EARTH BIOGENOME PROJECT

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

<u>CTGGAAAACATTTTTTGCCAATTTATGCCAAAAAAACATTTCCG</u>	<u>TGATTATGTGGCGATTTACCTATGCAATGGGTATTTAGGAATCA</u>
ACATTTATTATTATGGACAAGAGCATAGCTTCGATATTGTGTT	ATCGGAAACAGAAATGTTGAAAATGTTAAAAATTTCATCCGATT
TGTTGCTAAGACCCAAACTGGATTGTGCCTATCACGTAGTTG	<u>ICAATGCGGTAACATCAACCGTTTCCTCTCCTTTAACGTTTAAAG</u>
CGGAAAATGCTAAGAAGGAAATTGAAGAAGTCTTGATAAAACC	attaaaaggaggaaaggaagcatgttcagtgattttatcaggat
GAAAAACGTTATTTTTACATATAATACAAGAGTGTTTGC	TAAAGCTCTTTGGTTGTCGAAGAGAGAGTTTGAAGATTTTCTCG
TAAATTATGACCAGAAAATGGCGTTGTTGGTGGACGATTTCGA	CAACATGAAACAGTACAATGAAA AATTATGCAATTTTATGGATA
TCTCATTCGTATTAGCAGTACGTAATGTCGAAGCCATTGATC1	GCCATTGCGAGGAAGATTTCCGTGTGAATTAGAATTACTGGTGC
AAAGAATGGAAATTCTACATTTGTTGCTAAAAAACGAAACAGT	a a a tta t c c t c a g c t c a g a a c t g a t a a a g a t a t t g c t c a g a b
ATACGGGTGGTGATTTGAAAGCAATTGTTCATCGTGTCTGTAC	CAATTTTGAATTTACCGGTGATAAGAGTGAATTATTTCAATGTT
TGAAGCGCATACGTCCAACGGGTATTCGGCAGTTTGTACTTCA	GGTACCTGATATCAGTTGGGAAGACA TTGGAGGAAATCGGGAAT
TTGAACAGGCAATTTTATGGCCATATCGATACCCGGAAATTTT	CAAGEGTTTEGETTEGAAGEETEEATETGGTATATTATTATATG
GCAGCAAAACTTTAATAGCACGAGCAATTGCATCACACTCAC	CATGAATTITTAGCAGTAAAAGGACCG GAATTGTTTAGTAAAT
CGGAAAGAGCCATACGAGAACTTTTCAGGCGACCTCCCCACG	AGCTCCGGCTATTATCTTTTCGATGAAATCGATGCAGTAGGTG
ATCAAAACGAGAGCCATGTTGGCGAACGCGTATTGACACAGTT	GCTTACCGAACTTGATGGTTTGGAGGAAAA AAGGTGATGTACTGG
CCAATCGACCCGATAGATTAGACAGTGCATTACTAAGGCCAG	CAGATITIAATCTCACCATICATGTCCCTT TGCCAGACGAAGAGA
TTTTGCGTATTAGATTGAATCAAATGAAAGTTAGCATTGATT	GGA TGTGGAAGA TA TGGGTA GA A GA A CCA A AGGATTTTCTGGTG
AACTTTGTAATCAGGCCGTACGTGAAGCGTTGCTCGAGAATAG	AAATGCTCATAAGCTTCAATMTCCACATMTCGATCAAGCATTGA
CGCGAACTCCAAATTGGCTGTTGAGTCCCAATTGAAATATA	A COORDENE AND THE STAR TO A TAKEN A CANTATTATTGTAC
CGATTGTTACCAGTCAGAGTTTA C VICELIUS C VIU	
CAATGGCTAAACTTCACATATTACCTCGTGCACCTMCTTCGGT	AACAAATATCGGTGCTAATGGTGACAACAATATTATTGGTGTGC
TTTGGCCATGTGATGAGGTCAAGGAAATACATATTTTCTTTC	AAAGGCTCGAATTAATGTGAAGTATCGGTTA AATGTTGTGAGTA
GAAAAGTTTCAATTATACAATTGCGTCCATCATCTGGAAAACA	IMMMMMMGCCAA AMMMA FGCCAAAAAAACA MMMCCGTGATTATGTGG
GCAATGGGTATTTAGGAATCAATGTACCGGTATACATMTATTA	TTA TGGACAAGAGCATAGCTTCGATATTGTG TTATCGGAAACAG
TGTTAAAAATTTCATCCGATTTATCATCAGCCATGTTGCTAAG	ACCCAAACTGGATTGCCCTATCACGTAGTTGACAATGCGGTAA
CCTCTCCTTTAACGTTT AAAGA GTTT GCTCGCACGGAAAATCGC	TAAGAAGGAAATTGAAGAAGTCTTGATAAAACCATTAAAAAGGAG
GTTCAGTGATTTTTATCAGGA TATACAGGGTGTGGA A A A A CGTT	a thtmtacatataa a tacaaga grettirgcaa aataaagctcttt
GAGAGTTTGAAGATTTTCCCGAGAATGGCGCTGTAAATTATGA	CCAGAAAATGGCC7MCG7MCGC7GGACGA7MMCCGACAACATGAAAC
AATTATGCAATTTTATCGATACATATACGAATTTCCTCATTCGT	ATTAGCAGTACGTAATCTCGAAGCCATTGATCTGCCATTGCGAG
GTGAATTAGAATTACTCGTGCCATCATTACCCCAAAGAATGGA	AATTCTACATTCTTGCTAAAAACGAAACAGTTAAAATTATCCT
