iASeq Vignette

Yingying Wei,Hongkai Ji

October 27, 2020

1 Introduction

In diploid organisms, certain genes can be expressed, methylated or regulated in an allele-specific manner. These allele-specific events (AS) are of high interest for phenotypic diversity and disease susceptibility. Next generation sequencing technologies provide opportunities to study AS globally. However, little is known about the mechanism of AS. For instance, the patterns of allele-specific binding (ASB) across different Transcription Factors (TFs) and histone modifications (HMs) are unclear. Moreover, the limited number of reads on heterozygotic SNPs results in low-signal-to-noise ratio when calling AS. Here, we propose a Bayes hierarchical model to study ASB by jointly analyzing multiple ChIP-seq studies. The model is able to learn the patterns of ASB across studies and make substantial improvement in calling ASB. In principle, the model can also be applied to call AS for multiple RNA-seq and MeDIP-seq studies.

2 Data preparation

In order to adopt the *iASeq* model, one needs to call the function *iASeqmotif*. The first requirement, `exprs`, is the matrix containing the read counts data for heterozygotic SNPs that needs to be analyzed. Each row of the matrix corresponds to a heterozygotic SNP and each column of the matrix corresponds to the reads count for either the reference allele or non-reference allele in a replicate of a study.

The second arguement, `studyid`, identifies the group label of each column. All columns in the same study have the same `studyid`. Here we use data *sampleASE_exprs* as an illustration. *sampleASE_exprs* are combined from five studies for 5504 heterozygotic SNPs, each study having two replicates.

```
> library(iASeq)
> data(sampleASE)
> colnames(sampleASE_exprs)
```

```
 [1] "BroadH3k27ac_A"   "BroadH3k27ac_B"   "BroadH3k27acR_A"  "BroadH3k27acR_B"
 [5] "BroadH3k27me3_A"  "BroadH3k27me3_B"  "BroadH3k27me3R_A" "BroadH3k27me3R_B"
 [9] "UWH3k27me3_A"     "UWH3k27me3_B"     "UWH3k27me3R_A"    "UWH3k27me3R_B"
```

```
[13] "BroadH3k36me3_A"  "BroadH3k36me3_B"  "BroadH3k36me3R_A" "BroadH3k36me3R_B"
[17] "UWH3k36me3_A"     "UWH3k36me3_B"     "UWH3k36me3R_A"    "UWH3k36me3R_B"
```

```
> sampleASE_studyid
```

```
  BroadH3k27ac_A   BroadH3k27ac_B  BroadH3k27acR_A  BroadH3k27acR_B
               1                1                1                1
 BroadH3k27me3_A  BroadH3k27me3_B BroadH3k27me3R_A BroadH3k27me3R_B
               2                2                2                2
    UWH3k27me3_A     UWH3k27me3_B    UWH3k27me3R_A    UWH3k27me3R_B
               3                3                3                3
 BroadH3k36me3_A  BroadH3k36me3_B BroadH3k36me3R_A BroadH3k36me3R_B
               4                4                4                4
    UWH3k36me3_A     UWH3k36me3_B    UWH3k36me3R_A    UWH3k36me3R_B
               5                5                5                5
```

The third arguement, `repid`, represents the sample label for each column of `exprs` matrix. The two columns within the same replicate, one for reference allele and the other for non-reference allele, have the same `repid`. In other words, `repid` discriminates the different replicates within the same study. In *sampleASE*, `BroadH3k27ac_A, BroadH3k27ac_B, BroadH3k27acR_A,BroadH3k27acR_B` for example are two samples from the same study BroadH3k27ac:

```
> sampleASE_repid
```

```
  BroadH3k27ac_A   BroadH3k27ac_B  BroadH3k27acR_A  BroadH3k27acR_B
               1                1                2                2
 BroadH3k27me3_A  BroadH3k27me3_B BroadH3k27me3R_A BroadH3k27me3R_B
               1                1                2                2
    UWH3k27me3_A     UWH3k27me3_B    UWH3k27me3R_A    UWH3k27me3R_B
               1                1                2                2
 BroadH3k36me3_A  BroadH3k36me3_B BroadH3k36me3R_A BroadH3k36me3R_B
               1                1                2                2
    UWH3k36me3_A     UWH3k36me3_B    UWH3k36me3R_A    UWH3k36me3R_B
               1                1                2                2
```

The fourth arguement, `refid`, indicates the reference allele label for each column of `exprs` matrix. Please code 0 for reference allele columns and 1 for non-reference allele columns to make the interpretation of `over bound(or expressed in case of expression)` to be skewing to the reference allele. Otherwise, just interpret the other way round.

```
> sampleASE_refid
```

```
 BroadH3k27ac_A   BroadH3k27ac_B  BroadH3k27acR_A  BroadH3k27acR_B
              0                1                0                1
BroadH3k27me3_A  BroadH3k27me3_B BroadH3k27me3R_A BroadH3k27me3R_B
              0                1                0                1
   UWH3k27me3_A     UWH3k27me3_B    UWH3k27me3R_A    UWH3k27me3R_B
              0                1                0                1
BroadH3k36me3_A  BroadH3k36me3_B BroadH3k36me3R_A BroadH3k36me3R_B
              0                1                0                1
   UWH3k36me3_A     UWH3k36me3_B    UWH3k36me3R_A    UWH3k36me3R_B
              0                1                0                1
```

3 Model fitting

Once we have specified `studyid`, `repid` and `refid`, we are able to fit the *iASeq* model. We can fit the data with varying motif numbers and use information criterion BIC to select the best model. Here for *sampleASE*, we fit 5 models with total non-null motif patterns number varying from 1 to 5. Here is only a toy example, to get reseanable results, please run enough iterations for the EM algorithm.

