SPECIAL FEATURE: PERSPECTIVE



# Darwinian genomics and diversity in the tree of life

Taylorlyn Stephan[a,1], Shawn M. Burgess[a], Hans Cheng[b], Charles G. Danko[c], Clare A. Gill[d], Erich D. Jarvis[e,f], Klaus-Peter Koepfli[g,h], James E. Koltes[i], Eric Lyons[j], Pamela Ronald[k,l,m,n], Oliver A. Ryder[o,p], Lynn M. Schriml[q], Pamela Soltis[r], Sue VandeWoude[s], Huaijun Zhou[t], Elaine A. Ostrander[a], and Elinor K. Karlsson[u,v,w,2]

Edited by Gene Robinson, Entomology, University of Illinois at Urbana–Champaign, Urbana, IL; received September 16, 2021; accepted November 23, 2021

**Genomics encompasses the entire tree of life, both extinct and extant, and the evolutionary processes that shape this diversity. To date, genomic research has focused on humans, a small number of agricultural species, and established laboratory models. Fewer than 18,000 of ∼2,000,000 eukaryotic species (<1%) have a representative genome sequence in GenBank, and only a fraction of these have ancillary information on genome structure, genetic variation, gene expression, epigenetic modifications, and population diversity. This imbalance reflects a perception that human studies are paramount in disease research. Yet understanding how genomes work, and how genetic variation shapes phenotypes, requires a broad view that embraces the vast diversity of life. We have the technology to collect massive and exquisitely detailed datasets about the world, but expertise is siloed into distinct fields. A new approach, integrating comparative genomics with cell and evolutionary biology, ecology, archaeology, anthropology, and conservation biology, is essential for understanding and protecting ourselves and our world. Here, we describe potential for scientific discovery when comparative genomics works in close collaboration with a broad range of fields as well as the technical, scientific, and social constraints that must be addressed.**

comparative genomics | evolution | biodiversity | natural models | genomics

Genomics, from its inception, has encompassed evolutionary and interspecies comparisons (1), in a tacit acknowledgment that genome sequence is almost meaningless without context. Comparative genomics harnesses evolution to investigate genome function. The second genome sequenced for a free-living organism (*Mycoplasma genitalium*) was immediately compared to the first (*Haemophilus influenzae*) (2). The human genome was compared to mouse (3), chicken (4), dog (5), and then 28 mammals simultaneously (6), and recently to 240 mammals (7). The first plant genome, the model organism *Arabidopsis thaliana* (8), was compared to eight other crucifers (9). Genomic positions that resist change over long periods of time may be essential for survival, and those that accumulate changes unusually quickly in particular

[a]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20817; [b]Avian Disease and Oncology Laboratory, Agricultural Research Service, US Department of Agriculture, East Lansing, MI 48823; [c]Department of Biomedical Sciences, Baker Institute for Animal Health, Cornell University, Ithaca, NY 14850; [d]Department of Animal Science, Texas A&M University, College Station, TX 77843; [e]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY 10065; [f]HHMI, Chevy Chase, MD 20815; [g]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630; [h]Smithsonian Conservation Biology Institute, National Zoological Park, Washington, DC 20008; [i]Department of Animal Science, Iowa State University, Ames, IA 50011; [j]School of Plant Sciences, BIO5 Institute, University of Arizona, Tucson, AZ 85721; [k]Department of Plant Pathology, University of California, Davis, CA 95616; [l]The Genome Center, University of California, Davis, CA 95616; [m]The Innovative Genomics Institute, University of California, Berkeley, CA 94720; [n]Grass Genetics, Joint Bioenergy Institute, Emeryville, CA 94608; [o]San Diego Zoo Wildlife Alliance, Escondido, CA 92027; [p]Department of Evolution, Behavior, and Ecology, University of California San Diego, La Jolla, CA 92093; [q]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201; [r]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611; [s]Department of Micro-, Immuno-, and Pathology, Colorado State University, Fort Collins, CO 80532; [t]Department of Animal Science, University of California, Davis, CA 95616; [u]Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01655; [v]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01655; and [w]Broad Institute of MIT and Harvard, Cambridge, MA 02142

Author contributions: T.S., S.M.B., H.C., C.G.D., C.A.G., E.D.J., K.-P.K., J.E.K., E.L., P.R., O.A.R., L.M.S., P.S., S.V., H.Z., E.A.O., and E.K.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.



[1]Present address: Department of Biomolecular Engineering and Bioinformatics, University of California, Santa Cruz, CA 95060.

[2]To whom correspondence may be addressed. Email: elinor.karlsson@umassmed.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2115644119/-/DCSupplemental.

Published January 18, 2022.

lineages may be involved in development and propagation of advantageous phenotypes.

Evolutionary innovations in nonhuman species have already resulted in new therapeutics. Decades before the advent of genomics, the ovarian cancer drug paclitaxel (Taxol) was discovered in the Pacific yew tree, where it protected against pathogens (10). Transcription activator-like effectors, discovered in a plant pathogenic bacterium, led to the development of novel genome editing tools and a new therapeutic for acute lymphoblastic leukemia (11).

Despite this legacy, genomics has increasingly focused on humans (Fig. 1). The United Kingdom Biobank Project (12) and All Of Us Research Program (13) are scaling to millions of humans. Meanwhile, only 4% of animals and 2% of plants have a single representative genome assembly (14). Rather than advocating a shift away from humans, we propose broadening the scope to include more nonhuman data. By removing barriers that silo comparative genomics and human genomics into distinct disciplines, and integrating with nongenomic disciplines, we can transform every species into a "model organism" and accelerate discovery.

A broader focus is essential to protecting the ecosystems we depend on. Biodiversity is the unrecoverable foundation of comparative genomics. It is being lost at an alarming rate (15). Combining genomic tools with meticulous phenotyping and creative cross-disciplinary collaboration can help address this crisis (16, 17).

## Harnessing the Evolution of All Life

Evolution is an unparalleled tool for research. Functionally, it is somewhat analogous to a long-term clinical trial, initiated several billion years ago and enrolling all life on Earth. It includes species with evolutionary trajectories altered by human action, through both accelerated natural selection and experimental selection, creating populations we use as research models (Fig. 2 and *SI Appendix*, Table S1). As mutations arise, they are evaluated for their effect on survival and reproduction, as eloquently described by Charles Darwin more than 150 y ago:

> It may be said that natural selection is daily and hourly scrutinising, throughout the world, every variation, even the slightest; rejecting that which is bad, preserving and adding up all that is good; silently and insensibly working, whenever and wherever opportunity offers, at the improvement of each organic being in relation to its organic and inorganic conditions of life (18).

By comparing genomes within and between species, and connecting genomic variation to changes in cells, organisms, and ecosystems, we access the results of a natural experiment carried out on an unfathomable scale.

Genomic studies that include only humans capture just the last 50,000 y or so of evolution. Even so, naturally occurring human mutations guided the design of safe and effective drugs. Rare coding mutations that cause abnormally low cholesterol inspired the new class of PCSK9 inhibitor drugs (19), which reduce the risk of vascular events without major offsetting adverse events.

Other species routinely exhibit evolutionary adaptations that allow them to tolerate conditions that are disease-causing in humans. Hibernating mammals become obese and insulin-resistant in preparation for hibernation and, while hibernating, lose synaptic connectivity and suffer repeated episodes of ischemia and reperfusion (20). Yet they emerge healthy each spring in a physiological feat that holds clues for treating obesity, neurodegeneration, and heart disease (21).

