Working with ENCODE Data

Emily Pritchard
Ensembl

Bob Kuhn
UCSC

This course

Today

- Overview of the GENCODE reference geneset (EP)
- Overview of the ENCODE project and data part 1 (BK)
- Tea/coffee
- Overview of the ENCODE project and data part 2 (BK)
- Accessing ENCODE data with UCSC part 1 (BK)

Tomorrow

- Accessing ENCODE data with UCSC part 2 (BK)
- Accessing ENCODE data with Ensembl part 1 (EP)
- Tea/coffee
- Accessing ENCODE data with Ensembl part 2 (EP)
- General Q&A session (EP, BK)

Course materials

http://www.gencodegenes.org/workshops.html

- Slides
- Workbook
- Appendices

The GENCODE gene set

Dr Emily Pritchard

EMBL-EBI

wellcome trust sanger institute

This talk

- GENCODE: what and why?
- Genome assemblies
- Gene annotation in GENCODE
  - Automatic annotation
  - Manual annotation
  - The merge
  - CCDS
- Where to find GENCODE data

What is GENCODE?

ENCODE

- Project
- Phase 2 GENCODE Goals
- Data
- Statistics - Human
- Statistics - Mouse
- Participants
- Publications
- lncRNA microarray
- RGASP 1/2
- RGASP 3
- Blog
- GENCODE workshops
- Contact us

The GENCODE Project:

Encyclopædia of genes and gene variants

Current GENCODE version

The current version in **Human** is **Gencode 19**, released on the 10/12/2013.
For more information about the human releases please see the **README.txt** file.

The current version in **Mouse** is **Gencode M2**, released on the 10/12/2013.
For more information about the mouse releases please see the **README.txt** file.

** **NEW** ** Two publications now out on our RNASeq genome annotation assessment project (RGASP):

- **Assessment of transcript reconstruction methods for RNA-seq.**
  Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consortium, Abril JF, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, Bertone P, Bohnert R, Bucher P, Cloonan N, Derrien T, Djebali S, Du J, Dudoit S, Engström PG, Gerstein M, Gingeras TR, Gonzalez D, Grimmond SM, Guigó R, Habegger L, Harrow J, Hubbard TJ, Iseli C, Jean G, Kahles A, Kokocinski F, Lagarde J, Leng J, Lefebvre G, Lewis S, Mortazavi A, Niermann P, Rätsch G, Reymond A, Ribeca P, Richard H, Rougemont J, Rozowsky J, Sammeth M, Sboner A, Schulz MH, Searle SM, Solorzano ND, Solovyev V, Stanke M, Steijger T, Stevenson BJ, Stockinger H, Valsesia A, Weese D, White S, Wold BJ, Wu J, Wu TD, Zeller G, Zerbino D, Zhang MQ, Hubbard TJ, Guigó R, Harrow J and Bertone P
  Nature methods 2013;**10**;12;1177-84
  PUBMED: **24185837**; PMC: **3851240**; DOI: **10.1038/nmeth.2714**
- **Systematic evaluation of spliced alignment programs for RNA-seq data.**
  Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, RGASP Consortium, Alioto T, Behr J, Bertone P, Bohnert R, Campagna D, Davis CA, Dobin A, Engström PG, Gingeras TR, Goldman N, Grant GR, Guigó R, Harrow J, Hubbard TJ, Jean G, Kahles A, Kosarev P, Li S, Liu J, Mason CE, Molodtsov V, Ning Z, Ponstingl H, Prins JF, Rätsch G, Ribeca P, Seledtsov I, Sipos B, Solovyev V, Steijger T, Valle G, Vitulo N, Wang K, Wu TD, Zeller G, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R and Bertone P
  Nature methods 2013;**10**;12;1185-91
  PUBMED: **24185836**; DOI: **10.1038/nmeth.2722**

Who is involved?

- Ensembl (EBI/WTSI) - automatic gene annotation
- Havana (WTSI) - manual gene annotation
- Yale - pseudogene annotation
- CNIO - protein validation
- MIT - comparative genomics based validation
- Lausanne - experimental validation
- CRG - experimental validation
- UCSC - quality control

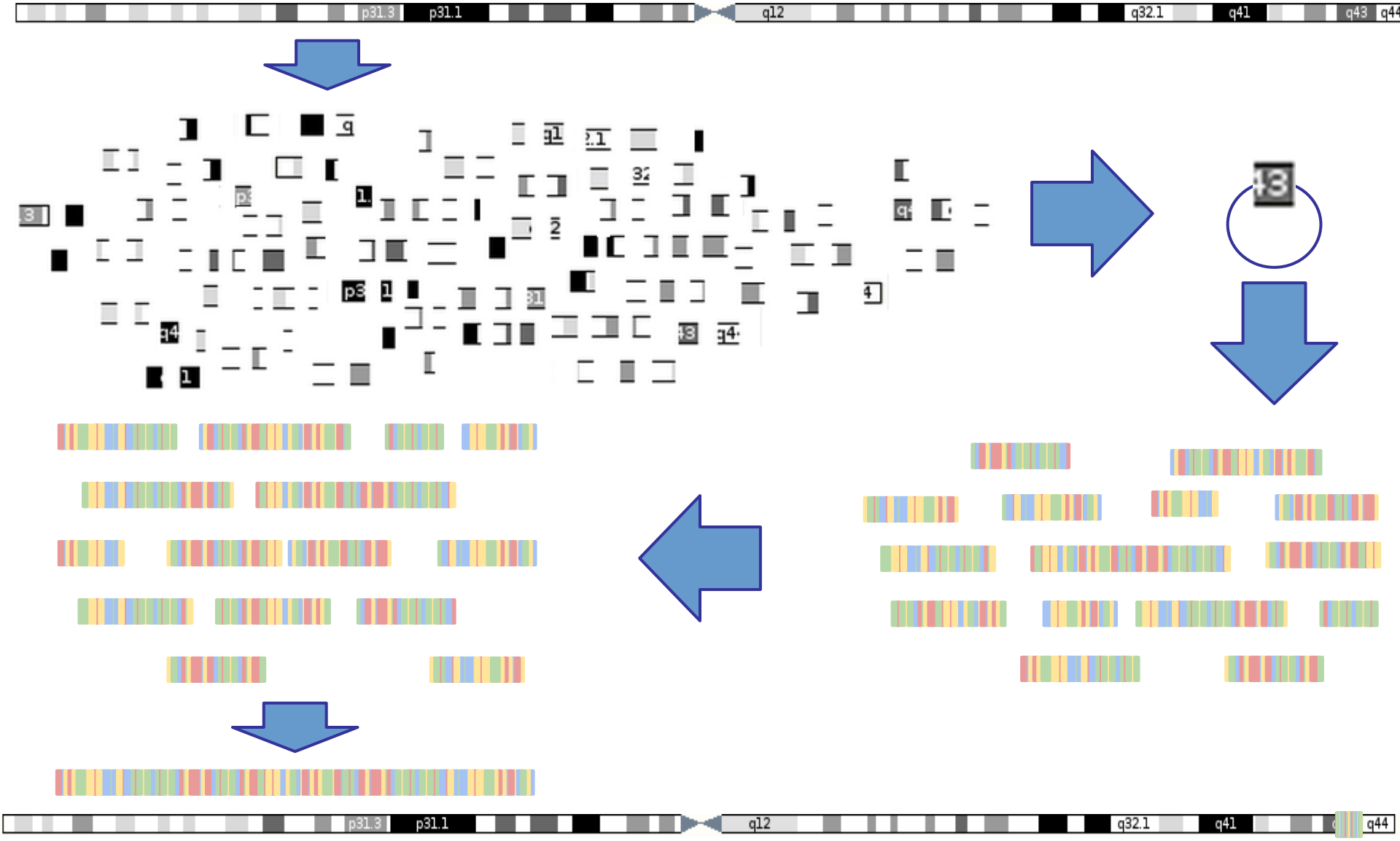
Why is a reference gene set important?

