6-3-2020

Deploying Big Data To Crack The Genotype To Phenotype Code

E. L. Westerman

S. E. J. Bowman

Bradley Justin Davidson , '91
Swarthmore College, bdavids1@swarthmore.edu

See next page for additional authors

Recommended Citation

E. L. Westerman; S. E. J. Bowman; Bradley Justin Davidson , '91; M. C. Davis; E. R. Larson; and C. Sanford. (2020). "Deploying Big Data To Crack The Genotype To Phenotype Code". *Integrative And Comparative Biology.* DOI: 10.1093/icb/icaa055
https://works.swarthmore.edu/fac-biology/603

Title: Deploying Big Data to Crack the Genotype to Phenotype Code

Running title: Cracking the Genotype to Phenotype Code

Erica L. Westerman[1*], Sarah E.J. Bowman[2#], Bradley Davidson[3#], Marcus C. Davis[4#], Eric R. Larson[5#], Christopher Sanford[6#]

[1]Department of Biological Sciences, University of Arkansas, Fayetteville, AR 72701
[2]Hauptman-Woodward Medical Research Institute, Buffalo, NY 14203
[3]Department of Biology, Swarthmore College, Swarthmore, PA 19081
[4]Department of Biology, James Madison University, Harrisonburg, VA 22807
[5]Department of Natural Resources and Environmental Sciences, University of Illinois, Urbana, IL 61801
[6]Department of Ecology, Evolution and Organismal Biology, Kennesaw State University, Kennesaw, GA 30144

*Corresponding Author
#Author order is alphabetical, all authors contributed equally

Corresponding Author Information: Erica L. Westerman
Department of Biological Sciences, University of Arkansas, Fayetteville, AR 72701
ewesterm@uark.edu
phone: 479-575-5348
fax: 479-575-4010

Total number of words: 5,214

Symposium Article

Key words: machine learning, genomics, integrative biology, big data

Abstract:

Mechanistically connecting genotypes to phenotypes is a longstanding and central mission of biology. Deciphering these connections will unite questions and datasets across all scales from molecules to ecosystems. Although high-throughput sequencing has provided a rich platform on which to launch this effort, tools for deciphering mechanisms further along the genome to phenome pipeline remain limited. Machine learning approaches and other emerging computational tools hold the promise of augmenting human efforts to overcome these obstacles. This vision paper is the result of a Reintegrating Biology Workshop, bringing together the perspectives of integrative and comparative biologists to survey challenges and opportunities in cracking the genotype to phenotype code and thereby generating predictive frameworks across biological scales. Key recommendations include: promoting the development of minimum “best practices” for the experimental design and collection of data; fostering sustained and long-term data repositories; promoting programs that recruit, train, and retain a diversity of talent and providing funding to effectively support these highly cross-disciplinary efforts. We follow this discussion by highlighting a few specific transformative research opportunities that will be advanced by these efforts.

Introduction:

Deciphering the mechanisms by which genotypes generate phenotypes is a central mission of biology. Historically this mission was hampered by a lack of sequence and expression data. Now, we are hindered by the daunting task of integrating large amounts of disparate data across multiple areas of expertise. Fully realizing these mechanisms will facilitate the integration of enormous datasets in organismal diversity research across molecular, morphological, behavioral, and ecosystem scales (Figure 1). Comprehensive, multi-scale data integration will impact broad reaching, interdisciplinary and integrative goals across biological disciplines (NRC, 2009). Although not a comprehensive list, some of these programs and goals include: 1) understanding the rules for signaling; 2) deciphering mechanisms underlying robustness and resilience; 3) predicting and ameliorating the impact of anthropogenic change to preserve biodiversity and ecosystem services; 4) integrating data across scale; 5) promoting proactive and personalized medicine designed around wellness instead of treating disease; and 6) effective deployment of synthetic biology approaches for health, energy, and environmental remediation applications.

While this unification of datasets has long been the goal of researchers, only now in the big data era are tools emerging that hold promise to augment human efforts (Camacho et al., 2018). Machine learning approaches now demonstrate their ability to make connections and find patterns at a pace that better aligns with the exponentially increasing rates of data collection. To fully exploit these advancements, the biological research community will need to invest significant resources towards 1) the development of data collection and storage standards; 2) the development of tools to overcome key bottlenecks in data acquisition and analysis; and 3) training initiatives and collaborative outreach to a diverse pool of existing and emerging talent (Hulsen et al., 2019). Here we discuss how sustained efforts in these areas can further catalyze biology's big data era for cracking the genotype to phenotype code. We follow this discussion

by highlighting a few specific transformative research opportunities that will be advanced by these efforts.

Context: This vision paper resulted from the authors participation in the Reintegrating Biology Jumpstart Workshop held in Austin, Texas in December of 2019. Reintegrating Biology is funded by the National Science Foundation through a grant to the University Corporation for Atmospheric Research (UCAR) and Knowinnovation (KI). The workshop consisted of virtual town halls, microlabs and jumpstart meetings designed to engage diverse participation from across the biological community. The primary objectives of Reintegrating Biology were to solicit input on key challenges and exciting opportunities on both long-standing and emerging biological research directions. After evaluating roughly 50 different proposals generated during the initial phase of the workshop, our working group (a structural biologist, an ecologist, a biomechanist, an integrative animal behaviorist, and two evolutionary developmental biologists) coalesced around the challenge of utilizing big data to elucidate the genome to phenome discovery pathway. This vision paper is the result of on-site discussions and group writing sessions followed by off-site collaborative writing. Our paper complements the focus of the ICB Building Bridges Special Issue on questions that cross multiple scales from molecules to whole organisms while offering the perspective of a suite of integrative and comparative biologists.

Challenges and their solutions:

Effective deployment of high throughput data to decode genotype to phenotype mechanisms will require extensive modification and resource allocation all along the pipeline from data collection to publication and storage. In this section, we provide an overview of some key challenges and potential solutions. These approaches align with the principles of making data Findable, Accessible, Interoperable and Reusable (FAIR) (Wilkinson et al., 2016).

1. Data collection and quality.

For a tool or repository to be useful there needs to be community defined and driven standards regarding experimental design, data collection and annotation. One example regards proper alignment of sequencing techniques (RadSeq, SNP arrays or whole genome resequencing) to match specific research goals. Whole genome resequencing provides complete genomic data for relatively few individuals, making it optimal for gene discovery (Xu & Bai, 2015). In contrast, because RadSeq and SNP arrays exploit widely spaced markers, they can be used to characterize and compare relatively large numbers of individuals, but with less information for each sample (Tam et al., 2019). Thus these techniques are optimal for high throughput characterization of populations. Additionally, experimental design tools, such as GWAPower (Feng et al., 2011), can be used to facilitate optimal selection of sample sizes and sequencing type (average distance between SNP and candidate gene) for new candidate gene identification projects. Another major difficulty involves proper annotation of functionally characterized genes. Many gene products are highly pleiotropic, shifting their function in a context-dependent manner within the same organism (Wagner & Zhang, 2011). Furthermore, orthologous genes are frequently re-deployed, leading to highly variable functions in different organisms. Overcoming these challenges will require a highly versatile system of annotation that can encompass functional variability without compromising overall utility.

