Preliminary Analysis of Actual Data II

October 15, 2012

1 Description of Work Done

1. Create training and test sets. Holdout 63 conditions across genes and use 251 conditions for training.

2. Set number of reps to be the 100*(prior probability) rounded to next highest 10. Set to 1 if prior probability is 0. Creates 1-10 reps. (most TFs get 1 rep.).

3. For illustration, chose 100 genes and ran BART. Used 1000 burn-in, 2000 posterior. Tried for ntree=5 and 10. Should add 20.

4. 100 Bootstrap Iterations where extra columns are random TFs and $\mathbf{y}$ is permuted to break all dependencies.

5. Considered selection for 95^{th} quantile using simultaneous coverage and point-wise coverage. FDR is probably better.

2 Inclusions

Pointwise

- 474 genes for 5 trees.
- 506 genes for 10 trees.

Is this sensible or counter-intuitive? Seems sensible since 5 may be bottlenecking too hard.

Simultaneous

- 35 for 5 trees.
- 52 for 10 trees.

Perhaps too stringent. Really want .05 simultaneous coverage? FDR.

3 Correlation Histograms

Between Model Correlation

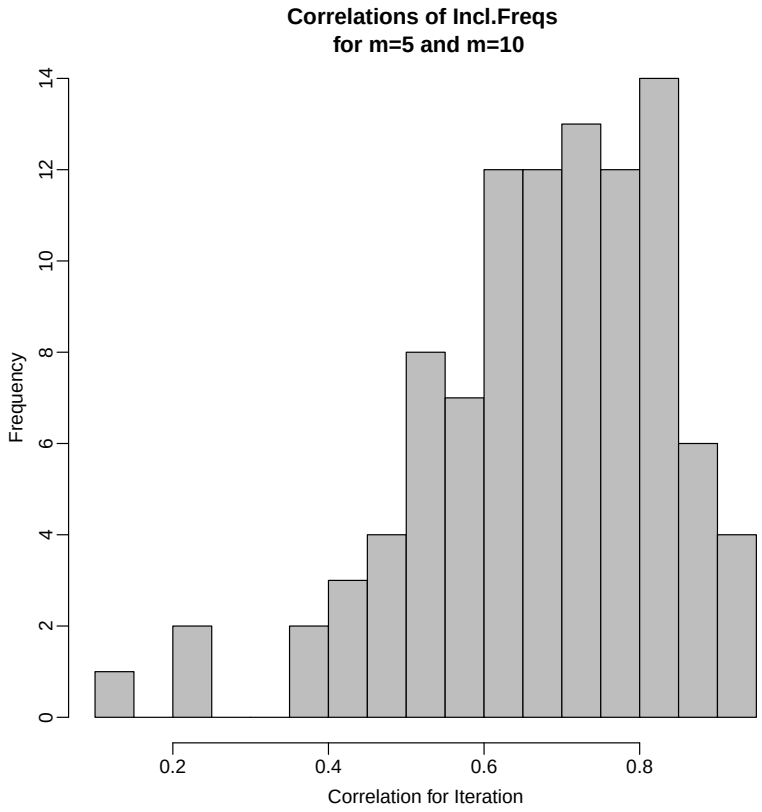

Figure 1: *Correlations of inclusion frequencies for 5 and 10 tree models*

Within Model Correlations

These are correlations between prior PROBABILITIES, not columns and TF inclusion frequency.

