Genome-Wide Association Studies

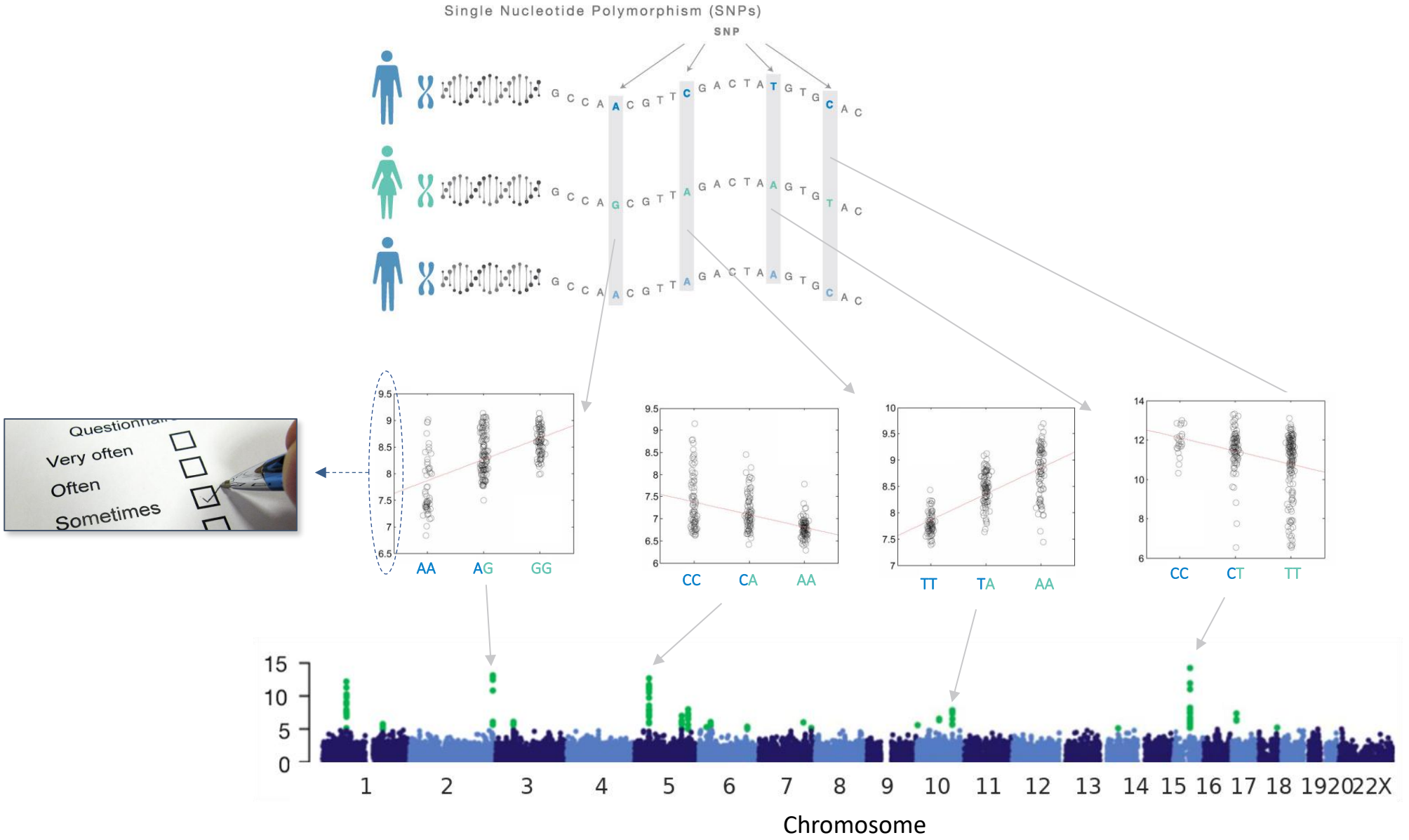
Part 2: signal versus noise

International Statistical Genetics Workshop – 2026

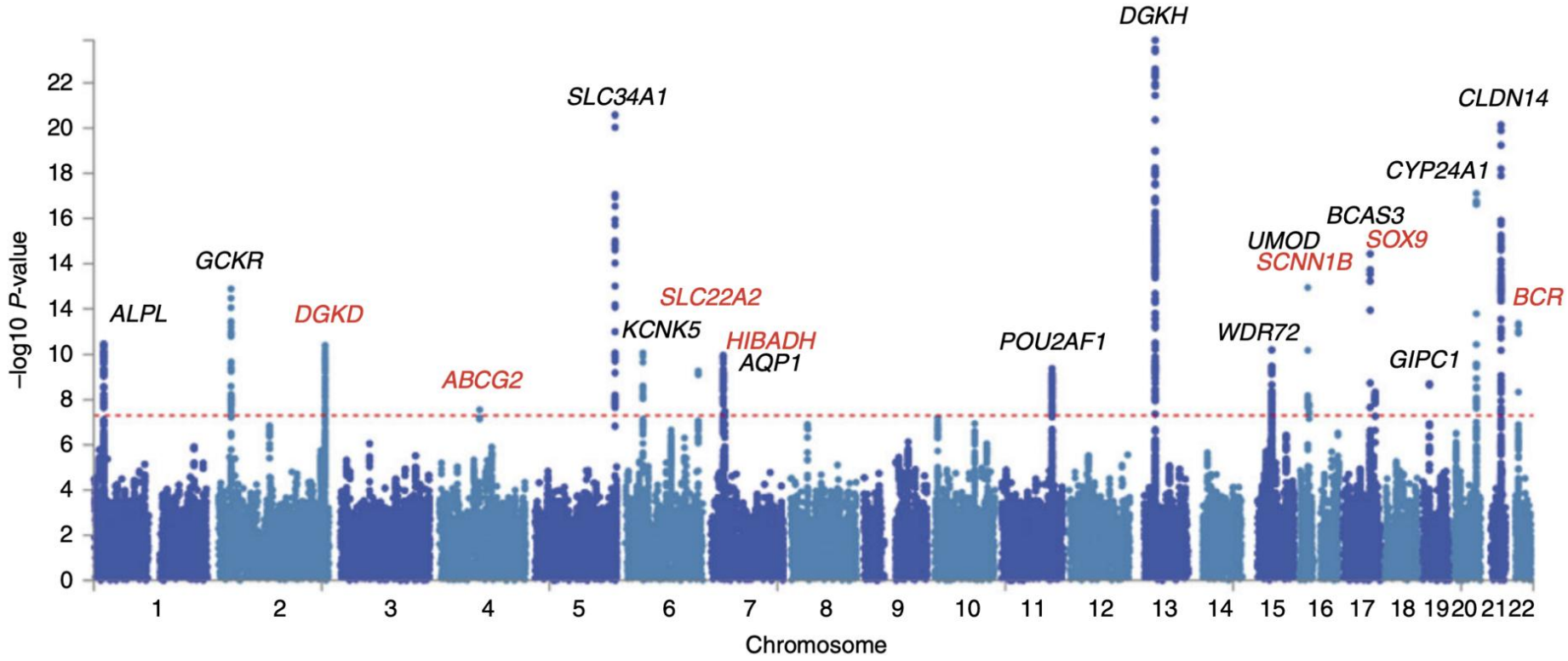
Abdel Abdellaoui

a.abdellaoui@amsterdamumc.nl

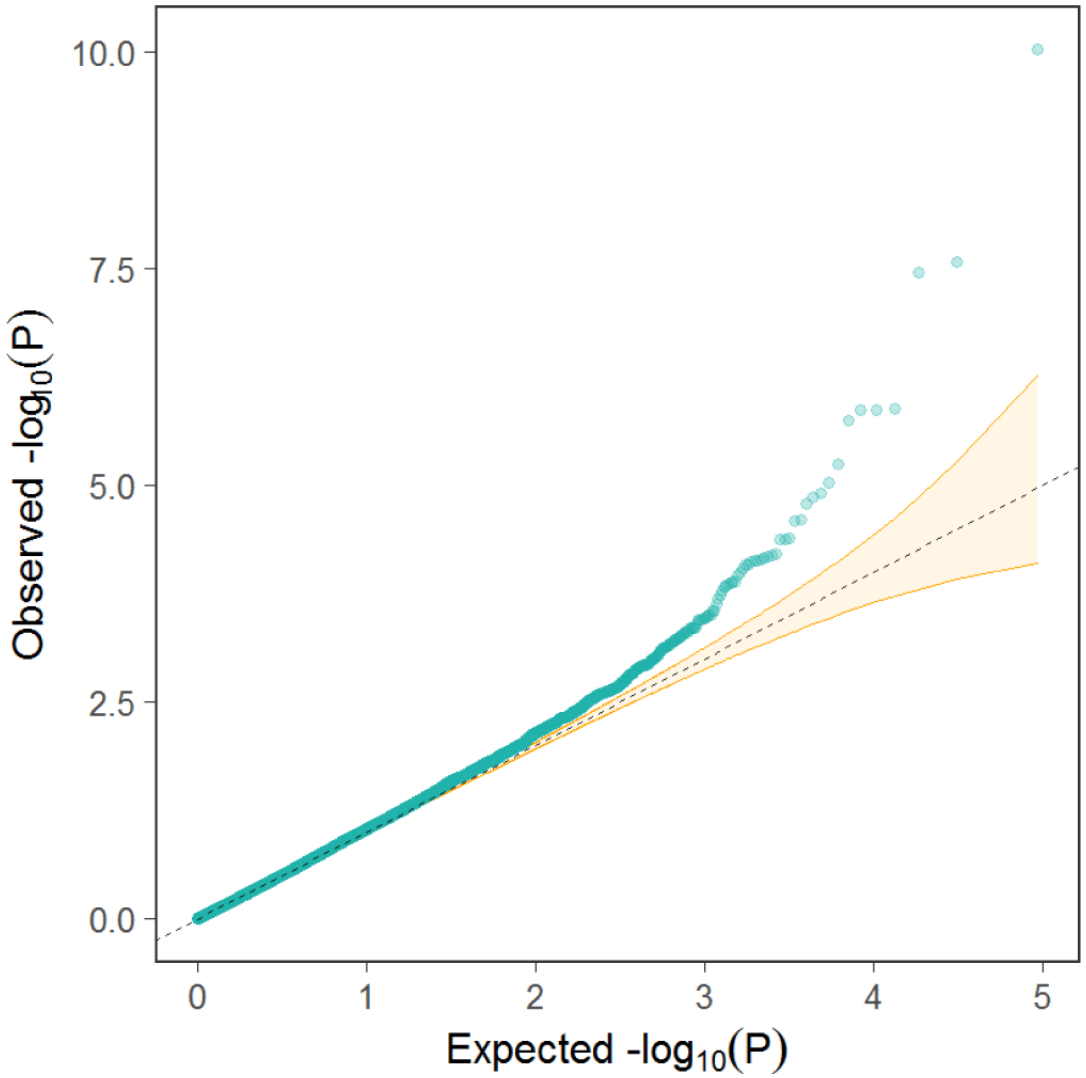

IN GENOME-WIDE ASSOCIATION STUDIES, OR GWAS, RESEARCHERS INTERROGATE SNPs THAT COMMONLY ARISE IN THE POPULATION... AND ASK WHETHER A VARIANT OF THAT SNP ASSOCIATES WITH DISEASE.


