Prediction for the individual: A practical workshop

Panum – 21.2.14

9 November 2021 14:15

Don’t worry guys,
I called Mom yesterday

(best son ever)

Andrew Schork

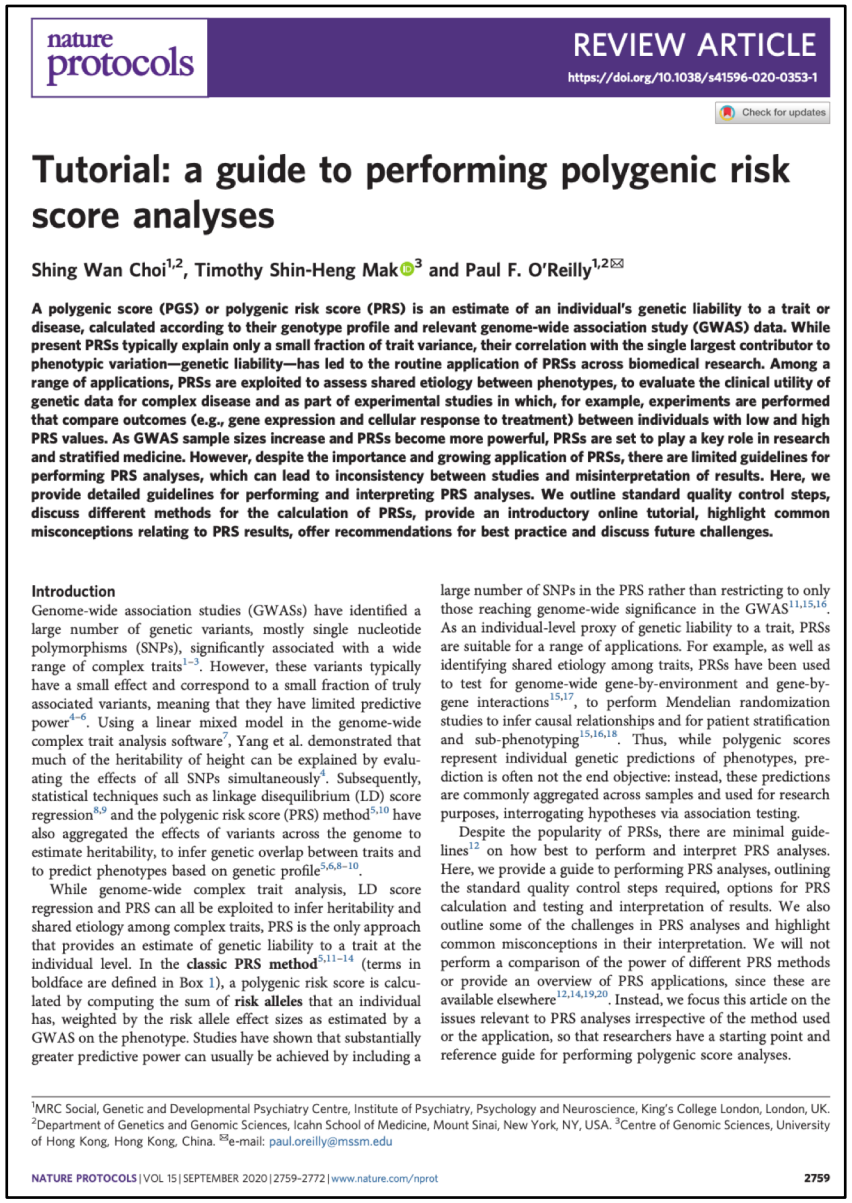
Institute for Biological Psychiatry
Andrew.Joseph.Schork@regionh.dk

UNIVERSITY OF COPENHAGEN

REGION H COPENHAGEN MENTAL HEALTH SERVICSES

Centre for GeoGenetics

iPSYCH



nature protocols

REVIEW ARTICLE

https://doi.org/10.1038/s41596-020-0353-1

Check for updates

Tutorial: a guide to performing polygenic risk score analyses

Shing Wan Choi[1,2], Timothy Shin-Heng Mak[3] and Paul F. O'Reilly[1,2]✉

A polygenic score (PGS) or polygenic risk score (PRS) is an estimate of an individual's genetic liability to a trait or disease, calculated according to their genotype profile and relevant genome-wide association study (GWAS) data. While present PRSs typically explain only a small fraction of trait variance, their correlation with the single largest contributor to phenotypic variation—genetic liability—has led to the routine application of PRSs across biomedical research. Among a range of applications, PRSs are exploited to assess shared etiology between phenotypes, to evaluate the clinical utility of genetic data for complex disease and as part of experimental studies in which, for example, experiments are performed that compare outcomes (e.g., gene expression and cellular response to treatment) between individuals with low and high PRS values. As GWAS sample sizes increase and PRSs become more powerful, PRSs are set to play a key role in research and stratified medicine. However, despite the importance and growing application of PRSs, there are limited guidelines for performing PRS analyses, which can lead to inconsistency between studies and misinterpretation of results. Here, we provide detailed guidelines for performing and interpreting PRS analyses. We outline standard quality control steps, discuss different methods for the calculation of PRSs, provide an introductory online tutorial, highlight common misconceptions relating to PRS results, offer recommendations for best practice and discuss future challenges.

Introduction

Genome-wide association studies (GWASs) have identified a large number of genetic variants, mostly single nucleotide polymorphisms (SNPs), significantly associated with a wide range of complex traits[1–3]. However, these variants typically have a small effect and correspond to a small fraction of truly associated variants, meaning that they have limited predictive power[4–6]. Using a linear mixed model in the genome-wide complex trait analysis software[7], Yang et al. demonstrated that much of the heritability of height can be explained by evaluating the effects of all SNPs simultaneously[4]. Subsequently, statistical techniques such as linkage disequilibrium (LD) score regression[8,9] and the polygenic risk score (PRS) method[5,10] have also aggregated the effects of variants across the genome to estimate heritability, to infer genetic overlap between traits and to predict phenotypes based on genetic profile[5,6,8–10].

While genome-wide complex trait analysis, LD score regression and PRS can all be exploited to infer heritability and shared etiology among complex traits, PRS is the only approach that provides an estimate of genetic liability to a trait at the individual level. In the **classic PRS method**[5,11–14] (terms in boldface are defined in Box 1), a polygenic risk score is calculated by computing the sum of **risk alleles** that an individual has, weighted by the risk allele effect sizes as estimated by a GWAS on the phenotype. Studies have shown that substantially greater predictive power can usually be achieved by including a large number of SNPs in the PRS rather than restricting to only those reaching genome-wide significance in the GWAS[11,15,16]. As an individual-level proxy of genetic liability to a trait, PRSs are suitable for a range of applications. For example, as well as identifying shared etiology among traits, PRSs have been used to test for genome-wide gene-by-environment and gene-by-gene interactions[15,17], to perform Mendelian randomization studies to infer causal relationships and for patient stratification and sub-phenotyping[15,16,18]. Thus, while polygenic scores represent individual genetic predictions of phenotypes, prediction is often not the end objective: instead, these predictions are commonly aggregated across samples and used for research purposes, interrogating hypotheses via association testing.

Despite the popularity of PRSs, there are minimal guidelines[12] on how best to perform and interpret PRS analyses. Here, we provide a guide to performing PRS analyses, outlining the standard quality control steps required, options for PRS calculation and testing and interpretation of results. We also outline some of the challenges in PRS analyses and highlight common misconceptions in their interpretation. We will not perform a comparison of the power of different PRS methods or provide an overview of PRS applications, since these are available elsewhere[12,14,19,20]. Instead, we focus this article on the issues relevant to PRS analyses irrespective of the method used or the application, so that researchers have a starting point and reference guide for performing polygenic score analyses.

[1]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [2]Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York, NY, USA. [3]Centre of Genomic Sciences, University of Hong Kong, Hong Kong, China. ✉e-mail: paul.oreilly@mssm.edu

https://choishingwan.github.io/PRS-Tutorial/

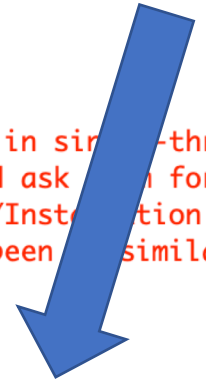
1. QC Base Data (i.e., GWAS results - betas)

2. QC Target Data (i.e., Genotype Data)

3. Calculate PRS

4. Visualize Results

```
> rm( list=ls() )
>
> require( data.table)
Loading required package: data.table
data.table 1.14.0 using 1 threads (see ?getDTthreads).  Latest news: r-datatable.com
**********
This installation of data.table has not detected OpenMP support. It should still work but in sin -threaded mode.
This is a Mac. Please read https://mac.r-project.org/openmp/. Please engage with Apple and ask  for support. Check r-datatable.com
for updates, and our Mac instructions here: https://github.com/Rdatatable/data.table/wiki/Inst tion. After several years of many
reports of installation problems on Mac, it's time to gingerly point out that there have been similar problems on Windows or Linux.
**********
>
> getwd()
[1] "/Users/AndrewSchork/Desktop/Teaching/KU Neurogenetics - 2021/3.5/C3.5 - Prediction/workdir2"
> list.files()
 [1] "ADHD.txt"                    "AndrewSchork.tfam"           "AndrewSchork.tped"           "AndrewSchork.traw"
 [5] "ASD.txt"                     "Height.txt"                  "IQ.txt"                      "KGP_GSA_PGS.traw"
 [9] "MDD.txt"                     "PGC3_SCZ_wave3_public.v2.tsv" "PGS.txt"                     "SCZ_hist_INFO.jpeg"
[13] "SCZ_hist_Neff.jpeg"          "SCZ_raw_sample.txt"          "SCZ.txt"                     "Session3.5.C_Workshop.pptx"
[17] "Session3.5.C_WorkshopCode.R"
>
> ######
> ## Load test GWAS "summary statistics"
> ######
>
>
```

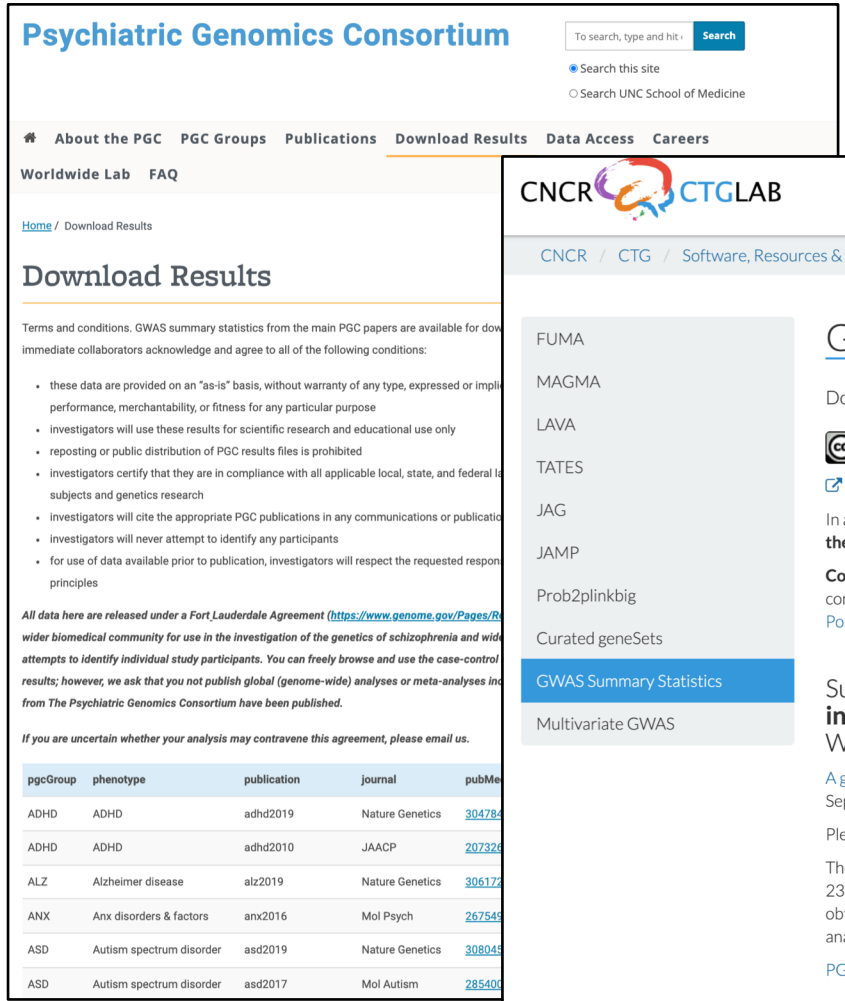


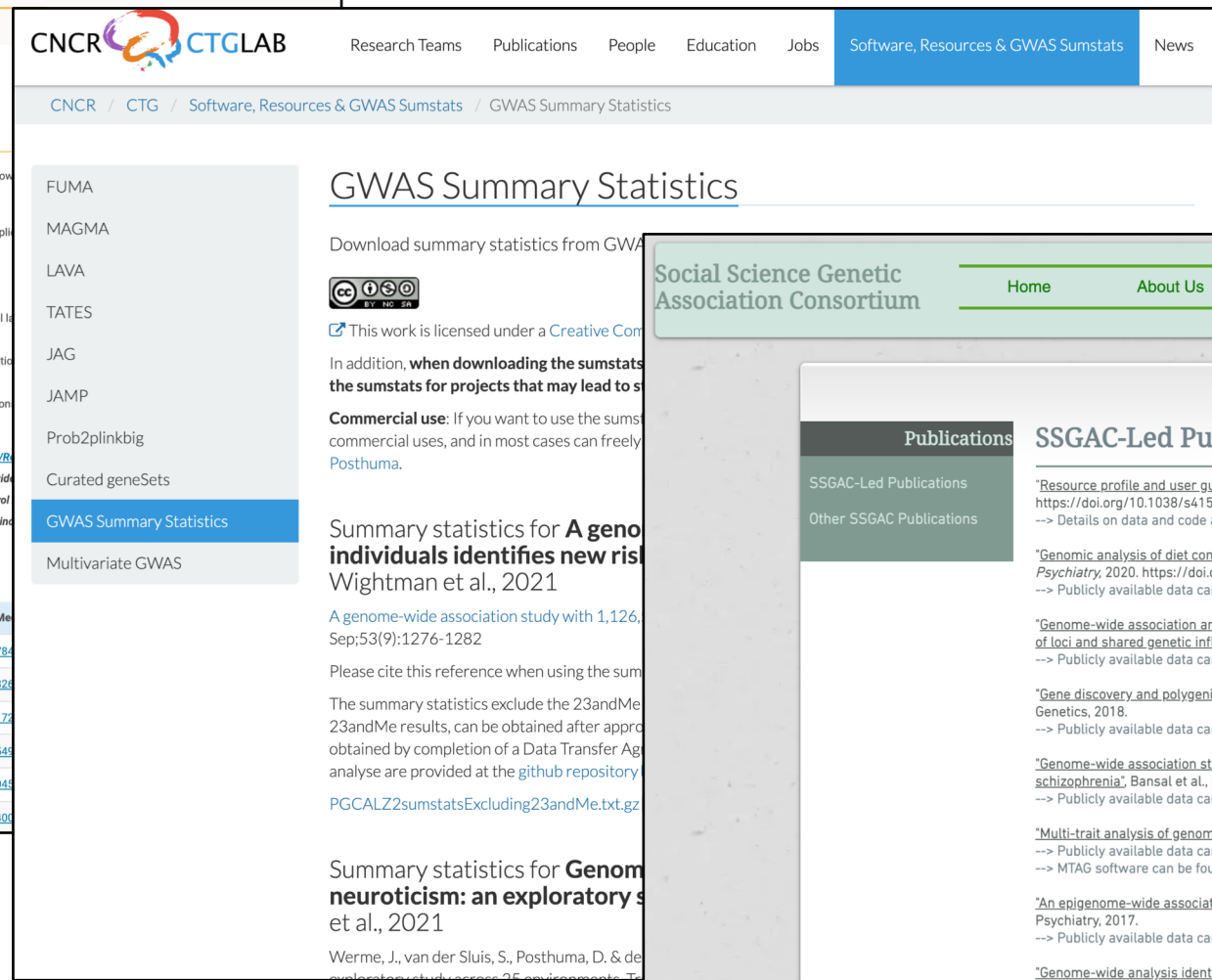
1. QC Base Data (i.e., GWAS results - betas)

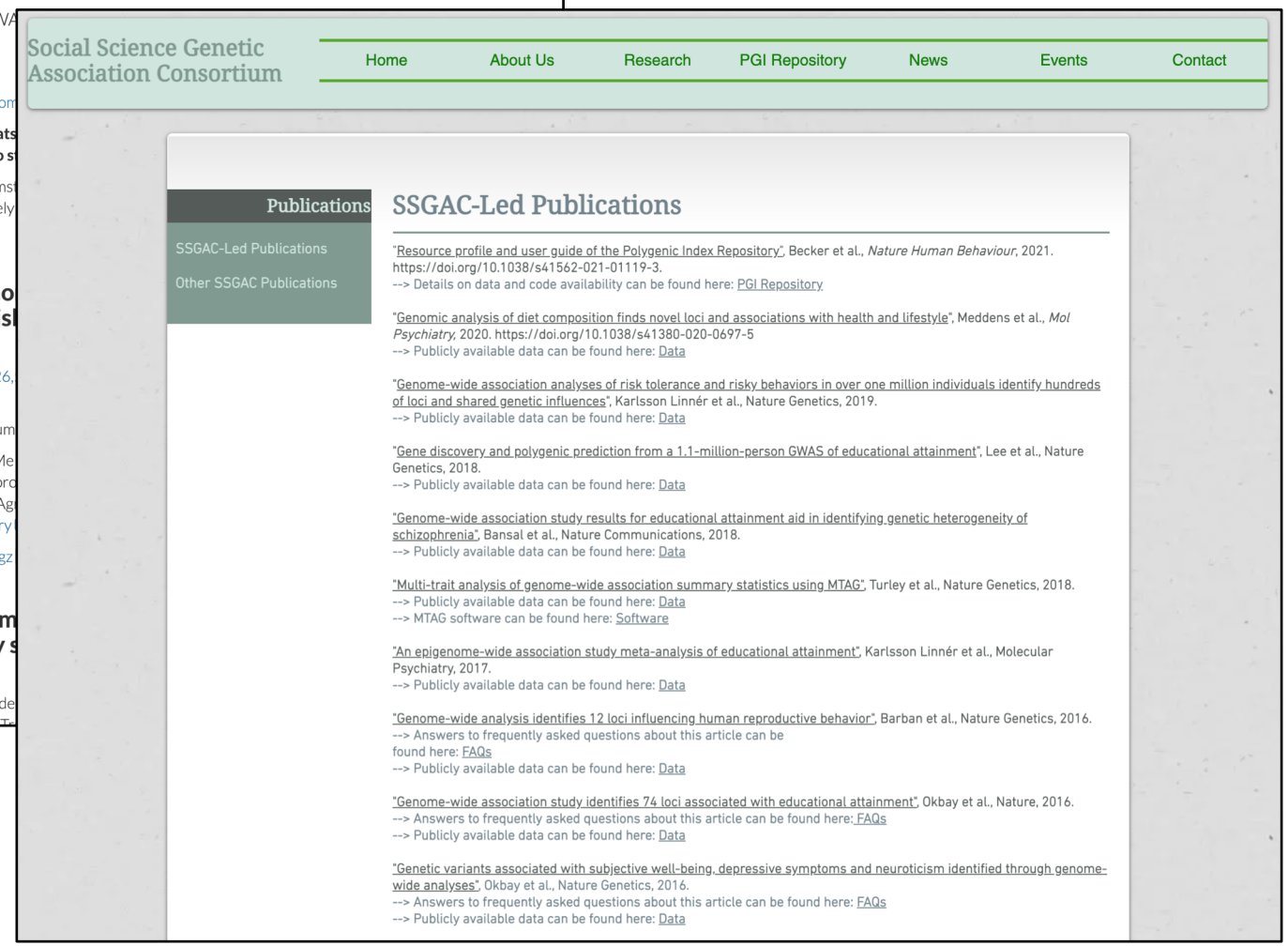
2. QC Target Data (i.e., Genotype Data)

