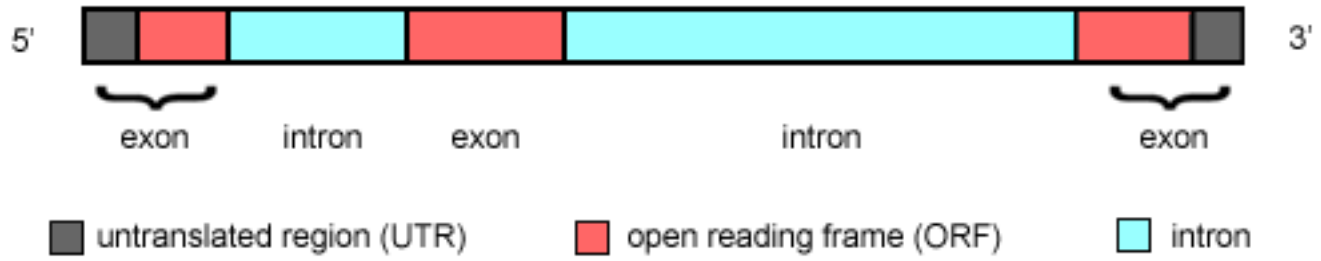
Combining everything

MAKER 2 pipeline

There are two major parts of annotation

- 1) Structural: Find out where the regions of interest (usually genes) are in the genome and what they look like. How many exons/introns? UTRs? Isoforms?

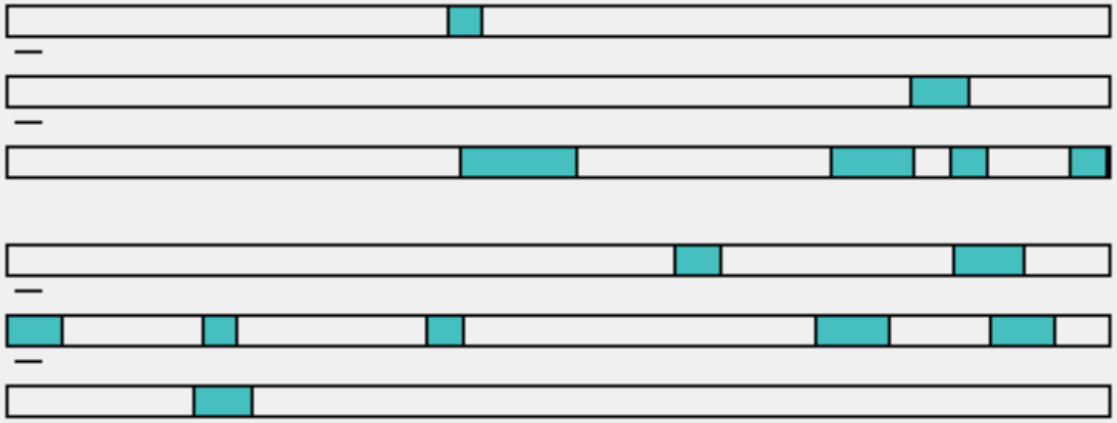


- 2) Functional: Find out what the regions do. What do they code for?

Open reading frames

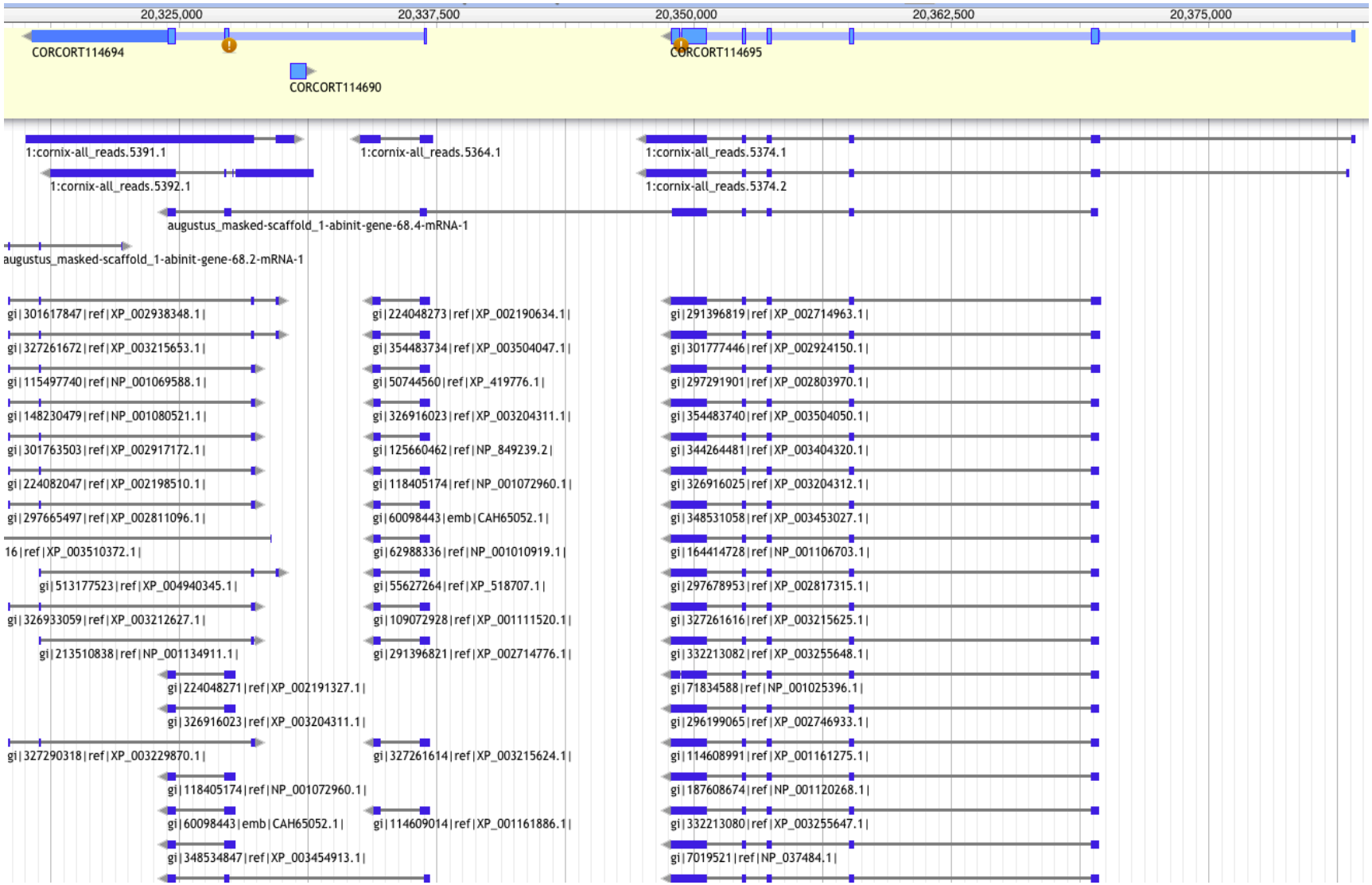
Anonymous

View | 1 GenBank | Redraw | 100 | SixFrames



Frame	from	to	Length
+3	1383	1736	354
+3	2511	2756	246
-2	2465	2686	222
-1	2880	3092	213
-2	2996	3187	192
-3	574	753	180
+2	2753	2929	177
-2	2	172	171
-1	2034	2171	138
+3	3237	3349	114
+3	2874	2984	111
-2	1283	1393	111
+1	1345	1446	102
-2	599	700	102

Difficult in practice



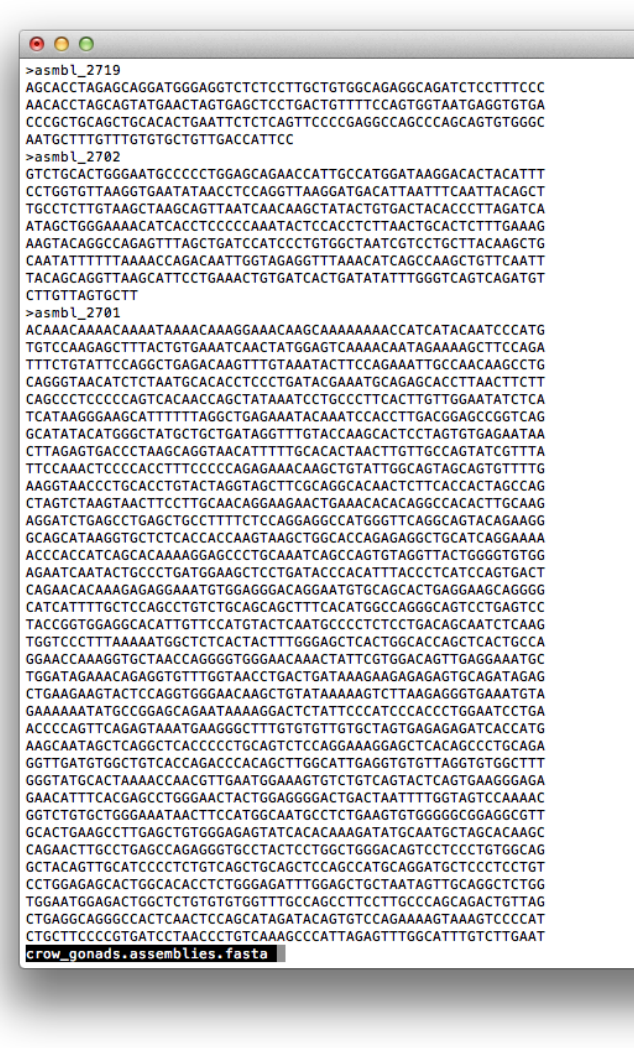
Combine data - use Maker!

