GENOME ANNOTATION WORKSHOP

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

WORKSHOP INFO

INTRODUCTIONS, INFO, CODE OF CONDUCT

WORKSHOP INFO

- GitHub: SmithsonianWorkshops
  https://github.com/SmithsonianWorkshops/2019-06-04-NMNH

- Etherpad: We will use to take notes
  https://pad.carpentries.org/2019-06-04-NMNH

- Helpers:
  - Rebecca Dikow
  - Mike Trizna
  - Vanessa Gonzalez
  - Maddy Bursell

WORKSHOP INFO

- Code of Conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html
  - Use welcoming and inclusive language
  - Be respectful of different viewpoints and experiences
  - Gracefully accept constructive criticism
  - Focus on what is best for the community. Show respect and courtesy towards other community members

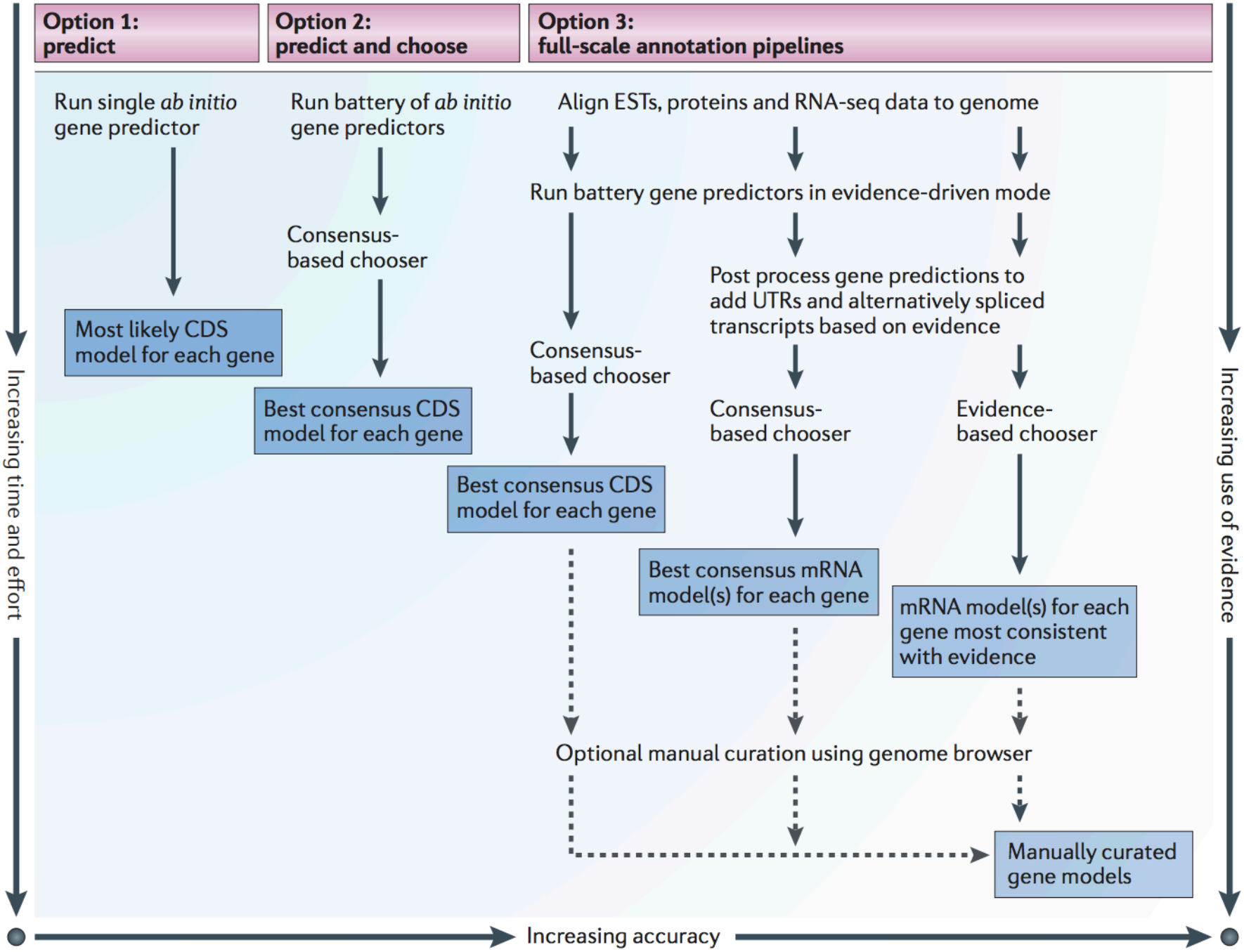
WHAT IS GENOME ANNOTATION?

- Genome annotation is the process of identifying different elements in a genome assembly:
  - Structural
    - Repetitive elements
    - Genes (with introns and exons)
  - Functional
    - What does each gene do?

GENE PREDICTION

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

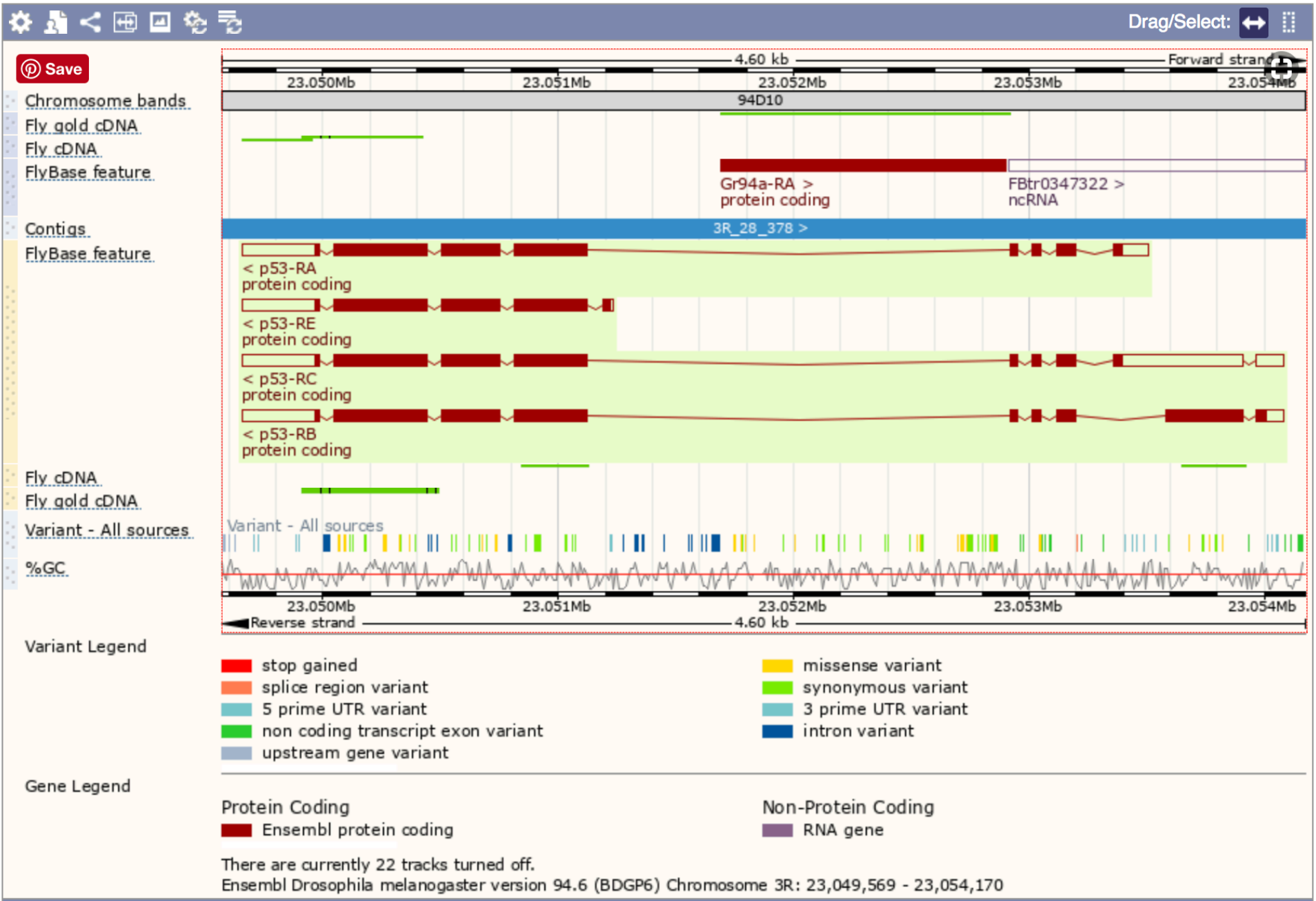
Combining both approaches is the best option


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

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



SOFTWARE

Augustus

SNAP

GlimmerHMM

Genemark-ES

FGenesh

Gnomon

OTHER ALTERNATIVES

NCBI

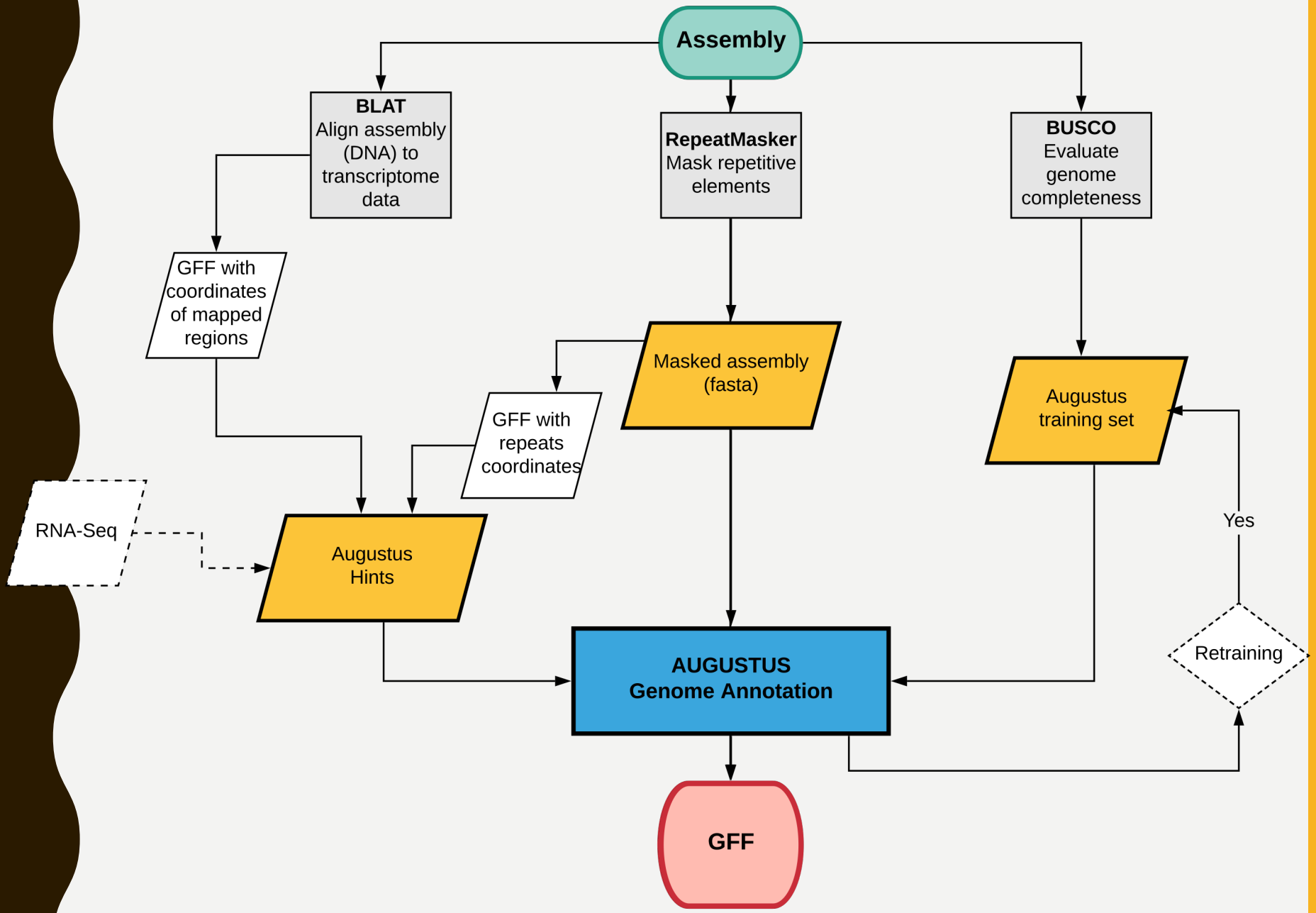
e!Ensembl

Pros

- Standardized
- Free

Limitations

- Quality requirements
- Taxonomic priorities


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
AUGUSTUS
Genome Annotation
Retraining
GFF

LET’S START!