3. Calculate PRS

4. Visualize Results


Psychiatric Genomics Consortium
To search, type and hit
Search
Search this site
Search UNC School of Medicine
About the PGC
PGC Groups
Publications
Download Results
Data Access
Careers
Worldwide Lab
FAQ
Home / Download Results
Download Results
Terms and conditions. GWAS summary statistics from the main PGC papers are available for dow
immediate collaborators acknowledge and agree to all of the following conditions:
these data are provided on an "as-is" basis, without warranty of any type, expressed or impli
performance, merchantability, or fitness for any particular purpose
investigators will use these results for scientific research and educational use only
reposting or public distribution of PGC results files is prohibited
investigators certify that they are in compliance with all applicable local, state, and federal la
subjects and genetics research
investigators will cite the appropriate PGC publications in any communications or publicatic
investigators will never attempt to identify any participants
for use of data available prior to publication, investigators will respect the requested respon
principles
All data here are released under a Fort Lauderdale Agreement (https://www.genome.gov/Pages/R
wider biomedical community for use in the investigation of the genetics of schizophrenia and wid
attempts to identify individual study participants. You can freely browse and use the case-control
results; however, we ask that you not publish global (genome-wide) analyses or meta-analyses inc
from The Psychiatric Genomics Consortium have been published.
If you are uncertain whether your analysis may contravene this agreement, please email us.
pgcGroup
phenotype
publication
journal
pubMe
ADHD
ADHD
adhd2019
Nature Genetics
304784
ADHD
ADHD
adhd2010
JAACP
207326
ALZ
Alzheimer disease
alz2019
Nature Genetics
306172
ANX
Anx disorders & factors
anx2016
Mol Psych
267549
ASD
Autism spectrum disorder
asd2019
Nature Genetics
308045
ASD
Autism spectrum disorder
asd2017
Mol Autism
285400


CNCR
CTGLAB
Research Teams
Publications
People
Education
Jobs
Software, Resources & GWAS Sumstats
News
CNCR / CTG / Software, Resources & GWAS Sumstats / GWAS Summary Statistics
FUMA
MAGMA
LAVA
TATES
JAG
JAMP
Prob2plinkbig
Curated geneSets
GWAS Summary Statistics
Multivariate GWAS
GWAS Summary Statistics
Download summary statistics from GWA
This work is licensed under a Creative Com
In addition, when downloading the sumstats
the sumstats for projects that may lead to s
Commercial use: If you want to use the sumst
commercial uses, and in most cases can freely
Posthuma.
Summary statistics for A geno
individuals identifies new risk
Wightman et al., 2021
A genome-wide association study with 1,126,
Sep;53(9):1276-1282
Please cite this reference when using the sum
The summary statistics exclude the 23andMe
23andMe results, can be obtained after appro
obtained by completion of a Data Transfer Ag
analyse are provided at the github repository
PGCALZ2sumstatsExcludingMe.txt.gz
Summary statistics for Genom
neuroticism: an exploratory s
et al., 2021
Werme, J., van der Sluis, S., Posthuma, D. & de


Social Science Genetic Association Consortium
Home
About Us
Research
PGI Repository
News
Events
Contact
Publications
SSGAC-Led Publications
Other SSGAC Publications
SSGAC-Led Publications
"Resource profile and user guide of the Polygenic Index Repository", Becker et al., Nature Human Behaviour, 2021. https://doi.org/10.1038/s41562-021-01119-3.
--> Details on data and code availability can be found here: PGI Repository
"Genomic analysis of diet composition finds novel loci and associations with health and lifestyle", Meddens et al., Mol Psychiatry, 2020. https://doi.org/10.1038/s41380-020-0697-5
--> Publicly available data can be found here: Data
"Genome-wide association analyses of risk tolerance and risky behaviors in over one million individuals identify hundreds of loci and shared genetic influences", Karlsson Linnér et al., Nature Genetics, 2019.
--> Publicly available data can be found here: Data
"Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment", Lee et al., Nature Genetics, 2018.
--> Publicly available data can be found here: Data
"Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia", Bansal et al., Nature Communications, 2018.
--> Publicly available data can be found here: Data
"Multi-trait analysis of genome-wide association summary statistics using MTAG", Turley et al., Nature Genetics, 2018.
--> Publicly available data can be found here: Data
--> MTAG software can be found here: Software
"An epigenome-wide association study meta-analysis of educational attainment", Karlsson Linnér et al., Molecular Psychiatry, 2017.
--> Publicly available data can be found here: Data
"Genome-wide analysis identifies 12 loci influencing human reproductive behavior", Barban et al., Nature Genetics, 2016.
--> Answers to frequently asked questions about this article can be found here: FAQs
--> Publicly available data can be found here: Data
"Genome-wide association study identifies 74 loci associated with educational attainment", Okbay et al., Nature, 2016.
--> Answers to frequently asked questions about this article can be found here: FAQs
--> Publicly available data can be found here: Data
"Genetic variants associated with subjective well-being, depressive symptoms and neuroticism identified through genome-wide analyses", Okbay et al., Nature Genetics, 2016.
--> Answers to frequently asked questions about this article can be found here: FAQs
--> Publicly available data can be found here: Data

```
AJSMacbook2020:sumstats AndrewSchork$ wc -l *
  13554551 MDD2018_ex23andMe
   2336270 Meta-analysis_Locke_et_al+UKBiobank_2018_UPDATED.txt
   7585078 PGC3_SCZ_wave3_public.v2.tsv
  12688340 PGCALZ2sumstatsExcluding23andMe.txt
   9295119 SavageJansen_2018_intelligence_metaanalysis.txt
   8094095 daner_adhd_meta_filtered_NA_iPSYCH23_PGC11_sigPCs_woSEX_2ell6sd_EUR_Neff_70.meta
   9112387 iPSYCH-PGC_ASD_Nov2017
```

IQ

SNP	UNIQUE_ID	CHR	POS	A1	A2	EAF_HRC	Zscore	stdBeta	SE	P	N_analyzed	minINFO	EffectDirection			
rs12184267	1:715265	1	715265	t	c	0.0408069	0.916	0.00688729787581148	0.007518884143899	0.3598	225955	0.805386	-???????????????++?			
rs12184277	1:715367	1	715367	a	g	0.9589313	-0.656	-0.00491449054466469	0.00749160144003763	0.5116	226215	0.808654	+???????????????--?			
rs12184279	1:717485	1	717485	a	c	0.0405759	1.05	0.00791160346381606	0.00753486044172958	0.2938	226224	0.807189	-???????????????++?			
rs116801199	1:720381	1	720381	t	g	0.042162	0.3	0.00221740320352237	0.00739134401174123	0.7644	226626	0.805329	-???????????????++?			
rs12565286	1:721290	1	721290	c	g	0.0423776	0.566	0.00417421538227414	0.00737493883794018	0.5711	226528	0.812657	-???????????????++?			
rs2977670	1:723891	1	723891	c	g	0.93688	-0.253	-0.00154984341088034	0.00612586328411202	0.8006	225312	0.836803	+???????????????--?			
rs28454925	1:726794	1	726794	c	g	0.9590545	-0.539	-0.00403872670010239	0.0074929994436037	0.5896	226782	0.809817	-???????????????--?			
rs116720794	1:729632	1	729632	t	c	0.0410995	0.27	0.00201855807601786	0.00747614102228837	0.7872	226989	0.809294	-???????????????++?			
rs4951859	1:729679	1	729679	c	g	0.183523	0.208	0.000805841472291884	0.00387423784755714	0.835	222312	0.725	--???????--?????+-?			

MDD

CHR	SNP	BP	A1	A2	FRQ_A_59851	FRQ_U_113154	INFO	OR	SE	P	ngt	Direction		HetISqt	HetDf	HetPVa	Nca	Nco	Neff
8	rs62513865	101592213	T	C	0.0747	0.0733	0.957	1.01461	0.0153	0.3438	0	---+++	0.0	5	0.7906	59851	113154	69115.85	
8	rs79643588	106973048	A	G	0.092	0.092	0.999	1.02122	0.0136	0.1231	0	++-+++	0.0	5	0.6847	59851	113154	69115.85	
8	rs17396518	108690829	T	G	0.562	0.565	0.98	1.00331	0.008	0.6821	6	--+--+	34.9	5	0.05637	59851	113154	69115.85	
8	rs6994300	102569817	A	G	0.00609	0.00556	0.466	0.88126	0.4243	0.7658	0	?-????	0.0	0	1	16823	25632	20313.61	
8	rs138449472	108580746	A	G	0.00965	0.00852	0.734	0.97181	0.0598	0.632	0	-+??-?	0.0	2	0.6663	41253	79756	47868.51	
8	rs983166	108681675	A	C	0.565	0.568	0.991	0.99144	0.008	0.2784	0	--+--+	10.1	5	0.1685	59851	113154	69115.85	
8	rs28842593	103044620	T	C	0.841	0.843	0.934	0.98926	0.0112	0.3381	1	-++--+	0.0	5	0.51	59851	113154	69115.85	
8	rs377046245	105176418	I	D	0.265	0.265	0.994	1.03004	0.0196	0.1311	-	?????-	0.0	0	1	14260	15480	14844.98	
8	rs7014597	104152280	C	G	0.166	0.164	0.996	1.01501	0.0106	0.1591	0	+-+-++	0.0	5	0.4818	59851	113154	69115.85	

Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia

Authors

The Schizophrenia Working Group of the Psychiatric Genomics Consortium

Corresponding Authors

Stephan Ripke[1,2], James TR Walters[3], Michael C O'Donovan[3]

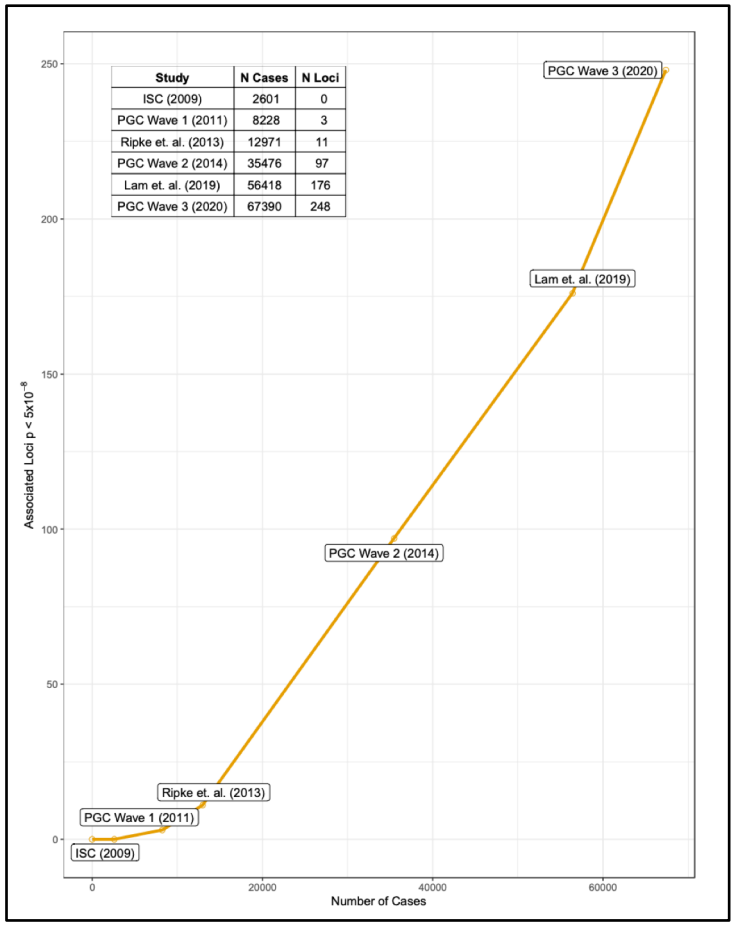
Affiliations

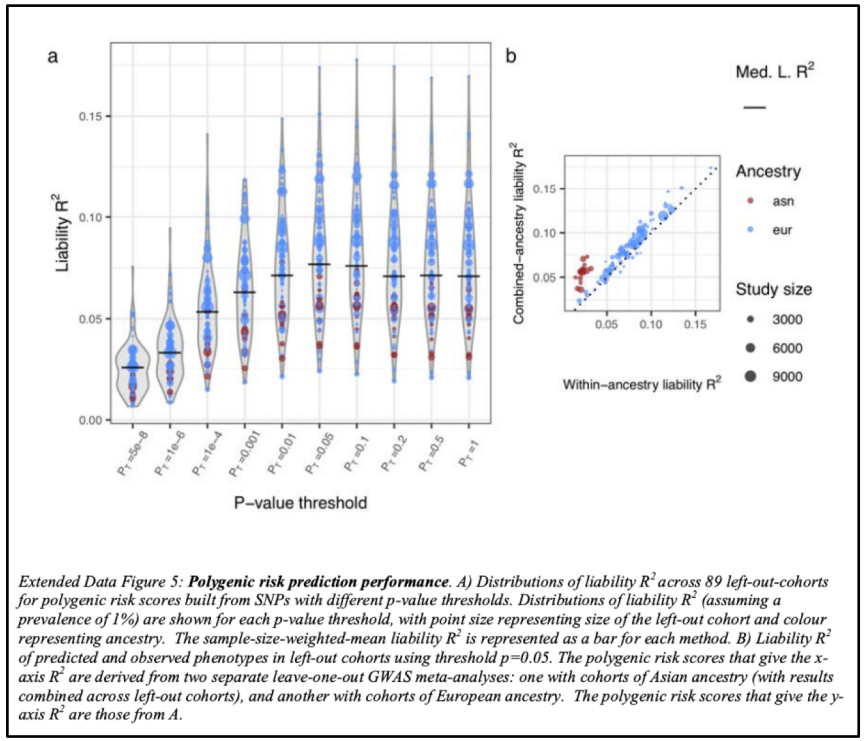
1. Dept. of Psychiatry and Psychotherapy, Charité - Universitätsmedizin, Berlin 10117, Germany
2. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston MA 02114, USA
3. MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychiatry and Clinical Neurosciences, Cardiff University, Hadyn Ellis Building, Maindy Road, Cardiff, CF24 4HQ

Contact

sripke@broadinstitute.org, waltersjt@cf.ac.uk, odonovanmc@cf.ac.uk

https://www.medrxiv.org/content/10.1101/2020.09.12.20192922v1.full.pdf





Extended Data Figure 5: ***Polygenic risk prediction performance****. A) Distributions of liability R^2 across 89 left-out-cohorts for polygenic risk scores built from SNPs with different p-value thresholds. Distributions of liability R^2 (assuming a prevalence of 1%) are shown for each p-value threshold, with point size representing size of the left-out cohort and colour representing ancestry. The sample-size-weighted-mean liability R^2 is represented as a bar for each method. B) Liability R^2 of predicted and observed phenotypes in left-out cohorts using threshold p=0.05. The polygenic risk scores that give the x-axis R^2 are derived from two separate leave-one-out GWAS meta-analyses: one with cohorts of Asian ancestry (with results combined across left-out cohorts), and another with cohorts of European ancestry. The polygenic risk scores that give the y-axis R^2 are those from A.*

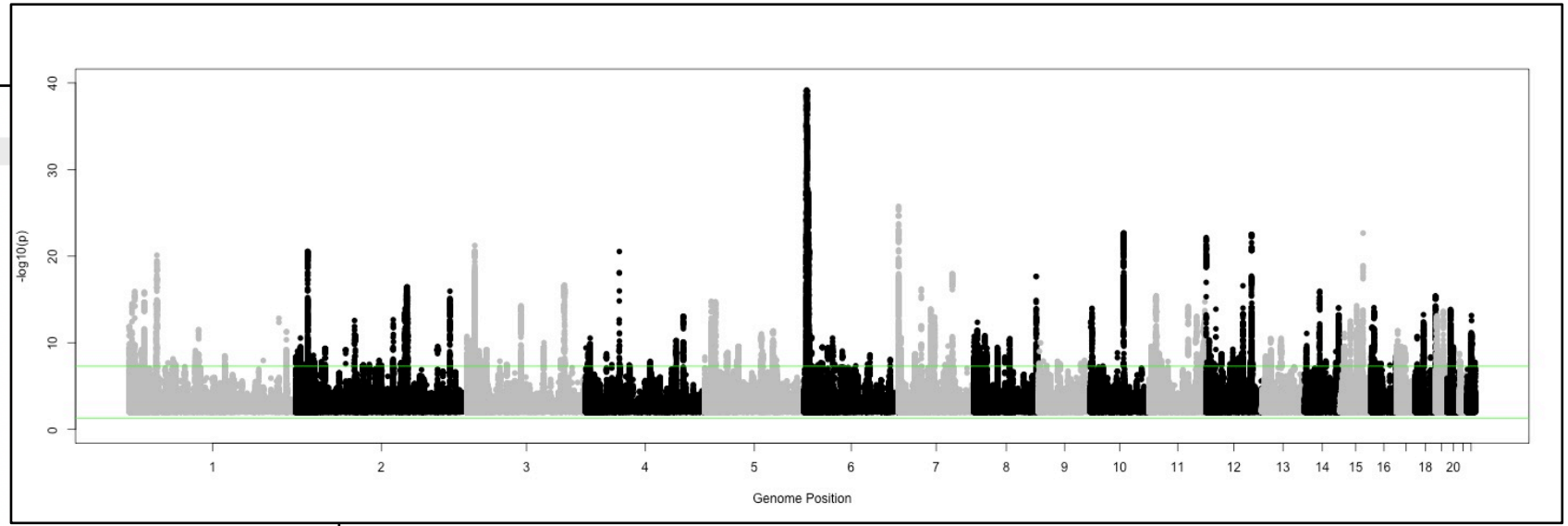
Most recent schizophrenia GWAS – still “unpublished”

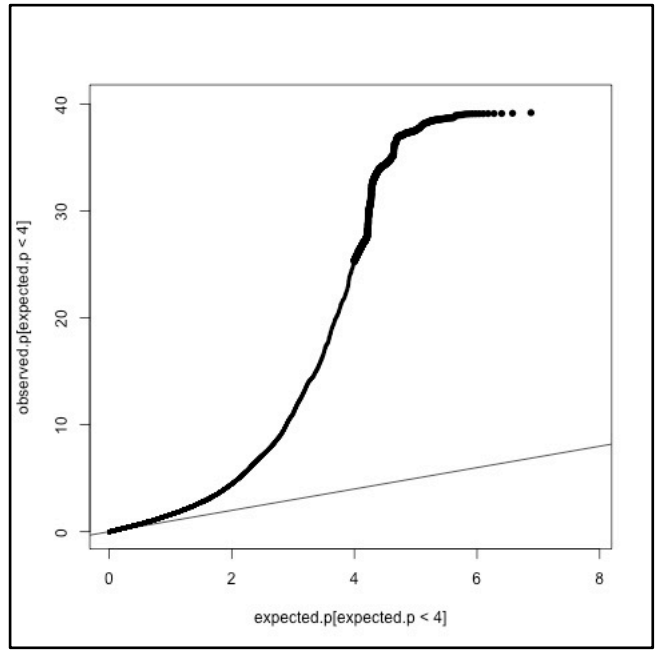
Typical quality steps

1. Visualize results for eye-ball tests
2. INFO scores
3. Highly missing
4. Overlap with target samples
5. Strand corrections

