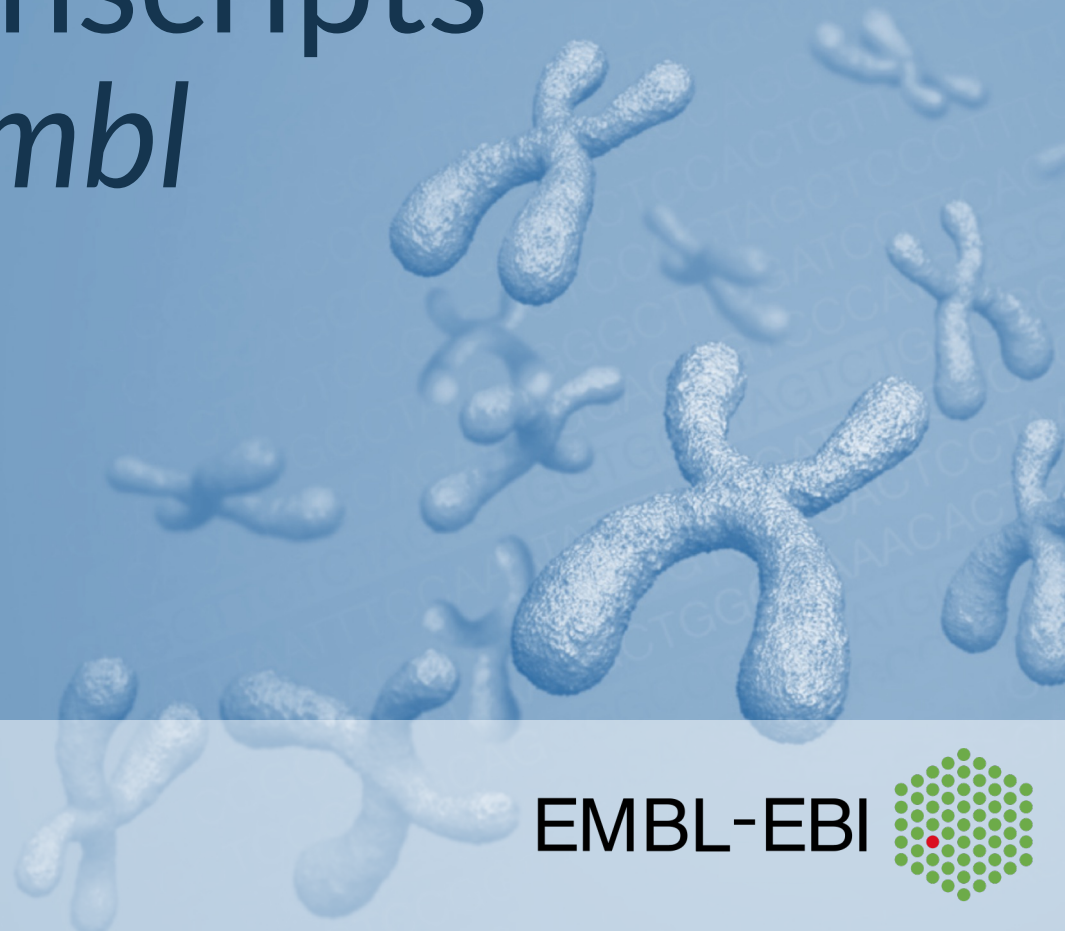
Genes and Transcripts in e!Ensembl



EMBL-EBI

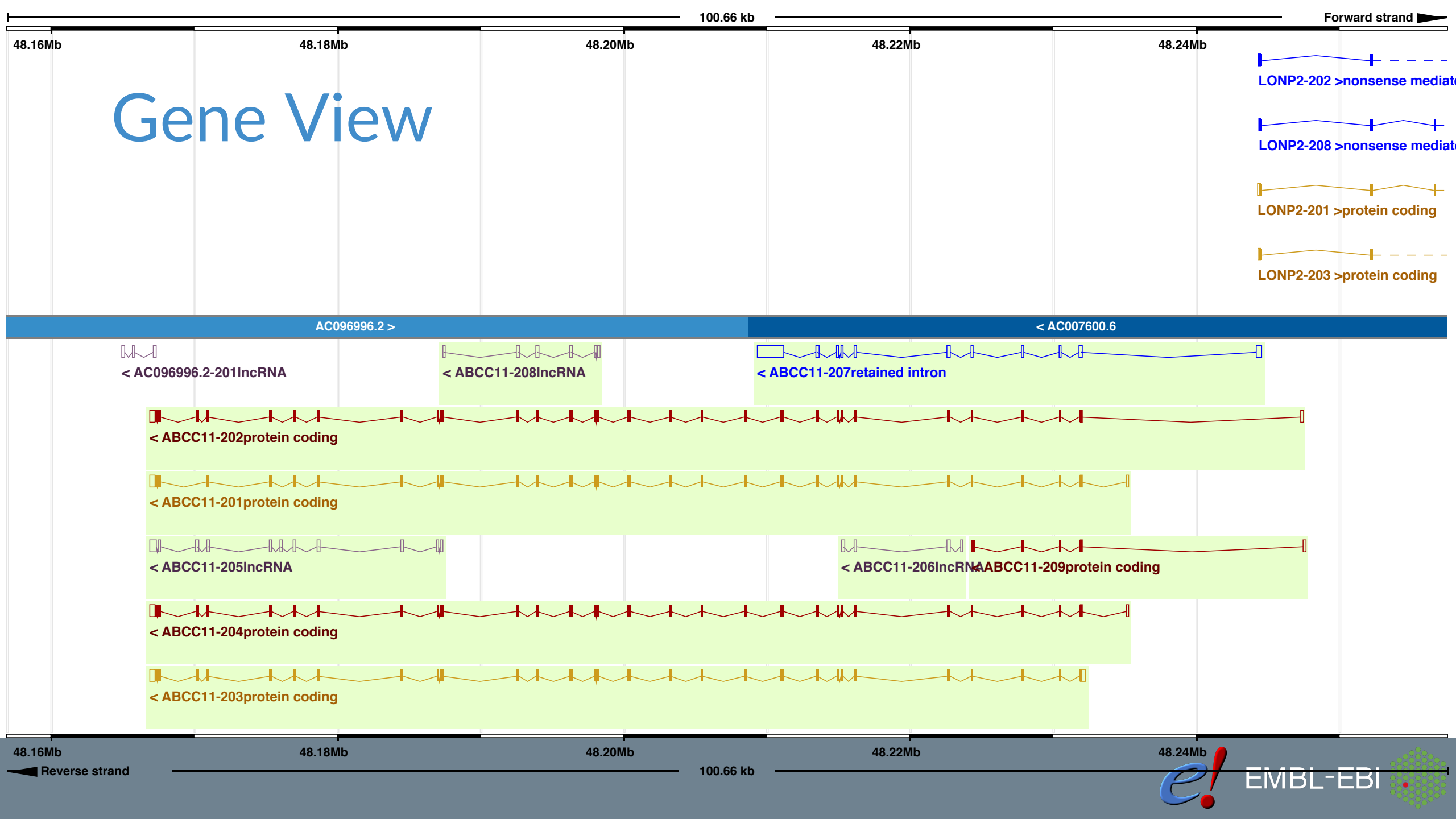
