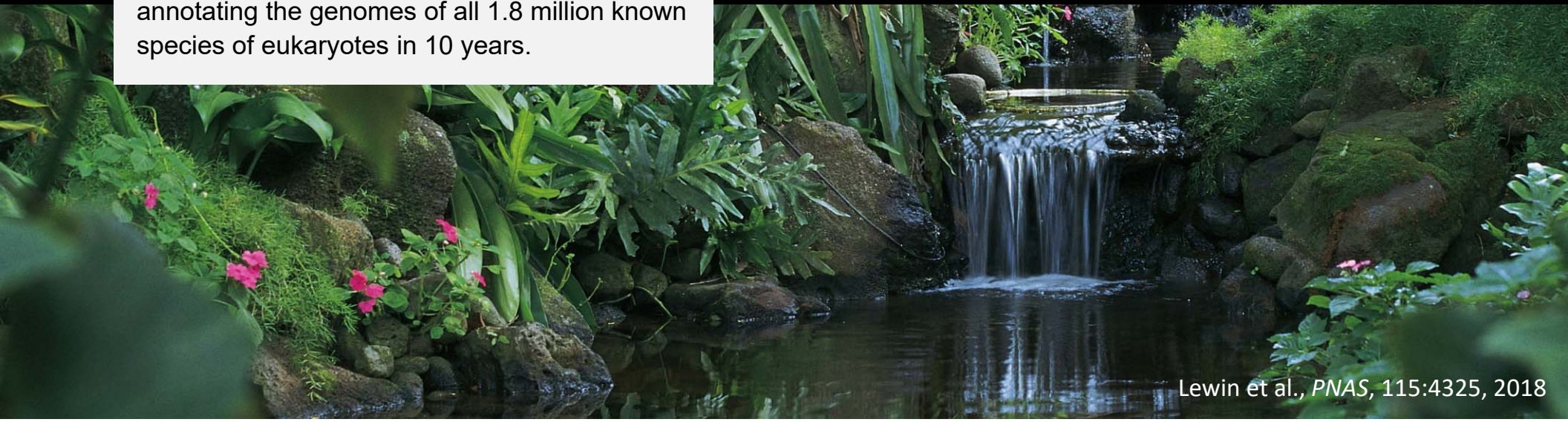
The Earth BioGenome Project: Progress and Challenges Ahead

Harris A. Lewin
University of California, Davis
Chair, EBP Working Group

www.earthbiogenome.org
Twitter: @EBPgenome

What is the Earth BioGenome Project?

The Earth BioGenome Project is a confederated international network-of-networks that has the common goal of sequencing and annotating the genomes of all 1.8 million known species of eukaryotes in 10 years.



Lewin et al., *PNAS*, 115:4325, 2018

bAdGer by Mark Blaxter

Why Sequence all Eukaryotes?

Blaxter et al., *PNAS* 2022

Building genomics-informed conservation

Understanding ecosystem function, stasis, and change

Revealing the deep logic of eukaryotic gene regulation

Tracking genomic changes in symbiosis

Decoding the genomics of complex traits

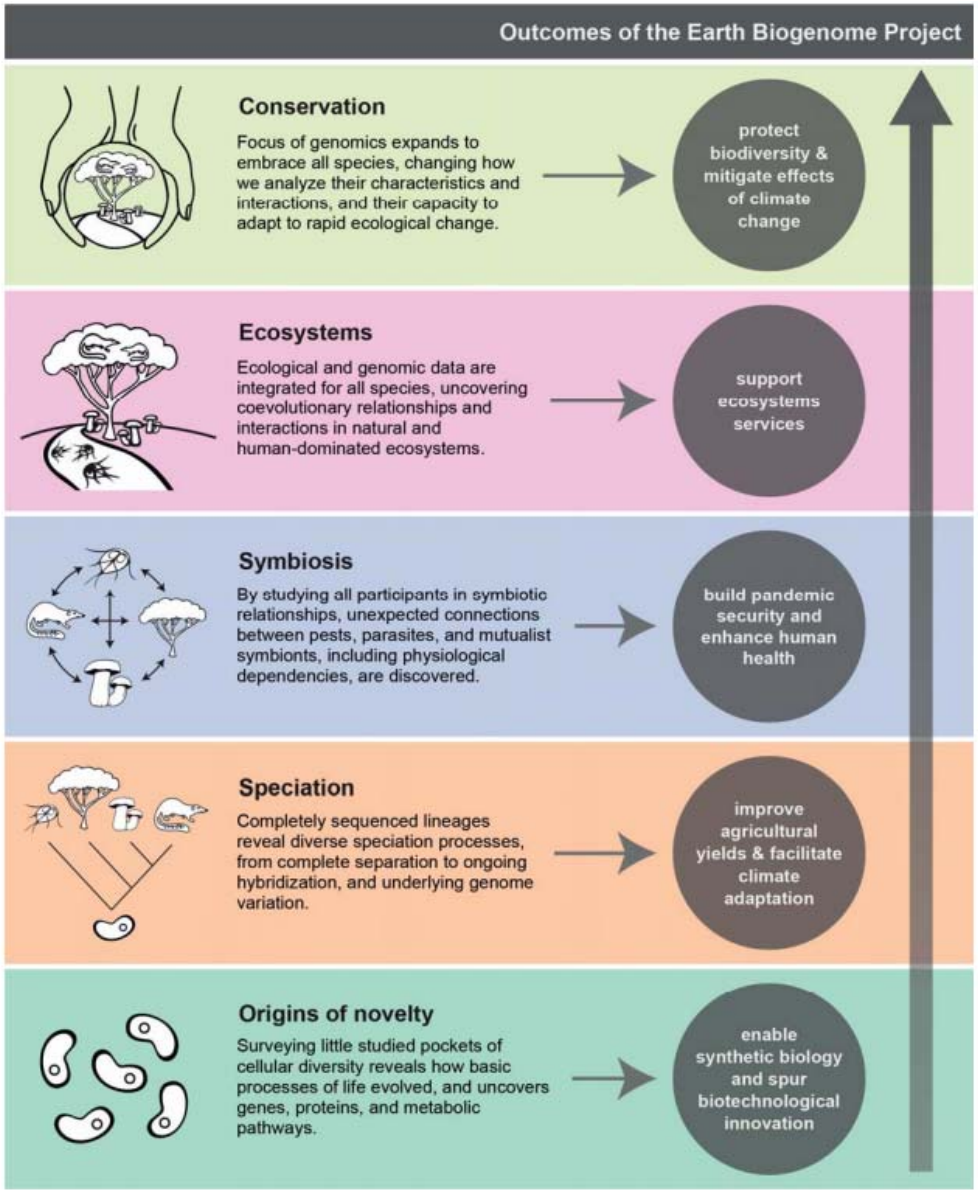
Probing the diversity of sexual systems

Decrypting chromosome evolution

Exploring diversity in the genomics of speciation

Defining the origin of eukaryotic cells

Discovering the trees of life



The Earth BioGenome Project:

A Confederated International **Network-of-Networks**

Global hub and spokes model

43
institutions members

50
Affiliated Projects

+5000
Participants

22
countries

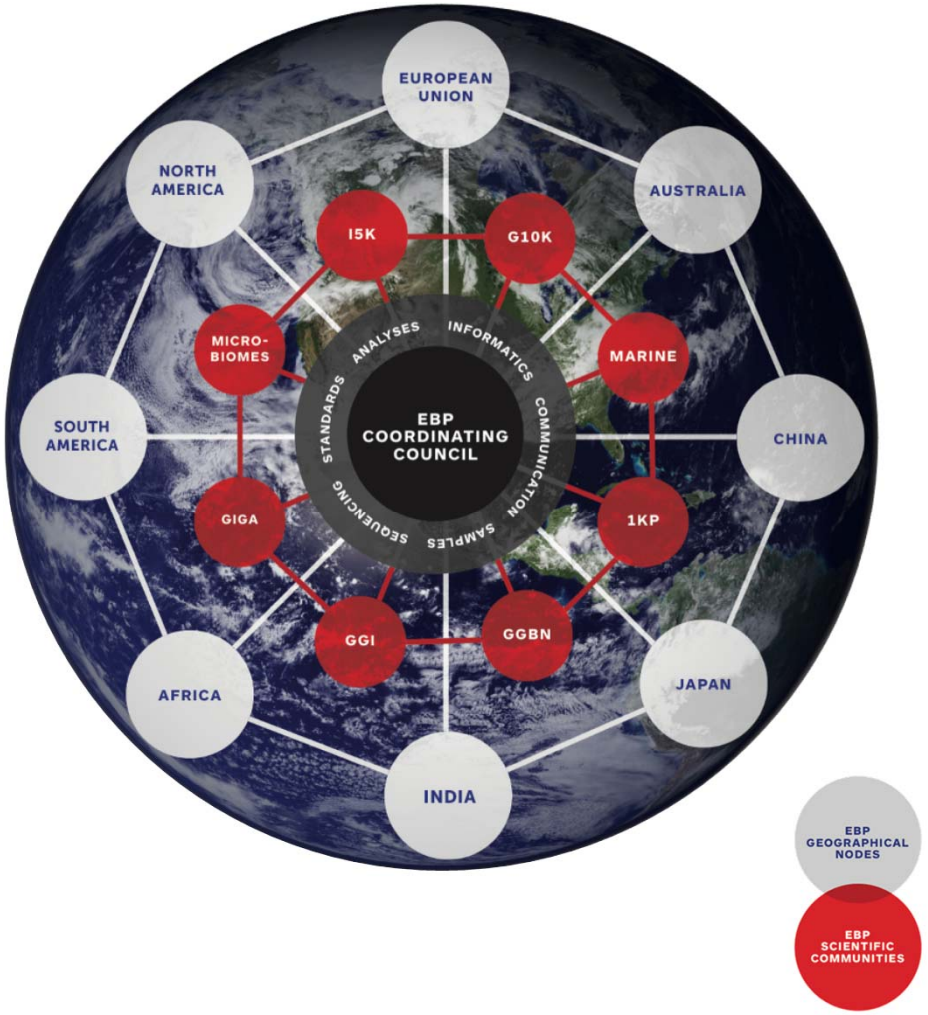
The Earth BioGenome Project:

A Confederated International **Network-of-Networks**

Committed to open data access and compliance with the Convention on Biological Diversity and the Nagoya Protocol on Access and Benefit Sharing (ABS)

Committed to diversity, equity, inclusion and justice among the EBP community of scientists and peoples of the world.

Model of EBP Network-of-Networks



International Scientific Committee (ISC)

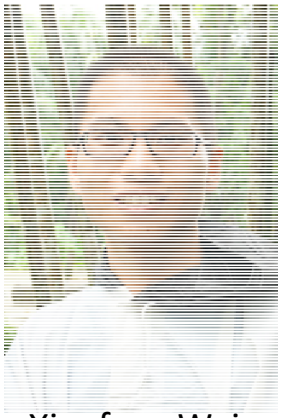
Federica Di Palma
ISC Chair

Mara Lawniczak
Sample Collection and Processing

Richard Durbin
Sequencing and Assembly

Paul Flicek
Annotation

Kerstin Lindblad-Toh
Data Analysis



Xiaofeng Wei
IT and Informatics

EBP Strategy 1: The Phylogenomic Wave

- Domains: 3 (Eubacteria, Archaea, Eukarya)
- Eukaryotic Kingdoms: 5 (animal, plant, fungi, chromista and protozoa)
- Eukaryotic Phyla: 70 (67) (33 animal; 8 plant; 5 fungi; 24 chromists+protozoa
- Eukaryotic Classes: 287 (281)
- Eukaryotic Orders: 1,305 (1,383)
- Eukaryotic Families: 9,302 (9,630) (Phase I; reference quality, Y1-Y3)
- Eukaryotic Genera: 160,000-200,000 (119,000) (Phase II, Y4-Y7)
- Eukaryotic Species: ~1.8 million known (1,550,000) (Phase III, Y8-Y10)

Sources: Global Genome Initiative (Jonathan Coddington & Mike Trizna); Catalog of Life (Luisa Abucay & Yuri Roskov)
NCBI, September, 2018

EBP Strategy 2: Ecosystem Sequencing; "Google Life"

- Location Sampling (e.g. Ocean Sampling Day Consortium; Genomic Observatories Network; NEON; Critical Zone Observatory; CAL*e*DNA)
- Sequence all organisms in a particular geographical area (e.g., within biodiversity hotspots); soil, land, water and air (e.g., Wytham Woods, Darwin Tree of Life Project)
- Enables studies of the effect of environmental change on biodiversity (bio-surveillance & genomic ecology)

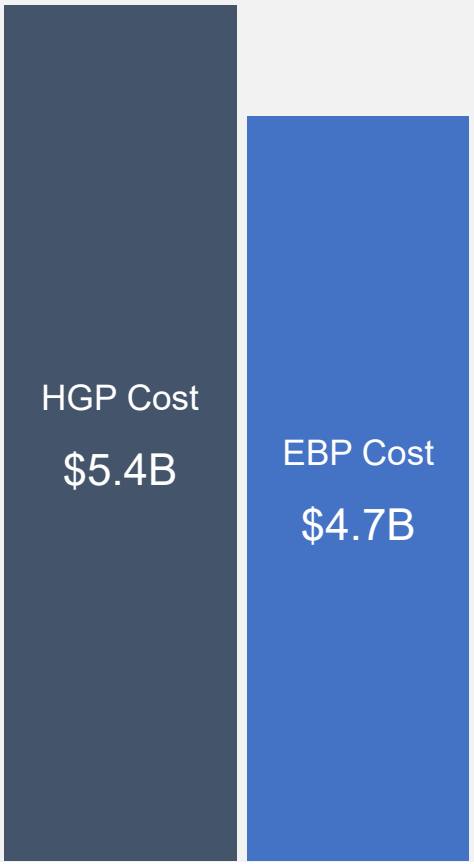
What will it cost and what are the expected returns?

The Human Genome Project (HGP)

Cost > $5.4 billion in 2017

+ other investments in genomics = $14.5 billion

ROI of the HGP estimated at ~$1 trillion, or 65:1 (Battelle Report, 2013) in the US alone



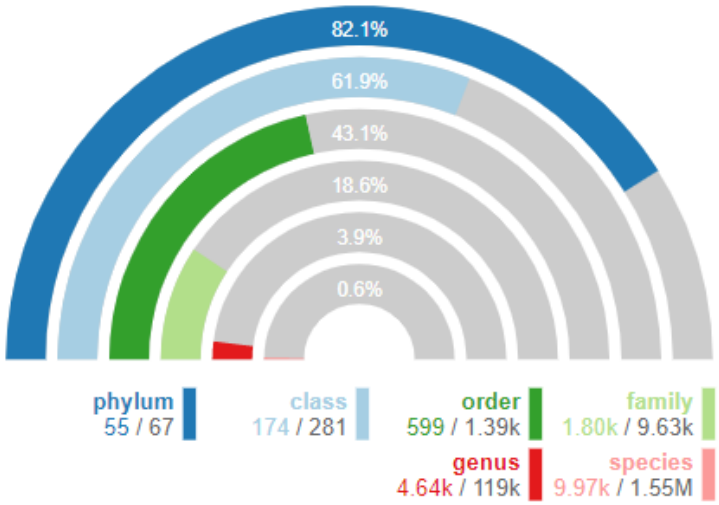
The Earth BioGenome Project (EBP)

$4.7 billion over 10 years (globally)

ROI of the EBP will likely be many times more than the HGP

Eukaryotic Genome Sequencing: Current Status

World Status



All INSDC taxa - Taxa with assemblies out of all Eukaryotic taxa in INSDC

EBP Umbrella - EBP taxa with assemblies out of all Eukaryotic taxa in INSDC

https://goat.genomehubs.org/

July 12, 2022

Genomes on a Tree
- GoaT -

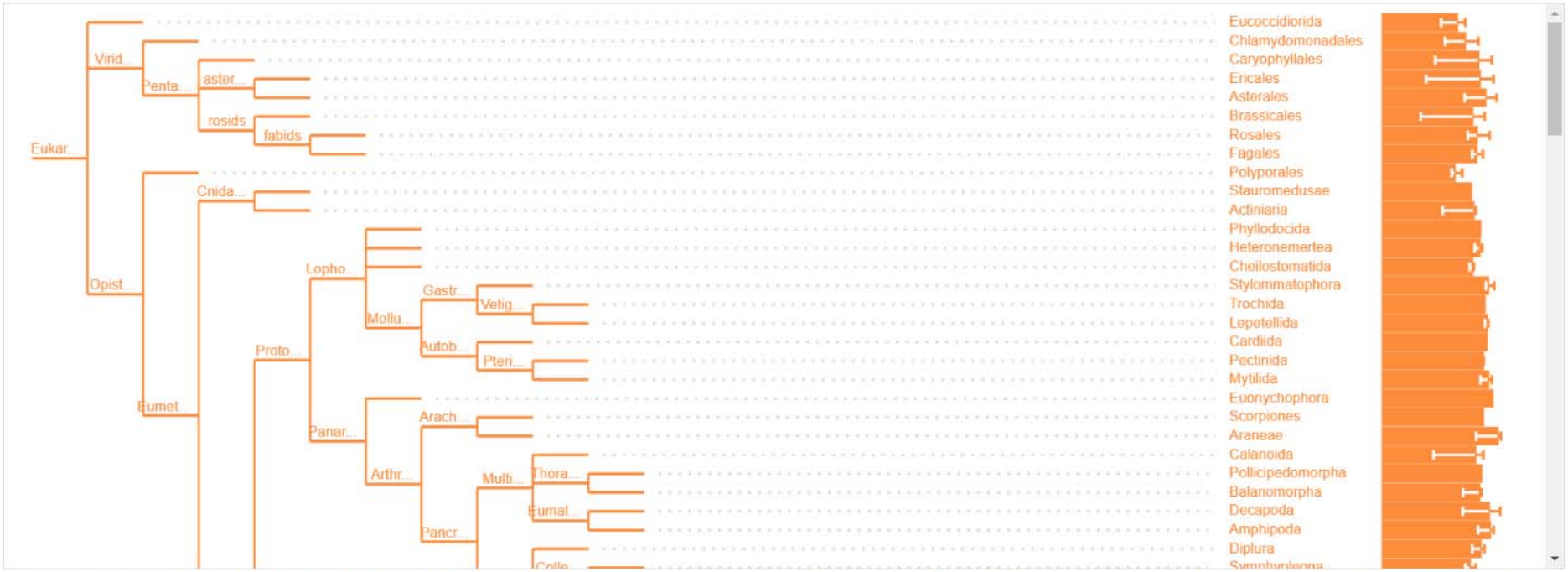
https://goat.genomehubs.org

GoaT is a platform that stores genome-relevant metadata for Eukaryotic species

Cibele Sotero-Caio, Sujai Kumar, Rich Challis, Mark Blaxter

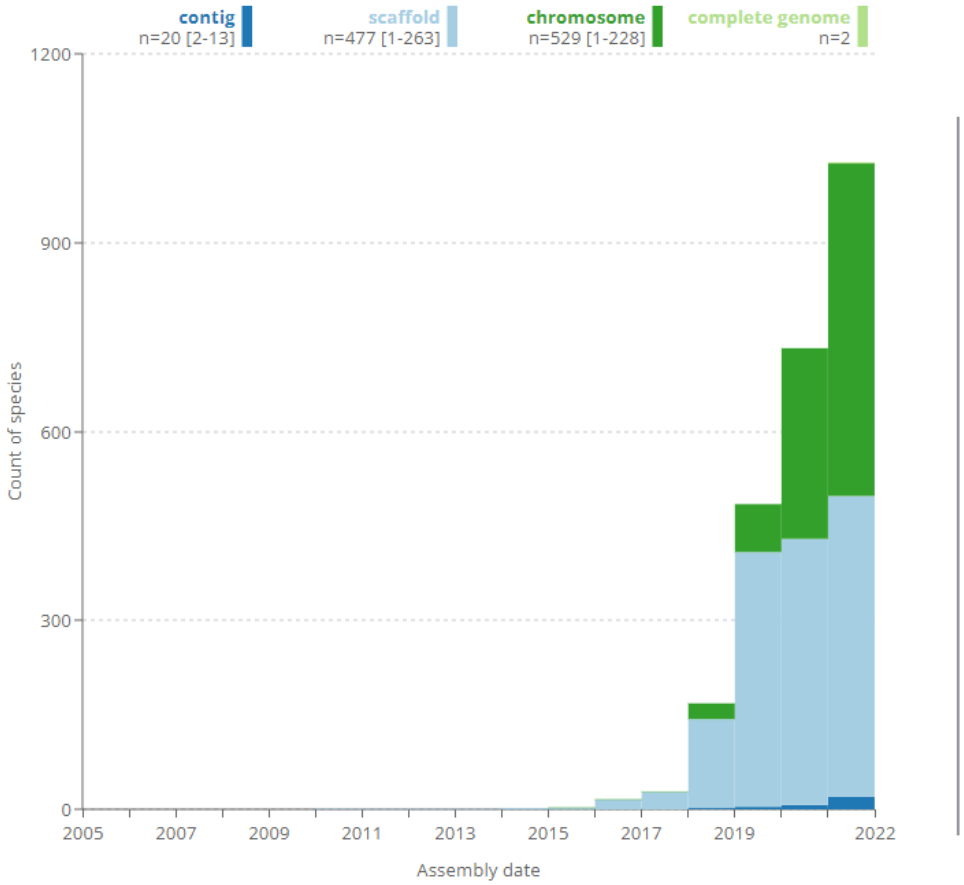
EBP Contribution to Eukaryotic Genome Sequencing

Tree representing the eukaryotic orders with at least one species sequenced by the EBP network

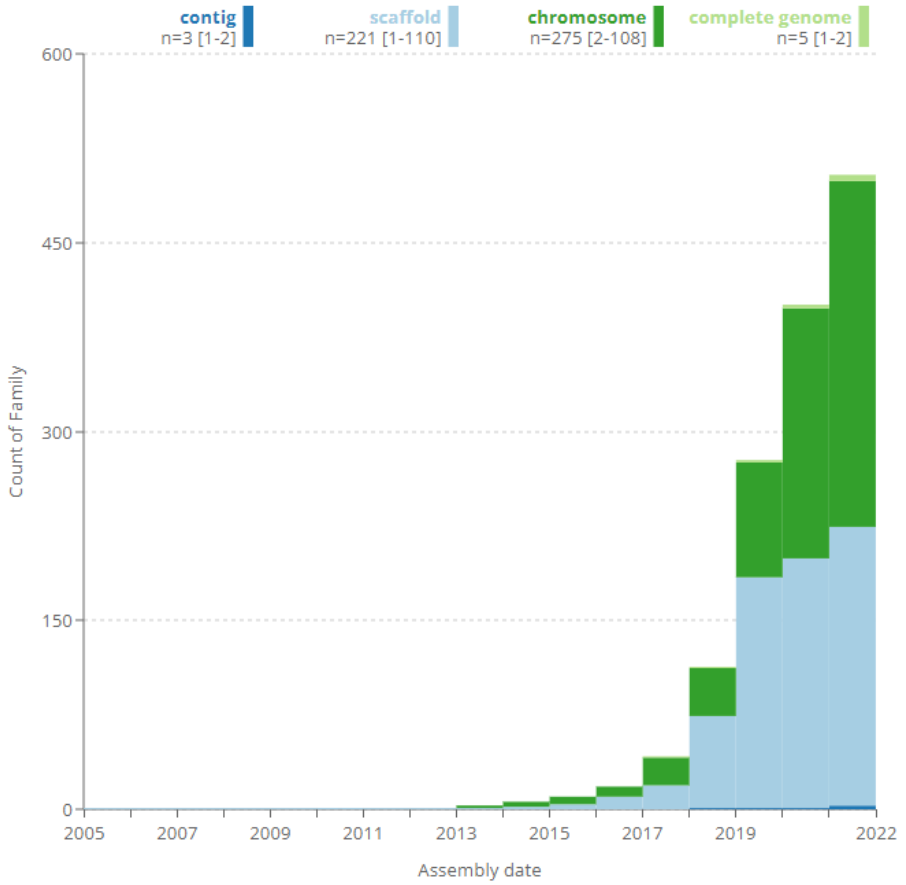


Orders with at least one species sequenced by the Earth Biogenome Project. Orange highlights represent clades with at least one assembly published under the EBP umbrella BioProject ID (PRJNA533106). Bars correspond to estimates (orange) or direct (green) assembly span values for each taxon. Tap tree nodes to see taxon records or long-press to expand each branch.

EBP Progress: Year over Year

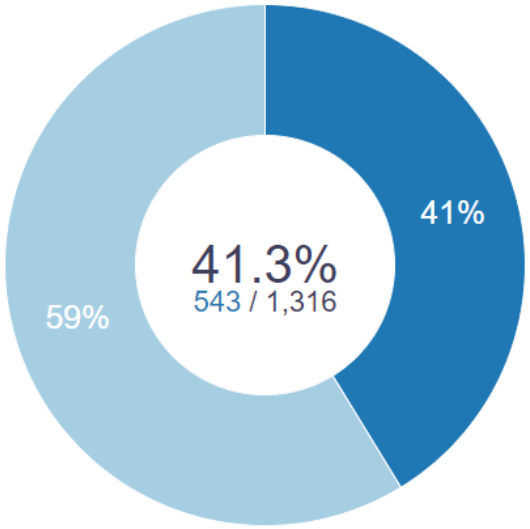


Cumulative number of assemblies for eukaryotic species generated by EBP affiliates over time

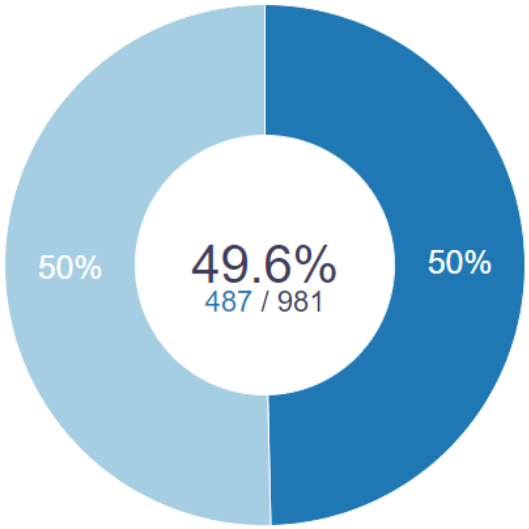


Cumulative number of eukaryotic families for which assemblies were generated by EBP affiliates over time

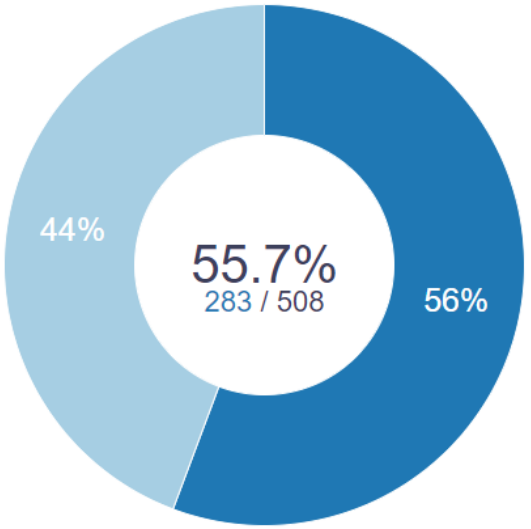
Contribution of EBP to Genome Assemblies Meeting EBP-Standard Metrics (contig n50 >= 1MB; scaffold n50 >= 10MB)



Contribution of EBP to total of **species** with assemblies meeting EBP metrics



Contribution of EBP to total of **genera** with assemblies meeting EBP metrics



Contribution of EBP to total of **families** with assemblies meeting EBP metrics

A taste of what's to come!

28 April 2021

Erich Jarvis, Chair

GENOME 10K

VGP VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

- Near-complete genomes for 16 vertebrate species
- Changed what's known about several species: discovered previously unknown chromosomes in platypus and zebra finch
- Discovered key differences between marmoset and human brain-related genes
- Discovered new insights into evolution of key neurochemicals oxytocin and vasotocin
- Gained new insights for conservation: Kākāpō, a critically endangered parrot from New Zealand; is able to purge deleterious mutations from its genome despite low genetic diversity (Dussex *et al*., *Cell Genomics,* 2021)

Conserve & regenerate biodiversity

Elinor Karlsson

Kerstin Lindblad-Toh

Zoonomia Consortium,
***Nature* 587:240-245, 2020**

Genomic diversity and threatened species status

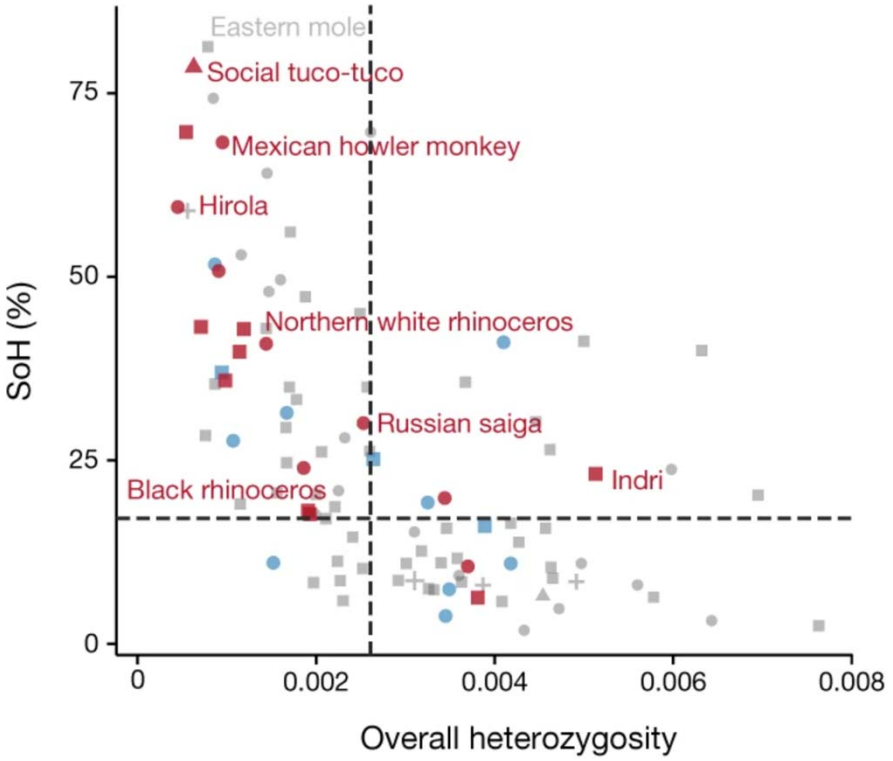
Fossa

Arctic fox

Hirola

Bumblebee bat

Snowshoe hare



Aye-aye

Geoffroy's spider monkey

Southern three-banded armadillo

Giant anteater

Brown-throated sloth

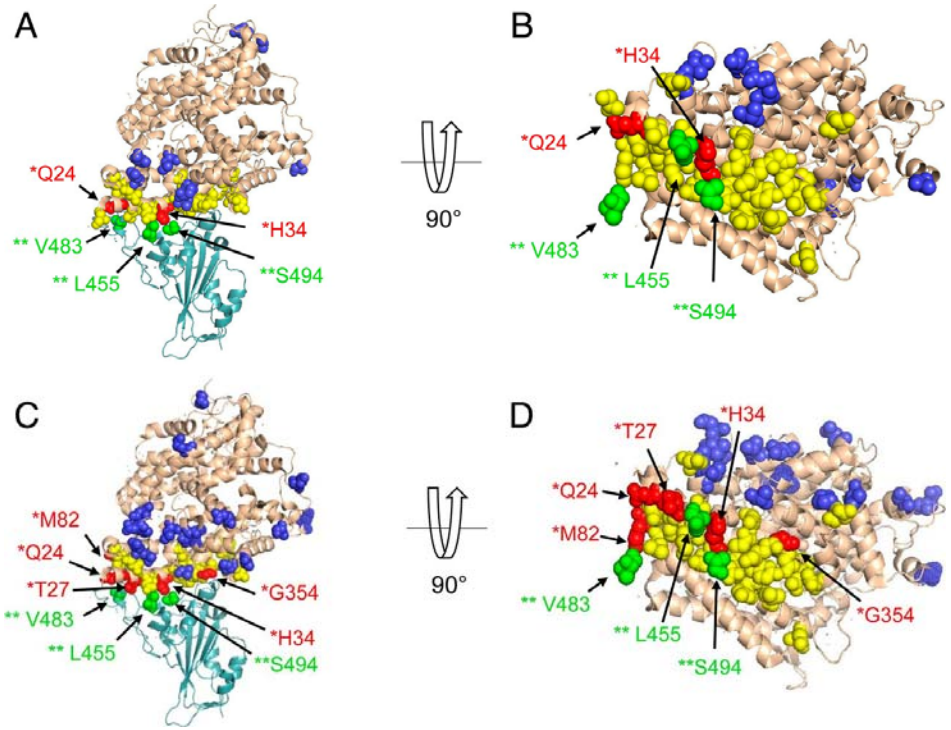
Improving pandemic prediction & responsiveness

PNAS

Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates

Joana Damas, Graham M. Hughes, Kathleen C. Keough, Corrie A. Painter, Nicole S. Persky, Marco Corbo, Michael Hiller, Klaus-Peter Koepfli, Andreas R. Pfenning, Huabin Zhao, Diane P. Genereux, Ross Swofford, Katherine S. Pollard, Oliver A. Ryder, Martin T. Nweeia, Kerstin Lindblad-Toh, Emma C. Teeling, Elinor K. Karlsson, Harris A. Lewin

doi: https://doi.org/10.1101/2020.04.16.045302



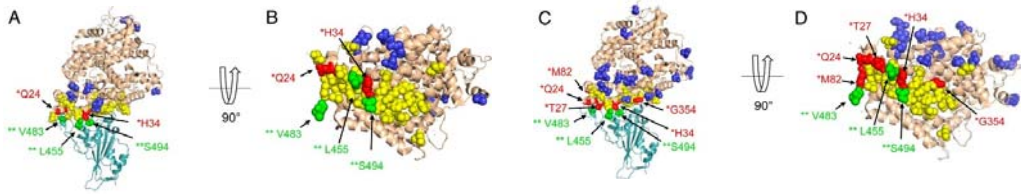
Improving pandemic prediction & responsiveness

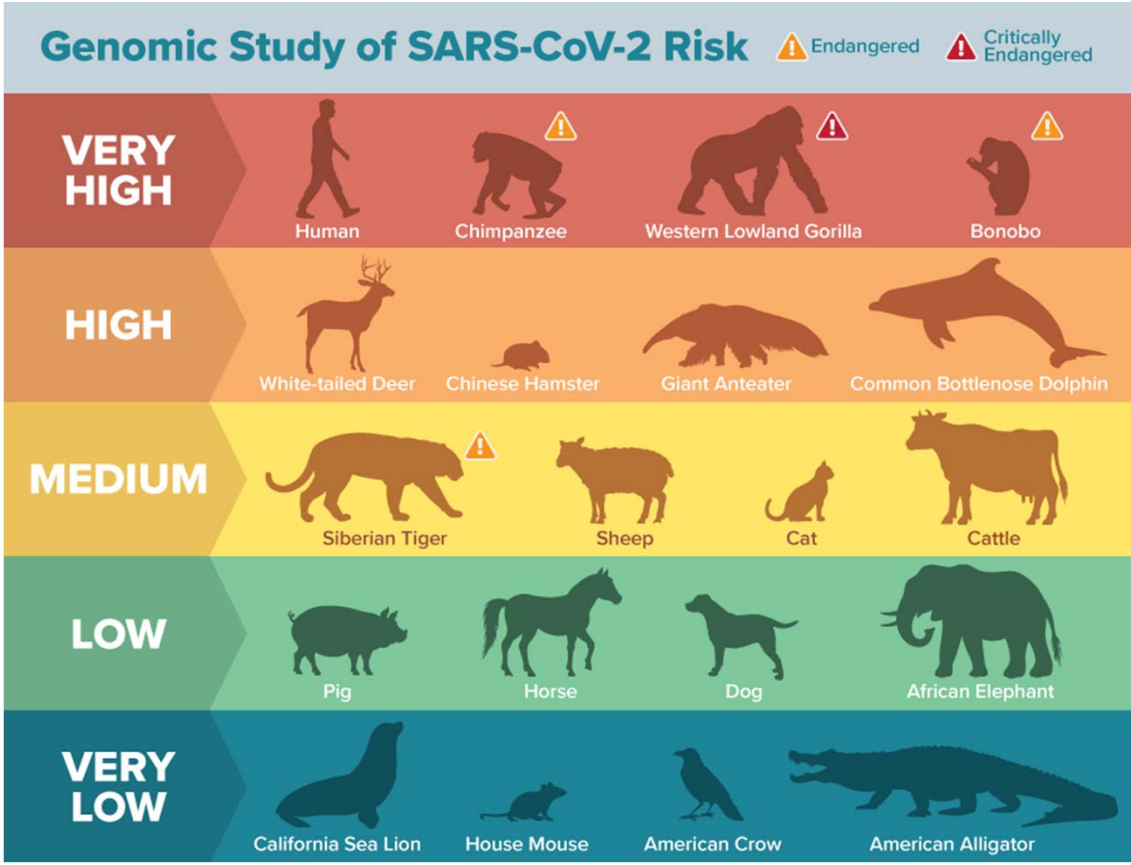
PNAS

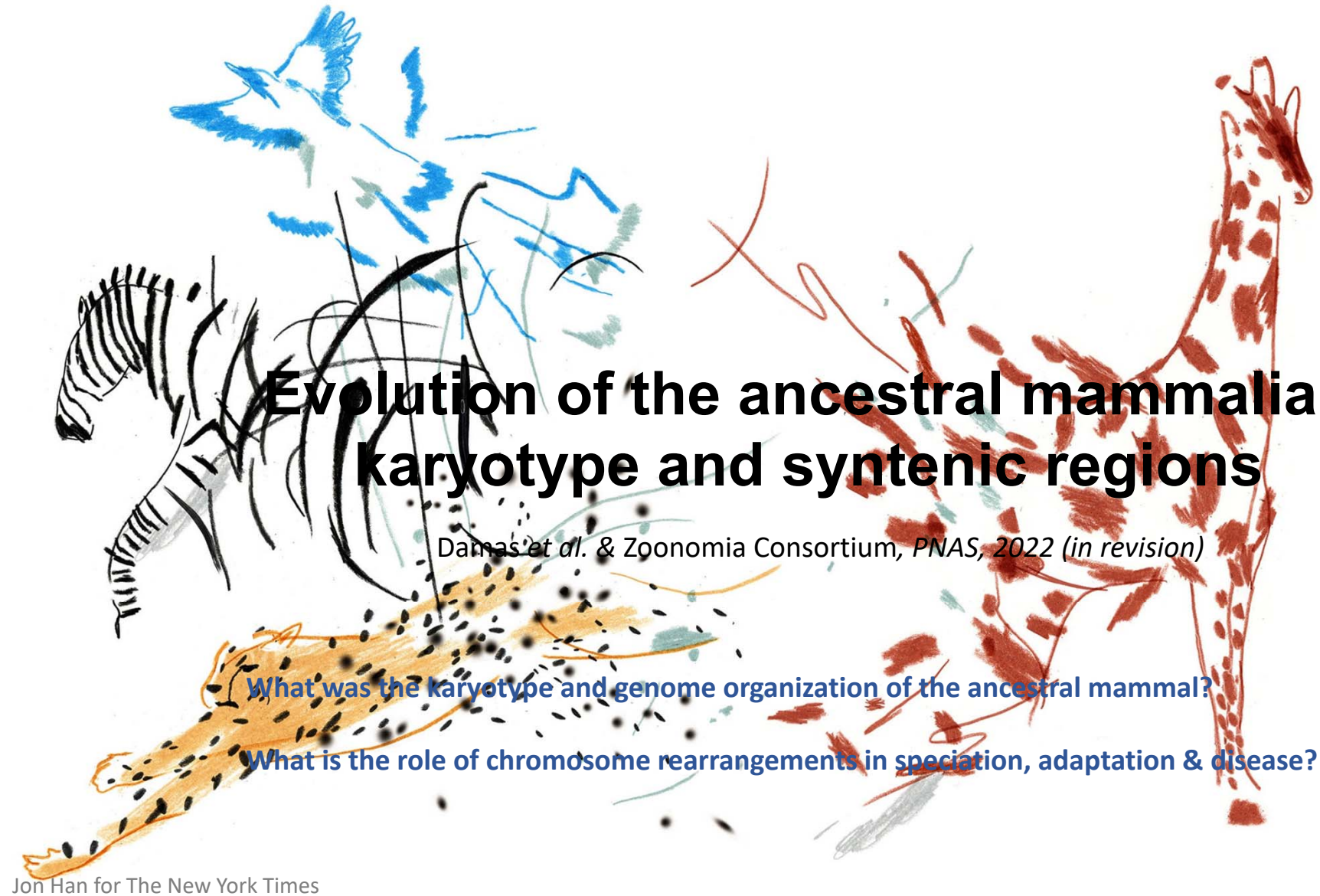
Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates

Joana Damas, Graham M. Hughes, Kathleen C. Keough, Corrie A. Painter, Nicole S. Persky, Marco Corbo, Michael Hiller, Klaus-Peter Koepfli, Andreas R. Pfenning, Huabin Zhao, Diane P. Genereux, Ross Swofford, Katherine S. Pollard, Oliver A. Ryder, Martin T. Nweeia, Kerstin Lindblad-Toh, Emma C. Teeling, Elinor K. Karlsson, Harris A. Lewin

doi: https://doi.org/10.1101/2020.04.16.045302







Evolution of the ancestral mammalian karyotype and syntenic regions

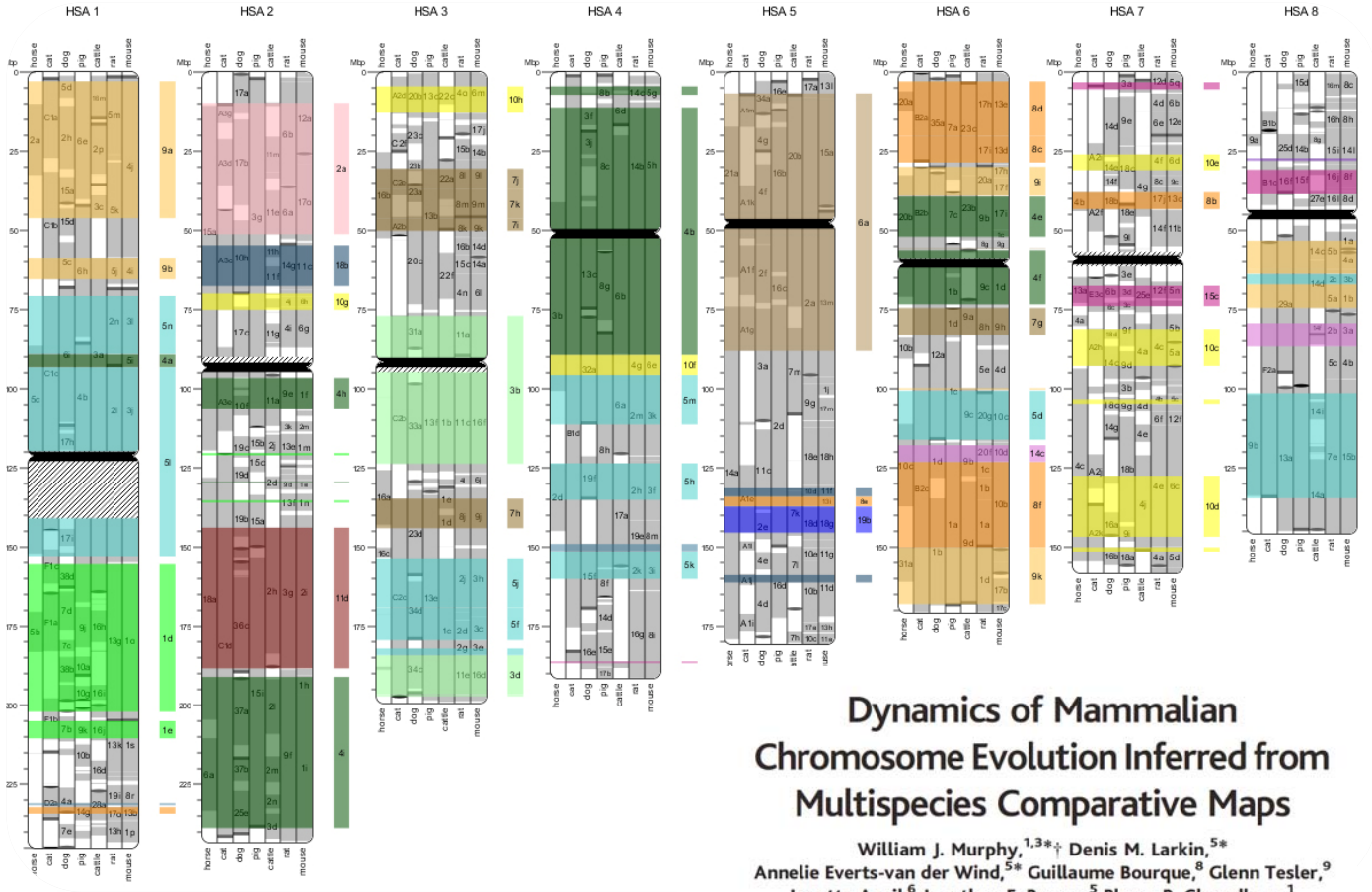
Damas *et al.* & Zoonomia Consortium, *PNAS, 2022 (in revision)*

What was the karyotype and genome organization of the ancestral mammal?

What is the role of chromosome rearrangements in speciation, adaptation & disease?

UCDAVIS GENOME CENTER

Jon Han for The New York Times

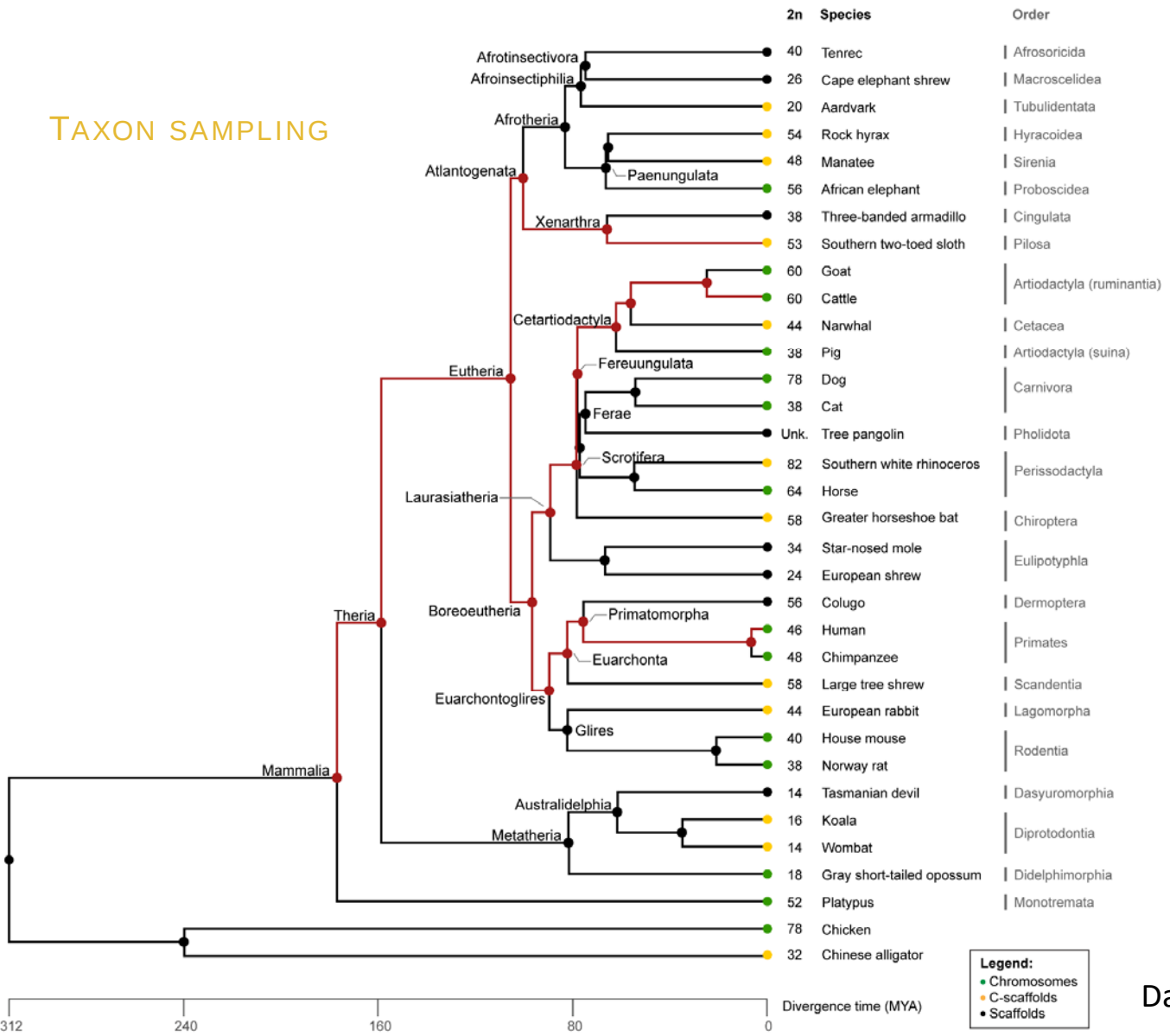


Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps

William J. Murphy,[1,3*†] Denis M. Larkin,[5*] Annelie Everts-van der Wind,[5*] Guillaume Bourque,[8] Glenn Tesler,[9] Loretta Auvil,[6] Jonathan E. Beever,[5] Bhanu P. Chowdhary,[1] Francis Galibert,[11] Lisa Gatzke,[6] Christophe Hitte,[11] Stacey N. Meyers,[5] Denis Milan,[12] Elaine A. Ostrander,[13] Greg Pape,[6] Heidi G. Parker,[13] Terje Raudsepp,[1] Margarita B. Rogatcheva,[5] Lawrence B. Schook,[5,7] Loren C. Skow,[1] Michael Welge,[6] James E. Womack,[2] Stephen J. O'Brien,[4] Pavel A. Pevzner,[10] Harris A. Lewin[5,7†]

Science 309, 2005

TAXON SAMPLING



32 mammalian genome assemblies

- 13 chromosome level assemblies
- 11 C-scaffold assemblies
- 8 scaffold assemblies

Representatives for 24 of 27 mammalian orders

- All (19) Eutherian orders represented
- 3 of 7 Metatherian orders represented
- Monotremata

Two outgroups

- Chicken
- Chinese alligator

Three independent reconstructions using human, sloth and cattle as the reference genome

