Eukaryotic Genome & Transcriptome Structural Annotation

Jessica Goodheart

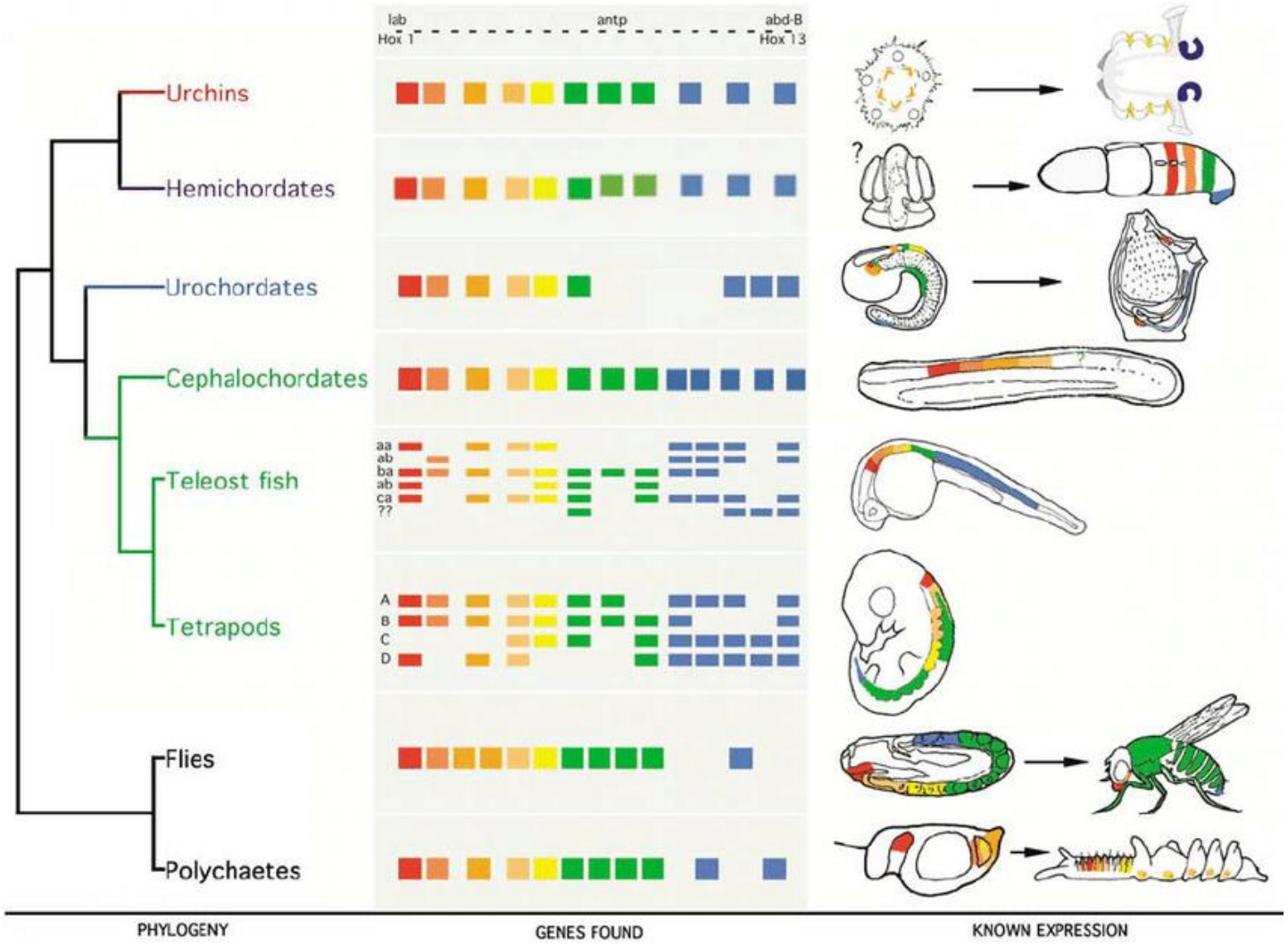
Comparative Genomics 2

1 November 2024

Expected outcomes

- You should be able to:
  - Describe the general goals and steps of structural genome and transcriptome annotation
  - Explain the purpose and process of repeat identification and masking
  - List and describe the types of gene prediction tools for genome and transcriptome annotation
  - Briefly describe the most common algorithms used in annotation
  - Select the appropriate gene prediction tool for your genome or transcriptome

Why annotate genomes and transcriptomes?



Swalla 2006

Genome and Transcriptome Annotation

- Genome annotation consists of five primary steps:
  1. Find and mask repeats
  2. Identifying genes in the genome (ORFs)
  3. Update gene models with alternative splicing events and UTRs
  4. Predict non-coding RNAs
  5. Attaching biological information to genes

- Transcriptome annotation consists of two primary steps:
  1. Identifying coding regions in the transcriptome
  2. Attaching biological information to genes

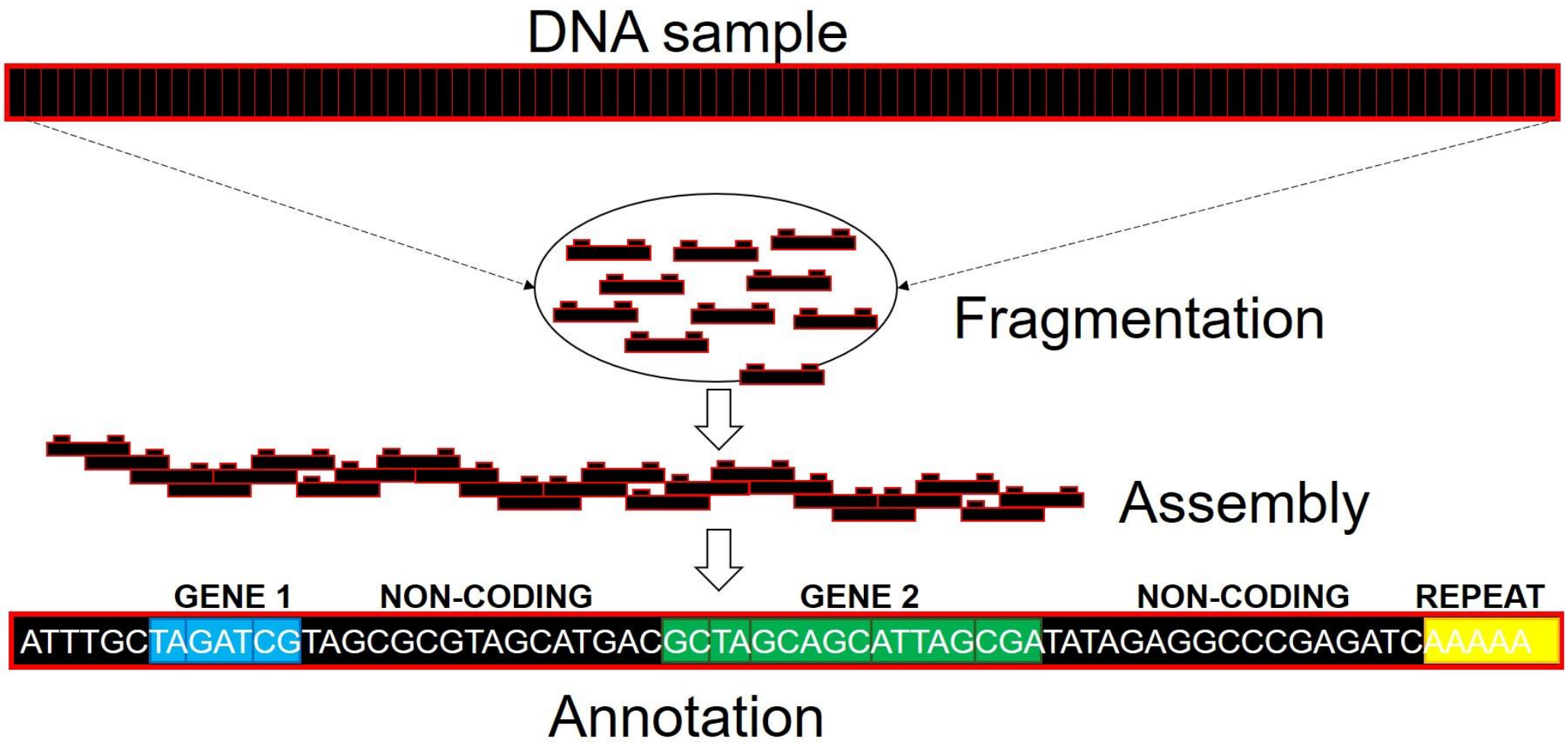
Goal to produce a GFF3 (Gene Feature Format) file

```
##gff-version 3
##Generated using GenSAS, Friday 27th of January 2017 03:31:25 PM
##Project Name : test_011317
##Job Name  : Annotations a1
##Tool      : Publish
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    gene    30779   49550   .       -       .       ID=Ps.00g000010-v2.0.a1;Name=Ps.00g000010;
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    mRNA    30779   49550   .       -       .       ID=Ps.00g000010.m01-v2.0.a1;Name=Ps.00g000010.m
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    49501   49550   .       -       .       ID=Ps.00g000010.m01.exon01-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     49501   49550   4.572   -       0       ID=Ps.00g000010.m01.CDS01-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    48853   48958   .       -       .       ID=Ps.00g000010.m01.exon02-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     48853   48958   8.308   -       1       ID=Ps.00g000010.m01.CDS02-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    42418   42490   .       -       .       ID=Ps.00g000010.m01.exon03-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     42418   42490   9.700   -       0       ID=Ps.00g000010.m01.CDS03-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    35796   35839   .       -       .       ID=Ps.00g000010.m01.exon04-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     35796   35839   9.505   -       2       ID=Ps.00g000010.m01.CDS04-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    34289   34402   .       -       .       ID=Ps.00g000010.m01.exon05-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     34289   34402   12.968  -       0       ID=Ps.00g000010.m01.CDS05-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    30779   30787   .       -       .       ID=Ps.00g000010.m01.exon06-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     30779   30787   10.870  -       0       ID=Ps.00g000010.m01.CDS06-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    gene    63826   74139   .       -       .       ID=Ps.00g000020-v2.0.a1;Name=Ps.00g000020;
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    mRNA    63826   74139   .       -       .       ID=Ps.00g000020.m01-v2.0.a1;Name=Ps.00g000020.m
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    74048   74139   .       -       .       ID=Ps.00g000020.m01.exon01-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     74048   74139   12.403  -       0       ID=Ps.00g000020.m01.CDS01-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    70576   70585   .       -       .       ID=Ps.00g000020.m01.exon02-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     70576   70585   3.360   -       1       ID=Ps.00g000020.m01.CDS02-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    70295   70380   .       -       .       ID=Ps.00g000020.m01.exon03-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     70295   70380   3.052   -       0       ID=Ps.00g000020.m01.CDS03-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    exon    63826   63835   .       -       .       ID=Ps.00g000020.m01.exon04-v2.0.a1;Name=Ps.00g0
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    CDS     63826   63835   0.530   -       1       ID=Ps.00g000020.m01.CDS04-v2.0.a1;Name=Ps.00g00
Ps_scaffold_3113    GenSAS_588bd2c415861-publish    gene    83242   83535   .       -       .       ID=Ps.00g000030-v2.0.a1;Name=Ps.00g000030;
```

Column 1:
Sequence names must match the names of sequences used in the GenSAS project

