How to perform a GWAS: MSK emphasis

Daniel S. Evans
Senior Scientist, CPMCRI
Adjunct Associate Professor, Department of Epidemiology and Biostatistics, UCSF

Feb 15, 2024

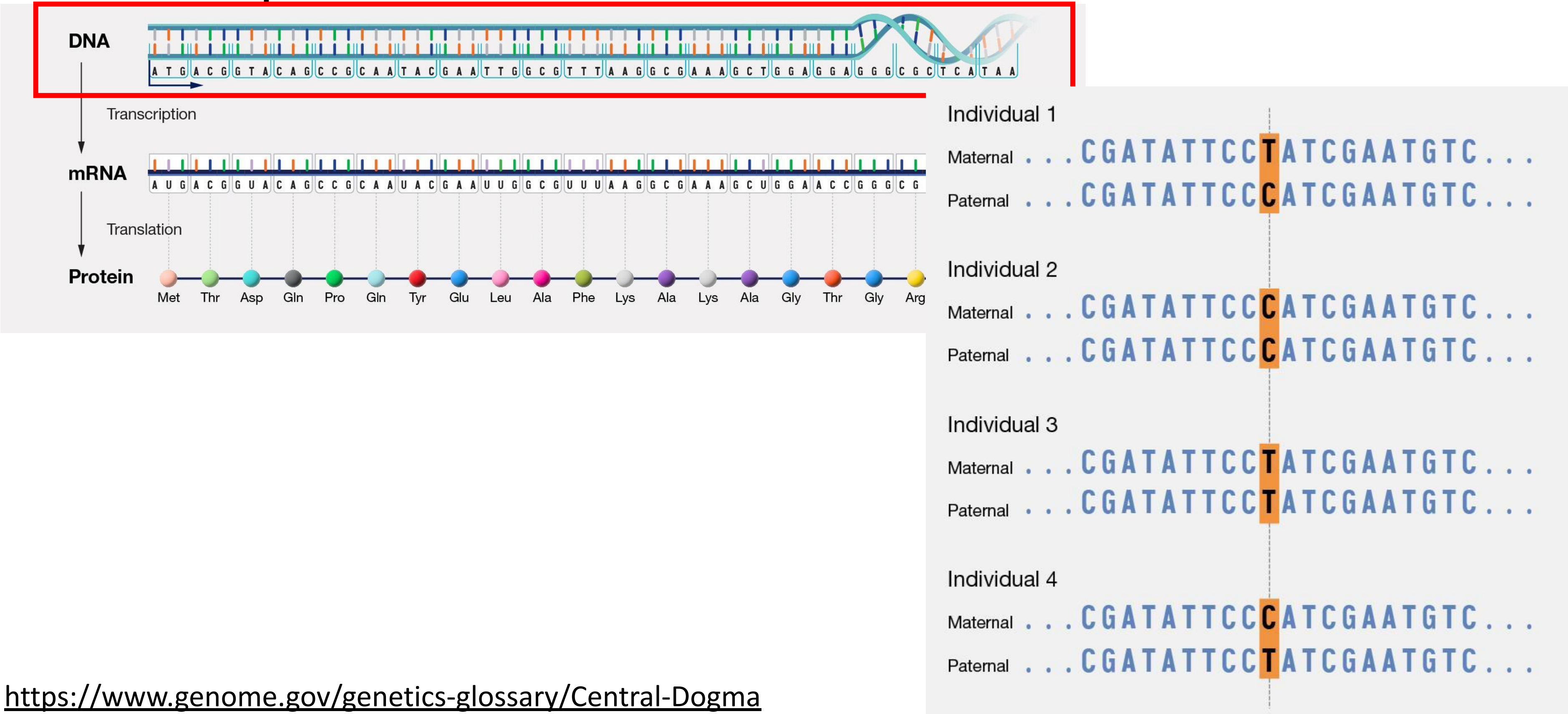
Outline

- GWAS background
- GWAS in the CCMBM
- GWAS at UCSF (regulatory, computing resources)
- Introduce the tools (linux, plink, R)
- Example using genome-wide genotypes from the Osteoarthritis Initiative (OAI)

Human genome

- 23 chromosome pairs (22 autosomes + sex chromosomes)
- Mitochondrial genome
- ~ 3 billion base pairs in haploid genome
- ~ 300 million single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) from TOPMed
  - ~1 variant every 10 bp

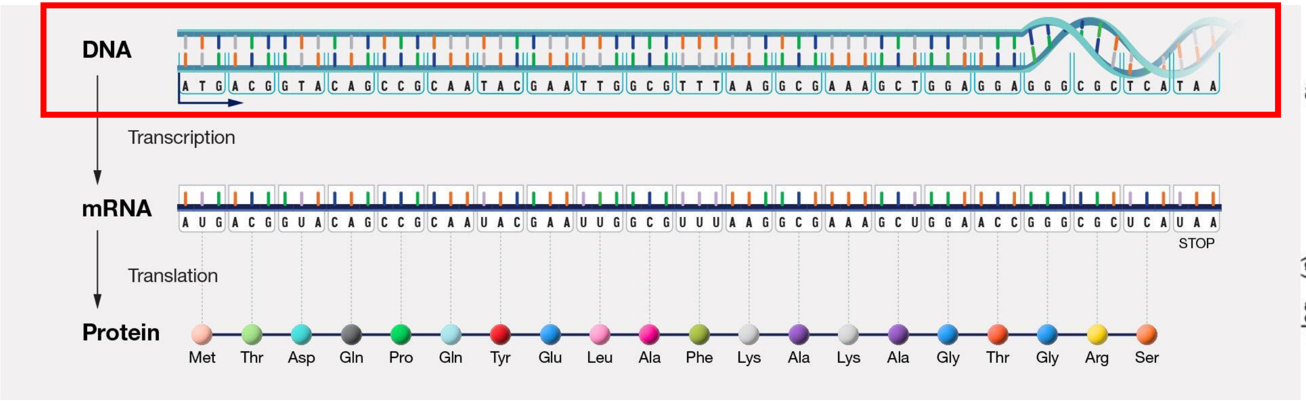
DNA sequence variation

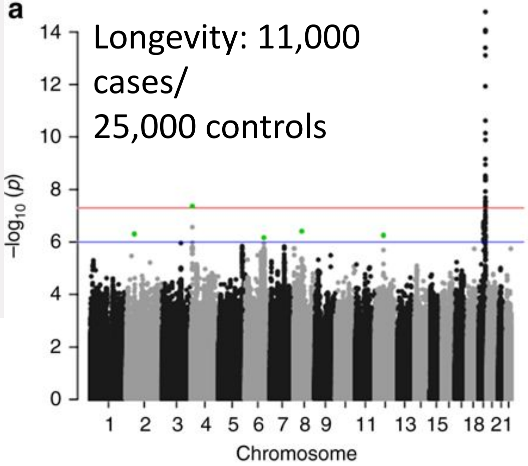

DNA
A T G A C G G T A C A G C C G C A A T A C G A A T T G G C G T T T A A G G C G A A A G C T G G A G G A G G G C G C T C A T A A
Transcription
mRNA
A U G A C G G U A C A G C C G C A A U A C G A A U U G G C G U U U A A G G C G A A A G C U G G A A C C G G G C G
Translation
Protein
Met
Thr
Asp
Gln
Pro
Gln
Tyr
Glu
Leu
Ala
Phe
Lys
Ala
Lys
Ala
Gly
Thr
Gly
Arg
Individual 1
Maternal . . . CGATATTCCTATCGAATGTC . . .
Paternal . . . CGATATTCCCATCGAATGTC . . .
Individual 2
Maternal . . . CGATATTCCCATCGAATGTC . . .
Paternal . . . CGATATTCCCATCGAATGTC . . .
Individual 3
Maternal . . . CGATATTCCTATCGAATGTC . . .
Paternal . . . CGATATTCCTATCGAATGTC . . .
Individual 4
Maternal . . . CGATATTCCCATCGAATGTC . . .
Paternal . . . CGATATTCCTATCGAATGTC . . .

https://www.genome.gov/genetics-glossary/Central-Dogma

https://www.genome.gov/genetics-glossary/Single-Nucleotide-Polymorphisms

Genome-wide association studies (GWAS)


DNA
A T G A C G G T A C A G C C G C A A T A C G A A T T G G C G T T T A A G G C G A A A G C T G G A G G A G G G C G C T C A T A A
Transcription
mRNA
A U G A C G G U A C A G C C G C A A U A C G A A U U G G C G U U U A A G G C G A A A G C U G G A A C C G G G C G C U C A U A A
STOP
Translation
Protein
Met
Thr
Asp
Gln
Pro
Gln
Tyr
Glu
Leu
Ala
Phe
Lys
Ala
Lys
Ala
Gly
Thr
Gly
Arg
Ser



Longevity: 11,000 cases/
25,000 controls

https://www.genome.gov/genetics-glossary/Central-Dogma
https://www.genome.gov/genetics-glossary/Single-Nucleotide-Polymorphisms

Deelen, J., Evans, D.S., Arking, D.E. *et al.* A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat Commun* 10, 3669 (2019). https://doi.org/10.1038/s41467-019-11558-2

