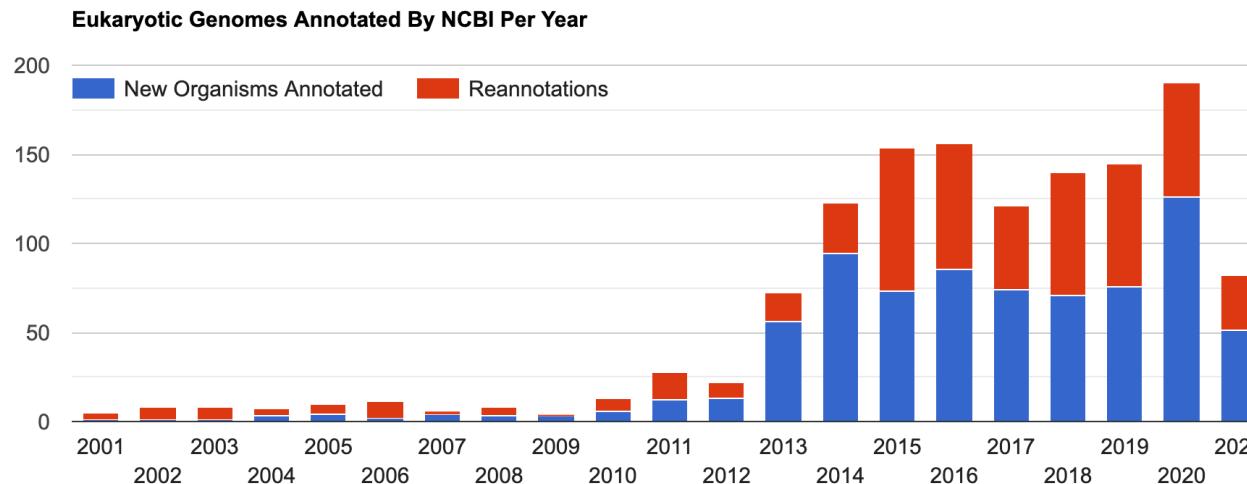
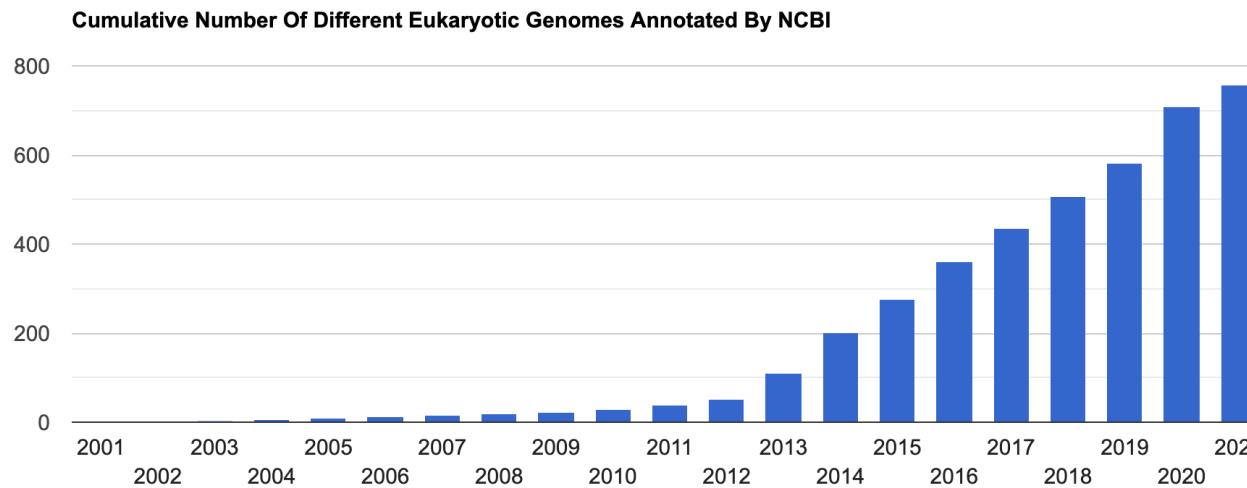


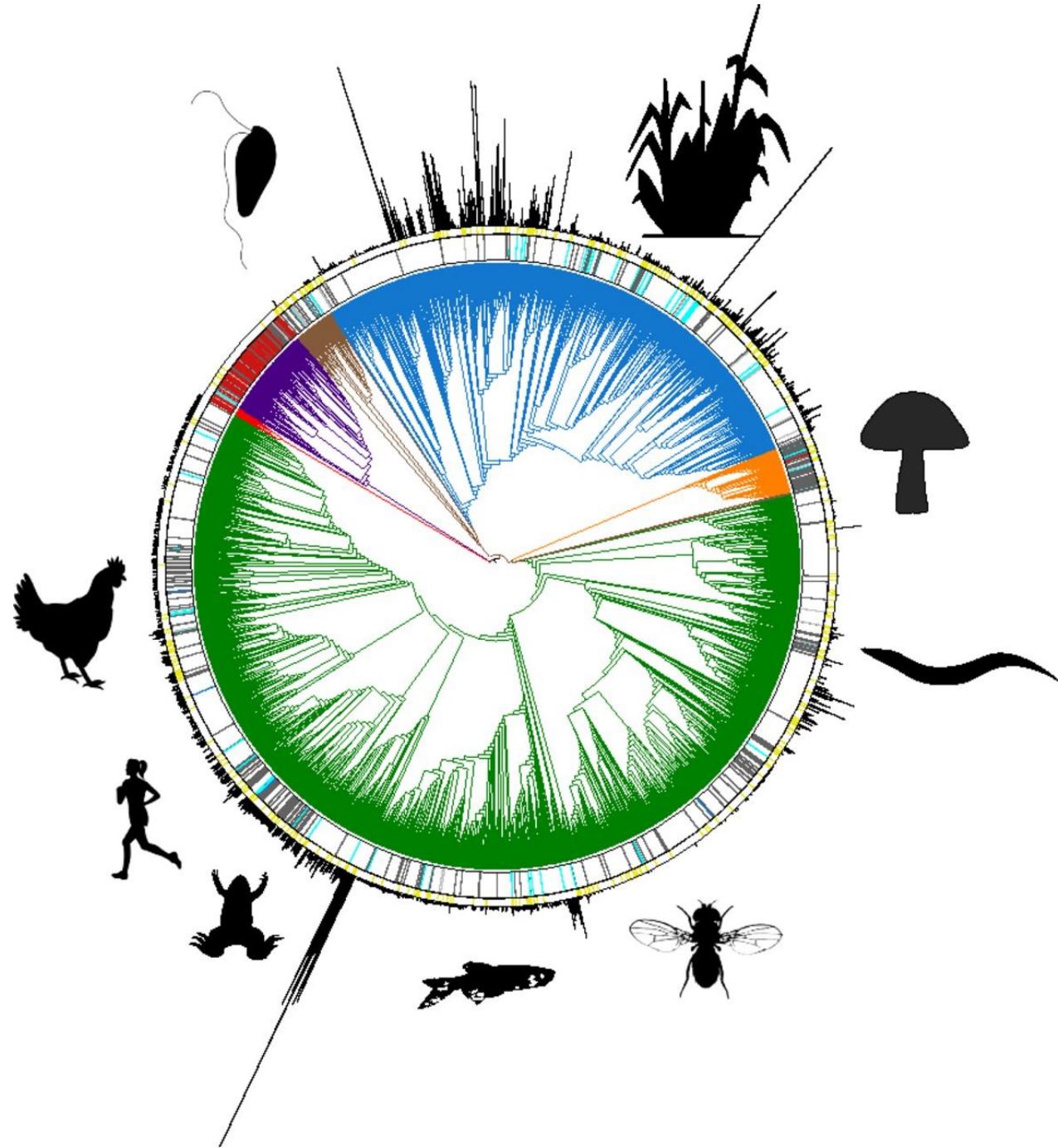
Whole Genome Alignment

Number of annotated eukaryotic genomes increasing



https://www.ncbi.nlm.nih.gov/genome/annotation_euk/

Status of sequencing of life



Lewin et al 2018

Earth BioGenome Project and others

PERSPECTIVE



Earth BioGenome Project: Sequencing life for the future of life

Harris A. Lewin^{a,b,c,d,1}, Gene E. Robinson^e, W. John Kress^f, William J. Baker^g, Jonathan Coddington^f, Keith A. Crandall^h, Richard Durbin^{i,j}, Scott V. Edwards^{k,l}, Félix Forest^g, M. Thomas P. Gilbert^{m,n}, Melissa M. Goldstein^o, Igor V. Grigoriev^{p,q}, Kevin J. Hackett^s, David Haussler^{r,t}, Erich D. Jarvis^u, Warren E. Johnson^v, Aristides Patrinos^w, Stephen Richards^x, Juan Carlos Castilla-Rubio^{y,z}, Marie-Anne van Sluys^{a,b,b}, Pamela S. Soltis^{c,c}, Xun Xu^{dd}, Huanming Yang^{e,e}, and Guojie Zhang^{dd,ff,gg}

Edited by John C. Avise, University of California, Irvine, CA, and approved March 15, 2018 (received for review January 6, 2018)

PERSPECTIVE

Article

Dense sampling of bird diversity increases power of comparative genomics

<https://doi.org/10.1038/s41586-020-2873-9>

Received: 9 August 2019

Accepted: 27 July 2020

Published online: 11 November 2020

Open access



A list of authors and affiliations appears at the end of the paper.

Whole-genome sequencing projects are increasingly populating the tree of life and characterizing biodiversity^{1–4}. Sparse taxon sampling has previously been proposed to confound phylogenetic inference⁵, and captures only a fraction of the genomic diversity. Here we report a substantial step towards the dense representation of avian phylogenetic and molecular diversity, by analysing 363 genomes from 92.4% of bird families—including 267 newly sequenced genomes produced for phase II of the Bird

<https://doi.org/10.1038/s41586-020-2876-6> Zoonomia Consortium*

Received: 17 April 2019

Accepted: 27 July 2020

Published online: 11 November 2020

Open access

The Zoonomia Project is investigating the genomics of shared and specialized traits in eutherian mammals. Here we provide genome assemblies for 131 species, of which all but 9 are previously uncharacterized, and describe a whole-genome alignment of 240 species of considerable phylogenetic diversity, comprising

Article

Towards complete and error-free genome assemblies of all vertebrate species

<https://doi.org/10.1038/s41586-021-03451-0>

Received: 22 May 2020

Accepted: 12 March 2021

Published online: 28 April 2021

Open access



A list of authors and their affiliations appears at the end of the paper.

High-quality and complete reference genome assemblies are fundamental for the application of genomics to biology, disease, and biodiversity conservation. However, such assemblies are available for only a few non-microbial species^{1–4}. To address this issue, the international Genome 10K (G10K) consortium^{5,6} has worked over a five-year period to evaluate and develop cost-effective methods for assembling highly accurate and nearly complete reference genomes. Here we present lessons learned from

Progressive Cactus

- We need:
 - Softmasked genome assemblies
 - Guide tree

Article

Progressive Cactus is a multiple-genome aligner for the thousand-genome era

<https://doi.org/10.1038/s41586-020-2871-y>

Received: 24 August 2019

Accepted: 27 July 2020

Published online: 11 November 2020

Open access

 Check for updates

Joel Armstrong¹, Glenn Hickey¹, Mark Diekhans¹, Ian T. Fiddes¹, Adam M. Novak¹, Alden Deran¹, Qi Fang^{2,3}, Duo Xie^{2,4}, Shaohong Feng^{2,5}, Josefin Stiller³, Diane Genereux⁶, Jeremy Johnson⁶, Voichita Dana Marinescu⁷, Jessica Alföldi⁶, Robert S. Harris⁸, Kerstin Lindblad-Toh^{6,7}, David Haussler^{1,9}, Elinor Karlsson^{6,10,11}, Erich D. Jarvis^{9,12}, Guojie Zhang^{3,5,13,14} & Benedict Paten¹

New genome assemblies have been arriving at a rapidly increasing pace, thanks to decreases in sequencing costs and improvements in third-generation sequencing

Soft masking

- Masking can be hard or soft:
 - Hard: replace repeats with Ns
 - Soft: lower case instead of capital
 - With hardmasking we lose information
 - Want softmasking of repeats
 - Alignments will not start inside masked regions, avoids trying to align one repeat copy to all others
- Original: AGGTCACTACTACTGACTTAC
Softmask: AGGT_CactactactactGACTTAC
Hardmask: AGGT_{NNNNNNNNNNNNN}GACTTAC

How to mask

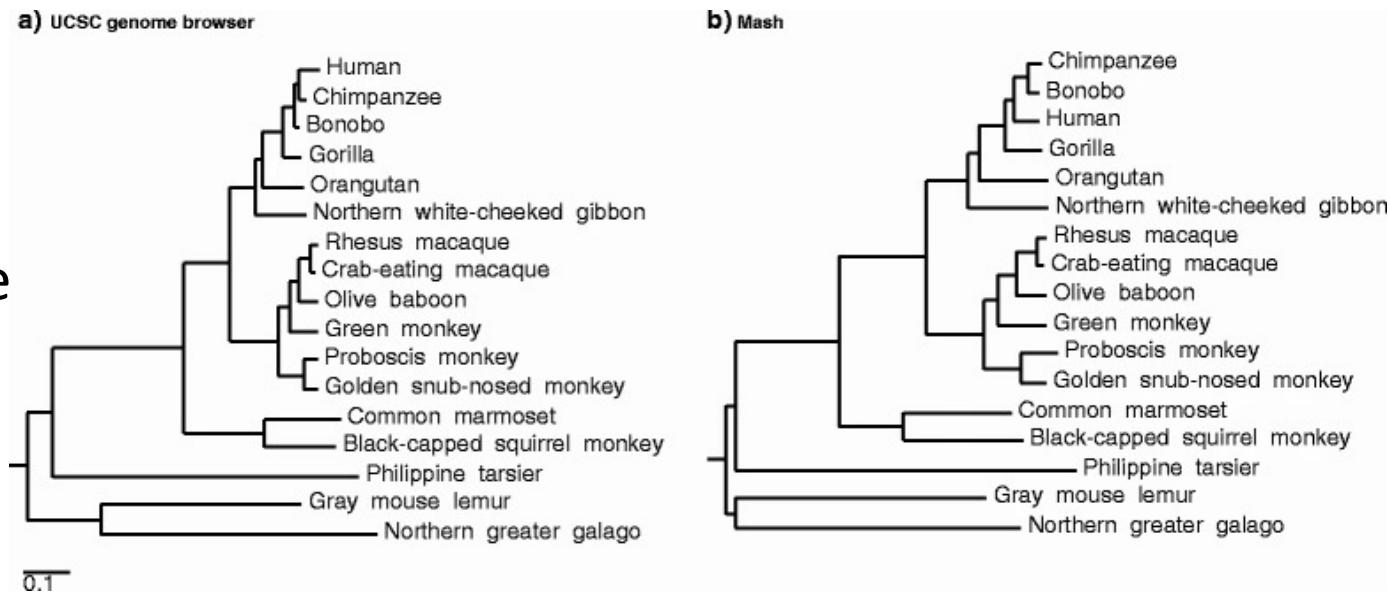
- RepeatModeler/RepeatMasker:
 - Use RepeatModeler to generate a *de novo* repeat library for your species
 - Use RepeatMasker to mask repeats based on repeat library or existing database
 - Pros: Characterizes/annotates most/many repeats (STRs, TEs, etc)
 - Cons: Take quite some time (a day or so for a medium sized genome (1 Gbp))
- Red:
 - Use Red to characterize and mask all repetitive sequence in a genome
 - Pros: Quick, and only characterises actual repetitive sequence (ignores TEs only occurring once or a few times)
 - Cons: Doesn't annotate the repeats (you don't know what it is)

Questions?



Obtaining a guide tree

- Often an complicated process
 - Find paralogs in some way
 - Align them to each other
 - Run some tree inference tool on the alignment
- Mash: fast genome distance estimation
 - Reduces sequences to sketches that are a thousandth as large
 - Can be used to quickly find distances between genomes



Ondov et al 2016

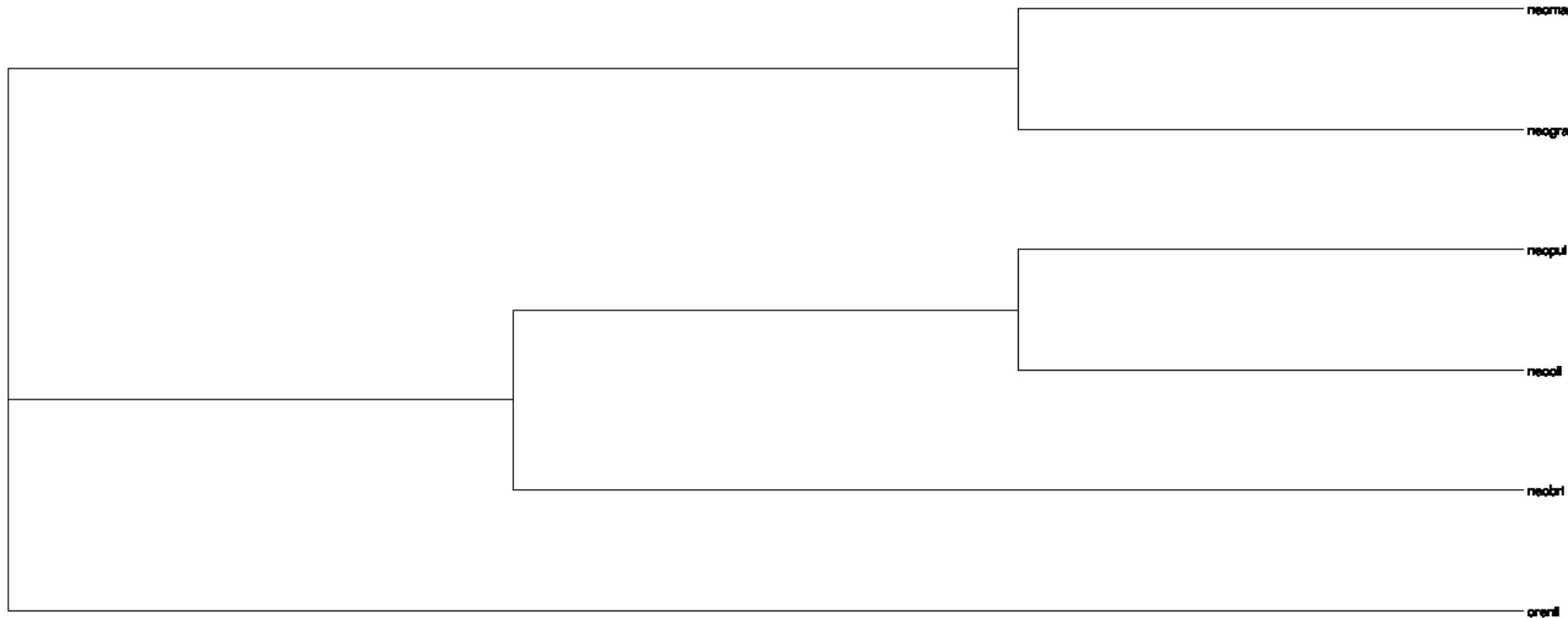
Distance matrix

- Mash triangle gives a distance matrix
- RapidNJ can generate a neighbour-joining tree based on a distance matrix

6					
neobri					
neogra	0.00981378				
neomar	0.00985499	0.00888862			
neooli	0.00904645	0.00948708	0.0090861		
neopul	0.00924552	0.00924552	0.00932572	0.00659353	
orenil	0.0519203	0.0509537	0.0515304	0.0509537	0.0511448

The guide tree

```
((neomar:0.0045736,neogra:0.004315):0.00021445,  
((neopul:0.0033453,neooli:0.0032482):0.00094829,  
neobri:0.0049009):0.00032849,orenil:0.046583);
```

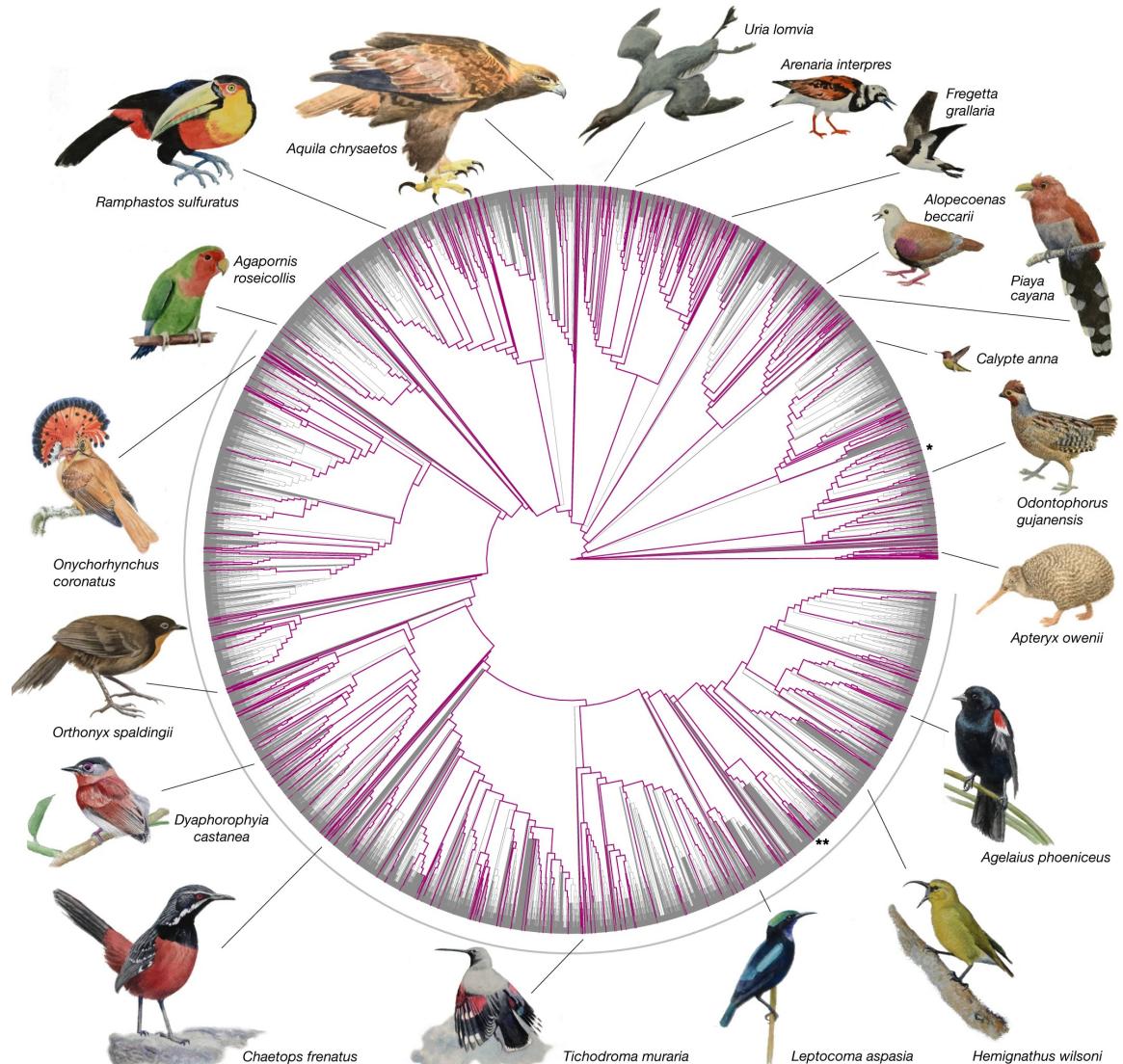


Questions?



Bird 10k project

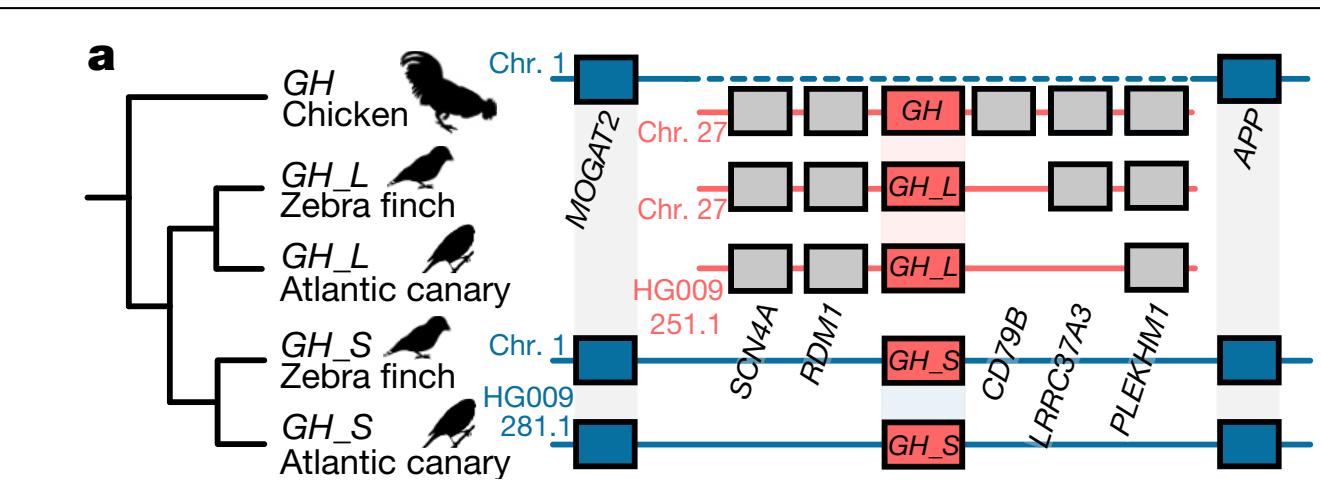
- 363 species (of 10,135 total), covering 92.4 % of families
- 267 new species
- Purple branches are with genomes