TASKS:

1. Go to the workshop page https://github.com/SmithsonianWorkshops/2019-06-04-NMNH
   a) Open the GenomeAnnotationGuide.md

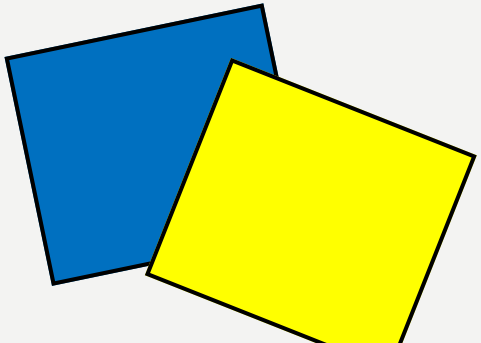
2. Open the Etherpad
   a) Write your name, unit, something you accomplished, favorite food.

3. Open another tab with the QSubGenerator https://hydra-4.si.edu/tools/QSubGenerator/

4. Login to Hydra:

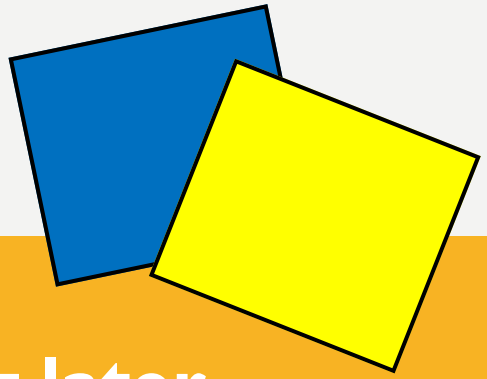
```
SSH USERNAME@HYDRA-LOGIN01.SI.EDU
```

When you are done, put a BLUE sticky note up.
If you have any issues, put the YELLOW one instead.



FOLDER STRUCTURE

1. Create the folder GAworkshop
2. Then, create the following folders inside GAworkshop
   - assembly
   - busco
   - repeatmasker
   - blat
   - augustus
   - jbrowse
   - logs
   - jobs

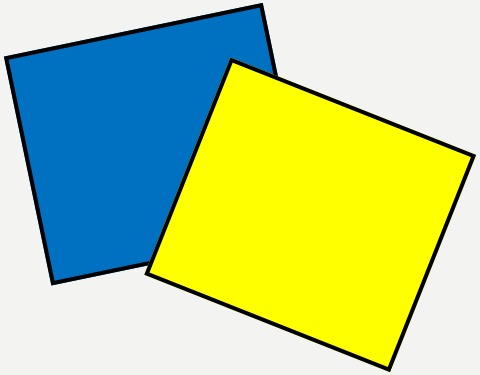


Why?
It is easier to find everything later.

ASSEMBLY STATS

```
GAworkshop
     |__ assembly
     |__ augustus
     |__ blat
     |__ busco
     |__ jbrowse
     |__ jobs
     |__ logs
     |__ repeatmasker
```

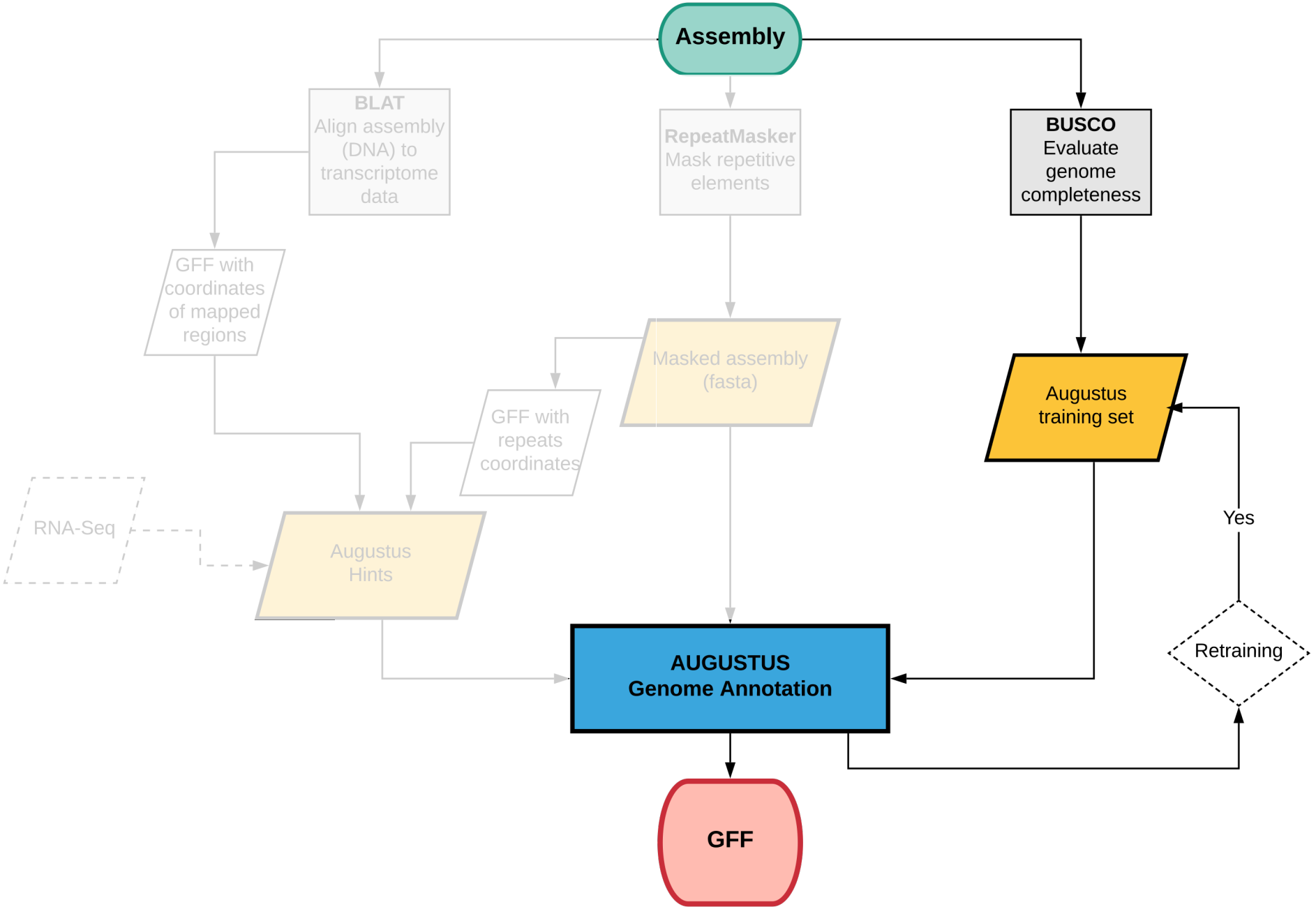
- ASSEMBLY
  - Species: *Drosophila hydei* (Dhydei_genome.fa)
  - Copy the file to your folder assembly. File location: /data/genomics/workshops/Gaworkshop/Dhydei_genome.fa
- We will use a python script (available in the module assembly_stats) to get some basic info about this genome.
- We will run this part from the interactive queue. Login to it using qrsh



ASSEMBLY STATS FROM YOUR GENOMES

BUSCO

BENCHMARKING UNIVERSAL SINGLE-COPY ORTHOLOGS


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

BUSCO

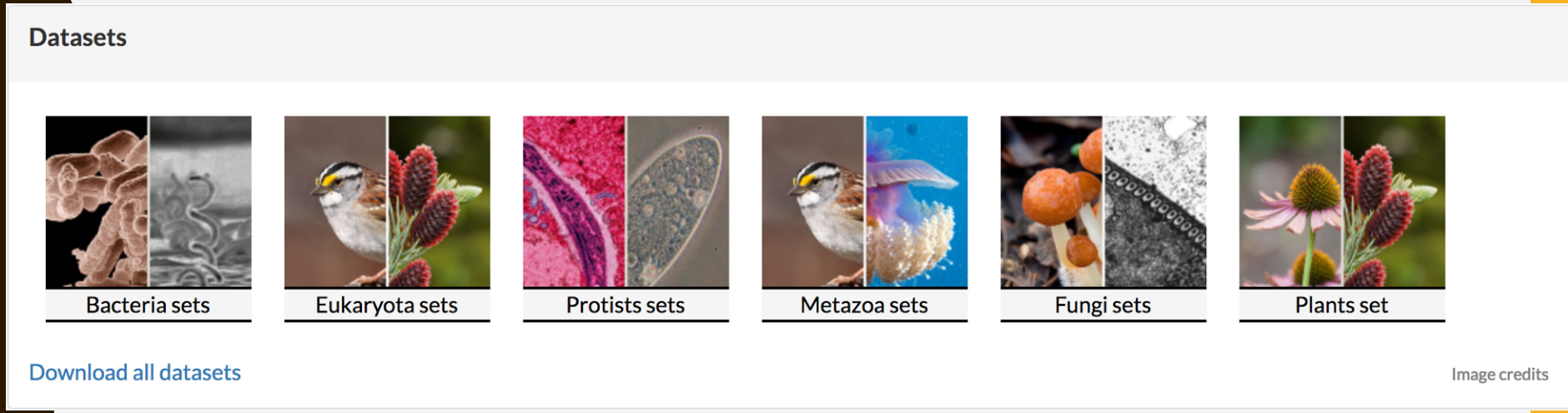
- Benchmarking Universal Single-Copy Orthologs

What is a ortholog?

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

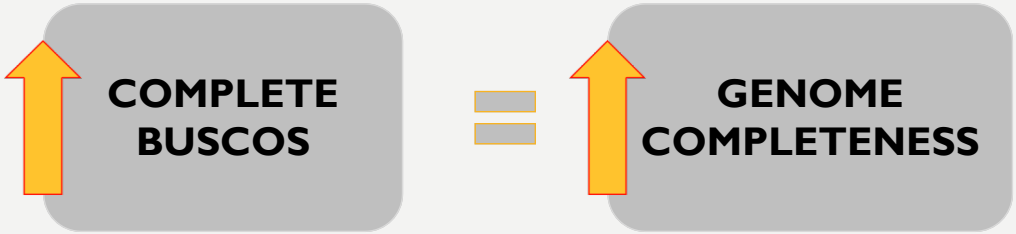
BUSCO:
HOW COMPLETE IS THE ASSEMBLY?

- Database: taxon-specific single copy orthologs



BUSCO:
HOW COMPLETE IS THE ASSEMBLY?

