Methods in genome annotation

Jacques Dainat, PhD
BILS genome annotation platform
Uppsala University

This lecture

1. Understanding gene annotation
2. The Maker2 annotation pipeline
3. The EnsEMBL annotation pipeline

The BILS annotation platform

1. Gene annotation

Jacques Dainat, PhD
BILS genome annotation platform

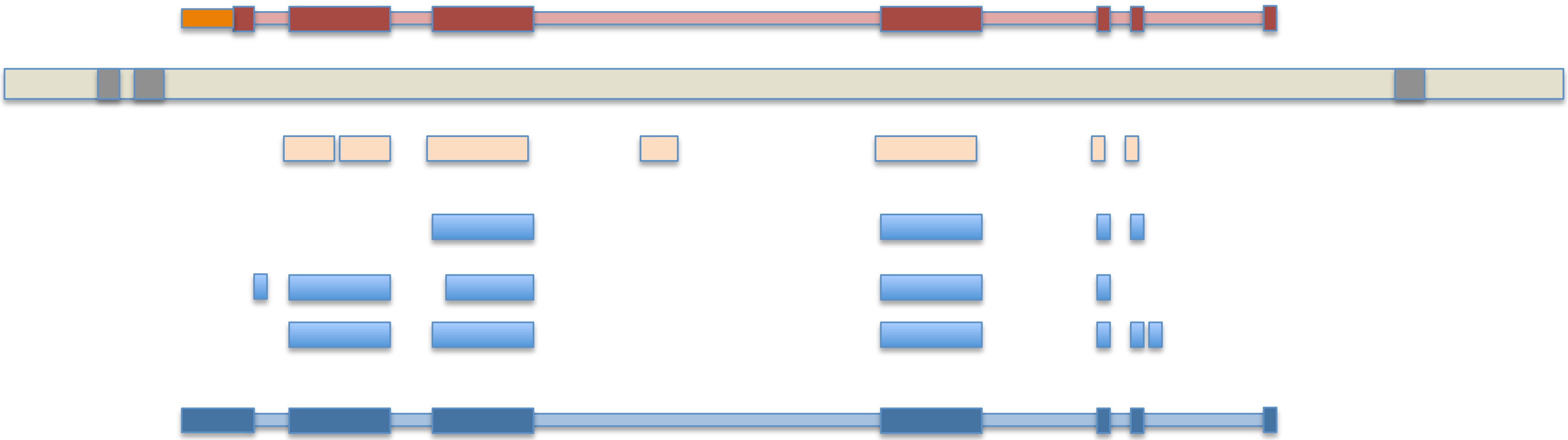
Overview

Annotation = combining different lines of evidence into gene models

Gene prediction – see the previous lecture

Evidence – see the previous lecture

Combining – the topic of this lecture

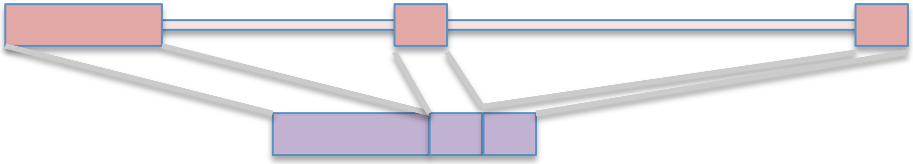


The BILS annotation platform

A bit of terminology first:

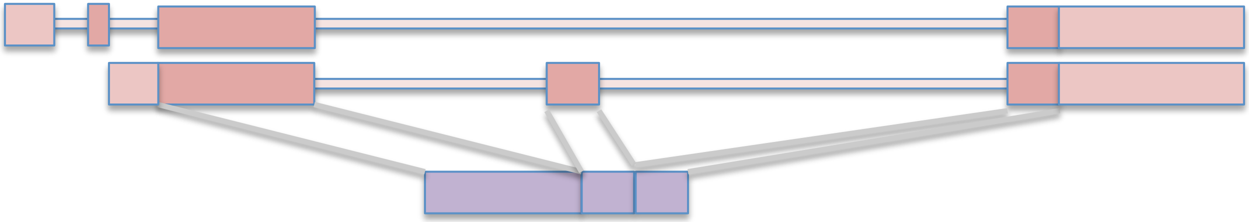
Gene prediction

Goal: Finding the single most likely coding sequence (CDS)



Gene annotation

Goal: Identify the entire gene structure



The BILS annotation platform

In recent years, the distinction between ab initio prediction and gene annotation has been blurred

Gene annotation ~ Gene building

2. The Maker2 annotation pipeline

Jacques Dainat, PhD
BILS genome annotation platform

Existing annotation pipelines – MAKER2

Maker – developed as an easy-to-use alternative to other pipelines

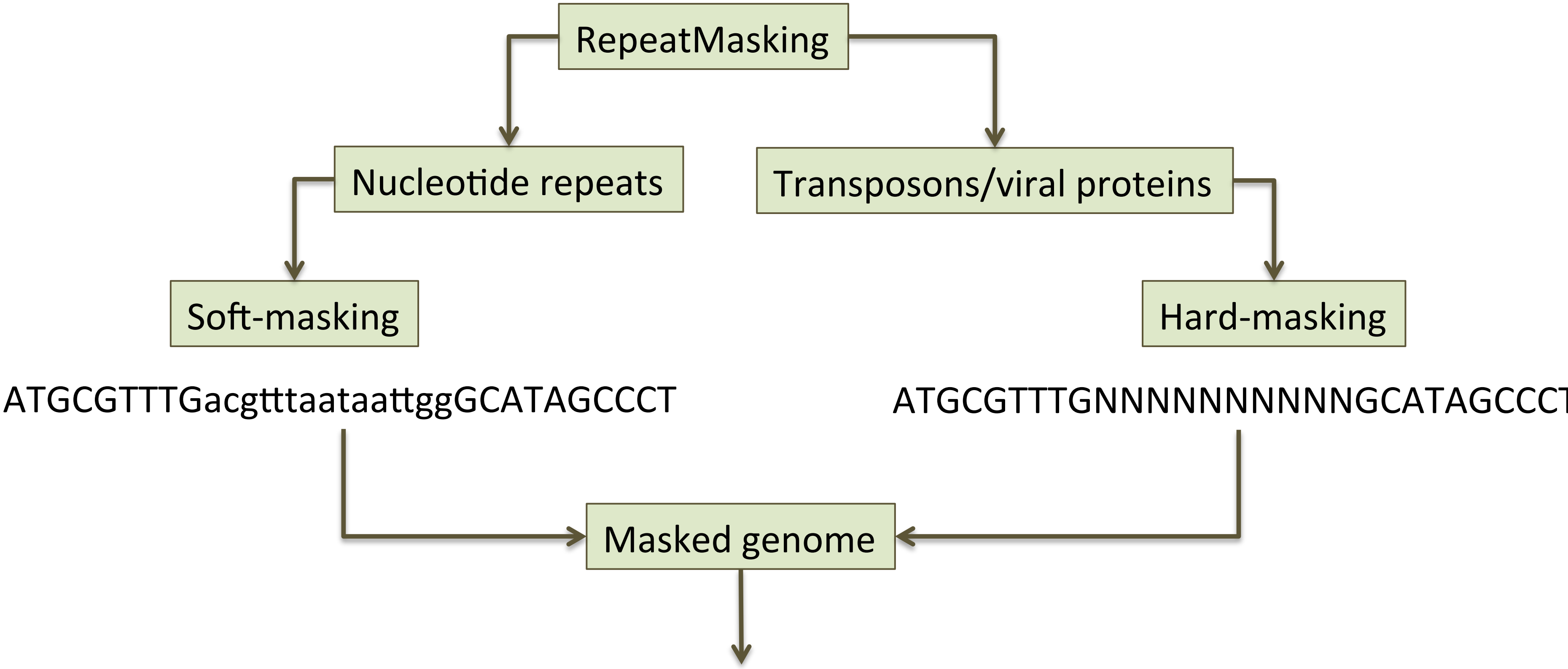
Advantages over competing solutions:

- Almost unlimited parallelism built-in (limited by data and hardware)
- Largely independent from the underlying system where it is run on
- Everything is run through one command, no manual combining of data/outputs
- Follows common standards, produces GMOD compliant output
- Annotation Edit Distance (AED) metric for improved quality control
- Provides a mechanism to train and retrain ab initio gene predictors
- Annotations can be updated by re-launching Maker with new evidences

But how does Maker work exactly?

