Getting lost in the reeds - a journey through the ordinal genomes

Tom Brown

Leibniz Institute for Zoo and Wildlife Research
IN THE FORSCHUNGSVERBUND BERLIN E.V.

BeGenDiv

ERGA
EUROPEAN REFERENCE GENOME ATLAS

Biodiversity Genomics Europe

VGP Phase 1 Data Freeze Manuscript

- Community effort to develop a roadmap for the creation high-quality reference genomes of representative species from every vertebrate order

- Figure 1: The taxonomic breadth covered, the amount/type of sequencing data generated, the number of assemblies at high standard

- **Figures 2-3: Demonstration of the quality of the generated assemblies and the variation across classes/orders/sequencing types/pipelines**

- Figures 4-5: Demonstration of what can now be investigated using long-read based, chromosome-scale assemblies across an entire sub-phylum and hundred of millions of years of evolution

VGP Phase 1 Data Freeze Manuscript

- Aim is to produce a summary manuscript detailing how *we* reached this point, the quality of the genomes and contextualising their use
- I will start quite anecdotal - dataset description and move a little towards some tentative conclusions or areas of further investigation
- I am hoping to spark some people to join us in the formation of the first draft
- Don’t like what you see and think you can do it better? Please do!!!

 # vgp-assembly-manuscript

genomeark.slack.com

Tom's naive planning

- Download all VGP genomes, extract metadata regarding datatype, software and versions, calculate statistics
- Demonstrate how assemblies have improved over time with switch from RSII -> CLR/ONT -> HiFi -> Haplotype-separated
- Show that the definition of a "high-quality" genome should be contextualised based on date, species, ploidy

Obtaining data and metadata

- Determining sequencing type has been non-trivial
- Methods and software are even more of a challenge

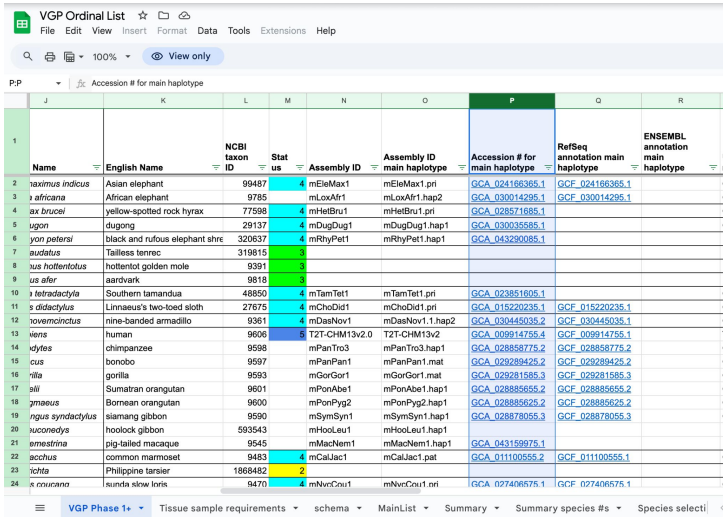
Scientific name	Ta…	Size…	Relea…	Sequencing technology
Scyliorhinus canicula (smaller sp…	7830	4,220	Dec, 2020	
Pungitius pungitius (ninespine sti…	134920	480.4	Apr, 2023	PacBio,Arima2
Cervus elaphus (red deer)	9860	2,887	Jul, 2021	
Anolis sagrei (Brown anole)	38937	1,951	Mar, 2024	PacBio Sequel II HiFi; Arima Hi-C v2
Melospiza melodia melodia	19149…	1,541	Jan, 2024	PacBio Sequel II HiFi; Bionano DLS…
Manis pentadactyla (Chinese pan…	143292	2,840	May, 2023	PacBio Sequel II HiFi; Dovetail Om…
Tachyglossus aculeatus (Australi…	9261	2,213	Dec, 2020	PacBio Sequel I CLR; Illumina Nova…
Scatophagus argus	75038	570.8	Oct, 2021	PacBio Sequel I CLR; Illumina Nova…
Acanthopagrus latus (yellowfin s…	8177	685.1	Oct, 2020	
Muntiacus reevesi (Reeves' muntj…	9886	2,656	Jan, 2024	PacBio,Arima2
Falco peregrinus (peregrine falcon)	8954	1,312	Jun, 2022	PacBio Sequel II HiFi; Arima Geno…
Thunnus thynnus (Atlantic bluefi…	8237	799.0	Jan, 2024	PacBio,Arima2
Eschrichtius robustus (grey whale)	9764	2,982	Jan, 2023	PacBio Sequel II HiFi; 3D-DNA Hi-C
Thunnus albacares (yellowfin tun…	8236	792.1	Oct, 2023	PacBio,Illumina,Arima

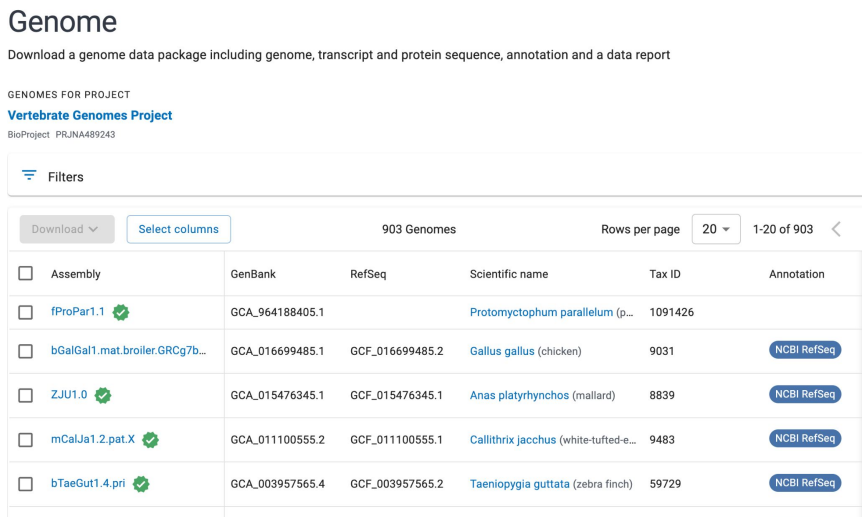
Assembly method various

```
if echo "$methods" | grep -q '[Hh]i[Ff]i'; then
    data='PacBio HiFi'
elif echo "$methods" | grep -q 'IPA'; then
    data='PacBio HiFi'
elif echo "$methods" | grep -q '[Hh]icanu'; then
    data='PacBio HiFi'
elif echo "$datatypes" | grep -q '[Hh]i[Ff]i'; then
    data='PacBio HiFi'
elif echo "$comments" | grep -q '[Hh]i[Ff]i'; then
    data='PacBio HiFi'
elif echo "$datatypes" | grep -q '[Oo]xford'; then
    data='Oxford Nanopore'
elif echo "$datatypes" | grep -q '[Nn]anopore'; then
    data='Oxford Nanopore'
elif echo "$datatypes" | grep -q 'ONT'; then
    data='Oxford Nanopore'
elif echo "$comments" | grep -q '[Oo]xford'; then
    data='Oxford Nanopore'
elif echo "$comments" | grep -q '[Nn]anopore'; then
    data='Oxford Nanopore'
elif echo "$comments" | grep -q 'ONT'; then
    data='Oxford Nanopore'
elif echo "$datatypes" | grep -q '[Cc][Ll][Rr]'; then
    data='PacBio CLR'
elif echo "$comments" | grep -q '[Cc][Ll][Rr]'; then
    data='PacBio CLR'
elif echo "$methods" | grep -q '[Cc][Ll][Rr]'; then
    data='PacBio CLR'
elif echo "$methods" | grep -q '[Ff]alcon'; then
    data='PacBio CLR'
elif echo "$methods" | grep -q 'FALCON'; then
    data='PacBio CLR'
elif echo "$comments" | grep -q '[Ff]alcon'; then
    data='PacBio CLR'
elif echo "$datatypes" | grep -q '[Rr][Ss][Ii][Ii]'; then
    data='PacBio RSII'
else
    data='missing'
    echo $gca
fi
```

