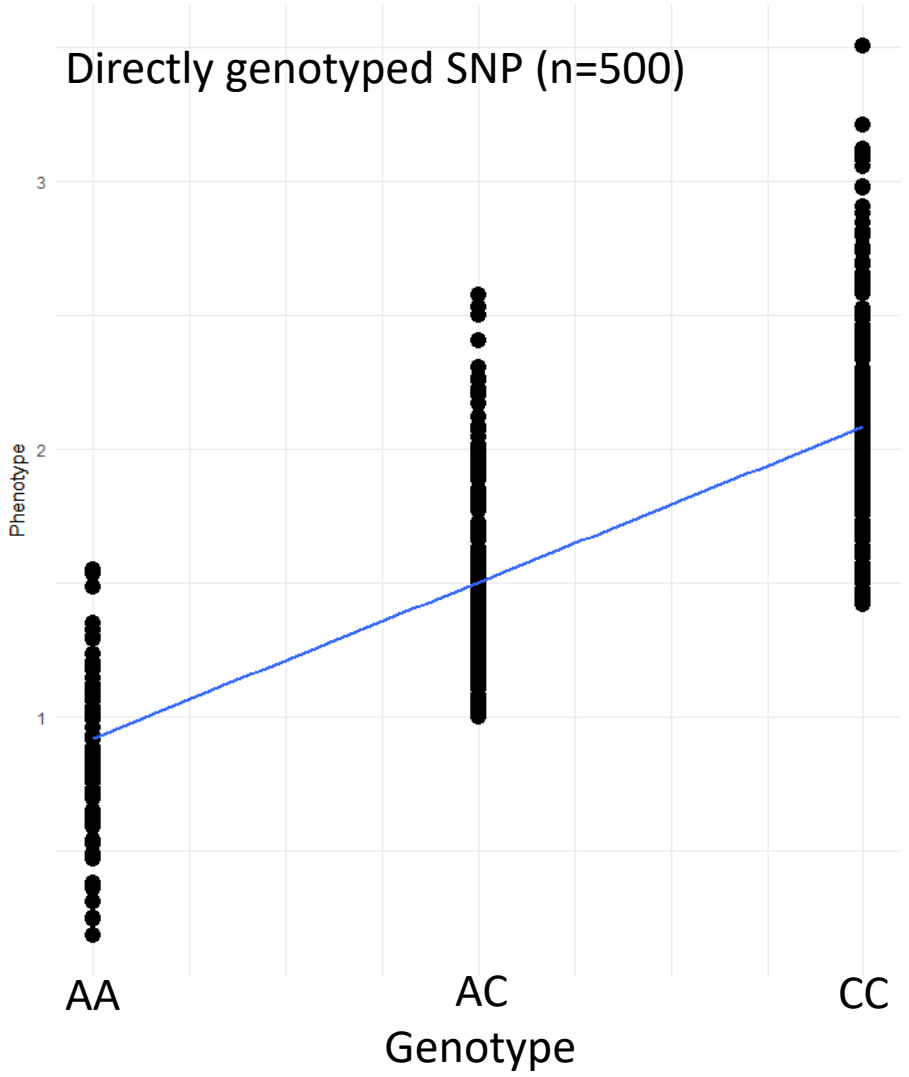
$$e^{\beta} = OR$$

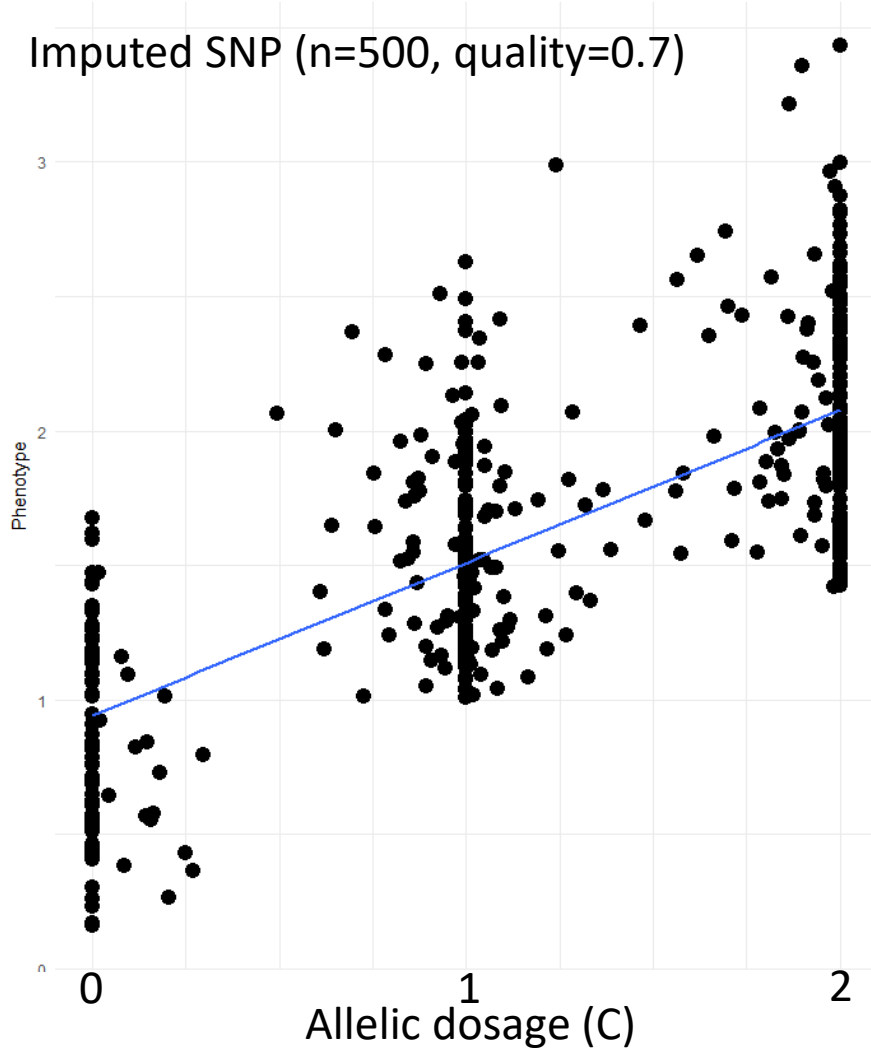
For example, if beta = 2.54, then the OR for that effect = $e^{2.54}$=12.68

Regression for SNPS?

- Genotype is a discrete variable in diploid organisms (humans)
  - e.g. SNP^{AC}: A/A, A/C, C/C
- However, the process of imputing unobserved SNPs is imperfect. Genotype probabilities are calculated given evidence provided in the chosen reference panel (posterior probabilities)
- Therefore we model genotype as a continuous variable (dosage), meaning the predicted 'dosage' of an allele at each SNP. The number ranges from 0-2.
- If the above SNP^{AC} was directly genotyped, it's C-allele dosages would be 0 (A/A), 1 (A/C), and 2 (C/C).

Regression for SNPS?





GWAS Key Considerations

camh
Krembil Centre for Neuroinformatics

- **Sample size**
  - Key influencer of statistical power in GWAS – that is, the ability of your study to detect genetic effects if they truly exist.

- **Minor allele frequency** (depends on sample size)
  - A SNP with **MAF=0.4** means that you would expect 160/1000 individuals to have the rare homozygous genotype
  - A SNP with **MAF=0.01** means that you would expect 0.1/1000 individuals to have the rare homozygous, and only ~20/1000 to have the heterozygous genotype.

- **Phenotype distribution**
  - For linear regression, normality of residuals is assumed
  - For logistic regression, you must have enough cases *and* controls (i.e. 1/0)

Summary of Key Statistics

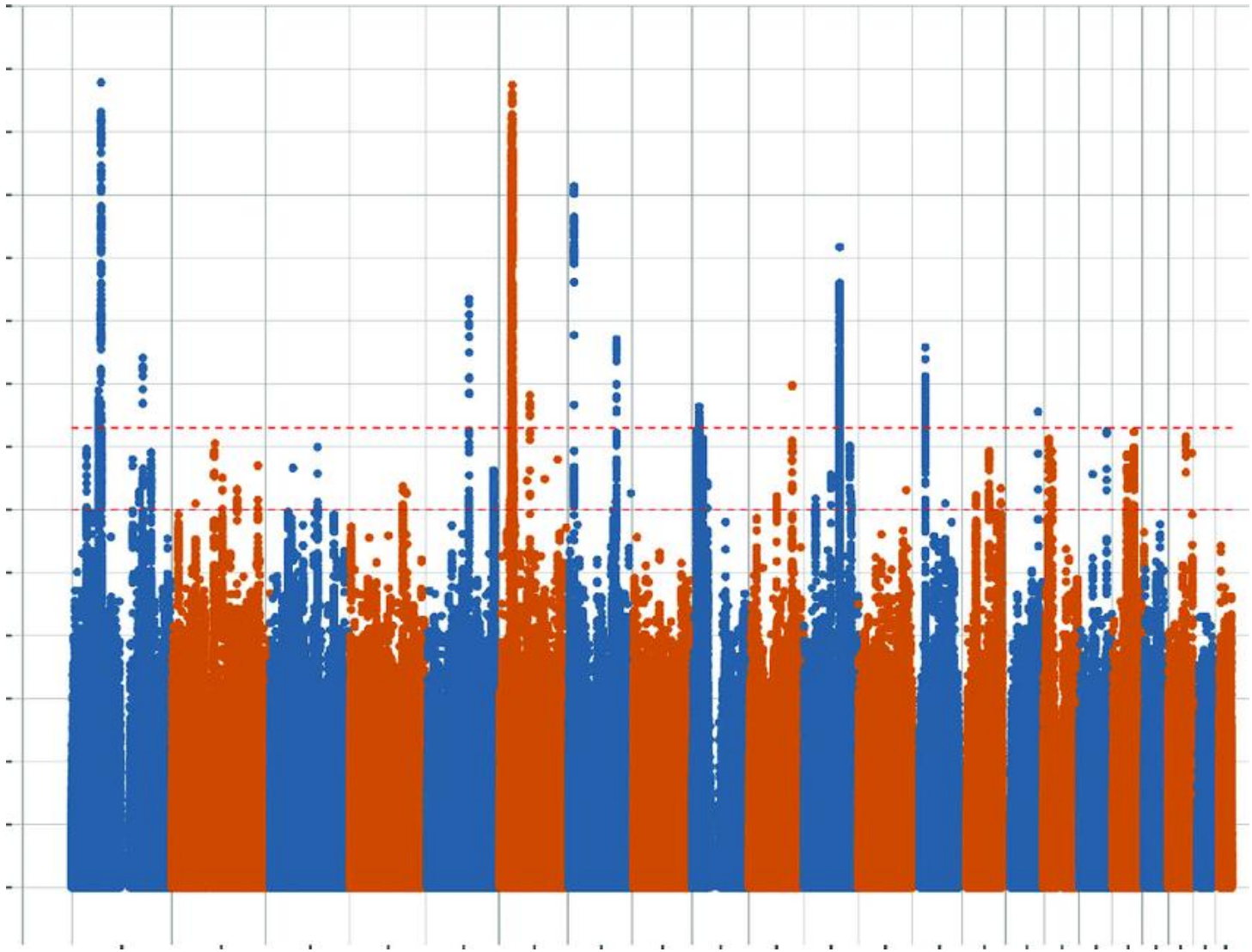
camh
Krembil Centre for Neuroinformatics

- **n** = sample size (important for degrees of freedom)
- **A1** = (usually) effect allele (vital for understanding direction of effect)
- **A2** = (usually) non-effect allele
- **Beta / OR** = effect coefficient (slope of line)
- **SE** = standard error of the effect coefficient (variability)
- **P-value** = probability of observing data if null is true (strength of evidence)

