PNAS
www.pnas.org

# Supplementary Information for

The Earth BioGenome Project 2020: Starting the Clock

Harris A. Lewin*[,1,2], Stephen Richards[3], Erez Lieberman Aiden[4], Miguel L. Allende[5,6], John M. Archibald[7], Miklós Bálint[8,9], Katharine B. Barker[10], Bridget Baumgartner[11], Katherine Belov[12], Giorgio Bertorelle[13], Mark L. Blaxter[14], Jing Cai[15], Nicolette D. Caperello[3], Keith Carlson[16], Juan Carlos Castilla-Rubio[17], Shu-Miaw Chaw[18], Lei Chen[15], Anna K. Childers[19], Jonathan A. Coddington[20], Dalia A. Conde[21,22], Montserrat Corominas[23,24], Keith A. Crandall[25,26], Andrew J. Crawford[27], Federica DiPalma[28], Richard Durbin[29,30], ThankGod E. Ebenezer[31], Scott V. Edwards[32,33], Olivier Federigo[34], Paul Flicek[35,30], Giulio Formenti[36], Richard A. Gibbs[37], M. Thomas P. Gilbert[38,39], Melissa M. Goldstein[40], Jennifer A.M. Graves[41,42], Henry T. Greely[43], Igor V. Grigoriev[44,45], Kevin J. Hackett[46], Neil Hall[47], David Haussler[48,49], Kristofer M. Helgen[50], Carolyn Hogg[12], Sachiko Isobe[51], Kjetill Sigurd Jakobsen[52], Axel Janke[9], Erich D. Jarvis[34,49], Warren E. Johnson[53,54], Steven J.M. Jones[55], Elinor K. Karlsson[56,57], Paul Kersey[58], Jin-Hyoung Kim[59], W. John Kress[60], Shigehiro Kuraku[61,62], Mara K.N. Lawniczak[14], James Leebens-Mack[63], Xueyan Li[64], Kerstin Lindblad-Toh[57,65], Xin Liu[66], Jose Lopez[67,68], Tomas Marques-Bonet[69,70,71,72], Sophie Mazard[73], Jonna A.K. Mazet[74], Camila J. Mazzoni[75,76], Gene Myers[77], Rachel J. O'Neill[78,79], Sadye Paez[34], Hyun Park[80], Gene E. Robinson[81], Cristina Roquet[82,83], Oliver A. Ryder[84,85], Jamal S.M. Sabir[86,87], H. Bradley Shaffer[88,89], Timothy M. Shank[90], Jacob S. Sherkow[91,81], Pamela S. Soltis[92,93], Boping Tang[94], Leho Tedersoo[95,96], Marcela Uliano-Silva[14], Kun Wang[15], Xiaofeng Wei[66], Regina Wetzer[97,98], Julia L. Wilson[30], Xun Xu[66], Huanming Yang[66], Anne D. Yoder[99,100] and Guojie Zhang[64,66,101,102].

Corresponding author: Harris A. Lewin
Email: lewin@ucdavis.edu

**This PDF file includes:**

**Supplementary Information Text**

# Brief Descriptions of 21/44 EBP-Affiliated Projects

## Table of Contents

**EBP Brief Affiliated Project Descriptions**

## *Genome project of crabs and related taxa to investigate their phylogeny and genetic basis of diversity*

*Project Contact:* Boping Tang boptang@163.com

*Scope and Goals:* The great morphological and ecological diversity of crabs have attracted the interest of many scientists (Gusmão et al., 2020; Tang et al., 2021). However, many relationships at higher taxonomic levels remain unresolved (Xin et al., 2017; Tang et al., 2020a and 2020b). To comprehensively promote our understanding in the evolution of crabs, we propose to sequence the genomes of crab species covering all (sub-)sections in the infraorder Brachyura and all superfamilies in Anomura using long-read sequencing technology. We will mainly focus on the two questions: 1) In Brachyura, re-examine the existing classifications of all (sub-)sections and infer their phylogenetic relationships. In Anomura, re-examine the accuracy of “hermit to king” and “king to hermit” hypotheses, and determine the occurrence times of carcinization events. 2) Study the genetic basis of the great morphological diversity of crabs, including the body size, leg number, and salinity adaptation.

*Progress:* We have successfully collected 11 crab species that cover all (sub-)sections in the infra-order Brachyura and two superfamilies in Anomura. Genome assembly and annotation have been done for most of them. The genome sizes range from 600 Mbp to 17 Gbp, and the scaffold N50 are comparable with the published high-quality genomes. In the follow-up work, we will continue to collect the remaining samples and employ comparative genomic analyses in depth. We hope this project will solve the controversies in crab taxonomy and phylogeny, and shed light on the genetic basis of morphological diversity (e.g. body form, body size, and leg number) and ecological adaptations (e.g. habitat) in crabs.

*References:*

1. Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. Trends in Genetics 34, 666-681.

2. Giribet, G., Edgecombe, G. D., and Wheeler, W. C. (2001). Arthropod phylogeny based on eight molecular loci and morphology. Nature 413, 157-161.

3. Gusmão, L. C., Van, D. V., Daly, M., and Rodríguez, E. (2020). Origin and evolution of the symbiosis between sea anemones (Cnidaria, Anthozoa, Actiniaria) and hermit crabs, with additional notes on anemone-gastropod associations. Molecular Phylogenetics and Evolution 148.

4. Luque, J., Feldmann, R. M., Vernygora, O., Schweitzer, C. E., Cameron, C. B., Kerr, K. A., Vega, F. J., Duque, A., Strange, M., Palmer, A. R. et al. (2019). Exceptional preservation of mid-Cretaceous marine arthropods and the evolution of novel forms via heterochrony. Science Advances 5.

5. Tang, B., Wang, Z., Liu, Q., Wang, Z., Ren, Y., Guo, H., Qi, T., Li, Y., Zhang, H., Jiang, S., et al. (2021). Chromosome-level genome assembly of Paralithodes

platypus provides insights into evolution and adaptation of king crabs. Molecular Ecology Resources 21, 511-525.

6. Tang, B., Wang, Z., Liu, Q., Zhang, H., Jiang, S., Li, X., Wang, Z., Sun, Y., Sha, Z., and Jiang, H., et al. (2020).High-quality genome assembly of Eriocheir japonica sinensis reveals its unique genome evolution. Frontiers in Genetics 10.

7. Tang, B., Zhang, D., Li, H., Jiang, S., Zhang, H., Xuan, F., Ge, B., Wang, Z., Liu, Y., Sha, Z., et al. (2020).Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (Portunus trituberculatus).Gigascience 9, 1-10.

8. Xin, Z., Liu, Y., Zhang, D., Wang, Z., Zhang, H., Tang, B., Zhou, C., Chai, X., and Liu, Q. (2017). Mitochondrial genome of Helice tientsinensis (Brachyura: Grapsoidea: Varunidae): Gene rearrangements and higher-level phylogeny of the Brachyura. Gene 627, 307-314.

## *Deep-Ocean Genomes Project: accelerating discovery of deep-sea adaptations and biodiversity*

*Project Contact:* Timothy M. Shank tshank@whoi.edu

*Scope and Goals:* Earth's ocean is the largest and most biodiverse ecosystem on Earth, hosting at least 33 known phyla from the tree of life, with ~410,000 named species. In partnership, Woods Hole Oceanographic Institution and the University of Connecticut have established the Deep-Ocean Genomes Project (DOG) to implement genomics technologies and address diverse ecological and evolutionary hypotheses within and across a myriad of species found in deep sea habitats (hydrothermal vents, methane seeps, oxygen minimum zones, seamounts, canyons and trenches). DOG uses genome sequencing and comparative genomics approaches to study species adaptations to extreme environments, including immense pressures (> 15,000 psi), near-freezing to near-boiling temperatures, absence of sunlight, toxic chemical conditions, and diverse energy sources. Our EBP-standard genome sequencing and assembly workflow is based on long and short-read sequencing applications that support chromosome-level assemblies, epigenomic profiling, and exploration of gene, regulatory, and noncoding content.

*Progress:* A cryo-repository containing > 500 species from > 17 phyla, spanning > 30 habitat types below 200 meters depth has been developed from > 80 deep-ocean expeditions. To date, 20 species of 100 projected for phase 1 are currently in our EBP standard genome sequencing pipeline including annelids from vents, fish from hadal regions and cnidarians from canyons and seamounts. Moving forward, phases 2 and 3 will support planned exploration and expanded genome sequencing, leading to phase 4 consisting of technology incubation programming for discovery-based applications and assessment of exploration beyond Earth. Additionally, questions derived from genome data will drive expeditions for targeted sampling and development of autonomous *in-situ* characterization of deep-sea ecosystems and genomes.

*References:*

1. Shank, T.M. (2010) Seamounts: deep-ocean laboratories of faunal connectivity, evolution, and endemism. *Oceanography* 23:108–122.

2. Sigwart, J.D., R. Blasiak, M. Jaspars, J-B. Jouffray, and Tasdemir, D. (2020) Unlocking the potential of marine biodiscovery. *Royal Society of Chemistry*. DOI: 10.1039/d0np00067a.

3. Thurber, A.R., A.K. Sweetman, B.E. Narayanaswamy , D.O.B. Jones , J.Ingels, and Hansman, R.L. (2014) Ecosystem function and services provided by the deep sea. Biogeosciences, 11, 3941–3963. DOI: 10.5194/bg-11-3941-2014.

4. Cui, J., Z, Yu, M. Mi, L. He, Z. Sha, P. Yao, J. Fang, and Sun, W. (2020) Occurrence of Halogenated Organic Pollutants in Hadal Trenches of the Western Pacific Ocean. *Environmental Science & Technology 54* (24), 15821-15828. DOI: 10.1021/acs.est.0c04995.

5. Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P. et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proceedings of

the National Academy of Sciences 115 (17) 4325-4333; DOI: 10.1073/pnas.1720115115

## *Diversity Initiative for Southern California Oceans (DISCO)*

*Project Contact:* Regina Wetzer rwetzer@nhm.org

*Scope and Goals:* To achieve its goal of sequencing genomes of eukaryotic biodiversity, the Earth BioGenome project requires specimens of all those species. The *Diversity Initiative for the Southern California Ocean* (DISCO) is helping to fill the need for specimens across the phylogenetically diverse and numerous species of the marine invertebrate community. DISCO is a cross-departmental working group at the Natural History Museum of Los Angeles County (NHMLA), which already houses one of the largest collections of marine invertebrates in North America. DISCO was initiated to explore the use of genetic techniques, specifically community metabarcoding, to discover and document marine invertebrate biodiversity.

