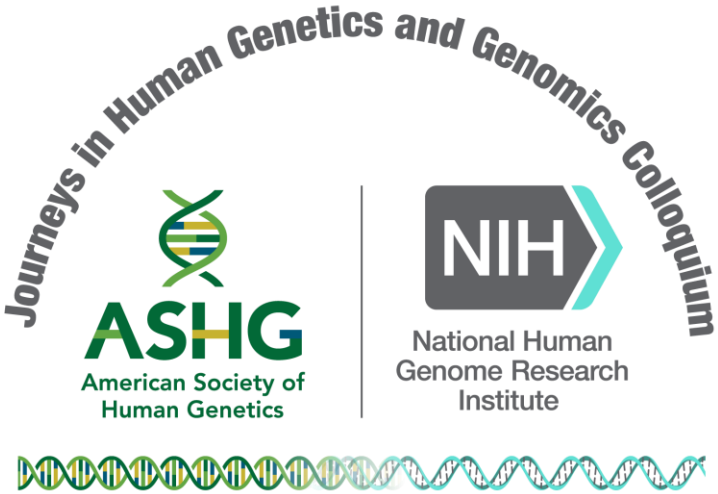
The Forefront of Genomics®

Genome Sequence Annotation

Jane Loveland, Ph.D.

October 2, 2024



Stratford-Upon-Avon

OXFORD BROOKES UNIVERSITY

STRATFORD GIRLS' GRAMMAR SCHOOL

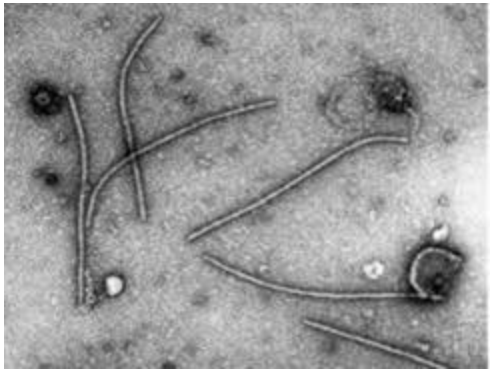
1989 First job in science: Beekeeping

National Bee Unit
Luddington Experimental Horticultural Station

A year off travelling

http://www.gcmap.com/

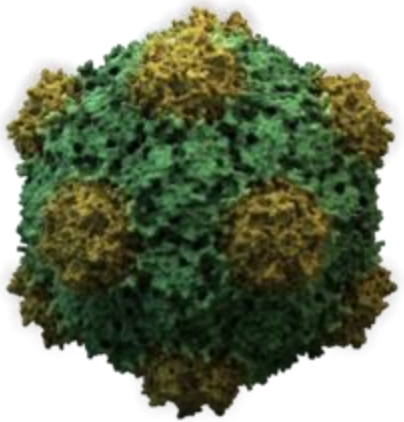
HRI Wellesbourne, Warwickshire, UK
(National Vegetable Research Station)



Zucchini yellow mosaic virus
Aphid transmitted potyvirus

1990-1991: Virus purification, electron microscopy

John Innes Centre

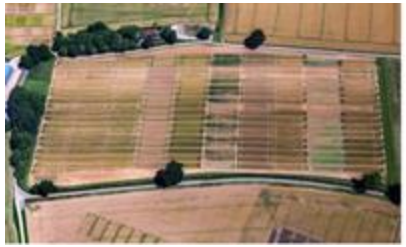


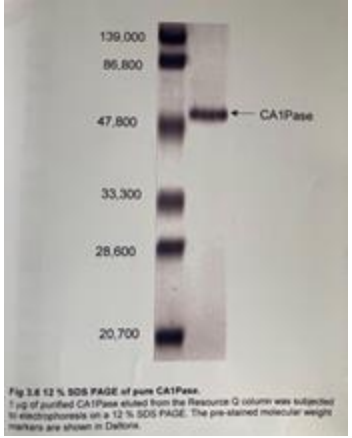
cowpea mosaic virus
(comovirus)

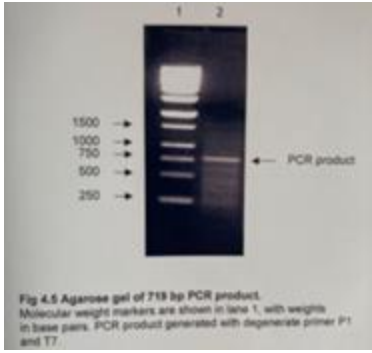
John Innes Centre, Norwich, UK
Professor George Lomonossoff, Protein expression in plants: HIV coat protein expression in cowpeas for vaccines. (PCR, DNA sequencing)

1991-1992

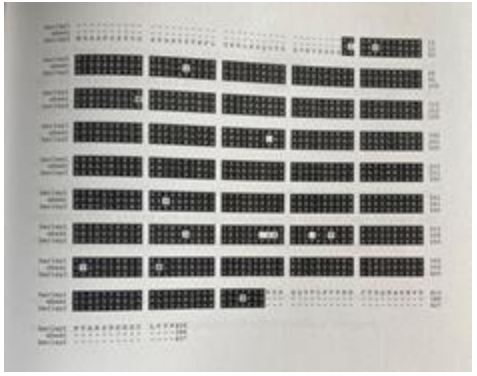
ROTHAMSTED RESEARCH







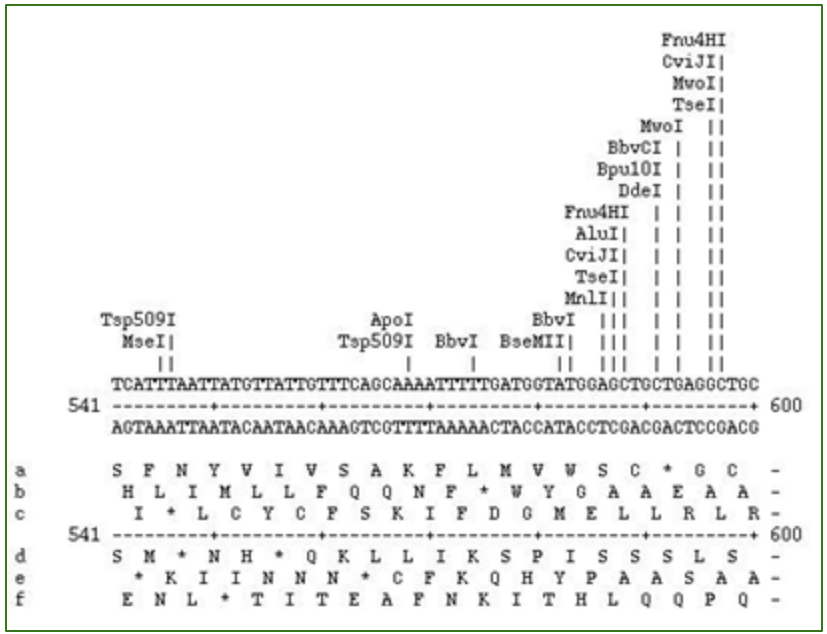
PhD: The purification and molecular analysis of CA1P-phosphatase in relation to the regulation of rubisco activity.



1992-2000: Biochemistry and Molecular Biology

BBSRC Bioscience IT Services: Bioinformatics training

Biotechnology and Biological Sciences Research Council



John Innes Centre
Unlocking Nature's Diversity

ROTHAMSTED RESEARCH

Babraham Institute

ROSLIN

THE Pirbright INSTITUTE

ibers
CNYDAU GWYDN RESILIENT CROPS

2000-2002: Unix, GCG

WELLCOME GENOME CAMPUS

wellcome sanger institute

Havana team

EMBL EBI

2002 - 2017

2017 onwards

Established 1992

Hinxton, UK

The Human Genome Project

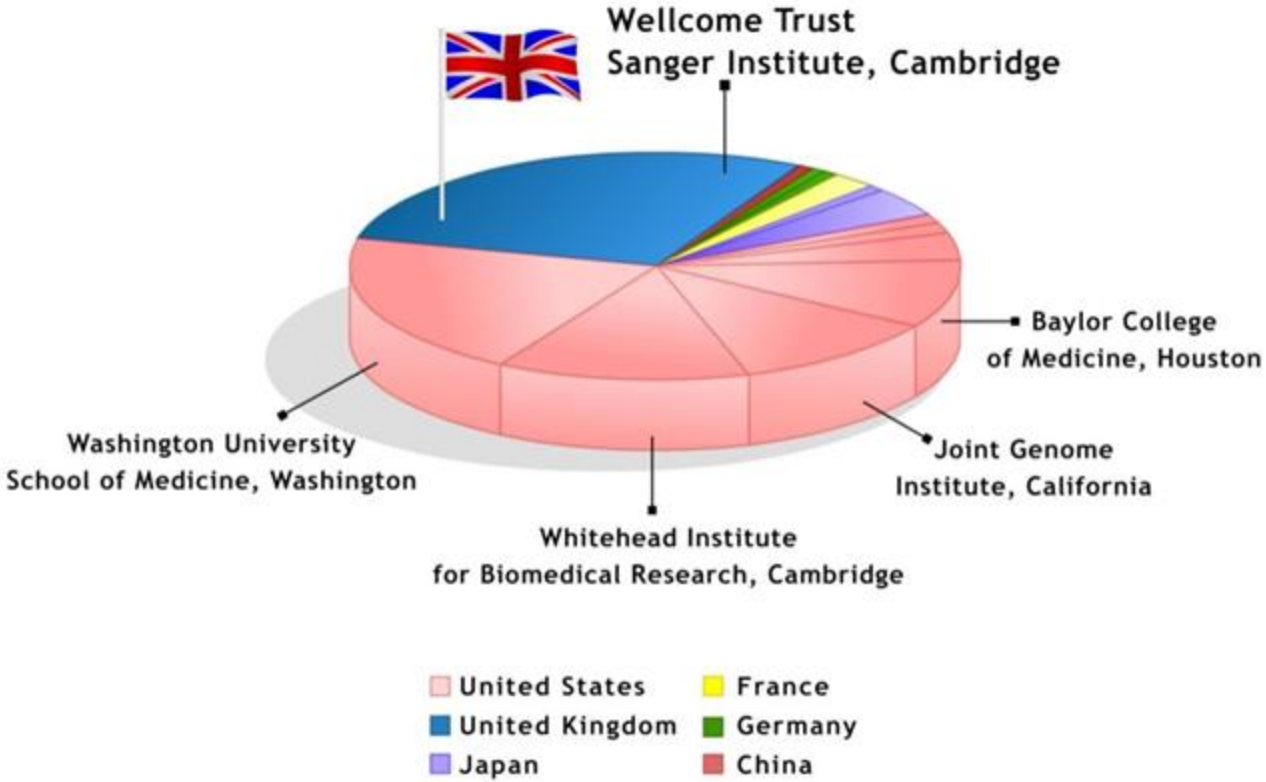
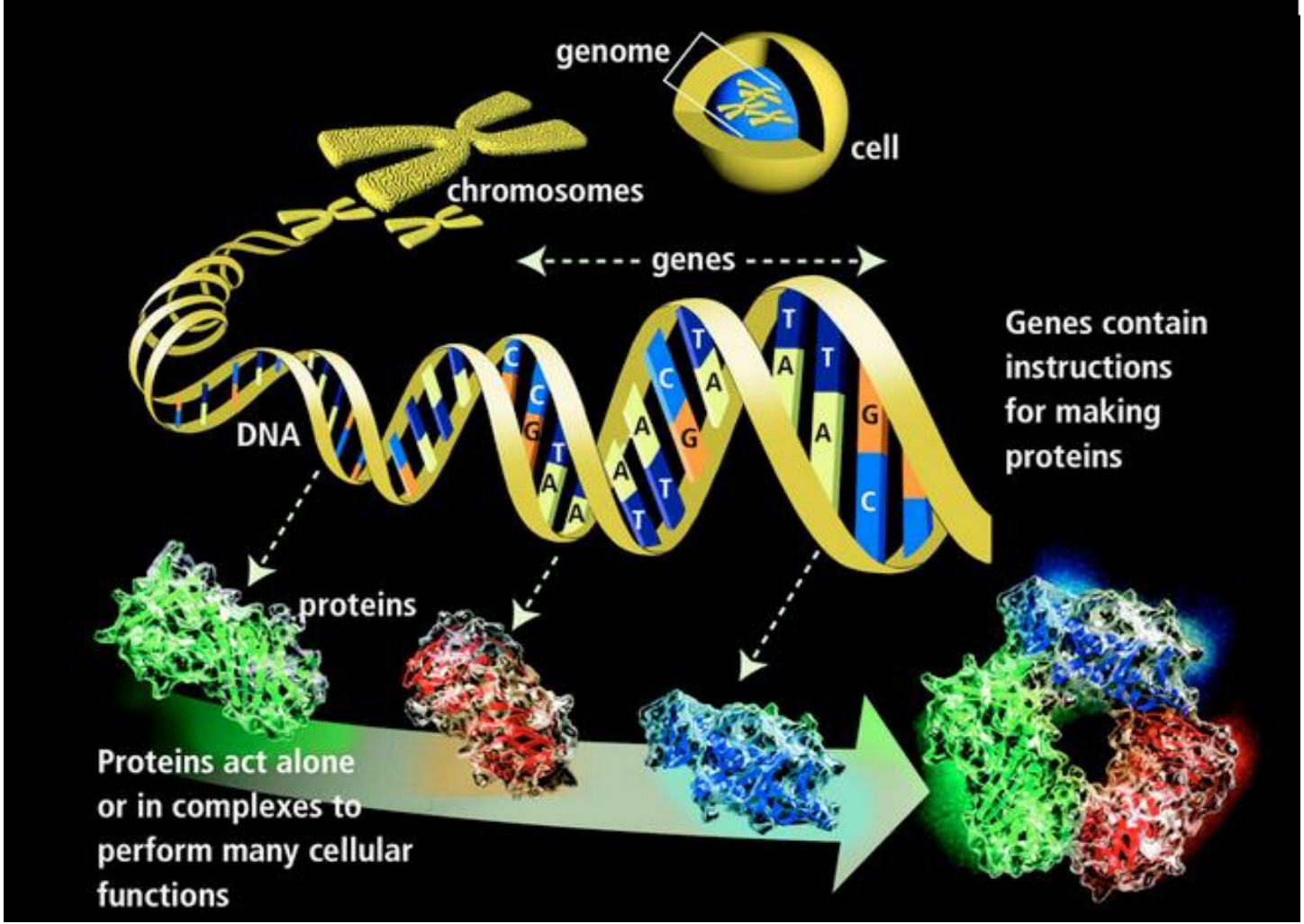


Image credit: Wellcome Trust Sanger Institute

EMBL-EBI

Open Door Workshops: Human genome freely available for everyone to use.

wellcome
connecting
science



http://chslearns.vail.k12.az.us/pluginfile.php/1738/course/section/616/central%20dogma.png

The Ensembl-HAVANA team

Human And Vertebrate ANalysis and Annotation

GENCODE

Whole genome or chromosome

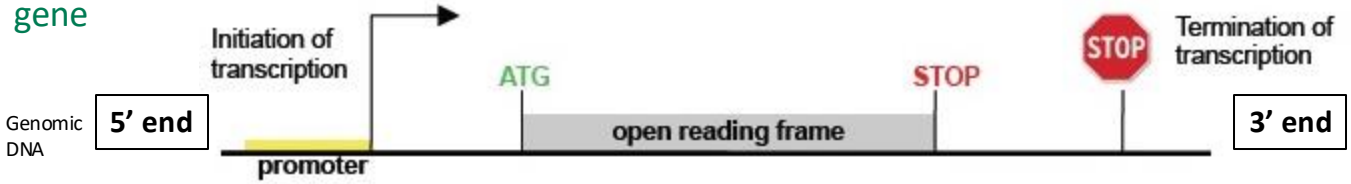
Targeted regions or genes

Community projects

EMBL-EBI

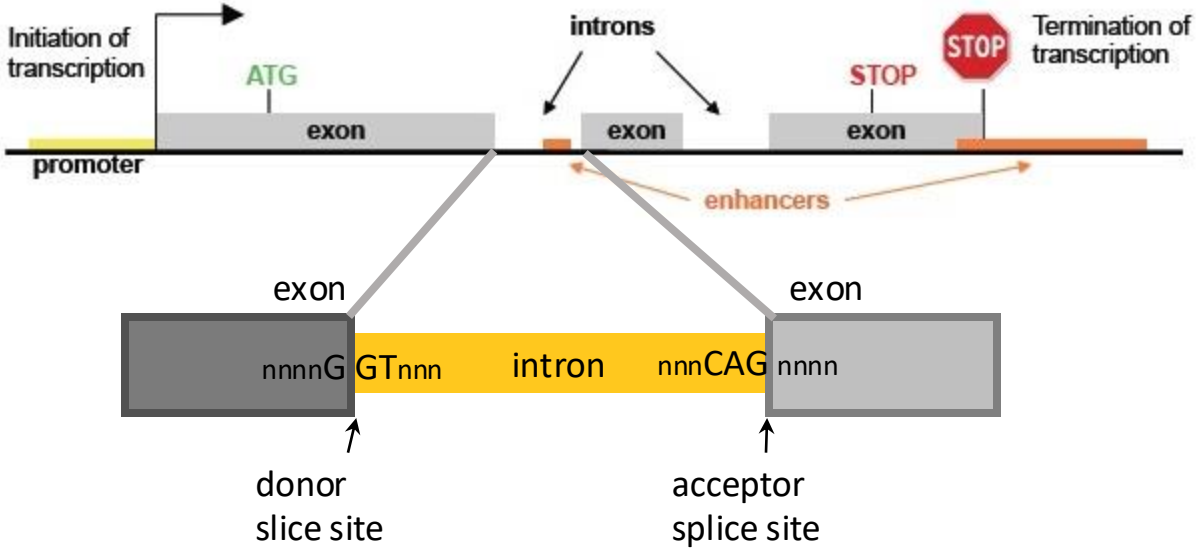
Prokaryotes:

Simple protein-coding gene



Eukaryotes:

More complex: Introns and Exons



EMBL-EBI

Gene Annotation

Definition: the plotting of genes and other genome features onto genome assemblies and indexing their genomic coordinates.

