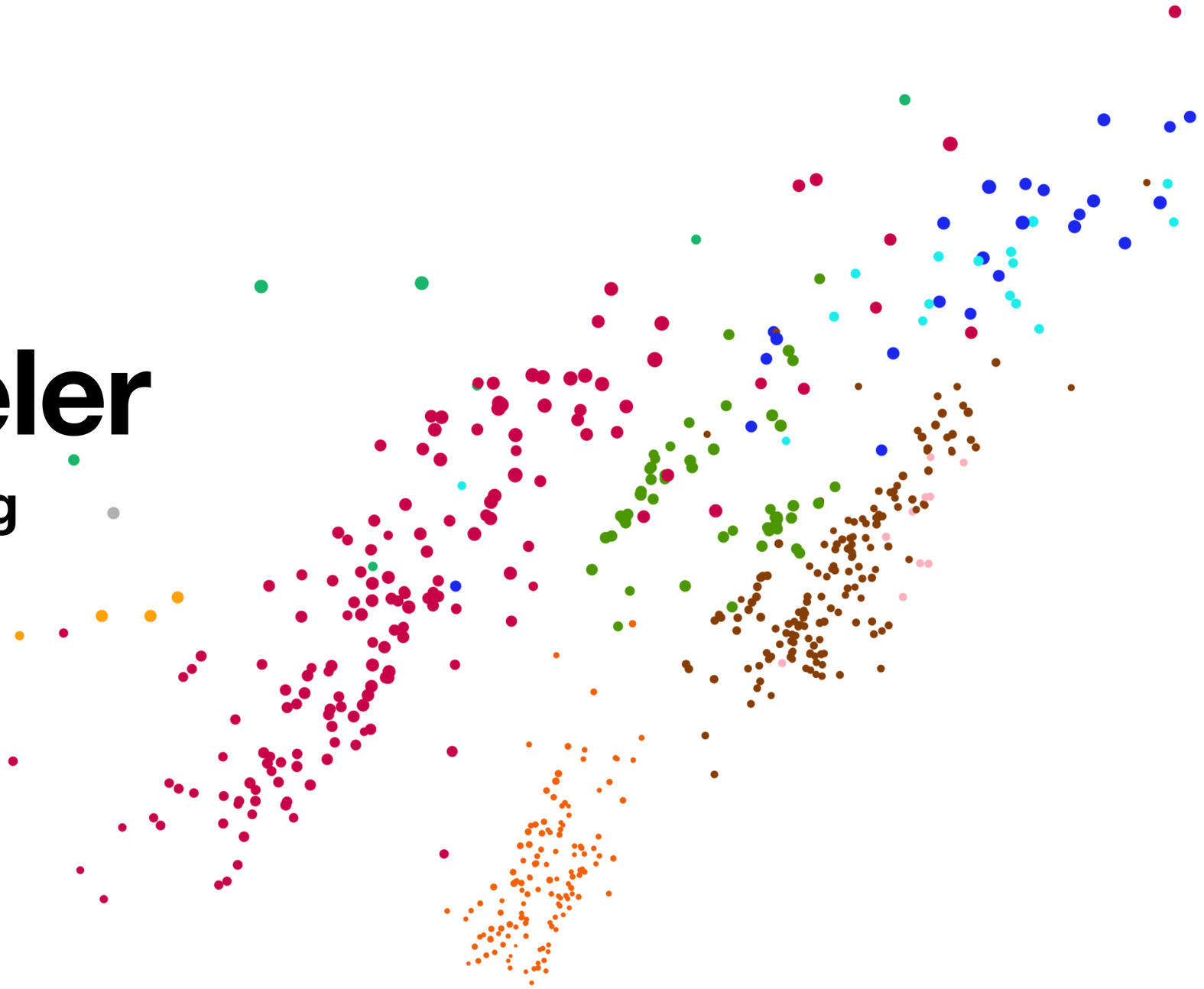
VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