Visualize

```r
head( data )

data$CHR <- as.numeric( data$CHR )
data <- data[ !is.na( data$CHR ), ]
data <- data[ order( data$CHR, data$BP ), ]

head( data )

data$ManPos <- cumsum( data$BP/1000 )
chr.mids <- cbind( 1:22, NA )
for ( i in 1:22 ) {
     chr.start <- min( data$ManPos[ data$CHR == i ] )
     chr.stop <- max( data$ManPos[ data$CHR == i ] )
     chr.length <- chr.stop-chr.start
     chr.mids[ i,2 ] <- chr.start + chr.length/2
}

jpeg( "SCZ_QuickManhattan.jpeg", width=1440, height=480)
     plot( data$ManPos[ data$P < 0.01 ], -log10( data$P[ data$P < 0.01 ] ),
               xaxt='n',
               col=c( 'black','grey' )[ data$CHR[ data$P < 0.01 ] %% 2 + 1 ],
               xlab="Genome Position", ylab="-log10(p)",
               ylim=c(0,40),
               pch=16 )
     abline( h=-log10( 0.05 ), col='green' )
     abline( h=-log10( 5e-8 ), col='green' )
     axis( 1, at=chr.mids[,2], labels=chr.mids[,1] )
dev.off()

# Mini qq-plot
jpeg( "SCZ_QuickQQ.jpeg", width=480, height=480)
     observed.p <- -log10( data$P[ order( data$P ) ] )
     expected.p <- -log10( ( 1:length( data$P ) ) / ( length( data$P ) ) )
     plot( expected.p[ expected.p < 4 ], observed.p[ expected.p < 4 ],
               type='l',
               lwd=4,
               ylim=c( 0, max(observed.p)+1),
               xlim=c( 0, max(expected.p)+1) )
     points( expected.p[ expected.p > 4 ], observed.p[ expected.p > 4 ], pch=16 )
     abline( 0,1 )
dev.off()
```



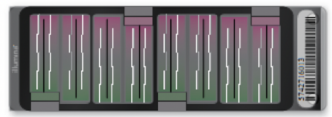


INFO

```
> data <- fread( "PGC3_SCZ_wave3_public.v2.tsv", fill=T )
|--------------------------------------------------|
|==================================================|
|--------------------------------------------------|
|==================================================|
> #data <- fread( "SCZ_raw_sample.txt" )
>
> dim( data )
[1] 7585077      19
> head( data )
   CHR        SNP        BP A1 A2 FRQ_A_67390 FRQ_U_94015  INFO      OR     SE      P ngt
1:   8 rs62513865 101592213  C  T       0.932       0.928 0.954 1.01349 0.0171 0.4319   0
2:   8 rs79643588 106973048  G  A       0.909       0.907 0.997 0.99084 0.0148 0.5361   0
3:   8 rs17396518 108690829  T  G       0.514       0.520 0.977 1.00060 0.0079 0.9409  16
4:   8   rs983166 108681675  A  C       0.547       0.549 0.983 1.00200 0.0078 0.7928   0
5:   8 rs28842593 103044620  T  C       0.862       0.860 0.935 0.99870 0.0116 0.9115   1
6:   8  rs7014597 104152280  G  C       0.845       0.840 0.986 1.00763 0.0117 0.5169   0
                                                                                Direction HetISqt
1: -?-+--+++???-+--+-+-++-++-+?????+++-+--+-??++-----++-------++++++-+-?+-+-++---++--+--++++     9.0
2: +?--+??-+???----+-++++++++?????++-+-+-++?++--+-----+----+++----++-+-?--++++++++++---+--+    21.3
3: ---+---+-+-+++++-++-+-+++-++-+++-+-++-++--++-+--+--+++++++--++++--+++++++--+++-++-----+-+     6.4
4: ---+---+-++---++-++-+-+---+--++--+-+--++-+++-+-+---++--++++++-++----+++++--+-++-++++---+++     0.0
5: +++++++--+++-----+++--++--++-+-----+--+-+-++-+---+++--+---++-+++--++-+-++++--+-++++-+-+-++     7.3
6: +?++-+---???---+--+-+--+--+?????++-+++--+??+---+++-+--+---+++-------+?+++---++--+-++++++++     0.0
   HetDf  HetPVa   Nca   Nco     Neff
1:    77 0.25900 55085 78957 60448.13
2:    76 0.05566 54376 78258 59744.11
3:    89 0.30900 67390 94015 73173.90
4:    89 0.57630 67390 94015 73173.90
5:    89 0.28700 67390 94015 73173.90
6:    77 0.90720 55085 78957 60448.13
>
```

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

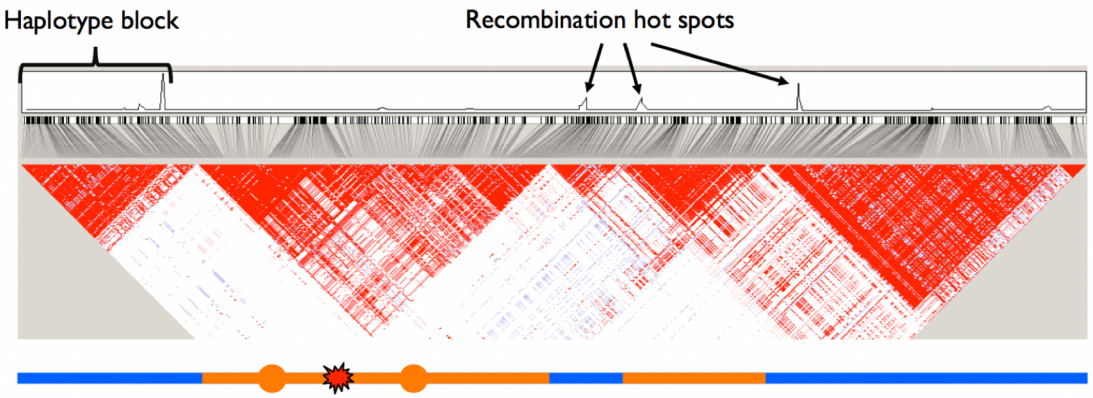
Common Variation Rare Mutations No Recorded Variation

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

Common Variation Mutations No Recorded Variation



- Red = regions of strong linkage; white = little or no linkage
- Haplotype blocks are on average 5–20 kilobases long

ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Genotype Imputation from Large Reference Panels

Sayantan Das,[1] Gonçalo R. Abecasis,[1] and Brian L. Browning[2]

[1]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, USA; email: sayantan@umich.edu, goncalo@umich.edu

[2]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington 98195-7720, USA; email: browning@uw.edu

Annu. Rev. Genom. Hum. Genet. 2018. 19:16.1–16.24

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-083117-021602

Keywords

imputation, genotype imputation, genome-wide association study, GWAS

Abstract

Genotype imputation has become a standard tool in genome-wide association studies because it enables researchers to inexpensively approximate whole-genome sequence data from genome-wide single-nucleotide polymorphism array data. Genotype imputation increases statistical power, facilitates fine mapping of causal variants, and plays a key role in meta-analyses of genome-wide association studies. Only variants that were previously observed in a reference panel of sequenced individuals can be imputed. However, the rapid increase in the number of deeply sequenced individuals will soon make it possible to assemble enormous reference panels that greatly increase the number of imputable variants. In this review, we present an overview of genotype imputation and describe the computational techniques that make it possible to impute genotypes from reference panels with millions of individuals.

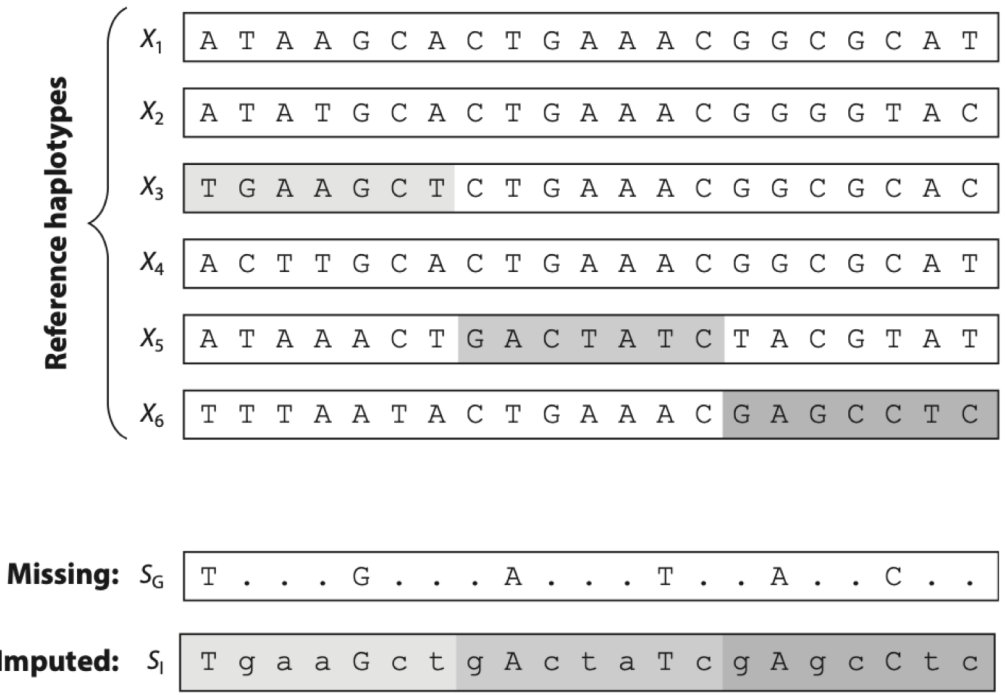
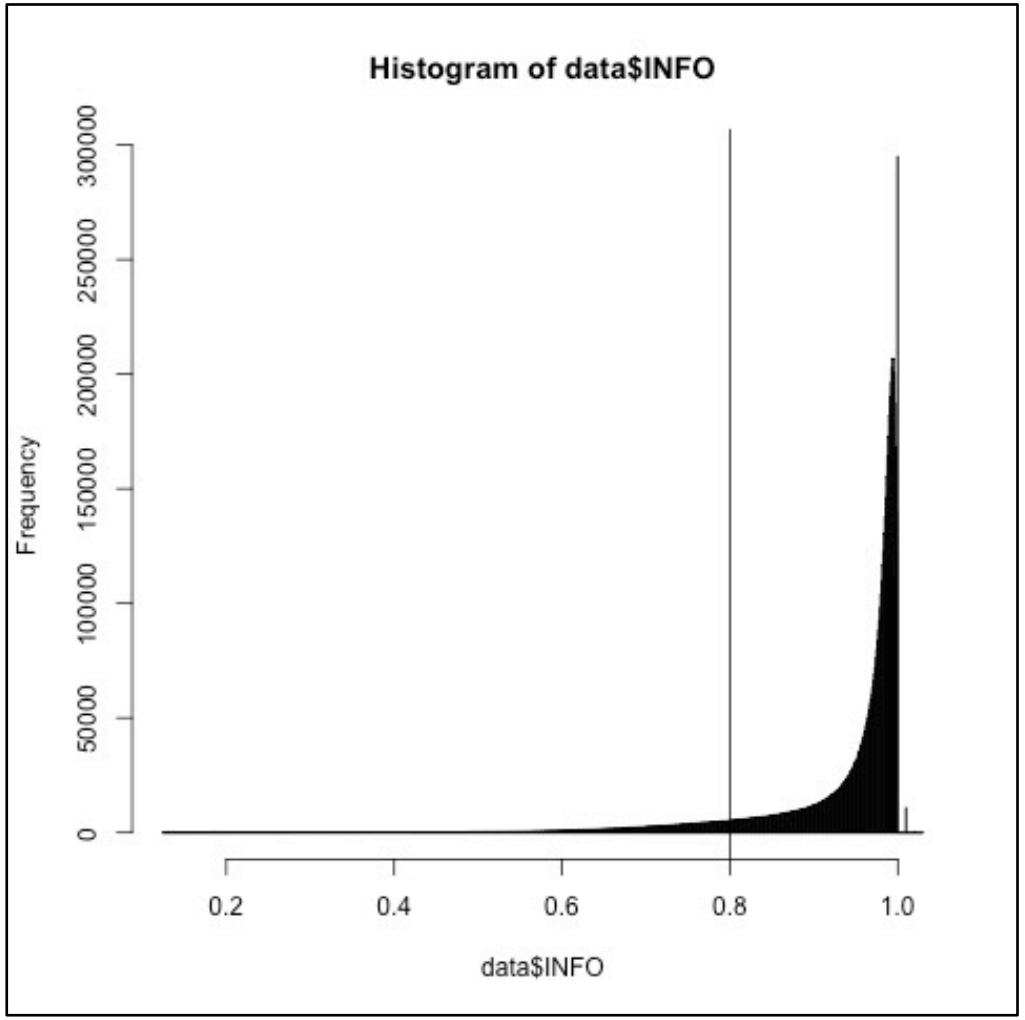


Figure 2

An illustration of genotype imputation, showing the process of imputation for a study haplotype (S_G) genotyped at 6 markers using a reference panel of sequenced haplotypes at 21 markers. The alleles in S_G are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype X_3. In the second segment, the alleles A and T imply that haplotype X_5 might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes X_3, X_5, and X_6. Consequently, the missing sites can be imputed to obtain the final imputed haplotype, S_I.

```
24  # Check INFO scores
25
26  summary( data$INFO )
27  sum( data$INFO < 0.8 )
28
29  jpeg( "SCZ_hist_INFO.jpeg", width= 480, height=480)
30      hist( data$INFO, breaks='fd' )
31      abline( v=0.8 )
32  dev.off()
33
```

```
> summary( data$INFO )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1240  0.9300  0.9770  0.9381  0.9910  1.0300
> sum( data$INFO < 0.8 )
[1] 654569
>
> jpeg( "SCZ_hist_INFO.jpeg", width= 480, height=480)
>  hist( data$INFO, breaks='fd' )
>  abline( v=0.8 )
> dev.off()
null device
          1
>
```



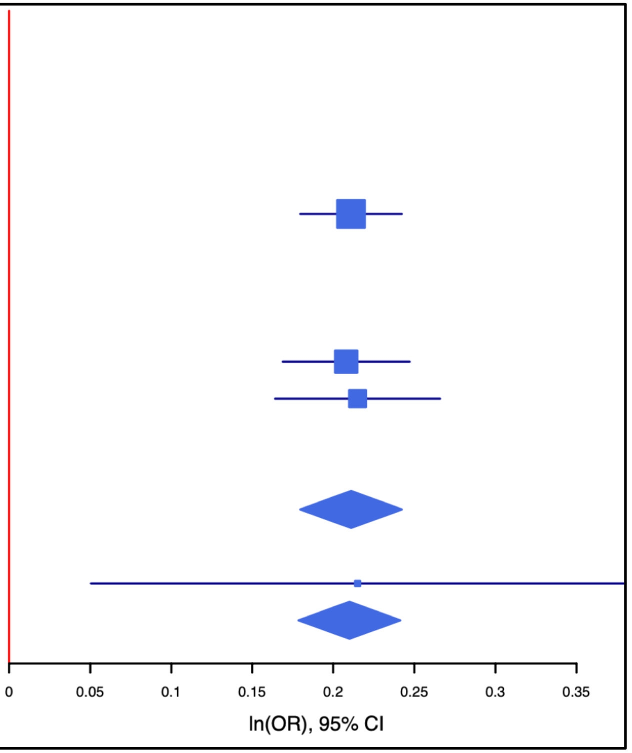
Missing

```
> head( data[ , c( 2,12,13,17,18,19,11 ) ] )
          SNP ngt                                                                                    Direction   Nca   Nco     Neff      P
1:  rs12238997   0 -+++++---+-++++-+++--+++++-+--++--+-++-++-+----+-+--+++--++--?--+??++---+++-?----+--+----+ 65792 91818 71533.14 0.7324
2:  rs55727773   0 ---+---+------++---+-+----+?-+?++--++-+----+++-+++-+++--++--+?---??+++++-+--?++-+-++-+++-- 64769 89260 70219.36 0.8476
3: rs149887893   0 -+-+++-+--++++??+++-+--??--++-+-+-+?-++??-+-+----+??--+---+?-?---??-----+?+-?++-++----+?-+ 54602 69851 58130.46 0.2294
4:  rs12184267   0 +-+---+-+??---??---+-++??++-??-?-+-?+-+??+++-++++-??++-+++-?+?+++??+++++-?-+?--+--++-++?+- 50225 64218 53377.68 0.2097
5:  rs12184277   0 -+-+++-+-??+++??+++-+--??+-+??+?+-+?-+-??---+----+??--+---+?-?---??-----+?+-?++-++--+--?-+ 50225 64218 53377.68 0.2335
6:  rs12184279   0 +-+---+-+??---??---+-++??++-??-?-+-?+-+??+++-++++-??++-+++-?+?+++??+++++-?-+?--+--++-++?+- 50225 64218 53377.68 0.2564
>
> head( data[ order(data$P), c( 2,12,13,17,18,19,11 ) ] )
          SNP ngt                                                                                    Direction   Nca   Nco     Neff         P
1:  rs13195636   0 +?+++??--???-+++-++++-++++-?????+++++-+++??-+++++++-+++++++++-+++++++?++++++++-+++++++++++ 53386 77258 58749.13 6.546e-40
2:  rs10484399  50 +?+++??--???-+++-++++-++++-?????+++++-+++??--++++++-+++++++++-+++++++?++++++++-+++++++++++ 53386 77258 58749.13 7.369e-40
3:  rs35501037   1 -?---??++???+---+----+----+???-?-----+-+-??+-------+---------+-------?--------+---------- 53933 77798 59292.61 7.726e-40
4: rs140365013   0 -?---??++???+---+----+----+?????-----+---??++------+---------+-------?--------+---------- 53386 77258 58749.13 7.856e-40
5:  rs66785117   0 +?+++??--???-+++-++++-++++-?????+++++-+++??--++++++-+++++++++-+++++++?++++++++-+++++++++++ 53386 77258 58749.13 7.861e-40
6:  rs34965299   0 +?+++??--???-+++-++++-++++-?????+++++-+++??--++++++-+++++++++-+++++++?++++++++-+++++++++++ 53386 77258 58749.13 8.070e-40
```

