Vega and Community Manual Annotation

Jane Loveland

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

Havana: Human and vertebrate analysis and annotation

- Manual annotation of human, mouse and zebrafish whole chromosomes or genomes
- Human ENCODE, mouse EUCOMM 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/pipeline/ access of data (VEGA)

- Community Manual Annotation –

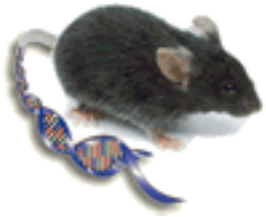
  Mouse (KOMP and NorCOMM)

  Swine autosomes (IRAG)

Do we know how many genes there are?

Gene counts

Known protein-coding genes:	20,442
Novel protein-coding genes:	434
Pseudogenes:	15,007
RNA genes:	12,523
Immunoglobulin/T-cell receptor gene segments:	562
Gene exons:	649,964
Gene transcripts:	181,744



MGSC

Gene counts

Known protein-coding genes:	21,879
Novel protein-coding genes:	826
Pseudogenes:	5,228
RNA genes:	6,695
Immunoglobulin/T-cell receptor gene segments:	481
Gene exons:	411,134
Gene transcripts:	95,883

Automatic Annotation vs Manual

Automatic Annotation

- Quick whole genome analysis ~ weeks
- Consistent annotation
- Use unfinished/illumina sequence/shotgun assembly
- No polyA sites/signals, pseudogenes
- Predicts ~75% loci

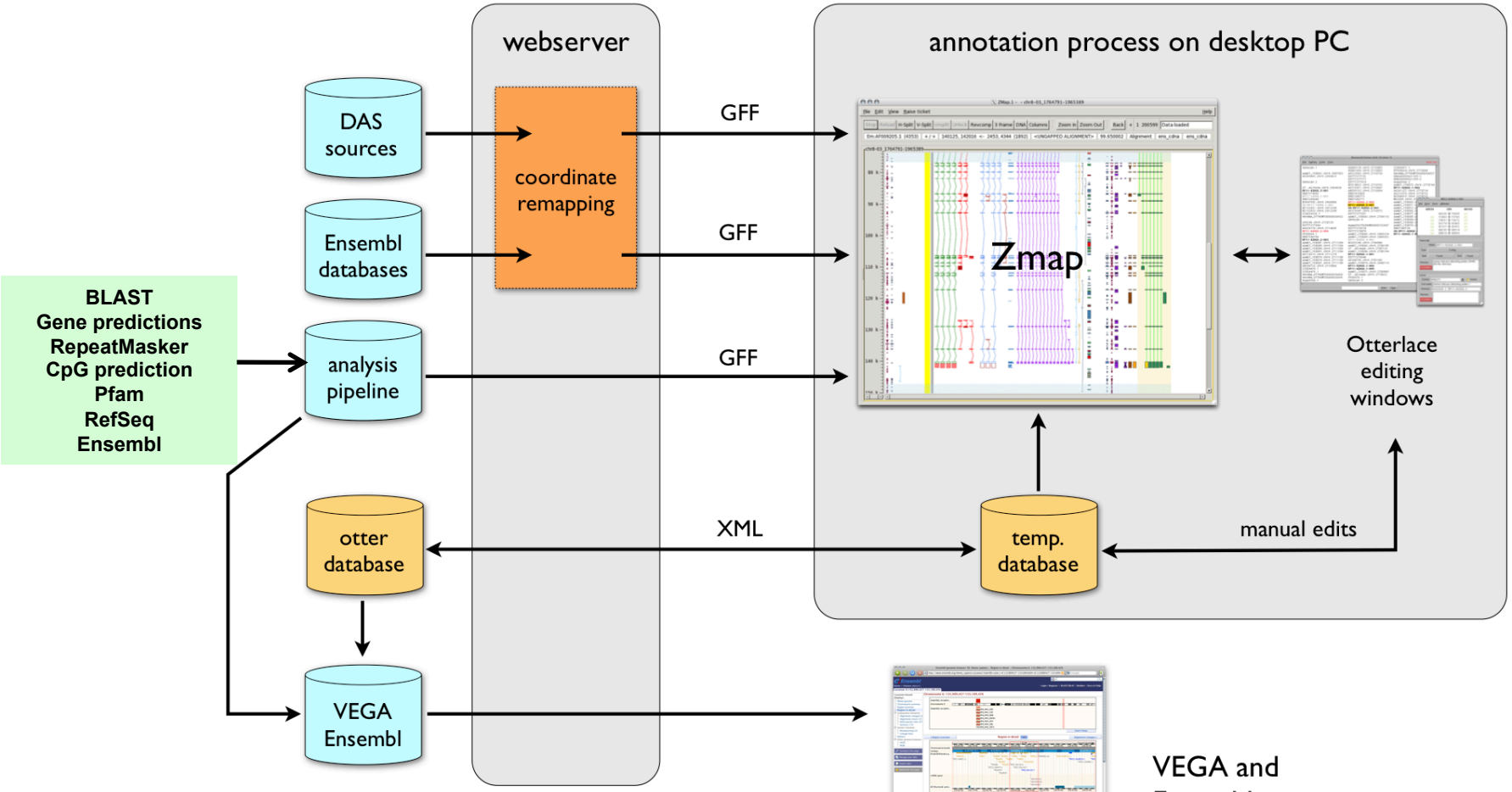
Manual Annotation

- Extremely slow~3 months Chr 6
- Need finished (high quality) seq
- Flexible, can deal with inconsistencies in data
- Most rules have exception
- Consult publications as well as databases

Manual annotation:

- manual annotation of genomic sequence (finished and unfinished)
- every exon of every transcript supported by homology (mRNA / EST / protein)
- splice variants
- pseudogenes
- nomenclature
- gene clusters
- interpretation of problematic evidence
- examination of literature

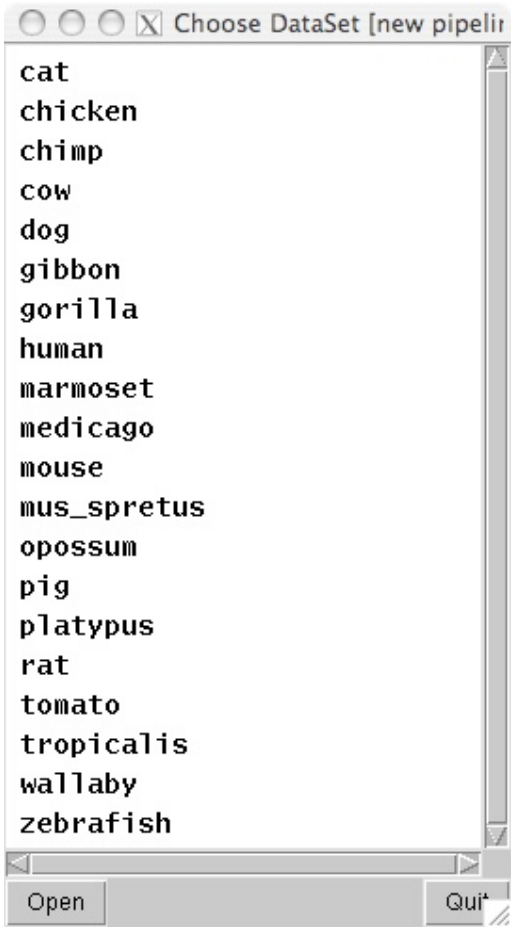
Analysis and Annotation pipeline: Otterlace/ZMap

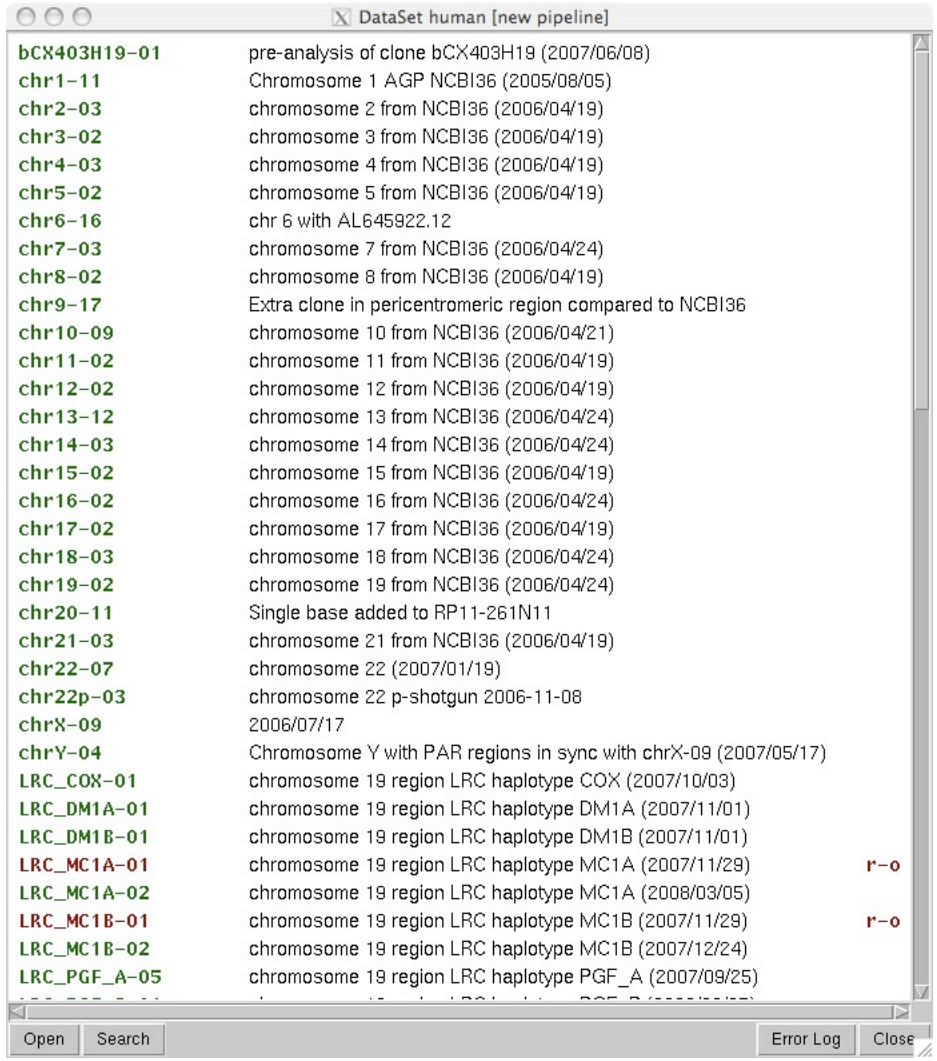


VEGA and Ensembl websites

DAS=Distributed Annotation system

Annotation interface: datasets





Ana_notes: interface to record annotation history

11	AC091729.4	AC091729.4	completed	2007-12-11	jpa	5' C7orf50, GPR30, ZFAND2A, novel transcript
12	AC073094.11	AC073094.11	completed	2008-02-14	jpa	Contains the UNCX gene for UNC homeobox, a putative novel transcript and ten CpG islands.
13	AC102953.5	AC102953.5	completed	2007-12-06	ds8	Contains the MICALL2 gene for MICAL-like 2, the 3' end of the INTS1 gene for integrator complex subunit 1, a novel gene and nine CpG islands.
14	AC093734.3	AC093734.3	completed	2007-12-06	ds8	Contains the 5' end of the INTS1 gene for integrator complex subunit 1, four novel genes, the MAFK gene for v-maf musculoaponeurotic fibrosarcoma oncogene homolog K (avian), the TMEM184A gene for transmembrane protein 184A, the PSMG3 gene for proteasome (prosome, macropain) assembly chaperone 3 and nine CpG islands.
15	AC074389.8	AC074389.8	completed	2007-12-11	jel	Contains a Myeloproliferative syndrome, transient (transient abnormal) (TAM) pseudogene, four novel genes, the gene for a novel protein and five CpG islands.
16	AC110781.3	AC110781.3	completed	2007-12-11	jel	Contains the 3' end of the MAD1L1 gene for MAD1 mitotic arrest deficient-like 1 (yeast), a novel gene, the 5' end of a novel gene and three CpG islands.
17	AC104129.4	AC104129.4	completed	2007-12-11	jel	Contains the 3' end of the gene for a novel protein, part of the MAD1L1 gene for MAD1 mitotic arrest deficient-like 1 (yeast) and a CpG island.
18	AC069288.7	AC069288.7	completed	2007-12-11	jel	Contains part of the MAD1L1 gene for MAD1 mitotic arrest deficient-like 1 (yeast) and six CpG islands.
19	AC006433.18	AC006433.18	completed	2007-12-11	jel	Contains part of the MAD1L1 gene for MAD1 mitotic arrest deficient-like 1 (yeast) and five CpG islands.
20	AC005282.4	AC005282.4	completed	2007-12-11	jel	Contains the 5' end of the MAD1L1 gene for MAD1 mitotic arrest deficient-like 1 (yeast), the 3' end of the FTSJ2 gene for FtsJ homolog 2 (E. coli) and three CpG islands.
21	AC004971.3	AC004971.3	completed	2008-02-19	jpa	Contains the NUDT1 gene for nudix (nucleoside diphosphate linked moiety X)-type motif 1, the SNX8 gene for sorting nexin 8, the 5' end of the EIF3S9 gene for eukaryotic translation initiation factor 3, subunit 9 eta, 116kDa and eight CpG islands.
22	AC004840.4	AC004840.4	completed	2008-02-19	jpa	Contains the 3'end of the EIF3S9 gene for eukaryotic translation initiation factor 3, subunit 9 eta, 116kDa, the CHST12 gene for carbohydrate (chondroitin 4) sulfotransferase 12, two novel genes and two CpG islands.