Feng et al 2020

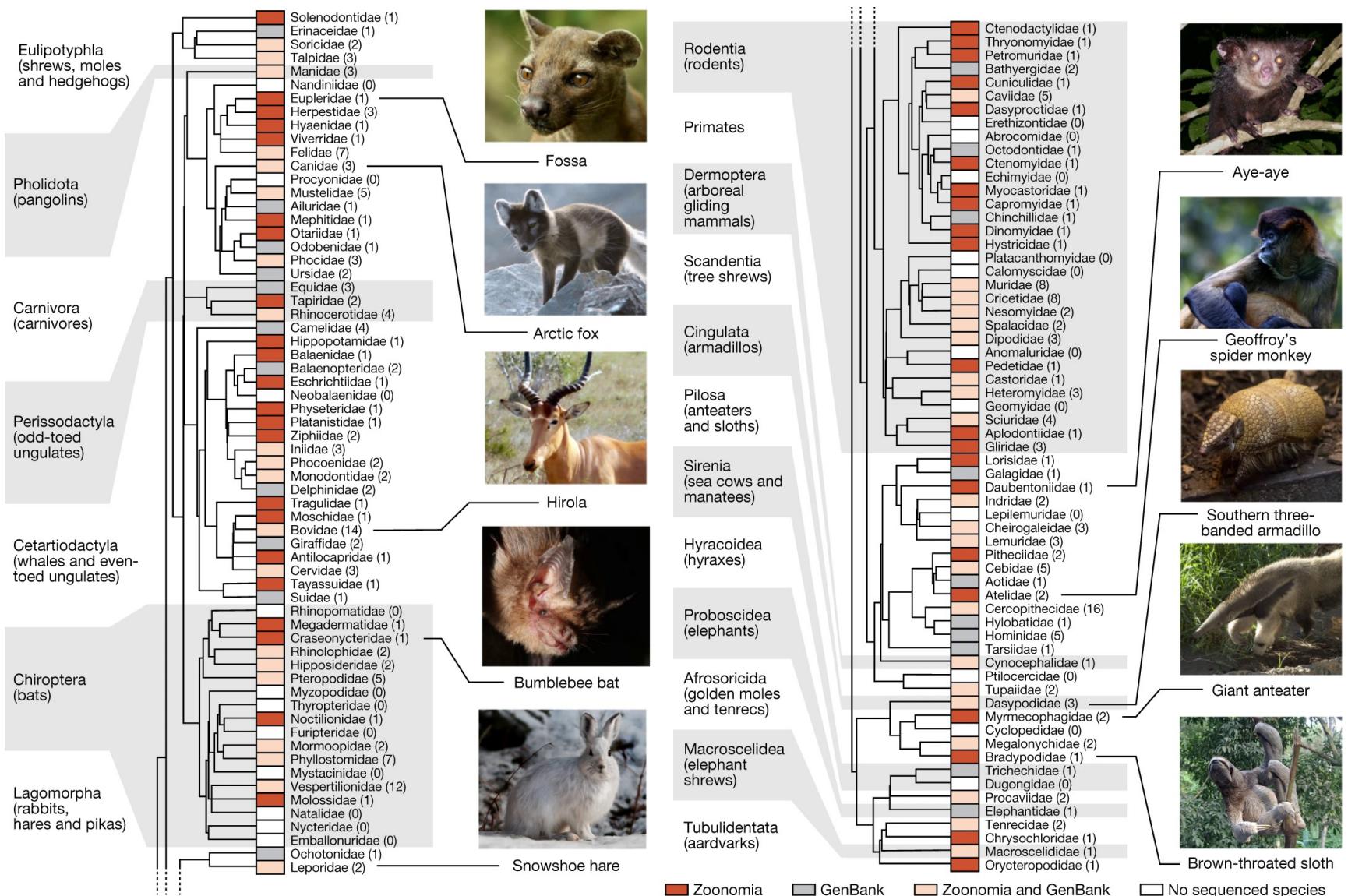
Cactus alignment informs synteny

- *GH* one copy in chicken, two in Passeriformes
- Based on synteny, *GH_L* is ancestral, and not *GH_S* even though it is highly similar.



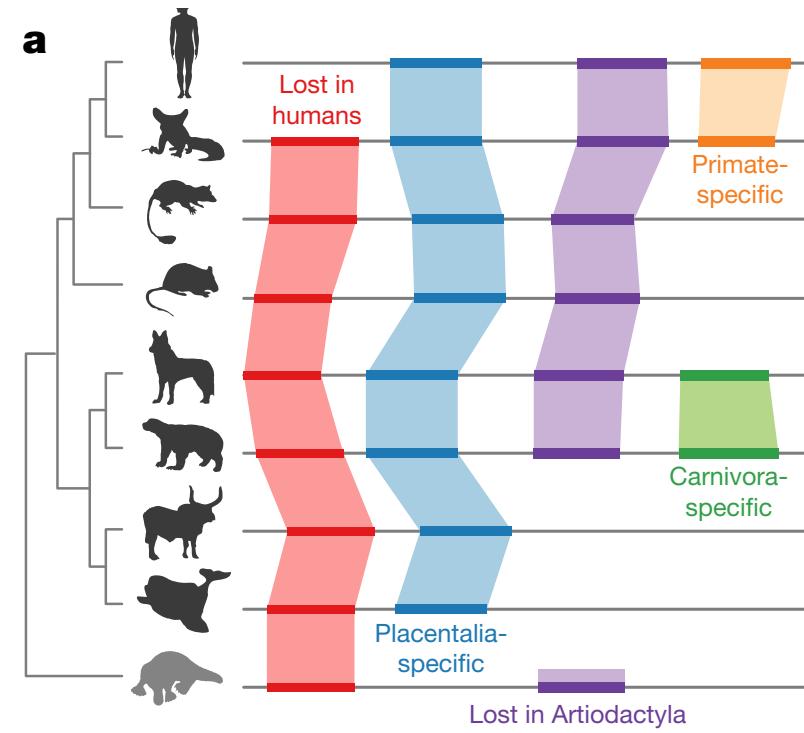
- 240 mammalian species, covering 83 % of families
- 131 new species

Zoonomia



Uses of the alignment

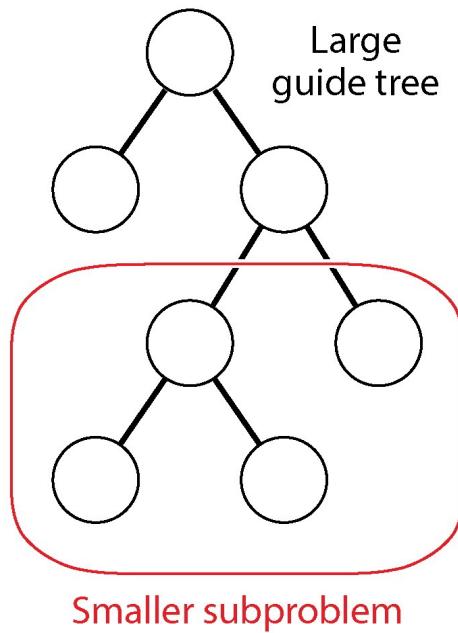
- Alignment has been used to compare the structure of ACE2 (receptor for SARS-CoV-2) and identified 47 mammals with high likelihood of being virus reservoirs (Damas et al 2020; PNAS)
- Reference free alignment detects sequence absent from human and other lineages



