



UNIVERSITY OF  
COPENHAGEN



# Prediction for the individual: A practical workshop

Panum – 21.2.14

9 November 2021 14:15

Andrew Schork

Institute for Biological Psychiatry  
[Andrew.Joseph.Schork@regionh.dk](mailto:Andrew.Joseph.Schork@regionh.dk)

Don't worry guys,  
I called Mom yesterday



(best son ever)



## Tutorial: a guide to performing polygenic risk score analyses

Shing Wan Choi<sup>1,2</sup>, Timothy Shin-Heng Mak<sup>1,3</sup> and Paul F. O'Reilly<sup>1,2</sup>✉

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<sup>1–3</sup>. 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<sup>4–6</sup>. Using a linear mixed model in the genome-wide complex trait analysis software<sup>7</sup>, Yang et al. demonstrated that much of the heritability of height can be explained by evaluating the effects of all SNPs simultaneously<sup>8</sup>. Subsequently, statistical techniques such as linkage disequilibrium (LD) score regression<sup>8,9</sup> and the polygenic risk score (PRS) method<sup>5,10</sup> 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<sup>5,6,8–10</sup>.

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<sup>5,11–14</sup> (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<sup>11,15,16</sup>. 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<sup>15,17</sup>, to perform Mendelian randomization studies to infer causal relationships and for patient stratification and sub-phenotyping<sup>15,16,18</sup>. 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<sup>12</sup> 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<sup>12,14,19,20</sup>. 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.

<sup>1</sup>MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.

<sup>2</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York, NY, USA. <sup>3</sup>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-ddatatable.com
*****
This installation of data.table has not detected OpenMP support. It should still work but in single-threaded mode.
This is a Mac. Please read https://mac.r-project.org/openmp/. Please engage with Apple and ask them for support. Check r-datable.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 similar problems on Windows or Linux.
*****
```

A large blue arrow points from the bottom of the screen down towards the error message.

```
>
> 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"                "KG_PGS_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

To search, type and hit:

- 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 download by 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 implied, performance, merchantability, or fitness for any particular purpose
- investigators will use these results for scientific research and educational use only
- reprinting or public distribution of PGC results files is prohibited
- investigators certify that they are in compliance with all applicable local, state, and federal laws and regulations regarding subjects and genetics research
- investigators will cite the appropriate PGC publications in any communications or publications
- investigators will never attempt to identify any participants
- for use of data available prior to publication, investigators will respect the requested responsible principles

All data here are released under a [Fort Lauderdale Agreement](#) (<https://www.genome.gov/Pages/Resources/PGC/PGC-Data-Release-Information.html>). This agreement is designed to facilitate wider biomedical community for use in the investigation of the genetics of schizophrenia and wider psychiatric disorders. It also 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 including results 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	pubMedID
ADHD	ADHD	adhd2019	Nature Genetics	<a href="#">304784</a>
ADHD	ADHD	adhd2010	JAACP	<a href="#">207320</a>
ALZ	Alzheimer disease	alz2019	Nature Genetics	<a href="#">306172</a>
ANX	Anx disorders & factors	anx2016	Mol Psych	<a href="#">267549</a>
ASD	Autism spectrum disorder	asd2019	Nature Genetics	<a href="#">308045</a>
ASD	Autism spectrum disorder	asd2017	Mol Autism	<a href="#">285400</a>



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 GWAS



This work is licensed under a [Creative Commons](#)

In addition, when [downloading the sumstats](#) the sumstats for projects that may lead to s

**Commercial use:** If you want to use the sums commercial uses, and in most cases can freely Posthuma.

**Summary statistics for A genome-wide association study with 1,126 individuals identifies new risk variants for neuroticism.** 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](#)

[PGCALZ2sumstatsExcluding23andMe.txt.gz](#)

**Summary statistics for Genom neuroticism: an exploratory study across 25 environments.** 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	-0.0040387267010239	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

1 **Mapping genomic loci prioritises genes and  
2 implicates synaptic biology in schizophrenia**

3 **Authors**

4 The Schizophrenia Working Group of the Psychiatric Genomics Consortium

5 **Corresponding Authors**

6 Stephan Ripke<sup>1,2</sup>, James TR Walters<sup>3</sup>, Michael C O'Donovan<sup>3</sup>

7 **Affiliations**

8 1. Dept. of Psychiatry and Psychotherapy, Charité - Universitätsmedizin, Berlin 10117,  
9 Germany

10 2. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston MA  
02114, USA

11 3. MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychiatry and  
12 Clinical Neurosciences, Cardiff University, Hadyn Ellis Building, Maindy Road, Cardiff,  
13 CF24 4HQ

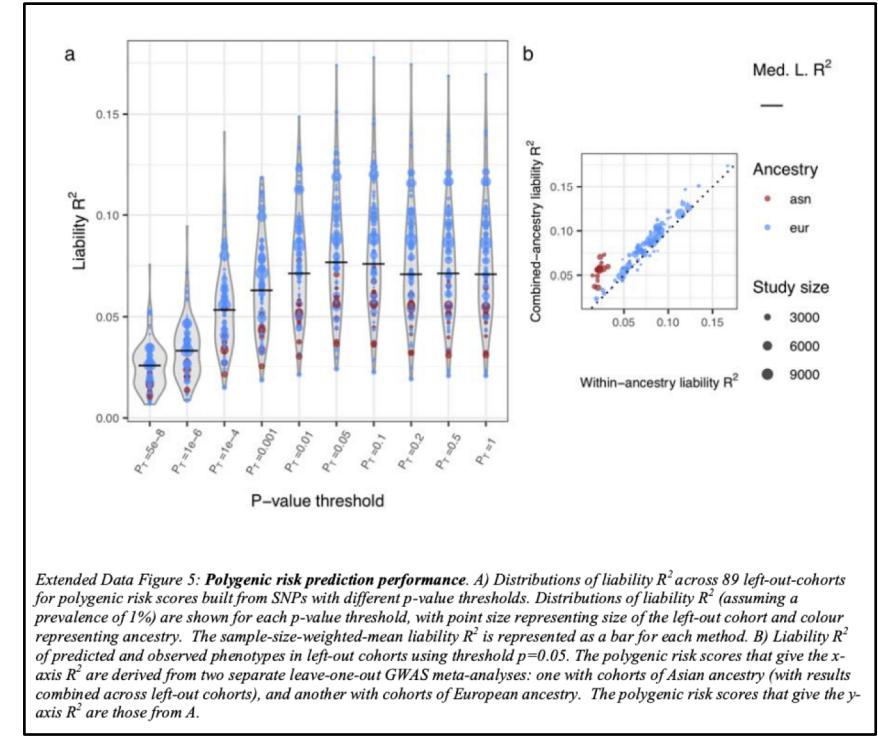
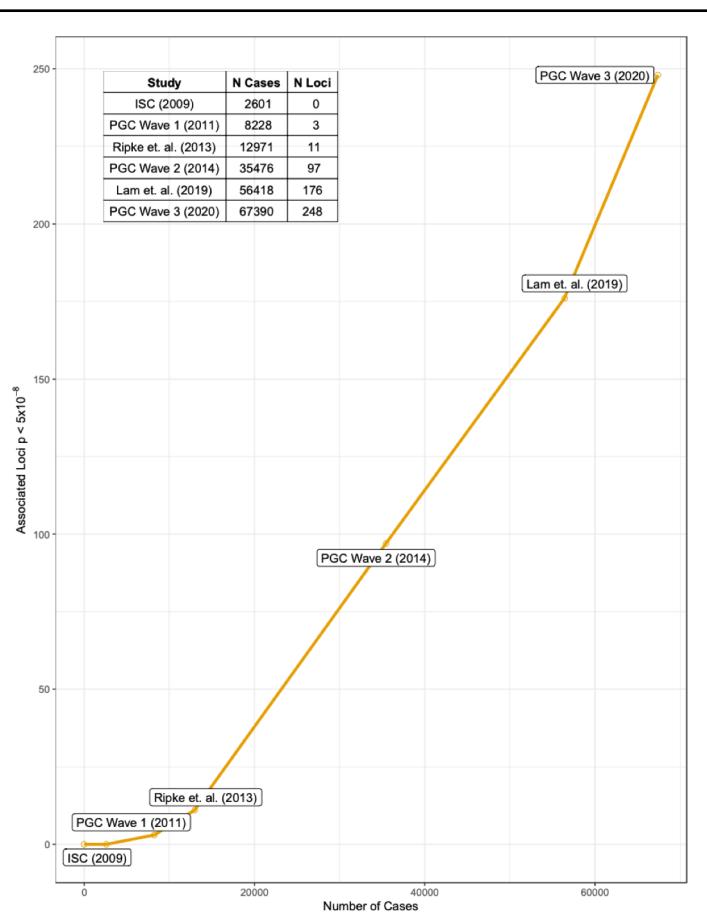
14

15 **Contact**

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

17

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



Most recent schizophrenia GWAS – still “unpublished”

# Typical quality steps

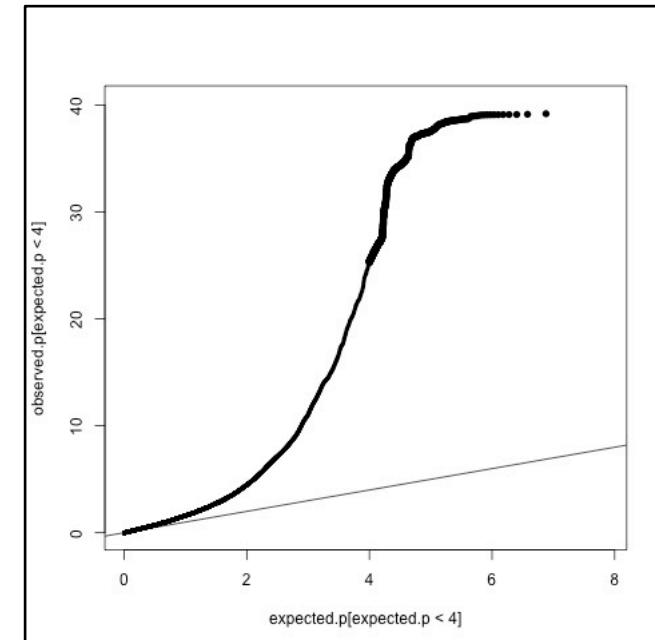
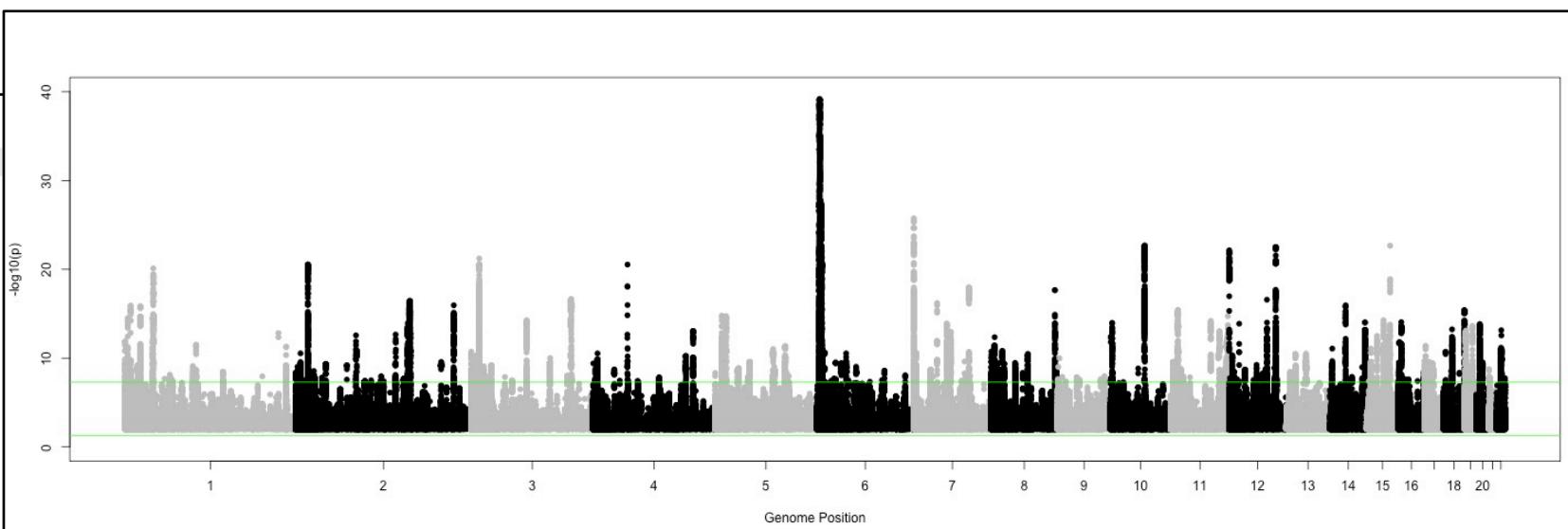
1. Visualize results for eye-ball tests
2. INFO scores
3. Highly missing
4. Overlap with target samples
5. Strand corrections