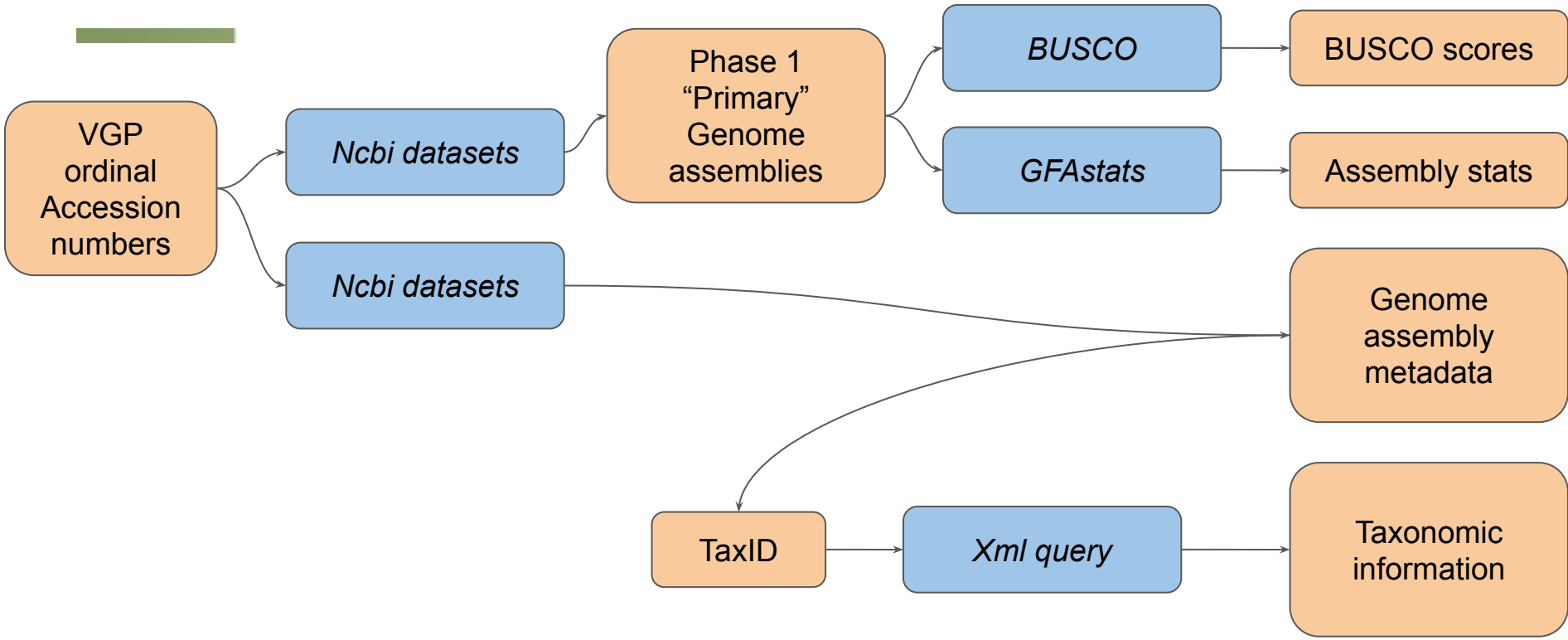
Obtaining data and metadata

- Very grateful to have Erich's google sheet to get a definitive list of which are the primary assemblies - Currently 903 assemblies under VGP BioProject, 415 marked as "reference"



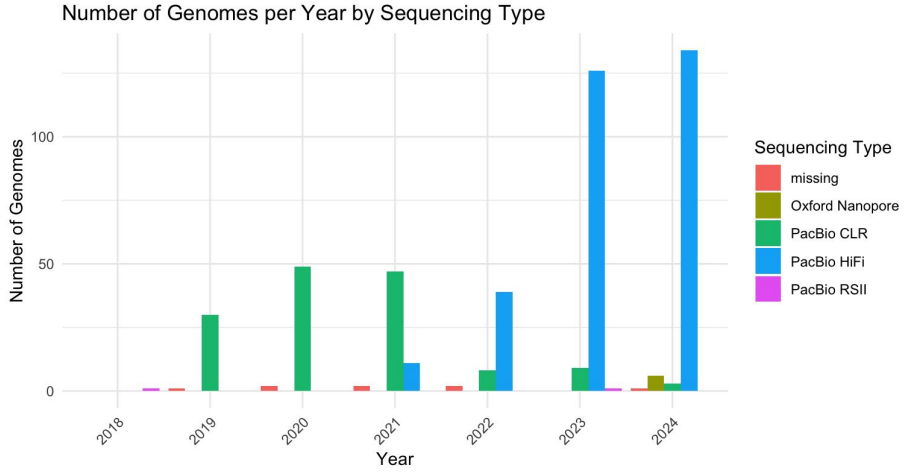


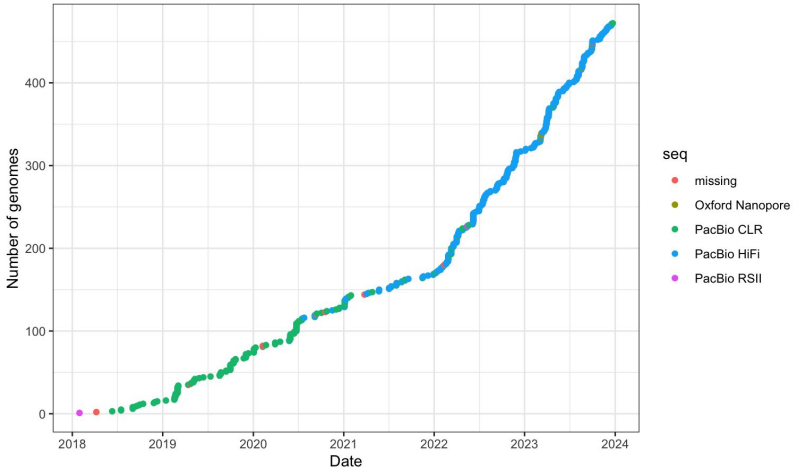
Obtaining data and metadata

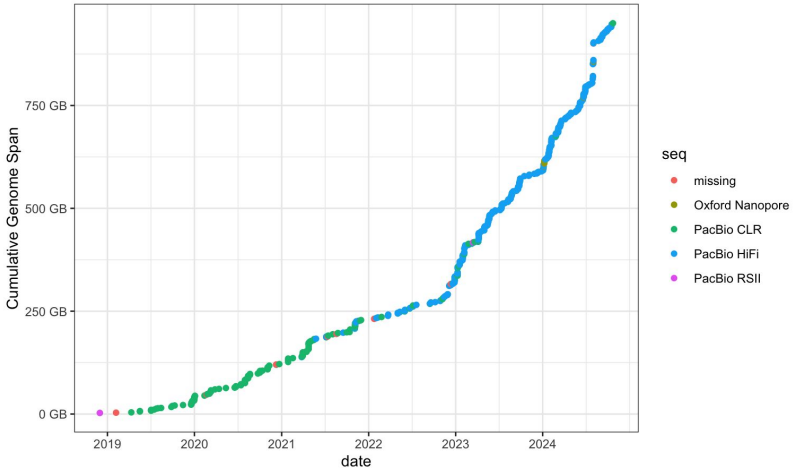


Phase1+ species removed (hagfish/lancelets/sea squirts)
A few "classes" have been filled in for species without them

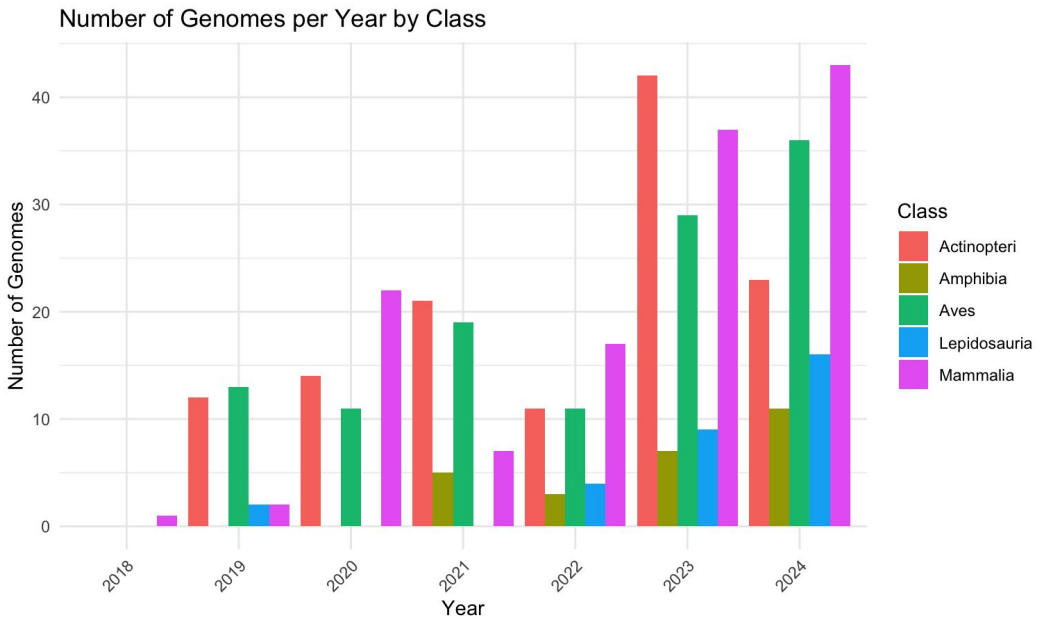
Rapid increase of HiFi genomes published over the last 2 years

