NGS Data Analysis Course

Association Analysis using PLINK

Marta Bleda, Ignacio Medina, David Montaner

2015-02-22

http://www.ngscourse.org

File formats: VCF

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

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=mvImputationProgramV3.1
##reference=file:///seg/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=GO.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
                                AIT
                                        QUAL FILTER INFO
                                                                                      FORMAT
                                                                                                  NACCOCC1
                                                                                                                 NACCOCC
                                                                                                                                NACCOCC
20
       14370 rs6054257 G
                                        29 PASS
                                                    NS=3:DP=14:AF=0.5:DB:H2
                                                                                      GT:GO:DP:HO 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
       17330
                                                    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
                                             a10
       1110696 rs6040355 A
                                G.T
                                        67 PASS
                                                  NS=2:DP=10:AF=0.333.0.667:AA=T:DB GT:GD:DP:HD 1|2:21:6:23.27 2|1:2:0:18.2
                                                                                                                                2/2:35:4
20
                                        47 PASS
20
       1230237 .
                                                   NS=3:DP=13:AA=T
                                                                                      GT:GD:DP:HD 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

Association Analysis

File formats: VCF

Each variant is described by **8 fields**

CHROM: chromosome

POS: position

ID: name

REF: reference base(s)

ALT: non-reference alleles

QUAL: quality score of the calls (phred scale)

FILTER: PASS / filtering_tag

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:

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 snpl 0 5000650
1 snpl 0 5000830
```

would be represented as TPED/TFAM files:

File formats: PED & MAP

PED file

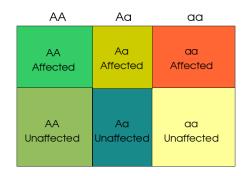
- Family ID
- Individual ID
- Paternal ID
- Maternal ID
- Sex (1=male; 2=female; other=unknown)
- Phenotype (1=unaffected; 2=affected; 0 missing; -9=missing)
- O ... genotypes ...

MAP file

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

Association to Disease / Phenotype

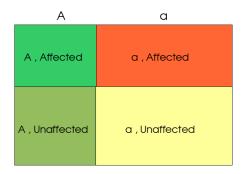




	AA	Aa	aa
Affected	15	8	4
Unaffected	5	3	2

Association to Disease / Phenotype





	Α	а
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