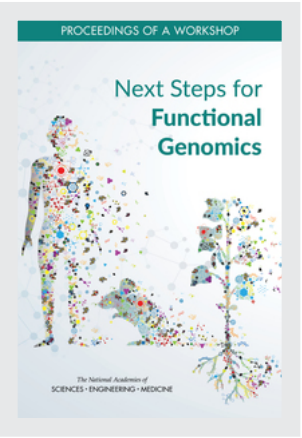

Next Steps for Functional Genomics: Proceedings of a Workshop (2020)

DETAILS

180 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-67673-1 | DOI 10.17226/25780

CONTRIBUTORS

Robert Pool, Steven M. Moss, and Frances Sharples, Rapporteurs; Board on Life Sciences; Division on Earth and Life Studies; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2020. *Next Steps for Functional Genomics: Proceedings of a Workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25780.

PREPUBLICATION COPY

Next Steps for Functional Genomics

PROCEEDINGS OF A WORKSHOP

Robert Pool, Steven M. Moss, and Frances Sharples, *Rapporteurs*

Board on Life Sciences

Division on Earth and Life Studies

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by the National Science Foundation under Award Number 1927620. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-XXXXX-X
International Standard Book Number-10: 0-309-XXXXX-X
Digital Object Identifier: https://doi.org/10.17226/25780

Additional copies of this publication are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2020. *Next Steps for Functional Genomics: Proceedings of a Workshop.* Washington, DC: The National Academies Press. https://doi.org/10.17226/25780.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

PLANNING COMMITTEE ON NEXT STEPS FOR FUNTIONAL GENOMICS: A WORKSHOP

Members

GENE E. ROBINSON (NAS, NAM), *Chair*, University of Illinois at Urbana-Champaign
PHILIP N. BENFEY (NAS), Duke University
CHARLES DANKO, Cornell University
EMMA FARLEY, University of California, San Diego
TRUDY F. C. MACKAY (NAS), Clemson University
TERRY MAGNUSON (NAM), University of North Carolina at Chapel Hill
LAUREN O'CONNELL, Stanford University
ANDREA SWEIGART, University of Georgia

Staff

STEVEN M. MOSS, Associate Program Officer, Board on Life Sciences
FRANCES SHARPLES, Board Director, Board on Life Sciences
KOSSANA YOUNG, Senior Program Assistant, Board on Life Sciences

BOARD ON LIFE SCIENCES

Staff

Reviewers

This Proceedings of a Workshop was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published proceedings as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this proceedings:

JULIA BAILEY-SERRES, University of California, Riverside
EMMA FARLEY, University of California, San Diego
MARC HALFON, University at Buffalo-SUNY
STEVEN HENIKOFF, Fred Hutchinson Cancer Research Center

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the proceedings nor did they see the final draft before its release. The review of this proceedings was overseen by **JASPER RINE,** University of California, Berkeley. He was responsible for making certain that an independent examination of this proceedings was carried out in accordance with standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the rapporteurs and the National Academies.

Contents

1

Introduction

It has been 25 years since the first complete genome of an organism was sequenced—that of the bacteria *Haemophilus influenzae* in 1995 (Fleishmann et al., 1995). It was followed quickly by the genomes of other bacteria and then organisms of increasing complexity—the nematode *Caenorhabditis elegans* in 1998 (*C. elegans* Sequencing Consortium, 1998), the fruit fly *Drosophila melanogaster* in 2000 (Adams et al., 2000), the laboratory mouse *Mus musculus* in 2002 (Chinwalla et al., 2002), and humans in 2004 (International Human Genome Sequencing, 2004). The ability to sequence whole genomes of organisms—a process that has become faster and less expensive with each passing year—has opened a golden age of biology, with researchers discovering much that had previously been out of their reach.

However, as pointed out by Donal Manahan, Director of the Division of Integrative Organismal Systems at the National Science Foundation (NSF), even after decades of intense work, researchers still do not understand the functions of all of the genes in any organism and exactly how the information in genes is transformed into an organism's phenotype, its physical characteristics. "Why has this been so complicated?" he asked. "It's fair to say that this might be viewed by many as the biggest gap in biological knowledge—our inability to predict the phenotype of a cell or an organism from what we know about its genome and its environment."

Manahan was speaking at the opening session of a workshop at the National Academies of Sciences, Engineering, and Medicine in Washington, DC. Held from February 10–12, 2020, the workshop was sponsored by NSF with the goal of determining the current state of of the science of functional genomics. Functional genomics was defined in the context of this workshop as the biological field concerned with understanding the connection between the information contained in an organism's genome and its physical characteristics. Understanding phenotypes from genomic information is increasingly done with broad genome-wide-, or "-omics-," style approaches. The workshop aimed to understand what is needed for the field of functional genomics to move forward (see the Statement of Task in Appendix A).

IMPORTANCE OF FUNCTIONAL GENOMICS

In his remarks Manahan mentioned an interesting editorial published in *Nature Genetics* titled "A Focus on Function" (2000). The article recommended a focus on developing a "molecular intelligence," the ability of researchers to understand and predict how genes contribute to physiological processes. But what is most striking about the article, he said, is that it was published 20 years ago, in July 2000. So even at the beginning of the 21st century, there was a community conversation about function. "And here we are 20 years later hopefully narrowing that down in many ways."

NSF funded the workshop, Manahan said, because it is interested in obtaining guidance from the scientific community concerning how to move the field of functional genomics forward. The information NSF sought from functional genomics researchers "will help us develop new

approaches as a community, to address how best to develop a predictive understanding of properties of living systems."

Gene Robinson of the University of Illinois at Urbana-Champaign, who chaired the workshop's organizing committee, then offered some additional background concerning the workshop. The field of genomics has two main thrusts: discovery genomics and functional genomics. Discovery genomics is "Roaring ahead," he said, and is well into the 21st century. Functional genomics, by contrast, has barely made it into this century. Most of the tools and techniques available in functional genomics allow researchers to work only at the level of single genes or small numbers of genes, while researchers in discovery genomics have the tools to see patterns at a much larger scale. It is one of the holy grails in biology, Robinson said, to be able to predict function from genetic sequence, including understanding the role of the environment and its interaction with the genome. "In order to be able to do this, we need these large-scale, discovery-oriented projects," and those in turn will require the development of much better tools. "That's what we're calling functional genomics."

WORKSHOP OUTLINE AND OBJECTIVES

Robinson then listed the workshop's objectives for the audience. They were to

- Understand the successes and failures in functional genomics research in a variety of research organisms—what tools helped investigators succeed and what tools still do not exist or need development for future research success.
- Discuss considerations for selecting experimental systems as well as research approaches that leverage all functional genomics–related scientific disciplines.
- Understand research strategies for determining factors (genetic, epigenetic, environmental, etc.) that influence phenotype.
- Discuss possible "rules of life" to guide baseline and comparative questions across the different realms of microbes, animals, and plants.
- Look at advantages and disadvantages of currently available consortia and databases. Also discuss emerging tools and databases that might not be widely available.
- Think critically about the training needs for future genotype-to-phenotype researchers.
- Discuss short- and medium-term research and knowledge goals, noting potential strategies to reach these goals.

With that, Robinson provided a quick outline of how the workshop would proceed. The first day would begin with a keynote address designed to set the stage for the rest of the workshop. It would be followed by individual sessions that focused on case studies of building and using functional genomics tools, defining "model systems," understanding the contributions of non-protein-coding DNA to phenotype, and the societal and ethical implications of functional genomics research.

