Genome-Wide Association Studies (GWAS): A practical workshop in R

Andrew Schork

Institute for Biological Psychiatry
Andrew.Joseph.Schork@regionh.dk

UNIVERSITY OF COPENHAGEN

nature neuroscience

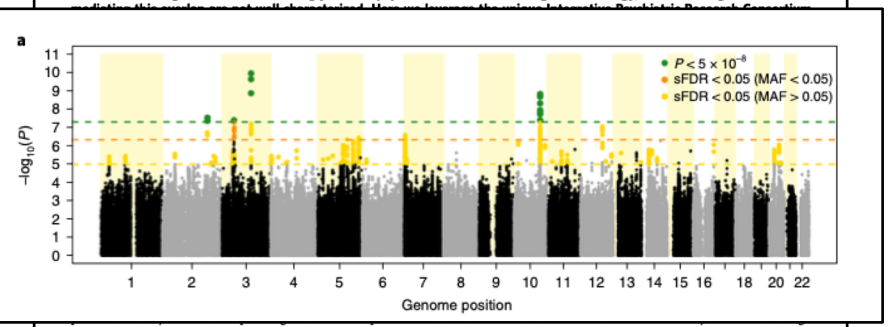
ARTICLES

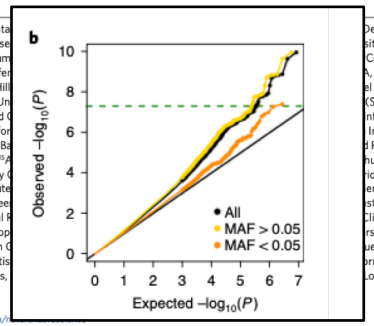
https://doi.org/10.1038/s41593-018-0320-0

A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment

Andrew J. Schork[1,2], Hyejung Won[3,4,5,6,7], Vivek Appadurai[1,2], Ron Nudel[1,2], Mike Gandal[3,4,5], Olivier Delaneau[8,9,10], Malene Revsbech Christiansen[11], David M. Hougaard[2,12], Marie Bækved-Hansen[2,12], Jonas Bybjerg-Grauholm[2,12], Marianne Giørtz Pedersen[2,13,14], Esben Agerbo[2,13,14], Carsten Bøcker Pedersen[2,13,14], Benjamin M. Neale[15,16,17], Mark J. Daly[15,16,17], Naomi R. Wray[18,19], Merete Nordentoft[2,20,21], Ole Mors[2,22], Anders D. Børglum[2,23,24], Preben Bo Mortensen[2,13,14,24], Alfonso Buil[1,2], Wesley K. Thompson[1,2,25], Daniel H. Geschwind[3,4,5,26] and Thomas Werge[1,2,21]*

There is mounting evidence that seemingly diverse psychiatric disorders share genetic etiology, but the biological substrates





[1]Institute of Biological Psychiatry, Menta... [2]The Lundbeck Foundation Initiative for Integrative Psychiatric Rese... sity of California, Los Angeles, Los Angeles, CA, USA. [4]Department of Hum... Center for Autism Research and Treatment, Semel Institute, David Geffer... A, USA. [6]Department of Genetics, University of North Carolina, Chapel Hill... el Hill, NC, USA. [8]Department of Genetic Medicine and Development, Un... (SIB), University of Geneva, Geneva, Switzerland. [10]Institute of Genetics and ... nformatics, Technical University of Denmark, Lyngby, Denmark. [12]Center for... Institut, Copenhagen, Denmark. [13]NCRR - National Centre for Register-Ba... d Register-based Research (CIRRAU), Aarhus University, Aarhus, Denmark. [15]A... husetts General Hospital and Harvard Medical School, Boston, USA. [16]Stanley ... idge, MA, USA. [19]Program in Medical and Population Genetics, Broad Institute... ence, University of Queensland, Brisbane, Queensland, Australia. [20]Queen... stralia. [22]Copenhagen Mental Health Center, Mental Health Services Capital R... Clinical Medicine, Faculty of Health and Medical Sciences, University of Cop... rsity Hospital, Risskov, Denmark. [23]Department of Biomedicine - Human G... uencing (iSEQ), Aarhus University, Aarhus, Denmark. [25]Division of Biostatis... rnia, San Diego, La Jolla, CA, USA. [26]Program in Neurobehavioral Genetics, ... Los Angeles, Los Angeles, CA, USA. *e-mail: Thomas.Werge@regionh.dk

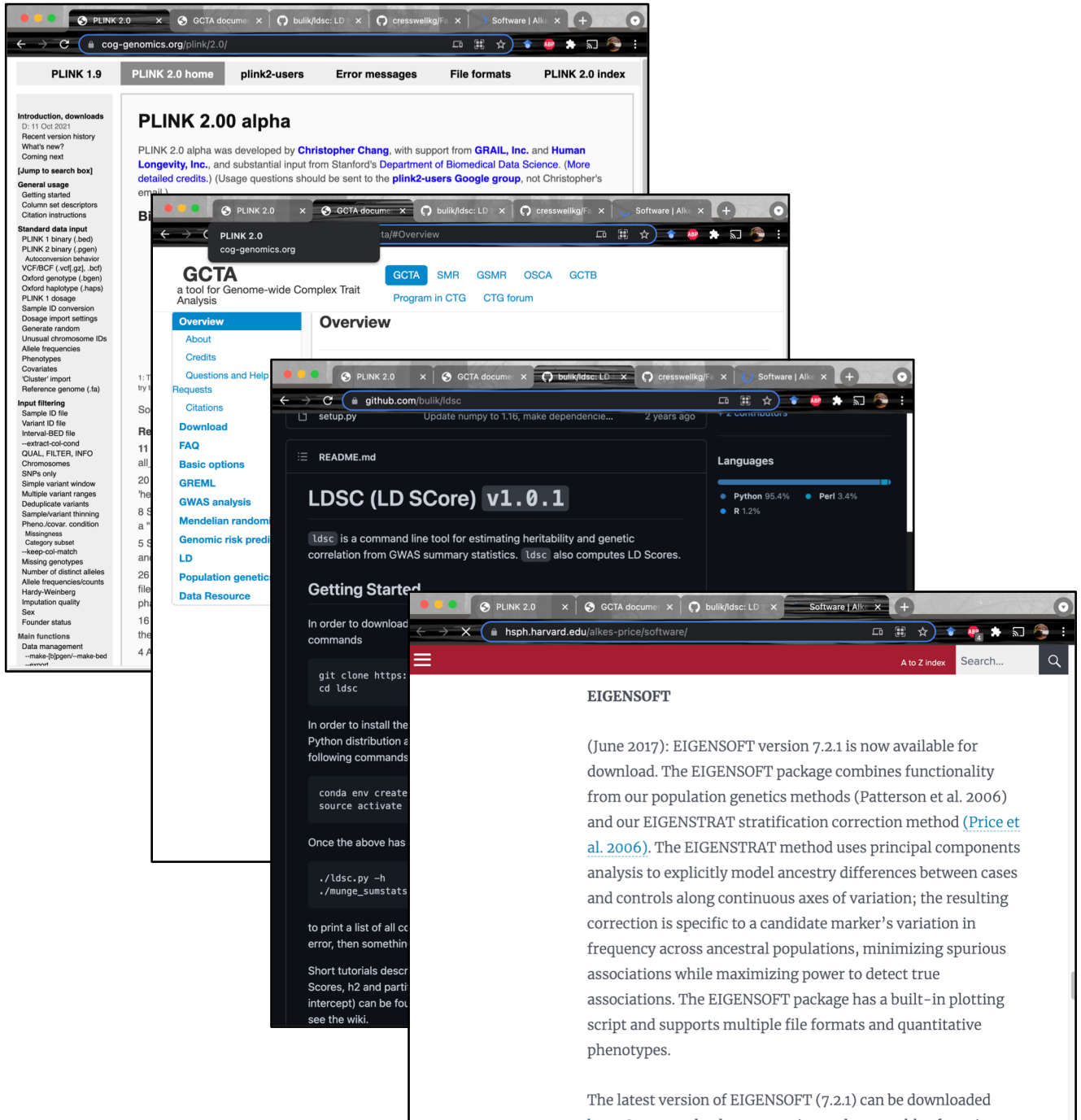
Initial genotyping was performed at the Broad Institute with amplified DNA extracted from dried blood spots and assayed on the Inifinium PsychChip v1.0 array[25]. In total, 78,050 subjects were successfully genotyped across 25 waves at approximately 550,000 SNPs. A subset of good-quality common SNPs ($n = 246{,}369$) were phased into haplotypes in a single batch using SHAPEIT3[59] and imputed in ten batches using Impute2 (ref. [60]) with reference haplotypes from the 1,000 genomes project phase 3 (ref. [61]). Imputed additive genotype dosages and best-guess genotypes were checked for imputation quality (INFO > 0.2), deviations from Hardy–Weinberg equilibrium (HWE; $P < 1 \times 10^{-6}$), association with genotyping wave ($P < 5 \times 10^{-8}$), association with imputation batch ($P < 5 \times 10^{-8}$; Supplementary Figs. 19–23) and differing imputation quality between subjects with and without psychiatric diagnoses ($P < 1 \times 10^{-6}$) as well as censored on MAF > 0.01). In total, 8,018,013 imputed dosages and best-guess genotypes were used for analysis.

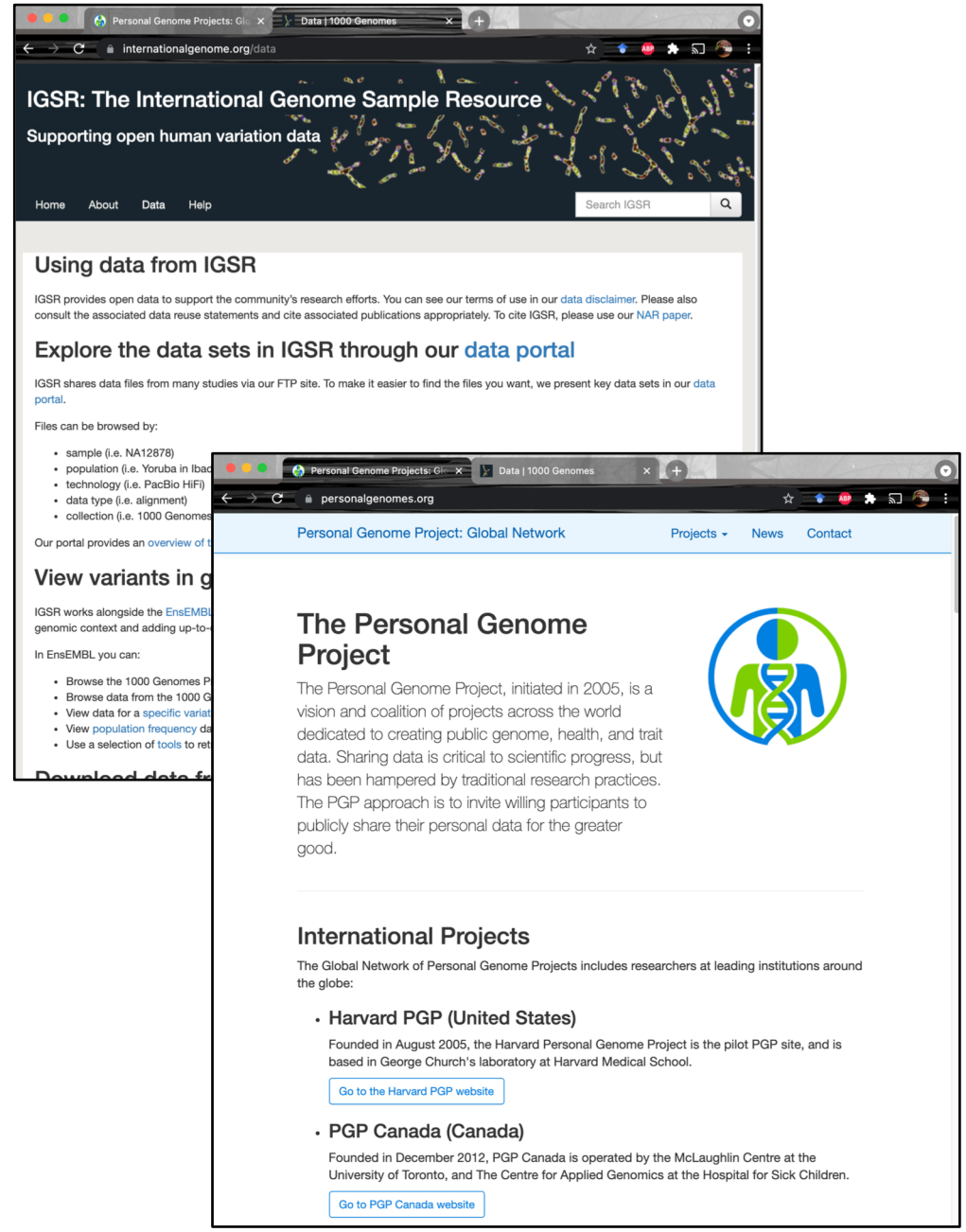
Three sets of cohorts of unrelated subjects with homogenous genetic ancestry were created by sub-setting the design cohorts, one for our primary GWAS analyses (GWAS cohorts) and two for heritability by either LDSC regression analysis (LDSC cohorts) or by GREML SNP heritability analysis (GCTA cohorts). Genetic ancestry for all cohorts was characterized using principal components analysis using smartPCA implemented in the Eigensoft package v6.0.1 (refs. [62,63]). We performed two iterations of censoring, removing subjects outlying from joint distribution of the first ten principle components defined in the subset of iPSYCH with four grandparents recorded in the Danish civil register as born in Denmark ($n = 6{,}474$ outliers removed; Supplementary Fig. 24), re-computing principle components on the remaining subjects and censoring again according to the same criteria ($n = 689$ outliers removed; Supplementary Fig. 25). Censored individuals were aggregated into a fourth ancestry diverse cohort (Replication cohort). For the GWAS and LDSC cohorts kinship was estimated using KING v1.9 (ref. [64]) and individuals were censored to ensure no pair had closer than third degree kinship ($n = 4{,}988$ removed). For the GCTA cohorts, kinship was more strictly filtered such that no pair had GCTA-based estimate greater than 0.034, the absolute value of the minimum estimated kinship ($n = 22{,}223$ removed). When possible cases were retained and the control relative was censored. All subject genotypes were flagged for abnormal sample heterozygosity, high levels of missing genotypes (>1%), sex concordance and inconsistencies among duplicate samples and those failing one or more tests were excluded ($n = 364$). In total, 65,534 subjects were retained in the GWAS and LDSC cohorts, 43,311 in the GCTA cohorts and 7,163 in the Replication cohort. A more detailed quality control protocol is available in our consortium white paper posted with our GWAS summary statistics (https://ipsych.au.dk/downloads/).

