

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
<i>Scyliorhinus canicula</i> (smaller sp...	7830	4,220	Dec, 2020	
<i>Pungitius pungitius</i> (ninespine sti...	134920	480.4	Apr, 2023	PacBio,Arima2
<i>Cervus elaphus</i> (red deer)	9860	2,887	Jul, 2021	
<i>Anolis sagrei</i> (Brown anole)	38937	1,951	Mar, 2024	PacBio Sequel II HiFi; Arima Hi-C v2
<i>Melospiza melodia melodia</i>	19149...	1,541	Jan, 2024	PacBio Sequel II HiFi; Bionano DLS...
<i>Manis pentadactyla</i> (Chinese pan...	143292	2,840	May, 2023	PacBio Sequel II HiFi; Dovetail Om...
<i>Tachyglossus aculeatus</i> (Australi...	9261	2,213	Dec, 2020	PacBio Sequel I CLR; Illumina Nova...
<i>Scatophagus argus</i>	75038	570.8	Oct, 2021	PacBio Sequel I CLR; Illumina Nova...
<i>Acanthopagrus latus</i> (yellowfin s...	8177	685.1	Oct, 2020	
<i>Muntiacus reevesi</i> (Reeves' muntj...	9886	2,656	Jan, 2024	PacBio,Arima2
<i>Falco peregrinus</i> (peregrine falcon)	8954	1,312	Jun, 2022	PacBio Sequel II HiFi; Arima Geno...
<i>Thunnus thynnus</i> (Atlantic bluefi...	8237	799.0	Jan, 2024	PacBio,Arima2
<i>Eschrichtius robustus</i> (grey whale)	9764	2,982	Jan, 2023	PacBio Sequel II HiFi; 3D-DNA Hi-C
<i>Thunnus albacares</i> (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 "$methods" | grep -q '[Hh]icanu'; then
    data='PacBio HiFi'
elif echo "$datatype" | grep -q '[Hh]i[FF]i'; then
    data='PacBio HiFi'
elif echo "$comments" | grep -q '[Hh]i[FF]i'; then
    data='PacBio HiFi'
elif echo "$datatype" | grep -q '[Oo]xford'; then
    data='Oxford Nanopore'
elif echo "$datatype" | grep -q '[Nn]anopore'; then
    data='Oxford Nanopore'
elif echo "$datatype" | 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 "$comments" | grep -q 'ONT'; then
    data='Oxford Nanopore'
elif echo "$datatype" | 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 "$comments" | grep -q '[Ff]alcon'; then
    data='PacBio CLR'
elif echo "$datatype" | grep -q '[Rr][Ss][Ii][Ii]'; then
    data='PacBio RSII'
else
    data='missing'
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"

The screenshot shows a Google Sheets document with the title 'VGP Ordinal List'. The table has columns for J, K, L, M, N, O, P, Q, and R. The first few rows show data for various species:

J	K	L	M	N	O	P	Q	R
Name	English Name	NCBI taxon ID	Stat	Assembly ID	Assembly ID main haplotype	Accession # for main haplotype	RefSeq annotation main haplotype	ENSEMBL annotation main haplotype
<i>maximus indicus</i>	Asian elephant	99487	4	mEleMax1	mEleMax1.pri	GCA_024188685.1	GCF_024188685.1	
<i>africana</i>	African elephant	9785		mLoxAfr1	mLoxAfr1.hap2	GCA_030014295.1	GCF_030014295.1	
<i>ax brucei</i>	yellow-spotted rock hyrax	77598	4	mHctBru1	mHctBru1.pri	GCA_028571685.1		
<i>uganda</i>	doguroo	29137	4	mDugDug1	mDugDug1.hap1	GCA_030035685.1		
<i>yon poteri</i>	bushy-tailed and rufous elephant shrew	320637	4	mRhyPetr1	mRhyPetr1.hap1	GCA_043290085.1		
<i>auduboni</i>	Taï forest lemur	319815	3					
<i>us hotentotus</i>	hotentot golden mole	9391	3					
<i>afer</i>	arder	9810						
<i>terradiadedyi</i>	Southern tamandua	48850	4	mTamTer1	mTamTer1.pri	GCA_023851605.1		
<i>c didactylus</i>	Linnaeus's two-toed sloth	27675	4	mChuDid1	mChuDid1.pri	GCA_015220238.1		
<i>coenomysinunctus</i>	mine-banded armadillo	6361	4	mDashNov1	mDasNov1.1.hap2	GCA_030445036.1		
<i>iensis</i>	human	9606	5	T2T-CHM13v2.0	T2T-CHM13v2	GCA_009914755.1		
<i>pygmaeus</i>	chimpanzee	9598		mPanTro3	mPanTro3.hap1	GCA_028858775.2	GCF_028858775.2	
<i>cus</i>	bonobo	9597		mPanPan1	mPanPan1.mat	GCA_029289425.2	GCF_029289425.2	
<i>nils</i>	gorilla	9593		mGorGor1	mGorGor1.mat	GCA_02921585.3	GCF_02921585.3	
<i>silv</i>	Sumatran orangutan	9601		mPonapeb1	mPonapeb1.hap1	GCA_028885655.2	GCF_028885655.2	
<i>griseus</i>	Bornean orangutan	9600		mPonPyg2	mPonPyg2.hap1	GCA_028885625.2	GCF_028885625.2	
<i>syrichta</i>	siamang gibbon	9590		mSymSyn1	mSymSyn1.hap1	GCA_028878055.3	GCF_028878055.3	
<i>wucongensis</i>	hoolock gibbon	593453		mHooLeu1	mHooLeu1.hap1			
<i>emarginata</i>	pig-tailed macaque	9545		mMacNem1	mMacNem1.hap1	GCA_043159975.1		
<i>acchensis</i>	common marmoset	9483	4	mCalUac1	mCalUac1.pat	GCA_011100555.2	GCF_011100555.1	
<i>richta</i>	Philippine tarsier	1868482	2					
<i>cruciana</i>	sunda slow loris	9470	4	mNvrCeu1	mNvrCeu1.nri	GCA_027409675.1	GCF_027409675.1	