Training materials

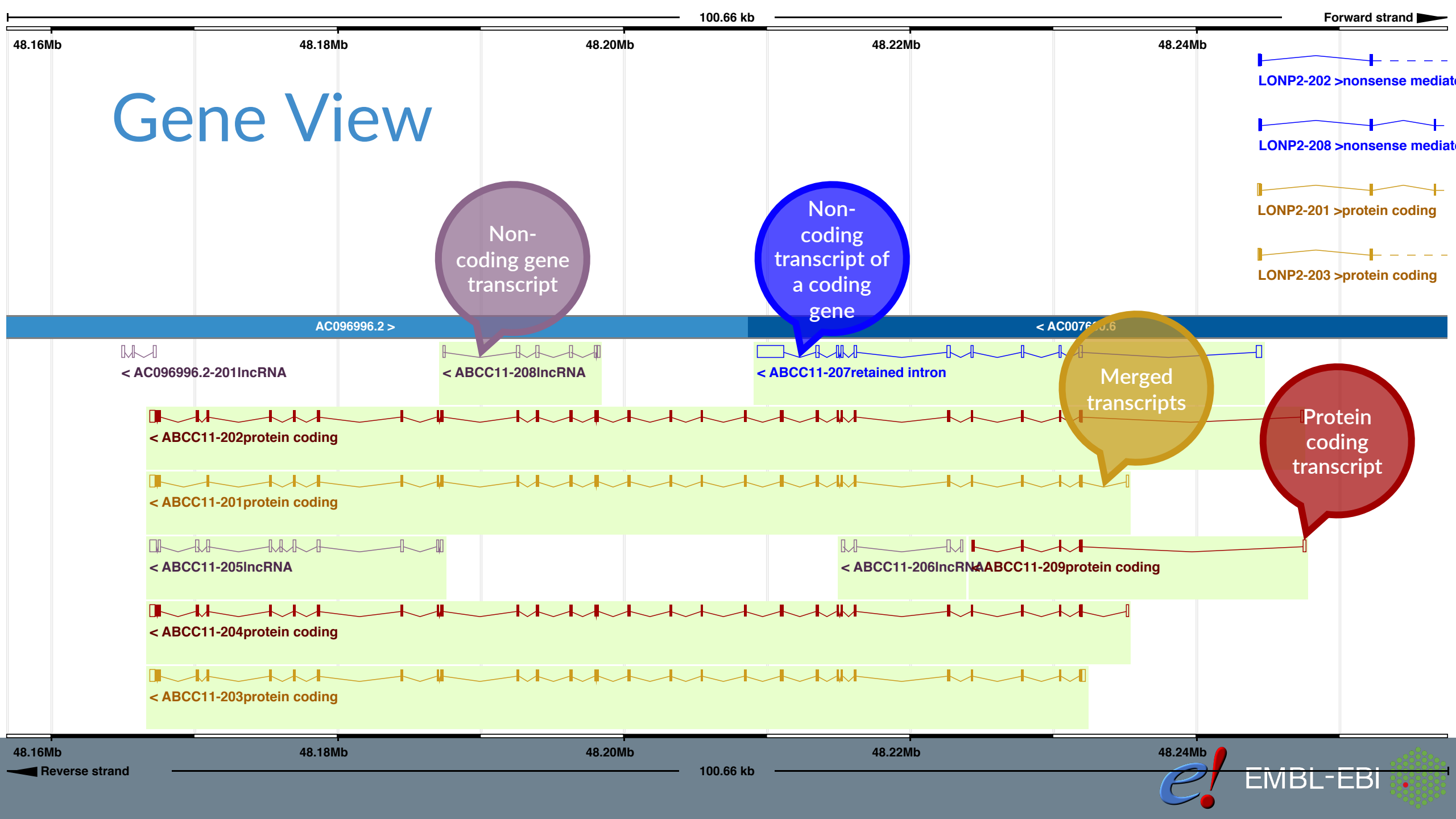
- Ensembl training materials are protected by a CC BY license:

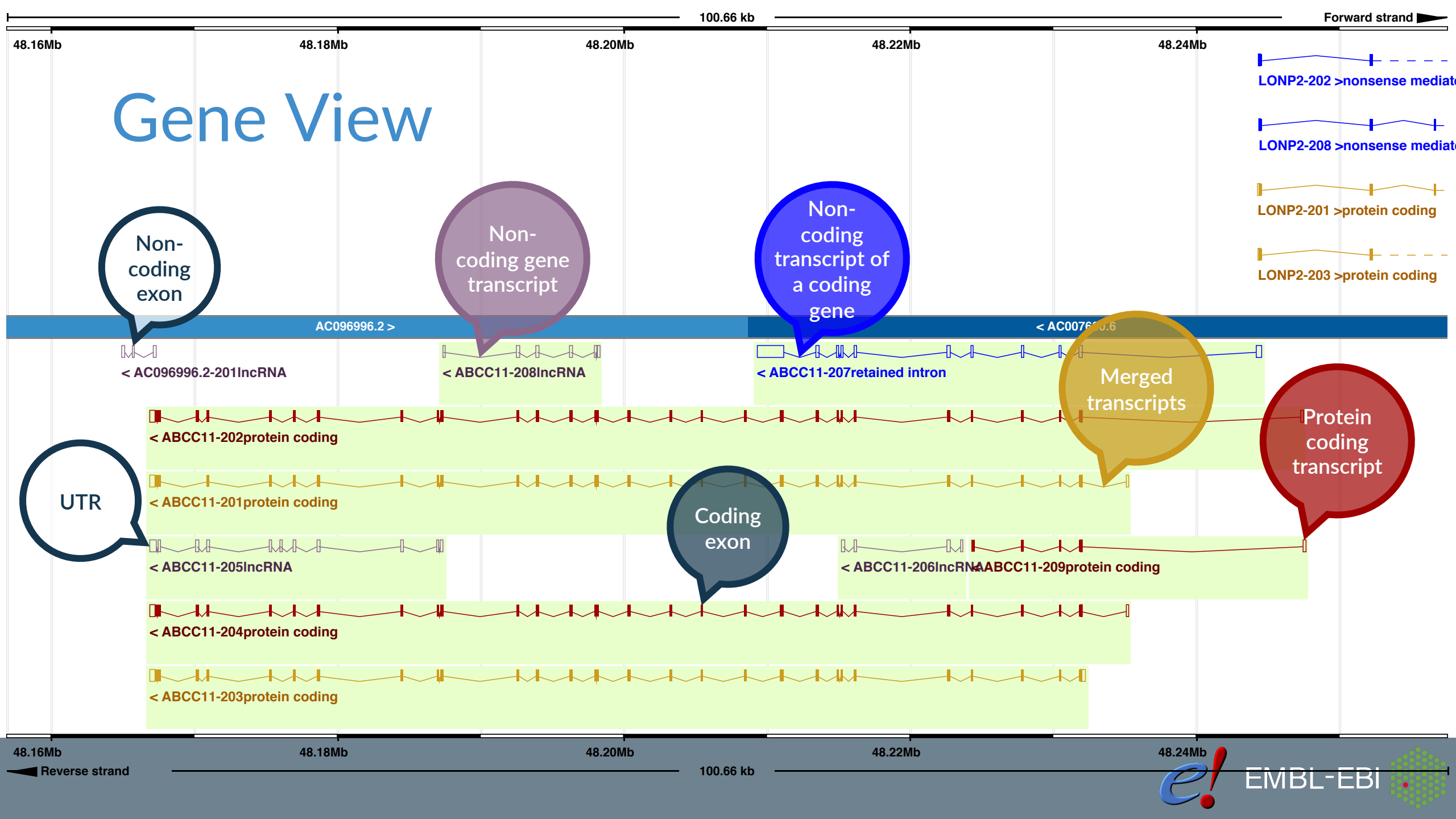
  creativecommons.org/licenses/by/4.0/

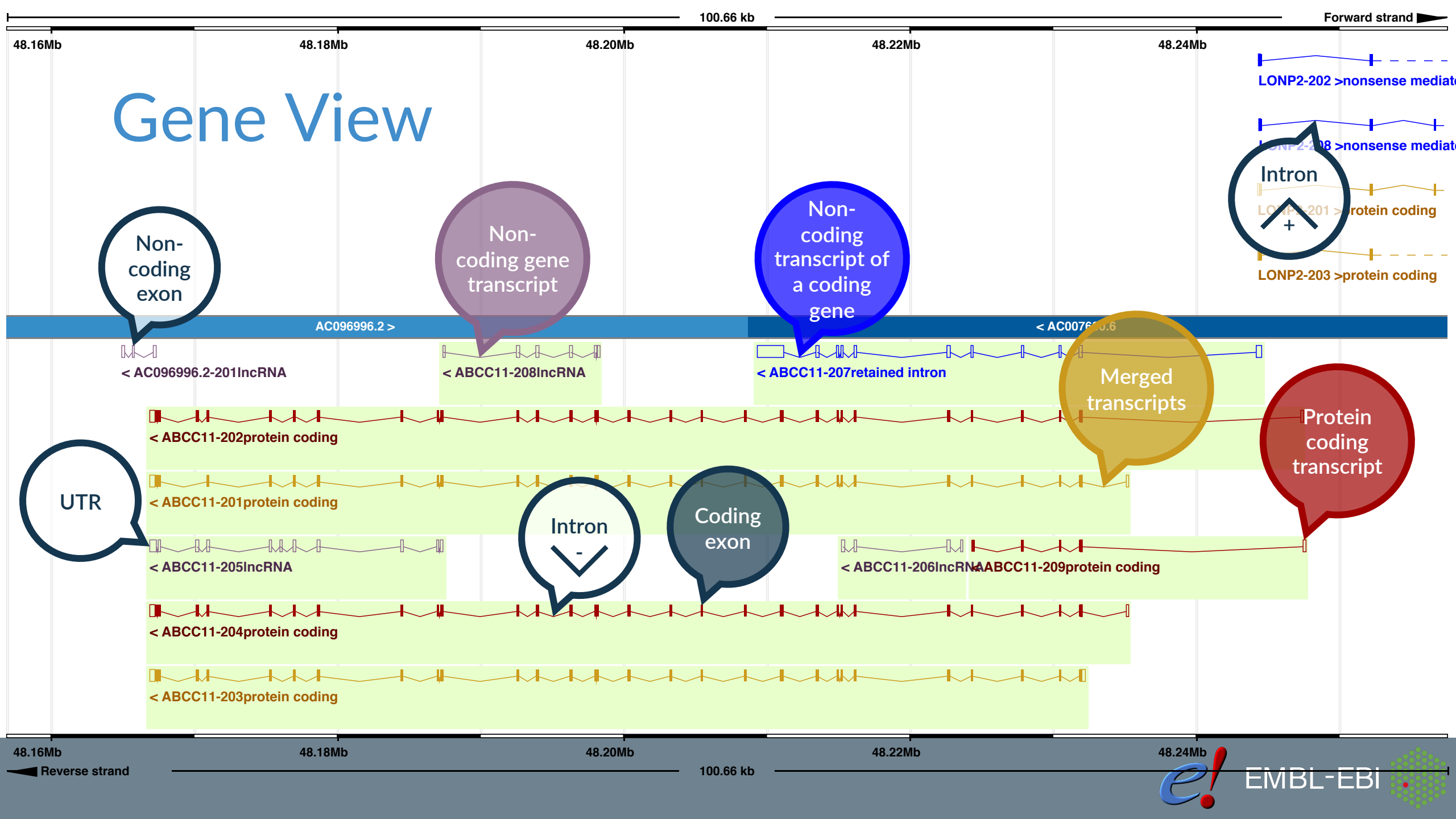
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers:

  ensembl.org/info/about/publications.html


100.66 kb
Forward strand
48.16Mb
48.18Mb
48.20Mb
48.22Mb
48.24Mb
Gene View
LONP2-202 >nonsense mediat
LONP2-208 >nonsense mediat
LONP2-201 >protein coding
LONP2-203 >protein coding
AC096996.2 >
< AC007600.6
< AC096996.2-201lncRNA
< ABCC11-208lncRNA
< ABCC11-207retained intron
< ABCC11-202protein coding
< ABCC11-201protein coding
< ABCC11-205lncRNA
< ABCC11-206lncRNA
< ABCC11-209protein coding
< ABCC11-204protein coding
< ABCC11-203protein coding
48.16Mb
48.18Mb
48.20Mb
48.22Mb
