1. Ensembl / GENCODE basics

Training materials

- Ensembl training materials are protected by a CC BY license:

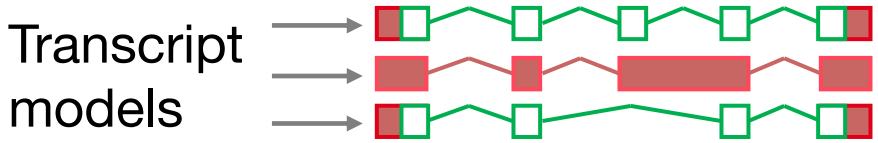
  creativecommons.org/licenses/by/4.0/

- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers:

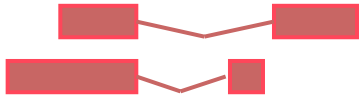
  ensembl.org/info/about/publications.html

GENCODE / Ensembl human gene annotation

60,651 genes in human v37, 234,485 transcript models



19,951 protein-coding genes

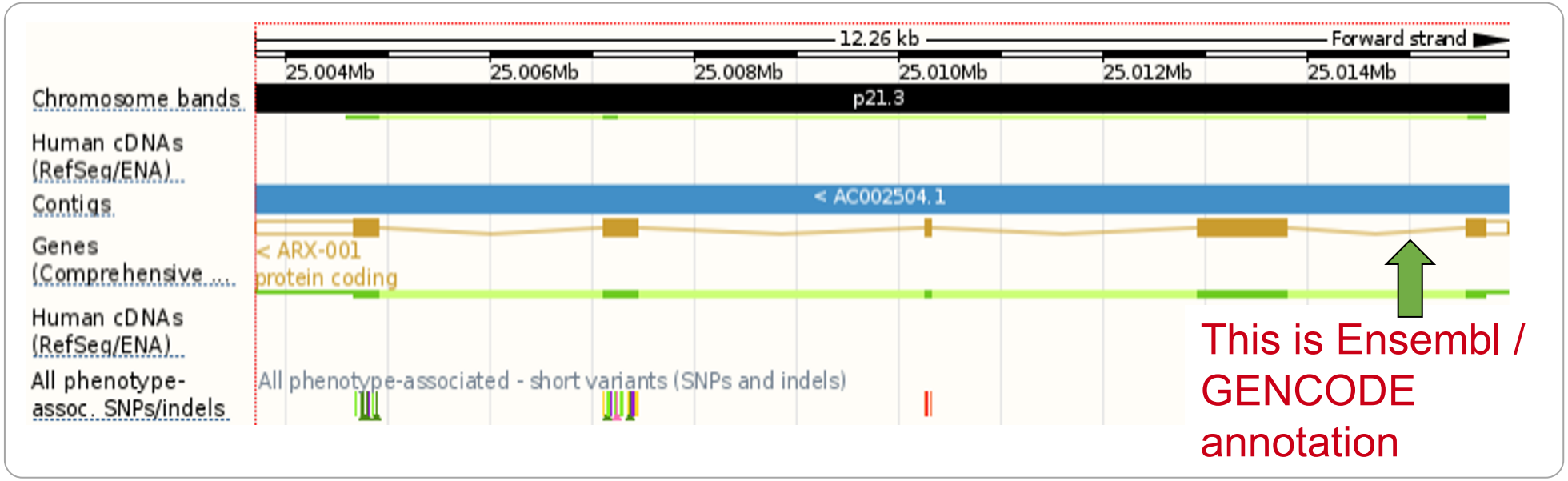


17,948 long non-coding RNA genes

… also pseudogenes and small RNAs

The GENCODE <–> Ensembl relationship

GENCODE human annotation = Ensembl human annotation



GENCODE also annotate mouse

Data access via Ensembl website

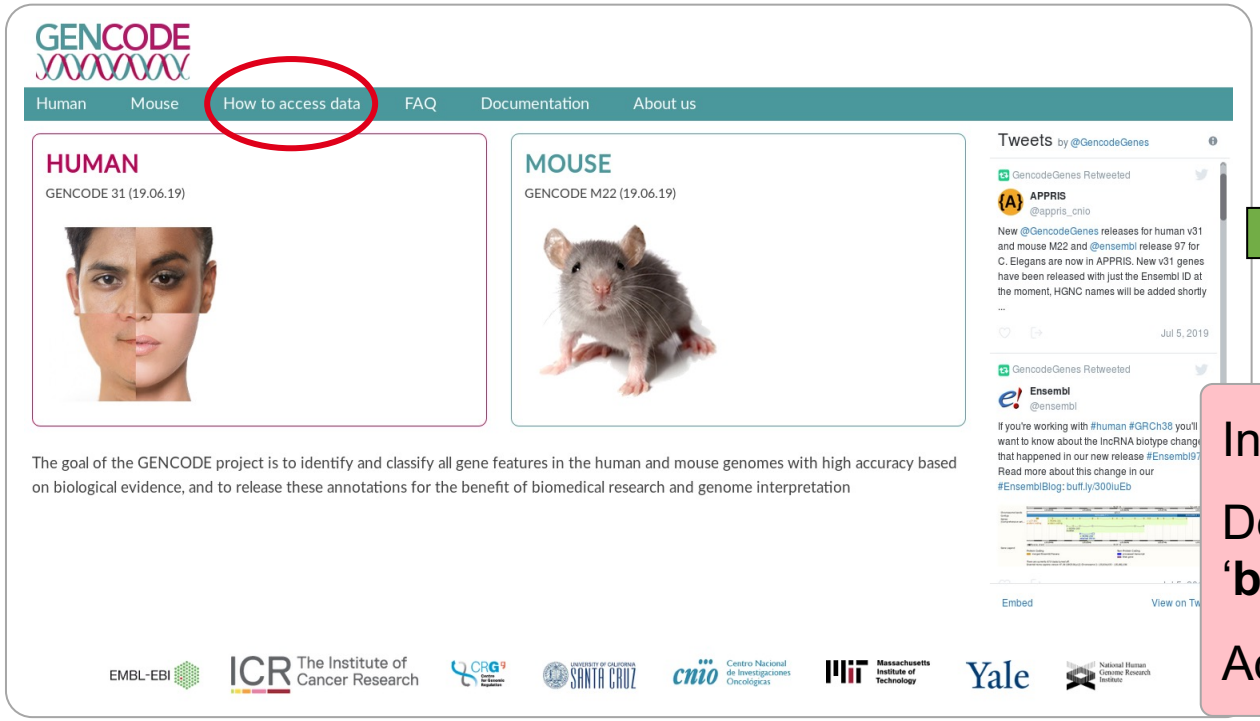
- BioMart data retrieval system
- FTP download site
- MySQL public server
- Perl interface
- Access to REST server

