

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

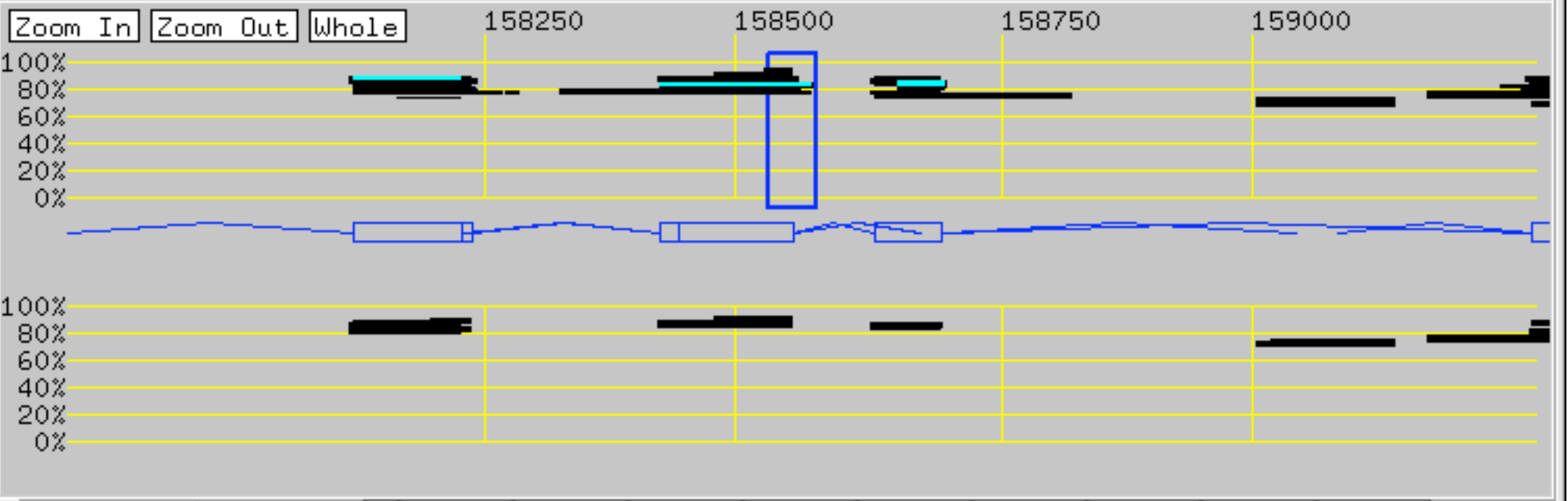
- Stored in acedb format
- Annotated clone by clone
- Transcripts that spanned several clones were fused during import into an Ensembl database:

Continued_from
Continues_as

otter - the new system

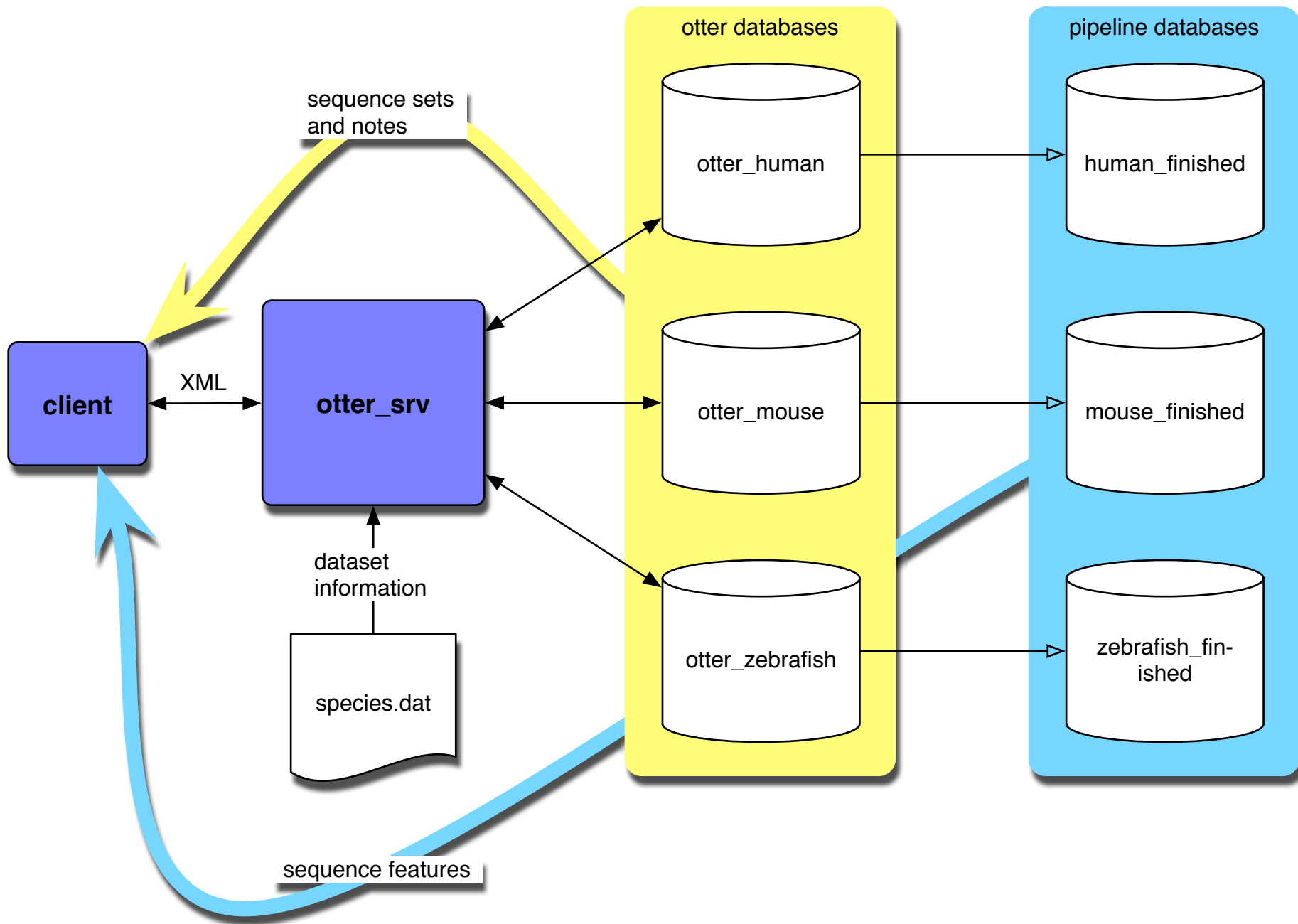
- still uses acedb xace front end on a local database, now driven by perl/Tk UI
- annotation stored in extended Ensembl schema
- annotators edit contiguous region of a chromosome
- improved viewing of gapped alignments

Blixem (nucleotide alignment): MANY.10000-743116



Sort HSPs by: identity Settings Goto < match match > << >> < > Strand^v EM:BI526

	Score	%ID	Start		End
MANY.10<>-743116(+)			158518	atgcagatgtccctgcagggagcaggtgagccagagcctgggtggtgc	158565
BI415285.1	81	86	761	atgcagatttccct.ca.ggaacagg	853
CK023426.1	79	86	711	.tgca.a.gt.cctgca.ggaacaagt	833
BY600752.1	86	86	1	atgcagatgtccctgcagggaaacag	58
BI408922.1	84	86	745	atgcagatgtccctgca.ggaacag	876
BI526698.1	81	85	568	atgcagatgtccctgcagggaaacaggtgagccaaagaatgggtg	727
BG963107.1	75	83	778	atg.aaat.tgcctgca.ggaccaagttagccaga	919
BI687156.1	78	80	57	atgcagatgtccctgcagggaaacag	284
BF608084.1	76	79	536	atccaaatttctnctncaaaaaaetttaeccaaaacc.eecte	687
MANY.10<>-743116(-)			158518	tacgtctacagggacgtccctcgtccactcgggtctcggaccaccacg	158565
AW545539.1	93	93	401	tacgtctacagggacgtcccttgtc	347
BG073583.2	92	92	425	tacgtctacagggacgtcccttgtc	348
AW541266.2	90	90	480	tacgtctacagggacgtcccttgtc	347
BG145138.1	89	89	387	tacgtctacagggacgtcccttgtc	256
BI438127.1	89	89	492	tacgtctacagggacgtcccttgtc	361
CF585021.1	89	89	508	tacgtctacagggacgtcccttgtc	377
AW536157.1	89	89	478	tacgtctacagggacgtcccttgtc	347
BF722546.1	89	89	497	tacgtctacagggacgtcccttgtc	256



otter XML

```
<otter>  
  <sequence_set>  
    <sequence_fragment>  
      <accession>  
    <locus>  
      <transcript>  
        <exon>  
      <feature_set>
```