2. Data storage.

Currently the tools and data associated with high throughput sequencing are inaccessible and unstable (i.e., often poorly maintained due to lack of support). The National Center for Biotechnology Information, NCBI (https://www.ncbi.nlm.nih.gov), is an excellent, supported repository for genomic and transcriptomic data (Coordinators, 2020), however its design for biomedical research makes it somewhat limited for non-model organisms. While international resources designed specifically for housing non-model organism genomic and transcriptomic

data and tools, such as Lepbase (http://lepbase.org), can suffer from limited financial support. Solutions will involve creating centralized, community edited (possibly open source), sustainable data and tool repositories (potentially modeled on ImageJ or the Brain Initiative). Additionally, robust infrastructures must be in place to maintain and oversee these repositories as they expand. There are already systems in place (da Veiga Leprevost et al., 2017; Barnett et al., 2019) from which we can learn best practices. Finally, establishing close links between research groups collecting and research groups analyzing data will be essential. A "hub and spoke" approach, as exemplified by data coordination centers used extensively in clinical research, may be an efficient model to foster this, with constant feedback from all stakeholders and advisory groups.

3. Data transparency.

Methods for data collection, management, and analysis are often opaque, making it difficult to critically evaluate datasets or efficiently redeploy them in different contexts. Agreement across fields on proper annotation of methodology, data and metadata could help overcome this issue. Data Carpentry (https://datacarpentry.org/semester-biology/syllabus/) may provide a framework to teach standard methods for data collection and management across biology. Data analysis and development platforms such as github (https://github.com) can also be excellent public repositories for code, as they allow others to easily recreate analyses and results. While open data challenges, such as Critical Assessment of Massive Data Analysis (CAMDA), Critical Assessment of Genome Interpretation (CAGI), and DREAM Challenges can facilitate opportunities for researchers to compare analysis methods and develop best practices.

4. Data sets are often incomplete.

Next generation sequencing is poised to promote the comprehensive collection of genomic data along with transcriptional and chromatin dynamics in organisms, tissues and cells. High

throughput mass spectrometry will allow comprehensive profiling of protein expression. Advances in imaging will enable pervasive characterization of cellular, organismal and population level phenotypes. For example, high throughput imaging approaches applied to classifying interspecific diversity (Lytle et al., 2010; Valan et al., 2019) may be improved over time to characterize intraspecific morphological or phenotypic diversity linked to genotypes. Tool development must keep pace with these technologies in order to provide efficient high throughput solutions for gathering and analyzing data at critical bottlenecks. These bottlenecks include candidate gene identification, mapping connections in gene regulatory and protein interaction networks, precise quantification of relevant biochemical processes such as signaling ligand diffusion, phenotypic profiling and mapping cross-species interactions. To fill in these gaps, new tools and data-collection efforts must be promoted, perhaps in the model of some existing “big science” initiatives from ecology or ecosystem science like the Long Term Ecological Research network (LTER; Hobbie et al., 2003), the National Ecological Observatory Network (NEON; Barnett et al., 2019), or hypothesis-driven coordinated distributed experiments (Fraser et al., 2012). Research that seeks to bridge the difficult divide from genotype to phenotype *in situ* and for non-model organisms may be more successful if leveraging data and researcher expertise from well-studied “model” ecosystems.

5. Data is exponentially increasing, unwieldy and noisy.

The potential benefits of high-throughput sequencing data and other large datasets are greatly limited by inherent difficulties in extracting signal from noise. Machine learning approaches could be employed as a possible solution. We define machine learning as the science of training a model from data, enabling the machine to perform specific tasks and generate predictions (Camacho et al., 2018). The development of such machine learning tools will require a highly interdisciplinary approach, engaging computer scientists, mathematicians, and teams of biologists with wide-ranging expertise.

6. Existing tools are often limited in applicability.

It is essential to provide resources and motivation to modify tools so they are more generally applicable. Decreasing barriers and increasing accessibility to tools and databases will provide resources to a broader user base that may not have developer or technical expertise. One component of increasing tool applicability is development of clearly defined and annotated instructions regarding the types of data taken as inputs, definitions of parameters (and how they can be tuned), the assumptions underlying the algorithms, and what is generated as output. This will often be most readily achieved by providing, along with the tool, a use-case, sample data, or vignette to serve as a tutorial for use and to exemplify performance.

7. Biologists using big data should share best practices across subdisciplines.

Scientists may apply machine learning approaches to resolve big data questions across the biological sciences, from mapping genotypes to phenotypes or structure to function, to predicting relationships between the distribution of species and their environments (Olden et al., 2008). These subfields working independently likely encounter some of the same challenges in applying and interpreting machine learning approaches: Are the big data sources we use reliable and well-maintained (Barnes et al., 2014)? Do machine learning predictions have mechanistic meaning, or are they occasionally over-fitting to noise (Walsh et al., 2016)? Biologists applying machine learning to big data questions across levels of the biological hierarchy might share experiences on best practices or discoveries, while also being aware that conventional statistical approaches can be more appropriate and interpretable for some purposes (Royle et al., 2012; Rudin, 2019). Shared challenges may include identifying cases of model over-fitting (Okser et al., 2014), improving interpretability of “black box” machine learning output (Olden and Jackson, 2002), quantifying or identifying uncertainty in predictions (Willcock et al., 2018), and sharing practices for independence of training and testing data (Kegerreis et

al., 2019). Biologists undoubtedly would benefit from more interactions with computer scientists and mathematicians in these fields, but may also have high potential to learn from innovations or experiences in other fields in biology using similar tools. For example, do ecologist's concerns about the independence of machine learning testing and training data, and associated implications for model transferability or generalizability (Wenger & Olden, 2012; Bahn & McGill, 2013), relate to machine learning in other fields of biology (Walsh et al., 2016; Kegerreis et al., 2019)? Further, machine learning results are often not reincorporated into subsequent models. Solutions would involve providing efficient avenues for scientists to identify models that are relevant to their data sets and vice versa and motivate them to incorporate relevant data.

Exciting Opportunities:

Here we detail a few exciting research opportunities across molecular, morphological, behavioral, and ecosystem scales that will be advanced by sustained big data and machine learning approaches.

Using big data to solve problems in molecular structure. A key part of solving the genotype to phenotype code is investigation of molecular structure, especially developing a better understanding of structural dynamics. Biomolecular structures are often envisioned as static; we generate structural maps from snapshots of biomolecules in specific conditions. We know, however, that the biochemical reactions that occur at a molecular level are dynamic. Parameters of a protein's environment (pH, temperature, physical location in the cell, presence/absence of binding partners, signaling molecules or ligands) can influence the fold and function of a protein. Similarly, RNA molecules can have different secondary structure folds despite the same nucleotide sequence. These dynamic modulations in structure can impact function and generate phenotypic changes at the cellular or organismal level (Nussinov et al., 2019). A fundamental problem is that while we are interested in generating movies of the

molecular machinery in action, we typically cannot access these ensemble dynamics. The predominant method used to investigate molecular structure is X-ray crystallography, which accounts for ~90% of the structural models available. These structures form the basis for generating questions about models for ligand binding, protein folding, and enzymatic function. These methods, however, depend on crystallizing the biomolecule, which necessitates finding chemical conditions in which a biomolecule will crystallize; this is a fundamental bottleneck in structural biology experiments, limiting our ability to structurally explore the dynamic ensemble of protein functional space. While advances have recently been made in developing a convolutional neural network to classify crystallization outcomes (Bruno et al., 2018), we currently have no working models for predicting what conditions will generate a crystal despite extensive attempts to use information about genetic sequences, homology modeling, and biomolecular parameter space to make predictions (Abrahams & Newman, 2019; Lynch, et al., 2020). Leveraging a big data and machine learning framework of data organization and annotation coupled with developing accessible repositories for full experimental details (including what doesn't work) and tools for using these data is critical for making predictive models. These big data approaches to molecular structural biology questions would enable a fuller exploration of the dynamics of protein function.