Automated Annotation

- Rapid annotation of genomes
- Broad range of approaches e.g. *Ab initio*, protein-genome alignments, projection
- Process vast amount of sequencing data easily

- Harder to project annotation over gene families/known difficult regions
- Genome assembly problems can be harder resolve
- Require set guidelines/rules for the definition of ‘features’ - biology cannot be defined easily which makes annotation difficult

Manual Annotation

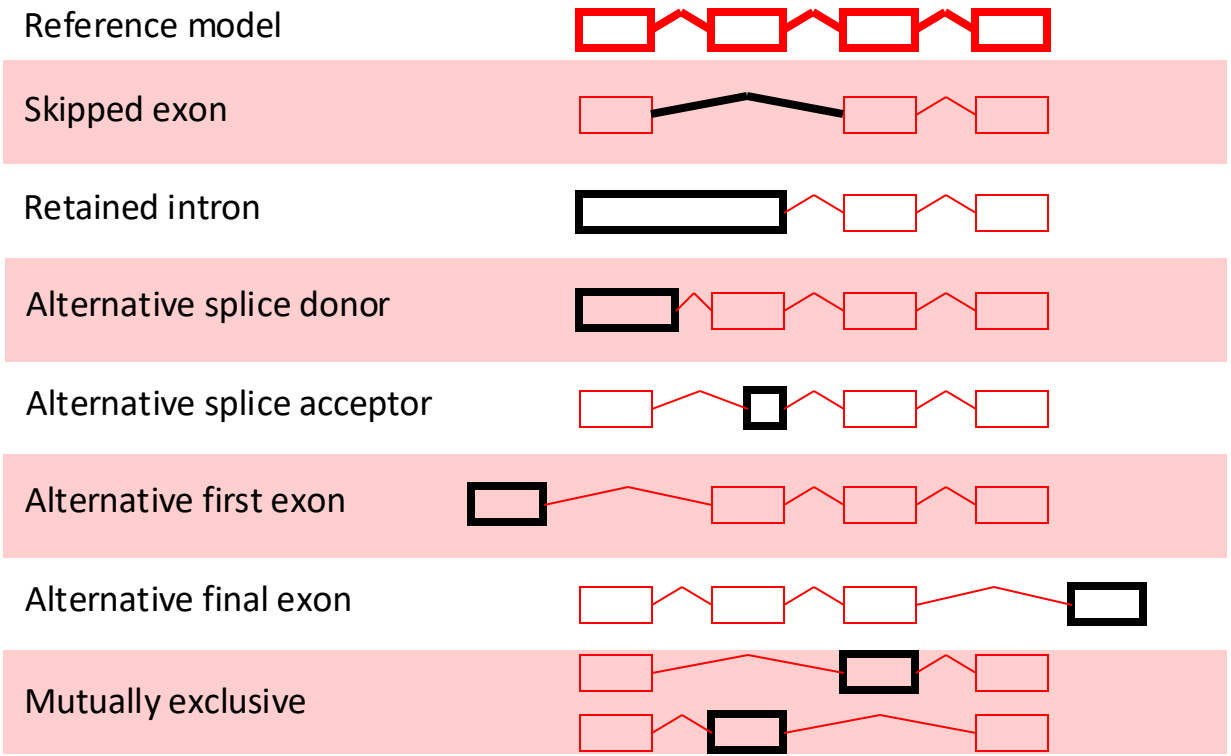
- Increased accuracy
- Easier to sort out difficult regions (e.g. MHC regions, gene families)
- Use variety of data sources and literature to define annotation
- Genes are so varied, flexibility on guidelines

- Time-consuming
- Expensive
- More difficult to work with vast amount of sequencing data

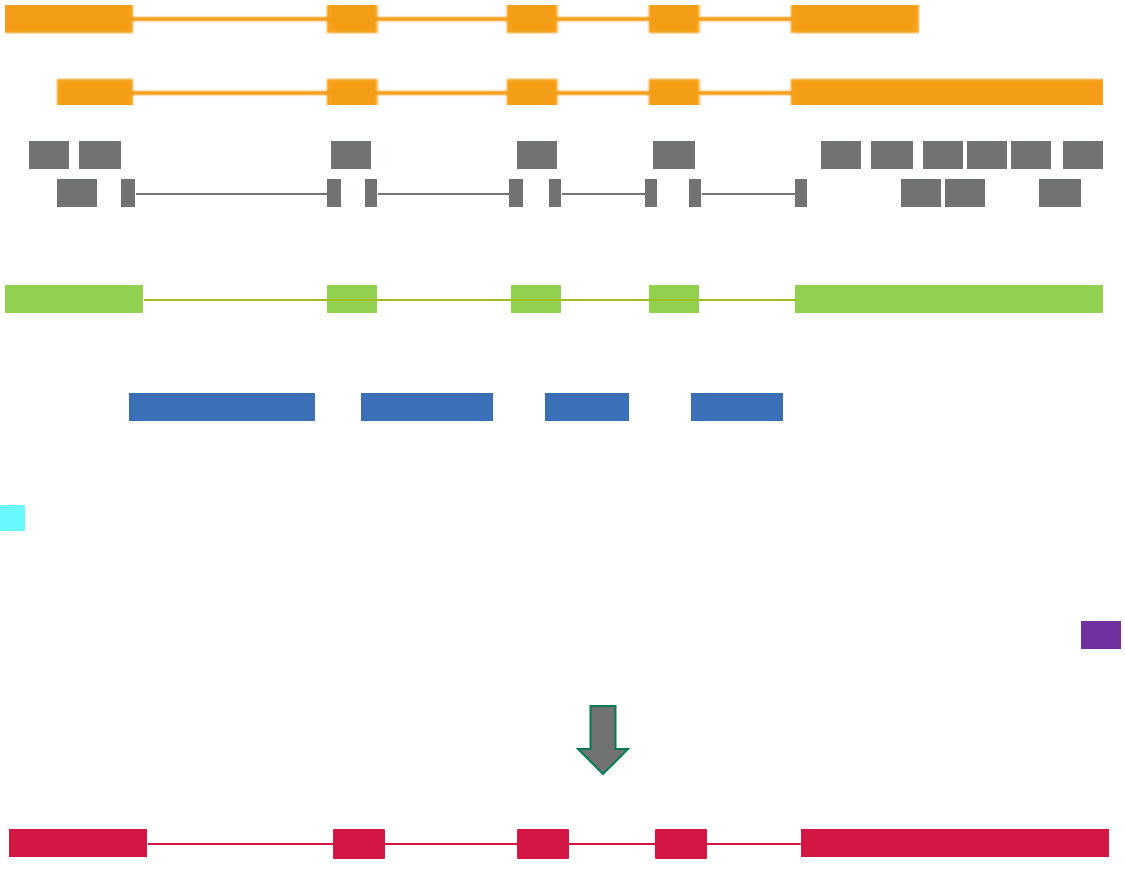
Both Manual and Automated Annotation are required for high quality reference genome annotation

EMBL-EBI

Alternative Splicing



HAVANA gene annotation basics



Long transcriptomic data

RNAseq data

Transcript models (tmerge)

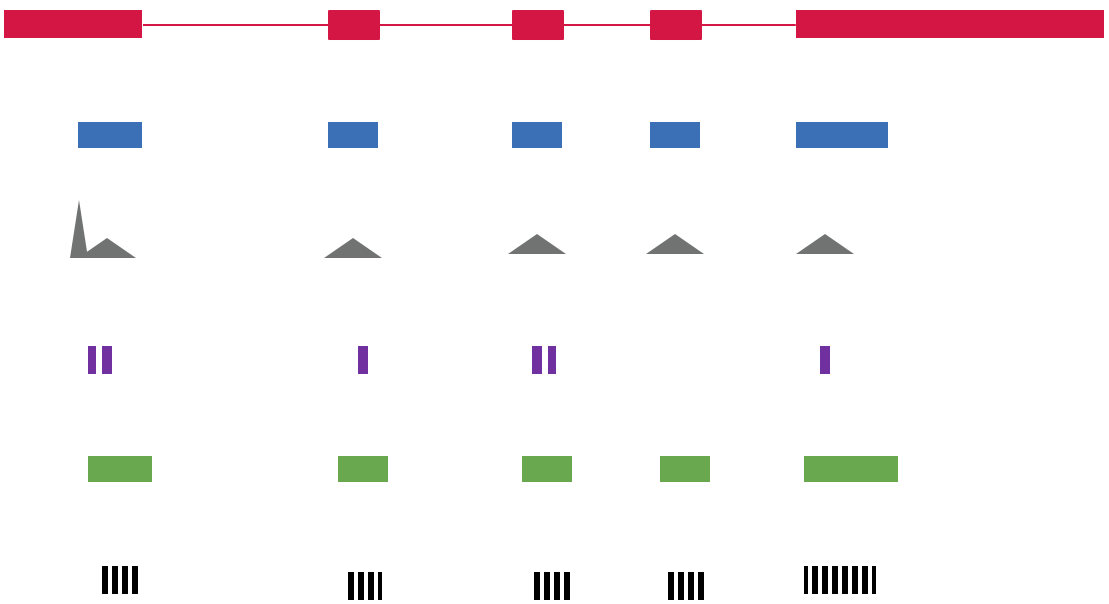
RNAseq introns (Recount3 data)

CAGE/RAMPAGE

polyAseq (signal and site)

Transcript annotation

HAVANA gene annotation basics



Transcript annotation

Protein homology

Ribo-seq data

Mass Spec data

Conservation/constraint
(PhyloCSF)

Human variation data

Literature, comparative annotation, functional annotation, external expert databases

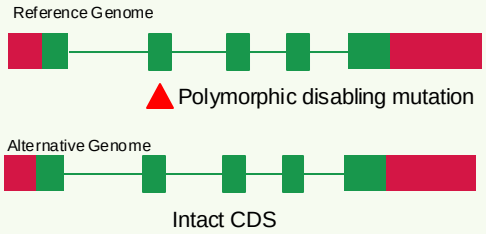
Biotypes

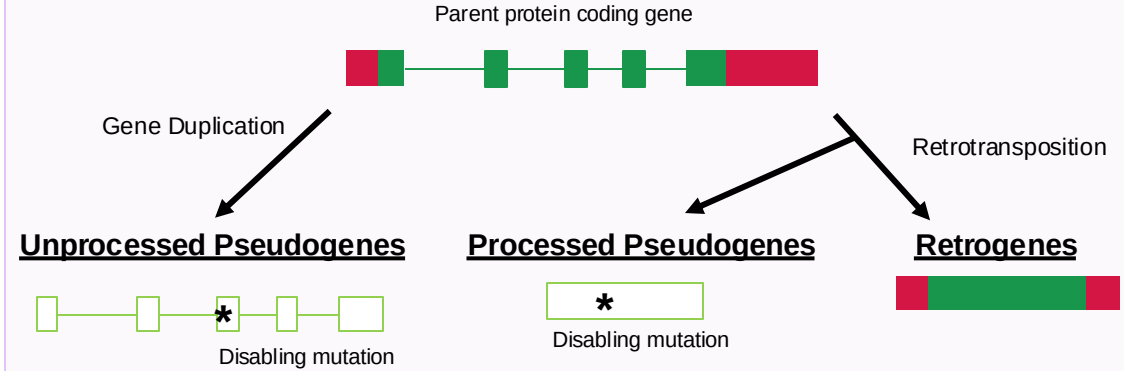
Protein Coding

NMD

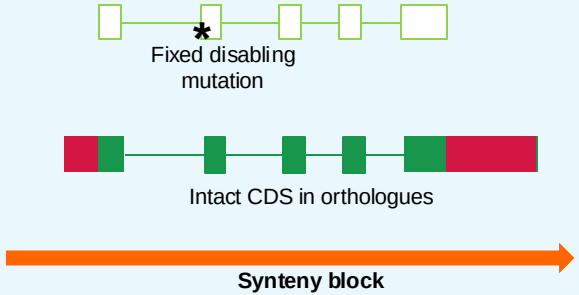
Retained Intron

Protein Loss of Function Genes





Unitary Pseudogenes



Small Non-Coding RNAs

miRNA
snoRNA
scaRNA
snRNA
rRNA
Vault RNA
tRNA

NOTE: Currently in the GENCODE geneset via automated pipelines

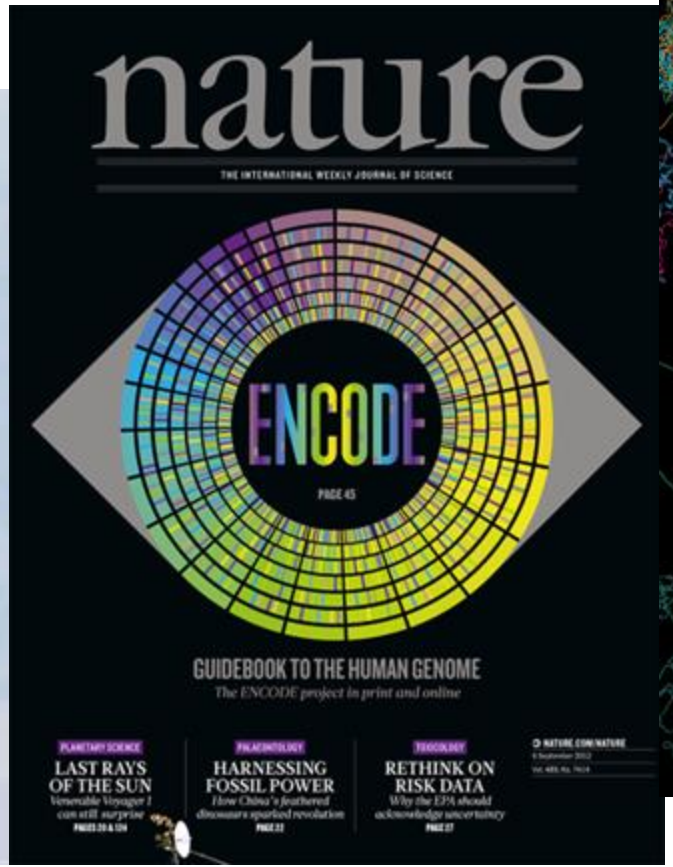
Long Non-Coding RNAs

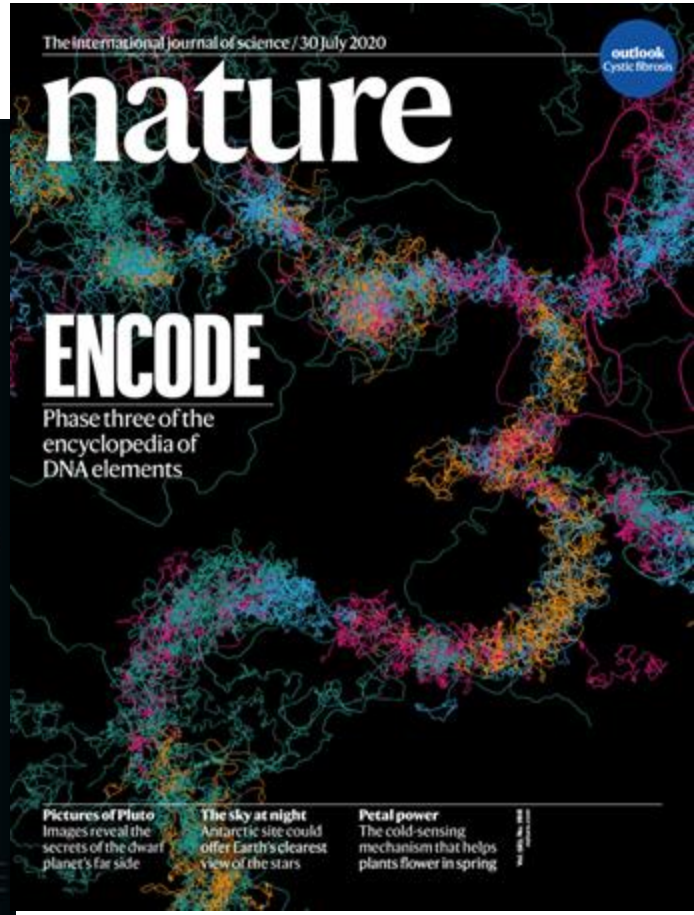
EMBL-EBI

Biotype	Definition
IG_C_gene IG_D_gene IG_J_gene IG_LV_gene IG_V_gene TR_C_gene TR_J_gene TR_V_gene TR_D_gene	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
IG_pseudogene IG_C_pseudogene IG_J_pseudogene IG_V_pseudogene TR_V_pseudogene TR_J_pseudogene	
Mt_rRNA Mt_tRNA miRNA misc_RNA rRNA scRNA snRNA snoRNA ribozyme sRNA scaRNA	
lncRNA	
TEC	To be Experimentally Confirmed. This is used for non-spliced EST clusters that have polyA features. This category has been specifically created for the ENCODE project to highlight regions that could indicate the presence of protein coding genes that require experimental validation, either by 5' RACE or RT-PCR to extend the transcripts, or by confirming expression of the putatively-encoded peptide with specific antibodies.
nonsense_mediated_decay	If the coding sequence (following the appropriate reference) of a transcript finishes >50bp from a downstream splice site then it is tagged as NMD. If the variant does not cover the full reference coding sequence then it is annotated as NMD if NMD is unavoidable i.e. no matter what the exon structure of the missing portion is the transcript will be subject to NMD.
non_stop_decay	Transcript that has polyA features (including signal) without a prior stop codon in the CDS, i.e. a non-genomic polyA tail attached directly to the CDS without 3' UTR. These transcripts are subject to degradation.
retained_intron	Alternatively spliced transcript believed to contain intronic sequence relative to other, coding, variants.
protein_coding	Contains an open reading frame (ORF).
protein_coding_LoF	Not translated in the reference genome owing to a SNP/DIP but in other individuals/haplotypes/strains the transcript is translated. Replaces the polymorphic_pseudogene transcript biotype.
protein_coding_CDS_not_defined	Transcript that belongs to a protein_coding gene and doesn't contain an ORF. Replaces the processed_transcript transcript biotype in protein_coding genes.
processed_transcript	
non_coding	
ambiguous_orf	
sense_intronic	
sense_overlapping	
antisense/antisense_RNA	
known_ncrna	
pseudogene	
processed_pseudogene	
transcribed_processed_pseudogene transcribed_unprocessed_pseudogene transcribed_unitary_pseudogene	Pseudogene where protein homology or genomic structure indicates a pseudogene, but the presence of locus-specific transcripts indicates expression.
translated_processed_pseudogene translated_unprocessed_pseudogene	Pseudogene that has mass spec data suggesting that it is also translated.
unitary_pseudogene	A species-specific unprocessed pseudogene without a parent gene, as it has an active orthologue in another species.
unprocessed_pseudogene	Pseudogene that can contain introns since produced by gene duplication.
artifact	Annotated on an artifactual region of the genome assembly.
lincRNA	Long, intervening noncoding (linc) RNA that can be found in evolutionarily conserved, intergenic regions.
macro_lncRNA	Unspliced lncRNA that is several kb in size.
3prime_overlapping_ncRNA	Transcript where ditag and/or published experimental data strongly supports the existence of short non-coding transcripts transcribed from the 3'UTR.
disrupted_domain	Otherwise viable coding region omitted from this alternatively spliced transcript because the splice variation affects a region coding for a protein domain.
vaultRNA/vault_RNA	Short non coding RNA gene that forms part of the vault ribonucleoprotein complex.
bidirectional_promoter_lncRNA	A non-coding locus that originates from within the promoter region of a protein-coding gene, with transcription proceeding in the opposite direction on the other strand.

https://www.gencodegenes.org/pages/biotypes.html

The **Encyclopedia of DNA Elements** (**ENCODE**) is a public research project which aims "to build a comprehensive parts list of functional elements in the human genome."





The Ensembl-HAVANA Team

HAVANA - **H**uman **A**nd **V**ertebrate **AN**alysis and **A**nnotation

GENCODE

The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation.

https://www.gencodegenes.org/

EMBL-EBI

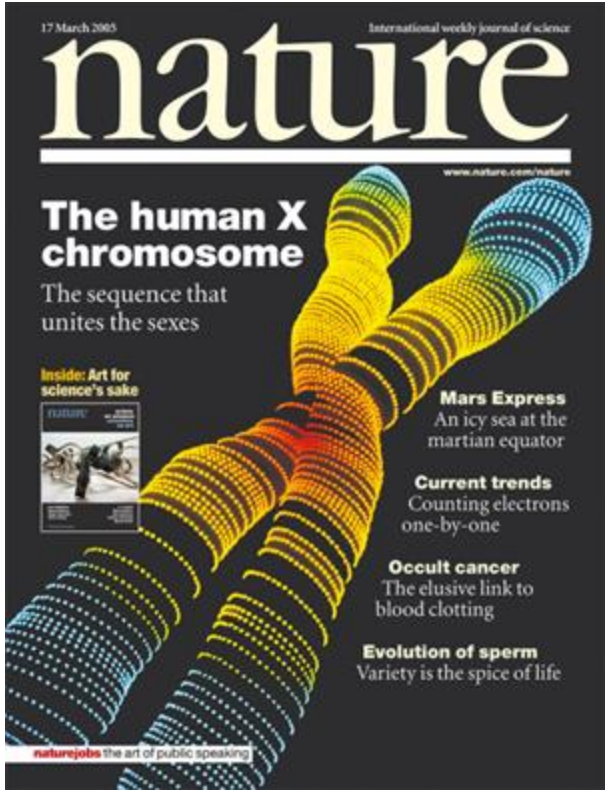
Reference genomes:

Human ~3Gb:
22 chromosomes + sex chromosomes
GRCh38p14

Mouse ~3 Gb:
19 chromosomes + sex chromosomes
GRCm39

Zebrafish ~1.4 Gb:
25 chromosomes, no specific sex chromosomes
GRCz11

Many chromosome publications

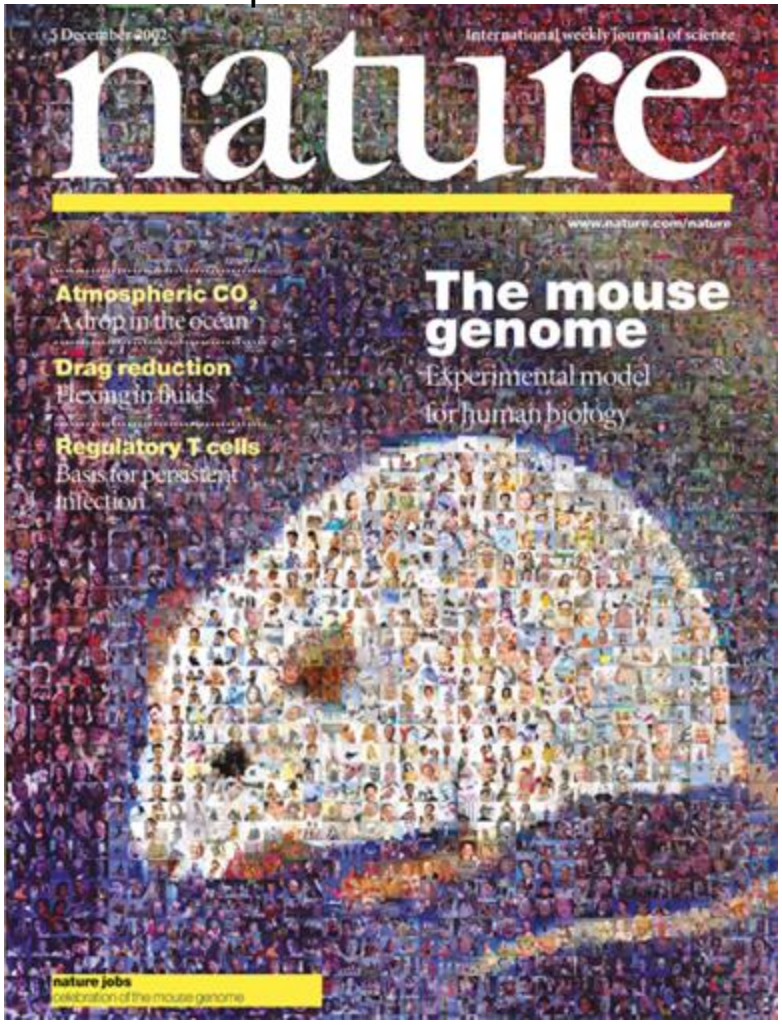


First pass manual annotation completed 2013

HGNC
HUGO Gene Nomenclature Committee
The resource for approved human gene nomenclature

https://www.genenames.org/

Genome published Dec 2002



Annotation and Knock Out designs

Knockout Mouse Project (KOMP)
Since 2006, scientists around the world have been working together to generate a targeted knockout mutation for every gene in the mouse genome.