# Visualize

```

16 head( data )
17
18 data$CHR <- as.numeric( data$CHR )
19 data <- data[ !is.na( data$CHR ), ]
20 data <- data[ order( data$CHR, data$BP ), ]
21
22 head( data )
23
24 data$ManPos <- cumsum( data$BP/1000 )
25 chr.mids <- cbind( 1:22, NA )
26 for ( i in 1:22 ) {
27   chr.start <- min( data$ManPos[ data$CHR == i ] )
28   chr.stop <- max( data$ManPos[ data$CHR == i ] )
29   chr.length <- chr.stop-chr.start
30   chr.mids[ i,2 ] <- chr.start + chr.length/2
31 }
32
33 jpeg( "SCZ_QuickManhattan.jpeg", width=1440, height=480 )
34 plot( data$ManPos[ data$P < 0.01 ], -log10( data$P[ data$P < 0.01 ] ),
35       xaxt='n',
36       col=c( 'black','grey')[ data$CHR[ data$P < 0.01 ] %% 2 + 1 ],
37       xlab="Genome Position", ylab="-log10(p)",
38       ylim=c(0,40),
39       pch=16 )
39 abline( h=-log10( 0.05 ), col='green' )
40 abline( h=-log10( 5e-8 ), col='green' )
41 axis( 1, at=chr.mids[,2], labels=chr.mids[,1] )
42 dev.off()
43
44 # Mini qq-plot
45 jpeg( "SCZ_QuickQQ.jpeg", width=480, height=480 )
46 observed.p <- -log10( data$P[ order( data$P ) ] )
47 expected.p <- -log10( ( 1:length( data$P ) ) / ( length( data$P ) ) )
48 plot( expected.p[ expected.p < 4 ], observed.p[ expected.p < 4 ],
49       type='l',
50       lwd=4,
51       ylim=c( 0, max(observed.p)+1 ),
52       xlim=c( 0, max(expected.p)+1 ) )
53 points( expected.p[ expected.p > 4 ], observed.p[ expected.p > 4 ], pch=16 )
54 abline( 0,1 )
55 dev.off()
56

```



INFO

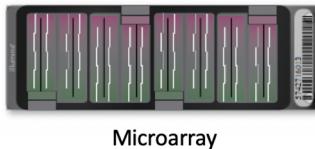
```

> 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 Disequilibrium*



## Microarray

**A**T CGAA **A**TG **X**TGAC **C**TTTGA **X**ATGA **T**CGGC **G**C**A**G**T**CAGC  
**T**TTCGAA **G**TG**A**TGACT **T**TTTGA **A**TGAG**C**GGC**G**CC**C**CA CAGC

## Common Variation

## Rare Mutations

### No Recorded Variation

## The Technology: Whole Genome Sequencing

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



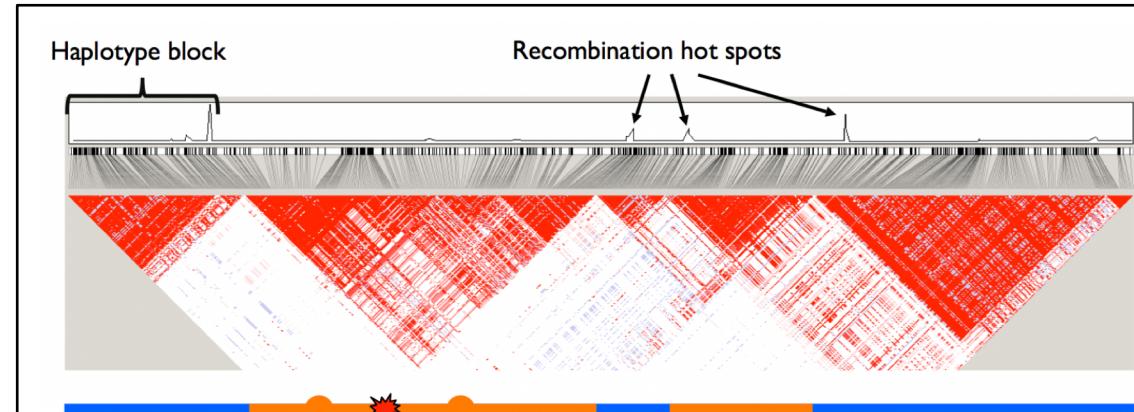
## Sequencer

**ATCGAAATGCATGACCTTTGATATGATCGGC TGCAGTCAGC**  
**TTCGAAGTGCATGACTTTGACATGAGCGGGCGGCCACAGC**

## Common Variations

## Mutations

#### No Recorded Variation



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



*Annual Review of Genomics and Human Genetics*

## Genotype Imputation from Large Reference Panels

Sayantan Das,<sup>1</sup> Gonçalo R. Abecasis,<sup>1</sup>  
and Brian L. Browning<sup>2</sup>

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

<sup>2</sup>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](http://genom.annualreviews.org)

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

Copyright © 2018 by Annual Reviews.  
All rights reserved.

### Keywords

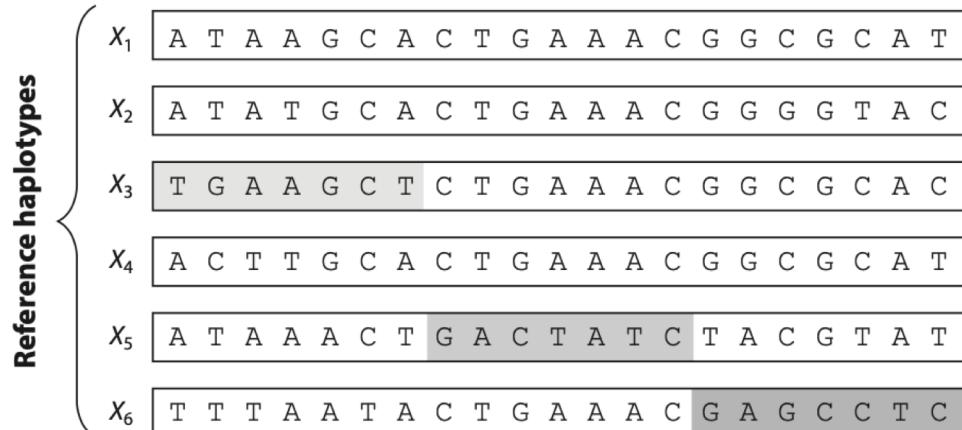
imputation, genotype imputation, genome-wide association study, GWAS

### 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.

Review in Advance first posted on May 23, 2018. (Changes may still occur before final publication.)

16.1



**Missing:**  $S_G$  T . . . G . . . A . . . T . . . A . . . C . . .

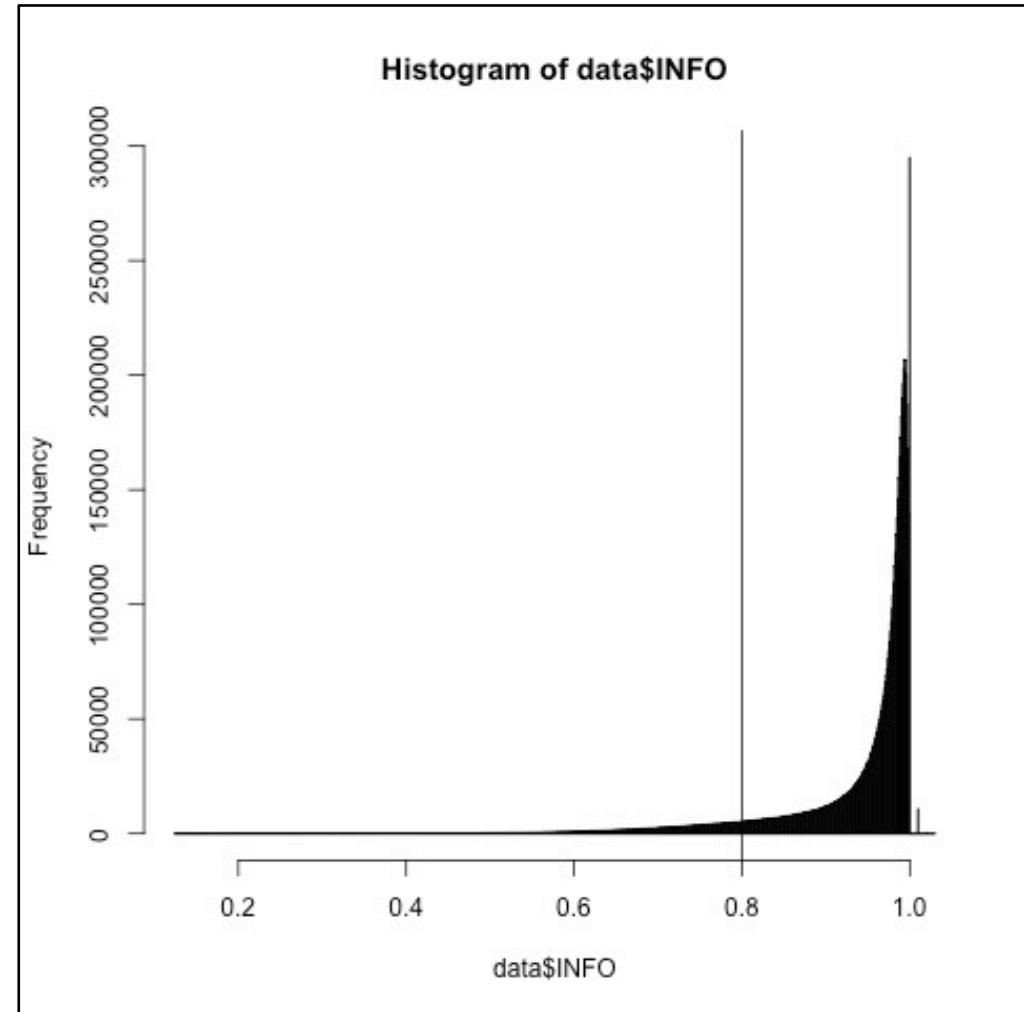
**Imputed:**  $S_I$  T g a a G c t g A c t a T c g A g c C t c

**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 ) ] )
```

Direction	Nca	Nco	Neff	P
-----------	-----	-----	------	---

Y

```
> head( data[ order(data$P), c( 2,12,13,17,18,19,11 ) ] )
```

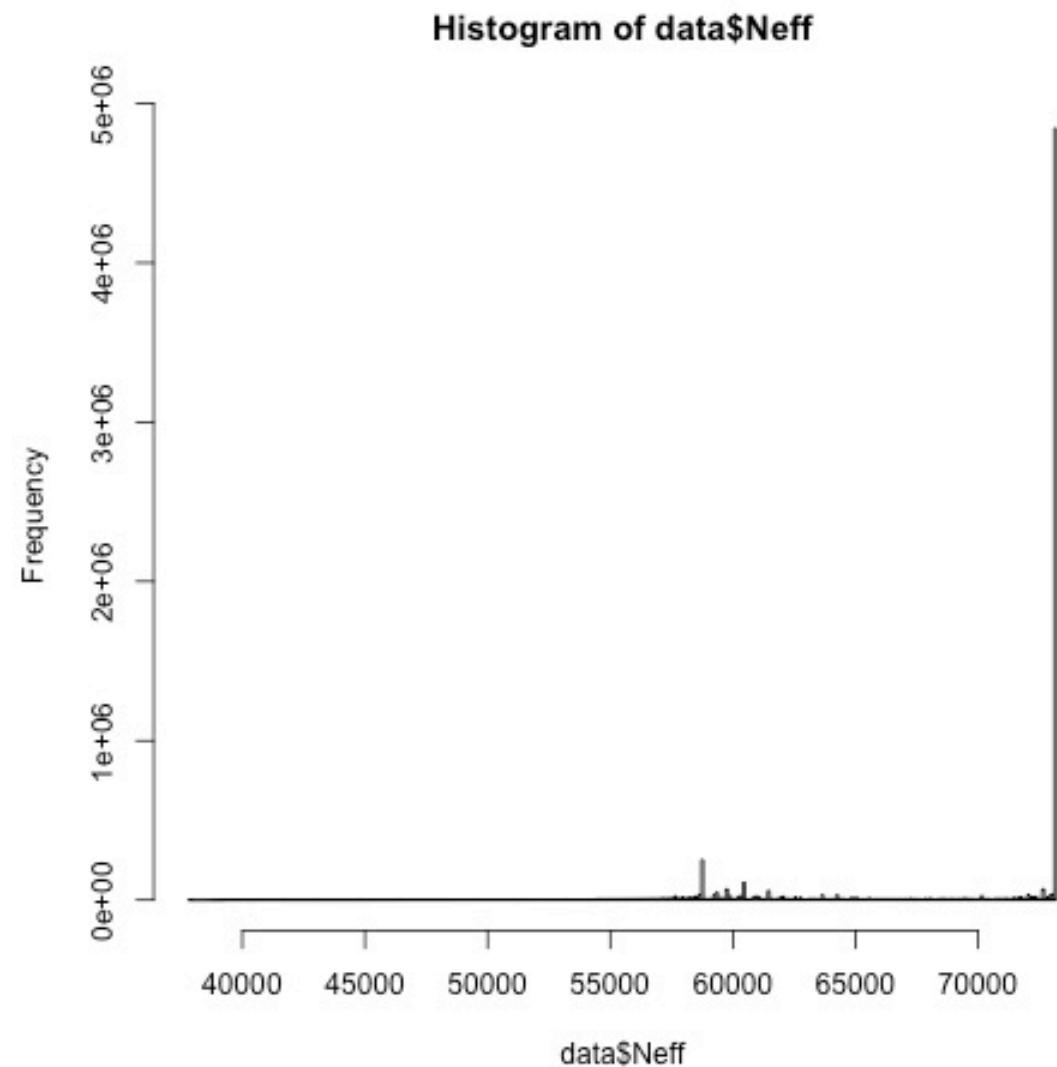
Direction	Nca	Nco	Neff	P
-----------	-----	-----	------	---

