Community Manual Genome Annotation and Vega

Jane Loveland
Havana group, Wellcome Trust Sanger Institute,
Hinxton, Cambridge, UK

Overview

- Manual annotation process
  - Tools
  - Biotypes
  - VEGA
    - Update tracks
  - Community Manual Annotation
    - Latest annotation
  - New data

Do we know how many genes there are?

Automatic Annotation vs Manual

Automatic Annotation
- Quick whole genome analysis ~ weeks
- Consistent annotation
- Use unfinished/Illumina sequence/shotgun assembly
- No polyA sites/signals, pseudogenes, lncRNAs
- Limited functional annotation
- Predicts ~75% loci

Manual Annotation
- Slow ~3 months per chromosome
- Prefer finished (high quality) seq
- Flexible, can deal with inconsistencies in data
- Most rules have exception
- Consult publications as well as databases
- Extensive Biotypes:
  - Excellent functional annotation
  - E.g. pseudogenes, lncRNA

Automated annotation alone is not sufficient for researchers needs

GENCODE geneset
GenCODE produces high quality reference gene annotation and experimental validation for human and mouse genomes.

Analysis and Annotation pipeline: Otter/ZMap

Otter works with Zmap to provide a complete integrated annotation workbench.

Annotation tools:
- BLAST
- Gene predictions
- RepeatMasker
- CpG prediction
- Pfam
- RefSeq
- Ensembl

Annotation process:
- Automated annotation
- Manual annotation
- Experimental verification

Otter works with Zmap to provide a complete integrated annotation workbench.

Annotation tools:
- Blixem: Interactive browser of pairwise alignments that have been stacked up in a "master-slave" multiple alignment
- Annotation tools:
  - Manual annotation
  - Tools
  - Biotypes
  - VEGA
  - Update tracks
  - Community Manual Annotation
  - Latest annotation
  - New data

Manual Annotation and Biotypes:
### Alternative Splicing

<table>
<thead>
<tr>
<th>Term</th>
<th>Diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference model</td>
<td></td>
</tr>
<tr>
<td>Skipped exon</td>
<td></td>
</tr>
<tr>
<td>Retained intron</td>
<td></td>
</tr>
<tr>
<td>Alternative splice donor</td>
<td></td>
</tr>
<tr>
<td>Alternative splice acceptor</td>
<td></td>
</tr>
<tr>
<td>Alternative first exon</td>
<td></td>
</tr>
<tr>
<td>Alternative final exon</td>
<td></td>
</tr>
<tr>
<td>Mutually exclusive</td>
<td></td>
</tr>
</tbody>
</table>

### Classifying functional transcripts within a protein coding gene:

- **RetSeq**
- **Gencode**
- **TSC2 (chr16)**
- **Nonsense-mediated decay**
- **Artifact**
- **Retained intron**

### 5' end annotation:

**GPR56** (Human G protein-coupled receptor 56 gene)

![Graph showing 5' end annotation of GPR56](image)

### RNAseq data to extend 3'UTRs

**GRIN2B**

- **Ensembl**
- **Bodymap**
- **Breast**
- **Brain**
- **polyA**
- **Extended 3' end**

### Improvements of lncRNA annotation:

**Understanding functionality**

![Diagram showing improvements of lncRNA annotation](image)

### LincRNA extended with 5' and 3' RACE 454 reads

- **454 reads**
- **CpG island**
- **Original lincRNA**
- **454 reads**
GAS5: growth arrest-specific 5 (non-protein coding) intronic snoRNAs

- Manual annotation process
  - Tools
  - Biotypes
- VEGA
  - Update tracks
  - Community Manual Annotation
  - Latest annotation
  - New data

Vega Homepage

Species Page

Gene View

Gene View (cont)
Evidence used to build transcripts:

- Manual annotation process
- Tools
- Biotypes
- VEGA
- Update tracks
- Community Manual Annotation
- Latest annotation
- New data

Ensembl view: GENCODE geneset

Gold (merged): agreed ensembl/havana

Red: coding (001 Havana, 201 Ensembl)

Blue: non-coding

http://www.gencodegenes.org

Community Annotation:

- Part of IKMC with EUCOMM annotation in mouse:
  - KOMP and NorCOMM annotation ("Blessed Annotator")
- Jamborees for species with strong community interest ("Gatekeeper")
- Xenopus tropicalis 2005 (cDNA)
- Pig 2008 (Genomic WGS)
- Rat manual annotation 2013, 2015 (BBSRC)
- Chicken 2016

Many transcript variants
- Found gene expansions and duplications
- Co-expression clustering analyses: some exhibited accelerated evolution

New genes in Pig (~1300 genes) manually annotated by community

http://www.gencodegenes.org
Community Annotation Approaches:
The value of a genome is only as good as its annotation

Rat whole genome annotation of Rnor 6.0

Otter/Zmap Annotation Software
Authentication:
Sanger single sign-on account (email)
Registered email for otterlace permitted users:
Access to our data and analysis pipeline

Mac and Linux:
Platforms of choice
Monthly updates/bugfixes

Windows:
Virtual machine image installed and run using VirtualBox
Runs an Ubuntu desktop with a bespoke Otter release

Community support: rat-annotation@sanger.ac.uk
Over 2000 genes annotated that have been chosen by the rat community

Manual annotation can be made available as a Trackhub (BigBED) that is viewable in Ensembl and UCSC

Rat Y chromosome annotated:

Vega: MHC comparison view

- Manual annotation process
  - Tools
  - Biotypes
  - VEGA
  - Update tracks
  - Community Manual Annotation
  - Latest annotation
  - New data
Ribosome Profiling

- PNRC2
  - human
  - mouse
  - zebrafish

Known CCDS
- Novel 56 aa CDS

CAGE supports multiple TSS in single exons

- BIRC6
  - human
  - mouse

- ATG 3
  - Non-conserved
  - No RP support

- ATG 1
  - RP supported

- ATG 2
  - RP supported

Different TSS access different ATGs
Gene appears to make 2 proteins

Proteomics: full re-analysis of draft proteome data

- Search space

Proteomics: Full reanalysis of draft proteome data

- 94% novel genes rejected as false positives (mostly pseudogenes)
- 17 novel genes found (11/17 single exon, 8/17 <200aa)

Mouse strains:

- Mouse strains annotation reveals new genes:
Long-read data:
Better for discovering novel alternatively spliced transcripts and full-length transcripts

Example of putative novel locus

Future plans:
Gencode geneset for mouse:
Expand and refine IncRNA and pseudogene annotation
Mouse strains

Expand community manual annotation:
Rat ongoing
Chicken
Development of Otter tool
Regions not available in Ensembl, update tracks

Incorporate new technology to find new genes?
CAGE, Mass Spec, PacBio (CaptureSeq), SLR-Seq

PAG sessions
Non-coding RNA: Sunday 4pm, Royal Palm Salon
FAANG long ncRNA analysis group: Monday, 10am Garden Salon 2