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

Manhattan Plot



QQ plot



Big part of doing a GWAS = separating **signal** from **noise**

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

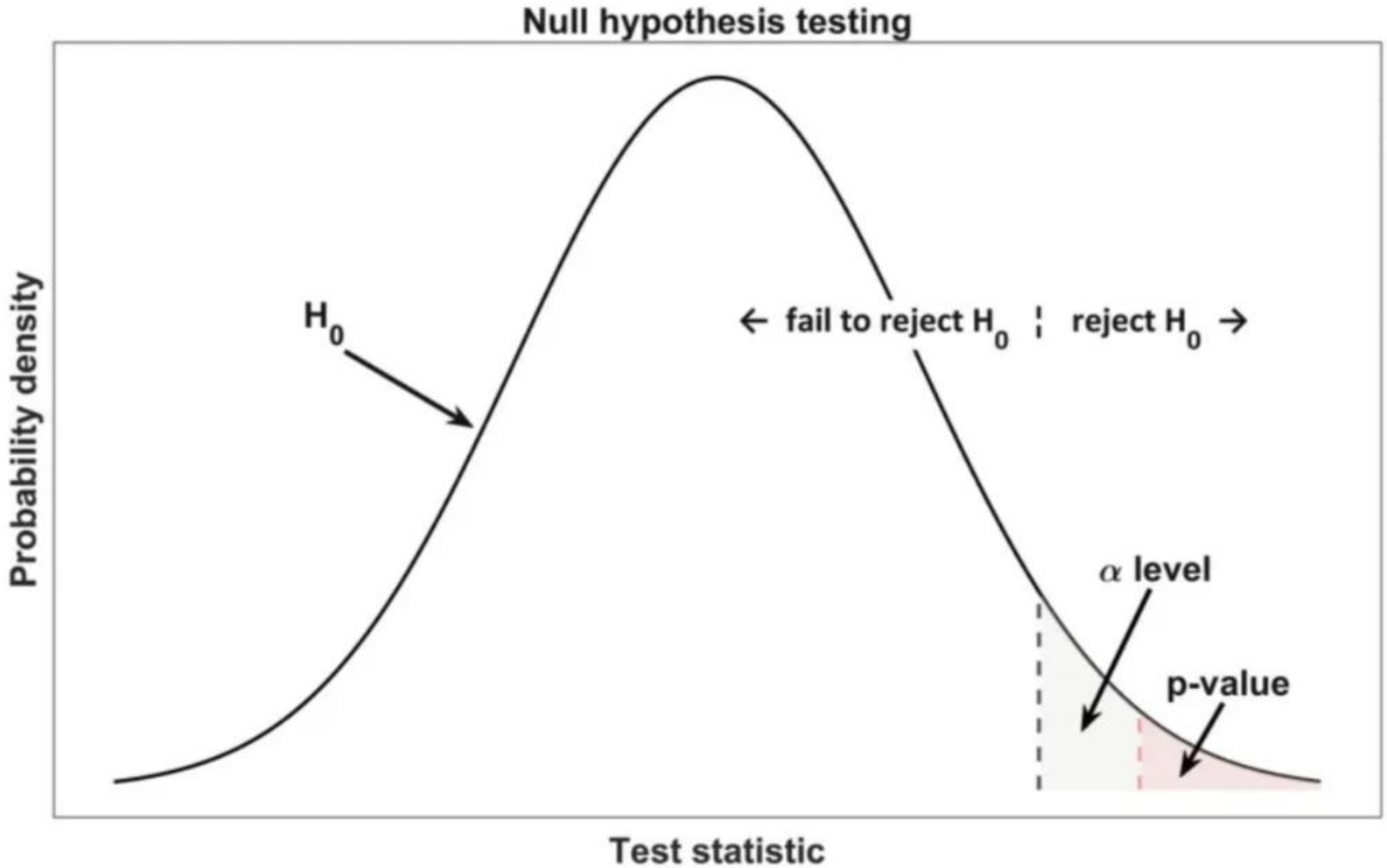
This video

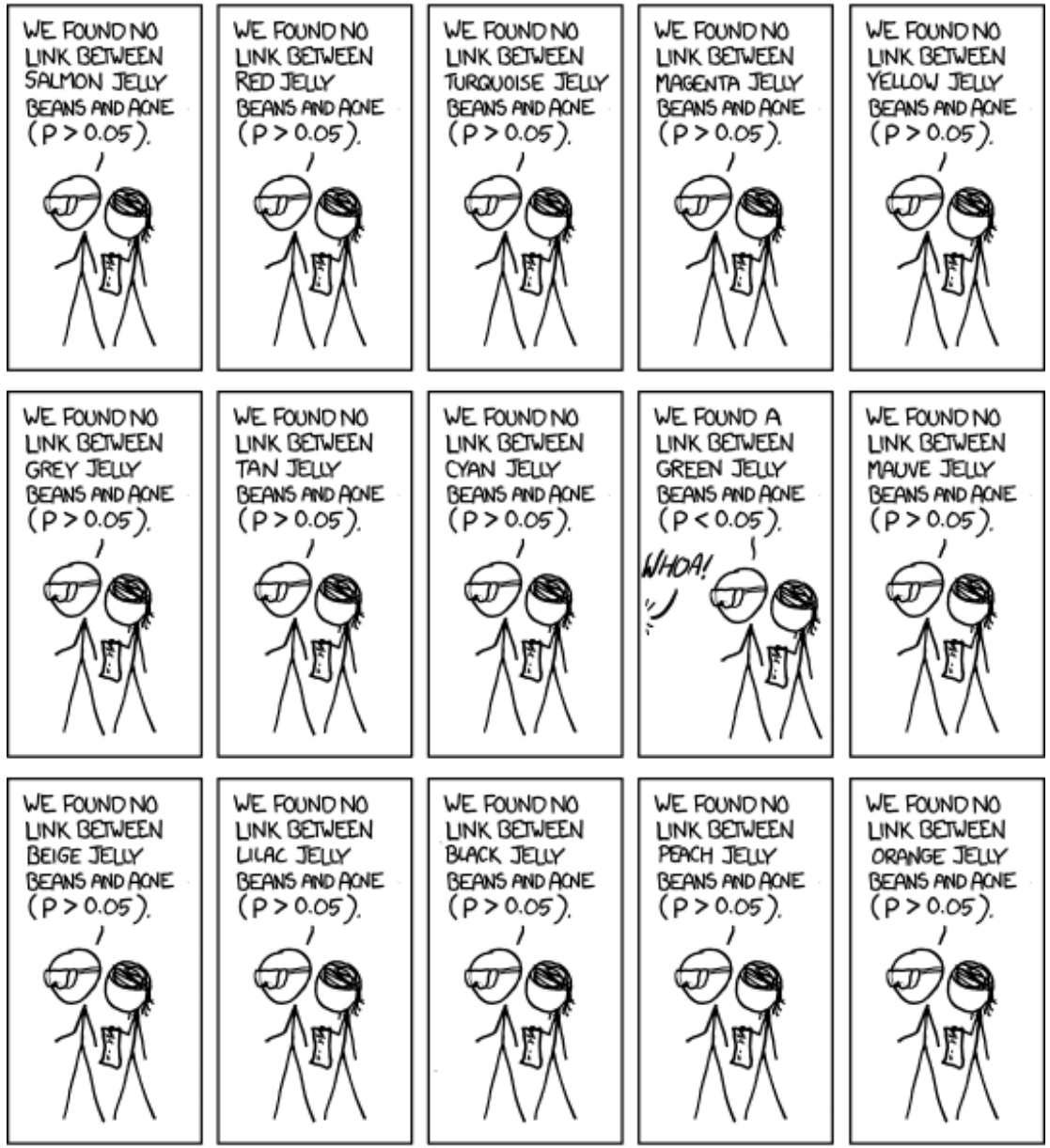
Real signal, but inflated or misunderstood:

- **Indirect genetic effects**: Parental genotype affects offspring via the environment