Association = trait levels differ by genotype

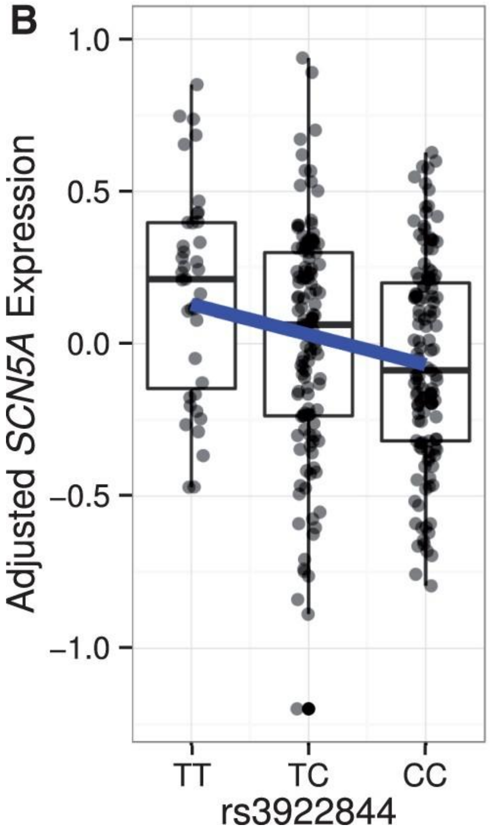
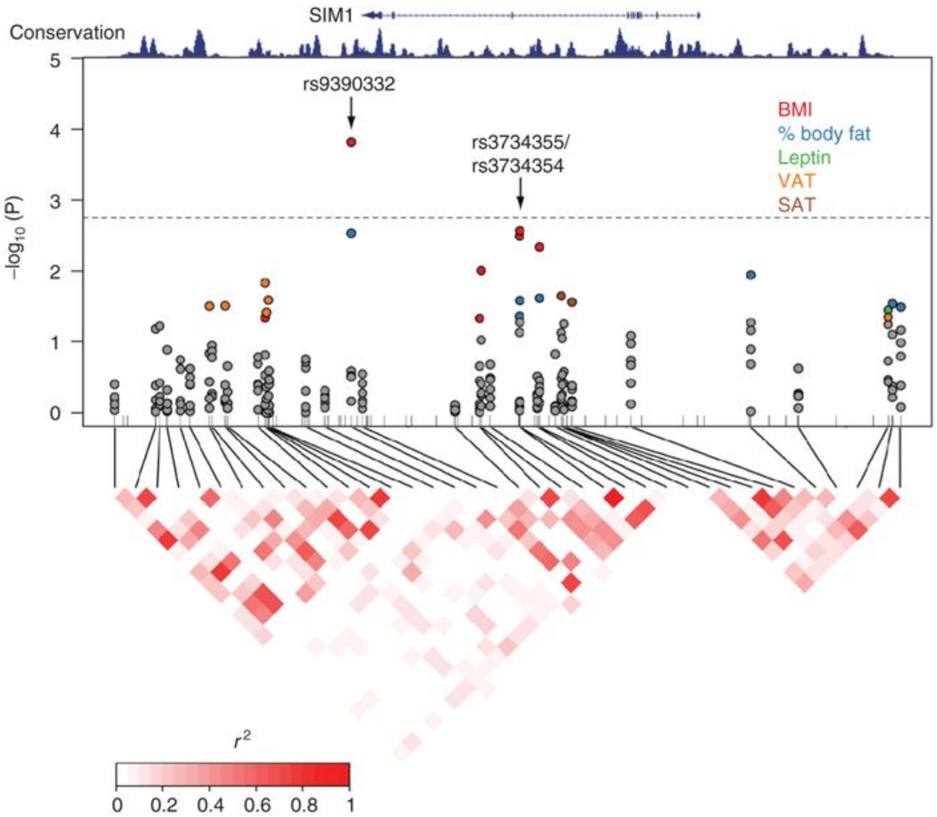


Figure 2. Association between rs3922844 genotype and SCN5A expression in human atrial tissue. Box plots display data ...

Hum Mol Genet, Volume 25, Issue 19, 1 October 2016, Pages 4350–4368, https://doi.org/10.1093/hmg/ddw284

Linkage Disequilibrium = linked markers on a chromosome



Replication and Extension of Association Between Common Genetic Variants in SIM1 and Human Adiposity

Obesity, Volume: 19, Issue: 12, Pages: 2394-2403, First published: 06 September 2012, DOI: (10.1038/oby.2011.79)

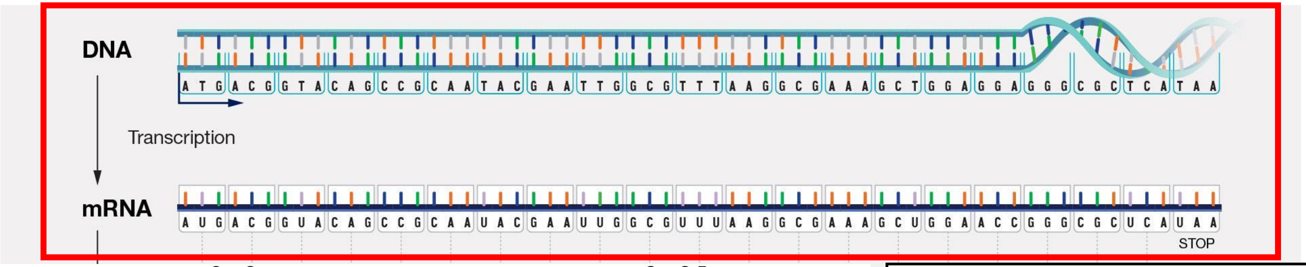
CCMBM cores

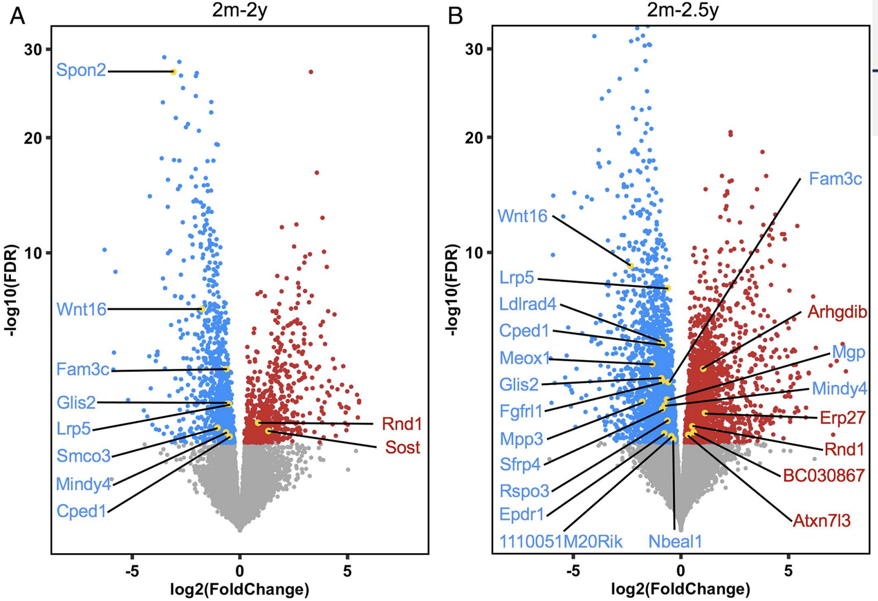
- EBSD Core - Studies with genome-wide genotype data. MrOS, SOF, OAI, Health ABC
- Imaging Core - genetic variants associated with hip shape

Baird DA, Evans DS, Kamanu FK, et al. Identification of Novel Loci Associated With Hip Shape: A Meta-Analysis of Genomewide Association Studies. *J Bone Miner Res*. 2019;34(2):241-251. doi:10.1002/jbmr.3605

- Skeletal biology Core - genes identified in model organisms, do these same genes have genetic associations with comparable trait in humans?

RNA and DNA integration, mouse -> human


DNA
A T G A C G G T A C A G C C G C A A T A C G A A T T G G C G T T T A A G G C G A A A G C T G G A G G A G G G C G C T C A T A A
Transcription
mRNA
A U G A C G G U A C A G C C G C A A U A C G A A U U G G C G U U U A A G G C G A A A G C U G G A A C C G G G C G C U C A U A A
STOP



- Genes with femur RNA expression that changes with age in mouse, are enriched for GWAS associations with BMD and fracture in humans
- Wnt16, Lrp5, Sost = known
- Glis2, Smco3, Spon2, Nbeal1 = not genome-wide significant, but evidence in mouse and human
- www.mouse2human.org

J of Bone & Mineral Res, Volume: 37, Issue: 4, Pages: 804-817, First published: 29 January 2022, DOI: (10.1002/jbmr.4516)

Regulatory compliance

- Plan ahead!
- Human research training and certifications
  - https://irb.ucsf.edu/human-research-protection-program-hrpp
- IRB
  - https://irb.ucsf.edu/human-research-protection-program-hrpp
- Data use agreement
  - Data provider. dbGaP, your colleague, ….
  - Agreement to transfer data between institutions

Wynton compute cluster

- https://wynton.ucsf.edu/hpc/index.html
- Genetics can be an identifier, so request a Wynton Protected Health Information (PHI) account
- https://it.ucsf.edu/standard-guideline/ucsf-policy-650-16-addendum-f-ucsf-data-classification-standard#phi

Wynton login

```
$ ssh user@plog1.wynton.ucsf.edu
user@plog1.wynton.ucsf.edu's password:
```

```
[user@plog1 ~]$ ssh pdev1
user@pdev1's password:
##################################################
#       All connections are monitored and recorded.          #
#  Disconnect IMMEDIATELY if you are not an authorized user!  #
##################################################
Welcome to the Wynton development nodes.
```

Wynton storage

- 500 GB free
- Directly genotyped data (~1 million markers, 4000 people) = 2 GB
- Will run out of space quickly after imputing. Buy more storage.

Linux system

- pwd = present working directory
- cd = change directory
- ls = list
- tar = pack/unpack tar archive
- gz/gunzip = compress/decompress file
- vi = text editor
- ssh = secure shell. Open in terminal, connect to remote computers
- exit = disconnect

Linux simple filters

- head = print first 10 lines of a file
- cut = select column from a file
- uniq = many options: print unique values, print duplicate values, count occurences
- grep = search for a string within a file. Regular expressions can make the search term powerful.

Shell, plink, and R

- Shell
  - View and manipulate large files, launch programs as single or parallel jobs
- Plink (https://www.cog-genomics.org/plink/1.9/)
  - Specialized C++ program designed for high dimensional genetic data
- R
  - Statistical software. Filter Plink output and reports, generate plots

Genotype file types in plink

- Plink format - PED and MAP files
  - The PED file is a white-space (space or tab) delimited file: the first six columns are mandatory:

Family ID
Individual ID
Paternal ID
Maternal ID
Sex (1=male; 2=female; other=unknown)
Phenotype
Biallelic genotypes

- PED file for 2 people and 3 SNPs

```
FAM001  1  0 0  1  2  A A  G G  A C
FAM001  2  0 0  1  2  A A  A G  C C
```

Plink MAP file

- MAP file is a SNP annotation file

chromosome (1-22, X, Y or 0 if unplaced)
rs# or snp identifier
Genetic distance (morgans)
Base-pair position (bp units)

- The autosomes should be coded 1 through 22. The following other codes can be used to specify other chromosome types:

X	X chromosome	-> 23
Y	Y chromosome	-> 24
XY	Pseudo-autosomal region of X	-> 25
MT	Mitochondrial	-> 26

Plink binary files

- Accommodates larger files
- Separates sample annotation from genotypes, allowing genotypes to be stored in binary format.
- .bed = binary genotypes
- .fam = sample annotation. First 6 columns of old ped file
- .bim = variant annotation.

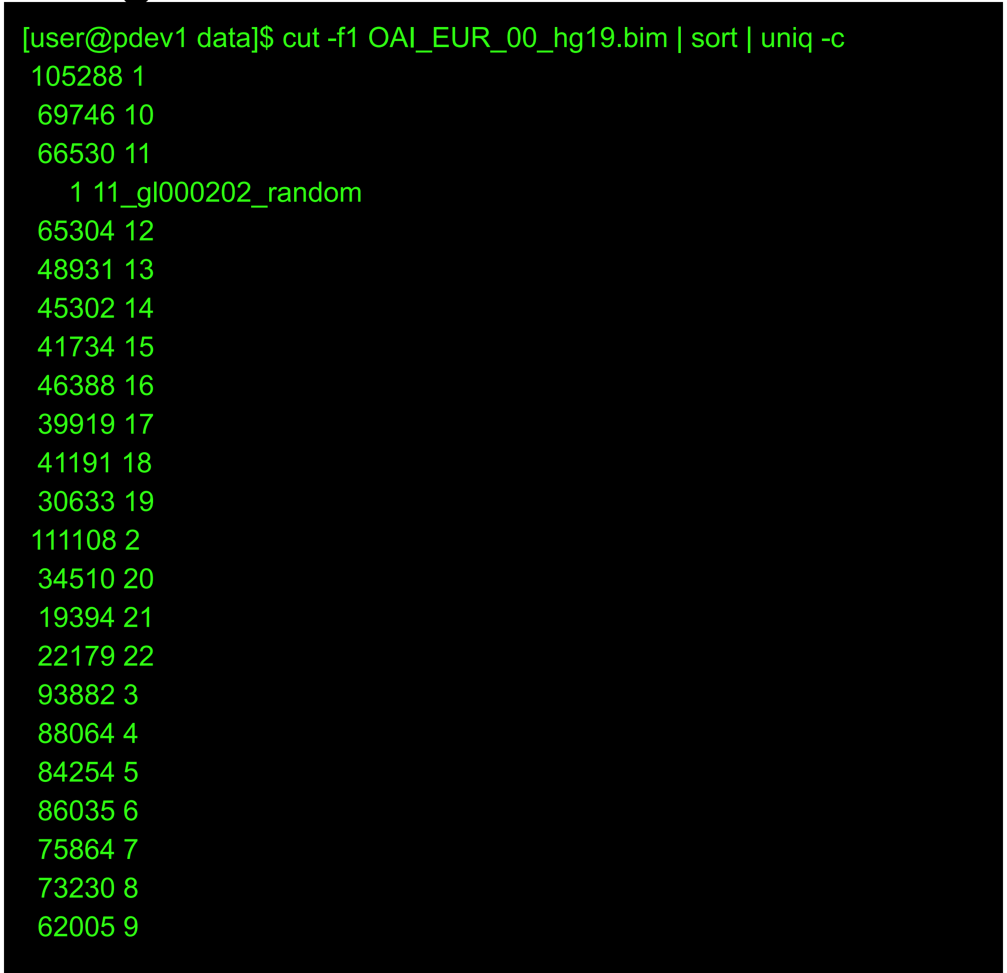
A text file with no header line, and one line per variant with the following six fields:

1. Chromosome code (either an integer, or 'X'/'Y'/'XY'/'MT'; '0' indicates unknown) or name
2. Variant identifier
3. Position in morgans or centimorgans (safe to use dummy value of '0')
4. Base-pair coordinate (1-based; limited to 231-2)
5. Allele 1 (corresponding to clear bits in .bed; usually minor)
6. Allele 2 (corresponding to set bits in .bed; usually major)

OAI GWAS as an example

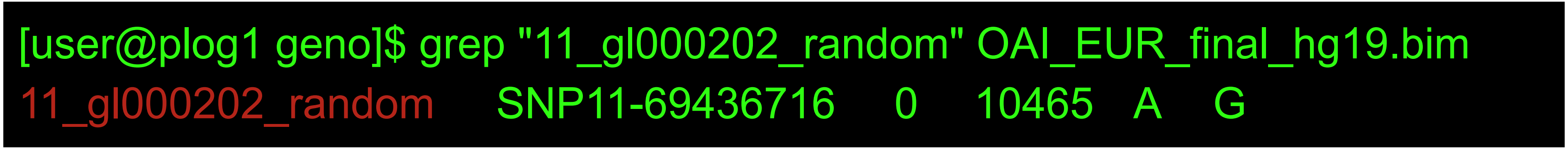
- The Osteoarthritis Initiative (OAI) is a prospective longitudinal cohort
- Primary aim = identifying risk factors for incidence and progression of tibiofemoral knee OA.
- Population-based recruitment to enroll 4,674 men and women between the ages of 45-79 years who either had radiographic symptomatic knee OA or who were without radiographic symptomatic OA in both knees but were considered high risk for OA because they had two or more known risk factors for knee OA.
- Subjects were recruited into the baseline phase of the OAI at multiple sites throughout the US between 2004 and 2006. All subjects were invited back for follow-up examinations to assess incidence or progression of OA annually, for up to 5 years.
- Illumina GWAS array on 4,129 participants

Browsing bim file



```
[user@pdev1 data]$ cut -f1 OAI_EUR_00_hg19.bim | sort | uniq -c
 105288 1
  69746 10
  66530 11
      1 11_gl000202_random
  65304 12
  48931 13
  45302 14
  41734 15
  46388 16
  39919 17
  41191 18
  30633 19
 111108 2
  34510 20
  19394 21
  22179 22
  93882 3
  88064 4
  84254 5
  86035 6
  75864 7
  73230 8
  62005 9
```

“random” variant?



```
[user@plog1 geno]$ grep "11_gl000202_random" OAI_EUR_final_hg19.bim
11_gl000202_random	SNP11-69436716	0	10465	A	G
```

Plink fam file

- 6 columns, no header, space delimited

  Family ID
  Individual ID
  Paternal ID
  Maternal ID
  Sex (1=male; 2=female; other=unknown)
  Phenotype

```
head -n2 OAI_EUR_00_hg19.fam | od -bc
0000000 071 060 060 060 060 071 071 040 061 063 062 067 063 061 060 060
          9   0   0   0   0   9   9       1   3   2   7   3   1   0   0
0000020 064 061 040 060 040 060 040 060 040 055 071 012 071 060 060 060
          4   1       0       0       0       -   9  \n   9   0   0   0
0000040 062 071 066 040 062 071 061 063 062 065 070 060 060 062 040 060
          2   9   6       2   9   1   3   2   5   8   0   0   2       0
0000060 040 060 040 060 040 055 071 012
              0       0       -   9  \n
```

