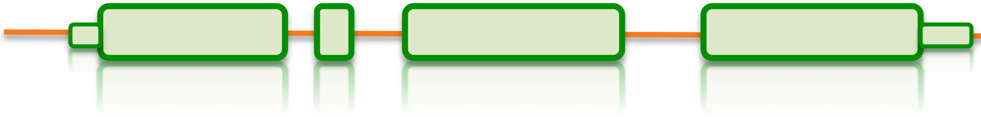


Methods in genome annotation 2016

Jacques Dainat, PhD
NBIS genome annotation service
Uppsala University

This lecture will mainly focus on eukaryote

1. Introduction - Understanding gene annotation
2. The different annotation approaches
3. Two pipelines cases (EnsEMBL MAKER2)
4. Quick word about Prokaryote annotation
5. Check an annotation
6. Closing remarks

1. Introduction

Overview

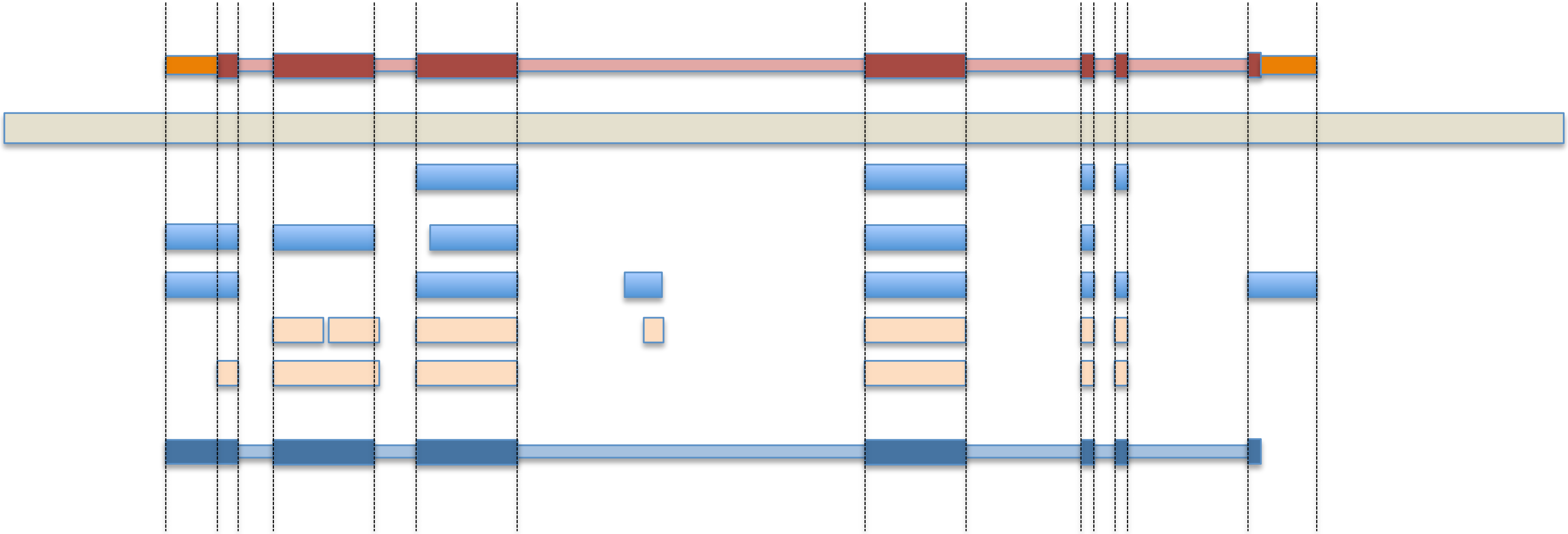
Annotation = combining different lines of evidence into gene models

Evidences: ESTs / Transcritps

Proteins

Ab-initio prediction

Combining



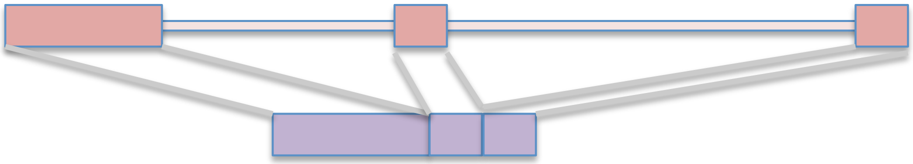
A bit of terminology first:

'gene prediction' and 'gene annotation' are often used as if they are synonyms, they are not

Mark Yandell & Daniel Ence
Nature Reviews Genetics **13**, 329-342 (May 2012) doi:10.1038/nrg3174

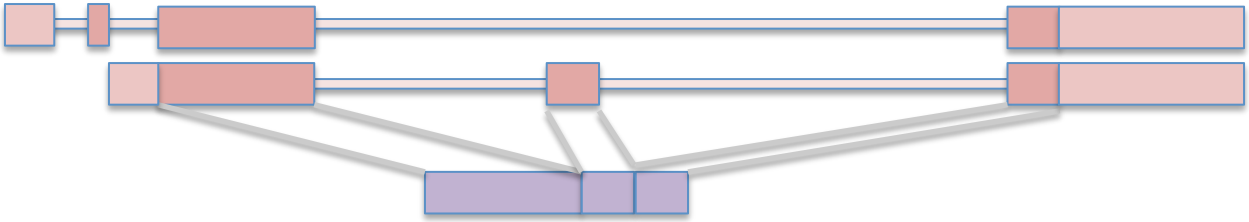
Gene prediction (*Ab-initio*)

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



Gene annotation

Goal: Identify the entire gene structure



Introduction

A bit of terminology first:

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

Gene prediction may : - predict UTRs
- predict isoforms
- use evidence information

Gene annotation can predict gene models without UTR

⇒Don’t be disturbed by the use of different terms, roughly mean the same thing : **determining the gene models.**

⇒**The most important is to know/understand the approach and the data used for the annotation.**

Gene prediction ~ **Gene annotation** ~ Gene building ~ Gene finding

The different approaches

- Similarity-based methods :

  These use similarity to annotated sequences like proteins, cDNAs, or ESTs
- *Ab initio* prediction:

  These don’t use external evidence to predict sequence structure
- Hybrid approaches :

  These are *ab initio* tools integrating multiple forms of evidence/hint
- Comparative (homology) based gene finders :

  These align genomic sequences from different species and use the alignments to guide the gene predictions
- Chooser, combiner approaches :

  These combine gene predictions of other gene finders
- Pipelines :

  These combine multiple approaches

2) The different annotation approaches

2.1) Annotation through similarity-based methods
“extrinsic approach”

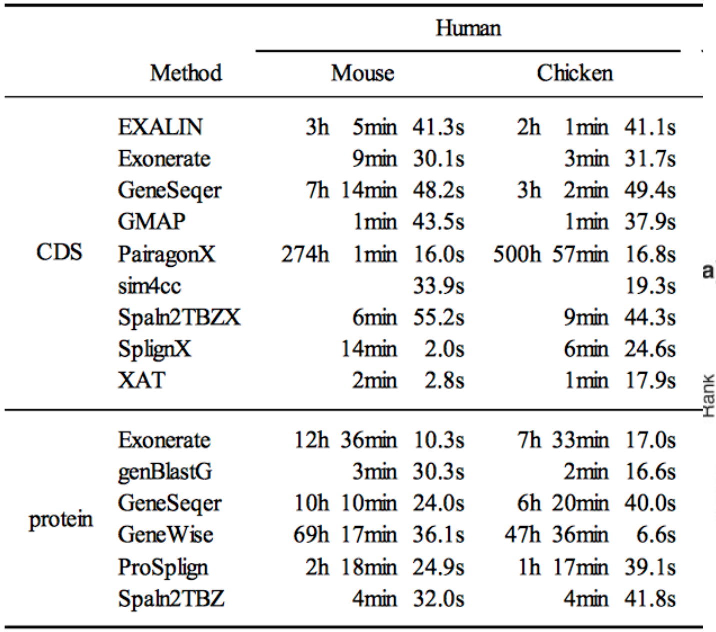
similarity-based method

- Rough approximation (fast)

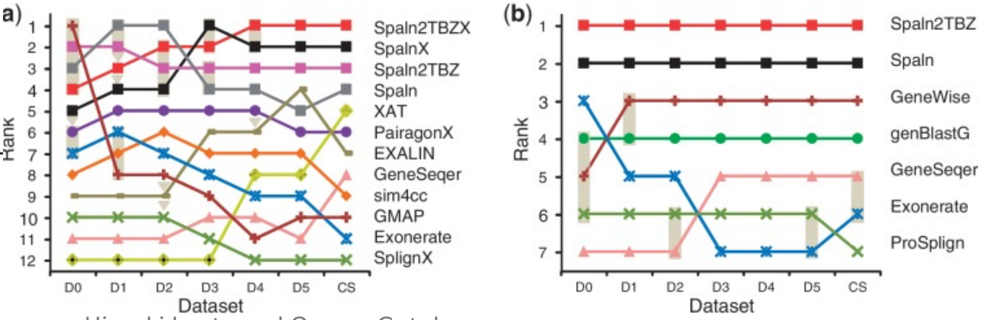
DNA	AA
Blastn	Pmatch
Vsearch	Blastx
NSimScan*	PSimScan*

- Splice-site aware alignment (slow – moderately slow)

DNA	AA
Exonerate	Exonerate
Gmap	Genewise
…	…



		Human	
	Method	Mouse	Chicken
CDS	EXALIN	3h 5min 41.3s	2h 1min 41.1s
	Exonerate	9min 30.1s	3min 31.7s
	GeneSeqer	7h 14min 48.2s	3h 2min 49.4s
	GMAP	1min 43.5s	1min 37.9s
	PairagonX	274h 1min 16.0s	500h 57min 16.8s
	sim4cc	33.9s	19.3s
	Spaln2TBZX	6min 55.2s	9min 44.3s
	SplignX	14min 2.0s	6min 24.6s
	XAT	2min 2.8s	1min 17.9s
protein	Exonerate	12h 36min 10.3s	7h 33min 17.0s
	genBlastG	3min 30.3s	2min 16.6s
	GeneSeqer	10h 10min 24.0s	6h 20min 40.0s
	GeneWise	69h 17min 36.1s	47h 36min 6.6s
	ProSplign	2h 18min 24.9s	1h 17min 39.1s
	Spaln2TBZ	4min 32.0s	4min 41.8s



Hiroaki Iwata and Osamu Gotoh
Nucleic Acids Res. 2012 Nov; 40(20): e161. doi: 10.1093/nar/gks708

similarity-based method

Limits for proteins:

- Related to pre-existing data
- Most of data are *Ab-initio* prediction (No verification of their existence)
- Consequently errors in databases can be transmitted
- No UTR

Limits for transcripts:

- Hard to catch low expressed / peculiar expressed (stage of life, condition, etc…)
- Not complete (EST)
- Transcriptome assembly errors
- Can even be difficult with long reads (error rate / frameshift)

General limits for transcripts:

- Difficult to find limits of similarity

Strengths => produce biologically relevant predictions
=> produce evidences useful for ab initio tools, combiners and pipelines

2) The different annotation approaches

2.2) *Ab-initio* annotation tools
“intrinsic approach”

from the beginning

Ab initio method

method based on **gene content :**
(statistical properties of protein-coding sequence)

- codon usage
- hexamer usage
- GC content
- compositional bias between codon positions
- nucleotide periodicity
- …

and on **signal detection**:

- Promoteur
- ORF
- Start codon
- Splice site (Donor and acceptor)
- Stop codon
- Poly(A) tail
- CpG islands
- …

=> *Ab initio* tool will combines these information through different Probabilistic models: HMM, GHMM, WAM, etc.