48.24Mb
Reverse strand
100.66 kb
EMBL-EBI

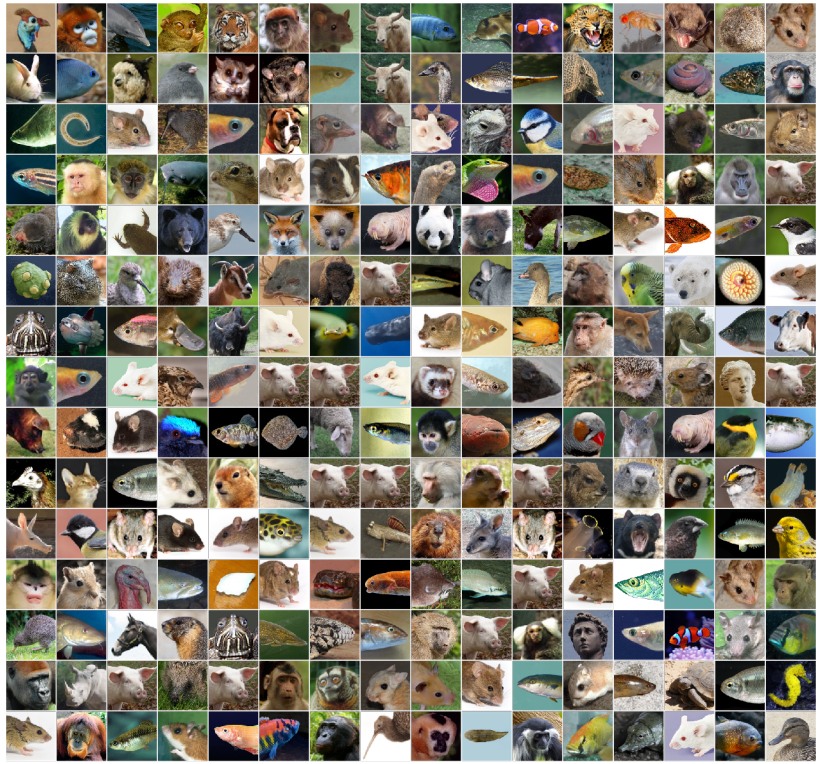

100.66 kb
Forward strand
48.16Mb
48.18Mb
48.20Mb
48.22Mb
48.24Mb
Gene View
LONP2-202 >nonsense mediat
LONP2-208 >nonsense mediat
LONP2-201 >protein coding
LONP2-203 >protein coding
Non-coding gene transcript
Non-coding transcript of a coding gene
AC096996.2 >
< AC096996.2-201lncRNA
< ABCC11-208lncRNA
< ABCC11-207retained intron
Merged transcripts
Protein coding transcript
< ABCC11-202protein coding
< ABCC11-201protein coding
< ABCC11-205lncRNA
< ABCC11-206lncRNA
< ABCC11-209protein coding
< ABCC11-204protein coding
< ABCC11-203protein coding
48.16Mb
48.18Mb
48.20Mb
48.22Mb
48.24Mb
Reverse strand
100.66 kb
EMBL-EBI