Genome browser

Extensive documentation available
… including video tutorials

www.ensembl.org

Data access via GENCODE web portal



ftp site for GTF / GFF3

Info. on content and format

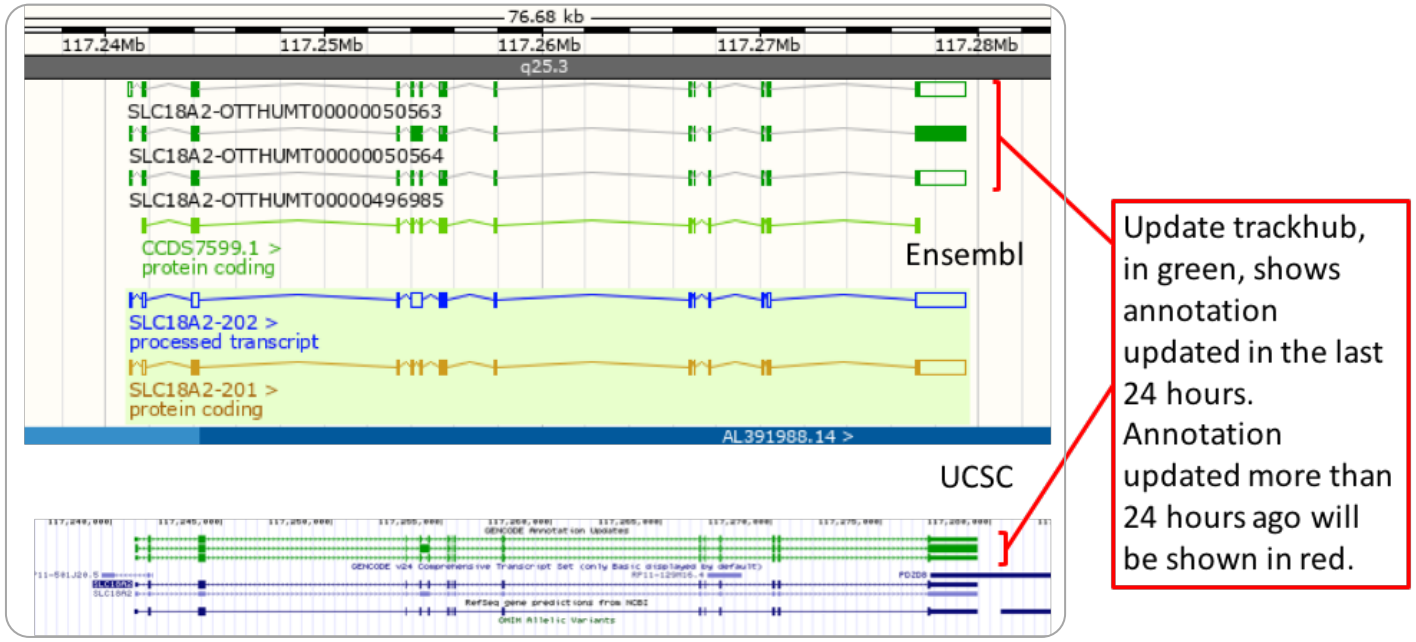
Descriptions of gene and transcript '**biotypes**'

Additional fixed-vocabulary '**attributes**'

www.gencodegenes.org

We provide release every three months

… we also provide an ‘update’ Ensembl and UCSC trackhub



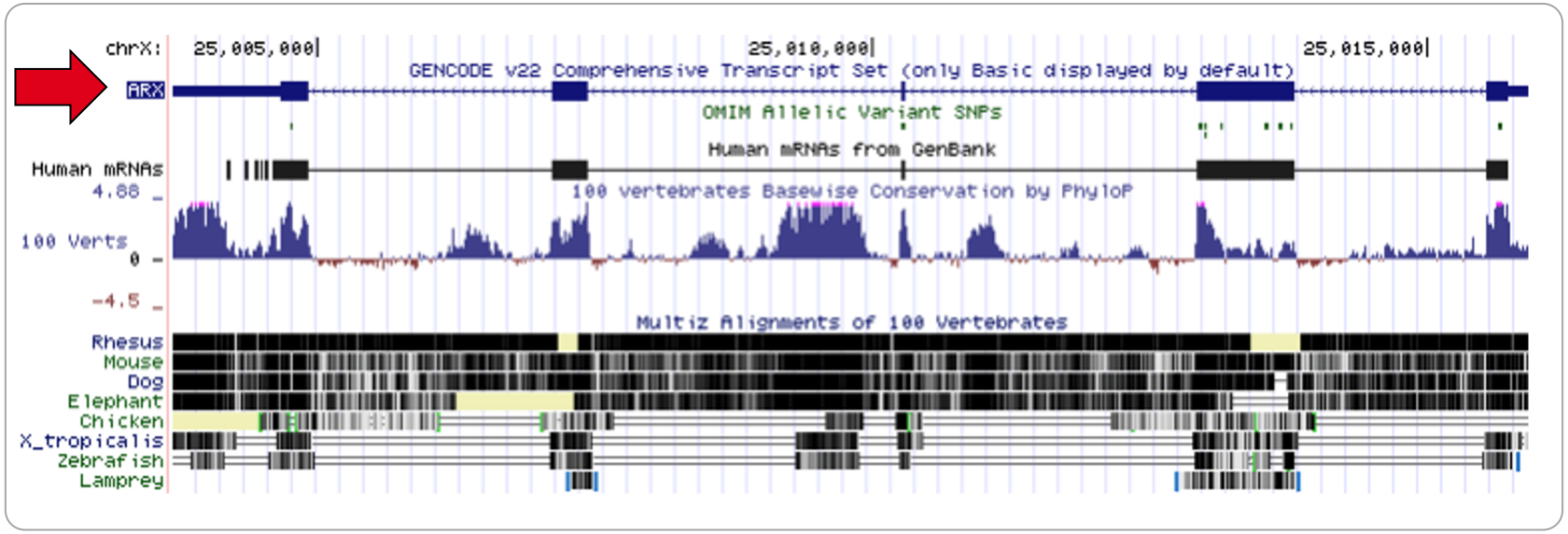
Connect to: http://ftp.ebi.ac.uk/pub/databases/gencode/update_trackhub/hub.txt