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

GENOMES FOR PROJECT

Vertebrate Genomes Project

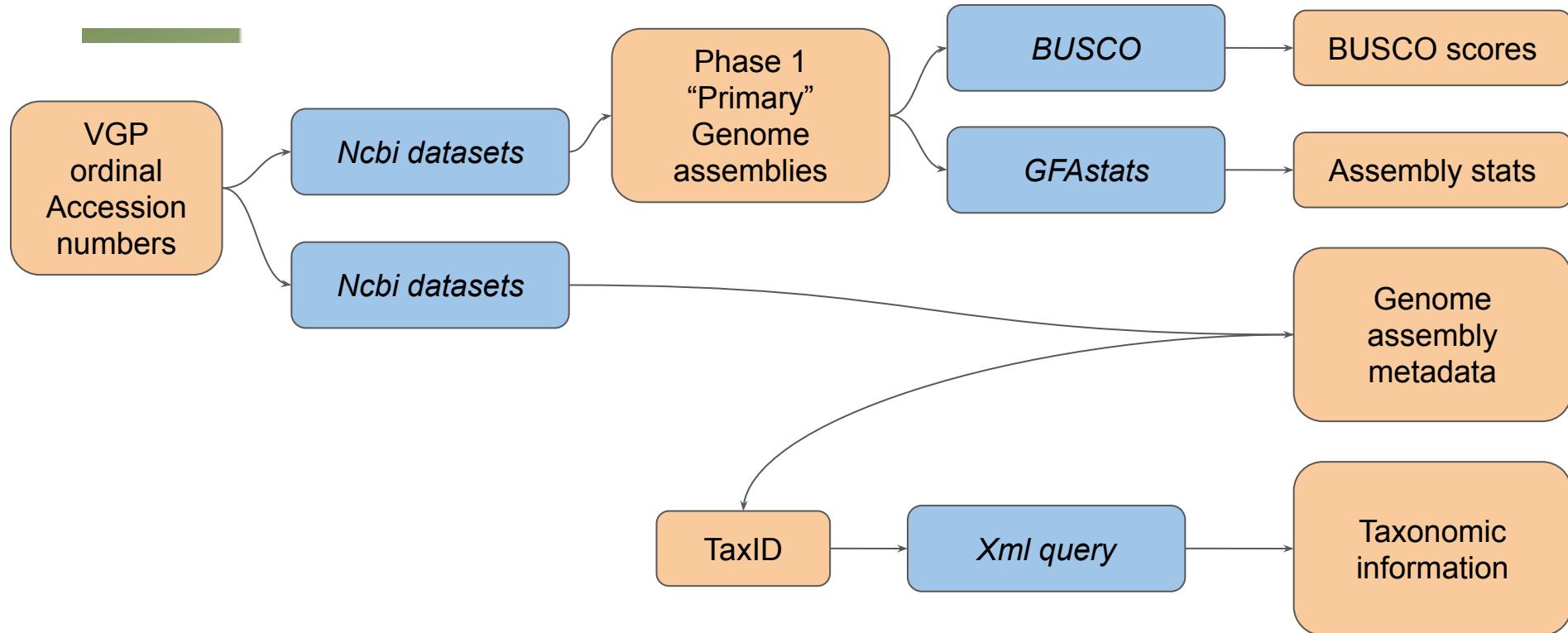
BioProject PRJNA489243

Filters

The screenshot shows a table of 903 genomes from the Vertebrate Genomes Project. The columns are: Assembly, GenBank, RefSeq, Scientific name, Tax ID, and Annotation.

Assembly	GenBank	RefSeq	Scientific name	Tax ID	Annotation
<input type="checkbox"/> fProPar1.1	GCA_964188405.1		Protomyctophum parallellum (p...)	1091426	NCBI RefSeq
<input type="checkbox"/> bGalGal1.mat.broiler.GRCg7b...	GCA_016699485.1	GCF_016699485.2	Gallus gallus (chicken)	9031	NCBI RefSeq
<input type="checkbox"/> ZJU1.0	GCA_015476345.1	GCF_015476345.1	Anas platyrhynchos (mallard)	8839	NCBI RefSeq
<input type="checkbox"/> mCalJa1.2.pat.X	GCA_011100555.2	GCF_011100555.1	Callithrix jacchus (white-tufted-e...	9483	NCBI RefSeq
<input type="checkbox"/> bTaeGut1.4.pri	GCA_003957565.4	GCF_003957565.2	Taeniopygia guttata (zebra finch)	59729	NCBI RefSeq

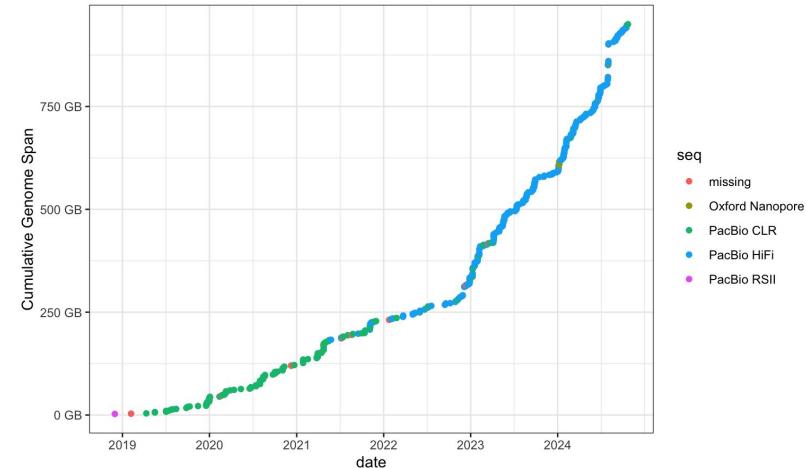
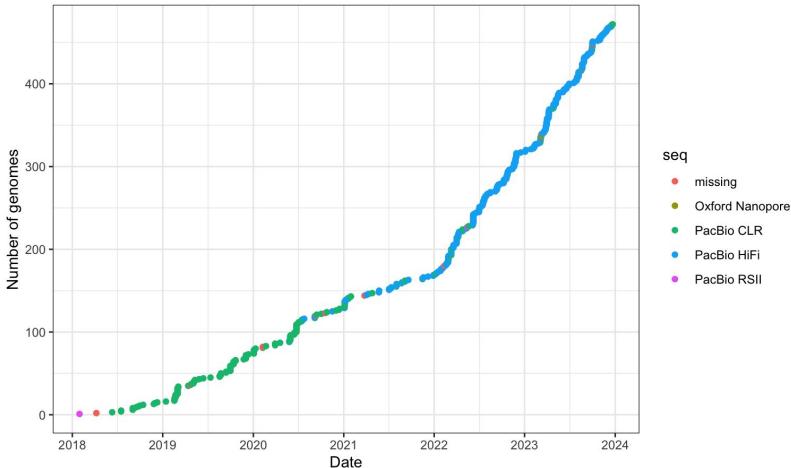
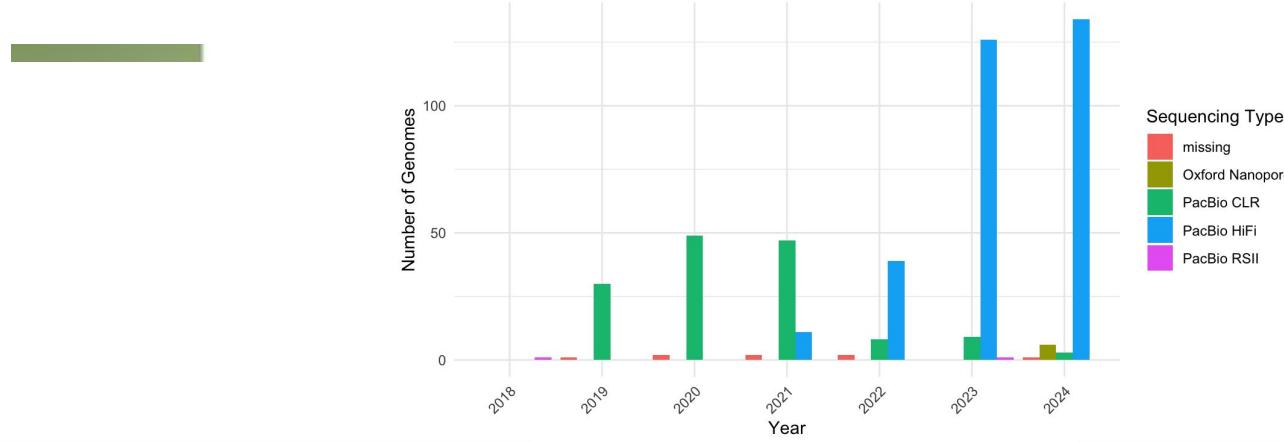
Obtaining data and metadata



