Analysis of Differential MNase Sensitivity in diploid and allopolyploid cotton

Discussion of iseg analysis results (2018)

**Outline**

[**August discussion**](#_h4ez8wroz6x8) **1**

[1. Higher MAD and SD values were found in D-genome versus A-genome, is it artefact or true biology?](#_too3nygt6pxz) 1

[2. Considering 1, how to choose the biological cutoff for segmentation across different genomes?](#_jmrv50wvxp) 4

[3. How to handle replicates?](#_v7vo0u9cratb) 5

[4. How to conduct motif discovery analysis?](#_1nnlhzws1nn4) 6

[**September Discussion**](#_k34b0a633jsi) **7**

[9-18-2018 Meeting summary](#_wacyxcw9c62l) 7

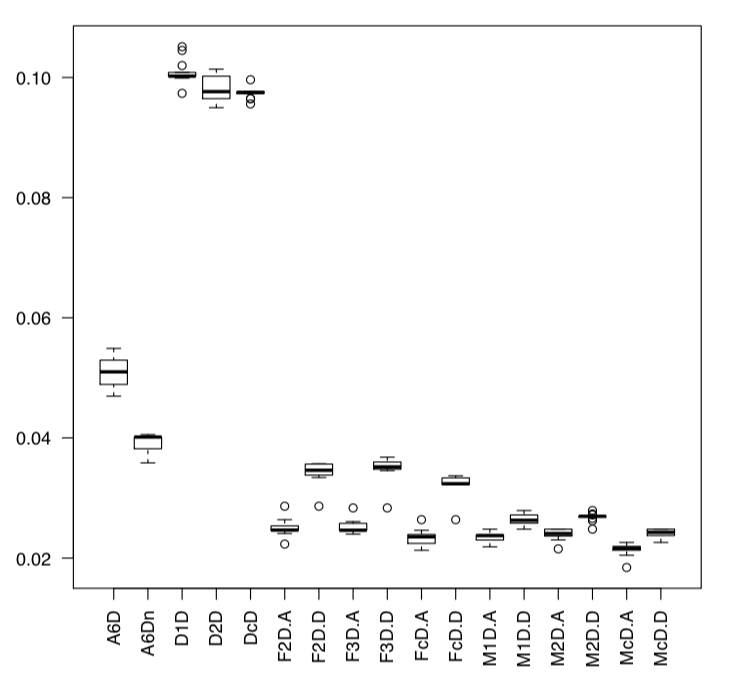
[RNA-seq analysis against individual reference transcriptomes vs. D5 with SNP index derived reference](#_yco4i6eglb4s) 8

# August discussion

## 1. Higher MAD and SD values were found in D-genome versus A-genome, is it artefact or true biology?

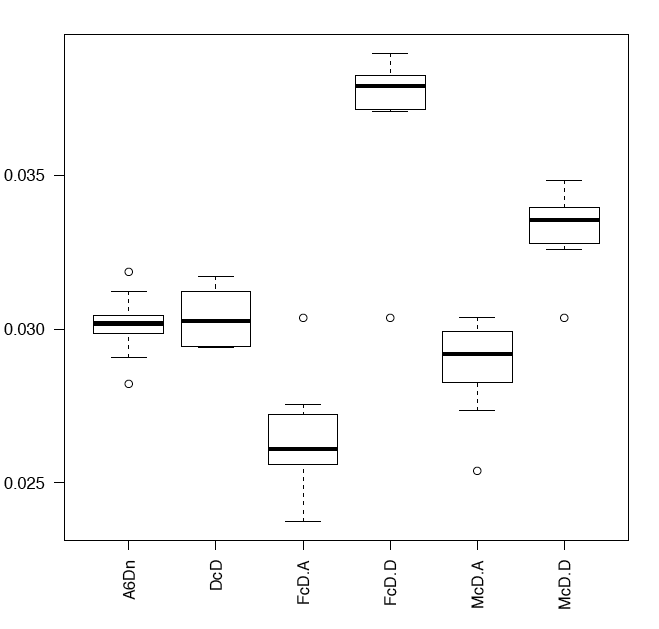
Differential nuclease sensitivity (DNS) profiles were generated by subtracting genome-wide coverages of heavily MNase digested sample from those of lightly digested samples (D= L-H), for each of four cotton genomes used: diploid A2 and D5 genome cottons, their F1 hybrid A2xD5, and the natural tetraploid cotton AD1. Prior to segmentation of MNase sensitive/resistance footprints (MSFs/MRFs), two statistical metrics were calculated for each DNS profiles (per chromosome) to measure the variability of genome-wide DNS - MAD refers to the median absolute deviation, and SD refers to standard deviation approximated by SD = 1.4826\*MAD.