… or search the ‘trackhub registry’

EMBL

GENCODE is the default UCSC human gene annotation

UCSC Genome Bioinformatics



Currently UCSC update for every other release

EMBL

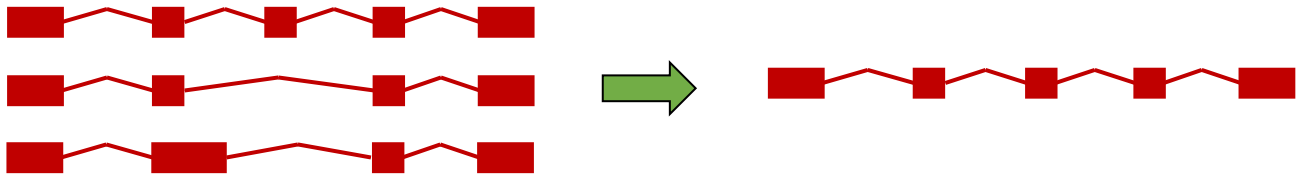
We offer ‘filtered’ transcript sets

‘**Comprehensive**’: the complete set of GENCODE annotations

‘**Basic**’: a smaller set based on filtering

Protein-coding genes contain only models with full-length CDS

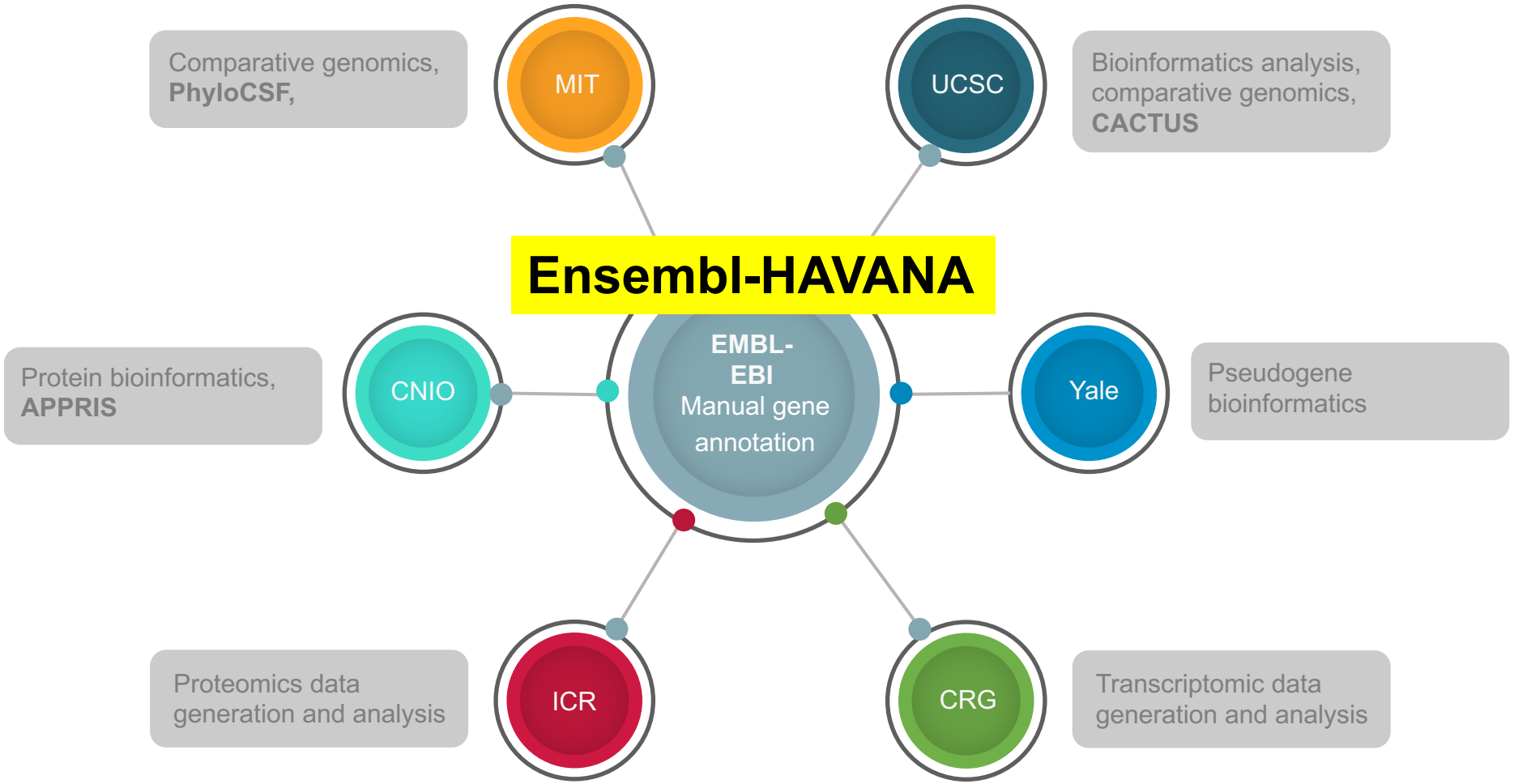
lncRNA genes contain minimum number of transcripts that provide 80% of the exonic coverage