Ab initio method

Popular tools:

- **SNAP** Works ok, easy to train, not as good as others especially on longer intron genomes.
- **Augustus** Works great, hard to train, but getting better).
- **GeneMark-ES** Self training, no hints, buggy, not good for fragmented genomes or long introns (Best suited for Fungus).
- **FGENESH** Works great, costs money even for training.

Supported by MAKER

http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial

- **GlimmerHMM** (Eukaryote)
- **GenScan**
- **Gnomon** (NCBI)

Ab initio method

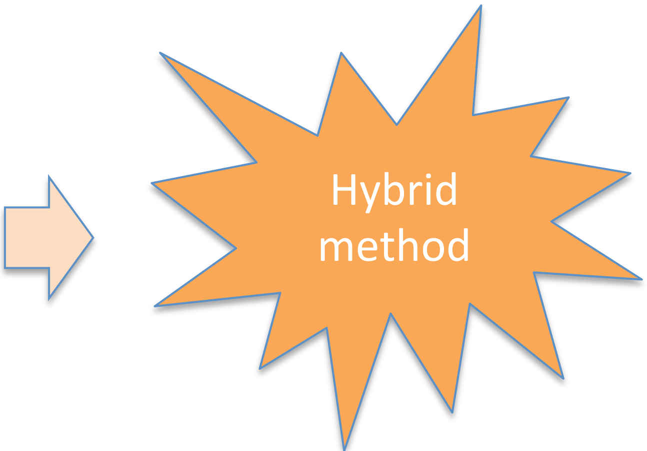
Strengths :

- Fast and easy means to identify genes
- Annotate unknown genes
- "Exhaustive" annotation
- Need no external evidence

Limits :

- No UTR*
- No alternatively spliced transcripts*
- Over prediction (exons or genes)
- **Training** needed to perform well in *terra incognita*'

- Split single gene into multiple predictions
- Fused with neighboring genes
- Less accurate than homology based method:
  - Exon extremities
  - Splicing sites



Hybrid method

Hybrid method

Hybrid (*evidence-drivable gene predictors*) approaches incorporate hints in the form of EST alignments or protein profiles to increase the accuracy of the gene prediction.

GenomeScan	Blast hit used as extra guide	
Augustus	16 types of hints accepted (gff): start, stop, tss, tts, ass, dss, exonpart, exon, intronpart, intron, CDSpart, CDS, UTRpart, UTR, irpart, nonexonpart.	
GeneMark-ET	EST-based evidence hints	Self training !
GeneMark-EP	Protein-based evidence hints	
SNAP	Accepts EST and protein-based evidence hints.	
Gnomon	Uses EST and protein alignments to guide gene prediction and **add UTRs**	
FGENESH+	Best suited for plant	
EuGene*	Any kind of evidence hints. Hard to configure (best suited for plant)	

Hybrid method

Strength : High accuracy

Limits :

- **Extra computation to generate alignments**

- **heterogeneous sequence quality** :

  Incomplete,
  Error during transcriptome assembly
  Contamination
  Sequence missing
  Orientation error

Hybrid method

The BRAKER1 gene finding pipeline:

BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS
Katharina J. Hoff et *al.*
Bioinformatics (2016) 32 (5): 767-769. doi: 10.1093/bioinformatics/btv661

- BRAKER1 was more accurate than MAKER2 when it is using RNA-Seq as sole source for training and prediction.

- BRAKER1 does not require pre-trained parameters or a separate expert-prepared training step.

2) The different annotation approaches

2.3) Annotation using comparative genomic approach

Comparative-based method

Comparative-based methods lie in the similarities shared by regions of two evolutionary related genomic sequences.
The main assumption of these methods is that the functional parts of an eukaryotic genomic sequence, the exons, tend to be more conserved than the non-functional ones, the introns.

Dual genome, de novo gene structure prediction:

- **Rosetta** (Pioneer – 2000)
- **SGP-2** (2001) – considered only the conservation in protein-coding regions
- **TWINSCAN** (2001) - included models of conservation in splice sites and start and stop codons
- **SLAM** (2003)
- **TWAIN** (2005)

More than 2 genomic sequences:

- **NSCAN** (2006)
- **Conrad** (CRF, 2007)
- **CONTRAST** (CRF, 2008)
- **Augustus-cgp** (new)

Comparative-based method

Strength : Good accuracy

CONTRAST the best de novo gene predictor for mammals. (Michael R. Brent in Nature reviews, 2008) => 58 % ORF structure correct in human

Limits

- Alignment errors bias
- biological function is not necessarily conserved
- Difficult to define limits of higher similarity
- Difficult to find optimal evolutionary distance (pattern of conservation differ between loci)
- Whole genome alignment is time/memory consuming

2) The different annotation approaches

2.4) Chooser / combiner

Chooser / combiner

Use battery of gene finders and evidence (EST, RNAseq, protein) alignments and:

Tool	Consensus based chooser	Evidence based chooser	weight of different sources	Comment
A) select the prediction whose structure best represents the consensus				
JIGSAW	X			
B) choose the best possible set of exons and combine them in a gene model				
EVM EVidenceModeler	X	X	X	User can set the expected evidence error rate manually or/and learn from a training set
Evigan	X		X	Unsupervised learning method
Ipred		X		Does not require any a priori knowledge Can also combine only evidences to create a gene model

Strength => They improve on the underlying gene prediction models

2) The different annotation approaches

2.5) Annotation of other genome features

Other genome features

Feature type	DB associated	Tool example	approach
ncRNA	Rfam	infernal	HMM + CM
tRNA	Sprinzl database	tRNAscan-SE	CM + WMA
snoRNA		snoscan	HMM + SCFG
miRNA	miRBase	Splign miR-PREFeR (for plant)	sequence alignment Based on expression patterns
Repeats	Repbase, Dfam	repeatMasker	HMM, blast
Pseudogenes		pseudopipe	homology-based (blast)
…			

3) Gene annotation pipelines
(The ultimate step)

Align evidences themselves, add UTRs and more

Annotation pipeline

PASA Produces evidence-driven consensus gene models

- - minimalist pipeline ()
- + good for detecting isoforms
- + biologically relevant predictions

=> associated to *Ab initio* tools and combined with **EVM** it gives pretty good job !

- - PASA + Ab initio + EVM not automatized

NCBI pipeline Evidence + *ab initio* (Gnomon), repeat masking, gene naming, data formatting, miRNAs, tRNAs

- - Not released by NCBI

Ensembl Evidence based only (comparative + homology) …

MAKER2 Evidence based and/or *ab initio* …

…

3) Gene annotation pipelines (The ultimate step)

3.1) EnsEMBL

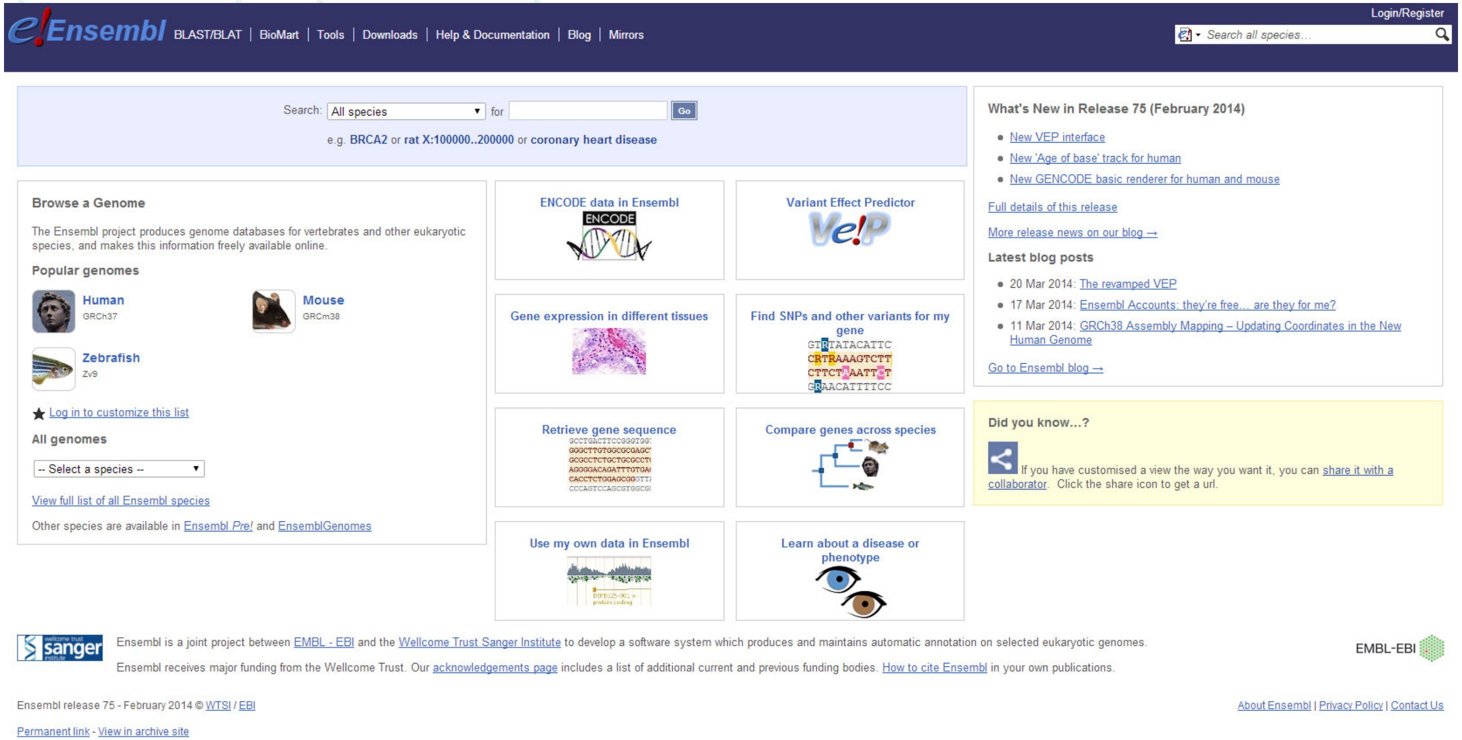
EnsEMBL

The NBIS annotation service

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Overview

Started in 1999



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.

EMBL-EBI

Ensembl release 75 - February 2014 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

Permanent link - View in archive site

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, …

EnsEMBL

Overview

3 major components

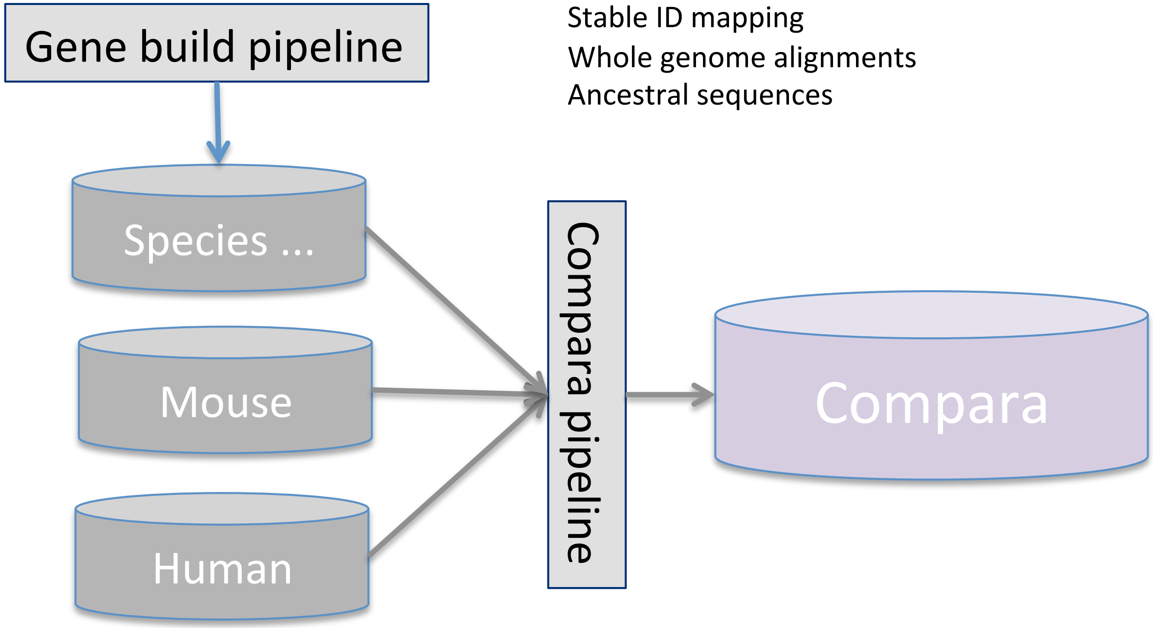
They also have **Variation** and **Regulatory regions** components

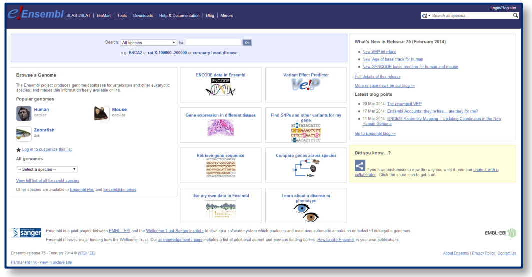
annotation

cross-species resources

Phylogenetic trees =>Z orthology / paralogy
TreeFam Families
Stable ID mapping
Whole genome alignments
Ancestral sequences

Data access




e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...
Search: All species for Go
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
sanger
EMBL-EBI

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

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.