- We want a reliable set of genes to study
- We need something to compare other genome-wide data to, eg:
  - Regulatory regions (ENCODE)
  - Variation (1000 genomes)

This talk

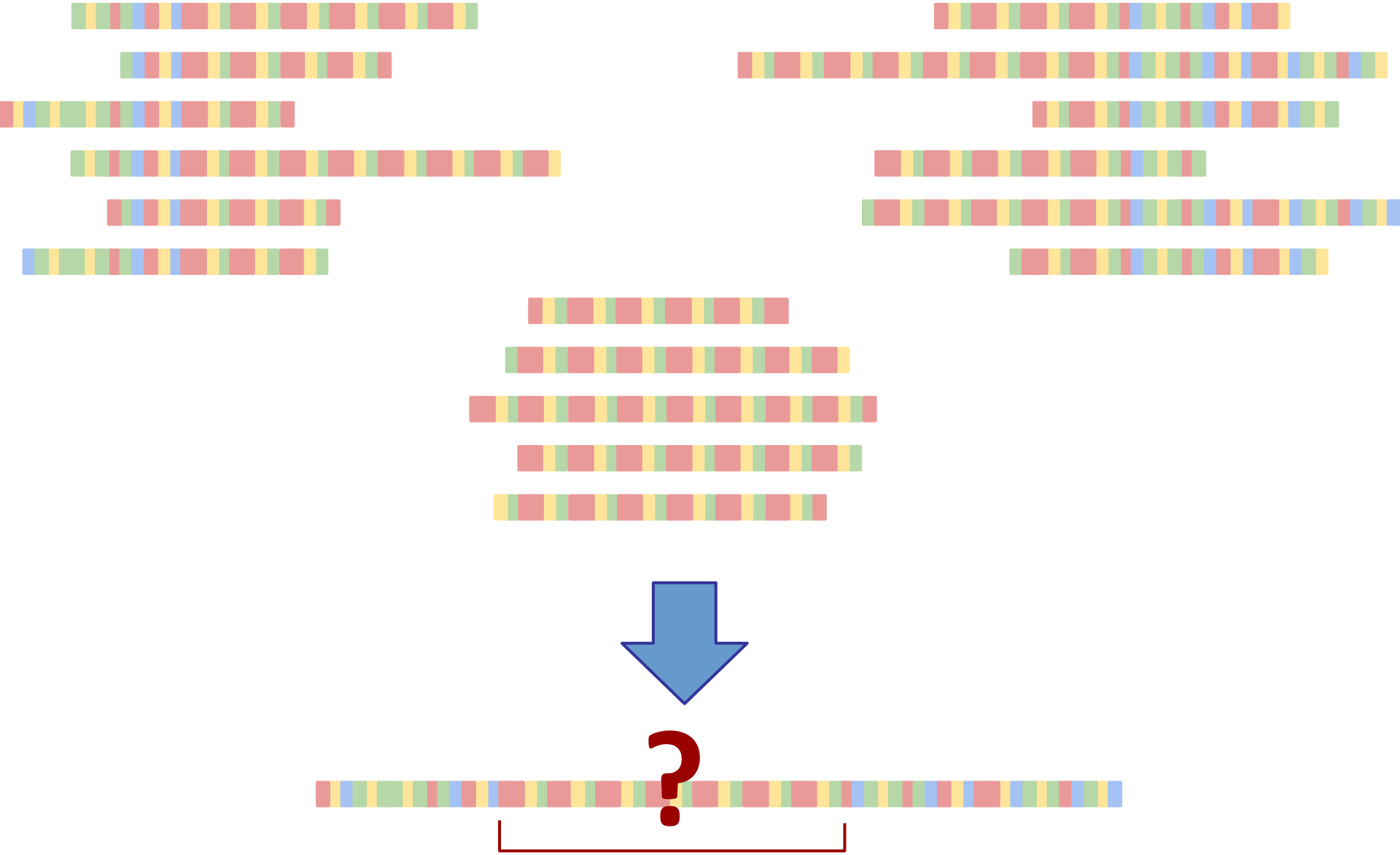
- GENCODE: what and why?
- Genome assemblies
- Gene annotation in GENCODE
  - Automatic annotation
  - Manual annotation
  - The merge
  - CCDS
- Where to find GENCODE data