Comprehensive mapping and analysis of gene regulatory networks. The developmental processes that generate diverse phenotypes (morphological, physiological and behavioral) are largely encoded by densely interconnected gene networks (Davidson & Erwin, 2006). Next generation sequencing is poised to identify nearly all of the components in these networks (coding genes, non-coding regulatory elements and associated chromatin states) in a wide range of organisms and cell types (Banf & Rhee, 2017; Das Gupta & Tsiantis, 2018; Lowe et al., 2017; Rebeiz & Tsiantis, 2017). However, we currently cannot leverage these sequencing data to accurately map the regulatory connections that link these elements in a high throughput

manner (Thompson et al., 2015; Fiers et al., 2018; Skinnider et al., 2019; Siahpirani et al., 2019; Huynh-Thu & Sanguinetti, 2019). Network mapping is particularly critical for efforts to characterize dynamic shifts in gene network connections that drive the temporal unfolding of developing patterning programs and mediate environmentally dependent variability in morphology or physiology. These mapping efforts may be supplemented by single cell sequencing, which can be used to enhance our understanding of cell fate and to connect transcriptional and epigenetic heterogeneity (Grün & Grün, 2020; Angermueller et al., 2016). Mapping will also facilitate characterization of key differences in network architecture or dynamics that generate diverse phenotypes at various biological scales from cells to super-organisms (Rebeiz et al., 2015). Additionally, mapping can promote characterization of genetically encoded intra and inter-specific interactions particularly within holobiont communities including microbe/metazoan, symbiotic or parasite/host interactions (Ferreiro et al., 2018). Mapping will also provide a productive framework for comparative approaches or targeted perturbations (CRISPR) used to test hypotheses regarding fundamental structure/function questions. In particular, these approaches can be used to elucidate architectural features or modules that are targeted by selection to produce novel phenotypes (Rebeiz et al., 2015; Nocedal & Johnson, 2015). These maps can also be used to identify key differences within heterologous cell populations within an individual that are associated with disease states (Chiquet et al., 2019). Broad characterization of these functionally critical network features or modules can then be used to search for shared properties which may facilitate predictive models or formulation of underlying principles. It is also possible that tools used to map or analyze gene network connections can be deployed in relation to other biological networks at different scales and thus exploit other poorly utilized data repositories (Yan et al., 2016).

A deep learning approach to gene expression analysis. In the continued aim to “reverse engineer” the gene regulatory networks (GRN) that generate organismal diversity (Cussat-Blanc

et al., 2019), researchers produce vast amounts of gene expression data. The literature is full of microscopy generated images of *in situ* hybridization assays for genes of interest, in both wild-type and experimental systems, across an ever-expanding range of organisms (Puniyani & Xing, 2013; Davis, 2013; Wu et al., 2016). *In situ* hybridization assays are open to subjective interpretation (Yang et al., 2019), and expression domain similarities, differences, and/or variation are rarely quantified within or across datasets (see excellent exceptions such as Mace et al., 2010; Patrushev et al., 2018). With the more recent practical availability of RNA-seq, tissue and cellular level transcriptomics provides a more quantitative approach for testing hypotheses about gene-gene interactions. However, transcriptomics has in no way replaced *in situ* assays. In particular, *in situ* assays validate expression data for genes of interest identified in high-throughput transcriptomic analysis. *In situ* assays also provide essential details on spatial and temporal patterns of expression. In well-characterized systems such as *Drosophila*, we are beginning to see intriguing synergies of these datasets (Karaiskos et al., 2017). How then do we continue this trend and bring a community's concerted efforts into a data pool to fully leverage the goal of understanding gene-gene interactions across organismal diversity?

Some deep learning approaches are demonstrating impressive abilities to recognize patterns within and between large datasets, and to make connections between visual and molecular datasets (Mobadersany et al., 2018). Deep learning algorithms are networked computational models that mimic the layered node-like, neuronal structure of organic brains (Goodfellow et al., 2016). Early variants of these algorithms relied on heavy processing of data before it went into the model in order for results to be meaningful. However, as big data gets even bigger, continued improvements in these algorithms have led to autonomous learning, in which the model itself is capable of finding meaningful patterns in the data (Webb, 2018). In particular, convoluted neural networks (CNNs), the form of deep neural networks behind rapid advancements in computer vision (Khan et al., 2018), hold significant promise for biology. CNNs

are ideal for processing data with two or more dimensions, such as -omics or image datasets (Camacho et al., 2018). Some of the ways in which CNN algorithms can be employed are already emerging, as recent studies yield promising returns in the automatic detection of positive *in situ* staining results (Dong et al., 2015), the screening of developmental stages and phenotypes (Ishaq et al., 2017; Cordero-Maldonado et al., 2019), the construction of "*in silico embryos"* (Shen et al., 2018) and the generation of GRN predictive models using expression data (Yang et al., 2019). Sustained progress in these areas will require community initiatives that 1) promote tool/algorithm development and sharing; and 2) foster long-term pan-taxa repositories for gene expression and associated transcriptomic datasets.

Characterizing complex phenotypes. If a phenotype is broadly defined, or influenced by a large number of genes with small effects, it can be incredibly difficult to have enough power to identify all, or any, of the genes involved.

One way to circumvent this problem is to design experiments that allow for the detection of genes associated with relatively simple, specific aspects of biologically relevant complex phenotypes. For example, if interested in the genes underlying mate selection, identify the genes associated with visual preference, olfactory preference, or vibratory preference separately, instead of searching for genes associated with a broadly characterized mate preference (Figure 2). This technique, of distilling complex phenotypes down to many simple phenotypes, has proven fruitful for identifying genes associated with behavior in model animals. The genes of large effect for conspecific pheromone detection and olfactory mate preference in *Drosophila* were identified by testing individual responses to the specific components of conspecific and heterospecific pheromones, with an array of gene knock-out lines (Xu et al., 2005; Datta et al, 2008; Jin et al., 2008; Billeter & Levine, 2013). The different genes associated with positive and negative memory formation in *Drosophila* fruit flies and *Aedes aegypti*

mosquitoes were identified using highly controlled experiments, with well-defined phenotypes (Schwaerzel et al., 2003, Vinauger et al., 2018). The same can be said for the genes involved in the circadian clock, and for genes involved in sex-specific responses to male pheromones in *Drosophila* (Curtin et al., 1995; Suri et al., 1999; Sarov-Blat et al., 2000; Demir & Dickson, 2005; Drapeau et al., 2006; Ruta et al., 2010). Indeed, the reason we know so much about the genetics of mate preference in *Drosophila* is because of decades of phenotype dissection and careful study of specific elements of the mate selection process (Keene & Waddell, 2007; Dickson, 2008). As illustrated by *Drosophila* mate preference genetics, these narrowly defined phenotypes are often the building blocks of the larger phenotype of interest, and once characterized, may scale up. While one can argue that this approach will only work for model animals, recent work on butterfly wing pattern genetics suggests otherwise. This approach has proven extremely fruitful for the identification of genes controlling specific color patterning elements of complex butterfly wing patterns in wild populations (Reed et al., 2011; Martin et al, 2012; Kronforst & Papa 2015; Nadeau et al., 2016; Westerman et al., 2018), and may also prove useful for identifying genes associated with other complex traits in wild populations, such as habitat selection and mate choice.

