Daphnia Genome Annotation

May 2007

Don Gilbert, gilbertd@indiana.edu
Annotation (TIGR)

- **genomics annotation**
  - gene product names
  - functional characteristics of gene products
  - physical characteristics of gene/protein/genome
  - overall metabolic profile of the organism

- **elements of annotation**
  - gene finding
  - homology searches
  - functional assignment
  - ORF management
  - data availability

- **manual vs. automatic**
  - computers do a fair job at preliminary annotation
  - high quality annotation requires manual review
wFleaBase Annotations

- **Gene Homology**
  - Nine well-annotated proteomes: Yeast, Worm, Mosquito, Fruitfly, Bee, Zebrafish, Mouse, Human, Arabidopsis

- **Gene Predictions**
  - SNAP - good ab-initio predictor
  - TwinScan in progress; Gnomon (NCBI) expected

- **PASA EST assembly analyses**
Genes Found in *Daphnia*
More small exons in *Daphnia*

<table>
<thead>
<tr>
<th></th>
<th>Daphnia</th>
<th>Fruitfly</th>
<th>Mouse</th>
<th>Worm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene span (ave/median)</td>
<td>2,500/1,700</td>
<td>5,500/2,000</td>
<td>32,000/8,000</td>
<td>3,300/2,100</td>
</tr>
<tr>
<td>Exons/gene</td>
<td>6 to 9</td>
<td>2.5</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Exon size</td>
<td>200</td>
<td>400</td>
<td>260</td>
<td>200</td>
</tr>
<tr>
<td>Intron size (ave/median)</td>
<td>150/70</td>
<td>900/70</td>
<td>4,800/1,200</td>
<td>310/66</td>
</tr>
<tr>
<td>CDS size (no_exons)</td>
<td>1,500</td>
<td>1,100</td>
<td>2,100</td>
<td>1,300</td>
</tr>
</tbody>
</table>
## Daphnia's "best" GO model

<table>
<thead>
<tr>
<th>Model</th>
<th>Gene Ontol Expt./Comp.</th>
<th>In Daphnia N (MOD%)</th>
<th>Daphnia Unique</th>
<th>Daphnia &quot;Best&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse</td>
<td>10,400 / 16,300</td>
<td>10,500 (55%)</td>
<td>1300</td>
<td>1750</td>
</tr>
<tr>
<td>Human</td>
<td>8,600 / 30,000</td>
<td>12,400 (55%)</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>Fruitfly</td>
<td>3,900 / 8,800</td>
<td>7,900 (59%)</td>
<td>1650</td>
<td>2300</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>1,000 / 12,000</td>
<td>15,200 (64%)</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>Worm</td>
<td>6,800 / 20,200</td>
<td>6,400 (33%)</td>
<td>na</td>
<td>547</td>
</tr>
<tr>
<td>Yeast</td>
<td>5,600 / 1,600</td>
<td>2,200 (39%)</td>
<td>na</td>
<td>97</td>
</tr>
</tbody>
</table>
Predictions and EST assemblies
Gene evidence types

- Protein homology
  - Include annotations of known genes
  - Beware of copy of copy of copy of copy of annotation

- Species EST
  - Strongest gene data for Daphnia
  - ESTs cover 2/3 to 3/4 of genes (15K to 18K)

- Gene Predictions
- Gene expression
Gene prediction types

- Prediction types
  - *Ab-initio* from trained HMM models (fgenesh, SNAP)
  - Protein gene mapping (GeneWise,…)
  - *Ab-initio* with EST, protein guide (Twinscan, Gnomon)
  - Combiners (Jigsaw, GLEAN, EvidenceModel)
- Combining several predictors is best
- Gene models weakest (15-40%); exons often good and similar among predictors
PASA Annotation

• Program to Assemble Spliced Alignments
  • Genome annotation pipeline from TIGR, used widely elsewhere
  • Exploits spliced alignments of ESTs to model gene structures
  • Maintains gene structure consistent with most experimental data.
  • Identifies all splicing variations supported by the transcripts.