Ensembl release 75 - February 2014 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

Permanent link - View in archive site

biomart

Perl API

The annotation pipeline

0. Setting up an annotation project

Config file needs to be written ’manually’

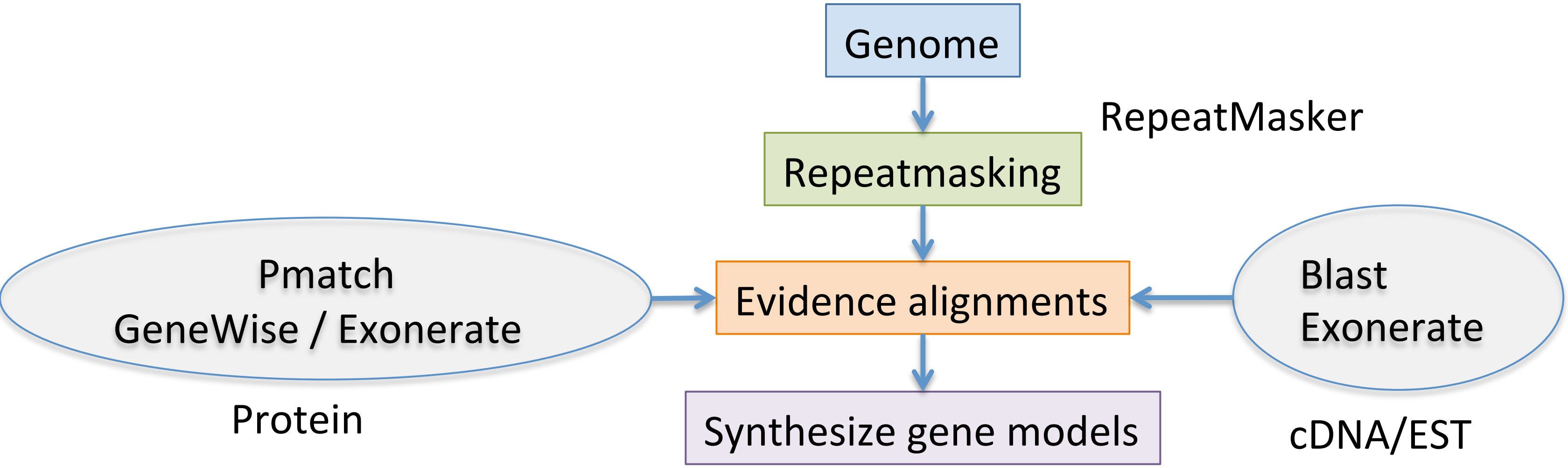
Pipeline logic needs to be specified ’manually’

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

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

EnsEMBL

The annotation pipeline



Genome

RepeatMasker

Repeatmasking

Pmatch
GeneWise / Exonerate

Protein

Evidence alignments

Blast
Exonerate

cDNA/EST

Synthesize gene models

no file output => data saved in database

EnsEMBL

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 (evidence hierachy)

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)

EnsEMBL

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

3. Output

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

→ Much more complex, but also more powerful

4. Miscellaneous

EnsEMBL provides no tools for manual curation

EnsEMBL

Strength : High accuracy
Training course exists
API to access the data
Compara

They re-designed the pipeline recently !
- Several months per species => less than two weeks now.
- Minimal human interaction
- lincRNA included
=> They plan to use it this year

Limits : Hard to configure
Complex data structure
Only limited training opportunities
Parallelization is done via SGE or LSF (cannot be run on just any system)
Documentation is very patchy

=> 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 your project

3) Gene annotation pipelines (The ultimate step)

3.2) MAKER2

MAKER2

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

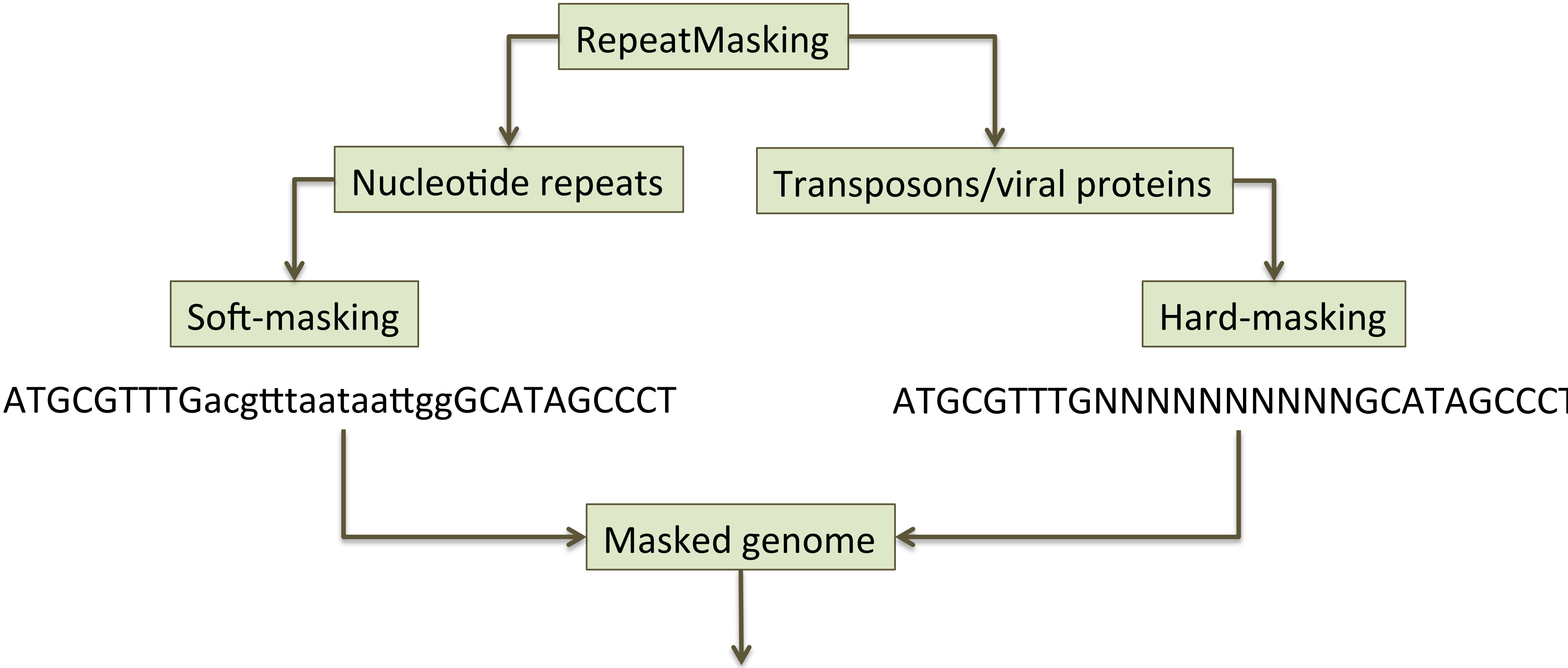
- can be used pure evidence-based, pure *ab initio*, or evidence-driven (on the fly) *ab initio.*
- add UTR when ESTs are supplied.
- Evidence based chooser : select post processed gene model the most consistent with evidences (protein / EST / RNAseq)

Advantages over competing solutions:

- Easy to use and to configure
- 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?

Step 1: Raw compute phase



RepeatMasking

Nucleotide repeats

Transposons/viral proteins

Soft-masking

Hard-masking

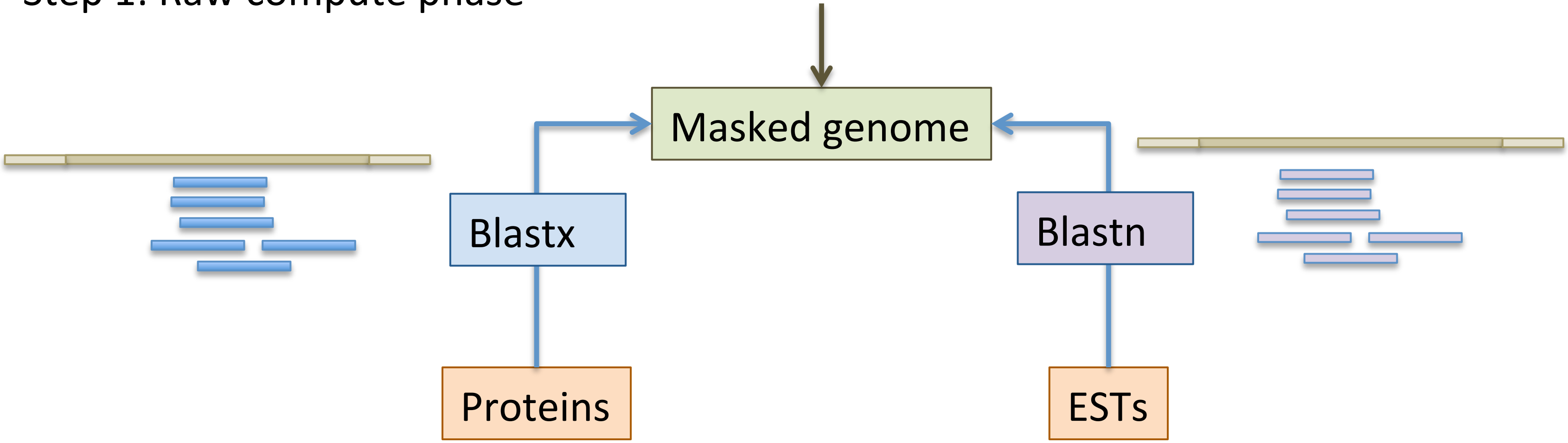
ATGCGTTTGacgtttaataattggGCATAGCCCT

ATGCGTTTGNNNNNNNNNNGCATAGCCCT

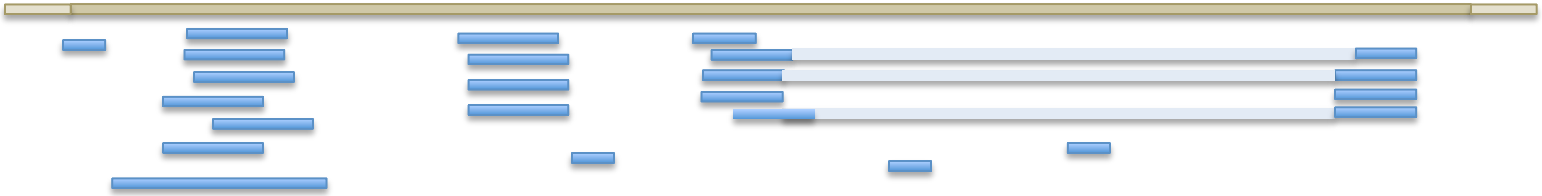
Masked genome

Existing annotation pipelines – MAKER2

Step 1: Raw compute phase



Step 2: Filter and cluster alignments

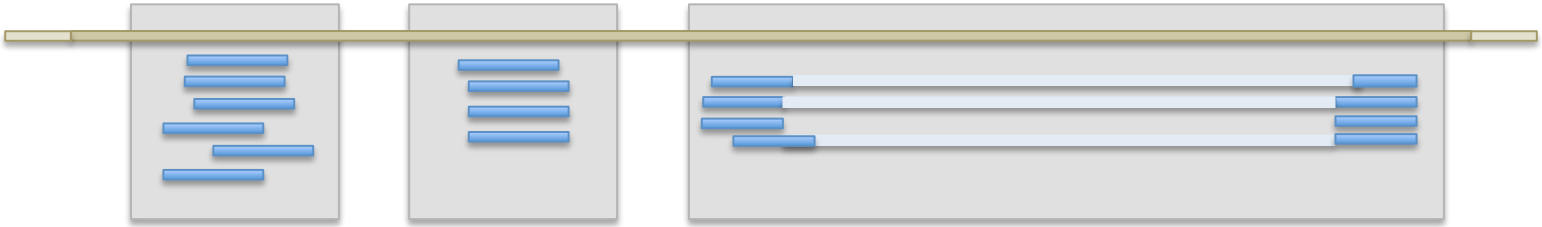


Filtering is based on rules defined in the Maker configuration for a given project