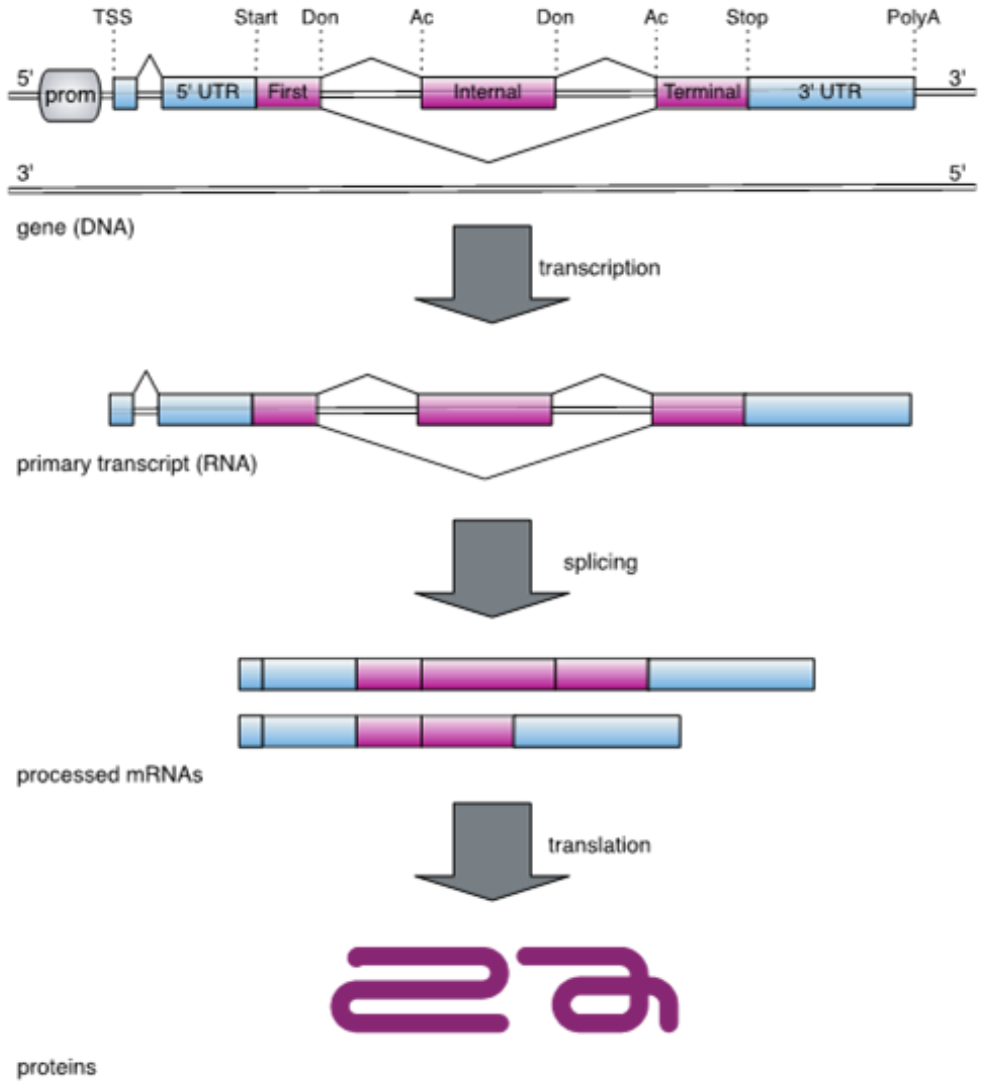
Column 3:
GenSAS looks at this column for the Feature type when importing

First: prediction of genome structures



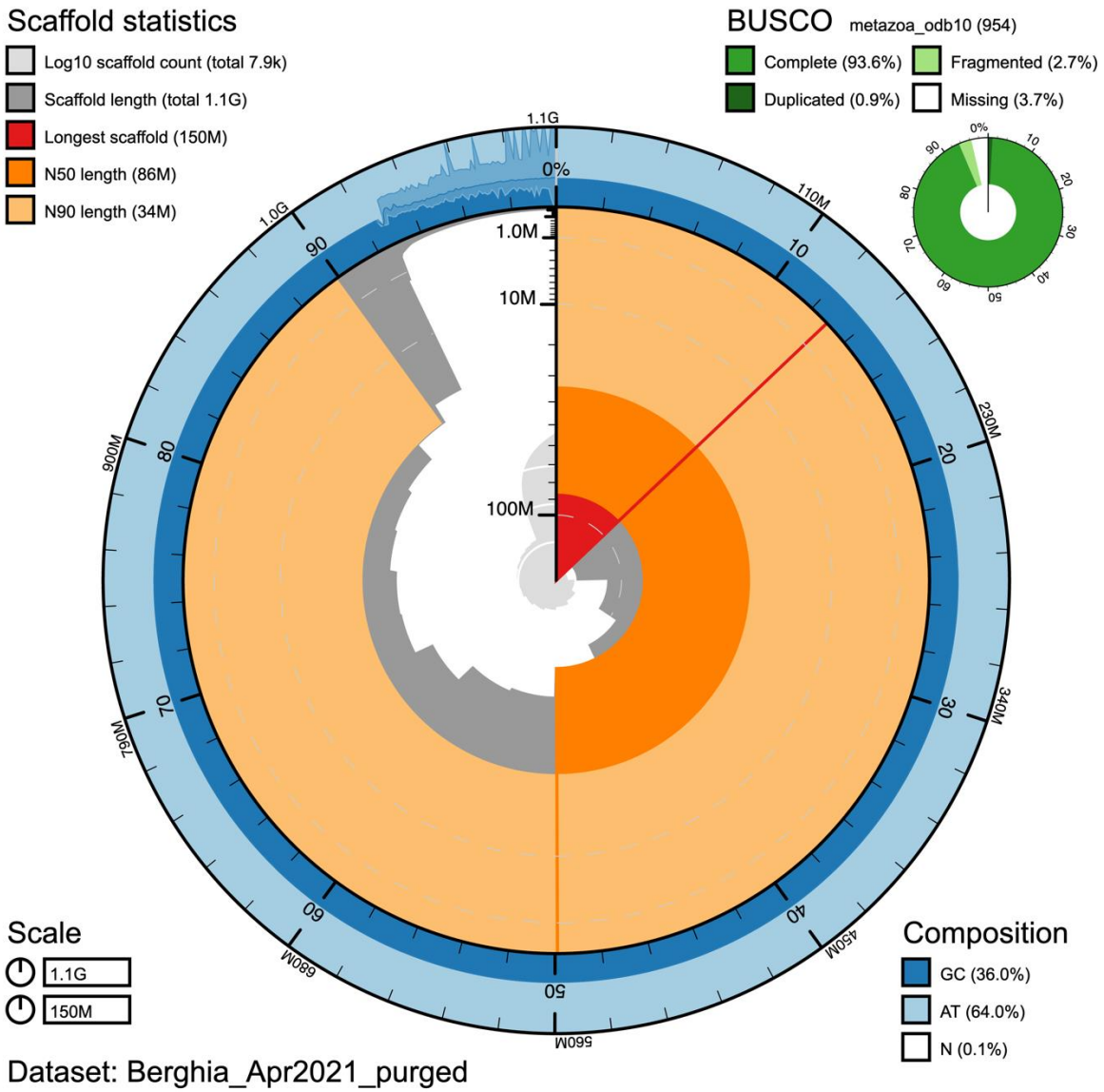
https://theg-cat.com/tag/genome-annotation/

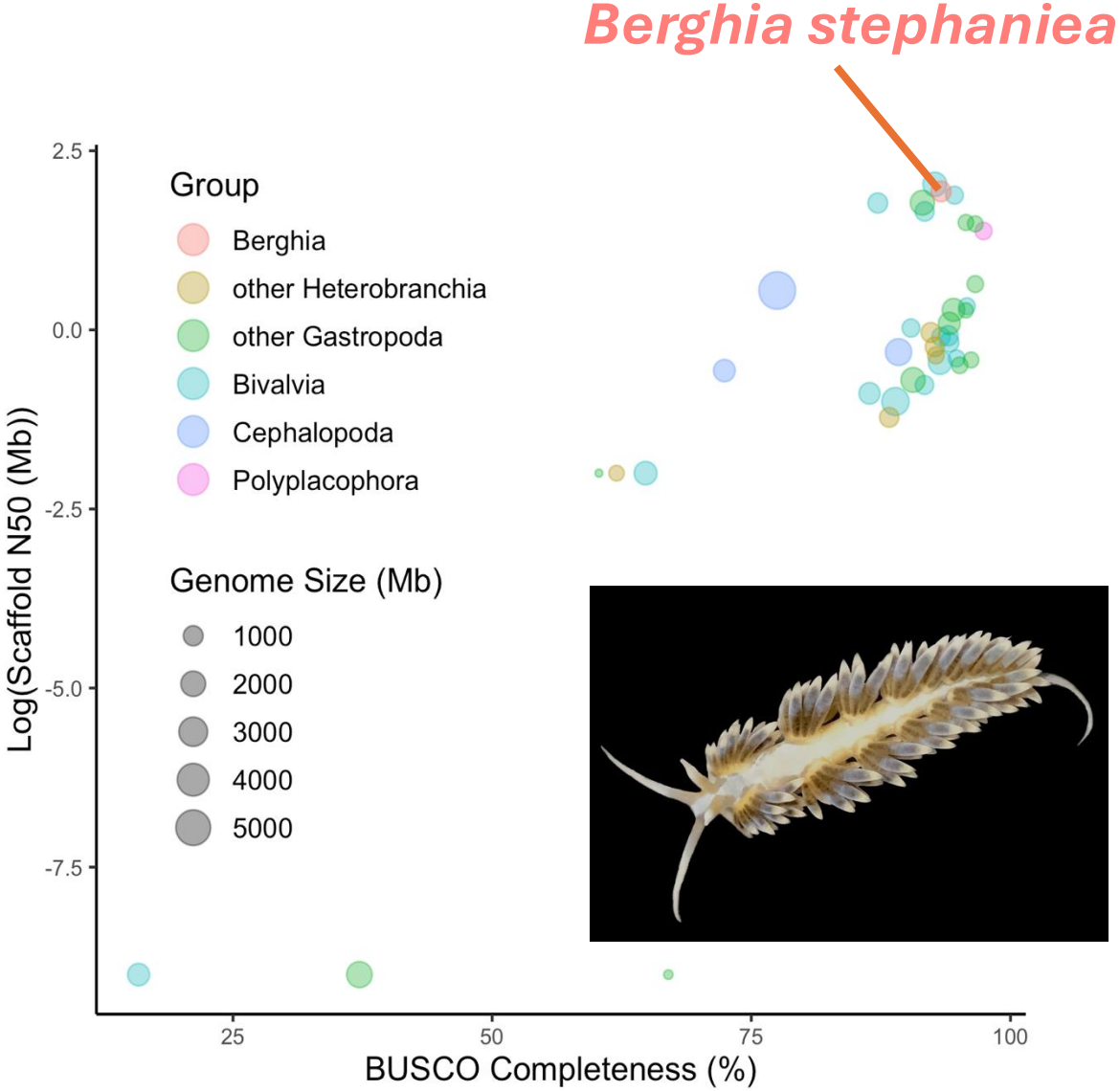
Including the components of gene structure



Tyler Alioto

Berghia genome as an example



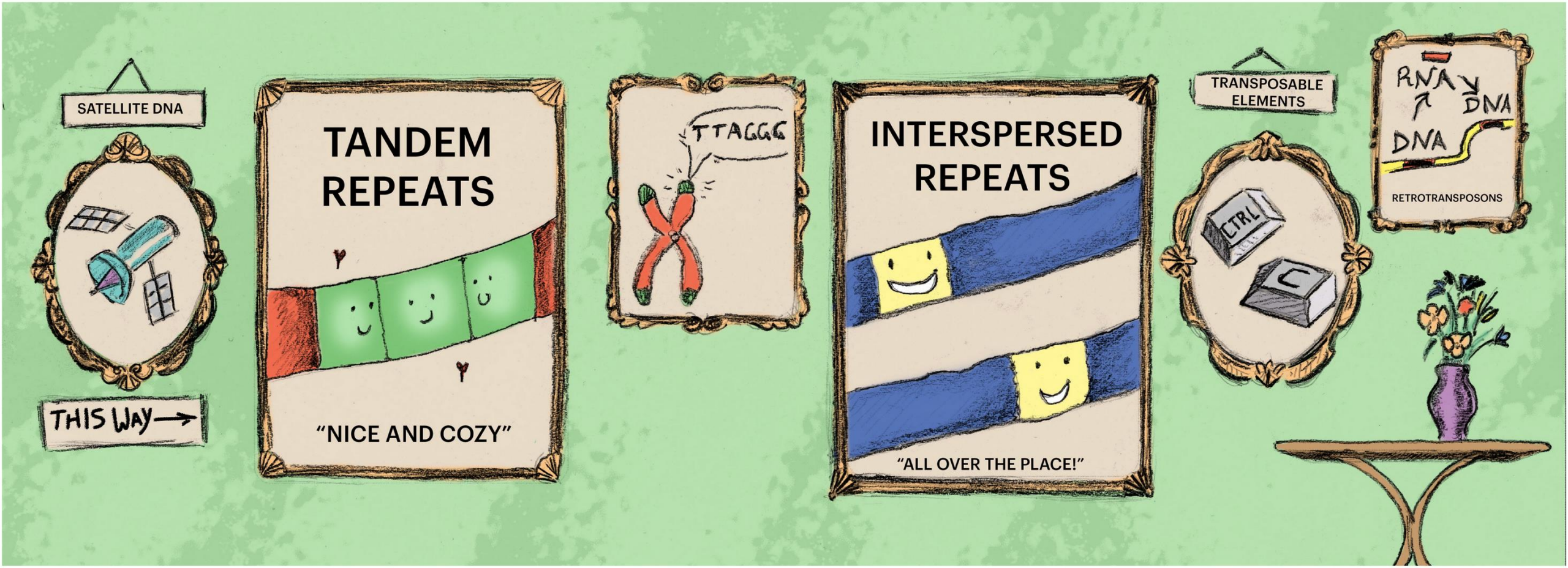


Genome and Transcriptome Annotation

- Genome annotation consists of five primary steps:
  1. Find and mask repeats
  2. Identifying genes in the genome (ORFs)
  3. Update gene models with alternative splicing events and UTRs
  4. Predict non-coding RNAs
  5. Attaching biological information to genes