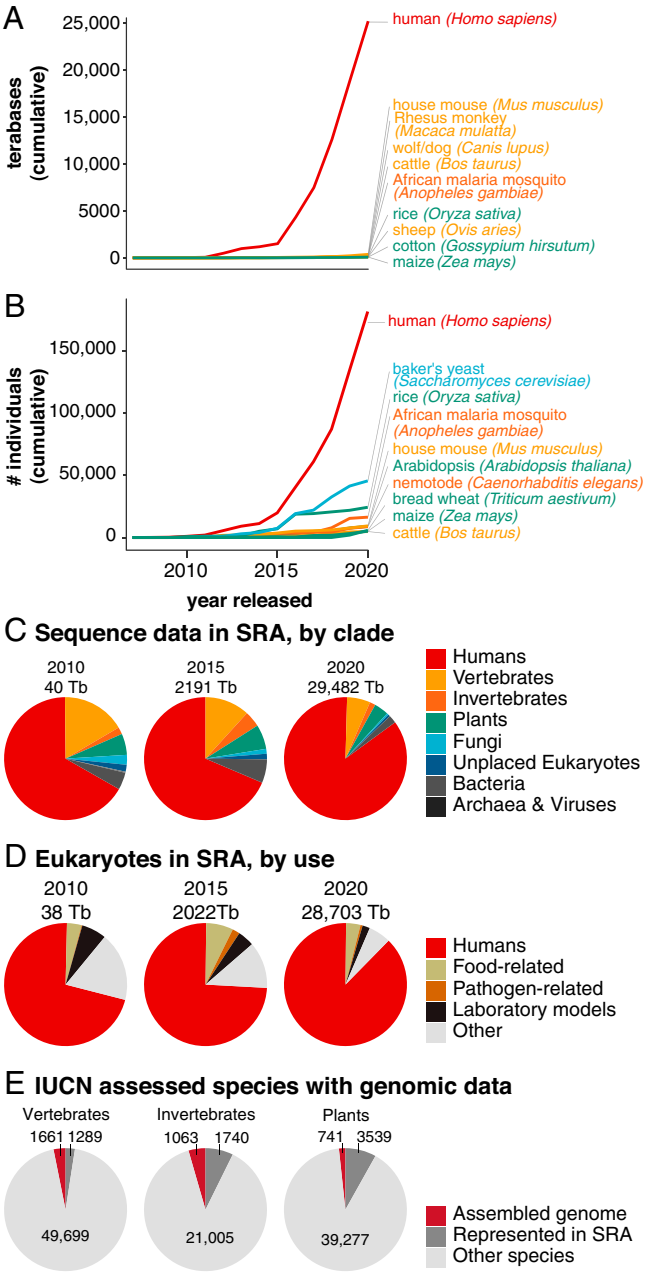


**Fig. 1. Species diversity in the Sequence Read Archive (SRA). The amount of human data exceeds that of the next top 10 species, measured as (*A*) terabases and (*B*) individuals sequenced. (*C*) The human proportion increased between 2010 and 2020, and (*D*) the proportion from species without known commercial/medical relevance ("other") dropped. (*E*) A tiny proportion of IUCN-recognized (80) species have a reference genome (red) or are otherwise represented in the SRA (dark gray). Retrieved November 14, 2020.**

Traits like hibernation are the outcome of a complex and iterative evolutionary process. Organisms adapt to changes in their environment, and by doing so, change that environment, driving adaptation in other species, and so on, ad infinitum. The substrate for this evolutionary arms race is mutation, both small (single nucleotide) and large-scale (structural variants and polyploidy), and the backdrop is a series of unpredictable natural events that constantly reset the stage. The mass extinction that marked the demise of nonavian dinosaurs opened up ecospace

for the diversification of mammals (22) and birds (23) into thousands of species extant today.

The sheer complexity of evolution may encourage a reductionist approach, but this is insufficient. Even when the mechanism of a single variant is known in great detail, its effect in the context of other genome variation can be unpredictable (24). Discovering the emergent properties of complex systems using large datasets is a more powerful approach, as demonstrated in biophysics (25), comparative genomics (7, 26), and human genomics (12).

We are poised to enter a new age of science heralded by new genome-editing technologies (27, 28). Scientists can directly edit DNA to achieve desired outcomes, whether curing heritable diseases, depleting invasive populations, reducing pathogen reservoirs, or engineering crops resilient to environmental stress. Even as we contemplate the role of genetic creators, we cannot yet predict the organismal impact of changing even simple genomes.

To understand how genomic variation shapes organismal variation and function, it is both possible and necessary for research to encompass the full scope of the evolution of life. We can now measure and modify the natural world with unprecedented precision, but researchers pursuing innovative and cross-disciplinary research encounter systemic and logistical barriers. By addressing these challenges, all species can contribute as genetic systems for understanding and protecting our world.

## A Darwinian Approach to Genomics

> There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved (18).

All organisms on Earth share a common origin, each being one of billions of variations on a common theme. Hundreds of genes shared between yeast and humans are so functionally similar that the human version can substitute in yeast (29). Sponge enhancers control cell-type–specific gene expression in zebrafish and mice, lineages that last shared a common ancestor 700 million y ago (30).

If genomes are the source code of life, then the interpretation is an interaction between that code, the cellular machinery that reads it, and the environment in which it is manifested. While a genome sequence may be essential, it is not sufficient to elucidate the complex processes underlying development, growth, differentiation, host defense, environmental responses, and countless other facets of biology. This requires transcriptomic and epigenomic data that vary by cell type and over time, samples from many individuals per species, and many samples per individual (31). It also requires new technology for collecting functional data, phenotypes, and environmental measurements at scale, including epigenomic assays (32), remote sensing [e.g., airborne lidar (33)], thermal and fluorescence imaging (34), passive environmental sampling (35), geographic information system mapping (36), and participatory science (37). Finally, it requires situating genomic change in the evolutionary timeline and in relation to geologic, ecologic, and anthropologic events.

Just as technology for large-scale sequencing transformed genomics, new technologies for large-scale data collection are transforming how we study the natural world. Biology is transitioning from a single-investigator, hypothesis-based endeavor to team-driven, discovery-based science. Collaborations that encompass biology, medicine, computer sciences, and historical sciences, as well as data-driven methods for studying complex systems, can support a more systems-based, and less reductionist, investigation of organisms and ecosystems.

Here, we call for a more Darwinian approach to genomics that considers all forms of life, their interactions, and the natural environment that shaped them. Charles Darwin developed his theory of evolution by natural selection by studying a wide range of species, including insects, plants, arthropods, and vertebrates. The groundbreaking first edition of *On the Origin of Species* (18) illustrates how a broader perspective enables discoveries not possible when focused on a single species. For scientists today, this requires collaborations that span diverse communities, within and outside of science, and the technology, scale, and skills to address multidimensional questions. Below, we review key discoveries that illustrate the potential of this approach, and propose strategies to support the cross-disciplinary integration essential to success (Box 1).

We use the term "Darwinian" after careful deliberation. For many scientists, Darwin's name, more than any other single word, evokes the connection between the processes of evolution and the organisms and ecosystems "most beautiful and most wonderful" (18) of the natural world. Since its publication, Darwin's work has been misused to lend a false veneer of scientific credibility to racist, ableist, and sexist beliefs that continue to cause immeasurable damage. We recognize our obligation to confront this history, and to work to undo the harm it has caused.

