Documentation of the BioVis 2011 contest data

Günter Jäger

1 Contest Data

The BioVis 2011 data set consists of three different data files containing information about the genotype of the studied patients (PED), the SNPs that are investigated (MAP) and the phenotypes/gene expression (PHEN) of the studied genes. All three files are specified in a way that they are directly applicable in the whole genome association analysis toolset PLINK [1].

1.1 PED file

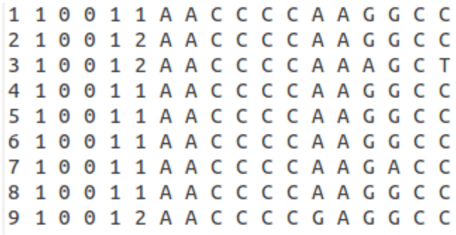
A PED file is a white space or tab delimited file without a header row containing the following information in the first six columns:

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex (1=male; 2=female; other=unknown)
6. Phenotype

A person is uniquely identified by the combination of family and individual ID. The PED file has precisely one phenotype in the sixth column, which can either be a quantitative trait or an affection status. The affection status is by default coded as follows:

- -9: missing
- 0: missing
- 1: unaffected
- 2: affected

Column seven onwards are also white-space or tab delimited and contain information about the persons' genotypes. Genotypes can be any character (e.g. 1,2,3,4 or A,C,G,T or anything else) except 0, which is by default the missing genotype character. All markers are biallic which means that all SNPs whether haploid of not must have two alleles specified. Either both alleles should be missing or neither. Figure 1 gives an example of a typical PED file.



```
1 1 0 0 1 1 A A C C C C A A G G C C
2 1 0 0 1 2 A A C C C C A A G G C C
3 1 0 0 1 2 A A C C C C A A A G C T
4 1 0 0 1 1 A A C C C C A A G G C C
5 1 0 0 1 1 A A C C C C A A G G C C
6 1 0 0 1 1 A A C C C C A A G G C C
7 1 0 0 1 1 A A C C C C A A G A C C
8 1 0 0 1 1 A A C C C C A A G G C C
9 1 0 0 1 2 A A C C C C G A G G C C
```

Figure 1: Example of a PED file.

1.2 MAP file

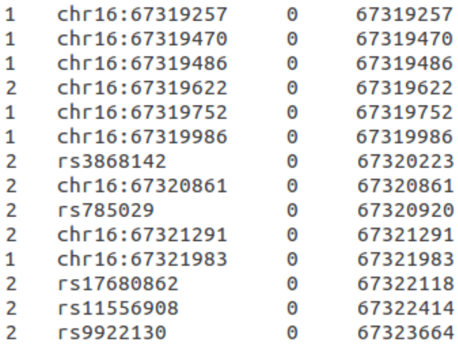
MAP files describe markers and contain exactly four columns:

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)

Base-pair positions are expected to be positive integers within the range of typical human chromosome sizes. SNP identifier can contain any characters except spaces or tabs. The MAP file must contain as many markers as are in the PED file. Thereby, the order of the markers in the PED file should align with the order in the MAP file. The autosomes are coded 1 through 22. The following other codes are used to specify other chromosome types:

- X: X chromosome
- Y: Y chromosome
- XY: Pseudo-autosomal region of X
- MT: Mitochondrial

Figure 2 shows an example of MAP file conaining 14 different markers.



```
1    chr16:67319257     0     67319257
1    chr16:67319470     0     67319470
1    chr16:67319486     0     67319486
2    chr16:67319622     0     67319622
1    chr16:67319752     0     67319752
1    chr16:67319986     0     67319986
2    rs3868142          0     67320223
2    chr16:67320861     0     67320861
2    rs785029           0     67320920
2    chr16:67321291     0     67321291
1    chr16:67321983     0     67321983
2    rs17680862         0     67322118
2    rs11556908         0     67322414
2    rs9922130          0     67323664
```

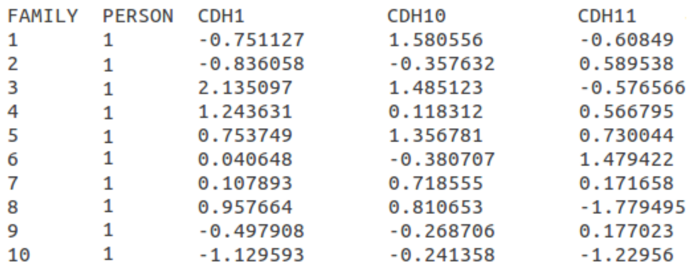
Figure 2: Example of a MAP file.

