GENOME ANNOTATION

MIRIAN T. N. TSUCHIYA
DATA SCIENCE POSTDOCTORAL FELLOW
DATA SCIENCE LAB - OCIO

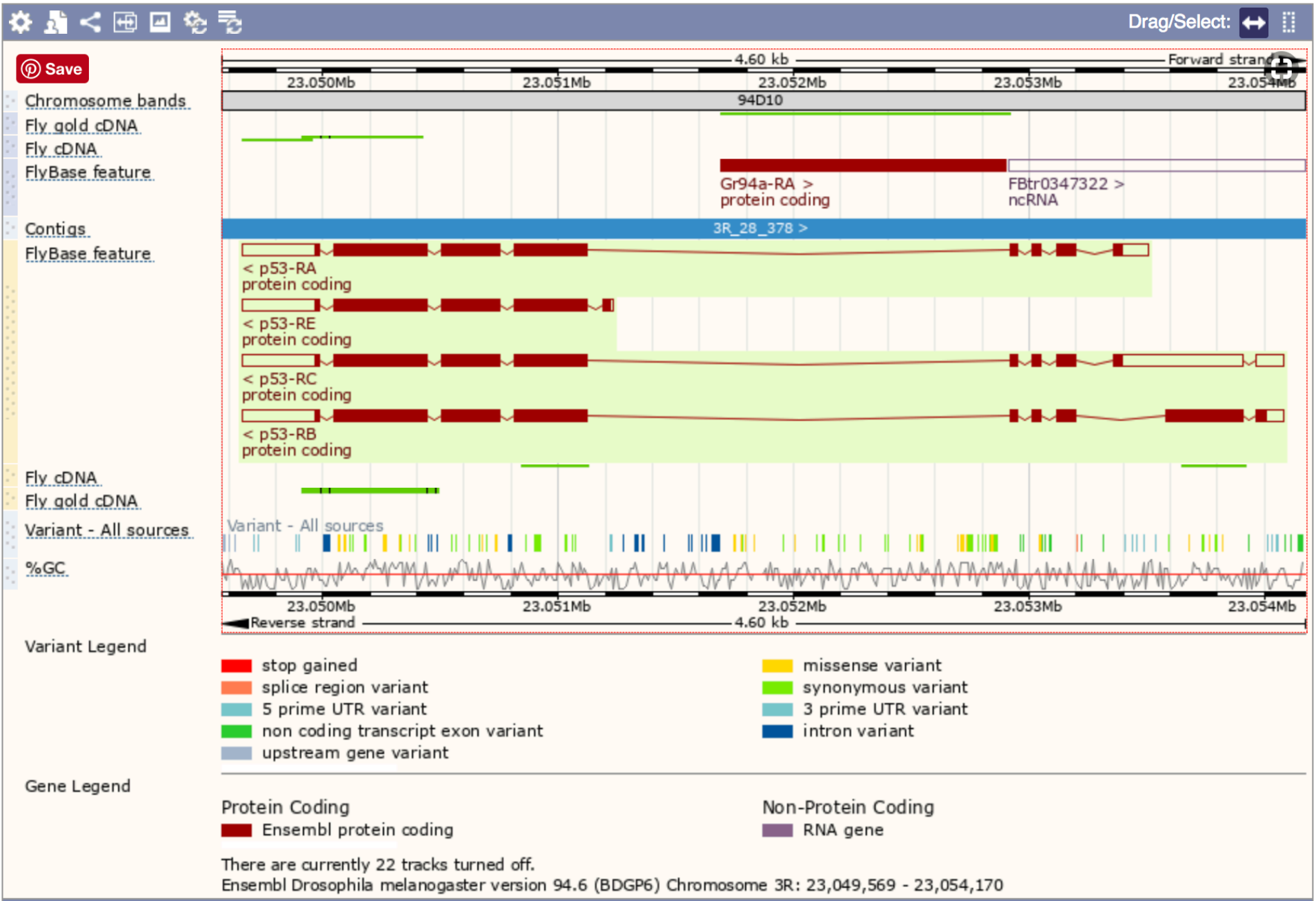
WHAT IS GENOME ANNOTATION?

Genome annotation is the process of identifying different elements in a genome assembly.

Chromosome 3R: 23,049,569-23,054,170

ATGTTTGCAAAATATGACAACTCTTTAATAAATTCACTTTTGGCTTTTTTGTACAGAGTATGTTTGTTGTGGTAGACAAAAGTTACGTAGTCACA
TAGAACTGCAGTTAACAACAGTATACTATCAAAAATATAAAAATGTTTTCTTTTCGGCCATCGAACATGCCAAAAGGGGAATATCACCGTTAAA
CGCCTTACGAACAAGGACTAATACAAACTTAAGATACTAATAGCTATGTAGTATACCCCACAGATGCAACACATTTCAAAGGTATCTGATACC
TGGGAGATTGTCGACCAGATCAGAAGTCATGGCAGCTCGTAGGCACGTTCTGAAAGGAAAGTTCGGTAAGTAAAGCTTAATCTCTTTGTCCT
CCGATTCACTGCTTACTCTTAAGGCTCAGCAATTTGTTGGCATGGCGACGTAGATTCTCTTGGTTGGGATTGCGCAGGACTTCAGCCGCCGC
CTCCTTAATCATGCCCTCGATGCTCTGCAGCAGCCATTCCTTATTGGGGCACGTAATAGCCAGACGGTAATCGCCATCCGGTGTCCGCGACA
CGTTCCACTCTGCGGCGGAGTCGTCGCAGTCTCGGCTATCATTGCTCTCCGTGTCCTCCGTCTTTATAGCAATGCACCGACGCACCTTGGAC
GGCTCATCTTCTTCGGCGGCTTCCGGCACGGACTTGCGCTTCTTGCTATTGAGCTGGCGTTCGTCTTGGATGCGATCCCGCTTGGGGCACGT
ACATATTTTAACATGTATAACATGCTGTCCCACGATATCGCCGCTGCAAGAAAGAAAACACGTTAAGCTTTGTTCTAGCCATCTAGAGTTTTGC
TGTCACCTTACCATGCTTTCTCCAGGCAGAAGACTAAGGAAGTTTCTTTTCGCCCGATACACGAGTTTTGGCAGACGAACTTGAAGGCCAGG
GTCTGGCGCGTGAGCCCACTGCGGGTTACAGACCGGCTCATGTTCAGGGGGACTACAACGGAAAAACGCTCGGAAATTCCCTTGCCCTGAG
CATTTCCACAATATACACTGTTGGGATTCTCGCTGCGCAGCAAGCTCTCGCGCATTTTTGCGTTATTGGCCGTCACTAAAGGTTTGTACAAAT
GATAGTTAAATAATCCTGTTCGATCTGTATTGTTATCTTCACTTACAAGGCTCAACGCTAAGGTGATTTTGACAGCGGACCACGGGAGCACTC
ACATCATTGGAGAAGCAAAGGAACACACGCAAATTAAGTGGTTGGATGGGCATTTTAGACTTGAACTGAACGTCCACGTTGAAGGCCTTGTT
CATCCGGATGTAGAGCTTGTTCAGCGGAATCGAGTACATCCAAAGAGACTTGGGCGGCTCATCCAGAACCATGCTGAAGCAATAACCACCGA
TGTTGTGATTCTCTAGCTTGGGCAGCGTGTTCGCCTGGATCTGAATGTCCTGCAGCATCATTTCGCGCAGCACGGATTGCTGTAAGCAAAAA
CAAGATATTATTTGAATGTCAACTTAACTCGCAATCATATAAGGGTTCACACTTTGCGCTCATAGAGCTGATAGAGCTTCATGTTCCAGCTTGC
TCGAGTTACAAATGGACTGGCGATTTGTTTTGTATTCTATTCTATTTTTCACAAGTGCTTGAAATTCTCGCTAAACGATGACGACGGGAAGGAG
TGCACCGTGCTAATCGGCTAGATAAGAGTGTTTTGTTCTGTCCTCCAACTGAAAGTGAACTTCGAAGCGACTTTGACGTCATGTCGCTCTTAT
GAAATTGCAGGCAGCGGCTGAGTCACGCAAGGAATGCCGCTTCCCTCACGCTTCTTATGCTCTTTTGTTTGTGGCTAGTAGTAGCACAACGC
ATGCTCAGTTGGTTTCTGGACGGCCGGTGAAATACATGCGTTTCGCGATGTGTCCCGATAGTGTCCGGCTACTTGACAGGTCCTCTAGATGG
CATCAGTTGGGTAGTTGGAAAACAATTAAGCAGATATTAATGCTGTGGAAAGTGCGCTCACAGTTTGTATTAATTGTTCGAGAGAAATGGACT
TCACCAGCGACTACGCGCATCGGCGTATGGTGAAATTTCTGACGATCATACTGATAGGCTTTATGACCGTCTTCGGACTCCTGGCCAATCGA
TATCGGGCGGGCCGTCGTGAAAGATTCCGCTTCTCAAAGGCAAATCTGGCCTTTGCTTCGCTGTGGGCAATTGCATTCAGCTTGGTTTACGG
GCGGCAAATCTACAAGGAGTACCAGGAGGGTCAGATCAACCTGAAGGACGCCACCACTCTGTACAGCTATATGAACATTACGGTGGCTGTTA
TTAACTATGTGTCGCAAATGATAATCAGTGACCATGTGGCCAAGGTGTTGAGCAAAGTGCCCTTCTTTGATACCCTAAAAGAATTCCGTCTGG
ACAGCAGGTCGCTGTACATATCCATCGTTTTGGCTCTGGTCAAGACCGTGGCTTTTCCTCTAACAATTGAAGTGGCTTTCATACTGCAACAGA
GGCGGCAGCATCCCGAGATGAGCTTGATCTGGACCTTGTACCGGCTGTTTCCCTTAATTATTTCGAATTTTCTCAATAACTGCTACTTTGGCG
CAATGGTGGTGGTGAAGGAGATTCTGTACGCTCTGAACAGACGGCTGGAAGCGCAGCTGCAGGAGGTGAATCTGCTGCAGAGGAAGGACCA
GCTAAAGTTGTACACTAAATACTACCGCATGCAGCGATTTTGCGCCTTGGCGGATGAACTCGACCAGCTGGCGTATCGCTATAGGTTGATATA
TGTGCATTCGGGAAAGTATCTGACCCCAATGTCCTTGTCCATGATTCTGTCGCTCATATGCCACCTGCTCGGAATAACGGTGGGTTTCTACAG
TCTGTACTATGCCATAGCGGACACCTTAATCATGGGCAAGCCGTACGATGGTCTTGGATCGCTGATCAATCTGGTTTTCCTCTCCATCTCGCT
GGCGGAGATCACATTGCTCACGCATTTGTGCAACCACCTATTGGTGGCCACCCGAAGATCGGCAGTCATTCTCCAGGAGATGAATCTCCAGC
ATGCGGACAGCCGCTACCGTCAGGCAGTCCACGGCTTTACTCTGCTGGTCACGGTGACCAAGTACCAAATTAAACCCTTGGGCTTGTACGAG
CTGGACATGCGACTGATCAGCAATGTCTTCTCGGCGGTGGCCAGCTTCCTGCTGATCCTCGTGCAGGCCGATCTGTCCCAGCGCTTCAAGAT
GCAATAGCTAATCGATGTTACCCACCTGGCTGAACTGCATCAGATTCCCGGACTGCGGGGAATAATATTAAATAGTTAGTAAGCTATAGCTTT

