Title: Deploying Big Data to Crack the Genotype to Phenotype Code (g2p w/ML)

Authors:

- Sarah EJ Bowman (sbowman@hwi.buffalo.edu) (Metalloproteins, Structural Biology)
- Bradley Davidson (bdavids1@swarthmore.edu) (Developmental Biology, Evolution)
- Marcus C Davis (davis4mc@jmu.edu) (Evolutionary Developmental Biology)
- Eric R Larson (erlarson@illinois.edu) (Ecology, Conservation Biology)
- Christopher Sanford (csanfo19@kennesaw.edu) (Biomechanics and Functional Morphology of Vertebrates)
- Erica Westerman (ewesterm@uark.edu) (Integrative Animal Behavior, Behavioral Genomics)

The Goal:
Use Big Data and Machine Learning approaches to crack the genotype to phenotype code and thereby generate predictive frameworks across biological scales.

Introduction:
Deciphering the mechanisms by which genotypes generate phenotypes is a central mission of biology. Fully realizing these mechanisms will facilitate integration of enormous datasets in organismal diversity research across molecular, morphological, behavioral, and ecosystem scales. Comprehensive, multi-scale data integration will impact broad reaching, interdisciplinary and integrative goals across biological disciplines including: 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; 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. Machine Learning (ML) 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) development of data collection and storage standards; 2) development of tools to overcome key bottlenecks in data acquisition and analysis; and 3) training initiatives and collaborative outreach. Here we discuss how sustained efforts in these areas can further catalyze the Big Data era for cracking the genotype to phenotype (g2p) code. We follow this discussion by detailing a few specific transformative research opportunities that will be advanced by these efforts.

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.

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, annotation and availability.

The need for experimental design standards**.** One of the challenges associated with using big data to crack the genotype to phenotype code is determining how to obtain "good" data. In particular, it is critical to determine what questions can be answered by data in hand, and what kinds of data are required to answer the key, unresolved questions. Additionally, the potential benefits of high-throughput sequencing are greatly limited by inherent difficulties in extracting signal from noise. One way to address these challenges involves revisiting experimental design, including reassessing the type of sequencing data used to characterize the genotype to phenotype code. For example, sequencing techniques (RadSeq, SNP arrays or whole genome resequencing) should be carefully chosen 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.

Important questions to think about before starting to collect data.

Extracting signal from noise can be related to two different aspects of the genotype to phenotype question. The data we are collecting may be noisy, but so may be the phenotype. 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 carefully define phenotypes, or to design experiments that allow for the detection of genes associated with specific aspects of biologically relevant 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. These narrowly defined phenotypes are the building blocks of the larger phenotype of interest, and once characterized, may scale up.

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). 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 (https://biocontainers.pro/#/, NEON; Barnett et al., 2019) from which we can

learn best practices. Finally, close links between research groups collecting and research groups analyzing data will be essential. A “hubs and spokes” approach may be the most 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.

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. 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 (in the model of the NEON initiative, Barnett et al., 2019).

5. Data is exponentially increasing, unwieldy and noisy.
It is critical to develop tools that can distinguish signal from noise across large datasets. **Deep learning** could be employed as a possible solution. The development of such deep 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 machine learning should share best practices across subdisciplines.
Scientists apply machine learning approaches to a variety of biological questions, from mapping genotypes to phenotypes or structure to function, to predicting relationships between the distribution of species and their environments. These subfields working independently likely

encounter some of the same challenges in applying and interpreting ML approaches: Are the big data sources we use reliable and well-maintained? Do machine learning predictions have mechanistic meaning, or are they occasionally over-fitting to noise? Biologists applying ML to big data questions across levels of the biological hierarchy might share experiences on best practices or lessons from their application and interpretation of their tools. Shared challenges may include identifying cases of model over-fitting, improving interpretability of “black box” machine learning output, quantifying or identifying uncertainty in predictions, and sharing practices for independence of training and testing data. 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. E.g., do ecologist’s concerns about the independence of machine learning testing and training data, and associated implications for model transferability or generalizability, relate to machine learning in other fields of biology (Wenger & Olden 2012, Bahn & McGill 2013)? Further, ML 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. 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. Leveraging a big

data and ML 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. 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). Additionally, these maps can 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. Much of these data are in the form of microscopy generated images, and are used to detect spatial and temporal co-expression of genes, in wild-type and experimental systems, across an ever-expanding range of organisms (Puniyani & Xing, 2013; Davis, 2013; Wu et al., 2016). However, these datasets go underutilized. Expression 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). Furthermore, gene expression patterns, like phenotypes, are open to

subjective interpretation (Yang et al., 2019). Machine learning approaches can potentially overcome these challenges, allowing for a more effective use of a comparative gene expression approach to generate hypotheses regarding GRN architecture and the ways network structure has shifted to generate novel developmental pathways and phenotypes. 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 to these algorithms are leading to autonomous learning, in which the model itself is required to find meaningful patterns in the data (Webb, 2018). Recent approaches at employing deep learning algorithms yield promising returns in the building of "*in silico embryos"* (Shen et al., 2018) and generation of GRN predictive models using expression data (Yang et al., 2019). Sustained progress in this area will require initiatives that 1) promote tool/algorithm development and sharing; and 2) foster long-term pan-taxa repositories for gene expression datasets.

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, from foraging to mating to habitat selection, in a wide range of species. These well defined phenotypes are ripe for genotype-phenotype discovery. 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). The challenge is to integrate genomic, proteomic, and network approaches (and scientists!) 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. 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).

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 or algorithmic 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.

Creating the human infrastructure for a Big Data and Machine Learning approach:
Ironically, leveraging Big Data and ML 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, 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. Addressing some of these challenges may involve 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.
- 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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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**.

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

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

Hoffmann, A., Griffin, P., Dillon, S., Catullo, R., Rane, R., Byrne, M., Jordan, R., Oakeshott, J., Weeks, A., Joseph, L., & Lockhart, P., 2015. A framework for incorporating evolutionary genomics into biodiversity conservation and management. *Climate Change Responses*, ***2,*** 1.

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.

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.

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.

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

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

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

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.

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* **45,** 115–123.

Shen, J., Liu, F., Tang, C.\., 2018. Toward deciphering developmental patterning with deep neural network. bioRxiv 374439. 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

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.

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.

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.

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.

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. 2016; 113(16):4290–4295. doi.org/10.1073/pnas.1521171113

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 Comput Biol 15(9): e1007324. doi. org/10.1371/journal.pcbi.1007324

APPENDIX

See overlapping themes with Vision Paper #15 - Austin