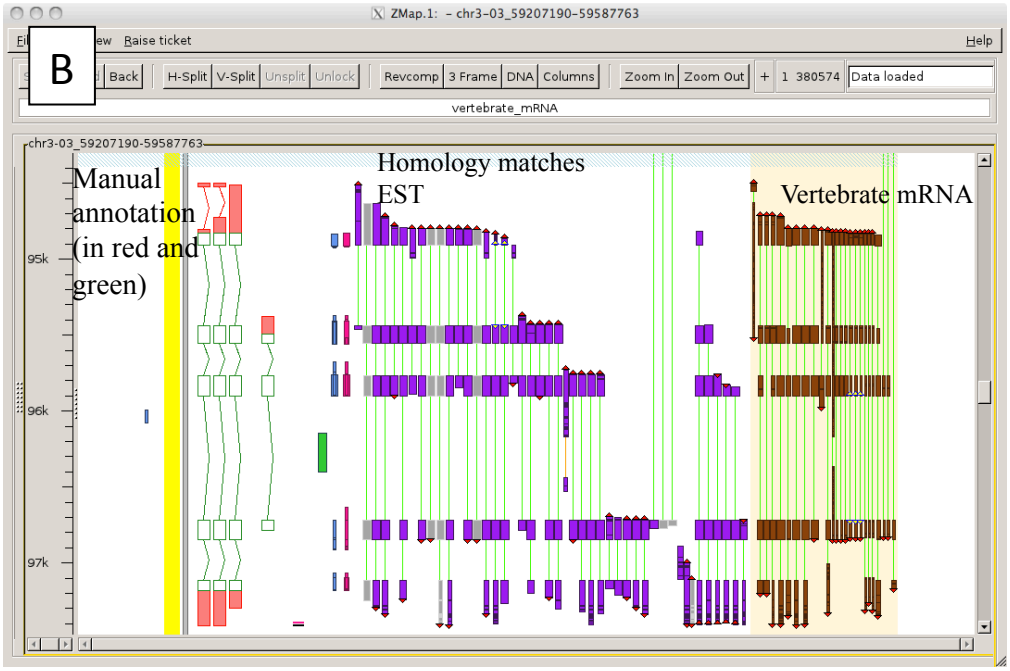
Note text: Set note clear

write access | Show Range [F7] | Hunt selection | Refresh Locks | Refresh Ana. Status | Open from chr coords | Load pipeline data | Run Iace | Close

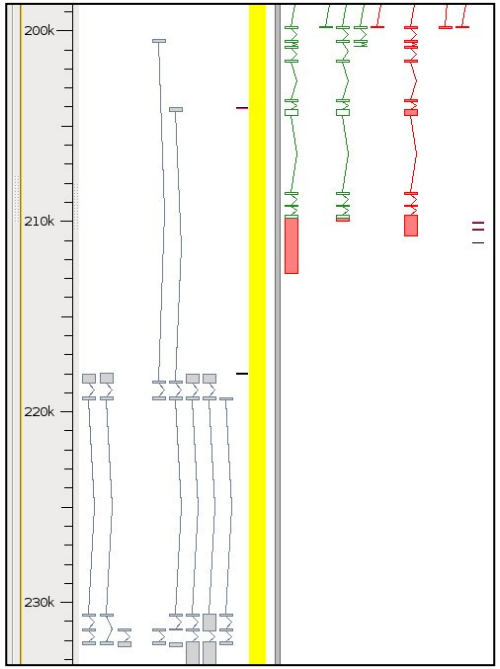
A

348	CU862069.1	CH242-512B1	completed		
349	FP102293.2	CH242-307A4	completed	🔒	jel
350	CU694523.1	CH242-64C23	completed	🔒	jel
351	CU468048.2	CH242-387C12	completed		

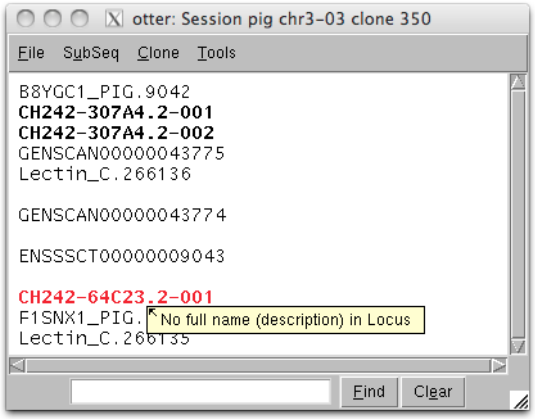
B



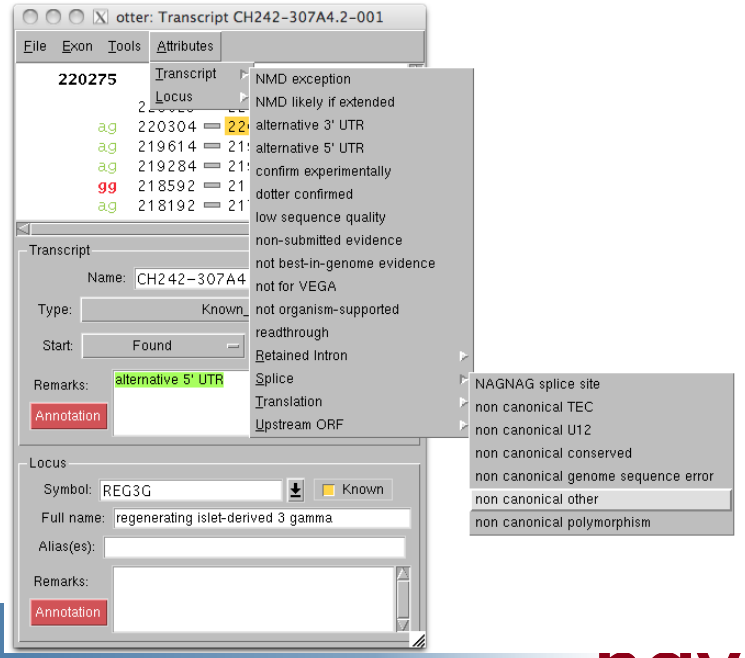
C



D



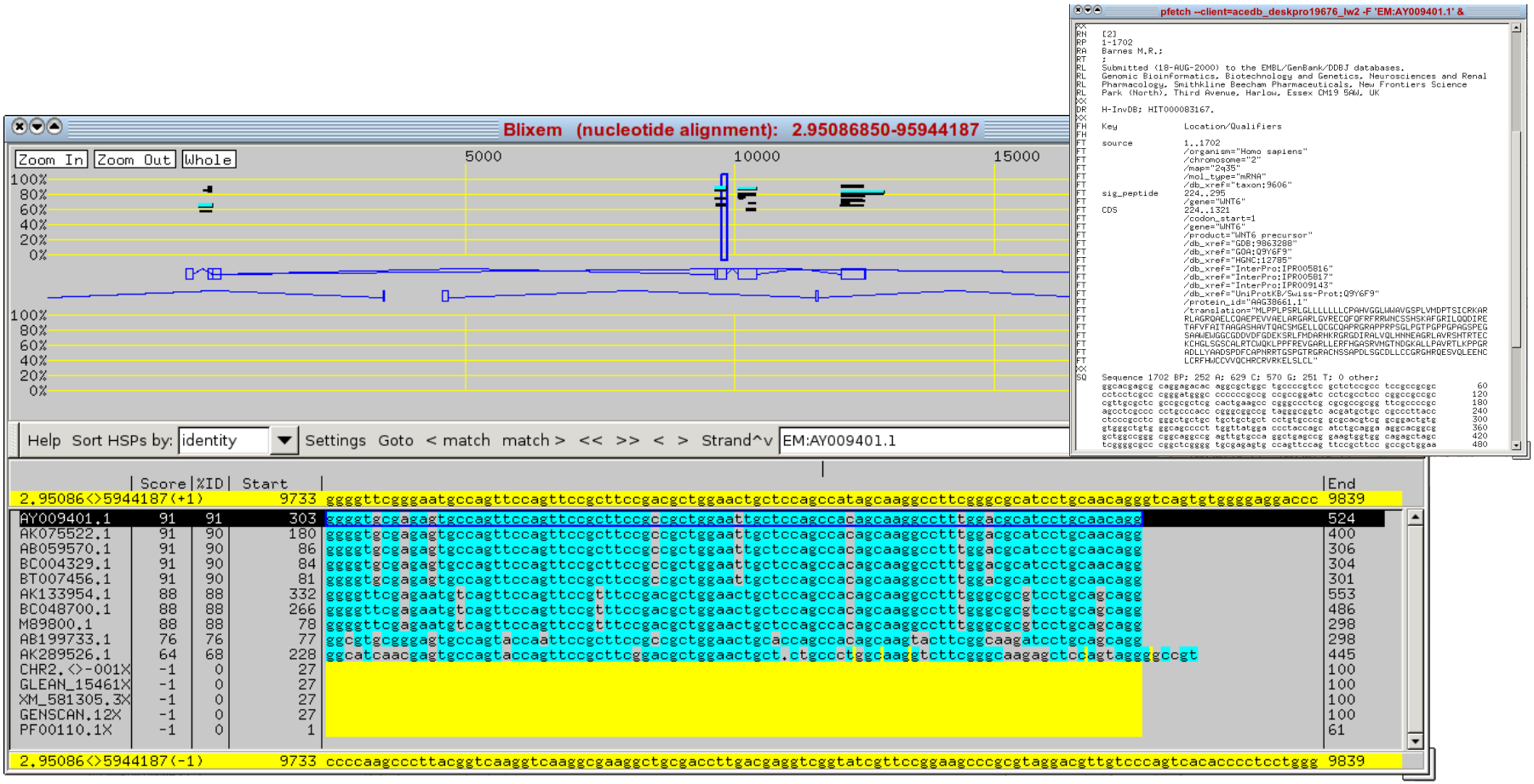
E



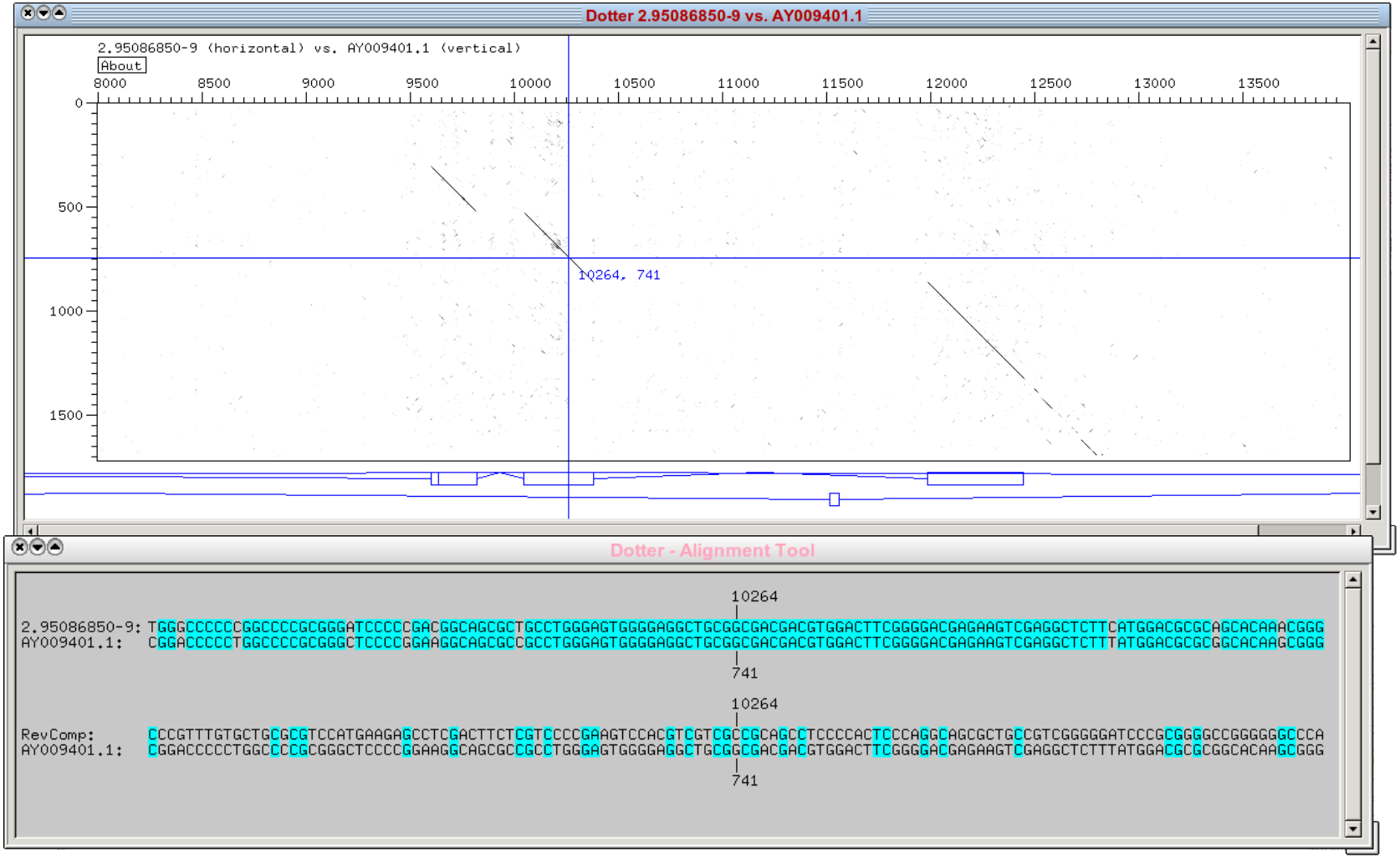
wellcome trust sanger institute

navana human and vertebrate analysis and annotation

Splicing checked via viewing cDNA alignments in “blixem”



Dotter can be used to align against unmasked sequence (reveal small exons)



DAS (distributed annotation system) source visible in Zmap

lace chr21-03, clones 486..490

File SubSeq Clone Tools

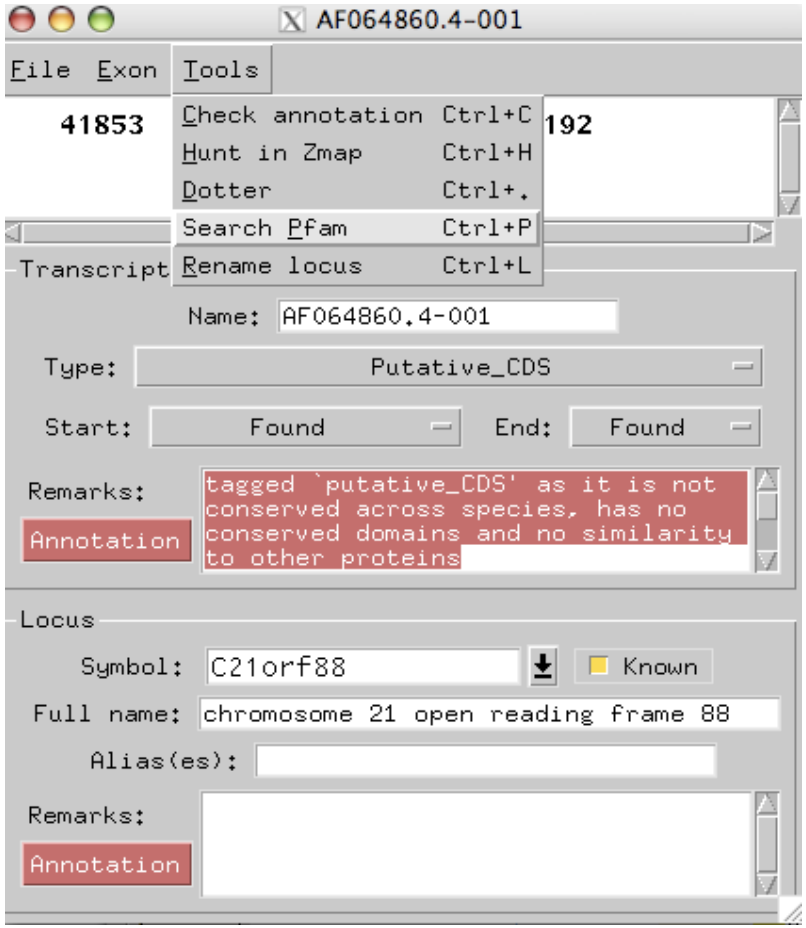
Launch ZMap	Ctrl+Z	
Launch In A ZMap		
Genomic Features	Ctrl+G	
Dotter Zmap hit	Ctrl+,	
Exonerate Zmap hit/Column	Ctrl+X	
Rename locus	Ctrl+Shift+L	
Re-authorize	Ctrl+Shift+A	
Load column data		