SNP	Idx	Direction	Neu	Nco	NCBI
rs13195636	0	+?++??--??-+++-++++-+++-?????+++++-??-+++++-+-----?+-----?+-----?+-----?+-----?+-----?+-----?+-----?	53386	77258	58749.13 6.546e-40
rs10484399	50	+?++??--??-+++-++++-+++-?????+++++-??-+-----?+-----?+-----?+-----?+-----?+-----?+-----?+-----?	53386	77258	58749.13 7.369e-40
rs35501037	1	-?--??++??+---+---+---+??-?-----+--?+-----+-----+-----+-----+-----+-----+-----+-----+-----?	53933	77798	59292.61 7.726e-40
rs140365013	0	-?--??++??+---+---+---+??-?-----+--?+-----+-----+-----+-----+-----+-----+-----+-----?	53386	77258	58749.13 7.856e-40
rs66785117	0	+?++??--??-+++-++++-+++-?????+++++-??-+-----?+-----?+-----?+-----?+-----?+-----?+-----?	53386	77258	58749.13 7.861e-40
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	<span style="color:blue;">□</span>
PGC_SCZ wave3 asn only	nan	nan	nan	nan(nan)	nan(nan)	nan	nan	<span style="color:blue;">—</span>
PGC_SCZ wave2 only	0	1.00	1.46e-24	0.925(32998)	0.912(43387)	0.208	0.02	<span style="color:blue;">□</span>
PGC_SCZ wave3 (excl. wave2)	0	0.99	5.28e-17	0.928(20388)	0.906(33871)	0.215	0.026	<span style="color:blue;">—</span>
PGC_SCZ wave3 (incl. wave2)	0	1.00	6.55e-40	0.926(53386)	0.909(77258)	0.211	0.016	<span style="color:blue;">diamond</span>
Decode – Replication	0	1.00	0.0164	0.938(1979)	0.932(142627)	0.215	0.084	<span style="color:blue;">—</span>
PGC_SCZ wave3 (incl. wave2 and replication)	0	1.00	2.9e-41	0.926(55365)	0.924(219885)	0.21	0.016	<span style="color:blue;">diamond</span>

```
> 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



Craig Venter

**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 initial sequencing and analysis of the human genome, which has been completed to a level of high quality. We also present an initial analysis of the data, describing some of the insights that can be drawn from the sequence.

The endeavour of Man's desire to understand the workings of the universe over the past two centuries – 'a scientific quest to understand the laws of nature and the principles of life and death' for the last hundred years – has produced the greatest masterpiece of all time, the human genome. The project involved the efforts of hundreds of thousands of scientists and engineers from four corners of the earth. The first established the genomic framework of the human genome, the second completed the sequencing of the DNA, the third conducted the intensive analysis of the sequence, and the fourth analysed the genome by which cells and the environment interact with genes and with each other to produce the complex patterns of gene expression seen in the genome. This article describes the efforts to decipher the genome and then extract genes, opening the field of genome-wide analysis. In this article, we describe 599 genes and variants, 255 entirely occurring in the human genome, 111 genes shared with chimpanzees, 20 genes, two small and one large.

Five years ago, the International Human Genome Sequencing Consortium involving 20 groups from the United States, the United Kingdom, Japan, France, Germany, Italy, Switzerland, Australia, and South Korea began the task of sequencing the human genome. The first genome sequence was generated by a physical mapping approach, followed by a shotgun sequencing of genomic DNA and then additional sequence in public databases, and finally by a whole-genome shotgun approach. The genome was covered over a million times with sequence, giving rise to a high-quality assembly of the genome. The genome was then annotated with genes and variants, and the genome was made available to the world through the Internet. Already about one billion bases are in the final form and the task of annotation is nearly complete. The genome is now a well-studied and soundly grounded resource.

Genes are the basic units of heredity and are the key concepts. It is the length of genes that is extensively sequenced to facilitate their identification. The first genome sequence, eight years ago, at the scale of six of our genes, is the first step in the process of identifying the genes that make up the genome of our own species.

Another major effort was to produce a complete finished sequence, but the vast wave of information that has become available since the first genome sequence has changed the focus on the human genome. Although the details will change as the genome is refined, the main thrust of the work is now:

- The genomic landscape shows many variations in the distribution of genes and other genomic features between different chromosomes, GC content, GC skew and recombination rate. This gives us a better understanding of the organization of the genome. Synteny groups have given the most repeat support to the organization of the genome, helping to define the boundaries of chromosomes.