- Transcriptome annotation consists of two primary steps:
  1. Identifying coding regions in the transcriptome
  2. Attaching biological information to genes

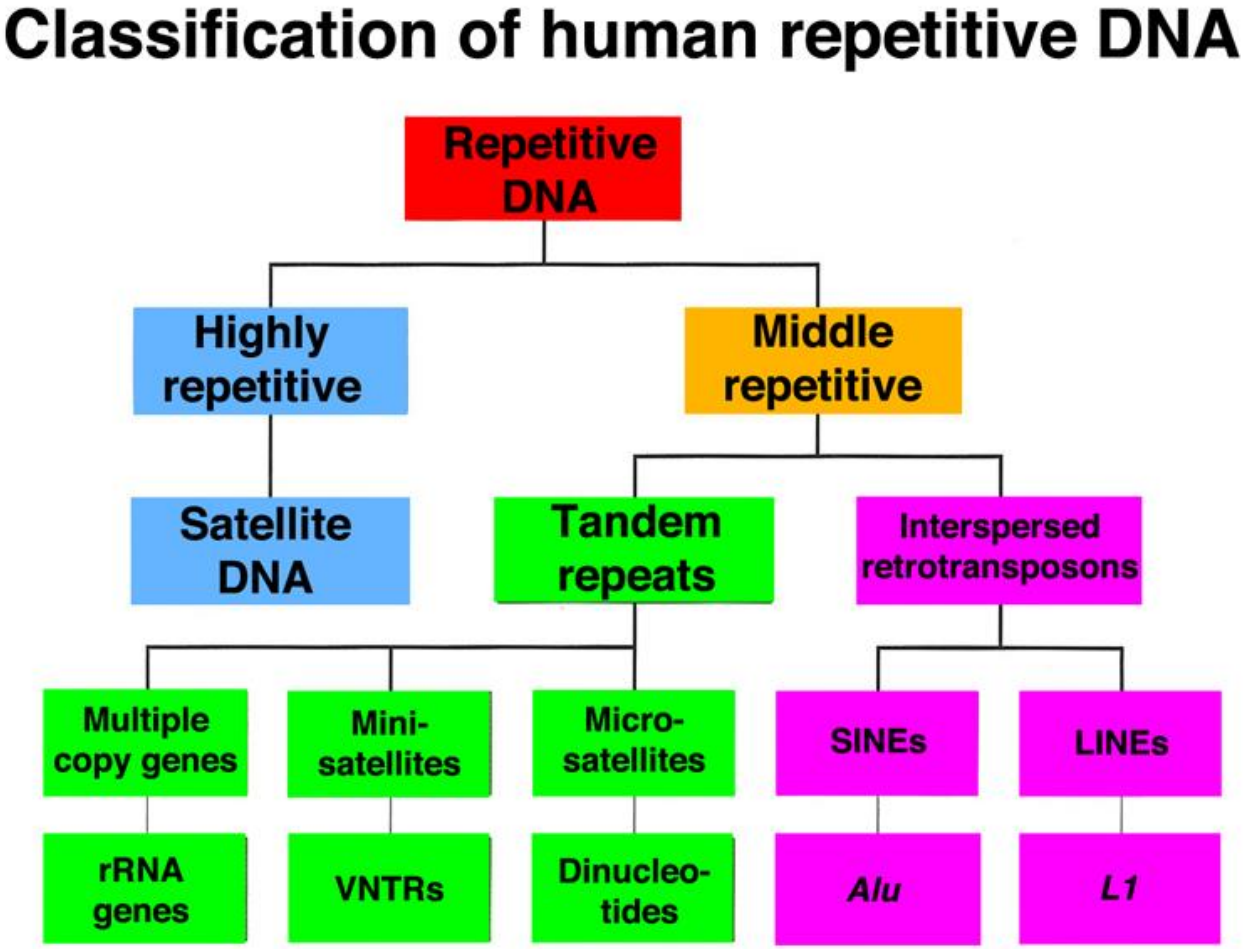
Multiple types of repeats exist



https://wi.mit.edu/many-roles-repetitive-dna-many-roles-repetitive-dna-cartoon-explainer

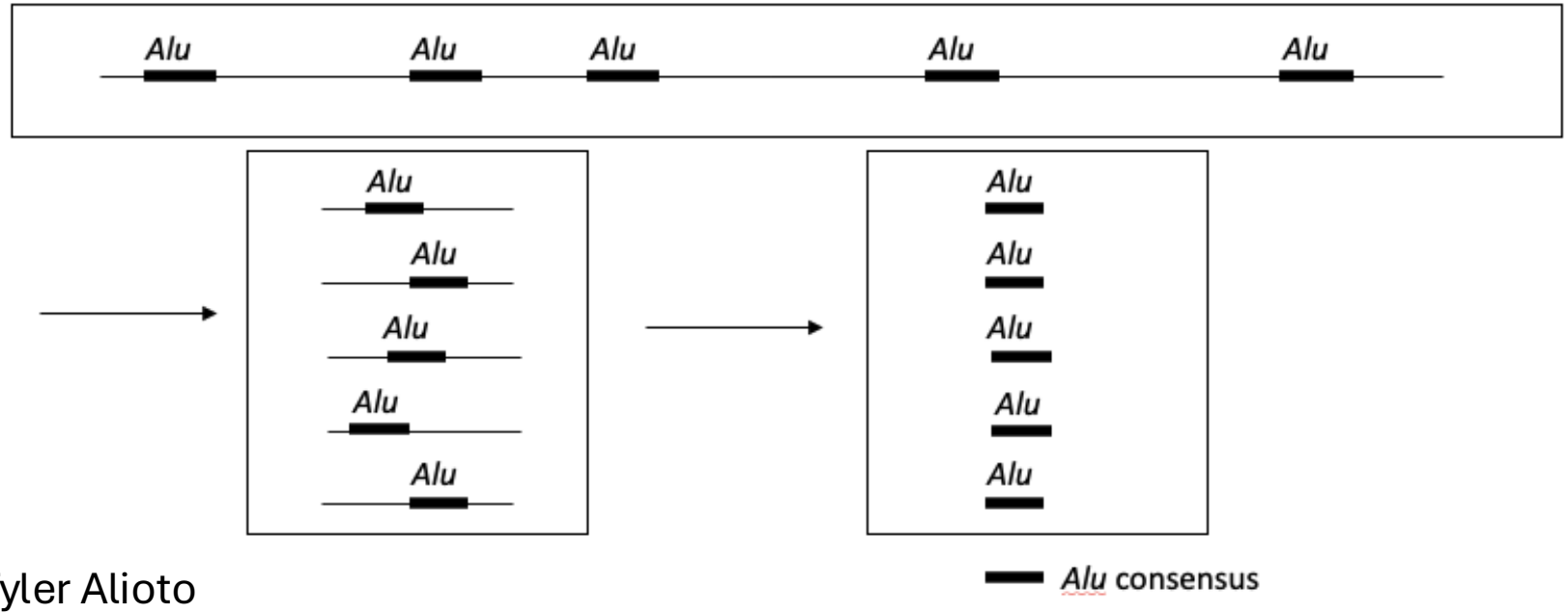
Repeats are common and can be misleading

- Repetitive sequences are found throughout genomes and are more likely to cause non-specific gene hits
- Can range from ~10-85% of a genome depending on the species
- BUT, repeats can be biologically meaningful



Repeats can be challenging to identify

- Regions containing repeats and repeat boundaries are not known *a priori*
- Some repeat occurrences appear as partial copies
- Risk of overmasking due to high-identity gene families



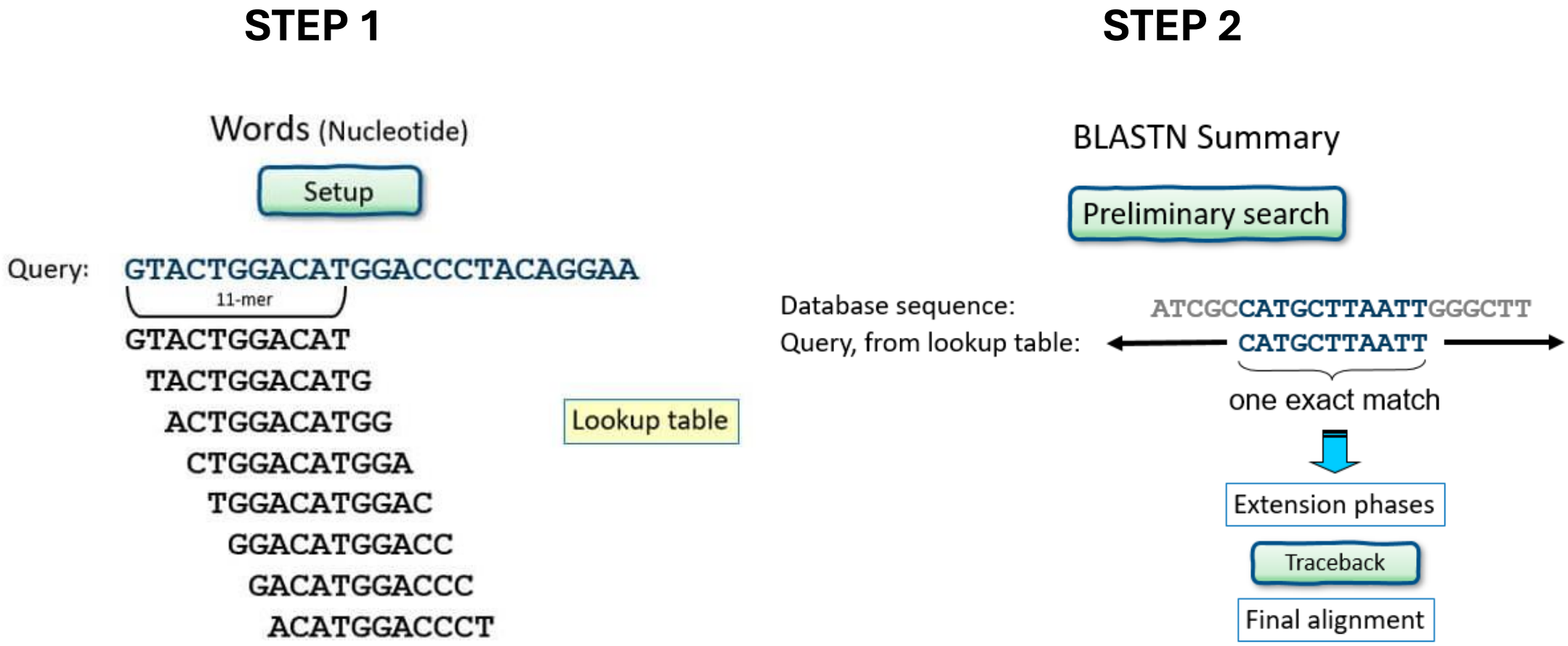
Tyler Alioto

Methods of finding repeats – Homology based

- Identification by finding sequences similar to known repeats
- Comparisons to databases like RepBase, Dfam, msRepDB, REXdb, and Pfam
- RepeatMasker is an example of such a tool, which uses Dfam or RepBase as the library and RMBLAST as the aligner
  - Other examples include Censor, TESeeker, Greedier, and T-le

Advantages: accuracy and high efficacy with small number of copies
Disadvantages: Cannot be used for new repeat discovery

BLAST compares query sequences to a DB



Methods of finding repeats – Structure based

- Repeats, and particularly TEs, often have specific structures
- Rely on prior knowledge of structural features of known repeats in a library
- Use a heuristic algorithm that uses prior info to identify repeats
- Examples include LTRharvest, MASiVE, SINE-finder, etc