- External data - proteins, rna-seq (incl. ESTs)
- Ab-initio gene finders
- (Lift-overs from closely related genomes)

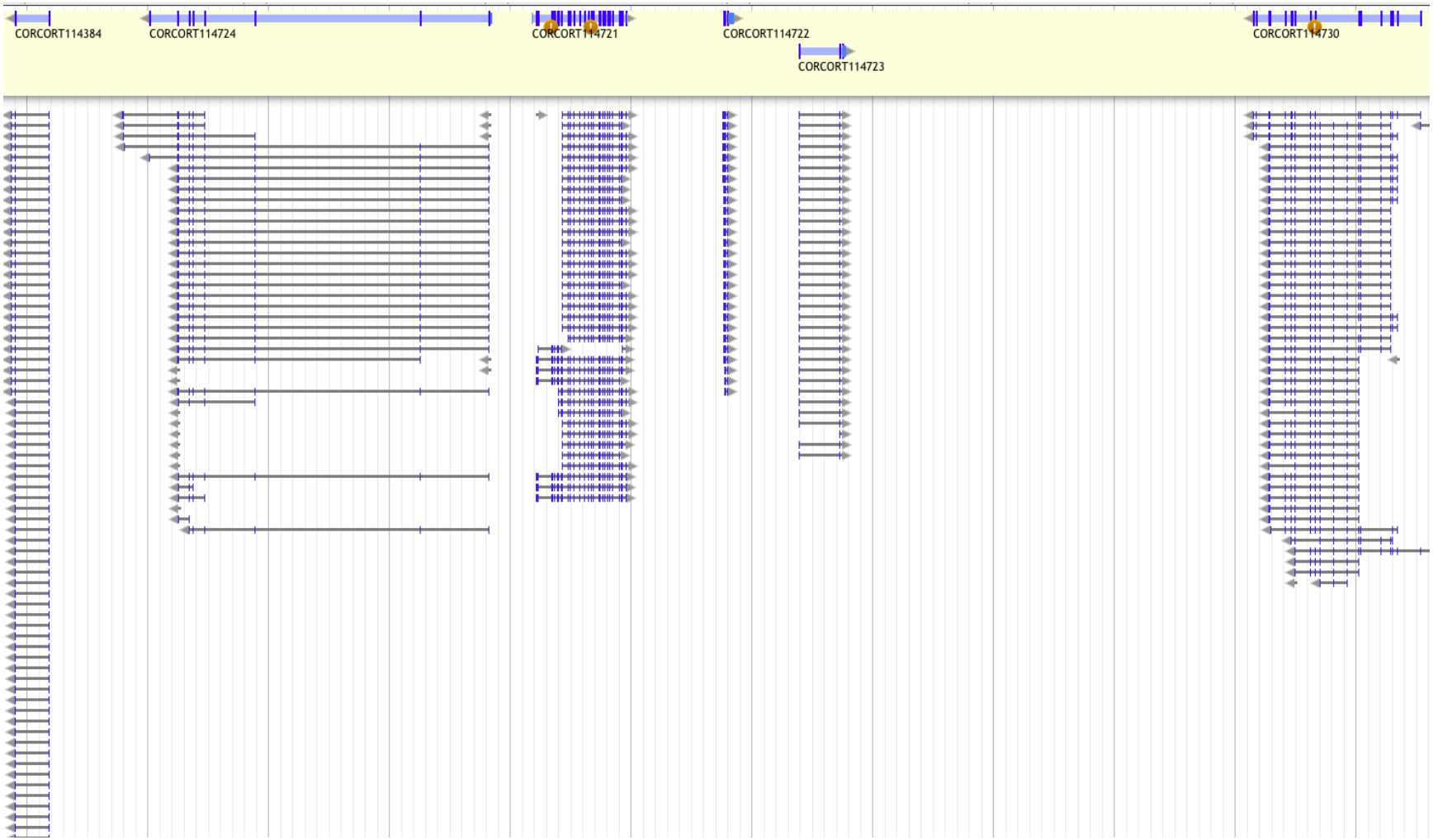
Combined annotation

Transcriptomes are different but have their own challenges

- No introns, but where are the start and stop codons?
- Still needs functional annotation



Data used - Proteins



Data used - Proteins

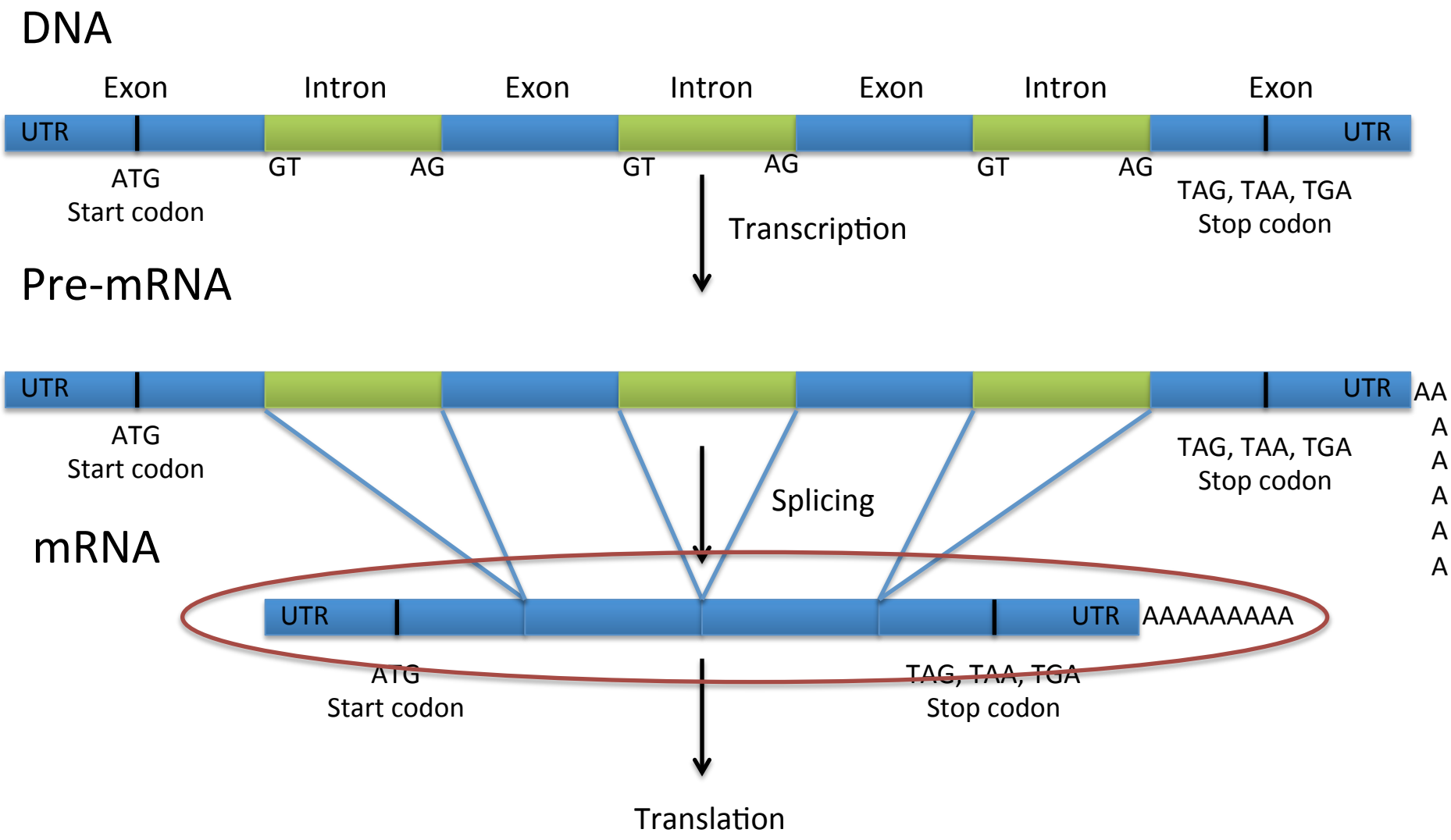
- Conserved in sequence => conserved annotation with little noise
- Proteins from model organisms often used => bias?
- Proteins can be incomplete => problems as many annotation procedures are heavily dependent on protein alignments

>ENSTGUP00000017616 pep:novel chromosome:taeGut3.2.4:8_random:2849599:2959678:-1 gene:ENSTGUG00000017338 transcript:ENSTGUT000000180
RSPNATEYNWHHLRYPKIPERLNPPAAAGPALSTAEGWMLPWGNGQHPLLARAPGKGRER
DGKELIKKPKTFKFTFLKKKKKKKKTFK
>ENSTGUP00000017615 pep:novel chromosome:taeGut3.2.4:23_random:205321:209117:1 gene:ENSTGUG00000017337 transcript:ENSTGUT00000018017
PDLRELVLMFEHLHRVRNGGFRNSEVKKWPDRSPPPYHSFTPAQKSFSLAGCSGESTKMG
IKERMRLSSSQRQGSRGRQQHLGPPLHRSPSPEDVAEATSPTKVQKSWSFNDRTRFRASL
RLKPRIPAEGDCPPEDSGEERSSPCDLTFEDIMPAVKTLIRAVRILKFLVAKRKFKETLR
PYDVKDVIEQYSAGHLDMLGRIKSLQTRVEQIVGRDRALPADKKVREKGEKPALEAELVD
ELSMMGRVVKVERQVQSIEHKLDLLLGLYSRCLRKGSANSLVLAAVRVPPGEPDVTSDYQ
SPVEHEDISTSAQSLSISRLASTNMD

Data used - Proteins

- Maker will align proteins for you: Blast -> Exonerate
- Blast is not structure aware, Exonerate is (splice sites, start/stop codons)
- Preferred file-format: fasta

