Mangrove: Genetic risk prediction on trees

Luke Jostins

March 8, 2012

Abstract

`Mangrove` is an R package for performing genetic risk prediction from genotype data. You can use it to perform risk prediction for individuals, or for families with missing data.

In general, you will require an Odds Ratio object, which contains the risk alleles, the frequencies and the odds ratios for all the risk variants. You will also require a pedigree object, which contains genotypes for each individual, and their relationship to each other.

This vignette goes through some examples that illustrate how to use the various functions that make up `Mangrove`.

1 The odds ratios object

We start by loading the Mangrove library:

```
> library(Mangrove)
```

We can read in odds ratios from a text file. For instance, the odds ratio file that corresponds to the `exampleORs` object contains:

```
rsID    RiskAllele      OR      Freq
SNP1    A       1.5     0.2
SNP2    C       1.3     0.4
SNP3    C       1.4     0.6
SNP4    A       2.0     0.01
```

In this case, as there is only one "OR" field, `Mangrove` assumes that the variants all have additive risk (i.e. ORhet=OR, ORhom=OR*OR). If you include two OR fields,"ORhet" and "ORhom", `Mangrove` will use both appropriately.

We can read in the odds ratio file using the `readORs` function. However, we are going to use the preloaded example object `exampleORs`:

```
> data(exampleORs)
> class(exampleORs)

[1] "MangroveORs" "data.frame"

> print(exampleORs)
```

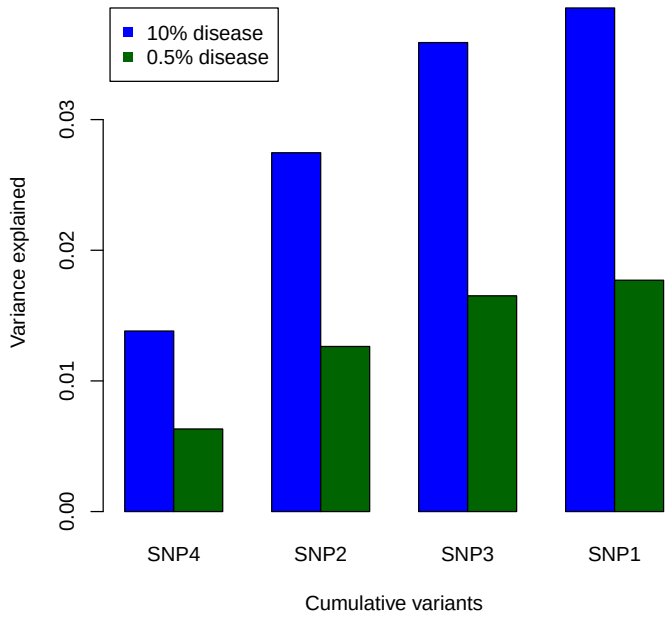
```
  rsID RiskAllele ORhet ORhom Freq
1 SNP1          A   1.5  2.25 0.20
2 SNP2          C   1.3  1.69 0.40
3 SNP3          C   1.4  1.96 0.60
4 SNP4          A   2.0  4.00 0.01
```

There are a few methods association with the `MangroveORs` class. `summary` gives you a general idea of how predictive the set of variants is, including the variance explained (on the liability scale) for a few example prevalences. `plot` gives the cumulative variance explained as variants are added in (in the order of most predictive first):

```
> summary(exampleORs)

A Mangrove Odds Ratios object:
Number of variants:4
Mean absolute het OR: 1.55
Mean absolute hom OR: 2.475
Mean risk allele frequency: 0.302
For a common (10%) disease, these variants explain 3.85% of variance
For a rare (0.5%) disease, these variants explain 1.77% of variance

> plot(exampleORs)
```

