Introduction to GWAS

Data Pre-Processing: Initial & Exploratory Data Analysis

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

HerrFalloppio

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

Oscar González-Recio

OscarGenomics

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)

Two major genotyping formats (but there are many…)

Plink format ped/map files

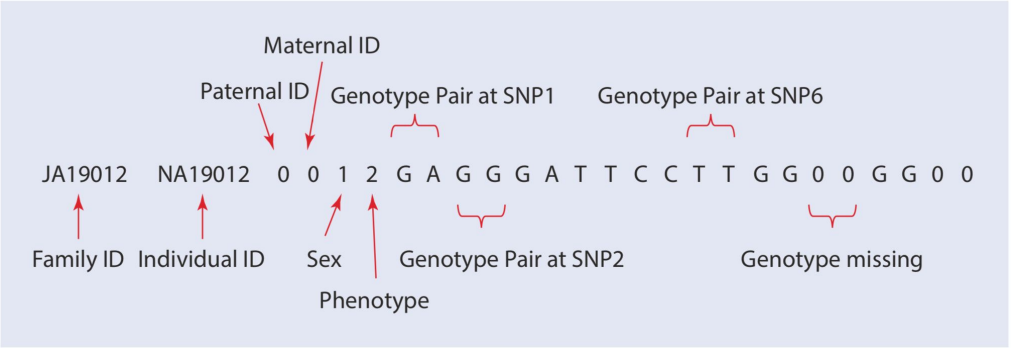
https://www.cog-genomics.org/plink2/

Variant calling format vcf files

https://www.internationalgenome.org/wiki/Analysis/vcf4.0/

https://vcftools.github.io/index.html

PLINK - the .ped file



Column 1 = Family ID
Column 2 = Individual ID
Column 3 = Paternal ID (zero for missing)
Column 4 = Maternal ID (zero for missing)
Column 5 = Sex
Column 6 = Phenotype (1=unaffected, 2=affected, and 0=missing)
Column 7, 8 = genotype pair of the first SNP1 (zero for missing)
Column 9, 10 = genotype pair of the second SNP2 (zero for means missing)
…
Column 457393, 457394 = genotype pair of the last SNP228694

Kim J.H. (2019) GWAS Data Analysis. In: Genome Data Analysis. Learning Materials in Biosciences. Springer, Singapore

PLINK - the .map file

genotype.ped

JA19012 NA19012 0 0 1 2 G A G G G A T T C C T T G G 0 0 G G 0 0

1 rs6681049 0 789870
1 rs4074137 0 1016570
1 rs7540009 0 1050098
1 rs1891905 0 1090080
1 rs9729550 0 1125105
1 rs3813196 0 1159244
1 rs6704013 0 1187454

genotype.map

Column 1 = chromosome number
Column 2 = SNP ID
Column 3 = Genetic Distance (morgans)
Column 4 = physical base-pair position (bp)

Kim J.H. (2019) GWAS Data Analysis. In: Genome Data Analysis. Learning Materials in Biosciences. Springer, Singapore

Variant calling format - the .vcf file

Data info

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

0/1 is not the same as **0|1** !! (unphased / phased)

CHROM → chromosome/contig

POS → position on chr/contig

ID → SNP name/ID

REF → reference genome allele

ALT → alternative allele (. / 1 or more)

QUAL → Phred scaled quality

-10log10 (call in Alt is wrong)

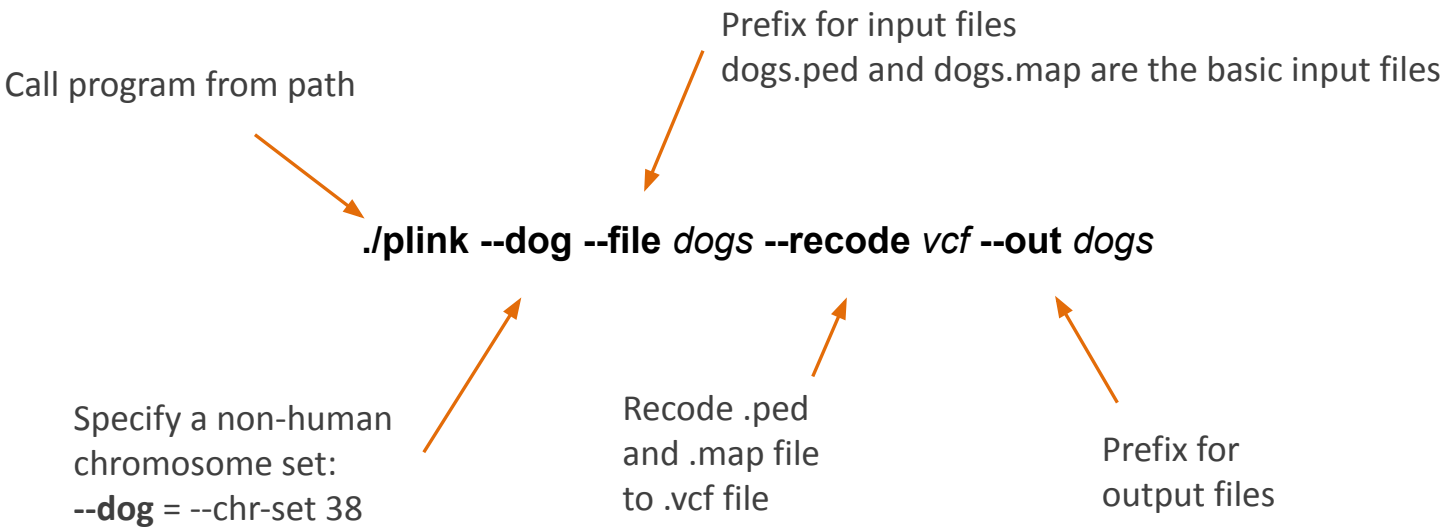
E.g. 1/10 chance of mistake → 10

FILTER □ quality filter (q10 → quality < 10)

INFO □ further information

Some basic data handling – plink (run in the shell)

Basic ***plink*** command structure: ***./plink --function specification***



Some basic data handling – plink (run in the shell)

Basic ***plink*** command structure: ***./plink --function specification***

When ped and map have different names:

plink **--dog** --ped dogs.ped --map dogs.map --recode vcf --out dogs

plink reads vcf too!

./plink --vcf dogs.vcf --recode --out dogs

Some basic data handling – vcftools (run in the shell)

Basic command structure: ***./vcftools --function specification***

./vcftools --vcf <path to vcf file> --plink --out <path to out file>

vcftools --vcf dogs.vcf --plink --out dogs_plink

(only biallelic markers will be in the output)