How to identify drivers of variation within biological systems?

Noa Pinter-Wollman, Joan Segura, Richard Gawne, Michael G. Hadfield

Summary:

Differences among entities in a biological system are ubiquitous, creating ample variation in nature. Because variation is a cornerstone of evolutionary processes and because it allows persistence and resilience in changing environments, uncovering the drivers of variation is critical. *Here we outline some of the barriers and solutions to studying the drivers of variation in biological systems.* We discuss what variation is and how it can be measured. We then detail the challenges and opportunities that may arise when studying the drivers of variation due to the multi-scale nature of biological systems. The study of the drivers of variation will lead to a reintegration of biology by producing a multi-scale integration of our understanding of biological systems. It will further forge interdisciplinary collaborations and open opportunities for training quantitative biologists. We hope that our suggestions will lead to the formulation of new questions and analytic tools to study the fundamental question of what drives variation in biological systems.

Disclaimer: *This document was produced over 1.5 days of a workshop. The authors are aware that many statements made here could be followed with multiple references. However, due to the time constraints, no references were incorporated.*

Goal: Design a framework for identifying the drivers of variation in biological systems

Why is this important?

1. General concept across biology:

It is axiomatic that the habitats of all living organisms are variable: in climate, terrain and co-inhabitants. It rains or not, the temperature goes up and down, earthquakes shake, lift it or cause it to descend, sometimes even underwater, and microbes, fungi, plants and animals come and go. Living organisms, to survive in such changing habitats, must themselves be variable; that is, (1) they must include intrinsic (genetic) variability and tolerance limits or (2) they must, as individuals within populations, be different from each other in ways that allow survival of some members. Thus, within species -- or populations -- variation must exist both for the persistence of species and to provide the 'meat' for evolution into forms that will survive critical change in their living and non-living environments.

We thus find variation among the genes and genomes of individuals within populations and species. To varying extents, this variation is reflected in the ways individual organisms respond to changes (e.g., drivers) in their living and non-living environments, and determines if they survive at all (selection). Variation in heritable traits across populations also has a major impact on survivorship and thus selection. At the community level, genetic variations within the

genomes of individuals belonging to a focal species determine whether that species is present or not in some habitats. This is reflected in variations in the compositions of communities.

2. ***Variation is important for evolution:***

Variation is fundamental for the processes of evolution. Natural selection acts on heritable variation. Selective pressures lead to the persistence of individuals with certain traits, yet variation in the population is maintained, providing the population with resilience against future changes in the environment. The shape of the distribution of traits in a population can itself help maintain variation in a population, for example because of frequency-dependent selection, or because extreme phenotypes can have a disproportionate impact on the system as a whole. Furthermore, variation can be adaptive, but it might also be stochastic, e.g., drift generates variation that can lead to new strategies and new phenotypes. Studying the drivers of variation may help elucidate principles of evolution and diversification that could help us better understand the ways in which variation in a population is maintained. Natural selection acts at multiple biological scales therefore the study of variation has to happen at multiple biological scales.

Phenotypic variation in a focal trait is often limited when it is under selection. In the absence of selective pressures, variation in the developmental processes that give rise to the trait will tend to increase within the population. This developmental variation may, in turn, result in an increase in phenotypic variation at the population level. In this way, the absence of selection can sometimes allow for an increase in phenotypic variation.

3. ***Variation allows acclimation to a changing environment:***

The phenotypic outcomes produced by developmental processes are seldom fully pre-determined. Many organisms exhibit developmental plasticity, which is traditionally defined as the ability of a single genotype to produce different phenotypes in response to changes in environmental conditions. For example, differences in temperature or photoperiod sometimes lead to changes in an animal's coloration, or a plant's height. Developmental plasticity increases the amount of phenotypic variation that can be produced with a single developmental 'toolkit', and is often adaptive because it allows developing organisms to rapidly respond to changes in environmental conditions.

Environments are often heterogenous at both large and small scales. When the range of a species or a population spans across these diverse environments, individuals will sometimes experience very different selective pressures. This variation in selective regimes can produce adaptive phenotypic change. In this way, variation in the environment can be viewed as a driver of variation in phenotypes.

Box 1: Some definitions:

Driver - an underlying cause (e.g., a modifier, a mechanism, determinant etc.) of variation

Variation - differences that can be observed and described, more details below.

Biological system - a group of entities that interact and influence each other within the context of life.

What is Variation?

Variation is the differences in an attribute that can be measured or described. While in statistics, the term 'variance' refers to the spread of a measure from a mean of a population, here we use variation to refer to the nature of the differences between entities in a biological system or changes in the system over time; i.e., we use the term variation interchangeably with difference or diversity. These differences can be measured at different biological, physical, and temporal scales. Furthermore, variation is specific to the properties of a system (or its entities) or its surroundings or environment. Some examples of variation include biodiversity, phenotypic differences in a population, different alleles of a gene, etc.