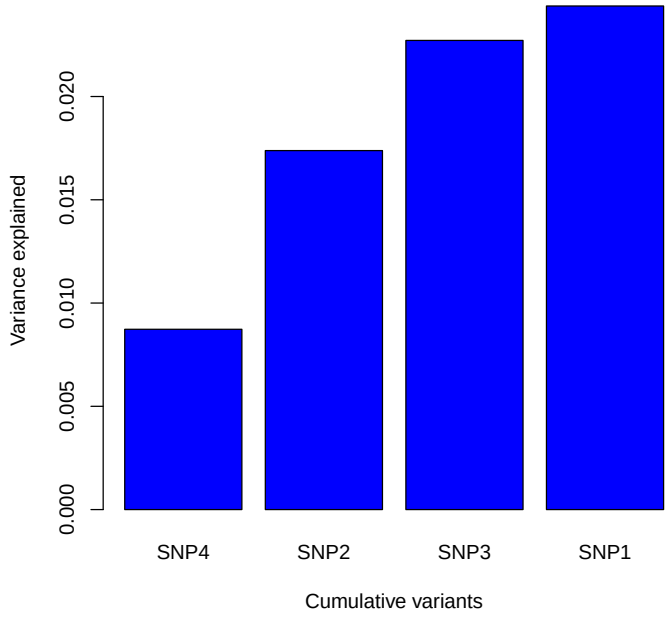


If you know the actual prevalence of your disease, you can get more accurate figures. For instance, suppose the prevalence of the disease that the odds ratios predict is $K = 0.02$:

```
> summary(exampleORs,K=0.02)

A Mangrove Odds Ratios object:
Number of variants:4
Mean absolute het OR: 1.55
Mean absolute hom OR: 2.475
Mean risk allele frequency: 0.302
Given a prevalence of 2% these variants explain 2.44% of variance

> plot(exampleORs,K=0.02)
```



2 Risk prediction in unrelated individuals

To perform genetic risk prediction, you need genotypes for your individuals. We read genetic data in from pedigree/map file pairs, such as those produced by the program `Plink`.

You can read in pedigree files using the `readPed` function. However, for this example we will use an example dataset from a large cohort of unrelated cases and controls (cases have the disease predicted by the odds ratios in `exampleORs`):

```
> data(ccped)
> class(ccped)

[1] "MangrovePed" "data.frame"
```

We can take a look at the contents of this pedigree object:

```
> head(ccped)
```

```
      Family    ID Mother Father Sex Phenotype SNP1.1 SNP1.2 SNP2.1 SNP2.2
ID106 FAM106 ID106      0      0   2         2      A      C      A      A
ID110 FAM110 ID110      0      0   2         2      C      C      C      C
ID170 FAM170 ID170      0      0   1         2      C      C      C      C
ID189 FAM189 ID189      0      0   2         2      C      C      A      A
ID200 FAM200 ID200      0      0   2         2      C      C      A      A
ID331 FAM331 ID331      0      0   2         2      C      C      C      A
      SNP3.1 SNP3.2 SNP4.1 SNP4.2
ID106      C      C      G      G
ID110      A      C      G      G
ID170      C      C      G      G
ID189      A      C      G      G
ID200      C      C      G      G
ID331      A      C      G      G
```

```
> summary(ccped)
```

```
A Mangrove pedigree.
Number of individuals: 20000
Number of genotyped individuals: 20000
Number of affecteds: 10000
Number of markers: 4
Allele counts:
 SNP1       SNP2      SNP3      SNP4
 AA: 1136   AA:6574   AA:2696   CC:    1
 AC: 7049   CA:9597   AC:9153   CG:  387
 CC:11815   CC:3829   CC:8151   GG:19612
```

This tells us that there are 20K individuals, split evenly between cases and controls, genotyped for 4 variants.