Linkage Disequilibrium

Linkage Disequilibrium

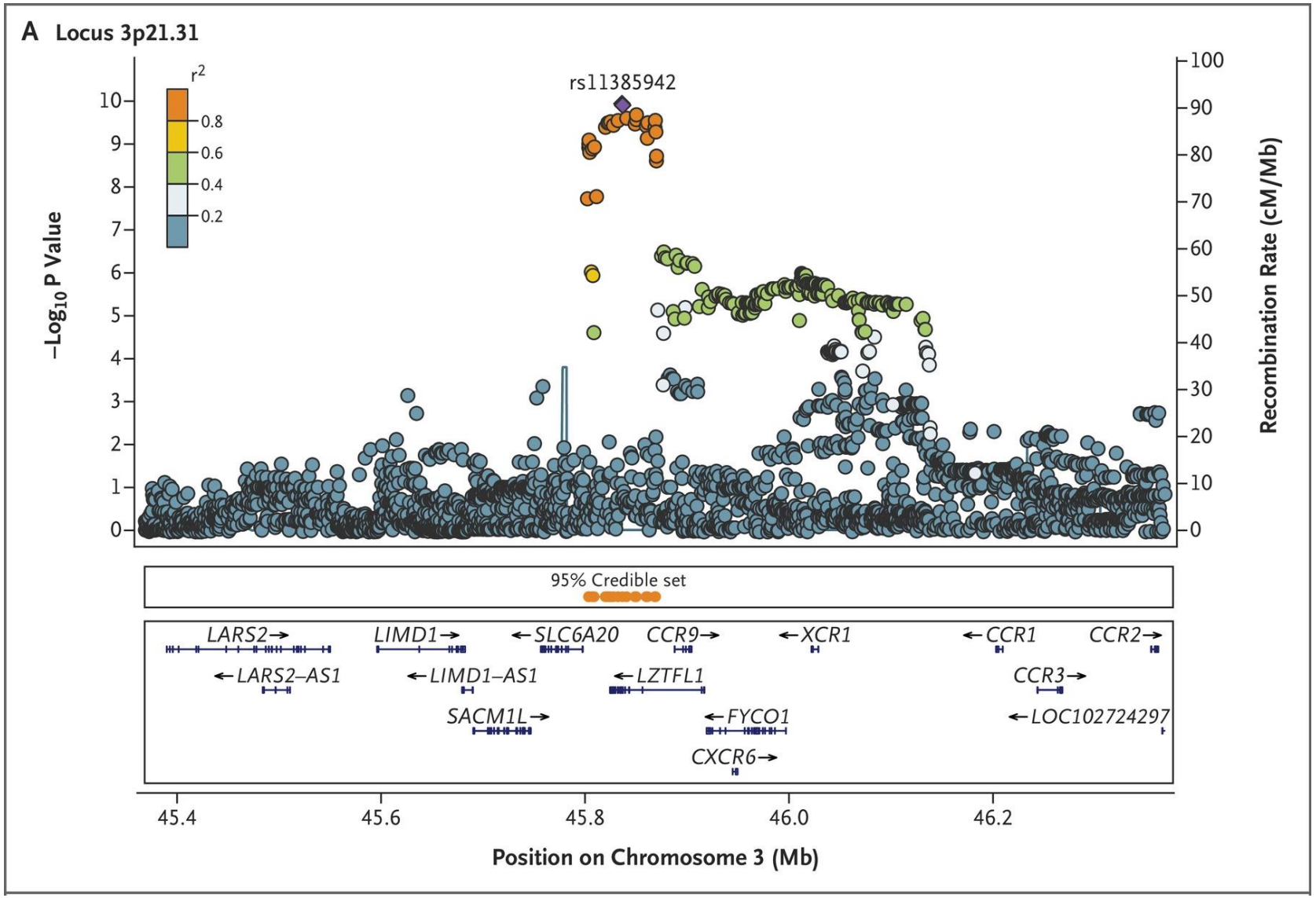
Notice peaks and valleys



Linkage Disequilibrium

camh
Krembil Centre for Neuroinformatics

Zoom in...



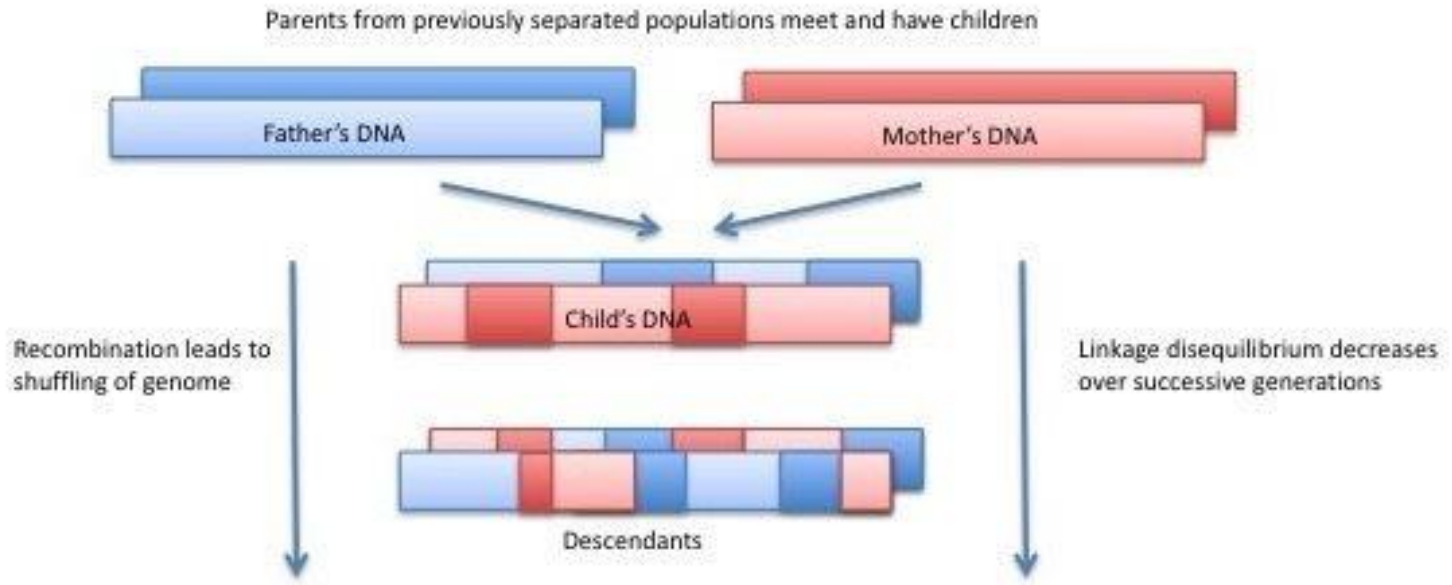
(Ellinghaus et al., 2020, NEJM)

Linkage Disequilibrium

SNPs are passed on non-randomly in blocks known as haplotype blocks.

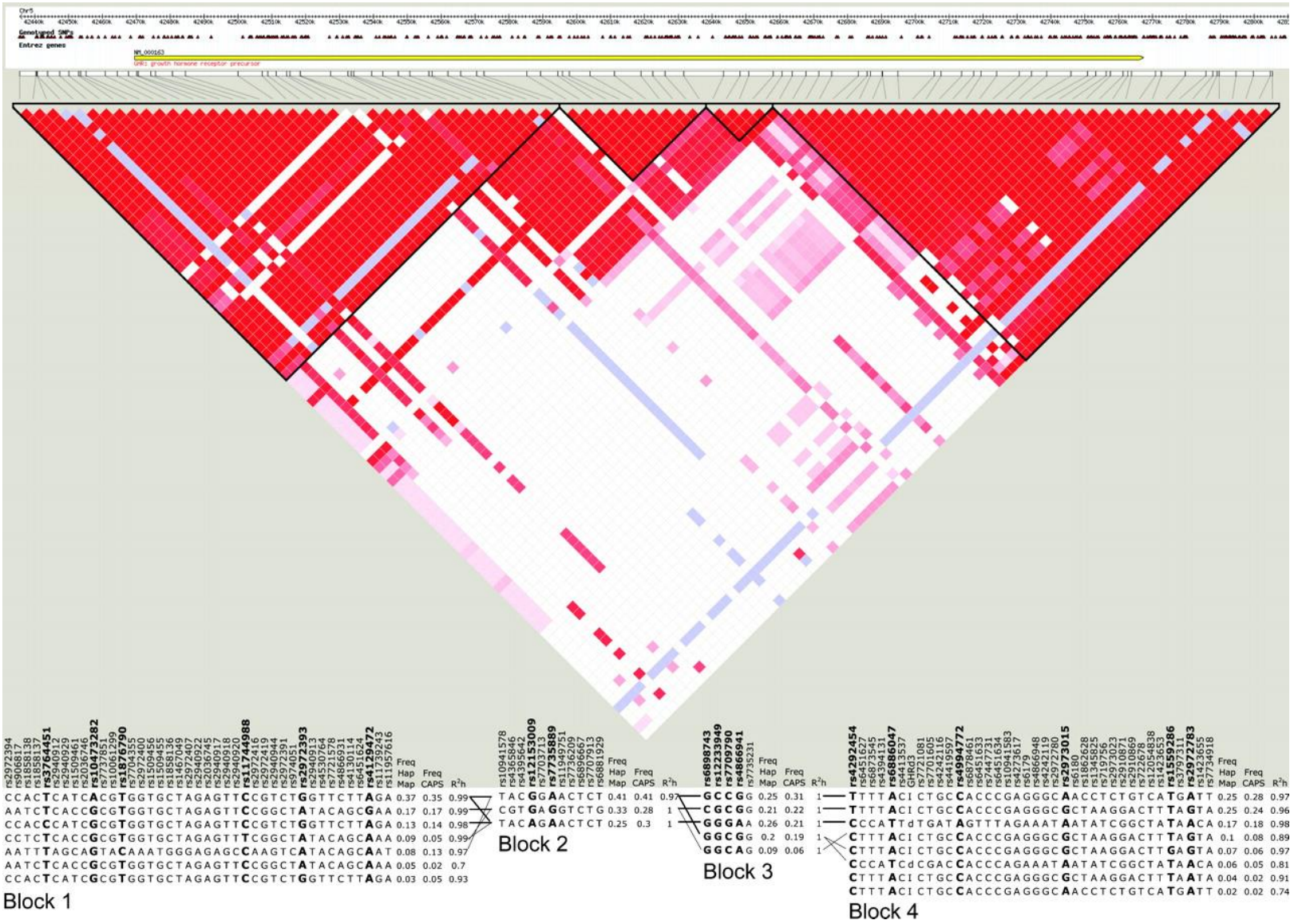
We can estimate the amount of disequilibrium (the non-randomness) by evaluating correlations of alleles in large populations.

Causal vs. in linkage?