Existing annotation pipelines – MAKER2

Step 1: Raw compute phase



RepeatMasking

Nucleotide repeats

Transposons/viral proteins

Soft-masking

Hard-masking

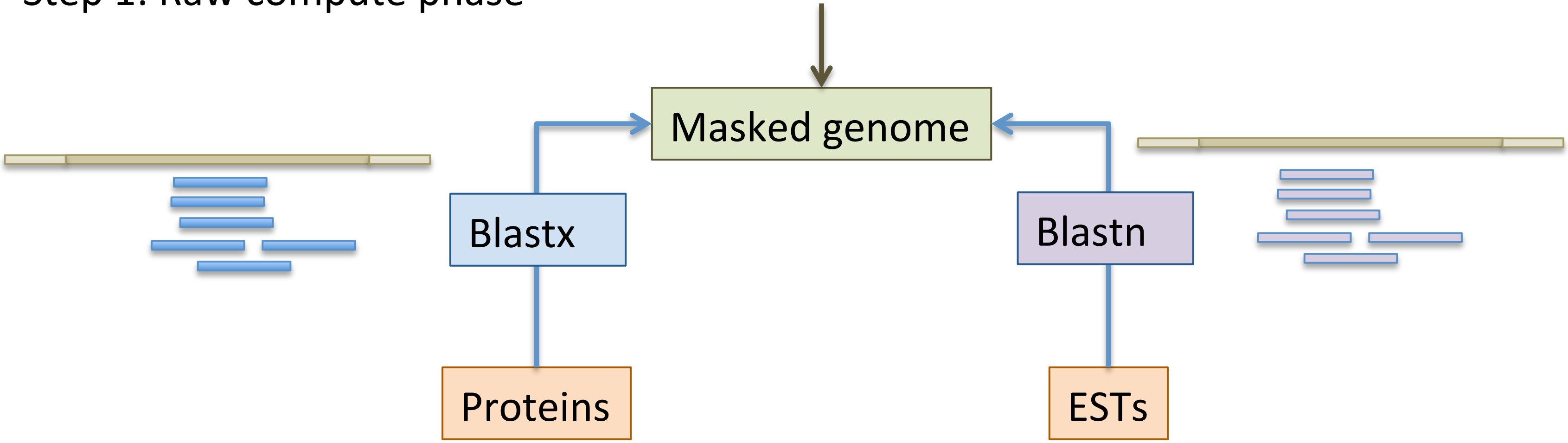
ATGCGTTTGacgtttaataattggGCATAGCCCT

ATGCGTTTGNNNNNNNNNNGCATAGCCCT

Masked genome

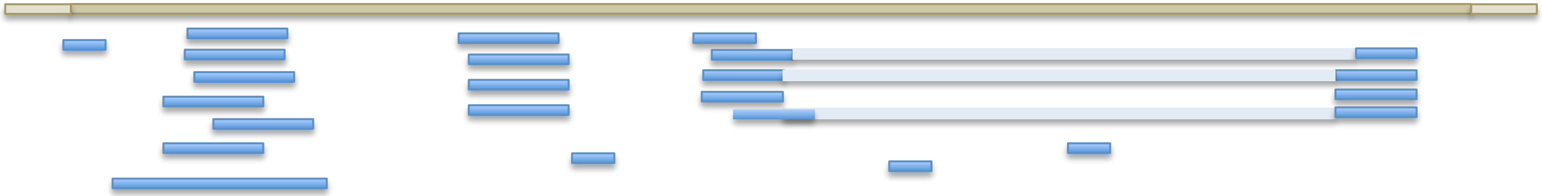
Existing annotation pipelines – MAKER2

Step 1: Raw compute phase



Existing annotation pipelines – MAKER2

Step 2: Filter and cluster alignments



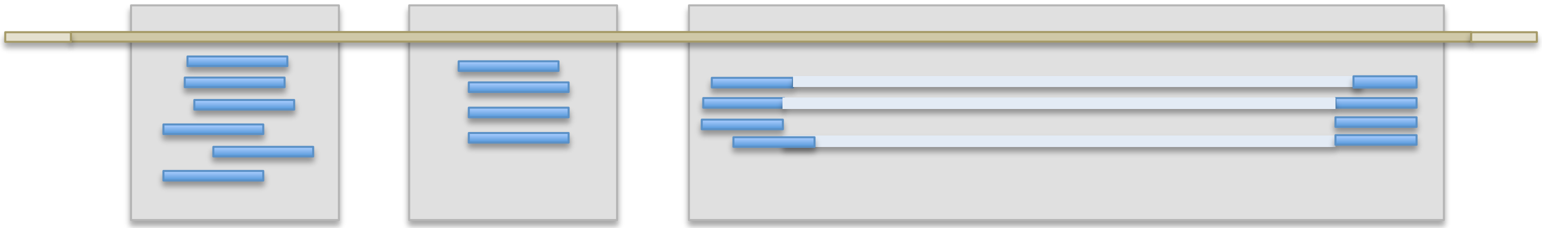
Filtering is based on rules defined in the Maker configuration for a given project

Example: EST alignment – 80% coverage and 85% identity

Default settings sensible for most projects, but can be changed!

Existing annotation pipelines – MAKER2

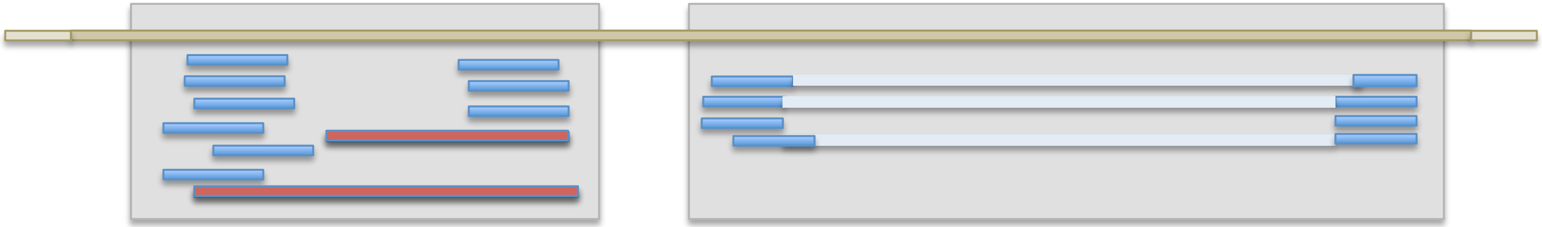
Step 2: Filter and cluster alignments



Clustering groups evidence alignments into ’loci’

