NGS data anlysis course

Association Analysis using PLINK

Ignacio Medina
David Montaner & Marta Bleda

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

File formats: VCF

Tab delimited text file with a **header section**

```
##fileformat=VCFv4.2
##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
```

May be compressed and indexed using `tabix`

File formats: VCF

Each variant is described by **8 fields**

1. CHROM: chromosome
2. POS: position
3. ID: name
4. REF: reference base(s)
5. ALT: non-reference alleles
6. QUAL: quality score of the calls (phred scale)
7. FILTER: PASS / filtering_tag
8. INFO: additional information

Genotype data for several samples may be included in a batch of additional columns (one for each sample) preceded by a FORMAT column which describes their format.

File formats: VCF INFO column

May include several semicolon separated fields containing information about the variants coded in key value style:

`<key>=<data>[,data]`

Some reserved (but optional) keys are:

- AA ancestral allele
- AC allele count in genotypes, for each ALT allele, in the same order as listed
- AF allele frequency
- CIGAR cigar string describing how to align an alternate allele to the reference allele
- DB dbSNP membership
- MQ RMS mapping quality, e.g. MQ=52
- MQ0 Number of MAPQ == 0 reads covering this record

File formats: PED & MAP

Classic format to represent genomic variants for several individuals

<---- normal.ped ---->

```
1 1 0 0 1  1  A A  G T
2 1 0 0 1  1  A C  T G
3 1 0 0 1  1  C C  G G
4 1 0 0 1  2  A C  T T
5 1 0 0 1  2  C C  G T
6 1 0 0 1  2  C C  T T
```

<--- normal.map --->

```
1  snp1   0  5000650
1  snp2   0  5000830
```

would be represented as TPED/TFAM files:

<------------ trans.tped ------------>

```
1 snp1 0 5000650 A A A C C C A C C C C C
1 snp2 0 5000830 G T G T G G T T G T T T
```

<- trans.tfam ->

```
1  1  0  0  1  1
2  1  0  0  1  1
3  1  0  0  1  1
4  1  0  0  1  2
5  1  0  0  1  2
6  1  0  0  1  2
```

File formats: PED & MAP

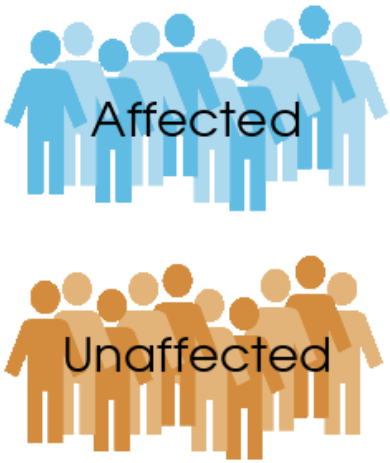
PED file

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex (1=male; 2=female; other=unknown)
6. Phenotype (1=unaffected; 2=affected; 0 missing; -9=missing)
7. . . . genotypes . . .

MAP file

1. chromosome (1-22, X, Y or 0 if unplaced)
2. rs. . . or SNP identifier
3. Genetic distance (Morgans)
4. Base-pair position (bp units)

Association to Disease / Phenotype

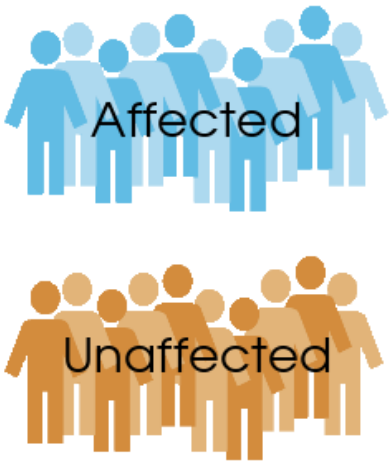


	AA	Aa	aa
Affected	AA Affected	Aa Affected	aa Affected
Unaffected	AA Unaffected	Aa Unaffected	aa Unaffected

	AA	Aa	aa
Affected	15	8	4
Unaffected	5	3	2

30+8 | 8+8
10+3 | 3+4

Association to Disease / Phenotype



	A	a
A , Affected	a , Affected	
A , Unaffected	a , Unaffected	

	A	a
Affected	38	16
Unaffected	13	7

Statistical tests

- Chi-squared test (χ^2)
- Fisher's exact
- Logistic regression models
- ...

Multiple-test correction is generally needed.
Many software available R, PLINK ...
`http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml`