Workshop Part III: EvoGeno Methods

Comparative Genomics and Phylogenomics

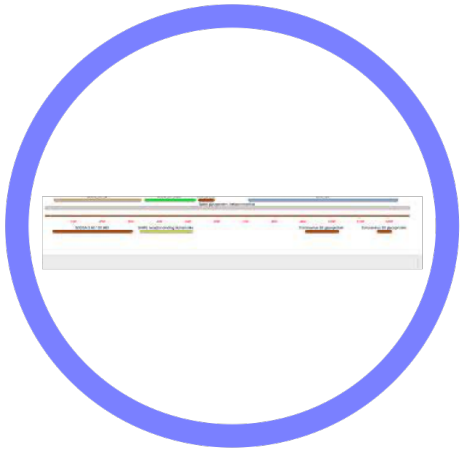
Yocelyn Gutiérrez Guerrero

ygutierrezgro@berkeley.edu

May 05, 2023

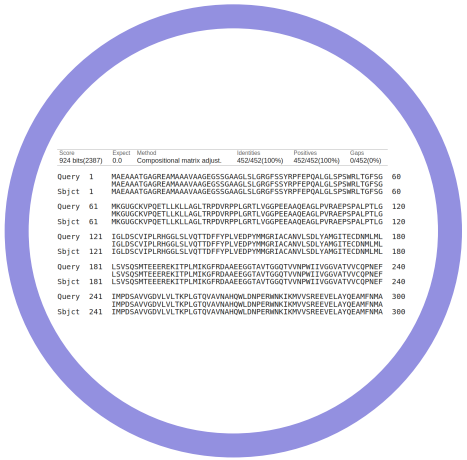
Overview

First Part



After assembly

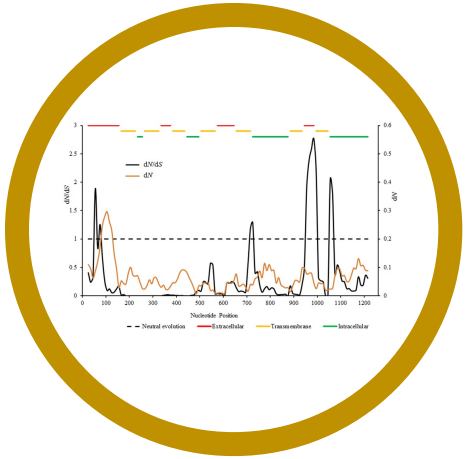
Databases



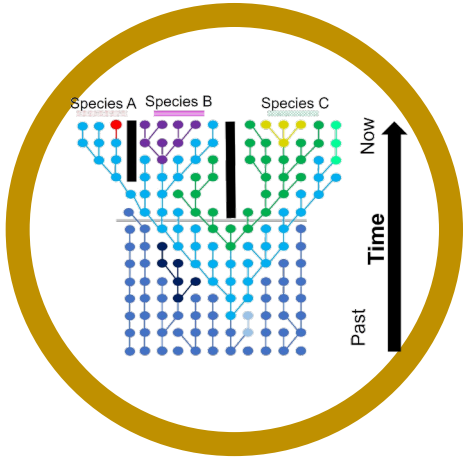
Orthologous identification

Second Part

Phylogenomics

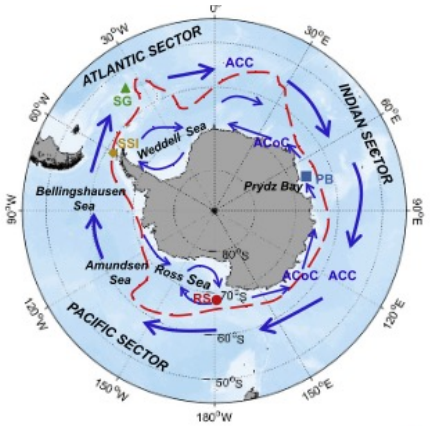


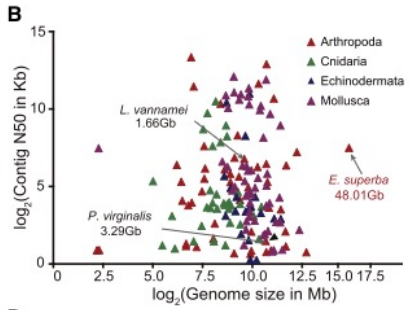
Selection Test

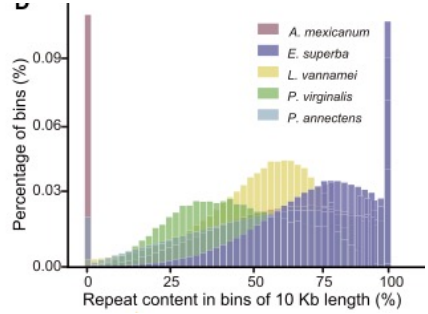


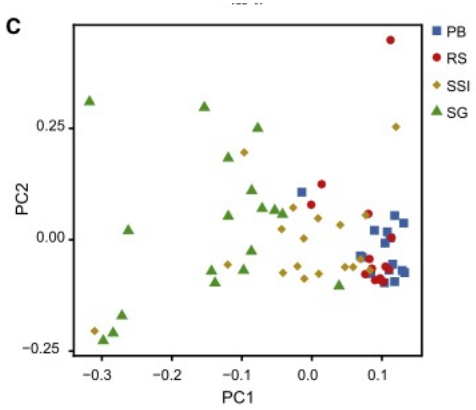
Incomplete lineage Sorting

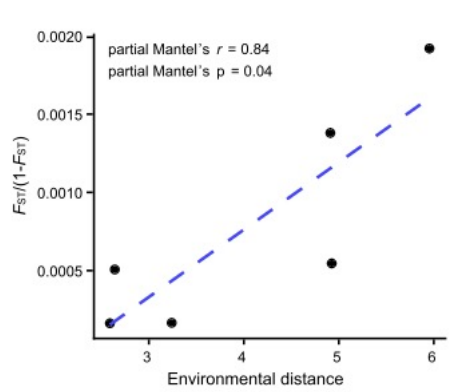
From Population to Comparative Genomics

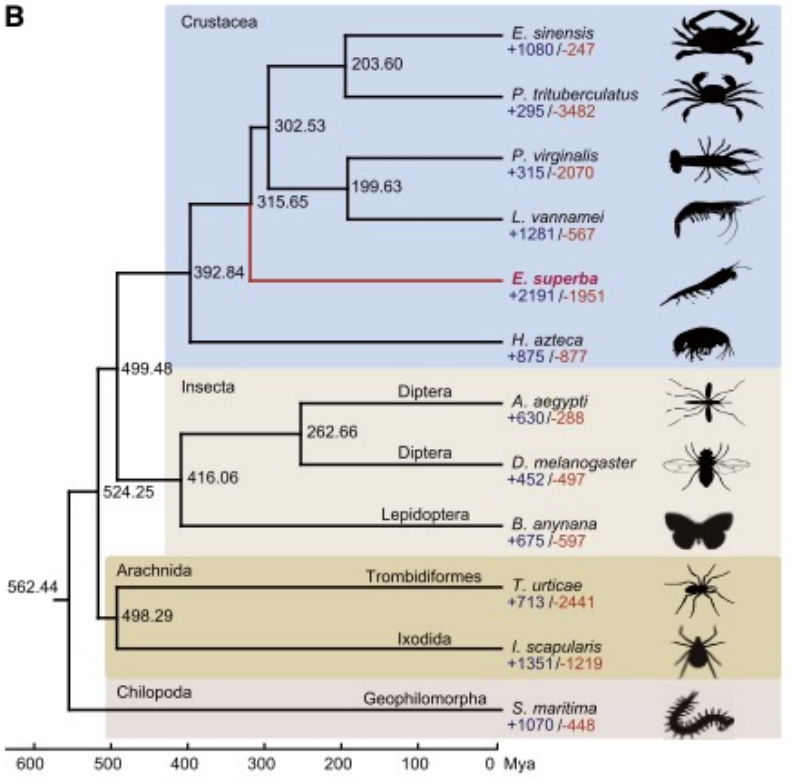


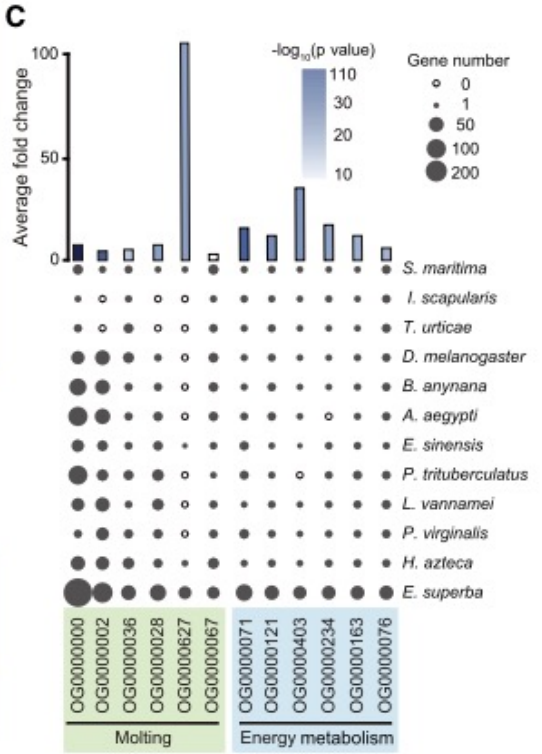












Shao et al., 2023

Comparative Genomics: Genomic features of different organism are compared

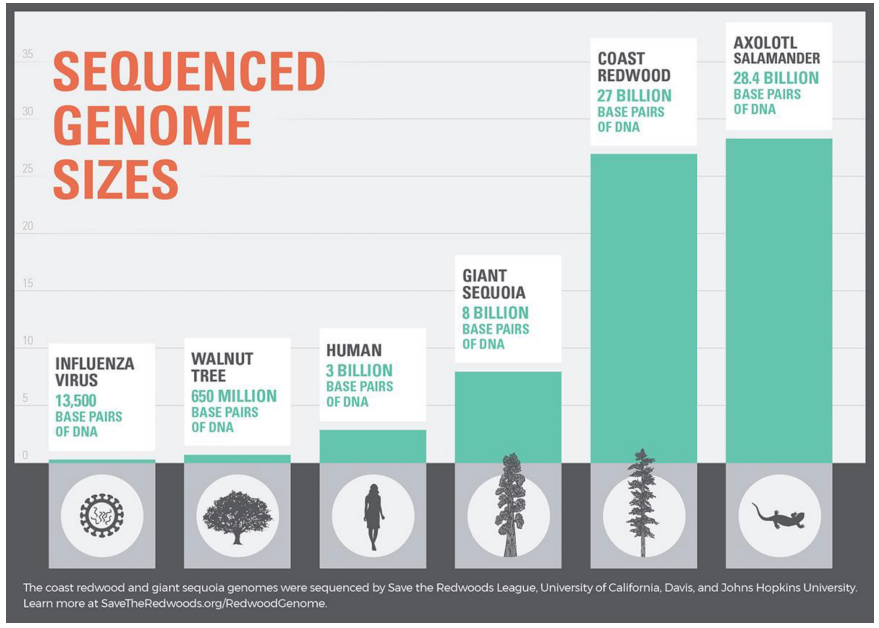
-Genome architecture

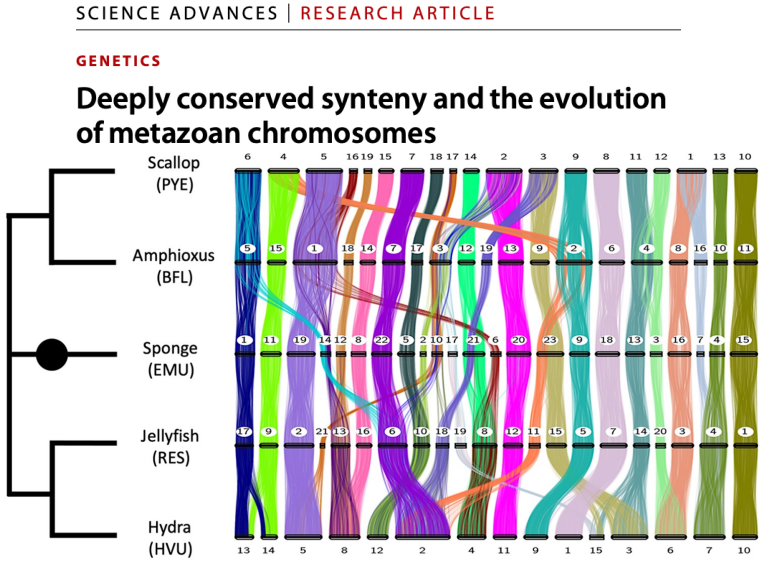
-Common and divergent features

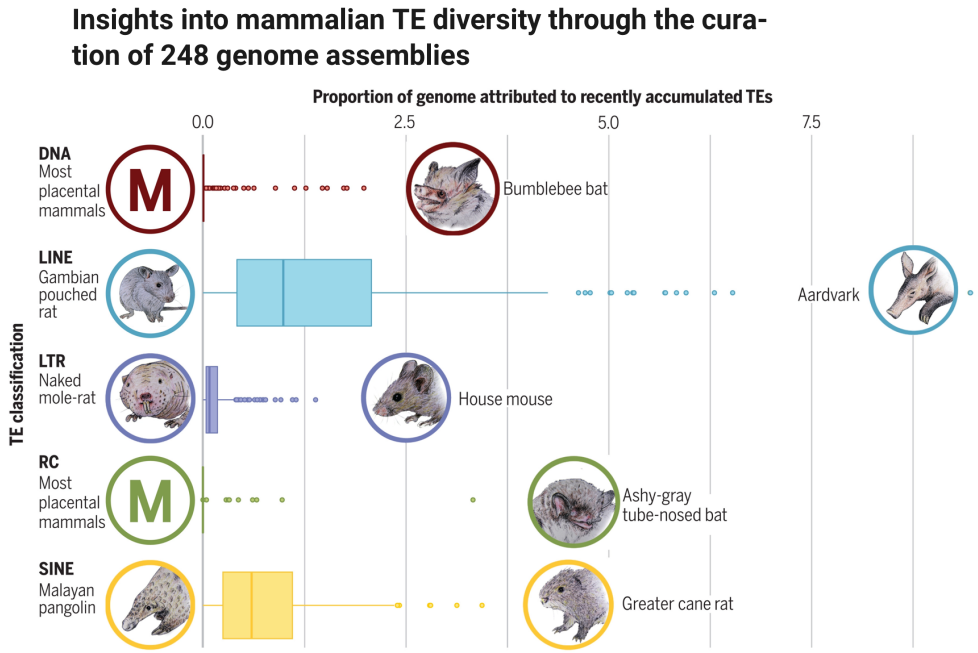
-Gene evolutionary rates

-Selection process

-Non coding elements evolution







earth microbiome project

CALIFORNIA CONSERVATION GENOMICS PROJECT CCGP

HMP

Avian Phylogenomics Project

UK 10K
Rare Genetic Variants in Health and Disease

VGP VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

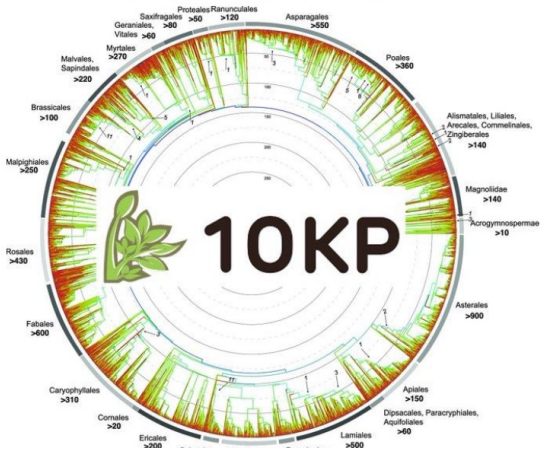
ENCODE

H3Africa
Human Heredity & Health in Africa

TARA OCEANS

BAT 1K

PCAWG
PanCancer Analysis of WHOLE GENOMES

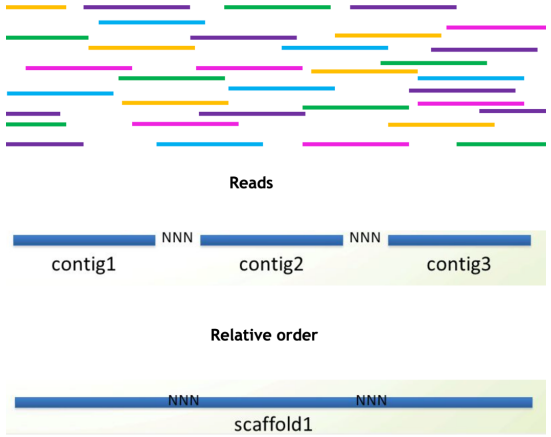


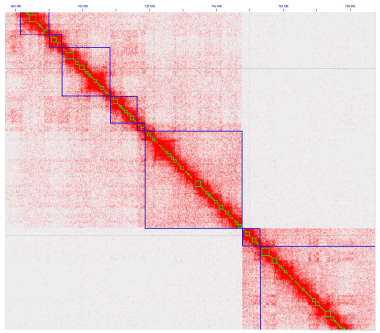
The 1KP Project

Plants transcriptome

1001 Genomes
A Catalog of Arabidopsis thaliana Genetic Variation

After genome assembly





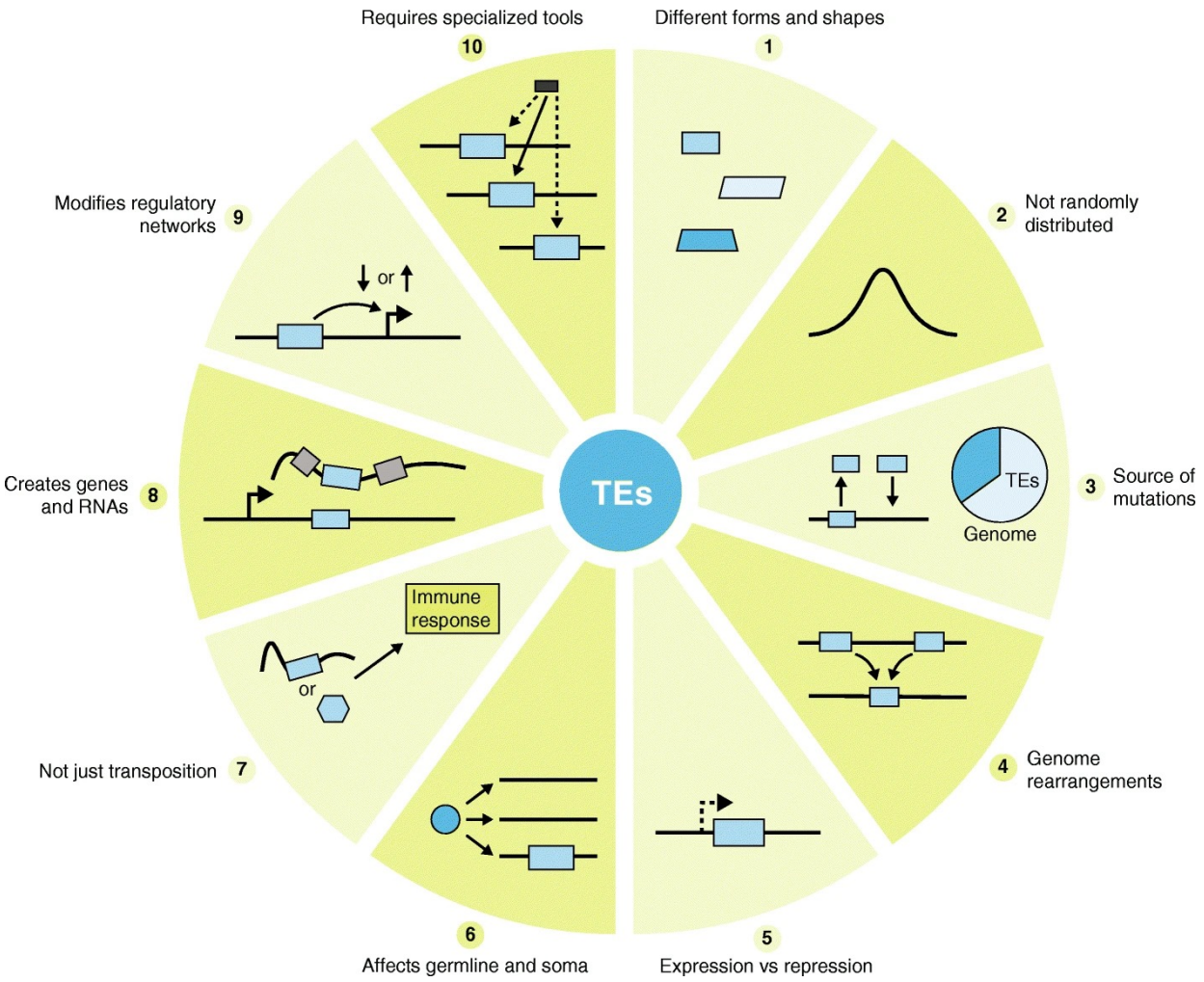
Chromosome

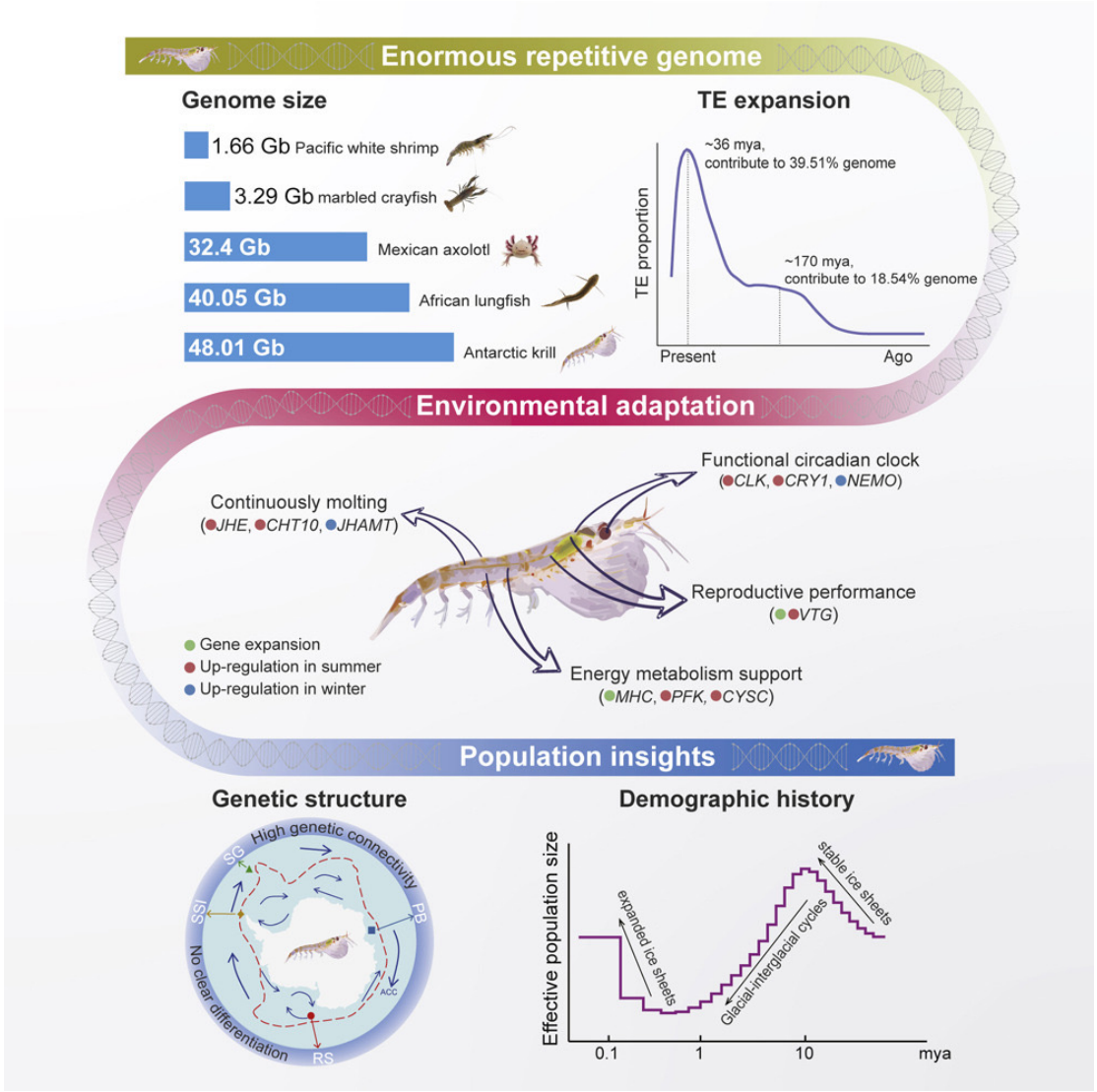
Repetitive elements identification

Mask genome

Gene annotation and genome features

Repetitive elements content

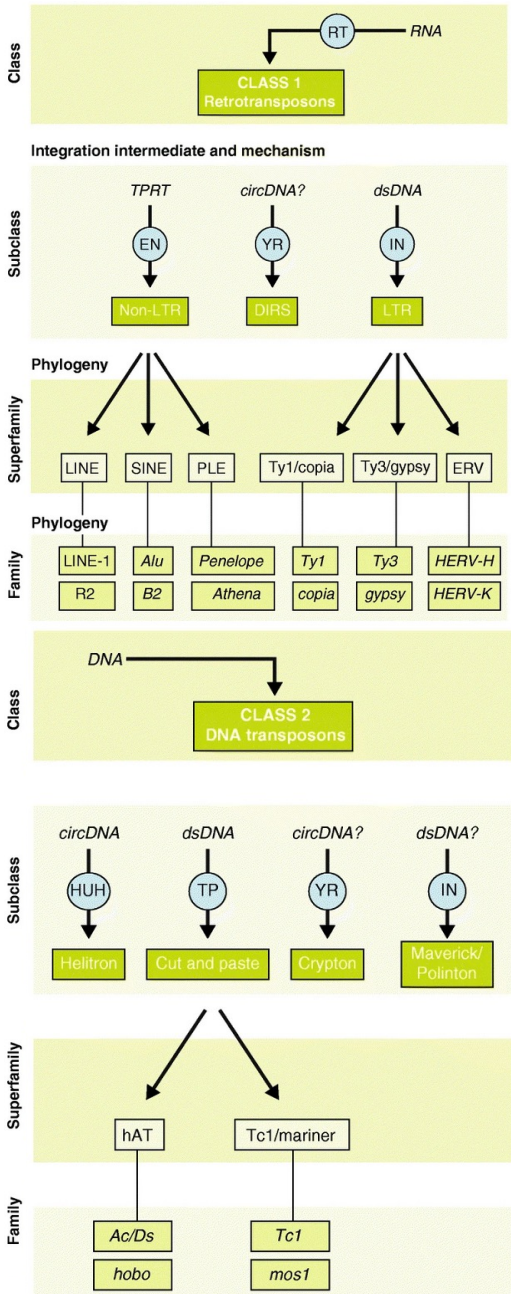




Tandem Repeats (TE) is over 92%, two expansion events

Elliot and Gregory, 2009; Shao et al., 2023

Repetitive elements identification



REPET

Create a Database for RepeatModeler

Perl

RepeatMasker

RepeatModeler

RECON (*de novo* Repeat Finder)

RepeatScout (*de novo* Repeat Finder)

TRF (Tandem Repeat Finder)

RMBlast

LTR_retriever (LTR discovery)

+

Dfam

giri REPBASE

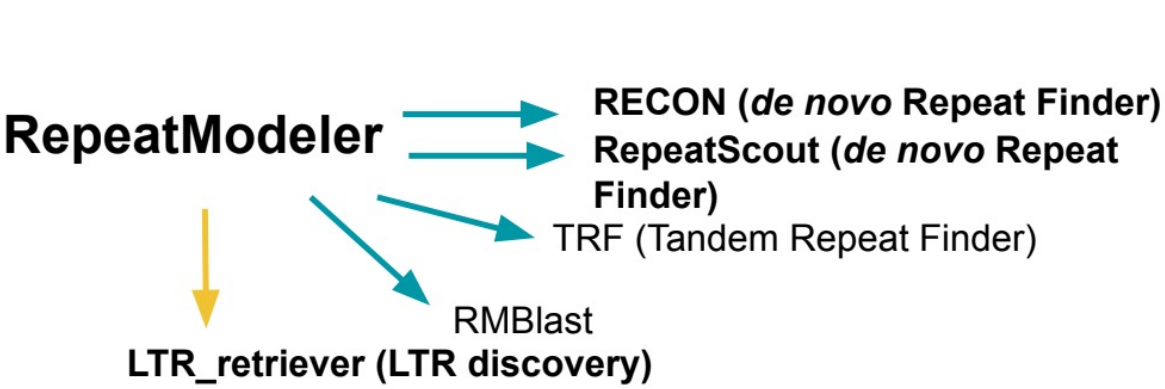
Fragments identified in sequence	Possible families
	MER58 or MER58B
	MER58 only
	MER58, MER58B or MER58A

```
# STOCKHOLM 1.0
#=GF ID    rnd-1_family-20
#=GF DE    RepeatModeler Generated - rnd-1_family-20, RepeatScout: [ Index = R=13,
RS Size = 240, Refiner Input Size = 100, Final Multiple Alignment Size = 100 ]
#=GF SQ    100
#=GC RF    xxxxx.xx.xxxxx.xx.xxxxxx.x..xxxx.xxx.........xxxx........xx..xx.xxxxxxx.
.xxxxxxx...........xxx..x.x.xxx.xx...xxxxxxx.xxxxxxxxxxxxxxxxx..xxxxx...x.xxx..x.
xxxx..xxxxx.xxxxx.x.xxxxxxx.xxxx
NC_044211.1_RaGOO:30307278-30307399    TGAGG.GG.CTGGA.GC.GTGTCC.A..GAGA.AGG.......
.GCAA........CA..GA.GCTGGGG..AAGGGTC............TGG..T.G.CAC.AA...GTCTTAT.GAGGAGTGG
CTGAGGGAG..CTGGG...G.GTA..T.TTAT..CTTGG.AGAAA.A.GGAGGATC.AGGG
NC_044211.1_RaGOO:69032180-69032300    TGAGG.AG.CTGGA.GT.GTGTCC.A..GAGA.AGG.......
.GCAA........TG..GA.ACGGGTG..AAGGGTC............TGG..A.G.CAC.AA...GTCCTGT.GAGGAGCAG
CTCAGGGAG..CTGAA...G.TTC..T.TTAG..CC.AG.AGAAA.A.GGAAGCTC.AGGG
NC_044213.1_RaGOO:139863354-139863474    TGAGG.TG.CTGGA.GT.GTGTCC.A..GAGA.AGG......
...A.AA........CA..GA.GCTGGGG..AAGGGGC............TGG..A.G.CAC.AG...GGCTGAT.GAGGAGC
AGCTGAGGGAG..CTGGG...G.GTG..C.TCAG..CCTGG.AGGAA.A.GGAGGCTC.AGGG
NC_044214.1_RaGOO:54951091-54951212    TGTGG.TG.CTGGA.TT.GTGCCC.A..GAGA.AGG.......
.GCAA........TG..GA.TGTGGTG..AAGGGTC............TGG..A.G.CAC.AA...GCCCTAC.GAGGAGCAC
TTGAGGGAG..CTGGG...G.GTG..T.TTAC..TCTGC.AGAAA.A.GGAGGCCC.TGGG
NC_044213.1_RaGOO:64961061-64961182    TGAGG.GG.CTGGA.GC.TTGTCC.A..GAGA.ATG.......
families.stk
```

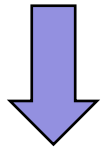
de novo Identification with Repeat Modeler

```
>rnd-1_family-29 ( RepeatScout Family Size = 117, Final Multiple Alignment Size = 1
00, Localized to 32 out of 32 contigs )
TGTCGCAGACATNTTTTATGAAAAATCCTTTCCTTAGGATTTTTCCTCCTGAGAAGCTGAGAGGCCTCAGAAACAAAATGNAA
ACAATGGTTATCTGCTGCTGTGGAATGCAACAGGTGCATCTGTGATTGGTCTCGTGTGGTTGTTTNTAATTAATGGCCAATCA
CGGCCAGCTGGCTCGGACTGAGAGTCCGAGCCACGAGCCTTTGTTATCATTCTTTTCTTTTCTATTCTTAGCTAGCCTTCTGA
TGAAATCCTTTCTTCTATTCTTTTAGTATAGTTTTAATATAATATATATCATAAAATAATAAATCAGCCTTCTGAAACATGGA
GTCAGATCCTCGTCTCTTCCCTCATCCTGGGACCCCTGCGAACACCGTCACA
```

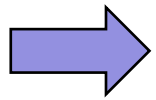
Consensus sequences identified



Genome assembly



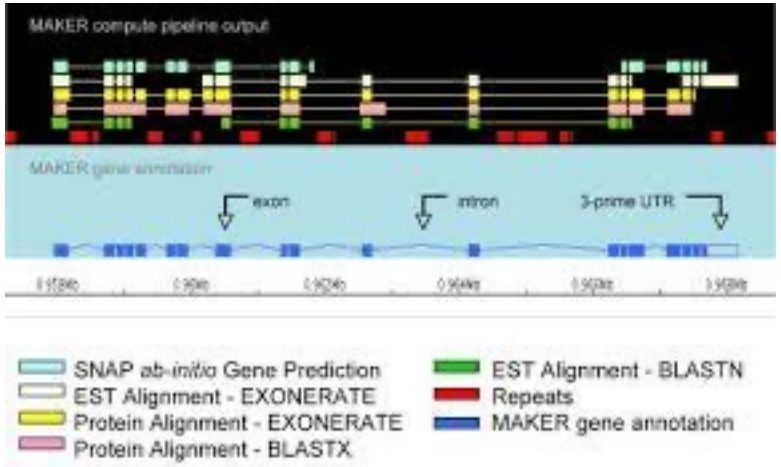
```
##date 2021-02-22
##sequence-region RaGOO_Dbar-Tgut_500.fasta
Chr0_RaGOO      RepeatMasker    similarity      860     1125    26.2    +       .       Target "Motif:rnd-5_family-980" 6
269
Chr0_RaGOO      RepeatMasker    similarity      1482    1584    22.6    +       .       Target "Motif:rnd-5_family-980" 65
9 761
Chr0_RaGOO      RepeatMasker    similarity      1516    1788    28.6    -       .       Target "Motif:rnd-2_family-34" 235
 868
Chr0_RaGOO      RepeatMasker    similarity      1789    2276    21.1    +       .       Target "Motif:rnd-5_family-2770" 4
6 511
Chr0_RaGOO      RepeatMasker    similarity      2277    2365    28.6    -       .       Target "Motif:rnd-2_family-34" 23
234
Chr0_RaGOO      RepeatMasker    similarity      2820    2884    20.2    +       .       Target "Motif:A-rich" 1 62
Chr0_RaGOO      RepeatMasker    similarity      3358    3402    19.4    +       .       Target "Motif:A-rich" 1 43
Chr0_RaGOO      RepeatMasker    similarity      4162    4224    26.5    +       .       Target "Motif:(GCA)n" 1 67
Chr0_RaGOO      RepeatMasker    similarity      4534    4576     9.3    -       .       Target "Motif:rnd-6_family-3719" 3
 45
Chr0_RaGOO      RepeatMasker    similarity      4700    4854    20.6    -       .       Target "Motif:rnd-2_family-7" 142
336
Chr0_RaGOO      RepeatMasker    similarity      4868    4907    15.0    -       .       Target "Motif:rnd-6_family-5569" 1
99 238
Chr0_RaGOO      RepeatMasker    similarity      5171    5259    27.0    -       .       Target "Motif:rnd-5_family-247" 15
4 242
Chr0_RaGOO      RepeatMasker    similarity      6335    6365    15.5    +       .       Target "Motif:(TA)n" 1 30
Chr0_RaGOO      RepeatMasker    similarity      6390    6529    17.4    +       .       Target "Motif:rnd-3_family-51" 367
 508
Chr0_RaGOO      RepeatMasker    similarity      9674    9708     9.1    +       .       Target "Motif:(A)n" 1 35
Chr0_RaGOO      RepeatMasker    similarity      11698   11733    5.9    +       .       Target "Motif:(AGTCT)n" 1 36
Chr0_RaGOO      RepeatMasker    similarity      13317   13472   23.1    -       .       Target "Motif:rnd-5_family-288" 1
```



```
GGAAAGGAATACCTTGCAAAAAAAAGTAAAGGAAGAACTTGAAAAGCTTG
AAAATGAGAAGAAAGAGGAAGCTGAAAAGGAGGGAAGGAAGAAAATTGCT
GCAGAAGAAATTTCTAAGTTTCAAGAGCATGTGCGTAATTTCACCTCCAG
TTTTGCATGTCGTTGTTTTGAATTTGATATTTCCTACTACAGTTATAATC
AGTTAAACAGGTTTTAAAGTCCAAGTACTATTCATCTAAATCCATTATTG
ATTCATTACAAATAACAGCATCTATTGCAGATAAAGAGTATGAAACTTGA
TTTTAGACAAAGATGCTCCTTTTCTTCAATAGATATTGAATTCTACAAGA
TTACTCAAGATCATATTTCTTTATCCATATGGAAGTGCTTTTATTTCTTC
AGTAAATAATTTTTTTGAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNAAAAACCCTCTGGGGACACTCTTGGAATAAT
CCTGGTGCTGCAGAAAACCAAGGGATAATTCTTGGCAGGACTGCACTTAA
TGAGAGCCCTTCCAATCCTGCTGCTTTATTAAACAGAGAAATTGCTCTGG
ATTCTCCTTCCTCCAAATCCCAGCCCTCTCCATGTCCATGGTGCAGCTCC
TTCCTTCCCCCAGCGCCCAGGCTGCCCCATAAATCACACTTGTACATCAA
ATAGGATTAATTACAGAGCCAGAAATATTCCTGCCCACATTCCCACCCCT
GCACAGCCCACAAAGCCCTGTGGCAATTCAAATATGTGGATTATGGGTTG
CTTCTCCAGGGTCACTGGGAAGCTAAGGACAGATGCCTGGAGTCCTCTGG
CTGTCAGGAATGCTGCTGCTCTTTTAGACATTAGTCCCAAATAAATAAAT
AATCTAAAGCCAAGCTGGTTTTTACTCTATTTTGAAACCTCACATTGATG
```

Masked genome

Genome Annotation



RNA/Transcript Evidence (the options are called EST for historic reasons)

```
est= #set of ESTs or assembled mRNA-seq in fasta format
altest= #EST/cDNA sequence file in fasta format from an alternate organism
est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file
altest_gff= #aligned ESTs from a closly relate species in GFF3 format
```

Protein Homology Evidence

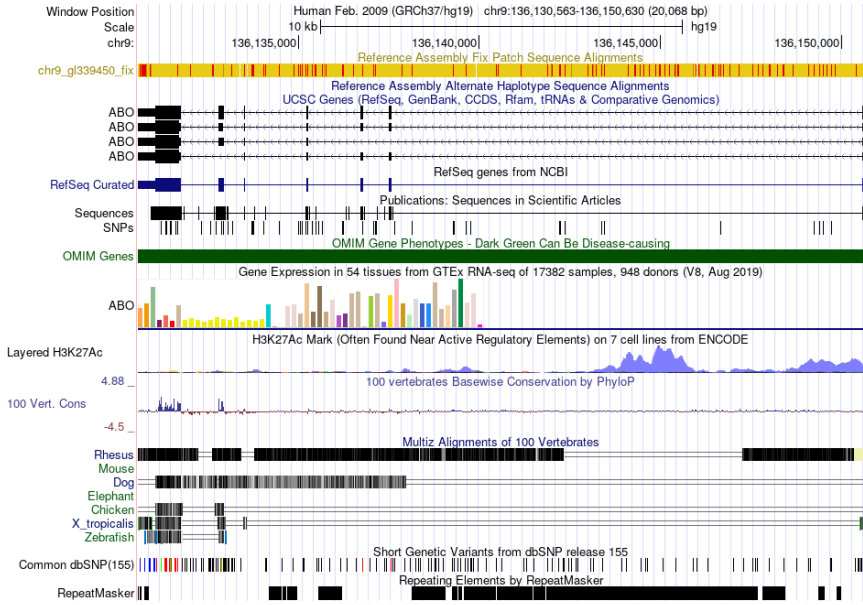
```
protein= #protein sequence file in fasta format (i.e. from mutiple oransisms)
protein_gff=  #aligned protein homology evidence from an external GFF3 file
```

Repeat Masking

```
model_org=all #select a model organism for RepBase masking in RepeatMasker
rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker
repeat_protein= #provide a fasta file of transposable element proteins for RepeatRunner
rm_gff= #pre-identified repeat elements from an external GFF3 file
prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no
softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and dust filtering)
```

Gene Prediction

```
snaphmm= #SNAP HMM file
gmhmm= #GeneMark HMM file
augustus_species= #Augustus gene prediction species model
fgenesh_par_file= #FGENESH parameter file
pred_gff= #ab-initio predictions from an external GFF3 file
model_gff= #annotated gene models from an external GFF3 file (annotation pass-through)
est2genome=1 #infer gene predictions directly from ESTs, 1 = yes, 0 = no
protein2genome=0 #infer predictions from protein homology, 1 = yes, 0 = no
trna=0 #find tRNAs with tRNAscan, 1 = yes, 0 = no
snoscan_rrna= #rRNA file to have Snoscan find snoRNAs
unmask=0 #also run ab-initio prediction programs on unmasked sequence, 1 = yes, 0 = no
```



Estimating phylogenies from genomes

Phylogenomics focuses largely on analyzing evolutionary histories to reconstruct relationships between taxa

A persistent problem in phylogenomics is the selection of appropriate genetic data or markers for a given taxonomic groups

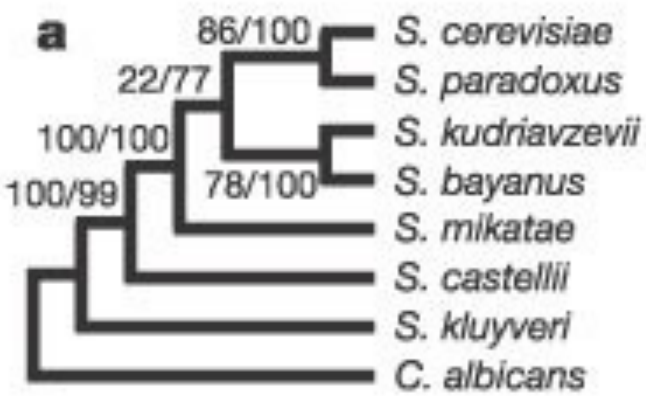
Practical limitations or considerations:

-Genome quality

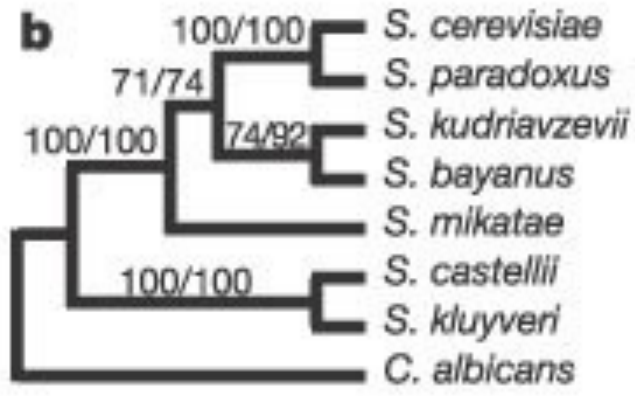
-Genome species sequenced
(Transcriptome alternative)

-Computational time

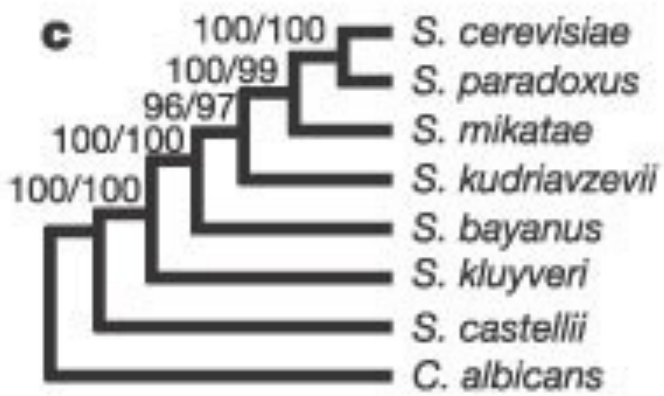
More genes with different evolutionary rates increase and fully resolve species tree with maximum support



6 genes



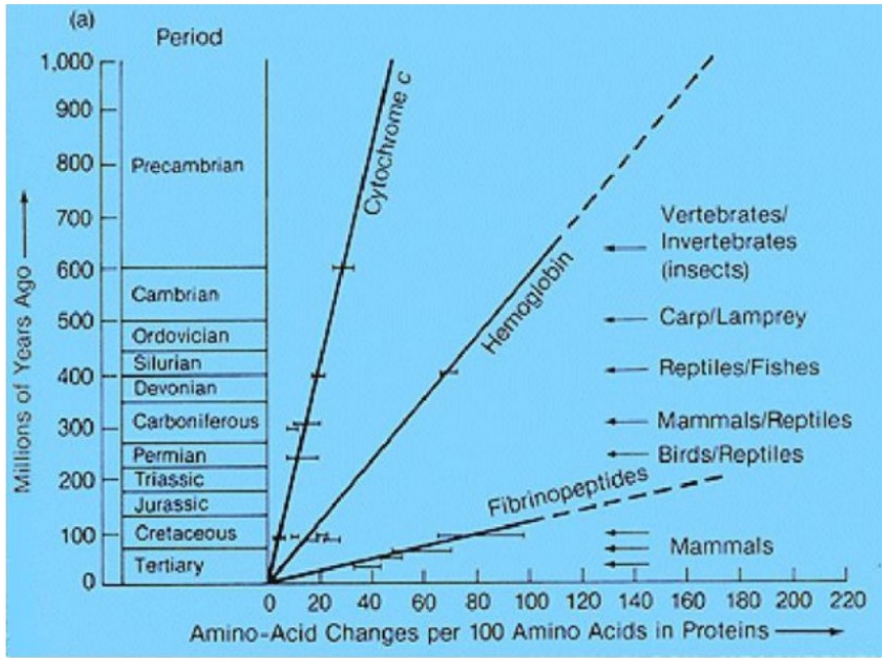
10 genes



106 genes

Gene	Description	Reference
EF-1α	Elongation factor-1α, Role in protein synthesis.	[52]
rpoA gene	Encoding the alpha subunit of RNA polymerase	[53]
atpB	Encode the beta subunit of ATP synthase	[54]
dnaA	involved in DNA synthesis initiation	[55]
ftsZ	Role in cell division	[56]
gapA	Codes for glyceraldehyde phosphate dehydrogenase	[57]
groEL	Encodes bacterial heat shock protein.	[58]
gltA	Encoding citrate synthase	[59]
ITS	Piece of non-functional RNA situated between structural	[60]
	ribosomal RNAs precursor transcript.	
lux Gene	encode proteins involved in luminescence	[61]
PEPCK	Codes for phosphoenolpyruvate carboxykinase	[62]
pyrH genes	Codes for uridine monophosphate (UMP) kinases	[63]
recA	Role in recombination	[64]
U2 snRNA	Component of the spliceosome	[65]
Wsp gene	Encodes a major cell surface coat protein	[66]
Nuclear H3	Codes for protein which is associated with DNA	[67]
trnH-psbA	Non-coding intergenic spacer region located in plastid genome	[68]
rpoB, rpoC1	Coding region located in plastid genome	[69]

Table 1: List of some other molecular markers used in phylogeny research.

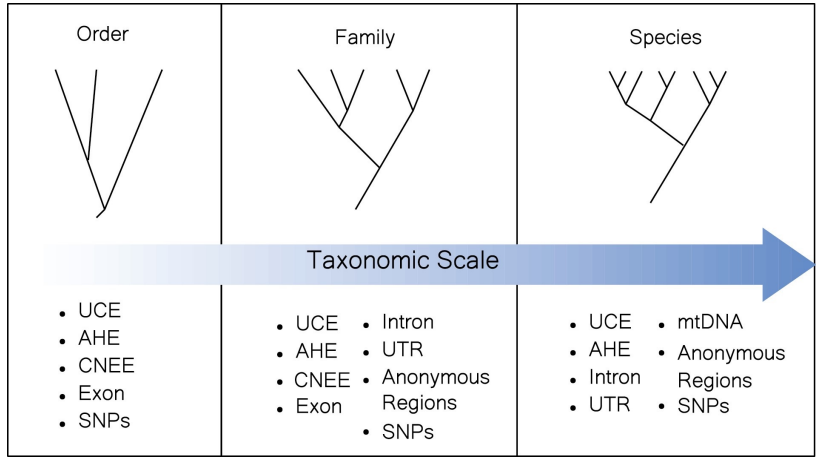


Hartl and Clark, 1997; Rokas et al., 2003

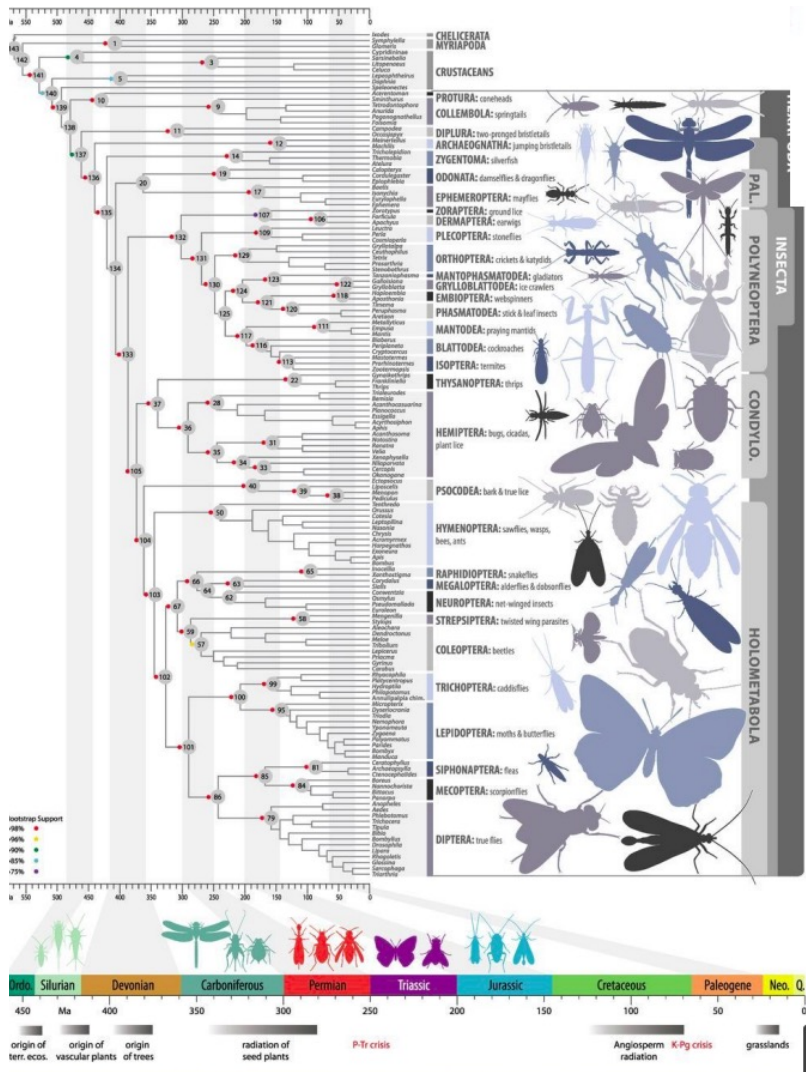
Marker types	Specimen samples needed/minimum quality needed	Evolutionary history estimate (deep-moderate-shallow)	Types of genetic and genomic method needed/reference genome requirement	Relative cost based on genomic method needed (see previous column)
UCEs = ultraconserved element flanking regions	Historic specimens, fresh tissue, blood samples/low quality	Deep to shallow time estimates	Target method, WGS data with computational target method/reference genome needed if designing probes	Low to moderate
AHE = anchor hybrid enrichment	Historic specimens, fresh tissue, blood samples/low quality	Deep to shallow time estimates	Target method, WGS data with computational target method/reference genome needed if designing probes	Low to moderate
CNEEs = conserved nonexonic elements	Historic specimens, fresh tissue, blood tissue/moderate quality	Moderate to deep time estimates	Target method or WGS data (with computational target method (preferred)/reference genome preferred	Low to moderate
Exon	Historic specimens, fresh tissue, blood sample, RNA samples/low to high quality	Mostly deep time estimates	WGS data, RNA/transcriptomic data, exonic target capture/reference genome preferred	Moderate to high
Introns	Historic specimens, fresh tissue, blood sample/moderate quality	Deep to moderate time estimates	WGS data/transcriptomic data, intronic target capture/reference genome preferred	Low to moderate
UTRs = untranslated regions	Fresh tissue, blood sample, RNA samples/high quality	Deep or shallow time estimates	WGS, RNA/transcriptomic data/reference genome preferred	Moderate to high
Mitochondrial DNA	Historic specimens, fresh tissue, blood samples/low quality	Shallow time estimates	mtDNA primers, WGS/no reference genome required	Low
Anonymous loci/regions (RADseq)	Historic specimens (not including RADseq), blood sample, tissue sample/moderate quality	Moderate to shallow time estimates; ultimately depends on the primers used	PCR primers, WGS/no reference genome required for PCR primers	Low

Estimating phylogenies from genomes: A beginners review of commonly used genomic data in vertebrate phylogenomics

Javan K Carter ✉, Rebecca T Kimball, Erik R Funk, Nolan C Kane, Drew R Schield, Garth M Spellman, Rebecca J Safran



Carter et al., 2023

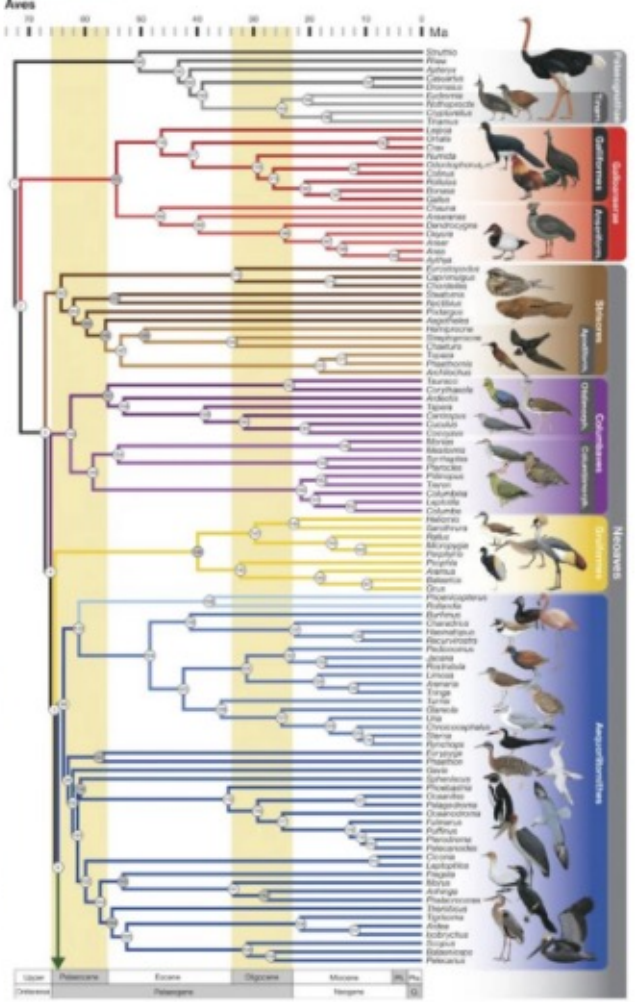


REPORT

Phylogenomics resolves the timing and pattern of insect evolution

Bernhard Misof[1,*,†], Shanlin Liu[2,3,*], Karen Meusemann[1,4,*], Ralph S. Peters[5,*], Alexander Donath[1,*], Christoph Mayer[1,*], Pau…

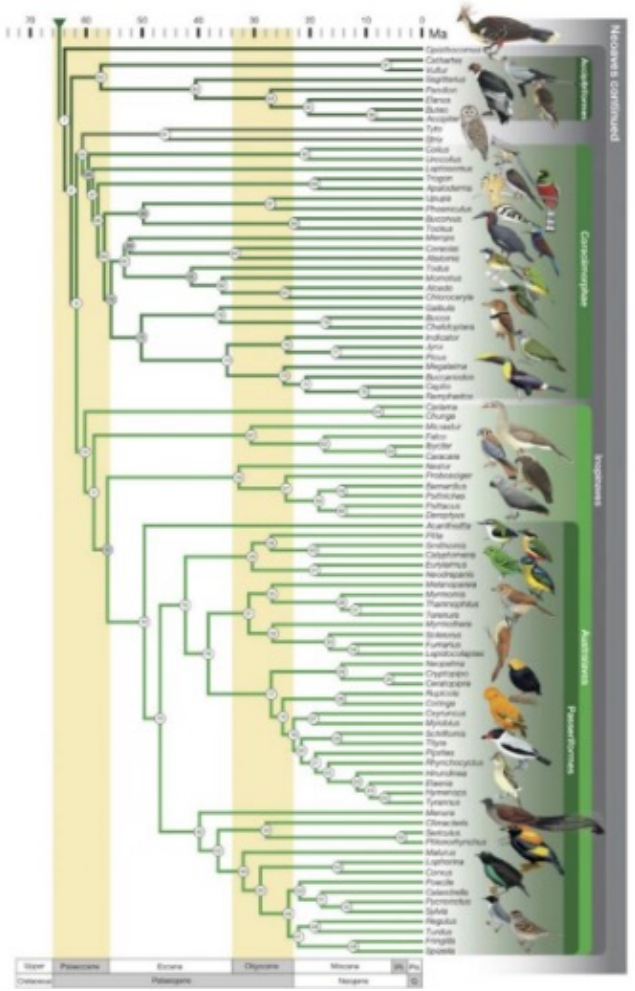
1,478 genes across species

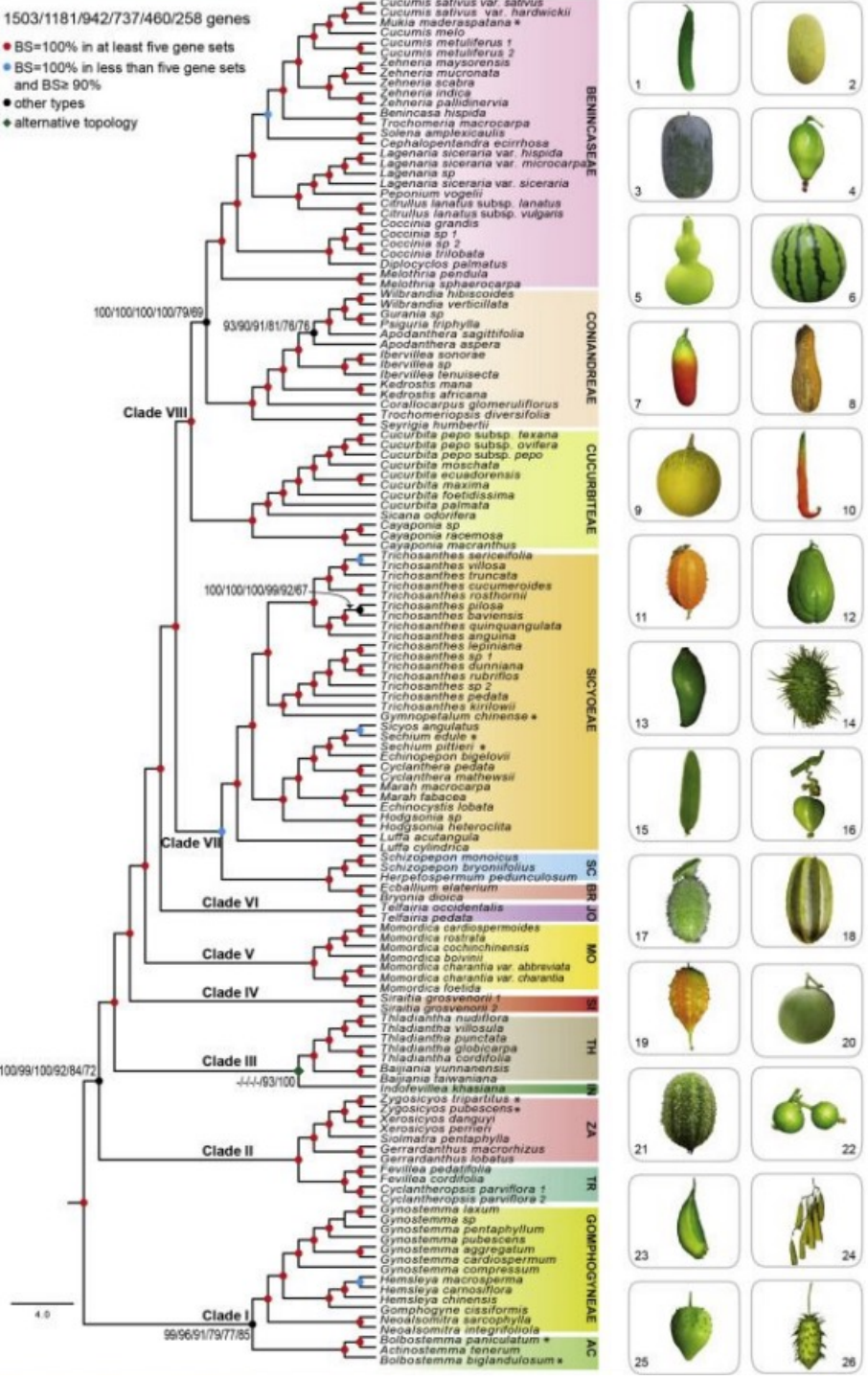


259 nuclear loci (1,523 bp)

A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing

Richard O. Prum, Jacob S. Berv, Alex Dornburg, Daniel J. Field, Jeffrey P. Townsend, Emily Moriarty Lemmon & Alan R. Lemmon





Molecular Plant
Volume 13, Issue 8, 3 August 2020, Pages 1117-1133

Research Article

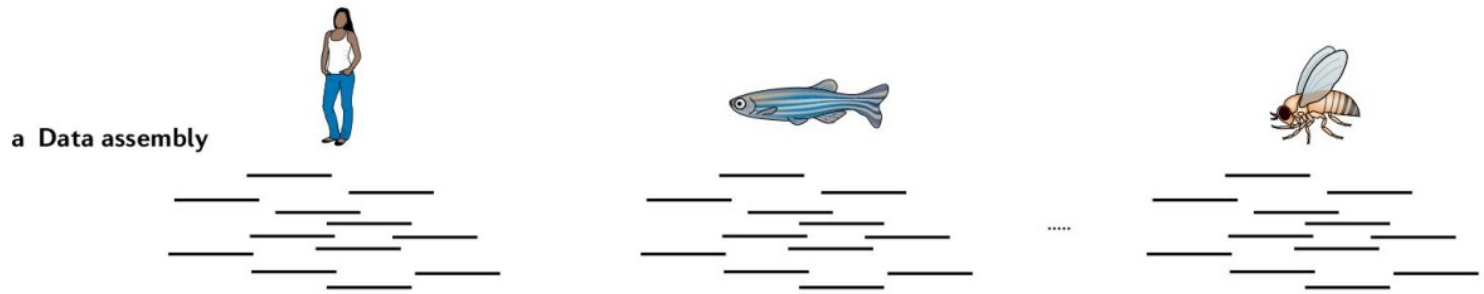
Phylotranscriptomics in Cucurbitaceae Reveal Multiple Whole-Genome Duplications and Key Morphological and Molecular Innovations

Jing Guo [1], Weibin Xu [2], Yi Hu [3], Jie Huang [1], Yiyong Zhao [1], Lin Zhang [1], Chien-Hsun Huang [1] ✉, Hong Ma [3] ✉

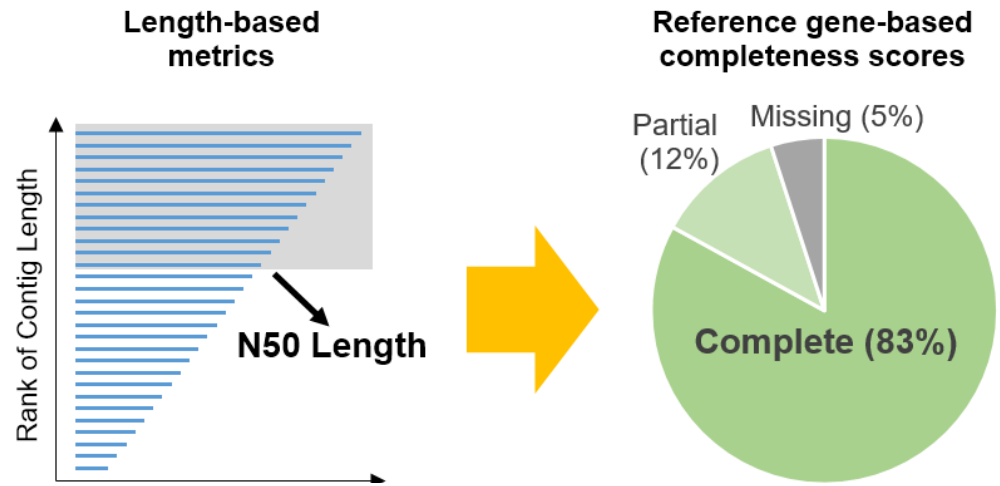
Transcriptomic data for phylogenetic reconstruction using only coding genes as an alternative to explore complex genomes

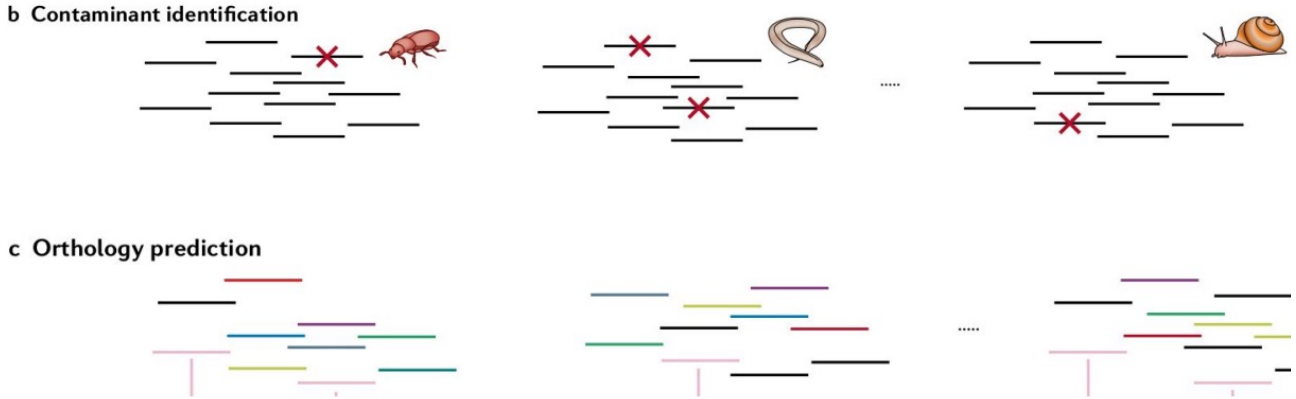
Phylogenetic tree building in the genomic age

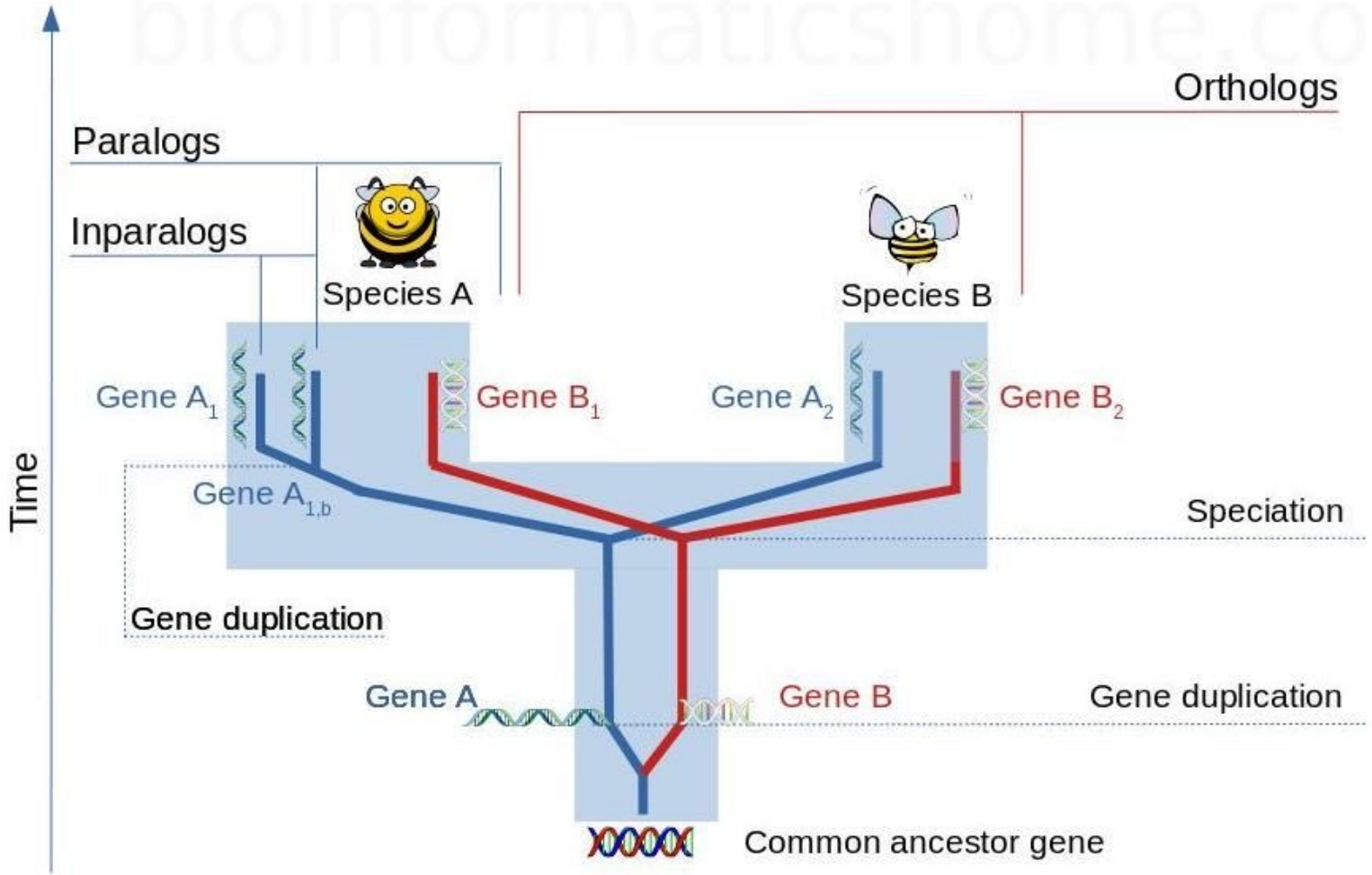
Paschalia Kapli, Ziheng Yang and Maximilian J. Telford ✉