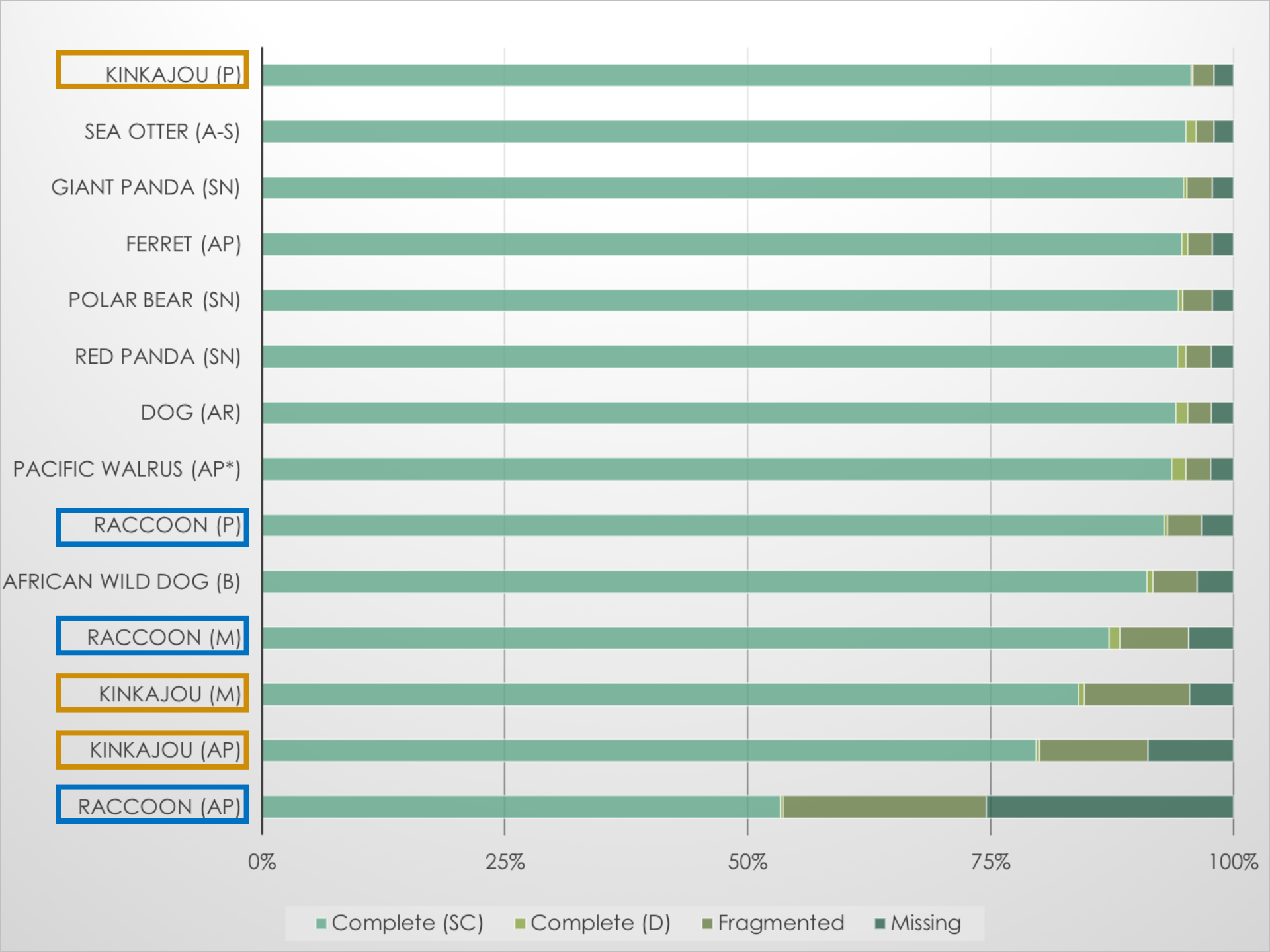
- Assessment:
  - Complete (single copy or duplicate)
  - Fragmented
  - Missing



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

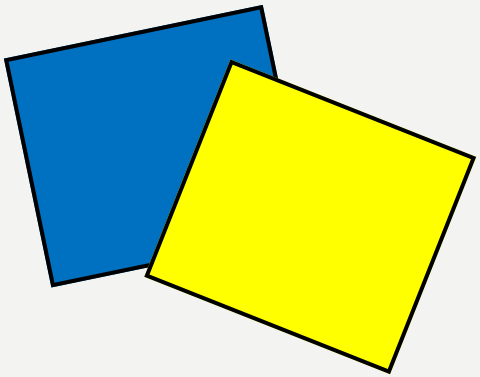
3 paired end libraries (350 bp) + 2 mate pair libraries (3 kb and 8 kb)


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

BUSCO: TASKS

```
GAworkshop
    |__ assembly
    |__ augustus
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repeatmasker
```

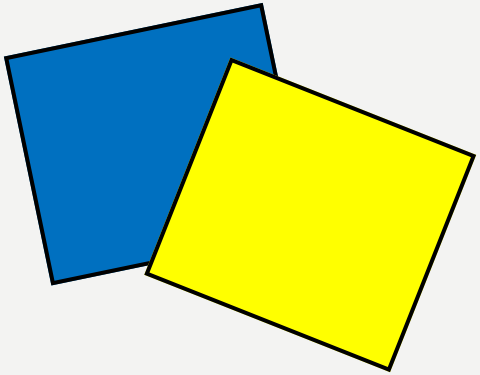
1. Copy the augustus config folder to YOUR augustus folder



BUSCO: TASKS

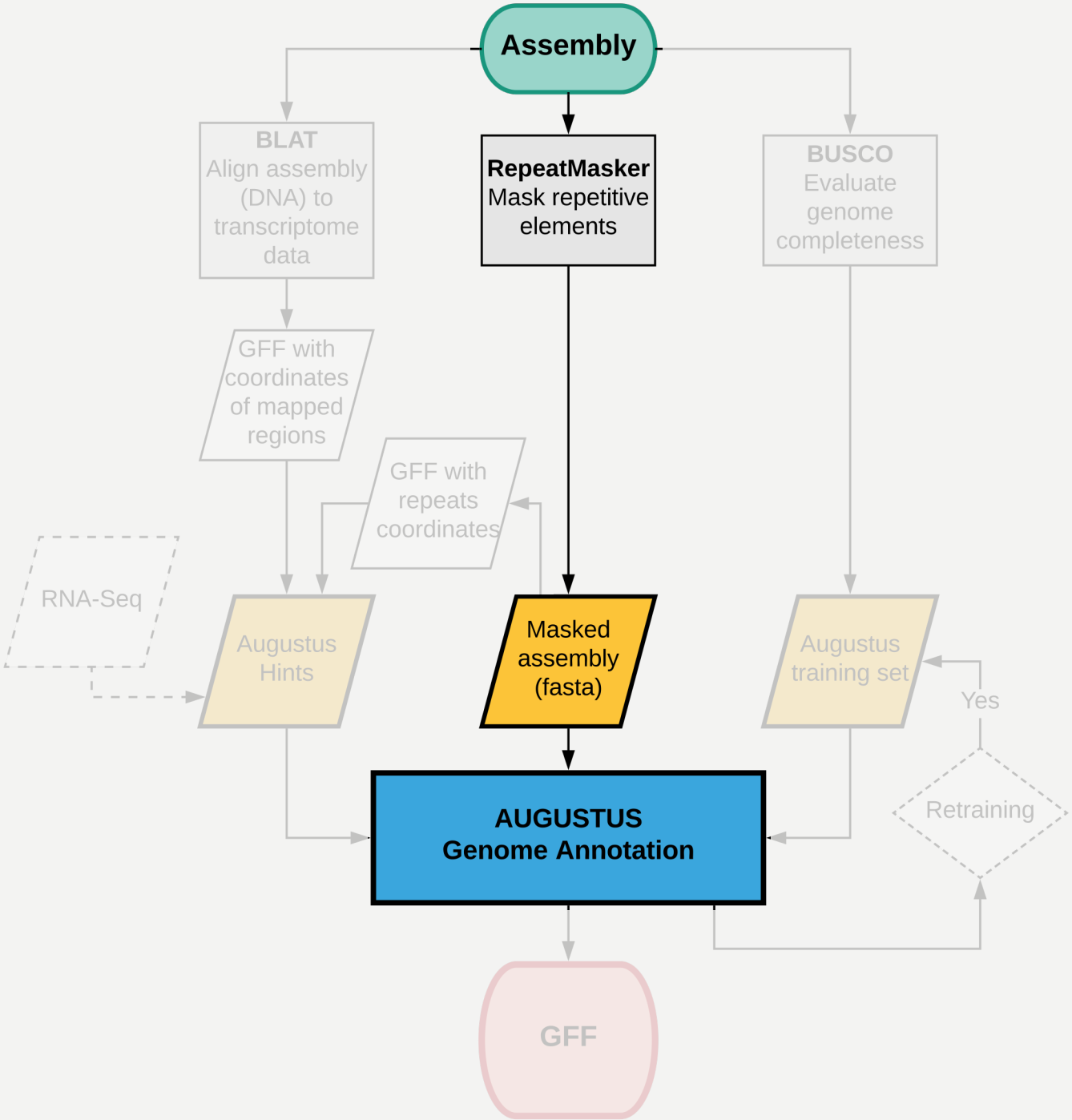
```
GAworkshop
    |__ assembly
    |__ augustus
          |__ config
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repeatmasker
```

1. Download the most appropriate database to your busco folder (in this case, we will use Diptera)
2. Create and submit the BUSCO job.