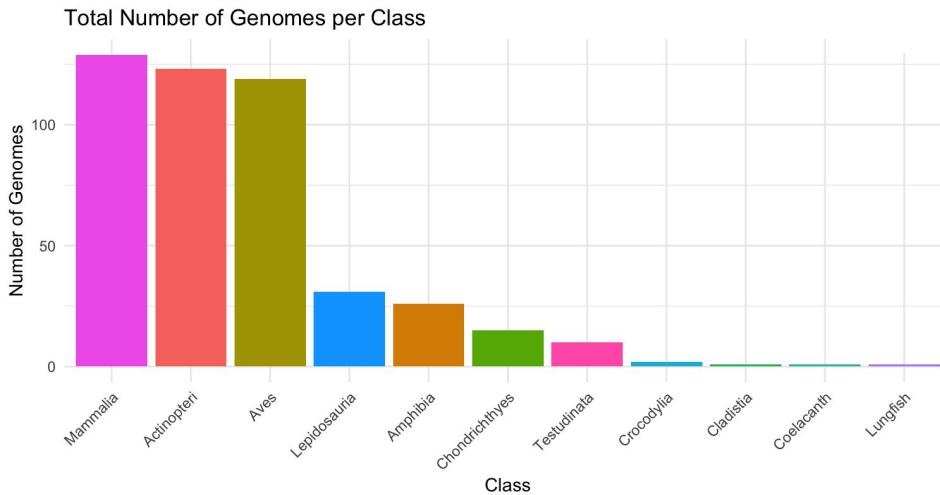
Phase1+ species removed (hagfish/lancelets/sea squirts)
A few “classes” have been filled in for species without them

Rapid increase of HiFi genomes published over the last 2 years

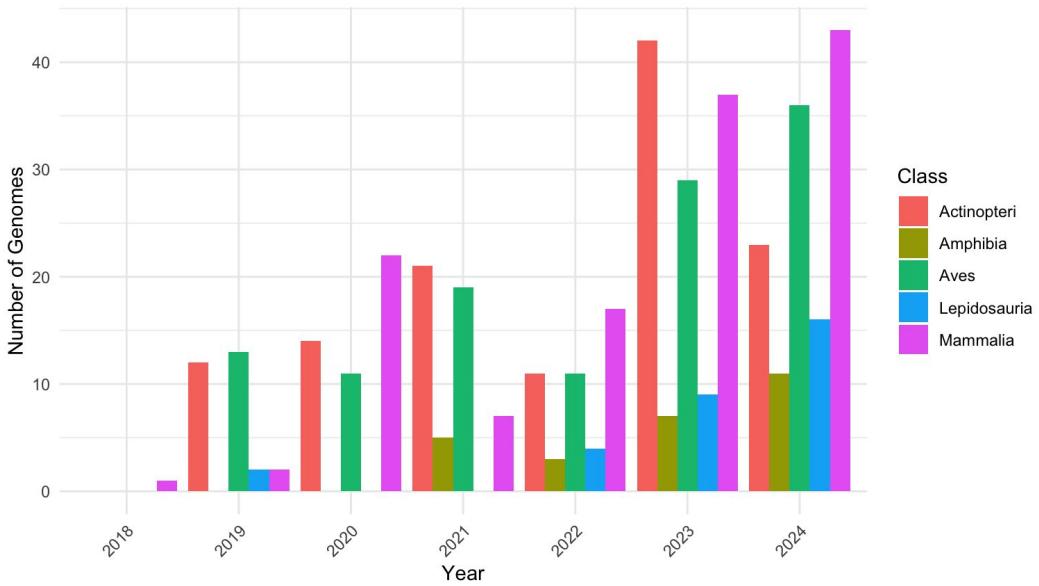
Number of Genomes per Year by Sequencing Type



Genomes per “class”