As shown in Figure 1, MAD values are significantly different between species: D5 > A2 > AD1/F1. Moreover, diploid D5 exhibits a much higher MAD than the diploid A2 species, and even within the hybrid and tetraploid AD1 genomes, the D subgenomes values are higher than A subgenome values (see [Table 1](https://docs.google.com/spreadsheets/d/1OrwVc4Uh03tEHUSZxTRRS-1FtNccrszU_mFXq3VxlPA/edit?usp=sharing) for actual MAD values ).



**Figure 1**. Boxplot of per-chromosome MAD values. Two biological replicates were used for each genome: A2 - A6D, A6Dn; D5 - D1D, D2D; AD1 - M1D and M2D. For tetraploid AD1 genome, the A-subgenome (At) and D-subgenome (Dt) chromosomes were separately presented, e.g. M1D\_At and M1D\_Dt.

After quantile normalization across genomes, the difference of Dt > At remains. How will this affect segmentation results between subgenome? More segments and higher genomic coverage in Dt than At?



**Figure 2**. Boxplot of per-chromosome MAD values after quantile normalization of pooled profiles: A2 - A6Dn; D5 - DcD; AD1 - McD; F1 - FcD.

## 

**Questions:**

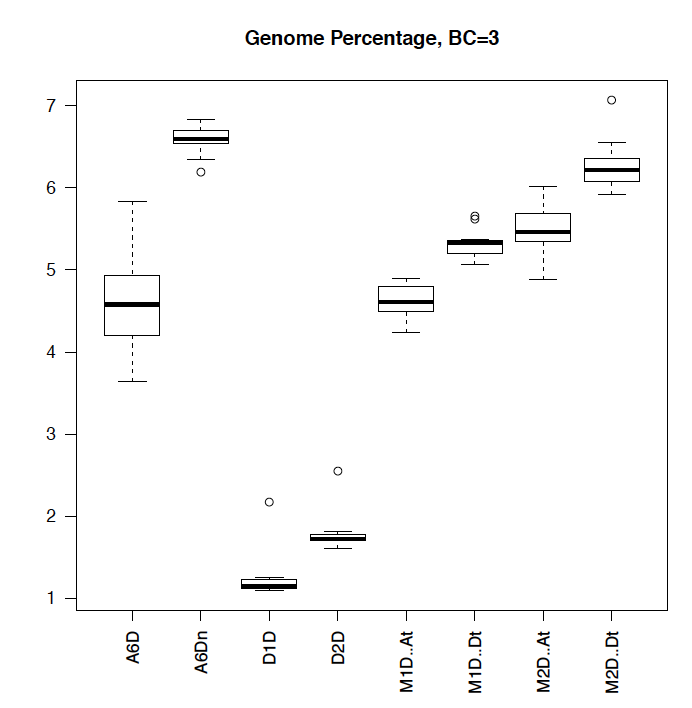
1. How to understand the inter-specific differences in MAD values? Is it possible that this pattern of MAD (D > A > M) related to genome size differences (D5 0.8G, A2 1.6Gbp, AD1 2.2G)? A statistical or technical artefact?
2. If this is true biology, does it mean the range of DNS on D-genome chromosomes (diploid and the subgenome) are more dynamic than that on A-genome chromosomes??