• Good intro to learn how to correct gene structure

• Find at http://wfleabase.org/prerelease/
  • User: dgcguest   Password: dgcguest

http://server2.eugenes.org/cgi-bin/PASA/cgi-bin/status_report.cgi?db=pasa_daphc
### Annotation Classification for Alignment Assemblies

<table>
<thead>
<tr>
<th></th>
<th>FL-assemblies</th>
<th>EST-assemblies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PASS</td>
<td>fail</td>
</tr>
<tr>
<td>Incorporated</td>
<td>558</td>
<td>2656</td>
</tr>
<tr>
<td>UTR addition</td>
<td>1682</td>
<td>2253</td>
</tr>
<tr>
<td>Gene extension</td>
<td>153</td>
<td>15</td>
</tr>
<tr>
<td>Internal gene structure rearrangement</td>
<td>0</td>
<td>1605</td>
</tr>
<tr>
<td>-passes homology tests</td>
<td>925</td>
<td>774</td>
</tr>
<tr>
<td>-fails homology, passes ORF span</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gene Merging</td>
<td>41</td>
<td>289</td>
</tr>
<tr>
<td>Gene Splitting</td>
<td>57</td>
<td>26</td>
</tr>
<tr>
<td>Alt Splicing Isoform</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>-passes homology test</td>
<td>639</td>
<td>273</td>
</tr>
<tr>
<td>-fails homology, passes ORF span</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New Gene</td>
<td>662</td>
<td>103</td>
</tr>
<tr>
<td>Alt splice of new gene</td>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FL-assembly fails gene requirements</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Antisense</td>
<td>853</td>
<td>0</td>
</tr>
<tr>
<td>Single-exon EST-assembly incompatible</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>delayed incorporation due to gene merging</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>delayed incorporation due to gene splitting</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>18444</td>
<td></td>
</tr>
</tbody>
</table>
Resolving Failed PASA

- Read PASA Report on failure
- View Genome Map evidence
  - Dappu1_FM5 vs. Other prediction models
  - EST assembly consistency
  - Protein Homology matches
- BLAST alternate protein models
Comment: -shorter protein is only 40.3890160183066 % of the original protein length. Insufficient. (FL_alt_splice_flag; 0) Stitched EST lacks compatibility with preexisting protein annotations; invalid and no alt-splice template available. Applied to Dappu1_FM5_220407,0 >asmbl_6313-based protein
MVVKFSRKLEIVSENLLKFHNCIIILEIMNLLPFKIMFMSMIIVFLCIAALNFGTGTKGAQIQ QHFNSDPGPDSVLFLQLFTRKNPGKQILQLEDITLLEQSNYNSSLPTKIFVHG...

Status: 18. EST assembly stitched into gene model fails validation.
Comment: -shorter protein is only 44.5378151260504% of the original protein length. Insufficient. (FL_alt_splice_flag; 0) Stitched EST lacks compatibility with preexisting protein annotations; invalid and no alt-splice template available. Applied to Dappu1_FM5_196379,0
>asmbl_7600-based protein
MSFIILLCLVAFASAAPQRAAVRVLQDLPVCLLPPVADPTQNCNNFSI...
PASA : Split needed

Filtered genes: JGI:FM5
Dapudu_FM5_102722 (Feerech_pg)

EST gene updates: PASA

Nscan: DGIL
CVM combiner: DGIL
SNAP_homlog: DGIL
DP_DGIL_SND_00010916

Clean: DGIL
EST/cDNA assembly: PASA

Daphnia EST

Fruitfly genes
C63103_01

Mouse genes
MGI:2303230_01

Human genes
WBGene00016140_01
F09F.3 (WBGene00017300_03)

Yeast genes
S000005577_01
RF22 (S000005577_02)

Pan (C633103_021)
Pan (C633103_026)

Tfp12 (MGI:130543_010)
Sprnt2 (MGI:1300004_016)

W01F3.3 (WBGene00012166_01)
V41D12.6 (WBGene00019646_01)
V43F8B.3 (WBGene00012314_015)
mc1 (WBGene0001150_02)

C37C3.5 (WBGene00016498_02)
W01F3.3 (WBGene00012186_018)
W01F3.3 (WBGene00012186_69)
PASA : New Gene???