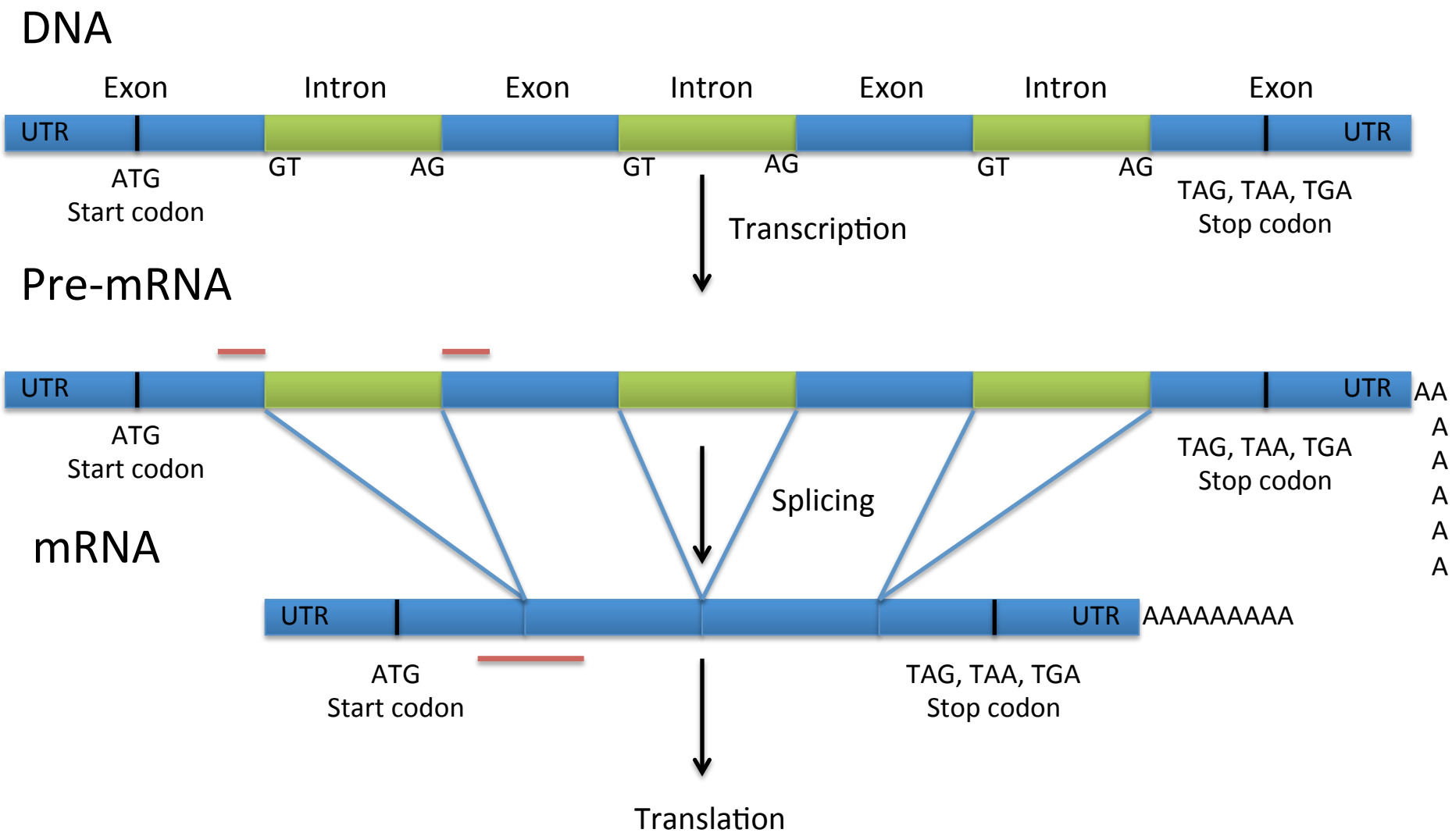
RNA-seq


DNA
Exon
Intron
Exon
Intron
Exon
Intron
Exon
UTR
UTR
GT
AG
GT
AG
GT
AG
ATG
Start codon
TAG, TAA, TGA
Stop codon
Transcription
Pre-mRNA
UTR
UTR
AA
A
A
A
A
A
ATG
Start codon
TAG, TAA, TGA
Stop codon
Splicing
mRNA
UTR
UTR
AAAAAAAAA
ATG
Start codon
TAG, TAA, TGA
Stop codon
Translation

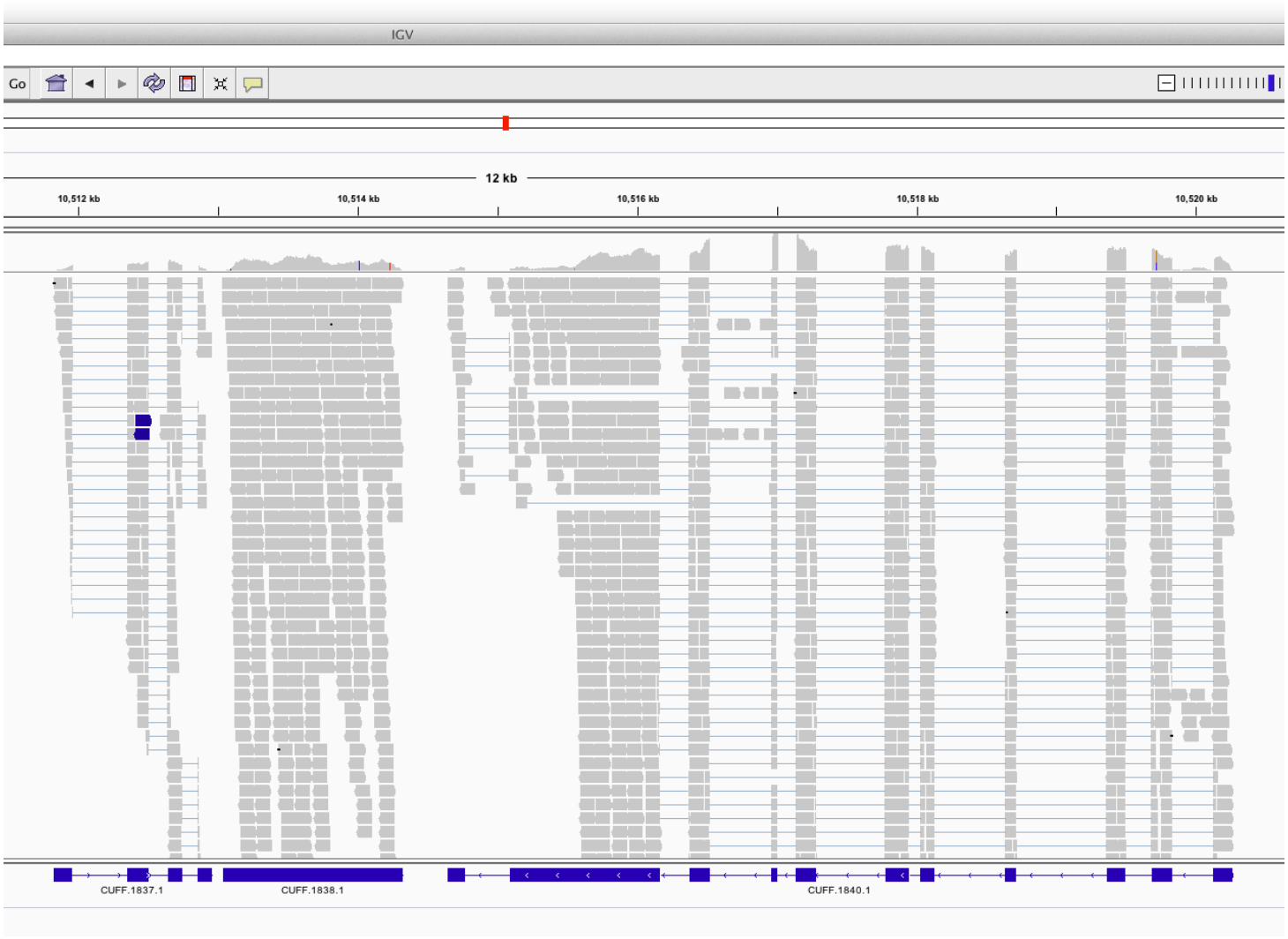
Data used - RNA-seq

- Should always be included in an annotation project
- From the same organism as the genomic data => unbiased
- Can be very noisy (tissue/species dependent), can include pre-mRNA
- PASA, or some other filtering method, often needed