The EUCOMM program

IMPC
International Mouse Phenotyping Consortium

First pass manual annotation started 2012 completed 2018

MGI
Mouse Genome Informatics

https://www.informatics.jax.org/

EMBL-EBI

Letter | Open access | Published: 17 April 2013

The zebrafish reference genome sequence and its relationship to the human genome

Kerstin Howe, Matthew D. Clark, Carlos F. Torroja, James Torrance, Camille Berthelot, Matthieu Muffato, John E. Collins, Sean Humphray, Karen McLaren, Lucy Matthews, Stuart McLaren, Ian Sealy, Mario Caccamo, Carol Churcher, Carol Scott, Jeffrey C. Barrett, Romke Koch, Gerd-Jörg Rauch, Simon White, William Chow, Britt Kilian, Leonor T. Quintais, José A. Guerra-Assunção, Yi Zhou, ... Derek L. Stemple

+ Show authors

Nature **496**, 498–503 (2013) | Cite this article

ZFIN

https://zfin.org/

EMBL-EBI

Genome Reference Consortium

- Correct regions in the genome
- To close as many gaps as possible
- To produce alternative assemblies of structurally variant loci where necessary
- Scientific community can report loci in need of review
- Human, mouse, zebrafish, rat and chicken

wellcome sanger institute

EMBL-EBI

ZFIN

Washington University School of Medicine in St. Louis

McDonnell Genome Institute

NIH National Library of Medicine
National Center for Biotechnology Information

RGD

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

Do we know how many genes there are?

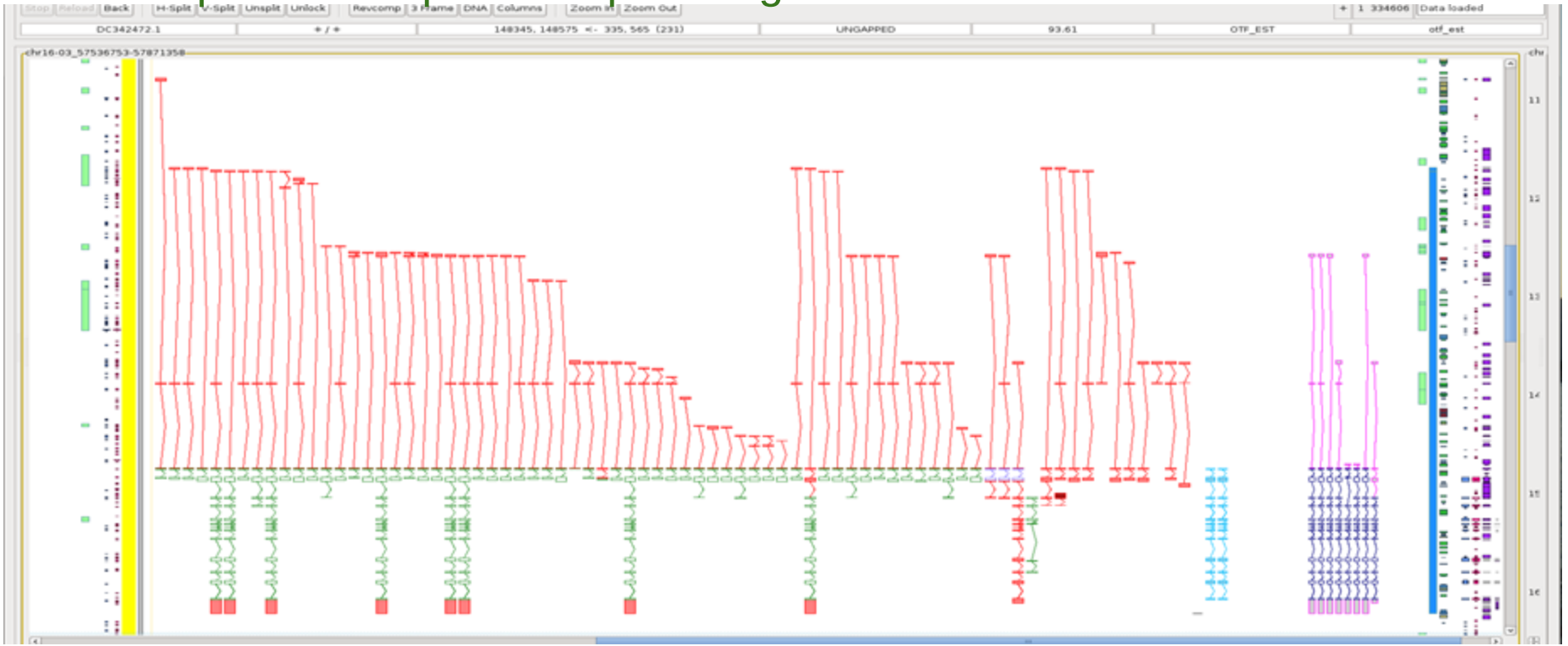
Protein coding genes

1980s	100,000
2000	40,000
Today	?

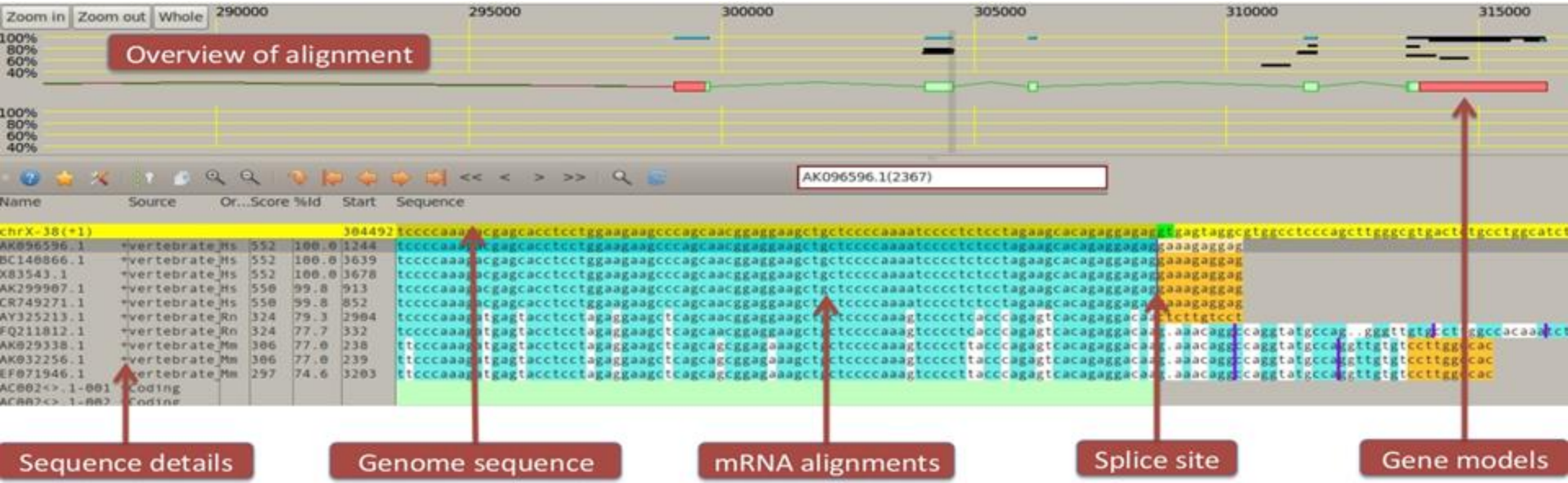
Otter/ZMAP

GPR56:

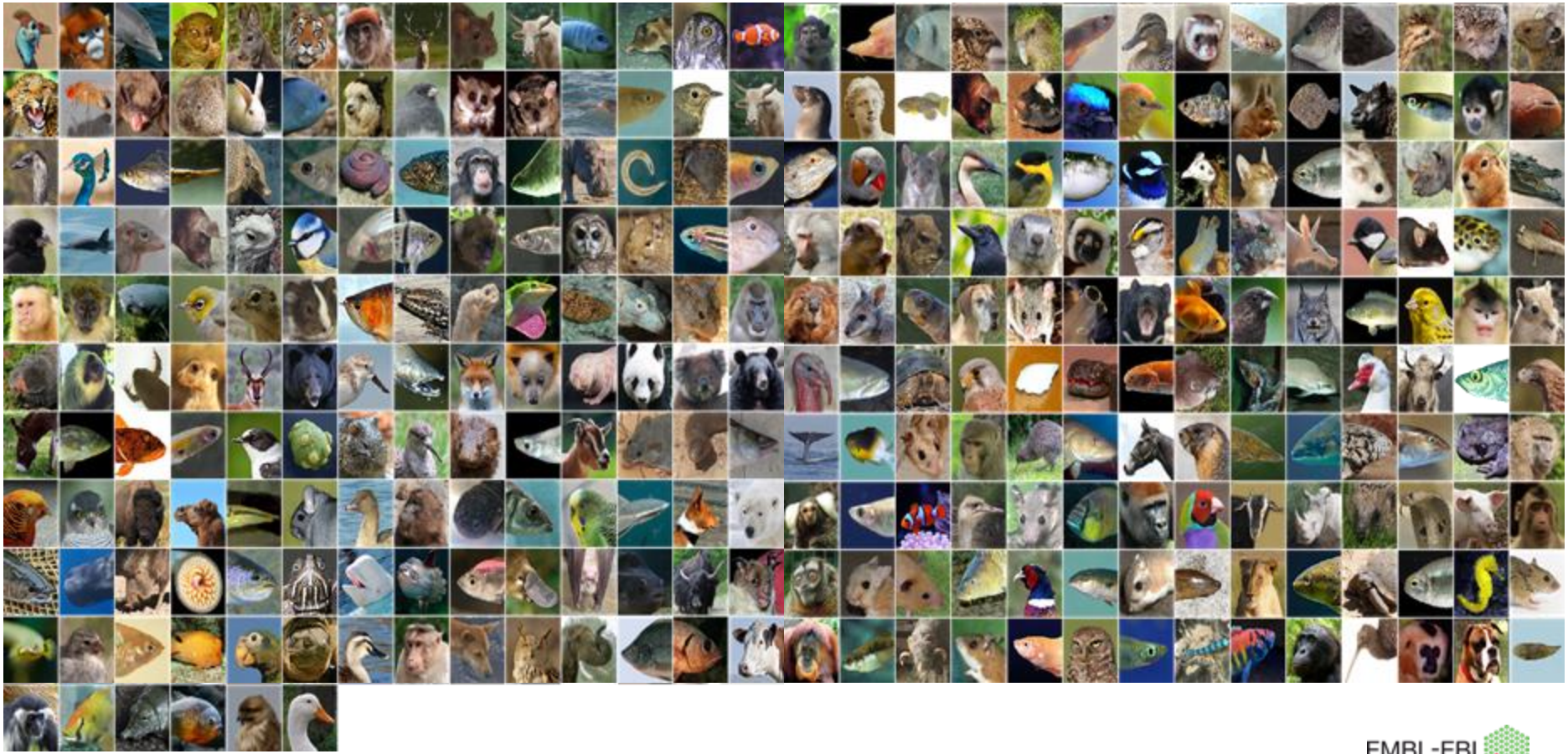
Human G protein-coupled receptor 56 gene



Blixem:
Interactive browser of pairwise alignments that have been stacked up in a multiple alignment



Automated Annotation

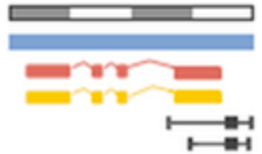


EMBL-EBI

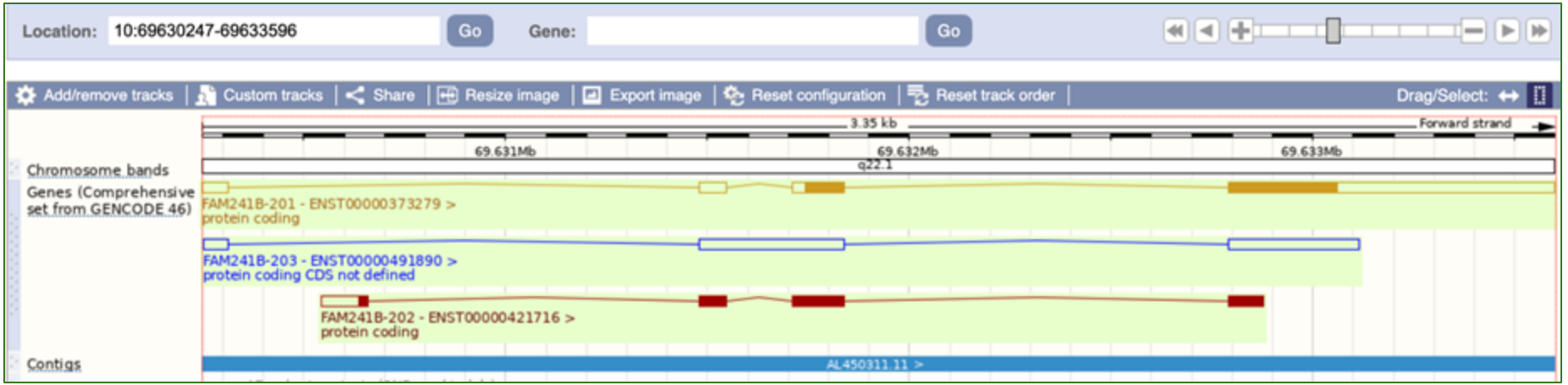
Manual Annotation

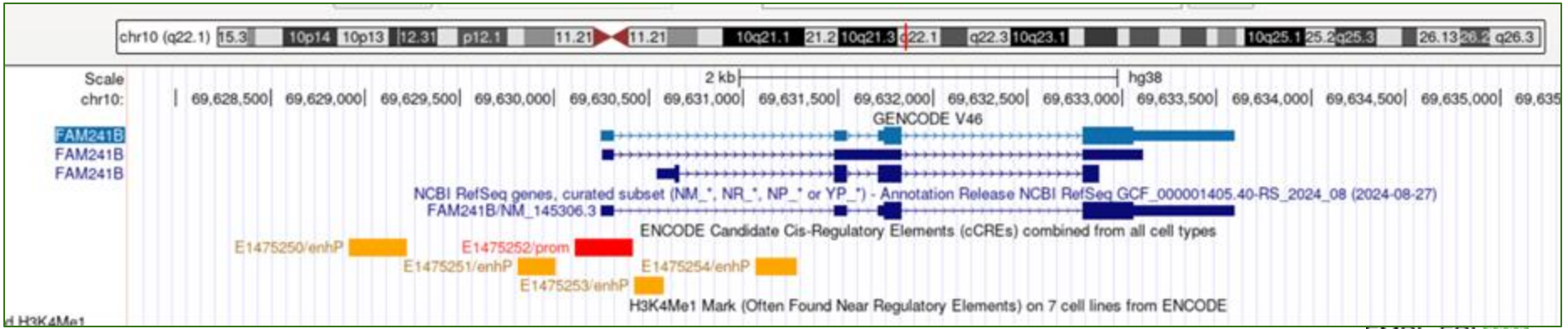
Ensembl features

- Genomes and gene builds for >300 species
- Variation data
- Compara (alignments, gene trees, homologues)
- Regulatory build
- BioMart (data export)
- Tools for data processing, e.g. VEP
- Display your own data
- Programmatic access via APIs
- Completely Open Source (FTP, GitHub)

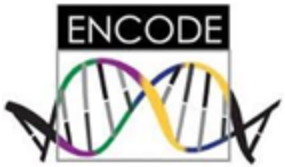




FAM241B in Ensembl and UCSC genome browsers





Consortia and resources that have utilized GENCODE data:



ClinVar (MANE)

Expression Atlas

GENCODE

Human | Mouse | How to access data | FAQ | Documentation | About us

Human

Statistics about the current GENCODE Release (version 46)

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

General stats

Total No of Genes	63086	Total No of Transcripts	254070
Protein-coding genes	19411	Protein-coding transcripts	89581
- readthrough genes (not included)	654	- full length protein-coding	64695
Long non-coding RNA genes	20310	- partial length protein-coding	24886
Small non-coding RNA genes	7565	Nonsense mediated decay transcripts	21774
Pseudogenes	14716	Long non-coding RNA loci transcripts	59927
- processed pseudogenes	10657		
- unprocessed pseudogenes	3564		
- unitary pseudogenes	258	Total No of distinct translations	65650
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13620
- protein coding segments	411		
- pseudogenes	237		

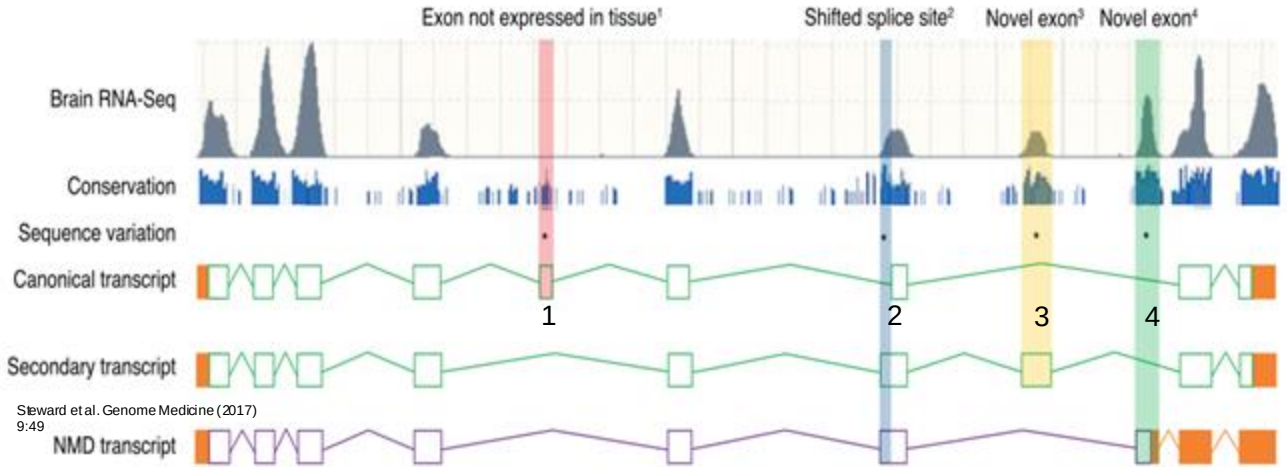
More about GENCODE Human

- Current human data
- Release history
- Statistics
- Data format
- FTP site

https://www.gencodegenes.org/human/stats.html

EMBL-EBI

- Transcript annotation is central to understanding variant effects



Steward et al. Genome Medicine (2017) 9:49

- How to report variant 3? Conflicting report depending on which transcript is used as reference?
- How to report variant 4? Do we need a new transcript?
- Annotation updated as new biology is uncovered

Human geneset refinement

Two comprehensive independent human reference transcript sets:

EMBL-EBI

NIH National Library of Medicine
National Center for Biotechnology Information

Why is this a problem?

- Resources use either Ensembl/GENCODE (EBI) or RefSeq (NCBI)
- Differences make it hard to for researchers to exchange data or translate co-ordinates
- Standardise transcript set across genomics browsers

What's the solution?