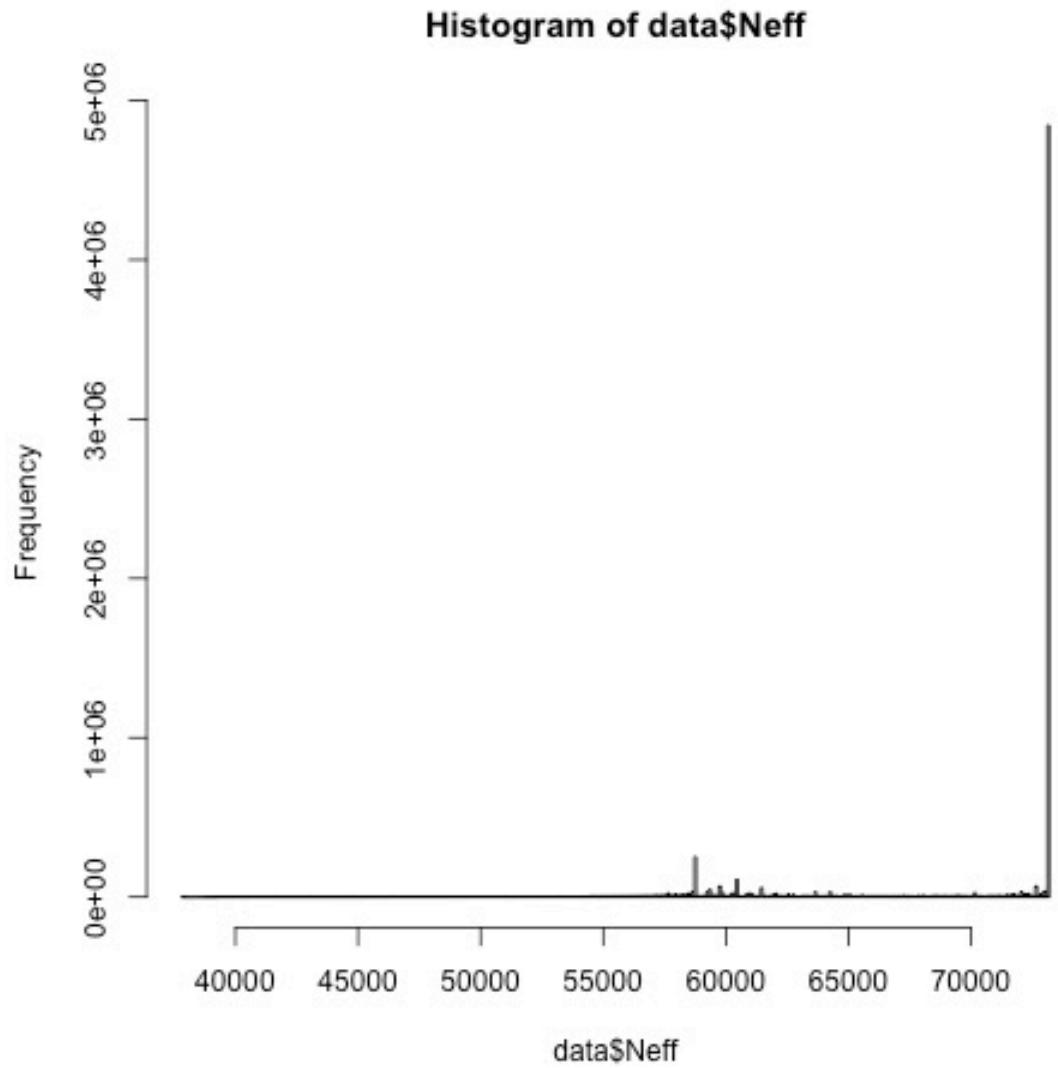
rs13195636 **A/C** **6:27509493**

++ het_P: 0.828 het_I: 0.0

	ngt	info	p_value	f_ca(n)	f_co(n)	ln(OR)	STDerr
PGC_SCZ wave3 eur only	0	1.00	6.55e−40	0.926(53386)	0.909(77258)	0.211	0.016
PGC_SCZ wave3 asn only	nan	nan	nan	nan(nan)	nan(nan)	nan	nan
PGC_SCZ wave2 only	0	1.00	1.46e−24	0.925(32998)	0.912(43387)	0.208	0.02
PGC_SCZ wave3 (excl. wave2)	0	0.99	5.28e−17	0.928(20388)	0.906(33871)	0.215	0.026
PGC_SCZ wave3 (incl. wave2)	**0**	**1.00**	**6.55e−40**	**0.926(53386)**	**0.909(77258)**	**0.211**	**0.016**
Decode – Replication	0	1.00	0.0164	0.938(1979)	0.932(142627)	0.215	0.084
PGC_SCZ wave3 (incl. wave2 and replication)	**0**	**1.00**	**2.9e−41**	**0.926(55365)**	**0.924(219885)**	**0.21**	**0.016**



```
> summary( data$Neff )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  37841   64415   73174   68827   73174   73174
>
> jpeg( "SCZ_hist_Neff.jpeg", width= 480, height=480)
>  hist( data$Neff, breaks='fd' )
>  abline( v=max(data$Neff)/2 )
> dev.off()
null device
          1
>
> sum( data$Neff < max(data$Neff)/2 )
[1] 0
>
>
```



Overlap

```
> tped <- fread( "AndrewSchork.tped" )
> names( tped ) <- c( 'CHR', 'SNP', 'cM', 'BP', 'AJS_A1', 'AJS_A2' )
> tfam <- fread( "AndrewSchork.tfam" )
> names( tfam ) <- c( 'FID', 'IID', 'PID', 'MID', 'SEX', 'PHENO' )
>
> dim( tped )
[1] 601886      6
> head( tped )
   CHR         SNP cM     BP AJS_A1 AJS_A2
1:   1  rs12564807  0 734462      A      A
2:   1   rs3131972  0 752721      G      G
3:   1 rs148828841  0 760998      C      C
4:   1  rs12124819  0 776546      A      A
5:   1 rs115093905  0 787173      G      G
6:   1  rs11240777  0 798959      G      G
>
> dim( tfam )
[1] 1 6
> head( tfam )
      FID    IID       PID        MID SEX        PHENO
1: Schork Andrew Nicholas Catherine   1 OneKewlDood
>
>
```

```
>
> ## Merge SNP info from subject data with GWAS data
>
> tped_new <- merge( tped, data[ ,c( 1,2,3,4,5,9,11 ) ], by="SNP" )
>
> dim( tped_new )
[1] 479797     12
> head( tped_new )
          SNP CHR.x cM      BP.x AJS_A1 AJS_A2 CHR.y      BP.y A1 A2      OR       P
1:  rs1000000    12  0 126890980      G      A    12 126890980  G  A 1.01888 0.05482
2: rs10000023     4  0  95733906      T      G     4  95733906  G  T 0.98393 0.03913
3: rs10000030     4  0 103374154      G      G     4 103374154  G  A 1.01126 0.30980
4: rs10000041     4  0 165621955      T      T     4 165621955  G  T 1.00833 0.43430
5: rs10000042     4  0   5237152      C      C     4   5237152  C  T 1.03956 0.01375
6:  rs1000007     2  0 237752054      T      C     2 237752054  T  C 1.00110 0.90230
>
> #mis-mapped positions
> tped_new[ tped_new$BP.x != tped_new$BP.y, ]
          SNP CHR.x cM     BP.x AJS_A1 AJS_A2 CHR.y     BP.y A1 A2      OR       P
1: rs61774271     5  0 69375211      A      A     1 95542929  T  C 1.04907 0.006703
>
> #mis-mapped alleles
> head( tped_new[ tped_new$AJS_A1 != tped_new$A1 & tped_new$AJS_A1 != tped_new$A2, ] )
          SNP CHR.x cM      BP.x AJS_A1 AJS_A2 CHR.y      BP.y A1 A2      OR       P
1:  rs5742913     5  0 133451683      T      C     5 133451683  C  A 1.00170 0.910300
2: rs61774271     5  0  69375211      A      A     1  95542929  T  C 1.04907 0.006703
>
>
```

The Human Genome Project

-publically funded sequencing of a representative genome
-Began in 1990 and was “completed” in 2001
-It cost 3 billion dollars

Gave us a reference genome:

-A representative example of a human genome
-An address book for the genome, its variability and its functional elements

Francis Collins

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

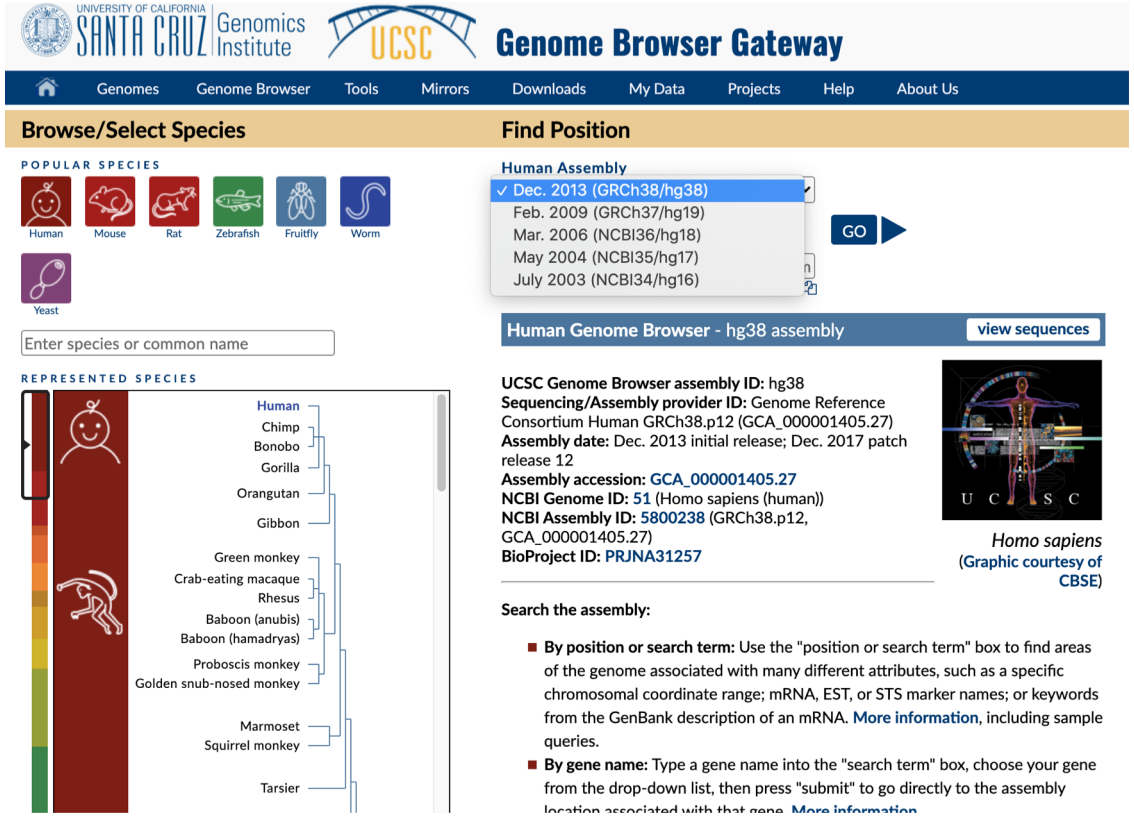
Craig Venter

THE HUMAN GENOME

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[4] Andrew G. Clark,[5] Joe Nadeau,[6] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weinia Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Allen Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[14] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Kerstelen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[15,16] Roderic Guigó,[15,16] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anuhya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcela Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

http://genome-euro.ucsc.edu/



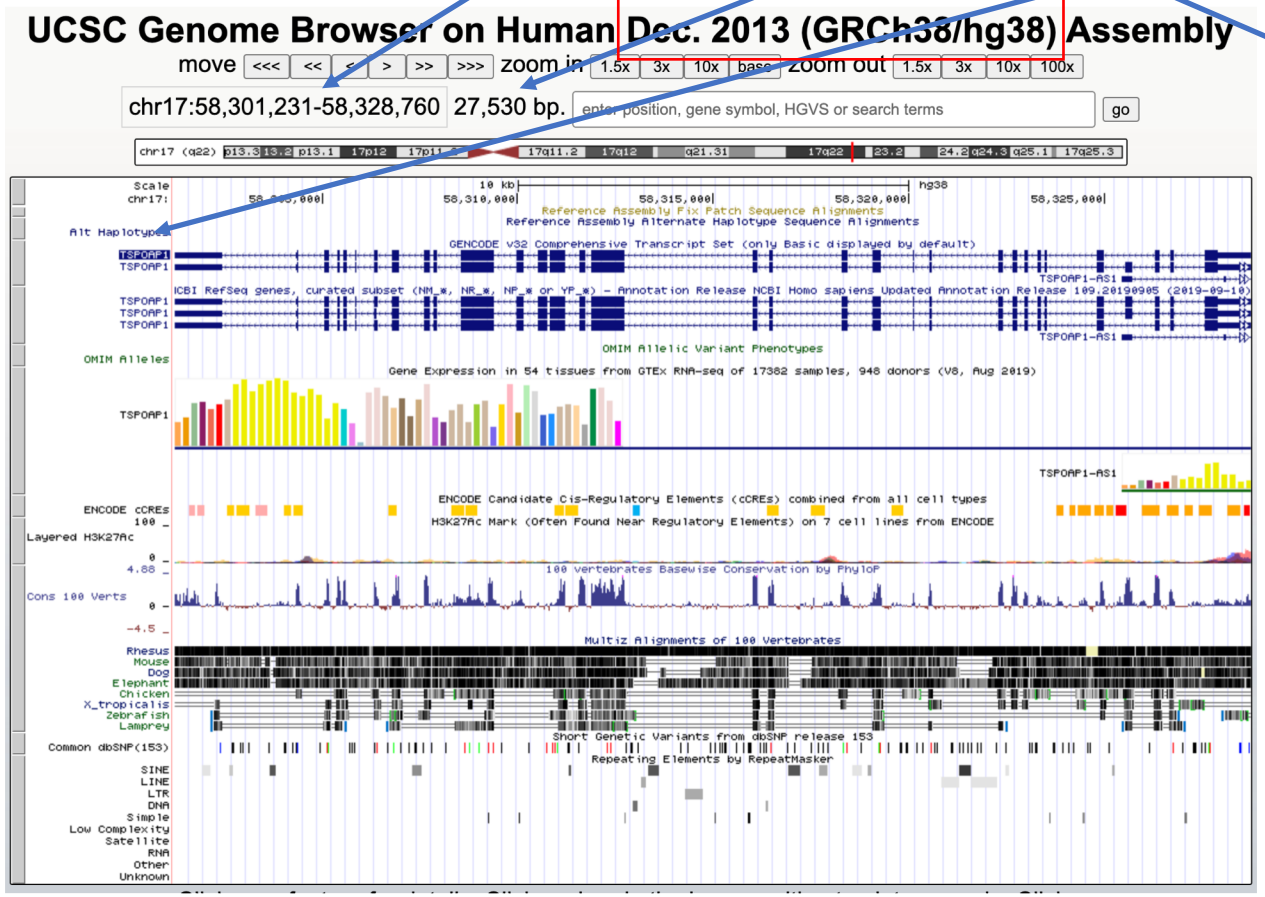
Genome build

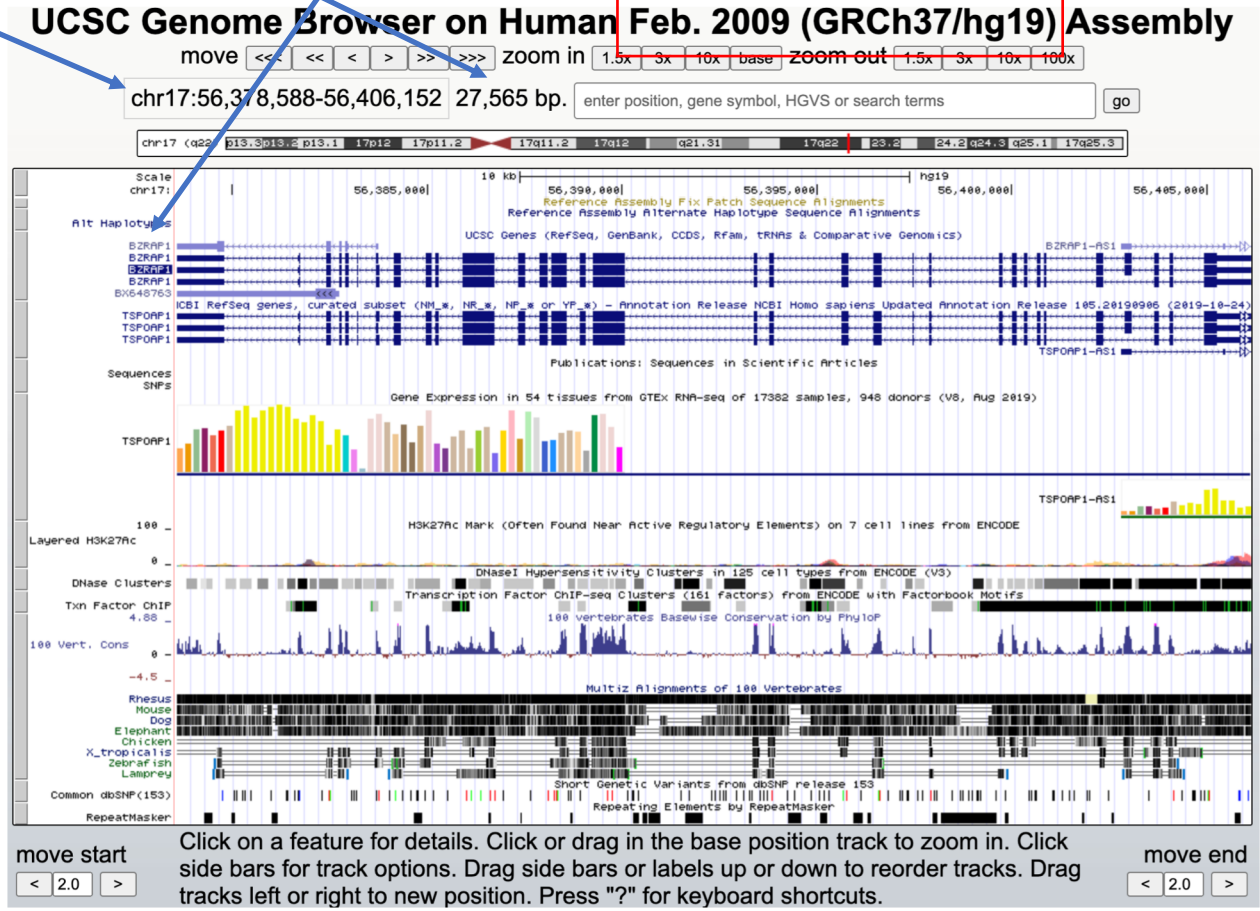
As stated in the base data section, the genome build for our base and target data is the same, as it should be.

Position changes

Length changes

Gene name changes



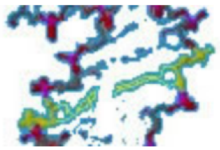


hg38 (human genome version 38)

hg19 (human genome version 19)

NCBI

dbSNP
Short Genetic Variations



dbVar | ClinVar | GaP | PubMed | Nucleotide | Protein

Search small variations in dbSNP or large structural variations in dbVar

Search Entrez dbSNP for [ ] Go

Have a question about dbSNP? Try searching the SNP FAQ Archive!

[ ] Go

GENERAL
- RSS Feed
- Contact Us
- Organism Data
- dbSNP Homepage
- NCBI Variation Resources
- Announcements
- dbSNP Summary
- FTP Download

SNP SUBMISSION

DOCUMENTATION

SEARCH

RELATED SITES

dbSNP Summary

RELEASE: NCBI dbSNP Build 155

dbSNP Component Availability Dates:

Component	Date available
dbSNP web query for build 155:	Jun 16, 2021
ftp data for build 155:	Jun 16, 2021
Entrez Indexing for build 155:	Jun 16, 2021

- The complete data for build 155 are available at https://ftp.ncbi.nlm.nih.gov/snp/ in multiple formats.
- All formats and conventions are described in https://ftp.ncbi.nlm.nih.gov/snp/00readme.txt.

- Please address any questions or comments regarding the data to snp-admin@ncbi.nlm.nih.gov.

