# Can we apply lessons learned from manual annotation in human and mouse to wheat?


Jane Loveland

Annotation Project leader

EnsemblI-HAVANA


PAG XXVI, 16th January 2018

EMBL-EBI

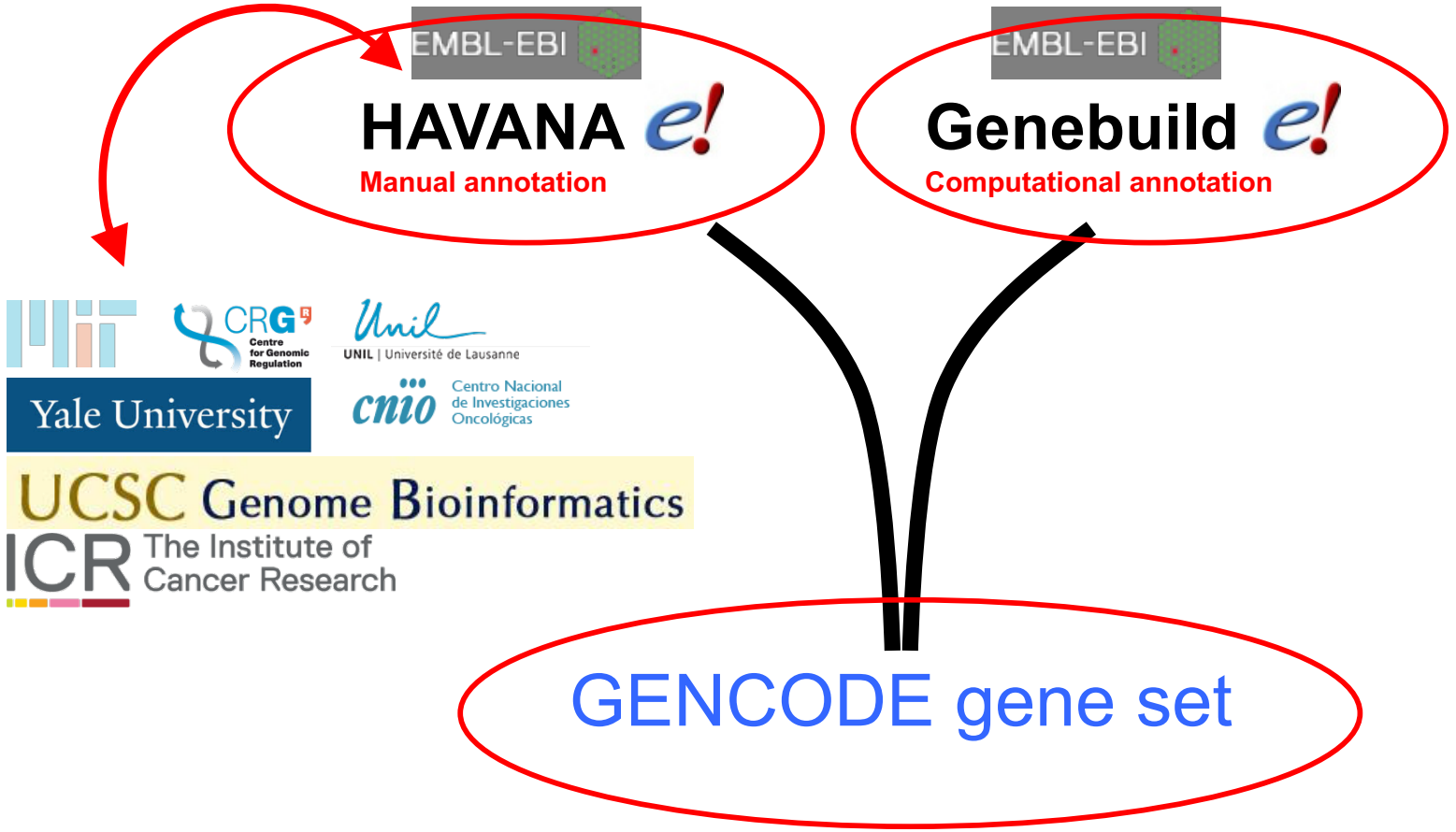
e!Ensembl



## 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,

# The GENCODE consortium



# The HAVANA team

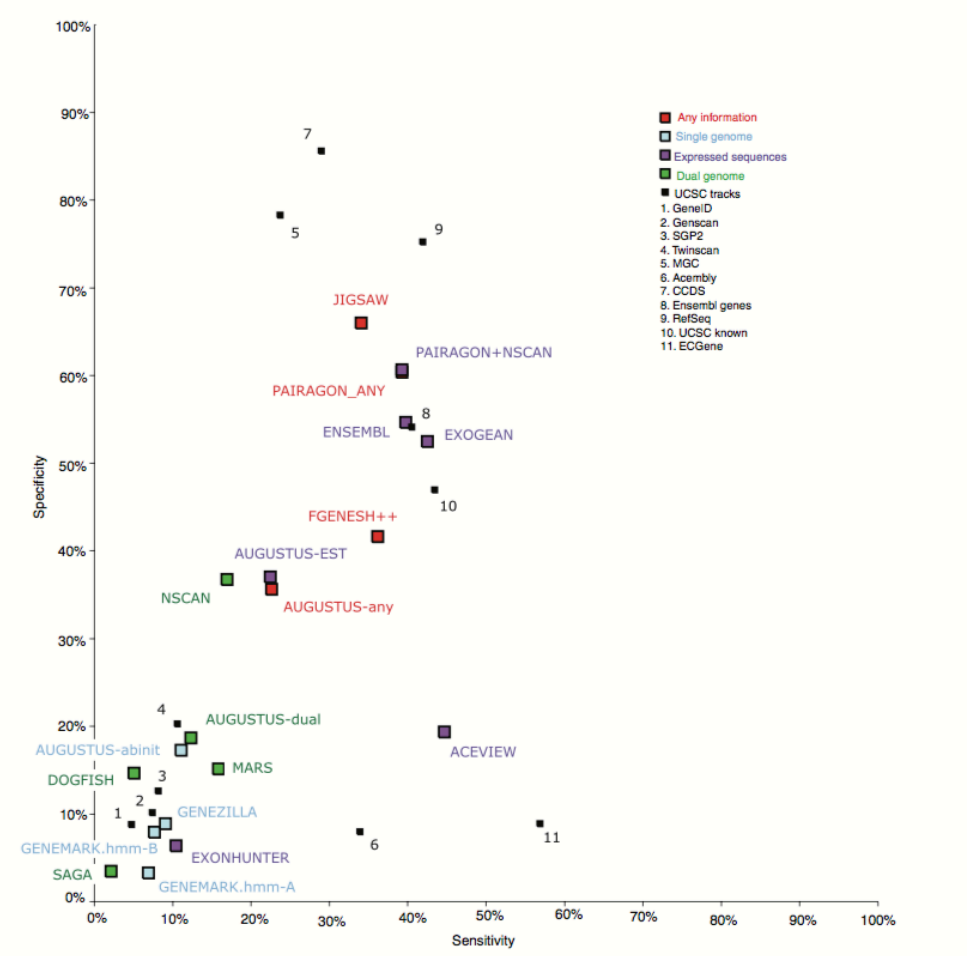
GENCODE

Whole Genome or chromosome

Targeted regions or genes

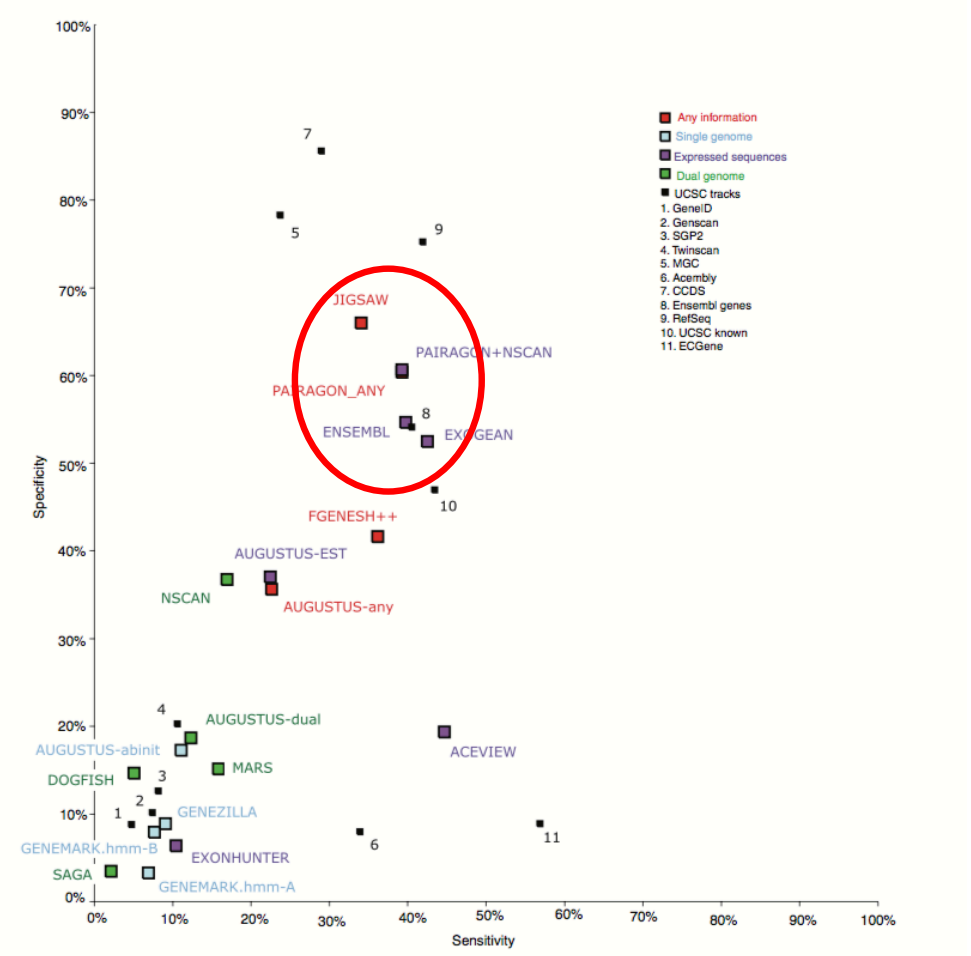
Community projects

# Requirement for manual annotation




EGASP - Guigó et al. Genome Biol. 7 Suppl 1:S2.1-31 (2006)

