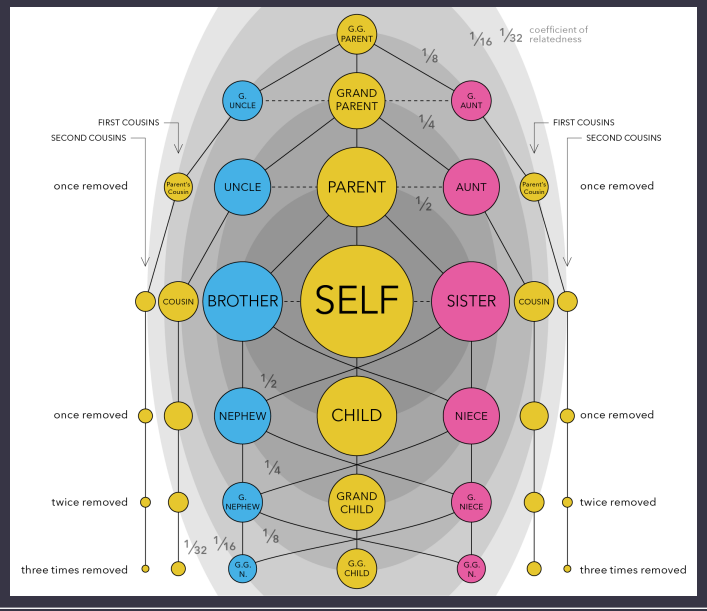
(superfast)

Overview of PLINK and Genetic Relatedness

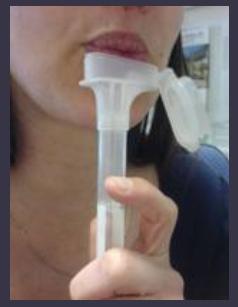


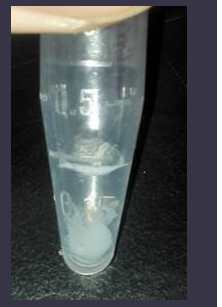
Sarah Medland and José Morosoli

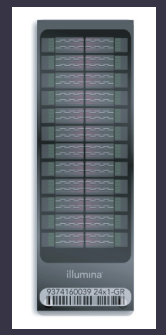
Today

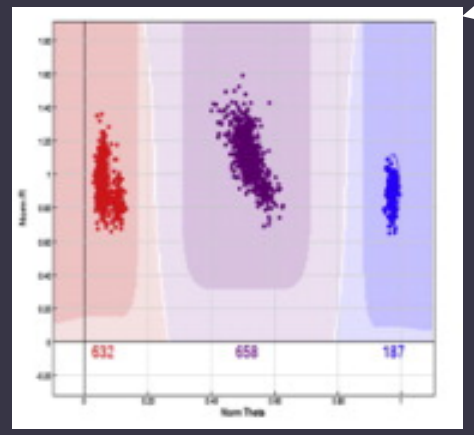
- Quick play with some QC functions in PLINK
- Calculate genetic relatedness

