AlphaPeel Quickstart

Contents

Introduction

AlphaPeel is a pedigree based phasing and imputation and calling algorithm. The program uses pedigree based information to call and impute low coverage sequence data invery large populations. A complete description of the methods is given in Whalen et al (Under review; https://doi.org/10.1101/228999).

Please report bugs or suggestions on how the program / user interface / manual could be improved or made more user friendly to John.Hickey@roslin.ed.ac.uk or Awhalen@roslin.ed.ac.uk.

Availability

AlphaPeel is available from AlphaGenes website. Material available comprises the compiled programs for Windows, Linux, and Mac OSX machines, together with this document and an example.

Conditions of use

AlphaPeel is available to the scientific community free of charge. Users are required, however, to credit its use in any publications. Commercial users should contact John Hickey (John.Hickey@roslin.ed.ac.uk).

Suggested Citation:

Whalen, A, Ros-Freixedes, R, Wilson, DL, Gorjanc, G, Hickey, JM. (2017). *Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees*. bioRxiv 228999; doi: https://doi.org/10.1101/228999

Disclaimer

While every effort has been made to ensure that **AlphaPeel** does what it claims to do, there is absolutely no guarantee that the results provided are correct. Use of **AlphaPeel** is entirely at your own risk!

Run commands and spec file

AlphaPeel is a pedigree based phasing and imputation algorithm. This algorithm implements a version of single locus iterative peeling and multi locus iterative peeling. Combined, the algorithms can perform hybrid peeling. To run **AlphaPeel** call `./|ap|` (linux) or `|ap|.exe` (windows) with the spec file as the first argument. **AlphaPeel** defaults to using a spec file called `|ap|Spec.txt`.

A sample spec file might look like

```
#Required paramaters
nsnp, 300                       #number of SNPs
inputfilepath, genotypes.txt    #input genotype file
pedigree, pedigree.txt          #input pedigree file
outputfilepath, output          #output file prefix
runtype, multi                  #Run type, either "multi" or "single"

#Optional paramaters
ncycles, 10                     #number of peeling cycles (default:10)
startsnp, 1                     #start SNP (default: 1)
endsnp, 2000                    #end snp (default: nsnp)

#Sequence paramaters
sequencefile, sequence.data     #input sequence files
useSequence, Yes                #Whether or not to use sequence data
mapfile, map.txt                #input map file
segfile, output.seg             #input segregation file
```

The program can either be run in a single locus peeling mode or a multi locus peeling mode. In both cases `nsnp` gives the number of markers in the file. `inputfilepath` provides the input genotype file, `pedigree` provides the input pedigree file, and for sequence data `sequencefile` provides the input sequence file. For sequence data use the following paramaters

```
inputfilepath, None
sequencefile, sequence.data
useSequence, Yes
```

AlphaPeel does not currently support using both sequence data and genotype data in a single run.

The behaviour of the algorithm can be changed via a number of optional parameters. `ncycles` is the number of iterative peeling cycles the algorithm will perform (default:10). **AlphaPeel** can also be run on only a subset of the chromosome. If only a subset needs to be analysed, use `startsnp` and `endsnp`, otherwise **AlphaPeel** will run on the entire chromosome.

For hybrid peeling, **AlphaPeel** can take in a custom segregation file, `segfile`, and a map file, `mapfile`, to determine the pattern of inheritance for each individual.

After each run, **AlphaPeel** produces a `.haps`, `.dosage`, `.params` file and if run in multi locus peeling mode, a `.seg` file. More details on each file type are below.

Input file formats

Genotype file

Genotype files contain the input genotypes for each individual. The first value in each line is the individual's id. The remaining values are the genotypes of the individual at each locus, either 0, 1, or 2 (or 9 if missing).

Example:

```
id1 1 1 2 0 1 1 1 1 1 0
id2 0 2 1 1 0 1 1 1 2 2
id3 1 2 0 1 2 1 0 1 2 0
id4 2 1 1 1 1 1 1 1 2 1
```

Sequence file

The sequence data file is in a similar Sequence data is given in a similar format to the genotype data. For each individual there are two lines. The first line gives the individual's id and the read counts for the reference allele. The second line gives the individual's id and the read counts for the alternative allele.