This proceedings was prepared by the workshop rapporteurs as a factual summary of what occurred at the workshop. The planning committee's role was limited to planning and convening the workshop. The views contained in the proceedings are those of individual workshop participants and do not necessarily represent the views of all workshop participants, the planning committee, or the National Academies.

The organization of this proceedings roughly follows the organization of the workshop itself. Chapter 2 contains a description of the first keynote talk by Aviv Regev of the Broad Institute and the Massachusetts Institute of Technology. Chapter 3 reports on a series of presentations and the ensuing discussion on case studies of building tools for use in functional genomics research in various organisms. The talks reported in Chapter 4 focus on understanding the contributions of non-protein-coding DNA to determining phenotype, while those in Chapter 5 describe research aimed at understanding the environmental regulation of gene function. Chapter 6 recounts the workshop's second keynote presentation, this one by Patricia Wittkopp of the University of Michigan. Chapter 7 covers a session that was devoted to interpreting and validating results from high-throughput screening approaches, while Chapter 8 describes presentations and a panel discussion focused on large databases and consortia. Chapter 9 describes several big-picture challenges in functional genomics research: education and training, determining and defining "model organisms," and the social and ethical implications of functional genomics research. Chapter 10 offers a brief wrap-up of the workshop and a look to the future.

2

The Genotype–Phenotype Challenge

"After the Human Genome Project one of the big promises was that we would be able to map genes for different types of diseases," Aviv Regev of the Massachusetts Institute of Technology told the workshop audience as she began her keynote address. Indeed, she continued, some 100,000 different regions in the human genome have now been associated with a range of diseases, everything from inflammatory bowel disease to schizophrenia. This has been a truly remarkable achievement by the human genetics research community, she said, but at the same time that achievement points to one of the biggest challenges facing that community today: How does one take these 100,000 variants scattered around the genome and understand what they do—first at the level of cells and then at the level of tissues, then organs, and then, finally, across entire humans beings?

Regev provided a big-picture view of the challenges facing the biological researchers attempting to understand the genotype–phenotype connection—that is, how the information coded in the genome leads to the physical characteristics of an organism. To do this she detailed results from various lines of research and described the multi-level approach that she argued will be necessary to discover the "rules of life."

OVERVIEW OF THE CHALLENGE

Regev began by discussing why it has proved so difficult to move from genotype to phenotype. The overarching reason, she said, is because biology is the integration of multiple different levels of organization. The fundamental unit of life is the cell, which is made up of molecules that are either encoded directly by genes or else obtained through chemical reactions that gene products catalyze. These cells interact with each other, and in multi-cellular organisms they combine to make tissues, which in turn are assembled into organs, which account for an individual's physiology. Increasing the complexity, individuals make up populations, which are part of ecologies, and it is at the ecological level that natural selection acts.

Research biologists interested in understanding the rules of life ask three different kinds of questions, Regev said. There is the "What?" question, which concerns structures. There is the "How?" question, which concerns genetics and mechanisms. And there is the "Why?" question, which is a functional question. Each of these questions must be asked at each of the organizational levels, including the levels of genes, tissue structure, physiology, and evolution.

The resulting complexity makes it difficult to answer any of these sorts of questions, Regev said. In addition, the interactions between the levels are complex and frequently nonlinear. This nonlinearity in particular made mapping and understanding these systems impossible for a long time.

"So it doesn't really matter in which direction you look," she said. "In theory, for each of these problems—and many other problems—the space of possibilities is enormous." Regev noted that the main issue is that researchers do not know, upfront, which connections matter and which do not.

The question then is, how does one study something systematically if it will never be possible to study it exhaustively? "So what I'm going to claim here today through many different examples," Regev said, "is that in this case the bigness of biology is no longer our problem. It's actually one of the best opportunities that has ever happened to us. We couldn't really seize it in the past. We didn't have the tools. But we do today."

In the past the basic approach was for researchers to look for intelligent ways to limit "the bigness" because the tools were not available to deal with the entire search space. So instead of testing all possible mutated sequences, for example, one might probe for those mutations that were known to exist. Researchers might look for interactions between genes that are known to likely interact to avoid testing all possible genetic interactions. Or instead of looking at sequences in general, design assays to look at ones that exist in nature or that relate to existing models.

"These are all great approaches, and they have actually served us incredibly well," Regev noted, however, that "they don't fully solve our problem. We're not sure that what we get after we limit the search space is a general answer or a rule of life."

In a search for a different approach, Regev offered a metaphor related to current approaches. Imagine that functional discovery in biology is like trying to figure out which painting is hidden behind a sheet of white tiles and every experiment in biology is like removing one small square from the sheet to see what is behind it. "Until recently," she said, "all we could do was remove a few tiles, so we had to focus." If, for instance, something interesting—a bit of red, say—was found by pulling off a tile, then one might uncover more tiles around that. That could lead to uncovering little patches in a sea of white where something is known about part of the underlying painting. But the painting as a whole remains a mystery.

If one uncovered the same number of tiles randomly, it would be hard to determine anything about the painting because most of what was uncovered was not interesting—background or some equally uninformative part of the painting. "But what if I were allowed a lot more tiles?" Regev asked. Then even if they were done randomly, one might well be able to discern enough pattern to recognize the painting underneath.

Doing this type of pattern recognition in experimental biology is not typical, Regev said. "It's a great new opportunity—and one for which we're already reaping benefits, as I'll try to show you today, in many different ways." The remainder of her talk consisted of examples related to cells, programs, and mechanisms.

CELLS

Cells are a key intermediate between genotype and phenotype, Regev said. For example, even though every cell in a body carries the same genetic variants that confer the risk of a particular disease, the disease will typically manifest only in those cells that express the gene product or are affected indirectly by it. "Knowing our cells is going to be essential for understanding gene function in humans," she said.

Subsequently, it is a problem that scientists do not really know how many different cells there are or what their molecular characteristics are. Historically, cells have been categorized in many different ways—according to their structure, their location, their function, and so on—so there is not a unified way to think about all of them. One way to address this would be to come up with a "map" or an "atlas" that had a unified set of coordinates for all human cells, Regev suggested. Gene expression levels could provide such a collection of coordinates so that each

cell would be a point in an extremely high-dimensional (20,000 or more dimensions) space, with one dimension for each gene.

This would not have worked in the past when it was not even possible to measure gene expression levels in single cells. In the past few years, breakthroughs in single-cell genomics have made it possible to measure expression, chromatin, and other molecular profiles in large numbers of individual cells. One technique in particular—single-cell RNA sequencing—has dramatically increased in recent years in the number of cells that can be examined, going from just over a dozen in 2011 to 30 million in 2019.

This capability makes it possible to carry out experiments that, in terms of Regev's painting metaphor, seek to reveal many random pieces of the big picture—an approach she calls "design for inference." She and her team decided to develop methods "that favor sparse and noisy data from massive numbers of cells over much richer and more precise data from a small number of cells"—in essence, uncovering lots of random tiles rather than focusing on one or two areas of interest. The reason for this decision, she said, was that her team knew they could handle the sparsity of data and data resolution, per cell, because gene expression and other characteristics of cells are highly structured both within and across cells. She noted that structure is key because it is possible to gain a lot more information from different types of data than one might expect.

That experimental design has proved successful, Regev said, although "it was extremely uncomfortable for the experimental biologists initially." Their experiments looked for patterns and structures in sparse and noisy data and uncovered a number of discrete cell types. In addition, they can look at cellular changes over time, such as through development, and how the cells respond to environmental changes. Regev noted that they "can even see the imprint of where a cell is located—its anatomical position—in the fine histology or the tissue structure that it's in and even the direct cellular neighbors that it has."