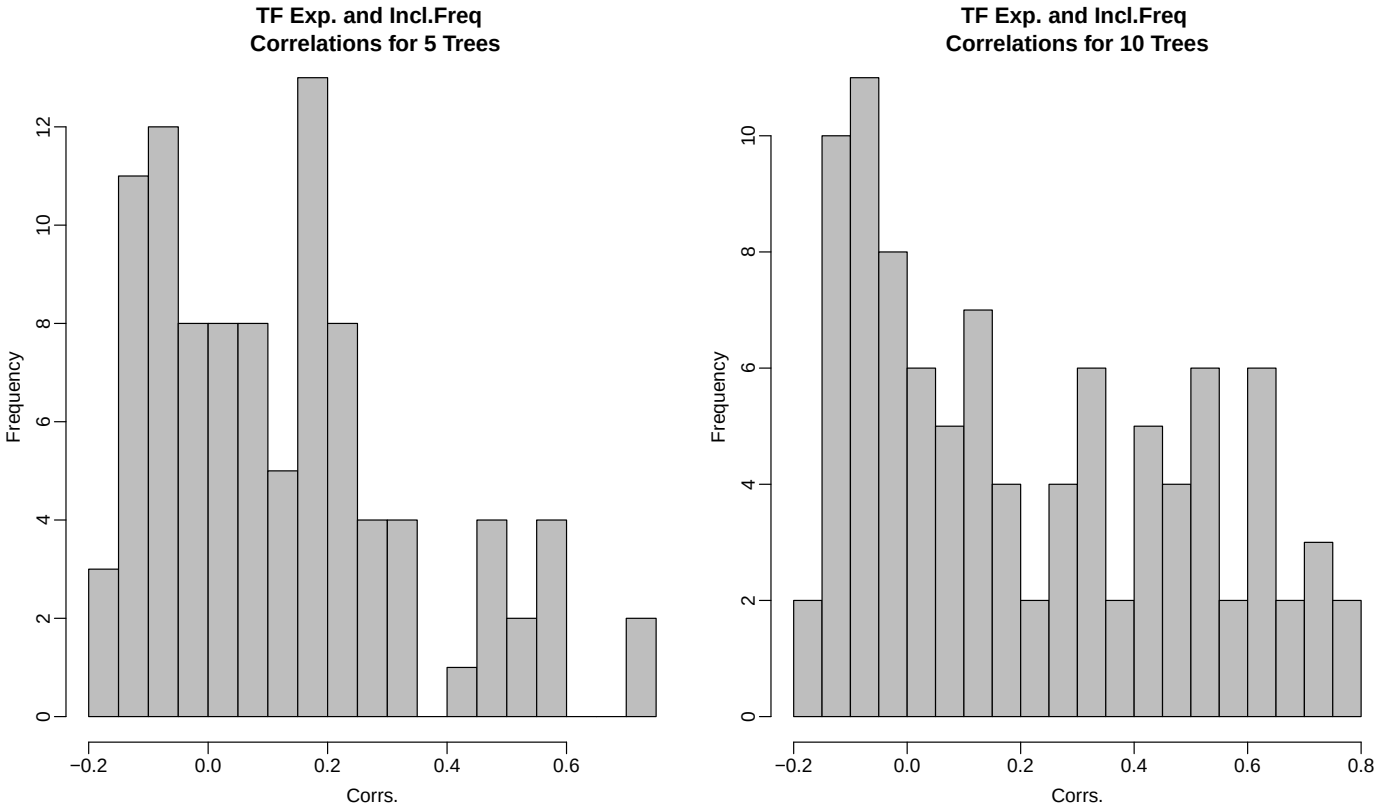

Figure 2: *Correlations between inclusion frequencies and prior probability.*

In above plot, notice that correlation goes up as number of trees goes up.

```
> summary(cor10,na.rm=T)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.15880 -0.04497  0.14450  0.21450  0.47160  0.76410
    NA's
       3
> summary(cor5,na.rm=T)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.16440 -0.05901  0.09428  0.12570  0.22480  0.74620
    NA's
       3
```