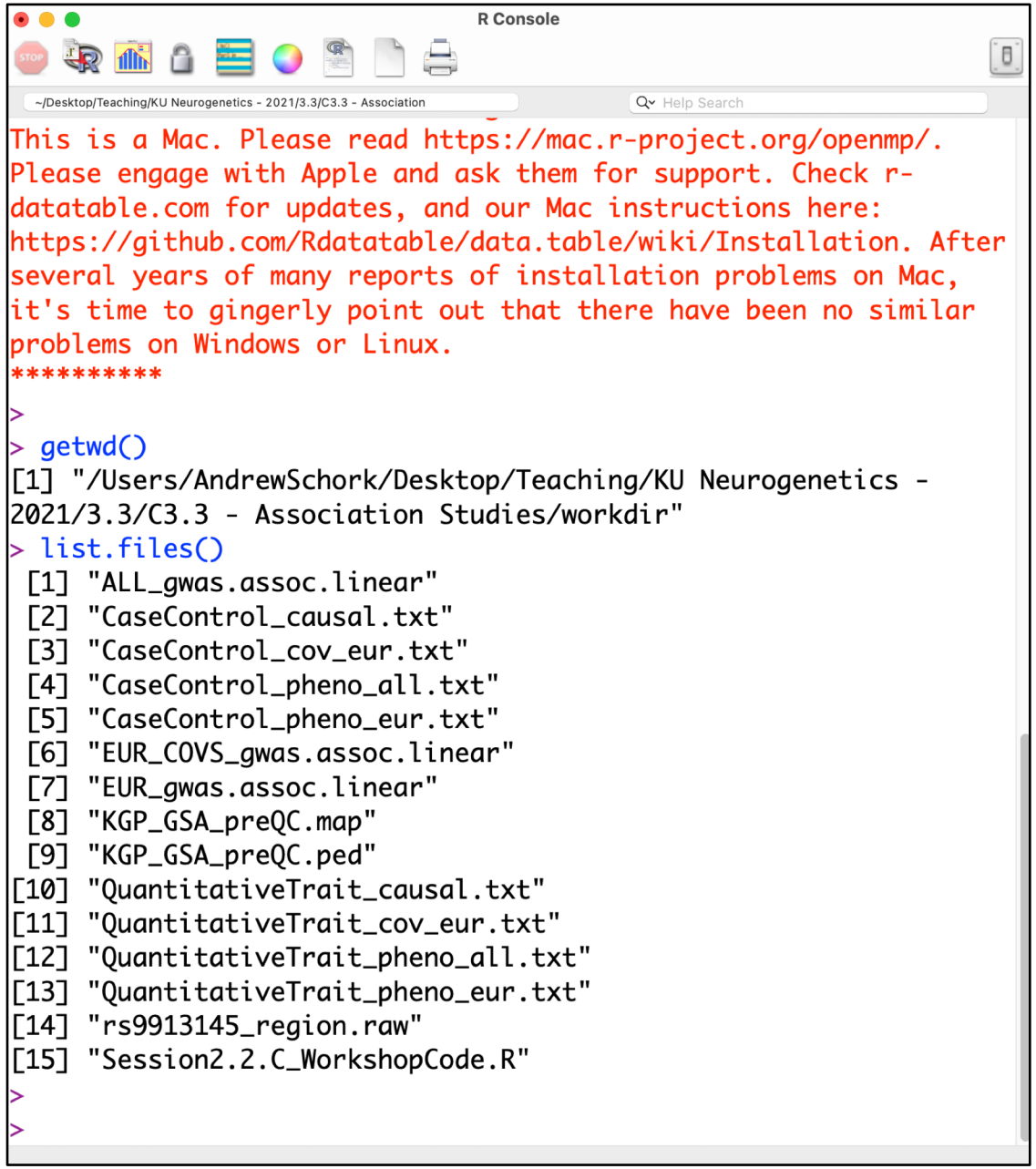
SNP heritability and genetic correlations. SNP heritability and genetic correlations were estimated in the GCTA cohorts with the GREML approach available in GCTA v1.25.2 (refs. [65–67]). Age, gender and ten principal components were included as fixed-effects covariates. Estimates were converted to the liability scale[67] according to estimates of lifetime risk take from Pedersen et al.[56] (Supplementary Table 2). Estimation of genetic correlation between indications was performed using bivariate GREML[66,68]. For each pair of phenotypes, subjects with both indications were excluded and controls were randomly and evenly split, creating two independent case-control groups. Splitting and estimation were repeated five times for each pair and the median values were retained.

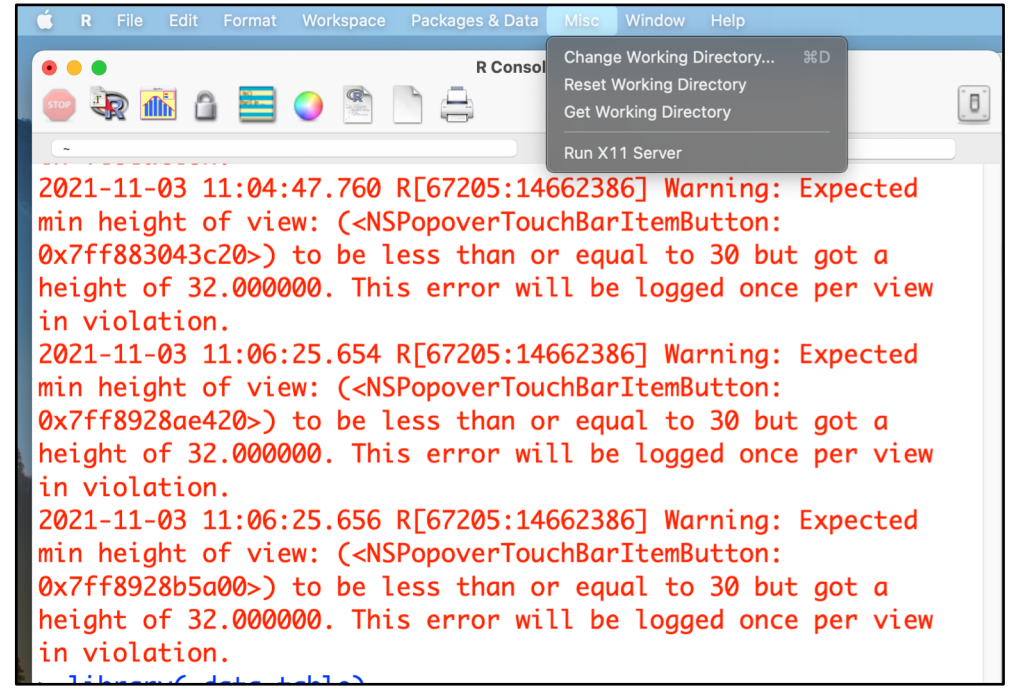
Published GCTA SNP heritability estimates for ADHD, AFF, ASD, BIP and SCZ were taken from Lee et al.[14]. GREML estimates of SNP for ANO and XDX were unavailable. GWAS statistics for eXDX[55], eADHD[69], eAFF[70], eANO[71], eASD[72] and eSCZ[73] were downloaded from the Psychiatric Genomics Consortium (PGC) repository (http://www.med.unc.edu/pgc/results-and-downloads). Statistics for eBIP (ref. [75]) were downloaded from the NHGRI-EBI GWAS catalog[75]. Linkage disequilibrium score regression (LDSC v1.0.0)[76] was used to estimate SNP heritability for these published studies and for each single iPSYCH indication. Reference LD scores and protocol were provided by the authors (https://github.com/bulik/ldsc/wiki). Genetic correlations between iPSYCH indications and published GWAS were estimated with LDSC[15] using the authors' protocols. For LDSC regression heritability and genetic correlation, single-indication iPSYCH GWAS were performed in the LDSC cohort according to the analysis approach described below. To facilitate comparisons a typical population prevalence was used for each liability scale transformation (XDX = 0.35, ADHD = 0.05, AFF = 0.15, ANO = 0.01, ASD = 0.01, BIP = 0.01, SCZ = 0.01)[14,71], including re-scaling the iPSYCH GREML estimates.

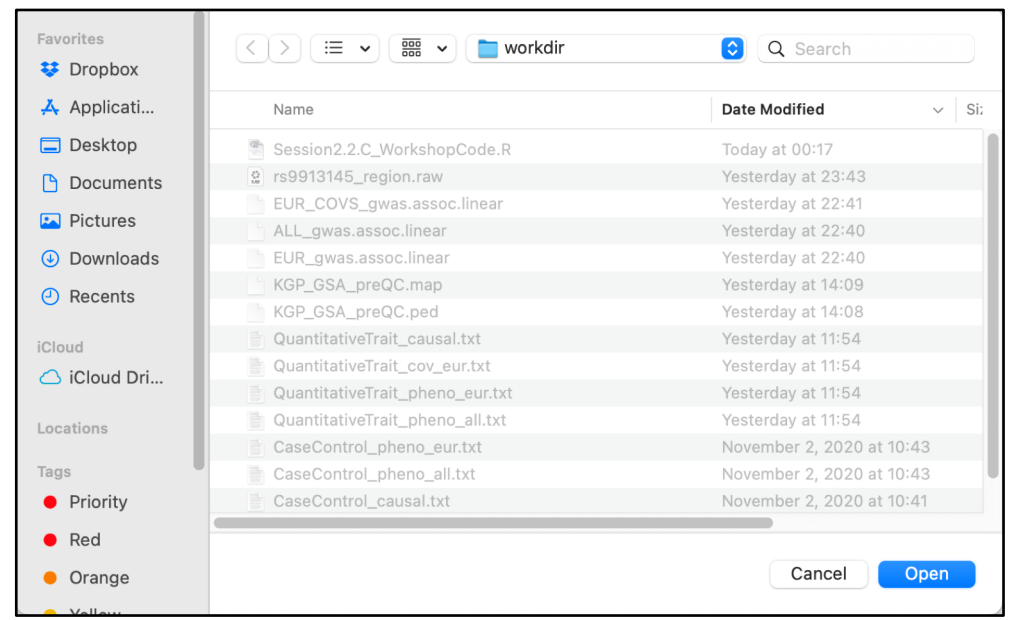
Association testing. GWAS were performed using imputed additive genotype dosages and logistic regression implemented in PLINK v1.90 (ref. [77]). The XDX GWAS included all subjects in the GWAS cohort (46,008 cases, 19,526 controls). Inflation was assed via genomic inflation factor (λ_{GC})[78] and LDSC[76]. Age, gender and ten principal components were included as fixed effects covariates. Stratified false discovery rates[79] (sFDR) were estimated according to Story's q value[80] and computed independently for common (MAF ≥ 0.05) and uncommon SNPs (0.01 < MAF < 0.05). The suggestive SNP threshold of sFDR q value < 0.05 corresponds to a P value less than 1.02×10^{-5} for common SNPs and less than 4.71×10^{-7} for uncommon SNPs. Single-indication odds ratios in Fig. 2c–f and XDX GWAS excluding each single indication used in Fig. 4 were performed to provide context for the XDX results in the GWAS cohort. For the internal replication cohort (7,163 individuals, 4,481 cases), association tests used best-guess genotypes and linear mixed models implemented in GCTA[66] accounting for relatedness and heterogeneity in genetic background with genome-wide estimates of empirical kinship. Gender and age were included as fixed-effects covariates.