Identifying the genetic basis of behavior. One of the major hurdles of behavioral ecology has been identifying the genetic basis of evolutionary and ecologically important behaviors. Scientists have spent decades carefully characterizing a vast array of behaviors using ethograms, from foraging to mating to habitat selection, in a wide range of species. These carefully characterized phenotypes are ripe for genotype-phenotype discovery, and the last decade has seen an uptick in behavioral genetics studies in a range of taxa and natural and semi-natural populations (Bubac et al., 2020). Importantly, the ecological and evolutionary underpinnings of these phenotypes are often known, so identifying the genetic basis of these traits will facilitate a dramatic advance in our understanding of how selective forces on whole

organisms translates to genomic change (as discussed in Bengston et al., 2018; Merlin & Liedvogel, 2019; Westerman, 2019). Additionally, many of the scientists studying these well-characterized behavioral traits are familiar enough with their study system that they can identify the most interesting and most accessible traits for gene identification. This drops the number of individuals that need to be sequenced for high quality candidate gene identification from the thousands needed in model animals and human populations to 70-120 individuals. This is primarily because we are looking for new genes of large effect in non-model animals (e.g. Westerman et al., 2018) instead of for new genes of small effect (which is what we are looking for in model animals and humans, e.g. Agrawal et al., 2016). These new genes of large effect are likely to be most relevant and tractable for management of responses to global change for non-model organisms (below). The genomic and translational tools necessary for identifying the genes underlying these behaviors now exist (Bentley, 2006; Visscher et al., 2012; Ran et al., 2013), and are starting to be incorporated into the study of behavior in a small set of species (Bubac et al., 2020). The challenge is to integrate genomic, proteomic, and network approaches (and scientists) more broadly into the study of behavior, and to expose data scientists to the wealth of behavioral phenotypic data and associated behavioral ecologists that can be utilized in our efforts to better understand the genotype to phenotype pathway.

Improved predictions for global change. Bridging the genotype to phenotype divide has high potential to improve management of species, communities, and ecosystems in response to global change challenges (climate, land use, invasive species), whether human-managed (e.g., agriculture; Abberton et al., 2016) or natural (e.g., endangered species, protected areas; Hoffman et al., 2015). Importantly, data deficiencies and uncertainties are likely to be most severe for wild species or remote ecosystems, relative to those upon which human societies are more dependent (Bland et al., 2015; Donaldson et al., 2016). As knowledge of genotypes has outpaced knowledge of phenotypes, researchers have called for high throughput phenotyping to

keep pace with genomic data (Kültz et al., 2013). Both phenotype and genotype data is urgently needed to guide adaptation and mitigation of global change effects on species and ecosystems. For example, current correlation-based predictions of species responses to global change (i.e., relating presence or distributions to environment conditions) inaccurately predict these relationships because they: 1) lack mechanism; 2) ignore biotic interactions; 3) omit potential for evolutionary response to change (Urban et al., 2016). Big data (both genetic and phenotypic) can improve these predictions by improving our understanding of organismal physiology, dispersal ability, or evolutionary potential. Phenotypic big data is being generated and improved through trait databases like TRY (Kattge et al. 2011), FishTraits (Frimpong & Angermeier 2009), and others. However, some specific data priorities to improve predictions of species, community, and ecosystem response to global change (land use, climate, invasive species) include: thermal, desiccation, and chemical tolerances; body mass; water and light requirements; life history traits; trophic position or diet; seed or larval size or dispersal traits; intra- and inter-specific interactions (mediated by behavior); and evolutionary or adaptive potential (Urban et al., 2016). Ecological genetic big data (transcriptomic and genomic) are being generated in field and common garden studies of increasingly diverse taxa under different climatic regimes (Pespeni et al, 2013; Hübner et al., 2015; Smith et al., 2013; Maor-Landaw et al., 2017). Increasing the taxonomic and ecological breadth of these data will enhance our understanding of genome to phenotype while improving the predictive power of our ecological models.

Calls to reintegrate organismal biology by collecting high throughput phenotypic data to compliment high throughput genomic data (Kültz et al., 2013) can leverage management and conservation needs for some similar data to guide more mechanistic models of species responses to climate change (Urban et al., 2016). Both needs and applications share a dependency on: 1) big data (e.g., van den Hoogen et al., 2019), often analyzed by 2) machine

learning approaches (e.g. Olden et al., 2008). Researchers might leverage funding opportunities by combining basic science questions in mapping the genotype to phenotype with applied science needs for both data sources to inform conservation and management of commercially important, invasive, or endangered species in natural ecosystems. This integration of basic and applied science requires choosing which organisms provide the most return on investment for both basic and applied science questions concurrently. Further, there are too many populations, species, and ecosystems to collect genotype and phenotype data for all biological entities that need management; rather, scientists and resource managers will need to prioritize representative systems that can generalize to similar taxa or ecosystems (Urban et al., 2016) - these may not be classical model organisms, but will still be surrogates or proxies for related organisms and ecosystems (Caro & O'Doherty 1999).

Creating the human infrastructure for a big data and machine learning approach:

Ironically, leveraging big data approaches and machine learning tools to crack the genotype to phenotype code will be about supporting people. There has been recognition that lack of data science proficiency and expertise is a fundamental roadblock in scientific research (Barone et al., 2017). Currently, exciting pioneering efforts are underway - in tool and research development, and in fundamental research. However, these efforts will likely remain insular, underutilized, and unavailable to the whole community - an inequitable situation - without broader development initiatives. Systematic top-down and bottom-up support structures are needed to: 1) attract, recruit, incentivize, and train a diverse group of students to these questions, many of which may never identify as biologists (i.e. they will remain data scientists, statisticians, etc.); 2) support and retrain biologists who are interested in developing these approaches; 3) develop sustained pan-disciplinary collaborations with experts in data science, mathematics, computer science, and related fields. These sustained pan-disciplinary collaborations can be particularly fruitful for pushing the boundaries of non-model organism

research, as illustrated by recent scientific advancements using *Heliconius* butterflies, most of which were achieved via multi-year collaborations between field biologists, bioinformaticians, developmental biologists, and population geneticists (as well as other sub-disciplines) (Merrill et al., 2015; Kronforst & Papa, 2015). Addressing some of these challenges may involve the development of interdisciplinary courses, programs, and degrees along with associated outreach to community colleges or other institutions that do not currently have access to resources. Formation of interdisciplinary teams who commit to attending and hosting each other's conferences will help build common languages and interest in the key questions in their fields. Programs such as NSF's Research Coordination Network (RCN) provide support pathways for human infrastructure and workforce development to achieve this goal. Ultimately, the results of these efforts can be seen as more than a reintegration - but instead the emergence of an augmented biology.

Recommendations:

- Promote the development of minimum "best practices" for the experimental design and collection of data - especially when these data are expected to be utilized as part of a community pool.
- Foster sustained and long-term initiatives for tool development and sharing.
- Promote data standards and annotations.
- Foster sustained and long-term data repositories, preferably those that would promote data sharing across scales and taxa.
- Support funding agency and publisher requirements that new datasets, tools and code be shared and made easily accessible to the community.
- Promote programs that recruit, train, and retain a diversity of talent - both new students and retrained biologists - that are interested in the use of these approaches.

- Promote collaborative pan-disciplinary exchange between biologists and data scientists and related fields.
- Identify opportunities where funding can be leveraged for basic and applied questions concurrently, including in response to management of natural and human-dependent species or ecosystems in response to global change.

Acknowledgements

We would like to thank the University Corporation for Atmospheric Research and Knowinnovation for organizing the Reintegrating Biology Jumpstart Workshops which facilitated this working group and manuscript, and NSF for funding these workshops. We would also like to thank the many participants of the Reintegrating Biology Jumpstart Workshop in Austin, TX whose thoughtful feedback improved this manuscript.

References

Abberton, M., Batley, J., Bentley, A., Bryant, J., Cai, H., Cockram, J., Costa de Oliveira, A., Cseke, L.J., Dempewolf, H., De Pace, C. & Edwards, D., 2016. Global agricultural intensification during climate change: a role for genomics. *Plant biotechnology journal* **14**: 1095-1098.

