Searching for Transcription Start Sites in *Drosophila*

Wilson Leung 08/2022

Outline

- Transcription start sites (TSS) annotation goals
- Promoter architecture in *D. melanogaster*
- Challenges with TSS annotations
- Protocol for defining TSS positions and TSS search regions in *D. kikkawai*

Goal: identify factors that enable expression of F Element genes in a heterochromatic environment

- F Element genes show a similar range of expression levels as euchromatic genes — how do they do it?
- Identify conserved, regulatory motifs which regulate the expression of F Element genes

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
Transcription factor binding sites (motifs) are short and degenerate

<table>
<thead>
<tr>
<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>
</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>
</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>
</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>
</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 sources of information, 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 profiles.

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

phyloP conservation scores surrounding the TSSs of D. melanogaster genes

Motif finding using multiple genes within a single species
Motif finding using single gene in multiple species

Genes
Conserved Elements
PhyloP

Multiple Sequence Alignment of 36 Drosophila 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.

Species analyzed by GEP students to identify conserved motifs on the F Element

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

Data from Table 1 of the modENCODE comparative genomics white paper

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 goals for the F Element project

- Research goal:
  - Identify motifs that enable 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

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

Outline

- Transcription start sites (TSS) annotation goals
- Promoter architecture in *D. melanogaster*
- Challenges with TSS annotations
- Protocol for defining TSS positions and TSS search regions in *D. kikkawai*

Peaked versus broad promoters

**Peaked promoter (Single strong TSS)**

**Broad promoter (Multiple weak TSS)**

![Peaked versus broad promoters](image)


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

![Basal transcription initiation factors binding sites](image)


Core promoter motifs can affect gene expression levels

![Core promoter motifs](image)


9-state chromatin model

![9-state chromatin model](image)

Three common types of core promoters

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

**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

5’ degraded products in PacBio Iso-Seq data
**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


---

**Techniques for defining the TSS**

- Identify sequences adjacent to the 5’ cap 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)
- RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression (RAMPAGE)

Available in *D. melanogaster, D. ananassae, D. bipectinata, D. kikkawai,* and *D. takahashii*

---

**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 6.46)
  - CAGE and RAMPAGE datasets
  - modENCODE TSS (Celniker) annotations
  - DNase I hypersensitive sites (DHS)
  - 9-state and 16-state chromatin models
  - Transcription factor binding site (TFBS) HOT spots

---

**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

**Promoter classifications based on promoter width and the Shape Index**

- **Peak:** SI > 1
- **Broad:** SI ≤ 1

---

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

\[
SI = \sum_i p_i \log_2 \left( \frac{p_i}{\bar{p}} \right)
\]

\( p_i = \text{Probability of observing a TSS at position } i \)

\( \bar{p} = \text{Set of base positions with at least one TSS tag} \)

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


Changes in the dominant TSS of Rad23 across different developmental stages

DNaseI Hypersensitive Sites (DHS) correspond to accessible regions

DNaseI Hypersensitive Sites (DHS) correspond to accessible regions

Additional DHS data from different stages of embryonic development

High turnover of DNase I Hypersensitive Sites (DHS) between human and mouse

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


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

D. kikkawai TSS annotation workflow

1. Note the gene structure of the D. melanogaster ortholog
   Characterize the promoter based on FlyBase annotations, modENCODE CAGE data, and TSRchitect analysis of the RAMPAGE data

2. Compare the initial exon of the D. melanogaster ortholog against the D. kikkawai scaffold with blastn
   Use more sensitive cross-species search parameters

3. Use the D. kikkawai RAMPAGE data to define the TSS positions and search regions

Evidence tracks for characterizing the promoters in D. melanogaster

FlyBase gene annotations (release 6.46)
Updated Transcriptome Tracks:
- RAMPAGE data (two biological replicates)
- TSRchitect Combined RAMPAGE TSS (Replicate 1 | 2)
- Strand-specific RNA-Seq data
  Combined modENCODE RNA-Seq (Development)
Expression and Regulation:
- Combined modENCODE CAGE TSS
- MACS2 DHS

Define Transcription Start Regions (TSRs) using the TSRchitect Bioconductor package

Outline

- Transcription start sites (TSS) annotation goals
- Promoter architecture in D. melanogaster
- Challenges with TSS annotations
- Protocol for defining TSS positions and TSS search regions in D. kikkawai
Determine the gene structure in *D. melanogaster*

FlyBase: JBrowse

Gene Record Finder: Transcript Details

Optimize alignment parameters based on expected levels of conservation

| TABLE 1: PAM Substitution Scores Based on the Uniform Mutation Model |
|------------------------|------------------------|------------------------|------------------------|------------------------|
| PAM | Forward | Reverse | Forward | Reverse | Forward | Reverse | Forward | Reverse |
| 50 | 0.50 | 0.20 | 0.55 | 0.15 | 0.50 | 0.25 | 0.55 | 0.25 |
| 55 | 0.45 | 0.16 | 0.50 | 0.12 | 0.45 | 0.20 | 0.50 | 0.20 |
| 60 | 0.40 | 0.12 | 0.45 | 0.10 | 0.40 | 0.15 | 0.45 | 0.15 |
| 65 | 0.35 | 0.08 | 0.40 | 0.06 | 0.35 | 0.10 | 0.40 | 0.10 |
| 70 | 0.30 | 0.06 | 0.35 | 0.04 | 0.30 | 0.08 | 0.35 | 0.08 |
| 75 | 0.25 | 0.04 | 0.30 | 0.02 | 0.25 | 0.06 | 0.30 | 0.06 |


Change **algorithm parameters** to perform a more sensitive cross-species *blastn* search