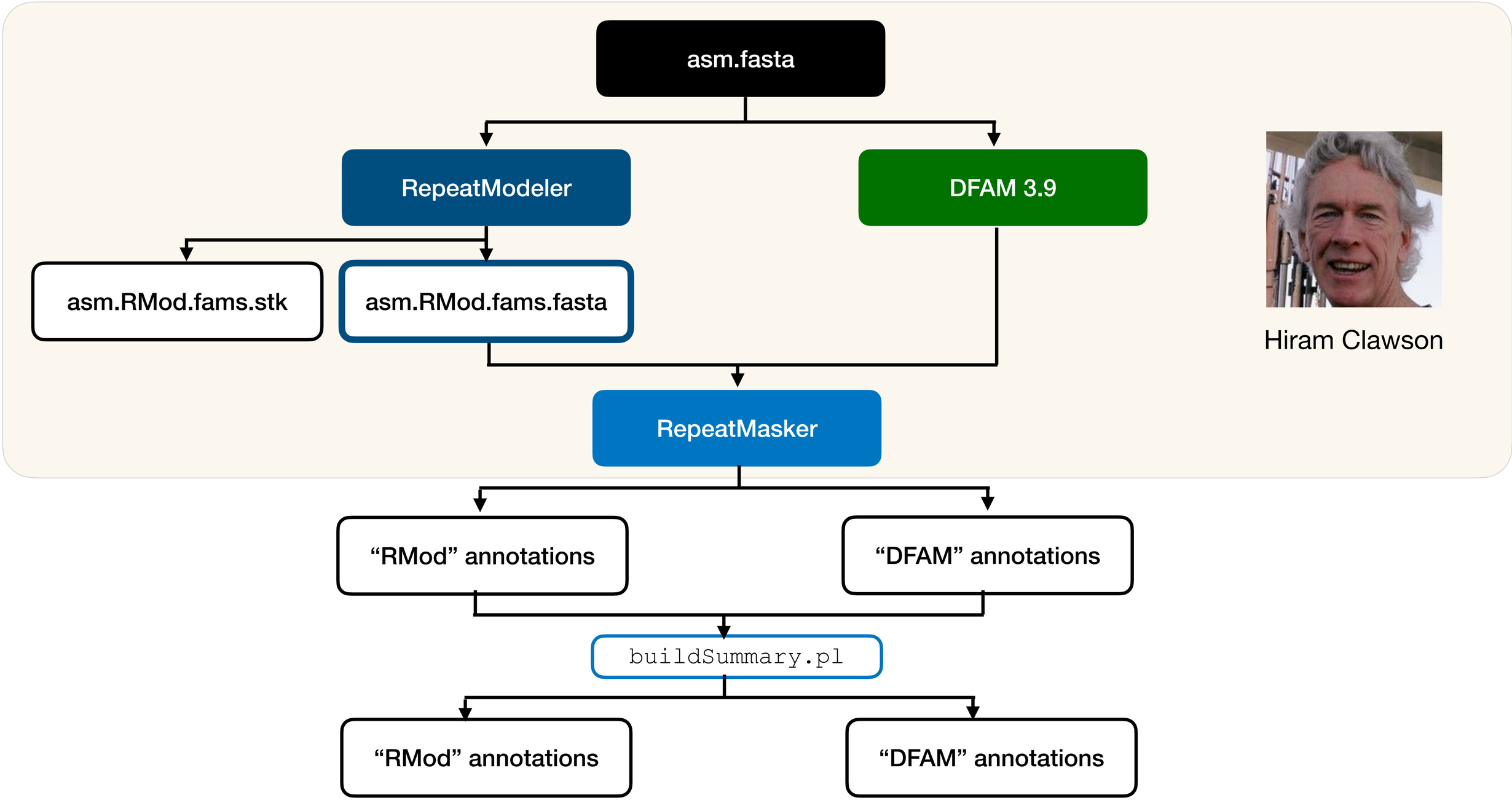
RepeatModeler

VGP repeat group meeting



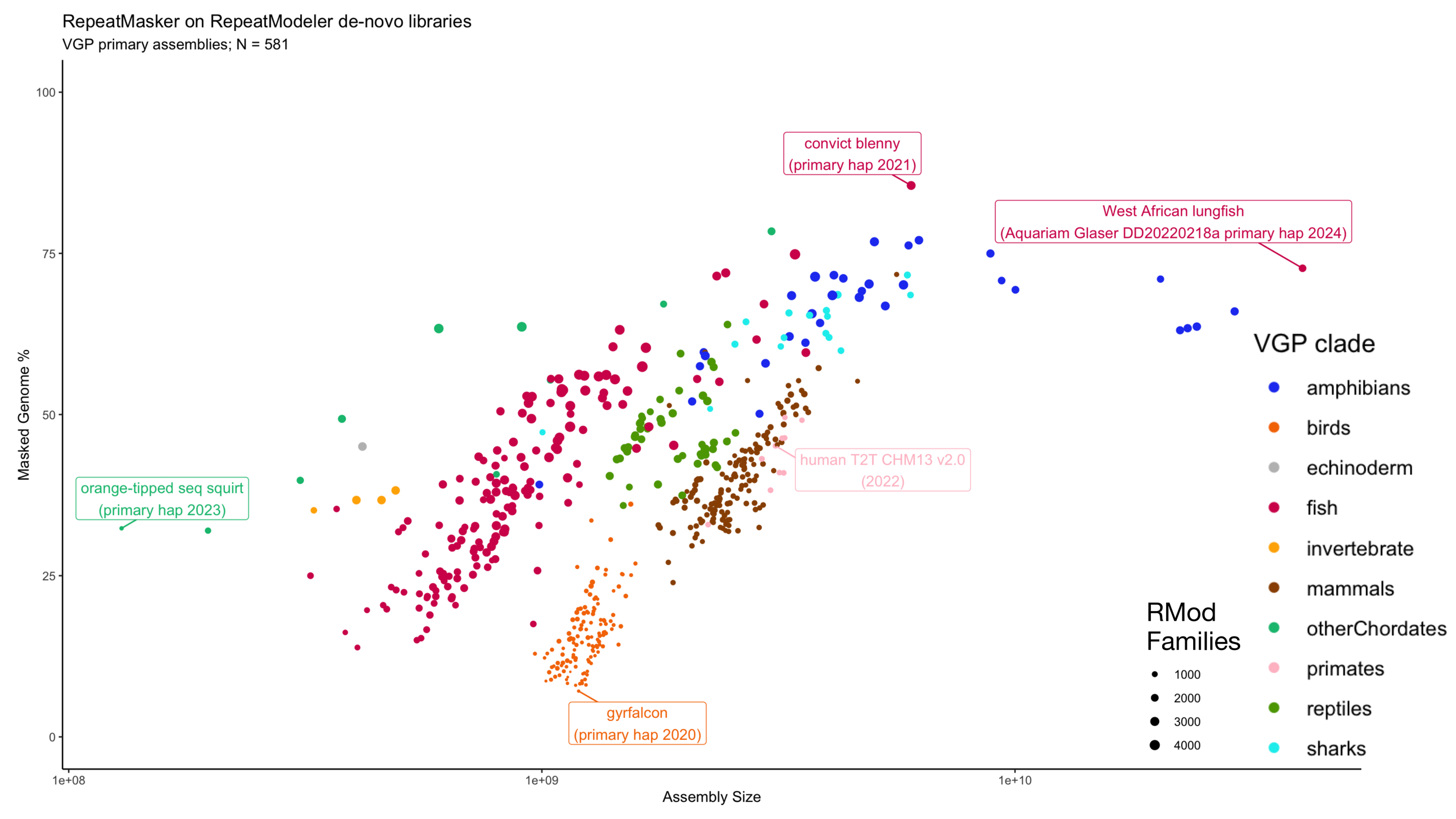
Clément Goubert | 07-24-25

RepeatModeler on 581 primary assemblies



Hiram Clawson

TE content ~ Assembly size


RepeatMasker on RepeatModeler de-novo libraries
VGP primary assemblies; N = 581
Masked Genome %
100
75
50
25
0
1e+08
1e+09
1e+10
Assembly Size
convict blenny
(primary hap 2021)
West African lungfish
(Aquariam Glaser DD20220218a primary hap 2024)
human T2T CHM13 v2.0
(2022)
orange-tipped seq squirt
(primary hap 2023)
gyrfalcon
(primary hap 2020)
VGP clade
amphibians
birds
echinoderm
fish
invertebrate
mammals
otherChordates
primates
reptiles
sharks
RMod
Families
1000
2000
3000
4000

RepeatMasker on RepeatModeler de-novo libraries

VGP primary assemblies; N = 581

convict blenny
(primary hap 2021)

West African lungfish
(Aquariam Glaser DD20220218a primary hap 2024)

orange-tipped sea squirt
(primary hap 2023)

human T2T CHM13 v2.0
(2022)

gyrfalcon
(primary hap 2020)

Masked Genome %

100
75
50
25
0

1e+08
1e+09
1e+10

Assembly Size

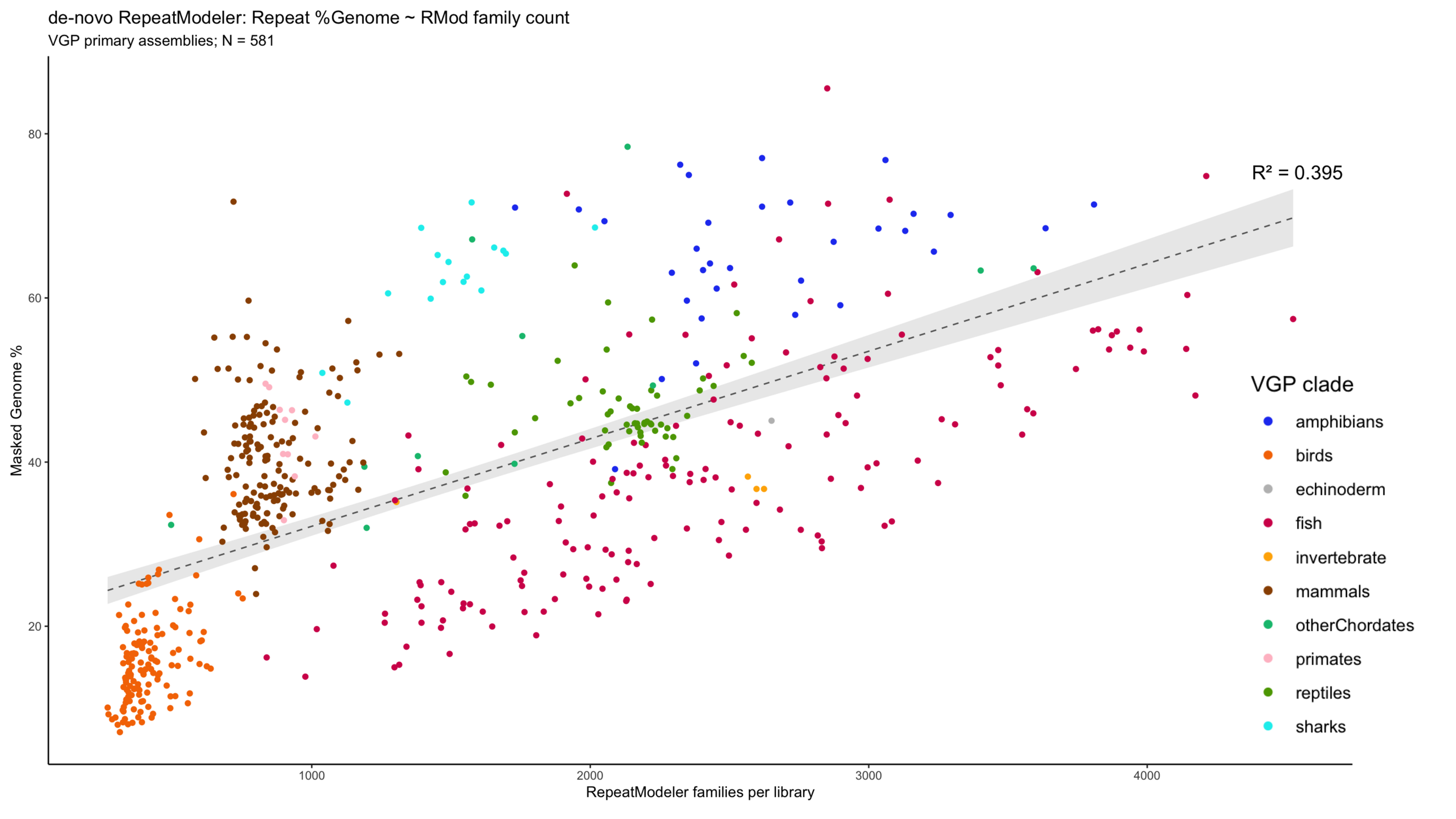
VGP clade

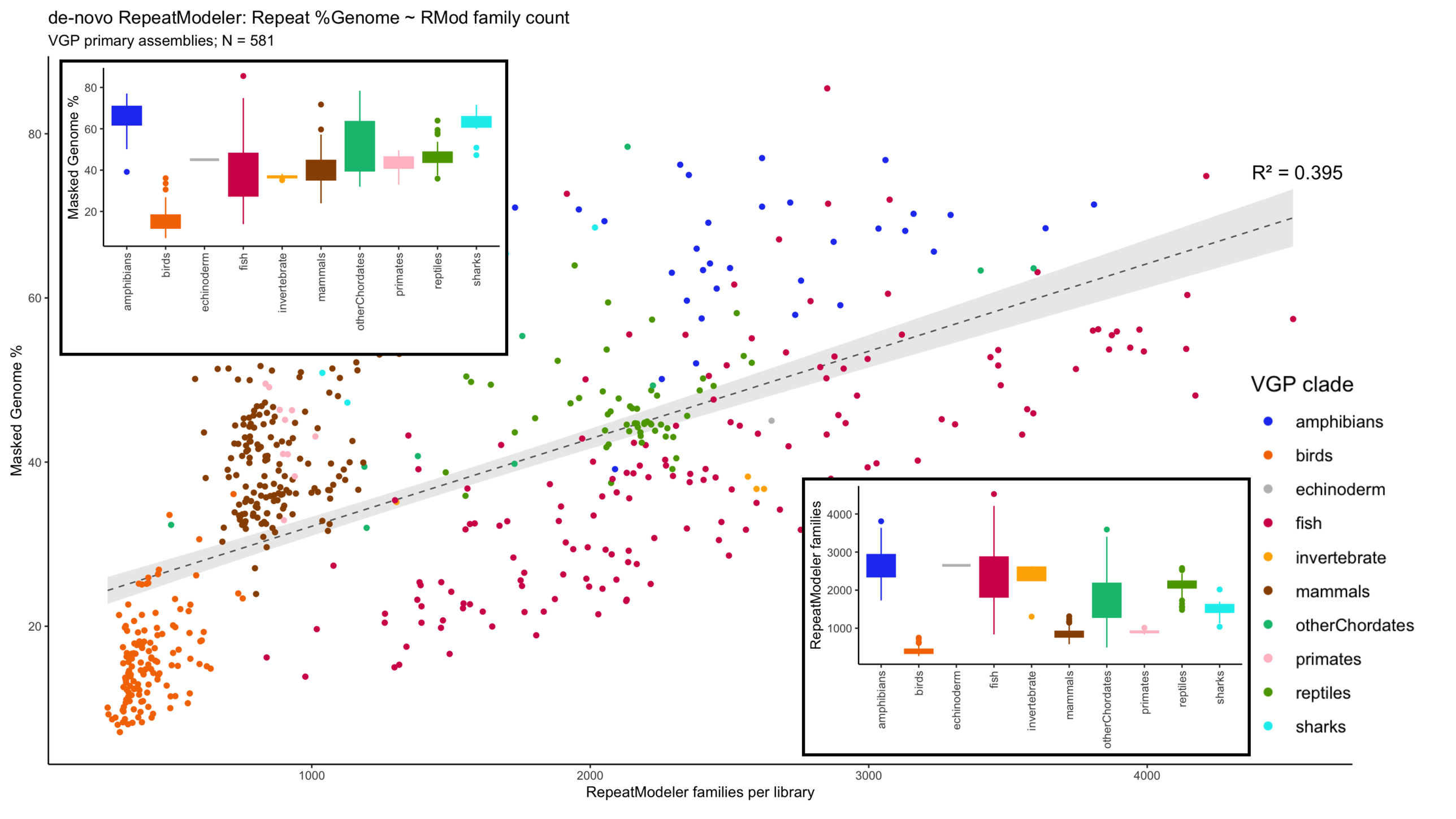
- amphibians
- birds
- echinoderm
- fish
- invertebrate
- mammals
- otherChordates
- primates
- reptiles
- sharks