Example:

```
id1 0 0 0 2 1 1 2 0 2 1
id1 3 1 2 1 2 2 1 4 4 3
id2 1 0 1 1 4 2 1 2 3 1
id2 0 1 2 1 1 1 3 2 3 2
id3 0 2 1 3 2 1 3 1 2 2
id3 2 3 3 1 2 2 0 2 1 2
id4 1 1 4 0 0 1 1 2 1 1
id4 1 3 2 3 2 1 2 2 1 3
```

Pedigree file

Each line of a pedigree file has three values, the individual's id, their father's id, and their mother's id. "0" represents an unknown id.

Example:

```
id1 0 0
id2 0 0
id3 id1 id2
id4 id1 id2
```

Map file

The map file specifies the distance between loci in the sequence file and the loci in the segregation file. Each line contains information on a single locus in the sequence file. The first column gives the index of the immediately preceding locus in the segregation file. The second column gives the index of the immediately proceeding locus. The third column gives the relative distance between the sequence locus and the preceding and proceeding locus.

Example:

```
62 63 0.7
62 63 0.8
62 63 0.9
63 63 0
63 64 0.1
63 64 0.2
63 64 0.3
```

Output file formats

Haplotype file

The haplotype file (*.haps*) provides the (phased) allele probabilities for each locus. There are four lines per individual containing the allele probability for the (aa, aA, Aa, AA) alleles where the paternal allele is listed first,

and where *a* is the reference (or major) allele and *A* is the alternative (or minor) allele.

Example:

```
id1    0.9998    0.0001    0.0001    1.0000
id1    0.0000    0.4999    0.4999    0.0000
id1    0.0000    0.4999    0.4999    0.0000
id1    0.0001    0.0001    0.0001    0.0000
id2    0.0000    1.0000    0.0000    1.0000
id2    0.9601    0.0000    0.0455    0.0000
id2    0.0399    0.0000    0.9545    0.0000
id2    0.0000    0.0000    0.0000    0.0000
id3    0.9998    0.0001    0.0001    1.0000
id3    0.0000    0.4999    0.4999    0.0000
id3    0.0000    0.4999    0.4999    0.0000
id3    0.0001    0.0001    0.0001    0.0000
id4    1.0000    1.0000    0.0000    1.0000
id4    0.0000    0.0000    0.0000    0.0000
id4    0.0000    0.0000    0.0000    0.0000
id4    0.0000    0.0000    1.0000    0.0000
```

Dosage file

The dosage file gives the expected allele dosage for the alternative (or minor) allele for each individual. The first value in each line is the individual ID. The remaining values are the allele dosages at each loci.

Example:

```
1    0.0003    1.0000    1.0000    0.0001
2    1.0000    0.0000    1.0000    0.0000
3    0.0003    1.0000    1.0000    0.0001
4    0.0000    0.0000    2.0000    0.0000
```

Segregation file

The segregation file gives the joint probability of each pattern of inheritance. There are four lines for each individual representing the probability of inheriting:

1. the grand **paternal** allele from the father and the grand **paternal** allele from the mother
2. the grand **paternal** allele from the father and the grand **maternal** allele from the mother
3. the grand **maternal** allele from the father and the grand **paternal** allele from the mother
4. the grand **maternal** allele from the father and the grand **maternal** allele from the mother

Example:

```
id1    1.0000    0.9288    0.9583    0.9834
id1    0.0000    0.0149    0.0000    0.0000
id1    0.0000    0.0554    0.0417    0.0166
id1    0.0000    0.0009    0.0000    0.0000
id2    0.9810    0.9842    1.0000    0.9971
id2    0.0174    0.0158    0.0000    0.0013
id2    0.0016    0.0000    0.0000    0.0016
id2    0.0000    0.0000    0.0000    0.0000
id3    0.0164    0.0149    0.0000    0.0065
id3    0.9259    0.9288    0.9582    0.9769
id3    0.0010    0.0009    0.0000    0.0001
id3    0.0567    0.0554    0.0417    0.0165
id4    0.0002    0.0000    0.0002    0.0004
id4    0.0015    0.0000    0.0019    0.0041
id4    0.1189    0.1179    0.1052    0.0834
id4    0.8794    0.8821    0.8927    0.9122
```

Parameter file

The parameter file (.params) gives the inferred model parameters for each locus. These parameters include an ancestral minor allele frequency (maf), a genotyping error rate and a recombination rate. For single locus peeling, only the first two columns are outputted.

Example:

```
maf genotypeError recombinationRate
0.1305504   0.0005969   0.0032816
0.2610893   0.0005378   0.0319264
0.4162576   0.0008617   0.0529236
```