*Progress:* DISCO has pursued a multi-year effort to collect, identify, genetically barcode, and put into storage 12,000 voucher specimens, representing 5,200 localities to build a comprehensive genetic barcode reference library for Southern California marine invertebrates. The program has involved dozens of agencies and academic organizations to arrange for access and acquisition, has negotiated unusual permitting arrangements to cover the unconventional taxonomic breadth of sampling, and has integrated the work of numerous taxonomists and collections staff across the US. The project is resulting in genetic barcodes for thousands of species, each accompanied by vouchered specimens available for future genetic work. The specimens collected by the DISCO program represent over 700 families of marine invertebrates, all of which are available for sequencing by EBP. Several of the current EBP target taxa being sequenced are based on specimens from the DISCO program. DISCO's work prior to the founding of the Earth BioGenome program fortuitously prepositioned it as a key resource available to the Earth BioGenome project, analogous to contributions DISCO has made to other genetic biodiversity projects [1, 2].

*References:*

1. McElroy, M.E., Dressler, T.L., Titcomb, G.C., Wilson, E.A., Deiner, K., Dudley, T.L., Eliason, E.J., Evans, N.T., Gaines, S.D., Lafferty, K.D., et al. (2020). Calibrating environmental DNA metabarcoding to conventional surveys for measuring fish species richness. Front. Ecol. Evol. *8*, 276. 10.3389/fevo.2020.00276.

2. McGee, K.M., Robinson, C.V., and Hajibabaei, M. (2019). Gaps in DNA-based biomonitoring across the globe. Front. Ecol. Evol. *7*, 337. 10.3389/fevo.2019.00337.

3. Meyer, R.S., Munguia Ramos, M., Curd, E.E., Schweizer, T.M., Gold, Z., Ruiz Ramos, D., Shirazi, S., Kandlikar, G., Kwan, W.-Y., Lin, M., et al. (2021). The CALeDNA program: Citizen scientists and researchers inventory California's biodiversity. Calif. Agric. *75*, 20–32. 10.3733/ca.2021a0001.

4. Gold, Z., Wall, A.R., Curd, E.E., Kelly, R.P., Pentcheff, N.D., Ripma, L., Barber, P.H., and Wetzer, R. (2020). eDNA metabarcoding bioassessment of endangered fairy shrimp (*Branchinecta* spp.). Conserv. Genet. Resour. *12*, 685–690. 10.1007/s12686-020-01161-9.

5. Ellwood, E.R., Kimberly, P., Guralnick, R., Flemons, P., Love, K., Ellis, S., Allen, J.M., Best, J.H., Carter, R., Chagnoux, S., et al. (2018). Worldwide engagement for digitizing biocollections (WeDigBio): The biocollections community's citizen-science space on the calendar. BioScience *68*, 112–124. 10.1093/biosci/bix143.

## *Butterfly genome project: explore the evolution of butterfly diversity*

*Project Contact:* Xueyan Li lixy@mail.kiz.ac.cn

*Scope and Goals:* Due to their extraordinarily diverse wing patterns and other diverse biological traits, butterflies have been of great interest to naturalists for centuries [3], and the study of butterflies has been an integral part of ecology and evolution ever since Darwin's times [4]. To date, the reference genomes of 63 butterfly species (six families, 14 subfamilies) are deposited in GenBank. Compared with about 18,000 described species (Papilionoidea, seven families, 35 subfamilies), however, genomic resources of butterflies are still lacking, and are unbalanced from a phylogenetic view. To investigate the genomic basis of butterfly morphological diversity in a comprehensively phylogenetic context, we proposed and started the butterfly genome project in August of 2017 to produce high-quality whole genomes of the representative species covering all subfamilies of all families using third generation sequencing technologies. Sequenced species were mainly selected based on their morphological traits, sample availability, host utilization, and life span etc., and thus can be feasibly used as models to further investigate more scientific questions by us and the whole butterfly research community.

*Progress:* We have finished assembling the genomes of 63 butterfly species in 29 subfamilies of six families (Papilionidae: 19; Hesperiidae: 9; Pieridae: 6; Nymphalidae: 21; Lycaenidae: 6; Roidinidae: 1), whose sizes range from 222 Mb to 1176 Mb. Among them, 16 subfamilies are assembled for the first time, and 26 assemblies are chromosome-level. Combining those butterfly reference genomes previously reported, more than 120 species in 30 subfamilies can be integrated into comparative genomics studies. We anticipate this project will clarify the genomic basis of morphological diversity evolution of butterflies in a comprehensively phylogenetic context, and provide data resources for a wide range of fields, including phylogeny, comparative genomics, evolutionary developmental biology, the relationship of insect and plant, and conservation genetics.

*References:*

1. Boogs, C.L., Ehrlich, P.R., and Watt, W.B. (2003). Butterflies: Ecology and Evolution Taking Flight (Chicago: University of Chicago Press).

2. Wallace, A. R. (1870). Contributions to the theory of natural selection: a series of essays (New York: Cambridge University Press).

3. Carroll, S. B. (2005). Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom (New York: W. W. Norton & Company).

4. Warren, M. S., Hill, J. K., Thomas, J. A., Asher, J., Fox, R., Huntley, B., Roy, D. B., Telfer, M. G., Jeffcoate, S., Harding, P., et al. (2001). Rapid responses of British butterflies to opposing forces of climate and habitat change. Nature 414(6859), 65-69.

5. Li, X., Fan, D., Zhang, W., Liu, G., Zhang, L., Zhao, L., Fang, X., Chen, L., Dong, Y., Chen,Y., et al. (2015). Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. Nat Commun 6, 8212.

## *Lilioid genomes illuminate monocot evolution*

*Project Contact:* Jing Cai, jingcai@nwpu.edu.cn

*Scope and Goals:* Monocots are one of the most important plant lineages, and contributes the majority of the agricultural biomass on earth. Species of core lilioid monocots have evolved spectacular flower appearance in adaptation to animal pollination and underground storage organs (e.g. geophytes). They are of high economic value with many species used as flavourings (such as the *Allium* species e.g. onion, garlic, leek, chives, etc.) and medicines in traditional medicine or cosmetics (e.g. *Veratrum*, *Paris polyphylla* and *Aloe*). The group also includes several important floral industry species used as cut flowers (e.g. gladiolus, orchids) or garden ornamentals (e.g. tulips, iris, lily-of-the-valley, hyacinths). In contrast to their high species diversity, the under-represented sampling of monocots, in particular the core lilioid monocots, leaves an enormous gap in our understanding of monocot evolution and function. Our goal is to sequence the genomes of representative species in each family of Liliales (ten families) and Asparagales (14 families) to the chromosomal level. This project complements the 10,000 Plant (10KP) sequencing project by the major gap in monocots.

*Progress:* We have completed the sequencing of two species in the Melanthiaceae family of Liliales and four species in the Amaryllidaceae family of Asparagales. Genome analysis was focused on genome size expansion, flower morphology evolution and biosynthesis of secondary metabolic compounds. These six genomes will be released in 2021 as the major result for the first phase of the lilioid genome project. Twenty more genomes will be sequenced over the next three years (2022-2024) in the second phase. Based on the genome sequences, we will reconstruct the phylogeny of the lilioid monocots core group and are investigating the genetic novelties contributing to the giant diversity of monocots. Follow-up genomic analysis on unique traits specific to each family of this group will greatly extend our understanding of the monocot evolution and provide novel insights into crop improvement.

*References:*

1. https://en.wikipedia.org/wiki/Monocotyledon
2. Christenhusz, M.J.M., and Byng, J.W. (2016). The number of known plants species in the world and its annual increase. Phytotaxa *261*, 201.
3. The Angiosperm Phylogeny Group (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. *181*, 1–20.
4. Yang, L., Su, D., Chang, X., Foster, C.S.P., Sun, L., Huang, C.-H., Zhou, X., Zeng, L., Ma, H., and Zhong, B. (2020). Phylogenomic Insights into Deep Phylogeny of Angiosperms Based on Broad Nuclear Gene Sampling. Plant Communications *1*, 100027.

## *EndemixIT: whole genome sequencing to study and protect Italian endangered endemics*

*Project Contact:* Giorgio Bertorelle, ggb@unife.it

*Scope and Goals:* Many endemic species are under extinction pressure in countries around the world but the use of genomics to guide protection of these species is rare. EndemixIT is a project funded in 2020 by the Italian Ministry of Education, University and Research (PRIN calls) to address this issue. EndemixIT goals are to a) obtain high-quality, highly contiguous, and annotated reference genomes for five Italian endangered endemics: the Adriatic sturgeon, *Acipenser naccarii*; the Apennine yellow-bellied toad, *Bombina pachypus*; the Marsican bear, *Ursus arctos marsicanus*; the Ponza grayling, *Hipparchia sbordonii*; the Aeolian lizard, *Podarcis raffonei*; b) produce whole-genome sequences for 20 to 30 individuals per species; c) use genomic variation data to reconstruct demographic trajectories and estimate the accumulation of deleterious mutations which may increase the susceptibility to extinction [5]; d) compare genomic patterns between small and large populations to test the relationship between genetic load and population size; e) perform computer simulations to predict the effects of genetic rescue plans [6]; f) perform functional assays to verify the deleterious effects predicted by bioinformatics approaches; g) promote the use of genomics in conservation biology [7].

*Progress:* As of February 2021, sampling is completed. The reference genome for the Marsican bear (2.3 GB) is assembled with a contig N50 of 71.5 MB and awaiting Omni-C scaffolding to reach a chromosomal level. The reference genome of the butterfly (0.5 GB) is almost done. The assembly of the toad genome (10GB) requires preliminary chromosome separations, and cell cultures have been established. The lizard and sturgeon genomes will be assembled in cooperation with the Vertebrate Genome Project at Rockefeller University (samples being processed). Libraries for the resequencing are ready for all species but the toad. Transfections for the functional study on the Marsican brown bear fibroblasts are in progress. Controlled crosses for the functional study on the sturgeon are concluded, and the progenies are under phenotypic and genotypic investigation. EndemixIT will produce new reference genomes, increase our understanding of the dynamic of genetic load accumulation, and significantly reduce the gap between theory and practice in conservation genomics. Our aim is to implement an informed conservation genomics approach, with these and other key, iconic, endangered and endemic Italian species, and aid other conservation groups around the world.

*References:*

1. Benazzo, A., Trucchi, E., Cahill, J.A., Maisano Delser, P., Mona, S., Fumagalli, M., Bunnefeld, L., Cornetti, L., Ghirotto, S., Girardi, M., et al. (2017) Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. Proc. Natl. Acad. Sci. USA. 114: E9589-E9597. doi.org/10.1073/pnas.1707279114.

2. van Oosterhout, C. (2020) Mutation load is the spectre of species conservation. Nat. Ecol. Evol. 4, 1004–1006. doi.org/10.1038/s41559-020-1204-8,

3. Harris, K., Zhang, Y.,and Nielsen, R. (2019). Genetic rescue and the maintenance of native ancestry. Cons. Genet. 20_59-64. doi.org/10.1007/s10592-018-1132-1.

4. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. 2010. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol. 6: e1001025 doi.org/10.1371/journal.pcbi.1001025

5. Hohenlohe, P.A., Funk, W.C. and Rajora, O.P. (2020). Population genomics fro wildlife conservation and management. Mol.Ecol. 30:62-82. doi.org/10.1111/mec.15720

## *Central-European Soil Invertebrate Genome Initiative (SIGI)*

*Project Contact:* Miklós Bálint ggb@unife.it

*Scope and Goals:* Soils harbor a tremendous diversity of organisms. These play major roles in ecosystem services central for human well-being, e.g. food production and climate regulation. However, much remains to be discovered about soil invertebrates. Organisms are often small, hard to observe, and challenging to study. We initiated SIGI to leverage emerging genomic tools to understand the ecology and evolution of soil invertebrates, their role in ecosystems and soil functions, to develop sustainable management procedures, and to provide resources for bioeconomy. We target all soil invertebrate groups: mites, springtails, enchytraeids, nematodes, myriapods, earthworms and others. Sampling considers two aspects: comprehensive representation of soil invertebrate groups, and significance to address global change impacts on soils. We prioritize samples authorized for molecular work: currently Central-European taxa. The involvement of organism experts ensures that samples are properly collected and identified. We employ non-destructive DNA extraction whenever possible; vouchers of sequenced specimens are deposited into museum collections.

*Progress:* Technical progress: We have successfully generated highly contiguous genomes of single mm-sized species, with contig N50s up to 8.7 Mbase [8] (Schneider et al., in press). Currently we are experimenting with low-input Hi-C scaffolding [9] and plan to start RNA sequencing soon. Genome progress: Genomes of ~250 species were so far generated using short read technologies within SIGI. Current sequencing has shifted efforts towards highly contiguous, annotated genomes. Ongoing analyses aim to link functional traits to genomic features, to improve the understanding of phylogenetic relationships, and to explore species identification possibilities from mixed/environmental samples. We plan to complement genome sequencing with automated imaging and machine learning, to improve specimen identifications and descriptions. Results will contribute to understanding the functional genomics of soil invertebrate adaptations. Analyzed genomes will be made public and are provided as a community resource to facilitate evolutionary and ecological work on soil invertebrates.

*References:*

1. Schneider, C., Woehle, C., Greve, C., D'Haese, C., Wolf, M., Hiller, M., Janke, A., Bálint, M., and Hüttel, B. (2021). Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola) In Press. GigaScience.

2. Díaz, N., Kruse, K., Erdmann, T., Staiger, A.M., Ott, G., Lenz, G., and Vaquerizas, J.M. (2018). Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. Nature communications *9*, 1-13.

3. FAO, I. (2020). State of knowledge of soil biodiversity - Status, challenges and potentialities: Report 2020 (Rome, Italy: FAO).

## *The Ungulate Genome Project*

*Project Contact:* Lei Chen chen_lei@nwpu.edu.cn

*Scope and Goals:* The ungulates comprise the mammalian orders Perissodactyla and Cetartiodactyla, spanning 13 extant families (excluding the Cetaceans). Ungulates make up the majority of large terrestrial mammals, which act as major drivers forming the shape and function of the terrestrial ecosystem (Baskin and Danell, 2003). Notably, the ungulates contain the majority of domesticated mammal livestock species, including the horse, donkey, pig, camel, cattle, goat and several others, thus have assumed important roles in most human cultures worldwide by contributing meat, milk, leather and draft. According to the IUCN, more than two thirds of the ~350 extant ungulates species have suffered serious population declines (IUCN, 2020), motivating the urgency of in-depth research and conservation studies for these animals. We initiated the Ungulate Genome Project to generate high quality chromosome-level genome sequence for 20 representative species covering all families and major subfamilies (combining with previously released genomes). The genomes will be assembled using the combined platforms of Nanopore, PacBio and Hi-C, to achieve or exceed the latest Earth BioGenome Project assembly standards.

*Progress:* We have completed the sequencing of 12 species, and reference genome completion is expected in 2022. The remaining species would be sequenced and released in 2023. We will dissect the phylogenetic tree of the ungulates, unveil the demographic history, uncover the genetic mechanism of unique traits (e.g. hoof origination, food habit change, limb development and hair loss) and special environment adaptations. We anticipate that the upcoming findings will shed new light on many aspects of ungulate biology and terrestrial mammalian evolution, and furthermore provide data for more informed conservation efforts involving the ungulates.

*References:*

1. Baskin, L., and Danell, K. (2003). Ecology of Ungulates: A Handbook of Species in Eastern Europe and Northern and Central Asia (Berlin Heidelberg: Springer-Verlag).

2. IUCN, The IUCN Red List of Threatened Species, Version 2020-3. (2020); www.iucn.org.

3. Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., et al. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science 364, eaav6202.

4. Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. PNAS 115, 4325–4333.

5. Meredith, R.W., Janecka, J.E., Gatesy, J., Ryder, O.A., Fisher, C.A., Teeling, E.C., Goodbla, A., Eizirik, E., Simao, T.L.L., Stadler, T., et al. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. Science 334, 521–524.

## *The European Reference Genome Atlas*

*Project Contact:* Camila J. Mazzoni (mazzoni@izw-berlin.de)

*Scope and Goals:* Addressing the current sixth mass extinction [10] will require significant scientific and conservation investment to understand, manage, conserve and restore biodiversity and ecosystems. With approximately one fifth of the ~200,000 European species at risk of extinction[11]. In response to the extinction risk and the urgent need to deepen our knowledge about existing biodiversity, a group of over 400 scientists from 33 European countries - including all 27 European Union members - has formed a large genomics initiative, the European Reference Genome Atlas (ERGA) (www.erga-biodiversity.eu). The goal of ERGA is to generate reference genomes representing all European biodiversity, from endangered species to species of importance for agriculture, fisheries, pest control, ecosystem functioning and stability. ERGA is the European hub of the Earth Biogenome Project and aligns to its goals, guidelines and principles.

*Progress:* As its first major action, ERGA's pilot project is generating high-quality reference genomes for at least 33 species, one per associated country, representing five biodiversity categories: endangered, iconic, marine, freshwater and pollinators. ERGA has made significant organizational progress as of March 2021, the ERGA consortium has a defined structure that includes a Council of country representatives, an executive board and committees that work on guidelines for scientific and beyond-science activities, such as knowledge transfer, citizen science and ethical, legal, inclusiveness, and social issues. ERGA established solid principles to govern its development, using state-of-the-art science in a democratic and socially fair manner. The six pillars of ERGA are: 1) scientific excellence in genome reconstruction and analysis, 2) distributed infrastructure and scientific expertise across Europe, 3) increase of taxonomic, geographical and habitat representation of genomes in a balanced manner, 4) inclusion of scientists in all socially diverse aspects, 5) data generation and release under the FAIR guidelines[12], 6) prioritization of species that require urgent protection. ERGA will set solid, accurate, scalable and comparable foundations for hundreds of scientific studies. Ultimately, these resources will enable a wide range of scientific analyses that will benefit ecosystems, society and humankind.

*References:*

1. Ceballos G, Ehrlich PR, Raven PH. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. Proc Natl Acad Sci U S A. 2020 Jun 16;117(24):13596–602.
2. IUCN. European Red List of Threatened Species [Internet]. 2020. Available from: https://www.iucn.org/regions/europe/our-work/species/european-red-list-threatened-species
3. European Commission. Communication on the European Green Deal, COM(2019)640. 2019 Dec;
4. European Commission. Communication on the Sustainable Europe Investment Plan – European Green Deal Investment Plan, COM(2020) 21 final. 2020 Jan;
5. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A. 2018 Apr 24;115(17):4325–33.

6. McMahon BJ, Teeling EC, Höglund J. How and why should we implement genomics into conservation? Evol Appl. 2014 Nov;7(9):999–1007.

7. Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations [Internet]. Vol. 26, Molecular Ecology. 2017. p. 5369–406. Available from: http://dx.doi.org/10.1111/mec.14264

8. Breed MF, Harrison PA, Blyth C, Byrne M, Gaget V, Gellie NJC, et al. The potential of genomics for restoring ecosystems and biodiversity. Nat Rev Genet. 2019 Oct;20(10):615–28.

9. Hoffmann A, Griffin P, Dillon S, Catullo R, Rane R, Byrne M, et al. A framework for incorporating evolutionary genomics into biodiversity conservation and management. Climate Change Responses. 2015 Jan 28;2(1):1.

10. Razgour O, Forester B, Taggart JB, Bekaert M, Juste J, Ibáñez C, et al. Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. Proc Natl Acad Sci U S A. 2019 May 21;116(21):10418–23.

11. Brandies P, Peel E, Hogg CJ, Belov K. The Value of Reference Genomes in the Conservation of Threatened Species. Genes [Internet]. 2019 Oct 25;10(11). Available from: http://dx.doi.org/10.3390/genes10110846

12. Hohenlohe PA, Funk WC, Rajora OP. Population genomics for wildlife conservation and management. Mol Ecol [Internet]. 2020 Nov 3; Available from: http://dx.doi.org/10.1111/mec.15720

13. Consortium Z, Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation [Internet]. Vol. 587, Nature. 2020. p. 240–5. Available from: http://dx.doi.org/10.1038/s41586-020-2876-6

14. Waldvogel A, Feldmeyer B, Rolshausen G, Exposito-Alonso M, Rellstab C, Kofler R, et al. Evolutionary genomics can improve prediction of species' responses to climate change [Internet]. Vol. 4, Evolution Letters. 2020. p. 4–18. Available from: http://dx.doi.org/10.1002/evl3.154

15. Capblancq T, Fitzpatrick MC, Bay RA, Exposito-Alonso M, Keller SR. Genomic Prediction of (Mal)Adaptation Across Current and Future Climatic Landscapes. Annu Rev Ecol Evol Syst. 2020 Nov 2;51(1):245–69.

16. Stange M, Barrett RDH, Hendry AP. The importance of genomic variation for biodiversity, ecosystems and people. Nat Rev Genet [Internet]. 2020 Oct 16; Available from: http://dx.doi.org/10.1038/s41576-020-00288-7

17. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018.

## *The Catalan initiative for the Earth Biogenome Project*

*Project Contact*: Montserrat Corominas, mcorominas@ub.edu

*Scope and Goals*: The Catalan Initiative for the Earth Biogenome Project (CBP, www.biogenoma.cat) is an EBP affiliated project with the aim of sequencing the genome of the more than 40,000 eukaryotic species estimated to live in the Catalan territories. These territories, spanning across Spain, France and the Principality of Andorra, have historically shared a strong cultural tradition, reflected most notably in the use of the Catalan language. They lay at the intersection of European and African plates and at the crossroad between the Euro-Siberian and the Mediterranean biogeographical regions. These territories represent a biodiversity hotspot: while covering less than 1% of Europe, they are home to about one fourth of all known European eukaryotic species. They are characterized by a high level of endemism. Many endemic species are threatened, a trend that will be aggravated in the future, as climate change will impact especially the Mediterranean Basin and mountain areas. Following the EBP model, the CBP is a networked organization under the umbrella of the Institute for Catalan Studies (IEC), the academy tasked with the promotion of science and culture in the Catalan territories, and of the Andorra Research and Innovation Foundation. IEC has provided seed funding to initiate the CBP activities and leadership has been delegated to the Catalan Society of Biology (SCB) and the Catalan Institution of Natural History (ICHN).