Parent Directory			-
Annotation_comparison/	2022-11-30	20:12	-
Evidence_alignments/	2022-11-30	20:12	-
GCF_000325575.1_ASM32557v1_assembly_structure/	2023-04-04	07:23	-
Gnomon_models/	2022-11-30	20:12	-
GCF_000325575.1_ASM32557v1_assembly_report.txt	2020-01-06	20:08	5.5M
GCF_000325575.1_ASM32557v1_assembly_stats.txt	2023-04-04	07:22	7.1K
GCF_000325575.1_ASM32557v1_cds_from_genomic.fna.gz	2020-01-06	20:08	15M
GCF_000325575.1_ASM32557v1_feature_count.txt.gz	2020-01-06	20:08	545
GCF_000325575.1_ASM32557v1_feature_table.txt.gz	2020-01-06	20:08	2.9M
GCF_000325575.1_ASM32557v1_genomic.fna.gz	2020-01-06	20:08	598M
GCF_000325575.1_ASM32557v1_genomic.gbff.gz	2022-11-30	20:12	809M
GCF_000325575.1_ASM32557v1_genomic.gff.gz	2022-11-30	20:12	15M
GCF_000325575.1_ASM32557v1_genomic.gtf.gz	2022-11-30	20:12	13M
GCF_000325575.1_ASM32557v1_genomic_gaps.txt.gz	2020-01-06	20:09	1.0M
GCF_000325575.1_ASM32557v1_protein.faa.gz	2022-11-30	20:12	9.9M
GCF_000325575.1_ASM32557v1_protein.gpff.gz	2022-11-30	20:12	22M
GCF_000325575.1_ASM32557v1_pseudo_without_product.fna.gz	2020-01-06	20:09	4.5M
GCF_000325575.1_ASM32557v1_rm.out.gz	2020-01-06	20:09	95M
GCF_000325575.1_ASM32557v1_rm.run	2020-01-06	20:09	889
GCF_000325575.1_ASM32557v1_rna.fna.gz	2020-01-06	20:09	30M
GCF_000325575.1_ASM32557v1_rna.gbff.gz	2022-11-30	20:12	75M
GCF_000325575.1_ASM32557v1_rna_from_genomic.fna.gz	2020-01-06	20:09	27M
GCF_000325575.1_ASM32557v1_translated_cds.faa.gz	2020-01-06	20:09	10M
GCF_000325575.1_knownrefseq_alns.bam	2020-01-06	20:09	411K
GCF_000325575.1_knownrefseq_alns.bam.bai	2020-01-06	20:09	609K
GCF_000325575.1_modelrefseq_alns.bam	2020-01-06	20:09	29M
GCF_000325575.1_modelrefseq_alns.bam.bai	2020-01-06	20:09	1.8M
Pteropus_alecto_AR102_annotation_report.xml	2020-01-06	20:08	55K
README.txt	2020-01-27	16:55	11K





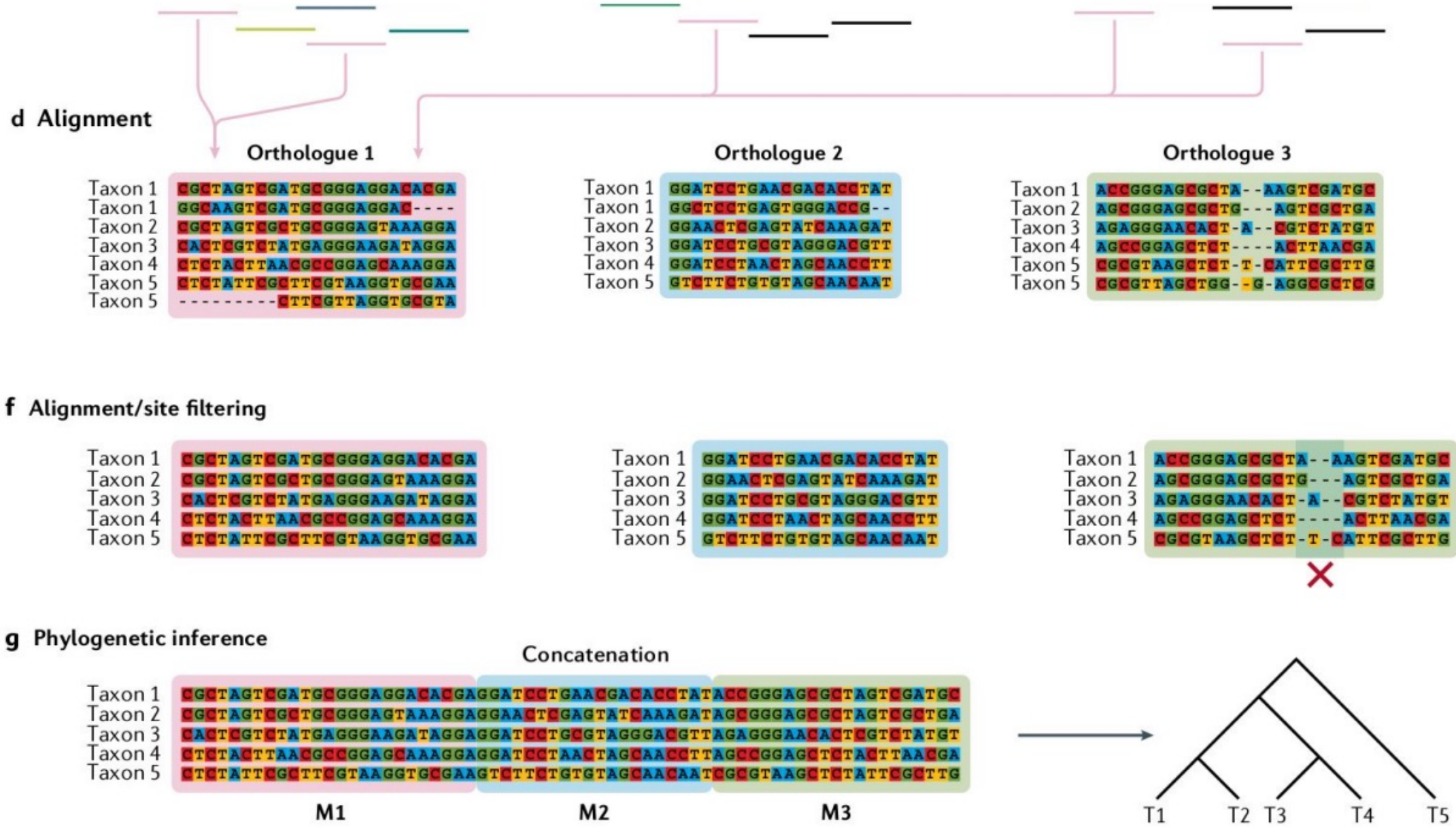


Orthologous genes are those genes that have diverged after a speciation event and are residing in different species

Single Copy Genes

Proportion	Small (~1.5%)
Rate of mutation	Low
Occurrence	Once in genome
Function	Makes protein
Identification	Similar b/w individuals Not used for profiling
Length	Long unique sequence

Kapli et al., 2020; Figure: bioinformaticshome.com

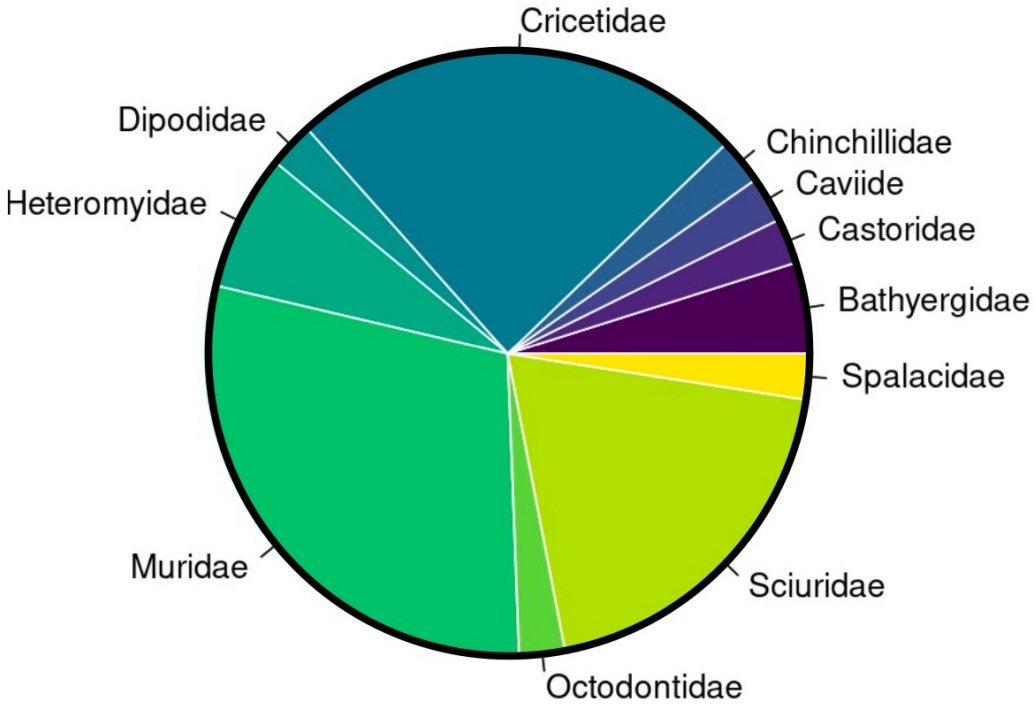

d Alignment
Orthologue 1
Taxon 1 CGCTAGTCGATGCGGGAGGACACGA
Taxon 1 GGCAAGTCGATGCGGGAGGAC----
Taxon 2 CGCTAGTCGCTGCGGGAGTAAAGGA
Taxon 3 CACTCGTCTATGAGGGAAGATAGGA
Taxon 4 CTCTACTTAACGCCGGAGCAAAGGA
Taxon 5 CTCTATTCGCTTCGTAAGGTGCGAA
Taxon 5 ---------CTTCGTTAGGTGCGTA
Orthologue 2
Taxon 1 GGATCCTGAACGACACCTAT
Taxon 1 GGCTCCTGAGTGGGACCG--
Taxon 2 GGAACTCGAGTATCAAAGAT
Taxon 3 GGATCCTGCGTAGGGACGTT
Taxon 4 GGATCCTAACTAGCAACCTT
Taxon 5 GTCTTCTGTGTAGCAACAAT
Orthologue 3
Taxon 1 ACCGGGAGCGCTA--AAGTCGATGC
Taxon 2 AGCGGGAGCGCTG---AGTCGCTGA
Taxon 3 AGAGGGAACACT-A--CGTCTATGT
Taxon 4 AGCCGGAGCTCT----ACTTAACGA
Taxon 5 CGCGTAAGCTCT-T-CATTCGCTTG
Taxon 5 CGCGTTAGCTGG--G-AGGCGCTCG
f Alignment/site filtering
Taxon 1 CGCTAGTCGATGCGGGAGGACACGA
Taxon 2 CGCTAGTCGCTGCGGGAGTAAAGGA
Taxon 3 CACTCGTCTATGAGGGAAGATAGGA
Taxon 4 CTCTACTTAACGCCGGAGCAAAGGA
Taxon 5 CTCTATTCGCTTCGTAAGGTGCGAA
Taxon 1 GGATCCTGAACGACACCTAT
Taxon 2 GGAACTCGAGTATCAAAGAT
Taxon 3 GGATCCTGCGTAGGGACGTT
Taxon 4 GGATCCTAACTAGCAACCTT
Taxon 5 GTCTTCTGTGTAGCAACAAT
Taxon 1 ACCGGGAGCGCTA--AAGTCGATGC
Taxon 2 AGCGGGAGCGCTG---AGTCGCTGA
Taxon 3 AGAGGGAACACT-A--CGTCTATGT
Taxon 4 AGCCGGAGCTCT----ACTTAACGA
Taxon 5 CGCGTAAGCTCT-T-CATTCGCTTG
g Phylogenetic inference
Concatenation
Taxon 1 CGCTAGTCGATGCGGGAGGACACGAGGATCCTGAACGACACCTATACCGGGAGCGCTAGTCGATGC
Taxon 2 CGCTAGTCGCTGCGGGAGTAAAGGAGGAACTCGAGTATCAAAGATAGCGGGAGCGCTAGTCGCTGA
Taxon 3 CACTCGTCTATGAGGGAAGATAGGAGGATCCTGCGTAGGGACGTTAGAGGGAACACTCGTCTATGT
Taxon 4 CTCTACTTAACGCCGGAGCAAAGGAGGATCCTAACTAGCAACCTTAGCCGGAGCTCTACTTAACGA
Taxon 5 CTCTATTCGCTTCGTAAGGTGCGAAGTCTTCTGTGTAGCAACAATCGCGTAAGCTCTATTCGCTTG
M1
M2
M3
T1
T2
T3
T4
T5