New Submission since previous build:

Organism	Current Build	New Submissions (ss#'s)	New RefSNP Clusters (rs#'s) (# validated)
Total: Organisms			()

**Submissions received after reclustering of current build will appear as new rs# clusters in the next build.*

Assigns rs numbers which may merge or move position with better data and new reference builds

Strand corrections - the real pain in the a$$

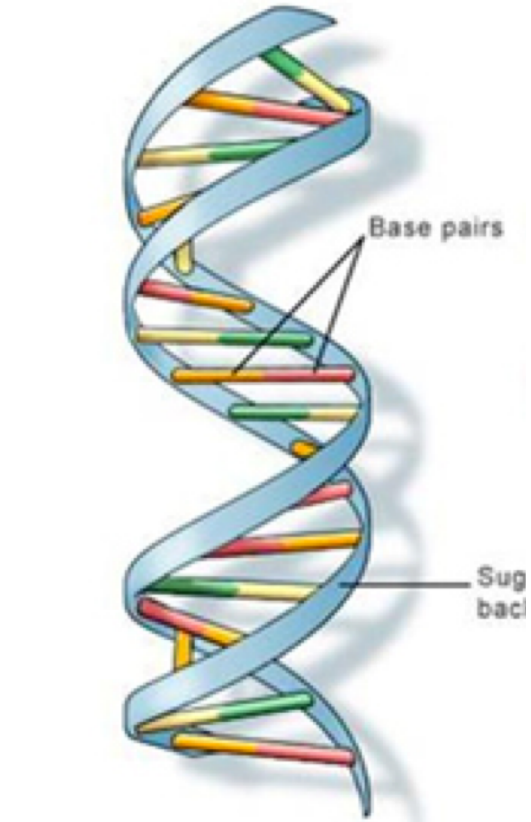
```
> ## strand ambiguity
> head( tped_new[ tped_new$A1 == 'A' & tped_new$A2 == 'T', ] )
          SNP CHR.x cM      BP.x AJS_A1 AJS_A2 CHR.y      BP.y A1 A2      OR          P
1:  rs1002229    11  0  32053007      T      A    11  32053007  A  T 0.99432 6.234e-01
2: rs10127775     1  0 230295789      T      T     1 230295789  A  T 0.99293 3.789e-01
3:  rs1013442    11  0  27578946      A      A    11  27578946  A  T 1.03873 4.773e-05
4: rs10140401    14  0  24888500      A      A    14  24888500  A  T 0.97707 1.399e-01
5: rs10203853     2  0 234687418      T      A     2 234687418  A  T 0.98886 1.605e-01
6:  rs1023568     2  0 191515442      T      A     2 191515442  A  T 0.98758 1.054e-01
> head( tped_new[ tped_new$A1 == 'C' & tped_new$A2 == 'G', ] )
          SNP CHR.x cM      BP.x AJS_A1 AJS_A2 CHR.y      BP.y A1 A2      OR       P
1:  rs1000735     9  0  93654616      G      C     9  93654616  C  G 1.00702 0.39270
2:  rs1001979     6  0 166585625      C      C     6 166585625  C  G 1.00401 0.73520
3: rs10028213     4  0 149652610      G      C     4 149652610  C  G 1.01217 0.24890
4:  rs1003582     6  0  29538403      G      G     6  29538403  C  G 0.98462 0.06801
5:  rs1005190    16  0    855717      G      C    16    855717  C  G 0.98876 0.21290
6:  rs1009592     1  0  11928714      C      C     1  11928714  C  G 1.01187 0.15940
>
>
```

DNA is **D**eoxyribo**N**ucleic **A**cid

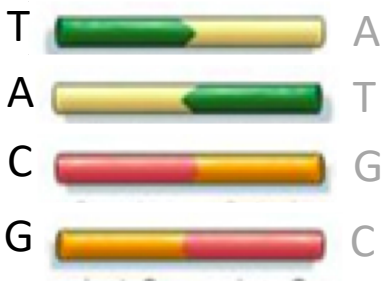
3.2 billion nucleobase pairs embedded in a deoxyribose sugar-phosphate backbone

4 nucleobases:

- **A**denine
- **T**hymine
- **C**ytosine
- **G**uanine



Four base-pair patterns

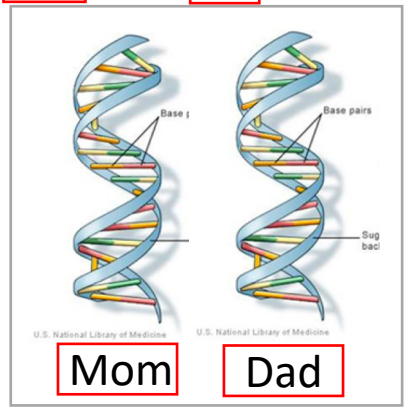


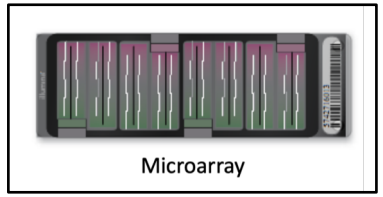
T	-	A	T	-	A
G	-	C	C	-	G
A	-	T	A	-	T
G	-	C	T	-	A
A	-	T	C	-	G
G	-	C	G	-	C
T	-	A	T	-	A
C	-	G	A	-	T
A	-	T	A	-	T
C	-	G	T	-	A
G	-	C	G	-	C
C	-	G	G	-	C
T	-	A	T	-	A
C	-	G	A	-	T
G	-	C	A	-	T
T	-	A	T	-	A
C	-	G	C	-	G

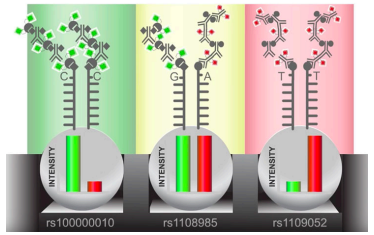
The DNA in any cell is over 2 meters long! (only 2 *nano*meters wide)

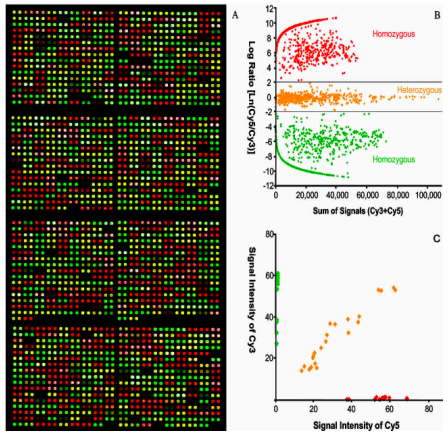
This is ~200,000 times longer than the diameter of the nucleus.

The DNA in your body would reach to the sun and back ~500 times









We can resolve this case, because if one study shows G/A we know to exchange G to C, and A to T

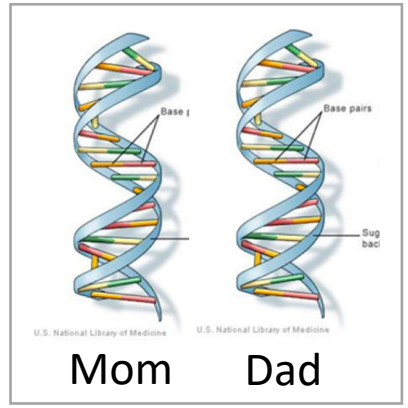
G / A

G			A
C	G	G	C
G	C	C	G
A	T	T	A
T	A	A	T
T	A	A	T
A	T	T	A
G	C	C	G

C / T

C			T
G	C	C	G
C	G	G	C
T	A	A	T
A	T	T	A
A	T	T	A
T	A	A	T
C	G	G	C

T - A	T - A
G - C	C - G
A - T	A - T
G - C	T - A
A - T	C - G
G - C	G - C
T - A	T - A
C - G	A - T
A - T	A - T
C - G	T - A
G - C	G - C
C - G	C - G
T - A	T - A
A - T	A - T
A - T	A - T
T - A	T - A
C - G	C - G

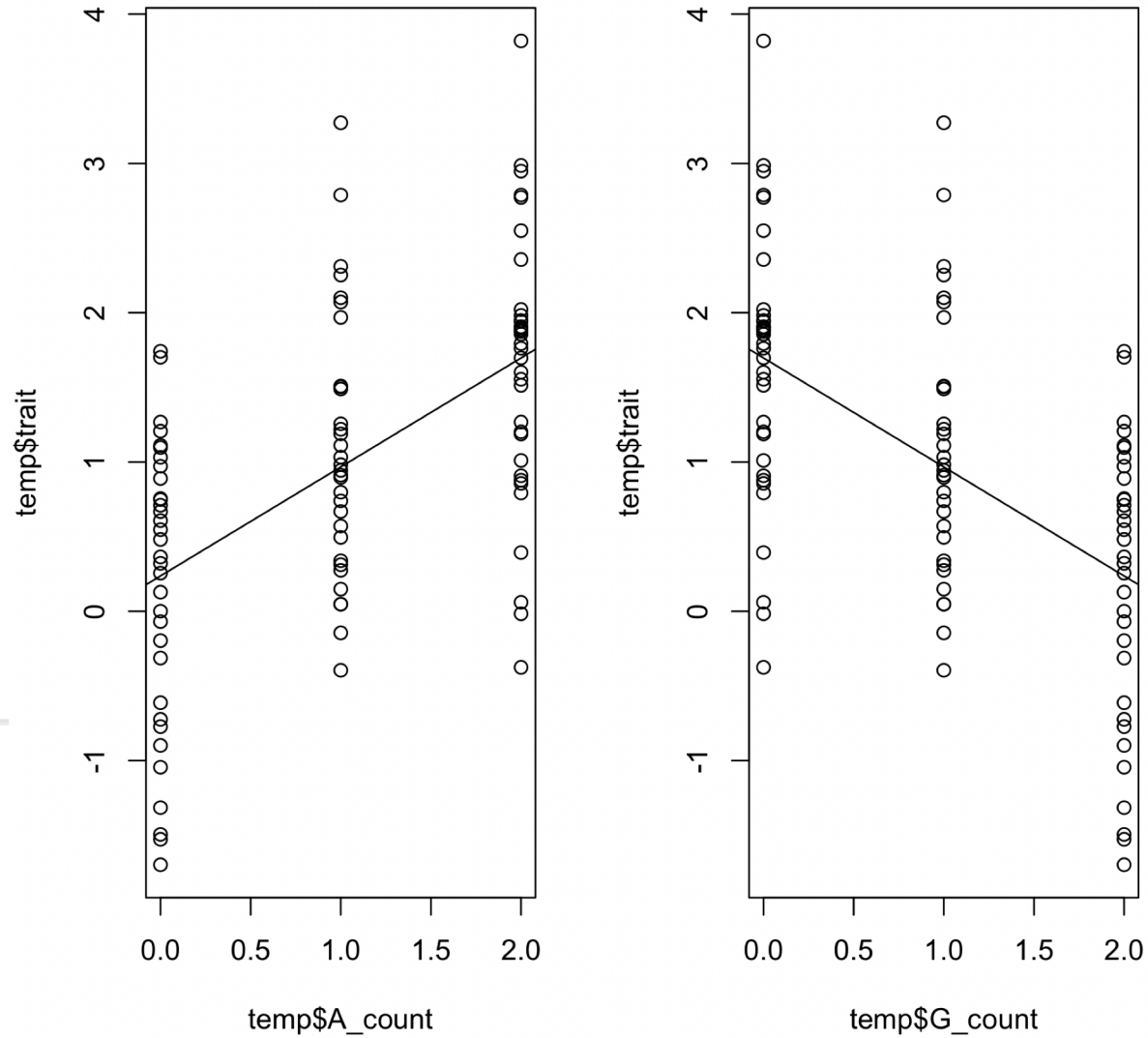


```
> ## strand ambiguity
> head( tped_new[ tped_new$A1 == 'A' & tped_new$A2 == 'T', ] )
          SNP CHR.x cM      BP.x AJS_A1 AJS_A2 CHR.y      BP.y A1 A2      OR          P
1:  rs1002229    11  0  32053007      T      A    11  32053007  A  T 0.99432 6.234e-01
2: rs10127775     1  0 230295789      T      T     1 230295789  A  T 0.99293 3.789e-01
3:  rs1013442    11  0  27                                3946  A  T 1.03873 4.773e-05
4: rs10140401    14  0  24                                3500  A  T 0.97707 1.399e-01
5: rs10203853     2  0 234                                7418  A  T 0.98886 1.605e-01
6:  rs1023568     2  0 191                                5442  A  T 0.98758 1.054e-01
> head( tped_new[ tped_new                                  ] )
          SNP CHR.x cM                                    BP.y A1 A2      OR       P
1:  rs1000735     9  0  93                                4616  C  G 1.00702 0.39270
2:  rs1001979     6  0 166                                5625  C  G 1.00401 0.73520
3: rs10028213     4  0 149                                2610  C  G 1.01217 0.24890
4:  rs1003582     6  0 29                                 3403  C  G 0.98462 0.06801
5:  rs1005190    16  0   855717      G      C    16    855717  C  G 0.98876 0.21290
6:  rs1009592     1  0 11928714      C      C     1  11928714  C  G 1.01187 0.15940
>
>
```

```
> # Simulate a genotype and a trait that has a genotype effect built in
>
> temp <- data.table( sample( c('AA', 'AG', 'GG'), 100, replace=T ) )
> temp$trait <- rnorm( 100 )
> temp$trait[ temp[[1]] == 'AG' ] <- temp$trait[ temp[[1]] == 'AG' ] + 1
> temp$trait[ temp[[1]] == 'AA' ] <- temp$trait[ temp[[1]] == 'AA' ] + 2
>
> # What is the 'true' genotype effect?  which allele and how much?  is it additive or dominant?
>
> ## Estimate the effect
> # additive count of A
> temp$A_count <- 0
> temp$A_count[ temp[[1]] == 'AG' ] <- 1
> temp$A_count[ temp[[1]] == 'AA' ] <- 2
>
> # additive count of G
> temp$G_count <- 0
> temp$G_count[ temp[[1]] == 'AG' ] <- 1
> temp$G_count[ temp[[1]] == 'GG' ] <- 2
>
> head( temp )
   V1      trait A_count G_count
1: GG -0.2473098       0       2
2: AA  2.5159164       2       0
3: AA  3.2711098       2       0
4: AG  0.8106688       1       1
5: AG  1.9287540       1       1
6: AA  0.6932917       2       0
```

```
> table( temp$A_count, temp$G_count )

    0  1  2
  0  0  0 35
  1  0 31  0
  2 34  0  0
>
> a.lm <- lm( trait ~ A_count, data=temp )
> summary( a.lm )$coefficient
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -0.1164449  0.1372122 -0.8486483 3.981452e-01
A_count      1.0763041  0.1061780 10.1367887 6.110571e-17
> g.lm <- lm( trait ~ G_count, data=temp )
> summary( g.lm )$coefficient
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  2.036163  0.1388458  14.66493 1.858291e-26
G_count     -1.076304  0.1061780 -10.13679 6.110571e-17
>
>
```

```
> par( mfrow=c(1,2) )
> plot( temp$A_count, temp$trait )
> abline( a.lm )
> plot( temp$G_count, temp$trait )
> abline( g.lm )
>
>
>
```



```
> sum( ( data$A1 == 'A' & data$A2 == 'T') )
[1] 0
> sum( ( data$A1 == 'T' & data$A2 == 'A') )
[1] 0
> sum( ( data$A1 == 'C' & data$A2 == 'G') )
[1] 0
> sum( ( data$A1 == 'G' & data$A2 == 'C') )
[1] 0
>
> data <- data[ !( data$A1 == 'A' & data$A2 == 'T'), ]
> data <- data[ !( data$A1 == 'T' & data$A2 == 'A'), ]
> data <- data[ !( data$A1 == 'C' & data$A2 == 'G'), ]
> data <- data[ !( data$A1 == 'G' & data$A2 == 'C'), ]
>
> dim( data )
[1] 5878401      19
> data.size
[1] 7585077      19
>
```

1. QC Base Data (i.e., GWAS results - betas)

2. QC Target Data (i.e., Genotype Data)

3. Calculate PRS

4. Visualize Results

Typical quality steps

1. Last time:

   Missing data, minor allele frequency, Hardy-Weinberg checks

2. But also:

   INFO scores, strand ambiguous SNPs

3. Good merge alignment / batch effects
4. Good ancestry alignment / reference selection

   Does my target fit these reference data?

   Is there a good ancestry reference?

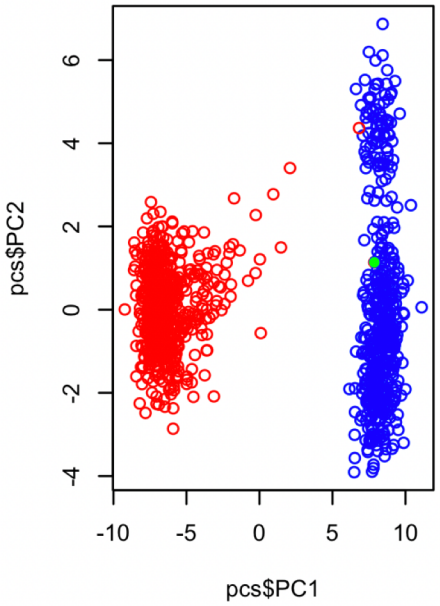
Merge target and reference data

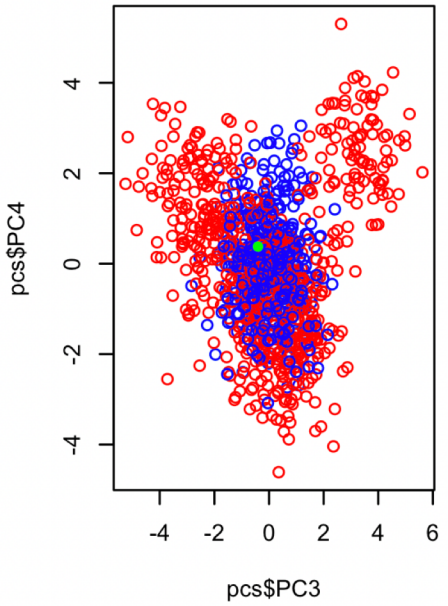
```
> # Load target individual
> ajs <- fread( 'AndrewSchork.traw' )
>
> dim( ajs )
[1] 65397     7
> head( ajs )
   CHR        SNP (C)M     POS COUNTED ALT Schork_Andrew
1:   1  rs9442373    0 1062638       C   0             2
2:   1  rs3813199    0 1158277       A 0,G             0
3:   1 rs17160669    0 1305561       T 0,C             0
4:   1  rs7525092    0 1810090       T   C             1
5:   1   rs424079    0 2071340       C 0,A             0
6:   1   rs884940    0 2223866       T 0,C             0
>
> # Load Reference individuals
> kgp <- fread( 'KGP_GSA_PGS.traw' )
>
> dim( kgp )
[1] 65397  1170
> head( kgp[,1:10] )
   CHR        SNP (C)M     POS COUNTED ALT EUR_HG00096 EUR_HG00097 EUR_HG00099 EUR_HG00100
1:   1  rs9442373    0 1062638       C   A           1           1           0           1
2:   1  rs3813199    0 1158277       A   G           1           0           0           0
3:   1 rs17160669    0 1305561       T   C           1           0           0           0
4:   1  rs7525092    0 1810090       T   C           1           0           1           2
5:   1   rs424079    0 2071340       C   A           1           0           1           1
6:   1   rs884940    0 2223866       T   C           1           1           2           0
>
> # Combine data
> tped <- cbind( ajs, kgp[ ,7:1170 ] )
>
> dim( tped )
[1] 65397  1171
> head( tped[,1:10] )
   CHR        SNP (C)M     POS COUNTED ALT Schork_Andrew EUR_HG00096 EUR_HG00097 EUR_HG00099
1:   1  rs9442373    0 1062638       C   0             2           1           1           0
2:   1  rs3813199    0 1158277       A 0,G             0           1           0           0
3:   1 rs17160669    0 1305561       T 0,C             0           1           0           0
4:   1  rs7525092    0 1810090       T   C             1           1           0           1
5:   1   rs424079    0 2071340       C 0,A             0           1           0           1
6:   1   rs884940    0 2223866       T 0,C             0           1           1           2
>
>
```

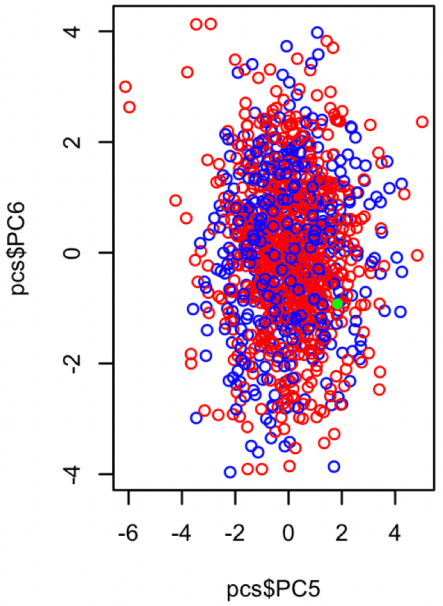
```
213 ## Assess fit to reference to test subject
214 # Get a smaller set of SNPs for computational speed
215 tped.small <- tped[ sample( dim(tped)[1], 2000 ) ,7:1171 ]
216
217 # Compute PCs to check ancestry
218 pca <- prcomp( t( tped.small[ ,7:1171 ] ) )
219 pcs <- cbind( do.call( rbind, strsplit( names(tped)[ 7:1171 ], split="_" ) ),
220                   pca$x[ ,1:25 ] )
221 pcs <- data.table( pcs )
222
```