RMod Families

- 1000
- 2000
- 3000
- 4000

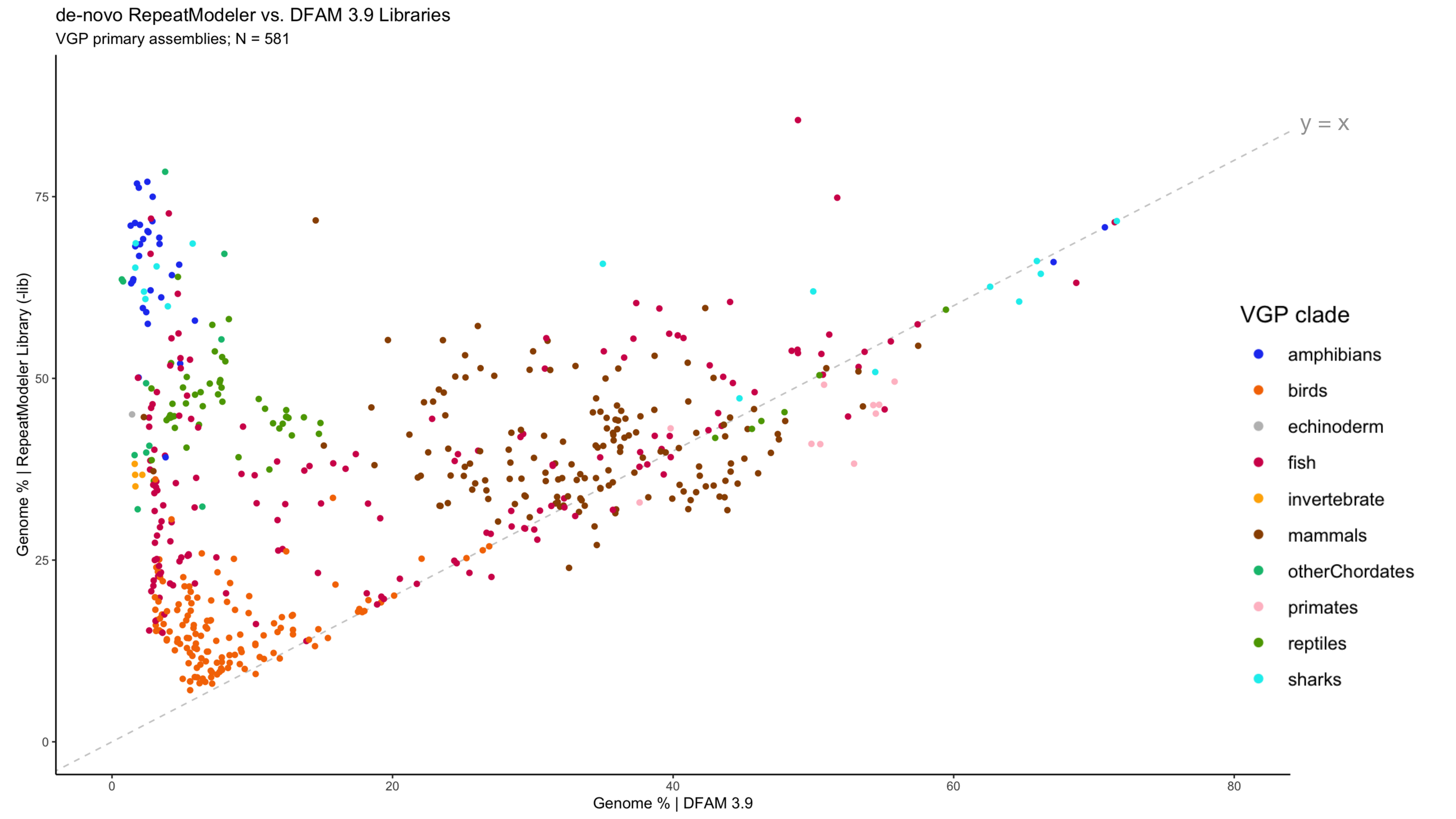
TE content ~ # RMod families

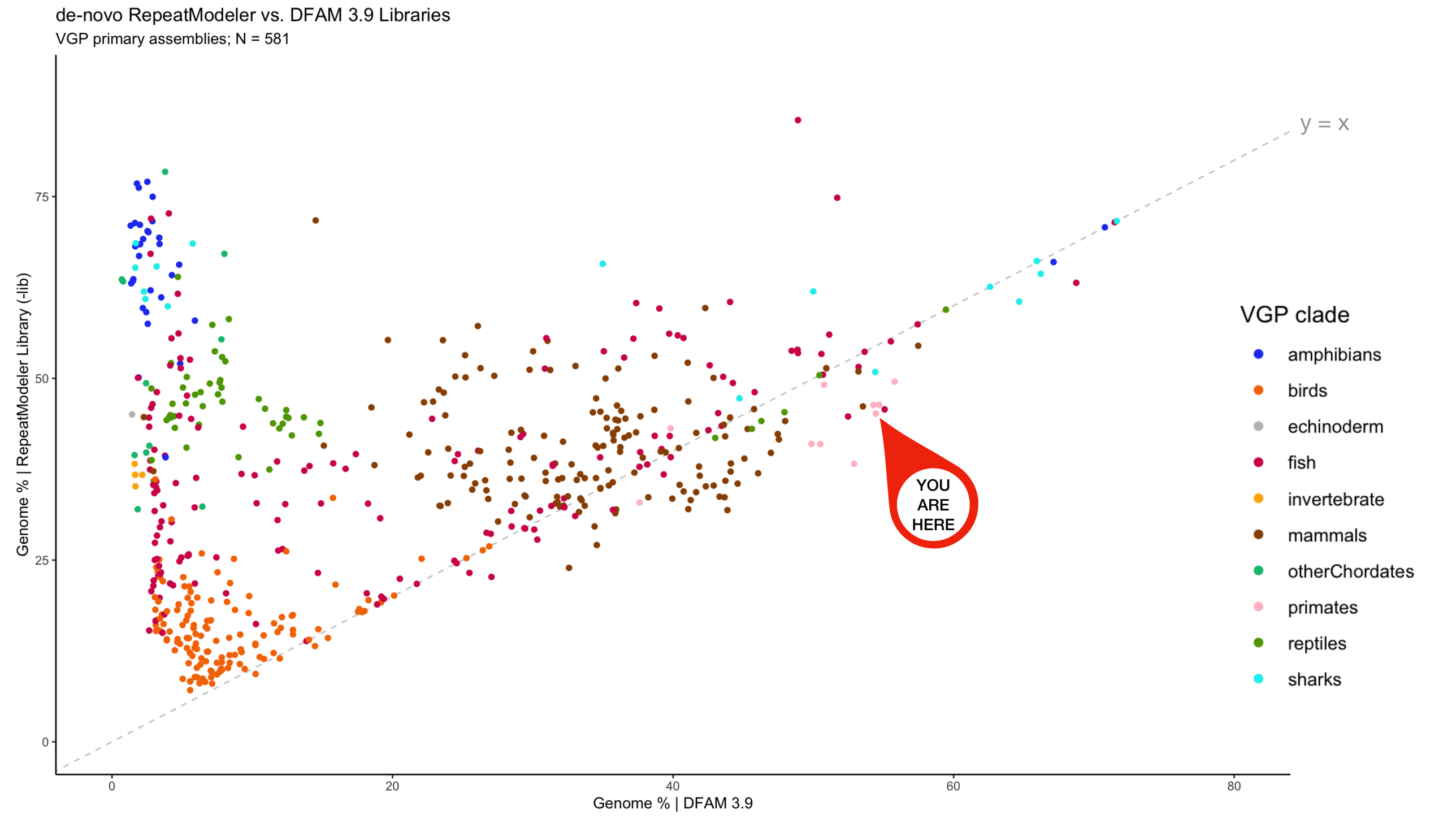

de-novo RepeatModeler: Repeat %Genome ~ RMod family count
VGP primary assemblies; N = 581
Masked Genome %
80
60
40
20
1000
2000
3000
4000
RepeatModeler families per library
R² = 0.395
VGP clade
amphibians
birds
echinoderm
fish
invertebrate
mammals
otherChordates
primates
reptiles
sharks


de-novo RepeatModeler: Repeat %Genome ~ RMod family count
VGP primary assemblies; N = 581
R² = 0.395
Masked Genome %
RepeatModeler families per library
VGP clade
amphibians
birds
echinoderm
fish
invertebrate
mammals
otherChordates
primates
reptiles
sharks

RepeatModeler vs DFAM 3.9

Where are we learning the most?


de-novo RepeatModeler vs. DFAM 3.9 Libraries
VGP primary assemblies; N = 581
y = x
Genome % | RepeatModeler Library (-lib)
Genome % | DFAM 3.9
0
25
50
75
0
20
40
60
80
VGP clade
amphibians
birds
echinoderm
fish
invertebrate
mammals
otherChordates
primates
reptiles
sharks


de-novo RepeatModeler vs. DFAM 3.9 Libraries
VGP primary assemblies; N = 581
y = x
YOU ARE HERE
VGP clade
amphibians
birds
echinoderm
fish
invertebrate
mammals
otherChordates
primates
reptiles
sharks
Genome % | RepeatModeler Library (-lib)
Genome % | DFAM 3.9
0
25
50
75
0
20
40
60
80

de-novo RepeatModeler vs. DFAM 3.9 Libraries

VGP primary assemblies; N = 581

y = x

YOU ARE HERE

VGP clade

- amphibians
- birds
- echinoderm
- fish
- invertebrate
- mammals
- otherChordates
- primates
- reptiles
- sharks