What is most important to understand about this is that even though most of her group's computational analysis methods are aimed at capturing one aspect at a time, the actual cell is all of those different things and identities at once. "It has a type and a location and a history and possible fates. It's always undergoing multiple transitions all at once. . . . And all of these aspects interact with each other."

The bottom line, she said, is that thinking about cells as the basic unit of biology can be useful for many purposes, but it can also be a limited representation because many of the behaviors that cells have do not obey those of the defined cell type. For that reason, she said, it is often helpful to carry out analyses where gene programs are a fundamental unit, rather than thinking only in terms of cells as the basic unit for analysis and understanding.

PROGRAMS

Gene programs are important in several ways, Regev said. First, they make it possible to better describe and understand cells that span the spectrum and do not obey the typical behavioral boundaries. Second, they help in studying the function of genes, that is, its phenotypic mapping. Finally, simply knowing that genes form structured programs helps with such problems as genetic interactions, which otherwise might appear intractable.

Understanding Cell Categories Using Gene Programs: An Example of Innate Lymphoid Cells in Psoriasis

She illustrated the idea that gene programs help describe cells that cannot be discretely categorized with an example related to the role of innate lymphoid cells (ILCs), a type of immune cell, in psoriasis. Initially in studies of psoriasis, it seemed that there were two distinct types of ILCs, labeled ILC2 and ILC3. ILC3s are thought to be the "first responders" that signal an immune response to T cells, which subsequently causes psoriatic skin. The issue, as Regev says, is that when you look at healthy skin, there are no ILC3s present, only ILC2s, which leaves the enduring question of where the ILC3s come from. While it is possible that these cells circulate from other parts of the body, or are too rare to be noticed, the Regev lab wanted to shed some light on the situation and did so through work with a mouse model.

When Regev's group investigated ILCs with single-cell RNA sequencing, they found that the ILCs were not discrete cell types but rather spanned a range of continuous cell states. This is difficult to capture when assuming that cells are the basic unit instead of considering the gene programs (Bielecki et al., 2018).

"So how can we capture this distinct biology?" Regev asked. To analyze what was happening in the cells in terms of gene programs she turned to an approach that is used in text analysis. To explain text analysis, she offered the example of a news article about the restaurant Chef Gordon Ramsay. "Now, this piece is related to different topics," she said. "There's food, there's business, and there's celebrity culture." And the words in the document will reflect the different topics. Some words, such as the names of dishes, are related only to one area—in this case, food and cooking. Others can be tied to multiple topics; "Gordon Ramsay," for instance is related both to the topic of food, because he is a chef, and to the topics of business and celebrities. The key point is that one can analyze the words in the article—independent of the meaning of the individual sentences and paragraphs—to get information about the topics covered in the article.

"In the same way that words in a document result from topics that the document covers, even though no one tells me up front what these topics are," Regev said, "gene transcripts in the cell are related to programs or processes in that cell even though we don't know these topics or programs up front." What is important is that by working from the gene transcripts in a cell, one can use various computational methods to capture the programs that are likely taking place in the cell. In her analysis of the role of ILCs in psoriasis, Regev used a particular method called weighted allocations. "They capture the topic [program] as a probability distribution over the genes where each gene has a weight of belonging to the topic," she explained, "and then each topic has a weight of being in the cell."

The analysis indicated that there were "co-expressing cells" that had both ILC2-like and ILC3-like programs or features and that one could better understand the ILCs in terms of the different programs running in the cells rather than in terms of different cell types. These co-expressing cells are poised to move in either the ILC2 or the ILC3 direction, depending on various epigenetic triggers (Bielecki et al., 2018). The lab was able to validate this result by looking at the single-cell expression profile under different conditions as well as testing it in a mouse model.

Thus, Regev concluded, "genetic programs are helpful in thinking better about the functionality of cells. They are just more fluid and more flexible than putting everything in these discrete categories."

Defining Function Through Gene Programs

The second advantage of programs that Regev discussed was in thinking about the functionality of genes. To illustrate this, Regev first described some of her work on ulcerative colitis, a form of inflammatory bowel disease (IBD). IBD is the "poster child" of human genetic studies because researchers have used genome-wide association studies (GWASs) to identify hundreds of loci that are associated with the disease, and for the vast majority of these, researchers do not know the function of the associated genes.

To study the genes responsible for ulcerative colitis, Regev and colleagues examined gene expression in different cell types and in particular looked for cells that were enriched for risk genes for ulcerative colitis (Smillie et al., 2019). Once they had those data, they began looking for patterns in gene expression. "One way of predicting gene function is asking which other genes co-express with it," she said. Because they had gene expression data by cell type, they could look for genes that were co-varying within the specific cell types in which those genes were expressed. In essence they were looking for gene programs to lead them to clues about gene function.

The result was a collection of gene modules made up of a set of genes whose expression co-varied in specific cell types, and, as it turned out, most of the modules they identified consisted of multiple genes identified by GWAS—that is, most of the genes that co-varied with a GWAS gene in a particular cell type were themselves GWAS genes. Out of about 100 different GWAS genes that had been implicated in risk for IBD, Regev and her colleagues formed about 10 different modules that include more than half of the GWAS genes in a cell-type-specific manner. These were gene programs implicated in ulcerative colitis.

The knowledge that such structures exist is useful in itself, Regev said. For example, since GWAS genes tend to aggregate with each other in modules, when examining candidate genes one could give preference to those candidates that co-vary with previously mapped GWAS genes or even just with other candidates.

Programs can also be used more directly in examining gene function, Regev said—not just by relying on inference, as in the work with ulcerative colitis, but through direct experiments. For example, she described genetic screens that make it possible to find all genes that individually can affect the expression of a particular target gene, call it Gene X. A great deal of biology has been learned from this technique, she said, but it does have some limitations. For one thing, there is a simple readout for each cell—the level of just one gene, Gene X. This means that one must know how to choose Gene X in advance of the experiment. It also means that all hits are going to look the same, because they all affect the level of Gene X. That makes it difficult to capture complex biology.

Regev described a new assay technique designed to get around these limitations called Perturb-seq, which involves pooled CRISPR screens with single-cell RNA-seq (RNA-sequencing) readout. One of its first applications was in dendritic cells, a type of immune cell, stimulated with lipopolysaccharide (LPS), a molecule found in the cell membrane of bacteria. What they found was the genes whose expression changed in the analysis partitioned into five programs (Dixit et al., 2016). "That means that all of the genes in the program are affected in a similar way across the perturbations," she explained.

What is particularly important about the Perturb-seq technique is that the presence of the programs makes the assays very scalable. For program-level effects, Regev said, it is enough to have as few as 30 to 50 cells per perturbation and a few hundred reads per cell. "The rest is just

given to you by structure." This in turn means that the screens can be done for many different purposes and will produce a unified readout. The technique also can be carried out with coding or noncoding variants in one cell type or in multiple cell types simultaneously.

Regev illustrated the power of Perturb-seq with a recent example where she used the technique to characterize the potential function of 35 genes with loss-of-function variants in autism (Jin et al., 2019). These are genes that are known to play a role in autism from human genetics, but researchers know nothing about the specific roles they play, the cell types in which they act, or the processes by which they work. "It's really hard to decide even which screen you should devise for them," she said. "So we devised this screen you do when you don't know anything."

Testing 35 genes known to be implicated in autism spectrum disorder against five major cell types, they first examined the effects of individual genes on individual cell types and found very little. Only 1 of the 35 genes, *dyrk1a*, had any significant phenotypic effects.