Now: working on transcript choice with RefSeq as **MANE project**

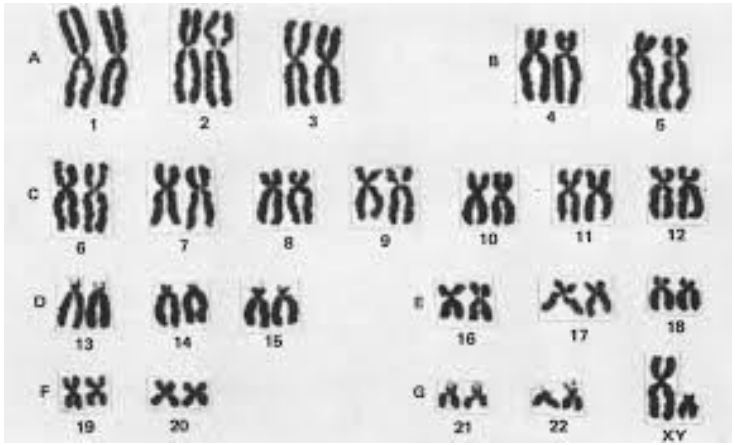
2. Producing Ensembl / GENCODE annotation

GENCODE is a multifaceted consortium



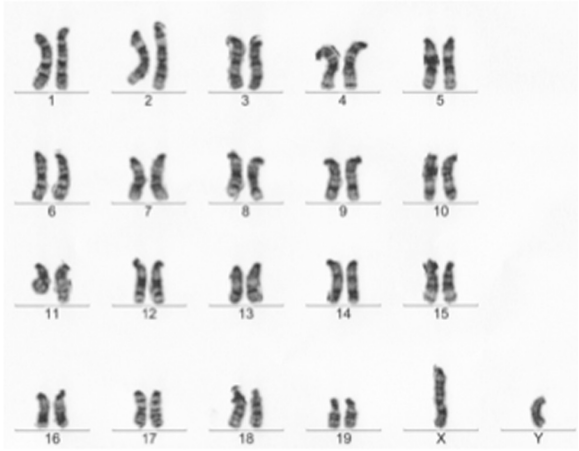
The ‘first pass’ annotation was completed