Chromosome 3R: 23,049,569-23,054,170



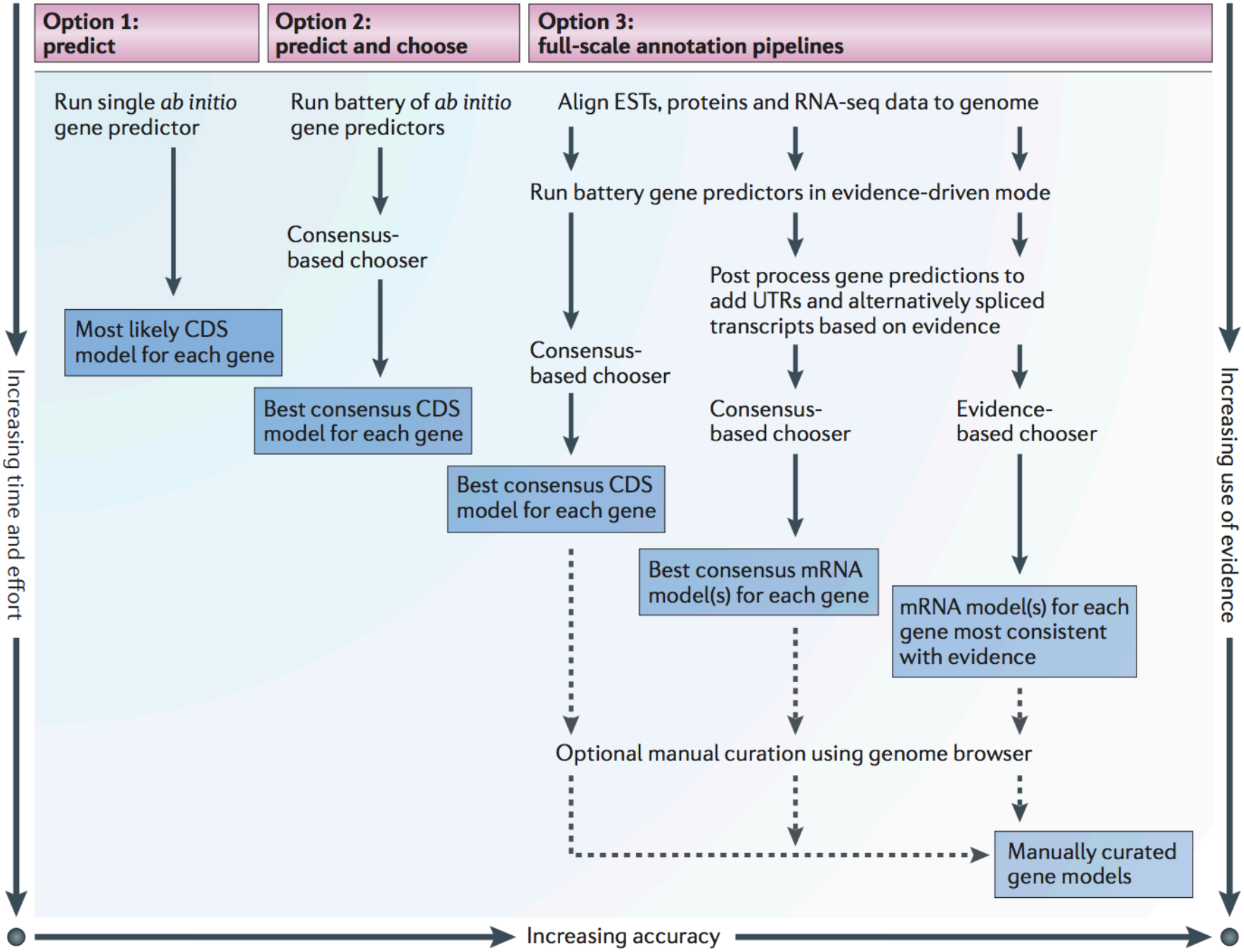
WHAT IS GENOME ANNOTATION?

Structural	• Repetitive elements • Genes (introns and exons)
Functional	• What does each gene do?

GENE PREDICTION

ab initio	evidence-driven
• use only the query sequence	• use external evidence

Combining both approaches is the best option


Option 1: predict
Option 2: predict and choose
Option 3: full-scale annotation pipelines
Run single ab initio gene predictor
Most likely CDS model for each gene
Run battery of ab initio gene predictors
Consensus-based chooser
Best consensus CDS model for each gene
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

SOFTWARE/WRAPPERS

Augustus	SNAP	Glimmer HMM
Genemark -ES	FGenesh	Gnomon

OTHER ALTERNATIVES

Pros

- Standardized
- Free

Limitations

- Quality requirements
- Taxonomic priorities

FUNCTIONAL ANNOTATION

Process of describing the biological identity of a gene, including molecular function, biological role, location, and expression domains.