Existing annotation pipelines – MAKER2

Step 2: Filter and cluster alignments

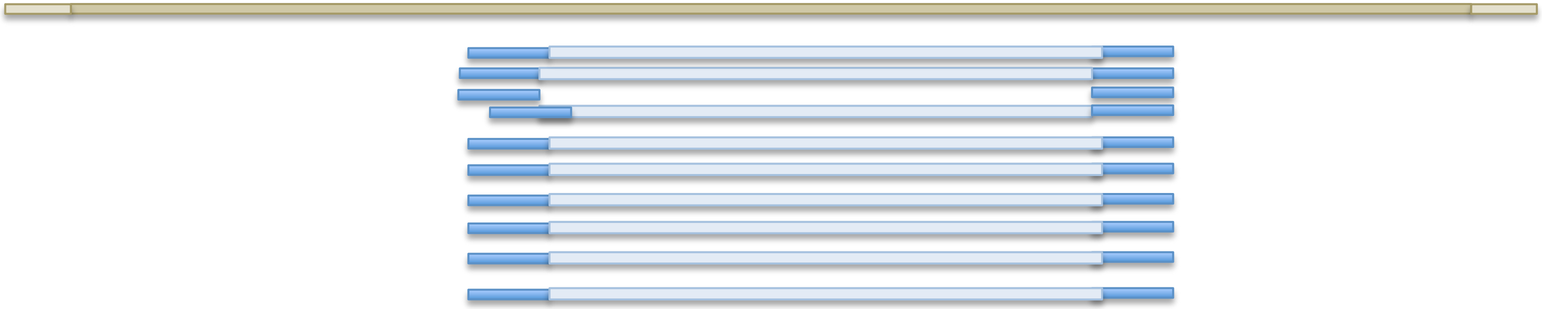


Problematic data can complicate clustering

Needs to be fixed by a) cleaner data or b) manual curation

Existing annotation pipelines – MAKER2

Step 2: Filter and cluster alignments



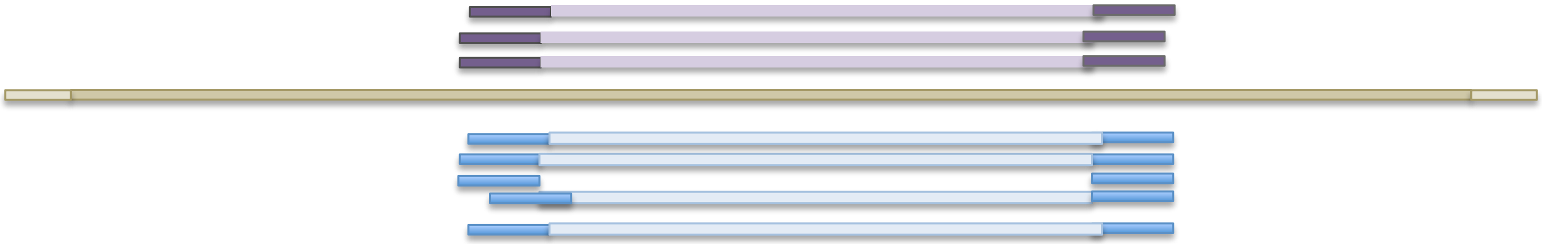
Clustering groups evidence alignments into ’loci’

Amount of data in any given cluster is then collapsed to remove redundancy

Threshold for the collapsing is also user-definable

Existing annotation pipelines – MAKER2

Step 2: Filter and cluster alignments



Clustering groups evidence alignments into ’loci’

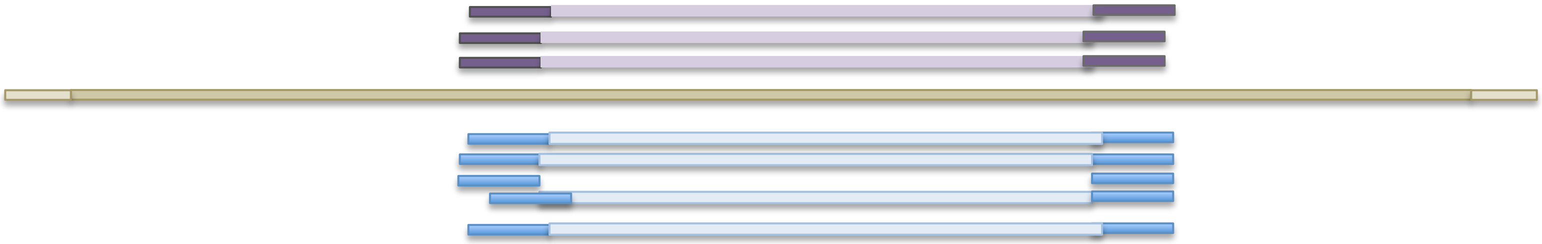
Amount of data in any given cluster is then collapsed to remove redundancy