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

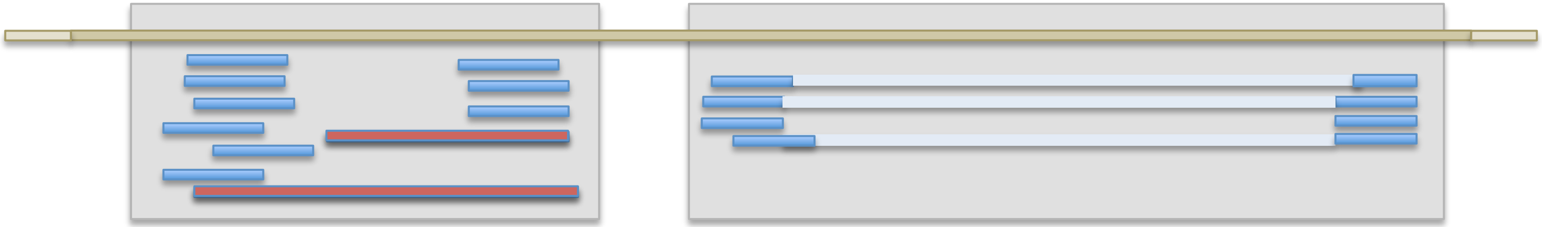
Default settings sensible for most projects, but can be changed!

Step 2: Filter and cluster alignments



Clustering groups evidence alignments into ’loci’

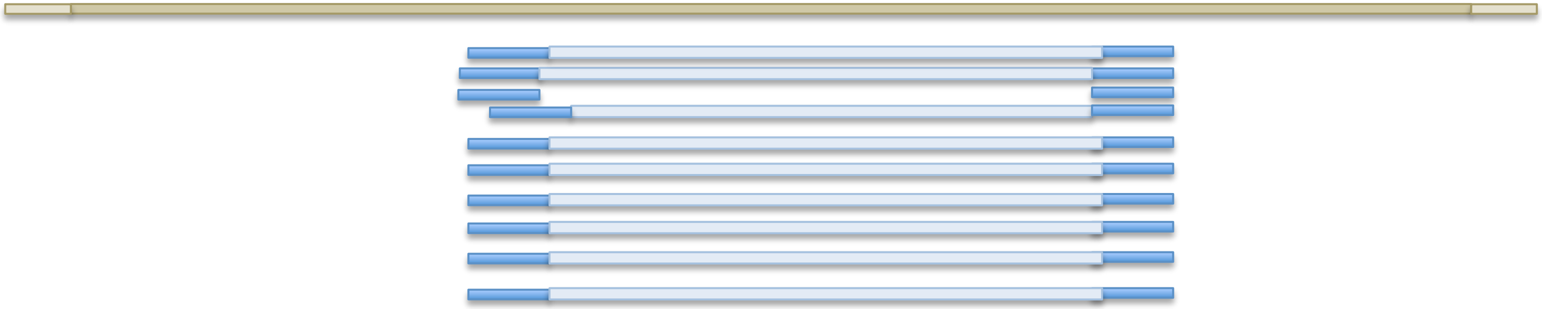
Step 2: Filter and cluster alignments



Problematic data can complicate clustering

Needs to be fixed by => cleaner data

Step 2: Filter and cluster alignments



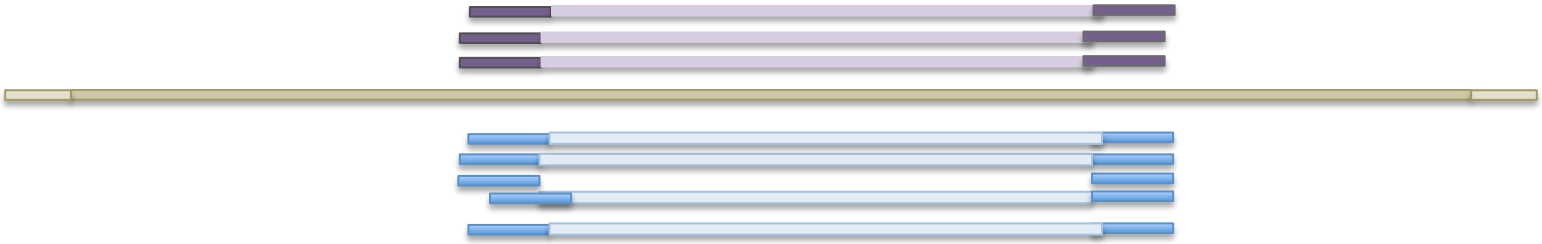
Clustering groups evidence alignments into ’loci’

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



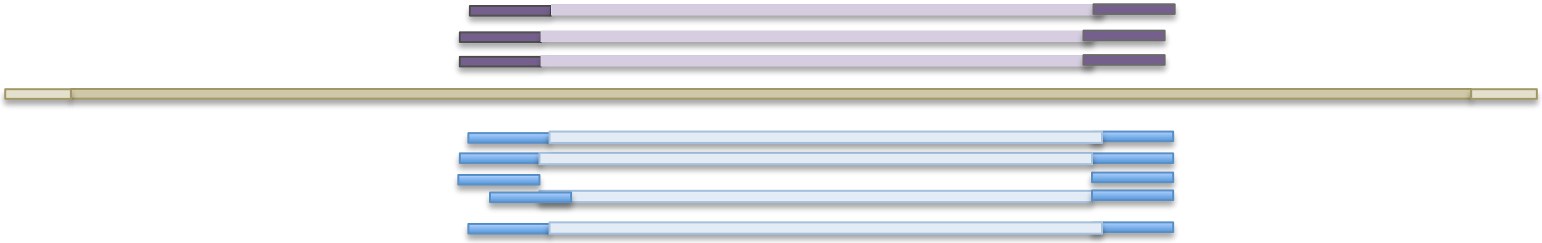
Clustering groups evidence alignments into ’loci’

