MAT

Shiksha Mantri
Mentor: Dr. Bino John
Model-based Analysis of Tiling-arrays for ChIP-chip

X. Shirley Liu et al., PNAS (2006) vol. 103 no. 33 12457–12462
Tiling Arrays

• Subtype of microarray chips

• Short fragments are designed to cover the entire genome

• Affymetrix tiling arrays have developed oligonucleotide arrays that tile all of the nonrepetitive genomic sequences of human and other eukaryotes.
NEED of effective algorithms...

- Chromatin immunoprecipitation (ChIP) coupled with DNA microarray analysis (ChIP-chip) studies have become very popular.

- unbiased genome-wide ChIP-chip experiments → massive amounts of data

- complex nature of the resulting data

- a need for effective and efficient analysis algorithms
MAT

a fast and powerful algorithm
WHY MAT?

Exciting features:

- fast and powerful analysis algorithm
- reliably detect enriched regions
- an innovative function to score regions for ChIP enrichment
- time and monetary costs minimized!
- available at http://chip.dfci.harvard.eduwliMAT

- FREE! FREE! FREE!
MAT modeling

• Baseline probe behavior modeled by:
  • considering the 25-mer sequence of each probe
  • copy number of all probes on a single tiling array

• MAT can standardize the signals of each probe in each array individually

• detect ChIP regions from:
  • a single ChIP sample,
  • multiple ChIP samples
  • multiple ChIP samples with controls

• And all this with increased accuracy!
How Does it Work?

- MAT applied to each array in a data set
- probe behavior model estimated by examining the
  - signal intensity,
  - sequence,
  - and copy number of all probes on an array.
MATh talk

\[
\log(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k \in \{A, C, G\}} \beta_{jk} I_{ijk}
\]

**Position-specific nucleotides:**

- Correlations of 0.53–0.60 between the predicted probe intensities and observed values
- A, C, G, and T contribute differently to the probe intensity
The position-specific nucleotide coefficients indicated two or three nucleotides at the two ends and in the middle of the probe tend to have the most variable effects.
more MATh talk

• Including the effect of the squared ACGT counts the correlation between model predictions and observations increased to 0.65–0.72

• 5% of the probes on the chr21 and chr22 tiling arrays mapped to multiple locations in the genome

• Therefore, their copy number effect was accounted for.

$$\log(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k \in \{A,C,G\}} \beta_{jk}I_{ijk} + \sum_{k \in \{A,C,G,T\}} \gamma_k n_{ik}^2 + \delta \log(n_i) + \varepsilon_i,$$
Probe standardization

- all of the probes on an array divided into 100 affinity bins
- each bin predicted to have similar intensities
- sample variance estimated for each bin
MAT uses the model-predicted intensity and bin variance to standardize every probe on the array according to:

$$t_i = \frac{\log(PM_i) - \hat{m}_i}{S_i \text{ affinity bin}}$$

$m^\wedge_i$: baseline intensity predicted by the model,
$s_i$: standard deviation of the affinity bin to which probe $i$ belongs

The distribution of $t$ values is approximately standard normal, and $t$ values may be compared across experiments without further normalization.
MAT score

• calculated for each window
• assigned to the probe at the center of the window

\[ \text{MATscore}(\text{region}) = \sqrt{n_p} \times TM(t \text{ values in region}). \]

\(TM\): trimmed-mean of all of the probe \(t\) values in the region
\(n_p\): number of observation points in the region used to calculate the \(TM\).

MAT scores are distributed approximately normal and allow different regions to be directly compared!
Test of Effectiveness

- Applied to the estrogen receptor (ER) ChIP-chip data covering chromosome (chr) 21 and 22.

- chip set:

  - each with 300,000 probe pairs
  - Three ChIP-chip replicates –represented as C1, C2, and C3
  - and three Input control replicates-represented as I1, I2, and I3
ChIP Region Detection

• MAT applied to detect ChIP regions in three different scenarios:
  • single sample
  • multiple samples (represented as 3C for ChIP triplicates and 3I for Input triplicates)
  • multiple ChIP samples with input controls (represented as 3C-3I)
  • MAT scores calculated for each
Results:

MATscore of all 600-bp windows across the Affymetrix human chr21 22 tiling A array. Shown are the MATscores calculated by using ChIP sample 1 (C1) (a), Input sample 1 (I1) (b), 3C (c), 3I (d), and all six samples (3C-3I) (e). The thickness of each horizontal band reflects the variance of the MATscores.

➢ All high-scoring windows chosen
How well did MAT perform?
How well did MAT perform?

- 83 regions have been qPCR-validated as ER ChIP regions on chr21-22

- A total of 77 ChIP regions were identified from (3C-3I), of which 68 were in common with the 3C without 3I.

- Other methods and algorithms but not as efficient as MAT
Conclusions

• MAT comes in handy:

  – Predict enriched ChIP regions

  – quickly test and optimize protocols and antibodies for new ChIP-chip experiments

  – run one ChIP sample on a single array for the different protocols or antibodies and use MAT to make predictions.

  – easily identify samples whose results are inconsistent with the other samples and with (3C-3I)
Drawback

DISCLAIMER:
Validated by many but unexplained physical basis

In search of better models…
THANK YOU!!!