Number of Genomes per Year by Class



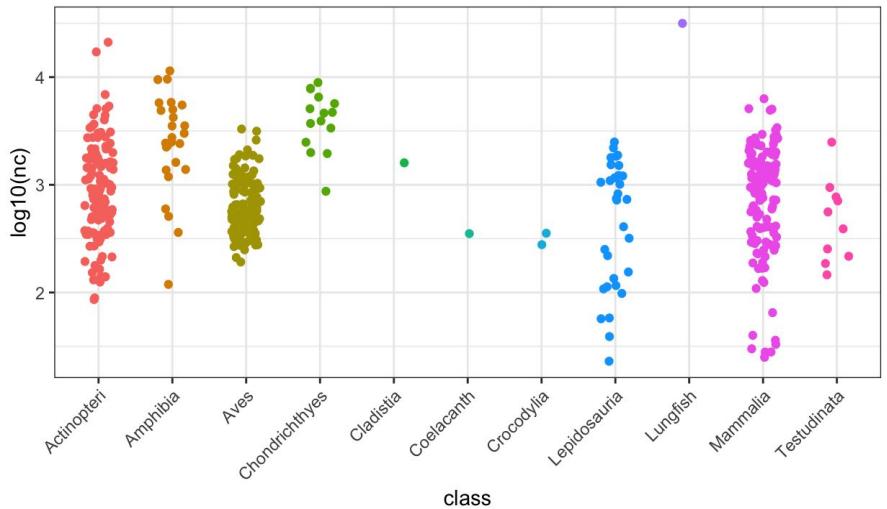
Class

- Actinopteri
- Amphibia
- Aves
- Lepidosauria
- Mammalia

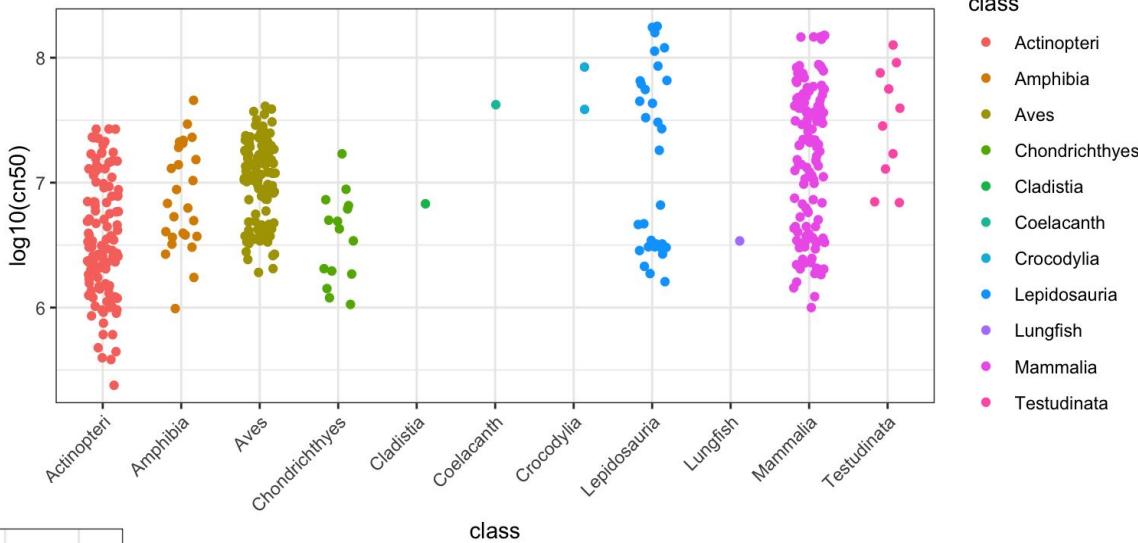
Contiguity by class



No. contigs per assembly