Getting to know the fam (file)

```
[user@pdev1 data]$ head -n3 OAI_EUR_00_hg19.fam
9000099 1327310041 0 0 0 -9
9000296 2913258002 0 0 0 -9
9000622 1328410042 0 0 0 -9
```

```
cut -d " " -f1 OAI_EUR_00_hg19.fam | sort | uniq -d
cut -d " " -f2 OAI_EUR_00_hg19.fam | sort | uniq -d
cut -d " " -f3 OAI_EUR_00_hg19.fam | sort | uniq -d
0
cut -d " " -f6 OAI_EUR_00_hg19.fam | sort | uniq -d
-9
```

Fam file updating

- Merged to additional OAI data to retrieve sex, race, and a phenotype for a GWAS, height (mm)
- Study design with unrelated participants.
- All participants in this GWAS file are self-reported white race

```
head -n5 OAI_EUR_01_hg19.fam
9000099 9000099 0 0 1 1812.5
9000296 9000296 0 0 1 -9
9000622 9000622 0 0 2 -9
9000798 9000798 0 0 1 1793
9001104 9001104 0 0 2 -9
```

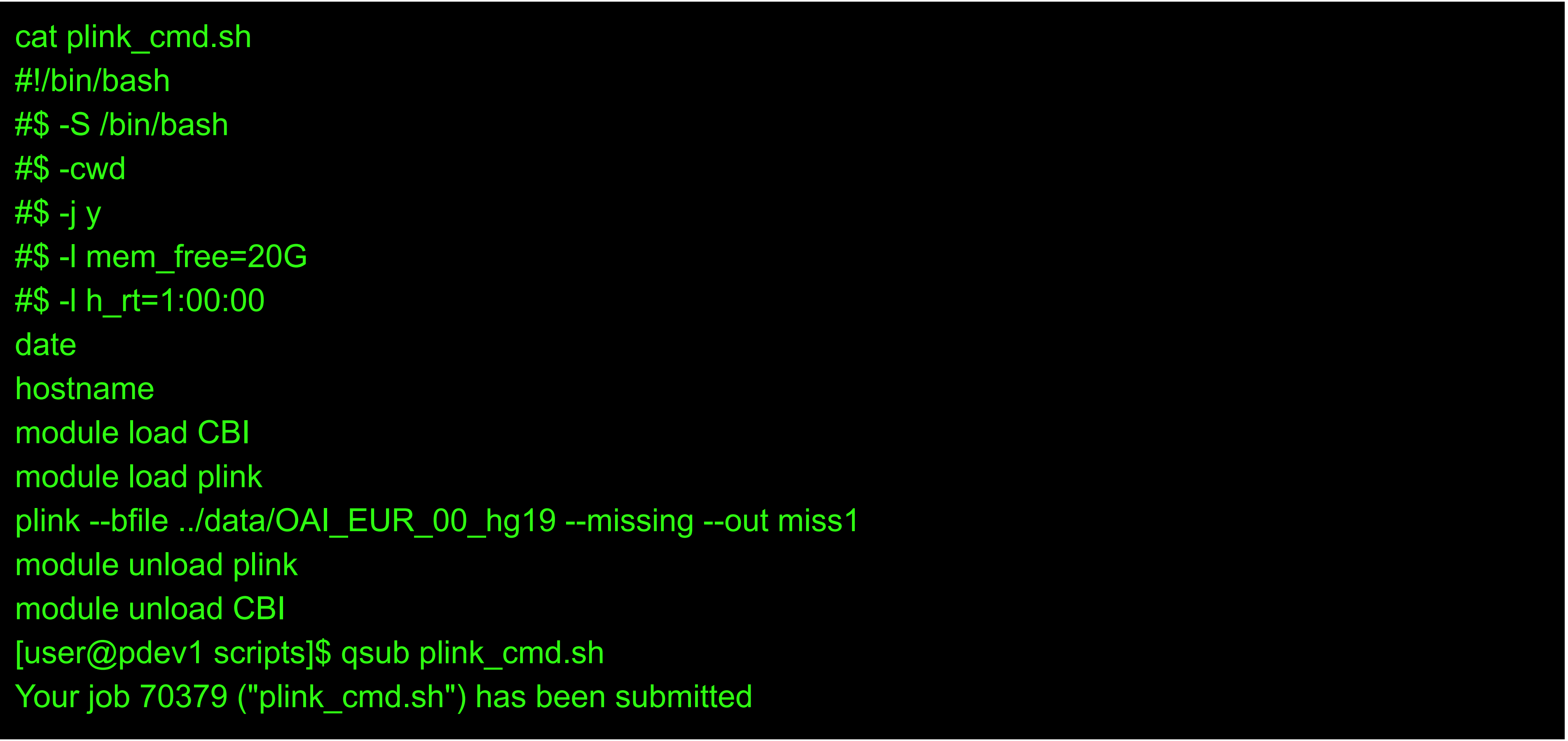
plink

```
module load CBI
```

```
module avail
```

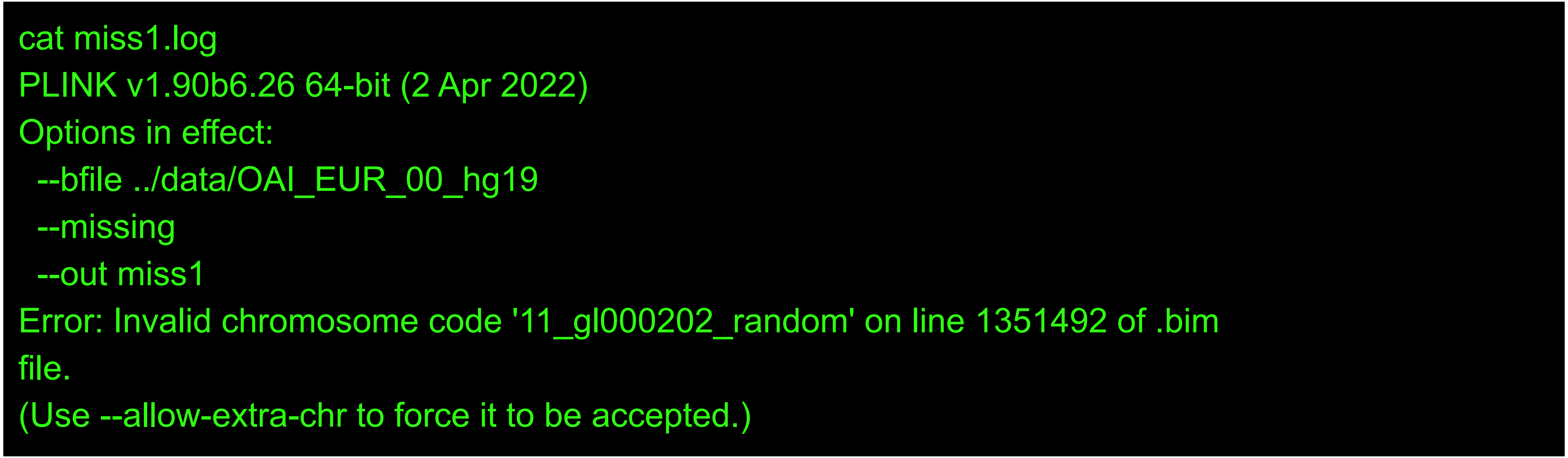
```
bedops/2.4.41          (D)    plink/1.07
   bedtools2/2.26.0              plink/1.90b6.10
   bedtools2/2.28.0              plink/1.90b6.16
   bedtools2/2.29.1              plink/1.90b6.18
   bedtools2/2.29.2              plink/1.90b6.21
   bedtools2/2.30.0              plink/1.90b6.24
   bedtools2/2.31.1       (D)    plink/1.90b6.25
   blast/2.9.0                plink/1.90b6.26             (D)
   blast/2.10.1               plink2/2.00a2LM
   blast/2.11.0               plink2/2.00a3LM             (D)
```

Plink missingness



```
cat plink_cmd.sh
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
#$ -j y
#$ -l mem_free=20G
#$ -l h_rt=1:00:00
date
hostname
module load CBI
module load plink
plink --bfile ../data/OAI_EUR_00_hg19 --missing --out miss1
module unload plink
module unload CBI
[user@pdev1 scripts]$ qsub plink_cmd.sh
Your job 70379 ("plink_cmd.sh") has been submitted
```

Strange chromosome causing error



```
cat miss1.log
PLINK v1.90b6.26 64-bit (2 Apr 2022)
Options in effect:
  --bfile ../data/OAI_EUR_00_hg19
  --missing
  --out miss1
Error: Invalid chromosome code '11_gl000202_random' on line 1351492 of .bim
file.
(Use --allow-extra-chr to force it to be accepted.)
```