Threshold for the collapsing is also user-definable

Performed for all lines of evidence

Existing annotation pipelines – MAKER2

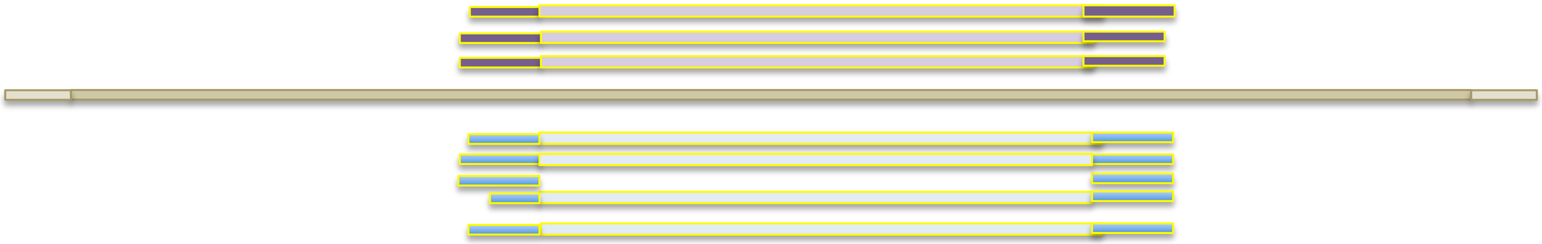
Step 3: Polishing alignments



Blast-based alignments are only approximations, need to be refined

Existing annotation pipelines – MAKER2

Step 3: Polishing alignments

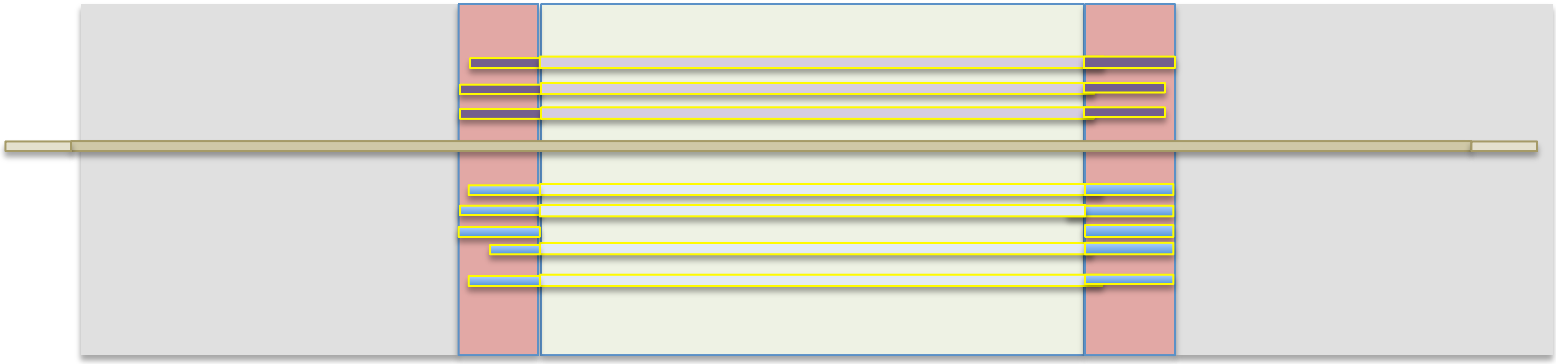


Blast-based alignments are only approximations, need to be refined

Exonerates is used to create splice-aware alignments

Existing annotation pipelines – MAKER2

Step 4: Synthesis



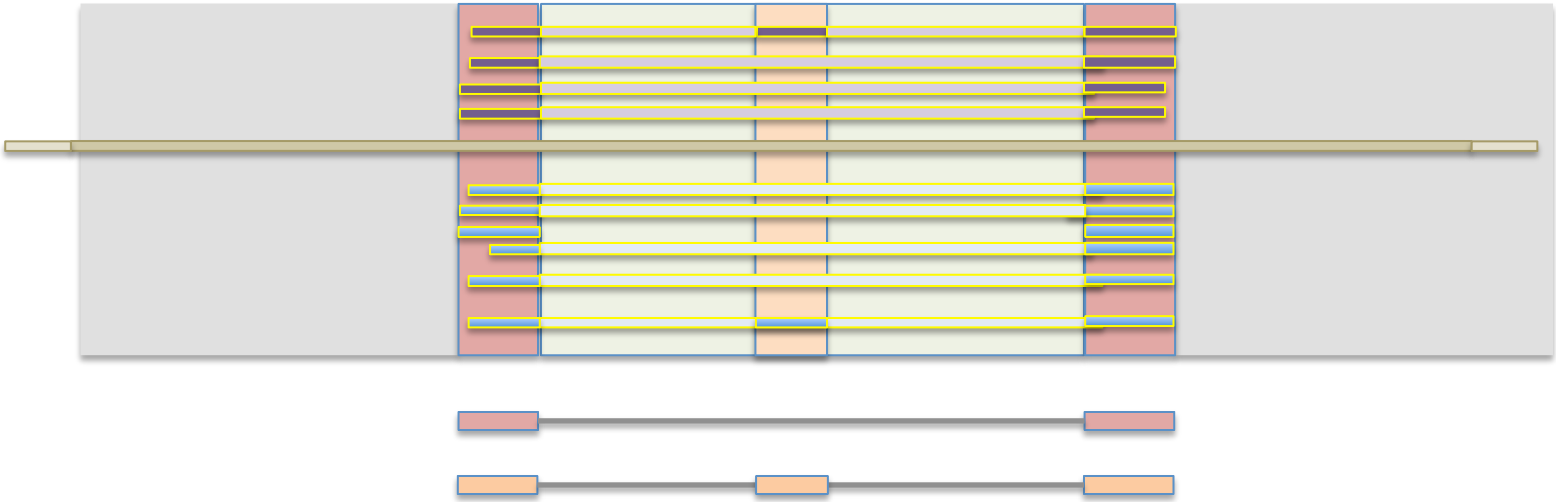
Synthesis refers to the extraction of information to generate evidence for annotations

Done by identifying genomic regions overlapping with sequence features

The BILS annotation platform

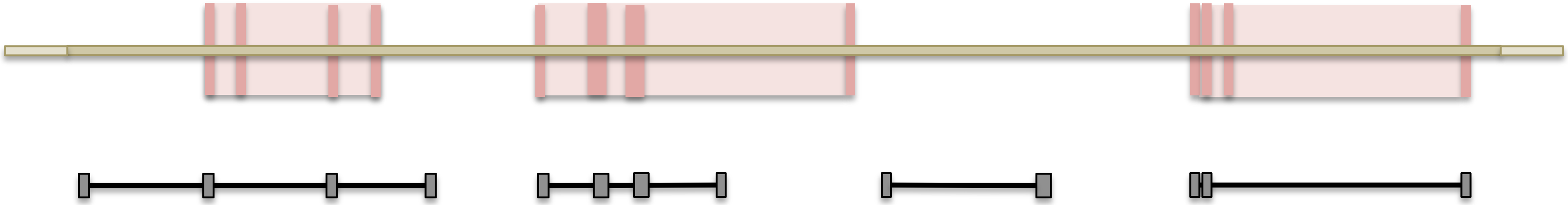
Existing annotation pipelines – MAKER2

Step 4: Synthesis



Existing annotation pipelines – MAKER2

Step 4: Synthesis...and ab-initio gene finding



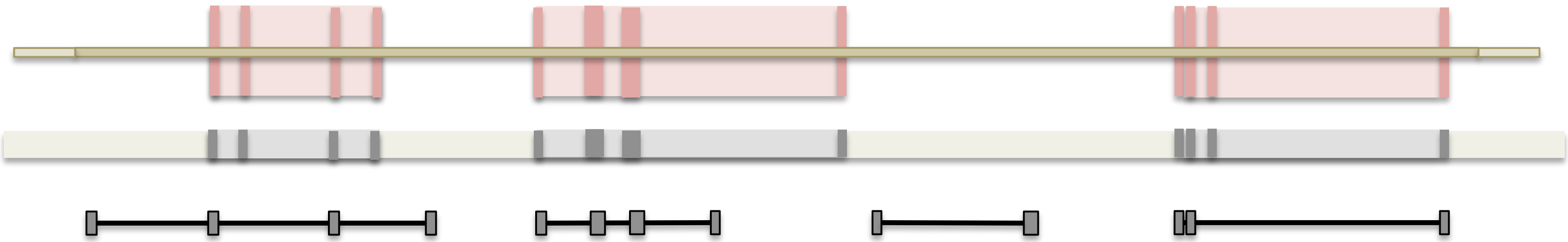
Evidence alignments provide support for the identifcation of gene loci

Ab-initio predictions can enhance these signals and fill gaps with no evidence

The BILS annotation platform

Existing annotation pipelines – MAKER2

Step 4: Synthesis...and ab-initio gene finding

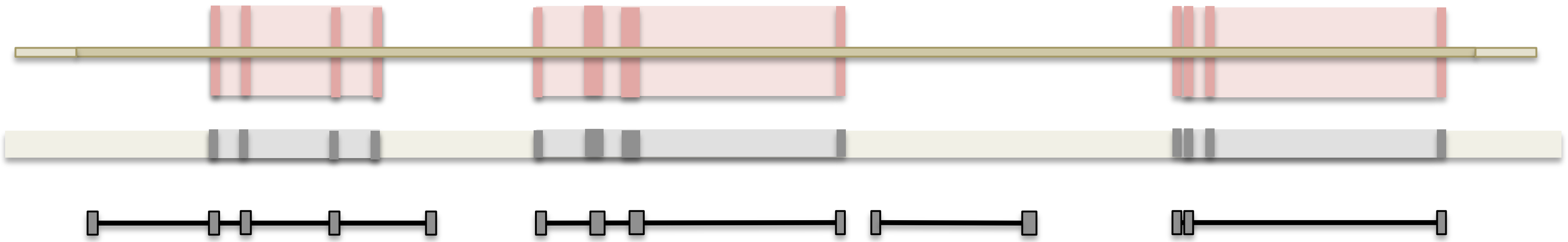


Ab-intio predictions can be improved when evidence is provided (hints)