Common variant genotyping





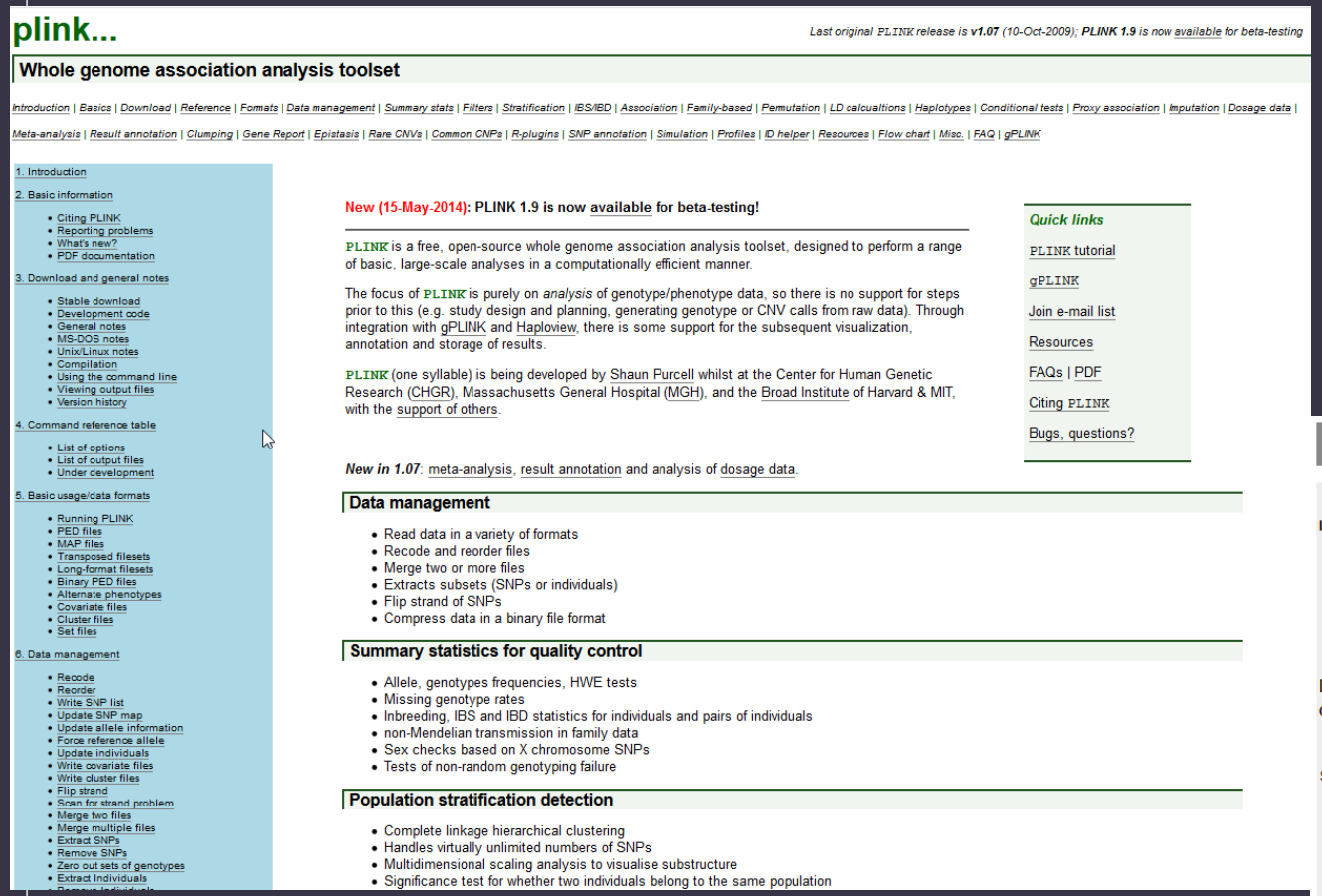




C	T	G	G	A	G	A	G	T	T	C	C	A	A	G	G	A	A	T	T	C	T	C	C
C	T	G	G	A	G	A	G	0	0	C	C	G	G	G	G	A	C	C	T	C	C	C	C
T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	C	C	T	C	C	C	C
T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	A	T	T	C	C	C	C
T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	T	A	A	T	T	T	T	C	C
T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	G	A	A	T	T	C	T	C	C
C	C	G	G	0	0	0	0	T	T	C	C	A	A	G	T	A	A	T	T	C	T	C	C
T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	A	C	C	C	C	C	C
T	T	G	G	A	A	G	G	0	0	C	C	A	A	G	G	A	C	C	T	C	C	C	C
T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	G	A	A	C	T	C	T	C	C
C	T	G	G	A	G	0	0	T	T	C	C	G	G	T	T	A	A	T	T	T	T	C	C
T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	T	A	A	T	T	C	C	C	C
T	T	G	G	A	A	G	G	T	T	C	C	G	G	G	G	A	C	T	T	C	C	C	C
C	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	A	T	T	C	C	C	C

PLINK…

- Designed to be a one stop shop for GWAS data handling and analysis
  - Limitations: families or dosage data



plink…

*Last original PLINK release is **v1.07** (10-Oct-2009); **PLINK 1.9** is now available for beta-testing*