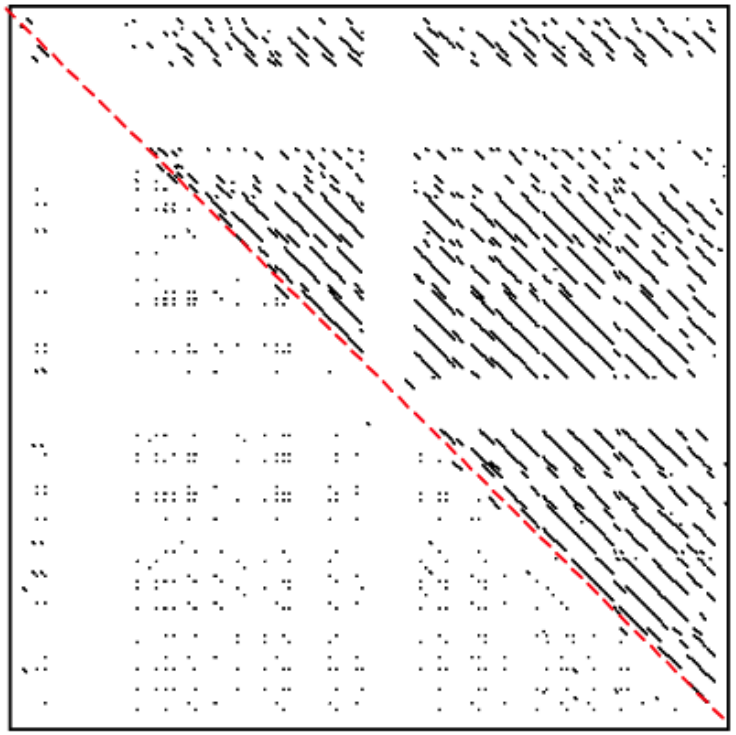
Advantages: high detection efficiency and lower false-positive rate, easy to verify and classify detected repeats
Disadvantages: Cannot be used to discover new repeats with unknown structure, relies heavily on precision and completeness of input sequences.

Methods of finding repeats – *de novo*

- Can be classified into three categories based on the core technology of the method:
  - Type 1 – Uses local multiple sequence alignments. Examples: PILER, RECON
  - Type 2 – High-frequency *k-mers* and space seed extension. Examples: RepeatModeler, RepeatScout
  - Type 3 – Sequence similarity networks built from de novo sequence assembly

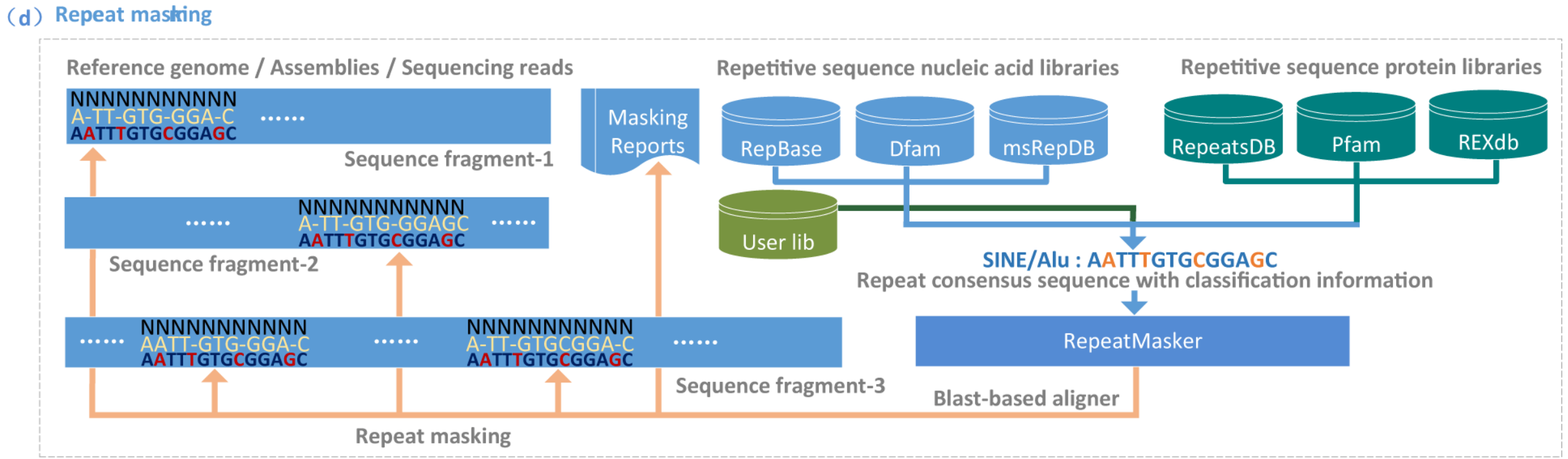
Each type has different advantages and disadvantages, but computing costs vary widely and detection can be inaccurate

Type 1



Edgar & Meyers 2003

Classification and masking repeats



Liao et al. 2023, Figure 5

Repeat modeler commands

Compile initial repeat library (with Classifications):

```
RepeatModeler -database berghia -pa 2 -LTRStruct
```

Blast proteome against RepeatMasker TE database:

```
blastp -query
../Berghia_alltissues_onerep_trinity291_transdecoder_cdhit95_noaliens_fulltranscripts_novectors_nocontaminants.fasta.trans
decoder.pep -db
/ocean/projects/bio210009p/shared/tools/miniconda3/envs/repeatmodeler/share/RepeatMasker/Libraries/RepeatPeps.lib -outfmt
'6 qseqid staxids bitscore std sscinames sskingdoms stitle' -max_target_seqs 25 -culling_limit 2 -num_threads 8 -evalue
1e-5 -out Bsteph_ref.pep.vs.RepeatPeps.25cul2.1e5.blastp.out
```

Remove TEs from proteome:

```
fastaqual_select.pl -f
../Berghia_alltissues_onerep_trinity291_transdecoder_cdhit95_noaliens_fulltranscripts_novectors_nocontaminants.fasta.trans
decoder.cds -e <(awk '{print $1}' Bsteph_re..pep.vs.RepeatPeps.25cul2.1e5.blastp.out | sort | uniq) >
Bsteph_ref.fa.no_tes.fa
```

Blast proteome against RepeatModeler library:

```
makeblastdb -in Bsteph_ref.fa.no_tes.fa -dbtype nucl

blastn -task megablast -query consensi.fa.classified -db Bsteph_ref.fa.no_tes.fa -outfmt '6 qseqid staxids bitscore std
sscinames sskingdoms stitle' -max_target_seqs 25 -culling_limit 2 -num_threads 8 -evalue 1e-10 -out
repeatmodeller_lib.v.Bsteph_ref.fa.no_tes.25cul2.1e10.megablast.out
```

Remove hits from RepeatModeler library:

```
fastaqual_select.pl -f consensi.fa.classified -e <(awk '{print $1}'
repeatmodeller_lib.v.Bsteph_ref.fa.no_tes.25cul2.1e25.megablast.out | sort | uniq) >
consensi.fa.classified.filtered_for_CDS_repeats.fa
```

Goodheart et al. 2024, Additional File 12

Table S2. RepeatModeler analysis for the *Berghia stephanieae* genome.

Category	Type	Number of Elements	Length Occupied (bp)	Percentage of Sequence
Retroelements		204933	54014008	5.16
	SINEs	17251	1753942	0.17
	Penelope	9522	2726192	0.26
	LINEs	180146	48023476	4.58
	CRE/SLACS	0	0	0
	L2/CR1/Rex	98626	23489586	2.24
	R1/LOA/Jockey	2230	920175	0.09
	R2/R4/NeSL	18134	4183768	0.4
	RTE/Bov-B	36219	11988222	1.14
	L1/CIN4	0	0	0
	LTR elements	7536	4236590	0.4
	BEL/Pao	386	384152	0.04
	Ty1/Copia	258	399159	0.04
	Gypsy/DIRS1	6892	3453279	0.33
	Retroviral	0	0	0
DNA transposons		91824	15755392	1.5
	hobo-Activator	10142	1790783	0.17
	Tc1-IS630-Pogo	19968	4791551	0.46
	En-Spm	0	0	0
	MuDR-IS905	0	0	0
	PiggyBac	3247	509014	0.05
	Tourist/Harbinger	0	0	0
	Other (Mirage, P-	25970	3071959	0.29
Rolling-circles	-	503	229700	0.02
Unclassified	-	1991187	287564488	27.45
Total interspersed	-		357333888	34.11
Small RNA	-	16841	1391635	0.13
Satellites	-	6364	998671	0.1
Simple repeats	-	1204558	121713746	11.62
Low complexity	-	103563	7316572	0.7

~50% of the *Berghia* genome is made up of repeats

RepeatMasker commands

Hardmask for STAR aligner:

```
RepeatMasker Berghia_Apr2021_hirise_purged.filtered.fasta -e ncbi -lib
RM_24730.MonJul121509252021/consensi.fa.classified.filtered_for_CDS_repeats
.fa -pa 20
```

Softmask for Augustus/BRAKER:

```
RepeatMasker Berghia_Apr2021_hirise_purged.filtered.fasta -e ncbi -lib
RM_24730.MonJul121509252021/consensi.fa.classified.filtered_for_CDS_repeats
.fa -xsmall -pa 20
```

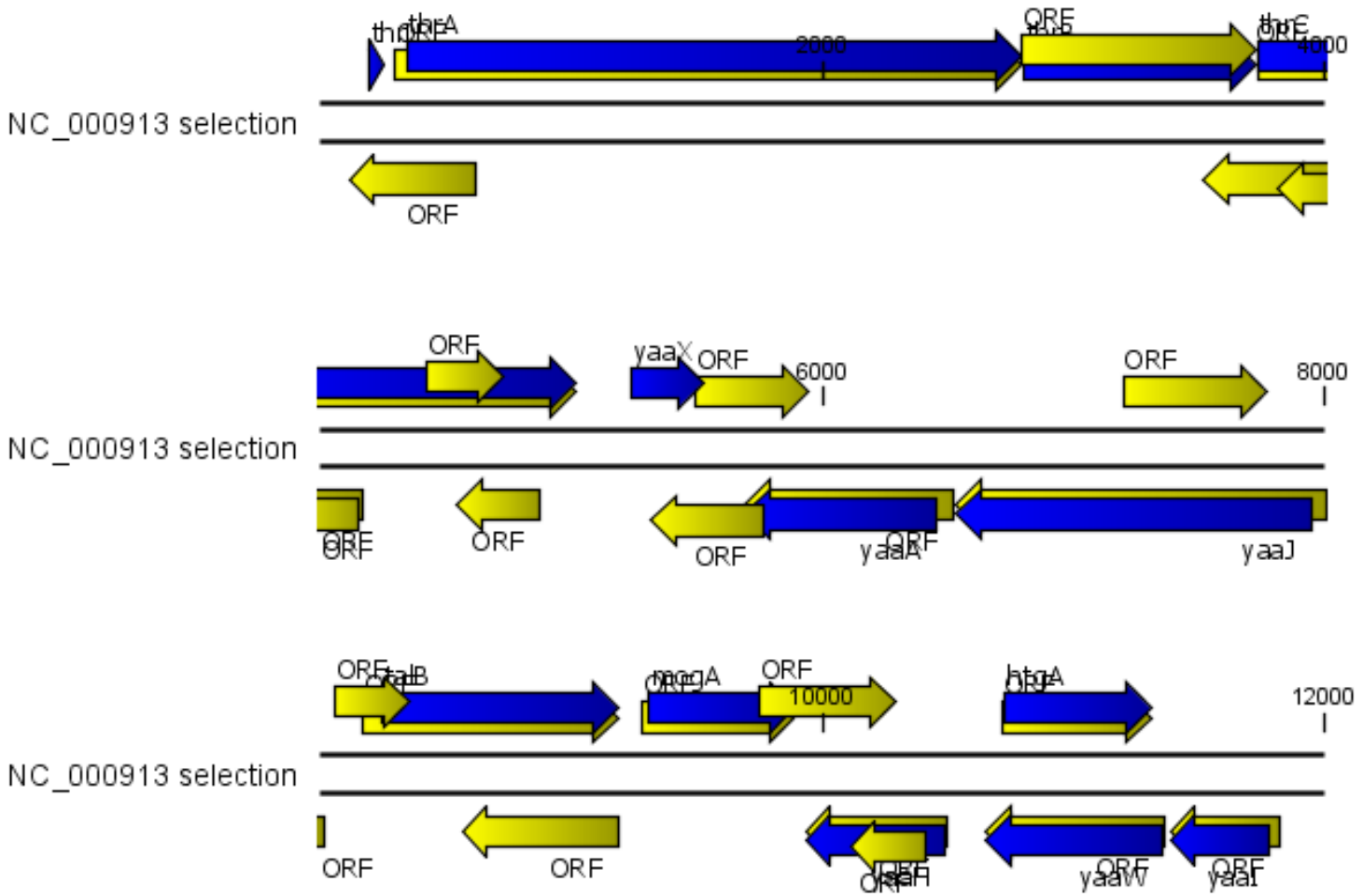
Goodheart et al. 2024, Additional File 12