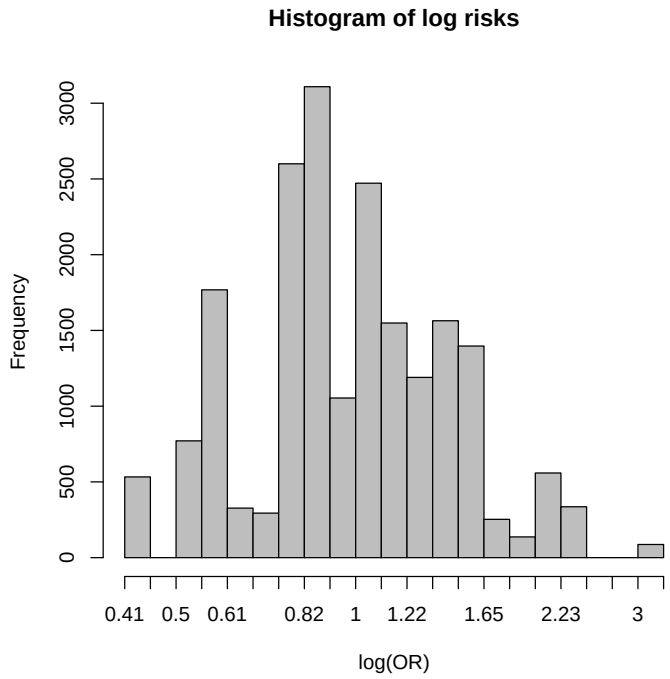
We can use our odds ratio object to perform risk prediction on these individuals:

```
> ccrisk <- calcORs(ccped,exampleORs)
> class(ccrisk)
```

```
[1] "MangroveRiskPreds"
```

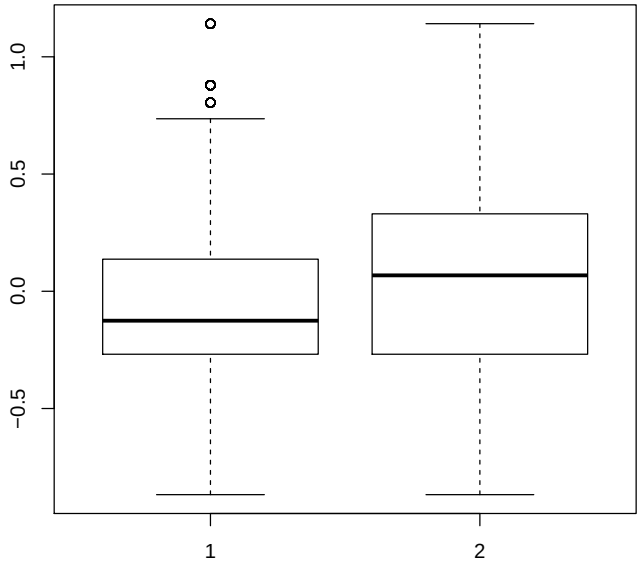
This object contains combined odds ratios, relative to population average, for every individual in the pedigree object. The `plot` method shows the distribution of this on the log scale, which should be approximately normally distributed (though for 4 variants, the approximation will be very approximate):

```
> plot(ccrisk)
```



Looking at the distribution in cases and controls shows that, as expected, the odds ratios are significantly higher in cases over controls:

```
> boxplot(log(ccrisk) ~ ccped[,6])
```

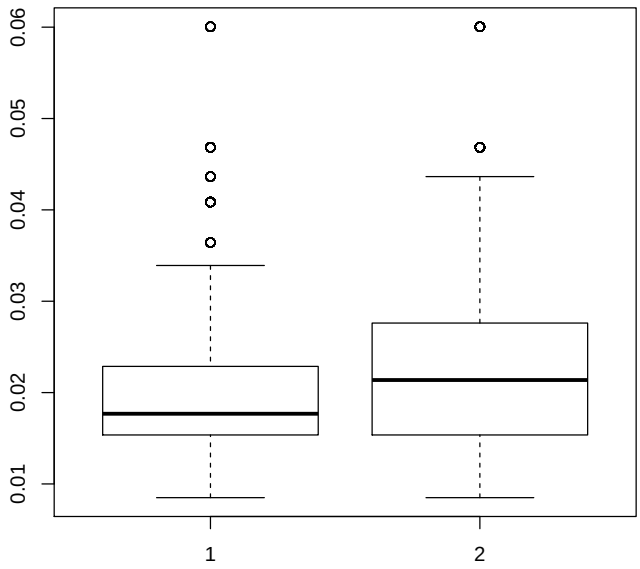


We can calculate posterior probabilities of disease incidence from the odds ratios, given a prevalence. Again assuming $K = 0.02$:

```
> ccprob <- applyORs(ccrisk,K=0.02)
> summary(ccprob)

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.008499 0.015360 0.019880 0.020960 0.024580 0.060050

> boxplot(ccprob ~ ccped[,6])
```