Genome assemblies

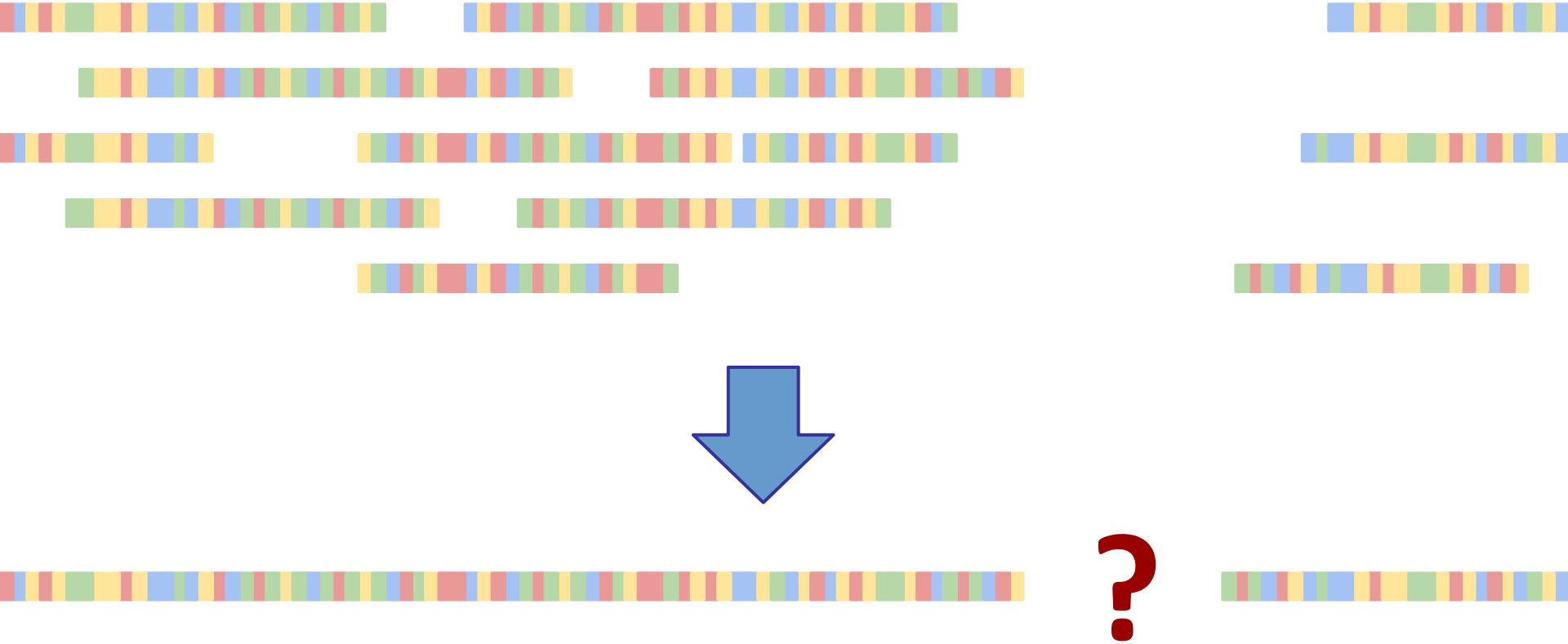


Assembling using overlap

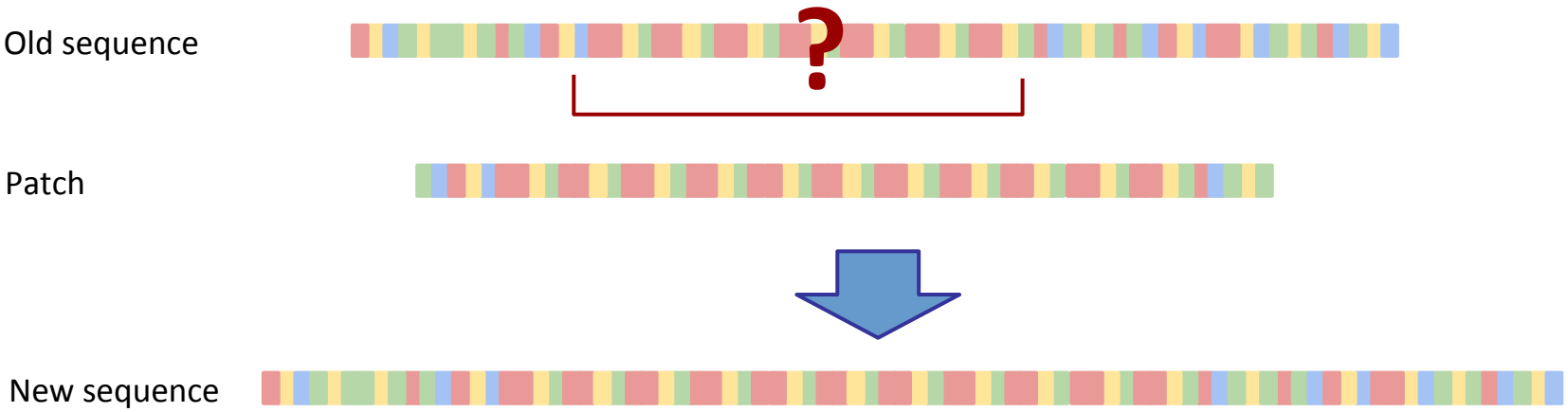
Repetitive sequences

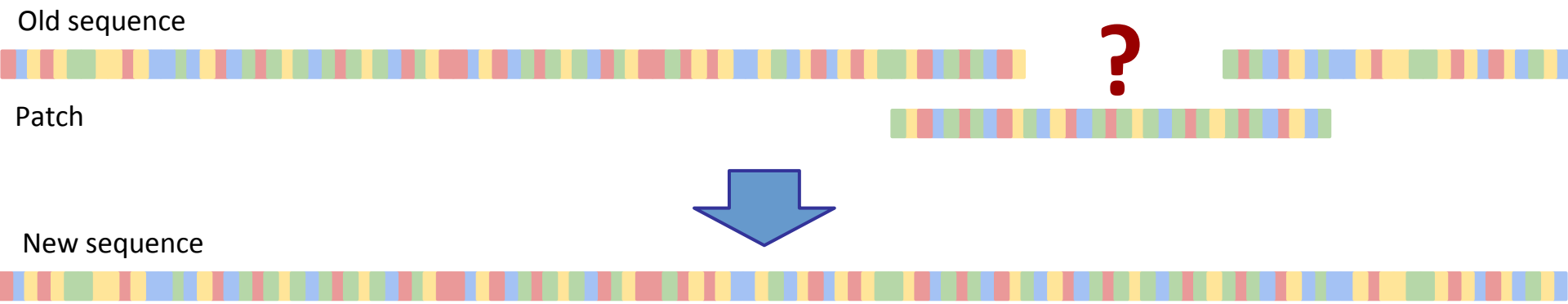


Gaps



Patches repair sequences





Patches and assembly updates

- A genome assembly will have errors.
- Fix patches will be added over time to repair these errors.
- The coordinate system of the genome is unaffected by the patches.
- The old sequence is retained as the primary assembly, with the patches placed on top.
- Every so often a new assembly comes out (eg GRCh38).
- All fix patches will be integrated into the assembly, fully replacing the old sequence.
- Genome coordinates change in a new assembly.

This talk

- GENCODE: what and why?
- Genome assemblies
- Gene annotation in GENCODE
  - Automatic annotation
  - Manual annotation
  - The merge
  - CCDS
- Where to find GENCODE data

The GENCODE set is made up of Ensembl and Havana annotation

Automatic annotation

Genome-wide determination using the Ensembl automated pipeline

Manual annotation

Gene determination on a case-by-case basis by a person

Both methods base their gene predictions on biological data.

Biological Evidence

- International Nucleotide Sequence databases

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

- NCBI RefSeq

  - Manually annotated proteins and mRNAs (NP, NM)

Ensembl automatic annotation –

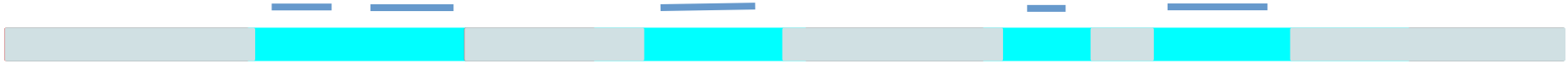
step 1: masking the genome

RepeatMasker

- RepeatMasker is used to find repetitive regions in the genome.
- These are masked for further analyses.
- Almost 50% of the genome is masked out.

Ensembl automatic annotation – step 2: finding gene regions

- Uniprot proteins are aligned to the genome using Pmatch.
- This is a high-sensitivity, high-specificity approach.
- It is used to find regions of the genome where exons might be.
- The regions identified (plus a bit of padding) are taken into the next analysis.