MASKING AND ANNOTATION OF REPETITIVE ELEMENTS

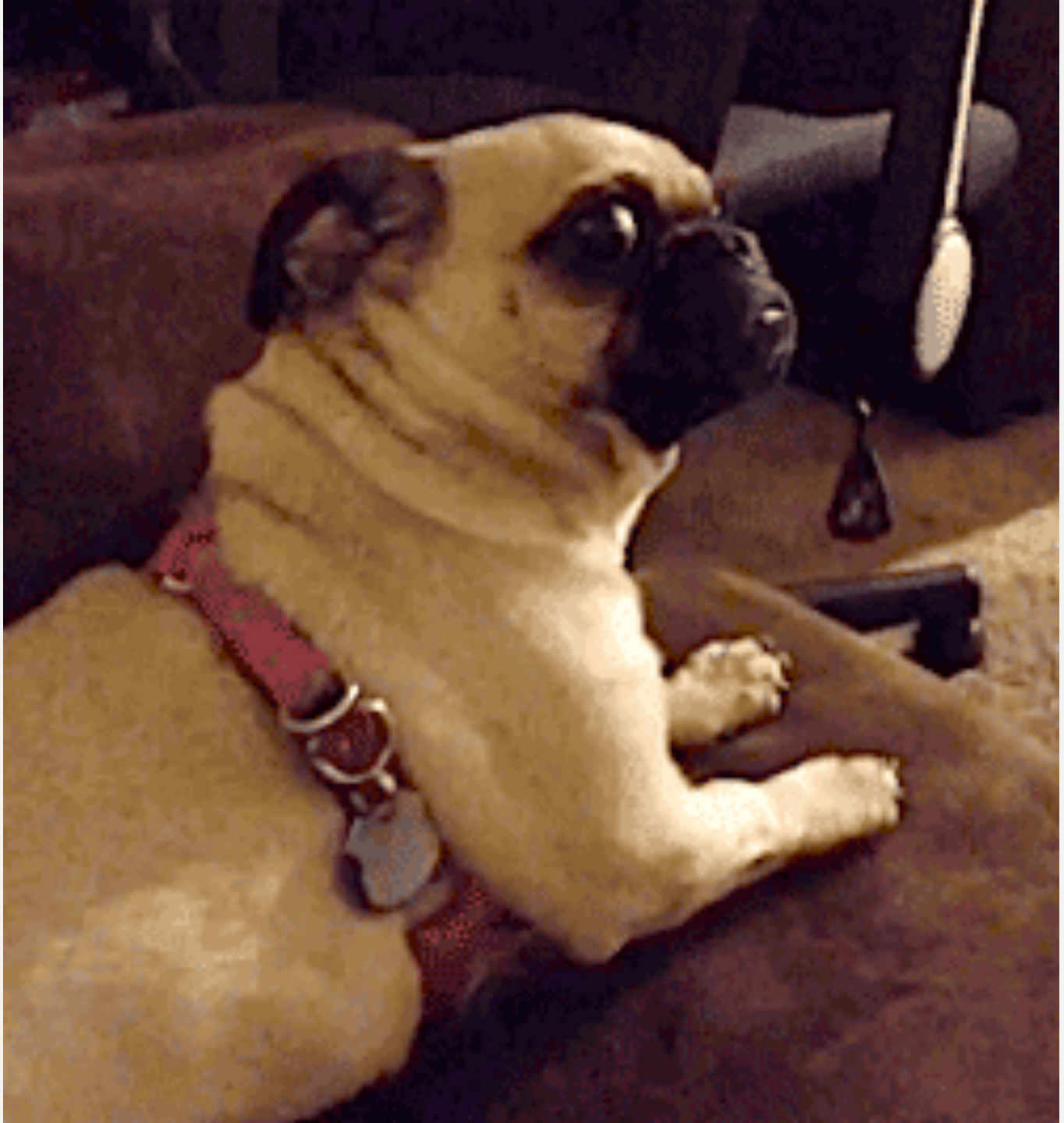
REPEATMASKER


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

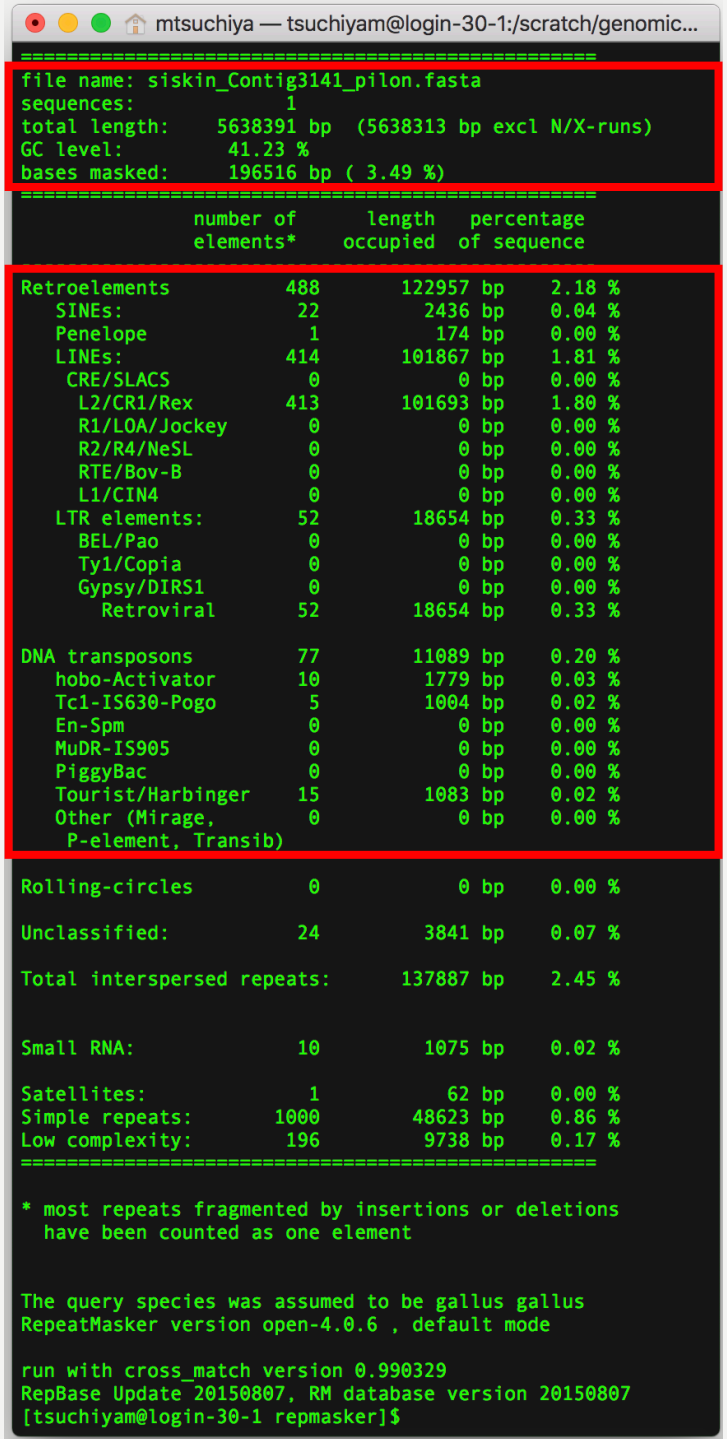
REPEATMASKER

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

(from the RepeatMasker website)




Desktop — less siskin_Contig3141_pilon.fasta.masked — 146×44
>Contig3141_pilon
siskin_Contig3141_pilon.fasta.masked

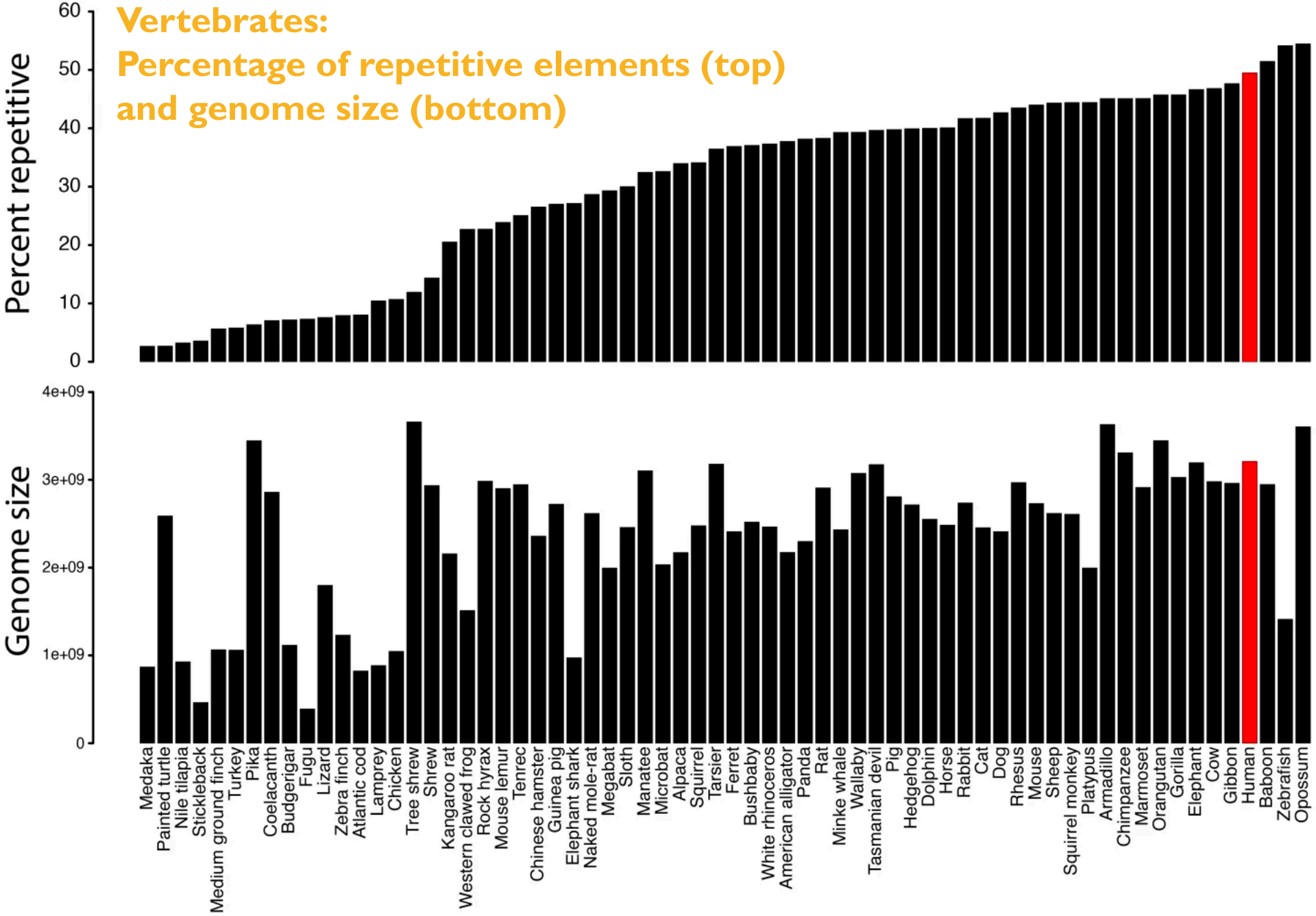


Run information: input file, total length, % bases masked

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

```
[[tsuchiyam@login-30-1 repmasker]$ head -n 200 *.out
```

SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin	end	(left)	matching repeat	repeat class/family	position in repeat begin	end	(left)	ID
17	8.8	0.0	0.0	Contig3141_pilon	604	627	(5637764) +	(CT)n	Simple_repeat	1	24	(0)	1
26	11.3	0.0	0.0	Contig3141_pilon	819	856	(5637535) +	(T)n	Simple_repeat	1	38	(0)	2
30	20.8	3.4	2.2	Contig3141_pilon	3372	3459	(5634932) +	(TGTT)n	Simple_repeat	1	89	(0)	3
12	8.2	3.6	3.6	Contig3141_pilon	3852	3879	(5634512) +	(GCCT)n	Simple_repeat	1	28	(0)	4
780	23.1	5.7	2.9	Contig3141_pilon	5845	6049	(5632342) +	CR1-X2	LINE/CR1	3923	4133	(6)	5
18	18.1	0.0	0.0	Contig3141_pilon	6253	6290	(5632101) +	(ATT)n	Simple_repeat	1	38	(0)	6
14	21.8	0.0	4.5	Contig3141_pilon	17123	17168	(5621223) +	G-rich	Low_complexity	1	44	(0)	7
459	24.3	0.0	1.8	Contig3141_pilon	23515	23627	(5614764) +	CR1-X1	LINE/CR1	4023	4133	(4)	8
1488	27.5	1.5	1.2	Contig3141_pilon	28192	28600	(5609791) C	CR1-F2	LINE/CR1	(16)	4497	4088	9
254	24.1	0.0	0.0	Contig3141_pilon	28926	28983	(5609408) C	UCON24	Unknown	(111)	263	206	10
20	22.2	0.0	3.4	Contig3141_pilon	30324	30384	(5608007) +	A-rich	Low_complexity	1	59	(0)	11
21	0.0	0.0	0.0	Contig3141_pilon	40281	40303	(5598088) +	(T)n	Simple_repeat	1	23	(0)	12
969	24.0	2.5	5.5	Contig3141_pilon	42725	43050	(5595341) C	CR1-X2	LINE/CR1	(41)	4098	3783	13
15	8.0	0.0	3.6	Contig3141_pilon	44280	44308	(5594083) +	(TCTTC)n	Simple_repeat	1	28	(0)	14
32	0.0	0.0	0.0	Contig3141_pilon	44314	44343	(5594048) +	(T)n	Simple_repeat	1	30	(0)	15
13	18.4	7.0	0.0	Contig3141_pilon	44531	44573	(5593818) +	(CTGCTG)n	Simple_repeat	1	46	(0)	16
13	18.0	0.0	3.0	Contig3141_pilon	47527	47560	(5590831) +	(CCTCCC)n	Simple_repeat	1	33	(0)	17
631	24.5	6.6	4.2	Contig3141_pilon	51189	51380	(5587011) +	CR1-H	LINE/CR1	4602	4798	(14)	18
19	0.0	0.0	3.7	Contig3141_pilon	52831	52858	(5585533) +	(AACA)n	Simple_repeat	1	27	(0)	19
15	5.6	0.0	0.0	Contig3141_pilon	57134	57152	(5581239) +	(T)n	Simple_repeat	1	19	(0)	20
1008	13.1	9.6	0.5	Contig3141_pilon	60288	60487	(5577904) +	CR1-C4	LINE/CR1	4289	4508	(3)	21
16	0.0	0.0	0.0	Contig3141_pilon	64723	64739	(5573652) +	(T)n	Simple_repeat	1	17	(0)	22
15	19.9	0.0	0.0	Contig3141_pilon	64868	64896	(5573495) +	A-rich	Low_complexity	1	29	(0)	23
12	3.5	5.7	8.8	Contig3141_pilon	65872	65906	(5572485) +	(CTTTA)n	Simple_repeat	1	34	(0)	24
47	0.0	0.0	0.0	Contig3141_pilon	72051	72093	(5566298) +	(T)n	Simple_repeat	1	43	(0)	25
47	39.1	0.7	2.2	Contig3141_pilon	72913	73192	(5565199) +	(GT)n	Simple_repeat	1	276	(0)	26
429	25.7	6.4	4.8	Contig3141_pilon	76071	76299	(5562092) C	GGLTR8B	LTR/ERVL	(837)	234	2	27
38	2.3	0.0	0.0	Contig3141_pilon	78999	79042	(5559349) +	(A)n	Simple_repeat	1	44	(0)	28
28	14.7	0.0	0.0	Contig3141_pilon	80244	80288	(5558103) +	(A)n	Simple_repeat	1	45	(0)	29
35	10.5	0.0	0.0	Contig3141_pilon	80590	80640	(5557751) +	(A)n	Simple_repeat	1	51	(0)	30
1021	23.5	6.7	1.7	Contig3141_pilon	84211	84687	(5553704) +	CR1-C4	LINE/CR1	3956	4516	(27)	31
235	29.2	3.0	0.0	Contig3141_pilon	90620	90684	(5547707) +	Chompy-2_Croc	DNA/PIF-Harbinger	6	72	(0)	32
13	9.8	0.0	0.0	Contig3141_pilon	92076	92097	(5546294) +	(GGA)n	Simple_repeat	1	22	(0)	33
513	15.5	1.1	0.0	Contig3141_pilon	93425	93538	(5544853) +	CR1-F2	LINE/CR1	3888	3982	(531)	34
12	8.1	3.6	3.6	Contig3141_pilon	97420	97447	(5540944) +	GA-rich	Low_complexity	1	28	(0)	35
12	12.7	2.5	7.9	Contig3141_pilon	105002	105041	(5533350) +	(TGTATA)n	Simple_repeat	1	38	(0)	36
17	12.5	0.0	0.0	Contig3141_pilon	109381	109406	(5528985) +	(T)n	Simple_repeat	1	26	(0)	37
13	11.1	0.0	6.5	Contig3141_pilon	110496	110528	(5527863) +	(TATGCA)n	Simple_repeat	1	31	(0)	38
263	28.0	9.1	1.0	Contig3141_pilon	123790	123890	(5514501) C	LFSINE_Vert	SINE/tRNA	(307)	152	43	39
20	10.0	0.0	0.0	Contig3141_pilon	124173	124204	(5514187) +	(ATA)n	Simple_repeat	1	32	(0)	40
916	19.7	11.7	0.5	Contig3141_pilon	132695	132913	(5505478) C	CR1-C4	LINE/CR1	(7)	4504	4258	41
13	11.7	3.6	0.0	Contig3141_pilon	143379	143406	(5494985) +	(CTGTG)n	Simple_repeat	1	29	(0)	42
13	0.0	0.0	0.0	Contig3141_pilon	146450	146466	(5491925) +	(GCA)n	Simple_repeat	1	17	(0)	43
871	26.1	0.0	4.6	Contig3141_pilon	154617	154877	(5483514) C	MER126	DNA	(199)	250	2	44
15	19.9	0.0	0.0	Contig3141_pilon	155007	155035	(5483356) +	A-rich	Low_complexity	1	29	(0)	45
20	16.4	0.0	0.0	Contig3141_pilon	158673	158706	(5479685) +	(T)n	Simple_repeat	1	34	(0)	46
19	4.5	0.0	0.0	Contig3141_pilon	162947	162969	(5475422) +	(T)n	Simple_repeat	1	23	(0)	47
19	0.0	0.0	0.0	Contig3141_pilon	163664	163684	(5474707) +	(TA)n	Simple_repeat	1	21	(0)	48
311	21.1	3.1	6.9	Contig3141_pilon	170755	170856	(5467535) C	CR1-8_Crp	LINE/CR1	(813)	2045	1948	49
16	18.9	0.0	0.0	Contig3141_pilon	173044	173073	(5465318) +	(AC)n	Simple_repeat	1	30	(0)	50
12	14.4	3.0	3.0	Contig3141_pilon	173908	173940	(5464451) +	(AATGAA)n	Simple_repeat	1	33	(0)	51
400	18.4	0.9	0.2	Contig3141_pilon	175213	175287	(5463104) C	CR1-X1	LINE/CR1	(1)	4136	4071	52
14	27.1	0.0	0.0	Contig3141_pilon	176602	176645	(5461746) +	A-rich	Low_complexity	1	44	(0)	53
12	7.8	3.2	6.7	Contig3141_pilon	187232	187262	(5451129) +	(ATAA)n	Simple_repeat	1	30	(0)	54
12	26.5	0.0	5.8	Contig3141_pilon	191939	191993	(5446398) +	(TTTC)n	Simple_repeat	1	52	(0)	55