Human



‘finished’ ~2015

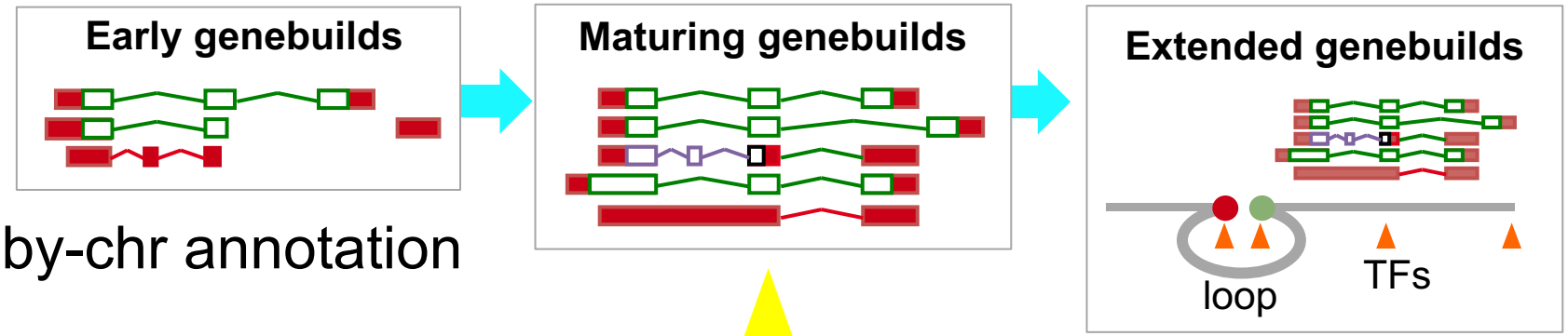
Mouse



‘finished’ ~2018

i.e. systematic chromosome-by-chromosome annotation

Gene annotation will continue for years



chr-by-chr annotation

e.g. linking enhancers to genes

Massive amounts of new data

Discovery of new features

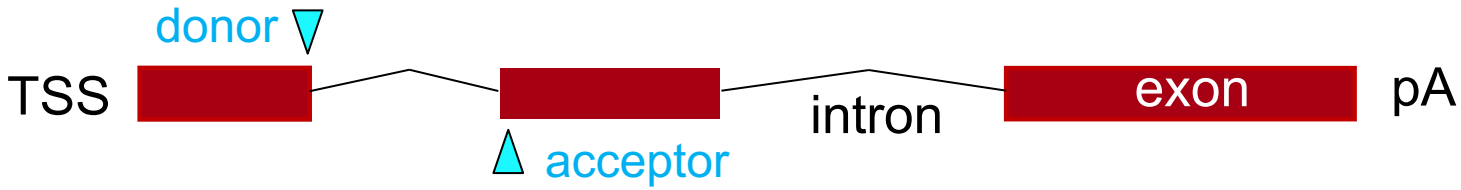
Extending incomplete models

Correct ‘functional’ annotation

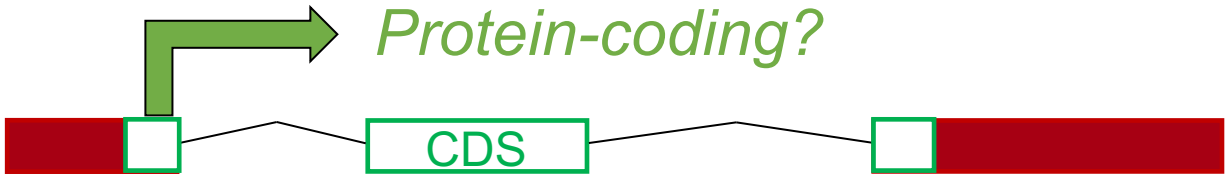
Manual annotation and analysis

The two principles of annotation

1. '**Structural**': making the exons and introns of the model



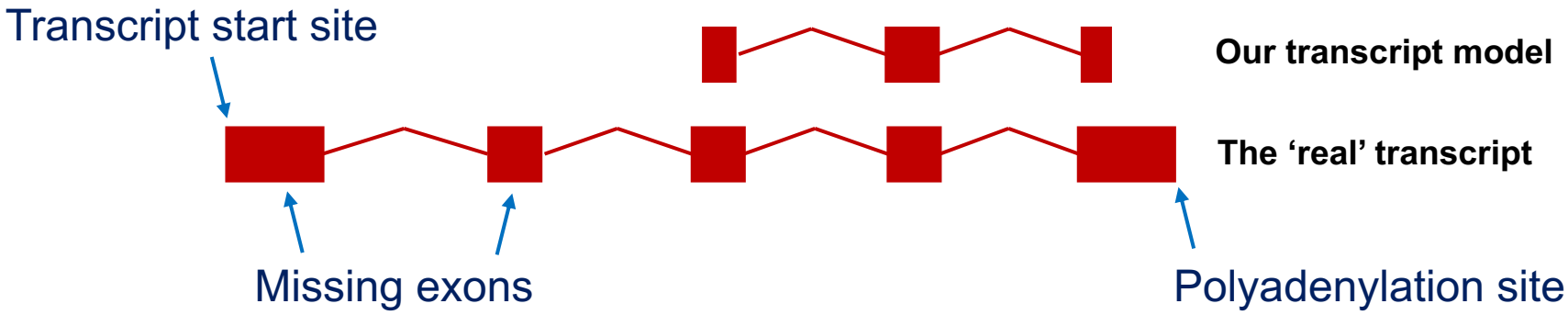
2. '**Functional**': assessing the biological nature of the model



Challenges in structural annotation

Transcript models are commonly ‘incomplete’, i.e. too short

… because the RNA evidence used to construct was not ‘full-length’

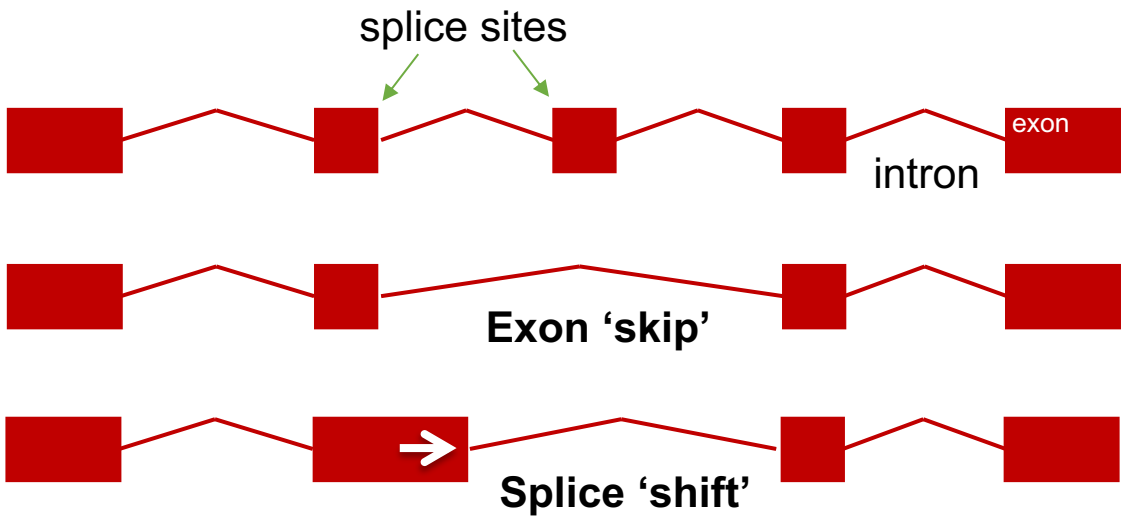


Incomplete models can lack correct biological features (e.g. CDS)

Challenges in structural annotation

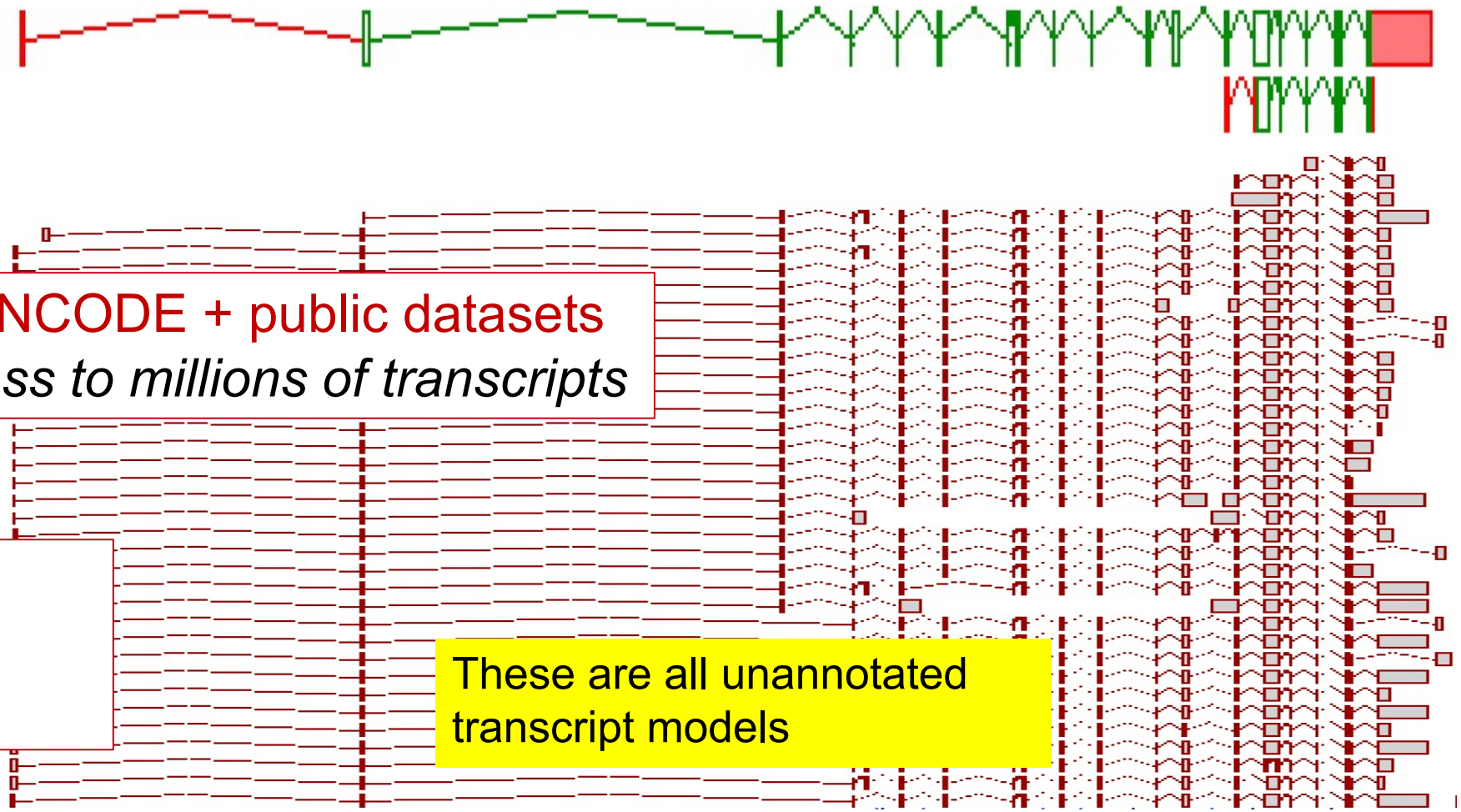
We can find novel transcripts within existing genes