Ensembl automatic annotation –

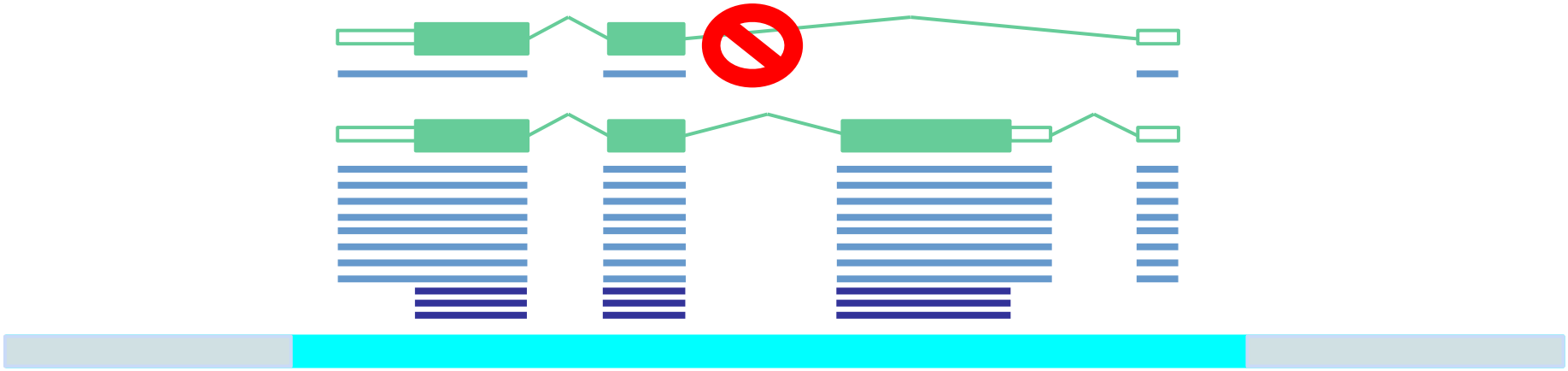
step 3: making transcript models

GeneWise

- Full protein sequences are used to weave together predicted exons.
- UTRs are added from cDNAs which overlap the protein sequence

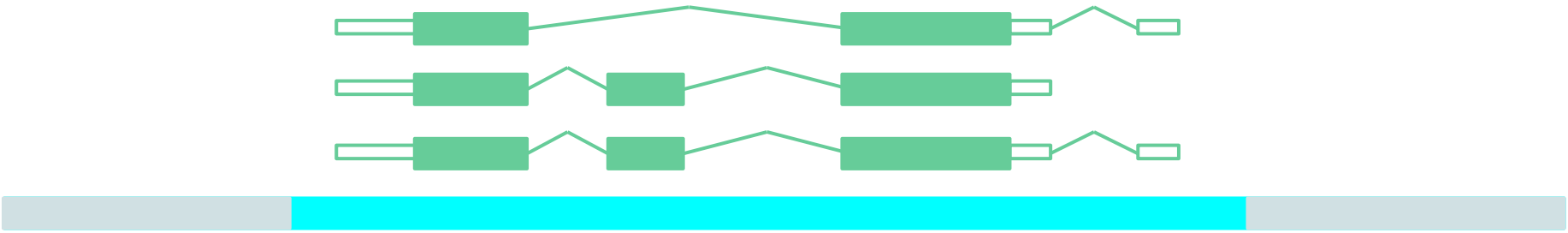
Ensembl automatic annotation – step 4: score hits

- There will be multiple hits from both cDNA and protein sequences repeated in the databases.
- We use the number of hits to score splice junctions and determine if transcripts are most well-supported.



Ensembl automatic annotation – step 5: group transcripts into genes

- If transcript share exons, they are grouped into genes

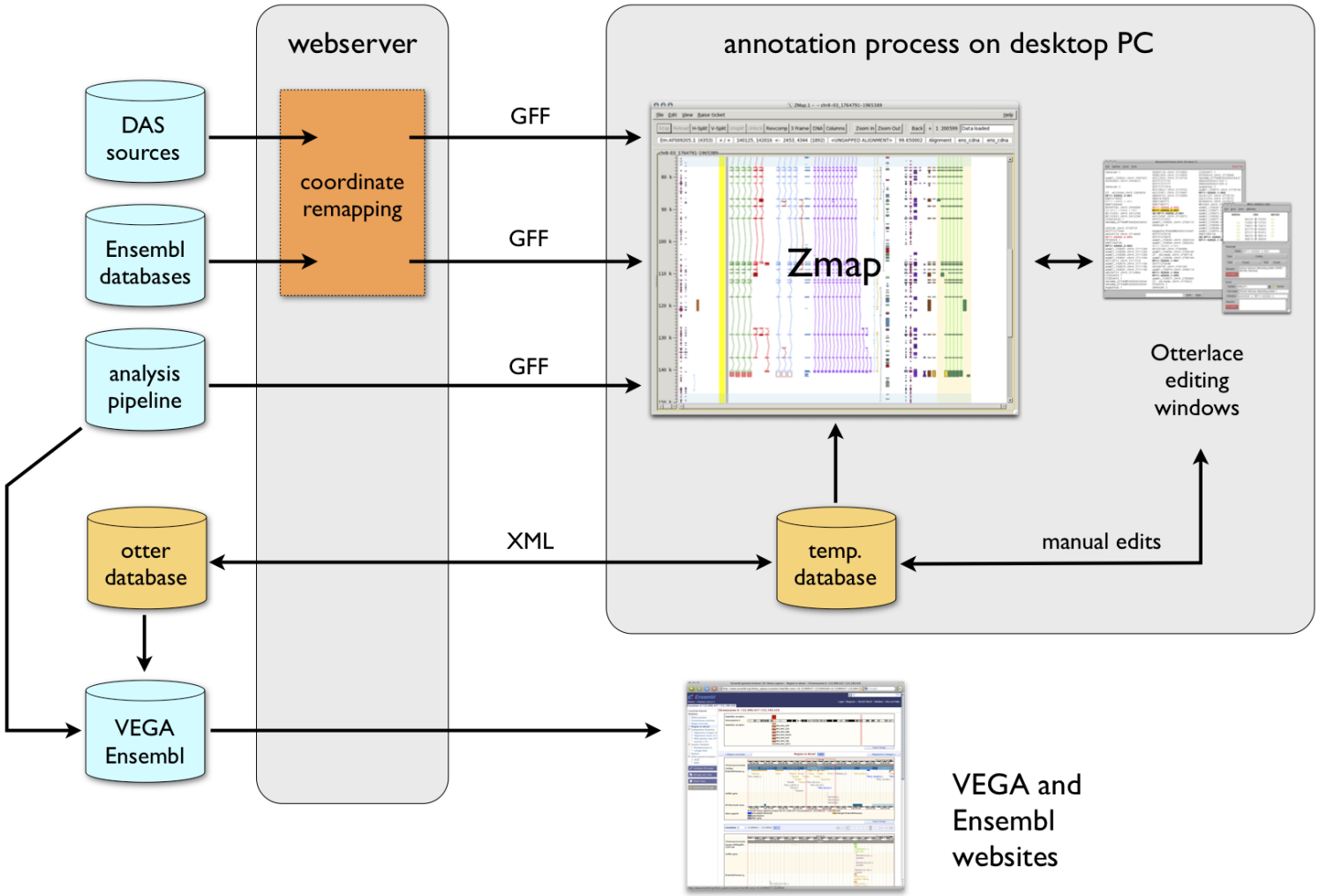


Havana manual annotation

- The main benefit of manual annotation is its flexibility.
- The data we have for each gene is of varying quality and quantity, which may need special attention.
- Pseudogenes and immunoglobulin genes have many biological exceptions, which also need special attention.
- Automated annotation can only determine ~75% of genes.
- Anything out of the ordinary: e.g. single-exon gene families such as olfactory receptors

- Data is taken from databases and publications.
- Particular attention is paid to splice sites and transcription start/stop sites.