Abrahams, G.J. & Newman, J. 2019. BLASTing away preconceptions in crystallization trials. *Acta Crystallographica Section F: Structural Biology Communications*, **75**(3): 184-192.

Agrawal, A., Edenberg, H.J. & Gelernter, J. 2016. Meta-analyses of genome-wide association data hold new promise for addiction genetics. *Journal of Studies on Alcohol and Drugs* **77**: 676-680.

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T., Kelsey, G., Stegle, O., Reik, W. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods* **13**(3): 229-232.

Bahn, V. & McGill, B.J. 2013. Testing the predictive performance of distribution models. *Oikos* **122**: 321-331.

Banf, M. & Rhee, S.Y. 2017. Computational inference of gene regulatory networks: Approaches, limitations and opportunities. *Biochimica et Biophysica Acta-Gene Regulatory Mechanisms* **1860** (1): 41-52.

Barnes, M.A., Jerde, C.L., Wittmann, M.E., Chadderton, W.L., Ding, J., Zhang, J., Purcell, M., Budhathoki, M., & Lodge, D.M. 2014. Geographic selection bias of occurrence data influences transferability of invasive *Hydrilla verticillata* distribution models. *Ecology and Evolution* **4**: 2584-2593.

Barnett, D.T., Duffy, P.A., Schimel, D.S., Krauss, R.E., Irvine, K.M., Davis, F.W., Gross, J.E., Azuaje, E.I., Thorpe, A.S., Gudex-Cross, D. & Patterson, M. 2019. The terrestrial organism and biogeochemistry spatial sampling design for the National Ecological Observatory Network. *Ecosphere* **10**: e02540.

Barone, L., Williams, J. & Micklos, D., 2017. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLoS computational biology* **13**: e1005755.

Bengston, S.E., Dahan, R.A., Donaldson, Z., Phelps, S.M., Van Oers, K., Sih, A. & Bell, A.M. 2018. Genomic tools for behavioural ecologists to understand repeatable individual differences in behaviour. *Nature Ecology & Evolution* **2:** 944-955.

Bentley, D.R. 2006. Whole-genome re-sequencing. *Current Opinion in Genetics & Development* **16**(6):545-552.

Billeter, J. & Levine, J.D. 2013. Who is he and what is he to you? Recognition in Drosophila melanogaster. *Current Opinion in Neurobiology* **23**:17-23.

Bland, L.M., Collen, B.E.N., Orme, C.D.L. & Bielby, J.O.N. 2015. Predicting the conservation status of data-deficient species. *Conservation Biology* **29**, 250-259.

Bruno, A.E., Charbonneau, P., Newman, J., Snell, E.H., So, D.R., Vanhoucke, V., Watkins, C.J., Williams, S. & Wilson, J. 2018. Classification of crystallization outcomes using deep convolutional neural networks. *PLoS ONE*, **13**(6).

Bubac, C.M., Miller, J.M., Coltman, D.W. 2020. The genetic basis of animal behavioural diversity in natural populations. *Molecular Ecology* Doi: 10.1111/MEC.15461.

Camacho, D.M., Collins K.M., Powers R.K., Costello J.C. & Collins J.J. 2018. Next-Generation Machine Learning for Biological Networks. *Cell* **14**: 1581-1592.

Caro, T.M. & O'Doherty, G. 1999. On the use of surrogate species in conservation biology. *Conservation biology*, **13**(4): 805-814.

Chen, J.W, Scaria J. & Chang Y. F. 2012. Phenotypic and Transcriptomic Response of Auxotrophic *Mycobacterium avium* Subsp. *paratuberculosis leuD* Mutant under Environmental Stress. *PLoS ONE* **7**(6): e37884

Chiquet J., Rigaill G. & Sundqvist M. 2019. A Multiattribute Gaussian Graphical Model for Inferring Multiscale Regulatory Networks: An Application in Breast Cancer. In: Sanguinetti G., Huynh-Thu V. (eds) Gene Regulatory Networks. Methods in Molecular Biology, vol 1883. Humana Press, New York, NY.

Coordinators, N.R. 2018. Database resources of the national center for biotechnology information. *Nucleic acids research*, **46**: D8.

Cordero-Maldonado M.L., Perathoner S., van der Kolk K-J., Boland R., Heins-Marroquin U., Spaink H.P., Meijer, A.H., Crawford A.D. & de Sonneville, J. 2019. Deep learning image recognition enables efficient genome editing in zebrafish by automated injections. *PLoS ONE* **14**: e0202377.

Curtin, K.D., Huang, Z.J. & Rosbash, M. 1995. Temporally regulated nuclear entry of the Drosophila period protein contributes to the circadian clock. *Neuron* **14**: 365-372.

Cussat-Blanc, S., Harrington, K. & Banzhaf, W. 2019. Artificial Gene Regulatory Networks - A Review. *Artificial Life* **24**(4): 296-328.

Das Gupta, M. & Tsiantis, M. 2018. Gene networks and the evolution of plant morphology, *Current Opinion in Plant Biology* **45**: 82-87.

Datta, S.R., Vasconcelos, M.L., Ruta, V., Luo, S., Wong, A., Demir, E., Flores, J., Balonze, K., Dickson, B.J., & Axel, R. 2008. The *Drosophila* pheromone cVA activates a sexually dimorphic neural circuit. *Nature*. **452**: 473-477.

da Veiga Leprevost, F., Grüning, B.A., Alves Aflitos, S., Röst, H.L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J. & Bai, M. 2017. BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, **33**: 2580-2582.

Davidson, E. H. & Erwin, D. H. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* **311:** 796–800.

Davis M.C. 2013. The Deep Homology of the Autopod: Insights from Hox Gene Regulation. *Integrative and Comparative Biology* **53**: 224-232.

Demir, E. & Dickson, B.J. 2005. Fruitless splicing specifies male courtship behavior in *Drosophila*. *Cell* **121**(5):785-794.

Dickson, B.J. 2008. Wired for sex: The neurobiology of *Drosophila* mating decisions. *Science* **322**:904-909.

Donaldson, M.R., Burnett, N.J., Braun, D.C., Suski, C.D., Hinch, S.G., Cooke, S.J. & Kerr, J.T. 2016. Taxonomic bias and international biodiversity conservation research. *Facets* **1**: 105-113.

Dong, B., Shao, L., Da Costa, M., Bandmann, O. & Frangi, A.F. 2015. Deep learning for automatic cell detection in wide-field microscopy zebrafish images. *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, New York, NY, 2015, pp. 772-776.

Drapeau, M.D, Cyran, S.A., Viering, M.M., Geyer, P.K. & Long, A.D. 2006. A cis-regulatory sequence within the yellow locus of *Drosophila melanogaster* required for normal male mating success. *Genetics* **172**:1009-1030.

Fiers M. W. E. J., Minnoye, L., Aibar, S., González-Blas, C. B., Atak, Z .K. & Aerts, S. 2018. Mapping gene regulatory networks from single-cell omics data, *Briefings in Functional Genomics*, **17**: 246–254.

Feng, S., Wang, S., Chen, C. & Lan, L. 2011. GWA Power: a statistical power calculation software for genome-wide association studies with quantitative traits. *BMC Genetics* **12**:12.

Ferreiro, A., Crook, N., Gasparrini, A. J. & Dantas, G. 2018. Multiscale Evolutionary Dynamics of Host-Associated Microbiomes. *Cell* **172:** 1216–1227.