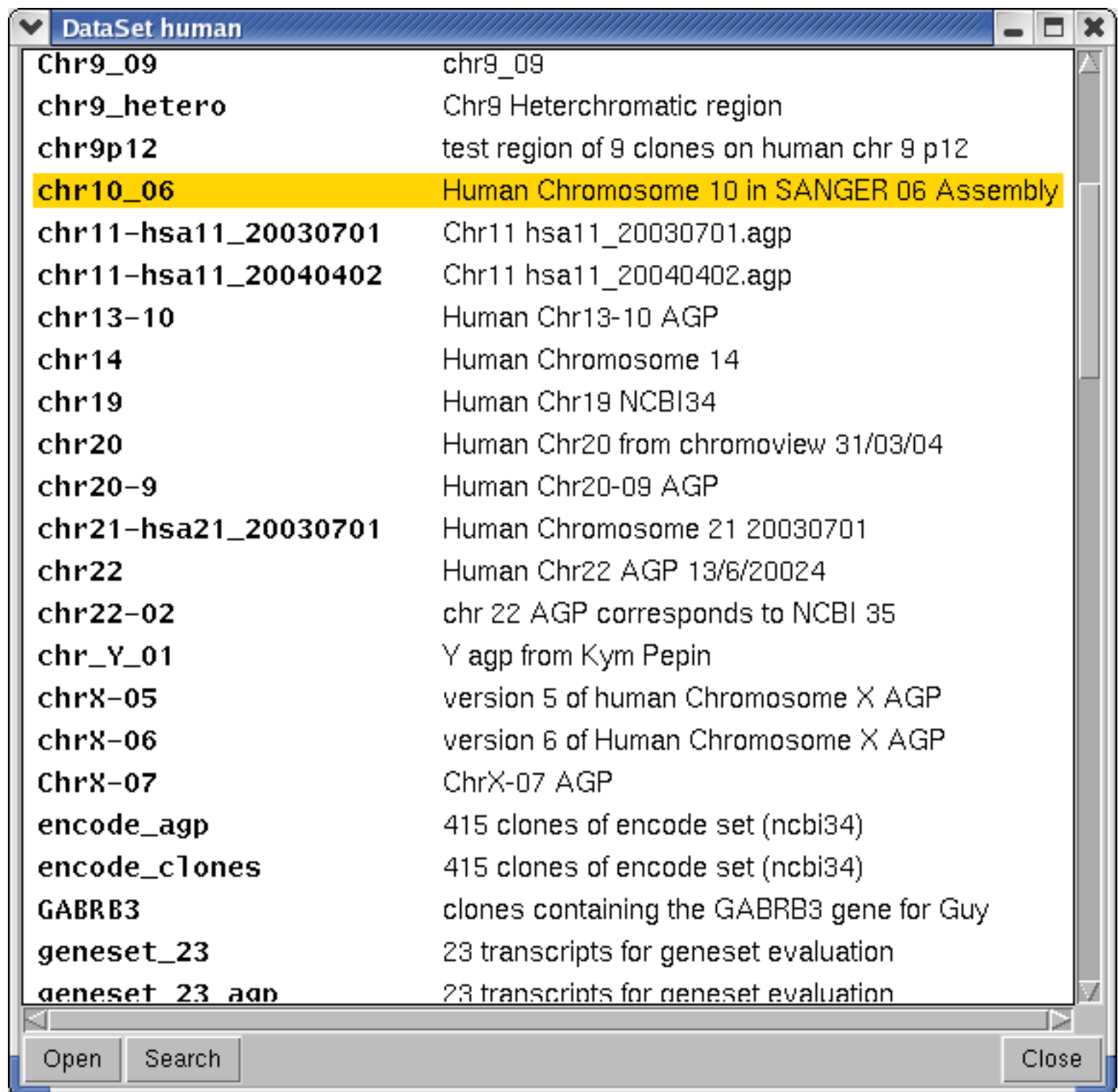
```
<stable_id>  
<author>  
<start> <end> <strand>
```