Compute ancestry

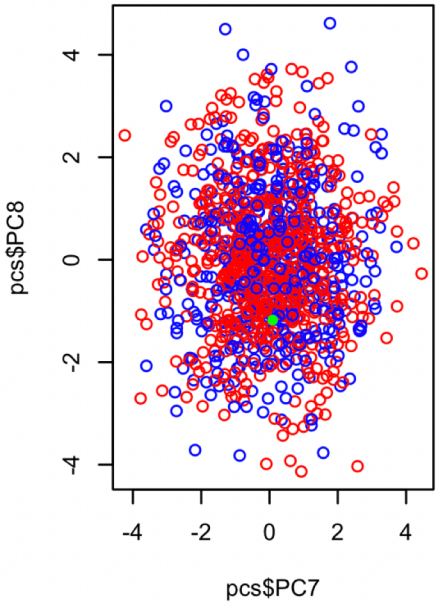
```r
# Plot reference and test subject ancestry PCAs
par( mfrow=c(2,3) )
plot( pcs$PC1, pcs$PC2,
        col=c( "red", "blue" )[ 1 + 1*(pcs[[1]] == 'EUR') ] )
points( pcs$PC1[1], pcs$PC2[1], col="green", pch=16 )
plot( pcs$PC3, pcs$PC4,
        col=c( "red", "blue" )[ 1 + 1*(pcs[[1]] == 'EUR') ] )
points( pcs$PC3[1], pcs$PC4[1], col="green", pch=16 )
plot( pcs$PC5, pcs$PC6,
        col=c( "red", "blue" )[ 1 + 1*(pcs[[1]] == 'EUR') ] )
points( pcs$PC5[1], pcs$PC6[1], col="green", pch=16 )
plot( pcs$PC7, pcs$PC8,
        col=c( "red", "blue" )[ 1 + 1*(pcs[[1]] == 'EUR') ] )
points( pcs$PC7[1], pcs$PC8[1], col="green", pch=16 )
plot( pcs$PC9, pcs$PC10,
        col=c( "red", "blue" )[ 1 + 1*(pcs[[1]] == 'EUR') ] )
points( pcs$PC9[1], pcs$PC10[1], col="green", pch=16 )
plot(0,0,type='n')
legend( "center", legend=c("AFR", "EUR", "KewlD00D" ), col=c( "red", "blue", "green" ), pch=c(1,1,16) )

# Looks ok - Falls where expected, no obvious bias
```

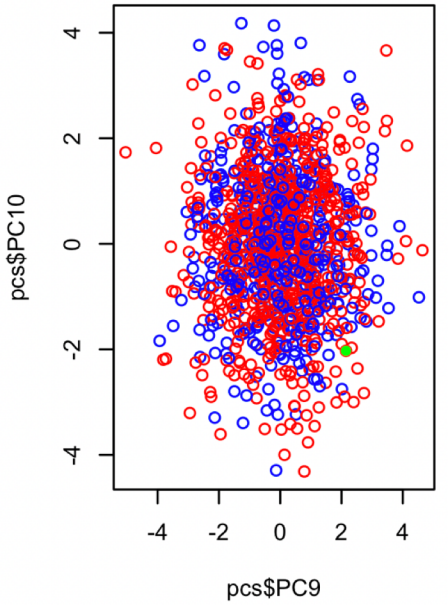






I'm European!





EUR

AFR

Andrew

1. QC Base Data (i.e., GWAS results - betas)

2. QC Target Data (i.e., Genotype Data)

3. Calculate PRS

4. Visualize Results

```
>
> ## Remember our scores are PGS_i = sum_j( beta_j * g_i_j )
> ## the PGS ( PGS_i ) for an individual ( i ) is equal to the sum, over all SNPs ( j )
> ##        of each SNP's aditive genotype count for that individual ( g_i_j ) times the
> ##        additive effect of that SNP taken from a GWAS ( beta_j )
>
> ## For our test subject, consider the first SNP:
>
> tped[ 1,1:7 ]
   CHR        SNP (C)M       POS COUNTED ALT Schork_Andrew
1:   4 rs10000030    0 103374154       A   G             0
> scz[ 1, ]
          SNP A1 A2        OR        BETA     SE      P
1: rs10000030  A  G 0.9888654 -0.01119708 0.0111 0.3098
>
> # How many risk alleles do they carry at this SNP?
>
> # What is their score considering only the first SNP?
>
>
>
```

```
> ## For our test subject, consider the first three SNPs:
>
> tped[ 1:3,1:7 ]
   CHR         SNP (C)M       POS COUNTED ALT Schork_Andrew
1:   4 rs10000030    0 103374154       A   G              0
2:   2  rs1000007    0 237752054       C   T              1
3:   4 rs10000092    0  21895517       C   T              1

> scz[ 1:3, ]
          SNP A1 A2        OR          BETA     SE      P
1: rs10000030  A  G 0.9888654 -0.011197078 0.0111 0.3098
2:  rs1000007  C  T 0.9989012 -0.001099395 0.0091 0.9023
3: rs10000092  C  T 0.9917388 -0.008295497 0.0086 0.3301
>
> # How many risk alleles do they carry across SNPs?
>
> # What is their score considering only these SNPs?
>
>
>
>
```

```
> # For our test subject, what is their score considering only SNP rs6985201?
>
> tped[ tped$SNP == "rs6985201", 1:7 ]
   CHR        SNP (C)M      POS COUNTED ALT Schork_Andrew
1:   8 rs6985201    0 18986961       C   T              2
> scz[ scz$SNP == "rs6985201", ]
         SNP A1 A2       OR         BETA     SE      P
1: rs6985201  C  T 1.001302 0.001300846 0.0099 0.8966
>
> # How many risk alleles do they carry at this SNP?
>
> # What is their score considering only this SNP?
>
>
```

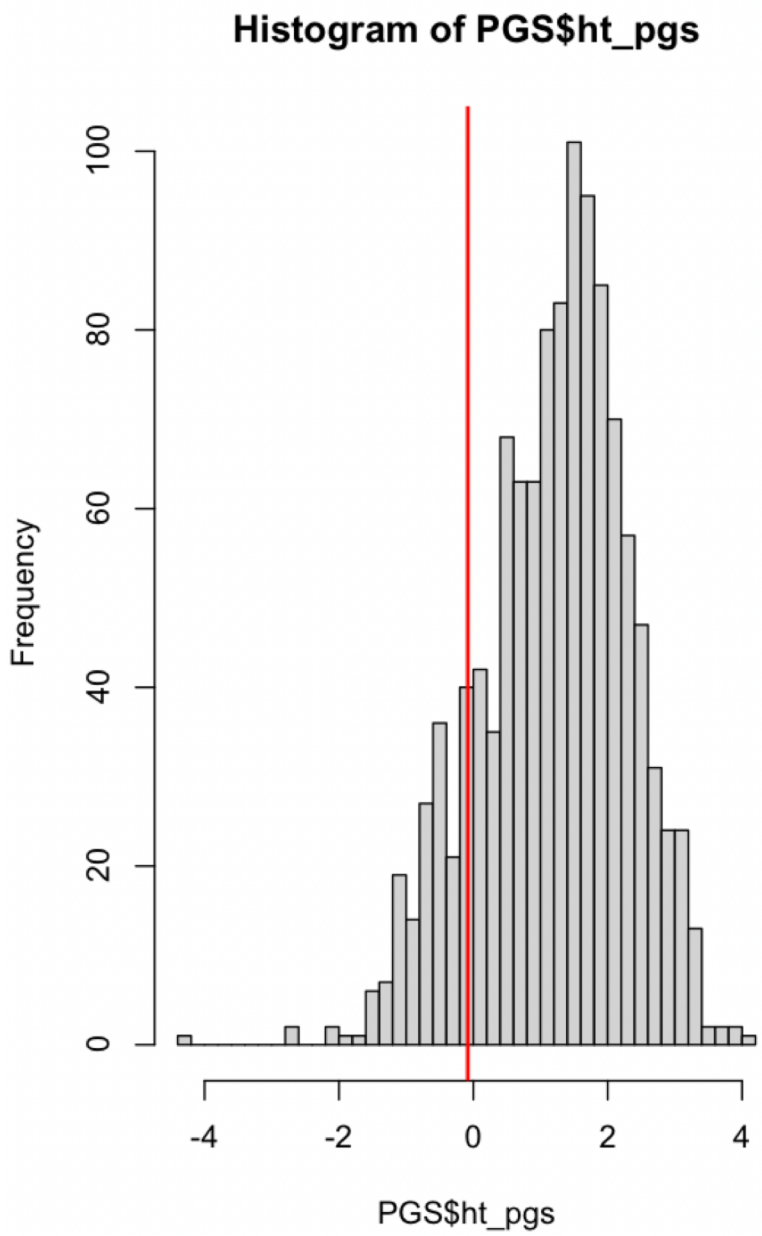
```
> # For our test subject, what is their score considering only SNP rs7746199?
>
> tped[ tped$SNP == "rs7746199", 1:7 ]
   CHR       SNP (C)M      POS COUNTED ALT Schork_Andrew
1:   6 rs7746199    0 27261324       T   C             0
> scz[ scz$SNP == "rs7746199", ]
         SNP A1 A2        OR       BETA     SE         P
1: rs7746199  T  C 0.8759943 -0.1323957 0.0115 7.285e-31
>
> # How many risk alleles do they carry at this SNP?
>
> # What is their score considering only this SNP?
>
>
```

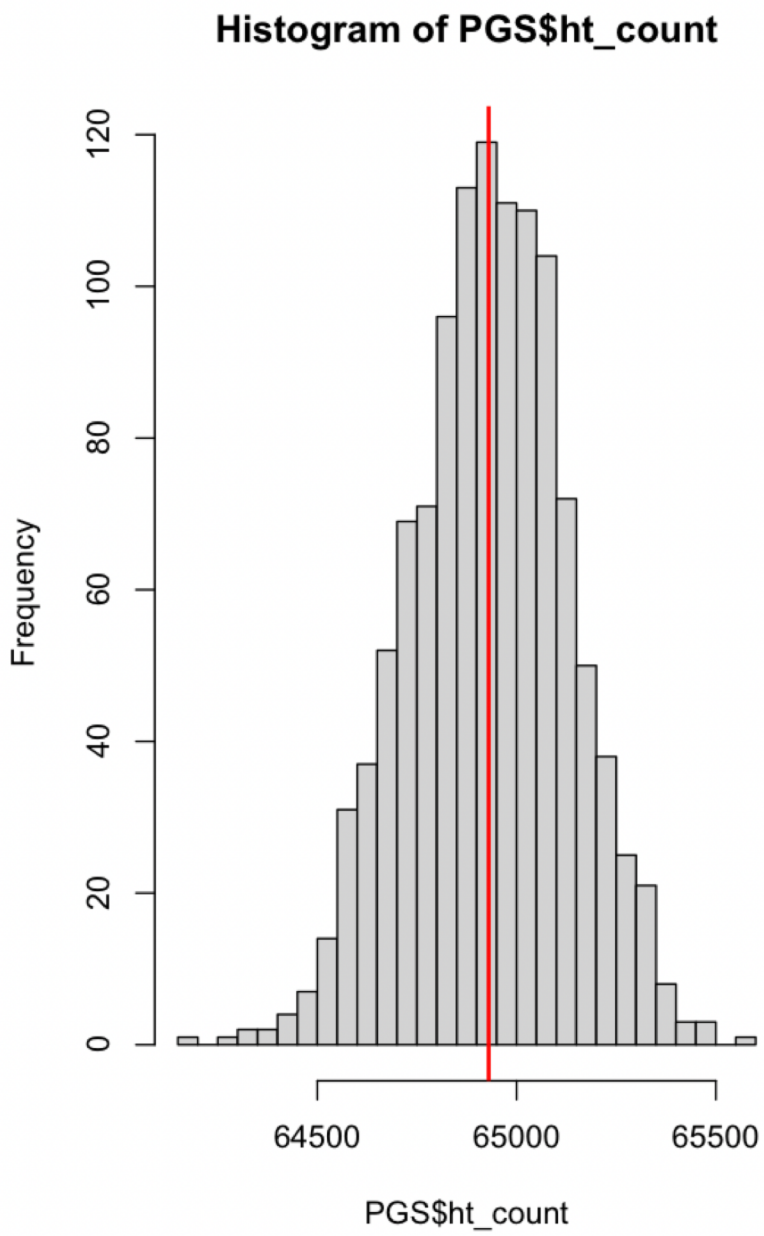
```
> # For our test subject, what is their score considering only SNP rs7746199 and rs17620999?
>
> tped[ tped$SNP %in% c( "rs7746199", "rs17620999" ), 1:7 ]
   CHR        SNP (C)M      POS COUNTED ALT Schork_Andrew
1:   3 rs17620999    0  2561556       G   A             0
2:   6  rs7746199    0 27261324       T   C             0
> scz[ scz$SNP %in% c( "rs7746199", "rs17620999" ), ]
          SNP A1 A2        OR        BETA     SE         P
1: rs17620999  G  A 1.0677943  0.06559508 0.0106 5.218e-10
2:  rs7746199  T  C 0.8759943 -0.13239575 0.0115 7.285e-31
>
> # How many risk alleles do they carry across SNPs?
>
> # What is their score considering only these SNPs?
>
>
>
```

```
> # Use this code for the last examaple and compare
>
> temp.g <- tped[[ 7 ]][ tped$SNP %in% c( "rs7746199", "rs17620999" ) ]
> temp.beta <- scz$BETA[ scz$SNP %in% c( "rs7746199", "rs17620999" ) ]
>
> temp.beta
[1]  0.06559508 -0.13239575
> temp.g
[1] 0 0
>
> temp.g.recoded <- temp.g
> temp.g.recoded[ sign( temp.beta ) == -1 ] <-
+   -1*( temp.g.recoded[ sign( temp.beta ) == -1 ] - 2 )
>
> temp.beta
[1]  0.06559508 -0.13239575
> temp.g
[1] 0 0
> temp.g.recoded
[1] 0 2
>
> temp.g %*% temp.beta
     [,1]
[1,]    0
> sum( temp.g.recoded )
[1] 2
>
>
```

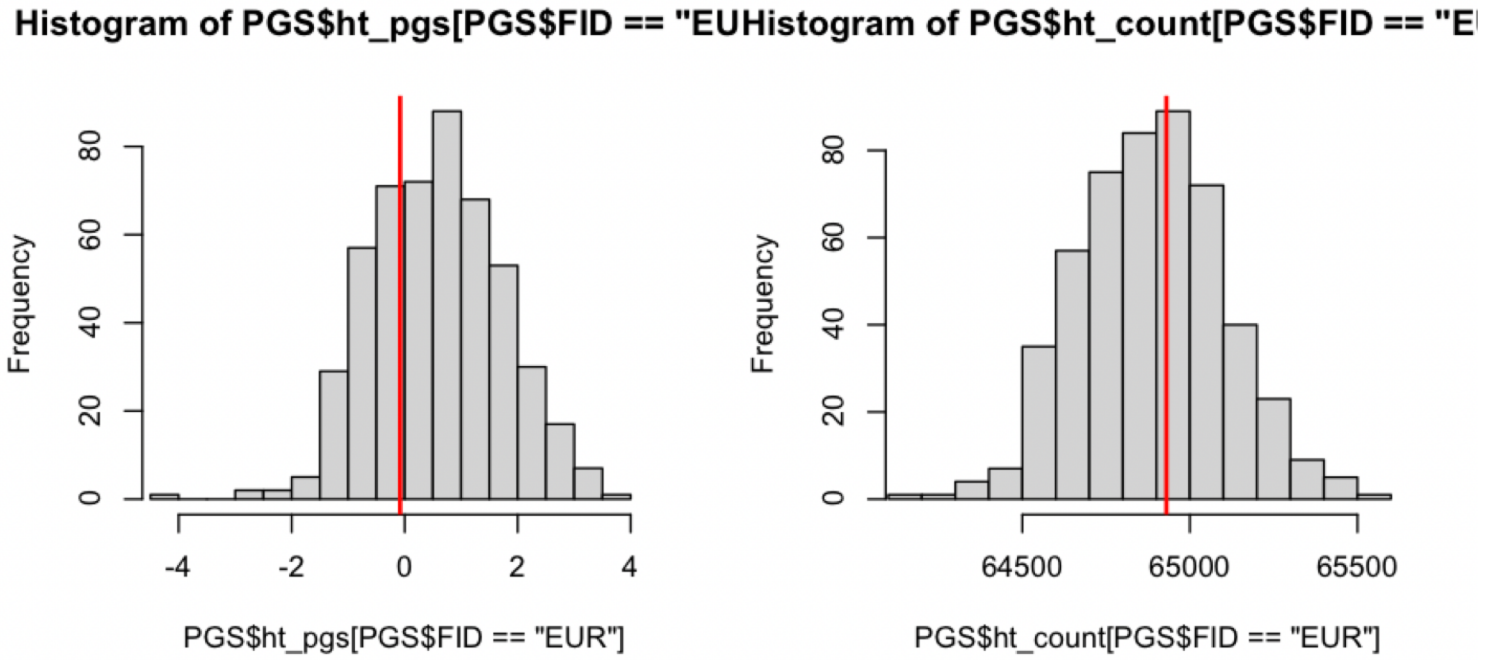
1. QC Base Data (i.e., GWAS results - betas)
2. QC Target Data (i.e., Genotype Data)
3. Calculate PRS
4. Visualize Results

```
> ######
> ## Let's explore concepts of PGS
> ######
>
> ## I got your back.  Delete everything, and load PGS file for all traits and disorders
>
> rm( list=ls() )
>
> PGS <- fread( 'PGS.txt' )
>
> head( PGS )
      FID     IID    scz_pgs scz_count  adhd_pgs adhd_count   asd_pgs asd_count   mdd_pgs mdd_count  ht_pgs ht_count     iq_pgs iq_count
1: Schork  Andrew 11.4456565     63936 15.752695      64374  7.926225     64452 5.2698454     64518 -0.0817    64930 -1.2775072    65372
2:    EUR HG00096 -1.2917829     63640 16.734479      64376 11.178241     64494 3.4933385     64454  2.4513    65482 -2.0358799    65048
3:    EUR HG00097  4.8083117     63795  5.589195      64053  7.068588     64331 3.0418330     64311 -0.9564    64543  1.1865658    65775
4:    EUR HG00099  2.5272816     63700 12.370220      64348  4.693693     64370 1.5946940     64176  1.0946    64966  0.5213370    65554
5:    EUR HG00100  3.2321754     63913 11.568494      64153  2.796782     64097 0.5176424     64155  1.1347    64985 -0.2276213    65297
6:    EUR HG00101 -0.4263195     63543  5.336059      63903  8.748369     64553 1.9113829     64193 -2.1611    64519  0.2935091    65567
>
```

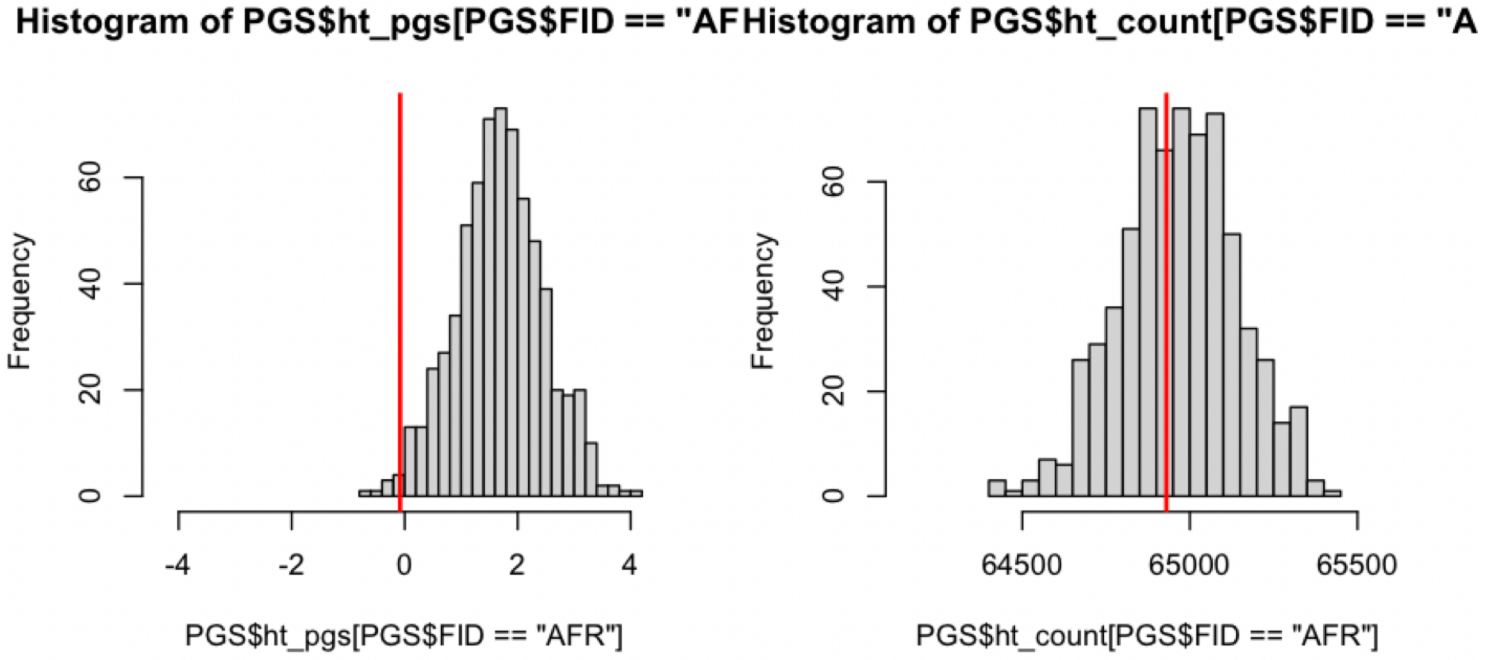

Histogram of PGS$ht_pgs
Frequency
0
20
40
60
80
100
-4
-2
0
2
4
PGS$ht_pgs


Histogram of PGS$ht_count
Frequency
0
20
40
60
80
100
120
64500
65000
65500
PGS$ht_count

Am I tall?