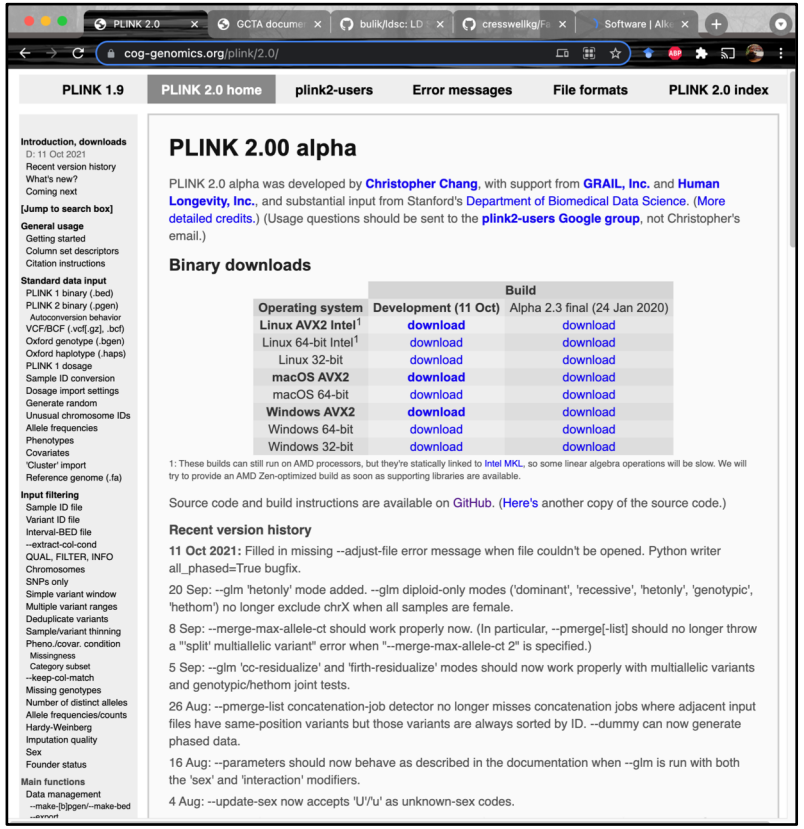

PLINK 2.0
cog-genomics.org/plink/2.0/
PLINK 1.9
PLINK 2.0 home
plink2-users
Error messages
File formats
PLINK 2.0 index
PLINK 2.00 alpha
PLINK 2.0 alpha was developed by Christopher Chang, with support from GRAIL, Inc. and Human Longevity, Inc., and substantial input from Stanford's Department of Biomedical Data Science. (More detailed credits.) (Usage questions should be sent to the plink2-users Google group, not Christopher's
GCTA
a tool for Genome-wide Complex Trait Analysis
GCTA
SMR
GSMR
OSCA
GCTB
Program in CTG
CTG forum
Overview
About
Credits
Questions and Help
Requests
Citations
Download
FAQ
Basic options
GREML
GWAS analysis
Mendelian randomi
Genomic risk predi
LD
Population genetic
Data Resource
github.com/bulik/ldsc
README.md
LDSC (LD SCore) v1.0.1
ldsc is a command line tool for estimating heritability and genetic correlation from GWAS summary statistics. ldsc also computes LD Scores.
Getting Started
Languages
Python 95.4%
Perl 3.4%
R 1.2%
hsph.harvard.edu/alkes-price/software/
A to Z index
Search...
EIGENSOFT
(June 2017): EIGENSOFT version 7.2.1 is now available for download. The EIGENSOFT package combines functionality from our population genetics methods (Patterson et al. 2006) and our EIGENSTRAT stratification correction method (Price et al. 2006). The EIGENSTRAT method uses principal components analysis to explicitly model ancestry differences between cases and controls along continuous axes of variation; the resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. The EIGENSOFT package has a built-in plotting script and supports multiple file formats and quantitative phenotypes.
The latest version of EIGENSOFT (7.2.1) can be downloaded


internationalgenome.org/data
IGSR: The International Genome Sample Resource
Supporting open human variation data
Home
About
Data
Help
Search IGSR
Using data from IGSR
IGSR provides open data to support the community's research efforts. You can see our terms of use in our data disclaimer. Please also consult the associated data reuse statements and cite associated publications appropriately. To cite IGSR, please use our NAR paper.
Explore the data sets in IGSR through our data portal
IGSR shares data files from many studies via our FTP site. To make it easier to find the files you want, we present key data sets in our data portal.
Files can be browsed by:
sample (i.e. NA12878)
population (i.e. Yoruba in Ibad
technology (i.e. PacBio HiFi)
data type (i.e. alignment)
collection (i.e. 1000 Genomes
Our portal provides an overview of t
View variants in g
IGSR works alongside the EnsEMBL
genomic context and adding up-to-
In EnsEMBL you can:
Browse the 1000 Genomes P
Browse data from the 1000 G
View data for a specific variat
View population frequency da
Use a selection of tools to ret
personalgenomes.org
Personal Genome Project: Global Network
Projects
News
Contact
The Personal Genome Project
The Personal Genome Project, initiated in 2005, is a vision and coalition of projects across the world dedicated to creating public genome, health, and trait data. Sharing data is critical to scientific progress, but has been hampered by traditional research practices. The PGP approach is to invite willing participants to publicly share their personal data for the greater good.
International Projects
The Global Network of Personal Genome Projects includes researchers at leading institutions around the globe:
Harvard PGP (United States)
Founded in August 2005, the Harvard Personal Genome Project is the pilot PGP site, and is based in George Church's laboratory at Harvard Medical School.
Go to the Harvard PGP website
PGP Canada (Canada)
Founded in December 2012, PGP Canada is operated by the McLaughlin Centre at the University of Toronto, and The Centre for Applied Genomics at the Hospital for Sick Children.
Go to PGP Canada website


R Console
~/Desktop/Teaching/KU Neurogenetics - 2021/3.3/C3.3 - Association
Help Search
This is a Mac. Please read https://mac.r-project.org/openmp/.
Please engage with Apple and ask them for support. Check r-
datatable.com for updates, and our Mac instructions here:
https://github.com/Rdatatable/data.table/wiki/Installation. After
several years of many reports of installation problems on Mac,
it's time to gingerly point out that there have been no similar
problems on Windows or Linux.
****
> getwd()
[1] "/Users/AndrewSchork/Desktop/Teaching/KU Neurogenetics -
2021/3.3/C3.3 - Association Studies/workdir"
> list.files()
[1] "ALL_gwas.assoc.linear"
[2] "CaseControl_causal.txt"
[3] "CaseControl_cov_eur.txt"
[4] "CaseControl_pheno_all.txt"
[5] "CaseControl_pheno_eur.txt"
[6] "EUR_COVS_gwas.assoc.linear"
[7] "EUR_gwas.assoc.linear"
[8] "KGP_GSA_preQC.map"
[9] "KGP_GSA_preQC.ped"
[10] "QuantitativeTrait_causal.txt"
[11] "QuantitativeTrait_cov_eur.txt"
[12] "QuantitativeTrait_pheno_all.txt"
[13] "QuantitativeTrait_pheno_eur.txt"
[14] "rs9913145_region.raw"
[15] "Session2.2.C_WorkshopCode.R"


R File Edit Format Workspace Packages & Data Misc Window Help
Change Working Directory... ⌘D
Reset Working Directory
Get Working Directory
Run X11 Server
2021-11-03 11:04:47.760 R[67205:14662386] Warning: Expected
min height of view: (<NSPopoverTouchBarItemButton:
0x7ff883043c20>) to be less than or equal to 30 but got a
height of 32.000000. This error will be logged once per view
in violation.
2021-11-03 11:06:25.654 R[67205:14662386] Warning: Expected
min height of view: (<NSPopoverTouchBarItemButton:
0x7ff8928ae420>) to be less than or equal to 30 but got a
height of 32.000000. This error will be logged once per view
in violation.
2021-11-03 11:06:25.656 R[67205:14662386] Warning: Expected
min height of view: (<NSPopoverTouchBarItemButton:
0x7ff8928b5a00>) to be less than or equal to 30 but got a
height of 32.000000. This error will be logged once per view
in violation.


Favorites
Dropbox
Applicati...
Desktop
Documents
Pictures
Downloads
Recents
iCloud
iCloud Dri...
Locations
Tags
Priority
Red
Orange
workdir
Search
Name
Date Modified
Session2.2.C_WorkshopCode.R Today at 00:17
rs9913145_region.raw Yesterday at 23:43
EUR_COVS_gwas.assoc.linear Yesterday at 22:41
ALL_gwas.assoc.linear Yesterday at 22:40
EUR_gwas.assoc.linear Yesterday at 22:40
KGP_GSA_preQC.map Yesterday at 14:09
KGP_GSA_preQC.ped Yesterday at 14:08
QuantitativeTrait_causal.txt Yesterday at 11:54
QuantitativeTrait_cov_eur.txt Yesterday at 11:54
QuantitativeTrait_pheno_eur.txt Yesterday at 11:54
QuantitativeTrait_pheno_all.txt Yesterday at 11:54
CaseControl_pheno_eur.txt November 2, 2020 at 10:43
CaseControl_pheno_all.txt November 2, 2020 at 10:43
CaseControl_causal.txt November 2, 2020 at 10:41
Cancel
Open



Open source tool for analyzing and manipulating large genotyping data sets

Map file: Describes each SNP

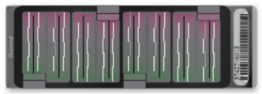
CHR	SNP	cM	BP
1	rs707582	0	5671786
1	rs74053039	0	7543000
1	rs523125	0	15430933
1	rs2270978	0	18023365
1	rs11801111	0	20060965
1	rs74060020	0	21935358
1	rs3932664	0	24511663
1	rs4653052	0	34992037
1	rs10493098	0	42261889

Ped file: Describes each Individual

FID	IID	MID	PID	Sex	Trait	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
EUR	HG00096	0	0	1	-9	A A	C C	A A	T T	A A	C C	T T	A A	C T
EUR	HG00097	0	0	2	-9	G A	C C	A A	C T	A A	C C	G T	A A	T T
EUR	HG00099	0	0	2	-9	G A	T C	G A	T T	A A	T C	G T	A A	C C
EUR	HG00100	0	0	2	-9	G A	C C	G A	C T	A A	C C	T T	A A	T T
EUR	HG00101	0	0	1	-9	G A	C C	G A	T T	A A	T C	T T	A A	C T
EUR	HG00102	0	0	2	-9	A A	C C	G A	T T	A A	T C	G T	A A	C T
EUR	HG00103	0	0	1	-9	G A	T C	A A	C T	A A	C C	T T	C C	C C
EUR	HG00105	0	0	1	-9	G A	C C	G G	T T	A A	T C	T T	A A	C T
EUR	HG00106	0	0	2	-9	A A	C C	G G	C T	G A	C C	T T	C A	C T
EUR	HG00107	0	0	1	-9	G G	C C	G A	C T	A A	C C	T T	A A	T T

The Technology: Genotyping

- Uses a microarray to measure a predefined set of SNPs
- We can measure chosen SNPs at 500,000 to 5,000,000 positions
- Depends on maps of known variation
- Relies on *Linkage Disequilibirum*



Microarray

ATCGAAATGCATGACCTTTGATATGATCGGCTGCAGTCAGC
TTCGAAGTGCATGACTTTTGACATGAGCGGCGGCCACAGC

Common Variation Rare Mutations No Recorded Variation

The Technology: Whole Genome Sequencing

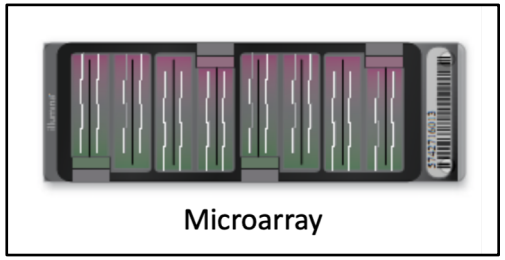
- Directly measure every base pair in the genome (3,200,000,000 x 2) and thus every possible SNP
- Currently it is relatively slow and expensive, but that is changing
- Can discover new "rare" mutations

Sequencer

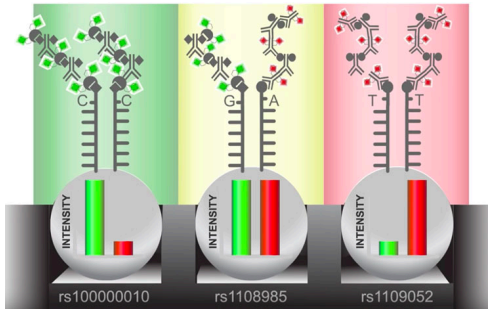
ATCGAAATGCATGACCTTTGATATGATCGGCTGCAGTCAGC
TTCGAAGTGCATGACTTTTGACATGAGCGGCGGCCACAGC