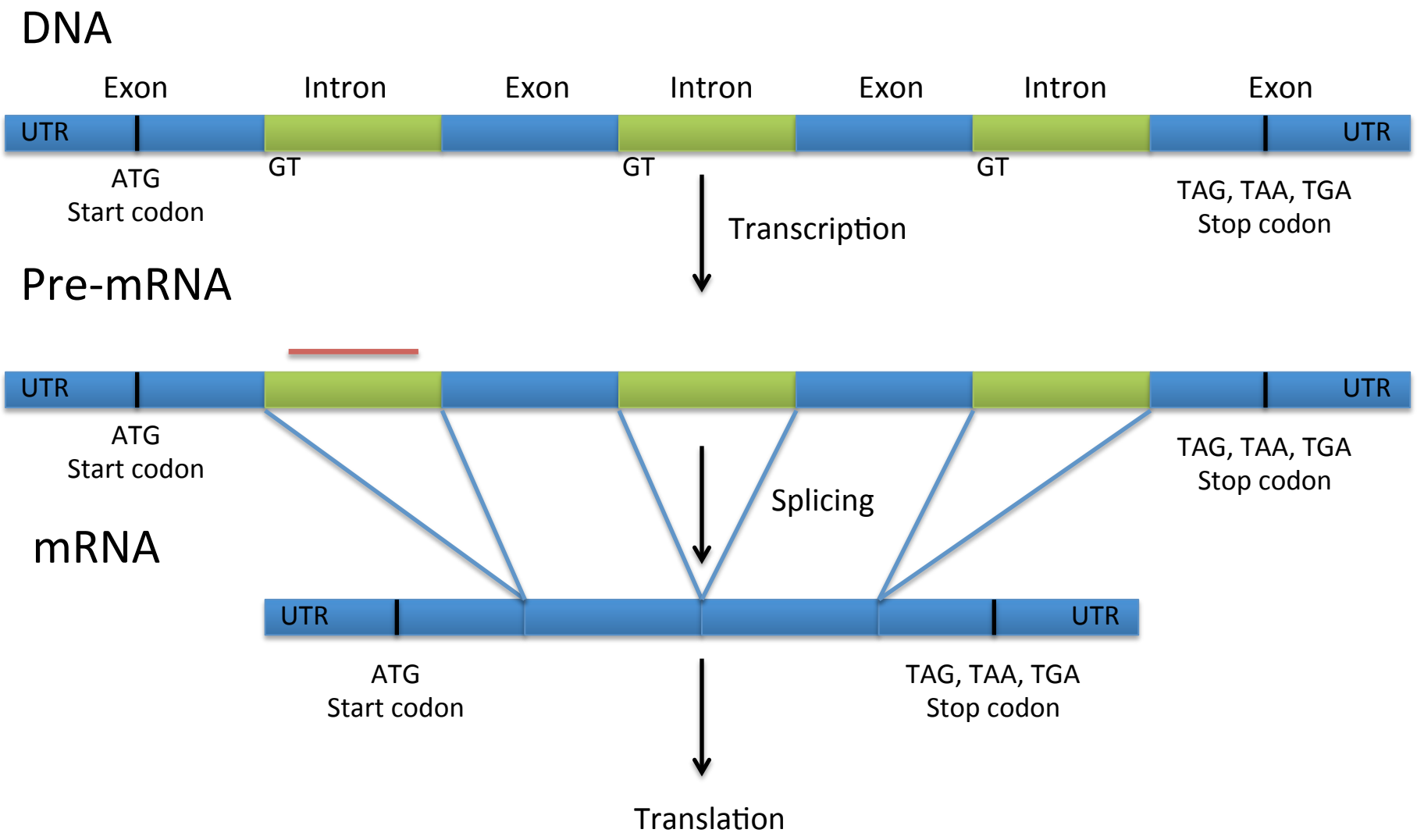
Spliced reads



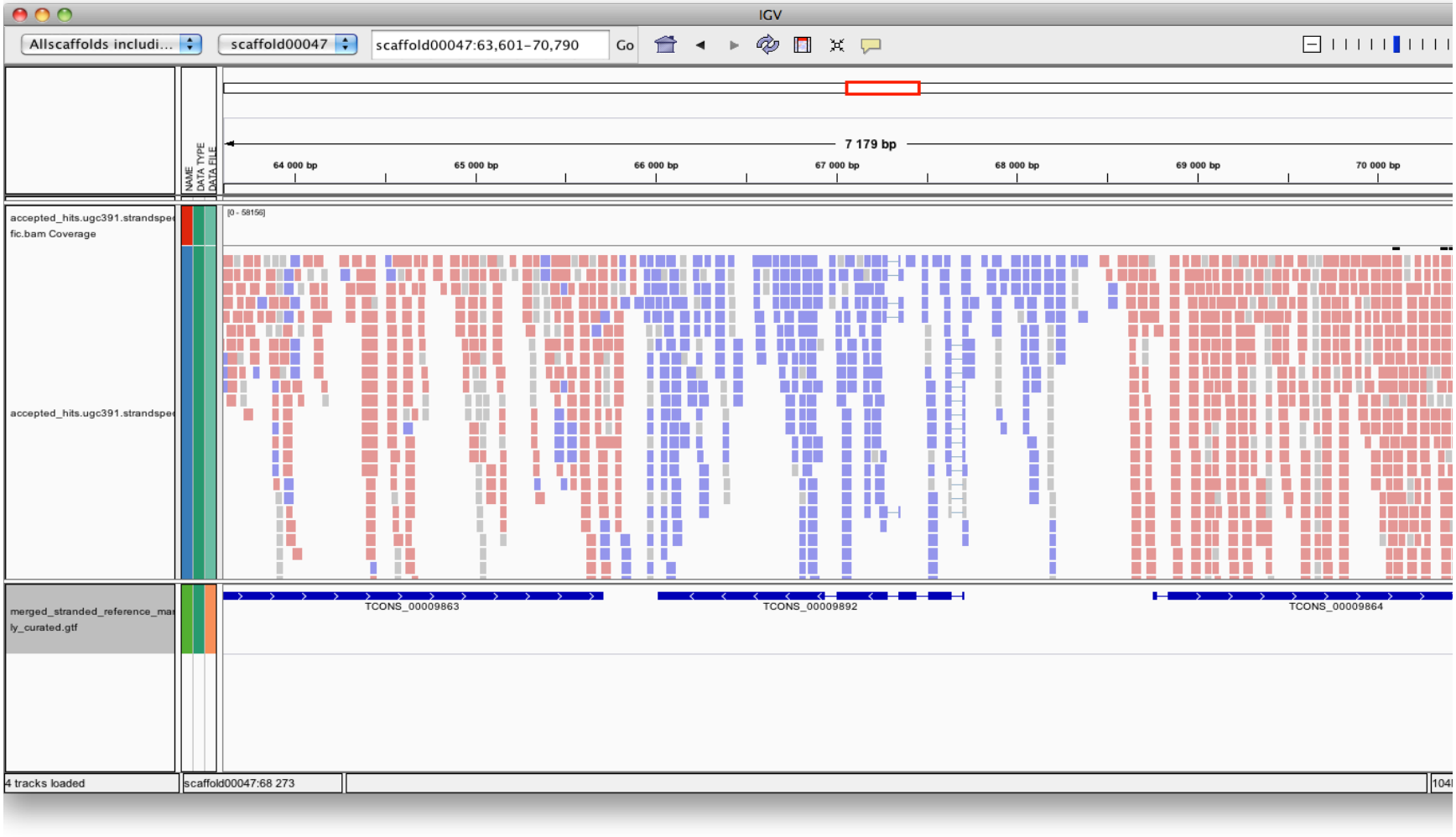
RNA-seq - Spliced reads



Pre-mRNA

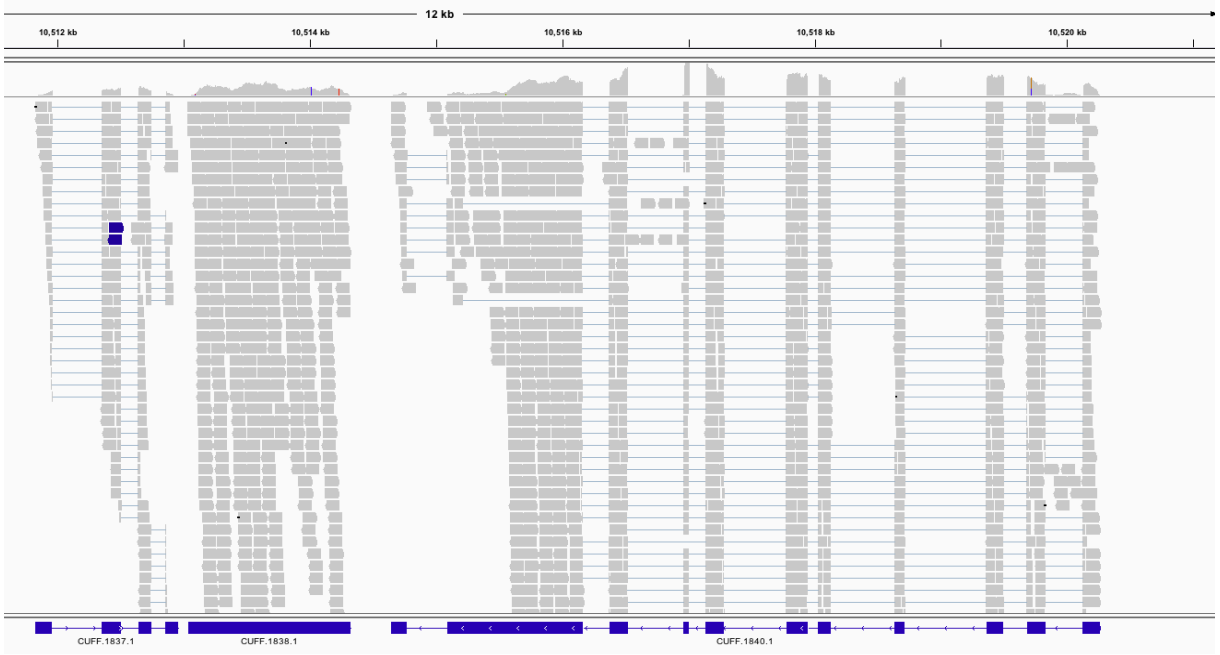

DNA
Exon
Intron
Exon
Intron
Exon
Intron
Exon
UTR
UTR
GT
GT
GT
ATG
Start codon
TAG, TAA, TGA
Stop codon
Transcription
Pre-mRNA
UTR
UTR
ATG
Start codon
TAG, TAA, TGA
Stop codon
Splicing
mRNA
UTR
UTR
ATG
Start codon
TAG, TAA, TGA
Stop codon
Translation

Stranded RNA-seq



How to use RNA-seq

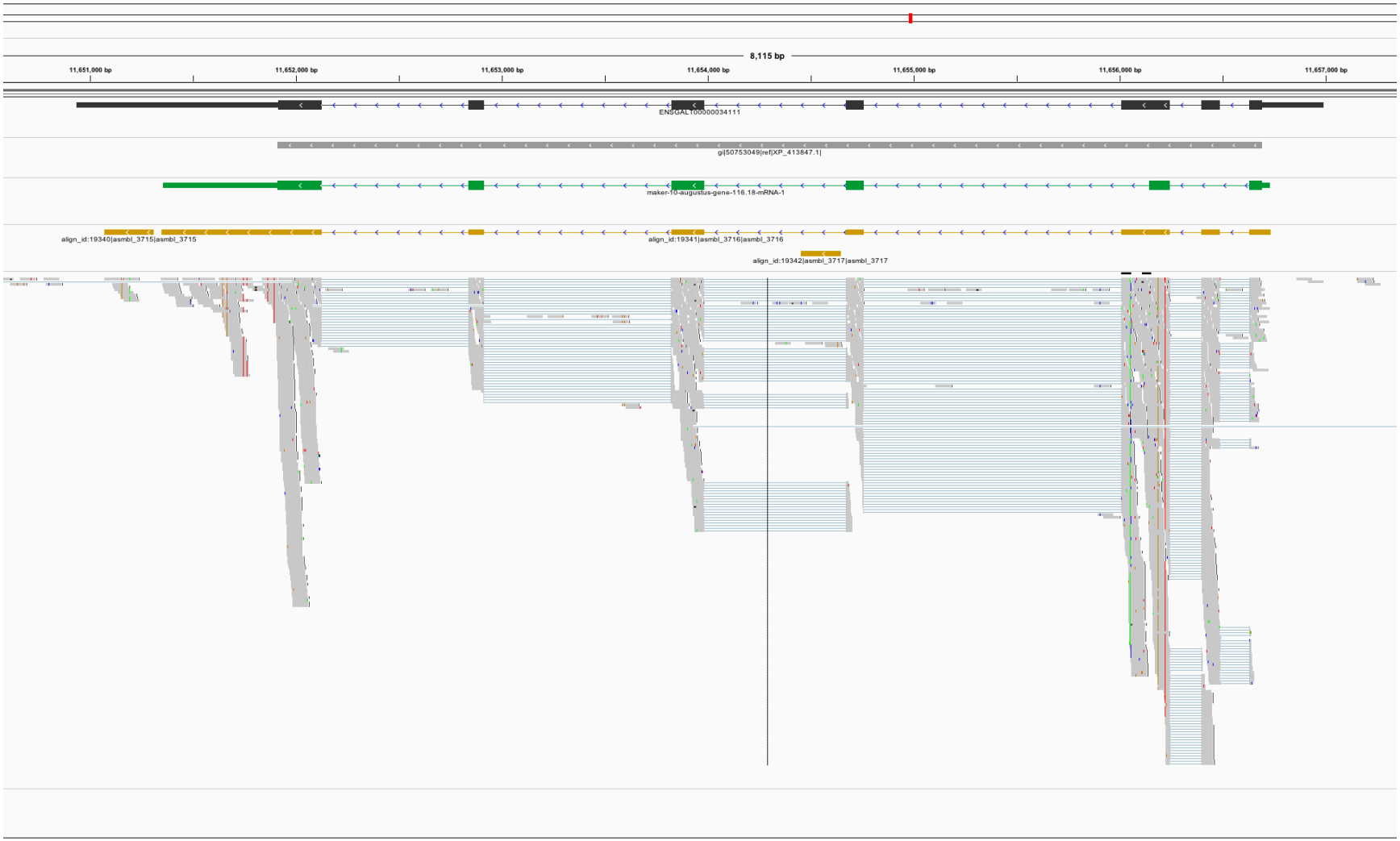
- Maker will align transcripts (ESTs), but these need to be assembled first.
- Cufflinks: mapped reads -> transcripts