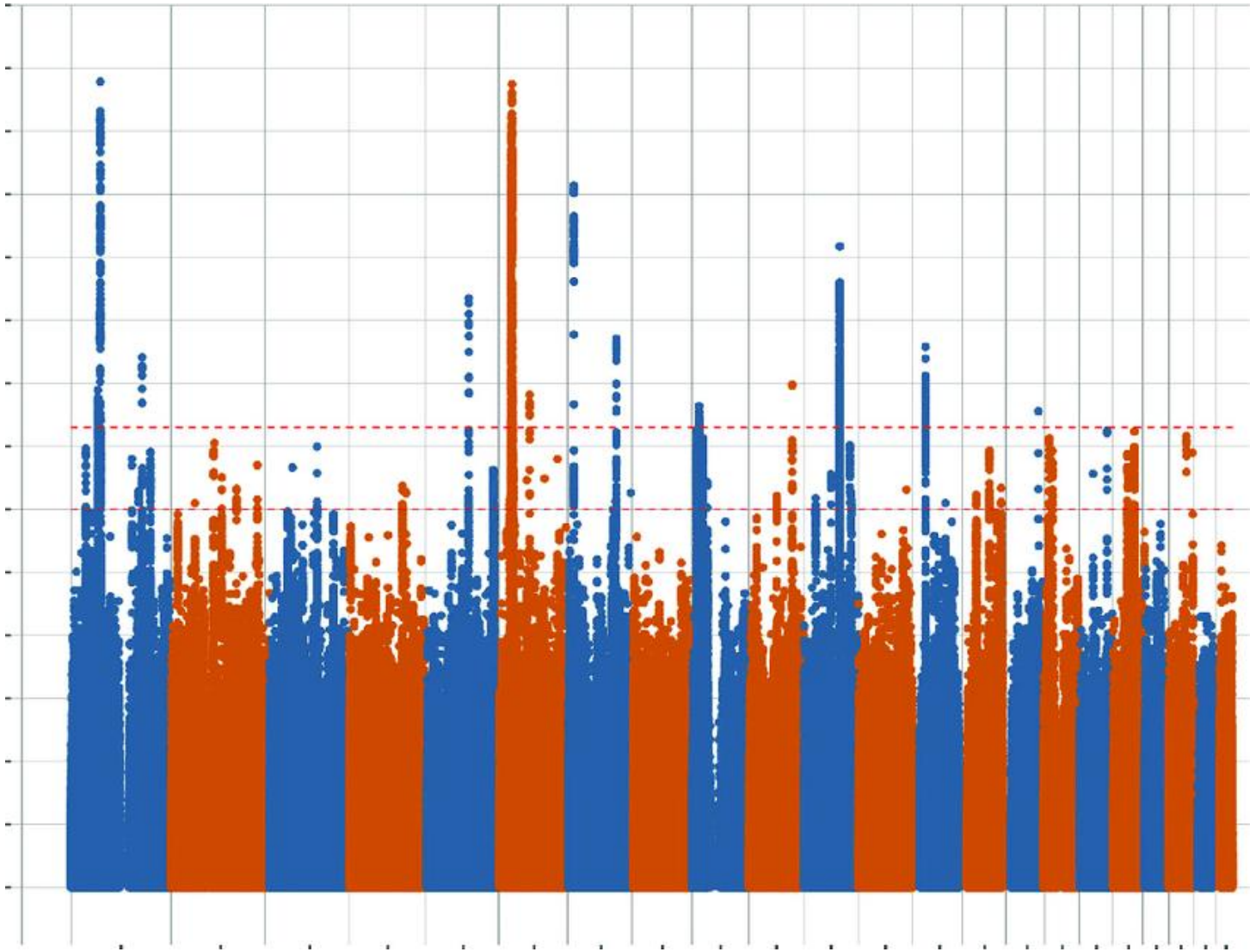


Linkage Disequilibrium

camh
Krembil Centre for Neuroinformatics

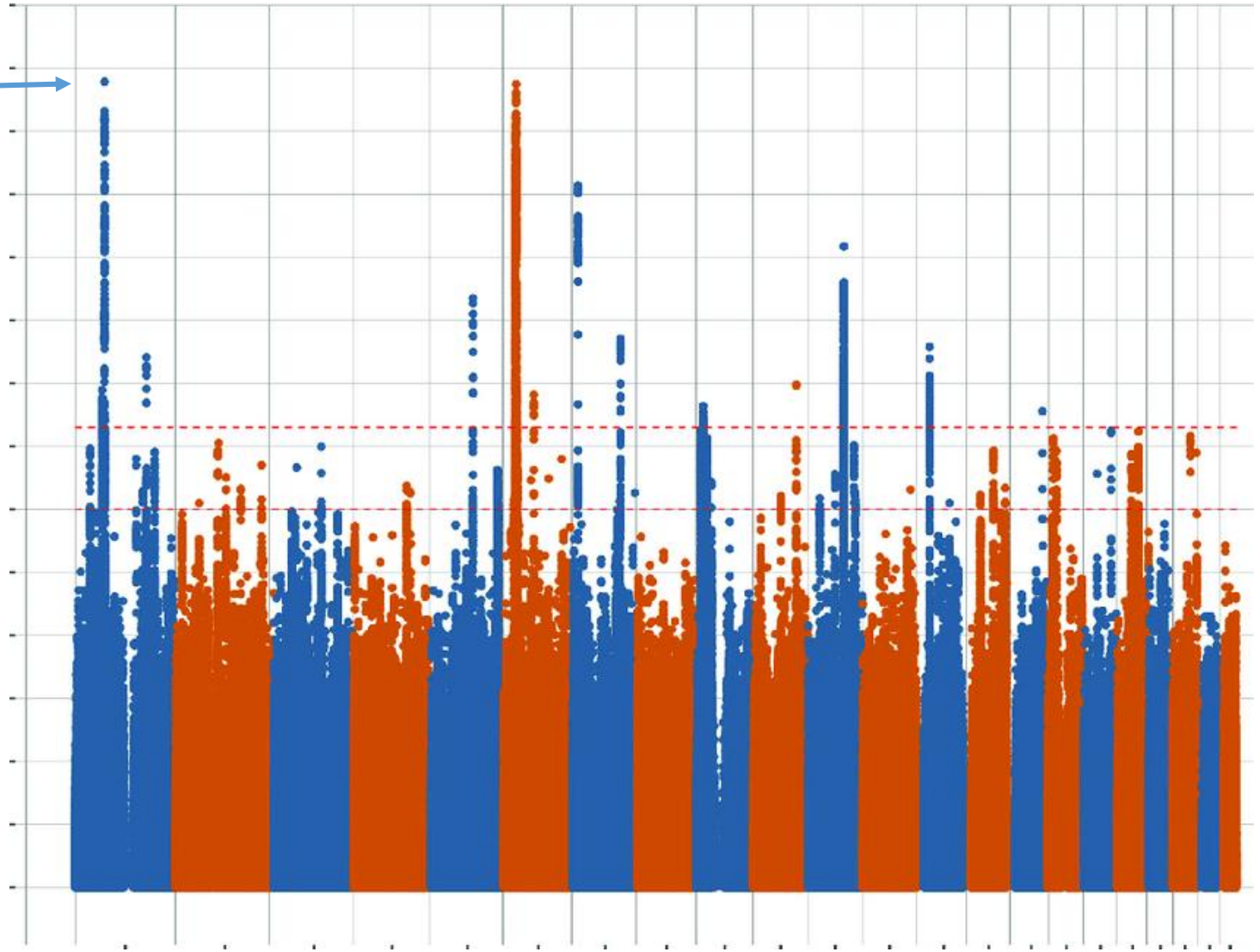


Linkage Disequilibrium

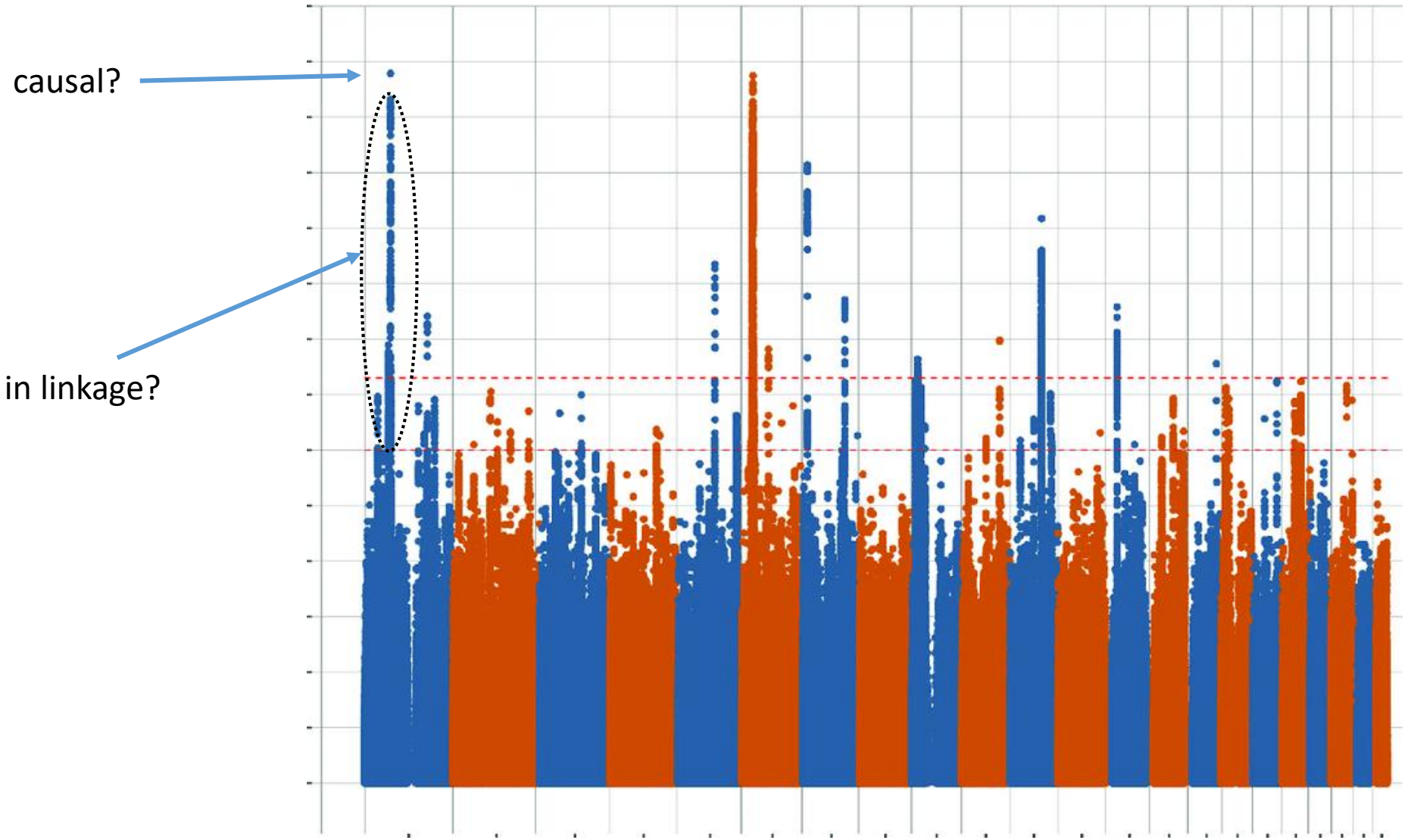


Linkage Disequilibrium

causal?



Linkage Disequilibrium



Polygenic Risk

Polygenic Risk Scores (PRS)

camh
Krembil Centre for Neuroinformatics

Nomenclature

Polygenic risk score (PRS)

Genetic risk score (GRS)

polygenic score (PGS? PS?)

- The above all refer to the same concept and are often exactly the same. The difference between PRS/GRS and a PGS/PS is that using the word “risk” implies that a higher score means higher risk of illness, and a lower score, a lower risk

- We may calculate a score for a trait where the word “risk” is not suitable. E.g. is a polygenic score determining height a ‘risk’ score?

Polygenic Risk Scores (PRS)

What are we trying to do?

- A polygenic score is just a way of *distilling* or *simplifying* data for many genetic variants in one person to a single number.
- In essence it is a method of *dimensionality reduction*, i.e. creating a single number from many.

Example: we have 100 SNPs associated with height in a GWAS. Can we use all 100 SNPs to generate a single score per subject that summarizes their genetically-determined 'risk' for taller or shorter height?

Shared Genetic Risk

camh
Krembil Centre for Neuroinformatics

RESEARCH ARTICLE

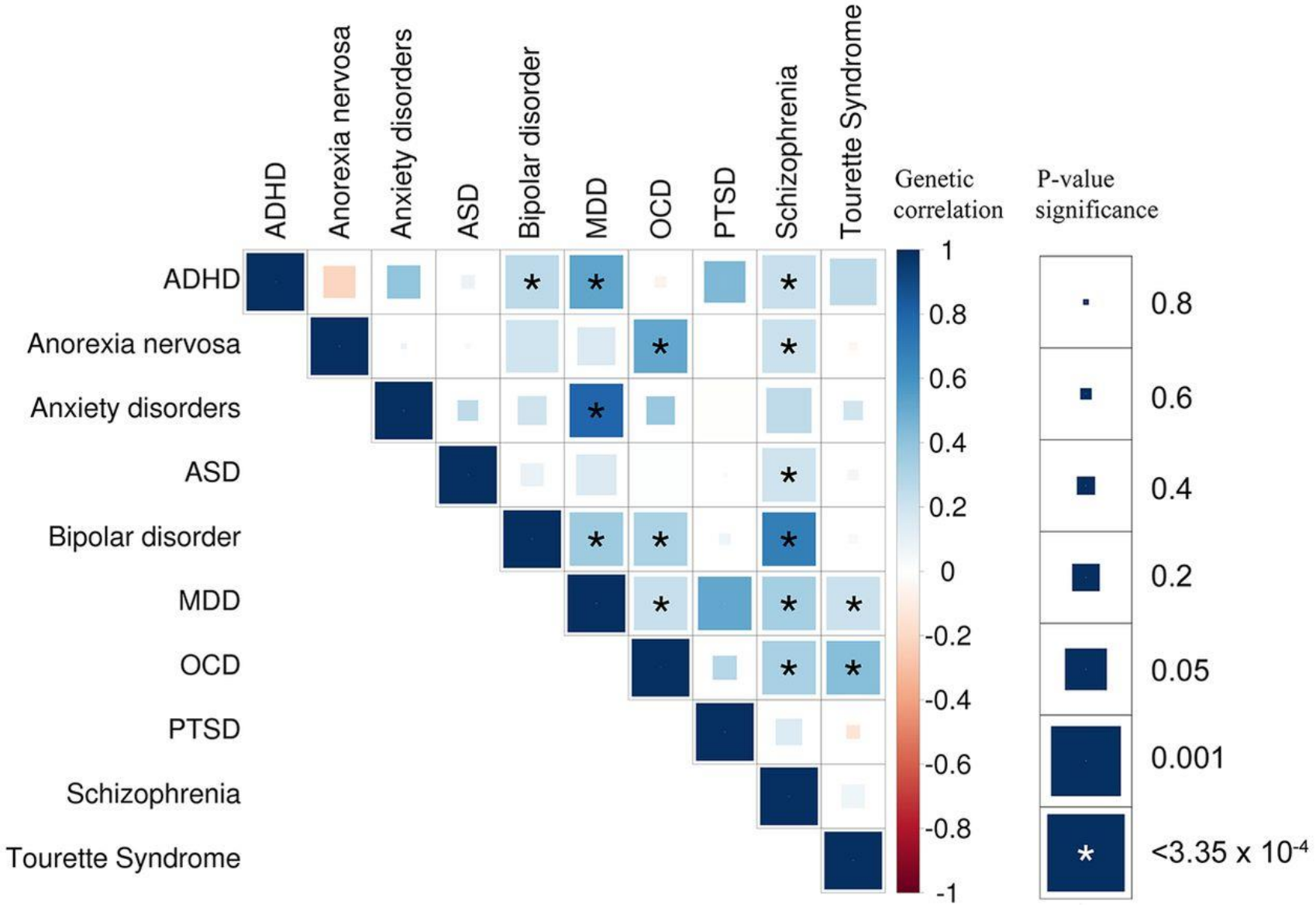
Analysis of shared heritability in common disorders of the brain

The Brainstorm Consortium, Verneri Anttila[1,2,3,*], Brendan Bulik-Sullivan[1,3], Hilary K. Finucane[2,3,4,5], Ra…