Whole genome association analysis toolset

Introduction | Basics | Download | Reference | Formats | Data management | Summary stats | Filters | Stratification | IBS/IBD | Association | Family-based | Permutation | LD calcualtions | Haplotypes | Conditional tests | Proxy association | Imputation | Dosage data | Meta-analysis | Result annotation | Clumping | Gene Report | Epistasis | Rare CNVs | Common CNPs | R-plugins | SNP annotation | Simulation | Profiles | ID helper | Resources | Flow chart | Misc. | FAQ | gPLINK

1. Introduction
2. Basic information
   - Citing PLINK
   - Reporting problems
   - What's new?
   - PDF documentation
3. Download and general notes
   - Stable download
   - Development code
   - General notes
   - MS-DOS notes
   - Unix/Linux notes
   - Compilation
   - Using the command line
   - Viewing output files
   - Version history
4. Command reference table
   - List of options
   - List of output files
   - Under development
5. Basic usage/data formats
   - Running PLINK
   - PED files
   - MAP files
   - Transposed filesets
   - Long-format filesets
   - Binary PED files
   - Alternate phenotypes
   - Covariate files
   - Cluster files
   - Set files
6. Data management
   - Recode
   - Reorder
   - Write SNP list
   - Update SNP map
   - Update allele information
   - Force reference allele
   - Update individuals
   - Write covariate files
   - Write cluster files
   - Flip strand
   - Scan for strand problem
   - Merge two files
   - Merge multiple files
   - Extract SNPs
   - Remove SNPs
   - Zero out sets of genotypes
   - Extract Individuals

New (15-May-2014): PLINK 1.9 is now available for beta-testing!

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

The focus of PLINK is purely on *analysis* of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). Through integration with gPLINK and Haploview, there is some support for the subsequent visualization, annotation and storage of results.

PLINK (one syllable) is being developed by Shaun Purcell whilst at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT, with the support of others.

Quick links

PLINK tutorial
gPLINK
Join e-mail list
Resources
FAQs | PDF
Citing PLINK
Bugs, questions?

New in 1.07: meta-analysis, result annotation and analysis of dosage data.

Data management

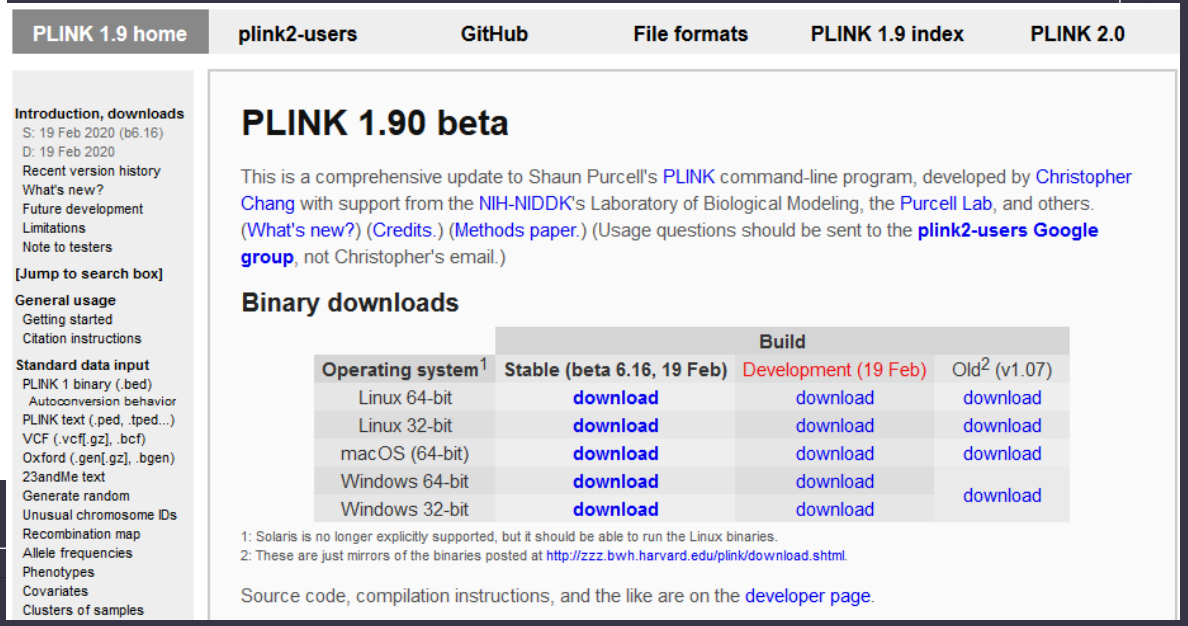
- Read data in a variety of formats
- Recode and reorder files
- Merge two or more files
- Extracts subsets (SNPs or individuals)
- Flip strand of SNPs
- Compress data in a binary file format