4 Likelihood vs. Prior Correlation Plots

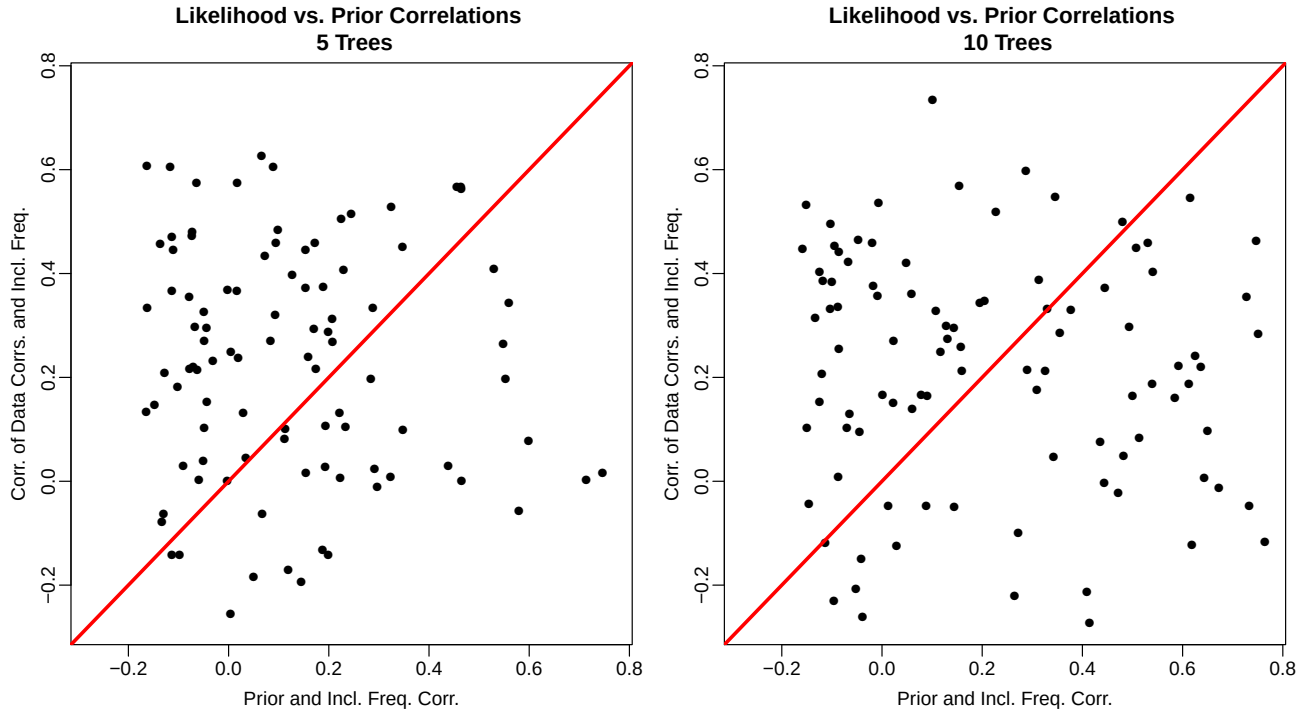

Figure 3: *Plots of correlations between correlation of gene and TF expression vs. correlations inclusion frequency and prior probability and inclusion frequency*

Due to the NAs, second plot removes any point that had a uniform prior in BART. Interesting to compare and see that the prior has more weight in this case. Sensible since these are the models where prior mattered more.