Genome and Transcriptome Annotation

- Genome annotation consists of five primary steps:
  1. Find and mask repeats
  2. Identifying genes in the genome (ORFs)
  3. Update gene models with alternative splicing events and UTRs
  4. Predict non-coding RNAs
  5. Attaching biological information to genes

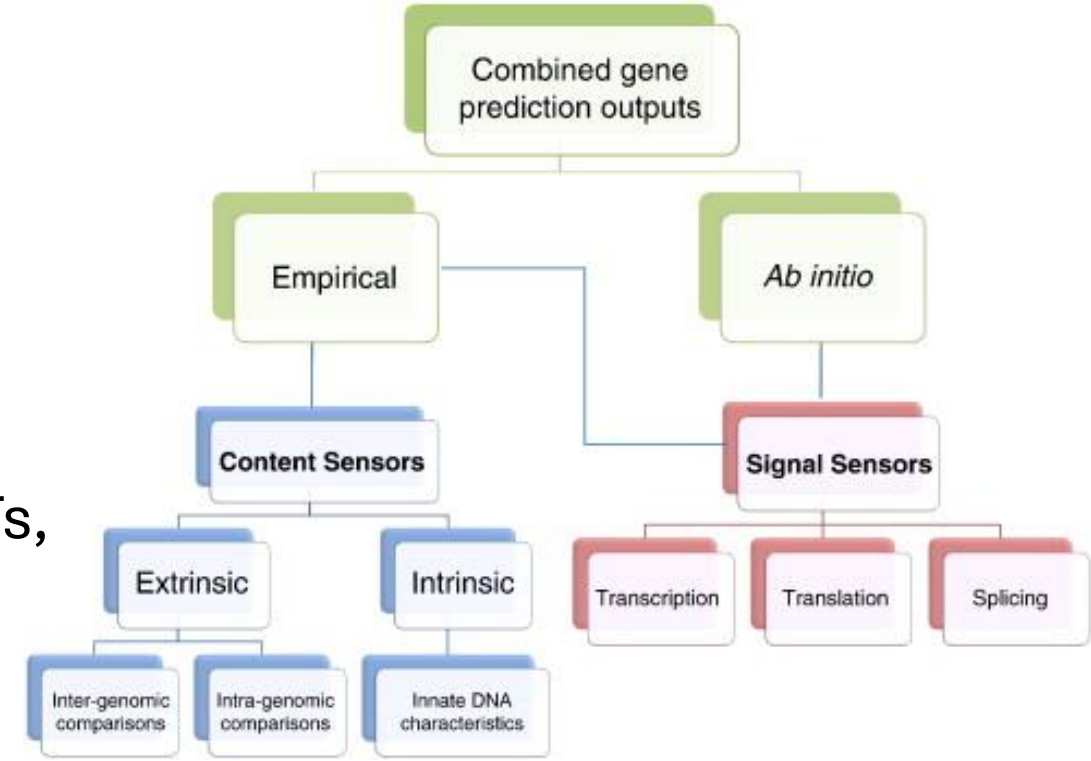
- Transcriptome annotation consists of two primary steps:
  1. Identifying coding regions in the transcriptome
  2. Attaching biological information to genes

Challenge: Lots of Open Reading Frames



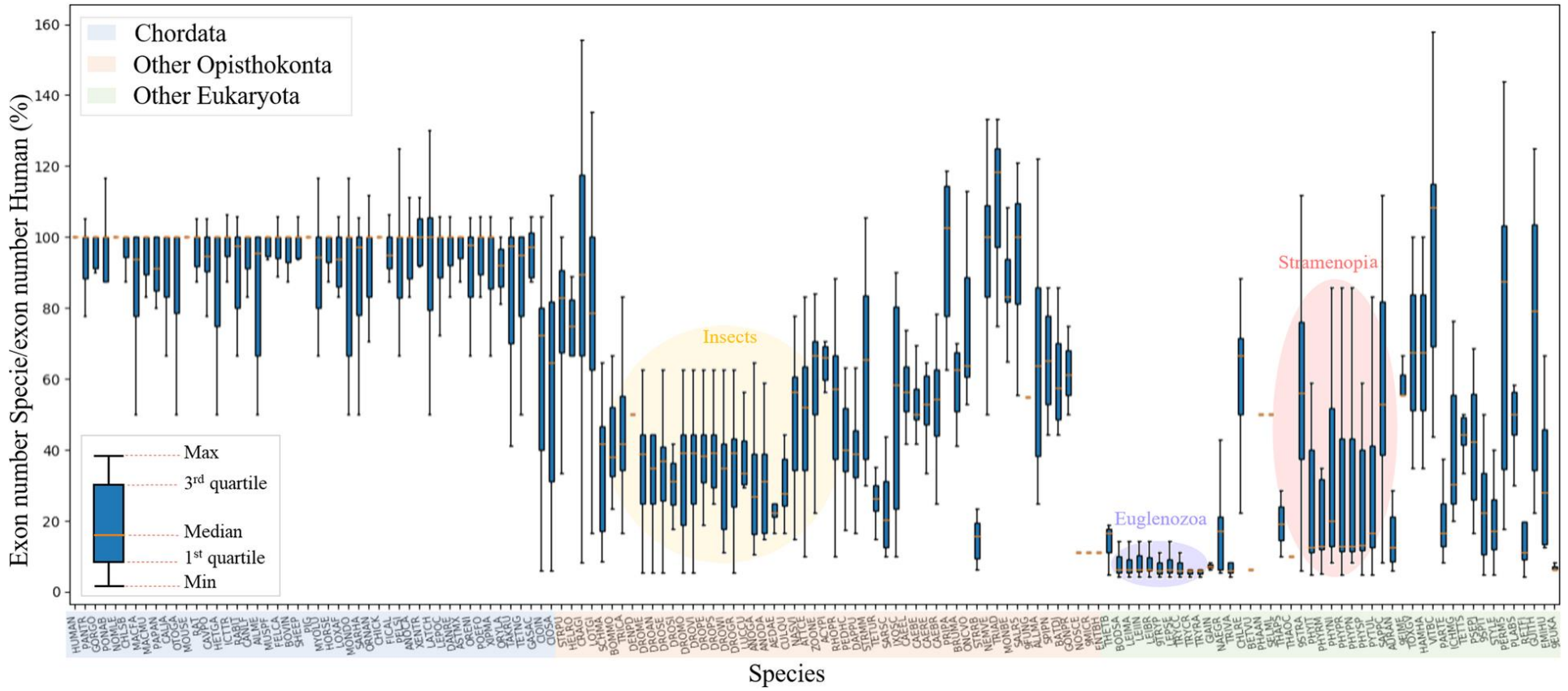
Gene Prediction – Different Approaches

- *Ab initio* gene finders
  - Intrinsic
  - *de novo*, rule-based
  - uses first principles
- Evidence-based approaches
  - Extrinsic
  - based on homology or similarity
  - External data – proteins, RNA-seq, ESTs, etc.
- Comparative approaches
  - Use of closely related genomes



Ab initio methods

- Statistical models need to be trained and optimized for your specific data set. Examples: GeneMark, GenomeScan



Scalzitti et al. 2020

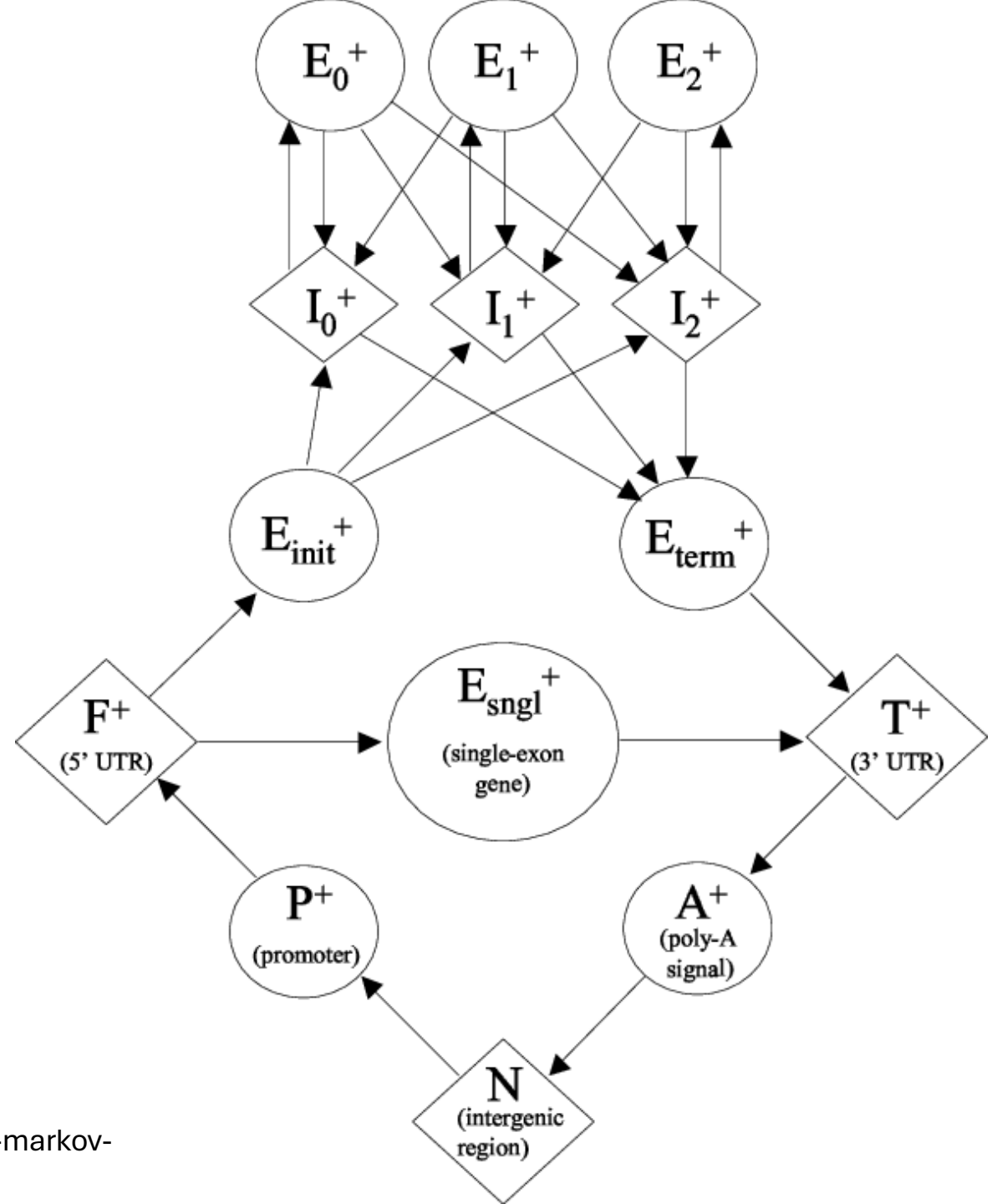
Ab initio methods - probabilistic models

- Often used Hidden Markov Models (HMMs) or Support Vector Machines (SVM)
- Example: Genscan
- HMMs – type of supervised machine learning algorithm using **Bayesian statistics**
  - Makes classifications based on characteristics of **training data**
  - Many types of applications for bioinformatics
    - Gene predictions
    - Sequence alignments
    - ChIP-seq analysis
    - Protein folding