- Identify a representative transcript that captures the most information about each protein-coding gene (not just the longest/first one)
- Will also help standardise clinical reporting

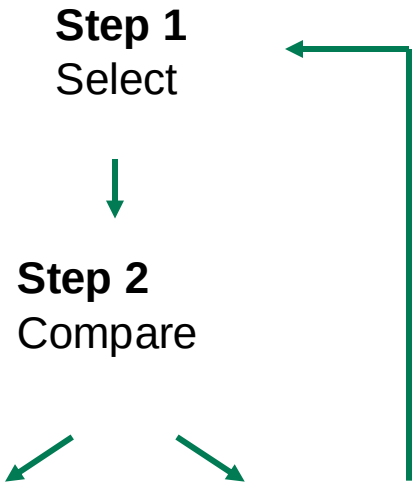
MANE project

Matched Annotation from NCBI and EMBL-EBI

- A transcript set with the following attributes:
  - Match to GRCh38
  - One MANE Select transcript per locus
  - 100% identical between the RefSeq and corresponding Ensembl transcript for 5'UTR, CDS, and 3'UTR
  - No new identifier
- Transcripts should be:
  - Well-supported, conserved, expressed
  - Representative of biology at each locus
- Fairly stable, but will allow updates when necessary

All the transcripts we annotate should always be considered and we are certainly NOT saying that biology can be simplified to a single transcript at each genomic locus

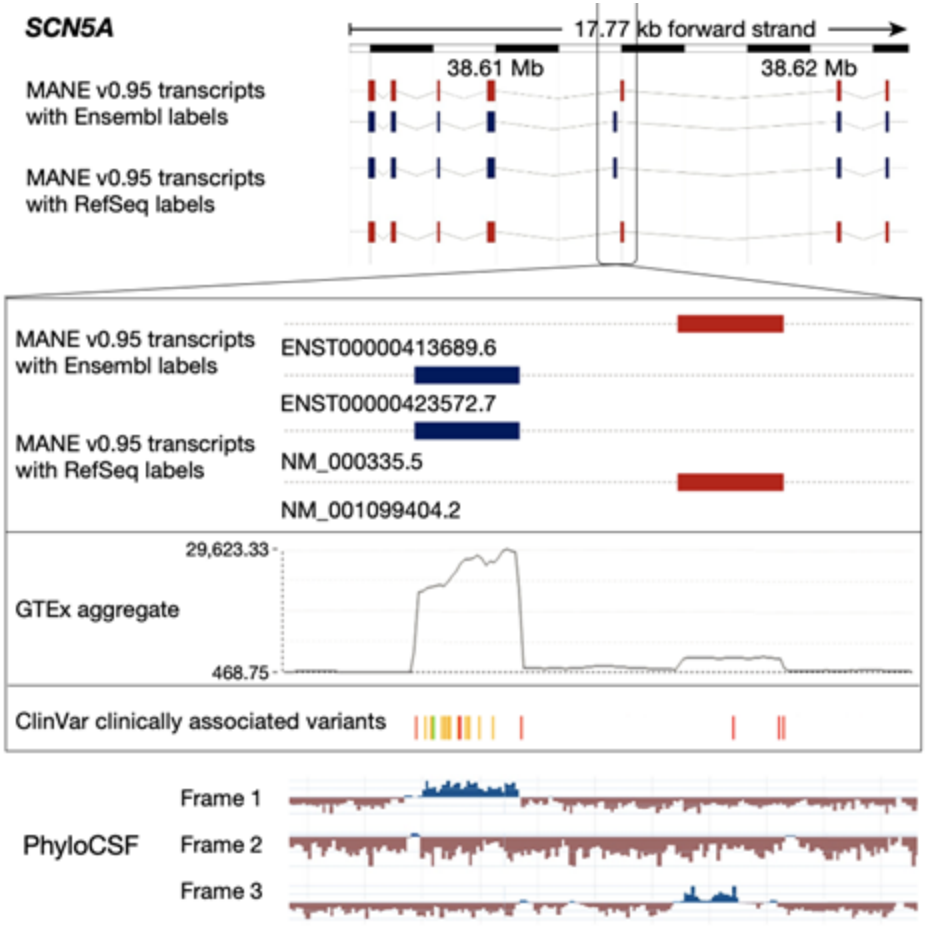
MANE Select methodology



Step 3a
Same splicing/CDS:
- Match ends
- Release

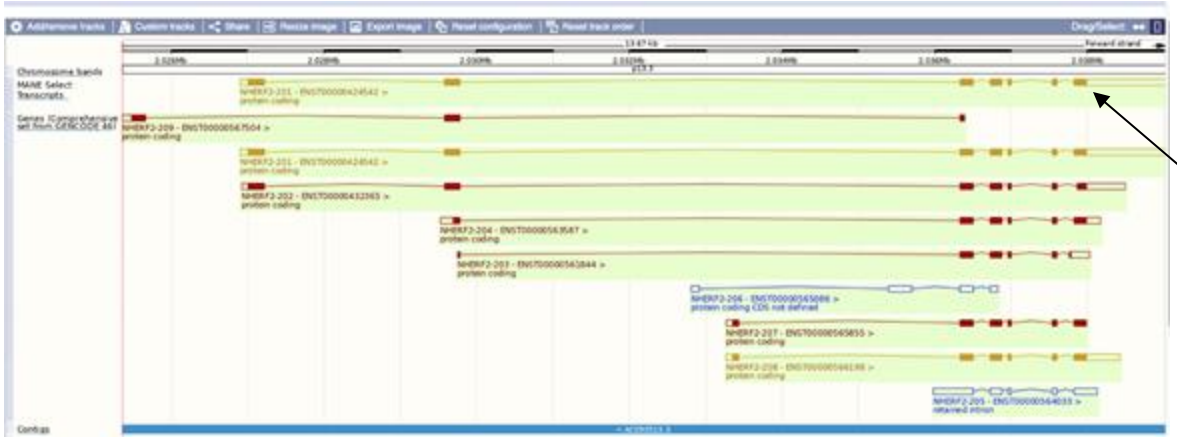
Step 3b
NOT same splicing/CDS:
- Manually curate
- Changes to pipeline

MANE Plus Clinical



EMBL-EBI

NHERF2



MANE Select

19,338 MANE Select , 66 MANE Plus Clinical v1.4

EMBL-EBI

The Forefront of Genomics

Questions & Answers #1

Too many transcripts?

Have we found all the genes?

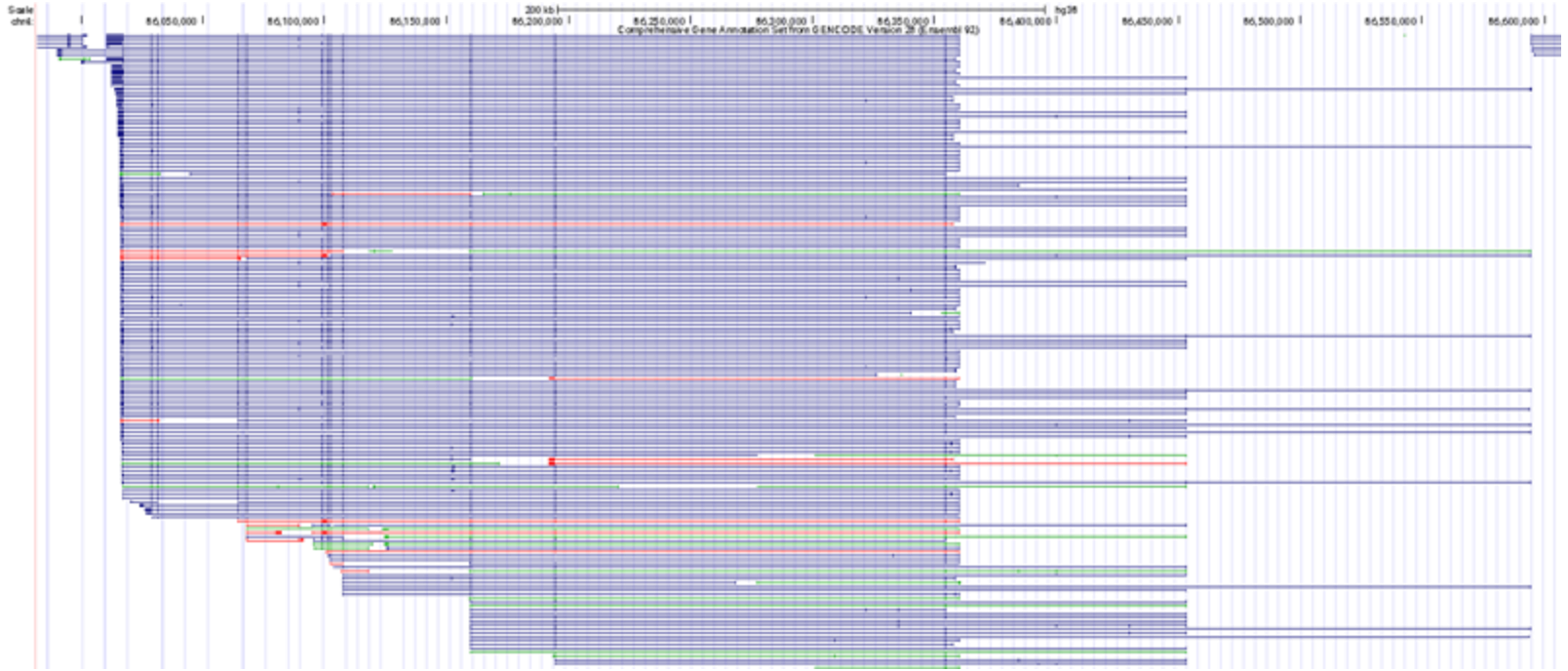
What about all the genomes?

Too many transcripts?

Have we found all the genes?

What about all the genomes?

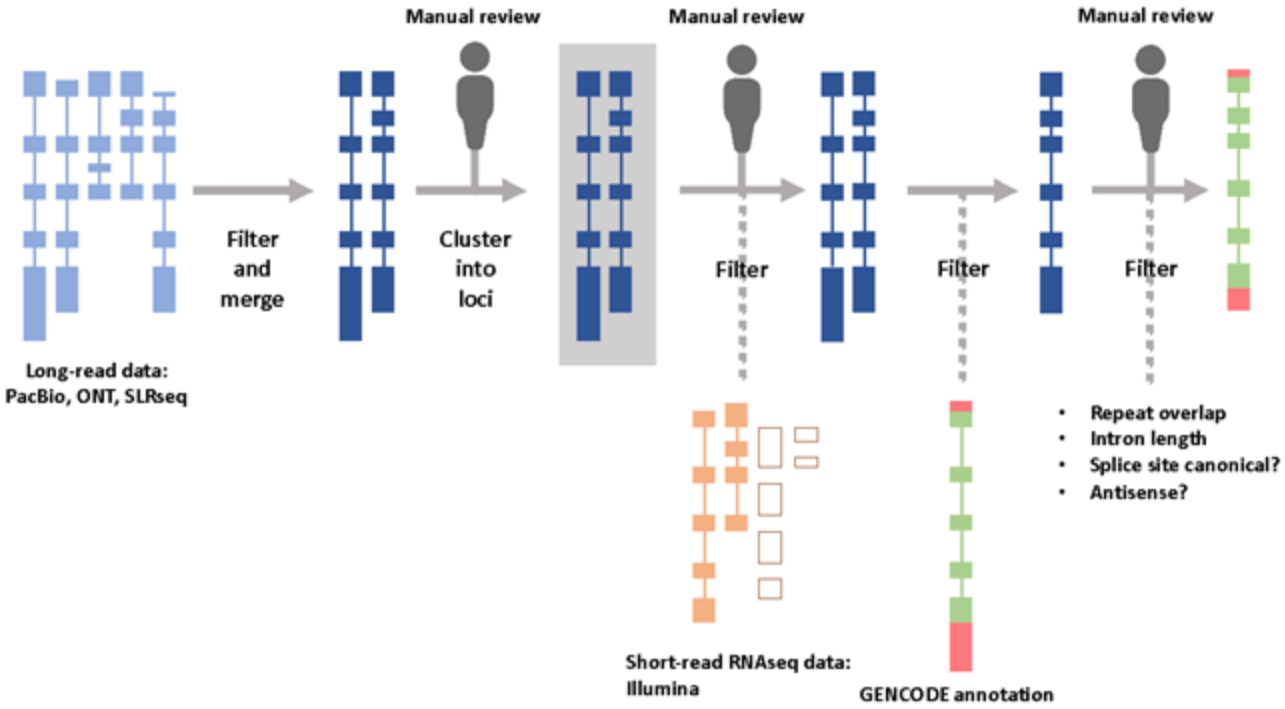
More data = more annotation



MAPK10 – 192 alternatively spliced transcripts

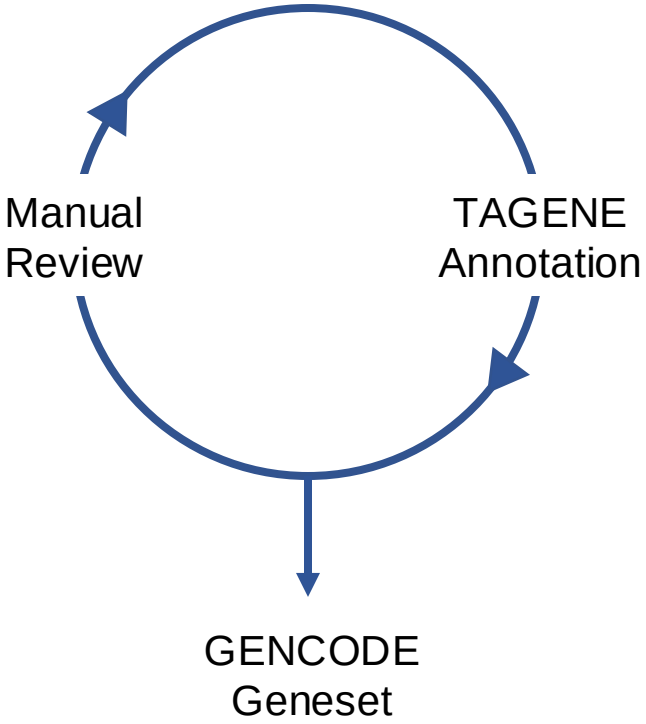
EMBL-EBI

TAGENE



EMBL-EBI

TAGENE development

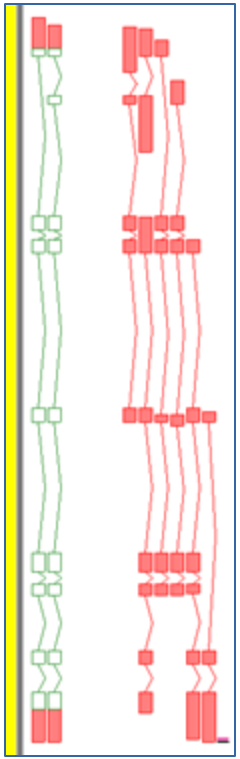


Improved biotype assignment

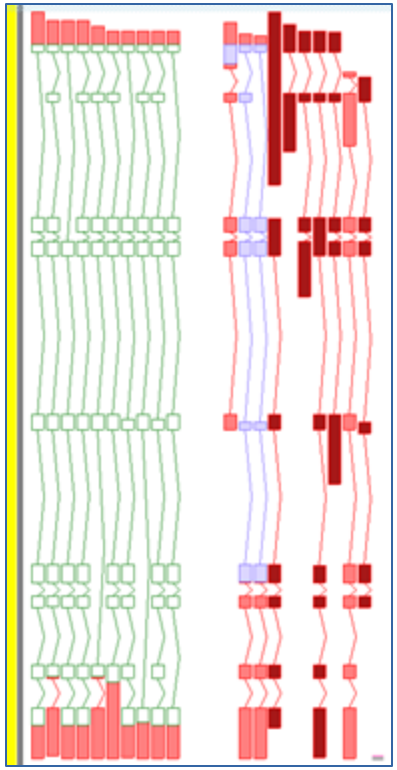
- better rules
- better filtering of confounders
  - readthrough transcripts

