Tools for Community Genome Annotation: Zmap and Otter

Jane Loveland

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

IWGSC – Standards and Protocols
PAG XXV, Tuesday 17th January 2017

Havana: Human and vertebrate analysis and annotation

- Manual annotation of human, mouse, zebrafish, pig and rat whole chromosomes or genomes
- Human GENCODE annotation and working on mouse GENCODE annotation
- Annotation of specific regions: human MHC & LRC haplotypes, multiple species MHCs & LRCs,

Vega: Vertebrate Genome Annotation

- Ensembl derived browser focusing on manual annotation

Overview

- Manual annotation process
  - Tools, data, biotypes
- Community Manual Annotation
  - Mouse, Swine autosomes (IRAG), Rat, Chicken
- New data and projects

- Manual annotation process
  - Tools, data, biotypes
- Community Manual Annotation
  - Mouse, Swine autosomes (IRAG), Rat, Chicken
- New data and projects

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

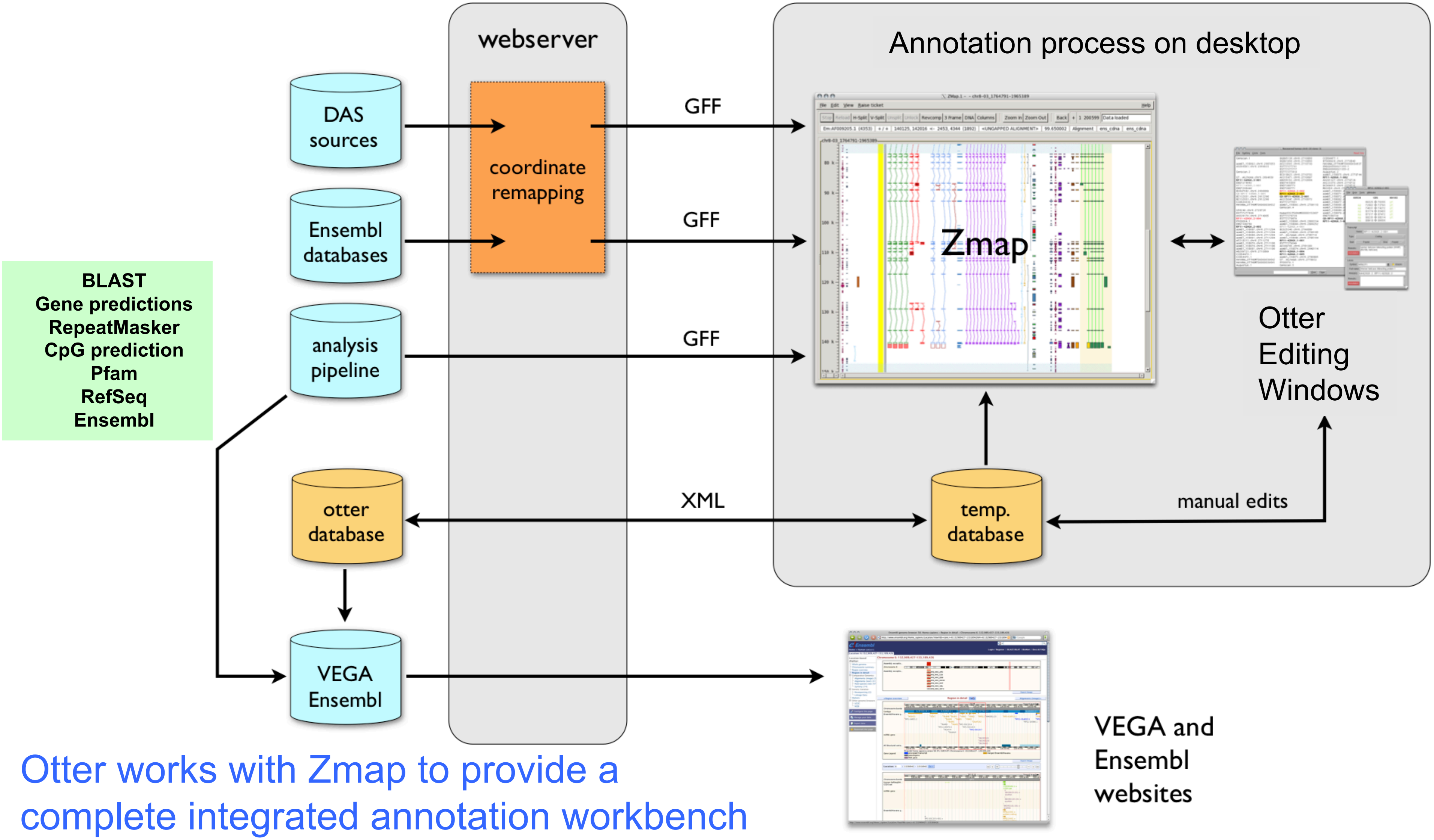
- Slow~3 months per chromosome
- Prefer finished (high quality) sequence
- Flexible, can deal with inconsistencies in data
- Most rules have exception
- Consult publications as well as databases
- Extensive Biotypes:
  - Excellent functional annotation
  - e.g. pseudogenes, lncRNA

Automated annotation alone is not sufficient for researchers needs

GENCODE geneset

Analysis and Annotation pipeline: Otter/ZMap

Annotators can work anywhere as software communicates with server over HTTP
The server controls data access and allows multi-user annotation



BLAST
Gene predictions
RepeatMasker
CpG prediction
Pfam
RefSeq
Ensembl

webserver

Annotation process on desktop

DAS sources

Ensembl databases

analysis pipeline

coordinate remapping

GFF

GFF

GFF

Zmap

Otter Editing Windows

otter database

XML

temp. database

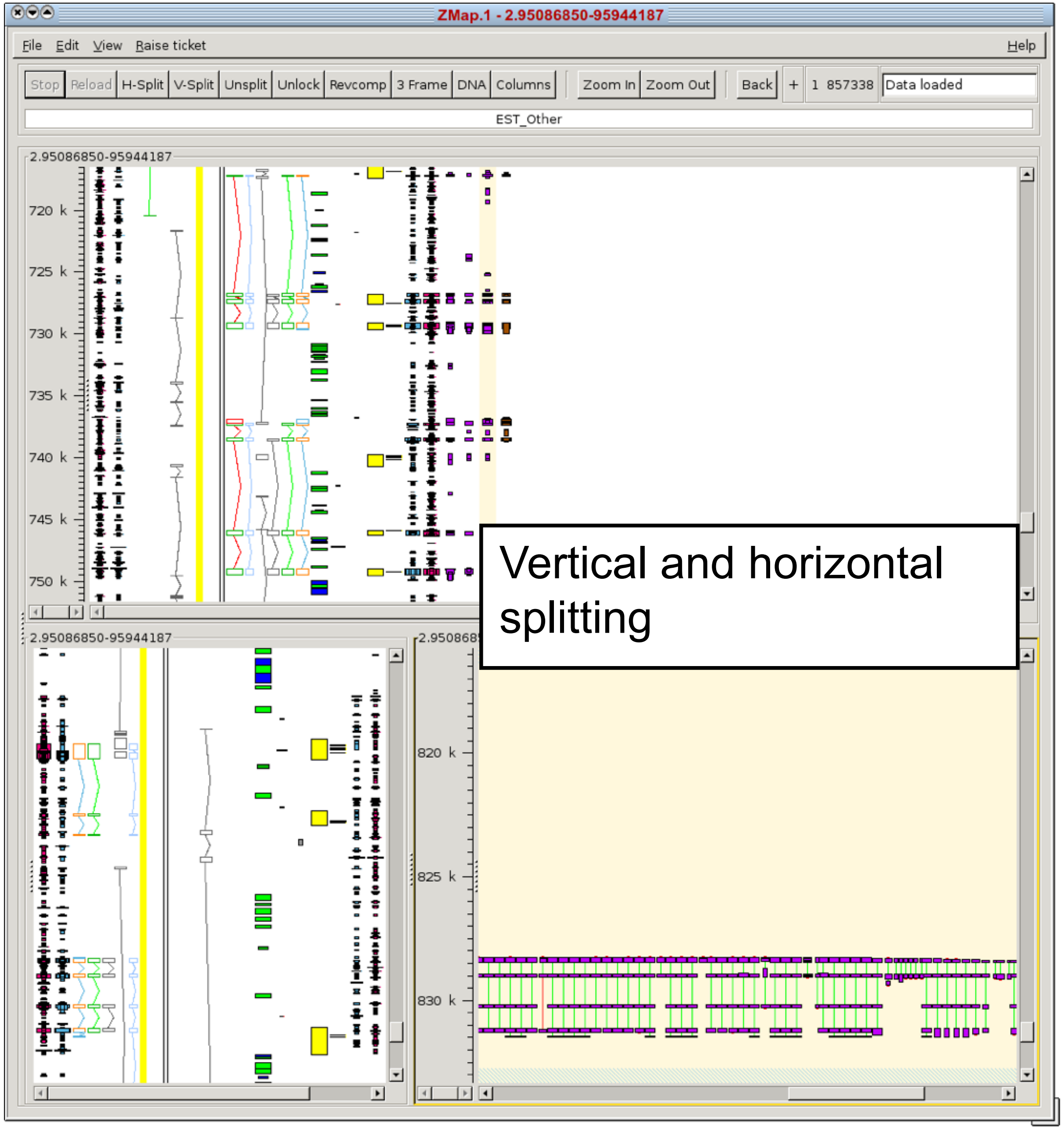
manual edits

VEGA Ensembl

VEGA and Ensembl websites

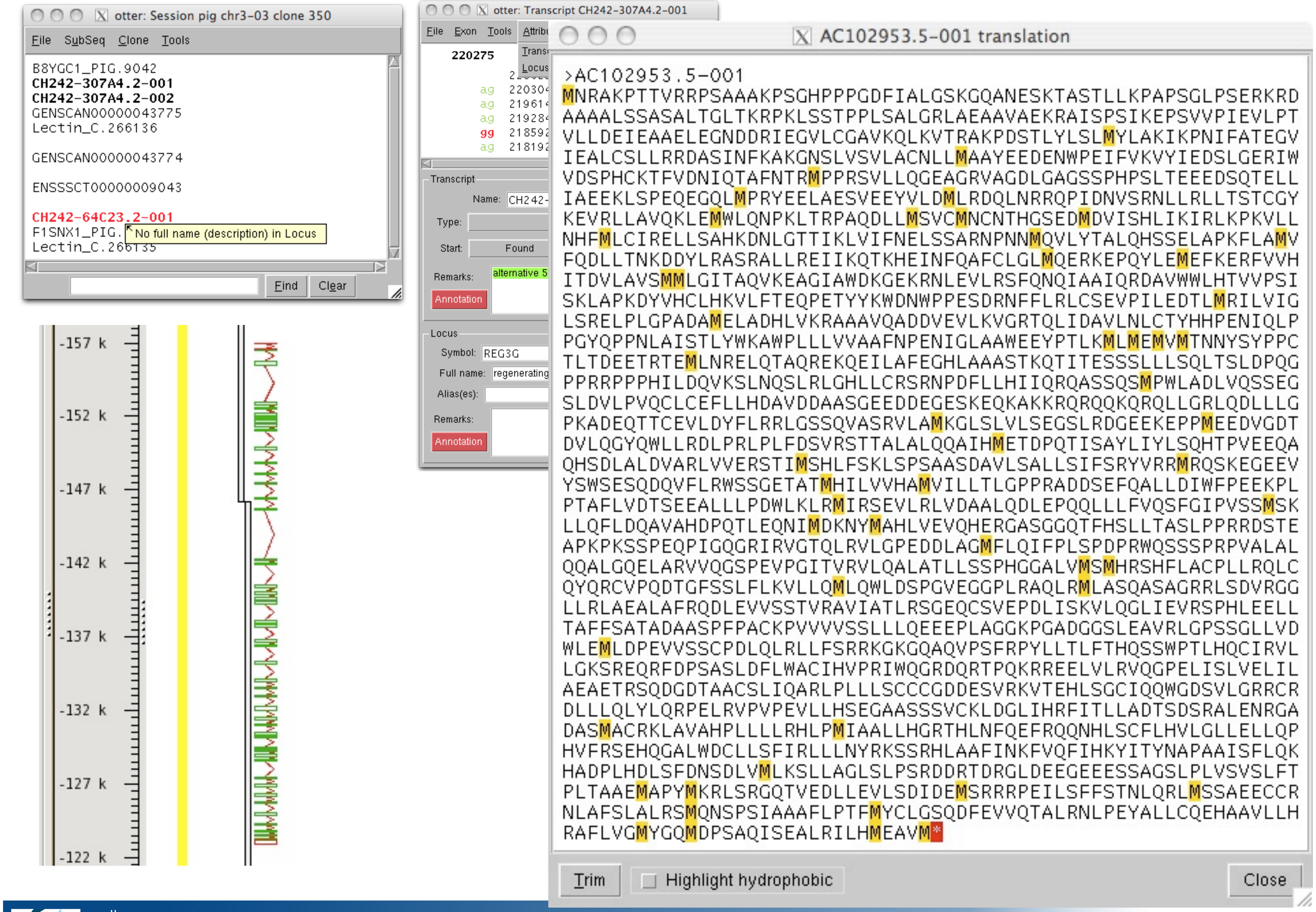
Otter works with Zmap to provide a complete integrated annotation workbench

Annotation tools: Zmap



Vertical and horizontal splitting

Annotation tools: Otter



otter: Session pig chr3-03 clone 350

```
B8YGC1_PIG.9042
CH242-307A4.2-001
CH242-307A4.2-002
GENSCAN00000043775
Lectin_C.266136

GENSCAN00000043774

ENSSSCT00000009043

CH242-64C23.2-001
F1SNX1_PIG.
Lectin_C.266135
```

No full name (description) in Locus

otter: Transcript CH242-307A4.2-001

```
220275
2
ag 22030
ag 21961
ag 21928
gg 21859
ag 21819
```

Transcript

Name: CH242-

Type:

Start: Found

Remarks: alternative 5

Locus

Symbol: REG3G

Full name: regenerating

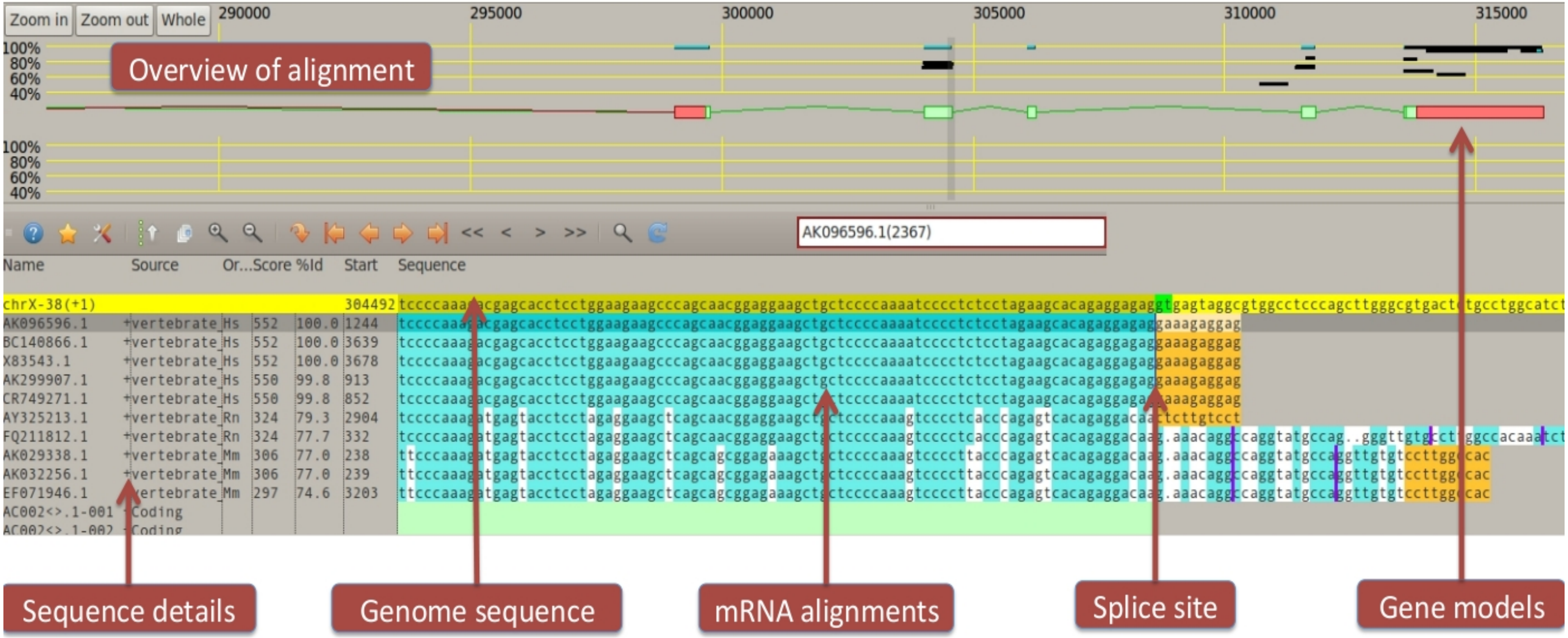
Alias(es):

Remarks:

AC102953.5-001 translation

```
>AC102953.5-001
MNRAKPTTVRRPSAAAKPSGHPPPGDFIALGSKGQANESKTASTLLKPAPSGLPSERKRD
AAAALSSASALTGLTKRPKLSSTPPLSALGRLAEAAVAEKRAISPSIKEPSVVPIEVLPT
VLLDEIEAAELEGNDDRIEGVLCGAVKQLKVTRAKPDSTLYLSLMYLAKIKPNIFATEGV
IEALCSLLRRDASINFKAKGNSLVSVLACNLLMAAYEEDENWPEIFVKVYIEDSLGERIW
VDSPHCKTFVDNIQTAFNTRMPPRSVLLQGEAGRVAGDLGAGSSPHPSLTEEEDSQTELL
IAEEKLSPEQEGQLMPRYEELAESVEEYVLDMLRDQLNRRQPIDNVSRNLLRLLTSTCGY
KEVRLLAVQKLEMWLQNPKLTRPAQDLLMSVCMNCNTHGSEDMDVISHLIKIRLKPKVLL
NHFMLCIRELLSAHKDNLGTTIKLVIFNELSSARNPNNMQVLYTALQHSSELAPKFLAMV
FQDLLTNKDDYLRASRALLREIIKQTKHEINFQAFCLGLMQERKEPQYLEMEFKERFVVH
ITDVLAVSMMLGITAQVKEAGIAWDKGEKRNLEVLRSFQNQIAAIQRDAVWWLHTVVPSI
SKLAPKDYVHCLHKVLFTEQPETYYKWDNWPPESDRNFFLRLCSEVPILEDTLMRILVIG
LSRELPLGPADAMELADHLVKRAAAVQADDVEVLKVGRTQLIDAVLNLCTYHHPENIQLP
PGYQPPNLAISTLYWKAWPLLLVVAAFNPENIGLAAWEEYPTLKMLMEMVMTNNYSYPPC
TLTDEETRTEMLNRELQTAQREKQEILAFEGHLAAASTKQTITESSSLLLSQLTSLDPQG
PPRRPPPHILDQVKSLNQSLRLGHLLCRSRNPDFLLHIIQRQASSQSMPWLADLVQSSEG
SLDVLPVQCLCEFLLHDAVDDAASGEEDDEGESKEQKAKKRQRQQKQRQLLGRLQDLLLG
PKADEQTTCEVLDYFLRRLGSSQVASRVLAMKGLSLVLSEGSLRDGEEKEPPMEEDVGDT
DVLQGYQWLLRDLPRLPLFDSVRSTTALALQQAIHMETDPQTISAYLIYLSQHTPVEEQA
QHSDLALDVARLVVERSTIMSHLFSKLSPSAASDAVLSALLSIFSRYVRRMRQSKEGEEV
YSWSESQDQVFLRWSSGETATMHILVVHAMVILLTLGPPRADDSEFQALLDIWFPEEKPL
PTAFLVDTSEEALLLPDWLKLRMIRSEVLRLVDAALQDLEPQQLLLFVQSFGIPVSSMSK
LLQFLDQAVAHDPQTLEQNIMDKNYMAHLVEVQHERGASGGQTFHSLLTASLPPRRDSTE
APKPKSSPEQPIGQGRIRVGTQLRVLGPEDDLAGMFLQIFPLSPDPRWQSSSPRPVALAL
QQALGQELARVVQGSPEVPGITVRVLQALATLLSSPHGGALVMSMHRSHFLACPLLRQLC
QYQRCVPQDTGFSSLFLKVLLQMLQWLDSPGVEGGPLRAQLRMLASQASAGRRLSDVRGG
LLRLAEALAFRQDLEVVSSTVRAVIATLRSGEQCSVEPDLISKVLQGLIEVRSPHLEELL
TAFFSATADAASPFPACKPVVVVSSLLLQEEEPLAGGKPGADGGSLEAVRLGPSSGLLVD
WLEMLDPEVVSSCPDLQLRLLFSRRKGKGQAQVPSFRPYLLTLFTHQSSWPTLHQCIRVL
LGKSREQRFDPSASLDFLWACIHVPRIWQGRDQRTPQKRREELVLRVQGPELISLVELIL
AEAETRSQDGDTAACSLIQARLPLLLSCCCGDDESVRKVTEHLSGCIQQWGDSVLGRRCR
DLLLQLYLQRPELRVPVPEVLLHSEGAASSSVCKLDGLIHRFITLLADTSDSRALENRGA
DASMACRKLAVAHPLLLLRHLPMIAALLHGRTHLNFQEFRQQNHLSCFLHVLGLLELLQP
HVFRSEHQGALWDCLLSFIRLLLNYRKSSRHLAAFINKFVQFIHKYITYNAPAAISFLQK
HADPLHDLSFDNSDLVMLKSLLAGLSLPSRDDRTDRGLDEEGEEESSAGSLPLVSVSLFT
PLTAAEMAPYMKRLSRGQTVEDLLEVLSDIDEMSRRRPEILSFFSTNLQRLMSSAEECCR
NLAFSLALRSMQNSPSIAAAFLPTFMYCLGSQDFEVVQTALRNLPEYALLCQEHAAVLLH
RAFLVGMYGQMDPSAQISEALRILHMEAVM*
```

Annotation tools: Blixem



```
Name            Source          Or...Score %Id   Start   Sequence

chrX-38(+1)                                      304492 tccccaaagacgagcacctcctggaagaagcccagcaacggaggaagctgctccccaaaatcccctctcctagaagcacagaggagaggtgagtaggcgtggcctcccagcttgggcgtgactctgcctggcatct
AK096596.1      +vertebrate_Hs  552   100.0      1244   tccccaaagacgagcacctcctggaagaagcccagcaacggaggaagctgctccccaaaatcccctctcctagaagcacagaggagaggaaagaggag
BC140866.1      +vertebrate_Hs  552   100.0      3639   tccccaaagacgagcacctcctggaagaagcccagcaacggaggaagctgctccccaaaatcccctctcctagaagcacagaggagaggaaagaggag
X83543.1        +vertebrate_Hs  552   100.0      3678   tccccaaagacgagcacctcctggaagaagcccagcaacggaggaagctgctccccaaaatcccctctcctagaagcacagaggagaggaaagaggag
AK299907.1      +vertebrate_Hs  550   99.8       913    tccccaaagacgagcacctcctggaagaagcccagcaacggaggaagctgctccccaaaatcccctctcctagaagcacagaggagaggaaagaggag
CR749271.1      +vertebrate_Hs  550   99.8       852    tccccaaagacgagcacctcctggaagaagcccagcaacggaggaagctgctccccaaaatcccctctcctagaagcacagaggagagaaagaggag
AY325213.1      +vertebrate_Rn  324   79.3       2904   tccccaaagatgagtacctcctagaggaagctcagcaacggaggaagctgctccccaaagtcccctcacccagagtcacagaggacaactcttgtcct
FQ211812.1      +vertebrate_Rn  324   77.7       332    tccccaaagatgagtacctcctagaggaagctcagcaacggaggaagctgctccccaaagtcccctcacccagagtcacagaggacaag.aaacaggccaggtatgccag..gggttgtgccttggccacaaatct
AK029338.1      +vertebrate_Mm  306   77.0       238    ttcccaaagatgagtacctcctagaggaagctcagcagcggagaaagctgctccccaaagtccccttacccagagtcacagaggacaag.aaacaggccaggtatgccaggttgtgtccttggccac
AK032256.1      +vertebrate_Mm  306   77.0       239    ttcccaaagatgagtacctcctagaggaagctcagcagcggagaaagctgctccccaaagtccccttacccagagtcacagaggacaag.aaacaggccaggtatgccaggttgtgtccttggccac
EF071946.1      +vertebrate_Mm  297   74.6       3203   ttcccaaagatgagtacctcctagaggaagctcagcagcggagaaagctgctccccaaagtccccttacccagagtcacagaggacaag.aaacaggccaggtatgccaggttgtgtccttggccac
AC002<>.1-001   Coding
AC002<>.1-002   Coding
```

Overview of alignment

Sequence details

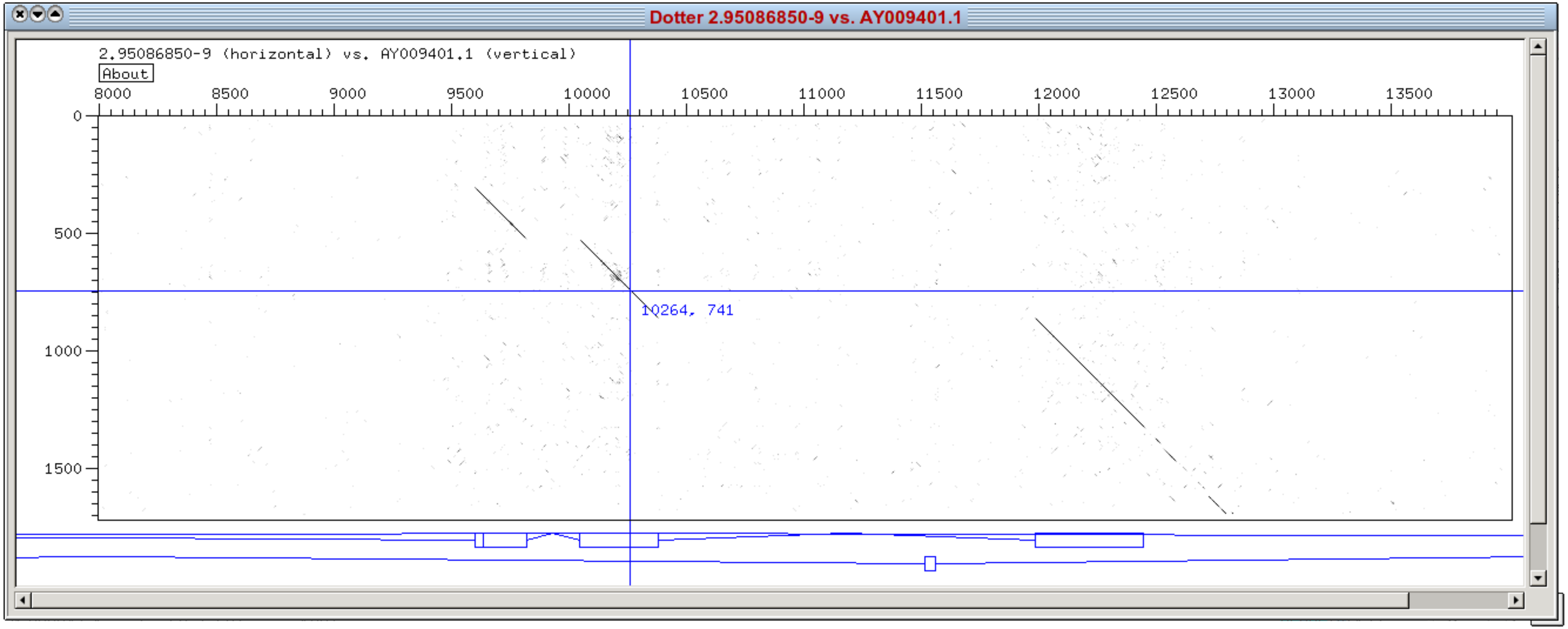
Genome sequence

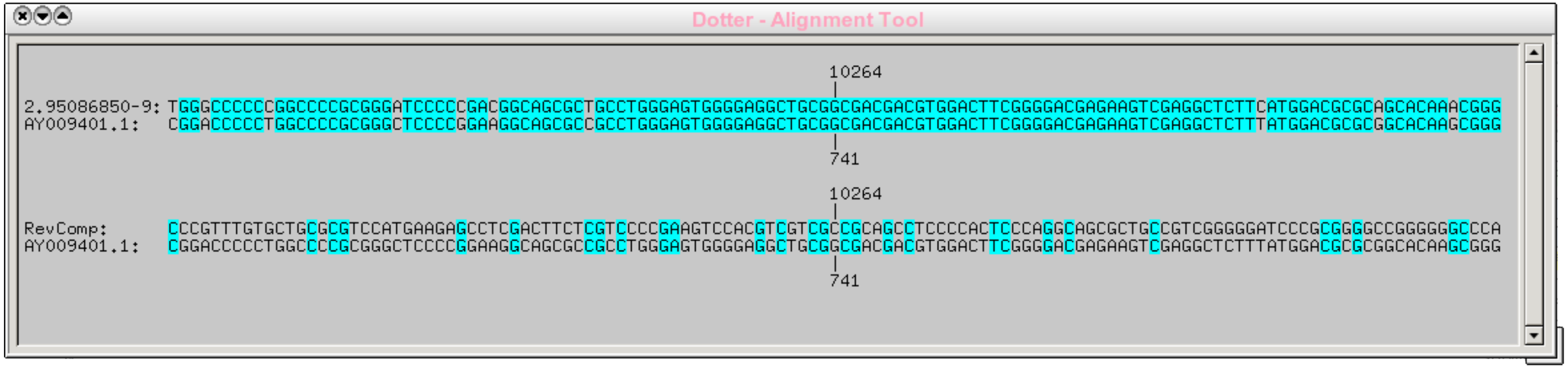
mRNA alignments

Splice site

Gene models

Annotation tools: Dotter


Dotter 2.95086850-9 vs. AY009401.1
2.95086850-9 (horizontal) vs. AY009401.1 (vertical)
About
8000
8500
9000
9500
10000
10500
11000
11500
12000
12500
13000
13500
0
500
1000
1500
10264, 741