But things looked different when they turned to the level of programs. Examining which programs were affected by the perturbations, they found that 15 of the 35 autism genes affected six programs across four different cell types. "This highlights that there are probably a limited number of cellular processes crossing different cell types that these genes actually converge into," she said. She added that this was an early screen done with a relatively small set of genes, but experimental improvements have now made it possible to do very large—and even genome-wide—screens.

Using Structured Programs to Understand Genetic Interactions

The third way that genetic programs can be of assistance is that simply knowing that genes form structured programs helps with such problems as understanding genetic interactions. For example, Regev said, "we can use this knowledge that there are expression programs not just to change our analysis of data we already measured, but also to change how we do measurements in the first place." This was, for instance, why she knew that she could get useful results from sparse and noisy data from a large number of cells—because there was an underlying structure.

"But we got greedier and greedier with time," she said. And they came up with a simple idea: Instead of measuring the expression of individual genes, they would measure what they called "composite genes," which were linear combinations of small subsets of genes. They could then use a mathematical technique called decompression to make individual gene measurements. The approach depended on the assumption that gene expression was structured—if it is structured, then the expression profile can be described by the linear combination of a small number of modules.

Regev described an application of this approach in the context of spatial imaging where they were limited in the number of measurement channels, "so it's a big deal if you can get rid of that limitation and get information about more genes without more experiments." They created composite genes by mixing probes against different genes but with the same label. The experiment was done in mouse motor cortex with nine composite genes, each including from 8 to 13 individual genes out of the 37 total genes covered by the study. They were able to successfully decompress the 9 composites to get patterns for all 37 genes (Cleary et al., 2017).

Summarizing her comments on programs, Regev said, "I showed you that many biological processes are best captured at the level of programs, not cells. This gives us an ability to handle spectrums of programs . . . and solves the problem that no single partitioning of cells is going to

capture all perspectives of biology." In the case of human genetics, most genes captured by GWAS are expressed in specific cell subsets, and they map into modules that vary within these subsets. The modules also help predict GWAS gene function. Programs provide multi-purpose, rich, robust, and diverse readouts for scaled pooled screens, and knowing that a structure exists makes it possible to perform more efficient experiments using compressed sensing.

MECHANISMS

To introduce her third topic, mechanisms, Regev asked the question, "How far can we go in chasing the really difficult problems?" There are certain problems in biology that seem to be so large and complicated that there are not enough cells in all of the humans in the world to do the necessary experiments and, indeed, where the number of necessary experiments is more than the total number of atoms in the universe. "So generally I think we all assume that these problems can never be fully tackled," she said. "And that might still be true, but I think there's some room for optimism." She illustrated her point with two examples of how such problems can be addressed by thinking about them differently.

Creating Models to Understand Gene Expression

The first problem she discussed was predicting how genetic sequence controls levels of gene expression. The basic approach is to work from examples where there are known regulatory sequences and associated expression. As an approach, they have taken the sequences and associated expression data from individuals in a population for many years. Or else researchers design the sequences, usually starting with something from nature, and modify them according to some understanding or hypothesis. With massively parallel reporter assays it is possible to get tens of thousands to hundreds of thousands of examples to work with.

The issue is that even though that sounds like a lot of data, it is not enough to take full advantage of machine learning, Regev said. She went on to ask how a much bigger dataset could be generated. One straightforward way would be to work with random sequences of DNA, which are available commercially as training data. This should work, Regev said, because "transcription factors bind in short degenerate sequences, and most transcription factor binding sites should exist in random DNA." According to a calculation she and her colleagues made in 2009, one transcription factor motif should appear in every 1,500 base pairs, so if one analyzed a library of 10 million 80-base-pair sequences, there should be more than 500,000 such motifs.

With this in mind, Carl de Boer, who was working in Regev's lab at the time, devised a simple assay: "You would measure the extremely noisy expression level of hundreds of millions of sequence examples (de Boer et al., 2020). For most of them we would get exactly 0 or 1 estimate of their expression level. You will see them not at all or once." It was an easy experiment to do, she said. "You put these in cells. You see how much expression they drive."

The result was a huge amount of data that could be used to train a very complex model, Regev said. In particular, the model she and her team worked with was mechanistic, with many biological details, designed to be interpretable. Such models can provide detailed mechanistic insight. "For example," she said, "they highlight very precise ways in which regulatory proteins might interact with DNA." Most importantly, the model and all the data that went into it made it possible to look at much smaller effect sizes than had been possible before—with some surprising results. "These models showed—and we confirmed this with experiments—that weak

interactions, which are usually completely ignored by models, . . . are actually a predominant way in which regulation happens."

Besides offering insight, a second main use of models is to predict, and Regev's team built a second model whose main purpose was to offer predictions rather than any insight into actual mechanisms. It worked well, Regev said. It accurately predicted the expression of both random sequences and native sequences from yeast (*S. cerevisiae*), their research organism. In particular, it predicted expression for random sequences with 98 percent accuracy and for native yeast sequences with 92 percent accuracy (de Boer et al., 2020). "It means we can now use the model based on random sequence to design sequences that have desired properties," she said. "And when you do that, you can ask for ones that give you particularly high or particularly low expression." Furthermore, she added, "if it's as predictive as that, you might start thinking that you have a full landscape of how sequence maps to expression, and if expression maps to fitness, you can say something meaningful about evolution."

Using Mechanisms to Understand Genetic Interactions

The second problem she discussed that can be approached in terms of mechanisms was the study of genetic interactions. This is an area in which the number of possible combinations is truly staggering. If, for example, one wished to test all possible five-gene combinations among the 20,000 or so human genes, there are not enough cells in the world to do the work. Furthermore, the interactions will manifest differently, depending on the readout gene.

At a small scale, genetic interactions can be studied by profiling-based methods such as Perturb-seq. Regev described one such study involving two genes for transcription factors, NF-κB1 and Rela, that jointly control a program, with Rela activating the program and NF-κB1 suppressing it. When perturbed together, her team found that in some interactions NF-κB1 was dominant over Rela and in others their interaction was perfectly additive.

Despite this success, a study such as this looking for two-, three-, four-, and five-way interactions among all human genes is never going to happen. "There aren't enough cells in the world." And that, Regev said, was motivation for thinking about how to do such experiments differently. "Is there some way of both doing the experiments more efficiently and learning more from the ones that we do?"

This is the current problem her lab is working on, and their approach is once again taking advantage of the fact that there is structure involved, so it is possible to detect patterns with much less data than would otherwise be needed. In this case, "the affected genes are structured in these co-regulated programs, and the targeted genes with genetic perturbations are structured into these co-functional modules." So the idea behind her approach was not to measure the effects of individual genes and individual perturbations, but instead measure the effects of their compositions.

"From the perturbation side," she explained, "it means we're going to sum up perturbations together. We can do it in different ways. We can perturb separately and only sum up at the measurement phase. Or we could squeeze a lot of perturbed cells into a single measurement and measure them together. We can also squeeze in a lot of perturbations into one cell."

As a test her team perturbed 600 genes with LPS in dendritic cells. "We did it either the traditional way, 1 cell at a time, 82,000 single cells, 19 channels for the traditional models, or we did it in a compressed way, squeezing 250,000 cells into 2 channels." After collecting the data, they used an algorithm to decompress the data from the second set and found that there was a 97

percent correlation between the results of the traditional and the compressed experiments. Similarly, when they examined the effects of five known major positive and negative regulators in this pathway on four major known targets, the results were consistent, showing that the approach works.

