Epigenomic enrichment

Mikhail Dozmorov

Spring 2018
Gene enrichment vs. genome enrichment

- **Gene set enrichment analysis** - summarizing many **genes** of interest, such as differentially expressed genes, with a few common **gene annotations** (molecular functions, canonical pathways)

- **Epigenomic enrichment analysis** - summarizing many **genomic regions** of interest, such as disease-associated genomic variants, with a few common **genome annotations** (chromatin states, transcription factor binding sites)
Genomic regions

- Gene/exon boundaries, promoters
- Single Nucleotide Polymorphisms (SNPs)
- Transcription Factor Binding Sites (TFBS)
- Differentially methylated regions
- CpG islands

Each genomic region has coordinates (unique IDs):

Chromosome, Start, End
Epigenomic (regulatory) regions - genomic regions annotated as carrying functional and/or regulatory potential

- DNasel hypersensitive sites
- Histone modification marks
- Transcription Factor Binding Sites
- DNA methylation
- Enhancers
- ...
Why “genomic region enrichment analysis”?

Enrichment = functional impact

- **Hypothesis**: SNPs in epigenomic regions may disrupt regulation
- More significant enrichment = more SNPs in epigenomic regions = more regulation is disrupted (SNP burden)
Regulatory marks are highly non-random

- Statistical analysis of pilot ENCODE regions showed highly non-random location of regulatory elements
- There are regulatory “hotspots” enriched in transcription factor binding sites, histone marks, as well as “deserts” of depleted regulatory marks
- Combinations of different types of regulatory marks matter

Statistics of epigenomic enrichments

- 6 out of 7 disease-associated SNPs overlap with epigenomic marks
- How likely this to be observed by chance? (Chi-square test/Binomial test/Permutation test)
Pearson correlation coefficient $r$: this quantity gives equal weight to co-binding (1,1) and co-non-binding (0,0). Hence, high values may not necessarily imply high levels of co-occurrence. For the above example, $r = 0.36$. 
### Statistics of epigenomic enrichments

<table>
<thead>
<tr>
<th></th>
<th>TF2 no</th>
<th>TF2 yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF1 no</td>
<td>n-k+t</td>
<td>m-t</td>
</tr>
<tr>
<td>TF1 yes</td>
<td>k-t</td>
<td>t</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>m</td>
</tr>
</tbody>
</table>

- **Hypergeometric test**: it tests for co-occurrence based on the contingency table, which can be re-written using random variables.

- Assume that the row and column sums \((m, n, k)\) are fixed. The probability of observing \(t\) is hypergeometric. The p-value for the example is
  
  \[
  p = Pr(T \geq 10 | H_0, m = 12, n = 8, k = 14) = 0.14
  \]
Statistics of epigenomic enrichments

<table>
<thead>
<tr>
<th></th>
<th>TF2 no</th>
<th>TF2 yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF1 no</td>
<td>n-k+t</td>
<td>m-t</td>
</tr>
<tr>
<td>TF1 yes</td>
<td>k-t</td>
<td>t</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>m</td>
</tr>
</tbody>
</table>

- **Chi-square test**: it tests for dependence (not co-occurrence) between TF1 and TF2, and applies to contingency tables with very large counts.
- The difference between observed and expected counts can be approximated by a chi-square distribution with one degree of freedom.

\[
D = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

where \(O_{ij}\) are the observed counts, and \(E_{ij}\) are the expected counts under the null hypothesis, and are computed under the fixed row and column sums, e.g. \(E_{22} = \frac{mk}{m+n}\)
Statistics of epigenomic enrichments

- **Poisson distribution**: it can be used to compute how likely it is for a single TF to have, say, three binding events in 1 kb with 300 events in 1 Mb. The formula is

\[ Pr(x = 3; L = 1\text{kb}; \rho = \frac{300}{1000\text{kb}}) = e^{\lambda \rho} \frac{(L\rho)^x}{x!} \]

where \( \rho \) is the binding rate per bp.

- **Fisher’s method for combining p-values**: one can calculate a p-value for each TF in a genomic region to assess whether that TF has more binding sites than expected in this region. To assess whether both TFs bind to more sites than expected, p-values can be combined using Fisher’s method

\[ P = -2 \sum_{i=1}^{n} \log p_i \]
Genomic features are nonrandomly distributed throughout the genome.

In permutation schemes, need to consider this to properly calculate observed and expected overlaps.

**Permutation test**: it tests for co-occurrence through repeatedly permuting observed enriched regions (or binding events) in one or both profiles many times.

A pre-defined co-occurrence score is calculated for each permutation.