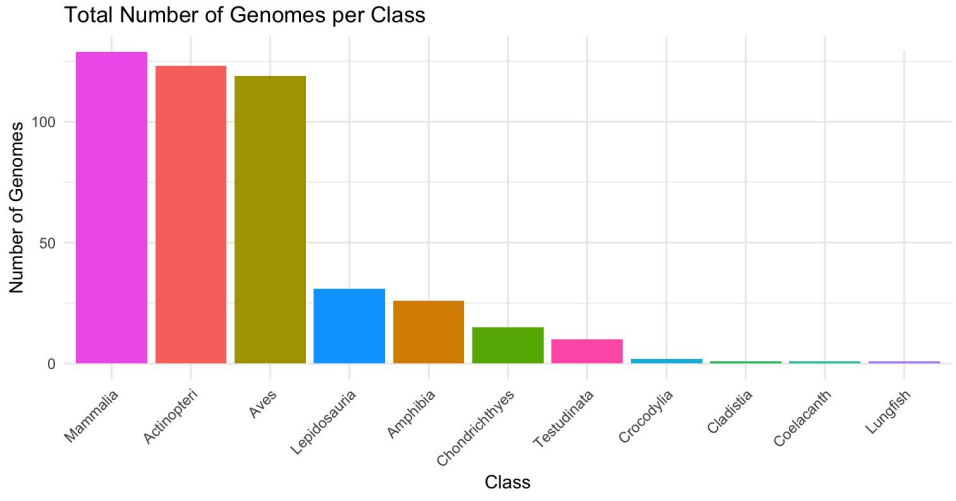




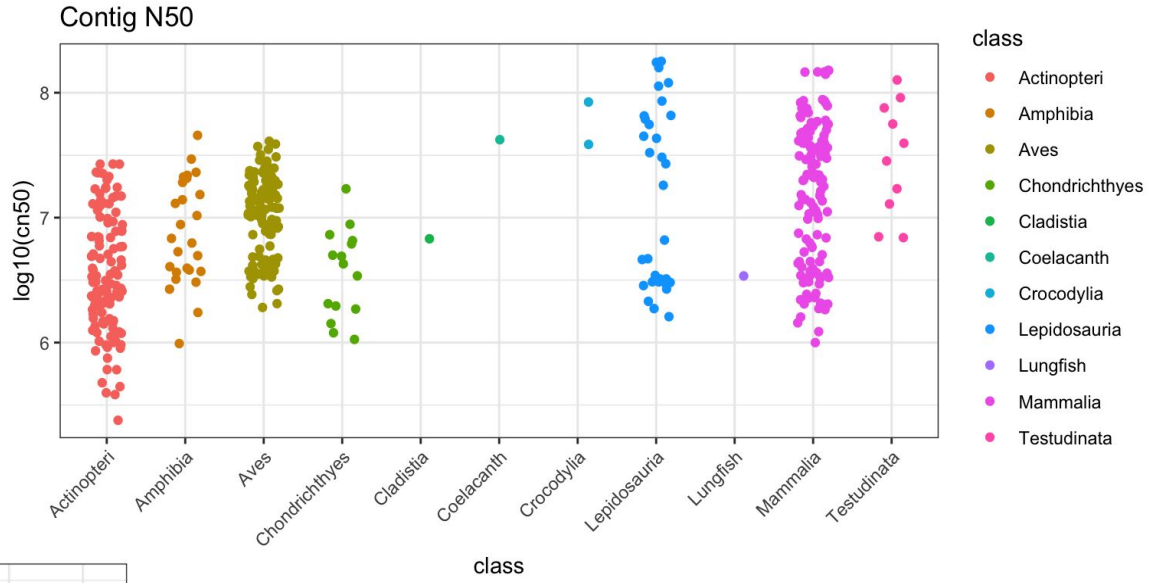


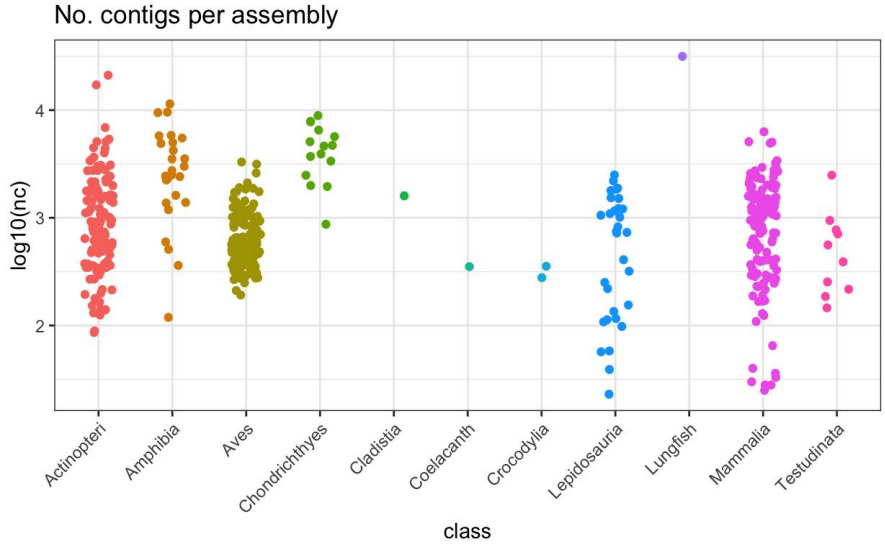
Genomes per “class”