TAGENE development

SCAMP3

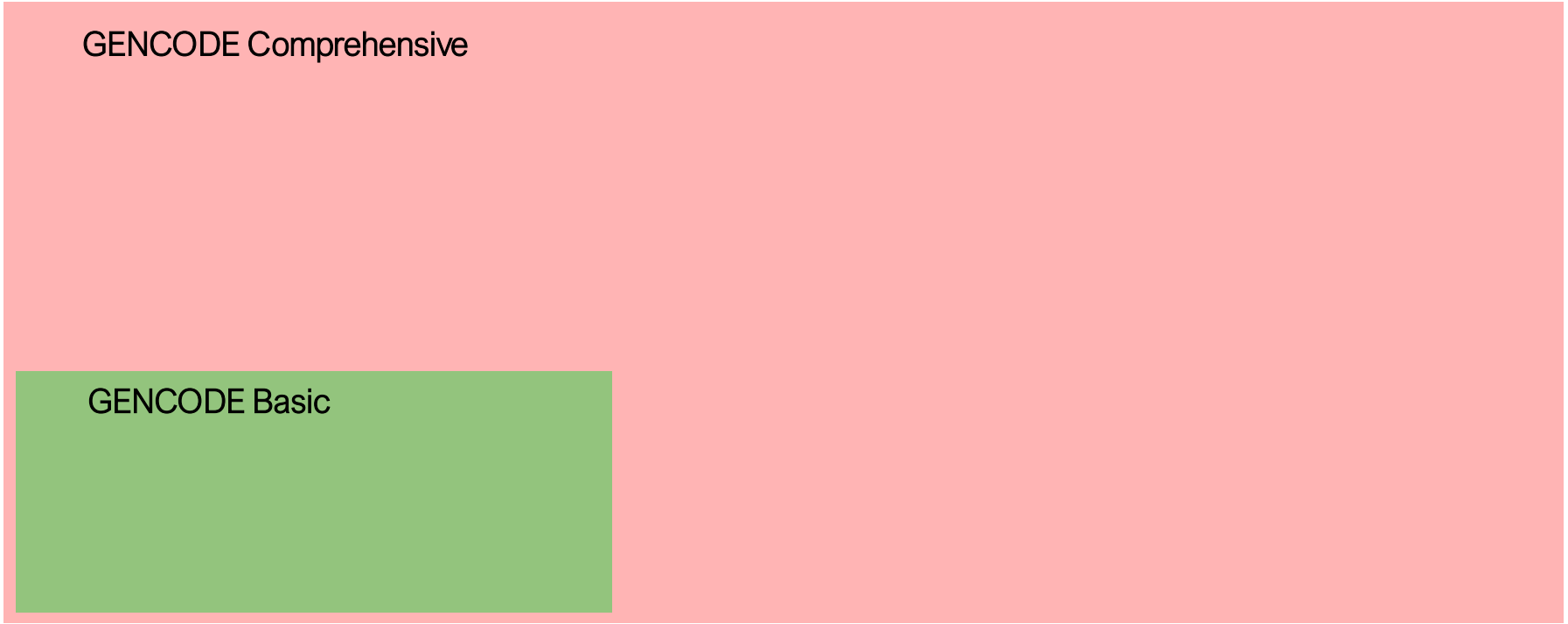


Before



After

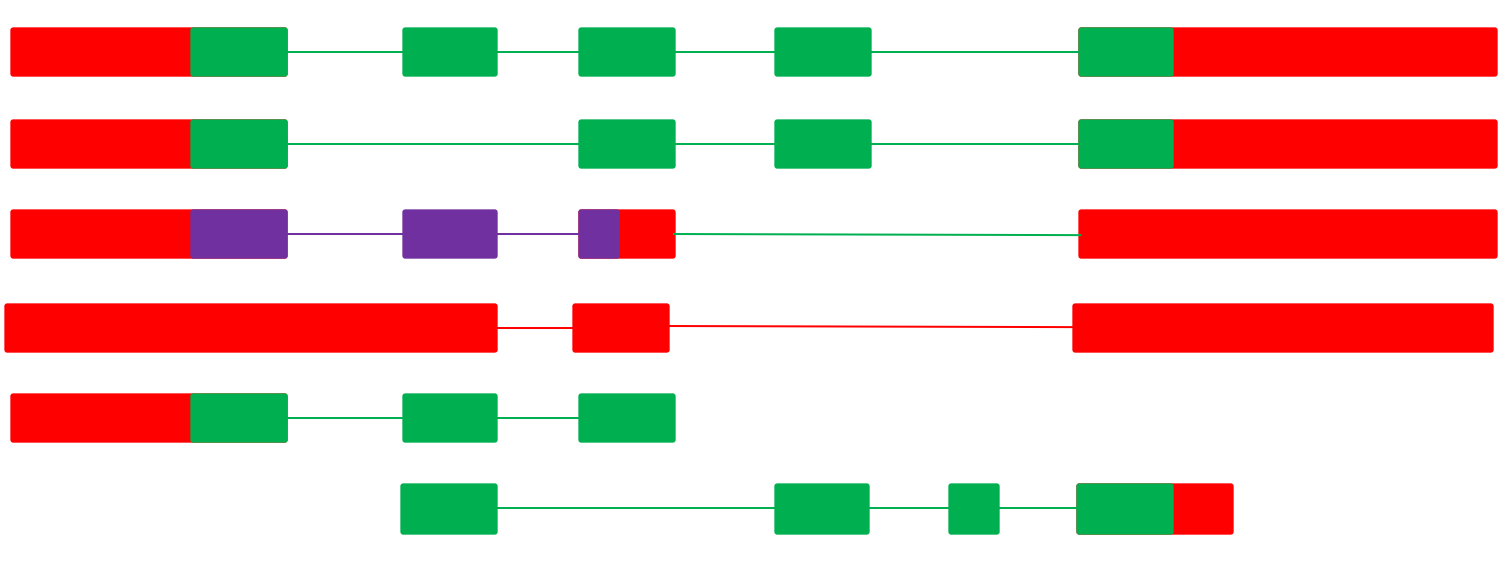
Current transcript classification



EMBL-EBI

Classifying transcripts

Basic

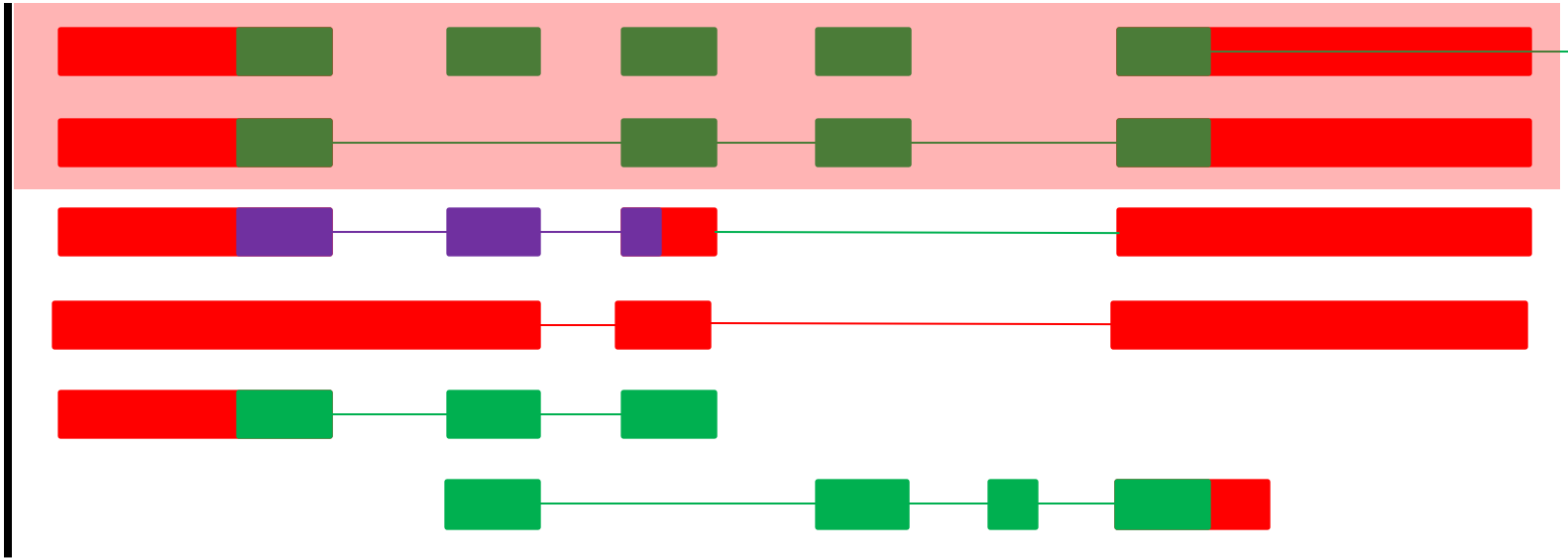


Comprehensive

EMBL-EBI

Classifying transcripts

Basic

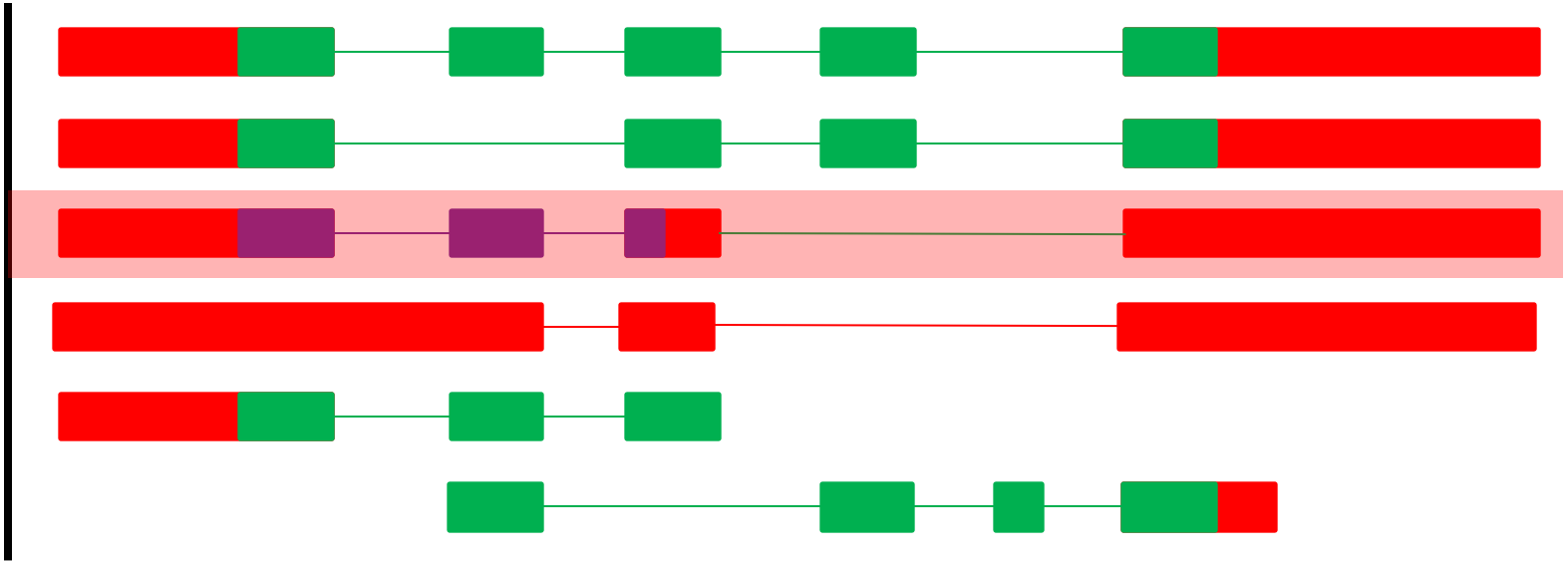


Comprehensive

EMBL-EBI

Classifying transcripts

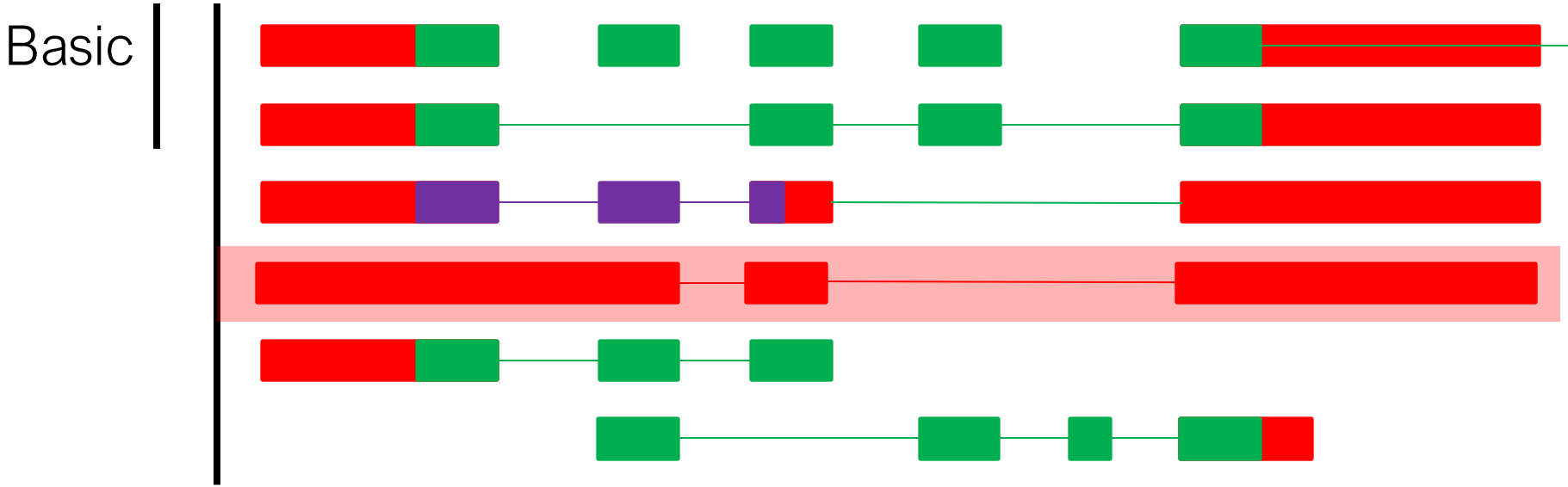
Basic



Comprehensive

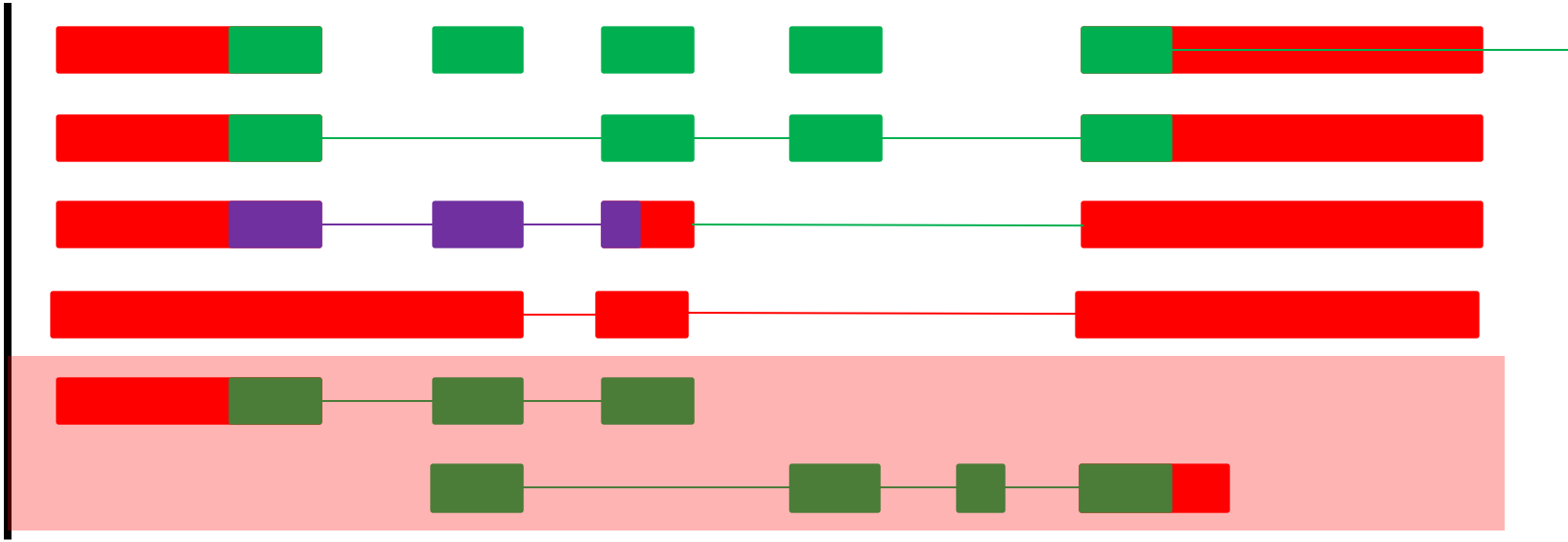
EMBL-EBI

Classifying transcripts



Classifying transcripts

Basic



Comprehensive

EMBL-EBI

New long transcriptomic data, breaks model

Basic

Comprehensive

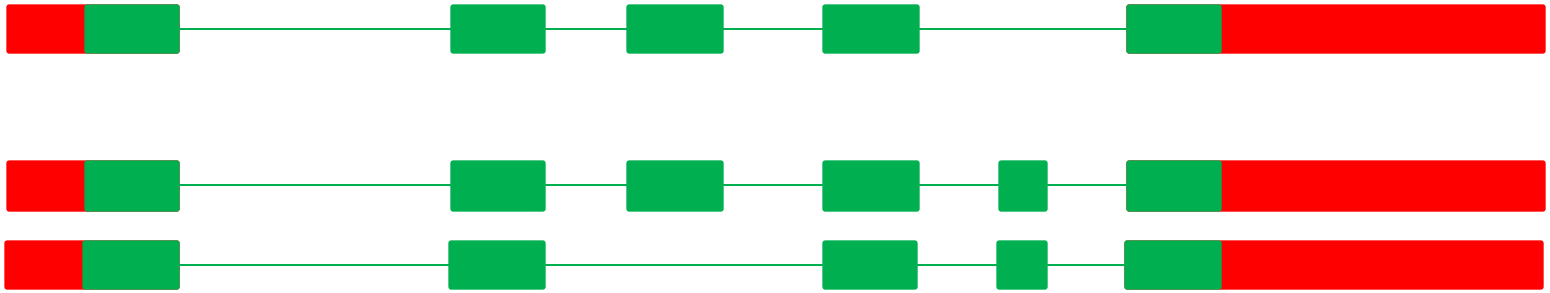
New long transcriptomic data, breaks model

How do we plan to manage this?

- MANE Select
- 1 representative transcript per coding gene
  - Limited extension to other biotypes in future
- Gives a reference transcript for coding genes
  - Ensembl canonical for other biotypes
- Having a reference essential in developing transcript hierarchy

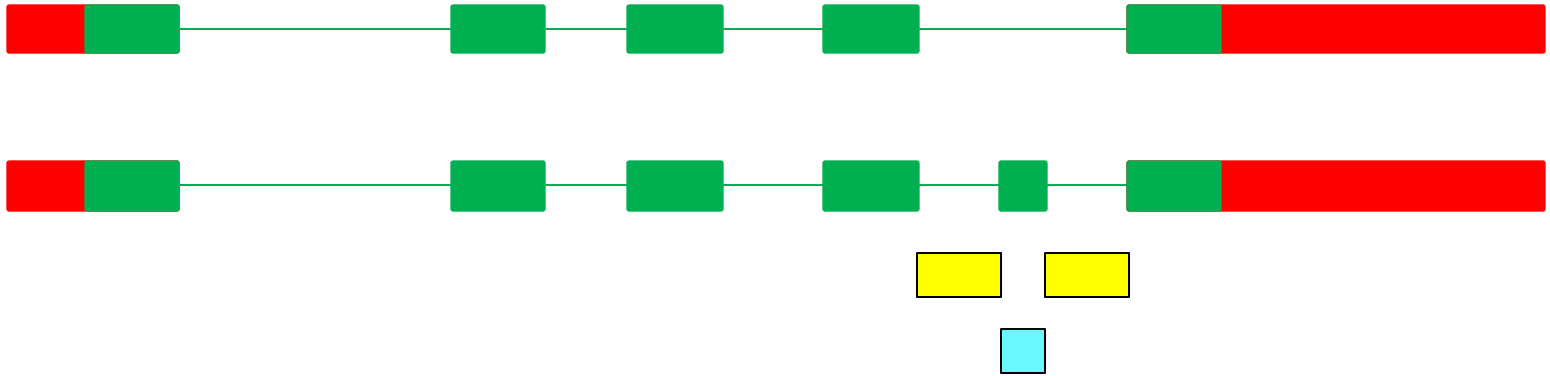
Reference transcript supports comparison

MS



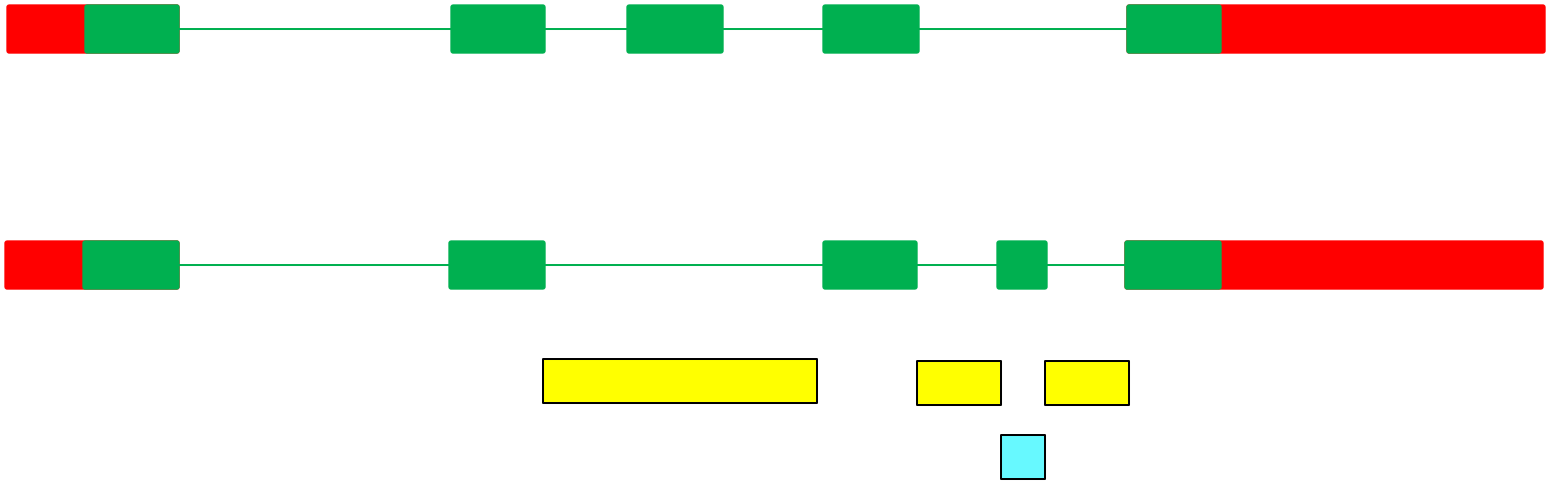
Reference transcript supports comparison

MS



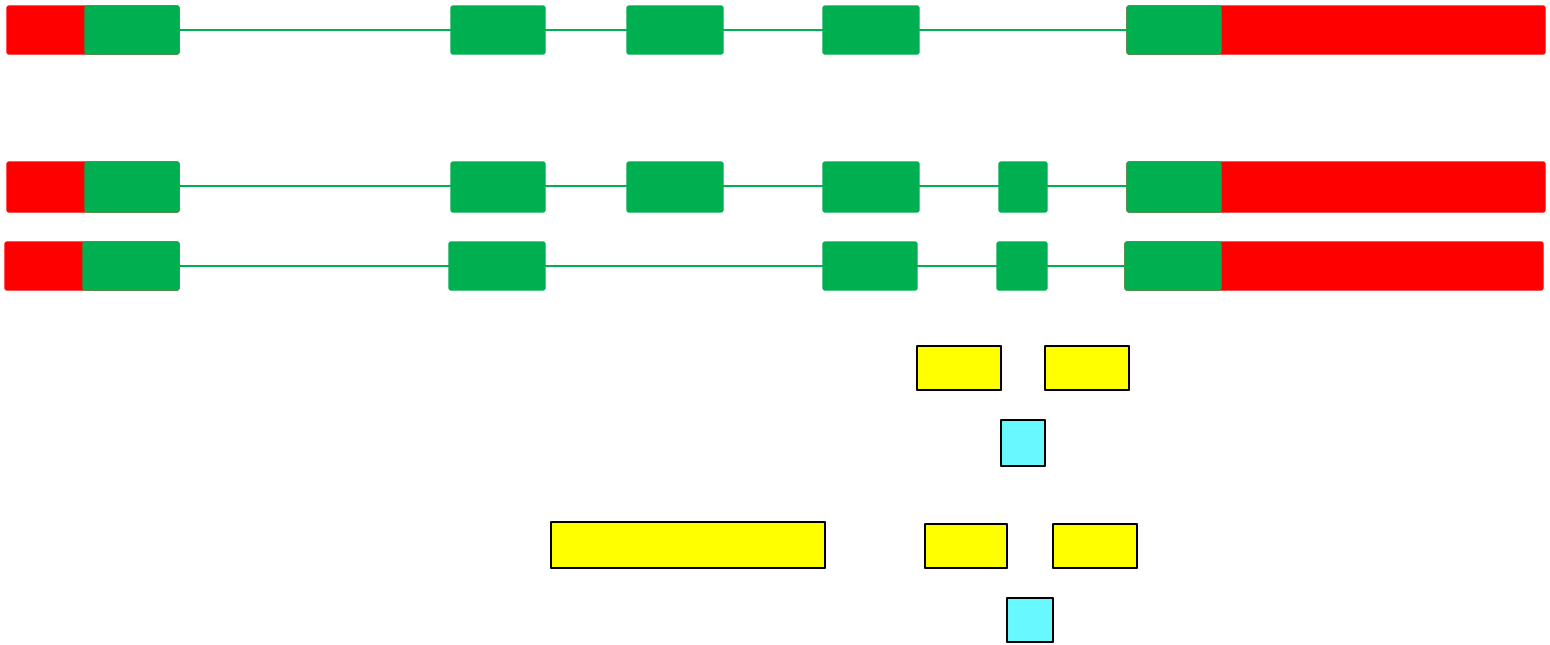
Reference transcript supports comparison