# Requirement for manual annotation



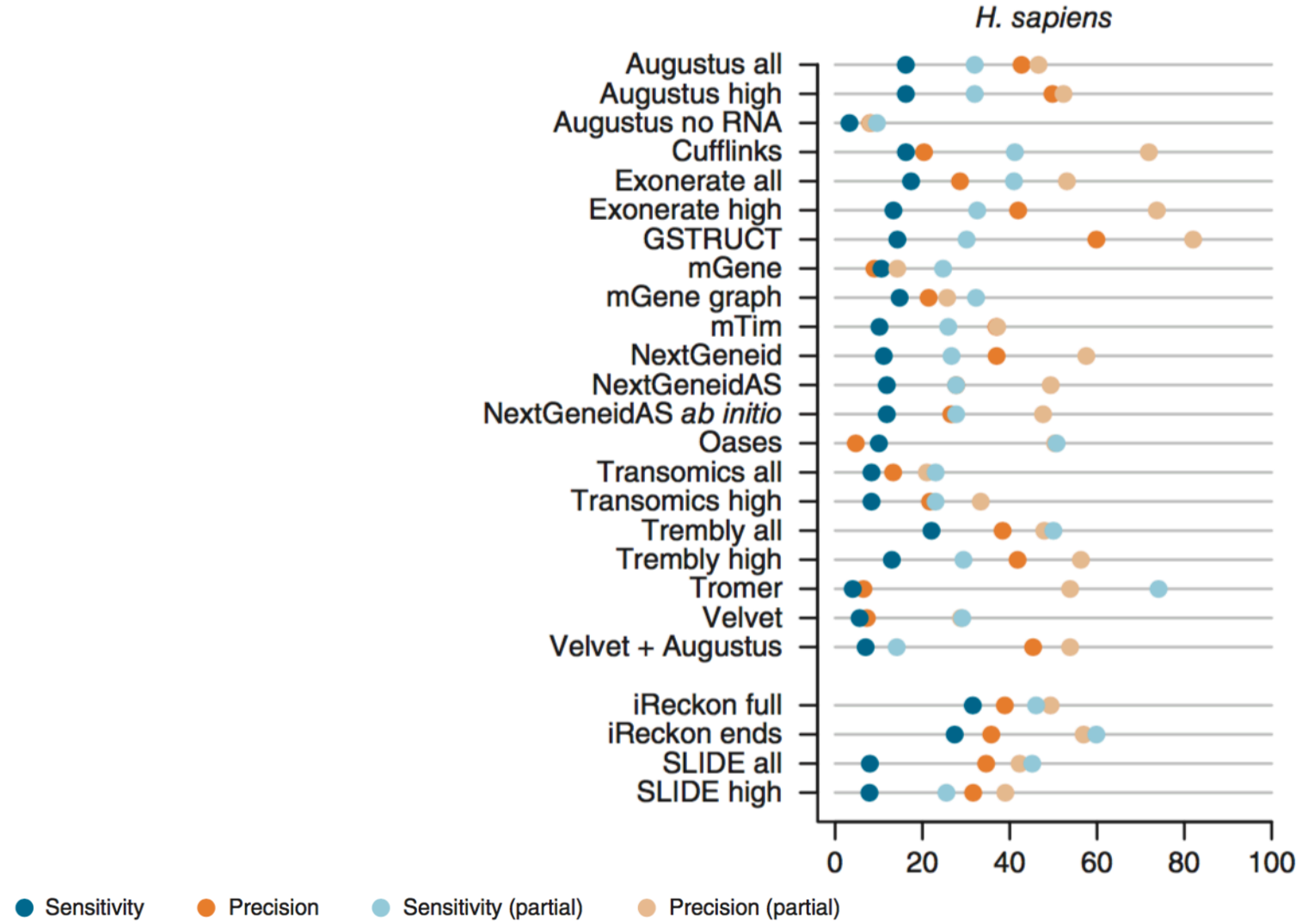

Best methods:

Sensitivity <50%

Specificity <70%

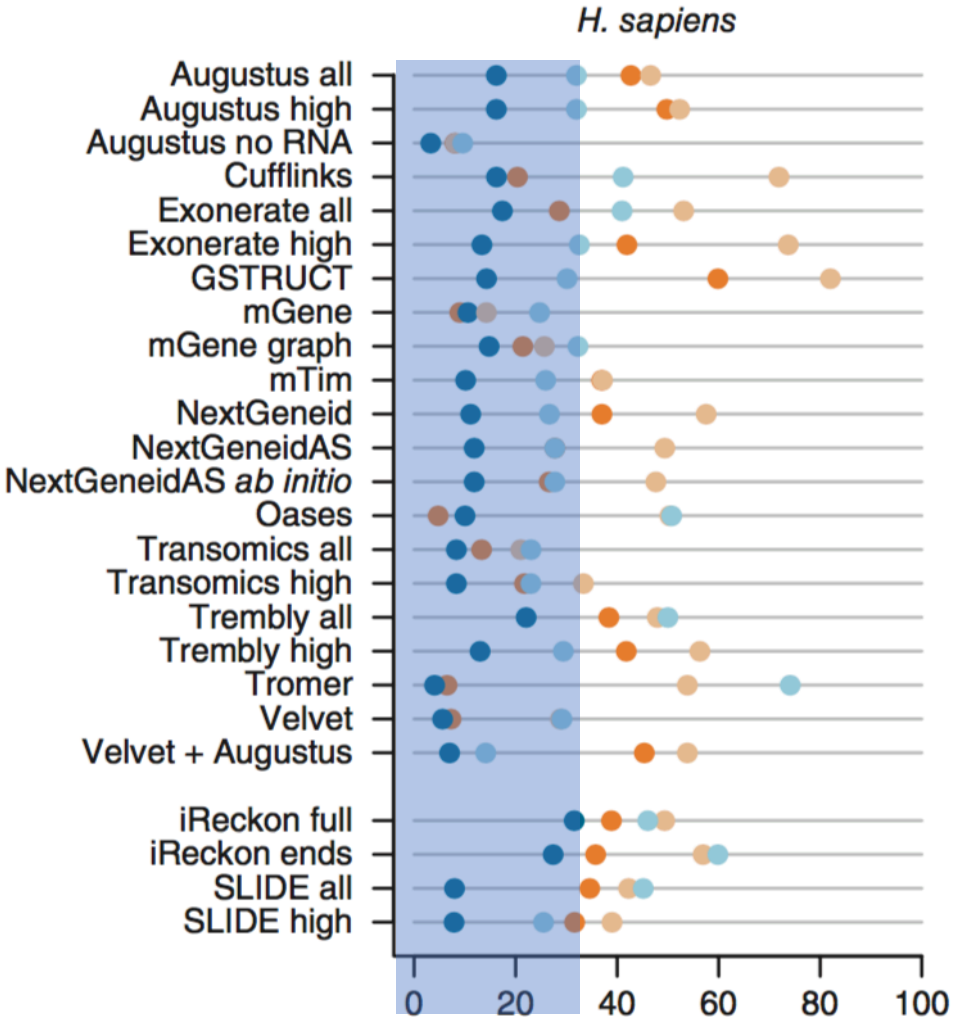
EGASP - Guigó et al. Genome Biol. 7 Suppl 1:S2.1-31 (2006)

# Requirement for manual annotation



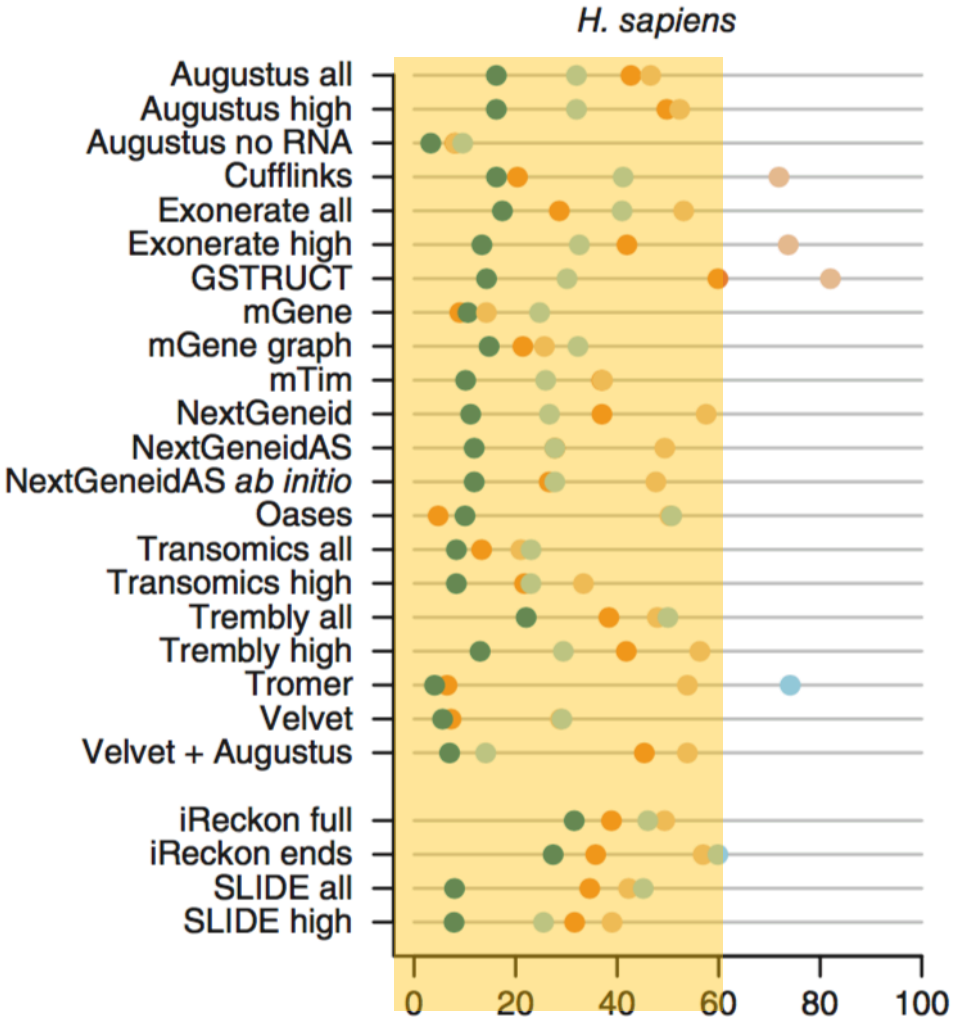

RGASP - Steijger et al. Nat Methods. 10(12):1177-84 (2013)

# Requirement for manual annotation




RGASP - Steijger et al. Nat Methods. 10(12):1177-84 (2013)

# Requirement for manual annotation




RGASP - Steijger et al. Nat Methods. 10(12):1177-84 (2013)

# Manual annotation supports automated annotation

**Gene: ENSBTAG00000000573**

| | |
|---|---|
| **Location** | Chromosome 19: 42,260,882-42,285,621 reverse strand.<br>UMD3.1:GK000019.2 |
| **About this gene** | This gene has 1 transcript (splice variant), 74 orthologues, 12 paralogues and is a member of 1 Ensembl protein family. |
| **Transcripts** | Hide transcript table |

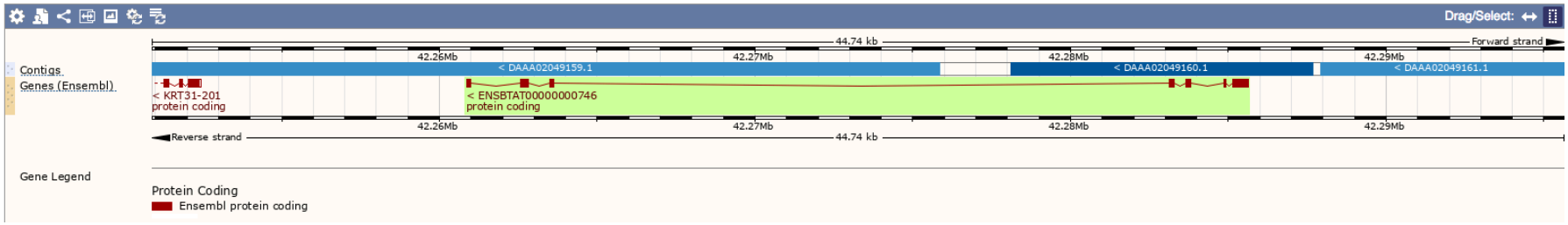
Show/hide columns  Filter

| Name | Transcript ID | bp | Protein | Translation ID | Biotype | UniProt | Flags |
|---|---|---|---|---|---|---|---|
| - | ENSBTAT00000000746.5 | 1359 | 452aa | ENSBTAP00000000746 | Protein coding | E1BFG1 | |

**Summary**

| | |
|---|---|
| **Ensembl version** | ENSBTAG00000000573.5 |
| **Gene type** | Protein coding |
| **Annotation method** | Annotation produced by the Ensembl genebuild. |

Go to Region in Detail for more tracks and navigation options (e.g. zooming)