i.e. due to **alternative splicing**



New models may contain additional biological features

Long-read sequencing data at *IL16*



PacBio data from GENCODE + public datasets

… *we now have access to millions of transcripts*

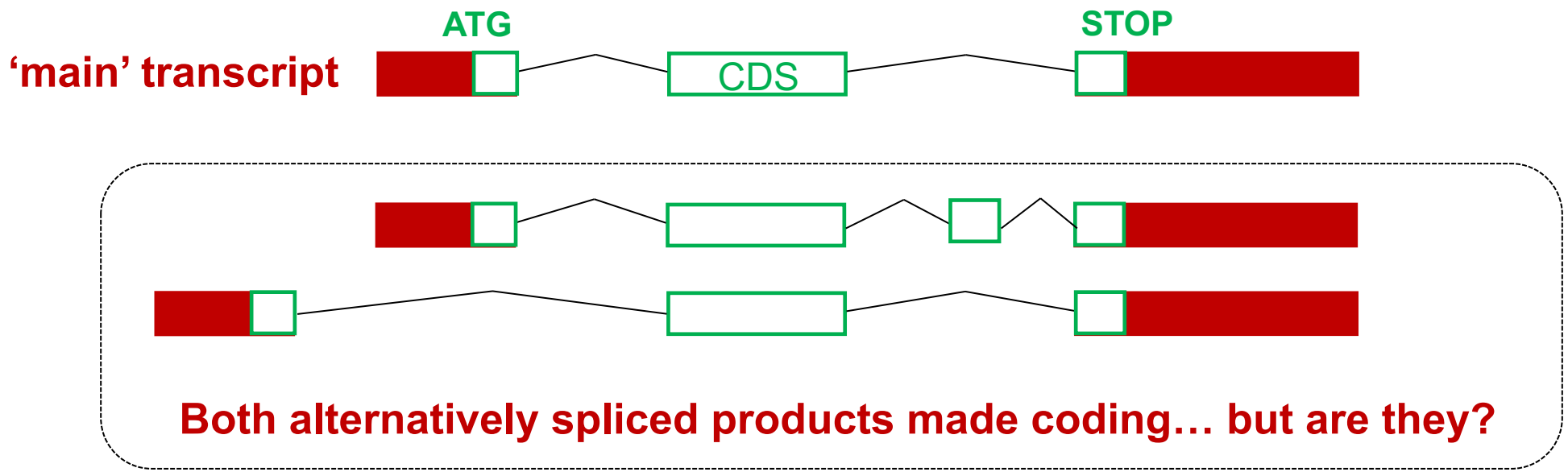
‘**TAGENE**’ pipeline for computational incorporation