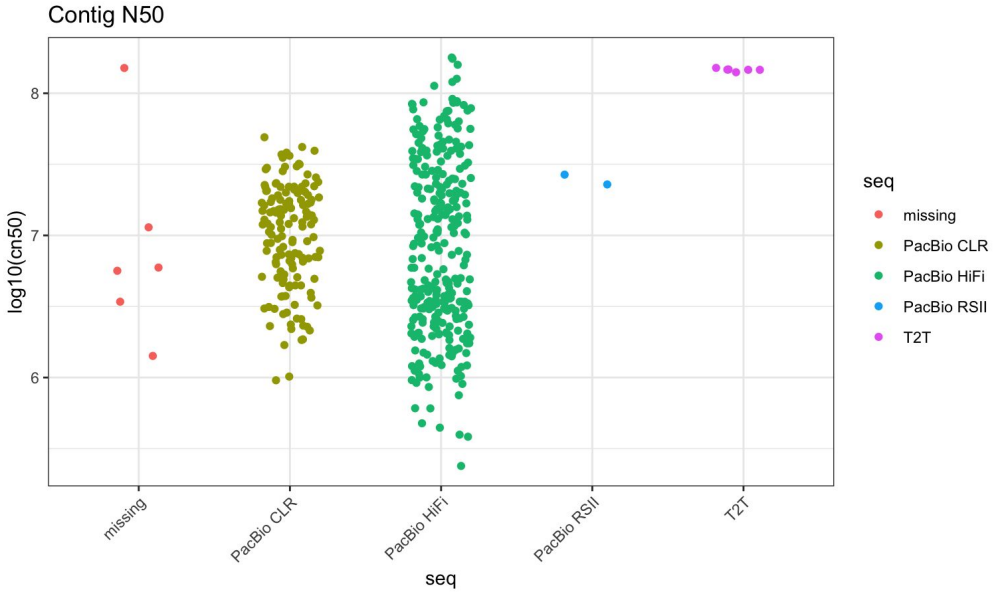


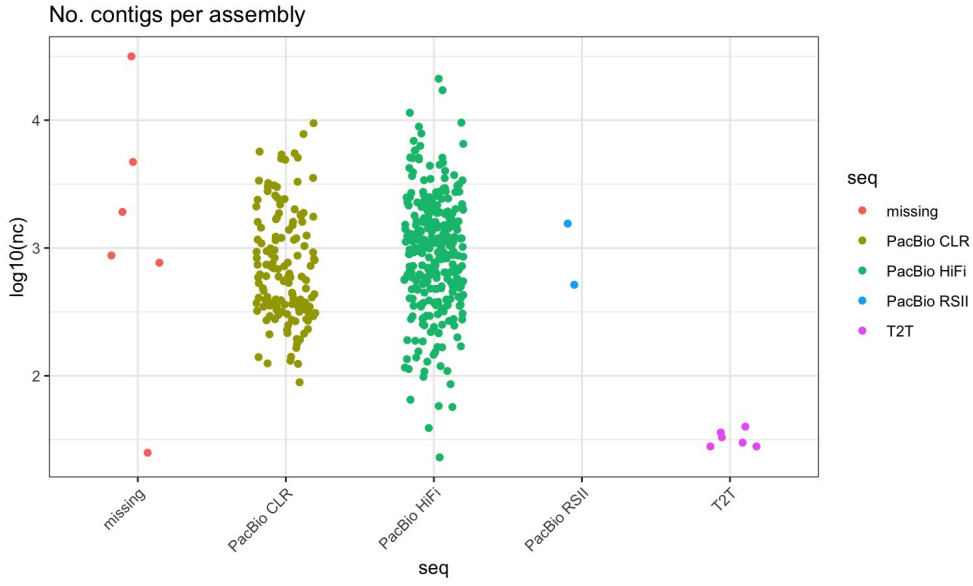
Contiguity by class



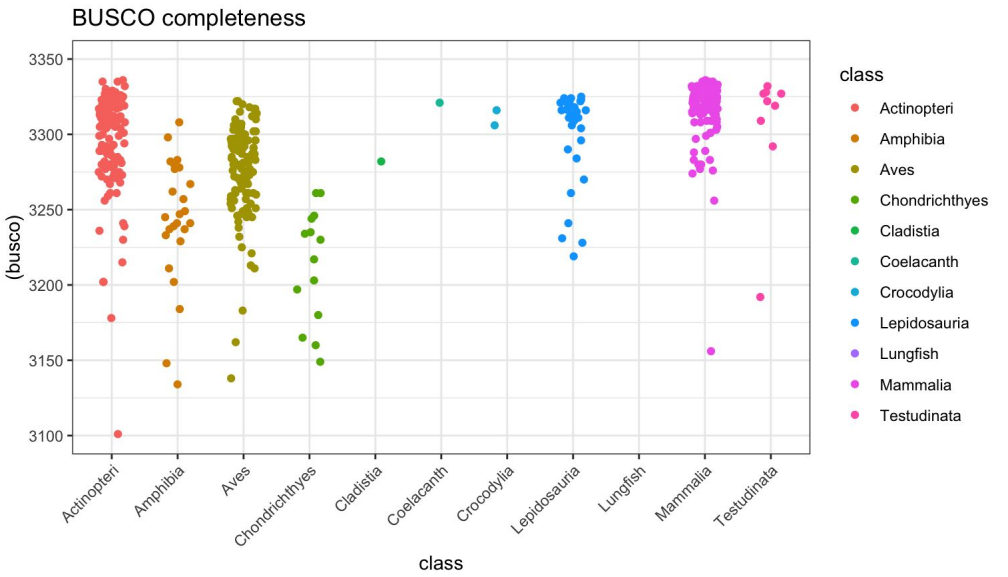


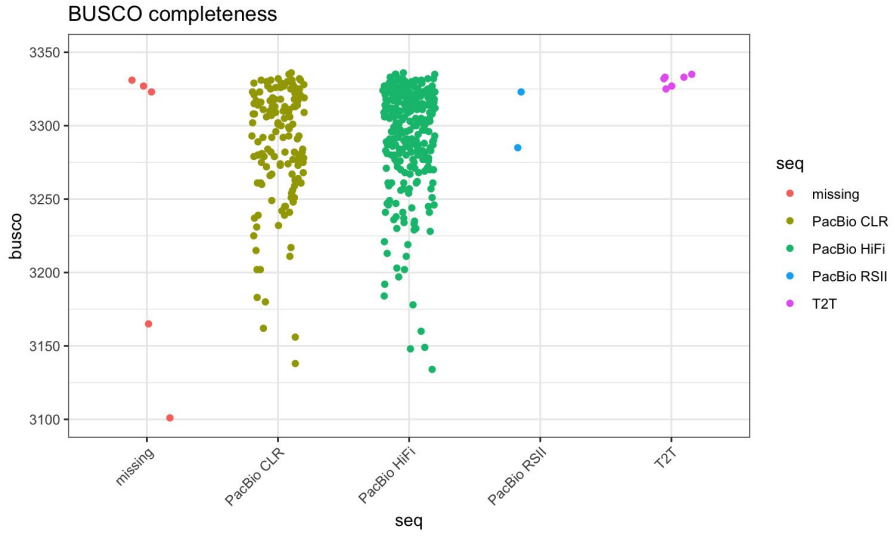
Contiguity by sequencing type

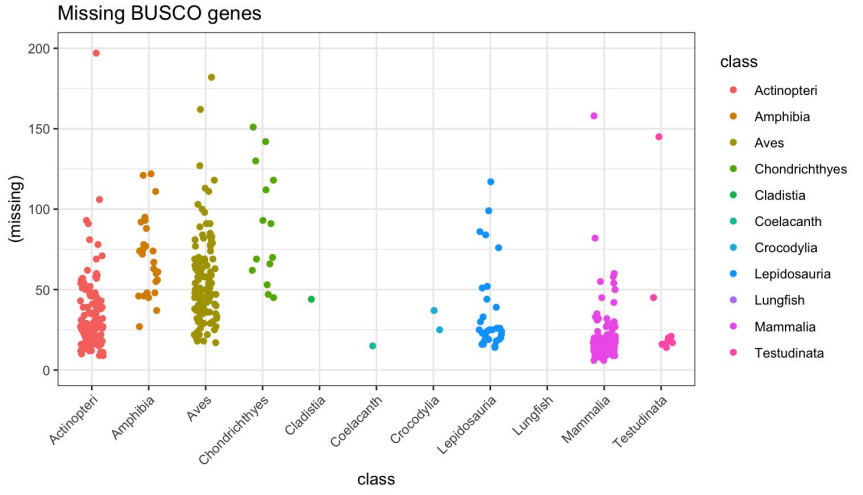




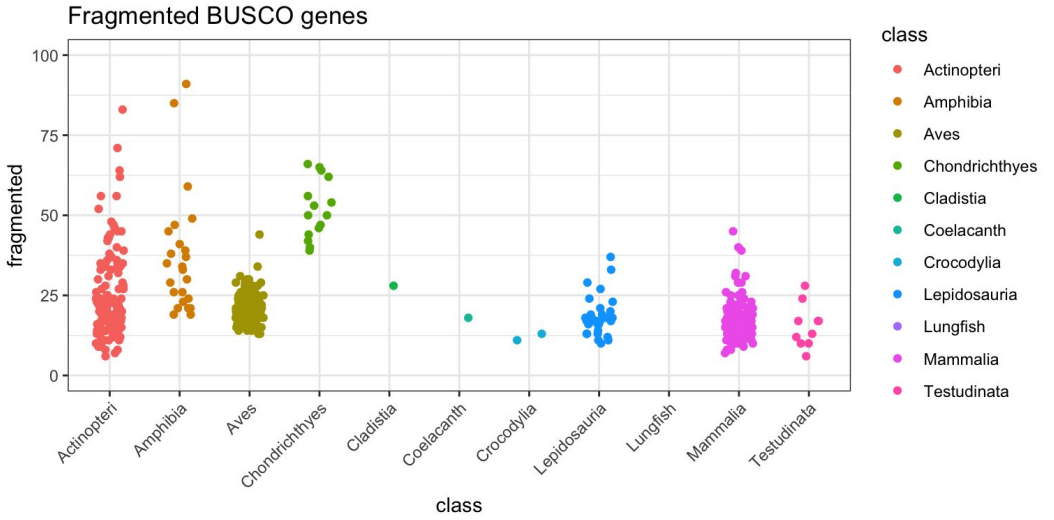