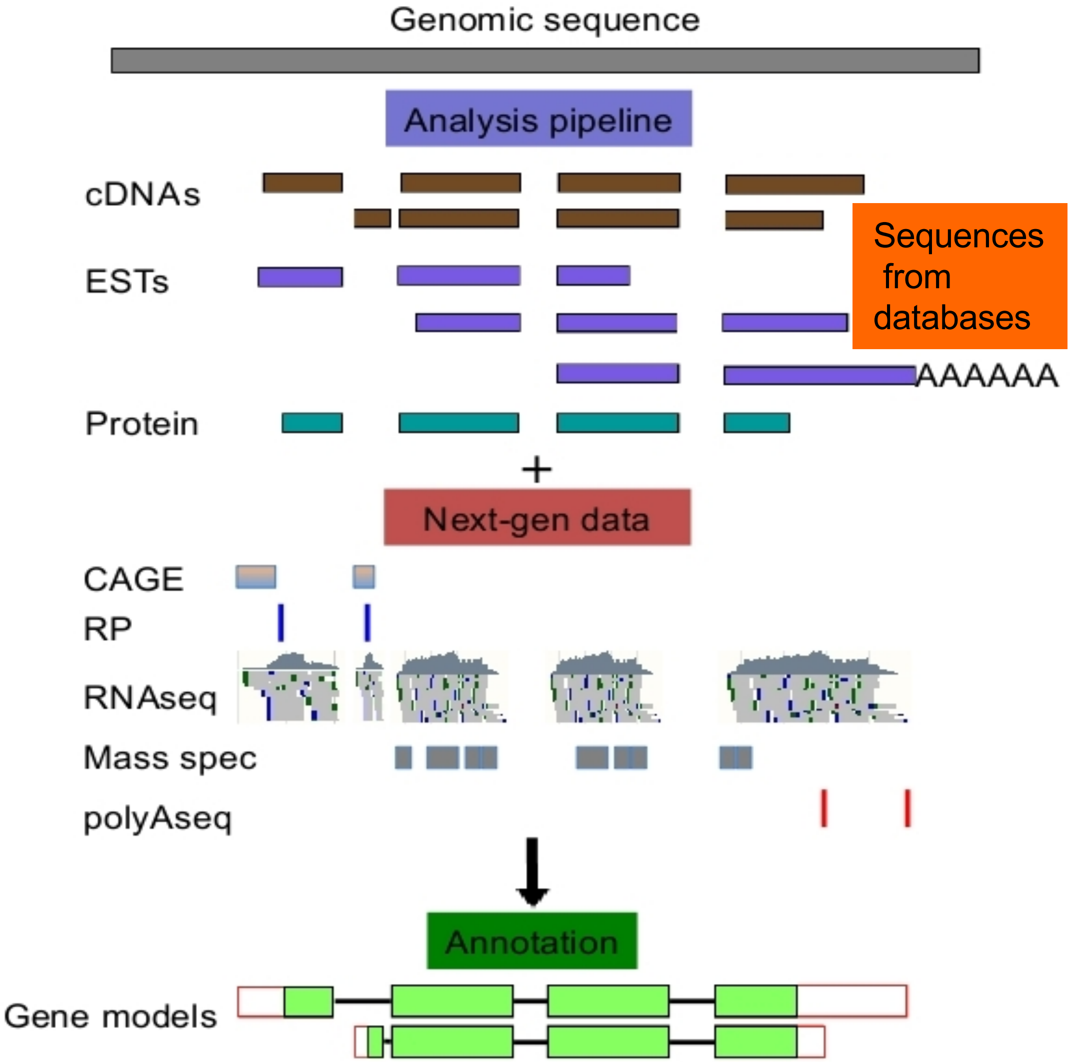


```
                                                                                     10264
                                                                                       |
2.95086850-9: TGGGCCCCCGGCCCCGCGGGATCCCCGACGGCAGCGCTGCCTGGGAGTGGGGAGGCTGCGGCGACGACGTGGACTTCGGGGACGAGAAGTCGAGGCTCTTCATGGACGCGCAGCACAAACGGG
AY009401.1:   CGGACCCCCTGGCCCCGCGGGCTCCCCGGAAGGCAGCGCCGCCTGGGAGTGGGGAGGCTGCGGCGACGACGTGGACTTCGGGGACGAGAAGTCGAGGCTCTTTATGGACGCGCGGCACAAGCGGG
                                                                                       |
                                                                                      741

                                                                                     10264
                                                                                       |
RevComp:      CCCGTTTGTGCTGCGCGTCCATGAAGAGCCTCGACTTCTCGTCCCCGAAGTCCACGTCGTCGCCGCAGCCTCCCCACTCCCAGGCAGCGCTGCCGTCGGGGATCCCGCGGGGCCGGGGGGCCCA
AY009401.1:   CGGACCCCCTGGCCCCGCGGGCTCCCCGGAAGGCAGCGCCGCCTGGGAGTGGGGAGGCTGCGGCGACGACGTGGACTTCGGGGACGAGAAGTCGAGGCTCTTTATGGACGCGCGGCACAAGCGGG
                                                                                       |
                                                                                      741
```

Manual Annotation: Biotypes

Annotation:
based on transcriptional evidence



Set of guidelines to help make annotation decisions

Biotypes

Protein Coding

- Known_CDS
- Novel_CDS
- Putative_CDS
- Nonsense_mediated_decay

Transcript retained intron
putative

Non-coding lincRNA
Antisense
Sense_intronic
Sense_overlapping
3'_overlapping_ncRNA

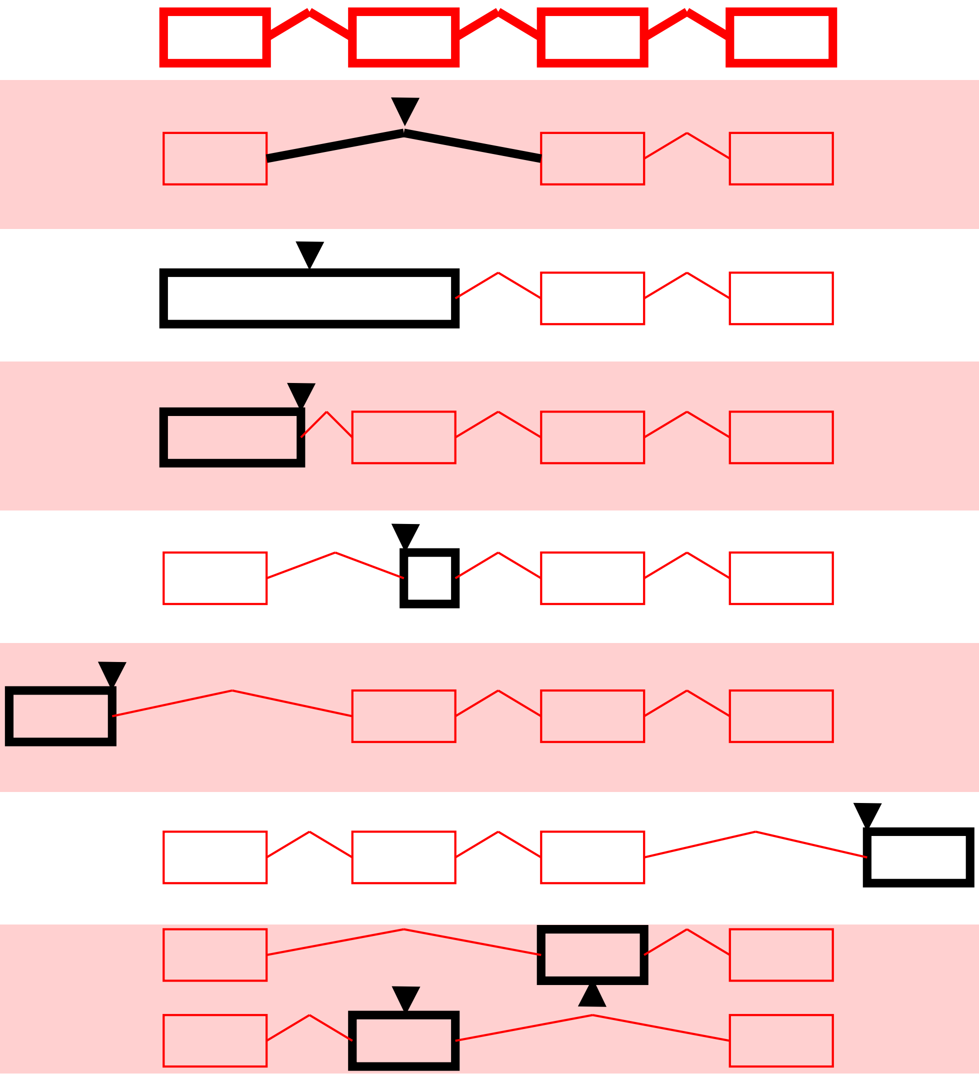
Pseudogene

- Processed
- Unprocessed
- Transcribed
- Translated
- Unitary
- Polymorphic

Immunoglobulin

- IG_pseudogene
- IG_Gene
- TR_Gene

Alternative Splicing:



Reference model

Skipped exon

Retained intron

Alternative splice donor

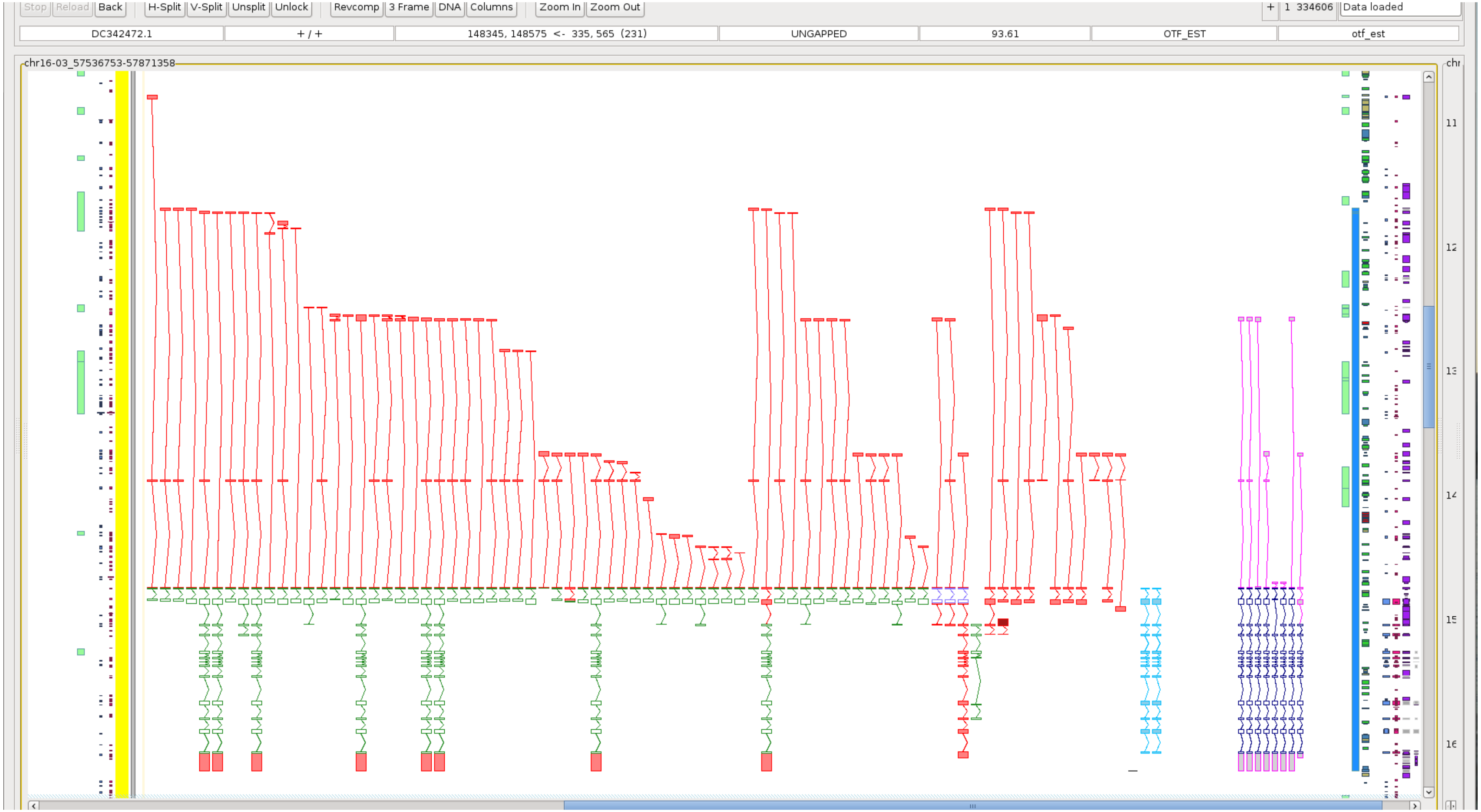
Alternative splice acceptor

Alternative first exon

Alternative final exon

Mutually exclusive

5’ end annotation:
GPR56 (Human G protein-coupled receptor 56 gene)