- **Gene-environment correlation**: Real genetic effects amplified by correlated environment
- **Assortative mating**: Creates genome-wide correlations that inflate polygenic signals

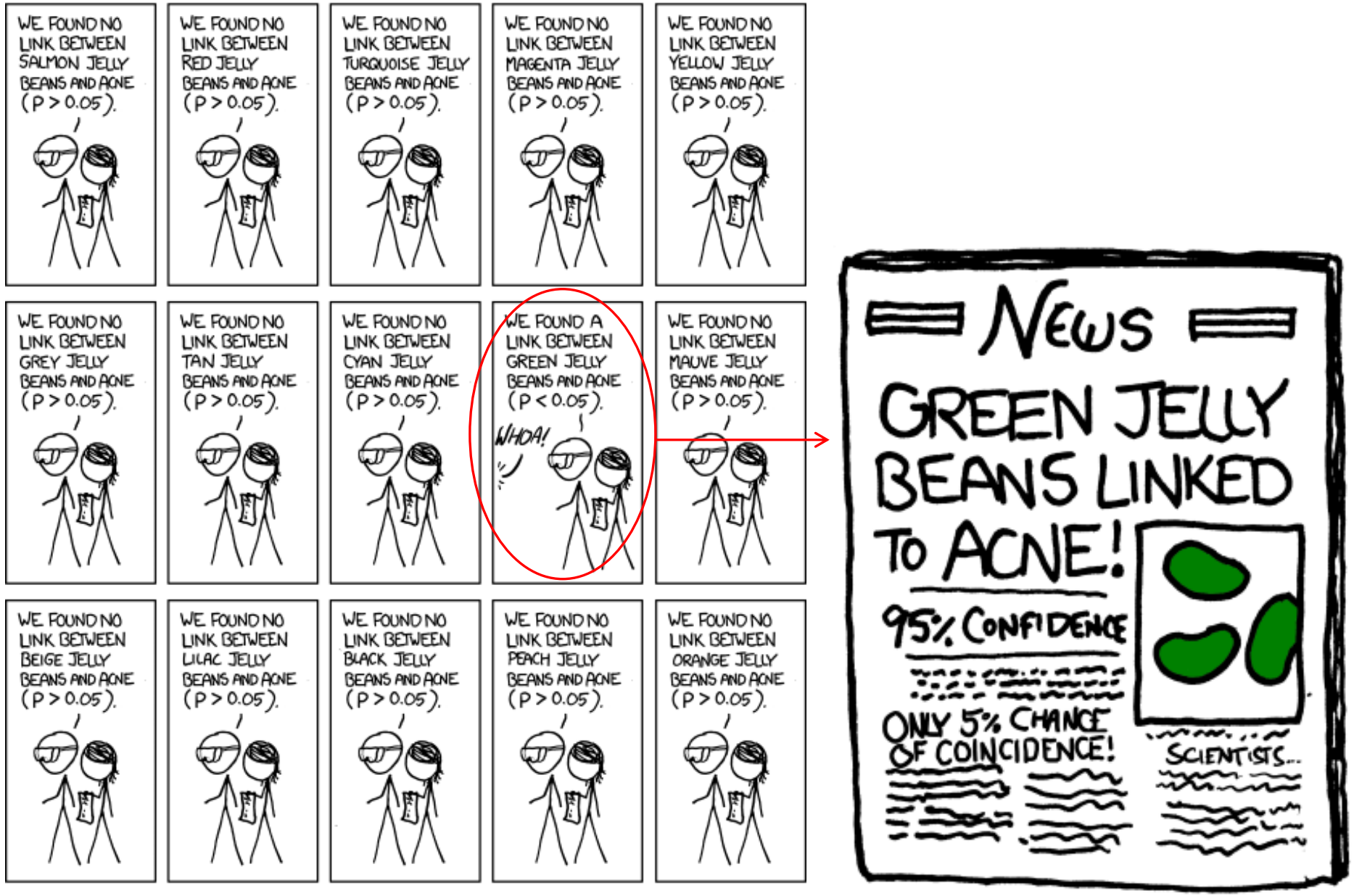
Will be discussed in Part 4

Multiple Testing

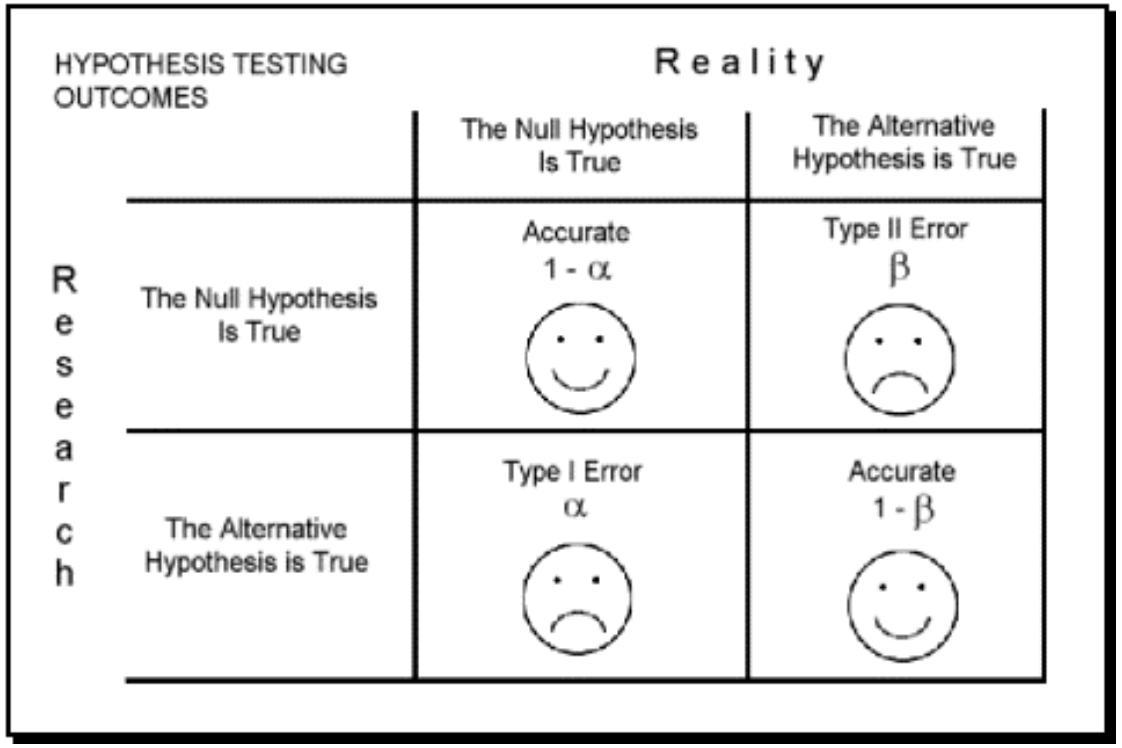

Null hypothesis testing
Probability density
H_0
← fail to reject H_0
reject H_0 →
α level
p-value
Test statistic



Do jelly beans cause acne?