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

otterlace - Datasets



Sequence Sets



Sequence Set Name	Description
Chr9_09	chr9_09
chr9_hetero	Chr9 Heterchromatic region
chr9p12	test region of 9 clones on human chr 9 p12
chr10_06	Human Chromosome 10 in SANGER 06 Assembly
chr11-hsa11_20030701	Chr11 hsa11_20030701.agp
chr11-hsa11_20040402	Chr11 hsa11_20040402.agp
chr13-10	Human Chr13-10 AGP
chr14	Human Chromosome 14
chr19	Human Chr19 NCBI34
chr20	Human Chr20 from chromoview 31/03/04
chr20-9	Human Chr20-09 AGP
chr21-hsa21_20030701	Human Chromosome 21 20030701
chr22	Human Chr22 AGP 13/6/20024
chr22-02	chr 22 AGP corresponds to NCBI 35
chr_Y_01	Y agp from Kym Pepin
chrX-05	version 5 of human Chromosome X AGP
chrX-06	version 6 of Human Chromosome X AGP
ChrX-07	ChrX-07 AGP
encode_agp	415 clones of encode set (ncbi34)
encode_cTones	415 clones of encode set (ncbi34)
GABRB3	clones containing the GABRB3 gene for Guy
geneset_23	23 transcripts for geneset evaluation
geneset_23_agp	23 transcripts for geneset evaluation

Open Search Close

Sequence Notes

ID	Accession	RP	Status	Date	User	Notes	Lock
1266	AL365260.11	RP11-433J22	complete	2004-06-25	hks	Checking checked - hks	
1267	AL445591.10	RP11-314N2	complete	2004-07-20	ds3	checking checked	
1268	BX537254.7	RP6-7406	complete	2004-07-20	ds3	checking checked	🔒
1269	BX842679.19	XXyac-YX155B6	complete	2004-07-20	ds3	checking checked	🔒
1270	AL451043.14	RP11-301M17	complete	2004-06-16	cas	Charlie checked	🔒
1271	AL592207.9	RP11-495P10	complete	2004-06-16	cas	Charlie checked	🔒
1272	AL691471.3	RP11-91G11	complete	2004-06-16	cas	Charlie checked	
1273	AL022240.8	RP3-328E19	complete	2004-06-16	cas	Charlie checked	
GAP (100 000 bp)							
1274	BX546486.21	RP11-89F3	missing	2004-07-13	ds3	checking checked	
GAP (50 000 bp)							
1275	AL954711.3	RP11-666A1	complete	2004-06-16	cas	Charlie checked	
GAP (50 000 bp)							
1276	AL592492.10	RP11-763B22	complete	2004-07-20	ds3	checking checked	
1277	AL513526.19	RP11-14N7	complete	2004-06-16	cas	Charlie checked	
1278	AL663102.4	RP11-427C16	complete	2004-06-16	cas	Charlie checked	

Note text:

Transcript editor

lace Chr1_final, clones 1268..1271

File	Show	SubSeq	PolyA
AL844549.1-005		Fgenesh.7	Genscan.15
AL844549.1-006		Genscan.13	XXyac-YX155B6.5-001
AL844549.1-007		PF06758.1	XXyac-YX155B6.5-002
Fgenesh.3		PF06758.2	
Genscan.7		PF06758.3	
		PF06758.4	
Fgenesh.4		PF06758.5	
Genscan.8		PF06758.6	
Genscan.9		PF06758.7	
PF00595.1		PF06758.8	
PF00595.2		XXyac-YX155B6.1-001	
PF00595.3		XXyac-YX155B6.1-002	

XXyac-YX155B6.1-001

Start	End	Start	End
193267	204184	204093	170276
	202195		202054
	201629		201506
	193674		193527
	193424		193215

Type: Coding

Name: XXyac-YX155B6.1-001

Locus: XXyac-YX155B6.1

Start: Found End: Not found

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	
Chao-Kung Chen	
Patrick Meidl	Ensembl
Steve Trevanion	
Tim Hubbard	ACeDB
Steve Searle *	
Michele Clamp	ISG

* Paper: [Genome Research 2004 May;14\(5\):963-70.](#)