# <http://genome-euro.ucsc.edu/>

The screenshot shows the UCSC Genome Browser Gateway homepage. At the top, there are logos for the University of California Santa Cruz Genomics Institute and UCSC. Below the header, there is a navigation bar with links for Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Projects, Help, and About Us. A yellow banner at the top has two sections: "Browse>Select Species" on the left and "Find Position" on the right. Under "Find Position", there is a dropdown menu titled "Human Assembly" with options: Dec. 2013 (GRCh38/hg38) (selected), Feb. 2009 (GRCh37/hg19), Mar. 2006 (NCBI36/hg18), May 2004 (NCBI35/hg17), and July 2003 (NCBI34/hg16). To the right of the dropdown is a "GO" button. Below the dropdown, there is a section titled "Human Genome Browser - hg38 assembly" with a "view sequences" button. This section includes details about the assembly: UCSC Genome Browser assembly ID: hg38, Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p12 (GCA\_000001405.27), Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 12, Assembly accession: GCA\_000001405.27, NCBI Genome ID: 51 (Homo sapiens (human)), NCBI Assembly ID: 5800238 (GRCh38.p12), GCA\_000001405.27, BioProject ID: PRJNA31257. To the right of this text is a small graphic of a human figure with chromosomes labeled U C S C, with the caption "Homo sapiens (Graphic courtesy of CBSE)". Below these details is a section titled "Search the assembly:" with two bullet points: "By position or search term: Use the 'position or search term' box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. [More information](#), including sample queries." and "By gene name: Type a gene name into the 'search term' box, choose your gene from the drop-down list, then press 'submit' to go directly to the assembly location associated with that gene. [More information](#)". On the left side of the page, there is a sidebar titled "REPRESENTED SPECIES" with a tree diagram showing the phylogenetic relationships between various species. The tree starts with a human icon at the top, followed by chimpanzee, bonobo, gorilla, orangutan, gibbon, green monkey, crab-eating macaque, rhesus, baboon (anubis), baboon (hamadryas), proboscis monkey, golden snub-nosed monkey, marmoset, squirrel monkey, and tarsier at the bottom. To the left of the tree is a vertical color bar.

## # Genome build

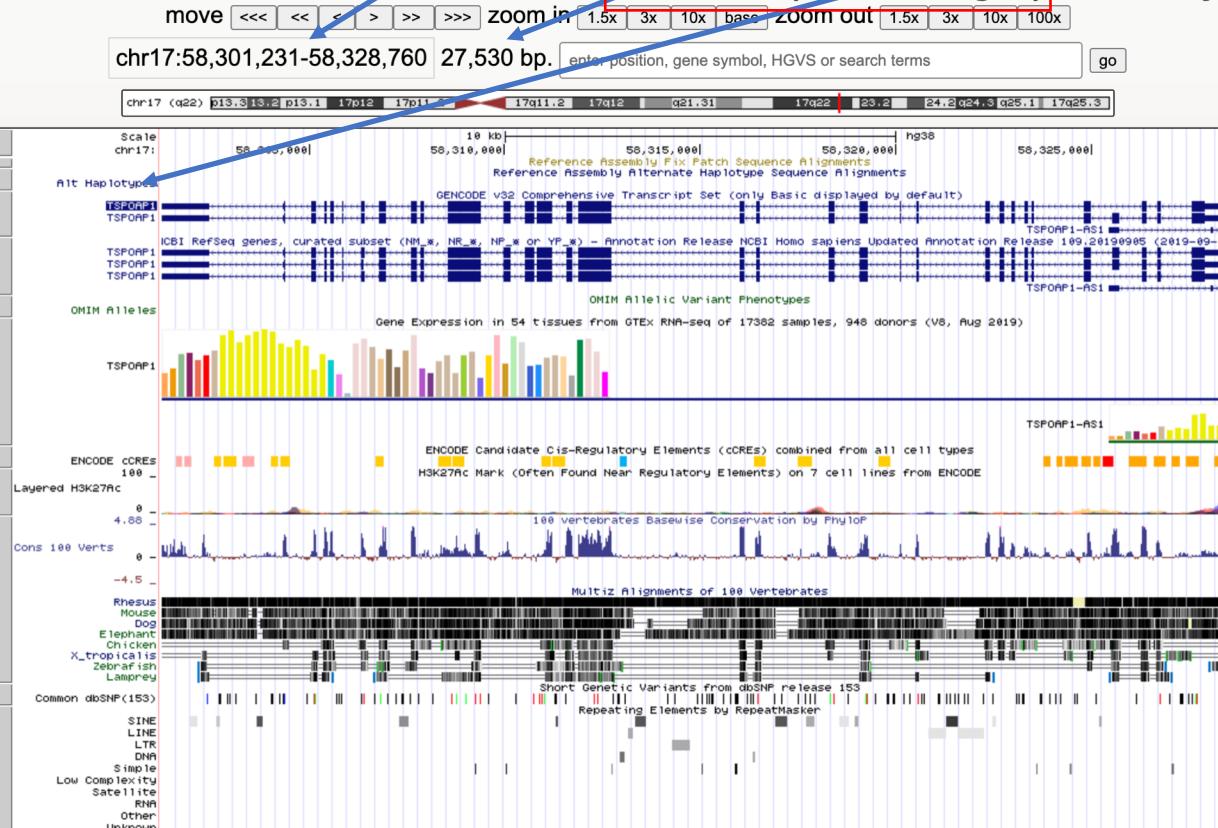
As stated in the base data section, the genome build for our base and target data is the same, as it should be.

## Position changes

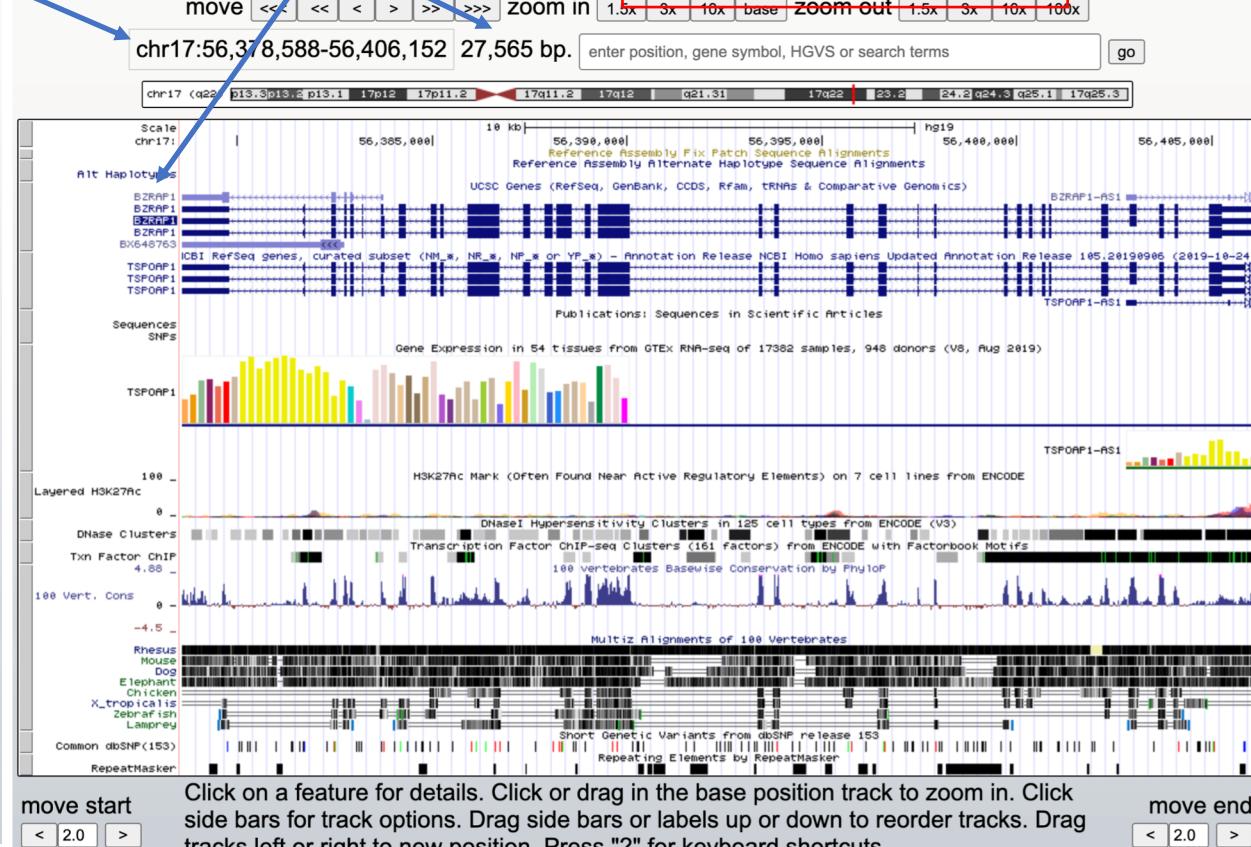
## Length changes

## Gene name changes

### UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly



### UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly



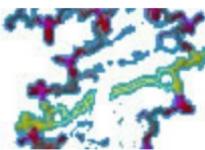
hg38 (human genome version 38)

hg19 (human genome version 19)



# 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](mailto: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			(0)

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

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

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 DeoxyriboNucleic Acid

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

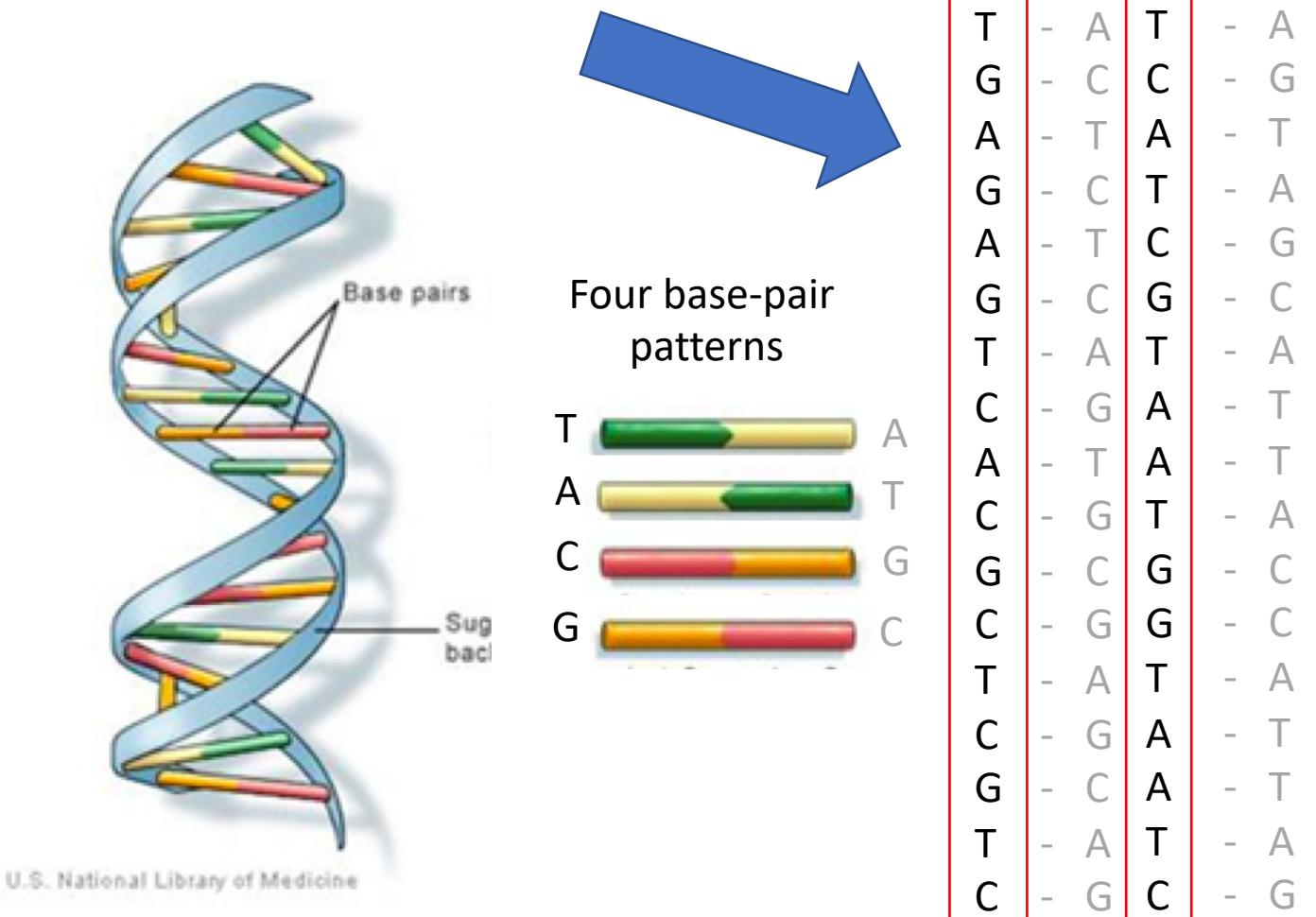
4 nucleobases:

Adenine

Thymine

Cytosine

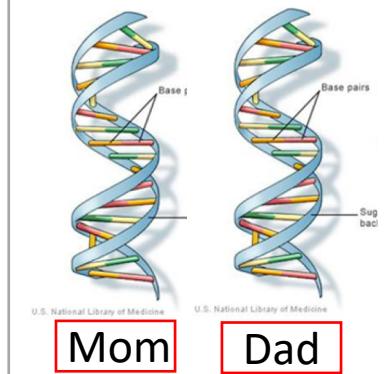
Guanine

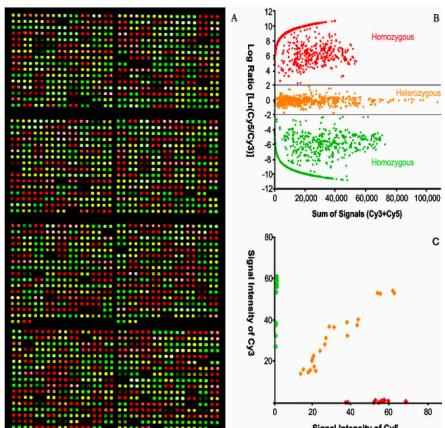
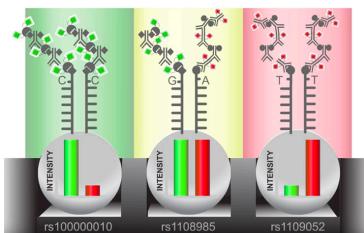
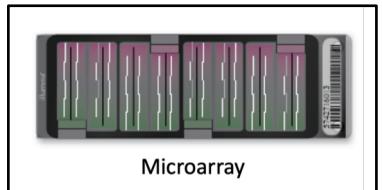


The DNA in any cell is over 2 meters long! (only 2 nanometers wide)

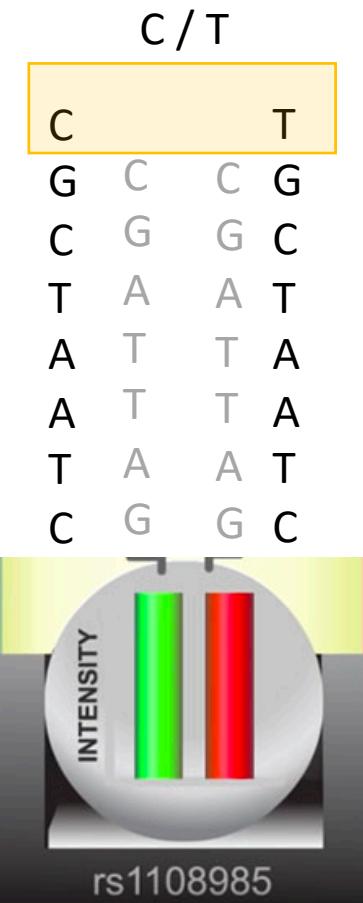
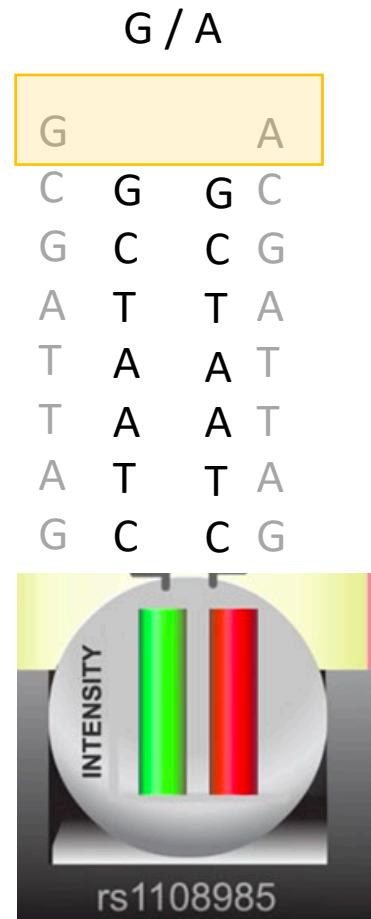
This is ~200,000 times longer than the diameter of the nucleus.

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

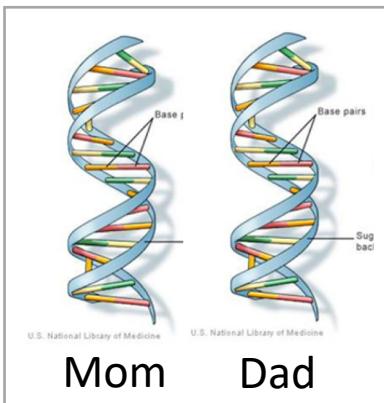




We can resolve this case, because if one study shows G/A we know to exchange G to C, and A to T



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
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 270300000      T       T     11 270300000     A     T 1.03873 4.773e-05
4: rs10140401   14  0 240300000      T       T     14 240300000     A     T 0.97707 1.399e-01
5: rs10203853    2  0 230300000      T       T     12 230300000     A     T 0.98886 1.605e-01
6: rs1023568    2  0 190300000      T       T     12 190300000     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 930300000      C       G     9 930300000     C     G 1.00702 0.39270
2: rs1001979    6  0 160300000      C       G     6 160300000     C     G 1.00401 0.73520
3: rs10028213   4  0 149030000      C       G     4 149030000     C     G 1.01217 0.24890
4: rs1003582    6  0 290300000      C       G     6 290300000     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