Otterlace pipeline



Havana 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

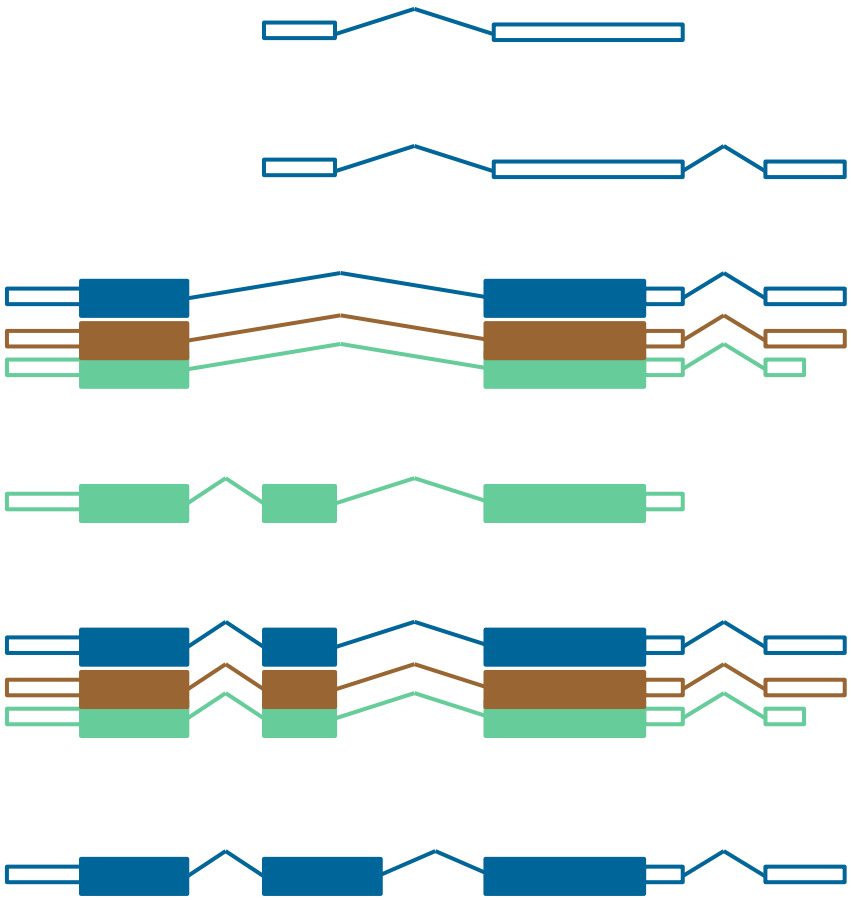
Immunoglobulin
IG_pseudogene
IG_Gene
TR_Gene

Differences between manual and automatic annotation

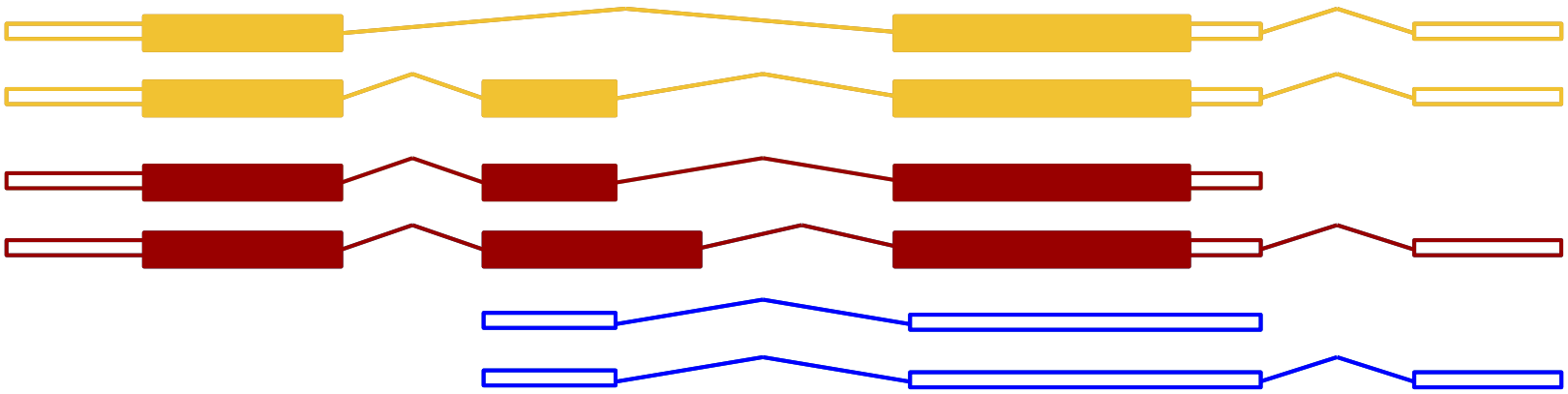
	Automatic	Manual
Speed	Fast – a complete human genome can be annotated in weeks	Slow – six months for a single chromosome
Sensitivity/ selectivity	More selective – aims to have the best supported transcripts	More sensitive – aims to annotate all transcripts
Data source	Databases	Databases and publications
Non-coding sequences	Good at sncRNAs and miRNAs	Good at lincRNAs, pseudogenes, splice variants.