If we wish to prioritise individuals for sequencing, our best bet is to pick the cases with the lowest genetic risk. For instance, to select 1000 cases:

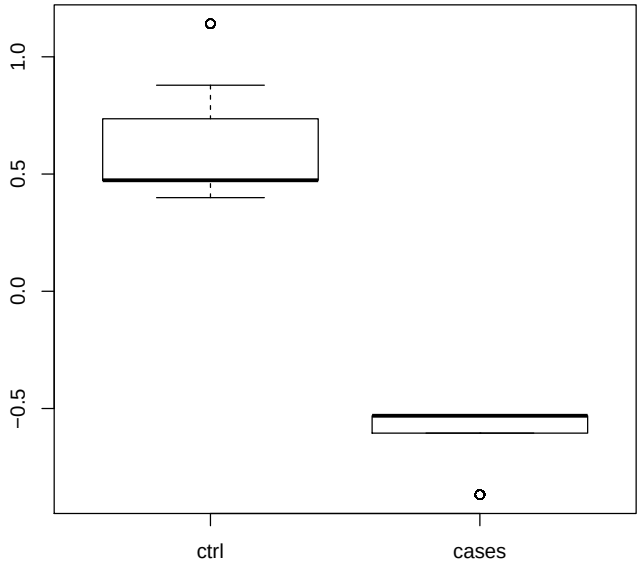
```
> selcases <- names(head(sort(ccrisk[ccped[,6] == 2]),1000))
```

We could also select 1000 matching control with the highest risk:

```
> selcontrols <- names(tail(sort(ccrisk[ccped[,6] == 1]),1000))
```

As expected, this pulls out a set of cases with much genetic risk than the set of controls:

```
> boxplot(list(ctrl=log(ccrisk[selcontrols]),cases=log(ccrisk[selcases])))
```



3 Quantitative trait prediction in unrelated individuals

You can also use Mangrove to perform continuous risk prediction on unrelated individuals. For continuous prediction we have a beta file, which contains beta-values (the equivalent of odds ratios in quantitative trait prediction). These have a similar format to the odds ratio file, and are read in using `readBetas`. All the same methods apply (plot, summary, print).

We will use an example file, which actually contains beta values for 179 SNPs that predict height:

```
> data(exampleBetas)
> class(exampleBetas)

[1] "MangroveBetas" "data.frame"

> summary(exampleBetas)

A Mangrove Betas object:
Number of variants:179
Mean absolute beta: 0.035
Mean risk allele frequency: 0.474
These variants explain 8.84% of variance
```

We will also look at 1000 (simulated) female individuals genotyped at these sites:

```
> data(contped)
> class(contped)
```

```
[1] "MangrovePed" "data.frame"
```

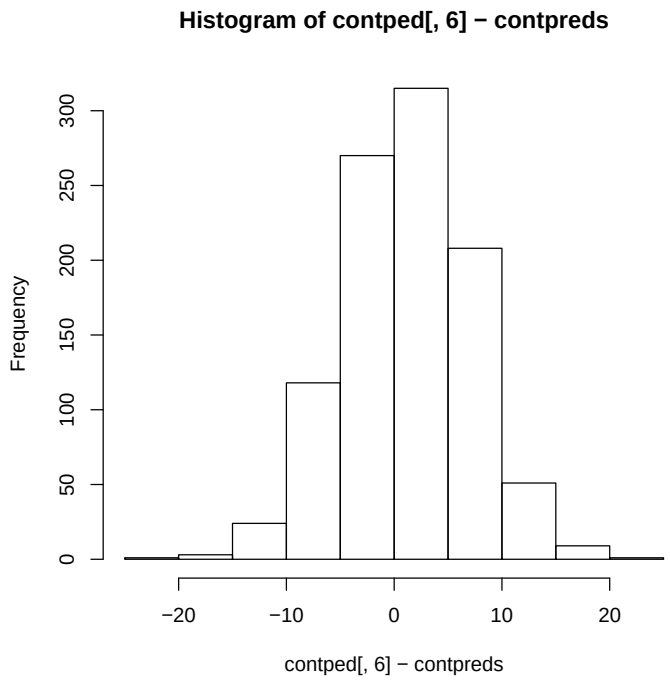
The mean female height is around 163cm, and the standard deviation is around 6.4cm, so we can perform .

```
> predbetas <- calcBetas(contped,exampleBetas)
> contpreds <- applyBetas(predbetas,162,6.4)
> summary(contpreds)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  156.5   160.7   162.0   162.0   163.2   167.4
```