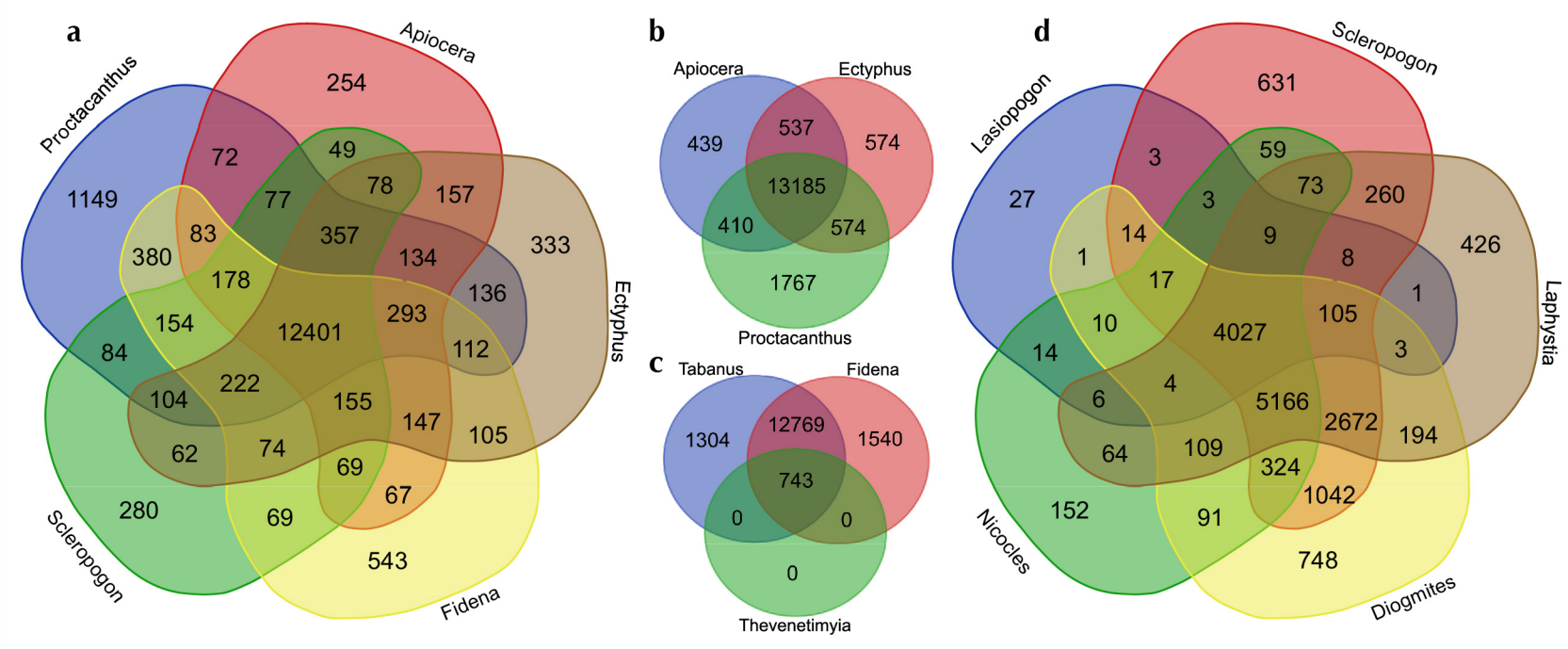
FUNCTIONAL ANNOTATION

- Gene Ontology:
  - Molecular Function (MF)
  - Cellular Component (CC)
  - Biological Process (BP)

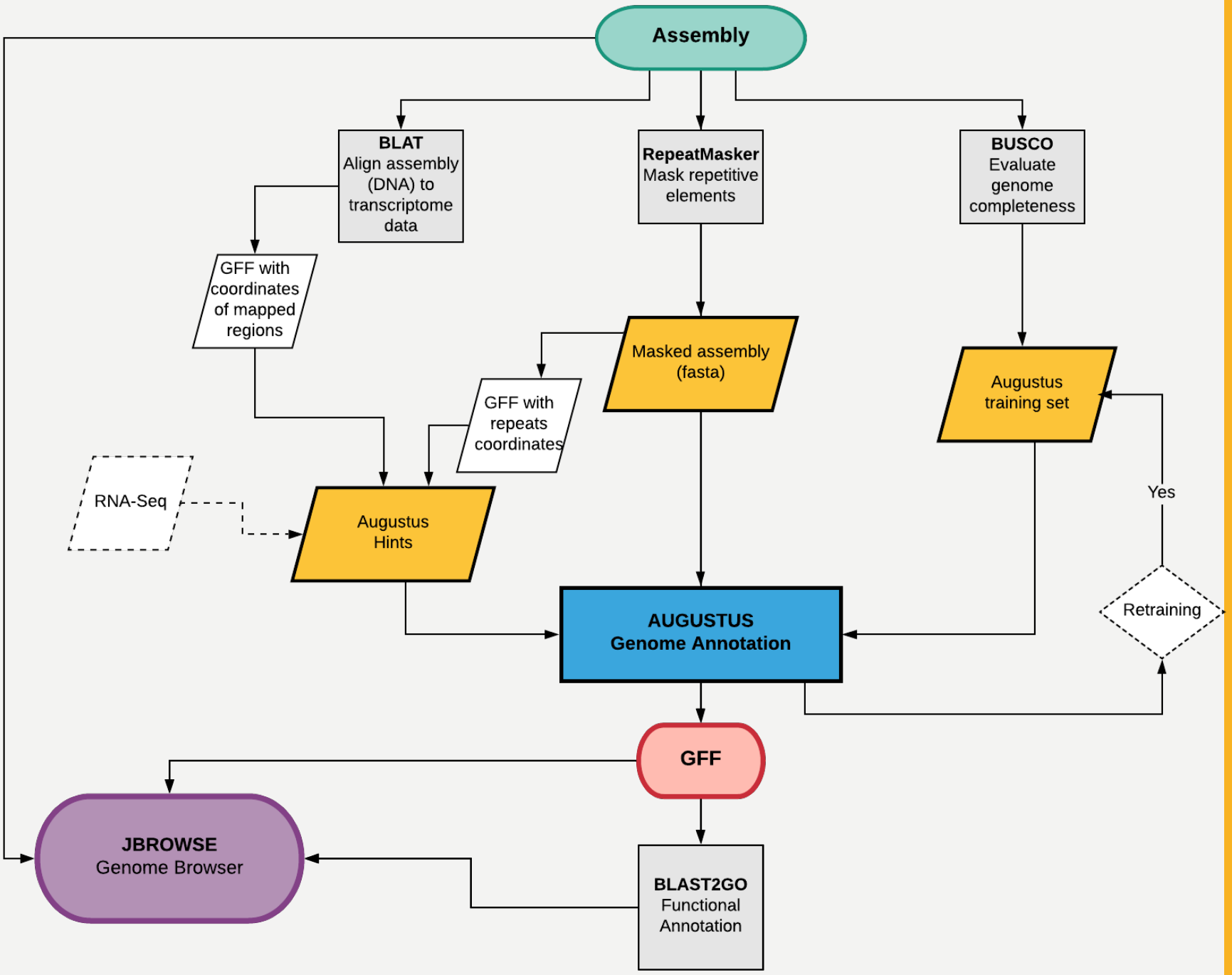
Example: cytochrome c
-molecular function: oxidoreductase activity
-biological process: oxidative phosphorylation
-cellular component: mitochondrial matrix.

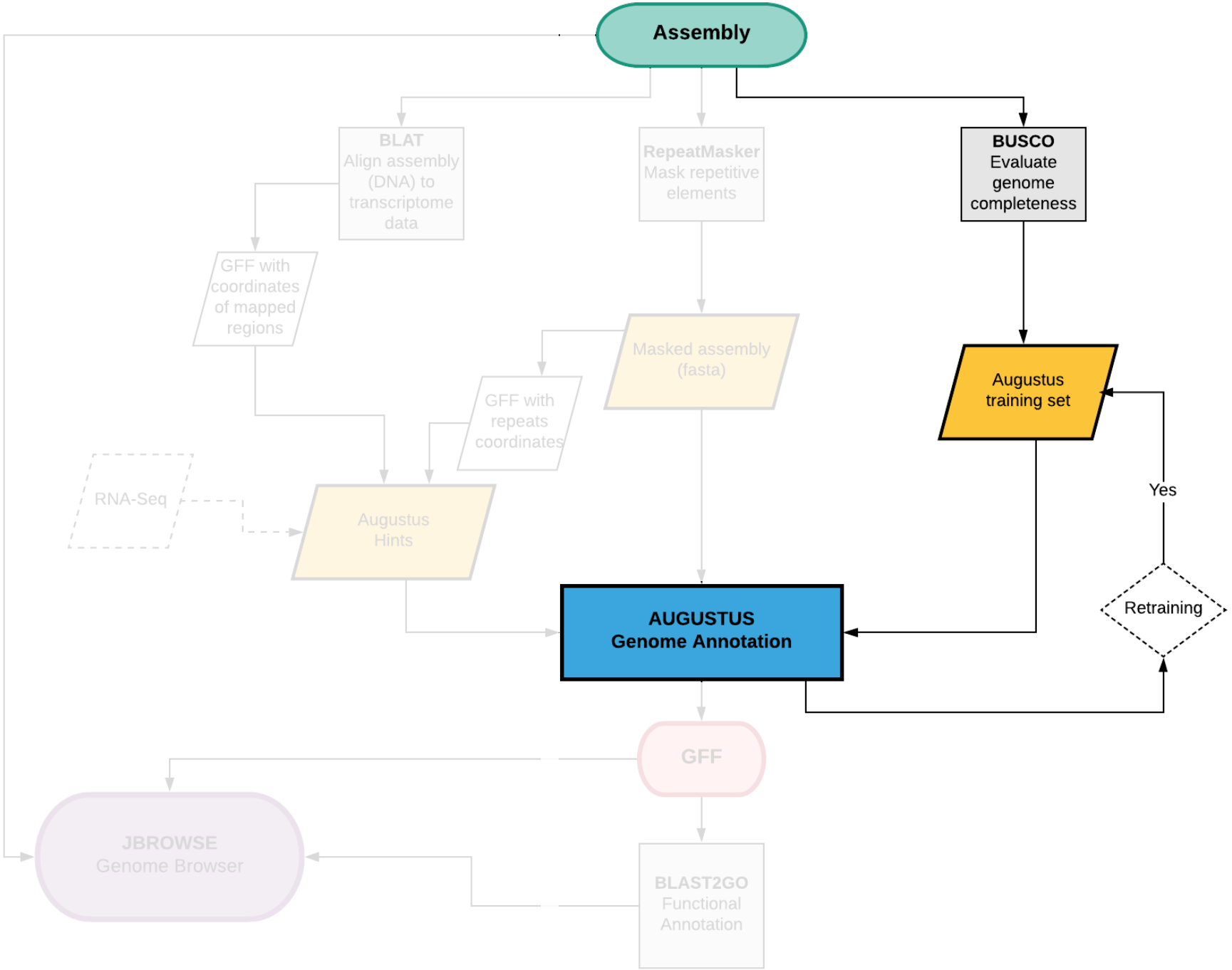
FUNCTIONAL ANNOTATION

- Software:
  - Blast2GO (available on Hydra; paid license)
  - Other (free) options: DAVID, GO FEAT