+ See all authors and affiliations

Science 22 Jun 2018:
Vol. 360, Issue 6395, eaap8757
DOI: 10.1126/science.aap8757

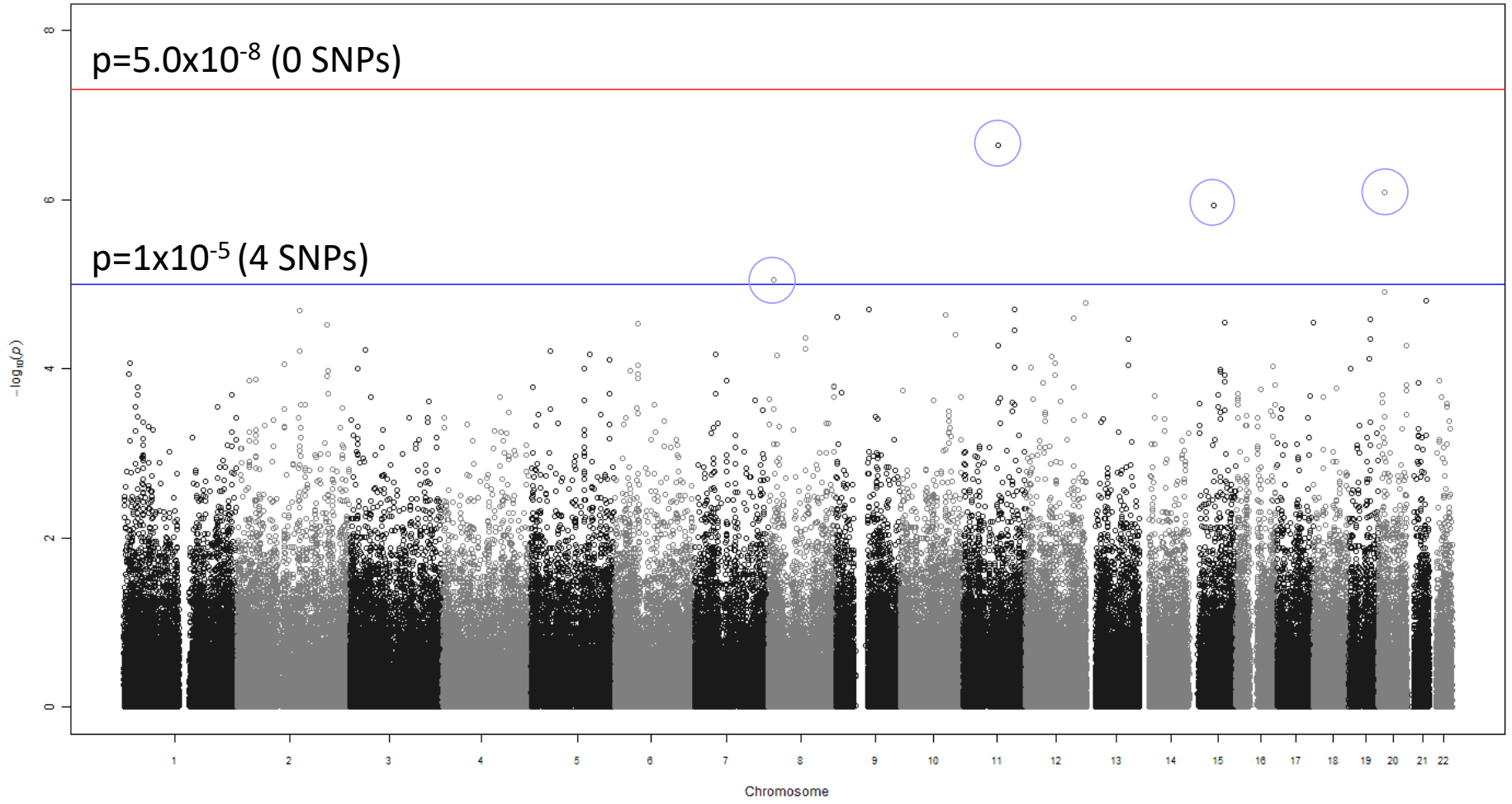
Shared Genetic Risk



Polygenic Risk Scores (PRS)

camh
Krembil Centre for Neuroinformatics

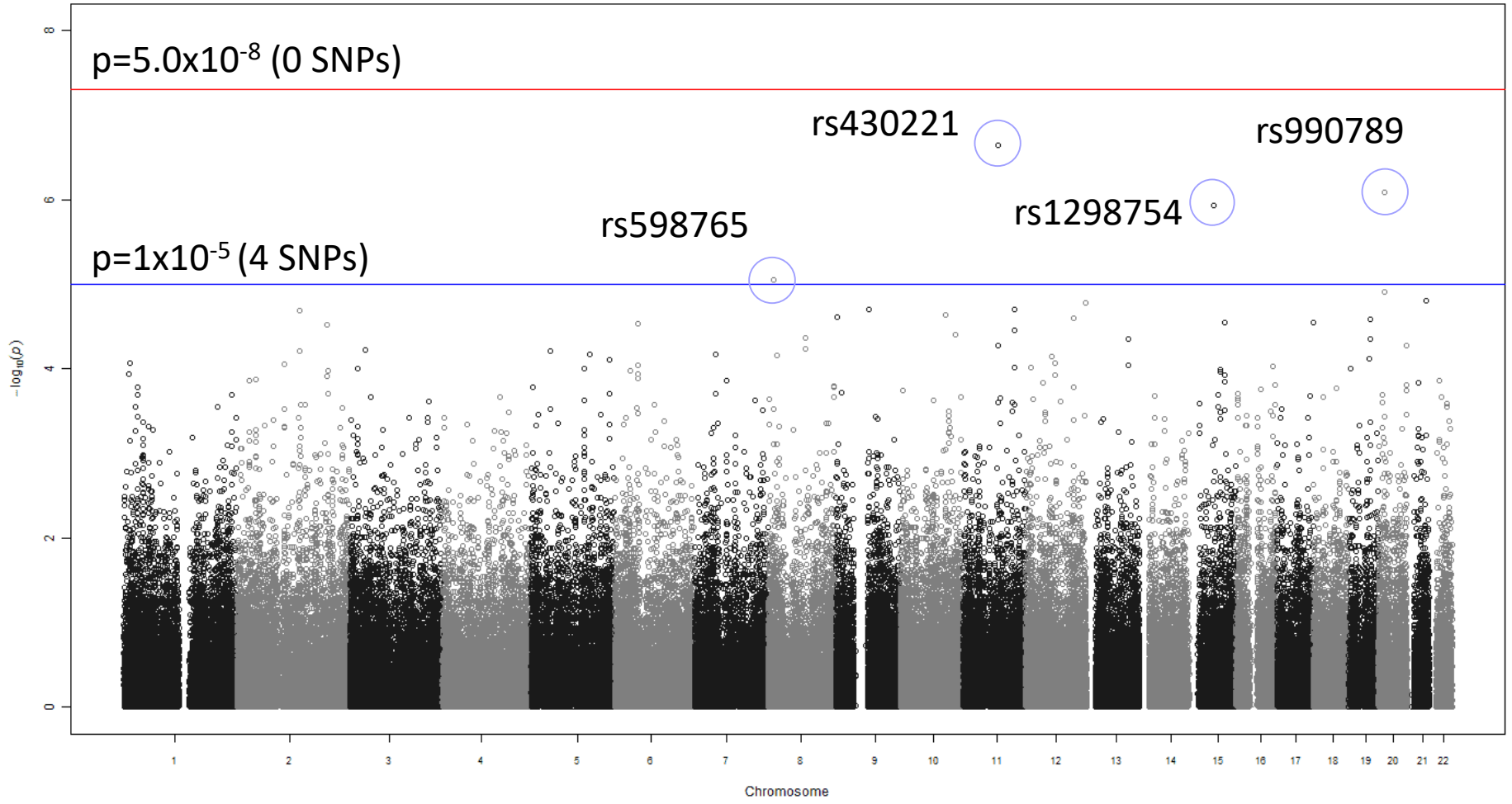
How might we do this simply?



Polygenic Risk Scores (PRS)

camh
Krembil Centre for Neuroinformatics

How might we do this simply?



Polygenic Risk Scores (PRS)

camh
Krembil Centre for Neuroinformatics

How might we do this simply?

- Threshold: $p=1.0x10^{-5}$

rs430221

rs990789

rs598765

rs1298754

	A1	A2	Beta	SE	P-value
	A	G	1.2	0.037	0.000008
	T	G	1.45	0.022	0.0000009
	C	T	1.26	0.0345	0.0000081
	A	G	1.6	0.025	0.0000031

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	GG	TG	CC	AA
Sub2	GA	TG	CT	AG
Sub3	GG	TT	TT	GG
Sub4	GA	TG	TT	GG
Sub5	AA	GG	TT	AA
Sub6	GA	TG	CT	AG
Sub7	GG	TG	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	GG	TG	CC	AA
Sub2	GA	TG	CT	AG
Sub3	GG	TT	TT	GG
Sub4	GA	TG	TT	GG
Sub5	AA	GG	TT	AA
Sub6	GA	TG	CT	AG
Sub7	GG	TG	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	GG (0+0=0)	TG	CC	AA
Sub2	GA (0+1=1)	TG	CT	AG
Sub3	GG (0+0=0)	TT	TT	GG
Sub4	GA (0+1=1)	TG	TT	GG
Sub5	AA (1+1=2)	GG	TT	AA
Sub6	GA (0+1=1)	TG	CT	AG
Sub7	GG (0=0=0)	TG	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	TG	CC	AA
Sub2	1	TG	CT	AG
Sub3	0	TT	TT	GG
Sub4	1	TG	TT	GG
Sub5	2	GG	TT	AA
Sub6	1	TG	CT	AG
Sub7	0	TG	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	TG (1+0=1)	CC	AA
Sub2	1	TG (1+0=1)	CT	AG
Sub3	0	TT (1+1=2)	TT	GG
Sub4	1	TG (1+0=1)	TT	GG
Sub5	2	GG (0+0=0)	TT	AA
Sub6	1	TG (1+0=1)	CT	AG
Sub7	0	TG (1+0=1)	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	1	CC	AA
Sub2	1	1	CT	AG
Sub3	0	2	TT	GG
Sub4	1	1	TT	GG
Sub5	2	0	TT	AA
Sub6	1	1	CT	AG
Sub7	0	1	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	1	2	2
Sub2	1	1	1	1
Sub3	0	2	0	0
Sub4	1	1	0	0
Sub5	2	0	0	2
Sub6	1	1	1	1
Sub7	0	1	1	2

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

$$\text{PRS} = X_1 + X_2 + X_3 + X_4$$

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789	∑
Sub1	0	1	2	2	5
Sub2	1	1	1	1	4
Sub3	0	2	0	0	2
Sub4	1	1	0	0	2
Sub5	2	0	0	2	4
Sub6	1	1	1	1	4
Sub7	0	1	1	2	4

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

$$PRS = (X_1 + X_2 + X_3 + X_4)/n_{SNPs}$$

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789	∑ / 4
Sub1	0	1	2	2	1.25
Sub2	1	1	1	1	1
Sub3	0	2	0	0	0.5
Sub4	1	1	0	0	0.5
Sub5	2	0	0	2	1
Sub6	1	1	1	1	1
Sub7	0	1	1	2	1

Polygenic Risk Scores (PRS)

Q: Can we improve on this?

- A: Yes! In two simple ways

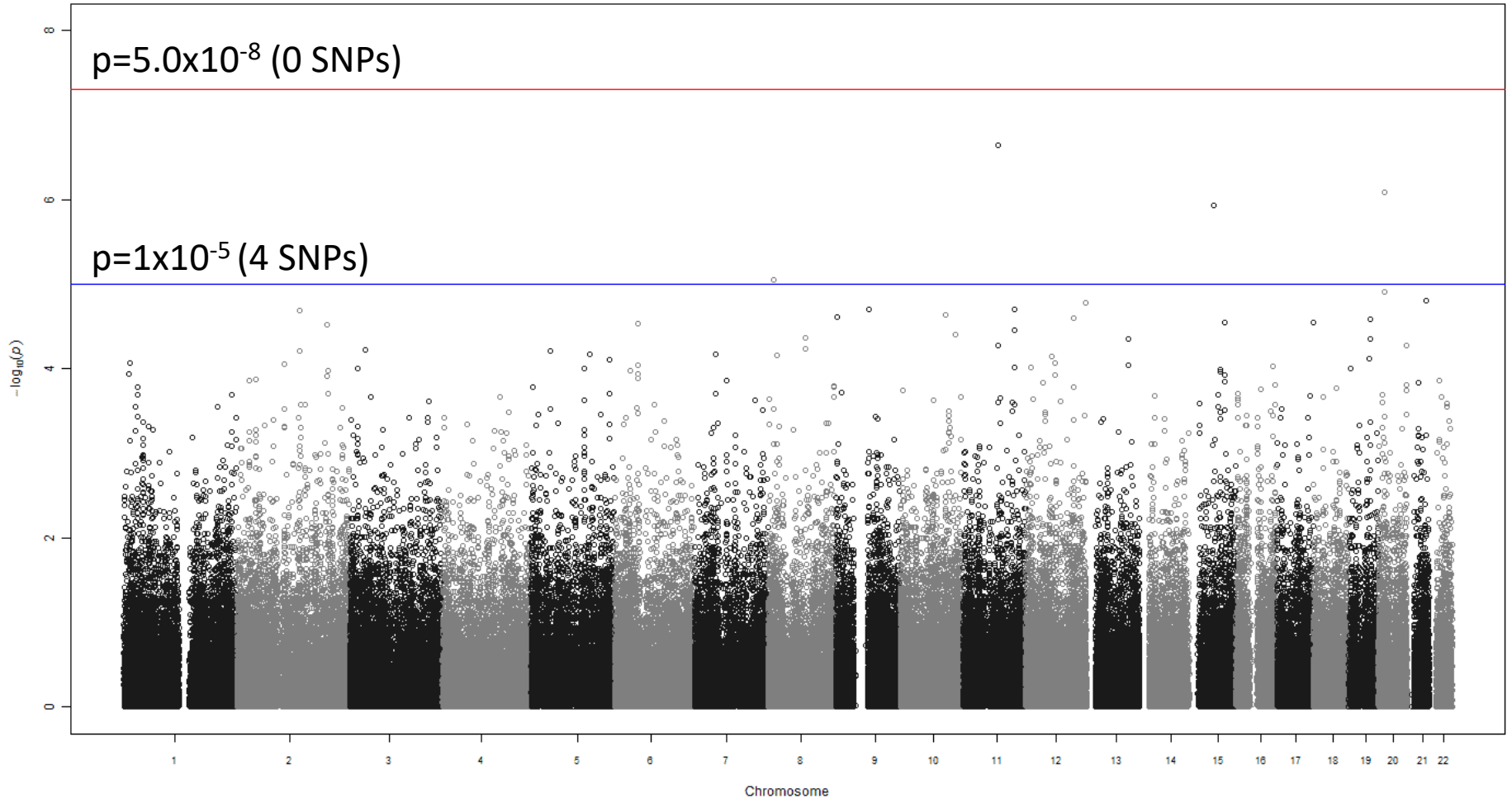
1. Raise the *p*-value threshold to include more SNPs
2. Take into account the strength of the effect (beta coefficient)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Polygenic Risk Scores (PRS)