Histogram of PGS$ht_pgs[PGS$FID == "EU
Frequency
0
20
40
60
80
-4
-2
0
2
4
PGS$ht_pgs[PGS$FID == "EUR"]
Histogram of PGS$ht_count[PGS$FID == "E
Frequency
0
20
40
60
80
64500
65000
65500
PGS$ht_count[PGS$FID == "EUR"]

Am I tall?


Histogram of PGS$ht_pgs[PGS$FID == "AF
Frequency
0
20
40
60
-4
-2
0
2
4
PGS$ht_pgs[PGS$FID == "AFR"]
Histogram of PGS$ht_count[PGS$FID == "A
Frequency
0
20
40
60
64500
65000
65500
PGS$ht_count[PGS$FID == "AFR"]

```
> all.pgs.ecdf( PGS$ht_pgs[1] )
[1] 0.1381974
> all.ct.ecdf( PGS$ht_count[1] )
[1] 0.4927039
> eur.pgs.ecdf( PGS$ht_pgs[1] )
[1] 0.3001988
> eur.ct.ecdf( PGS$ht_count[1] )
[1] 0.5864811
> afr.pgs.ecdf( PGS$ht_pgs[1] )
[1] 0.01361573
> afr.ct.ecdf( PGS$ht_count[1] )
[1] 0.4205749
>
```

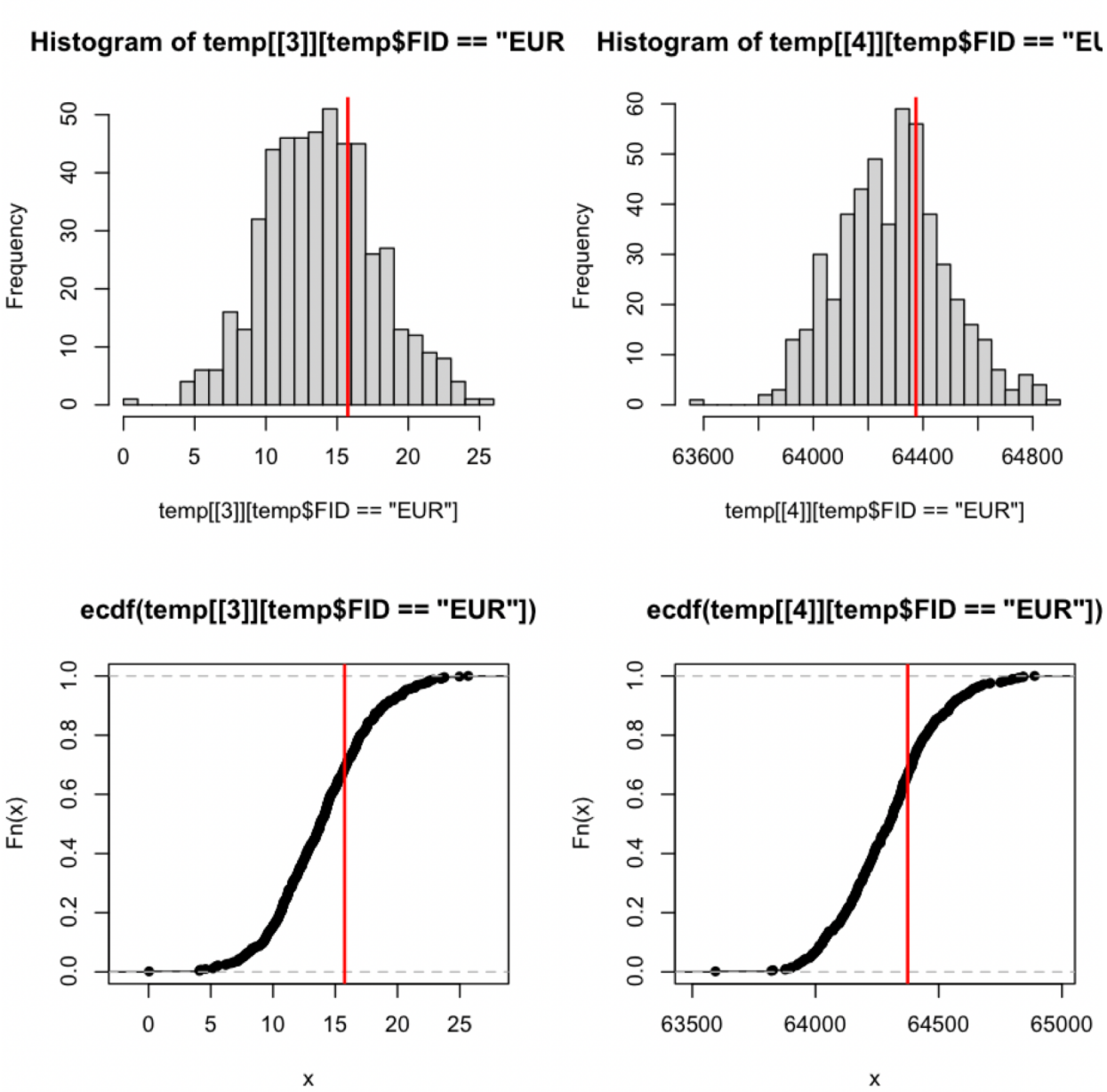
Am I tall?!?!

What do we notice re: ancestry?

What do we notice re: counts vs. scores

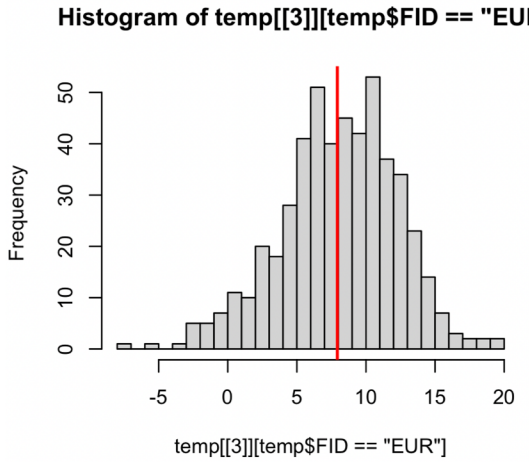
ADHD?

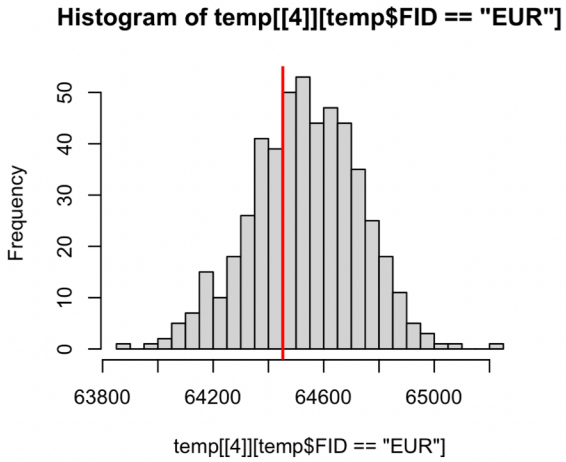
```
> #ADHD
> temp <- PGS[ ,c(1,2,5,6) ]
>
> par( mfrow=c(2,2) )
>
> hist(  temp[[3]][ temp$FID == 'EUR' ], breaks='fd' )
> abline( v=temp[[3]][1], col='red', lwd=2 )
> hist(  temp[[4]][ temp$FID == 'EUR' ], breaks='fd' )
> abline( v=temp[[4]][1], col='red', lwd=2 )
>
> pgs.ecdf <- ecdf( temp[[3]][ temp$FID == 'EUR' ] )
> ct.ecdf <- ecdf( temp[[4]][ temp$FID == 'EUR' ] )
>
> plot( pgs.ecdf )
> abline( v=temp[[3]][1], col='red', lwd=2 )
> plot( pgs.ecdf )
> abline( v=temp[[4]][1], col='red', lwd=2 )
>
> pgs.ecdf( temp[[3]][1] )
[1] 0.6878728
> ct.ecdf( temp[[4]][1] )
[1] 0.6640159
```

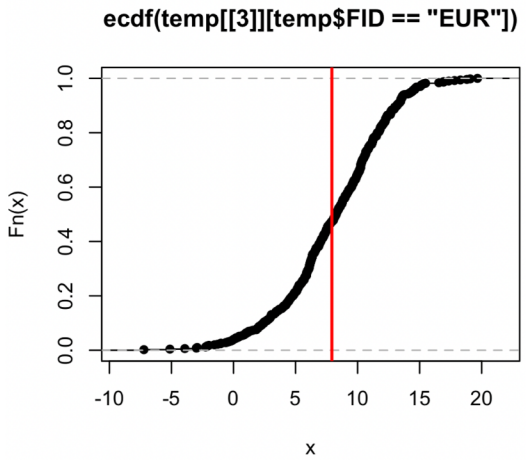


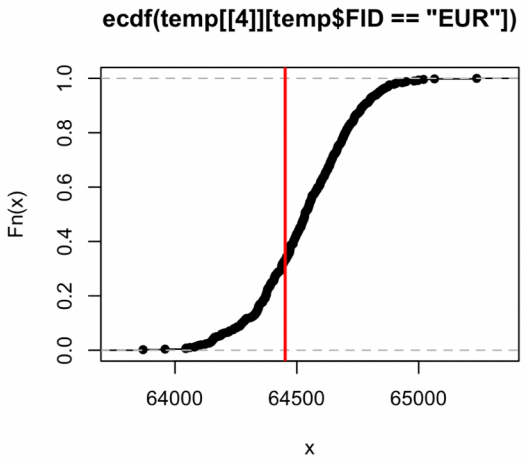
ASD?

```
> #ASD
> temp <- PGS[ ,c(1,2,7,8) ]
>
> par( mfrow=c(2,2) )
>
> hist(  temp[[3]][ temp$FID == 'EUR' ], breaks='fd' )
> abline( v=temp[[3]][1], col='red', lwd=2 )
> hist(  temp[[4]][ temp$FID == 'EUR' ], breaks='fd' )
> abline( v=temp[[4]][1], col='red', lwd=2 )
>
> pgs.ecdf <- ecdf( temp[[3]][ temp$FID == 'EUR' ] )
> ct.ecdf <- ecdf( temp[[4]][ temp$FID == 'EUR' ] )
>
> plot( pgs.ecdf )
> abline( v=temp[[3]][1], col='red', lwd=2 )
> plot( ct.ecdf )
> abline( v=temp[[4]][1], col='red', lwd=2 )
>
> pgs.ecdf( temp[[3]][1] )
[1] 0.471173
> ct.ecdf( temp[[4]][1] )
[1] 0.332008
>
>
```





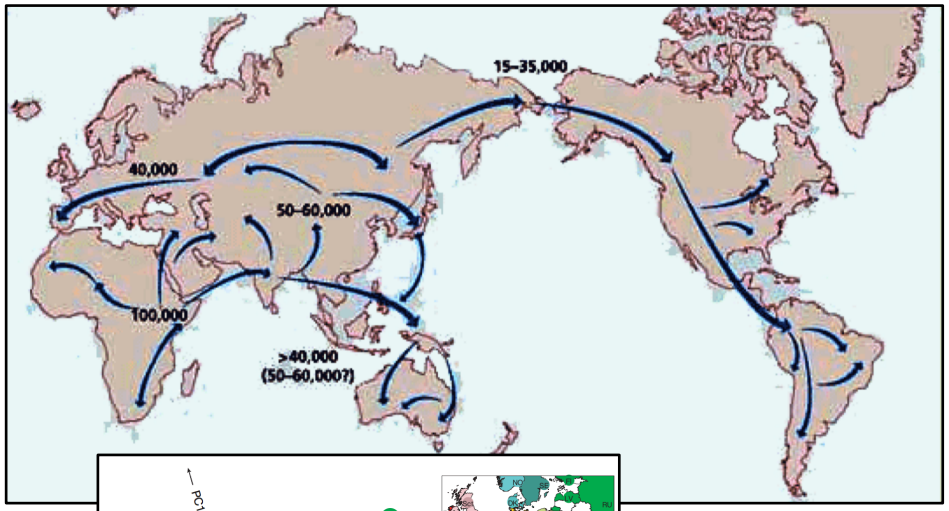


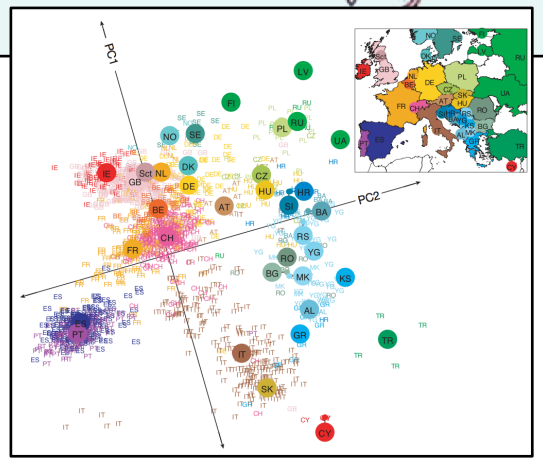


MDD?

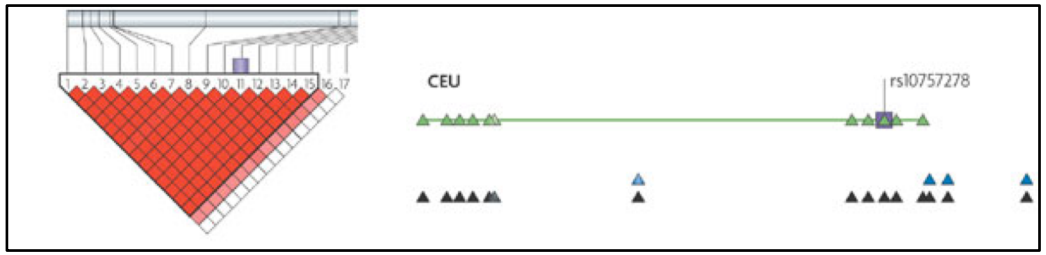
~~SCZ?~~ Creative

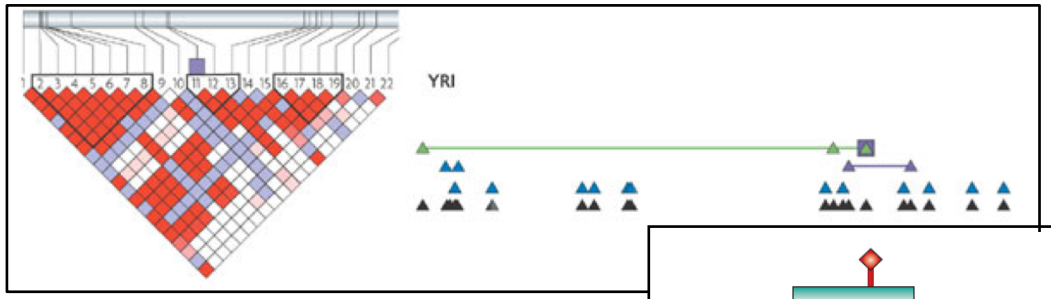
IQ?



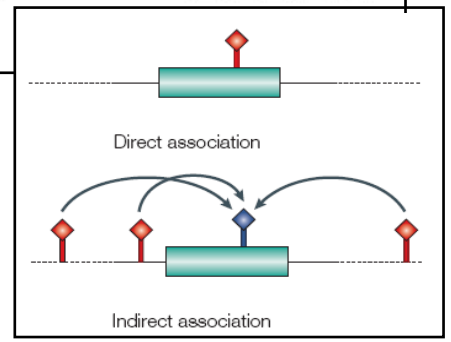


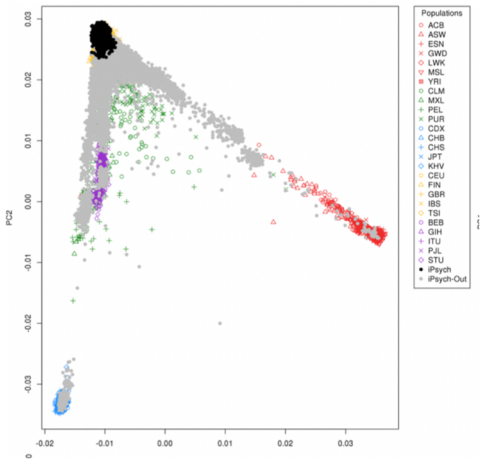
- Frequencies vary by population
- Power varies by frequency
- The most important SNPs vary

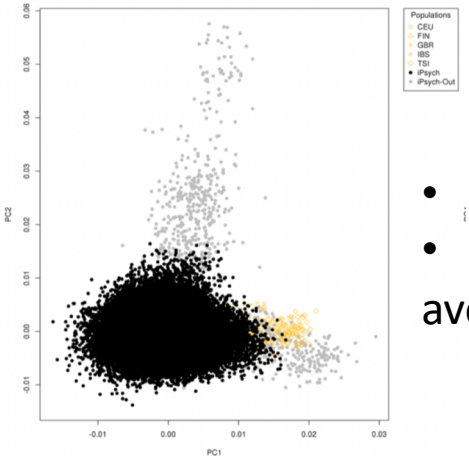




- LD varies
- Beta is a function of LD
- Betas can be wrong







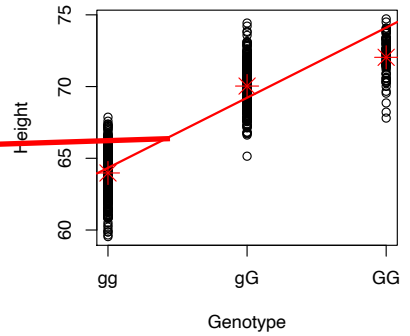
- European studies are the biggest
- Non-Europeans are excluded to avoid population stratification

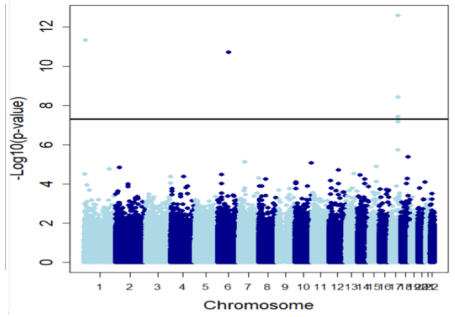
“Polygenic (risk) scores (PRS)”

$$\sum_{all\ g} \beta g_i$$

g_i the number of “g” alleles carried by i and a given ~~causative~~ *genotyped SNP*

β the ~~true~~ estimated phenotype change caused by each “g” allele





The notion that we can genotype an individual to measure the allelic state for large set of SNPs (g’s) and combine these with estimated allelic effects from a large GWAS to try and measure an individuals additive genetic value (aka breeding value) as a PRS.