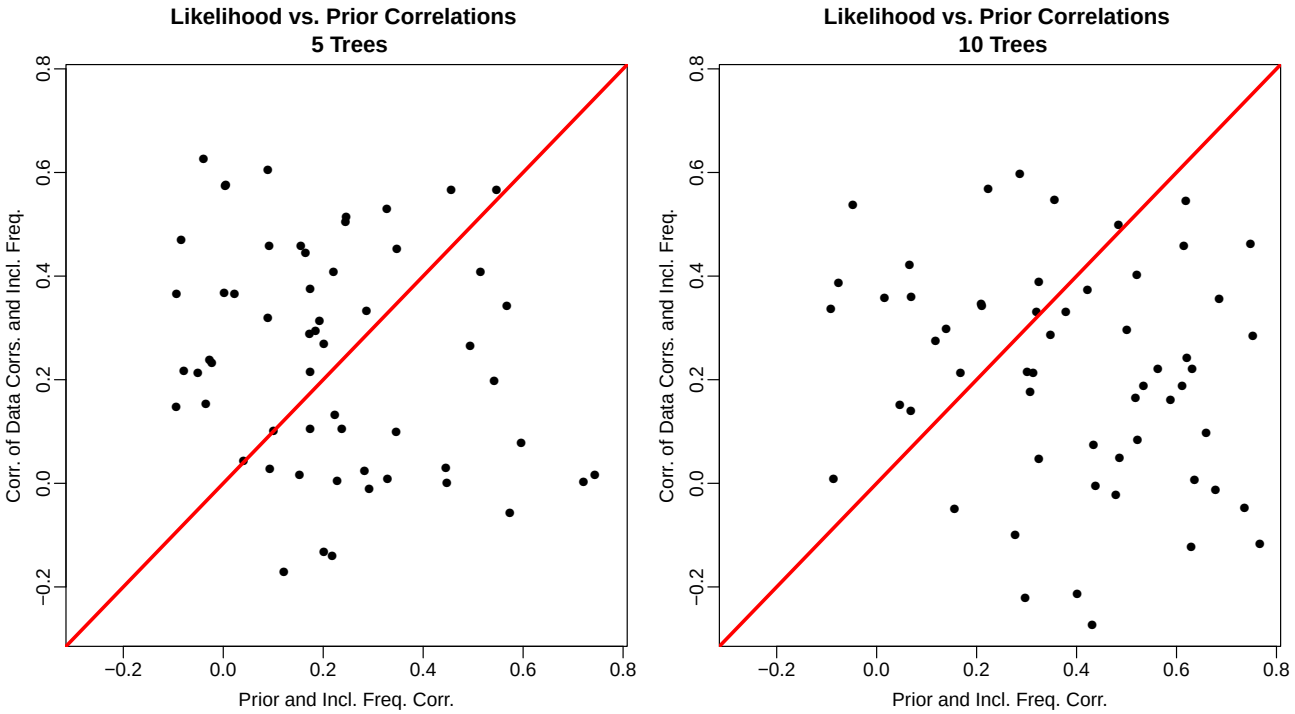

Figure 4: *Plots of correlations between correlation of gene and TF expression vs. correlations inclusion frequency and prior weights and inclusion frequency*

5 Sample Histograms of Null Distributions

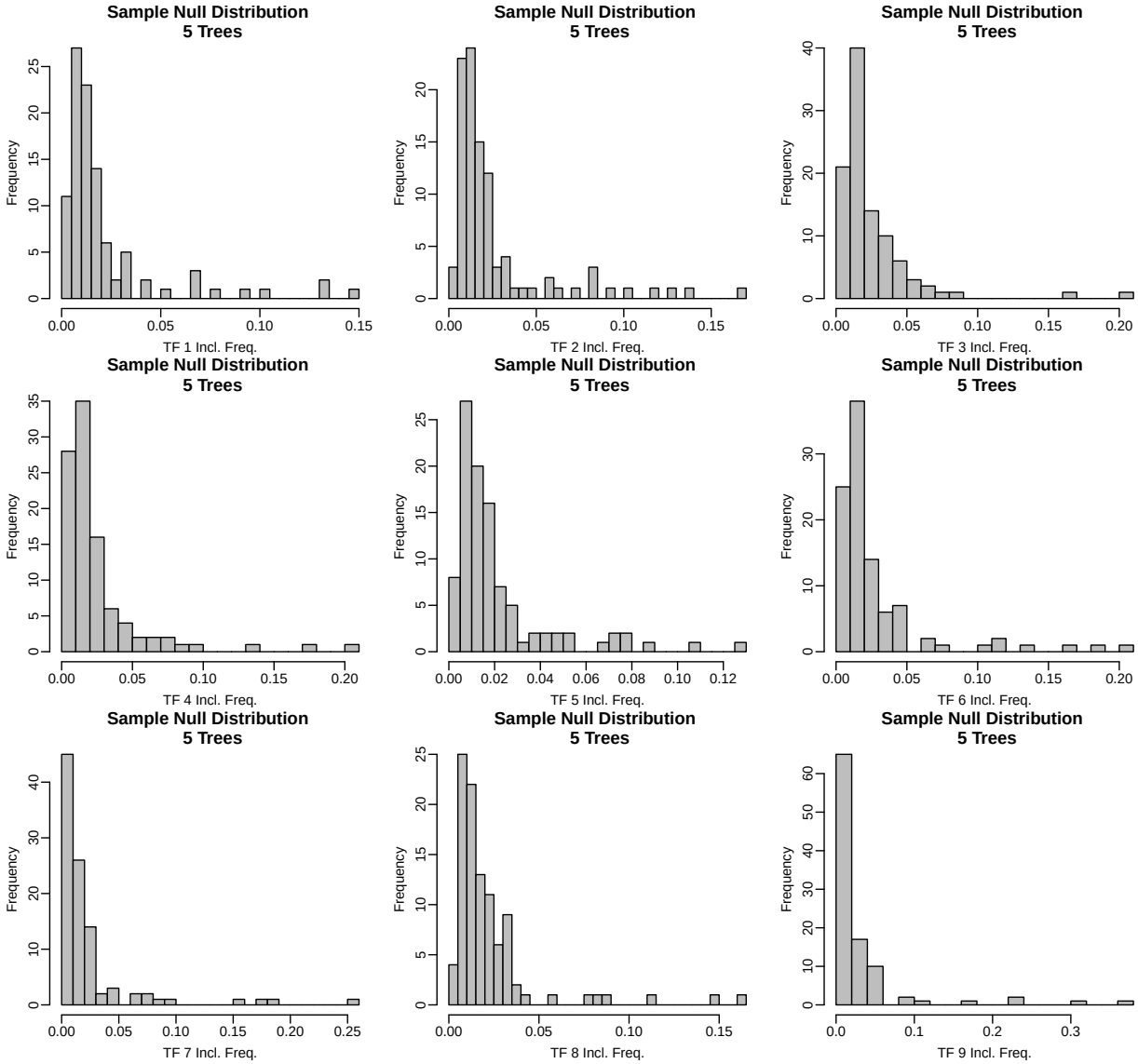

Figure 5: *Null Distributions for first 9 TFs for gene YAL001C (using ORF) in 5 tree model.*