**Mutual Affinities.** Collaboration is essential for expanding the scope of comparative genomics; this requires overcoming traditional barriers separating disciplines and scientists from communities. To reconstruct the historic dispersal of *Oryza sativa* ssp. *japonica*, the progenitor of much of our domesticated rice, sequence data for 1,400 strains was insufficient. Combining geographic, environmental, archaeobotanical, and paleoclimate information revealed that rice diversified into temperate and tropical *japonica* rice during a global cooling event 4,200 y ago, suggesting that further research might find adaptations to changing climates (38).

Collaborations that span ethnic, geographic, and socioeconomic backgrounds improve productivity and data richness (39), but require communication, leadership, open thinking, and appreciation for all participants (40). Particularly when collaborations span fields with different norms, or include remote study locations, success depends on trust and ensuring all participants are acknowledged (41). Funding agencies, journal editors, and academic institutions can encourage collaborations with reward structures that credit all team members (42) . Scientists sequencing the genome of the tuatara, a reptile endemic to New Zealand and the only living member of its order, partnered with Ngātiwai, the Māori *iwi* (tribe) holding guardianship over the individual tuatara studied (43). Their successful collaboration, recognized with authorship, was guided by common goals of increasing knowledge and supporting conservation, with Ngātiwai participating in data-use and benefit-sharing discussions. People working within Indigenous or traditional knowledge systems can offer information on species behavior, habitats, and conservation issues unfamiliar to scientists working within "Western" knowledge systems (44). Using DNA barcoding technology, scientists in the Velliangiri Hills of India identified three species of herbaceous plants new to science, but already classified as distinct species in the local traditional knowledge system (45).

***Box 1. Perspectives in Comparative Genomics and Evolution Workshop.***

In August 2019, three funding agencies—the National Human Genome Research Institute (NIH), the National Institute of Food and Agriculture (US Department of Agriculture), and the National Science Foundation—convened a 2-d workshop on Perspectives in Comparative Genomics and Evolution, where 120 participants evaluated the state of the field, focusing on commonalities across humans, model organisms (traditional and nontraditional), agricultural and wildlife species, and microbes. For this paper, the authors synthesized common themes, roadblocks, and strategies that emerged from the workshop.

Engaging community members directly in research can facilitate collection of large and geographically disparate datasets needed to explore real-world evolutionary processes, while positively impacting communities. Using eBird, a community science project whose participants have collected over 915 million bird observations (46), scientists had sufficient data to assess whether speciation is associated with niche divergence in *Aphelocoma* jays (47). The spread of the cabbage white butterfly, *Pieris rapae*, a destructive agricultural pest, was traced using samples collected by over 150 volunteers from 32 countries, which implicated specific human activities as possible drivers (48). Children in India, Kenya, Mexico, and the United States surveyed mammalian biodiversity near their schools using camera traps, collecting high-quality data while learning to value their local natural history (49).

Such research should align with the Convention on Biological Diversity, ensuring local knowledge is included and attributed, that data are correctly interpreted, and that cultural practices are respected (44). All stakeholders, including local communities, should benefit (50). Full partnership with field scientists is vital. Their meticulous observations and careful sample collection, along with the curation and annotation of the specimens in both living and natural history collections, are the keystone of interdisciplinary research.

**Geometrical Powers of Increase.** Our conception of a more collaborative approach to comparative genomics is rooted in the open-data culture of genomics, exemplified by the Human Genome Project (51) and the sometimes controversial (52) shift to team projects that generate and analyze multidimensional datasets (53). Today, genomic data dominate, but other data types are expanding [imaging, personal wearable devices, remote sensing, and electronic medical records (54)]. Resources like the Global Biodiversity Information Facility (GBIF) (55) and the Integrated Digitized Biocollections (iDigBio) provide standards and open-source tools for unifying disparate organismal occurrence data (56). The Genomic Observatories Metadatabase (GEOME) (57) links the Sequence Read Archive (SRA) (58) to ecological data repositories not configured for genomic information.

A single reference genome is rarely sufficient for answering biological questions, but when shared, supports many different studies (53). Historically, researchers were forced to weigh the often considerable cost of generating a reference against the value of other data that could be collected instead. Today, falling costs and new technology are making high-quality reference genomes more achievable (59). The Earth BioGenome Project proposes producing reference genomes for ∼2 million known eukaryotic species in the next 10 years (60).

High-quality reference genomes can lead to discoveries even in well-studied organisms. Using the highly contiguous genome for the bioenergy crop switchgrass (*Panicum virgatum*), scientists compared hundreds of plants grown in common gardens spanning 1,800 km of latitude. They discovered genetic variation accumulating on the less constrained subgenome, suggesting a polyploid genome may enhance adaptive potential (36). Comparing high-contiguity genomes for six bat species revealed positive selection at hearing-related genes, suggesting echolocation is an ancestral trait lost in the nonecholocating bats (61).

Data structures that accommodate genetic diversity within species are still under development. The traditional linear genome structure struggles even with human data, introducing pervasive reference biases (62). For species with more genetic diversity, like gorillas and butterflies (63, 64), new representations, like graph-based pangenomes, are essential (65).

With falling sequencing costs, functional genomic assays [e.g., RNA-sequencing, chromatin accessibility assays, Hi-C, PRO-seq, and ribosome profiling (66–69)] can capture cellular change over time, by cell and tissue types, and with environment. Comparing the epigenomic landscape in 10 mammalian species using chromatin immunoprecipitation-sequencing uncovered unexpected plasticity in regulatory elements, including switching from promoter to enhancer, and vice versa (70).

Functional genomic assays are essential for investigating mechanisms of action. To pinpoint a variant conferring increased obesity risk in humans, scientists combined long-range chromatin interactions, expression quantitative-trait locus analysis, luciferase reporter assays, and directed perturbations in primary cells (71). Joint analysis with comparative genomic data identified an endogenous retrovirus insertion that encoded an enhancer involved in activating the inflammasome, and may be a pathogen-response adaptation (72).

In more easily manipulated laboratory models, single-cell, single-nucleus, and spatial sequencing methods are revealing the fundamental biology of the cell. By embedding sequence barcodes in fertilized zebrafish eggs, and editing them with each cell division, cell lineages were tracked throughout embryo development and the lineage tree reconstructed (73).

For single-cell organisms—including bacteria, archaea, and protists—single-cell genomics captures culture-independent diversity. Single-cell transcriptomics on organisms from the hindgut of wood-feeding termites showed four protist species with distinct roles in wood degradation, suggesting microbiome diversity is essential for termite survival (74).

Cloud-computing resources, which offer massive compute and storage capacity, are essential as sequence datasets grow (75). When cohorts reach half a million, and phenotypes number over 7,000, correlating genotype and phenotype requires millions of CPU hours. Using cloud-based clusters, such jobs are completed in a week (76). Today, the compute time required to align genomes, essential for comparative genomics, scales quadratically with genome size (77), although algorithmic advances could improve efficiency. To make protein structure prediction more accurate and efficient, AlphaFold's neural network-based algorithm predicts energy landscapes rather than calculating binary contact maps (21, 74).

**Collecting All Forms in Time and Space.** Extending genomics to consider all forms of life requires prioritizing sample collection in challenging environments. Long-read sequencing technology

is of little use if the input DNA is fragmented due to sample degradation. Chromatin conformation capture can measure the three-dimensional structure of the genome only if samples have intact nuclei. To measure the response of cells to stimuli, living cell cultures are needed, an expensive and labor-intensive resource to establish (*SI Appendix*, Fig. S1).

Collecting high-quality samples from species living in regions remote from scientists is particularly challenging. Sampling three highland wild dogs in New Guinea required field biology studies, GPS tagging, video, and collaboration with local scientists, but rediscovered a population of free-living dogs long thought extinct (78). While captive populations may be easier to sample, zoos house representatives of only 12% of the ~31,771 terrestrial vertebrate species (79, 80), and botanical gardens capture only a fraction of plant species (81).

