Web Databases for Drosophila
An introduction to web tools, databases and NCBI BLAST

Wilson Leung 08/2019

Agenda

• GEP annotation project overview
• Web databases for Drosophila annotation
  – UCSC Genome Browser
  – NCBI / BLAST
  – FlyBase
  – Gene Record Finder

Annotation – adding labels to a sequence

• Genes: Novel or known genes, pseudogenes
• Regulatory Elements: Promoters, enhancers, silencers
• Non-coding RNA: tRNAs, miRNAs, siRNAs, snoRNAs
• Repeats: Transposable elements, simple repeats
• Structural: Origins of replication
• Experimental Results:
  – DNase I Hypersensitive sites
  – ChIP-chip and ChIP-Seq datasets (e.g., modENCODE)

Gene annotation workflow

Visualize a genomic region with evidence tracks
GEP UCSC Genome Browser

Identify interesting features and putative orthologs
NCBI BLAST

Learn about the putative D. melanogaster ortholog
NCBI / FlyBase

Understand the gene and isoform structure
Gene Record Finder

GEP Drosophila Annotation Projects

Informant Species
GEP Publications
Motif Project
Expanded F Project

The Pathways Project analyze genes from 27 Drosophila species
UCSC Genome Browser

- Provide graphical view of genomic regions
  - Sequence conservation
  - Gene and splice site predictions
  - RNA-Seq and splice junction predictions

- BLAT – BLAST-Like Alignment Tool
  - Map protein or nucleotide sequences against an assembly
  - Faster but less sensitive than BLAST

- Table Browser
  - Access data used to create the graphical browser

Control how evidence tracks are displayed on the Genome Browser

- Most evidence tracks have five display modes:
  - Hide: track is hidden
  - Dense: all features (including overlapping features) are displayed on a single line
  - Squish: overlapping features are drawn on separate lines, features are half the height compared to full mode
  - Pack: overlapping features are drawn on separate lines, features are the same height as full mode
  - Full: Each feature is displayed on its own line

- Some annotation tracks (e.g., RepeatMasker) only have a subset of these display modes

Two different versions of the UCSC Genome Browser

- Official UCSC Version
  - http://genome.ucsc.edu
  - Published data, lots of species, whole genomes; used for “Chimp Chunks”

- GEP Version
  - http://gander.wustl.edu
  - GEP projects, Drosophila genome assemblies; used for Drosophila annotations

Additional resources for the UCSC Genome Browser

- Training section on the UCSC web site
  - http://genome.ucsc.edu/training/index.html
  - Video tutorials
  - User guides
  - Mailing lists

- Biostars
  - https://www.biostars.org/t/ucsc/
  - Questions with the “ucsc” tag

UCSC GENOME BROWSER DEMO
Use **BLAST** to detect sequence similarity

- **BLAST** = **B**asic **L**ocal **A**lignment **S**earch **T**ool
- Why is **BLAST** popular?
  - Provide statistical significance for each match
  - Good balance of sensitivity and speed
- Find local regions of similarity irrespective of where they are in the sequence

**Common types of BLAST programs**

- Except for **BLASTN**, all alignments are based on comparisons of protein sequences
- Decide which **BLAST** program to use based on the type of query and subject sequences:

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database (Subject)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BLASTN</strong></td>
<td>Nucleotide</td>
<td>Nucleotide</td>
</tr>
<tr>
<td><strong>BLASTP</strong></td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td><strong>BLASTX</strong></td>
<td>Nucleotide → Protein</td>
<td>Protein</td>
</tr>
<tr>
<td><strong>TBLASTN</strong></td>
<td>Protein</td>
<td>Nucleotide → Protein</td>
</tr>
<tr>
<td><strong>TBLASTX</strong></td>
<td>Nucleotide → Protein</td>
<td>Nucleotide → Protein</td>
</tr>
</tbody>
</table>