Summary statistics for quality control

- Allele, genotypes frequencies, HWE tests
- Missing genotype rates
- Inbreeding, IBS and IBD statistics for individuals and pairs of individuals
- non-Mendelian transmission in family data
- Sex checks based on X chromosome SNPs
- Tests of non-random genotyping failure

Population stratification detection

- Complete linkage hierarchical clustering
- Handles virtually unlimited numbers of SNPs
- Multidimensional scaling analysis to visualise substructure
- Significance test for whether two individuals belong to the same population



PLINK 1.9 home | plink2-users | GitHub | File formats | PLINK 1.9 index | PLINK 2.0

Introduction, downloads
S: 19 Feb 2020 (b6.16)
D: 19 Feb 2020
Recent version history
What's new?
Future development
Limitations
Note to testers
[Jump to search box]
General usage
Getting started
Citation instructions
Standard data input
PLINK 1 binary (.bed)
Autoconversion behavior
PLINK text (.ped, .tped...)
VCF (.vcf[.gz], .bcf)
Oxford (.gen[.gz], .bgen)
23andMe text
Generate random
Unusual chromosome IDs
Recombination map
Allele frequencies
Phenotypes
Covariates
Clusters of samples

PLINK 1.90 beta

This is a comprehensive update to Shaun Purcell's PLINK command-line program, developed by Christopher Chang with support from the NIH-NIDDK's Laboratory of Biological Modeling, the Purcell Lab, and others. (What's new?) (Credits.) (Methods paper.) (Usage questions should be sent to the **plink2-users Google group**, not Christopher's email.)

Binary downloads

	Build		
Operating system[1]	Stable (beta 6.16, 19 Feb)	Development (19 Feb)	Old[2] (v1.07)
Linux 64-bit	download	download	download
Linux 32-bit	download	download	download
macOS (64-bit)	download	download	download
Windows 64-bit	download	download	download
Windows 32-bit	download	download	

1: Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.
2: These are just mirrors of the binaries posted at http://zzz.bwh.harvard.edu/plink/download.shtml.

Source code, compilation instructions, and the like are on the developer page.

Plink…

- Brilliant well explained online manual

Missing genotypes

To generate a list genotyping/missingness rate statistics:

```
plink --file data --missing
```

This option creates two files:

```
plink.imiss
plink.lmiss
```

which detail missingness by individual and by SNP (locus), respectively. For individuals, the format is:

```
FID             Family ID
IID             Individual ID
MISS_PHENO      Missing phenotype? (Y/N)
N_MISS          Number of missing SNPs
N_GENO          Number of non-obligatory missing genotypes
F_MISS          Proportion of missing SNPs
```

For each SNP, the format is:

```
SNP             SNP identifier
CHR             Chromosome number
N_MISS          Number of individuals missing this SNP
N_GENO          Number of non-obligatory missing genotypes
F_MISS          Proportion of sample missing for this SNP
```

HINT To test for case/control differences in missingness, see the `--test-missing` option.

HINT To produce summary of missingness that is stratified by a categorical cluster variable, use the `--within` *filename* option as well as `--missing`. In this way, the missing rates will be given separately for each level of the categorical variable. For example, the categorical variable could be which plate that sample was on in the genotyping. Details on the format of a cluster file can be found here.