# Manual annotation supports automated annotation

**Gene: ENSBTAG00000000573**

| | |
|---|---|
| **Location** | Chromosome 19: 42,260,882-42,285,621 reverse strand.<br>UMD3.1:GK000019.2 |
| **About this gene** | This gene has 1 transcript (splice variant), 74 orthologues, 12 paralogues and is a member of 1 Ensembl protein family. |
| **Transcripts** | Hide transcript table |

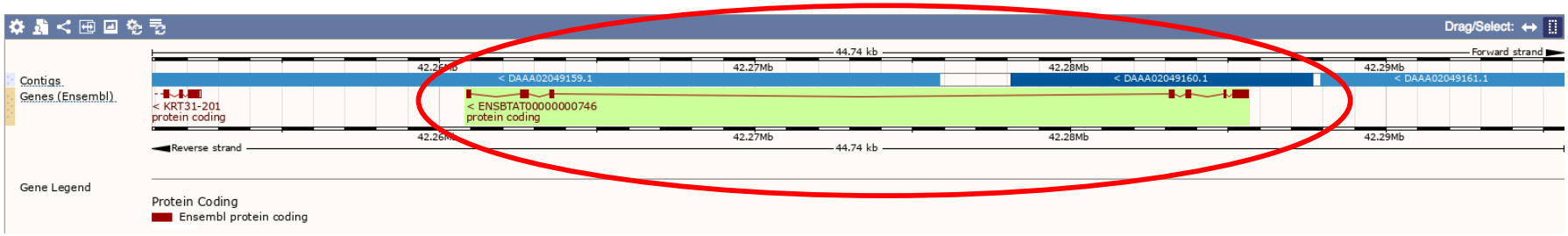
Show/hide columns | Filter

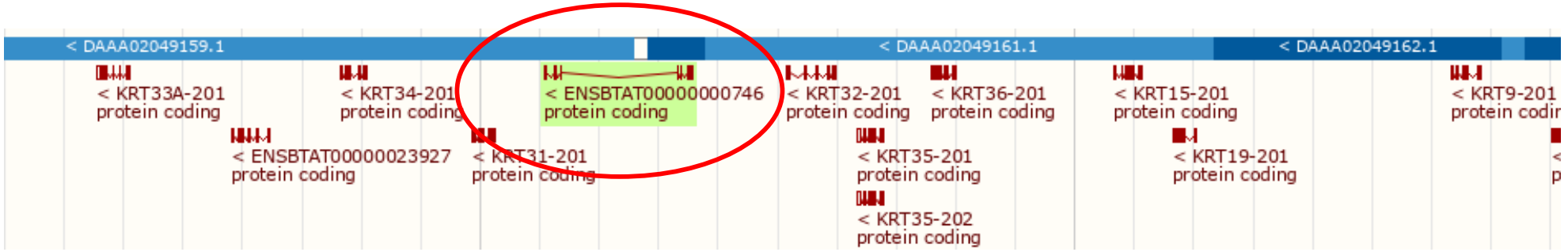
| Name | Transcript ID | bp | Protein | Translation ID | Biotype | UniProt | Flags |
|---|---|---|---|---|---|---|---|
| - | ENSBTAT00000000746.5 | 1359 | 452aa | ENSBTAP00000000746 | Protein coding | E1BFG1 | |

**Summary**

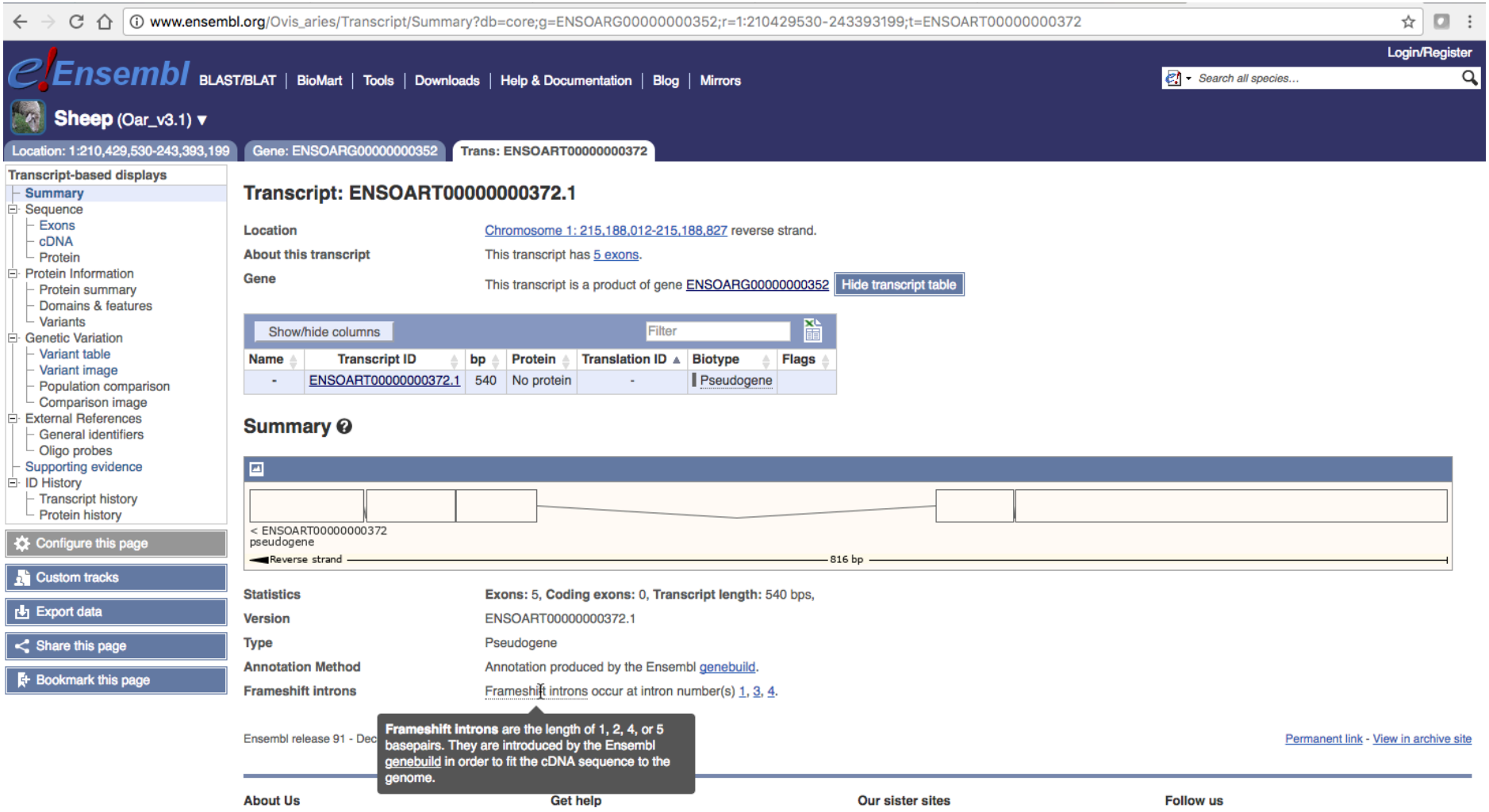
| | |
|---|---|
| **Ensembl version** | ENSBTAG00000000573.5 |
| **Gene type** | Protein coding |
| **Annotation method** | Annotation produced by the Ensembl genebuild. |

Go to Region in Detail for more tracks and navigation options (e.g. zooming)



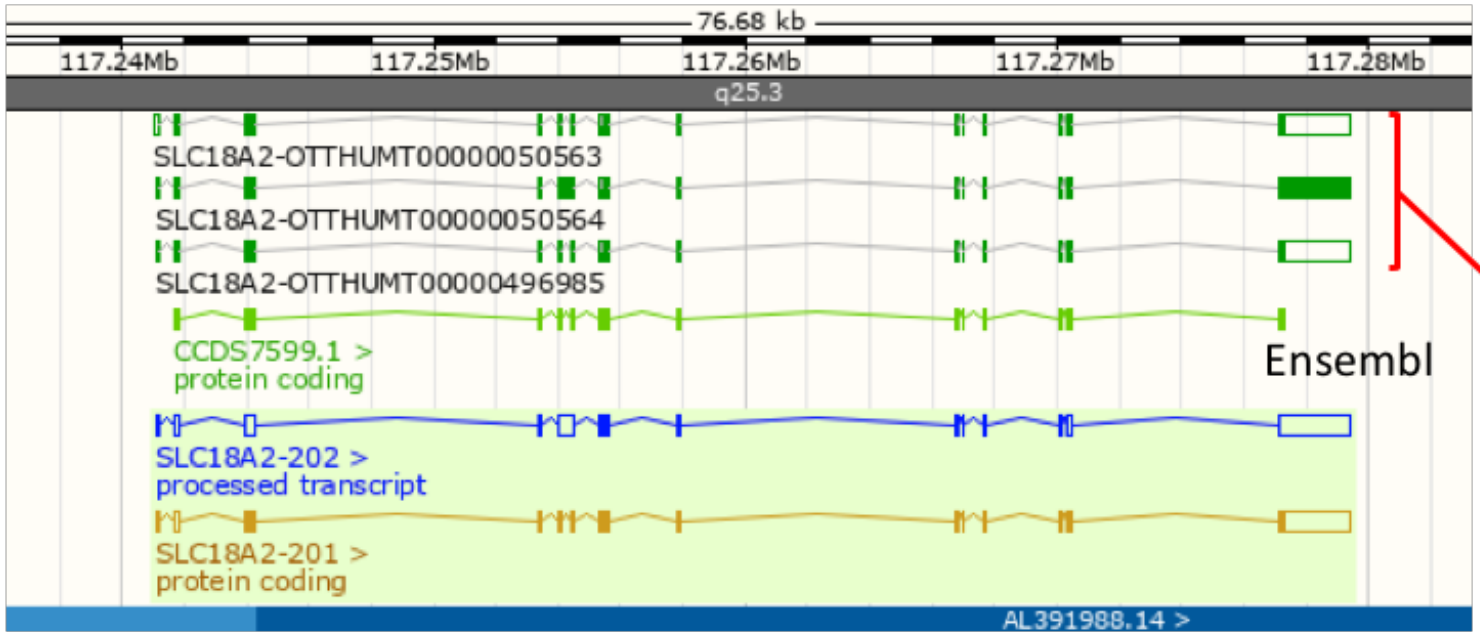
# Automated annotation problems:



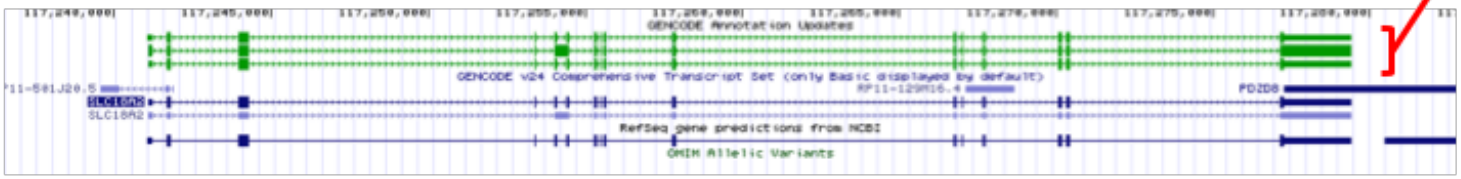

## Annotation issues:

Gaps, clusters and duplications, haplotype vs duplication, concatenation, overlapping genes, expansions and deletions, pseudogenes vs lof genes: biotype uncertainty, nomenclature, artefacts, transposons……