OUR PIPELINE


Assembly
BLAT
Align assembly (DNA) to transcriptome data
RepeatMasker
Mask repetitive elements
BUSCO
Evaluate genome completeness
GFF with coordinates of mapped regions
Masked assembly (fasta)
Augustus training set
GFF with repeats coordinates
RNA-Seq
Augustus Hints
Yes
Retraining
AUGUSTUS
Genome Annotation
GFF
JBROWSE
Genome Browser
BLAST2GO
Functional Annotation


Assembly
BLAT
Align assembly (DNA) to transcriptome data
RepeatMasker
Mask repetitive elements
BUSCO
Evaluate genome completeness
GFF with coordinates of mapped regions
Masked assembly (fasta)
Augustus training set
GFF with repeats coordinates
RNA-Seq
Augustus Hints
Yes
Retraining
AUGUSTUS
Genome Annotation
GFF
JBROWSE
Genome Browser
BLAST2GO
Functional Annotation

BUSCO

BENCHMARKING UNIVERSAL SINGLE-COPY ORTHOLOGS

BUSCO

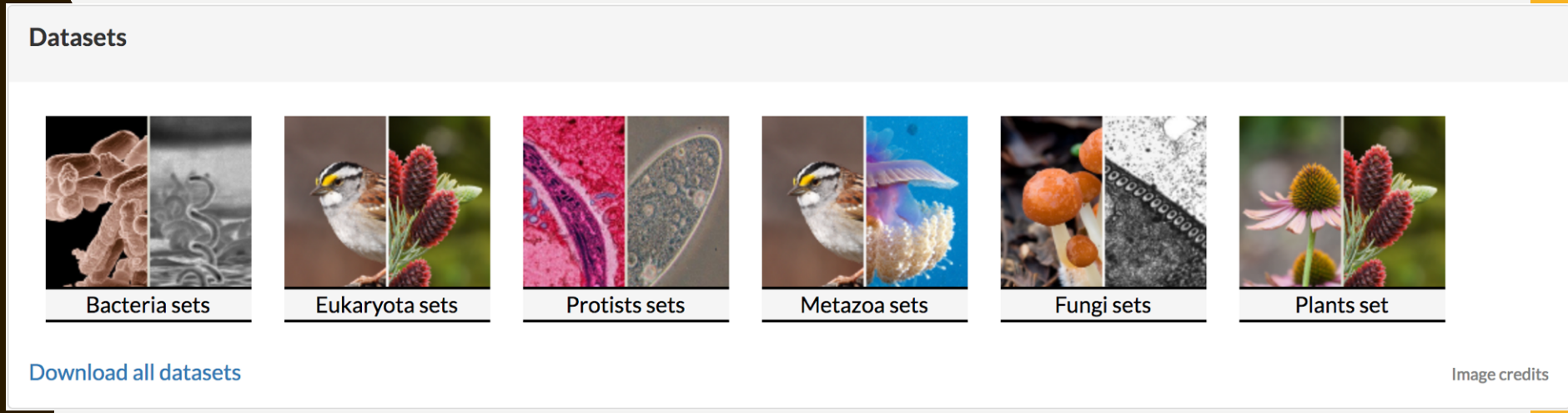
- Benchmarking Universal Single-Copy Orthologs

What is a ortholog?

Orthologs are genes in different species that evolved from a common ancestral gene by speciation.

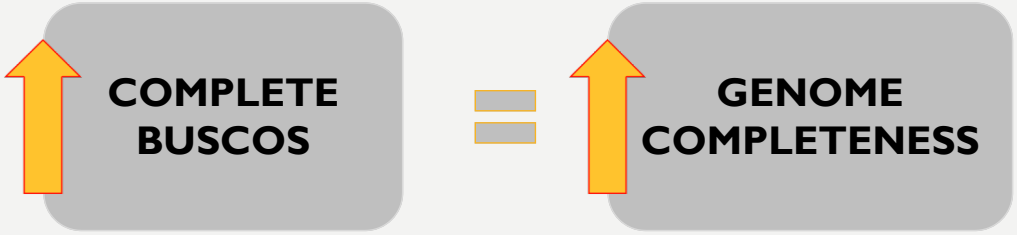
BUSCO:
HOW COMPLETE IS THE ASSEMBLY?

- Database: taxon-specific single copy orthologs



BUSCO:
HOW COMPLETE IS THE ASSEMBLY?

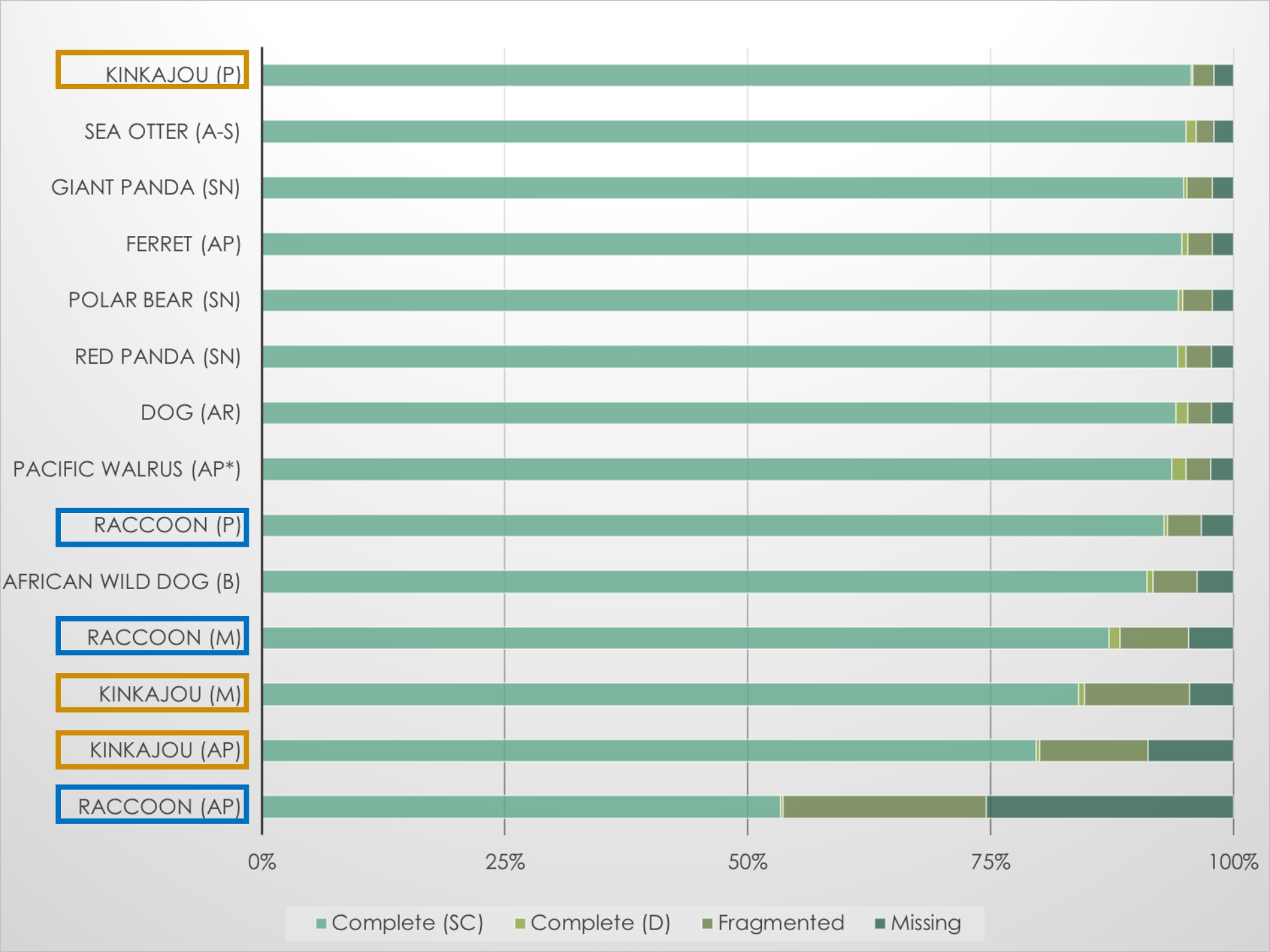
- Assessment:
  - Complete (single copy or duplicate)
  - Fragmented
  - Missing