How to use RNA-seq

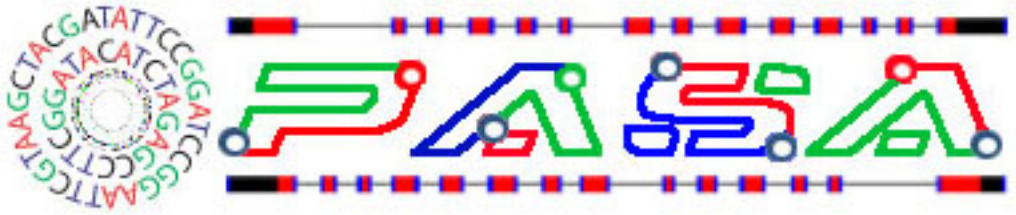
- Maker will align transcripts (ESTs), but these need to be assembled first.
- Cufflinks: mapped reads -> transcripts
- Trinity: assembles transcripts without a genome

Mapped Trinity-assembled transcripts



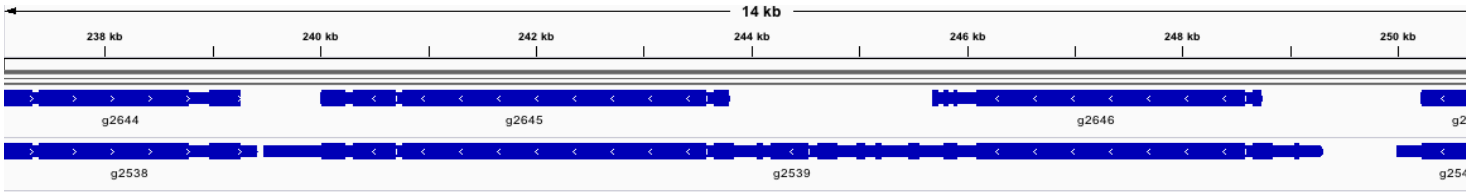
How to use RNA-seq

- Maker will align transcripts (ESTs), but these need to be assembled first.
- Cufflinks: mapped reads -> transcripts
- Trinity: assembles transcripts without a genome
- PASA can be used to improve transcript quality



Ab initio gene finders are used in Maker

- Commonly used programs: Augustus, Snap, Genemark-ES, FGENESH, Genscan, Glimmer-HMM,…
- Uses HMM-models to figure out how introns, exons, UTRs etc. are structured
- These HMM-models need to be trained!



General recommendations

- Always combine different types of evidence!
- One single method is not enough!
- Use Maker!

Overview

Annotation = combining different lines of evidence into gene models

Gene prediction

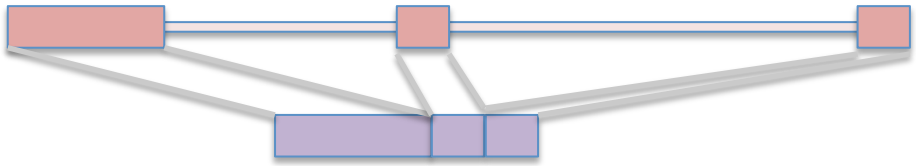
Evidence

Combining – MAKER

A bit of terminology first:

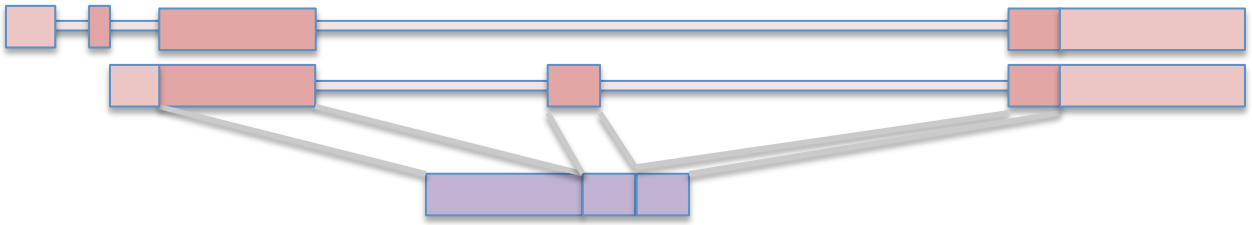
Gene prediction

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



Gene annotation

Goal: Identify the entire gene structure



Existing annotation pipelines – MAKER2

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

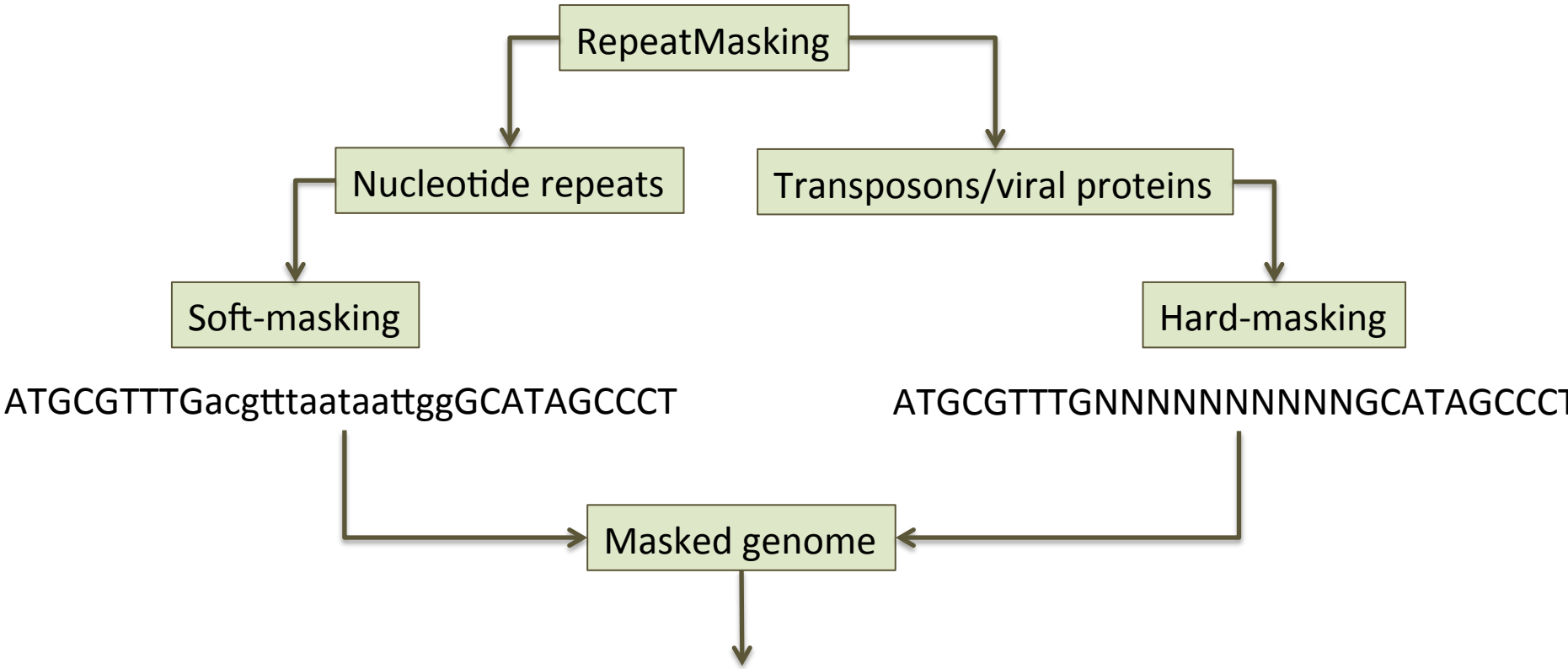
Advantages over competing solutions:

- 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?