Obligatory missing genotypes

Often genotypes might be missing obligatorarily rather than because of genotyping failure. For example, some proportion of the sample might only have been genotyped on a subset of the SNPs. In these cases, one might not want to filter out SNPs and individuals based on this type of missing data. Alternatively, genotypes for specific plates (sets of SNPs/individuals) might have been blanked out with the `--zero-cluster` option, but you still might want to be able to sensibly set missing data thresholds.

Calculating genetic relatedness

- Estimate relatedness from genomic data
  - Typically common variant GWAS data
  - Doesn't need to be very dense genotyping but you do want whole genome coverage
- Two main approaches
  - Identity by state
  - Identity by descent

Calculating genetic relatedness

- Identity by state
  - Do individuals share the same genetic variants at a given locus?
- Identity by descent
  - Do individuals share the same genetic variants at a given locus AND was it inherited from the same ancestor?

Calculating genetic relatedness

- Identity by state
  - Does not require pedigree information or data from other relatives
- Identity by descent
  - Does

Calculating genetic relatedness

- Identity by state
  - Information is limited to the observed locus (and variants that are very, very close by)
- Identity by descent
  - Provides more information about surrounding loci

Calculating genetic relatedness

- Identity by state
  - Produces a coefficient of relatedness known as ***r***
  - Implementation in GRMs is typically calculated as the average correlation between individuals across genotyped snps

$$A_{ij} = \frac{1}{m}\sum_{k}^{m}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)} \quad (1)$$

where m is the number of SNPs, x_{jk} is the genotype (coded as 0, 1, or 2) of individual j at the kth locus, and p_k is the minor allele frequency (MAF) of the kth locus. The

Calculating genetic relatedness

- Identity by descent
  - Estimation typically requires Monte Carlo Markov Chain methods
  - Estimate is known as $\hat{\pi}$ (pi-hat)

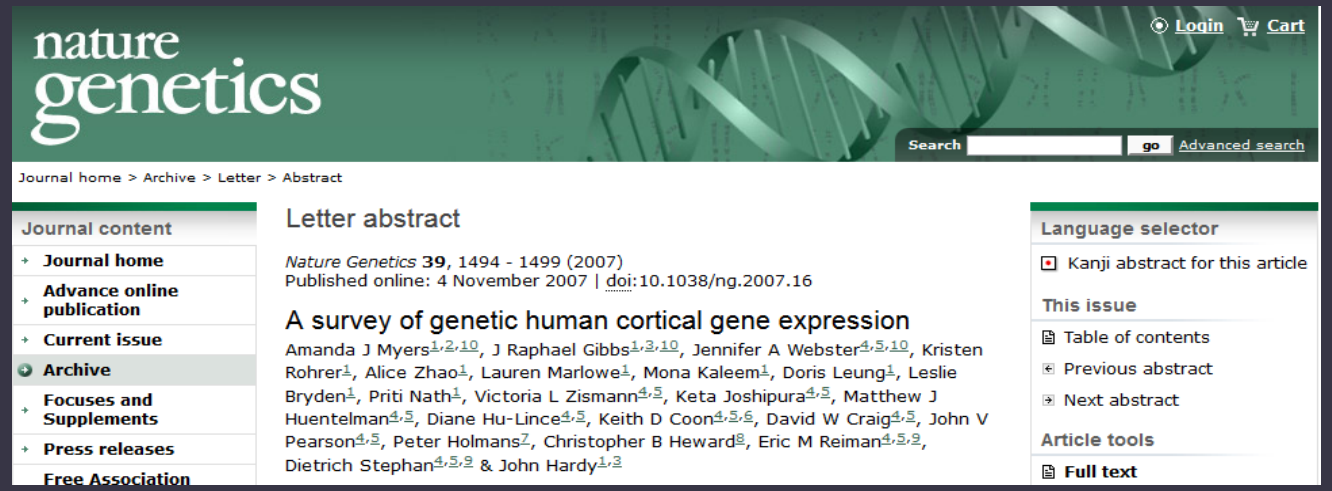
Plink and GRM practical

In this practical, we will:

- Learn about genetic relatedness / relationship matrices by estimating one.
- Learn about some concepts and methods from molecular genetic methods.