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

Threshold for the collapsing is also user-definable

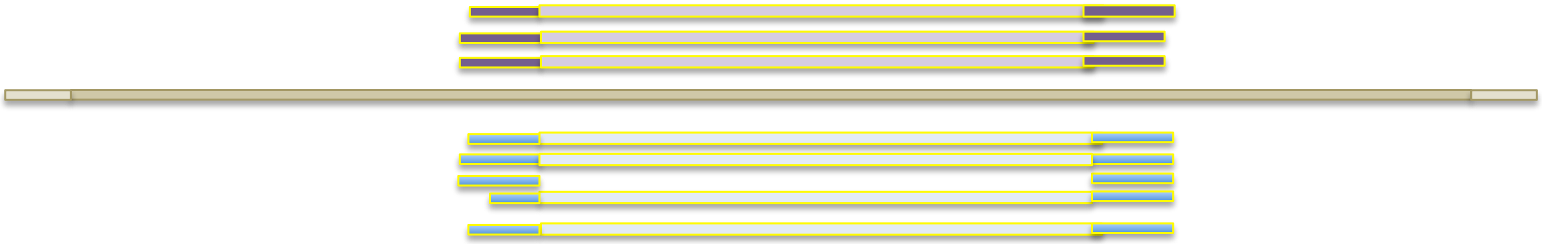
Performed for all lines of evidence

Step 3: Polishing alignments



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

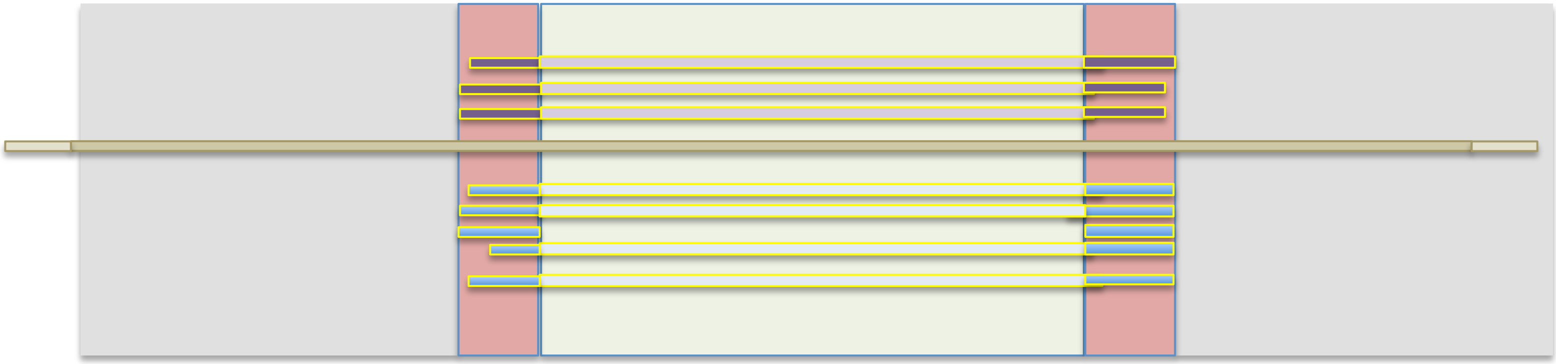
Step 3: Polishing alignments



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

Exonerates is used to create splice-aware alignments

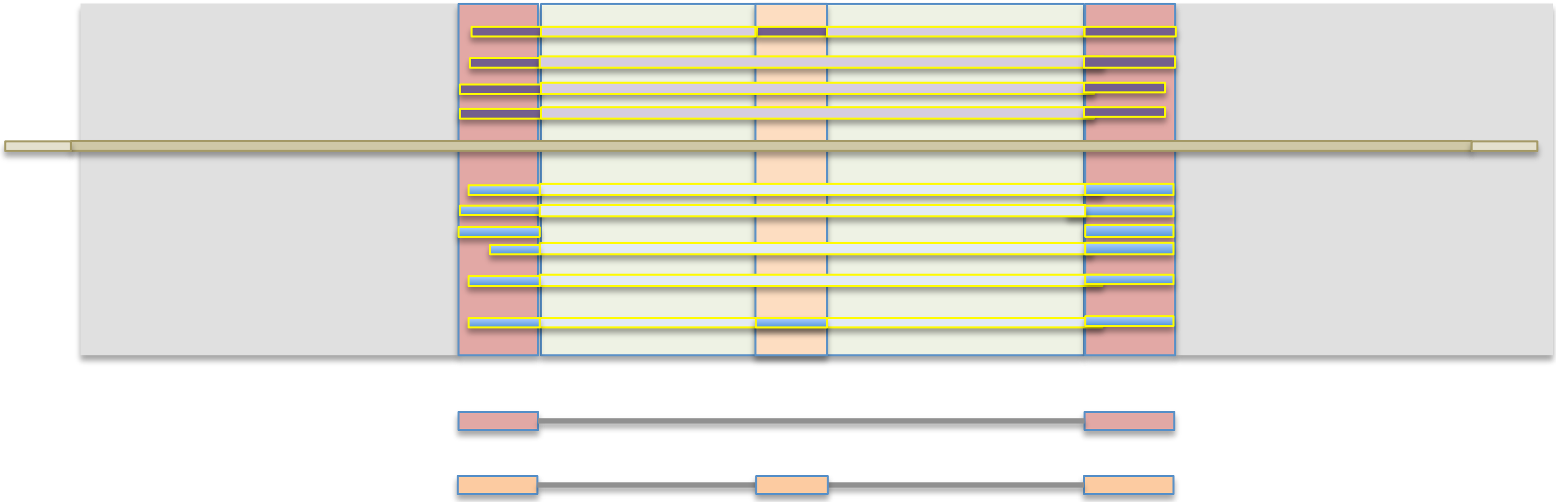
Step 4: Synthesis



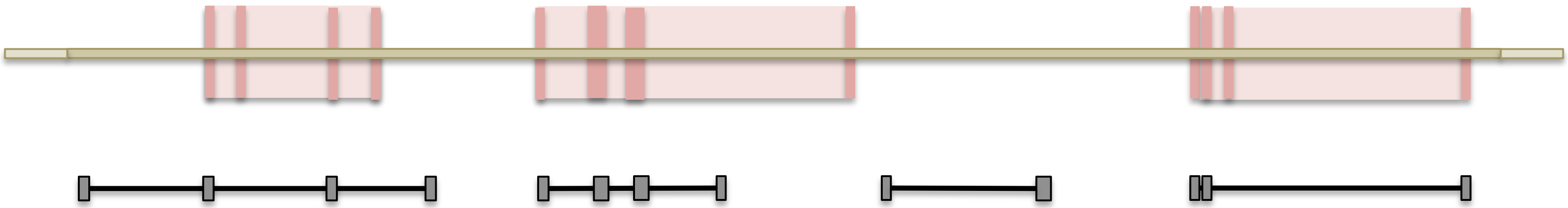
Synthesis refers to the extraction of information to generate evidence for annotations

Done by identifying genomic regions overlapping with sequence features

Step 4: Synthesis



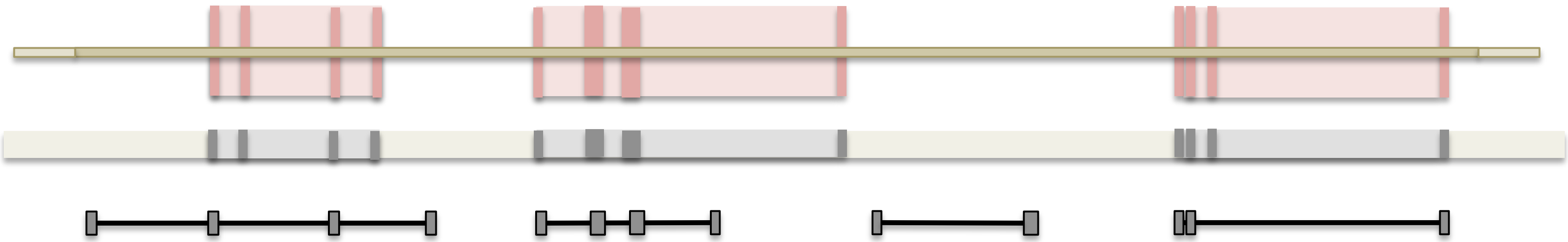
Step 4: Synthesis...and ***ab-initio*** gene finding



Evidence alignments provide support for the identifcation of gene loci

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

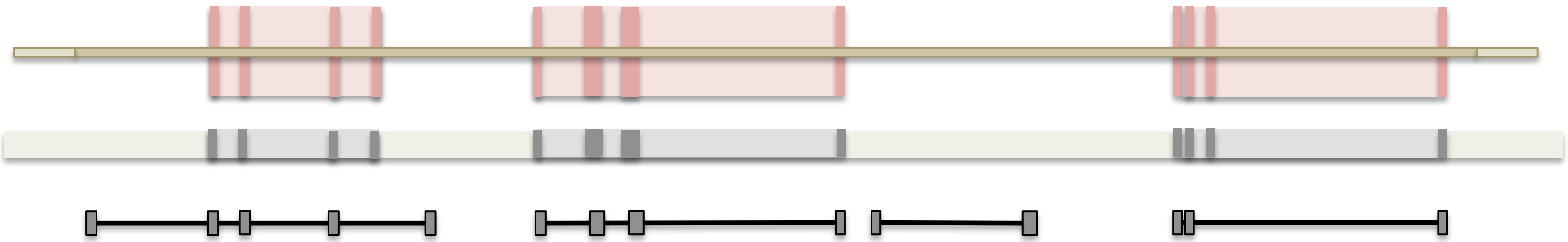
Step 4: Synthesis...and *ab-initio* gene finding



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

Help refine and calibrate a computational inference for a given locus

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

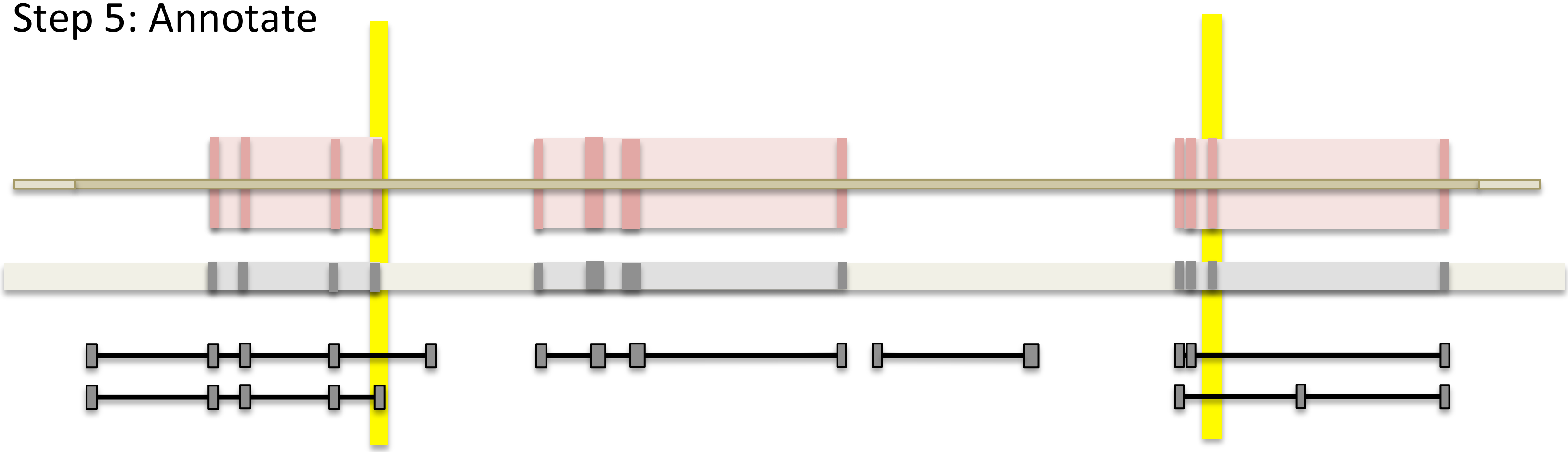


Ab-intio predictions can be improved when evidence is provided (hints)

Help refine and calibrate a computational inference for a given locus

Hints: Introns, intergenic sequence, CDS

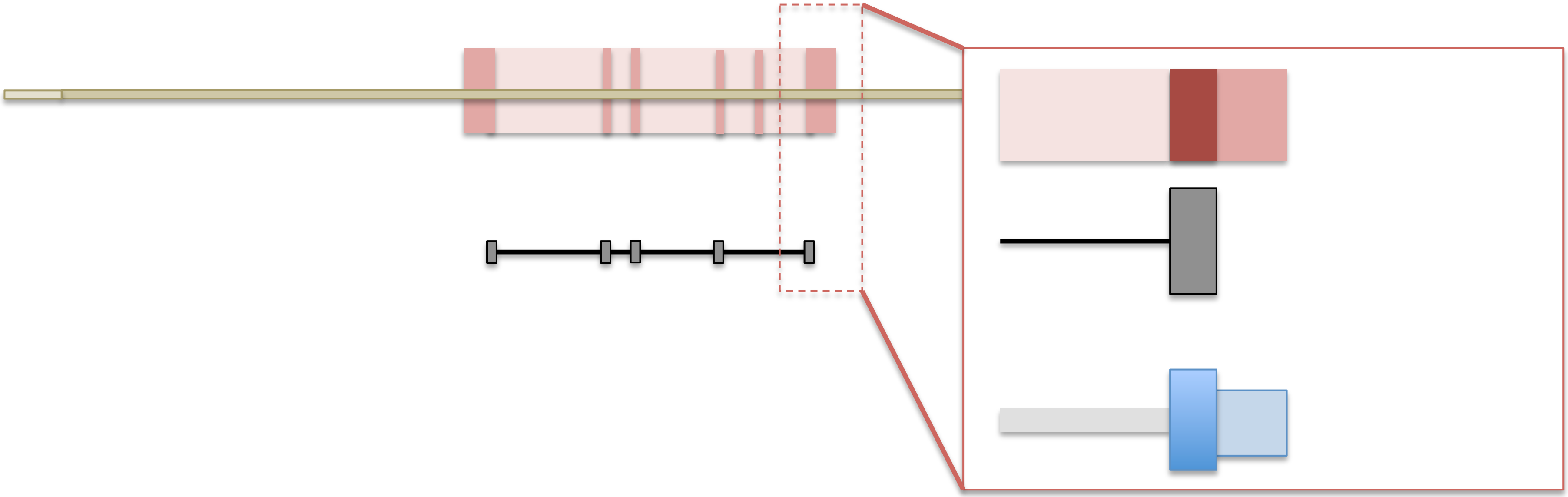
Step 5: Annotate



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

The gene models will be selected in agreement with the available evidences
-> The minimum agreement threshold can be chosen