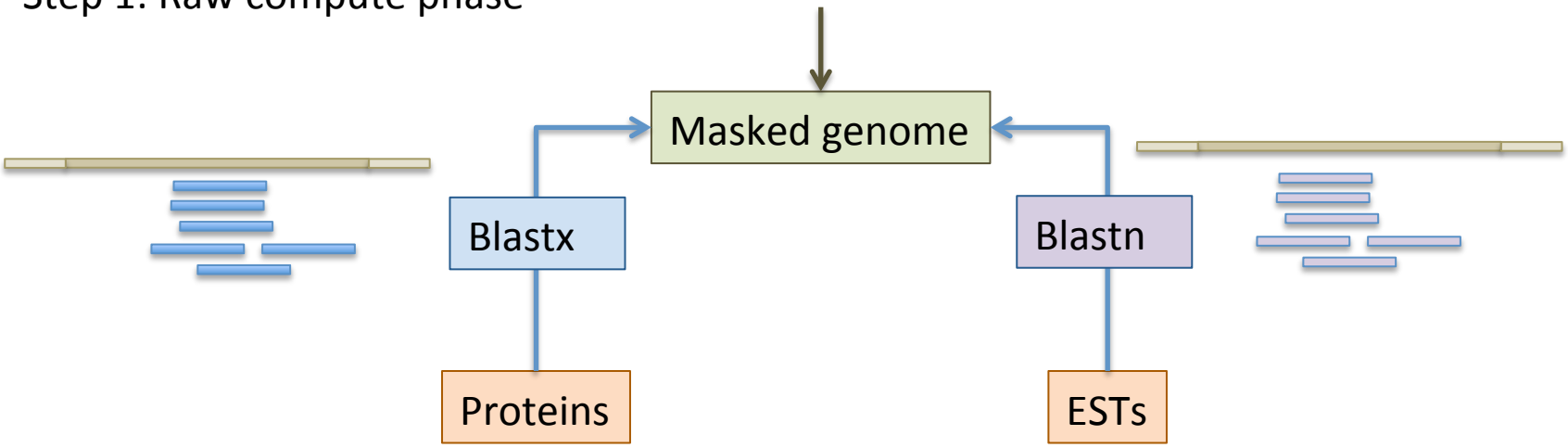
Existing annotation pipelines – MAKER2

Step 1: Raw compute phase



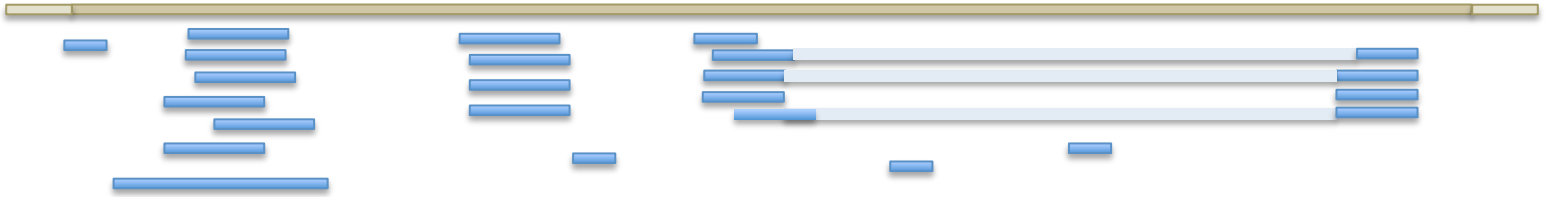
Existing annotation pipelines – MAKER2

Step 1: Raw compute phase



Existing annotation pipelines – MAKER2

Step 2: Filter and cluster alignments



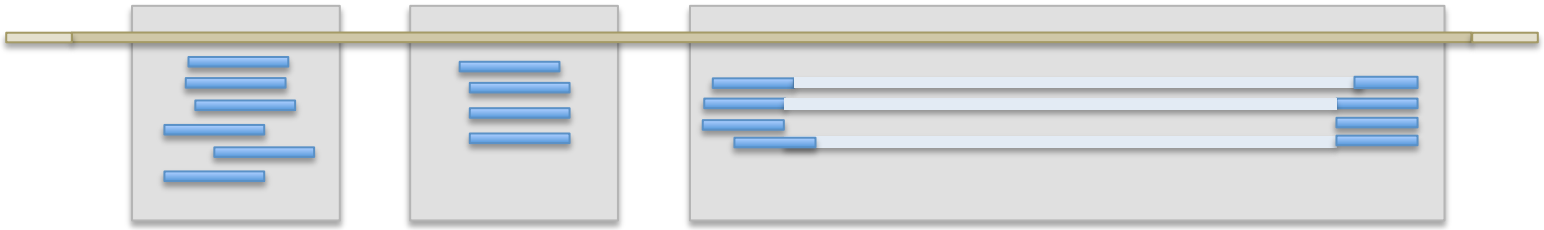
Filtering is based on rules defined in the Maker configuration for a given project

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

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

Existing annotation pipelines – MAKER2

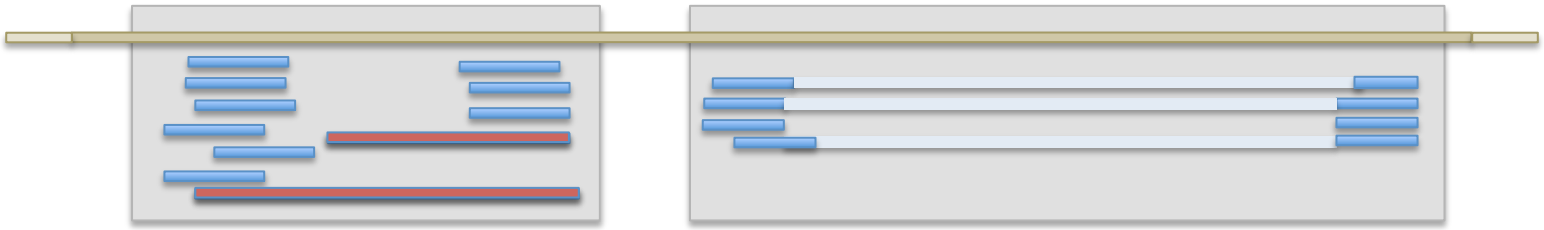
Step 2: Filter and cluster alignments



Clustering groups evidence alignments into ’loci’

Existing annotation pipelines – MAKER2

Step 2: Filter and cluster alignments

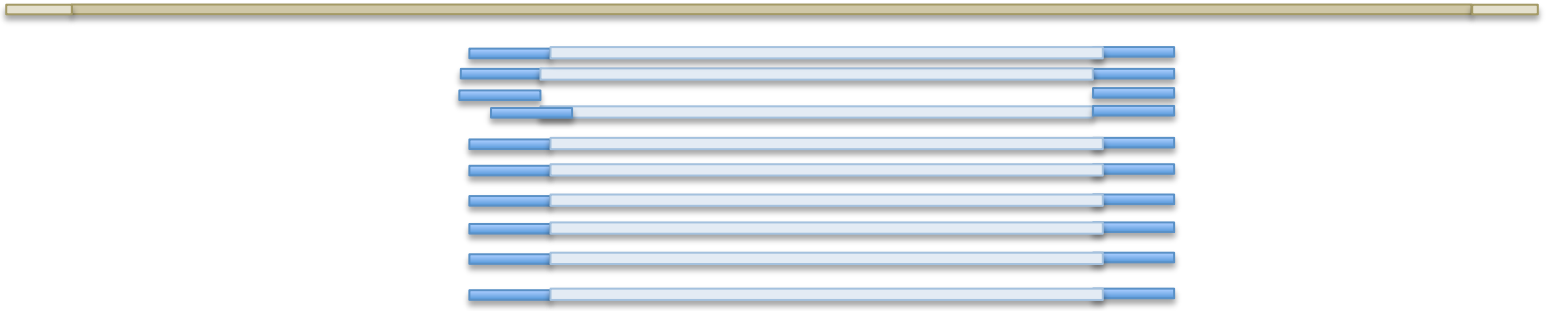


Problematic data can complicate clustering

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

Existing annotation pipelines – MAKER2

Step 2: Filter and cluster alignments

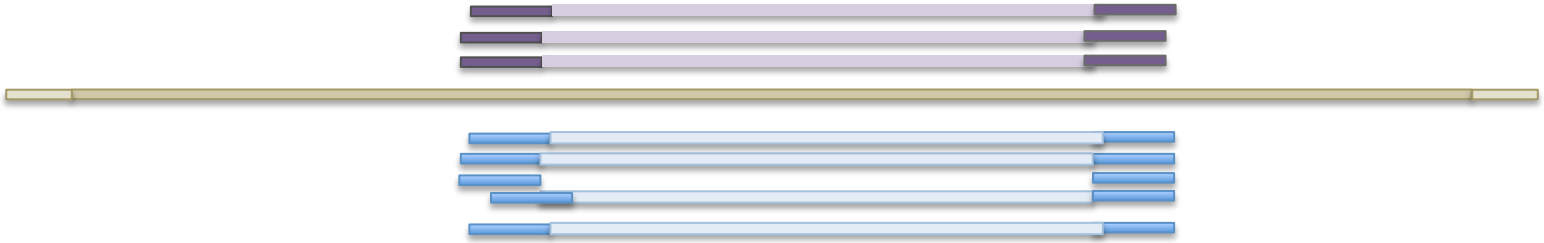


Clustering groups evidence alignments into 'loci'

- 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'

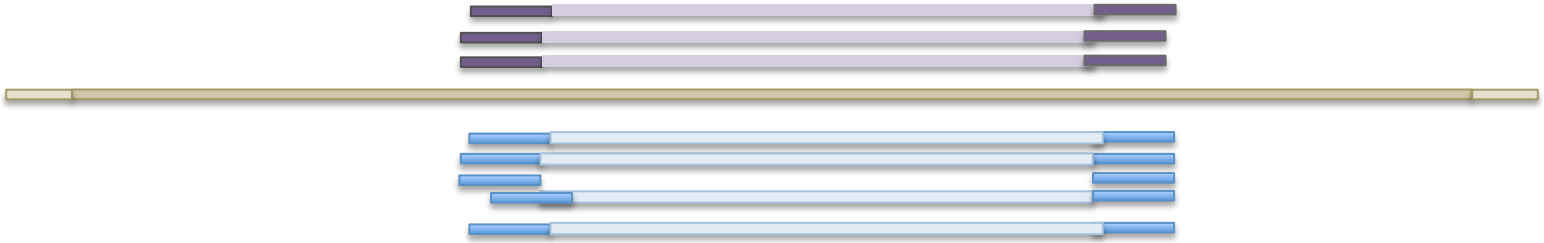
Amount of data in any given cluster is then collapsed to remove redundancy