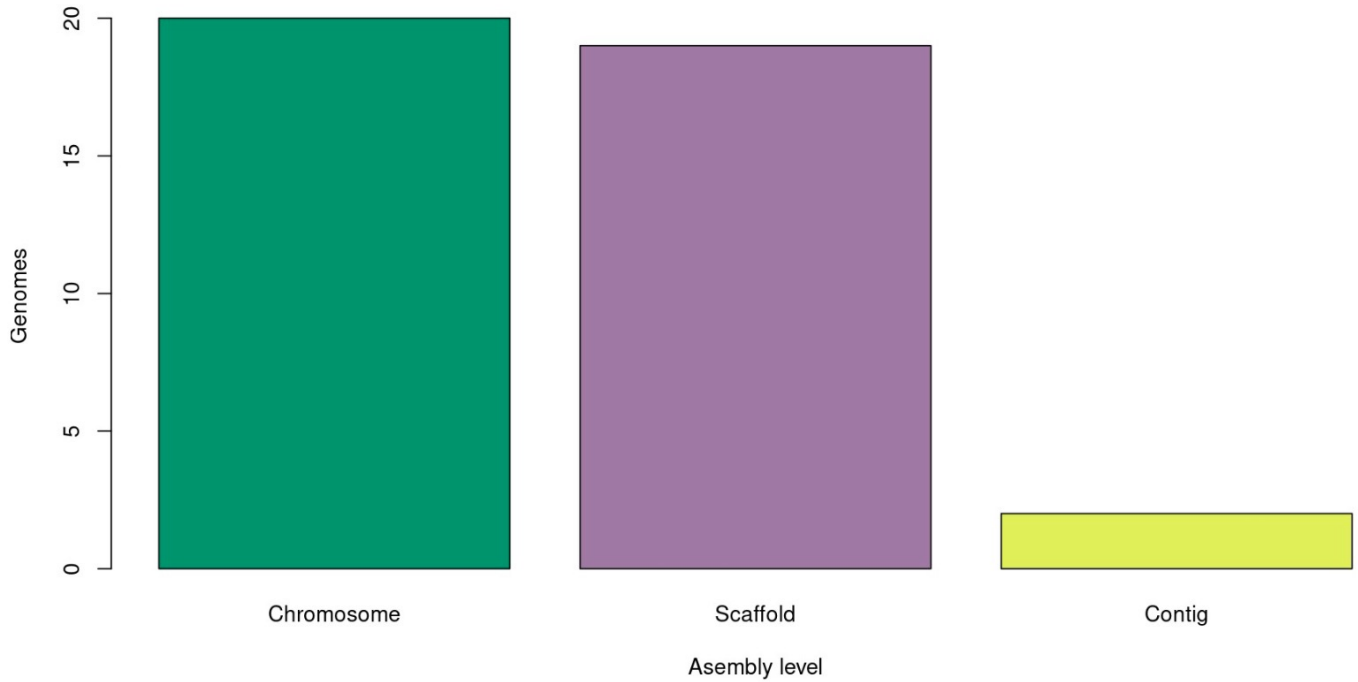
Phylogenomic exercise

First Part

Photo credits: inaturalista; mammal.org; wildlife; Wikispecies

41 rodent species
11 families





Genome size

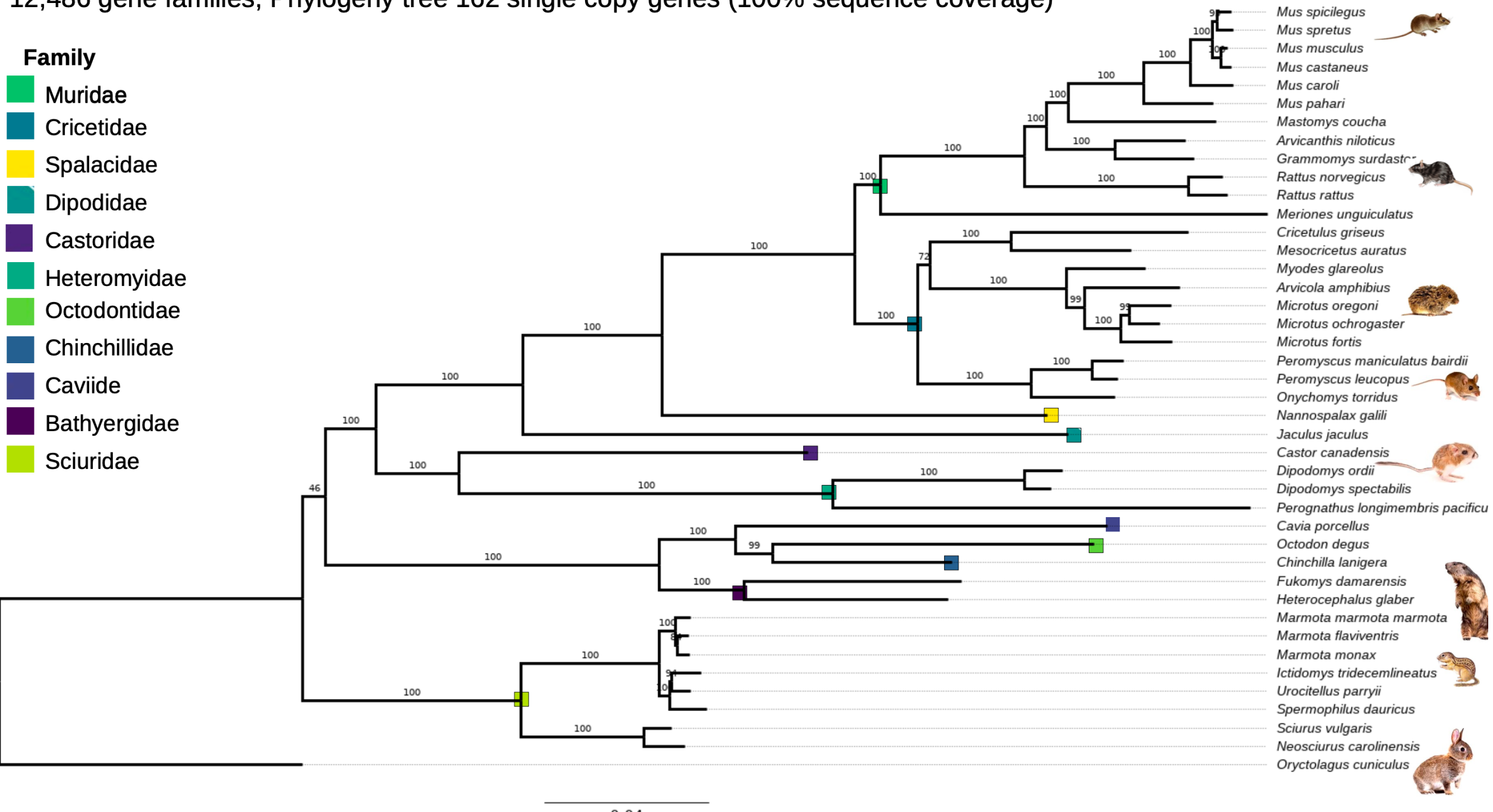
2.18 Gbp

Microtus fortis

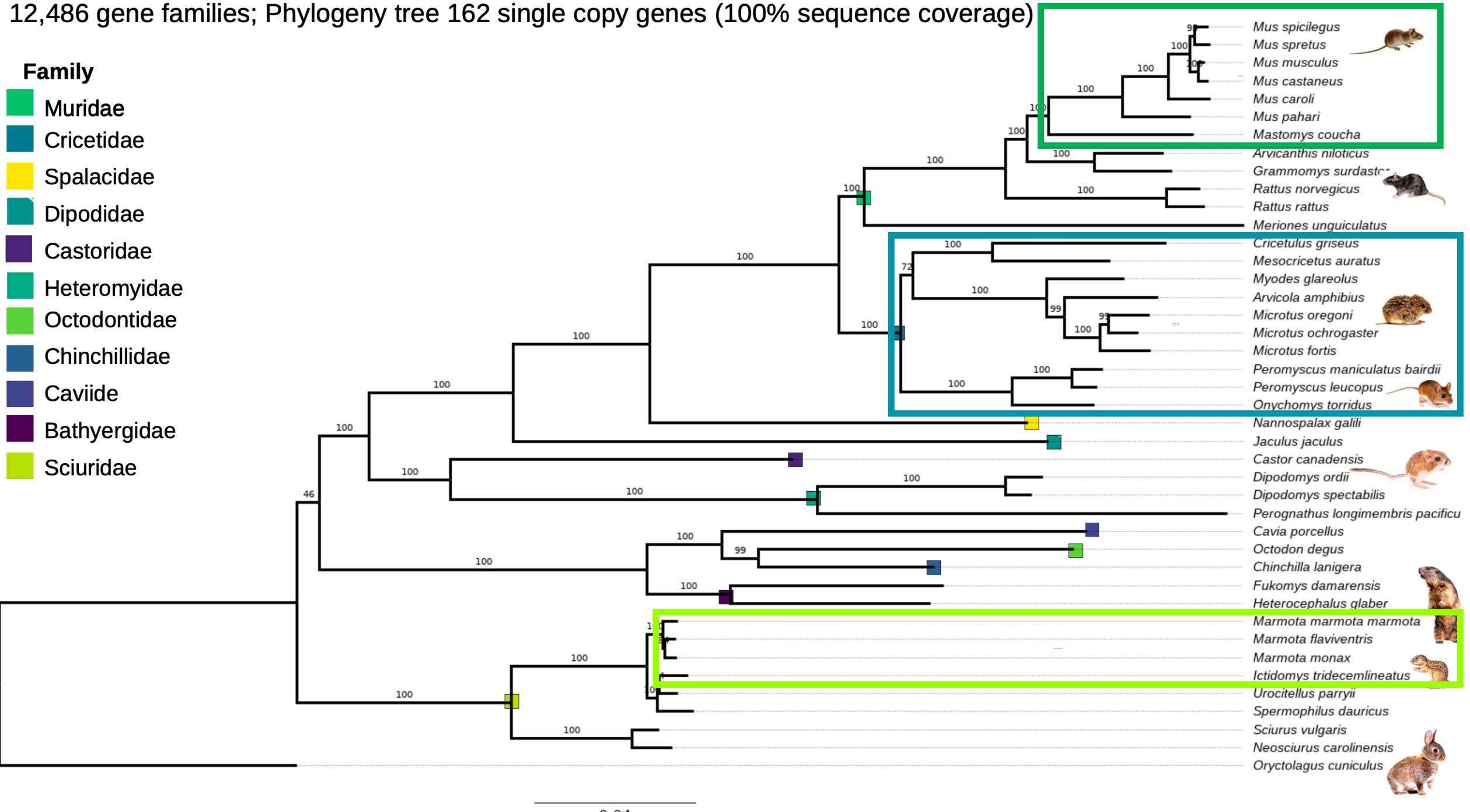
3.1 Gbp

Spermophilus dauricus

12,486 gene families; Phylogeny tree 162 single copy genes (100% sequence coverage)



12,486 gene families; Phylogeny tree 162 single copy genes (100% sequence coverage)

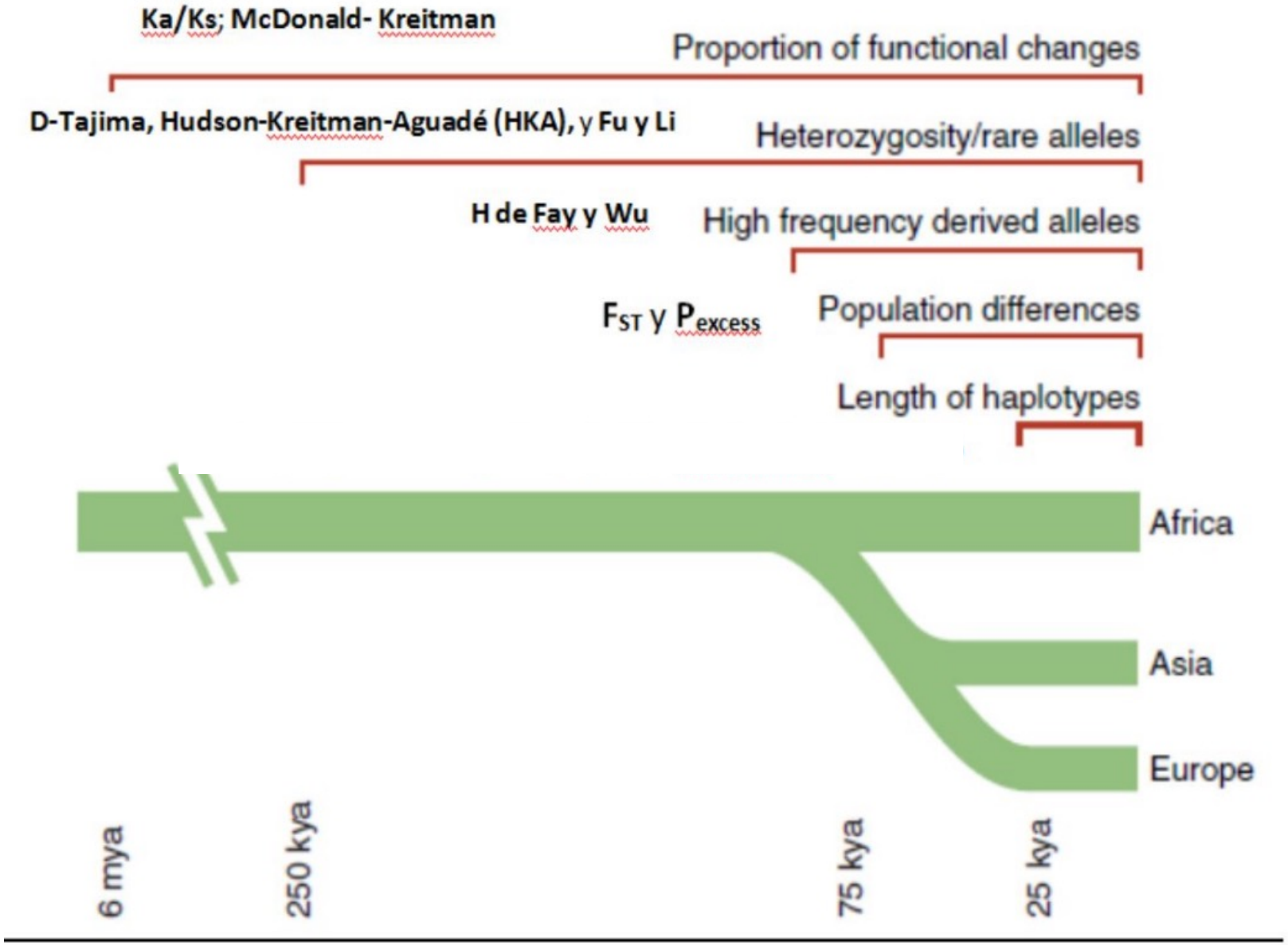


Selection test over long evolutionary time scales (*dN/dS*)