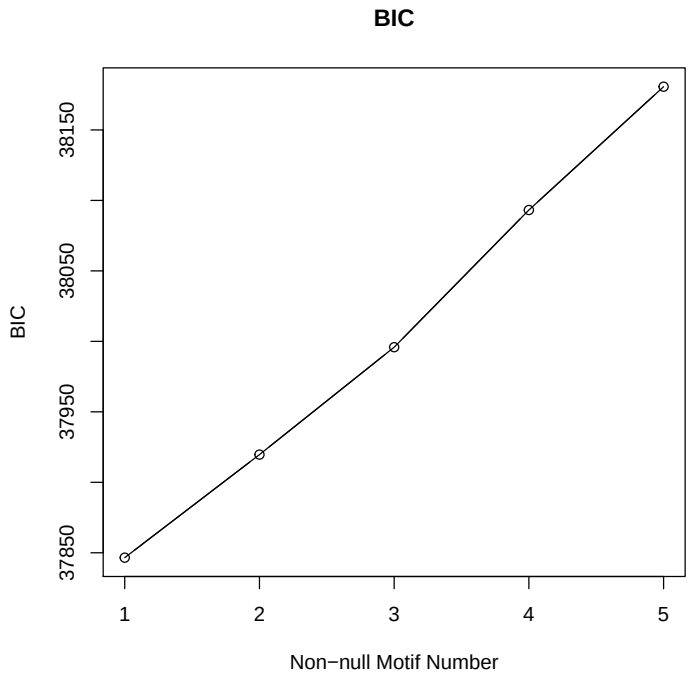
```
> motif.fitted<-iASeqmotif(sampleASE_exprs,sampleASE_studyid,sampleASE_repid,
+         sampleASE_refid,K=1:5,iter.max=5,tol=1e-3)
```

After fitting the *iASeq* model, we can check the BIC values obtained by all cluster numbers:

```
> motif.fitted$bic
```

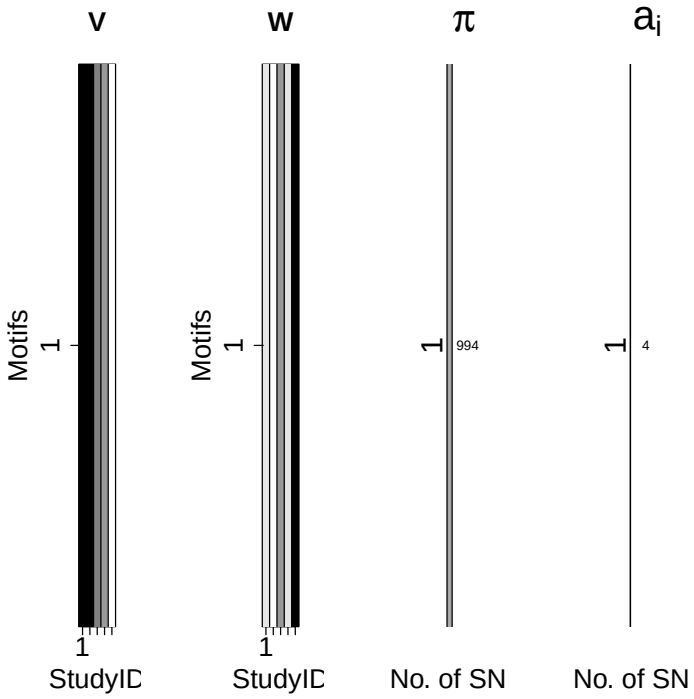
```
     K      bic
[1,] 1 37846.54
[2,] 2 37919.65
[3,] 3 37995.85
[4,] 4 38093.19
[5,] 5 38180.68
```

```
> plotBIC(motif.fitted)
>
```



To picture the motif patterns learned by the algorithm, we can use function `plotMotif`. Each row in all four graphs corresponds to the same one motif pattern. We call the left two graphs *pattern graphs* and the right two bar charts *frequency graphs*. In the pattern graphs, each row indicates a motif pattern and each column represents a study. The grey scale of the cell (k, d) demonstrates the probability of skewness to the reference allele or skewness to the non-reference allele in study d for pattern k, and the values are stored in `motif.fitted$bestmotif$motif.qup` and `motif.fitted$bestmotif$motif.qdown`. Each row of the frequency graph corresponds to the motif pattern in the same row of the left pattern graphs. The length of the bar in the first frequency graphs estimates the number of SNPs of the given pattern in the dataset according to motif frequency, which is equal to `motif.fitted$bestmotif$motif.prior`, multiplying the number of total SNPs. The length of the bar in the second bar chart shows the number of SNPs called for the given pattern according to `cutoff` of posterior probability.

```
> plotMotif(motif.fitted$bestmotif,cutoff=0.9)
>
```



The posterior probability for each SNP to be allele-specific event is stored in `bestmotif$p.post`.

```
> head(motif.fitted$bestmotif$p.post)

[1] 0.18133498 0.22025682 0.10480058 0.12419842 0.05427786 0.47551213

>
```

References

[1] Yingying Wei, Xia Li, Qianfei Wang, Hongkai Ji (2012) iASeq:integrating multiple ChIP-seq datasets for detecting allele-specific binding. *In preparation.*