The number of samples is sometimes more critical than sample quality, particularly when a high-quality reference genome is available. Pairing samples with metadata, such as collection dates, locations, and phenotypes, makes it possible to evaluate population demography, and identify mutations that can impact fitness. Whole-genome sequencing of century-old gorilla specimens, annotated with collection dates, revealed a drop in genetic diversity associated with increased inbreeding in the critically endangered Grauer's gorillas, but not in the mountain gorilla, which did not experience the same population declines (82).

New methods for extracting and analyzing DNA allow samples in less-than-ideal condition to be used. The oldest DNA sequence, recovered from wooly mammoths living in Siberia 1 million y ago, shows that North American mammoths likely descended from a hybridization event, with cold climate adaptations already present (83). By sequencing slow-degrading structural proteins in samples 3.5 million y old, the origin of modern camels was traced to the forested Arctic of the Mid-Pliocene (84). Sequencing can characterize complex mixes of species in paleo-samples. Fossil rodent middens are mixtures of plant and animal remains, collected by foraging rodents ranging ~100 m, and preserved for thousands of years. Sequencing them captures the community of plants, animals, bacteria, and fungi at a single location in the past with exquisite resolution (85). Epigenomic profiling of ancient specimens, while technically challenging, could improve predictions of species resilience (86).

Methods developed for old or degraded samples support studies of natural populations where invasive collections are not possible. Methods that enrich host DNA make feces samples, dominated by microbes, more useful (87). DNA extracted from elephant tusks traced samples to their source, helping law enforcement disrupt poaching activities (88).

Portable sequencing technology, deployable in remote locations, could be transformative by eliminating shipping risks and supporting field-based training with local scientists leading environmental efforts (89). In the Ecuadorian Chocó rainforest, one of the world's most imperiled biodiversity hotspots, on-site sequencing distinguished species through DNA barcoding (90). In Hawaii, long ribosomal DNA sequencing in the field yielded a phylogeny of 83 spiders that captured the adaptive radiation of the genus *Tetragnatha* (91).

Genomic, epigenomic, and proteomic assays all require destructive sampling, and this cost should be carefully considered. The scientists who identified the first archaic human from the Denisovan lineage did so by destroying part of a tiny sliver of bone, the only sample available for DNA extraction (92). Their work showed Denisovans were evolutionarily distinct from Neanderthals and modern humans, transforming our understanding of human evolution.

Destructive sampling puts museums in the difficult position of judging which projects are worthy. Genomic data offers a window into the past unattainable through other technology (93). Sequencing of 28 fossils, including 7 from museums, discovered a now-extinct horse genus endemic to North America, adding a branch to the phylogeny of mammals (94). Museums may be reluctant to authorize damage to specimens in their care (95), but collecting genomic data could also mitigate, somewhat, loss of collections in the future. Even minimal genomic data from the 20 million samples lost when Brazil's National Museum burned down in 2018 (96) would comprise an unparalleled scientific resource. Further complicating the question, the same sample may yield more information with time. Two years after the first Denisovan paper (92), a subsequent paper described a DNA library preparation method requiring half as much input (97). Guidelines are needed for researchers, museums, and journals to ensure samples are used responsibly, projects are high quality and ethically executed, and that data and specimen information are shared (98).

**Disentangle the Inextricable Web of Affinities.** Collecting, quantifying, and comparing complex phenotypes in diverse species, at scale, is perhaps the greatest challenge in comparative genomics (99). The observable phenotype of an organism reflects the interaction of "preprogrammed" traits encoded in a genome with its environment, suggesting we could, in theory, predict its structure and function from its genome. To understand how phenotypes evolve, we must compare the same species in differing environments (36), different species with shared traits (100), and outliers with incredible adaptations.

In laboratory models, phenotyping technology is well developed and genomic resources are robust, elevating species such as yeast, fruit fly, nematodes, zebrafish, rat, mouse, *Arabidopsis*, rice, and others as primary models for fundamental biological questions. Using an experimental design that inverts traditional gene mapping, the International Mouse Phenotyping Consortium disrupted 3,328 genes and produced models for 360 human diseases, including the first for some bleeding disorders and ciliopathies (101). Deeply sequencing 1,504 mutant lines of the model rice cultivar Kitaake (*O. sativa* ssp. *japonica*) found 90,000 mutations affecting 58% of genes, including a causal mutation for short-grain rice (102).

Laboratory models are diversifying with the emergence of versatile, species-agnostic gene knockout technology. Making a primate model carrying even one biallelic mutation through breeding is difficult, given long maturation times and low reproduction rates. With CRISPR-based genome editing, multiple variants can be engineered in parallel, producing new models for human polygenic diseases (103). Integrating large DNA constructs into mammalian stem cells allows systematic locus-scale analysis of genome function (104). In the future, editing ancient DNA sequences into living cells could enable paleoepigenomics.

Domesticated species are natural models for linking phenotypes, many from intentional and inadvertent selective breeding, to genomic changes. The phenotypically diverse food crops—cabbage, kale, collards, Brussels sprouts, broccoli, and cauliflower—were developed from a single plant species, *Brassica oleracea*, primed for a dramatic response to breeding by an ancient whole-genome triplication (105). Strong, recent selective breeding, as in ornamental goldfish (106) and dog breeds

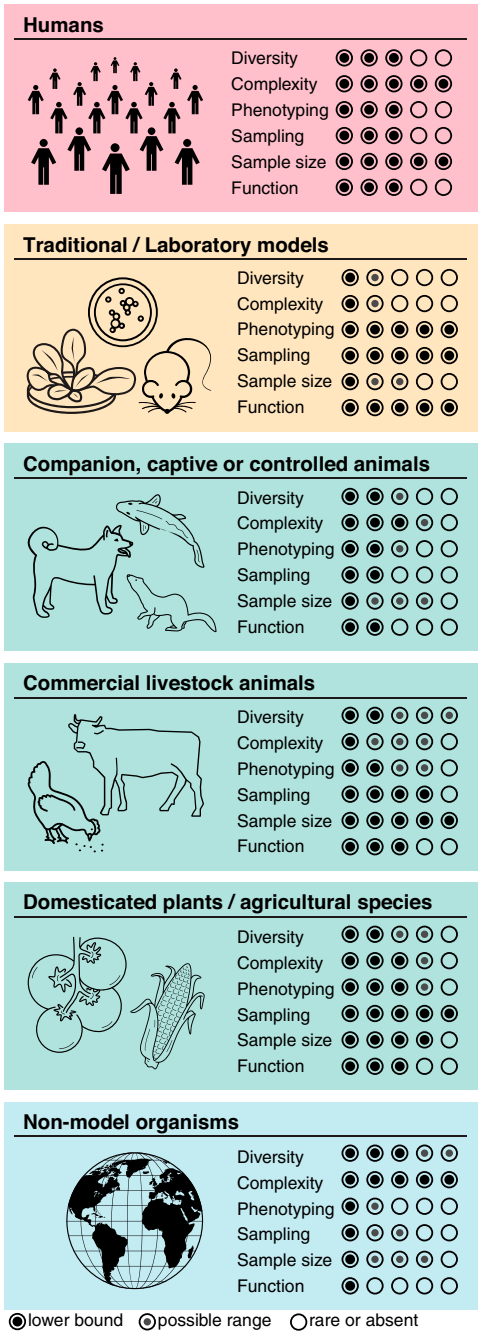


**Fig. 2. Different types of study populations have different strengths. "Diversity": genetic diversity in populations, ranging from inbred (e.g., laboratory mice) to outbred/highly diverse. Humans (midpoint) are outbred but less diverse than many species. "Complexity": genetic complexity of traits; low in the laboratory mouse, with controlled genetic background and environment, and high in humans, where most traits are complex. "Phenotyping": ease of collecting phenotype data, ranging from only noninvasive phenotyping in natural environments, to invasive laboratory phenotyping. In humans (midpoint), resources like electronic medical records make it possible, but not easy, to collect detailed phenotypes at scale. "Sampling": ease of collecting samples, ranging from only minimally invasive sampling in wild-caught individuals, to populations where euthanasia and tissue collection are feasible. "Sample size": number of individuals that can be sampled, ranging from <100 (endangered species or laboratory animals requiring costly care) to millions (humans). "Function": potential for functional genomics (epigenomics, cellular and organoid models, genetic engineering, and so forth). In humans, cellular models are well developed, but organism-level experimentation is not possible.**