camh
Krembil Centre for Neuroinformatics

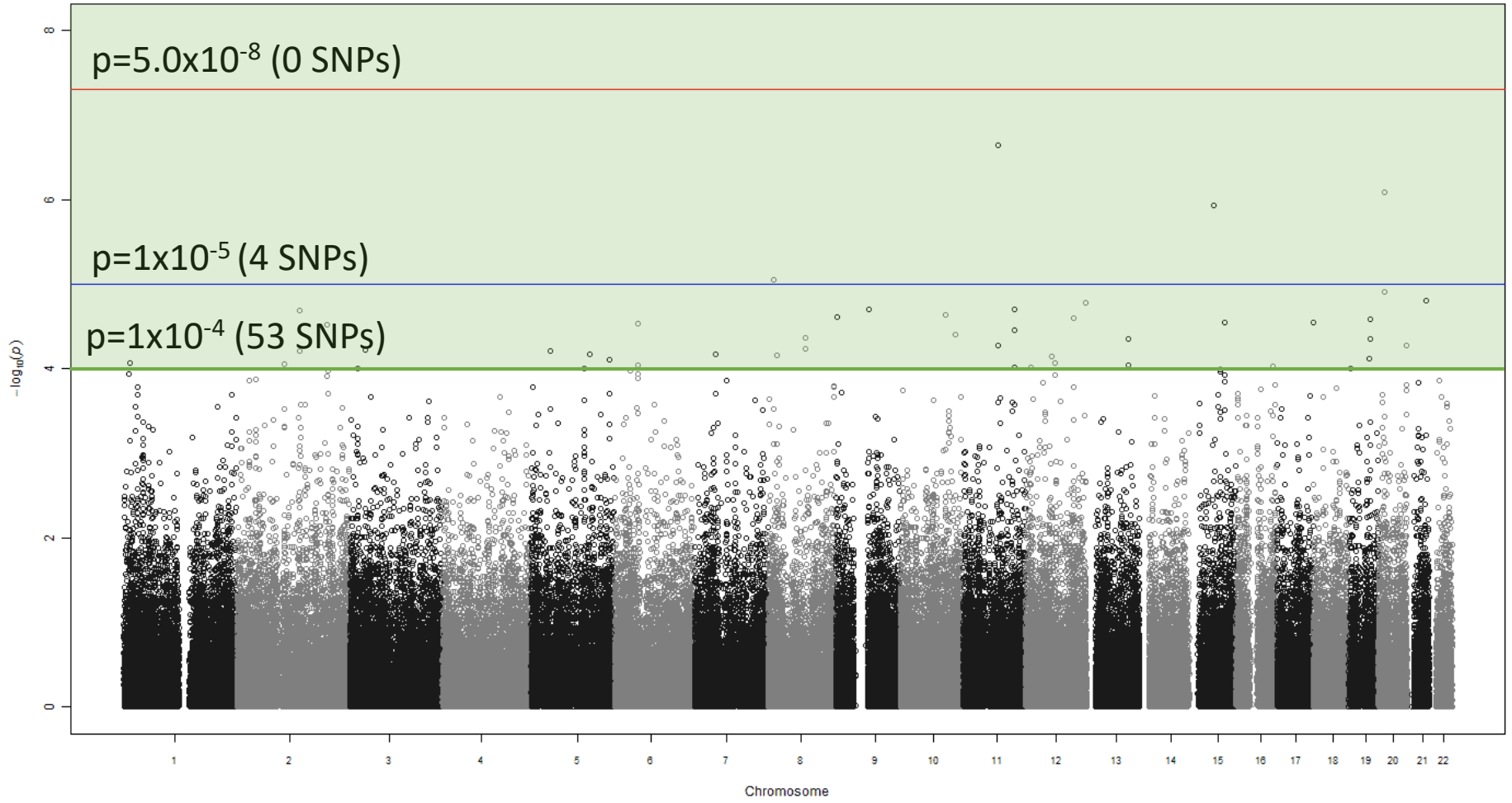
Including more SNPs



Polygenic Risk Scores (PRS)

camh
Krembil Centre for Neuroinformatics

Including more SNPs



Polygenic Risk Scores (PRS)

Weighting by effect

$$\text{PRS} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{n_{SNPs}} X_{n_{SNPs}}$$

We input the # of effect alleles for $X_1 \ldots X_{nSNPS}$ and use beta coefficients from a GWAS as $\beta_1 \ldots \beta_{nSNPs}$

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	GG	TG	CC	AA
Sub2	GA	TG	CT	AG
Sub3	GG	TT	TT	GG
Sub4	GA	TG	TT	GG
Sub5	AA	GG	TT	AA
Sub6	GA	TG	CT	AG
Sub7	GG	TG	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	TG	CC	AA
Sub2	1	TG	CT	AG
Sub3	0	TT	TT	GG
Sub4	1	TG	TT	GG
Sub5	2	GG	TT	AA
Sub6	1	TG	CT	AG
Sub7	0	TG	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0*1.2=0	TG	CC	AA
Sub2	1*1.2=1.2	TG	CT	AG
Sub3	0*1.2=0	TT	TT	GG
Sub4	1*1.2=1.2	TG	TT	GG
Sub5	2*1.2=2.4	GG	TT	AA
Sub6	1*1.2=1.2	TG	CT	AG
Sub7	0*1.2=0	TG	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	TG	CC	AA
Sub2	1.2	TG	CT	AG
Sub3	0	TT	TT	GG
Sub4	1.2	TG	TT	GG
Sub5	2.4	GG	TT	AA
Sub6	1.2	TG	CT	AG
Sub7	0	TG	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	1	CC	AA
Sub2	1.2	1	CT	AG
Sub3	0	2	TT	GG
Sub4	1.2	1	TT	GG
Sub5	2.4	0	TT	AA
Sub6	1.2	1	CT	AG
Sub7	0	1	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	1*1.45	CC	AA
Sub2	1.2	1*1.45	CT	AG
Sub3	0	2*1.45	TT	GG
Sub4	1.2	1*1.45	TT	GG
Sub5	2.4	0*1.45	TT	AA
Sub6	1.2	1*1.45	CT	AG
Sub7	0	1*1.45	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789
Sub1	0	1.45	CC	AA
Sub2	1.2	1.45	CT	AG
Sub3	0	2.9	TT	GG
Sub4	1.2	1.45	TT	GG
Sub5	2.4	0	TT	AA
Sub6	1.2	1.45	CT	AG
Sub7	0	1.45	CT	AA

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

$$\text{PRS} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{n_{SNPs}} X_{n_{SNPs}}$$

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789	∑
Sub1	0	1.45	2.52	3.2	7.17
Sub2	1.2	1.45	1.26	1.6	5.51
Sub3	0	2.9	0	0	2.9
Sub4	1.2	1.45	0	0	2.65
Sub5	2.4	0	0	3.2	5.6
Sub6	1.2	1.45	1.26	1.6	5.51
Sub7	0	1.45	1.26	3.2	5.91

GWAS results
(PLINK output)

	A1	A2	Beta	SE	P-value
rs598765	A	G	1.2	0.037	0.000008
rs430221	T	G	1.45	0.022	0.0000009
rs1298754	C	T	1.26	0.0345	0.0000081
rs990789	A	G	1.6	0.025	0.0000031

$$\text{PRS} = (\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{n_{SNPs}} X_{n_{SNPs}})/n_{SNPs}$$

Subject data
(PLINK input)

Subject	rs598765	rs430221	rs1298754	rs990789	∑ / 4
Sub1	0	1.45	2.52	3.2	1.7925
Sub2	1.2	1.45	1.26	1.6	1.3775
Sub3	0	2.9	0	0	0.725
Sub4	1.2	1.45	0	0	0.6625
Sub5	2.4	0	0	3.2	1.4
Sub6	1.2	1.45	1.26	1.6	1.3775
Sub7	0	1.45	1.26	3.2	1.4775

Polygenic Risk Scores (PRS)

camh
Krembil Centre for Neuroinformatics

Compare

$$\text{PRS} = (X_1 + X_2 + X_3 + X_4)/n_{SNPs}$$

$$\text{PRS} = (\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{n_{SNPs}} X_{n_{SNPs}})/n_{SNPs}$$

Subject	Unweighted		Weighted	
	PGS	Rank	PGS	Rank
Sub1	1.25	1	1.7925	1
Sub2	1	2	1.3775	4
Sub3	0.5	3	0.725	5
Sub4	0.5	3	0.6625	6
Sub5	1	2	1.4	3
Sub6	1	2	1.3775	4
Sub7	1	2	1.4775	2

Polygenic Risk Scores (PRS)

Key considerations

- Resolution
  - The fewer SNPs in the score, the less normal the distribution of the score. Some methods use feature selection (e.g. LASSO regression) to identify optimal numbers of SNPs.

- Ancestry
  - A PRS cannot be calculated in one ancestry and applied to another
  - Most available GWAS summary statistics are from Caucasian populations!

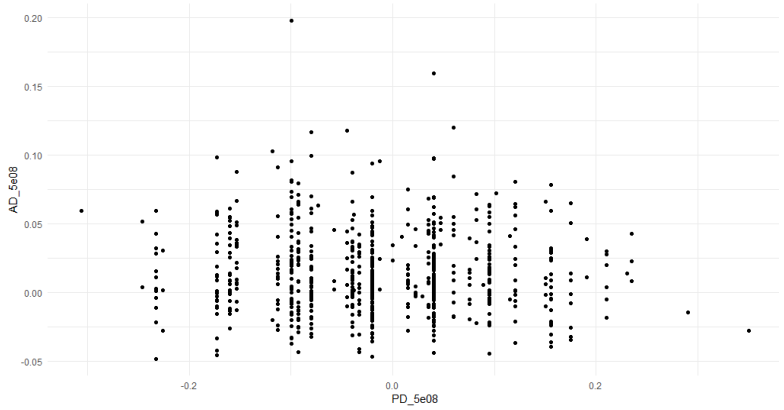
Polygenic Risk Scores (PRS)

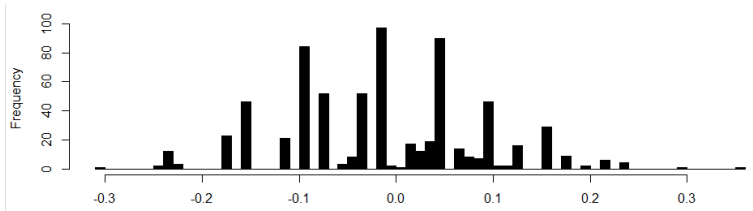
camh
Krembil Centre for Neuroinformatics

Resolution (PD example)

$p<5x10^{-8}$

$N_{snps}=4$





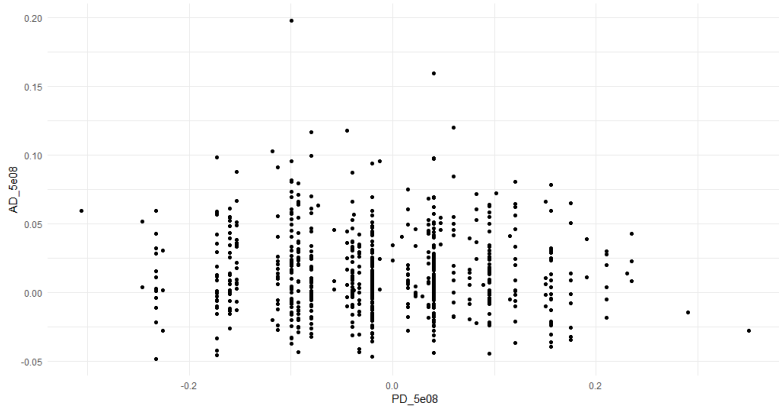
Polygenic Risk Scores (PRS)

