Assessment of assemblies

Mirian T. N. Tsuchiya

October 11, 2018

BUSCO

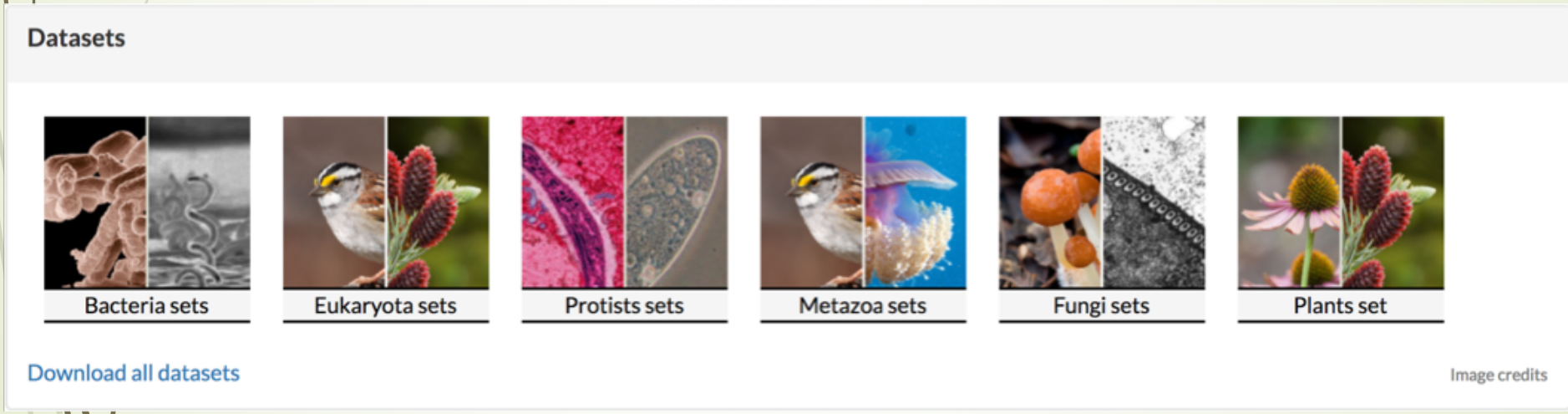
- **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

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



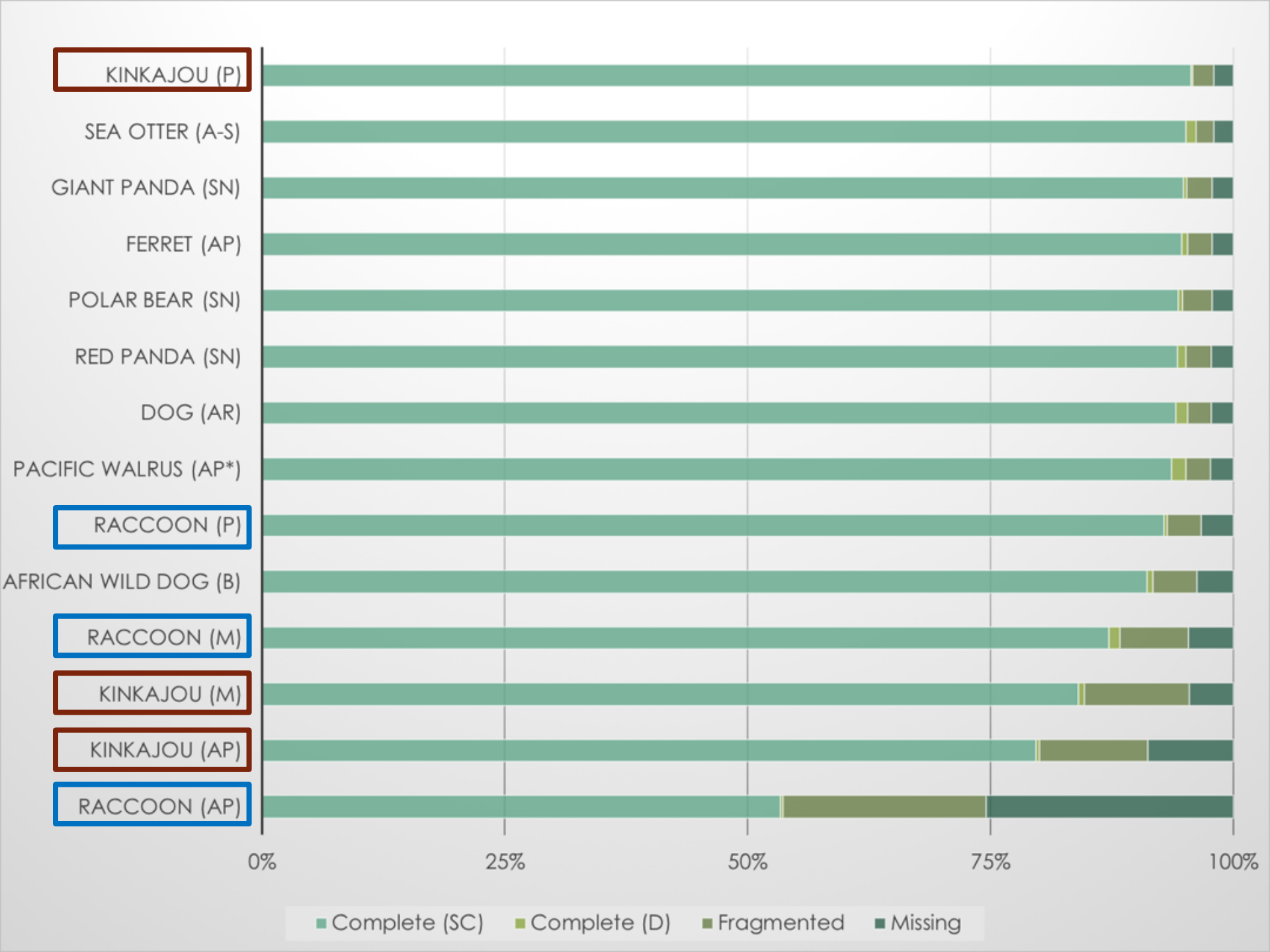
BUSCO: How complete is the assembly?