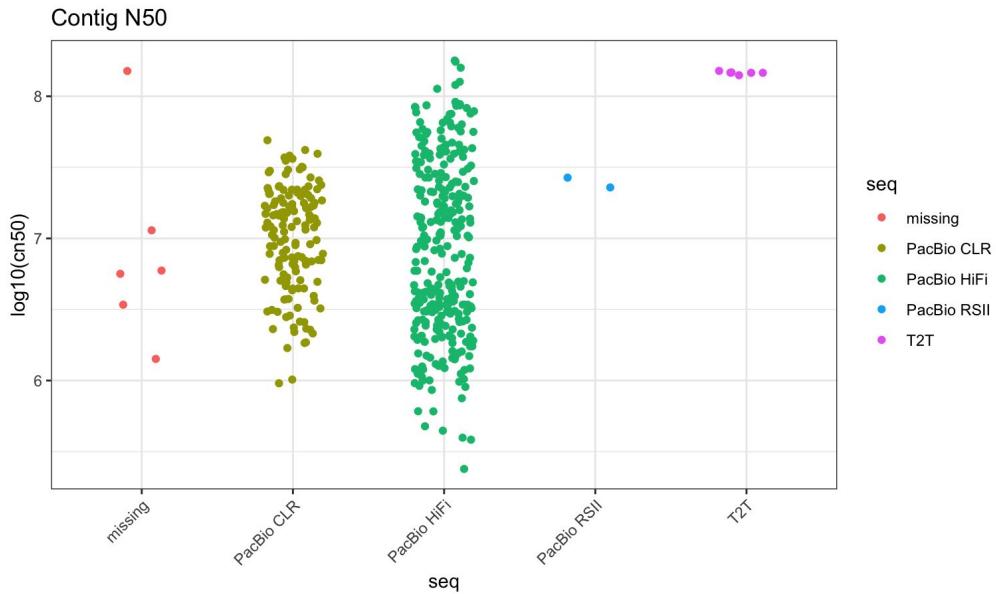
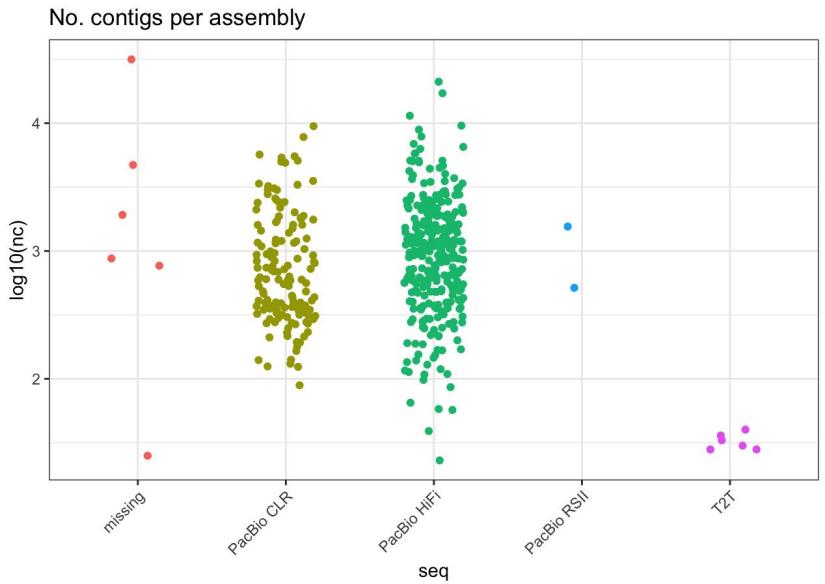
Contig N50



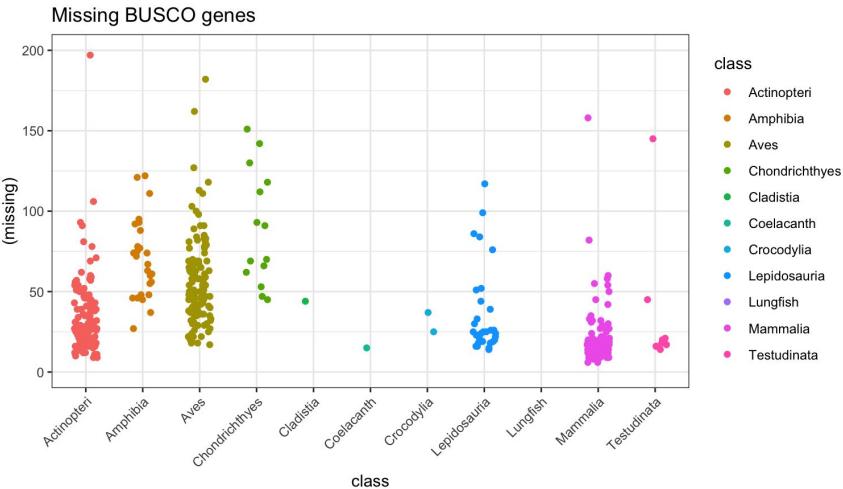
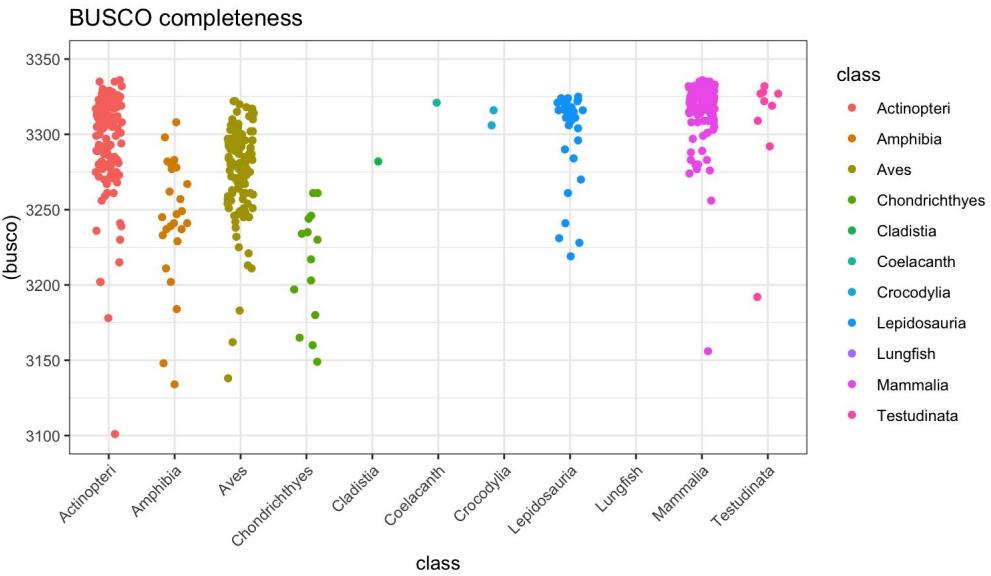
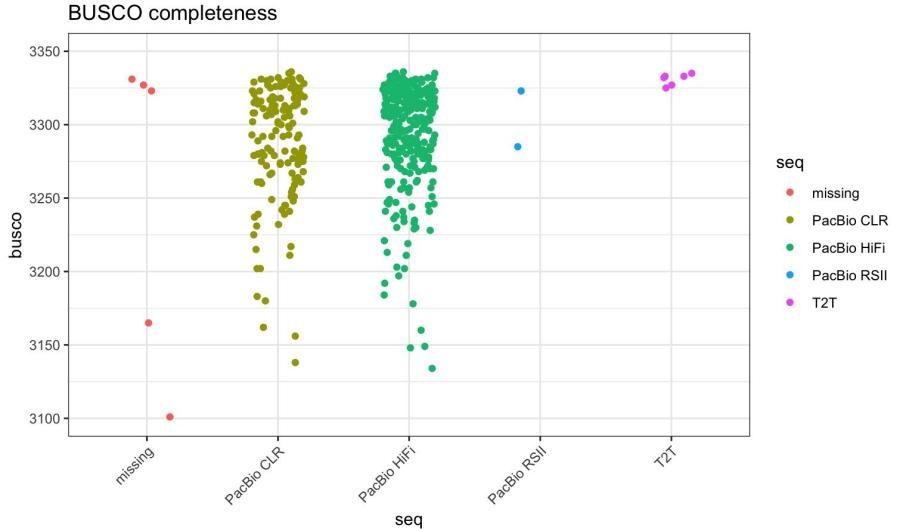
class