100.66 kb
Forward strand
48.16Mb
48.18Mb
48.20Mb
48.22Mb
48.24Mb
Gene View
LONP2-202 >nonsense mediat
LONP2-208 >nonsense mediat
LONP2-201 >protein coding
LONP2-203 >protein coding
Non-coding exon
Non-coding gene transcript
Non-coding transcript of a coding gene
AC096996.2 >
< AC096996.2-201lncRNA
< ABCC11-208lncRNA
< ABCC11-207retained intron
Merged transcripts
Protein coding transcript
UTR
< ABCC11-202protein coding
< ABCC11-201protein coding
Coding exon
< ABCC11-205lncRNA
< ABCC11-206lncRNA
ABCC11-209protein coding
< ABCC11-204protein coding
< ABCC11-203protein coding
48.16Mb
48.18Mb
48.20Mb
48.22Mb
48.24Mb
Reverse strand
100.66 kb
EMBL-EBI

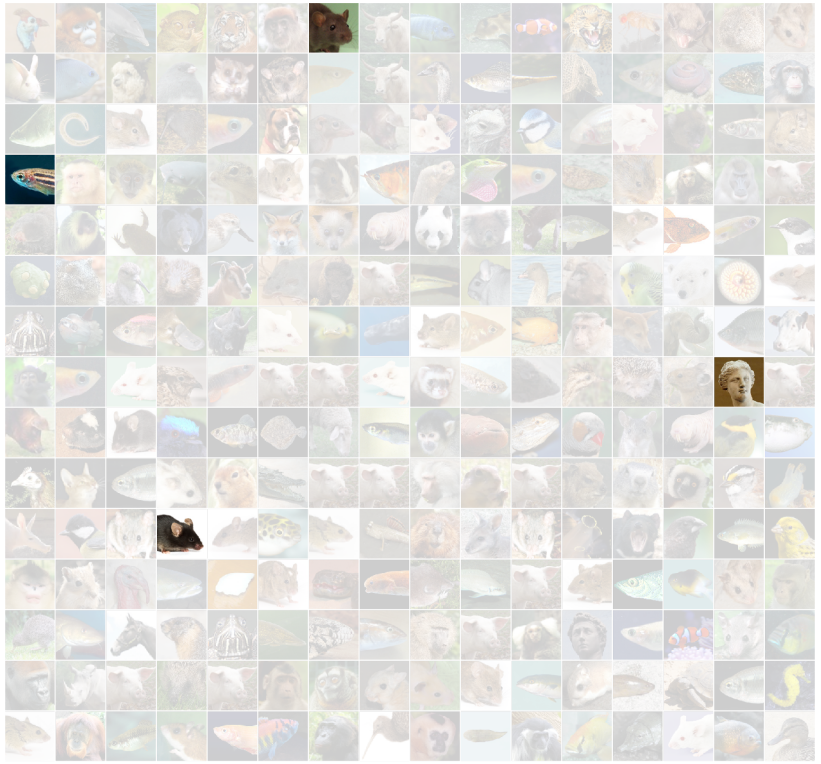

100.66 kb
Forward strand
48.16Mb
48.18Mb
48.20Mb
48.22Mb
48.24Mb
Gene View
LONP2-202 >nonsense mediate
LONP2-208 >nonsense mediate
Intron
+
LONP2-201 >protein coding
LONP2-203 >protein coding
Non-coding exon
Non-coding gene transcript
Non-coding transcript of a coding gene
AC096996.2 >
< AC096996.2-201lncRNA
< ABCC11-208lncRNA
< ABCC11-207retained intron
Merged transcripts
Protein coding transcript
UTR
< ABCC11-202protein coding
< ABCC11-201protein coding
Intron
-
Coding exon
< ABCC11-205lncRNA
< ABCC11-206lncRNA
ABCC11-209protein coding
< ABCC11-204protein coding
< ABCC11-203protein coding
48.16Mb
48.18Mb
48.20Mb
48.22Mb
48.24Mb
Reverse strand
100.66 kb
EMBL-EBI