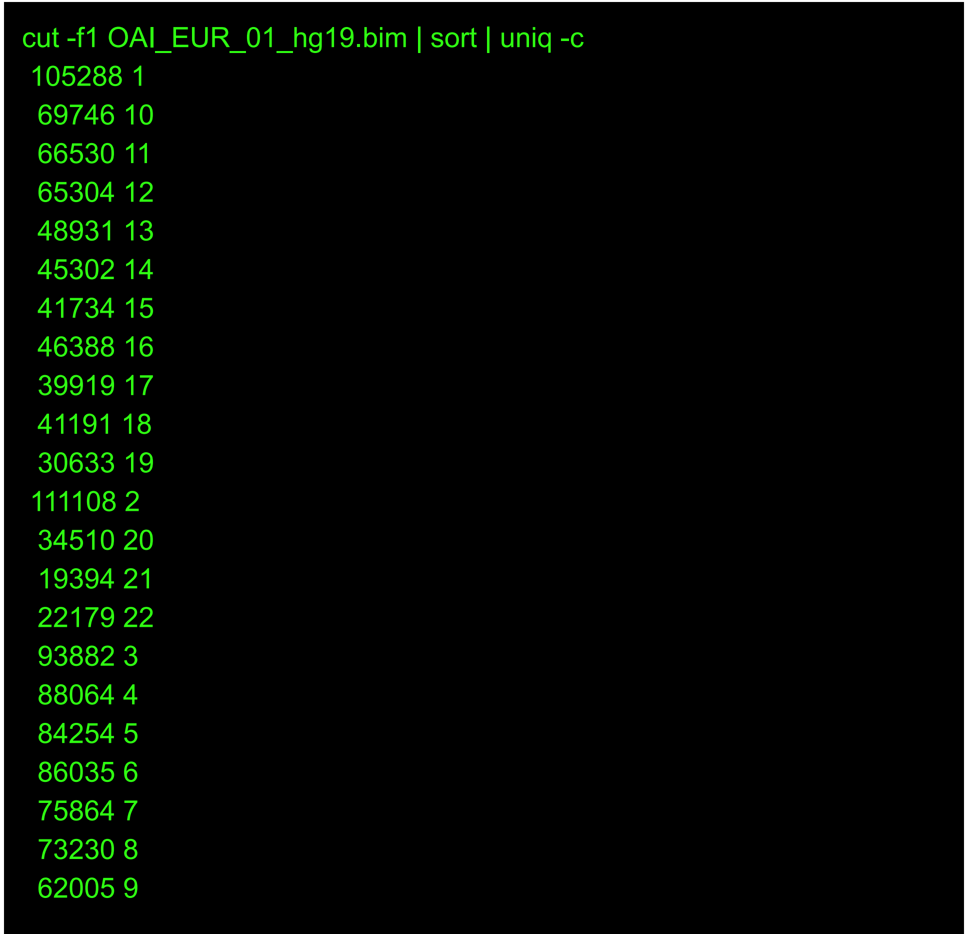
Remove single variant on “11_random” chr

- Submitted as job

```
plink --bfile ../data/OAI_EUR_00_hg19 --allow-extra-chr
--not-chr 11_gl000202_random --make-bed --out
../data/OAI_EUR_01_hg19
```

```
qstat -u $user
job-ID  prior   name       user         state submit/start at
queue                          slots ja-task-ID
----------------------------------------------------------------------
-------------------------------------------
  70396 0.00000 plink_cmd. $user       qw    02/11/2024 00:06:02
1
```

11_random is removed



```
cut -f1 OAI_EUR_01_hg19.bim | sort | uniq -c
 105288 1
  69746 10
  66530 11
  65304 12
  48931 13
  45302 14
  41734 15
  46388 16
  39919 17
  41191 18
  30633 19
 111108 2
  34510 20
  19394 21
  22179 22
  93882 3
  88064 4
  84254 5
  86035 6
  75864 7
  73230 8
  62005 9
```

Missingness, allele frequency, HWE analysis

- Missing. Failed genotype calls.
  - Remove SNPs with missing rate > 0.05
  - Remove samples with missing rate > 0.05
- Allele frequency
  - SNPs with minor allele frequency (MAF) much less that 1% have more technical variability from the genotyping array, less power for statistical association tests
  - Typically can remove SNPs with MAF < 0.01 or 0.001. Judgement call.
- Hardy-Weinberg Equilibrium (HWE). Relates allele to genotype frequency. (1) random mating (i.e, low population structure), (2) no natural selection, (3) a very large population size, (4) no gene flow or migration, (5) no mutation, and (6) the locus is autosomal. HWE violations can indicate genotyping errors. Remove SNPs with HWE P-value < 10-6

Missingness, allele frequency, HWE analysis

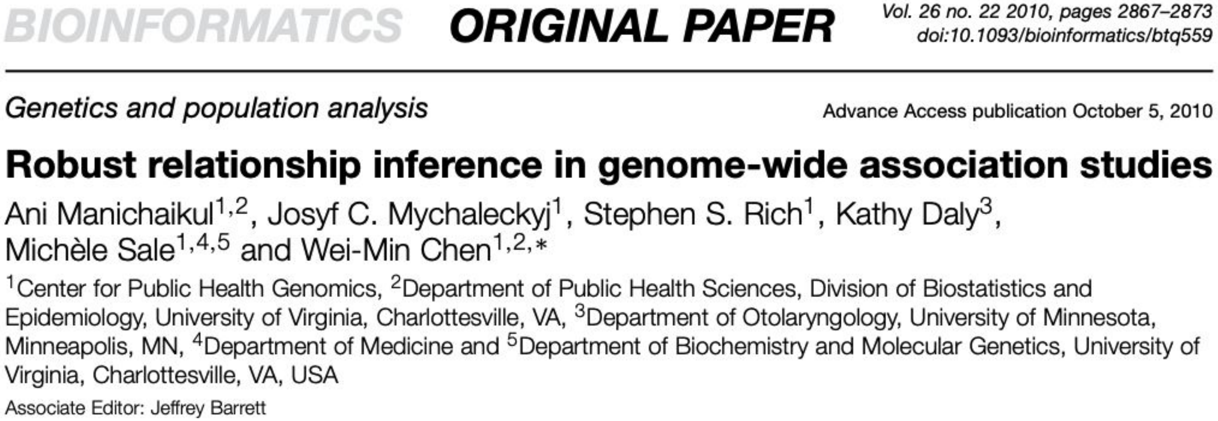
```
cat plink_cmd.sh
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
#$ -j y
#$ -l mem_free=20G
#$ -l h_rt=1:00:00
date
hostname
module load CBI
module load plink
plink --bfile ../data/OAI_EUR_01_hg19 --missing --out geno01 \
     --freq \
     --hardy
module unload plink
module unload CBI
```

Launch Rstudio in Wynton

- https://wynton.ucsf.edu/hpc/howto/rstudio.html
- ssh -Y -C plog1.wynton.ucsf.edu
- ssh -X pdev1
- module load CBI rstudio
- rstudio
- As long as you have xterm installed on mac, rstudio starts up in an Xterm window!
- Missing, MAF, and HWE were cleaned

Genetic relatedness

- OAI participants are not supposed to be related, but we can check with genetics
- Association testing assumes observations are independent
- Genetic ancestry can be skewed by related samples
- Unfortunately, some genetic relatedness algorithms assume no population structure. Best to use relatedness inference algorithms robust to stratification.



BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 22 2010, pages 2867–2873
doi:10.1093/bioinformatics/btq559

Genetics and population analysis

Advance Access publication October 5, 2010

Robust relationship inference in genome-wide association studies

Ani Manichaikul[1,2], Josyf C. Mychaleckyj[1], Stephen S. Rich[1], Kathy Daly[3], Michèle Sale[1,4,5] and Wei-Min Chen[1,2,*]

[1]Center for Public Health Genomics, [2]Department of Public Health Sciences, Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, VA, [3]Department of Otolaryngology, University of Minnesota, Minneapolis, MN, [4]Department of Medicine and [5]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA, USA

Associate Editor: Jeffrey Barrett

King-robust

```
cat plink_cmd_fam.sh
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
#$ -j y
#$ -l mem_free=20G
#$ -l h_rt=3:00:00
date
hostname
module load CBI
module load plink2
plink2 --bfile ../data/OAI_EUR_01_hg19 \
     --make-king \
     --make-king-table \
     --out geno01
module unload plink2
module unload CBI
```

No 2nd degree or higher relationships

```
dat <- read_delim("geno01.kin0" )
dat %>%
  filter(KINSHIP >= 0.125) %>%
  summarize(n = n())
```

```
# A tibble: 1 × 1
      n
  <int>
1     0
```