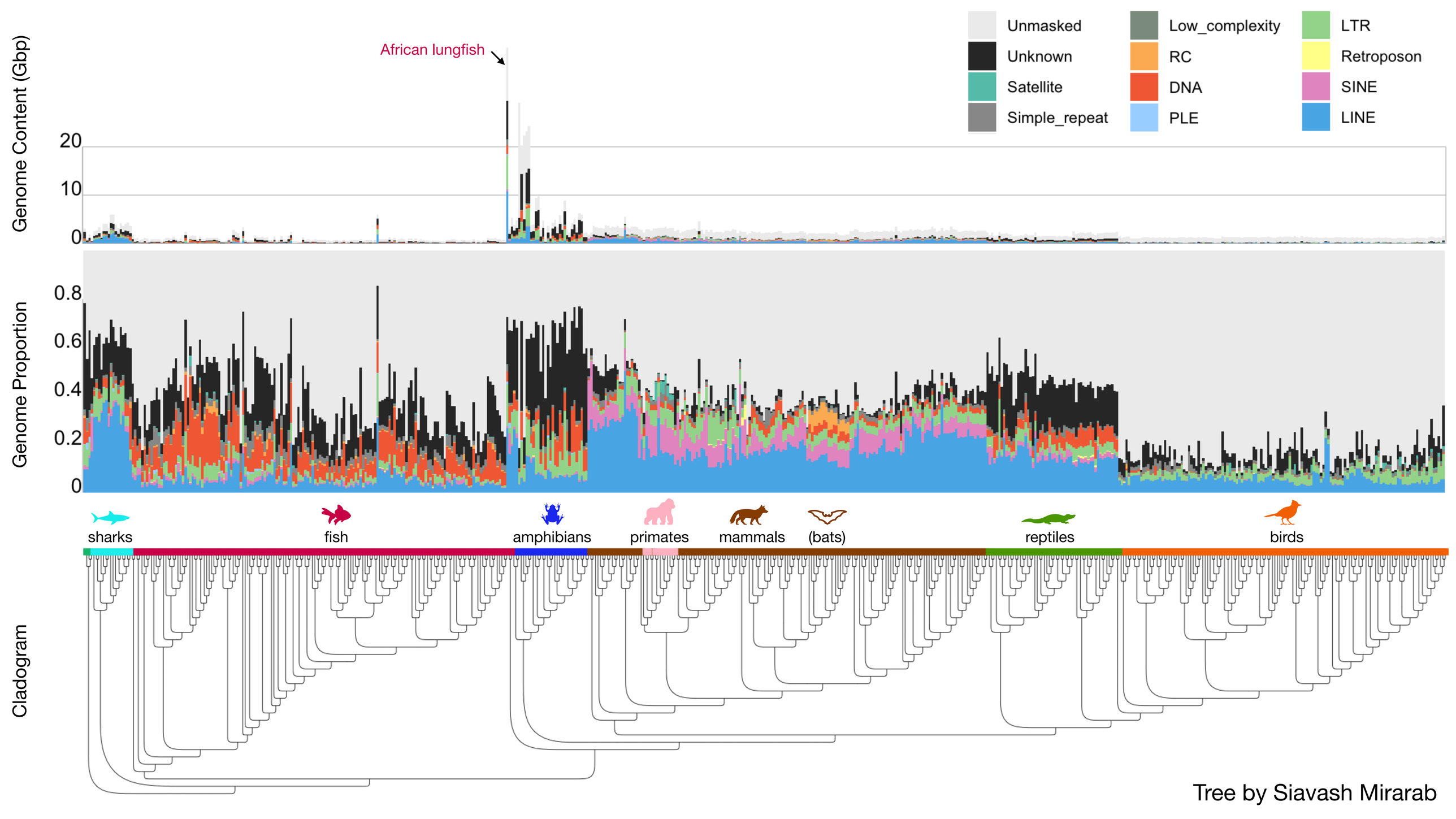
Genome % | RepeatModeler Library (-lib)

0
25
50
75

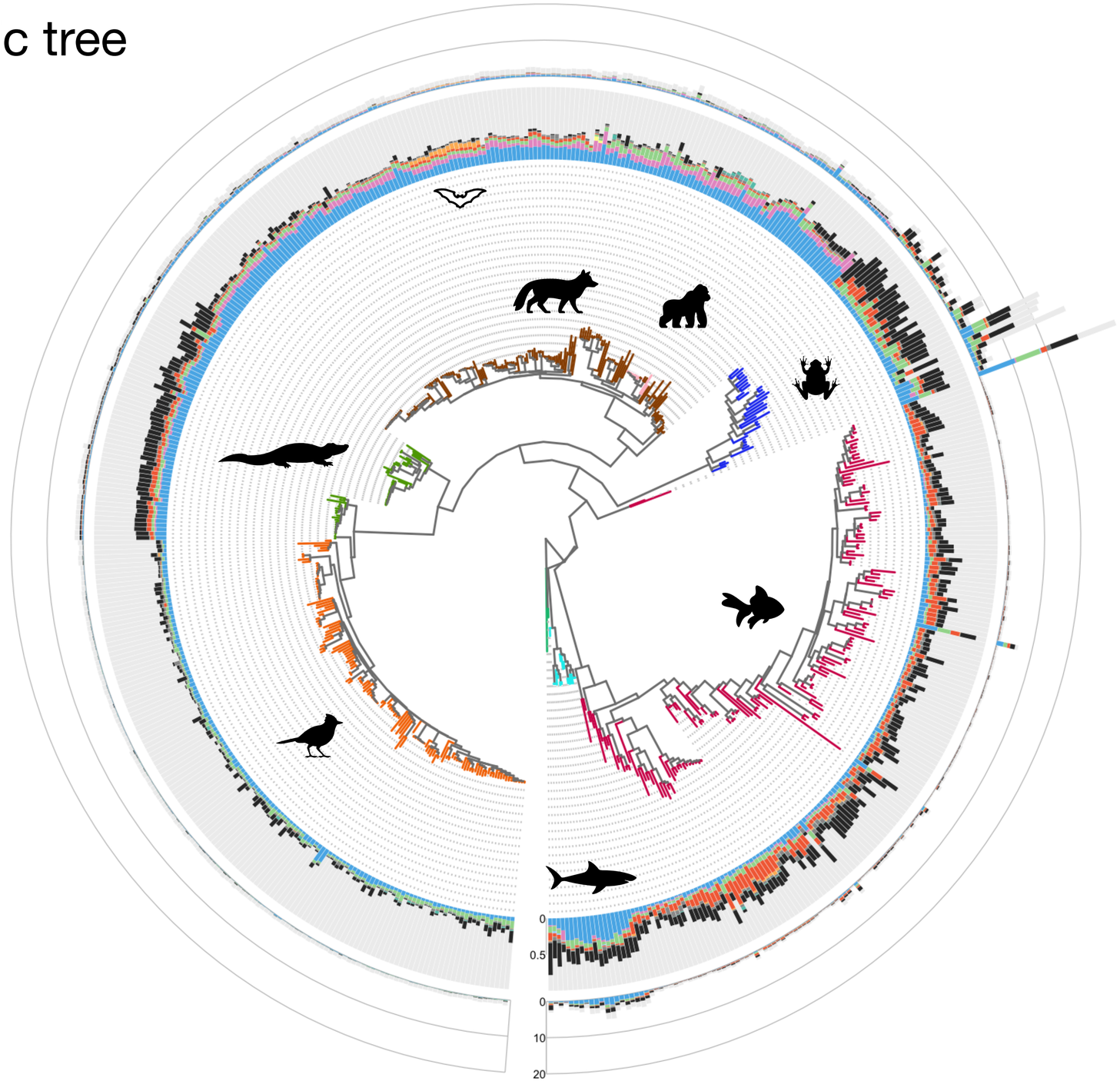
Genome % | DFAM 3.9

0
20
40
60
80

Per-class breakdown


Genome Content (Gbp)
African lungfish
Unmasked
Unknown
Satellite
Simple_repeat
Low_complexity
RC
DNA
PLE
LTR
Retroposon
SINE
LINE
20
10
0
Genome Proportion
0.8
0.6
0.4
0.2
0
sharks
fish
amphibians
primates
mammals
(bats)
reptiles
birds
Cladogram
Tree by Siavash Mirarab

VGP phylogenetic tree



VGPLineage

- amphibians
- birds
- fish
- mammals
- otherChordates
- primates
- reptiles
- sharks
- NA