However, even with this sort of compression, examining all gene interactions in humans is still out of reach, she said. There are just too many possible interactions. So she is working on yet another approach that uses simple modular structures to help think about which genes are likely to interact. Using the GWAS genes from the ulcerative colitis work as seeds, they are building modules of two types: cell-type-specific modules which vary across all of the cell types, and program modules in which genes co-vary within a cell type. Then they will examine genetic interactions either where the genes in the same module interact or where the interactions are between genes in different modules. Some early results are just emerging from that work.

With that, Regev summarized the third part of her talk: "Responding genes form co-regulated programs, perturbed genes form co-functional modules, and this coupled structure can be leveraged in many different ways to tackle genetic interactions." Furthermore, random experiments can be used to tackle such questions as the effects of sequence on gene expression. Indeed, she said, researchers have been doing random experiments for decades. "They just didn't call them by that name." But random experiments get better and better as the amount of data increases, she said. "So I think we have a lot of room for optimism."

CONCLUDING REMARKS AND SUMMARY

To conclude her talk, Regev mentioned the various large scientific initiatives that made the work she discussed possible. Details of this part of the talk are explained in Chapter 8 in relation to the consortia and databases with which other speakers work.

In addition to the topic of scientific initiatives, Regev's talk touched on other topics that came up as themes throughout the workshop. These include understanding epigenetics and gene regulation, learning about how environmental interactions and perturbations affect gene expression, and the use of research organisms as models for laboratory work. Her talk also covered some conceptual ideas that other speakers brought up in their talks, including the non-linearity of biological interactions and gene expression, the complexity of understanding genetic function, and the inherent structure present in biological systems. Over the remainder of the 2.5-day workshop, speakers and panelists continually referred to the principles brought up in Regev's talk as they related these important concepts to their own work.

3

Case Studies on Building Functional Genomics Tools in Diverse Systems

As Gene Robinson of the University of Illinois at Urbana-Champaign commented in his remarks in the workshop's opening session, one of the keys to developing functional genomics will be developing new sets of tools that make it possible to carry out "large-scale, discovery-oriented projects" that will derive knowledge in ways that are not yet possible and at scales that will massively accelerate researchers' ability to explore the underpinnings of life. Thus, the workshop's first set of talks was devoted to descriptions of some of the current tools and systems being developed and used at the cutting edge of functional genomics. In particular, as session moderator Lauren O'Connell of Stanford University said, the organizing committee wanted to hear researchers talk about their successes and failures in starting research with new organisms and about the tools they developed to map the genotype-to-phenotype landscape. O'Connell also mentioned the committee's interest in hearing about successful examples of investigators moving tools from established researched organisms to new ones.

The session's speakers were Andrea Sweigart from the University of Georgia, who spoke about her work with monkeyflowers; Rachel Dutton of the University of California, San Diego, who described her work with microbial communities living on cheese rinds and the tools she has developed to study those communities; Zoe Donaldson of the University of Colorado Boulder, whose presentation dealt with the neurogenetics of sociality in voles; Dominique Bergmann of Stanford University who explained how she has moved among diverse species of plants, using what is learned with one to working on another and piecing together a broader picture than would be possible using just one species; and Steven Henikoff of the Fred Hutchinson Cancer Research Center, who described a method of low-cost, high-resolution chromatin profiling that can be applied to a wide variety of organisms.

UNDERSTANDING THE GENOTYPE–PHENOTYPE CONNECTION IN MONKEYFLOWERS

Andrea Sweigart opened her presentation by saying that she would be describing a community that is just beginning to move into functional genomics, so she would be touching on some of the considerations facing researchers getting started in this area.

With her work in monkeyflowers, she is studying genotype to phenotype, she said, and, in particular, seeking to understand how the environment changes that relationship. "And, as an evolutionary biologist focused on speciation," she added, "I also want to understand how variation is maintained within populations and then eventually leads to divergence between populations and species."

Sweigart first touched on the issue of how a researcher selects an organism to work on. Through her talk, she listed the following considerations:

- **The phenotypes of interest should be present in the organism.** "We want to have the full range of diverse phenotypes present," she said, and not just whatever phenotypes are present in traditional model systems.
- **The experiment should be tractable.** You need organisms that can be grown in the lab or greenhouse or perhaps in experimental plots.
- **Having access to natural populations is valuable.** This is particularly true for those researchers studying ecology and evolution, and it is useful to be able to carry out experiments in the field.
- **It is preferable to have a rich history of research in the organism,** if possible, and a diverse and interactive scientific community working with it.

All of those attributes are present in monkeyflowers, Sweigart said, which is an incredibly diverse genus of wildflowers. The genus, *Mimulus*, includes about 150 species, with its center of divergence in western North America. Rapid adaptive divergence is a key feature of the genus, and most of its taxonomic groups contain largely interfertile taxa that are phenotypically distinct. Much of the phenotypic diversity in the group is driven by divergence in pollinator attraction and mating systems.

One highly studied group of monkeyflowers is the species complex *Mimulus guttatus*, which Sweigart illustrated with a slide showing the flowers from half a dozen members of this group (see Figure 3-1). They are similar in shape but have about a fivefold variation in size, which is mainly due to differences in how they breed. The smallest flowers occur in self-fertilizing plants that do not need large, showy flowers to attract pollinators.

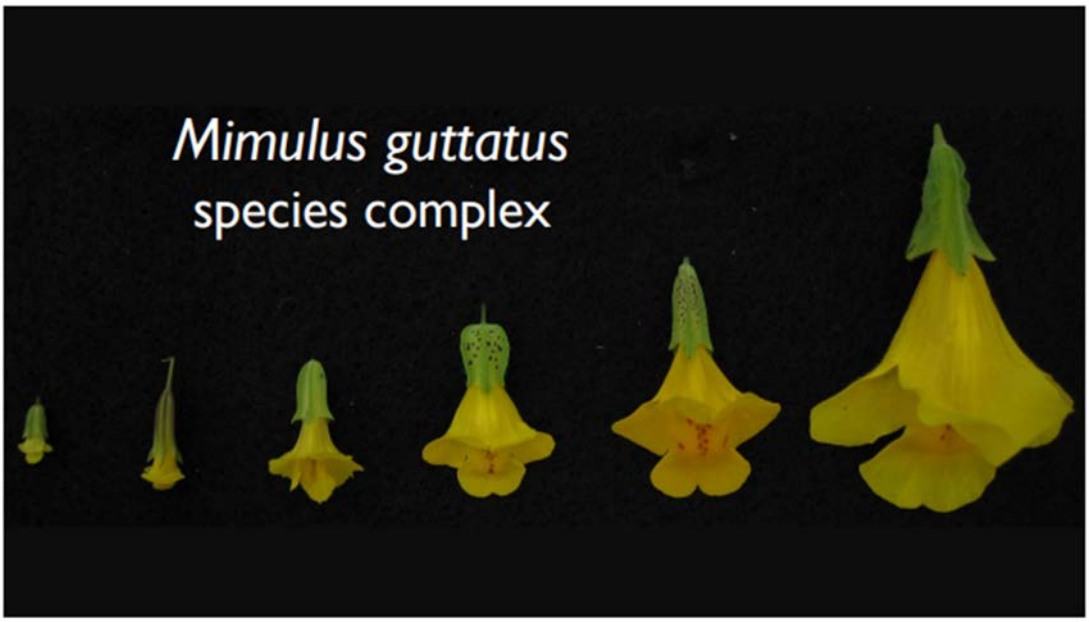

FIGURE 3-1 Display of the different sizes and shapes of flowers from the *Mimulus guttantus* species complex.
SOURCE: Andrea Sweigart presentation, slide 5.