Fraser, L.H., Henry, H.A., Carlyle, C.N., White, S.R., Beierkuhnlein, C., Cahill Jr., J.F., Casper, B.B., Cleland, E., Collins, S.L., Dukes, J.S. & Knapp, A.K. 2013. Coordinated distributed experiments: an emerging tool for testing global hypotheses in ecology and environmental science. *Frontiers in Ecology and the Environment* **11**: 147-155.

Frimpong, E.A., & Angermeier, P.L. 2009. Fish traits: a database of ecology and life-history traits of freshwater fishes of the United States. *Fisheries* **34**: 487-495.

Goodfellow, I,. Bengio, Y &, Courville, A. 2016. Deep Learning. MIT Press.

Grün, S., Grün, D. 2020. Deciphering cell fate decision by integrated single-cell sequencing analysis. *Annual Review of Biomedical Data Science* **3**: 1-22.

Hobbie, J.E., Carpenter, S.R., Grimm, N.B., Gosz, J.R., & Seastedt, T.R. 2003. The US long term ecological research program. *BioScience* **53**: 21-32.

Hübner, S., Korol, A.B., Schmid, K.J. 2015. RNA-Seq analysis identifies genes associated with differential reproductive success under drought-stress in accessions of wild barley *Hordeum spontaneum*. *BMC Plant Biology* **15**: 134.

Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., Spreafico, R., Hafler, D. A., & McKinney, E. F. 2019. From Big Data to Precision Medicine. *Frontiers in Medicine* **6**: 34.

Huynh-Thu, V.A. & Sanguinetti, G., 2019. Gene Regulatory Network Inference: An Introductory Survey. In: Sanguinetti G., Huynh-Thu V. (eds) Gene Regulatory Networks. Methods in Molecular Biology, vol 1883. Humana Press, New York, NY.

Ishaq, O., Sadanandan, S.K. & Wählby C. 2017. Deep Fish: Deep Learning–Based Classification of Zebrafish Deformation for High-Throughput Screening. *SLAS DISCOVERY:* ***22***: 102–107.

Jin, X., Ha, T.S. & Smith, D.P. 2008. SNMP is a signaling component required for pheromone sensitivity in *Drosophila*. *Proceedings of the National Academy of Sciences* **105**: 10996-11001.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N. & Zinzen, R.P. 2017. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**: 194-199.

Kattge, J., Diaz, S., Lavorel, S., Prentice, I.C., Leadely, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P.B., Wright, I.J. & Cornelissen, J.H.C. 2011. TRY - a global database of plant traits. *Global Change Biology* **17**: 2905-2935.

Keene, A.C. & Waddell, S. 2007. *Drosophila* olfactory memory: single genes to complex neural circuits. *Nature Reviews Neuroscience* **8**: 341-354.

Kegerreis, B., Catalina, M.D., Bachali, P., Geraci, N.S., Labonte, A.C., Zeng, C., Stearrett, N., Crandall, K.A., Lipsky, P.E., & Grammer, A.C. 2019. Machine learning approaches to predict lupus disease activity from gene expression data. *Scientific Reports* **9**: 1-12.

Khan, S., Rahmani H., Shah, S.A.A. & Bennamoun, M. 2018. A Guide to Convolutional Neural Networks for Computer Vision. *Synthesis Lectures on Computer Vision* **8**: 1-207.

Kronforst, M.R. & Papa, R. 2015 The functional basis of wing patterning in *Heliconius* butterflies: The molecules behind the mimicry. *Genetics* **200** 1-19.

Kültz, D., Clayton, D.F., Robinson, G.E., Albertson, C., Carey, H.V., Cummings, M.E., Dewar, K., Edwards, S.V., Hofmann, H.A., Gross, L.J. & Kingsolver, J.G. 2013. New frontiers for organismal biology. *BioScience* **63**: 464-471.

Lowe, E. K., Cuomo, C. & Arnone, M. I. 2017. Omics approaches to study gene regulatory networks for development in echinoderms. *Brief Funct Genomics* **16:** 299–308.

Lynch, M.L., Dudek, M.F. & Bowman, S.E.J. 2020. A Searchable Database of Crystallization Cocktails in the PDB: Analyzing the Chemical Condition Space. *Patterns*. doi.org/10.1016/j.patter.2020.100024.

Lytle, D.A., Martínez-Muñoz, G., Zhang, W., Larios, N., Shapiro, L., Paasch, R., Moldenke, A., Mortensen, E.N., Todorovic, S., & Dietterich, T.G. 2010. Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society* **29**: 867-874.

Martin, A., Papa, R., Nadeau, N.J., Hill, R.I., Counterman, B.A., Halder, G., Jiggins, C.D., Kronforst, M.R., Long, A.D., McMillan, W.O. & Reed, R.D. 2012. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a *Wnt* ligand. *Proceedings of the National Academy of Sciences* **109** (31):12632-12637.

Mace, D.L., Varnado, N., Zhang, W., Frise, E. & Ohler, U. 2010. Extraction and comparison of gene expression patterns from 2D RNA in situ hybridization images. *Bioinformatics* **26:** 761-769.

Maor-Landaw, K., Ben-Asher, H.W., Karako-Lambert, S., Salmon-Divon, M., Prada, F., Caroselli, E., Goffredo, S., Falini, G., Dubinsky, Z., Levy, O. 2017. Mediterranean versus Red sea corals facing climate change, a transcriptome analysis. *Scientific Reports* **7**:42405.

Merlin, C. & Liedvogel, M. 2019. The genetics and epigenetics of animal migration and orientation: birds, butterflies and beyond. *Journal of Experimental Biology* **222**: jeb191890. doi.org/10.1242/jeb.191890.

Merrill, R.M., Dasmahaptra, K.K., Davey, J.W., Dell'Aglio, D.D., Hanly, J.J., Huber, B., Jiggins, C.D., Joron, M., Kozak, K.M., Llaurens, V., Marin, S.H., Montgomery, S.H., Morris, J., Nadeau, N.J., Pinharanda, A.L., Rosser, N., Thompson, M.J., Vanjari, S., Wallbank, R.W.R., Yu, Q. 2015. The diversification of *Heliconius* butterflies: what have we learned in 150 years? *Journal of Experimental Biology* **28 (8)**: 1417-1438.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Velázquez Vega, J.E., Brat, D.J. & Cooper., L.A.D. 2018. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences* **115**: E2970-E2979.

Nadeau, N.J., Pardo-Diaz, C., Whibley, A., Supple, M.A., Saenko, S.V., Wallbank, R.W.R., Wu, G.C., Maroja, L., Ferguson, L., Hanly, J.J., Hines, H., Salazar, C., Merrill, R.M., Dowling, A.J., ffrench-Constant, R.H., Llaurens, V., Joron, M., McMillan, W.O. & Jiggins, C.D. 2016. The gene *cortex* controls mimicry and crypsis in butterflies and moths. *Nature* **534**:106-110.

Nocedal, I. & Johnson, A. D. 2016. How transcription networks evolve and produce biological novelty. *Cold Springs Harbor Symposia on Quantitative Biology* **80:** 265–274.

National Research Council (US) Committee on a New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution. 2009. A New Biology for the 21st

Century: Ensuring the United States Leads the Coming Biology Revolution. Washington (DC): *National Academies Press (US)* **4**: pp 112.

Nussinov, R., Tsai, C. J. & Jang, H. 2019. Protein ensembles link genotype to phenotype. *PLoS Computational Biology* **15**: e1006648.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., & Aittokallio, T. 2014. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genetics* **10**: e1004754.

Olden, J.D., & Jackson, D.A. 2002. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* **154**: 135-150.

Olden, J.D., Lawler, J.J. & Poff, N.L. 2008. Machine learning methods without tears: a primer for ecologists. *Quarterly Review of Biology* **83**: 171-193.