These estimated PRS (or in pure statistical genetics - eBV) can be used to predict traits in individuals

nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0183-z

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4*]

A key public health need is to identify individuals at high risk for a given disease to enable enhanced screening or preventive therapies. Because most common diseases have a genetic component, one important approach is to stratify individuals based on inherited DNA variation[1]. Proposed clinical applications have largely focused on finding carriers of rare monogenic mutations at several-fold increased risk. Although most disease risk is polygenic in nature[2–5], it has not yet been possible to use polygenic predictors to identify individuals at risk comparable to monogenic mutations. Here, we develop and validate genome-wide polygenic scores for five common diseases. The approach identifies 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively. For coronary artery disease, this prevalence is 20-fold higher than the carrier frequency of rare monogenic mutations conferring comparable risk[6]. We propose that it is time to contemplate the inclusion of polygenic risk prediction in clinical care, and discuss relevant issues.

For various common diseases, genes have been identified in which rare mutations confer several-fold increased risk in heterozygous carriers. An important example is the presence of a familial hypercholesterolemia mutation in 0.4% of the population, which confers an up to threefold increased risk for coronary artery disease (CAD)[6]. Aggressive treatment to lower circulating cholesterol levels among such carriers can significantly reduce risk[7]. Another example is the p.Glu508Lys missense mutation in *HNF1A*, with a carrier frequency of 0.1% of the general population and 0.7% of Latinos[8], which confers up to fivefold increased risk for type 2 diabetes[9]. Although the ascertainment of monogenic mutations can be highly relevant for carriers and their families, the vast majority of disease occurs in those without such mutations.

For most common diseases, polygenic inheritance, involving many common genetic variants of small effect, plays a greater role than rare monogenic mutations[2–5]. However, it has been unclear whether it is possible to create a genome-wide polygenic score (GPS) to identify individuals at clinically significantly increased risk—for example, comparable to levels conferred by rare monogenic mutations[10,11].

Previous studies to create GPSs had only limited success, providing insufficient risk stratification for clinical utility (for example, identifying 20% of a population at 1.4-fold increased risk relative to the rest of the population)[12]. These initial efforts were hampered by three challenges: (1) the small size of initial genome-wide association studies (GWASs), which affected the precision of the estimated impact of individual variants on disease risk; (2) limited computational methods for creating GPSs; and (3) a lack of large datasets needed to validate and test GPS.

Using much larger studies and improved algorithms, we set out to revisit the question of whether a GPS can identify subgroups of the population with risk approaching or exceeding that of a monogenic mutation. We studied five common diseases with major public health impact: CAD, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer.

For each of the diseases, we created several candidate GPSs based on summary statistics and imputation from recent large GWASs in participants of primarily European ancestry (Table 1). Specifically, we derived 24 predictors based on a pruning and thresholding method, and 7 additional predictors using the recently described LDPred algorithm[13] (Methods, Fig. 1 and Supplementary Tables 1–6). These scores were validated and tested within the UK Biobank, which has aggregated genotype data and extensive phenotypic information on 409,258 participants of British ancestry (average age: 57 years; 55% female)[14,15].

We used an initial validation dataset of the 120,280 participants in the UK Biobank phase 1 genotype data release to select the GPSs with the best performance, defined as the maximum area under the receiver-operator curve (AUC). We then assessed the performance in an independent testing dataset comprised of the 288,978 participants in the UK Biobank phase 2 genotype data release. For each disease, the discriminative capacity within the testing dataset was nearly identical to that observed in the validation dataset.

Taking CAD as an example, our polygenic predictors were derived from a GWAS involving 184,305 participants[16] and evaluated based on their ability to detect the participants in the UK Biobank validation dataset diagnosed with CAD (Table 1). The predictors had AUCs ranging from 0.79–0.81 in the validation set, with the best predictor (GPS_{CAD}) involving 6,630,150 variants (Supplementary Table 1). This predictor performed equivalently well in the testing dataset, with an AUC of 0.81.

[1]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [2]Cardiology Division of the Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Cardiovascular Disease Initiative of the Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]These authors contributed equally: Amit V. Khera, Mark Chaffin. *e-mail: skathiresan1@mgh.harvard.edu

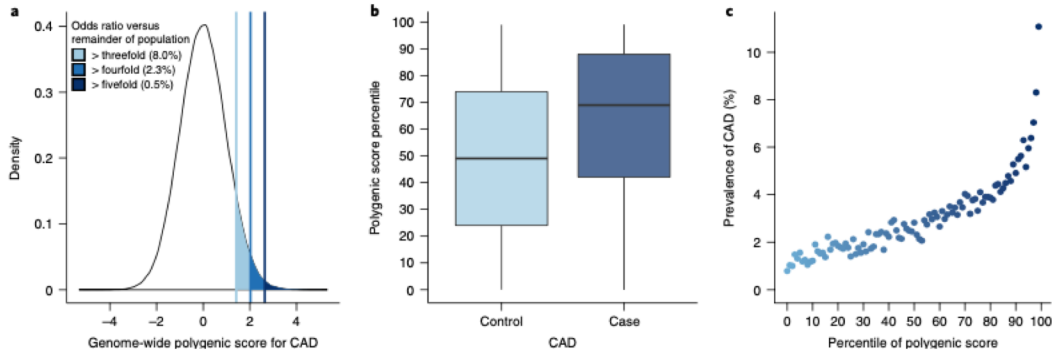


Fig. 2 | Risk for CAD according to GPS. a, Distribution of GPS_{CAD} in the UK Biobank testing dataset ($n = 288,978$). The *x* axis represents GPS_{CAD}, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation. Shading reflects the proportion of the population with three-, four-, and fivefold increased risk versus the remainder of the population. The odds ratio was assessed in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. **b**, GPS_{CAD} percentile among CAD cases versus controls in the UK Biobank testing dataset. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range, and the whiskers reflect the maximum and minimum values within each grouping. **c**, Prevalence of CAD according to 100 groups of the testing dataset binned according to the percentile of the GPS_{CAD}.

The top 8% of PRS for coronary artery disease (CAD) had risk equivalent to a known, clinically actionable, rare variant present in only 0.4% of the population.

Should interventions be prescribed on this basis?

- 20 times more targets -> 20 times more prevention?
- Would the intervention (statins) be as effective?

nature COMMUNICATIONS

ARTICLE

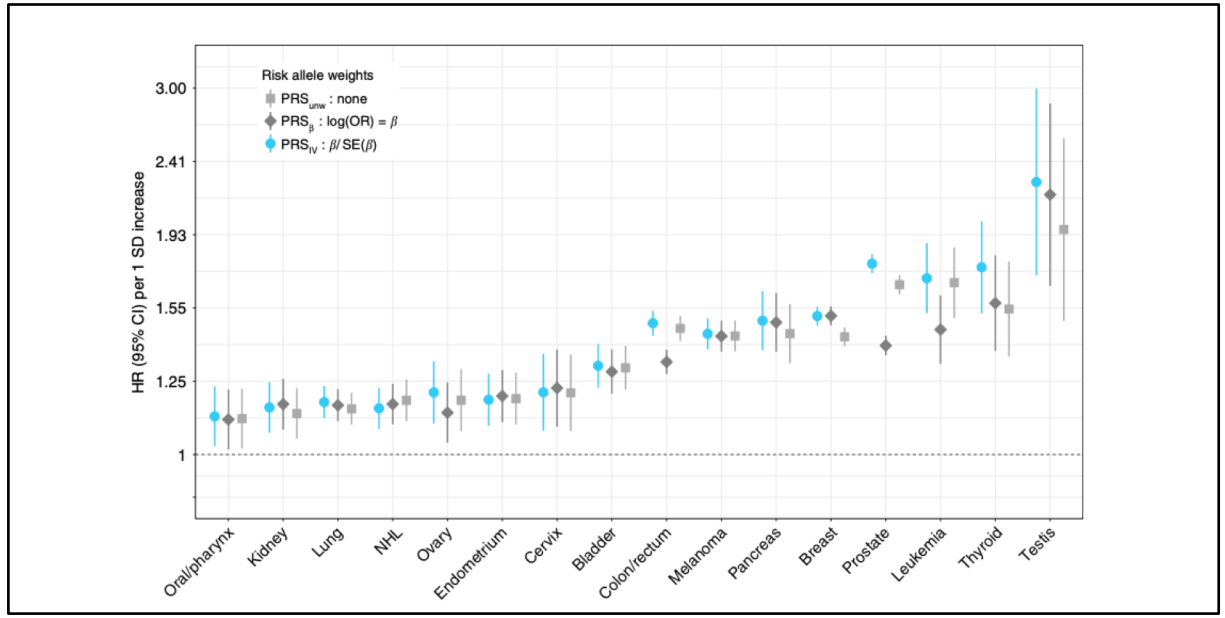
https://doi.org/10.1038/s41467-020-19600-4 OPEN

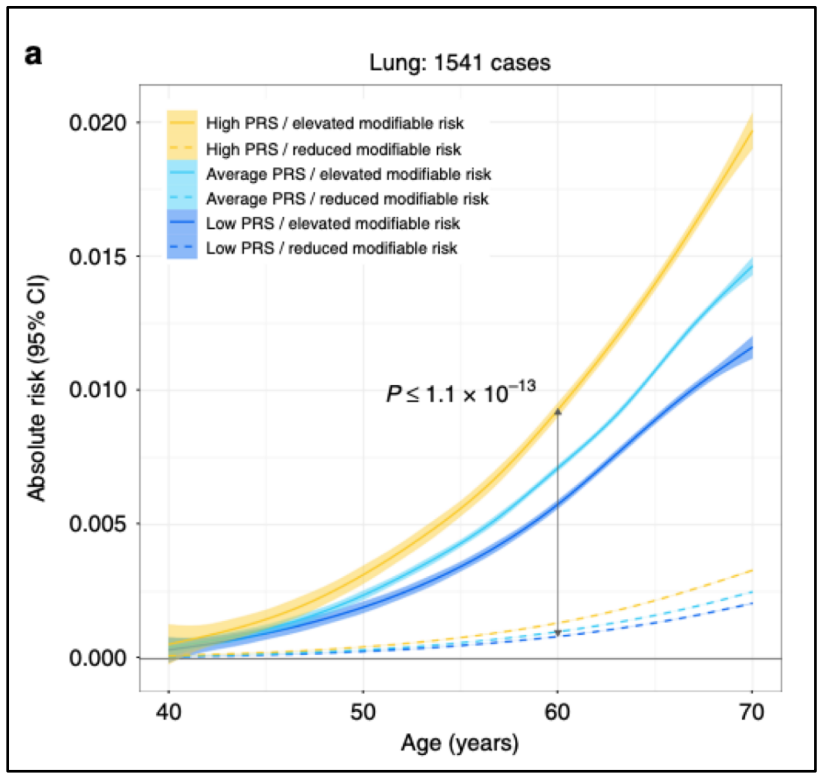
Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction

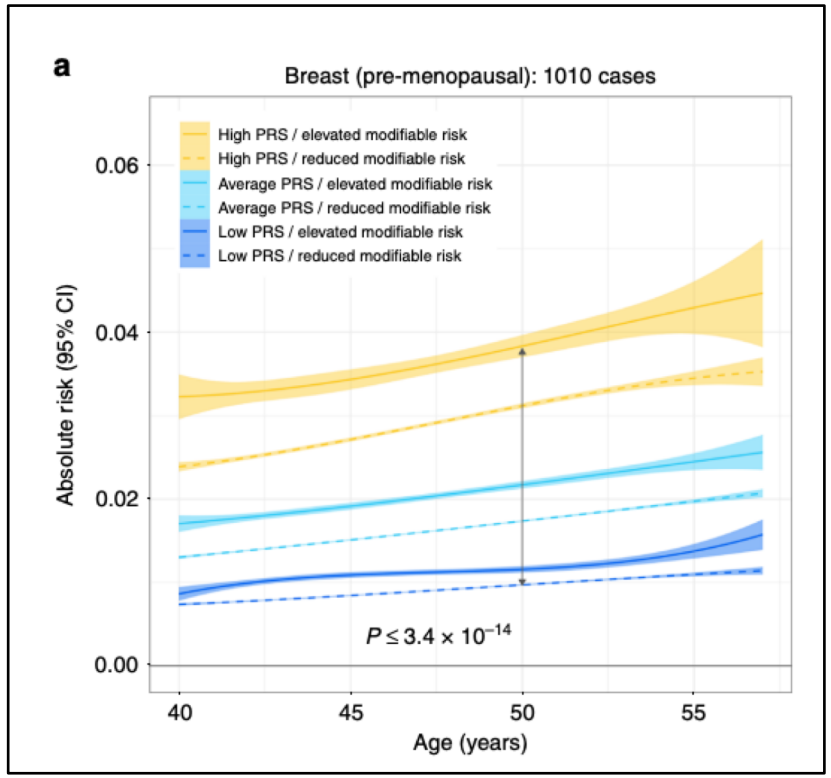
Linda Kachuri[1], Rebecca E. Graff[1], Karl Smith-Byrne[2], Travis J. Meyers[1], Sara R. Rashkin[1], Elad Ziv[3,4,5], John S. Witte[1,4,5,6,7✉] & Mattias Johansson[2,7✉]

Cancer risk is determined by a complex interplay of environmental and heritable factors. Polygenic risk scores (PRS) provide a personalized genetic susceptibility profile that may be leveraged for disease prediction. Using data from the UK Biobank (413,753 individuals; 22,755 incident cancer cases), we quantify the added predictive value of integrating cancer-specific PRS with family history and modifiable risk factors for 16 cancers. We show that incorporating PRS measurably improves prediction accuracy for most cancers, but the magnitude of this improvement varies substantially. We also demonstrate that stratifying on levels of PRS identifies significantly divergent 5-year risk trajectories after accounting for family history and modifiable risk factors. At the population level, the top 20% of the PRS distribution accounts for 4.0% to 30.3% of incident cancer cases, exceeding the impact of many lifestyle-related factors. In summary, this study illustrates the potential for improving cancer risk assessment by integrating genetic risk scores.

[1] Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA. [2] Genetic Epidemiology Group, Section of Genetics, International Agency for Research on Cancer, Lyon, France. [3] Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. [4] Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA. [5] Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. [6] Department of Urology, University of California, San Francisco, San Francisco, CA, USA. [7] These authors jointly supervised this work: John S. Witte, Mattias Johansson. ✉email: jwitte@ucsf.edu; JohanssonM@iarc.fr







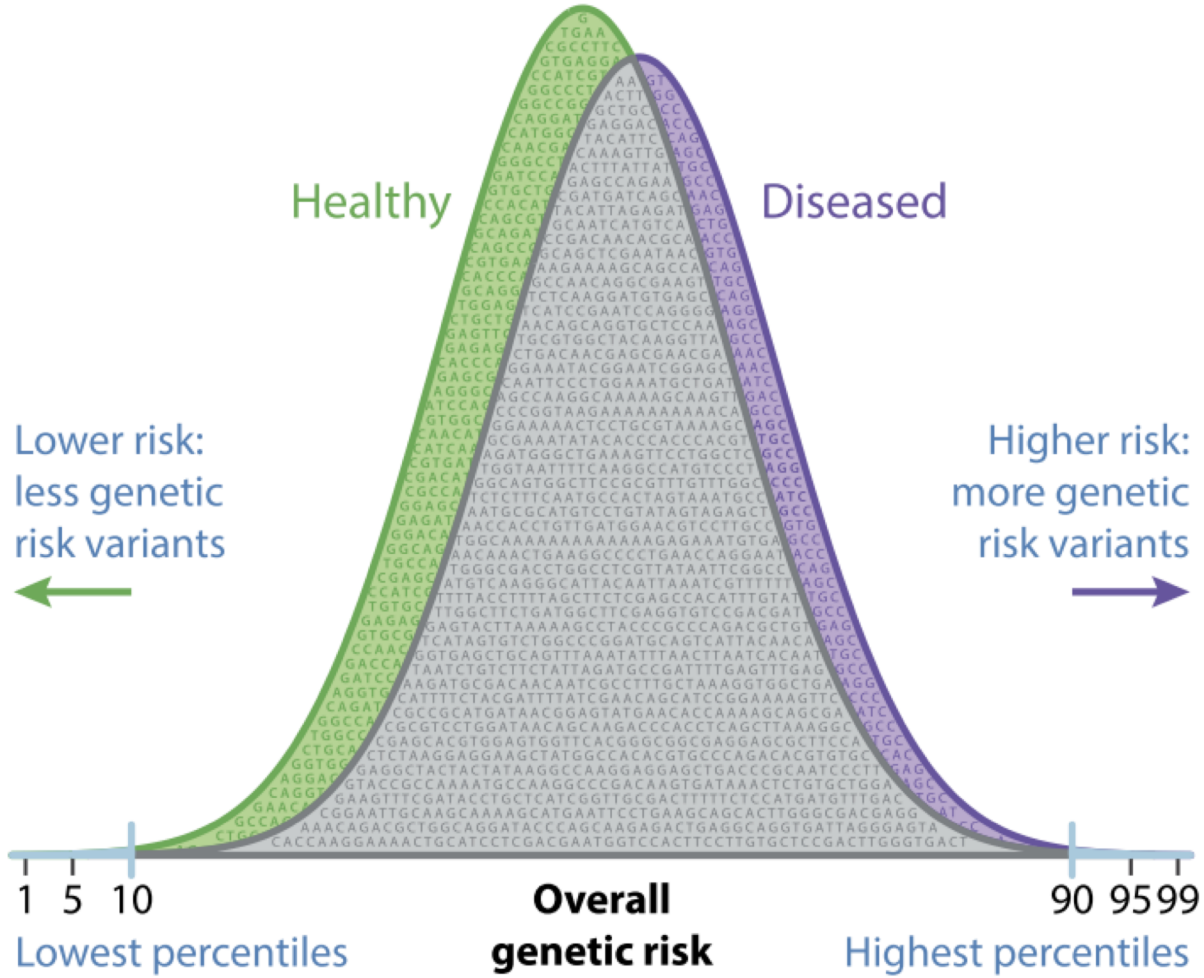


Figure 2: Polygenic scores typically follow a normal distribution on a population level. The distributions of cases and controls show a clear overlap. Thus, meaningful risk predictions can be only expected for extreme quantiles at the top and bottom of the PGS distributions.