**Common BLAST programs use cases**

- **BLASTN**: Map mRNAs against genomic assemblies
- **BLASTP**: Search for proteins similar to predicted genes
- **BLASTX**: Map proteins against genomic sequence
- **TBLASTN**: Map proteins against genomic assemblies
- **TBLASTX**: Identify genes in unannotated sequences

See the “Guide to BLAST home and search pages” document for details:


**NCBI BLAST nucleotide databases**

- GenBank Non-Redundant Nucleotide Database (nr/nt)
  - Most comprehensive but some entries are low quality
  - Exclude sequences from whole genome assemblies
- RefSeq RNA Database
  - mRNA and non-coding RNA entries from the NCBI Reference Sequence Project
  - Include real and computationally predicted sequences
- Transcriptome Shotgun Assembly (TSA) Database
  - Transcripts assembled from expressed sequence tags (ESTs) and RNA-Seq reads

**NCBI BLAST protein databases**

- GenBank Non-Redundant Protein Database (nr)
  - Most comprehensive but some entries are low quality
  - Include sequences from RefSeq and UniProtKB
- RefSeq Protein Database
  - Sequences from the NCBI Reference Sequence Project
  - Higher quality than the nr database
  - Include real and computationally predicted sequences
- UniProtKB / Swiss-Prot Protein Database
  - Manually curated proteins from literature
  - Real proteins with known functions
  - Much smaller database than either RefSeq or nr

**Where can I run BLAST?**

- NCBI BLAST web service
- EBI BLAST web service
  - https://www.ebi.ac.uk/Tools/sss/
- FlyBase BLAST (Drosophila and other insects)
  - http://flybase.org/blast/

**Where can I run BLAST?**

- NCBI BLAST web service
- EBI BLAST web service
  - https://www.ebi.ac.uk/Tools/sss/
- FlyBase BLAST (Drosophila and other insects)
  - http://flybase.org/blast/
NCBI BLAST DEMO

National Center for Biotechnology Information (NCBI)


Key features of NCBI

• Strengths
  – Most comprehensive among publicly available databases
  – PubMed for literature searches
  – Comprehensive BLAST web service

• Weaknesses
  – Web site is large and complex
  – Quality of GenBank records may vary

• Use cases
  – Perform BLAST searches against Refseq, nr/nt databases
  – Use BLAST (bi2seq) to align two or more sequences

FlyBase — Database for the Drosophila research community

http://flybase.org/

Key Features of FlyBase

• Lots of ancillary data for each gene in Drosophila
• Curation of literature for each gene
• Reference Drosophila annotations for all the other databases (including NCBI)
• Fast release cycle (6-8 releases per year)

• Use cases:
  – BLAST searches against Drosophila genome assemblies
  – Genome browsers (GBrowse/JBrowse)
  – RNA-Seq expression profile and similarity searches

Web databases and tools

• Many genome databases available
  – Be aware of different annotation releases
  – Use FlyBase as the canonical reference

• Web databases are being updated constantly
  – Update GEP materials before semester starts
  – Discrepancies in exercise screenshots
  – Minor changes in search results
  – Let us know about errors or revisions
FLYBASE DEMO

Key features of the Gene Record Finder

• List of unique coding and non-coding exons for each gene in D. melanogaster
• CDS and exon usage maps for each isoform
• Optimized for exon-by-exon annotation strategy
• Slower update release cycle than FlyBase
  — Database is updated every semester

• Use cases:
  — Get amino acid sequences and nucleotide sequences of each exon for BLAST 2 Sequences (bl2seq) searches

Summary

• GEP annotation project seeks to generate high quality manually curated gene models for multiple Drosophila species

• Use BLAST to characterize a genomic sequence

• Use web databases to gather information on a gene
  — UCSC Genome Browser
  — NCBI
  — FlyBase
  — Gene Record Finder

Questions?

https://www.flickr.com/photos/jac_opo/240254763/sizes/l/