- Actinopteri
- Amphibia
- Aves
- Chondrichthyes
- Cladistia
- Coelacanth
- Crocodylia
- Lepidosaurs
- Lungfish
- Mammalia
- Testudinata

Contiguity by sequencing type



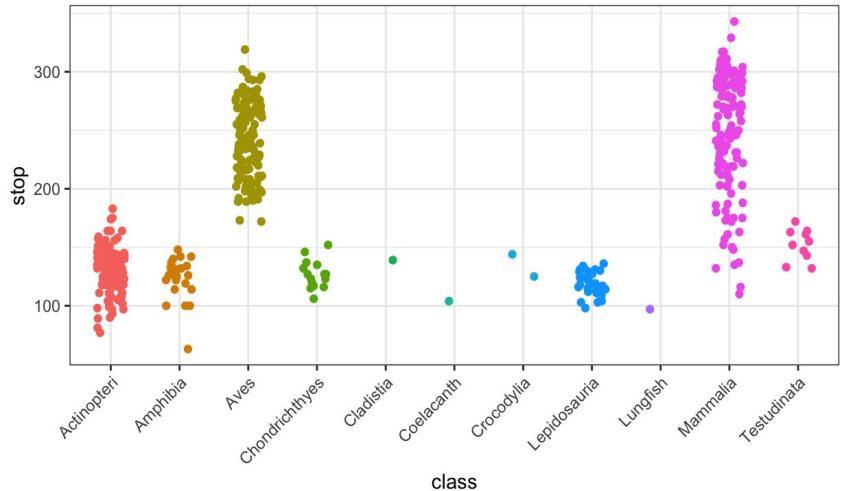
BUSCO completeness



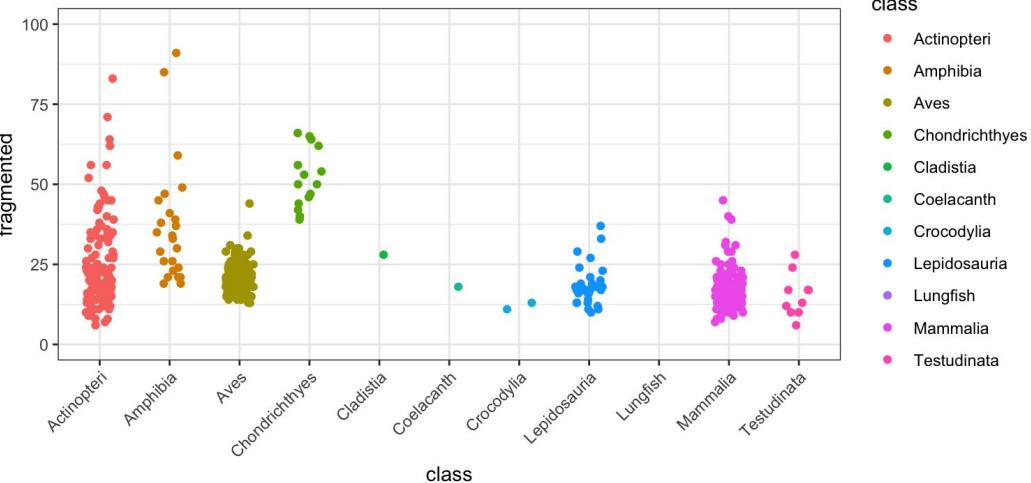
and more BUSCO



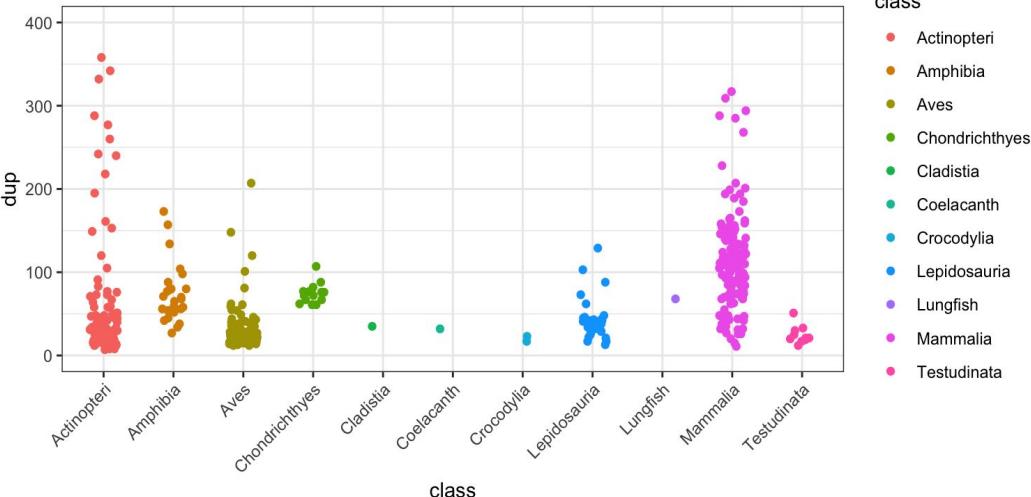
BUSCO genes containing STOP codons



Fragmented BUSCO genes



Duplicated BUSCO genes



Comparing with other large vertebrate genome projects

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

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

GENOMES FOR PROJECT

DNA Zoo

BioProject PRJNA944680

Filters

52 Genomes					
Rows per page 20 1-20 of 52					
Assembly	GenBank	RefSeq	Scientific name	Tax ID	Annotation
<input type="checkbox"/> Grampus_griseus_HiC	GCA_028646425.1		Grampus griseus (Risso's dolphin)	83653	
<input type="checkbox"/> Panthera_onca_HiC	GCA_028533385.1	GCF_028533385.1	Panthera onca (jaguar)	9690	NCBI RefSeq
<input type="checkbox"/> Papio_papio_HiC	GCA_028645565.1		Papio papio (Guinea baboon)	100937	
<input type="checkbox"/> pl-1k	GCA_028646535.1		Procyon lotor (raccoon)	9654	
<input type="checkbox"/> Cryptoprocta_ferox_HiC	GCA_028646485.1		Cryptoprocta ferox (fossa)	94188	
<input type="checkbox"/> Tremarctos_ornatus_HiC	GCA_028551375.1		Tremarctos ornatus (spectacled...)	9638	
<input type="checkbox"/> Rhinoceros_unicornis_HiC	GCA_028646465.1		Rhinoceros unicornis (greater...)	9809	

NCBI Datasets

Taxonomy

Genome

Gene

Command-line tools

Documentation

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

GENOMES FOR PROJECT

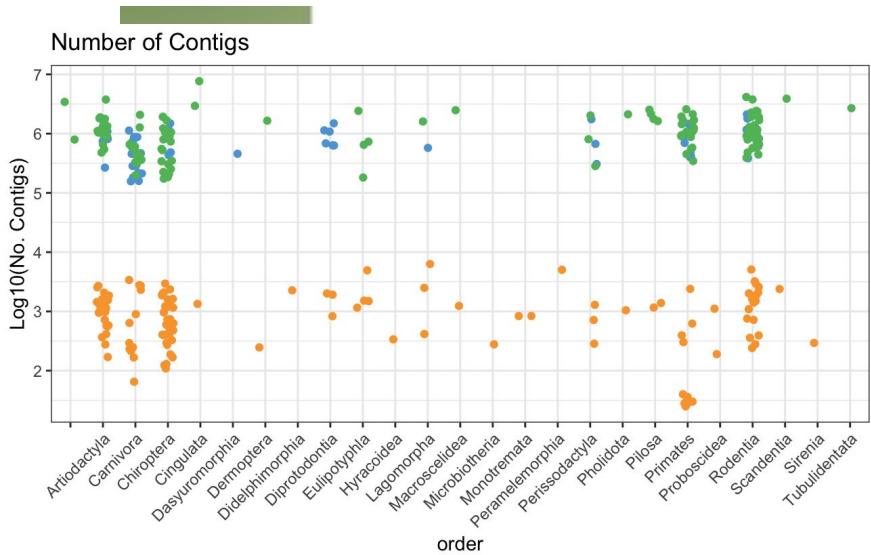
200 Mammals

BioProject PRJNA312960

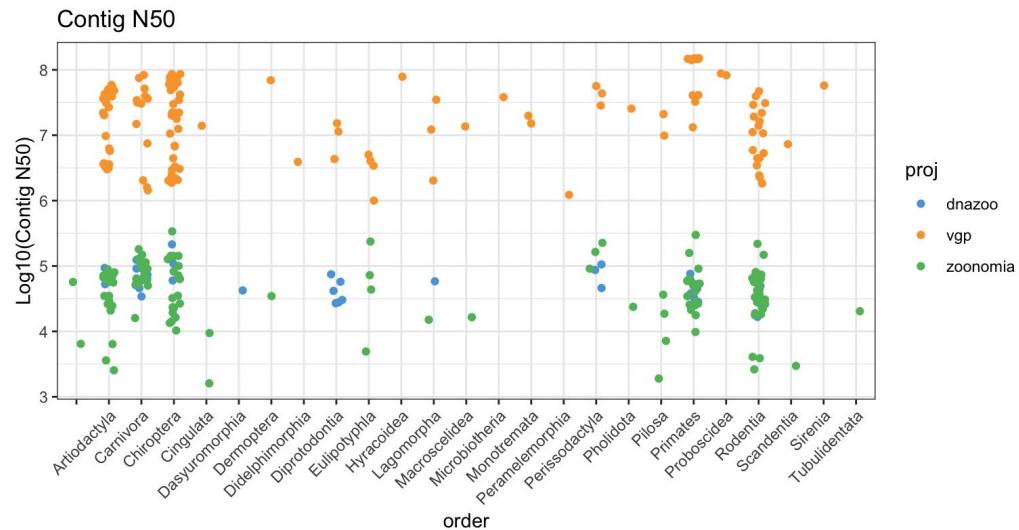
Filters

131 Genomes					
Rows per page 20 1-20 of 131					
Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation
<input type="checkbox"/> OndZib_v1_BIUU	GCA_004026605.1		Ondatra zibethicus (muskrat)	US110 (isolate)	
<input type="checkbox"/> MosMos_v2_BIUU_UCD	GCA_004024705.2		Moschus moschiferus (Siberian...)	BS20 (isolate)	
<input type="checkbox"/> SolPar_v1_BIUU	GCA_004363575.1		Solenodon paradoxus (Hispanio...)	US097 (isolate)	
<input type="checkbox"/> TapTer_v1_BIUU	GCA_004025025.1		Tapirus terrestris (Brazilian tapir)	BS40 (isolate)	
<input type="checkbox"/> HydHyd_v1_BIUU	GCA_004027455.1		Hydrochoerus hydrochaeris (ca...	US065 (isolate)	

