TUTORIAL - TASKS

BUSCO: TASKS

```
genome_annot
    |__ assembly
    |__ augustus
    |__ blast
    |__ b2go
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repmasker
```

1. Copy the augustus config folder to YOUR augustus folder

```
cp -r /share/apps/bioinformatics/augustus/conda/3.3.2/config/ .
```

BUSCO: TASKS

```
genome_annot
    |__ assembly
    |__ augustus
          |__ config
    |__ blast
    |__ b2go
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repmasker
```

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

TASKS

- Figure out which species/taxon to use for siskin using `Repbase_RepeatQuery_Taxonomy_MTNT.sh`
  - Hint: siskin genus is *Spinus*. Compare to a more common bird species.

- Create the repeatmasker job file in the **jobs** folder following the info from the Github guide.

PARAMETERS

RepeatMasker

-species chicken: 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 ../repmasker/ : output the results to the specified folder

../assembly/siskin_10largest.fasta: input file

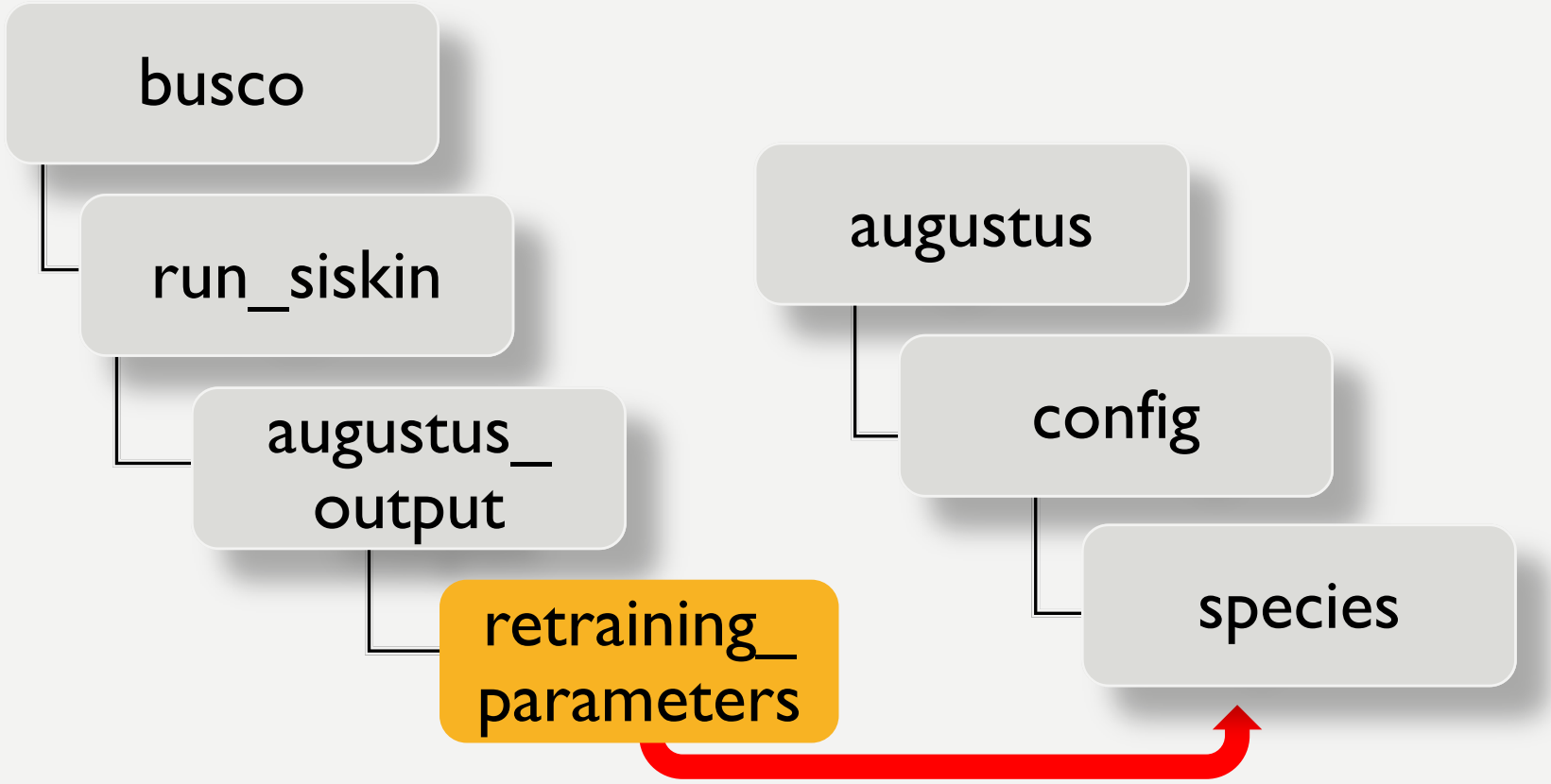
BLAT: TASKS

```
genome_annot
    |__ assembly
    |__ augustus
    |__ blast
    |__ b2go
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repmasker
```

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

FROM BUSCO TO AUGUSTUS

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



FROM BUSCO TO AUGUSTUS

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

You can find this info by looking at the file prefix inside the folder. In this case, we will rename the folder BUSCO_siskin_3415293029

AUGUSTUS: FOLDER STRUCTURE

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

AUGUSTUS: FOLDER STRUCTURE

- 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

```
[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: BLAT

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

```
genome_annot
    |__ assembly
    |__ augustus
          |__ config
          |__ hints
          |__ output
          |__ scaffolds
    |__ blast
    |__ b2go
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repmasker
```

AUGUSTUS HINTS: REPEATMASKER

```
genome_annot
    |__ assembly
    |__ augustus
          |__ config
          |__ hints
          |__ output
          |__ scaffolds
    |__ blast
    |__ b2go
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repmasker
```

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: COMBINING HINTS

```
genome_annot
    |__ assembly
    |__ augustus
          |__ config
          |__ hints
          |__ output
          |__ scaffolds
    |__ blast
    |__ b2go
    |__ blat
    |__ busco
    |__ jbrowse
    |__ jobs
    |__ logs
    |__ repmasker
```

- Merge both files:

```
cat siskin_RM_hints.out
siskin_blat_hints.out | sort -k1,1V -
k4,4n > siskin_hints_RM_E.gff3
```