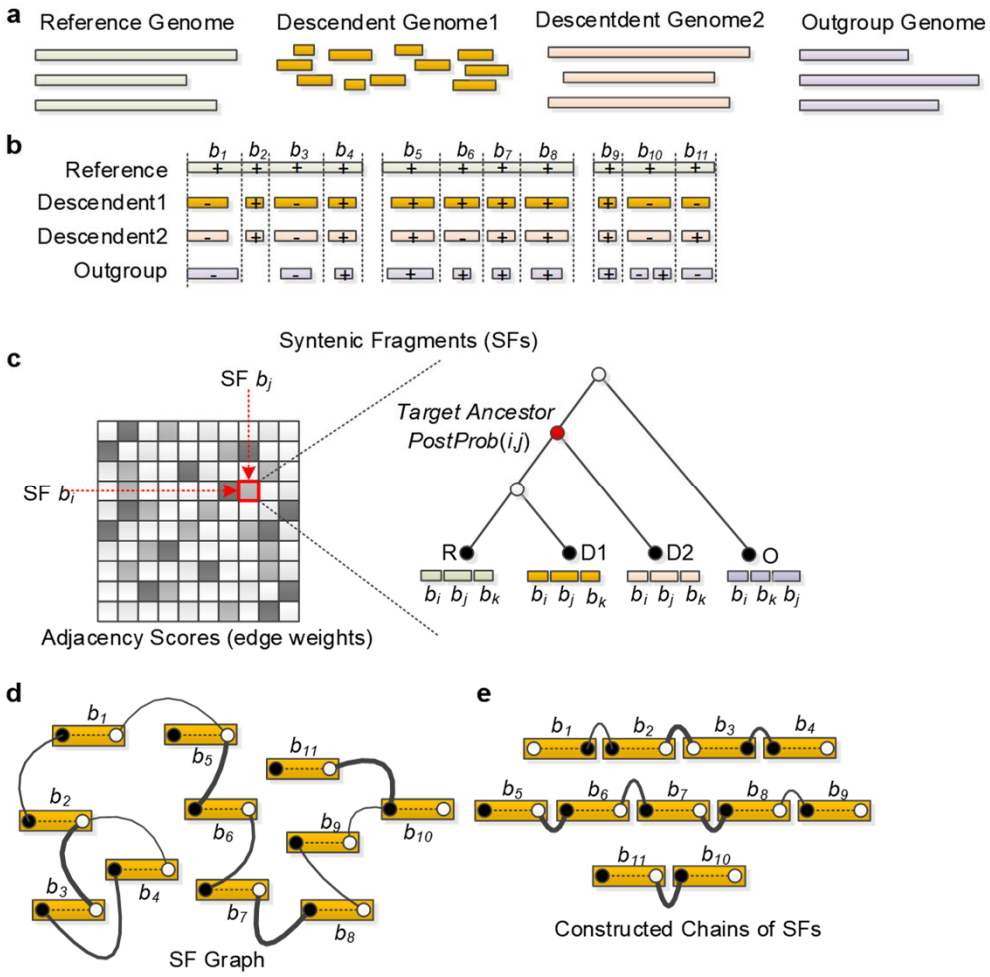
DESCHRAMBLER algorithm
Kim et al., *PNAS* **114**:E5379-E5388, 2017

Damas *et al., PNAS (in revision)*

UCDAVIS GENOME CENTER

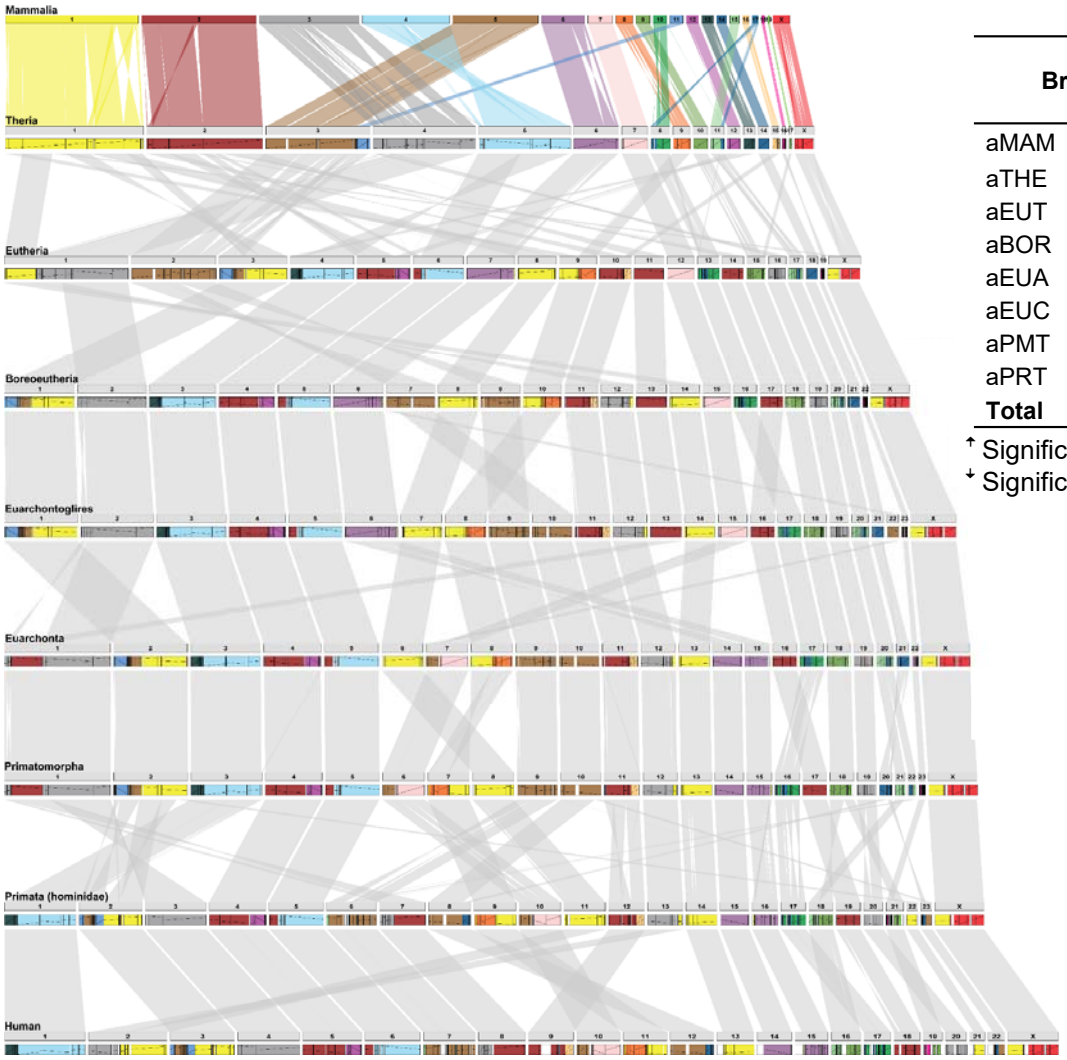
Ancestral Chromosome Reconstruction (DESCHRAMBLER)

derived from RACA method; Kim et al., 2013



(Kim *et al*. *PNAS,* 114: 2017)

EVOLUTION OF MAMMALIAN CHROMOSOMES



Branch			Branch length (MY)	MY from present	No. rearrangements			
					Inversions	Fissions	Fusions	Total
aMAM	→	aTHE	18	177	90 ↑	3	3	96 ↑
aTHE	→	aEUT	53	159	94	16	14 ↑	124
aEUT	→	aBOR	9	106	1 ↓	3 ↓	0 ↓	4
aBOR	→	aEUA	7	97	4	1	0 ↓	5
aEUA	→	aEUC	8	90	9	1	2 ↑	12
aEUC	→	aPMT	6	82	24 ↑	2 ↑	1	27 ↑
aPMT	→	aPRT	69	76	73	4	4	81
aPRT	→	aHSA	7	7	15	0 ↓	1	16
Total					**310**	**30**	**25**	**365**

↑ Significantly higher than average across all branches for respective lineage (FDR $P<0.05$).
↓ Significantly lower than average across all branches for respective lineage (FDR $P<0.05$).

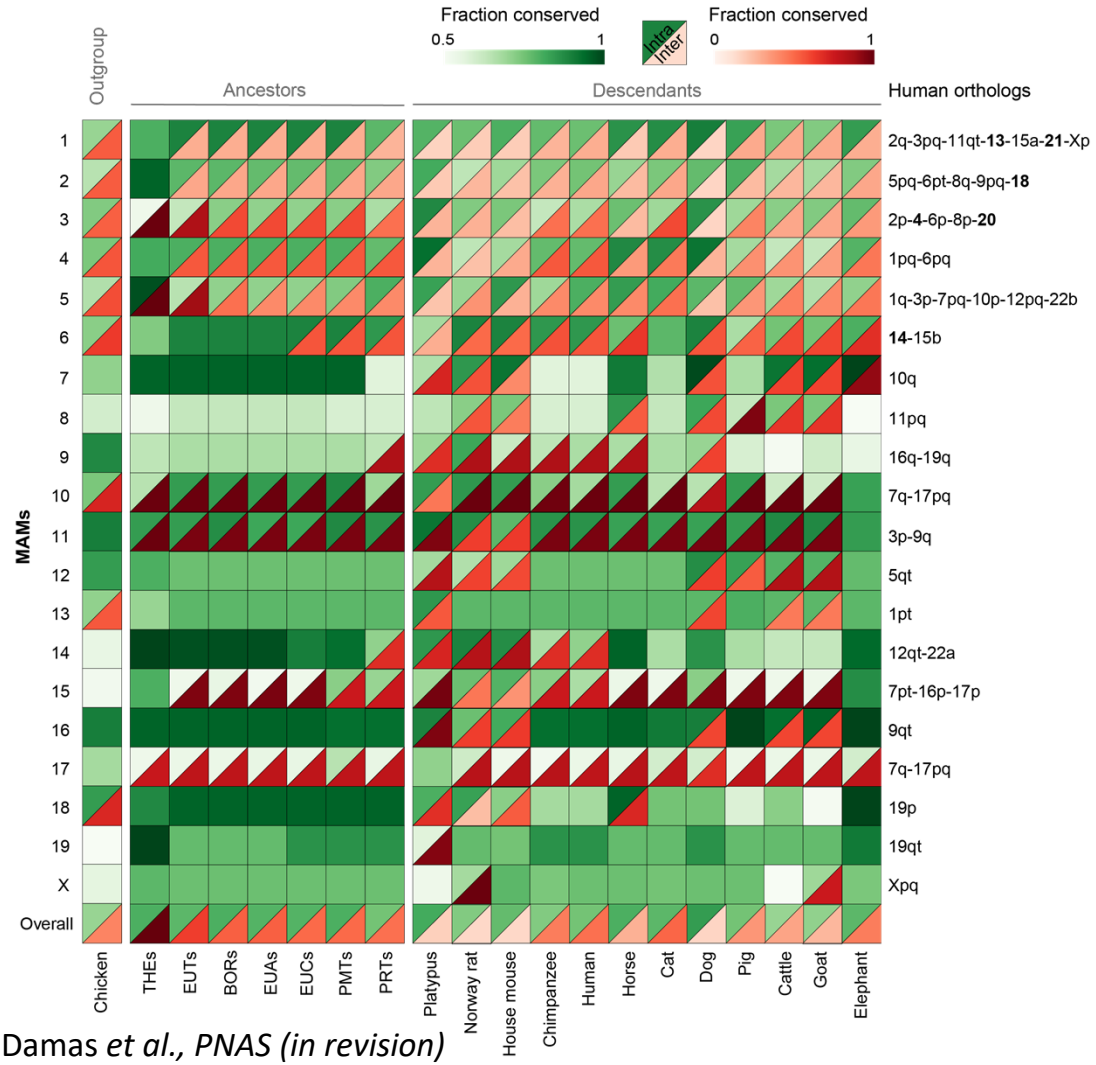
2,557 syntenic segments on average 880 Kbp