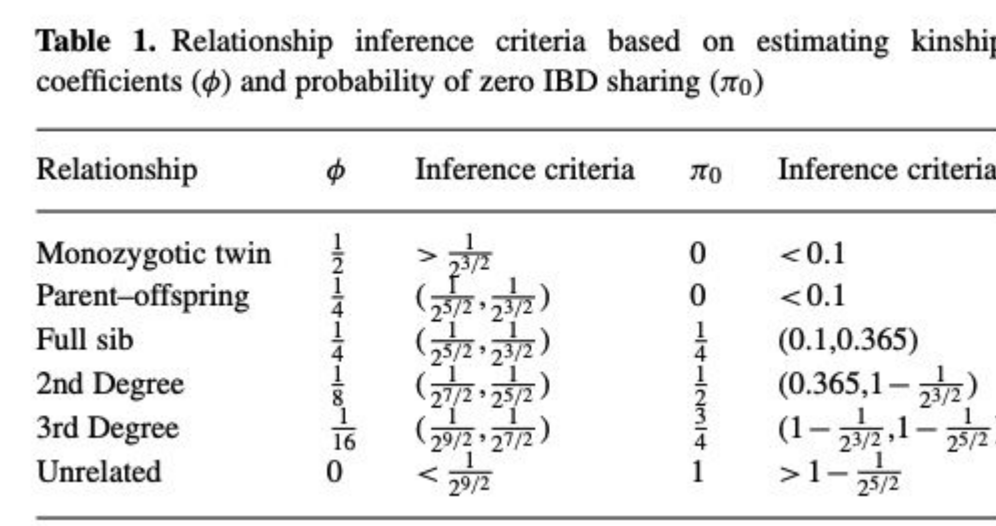


Table 1. Relationship inference criteria based on estimating kinshi
coefficients (ϕ) and probability of zero IBD sharing (π_0)

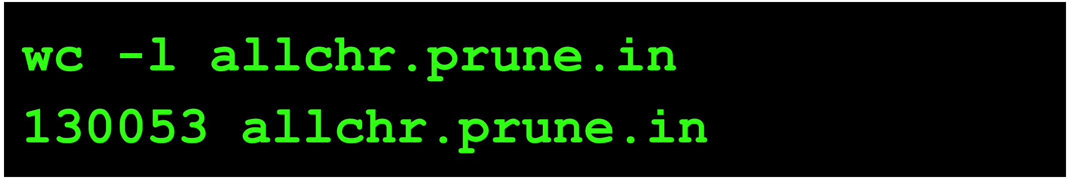
Relationship	ϕ	Inference criteria	π_0	Inference criteria
Monozygotic twin	$\frac{1}{2}$	$> \frac{1}{2^{3/2}}$	0	< 0.1
Parent–offspring	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	0	< 0.1
Full sib	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	$\frac{1}{4}$	$(0.1, 0.365)$
2nd Degree	$\frac{1}{8}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$	$\frac{1}{2}$	$(0.365, 1 - \frac{1}{2^{3/2}})$
3rd Degree	$\frac{1}{16}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}})$	$\frac{3}{4}$	$(1 - \frac{1}{2^{3/2}}, 1 - \frac{1}{2^{5/2}}$
Unrelated	0	$< \frac{1}{2^{9/2}}$	1	$> 1 - \frac{1}{2^{5/2}}$

Ancestry estimation

- PCA on the genetic relationship matrix
- Sensitive to LD. Regions with high LD can have a larger impact on ancestry estimates.
- LD prune - remove SNPs in LD over a threshold in a region

```
cat plink_cmd_ld.sh
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
#$ -j y
#$ -l mem_free=20G
#$ -l h_rt=2:00:00
#$ -t 1-22
date
hostname
echo $SGE_TASK_ID
module load CBI
module load plink
plink --bfile ../data/OAI_EUR_01_hg19 \
     --snps-only \
     --chr $SGE_TASK_ID \
     --maf 0.05 \
     --indep-pairwise 10000 50 0.2 \
     --out prune/LDchr$SGE_TASK_ID
module unload plink
module unload CBI
```

Concatenate chromosome-specific pruned SNPs

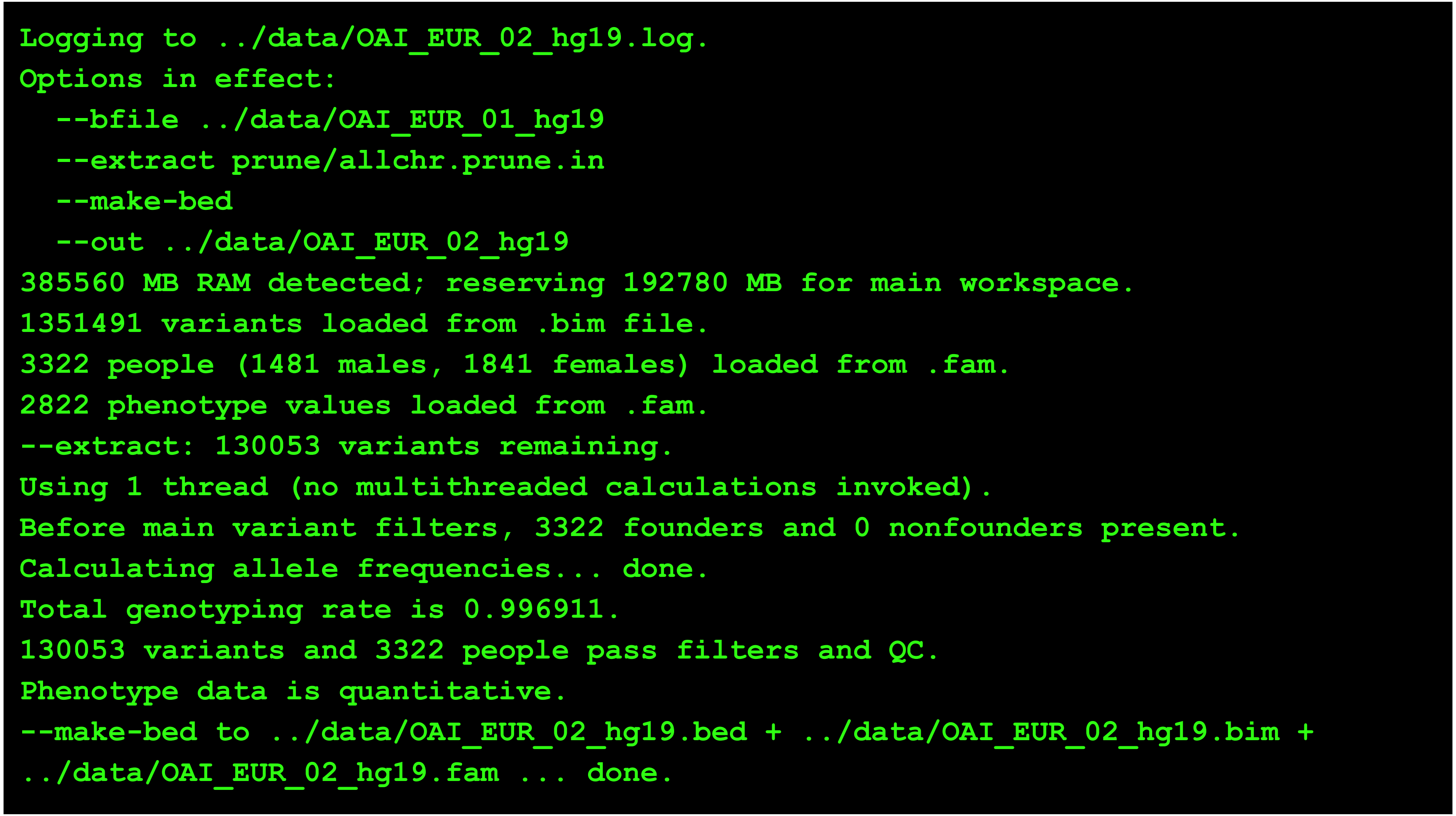
- cat *.prune.in > allchr.prune.in
- 

```
wc -l allchr.prune.in
130053 allchr.prune.in
```