Long-read genomes more contiguous 😊

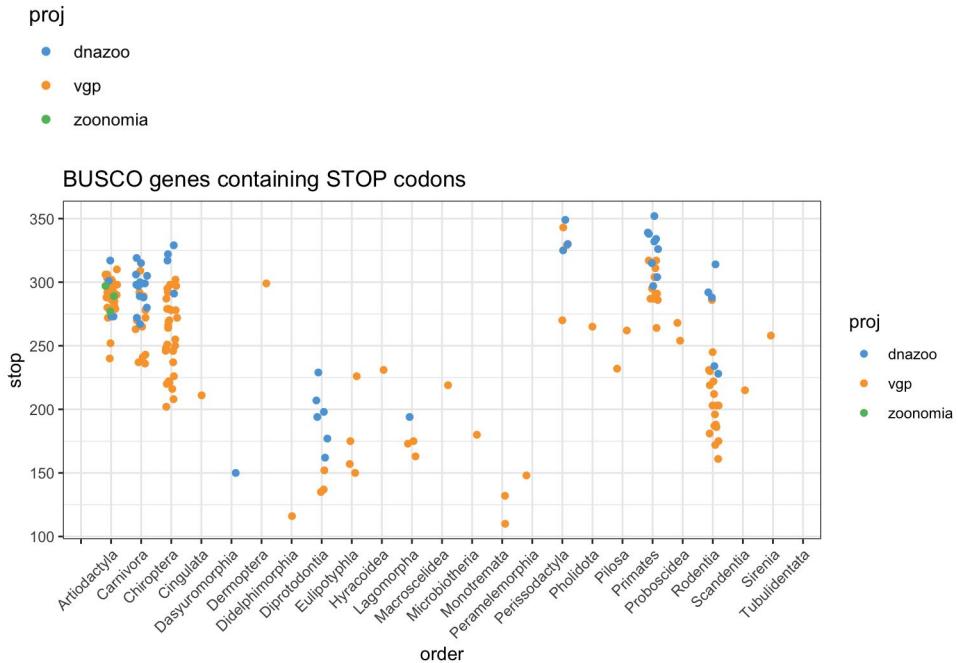
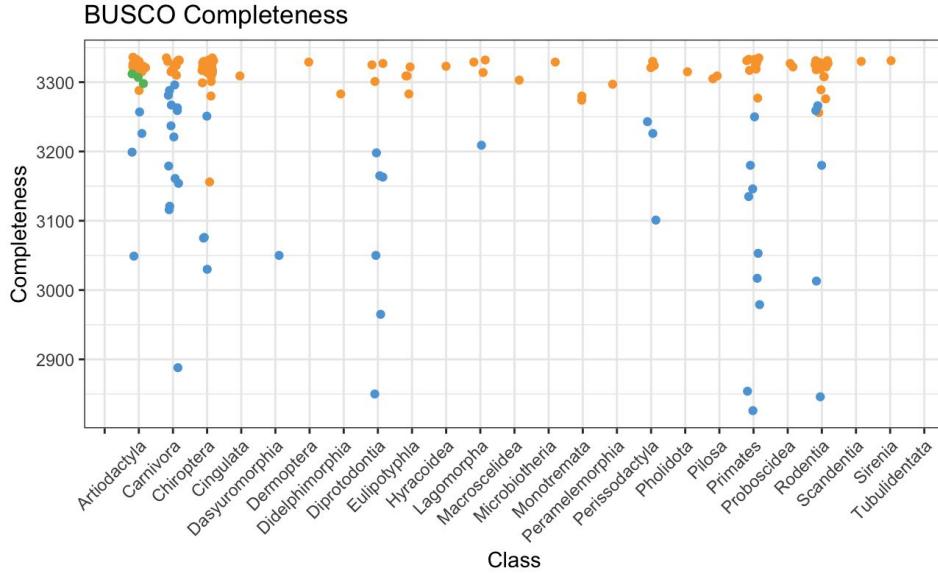


proj
● dnazoo
● vgp
● zoonomia



proj
● dnazoo
● vgp
● zoonomia

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



Initial observations

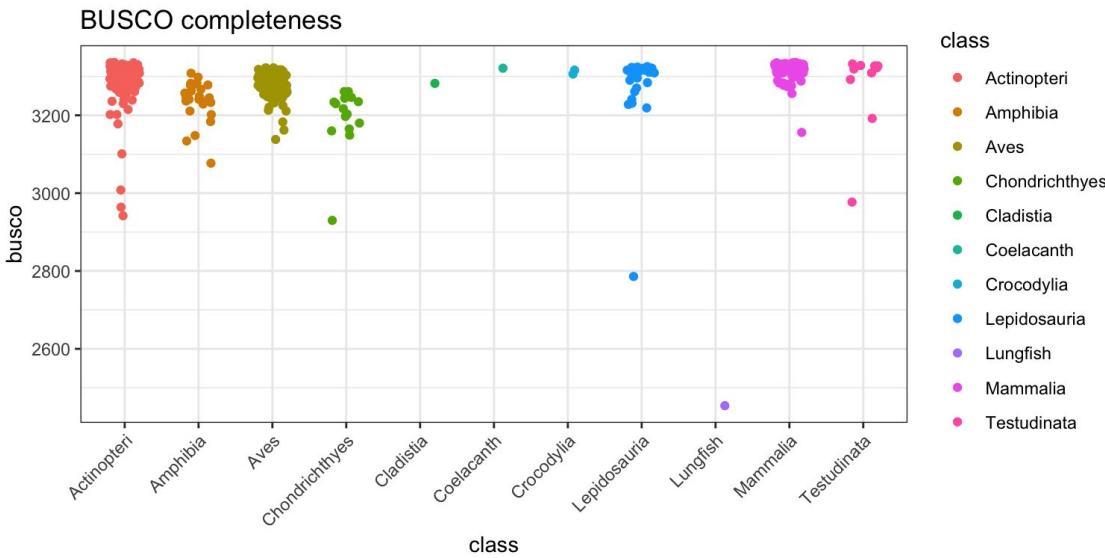
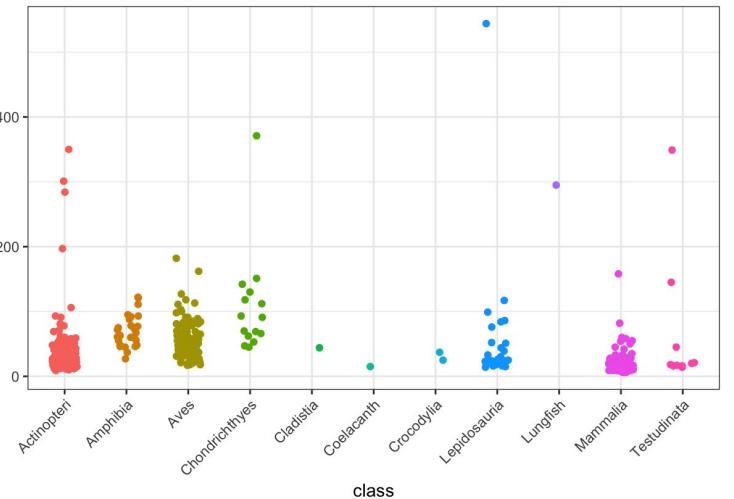
- At this level, superficial statistics not able to see a difference in CLR/HiFi (can we think about (compute) cost per genome?)
- Statistics fairly consistent apart from lower contiguity/completeness in Rays & Sharks?
- Elevated BUSCO genes with stop codons in mammals and birds
- Assemblies are more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