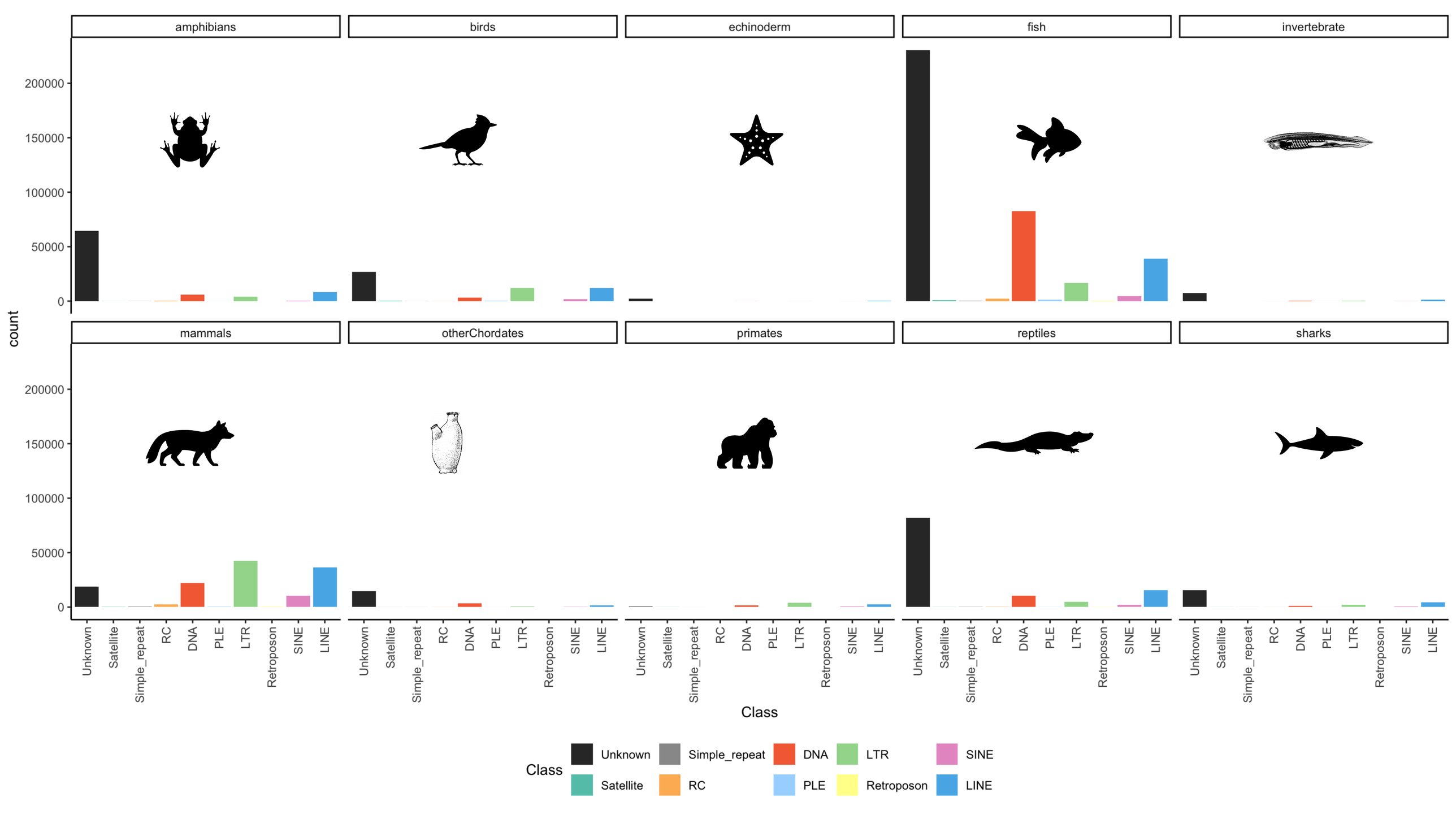
Class

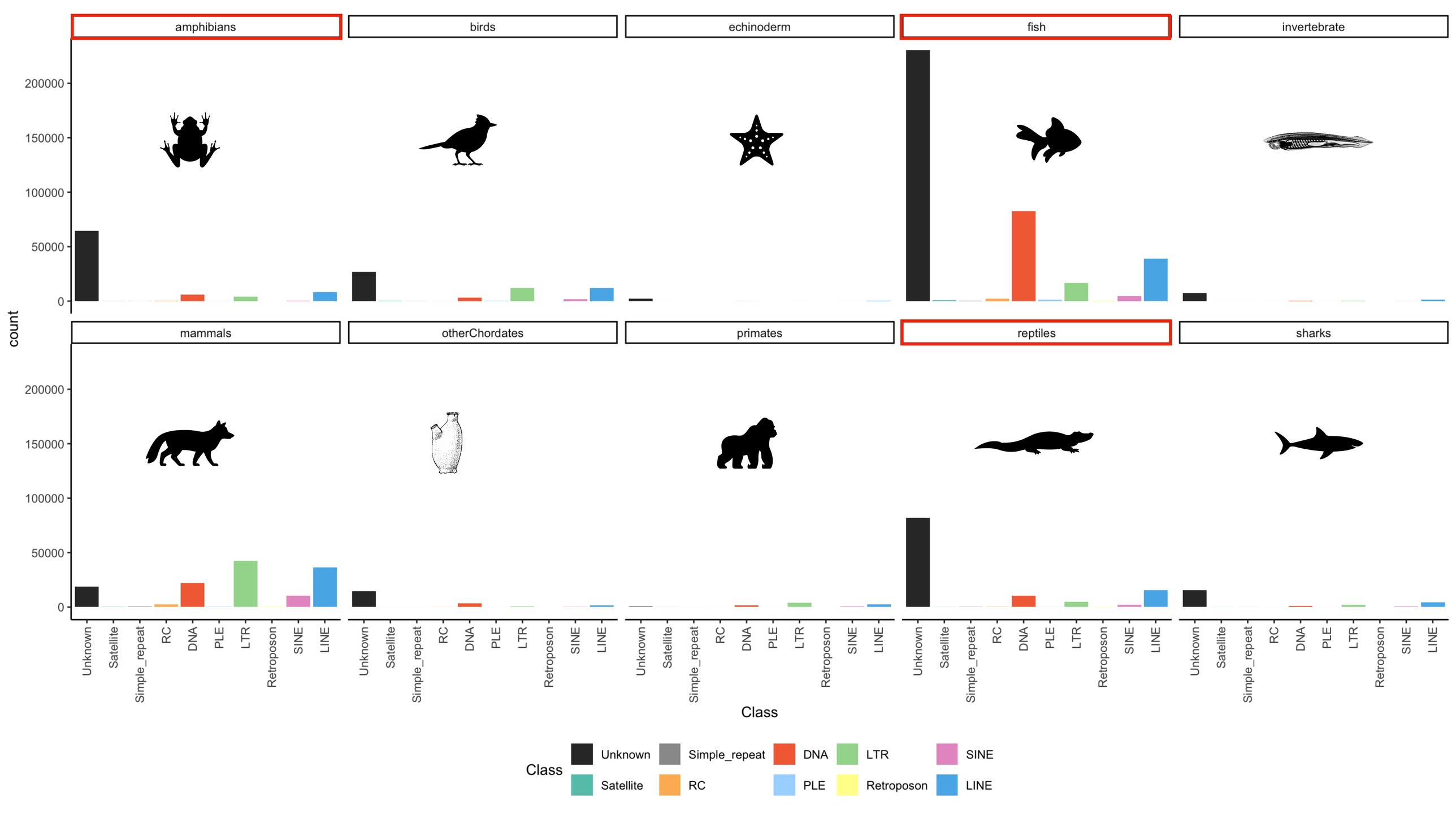
- Unmasked
- Unknown
- Satellite
- Simple_repeat
- Low_complexity
- RC
- DNA
- PLE
- LTR
- Retroposon
- SINE
- LINE