-Different hypothesis and resolution

-Depend on the questions

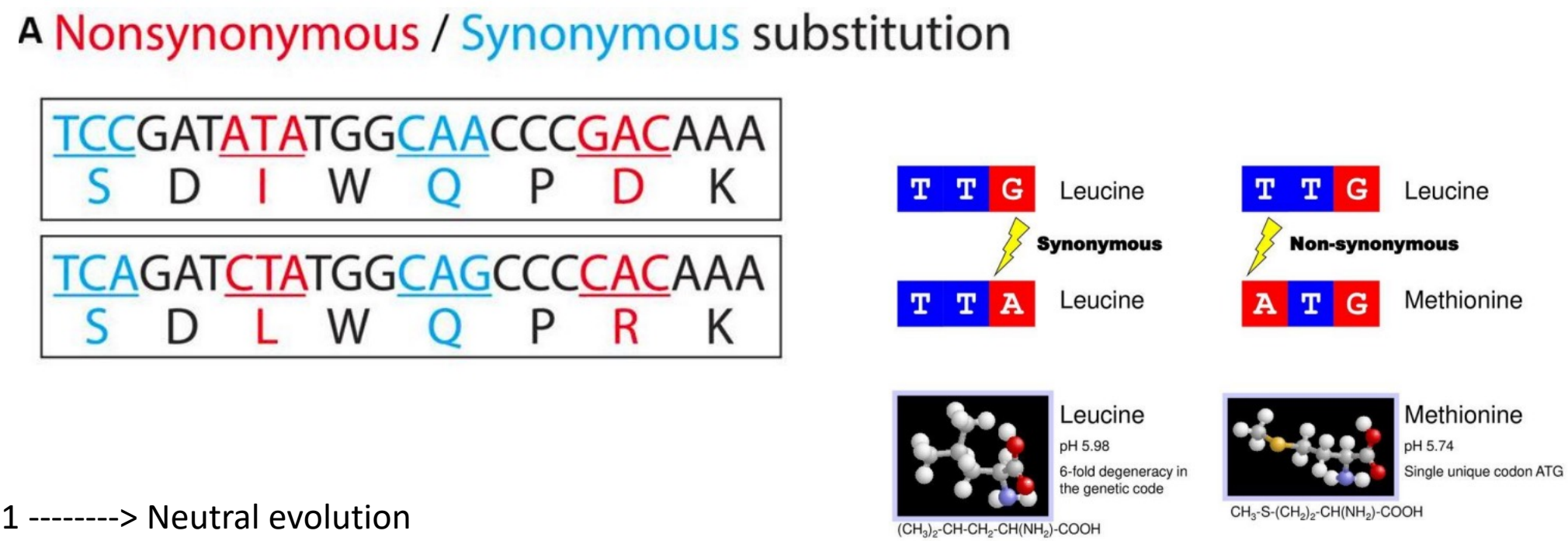
-Timing and divergence



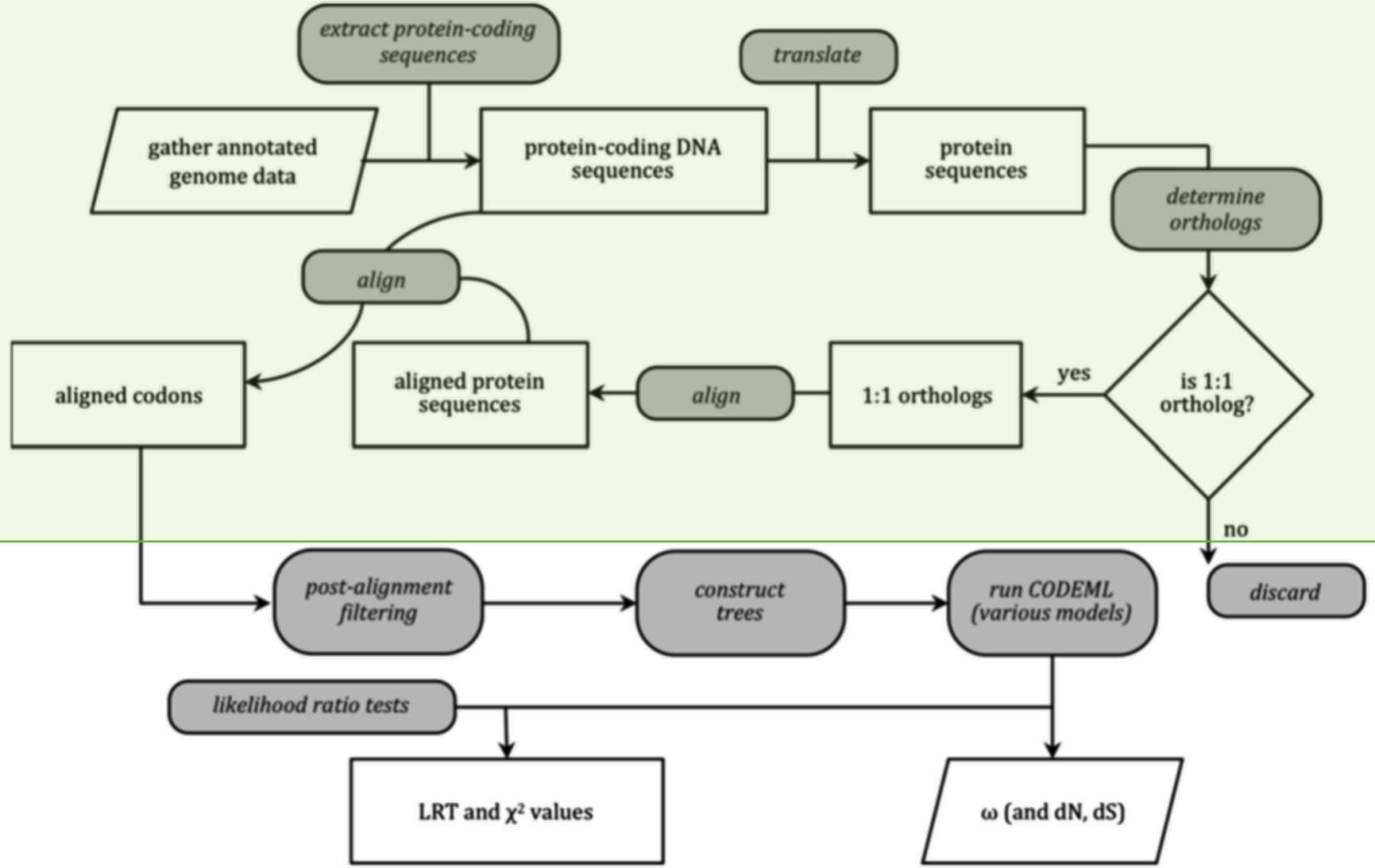
Sanders et al., 2013

The ratio of non-synonymous to synonymous substitutions (*dN*/*dS*) is a statistic measure of the strength and mode of natural selection acting on protein-coding genes in species that have diverged

Non-synonymous substitutions are nucleotide changes that alter the protein sequence and might modify protein structure
Synonymous substitutions do not change the protein sequence



dN/dS = 1 --------> Neutral evolution
dN/dS > 1 --------> Positive Selection
dN/dS < 1 --------> Purifying Selection



Phylogenomic

CODEML calculates *dN / dS* using the observed changes present in a multiple alignment of protein coding gene sequences from several species in a phylogeny

Statistical estimation of *dN / dS* ratio uses maximum likelihood, and correct for multiple changes, accounting for the different numbers of non-synonymous and synonymous sites

Yang and Nielsen, 2000; Jeffares et al., 2015

CODEML

Models		k	Hypothesis tested
Site models:	M2a vs. M1a	2[a]	Does adding a third class of sites with $\omega > 1$ (adaptive evolution) fit the data better than a model with two classes $\omega < 1$, $\omega = 1$?
	M8 vs. M7	2[a]	Does adding an extra class of sites with $\omega > 1$ (adaptive evolution) fit the data better than a model with ten classes with flexible normalized non-synonymous ratio distribution?
Branch models:	*Free-ratio* vs. *one-ratio* model	$2s-4$	For a tree of s species, is ω different among lineages?
	Two-ratio vs. *one-ratio* model	1	Are the foreground branches that you specify more likely to have different ω from background branches?
Branch-site models:	MA($\omega > 1$) vs. MA($\omega = 1$)	1[b]	Is the defined "foreground branch" more likely to contain sites with $\omega > 1$

HyPhy: hypothesis testing using phylogenies

Sergei L. Kosakovsky Pond ✉, Simon D. W. Frost, Spencer V. Muse

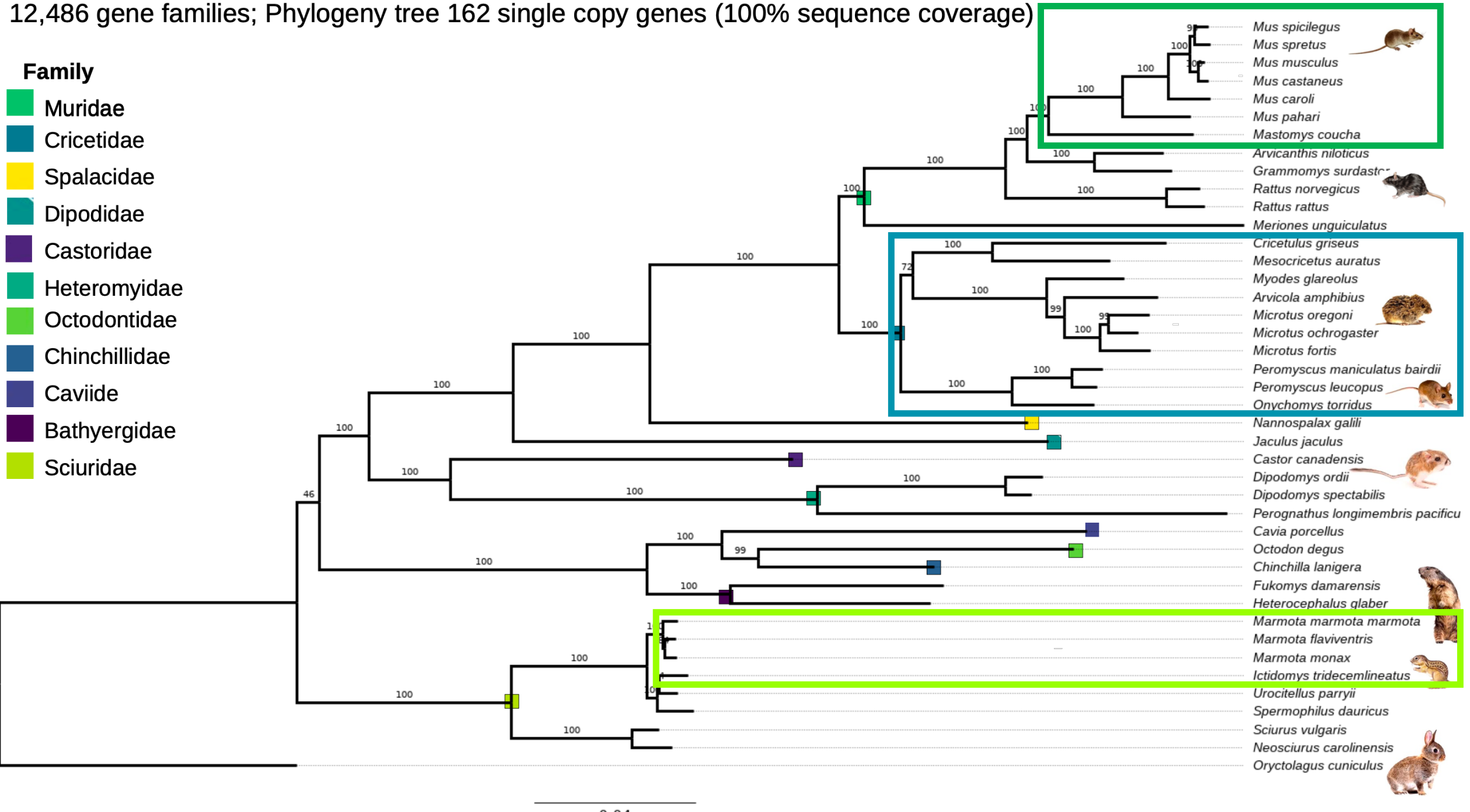
The Site-Wise Log-Likelihood Score is a Good Predictor of Genes under Positive Selection

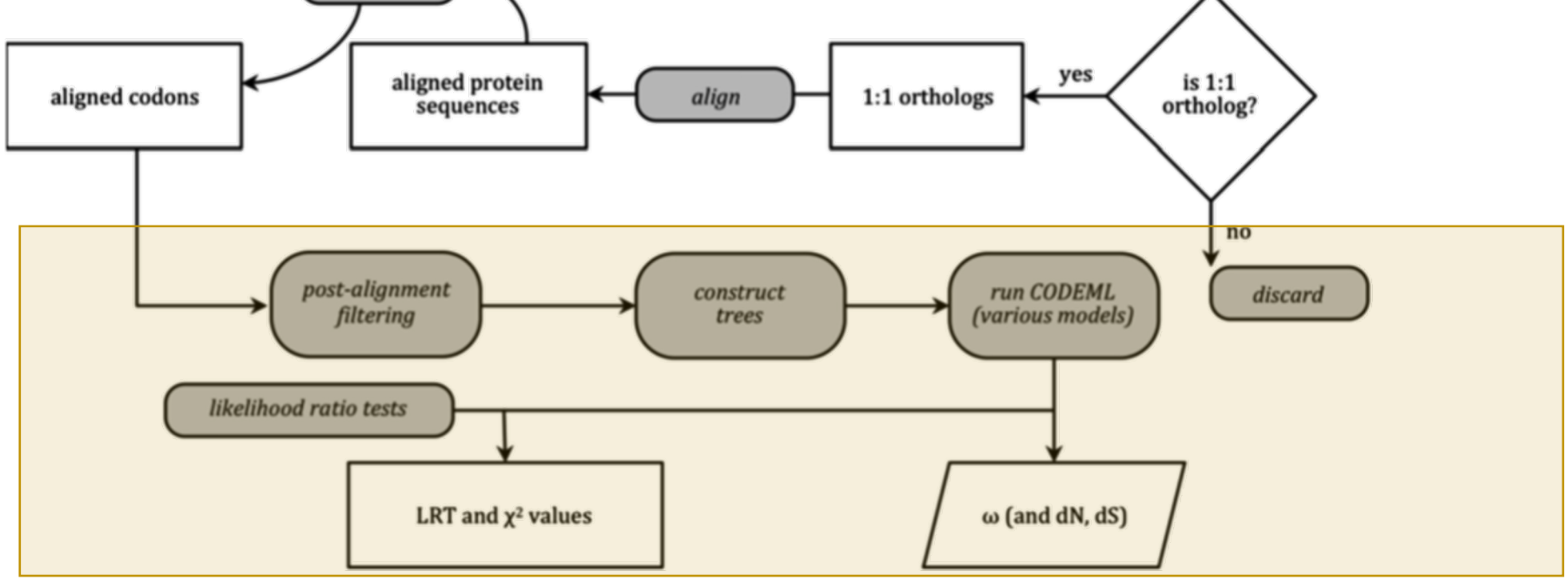
Huai-Chun Wang ✉, Edward Susko & Andrew J. Roger

dN/dS selection test exercise

Second Part

12,486 gene families; Phylogeny tree 162 single copy genes (100% sequence coverage)





Phylogenomic

***dN/dS* test with CODEML**

Annotation
and
Gene Enrichment analysis

(topGO, R library)

Figure adapted from Jeffares et al., 2015