Motif Searching

Simon Andrews
simon.andrews@babraham.ac.uk
@simon_andrews

v1.0
Rationale

Hit A
GGATCC

Hit B
GGATCC

Hit C
GGATCC

Prom A
Gene A

Prom B
Gene B

Prom C
Gene C
Basic Questions

• Does the sequence around my hits look unusual?

• Do specific sequences turn up more often than expected in my hits?

• If so, do the sequences look like any known functional sequence?

• Are there sequences which can distinguish between two or more groups of hits?
Basic Workflow

1. Hit regions
   Genes, CDS, Positions, Whatever

2. Extract Sequences

3. Check for artefacts

4. Check for enriched sequences

5. Check for composition

6. Try to identify enriched sequences
Deciding what to extract

- Hit
- Hit plus context
- Fixed width, centred on hit

Gene A

- Promoter
- Gene Body / CDS
- 3’ UTR
- 5’ UTR
Extracting Sequence

• From positions
  – BEDTools
  – Genome Browsers*
  – Custom scripts

• From features
  – Genome Browsers*
  – BioMart

*not easily automatable for multiple sequences
BioMart – Selecting Assembly

http://ensembl.org/biomart
BioMart – Specifying features
BioMart – selecting seq region
## BioMart – header info

<table>
<thead>
<tr>
<th>Header Information</th>
<th>Gene Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Start (bp)</td>
<td>Gene End (bp)</td>
</tr>
<tr>
<td>Gene type</td>
<td>Ensembl Protein ID</td>
</tr>
<tr>
<td>Ensembl Protein Family ID(s)</td>
<td>Transcript type</td>
</tr>
<tr>
<td>Chromosome Name</td>
<td>Strand</td>
</tr>
<tr>
<td>Associated Gene Name</td>
<td>Transcript Start (bp)</td>
</tr>
<tr>
<td>Associated Gene Source</td>
<td>Transcript End (bp)</td>
</tr>
<tr>
<td>5' UTR Start</td>
<td>Transcript Start Site (TSS)</td>
</tr>
<tr>
<td>5' UTR End</td>
<td>Transcript length</td>
</tr>
<tr>
<td>3' UTR Start</td>
<td></td>
</tr>
<tr>
<td>3' UTR End</td>
<td></td>
</tr>
<tr>
<td>Ensembl Transcript ID</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transctipt Information</th>
<th>Transcript Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS start (within cDNA)</td>
<td>Ensembl Protein ID</td>
</tr>
<tr>
<td>CDS end (within cDNA)</td>
<td>Transcript type</td>
</tr>
<tr>
<td>5' UTR Start</td>
<td>Strand</td>
</tr>
<tr>
<td>5' UTR End</td>
<td>Transcript Start (bp)</td>
</tr>
<tr>
<td>3' UTR Start</td>
<td>Transcript End (bp)</td>
</tr>
<tr>
<td>3' UTR End</td>
<td>Transcript Start Site (TSS)</td>
</tr>
<tr>
<td>Ensembl Transcipt ID</td>
<td>Transcript length</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exon Information</th>
<th>Transcript Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS Length</td>
<td>Ensembl Protein ID</td>
</tr>
<tr>
<td>CDS Start</td>
<td>Transcript type</td>
</tr>
<tr>
<td>CDS End</td>
<td>Strand</td>
</tr>
<tr>
<td>Ensembl Exon ID</td>
<td>Transcript Start (bp)</td>
</tr>
<tr>
<td>Exon Chr Start (bp)</td>
<td>Transcript End (bp)</td>
</tr>
<tr>
<td>Exon Chr End (bp)</td>
<td>Transcript Start Site (TSS)</td>
</tr>
<tr>
<td>Strand</td>
<td>Transcript length</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exon Information</th>
<th>Transcript Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon Rank in Transcript</td>
<td>Ensembl Protein ID</td>
</tr>
<tr>
<td>phase</td>
<td>Transcript type</td>
</tr>
<tr>
<td>cDNA coding start</td>
<td>Strand</td>
</tr>
<tr>
<td>cDNA coding end</td>
<td>Transcript Start (bp)</td>
</tr>
<tr>
<td>Genomic coding start</td>
<td>Transcript End (bp)</td>
</tr>
<tr>
<td>Genomic coding end</td>
<td>Transcript Start Site (TSS)</td>
</tr>
<tr>
<td>Constitutive Exon</td>
<td>Transcript length</td>
</tr>
</tbody>
</table>
BioMart - exporting
Deciding on a comparison