# The GENCODE update trackhub: Ensembl and UCSC genome browsers

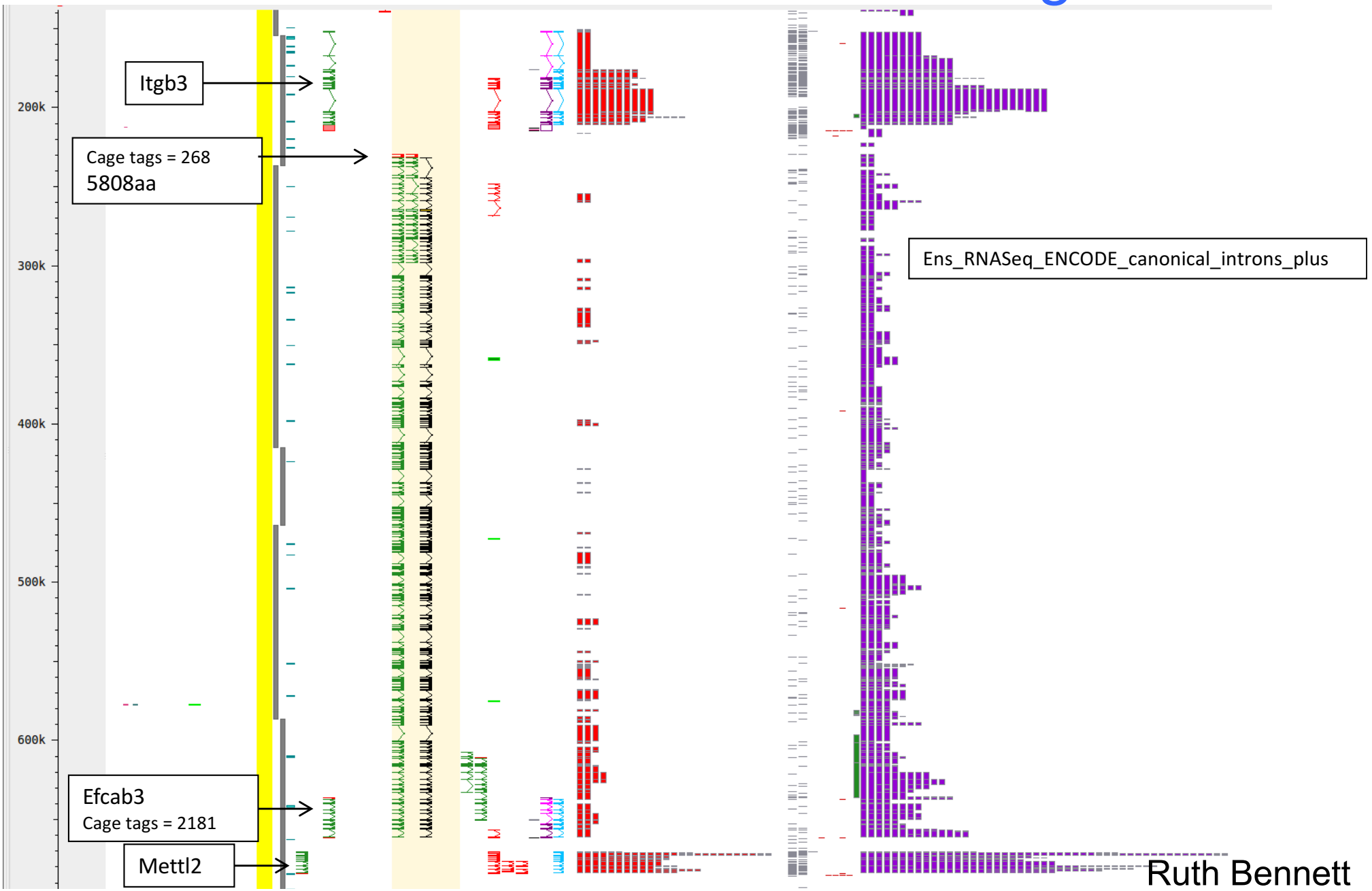



UCSC



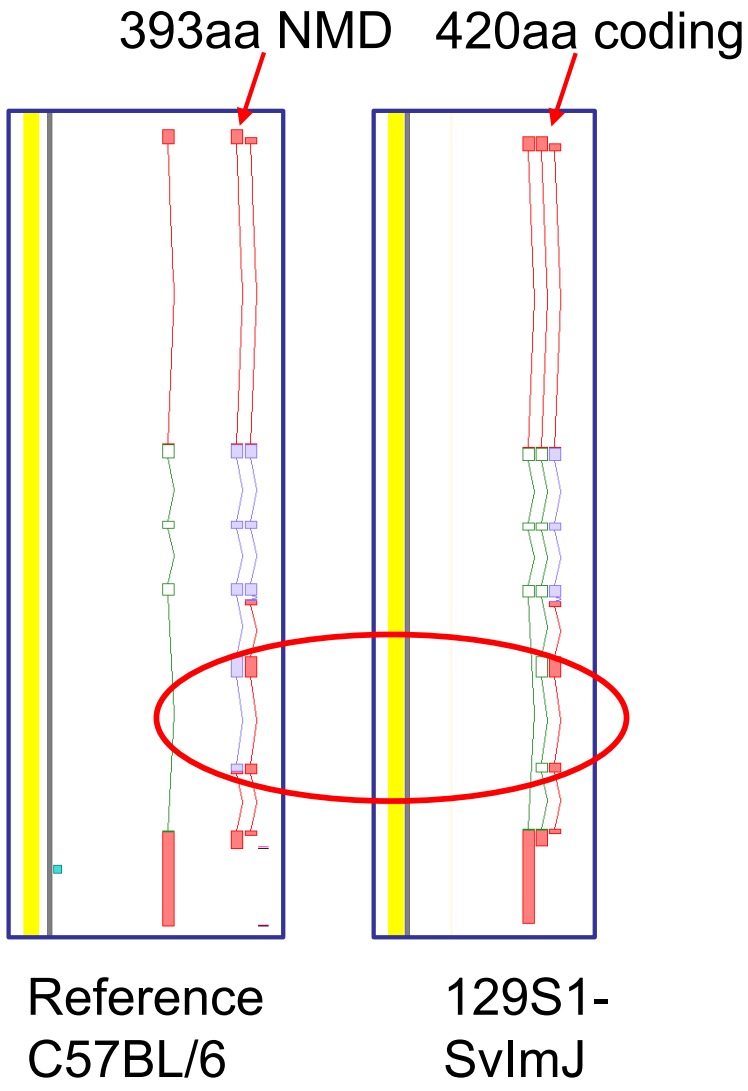
Update trackhub, in green, shows annotation updated in the last 24 hours. Annotation updated more than 24 hours ago will be shown in red.

# Mouse strain annotation reveals new genes:




Ruth Bennett

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




>Reference longest NMD transcript 393 AA.

MVNEYKRIVLLTGLMGINDHDFRMVKSLLSKELKLNKMQDEYDRVKI
ADLMEDKFPKDAGVVQLIKLYKQIPGLGDIANKLKNEKAKAKRKGKG
KRKTAAKRQRQEEPSTSQPMSTTNEDAEPESGRSTPDTQVAQLSLPT
ASRRNQAIQISPTIASSSGQTSSRSSETLQSIIQSPETPTRSSSRIL
DPPVSPGTAYSSAQALGVLLATPAKRQRLKNVPKEPSEENGYQQGSK
KVMVLKVTEPFAYDMKGEKMFHATVATETEFFRVKVFDIVLKEKFIP
NKVLTISNYVGCNGFINIYSASSVSEVNDGEPMNIPLSLRKSANRTP
KINYLCSKRRGIFVNGVFTVCKKEERGYYICYEIGDDTGMMEVEVYG
RLTNIACNPGDKLRLML*

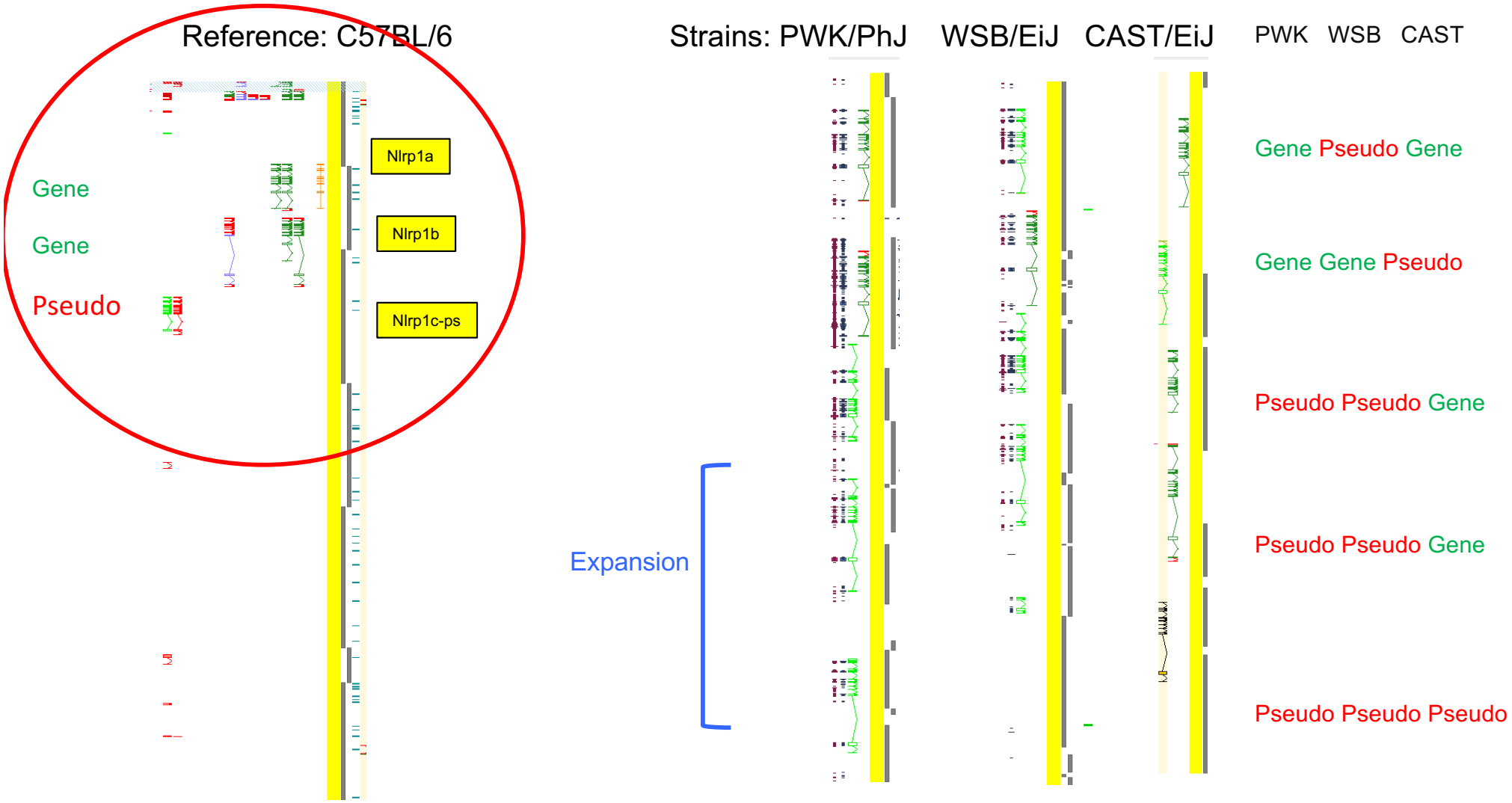
Stop codon isn't a SNP. Caused by a much larger disruption.

>129 Patch long coding transcript (equivalent to ref NMD) 420 AA.