In relation to the first requirement on Sweigart's list of characteristics for choosing a research organism, Sweigart noted that, "this genus is really famous for its ecological breadth." *M. guttatus* lives in almost every imaginable habitat in the western United States. It is found on high alpine rocky outcrops, on dunes along the Pacific, in serpentine soils, in abandoned copper mines, and even in 60°C thermal soils in Yellowstone National Park. These varied habitats require significant adaptation, Sweigart said. For example, the monkeyflowers along the coast must be able to tolerate salt spray, and those in abandoned copper mines can thrive with levels of copper that are toxic to most other plants. Furthermore, monkeyflowers can adapt to these conditions with surprising speed. The copper mines, she noted, were established in the mid-19th century and abandoned by the 20th century, and so *M. guttatus* found a way to adapt in fewer than 150 generations in conditions where most other plants could not survive.

The genus is also experimentally tractable and has accessible natural populations, Sweigart said. They are easy to grow in the greenhouse and are highly fecund. Researchers can bring seeds from the plants collected in the wild into the greenhouse and carry out various experiments there.

After providing the background on monkeyflowers, Sweigart described two case studies from her own research to illustrate some of the challenges in functional genomics. In the first she examined hybrid lethality between two species of *Mimulus* whose habitats overlapped. After bringing the two species into the lab and creating inbred lines, she crossed them to create hybrids. When these hybrids were bred to create an F2 generation, one-sixteenth of the plants died. To identify the genes responsible, she said, they used brute-force mapping and positional cloning as well as RNA-seq (RNA-sequencing), and ultimately traced the lethality to duplicates of the gene *pTAC14* (Zuellig and Sweigart, 2018). In *Arabidopsis thaliana*, a well-characterized plant, the gene seems to be essential for proper chloroplast development.

This is what seems to be happening, Sweigart said: The gene was duplicated in one of the two species (*M. guttatus*), and subsequently the ancestral copy developed a one-base-pair deletion, which knocked out its function, so only one of the two once-identical genes now work. The other species, *M. nasutus*, never had the duplication. So when the two species are crossed, second-generation progeny that inherit only non-functional copies of one of the genes or a missing copy will lack chlorophyll and will die early during development.

Gaining this understanding is just the first step. "What we really want to do is understand the evolutionary history of those genes," Sweigart said. "We want to know about the evolutionary dynamics of those lethality alleles in natural populations." Is the one-base-pair deletion neutral, for instance? Could it be evolutionarily advantageous? What is nice about this system, she said, is that it is one of the first where hybrid incompatibility genes have been identified and where their effects can be studied in populations in the wild (Sweigart et al., 2006; Sweigart and Flagel, 2015; Kerwin and Sweigart, 2020).

As it turns out, Sweigart said, the *pTAC14* story might have been the best-case scenario, or at least not the worst. She has been working on another case of hybrid incompatibility where she has found it difficult to identify the causal variants. There are several reasons for this: the phenotype is more complex, there are many more functional candidates for this phenotype, and there is quite a bit of copy number and microstructural variation in the relevant regions. "So we've realized as a community," she said, "that we're only going to get so far with this kind of classic approach of brute-force positional cloning."

Fortunately, she said, her team has been awarded a grant through the National Science Foundation's Enabling Discovery through Genomic Tools (EDGE) program to develop robust, repeatable transgenic techniques for use in many genotypes and species and mutant libraries to

be of use in multiple species. Already, she said, there have been some preliminary successes by researchers using some of these transgenic approaches, but the methods have been hard to apply in individual labs that do not have all the necessary technical skills to do transgenic work.

In closing, Sweigart offered suggestions for what is needed to move forward. The field continues to need more functional tools. "We want the ability to do CRISPR homologous recombination for things like adaptation and speciation," she said. "This is key because we want to be able to compare individual alleles from different lines." She also mentioned reporter gene/sensor lines, as well as ATAC-seq (assay for transposase-accessible chromatin using sequencing) and ChIP-seq (chromatin immunoprecipitation sequencing). Furthermore, "we need more genomics resources, generally. We need whole genomes, we need the pan-genome." Because there is such tremendous variation in the monkeyflower genus, one cannot work with a single reference sequence and expect to see the whole picture, she said. Finally, she named community resources, such as seed banks and distributors, as being critical for her work.

MICROBIAL COMMUNITIES AND THEIR INTERACTIONS ON CHEESE RINDS

Rachel Dutton began studying microbial communities living on cheese rinds, she said, because of a conviction that a vast amount of biology was being missed by studying individual organisms in isolation. "My background is in *E. coli* genetics," she explained, "but I started to get really interested in this idea that microbes don't grow in isolation; they grow as parts of complex communities. So, what is the biology that is yet to be discovered if we put organisms into a more natural context?"

She set out to find a natural microbial community that could be studied in situ but that could also be separated and studied as individual components. A major challenge in the study of microbiomes is that researchers cannot culture most of the diversity present in a particular environment. In addition, many microbiomes consist of hundreds or even thousands of individual species living together, so it is not feasible to work with them experimentally. "I was looking for a system where we could actually take it apart into its individual components, and then ideally . . . put it back together in a lab in an in vitro system where we can manipulate the membership and manipulate the conditions under which the community is growing," Dutton said. This would allow her to identify some of the mechanisms and functions operating in these communities.

She ended up focusing on the microbial communities that are found in fermented foods, such as cheese, beer, or wine. Most such foods form through the rapid growth of relatively simple microbial communities. She reasoned that she should be able to culture these communities because they grow rapidly and in well-defined systems.

In particular, she decided to study cheese rind biofilms, which are microbial communities that form on the surface of cheese during the aging process. Typically, the complexity of these communities ranges from low to medium, she said, or from about 3 to 10 member species. However, these members are phylogenetically diverse, with fungi and many different types of bacterial species from across different phyla.

"We showed that these systems are completely culturable," she continued. "We can completely deconstruct them into their individual components," she said, and can put them back together in vitro to create systems that represent the natural behavior found in the real world. "We have in vitro cheese in the lab" (Wolfe et al., 2014).

After spending about 5 years developing the system, Dutton set out to explore the biology of these communities. One of the advantages of working with a system where one has access to

both in situ communities and in vitro communities is that it is possible to take both top-down and bottom-up approaches to studying their biology. The top-down approaches include such things as comparative genomics and metagenomics, while the bottom-up approaches include genetic screens and in vitro community manipulation. She described a high-throughput genetic screening approach in which her team manipulates in vitro communities to gain insight into what is happening in the natural system (Morin et al., 2018).

The work has been done on the simplest of cheese communities, a brie- or camembert-style community with three members: one bacterial species, *Hafnia alvei*, and two fungal species, a yeast, *Geotrichum candidum*, and a fungus, *Penicillin camemberti*. "Even within just the simplest community," she noted, "we have quite a bit of phylogenetic diversity here."

The approach they take is to use large barcoded transposon libraries (Wetmore et al., 2015). "You're making large random mutant libraries," she explained, "but each of the transposons in the library has a random barcode, so you can associate each insertion in the genome with a barcode sequence and follow the population changes in the library just by sequencing barcode abundances." The library they used had a pool of about 150,000 mutants, which represented about 15 different insertions and every non-essential gene in the genome, she said.

The team uses the sequencing approach to measure the barcode abundances in the starting population, and then looks for genes that are defective in a certain type of environment, that is, genes required for growth in that particular environment. The strategy then is to grow the libraries under different conditions—by themselves in culture, on cheese alone, on cheese with partners, or on cheese with the entire community—and compare the different outcomes (see Figure 3-2). If a gene has a barcode that drops out of the population under growth alone and also in the interaction condition, this shows that the gene is always required. Another category consists of genes that are not required for growth in the "alone" condition but are required in the presence of interacting partners. A third category, which Dutton said they had not considered until they saw the data, is called interaction-alleviated genes, which consists of genes that are required in the growing-alone condition but not when a community is present.