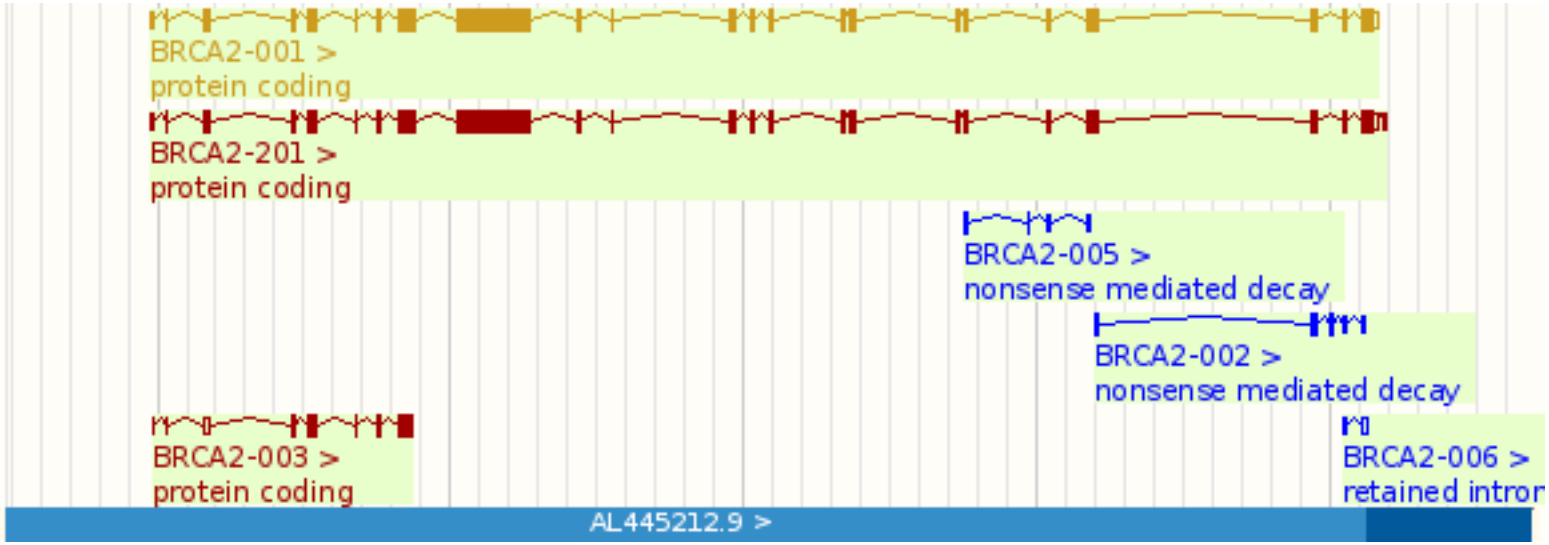
Ensembl/Havana merge

- Clusters of transcripts are compared, all against all
- Identical transcripts are merged
- Only intron junctions are compared, so differing UTRs are still merged
- Havana-annotated biotypes are assigned to transcripts



Ensembl/Havana merge



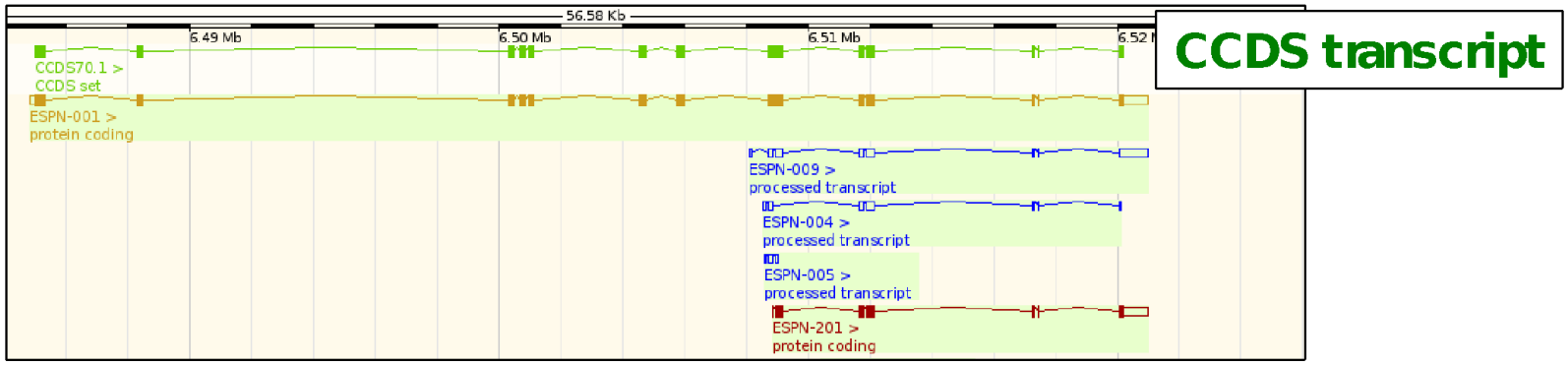


CCDS transcripts

Consensus CDS protein set

CCDS Database

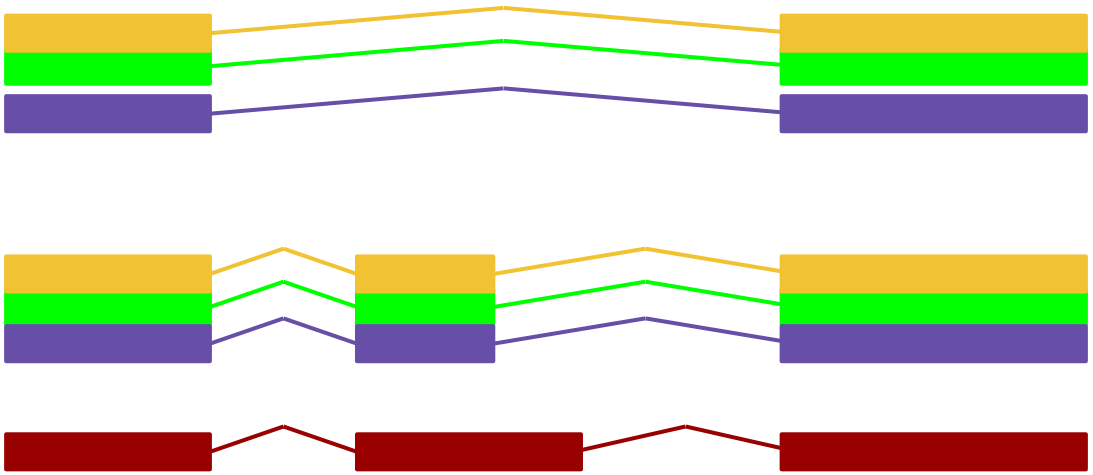
- Consensus coding DNA sequence set
- Agreement between EBI, WTSI, UCSC and NCBI
- http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi



CCDS annotation –

step 1: NCBI and Ensembl analyse all transcripts

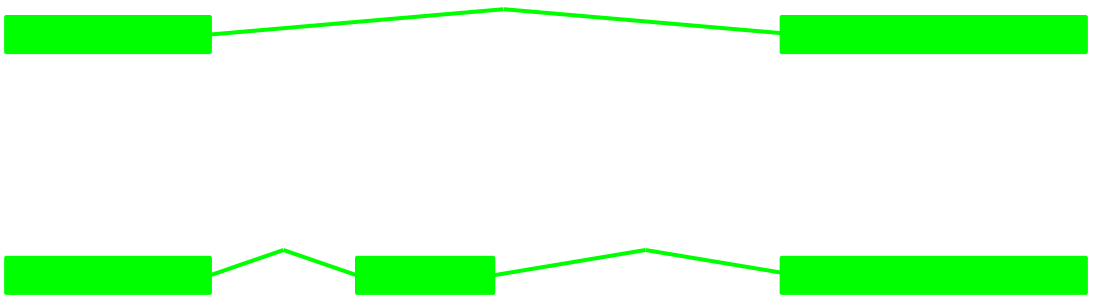
- NCBI and Ensembl both do the same analysis.
- They walk along the chromosome, checking if the same transcripts appear in the same places from both sources
- Only coding sequences are checked.
- This analysis is automatic



