White Paper: Detect Regions of Significance

Detect Regions of Significance Overview

The MAT algorithm (Johnson et al., 2006) is used to find regions of binding in ChIP-chip experiments for a single, or multiple samples. This was done by estimating t-statistics for each probe by using a linear model fit on a subset of probe intensities taking into consideration multiple factors such as GC content and the number of times a sequence maps to the genome. These probe-level t-statistics are used to generate MAT scores using the trimmed mean of probe-level t-statistics in a window of fixed genomic length. An empirical distribution is used to determine MAT score significance by sampling windows from the original data. After identifying regions of a specified target length as significant, they were the combined with other close regions.

To make the method more flexible and able to handle multiple factors and contrasts, Partek uses ANOVA contrast t-statistics on each probe, then identifies regions of significance using a method based on the methodology above. The main difference from those in MAT is the empirical distribution is estimated by sampling non-overlapping windows of permuted (rather than original) data.

Using the Detect Regions of Significance Dialog

- After creating an ANOVA probe level result with t-statistics added for the contrast of interest, select **Detect Regions of Significance** from the *Tiling* workflow in *Analysis* section. The *Detect Regions of Significance* dialog will appear (Figure 1)

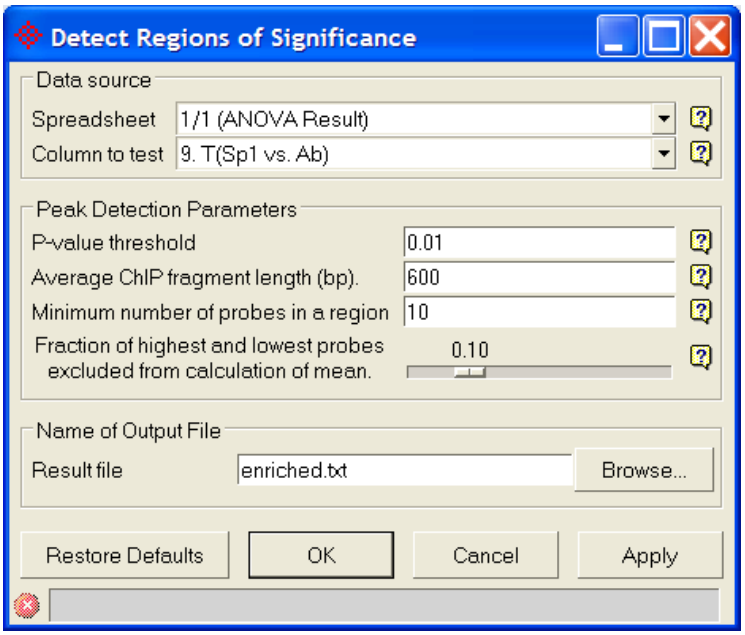

Figure 1: Configuring the Detect Regions of Significance dialog

Descriptions for the options contained in the *Detect Regions of Significance* dialog are shown below:

- *Spreadsheet*: the previously created ANOVA result containing the contrast of interest
- *Column to test*: the column containing the contrast's t-statistics
- *P-value threshold*: (Default 0.001) this p-value cut off is used when determining if a region's MAT score is significant.
- *Average ChIP fragment length (bp)*: (Default 600 bp) the length of an expected ChIP region
- *Minimum number of probes in a region*: (Default 10) excluding windows with very low probe coverage from consideration can improve the specificity of the results
- *Fraction of highest and lowest probes excluded from calculation of the mean*: (Default 0.1) this represents the fraction of extreme t-statistics that will be excluded from each region when calculating the MAT score. For example, using a value of 0.1 will exclude the upper and lower 10% of the data, using the central 80% to calculate the MAT score for the region.

Descriptions of the resulting spreadsheet columns:

- *Chromosome*: the chromosome of the detected enrichment region
- *Start*: the position in base pairs of the first base of the region
- *Stop*: the position of the last base included in the detected region
- *length(bps)*: the length of the detected region in base pairs
- *probes in region*: the number of probes included in this region

- *p-value(region)*: the empirical p-value of the most significant window contained in the region
- *Fraction of negatively enriched:* represents the proportion of false positive probes included in this region. This is calculated as the # probes not significant / # probes in reported region. Regions with a high value may be less confident or only caused by a large number of outliers within the data rather than a true discovery
- *MAT-score*: the maximum MAT score for this region. A positive value means the trimmed mean of t-statistics from the specified contrast was positive, negative scores result from a negative trimmed mean of t-statistics

References

Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M and Liu XS. Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. USA* 103 (2006) 12457-12462.