https://davetang.org/muse/2014/05/22/repetitive-elements-in-vertebrate-genomes/

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

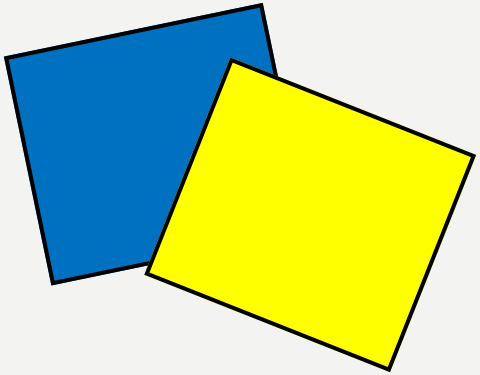
- To query the RepeatMasker database by taxonomy, you can use the following command:

```
queryTaxonomyDatabase.pl -species cat
```

PARAMETERS

```
GAworkshop
    |__ assembly
    |__ augustus
          |__ config
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repeatmasker
```

- RepeatMasker
- -species drosophila: RepBase species
- -xsmall : soft-masking (repetitive elements are masked in low caps instead of replaced by N)
- -gff: additional output in gff2 format
- -pa $NSLOTS: number of cpus
- -dir .: output the results to the current folder
- ../assembly/Dhydei_genome.fa: input file



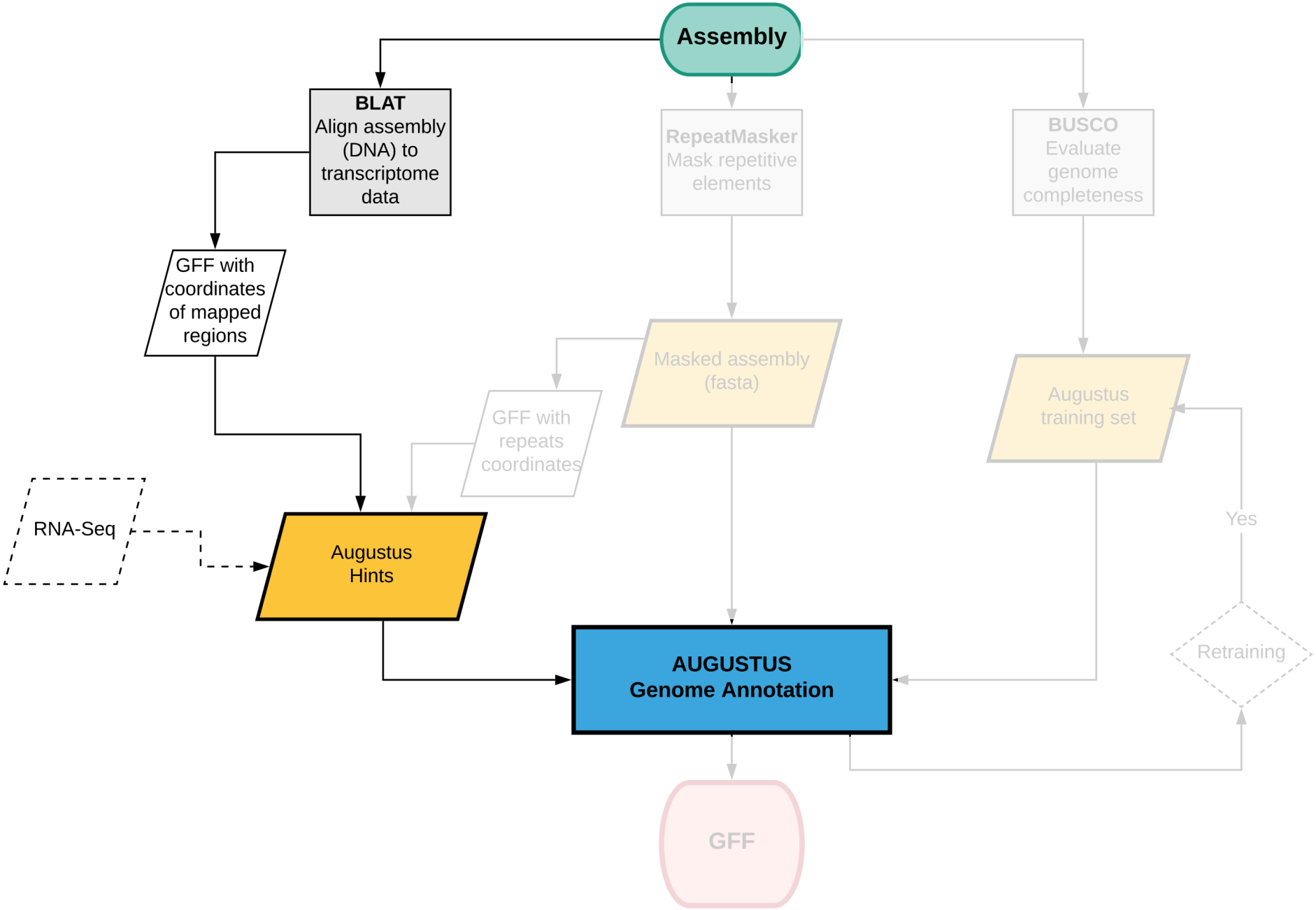
BLAT

BLAST-LIKE ALIGNMENT TOOL

OTHER SOURCES OF EVIDENCE

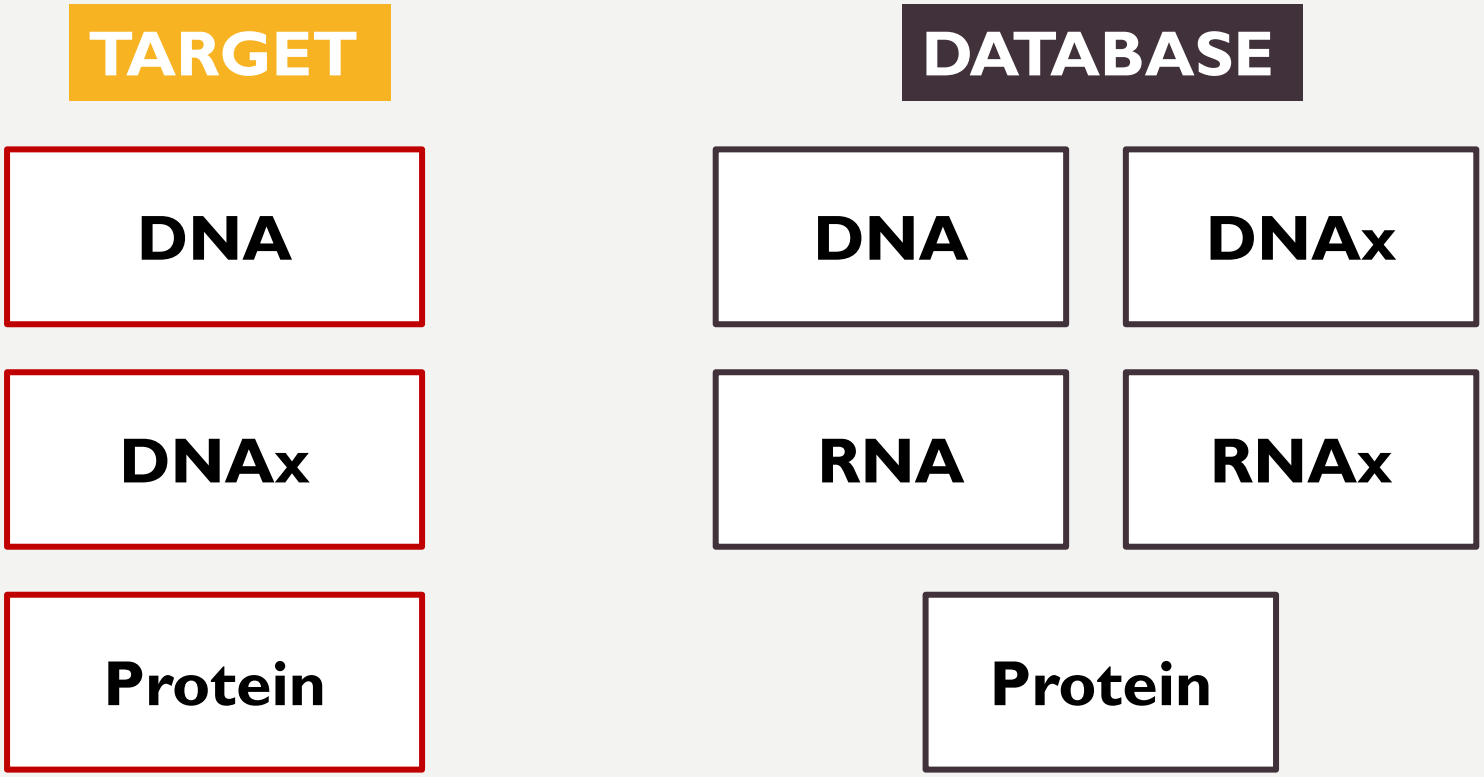
- RNA-Seq
- Transcriptomes

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

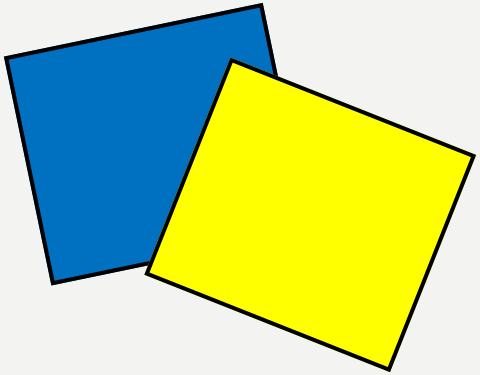
BLAT: WHAT DOES IT TELL US?