ASSEMBLERS

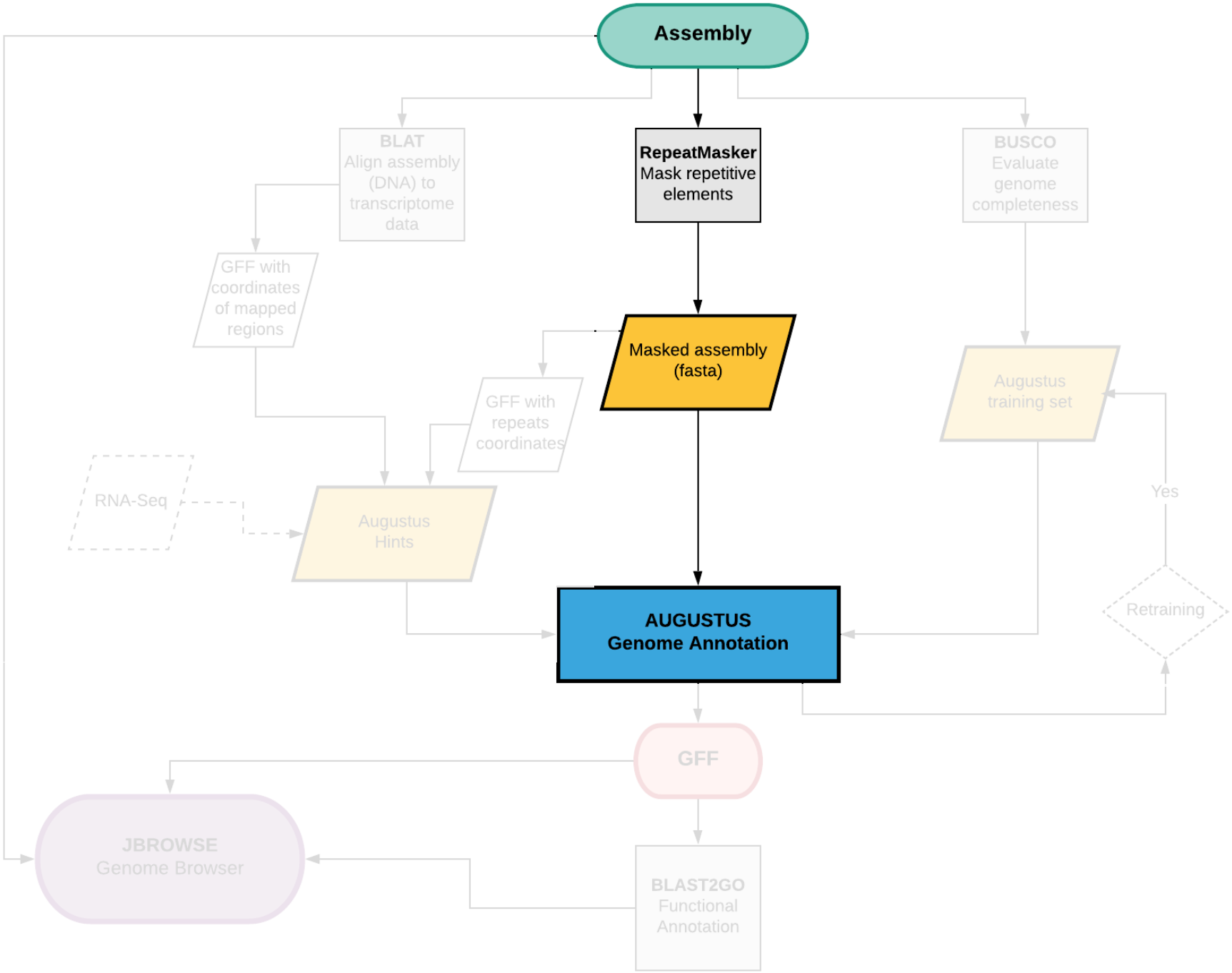
		ALLPATHS-LG	Platanus	MaSuRCA
Raccoon (34X)	Number	42,696	50,007	277,099
	N50 (Mb)	0.11	1.45	0.38
	Longest (Mb)	1.83	10.59	3.43
	Total Length (Gb)	1.79	2.25	2.78
Kinkajou (48X)	Number	23,505	15,879	67,074
	N50 (Mb)	0.29	3.55	0.12
	Longest (Mb)	3.91	15.44	1.01
	Total Length (Gb)	2.05	2.21	2.3

3 paired end libraries (350 bp) + 2 mate pair libraries (3 kb and 8 kb)


KINKAJOU (P)
SEA OTTER (A-S)
GIANT PANDA (SN)
FERRET (AP)
POLAR BEAR (SN)
RED PANDA (SN)
DOG (AR)
PACIFIC WALRUS (AP*)
RACCOON (P)
AFRICAN WILD DOG (B)
RACCOON (M)
KINKAJOU (M)
KINKAJOU (AP)
RACCOON (AP)
0%
25%
50%
75%
100%
Complete (SC)
Complete (D)
Fragmented
Missing

MASKING AND ANNOTATION OF REPETITIVE ELEMENTS

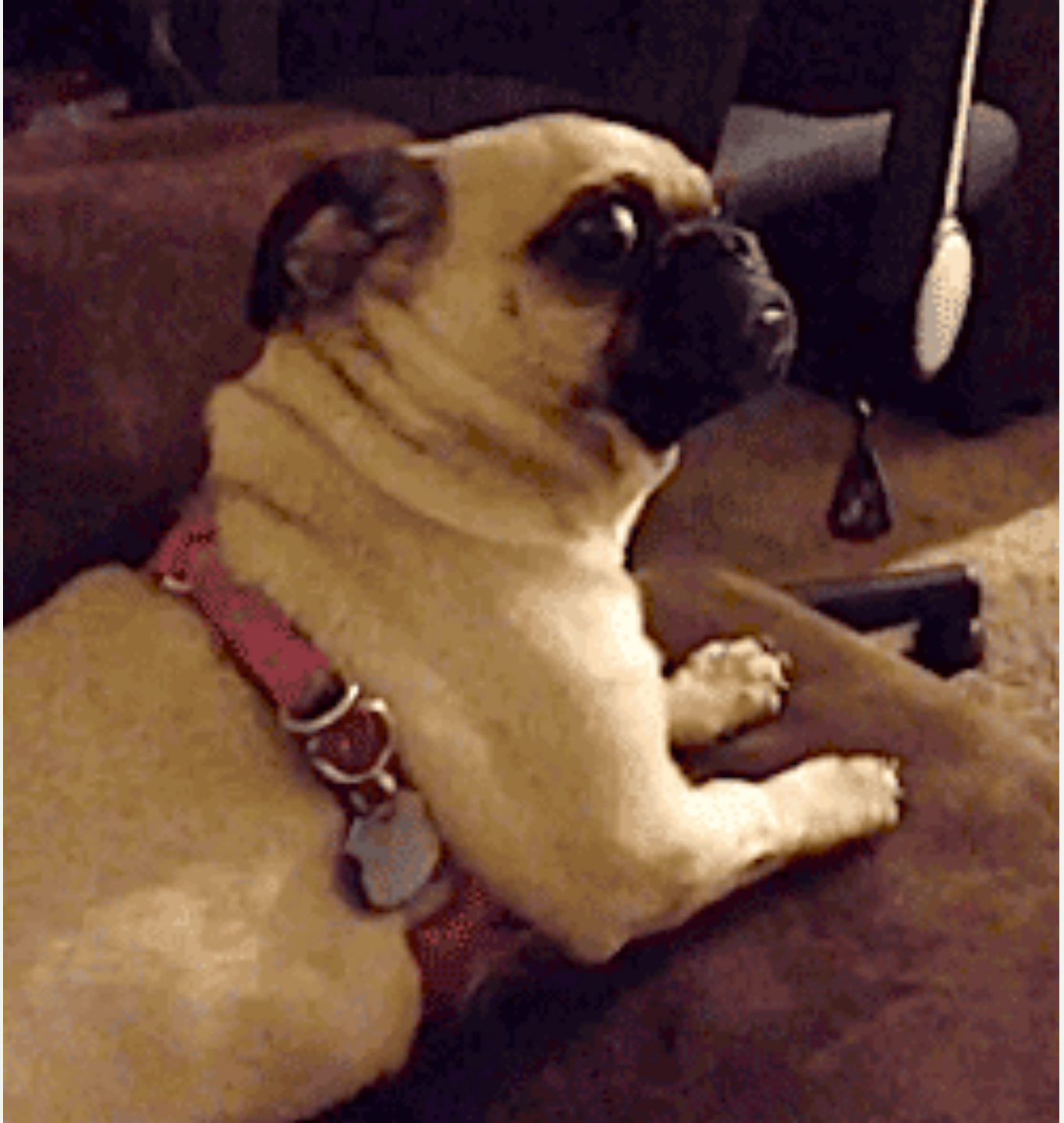
REPEATMASKER


Assembly
BLAT
Align assembly (DNA) to transcriptome data
RepeatMasker
Mask repetitive elements
BUSCO
Evaluate genome completeness
GFF with coordinates of mapped regions
Masked assembly (fasta)
Augustus training set
GFF with repeats coordinates
RNA-Seq
Augustus Hints
Yes
Retraining
AUGUSTUS
Genome Annotation
GFF
JBROWSE
Genome Browser
BLAST2GO
Functional Annotation

REPEATMASKER

- RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences.