*Progress*: A white paper was released in October 2018. A number of organizational meetings with local stakeholders took place, and additional documents were elaborated before an international meeting on Genomics for Biodiversity was organized in September 2019. Since then, the ICHN is working on an updated digitized catalogue of the eukaryotic species living in the Catalan territories. As of August 2021, this catalogue includes 21,457 species, and will serve as a reference to prioritize the genomes to be sequenced. During the pilot phase, working groups have been established, mirroring partially those at the EBP. The CBP has also launched two calls for projects, in 2020 and 2021. Criteria for prioritization include phylogenetic position and novelty, interest to local research groups, degree of endemism and conservation, biomedical, agricultural and industrial interest. Building on this and on additional funding, high-quality reference genomes of 40 species are currently being generated under the CBP umbrella. The CBP aims to play a central role in the biodiversity genomics projects in Europe, which are now being organized under the European Reference Genome Atlas (ERGA). Locally, the CBP represents an excellent opportunity to bring together research communities that have been traditionally isolated from each other. It will enhance the natural history institutions with the state-of-the-art infrastructures and human resources required to guarantee cataloguing, preservation of specimens, tissues and DNA and curation of vouchers for future generations. Moreover, by building on the perception of a strongly shared cultural background, the CBP aims to engage the society as a whole. Beyond a pure scientific endeavor, we see the CBP as part of a worldwide transformative movement that raises social awareness about the threat posed by biodiversity loss on human well-being, and that globally engages the society into a different, more balanced, relationship with nature.

*References:*

1. Camarasa J.M., and Casassas O., "Cent anys de la Societat Catalana de Biologia, la primera societat filial de l'Institut d'Estudis Catalans", Institut d'Estudis Catalans (2020), ISBN: 978-84-9965-556-7

2. Casas-Sainz, A.M., and de Vicente, G. (2009). On the tectonic origin of iberian topography. Technophysics *474*, 214–235.

3. Cramer, W., Guiot, J., Fader, M., Garrabou, J., Gattuso, J.-P., Iglesias, A., Lange, M.A., Lionello, P., Llasat, M.C., Paz, S., et al. (2018). Climate change and interconnected risks to sustainable development in the Mediterranean. Nat. Clim. Chang. *8*, 972–980.

4. Crisci, J.V., Katinas, L., Apodaca, M.J., and Hoch, P.C. (2020). The end of botany. Trends Plant Sci. *25*, 1173–1176.

5. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319.

6. Folch i Guillèn R. Director (1984-2012). Història natural dels Països Catalans, Enciclopèdia Catalana, Barcelona.

7. Hoegh-Guldberg, O., Jacob, D., Taylor, M., Guillén Bolaños, T., Bindi, M., Brown, S., Camilloni, I.A., Diedhiou, A., Djalante, R., Ebi, K., et al. (2019). The human imperative of stabilizing global climate change at 1.5°C. Science *365*.

8. Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci USA *115*, 4325–4333.

9. Pepin, N., Bradley, R.S., Diaz, H.F., Baraer, M., Caceres, E.B., Forsythe, N., Fowler, H., Greenwood, G., Hashmi, M.Z., Liu, X.D., et al. (2015). Elevation-dependent warming in mountain regions of the world. Nat. Clim. Chang. *5*, 424–430.

10. Societat Catalana de Biologia (2018). The Catalan Biogenome Project. (https://drive.google.com/file/d/1hjU5TheoEd2yC6K9euUvr4DtyLOiq0vM/view?usp=sharing)

11. The Darwin Tree of Life Project Consortium (2021). Sequence locally, think globally: The Darwin Tree of Life Project. Cell Genomics

12. Tuel, A., and Eltahir, E.A.B. (2020). Why is the Mediterranean a climate change hot spot? J. Clim. *33*, 5829–5843.

## *The California Conservation Genomics Project*

*Project Contact:* Brad Shaffer (brad.shaffer@ucla.edu), Erin Toffelmier (etoff@ucla.edu)

*Scope and Goals*: California is home to an extraordinary number of species, many of which are under threat of extinction (Calsbeek et al., 2003; Mittermeier et al, 2004; Loarie et al, 2008). The state is similarly diverse ecologically, with 19 terrestrial and three marine ecoregions that range from deserts to temperate rainforests to alpine habitats (Goudey and Smith, 1994). California's landscapes are inherently dynamic, with climate, fire, and human population pressure leading to an ever-shrinking mosaic of healthy ecosystems, embedded in a patchwork of urban and working lands. Consistent with the California Biodiversity Initiative (2018), the California Conservation Genomics Project (CCGP; https://www.ccgproject.org) was funded and launched in 2019 to assemble a multi-species resource of genomic variation for ~250 species of marine and terrestrial plants, vertebrates and invertebrates to enable proactive biodiversity management. For each of the 147 genus-level projects that have been funded, the CCGP is producing one high-quality reference genome, and providing support to University of California (UC) investigators and their collaborators to fully resequence the genomes of 100-150 individuals from geographically dispersed localities across each species' California range. The resulting data will be analyzed with a common set of CCGP-developed bioinformatic pipelines to identify genetic hotspots, natural barriers to gene flow, and regions with genetically depauperate populations. This landscape genomic information is essential for ongoing studies of the impact of climate and landscape change on biodiversity stewardship.

*Progress:* Investigator interest in the CCGP has been extremely high. We are now working with 246 species that fall into 147 genera, and 106 principal investigators representing all 10 UC campuses. The CCGP reference genome pipeline is now in full production phase. Flash frozen reference genome tissues for HiFi, Omni-C and RNA sequencing have been submitted for 101 species; many more have been submitted for at least one sequencing component. Of those, we have generated HiFi data for 38, completed draft HiFi assemblies for 28, and we are approaching our target goal of completing 2-3 chromosome-level reference genomes per week. Our landscape sampling now includes reasonable coverage of all California ecoregions and major taxonomic groups. Approximately 67% of ~18,000 proposed resequencing samples are in the hands of individual investigators, and sampling should be complete by the end of 2021 or very early 2022. All data collection is targeted for completion by June 2022.

Our next milestone is developing CCGP bioinformatic pipelines to ensure that genomic coverage, filtering, base calling, and assembly are standardized across species for maximal data compatibility. In parallel, a set of standard landscape genomic analyses and mapping tools, applicable to both terrestrial and marine species, are being developed and implemented. As these resources are refined, the CCGP will move into an outreach phase, working with state, federal and non-governmental stakeholders to implement strategies that best utilize our landscape genomic output to complement and further modernize conservation efforts and species protection across California.

The CCGP is supported with funding provided to the University of California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224].

*References:*

1. California Biodiversity Initiative (2018). Available at https://www.californiabiodiversityinitiative.org/pdf/california-biodiversity-action-plan.pdf

2. Calsbeek, R., Thompson, J. N., & Richardson, J. E. (2003). Patterns of molecular evolution and diversification in a biodiversity hotspot: the California Floristic Province. *Molecular Ecology, 12*:1021-1029. doi:10.1046/j.1365-294X.2003.01794.x

3. Goudey, C.B., and D.W. Smith, eds. 1994. Updated with ECOMAP 2007: Cleland, D.T.; Freeouf, J.A.; Keys, J.E., Jr.; Nowacki, G.J.; Carpenter, C; McNab, W.H. 2007. Ecological Subregions: Sections and Subsections of the Conterminous United States [1:3,500,000] [CD-ROM]. Sloan, A.M., cartog. Gen. Tech. Report WO-76. Washington, DC: U.S. Department of Agriculture, Forest Service.

4. Loarie, S. R., Carter, B. E., Hayhoe, K., McMahon, S., Moe, R., Knight, C. A., & Ackerly, D. D. (2008). Climate Change and the Future of California's Endemic Flora. *PLoS ONE, 3*(6), Article No.: e2502.

5. Mittermeier, R. A., Gil, P. R., Hoffmann, M., Pilgrim, J., Brooks, T., Mittermeier, C. G., Lamoreux, J., & da Fonseca, G. A. B. (2004). Hotspots revisited: Earth's biologically richest and most endangered terrestrial ecoregions. Cemex: Mexico. 391 pages.

## *10KP: 10,000 Plants Genome Sequencing Project*

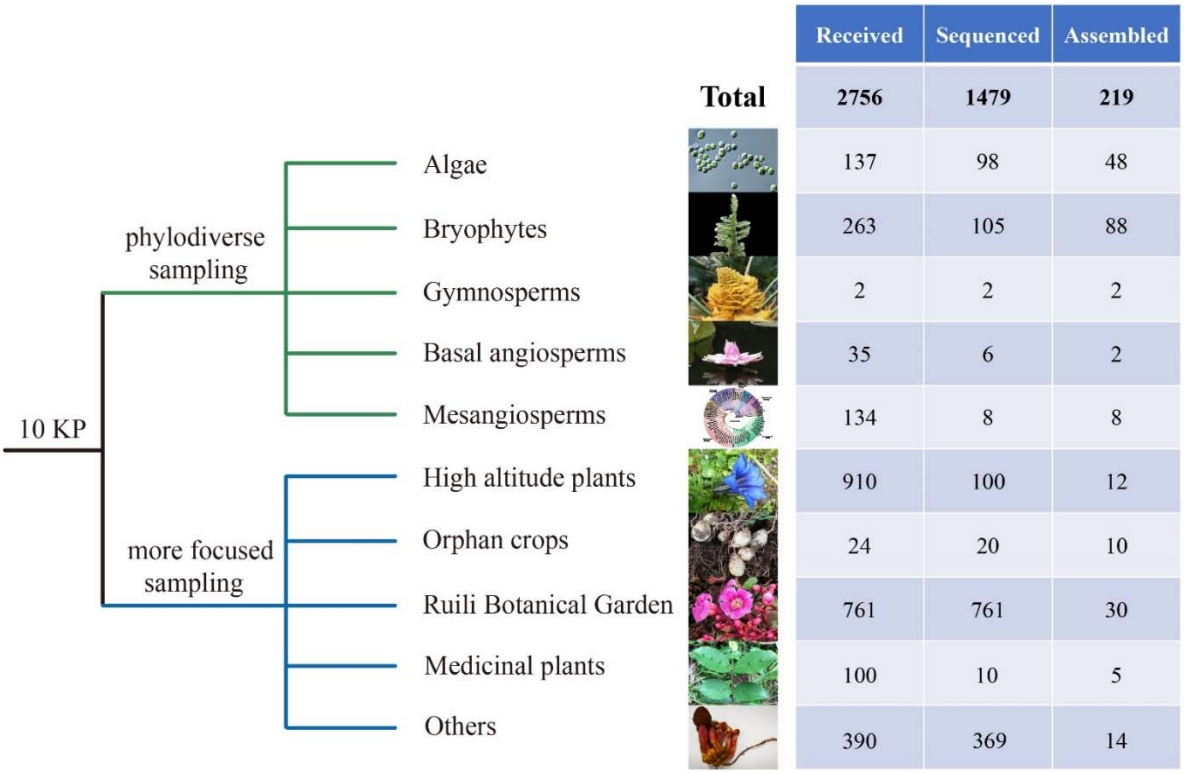
*Project Contacts*: Xun Xu, xuxun@genomics.cn and Xin Liu liuxin@genomics.cn

*Scope and Goals:* Applications for genomics continue to expand due to ongoing advances in low-cost, high-throughput technologies and methodologies. Plant diversity research has been one of many beneficiaries of these advances. However, understanding evolution and diversity in a phylogenomics context remains a great challenge, given the limited availability of genome-scale data across phylodiverse species (1). To address this gap, the 10KP project (10,000 Plants Genome Sequencing Project), a part of the Earth BioGenome Project (EBP), was launched in July 2017 (2). The project aims to sequence and characterize representative genomes from every major clade of green plants (Viridiplantae), including embryophytes, streptophyte and chlorophyte algae, other members of Archaeplastida such as red algae, and a select number of photosynthetic and heterotrophic protists (3). It is structured as an international consortium, open to the scientific community worldwide, including botanical gardens, culture collection centers, plant research institutes, universities, and private industry groups.