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Multiple testing



With an α of .05, a GWAS of 1,000,000 SNPs is expected to give 50,000 (!) significant signals by chance.

Bonferroni correction: divide α by number of tests

Genetic Epidemiology 32: 227–234 (2008)

Estimation of Significance Thresholds for Genomewide Association Scans

Frank Dudbridge* and Arief Gusnanto

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

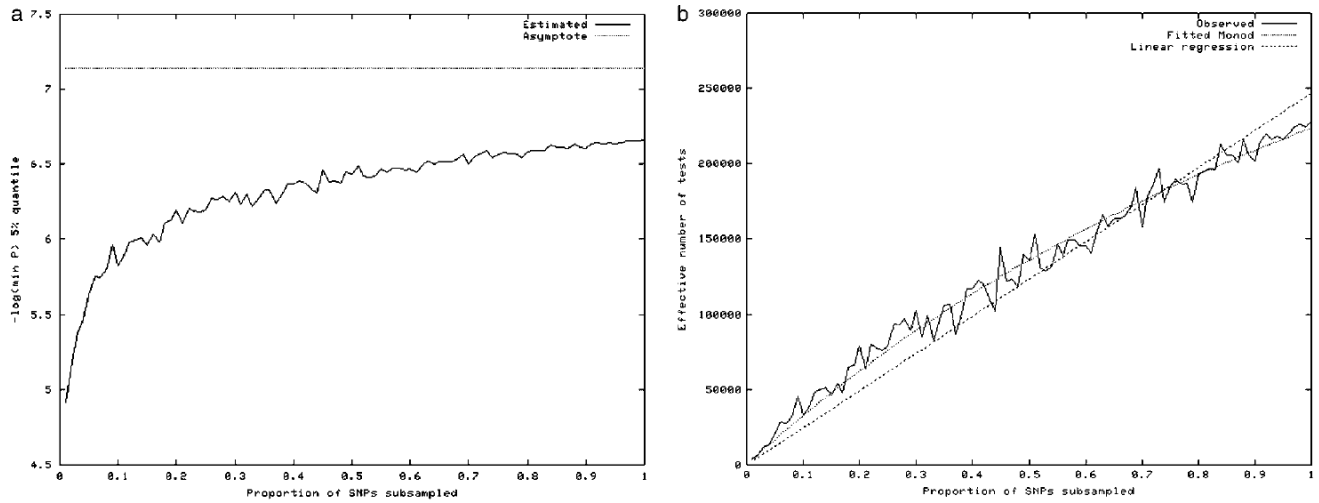


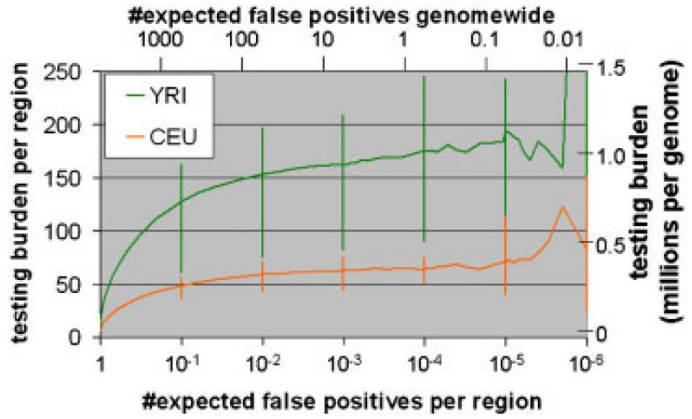
Fig. 1. (a) Significance threshold as a function of marker density in combined NBS and 58BC sample from permutation procedure. At current density (359K single nucleotide polymorphisms typed) the significance threshold is about 2.2×10^{-7}. The dotted line shows the estimated asymptote of 7.2×10^{-8}. (b) Fitted Monod function to the effective number of tests associated with the significance threshold. At infinite density the number of tests is estimated at 693,138 giving the asymptote in (a).

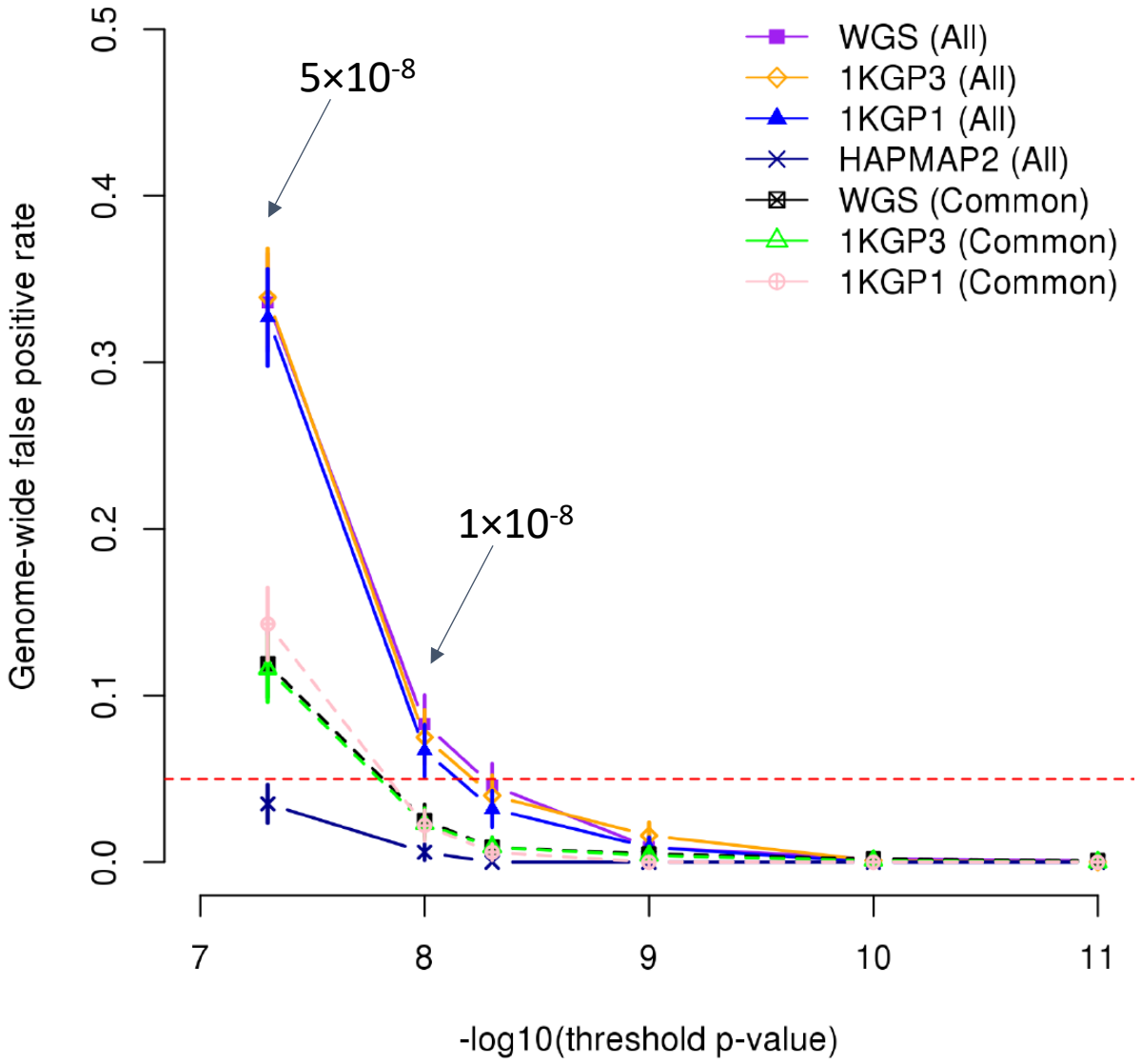
How many tests genome-wide?
With very common SNPs, ~1 million
→ $\alpha = .05/1M = 5 \times 10^{-8}$

Genetic Epidemiology 32: 381–385 (2008)