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

BLAT: TASKS

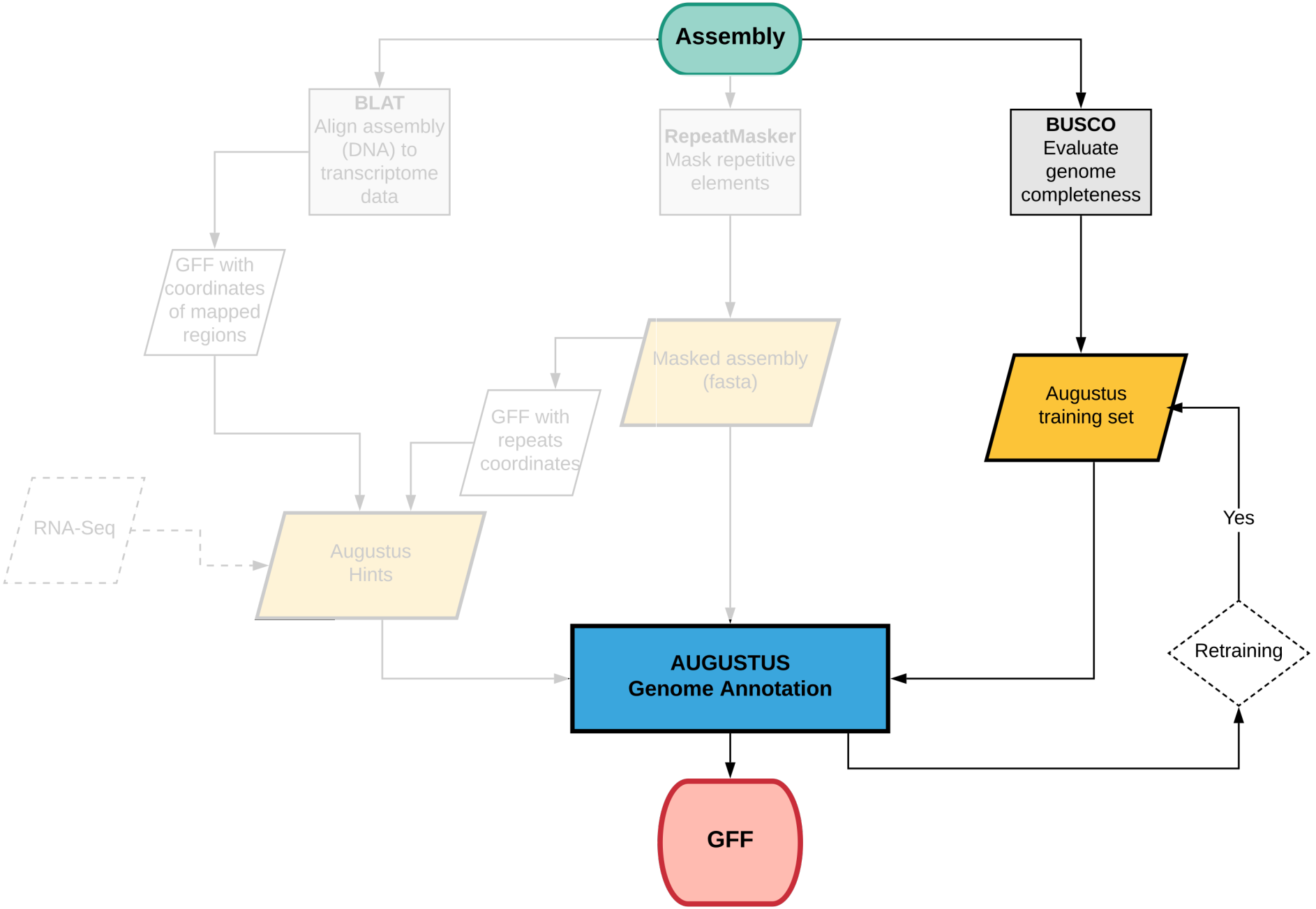
```
GAworkshop
    |__ assembly
    |__ augustus
          |__ config
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repeatmasker
```

1. Download the transcriptome of *Drosophila melanogaster* from Genbank. Extract the file.
2. Create the BLAT job
3. Submit the job

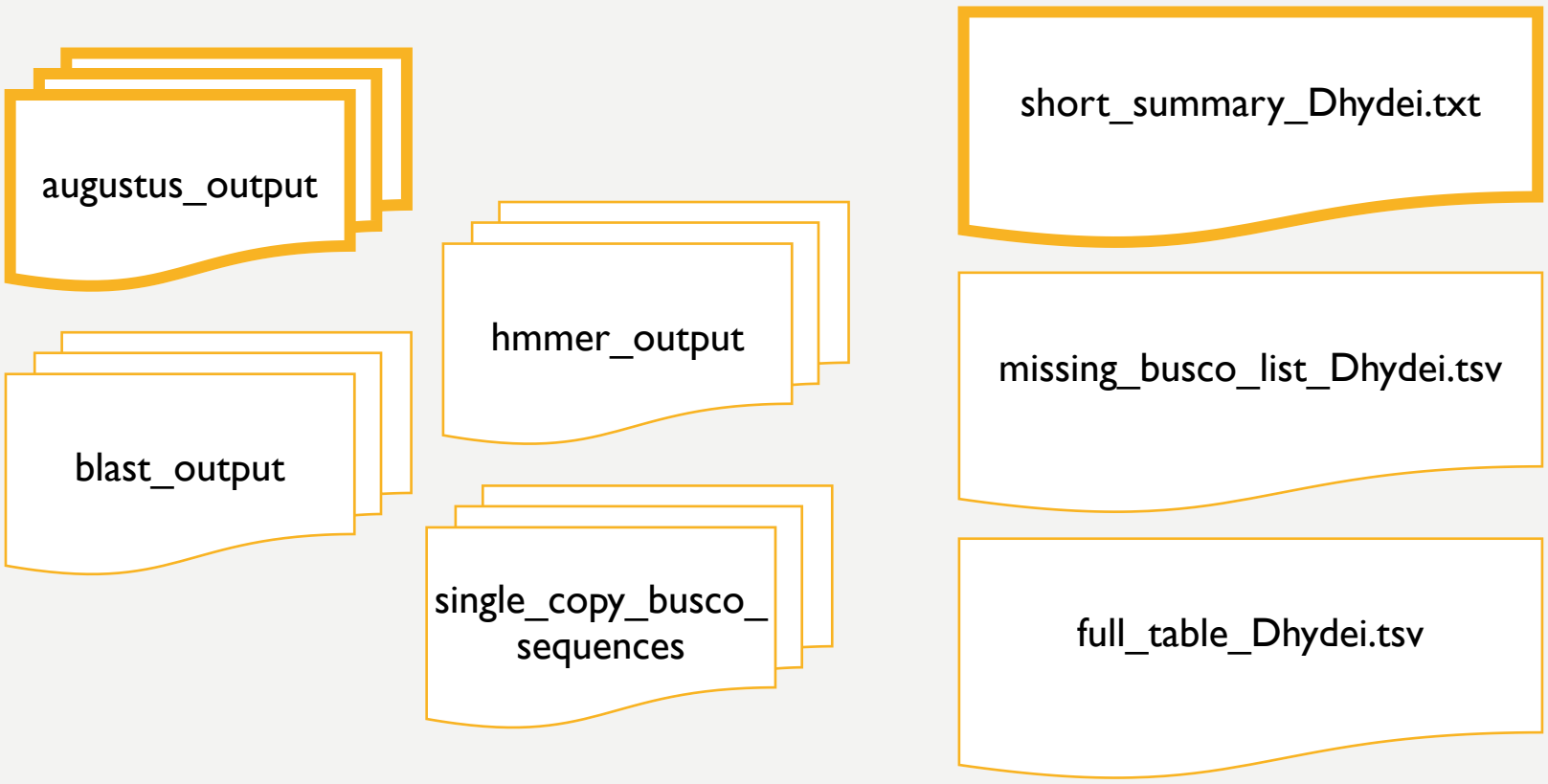


BACK TO BUSCO

STEP 5A


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

BUSCO OUTPUT: RUN_DHYDEI

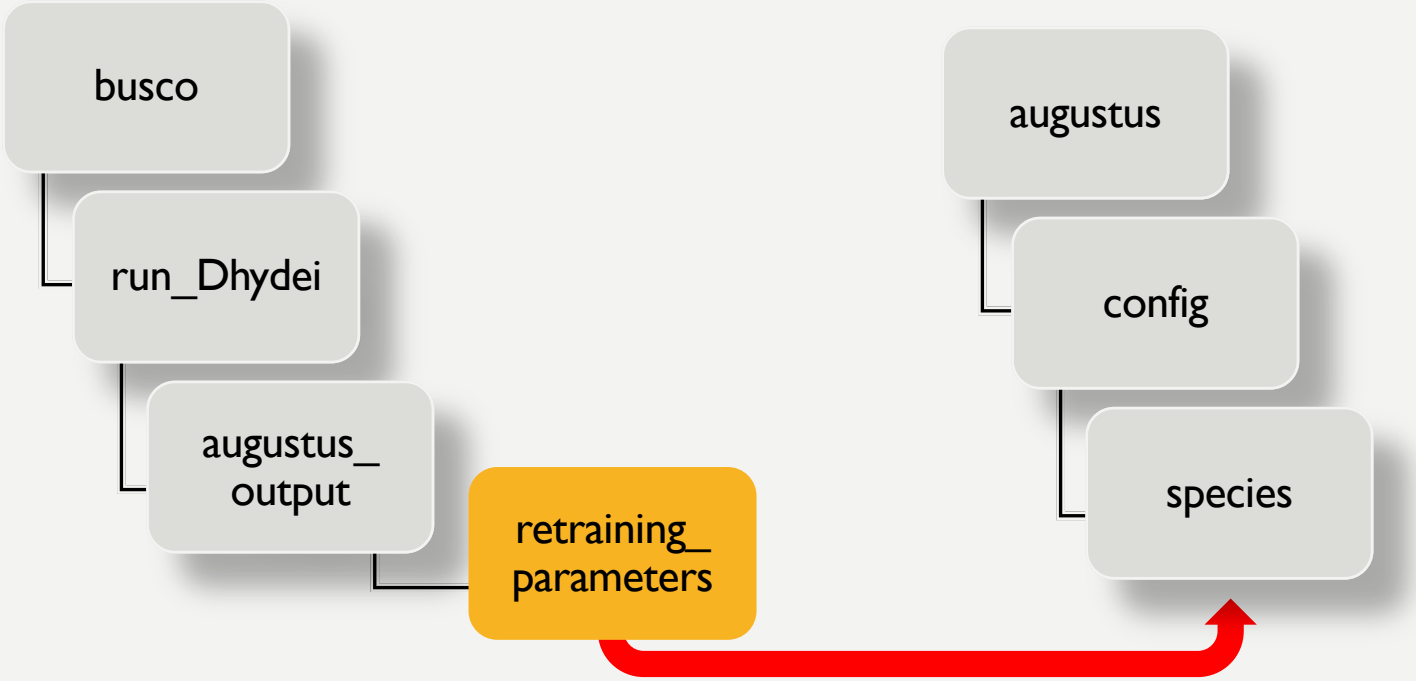


BUSCO - RESULTS

- Check the results of your BUSCO run in the short_summary_Dhydei.txt
  - How many Complete (C), Duplicated (D), Fragmented (F) and Missing (M)?
  - Do you think this is a good or a bad assembly?