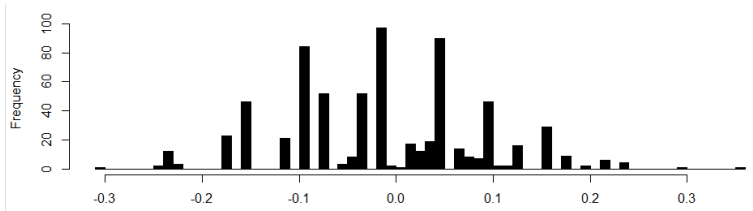
camh
Krembil Centre for Neuroinformatics

Resolution (PD example)

$p<5x10^{-8}$

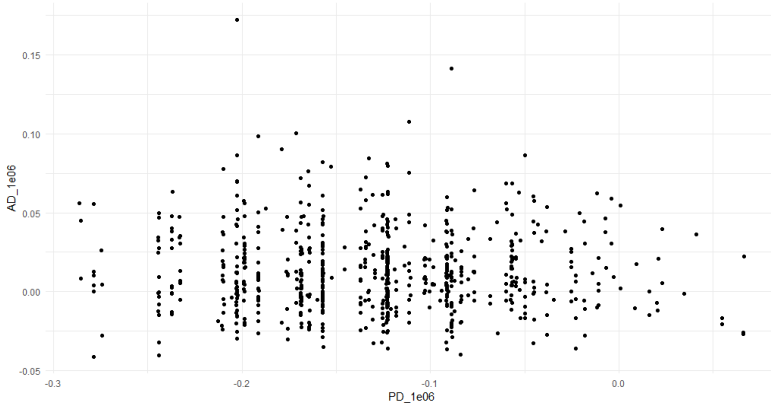
$N_{snps}=4$

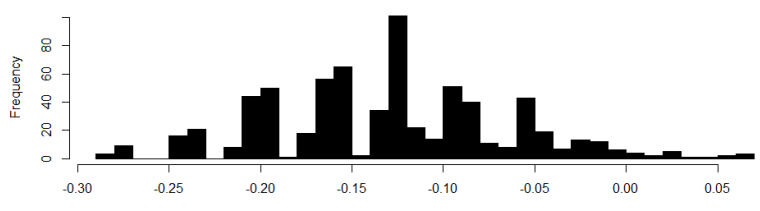




$p<1x10^{-6}$

$N_{snps}=7$



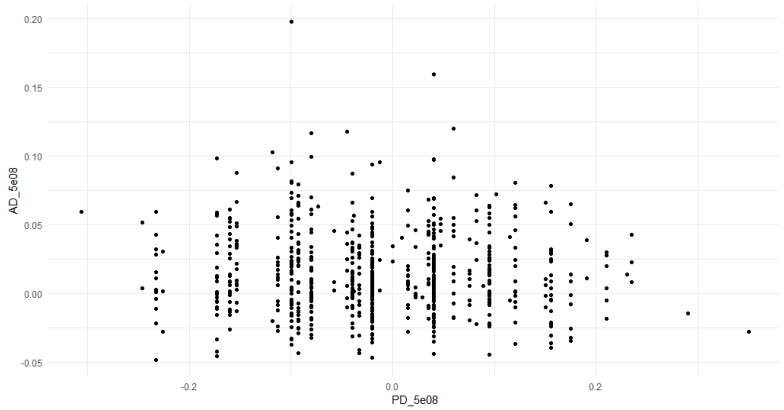


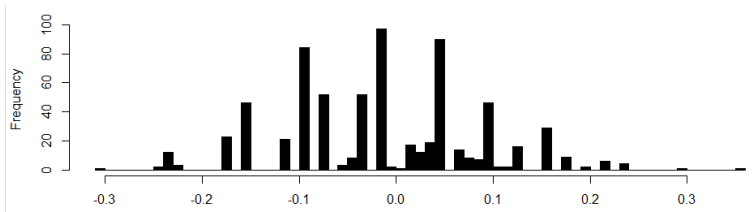
Polygenic Risk Scores (PRS)

camh
Krembil Centre for Neuroinformatics

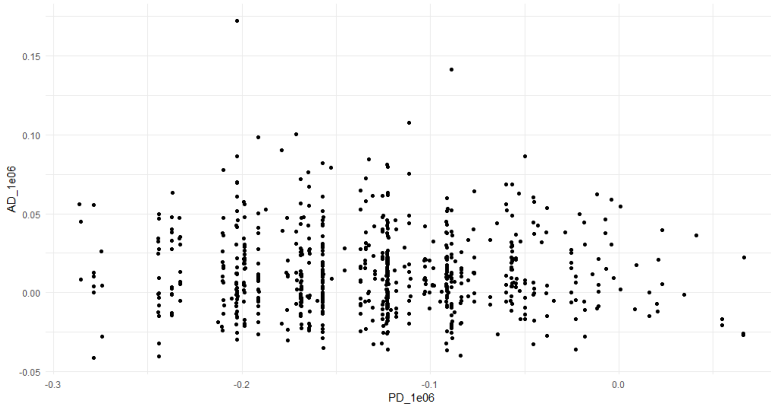
Resolution (PD example)

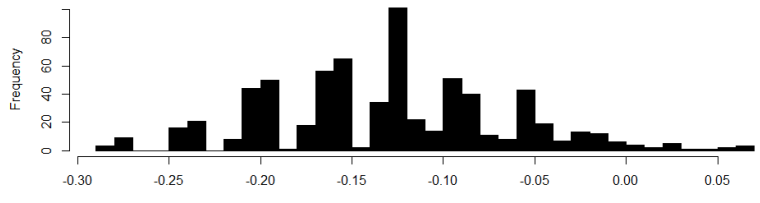
$p<5x10^{-8}$
$N_{snps}=4$



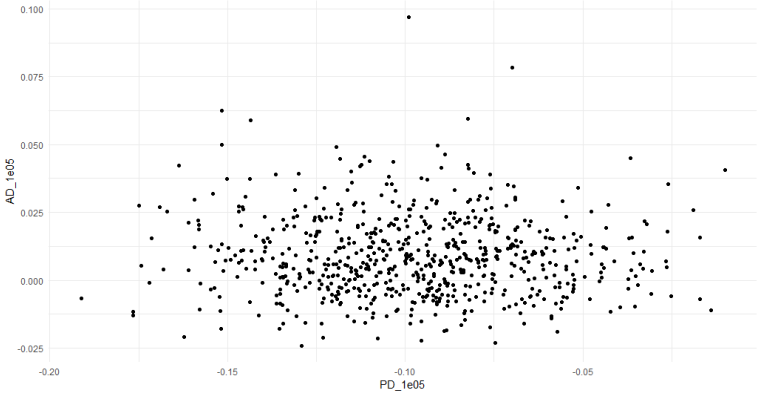


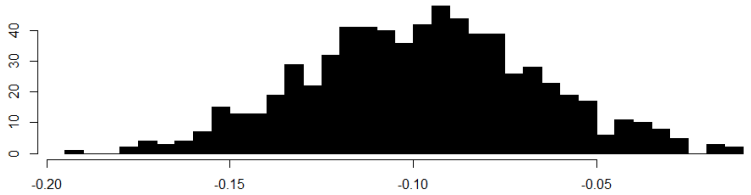
$p<1x10^{-6}$
$N_{snps}=7$





$p<1x10^{-5}$
$N_{snps}=22$

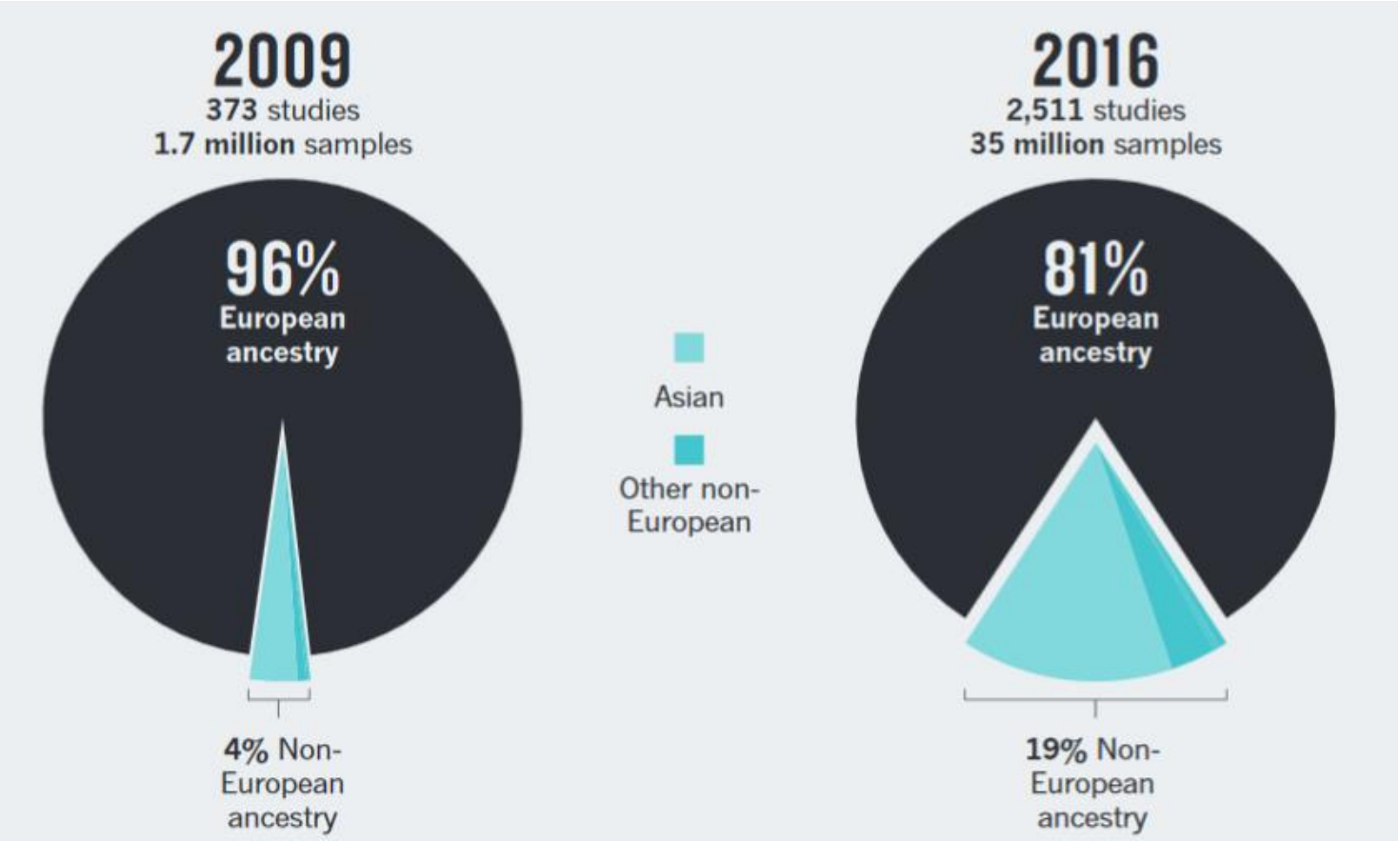




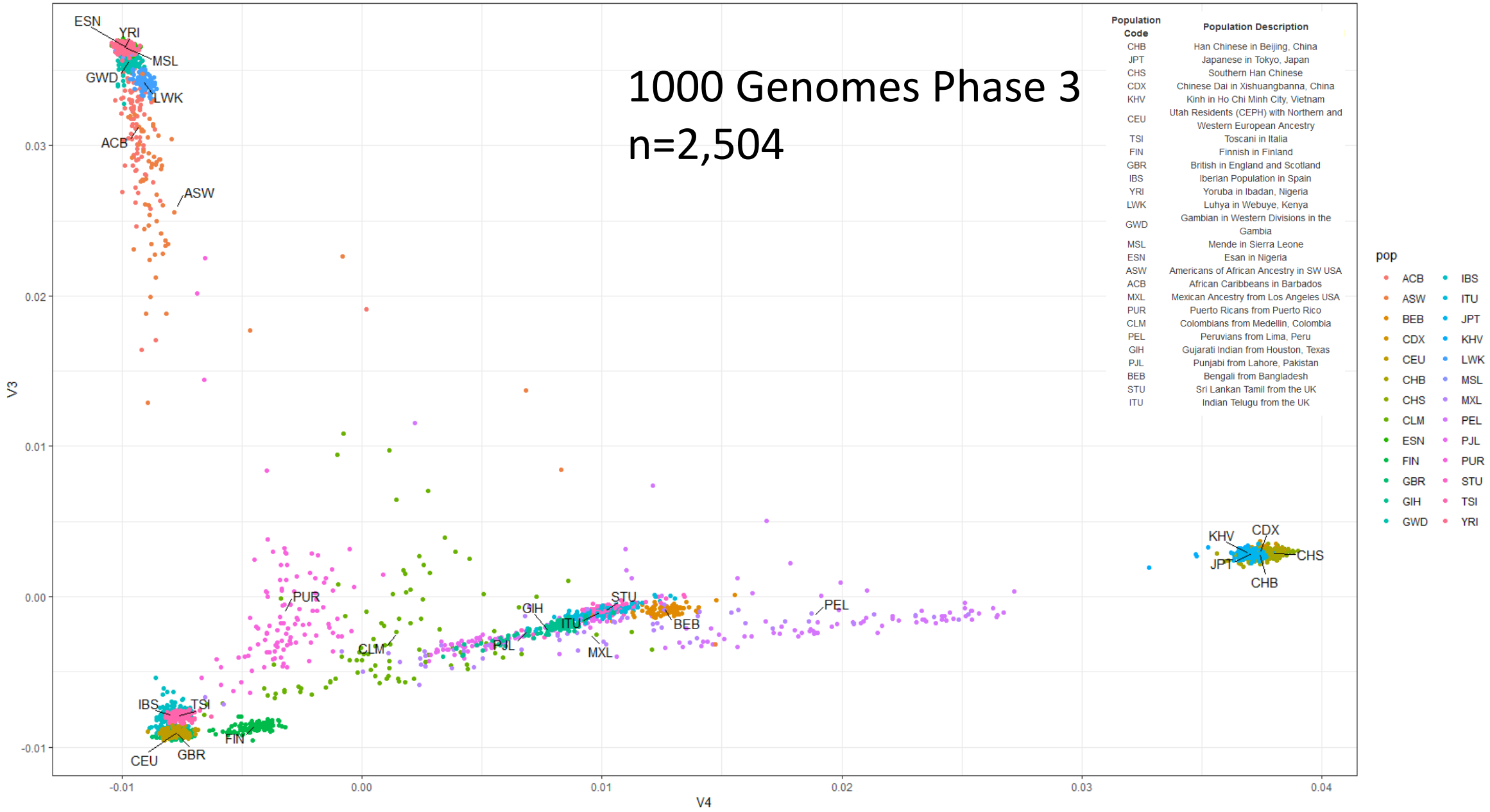
Polygenic Risk Scores (PRS)