Our data

- Simulated dataset based on real data.



- 23 MZ pairs, 21 DZ pairs, 62 unrelated individuals

1	WGACON 1	0	0	1	1	C	T	G	G	A	G	A	G	T	T	C	C	A	A	G	G	A	A	T	T	C	T	C	C	C	C	G	G	G	T	A	G
2	WGACON 2	0	0	1	1	C	T	G	G	A	G	A	G	0	0	C	C	G	G	G	G	A	C	C	T	C	C	C	C	C	C	G	G	G	T	A	G
3	WGACON 3	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	C	C	T	C	C	C	C	C	T	G	G	G	T	A	G
4	WGACON 4	0	0	2	1	T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	A	T	T	C	C	C	C	C	C	G	G	0	0	A	G
5	WGACON 5	0	0	2	1	T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	T	A	A	T	T	T	T	C	C	C	C	G	T	G	T	A	A
6	WGACON 6	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	G	A	A	T	T	C	T	C	C	C	C	G	G	G	T	A	A
7	WGACON 7	0	0	1	1	C	C	G	G	0	0	0	0	T	T	C	C	A	A	G	T	A	A	T	T	C	T	C	C	C	C	G	G	G	T	G	G
8	WGACON 8	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	A	C	C	C	C	C	C	C	C	G	G	G	T	A	A
9	WGACON 9	0	0	1	1	T	T	G	G	A	A	G	G	0	0	C	C	A	A	G	G	A	C	C	T	C	C	C	C	C	C	G	G	0	0	A	G
10	WGACON 10	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	G	A	A	C	T	C	T	C	C	C	C	G	G	G	T	A	G
11	WGACON 11	0	0	2	1	C	T	G	G	A	G	0	0	T	T	C	C	G	G	T	T	A	A	T	T	T	T	C	C	C	C	G	G	G	T	A	A
12	WGACON 12	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	T	A	A	T	T	C	C	C	C	C	C	G	G	G	T	G	G
13	WGACON 13	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	G	G	G	G	A	C	T	T	C	C	C	C	C	C	G	G	0	0	A	A
14	WGACON 14	0	0	2	1	C	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	A	T	T	C	C	C	C	C	C	G	G	G	T	A	A
15	WGACON 15	0	0	1	1	C	T	G	G	A	A	A	G	C	T	C	C	A	G	G	G	A	C	C	T	C	C	C	C	C	C	G	G	G	T	G	G
16	WGACON 16	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	T	A	A	C	C	C	C	C	C	C	C	G	T	G	T	A	A
17	WGACON 17	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	A	T	T	C	C	C	C	C	C	G	G	G	T	A	A
18	WGACON 18	0	0	2	1	T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	G	A	C	T	T	C	C	C	C	C	C	G	G	G	T	A	A
19	WGACON 19	0	0	2	1	C	T	G	G	A	G	A	G	0	0	C	C	A	A	G	G	A	A	C	T	C	C	C	C	C	C	G	G	0	0	A	A
20	WGACON 20	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	G	0	0	T	T	C	C	C	T	C	C	G	G	G	T	A	G
21	WGACON 21	0	0	1	1	C	C	G	G	G	G	0	0	T	T	C	C	A	A	G	G	A	C	T	T	C	C	C	C	C	C	G	G	G	T	A	A
22	WGACON 22	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	A	G	G	A	A	C	T	C	C	C	C	C	T	G	G	G	T	A	G
23	WGACON 23	0	0	2	1	C	T	G	G	A	A	G	G	T	T	C	C	0	0	G	G	A	A	T	T	C	C	C	C	C	C	G	G	0	0	A	A
24	WGACON 24	0	0	2	1	T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	C	T	T	C	C	C	C	C	C	G	G	G	T	A	G
25	WGACON 25	0	0	1	1	T	T	G	G	A	A	A	G	T	T	C	C	A	G	G	G	A	A	T	T	C	C	C	C	C	C	G	G	G	T	G	G
26	WGACON 26	0	0	1	1	C	T	G	G	A	G	A	G	T	T	C	C	A	A	G	G	A	A	T	T	C	T	C	C	C	C	G	G	G	T	A	G
27	WGACON 27	0	0	1	1	C	T	G	G	A	G	A	G	0	0	C	C	G	G	G	G	A	C	C	T	C	C	C	C	C	C	G	G	G	T	A	G
28	WGACON 28	0	0	1	1	T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	C	C	T	C	C	C	C	C	T	G	G	G	T	A	G
29	WGACON 29	0	0	2	1	T	T	G	G	A	A	G	G	T	T	C	C	A	G	G	G	A	A	T	T	C	C	C	C	C	C	G	G	0	0	A	G
30	WGACON 30	0	0	2	1	T	T	G	G	A	A	G	G																								
31	WGACON 31	0	0	1	1	T	T	G	G	A	A	G	G																								
32	WGACON 32	0	0	1	1	C	C	G	G	0	0	0	0																								
33	WGACON 33	0	0	1	1	T	T	G	G	A	A	G	G																								