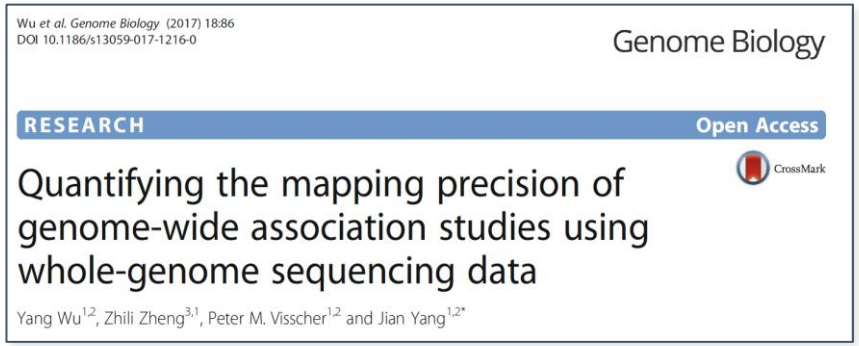
Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants

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





With more SNPs with lower allele frequencies, α should be 1×10^{-8}


Wu et al. Genome Biology (2017) 18:86
DOI 10.1186/s13059-017-1216-0

Genome Biology

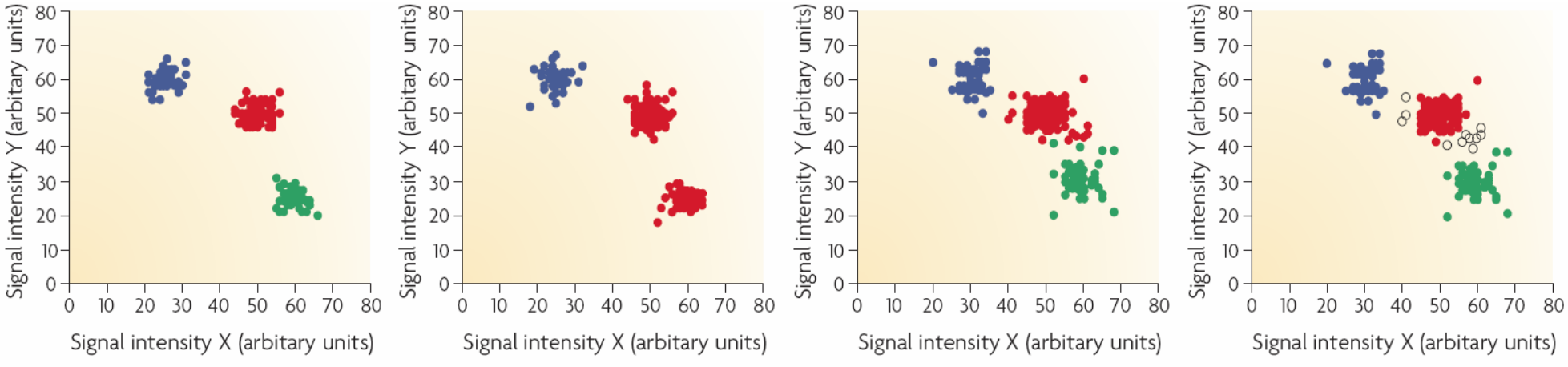
RESEARCH Open Access

Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data

Yang Wu[1,2], Zhili Zheng[3,1], Peter M. Visscher[1,2] and Jian Yang[1,2*]

Technical Artefacts

Genotyping errors



Intensity plots

Blue = AA

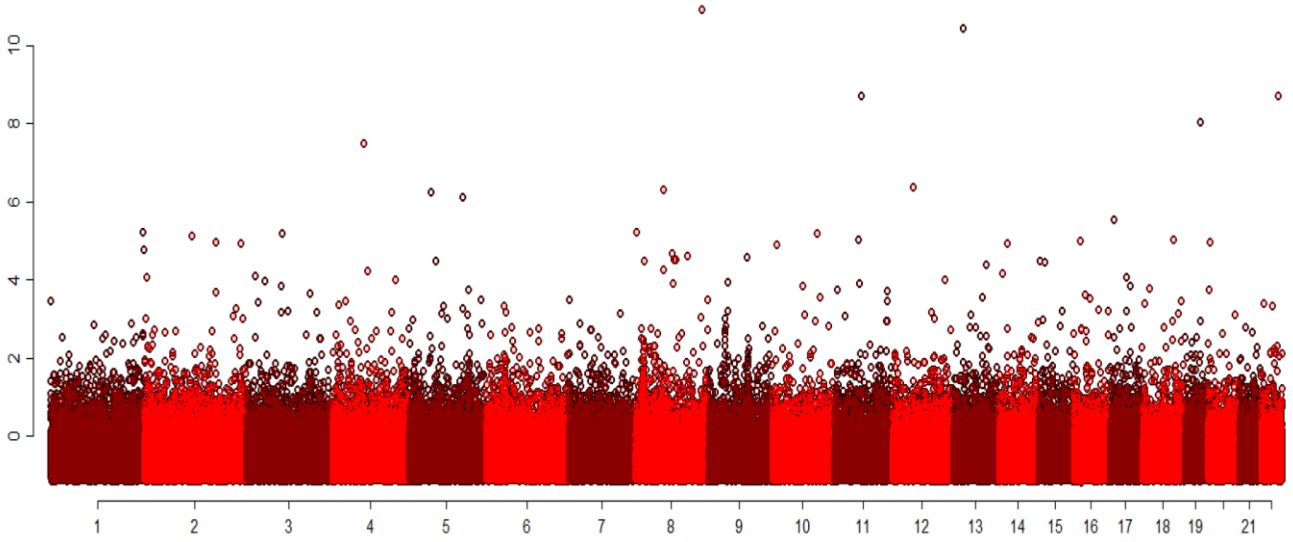
Red = AC

Green = CC

Batch Effects

Sciencexpress Report

Genetic Signatures of Exceptional Longevity in Humans



Editorial Expression of Concern

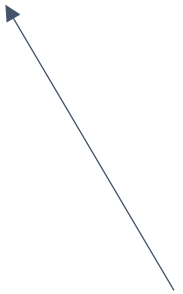
RESEARCHERS INVOLVED IN GENOME-WIDE association studies have expressed technical concerns about a Report by P. Sebastiani *et al.*, "Genetic signatures of exceptional longevity in humans," published in *Science* Express on 1 July 2010. In their study (*1*), Sebastiani *et al.* used a number of different genotyping platforms and neglected to perform data quality-control steps, which resulted in their reporting several false-positive single-nucleotide polymorphism (SNP) associations.

How to deal with technical artefacts?

Good study design and good QC

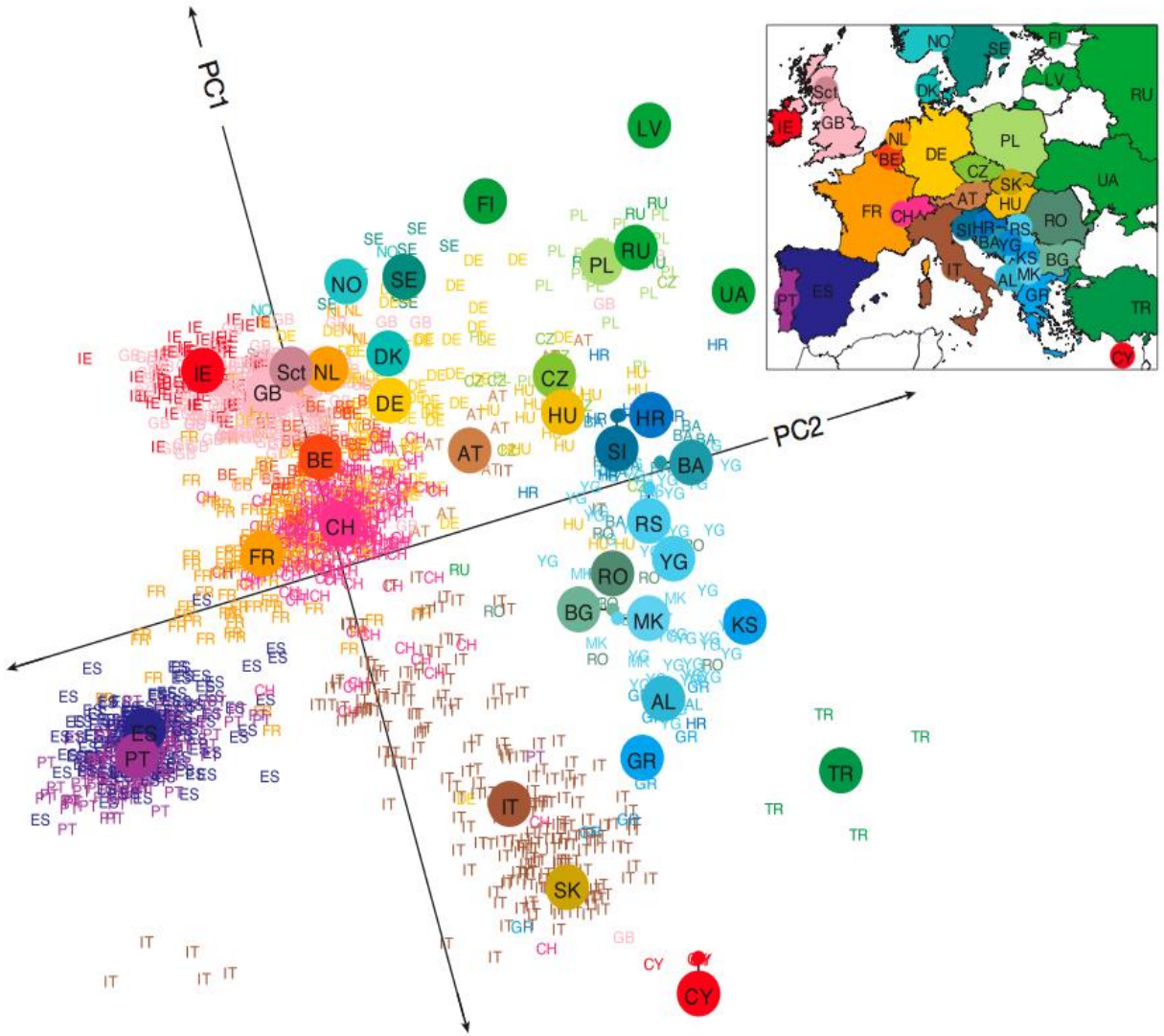
How to deal with technical artefacts?