Many permutations produce a null distribution of the co-occurrence score. One can then use this null distribution to compute a p-value for the observed co-occurrence score.
Permutation on steroids

Different randomization strategies:

i. Shuffling across the genome

ii. Shuffling within chromosomes

iii. Shuffling within additional constraints

iv. Sub-sampling within segmented regions
## Permutation strategies

<table>
<thead>
<tr>
<th>Randomization method</th>
<th>Description</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome-wide randomization</td>
<td>Shuffling one or more features unconstrained throughout the genome</td>
<td>Simple to implement. Assumes uniform distribution of features across the genome</td>
<td>Ignores chromosome-wide or local biases in the distribution</td>
</tr>
<tr>
<td>Chromosome-wide randomization</td>
<td>Shuffling one or more features unconstrained within respective chromosomes</td>
<td>Simple to implement. Accommodates chromosome-specific biases in the distribution</td>
<td>Ignores local or domain-level biases in the distribution</td>
</tr>
<tr>
<td>Randomization (dis)allowing overlaps</td>
<td>Overlap is allowed (or prohibited) among shuffled features on the genome</td>
<td>Biologically relevant in some scenarios (e.g. sites of amplification and deletions within a cancer genome cannot overlap)</td>
<td>Long run-time. Requires informed assumptions</td>
</tr>
<tr>
<td>Randomization with additional constraints</td>
<td>User-specific constraints are included in the model</td>
<td>Can accommodate case-specific biological or technical constraints</td>
<td>Long run-time. Requires informed assumptions</td>
</tr>
<tr>
<td>Randomization with fixed location model</td>
<td>Generating expected distribution by probabilistically sampling from the observed distribution</td>
<td>Biologically relevant in several scenarios (e.g. when analyzing transcription factor binding site co-occurrence)</td>
<td>Higher order organization of the features might be ignored.</td>
</tr>
<tr>
<td>Randomization with fixed locations fixed event type model</td>
<td>Shuffling location of the first feature, while keeping the location of the second feature unchanged</td>
<td>Preserves higher order structure of the second feature</td>
<td>The chromosome or domain-specific biases in the first feature are not considered</td>
</tr>
<tr>
<td>Randomization with sub-sampling accounting for genomic structure</td>
<td>Shuffling within respective segments</td>
<td>Highly powerful if correctly implemented. Segments can be generated based on sequence composition or biologically relevant assumptions</td>
<td>Potentially longer run time than others. Determining the segment boundaries is nontrivial</td>
</tr>
</tbody>
</table>
### Evaluating overlap between sets of genomic regions

**Table 1.** Methods for scoring overlapping and adjacent signals in two or more ChIP (or DamID) profiles. See text for details of these methods.

<table>
<thead>
<tr>
<th>Number of profiles under comparison</th>
<th>Accounting for spatial variability of events (Yes/No)</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two</td>
<td>No</td>
<td>Simple counting(^{17,18,33,34})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pearson correlation coefficient(^{14,35–37})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hypothesis tests based on a single score</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hypergeometric test(^{3,5,20,38})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chi-square test(^{36})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Log-linear model(^{39})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Permutation test(^{40–42})</td>
</tr>
<tr>
<td>Two</td>
<td>Yes</td>
<td>Poisson hierarchical model(^{43})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hidden Markov model(^{44})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Standard gene’(^{45})</td>
</tr>
<tr>
<td>Many</td>
<td>Yes</td>
<td>Overall assessment of co-occurrence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Permutation test(^{4,46,47})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identification of ‘co-localisation’ hotspots:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple testing based on Poisson distribution(^{46})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clustering(^{14,37,48–51})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identification of (cis)-regulatory modules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Factor regression(^{52})</td>
</tr>
</tbody>
</table>

We can look at biological significance of our peaks using Gene Ontologies (GO) terms genome annotations

- GO: Set of structured, controlled vocabularies for community use in annotating genes, gene products and sequences

- Popular tool: the Genomic Regions Enrichment of Annotations Tool (GREAT)

Binding sites are often not located in the proximal region of the gene of interest
GREAT looks beyond this proximal region
Input: BED file with regions of interest
Output: Matching GO terms for Molecular Functions, Biological Processes, Phenotypes, Diseases, etc.
GREAT: Cis-regulatory regions functions prediction

GREAT: Cis-regulatory regions functions prediction

Step 4: Perform hypergeometric test over genes

\[ N = 8 \text{ genes in genome} \]
\[ K_\pi = 3 \text{ genes in genome carry annotation } \pi \]
\[ n = 2 \text{ genes selected by proximal genomic regions} \]
\[ k_\pi = 1 \text{ gene selected carries annotation } \pi \]
\[ P = \Pr_{\text{hyper}} (k \geq 1 \mid N = 8, K = 3, n = 2) \]

Step 4: Perform binomial test over genomic regions

\[ n = 6 \text{ total genomic regions} \]
\[ p_\pi = 0.6 \text{ fraction of genome annotated with } \pi \]
\[ k_\pi = 5 \text{ genomic regions hit annotation } \pi \]
\[ P = \Pr_{\text{binom}} (k \geq 5 \mid n = 6, p = 0.6) \]