Help refine and calibrate a computational inference for a given lovis

Existing annotation pipelines – MAKER2

Step 4: Synthesis...and ab-initio gene finding



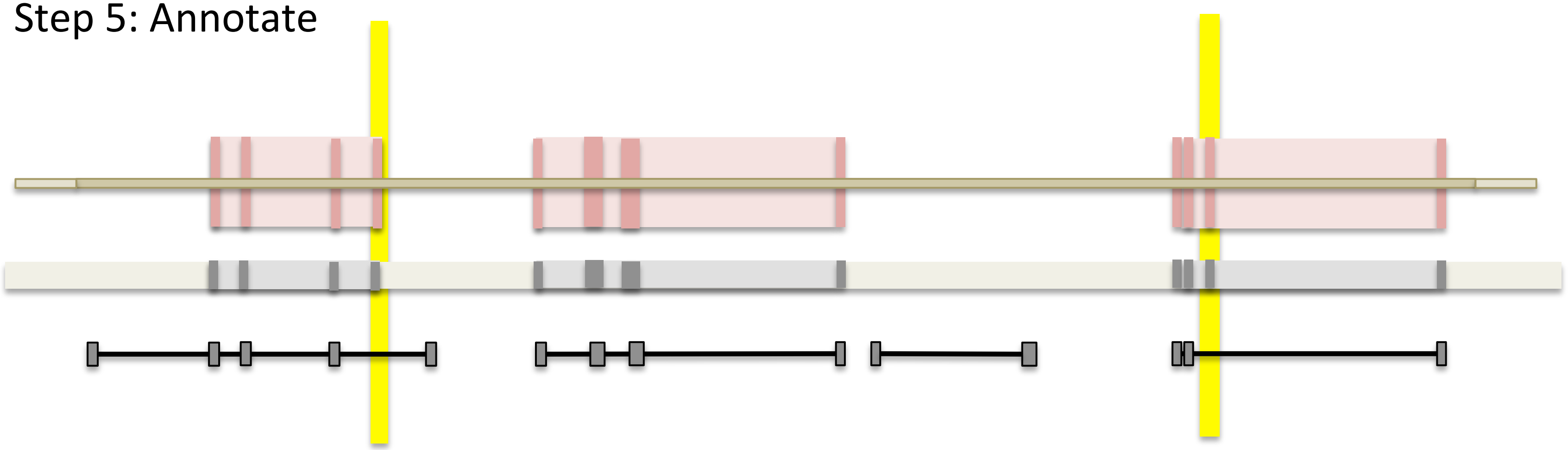
Ab-intio predictions can be improved when evidence is provided (hints)

Help refine and calibrate a computational inference for a given lovis

Hints: Introns, intergenic sequence, CDS

Existing annotation pipelines – MAKER2

Step 5: Annotate



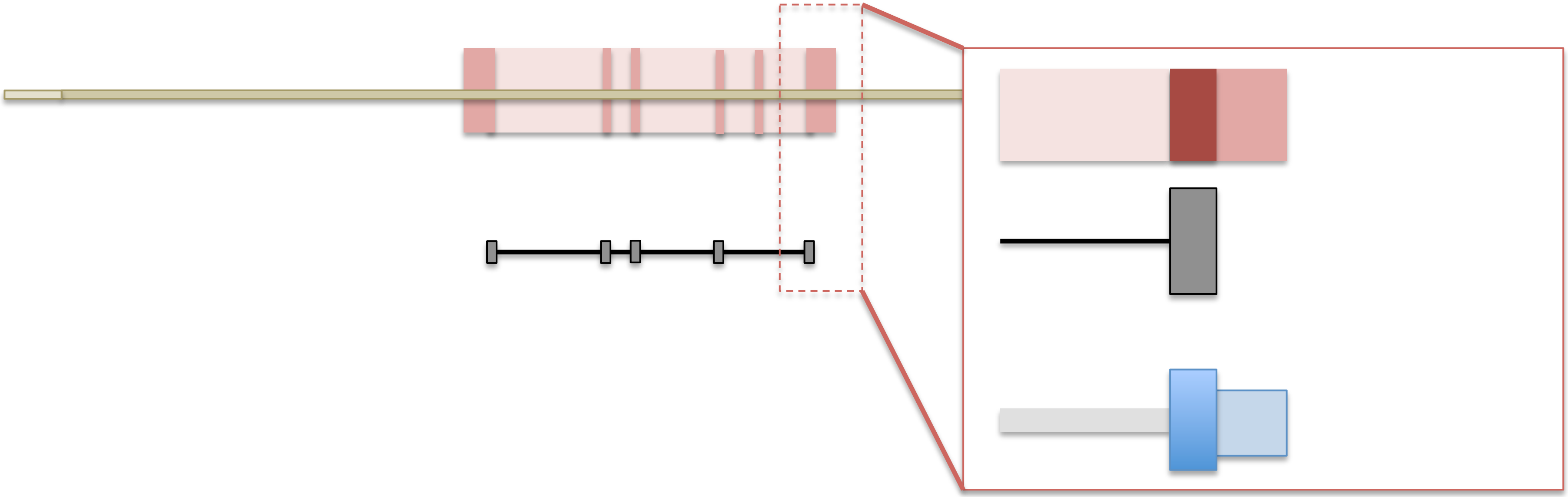
Refined ab-initio models may still be incomplete / partially wrong