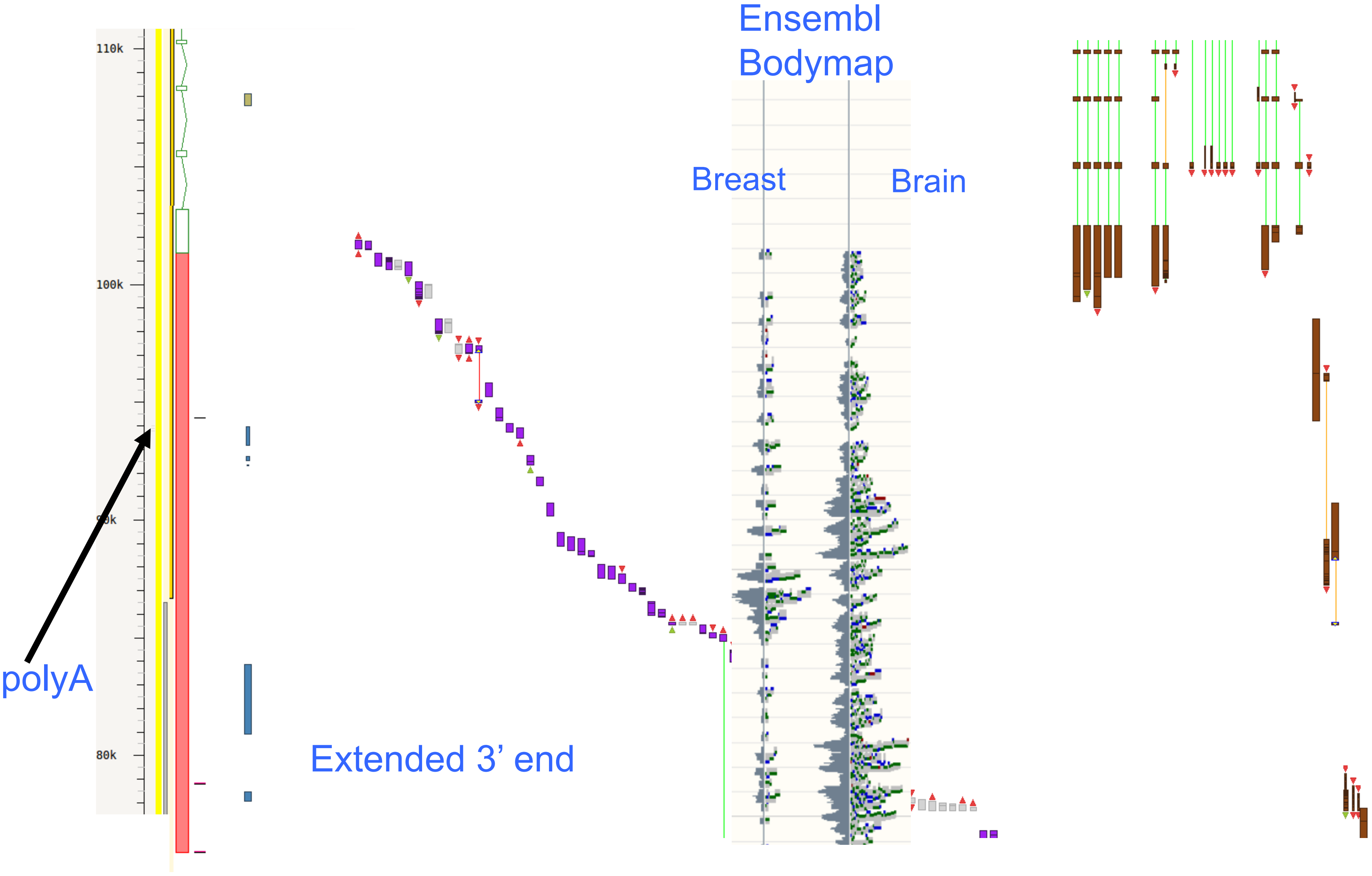
1 334606 | Data loaded

DC342472.1 | + / + | 148345, 148575 <- 335, 565 (231) | UNGAPPED | 93.61 | OTF_EST | otf_est

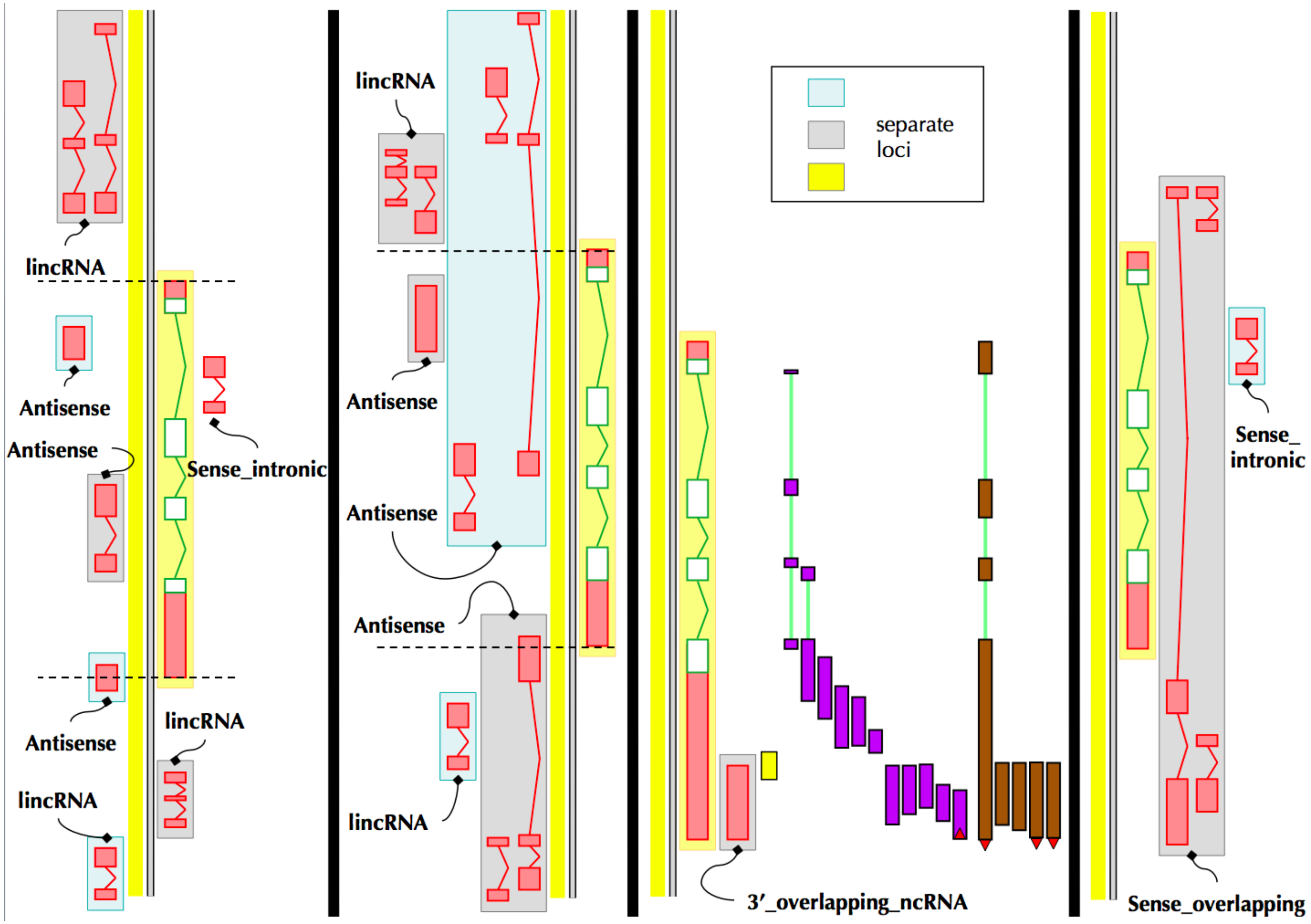
chr16-03_57536753-57871358

RNAseq data to extend 3'UTRs

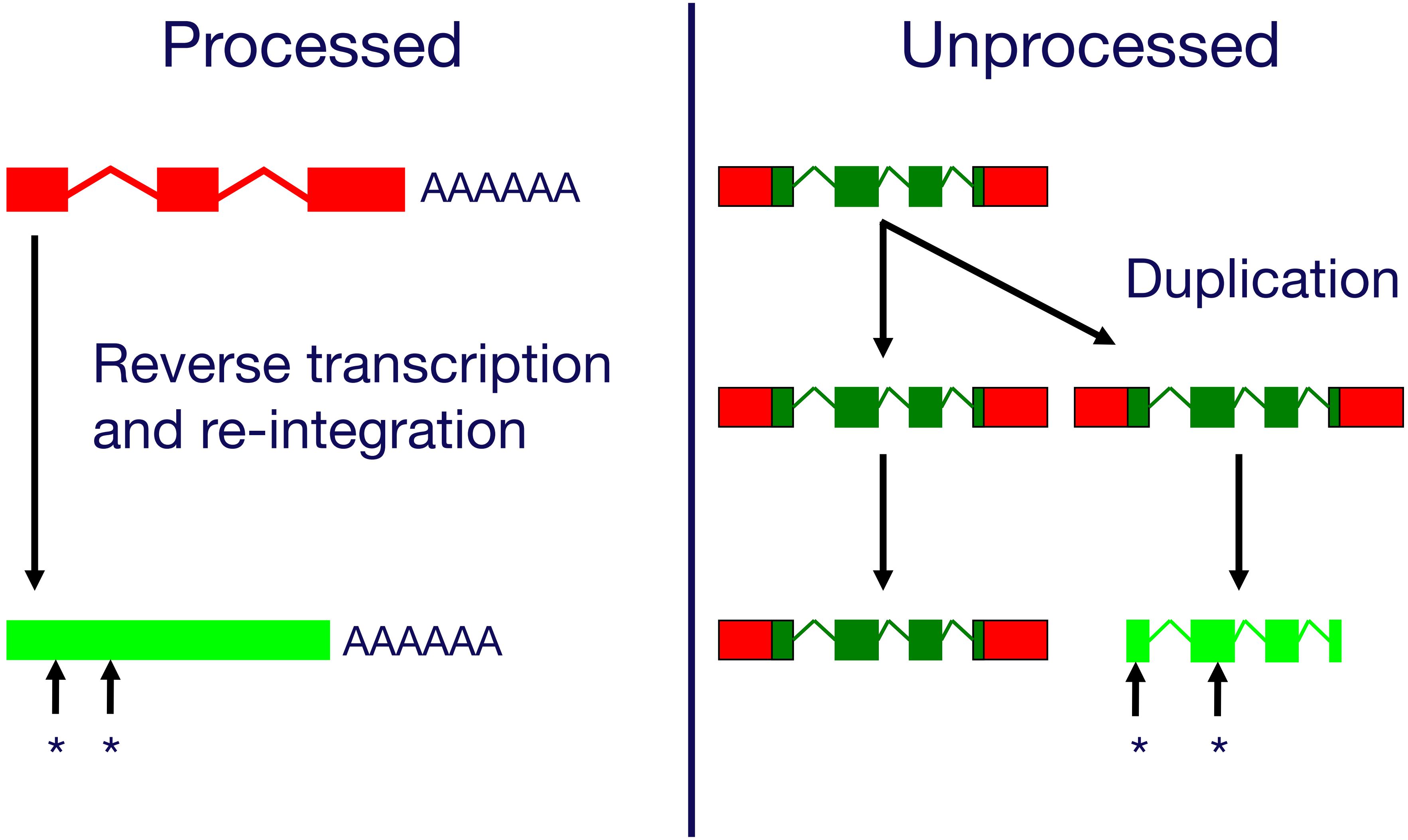
GRIN2B



Improvements of lncRNA annotation: understanding functionality



HAVANA Pseudogene Loci:



Processed

Unprocessed

AAAAAA

Reverse transcription and re-integration

AAAAAA

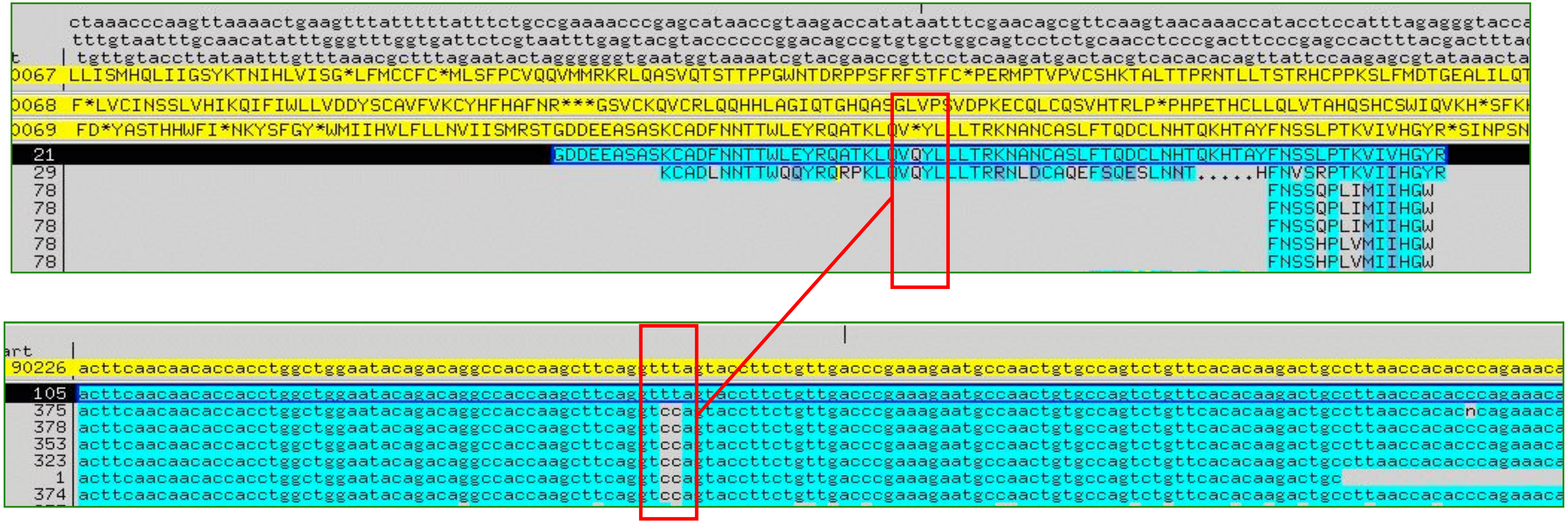
* *

Duplication

* *

Sequencing error, pseudogene or polymorphic pseudogene?