**Discussion Log (08-21) by Hank and Jing)**

* good question - can you send me screen shots of three homeologous gene pair regions?
  + The current observation is based on chromosome-level MAD, not gene-level yet. I will do so next and share what I find.
* and what are the units of your coverage (input bedgraphs)
  + rpm normalized 20bp coverage. Before getting the DNS profile, H and L were quantile normalized.
* we usually quantile normlize data distributions we want to compare - if you do that, does it still show that?
  + I will try that. Just to be clear, for comparisons across genomes, do you suggest qnorm all the D=L-H coverage files (will be different lines due to genome size and reference genome difference) before iSeg? In this way, we have to assume the unequal variations across genomes are technical, not biological; is it correct?
* an do you see it in D vs. A in both diploids and polyploids? - if so, maybe biological - the polyploids are internally controlled for a lot of the technical artifcacts -but mapping biases & genome assembly/quality differences
  + Yes this is in polyploid too, see Figure 1. About mapping biases, the A-subgenome is much bigger, has more repetitive sequence regions, can this lead to the smaller dynamic range of DNS on At versus Dt? My guess it that the more repetitive regions in At are not uniquely mappable, even we take only the top alignment reported per read, the alignment may not be true; then considering this for both H and L profiles, their differences become averaged out, or flattened with smaller dynamic range than Dt. I wondering if I can prove this by comparing MDS for only gene regions, which would have comparable dynamic ranges if my guess is correct. This also goes back to your comment on screen shots of homoeolog gene pair regions.

## 2. Considering 1, how to choose the biological cutoff for segmentation across different genomes?

During segmentation, biological cutoffs (BCs) were used to extract significant signals above certain threshold from baseline. The actual threshold of segment DNS is BC\*SD. Due to the differences in SD values across genomes, the choice of same BC in fact applies different actual threshold: D > A > M. For example with BC=3,1.2%-1.8% of D5 genome were detected as significant MSFs/MRFs, 4.6-6.6% were found for A2 genome, and 4.88-5.78% were found for AD1 genome ([Table 3](https://drive.google.com/open?id=1OrwVc4Uh03tEHUSZxTRRS-1FtNccrszU_mFXq3VxlPA) ).



**Questions:**

The goal of this project is to compare chromosome features across genomes. I can think of different ways to choose BC values, but not sure what is the most reasonable:

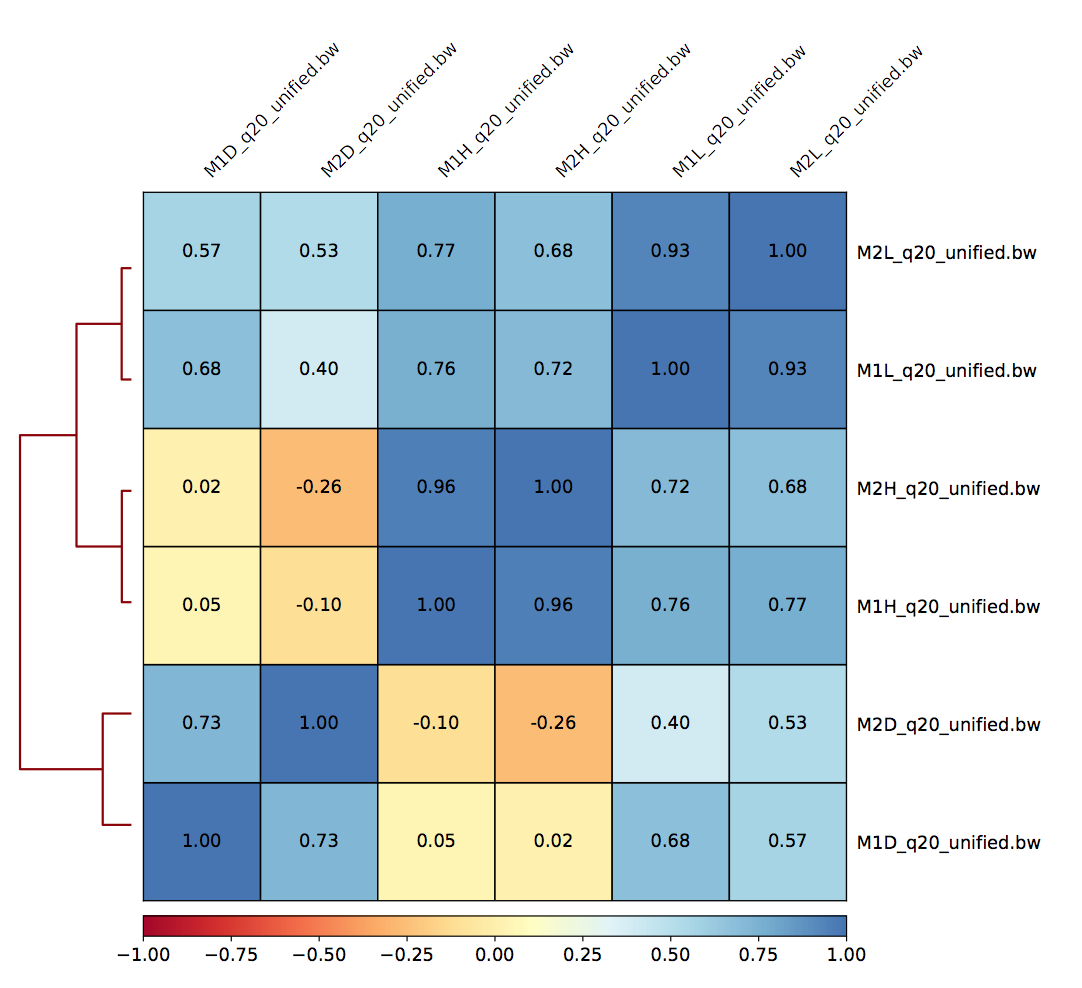
* Using the same BC values as above - fewer segments and genomic regions were detected in D5 than other genome, does it reflect true biology?
* Choose the BC values resulting in similar genome percentages of significant regions

Hank: another good question - it is a significance value, but BC2 doesn't always = BC2 for datasets with different structures ... so we do a series (BC0.5, 1.0, 2.0, 3.0, 5.0, 7.0, 9.0) for example, then to compare - we like to use "whatever BC calls the a similar proportion of the data" in the example below, BC=5.0 calls ~ 1% of the data - so if you want to ask about the top 1% of peaks ..., use BC=5.0 for this data, but the 1% may be in BC=3.0 in another dataset.

* Understood! So in this way, our downstream comparative analysis is structured as "how do the top 1% significant segments look like or differ between species", rather that "how are the percentages of significant regions different between species."
* After pooling replicates and then quantile normalization across genome, BC=6.0 calls 1.6% of genomic regions as significant segments in all genomes.

## 3. How to handle replicates?

Two biological replicates were used for each genome. To ensure the reproducibility of assays, I used the deeptools2 to plot sample correlation among H, L and D profiles. Pearson correlations above 0.9 were observed between replicates.



But how should we handle replicates for following analyses including segmentation? Some options are as below. Suggestions?

1. Combine mapping results (BAM) between replicates prior to iSeg analysis, thus only one pooled sample per genome. This is the simplest way, but not really benefit from the duplicated experimental design.
2. After iSeg analysis of each replicates, identify consensus regions between segment lists sharing a minimum absolute overlap or minimum relative overlap, using bedtools or findOverlaps in R. I gave this a try, and found less than 20% of segments share overlaps between replicates at BC=3, over 50% at BC=1, and 30-40% at BC=2.
3. Also after individual iSeg runs, combine replicates by only using highly reproducible peaks using the IDR (irreproducibility Discovery Rate) method, which is available for ChIP-seq dataset (used by [ENCODE](https://sites.google.com/site/anshulkundaje/projects/idr)). The key is avoiding initial cutoffs of peak calling, but using ranked lists of all possible peaks. Quite complicated, not sure if it is suitable here.

if the replicates pass basic QC and similar fragment size distribution & other general features, we pool them - but often we iSeg reps & combos & inspect all - if nothing looks weird, we go with the combo.

so for combinations, you may need to quantile normalize the combos to each other, then iSeg the combos.