(107), leaves distinctive "signals" around causal variants. Testing for signals of selection in 82 strains of budding yeast connected the unique ability of cheese-making strains to grow quickly on galactose to the replacement of the *GAL1*, *GAL7*, and *GAL10* genes with orthologs from another species (108).

The very large population sizes and, for commercially relevant traits, rigorous phenotyping in modern commercial livestock make them useful genomic models. One million chickens are vaccinated every hour against an oncogenic herpesvirus using a vaccine repeatedly reformulated for more virulent strains (109), making commercial chicken farms a model for intersecting host genomics, viral evolution, and disease epidemiology. The vaccine prevents severe disease but not transmission, and effectively controls outbreaks (110), reassuring for humans suffering through the COVID-19 pandemic.

In natural populations, genomic studies focused on dissecting the etiology of traits are challenged by the need for large numbers of well-phenotyped samples (111), yet technologies like Google Earth (112) can provide rich new data sources. To detect systems-level patterns in ecological diversity, and the impact of environmental change, researchers paired sequencing of samples collected by community scientists with habitat, bioclimate, soil, topography, and vegetation data (113). To collect tick samples with the geographic, temporal, and image data needed to study pathogen transmission dynamics, scientists used social media to enlist the help of thousands of community scientists (114).

Combining genomic and nongenomic data can identify drivers of disease spread, thereby informing the design of effective interventions. Phylogenomic analysis of 772 complete SARS-CoV-2 genomes, when paired with epidemiology data, showed how superspreader events shaped the course of the COVID-19 pandemic (115).

A perspective that considers all species, rather than focusing on humans or a few familiar models, provides more options for selecting the optimal model for the scientific question at hand (Fig. 2). The protein CD163 was identified as the likely host receptor for the porcine virus PRRSV (116) using cells from African green monkey cells (116), leading to the production of PRRSV-resistant pigs that could save hundreds of millions of dollars per year (117).

## Natural Variation of Form and Function

> We see these beautiful co-adaptations most plainly in the woodpecker and mistletoe; and only a little less plainly in the humblest parasite which clings to the hairs of a quadruped or feathers of a bird; in the structure of the beetle which dives through the water; in the plumed seed which is wafted by the gentlest breeze; in short, we see beautiful adaptations everywhere and in every part of the organic world (18).

Through genomic technology, we can read the results of the biological experiment that is life on Earth. Billions of years of selection on random alterations in the genetic code have produced species that thrive in a huge range of niches (118). The tiny tardigrade can survive temperatures from −272 °C to 151 °C, a vacuum, and exposure to gamma rays (119), and has recently been recovered from a 16-million-y-old piece of amber (120). Evolution is a powerful guide for developing safe and effective therapeutics, as it favors adaptations that avoid fitness-reducing pleiotropic effects.

While we can't know all the forces that shaped life on Earth, the outcomes are observable. Comparing 72 fungal genomes, scientists discovered that multicellularity in filamentous fungi arose through different mechanisms than in other multicellular lineages (121). Characterizing vocal learning ability in dozens of bird species and comparing their genomes revealed the trait likely evolved three different times (23). Populations of humans (122), dogs (123), horses (124), deer mice (125), and ducks (126) adapted to high altitudes through selection on the *EPAS1* gene. In stickleback fish, a freshwater adaptation for reduced armor plating was mapped to the gene *GDF6*; a deletion of a conserved regulatory element controlling GDF6 may explain humans' unusually short toes (127).

Interpreting the results of Earth's evolutionary experiment will be challenging. Species need to be considered in aggregate, in the context of the physical environment and all the eukaryotic and prokaryotic commensal, competitive, and parasitic relationships that comprise complex ecosystems. Rather than using a reductionist approach that eliminates complexity, genomic research can use that complexity to investigate phenotypes well beyond those tractable in traditional laboratory model organisms.

## Addressing Roadblocks

> After my return to England it appeared to me that … collecting all facts which bore in any way on the variation of animals & plants under domestication & nature, some light might perhaps be thrown on the whole subject ( 128).

Darwin developed his theory of natural selection by considering patterns shared across seemingly very different species. His input data were a naturalist's observations, but adopting this approach in genomics requires far more complex resources. We must go beyond the obvious (e.g., integrating genetics, bioinformatics, and medicine), and engage with anthropology and other historical sciences, experts using different knowledge systems, and the public. In the process, it is critical to address the systemic racism, sexism, and ableism that has been reinforced by twisted interpretations of Darwin's evolutionary theory. Collaborations where each field retains its unique strengths, rather than developing a single perspective, are essential, as are new modalities for communicating across skill sets that are currently "domain specific" (*SI Appendix*, Table S1). We suggest six pillars for accomplishing this.

First, we propose that biology is the starting point for developing a common dialogue. In genomics, the work of biologists is too often perceived as the "sample-collecting" prelude to the main project, but connecting genomic variation to changes in organisms and ecosystems is fundamentally a biological research challenge. Thus, the contribution of biologists, particularly nonmolecular and noncomputational biologists, should be carefully considered and appropriately resourced when setting funding and sample dispersal priorities.

Second, increasing the number of and training for computational biologists is critical. The field is understaffed and underfunded, and those in it struggle with conflicting priorities. We need to recruit computational experts into the biological sciences, and provide the training in biology and biomedicine tailored to their area of interest, ranging from laboratory work to field biology (129).

Third, comprehensive training in computational biology should be a requirement for all fields. While not reducing the need for highly skilled computational biologists, it will enable field and laboratory-based scientists to do crucial initial analyses. Better computational and data literacy, taught as an integral part of science education (130), will facilitate collaborations between those collecting data and those doing much of the analysis. Existing training opportunities [e.g., Data Carpentry workshops (131); weeklong NSF-sponsored Genomics of Diseases of Wildlife courses (132)] should be expanded globally, and more extended programs developed (e.g., "embedding" in another research group for a semester).

Fourth, training opportunities in science communication should be expanded (133). Genomics is a global science, and as such requires engagement between scientists and nonscientists alike. Education programs that embrace narrative, social learning, digital media, and gamification reach hundreds of thousands of people (134). Ongoing, effective communication between all stakeholders will help ensure that research ultimately benefits public health, sustainable agriculture, and biodiversity conservation.

Fifth, we call for data-sharing with minimal restriction and delay, and adherence to the FAIR (findability, accessibility, interoperability, and reuse of digital assets) data principles (135). The FAIR principles are followed by major genomic consortia including ENCODE (136), FAANG (Functional Annotation of Animal Genomes) (137), the Alliance of Genome Resources (138), and the Genomic Standards Consortium (139). When necessary, we should modify existing data standards to support cross-species comparisons.