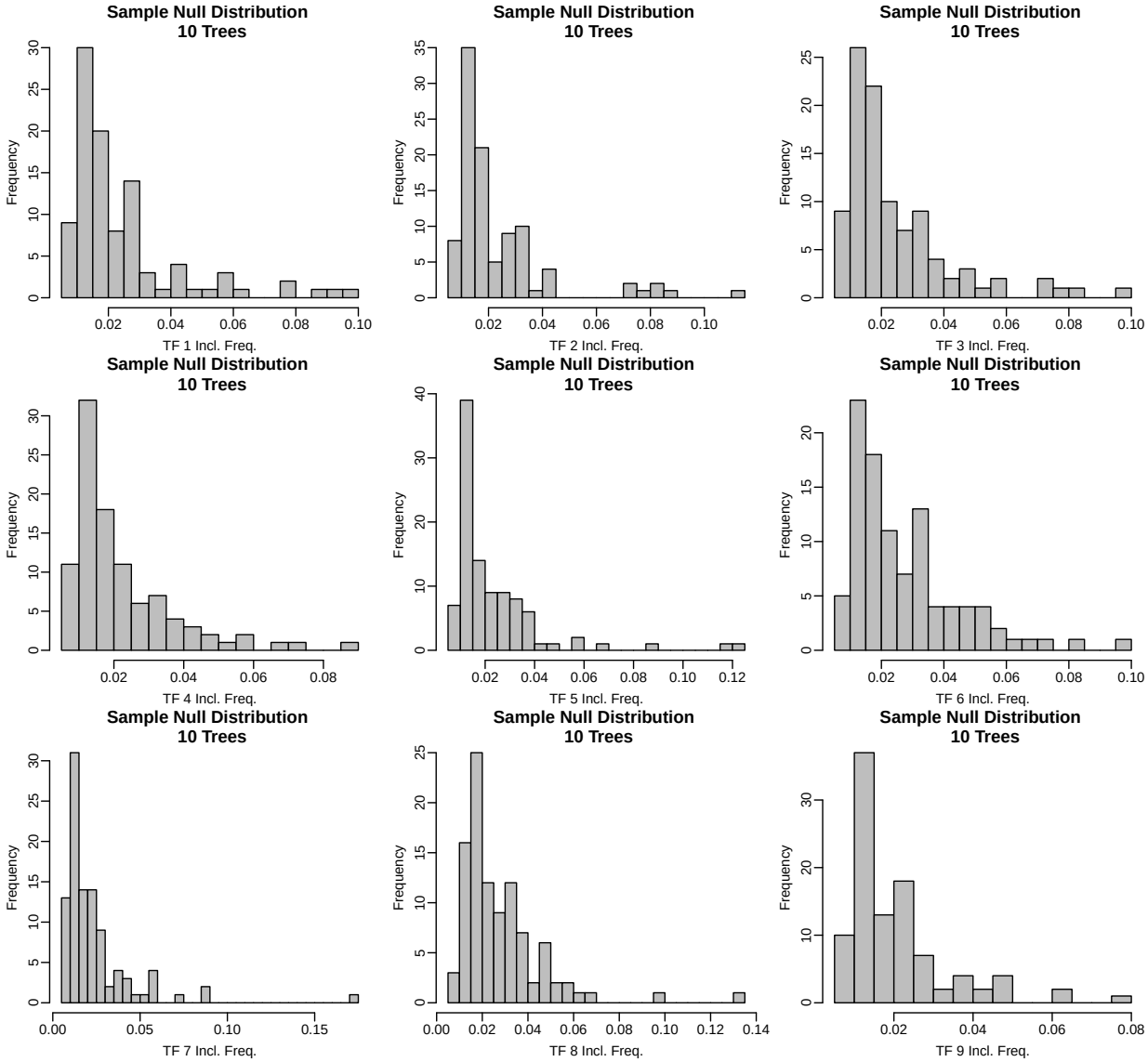

Figure 6: *Null Distributions for first 9 TFs for gene YAL001C (using ORF) in 10 tree model.*