Modernizing the Phenotype: Defining phenotypes across scales

Authors:
Ramona Walls, The University of Arizona
John Cort, Pacific Northwest National Laboratory
Dina Navon, University of California Riverside
Sladjana Prisic, University of Hawaii
Sanja Roje, Washington State University
Jackie Rose, Washington State University

Summary

Variation in the phenotype – the entire set of observable traits characterizing a biological entity that are the product of genes interacting with the environment – forms the basis of much basic and applied biological research and is central to fields ranging from biodiversity to human health. Phenotypic data can be generated across levels of organization from molecular to cellular, organismal, population, community, and ecosystem, and researchers working at each level have their own concepts of what a phenotype is. Though the genotype-phenotype paradigm has evolved since its emergence over one hundred years ago, we argue that it has not kept pace with modern high throughput molecular analyses, nor has it been expanded to be fully inclusive of non-traditional phenotypes such as behavior and microbial community function. A holistic approach that unifies phenotypes from the molecular to the ecosystem level via explicit consideration of the mechanisms that connect levels of organization could provide a common framework for all levels of biological research and integrate currently fragmented fields of biology in much the same way genomic data are now used across all biology. The timing is ripe for modernizing the study of phenomics, thanks to the growth of high throughput phenotyping technologies, coupled with high throughput molecular approaches such as proteomics and metabolomics, data integration tools such as phenotype ontologies, and new analysis methods such as machine/deep learning. Many of the intellectual, technical, and social challenges that have constrained phenomics to a descriptive approach have been or are on the verge of being overcome. A modernized conceptualization of phenotype that incorporates the innumerable ways that biological systems can be observed could be instrumental in reintegrating biology. The tools and infrastructure needed to support the integrated phenotype will enhances our ability to manipulate biological systems at different scales and predict systems' responses to change and provide intellectual capital that can support many other areas of research. Societal benefits would include applications in sustainability, food security, curing diseases, bioengineering, and renewable energy.

Introduction

The concept of phenotype originated in the field of genetics (Johannsen 1903) more than one hundred years ago — prior to the development of mechanistic biochemistry and molecular biology. It has not kept pace with modern high throughput molecular analyses, nor has it been expanded to be fully inclusive of non-traditional (e.g., behavioral or microbial community) phenotypes. For example, at the molecular scale, global analyses such as metabolomics or proteomics aspire to identify and quantify tens of thousands of distinct molecules present in a single cell, pure culture, or a mixed population. Such data are not usually included in description of the phenotype, though fields like evo-devo are beginning to call these kinds of phenotypes into our standard definition. As these sets of molecules and the network of interactions among them over time constitute the very machinery responsible for transducing the genotype to the phenotype, their inclusion in our understanding of the phenotype is essential. A hypothetical "complete" description and understanding of molecular interactions and their outcomes, plus how they respond to the environment, might someday be able to predict cellular, organismal, or larger scale phenotypes. However, many phenotypes are cryptic, and manifest only at the molecular level or in the context of particular environments (e.g., Rohner et al. 2013). Recent efforts to build ontologies are only beginning to take on the phenotype — data that have been traditionally a descriptive characterization of traits observed visually — hindering quantitative comparisons. Standard, machine interpretable descriptions of phenotypes are essential for high-throughput "phenomics" (e.g., plant-growth parameters) that mirror analagous high-throughput molecular approaches such as proteomics and metabolomics. We propose that the concept of the phenotype should be updated to reflect how biology is understood today: an integrative, multiscale, quantitative, cross-disciplinary, and complete description of life forms.

Modernizing the phenotype

The current view of the phenotype is fragmented among disciplines and levels of biological organization. The concept of phenotype was originally coined to distinguish the observable characteristics of an organism from the (at that time) unobservable components that determined them (the genotype, Johannsen 1903). As researchers began to understand the role of the environment in shaping an organism, the notion of phenotype as a product of genotype interacting with the environment (GxE=P) became a paradigm in many fields of biology, with some disagreement. For example, some developmental geneticists imply that G -> P and E contributes only noise/messiness in that signal, whereas evolutionary-developmentalists might argue that that's a limited view of the world, as the environment can significantly predict the phenome, to a similar degree as the genome. Scientists have begun to deconstruct and rethink the GxE=P paradigm to accommodate phenotype by phenotype interactions, temporal changes (both over development and in response to environment), epigenetics, etc. For example, one new formulation would be [[GxG]xE](T1) x [[GxG]xE](T2)...=P (pers comm. Dina Navon).