- 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 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

Population Stratification



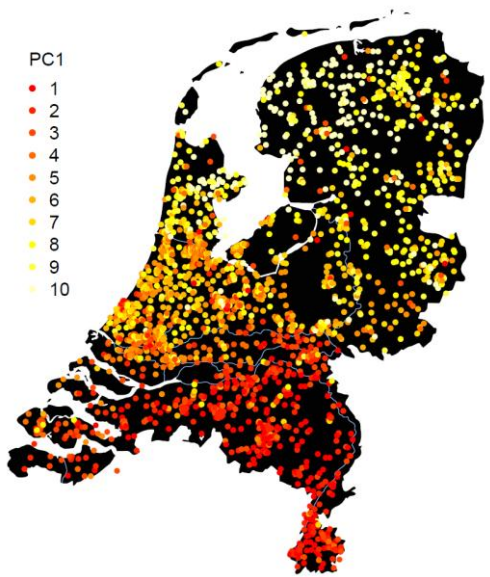
nature

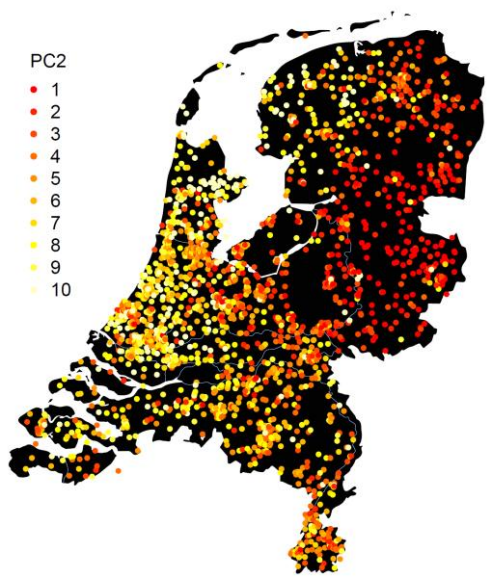
Vol 456|6 November 2008|doi:10.1038/nature07331

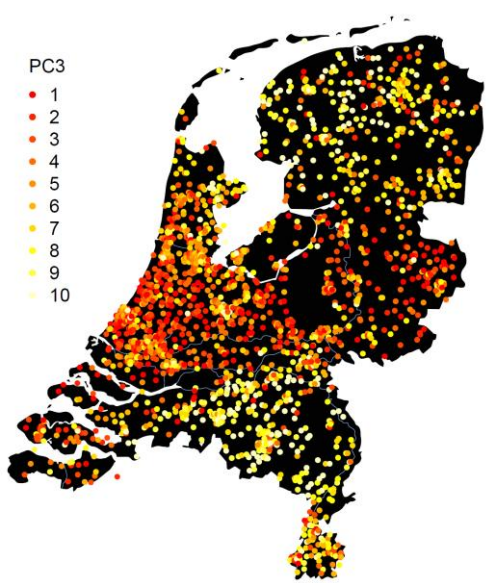
LETTERS

Genes mirror geography within Europe

John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zoltán Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]





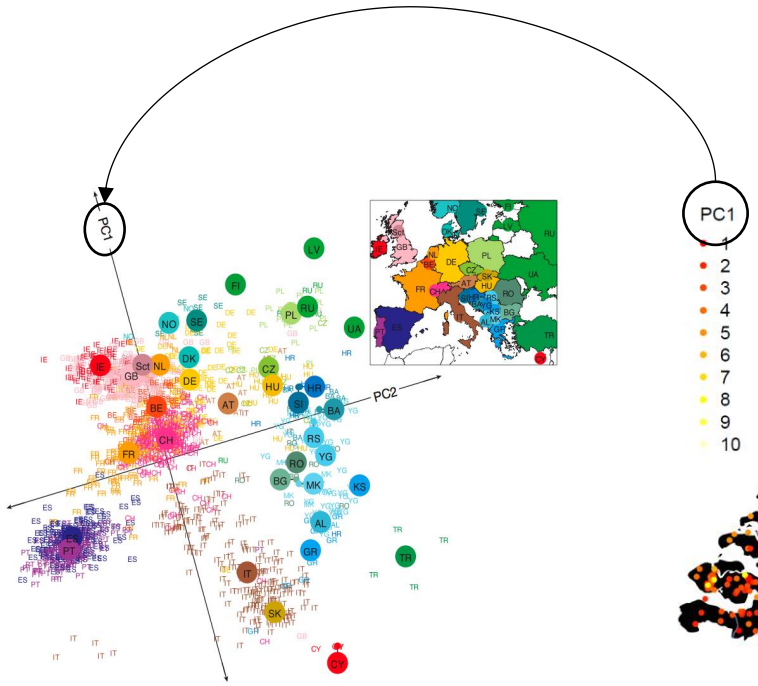


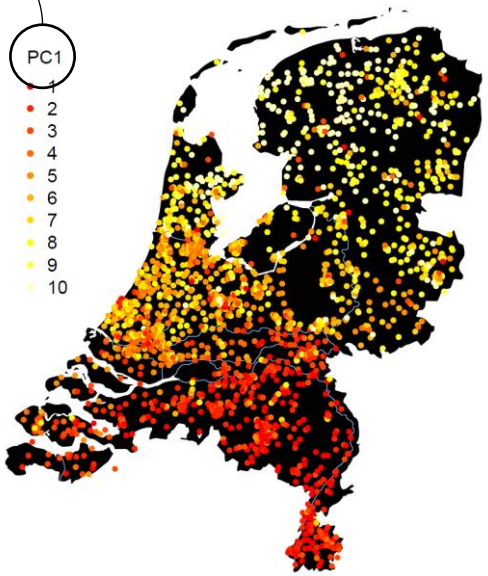
European Journal of Human Genetics (2013) 21, 1277–1285
© 2013 Macmillan Publishers Limited All rights reserved 1018-4813/13
www.nature.com/ejhg

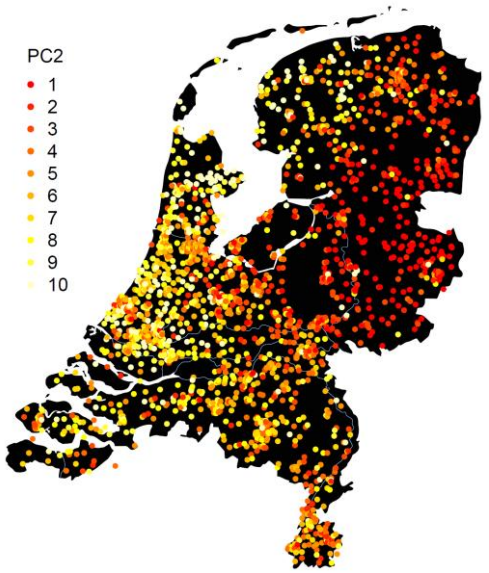
ARTICLE

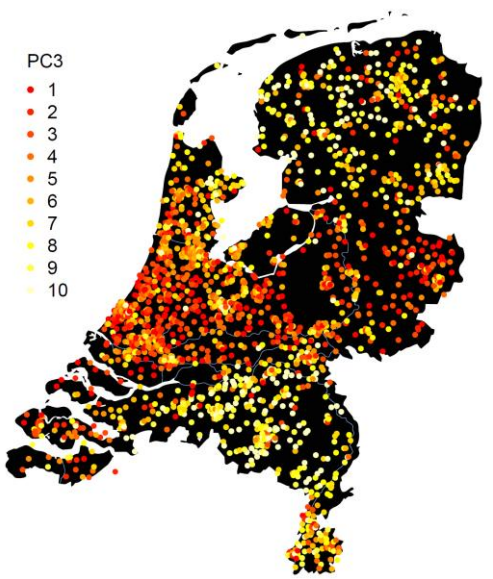
Population structure, migration, and diversifying selection in the Netherlands

