Searching for Transcription Start Sites in *Drosophila*

Wilson Leung 08/2021

Outline
- Transcription start sites (TSS) annotation goals
- Promoter architecture in *D. melanogaster*
- Define TSS positions and TSS search regions in *D. ananassae*

GEP Annotation Projects
- **F Element Projects**
  - Informant Species
  - GEP Publications
  - Motif Project
  - Expanded F Project
- **The Pathways Project**
  - annotates genes from 27 *Drosophila* species
- **The Parasitoid Wasp Project**
  - annotates genes from 4 wasp species

Research goals for the *Drosophila* Muller F element motif project
- **Identify conserved, regulatory motifs** that enable F element genes to be expressed in a heterochromatic environment

Use comparative genomics to identify **factors that enable gene expression** on the expanded F element

**D. ananassae** and **D. melanogaster** F element genes show **similar range of expression levels**

\[ \text{Expression levels (rlog)} \]

LOESS Regression Line

CAI (Codon Bias)
Transcription factor binding sites (motifs) are short and degenerate

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>6</td>
<td>12</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>13</td>
<td>5</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>C</td>
<td>22</td>
<td>23</td>
<td>6</td>
<td>64</td>
<td>0</td>
<td>70</td>
<td>0</td>
<td>54</td>
<td>13</td>
<td>33</td>
</tr>
<tr>
<td>G</td>
<td>9</td>
<td>11</td>
<td>30</td>
<td>3</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>T</td>
<td>36</td>
<td>31</td>
<td>23</td>
<td>4</td>
<td>54</td>
<td>0</td>
<td>56</td>
<td>5</td>
<td>40</td>
<td>19</td>
</tr>
</tbody>
</table>

Futility theorem: difficult to find functional motif instances in the entire genome

- Proposed by Wasserman WW and Sandelin A
- Nat Rev Genet. 2004 Apr;5(4):276-87

- Most motif instances in the genome are false positives
  - Multiple testing correction reduce power to detect biologically interesting motifs as the search space increase

- Does it make sense to scan an entire genome with transcription factor motifs using FIMO?
  - This is probably a bad idea for several reasons, and it may be preferable to focus on shorter regions such as promoters, enhancers or ChIP-seq peak regions. Also, if you have them, FIMO can make use of additional pattern if available, like DNase hypersensitivity, to help distinguish chance matches from the biologically significant ones. More information about that is available in the FIMO documentation about using position-specific priors.


TSS of F element genes show lower levels of H3K9me3 and HP1a

Three strategies for motif finding

- Multiple genes in a single species
  - Genes with common expression pattern
  - Sequences associated with ChIP-Seq peaks
- Single gene in multiple species
  - Phylogenetic footprinting
- Multiple genes in multiple species
  - Compare multiple sequence alignment profiles of multiple genes (Magma)

Motif finding using multiple genes within a single species

Motif finding using single gene in multiple species
Motif finding using multiple genes in multiple species (*PhyloNet*)

1. Identify conserved regions (profiles) in whole genome multiple sequence alignments
2. Identify multiple genes in the genome with similar alignment profiles

Based on Figure 1 from Wang T and Stormo GD. PNAS 2005 Nov 29;102(48):17400-5.

Promoter sequences  Conserved motifs

**Magma**: Multiple Aligner of Genomic Multiple Alignments

- Key features of *Magma*:
  - Runs ~70x faster than *PhyloNet*
  - Analyze multiple sequence alignments with gaps
  - Use set-covering approach to minimize redundancy in discovered motifs

**Drosophila** species used to identify conserved regulatory motifs on the F element

<table>
<thead>
<tr>
<th>Species</th>
<th>Substitutions per neutral site</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>D. ficusphila</em></td>
<td>0.80</td>
</tr>
<tr>
<td><em>D. eugracilis</em></td>
<td>0.76</td>
</tr>
<tr>
<td><em>D. biarmipes</em></td>
<td>0.70</td>
</tr>
<tr>
<td><em>D. takahashii</em></td>
<td>0.65</td>
</tr>
<tr>
<td><em>D. elegans</em></td>
<td>0.72</td>
</tr>
<tr>
<td><em>D. hoogstra</em></td>
<td>0.66</td>
</tr>
<tr>
<td><em>D. kikkawai</em></td>
<td>0.89</td>
</tr>
<tr>
<td><em>D. bipectinata</em></td>
<td>0.99</td>
</tr>
</tbody>
</table>