Researchers' conception of the phenotype also vary greatly among those studying different levels of biological organization. Biologists working at the organismal or population level in

multicellular organisms tend to focus on morphological, behavioral, or physiological and phenotypes. Molecular level characterization of biological systems as practiced today aspires to provide a complete, quantitative, and time-resolved description of the entire complement of molecules in (ideally) a single cell or at least a small number of homogeneous types of cells. Usually these approaches have the suffix “omics” appended: genomics, transcriptomics, proteomics, and metabolomics are currently in different stages of active development and in principle can be applied simultaneously to the same sample. Together, such measurements could be described as a “molecular phenotype”, insofar as the cellular or organismal phenotype is the outcome of all biochemical processes, mechanisms, and interactions occurring at the molecular-scale within and among the cell being observed. To consider the phenotype as a multifaceted description that spans scales, comprising many types of measurements or observations across many scales, would simply bring the phenotype up to date -- modernizing it. This modernized phenotype would actually be analogous to the descriptive phenotype that Johannsen would have had in mind over 100 years ago, had such a variety of measurements been available. It is the complete description of what is there and how it behaves over the course of the observation.

The modernized phenotype need not be restricted to one cell, and it does not need to be directed only toward the smaller components of the cell. Microbiologists and microbial/viral ecologists study molecular functions of entire communities of populations based on levels of gene expression or the presence of certain molecules in environmental samples. At a broad scale, ecosystem ecologists may view the collective characteristics of an ecosystem or biome to be its “phenotype”. Even at this macroscopic level however, molecular-scale “meta-omics” data are being used to understand ecosystem-level responses to environmental perturbations, for example changes in molecular process responsible for elemental cycling by microbial communities due to warming, acidification, desiccation, or sea level rise.

The phenotype is what we are able to observe and our ability to “observe”, or measure, a variety of traits has dramatically improved in the 21st century. In addition, our ability to modify genotype and manipulate environment has also been improved, compared to the 20th century methodology. At the smallest scales, we are now able to regulate expression of almost any gene in virtually any organism using CRISPR/Cas9, identify and quantify small amounts of protein and metabolites in complex mixtures using mass spectrometry, and track single molecules in the cell and measure their interactions using sophisticated microscopy methods. At larger scales, we can use multispectral high throughput phenotyping and satellite imagery, combined with complex image processing algorithms and machine learning to automatically score phenotypes one entire populations or ecosystems in a matter of hours. These new approaches provide new opportunities for understanding the phenotype in modern terms, but also increases the complexity of data analysis.

We propose a holistic approach that unifies phenotypes from the molecular to the ecosystem level via explicit consideration of the mechanisms that connect levels of organization (Fig. 1). Driving our approach is the goal of answering questions such as how do changes in the genome affect the transcriptome, proteome, and metabolome? How do those molecular phenotypes then

determine cellular and organismal phenotypes? How do organismal phenotypes affect the characteristics of populations, communities, and ecosystems? How do environmental conditions (at all scales) factor into these translations across scales? By defining phenotypes through parameters that we can measure at each level and focusing on the design of better experimental and computational tools to predict function using measurable parameters, we could delimit the minimum information needed to integrate phenotypes across scales from one level of organization to another and bridge gaps in integration across different levels/scales.

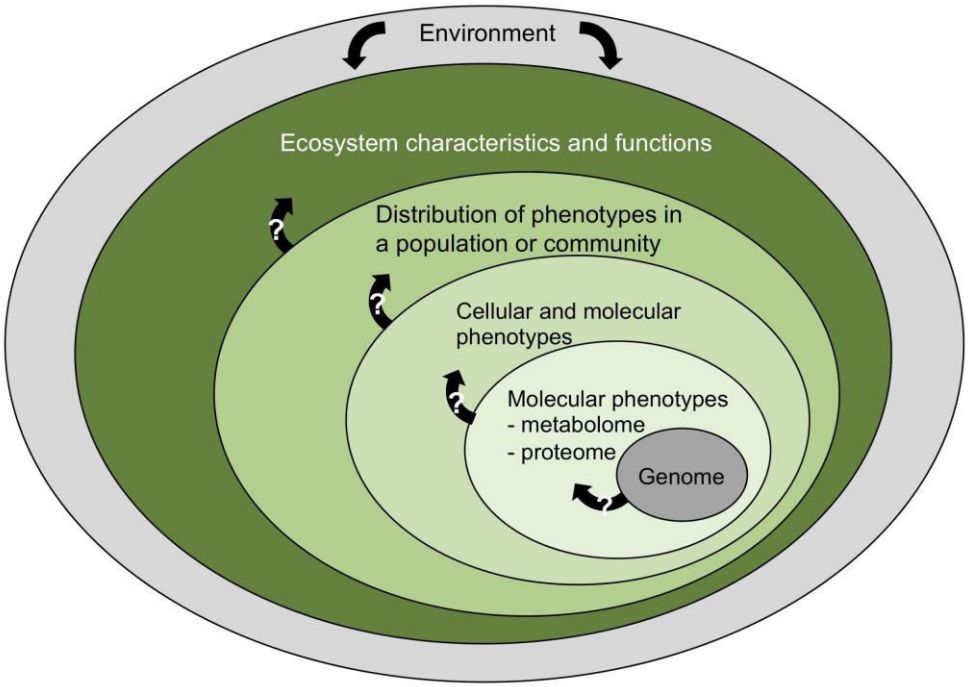

Fig. 1. *An integrated approach to the study of phenotypes. Phenotypes at different levels of biological organization are represented as nested green ovals, each level depending on all of the lower levels (black arrows with question marks) as well as the most basic level of the genome (dark gray oval). The environment (light gray oval) can influence how information is transferred between any levels. For simplicity, this figure omits influences such as interactions within any level or feedback from higher levels.*