(from the RepeatMasker website)



>Contig3141_pilon
GACGGTCTGTGCTAGTTCAGCACTCATCCACTGTTCCAGCTGGCAGCCCG
CTCGAAGGTGAAAAAATACTATAGATAGATATGTAGATGTAAAAAGCAGA
TGAGAACCGGTCGCATATATAGACAAGCTCGCGTAACTCTGGAGGAGGAG
AGTGAGCCCTGCGCTCCAGAGATGTCCTTAGAGACTCCGCCGCTCTGATA
CTGTGTCTAGCCATATATAGACAACCAACGAGTCAACCCACTACCGAAAA
GTTGCGTGGTACTGAACGCTCATTCGCGCTTGCCACGCTCTCTCTCGAGA
ACAGCTCTAGCACCAAATATAAGGACCATCCGCACATCGAGCACCTATGG
AGGCTATCTCACCCGCGTGGTGGCGAACACTCAGCGCTCACACACTATGC
GTAAGCCTCTGATTCTACTAGCCAATTGTAACATCAGCCCAAAACAGCAT
TGTTTGAGAACTTGCCCTCTACTTCTACAGATTATCGCTCAAATCTAACA
TATGCTGAGGTTCAATCTGCCCTCCGCTACTTTAGTGGAGACACATGTAA
ATACAGTCGAGGTGCATCACATTAGTAACTCTCGAGAGACTGAGTCTCTA
CAGctctctctctctccctctatctATGCCTATATGCGAGTGCTGAGA
GGACTAGCTAGTGACACGTGCTGGTATCTCTATGCTCTCTCTCATTGCTA
GATGGACACATCGTGGGTGGCAGCGCACGCGAGAGACGCATACTTAGCTG
AACGTCTCCATTATTCACTCCCTTACATCAGATAGCATAACATCTAGAAG
TTACTGTGGTAGCACCCAtttctatttatttttttttttttatttttttt
ttttttCCTACAGAAGTGCTATCTTCAGCACTGAGCGGCGAAAACAGCTA
TTTTGCAGATATTCCATCGAAAATAGACTCCAAAAACATCATCATGAAGA
TCGACCACTTCCCCCTGATATAGGATTTCACAAGTTAAGATGGAAGATGG
CTATTTTCAGGTAGCAGTGGAAGCTGCTTCCTCTTCTCTCGCCTTGGTAG
CAGCCGACGGCTGGATTTGGGCCTCCACATTGCACCCCCAGGTCTTGGCT
GCAAGGACAAGAGAGAGATACTTGGCAAAAAAATCCACGGTTGGTGCTCT
CTGTCCCCCTCTCGGCCTTTGACCTTGTTATGACTTCCTGAGGTATATAT

```
file name: siskin_10largest.fasta
sequences:            10
total length:   41463291 bp  (41459480 bp excl N/X-runs)
GC level:         42.07 %
bases masked:    1027439 bp ( 2.48 %)
================================================
```

	number of elements*	length occupied	percentage of sequence
DNA transposons	256	40450 bp	0.10 %
hobo-Activator	41	7588 bp	0.02 %
Tc1-IS630-Pogo	13	2092 bp	0.01 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	31	5322 bp	0.01 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %
Rolling-circles	1	108 bp	0.00 %
Unclassified:	100	15272 bp	0.04 %
Total interspersed repeats:		538158 bp	1.30 %
Small RNA:	110	13149 bp	0.03 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	8441	395878 bp	0.95 %
Low complexity:	1650	88173 bp	0.21 %

```
* most repeats fragmented by insertions or deletions
  have been counted as one element

The query species was assumed to be gallus gallus
RepeatMasker Combined Database: Dfam_3.0

run with rmblastn version 2.9.0+
```

Run information: input file, total length, % bases masked

Types of repetitive elements, quantity, length and percentage of sequence

Repeat database and program version

```
[tsuchiyam@hydra-login01 repmasker]$ cat siskin_10largest.fasta.out
   SW   perc perc perc  query             position in query           matching            repeat              position in repeat
score   div. del. ins.  sequence          begin   end        (left)   repeat              class/family      begin  end    (left)    ID

   35    5.1  0.0  0.0  Contig1977_pilon      392     432 (3768949) + (GCCGCT)n             Simple_repeat         1     41    (0)      1
   24   14.2  3.5  3.5  Contig1977_pilon     6593    6650 (3762731) + (CCG)n                Simple_repeat         1     58    (0)      2
   18   13.3  7.0  2.2  Contig1977_pilon     7122    7174 (3762207) + (AT)n                 Simple_repeat         1     55    (0)      3
   15    5.6  0.0  0.0  Contig1977_pilon     9974    9992 (3759389) + (T)n                  Simple_repeat         1     19    (0)      4
   15   19.7  0.0  0.0  Contig1977_pilon    20766   20794 (3748587) + (T)n                  Simple_repeat         1     29    (0)      5
   17    3.8  3.7  0.0  Contig1977_pilon    27637   27663 (3741718) + (TTTTG)n              Simple_repeat         1     28    (0)      6
   12   12.5  0.0  3.7  Contig1977_pilon    31589   31616 (3737765) + (TTTGG)n              Simple_repeat         1     27    (0)      7
  230   29.3  5.8  6.9  Contig1977_pilon    32387   32591 (3736790) + CR1-3_Croc            LINE/CR1            950   1152 (2453)     8
  236   30.6 11.1  0.0  Contig1977_pilon    32767   32874 (3736507) + CR1-L3A_Croc          LINE/CR1           3284   3403  (885)     9
   12   14.7  3.3  0.0  Contig1977_pilon    33646   33675 (3735706) + (CTGCTG)n             Simple_repeat         1     31    (0)     10
   13   17.0  0.0  5.7  Contig1977_pilon    40521   40557 (3728824) + (CCGCC)n              Simple_repeat         1     35    (0)     11
   14    4.0  7.7  0.0  Contig1977_pilon    42391   42416 (3726965) + (CACA)n               Simple_repeat         1     28    (0)     12
   13   18.9  0.0  0.0  Contig1977_pilon    43596   43625 (3725756) + (TTA)n                Simple_repeat         1     30    (0)     13
   13   26.2  4.9  1.6  Contig1977_pilon    46810   46870 (3722511) + (GCCCG)n              Simple_repeat         1     63    (0)     14
   14   16.3  4.7  2.3  Contig1977_pilon    46897   46939 (3722442) + (GGCGG)n              Simple_repeat         1     44    (0)     15
   13   15.9  2.7  2.7  Contig1977_pilon    53355   53391 (3715990) + (TTTCTT)n             Simple_repeat         1     37    (0)     16
  264   16.3  1.4 17.2  Contig1977_pilon    58730   58803 (3710578) C CR1-3_Croc            LINE/CR1             (1)   3604   3541     17
   24   16.4  0.0  0.0  Contig1977_pilon    83019   83059 (3686322) + (T)n                  Simple_repeat         1     41    (0)     18
   22   21.9  4.6  1.5  Contig1977_pilon    83423   83487 (3685894) + (AT)n                 Simple_repeat         1     67    (0)     19
   30    0.0  0.0  0.0  Contig1977_pilon    84441   84468 (3684913) + (T)n                  Simple_repeat         1     28    (0)     20
   88    0.0  0.0  0.0  Contig1977_pilon    85203   85279 (3684102) + (AAT)n                Simple_repeat         1     77    (0)     21
   12   10.3  6.1  2.9  Contig1977_pilon    87513   87545 (3681836) + (ACTC)n               Simple_repeat         1     34    (0)     22
   32   23.3  1.5  7.1  Contig1977_pilon    91116   91248 (3678133) + (GCCCCG)n             Simple_repeat         1    126    (0)     23
  125    0.0  3.1  7.6  Contig1977_pilon    94671   94863 (3674518) + (TCCTT)n              Simple_repeat         1    185    (0)     24
   11   10.6  6.2  8.5  Contig1977_pilon    98745   98792 (3670589) + (TCTGATAA)n           Simple_repeat         1     47    (0)     25
   78   21.1  2.8  2.3  Contig1977_pilon   105934  106149 (3663232) + (GCAGG)n              Simple_repeat         1    217    (0)     26
   24   13.6  9.2  0.0  Contig1977_pilon   106238  106302 (3663079) + (CCCGCGC)n            Simple_repeat         1     71    (0)     27
   21   20.5  0.0  0.0  Contig1977_pilon   110256  110295 (3659086) + A-rich               Low_complexity        1     40    (0)     28
  267   22.9  4.3  0.0  Contig1977_pilon   111594  111663 (3657718) + CR1-3_Croc           LINE/CR1           2683   2755  (850)     29
   92    0.0  0.0  0.0  Contig1977_pilon   116028  116108 (3653273) + (AAAAT)n             Simple_repeat         1     81    (0)     30
   12   11.1  0.0  9.4  Contig1977_pilon   119685  119719 (3649662) + (GCTGA)n             Simple_repeat         1     32    (0)     31
   13   28.4  0.0  0.0  Contig1977_pilon   122110  122152 (3647229) + A-rich               Low_complexity        1     43    (0)     32
   11    9.7  8.6  2.7  Contig1977_pilon   126315  126349 (3643032) + (GGAGGCT)n           Simple_repeat         1     37    (0)     33
   43    2.5  0.0  0.0  Contig1977_pilon   136676  136716 (3632665) + (TA)n                Simple_repeat         1     41    (0)     34
   13   10.8  3.1  3.1  Contig1977_pilon   154695  154726 (3614655) + (TTGGT)n             Simple_repeat         1     32    (0)     35
   13    0.0  7.7  3.7  Contig1977_pilon   157644  157669 (3611712) + (CTAAT)n             Simple_repeat         1     27    (0)     36
   15    5.5  0.0  0.0  Contig1977_pilon   157828  157846 (3611535) + (AC)n                Simple_repeat         1     19    (0)     37
   12    5.5  0.0  0.0  Contig1977_pilon   176097  176115 (3593266) + (CCA)n               Simple_repeat         1     19    (0)     38
   15   24.4  5.8  1.4  Contig1977_pilon   176667  176735 (3592646) + A-rich               Low_complexity        1     72    (0)     39
   13    7.8  7.4  0.0  Contig1977_pilon   177664  177690 (3591691) + (CTC)n               Simple_repeat         1     29    (0)     40
  225   11.3  6.5  4.3  Contig1977_pilon   179927  179972 (3589409) + CR1-3_Croc           LINE/CR1           1299   1345 (2260)     41
   15   24.1  0.0  2.2  Contig1977_pilon   182585  182630 (3586751) + G-rich               Low_complexity        1     45    (0)     42
```

