The Otter Annotation System

James Gilbert

jgrg@sanger.ac.uk
Man versus Machine

- Ensembl and similar systems provide excellent even coverage of genome
- (good) human annotator still wins gene by gene
Sanger annotation

• Annotating:
  our ⅓ of human + MHC haplotypes + encode mouse (chromosomes 2, 4, 11, X)
  all of zebrafish
  miscellaneous other vertebrates

• Available from Vega, Ensembl, EMBL / Genbank
  http://vega.sanger.ac.uk

• Import quality annotation (GTF or XML)
fox - the old system

- Stored in aceddb format
- Annotated clone by clone
- Transcripts that spanned several clones were fused during import into an Ensembl database:

Continued_from
Continues_as
otter - the new system

• still uses acedb xace front end on a local database, now driven by perl/Tk UI
• annotation stored in extended Ensembl schema
• annotators edit contiguous region of a chromosome
• improved viewing of gapped alignments
otter XML

<otter>
  <sequence_set>
    <sequence_fragment>
      <accession>
        <locus>
          <transcript>
            <exon>
            <feature_set>
          <stable_id>
          <author>
  <start> <end> <strand>

http://www.sanger.ac.uk/~jgrg/otter_xml.html
otterlace - Datasets

cat
cat
chimp
dog
human
misc_human
misc_mouse
mouse
pig
platypus
rat
zebrafish
Sequence Sets

<table>
<thead>
<tr>
<th>Sequence Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr9_09</td>
<td>chr9_09</td>
</tr>
<tr>
<td>chr9_hetero</td>
<td>Chr9 Heterochromatic region</td>
</tr>
<tr>
<td>chr9p12</td>
<td>test region of 9 clones on human chr 9 p12</td>
</tr>
<tr>
<td><strong>chr10_06</strong></td>
<td><strong>Human Chromosome 10 in SANGER 06 Assembly</strong></td>
</tr>
<tr>
<td>chr11-hsa11_20030701</td>
<td>Chr11 hsa11_20030701.agp</td>
</tr>
<tr>
<td>chr11-hsa11_20040402</td>
<td>Chr11 hsa11_20040402.agp</td>
</tr>
<tr>
<td>chr13-10</td>
<td>Human Chr13-10 AGP</td>
</tr>
<tr>
<td>chr14</td>
<td>Human Chromosome 14</td>
</tr>
<tr>
<td>chr19</td>
<td>Human Chr19 NCBI34</td>
</tr>
<tr>
<td>chr20</td>
<td>Human Chr20 from chromoview 31/03/04</td>
</tr>
<tr>
<td>chr20-9</td>
<td>Human Chr20-09 AGP</td>
</tr>
<tr>
<td>chr21-hsa21_20030701</td>
<td>Human Chromosome 21 20030701</td>
</tr>
<tr>
<td>chr22</td>
<td>Human Chr22 AGP 13/6/20024</td>
</tr>
<tr>
<td>chr22-02</td>
<td>chr 22 AGP corresponds to NCBI 35</td>
</tr>
<tr>
<td>chr_Y_01</td>
<td>Y agp from Kym Pepin</td>
</tr>
<tr>
<td>chrX-05</td>
<td>version 5 of human Chromosome X AGP</td>
</tr>
<tr>
<td>chrX-06</td>
<td>version 6 of Human Chromosome X AGP</td>
</tr>
<tr>
<td>ChrX-07</td>
<td>ChrX-07 AGP</td>
</tr>
<tr>
<td>encode_agp</td>
<td>415 clones of encode set (ncbi34)</td>
</tr>
<tr>
<td>encode_clones</td>
<td>415 clones of encode set (ncbi34)</td>
</tr>
<tr>
<td>GABRB3</td>
<td>clones containing the GABRB3 gene for Guy</td>
</tr>
<tr>
<td>geneset_23</td>
<td>23 transcripts for geneset evaluation</td>
</tr>
<tr>
<td>geneset_23_agp</td>
<td>23 transcripts for geneset evaluation</td>
</tr>
</tbody>
</table>
## Sequence Notes

![SequenceSetChr1_final](image)