Missing BUSCO genes



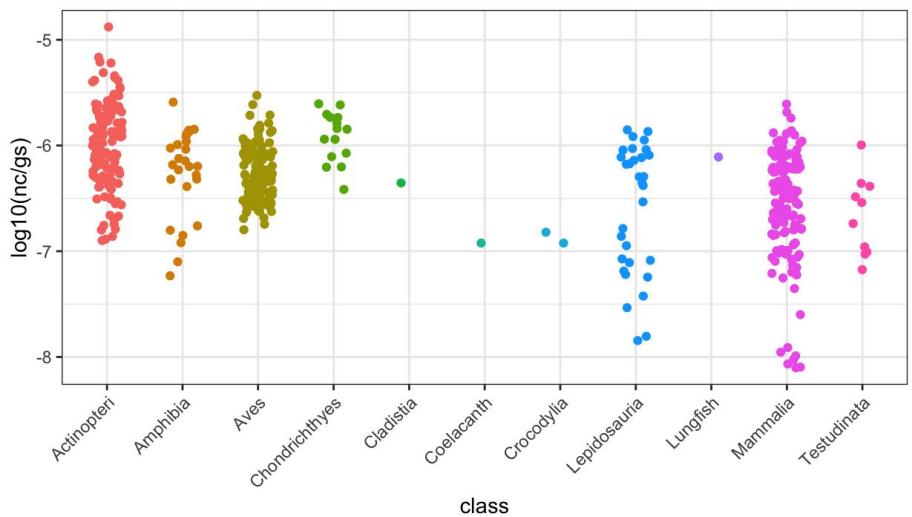
class

- Actinopteri
- Amphibia
- Aves
- Chondrichthyes
- Cladistia
- Coelacanth
- Crocodylia
- Lepidosauria
- Lungfish
- Mammalia
- Testudinata

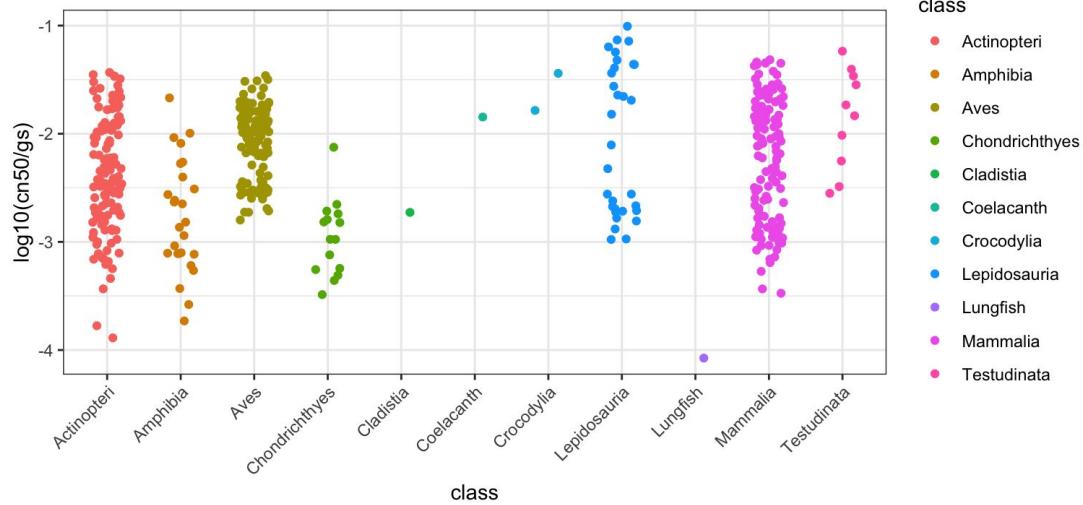
Appendix



No. contigs per assembly per Gb



Contig N50 per Gb



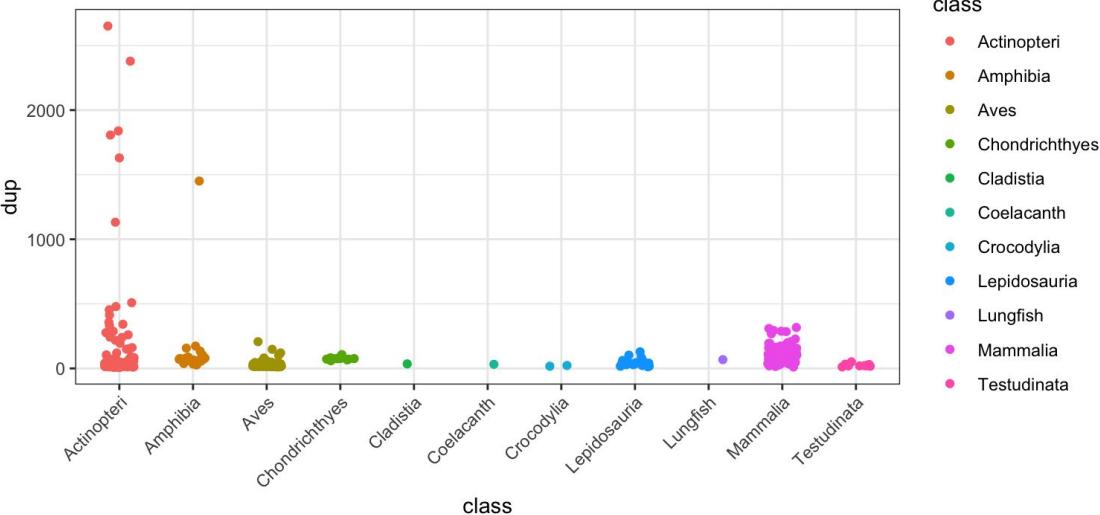
class

- Actinopteri
- Amphibia
- Aves
- Chondrichthyes
- Cladistia
- Coelacanth
- Crocodylia
- Lepidosauria
- Lungfish
- Mammalia
- Testudinata

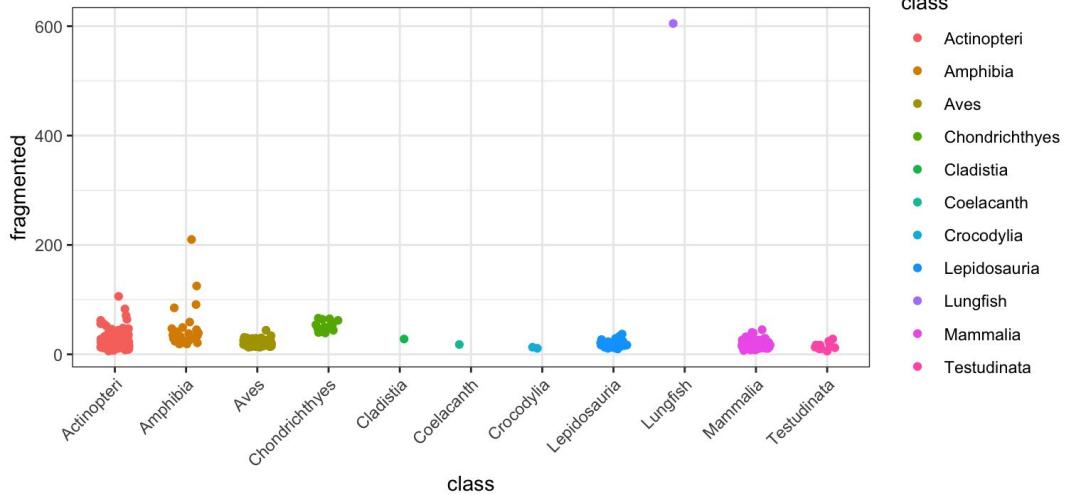
Appendix



Duplicated BUSCO genes



Fragmented BUSCO genes



class

- Actinopteri
- Amphibia
- Aves
- Chondrichthyes
- Cladistia
- Coelacanth
- Crocodylia
- Lepidosauria
- Lungfish
- Mammalia
- Testudinata

Appendix