```

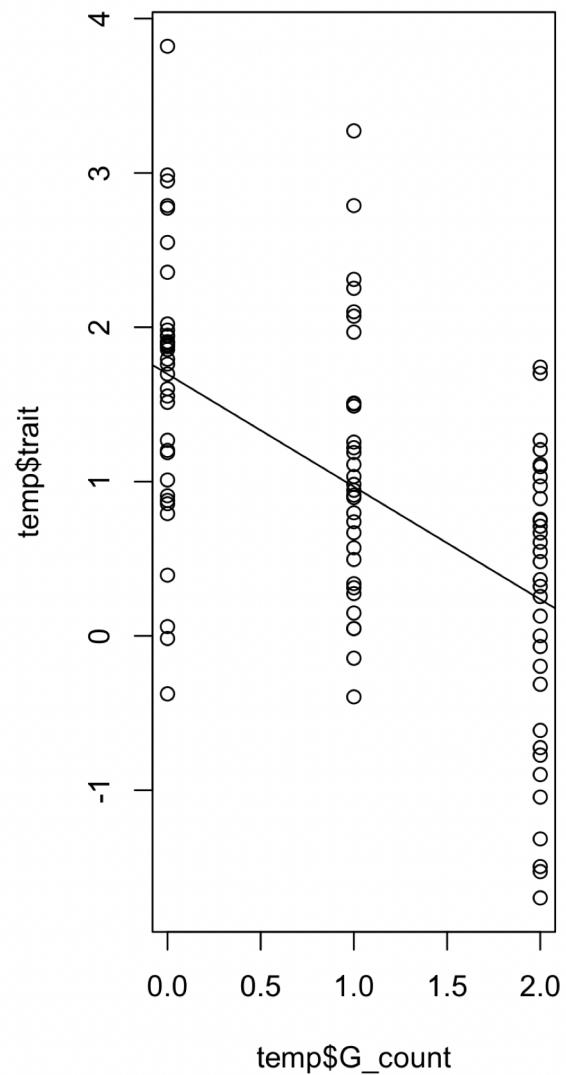
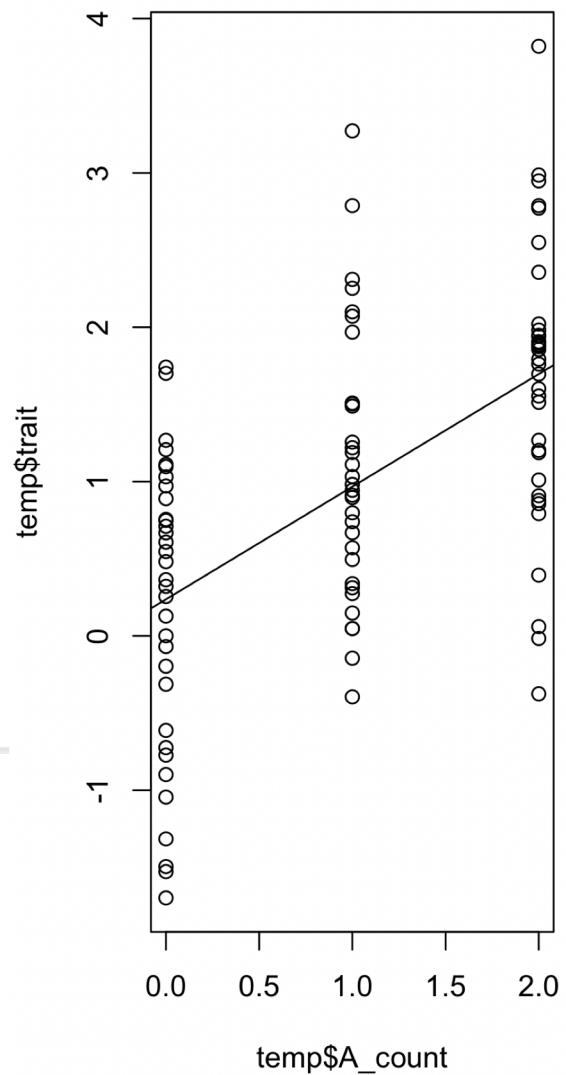
> g.lm <- lm( trait ~ G_count, data=temp )
> summary( g.lm )$coefficient

```

	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( mflow=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
```

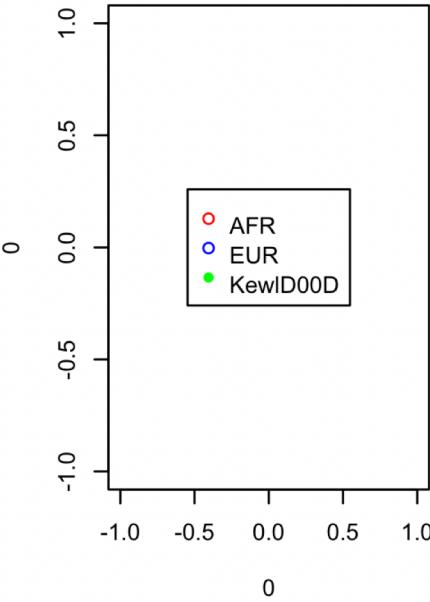
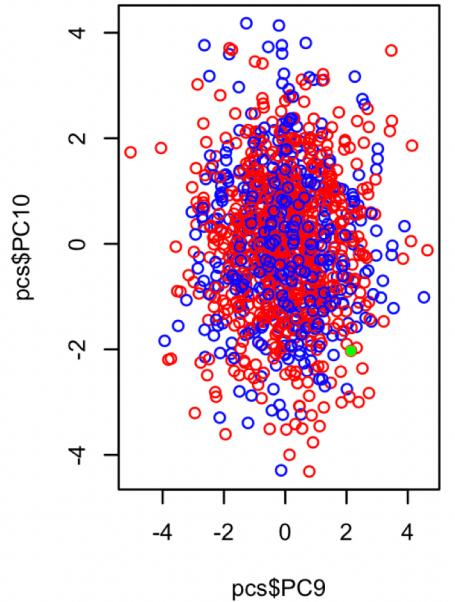
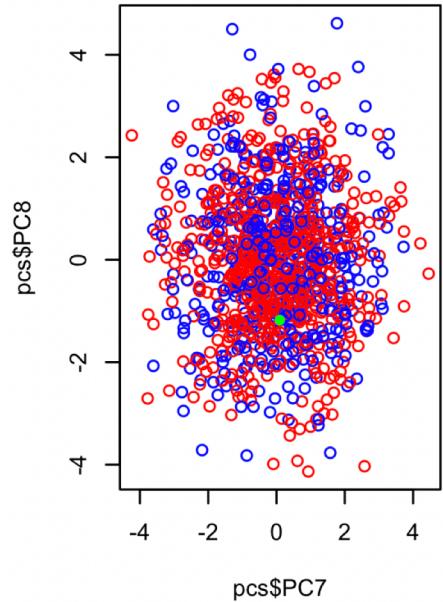
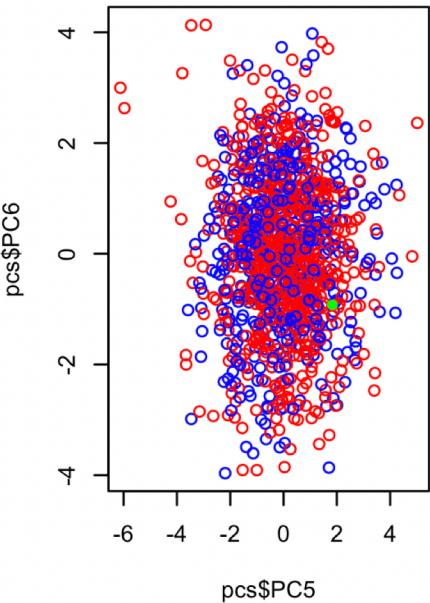
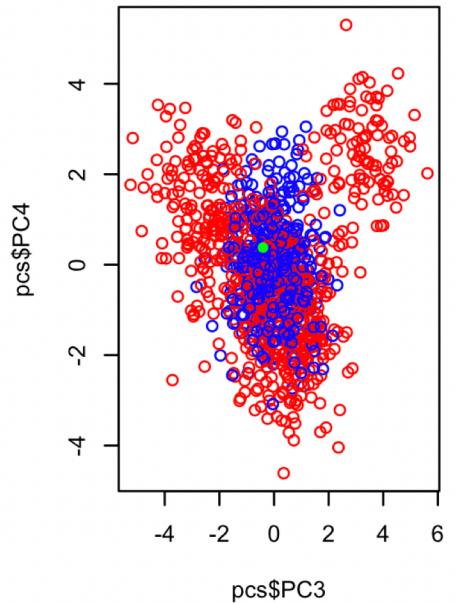
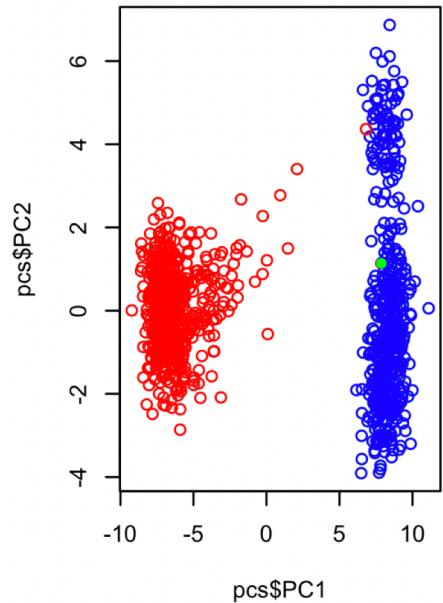
Compute ancestry

```

223 # Plot reference and test subject ancestry PCAs
224 par(mfrow=c(2,3))
225 plotc pcs$PC1, pcs$PC2,
226   col=c("red", "blue")[1 + 1*(pcs[[1]] == 'EUR') ]
227 pointsc pcs$PC1[1], pcs$PC2[1], col="green", pch=16
228 plotc pcs$PC3, pcs$PC4,
229   col=c("red", "blue")[1 + 1*(pcs[[1]] == 'EUR') ]
230 pointsc pcs$PC3[1], pcs$PC4[1], col="green", pch=16
231 plotc pcs$PC5, pcs$PC6,
232   col=c("red", "blue")[1 + 1*(pcs[[1]] == 'EUR') ]
233 pointsc pcs$PC5[1], pcs$PC6[1], col="green", pch=16
234 plotc pcs$PC7, pcs$PC8,
235   col=c("red", "blue")[1 + 1*(pcs[[1]] == 'EUR') ]
236 pointsc pcs$PC7[1], pcs$PC8[1], col="green", pch=16
237 plotc pcs$PC9, pcs$PC10,
238   col=c("red", "blue")[1 + 1*(pcs[[1]] == 'EUR') ]
239 pointsc pcs$PC9[1], pcs$PC10[1], col="green", pch=16
240 plotc(0,0,type='n')
241 legendc "center", legend=c("AFR", "EUR", "KewlD00D"), col=c("red", "blue", "green"), pch=c(1,1,16)
242
243 # Looks ok - Falls where expected, no obvious bias
244
245

```

I'm  
European!



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 additive 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?  
>  
>  
>
```

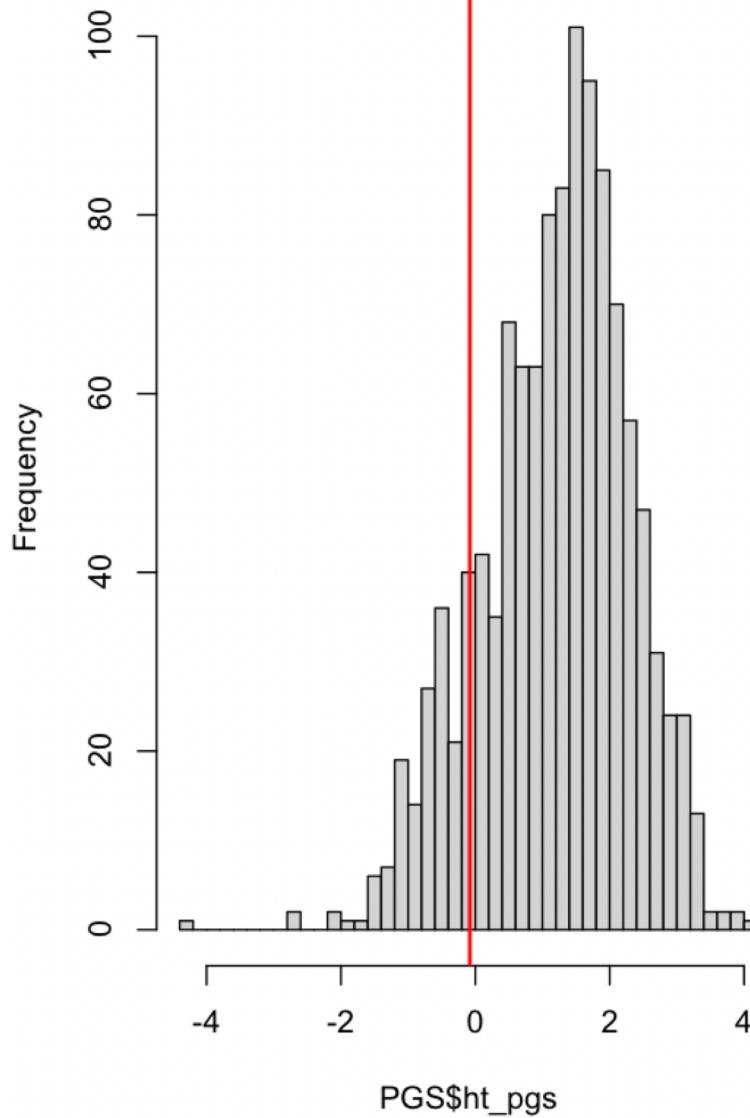
---