These are all unannotated transcript models

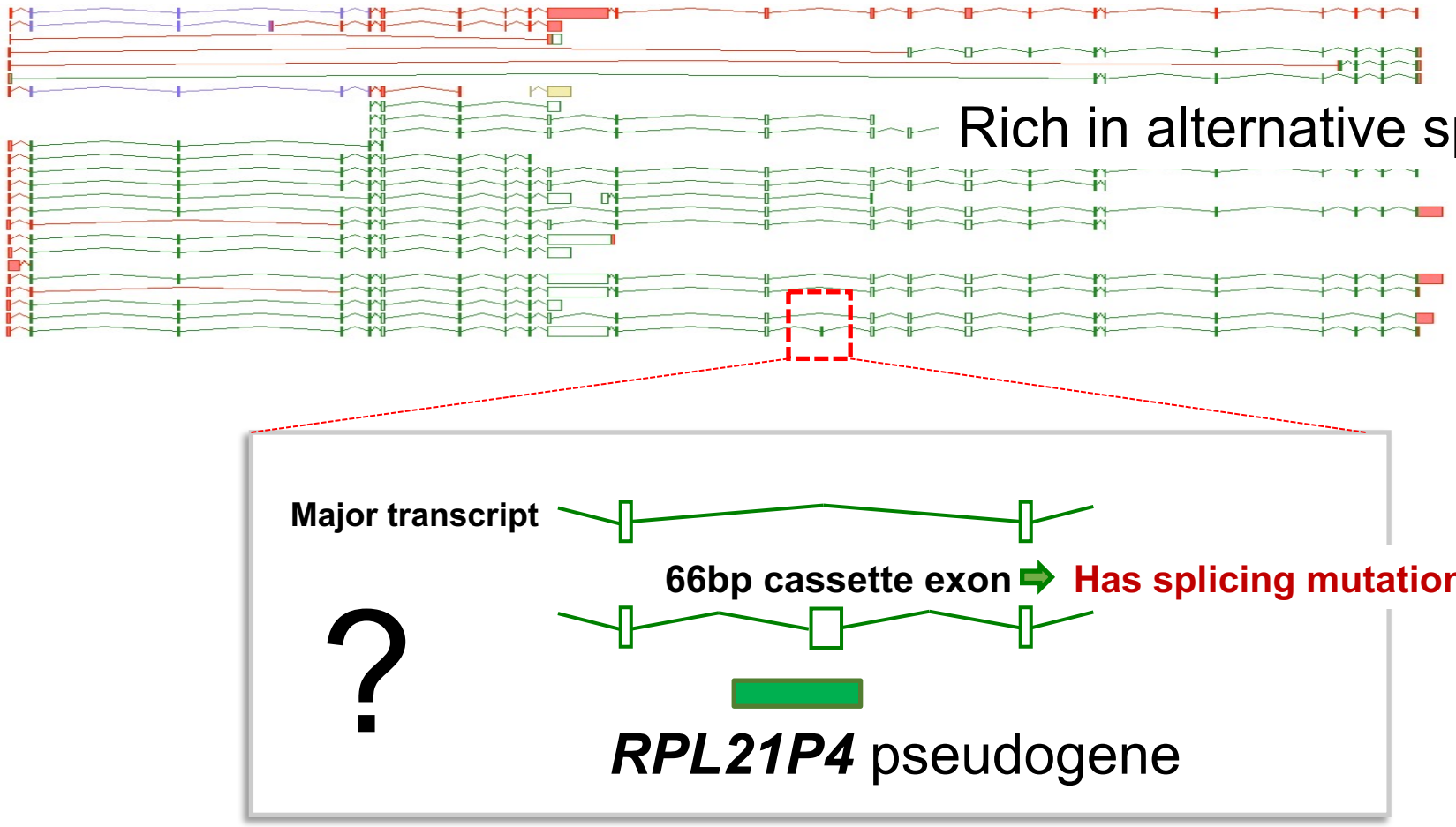
‘Functional annotation’

Major question: which transcripts encode proteins?

Traditionally: CDS annotation is predictive, based on first principles



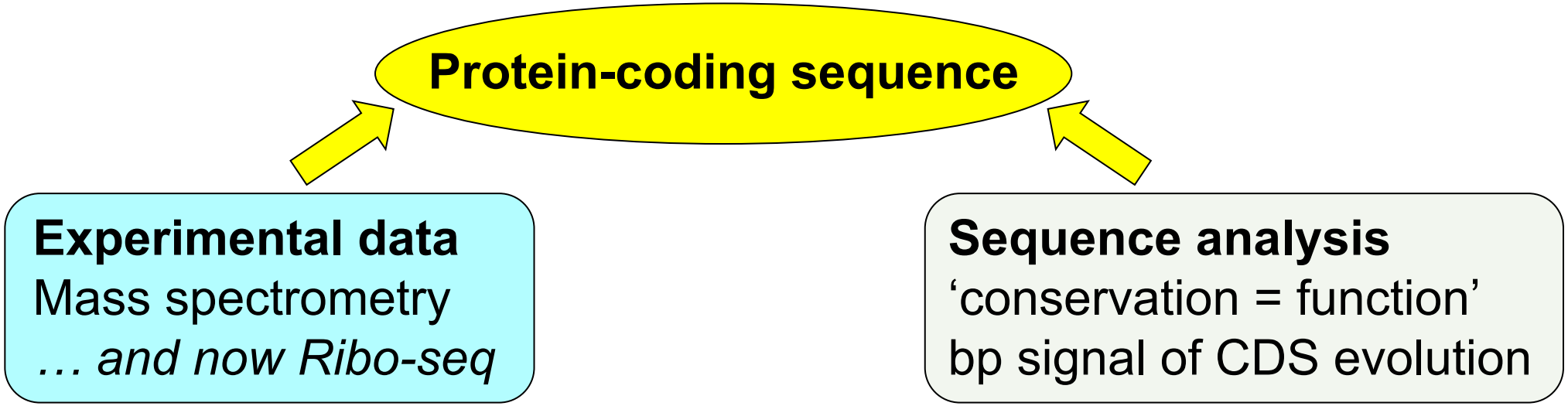
BRCA1 putative coding exon



Confirming protein-coding transcripts

Fundamental problem: sequencing protein remains very challenging

… protein-coding annotation is highly predictive



Both used for **discovery** as well as **validation**

Inverted formin 2 (INF2)

Existing model

PhyloCSF: identifies protein-coding evolution

new

A gene linked to focal segmental glomerulosclerosis

Structural and **functional** annotation combine to identify a new biological feature of presumed importance

3. Ensembl / GENCODE annotation strategy

The ‘smart’ deployment of expert manual annotation

Genome-wide improvements

- E.g. integrating massive new datasets
  i.e. 1,000,000s of long reads, Ribo-seq
- New paradigm: manual ‘**curation**’ of computational models

Q: How much manual annotation?

‘Deep dive’ on a single gene

- Striving to ‘perfect’ annotation
- Identifying additional features
- In depth functional assessment

Q: Which genes to annotate?

Disease genes

Targeted expert annotation

‘Re-annotating’ disease-linked genes

‘**Initiate**’ is being done in collaboration with clinical partners

Major focus on brain disorders, especially epilepsy

204 genes re-annotated

- 1,616 new transcript models, 568 of them coding
- Averages of 7.9 / 2.8 per gene
- **122 novel coding features**