The “building blocks” of all extant mammal genomes

Damas *et al., PNAS (in revision)*

UCDAVIS GENOME CENTER

EVOLUTION OF MAMMALIAN CHROMOSOMES



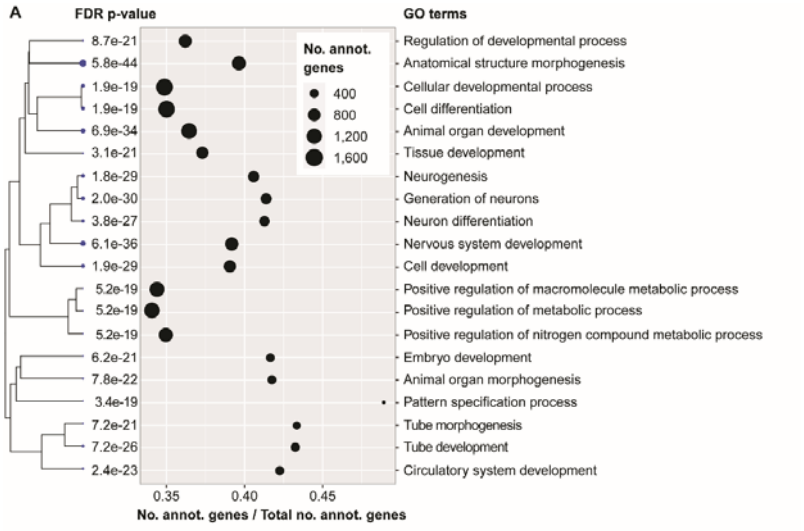
Damas *et al., PNAS (in revision)*

- Larger ancestral mammalian chromosomes (>100 Mbp; MAM 1-6) were more affected by chromosomal rearrangements (fissions & inversions).
- 9/14 smaller ancestral mammalian chromosomes (<100 Mbp; MAM 7-19 and X) had 1:1 orthology with chicken & previously reconstructed avian and amniote ancestors (conserved ~130 My).
- MAM7, MAM14 & MAM19 most conserved.
  - MAM7 was maintained as a single chromosome for >76 My in mammalian evolution, with >95% of its length unaffected by inversions.
- Several MAMs conserved for ~318 My of vertebrate evolution.

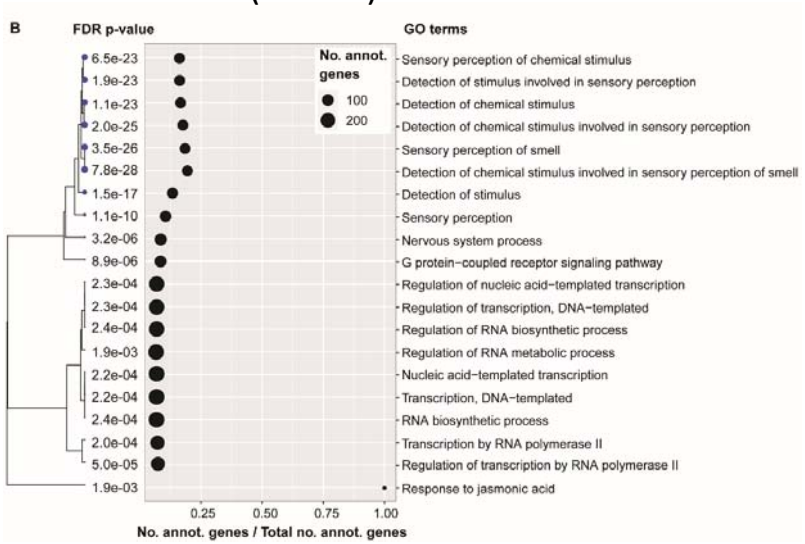
UCDAVIS GENOME CENTER

GENE ONTOLOGY ENRICHMENT FOR MSHSBS AND EBRS

msHSBs (*N*=522 > 1Mbp)



EBRs (*N*=323)

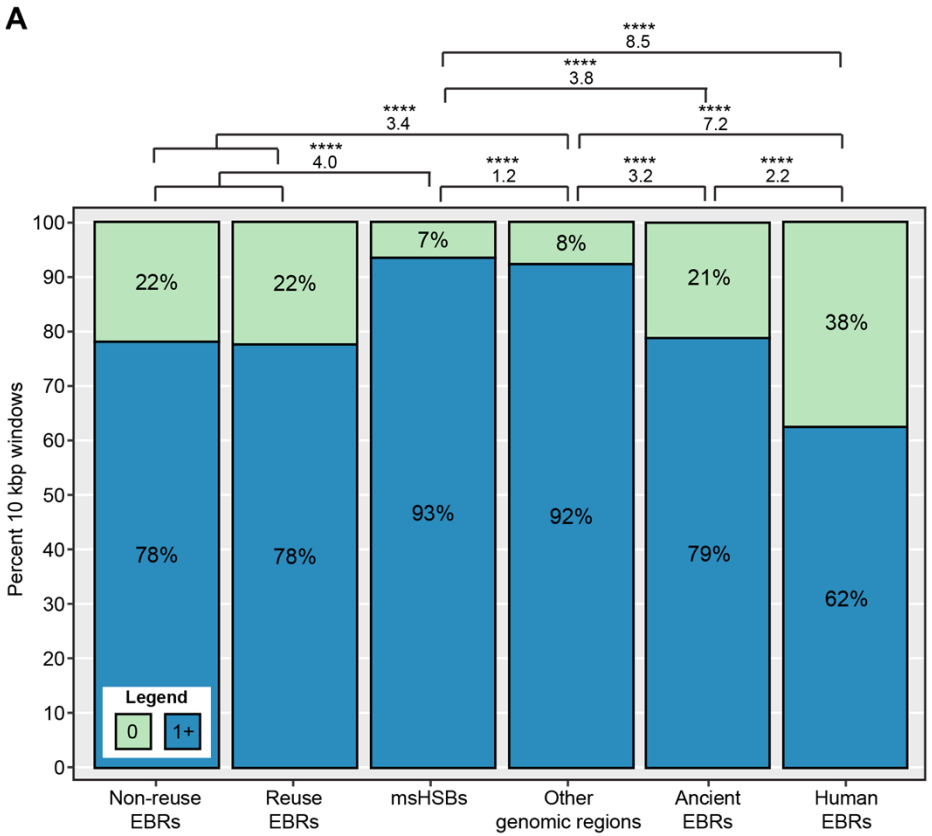


- msHSBs are enriched for genes that have functions related to anatomical and central nervous system development.
- Genes within EBRs are enriched for functions primarily related to sensory perception and regulation of transcription

Damas *et al., PNAS (in revision)*

UCDAVIS GENOME CENTER

TADs in msHSBs, EBRs, and other genomic regions

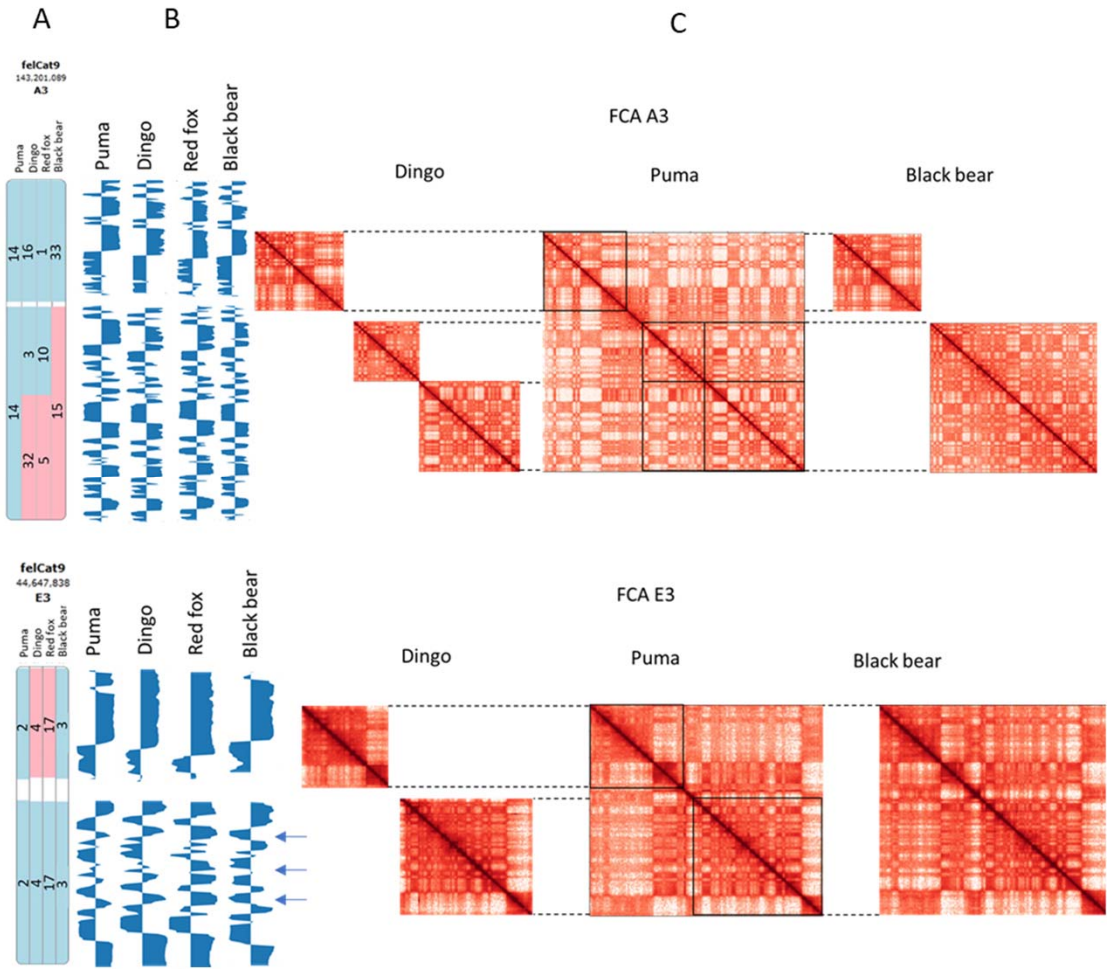


- Human TADS (GM12878 dataset)
- OR of 10 kbp windows in EBRs not having a TAD >4.0x than msHSBs
  - in human-specific EBRs, >8.5x than msHSBs
- msHSBs mostly overlap with TADS and thus appear to be core functional units of chromatin structure and organization, and play a role in coordinated transcriptional control of their internal genes
- EBRs tend to locate between TADs.