The scale and sampling density at which one studies a biological system can influence the type of variation they can examine, as we detail in the next section. Furthermore, the magnitude and range of differences between entities that is detectable can differ across biological systems based on our ability to study them. For example, differences between entities might exist but the resolution of our measurement is not fine enough to detect such differences.

The study of variation requires one to define the entity, or scale at which differences are being examined and the system within which variation will occur. For example, when examining variation in a trait such as beak morphology of Galapagos finches or aggressive behavior of spiders, the entity would be a finch or a spider and the system would be a population of birds or a group of spiders; when examining variation in the response of cells to signaling molecules, the entity would be a cell and the system would be a tissue; when examining variation among trophic levels, the entity would be a species and the system would be an ecosystem.

The timescale over which differences among entities transpire can determine the nature and drivers of variation. Some variation may seem stable and in steady state. For example, the amount of variation in a population will not change over the lifetime of an entity. In contrast, variation can be dynamic, traits may change at a pace that is faster than the lifetime of an entity, leading to dynamics in the variation that characterizes a biological system. The transition between dynamic and static variation depends on the scale at which variation is measured and/or the nature/type of variation that is being measured. For example, traits that are plastic, such as the color of the fur of a snow hare can seem both dynamic and static - depending on when and for how long they are studied. In the fall, when hares change their coat color from brown to white, there will be a large amount of variation in coat color in the population because

individuals will differ in how quickly they replace their fur. Furthermore, the magnitude of this variation will change over time, and therefore seem dynamic. Once all hares in the population turn white, the variation of coat color in the population may seem stable, if it is examined only over a few winter months. Furthermore, in the midst of winter, what at first may seem like very little variation in coat color, my turn out to be a large amount of variation if measurement tools that allow differentiating between different shades of white are used.

More variation could be beneficial for some systems, and detrimental to others. There might be an optimal amount of variation for certain systems. For example, the mix of behavioral types in a honeybee colony can determine its collective foraging behavior. This collective, system level foraging behavior can be optimised when there is a particular amount of behavioral variation, i.e., colonies with certain behavioral composition out-perform colonies with a different amount of behavioral variation. The consequences of variation might determine how we study its drivers, as we detail below.

Key barriers:

1. ***How do we measure variation?***

There is no universal approach to measure variation across different scales or across different biological systems. For different biological systems and scales, different methodologies and protocols have been developed to measure variation, depending on their specific needs. The different technologies developed for these purposes cover ranges from molecular scales to populations of organisms. Therefore, the diversity of experimental protocols used to collect samples and the computational pipelines to process and analyze variation will comprise a disparate collection of resources. The lack of a general approach for measuring variation leads to the development of system-dependent approaches for measuring and analyzing variation. However, depending on the magnitude or attribute data type that is being measured we can find similar statistical analysis tools and data processing pipelines that could be integrated in a common framework.

2. ***The scale problem:***

As we noted above, the study of variation and its drivers is subject to biological scales. First, a driver of variation at one scale may not seem like a driver of variation in another scale. For example, drift can cause variation among populations but eliminate variation within a population. Second, variation at one biological scale may, or may not, influence variation at another scale. For example, mutations in a gene at the molecular level, often leads to variation in a phenotype at the organismal level. However, a mutation in a non-coding region of the genome will produce variation among genomes, but these differences will not result in variation among organisms in an observable trait.

The fact that variation at one level can have consequences that reverberate to higher and lower organizational levels is a challenge because it forces us to rethink the tendency to rely on simplified scientific explanations that aim to explain a focal phenomenon by pointing to a single

causally relevant factor at a single scale. At the same time, the fact that biological scales are interlinked can be viewed as an opportunity because it allows us to step outside of our narrowly defined areas of professional specialization, and collaborate with colleagues in other areas of biology that we might not currently interact with on a regular basis. The potential for causal feedback between variation at different scales allows for an integration of research programs that focus on a multi-scale perspective of biology.

Considering the multi-scale nature of biology and the potential cascading effects of variation at one scale onto the performance of processes at a different scale, the search for drivers of variation needs to account for the question of scale. First, we need to determine how 'deep' we want to probe to identify a driver of variation. For example, identifying a mutation in a genome that led to variation in a certain morphological trait might sometimes be a sufficient explanation for the driver of variation in the morphological trait. However, we might want to probe further and deeper into the molecules that form the genes, patterns of polarity, cell-cell interactions, tissue-specific responses to signaling molecules, or competition between anatomical characters. Alternatively, we might want to examine top-down drivers of variation in the morphological trait, such as the environment and selective pressures that shape and differentially select between the proximate mechanisms mentioned above. The consequences of variation in a system may influence what drivers we search for and where (i.e., at which scale) we search for them.