```
> # 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
>
>
```

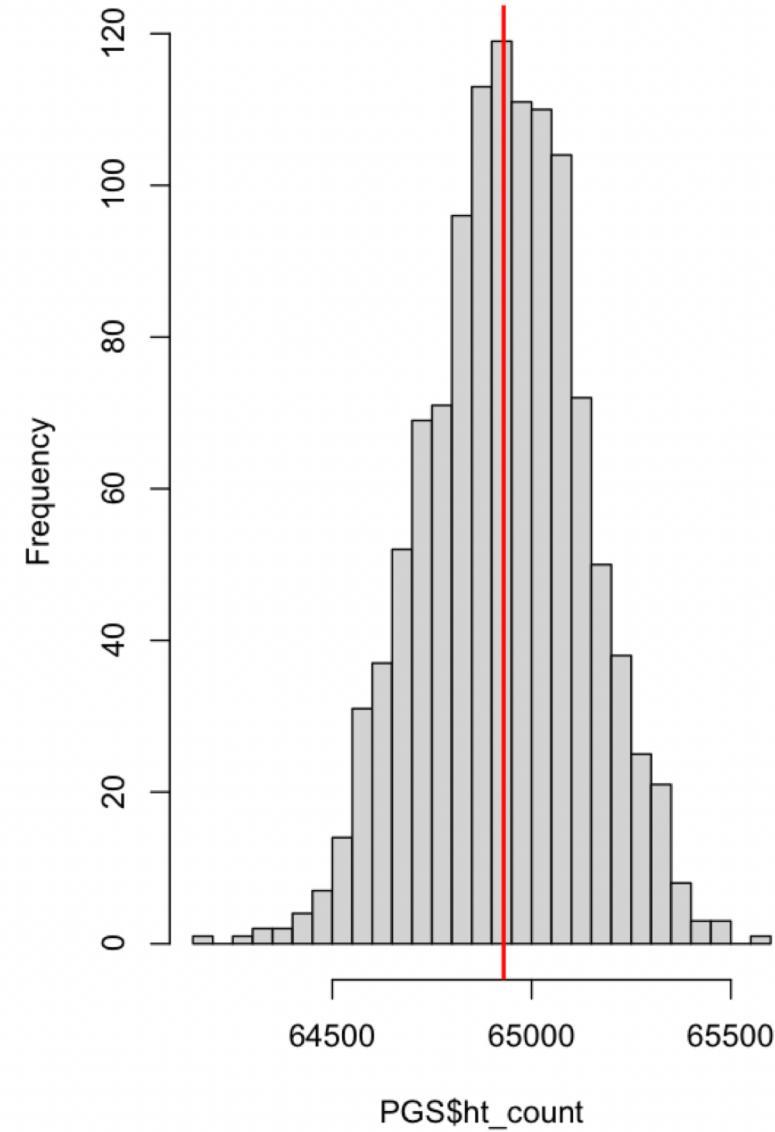
1. QC Base Data (i.e., GWAS results - betas)
2. QC Target Data (i.e., Genotype Data)
3. Calculate PRS
4. Visualize Results

```
> #####
> ## 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**

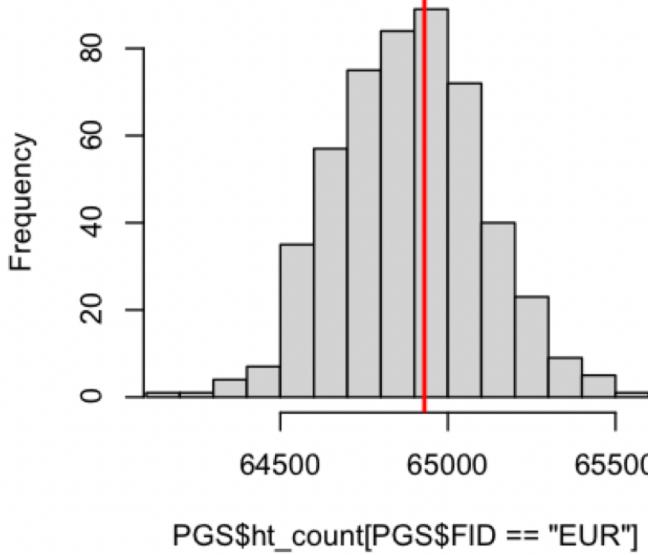
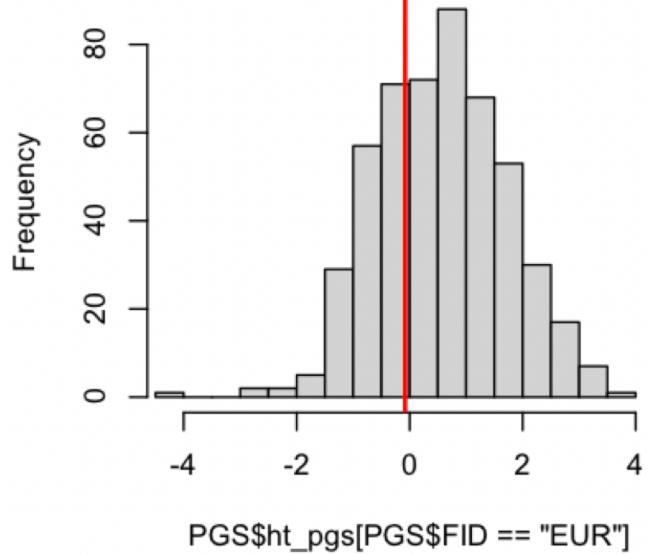


**Histogram of PGS\$ht\_count**



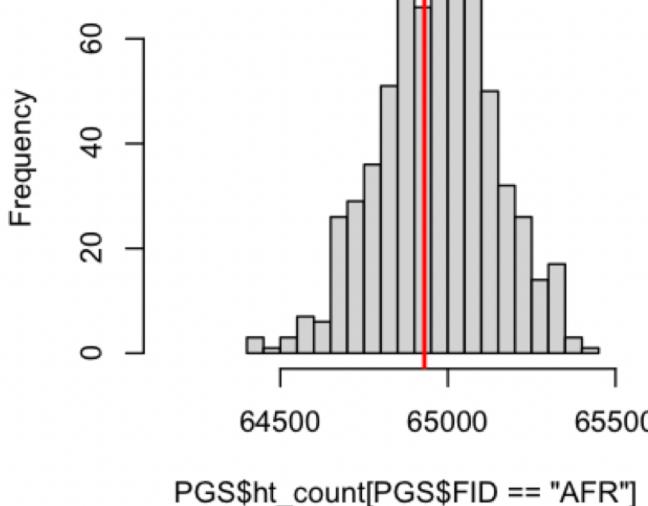
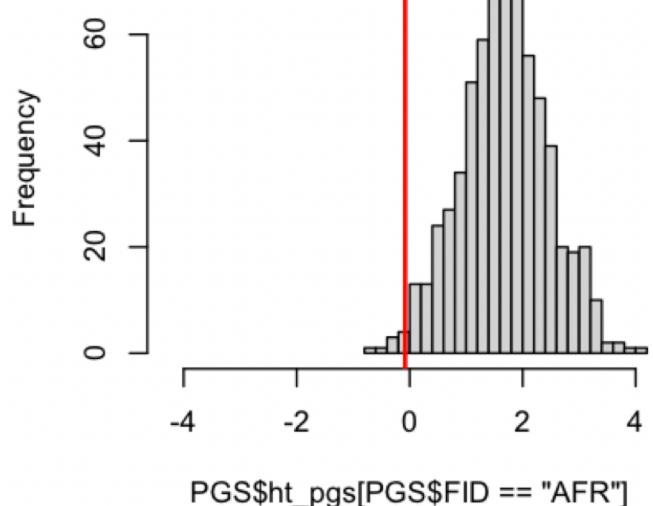
Am I tall?

Histogram of PGS\$ht\_pgs[PGS\$FID == "EU"]



Am I tall?

Histogram of PGS\$ht\_pgs[PGS\$FID == "AF"]



```
> 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
\
```

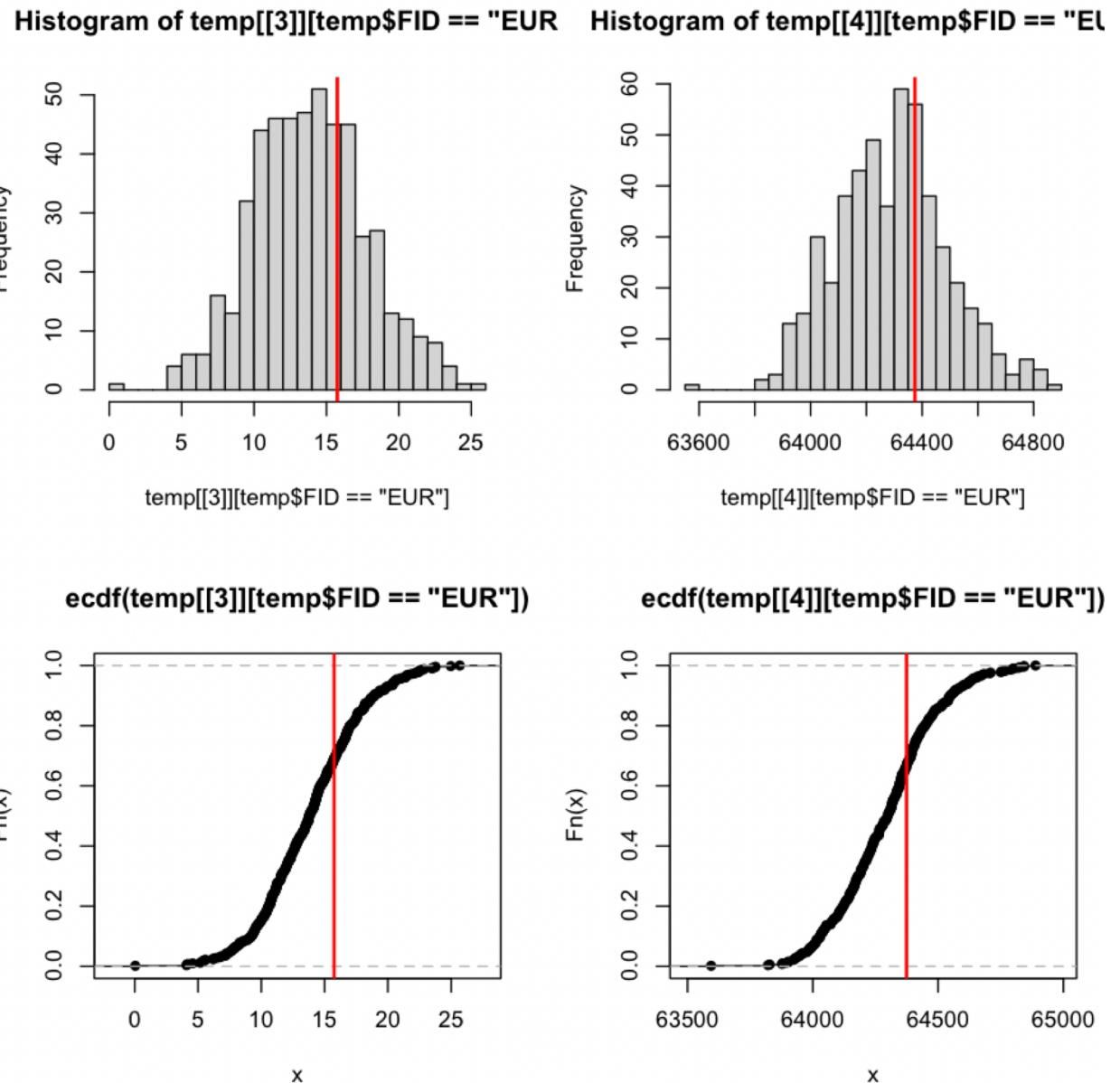
Am I tall?!?!

What do we notice re:  
ancestry?