*Progress:* To date, we have received 2,756 plant specimens (including freeze-dried tissues/cells and DNA samples) covering 280+ families, and generated whole-genome-shotgun sequence for about 1,500 species. Among these, we have successfully assembled high-quality genomes for 219 species of streptophyte and chlorophyte algae, non-flowering embryophytes (liverworts, mosses, hornworts, ferns, and gymnosperms), and flowering plants (basal angiosperms, magnoliids, monocots/grasses, asterids, and rosids) (Figure 1). 10KP has published 20+ research papers describing many novel insights on the phylogeny and evolution of plants (4-8) (https://db.cngb.org/10kp/). As the most exciting findings will only become apparent after the research community has access to these genomes, data are periodically made public at https://db.cngb.org/10kp/database/.

Future directions include careful integration of genomic information with phenotypic traits, ecological, and metagenomic or multi-omics data. We remain open to all manner of collaborations.

Looking ahead, some of the biggest challenges for 10KP are the collection of samples from geographically restricted lineages, and cross-border transport of materials in compliance with the relevant international regulations (e.g. Nagoya protocol). Another major challenge is to obtain and transport HMW DNA at the requisite quantities and qualities. Finally, the assembly of highly-heterozygous and/or large and/or polyploid genomes remains difficult (Kress et al., 2021).




| | Received | Sequenced | Assembled |
|---|---|---|---|
| **Total** | **2756** | **1479** | **219** |
| Algae | 137 | 98 | 48 |
| Bryophytes | 263 | 105 | 88 |
| Gymnosperms | 2 | 2 | 2 |
| Basal angiosperms | 35 | 6 | 2 |
| Mesangiosperms | 134 | 8 | 8 |
| High altitude plants | 910 | 100 | 12 |
| Orphan crops | 24 | 20 | 10 |
| Ruili Botanical Garden | 761 | 761 | 30 |
| Medicinal plants | 100 | 10 | 5 |
| Others | 390 | 369 | 14 |

**Figure 1:** Current status of the 10KP project, showing the number of samples received, sequenced, and assembled. (Note: The left panel is not a species tree, it displays the sub-projects of 10KP)

*References:*


1. Twyford AD (2018) The road to 10,000 plant genomes. *Nat Plants* 4(6):312-313.
2. Normile D (2017) Plant scientists plan massive effort to sequence 10,000 genomes. *Science* 27:1-2.
3. Cheng S*, et al.* (2018) 10KP: A phylodiverse genome sequencing plan. *Gigascience* 7(3):1-9.
4. Wang S*, et al.* (2020) Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. *Nat Plants* 6(2):95-106.
5. Fan Y*, et al.* (2020) Dissecting the genome of star fruit (Averrhoa carambola L.). *Hortic Res* 7:1-10.
6. Liu H*, et al.* (2019) Molecular digitization of a botanical garden: high-depth whole-genome sequencing of 689 vascular plant species from the Ruili Botanical Garden.

## *The Global Genome Biodiversity Network (GGBN)*

*Project Contacts*: Ole Seberg, oles@snm.ku.dk  Katharine B. Barker, barkerk@si.edu

*Scope and Goals:* The Global Genome Biodiversity Network (GGBN) seeks to build Nagoya-compliant, universally discoverable frozen collections representing Earth's biodiversity, and available for research and non-commercial use. GGBN (Droege et al., 2016) is the only international organization tracking the existence, location, metadata, and accessibility of physical genomic resources for Eukaryotes. GGBN members maintain legal ownership of their collections and control their use. Establishing a national, centrally funded, enterprise-level biorepository is hard work, especially in the Global South, where GGBN tries to focus its funding. Nevertheless, the genomic revolution in biodiversity science will lead to creation of institutionally supported, legal collections of shared, genomic resources that conserve genome-quality samples. However, the insecure status of free access to digital sequence information (DSI) threatens this endeavor (Rohden et al., 2020). GGBN and its task forces also provide resources for the benefit of the collections and research community, including the GGBN Document Library, which exemplifies best practices in sample preservation, biorepository management, and legal considerations. GGBN is a model for responsible, accountable, trackable biodiversity genomics research. It partners with CryoArks, DiSSCo, EAZA Biobank, EBP, ECN, Swedish ESB, ESBB, Frozen Ark, GBIF, GCM, ISBER, SYNTHESYS+, Species360, and WFCC.

*Progress:* In 2007, GGBN (and its antecedents) created the data model for physical genomic samples (fixed or viable tissues, DNA, RNA, amplicons, eDNA, environmental samples, etc.), that links samples to sequences and taxa. The Network currently includes 100 members from 36 countries. To date, GGBN has made over 4,600 families, 28,000 genera, 79,000 species and 5.7 million samples discoverable for research. GGBN and its major partner, the Global Genome Initiative at the Smithsonian Institution has funded approximately 300 genomic projects, and continues to build collections of "dark" taxa, as well as funding efforts to grow the global network. GGBN strategically funds major biological specimen databases/models such as Symbiota, Specify, Arctos, Species360, and EMu to incorporate the GGBN data model, enabling hundreds of institutions to make their growing genomic collections discoverable.

*Acknowledgments*. JAC and KBB acknowledge GGI for funding.

*References:*

1. Droege, G, K. B. Barker, O. Seberg, J. A. Coddington, E. Benson, W. Berendsohn, B. Bunk, C. Butler, E. Cawsey, J. Deck, M. Döring, P. Flemons, B. Gemeinholzer, T. Hollowell, P. Kelbert, I. Kostadinov, R. Kottmann, R. Lawlor, C. Lyal, J., Mackenzie-Dodds, C. Meyer, D. Mulcahy, S. Nussbeck, E. Ó Tuama, T. Orrell, G. Petersen, T. Robertson, C. Söhngen, J. Whitacre, J. Wieczorek, P. Yilax, H. Zetschem, Y. Zhang, X. Zhou. 2016. The Global Genome Biodiversity Network (GGBN) Data Standard specification. Database (2016) 2016 : baw125 doi: 10.1093/database/baw125.

2. Rohden, F., S. Huang, G. Dröge, A.H. Scholz, K. Barker, W. G. Berendsohn, J. A. Coddington, M. da Silva, J. Overmann, O. Seberg, M. van der Bank, X. Xu. 2020. Combined study on digital sequence information in public and private databases and traceability. https://www.cbd.int/meetings/DSI-AHTEG-2020-01.

## *The Ag100Pest Initiative*

*Project Contact*: Anna K. Childers, anna.childers@usda.gov

*Scope and Goals*: Arthropod pests are responsible for significant economic losses and health damages across all agricultural sectors and threaten global food security. Genomics holds promise to enable new strategies for pest control, and genome sequencing has emerged as a *de facto* foundational step toward targeted arthropod pest management[1,2].

The mission of the United States Department of Agriculture (USDA) Agricultural Research Service's (ARS) Ag100Pest Initiative[3] is the generation of reference-quality genome assemblies and annotations for 100 top agriculturally relevant arthropods within a five-year period. The Initiative's vision is to establish critical genomic infrastructure for molecular research, leveraging the USDA ARS's unique expertise in arthropod pest management and agricultural genomics research, to develop specific and effective pest control measures that reduce damage to other insects. Ag100Pest also serves to enhance the USDA ARS's contribution to two international genome sequencing projects – the i5K Initiative[4] (http://i5k.github.io/) and the Earth BioGenome Project.

*Progress*: The Ag100Pest Initiative began in the fall of 2018 with a call for nominations of candidate pest species from the Animal and Plant Health Inspection Service (APHIS), the Federal Interagency Committee on Invasive Terrestrial Animals and Pathogens (ITAP), the Cooperative Agricultural Pest Survey (CAPS), and ARS researchers, as well as the broader arthropod research community. Diverse pest candidates were sought across agricultural stakeholder groups, including field crops, animal production, beekeeping, forests, rangeland, and stored products. While selection was prioritized for agricultural pests in the U.S., species with high potential to become established as invasive pests were also included. Priority was given to species where a genome would most impact active research projects.

Due to early successes and great demand, the Ag100Pest Initiative has expanded over time beyond its original goal of 100 species. More than 170 genome projects are currently in progress or planned, with teams fully equipped to support DNA extraction, library preparation, sequencing, assembly, and data submission to the National Center for Biotechnology Information (NCBI) under the umbrella project PRJNA533106. The i5K Workspace@NAL[5] (https://i5k.nal.usda.gov/) team completes the end-to-end pipeline with an arthropod community-focused genome database supporting data visualization through genome browsers, data query tools, functional gene annotation[6], the Apollo manual annotation platform[7], and generation of official gene sets.

The Ag100Pest Initiative utilizes a single specimen for long-read sequencing. Continuous methods development over the timeline of the project has allowed the Ag100Pest Initiative to push the boundaries of arthropod sequencing to the extremes, achieving success with species barely visible to the naked eye and species with haploid genome sizes over 8.5 Gb, all the way to the ultimate frontier for arthropod genomics – species with small physical size, but large genomes.

The project to date covers 8 arthropod orders, 54 families and 159 species. Some species will have more than one assembly where there is value in comparing sexes,

populations, or subspecies. A list of species and general information about the Ag100Pest Initiative can be viewed at http://i5k.github.io/ag100pest.

*References:*


1. Coates, B., Poelchau, M., Childers, C., Evans, J., Handler, A., Guerrero, F., Skoda, S., Hopper, K., Wintermantel, W., Ling, K., Hunter, W., Oppert, B., Pérez de León, A., Hackett, K., and Shoemaker, D. (2015). Arthropod genomics research in the United States Department of Agriculture-Agricultural Research Service: Current impacts and future prospects. Trends in Entomology. 11. 1-27.