Finally, one might identify common drivers that act on variation at multiple scales. For example, differences among alleles of a gene are by definition variation at the scale of the genome. They are also drivers of variation in the phenotypic traits that they code for. This variation among phenotypes can lead to variation among groups of organisms, different populations, or across entire ecosystems. Similarly, variation in the environment can drive variation among populations, organisms within a population, and plastic developmental responses such as circulating hormonal titres. Thus, variation at one scale might influence variation at many other scales or in many systems within a certain scale.

3. The causation problem or, what is a driver?

As with any study, correlation does not imply causation and statistical approaches must be employed to resolve the direction of a relationship to determine what is a driver, and what simply co-varies with variation. In many cases, statistical analysis of experimental measures will not definitively resolve the direction between driver and variation and thus, driver and variation could not be distinguished from each other. For example, in a colony of organisms, the proportion of a particular type of organism might be highly correlated with the pH of the environment. However, it might not be possible to determine whether the pH is indeed the cause of the observed variation or, for example, a result or the biochemical processes of the organism.

In addition, certain types of variation might be explained by more than one driver. Finding out how many drivers are underlying a particular type of variation and their relative importance could be a challenge that opens up opportunities for discovery. Drivers may further interact with one another and disentangling their effect could be a challenge worthy of investigation.

4. Resources

Clearly, studies of genetic/organismic/population/community variation are very different in their spatial focus, and instrumental requirements. Some may be totally laboratory based, but even the shapes and spaces will vary with the questions being answered. Bacteria can be studied on agar-filled Petri dishes (and require autoclaves, centrifuges, etc.) in relatively small labs, while studies of tigers may require access to forests via travel (even international), cameras and other detection devices. The bases of costs, as well as their magnitudes, will vary among research questions and the approaches required to answer them. Thus, costs alone will be a limiting factor in research. It will be cheaper to study mice than tigers, but choosing only mice will leave many scientific questions unanswered.

The availability of individuals to carry out research, including undergraduates, graduate students, postdocs and PIs, is often limiting. Each career stage presumes a certain amount of training, and the costs of that training. Thus, support for training of "resource" individuals is a barrier to be faced by funding agencies (e.g., the NSF), institutions (from community colleges to R1 universities), departments, and faculty.

While the warehouses of supply and technical companies are filled with a huge spectrum of chemicals, reagents, antibodies, and equipment to measure, analyze and visualize just about anything, new needs still arise for very specialized instruments for new investigations. These needs may arise most often when the biologist wishes to measure things at very large scales: e.g., frogs in the Amazon jungles. Development of drone technologies, environmental DNA monitoring, and remote sensing may all be brought together, even integrated, for such endeavors. Needless to say, costs will be large, and peer-reviewers of research proposals will have to be convinced of the importance of the research. Thus, essential "resources" may present demands at many levels, including physical, financial and human interest.

Suggestions for overcoming barriers

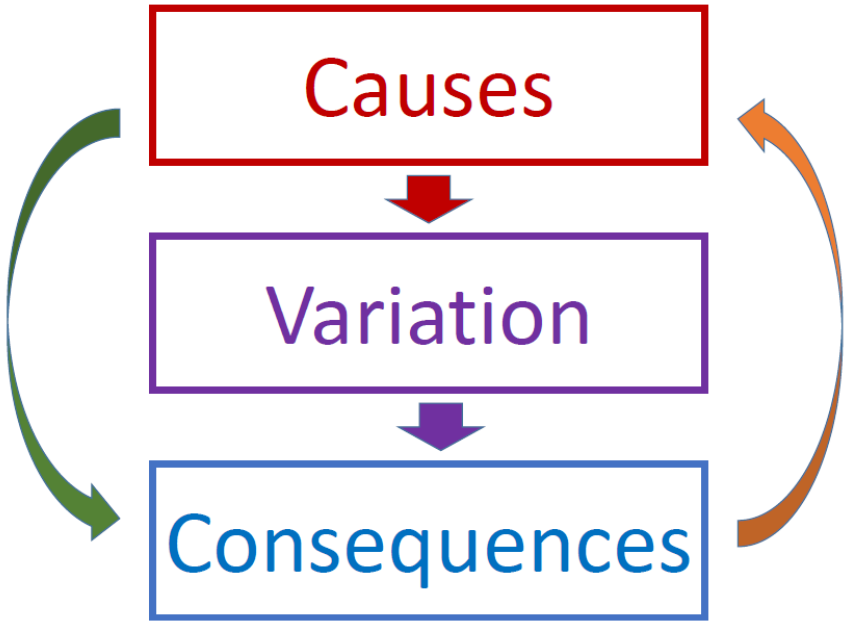
1. The feedback solution:

One solution we propose to help tackle the barriers we listed above is to study the drivers of variation within the context of their consequences. While our goal focuses on detecting the drivers of variation, often, variation also has consequences. These consequences can help guide researchers to the types of drivers they might seek. For example, if the goal is to study the evolutionary consequences of variation in a population, one might choose to focus on probing the heritable aspects that underlie variation. Alternatively, if one is interested in the effects of variation on how populations of different species interact, they might focus on studying the environmental (physical and/or social) features that drive variation.