BUSCO completeness

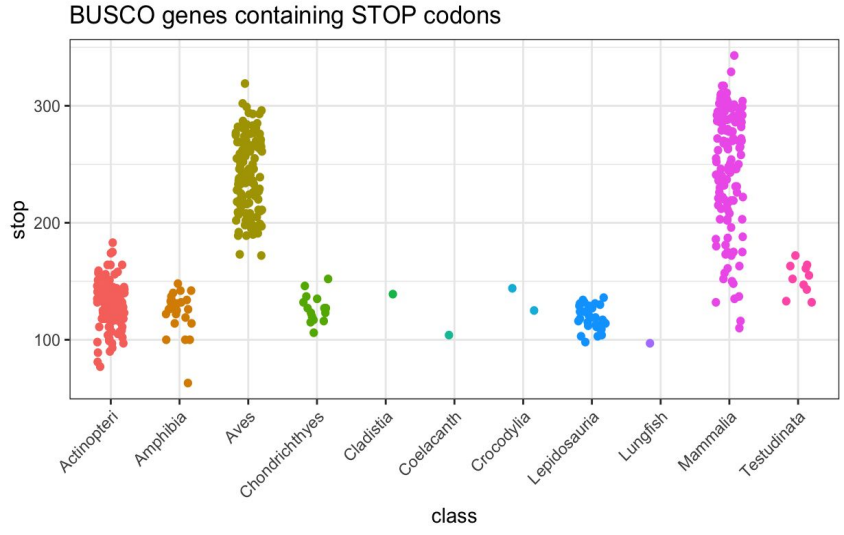


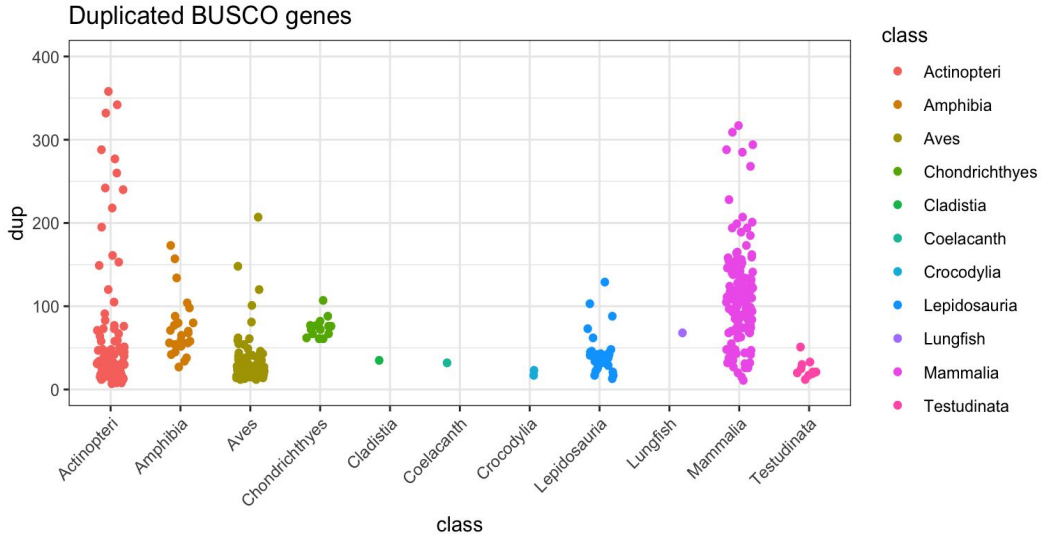




and more BUSCO

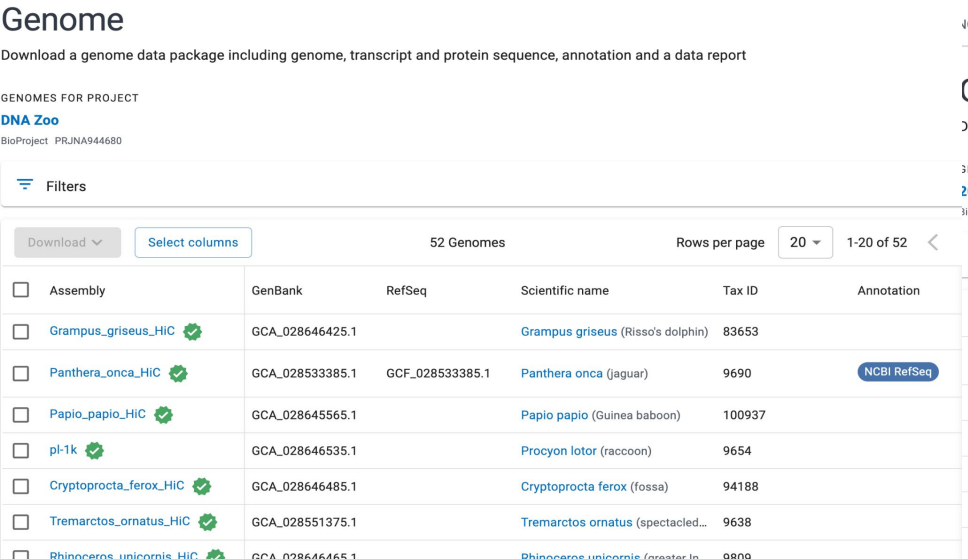


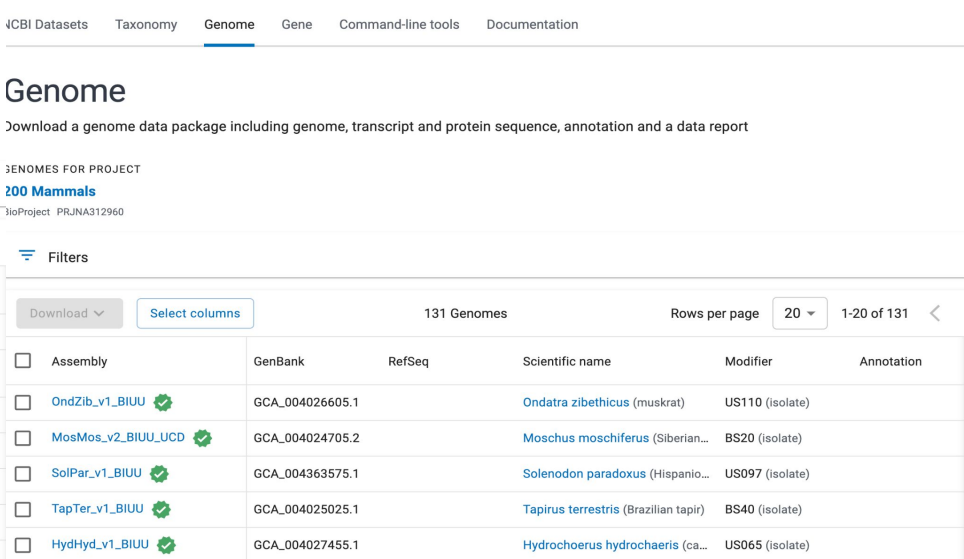




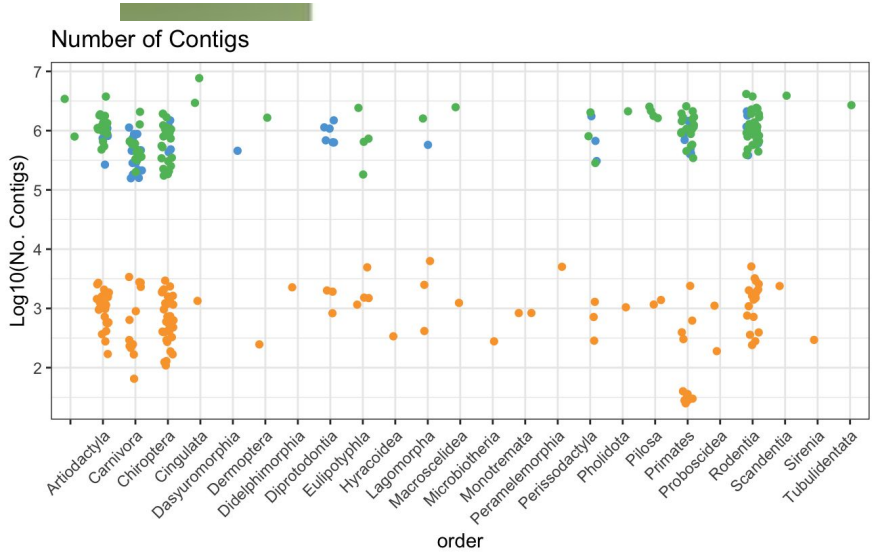
Comparing with other large vertebrate genome projects