gwas_plinkdata.ped

- 150 individuals
- 23 MZ pairs, 21 DZ pairs, 62 unrelated individuals

1) Change working directory

First, open Unix terminal. Copy the files to a folder of your preference.

Then, we need to make sure we are working in the right folder

For example:

```
cd /home/jose/2020
```

2) Check file format

PLINK-friendly formats (.bed and .ped):

- PED: pedigree information standard
- BED: compressed (binary) version of PED

```
plink --file gwas_plinkdata --make-bed

--out gwas_plinkdata
```

3) Clean the data (quality control)

- PLINK includes several options to *clean* genetic data.
- This means filtering out low quality data or outliers.
- We are going to run a very basic quality control (to learn more: next year at the IBG GWAS workshop!)

```
plink --bfile gwas_plinkdata --geno 0.05 --mind 0.05 --hwe

1e-6 --maf 0.1 --make-bed --out gwas_plinkdata_clean
```

4) Estimate the genetic relatedness matrix

Output style (A) -- matrix

```
plink --bfile gwas_plinkdata_clean --make-rel triangle
```

Take a look to the results in Unix…

```
zless -S plink.rel
```

4) Estimate the genetic relatedness matrix

- Open the file 'GRM_highlighted.pdf' to see how it looks.

- Zoom in to find which individuals are likely to be MZ twins (**in green**), DZ twins (**in red**), and genetically unrelated (not highlighted).

4) Estimate the genetic relatedness matrix

0.994539				
-0.02333	1.00109			
-0.02266	-0.02092	0.995978		
-0.01955	-0.01997	-0.01365	1.02602	
-0.02774	-0.0197	-0.0211	-0.0292	0.985955
-0.02373	-0.02732	-0.02697	-0.0151	-0.01919
-0.02781	-0.02424	-0.02777	-0.0025	-0.01934
-0.01162	-0.01133	-0.01948	-0.02476	-0.01486
-0.01204	-0.0176	-0.02524	-0.02264	-0.02792

4) Estimate the genetic relatedness matrix

Output style (B) – relatedness pair by pair

```
plink --bfile gwas_plinkdata_clean -make-grm-gz no-gz
```

Take a look to the results in Unix…

```
zless -S plink.grm
```

4) Estimate the genetic relatedness matrix

- Open the file 'grel_highlighted.xls' to see how it looks.

- You can sort the data by gen. relatedness and find which pairs are likely to be MZ twins (**in green**), DZ twins (**in red**), or unrelated individuals (not highlighted).

4) Estimate the genetic relatedness matrix

ID1	ID2	common SNPs	gen relatedness
2	1	247262	-0.0233
3	1	250987	-0.0227
3	2	247011	-0.0209
4	1	247683	-0.0195
26	3	251499	0.9960
73	51	247061	0.5048

USING these estimates in our models

- Today – within family ‘ACE’ models
  - Sarah and Lucia
- Friday – across family ‘REML’ models
  - Rob
- Next year – across family ‘GCTA’ analysis ☺

Questions?