Searching for Transcription Start Sites in *Drosophila*

Wilson Leung  01/2018

Outline
- Transcription start sites (TSS) annotation goals
- Promoter architecture in *D. melanogaster*
- New *D. melanogaster* TSS datasets
- Find the initial transcribed exon
- Annotate putative transcription start sites
- Search for core promoter motifs

Muller F element, heterochromatic, and euchromatic genes show similar expression levels

Muller 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

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

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

Computationally tractable to analyze conserved motifs in multiple eukaryotic genomes

Goals for the transcription start sites (TSS) annotations

- 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

Estimated evolutionary distances with respect to D. melanogaster

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
- Low percent identity (~50-70%) between D. biarmipes contigs and D. melanogaster UTRs
- Most gene finders do not predict UTRs
- Lack of experimental data
- Cannot use RNA-Seq data to precisely define the TSS
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. Identify any core promoter motifs in region
7. Define TSS positions or TSS search regions

### RNA Polymerase II core promoter

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

![RNA Polymerase II core promoter diagram](image)


### Peaked versus broad promoters

**Peaked promoter** (Single strong TSS)

**Broad promoter** (Multiple weak 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


### 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

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


Peaked and broad promoters are enriched in different core promoter motifs

9-state chromatin model

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 6.16)
  - modENCODE TSS (Celniker) annotations
  - DNase I hypersensitive sites (DHS)
  - CAGE and RAMPAGE TSS datasets
  - 9-state and 16-state chromatin models
  - Transcription factor binding site (TFBS) HOT spots

DNaseI Hypersensitive Sites (DHS) correspond to accessible regions

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></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Intermediate</td>
<td>≤ 1</td>
<td>&gt; 1</td>
</tr>
<tr>
<td></td>
<td>&gt; 1</td>
<td>≤ 1</td>
</tr>
<tr>
<td>Broad</td>
<td>&gt; 1</td>
<td>&gt; 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

---

**DEMO: Classify the core promoter of *D. melanogaster* Rad23**

---

**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)


---

**Additional TSS data available in FlyBase release 6.11**


---

**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

---

---
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

Evidence for TSS annotations
(in general order of importance)

1. Experimental data
   - RNA-Seq
   - RNA Pol II ChIP-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 (Multiz)

3. Core promoter motifs
   - Inr, TATA box, etc.


Stages of Development

Changes in the dominant TSS of Rad23 across different developmental stages

Determine the gene structure in D. melanogaster

FlyBase: GBrowse

Gene Record Finder: Transcript Details

Retrieve the sequences of the initial exons from the Transcript Details tab of the Gene Record Finder

Use placement of the flanking exons to reduce the size of the search region if possible

Increase sensitivity of nucleotide searches
- Change Program Selection to blastn
- Change Word size to 7
- Change Match/Mismatch Scores to +1, -1
- Change Gap Costs to Existence: 2, Extension: 1

Identify the initial transcribed exon using NCBI blastn
Extrapolate TSS position based on blastn alignment of the initial transcribed exon

Assume the length of the initial transcribed exon is conserved between *D. melanogaster* and *D. biarmipes*

Core promoter motifs can affect gene expression levels

**SCPI:**


Use core promoter motifs to support TSS annotations

- Some sequence motifs are enriched in the region (~300 bp) surrounding the TSS
- Some motifs (e.g., Inr, TATA) are well-characterized
- Other motifs are identified based on computational analysis
- Presence of core promoter motifs can be used to support the TSS annotations
- Inr motif (TCAKTY) overlaps with the TSS (-2 to +4)
- Absence of core promoter motifs is a negative result
- Most *D. melanogaster* TSS do not contain the Inr motif

Use UCSC Genome Browser Short Match to find *Drosophila* core promoter motifs

**TATA box**

**Initiator (Inr)**


Available under "Projects" ➔ "Annotation Resources" ➔ "Core Promoter Motifs" on the GEP web site: http://gander.wustl.edu/~wilson/core_promoter_motifs.html

Core Promoter Motifs tracks

- Show core promoter motif matches for each contig
- Separated by strand
- Visualize matches to different core promoter motifs
- Use UCSC Table Browser (or other means) to export the list of motif matches within the search region

**DEMO:** Use the Inr motif to support the TSS position of *Rad23*
RNA PolII ChIP-Seq tracks
(available for D. biarmipes, D. elegans, and D. ficusphila)