Data from Table 1 of the modENCODE comparative genomics white paper

TSS annotation goals for the F element project

- Research goal:
  - Identify motifs that enable Muller F element genes to function within a heterochromatic environment

- Annotation goals:
  - Define search regions enriched in regulatory motifs
  - Define precise location of TSS if possible
  - Define search regions where TSS could be found
  - Document the evidence used to support the TSS annotations

Strategies for utilizing the GEP TSS annotations

- Use TSS annotations to anchor the whole-genome multiple sequence alignments
- Multiple sequence alignment is hard (NP-hard)
- Example: anchored multiple alignment with DIALIGN

- Align orthologous promoters to identify conserved regulatory motifs:
  - Examples: EvoPrinterHD, Pro-Coffee

- De novo and discriminatory motif discovery
  - Analyze TSS search regions to identify over-represented motifs
  - Examples: MEME Suite, HOMER

TSS annotation workflow

1. Identify the ortholog
2. Note the gene structure in *D. melanogaster*
3. Annotate the coding exons
4. Classify the type of core promoter in *D. melanogaster*
5. Annotate the initial transcribed exon
6. Define TSS positions or TSS search regions
RNA Polymerase II core promoter

- Initiator motif (Inr) contains the TSS
- TFIID binds to the TATA box and Inr to initiate the assembly of the pre-initiation complex (PIC)


Peaked versus broad promoters


Basal transcription initiation factors binding sites enriched in peaked D. melanogaster core promoters


Core promoter motifs can affect gene expression levels


High turnover of core promoter motifs between D. erecta and D. ananassae


Challenges with TSS annotations

- Fewer constraints on untranslated regions (UTRs)
- UTRs evolve more quickly than coding regions
- Open reading frames, compatible phases of donor and acceptor sites do not apply to UTRs
- Most genes on the expanded F element have large introns
- Low percent identity (~50-70%) between the target genome and the D. melanogaster UTRs
- Most ab initio gene finders do not predict UTRs
- Cannot use RNA-Seq data to precisely define the TSS
**N-SCAN** uses **hidden Markov models** to predict 5’ UTRs

**Twinscan**

**N-SCAN**

True positive: prediction falls within -500 to +200 bp of the annotated TSS


**RNA-Seq biases introduced by library construction**

- cDNA fragmentation
- Strong bias at the 3’ end
- RNA fragmentation
- More uniform coverage
- Miss the 5’ and 3’ ends of the transcript


**3’ bias and low coverage near the TSSs in Nanopore direct RNA sequencing data**

- Arabidopsis thaliana gene AT2G39730
- No adapter
- Adapter


**Techniques for finding TSS**

- Identify the 5’ cap at the beginning of the mRNA
- Cap Analysis of Gene Expression (CAGE)
- RNA Ligase Mediated Rapid Amplification of cDNA Ends (RLM-RACE)
- Cap-trapped Expressed Sequence Tags (5’ ESTs)

- More information on these techniques:

**Promoter architecture in Drosophila**

- Classify core promoter based on the **Shape Index (SI)**
- Determined by the distribution of CAGE and 5’ RLM-RACE reads
- Shape index is a **continuum**
- Most promoters in *D. melanogaster* contain multiple TSS
  - Median width = 162 bp
- ~70% of vertebrate genes have broad promoters


**Calculate SI based on distributions of ESTs, CAGE, and 5’ RACE reads**

- Peak
- Unclassified
- Broad

Genes with peaked promoters show stronger spatial and tissue specificity

- 46% of genes with broad promoters are expressed in all stages of embryonic development
- 19% of genes with peaked promoters are expressed in all stages