Need to reconcile with evidence so we don’t miss information

-> Limited by agreement between ab-initio profile and evidence

Existing annotation pipelines – MAKER2

Step 5: Annotate

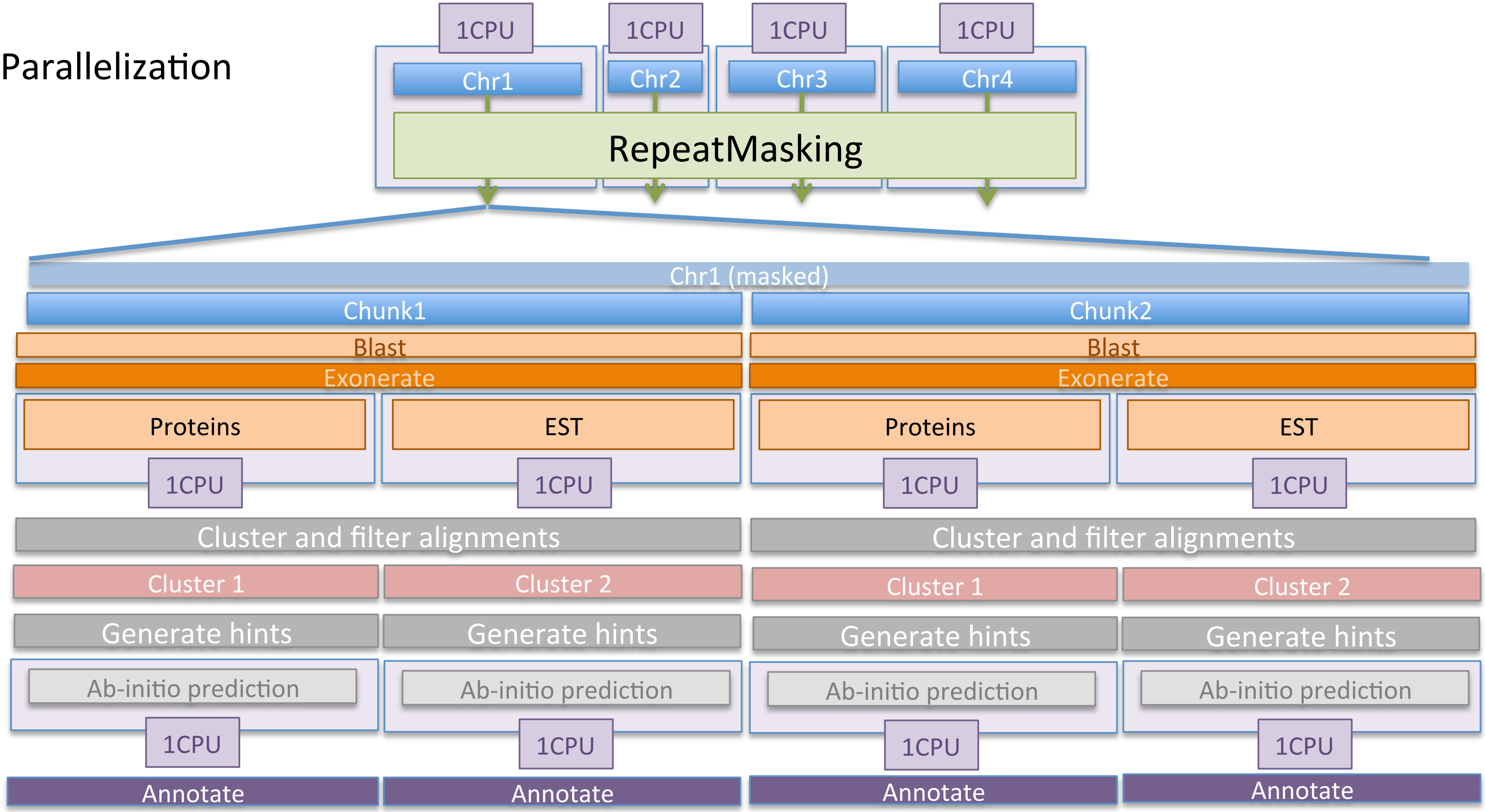


Synthesized transcript structures are compared against evidence to find UTRs

The BILS annotation platform

Existing annotation pipelines – MAKER2

Parallelization



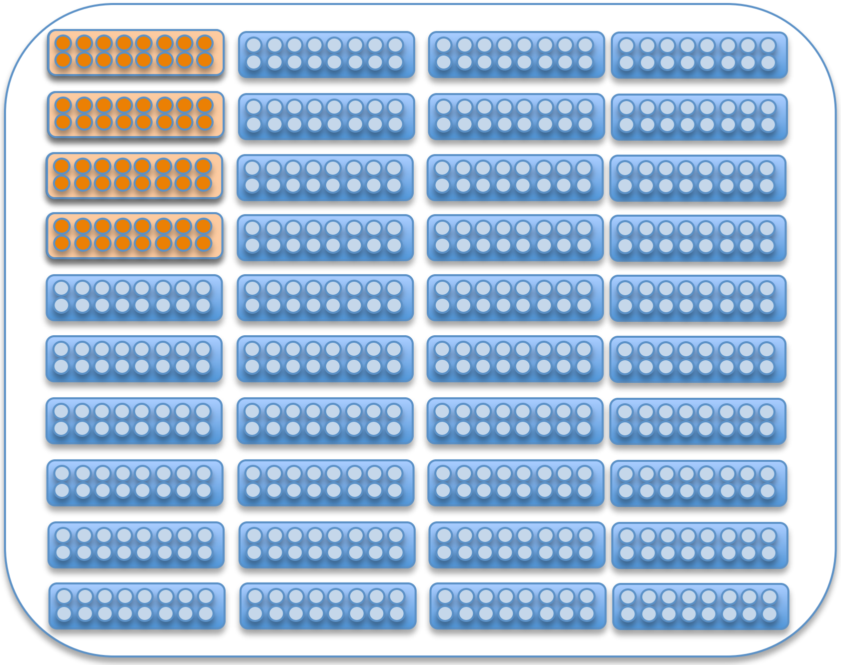
Existing annotation pipelines – MAKER2

Parallelization – Running on Uppmax

Maker uses MPI for job distributon

- runs on almost all computing platforms

- Operates on cores, not nodes



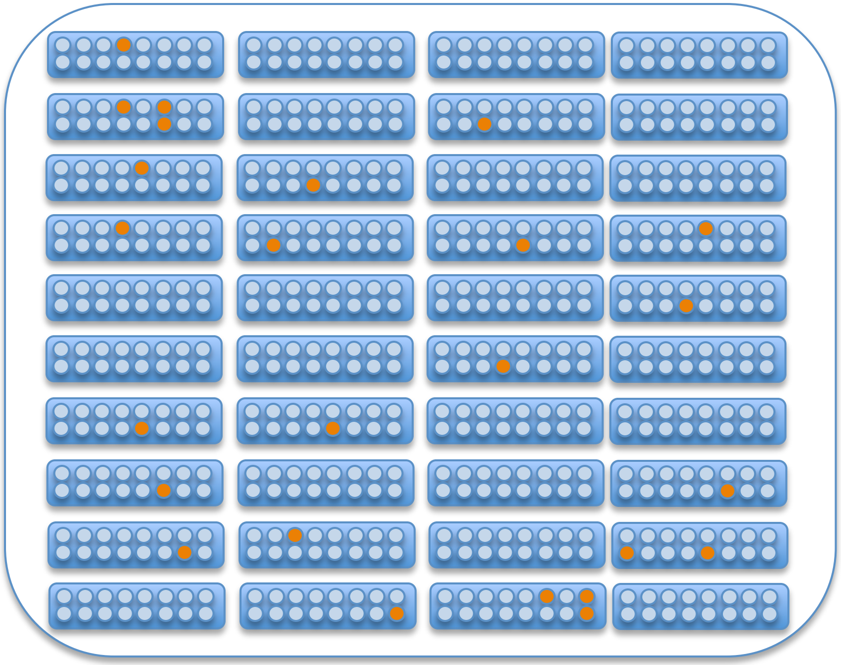
Existing annotation pipelines – MAKER2

Parallelization – Running on Uppmax

Maker uses MPI for job distributon

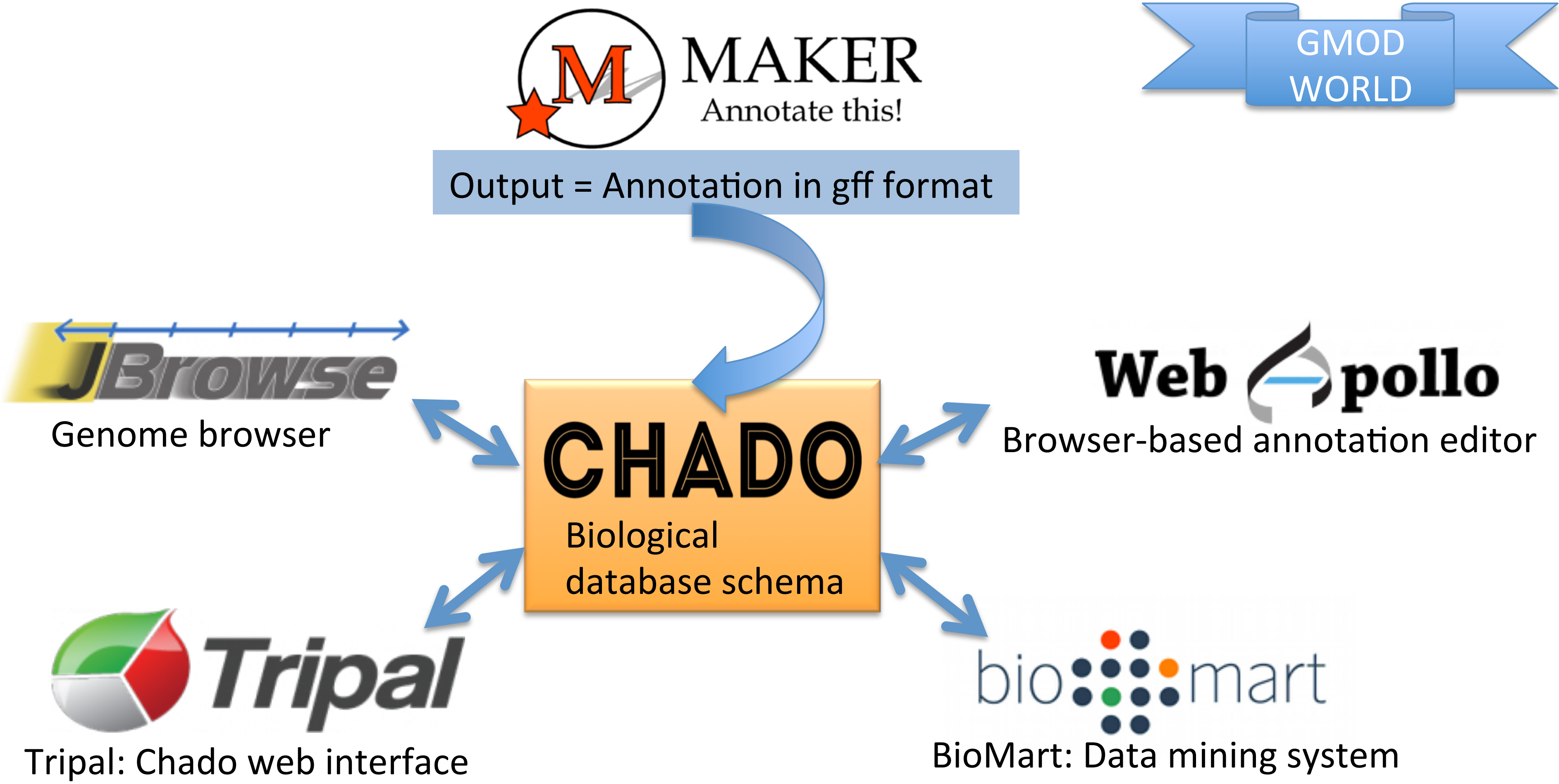
- runs on almost all computing platforms

- Operates on cores, not nodes



The BILS annotation platform

Existing annotation pipelines – MAKER2



Output = Annotation in gff format

CHADO

Biological database schema

Genome browser

Browser-based annotation editor

Tripal: Chado web interface

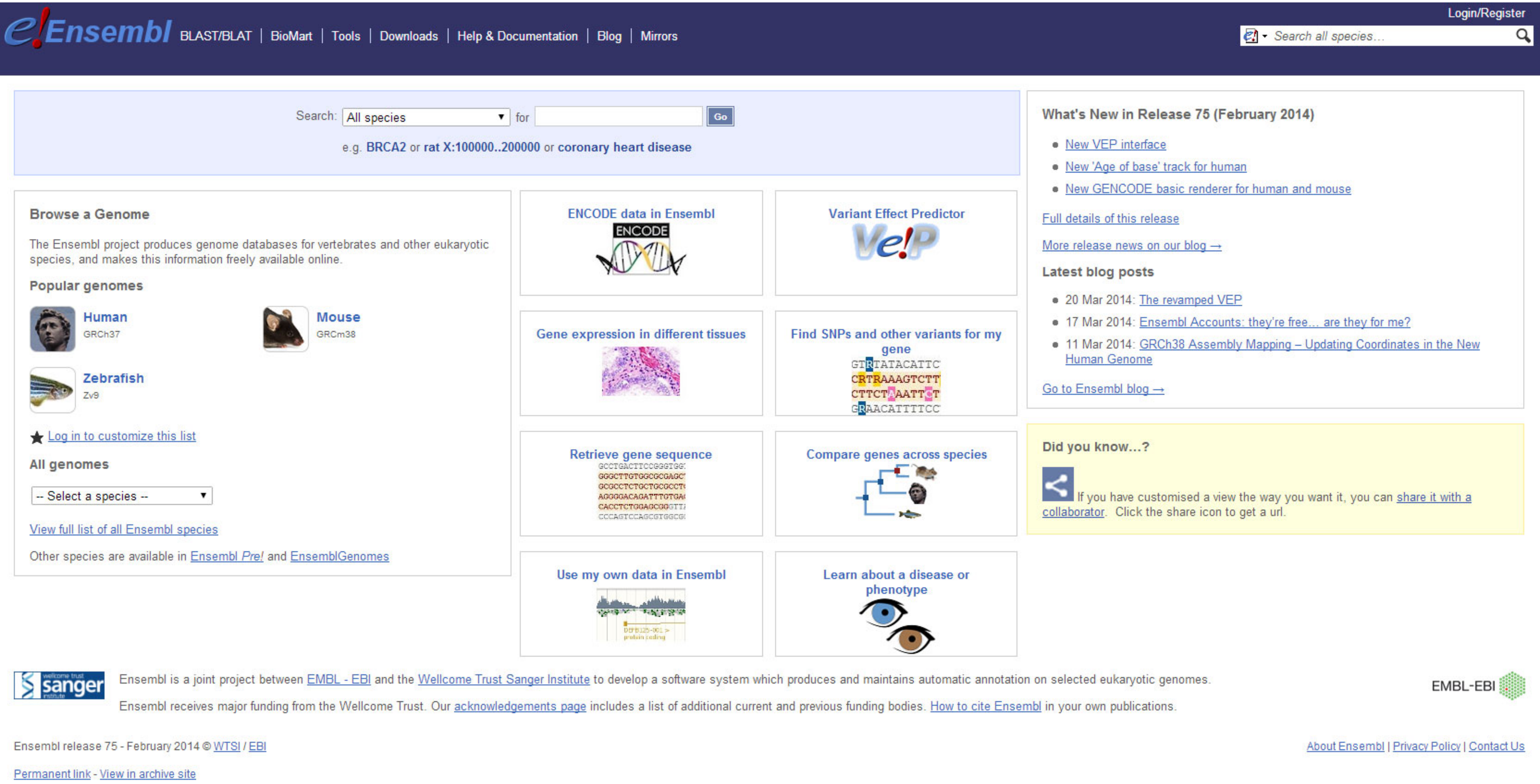
BioMart: Data mining system

3. The EnsEMBL gene annotation pipeline

Jacques Dainat, PhD
BILS genome annotation platform

The BILS annotation platform

BILS
Bioinformatics Infrastructure for Life Sciences



e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login/Register

Search all species...

Search: All species ▾ for [ ] Go

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh37	Mouse GRCm38
Zebrafish Zv9	

★ Log in to customize this list

All genomes

-- Select a species -- ▾

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

ENCODE data in Ensembl

Variant Effect Predictor

Gene expression in different tissues

Find SNPs and other variants for my gene

Retrieve gene sequence

Compare genes across species

Use my own data in Ensembl

Learn about a disease or phenotype

What's New in Release 75 (February 2014)

- New VEP interface
- New 'Age of base' track for human
- New GENCODE basic renderer for human and mouse

Full details of this release

More release news on our blog →

Latest blog posts

- 20 Mar 2014: The revamped VEP
- 17 Mar 2014: Ensembl Accounts: they're free… are they for me?
- 11 Mar 2014: GRCh38 Assembly Mapping – Updating Coordinates in the New Human Genome

Go to Ensembl blog →

Did you know...?

If you have customised a view the way you want it, you can share it with a collaborator. Click the share icon to get a url.

Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies. How to cite Ensembl in your own publications.

sanger EMBL-EBI

Ensembl release 75 - February 2014 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

Permanent link - View in archive site

Existing annotation pipelines – EnsEMBL gene build pipeline

EnsEMBL – an overview

Perhaps the largest project in the world to deliver annotations

Originally created to support the annotation effort for the human genome

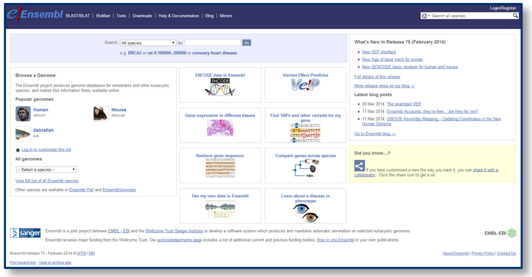
Pipelines and infrastructure have since been applied to a range of other species

- Strong focus on vertebrates
- Forked projects include Gramene (plant annotation), Wormbase, ...

The BILS annotation platform

Existing annotation pipelines – EnsEMBL gene build pipeline

EnsEMBL – an overview



e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...
Search: All species for Go
ENCODE
Ve!P

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh37

Mouse GRCm38

Zebrafish Zv9

★ Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl Pre! and EnsemblGenomes

ENCODE data in Ensembl

Variant Effect Predictor

Gene expression in different tissues

Find SNPs and other variants for my gene

Retrieve gene sequence

Compare genes across species

Use my own data in Ensembl

Learn about a disease or phenotype

What's New in Release 75 (February 2014)

- New VEP interface
- New 'Age of base' track for human
- New GENCODE basic renderer for human and mouse

Full details of this release

More release news on our blog →

Latest blog posts

- 20 Mar 2014: The revamped VEP
- 17 Mar 2014: Ensembl Accounts: they're free... are they for me?
- 11 Mar 2014: GRCh38 Assembly Mapping – Updating Coordinates in the New Human Genome

Go to Ensembl blog →

Did you know...?

If you have customised a view the way you want it, you can share it with a collaborator. Click the share icon to get a url.

sanger — Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies. How to cite Ensembl in your own publications. EMBL-EBI

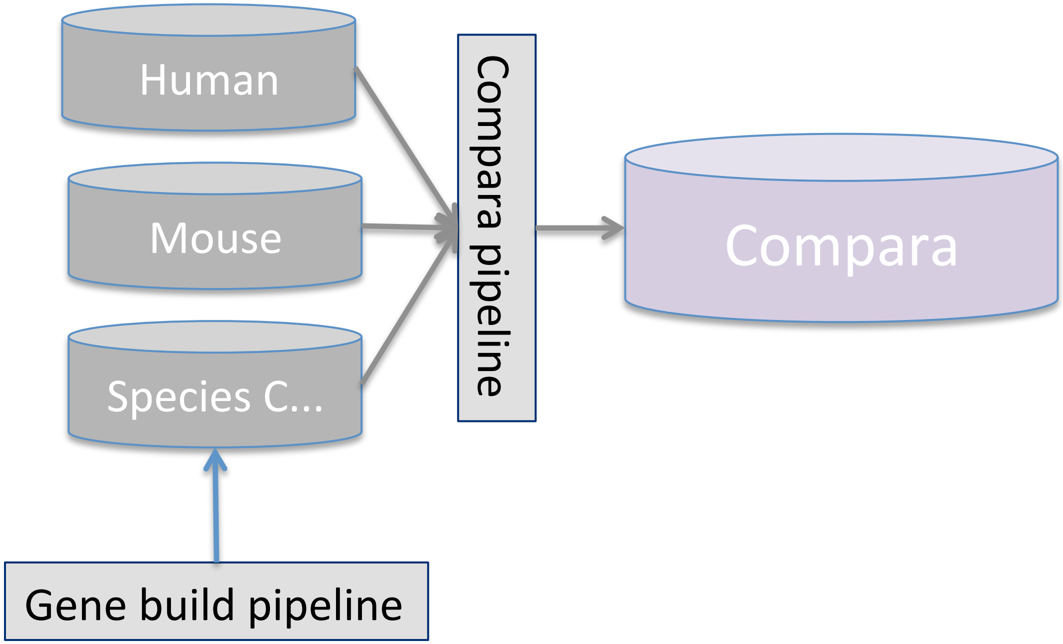
Ensembl release 75 - February 2014 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

Permanent link - View in archive site

biomart

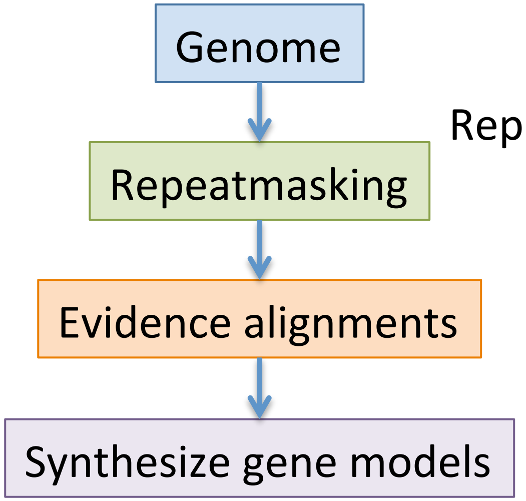
Perl API



Existing annotation pipelines – EnsEMBL gene build pipeline

Comparing EnsEMBL with Maker (and other pipelines)

A lot of commonalities



Genome

Repeatmasking

Evidence alignments

Synthesize gene models

RepeatMasker

Pmatch

GeneWise / Exonerate

Blastx, Blastn
And
Exonerate

Existing annotation pipelines – EnsEMBL gene build pipeline

How does EnsEMBL differ from e.g. Maker?

0. Setting up an annotation project

Config file needs to be written ’manually’

Pipeline logic needs to be specified ’manually’

Requires a total of 3 MySQL databases to be set up prior to starting

Stores assembly in layers (contigs, scaffold, chromosomes – via AGP file)

Existing annotation pipelines – EnsEMBL gene build pipeline

How does EnsEMBL differ from e.g. Maker?

1. Gene building

Uses reference gene sequences as additional evidence

Does NOT use ab-initio gene predictions during gene building (in most cases...)
= purely evidence-based

Combining and clustering of evidence is layered

Automatically patches suspected sequencing errors (cDNA read-through)

Generally does not try to annotate isoforms

Pipeline for ncRNA annotation is available (for select taxonomic groups)

Existing annotation pipelines – EnsEMBL gene build pipeline

How does EnsEMBL differ from e.g. Maker?

2. Additional analyses

Can be configured to perform down-stream analyses

- Annotation of protein domains
- Mapping of gene names
- Cross-referencing with other databases

Existing annotation pipelines – EnsEMBL gene build pipeline

How does EnsEMBL differ from e.g. Maker?

3. Output

Annotation file not a primary output, but a database filled with information

→ Much more complex, but also more powerful

Existing annotation pipelines – EnsEMBL gene build pipeline

How does EnsEMBL differ from e.g. Maker?

4. Miscellaneous

EnsEMBL provides no tools for manual curation

Parallelization is done via SGE or LSF (cannot be run on just any system)

Documentation is very patchy

Only limited training opportunities

Existing annotation pipelines – EnsEMBL gene build pipeline

When to use EnsEMBL

- You need access to the EnsEMBL infrastructure (webcode, API, data structure)
- You have access to a cluster with LSF/SGE
- Investing weeks/months into learning the pipeline is ’worth it’ for you project

Closing remarks / What’s next?

The BILS annotation platform

What’s next

Computational pipelines make mistakes

- Need to be run very conservatively (EnsEMBL) or require **manual curation**

- Pipelines like Maker only build gene models, no **functional inference**