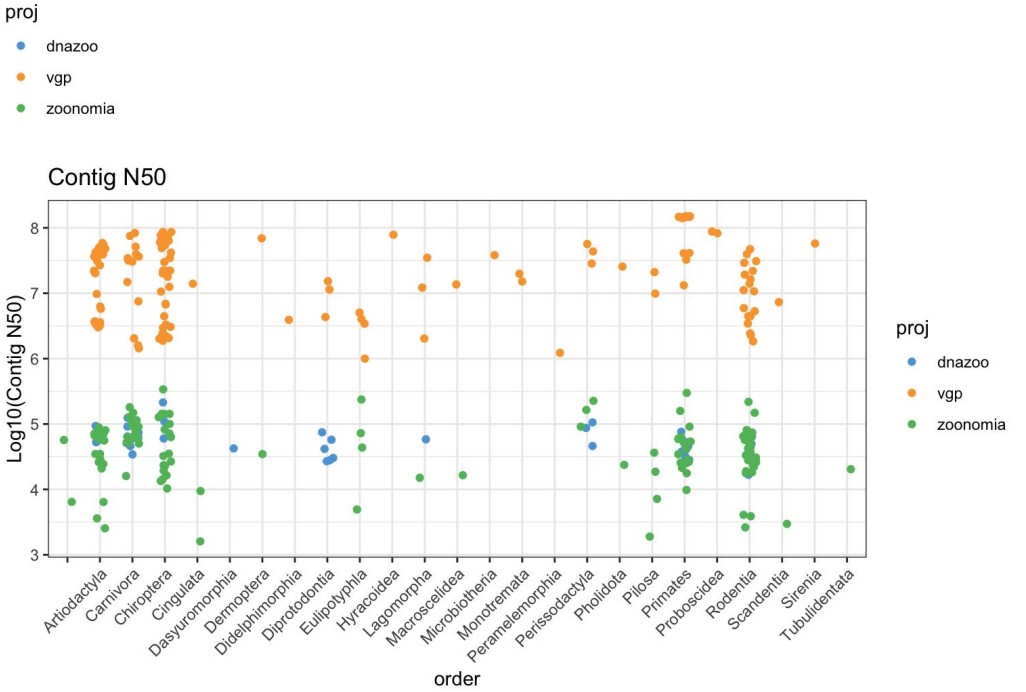
- Restricted to genomes published on GenBank - for DNAZoo this included only mammals