Ensembl and Havana annotation

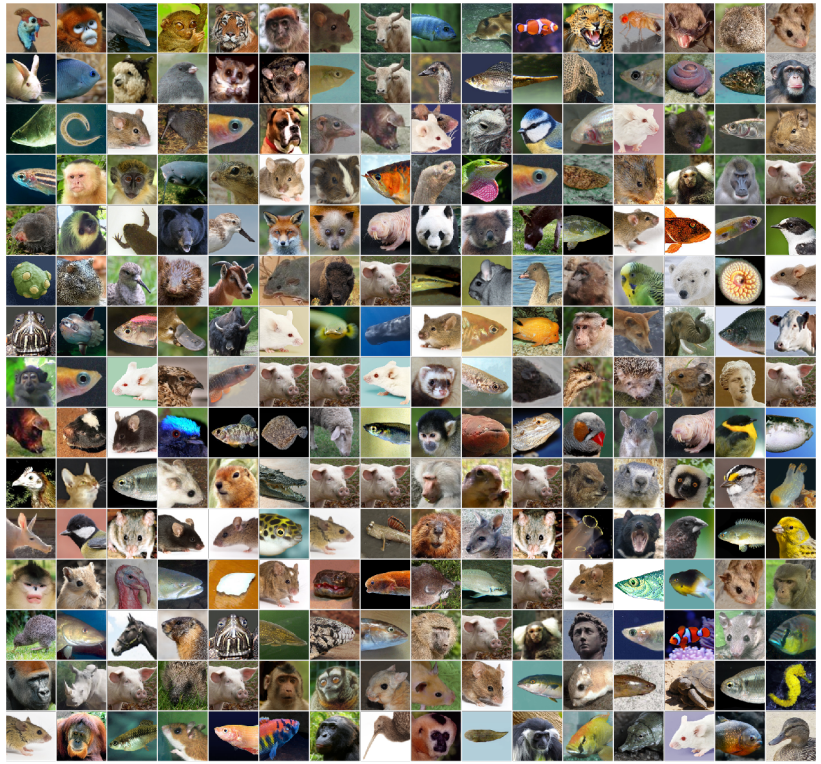
Ensembl



havana
human and vertebrate analysis and annotation



Automatic annotation



- Genome-wide determination using automated pipeline
- Predictions based on experimental data: known proteins, cDNAs plotted onto the genome using sequence matching
- One genome in two weeks
- Not *ab initio* ORFs prediction!

Biological evidence

- INSDC: International Nucleotide Sequence Database Collaboration
  - Collaboration between:
  - cDNAs
  - RNAseq
  - ESTs

NCBI

DDBJ DNA Data Bank of Japan

ENA European Nucleotide Archive

- Protein sequence databases
  - Swiss-Prot: manually curated
  - TrEMBL: unreviewed translations

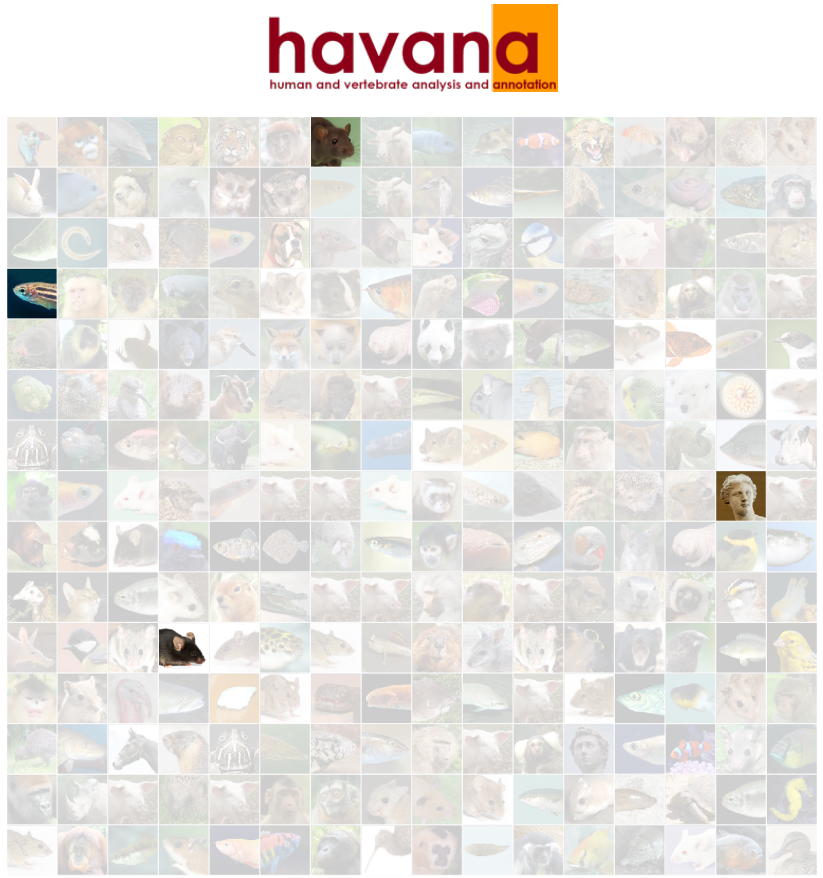
UniProt

- Homology based inference: e.g. predicting genes in

 by mapping cDNAs/proteins from  to the  genome

Manual annotation

- Genome-wide annotation
- Gene determination on a case-by-case basis by a person (annotator)
- Uses data from databases and papers
- One gene in half a day
- One genome in several years



Manual annotation

- Genome-wide annotation
- Gene determination on a case-by-case basis by a person (annotator)
- Uses data from databases and papers
- One gene in half a day
- One genome in several years

- RNA-seq transcriptome data
  - Illumina short reads
  - Long reads (PacBio and Oxford Nanopore)
- Transcript structure data
  - Introns
  - CAGE transcription start sites
  - PolyA-Seq transcription ends
- Mass-Spec protein data
- Publications