2. Poelchau, M.F., Coates, B.S., Childers, C.P., de Leon, A.A.P., Evans, J.D., Hackett, K. and Shoemaker, D. (2016) Agricultural applications of insect ecological genomics. Current Opinion in Insect Science, 13, pp.61-69. 10.1016/j.cois.2015.12.002

3. Childers, A.K., Geib, S.M., Sim, S.B., Poelchau, M.F., Coates, B.S., Simmonds, T.J., Scully, E.D., Smith, T.P.L., Childers, C.P., Corpuz, R.L., Hackett, K., and Scheffler, B. (2021) The USDA-ARS Ag100Pest Initiative: High-Quality Genome Assemblies for Agricultural Pest Arthropod Research. Insects, 12, 626. 10.3390/insects12070626

4. i5K Consortium. (2013) The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. Journal of Heredity 104 (5), 595-600. 10.1093/jhered/est050

5. Poelchau, M., Childers, C., Moore, G., Tsavatapalli, V., Evans, J., Lee, C.Y., Lin, H., Lin, J.W. and Hackett, K. (2015) The i5k Workspace@ NAL—enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Research, 43(D1), pp.D714-D719. 10.1093/nar/gku983

6. Saha, S., Cooksey, A.M., Childers, A.K., Poelchau, M.F., and McCarthy, F.M. (2021) Workflows for Rapid Functional Annotation of Diverse Arthropod Genomes. Insects, 12, 748. 10.3390/insects12080748

7. Dunn, N.A., Unni, D.R., Diesh, C., Munoz-Torres, M., Harris, N.L., Yao, E., Rasche, H., Holmes, I.H., Elsik, C.G. and Lewis, S.E. (2019) Apollo: Democratizing genome annotation. PLoS Computational Biology, 15(2), p.e1006790. 10.1371/journal.pcbi.1006790

## *Plant GARDEN ：a portal web site for accessing plant genome, DNA marker and SNP information*

Project Contact: Sachiko Isobe, sisobe@kazusa.or.jp

*Scope and Goals:* With the advance in NGS technology, *de novo* genome sequence assembly and resequencing have been conducted in various plant species. Development of a database, which stores references genome and variants information is important to manage the large amount of sequence data generated each day. Many plant genome databases have been developed for genome and gene sequences in diverse plant species, or genetic diversity data for specific families or species. However, there was no database able to compare genome, gene and variants information across species on a single database. Hence, we have developed a plant genome portal site, Plant GARDEN (Genome And Resource Database Entry; https://plantgarden.jp/en/index), to provide diverse information relating plant genomics and genetics in divergent plant species.

The EBP aims to create genome sequence information for whole organisms on Earth in order to preserve biodiversity. Producing genome sequence *per se* is an important task, however, we should consider how the genome sequence information is used at the same time. To achieve the aim, our mission is to create an environment that connects genome information for the people in society. Therefore, we have created a simple and user-friendly WUI (Web-based User Interface) for Plant GARDEN to remove a barrier between scientists and non-experts. We expect that Plant GARDEN will open the genome information to the broad society, and contribute to save the planet's diversity with this knowledge.

*Progress:* Several types of data are stored in Plant GARDEN: Reference genome, gene sequences, PCR-based DNA markers, trait-linked DNA markers identified in genetic studies, SNPs, and In/dels on publicly available sequence read archive (SRA, Table). The data registered in Plant GARDEN currenty (2021 Aug) is 169 assembled genome sequences, 8,136,812 gene sequences, 313,115 DNA markers, 8,225 QTLs, and 3,812 SNP list (gvcf files). In addition, we have re-annotated all the genes registered in Plant GARDEN by using a functional annotation tool, Hayai-Annotation (Ghelfi et al., 2019), based on KusakiDB (https://github.com/aghelfi/kusakiDB) to compare orthologous relationships of genes. KusakiDB is a database of orthologous genes in plant, consisting of sequence information derived from OrthoDB (https://www.orthodb.org/), UniProt (https://www.uniprot.org/), and RefSeq (https://www.ncbi.nlm.nih.gov/refseq/). The re-annotated genes are associated with OG IDs on OrthoDB, which enable cross-species gene comparison.

This research is supported by the Database Integration Coordination Program (DICP) of JST/NBDC.

*Reference:*

1. Ghelfi A., Shirasawa K., Hirakawa H. and Isobe S. (2019), Hayai-Annotation Plants: an ultra-fast and comprehensive functional gene annotation system in plants. Bioinformatics 35, 4427–4429.

## *Squalomix: shark and ray genome sequencing to analyze their diversity and evolution*

*Project contact*: Shigehiro Kuraku, skuraku@nig.ac.jp

*Scope and goals*: Cartilaginous fishes (chondrichthyans) form a distinct class of vertebrates with more than 1,200 species, known mostly as sharks or rays. Among vertebrates, they, as a taxonomic class, have the longest evolutionary history of about 400 million years, in terms of the divergence of extant members. Different species in this taxon are characterized by unique traits including electromagnetic sensing, electricity generation, varying morphology sometimes with a flattened body and/or a toothed rostrum. The highlight of their biological enigmas is in their reproductive modes with high plasticity between oviparity and viviparity, while they occasionally exhibit parthenogenesis and intersexuality. Mainly because of overfishing, many cartilaginous fish populations are declining [1], which necessitates genomic platforms for evidence-based management. Despite its outstanding biological importance, genomic approaches have not been applied to this taxon until recently (reviewed in [2]).

To sequence the genomes of chondrichthyans, the project Squalomix (https://github.com/Squalomix/info) was launched in 2020 and is conducted mainly in RIKEN Kobe, Japan. The Squalomix project aims to provide genomic sequences and other genome-wide data including transcriptomes and epigenomes, and its network of collaboration includes researchers with diverse backgrounds and locations. Squalomix aims to tightly interact with other EBP-affiliated projects whose target species list includes cartilaginous fishes. Inclusive cooperation respecting complementary expertise is expected to overcome the long-standing difficulty in studying elasmobranchs sustainably.

*Progress*: In Squalomix, sample collection is performed cautiously to minimize the sacrifice of wildlife and is characterized by a rich marine fauna in Japan's neighboring temperate waters, with occasional sources from death stranding of elusive species. The project collaborates with aquariums oriented toward academic research, which play indispensable roles in relaying offshore sampling and enable sustainable sampling of embryos and blood from live individuals.

Another specialty of Squalomix is its expertise in laboratory solutions that are not confined to DNA sequencing. Access to fresh tissues from local aquariums facilitates embryological analysis, flow cytometry-based genome size quantification , and karyotyping using cultured cells. Cell culture in cartilaginous fishes, widely thought difficult because of their high body fluid osmolarity, was enabled by modifying the culture medium with balancing osmolytes [3]. Our cytological expertise also allowed various epigenomic analyses [4].

The sequencing strategy for Squalomix is designed to accommodate genomic characteristics of cartilaginous fishes, mostly with large, repetitive genomes. In the standard protocol formulated in January 2021, we start by estimating genome size using flow cytometry and karyotyping as well as by transcriptome sequencing. We then proceed to genome sequencing, which employs both short-read and long-read high-fidelity sequencing platforms, together with Hi-C data production for chromosome-scale scaffolding. The outputs will be curated with reference to genome size and chromosomal organization obtained separately. These validations allow us to scrutinize the inclusion of those genomic regions that are difficult to sequence and assemble, such as the Hox C genes [5]. Complete genome assembly is also demanded in corroborating scarce gene repertoires suggested for visual opsins and conventional olfactory receptors [5]. The

standard procedure outlined above has been applied to several species including the brown banded bamboo shark for which a draft genome assembly was already released and the zebra shark, one of the egg-laying species with the smallest known genome size among sharks.

*References:*


1. Pacoureau, N., Rigby, C.L., Kyne, P.M., Sherley, R.B., Winker, H., Carlson, J.K., Fordham, S.V., Barreto, R., Fernando, D., Francis, M.P., Jabado, R.W., Herman, K.B., Liu, K.M., Marshall, A.D., Pollom, R.A., Romanov, E.V., Simpfendorfer, C.A., Yin, J.S., Kindsvater, H.K., Dulvy, N.K., 2021. Half a century of global decline in oceanic sharks and rays. Nature 589, 567-571.

2. Kuraku, S., 2021. Shark and ray genomics for disentangling their morphological diversity and vertebrate evolution. Dev Biol 477: 262-272.

3. Uno, Y., Nozu, R., Kiyatake, I., Higashiguchi, N., Sodeyama, S., Murakumo, K., Sato, K., Kuraku, S., 2020. Cell culture-based karyotyping of orectolobiform sharks for chromosome-scale genome analysis. Commun Biol 3, 652.

4. Onimaru, K., Tatsumi, K., Tanegashima, C., Kadota, M., Nishimura, O., Kuraku, S., 2021. Developmental hourglass and heterochronic shifts in fin and limb development. eLife 10, e62865.

5. Hara, Y., Yamaguchi, K., Onimaru, K., Kadota, M., Koyanagi, M., Keeley, S.D., Tatsumi, K., Tanaka, K., Motone, F., Kageyama, Y., Nozu, R., Adachi, N., Nishimura, O., Nakagawa, R., Tanegashima, C., Kiyatake, I., Matsumoto, R., Murakumo, K., Nishida, K., Terakita, A., Kuratani, S., Sato, K., Hyodo, S., Kuraku, S., 2018. Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. Nat Ecol Evol 2, 1761-1771.

## *The Aquatic Symbiosis Genomics Project*

*Project Contact*: Mark Blaxter, mb35@sanger.ac.uk

*Scope and Goals:* Symbiosis shapes the natural world and has been fundamental to diversification of life on Earth, from the origin of eukaryotic cells and photosynthetic eukaryotes, through to ongoing evolution of new associations (Oulhen, Schulz, and Carrier 2016; Archibald 2014). A symbiotic partnership can draw from the independent evolutionary histories of both partners: each "inherits" the others' accumulated genomic wisdom. Symbiosis allows aerobic eukaryotes to colonise anaerobic habitats and heterotrophic animals to become photosynthetic. The Aquatic Symbiosis Genomics project (ASG) is a global initiative to generate reference genome sequences for ~500 eukaryotic symbiotic organisms, including "hosts" and microbial symbionts. We will use these genomes to understand the dynamics of genomic change in symbiosis and to support a growing community using genetic and genomic methods to promote symbiont diversity and survival. Community training in genomics and bioinformatics is an important part of our project, and we work with the European Bioinformatics Institute, Wellcome Connecting Science and The Carpentries to deliver custom training to partners.