Polymorphic pseudogene

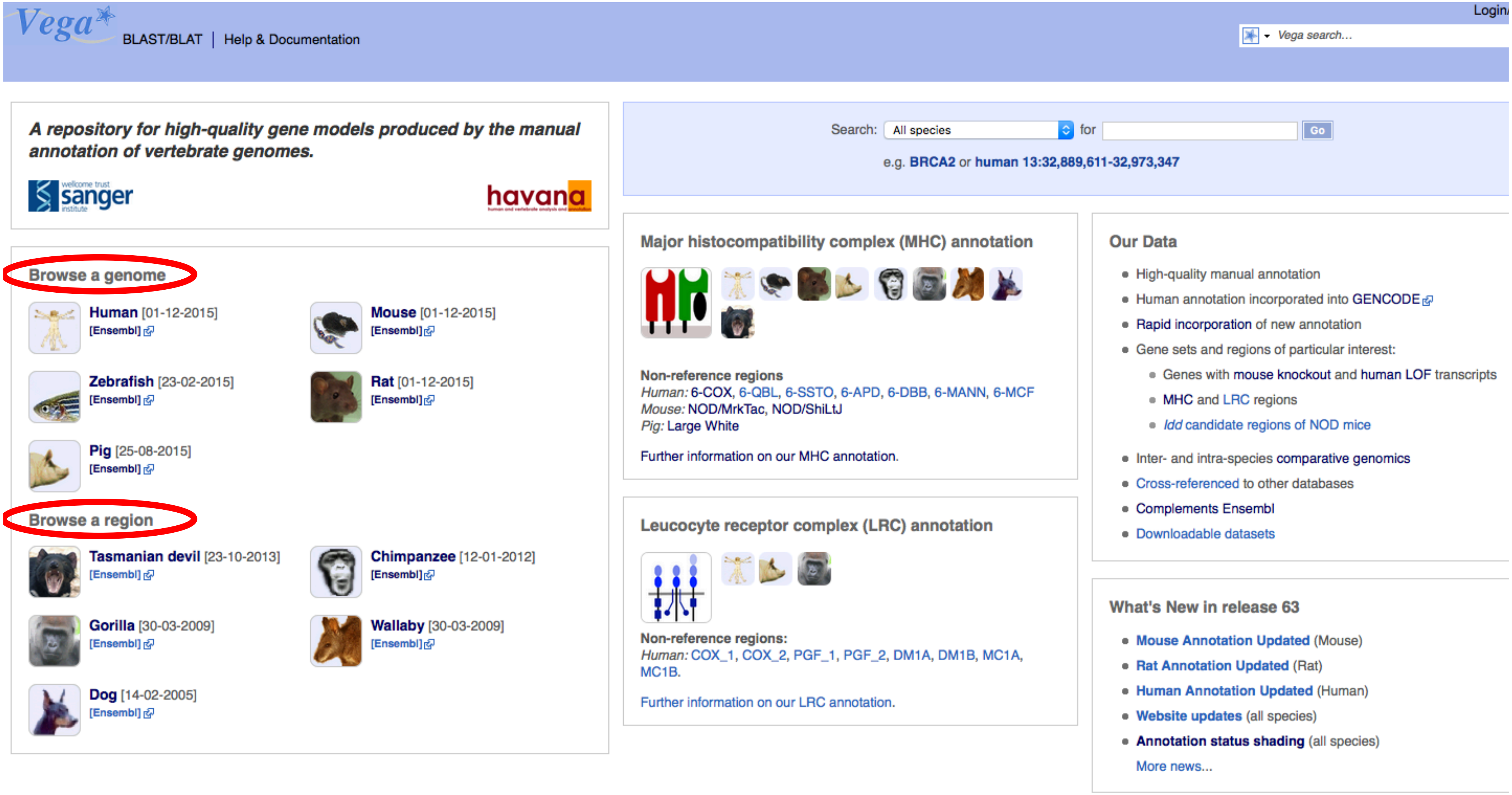


```
      ctaaacccaagttaaaactgaagtttatttttatttctgccgaaaacccgagcataaccgtaagaccatataatttcgaacagcgttcaagtaacaaaccatacctccatttagagggtacca
       tttgtaatttgcaacatatttgggtttggtgattctcgtaatttgagtacgtaccccccggacagccgtgtgctggcagtcctctgcaacctcccgacttcccgagccactttacgacttta
t     | tgttgtaccttataatttgtttaaacgctttagaatactagggggggtgaatggtaaaatcgtacgaaccgttcctacaagatgactacgtcacacacagttattccaagagcgtataaacta
0067 LLISMHQLIIGSYKTNIHLVISG*LFMCCFC*MLSFPCVQQVMMRKRLQASVQTSTTPPGWNTDRPPSFRFSTFC*PERMPTVPVCSHKTALTTPRNTLLTSTRHCPPKSLFMDTGEALILQT

0068  F*LVCINSSLVHIKQIFIWLLVDDYSCAVFVKCYHFHAFNR***GSVCKQVCRLQQHHLAGIQTGHQASGLVPSVDPKECQLCQSVHTRLP*PHPETHCLLQLVTAHQSHCSWIQVKH*SFKH
0069   FD*YASTHHWFI*NKYSFGY*WMIIHVLFLLNVIISMRSTGDDEEASASKCADFNNTTWLEYRQATKLQV*YLLLTRKNANCASLFTQDCLNHTQKHTAYFNSSLPTKVIVHGYR*SINPSN
  21                                                GDDEEASASKCADFNNTTWLEYRQATKLQVQYLLLTRKNANCASLFTQDCLNHTQKHTAYFNSSLPTKVIVHGYR
  29                                                         KCADLNNTTWQQYRQRPKLQVQYLLLTRRNLDCAQEFSQESLNNT.....HFNVSRPTKVIIHGYR
  78                                                                                                                FNSSQPLIMIIHGW
  78                                                                                                                FNSSQPLIMIIHGW
  78                                                                                                                FNSSQPLIMIIHGW
  78                                                                                                                FNSSHPLVMIIHGW
  78                                                                                                                FNSSHPLVMIIHGW
```

```
art   |
90226 acttcaacaacaccacctggctggaatacagacaggccaccaagcttcaggtttagtaccttctgttgacccgaaagaatgccaactgtgccagtctgttcacacaagactgccttaaccacacccagaaaca
  105 acttcaacaacaccacctggctggaatacagacaggccaccaagcttcaggtttagtaccttctgttgacccgaaagaatgccaactgtgccagtctgttcacacaagactgccttaaccacacccagaaaca
  375 acttcaacaacaccacctggctggaatacagacaggccaccaagcttcaggtccagtaccttctgttgacccgaaagaatgccaactgtgccagtctgttcacacaagactgccttaaccacacncagaaaca
  378 acttcaacaacaccacctggctggaatacagacaggccaccaagcttcaggtccagtaccttctgttgacccgaaagaatgccaactgtgccagtctgttcacacaagactgccttaaccacacccagaaaca
  353 acttcaacaacaccacctggctggaatacagacaggccaccaagcttcaggtccagtaccttctgttgacccgaaagaatgccaactgtgccagtctgttcacacaagactgccttaaccacacccagaaaca
  323 acttcaacaacaccacctggctggaatacagacaggccaccaagcttcaggtccagtaccttctgttgacccgaaagaatgccaactgtgccagtctgttcacacaagactgccttaaccacacccagaaaca
    1 acttcaacaacaccacctggctggaatacagacaggccaccaagcttcaggtccagtaccttctgttgacccgaaagaatgccaactgtgccagtctgttcacacaagactgc
  374 acttcaacaacaccacctggctggaatacagacaggccaccaagcttcaggtccagtaccttctgttgacccgaaagaatgccaactgtgccagtctgttcacacaagactgccttaaccacacccagaaaca
```

Loss of function gene



Vega
BLAST/BLAT | Help & Documentation
Login
Vega search...

A repository for high-quality gene models produced by the manual annotation of vertebrate genomes.

wellcome trust sanger institute
havana

Search: All species for Go

e.g. **BRCA2** or **human 13:32,889,611-32,973,347**

Browse a genome

Human [01-12-2015] [Ensembl]

Mouse [01-12-2015] [Ensembl]

Zebrafish [23-02-2015] [Ensembl]

Rat [01-12-2015] [Ensembl]

Pig [25-08-2015] [Ensembl]

Browse a region

Tasmanian devil [23-10-2013] [Ensembl]

Chimpanzee [12-01-2012] [Ensembl]

Gorilla [30-03-2009] [Ensembl]

Wallaby [30-03-2009] [Ensembl]

Dog [14-02-2005] [Ensembl]

Major histocompatibility complex (MHC) annotation

Non-reference regions
Human: 6-COX, 6-QBL, 6-SSTO, 6-APD, 6-DBB, 6-MANN, 6-MCF
Mouse: NOD/MrkTac, NOD/ShiLtJ
Pig: Large White

Further information on our MHC annotation.

Leucocyte receptor complex (LRC) annotation

Non-reference regions:
Human: COX_1, COX_2, PGF_1, PGF_2, DM1A, DM1B, MC1A, MC1B.

Further information on our LRC annotation.

Our Data

- High-quality manual annotation
- Human annotation incorporated into GENCODE
- Rapid incorporation of new annotation
- Gene sets and regions of particular interest:
  - Genes with mouse knockout and human LOF transcripts
  - MHC and LRC regions
  - *Idd* candidate regions of NOD mice
- Inter- and intra-species comparative genomics
- Cross-referenced to other databases
- Complements Ensembl
- Downloadable datasets

What's New in release 63

- **Mouse Annotation Updated** (Mouse)
- **Rat Annotation Updated** (Rat)
- **Human Annotation Updated** (Human)
- **Website updates** (all species)
- **Annotation status shading** (all species)

More news...

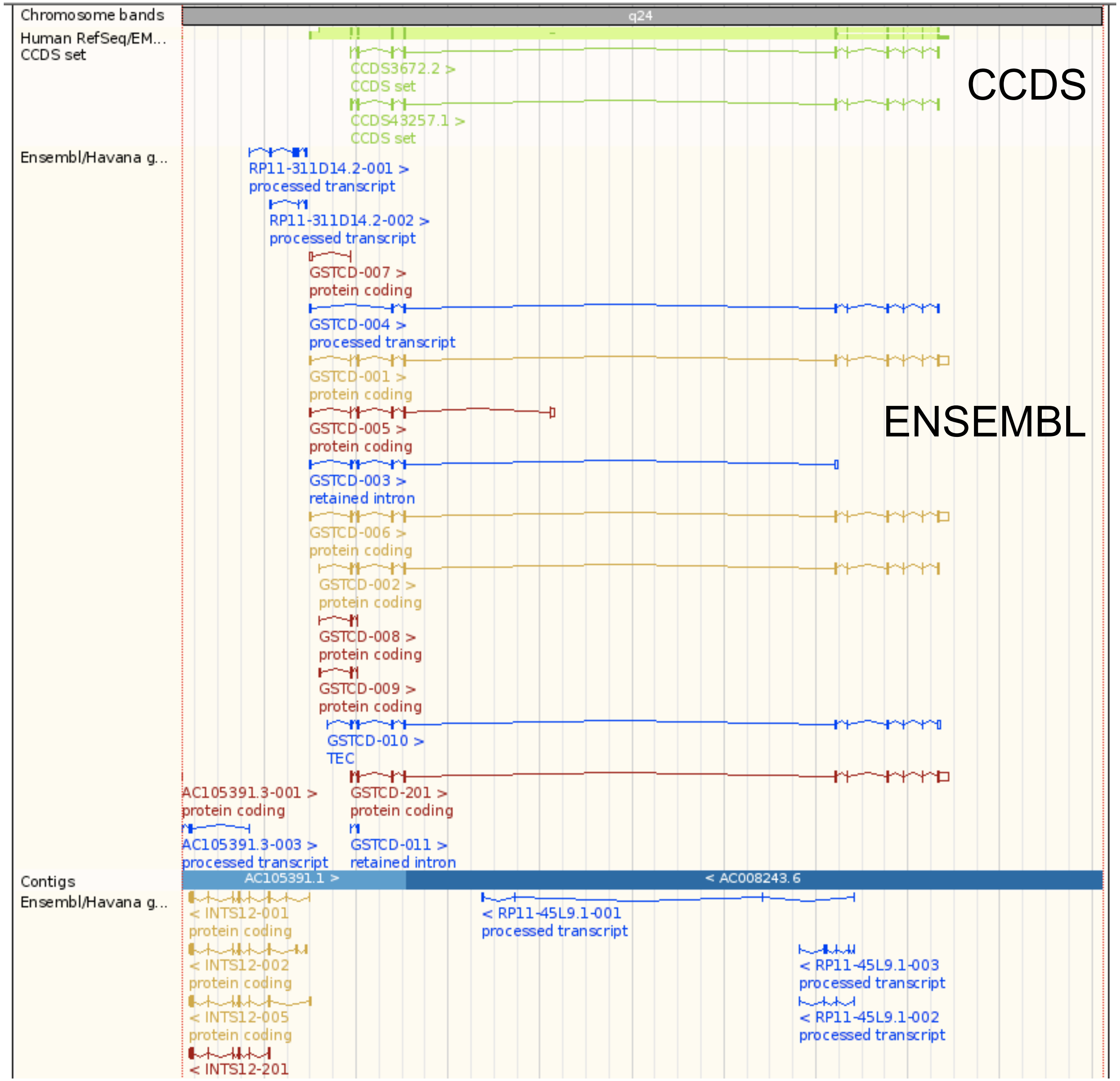
Havana team move to EBI April 2017
VEGA is being archived and final release will be February 2017

Ensembl view: GENCODE geneset

Gold (merged): agreed ensembl/havana

Red: coding (001 Havana, 201 Ensembl)