Furthermore, there is feedback between the causes (drivers) and consequences of variation that could help guide the way in which variation is studied (Figure 1). Drivers at one scale might be influenced by forces and consequences at a different scale. For example, genetic variation changes as a function of selection on the traits that the genes code. This feedback across

scales can help create an integrative, multi-scale approach to the study of the drivers of variation.

Figure 1: Considering feedback between the causes and consequences of variation may help uncover the drivers of variation in a multi-scale framework.



2. ***Design a scale-specific framework:***

Although we cannot describe a universal solution or develop a general approach that might measure variation in any scale or biological system, we can provide a methodology to identify proper statistical approaches and processing pipelines for collection of data and their analyses. Variation can be classified in terms of its data type in order to indicate what set of statistical tools and data-processing pipelines might be more useful for its analysis. In this way, variation can be classified as numerical or categorical. At the same time, numerical variation can be continuous (*e.g.,* differences in molecular weight) or discrete (*e.g.,* number of mutations). Categorical variation may include any measure that takes one of a limited number of values (*e.g.,* different atom types, residue mutations, color). Depending on the data type, there are different statistical approaches to quantify variation. A partial generalization will include the design and development of a statistical framework where specific approaches to measure and find variation will be integrated and classified in terms of data types. This will not only offer a centralized statistical platform integrating relevant tools but also a guide to what methods to use for different scales and biological systems.

At this point, it is worth noting that the first bottleneck that any statistical analysis might face is a poor relation between signal and noise; *i.e.,* the accuracy of any statistical pipeline primarily depends on the quality of the experimental sampled data. Therefore, knowledge of what experimental features are relevant for detection and measure of variation and the right tools for

its accurate measure will be the most important step in the analysis pipeline. This critical point depends mainly on the studied biological system itself and the scale at which variation is measured; it is the main reason why a universal method it is not proposed.

3. Use perturbation to get at mechanism

Very often variation is detected, but the cause of it is uncertain. A common approach to sorting the sources of variation is to alter or change the suspect of cause. For example, we could change the nutrients in the media for a bacterium, light exposure for a plant, or species composition in a community. If the predicted variation does not emerge, we must try another suspected factor and perturb its occurrence, abundance, or character. This process is clearly iterative, as has been experienced by most biologists.

A type of 'natural perturbation' are constraints. Examining the constraints on the amount of variation in a system can assist in investigating the drivers of variation. Uncovering the physiological, chemical, and environmental constraints on the range of variation of traits can assist in identifying the drivers of variation and their scope. For example, an organism can only reach a certain maximum size, implying that there is a threshold to variation. Identifying the drivers that limit the size of an organism can assist in identifying the causes that underlie variation in size.

Exciting scientific opportunities, or what is the potential impact?

Scientific education tends to be very broad initially, but over time professional pressures encourage researchers to become extremely specialized, focusing on a particular organism, or biological scale at the expense of others. This results in a compartmentalization of knowledge, and a lack of cross-talk between fields that could hinder scientific progress. Many biologists would agree that there is a need to re-integrate the field, but it remains unclear how this should be done. One way to proceed that acknowledges the interconnectedness of the various sub-disciplines of biology, and allows for fruitful interactions between seemingly disparate research programs is to focus on describing, and identifying the causes of variation across scales.

The research program we have described will allow biologists to step outside of their narrowly defined professional specializations by enabling them to seek collaborations with colleagues in other areas, and actively encourage them to pursue projects that do not neatly fall into existing disciplines. The fact that the proposed program will facilitate truly integrative research will allow researchers to view existing data in a new light, and encourage the formulation of new questions that might not otherwise be posed. This, in turn, will lead to new experimental designs that have the potential to spur novel scientific advances.

In addition to bringing disciplines together, the study of the drivers of variation will advance our general understanding of biology. As we detailed above, the study of variation will require a multi-scale approach to integrating studies across systems. Such an integration requires not only breadth of thought and expertise but also new tools and analytical approaches that still need to be developed. The importance of uncovering the drivers of variation is clear when

considering the changing world in which we live. Systems can go through abrupt ‘tipping points’ or phase transitions, that can lead to their collapse. Understanding the causes of variation can aid in identifying early indicators of tipping points because potentially a reduction in variation or the rise of particular variants may expedite certain feedbacks that eventually lead to collapse. Thus, the consequences of variation are wide reaching and uncovering the drivers of variation is fundamental for reintegrating biology.