*Progress:* ASG will sequence a wide breadth of target taxa and systems. We will interrogate photosymbioses, including cnidarians, acoels, annelids and molluscs. Symbioses with hydrogen sulphide metabolising prokaryotes that allow animals to colonise deep sea and vent habitats will be explored in annelids and molluscs. While most symbiont hosts are sessile or benthic some live in the water column: we will sequence pelagic molluscs, tunicates, cnidaria and ctenophores, and probe the diversity of light-emitting organs in cephalopods. Sponges are inhabited by communities of microbes that support nutrition and defence, and by sequencing diverse species we will identify microbes with long symbiotic association with sponges. Algal-fungal symbiosis in littoral lichens will be probed to understand adaptation to marine environments. New symbioses have been established continuously through evolutionary time, and we will also sequence the partners in some recently emerged symbioses between diverse single-celled protists and their prokaryotic and eukaryotic endosymbionts.

ASG is a collaboration between conservationists, ecologists, evolutionary biologists and experimentalists (see doi: https://doi.org/10.5281/zenodo.5385621 for a list of lead partners) exploring the biology of symbiosis. ASG is funded by the Gordon and Betty Moore Foundation's Symbiosis in Aquatic Systems initiative and the Wellcome Trust. ASG is organised in a hub and spokes model. Ten hubs lead the collection of specimens from pre-agreed lists. All collections follow rigorous ethical and legal codes. Specimens are shipped to the Wellcome Sanger Institute for sequencing and assembly, using strategies designed to assemble high-quality genomes of hosts and cobionts independently. Gene finding and annotation will be performed on each genome. The genome assemblies will be released openly to the European Nucleotide Archive, where a dedicated portal shows project progress (https://portal.aquaticsymbiosisgenomics.org/). Each submitted assembly will be accompanied by a Genome Note, crediting the full chain of work from collection to submission, and announcing the genomes' open availability for others to use.

*Acknowledgements:*
The Aquatic Symbiosis Genomics project is funded by the Gordon and Betty Moore Foundation (GBMF8897) and Wellcome Trust Grant 206194. For the purpose of Open

 The Aquatic Symbiosis Genomics project relies on engagement and support from the whole of the Tree of Life production genomics team and of many colleagues who are participants in the ten Hubs.

*References:*

1. Archibald, John. 2014. *One Plus One Equals One: Symbiosis and the Evolution of Complex Life*. Oxford University Press, USA.

2. Oulhen, Nathalie, Barbara J. Schulz, and Tyler J. Carrier. 2016. “English Translation of Heinrich Anton de Bary’s 1878 Speech, ‘Die Erscheinung Der Symbiose’ (‘De La Symbiose’).” *Symbiosis*. https://doi.org/10.1007/s13199-016-0409-8.

## *The European Innovative Training Network "Comparative Genomics of Non-Model Invertebrates" (ITN IGNITE)*

*Project contact:* Gert Wörheide, woerheide@lmu.de

*Scope and goals:* ITN IGNITE (www.itn-ignite.eu), funded through the Marie Skłodowska-Curie Actions of the European Horizon 2020 framework program, bundles the expertise of 13 lead investigators ("beneficiaries") and partners across Europe and Australia to jointly train 15 internationally recruited Early-Stage Researchers (ESRs) in all aspects of genomics. Its overarching research goals are to improve the sampling and analysis of invertebrate genomes, especially from undersampled branches of the animal tree of life, and to extend the toolbox for their analysis, including the development and deployment of innovative production-level software beyond the current state-of-the-art.

IGNITE's core training objectives are:

1. to convey broad interdisciplinary knowledge in organismal biology, animal physiology, ecology, biogeography, evolution, and genomics;
2. to provide the technical laboratory skills required for genomics;
3. to impart technical computing skills in programming and bioinformatic pipeline development;
4. to build experience in effective dissemination and communication of results to target audiences;
5. to provide young researchers with a unique set of relevant academic and non-academic transferable skills to enhance human resource development and entrepreneurship; and
6. to establish a strong and long-lasting European and international network in invertebrate genomics.

*Progress:* IGNITE's training objectives are being reached by means of three levels of training. First, in-depth local training-by-research through individual scientific doctoral projects is provided and additionally through participation in complementary skills and scientific course-level training offered at each of the beneficiaries' institutions. The 15 individual research projects revolve around four broad topics: i) Function – genomes and the organism, ii) Ecology – genomes in the environment (1), iii) Evolution – genomes through Earth's history (2), iv) Bioinformatics – new tools to study genomes and to publish genomic data (3–5).
Second, in so-called yearly "network-wide training events" (NTEs, a sort of "summer schools"), inter/multidisciplinary, intersectoral, and transferable-skill training is provided to the whole group of ESRs by different beneficiaries (depending on topic) and/or external trainers. For example, the first NTE centered around sample collection and DNA/RNA extraction, a second one - self-organized by the ESRs - focused on bioinformatics and programming.
Third, interdisciplinary and/or intersectoral secondments, where each ESR carries out topical work related to his/her doctoral project in another beneficiaries' lab for a period of normally one month, round up the training program.
Through their excellent interdisciplinary and intersectoral training spanning from biology to bioinformatics and computer science, IGNITE's graduates will be in a prime position to take up leadership roles in both academia and industry. IGNITE is fully devoted to FAIR (Findable, Accessible, Interoperable, Reusable) data and open science principles and promotes, through various measures and activities, gender equality and diversity.

Although the IGNITE program, as the first of its kind, goes a long way in training the next generation of highly skilled genomicists, large international endeavors such as the Earth Biogenome Project (EBP) or the European Reference Genome Atlas (ERGA) initiative call for additional in-depth coordinated and focused training programs in genomics. However, the training measures implemented and experiences made in IGNITE may serve as a template for the future development of such training programs.

*Acknowledgements:* IGNITE has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 764840. We wish to thank Joe Lopez for continuously supporting IGNITE as a scientific advisor.

*References:*

1. Z. Chen, Ö. Doğan, N. Guiglielmoni, A. Guichard, M. Schrödl, The *de novo* genome of the "Spanish" slug *Arion vulgaris* Moquin-Tandon, 1855 (Gastropoda: Panpulmonata): massive expansion of transposable elements in a major pest species. *BioRxiv*, 2020.11.30.403303 (2020).

2. B. Bettisworth, A. Stamatakis, Root Digger: a root placement program for phylogenetic trees. *BMC Bioinformatics* **22**, 225 (2021).

3. R. E. Rivera-Vicéns, C. Garcia Escudero, N. Conci, M. Eitel, G. Wörheide, TransPi – a comprehensive TRanscriptome ANalysiS PIpeline for *de novo* transcriptome assembly. *bioRxiv*, 2021.02.18.431773 (2021).

4. N. Guiglielmoni, A. Houtain, A. Derzelle, K. van Doninck, J.-F. Flot, Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* **22**, 303 (2021).

5. M. Dimitrova, *et al.*, A streamlined workflow for conversion, peer review, and publication of genomics metadata as omics data papers. *GigaScience* **10**, giab034 (2021).

## *The 10,000 fish genomes project (Fish10k)*

*Project Contact:* Kun Wang, wangkun@nwpu.edu.cn; Guangyi Fan, fanguangyi@genomics.cn

*Scope and Goals:* Following the breakthroughs in DNA sequencing technology, researchers are committed to bridge the gap between the diverse phenotypes and genotypes to understand the mystery of evolution [1, 2]. Vertebrates occupy the top ecological niches of the sea and land and have shown vigorous vitality since the Cambrian explosion [3]. Until now, about half of vertebrates are water-dwelling fishes, with more than 34,000 described species from ~80 orders, ~529 families, and ~5,000 genera, including jawless, cartilaginous, and bony fishes, of which bony fishes can be divided into ray-finned and lobe-finned fishes [4]. The ray-finned fishes and their subclassification teleost comprise about 99% and 95% of living fish species, respectively. Fishes are the most successfully evolved taxa in the kingdom of zoology that occupy almost all the aquatic biomes and thrive among various habitats from freshwater to ocean, from tropics to polar, from highlands to the hadal zone, etc. [4, 5]. Notwithstanding their dramatic diversity and evolutionary importance, only a few hundred fish genomes have been resolved to date [6], and the largest phylogenetic work including 303 species [7] still cannot meet the growing needs of academics. There is a pressing need for more high-quality genome resources of fishes. Here, we introduce the Fish10k Genome Project, which will sequence all the genera from cartilaginous fishes, ray-finned and lobe-finned fishes through three phases. This project will provide a rich genetic resource that will greatly advance our understanding of vertebrate evolution.

*Progress:* The fish10k project will be carried out in three phases, covering all orders (~ 500 species, in three years), all families (~ 3,000 species, in three years), and all genera (~6,500 species, in four years), respectively. We are currently working on the first stage and going to sequence 450 bony fishes and 50 cartilaginous fish species. About one-third of the bony fish samples have been collected, the majority of the cartilaginous fish samples have been collected, most of the sequencing work has been done, and the genome assembly process is underway. Our project also facilitates the genome analysis of African lungfish [8] and non-teleost ray-finned fishes [9], as well as the evolution of endothermy in teleosts [10], which has led to a deeper understanding of the vertebrate evolution. For the next phases, we will seek more collaborators for sample collection and analysis, and adapt sequencing analysis protocols to technological developments.

*References:*

1. Koepfli, K.P., B. Paten, and S.J. O'Brien, *The Genome 10K Project: a way forward.* Annu Rev Anim Biosci, 2015. **3**: p. 57-111.

2. Lewin, H.A., et al., *Earth BioGenome Project: Sequencing life for the future of life.* Proc Natl Acad Sci U S A, 2018. **115**(17): p. 4325-4333.

3. Shu, D., *Cambrian explosion: Birth of tree of animals.* Gondwana Research, 2008. **14**(1): p. 219-240.
4. Nelson, J.S., T.C. Grande, and M.V.H. Wilson, *Fishes of the World*. 5 ed. 2016: Wiley.
5. Wang, K., et al., *Morphology and genome of a snailfish from the Mariana Trench provide insights into deep-sea adaptation.* Nat Ecol Evol, 2019. **3**(5): p. 823-833.

6. Fan, G., et al., *Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K).* Gigascience, 2020. **9**(8).

7. Hughes, L.C., et al., *Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data.* Proc Natl Acad Sci U S A, 2018. **115**(24): p. 6249-6254.

8. Wang, K., et al., *African lungfish genome sheds light on the vertebrate water-to-land transition.* Cell, 2021. **184**(5): p. 1362-1376.e18.

9. Bi, X., et al., *Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes.* Cell, 2021. **184**(5): p. 1377-1391.e14.

10. Wu, B., et al., *The Genomes of Two Billfishes Provide Insights into the Evolution of Endothermy in Teleosts.* Mol Biol Evol, 2021. **38**(6): p. 2413-2427.