TGATAAAAGATATTCCTCAGAAAAGTCACCGTTATACGGGTCC	TGA TTTGA AA GCAA TTGTTCA TCGTGTCTGTACCAATTTTGAAT
AGAGTGAATTATTTCAATGTTTCGATCTTCCACTGAAGCCCAT	CACGTCCAACGGGTATTCCGGCAGTTTCACTTCAGGTACCTGATA
ACATTGGAGGAAATCGGGGAATTGAAGATGAAAATTGAACAGGC	a a time ta regela ta reg atac ee ca a a tittea agegtite
CATCTGGTATATTATTATATGGACCACCCGGTTGCAGCAAAAC	TMMTAATAGCACGAGCAA TTGCAT CACAGTC ACGCATGAATTTTT
GACCGGAATTGTTTAGTAAATCGGGTTCGCGAATCGGAAAGAGC	CATACGAGAACTMMTCAGGCGAGCTCGGCAGGTAGCTCCGGCTA
ATGAAATCGATCCAGTAGGTCCAAATCGAGGAAATCAAAACGA	GA GCCATGTTGGCCGAACGCGTATTGACACAGTTGCTTACCGAAC
AGGAAAAAGGTGATGTACTGGTGCTAGCAGCTACCAATCGACC	CGA TAGA TTAGA CAGTGCATTACTA A GCCCA GGCAGATTTAATC
TCCCTTTGCCAGACGAAGAGACTAGGCTTGATATTMCGCGTAT	TAGATTGAATGAAAGTTAGCATTGATTTGGATGTGGAAG
GAACCAAAGGATTTTTCTGGTGCTGAAGTAGTCGAACTTTGTAA	TCAGCCGTACGTGAAGCGTTGCTCGAGAATAGAAATGCTGATA
GACATTTCGATCAAGCATTGAACGAGATTATTCCGCGAACTCC	A A TTGGCTGTTGAGTATTTATGAAGAATTTAAAAGAGGAGCCA
TCTAACTCCAAATTGGCTGTTCAGTATTTATGAAGAATTTAAA	A CACCACATTGGATGTGATCTAATGCCATTGCGAGGAAGA
CAAACTTTTTTACCACCTGACGTAATCGGTTGGGAAAAATGGAA	TACTONTTAATGCATCCAGAAACAATGGCTAAACTTCACATAT
CTTGTTGGTTAACAAATATCCGTGCTAATGGTCACAACAATAT	TATTCGTCTGCAATGCGTAACAGTTTGGCCATGTGATGAGGTCA
TTTTCTTTCAAA GGCTCCAATTAATGTGAAGTATCGGTTAAA	TETTETEAGTACCGAAAACATTAGAAAAGTTTCAATTATACAAT
CTGGAAAACATTTTTTGCCAATTTATGCCAAAAAACATTTCCCG	TGATA TGTGGCGATTTACCTATGCAATGGGTATTTAGGAATCA
ACATTTATTATTATCGACAACAGCATAGCTTCGATATTGTGTGT	A CCCAAACAGAAATGTTGAAAATGTTAAAAATTTCATCCGATT
TGTTGCTAAGACCCAAACTCGATTCTCCCCTATCACCTACTTCA	CAMTGCGGTAACATCAACCGTTTCCTCTCCTTTAACGTTTAAAG
CGGAAAATGCTAACAAGGAAATTGAAGAAGTCTTGAAAAAAC	MTTAAAAGGAGGAAAGGAAGCATGTTCAGTGATTTTATCAGGAT
GAAAAACGTTATTTTTACATATAATACAAGAGTGTTTTCCCAAAA	TAAAGCTCTTTGGTTGTCGAAGAGAGAGTTTGAAGATTTTCTCCC
TAAATTATGACCAGAAAATGGCGTTGTTGGTGCACCATTATCCA	CAACATGAAACAGTACAATGAAAAATTATGCAATTTTATGGATA
TCTCATTCGTATTAGCAGTACGTAATGTCGAAGCCATTCATCT	GCCATTGCGAGGAAGATTTCCGTGTGAATTAGAATTACTGGTGC
AAAGAATGGAAATTCTACATTTGTTGCTAAAAACGAAACAGTT	AAAATTATCCTCAGCTCAAGAACTGATAAAAGATATTGCTCAGA

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





The Earth BioGenome Project:

A Confederated International **Network**of-Networks

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

<u>Committed to diversity, equity,</u> <u>inclusion and justice</u> 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



Kerstin Lindblad-Toh Data Analysis



Richard Durbin Sequencing and Assembly



IT and Informatics





Paul Flicek Annotation





- 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)
- (Fliase I, Telefence quality,
- 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; CALeDNA)
- 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?