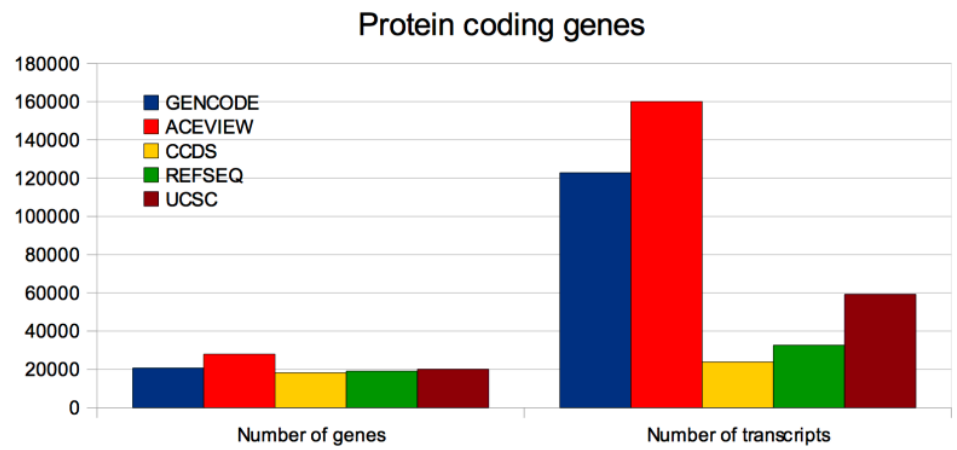
CCDS annotation –

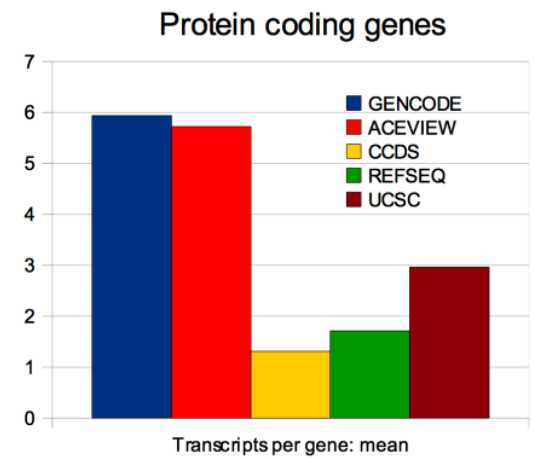
step 2: UCSC and Havana check new transcripts

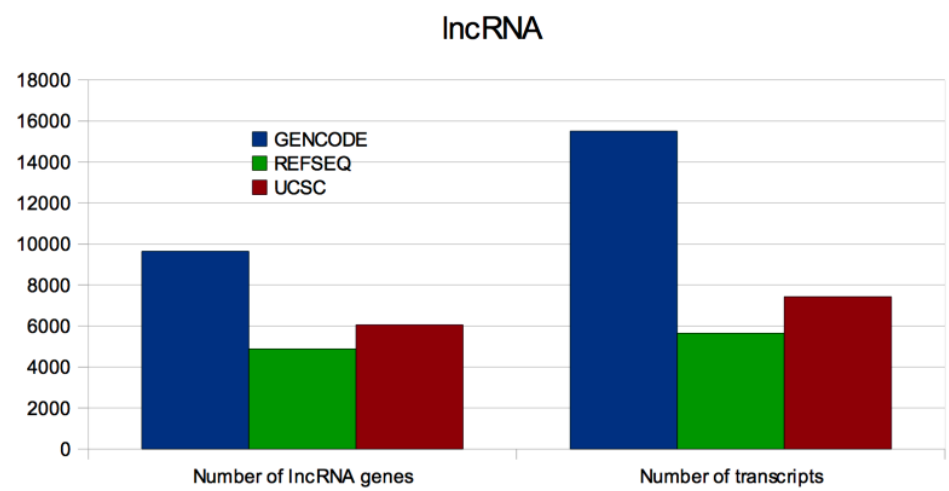
- Transcripts present in the last CCDS release are kept automatically.
- New transcripts are sent to UCSC and Havana who check them manually.
- Transcripts lost since the last release are checked too.

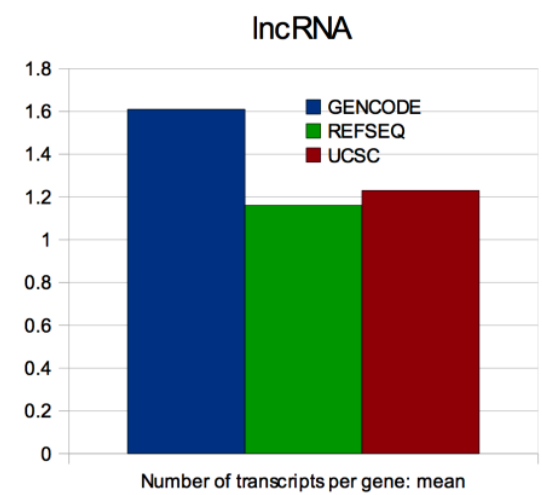


How GENCODE compares against other datasets?







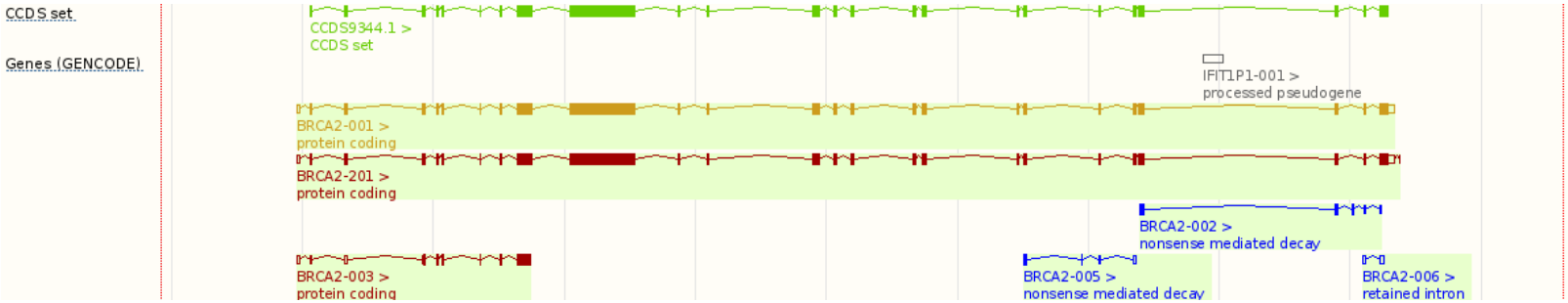


This talk

- GENCODE: what and why?
- Genome assemblies
- Gene annotation in GENCODE
  - Automatic annotation
  - Manual annotation
  - The merge
  - CCDS
- Where to find GENCODE data

View GENCODE data in...

Ensembl



UCSC

You can also download GTF files from GENCODE:
http://www.gencodegenes.org/releases/19.html

Havana annotation can also be browsed in Vega

Vega BLAST/BLAT | Help & Documentation

Login/Register

Vega search...

A repository for high-quality gene models produced by the manual annotation of vertebrate genomes.

wellcome trust sanger institute

havana

Search: All species for Go

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

Browse a genome

- Zebrafish [21-01-2014] [Ensembl]
- Human [23-10-2013] [Ensembl]
- Mouse [23-10-2013] [Ensembl]
- Pig [23-10-2013] [Ensembl]
- Rat [21-08-2013] [Ensembl]

Browse a region

- Tasmanian devil [23-10-2013] [Ensembl]
- Chimpanzee [12-01-2012] [Ensembl]
- Gorilla [30-03-2009] [Ensembl]
- Wallaby [30-03-2009] [Ensembl]
- Dog [14-02-2005] [Ensembl]

Major histocompatibility complex (MHC) annotation

Non-reference regions
Human: 6-COX, 6-QBL, 6-SSTO, 6-APD, 6-DBB, 6-MANN, 6-MCF
Mouse: NOD/MrkTac, NOD/ShiLtJ
Pig: Large White

Further information on our MHC annotation.