Blue: non-coding

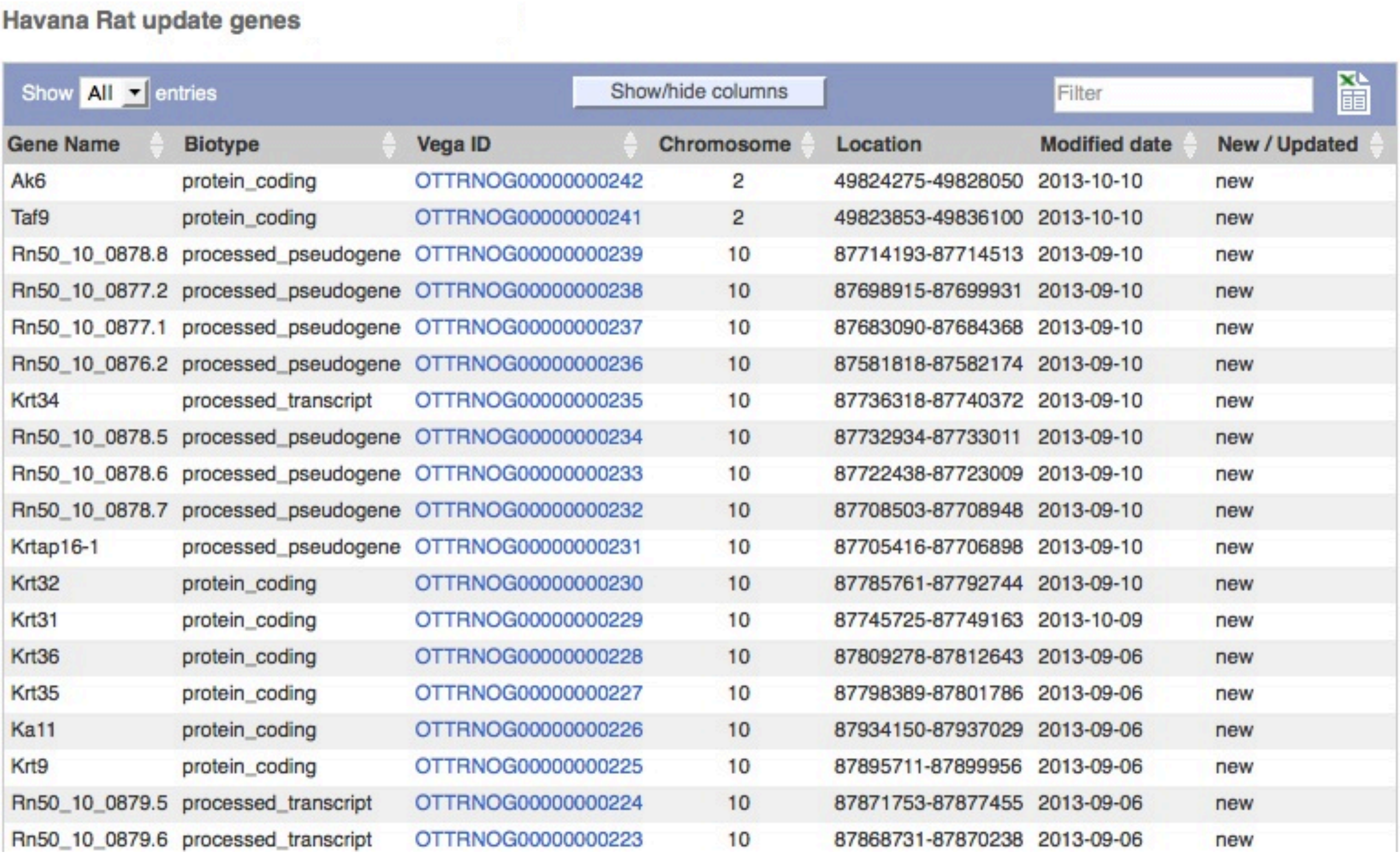


http://www.gencodegenes.org

ENCODE

UCSC

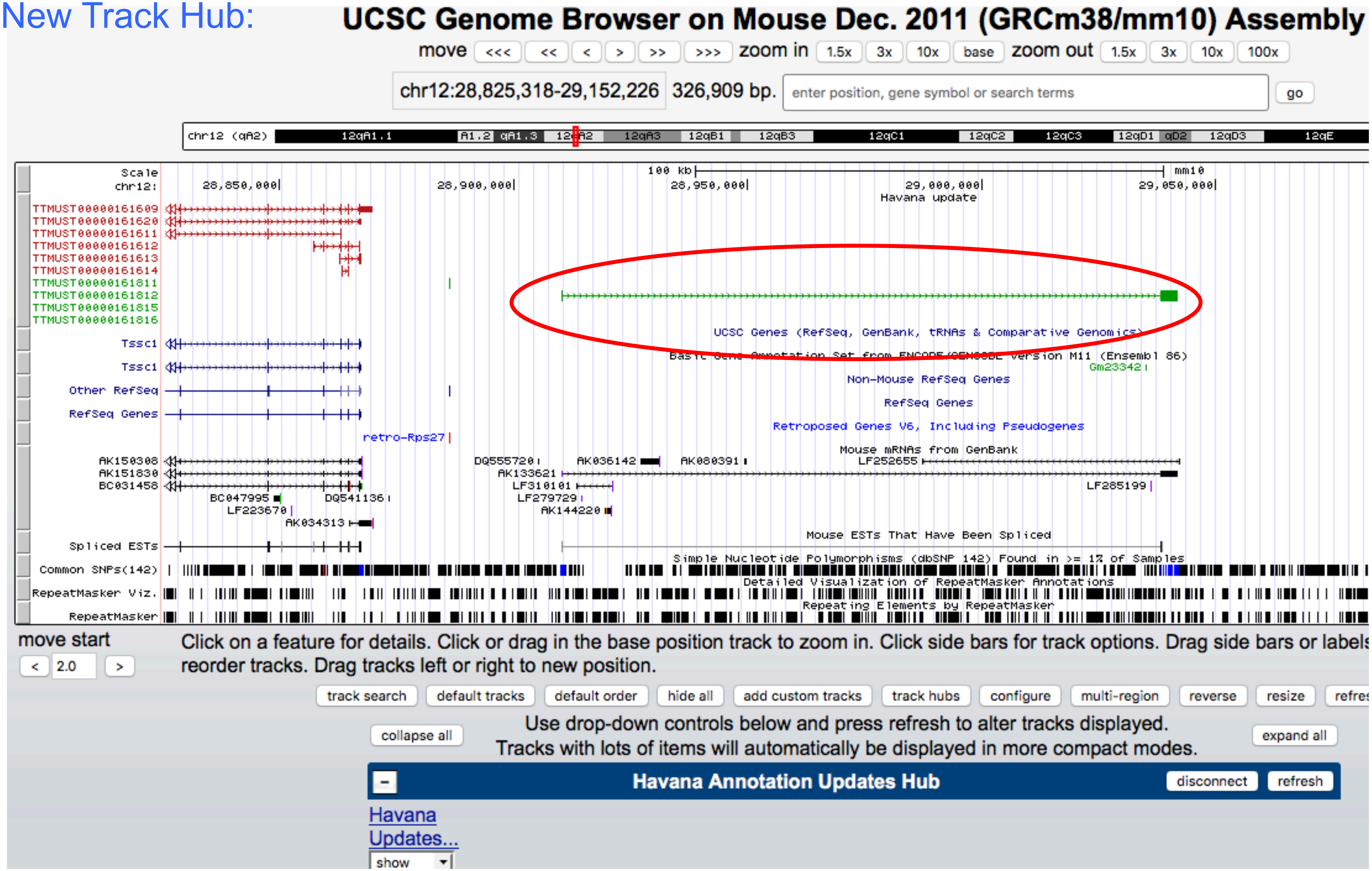
Update genes:



Havana Rat update genes

Gene Name	Biotype	Vega ID	Chromosome	Location	Modified date	New / Updated
Ak6	protein_coding	OTTRNOG00000000242	2	49824275-49828050	2013-10-10	new
Taf9	protein_coding	OTTRNOG00000000241	2	49823853-49836100	2013-10-10	new
Rn50_10_0878.8	processed_pseudogene	OTTRNOG00000000239	10	87714193-87714513	2013-09-10	new
Rn50_10_0877.2	processed_pseudogene	OTTRNOG00000000238	10	87698915-87699931	2013-09-10	new
Rn50_10_0877.1	processed_pseudogene	OTTRNOG00000000237	10	87683090-87684368	2013-09-10	new
Rn50_10_0876.2	processed_pseudogene	OTTRNOG00000000236	10	87581818-87582174	2013-09-10	new
Krt34	processed_transcript	OTTRNOG00000000235	10	87736318-87740372	2013-09-10	new
Rn50_10_0878.5	processed_pseudogene	OTTRNOG00000000234	10	87732934-87733011	2013-09-10	new
Rn50_10_0878.6	processed_pseudogene	OTTRNOG00000000233	10	87722438-87723009	2013-09-10	new
Rn50_10_0878.7	processed_pseudogene	OTTRNOG00000000232	10	87708503-87708948	2013-09-10	new
Krtap16-1	processed_pseudogene	OTTRNOG00000000231	10	87705416-87706898	2013-09-10	new
Krt32	protein_coding	OTTRNOG00000000230	10	87785761-87792744	2013-09-10	new
Krt31	protein_coding	OTTRNOG00000000229	10	87745725-87749163	2013-10-09	new
Krt36	protein_coding	OTTRNOG00000000228	10	87809278-87812643	2013-09-06	new
Krt35	protein_coding	OTTRNOG00000000227	10	87798389-87801786	2013-09-06	new
Ka11	protein_coding	OTTRNOG00000000226	10	87934150-87937029	2013-09-06	new
Krt9	protein_coding	OTTRNOG00000000225	10	87895711-87899956	2013-09-06	new
Rn50_10_0879.5	processed_transcript	OTTRNOG00000000224	10	87871753-87877455	2013-09-06	new
Rn50_10_0879.6	processed_transcript	OTTRNOG00000000223	10	87868731-87870238	2013-09-06	new

New Track Hub:



UCSC Genome Browser on Mouse Dec. 2011 (GRCm38/mm10) Assembly

chr12:28,825,318-29,152,226 326,909 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Havana Annotation Updates Hub

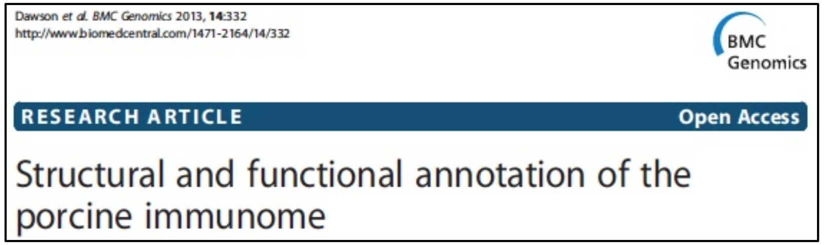
ftp://ngs.sanger.ac.uk/production/gencode/update_trackhub/hub.txt

- Manual annotation process
  - Tools, data, biotypes
- Community Manual Annotation
  - Mouse, Swine autosomes (IRAG), Rat, Chicken
- New data and projects

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)
  - Cow 2007 (Genomic WGS)



Dawson *et al. BMC Genomics* 2013, **14**:332
http://www.biomedcentral.com/1471-2164/14/332

BMC
Genomics

RESEARCH ARTICLE Open Access

Structural and functional annotation of the porcine immunome

  - Pig
    2008 (Genomic WGS)
    2010 - 2013
    - IR genes in Pig (~1300 genes) manually annotated by community

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

- Rat manual annotation 2013, 2015 (BBSRC)

- Chicken MHC 2016

Community Annotation Approaches:

The value of a genome is only as good as its annotation

Rat whole genome annotation of Rnor 6.0

Chicken MHC

Otter/Zmap Annotation Software

Authentication:

Sanger single sign-on account (email)

Registered email for Otter permitted users:

Access to our data and and analysis pipeline

Mac and Linux: Platforms of choice

Monthly updates/bugfixes

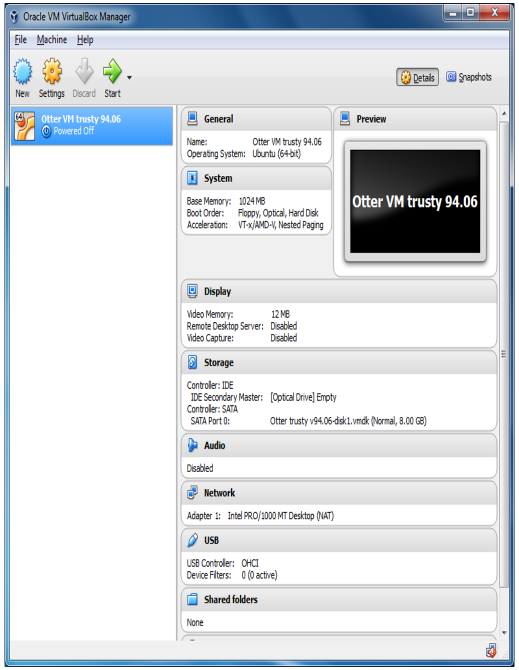
Windows:

Virtual machine image installed and run using VirtualBox

Runs an Ubuntu desktop with a bespoke Otter release

Over 2600 genes manually annotated that have been chosen by the rat community and RGD. Targeted annotation.

Final rat Vega and Ensembl merge in progress



General

Name: Otter VM trusty 94.06
Operating System: Ubuntu (64-bit)

System

Base Memory: 1024 MB
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging

Display

Video Memory: 12 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage

Controller: IDE
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: Otter trusty v94.06-disk1.vmdk (Normal, 8.00 GB)

Audio

Disabled

Network

Adapter 1: Intel PRO/1000 MT Desktop (NAT)

USB

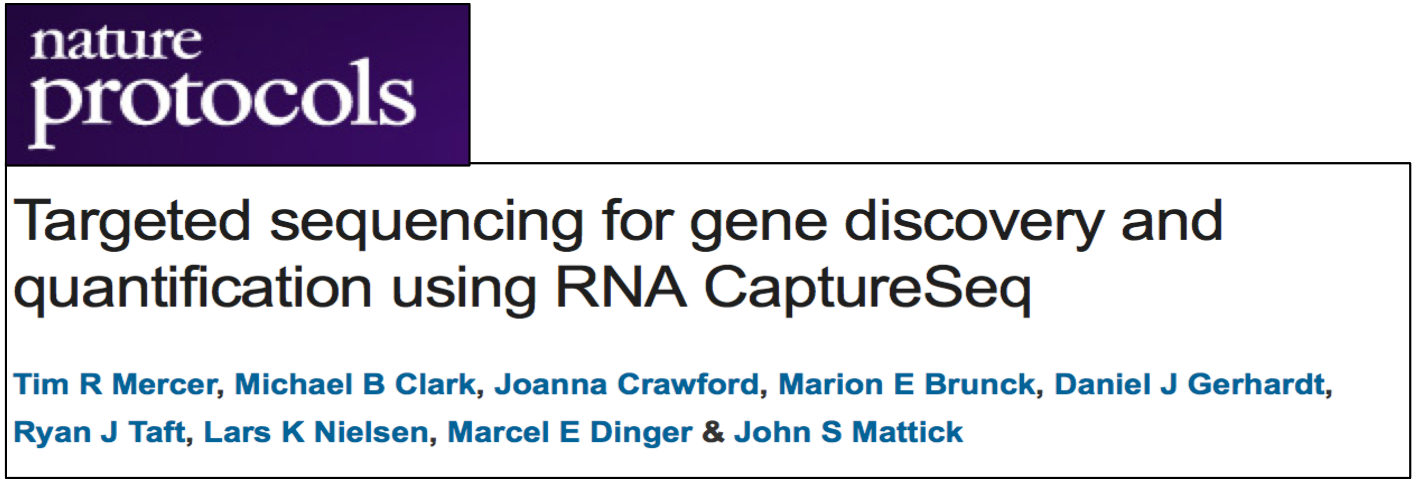
USB Controller: OHCI
Device Filters: 0 (0 active)

Shared folders

None

- Manual annotation process
  - Tools, data, biotypes
- Community Manual Annotation
  - Mouse, Swine autosomes (IRAG), Rat, Chicken
- New data and projects

Long-read data:
Better for discovering novel alternatively spliced transcripts and full-length transcripts



nature protocols

Targeted sequencing for gene discovery and quantification using RNA CaptureSeq

Tim R Mercer, Michael B Clark, Joanna Crawford, Marion E Brunck, Daniel J Gerhardt, Ryan J Taft, Lars K Nielsen, Marcel E Dinger & John S Mattick

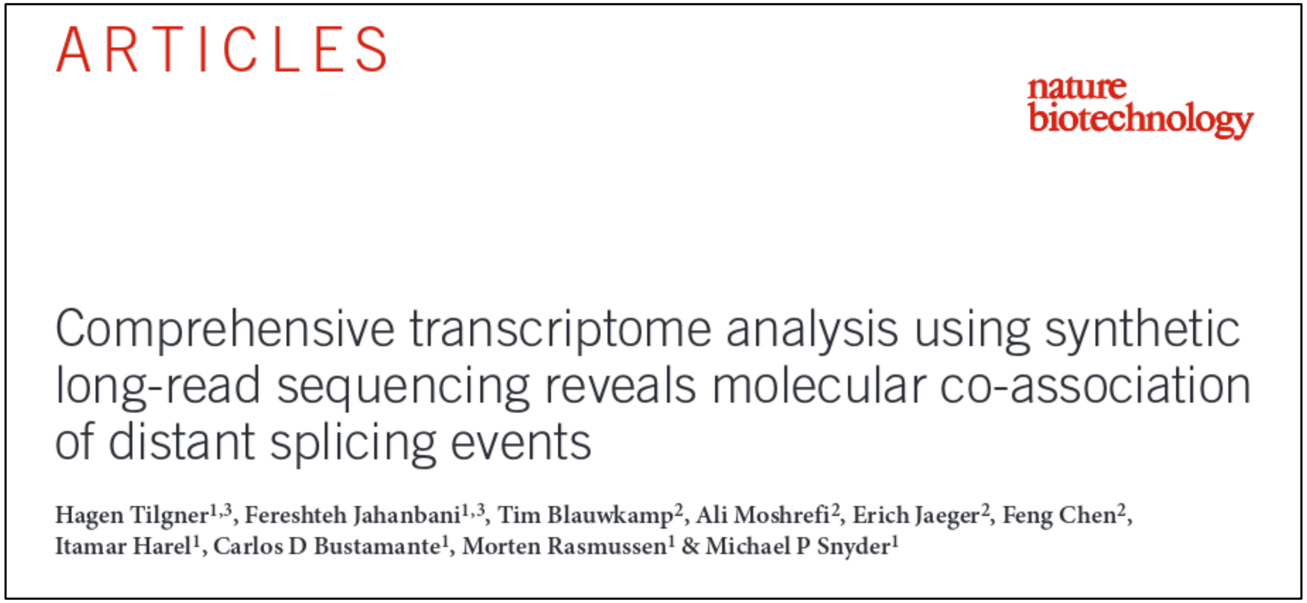
PacBio-CaptureSeq

(human: brain, testis, heart, liver, HeLa, K562)