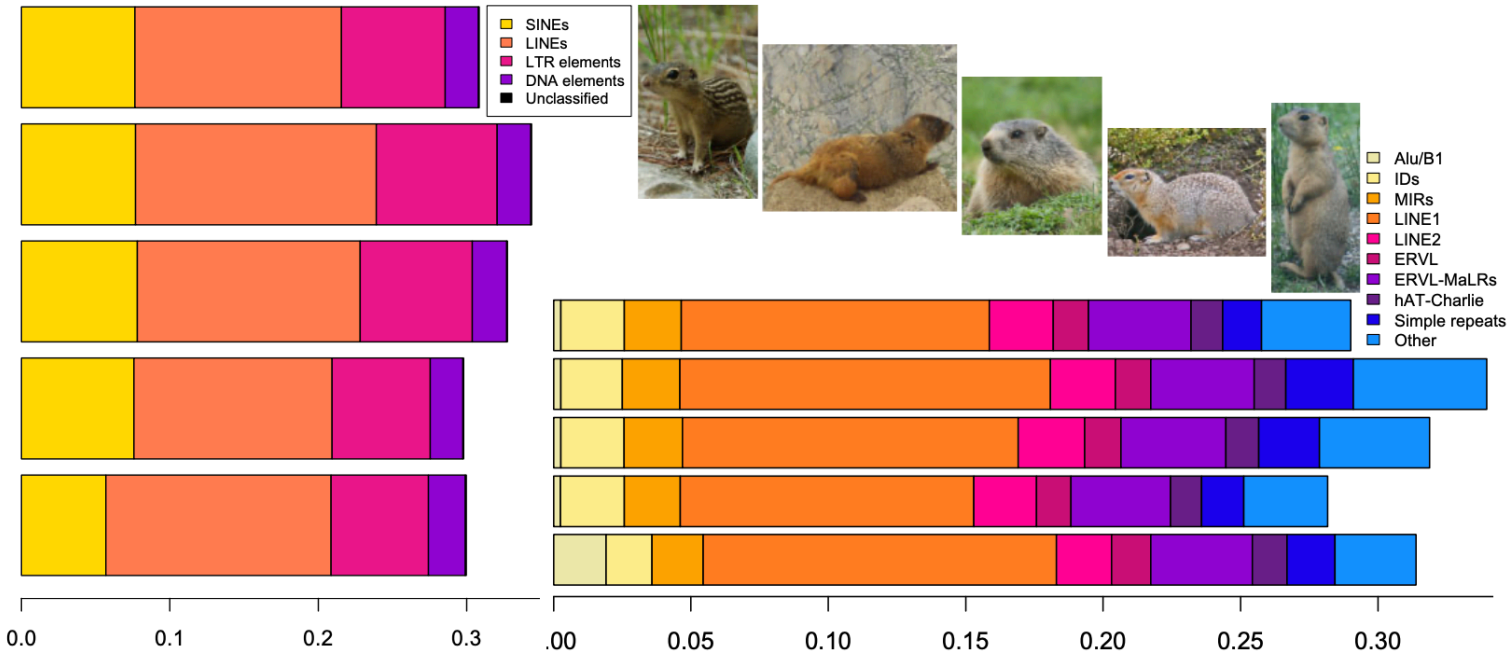


Fig. 2.—Percent repeat content (repeat classes, left; repeat subclasses, right) in ground squirrel genomes. Top to bottom, and pictured left to right: *I. tridecemlineatus*, *M. flaviventris*, *M. marmota*, *Urocitellus parryii*, *C. gunnisoni*. *M. flaviventris*, and *C. gunnisoni* images; copyright: Loren Cassin-Sackett; others publicly available from Wiki Commons.

REPEATMASKER

- RepeatMasker has several repetitive elements databases (Eukaryotic species only)

- Commonly used species include: mammal, carnivore, rodentia, rat, cow, pig, cat, dog, chicken, fugu, danio, "ciona intestinalis" drosophila, anopheles, elegans, diatoaea, artiodactyl, arabidopsis, rice, wheat, and maize

HOW TO KNOW WHICH SPECIES TO USE?

REPEATMASKER

- To query the RepeatMasker database by taxonomy, you can use the following command:

```
queryTaxonomyDatabase.pl -species Spinus
```

- You can also see the repeat library available for your species

```
queryRepeatDatabase.pl -species Spinus
```

- Figure out which species/taxon to use for siskin using `Repbase_RepeatQuery_Taxonomy_MTNT.sh`