Damas *et al., PNAS (in revision)*

UCDAVIS GENOME CENTER

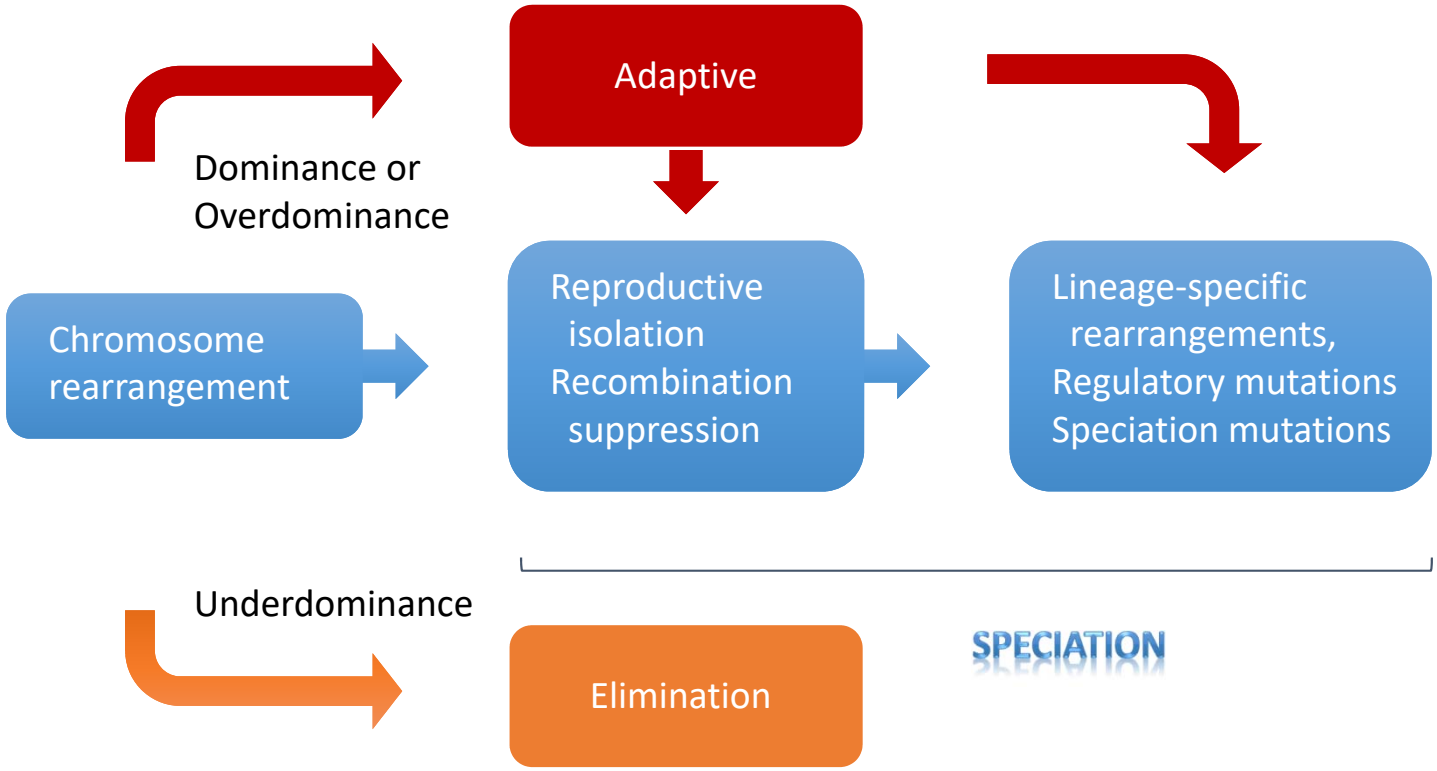
Conservation of chromatin structure for 54 My of carnivore evolution



Corbo *et. al*., 2022. *PNAS,* 119:e2120555119

Chromosome Rearrangements and Speciation

(Chromosome Speciation Model; Quantum Speciation, Simpson, 1944)



Technical & Scientific Challenges Ahead

- **Sourcing, acquiring and vouchering thousands of specimens**
- HMW DNA and RNA isolation at scale
- Managing data workflows internationally
- Sequencing capacity and throughput
- **Capturing and integrating sample metadata**
- **Assembly and curation at scale**
- **Annotation at scale**
- **Whole genome alignments at scale**
- **Resolving phylogenetic relationships**
- **Comparative genomic analysis and data visualization at scale**

- Sequencing capacity & throughput
  - Phase I (9,300 in Y1-3)
    - 9 genomes/day √
  - Phase II (~180,000 in Y4-7)
    - 123 genomes/day (↑14x)
  - Phase III (1.32 M in Y8-10)
    - 1,205/day (↑9.8x)

January 25, 2022 | vol. 119 | no. 4

PNAS

Proceedings of the National Academy of Sciences of the United States of America www.pnas.org

Earth BioGenome Project

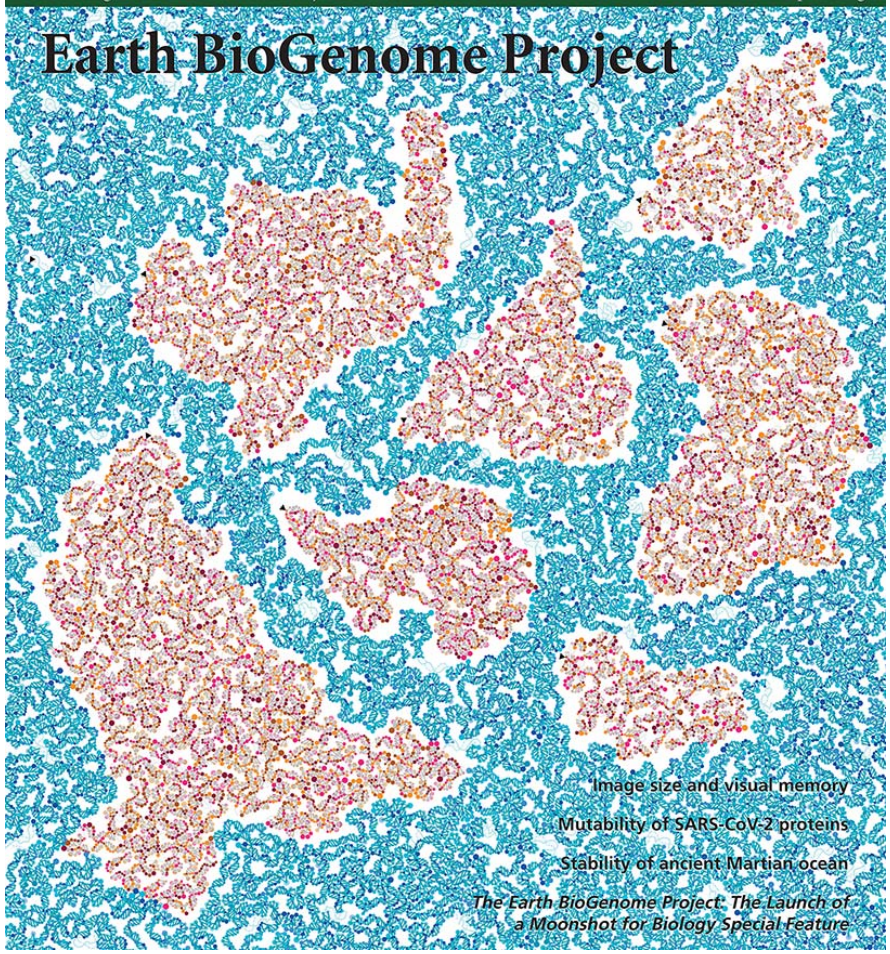


Image size and visual memory

Mutability of SARS-CoV-2 proteins

Stability of ancient Martian ocean

The Earth BioGenome Project: The Launch of a Moonshot for Biology Special Feature

W. John Kress
Co-Chair, Earth BioGenome Project Working Group
Smithsonian Institution

Gene E. Robinson
Co-Chair, Earth BioGenome Project Working Group
University of Illinois at Urbana-Champaign

EARTH BIOGENOME PROJECT
sequencing life for the future of life

Hank Greely
Chair, ELSI Committee

Melissa Goldstein
Co-Chair, ELSI Committee

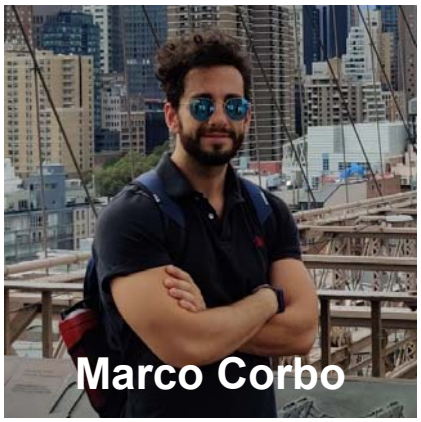
Federica Di Palma
Chair, International Scientific Committee

Sadye Paez & Marcela Uliano da Silva
Co-Chairs, JEDI Committee

Nicolette Caperello
Chair, Communications Committee

ACKNOWLEDGEMENTS

UCDAVIS GENOME CENTER



jmdamas@ucdavis.edu

Chromosome evolution working group
Jaebum Kim
Marta Farré
Denis M. Larkin
Jian Ma

Koala Genome
Rebecca N. Johnson

Narwhal Genome
Martin T. Nweeia

Sloth Genome
Camila Mazzoni
Marcela Uliano-Silva

Rock hyrax Genome
Olga Dudchenko
Erez Leiberman-Aiden

zoonomia

Elinor K. Karlsson
Kirsten Lindblad-Toh
Diane P. Genereux
Jason Turner-Maier
Jeremy Johnson
Bruce Birren
+ others

SAN DIEGO ZOO WILDLIFE ALLIANCE

Oliver A. Ryder
Cynthia Steiner
Marlys Houck

EARTH BIOGENOME PROJECT
sequencing life for the future of life

BAT 1K

VGP VERTEBRATE GENOMES PROJECT
DIGITAL NOAH'S ARK GENOME LIBRARY

PACBIO®

Dovetail GENOMICS


EARTH
BIOGENOME
PROJECT
Sequencing life for the future of life
www.earthbiogenome.org
Twitter: @EBPgenome