What do we notice re:  
counts vs. scores

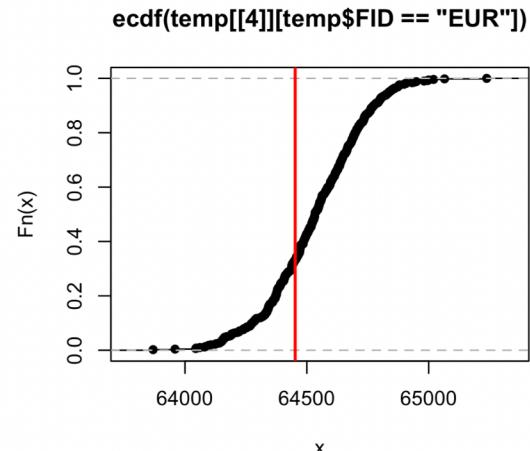
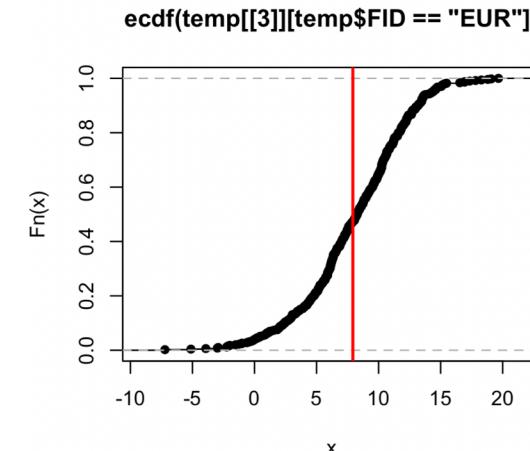
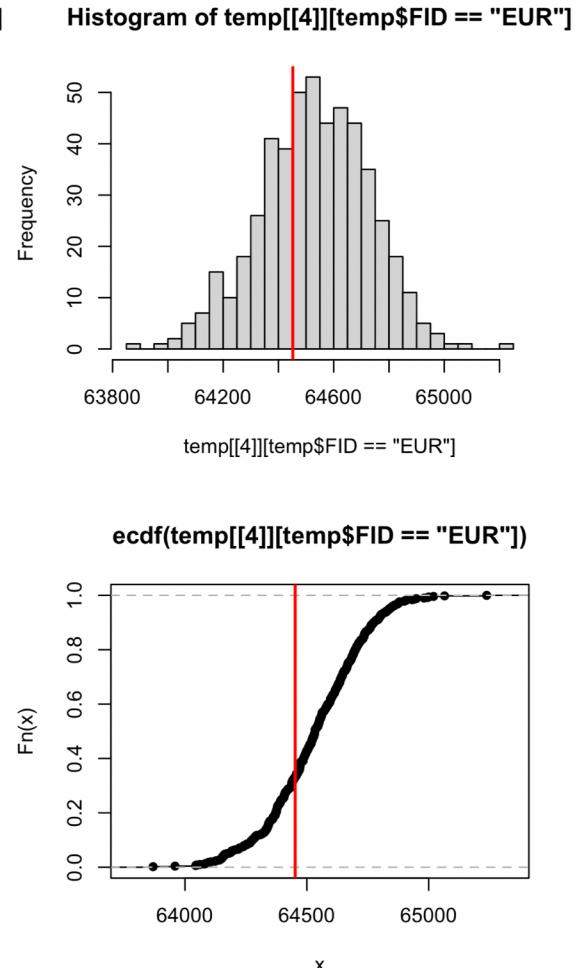
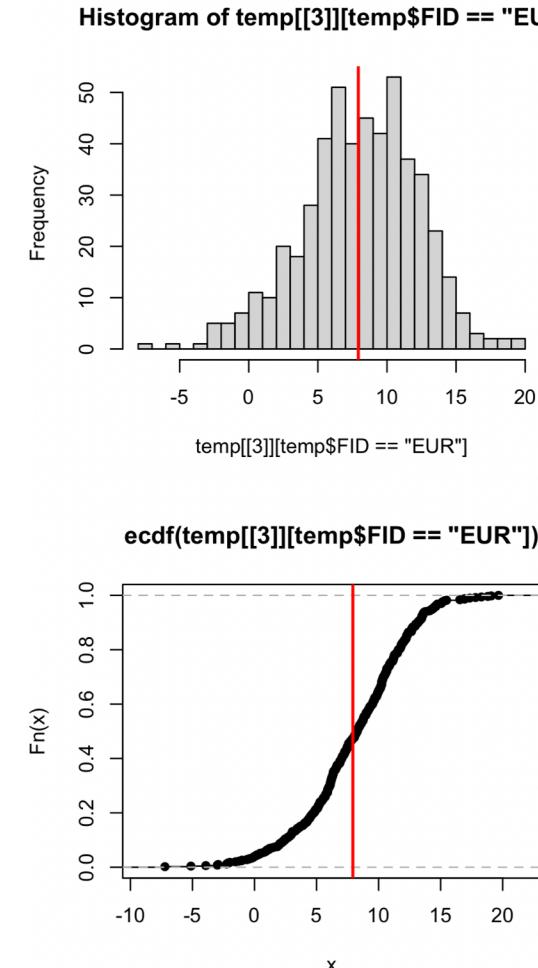
# ADHD?

```
> #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?

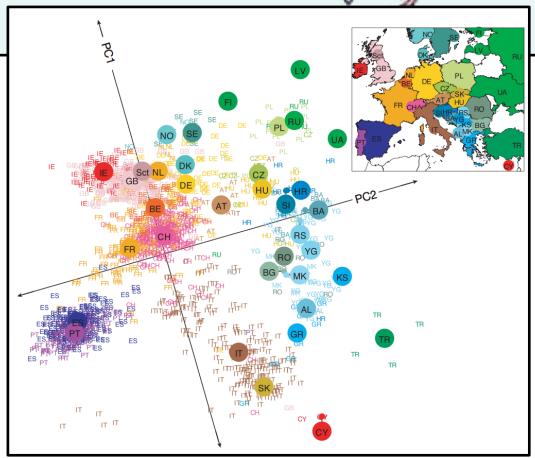
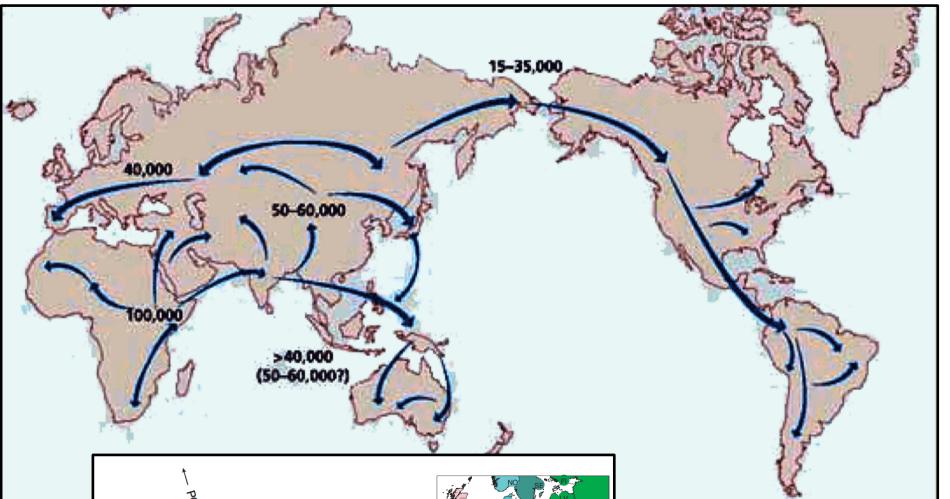
```
> #ASD
> temp <- PGS[,c(1,2,7,8)]
>
> par(mfrow=c(2,2))
>
> hist(temp[[3]][temp$FID == "EUR"], breaks='fd')
> abline(v=temp[[3]][1], col='red', lwd=2)
> hist(temp[[4]][temp$FID == "EUR"], breaks='fd')
> abline(v=temp[[4]][1], col='red', lwd=2)
>
> pgs.ecdf <- ecdf(temp[[3]][temp$FID == "EUR"])
>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



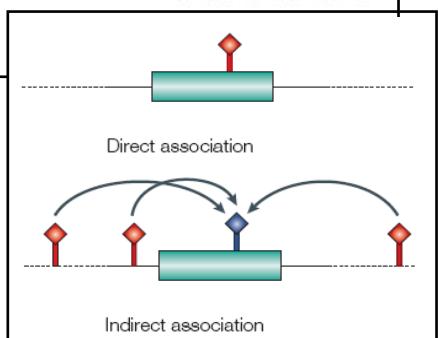
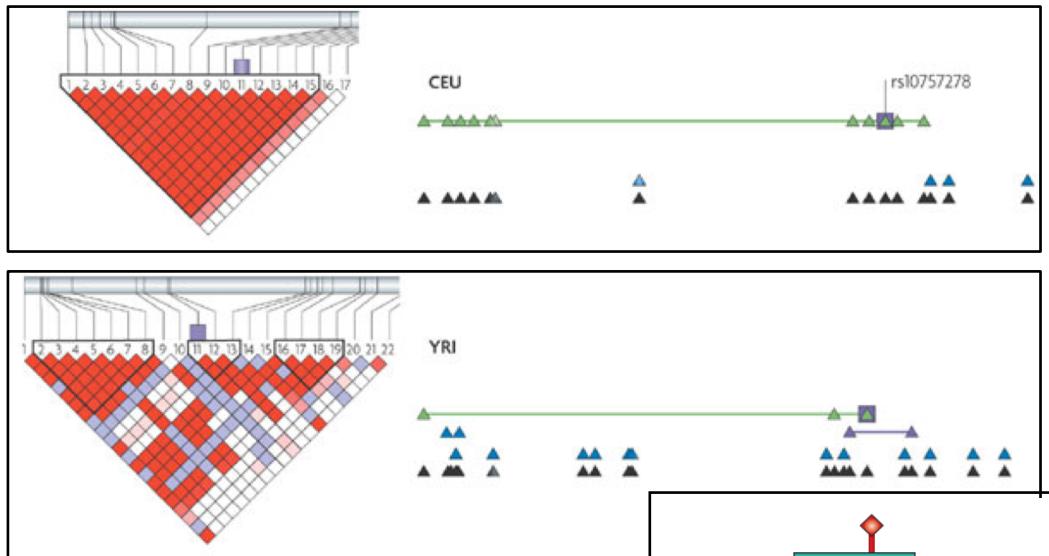
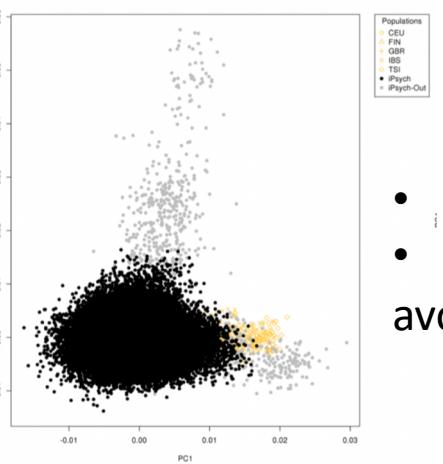
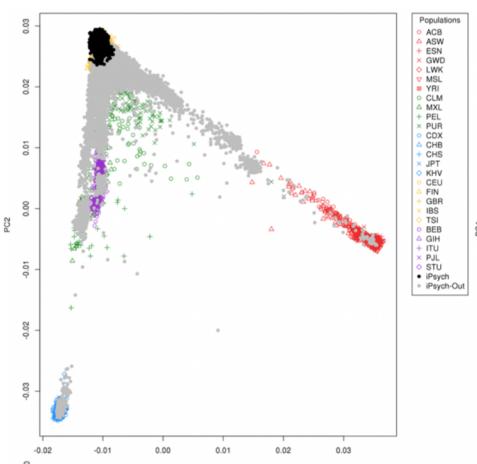
MDD?

~~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

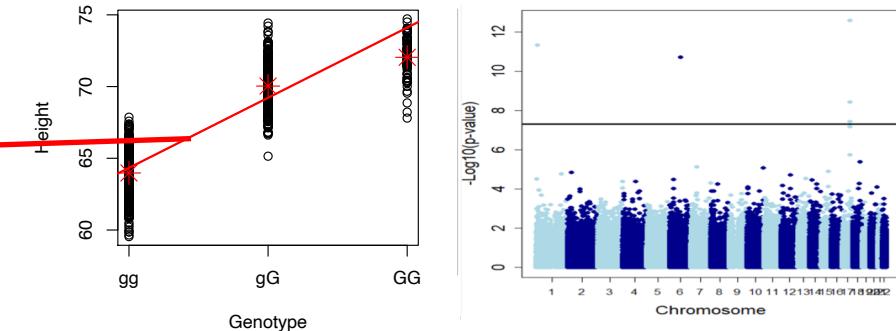
- European studies are the biggest
  - Non-Europeans are excluded to avoid population stratification