FROM BUSCO TO AUGUSTUS

1. Copy the folder retraining_parameters to augustus/config/species



2. Rename the folder retraining_parameters using the prefix that appears in all files.

CREATING HINTS FOR AUGUSTUS

AUGUSTUS: FOLDER STRUCTURE

- Create the following folders in your augustus directory:
  - hints
  - output
  - scaffolds

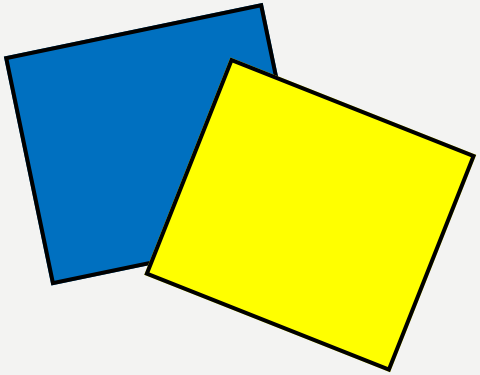
- The command ls should return the following result:

  config

  hints

  output

  scaffolds



IMPORTANT QUESTIONS

- What sources of extrinsic evidence do we have?
- How do those files look like?

We need to convert the hints into a format that augustus can read

GFF

- General Feature Format. Fields:
  - **seqname**
  - **source**
  - **feature**
  - **start**
  - **end**
  - **score**
  - **strand**
  - **frame**
  - **attribute**

GFF3 - REPEATMASKER

mtsuchiya — tsuchiyam@login-30-1:/scratch/genomics/tsuchiyam/RepeatMasker/RM_hints — ssh tsuchiyam@hydra-login01.si.edu — 139×42

```
-rw-rw-r-- 1 tsuchiyam tsuchiyam 5.5M Oct 10 15:37 siskin_Contig3141_pilon.fasta.masked
-rw-rw-r-- 1 tsuchiyam tsuchiyam 2.5K Oct 10 15:37 siskin_Contig3141_pilon.fasta.tbl
-rw-r--r-- 1 tsuchiyam tsuchiyam  20K Oct 10 15:37 GA01_repmask_soft.log
-rw-rw-r-- 1 tsuchiyam tsuchiyam    0 Oct 10 16:11 siskin_Contig3141_pilon.fasta.gff3
-rw-rw-r-- 1 tsuchiyam tsuchiyam 162K Oct 10 16:19 siskin_Contig3141_pilon.gff3
-rwxrwxr-x 1 tsuchiyam tsuchiyam  27K Oct 15 11:37 gff2hints.pl
-rw-rw-r-- 1 tsuchiyam tsuchiyam 162K Oct 15 11:41 test.gff3
[tsuchiyam@compute-8-31 repmasker]$ cat test.gff3
##gff-version 3
##sequence-region Contig3141_pilon 1 5638391
Contig3141_pilon        RepeatMasker    dispersed_repeat        604     627     17      +       .       Target=(CT)n 1 24
Contig3141_pilon        RepeatMasker    dispersed_repeat        819     856     26      +       .       Target=(T)n 1 38
Contig3141_pilon        RepeatMasker    dispersed_repeat        3372    3459    30      +       .       Target=(TGTT)n 1 89
Contig3141_pilon        RepeatMasker    dispersed_repeat        3852    3879    12      +       .       Target=(GCCT)n 1 28
Contig3141_pilon        RepeatMasker    dispersed_repeat        5845    6049    780     +       .       Target=CR1-X2 3923 4133
Contig3141_pilon        RepeatMasker    dispersed_repeat        6253    6290    18      +       .       Target=(ATT)n 1 38
Contig3141_pilon        RepeatMasker    dispersed_repeat        17123   17168   14      +       .       Target=G-rich 1 44
Contig3141_pilon        RepeatMasker    dispersed_repeat        23515   23627   459     +       .       Target=CR1-X1 4023 4133
Contig3141_pilon        RepeatMasker    dispersed_repeat        28192   28600   1488    -       .       Target=CR1-F2 4088 4497
Contig3141_pilon        RepeatMasker    dispersed_repeat        28926   28983   254     -       .       Target=UCON24 206 263
Contig3141_pilon        RepeatMasker    dispersed_repeat        30324   30384   20      +       .       Target=A-rich 1 59
Contig3141_pilon        RepeatMasker    dispersed_repeat        40281   40303   21      +       .       Target=(T)n 1 23
Contig3141_pilon        RepeatMasker    dispersed_repeat        42725   43050   969     -       .       Target=CR1-X2 3783 4098
Contig3141_pilon        RepeatMasker    dispersed_repeat        44280   44308   15      +       .       Target=(TCTTC)n 1 28
Contig3141_pilon        RepeatMasker    dispersed_repeat        44314   44343   32      +       .       Target=(T)n 1 30
Contig3141_pilon        RepeatMasker    dispersed_repeat        44531   44573   13      +       .       Target=(CTGCTG)n 1 46
Contig3141_pilon        RepeatMasker    dispersed_repeat        47527   47560   13      +       .       Target=(CCTCCC)n 1 33
Contig3141_pilon        RepeatMasker    dispersed_repeat        51189   51380   631     +       .       Target=CR1-H 4602 4798
Contig3141_pilon        RepeatMasker    dispersed_repeat        52831   52858   19      +       .       Target=(AACA)n 1 27
Contig3141_pilon        RepeatMasker    dispersed_repeat        57134   57152   15      +       .       Target=(T)n 1 19
Contig3141_pilon        RepeatMasker    dispersed_repeat        60288   60487   1008    +       .       Target=CR1-C4 4289 4508
Contig3141_pilon        RepeatMasker    dispersed_repeat        64723   64739   16      +       .       Target=(T)n 1 17
Contig3141_pilon        RepeatMasker    dispersed_repeat        64868   64896   15      +       .       Target=A-rich 1 29
Contig3141_pilon        RepeatMasker    dispersed_repeat        65872   65906   12      +       .       Target=(CTTTA)n 1 34
Contig3141_pilon        RepeatMasker    dispersed_repeat        72051   72093   47      +       .       Target=(T)n 1 43
Contig3141_pilon        RepeatMasker    dispersed_repeat        72913   73192   47      +       .       Target=(GT)n 1 276
Contig3141_pilon        RepeatMasker    dispersed_repeat        76071   76299   429     -       .       Target=GGLTR8B 2 234
Contig3141_pilon        RepeatMasker    dispersed_repeat        78999   79042   38      +       .       Target=(A)n 1 44
Contig3141_pilon        RepeatMasker    dispersed_repeat        80244   80288   28      +       .       Target=(A)n 1 45
Contig3141_pilon        RepeatMasker    dispersed_repeat        80590   80640   35      +       .       Target=(A)n 1 51
Contig3141_pilon        RepeatMasker    dispersed_repeat        84211   84687   1021    +       .       Target=CR1-C4 3956 4516
Contig3141_pilon        RepeatMasker    dispersed_repeat        90620   90684   235     +       .       Target=Chompy-2_Croc 6 72
```

GFF3 - BLAT

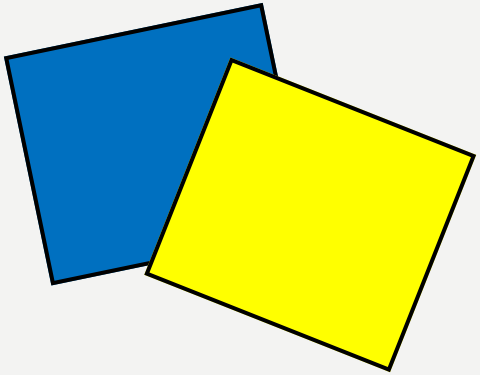
mtsuchiya — tsuchiyam@login-30-1:/scratch/genomics/tsuchiyam/RepeatMasker/RM_hints — ssh tsuchiyam@hydra-login01.si.edu — 144×44

```
Contig3141_pilon        b2h     ep      617     623     0       .       .       grp=XR_003076074.1;pri=4;src=E
Contig3141_pilon        b2h     ep      617     623     0       .       .       grp=XR_003076075.1;pri=4;src=E
Contig3141_pilon        b2h     ep      617     623     0       .       .       grp=XR_003076077.1;pri=4;src=E
Contig3141_pilon        b2h     ep      617     623     0       .       .       grp=XR_003076078.1;pri=4;src=E
Contig3141_pilon        b2h     ep      3406    3423    0       .       .       grp=XM_003641786.3;pri=4;src=E
Contig3141_pilon        b2h     ep      3409    3435    0       .       .       grp=XM_015291998.2;pri=4;src=E
Contig3141_pilon        b2h     ep      3409    3435    0       .       .       grp=XM_025153839.1;pri=4;src=E
Contig3141_pilon        b2h     ep      3409    3435    0       .       .       grp=XM_025153840.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=NM_001278026.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=NM_001278028.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_015296983.2;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_015296984.2;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_015296985.2;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_015296986.2;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_015296988.2;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_015296989.2;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_025142718.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_025142719.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_025142720.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_025142721.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_025142722.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_025142723.1;pri=4;src=E
Contig3141_pilon        b2h     ep      5916    5932    0       .       .       grp=XM_417523.6;pri=4;src=E
Contig3141_pilon        b2h     intron  28014   28314   0       .       .       mult=5;pri=4;src=E
Contig3141_pilon        b2h     ep      28316   28334   0       .       .       grp=XR_001469930.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28316   28334   0       .       .       grp=XR_003071673.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28336   0       .       .       grp=XR_001464165.2;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28340   0       .       .       grp=XR_003071393.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28340   0       .       .       grp=XR_003071394.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28341   0       .       .       grp=XR_003074136.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28345   0       .       .       grp=XR_003072657.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28348   0       .       .       grp=XR_003074120.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28335   28348   0       .       .       grp=XR_003073844.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28328   28349   0       .       .       grp=XR_003073482.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28328   28349   0       .       .       grp=XR_003073483.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28328   28349   0       .       .       grp=XR_003073484.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28315   28350   0       .       .       grp=XR_003074104.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28315   28350   0       .       .       grp=XR_003074105.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28315   28350   0       .       .       grp=XR_003074106.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28315   28350   0       .       .       grp=XR_003074107.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28315   28350   0       .       .       grp=XR_003074108.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28350   0       .       .       grp=XM_015286437.2;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28350   0       .       .       grp=XM_025148848.1;pri=4;src=E
Contig3141_pilon        b2h     ep      28325   28354   0       .       .       grp=XM_025149296.1;pri=4;src=E
```

AUGUSTUS HINTS: REPEATMASKER

```
GAworkshop
    |__ assembly
    |__ augustus
          |__ config
          |__ hints
          |__ output
          |__ scaffolds
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repeatmasker
```

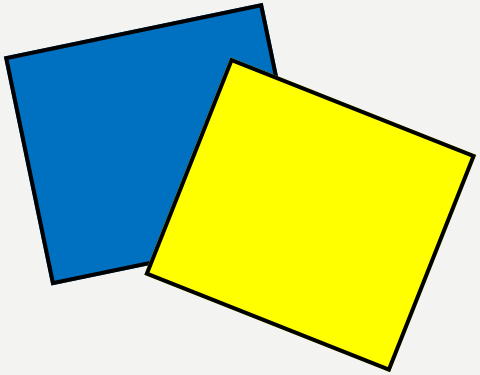
1. Load the module repeatmasker
2. Use the script rmOutToGFF3.pl to convert your .out file into GFF3
3. Use the script gff2hints to make the final conversion



AUGUSTUS HINTS: BLAT

```
GAworkshop
    |__ assembly
    |__ augustus
          |__ config
          |__ hints
          |__ output
          |__ scaffolds
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repeatmasker
```

1. Log to the interactive queue
2. Sort the .psl file
3. Load the augustus/3.3 module
4. Run the script blat2hints.pl

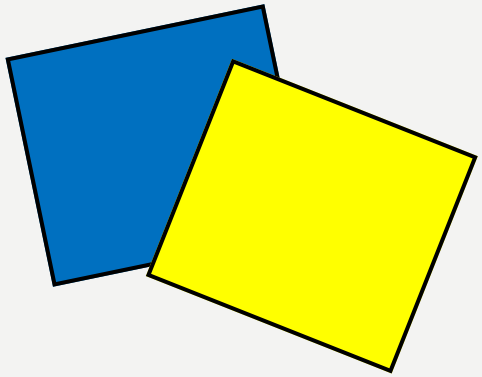


AUGUSTUS: COMBINING HINTS

```
GAworkshop
    |__ assembly
    |__ augustus
          |__ config
          |__ hints
          |__ output
          |__ scaffolds
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repeatmasker
```

- Merge both files:

cat Dhydei_RM_hints.out Dhydei_blat_hints.out | sort -k1,1 -k4,4n > Dhydei_hints_RM_E.gff3