MVNEYKRIVLLTGLMGINDHDFRMVKSLLSKELKLNKMQDEYDRVKI
ADLMEDKFPKDAGVVQLIKLYKQIPGLGDIANKLKNEKAKAKRKGKG
KRKTAAKRQRQEEPSTSQPMSTTNEDAEPESGRSTPDTQVAQLSLPT
ASRRNQAIQISPTIASSSGQTSSRSSETLQSIIQSPETPTRSSSRIL
DPPVSPGTAYSSAQALGVLLATPAKRQRLKNVPKEPSEENGYQLGSK
KVMVLKVTEPFAYDMKGEKMFHATVATETEFFRVKVFDIVLKEKFIP
NKVLTISNYVGCNGFINIYSASSVSEVNDGEPMNIPLSLRKSANRTP
KINYLCSKRRGIFVNGVFTVCKKEERGYYICYEIGDDTGMMEVEVYG
RLTNIACNPGDKLRLICFELTPDEETAWLRSTTHSNMQVIKARN*

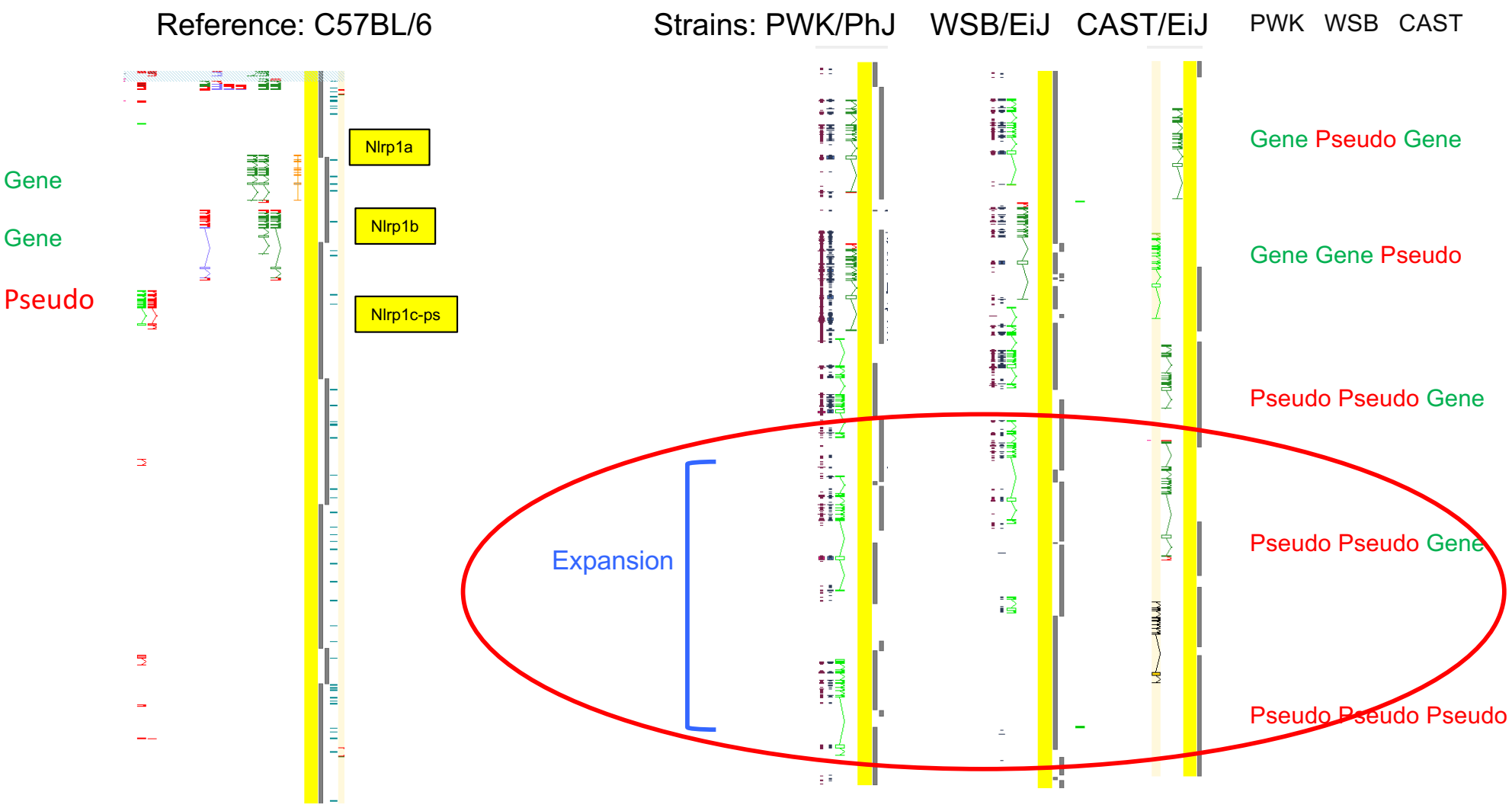
Ruth Bennett

# Mouse strain annotation reveals strain specific expansions at the Nlrp locus



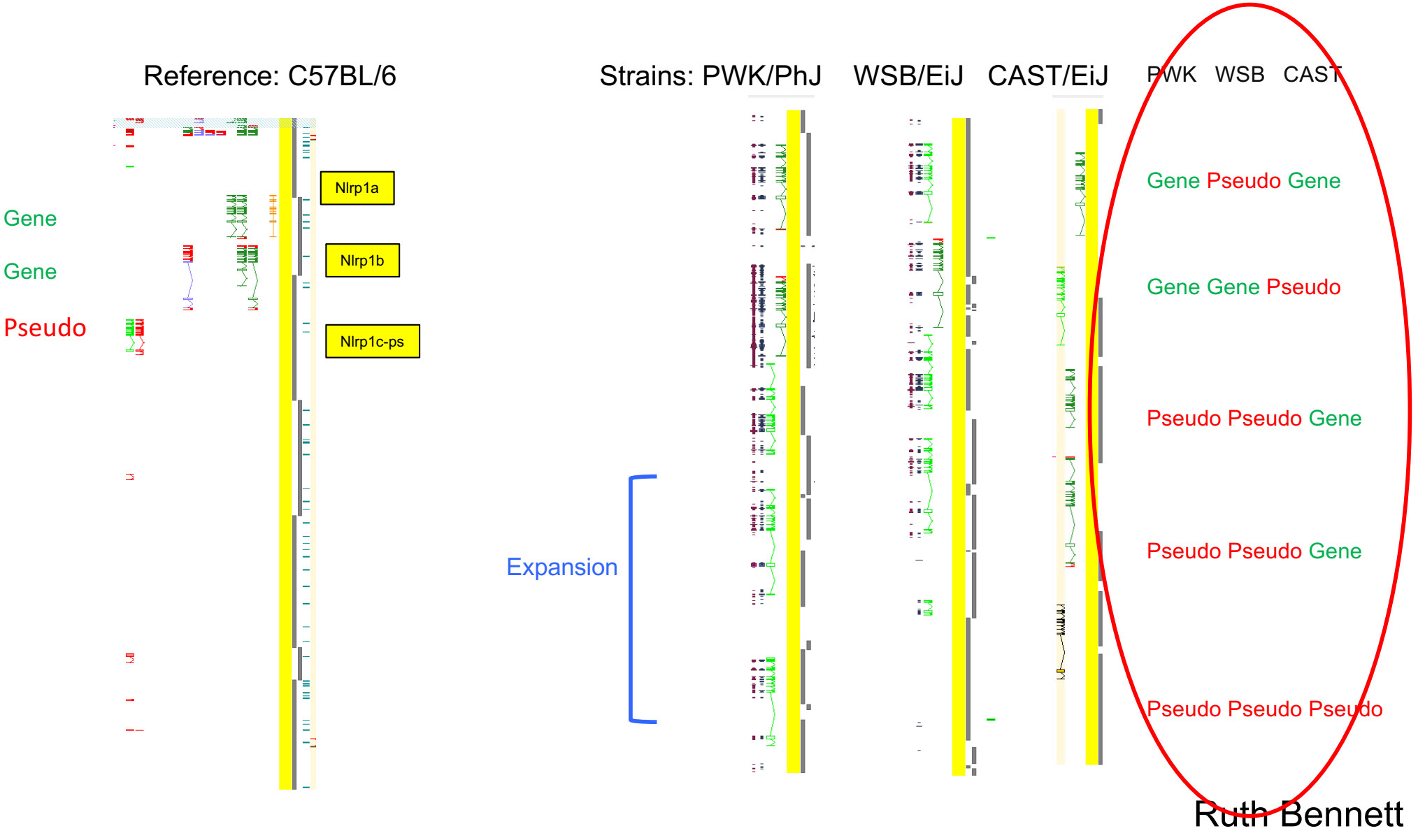

Ruth Bennett

# Mouse strain annotation reveals strain specific expansions at the Nlrp locus



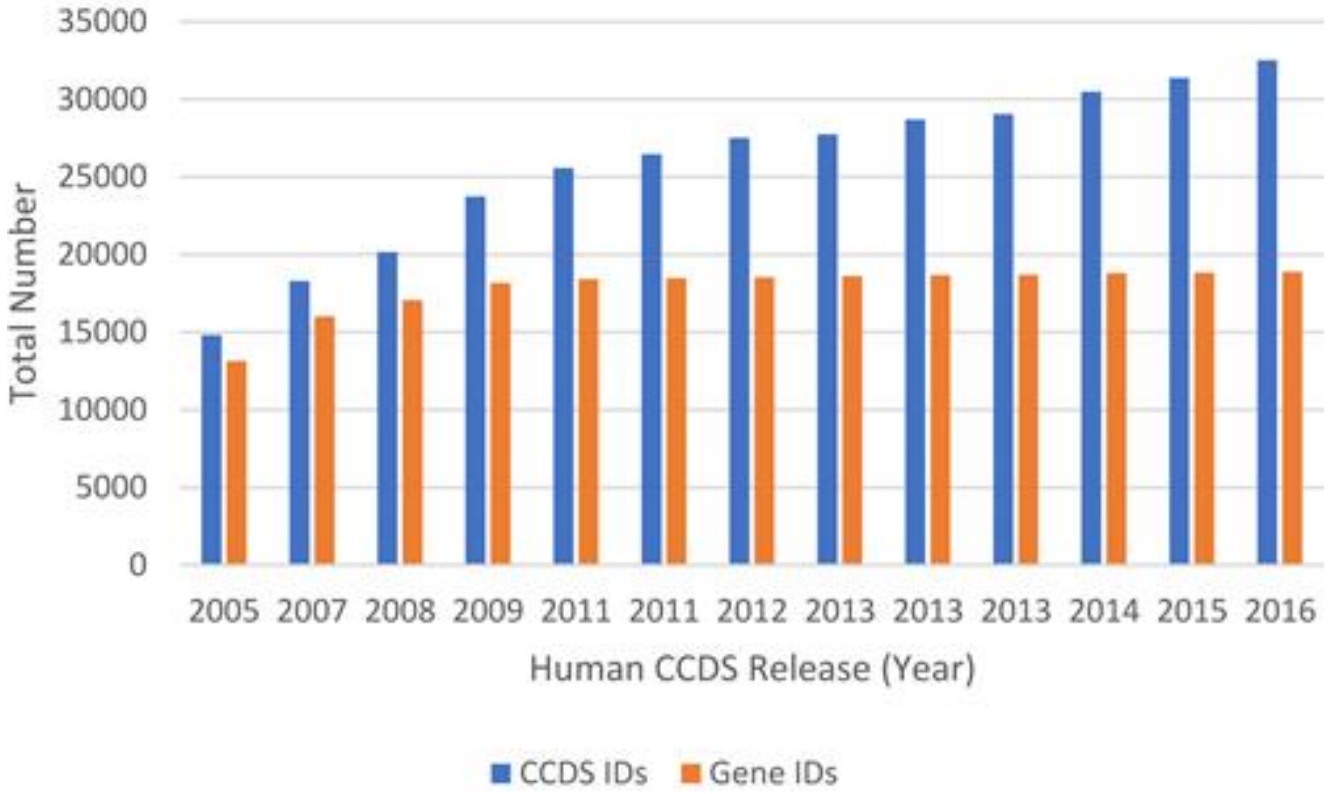

Ruth Bennett

# Mouse strain annotation reveals strain specific expansions at the Nlrp locus




Ruth Bennett

# Differences in manual annotation approach matter


| CCDS Totals | |
|---|---|
| **Category** | **Count** |
| CCDS IDs | 32,524 |
| Gene IDs | 18,894 |
| Sequence IDs | 79,349 |

| Status | |
|---|---|
| Public | 32,443 |
| Reviewed, update pending | 18 |
| Reviewed, withdrawal pending | 34 |
| Under review, update | 6 |
| Under review, withdrawal | 23 |