(mouse: brain, testis, heart, liver, E7, E15)

SLR-RNAseq

(human/mouse brain)



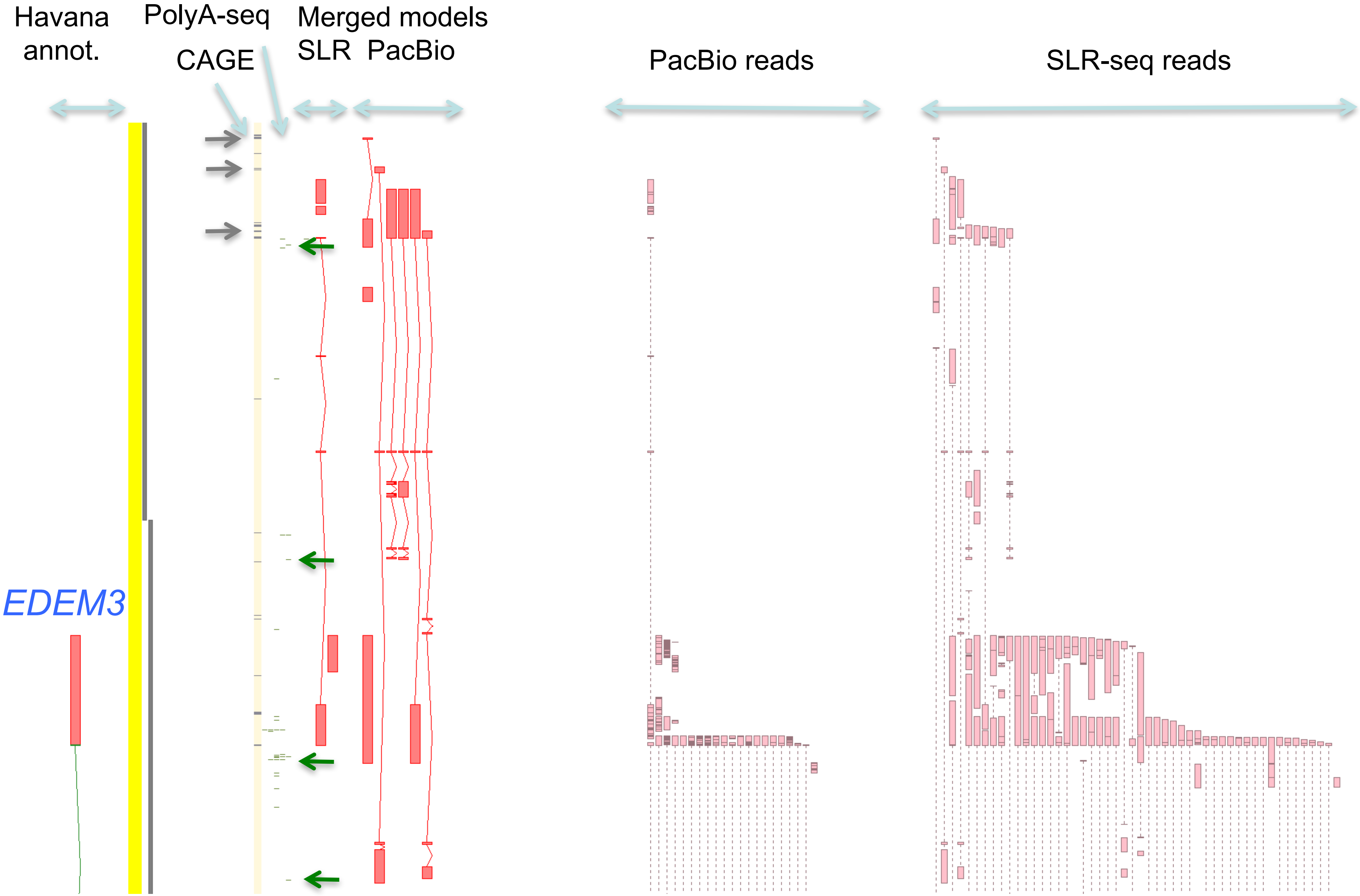
ARTICLES

nature biotechnology

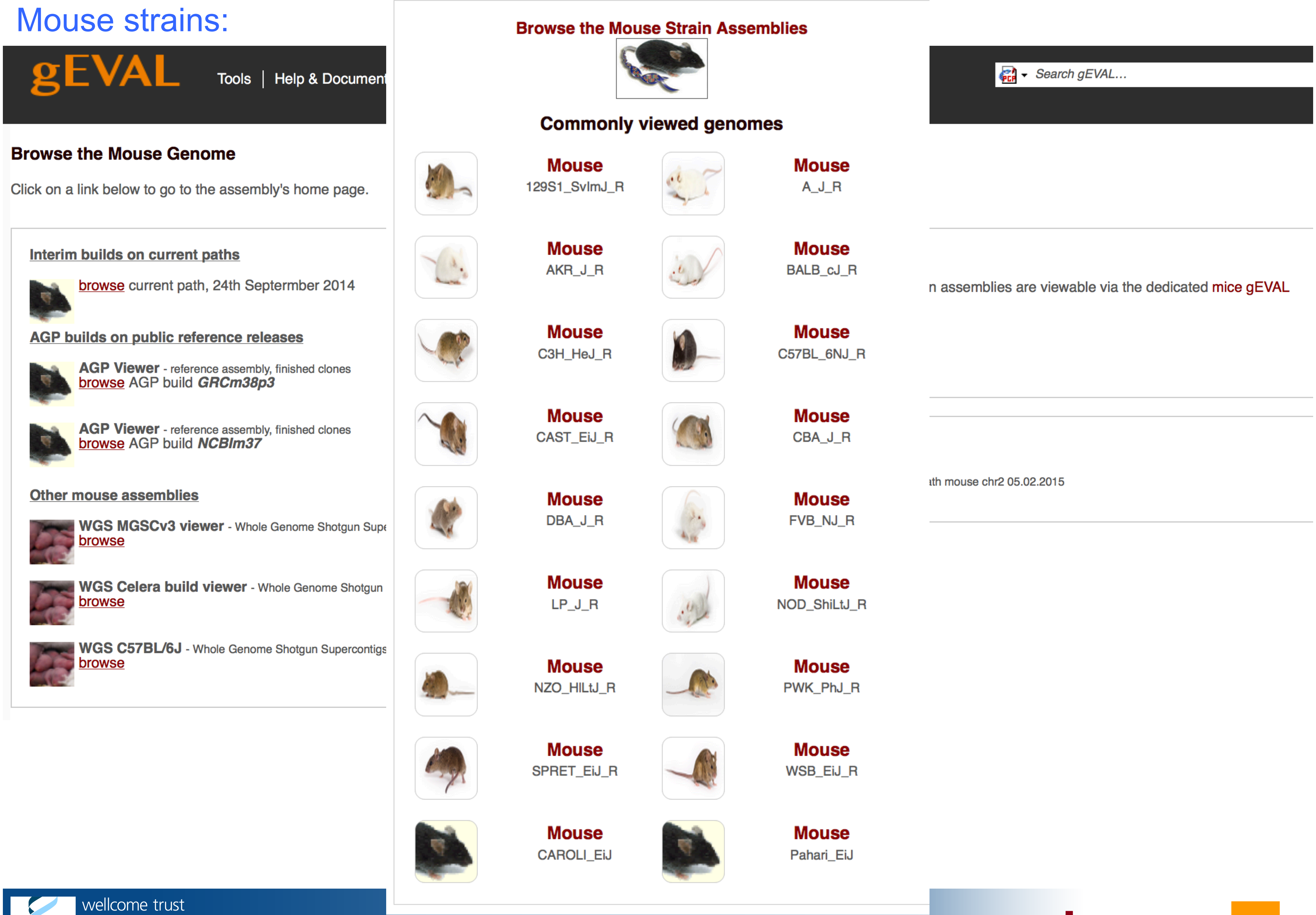
Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events

Hagen Tilgner[1,3], Fereshteh Jahanbani[1,3], Tim Blauwkamp[2], Ali Moshrefi[2], Erich Jaeger[2], Feng Chen[2], Itamar Harel[1], Carlos D Bustamante[1], Morten Rasmussen[1] & Michael P Snyder[1]

Computationally combining evidence:



Mouse strains:



Browse the Mouse Genome

Click on a link below to go to the assembly's home page.

Interim builds on current paths

browse current path, 24th Septermber 2014

AGP builds on public reference releases

AGP Viewer - reference assembly, finished clones
browse AGP build ***GRCm38p3***

AGP Viewer - reference assembly, finished clones
browse AGP build ***NCBIm37***

Other mouse assemblies

WGS MGSCv3 viewer - Whole Genome Shotgun Supe
browse

WGS Celera build viewer - Whole Genome Shotgun
browse

WGS C57BL/6J - Whole Genome Shotgun Supercontigs
browse

n assemblies are viewable via the dedicated mice gEVAL

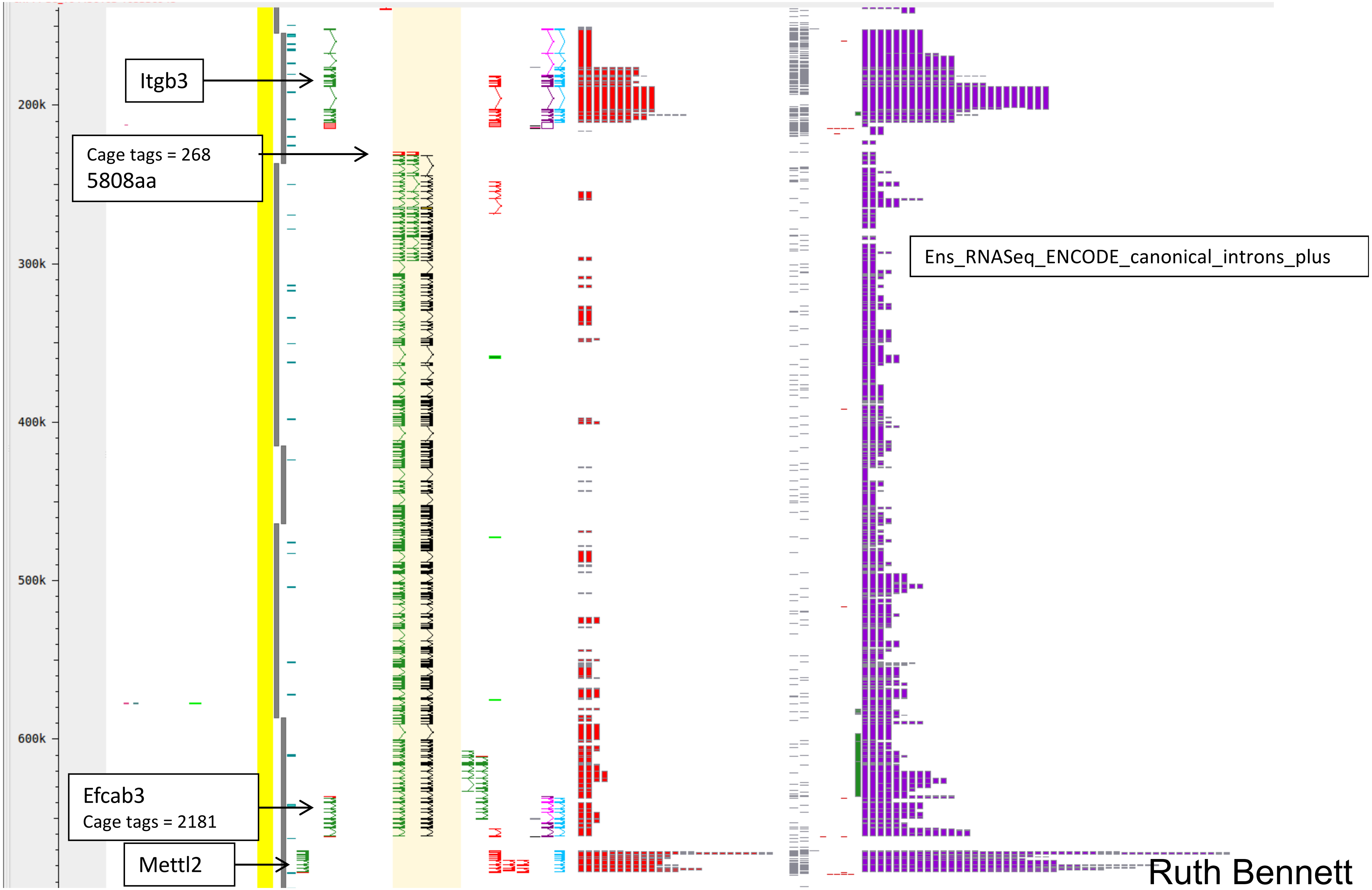
ıth mouse chr2 05.02.2015

Browse the Mouse Strain Assemblies

Commonly viewed genomes

Mouse 129S1_SvImJ_R	**Mouse** A_J_R
Mouse AKR_J_R	**Mouse** BALB_cJ_R
Mouse C3H_HeJ_R	**Mouse** C57BL_6NJ_R
Mouse CAST_EiJ_R	**Mouse** CBA_J_R
Mouse DBA_J_R	**Mouse** FVB_NJ_R
Mouse LP_J_R	**Mouse** NOD_ShiLtJ_R
Mouse NZO_HlLtJ_R	**Mouse** PWK_PhJ_R
Mouse SPRET_EiJ_R	**Mouse** WSB_EiJ_R
Mouse CAROLI_EiJ	**Mouse** Pahari_EiJ

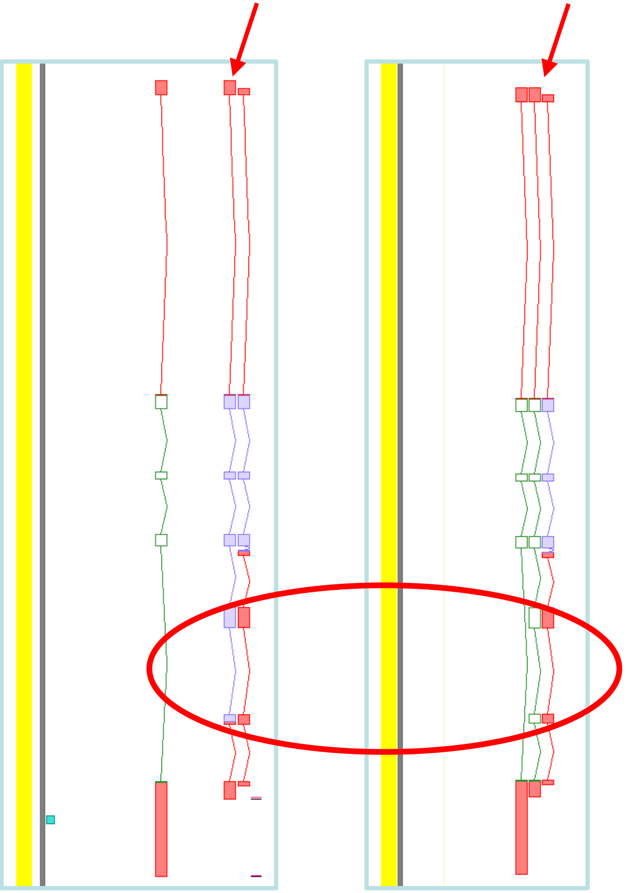
Mouse strain annotation reveals new genes:



Ruth Bennett

Mouse strain annotation reveals strain specific coding transcripts: *Ifi214*

393aa NMD 420aa coding



Reference C57BL/6

129S1-SvImJ

```
>Reference longest NMD transcript 393 AA.
MVNEYKRIVLLTGLMGINDHDFRMVKSLLSKELKLNKMQDEYDRVKI
ADLMEDKFPKDAGVVQLIKLYKQIPGLGDIANKLKNEKAKAKRKGKG
KRKTAAKRQRQEEPSTSQPMSTTNEDAEPESGRSTPDTQVAQLSLPT
ASRRNQAIQISPTIASSSGQTSSRSSETLQSIIQSPETPTRSSSRIL
DPPVSPGTAYSSAQALGVLLATPAKRQRLKNVPKEPSEENGYQQGSK
KVMVLKVTEPFAYDMKGEKMFHATVATETEFFRVKVFDIVLKEKFIP
NKVLTISNYVGCNGFINIYSASSVSEVNDGEPMNIPLSLRKSANRTP
KINYLCSKRRGIFVNGVFTVCKKEERGYYICYEIGDDTGMMEVEVYG
RLTNIACNPGDKLRLML*
```

↑

Stop codon isn’t a SNP. Caused by a much larger disruption.

```
>129 Patch long coding transcript (equivalent to ref NMD) 420
AA.
MVNEYKRIVLLTGLMGINDHDFRMVKSLLSKELKLNKMQDEYDRVKI
ADLMEDKFPKDAGVVQLIKLYKQIPGLGDIANKLKNEKAKAKRKGKG
KRKTAAKRQRQEEPSTSQPMSTTNEDAEPESGRSTPDTQVAQLSLPT
ASRRNQAIQISPTIASSSGQTSSRSSETLQSIIQSPETPTRSSSRIL
DPPVSPGTAYSSAQALGVLLATPAKRQRLKNVPKEPSEENGYQLGSK
KVMVLKVTEPFAYDMKGEKMFHATVATETEFFRVKVFDIVLKEKFIP
NKVLTISNYVGCNGFINIYSASSVSEVNDGEPMNIPLSLRKSANRTP
KINYLCSKRRGIFVNGVFTVCKKEERGYYICYEIGDDTGMMEVEVYG
RLTNIACNPGDKLRLICFELTPDEETAWLRSTTHSNMQVIKARN*
```

Ruth Bennett

Mouse strain annotation reveals strain specific expansions and biotype differences: Nlrp locus

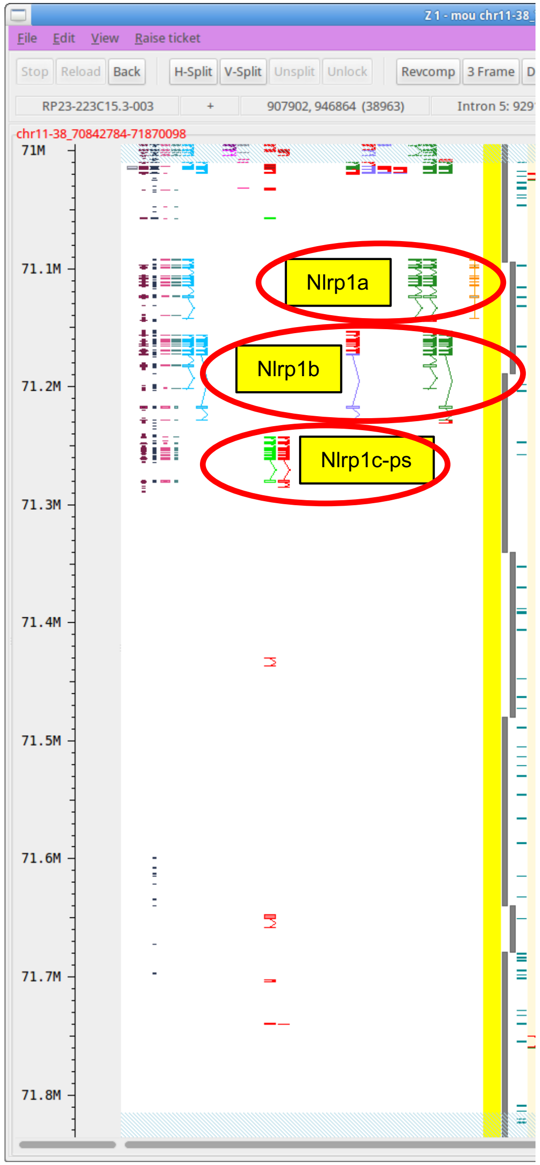
Reference: C57BL/6

Strains: PWK/PhJ WSB/EiJ CAST/EiJ

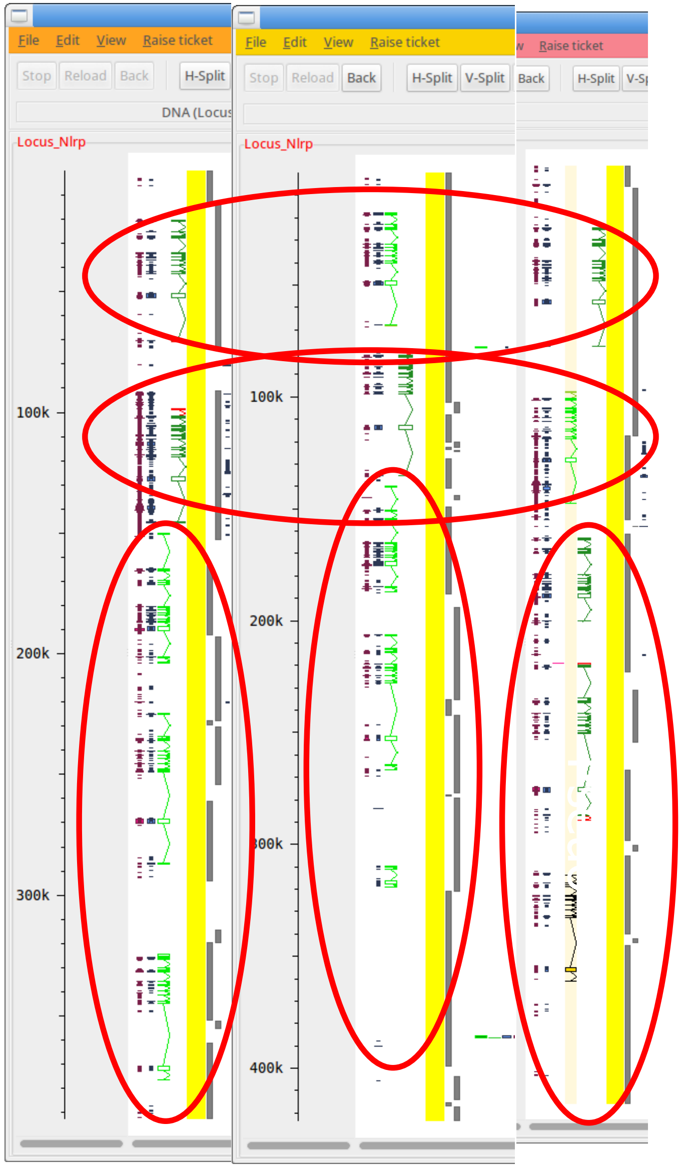
Gene

Gene

Pseudogene


Z 1 - mou chr11-38
File Edit View Raise ticket
Stop Reload Back H-Split V-Split Unsplit Unlock Revcomp 3 Frame
RP23-223C15.3-003 + 907902, 946864 (38963) Intron 5: 929
chr11-38_70842784-71870098
71M
71.1M
71.2M
71.3M
71.4M
71.5M
71.6M
71.7M
71.8M
Nlrp1a
Nlrp1b
Nlrp1c-ps

Expansion



PWK WSB CAST

Gene Pseudo Gene

Gene Gene Pseudo

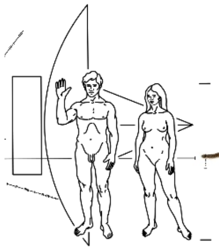
Pseudo Pseudo Gene

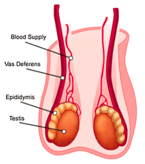
Pseudo Pseudo Gene

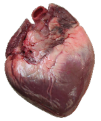
Pseudo Pseudo Pseudo

Ruth Bennett

Clinical data: RNA capture-seq across a range of tissues





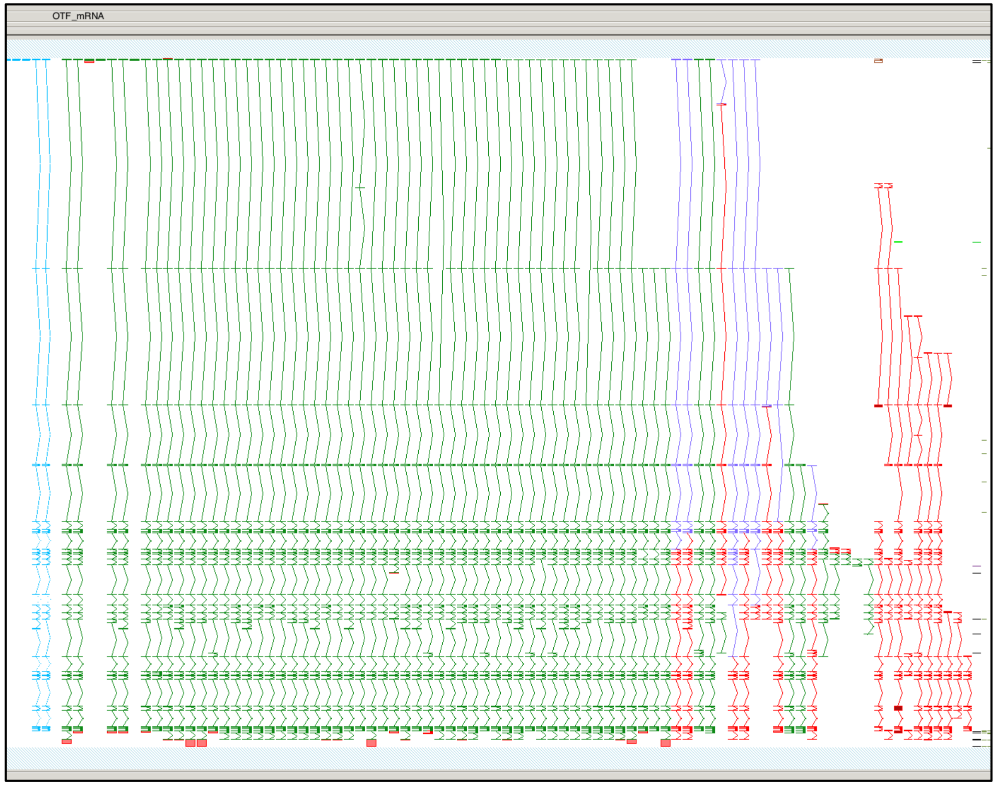


HUMAN	MOUSE
1. Brain	**1. Brain**
2. Testis	**2. Testis**
3. Heart	**3. Heart**
4. Liver	**4. Liver**
5. HeLa	5. Embryo 7d
6. K562	6. Embryo 15d

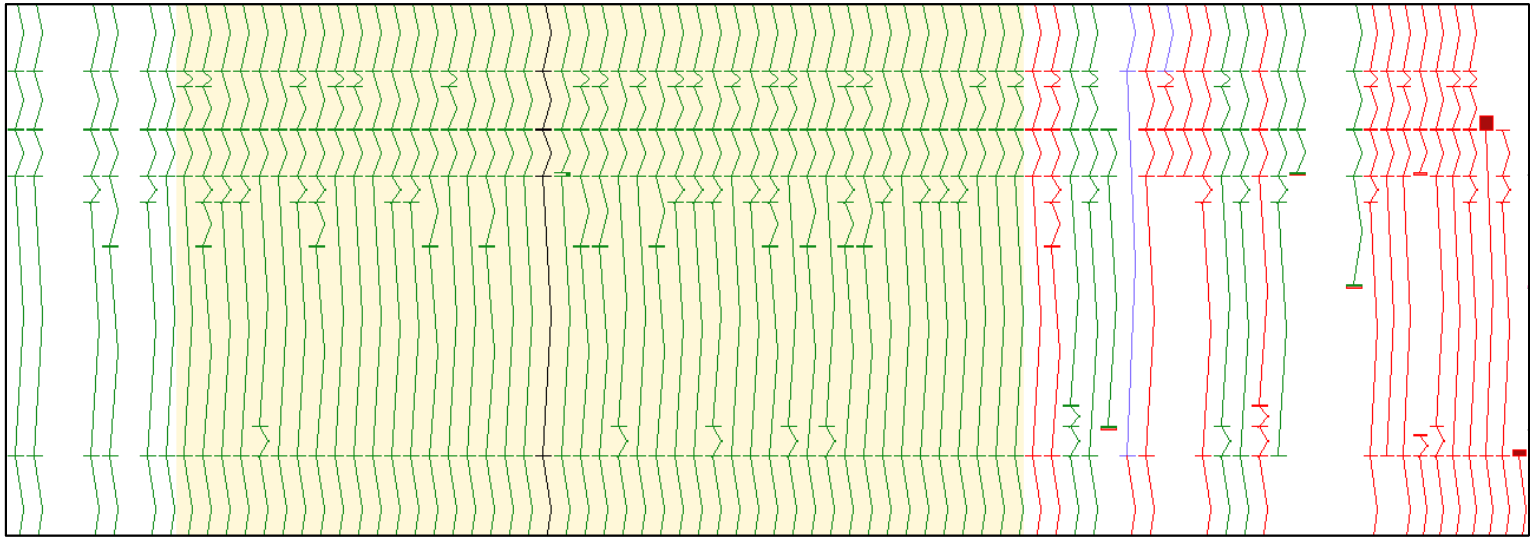
Clinical Data

Human *KCNMA1*:

22 original transcripts
Now 92 transcripts
25 from SLR-Seq
SLR-Seq also extended other transcripts



Micro exons, NAGNAG, alternate exon use



Marie-Marthe Suner

Final Comments

Zmap/otter tool

- Can handle large datasets
- Training, QC and feedback

Let the pipelines take the strain

- Targeted manual annotation
- Spend time on the tricky things

Feedback genome quality

- Report errors
- Improve the assembly

Acknowledgements

Havana:

Adam Frankish
If Barnes
Ruth Bennett
Andrew Berry
Alex Bignell
Claire Davidson
Gloria Despacio-Reyes
Sarah Donaldson
Matthew Hardy
Toby Hunt
Mike Kay
Jane Loveland
Deepa Manthravadi
Gaurab Mukherjee
Jonathan Mudge
Marie-Marthe Suner
Mark Thomas

Gencode:

Jose Gonzalez
Stephen Fitzgerald

Annosoft:

Ed Griffiths
Steve Miller

Vega:

Stephen Trevanion
Dan Sheppard

ftp://ngs.sanger.ac.uk/production/gencode/update_trackhub/hub.txt