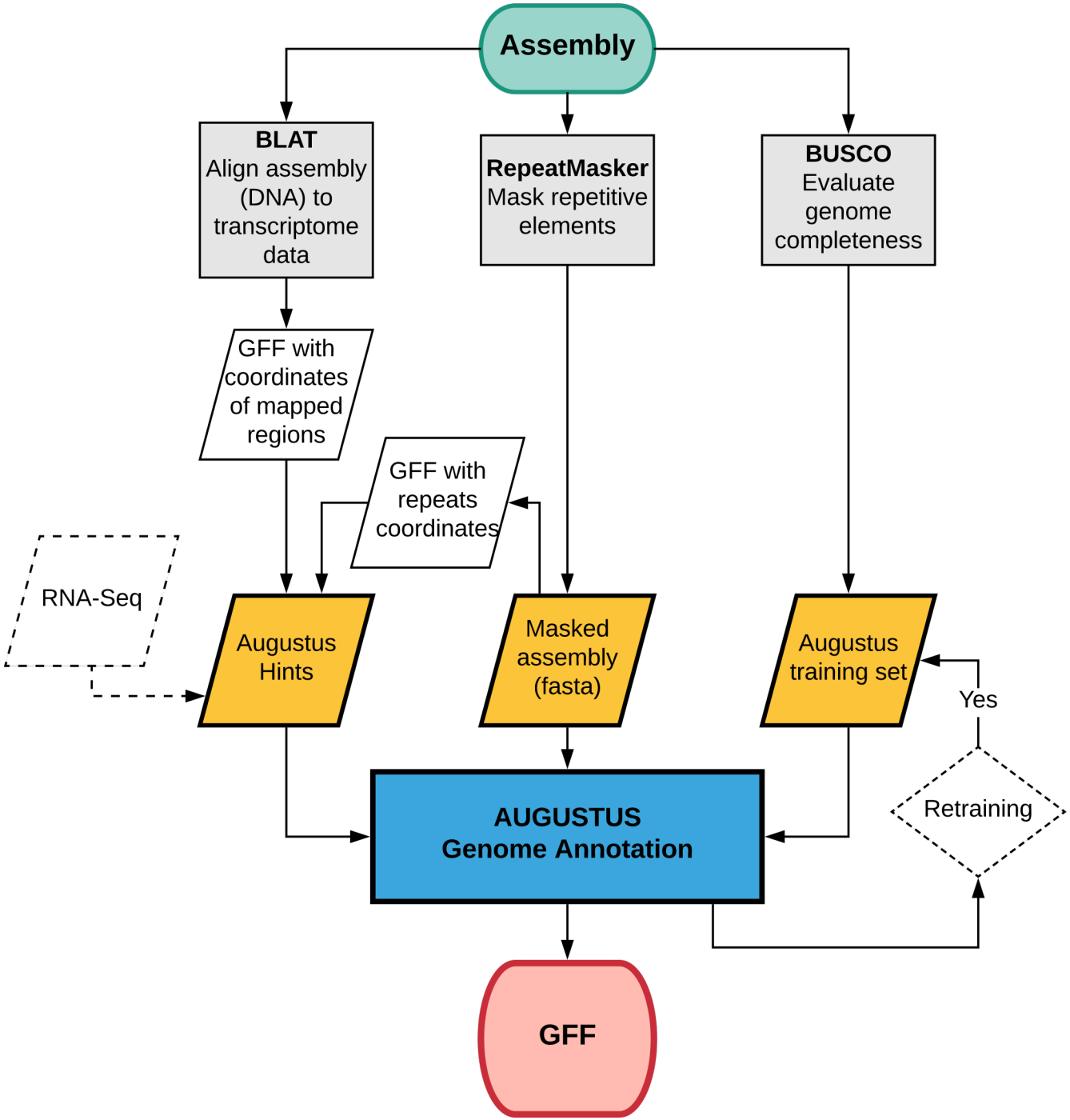


PHEW…

- What do we have now?
  - Masked fasta from RepeatMasker
  - Hints file
  - Training set from BUSCO
- What else do we need?

AUGUSTUS

FINALLY!


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

AUGUSTUS

- ab initio (internal) + evidence-driven(external)

AUGUSTUS is based on a generalized hidden Markov model (GHMM), which defines probability distributions for the various sections of genomic sequences. Introns, exons, intergenic regions, etc. correspond to states in the model and each state is thought to create DNA sequences with certain pre-defined emission probabilities.

AUGUSTUS

- Augustus needs to be trained:
  - Training consists on the generation of a training set that will more accurately predict genes.

  (BUSCO solved the training issue for us)

AUGUSTUS EXTRINSIC FILE

- Defines how the information from the hints will be weighted:

- It assigns "bonus" and "malus" (penalty) values to each hint used
  - M: manual annotation
  - W: RNA-Seq coverage information
  - E: EST/cDNA database hit
  - R: retroposed genes
  - RM: repeat masking

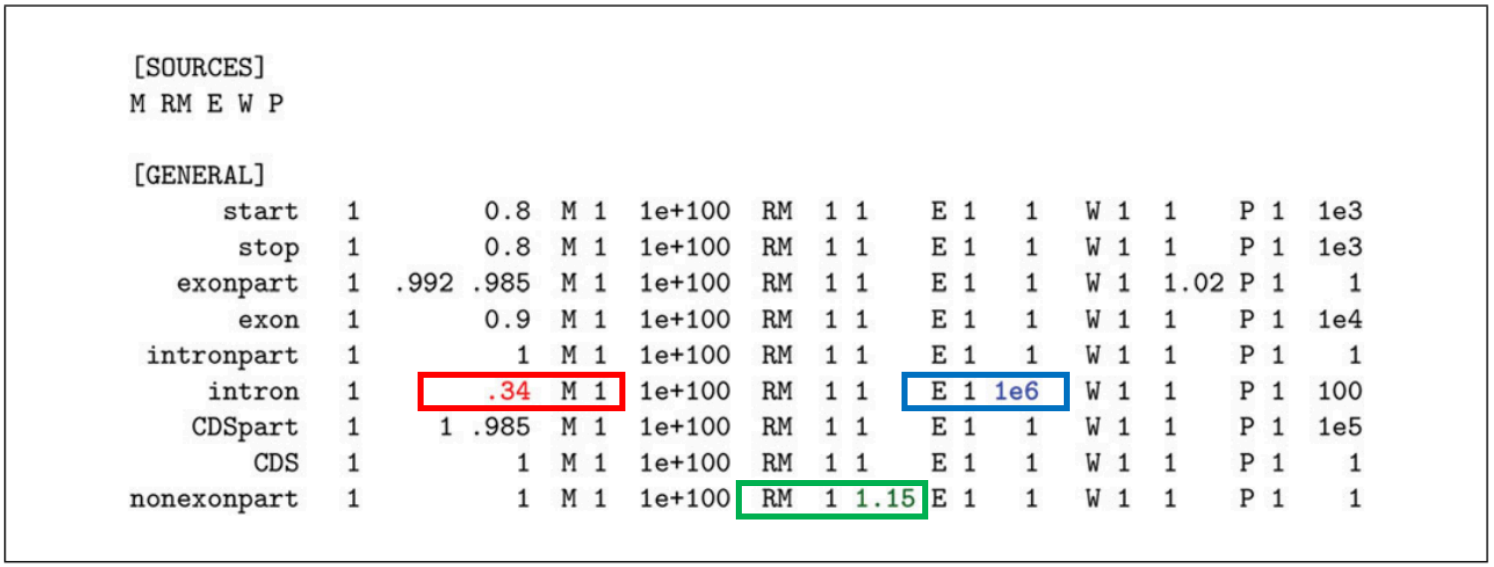
```
[SOURCES]
M RM E

#
# individual_liability: Only unsatisfiable hints are disregarded. By default this flag is not set
# and the whole hint group is disregarded when one hint in it is unsatisfiable.
# 1group1gene: Try to predict a single gene that covers all hints of a given group. This is relevant for
# hint groups with gaps, e.g. when two ESTs, say 5' and 3', from the same clone align nearby.
#
[SOURCE-PARAMETERS]


#   feature        bonus         malus   gradelevelcolumns
#               r+/r-
#
# the gradelevel colums have the following format for each source
# sourcecharacter numscoreclasses boundary    ...  boundary    gradequot  ...  gradequot
#

[GENERAL]
      start      1          1  M    1  1e+100  RM  1     1    E 1    1
       stop      1          1  M    1  1e+100  RM  1     1    E 1    1
        tss      1          1  M    1  1e+100  RM  1     1    E 1    1
        tts      1          1  M    1  1e+100  RM  1     1    E 1    1
        ass      1      1 0.1  M    1  1e+100  RM  1     1    E 1    1
        dss      1      1 0.1  M    1  1e+100  RM  1     1    E 1    1
   exonpart      1  .992 .985  M    1  1e+100  RM  1     1    E 1  1e2
       exon      1          1  M    1  1e+100  RM  1     1    E 1  1e4
 intronpart      1          1  M    1  1e+100  RM  1     1    E 1    1
     intron      1        .34  M    1  1e+100  RM  1     1    E 1  1e6
    CDSpart      1     1 .985  M    1  1e+100  RM  1     1    E 1    1
        CDS      1          1  M    1  1e+100  RM  1     1    E 1    1
    UTRpart      1     1 .985  M    1  1e+100  RM  1     1    E 1    1
        UTR      1          1  M    1  1e+100  RM  1     1    E 1    1
     irpart      1          1  M    1  1e+100  RM  1     1    E 1    1
nonexonpart      1          1  M    1  1e+100  RM  1     1.15 E 1    1
  genicpart      1          1  M    1  1e+100  RM  1     1    E 1    1

#
# Explanation: see original extrinsic.cfg file
```



```
[SOURCES]
M RM E W P

[GENERAL]
      start   1         0.8  M 1  1e+100  RM  1 1    E 1    1   W 1  1     P 1  1e3
       stop   1         0.8  M 1  1e+100  RM  1 1    E 1    1   W 1  1     P 1  1e3
   exonpart   1  .992 .985  M 1  1e+100  RM  1 1    E 1    1   W 1  1.02  P 1    1
       exon   1         0.9  M 1  1e+100  RM  1 1    E 1    1   W 1  1     P 1  1e4
 intronpart   1           1  M 1  1e+100  RM  1 1    E 1    1   W 1  1     P 1    1
     intron   1        .34   M 1  1e+100  RM  1 1    E 1 1e6    W 1  1     P 1  100
    CDSpart   1     1 .985   M 1  1e+100  RM  1 1    E 1    1   W 1  1     P 1  1e5
        CDS   1           1  M 1  1e+100  RM  1 1    E 1    1   W 1  1     P 1    1
nonexonpart   1           1  M 1  1e+100  RM  1 1.15 E 1    1   W 1  1     P 1    1
```

Figure 5 An excerpt of an extrinsic configuration file. In this example, each number to the right of the column filled with M's that is different from 1 specifies a *bonus*. A bonus is a relative factor that the un-normalized joint probability of gene structure candidate gets for being compatible with a hint of that type and source. For example, the blue *1e6* in the intron row after the source letter E means that for each intron hint with source tag E (src=E), gene structures that have an intron with both boundaries given as in the hint are rewarded by a factor of 10^6 relatively to gene structures disregarding the intron hint. A high bonus has the effect that many of the respective hints are respected by AUGUSTUS. The green 1.15 in the non-exonpart row after the tag RM (repeat masking) specifies that for each non exonpart hint, every gene structure gets a relative bonus factor of 1.15 *for each base* that is not an exon and not in a repeat. This discourages—but does not exclude—the overlap of exons and repeats. Repeat masking evidence can be given explicitly with hints of source RM, or implicitly with a soft-masked genome and the option *softmasking* turned on. The number(s) immediately to the left of the M column other than 1 specifies a penalty (malus) for gene structures with unsupported features. For example, the red .34 in the intron row means that every intron candidate that has no intron hints supporting it is penalized by multiplying its unnormalized probability with the factor 0.34. If you decrease this number even more (say from .3 to .001) then fewer introns unsupported by hints should be predicted. This would likely decrease the false positive intron rate, but, also, more true unsupported introns would be missed. For more information, see the file *config/extrinsic/extrinsic.cfg*.

EXTRINSIC FILE

- For practical purposes, copy the extrinsic file below to your augustus/config/extrinsic folder:

/data/genomics/workshops/GAworkshop/extrinsic.M.RM.E.cfg