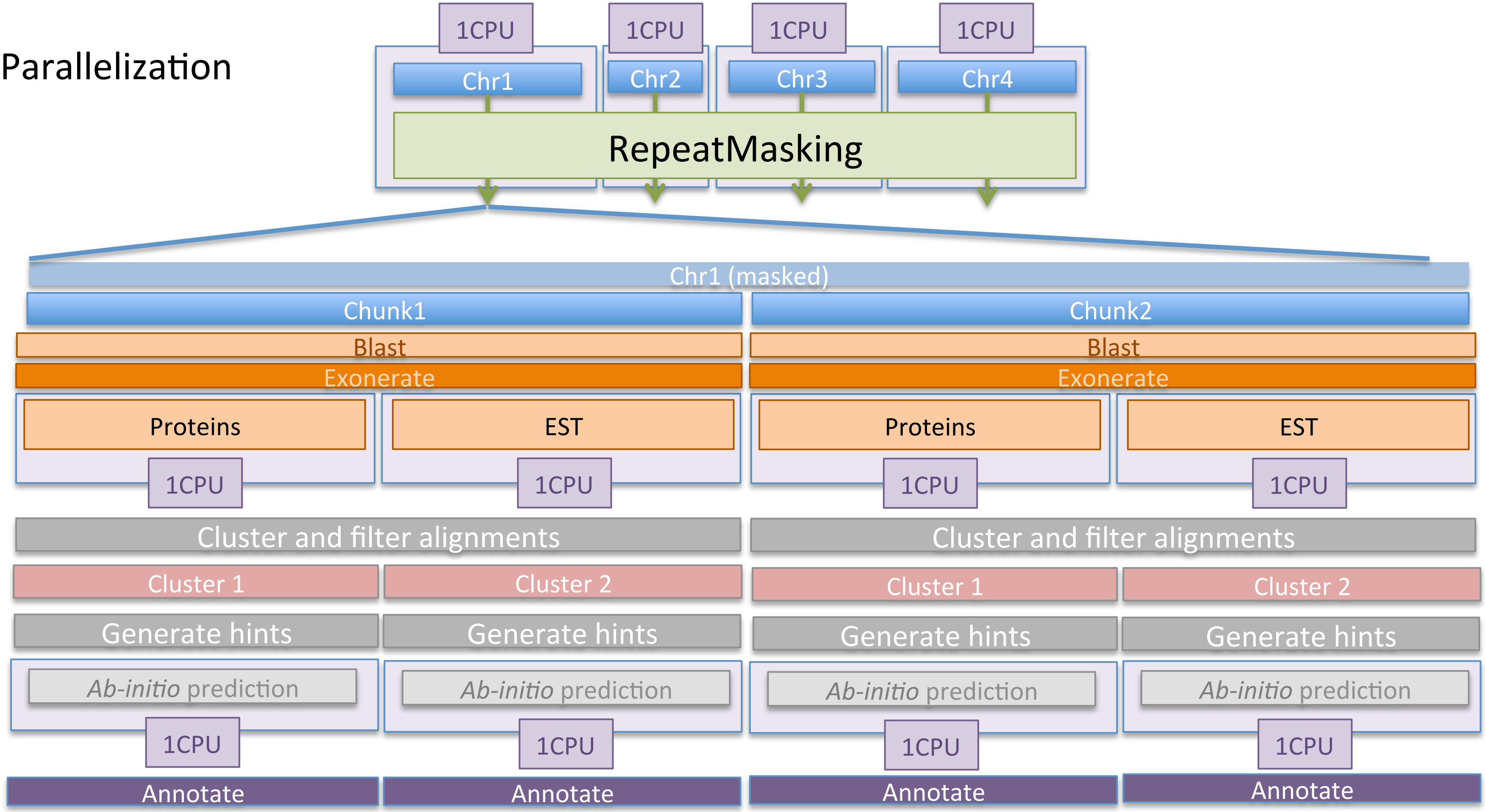
Step 5: Annotate



Synthesized transcript structures are compared against evidence to find UTRs

MAKER2

Parallelization



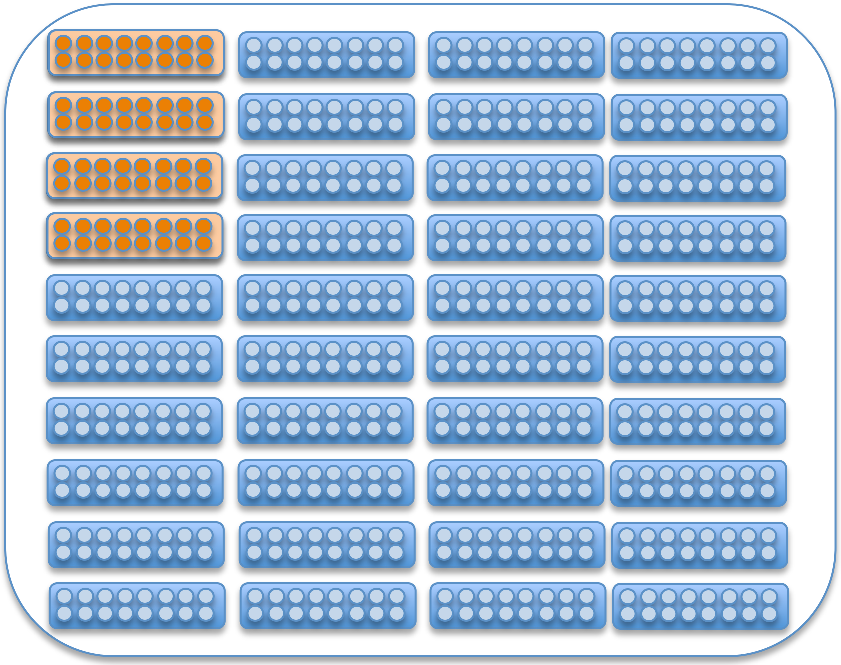
Existing annotation pipelines – MAKER2

Parallelization – Running on Uppmax

Maker uses MPI for job distributon

- runs on almost all computing platforms

- Operates on cores, not nodes



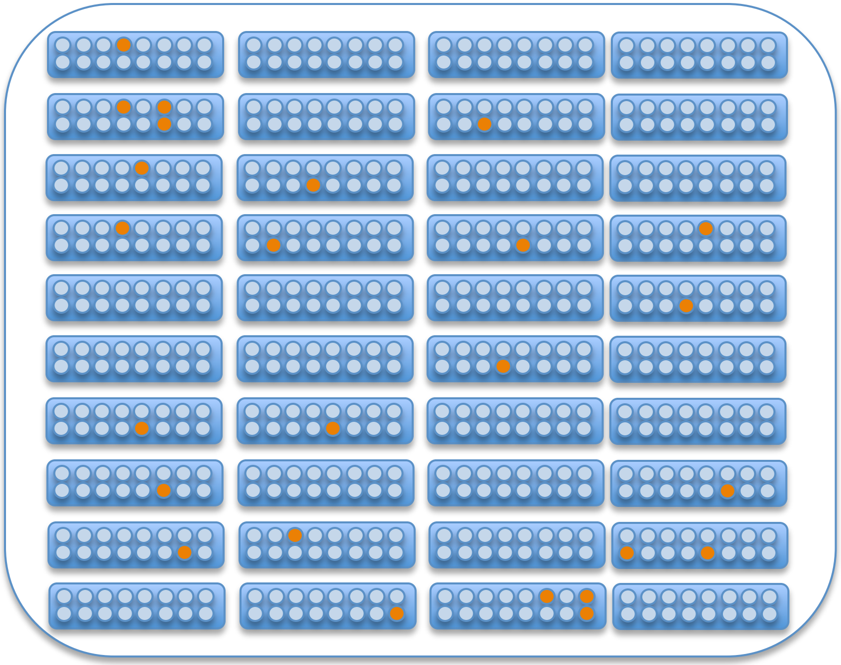
Existing annotation pipelines – MAKER2

Parallelization – Running on Uppmax

Maker uses MPI for job distributon

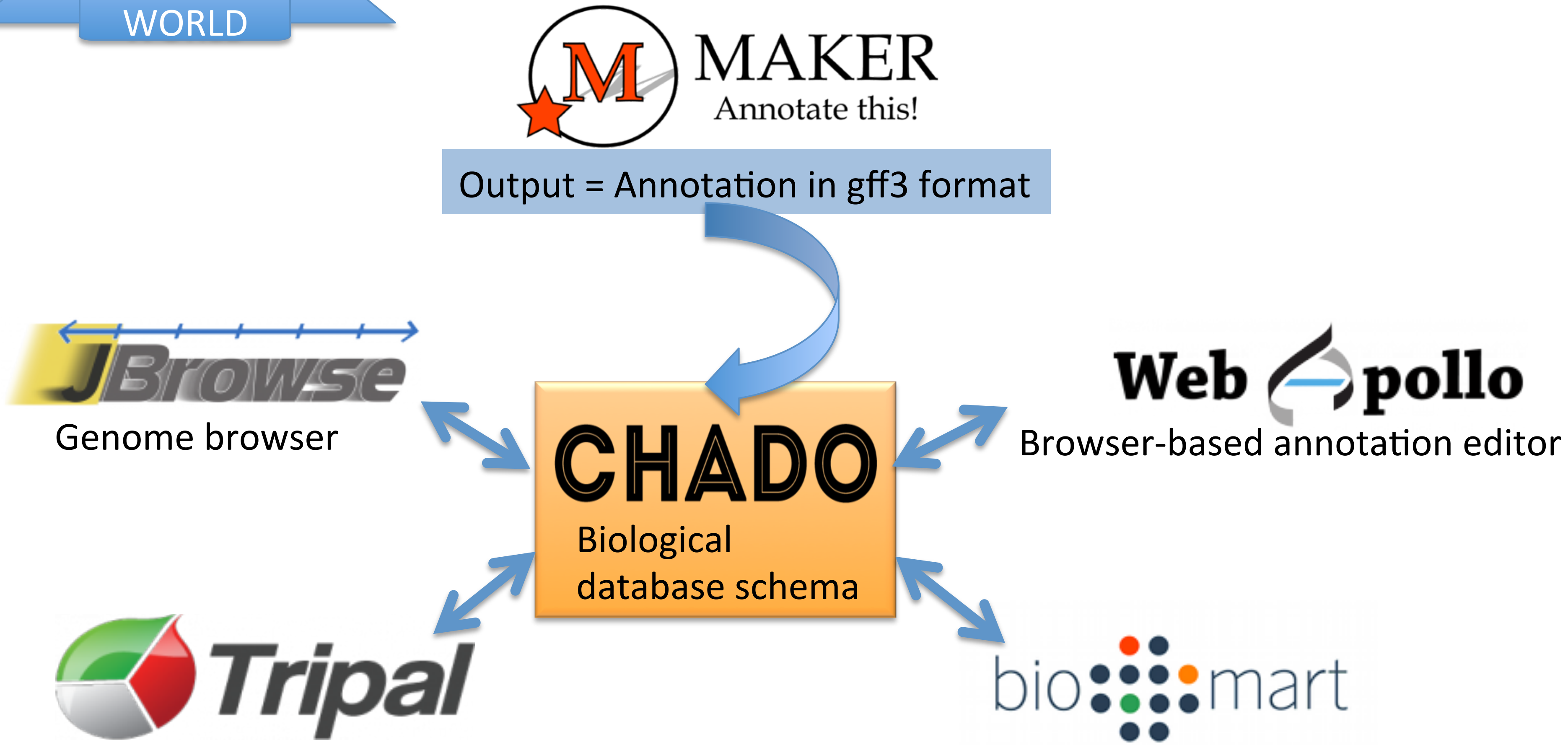
- runs on almost all computing platforms

- Operates on cores, not nodes



MAKER2

GMOD WORLD



Output = Annotation in gff3 format

CHADO

Biological database schema

Genome browser

Browser-based annotation editor

Tripal: Chado web interface

BioMart: Data mining system

4) Prokaryote annotation

Prokaryote annotation

- Prokaryotes have relatively simple gene structure
  - Single open reading frame
  - Alternative start codons: AUG, GUG, UUG

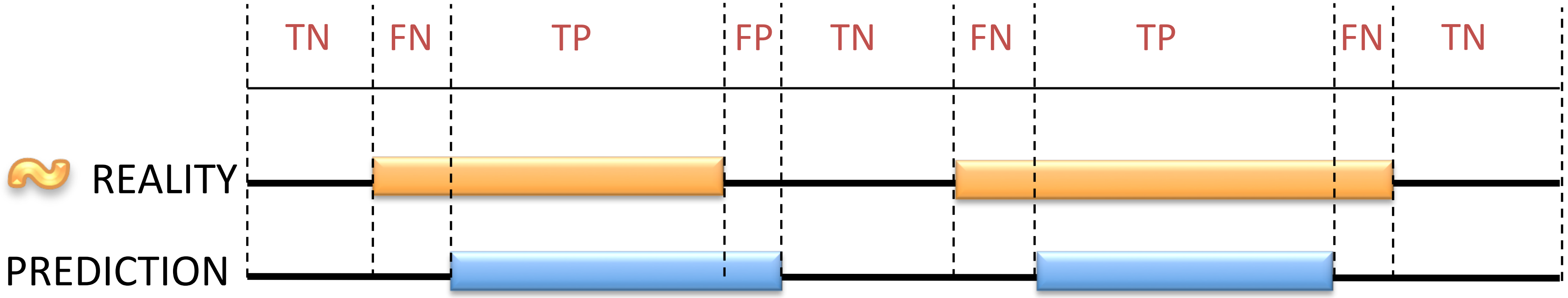
- Gene finders can predict most prokaryotic genes accurately (> 90% sensitivity and specificity)
  - Glimmer
  - Prodigal
  - Genemark-P
  - Eugene-PP

- Pipelines
  - **Prokka**
    - Barrnap (rRNA).
    - Aragorn (tRNA).
    - Infernal ncRNA family profile (misc_RNA).
    - Prodigal used to detect the protein-coding features.
    - SignalP used to detect the signal peptide.