MS

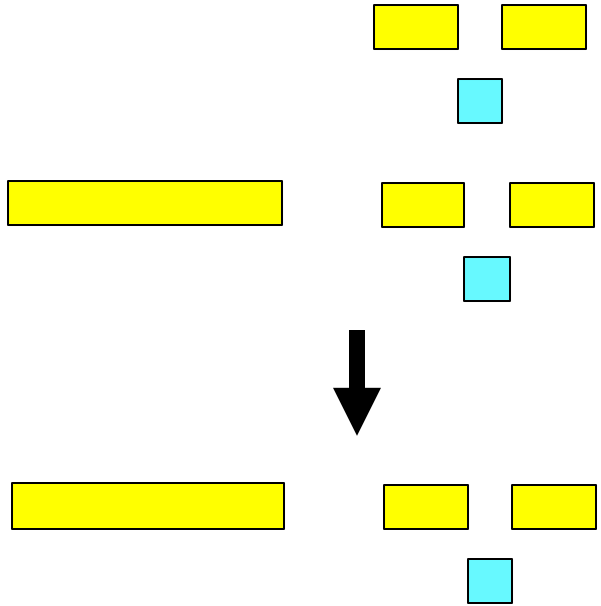


Reference transcript supports comparison

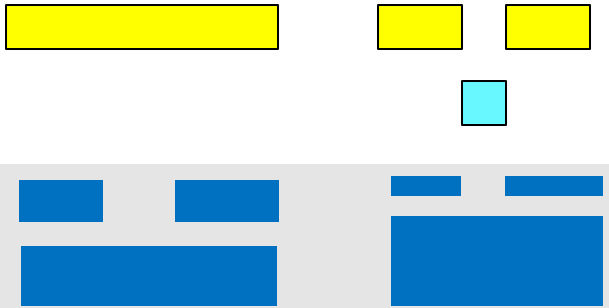
MS



Collapse define NR set of non-reference features



Test non-reference features to identify those with functional significance



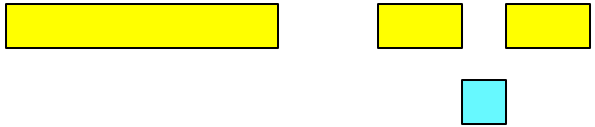
Introns:

RNAseq (Recount3)

Compare junction inclusion ratio (JIR) reference vs alt

EMBL-EBI

Test non-reference features to identify those with functional significance



Exons:

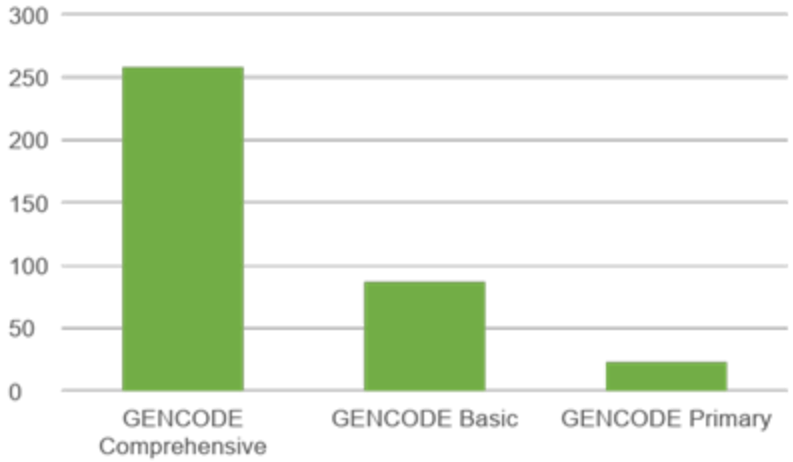
Conservation and constraint

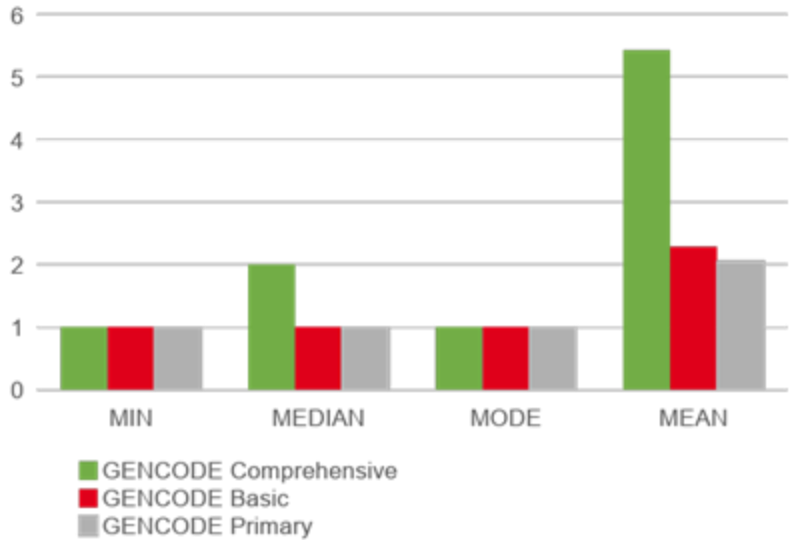
PhastCons and PhyloCSF

From scoring features to GENCODE Primary

1. Identify set of features (introns and exons) that exceed threshold
2. Identify the transcripts that those features are part of
3. Use ‘Ensembl Select’ pipeline to generate per transcript scores and ranking
4. Add highest scoring feature-containing transcripts to GENCODE Primary
5. Retain rankings for GENCODE Comprehensive transcripts

GENCODE Primary initial set





EMBL-EBI

GENCODE Primary, MANE and Ranking

GENCODE Primary

MANE Select
MANE Plus Clinical

GENCODE transcripts 3-5

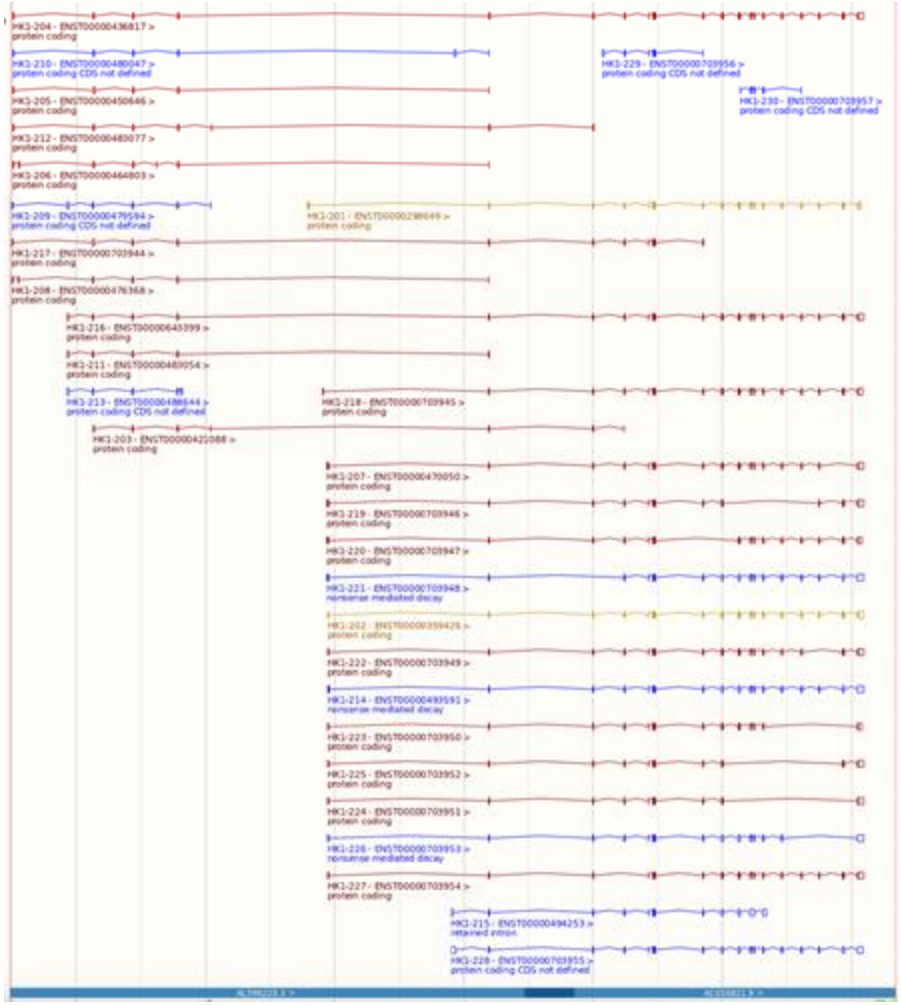
GENCODE transcripts 6-100

GENCODE Comprehensive

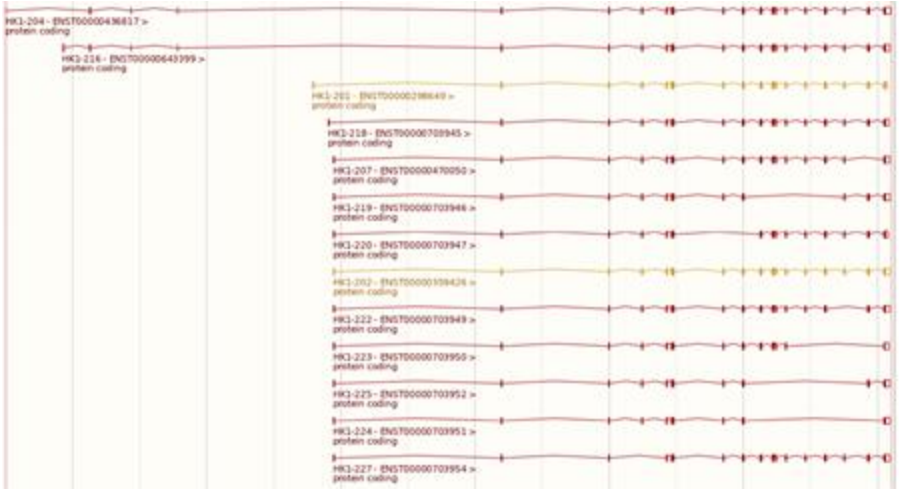
EMBL-EBI

Ensembl location view of the *HK1* gene on human chromosome 10 (e112)

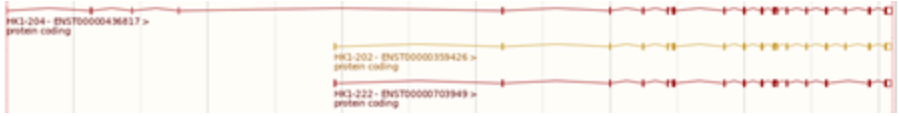
GENCODE Comprehensive (30 transcripts)



GENCODE Basic (13 transcripts)



GENCODE Primary



MAPK10: 192, 87, 18

EMBL-EBI

Too many transcripts?

Have we found all the genes?

What about all the genomes?

Microproteins

Under 100aa: not usually looking for proteins that small

The major approaches so far used by GENCODE:

- Evolutionary analysis
- Mass spectrometry-based searches for smORFs
- Ribo-seq / ribosome profiling

GENCODE microprotein discovery: evolution

Genome Res. 2019 Dec;29(12):2073-2087. doi: 10.1101/gr.246462.118. Epub 2019 Sep 19.

Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci

Jonathan M Mudge [# 1], Irwin Jungreis [# 2 3], Toby Hunt [1], Jose Manuel Gonzalez [1], James C Wright [4], Mike Kay [1], Claire Davidson [1], Stephen Fitzgerald [5], Ruth Seal [1 6], Susan Tweedie [1], Liang He [2 3], Robert M Waterhouse [7 8], Yue Li [2 3], Elspeth Bruford [1 6], Jyoti S Choudhary [4], Adam Frankish [1], Manolis Kellis [2 3]

We annotated 144 new human protein-coding genes based on observation of protein constraint

50 are microproteins

BIOINFORMATICS

Vol. 27 ISMB 2011, pages i275–i282
doi:10.1093/bioinformatics/btr209

PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions

Michael F. Lin[1,2,*], Irwin Jungreis[1,2] and Manolis Kellis[1,2,*]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street 32-D510, Cambridge, MA 02139 and [2]The Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

GENCODE microprotein discovery: evolution

FAM240B

78aa microprotein

Gene completely missed during first pass human and mouse manual annotation

- Following PhyloCSF identification, annotation was possible with short-read data
- **Expression of the gene is specific to fetal eye in human and mouse**

GENCODE microprotein discovery: evolution

e.g. ***TINCR***, once a famous lncRNA

TINCR ubiquitin domain containing [Eublepharis macularius]

Sequence ID: XP_054835179.1 Length: 87 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 87 GenPept Graphics Next Match Previous

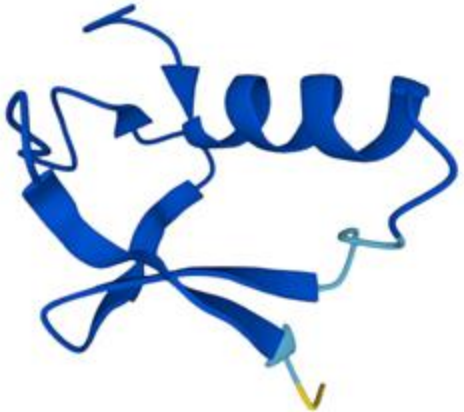
Score	Expect	Method	Identities	Positives	Gaps
133 bits(334)	3e-38	Compositional matrix adjust.	61/87(70%)	76/87(87%)	0/87(0%)

```
Query  1   MEGLRRGLSRWKRYHIKVHLADEALLLPLTVRPRDTLSDLRAQLVGQGVSSWKRAFYYNA  60
           ME LRR LSRWKRYHIKVHLA++ +LLPLTVRP D + DLRA LV +GV+SWK+ FYYNA
Sbjct  1   METLRRSLSRWKRYHIKVHLAEDDVLLPLTVRPTDNMMDLRAHLVREGVTSWKKTFYYNA  60

Query  61  RRLDDHQTVRDARLQDGSVLLLVSDPR  87
           R+L +H+TV++A++Q+GSVLLLVSD R
Sbjct  61  RQLGEHETVKEAKIQNGSVLLLVSDER  87
```

A ubiquitin-like protein encoded by the “noncoding” RNA TINCR promotes keratinocyte proliferation and wound healing

Akihiro Nita, Akinobu Matsumoto, Ronghao Tang, Chisa Shiraishi, Kazuya Ichihara, Daisuke Saito, Mikita Suyama, Tomoharu Yasuda, Gaku Tsuji, Masutaka Furue, Bumpei Katayama, Toshiyuki Ozawa, Tetsuasa Murata, [...], Keiichi I. Nakayama



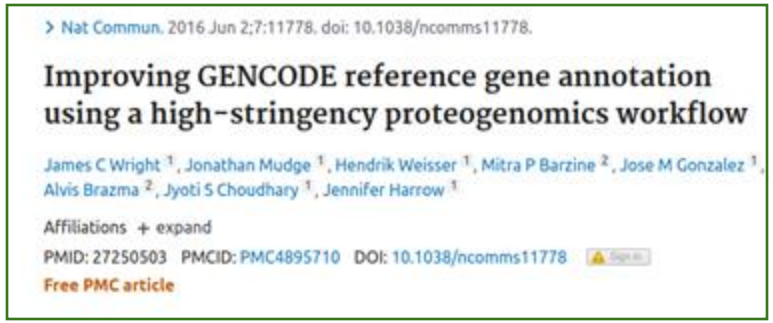
Alphafold structure prediction

GENCODE microprotein discovery: proteomics

Proteomics:

2016 reanalysis of the Pandey / Kuster lab proteomics datasets

Mass spectrometry analysis done by J. Choudhary & J. Wright



Found 16 missing protein-coding genes

We have not yet found any microproteins using MS-first approaches

EMBL-EBI

GENCODE microprotein discovery: Ribo-seq

> Nat Biotechnol. 2022 Jul;40(7):994-999. doi: 10.1038/s41587-022-01369-0.

Standardized annotation of translated open reading frames

Jonathan M Mudge [#1], Jorge Ruiz-Orera [#2], John R Prensner [#3,4,5], Marie A Brunet [6], Ferriol Calvet [7], Irwin Jungreis [8,9], Jose Manuel Gonzalez [7], Michele Magrane [7], Thomas F Martinez [10,11], Jana Felicitas Schulz [12], Yucheng T Yang [13,14], M Mar Albà [15,16], Julie L Aspden [17,18], Pavel V Baranov [19], Ariel A Bazzini [20,21], Elspeth Bruford [7,22], Maria Jesus Martin [7], Lorenzo Calviello [23,24], Anne-Ruxandra Carvunis [25,26], Jin Chen [27], Juan Pablo Couso [28], Eric W Deutsch [29], Paul Flicek [7], Adam Frankish [7], Mark Gerstein [13,30,31,32], Norbert Hubner [12,33,34], Nicholas T Ingolia [35], Manolis Kellis [8,9], Gerben Menschaert [36], Robert L Moritz [29], Uwe Ohler [37,38,39], Xavier Roucou [40], Alan Saghatelian [10], Jonathan S Weissman [41,42,43], Sebastiaan van Heesch [#44]

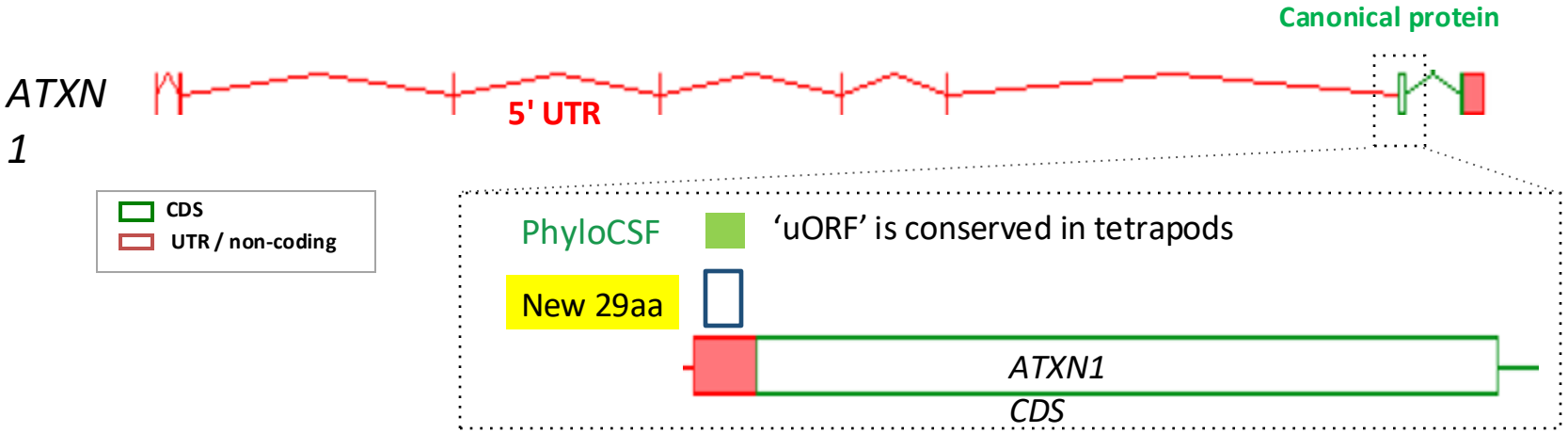
Affiliations + expand

PMID: 35831657 PMCID: PMC9757701 DOI: 10.1038/s41587-022-01369-0

6,885 Ribo-seq ORFs are <100aa (95%)
'microtranslations'

These are *NOT* annotated as proteins, they're just 'Ribo-seq ORFs'

GENCODE microprotein discovery efforts: Ribo-seq



Concurrent transcript annotation tries to infer how such proteins are expressed

- some uORFs are differentially transcribed compared with the canonical CDS
- some, like this case, seem to be part of the same transcript structure

Too many transcripts?