19,766 protein coding

From: Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation
Nucleic Acids Res. Published online November 06, 2017. doi:10.1093/nar/gkx1031

# Differences in manual annotation approach matter

McCarthy *et al. Genome Medicine* 2014, **6**:26
http://genomemedicine.com/content/6/3/26

Genome Medicine

RESEARCH Open Access

## Choice of transcripts and software has a large effect on variant annotation

Davis J McCarthy[1,2*], Peter Humburg[2], Alexander Kanapin[2], Manuel A Rivas[2], Kyle Gaulton[2], The WGS500 Consortium, Jean-Baptiste Cazier[3] and Peter Donnelly[1,2]

Frankish *et al. BMC Genomics* 2015, **16**(Suppl 8):S2
http://www.biomedcentral.com/1471-2164/16/S8/S2

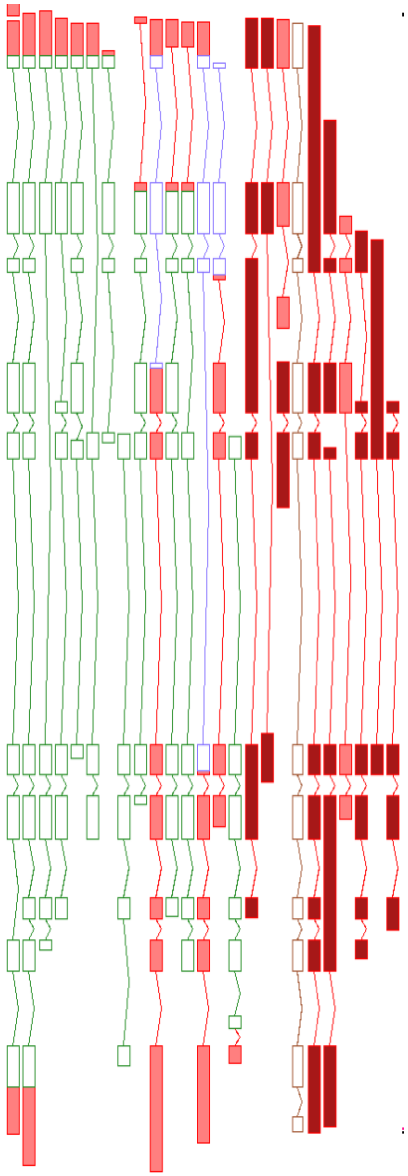
BMC Genomics

RESEARCH Open Access

## Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction

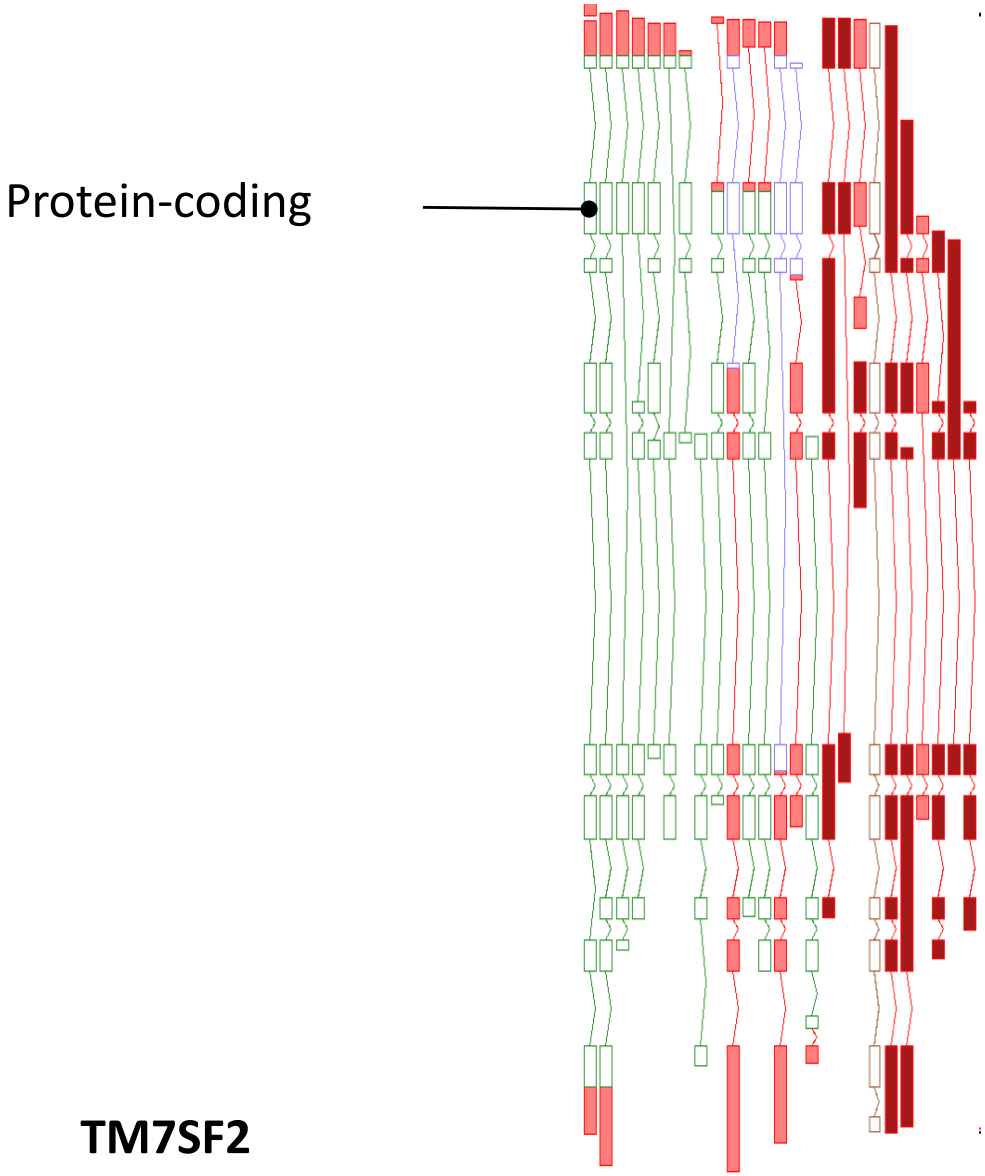
Adam Frankish[1*], Barbara Uszczynska[2], Graham RS Ritchie[1,3], Jose M Gonzalez[1], Dmitri Pervouchine[2,4], Robert Petryszak[3], Jonathan M Mudge[1], Nuno Fonseca[3], Alvis Brazma[3], Roderic Guigo[2], Jennifer Harrow[1*]

# HAVANA annotation guidelines: protein-coding genes

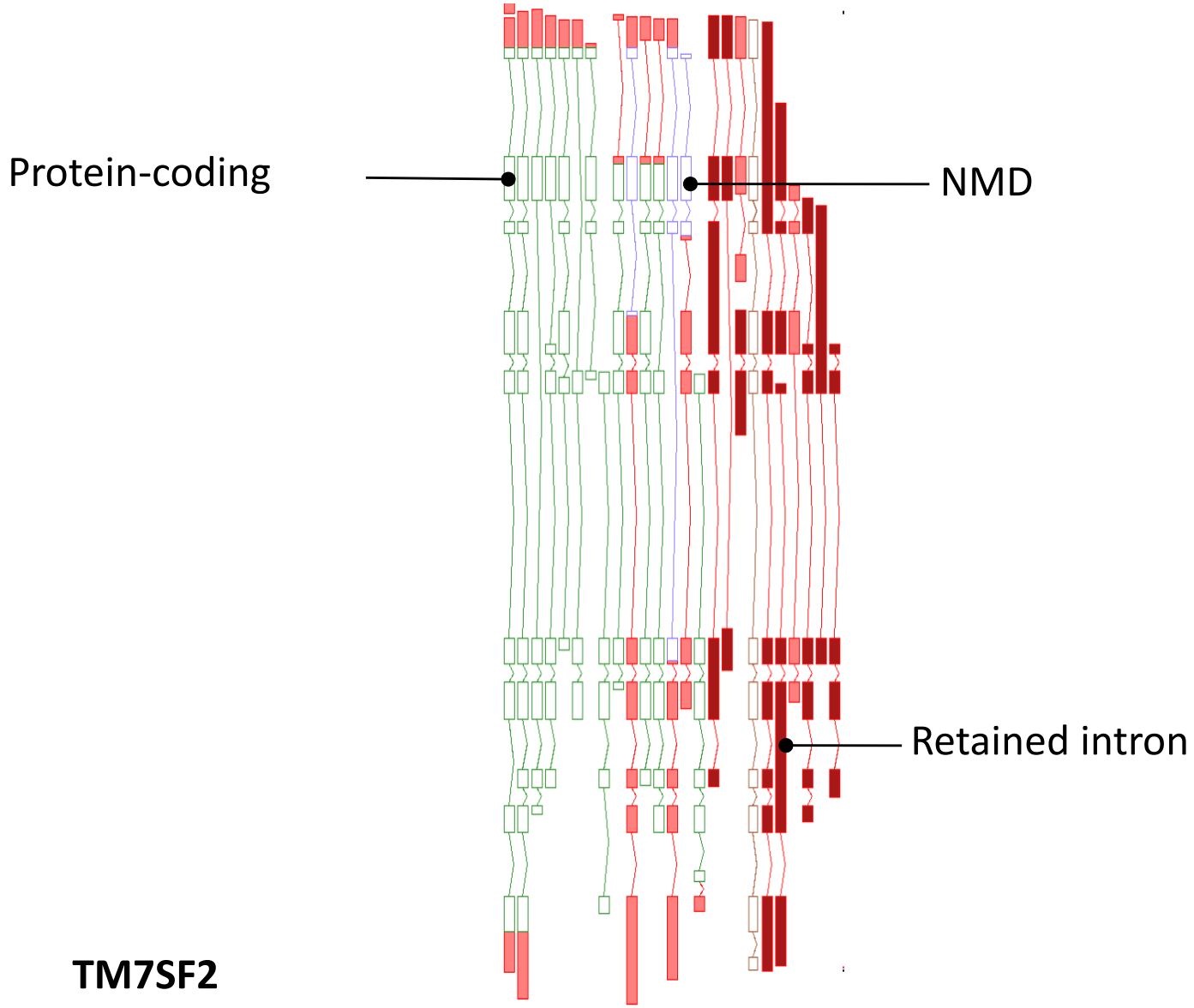


**TM7SF2**

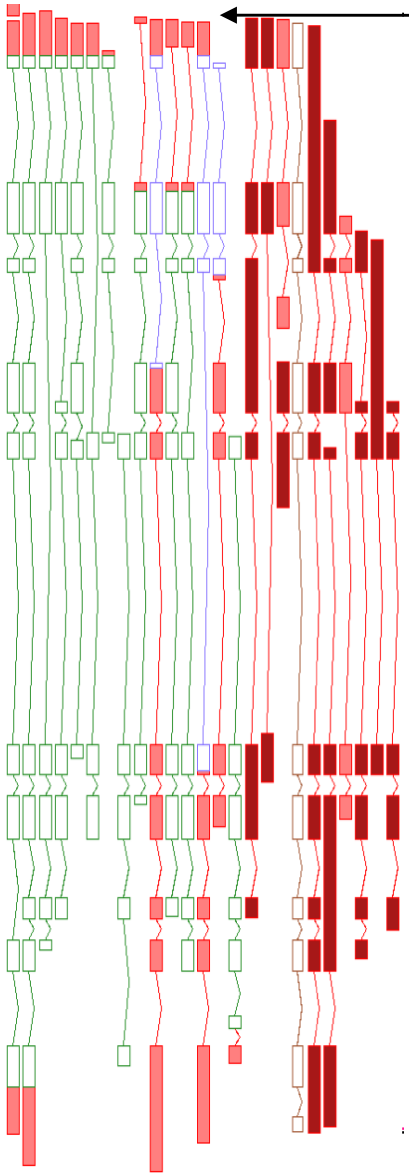
# HAVANA annotation guidelines: protein-coding genes