Finally, more support for museums, including zoos, aquaria, and botanical gardens, is an absolute necessity (140). Museums are irreplaceable reservoirs of specimens, history, and ideas, and communicate the value of science. They are essential partners in any effort to understand all of the world's species. Rather than sample providers, we envision museums as something akin to a public library, where information is shared, specimens are protected, and safeguards supporting responsible access are in place.

## Conclusion

We stand at the precipice of a new genomic age, with the power to both read and write DNA. Even as therapeutics based on genome editing save lives (141), we grapple with the ethical dilemmas inherent in editing germline cells (142). The most useful guidebook to this brave new world is the evolutionary past, and its constant testing of new variants through natural selection. With the technology to sequence DNA, assay cellular activity, and measure phenotypes at massive scales, we can read the results of that grand experiment.

To understand how genomes shape organisms and ecosystems, we must look outside our own species to all life on Earth. The conceptual foundation is basic evolutionary theory, some of it first described by Charles Darwin, but it requires scale and scope that would have been difficult for the 19th century naturalist to grasp, yet is now achievable. It is incumbent on us to figure out how we use these tools effectively for scientific discovery, for advancing medicine, and for protecting our world.

To illustrate the potential, we return to the Galapagos for a thought experiment with Darwin's finches (143). Imagine we could collect genome sequences not just for every bird on those islands, but for all the animals, plants, and microbes interacting with each bird, and imagine we could do so for every generation since the birds first colonized the islands. Our data collection continues to the present day, and we capture the disruption of the Industrial

Age, and know the history of geopolitical events. We measure organismal phenotypes, from morphology to health to feeding behavior to reproduction, and record all interactions between species, and changes with each generation, with incredible precision. Finally, we collect detailed data on rainfall, sea and air temperatures, and other meteorological events.

In reality, in-depth monitoring can inflict unacceptable damage on fragile ecosystems, illustrating the need for careful study design, and technology that minimizes harm. Any project so broad in scope raises complicated ethical, legal, and social issues that must be carefully addressed (144). The potential for discovery in such rich datasets, extending far beyond genomics, encapsulates the vision of a more extensive, inclusive, Darwinian approach to genomics.

***Data Availability.*** There are no data underlying this work.


## Acknowledgments

We thank the members of the Perspectives in Comparative Genomics and Evolution Workshop organizing committee, including Jennifer Troyer, Theodore Morgan, Lakshmi Matukumalli, Claudio Mello, Cyril Gay, Lorjetta Schools, Heidi Sofia, and Kris Wetterstrand. The workshop Perspectives in Comparative Genomics and Evolution was supported by the National Human Genome Research Institute's 2020 Strategic Planning process and the following grants: NSF-1939343, NIFA-AFRI-2019-67015-29490, and NIH GM122968.



1 B. Kuska, Beer, Bethesda, and biology: How "genomics" came into being. *J. Natl. Cancer Inst.* **90**, 93 (1998).
2 C. M. Fraser *et al.*, The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
3 E. S. Lander *et al.*, International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4 International Chicken Genome Sequencing Consortium, Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
5 K. Lindblad-Toh *et al.*, Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
6 K. Lindblad-Toh *et al.*, Broad Institute Sequencing Platform and Whole Genome Assembly Team; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team; Genome Institute at Washington University, A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
7 Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
8 Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).
9 A. Haudry *et al.*, An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45**, 891–898 (2013).
10 M. C. Wani, H. L. Taylor, M. E. Wall, P. Coggon, A. T. McPhail, Plant antitumor agents. VI. The isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. *J. Am. Chem. Soc.* **93**, 2325–2327 (1971).
11 R. Benjamin *et al.*, UCART19 Group, Genome-edited, donor-derived allogeneic anti-CD19 chimeric antigen receptor T cells in paediatric and adult B-cell acute lymphoblastic leukaemia: Results of two phase 1 studies. *Lancet* **396**, 1885–1894 (2020).
12 C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
13 J. C. Denny *et al.*, All of Us Research Program Investigators, The "All of Us" research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
14 N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
15 E. Stokstad, Global efforts to protect biodiversity fall short. *Science* **369**, 1418 (2020).
16 M. Stange, R. D. H. Barrett, A. P. Hendry, The importance of genomic variation for biodiversity, ecosystems and people. *Nat. Rev. Genet.* **22**, 89–105 (2021).
17 R. Phelan, B. Baumgartner, S. Brand, E. Brister, Intended consequences statement. *Conservat. Sci. Prac.* **3**, e371 (2021).
18 C. Darwin, *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* (J. Murray, 1859).
19 J. C. Cohen, E. Boerwinkle, T. H. Mosley Jr., H. H. Hobbs, Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
20 S. M. Mohr, S. N. Bagriantsev, E. O. Gracheva, Cellular, molecular, and physiological adaptations of hibernation: The solution to environmental challenges. *Annu. Rev. Cell Dev. Biol.* **36**, 315–338 (2020).
21 H. T. Jansen *et al.*, Hibernation induces widespread transcriptional remodeling in metabolic tissues of the grizzly bear. *Commun. Biol.* **2**, 336 (2019).
22 R. W. Meredith *et al.*, Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524 (2011).
23 E. D. Jarvis *et al.*, Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
24 X. Li, B. Lehner, Biophysical ambiguities prevent accurate genetic prediction. *Nat. Commun.* **11**, 4923 (2020).
25 A. W. Senior *et al.*, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
26 S. Feng *et al.*, Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257 (2020).
27 H. Shivram, B. F. Cress, G. J. Knott, J. A. Doudna, Controlling and enhancing CRISPR systems. *Nat. Chem. Biol.* **17**, 10–19 (2021).
28 P. D. Hsu, E. S. Lander, F. Zhang, Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
29 A. H. Kachroo *et al.*, Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348**, 921–925 (2015).
30 E. S. Wong *et al.*, Deep conservation of the enhancer regulatory code in animals. *Science* **370**, eaax8137 (2020).
31 R. M. Sherman, S. L. Salzberg, Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
32 E. Giuffra, C. K. Tuggle; FAANG Consortium, Functional Annotation of Animal Genomes (FAANG): Current achievements and roadmap. *Annu. Rev. Anim. Biosci.* **7**, 65–88 (2019).
33 G. D. Pearse, J. Morgenroth, M. S. Watt, J. P. Dash, Optimising prediction of forest leaf area index from discrete airborne lidar. *Remote Sens. Environ.* **200**, 220–239 (2017).
34 T. Roitsch *et al.*, Review: New sensors and data-driven approaches—A path to next generation phenomics. *Plant Sci.* **282**, 2–10 (2019).
35 C. M. Poutasse *et al.*, Discovery of firefighter chemical exposures using military-style silicone dog tags. *Environ. Int.* **142**, 105818 (2020).
36 J. T. Lovell *et al.*, Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* **590**, 438–444 (2021).
37 M. Chandler *et al.*, Contribution of citizen science towards international biodiversity monitoring. *Biol. Conserv.* **213**, 280–294 (2017).
38 R. M. Gutaker *et al.*, Genomic history and ecology of the geographic spread of rice. *Nat. Plants* **6**, 492–502 (2020).
39 H. A. Valantine, F. S. Collins, National Institutes of Health addresses the science of diversity. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12240–12242 (2015).
40 National Research Council (U.S.), Committee on the Science of Team Science, National Research Council (U.S.); Division of Behavioral and Social Sciences and Education, *Enhancing the Effectiveness of Team Science* (National Academies Press, 2015).
41 D. Stokols, S. Misra, R. P. Moser, K. L. Hall, B. K. Taylor, The ecology of team science: Understanding contextual influences on transdisciplinary collaboration. *Am. J. Prev. Med.* **35**(suppl.) S96–S115 (2008).
42 Institute of Medicine, *Bridging Disciplines in the Brain, Behavioral, and Clinical Sciences* (The National Academies Press, Washington, DC, 2000).
43 N. J. Gemmell *et al.*, Ngatiwai Trust Board, The tuatara genome reveals ancient features of amniote evolution. *Nature* **584**, 403–409 (2020).
44 V. Hayes, Indigenous genomics. *Science* **332**, 639 (2011).
45 S. G. Newmaster, S. Ragupathy, Ethnobotany genomics—Discovery and innovation in a new era of exploratory research. *J. Ethnobiol. Ethnomed.* **6**, 2 (2010).
46 B. L. Sullivan *et al.*, eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142**, 2282–2292 (2009).