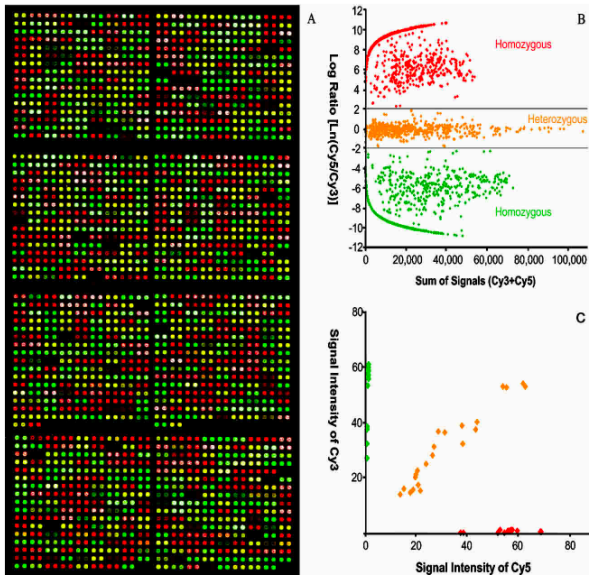
Common Variation Mutations No Recorded Variation

Allele Frequency Bin	Concordance between genotyping array and Exome Sequencing Data	Number of SNPs
0.00001 - 0.0001	0.4085	20701
0.0001 - 0.001	0.7976	30367
0.001 - 0.01	0.9676	14145
0.01 - 0.1	0.9966	6795
0.1 - 0.5	0.999	5081
0.5 - 1	0.9991	28



Microarray





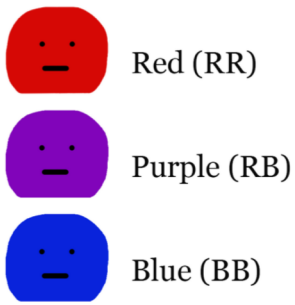
In order for a population to be in Hardy-Weinberg Equilibrium, the following conditions must be met:

1) Organism is diploid (2 copies of chromosomes)
2) Only sexual reproduction occurs (no selfing)
3) Mating is random (no assortment on genotype)
4) Population is infinite
5) No sex linked genotypes
6) No migration,
7) No admixture
8) No mutation
9) No selection
10) No inbreeding

How to use chi-squared to test for Hardy-Weinberg equilibrium

Updated: Apr 29

This post demonstrates the use of chi-squared to test for Hardy-Weinberg equilibrium. There is a question on a recent (February 2020) AP Biology practice test that required this calculation. The question is a secure item, so the exact question **will not** be discussed here. There is a previous post on this blog explaining how to test for evolution using the null hypothesis and chi-squared.



Phenotypes and genotypes for examples

For our examples, we'll use the fictional species featured in many of the evolution simulations. The population demonstrates incomplete dominance for color. There are two alleles; red and blue. Heterozygotes have a purple phenotype.

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

Chi-squared equation

Chi-squared is a statistical test used to determine if observed data (o) is equivalent to expected data (e). A population is at Hardy-Weinberg equilibrium for a gene if five conditions are met; random mating, no mutation, no gene flow, no natural selection, and large population size. Under these circumstances, the allele frequencies for a population are expected to remain consistent (equilibrium) over time. The H-W equations are expected to estimate genotype and allele frequencies for a population that is at equilibrium. The equations may not accurately predict the frequencies if the population is not at equilibrium (for example, if selection is occurring). However, it is possible that, even with the presence of an evolutionary force, a population may still demonstrate the expected H-W data.

$$p + q = 1$$
$$p^2 + 2pq + q^2 = 1$$

$p = R\ allele\ frequency$
$q = B\ allele\ frequency$
$p^2 = RR\ genotype\ frequency$
$2pq = RB\ genotype\ frequency$
$q^2 = BB\ genotype\ frequency$

Hardy-Weinberg equations

Once we have the frequencies for each genotype, we can then find the expected numbers by multiplying the frequencies by the total number of individuals (50).

$$\text{Red (RR) expected number} = .09 \times 50 = 4.5$$
$$\text{Purple (RB) expected number} = .42 \times 50 = 21$$
$$\text{Blue (BB) expected number} = .49 \times 50 = 24.5$$

Expected value calculations

Now that we have both observed and expected values, we can plug them into the chi-squared equation.

Phenotype (Genotype)	Observed (o)	Expected (e)
Red (RR)	10	4.5
Purple (RB)	10	21
Blue (BB)	30	24.5

$$\chi^2 = \frac{(10-4.5)^2}{4.5} + \frac{(10-21)^2}{21} + \frac{(30-24.5)^2}{24.5} =$$
$$6.72 + 5.76 + 1.23 = 13.71$$

chi-squared calculation

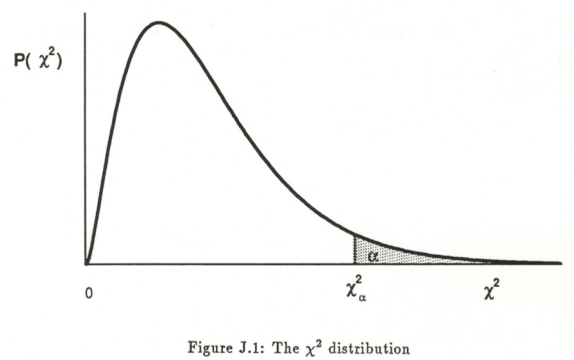
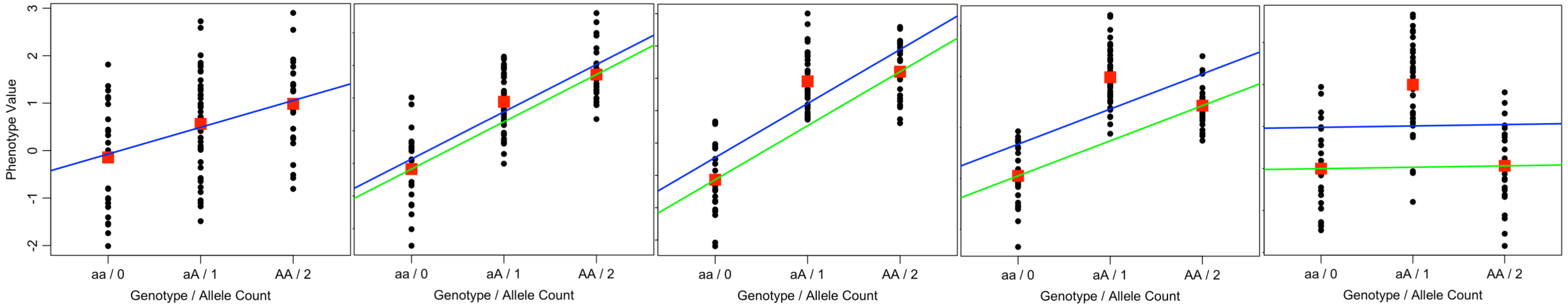


Figure J.1: The χ^2 distribution

```
# chi.sq.p <- pchisq( chi.sq, df=1, lower.tail=F )
```

https://www.biologysimulations.com/post/how-to-use-chi-squared-to-test-for-hardy-weinberg-equilibrium

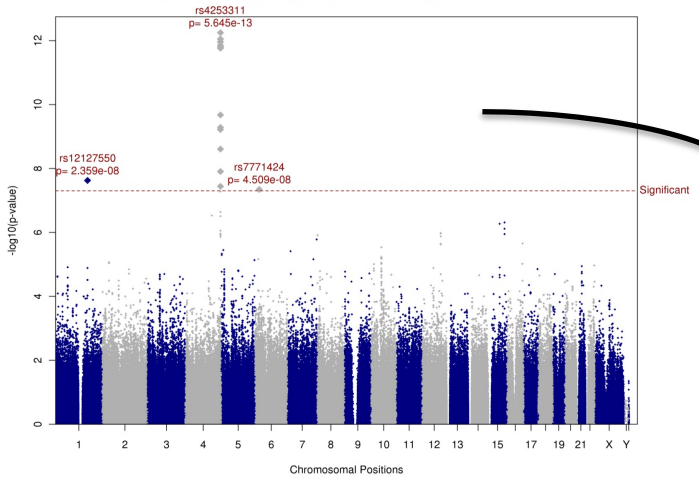
In practice, we typically code the additive model



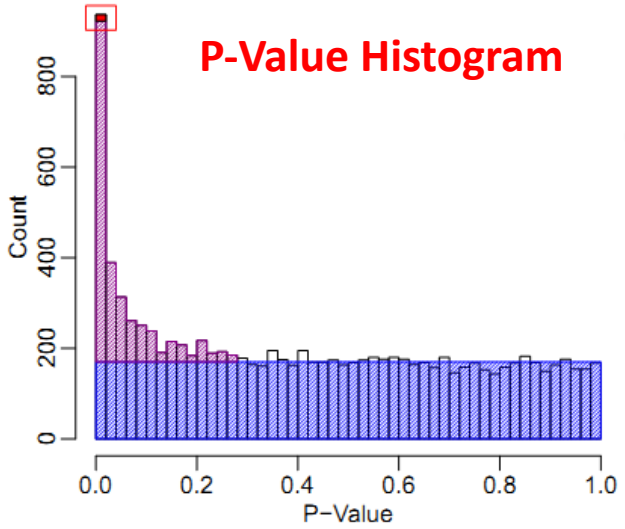
The statistical effect observed when testing an additive genotype coding is useful (i.e., not measured as 0) under the widest range of models, even when "wrong."

Even though few variants are believed to be *purely* additive, very few have been proven to show large deviations of the sort where an additive model is not *useful*.

Manhattan Plot



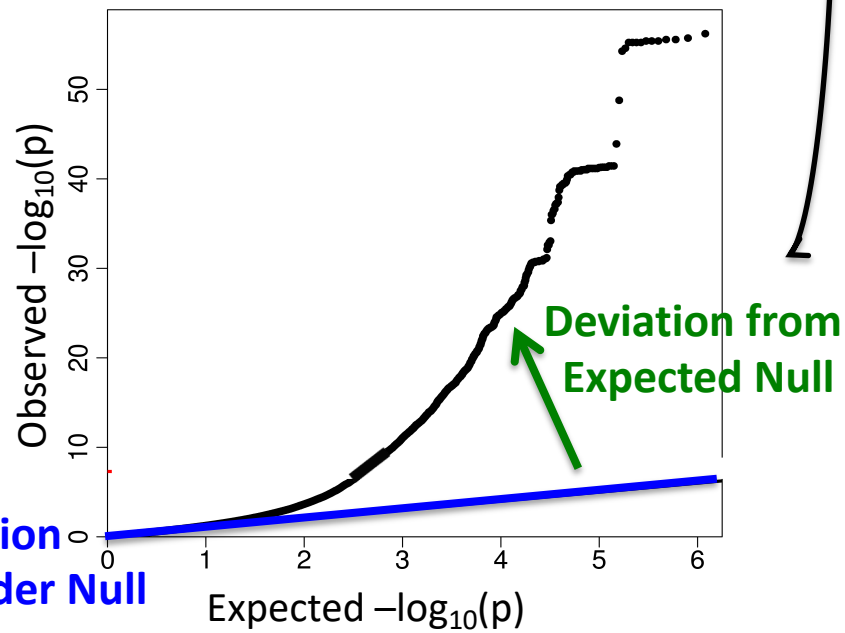
P-Value Histogram



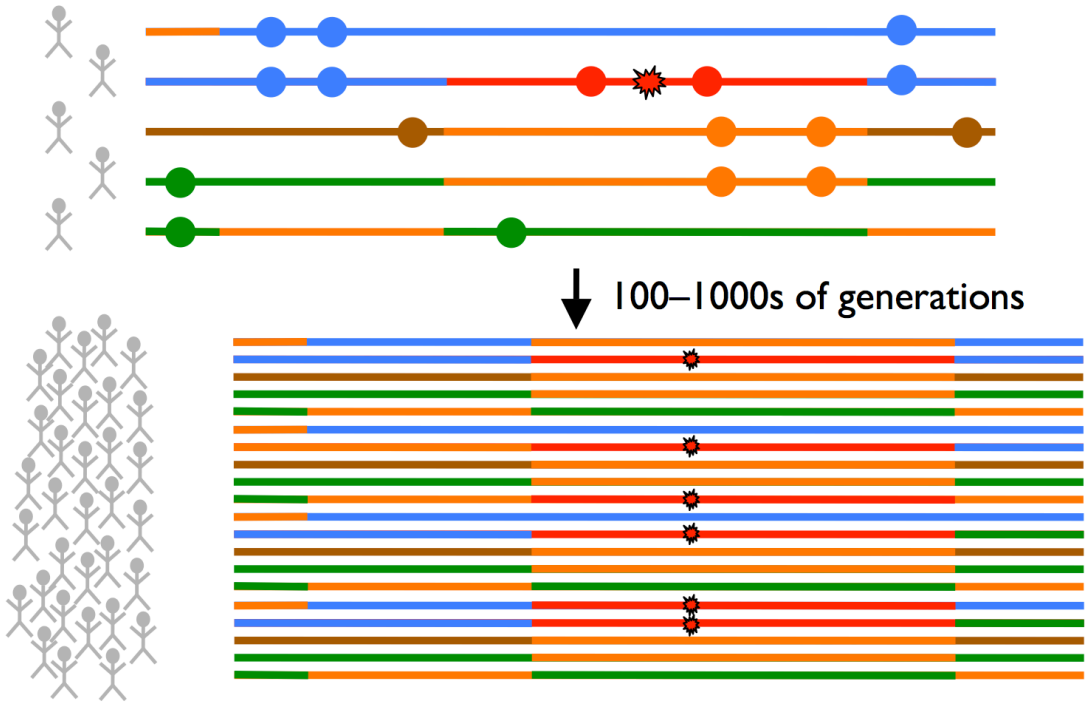
Manhattan Plots show us the significance of SNPs at different locations