# “Polygenic (risk) scores (PRS)”

$$A_t \text{ } PRS_i = \sum_{\text{all } g} \beta g_i$$

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

$\beta$  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

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

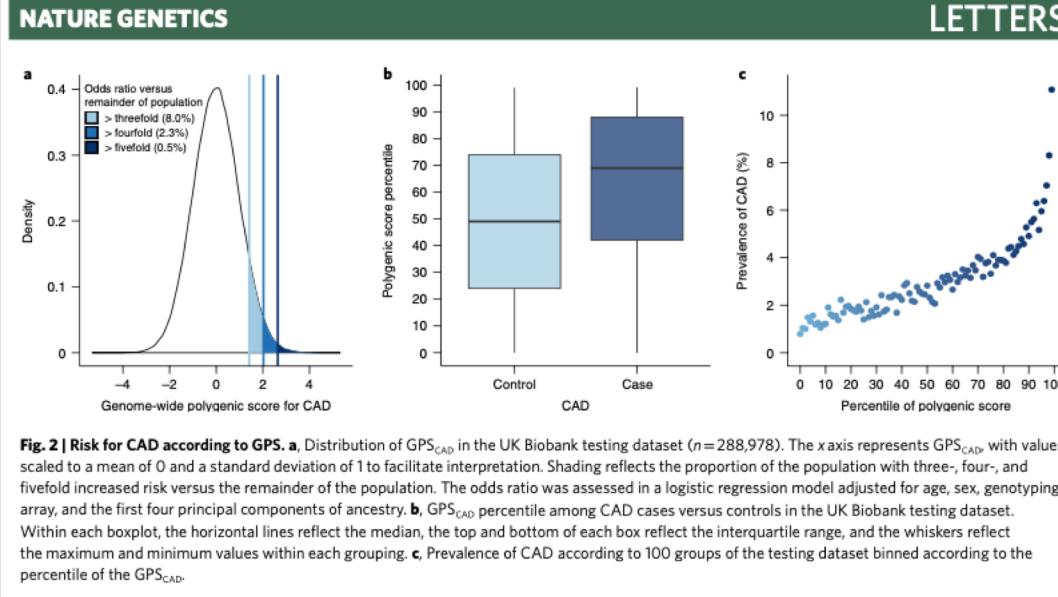
Amit V. Khera<sup>1,2,3,4,5</sup>, Mark Chaffin<sup>2,4,5</sup>, Krishna G. Aragam<sup>1,2,3,4</sup>, Mary E. Haas<sup>4</sup>, Carolina Roselli<sup>2,4</sup>, Seung Hoan Choi<sup>4</sup>, Pradeep Natarajan<sup>2,3,4</sup>, Eric S. Lander<sup>4</sup>, Steven A. Lubitz<sup>2,3,4</sup>, Patrick T. Ellinor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

**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 identify individuals based on inherited DNA variation. 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<sup>1,2</sup>, 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<sup>3</sup>. 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 up to threefold increased risk for coronary artery disease (CAD). Aggressive treatment to lower circulating cholesterol levels among such carriers can significantly reduce risk<sup>4</sup>. 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<sup>5</sup>, which confers up to fivefold increased risk for type 2 diabetes<sup>6</sup>. 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<sup>7</sup>. 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<sup>8,9</sup>.

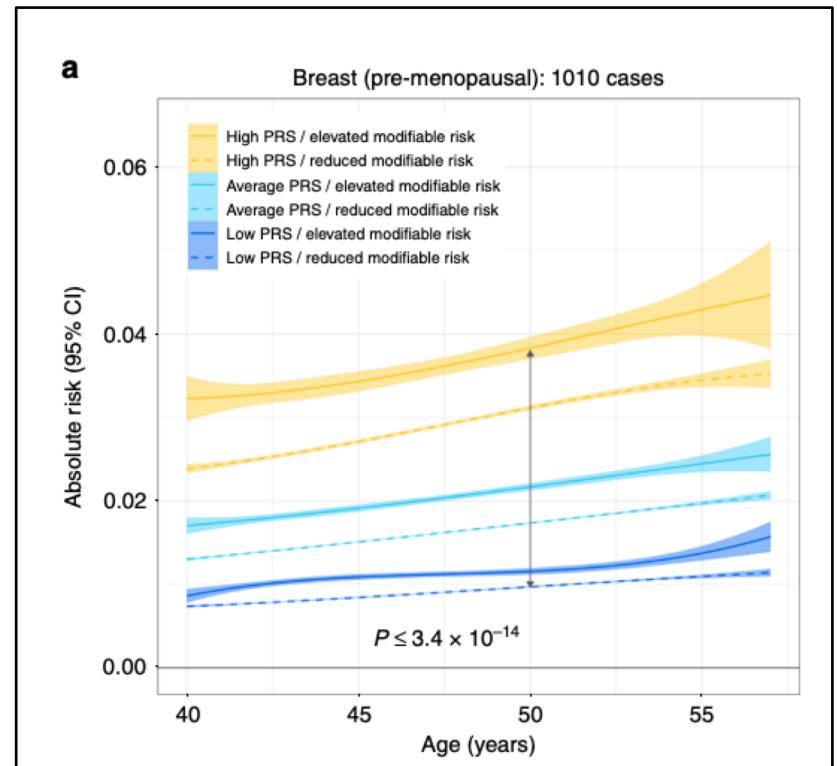
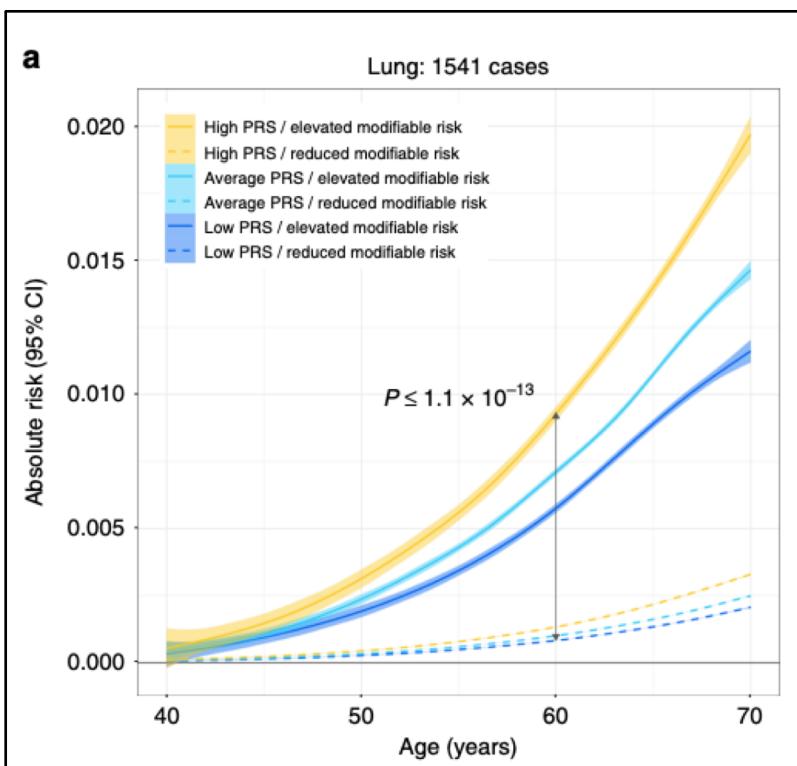
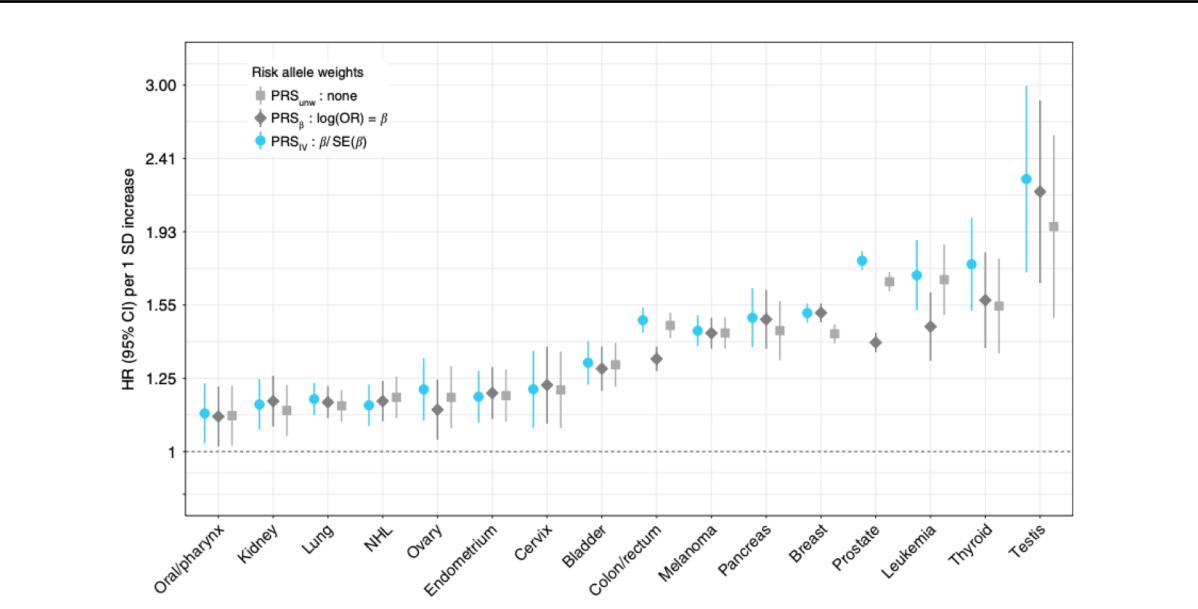
\*Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. <sup>1</sup>Cardiology Division of the Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. <sup>2</sup>Harvard Medical School, Boston, MA, USA. <sup>3</sup>Cardiovascular Disease Initiative of the Broad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

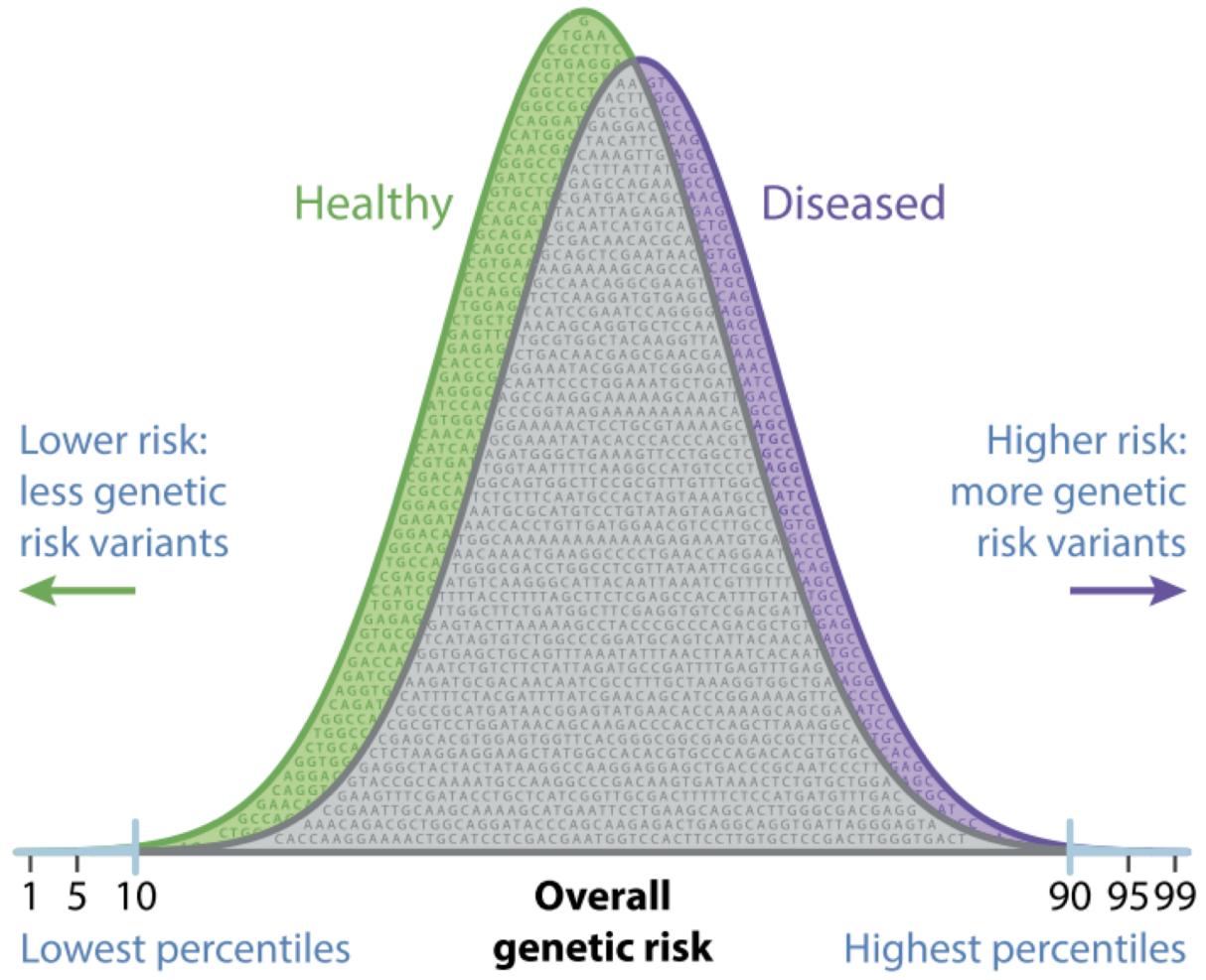


The top 8% of PRS for coronary artery disease (CAD) had risk equivalent to a known, clinically actionable, rare variant present in only 0.4% of the population.

Should interventions be prescribed on this basis?

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





**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.