Eukaryotic Genome Sequencing: Current 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

er

https://goat.genomehubs.org/

wellcome

san

institute

July 12, 2022

н

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



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



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









- 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



Aye-aye

Geoffroy's spider monkey



Southern threebanded armadillo



Giant anteater



Brown-throated sloth

Snowshoe hare

Improving pandemic prediction & responsiveness



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, Kaus-Peter Koepfli,
 Andreas R. Pfenning, Huabin Zhao, Diane P. Genereux, Ross Swofford,
 Katherine S. Pollard, Oliver A. Ryder, Martin T. Nweeia, Kerstin Lindblad-Toh,
 Elinor K. Karlsson, Harris A. Lewin
 dol: https://doi.org/10.1101/2020.04.16.045302









Improving pandemic prediction & responsiveness



Genomic Study of SARS-CoV-2 Risk A Endangered Critically Endangered



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, Kathleen C. Keough, Corrie A. Painter,
 Andreas R. Pfenning, Huabin Zhao, Diane P. Genereux, Ross Swofford,
 Katherine S. Pollard, Oliver A. Ryder, Martin T. Nweeia, Karstin 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)

GENOME CENTER

(What was the karyotype and genome organization of the ance tral mammal?)

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

Jon Han for The New York Times



Science 309, 2005



Ancestral Chromosome Reconstruction (DESCHRAMBLER) derived from RACA method; Kim et al., 2013



(Kim et al. PNAS, 114: 2017)

EVOLUTION OF MAMMALIAN CHROMOSOMES



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





EVOLUTION OF MAMMALIAN CHROMOSOMES

- 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.

GENE ONTOLOGY ENRICHMENT FOR MSHSBS AND EBRS

msHSBs (<i>N</i> =522 > 1Mbp)			EBRs (<i>N</i> =323)				
Α	FDR p-value		GO terms	в	FDR p-value		GO terms
	FDR p-value 8.7e-21 5.8e-44 1.9e-19 6.9e-34 3.1e-21 2.0e-30 3.8e-27 6.1e-36 1.9e-29 5.2e-19 5.2e-20 5.2e-19 5.2e-19 5.2e-19 5.2e-19 5.2e-19 5.2e-19 5.2e-19 5.2e-19 5.2e-19 5.2e-19 5.2e-20	No. annot. genes ● 400 ● 800 ● 1,200 ● 1,600	GO terms Regulation of developmental process Anatomical structure morphogenesis Cellular developmental process Cell differentiation Animal organ development Tissue development Tissue development Neurogenesis Generation of neurons Neuron differentiation Nervous system development Cell development Positive regulation of macromolecule metabolic process Positive regulation of nitrogen compound metabolic process Embryo development Animal organ morphogenesis Pattern specification process Tube morphogenesis Tube development	В	FDR p-value 6.5e-23 1.9e-23 2.0e-25 3.5e-26 7.8e-28 1.1e-10 3.2e-06 8.9e-06 2.3e-04 2.4e-04 2.2e-04	No. annot. genes ● 100 ● 200	GO terms - Sensory perception of chemical stimulus - Detection of stimulus involved in sensory perception - Detection of chemical stimulus - Detection of chemical stimulus involved in sensory perception - Sensory perception of smell - Detection of stimulus - Sensory perception - Detection of stimulus - Sensory perception - Nervous system process - G proteincoupled receptor signaling pathway - Regulation of runceic acid-templated transcription - Regulation of RNA biosynthetic process - Nucleic acid-templated transcription - Transcription, DNA-templated - RNA biosynthetic process - Transcription by RNA polymerase II - Regulation of transcription by RNA polymerase II
	2.4e-23 0.35 0.40 No. annot. genes / Total n	0.45 no. annot. gene	Circulatory system development		0.25 0.25 No. annot. gener	0.50 0.75 1.0 6 / Total no. annot. gen	Response to jasmonic acid o es

- 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)

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)

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)









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



Hank Greely Chair, ELSI Committee



Melissa Goldstein Co-Chair, ELSI Committee



F**ederica 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



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



Oliver A. Ryder Cynthia Steiner Marlys Houck









Dovetail

GENOMICS

EARTH BIOGENOME PROJECT

Sequencing life for the future of life



www.earthbiogenome.org Twitter: @EBPgenome