Manual annotation

Advantages

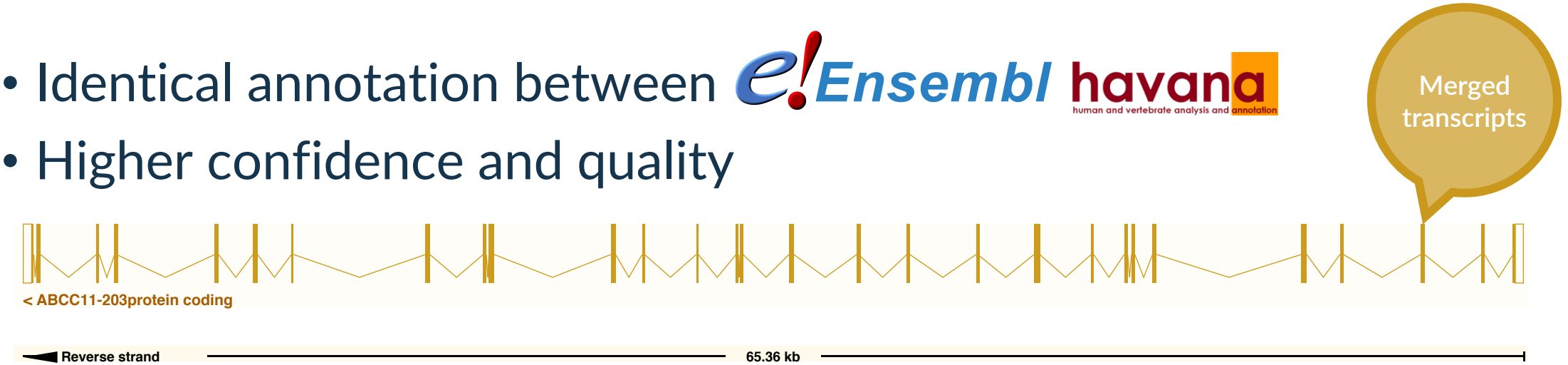
- More comprehensive (especially for non-coding features)
  - More genes
  - More transcripts per gene
  - More biotypes
- Requires less evidence
- More accurate for difficult regions (e.g. UTRs, splice sites, exceptions: immunoglobulins, stop codon readthroughs)

Disadvantages

- Slower
- Small scale

Golden transcripts

- Identical annotation between Ensembl and havana (human and vertebrate analysis and annotation)
- Higher confidence and quality



Imported annotation

wellcome sanger institute

horizon™ INSPIRED CELL SOLUTIONS USCS Comparative Annotation Toolkit

- We have imported annotation for small number of species from external sources:
  - Chinese hamster (ovary cell line: CHOK1GS)
  - 18 mouse strains/species
- Quality ensured by:
  - Our rigorous QC
  - Trusted sources

GENCODE:

Ensembl human and mouse genes

- The GENCODE gene set is the merged set of
  - Ensembl automatically annotated genes
  - Havana manually annotated genes
- It is the default gene set used by major projects such as:

gnomAD genome aggregation database

HUMAN CELL ATLAS

1000 Genomes Project

International Cancer Genome Consortium for Medicine

BLUEPRINT epigenome

THE CANCER GENOME ATLAS

GTEx

ENCODE

Genomics england

MANE: Matched Annotation from the NCBI and EMBL-EBI

- 100% identity (TSS, 5' UTR, CDS, 3' UTR, 3' end) between RefSeq (NM) and Ensembl (ENST) transcript
- MANE Select: one well-supported transcript for every protein-coding gene, agreed to be the most biologically relevant of that gene (based on expression, conservation and clinical variation)
- MANE Plus Clinical: second transcript where there are mutually exclusive clinically important exons (43 additional transcripts)

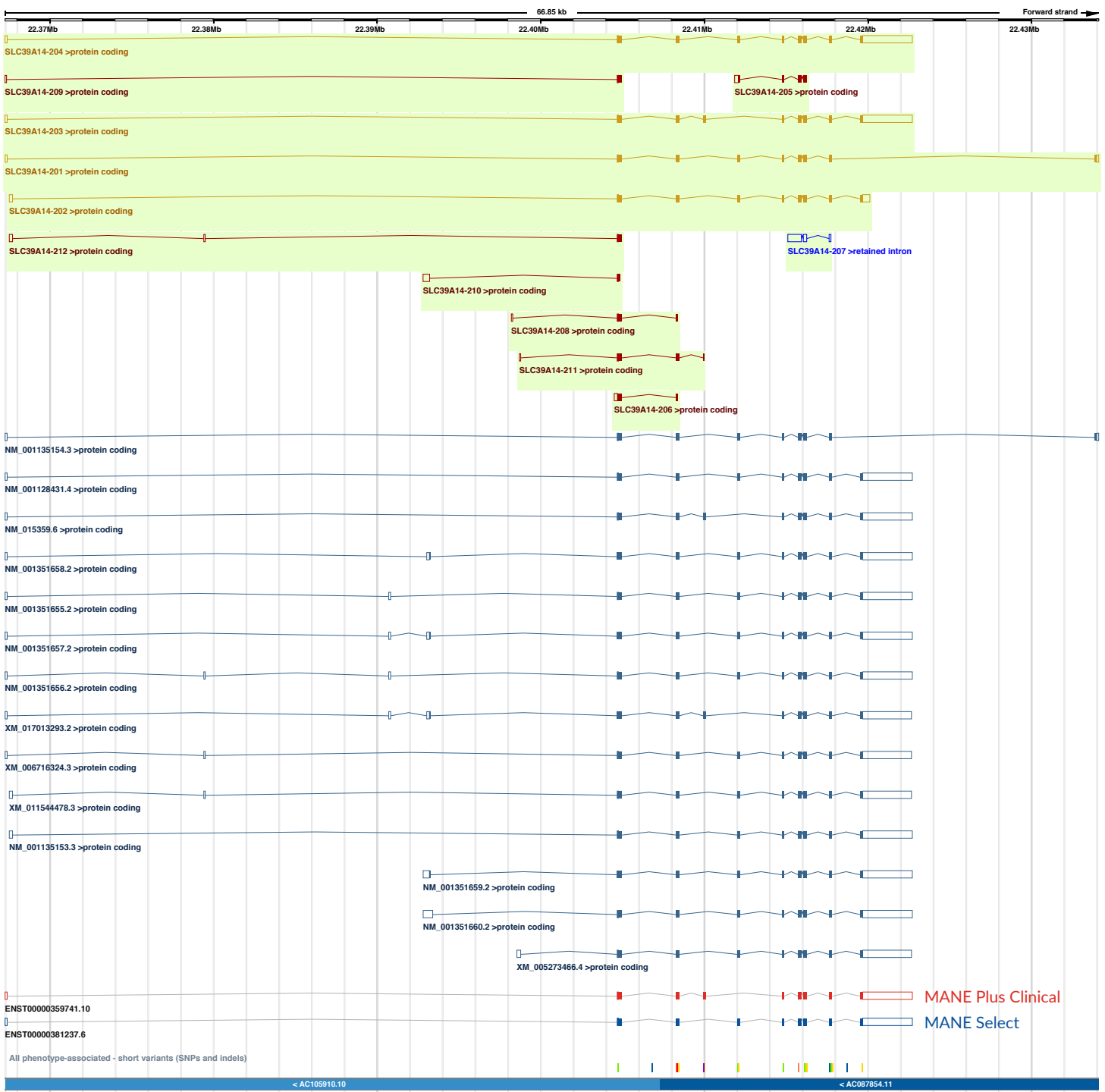
https://www.youtube.com/watch?v=43p_73zJfIo&feature=youtu.be
http://www.ensembl.info/2019/03/12/coming-soon-mane-select-v0-5/