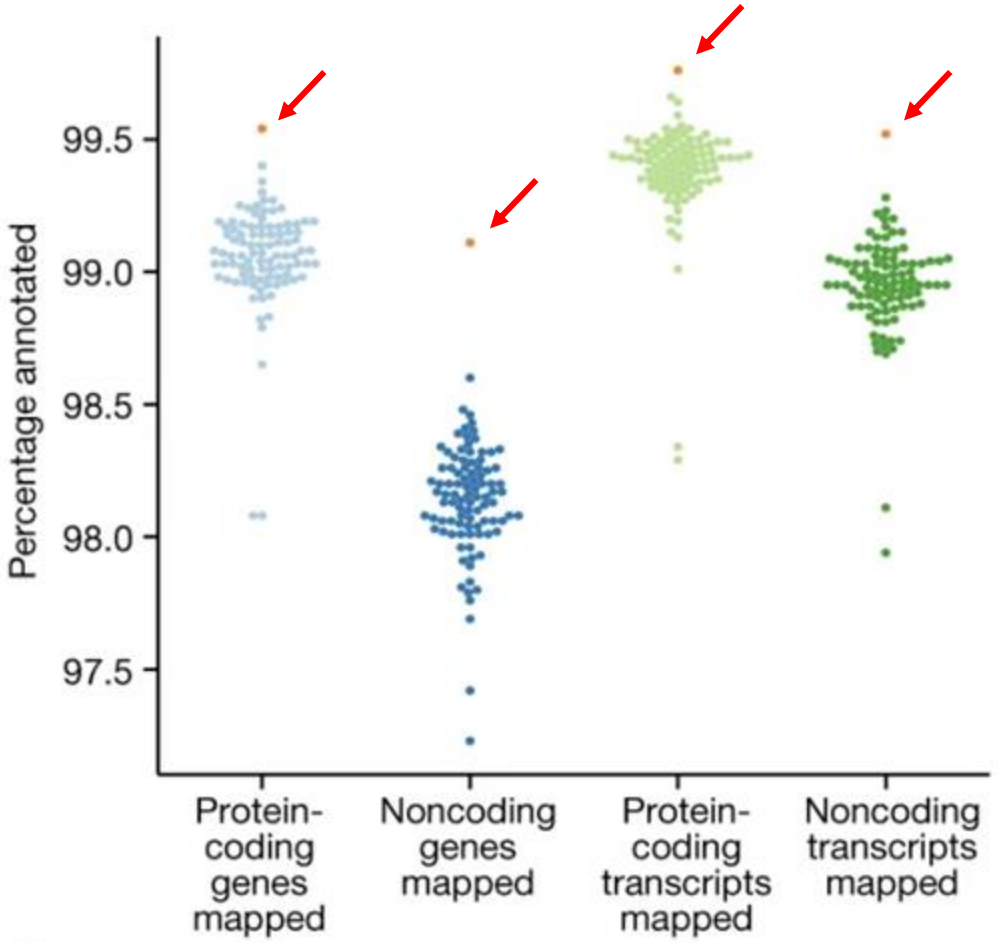
Have we found all the genes?

What about all the genomes?


Earth's heart of iron begins to yield its secrets p. 18
Microglia in chronic pain recovery and relapse pp. 33 & 86
Particle acceleration in a nova explosion p. 77
Science
$15
1 APRIL 2022
SPECIAL ISSUE
science.org
AAAS
FILLING THE GAPS
Closing in on a complete human genome p. 42


The international journal of science / 11 May 2023
nature
HUMAN PANGENOME
Data from 47 individuals combine to create reference resource that reflects human diversity

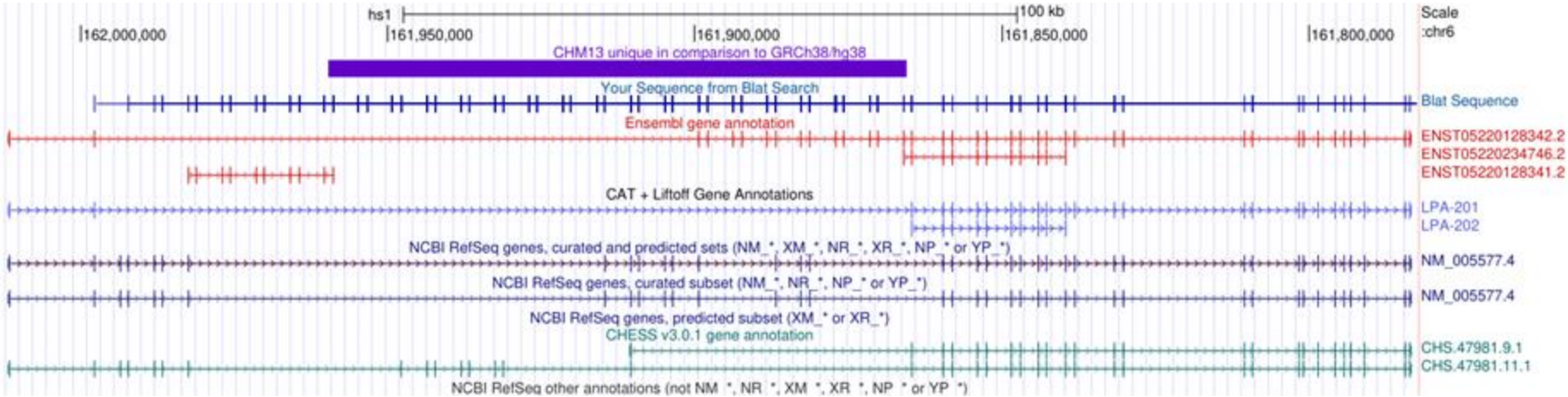
Mapping GENCODE annotation to HPRC Haplotypes



Liao, et al. 2023

Where do we need to improve - novel sequence?

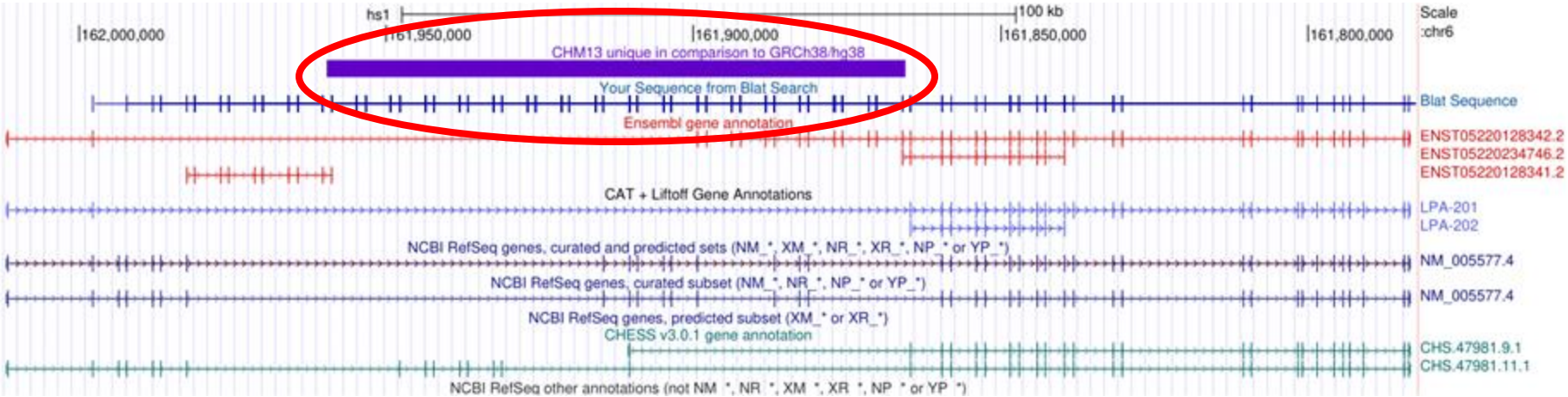
LPA on T2T:CHM13



EMBL-EBI

Where do we need to improve - novel sequence?

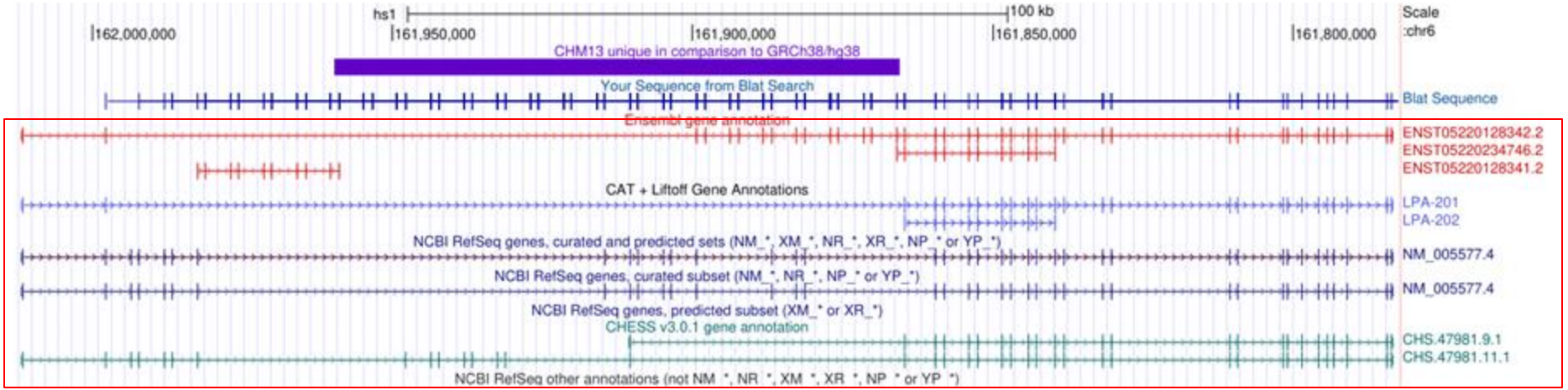
LPA on T2T:CHM13



EMBL-EBI

Where do we need to improve - novel sequence?

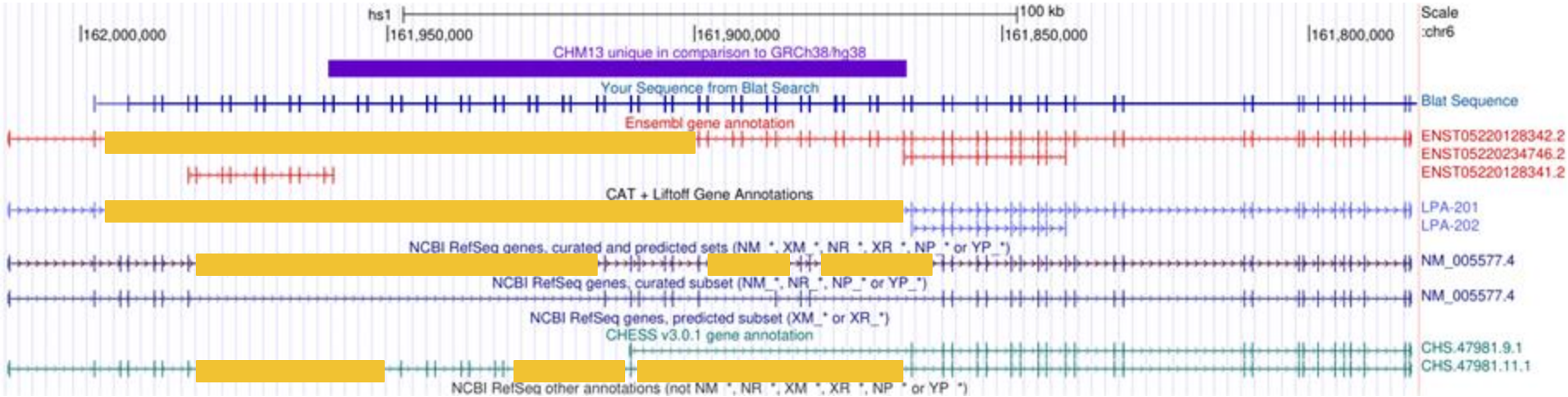
LPA on T2T:CHM13



EMBL-EBI

Where do we need to improve - novel sequence?

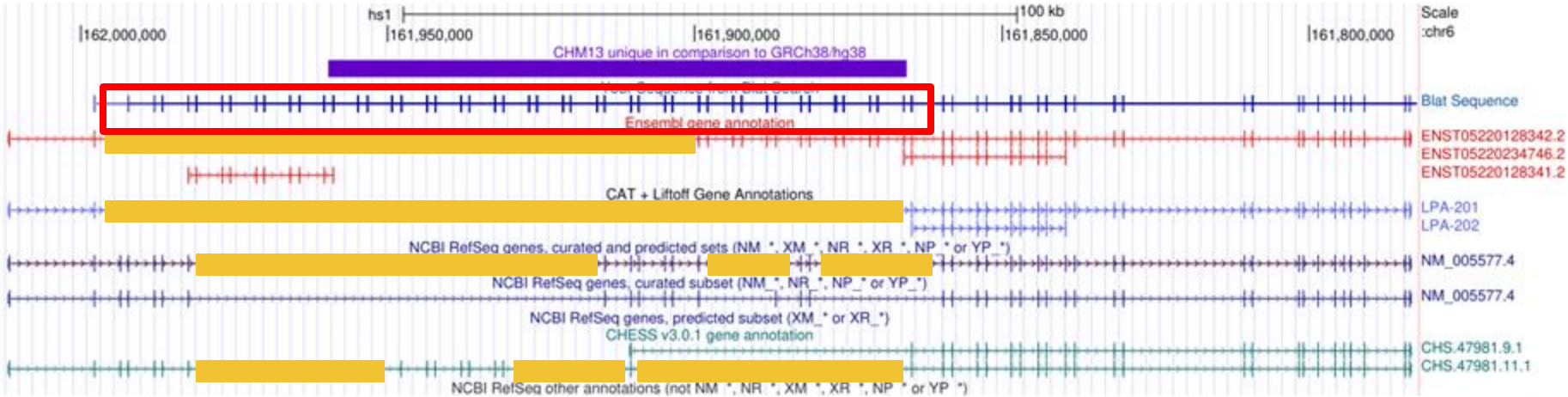
LPA on T2T:CHM13



EMBL-EBI

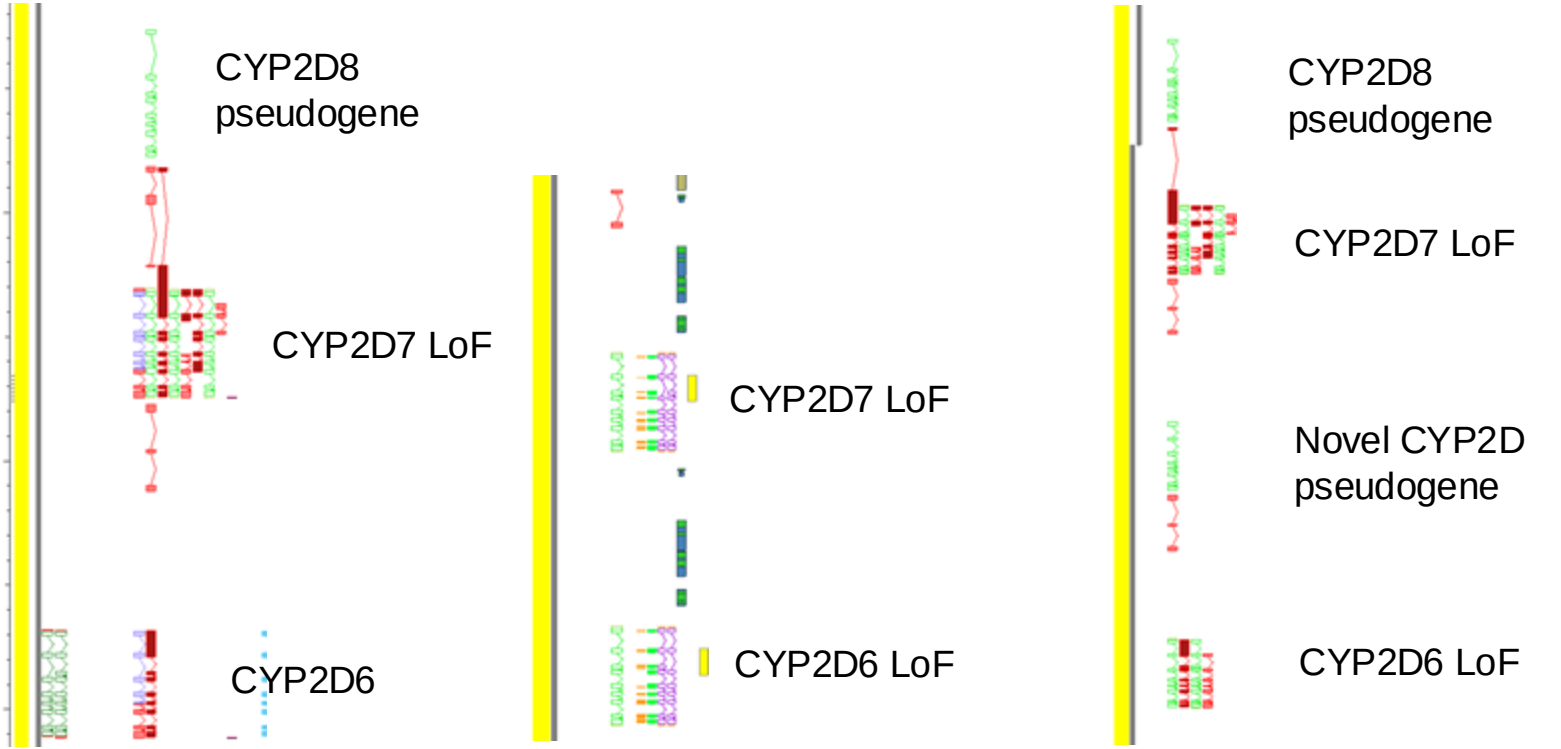
Where do we need to improve - novel sequence?