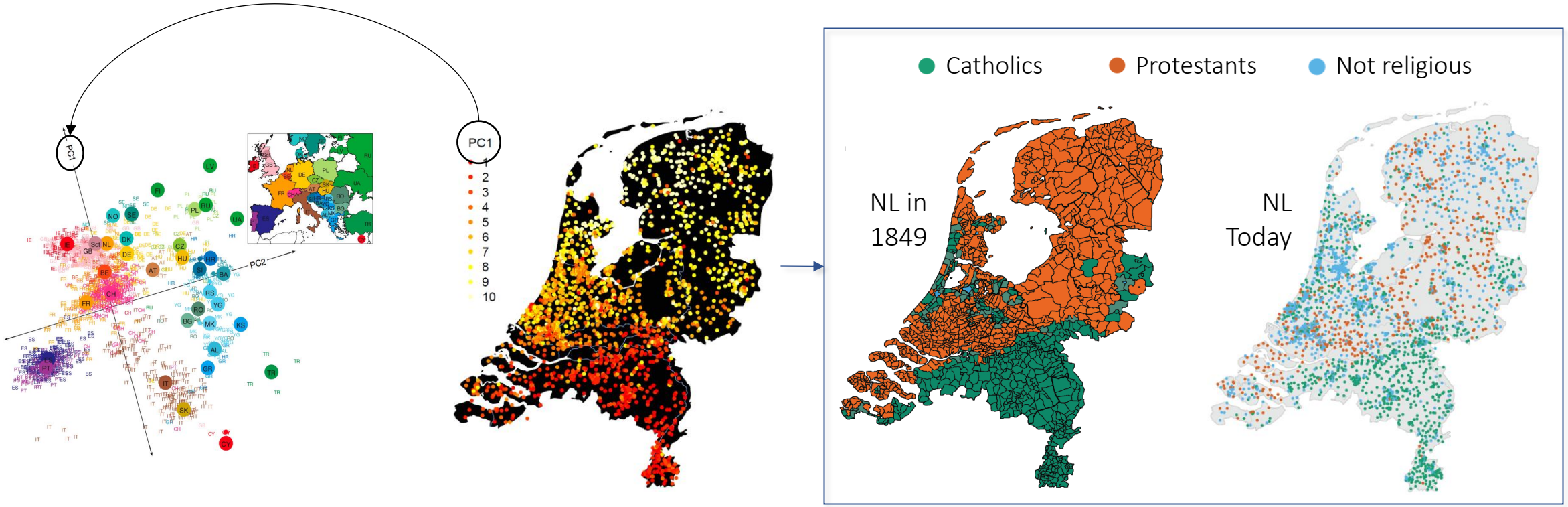
Abdel Abdellaoui*[,1], Jouke-Jan Hottenga[1], Peter de Knijff[2], Michel G Nivard[1], Xiangjun Xiao[3], Paul Scheet[3], Andrew Brooks[4], Erik A Ehli[5], Yueshan Hu[5], Gareth E Davies[5], James J Hudziak[6], Patrick F Sullivan[7], Toos van Beijsterveldt[1], Gonneke Willemsen[1], Eco J de Geus[1], Brenda WJH Penninx[8] and Dorret I Boomsma[1]


PC1
PC2


PC1
1
2
3
4
5
6
7
8
9
10


PC2
1
2
3
4
5
6
7
8
9
10


PC3
1
2
3
4
5
6
7
8
9
10


PC1
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Catholics
Protestants
Not religious
NL in 1849
NL Today

Cryptic Relatedness

Cryptic Relatedness

- Unmodelled family relationships can lead to attenuated standard error estimates of β, increasing number of false positives.
- Related individuals share environmental conditions that may be causal, thus estimated effect sizes may be biased.

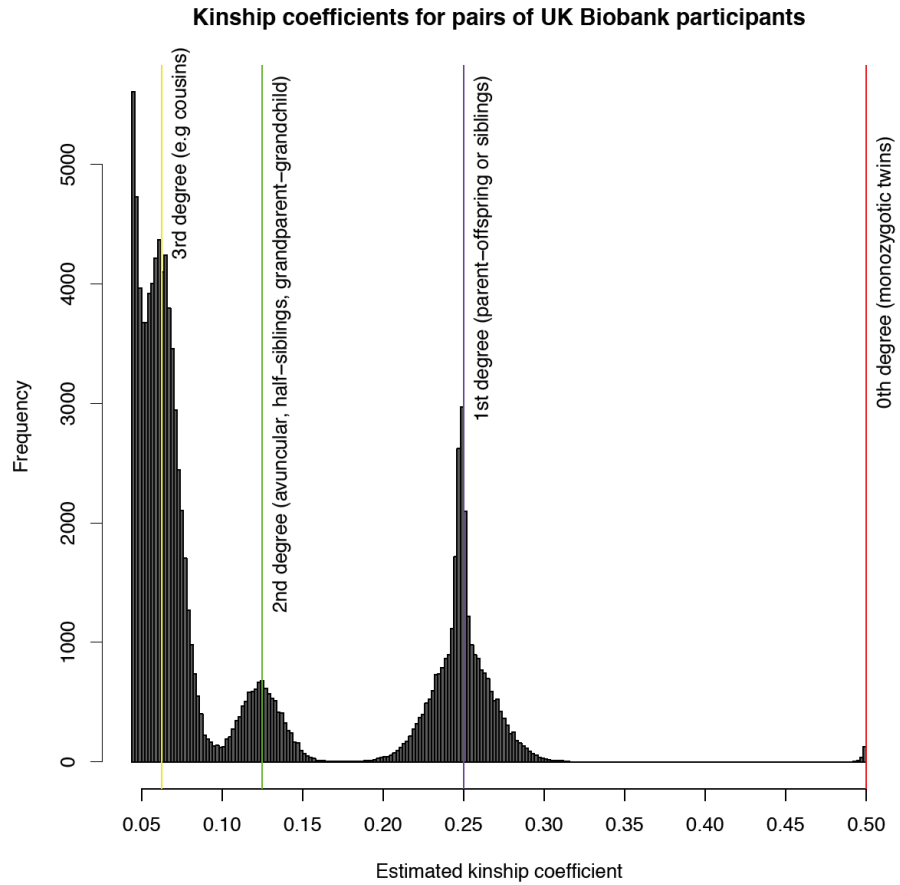

ARTICLE

OPEN

https://doi.org/10.1038/s41586-018-0579-z

The UK Biobank resource with deep phenotyping and genomic data

Clare Bycroft[1,13], Colin Freeman[1,13], Desislava Petkova[1,12,13], Gavin Band[1], Lloyd T. Elliott[2], Kevin Sharp[2], Allan Motyer[3], Damjan Vukcevic[3,4], Olivier Delaneau[5,6,7], Jared O'Connell[8], Adrian Cortes[1,9], Samantha Welsh[10], Alan Young[11], Mark Effingham[10], Gil McVean[1,11], Stephen Leslie[3,4], Naomi Allen[11], Peter Donnelly[1,2,14] & Jonathan Marchini[1,2,14]✉



Both cryptic relatedness and population stratification are confounds arising from kinship (**shared ancestry**) that is not accounted for