EMBL-EBI

MANE: Matched Annotation from the NCBI and EMBL-EBI

NCBI

EMBL-EBI



Which transcript do I use?

1. MANE Select
2. APPRIS principal isoform:
   - Major isoforms(s) from combining protein structural information, functionally important residues and evidence from cross-species alignments: scored P1-P5 (1 = best)
3. GENCODE basic:
   - 5' and 3' complete GENCODE transcripts
4. Transcript Support Level (TSL):
   - Scored 1-5 for quality (1 = best)
5. CCDS:
   - Consensus Coding DNA Sequence: matching CDS from NCBI and EMBL-EBI
6. Golden transcripts

http://www.ensembl.org/Help/Faq?id=276
http://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html

1. MANE Select
2. APPRIS principal isoform:
   - Major isoforms(s) from combining protein structural information, functionally important residues and evidence from cross-species alignments: scored P1-P5 (1 = best)
3. GENCODE basic:
   - 5’ and 3’ complete GENCODE transcripts
4. Transcript Support Level (TSL):
   - Scored 1-5 for quality (1 = best)
5. CCDS:
   - Consensus Coding DNA Sequence: matching CDS from NCBI and EMBL-EBI
6. Golden transcripts

http://www.ensembl.org/Help/Faq?id=276
http://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html

Ensembl Canonical transcript

- Single, representative transcript identified at every locus
- Selection for human coding genes:
  - Highly conserved CDS (PhyloCSF)
  - Highly expressed (RNAseq reads spanning introns, CAGE)
  - Concordant with APPRIS P1 isoform
  - Concordant with UniProt canonical isoform
  - Longest CDS
  - Covers the largest number of clinical variants
  - Complete
- Where available Ensembl Canonical = MANE Select

Ensembl Canonical

A single transcript chosen for a gene which is the most conserved, most highly expressed, has the longest coding sequence and is represented in other key resources, such as NCBI and UniProt.

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
BRCA2-201	ENST00000380152.8	11954	3418aa	Protein coding	CCDS9344	P51587	NM_000059.4	MANE Select v0.93, Ensembl Canonical, GENCODE basic, APPRIS P1, TSL:5
BRCA2-210	ENST00000680887.1	11880	3418aa	Protein coding	CCDS9344	-	-	APPRIS P1
BRCA2-206	ENST00000544455.6	11854	3418aa	Protein coding	CCDS9344	P51587	-	GENCODE basic, APPRIS P1, TSL:1
BRCA2-204	ENST00000530893.6	2011	481aa	Protein coding	-	A0A590UJI7	-	TSL:1, CDS 3' incomplete
BRCA2-207	ENST00000614259.2	11763	2649aa	Nonsense mediated decay	-	-	-	TSL:2
BRCA2-208	ENST00000665585.1	2598	438aa	Nonsense mediated decay	-	A0A590UJU6	-	CDS 5' incomplete
BRCA2-202	ENST00000470094.1	842	186aa	Nonsense mediated decay	-	H0YE37	-	TSL:5, CDS 5' incomplete
BRCA2-209	ENST00000666593.1	523	58aa	Nonsense mediated decay	-	A0A590UJ24	-	CDS 5' incomplete
BRCA2-203	ENST00000528762.1	495	64aa	Nonsense mediated decay	-	H0YD86	-	TSL:4, CDS 5' incomplete
BRCA2-205	ENST00000533776.1	523	No protein	Retained intron	-	-	-	TSL:3

Ensembl stable IDs

- ENSG##########.# Ensembl Gene ID
- ENST##########.# Ensembl Transcript ID
- ENSP##########.# Ensembl Protein ID
- ENSE##########.# Ensembl Exon ID
- ENSR##########.# Ensembl Regulatory region ID
- For non-human species a suffix is added:
  - MUS (*Mus musculus*) for mouse: ENSMUSG###
  - DAR (*Danio rerio*) for zebrafish: ENSDARG###

More information

Aken *et al.*

The Ensembl gene annotation system

Database: the Journal of Biological Databases and Curation (2016)

europepmc.org/articles/PMC4919035