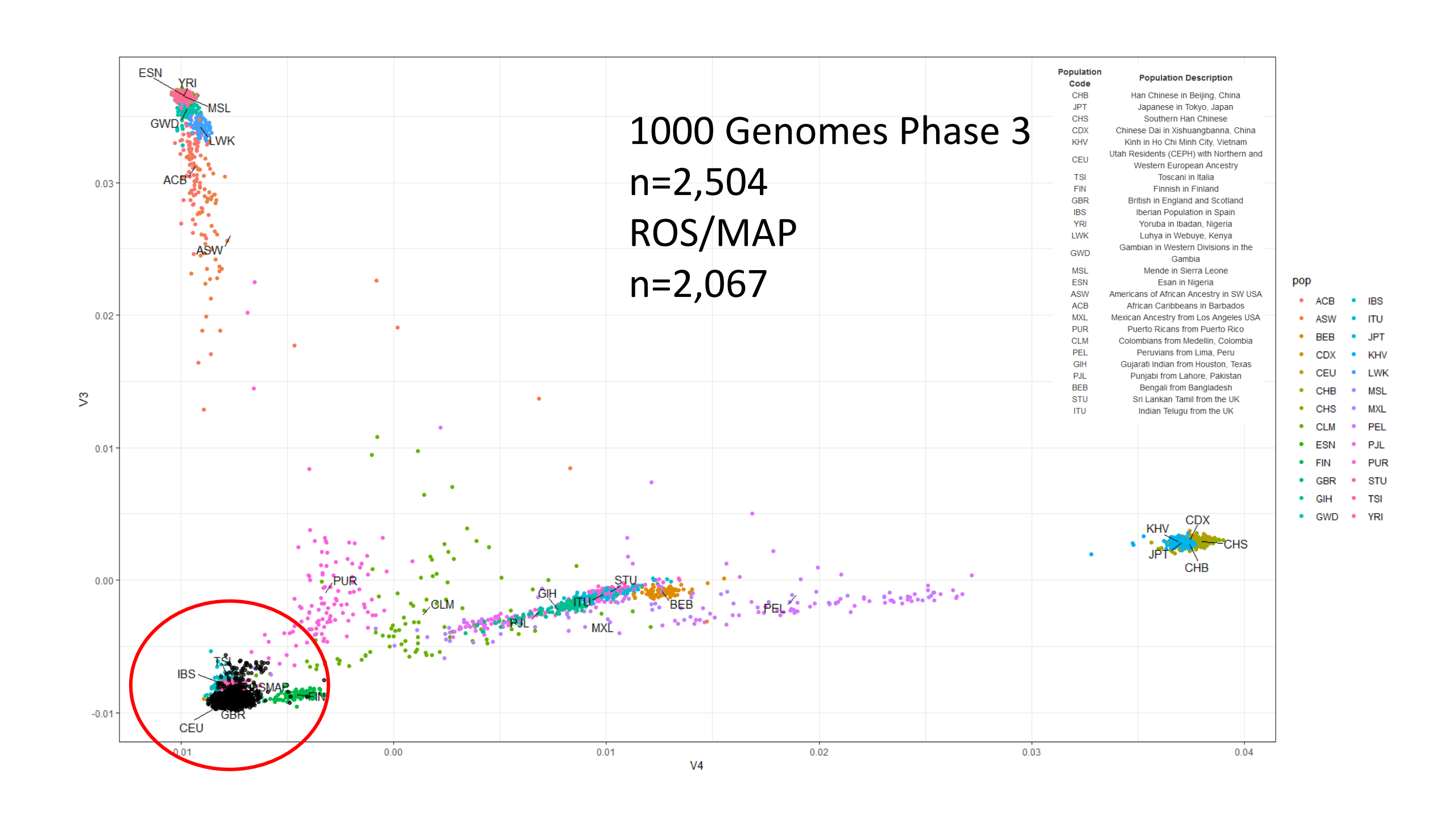
camh
Krembil Centre for Neuroinformatics