Leucocyte receptor complex (LRC) annotation

Non-reference regions:
Human: COX_1, COX_2, PGF_1, PGF_2, DM1A, DM1B, MC1A, MC1B.

Further information on our LRC annotation.

Our Data

- High-quality manual annotation
- Human annotation incorporated into GENCODE
- Rapid incorporation of new annotation
- Gene sets and regions of particular interest:
  - Genes with mouse knockout and human LOF transcripts
  - MHC and LRC regions
  - *Idd* candidate regions of NOD mice
- Inter- and intra-species comparative genomics
- Cross-referenced to other databases
- Complements Ensembl
- Downloadable datasets

What's New in release 55

- **Zebrafish Annotation Updated** (Zebrafish)
- **Regular Zebrafish Updates** (Zebrafish)

More news...

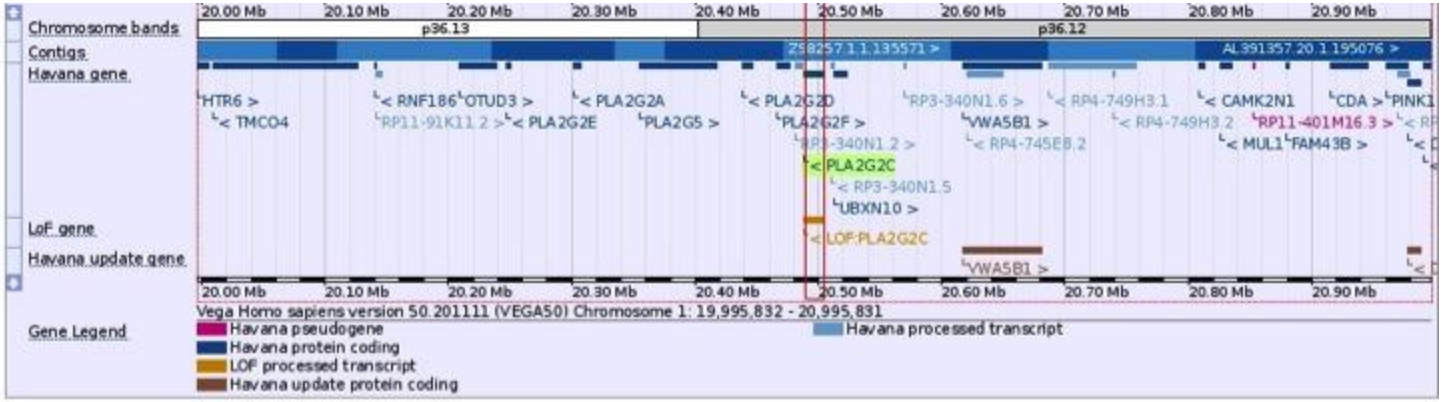
http://vega.sanger.ac.uk/index.html

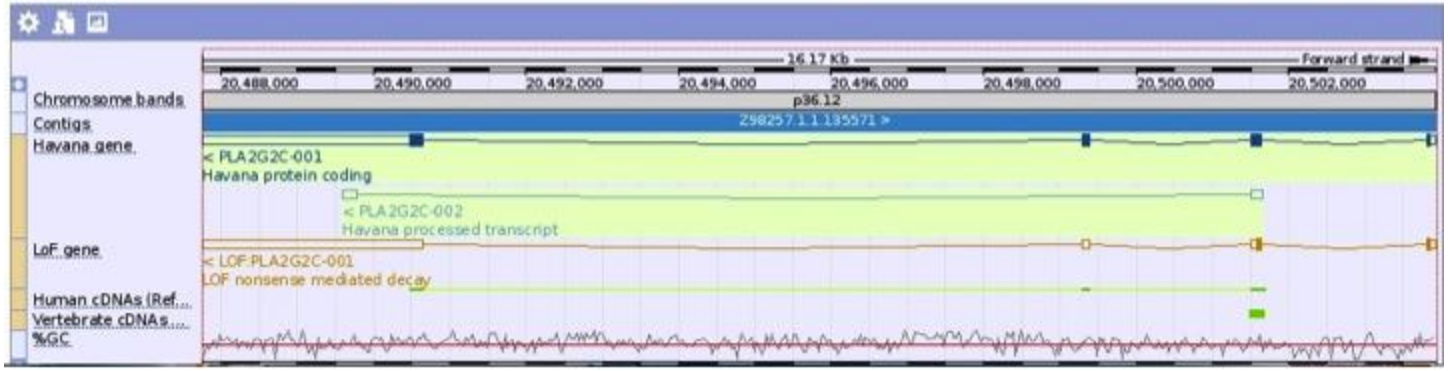
Havana annotation can also be browsed in Vega

Havana is an Ensembl style browser with further information on annotation including:

- Havana update
- Loss of function variants
- Regions not available in Ensembl e.g. MHC of Large White Pig, NOD mouse sequence.

Havana LoF variants





Havana update genes

Acknowledgements

Ensembl 2014

Paul Flicek[1,2,*], M. Ridwan Amode[2], Daniel Barrell[2], Kathryn Beal[1], Konstantinos Billis[2], Simon Brent[2], Denise Carvalho-Silva[1], Peter Clapham[2], Guy Coates[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García Girón[2], Leo Gordon[1], Thibaut Hourlier[2], Sarah Hunt[1], Nathan Johnson[1], Thomas Juettemann[1], Andreas K. Kähäri[2], Stephen Keenan[1], Eugene Kulesha[1], Fergal J. Martin[2], Thomas Maurel[1], William M. McLaren[1], Daniel N. Murphy[2], Rishi Nag[2], Bert Overduin[1], Miguel Pignatelli[1], Bethan Pritchard[2], Emily Pritchard[1], Harpreet S. Riat[2], Magali Ruffier[1], Daniel Sheppard[2], Kieron Taylor[1], Anja Thormann[1], Stephen J. Trevanion[2], Alessandro Vullo[1], Steven P. Wilder[1], Mark Wilson[2], Amonida Zadissa[1], Bronwen L. Aken[2], Ewan Birney[1], Fiona Cunningham[1], Jennifer Harrow[2], Javier Herrero[1], Tim J.P. Hubbard[2], Rhoda Kinsella[1], Matthieu Muffato[1], Anne Parker[2], Giulietta Spudich[1], Andy Yates[1], Daniel R. Zerbino[1] and Stephen M.J. Searle[2]

+ Jane Loveland and Jen Harrow from Havana

Funding

wellcome trust

European Commission Framework Programme 7

EMBL

National Human Genome Research Institute

BLUEPRINT epigenome

BBSRC bioscience for the future

EURATRANS

HELIX NEBULA THESCIENCECLOUD

SEVENTH FRAMEWORK PROGRAMME