47 J. E. McCormack, A. J. Zellmer, L. L. Knowles, Does niche divergence accompany allopatric divergence in *Aphelocoma jays* as predicted under ecological speciation? Insights from tests with niche models. *Evolution* **64**, 1231–1244 (2010).
48 S. F. Ryan *et al.*, Global invasion history of the agricultural pest butterfly *Pieris rapae* revealed with genomics and citizen science. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 20015–20024 (2019).
49 S. G. Schuttler *et al.*, Citizen science in schools: Students collect valuable mammal data for science, conservation, and community engagement. *Bioscience* **69**, 69–79 (2018).
50 A. M. Mc Cartney *et al.*, Balancing openness with Indigenous data sovereignty: An opportunity to leave no one behind in the journey to sequence all of life. *Proc. Natl. Acad. Sci. U.S.A.*, e2115860119 (2022).
51 R. Cook-Deegan, R. A. Ankeny, K. Maxson Jones, Sharing data to build a medical information commons: From Bermuda to the global alliance. *Annu. Rev. Genomics Hum. Genet.* **18**, 389–415 (2017).
52 W. L. Kraus, Editorial: Would you like a hypothesis with those data? Omics and the age of discovery science. *Mol. Endocrinol.* **29**, 1531–1534 (2015).
53 R. A. Gibbs, The Human Genome Project changed everything. *Nat. Rev. Genet.* **21**, 575–576 (2020).
54 F. C. P. Navarro *et al.*, Genomics and data science: An application within an umbrella. *Genome Biol.* **20**, 109 (2019).
55 J. L. Edwards, Research and societal benefits of the global biodiversity information facility. *Bioscience* **54**, 485–486 (2004).
56 L. M. Page, B. J. MacFadden, J. A. Fortes, P. S. Soltis, G. Riccardi, Digitization of biodiversity collections reveals biggest data on biodiversity. *Bioscience* **65**, 841–842 (2015).
57 J. Deck *et al.*, The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biol.* **15**, e2002925 (2017).
58 R. Leinonen, H. Sugawara, M. Shumway; International Nucleotide Sequence Database Collaboration, The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
59 A. Rhie *et al.*, *Towards complete and error-free genome assemblies of all vertebrate species*. Nature **592**, 737–746 (2021).
60 M. Blaxter *et al.*, Why sequence all eukaryotes? *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115636118 (2022).
61 D. Jebb *et al.*, Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583**, 578–584 (2020).
62 G. Lunter, M. Goodson, Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
63 S. H. Martin *et al.*, Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. *Genetics* **203**, 525–541 (2016).
64 J. Prado-Martinez *et al.*, Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).
65 E. Garrison *et al.*, Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
66 D. B. Mahat *et al.*, Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).
67 L. Calviello, U. Ohler, Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. *Trends Genet.* **33**, 728–744 (2017).
68 T. S. Mikkelsen *et al.*, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
69 E. Lieberman-Aiden *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
70 M. Roller *et al.*, LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol.* **22**, 62 (2021).
71 M. Claussnitzer *et al.*, FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
72 E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
73 A. McKenna *et al.*, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
74 Y. Nishimura *et al.*, Division of functional roles for termite gut protists revealed by single-cell transcriptomes. *ISME J.* **14**, 2449–2460 (2020).
75 B. Langmead, A. Nellore, Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.* **19**, 208–219 (2018).
76 K. Karczewski, "Just dropped 12T of summary statistics today." *Twitter* (2020). https://twitter.com/konradjk/status/1272961093330239492. Accessed 18 November 2020.
77 J. Armstrong *et al.*, Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
78 S. Surbakti *et al.*, New Guinea highland wild dogs are the original New Guinea singing dogs. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 24369–24376 (2020).
79 D. A. Conde *et al.*, Zoos through the lens of the IUCN Red List: A global metapopulation approach to support conservation breeding programs. *PLoS One* **8**, e80311 (2013).
80 IUCN, The IUCN Red List of Threatened Species. Version 2019-2 (2019). https://www.iucnredlist.org/. Accessed 6 January 2021.
81 W. J. Kress *et al.*, Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115640118 (2022).
82 T. van der Valk, D. Díez-Del-Molino, T. Marques-Bonet, K. Guschanski, L. Dalén, Historical genomes reveal the genomic consequences of recent population decline in eastern gorillas. *Curr. Biol.* **29**, 165–170.e6 (2019).
83 T. van der Valk *et al.*, Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* **591**, 265–269 (2021).
84 N. Rybczynski *et al.*, Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nat. Commun.* **4**, 1550 (2013).
85 G. Moore, M. Tessler, S. W. Cunningham, J. Betancourt, R. Harbert, Paleo-metagenomics of North American fossil packrat middens: Past biodiversity revealed by ancient DNA. *Ecol. Evol.* **10**, 2530–2544 (2020).
86 E. E. Hahn, A. Grealy, M. Alexander, C. E. Holleley, Museum epigenomics: Charting the future by unlocking the past. *Trends Ecol. Evol.* **35**, 295–300 (2020).
87 K. L. Chiou, C. M. Bergey, Methylation-based enrichment facilitates low-cost, noninvasive genomic scale sequencing of populations from feces. *Sci. Rep.* **8**, 1975 (2018).
88 S. K. Wasser *et al.*, Conservation. Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. *Science* **349**, 84–87 (2015).
89 M. Watsa, G. A. Erkenswick, A. Pomerantz, S. Prost, Portable sequencing as a teaching tool in conservation and biodiversity research. *PLoS Biol.* **18**, e3000667 (2020).
90 A. Pomerantz *et al.*, Real-time DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **7**, giy033 (2018).
91 H. Krehenwinkel *et al.*, Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *Gigascience* **8**, giz006 (2019).
92 D. Reich *et al.*, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
93 D. C. Card, B. Shapiro, G. Giribet, C. Moritz, S. V. Edwards, Museum genomics. *Annu. Rev. Genet.* **55**, 633–659 (2021).
94 P. D. Heintzman *et al.*, A new genus of horse from Pleistocene North America. *eLife* **6**, e2994 (2017).
95 J. Freedman, L. B. van Dorp, S. Brace, Destructive sampling natural science collections: An overview for museum professionals and researchers. *Journal of Natural Science Collections* **5**, 21–34 (2018).
96 H. Escobar, In a 'foretold tragedy,' fire consumes Brazil museum. *Science* **361**, 960 (2018).
97 M. Meyer *et al.*, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
98 A. H. Pálsdóttir, A. Bläuer, E. Rannamäe, S. Boessenkool, J. H. Hallsson, Not a limitless resource: Ethics and guidelines for destructive sampling of archaeofaunal remains. *R. Soc. Open Sci.* **6**, 191059 (2019).
99 D. Houle, D. R. Govindaraju, S. Omholt, Phenomics: The next challenge. *Nat. Rev. Genet.* **11**, 855–866 (2010).
100 M. Hiller *et al.*, A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep.* **2**, 817–823 (2012).