# HAVANA annotation guidelines: protein-coding genes



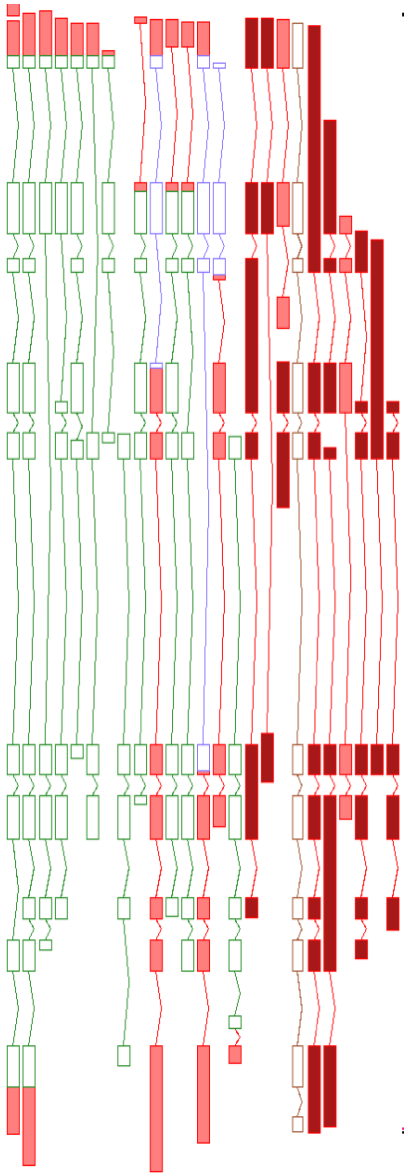
# HAVANA annotation guidelines: protein-coding genes




**TM7SF2**

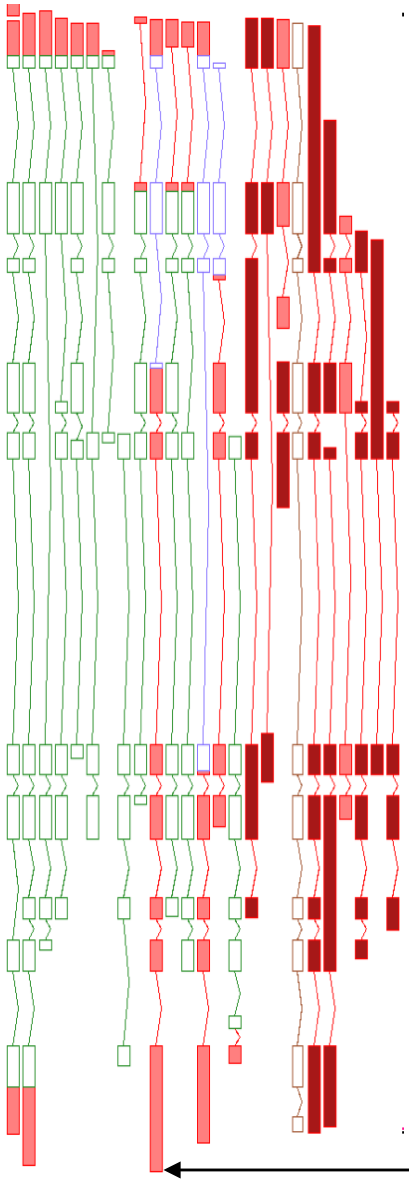
# HAVANA annotation guidelines: protein-coding genes

Translation initiation site



**TM7SF2**

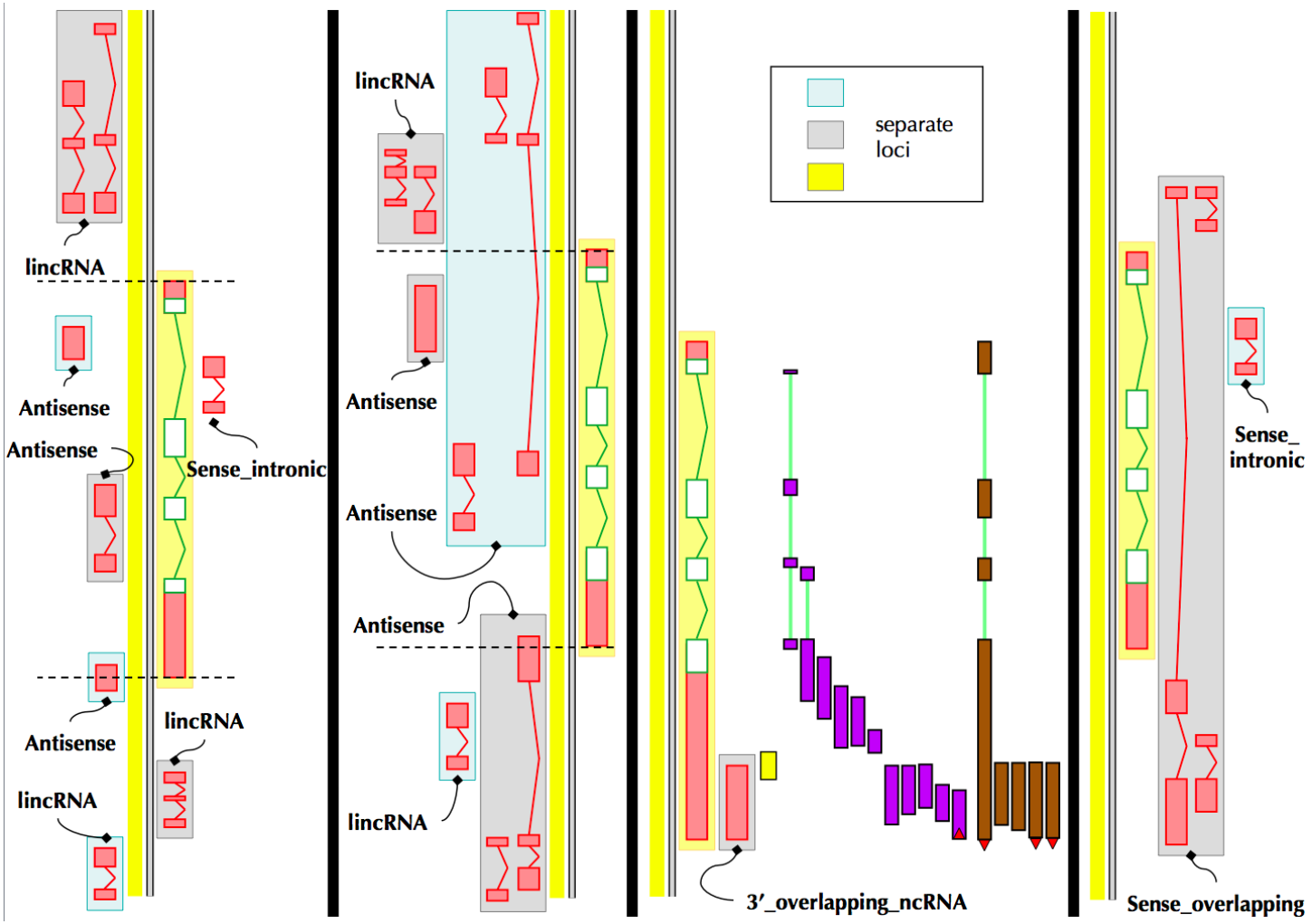
# HAVANA annotation guidelines: protein-coding genes