Create new LD-pruned dataset

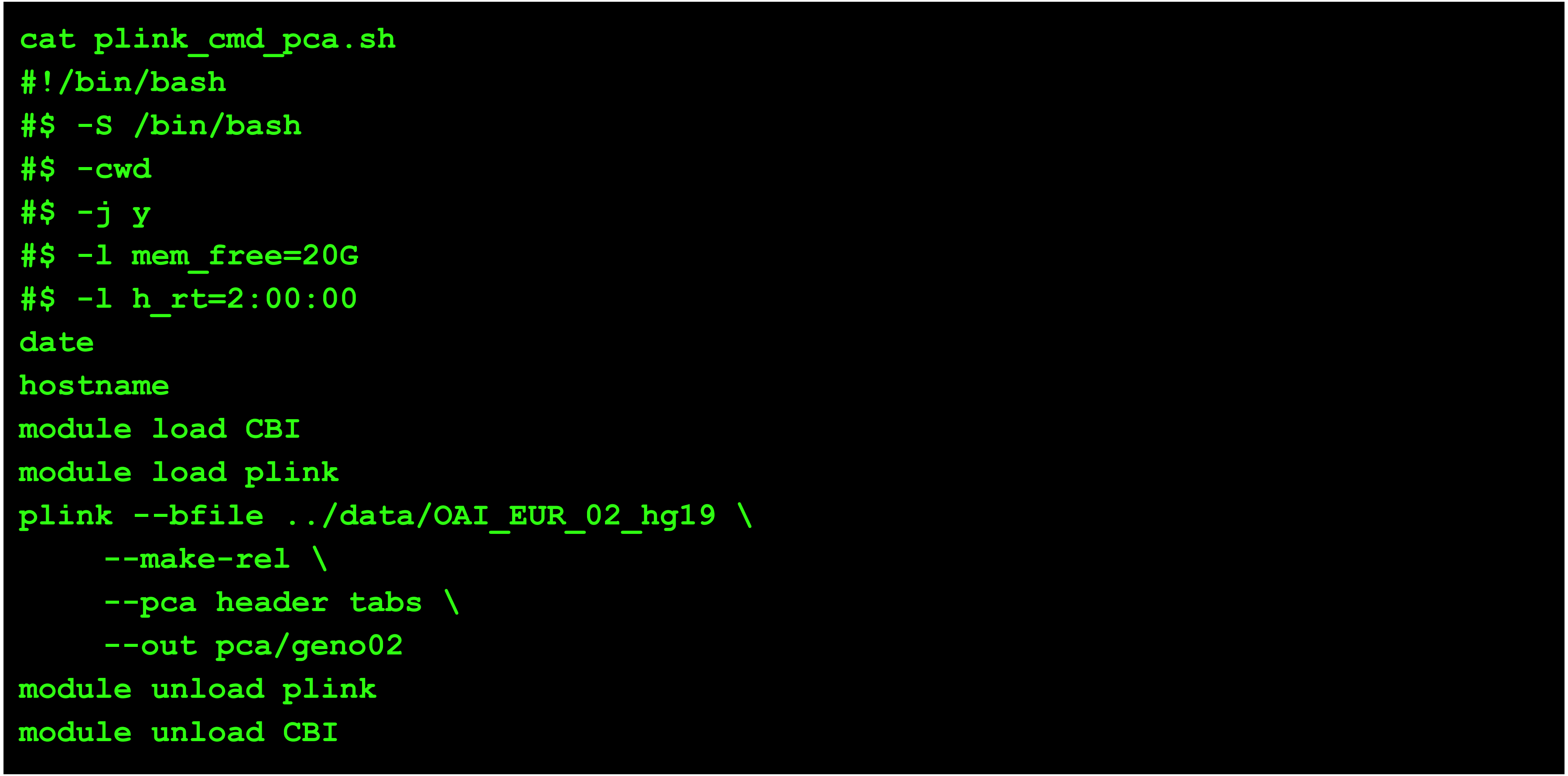
```
cat plink_cmd_prune.sh
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
#$ -j y
#$ -l mem_free=20G
#$ -l h_rt=1:00:00
date
hostname
module load CBI
module load plink
plink --bfile ../data/OAI_EUR_01_hg19 \
     --extract prune/allchr.prune.in \
     --make-bed \
     --out ../data/OAI_EUR_02_hg19
module unload plink
module unload CBI
```

Check sample size and variant count



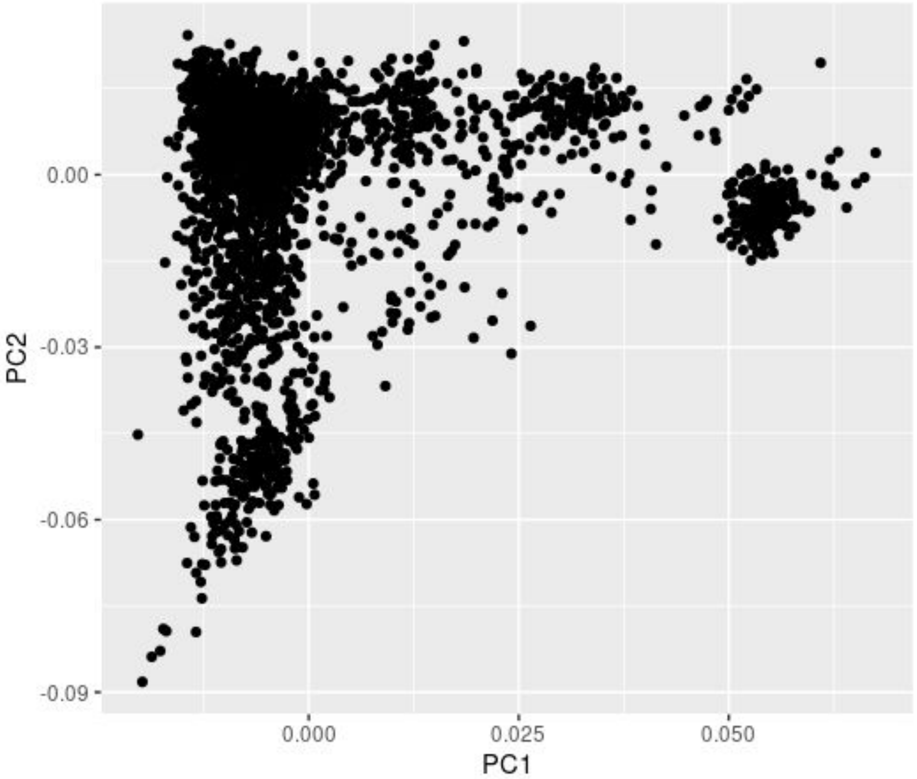
```
Logging to ../data/OAI_EUR_02_hg19.log.
Options in effect:
  --bfile ../data/OAI_EUR_01_hg19
  --extract prune/allchr.prune.in
  --make-bed
  --out ../data/OAI_EUR_02_hg19
385560 MB RAM detected; reserving 192780 MB for main workspace.
1351491 variants loaded from .bim file.
3322 people (1481 males, 1841 females) loaded from .fam.
2822 phenotype values loaded from .fam.
--extract: 130053 variants remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 3322 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.996911.
130053 variants and 3322 people pass filters and QC.
Phenotype data is quantitative.
--make-bed to ../data/OAI_EUR_02_hg19.bed + ../data/OAI_EUR_02_hg19.bim +
../data/OAI_EUR_02_hg19.fam ... done.
```

PCA on LD-pruned data



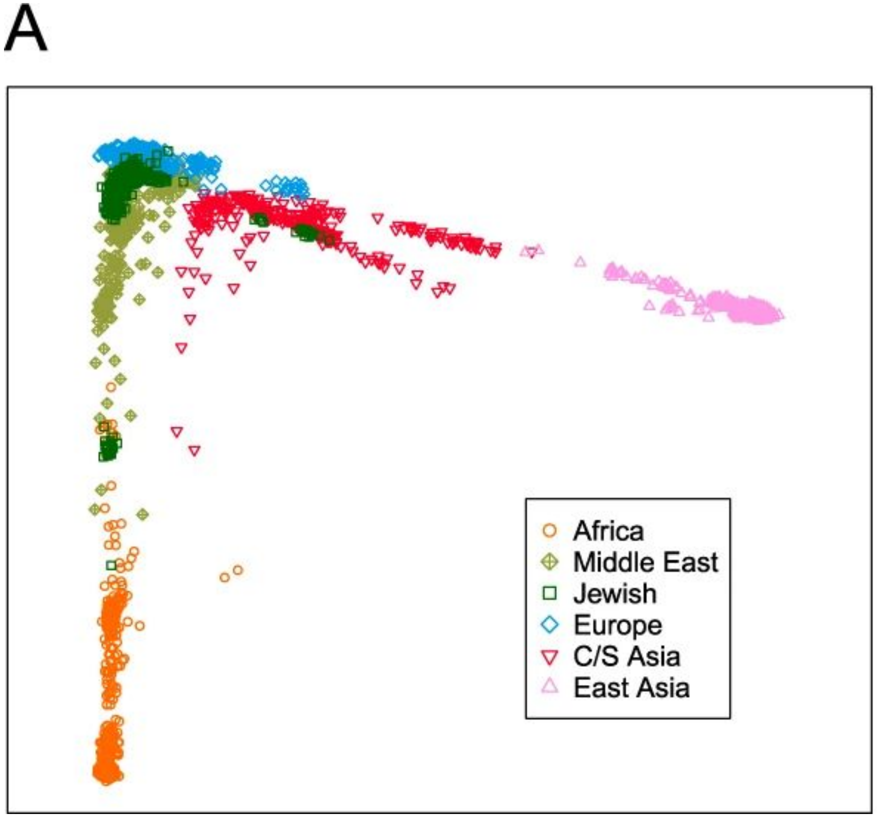
```
cat plink_cmd_pca.sh
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
#$ -j y
#$ -l mem_free=20G
#$ -l h_rt=2:00:00
date
hostname
module load CBI
module load plink
plink --bfile ../data/OAI_EUR_02_hg19 \
     --make-rel \
     --pca header tabs \
     --out pca/geno02
module unload plink
module unload CBI
```

PCA self-identified white participants



- PCA with reference populations. Are apparent clusters more similar to reference EUR samples than other samples?
- Potentially exclude ancestry outliers
- Adjust for PCs in GWAS