Challenges and solutions

There are many intellectual, technical, and social barriers and challenges to holistically describing phenotypes, yet we are at a point where solutions to many of these challenges are possible. **Intellectually**, the biggest challenge to overcoming the current fragmented view of a phenotype may be developing conceptual and mathematical models that can incorporate molecular ‘omic data; morphological, behavioral, and physiological observations; and environmental parameters. Additional challenges arise from changes in phenotypes over time within a single organism (due to ontogeny, behavior, or environment) and the need to understand evolutionary constraints. One way to overcome these challenges is to reduce the

information required by identifying missing components that span levels or systems. Having well defined phenotypes that one could work with across levels/scales would substantially facilitate progress. Network analyses that includes genotypes, environments, and various aspects of phenotypes in an ecologically relevant system also would be extremely useful.

Technical challenges center around the collection and management of phenotypic data, because a great deal of data are needed to analyze phenotypes holistically. This presents challenges in scaling up data collection, storage, and analysis. Harmonizing and standardizing phenotype descriptions across scales and domains also presents major challenges. Fortunately, we are on the cusp of a state where advances in computing power and data technology can free phenotype data from domain or taxon-specific siloes to become part of the web of linked data. A key to the success of our approach is the standardized description of phenotypic data across domains, taxa, and scales. Ontologies to describe phenotypes for most organisms (e.g., UBERON, PO, OBA, MICRO, MP, PATO, all available at http://www.obofoundry.org/) exist at various levels of detail. Augmenting free-text descriptions of phenotypes with machine-readable ontology terms is essential. We also must develop methods for integrating phenotypes that do not easily map to existing ontologies (e.g., features extracted from multispectral images via machine learning that have no analog in traditional trait descriptors). Shared semantic models for managing phenotypic and environmental data (e.g., Walls et al. 2018, Madin et al. 2007) provide the backbone for large-scale data integration. Schema.org provides a method of making data discoverable through machine readable metadata and interpretable if ontology terms are included in the metadata. FAIR data principles (Wilkinson et al. 2016) provide guidance on how to make data findable, accessible, interoperable, and reusable, such as using permanent, globally unique identifiers, standardized metadata, and appropriate licensing.

As with any change in paradigm, **social challenges** often prove the most difficult to overcome. Communication among researchers working in different fields is not always easy to establish, as we presently rely to a large extent on physical proximity and word of mouth. Some of the collaborations needed to integrate phenotypes would require interactions among scientists that seldom or never have a chance to meet. Simply deciding which suites of phenotypes to prioritize is another challenge. Research that is beneficial to human survival -- medicine, food security, environmental protection, clean energy, adjustment to the changing environmental conditions -- provide one way to evaluate priorities, but we do not necessarily have *a priori* knowledge of which phenotypes are important for those problems. Finally, the need for data from many sources raises the challenges associated with data sharing, such as how to share data an equitable way that ensures proper credit for work and addresses current inequities in data and knowledge access.

Many of these challenges could be overcome with targeted funding not only for integrative research, but for social and educational networks. To ensure that future generations are set on the right educational path toward collaborative research, we need to provide students with opportunities for problem-based learning. This would enable those who are interested in pursuing careers in science to develop the set of intellectual and communication skills needed to tackle complex scientific problems in a collaborative environment. FAIR data and open

science principles provide guidelines for researchers, funders, and other stakeholders, but much more education in this area is needed. Funders and can academic institutions can play a role in setting policies that support data sharing, proper citation and credit, and collaboration.

Potential impacts

A modernized conceptualization of phenotype that incorporates the innumerable ways that biological systems can be observed today could be instrumental in reintegrating biology. Our proposed approach requires access to standardized data and facilitated communication with researchers working at different levels of organization. Developing the tools and infrastructure for this approach (e.g., biological models, ontologies, distributed data management systems) will provide intellectual capital that can support many other areas of research. Likewise, multidisciplinary training focused on development of problem-solving skills will ensure that future generations of scientists are prepared to tackle not only questions about integrated phenotypes, but many other challenges as well.