**101** T. F. Meehan *et al.*, International Mouse Phenotyping Consortium, Disease model discovery from 3,328 gene knockouts by the International Mouse Phenotyping Consortium. *Nat. Genet.* **49**, 1231–1238 (2017).
**102** G. Li *et al.*, The sequences of 1504 mutants in the model rice variety Kitaake facilitate rapid functional genomic studies. *Plant Cell* **29**, 1218–1231 (2017).
**103** W. Zhang *et al.*, Multiplex precise base editing in cynomolgus monkeys. *Nat. Commun.* **11**, 2325 (2020).
**104** R. Brosh *et al.*, A versatile platform for locus-scale genome rewriting and verification. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023952118 (2021).
**105** F. Cheng *et al.*, Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. *Nat. Genet.* **48**, 1218–1224 (2016).
**106** K. G. Ota, G. Abe, Goldfish morphology as a model for evolutionary developmental biology. *Wiley Interdiscip. Rev. Dev. Biol.* **5**, 272–295 (2016).
**107** A. R. Boyko *et al.*, A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol.* **8**, e1000451 (2010).
**108** J.-L. Legras *et al.*, Adaptation of *S. cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Mol. Biol. Evol.* **35**, 1712–1727 (2018).
**109** R. L. Witter, The changing landscape of Marek's disease. *Avian Pathol.* **27**, S46–S53 (1998).
**110** R. I. Bailey *et al.*, Pathogen transmission from vaccinated hosts can cause dose-dependent reduction in virulence. *PLoS Biol.* **18**, e3000619 (2020).
**111** M. R. Robinson, N. R. Wray, P. M. Visscher, Explaining additional genetic variation in complex traits. *Trends Genet.* **30**, 124–132 (2014).
**112** N. Gorelick *et al.*, Google Earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).
**113** M. Lin *et al.*, Landscape analyses using eDNA metabarcoding and Earth observation predict community biodiversity in California. *Ecol. Appl.* **31**, e02379 (2021).
**114** G. Chauhan *et al.*, Combining citizen science and genomics to investigate tick, pathogen, and commensal microbiome at single-tick resolution. *Front. Genet.* **10**, 1322 (2020).
**115** J. E. Lemieux *et al.*, Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, eabe3261 (2021).
**116** W. Van Breedam *et al.*, Porcine reproductive and respiratory syndrome virus entry into the porcine macrophage. *J. Gen. Virol.* **91**, 1659–1667 (2010).
**117** K. M. Whitworth *et al.*, Gene-edited pigs are protected from porcine reproductive and respiratory syndrome virus. *Nat. Biotechnol.* **34**, 20–22 (2016).
**118** P. H. Rampelotto, Extremophiles and extreme environments. *Life (Basel)* **3**, 482–485 (2013).
**119** R. C. Neves, L. K. B. Hvidepil, T. L. Sørensen-Hygum, R. M. Stuart, N. Møbjerg, Thermotolerance experiments on active and desiccated states of *Ramazzottius varieornatus* emphasize that tardigrades are sensitive to high temperatures. *Sci. Rep.* **10**, 94 (2020).
**120** M. A. Mapalo, N. Robin, B. E. Boudinot, J. Ortega-Hernández, P. Barden, A tardigrade in Dominican amber. *Proc. Biol. Sci.* **288**, 20211760 (2021).
**121** E. Kiss *et al.*, Comparative genomics reveals the origin of fungal hyphae and multicellularity. *Nat. Commun.* **10**, 4080 (2019).
**122** C. M. Beall *et al.*, Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 11459–11464 (2010).
**123** X. Gou *et al.*, Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Res.* **24**, 1308–1315 (2014).
**124** X. Liu *et al.*, EPAS1 gain-of-function mutation contributes to high-altitude adaptation in Tibetan horses. *Mol. Biol. Evol.* **36**, 2591–2603 (2019).
**125** R. M. Schweizer *et al.*, Physiological and genomic evidence that selection on the transcription factor Epas1 has altered cardiovascular function in high-altitude deer mice. *PLoS Genet.* **15**, e1008420 (2019).
**126** A. M. Graham, K. G. McCracken, Convergent evolution on the hypoxia-inducible factor (HIF) pathway genes EGLN1 and EPAS1 in high-altitude ducks. *Heredity* **122**, 819–832 (2019).
**127** V. B. Indjeian *et al.*, Evolving new skeletal traits by *cis*-regulatory changes in bone morphogenetic proteins. *Cell* **164**, 45–56 (2016).
**128** C. Darwin, *Life and Letters of Charles Darwin* (John Murray, London, 1887), vol. 1.
**129** B. Knapp *et al.*, Ten simple rules for a successful cross-disciplinary collaboration. *PLoS Comput. Biol.* **11**, e1004214 (2015).
**130** E. R. Ellwood, A. Monfils, L. White, D. Linton, N. Douglas, Developing a data-literate workforce through BLUE: Biodiversity literacy in undergraduate education. *Bioinformation Science and Standards* **3**, e37339 (2019).
**131** T. K. Teal *et al.*, Data carpentry: Workshops to increase data literacy for researchers. *Int. J. Digit. Curation* **10**, 135–143 (2015).
**132** R. R. Fitak *et al.*, The expectations and challenges of wildlife disease research in the era of genomics: Forecasting with a horizon scan-like exercise. *J. Hered.* **110**, 261–274 (2019).
**133** National Academies of Sciences, Engineering, and Medicine, Division of Behavioral and Social Sciences and Education, Committee on the Science of Science Communication: A Research Agenda, *Communicating Science Effectively: A Research Agenda* (National Academies Press, 2017).
**134** K. Hinde *et al.*, March mammal madness and the power of narrative in science outreach. *eLife* **10**, e65066 (2021).
**135** M. D. Wilkinson *et al.*, The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
**136** J. E. Moore *et al.*, ENCODE Project Consortium, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
**137** L. Andersson *et al.*, FAANG Consortium, Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **16**, 57 (2015).
**138** Alliance of Genome Resources Consortium, Alliance of Genome Resources portal: Unified model organism research platform. *Nucleic Acids Res.* **48**, D650–D658 (2020).
**139** D. Field *et al.*, The genomic standards consortium. *PLoS Biol.* **9**, e1001088 (2011).
**140** National Academies of Sciences, Engineering, and Medicine, Division on Earth and Life Studies, Board on Life Sciences, Committee on Biological Collections: Their Past, Present, and Future Contributions and Options for Sustaining Them, *Biological Collections: Ensuring Critical Research and Education for the 21st Century* (National Academies Press, 2021).
**141** H. Frangoul *et al.*, CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. *N. Engl. J. Med.* **384**, 252–260 (2021).
**142** E. S. Lander *et al.*, Adopt a moratorium on heritable genome editing. *Nature* **567**, 165–168 (2019).
**143** F. D. Steinheimer, Charles Darwin's bird collection and ornithological knowledge during the voyage of HMS "Beagle", 1831–1836. *J. Ornithol.* **145**, 300–320 (2004).
**144** J. S. Sherkow *et al.*, Ethical, legal, and social issues in the Earth BioGenome Project (2021). https://doi.org/10.2139/ssrn.3840280. Accessed 13 November 2021.