Tree by Siavash Mirarab

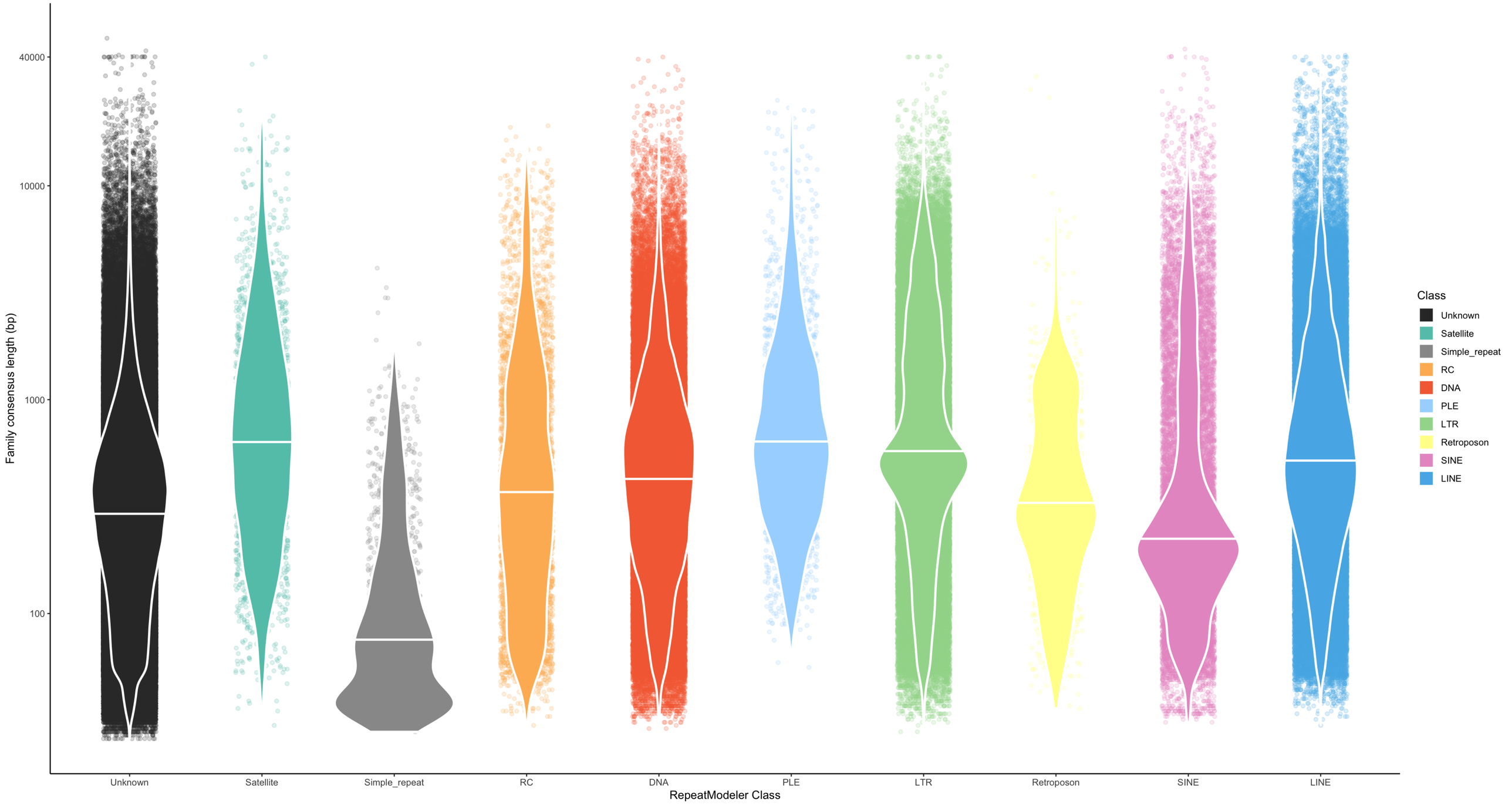
Where there is work to do

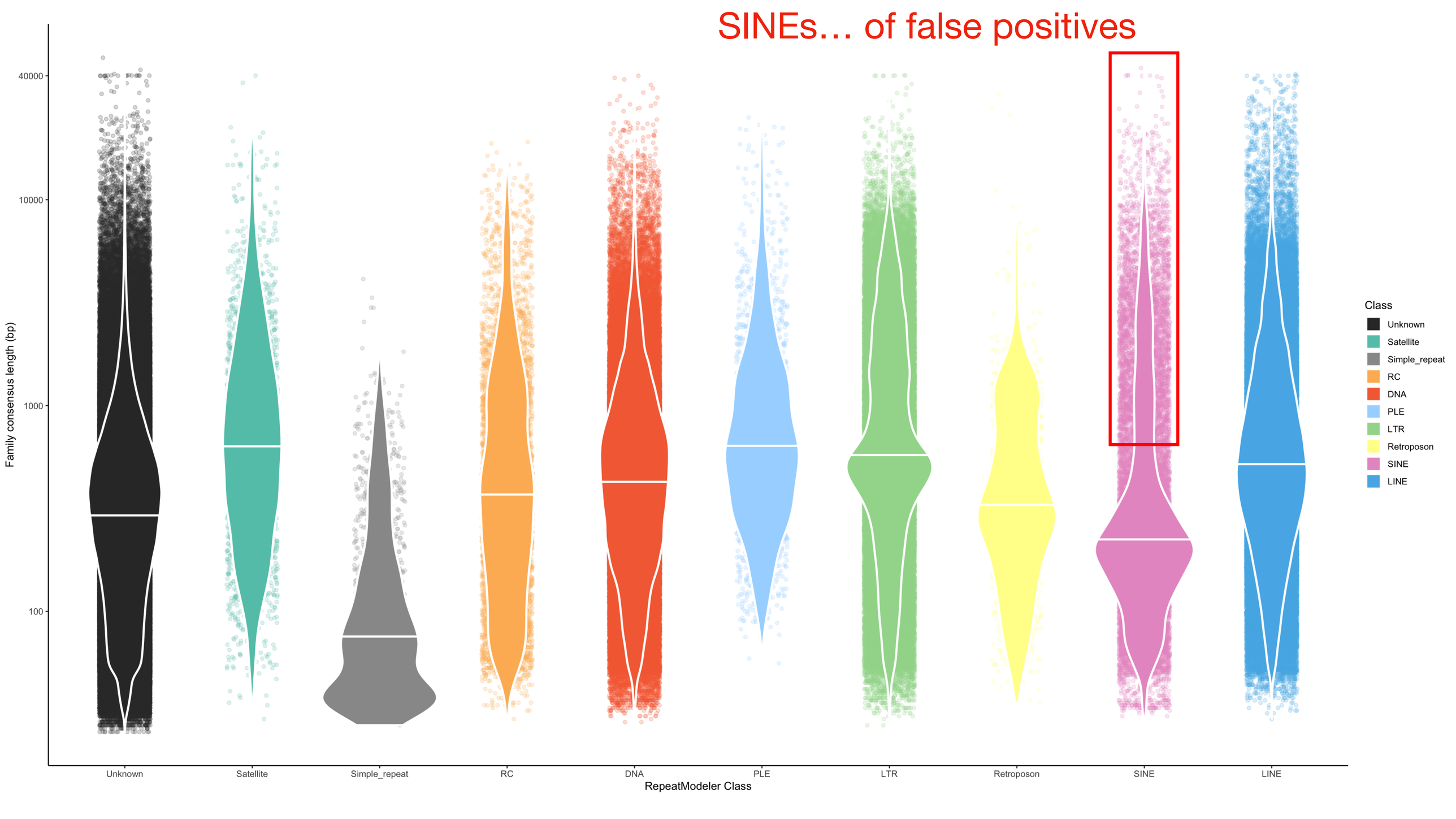
Counts of Unknown families

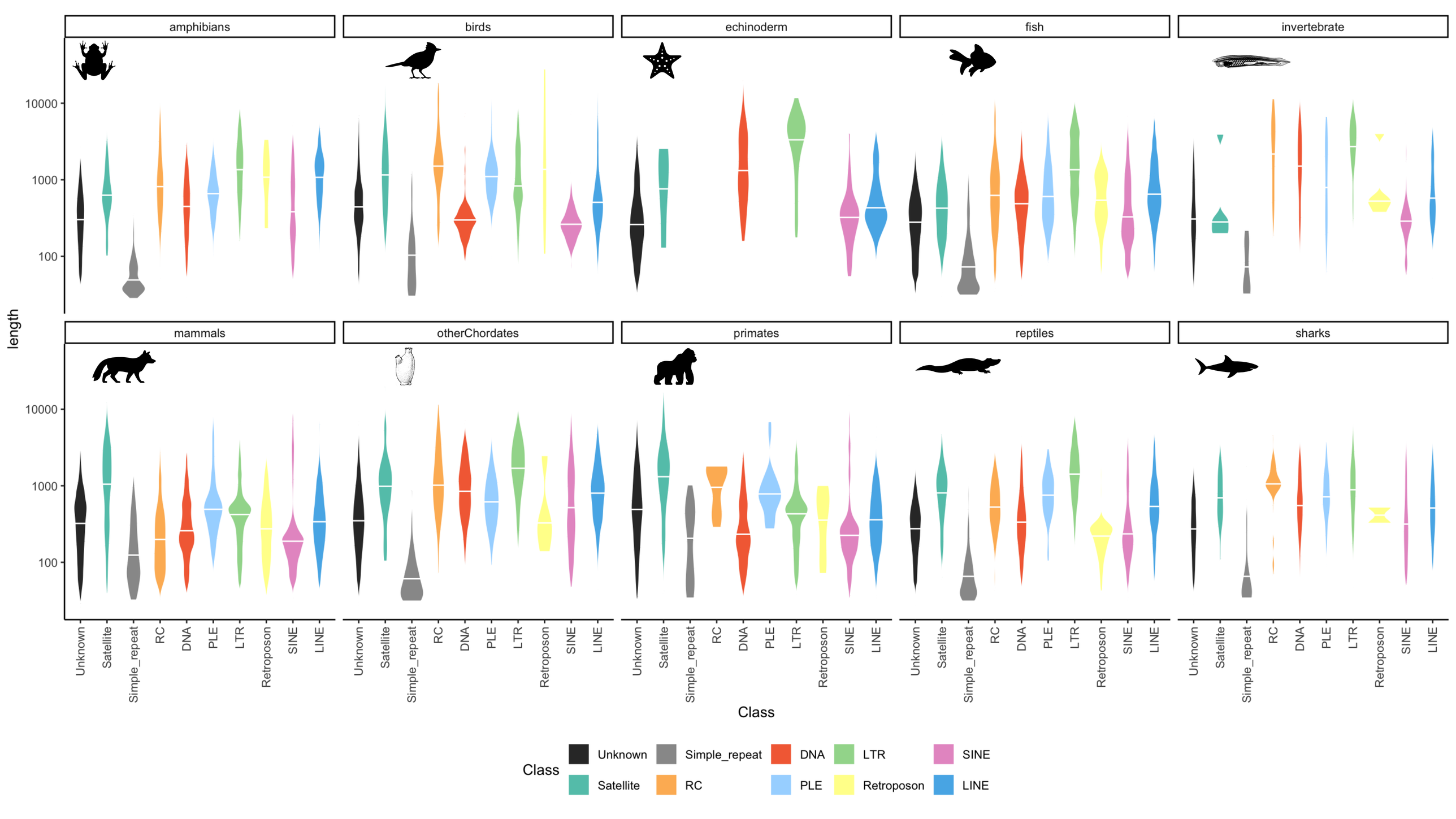

amphibians
birds
echinoderm
fish
invertebrate
mammals
otherChordates
primates
reptiles
sharks
count
200000
150000
100000
50000
0
Unknown
Satellite
Simple_repeat
RC
DNA
PLE
LTR
Retroposon
SINE
LINE
Class


amphibians
birds
echinoderm
fish
invertebrate
mammals
otherChordates
primates
reptiles
sharks
count
200000
150000
100000
50000
0
Unknown
Satellite
Simple_repeat
RC
DNA
PLE
LTR
Retroposon
SINE
LINE
Class

Family consensus length


40000
10000
1000
100
Family consensus length (bp)
Unknown
Satellite
Simple_repeat
RC
DNA
PLE
LTR
Retroposon
SINE
LINE
RepeatModeler Class
Class
Unknown
Satellite
Simple_repeat
RC
DNA
PLE
LTR
Retroposon
SINE
LINE


SINEs… of false positives
Family consensus length (bp)
40000
10000
1000
100
Unknown
Satellite
Simple_repeat
RC
DNA
PLE
LTR
Retroposon
SINE
LINE
RepeatModeler Class
Class
Unknown
Satellite
Simple_repeat
RC
DNA
PLE
LTR
Retroposon
SINE
LINE


amphibians
birds
echinoderm
fish
invertebrate
mammals
otherChordates
primates
reptiles
sharks
length
10000
1000
100
Unknown
Satellite
Simple_repeat
RC
DNA
PLE
LTR
Retroposon
SINE
LINE
Class

Outliers

rnd-5_family-2087

(Unknown)