QQ-Plots show us how many low p-values there are in the study, compared to the null distribution

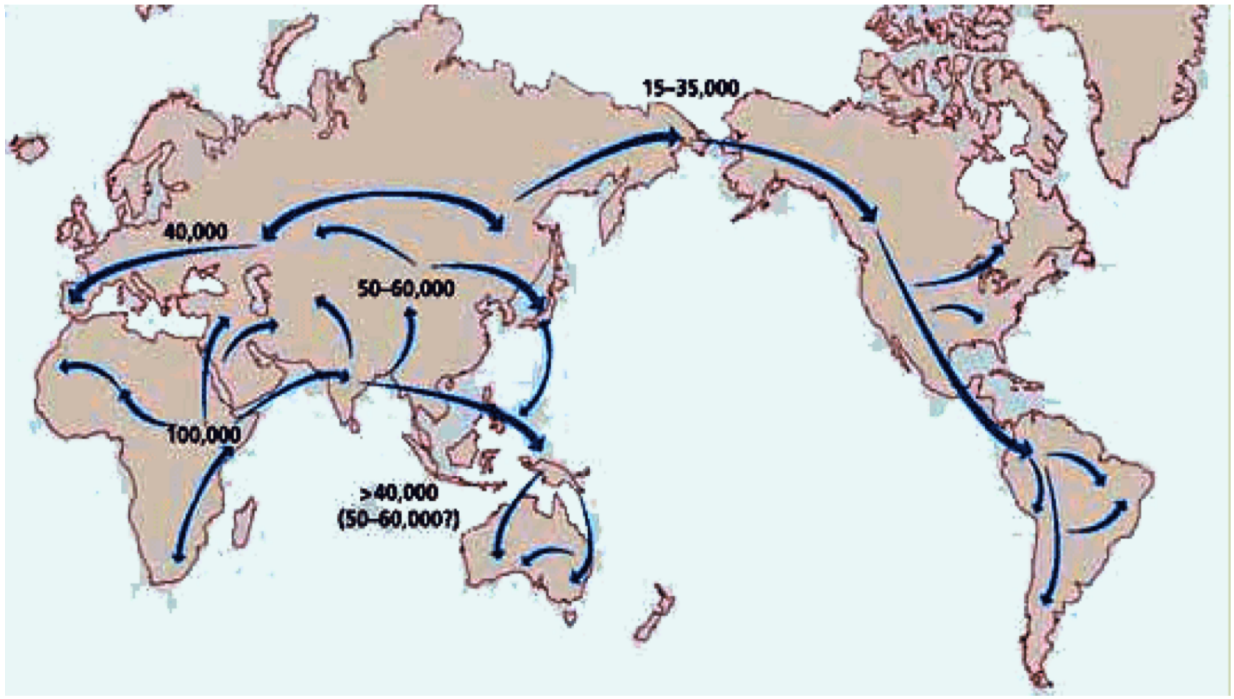
Quantile-Quantile Plot



SNP start as a mutation in one person,
and then spread through their offspring



Not all offspring go everywhere



Different populations have different frequencies of different SNPs

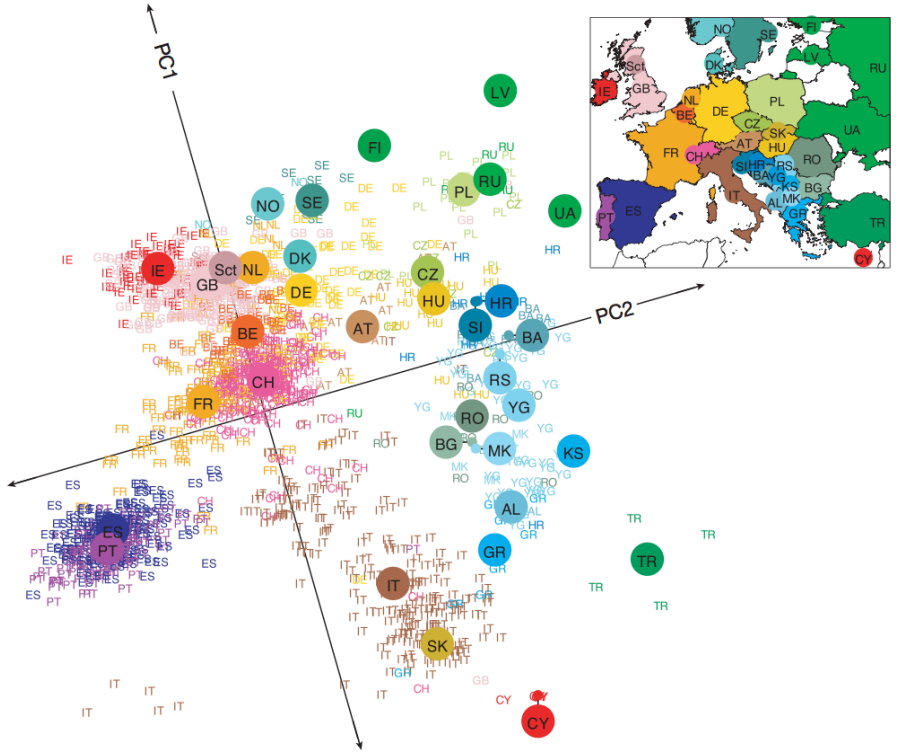
A Cautionary Note: Population Stratification

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

LETTERS

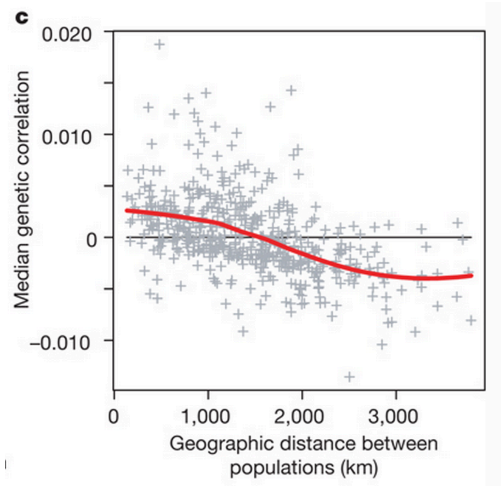
Genes mirror geography within Europe

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]

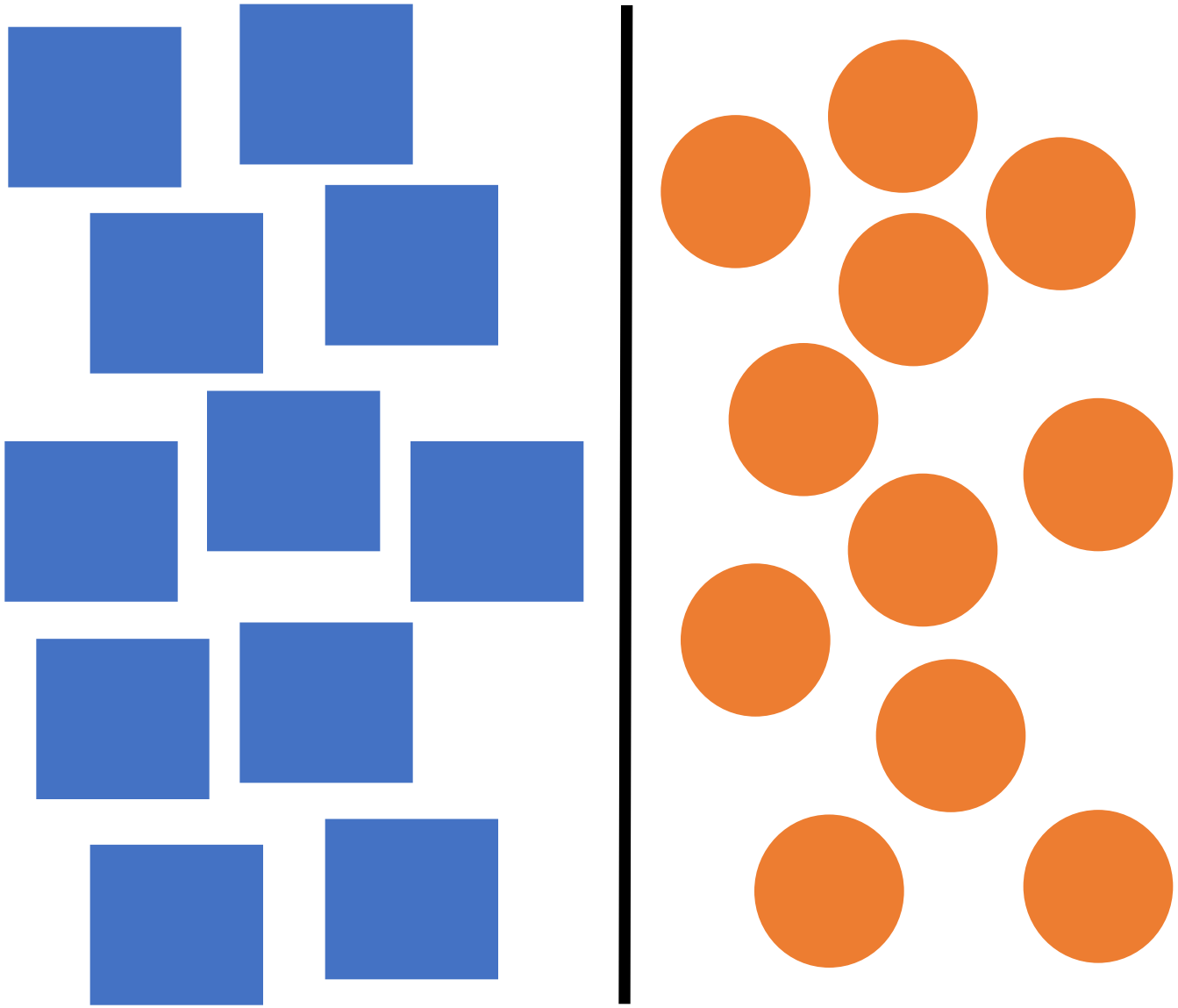


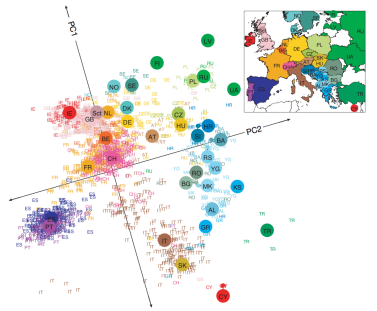
Given an individuals genotypes, you can predict their longitude and latitude or country of origin to a very fine scale.

In other words, the allele frequencies of SNP variants change depending on the population you are studying