Scalzitti et al. 2020

Genscan HMM

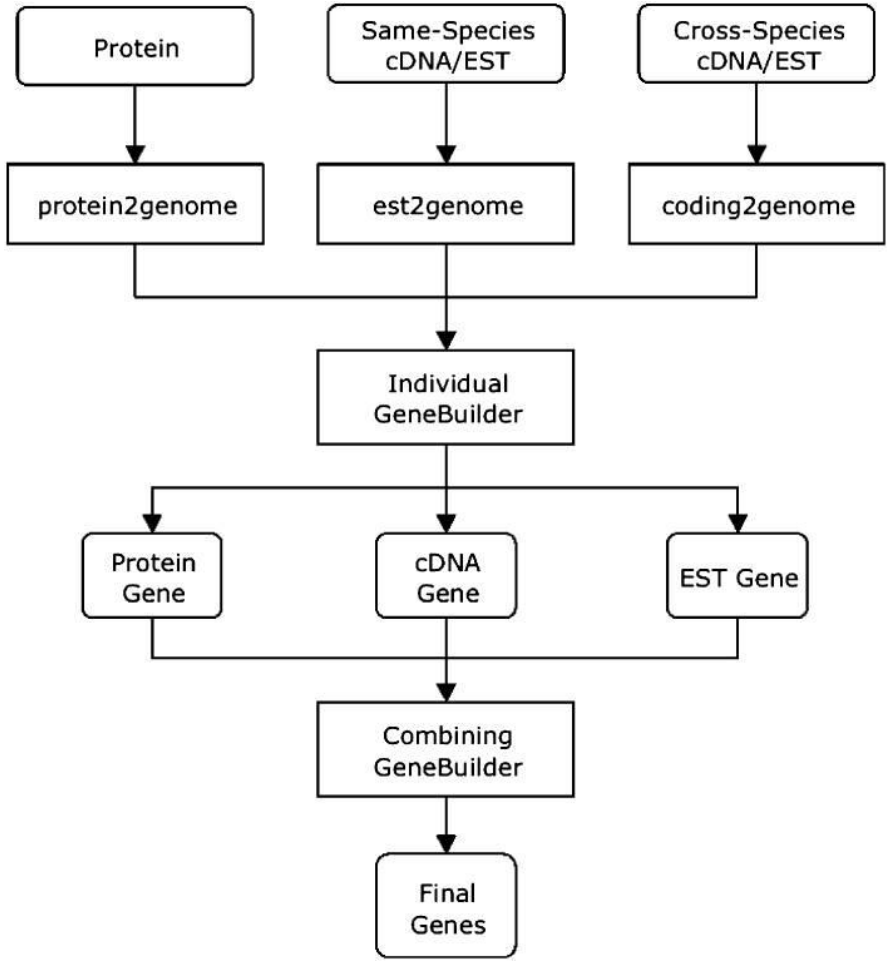
- Each shape represents a functional unit of a gene or genomic region
- Pairs of intron/exon units represent the different ways an intron can interrupt a coding sequence (after 1st base in codon, after 2nd base or after 3rd base)
- Complementary submodel (not shown) detects genes on opposite DNA strand



Great description of HMM models here: https://towardsdatascience.com/hidden-markov-model-hmm-simple-explanation-in-high-level-b8722fa1a0d5

Evidence-based approaches

- Use experimental data such as RNA-seq or proteomes to provide direct evidence for exon boundaries and gene length
- More universally applicable, but introns can sometimes still be present in transcripts
- Examples: Stringtie, Scallop



Liang et al. 2009

Gene prediction commands using StringTie

Run StringTie for each tissue separately

Short reads

```
stringtie -o Bs_genome_V1_18seqs_stringtie_illumina_Bb7.gtf --rf
../../mapping_files/illumina/Bb7.sorted.names.bam -p 12 -l Bs
```

Long reads

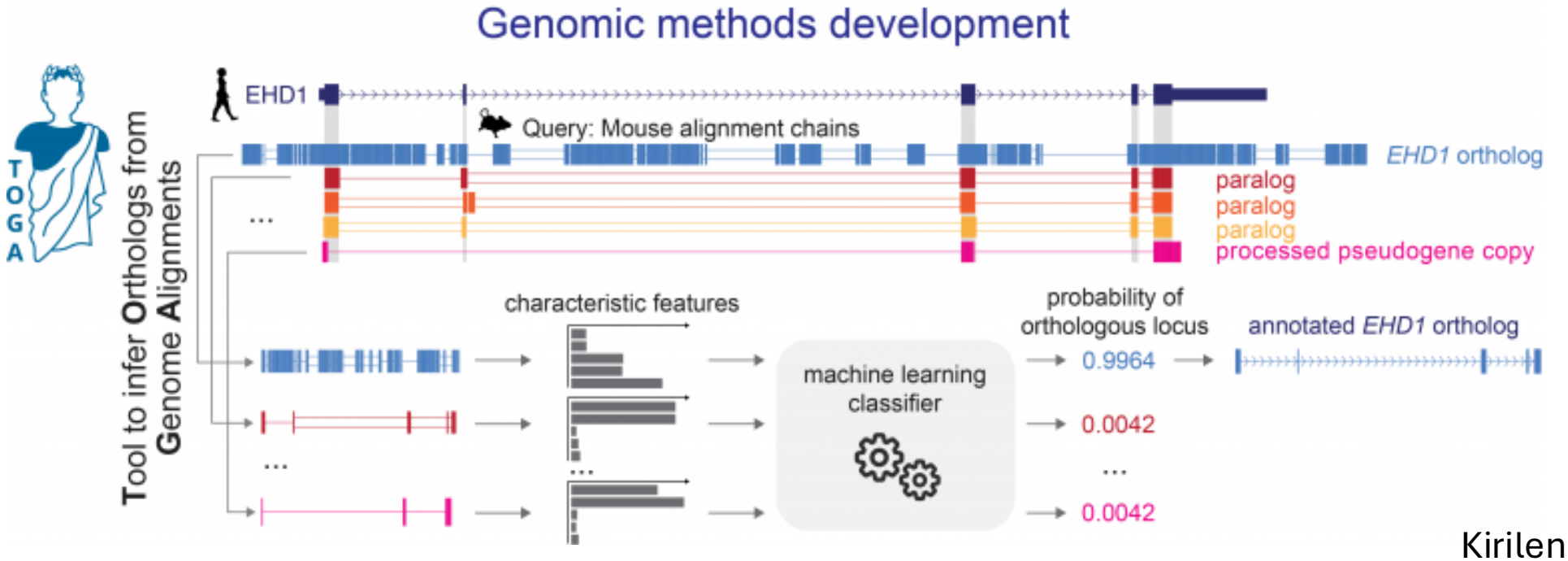
```
stringtie -o Bs_genome_V1_18seqs_stringtie_long_BSDEV.gtf -L
../../mapping_files/isoseq/BSDEV_flnc_isoforms.sorted.names.bam -p 12 -l Bs
```

Merge StringTie outputs

```
stringtie --merge -p 6 -o Bs_genome_V1_18seqs_stringtie_merged.gtf gtf_files.txt
```

Comparative approaches - TOGA

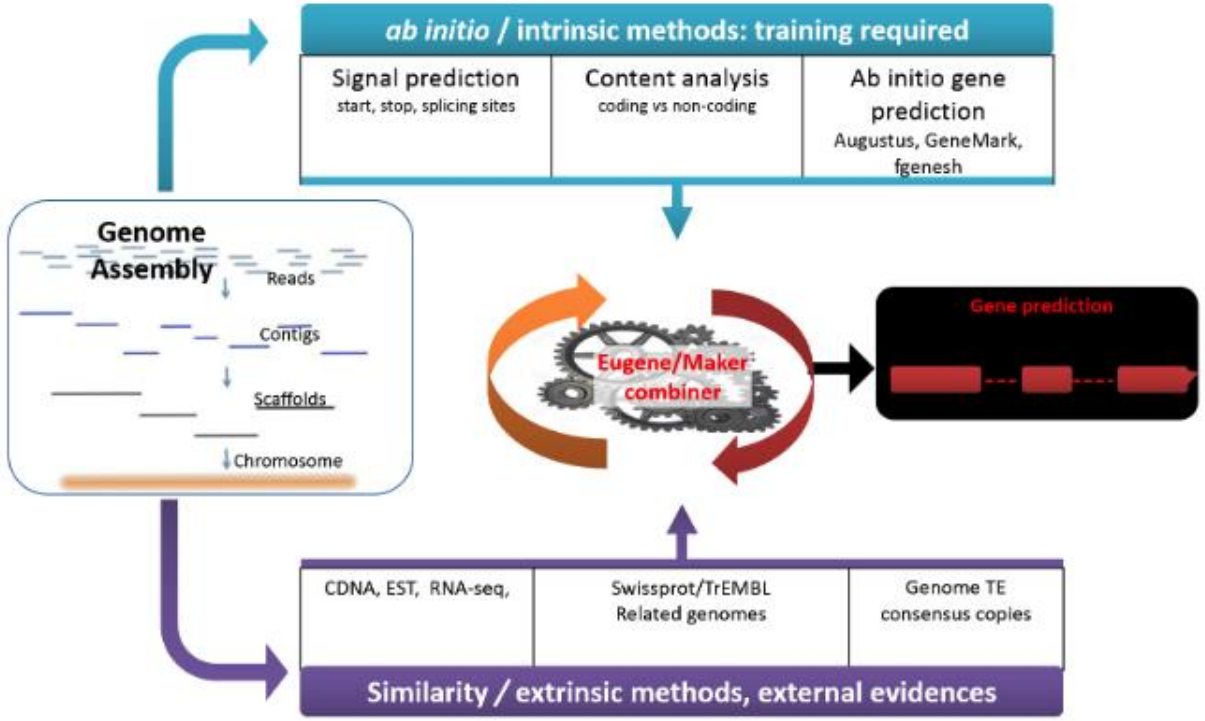
- TOGA (Tool to infer Orthologs from Genome Alignments) is the first method that integrates gene annotation, inferring orthologous genes and classifying genes as intact or lost
- Relies on existing closely related, well-annotated genomes



Kirilenko et al. 2023

Combined approaches

- Combine *ab initio* and evidence-based gene predictions
- Most popular and widely used
- Examples: EvidenceModeler, Jigsaw, GAZE, GLEAN, EuGene, MAKER, etc
- Not all combiners are the same – must decide what type of prediction is preferred



Dominguez Del Angel et al. 2019

Gene prediction commands using BRAKER2

Initial Annotation

```
braker.pl --species=Bsteph --
genome=Berghia_Apr2021_hirise_purged.filtered.fasta_edited.softmasked --prot_seq
mollusca_odb10_with_berghiabuscos.fasta --
bam=Berghia_Apr2021_repeatmodeler_masked_starmappedtwopassAligned.sortedByCoord.out.
bam,Berghia_Apr2021_repeatmodeler_masked_hisat2mapped_sort_shortheaders.bam,berghia_
flnc_isoforms_bestlocus.sorted.bam --etpmode --softmasking -cores 12 --useexisting -
-gff3
```