* Done!

## 4. How to conduct motif discovery analysis?

I would like to ask what motifs are enriched on MSF and MRF regions, and whether there are motifs differentially enriched between species, and between subgenomes within the tetraploid.

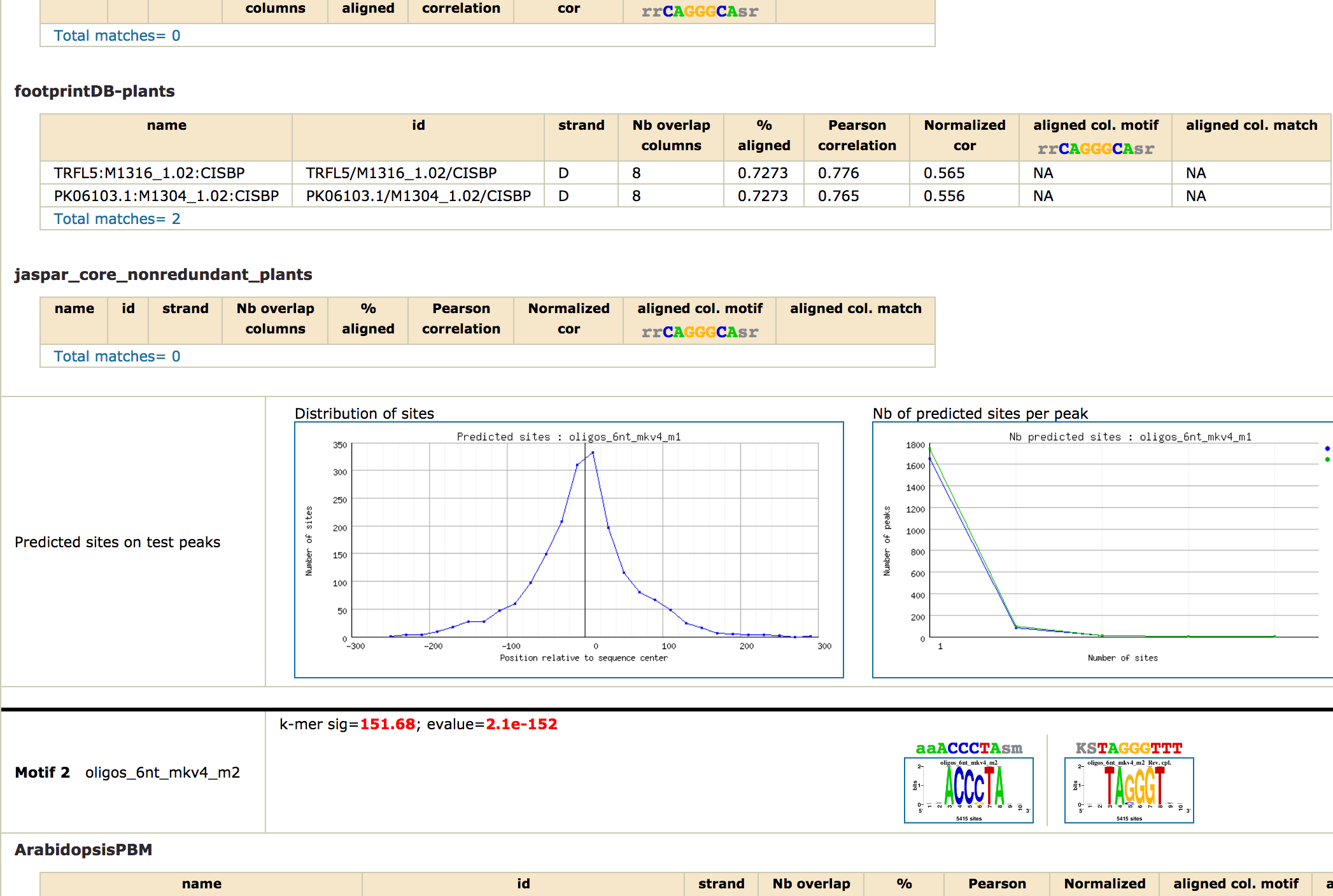
Any suggestions on which program to use? RSAT, HOMMER, MEME?