A Cautionary Note: Population Stratification





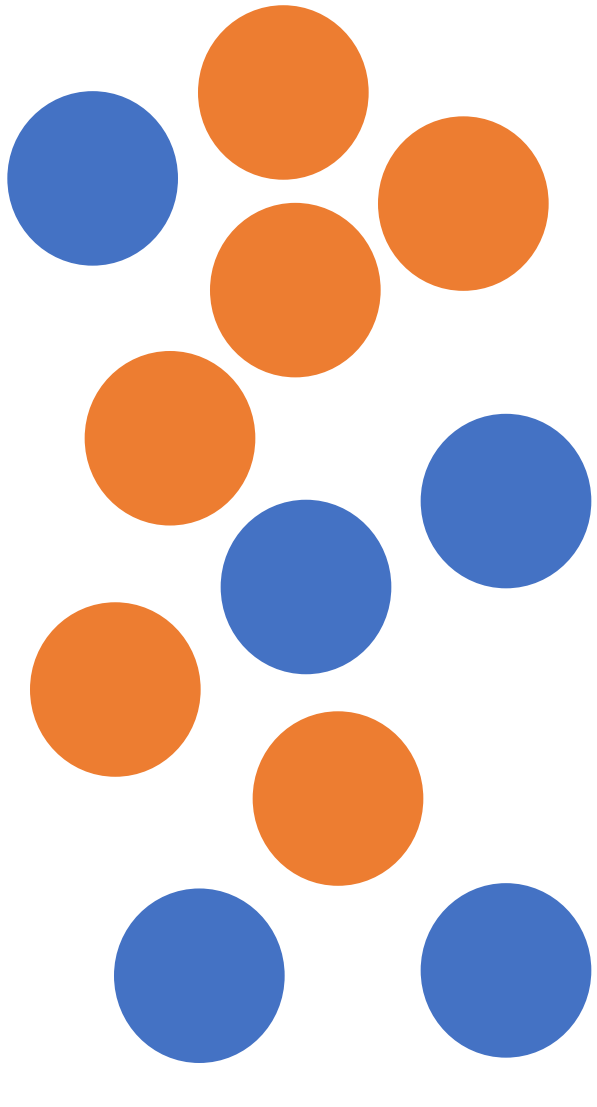
Case

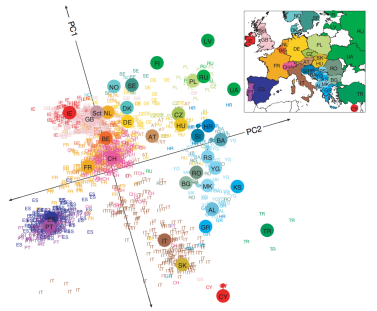
Control

African

European

A Cautionary Note: Population Stratification





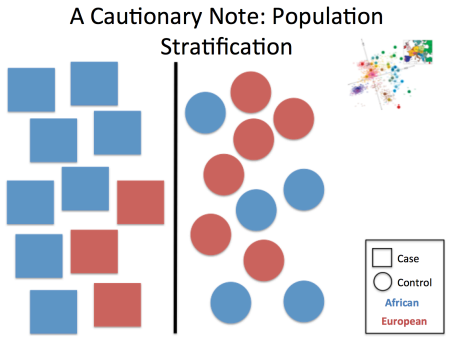
Case

Control

African

European

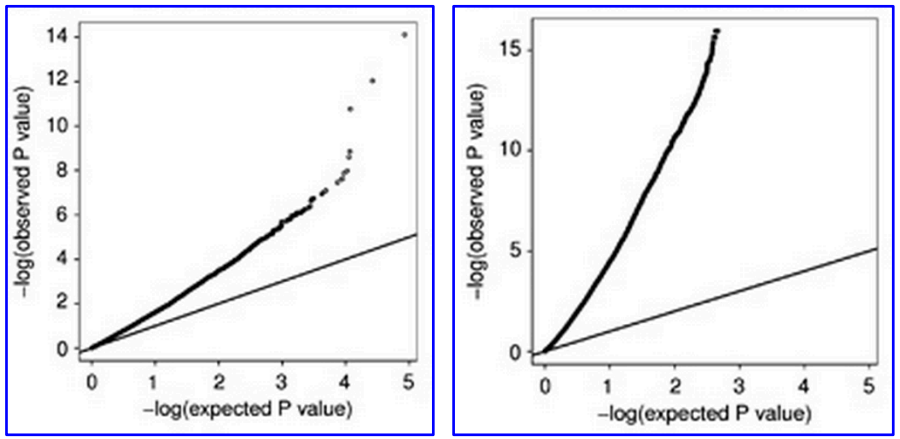
What is our test statistic measuring when we test for differences in the frequency of an allele among cases and controls?

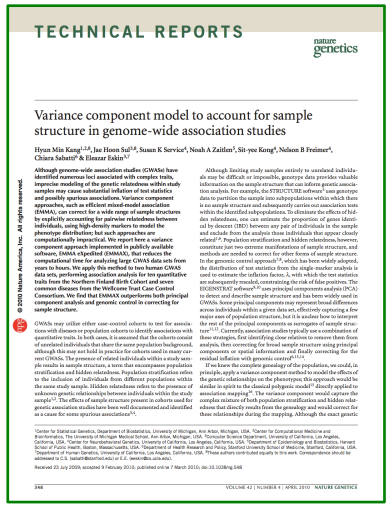


The same logic applies with quantitative traits if the mean varies among populations

Avoiding population stratification

Test if you have too many low p-values.




TECHNICAL REPORTS

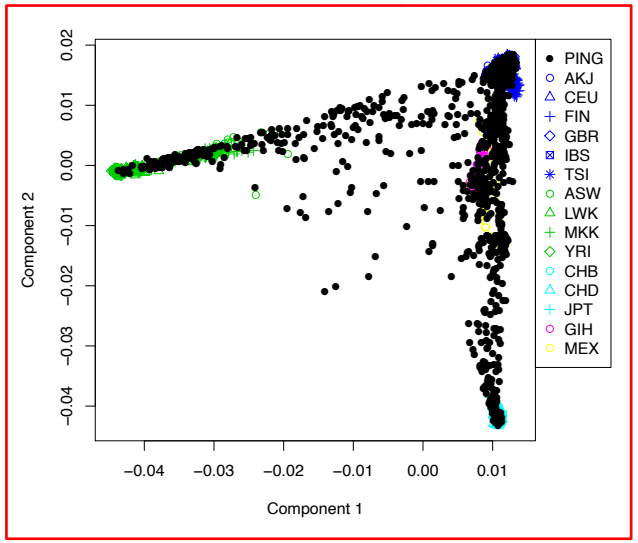
nature genetics

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]

Use advanced statistical methods to correct for it

Perform GWAS in homogenous groups



It is *absolutely critical* to investigate population stratification in association studies

… TGAGAGTCACTCGTCAAATCCGGGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCAAATCCGTGCCTGCTACTCGATCGGAAT …
… TGAGAGCCACTCGTCAAATCCGGGCCTGCTACTCGATCGGAAT …
… TGAGAGTCACTCGTCAAATCCGTGCCTGCTAATCGATCGGAAT …
… TGAGAGCCACTCGTCGAATCCGGGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCAAATCCGGGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCGAATCCGTGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCAAATCCGTGCCTGCTAATCGATCAGAAT …

… TGAGAGCCACTCGTCGAATCCGTGCCTGCTAATCGATCAGAAT …
… TGAGAGCCACTCGTCAAATCCGTGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCGAATCCGGGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCAAATCCGGGCCTGCTACTCGATCGGAAT …
… TGAGAGCCACTCGTCGAATCCGTGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCGAATCCGGGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCGAATCCGGGCCTGCTAATCGATCGGAAT …
… TGAGAGTCACTCGTCAAATCCGTGCCTGCTAATCGATCGGAAT …

Single Nucleotide Polymorphisms (SNPs): A very easy and cheap to measure, extremely abundant form of genetic variation. Definitely not the only kind

rs1	rs2	rs3	rs4	rs5
0	2	1	1	2
1	2	1	1	2
1	1	2	0	2
0	1	0	0	1
2	1	0	0	1
0	1	2	1	2
1	0	1	0	2
0	1	1	0	2

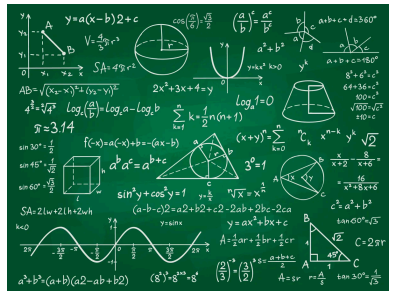
Compute a similarity (correlation) for two people based on all observed SNPs

1	-0.03	0.02	0.03	0.01	0	-0.02	-0.01
-0.03	1	0.03	0.01	0	-0.02	0.01	-0.01
0.01	0.03	1	0.04	0.03	0.01	0.01	0.02
0.03	0.03	0.04	1	-0.04	0.01	0.01	0
0.01	0	0.03	-0.04	1	0.03	0.02	0.03
0	-0.02	0.01	0.01	0.03	1	-0.03	0.01
-0.02	0.01	0.01	0.01	0.01	-0.03	1	-0.04
-0.01	-0.01	0.02	0	0.03	0.03	-0.04	1

rs1	rs2	rs3	rs4	rs5
0	2	1	1	2
1	2	1	1	2
1	1	2	0	2
0	1	0	0	1
2	1	0	0	1
0	1	2	1	2
1	0	1	0	2
0	1	1	0	2

Variability in SNPs across people

1	-0.03	0.02	0.03	0.01	0	-0.02	-0.01
-0.03	1	0.03	0.01	0	-0.02	0.01	-0.01
0.01	0.03	1	0.04	0.03	0.01	0.01	0.02
0.03	0.03	0.04	1	-0.04	0.01	0.01	0
0.01	0	0.03	-0.04	1	0.03	0.02	0.03
0	-0.02	0.01	0.01	0.03	1	-0.03	0.01
-0.02	0.01	0.01	0.01	0.01	-0.03	1	-0.04
-0.01	-0.01	0.02	0	0.03	0.03	-0.04	1

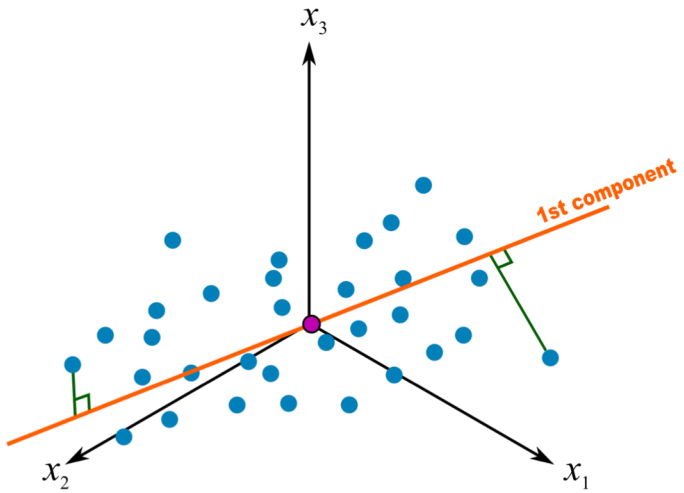


Principal Components Analysis

PC1	PC2	PC3	PC4	PC5
-1.2	0.25	1	1.2	-2
1.3	0.33	1.1	0.8	-2.3
2.0	0.75	-4	0.7	3.1
2.1	-1.3	-4	0.6	2.3
1.1	-2.5	-3.2	0.3	0.4
0.9	0.6	0.4	-1	-0.4
-0.2	1.4	-0.4	-1	0.76
-3	1.6	2	-1	1.1

Re-summarized variability in SNPs across people

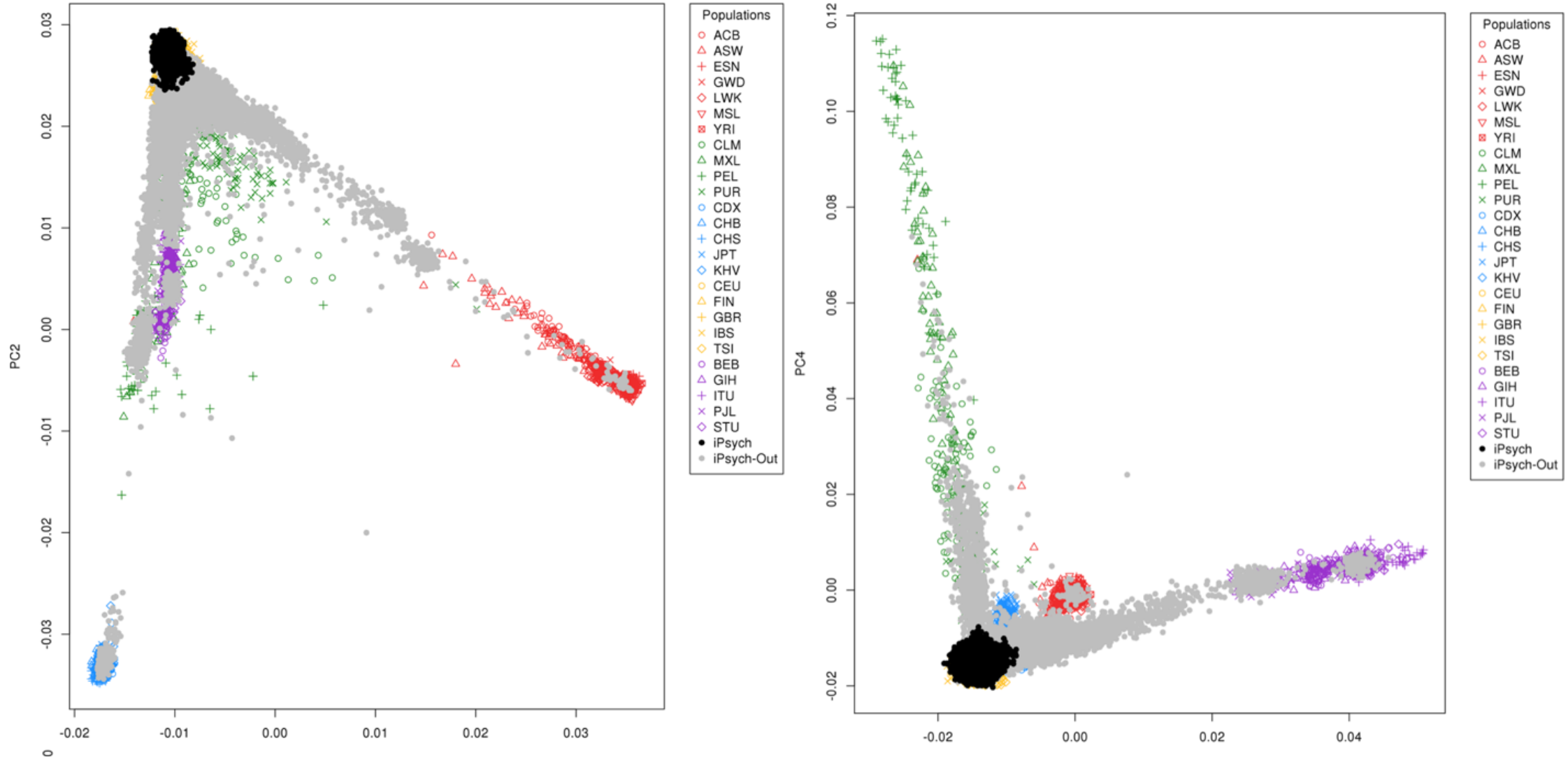
Each individual has three "measures" x1, x2, x3 on which they *covary*