Patrushev, I., James-Zorn, C., Ciau-Uitz, A., Patient, R. & Gilchrist, M.J. 2018. New methods for computational decomposition of whole-mount *in situ* images enable effective curation of a large, highly redundant collection of *Xenopus* images. *PLoS Computational Biology* **14**: e1006077.

Pespeni, M.H., Barney, B.T., Palumbi, S.R. 2013. Differences in the regulation of growth and biomineralization genes revealed through long-term common-garden acclimation and experimental genomics in the purple sea urchin. *Evolution* **67**(7): 1901-1914.

Puniyani, K. & Xing, E.P. 2013. GINI: From ISH images to gene interaction networks. *PLoS Computational Biology* **9:** 1003227.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A. & Zhang, F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**:2281-2308.

Rebeiz, M., Patel, N. H. & Hinman, V. F. 2015. Unraveling the Tangled Skein: The Evolution of Transcriptional Regulatory Networks in Development. *Annu Rev Genomics Hum Genet* **16:** 103–131.

Rebeiz, M. & Tsiantis, M. 2017. Enhancer evolution and the origins of morphological novelty. *Curr Opin Genet Dev* **4:** 115–123.

Reed, R.D., Papa, R., Martin, A., Counterman, B.A., Pardo-Diz, C., Jiggins, C.D., Chamberlain, N.L., Kronforst, M.R., Chen, R., Halder, G., Nijhout, H.F. & McMillan, W.O. 2011. Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333**:1137-1141.

Royle, J.A., Chandler, R.B., Yackulic, C., & Nichols, J.D. 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution* **3**: 545-554.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1:** 206-215.

Ruta, V., Datta, S.R., Vasconcelos, M.L., Freeland, J., Looger, L.L. & Axel, R. 2010. A dimorphic pheromone circuit in *Drosophila* from sensory input to descending output. *Nature*. **468**:686-692.

Sarov-Blat, L., So, W.V., Liu, L. & Rosbash, M. 2000.The *Drosophila takeout* gene is a novel molecular link between circadian rhythms and feeding behavior. *Cell* **101**: 647-656.

Schwaerzel, M., Monastirioti, M., Scholz, H., Friggi-Grelin, F., Birman, S. & Heisenberg, M. 2003. Dopamine and octopamine differentiate between aversive and appetitive olfactory memories in *Drosophila*. *Journal of Neuroscience* **23**(33): 10495-10502.

Shen, X., Shen, S., Li, J., Hu, Q., Nie, L., Tu, C., Wang, X., Poulsen, D. J., Orsburn, B. C., Wang, J. & Qu, J. 2018. IonStar enables high-precision, low-missing-data proteomics quantification in large biological cohorts. *Proceedings of the National Academy of Sciences* **115** (21): E4767-E4776.

Shen, J., Petkova, M.D., Tu, Y., Liu, F. & Tang, C. 2020. Deciphering gene regulation from gene expression dynamics using deep neural network. *bioRxiv* doi: https://doi.org/10.1101/374439.

Siahpirani A.F., Chasman D. & Roy S. 2019. Integrative Approaches for Inference of Genome-Scale Gene Regulatory Networks. In: Sanguinetti G., Huynh-Thu V. (eds) Gene Regulatory Networks. Methods in Molecular Biology, vol 1883. Humana Press, New York, NY.

Skinnider, M.A., Squair, J.W. & Foster, L.J. 2019. Evaluating measures of association for single-cell transcriptomics. *Nat Methods* **16**: 381–386 doi:10.1038/s41592-019-0372-4.

Smith, S., Bernatchez, L., Beheregaray, L.B. 2013. RNA-seq analysis reveals extensive transcriptional plasticity to temperature stress in a freshwater fish species. *BMC Genomics* **14**: 375.

Suri, V., Lanjuin, A. & Rosbash, M. 1999. TIMELESS-dependent positive and negative autoregulation in the *Drosophila* circadian clock. *The EMBO Journal* **18** (3): 675-686.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. & Meyre, D. 2019. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**: 467-484. do.org/10.1038/s41576-019-0127-1.

Thompson, D., Regev, A. & Roy, S. 2015. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution. *Annu Rev Cell Dev Biol* **31**: 399–428.

Urban, M.C., Bocedi, G., Hendry, A.P., Mihoub, J.B., Pe'er, G., Singer, A., Bridle, J.R., Crozier, L.G., De Meester, L., Godsoe, W. & Gonzalez, A. 2016. Improving the forecast for biodiversity under climate change. *Science* **353**: p.aad8466.

Valan, M., Makonyi, K., Maki, A., Vondráček, D., & Ronquist, F. 2019. Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Systematic Biology* **68**: 876-895.

van den Hoogen, J., Geisen, S., Routh, D., Ferris, H., Traunspurger, W., Wardle, D.A., De Goede, R.G., Adams, B.J., Ahmad, W., Andriuzzi, W.S. & Bardgett, R.D. 2019. Soil nematode abundance and functional group composition at a global scale. *Nature* **572**: 194-198.

Vinauger, C., Lahondere, C., Wolff, G.H., Liaw, J.E., Parrish, J.Z., Akbari, O.S., Dickinson, M.H. & Riffell, J.A. 2018. Modulation of host learning in *Aedes aegypti* mosquitoes. *Current Biology* **28** (3): 333-344.

Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. 2012. Five years of GWAS discovery. *The American Journal of Human Genetics* **90**:7-24.

Wagner, G.& Zhang, J. 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics* **12**: 204–213.

Walsh, I., Pollastri, G., & Tosatto, S.C. 2016. Correct machine learning on protein sequences: a peer-reviewing perspective. *Briefings in Bioinformatics* **17**: 831-840.

Webb, S., 2018. Deep learning for biology. *Nature* **554**: 555–557.

Wenger, S.J. & Olden, J.D. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution* **3**: 260-267.

Westerman, E.L., VanKuren, N.W., Massardo, D., Tenger-Trolander, A., Zhang, W., Hill, R.I., Perry, M., Bayala, E., Barr, K., Chamberlain, N., Douglas, T.E., Buerkle, N., Palmer, S.E. & Kronforst, M.R. 2018. *Aristaless* controls butterfly wing color variation used in mimicry and mate choice. *Current Biology* **28**(21):3469-3474.e4 doi.org/10.1016/j.cub.2018.08.051.

Westerman, E.W. 2019. Searching for the genes driving assortative mating. *PLoS Biology* **17**(2):e3000108. doi.org/10.1371/journal.pbio.3000108.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes A.J., Clark T.,

Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C.T., Finkers R., Gonzalez-Beltran A., Gray A.J., Groth P., Goble C., Grethe J.S., Heringa J., 't Hoen P.A., Hooft R., Kuhn T., Kok R., Kok J., Lusher S.J., Martone M.E., Mons A., Packer A.L., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S.A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M.A., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. PMID: 26978244; PMCID: PMC4792175.

Willcock, S., Martínez-López, J., Hooftman, D.A., Bagstad, K.J., Balbi, S., Marzo, A., Prato, C., Sciandrello, S., Signorello, G., Voigt, B., & Villa, F. 2018. Machine learning for ecosystem services. *Ecosystem Services* **33**: 165-174.

Wu, S., Joseph, A., Hammonds, A. S., Celniker, S. E., Yu, B. & Frise, E. 2016. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences*. **113**(16):4290–4295. doi.org/10.1073/pnas.1521171113.

Xu, P., Atkinson, R., Jones, D.N.M. & Smith, D.P. 2005. *Drosophila* OBP LUSH is required for activity of pheromone-sensitive neurons. *Neuron* **45**:193-200.