Resources for classifying the type of core promoter in *D. melanogaster*

- Only a subset of the modENCODE data are available through FlyBase
- *D. melanogaster* GEP UCSC Genome Browser [Aug. 2014 (BDGP Release 6) assembly]
- FlyBase gene annotations (release 6.40)
- modENCODE TSS (Celniker) annotations
- DNase I hypersensitive sites (DHS)
- CAGE and RAMPAGE datasets
- 9-state and 16-state chromatin models
- Transcription factor binding site (TFBS) HOT spots

9-state chromatin model

- DNaseI Hypersensitive Sites (DHS) correspond to accessible regions
- High turnover of DNase I Hypersensitive Sites (DHS) between human and mouse
- Three common types of core promoters


modENCODE TSS annotations

- Two sets of modENCODE TSS predictions
  - TSS (Celniker)
    - Most recent dataset produced by modENCODE
    - Available on the GEP UCSC Genome Browser
  - TSS (Embryonic)
    - Older dataset available from FlyBase GBrowse
- Use TSS (Celniker) dataset as the primary evidence


Classify the D. melanogaster core promoter based on (TSS) Celniker annotations and DHS positions

<table>
<thead>
<tr>
<th>TSS classification</th>
<th># Annotated TSS</th>
<th># DHS positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peaked</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Intermediate</td>
<td>≤ 1</td>
<td>&gt; 1</td>
</tr>
<tr>
<td>Broad</td>
<td>&gt; 1</td>
<td>≤ 1</td>
</tr>
<tr>
<td>Insufficient evidence</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Consider DHS positions within a 300bp window surrounding the start of the D. melanogaster transcript

Additional DHS data from different stages of embryonic development

- DHS data produced by the BDTNP project
- Evidence tracks:
  - Detected DHS Positions (Embryos)
  - DHS Read Density (Embryos)


Standardize analysis of MachiBase and modENCODE CAGE data using CAGEr

- Bioconductor package developed by RIKEN
- Map datasets against release 6 assembly
  - 37 modENCODE CAGE samples, 7 MachiBase samples
  - Define TSS and promoters for each sample
  - Define consensus promoters across all samples


TSS classifications based on CAGEr

Peaked
- FlyBase Genes
- modENCODE CAGE Peaks
- modENCODE CAGE (Plus)

Intermediate
- FlyBase Genes
- modENCODE CAGE Peaks
- modENCODE CAGE (Plus)

Broad
- FlyBase Genes
- modENCODE CAGE Peaks
- modENCODE CAGE (Minus)
Benefits of RAMPAGE

- **RAMPAGE = RNA Annotation and Mapping of Promoters for Analysis of Gene Expression**
- CAGE only allows sequencing of short sequence tags (~27 bp) near the 5' cap
- Ambiguous read mapping to large parts of the genome
- RAMPAGE produces long paired-end reads instead of short sequence tags
- Developed novel algorithm to identify TSS clusters
- Used paired-end information during peak calling
- Used Cufflinks to produce partial transcript models


RAMPAGE results on the GEP UCSC Genome Browser

- Lifted RAMPAGE results from release 5 to release 6
- Results from 36 developmental stages
- Combined TSS peak call from all samples

Available under the “Expression and Regulation” section

Evidence for TSS annotations (in general order of importance)

1. Experimental data
   - RAMPAGE
   - ATAC-Seq
   - RNA-Seq
2. Conservation
   - Type of TSS (peaked/intermediate/broad) in *D. melanogaster*
   - Sequence similarity to initial exon in *D. melanogaster*
   - Sequence similarity to other *Drosophila* species (ROAST)
3. Core promoter motifs
   - Inr, TATA box, etc.

Evidence tracks for TSS annotations in the *D. ananassae* GEP UCSC Genome Browser

- Combined RAMPAGE TSS
- Eye disc ATAC-Seq
- Histone modifications ChIP-Seq (H3K4me2)
- TSS predictions from Gnomon, N-SCAN PASA-EST, and Augustus
- RNA-Seq data from multiple stages and tissues
- Combined splice junction predictions

TSRchitect analysis results for the *Drosophila* RAMPAGE datasets

- *D. erecta, D. ananassae, and D. pseudoobscura*
- Samples collected at 1-hour intervals throughout embryonic development (1-hour to 23-hour embryos)