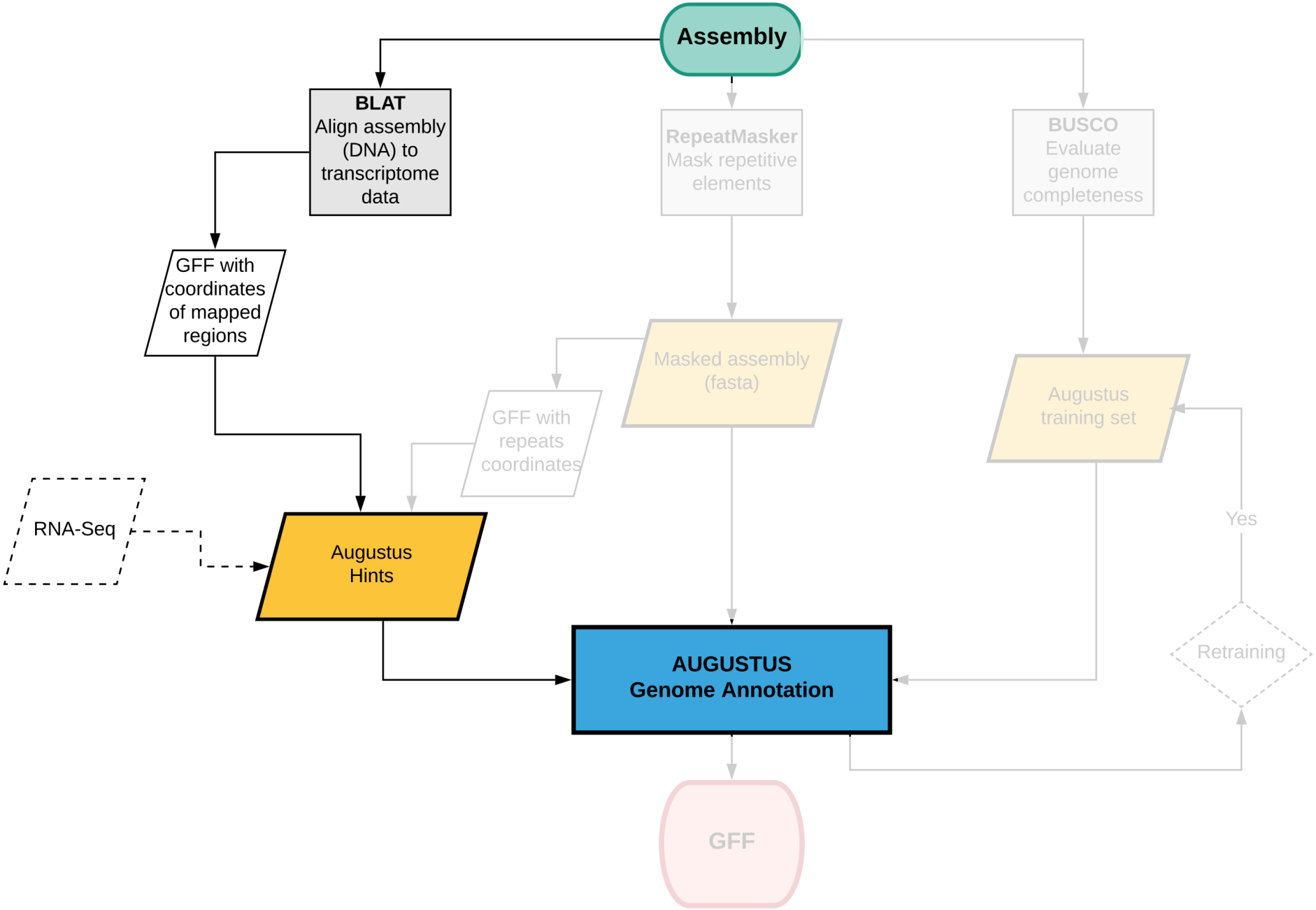
BLAT

BLAST-LIKE ALIGNMENT TOOL

OTHER SOURCES OF EVIDENCE

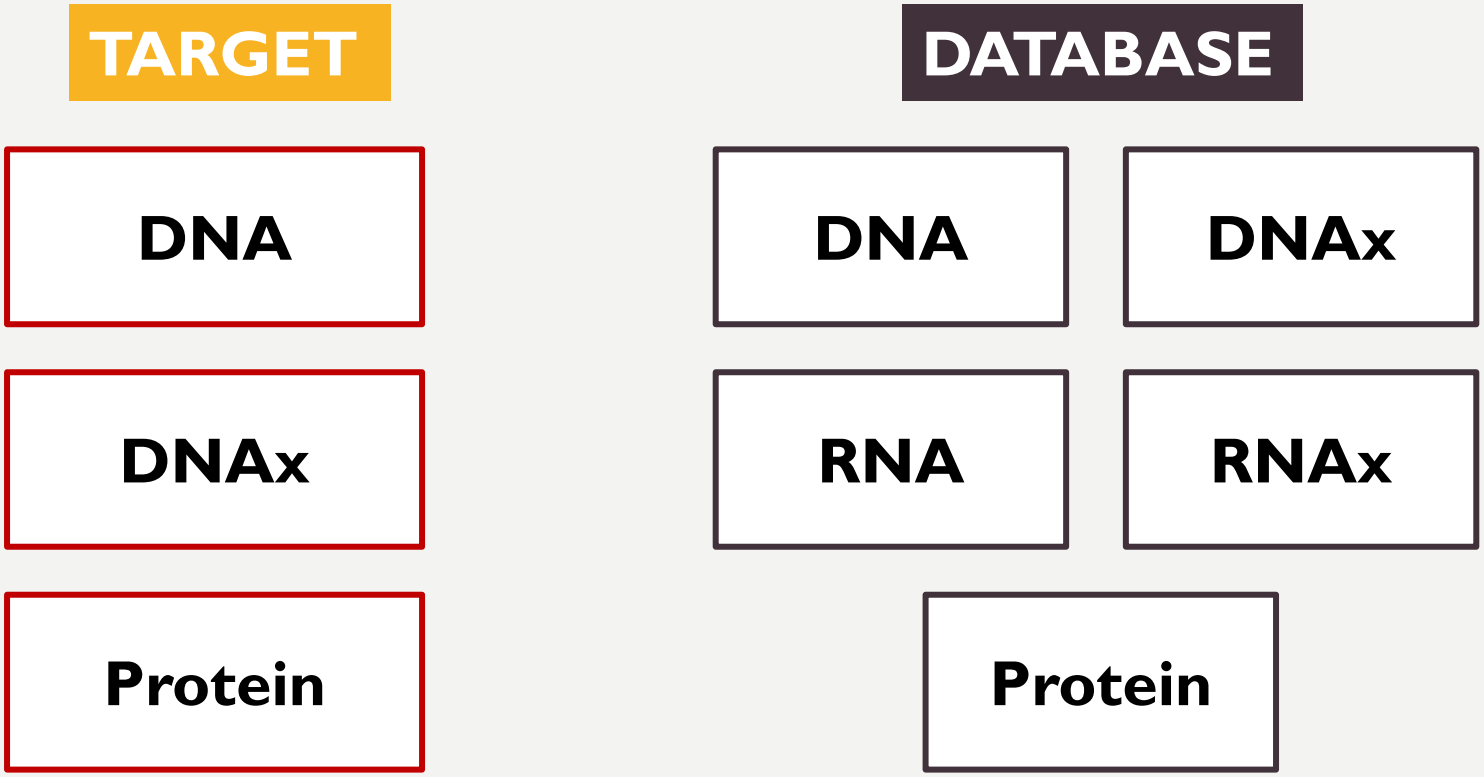
- RNA-Seq
- Transcriptomes

Today we will use the transcriptome of a different species to generate another source of information for the annotation


Assembly
BLAT
Align assembly (DNA) to transcriptome data
RepeatMasker
Mask repetitive elements
BUSCO
Evaluate genome completeness
GFF with coordinates of mapped regions
Masked assembly (fasta)
Augustus training set
GFF with repeats coordinates
RNA-Seq
Augustus Hints
Yes
Retraining
AUGUSTUS
Genome Annotation
GFF

BLAT

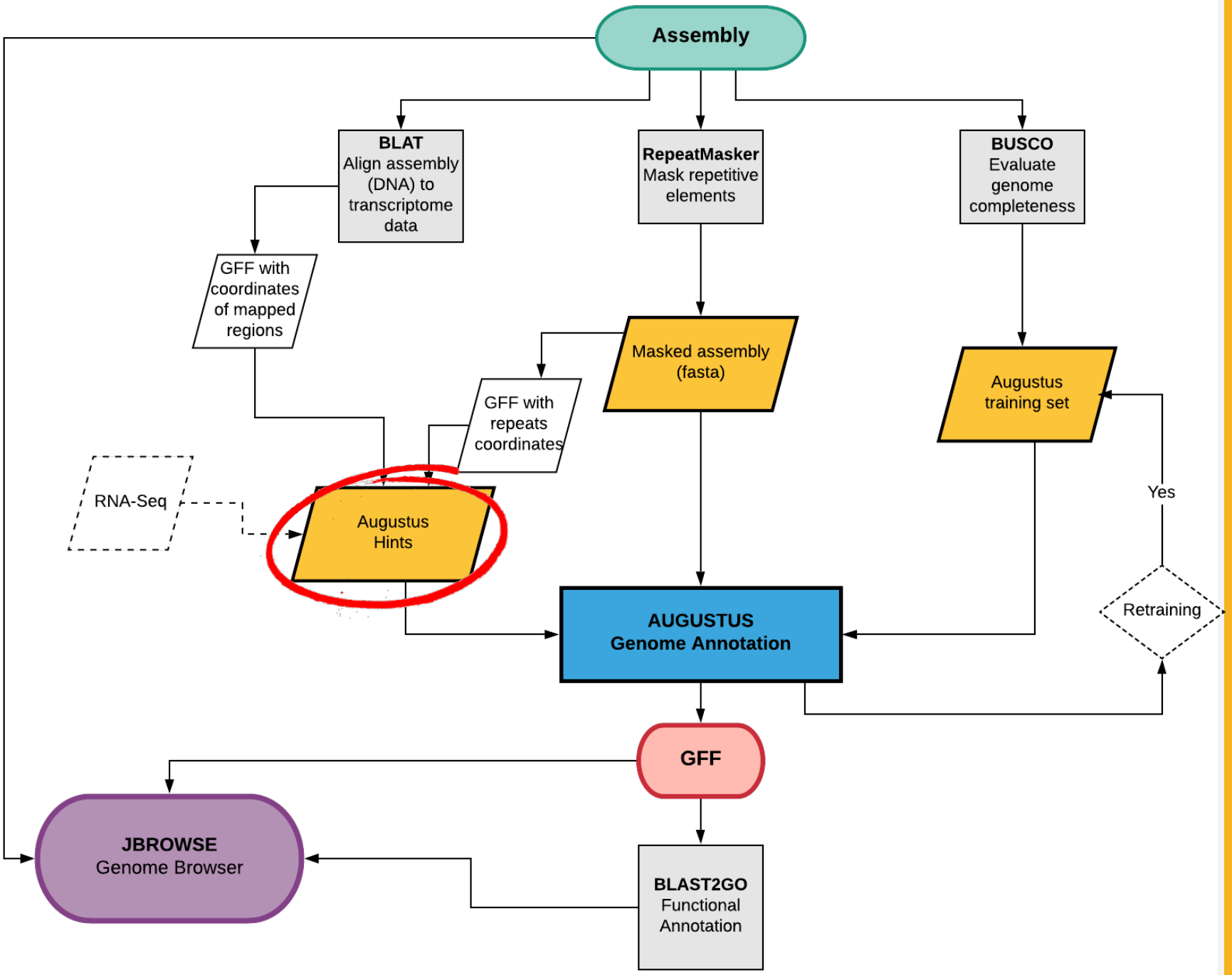
- BLAST-like Alignment Tool



DNAx and RNAx correspond to 6-frame translated sequences

BLAT: WHAT DOES IT TELL US?

- It provides information regarding exons and introns, based on the alignment of the transcriptome sequence to our assembly.


Assembly
BLAT
Align assembly (DNA) to transcriptome data
RepeatMasker
Mask repetitive elements
BUSCO
Evaluate genome completeness
GFF with coordinates of mapped regions
Masked assembly (fasta)
Augustus training set
GFF with repeats coordinates
RNA-Seq
Augustus Hints
Yes
Retraining
AUGUSTUS
Genome Annotation
GFF
JBROWSE
Genome Browser
BLAST2GO
Functional Annotation

TUTORIAL

GITHUB