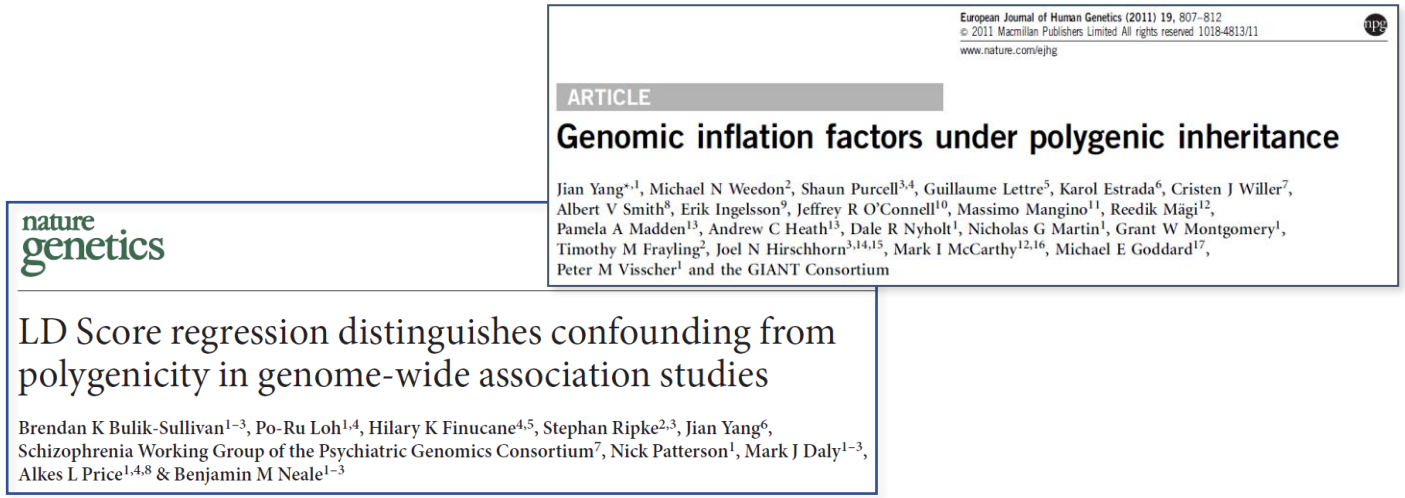
Dealing with population stratification & cryptic relatedness

- 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
© 2011 Macmillan Publishers Limited All rights reserved 1018-4813/11
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

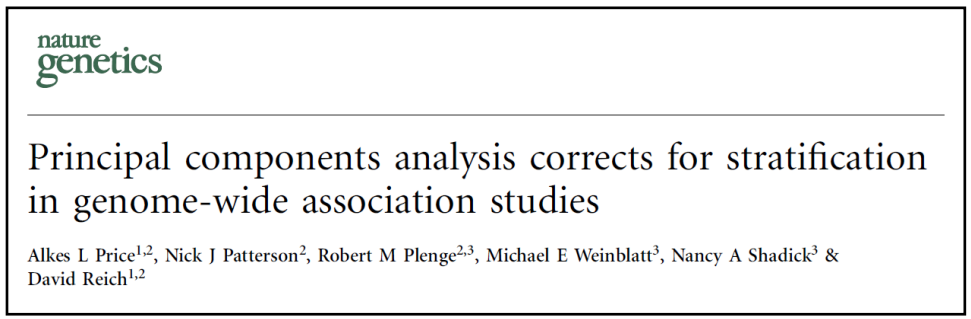
nature genetics

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]

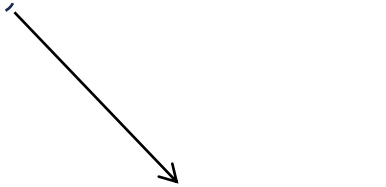
Dealing with population stratification & cryptic relatedness

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



Dealing with population stratification & cryptic relatedness

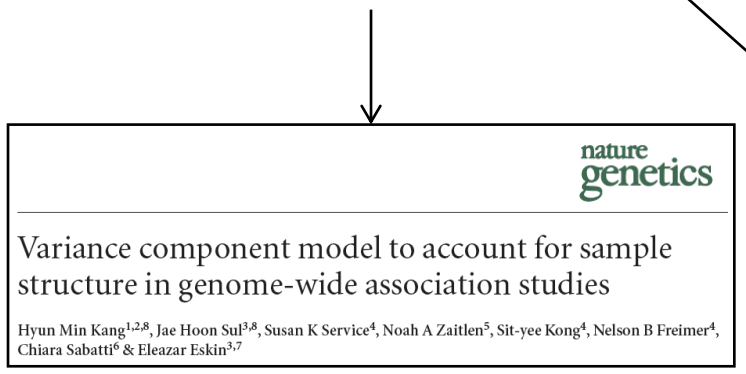
- Genomic Control (GC)
- Principal Component Analysis (does not deal with cryptic relatedness)
- Within Family Association
- Linear Mixed Modeling



Video 4

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

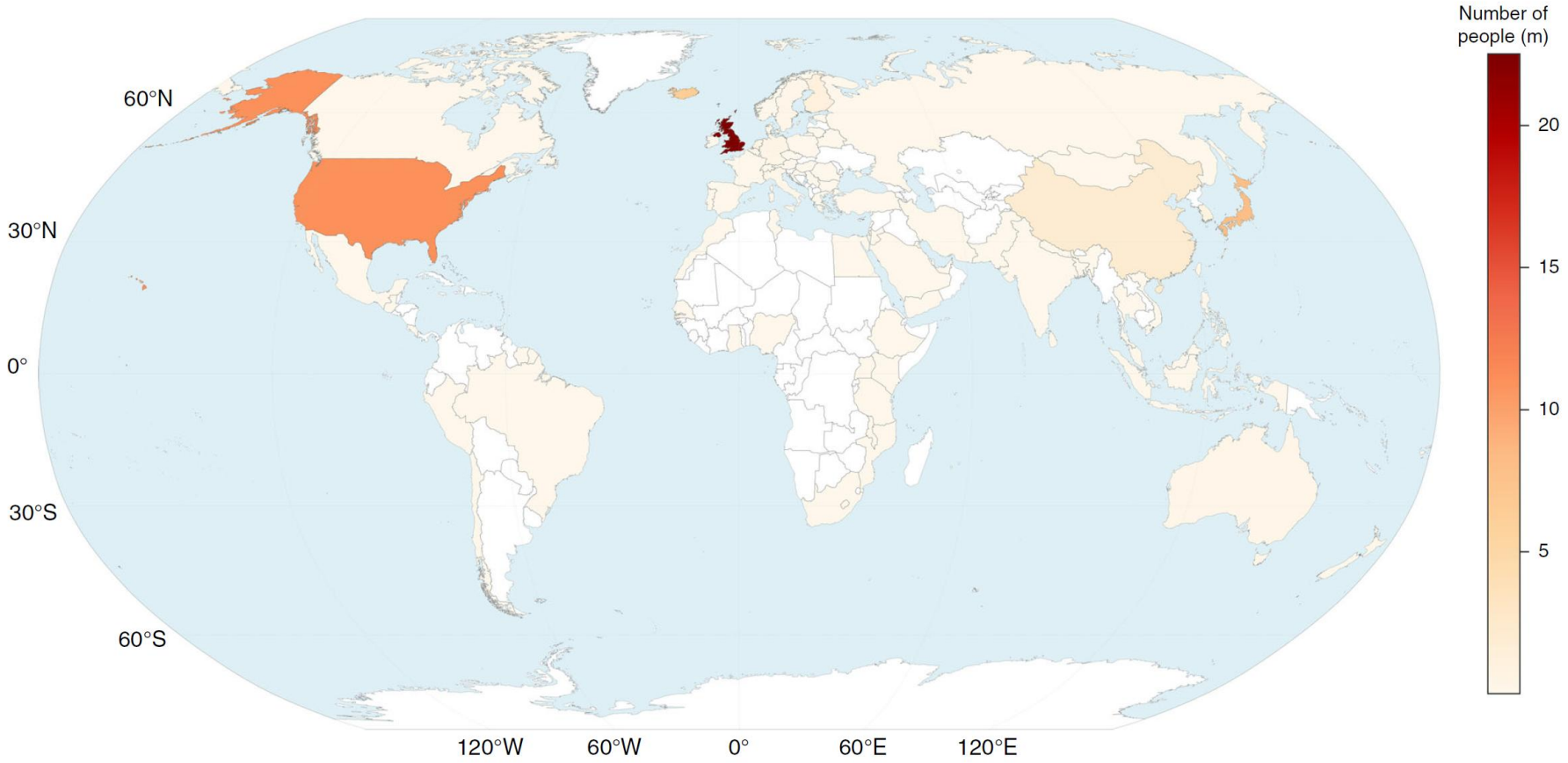
Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[4], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]


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]

Another way to minimize population stratification is to keep the sample homogeneous, which comes with another major problem:



COMMUNICATIONS BIOLOGY
REVIEW ARTICLE
https://doi.org/10.1038/s42003-018-0261-x OPEN

A scientometric review of genome-wide association studies

Melinda C. Mills[1] & Charles Rahal[1]

Next video: analyzing smaller cohorts, big biobanks, and cohorts combined