<table>
<thead>
<tr>
<th>For Algorithm Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Parameters:</td>
</tr>
<tr>
<td>Word size: 7</td>
</tr>
<tr>
<td>Gap Costs: Existence 2, Extension 1</td>
</tr>
<tr>
<td>Match / Mismatch Scores: 1, -1</td>
</tr>
</tbody>
</table>

Identify the initial exon in *D. kikkawai*

- Retrieve the sequences of the initial exons of the *D. melanogaster* gene from the Transcript Details tab of the Gene Record Finder
- Use placement of the flanking exons to reduce the size of the search region in *D. kikkawai* 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

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

- 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

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

- Assume the length of the initial exon is conserved between *D. melanogaster* and *D. kikkawai*
Evidence for *D. kikkawai* TSS annotations (in order of general importance)

1. Experimental data
   - RAMPAGE
   - RNA-Seq
2. Conservation
   - Sequence similarity to initial exon in *D. melanogaster*
   - Sequence similarity to other *Drosophila* species (ROAST)
3. Promoter predictions
   - RefSeq Genes, N-SCAN PASA-EST, Augustus
4. Core promoter motifs
   - Intr, TATA box, etc.

Evidence tracks for TSS annotations available in the *D. kikkawai* Genome Browser

- RAMPAGE data from adult females and embryos
- RNA-Seq data from adult females, adult males, and mixed embryos
- Combined splice junction predictions
- TSS predictions from RefSeq Genes, N-SCAN PASA-EST, and Augustus
- BLAT alignments against the RefSeq transcripts from 36 *Drosophila* species

**TSRchitect** analysis results for the *D. kikkawai* RAMPAGE datasets

- RAMPAGE data produced by the Gingeras Lab at Cold Spring Harbor Laboratory (CSHL)
- Embryos and sexed flies from the Ellison Lab at Rutgers University

Define TSS position and search regions based on *D. kikkawai* RAMPAGE data

- Rad23 promoter shape in *D. melanogaster*
  - Shape: Broad; SI: -1.73
- Rad23 promoter in *D. kikkawai*
  - Shape: Broad; SI: -1.36
  - TSS Position estimated from blastn search: 546,863
  - TSS Position estimated from RAMPAGE read density:
    - Adult Females: 546,844
    - Mixed Embryos: 546,844
  - TSS Search Region from Combined RAMPAGE peak: 546,797-546,863

Overview of the *D. kikkawai* RAMPAGE evidence tracks

- Extrapolated start position for Rad23 based on the blastn alignment
TSS annotation summary

- Most *D. melanogaster* promoters have multiple TSSs
- Use the Shape Index to characterize the promoter shape
- Use multiple lines of evidence to infer the TSS positions and TSS search regions in *D. kikkawai*
  - blastn search against initial exon from *D. melanogaster*
  - Use more sensitive cross-species search parameters
- RAMPAGE
- RNA-Seq coverage
- TSS predictions from gene predictors (e.g., RefSeq Genes)
- Whole genome multiple sequence alignments (ROAST)
- *D. melanogaster* Drosophila Conservation (36 Species)
- *D. bipectinata* Flyseq Cons. (17 Species)

Questions?

PhyloP scores: 
- Under negative selection
- Fast-evolving

Structure of a typical mRNA

Pesole G. et al. Genome Biology. 2002; 3(3) reviews0004.1-reviews0004.10.

RAMPAGE protocol


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

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

- Combined RAMPAGE TSS
- Eye discs ATAC-Seq
- Histone modifications ChIP-Seq (H3K4me2)
- TSS predictions from RefSeq Genes, 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

Conservation tracks on the *D. melanogaster* GEP UCSC Genome Browser

- Whole genome alignments of multiple *Drosophila* species
- *Drosophila* Chain/Net composite track
- Generate multiple sequence alignments from these pairwise alignments (ROAST)
- Identify conserved regions from ROAST alignments
  - PhastCons: identify conserved elements
  - PhyloP: measure level of selection at each nucleotide
- *Multiz* alignment of 124 insect species available on the official UCSC Genome Browser
  - Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) assembly

Divergent promoters are common in mammalian genomes

- GRO-seq data shows 13,633 genes in human fibroblast nuclei exhibit significant divergent transcription
- 77% of active genes

Use SQANTI with CAGE data to identify true TSSs in Iso-Seq data

modENCODE TSS annotations

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

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 (TCAKY) 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

- Available at https://gander.wustl.edu/~wilson/core_promoter_motifs.html

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

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

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

Define the TSS search region based on blastn alignment to the initial transcribed exon

TSS annotation for Rad23

Additional TSS annotation resources

RNA PolII ChIP-Seq tracks

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

If the blastn alignment satisfies these criteria:

- Has an E-value less than 1e-5
- Covers more than half of the initial transcribed exon
- Requires extrapolation of less than 150bp to estimate the TSS position
- Is in congruence with other evidence tracks
  - Examples: RNA-Seq read coverage, RNA PolyII ChIP-Seq data

Define TSS search region as ±300bp surrounding the TSS position estimated from the blastn alignment

See Module TSS4 for details

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

Resources that could be useful if the D. melanogaster evidence is ambiguous

- Multiple sequence alignments of 36 Drosophila species
- PhastCons and PhyloP conservation scores
- Genome browsers for 35 Drosophila species
- RNA Pol II ChIP-Seq data
- CAGE data for D. pseudoobscura
- RAMPAGE data for D. erecta, D. ananassae, D. bipectinata, D. kikkawai, D. takahashii, and D. pseudoobscura
- ATAC-Seq data for 10 Drosophila species
- RNA-Seq coverage, splice junctions, assembled transcripts
- RefSeq Genes, Augustus, and N-SCAN P4x EST gene predictions