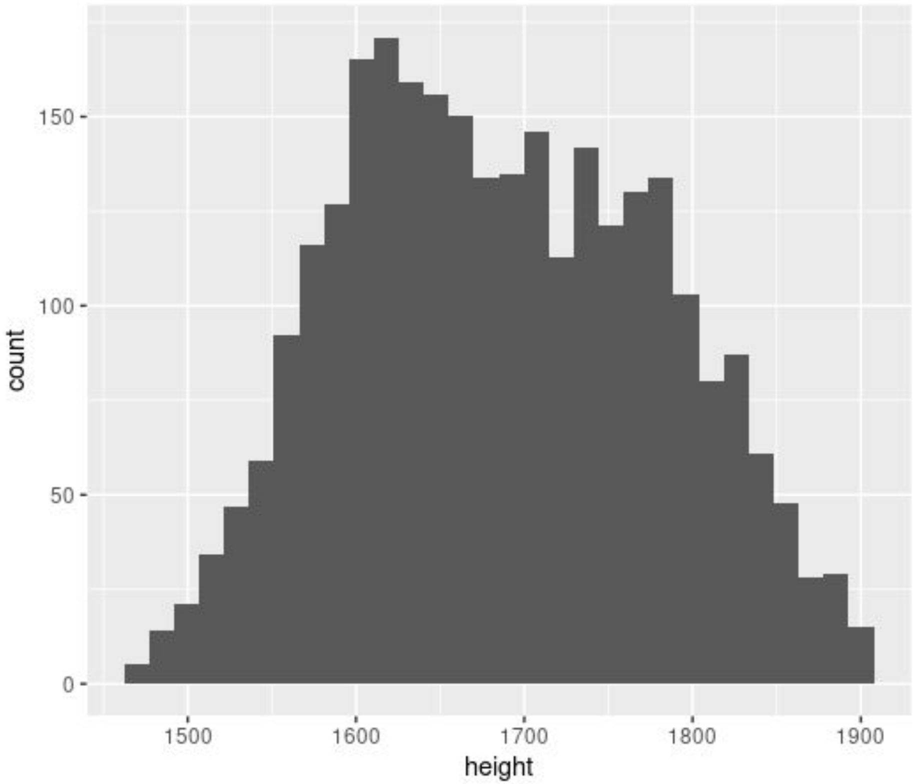
PCA with reference populations



Kopelman, N.M., Stone, L., Hernandez, D.G. *et al.* High-resolution inference of genetic relationships among Jewish populations. *Eur J Hum Genet* 28, 804–814 (2020). https://doi.org/10.1038/s41431-019-0542-y

Fig 1. MDS of jewish samples with ancestry references

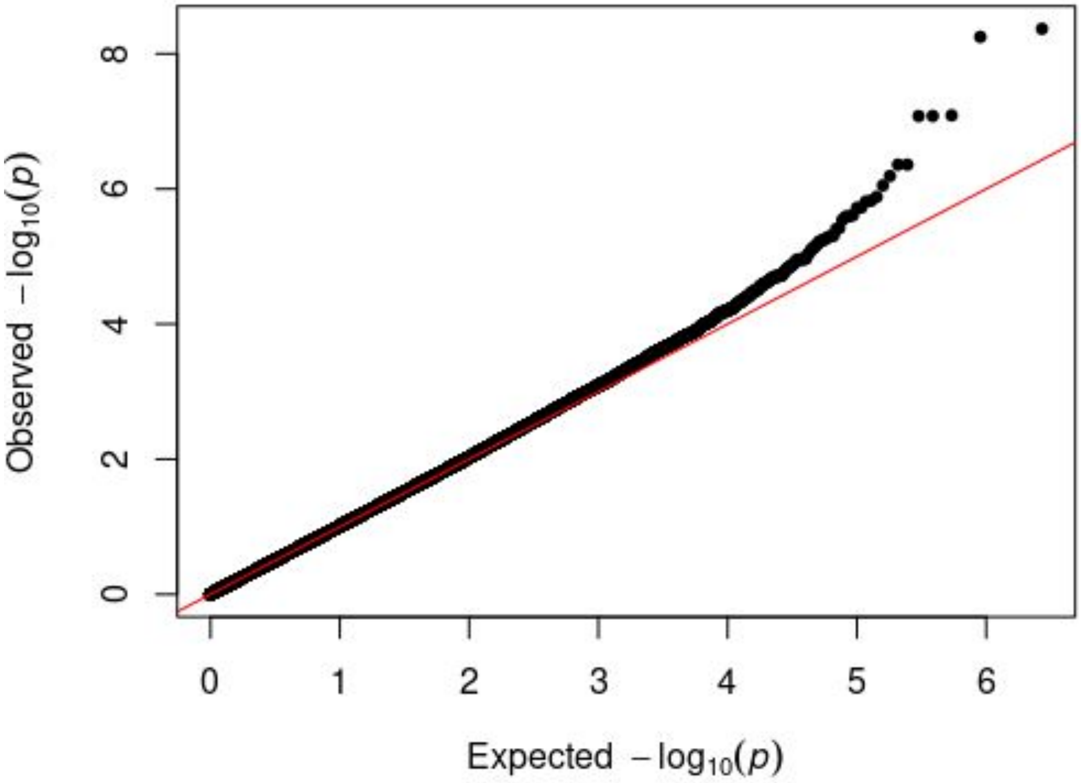
GWAS of height



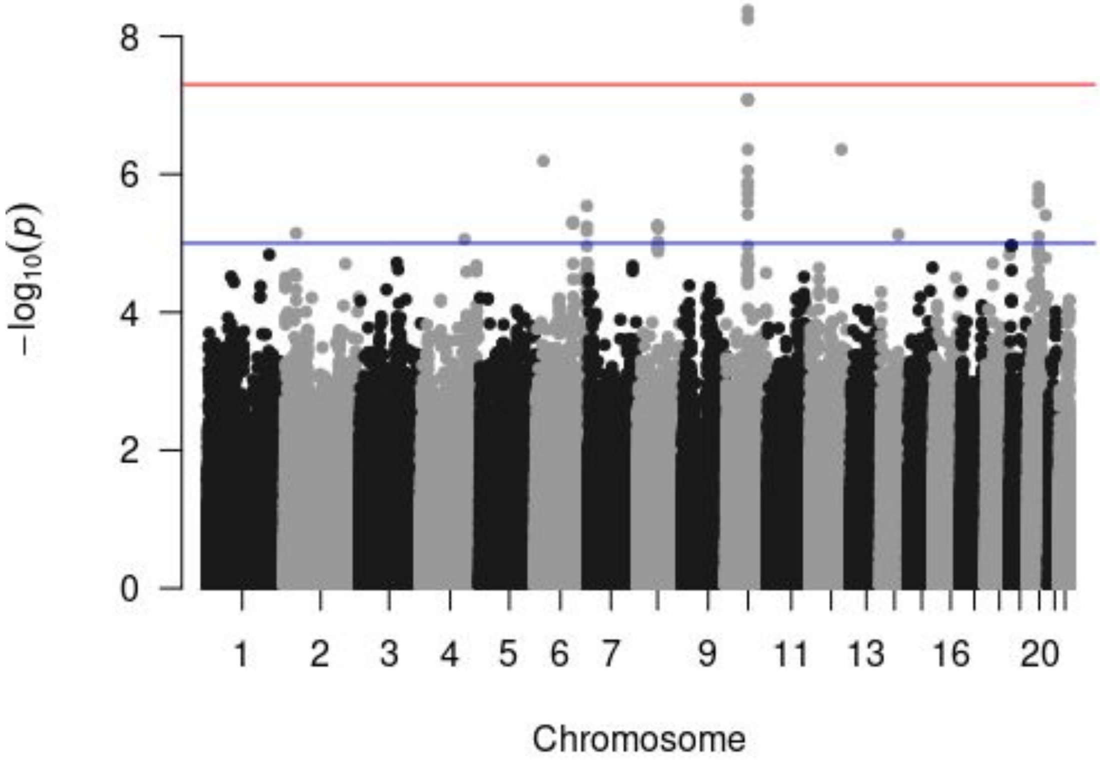
GWAS of height, adjusted for sex and PC1-PC3

```
cat plink_cmd_assoc.sh
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
#$ -j y
#$ -l mem_free=20G
#$ -l h_rt=1:00:00
date
hostname
module load CBI
module load plink
plink --bfile ../data/OAI_EUR_01_hg19 \
    --covar pca/geno02.eigenvec \
    --covar-name PC1,PC2,PC3 \
    --linear hide-covar sex \
    --out assoc/height
module unload plink
module unload CBI
```

QQ plot, qqman R package



Manhattan,
qqman R package



Next steps

- Study design, epidemiology
- Apply appropriate association test for your data
- Genotype imputation
  - HRC = 40 million variants
  - TOPMed = 300 million variants
- Association analysis of genotype probabilities
- Multiple testing
- Meta-analysis
- Power calculations
- Post-GWAS annotation and result interpretation

Interested in learning more?

Daniel [dot] Evans [at] ucsf [dot] edu

Genetics and genomics of aging

Serra [dot] kaya [at] ucsf [dot] edu and I are performing GWAS with OAI.

Illumina annotation file

- Lookup the genotype array here: https://www.strand.org.uk/
- OAI used Illumina HumanOmni2.5-4v1_B
  - https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000955.v1.p1
-