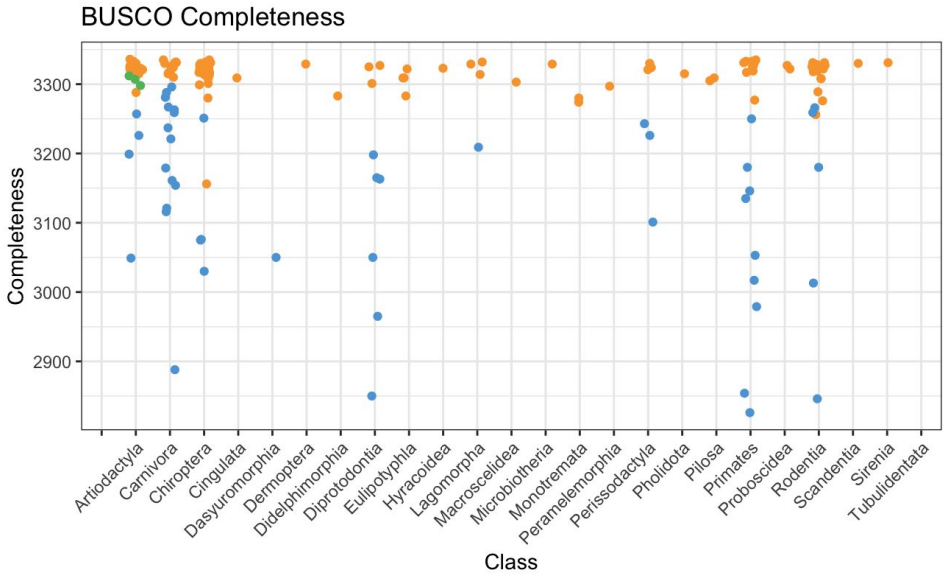


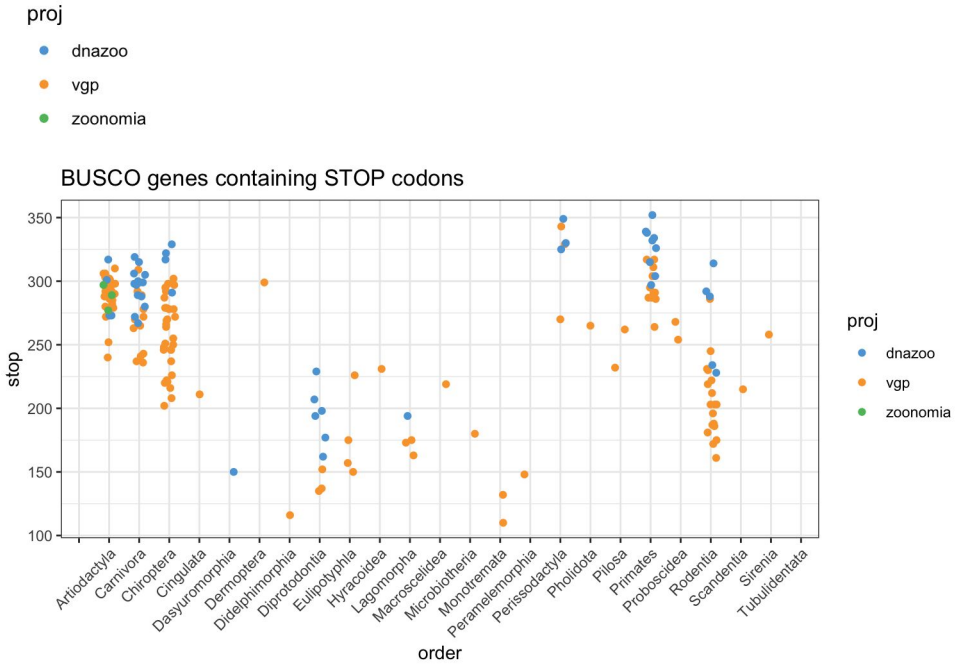
Long-read genomes more contiguous 🤯





… but also more complete and few stop-codon-containing BUSCO genes





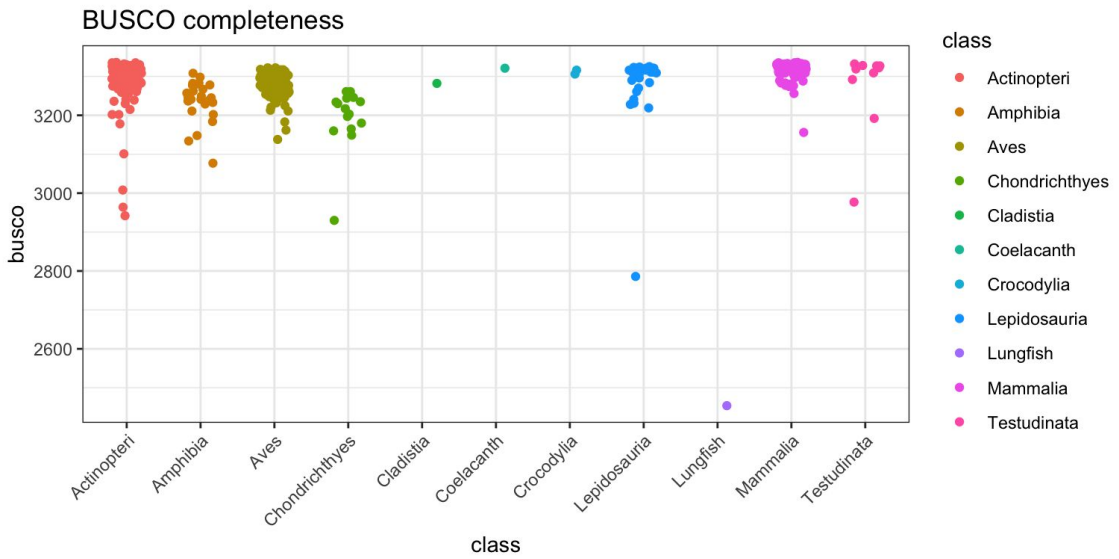
Initial observations

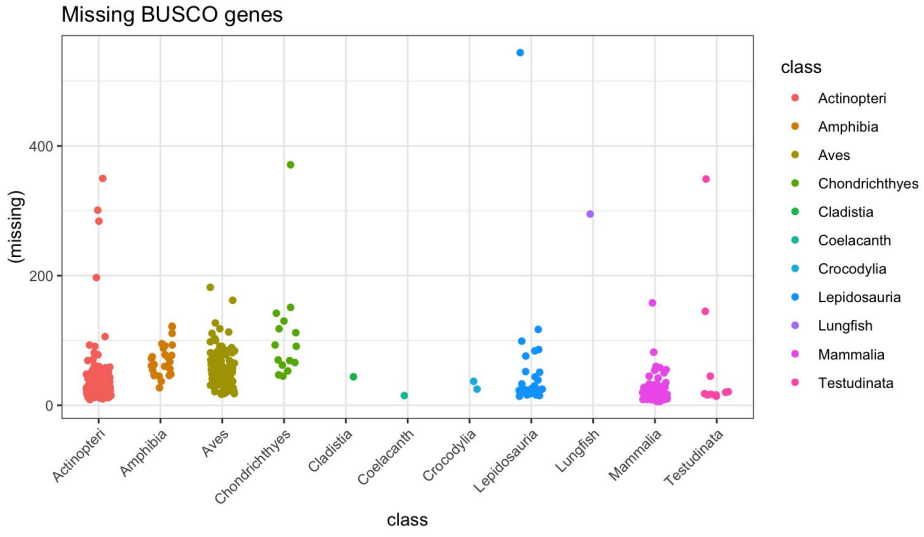
- At this level, superficial statistics not able to see a difference in CLR/HiFi (can we think about (compute) cost per genome?)
- Statistics fairly consistent apart from lower contiguity/completeness in Rays & Sharks?
- Elevated BUSCO genes with stop codons in mammals and birds
- Assemblies are more contiguous and complete on the gene level than DNAzoo/Zoonomia assemblies

Potential next steps

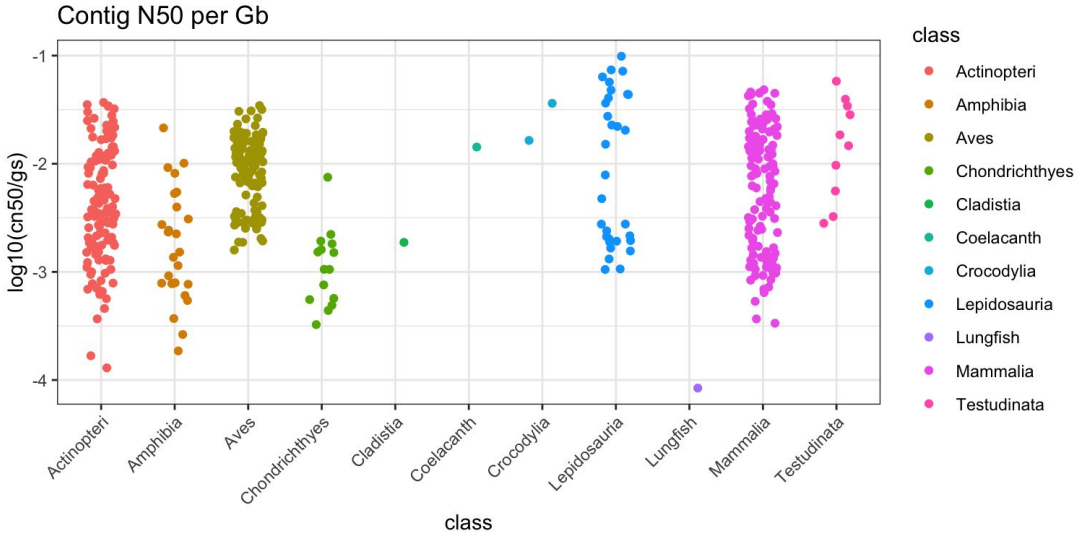
- Rerun BUSCO using the –metaeuk mapper to hopefully get a more complete list of scores
- Expand list of genomes to all vertebrate “reference” genomes (currently ~4,000)
- Make a list of class-specific reference genome metrics (not only BUSCO > 9X%, but should also contain *this set* of genes)
- Start thinking about ways to demonstrate the power of the fact that assemblies are chromosome-scale and curated

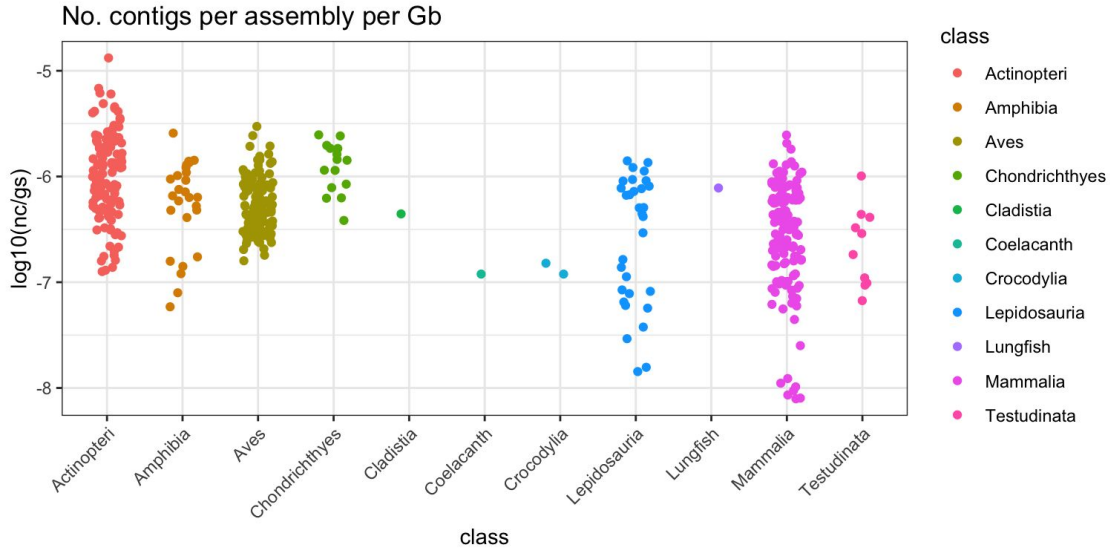
Appendix


BUSCO completeness
busco
3200
3000
2800
2600
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata
class
class
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata

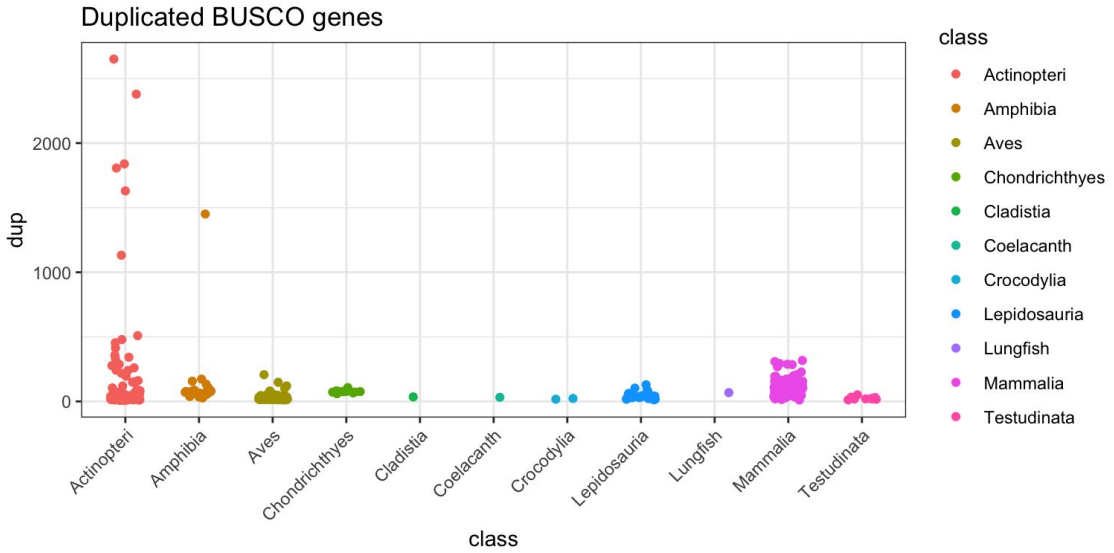

Missing BUSCO genes
(missing)
400
200
0
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata
class
class
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata

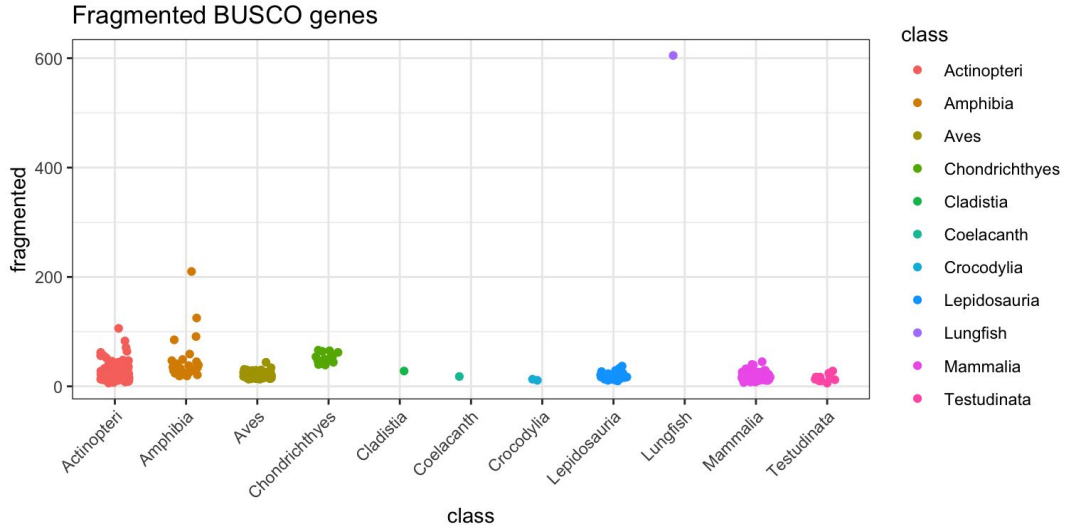
Appendix


Contig N50 per Gb
log10(cn50/gs)
-1
-2
-3
-4
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata
class
class
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata

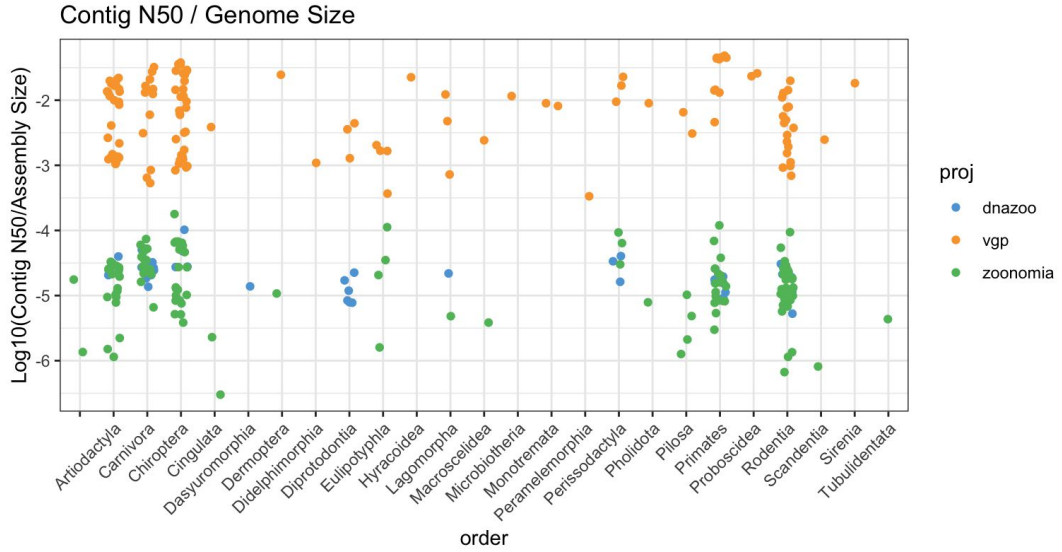

No. contigs per assembly per Gb
log10(nc/gs)
-5
-6
-7
-8
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata
class
class
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata

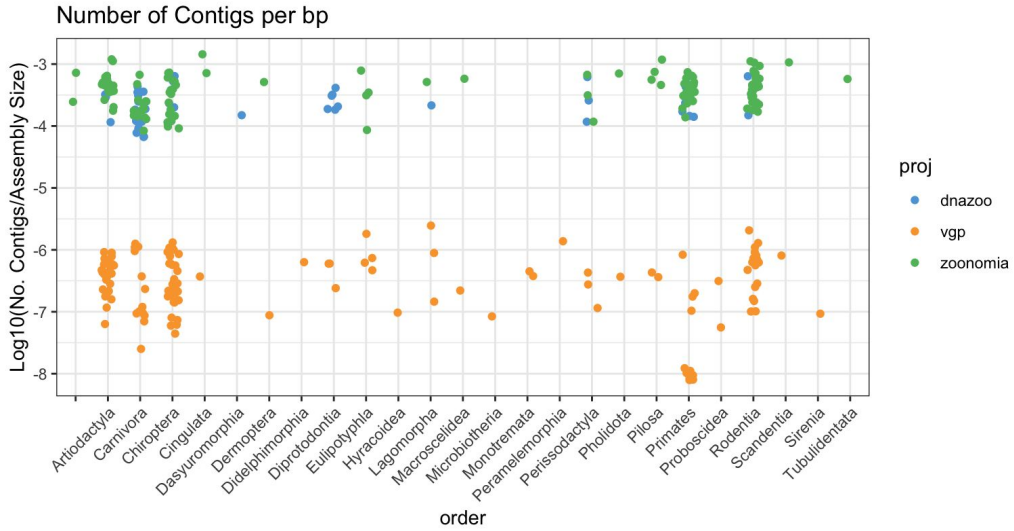
Appendix


Duplicated BUSCO genes
dup
2000
1000
0
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata
class
class
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata


Fragmented BUSCO genes
fragmented
600
400
200
0
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata
class
class
Actinopteri
Amphibia
Aves
Chondrichthyes
Cladistia
Coelacanth
Crocodylia
Lepidosauria
Lungfish
Mammalia
Testudinata

Appendix


Contig N50 / Genome Size
Log10(Contig N50/Assembly Size)
-2
-3
-4
-5
-6
Artiodactyla
Carnivora
Chiroptera
Cingulata
Dasyuromorphia
Dermoptera
Didelphimorphia
Diprotodontia
Eulipotyphla
Hyracoidea
Lagomorpha
Macroscelidea
Microbiotheria
Monotremata
Peramelemorphia
Perissodactyla
Pholidota
Pilosa
Primates
Proboscidea
Rodentia
Scandentia
Sirenia
Tubulidentata
order
proj
dnazoo
vgp
zoonomia


Number of Contigs per bp
Log10(No. Contigs/Assembly Size)
-3
-4
-5
-6
-7
-8
Artiodactyla
Carnivora
Chiroptera
Cingulata
Dasyuromorphia
Dermoptera
Didelphimorphia
Diprotodontia
Eulipotyphla
Hyracoidea
Lagomorpha
Macroscelidea
Microbiotheria
Monotremata
Peramelemorphia
Perissodactyla
Pholidota
Pilosa
Primates
Proboscidea
Rodentia
Scandentia
Sirenia
Tubulidentata
order
proj
dnazoo
vgp
zoonomia