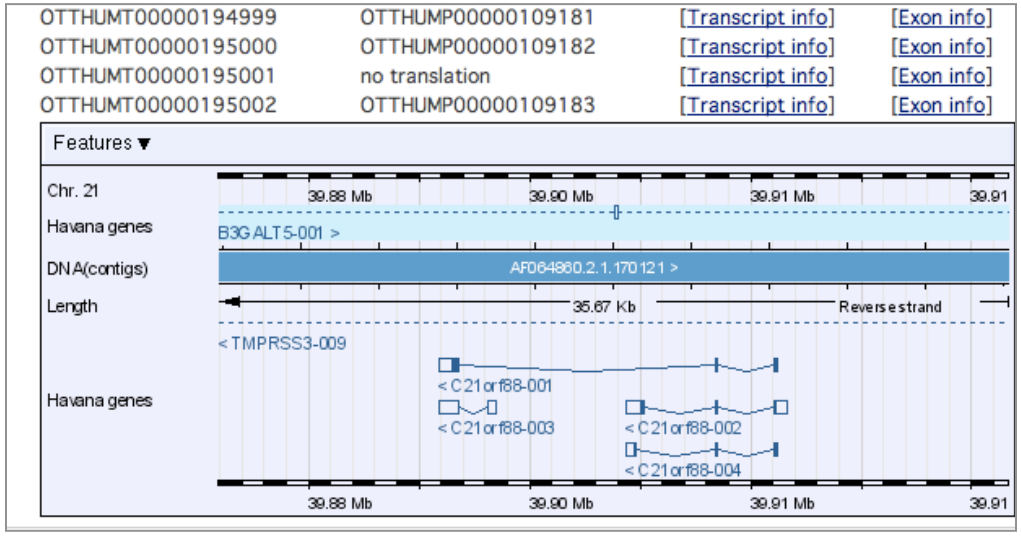
AP001469.6-002		
ESTT60480		
augustus.2		
ENST397708		
AP001469.6-001		
ENST291688		
ERI:AP001469.3-00		
GD:AP001469.9-001		
MPI:AP001469.2-003	AP000471.3-010	ESTT60502
genscan.5	AP000471.3-009	augustus.1
CCDS13734.1	ENST310126	ESTT60498
ESTT60605	CCDS13735.1	ESTT60499
ESTT60603	ESTT60583	AP000337.2-005
ESTT60600	ESTT60585	ESTT60495
ESTT60599	ESTT60588	ESTT60496
ESTT60607	AP000471.3-008	augustus.5
AP001469.6-006	OTTHUMT00000207282	AP000337.2-004
PF03399.1	AP000471.3-003	genscan.1
	AP000471.3-004	AP000337.2-002
ESTT60474	OTTHUMT00000207283	ESTT60491
	ERI:AP000471.59-001	CCDS33592.1
AP001469.6-008	AP000471.3-002	MPI:AP000337.1-001
AP001469.6-007	AP000471.3-001	ERI:AP000337.1-001
	MPI:AP000471.60-001	GD:AP000337.1-001
PF02130.1	ESTT60587	OTTHUMT00000207336
CCDS33591.1	AP000471.3-007	ENST337772
ESTT60487	MPI:AP000471.60-003	AP000337.2-003
OTTHUMT00000207272	OTTHUMT00000207286	AP000337.2-001

Load column data

- augustus
- cpg
- das_aspic
- das_comparacons_10way
- das_congo_exons
- das_evigan
- das_exonify
- das_gerp_23_way_constrelem
- das_phastcons_17way
- das_phastcons_28way
- das_siepel_novelloci
- das_transmap_mrna
- das_transmap_refseq
- das_transmap_splicedest
- das_transmap_ucscgenes
- das_ucsc_retroali3
- das_washu_human_pasa_ests
- das_washu_mrnas
- das_washu_nscan1
- das_yale_pseudogene
- ditag_chip_pet
- ditag_gis_pet
- ditag_gis_pet_encode
- ens_ccds_from_ensembl
- ens_ensembl
- ens_ensembl_from_ensembl_havana
- ens_ensembl_havana
- ens_estgenes
- ens_ncrna
- ens_separate_ccds
- eponine
- est2genome_human
- est2genome_mouse
- est2genome_other
- genscan
- halfwise
- refseq_human
- repeatmasker
- trf
- uniprot_sw
- uniprot_tr

Search Pfam on the fly in otterlace



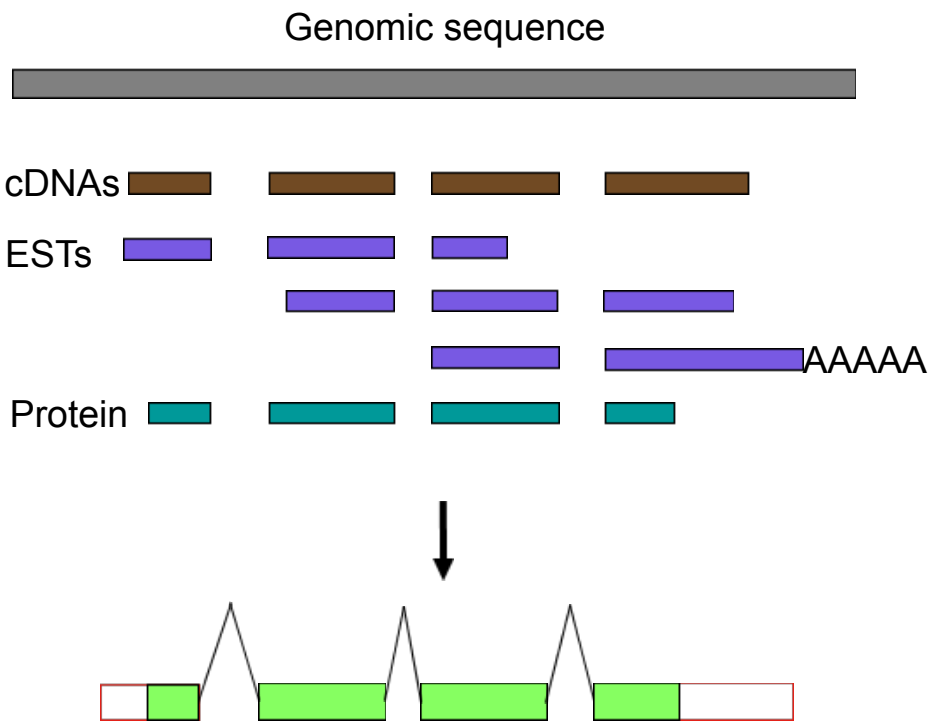


Transcript Name	**C21orf88-001** (Vega_transcript) Ensembl transcript sharing CDS with Havana: ENST00000380612
Transcript information	**Exons:** 3 **Transcript length:** 1,071 bps **Protein length:** 145 residues [Further Transcript info] [Exon information] [Protein information]
Transcript Class	Putative protein coding [Definition]
Transcript structure	15.29 Kb Reverse strand
Protein features	Peptide Prosite profiles PS51257 Scale (aa) 20 40 60 80 100 120 145

C21orf88 no pfamA domains

Manual Annotation and Biotypes:

Annotation based on transcriptional evidence.



Protein Coding

- Known_CDS
- Novel_CDS
- Putative_CDS
- Nonsense_mediated_decay

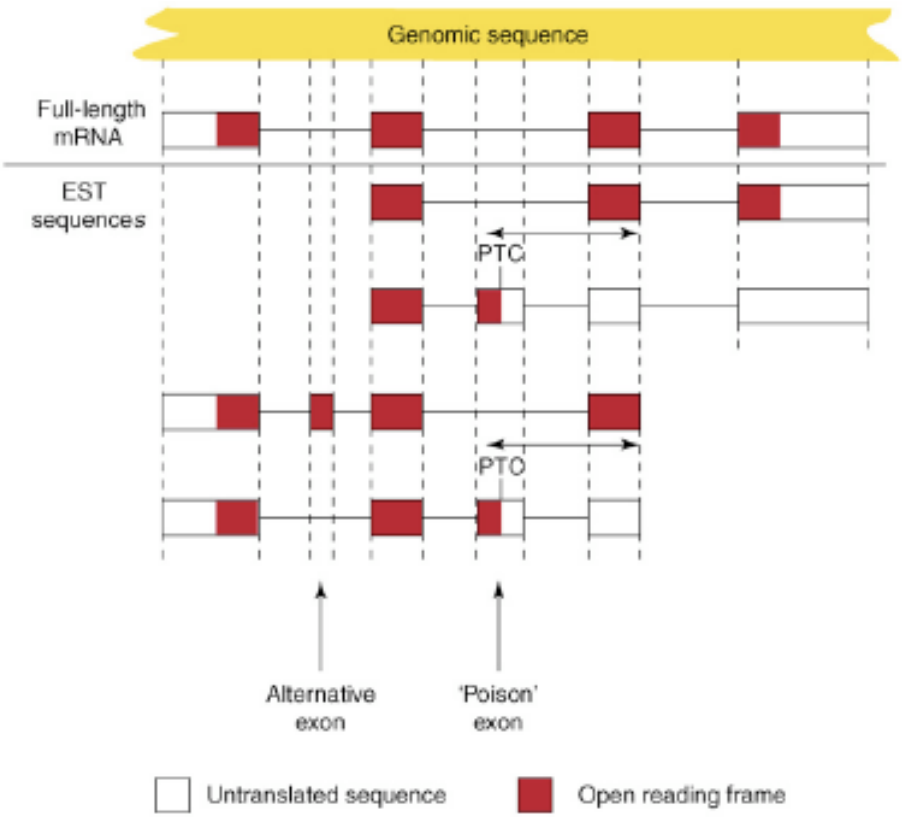
Transcript

- Non_coding
- LincRNA
- Antisense
- Sense intronic
- Sense overlapping
- 3' overlapping ncRNA
- Retained_intron
- Putative
- Artefact

Pseudogene

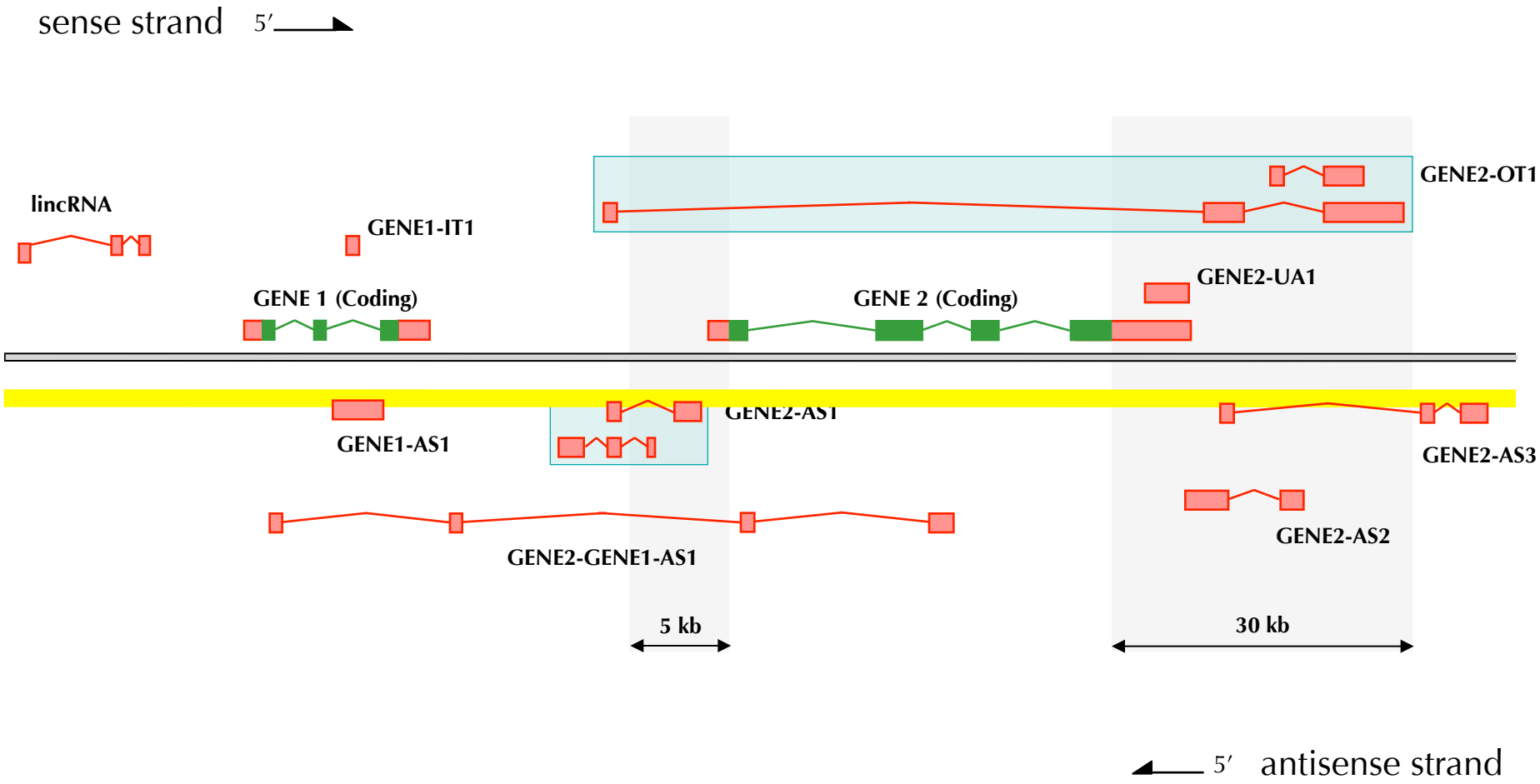
- Processed_pseudogene
- Unprocessed_pseudogene
- Transcribed_processed
- Transcribed_unprocessed
- Unitary_pseudogene
- Polymorphic_pseudogene

Identification of NMD:

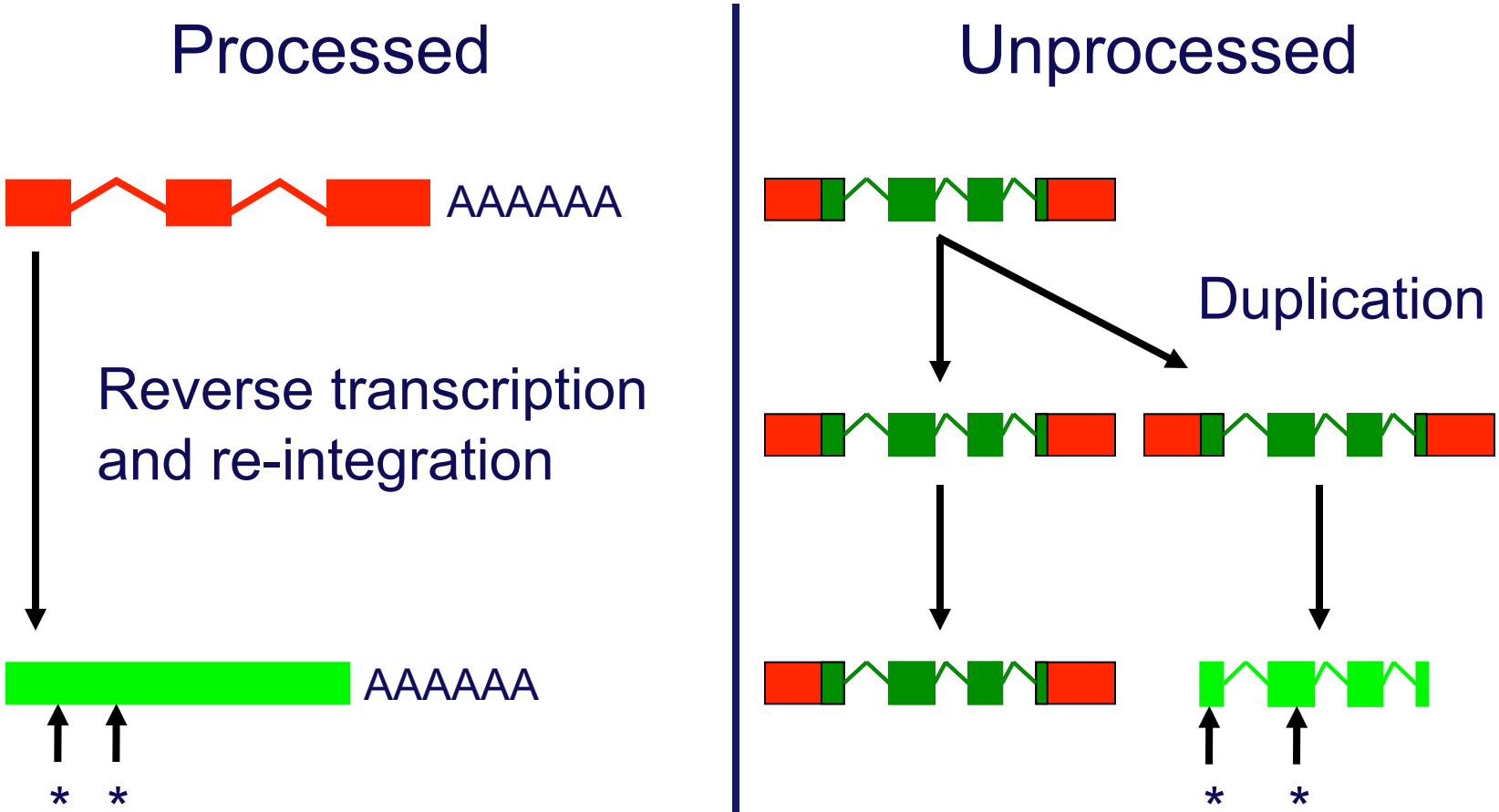


TIBs Vol 33:8

Transcript biotypes:
Schematic of lncRNA

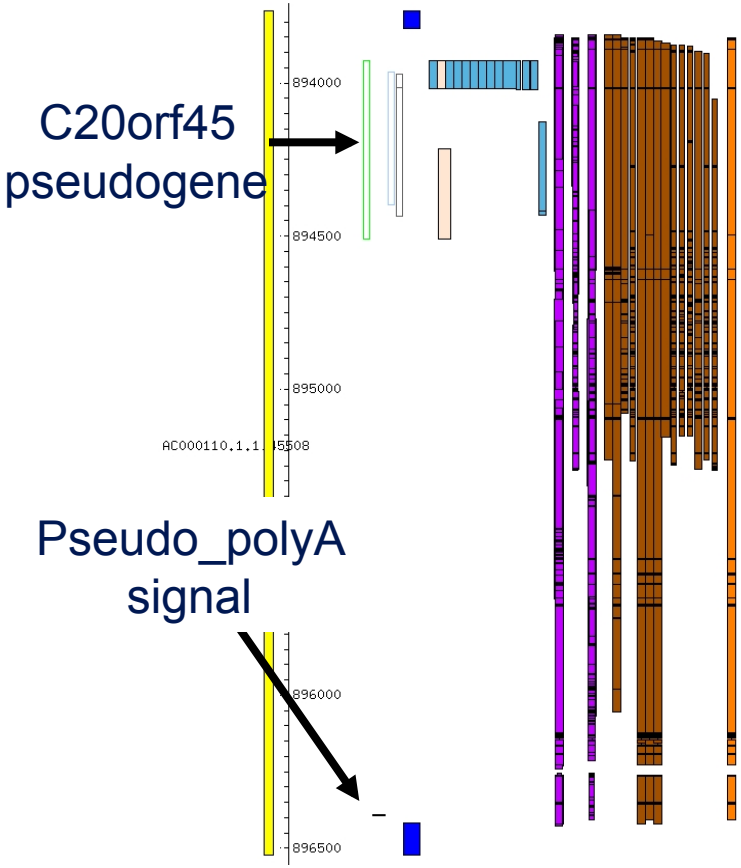


Pseudogene Loci:

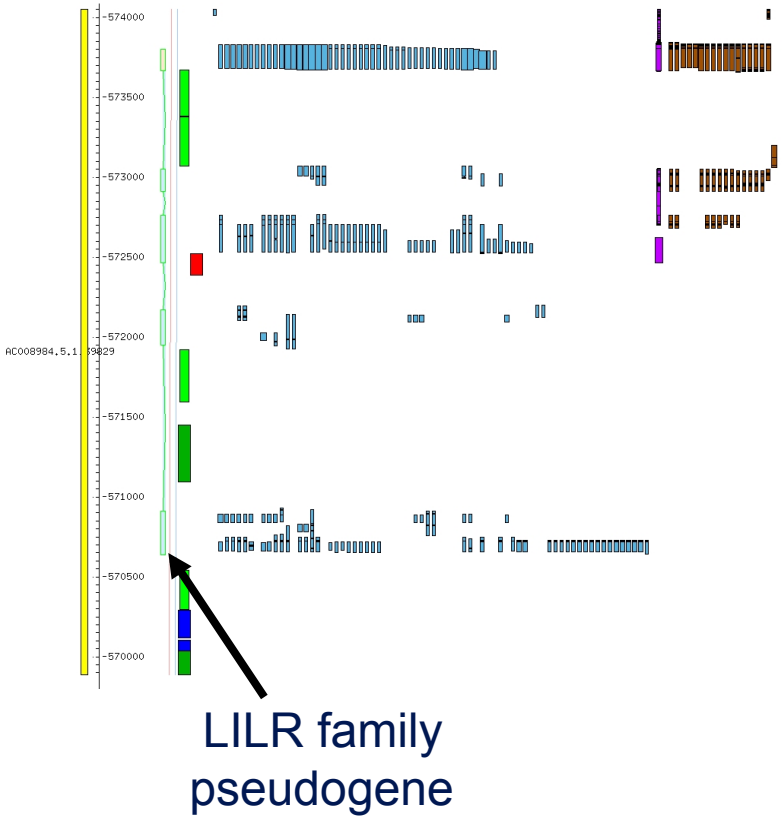


HAVANA Pseudogene Loci examples:

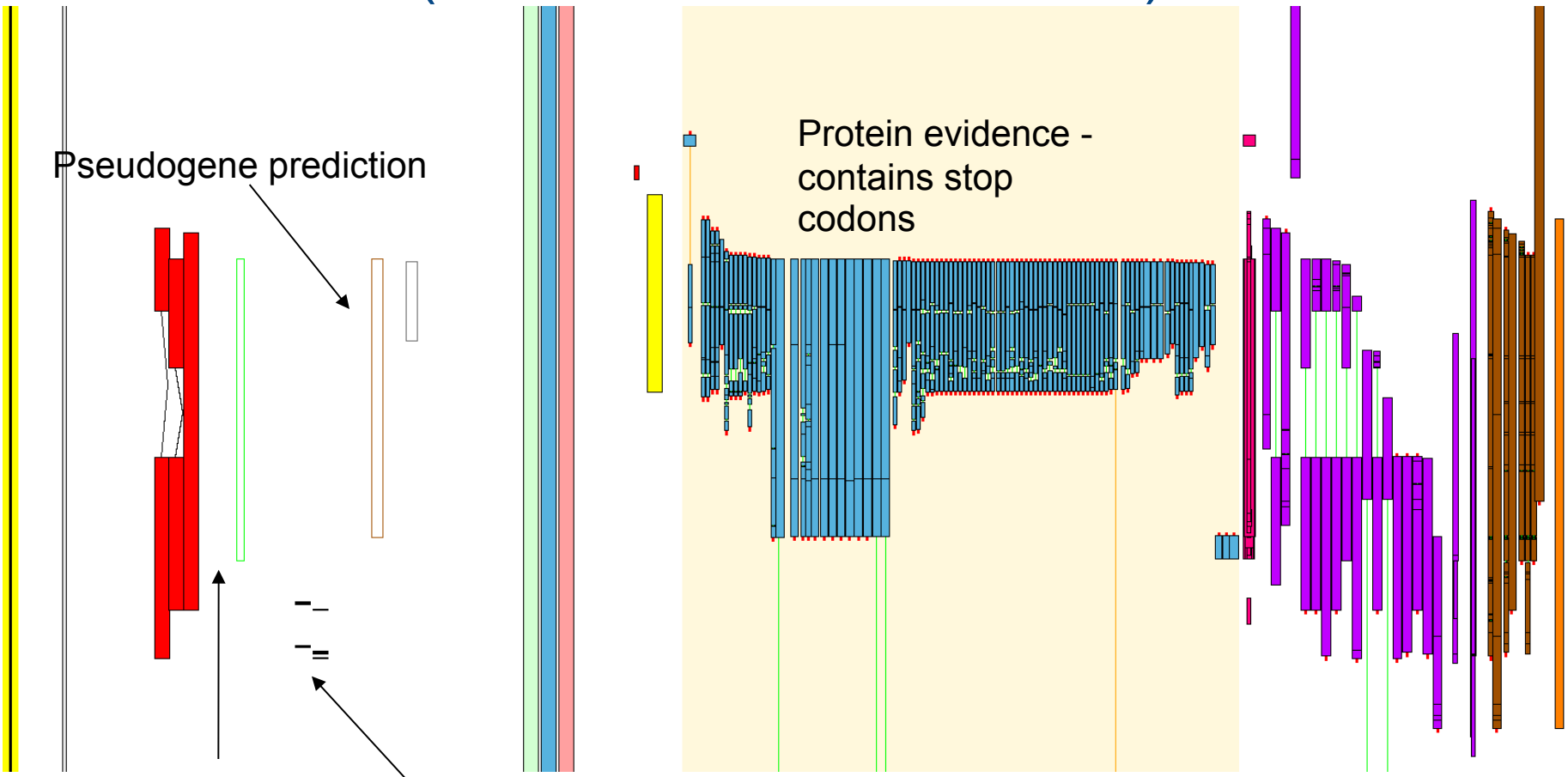
Processed



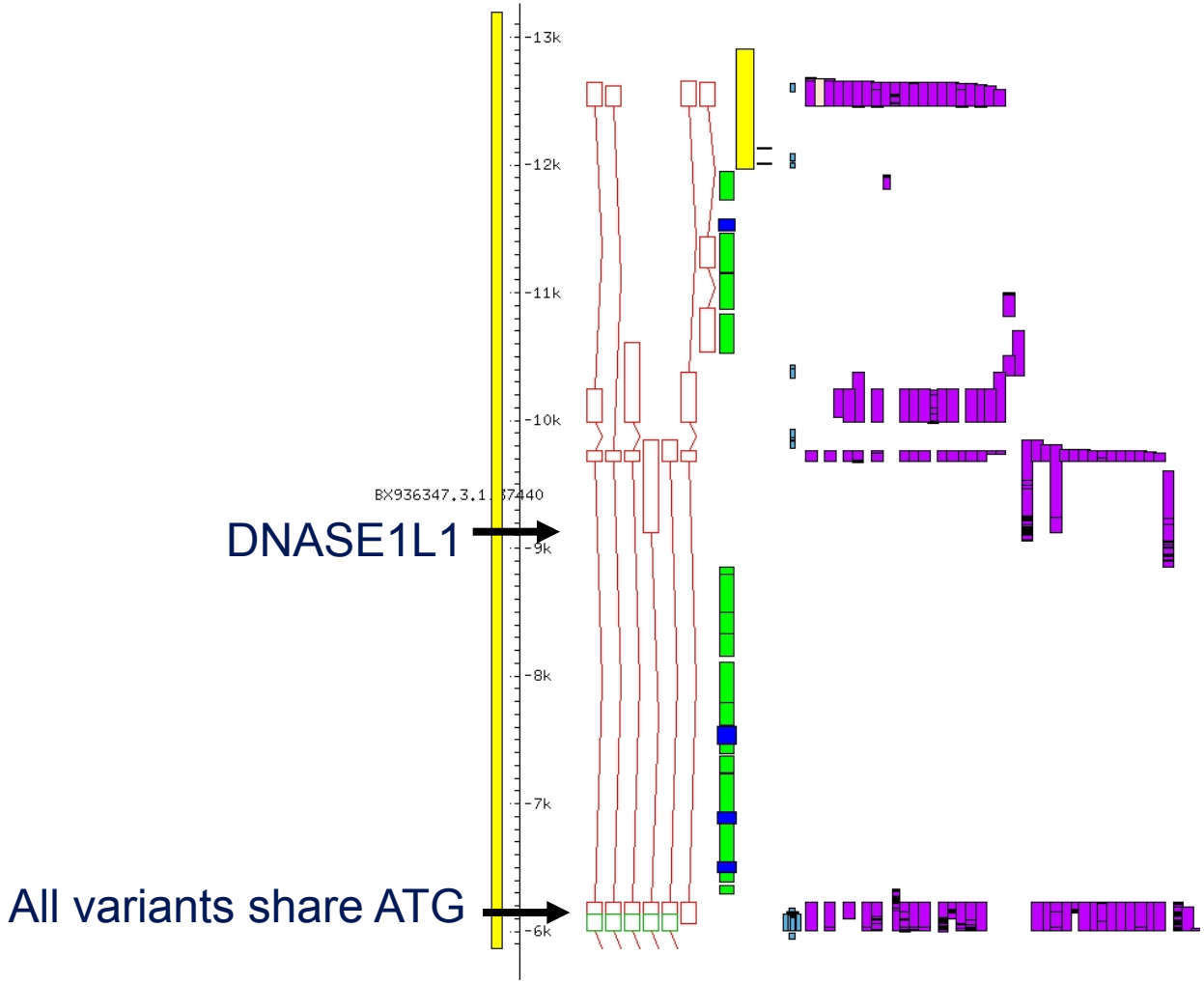
Unprocessed



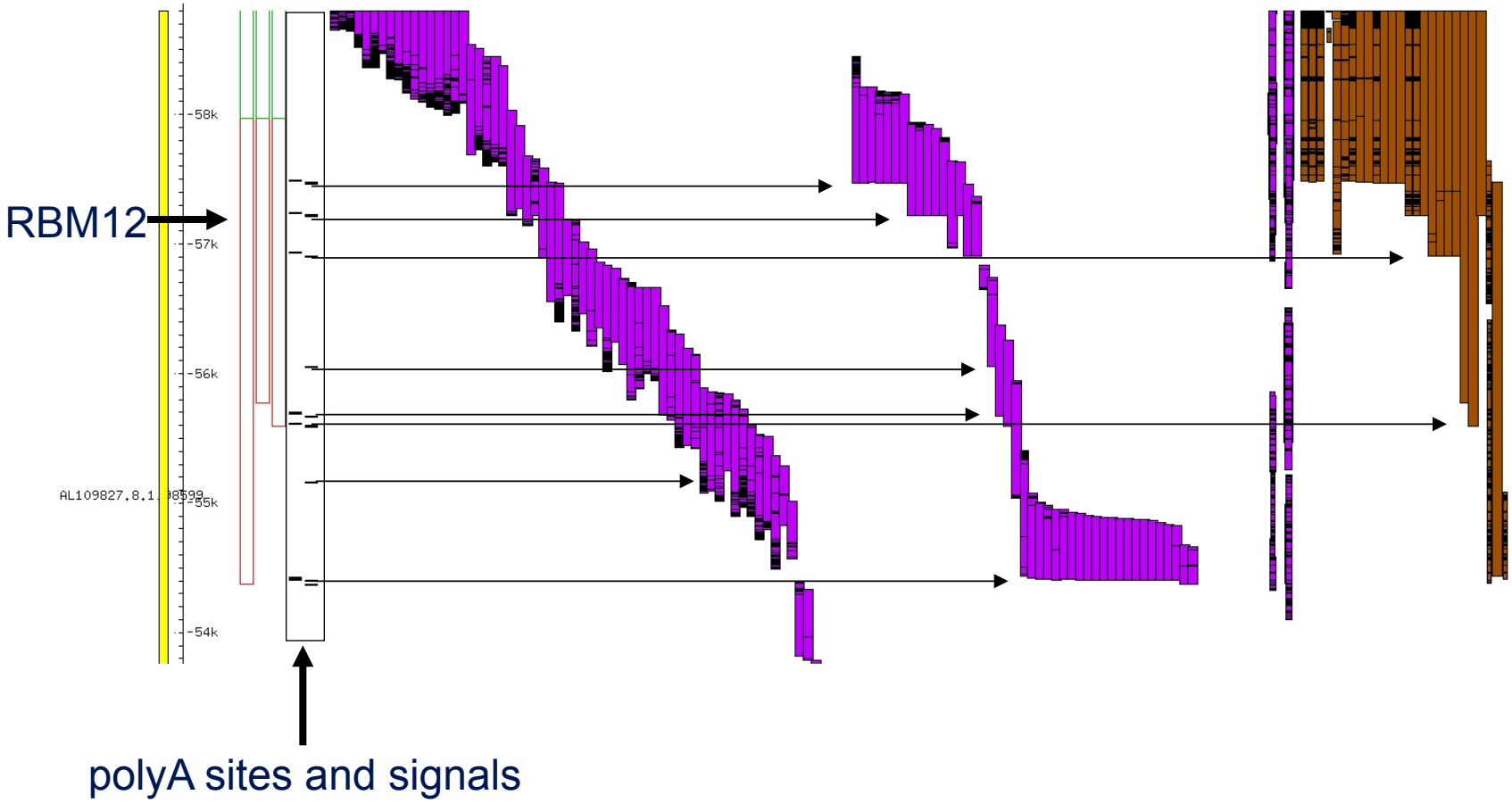
Transcribed processed pseudogene: functional ? (OTTHUMT00000130640)



Gene Structure - 5’ End



Gene Structure - 3’ End



Vega: Portal for the data

Vega BLAST/BLAT | Help & Documentation

Login · Register

wellcome trust sanger institute

The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality manual annotation of vertebrate finished genome sequence. Human, mouse and zebrafish are in the process of being completely annotated, whereas for other species the annotation is only of specific genomic regions of particular biological interest. The majority of the annotation is from the HAVANA group at the Welcome Trust Sanger Institute

havana

The website is built upon code from the Ensembl project.

Search: All species for Go

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

Browse a genome

- ***Human*** [12-01-2012] Ensembl
- ***Mouse*** [12-01-2012] Ensembl
- ***Zebrafish*** [24-10-2011] Ensembl
- ***Chimpanzee*** [12-01-2012] Ensembl
- ***Gorilla*** [30-03-2009] Ensembl
- ***Wallaby*** [30-03-2009] Ensembl
- ***Pig*** [16-05-2007] Ensembl
- ***Dog*** [14-02-2005] Ensembl

What's New in Release 46 (12 January 2012)

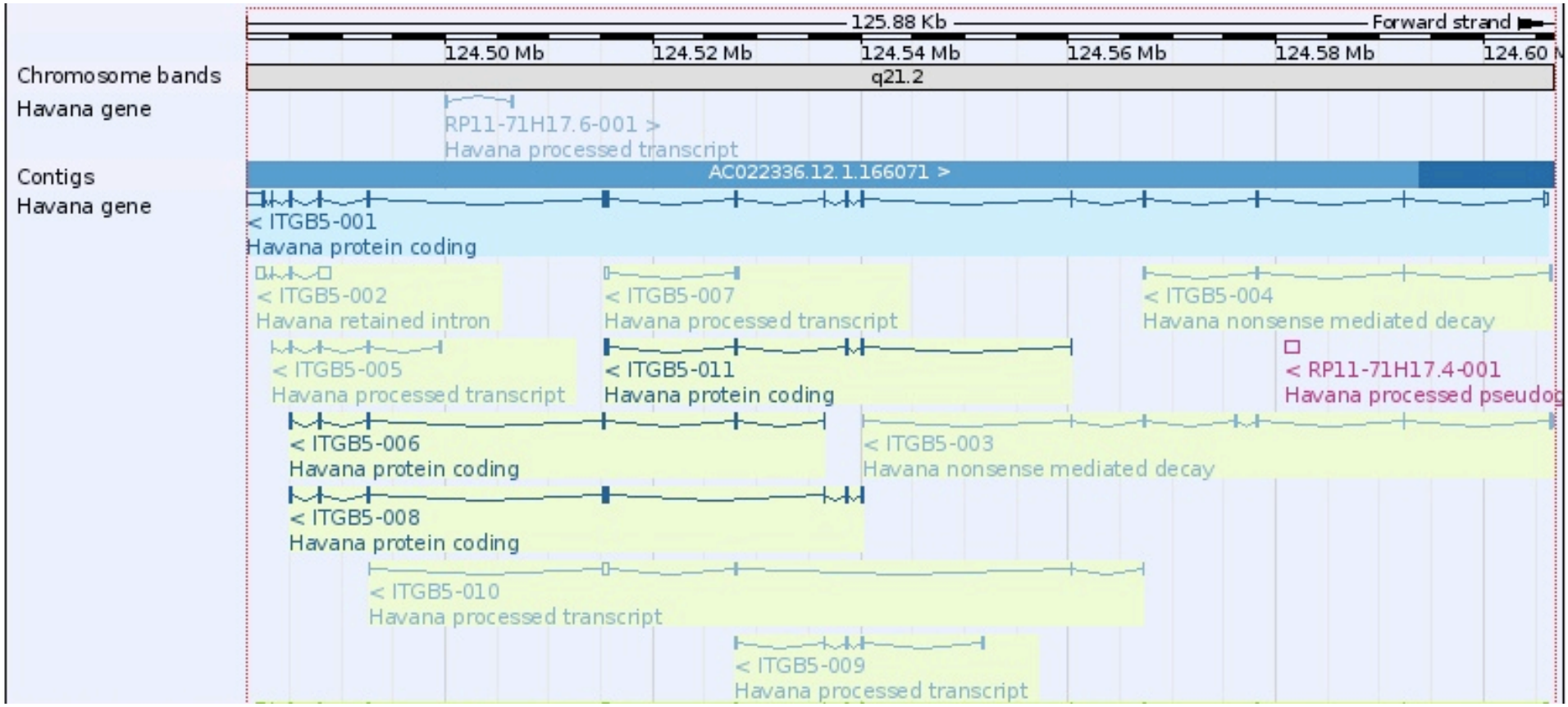
- **Update to human annotation** (Human)
- **Update to mouse annotation** (Mouse)
- **Chimpanzee annotation** (Chimpanzee)
- **Website enhancements** (all species)
- **Schema change** (all species)

More news...

What's New in Release 45 (24 October 2011)

- **Update to human annotation** (Human)
- **Update to zebrafish annotation** (Zebrafish)
- **Update to MGI links** (Mouse)
- **Website enhancements** (all species)
- **Schema change** (all species)

More news...


125.88 Kb
Forward strand
124.50 Mb
124.52 Mb
124.54 Mb
124.56 Mb
124.58 Mb
124.60
Chromosome bands
q21.2
Havana gene
RP11-71H17.6-001 >
Havana processed transcript
Contigs
AC022336.12.1.166071 >
Havana gene
< ITGB5-001
Havana protein coding
< ITGB5-002
Havana retained intron
< ITGB5-007
Havana processed transcript
< ITGB5-004
Havana nonsense mediated decay
< ITGB5-005
Havana processed transcript
< ITGB5-011
Havana protein coding
< RP11-71H17.4-001
Havana processed pseudog
< ITGB5-006
Havana protein coding
< ITGB5-003
Havana nonsense mediated decay
< ITGB5-008
Havana protein coding
< ITGB5-010
Havana processed transcript
< ITGB5-009
Havana processed transcript

Locus summary:

Gene: Elk1 (OTTMUSG00000017185)

ELK1, member of ETS oncogene family

Location Chromosome X: 20,510,521-20,527,734 reverse strand.

Transcripts There are 2 transcripts in this gene: **hide transcripts**

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
Elk1-001	OTTMUST00000041618	3548	OTTMUSP00000018693	429	Protein coding	CCDS30048
Elk1-002	OTTMUST00000041619	2246	No protein product	-	Retained intron	-

Gene summary help

Curated Locus Elk1 (MGI Symbol)

CCDS This gene is a member of the Mouse CCDS set: CCDS30048

Gene type Known protein coding [Definition]

Author This transcript was annotated by Havana <vega@sanger.ac.uk>

Version & date Version 2, last modified on 04/06/2008 (Created on 25/05/2006)

Other assemblies This gene maps to to 20,510,521-20,527,734 in NCBIM37 (Ensembl) coordinates.
Jump to this stable ID in Ensembl

MGI Symbol: Elk1 [view all locations]

Ensembl Mouse Gene: ENSMUSG00000009406 [view all locations]

UniProtKB/Swiss-Prot: A9L8T0 [align] [view all locations]

RefSeq peptide: NP_031948 [align] [view all locations]

RefSeq DNA: NM_007922 [align] [view all locations]

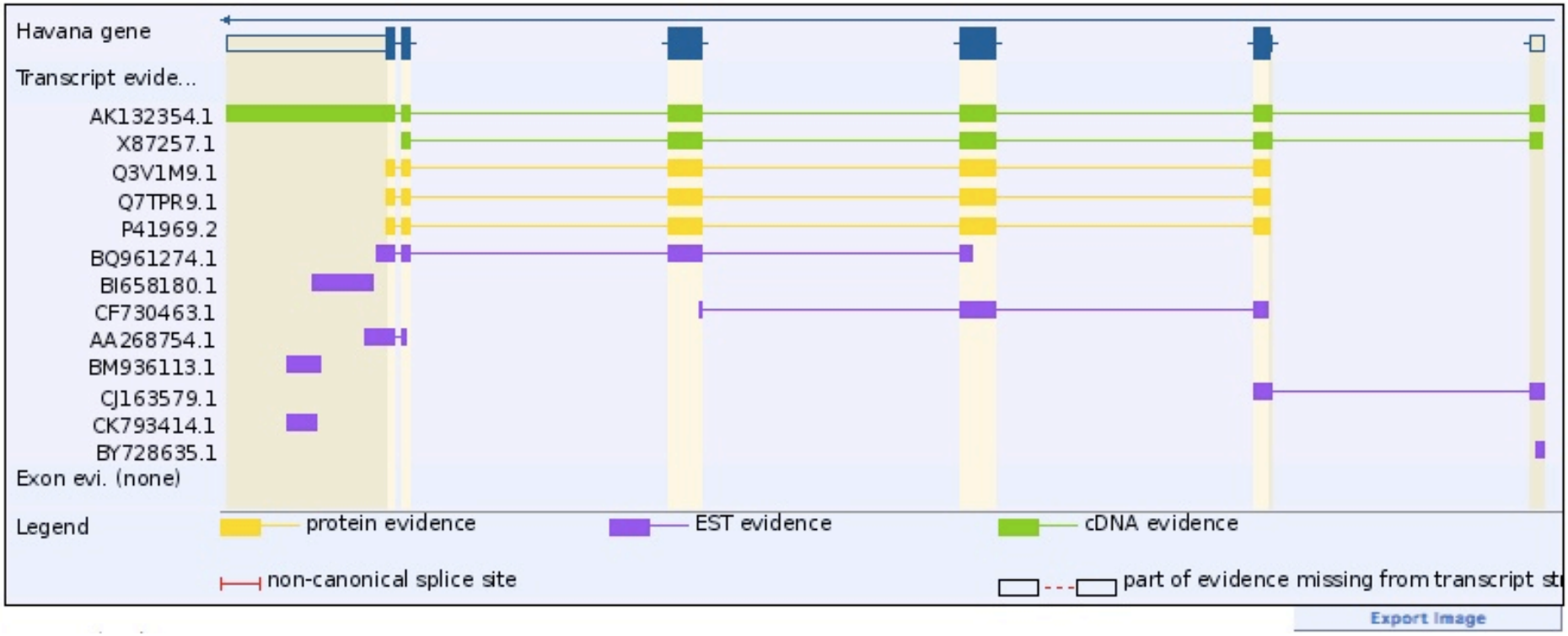
EntrezGene: 13712 [view all locations]

