Annotation of Other Genomic Features within the F Element Project

Wilson Leung  07/08/2022

Checking for additional features

- Analyze each Genscan gene prediction that does not overlap with the genes you have annotated.
- Analyze regions with RNA-Seq read coverage that do not overlap with the features you have annotated or with transposon remnants identified by RepeatMasker.

Analysis of Genscan predictions

- FlyBase blastp search of the predicted protein against the D. melanogaster "Annotated proteins" database
  - Analyze matches with E-values < 1e-5
- NCBI blastp search of the predicted protein against the NCBI "Reference proteins (refseq_protein)" database
  - Analyze matches to curated RefSeq gene models (i.e., accession numbers with the "NP" prefix)
- Determine if the feature is a protein-coding gene (e.g., novel paralog), pseudogene, or partial gene duplication

Analysis of regions with substantial RNA-Seq read coverage

- The FANTOM consortium uses an arbitrary length cutoff of 300 nt (100aa) to identify putative mRNAs
  - Reviewed in Dinger ME et al., 2008
- Identify regions with contiguous RNA-Seq read coverage
  - Perform a NCBI blastx search of each region against the refseq_protein database
    - Analyze matches to curated RefSeq gene models with E-values < 1e-5


Goals for the analysis of additional genomic features

- Analysis of the features might require additional techniques and tools beyond the standard annotation protocol
- Document the genomic features in the annotation report for the F Element Project
  - Reconcilers can follow up on the analysis of the feature
- Consult with Dr. Arrigo and Wilson during Office Hours for the F Element Project

Examples of additional features in the expanded F Elements

- Paralogs (Chris Hemauer)
- Pseudogenes (Ishtar Olaveja)
- Retrogenes (Meher Arora)
- Partial gene duplications (Meher Arora)
- Pseudogene clusters (Ethan Cordes)
- Nuclear mitochondrial DNA segments (NUMT) (Martin Dalling)
Novel paralog of PMCA in the D. kikkawai F Element scaffold tig00000049

Region with similarity to yin in the D. bipectinata F Element project contig16

The D. melanogaster yin gene is located on the X chromosome

Putative ortholog of yin is placed at scaffold_661 in the D. bipectinata assembly

The tblastn alignment of yin-PD against contig16 include frame-shifts and in-frame stop codons

Analysis of a mael retrogene in the D. ananassae F Element project contig4
The *D. melanogaster mael* gene is located on the Muller D Element (chr3L).

Putative ortholog of *mael* is placed at scaffold QMES02000178 in the *D. ananassae* assembly.

The tblastn alignment of *mael*-PC against contig4 shows one open reading frame.

Partial duplication of *Dyrk3* in the *D. ananassae* F Element project contig24.

Multiple blastx matches to *Dyrk3* CDS 5_2346_2 in the *D. ananassae* F Element contig24.

Additional matches to CDS 5_2346_2 in contig24 contain in-frame stop codons.
Region with multiple matches to CDS 5_2346_2 contain multiple copies of Helitron-N1 transposons

FlyBase blastx search shows three potential genes within this region of contig24

Frame shift and in-frame stop codon in the blastx alignments between the msk CDS 6_2209_2 and D. ananassae F Element contig24

A previous study shows the transfer of mitochondrial DNA to the D. ananassae F Element

Putative ortholog of msk is placed at scaffold QMES02000178 in the D. ananassae assembly
**D. ananassae** F Element scaffold QMES02000012 shows similarity to **mitochondrial proteins**

![Graphical representation of similarity](image)

Insertion of the *D. ananassae* mitochondrial genome into the F Element scaffold QMES02000012

- Alignment Scores:
  - < 40
  - 40 - 50
  - 50 - 80
  - 80 - 200
  - ≥ 200

- Distribution of the top 7 Blast Hits on 1 subject sequences

Analysis summary of other genomic features on the expanded F Elements

- Transposons are major contributors to the expansion of the *Drosophila* F Elements
- Active transposons might introduce unusual genomic features to the F Element
  - Novel paralogs, pseudogenes, retrogenes, NUMTs, ...
- Consult with F Element Project staff or the GEP Virtual TAs on the analysis of these genomic features
- F Element Project **Office Hours**

Questions?

- Transposon density = 78%

Nicole Torosin (Ellison Lab)