Xu, X. & Bai, G. 2015. Whole-genome resequencing: changing the paradigms of SNP detection, molecular mapping and gene discovery. *Molecular Breeding* **35**:33. doi.org/10.1007/s11032-015-0240-6.

Yan, K. K., Wang, D., Sethi, A., Muir, P., Kitchen, R., Cheng, C. & Gerstein, M. 2016. Cross-Disciplinary Network Comparison: Matchmaking between Hairballs. *Cell Systems* **2:** 147–157.

Yang Y., Fang Q. & Shen H-B. 2019. Predicting gene regulatory interactions based on spatial gene expression data and deep learning. *PLoS Computational Biology* **15**(9): e1007324. doi. org/10.1371/journal.pcbi.1007324.

Figure Legends

Figure 1: A molecular cascade largely encoded by the genome (top row) generates observed phenotypic variation on multiple scales from cells to ecosystems (bottom row). Inter-organismal phenotype refers to holobiont, populations, and communities in this context. Environmental factors can have epigenetic impacts on the genetic program (dashed arrows). Although high-throughput sequencing has led to a rapid increase in available genomes and transcriptomes (large diameter pipelines), the ability to leverage this data to understand the emergence of diverse phenotypes is extremely limited (smaller diameter pipelines). Deciphering the genome to phenome pathway necessitates a multidisciplinary approach including scientists with expertise at each of these scales. Integration of skilled data scientists will be critical for increasing the rate of productive analyses at each of these chokepoints. Image credits: Sequencing image - www.genomicseducation.hee.nhs.uk/, DNA strand-Tracey Saxby, Integration and Application Network, Proteome image - modified from Shen et al., 2018 Gene network - modified from Chen et al., 2015, Epidermis -http://blogs.ubc.ca/, Wild flower filled prairie- Grace Hirzel, by permission. All other images by authors.

Figure 2: Complex phenotypes such as mate preference and color pattern are often a mosaic of smaller, simpler elements, whose genetic underpinnings are easier to identify independently

than when assessed as a group. This approach has proven quite fruitful for identifying causative genes for color patterning elements, and can be used for other complex traits such as mate preference, as illustrated here. Once the genetic underpinnings of these elements are known, their combinatorial effects can be explored, as well as pleiotropic effects of genetic background and any effects of environment.

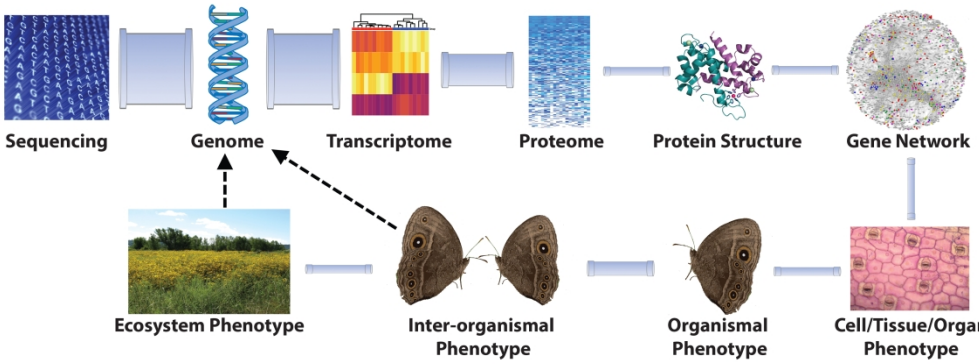

Figure 1: A molecular cascade largely encoded by the genome (top row) generates observed phenotypic variation on multiple scales from cells to ecosystems (bottom row). Inter-organismal phenotype refers to holobiont, populations, and communities in this context. Environmental factors can have epigenetic impacts on the genetic program (dashed arrows). Although high-throughput sequencing has led to a rapid increase in available genomes and transcriptomes (large diameter pipelines), the ability to leverage this data to understand the emergence of diverse phenotypes is extremely limited (smaller diameter pipelines). Deciphering the genome to phenome pathway necessitates a multidisciplinary approach including scientists with expertise at each of these scales. Integration of skilled data scientists will be critical for increasing the rate of productive analyses at each of these chokepoints. Image credits: Sequencing image - www.genomicseducation.hee.nhs.uk/, DNA strand-Tracey Saxby, Integration and Application Network, Proteome image - modified from Shen et al., 2018 Gene network - modified from Chen et al., 2015, Epidermis -http://blogs.ubc.ca/, Wild flower filled prairie- Grace Hirzel, by permission. All other images by authors.

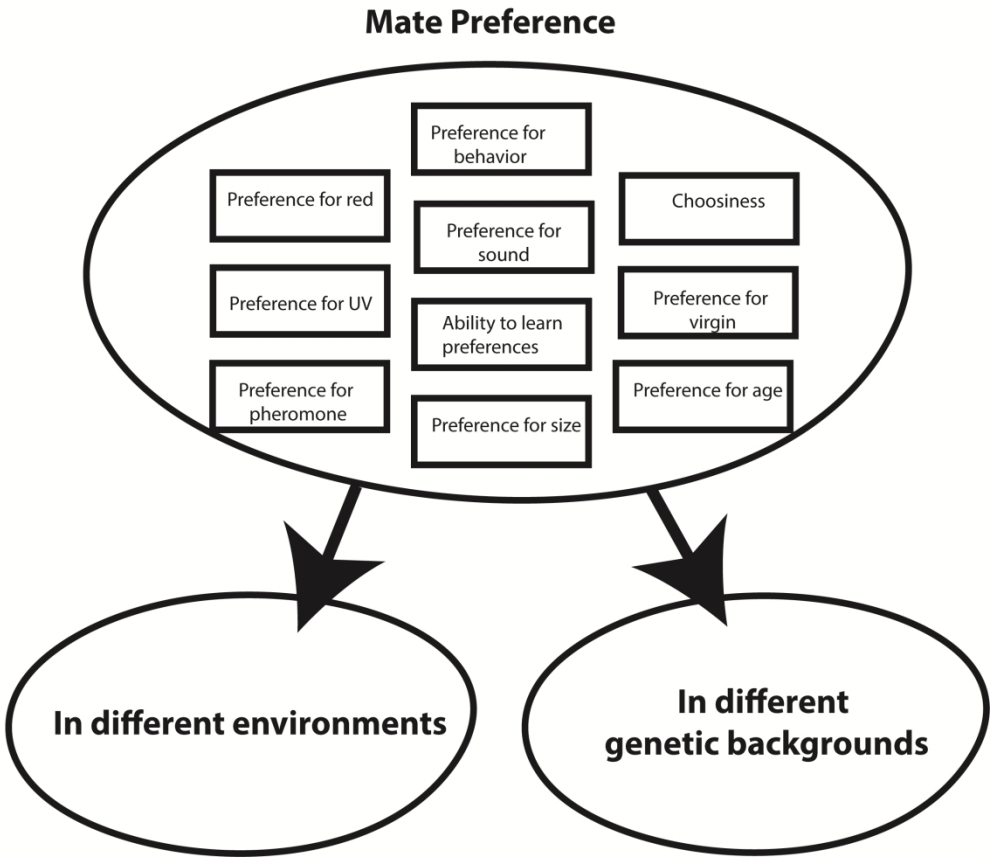

Figure 2: Complex phenotypes such as mate preference and color pattern are often a mosaic of smaller, simpler elements, whose genetic underpinnings are easier to identify independently than when assessed as a group. This approach has proven quite fruitful for identifying causative genes for color patterning elements, and can be used for other complex traits such as mate preference, as illustrated here. Once the genetic underpinnings of these elements are known, their combinatorial effects can be explored, as well as pleiotropic effects of genetic background and any effects of environment.