- Show regions that are enriched in RNA Polymerase II compared to input DNA

Using RNA-Seq and RNA PolII ChIP-Seq data to define the TSS search region

TSS annotation for Rad23

- TSS position: 28,936
  - Conservation with D. melanogaster
    - blastn search of initial exon
    - "D. mel Transcripts" track
    - Location of the Inr motif

- TSS search region: 28,716-28,936
  - Enrichment of RNA PolII upstream of the TSS position
  - RNA-Seq read coverage upstream of the TSS position
  - Search region defined by the extent of the RNA PolII peak

TSS annotation resources

- Walkthroughs:
  - Annotation of Transcription Start Sites in Drosophila
  - Sample TSS report for onecut

- Reference:
  - TSS Annotation Workflow

- GEP Annotation Report:
  - Classify the type of core promoter
  - Evidence that supports or refutes the TSS annotation
  - Distribution of core promoter motifs

Additional TSS annotation resources

- The D. melanogaster gene annotations are the primary source of evidence

- Resources that could be useful if the D. melanogaster evidence is ambiguous
  - Whole genome alignments of 26 Drosophila species
  - PhastCons and PhyloP conservation scores

- Genome browsers for 26 Drosophila species
  - RNA Pol II ChIP-Seq (D. biarmipes, D. elegans, and D. ficusphila)
  - CAGE data for D. pseudoobscura
  - RNA-Seq coverage, splice junctions, assembled transcripts
  - Gnomon and Augustus gene predictions

Phylogenetic tree for 26 Drosophila species

Tree scale: 0.1
Genome alignments of 25 Drosophila species against D. melanogaster

<table>
<thead>
<tr>
<th>Genomes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. simulans</td>
<td>2017 (University of Arizona)</td>
</tr>
<tr>
<td>D. yakuba</td>
<td>2011 (Cold Spring Harbor Laboratory)</td>
</tr>
<tr>
<td>D. erecta</td>
<td>2014 (Cold Spring Harbor Laboratory)</td>
</tr>
<tr>
<td>D. antiqua</td>
<td>2013 (University of Arizona)</td>
</tr>
<tr>
<td>D. buzzatii</td>
<td>2018 (SOCL244)</td>
</tr>
<tr>
<td>D. eugenei</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. miranda</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. virilis</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. azteca</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. nigrispars</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. pseudoobscura</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. persimilis</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. santomeana</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. immigrans</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>2013 (SOCL244)</td>
</tr>
</tbody>
</table>

Use the conservation tracks to identify regions under selection

PhyloP scores:
- Under negative selection
- Fast-evolving

Examine the ROAST alignments to identify the orthologous TSS regions

ROAST Alignments for 26 Drosophila Species

<table>
<thead>
<tr>
<th>Genomes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. melanogaster</td>
<td>Core promoters</td>
</tr>
<tr>
<td>D. simulans</td>
<td>2017 (University of Arizona)</td>
</tr>
<tr>
<td>D. yakuba</td>
<td>2011 (Cold Spring Harbor Laboratory)</td>
</tr>
<tr>
<td>D. erecta</td>
<td>2014 (Cold Spring Harbor Laboratory)</td>
</tr>
<tr>
<td>D. antiqua</td>
<td>2013 (University of Arizona)</td>
</tr>
<tr>
<td>D. buzzatii</td>
<td>2018 (SOCL244)</td>
</tr>
<tr>
<td>D. eugenei</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. miranda</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. virilis</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. azteca</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. nigrispars</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. pseudoobscura</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. persimilis</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. santomeana</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. immigrans</td>
<td>2013 (SOCL244)</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>2013 (SOCL244)</td>
</tr>
</tbody>
</table>

TSS annotation summary

- Most of the D. melanogaster core promoters have multiple TSS
- Classify the type of promoter (peaked/intermediate/broad) based on the transcriptome evidence from D. melanogaster
- Define search regions that might contain TSS
- Use multiple lines of evidence to infer the TSS region
- Identify initial exon
  - RNA-Seq coverage
  - blastn (change search parameters)
- Distribution of core promoter motifs (e.g., Inr)
- RNA PolII ChIP-Seq peaks
- Maintain conservation compared to D. melanogaster

Questions?

- TSS Annotation Workflow
  - Identify strong TSS in D. melanogaster orthologs
  - Identify isoform with closest TSS
  - Characterize TSS in the D. melanogaster orthologs
  - Profile genomic region around TSS
  - Annotate TSS in the target species (e.g., D. simulans)

- Distribution of core promoter motifs (e.g., Inr)
- RNA PolII ChIP-Seq peaks
- Maintain conservation compared to D. melanogaster