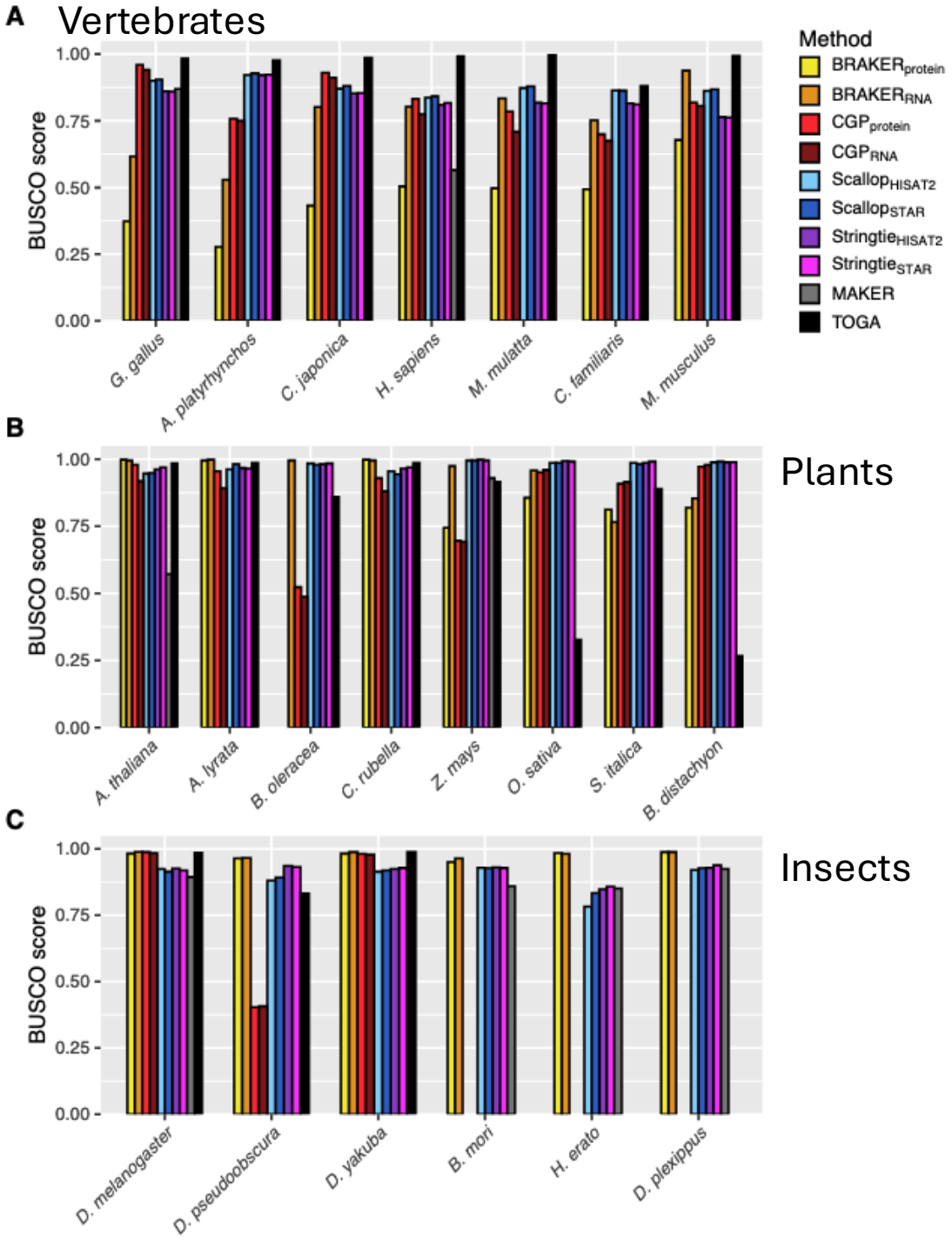
Filter Annotations

```
python selectSupportedSubsets.py --anySupport augustus.anysupport.gtf --fullSupport
augustus.fullsupport.gtf --noSupport augustus.nosupport.gtf augustus.hints.gtf
hintsfile.gff
```

Goodheart et al. 2024, Additional File 12

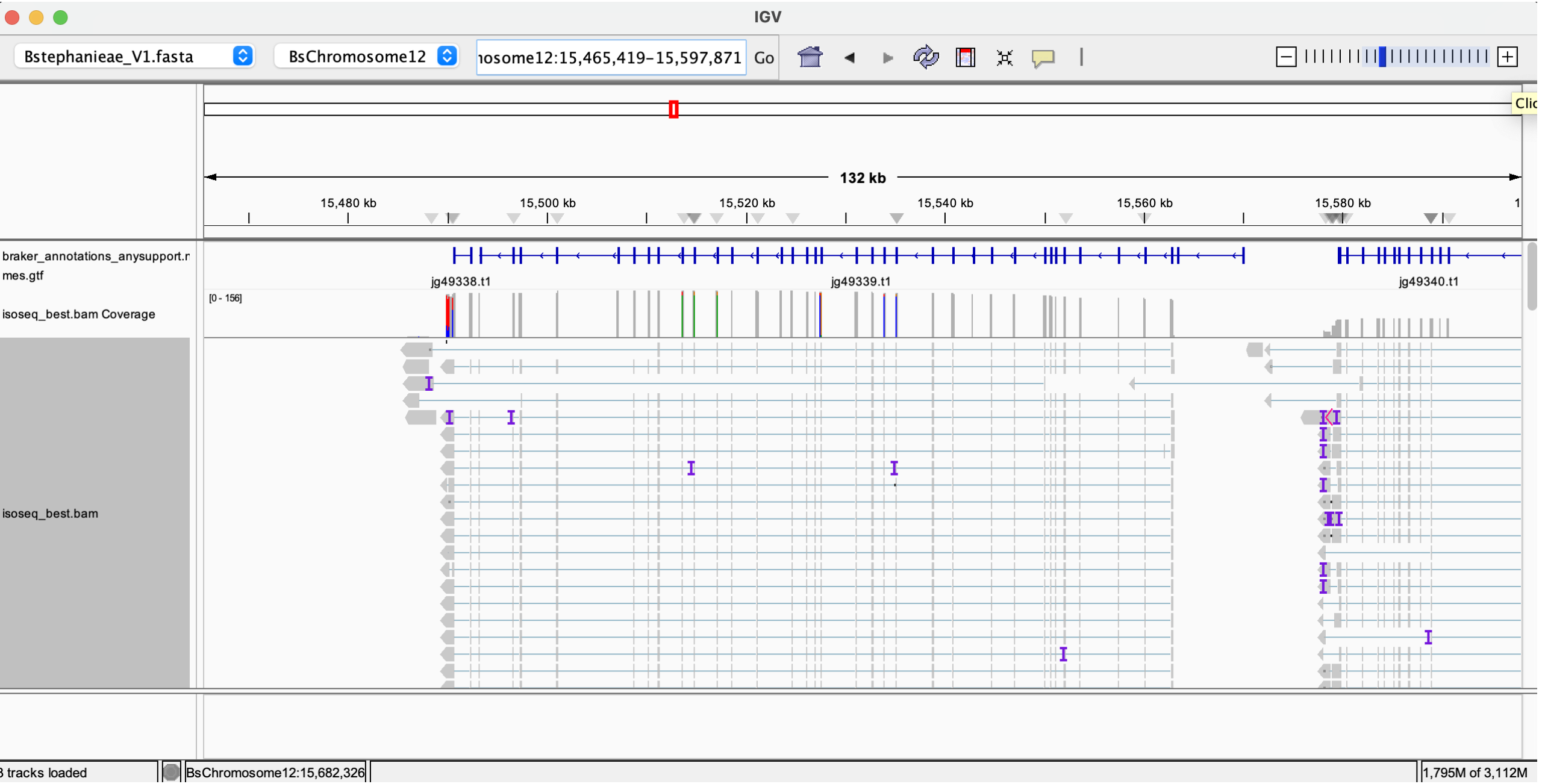
Certain types of annotation approaches are more effective for some types of organisms

- Moral of the story: Be thoughtful about your choices for which programs and pipelines to use



Gene prediction outputs for *Berghia*

	Initial Prediction (BRAKER2)	Filtered Predictions (BRAKER2, any support)	StringTie predictions
No. Gene Models	61,662	26,595	91,490
Avg. Protein Length	327.1 AA	441.1 AA	5,940 AA


IGV
Bstephanieae_V1.fasta
BsChromosome12
osome12:15,465,419-15,597,871
Go
132 kb
15,480 kb
15,500 kb
15,520 kb
15,540 kb
15,560 kb
15,580 kb
braker_annotations_anysupport.r
mes.gtf
jg49338.t1
jg49339.t1
jg49340.t1
[0 - 156]
isoseq_best.bam Coverage
isoseq_best.bam
tracks loaded
BsChromosome12:15,682,326
1,795M of 3,112M

Challenges with gene prediction

- Eukaryotic gene predictions have **high error rates**
- Can be caused by:
  - Errors in sequencing and assembly
  - Mis-modeling of complex genes
  - Diversity of eukaryotic gene structure
  - Errors propagated by public databases

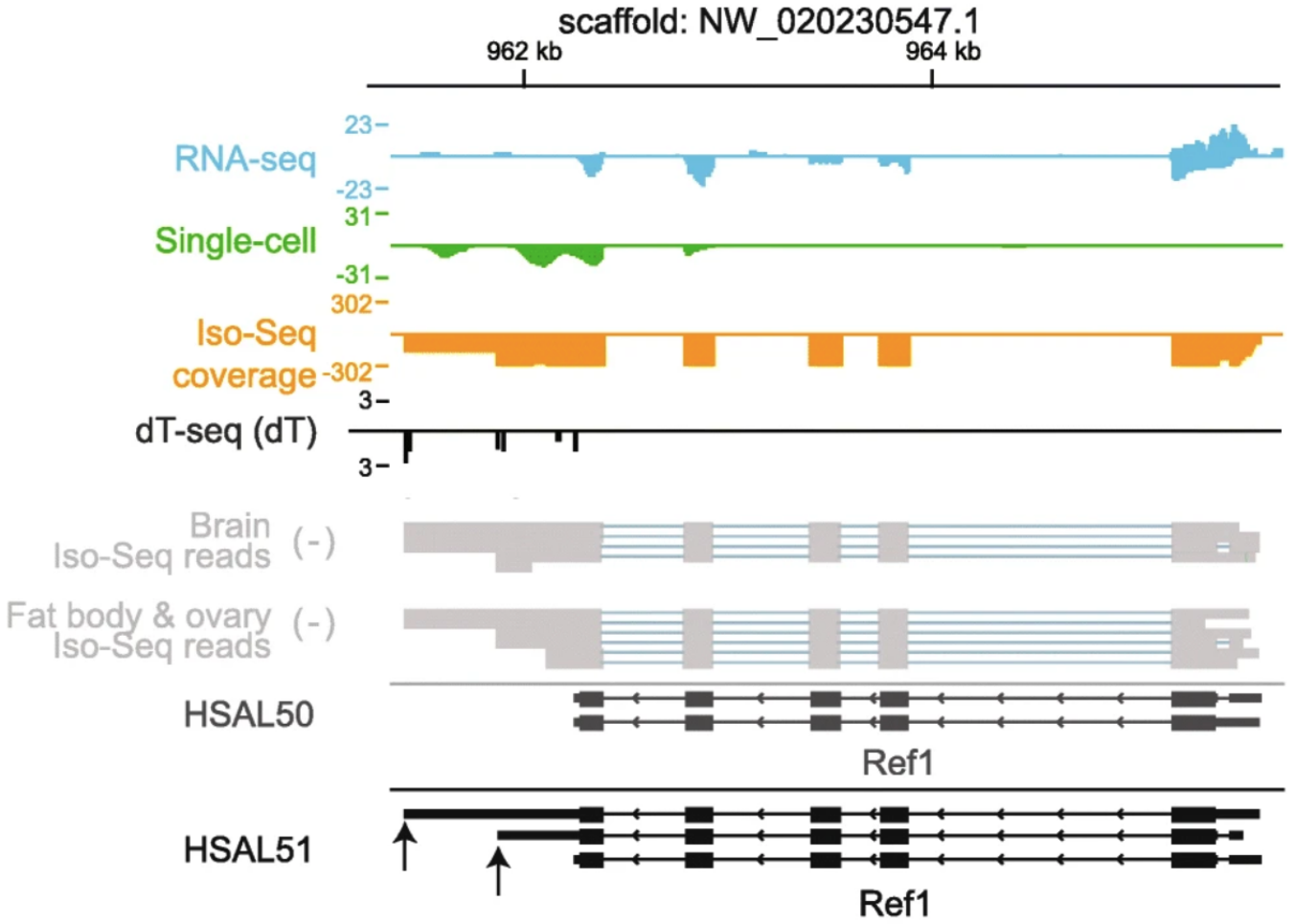
Genome and Transcriptome Annotation

- Genome annotation consists of five primary steps:
  1. Find and mask repeats
  2. Identifying genes in the genome (ORFs)
  3. Update gene models with alternative splicing events and UTRs
  4. Predict non-coding RNAs
  5. Attaching biological information to genes

- Transcriptome annotation consists of two primary steps:
  1. Identifying coding regions in the transcriptome
  2. Attaching biological information to genes

UTRs and Alternative Splicing

- Both PASA and BRAKER3 have mechanisms for including UTR annotations, but BRAKER3 is in beta and not recommended
- UTRs and alternative splicing helpful for analyzing different types of data