1.3 PHEN file

The PHEN file is a txt file that contains at least three columns:

1. Family ID
2. Individual ID
3. Phenotype

As in the PED file the first two columns uniquely identify a person. If an individual is listed in the PED file, but not in the PHEN file, that persons phenotype will be set to missing. Likewise, a persons phenotype will be ignored if it is listed in the PHEN file, but the person is not contained in the PED file. The PHEN file must contain at least one phenotype. If more than one phenotype is given, column four onwards are used for the others. Variable names are not allowed to have whitespaces in them. The PHEN file can have a header row, which must be specified when loading the file into Reveal. Figure 3 shows an example of a PHEN file with three different phenotypes (i.e., five columns in total).



```
FAMILY  PERSON  CDH1            CDH10           CDH11
1       1       -0.751127       1.580556        -0.60849
2       1       -0.836058       -0.357632       0.589538
3       1       2.135097        1.485123        -0.576566
4       1       1.243631        0.118312        0.566795
5       1       0.753749        1.356781        0.730044
6       1       0.040648        -0.380707       1.479422
7       1       0.107893        0.718555        0.171658
8       1       0.957664        0.810653        -1.779495
9       1       -0.497908       -0.268706       0.177023
10      1       -1.129593       -0.241358       -1.22956
```

Figure 3: Example of a PHEN file with three different phenotypes.

2 PLINK results

2.1 Single Locus

PLINK allows one to test quantitative traits for association using e.g. asymptotic (Wald test) significance values. If the phenotypes (`--pheno` option in PLINK) in the PHEN file are quantitative (i.e. contain values other than 1, 2, 0 or missing) then PLINK will automatically treat the analysis as a quantitative trait analysis and it will generate for each phenotype a file called `plink.phenotype.qassoc` where `phenotype` is the name of the phenotype in the PHEN file. Such a single locus result file contains the following columns:

1. CHR: Chromosome number
2. SNP: SNP identifier
3. BP: Physical position (base-pair)
4. NMISS: Number of non-missing genotypes

5. BETA: Regression coefficient

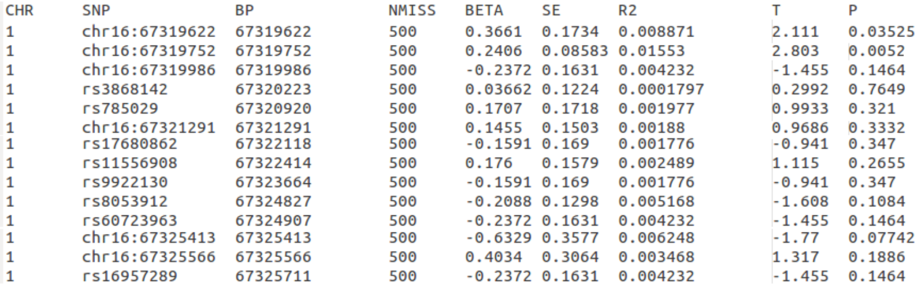
6. SE: Standard error

7. R2: Regression r-squared

8. T: Wald test (based on t-distribution)

9. P: Wald test asymptotic p-value

An example of a single locus result file is shown in figure 4.



```
CHR     SNP              BP              NMISS   BETA    SE      R2              T       P
1       chr16:67319622  67319622         500     0.3661  0.1734  0.008871        2.111   0.03525
1       chr16:67319752  67319752         500     0.2406  0.08583 0.01553         2.803   0.0052
1       chr16:67319986  67319986         500     -0.2372 0.1631  0.004232        -1.455  0.1464
1       rs3868142       67320223         500     0.03662 0.1224  0.0001797       0.2992  0.7649
1       rs785029        67320920         500     0.1707  0.1718  0.001977        0.9933  0.321
1       chr16:67321291  67321291         500     0.1455  0.1503  0.00188         0.9686  0.3332
1       rs17680862      67322118         500     -0.1591 0.169   0.001776        -0.941  0.347
1       rs11556908      67322414         500     0.176   0.1579  0.002489        1.115   0.2655
1       rs9922130       67323664         500     -0.1591 0.169   0.001776        -0.941  0.347
1       rs8053912       67324827         500     -0.2088 0.1298  0.005168        -1.608  0.1084
1       rs60723963      67324907         500     -0.2372 0.1631  0.004232        -1.455  0.1464
1       chr16:67325413  67325413         500     -0.6329 0.3577  0.006248        -1.77   0.07742
1       chr16:67325566  67325566         500     0.4034  0.3064  0.003468        1.317   0.1886
1       rs16957289      67325711         500     -0.2372 0.1631  0.004232        -1.455  0.1464
```

Figure 4: Example of a single locus result file generated by PLINK.