Each individual can be described in three new measures that capture how they *covary* on the original x1, x2, x3

… but the new measures are more efficient

… math has constructed them to concentrate the most possible covariance in the fewest new measures



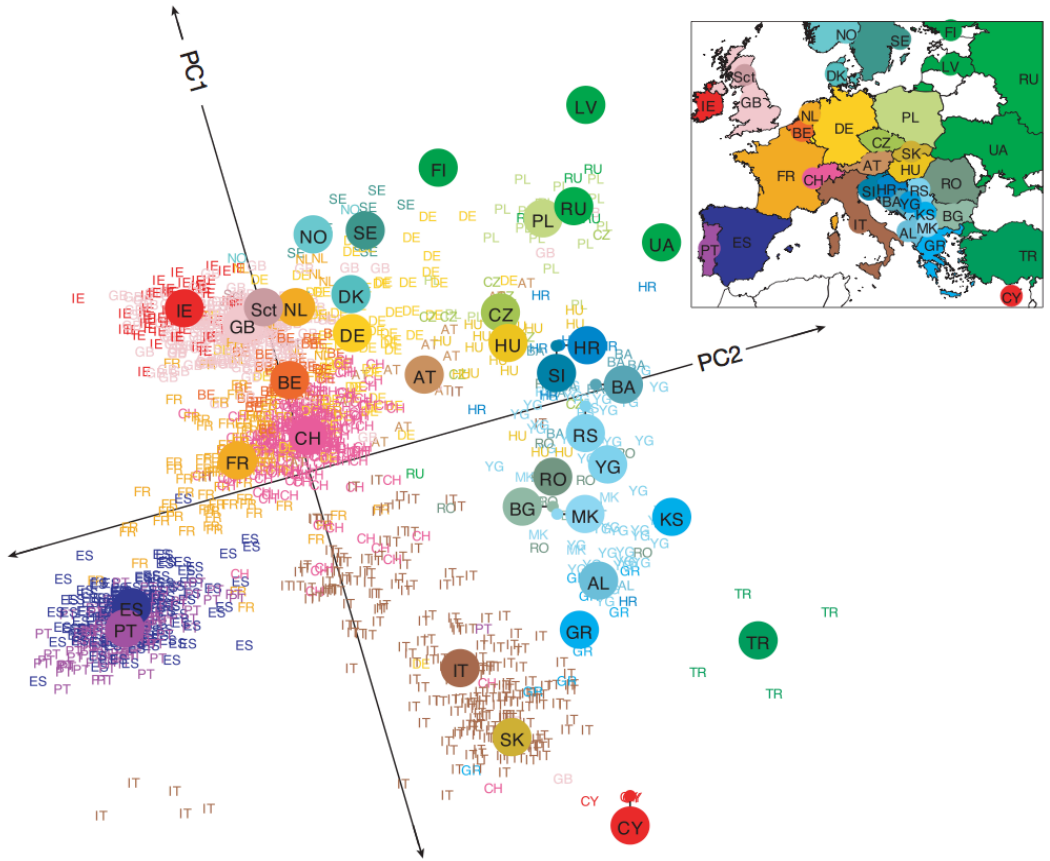
Summarize similarity in 12,000,000 SNPs among 65,000 Danes using 4 new measures (Principal Components)

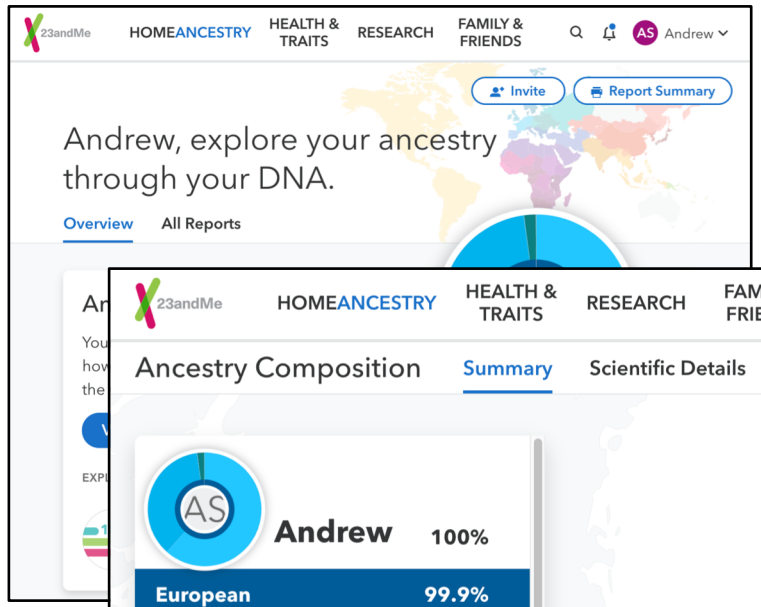
nature Vol 456|6 November 2008|doi:10.1038/nature07331

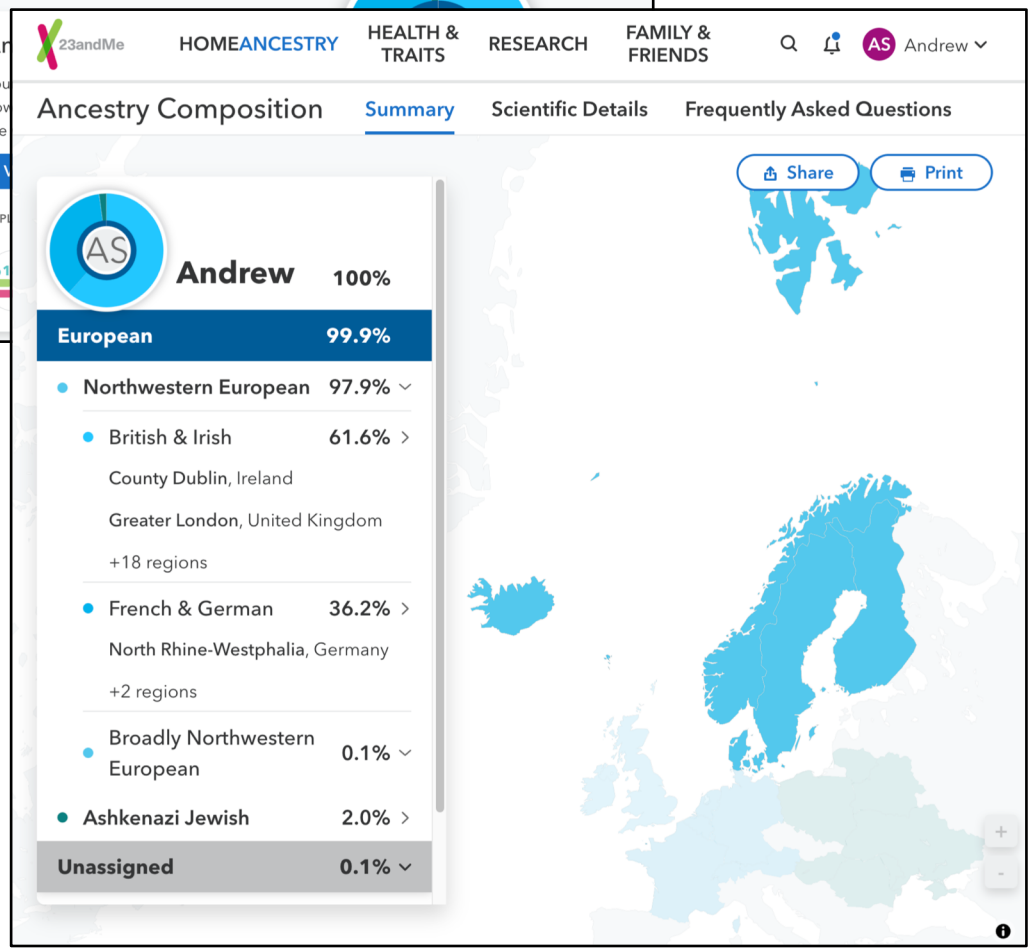
LETTERS

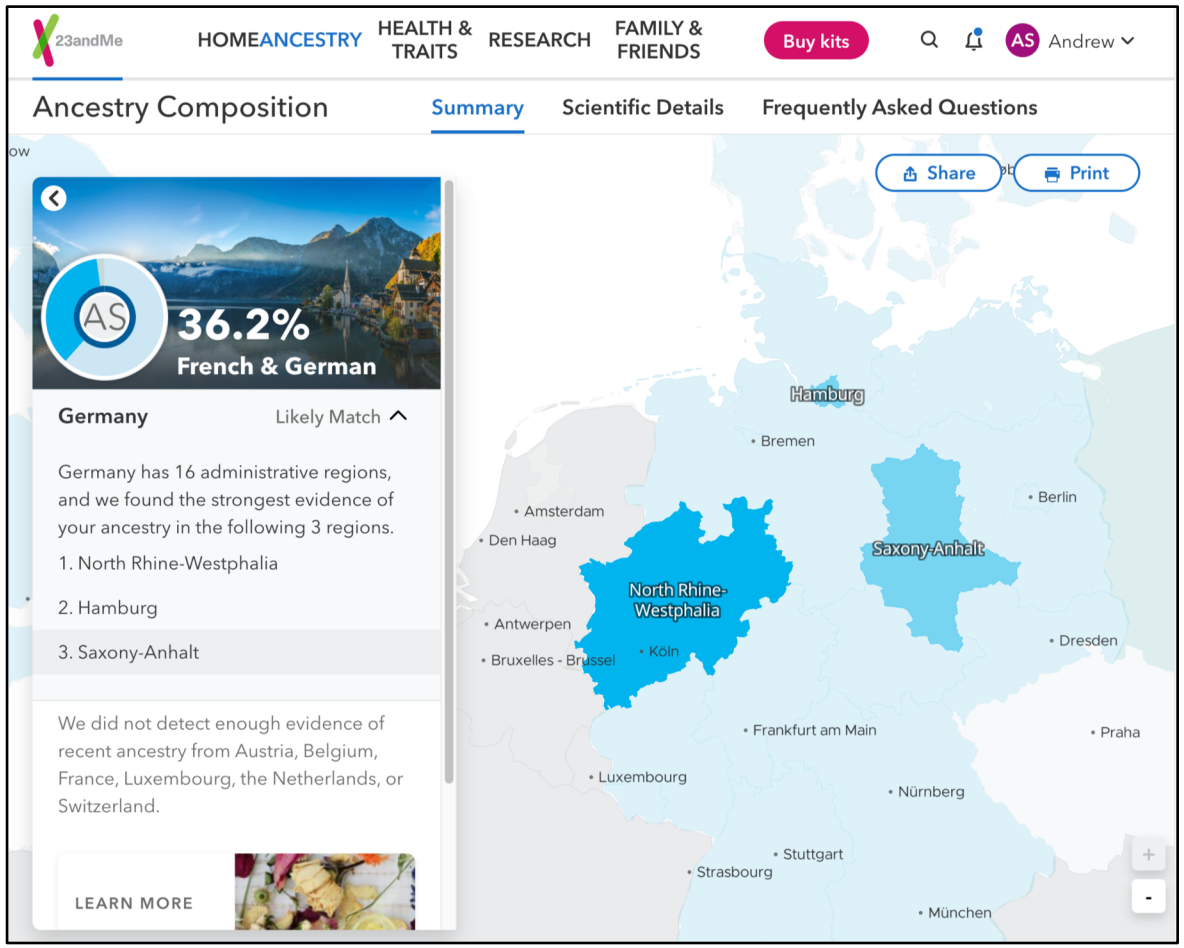
Genes mirror geography within Europe

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]



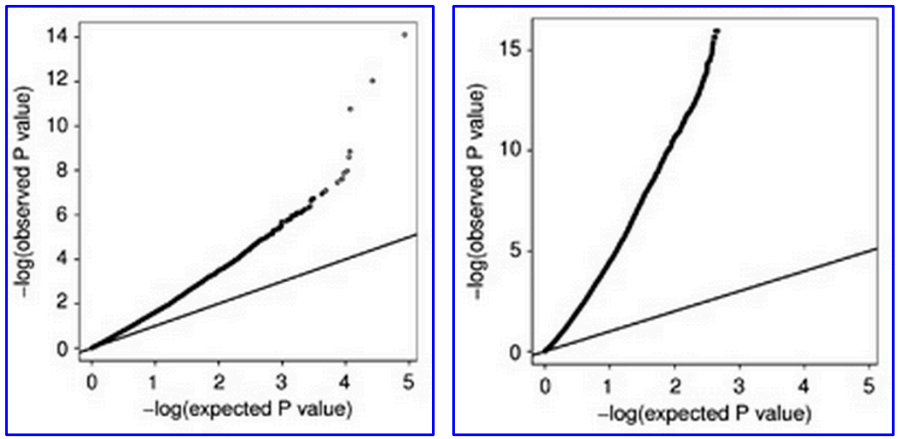

23andMe
HOMEANCESTRY
HEALTH & TRAITS
RESEARCH
FAMILY & FRIENDS
Andrew
Invite
Report Summary
Andrew, explore your ancestry through your DNA.
Overview
All Reports