Shields et al. 2021

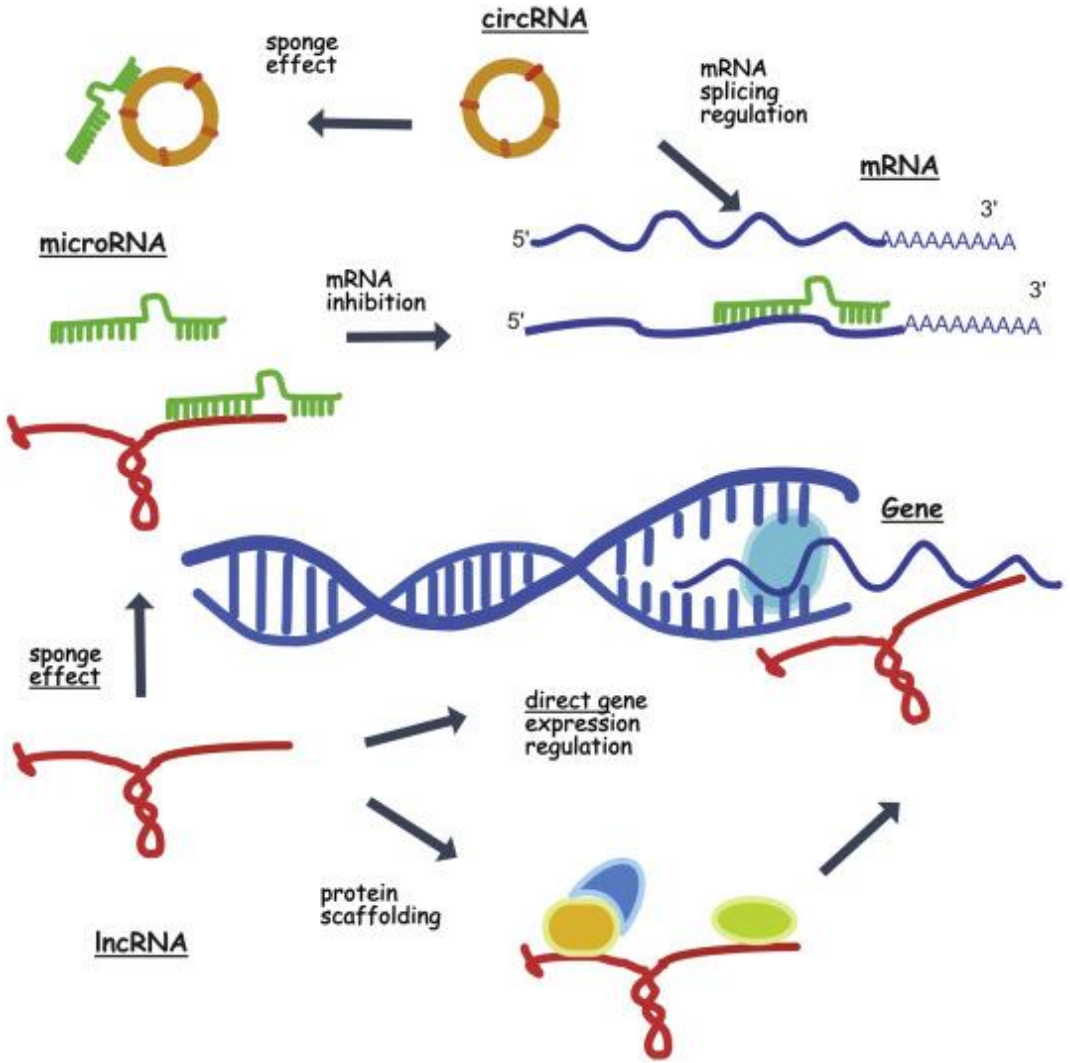
Genome and Transcriptome Annotation

- Genome annotation consists of three primary steps:
  1. Find and mask repeats
  2. Identifying genes in the genome (ORFs)
  3. Update gene models with alternative splicing events and UTRs
  4. Predict non-coding RNAs
  5. Attaching biological information to genes

- Transcriptome annotation consists of two primary steps:
  1. Identifying genes in the genome
  2. Attaching biological information to genes

Non-coding RNAs

- Non-coding RNAs are functional RNA molecules that regulate gene expression and other processes without being translated into proteins
- Majority of many genomes (e.g., human genome at 76-97% ncRNAs)
- Many programs available to use:
  - tRNAscan, snoScan, INFERNAL
  - lncRNAs via BLAST or RNAseq (Cufflinks models or PASA assemblies with no predicted CDS)



Panni et al. 2020

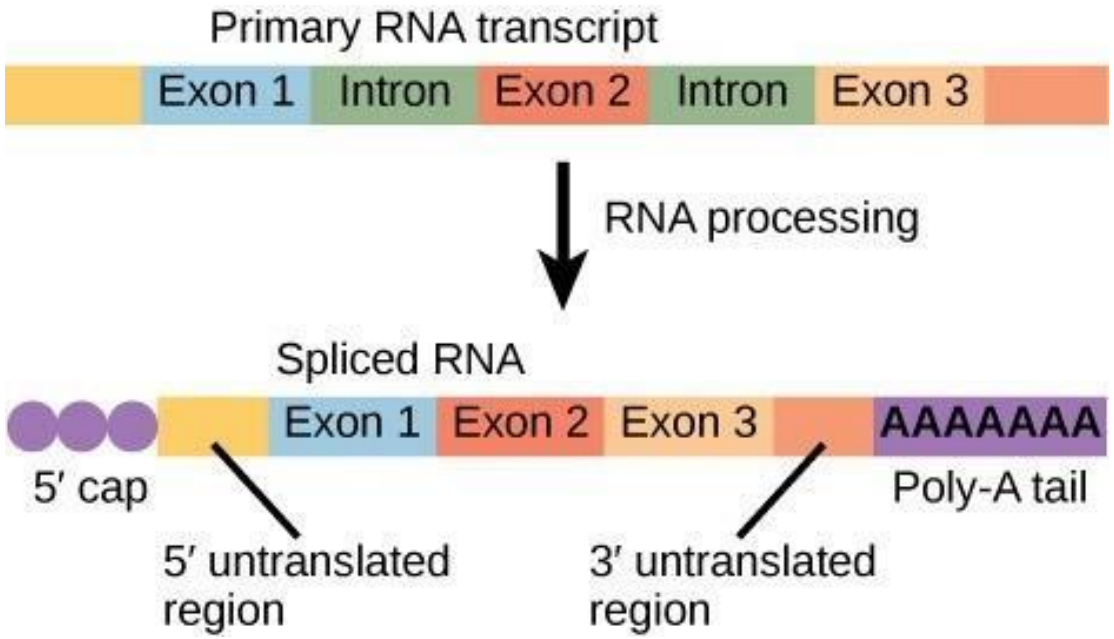
Genome and Transcriptome Annotation

- Genome annotation consists of five primary steps:
  1. Find and mask repeats
  2. Identifying genes in the genome (ORFs)
  3. Update gene models with alternative splicing events and UTRs
  4. Predict non-coding RNAs
  5. Attaching biological information to genes

- Transcriptome annotation consists of two primary steps:
  1. Identifying coding regions in the transcriptome
  2. Attaching biological information to genes

Transcriptome ORF/CDS prediction

- Important for determining coding sequence in transcripts as opposed to UTRs
- TransDecoder has been the primary means for CDS detection in transcriptome data
  - Used for *de novo* ORF detection using the presence of start and stop codons combined with sequence length data
- Other option is ORFanage, which uses a set of reference ORFs to find ORFs in query transcripts



TransDecoder for use in *Berghia*

ORF Detection

```
TransDecoder.LongOrfs -t Berghia_alltissues_onerep_trinity291.fasta
```

```
TransDecoder.Predict -t Berghia_alltissues_onerep_trinity291.fasta
```

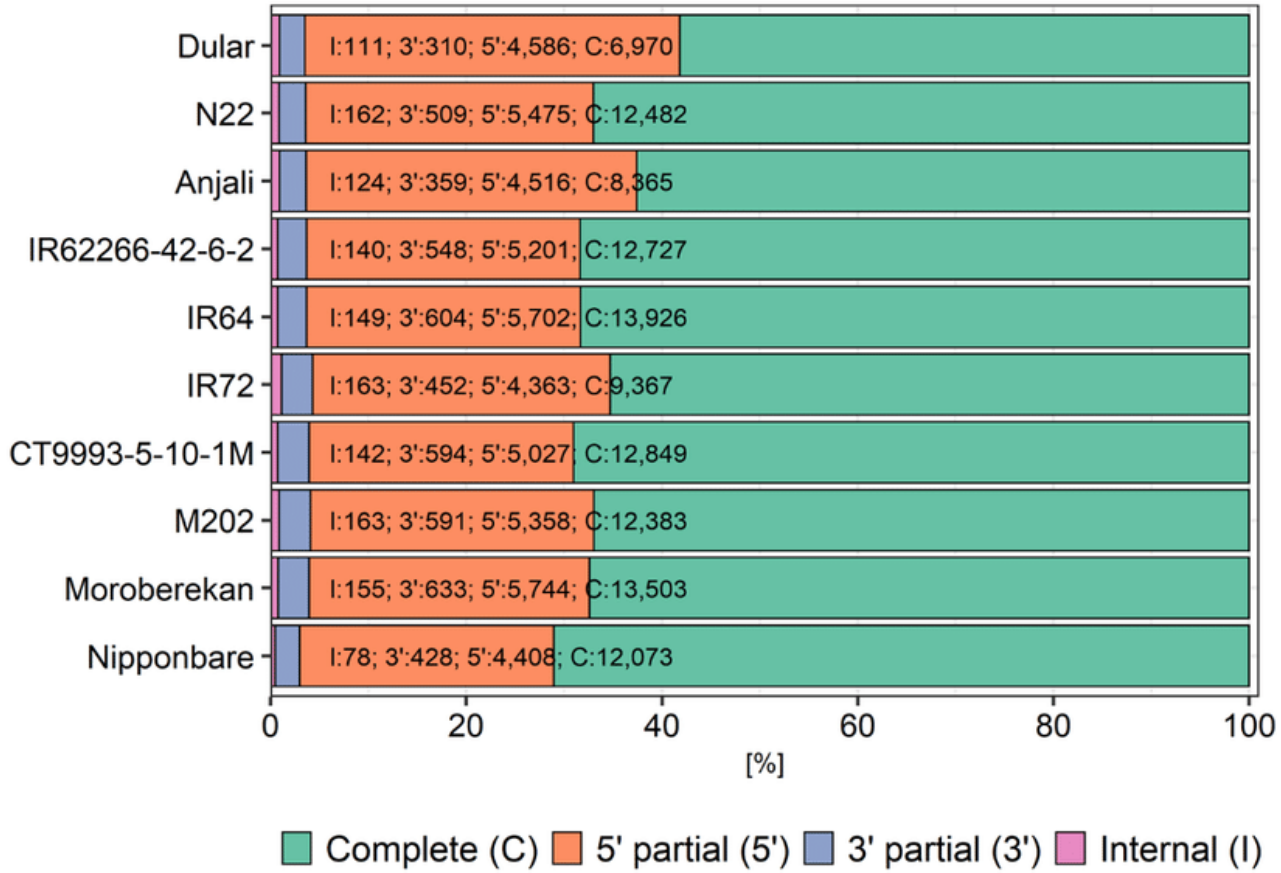
NOTE: By default, TransDecoder.LongOrfs will identify ORFs that are at least 100 amino acids long. You can lower this via the '-m' parameter, but know that the rate of false positive ORF predictions increases drastically with shorter minimum length criteria.

Transcript Clustering

```
cd-hit-est -i Berghia_alltissues_onerep_trinity291.fasta.transdecoder.cds -o
Berghia_alltissues_onerep_trinity291_transdecoder_cdhit95.fasta -c 0.95 -n 11 -M
96000 -T 24
```

Challenges with TransDecoder ORF prediction

- False positive predictions
- **Incomplete transcripts**
- Upstream amino acid codons
- ORFs on the opposite strand



Schaarschmidt et al. 2020

Summary

- Eukaryotic genomes and transcriptomes are complex and diverse
- Predicting genes in eukaryotic genomes is important, but still incredibly challenging
- Use as much data as possible for building gene models – including RNAseq data in constructing predictions is incredibly valuable
- Use gene predictions with care

Additional References

- Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O et al. Ten steps to get started in Genome Assembly and Annotation. F1000Res. 2018;7(ELIXIR):148. DOI: 10.12688/f1000research.13598.1
- Freeman AH, Sackton TB. Building better genome annotations across the tree of life. bioRxiv. 2024.04.12.589245; https://doi.org/10.1101/2024.04.12.589245
- Salzberg, S.L. Next-generation genome annotation: we still struggle to get it right. *Genome Biol* **20**, 92 (2019). https://doi.org/10.1186/s13059-019-1715-2
- Scalzitti, N., Jeannin-Girardon, A., Collet, P. *et al*. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics* **21**, 293 (2020). https://doi.org/10.1186/s12864-020-6707-9
- Sun, YM., Chen, YQ. Principles and innovative technologies for decrypting noncoding RNAs: from discovery and functional prediction to clinical application. *J Hematol Oncol* **13**, 109 (2020). https://doi.org/10.1186/s13045-020-00945-8
- Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. 2009 Sep;10(6):402-15. doi: 10.2174/138920209789177575. PMID: 20190955; PMCID: PMC2766791.