LPA on T2T:CHM13



EMBL-EBI

Gene clusters - CYP2D6



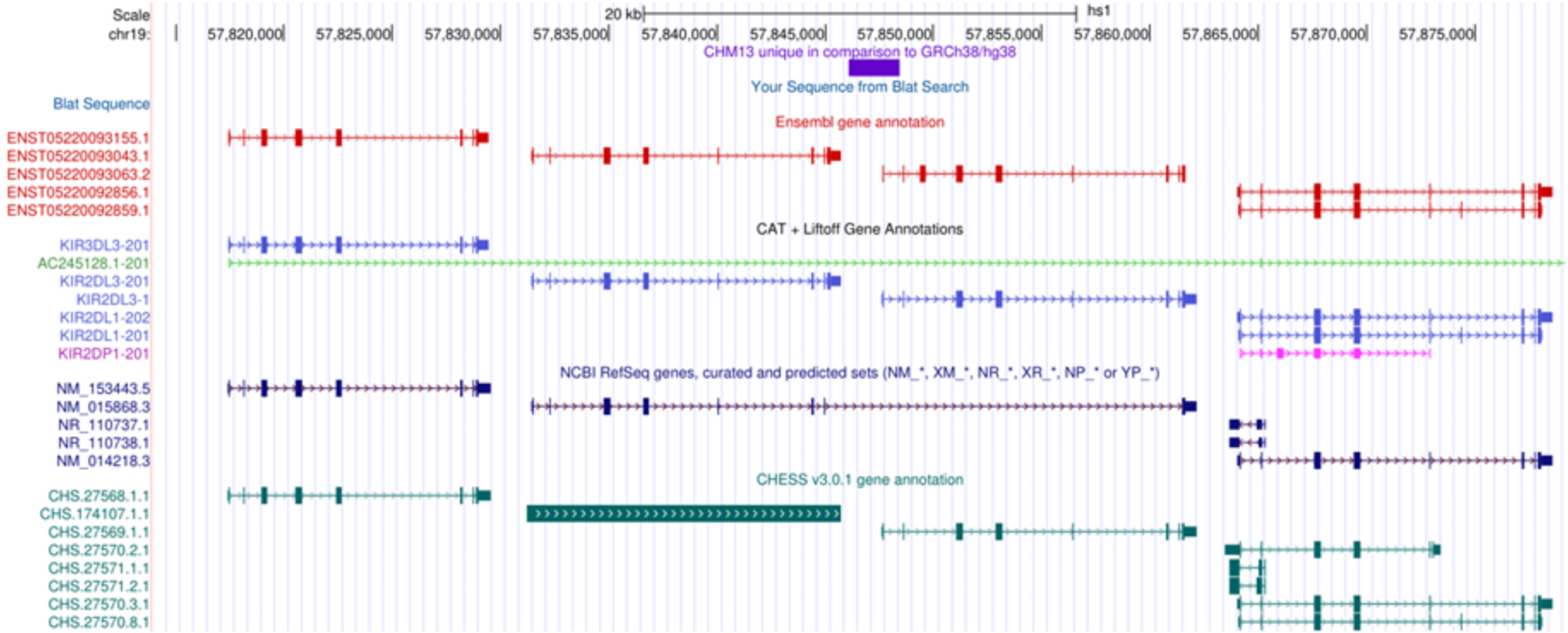
Reference - GRCh38

CHR_HSCHR22_2_CTG1

CHR_HSCHR22_8_CTG1

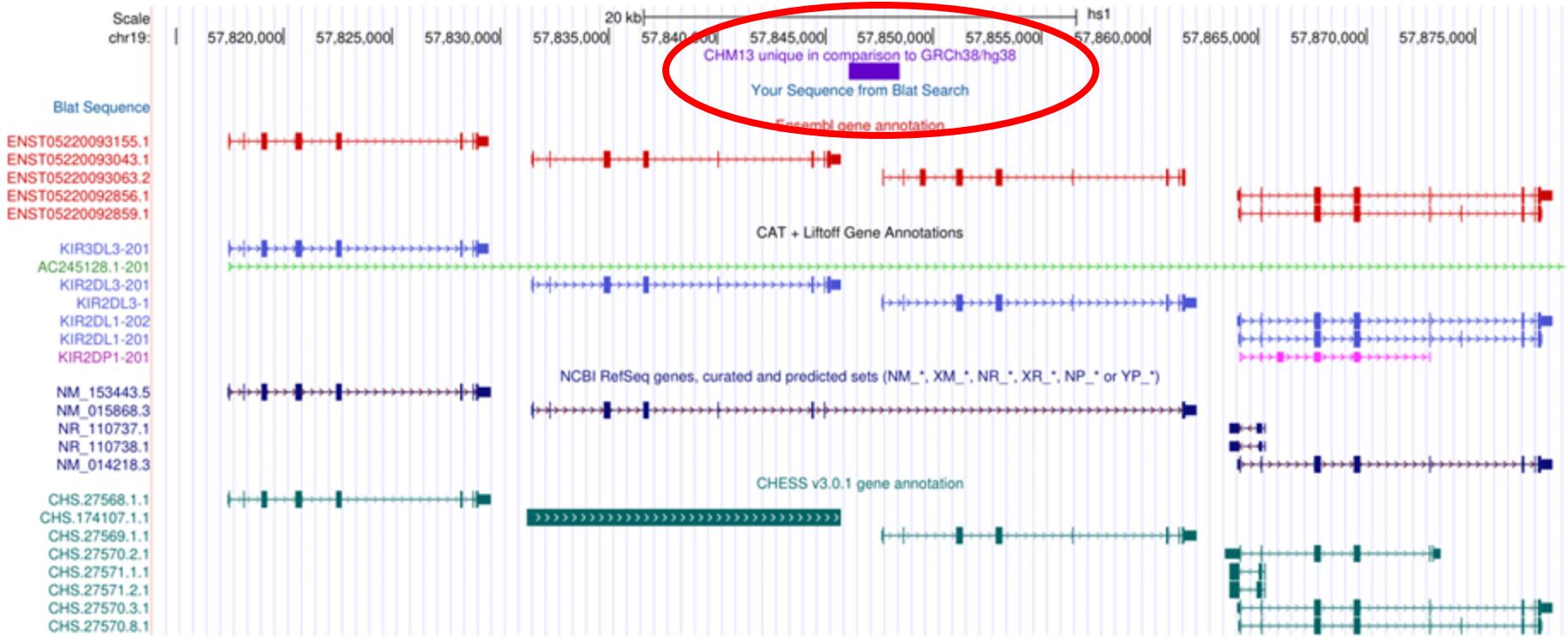
EMBL-EBI

Gene clusters - KIR - T2T-CHM13



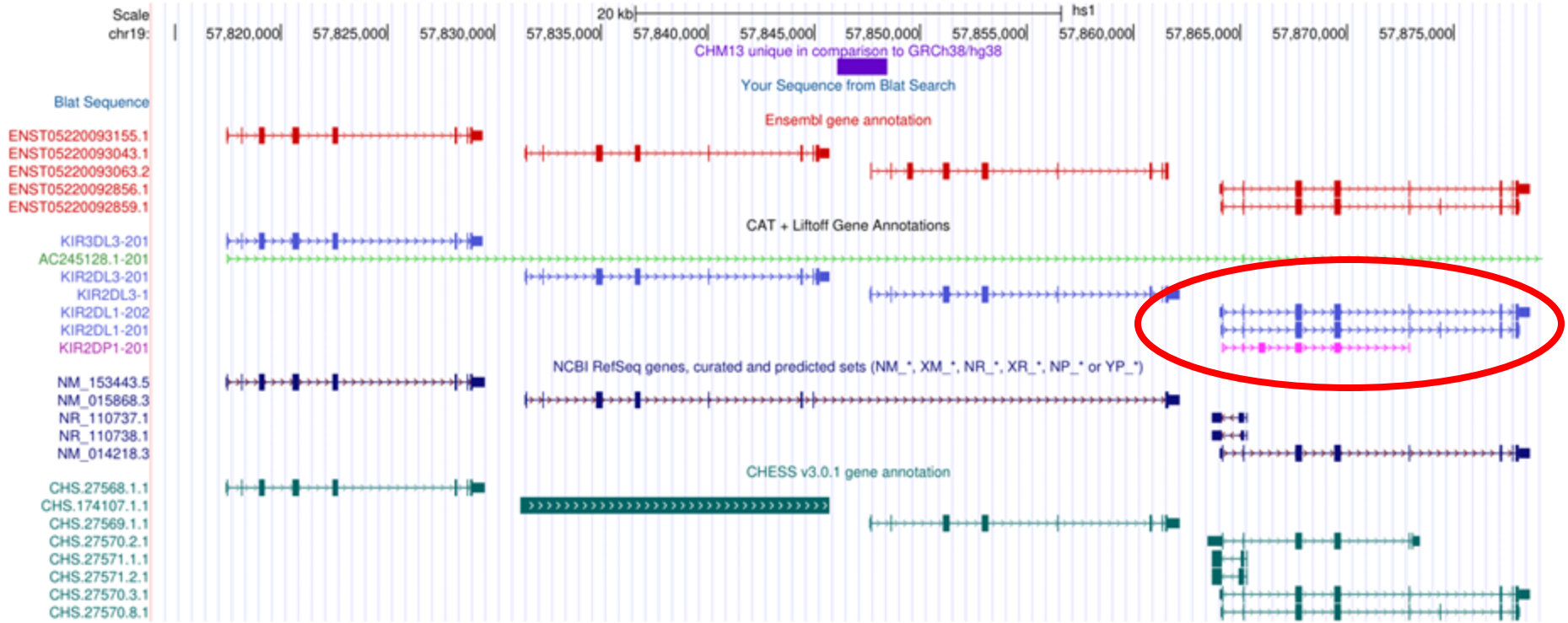
EMBL-EBI

Gene clusters - KIR - T2T-CHM13



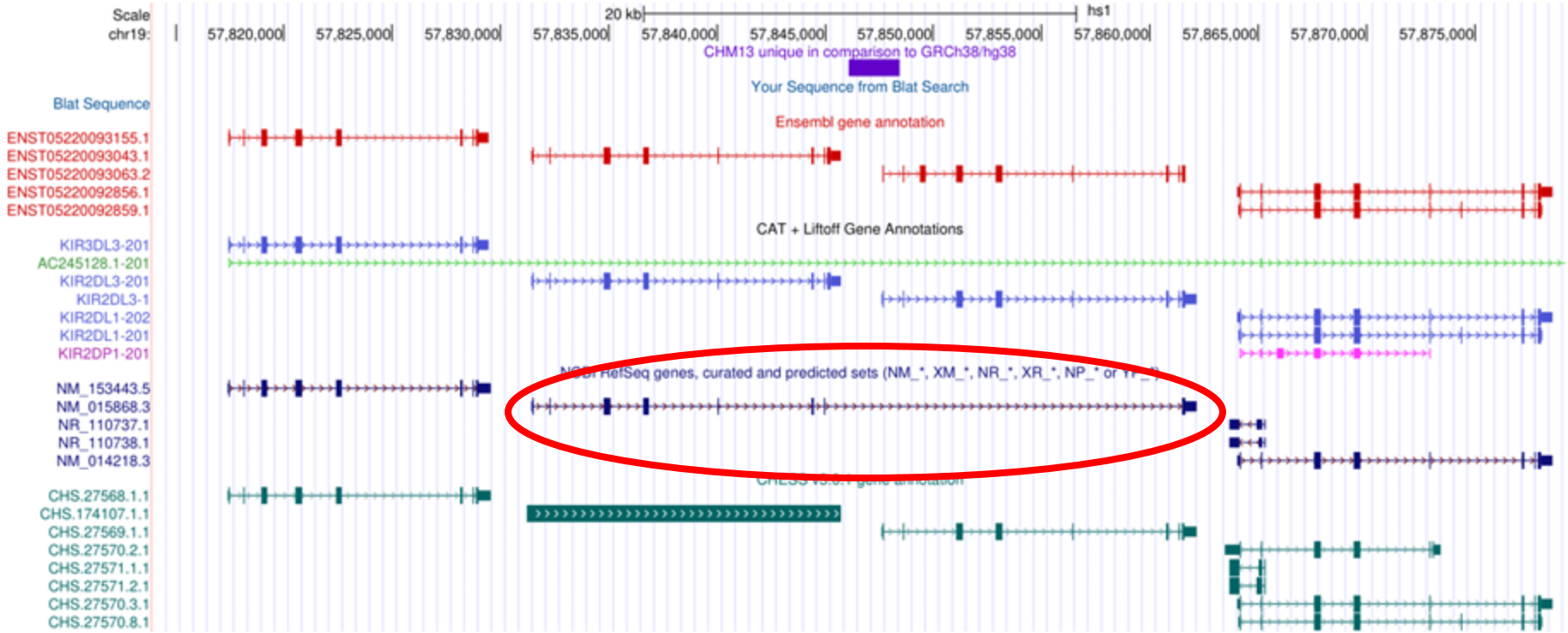
EMBL-EBI

Gene clusters - KIR - T2T-CHM13



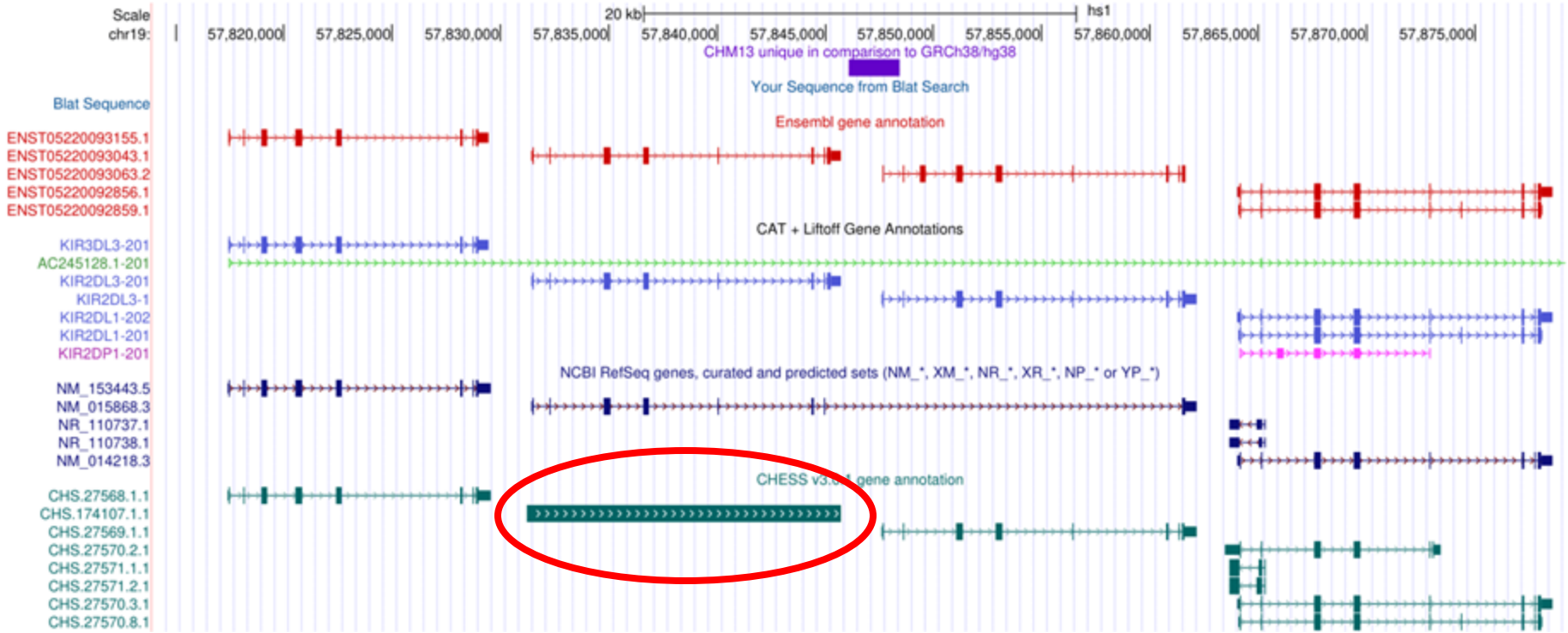
EMBL-EBI

Gene clusters - KIR - T2T-CHM13



EMBL-EBI

Gene clusters - KIR - T2T-CHM13



EMBL-EBI

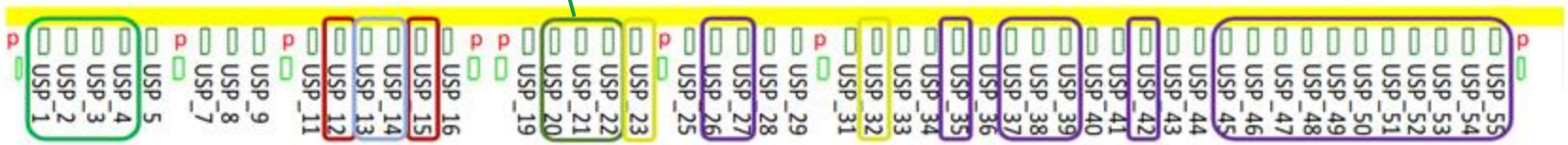
Clusters highlight another problem

GRCh38

Only 1 gene is 100% identical between the genomes
1:3 relationship

Almost all other GRCh38 – T2TCHM13 pairings have a 2+ amino acid difference

T2T CHM13



EMBL-EBI

Changing the concept of a gene

- Genes on reference treated as individual entities
  - Unique HGNC gene symbol
  - Systemic labelling in clusters
    - USP17L10, USP17L11, etc
- This does not scale across pangenome
- Root gene symbol + suffix
  - Capture information about haplotype, position, more?
  - Human readability vs computational utility
  - USP17L_******
  - Apply to reference genome as an alias?

The Forefront of Genomics

Questions & Answers #2

Acknowledgements

EMBL-EBI, Cambridge, UK

- Ewan Birney
- Fergal Martin
- Adam Frankish
- Jose Manuel Gonzalez
- Toby Hunt
- Jonathan Mudge
- Jane Loveland

Center for Genomic Regulation (CRG), Barcelona, Spain

- Roderic Guigó
- Carme Arnan
- Sílvia Carbonell-Sala
- Gazaldeep Kaur
- Tamara Perteghella
- Emilio Palumbo

Yale University, New Haven, USA

- Mark Gerstein
- Cagatay Dursun
- Pengyu Ni
- Cristina Sisu
- Dingyao Zhang
- Yunzhe Jiang

Spanish National Cancer Research Centre (CNIO), Madrid, Spain

- Michael Tress
- Daniel Cerdán-Vélez
- Miguel Maquedano

University of California, Santa Cruz, California, USA

- Benedict Paten
- Mark Diekhans

Stanford University, California, USA

- Anshul Bharat Kundaje
- Kelly Cochran

Massachusetts Institute of Technology (MIT), Boston, USA

- Manolis Kellis
- Irwin Jungreis

Gencode-CRG Collaborators and Alumni:

- *Rory Johnson (University College Dublin)*
- *Julien Lagarde (Flomics, Barcelona)*
- *Andrea Tanzer (University of Vienna)*
- *Barbara Uszczynska (IBCH PAS, Poznan)*
- *Hiromi Nishiyori and Piero Carninci (Riken Institute)*

EMBL-EBI

Acknowledgements

Ensembl-HAVANA
If Barnes
Ruth Bennett
Andrew Berry
Claire Davidson
Sarah Donaldson
Reham Fatima
Jose Gonzalez
Matt Hardy
Zoe Hollis
Toby Hunt
Mike Kay
Jane Loveland
Ryan Merritt
Jonathan Mudge
Marie-Marthe Suner
Lucas Tidmarsh Cortes

RefSeq
Terence Murphy
Shashi Pujar
Eric Cox
Catherine Farrell
Tamara Goldfarb
John Jackson
Vinita Joardar
Kelly McGarvey
Michael Murphy
Nuala O'Leary
Bhanu Rajput
Sanjida Rangwala
Lillian Riddick
David Webb
Alex Astashyn
Olga Ermolaeva
Vamsi Kodali
Craig Wallin

Pangenome Annotation
Ewan Birney
Elspeth Bruford
Cagatay Dursun
Rob Finn
Erik Garrison
Mark Gerstein
Roderic Guigo
Toby Hunt
Irwin Jungreis
Manolis Kellis
Anshul Kundaje
Jane Loveland
Fergal Martin
Jonathan Mudge
Terence Murphy
Benedict Paten
Adam Phillippy
Heidi Rehm
Arang Rhie
Michael Tress
Joel Rozowsky
Cristina Sisu
Ting Wang

EMBL-EBI

2022 Apr;604(7905):310-315. doi: 10.1038/s41586-022-04558-8. Epub 2022 Apr 6.

A joint NCBI and EMBL-EBI transcript set for clinical genomics and research

Joannella Morales #1, Shashikant Pujar #2, Jane E Loveland1, Alex Astashyn2, Ruth Bennett1, Andrew Berry1, Eric Cox2, Claire Davidson1, Olga Ermolaeva2, Catherine M Farrell2, Reham Fatima1, Laurent Gil1, Tamara Goldfarb2, Jose M Gonzalez1, Diana Haddad2, Matthew Hardy1, Toby Hunt1, John Jackson2, Vinita S Joardar2, Michael Kay1, Vamsi K Kodali2, Kelly M McGarvey2, Aoife McMahon1, Jonathan M Mudge1, Daniel N Murphy1, Michael R Murphy2, Bhanu Rajput2, Sanjida H Rangwala2, Lillian D Riddick2, Françoise Thibaud-Nissen2, Glen Threadgold1, Anjana R Vatsan2, Craig Wallin2, David Webb2, Paul Flicek1, Ewan Birney1, Kim D Pruitt2, Adam Frankish1, Fiona Cunningham1, Terence D Murphy3
Affiliations Expand

PMID: 35388217

GENCODE: reference annotation for the human and mouse genomes in **2023**.

Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I, Banerjee A, Bennett R, Berry A, Bignell A, Boix C, Calvet F, Cerdán-Vélez D, Cunningham F, Davidson C, Donaldson S, Dursun C, Fatima R, Giorgetti S, Giron CG, Gonzalez JM, Hardy M, Harrison PW, Hourlier T, Hollis Z, Hunt T, James B, Jiang Y, Johnson R, Kay M, Lagarde J, Martin FJ, Gómez LM, Nair S, Ni P, Pozo F, Ramalingam V, Ruffier M, Schmitt BM, Schreiber JM, Steed E, Suner MM, Sumathipala D, Sycheva I, Uszczynska-Ratajczak B, Wass E, Yang YT, Yates A, Zafrulla Z, Choudhary JS, Gerstein M, Guigo R, Hubbard TJP, Kellis M, Kundaje A, Paten B, Tress ML, Flicek P.

Nucleic Acids Res. 2023 Jan 6;51(D1):D942-D949. doi: 10.1093/nar/gkac1071.
PMID: 36420896

Mol Cell Proteomics
. 2023 Sep;22(9):100631. doi: 10.1016/j.mcpro.2023.100631. Epub 2023 Aug 11.

What Can Ribo-Seq, Immunopeptidomics, and Proteomics Tell Us About the Noncanonical Proteome?

John R Prensner 1, Jennifer G Abelin 2, Leron W Kok 3, Karl R Clauser 2, Jonathan M Mudge 4, Jorge Ruiz-Orera 5, Michal Bassani-Sternberg 6, Robert L Moritz 7, Eric W Deutsch 7, Sebastiaan van Heesch

PMID: 37572790

The Forefront of Genomics®

Thank you for attending!


The Forefront
of Genomics®