Castor canadensis

48947 bp

List of domain hits

Name	Accession	Description	Interval	E-value
[+] PHA03142 super family	cl30018	helicase-primase primase subunit BSLF1; Provisional	19198-21156	0e+00
[+] Herpes_UL31 super family	cl27388	Herpesvirus UL31-like protein; This is a family of Herpesvirus proteins including UL31, UL53, ...	46828-47622	1.06e-70
[+] PHA03131 super family	cl33715	dUTPase; Provisional	17020-17871	6.67e-45
[+] Herpes_heli_pri super family	cl27948	Herpesvirus helicase-primase complex component; This family consists of several ...	4864-5169	3.21e-21
[+] Herpes_U30 super family	cl20269	Herpes virus tegument protein U30; This family is named after the human herpesvirus protein, ...	31061-33763	3.34e-173
[+] YqaJ super family	cl09232	YqaJ-like viral recombinase domain; This protein family is found in many different bacterial ...	731-1477	6.69e-88
[+] PHA03142 super family	cl30018	helicase-primase primase subunit BSLF1; Provisional	18716-19153	5.08e-44
[+] Herpes_UL69 super family	cl27025	Herpesvirus transcriptional regulator family; This family includes UL69 and IE63 that are ...	22466-23041	8.15e-20
[+] Herpes_TAF50 super family	cl25754	Herpesvirus transcription activation factor (transactivator); This family includes EBV BRLF1 ...	14537-15520	2.29e-12
[+] Herpes_env super family	cl28122	Herpesvirus putative major envelope glycoprotein; This family consists of probable major ...	45341-45664	8.25e-09
[+] Herpes_teg_N super family	cl04795	Herpesvirus tegument protein, N-terminal conserved region;	33926-34036	2.03e-08
[+] PHA03311	PHA03311	helicase-primase subunit BBLF4; Provisional	7707-10133	0e+00
[+] PHA03312 super family	cl28610	helicase-primase subunit BBLF2/3; Provisional	3156-4508	1.37e-76
[+] YqaJ super family	cl09232	YqaJ-like viral recombinase domain; This protein family is found in many different bacterial ...	153-722	2.35e-57
[+] Herpes_teg_N super family	cl04795	Herpesvirus tegument protein, N-terminal conserved region;	34029-34448	2.00e-28
[+] Herpes_env super family	cl28122	Herpesvirus putative major envelope glycoprotein; This family consists of probable major ...	45858-46658	1.60e-21
[+] PHA03142 super family	cl30018	helicase-primase primase subunit BSLF1; Provisional	21111-21260	7.10e-10
[+] RNR_PFL super family	cl38938	Ribonucleotide reductase and Pyruvate formate lyase; Ribonucleotide reductase (RNR) and ...	27771-29636	7.08e-173
[+] Herpes_UL49_1 super family	cl03898	UL49 family; Members of this family, found in several herpesviruses, include EBV BFRF2 and ...	42681-43349	1.22e-59
[+] Herpes_U44 super family	cl28003	Herpes virus U44 protein; This is a family of proteins from dsDNA beta-herpesvirinae and ...	18069-18671	2.72e-53
[+] Herpes_UL33 super family	cl28035	Herpesvirus UL33-like protein; This is a family of Herpesvirus proteins including UL33,UL51. ...	44844-45059	1.18e-19
[+] BI-1-like super family	cl00473	BAX inhibitor (BI)-1/YccA-like protein family; Mammalian members of the BAX inhibitor (BI)-1 ...	47853-48464	1.35e-17
[+] Herpes_glycop super family	cl28128	Herpesvirus glycoprotein M; The herpesvirus glycoprotein M (gM) is an integral membrane ...	2070-2603	2.10e-04
[+] Herpes_UL6 super family	cl20225	Herpesvirus UL6 like; This family consists of various proteins from the herpesviridae that are ...	6068-7681	0e+00
[+] Herpes_U34 super family	cl29790	Herpesvirus virion protein U34; The virion proteins in this family include membrane ...	44246-44776	4.09e-50
[+] Herpes_glycop super family	cl28128	Herpesvirus glycoprotein M; The herpesvirus glycoprotein M (gM) is an integral membrane ...	2681-2974	2.37e-07
[+] UDG-like super family	cl00483	uracil-DNA glycosylases (UDG) and related enzymes; Uracil-DNA glycosylases (UDGs) initiate ...	11431-12189	1.21e-134
[+] PHA03263 super family	cl19801	Capsid triplex subunit 1; Provisional	29995-30993	9.21e-128
[+] Ribonuc_red_sm	pfam00268	Ribonucleotide reductase, small chain;	26671-27423	2.64e-60
[+] Herpes_DNAp_acc super family	cl15563	Herpes DNA replication accessory factor; Replicative DNA polymerases are capable of ...	25063-26358	2.46e-56
[+] Herpes_UL7 super family	cl28121	Herpesvirus UL7 like; This family consists of various functionally undefined proteins from the ...	5266-5898	3.08e-44
[+] Herpes_BBRF1 super family	cl27986	BRRF1-like protein; Family of herpesvirus proteins including Epstein-barr virus protein BBRF1.	13465-14208	1.75e-17
[+] Herpes_UL73 super family	cl29774	UL73 viral envelope glycoprotein; This family groups together the viral proteins BLRF1, U46, ...	16648-16851	5.75e-07

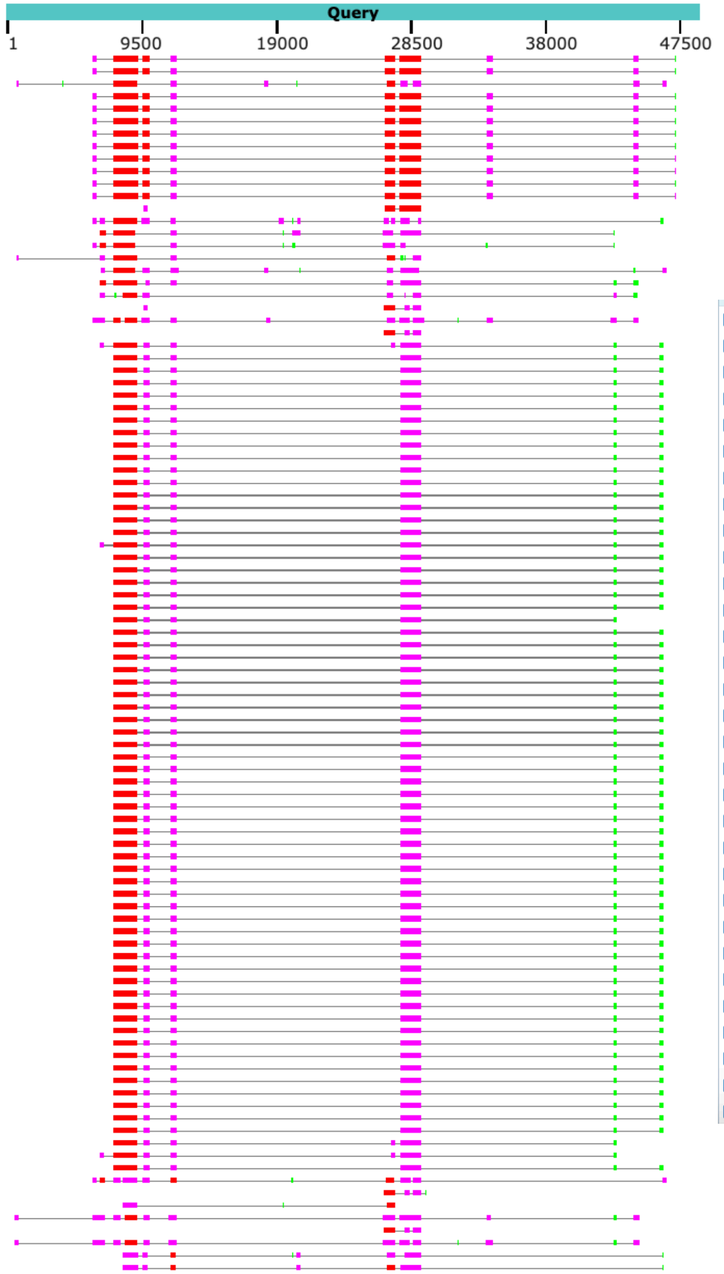
rnd-5_family-2087

(Unknown)

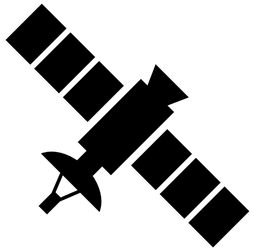
Castor canadensis

48947 bp

Distribution of the top 664 Blast Hits on 100 subject sequences



Bovine gammaherpesvirus 4 isolate 07-435, partial genome	Bovine gammaherpesvirus 4	433	1884	13%	3e-114	66.07%	108381	OQ709765.1
Bovine gammaherpesvirus 4 isolate 10-154, partial genome	Bovine gammaherpesvirus 4	419	1935	13%	2e-110	65.90%	108667	OQ709766.1
Harp seal herpesvirus isolate FMV04-1493874, partial genome	Harp seal herpesvirus	416	1494	11%	2e-109	66.05%	117276	NC_055139.1
Bovine gammaherpesvirus 4 SG20 DNA, complete genome	Bovine gammaherpesvirus 4	413	1924	13%	3e-108	65.78%	108819	LC785510.1
Bovine gammaherpesvirus 4 isolate HB-ZJK long unique region, complete sequence	Bovine gammaherpesvirus 4	413	1924	13%	3e-108	65.78%	109811	OP631674.1
Bovine gammaherpesvirus 4 isolate SD16-49, partial genome	Bovine gammaherpesvirus 4	413	1952	13%	3e-108	65.78%	108478	MN551084.1
Bovine gammaherpesvirus 4 isolate BoHV-4/9746/Bursa-TR/Cattle/Lung/2018, partial genome	Bovine gammaherpesvirus 4	413	1938	13%	3e-108	65.78%	86870	PV135575.1
Bovine herpesvirus 4 strain V.test long unique region, complete sequence	Bovine gammaherpesvirus 4	413	1938	13%	3e-108	65.78%	108241	JN133502.1
Bovine gammaherpesvirus 4 isolate SD16-38, partial genome	Bovine gammaherpesvirus 4	413	1970	13%	3e-108	65.78%	108669	MN551083.1
Bovine gammaherpesvirus 4 strain 85/16 TV BoGHV-4, complete genome	Bovine gammaherpesvirus 4	413	1961	13%	3e-108	65.78%	108868	PQ557568.1
Bovine herpesvirus 4 isolate FMV09-1180503, partial genome	Bovine gammaherpesvirus 4	413	1952	13%	3e-108	65.78%	108349	KC999113.1
Bovine herpesvirus 4 long unique region, complete sequence	Bovine gammaherpesvirus 4	405	1961	13%	4e-106	66.93%	108873	NC_002665.1
Bovine gammaherpesvirus 4 isolate Tver1 hypothetical protein gene, partial cds	Bovine gammaherpesvirus 4	400	683	4%	2e-104	66.86%	3330	OQ941875.1
Rhinolophus gammaherpesvirus 1 BV1 DNA, complete genome	Rhinolophus gammaherpesvir…	374	1701	12%	7e-97	66.07%	147790	NC_040539.1
Saimiriine herpesvirus 2 complete genome	Saimiriine gammaherpesvirus 2	374	1226	11%	7e-97	66.10%	112930	NC_001350.1
Saimiriine herpesvirus 2 complete L-DNA sequence, strain C488	Saimiriine gammaherpesvirus 2	352	1331	10%	2e-90	65.82%	113027	AJ410493.1
Felis catus gammaherpesvirus 1 isolate 31286, complete genome	Felid gammaherpesvirus 1	350	1164	9%	3e-89	65.47%	122522	NC_028099.1
MAG TPA_asm: Herpesvirus ursus25482 isolate Herpes25482 genomic sequence	Herpesvirus ursus25482	333	1382	12%	2e-84	65.43%	112201	BK066690.1
Ateline herpesvirus 3 complete genome	Ateline gammaherpesvirus 3	292	1143	11%	7e-72	64.98%	108409	NC_001987.1
Common bottlenose dolphin gammaherpesvirus 1 strain Sarasota, complete genome	Common bottlenose dolphin g…	288	1117	7%	9e-71	66.70%	167212	NC_035117.1
MAG: Bat gammaherpesvirus 1 isolate HB2020_015_1367, partial genome	Bat gammaherpesvirus 1	261	543	4%	1e-62	67.67%	7374	OR998733.1
Marmot herpesvirus 1 strain HW99, complete genome	Marmot herpesvirus 1	259	1820	15%	4e-62	66.60%	154413	OK337615.1
MAG: Bat gammaherpesvirus 1 isolate HB2020_019_45214, partial genome	Bat gammaherpesvirus 1	257	541	4%	1e-61	67.54%	5370	OR998734.1
MAG: Human gammaherpesvirus 8 isolate U210-B, partial genome	Human gammaherpesvirus 8	255	1127	10%	5e-61	64.51%	137780	MZ923817.1
Human gammaherpesvirus 8 isolate FNL0054, partial genome	Human gammaherpesvirus 8	252	851	9%	6e-60	64.41%	136987	OR829374.1
Human gammaherpesvirus 8 isolate FNL0046, partial genome	Human gammaherpesvirus 8	251	850	9%	6e-60	64.49%	137074	OR829366.1
Human gammaherpesvirus 8 isolate FNL0033, partial genome	Human gammaherpesvirus 8	251	855	9%	6e-60	64.49%	136757	OR829355.1
Human gammaherpesvirus 8 isolate FNL0040, partial genome	Human gammaherpesvirus 8	251	855	9%	6e-60	64.49%	137030	OR829362.1
Human gammaherpesvirus 8 isolate FNL0034, partial genome	Human gammaherpesvirus 8	251	855	9%	6e-60	64.49%	137130	OR829356.1
Human gammaherpesvirus 8 isolate FNL0055, partial genome	Human gammaherpesvirus 8	251	855	9%	6e-60	64.49%	136736	OR829375.1
Human gammaherpesvirus 8 strain UNC_BCBL-1, complete genome	Human gammaherpesvirus 8	251	850	9%	6e-60	64.49%	137969	MZ712172.1