+ Gene name and function inferred by best blast hit
+ Output in different formats : NCBI, gff, etc.

5) Check an annotation

Assess the quality of an annotation:



Sensitivity is the proportion of true predictions to total number of correct genes (including missed predictions)

Specificity is the proportion of true predictions among all predicted genes (including incorrectly predicted ones)

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TP}{TP + FP}$$

Ab Initio methods can approach 100% sensitivity, however as the sensitivity increases, accuracy suffers as a result of increased false positives.

Visualization / Manual curation

Selection of most common visualization or/and Manual curation tools

Name	Standalone	Web tool	Manual curation	year	comment
Artemis	X		X	2000	Can save annotation in EMBL format
IGV	X			2011	Popular
Savant	X			2010	Sequence Annotation, Visualization and ANalysis Tool. enable Plug-ins
Tablet	X		X	2013	
IGB	X			2008	enable Plug-ins. Can load loacl and remote data (dropbox, UCSC genome, etc)
Jbrowse		X		2010	GMOD (successor of Gbrowse)
Web Apollo		X	X	2013	Active community (gmod). Based on Jbrowse. Real-time collaboration
UCSC		X		2000	A large amount of locally stored data must be uploaded to servers across the internet
Ensembl genome browsers		X		2002	A large amount of locally stored data must be uploaded to servers across the internet

FOR AN EXHAUSTIVE LIST: https://en.wikipedia.org/wiki/Genome_browser

6) *To resume / Closing remarks*

Closing remarks

Plethoric choice of methods

year	Gene finder Name	Type	Nb citation	Comments
1991	GRAIL	*Ab initio*		No longer supported
1992	GeneID	*Ab initio*		
1993	GeneParser	*Ab initio*		
1994	Fgeneh	*Ab initio*		Finds single exon only
1996	Genie	Hybrid		
1996	PROCRUSTES	Evidence based		
1997	Fgenes	Hybrid		No download version
1997	GeneFinder	*Ab initio*		Unpublished work
1997	GenScan	*Ab initio*		
1997	HMMGene	*Ab initio*		No download version
1997	GeneWise	Evidence based		
1998	GeneMark.hmm	*Ab initio*		
2000	GenomeScan			
2001	Twinscan			

H

C

C

Closing remarks

Plethoric choice of methods

year	Gene finder Name	Type	Nb citation	Comments
1998	GeneMark.hmm	*Ab initio*		
2000	GenomeScan			
2001	Twinscan			
2002	GAZE			
2004	Ensembl			
2004	GeneZillq/TIGRSCAN	*Ab initio*		No longer supported
2004	GlimmerHMM	*Ab initio*		
2004	SNAP	*Ab initio*		
2006	AUGUSTUS+			
2006	N-SCAN			
2006	TWINSCAN_EST			
2006	N_Scan_EST	Comparative+ Evidence		
2007	Conrad	*Ab initio*		

Closing remarks

Plethoric choice of methods

year	Gene finder Name	Type	Nb citation	Comments
2007	Contrast	Comparative	90	can also incorporate information from EST alignment
2008	Maker			
2009	mGene	*Ab initio*		No longer supported
2015	Ipred	Combiner evidence-based		
2016	BRAKER1	Hybrid		

Hybrid = ab initio and evidence based;
Comparative = genome sequence comparison

List not exhaustive !!

H
C

Closing remarks

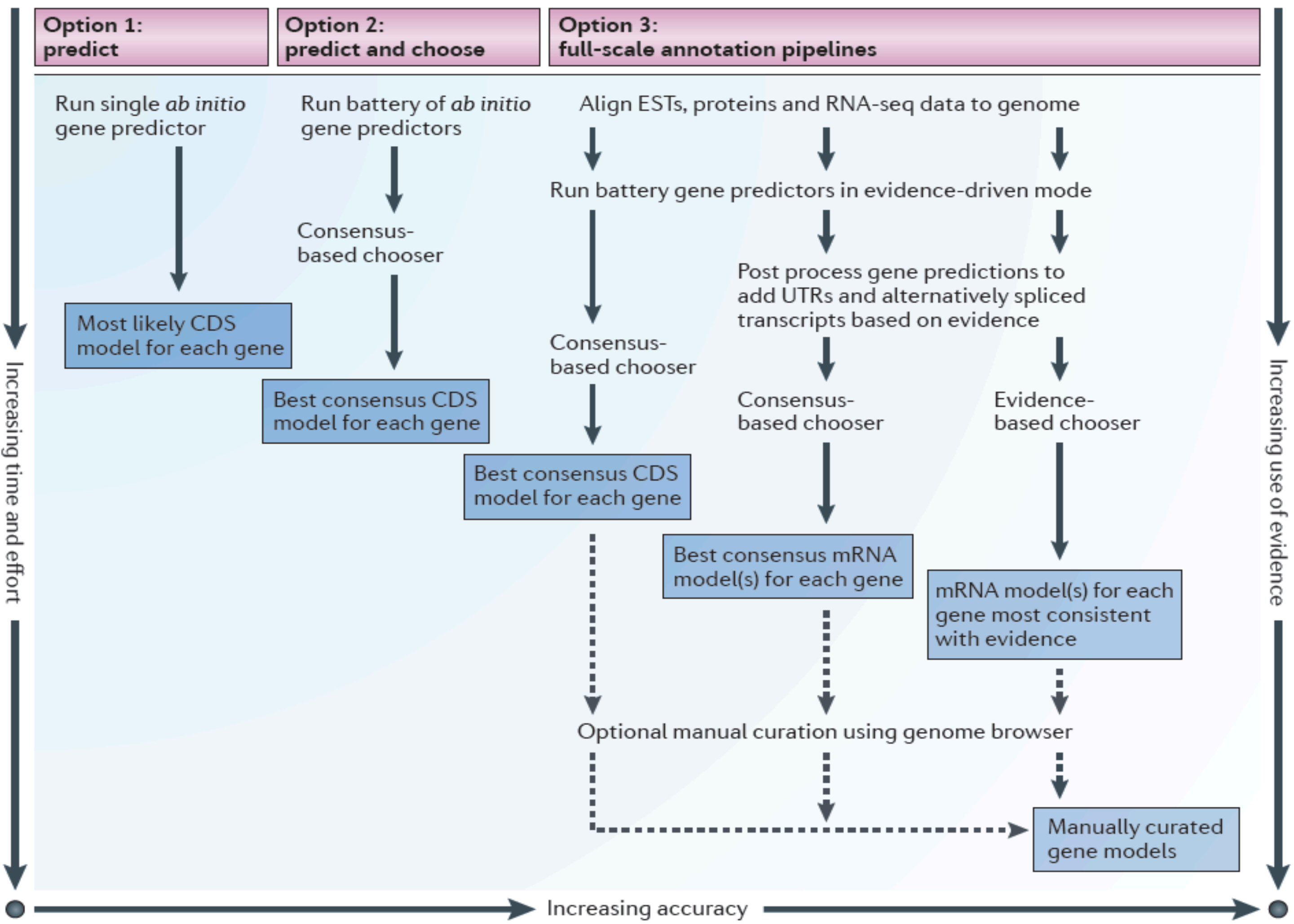
How to choose Method:

- Scientific question behind (need of a conservative annotation vs exhaustive)
- Species dependent (plant / Fungi / eucaryotes)
- phylogenetic relationship of the investigated genome to other annotated genomes (Tera incognita, close, already annotated).
- Data available (hmm profile, RNAseq, etc…)
- Depending on computing ressources (*ab initio* ~ hours < vs > pipeline ~ weeks)
- effort versus accuracy

Closing remarks

How to choose Method:

Mark Yandell & Daniel Ence
Nature Reviews Genetics **13**,329-342
(May 2012) doi:10.1038/nrg3174



Option 1: predict

Run single *ab initio* gene predictor

Most likely CDS model for each gene

Option 2: predict and choose

Run battery of *ab initio* gene predictors

Consensus-based chooser

Best consensus CDS model for each gene

Option 3: full-scale annotation pipelines

Align ESTs, proteins and RNA-seq data to genome

Run battery gene predictors in evidence-driven mode

Consensus-based chooser

Best consensus CDS model for each gene

Post process gene predictions to add UTRs and alternatively spliced transcripts based on evidence

Consensus-based chooser

Best consensus mRNA model(s) for each gene

Evidence-based chooser

mRNA model(s) for each gene most consistent with evidence

Optional manual curation using genome browser

Manually curated gene models

Increasing time and effort

Increasing use of evidence

Increasing accuracy

Figure 2 | Three basic approaches to genome annotation and some common variations. Approaches are compared on the basis of relative time, effort and the degree to which they rely on external evidence, as opposed to ab initio gene models. The y axis shows increasing time and effort; the x axis shows increasing use of external evidence and, consequently, increasing accuracy and completeness of the resulting gene models. The type of final product produced by each kind of pipeline is shown in the dark blue boxes. Relative positions in the figure are for summary purpose