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

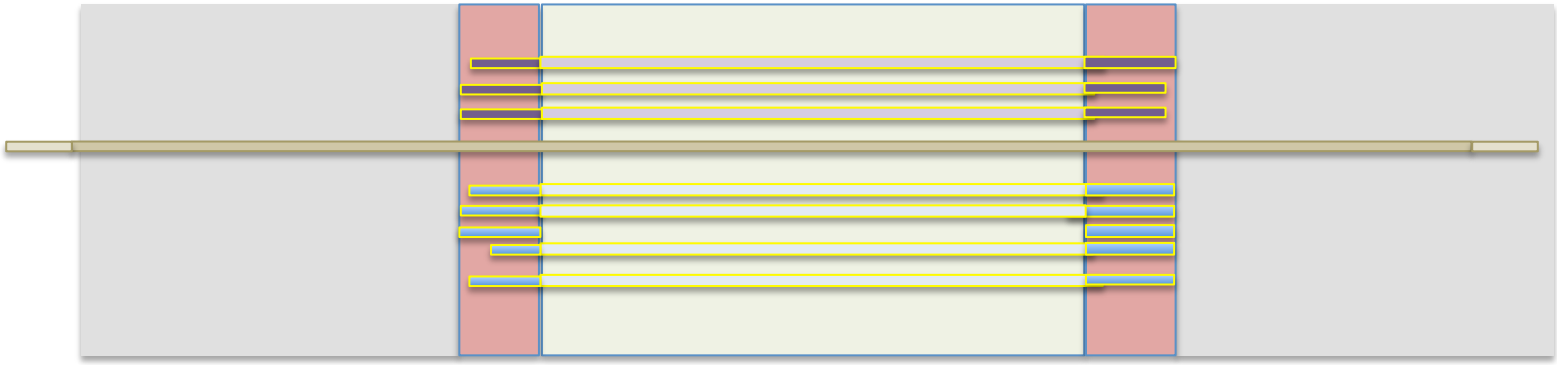
Step 3: Polishing alignments

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

Exonerate is used to create splice-aware alignments

Existing annotation pipelines – MAKER2

Step 4: Synthesis

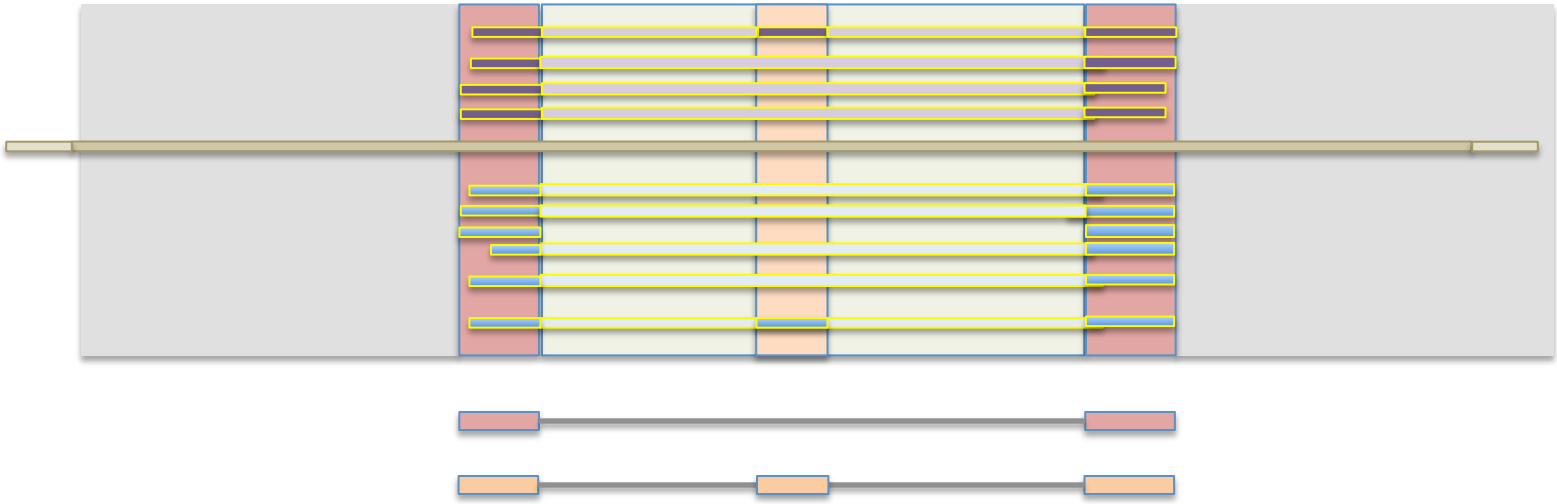


Synthesis refers to the extraction of information to generate evidence for annotations

Done by identifying genomic regions overlapping with sequence features

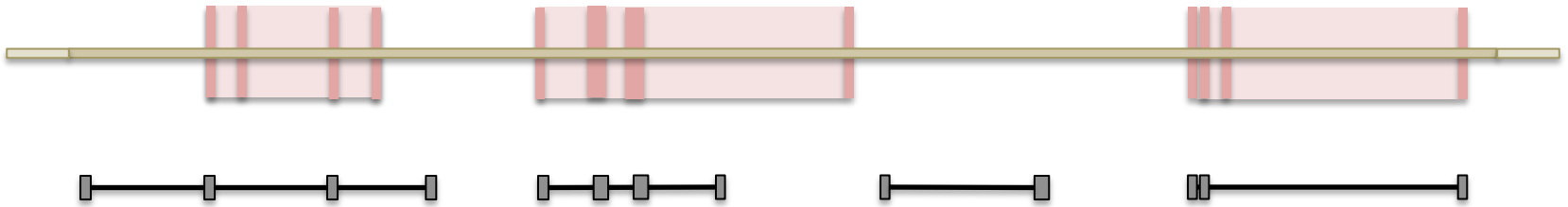
Existing annotation pipelines – MAKER2

Step 4: Synthesis



Existing annotation pipelines – MAKER2

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

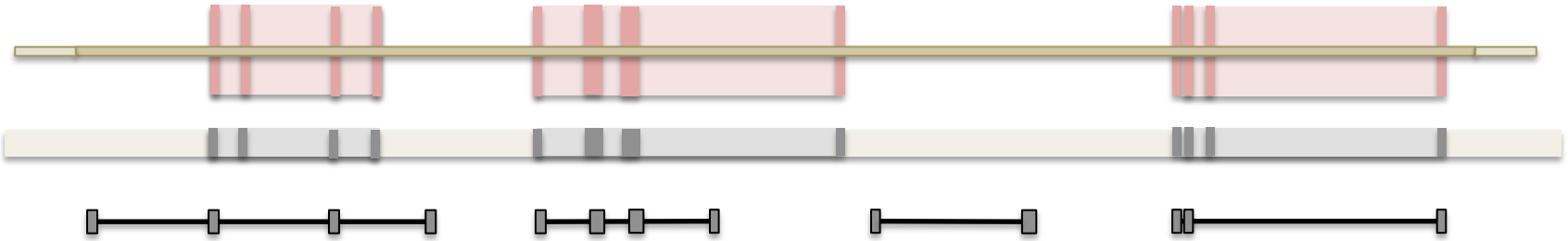


Evidence alignments provide support for the identifcation of gene loci

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

Existing annotation pipelines – MAKER2

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

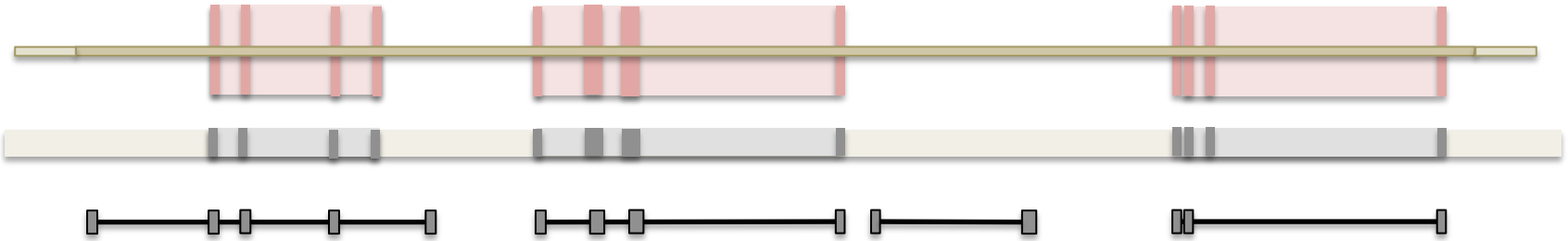


Ab-intio predictions can be improved when evidence is provided (hints)

Help refine and calibrate a computational inference for a given locus

Existing annotation pipelines – MAKER2

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



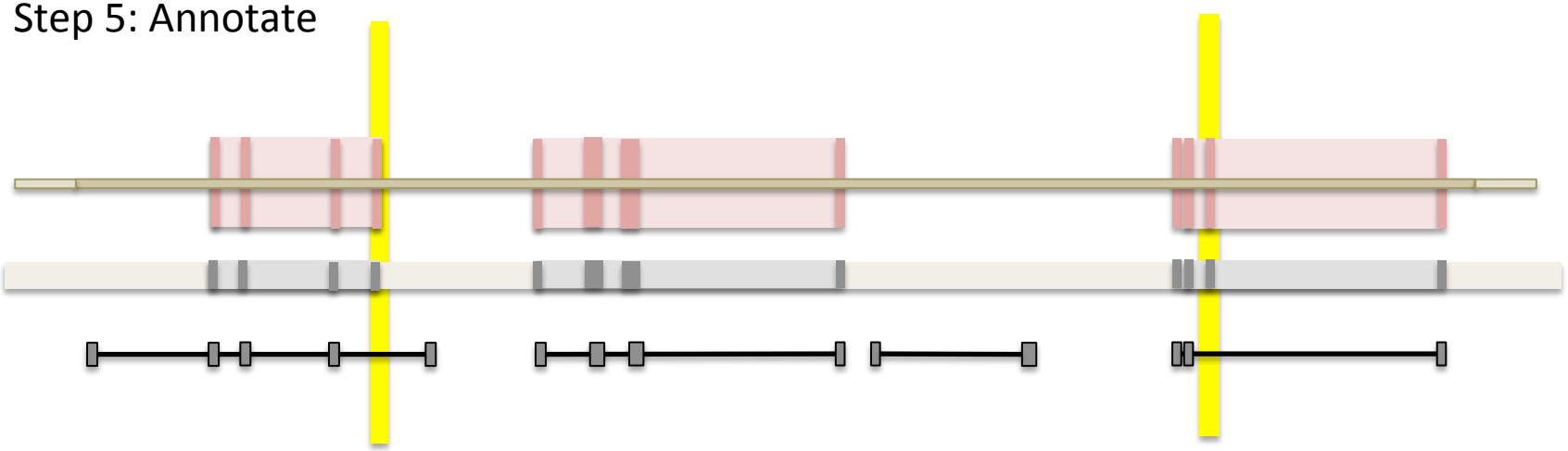
Ab-intio predictions can be improved when evidence is provided (hints)