CCDS

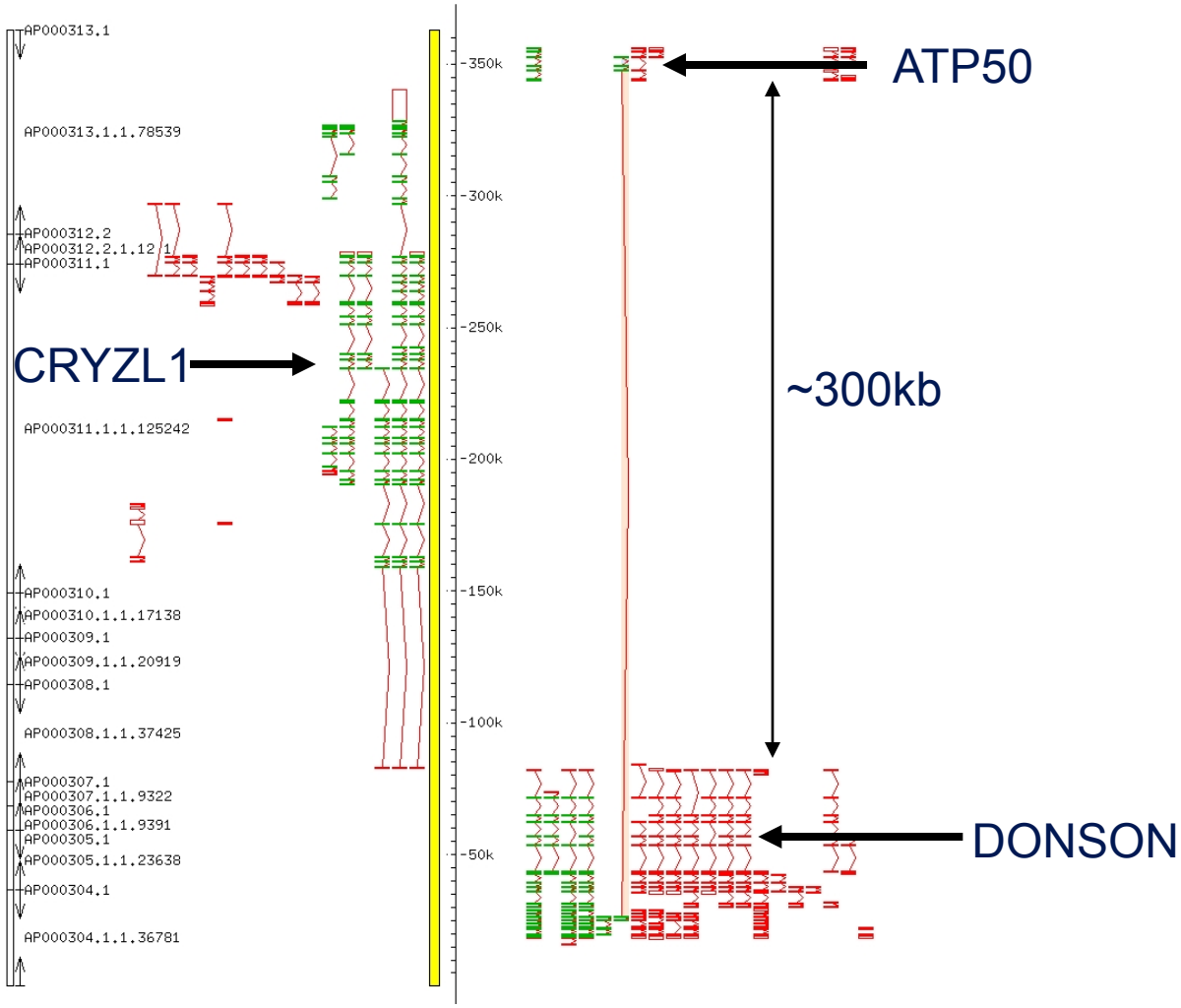
Annotation date

xrefs

Evidence used to build transcripts



Linked loci


AP000313.1
AP000313.1.1.78539
AP000312.2
AP000312.2.1.12/1
AP000311.1
CRYZL1
AP000311.1.1.125242
AP000310.1
AP000310.1.1.17138
AP000309.1
AP000309.1.1.20919
AP000308.1
AP000308.1.1.37425
AP000307.1
AP000307.1.1.9322
AP000306.1
AP000306.1.1.9391
AP000305.1
AP000305.1.1.23638
AP000304.1
AP000304.1.1.36781
-350k
-300k
-250k
-200k
-150k
-100k
-50k
ATP50
~300kb
DONSON

HOXA gene cluster

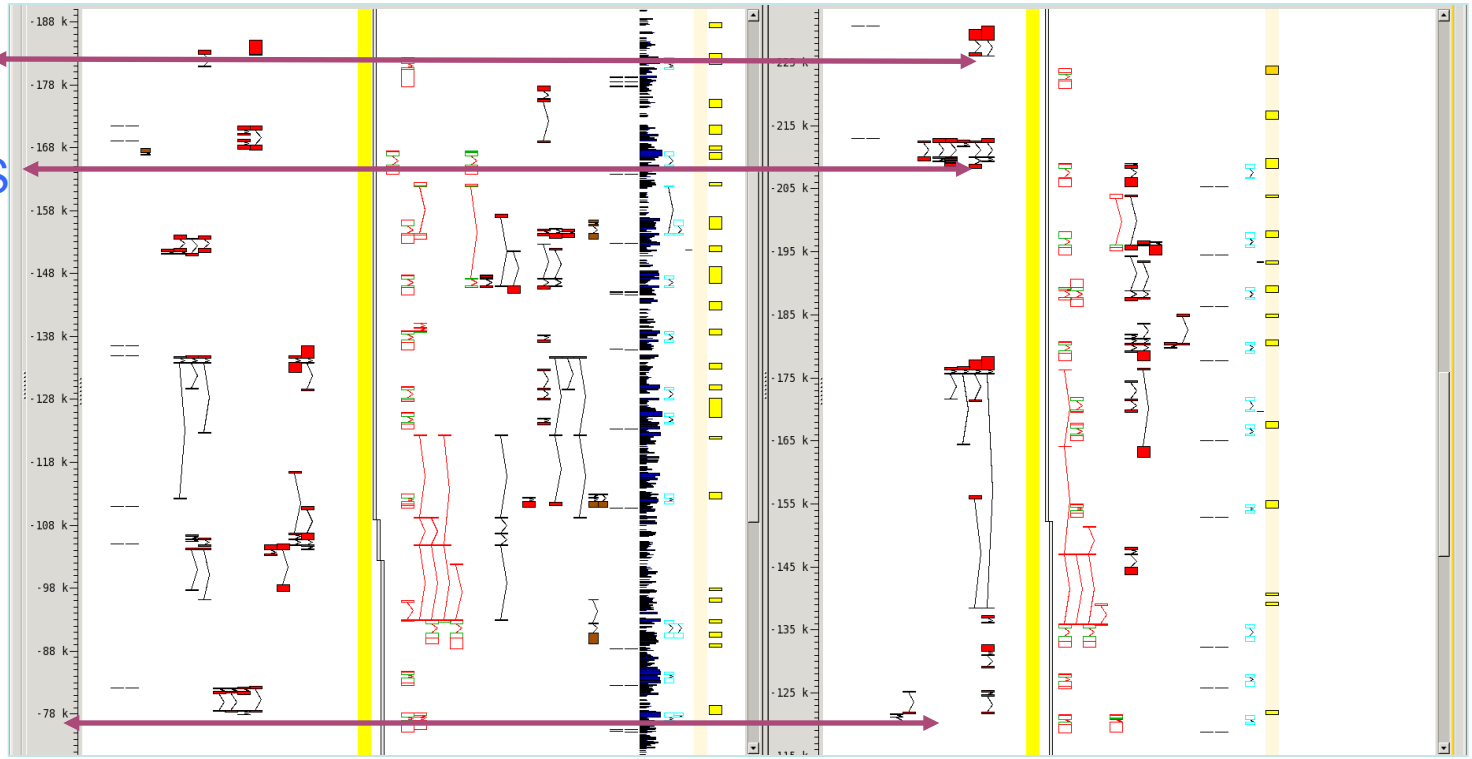
Human chr 7p15.2

Mouse chr 6qB3

HIT18844

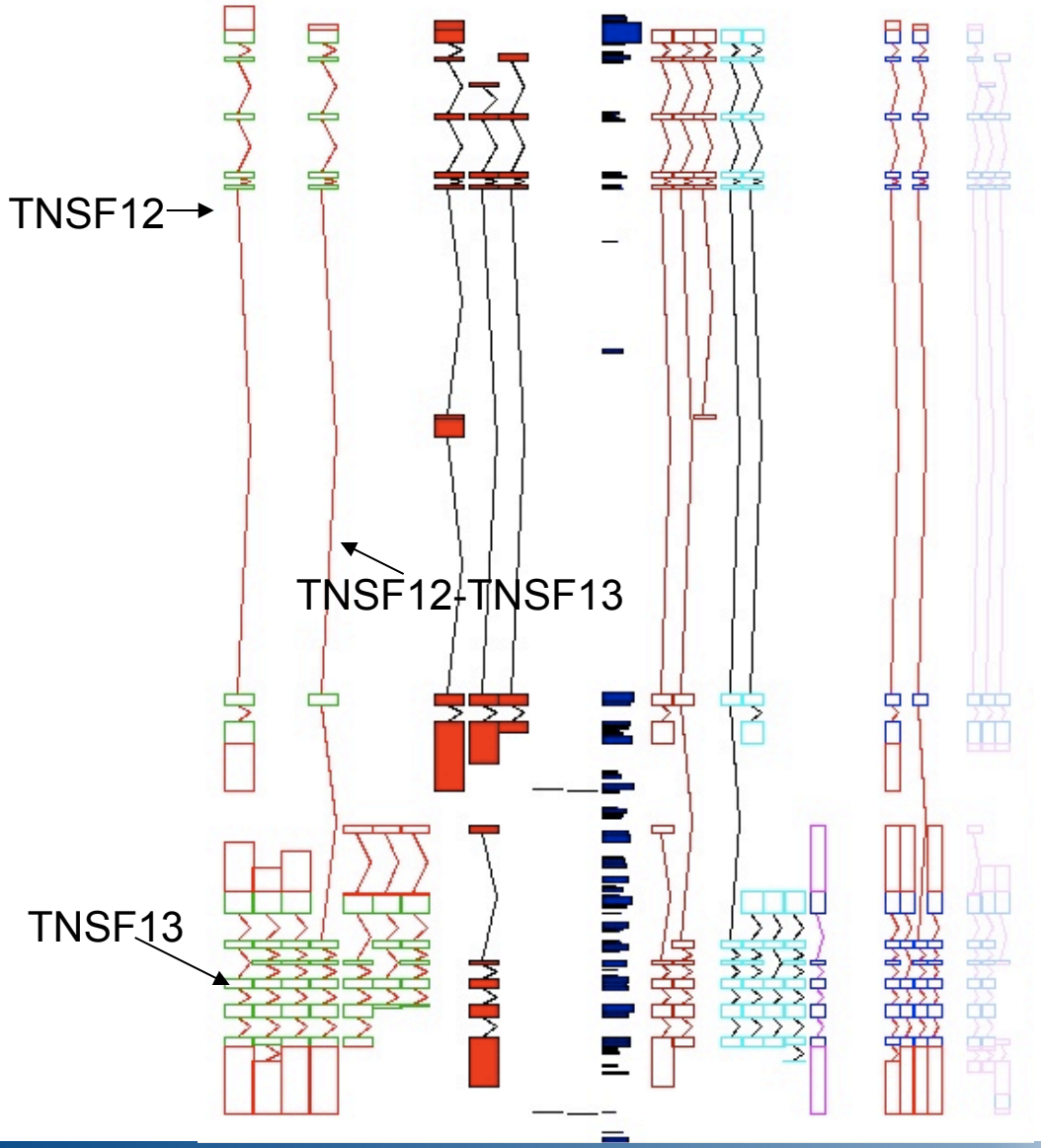
HOXA11AS

HOTAIRM1



Long non-coding transcripts are conserved across species and regulate expression of HOX genes

Readthroughs/fusion proteins:



Human haplotypes in VEGA:

MHC:

- Reference (PGF)
- 6-COX
- 6-QBL
- 6-SSTO
- 6-APD
- 6-DBB
- 6-MANN
- 6-MCF

LRC:

- 19-COX
- 19-PGF_1

Other species MHC:

Gorilla [30-03-2009]
Ensembl

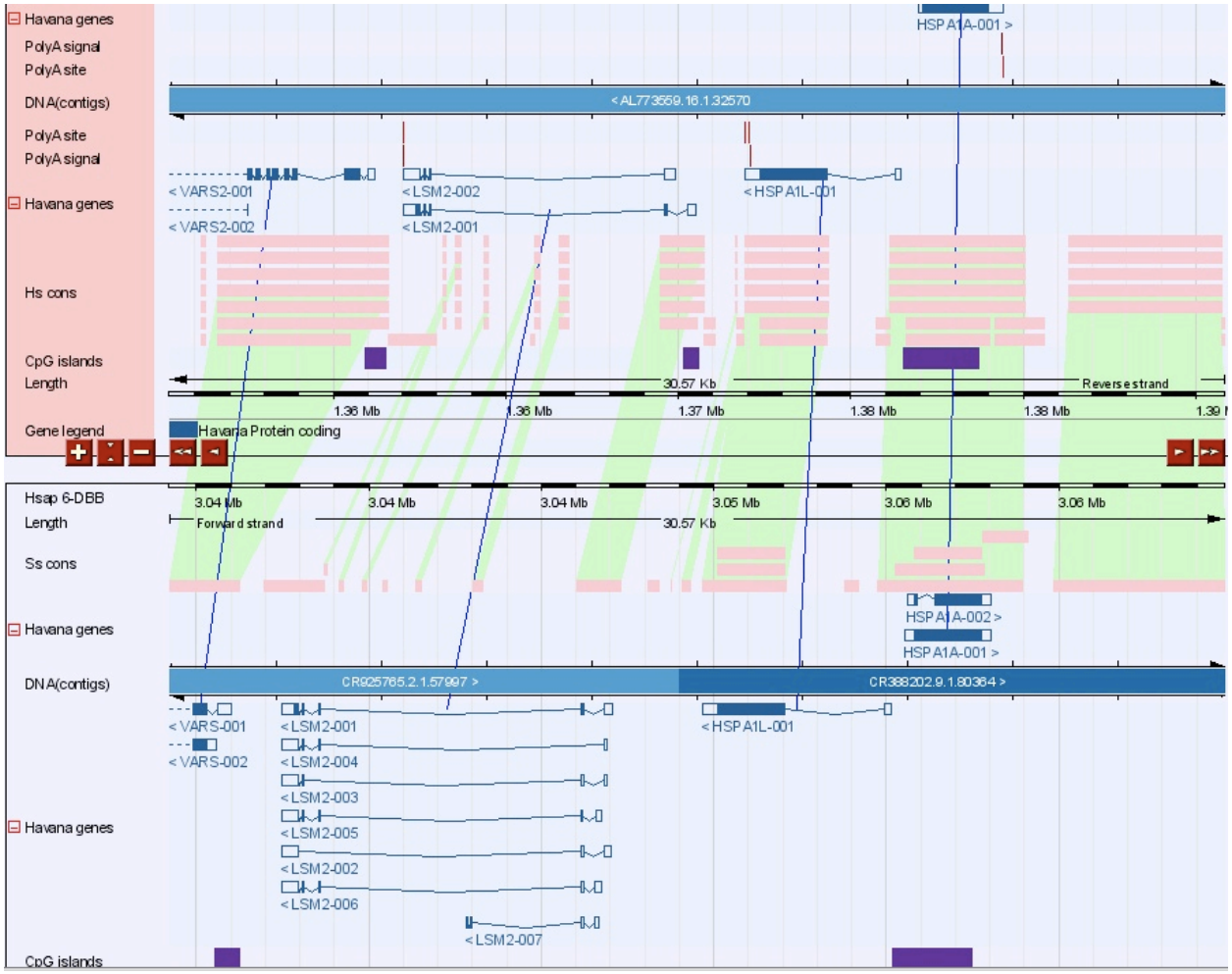
Wallaby [30-03-2009]
Ensembl

Pig [16-05-2007]
Ensembl

Dog [14-02-2005]
Ensembl

Chimpanzee [12-01-2012]
Ensembl

Multicontigview:
Compare regions in MHC between pig and human



Community Annotation:

- Otterlace/Zmap software available for Mac and Linux

- Part of IKMC with EUCOMM annotation in mouse:
    - KOMP and NorCOMM annotation

- Jamborees for species with strong community interest:

    - *Xenopus tropicalis* 2005 (cDNA)

    - Cow 2007 (Genomic WGS) Publication

    - Pig
    - 2008 (Genomic WGS)
    - 2010 - 2012
        - IR genes in Pig (~2000 genes) manually annotated by community
        - Chromosomes X and Y to be fully finished and annotated by Havana

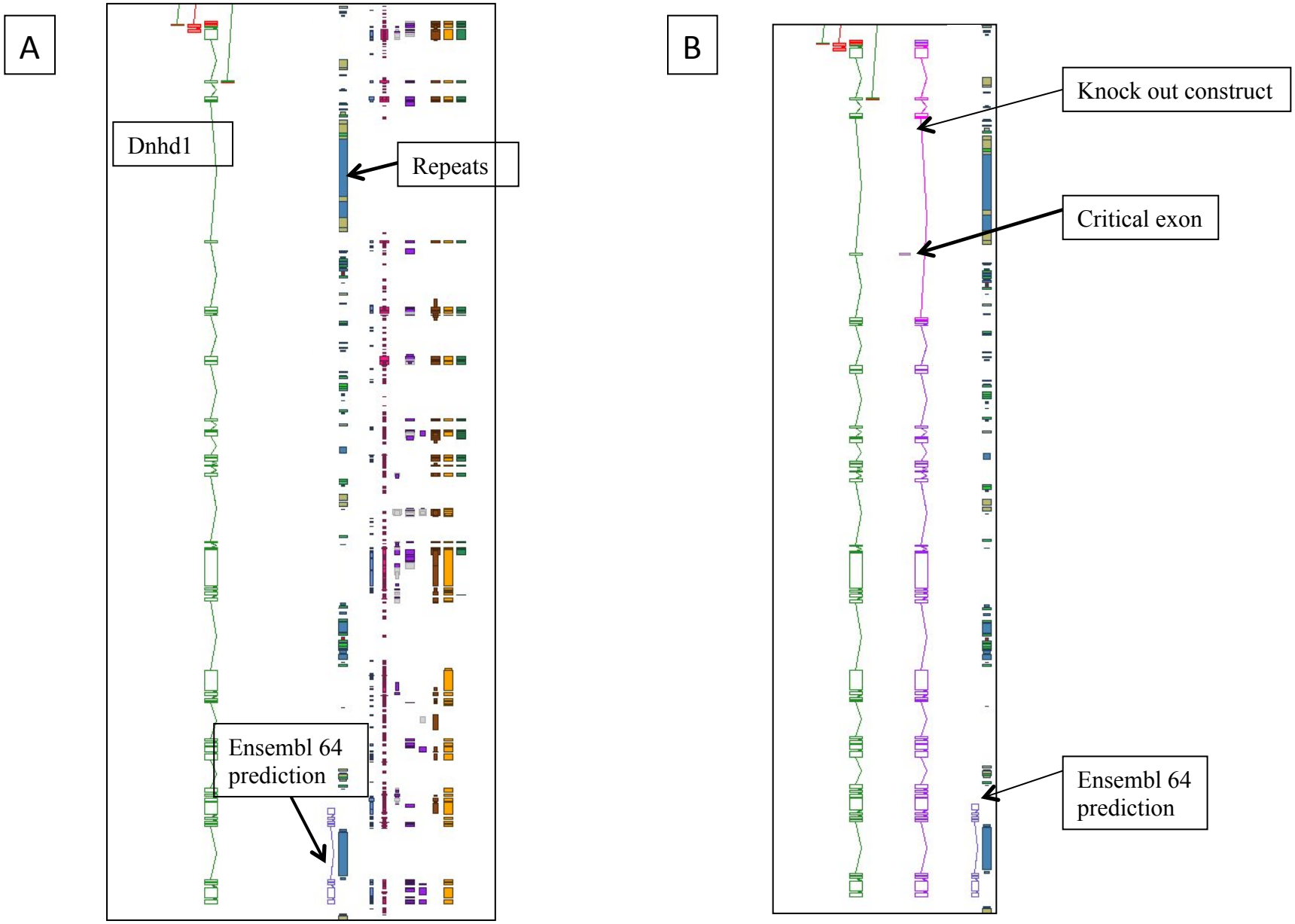
Community Annotation Approaches:
The value of a genome is only as good as its annotation

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

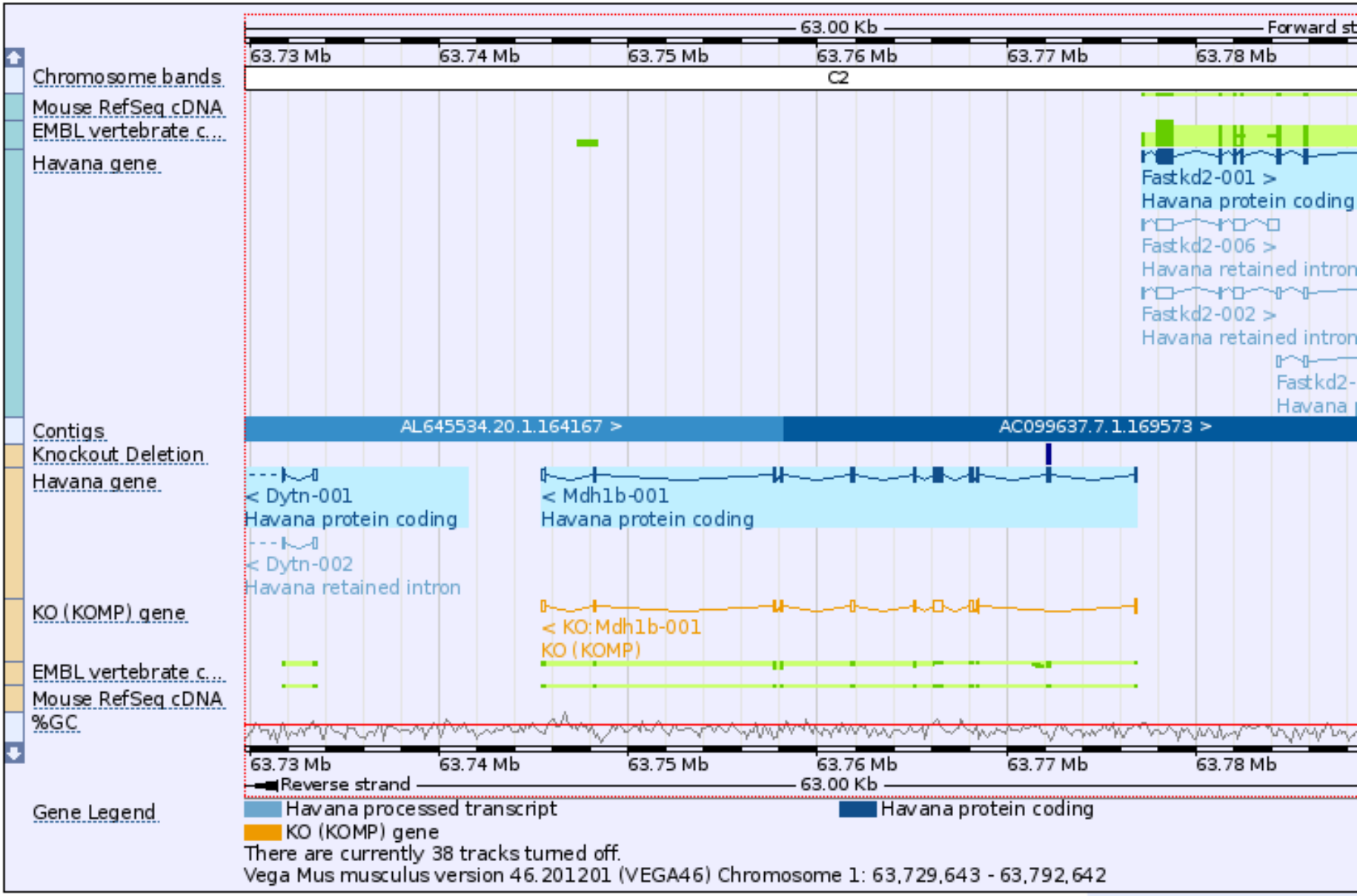
"Blessed Annotator"
Mouse KOMP and NorCOMM (part of IKMC)
External annotators (3) trained from Wash U and U Manitoba
Identifying critical exons to make knock-out constructs
6 months of training and QC – Integrated into mouse gene build

"Gatekeeper"
Swine autosomes
External annotators worldwide (~30)
Short training for European and US groups, plus regular calls and WebEx
Guidance and QC by WTSI

Mouse KOMP annotation



View KO's in Vega

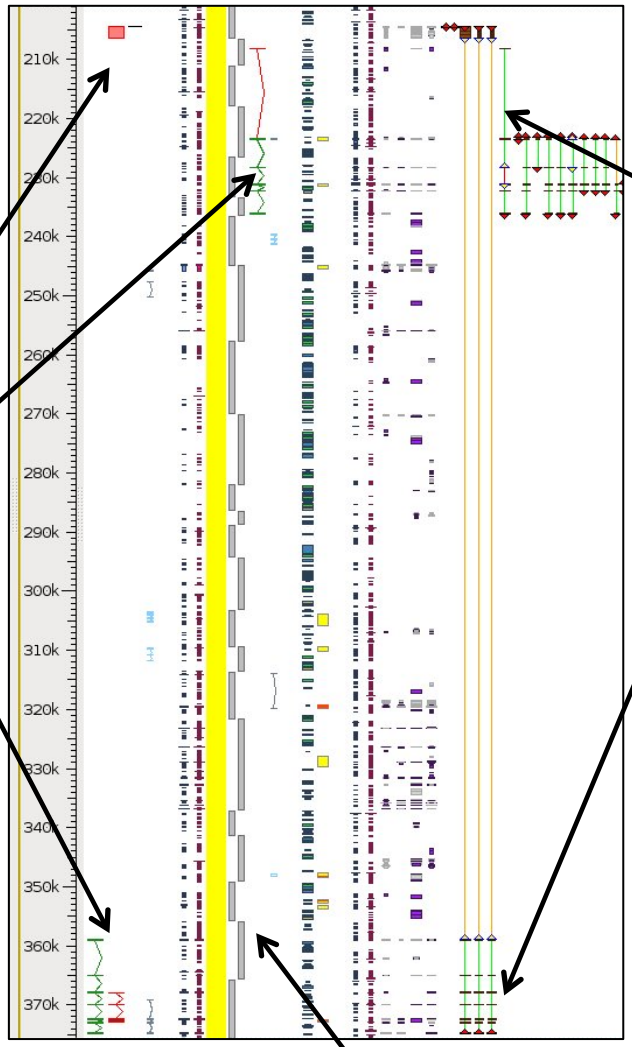


Swine Immune Response Annotation Group (IRAG)

Chris Tuggle (Iowa State), Claire Rogel-Gaillard (INRA)

USA:	Iowa State	China:	Huazong Agricultural University
	USDA		
	Michigan State	Europe:	INRA
	Univ Minnesota		Parco Tecnologico Padano
	Oaklahoma State		Roslin
	Kansas State		UCL
			WTSI
Japan:	AFFRC		
	STAFF		~30 annotators!

Unordered contigs in pig



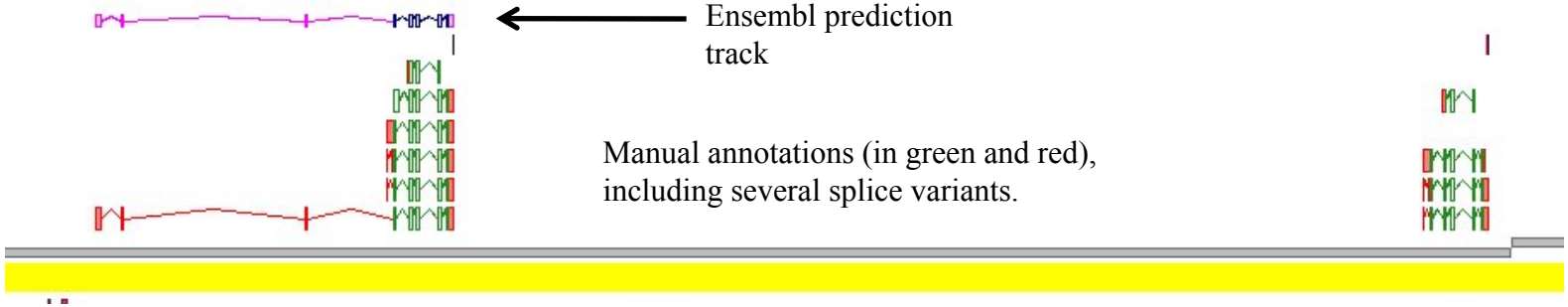
Fragmented manual Annotation of CRISPLD2 (in red and green).