In the first pass of their experiments, Dutton said, her team used a pre-built library that was transformed into *E. coli. E. coli* was used for practical reasons, because of all the knowledge and tools available for this well-studied organism. "So we grew *E. coli* alone in these conditions, we grew it with individual partners, in three-member communities with two pairs of cheese partners, and then in a complete community, and compared all of these results," Dutton explained.

They calculated the gene fitness data based on the barcode abundance of the inoculums versus growth on their in vitro medium, which is a cheese curd agar. What they found was a large number of genes that were required in the grow-alone condition and a somewhat smaller set of genes that were required to grow in the presence of a community. The overlap between these two sets was characterized as "core requirements." Some genes were not important in the alone condition but became important when they were growing in the context of a community, and there was a relatively large set of genes that were important alone but no longer needed when growing in a community.

The latter type of gene, the community-alleviated genes, frequently map to amino acid biosynthesis pathways. "What this tells us," Dutton said, "is that *E. coli*, when it's growing alone on cheese, . . . has to make its own amino acids. If you knock out any genes in amino acid biosynthesis, when *E. coli* is growing by itself, it will die." However, if it is growing in a community, the data imply that some other members of the community are providing it with the required amino acids. When they looked at their data in more detail and examined individual

pair-wise contributions to the fitness effects, they found that it was only the fungal species that were producing this cross-feeding effect. What they believe is happening is that fungal species are secreting proteases and breaking down proteins into peptide chains and amino acids which *E. coli* is then able to use.

Finally, Dutton described some high-order interactions that they were able to detect in their data by looking at the patterns of genetic requirements of *E. coli* grown alone versus in pair-wise combinations versus in the community. Strangely, there were situations where a gene was not required when *E. coli* was growing alone, was required in paired conditions, but was not required when growing in a community. There were also combinations of genes required in the alone condition, not required in the paired condition, and required again in the community.

By looking at the data more closely and working with people who were more familiar with quantitative epistasis in gene patterns, Dutton was able to determine what was going on in these higher-order interactions. She described one such situation: "So there's a multi-drug efflux pump in *E. coli* made up by acrA and B proteins," she said. "When *E. coli* is grown alone, these genes are not required. It doesn't need a drug efflux pump when it's growing by itself." However, when *E. coli* is grown in the presence of *Geotrichum*, the pump is now required, which indicates that *Geotrichum* is producing some antimicrobial that *E. coli* needs to pump out in order to survive. But when *E. coli* is grown in the presence of the community, again the pump is not required because *Hafnia* was somehow negating the effect of the antimicrobial produced by *Geotrichum*, allowing *E. coli* to survive without the drug efflux pump.

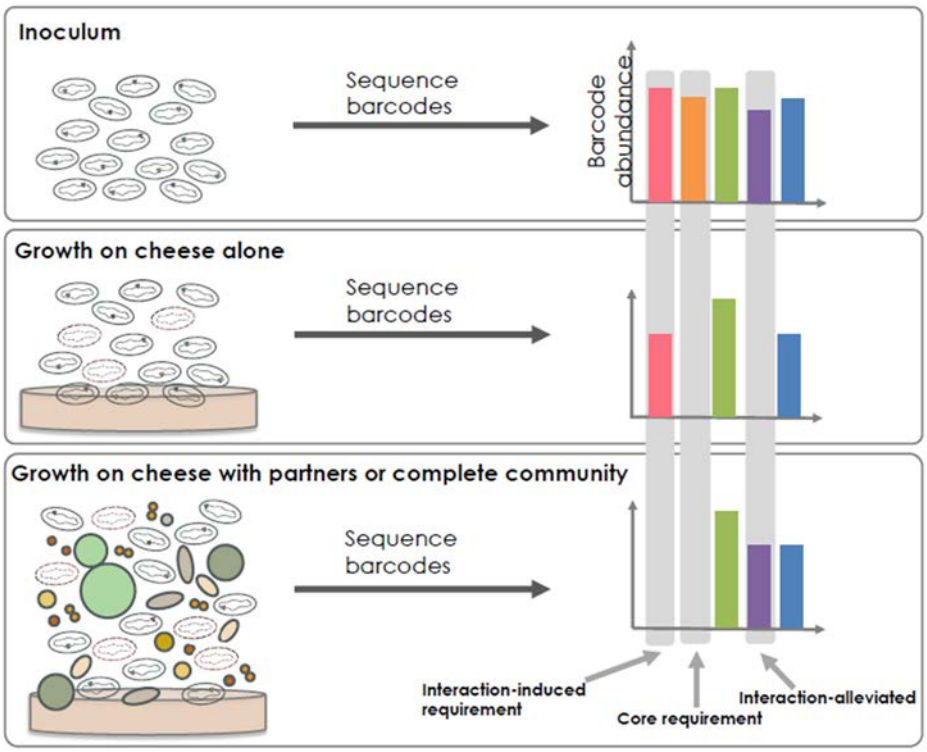

FIGURE 3-2 Different growth conditions for a microbe being tested for genes that are present or absent when grown in a microbial community.
SOURCE: Rachel Dutton presentation, slide 2.

In closing, Dutton listed the successes of her team and the roadblocks they still face in their work. They have successfully implemented some high-throughput screening techniques in a relatively new model system by using *E. coli* as a proxy for what was happening in the environment. Since performing those experiments, they have made libraries in cheese-related microbial species and have detected many types of interactions among the organisms, even in a simple community.

One of the greatest experimental challenges, she said, is that about 30 percent of the genes they identify, even in *E. coli*, have no known function. A second challenge is that the biological insights they obtained from their screens was mostly limited to well-characterized areas of biology, such as metabolism.

Moving forward, she said, what they really need are medium- and high-throughput ways to efficiently categorize, prioritize, and characterize the genes that come out of the screens. What, for instance, are the genes that represent new biology that is only happening in communities, and how can comparisons be made across systems?

"We are not really interested in learning about cheese," Dutton said in conclusion. "We want to learn general principles, general mechanisms. How do we take our findings in this relatively simple and tractable system and compare it to other systems to figure out what are the generally important pathways and processes?"

NEUROGENETICS OF SOCIALITY AND RELATIONSHIPS

The vole is a small rodent related to hamsters and lemmings, and, as Zoe Donaldson explained, it offers an example of how vast behavioral differences can exist even between closely related species. In particular, the prairie vole and the meadow vole live in the same habitats and appear to be nearly identical, but behaviorally they are distinct from one another. Prairie voles are monogamous. A male and a female will mate, share a territory, and raise their offspring together. Meadow voles, by contrast, are promiscuous. "Females will become behaviorally receptive to males, will mate sometimes with multiple males, and then go off and raise the offspring by themselves," Donaldson said.

The question Donaldson asked was what genetic differences underlie this behavioral difference and makes the brains of the prairie voles, but not the meadow voles, capable of forming long-term pair bonds. It is a question that cannot be answered in a mouse model, Donaldson noted, because "mice are not monogamous and you can't study pair bonding if it doesn't exist within your species." So, it was necessary to develop the tools to study the question in voles.