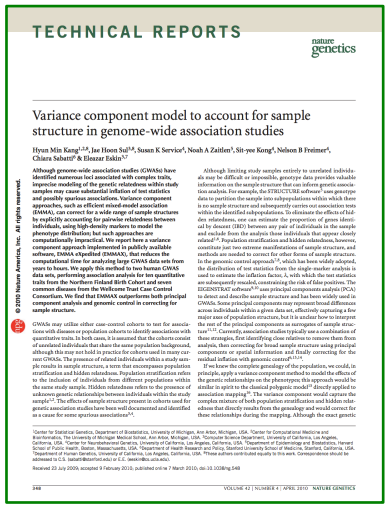

23andMe
HOMEANCESTRY
HEALTH & TRAITS
RESEARCH
FAMILY & FRIENDS
Andrew
Ancestry Composition
Summary
Scientific Details
Frequently Asked Questions
Share
Print
Andrew 100%
European 99.9%
Northwestern European 97.9%
British & Irish 61.6%
County Dublin, Ireland
Greater London, United Kingdom
+18 regions
French & German 36.2%
North Rhine-Westphalia, Germany
+2 regions
Broadly Northwestern European 0.1%
Ashkenazi Jewish 2.0%
Unassigned 0.1%


23andMe
HOMEANCESTRY
HEALTH & TRAITS
RESEARCH
FAMILY & FRIENDS
Buy kits
Andrew
Ancestry Composition
Summary
Scientific Details
Frequently Asked Questions
Share
Print
36.2%
French & German
Germany
Likely Match
Germany has 16 administrative regions, and we found the strongest evidence of your ancestry in the following 3 regions.
1. North Rhine-Westphalia
2. Hamburg
3. Saxony-Anhalt
We did not detect enough evidence of recent ancestry from Austria, Belgium, France, Luxembourg, the Netherlands, or Switzerland.
LEARN MORE
Hamburg
Bremen
Amsterdam
Den Haag
Berlin
Saxony-Anhalt
North Rhine-Westphalia
Antwerpen
Köln
Bruxelles - Brussel
Dresden
Frankfurt am Main
Praha
Luxembourg
Nürnberg
Stuttgart
Strasbourg
München

Avoiding population stratification

Test if you have too many low p-values.




TECHNICAL REPORTS

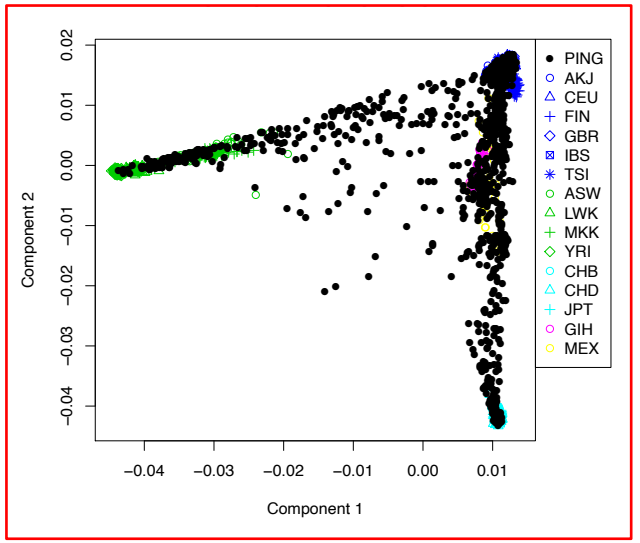
nature genetics

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

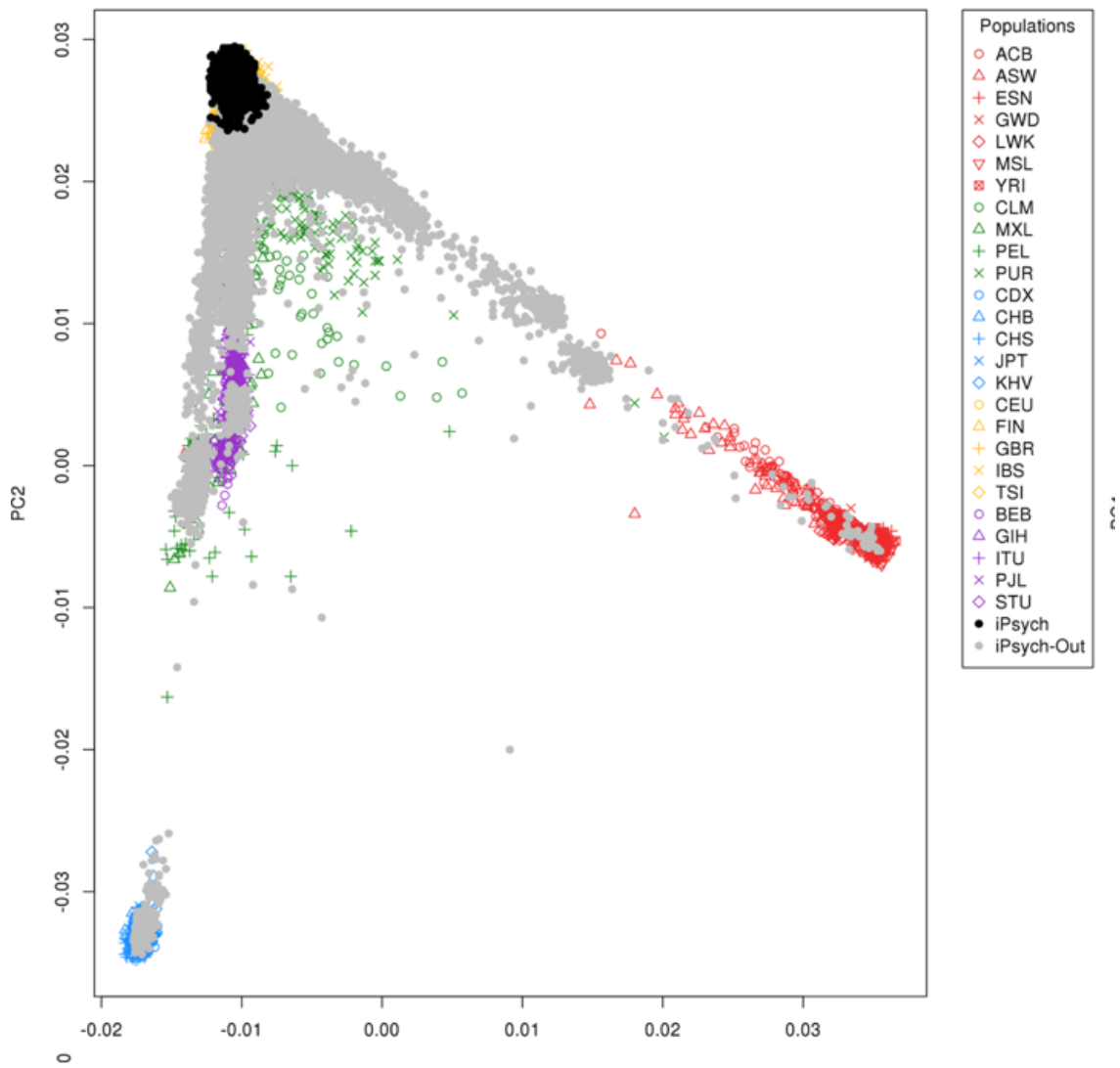
Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[6], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]

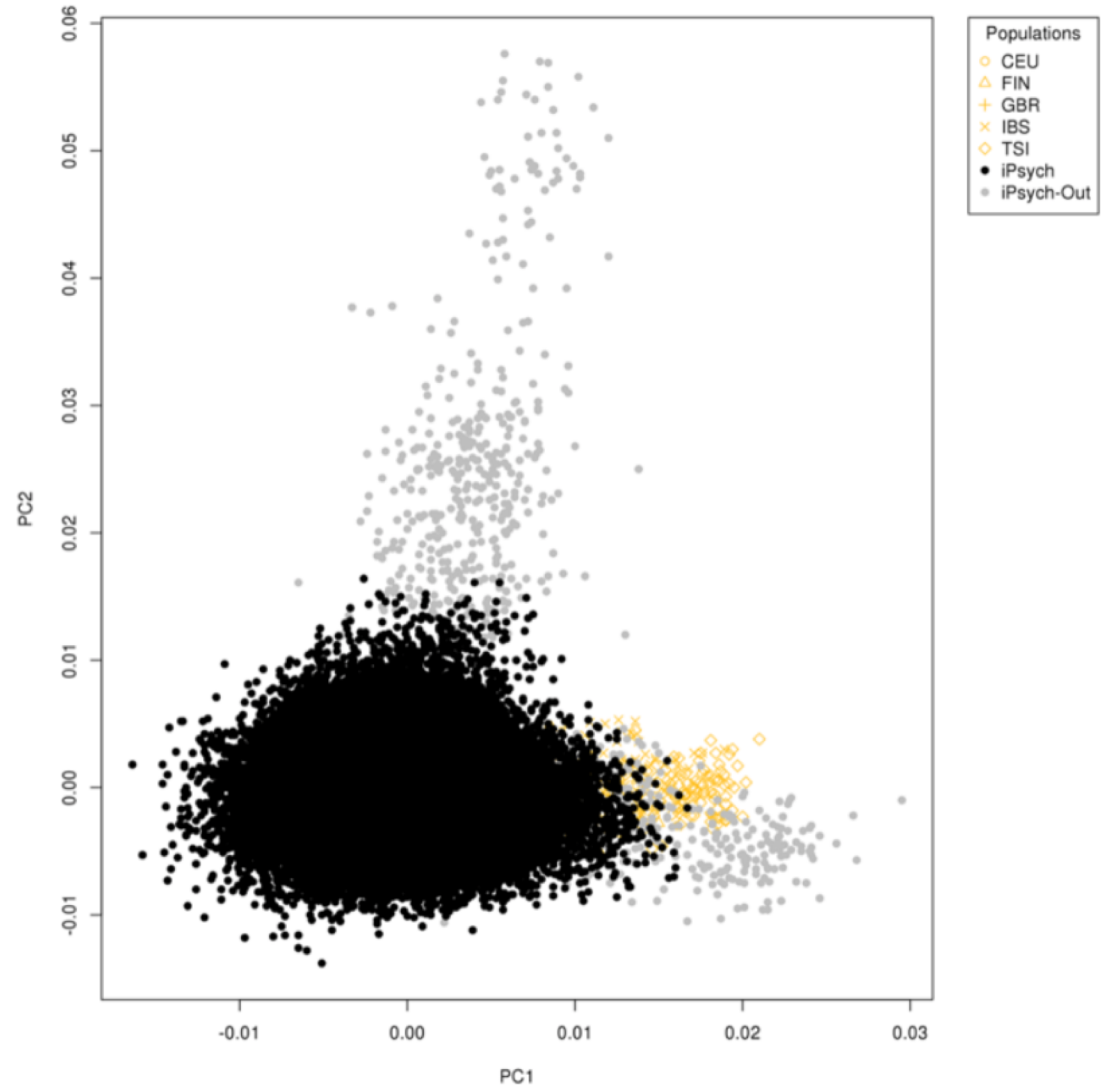
Use advanced statistical methods to correct for it

Perform GWAS in homogenous groups



It is *absolutely critical* to investigate population stratification in association studies


PC2
PC1
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
-0.02
-0.01
0.00
0.01
0.02
0.03
Populations
ACB
ASW
ESN
GWD
LWK
MSL
YRI
CLM
MXL
PEL
PUR
CDX
CHB
CHS
JPT
KHV
CEU
FIN
GBR
IBS
TSI
BEB
GIH
ITU
PJL
STU
iPsych
iPsych-Out


PC2
PC1
PC4
0.06
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
-0.01
0.00
0.01
0.02
0.03
Populations
CEU
FIN
GBR
IBS
TSI
iPsych
iPsych-Out

- (g)LMM based GWAS programs
  - BOLT-LMM: https://alkesgroup.broadinstitute.org/BOLT-LMM/
  - GCTA: https://yanglab.westlake.edu.cn/software/gcta/#fastGWA
  - FaST-LMM: https://www.microsoft.com/en-us/research/project/fastlmm/
  - regenie: https://rgcgithub.github.io/regenie/
  - SAIGE: https://www.leelabsg.org/software

- Programs that can do PCA on genotype data at scale:
  - smartpca: https://alkesgroup.broadinstitute.org/EIGENSOFT/
    https://github.com/chrchang/eigensoft
  - flashpca: https://github.com/gabraham/flashpca
  - plink: https://www.cog-genomics.org/plink/1.9/strat
  - plink2: https://www.cog-genomics.org/plink/2.0/strat
  - GCTA: https://yanglab.westlake.edu.cn/software/gcta/#PCA