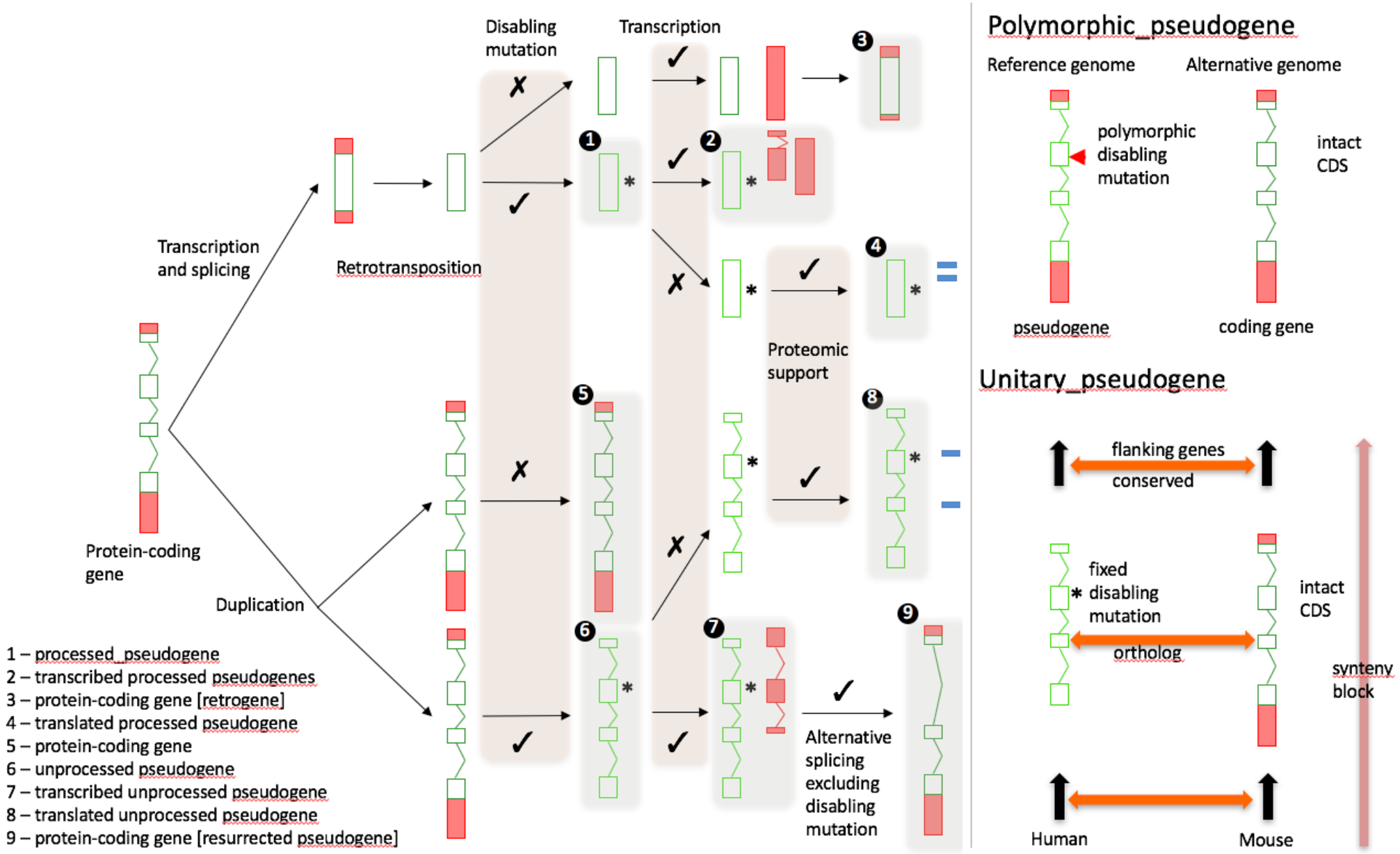
**TM7SF2**

Transcript end

# Improvements of lncRNA annotation: understanding functionality

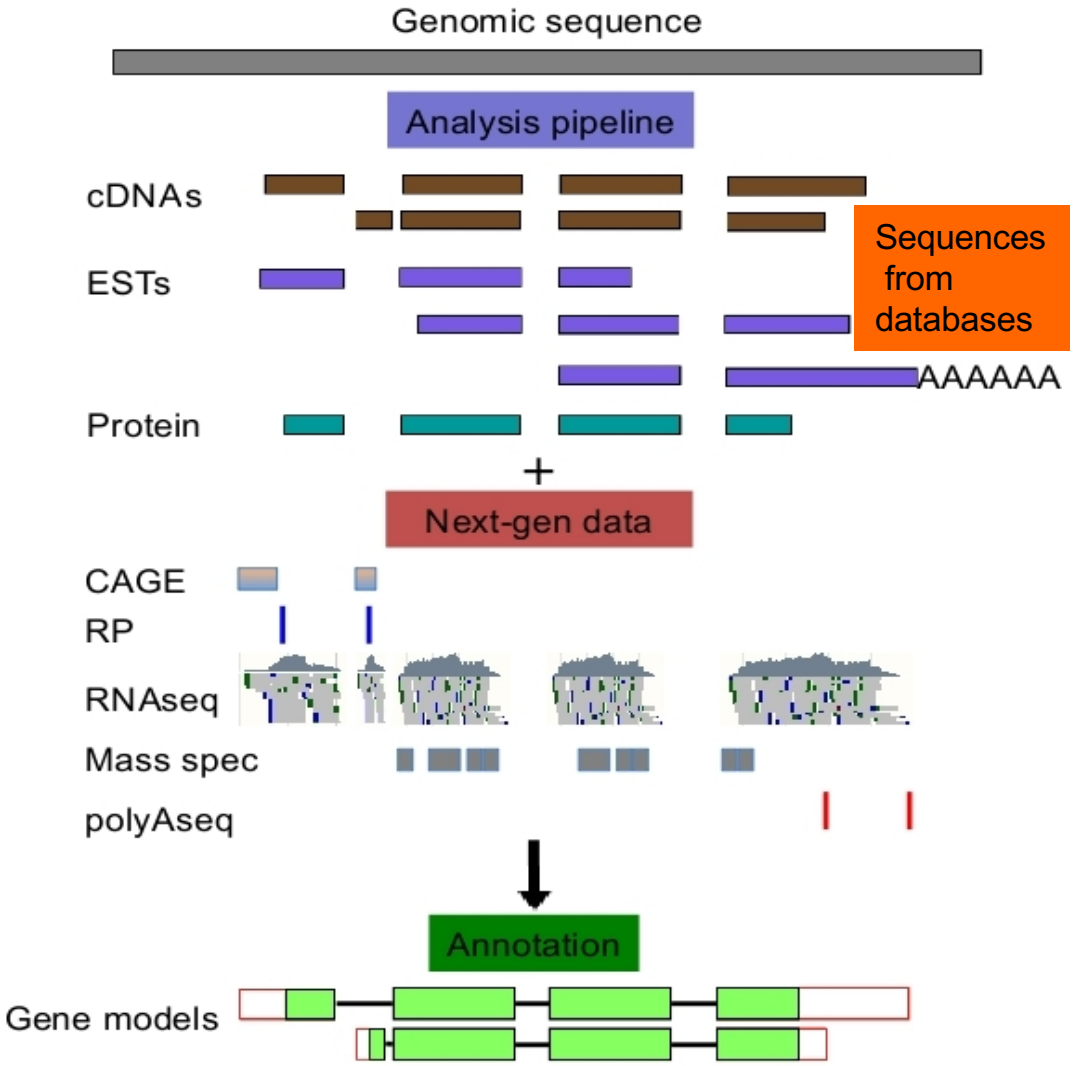


# HAVANA annotation guidelines: pseudogenes



# 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

# GENCODE Biotypes

## Version 27 (January 2017 freeze, GRCh38) - Ensembl 90, 91

### General stats

| | |
|---|---|
| Total No of Genes | 58288 |
| Protein-coding genes | 19836 |
| Long non-coding RNA genes | 15778 |
| Small non-coding RNA genes | 7569 |
| Pseudogenes | 14694 |
| - processed pseudogenes: | 10704 |
| - unprocessed pseudogenes: | 3469 |
| - unitary pseudogenes: | 206 |
| - polymorphic pseudogenes: | 63 |
| - pseudogenes: | 18 |
| Immunoglobulin/T-cell receptor gene segments | |
| - protein coding segments: | 410 |
| - pseudogenes: | 234 |
| Total No of Transcripts | 200401 |
| Protein-coding transcripts | 80930 |
| - full length protein-coding: | 55406 |
| - partial length protein-coding: | 25524 |
| Nonsense mediated decay transcripts | 14208 |
| Long non-coding RNA loci transcripts | 27908 |
| Total No of distinct translations | 60297 |
| Genes that have more than one distinct translations | 13580 |

| biotype | genes | transcripts |
|---|---|---|
| 3prime_overlapping_ncRNA | 31 | 35 |
| antisense_RNA | 5521 | 11050 |
| bidirectional_promoter_lncRNA | 19 | 40 |
| IG_C_gene | 14 | 23 |
| IG_C_pseudogene | 9 | 9 |
| IG_D_gene | 37 | 37 |
| IG_J_gene | 18 | 18 |
| IG_J_pseudogene | 3 | 3 |
| IG_pseudogene | 1 | 1 |
| IG_V_gene | 144 | 144 |
| IG_V_pseudogene | 188 | 188 |
| lincRNA | 7499 | 13348 |
| macro_lncRNA | 1 | 1 |
| miRNA | 1881 | 1881 |
| misc_RNA | 2213 | 2227 |
| Mt_rRNA | 2 | 2 |
| Mt_tRNA | 22 | 22 |
| non_coding | 3 | 3 |
| non_stop_decay | 0 | 84 |
| nonsense_mediated_decay | 0 | 14208 |
| polymorphic_pseudogene | 63 | 89 |
| processed_pseudogene | 10240 | 10243 |
| processed_transcript | 544 | 28230 |
| protein_coding | 19836 | 80930 |
| pseudogene | 18 | 37 |
| retained_intron | 0 | 27239 |
| ribozyme | 8 | 8 |
| rRNA | 544 | 544 |
| scaRNA | 49 | 49 |
| scRNA | 1 | 1 |
| sense_intronic | 905 | 963 |
| sense_overlapping | 189 | 339 |
| snoRNA | 943 | 955 |
| snRNA | 1900 | 1900 |
| sRNA | 5 | 5 |
| TEC | 1066 | 1165 |
| TR_C_gene | 6 | 6 |
| TR_D_gene | 4 | 4 |
| TR_J_gene | 79 | 79 |
| TR_J_pseudogene | 4 | 4 |
| TR_V_gene | 108 | 108 |
| TR_V_pseudogene | 30 | 30 |
| transcribed_processed_pseudogene | 462 | 462 |
| transcribed_unitary_pseudogene | 111 | 113 |
| transcribed_unprocessed_pseudogene | 830 | 836 |
| translated_processed_pseudogene | 2 | 2 |
| unitary_pseudogene | 95 | 95 |
| unprocessed_pseudogene | 2639 | 2640 |

# IWGSC RefSeq v1.0 Gene stats:

~4000 manually curated genes

107,886 high confidence genes

Duplicated genes (Inparalogs): 27% of the high confidence gene set
Lineage specific duplications

~162,000 low confidence genes: unsure if pseudogene or gene or genomic error

~300,000 pseudogenes

~4 million transposable elements (85% genome)

850 RNAseq sets available

# Please speak to us about manual annotation

Guidelines can be found at:

ftp://ftp.sanger.ac.uk/pub/project/havana/Guidelines/Guidelines_March_2016.pdf

Please contact us at:

gene-annotation@ebi.ac.uk

# GENCODE Acknowledgements


**Ensembl-HAVANA:**

Adam Frankish
If Barnes
Andrew Berry
Alex Bignell
Sarah Donaldson
Matt Hardy
Toby Hunt
Jane Loveland
Jonathan Mudge
Gaurab Mukherjee
Marie-Marthe Suner
Mark Thomas

**TGMI:**

Joannella Morales
Ruth Bennett
Claire Davidson
Mike Kay

**Annotrack/GENCODE:**

Jose Manuel Gonzalez

**Ensembl:**

Paul Flicek
Bronwen Aken
Fiona Cunningham
Thibaut Hourlier
Carlos García Girón
Fergal Martin

**GENCODE Consortium**

**Roderic Guigo, CRG**
Julien Legarde
Barbara Uszczynski
Rory Johnson
**Manolis Kellis, MIT**
Irwin Jungreis
**Michael Tress, CNIO**
**Alex Reymond, UNIL**
Anne-Maude Ferreira
**Mark Gerstein, Yale**
Cristina Sisu
Fabio Navara
**Benedict Paten, UCSC**
Mark Diekhans
**Tim Hubbard, KCL**


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

# Ensembl Acknowledgements

## The Entire Ensembl Team

**Daniel R. Zerbino[1], Premanand Achuthan[1], Wasiu Akanni[1], M. Ridwan Amode[1], Daniel Barrell[1,2], Jyothish Bhai[1], Konstantinos Billis[1], Carla Cummins[1], Astrid Gall[1], Carlos García Giroń[1], Laurent Gil[1], Leo Gordon[1], Leanne Haggerty[1], Erin Haskell[1], Thibaut Hourlier[1], Osagie G. Izuogu[1], Sophie H. Janacek[1], Thomas Juettemann[1], Jimmy Kiang To[1], Matthew R. Laird[1], Ilias Lavidas[1], Zhicheng Liu[1], Jane E. Loveland[1], Thomas Maurel[1], William McLaren[1], Benjamin Moore[1], Jonathan Mudge[1], Daniel N. Murphy[1], Victoria Newman[1], Michael Nuhn[1], Denye Ogeh[1], Chuang Kee Ong[1], Anne Parker[1], Mateus Patricio[1], Harpreet Singh Riat[1], Helen Schuilenburg[1], Dan Sheppard[1], Helen Sparrow[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Brandon Walts[1], Amonida Zadissa[1], Adam Frankish[1], Sarah E. Hunt[1], Myrto Kostadima[1], Nicholas Langridge[1], Fergal J. Martin[1], Matthieu Muffato[1], Emily Perry[1], Magali Ruffier[1], Dan M. Staines[1], Stephen J. Trevanion[1], Bronwen L. Aken[1], Fiona Cunningham[1], Andrew Yates[1] and Paul Flicek[1,3]**

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Eagle Genomics Ltd., Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK and [3]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

## Funding

Wellcome
EMBL
National Human Genome Research Institute
SAVE THE TASMANIAN DEVIL PROGRAM APPEAL
Open Targets
BBSRC bioscience for the future
TRANSFORMING GENETIC MEDICINE INITIATIVE
MedBioinformatics
EMBL-EBI
Co-funded by the European Union