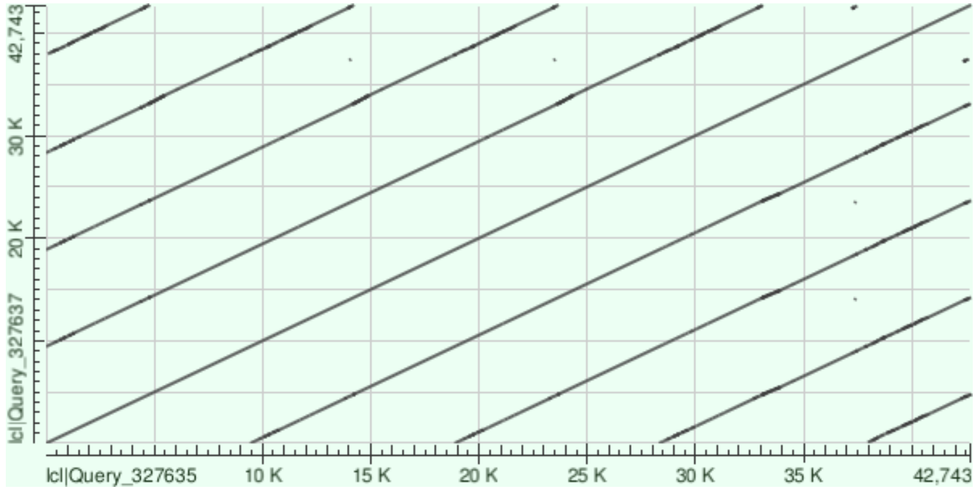
Satellites?

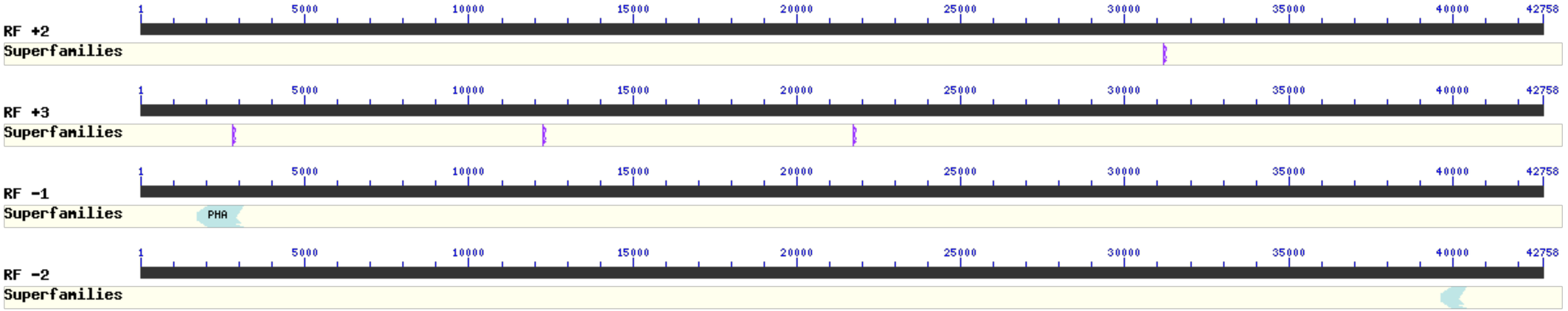
Segmental Duplication?

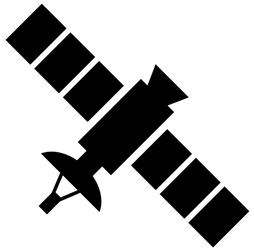
Desmodus rotundus

rnd-5_family-6186#Unknown

42743 bp







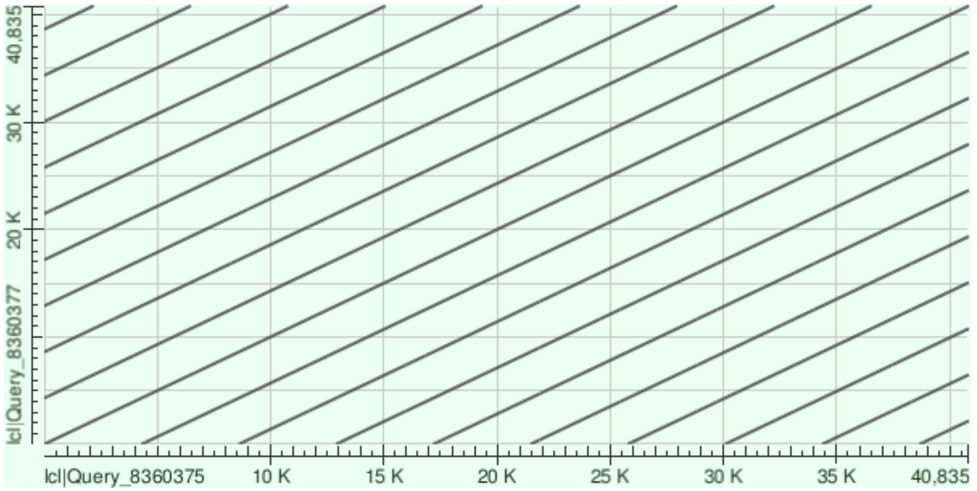
Satellites?

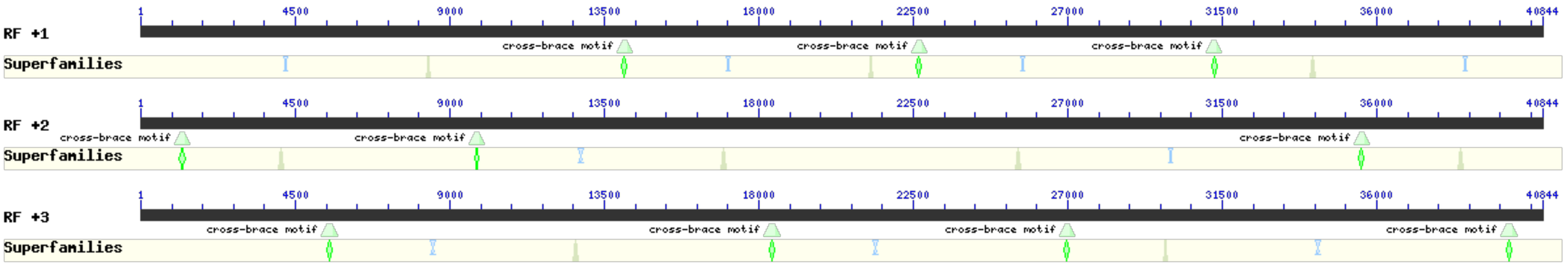
Segmental Duplication?

Notamacropus eugenii

rnd-5_family-3860

40835 bp





Future Directions

- Distribution of copy numbers / Model support
- Redundancy evaluation
- Fragmentation *vs.* over-extension?
- Primary *vs.* Secondary assemblies

```
~/tools/mmseqs/bin/mmseqs easy-cluster
all.primary.repeatModeler.fasta all.primary.clustR tmp --cov-
mode 1 -c 0.8 --min-seq-id 0.8 -s 100 --exact-kmer-matching 1
--threads 8
```

```
===== Cluster Summary =====

Total number of clusters: 553555

Total number of singletons: 461712

Average cluster size: 1.51
```

CG General Info Analysis

376 497 Nodes

284 654 Links

rnd-5_family-1878_GCF_036370855.1

Selected node

0 Input links

831 Output links

NODE DATA

ID
rnd-5_family-1878_GCF_036370855.1

ACCESSION
GCF_036370855.1

FAMILY
rnd-5_family-1878

CLASS
DNA

FOUND 831 RECORDS

Open table

EXPORT SELECTED DATA

Records Metadata

save current layout