Reannotation of human genes linked to COVID

We will improve the annotation of **any** gene linked to COVID-19

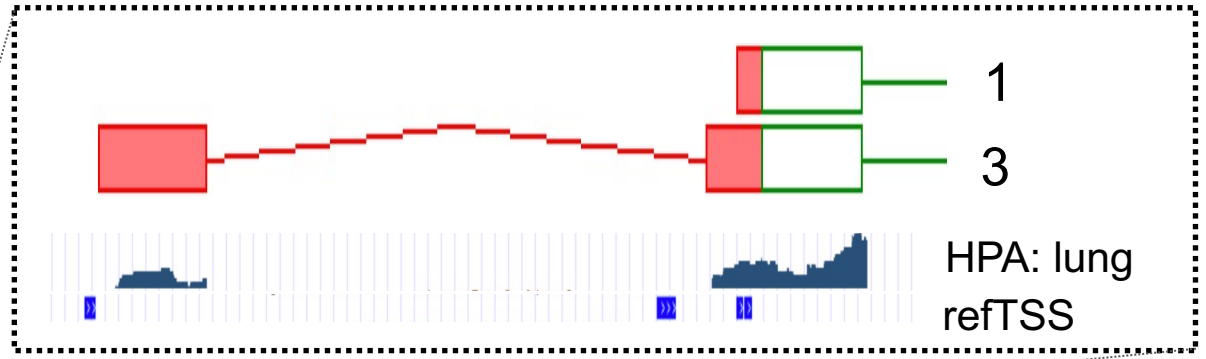
We will continue for as long as necessary

The current list is 556 genes, including:

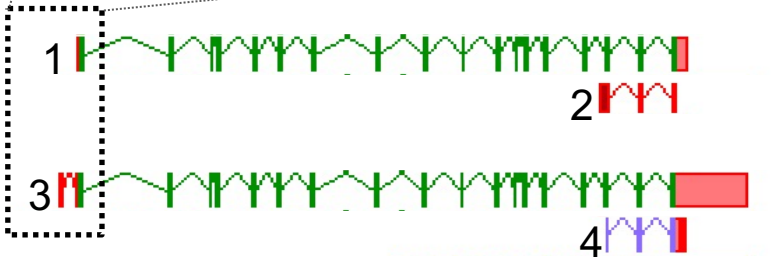
- 118 from Zhou *et al* (PMID:32194980)
- 371 from Gordon *et al* (PMID:32353859)
- More genes from collaborative projects, e.g. UniProt and Cell Atlas
- Genes from a GWAS (e.g. *LDLRAD4*)

Nearly 200 genes now reannotated: 1,300 models added and 300 updated

Updated *ACE2* annotation

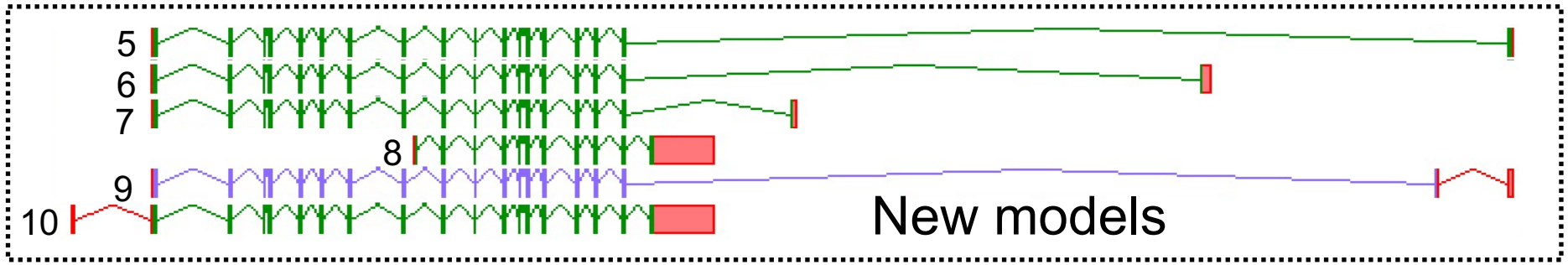


ACE2



Unchanged models

Modified models



We’re targeting unannotated PhyloCSF regions

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

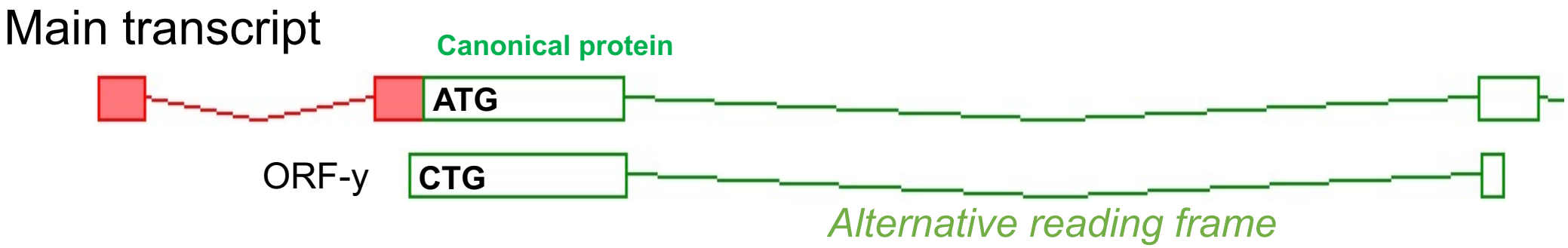
Jonathan M Mudge [# 1], Irwin Jungreis [# 2 3], Toby Hunt [1], Jose Manuel Gonzalez [1], James C Wright [4], Mike Kay [1], Claire Davidson [1], Stephen Fitzgerald [5], Ruth Seal [1 6], Susan Tweedie [1], Liang He [2 3], Robert M Waterhouse [7 8], Yue Li [2 3], Elspeth Bruford [1 6], Jyoti S Choudhary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

Added 144 deeply conserved human protein-coding genes

Added new coding exons to 236 known protein-coding genes

This is an ongoing project

A novel dual frame translation in *POLG*



Research article | Open Access | Published: 06 March 2020

Evidence for a novel overlapping coding sequence in *POLG* initiated at a CUG start codon

Yousuf A. Khan, Irwin Jungreis, James C. Wright, Jonathan M. Mudge, Jyoti S. Choudhary, Andrew E. Firth & Manolis Kellis

BMC Genetics **21**, Article number: 25 (2020) | Cite this article

47k Accesses | **2** Citations | **47** Altmetric | Metrics

ORF-y found by high PhyloCSF score

… **it's a bona fide protein**

41 ClinVar variants that don't change the canonical protein are non-syn in ORF-y

EMBL