Choosing the appropriate comparison is the hardest part!
Filtering list of hits

- High specificity
- Quick run times
- Potentially lower power
- Highest hit artefacts

- More power
- Long run times
- More noise

- Don’t need all hits to generate motif
- Often better to have a clean sequence set
- Remove sequences which look unusual
Artefacts

- Exclude common repeats
  - Simple repeats (poly-A, SerThr repeats etc)
  - Complex repeats (retroviral etc)
  - Exclude hits with repeats
  - Repeatmasked sequence

- Check composition
  - Analyse compositionally biased regions explicitly
Software

- The MEME Suite: meme-suite.org
- XXmotif: xxmotif.genzentrum.lmu.de/
- CisFinder: lgsun.grc.nia.nih.gov/CisFinder/
- CREAD: cb.utdallas.edu/cread/
- HOMER: homer.salk.edu/homer/motif/
MEME Suite
MEME Motif Discovery

• MEME
  – Original motif enrichment program
  – PWM based motifs
  – Long ungapped motifs, sensitive search, slow!

• DREME
  – Short ungapped discriminatory motifs
  – Degeneracy based motifs
  – Quick!

• GLAM2
  – Gapped motifs
Main Parameters:
- Sequences (multi-fasta)
- Expected sites
- How many motifs to find

Advanced
- Custom background
- Negative set
- Motif size restriction

NB:
Query size limited to 60kb

Local installations don’t have this limit
Good Result
Good Result - Motif
For ‘peak’ data, expect motifs to be roughly centred
For promoter data there may be no pattern.
Artefactual Result - Composition

MEME tends to favour long compositionally biased motifs
Real motifs can be further down the list
Artefactual Result - Duplication

Multiple transcripts with the same promoter
Overlapping regions
AME – Known motif search

- Quicker / easier than de-novo discovery
- Limited to characterised binding sites
- Can choose from common motif sources
- Good place to start
AME Result

- No additional detail
- Could check for positional bias with CentriMo

Beware similar motifs from different factors
Discriminatory Motifs

MEME can run in discriminatory mode
DREME is designed for this specifically
DREME discovers short, ungapped motifs (recurring, fixed-length patterns) that are relatively enriched in your nucleotide sequences compared with shuffled sequences or your control sequences (sample output from sequences). See this Manual or this Tutorial for more information.
For further information on how to interpret these results or to get a copy of the MEME software please access [http://meme-suite.org](http://meme-suite.org).

If you use TOMTOM in your research, please cite the following paper:

### QUERY MOTIFS

<table>
<thead>
<tr>
<th>Name</th>
<th>Alt. Name</th>
<th>Preview</th>
<th>Matches</th>
<th>List</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCCTCTAA</td>
<td>DREME</td>
<td><img src="image" alt="GCCTCTAA" /></td>
<td>2</td>
<td>MA0503.1 (Nico2.5), MA0122.1 (Nico2-2), MA0504.1 (NR2C2)</td>
</tr>
</tbody>
</table>

### TARGET DATABASES

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of Motifs</th>
<th>Motifs Matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>JASPAR_CORE_2014_vertebrates.meme</td>
<td>205</td>
<td>3</td>
</tr>
</tbody>
</table>

### MATCHES TO QUERY MOTIF GCCTCTAA (DREME)

<table>
<thead>
<tr>
<th>Name</th>
<th>MA0503.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alt. Name</td>
<td>Nico2-5</td>
</tr>
<tr>
<td>Database</td>
<td>JASPAR_CORE_2014_vertebrates.meme</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0152266</td>
</tr>
<tr>
<td>E-value</td>
<td>3.12146</td>
</tr>
<tr>
<td>q-value</td>
<td>1</td>
</tr>
<tr>
<td>Overlap</td>
<td>0</td>
</tr>
<tr>
<td>Offset</td>
<td>1</td>
</tr>
<tr>
<td>Orientation</td>
<td>Normal</td>
</tr>
</tbody>
</table>