## *The Zoonomia Project*

*Project Contact:* Kerstin Lindblad-Toh kersli@broadinstitute.org; Elinor Karlsson elinor@broadinstitute.org

*Scope and goals:* The Zoonomia Project, by comparing the genomes of hundreds of different mammals, is investigating mammalian evolution in exceptional detail, and building a powerful new resource that will accelerate the search for genetic variants underlying common and rare human diseases. Since the explosion of mammalian diversity starting 100 million years ago, trillions of mutations have been tested by natural selection. By harnessing the power of this natural experiment, we achieve unprecedented insight into mammalian genome function, connecting evolutionary patterns of conservation and acceleration to changes in organismal phenotypes.

With our whole-genome alignment of 241 diverse mammals, the largest such alignment to date, now complete, we are now exploring how this resource can be used. We are tracing the genomic origins of morphological, physiological and behavioral traits that differ between mammals adapted to widely varying ecological niches. We are addressing long-standing questions about the origins of mammals, and examining how chromosomal translocations, transposable elements, and other types of variation have shaped the genomes of mammalian clades. In parallel, we are exploring how patterns of evolutionary constraint can be leveraged to accelerate the search for new therapeutics, by providing a novel perspective on the search for the functional variation underlying human diseases.

*Progress:* The project data, consisting of 131 new mammalian genomes combined with 110 previously sequenced genomes, and aligned in a reference-free Cactus alignment, has been published **(Zoonomia Consortium 2020)**(https://zoonomiaproject.org). As part of this work, we showed that analyzing the genetic diversity of reference genomes, can help identify at-risk populations, and could be a valuable tool for prioritizing species for more detailed biodiversity assessment. Endangered species from the IUCN list were significantly less diverse than species categorized as of least concern (Zoonomia Consortium 2020). Illustrating its versatility. The Zoonomia project data has also already been used to find adaptations in sea otters (Beichman et al. 2019), evasion of cancer in capybara (Herrera-Álvarez et al. 2020), speciation in howler monkeys (Baiz et al. 2020) and to predict whether species are susceptible to SARS-CoV-2 infection through the ACE2 receptor (Damas et al. 2020). We are now resolving persistently challenging nodes in the eutherian phylogeny (Murphy et al. 2020), and identifying overarching patterns in the evolution of genome structure, at scales ranging from short repetitive elements, to gene families, to entire chromosomes. By examining classes of regulatory elements using single-base constraint, we can identify those common to all mammals and those found only in some lineages. Lineage-specific conservation or acceleration may indicate a direct role in innovation, allowing us to discern the genomic basis of structural and physiological phenotypes like hibernation. Finally, we are investigating the genomic risk factors for human diseases. Our analysis shows that highly constrained bases are strongly enriched for variants explaining common disease heritability, and that constraint is more informative than most types of functional annotations. With genomes from 241 mammals, we have a unique perspective on human genome function, unattainable even in studies with hundreds of thousands of humans.

*References:*


1. Baiz, Marcella D., Priscilla K. Tucker, Jacob L. Mueller, and Liliana Cortés-Ortiz. 2020. “X-Linked Signature of Reproductive Isolation in Humans Is Mirrored in a Howler Monkey Hybrid Zone.” *The Journal of Heredity* 111 (5): 419–28.

2. Beichman, Annabel C., Klaus-Peter Koepfli, Gang Li, William Murphy, Pasha Dobrynin, Sergei Kliver, Martin T. Tinker, et al. 2019. “Aquatic Adaptation and Depleted Diversity: A Deep Dive into the Genomes of the Sea Otter and Giant Otter.” *Molecular Biology and Evolution* 36 (12): 2631–55.

3. Damas, Joana, Graham M. Hughes, Kathleen C. Keough, Corrie A. Painter, Nicole S. Persky, Marco Corbo, Michael Hiller, et al. 2020. “Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (36): 22311–22.

4. Herrera-Álvarez, Santiago, Elinor Karlsson, Oliver A. Ryder, Kerstin Lindblad-Toh, and Andrew J. Crawford. 2020. “How to Make a Rodent Giant: Genomic Basis and Tradeoffs of Gigantism in the Capybara, the World’s Largest Rodent.” *Molecular Biology and Evolution*, November. https://doi.org/10.1093/molbev/msaa285.

5. Murphy, William J., Nicole M. Foley, Kevin R. Bredemeyer, John Gatesy, and Mark S. Springer. 2020. “Phylogenomics and the Genetic Architecture of the Placental Mammal Radiation.” *Annual Review of Animal Biosciences*, November. https://doi.org/10.1146/annurev-animal-061220-023149.

6. Zoonomia Consortium. 2020. “A Comparative Genomics Multitool for Scientific Discovery and Conservation.” *Nature* 587 (7833): 240–45.

## *The Chilean 1000 Genomes Initiative*

*Project Contact:* Miguel L Allende mallende@uchile.cl

*Scope and Goals:* This national initiative is led by five centers of excellence in the areas of genomics, mathematics and systems biology. The Chilean 1000 genomes initiative aims to obtain genome sequences of endemic organisms from many of the unique habitats present along the country. The work has focalized, for the most part, on carrying out a comprehensive metagenomic survey of the animal, plant and microbial species inhabiting the Atacama Desert, the driest on earth. In this region, life is possible in the high altitude Andean steppes (the *Altiplano*) where there is seasonal precipitation. It is an ecosystem of concern due to its high sensitivity to climate change[1].

*Progress:* In Atacama, genome sequencing efforts are underway to characterize many of the plant species inhabiting an altitudinal gradient, from 2,000 to 4,000 meters above sea level (masl)[2]. These plants, belonging to several diverse taxa, have convergently adapted to withstand severely harsh conditions that include poor nutrient availability, large temperature variations, high UV radiation and extreme aridity. Together with the sampling and characterization of about 75 species of plants, we have also surveyed the metagenome of the associated soils, in order to correlate plant adaptations with recruitment of specific cohorts of microorganisms in their rhizosphere. The sequencing initiative has also focused on plant species that live in the hyperarid Atacama lowlands, which bloom on the rare occasions (once every ~10 years) when rainfall occurs. For instance, the main flowering desert species, *Cistanthe longiscapa* (Montiaceae; Caryophyllales)[3], shows genomic and physiological adaptations that allow it to thrive and flower for a few weeks, after long periods of dormancy.
In the case of animals, we have sequenced the genomes of four species of critically endangered pupfish (Cyprindontiformes; Cyprinodontidae) of the genus *Orestias*, which inhabit both freshwater lakes and salt pans in the *Altiplano*, at up to 4,000 masl[4]. Members of this group have undergone evolutionarily recent allopatric speciation after the large paleolake of the Central Andes became fragmented when this mountain range emerged; each lake or saltpan harbors a single species. Two of the Orestias species we have selected became adapted to very high water salinity and to extreme variations in temperature. In this case, we hypothesize that the rapid evolutionary divergence of their remarkable physiology can be revealed through comparative and functional genomic analyses.

Other areas of interest of the initiative and where surveys and species collection are currently underway are the Antarctic continent and the Eastern South Pacific coast. Overall, our aim is to characterize species from the unique environments found in Chile with particular emphasis on those that impose challenges to life or that display an increased threat from climate change and habitat degradation. All of the generated data will be made available in public repositories. In addition to research activities, since 2018 the initiative has promoted genomics education

through the “Chile sequences Chile” program, where high school students carry out genomic sequencing in their schools.

*Acknowledgments*
The author wishes to acknowledge the contribution of the five centers of excellence that lead the Chilean 1000 genomes initiative: The Center for Genome Regulation (ANID/FONDAP/15200002), the Center for Mathematical Modeling (Basal Grant AFB170001), the Millennium Institute for Integrative Biology (Iniciativa Científica Milenio-MINECON), the Advanced Center for Chronic Diseases (ANID/FONDAP/15130011) and the Geroscience Center for Brain Health and Metabolism (ANID/FONDAP/15150012).

*References:*

1. Díaz FP, Latorre C, Carrasco-Puga G, Wood JR, Wilmshurst JM, Soto DC, Cole TL, Gutiérrez RA. (2019). Multiscale climate change impacts on plant diversity in the Atacama Desert. Glob Chang Biol. 25(5):1733-1745. doi: 10.1111/gcb.14583.

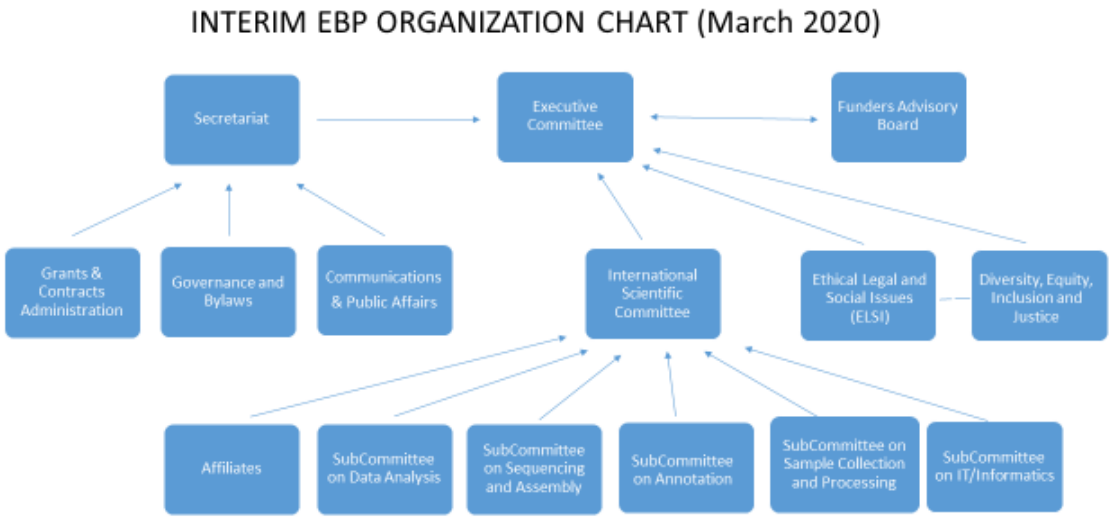
2. Eshel et al. (2021). Plant ecological genomics at the limits of life in the Atacama Desert. Proc Natl Acad Sci USA. (Accepted for Publication [MS# 2021-01177PRR]).

3. Hershkovitz, M.A. (1991). Taxonomic notes in Cistanthe, Calandrinia and Talinum (Portulaceae). Phytologia 70(3),209-225.

4. Vila I, Morales P, Scott S, Poulin E, Véliz D, Harrod C, Méndez MA. (2013). Phylogenetic and phylogeographic analysis of the genus Orestias (Teleostei: Cyprinodontidae) in the southern Chilean Altiplano: the relevance of ancient and recent divergence processes in speciation. J Fish Biol. 82(3):927-43. doi: 10.1111/jfb.12031.

**Additional Figure and Dataset**

***Fig. S1. Interim governance structure of the Earth BioGenome Project.*** See www.earthbiogenome.org for further details on committee roles and responsibilities.



***Dataset S1. Progress of EBP-affiliated projects in whole genome sequencing and the production of reference genomes.*** Total numbers in each column may include species that overlap between projects.