2.2 Two Locus

PLINK also allows one to test SNP x SNP epistasis for case/control population based samples. The default test uses linear regression if the phenotype is a quantitative trait. The test only considers allelic by allelic epistasis. If the `--phen` option is specified PLINK produces for each phenotype a file called `plink.phenotype.epi.qt` where `phenotype` equals to the phenotype name in the PHEN file. Such a two locus result file is in the form:

1. CHR1: Chromosome of the first SNP (or index of the quantitative trait associated with the first SNP if `--pheno` is specified)

2. SNP1: Identifier of the first SNP

3. CHR2: Chromosome of the second SNP (or index of the quantitative trait associated with the second SNP if `--pheno` is specified)

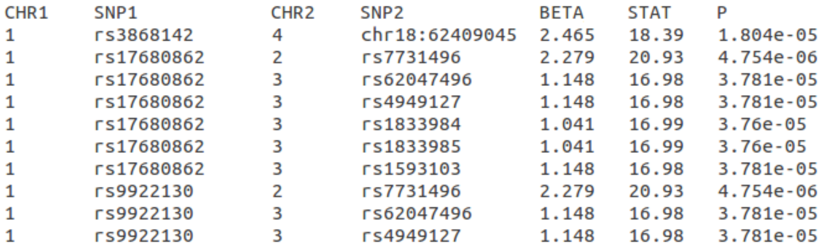
4. SNP2: Identifier of the second SNP

5. BETA: Odds ratio for the interaction

6. STAT: χ^2 statistic with one degree of freedom

7. P: Asymptotic p-value

An example of a two locus result file is shown in figure 5.



```
CHR1    SNP1          CHR2    SNP2             BETA    STAT    P
1       rs3868142     4       chr18:62409045   2.465   18.39   1.804e-05
1       rs17680862    2       rs7731496        2.279   20.93   4.754e-06
1       rs17680862    3       rs62047496       1.148   16.98   3.781e-05
1       rs17680862    3       rs4949127        1.148   16.98   3.781e-05
1       rs17680862    3       rs1833984        1.041   16.99   3.76e-05
1       rs17680862    3       rs1833985        1.041   16.99   3.76e-05
1       rs17680862    3       rs1593103        1.148   16.98   3.781e-05
1       rs9922130     2       rs7731496        2.279   20.93   4.754e-06
1       rs9922130     3       rs62047496       1.148   16.98   3.781e-05
1       rs9922130     3       rs4949127        1.148   16.98   3.781e-05
```

Figure 5: Example of a two locus result file generated by PLINK.

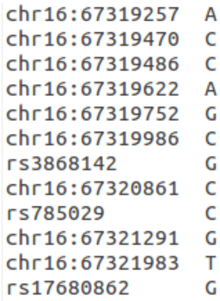
3 Additional files

3.1 SNP reference file (REF)

The SNP reference file is a simple text file without a header row containing exactly two tab delimited columns. These are:

1. SNP: SNP identifier as specified in the MAP file
2. REF: Reference nucleotide

The reference column should contain one of the following characters: A, G, C, T or N if the reference is unknown. An example of a SNP reference file is shown in figure 6.



```
chr16:67319257  A
chr16:67319470  C
chr16:67319486  C
chr16:67319622  A
chr16:67319752  G
chr16:67319986  C
rs3868142       G
chr16:67320861  C
rs785029        C
chr16:67321291  G
chr16:67321983  T
rs17680862      G
```

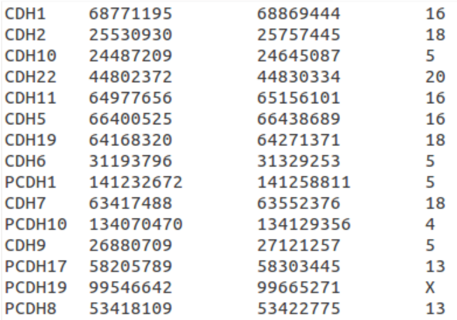
Figure 6: Example of a SNP reference file.

3.2 Quantitatice trait locations file (LOC)

The trait locations file (LOC file) is a tab delimited text file without a header row that contains exactly the following four columns:

1. PHEN: the quantitative trait identifier as specified in the PHEN file
2. START: the start position of the quantitative trait (base-pair)
3. STOP: the end position of the quantitative trait (base-pair)
4. CHR: the chromosome the quantitative trait lies on

The base at the end position of the quantitative trait is included in the quantitative trait. Thus, the length of the quantitative trait is stop $-$ start $+1$. The chromosome should be 1-22 for autosomes, X, Y or MT (mitochondrial). An example of a LOC file is shown in figure 7.



```
CDH1     68771195      68869444      16
CDH2     25530930      25757445      18
CDH10    24487209      24645087      5
CDH22    44802372      44830334      20
CDH11    64977656      65156101      16
CDH5     66400525      66438689      16
CDH19    64168320      64271371      18
CDH6     31193796      31329253      5
PCDH1    141232672     141258811     5
CDH7     63417488      63552376      18
PCDH10   134070470     134129356     4
CDH9     26880709      27121257      5
PCDH17   58205789      58303445      13
PCDH19   99546642      99665271      X
PCDH8    53418109      53422775      13
```

Figure 7: Example of a LOC file containing genomic positions for quantitative traits.

References

[1] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. De Bakker, M. Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.