- 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 – pre-tutorial prep

- Create the following folders in your pool/genomics folder:
  - busco (**mkdir busco**)
  - augustus (**mkdir augustus && cd augustus**)
- Copy the augustus config folder to YOUR augustus folder

cp -r /share/apps/bioinformatics/augustus/gcc/4.9.2/3.3/config .

You should be in your augustus folder to do this.

BUSCO – tutorial

- Go to BUSCO website: https://buscos.ezlab.org/
- Find the best dataset for the species we're studying and download it.
  - Right-click in the selected dataset and copy link address
  - On terminal, **cd to the busco** folder you created and download it there.

    wget https://buscos.ezlab.org/datasets/aves_odb9.tar.gz

  - Extract the file using

    tar -xzf aves_odb9.tar.gz

Create a job file

- Job parameters:
  - Medium
  - mthread 4
  - 6 GB memory
  - Module: bioinformatics/busco

COMMAND:

```
export AUGUSTUS_CONFIG_PATH="/pool/genomics/user_id/augustus/config"
#
run_BUSCO.py --long -o siskin -i /data/genomics/dikowr/SMSC/finished_assembly/masurca_siskin.fasta -l ./aves_odb9 -c 1 -m genome
```

Command parameters

run_BUSCO.py

--long : Augustus optimization mode

-o siskin : output name

-i input.fasta : input fasta

-l ./aves_odb9 : lineage path

-c $NSLOTS : number of CPUs

-m genome : mode (in this case, genome)

```
# /bin/sh
# ----------------Parameters---------------------- #
#$ -S /bin/sh
#$ -pe mthread 4
#$ -q mThC.q
#$ -l mres=6G,h_data=6G,h_vmem=6G
#$ -cwd
#$ -j y
#$ -N GA02_busco
#$ -o GA02_busco.log
#
# ----------------Modules------------------------- #
module load bioinformatics/busco/3.0
#
# ----------------Your Commands------------------- #
#
echo + `date` job $JOB_NAME started in $QUEUE with jobID=$JOB_ID on $HOSTNAME
echo + NSLOTS = $NSLOTS
#
export AUGUSTUS_CONFIG_PATH="/scratch/genomics/tsuchiyam/augustus/config"
#
run_BUSCO.py --long -o siskin -i /data/genomics/dikowr/SMSC/finished_assembly/
masurca_siskin.fasta -l ./aves_odb9 -c 1 -m genome
#
echo = `date` job $JOB_NAME done
```