“Combined RAMPAGE TSS” track on the GEP UCSC Genome Browser

Calculate promoter Shape Index (SI) with TSRchitect

ATAC-Seq: Assay for Transposase-Accessible Chromatin with high-throughput Sequencing

ATAC-Seq data for D. ananassae and 10 other Drosophila species

Determine the gene structure in D. melanogaster

Optimize alignment parameters based on expected levels of conservation

Identify the initial transcribed exon using NCBI blastn

- Derive alignment scores using information theory
- Relative entropy of target and background frequencies
- Match +2, Mismatch -3 optimized for 90% identity
- Match +1, Mismatch -1 optimized for 75% identity

<p>| TABLE 1: PAM Substitution Scores Based on the Uniform Matrix Model |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>PAM</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>4.0</td>
<td>5.0</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
<td>3.5</td>
<td>4.5</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td>3</td>
<td>4.0</td>
<td>5.0</td>
<td>6.0</td>
<td>7.0</td>
<td>8.0</td>
</tr>
<tr>
<td>4</td>
<td>5.5</td>
<td>6.5</td>
<td>7.5</td>
<td>8.5</td>
<td>9.5</td>
</tr>
<tr>
<td>5</td>
<td>7.0</td>
<td>8.0</td>
<td>9.0</td>
<td>10.0</td>
<td>11.0</td>
</tr>
<tr>
<td>6</td>
<td>8.5</td>
<td>9.5</td>
<td>10.5</td>
<td>11.5</td>
<td>12.5</td>
</tr>
<tr>
<td>7</td>
<td>10.0</td>
<td>11.0</td>
<td>12.0</td>
<td>13.0</td>
<td>14.0</td>
</tr>
<tr>
<td>8</td>
<td>11.5</td>
<td>12.5</td>
<td>13.5</td>
<td>14.5</td>
<td>15.5</td>
</tr>
<tr>
<td>9</td>
<td>13.0</td>
<td>14.0</td>
<td>15.0</td>
<td>16.0</td>
<td>17.0</td>
</tr>
<tr>
<td>10</td>
<td>14.5</td>
<td>15.5</td>
<td>16.5</td>
<td>17.5</td>
<td>18.5</td>
</tr>
<tr>
<td>11</td>
<td>16.0</td>
<td>17.0</td>
<td>18.0</td>
<td>19.0</td>
<td>20.0</td>
</tr>
<tr>
<td>12</td>
<td>17.5</td>
<td>18.5</td>
<td>19.5</td>
<td>20.5</td>
<td>21.5</td>
</tr>
<tr>
<td>13</td>
<td>19.0</td>
<td>20.0</td>
<td>21.0</td>
<td>22.0</td>
<td>23.0</td>
</tr>
<tr>
<td>14</td>
<td>20.5</td>
<td>21.5</td>
<td>22.5</td>
<td>23.5</td>
<td>24.5</td>
</tr>
<tr>
<td>15</td>
<td>22.0</td>
<td>23.0</td>
<td>24.0</td>
<td>25.0</td>
<td>26.0</td>
</tr>
<tr>
<td>16</td>
<td>23.5</td>
<td>24.5</td>
<td>25.5</td>
<td>26.5</td>
<td>27.5</td>
</tr>
<tr>
<td>17</td>
<td>25.0</td>
<td>26.0</td>
<td>27.0</td>
<td>28.0</td>
<td>29.0</td>
</tr>
<tr>
<td>18</td>
<td>26.5</td>
<td>27.5</td>
<td>28.5</td>
<td>29.5</td>
<td>30.5</td>
</tr>
<tr>
<td>19</td>
<td>28.0</td>
<td>29.0</td>
<td>30.0</td>
<td>31.0</td>
<td>32.0</td>
</tr>
<tr>
<td>20</td>
<td>29.5</td>
<td>30.5</td>
<td>31.5</td>
<td>32.5</td>
<td>33.5</td>
</tr>
</tbody>
</table>

Perform `blastn` search of Rad23:1 against the *D. ananassae* contig

Extrapolate TSS position based on `blastn` alignment of the initial transcribed exon

TSS annotation for *Rad23* in *D. ananassae*

TSS annotation summary

Questions?