There are 3000 EST assemblies in this ambiguous category. They lack enough evidence to be sure, but they may be genes. Some have more evidence …
PASA : New Gene???

Like this among the 3000, has strong EST, some homology, and a matching prediction model -- I’d call it as a gene.
PASA Status(19): EST assembly aligns to intergenic region.

This is where annotators can find a table with 3000 ambiguous EST assemblies, view related evidence, and call them in minutes as likely genes or not likely.

<table>
<thead>
<tr>
<th>#</th>
<th>subcluster id</th>
<th>assembly acc</th>
<th>Map view</th>
<th>Chromosome Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>asmbl_2</td>
<td>scaffold_1:81185..81428</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>asmbl_3</td>
<td>scaffold_1:100774..101479</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>asmbl_5</td>
<td>scaffold_1:107202..107754</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>asmbl_6</td>
<td>scaffold_1:118959..119418</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>asmbl_8</td>
<td>scaffold_1:121929..122512</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>asmbl_10</td>
<td>scaffold_1:149923..150708</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>asmbl_19</td>
<td>scaffold_1:172179..173054</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>asmbl_20</td>
<td>scaffold_1:173179..173921</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>19</td>
<td>asmbl_22</td>
<td>scaffold_1:181610..182405</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>asmbl_23</td>
<td>scaffold_1:195958..196604</td>
<td></td>
</tr>
</tbody>
</table>
BLAST compare proteins

- Often quick map view won’t resolve ambiguous gene models
- Protein BLAST will answer: does one model match others?
- Enter primary and alternate protein models
  >Dappu1_FM5_232463
  MDHKDHHTVAHGKKGHEHTDSKHQAEENQAPRAGFQLGQMEKRVNTRLRSLKKT
  LRVLVSSLVRRRPALGPTLLSFRLSPIRKTSMWKLHNHAATADTMKSTVSTTTKMTK
  TTTLVNASKSLVIFLFAIF*
  >asmb_483-based protein
  MEKVTNRLRSLKKLKTRLVSSLVRRRPALGPTKLLSFRLSPIRKTSMWKLHNAA
  TADTMKSTVSTTTKMKTKTTLKNASKSLVIFLFAIF*
- Results: standard model has weak match, alternate doesn’t
  - Dappu1_FM5_232463: hypothetical protein RPA2457 [Rhodopseudomonas palustris]
  - asmb_483-based protein: No significant similarity found.
Annotation Strategy
Annotation Strategy

- **Web windows on**
  - **Genome Maps** - JGI and wFleaBase use same locations
  - **PASA EST Evidence**
    - Search PASA for JGI EST id (e.g. JGI_CANY881)
    - link via GBrowse detail page
  - **NCBI BLAST**
  - **Overview of genes via Gene Ontology, PASA summary, other ..**
Annotation Strategy II

- Prioritizing genes to check
  - Biological interest (gene family, function,..)
  - PASA Status: PASA RED needs human check

- Easy versus hard calls
  - Multiple evidence types support one model?
  - Alternate evidence is in conflict, or missing

- Sub-divide tasks with others
  - Many genes briefly, or fewer in detail
  - Set goal of genes per hour, week
Problems to watch

- Tandem genes
  - GeneWise tends to fail, *Ab initio* may be better
  - Tandem examples
    - scaffold_2:878732..888731: a Dappu1_FM5 model joins 3 mouse lipoprotein genes
    - scaffold_23:187883..194536: a GeneWise model joins 3 tandem genes (fly, mouse, and EST)
    - scaffold_79:251200-267599: 4 duplicates with homology, similar SNAP predictions

- Automated EST extension errors
  - Models with “est” in name: est+GeneWise, est+fgenesh
  - EST (PASA) models split into multiple genes
  - Protein homology can suggest multiple genes