Vertebrate mRNA homology matches

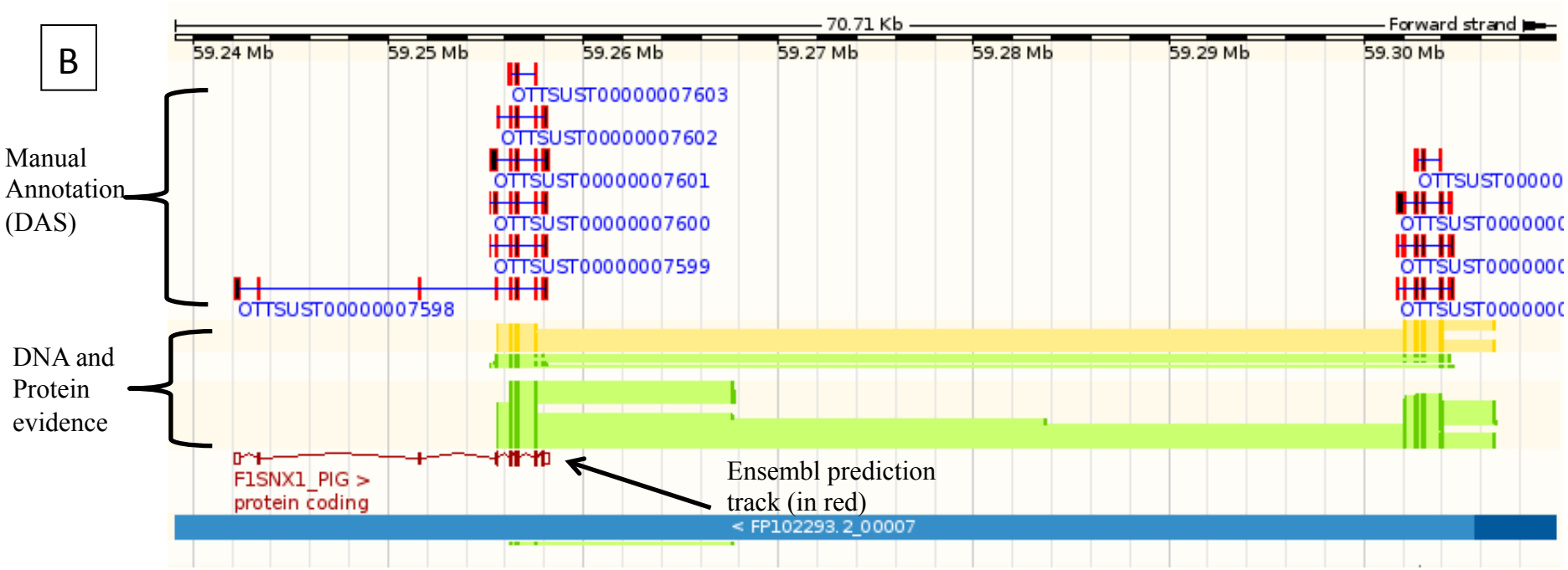
Grey vertical bars represent unordered contigs.

Gene duplications in pig

A

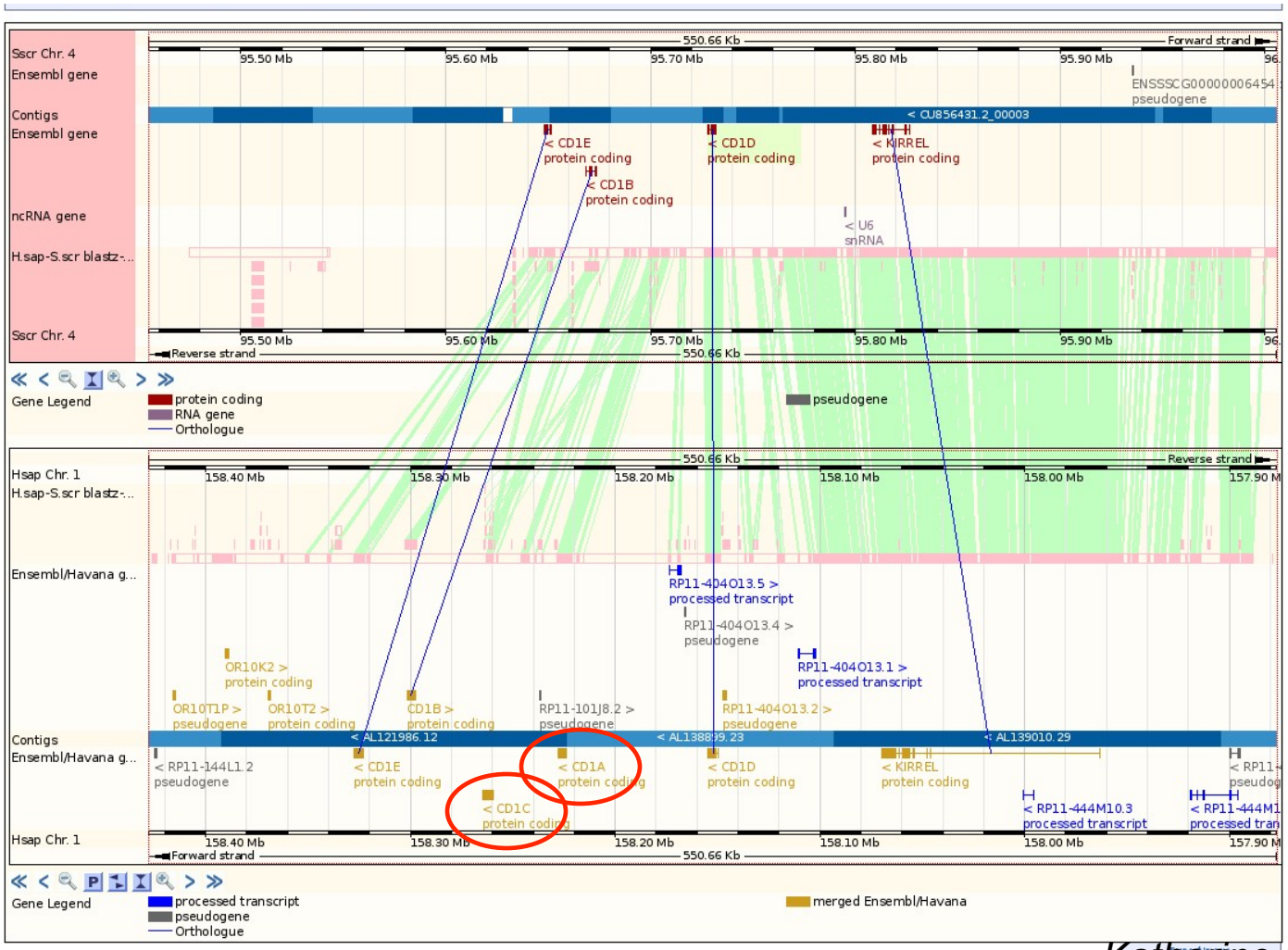


B



REG3A gene

Missing genes from pig?
Ensembl multi-species view



Family of glycoproteins, related to class 1 MHC. Activate natural kller T cells

Katherine Mann

wellcome trust sanger institute

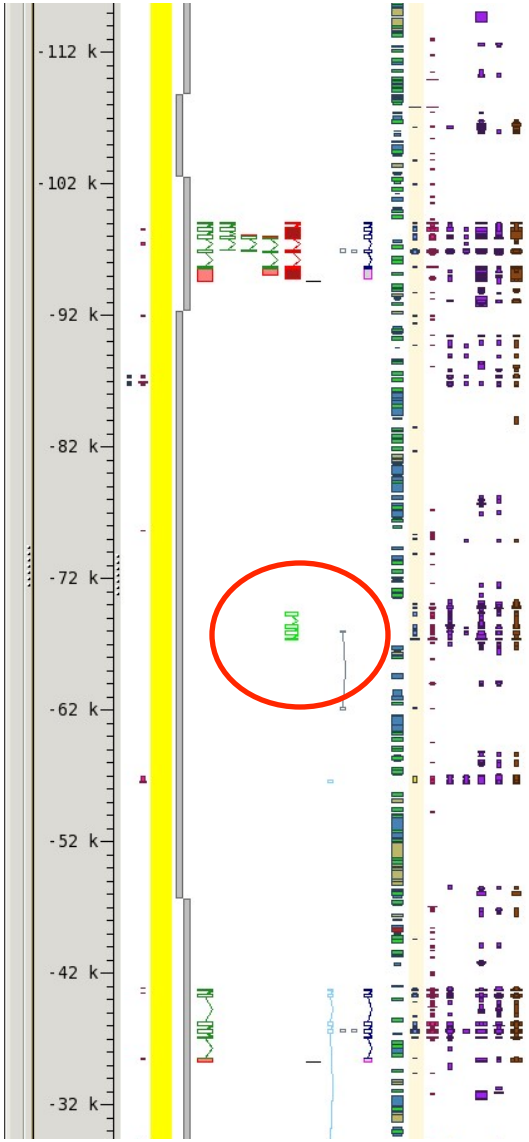
havana human and vertebrate analysis and annotation

Manual annotation:

CD1D

CD1 family pseudogene

CD1B



Expansions in cow and dog, but not pig?

Comparative annotation:

Denise Carvalho-Silva

Community Annotation: Summary

“Blessed Annotator”:

Extended training means less QC
Wide range of biotypes
Annotation figures: KOMP 1876 genes, NorCOMM 378 genes

“Gatekeeper”:

Shorter training means more QC
Annotation figures: Pig IRAG 1276 genes

Lessons Learned:

QC

How to maintain quality with diverse annotation expertise

Training

Essential, plus regular updates (WebEx)

Nomenclature

Swine Nomenclature Committee

What next:

Merge the IRAG manual annotation with the automated Ensembl annotation:

~ 8% of genome

Extend annotation / collaboration:

Refined gene list for IRAG

QC: Complex and partial genes

Publications

Acknowledgements

Havana:
Jen Harrow
If Barnes
Ruth Bennett
Alex Bignell
Veronika Boychenko
Gloria Despacio-Reyes
Sarah Donaldson
Adam Frankish
Matthew Hardy
Toby Hunt
Mike Kay
Gavin Laird
David Lloyd
Jane Loveland
Deepa Manthravadi
Gaurab Mukherjee
Jonathan Mudge
Jeena Rajan
Liam Redgrave
Gary Saunders
Catherine Snow
Charles Steward
Marie-Marthe Suner
Mark Thomas
Laurens Wilming

Anacode:
James Gilbert
Matthew Astley
Michael Gray
Jeremy Henty

Vega:
Stephen Trevanion
Maurice Hendrix

Zmap:
Ed Griffiths
Gemma Barson
Malcolm Hinsley

Mouse annotation:

KOMP:
Amy Horton

NorCOMM:
Molly Pind

IRAG annotators:

USA:
Jim Reecy
Chris Tuggle
Daniel Berman
Frank Blecha
Ryan Chen
Celine Chen
Daniel Ciobanu
Harry Dawson
Cathy Earnst
Zhiliang Hu
Joan Lunney
Katherine Mann
Michael Murtaugh
Yongming Sang
John Schwartz

China:
Shuhong Zhao

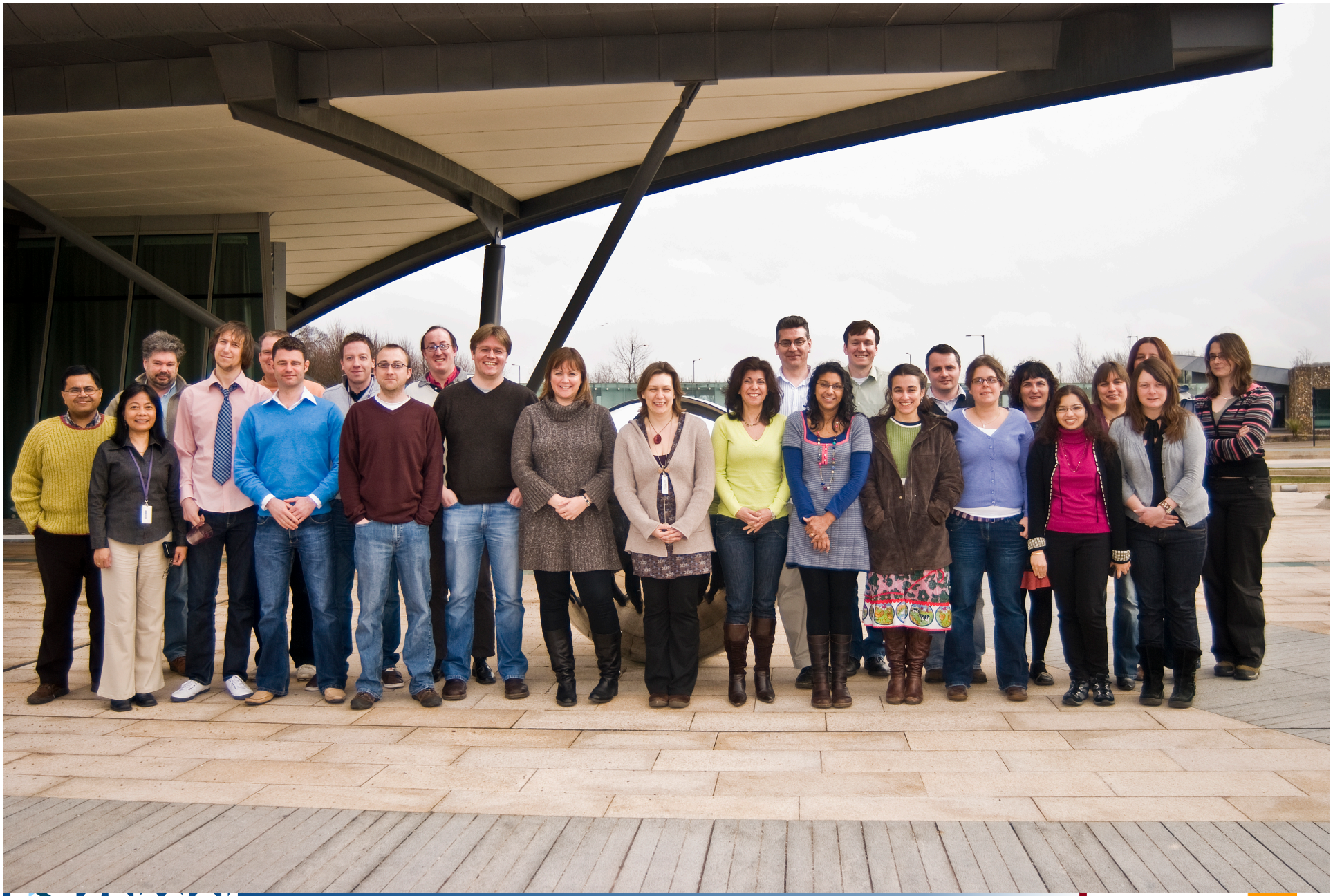
Japan:
Hirohide Huenishi
Takeya Morozumi
Hiroke Sinkai
Diasuke Toki

Europe:
Alan Archibald
Claire Rogel-Gaillard
Anna Anselmo
Bouabid Badauoi
Betrand Bed'Hom
Dario Beraldi
Lynsey Fairbairn
Elisabetta Giuffra
David Hume
Ronan Kapetanovich
Dennis Prickett
Christelle Robert
Yasu Takeuchi

http://vega.sanger.ac.uk

wellcome trust sanger institute

havana human and vertebrate analysis and annotation



sanger institute

havana
human and vertebrate analysis and annotation