If you are prioritising people for sequencing to find eQTLs, the best method is to sample people who have significantly larger and significantly smaller values than predicted from known genetics:

```
> hist(contped[,6] - contpreds)
```



4 Risk prediction in families

Finally, we can consider the case of risk prediction in families. Suppose we have a family that we are considering sequencing, due to their higher-than-expected prevalence of the disease. We have genotyped them, and can get their pedigree file:

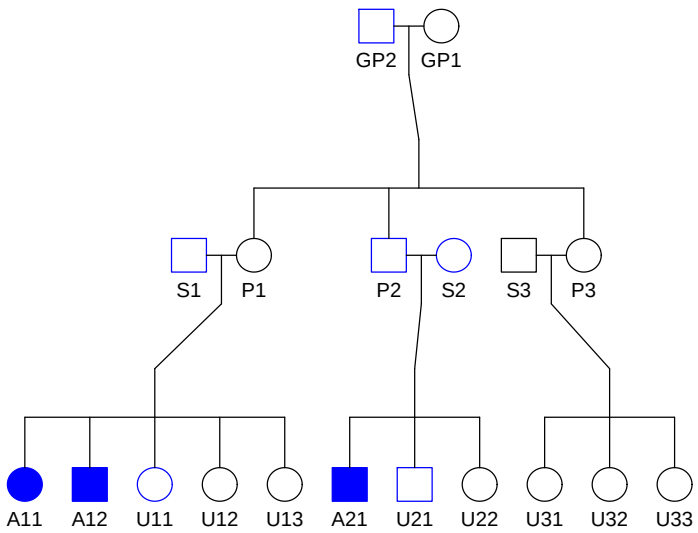
```
> data(famped)
> summary(famped)

A Mangrove pedigree.
Number of individuals: 19
Number of genotyped individuals: 9
Number of affecteds: 3
Number of markers: 4
Allele counts:
 SNP1   SNP2   SNP3   SNP4
 AA:4   AA:1   AA:1   AG:6
 AC:4   AC:3   AC:5   GG:3
 CC:1   CC:5   CC:3
```

We can see that the family has three affected individuals, out of a total of 19. 9 family members have been genotyped for the same 4 variants as above.

We can use the package `kinship2` to plot the pedigree, colouring individuals we have genotypes for blue:

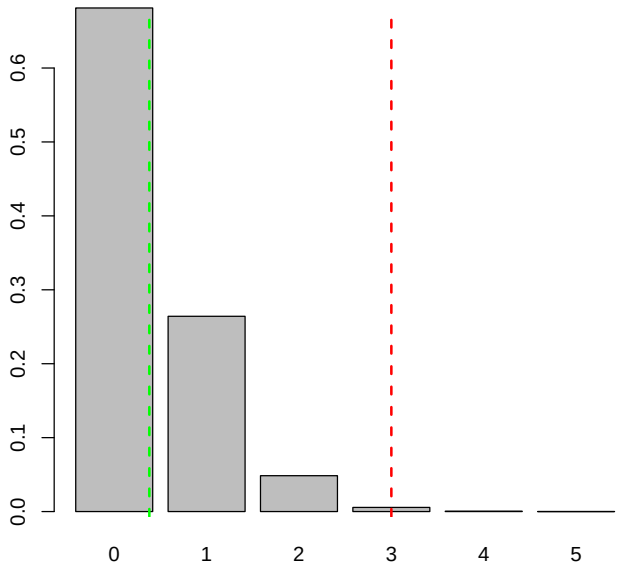
```
> library(kinship2)
> missing <- (apply(famped[,-c(1:6)] == 0,1,sum) == 0)
> kinped <- pedigree(famped$ID,famped$Father,famped$Mother,famped$Sex,famped$Phenotype,mis
> plot(kinped,col=missing*3 + 1)
```