Closing remarks

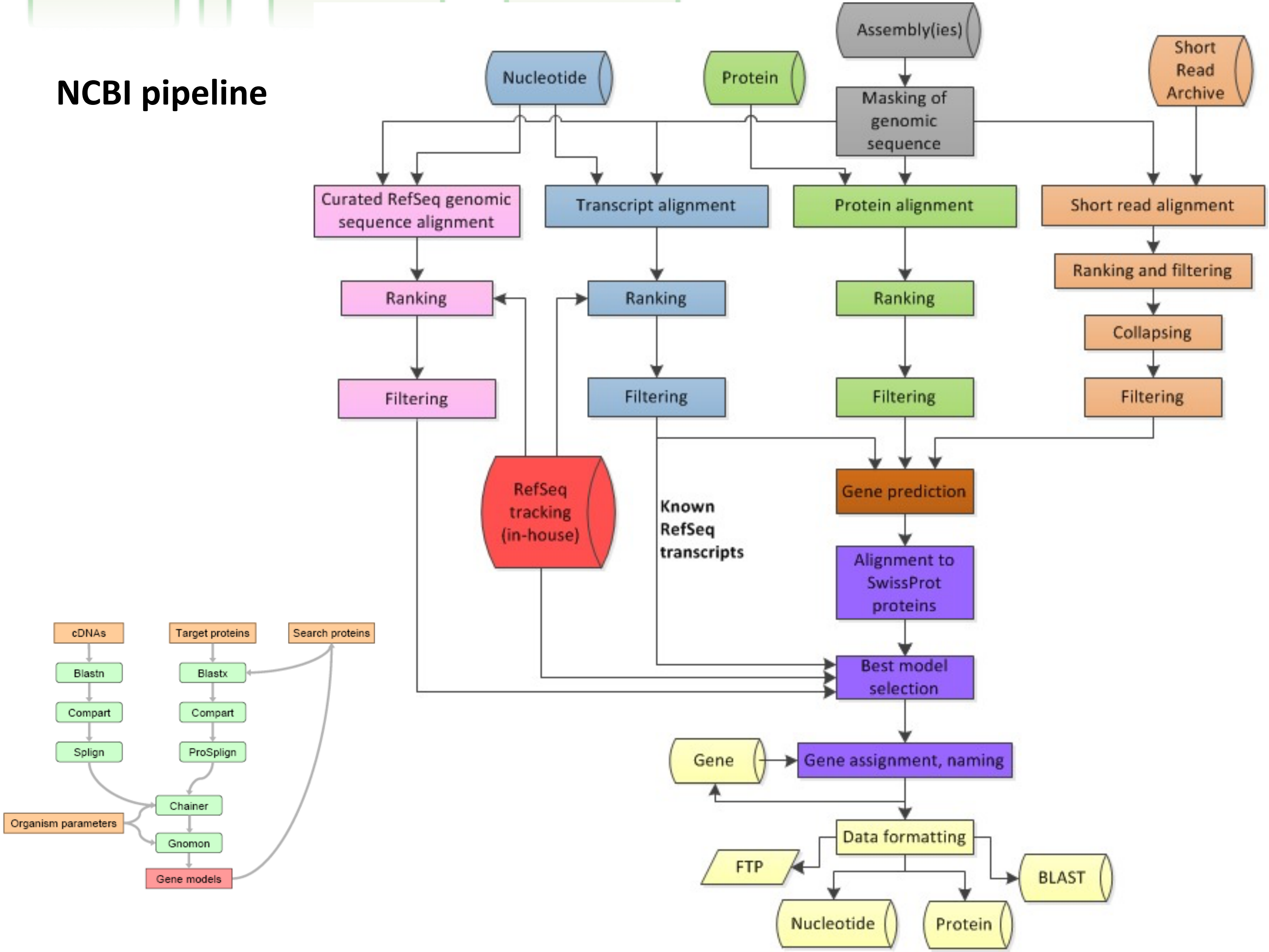
- Pipelines give good results
  MAKER2 the most flexible, adjustable
- Most of methods only build gene models, no **functional inference**
- Computational pipelines make mistakes !!
- Annotation requires **manual curation**
- As for assembly an annotation is never ending, can always be improved (e.g. Human)

=> Practical session will focus on the MAKER2 pipeline

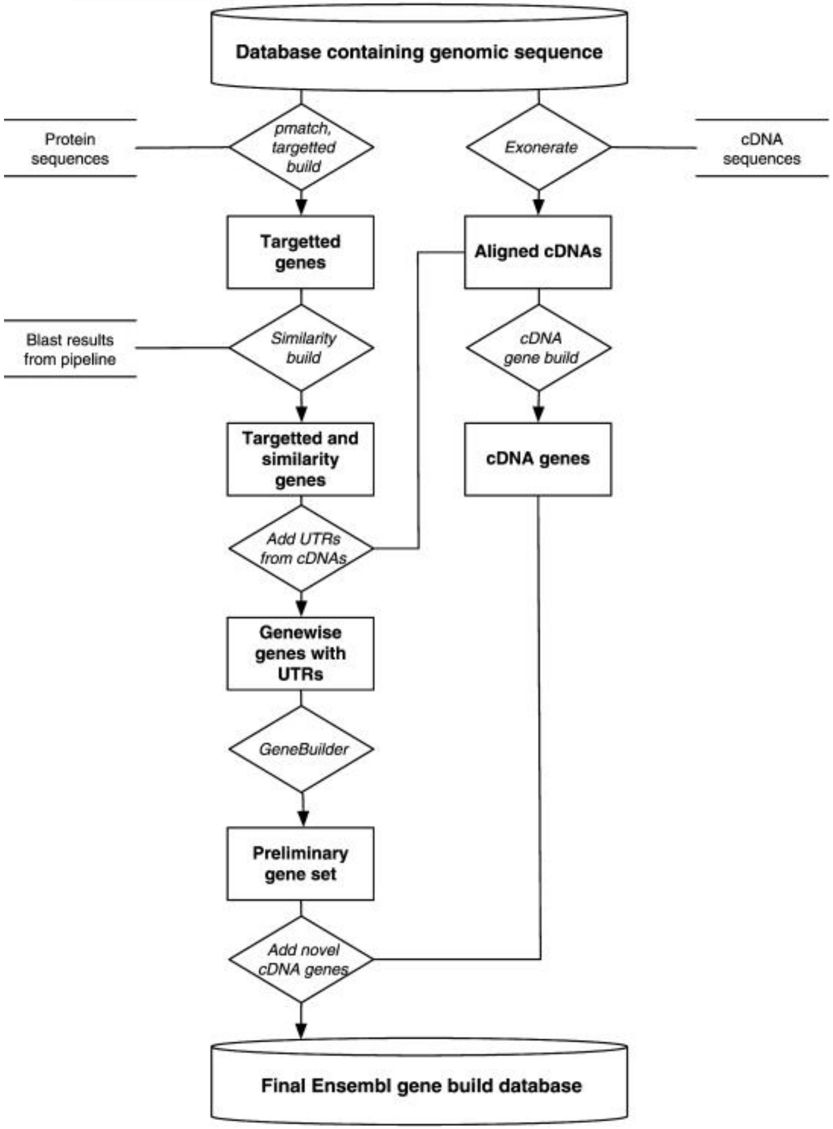
THE END

Supplement

NCBI pipeline



Ensembl pipeline



Database containing genomic sequence

Protein sequences

pmatch, targetted build

Exonerate

cDNA sequences

Targetted genes

Aligned cDNAs

Blast results from pipeline

Similarity build

cDNA gene build

Targetted and similarity genes

cDNA genes

Add UTRs from cDNAs

Genewise genes with UTRs

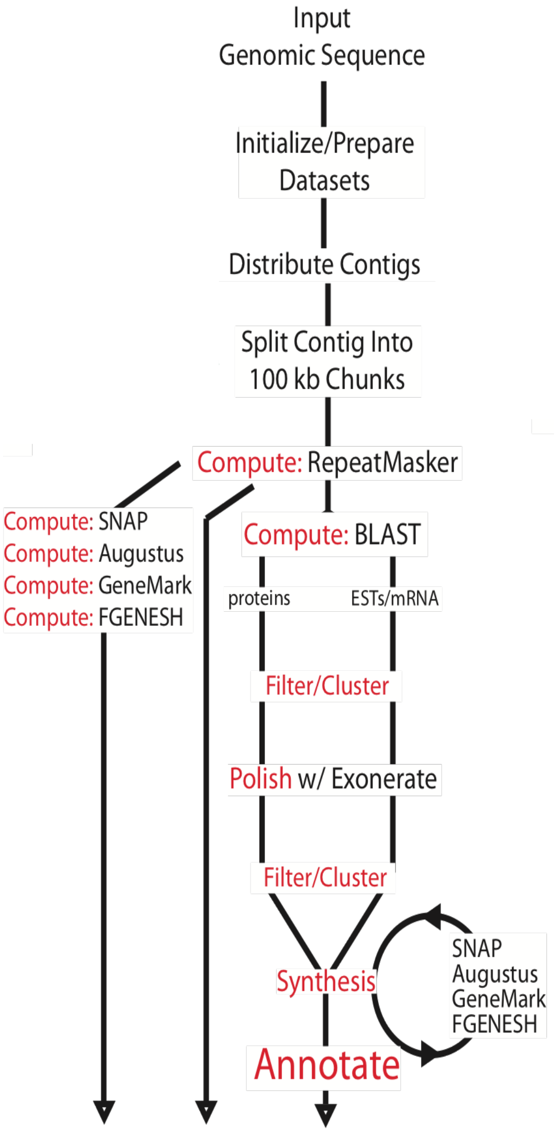
GeneBuilder

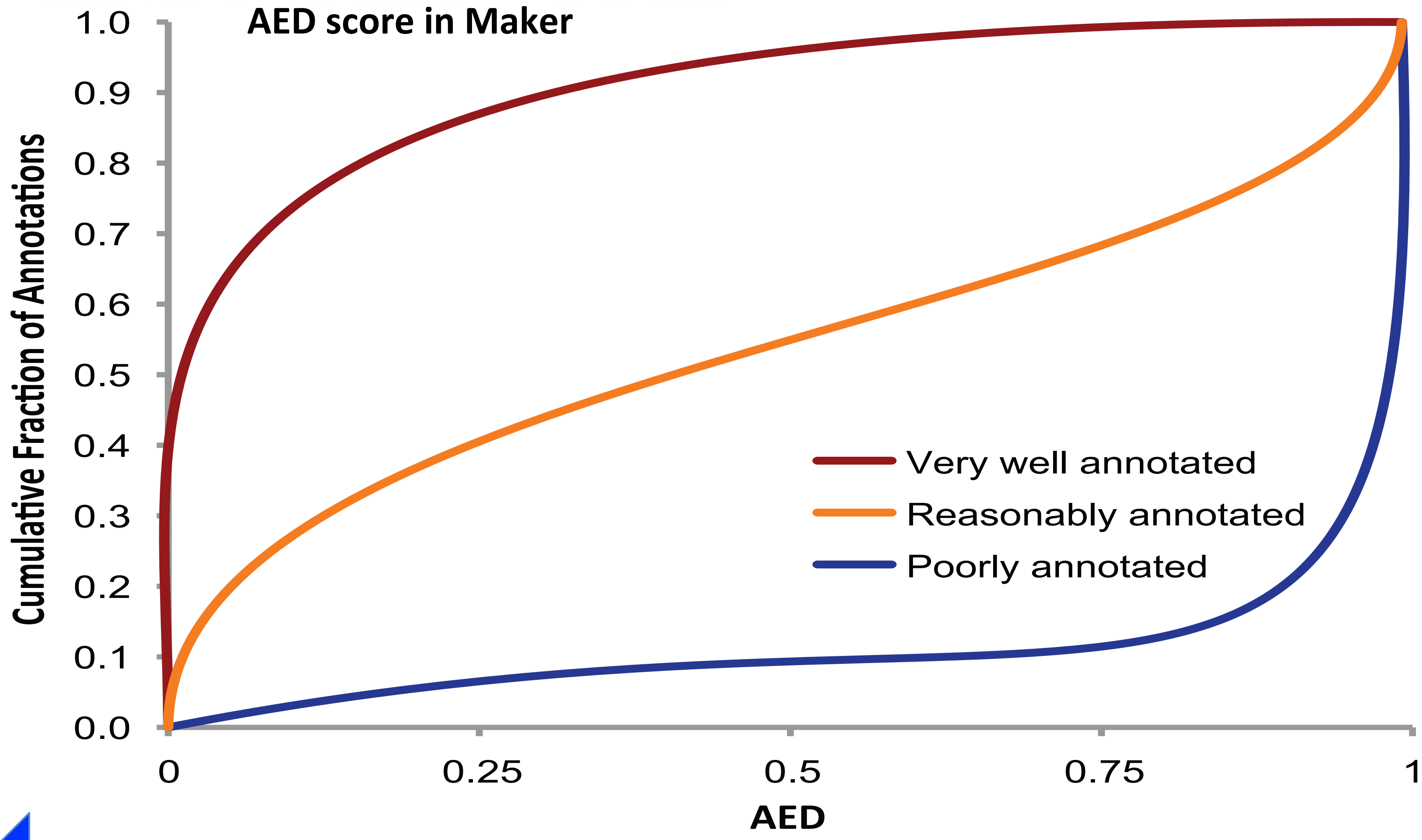
Preliminary gene set

Add novel cDNA genes

Final Ensembl gene build database

Maker pipeline




AED score in Maker
Cumulative Fraction of Annotations
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
0.25
0.5
0.75
1
AED
Very well annotated
Reasonably annotated
Poorly annotated

Better
Quality
Worse