<table>
<thead>
<tr>
<th>ID</th>
<th>Reference</th>
<th>Name</th>
<th>Status</th>
<th>Date</th>
<th>Accession</th>
<th>Check</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>1266</td>
<td>AL365260.11</td>
<td>RP11-433J22</td>
<td>complete</td>
<td>2004-06-25</td>
<td>hks</td>
<td>Checking</td>
<td>Charlie</td>
</tr>
<tr>
<td>1267</td>
<td>AL445591.10</td>
<td>RP11-314N2</td>
<td>complete</td>
<td>2004-07-20</td>
<td>ds3</td>
<td>checking</td>
<td>Charlie</td>
</tr>
<tr>
<td>1268</td>
<td>BX537254.7</td>
<td>RP6-7406</td>
<td>complete</td>
<td>2004-07-20</td>
<td>ds3</td>
<td>checking</td>
<td>Charlie</td>
</tr>
<tr>
<td>1269</td>
<td>BX842679.19</td>
<td>XXyac-YX155B6</td>
<td>complete</td>
<td>2004-07-20</td>
<td>cas</td>
<td>Charlie</td>
<td>Charlie</td>
</tr>
<tr>
<td>1270</td>
<td>AL451043.14</td>
<td>RP11-301M17</td>
<td>complete</td>
<td>2004-06-16</td>
<td>cas</td>
<td>Charlie</td>
<td>Charlie</td>
</tr>
<tr>
<td>1271</td>
<td>AL592207.9</td>
<td>RP11-495P10</td>
<td>complete</td>
<td>2004-06-16</td>
<td>cas</td>
<td>Charlie</td>
<td>Charlie</td>
</tr>
<tr>
<td>1272</td>
<td>AL691471.3</td>
<td>RP11-91G11</td>
<td>complete</td>
<td>2004-06-16</td>
<td>cas</td>
<td>Charlie</td>
<td>Charlie</td>
</tr>
<tr>
<td>1273</td>
<td>AL022240.8</td>
<td>RP3-328E19</td>
<td>complete</td>
<td>2004-06-16</td>
<td>cas</td>
<td>Charlie</td>
<td>Charlie</td>
</tr>
</tbody>
</table>

**GAP (100,000 bp)**

<table>
<thead>
<tr>
<th>ID</th>
<th>Reference</th>
<th>Name</th>
<th>Status</th>
<th>Date</th>
<th>Accession</th>
<th>Check</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>1274</td>
<td>BX546486.21</td>
<td>RP11-89F3</td>
<td>missing</td>
<td>2004-07-13</td>
<td>ds3</td>
<td>checking</td>
<td>Charlie</td>
</tr>
</tbody>
</table>

**GAP (50,000 bp)**

<table>
<thead>
<tr>
<th>ID</th>
<th>Reference</th>
<th>Name</th>
<th>Status</th>
<th>Date</th>
<th>Accession</th>
<th>Check</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>1275</td>
<td>AL954711.3</td>
<td>RP11-666A1</td>
<td>complete</td>
<td>2004-06-16</td>
<td>cas</td>
<td>Charlie</td>
<td>Charlie</td>
</tr>
</tbody>
</table>

**GAP (50,000 bp)**

<table>
<thead>
<tr>
<th>ID</th>
<th>Reference</th>
<th>Name</th>
<th>Status</th>
<th>Date</th>
<th>Accession</th>
<th>Check</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>1276</td>
<td>AL592492.10</td>
<td>RP11-763B22</td>
<td>complete</td>
<td>2004-07-20</td>
<td>ds3</td>
<td>checking</td>
<td>Charlie</td>
</tr>
<tr>
<td>1277</td>
<td>AL513526.19</td>
<td>RP11-14N7</td>
<td>complete</td>
<td>2004-06-16</td>
<td>cas</td>
<td>Charlie</td>
<td>Charlie</td>
</tr>
<tr>
<td>1278</td>
<td>AL663102.4</td>
<td>RP11-427C16</td>
<td>complete</td>
<td>2004-06-16</td>
<td>cas</td>
<td>Charlie</td>
<td>Charlie</td>
</tr>
</tbody>
</table>
Transcript editor

<table>
<thead>
<tr>
<th>File</th>
<th>Show</th>
<th>SubSeq</th>
<th>PolyA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL844549.1-005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL844549.1-006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL844549.1-007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fgenesh.3</td>
<td>PF06758.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genscan.7</td>
<td>PF06758.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genscan.8</td>
<td>PF06758.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genscan.9</td>
<td>PF06758.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF00595.1</td>
<td>PF06758.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF00595.2</td>
<td><strong>XXyac-YX155B6.1-001</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF00595.3</td>
<td><strong>XXyac-YX155B6.1-002</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Image of transcript editor window with data and annotations]
Improvements

- Schema glitches
- Ensemble core schema + API catchup
- Smarter filtering of vast amount of data presented to annotator
- Speed – GUI and feature fetching
- Easy to install externally and tunnel over SSH
- Acedb replacement + Gtk
Acknowledgments

Roy Storey                Havana
Mike Croning              Ensembl
Chao-Kung Chen            Ensembl
Patrick Meidl             ACeDB
Steve Trevanion           ACeDB
Tim Hubbard               ACeDB

Steve Searle *           ISG
Michele Clamp