Darwinian genomics and diversity in the tree of life

Taylorlyn Stephen,a,1 Shawn M. Burgess,b, Hans Cheng,b, Charles G. Danko,c, Clare A. Gill,d, Erich D. Jarvis,e,f, Klaus-Peter Koepfi,g,h, James E. Kolte,i, Eric Lyons,j, Pamela Ronald,k,l,m,n, Oliver A. Ryder,o,p, Lynn M. Schrimlq, Pamela Softs,2 Sue VandeWoude,q, Huaijun Zhou,1 Elaine A. Ostrandera,q and Elinor K. Karlssonu,v,w,2

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 discovering 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.

Keywords: 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...
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 development of 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 years 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).

Stephan et al.
Darwinian genomics and diversity in the tree of life

https://doi.org/10.1073/pnas.2115644119
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 (DiGibio) 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 inflamasome, 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

---

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

In August 2019, three funding agencies—the National Human Genome Research Institute (NHGRI), 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.
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 Miocene (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 private 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 paleopropionomics.

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 samples by millions (humans). “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.

Humans

<table>
<thead>
<tr>
<th>Function</th>
<th>Diversity</th>
<th>Complexity</th>
<th>Phenotyping</th>
<th>Sampling</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★★★★★★☆</td>
<td>★★★★★★☆</td>
<td>★★★★★★☆</td>
<td>★★★★★★☆</td>
<td>★★★★★★☆</td>
</tr>
</tbody>
</table>

Traditional / Laboratory models

<table>
<thead>
<tr>
<th>Function</th>
<th>Diversity</th>
<th>Complexity</th>
<th>Phenotyping</th>
<th>Sampling</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
</tr>
</tbody>
</table>

Companion, captive or controlled animals

<table>
<thead>
<tr>
<th>Function</th>
<th>Diversity</th>
<th>Complexity</th>
<th>Phenotyping</th>
<th>Sampling</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
</tr>
</tbody>
</table>

Commercial livestock animals

<table>
<thead>
<tr>
<th>Function</th>
<th>Diversity</th>
<th>Complexity</th>
<th>Phenotyping</th>
<th>Sampling</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
</tr>
</tbody>
</table>

Domesticated plants / agricultural species

<table>
<thead>
<tr>
<th>Function</th>
<th>Diversity</th>
<th>Complexity</th>
<th>Phenotyping</th>
<th>Sampling</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
</tr>
</tbody>
</table>

Non-model organisms

<table>
<thead>
<tr>
<th>Function</th>
<th>Diversity</th>
<th>Complexity</th>
<th>Phenotyping</th>
<th>Sampling</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
<td>★★★★★☆☆</td>
</tr>
</tbody>
</table>

(lower bound) ★possible range (rare or absent) (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 yeasts 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 “thought experiment with Darwin’s finches” (143). Imagine we could do so for every generation since the 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.

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.

Finally, more support for museums, including zoos, aquariums, 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, Lorjeta 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-AR0-2019-67015-29490, and NIH GM122968.