I have not done this, but a collaborator just did this with our small fragment data (MOA-seq) - you can take all the fragments < 130bp and iSeg them, then look as fasta seqs under the peaks (the iSeg that calls top 1%, for ex) and submit those to RSAT.

You can also use the MSF peaks - for those, I'd use the BC that calls the top 3% of the data - those will be larger peaks than the small fragment coverage, but at least you already have those iSeg peak call bed files

RSAT can be loaded up with multiple motif DBs for comparison and do a de novo discovery from the raw data -

here's a screenshot of the very nice output RSAT makes:



# September Discussion

## 9-18-2018 Meeting summary

1. Examine fragment size distribution and sequence coverage for each species

* Heavy digestion fragments show prominent secondary peaks around 135 bp in addition to the main peaks around 147bp - what are those small fragments? are they derived from different genomic regions in comparison with 147bp fragments? any biological meanings? **TO-DO: partition 135bp fragments, examine proportion and genomic location.**
* one A2 technical replicate (H&L) is problematic, will be excluded from downstream analysis
* technical replicates show good reproducibilities according to PCA and clustering analysis, except A2, therefore ok to pool.

2. Examine nucleosome occupancy (H&L) and differential sensitivity profiles at genes, in association with expression quantiles

* -Nuclease sensitivity (D) is positively correlated with expression quantiles.
* Nucleosomal occupancy (H&L) is anti-correlated with gene expression, except for L in tetraploid AD1. What is the cause of the exception of L in AD1? One possibility is that smaller fragments (0-130bp) tend to display positive correlation between L and expression. Is it possible AD1 L have a higher portion of smaller fragments, thereby displaying different patterns? **TO-DO: run size partition for all genomes, repeat analysis and examine quantile patterns.**

3. Discuss the cause of different MAD between diploid and between subgenomes within F1 and AD1.

* MAD Dt>At persists in AD1, even after quantile normalization across genomes, can rule out the reference genome quality difference between sub genome.
* The genome size difference between A and D genome, thereby different uniquely mapped region portions, is most likely the cause. **TO-DO: to account for the 2-fold genome size difference in A vs D, reduce full genome reference to gene-centric reference (gene body plus 3kb up- and down-stream), re-calculate D profiles, see if the AvsD differences in MAD disappear.**

4. Examine genomic locations of iSeg detected segments

1. MRFs are mostly located in distal intergenic regions.
2. In compared to MRFs, MSFs are more likely to be detected near genes, and more >1kb promoter MSFs were found in AD1 than in diploid or hybrid, considering both peak number and peak region (bp).
3. the presence and absence of MSFs in promoter do NOT seem to correlate with transcriptomic gene expression patterns. **TO-DO: instead of binary data of MSF presence/absence, use the 300bp D profile right before TSS to extract quantitative data (such as mean coverage, under peak area, etc.), and then study their relationships with gene expression patterns.**

## RNA-seq analysis against individual reference transcriptomes vs. D5 with SNP index derived reference

Expression datasets by mapped against Individual transcriptomes resulted into weird DE resulted for A2vsD5 and F1.AtvsF1vDt, while normal for AD1.AtvsAD1.Dt. **TO-DO: try mapping against D5 with SNP index derived reference.**