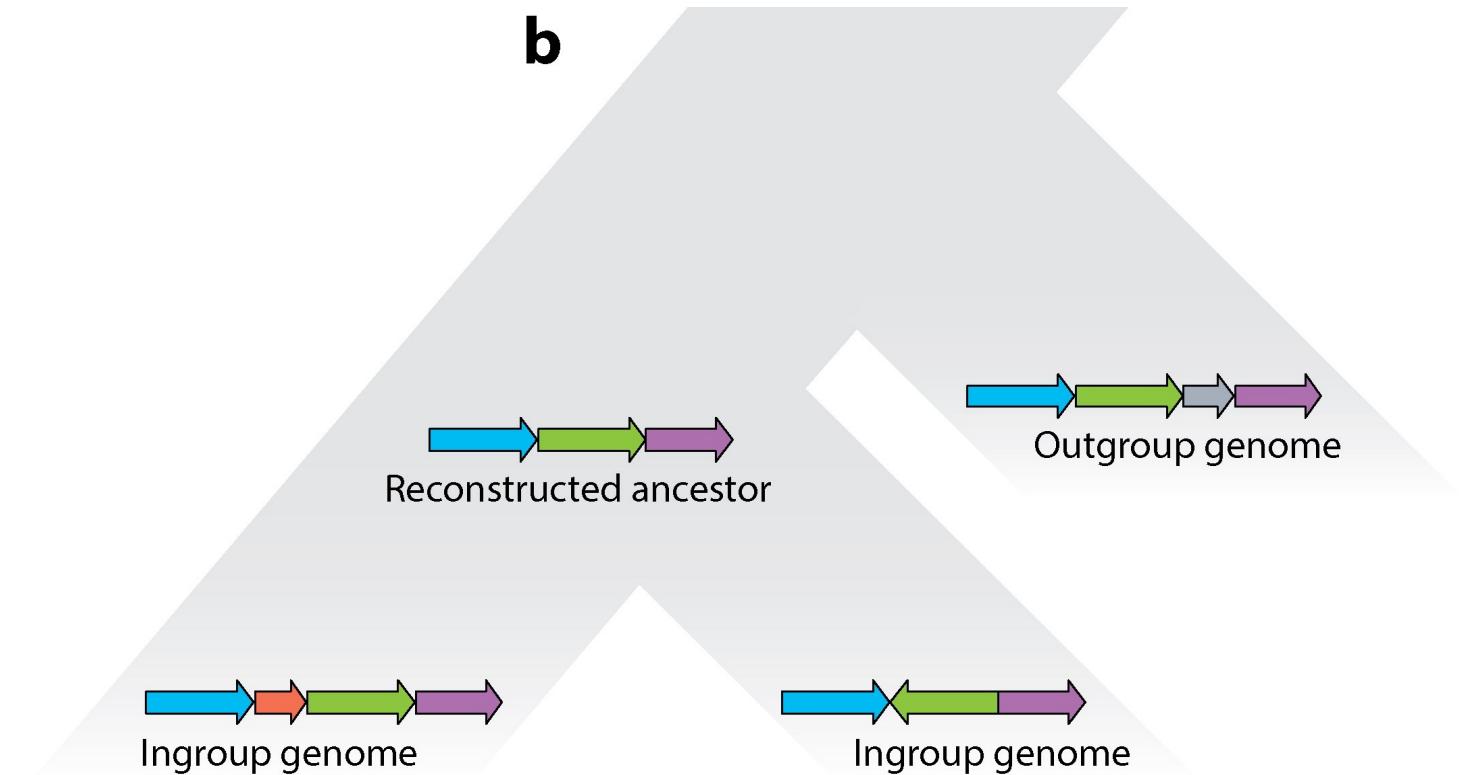
Zoonomia Consortium et al 2020

Progressive Cactus

a



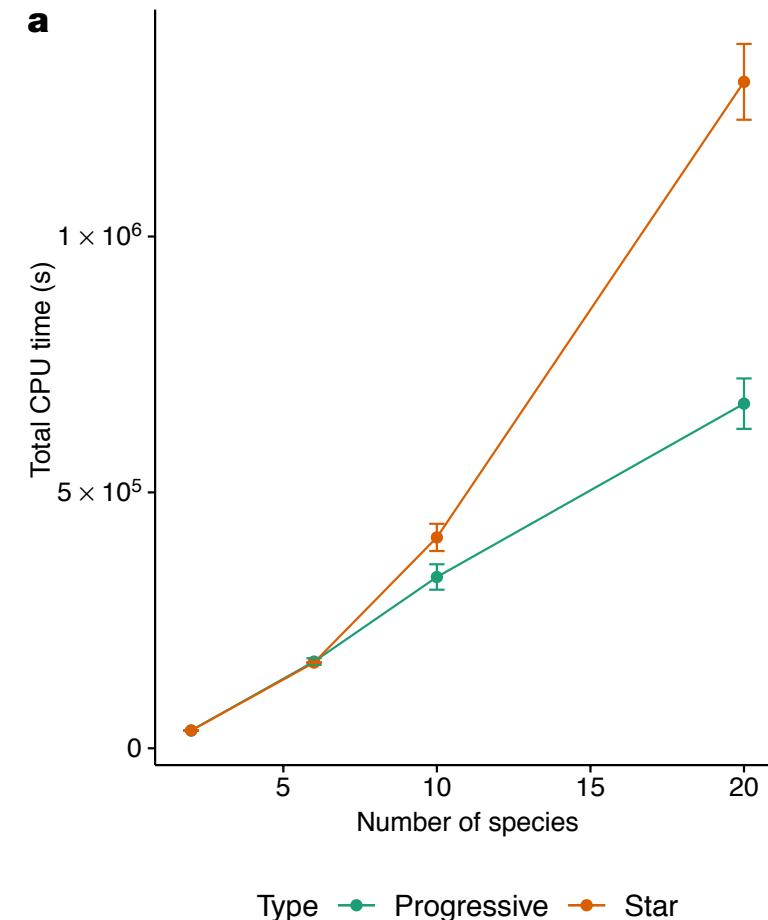
b



Armstrong J, et al. 2019.
Annu. Rev. Anim. Biosci. 7:41–64

Progressive vs star

- Progressive uses less time than star
- Without progressive alignment in some way, aligning many genomes would take too much time/effort/computing power

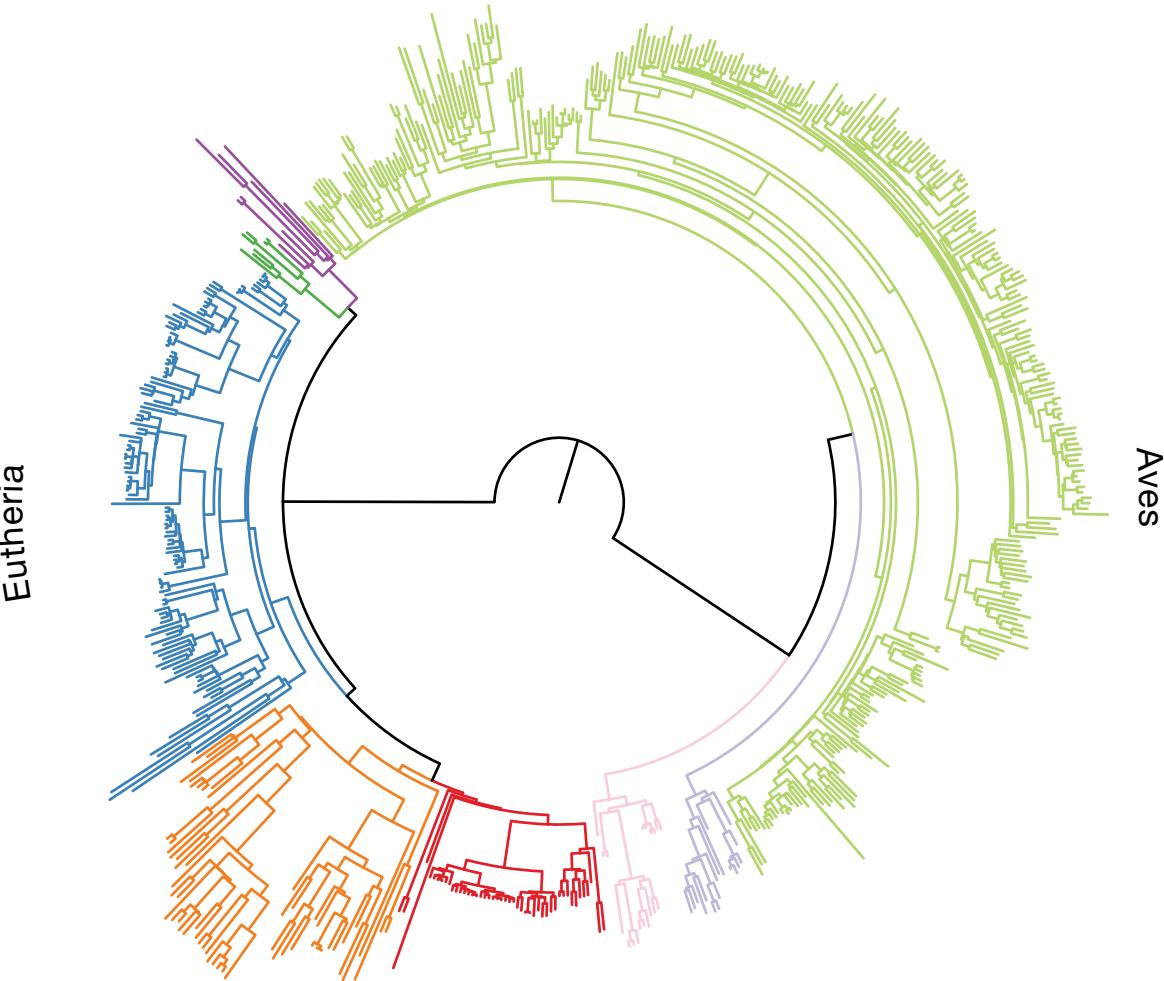


605-way alignment

Alignment	No. of genomes	Total bases	Instance-hours	Core-hours	Common ancestor size
Zoonomia	242	669 billion	68,166	1.9 million	1.73 Gb
B10K	363	400 billion	5,302	0.2 million	1.13 Gb
Combined	605	1.07 trillion	73,692	2.1 million	181 Mb

The increase in computational work for the mammal alignment compared with the bird alignment is largely caused by the increase in the pairwise alignment phase runtime because it scales quadratically with the size of the genomes being aligned.

a



Questions?