Analyzing and integrating phenotypes across scales is key to elucidating underlying biological mechanisms that are inherently linked to scale. Such knowledge would enhance our ability to manipulate biological systems at different scales and predict systems' responses to change. We envision that societal benefits would include applications in sustainability, food security, curing diseases, bioengineering, and renewable energy. Adopting a holistic approach to the study of phenotypes will not only allow a new, integrated, cross-scale comprehension of biology to solve big problems, it will enhance the individual disciplines that are the building blocks of interdisciplinary questions. In this context, subdisciplines within biology are appreciated not only as resources for new tools of investigation, but as the source of basic science discoveries that can be applied to a broader understanding beyond the original scope of study.

As a specific example of how our approach interacts with traditional disciplinary biology, consider the field of neuroscience. Behavioral neuroscience measures dynamic phenotypes, characterized by change, to describe underlying neuronal processes of behavior. Manipulating environmental and/or internal conditions in order to measure the effects of these conditions on behavior requires an understanding of the baseline behavioral response and mechanism(s) to discern that some manipulation reliably produced or influenced a change in that response. Consistent descriptions of even basic behavioral components are needed across experiments and researchers to discover neuronal mechanisms for agreed upon measures of behaviors. However, consistency in behavioral measures (and therefore phenotype descriptions) is not easily achieved, given the diversity of protocol variations that occur even in the most prescribed behavioral assays. This variability in behavioral measures and thus descriptions impacts the field by making it difficult to recognize molecular as well as ecological research findings that may align along the same phenotype trajectory at different levels of scale or across time.

Establishing an open, standardized, consistent resource of well-described behavioral phenotypes using agreed upon guidelines (e.g., FAIR) would facilitate making novel predictions by allowing for a behavioral ontology to be developed across levels of organismal and functional

complexity. A major limitation preventing the coalescing of behavioral phenotypes is the current lack of published stand-alone phenotype descriptions; most publishing outlets require a series of experiments uncovering mechanism for work to be deemed suitable for publication. Further, mechanism is appreciated to the extent that it can be demonstrated to be 'necessary' and 'sufficient' to produce a change in behavior. Together, these emphases do not take into account that the behavior may be a part of a much broader ecological phenomenon or that the mechanism of that behavior is subject to limitations of molecular dynamics. Other areas of biology that value describing phenotypes have developed a rich tradition of publishing phenotype descriptions allowing for categorization and thus connections to be made along multiple lines; however, this approach is not frequently seen with regards to studies of behavioral mechanisms in animal models. As a first step, we propose the development of standardized behavioral phenotype descriptors that can be used to publish incremental phenotype data.

Conclusion

Most biologists have a working definition of phenotype, but those definitions vary widely from molecular biologists to ecosystem scientists. We posit that there is a unified view of the phenotype that can span all levels of biological organization and spatial scales, and that a shared information and data framework, including shared terminology and formats, will reintegrate biological sciences and support a more sophisticated reworking of the biological paradigm GxE=P. Our vision is that any biologist working with phenotypes (which is nearly all biologists), could contribute to the shared pool of phenotypic knowledge across scales. A holistic view of the phenotype requires input from molecular, developmental, and organismal biologists, ecologists, ecosystem scientists, data scientists, and theoreticians. With a holistic knowledge of phenotypes, we could begin to understand how changes in the genome translate into cellular and organismal level phenotypic changes and how organismal phenotypes collectively impact ecosystem function.

References

Johannsen, W. (1903) *Om arvelighed i samfund og i rene linier. Oversigt over det Kongelige Danske Videnskabernes Selskabs Forhandlinger*, vol. 3: 247-270. German ed. Erblichkeit in Populationen und in reinen Linien (1903) Gustav Fischer, Jena.

Madin, J., S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa (2007) An ontology for describing and synthesizing ecological observation data. Ecological Informatics 2:279–296. doi:10.1016/j.ecoinf.2007.05.004

Reilly, S., & Schachtman, T. R. (Eds.) (2008) Conditioned taste aversion: neural and behavioral processes. Oxford University Press.

Rohner, Nicolas, Jarosz, Dan F., Kowalko, Johanna E., Yoshizawa, Masato, Jeffery, William R., Borowsky, Richard L., Lindquist, Susan, Tabin, Clifford J. (2013) Cryptic Variation in Morphological Evolution: HSP90 as a Capacitor for Loss of Eyes in Cavefish. Science 342:1372--1375. https://doi.org/10.1126/science.1240276.

Walls, R.L., P.L. Buttigieg, J. Deck, R.P. Guralnick, and J. Wieczorek (2018) Integrating and Managing Biodiversity Data with the Biocollections Ontology, in Application of Semantic Technologies in Biodiversity Science, Anne Thessen, editor. IOS Press. ISBN978-1-61499-853-2 (print) | 978-1-61499-854-9 (online).

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. doi:10.1038/sdata.2016.18