We can ask, given a naive binomial model, what the distribution of number of affecteds would be in a family of this size, assuming no genetic effects, using the `plotNaivePrev` function. The green line is the expected number in a family

of this size, the red is the observed number, and the grey bars of the expected distribution.

```
> plotNaivePrev(famped,K=0.02)
```



We can see that the prevalence is far higher than would be expected by chance. We can verify this by calculating a p-value:

```
> p <- 1 - pbinom(2,19,0.02)
> print(p)

[1] 0.006098341
```

However, the higher prevalence could by simply due to a higher load of known genetic risk factors. As we saw when we ran `summary`, the family does seem to have a pretty high frequency of the low-frequency risk allele of SNP4. Can this explain the number of cases we observe?

We can use the Inside-Outside Algorithms to sample from the posterior distribution of number of affecteds. First of all, we initialise a tree object, and load the genetic data into it:

```
> tree <- initialiseTree()
> class(tree)

[1] "MangroveTree"

> tree$addPed(famped,exampleORs)
> summary(tree)
```

```
A mangrove tree
Number of nodes: 15
Max depth: 3
Genotype data: loaded
Model parameters: not calculated
Sampling: not performed
```

```
> print(tree)
```

```
A Mangrove tree with attached data:
-GP2-GP1
--P1-S1
---A11
---A12
---U11
---U12
---U13
--P2-S2
---A21
---U21
---U22
--P3-S3
---U31
---U32
---U33
```

We can then perform the sampling, given the odds ratios and prevalences:

```
> sam <- tree$getPrevs(exampleORs,K=0.02)
```

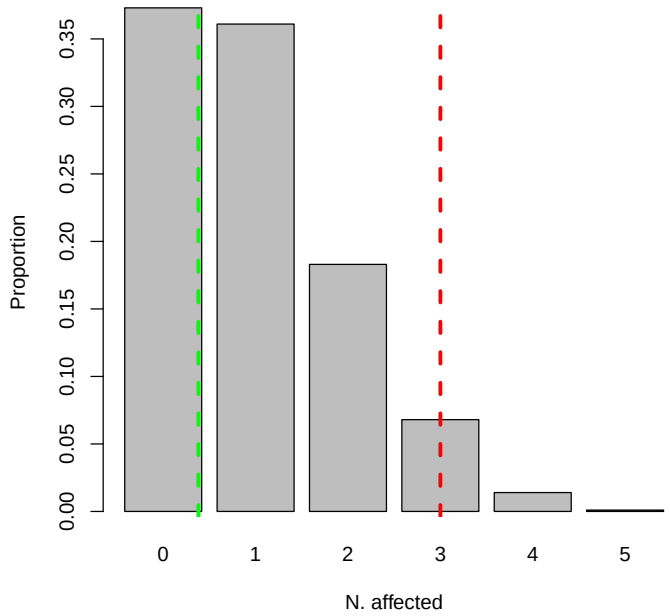
```
Running Inside-Outside Algorithm
Calculating likelihoods.....done
Calculating inside parameters.....done
Calculating outside parameters.....done
Calculating posteriors.....done
Sampling variants from posterior....done
Sampling cases given sampled variants.....done
```

```
> class(sam)
```

```
[1] "MangroveSample"
```

We can view the expected distribution of affecteds given the observed genotypes using the `plot` method

```
> plot(sam)
```



We can see that the number of affecteds we see no longer looks that unexpected. Once again, we can quantify this with a p-value, in this case using the `summary` method:

```
> summary(sam)

A Mangrove family prevalence simulation.
The pedigree has 3 cases out of 19 individuals.
The prevalence is 0.02.
Expected number of cases given common variants (95% CI): 0.992 (0 - 3)
Probability of seeing at least 3 cases given common variants: 0.083
```

This family is not significantly enriched for cases over-and-above what we would expect from the known risk loci. Thus the family is probably not a good candidate for further study.

If we did want to pursue this family, we could look at the risk predictions for the affected family members:

```
> famrisk <- calcORs(famped,exampleORs)
> print(famrisk[famped[,6] == 2])

     A11      A12      A21
2.981456 4.472185 1.146714
```

We can see that A21 does not have a particularly high genetic risk, and would be our best candidate for sequencing if we wanted to push forward with this family.