Help refine and calibrate a computational inference for a given locus

Hints: Introns, intergenic sequence, CDS

Existing annotation pipelines – MAKER2

Step 5: Annotate



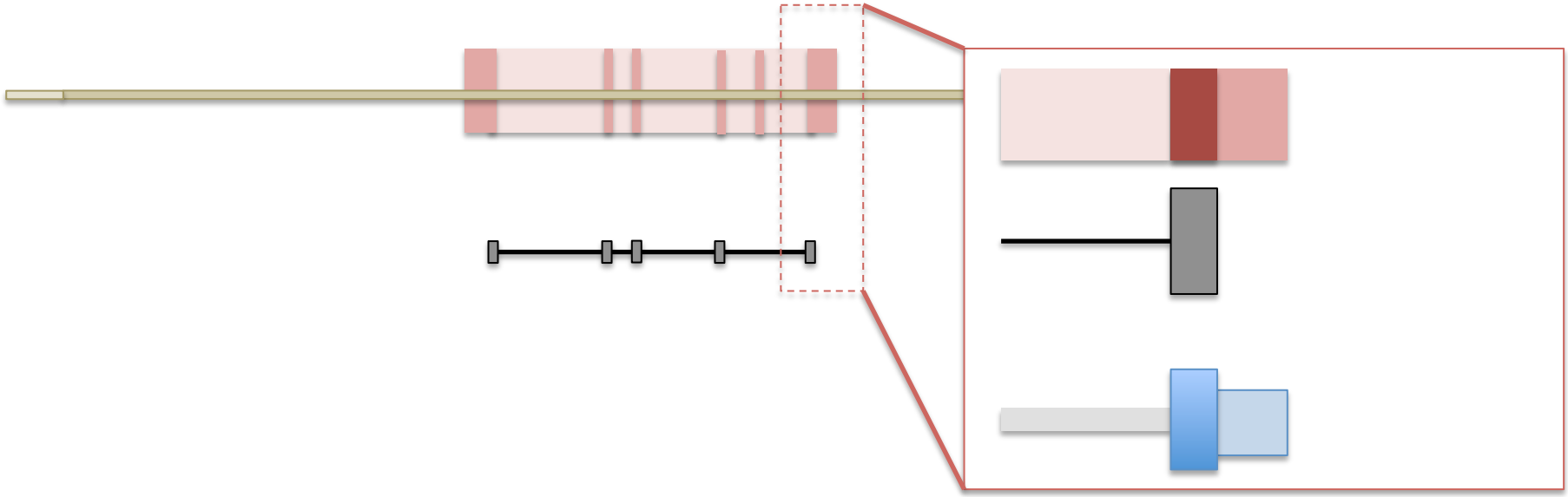
Refined ab-initio models may still be incomplete / partially wrong

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

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

Existing annotation pipelines – MAKER2

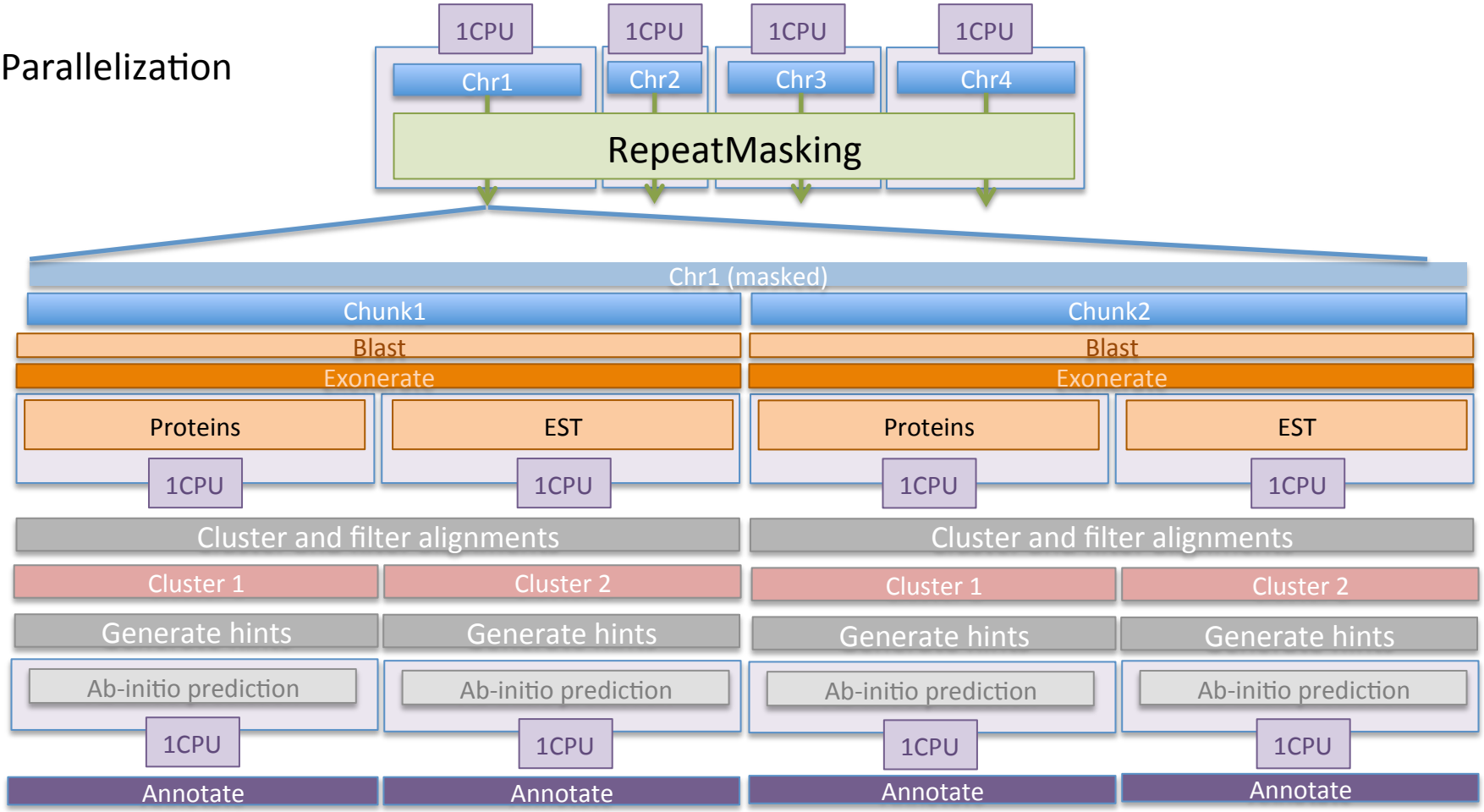
Step 5: Annotate



Synthesized transcript structures are compared against evidence to find UTRs

Existing annotation pipelines – MAKER2

Parallelization

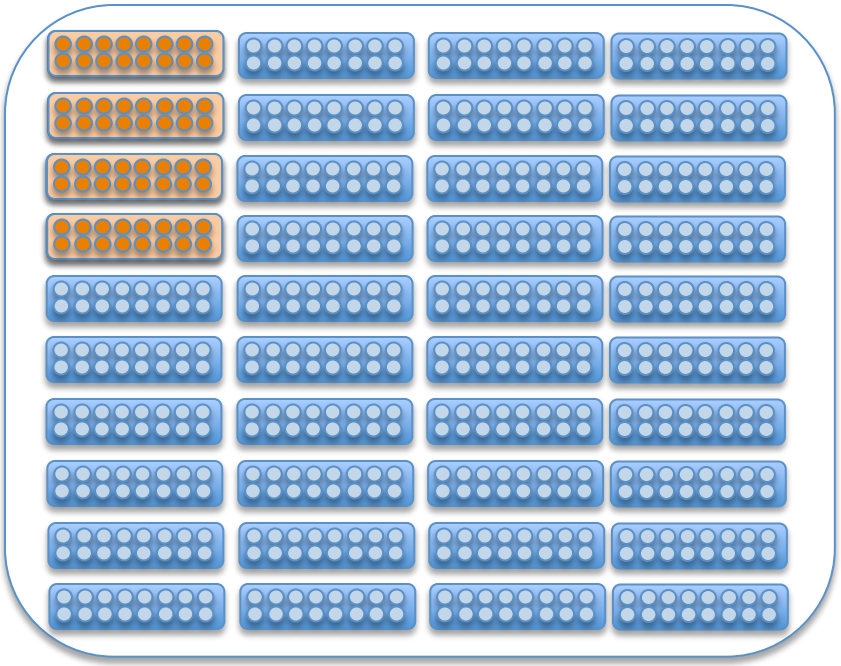


Existing annotation pipelines – MAKER2

Parallelization – Running on a cluster

Maker uses MPI for job distributon

- runs on almost all computing platforms
- Operates on cores, not nodes

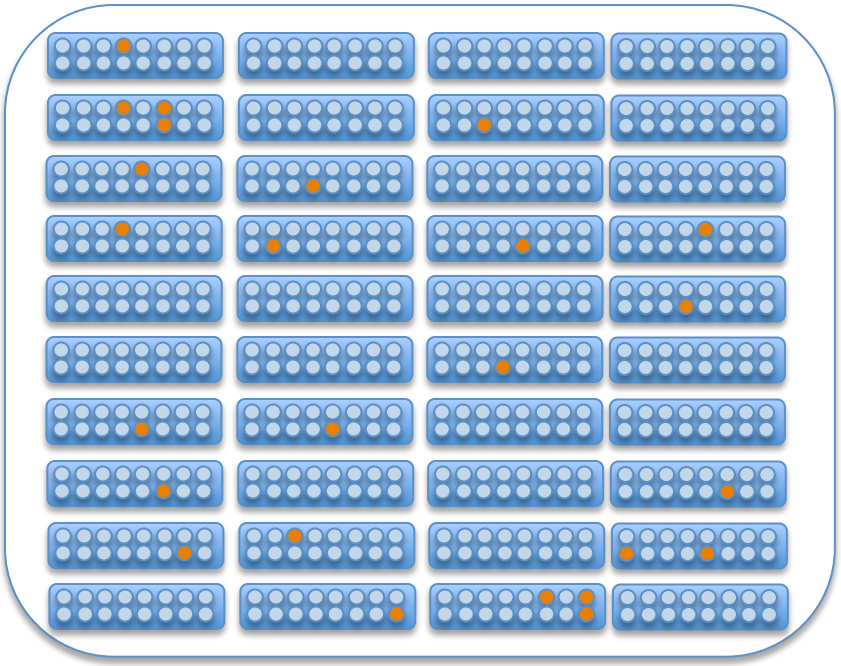


Existing annotation pipelines – MAKER2

Parallelization – Running on a cluster

Maker uses MPI for job distributon

- runs on almost all computing platforms



What's next

Computational pipelines make mistakes

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