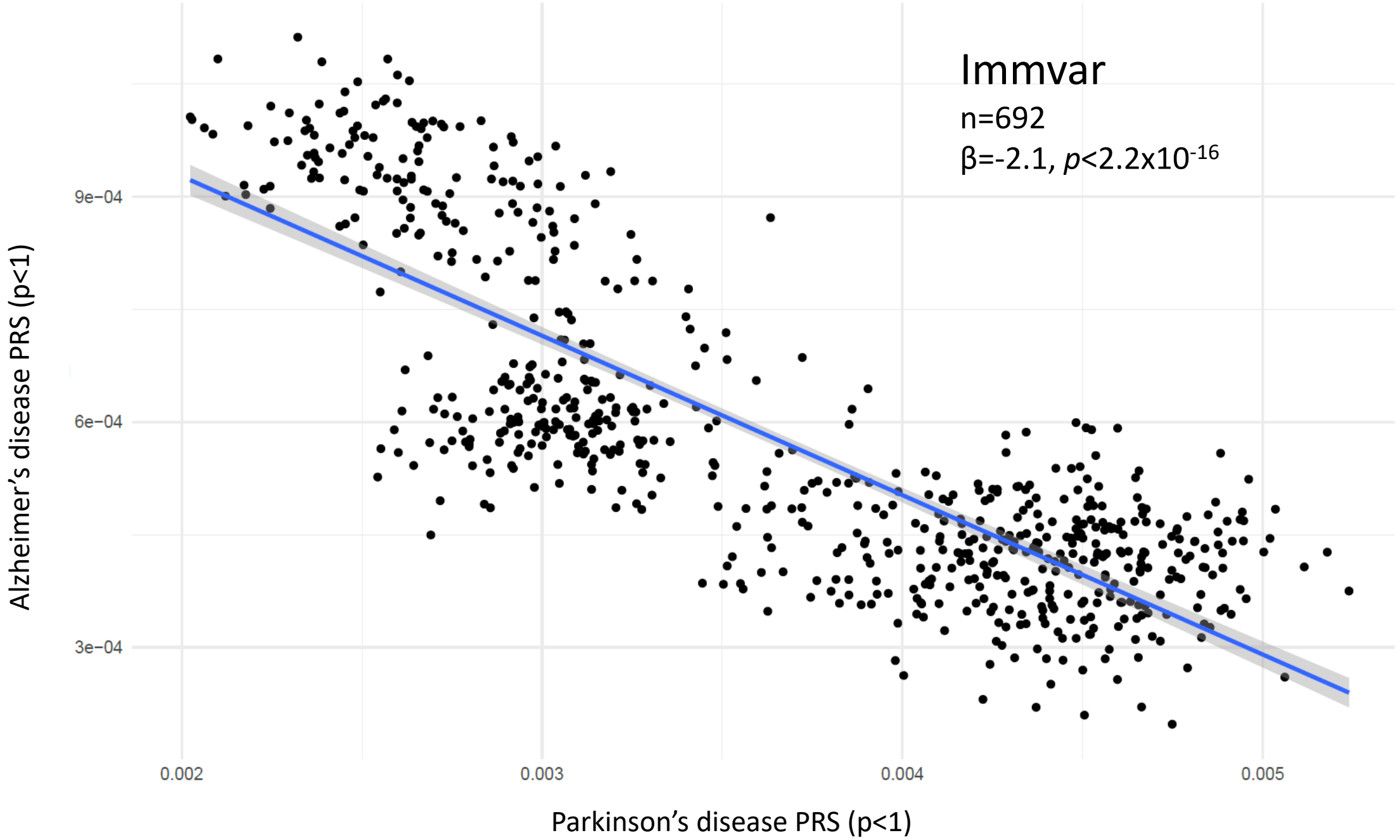
Ethnicity

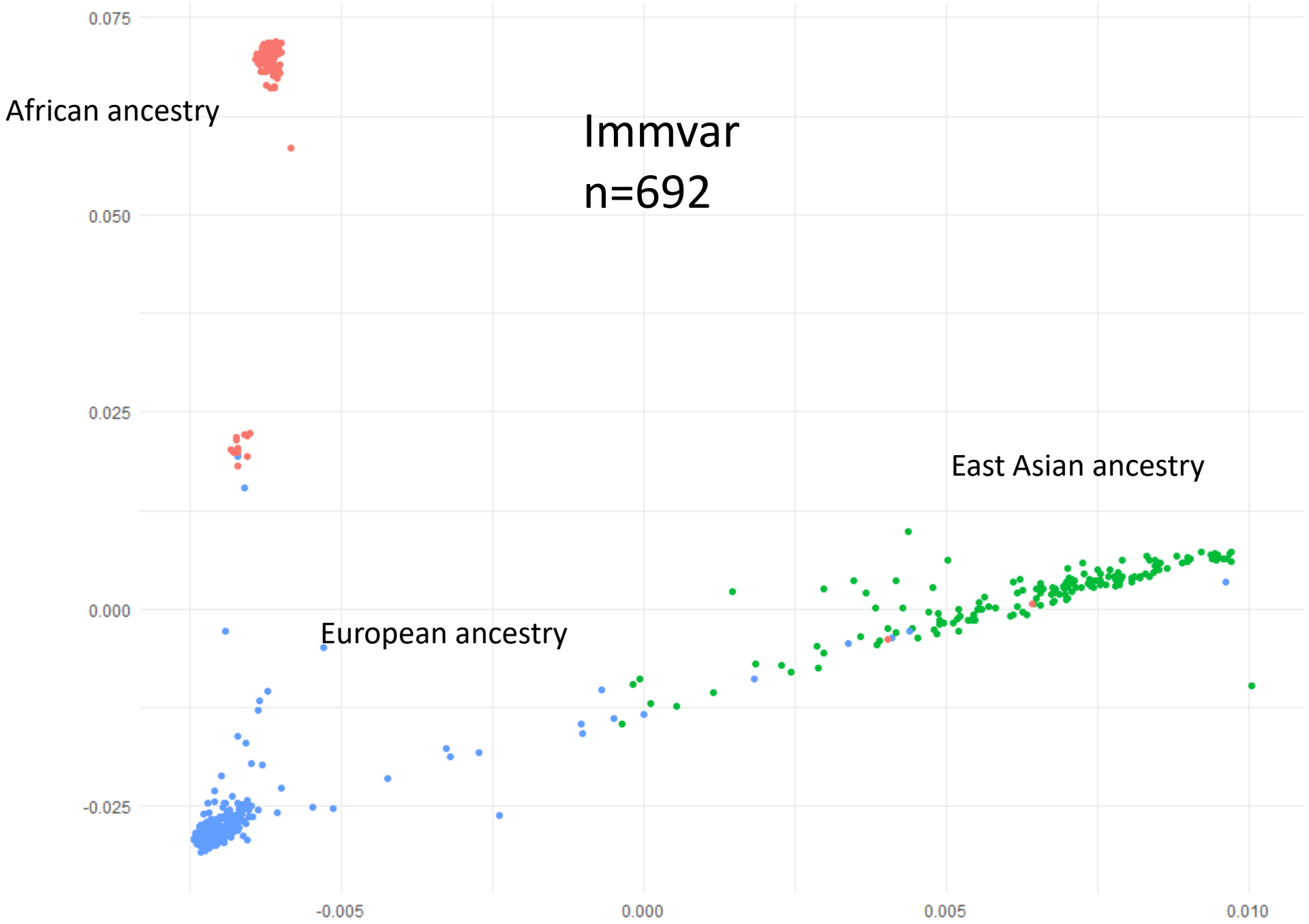


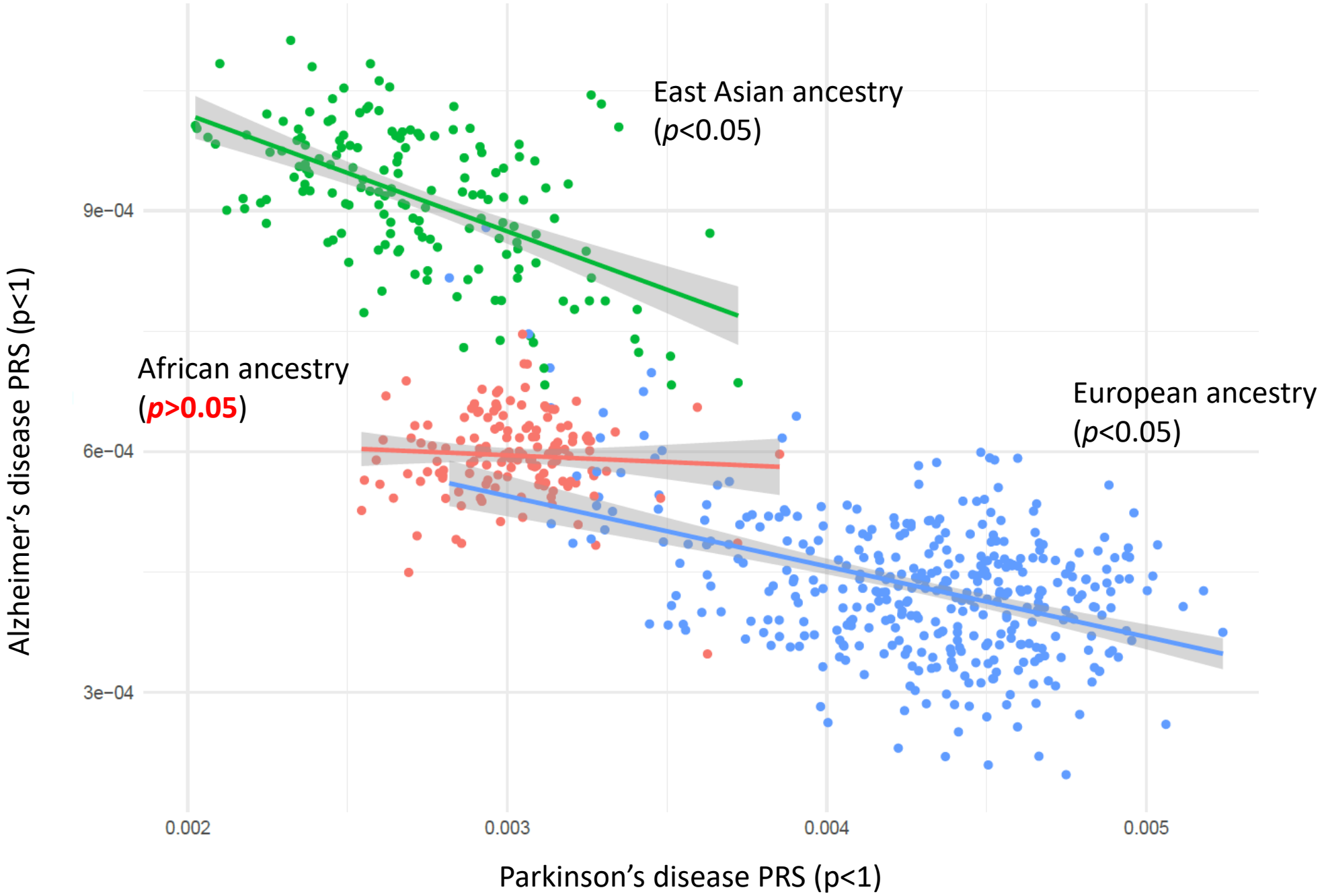
(Popejoy and Fullerton, Nature, 2016)


1000 Genomes Phase 3
n=2,504
ESN
YRI
MSL
GWD
LWK
ACB
ASW
PUR
CLM
GIH
PJL
ITU
STU
MXL
BEB
PEL
IBS
TSI
FIN
CEU
GBR
KHV
CDX
JPT
CHS
CHB
V3
V4
-0.01
0.00
0.01
0.02
0.03
0.04
Population Code
Population Description
CHB Han Chinese in Beijing, China
JPT Japanese in Tokyo, Japan
CHS Southern Han Chinese
CDX Chinese Dai in Xishuangbanna, China
KHV Kinh in Ho Chi Minh City, Vietnam
CEU Utah Residents (CEPH) with Northern and Western European Ancestry
TSI Toscani in Italia
FIN Finnish in Finland
GBR British in England and Scotland
IBS Iberian Population in Spain
YRI Yoruba in Ibadan, Nigeria
LWK Luhya in Webuye, Kenya
GWD Gambian in Western Divisions in the Gambia
MSL Mende in Sierra Leone
ESN Esan in Nigeria
ASW Americans of African Ancestry in SW USA
ACB African Caribbeans in Barbados
MXL Mexican Ancestry from Los Angeles USA
PUR Puerto Ricans from Puerto Rico
CLM Colombians from Medellin, Colombia
PEL Peruvians from Lima, Peru
GIH Gujarati Indian from Houston, Texas
PJL Punjabi from Lahore, Pakistan
BEB Bengali from Bangladesh
STU Sri Lankan Tamil from the UK
ITU Indian Telugu from the UK
pop
ACB
ASW
BEB
CDX
CEU
CHB
CHS
CLM
ESN
FIN
GBR
GIH
GWD
IBS
ITU
JPT
KHV
LWK
MSL
MXL
PEL
PJL
PUR
STU
TSI
YRI


1000 Genomes Phase 3
n=2,504
ROS/MAP
n=2,067
Population Code
Population Description
CHB Han Chinese in Beijing, China
JPT Japanese in Tokyo, Japan
CHS Southern Han Chinese
CDX Chinese Dai in Xishuangbanna, China
KHV Kinh in Ho Chi Minh City, Vietnam
CEU Utah Residents (CEPH) with Northern and Western European Ancestry
TSI Toscani in Italia
FIN Finnish in Finland
GBR British in England and Scotland
IBS Iberian Population in Spain
YRI Yoruba in Ibadan, Nigeria
LWK Luhya in Webuye, Kenya
GWD Gambian in Western Divisions in the Gambia
MSL Mende in Sierra Leone
ESN Esan in Nigeria
ASW Americans of African Ancestry in SW USA
ACB African Caribbeans in Barbados
MXL Mexican Ancestry from Los Angeles USA
PUR Puerto Ricans from Puerto Rico
CLM Colombians from Medellin, Colombia
PEL Peruvians from Lima, Peru
GIH Gujarati Indian from Houston, Texas
PJL Punjabi from Lahore, Pakistan
BEB Bengali from Bangladesh
STU Sri Lankan Tamil from the UK
ITU Indian Telugu from the UK
pop
ACB ASW BEB CDX CEU CHB CHS CLM ESN FIN GBR GIH GWD IBS ITU JPT KHV LWK MSL MXL PEL PJL PUR STU TSI YRI
V3
V4
0.03
0.02
0.01
0.00
-0.01
0.01
0.00
0.01
0.02
0.03
0.04

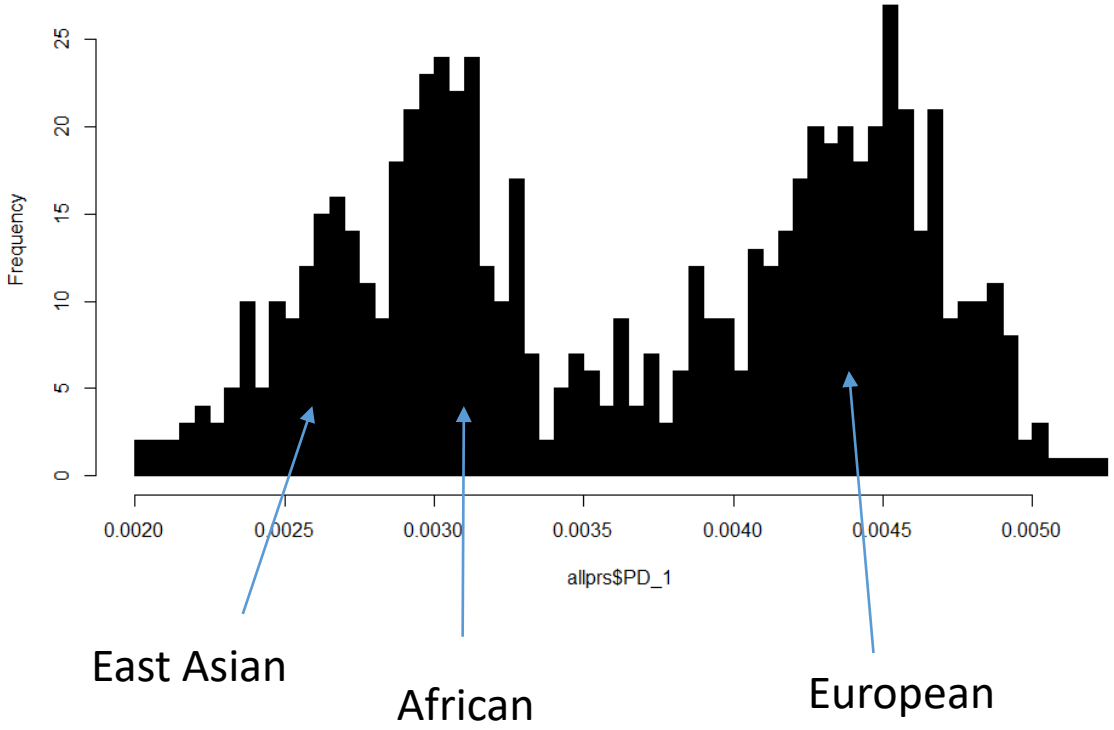

Immvar
n=692
β=-2.1, p<2.2x10-16
Alzheimer's disease PRS (p<1)
9e−04
6e−04
3e−04
0.002
0.003
0.004
0.005
Parkinson's disease PRS (p<1)


African ancestry
Immvar
n=692
East Asian ancestry
European ancestry
0.075
0.050
0.025
0.000
-0.025
-0.005
0.000
0.005
0.010

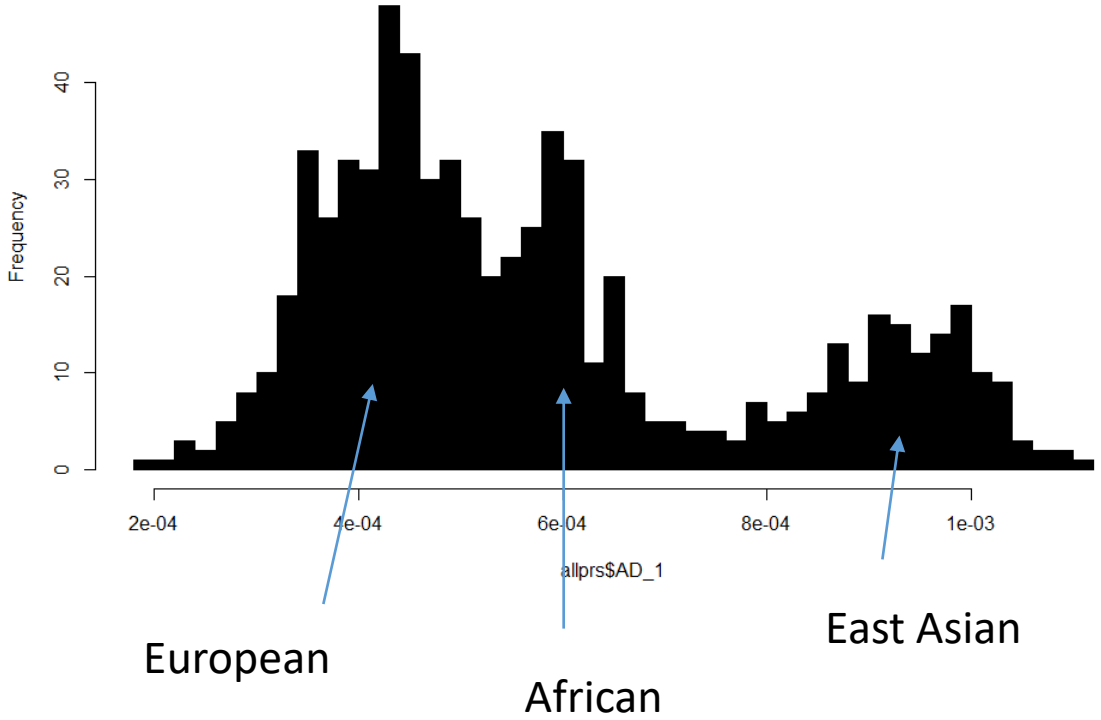

East Asian ancestry
(p<0.05)
African ancestry
(p>0.05)
European ancestry
(p<0.05)
Alzheimer's disease PRS (p<1)
9e−04
6e−04
3e−04
0.002
0.003
0.004
0.005
Parkinson's disease PRS (p<1)

Ethnicity (Immvar example)

Parkinson's (p<1), Nalls et al, 2009



Alzheimer's (p<1), Lambert et al., 2013



Calculation of polygenic risk for MDD

Steps and Aims

Step 1: perform quality control of genetic data

Step 2: perform quality control of GWAS summary statistics

Step 3: identify parameters for PRS calculation

Step 4: perform a high-resolution PRS analysis with depression phenotype

Step 5: calculate a set of PRS to export for further analyses