Skipping over much of the history of how potentially relevant genes were discovered, Donaldson identified the gene for vasopressin receptor 1a (V1aR) as crucial to the difference in behavior (Lim et al., 2004). In particular, there are striking differences in how the gene is expressed in the brains of the two species. One brain region with a particularly large difference is the ventral pallidum, where there are high levels of gene expression in the prairie vole (the monogamous species) and lower levels in the meadow vole (the promiscuous species) (Phelps and Young, 2003). Researchers have demonstrated that these expression levels are critical for some of the differences in social behavior by using a viral vector to increase the expression of the gene in the ventral pallidum of the promiscuous species, after which they "show an affiliative preference for the animal they're mating with, which is the basic hallmark behavior that you need in order to be monogamous" (Hammock and Young, 2003).

This fascinating transformation shows "that you can change the expression pattern of a single gene, and the architecture's already in place within the brain to completely transition your mating," Donaldson noted. She set out to find the genetic basis of the difference in expression patterns in the vasopressin receptors in these animals.

Her early work on this system took place in the pre-genome era, Donaldson said, so she had to use the tools that were available at that time. "We fished out the gene from a phage library and found that the coding region of this gene was nearly identical between these species," she said. In contrast, there is a length of repetitive DNA upstream of the gene that is nearly absent in the promiscuous species but more than 600 base pairs long in the monogamous species. Furthermore, there is an allele-like variation in the lengths of that repeat-containing element in the monogamous species that was shown in breeding studies to be associated with individual differences in social behavior, such as how attentive a father is toward his offspring. This information led her to hypothesize that variation in this repeat-containing element directly contributes to both species-specific and individual differences in patterns of expression of the vasopressin receptor gene.

At the time, she did not have the tools to test this hypothesis in voles, and so she turned to mice, where it was possible to manipulate the genome in a specific way. This is reflective of how Dutton used *E. coli* as a tractable system with the eventual goal of moving to more complicated and less studied yeast and fungal communities. To start her mouse work, Donaldson took a 3.5-kb region upstream of the gene encoding the V1aR from voles—the region that contained the repeat element—and put it into mice, replacing the corresponding region of the mouse genome (Donaldson and Young, 2013). She did this for three versions of the region, one from the promiscuous species and two different versions, one with a longer repeat than the other, from the monogamous species. Because the genetic differences in these mice were limited to the repeat-containing element, Donaldson could assess the effects of the repeats cleanly, without having to worry that the actual functional variants were something else nearby in the genome that were causing the differences in expression patterns.

The expression of the V1aR gene in the transgenic mice looked much closer to the expression of the gene in normal mice than the expression in voles. However, there were particular brain regions with clear changes in the gene expression pattern, including the dentate gyrus, part of the thalamus, and the central amygdala. "This wound up being a silver lining," she said, "because now instead of having to look at the entire brain and the complexity of gene expression within multiple brain regions, I was able to focus on what was going on within these three separate brain regions." Also, she added, the similarity in gene expression patterns in other regions of the brain implied that there were other regulatory elements outside of that 3.5-kb region that were driving the gene expression in those other parts of the brain.

When Donaldson examined how patterns of gene expression differed among the three types of transgenic mice in those three relevant brain regions, she found that the patterns differed in the same direction as the patterns in the three voles—that is, the promiscuous species and the two versions of the monogamous species with the shorter and longer repeats. For example, there were higher levels of gene expression in the dentate gyrus in a prairie vole than in a meadow vole, and that was recapitulated in the mice carrying the prairie vole or the meadow vole versions of the repeat-containing element.

"If we put all of this together," Donaldson said, "what we've essentially learned is that DNA diversity in these regulatory elements leads to species and individual differences in the

expression of this gene, but in a brain region-specific way, such that we have brain regions that are essentially untouched by this manipulation."

An important aspect of this, Donaldson said, is that repeats of the sort responsible for the differences in behavior mutate at a much faster rate than the rest of the genome. If such repeats are in the right areas of the genome, they can provide an evolutionary mechanism for generating diversity in gene expression, acting as a sort of "tuning knob" in a specific brain region.

One aspect of the work that concerned her, Donaldson said, was that it had been done in mice, not voles. How might the results have been affected by the fact that the repeat-containing elements were put into mice? What effect might the genomic milieu of the mice have had on the regulatory elements from the voles? "One of my goals has been to develop techniques that will allow us to eventually answer this question," she said.

Her first step in that direction was to develop the first germline transgenic prairie voles. She did this by injecting a lentivirus into embryos, allowing the lentivirus to infect the embryo and place DNA into its genome (Donaldson et al., 2009). While useful, this technique only allows one to add genes, not delete them, and furthermore, the lentiviral constructs become repressed after a few generations.

She has also teamed up with Devanand Manoli to use CRISPR to do germline knockouts in voles. It is a powerful technique, Donaldson said, but it is also challenging for a number of reasons. In addition, both CRISPR and the lentivirus technique require a great deal of optimization for each species. For instance, whereas researchers have learned how to cause mice to superovulate, the ability to do so in voles is still extremely limited, which means that it is necessary to use a tremendous number of the animals to get just one knockout vole and is labor-intensive. "We are working [on] moving from prairie voles to meadow voles," she added, "so we'll soon get a sense of how much of a challenge there is even moving these techniques across closely related species."

There is also what Donaldson called a "conceptual limitation" of the CRISPR approach. She explained that when a gene is knocked out, it gets rid of the expression of that gene throughout the life course of the organism and across the entire organism. This is particularly important in the case of V1aR, which is encoded by a multi-faceted gene that mediates many different aspects of monogamous behavior but does so via its activity in different brain regions. For example, the vasopressin receptor is known from various studies to act within the pallidum to influence not only mating behavior but also parental care—in essence, whether fathers take care of their offspring. In the hypothalamus, however, it acts to influence mate guarding, a behavioral characteristic that reinforces the bond with the mate, and in the retrosplenial cortex it may be involved in use of space and, therefore, male fidelity in the species.

"The question then," Donaldson said, "is how we begin to parse out the pleiotropic effects of these genes, and that is where my lab is currently making a lot of effort to develop ways to go in and selectively manipulate gene expression in adult animals in specific brain regions." She described one success that her lab has had in this effort, which involved injecting short-hairpin RNAs (shRNAs) in particular areas of the brain to decrease the expression of vasopressin receptors. And at present, she said, her lab is working with a slightly modified version of that approach using CRISPR to either inhibit or activate the transcription of these genes within specific brain regions.

She closed with a look to the future. First, she said, her lab is still trying to identify the genetic elements that contribute to species and individual differences in social behavior. To do this, it would be helpful to have high-quality genomes. Having a reference genome is not

enough, she said. "There's a huge amount of SNP (single nucleotide polymorphism) variation within these voles, so even something as simple as developing a guide RNA can be incredibly sensitive if you have a SNP within the binding of your guide RNA region." She also reemphasized the need to be able to manipulate the genetic elements in the appropriate genomic or organismal context.

Second, she said, she needs strategies for parsing pleiotropy. In particular, she needs to be able to manipulate gene expression in a regional- and temporal-specific manner.

Finally, she said, "as a neuroscientist I feel the need to point out that genes don't encode behavior. Genes encode mRNAs that make proteins that alter your neuronal function, and ultimately, you get differences in behavior." So, it is important to think in terms of what might be called "neural intermediate phenotypes" and to look for general principles related to how genes can affect neurons.

GENETIC INTERROGATION OF DIVERSE PLANTS

While much can be learned by studying the functional genomics of one species or a group of closely related species, Dominique Bergmann of Stanford University said, there are also benefits to using what is learned in one species when working on another. In her presentation she described how she studies the genetics of diverse plants to uncover rules of developmental fate, pattern, and resilience.

Her research focus is the development of plants, Bergmann said, and she seeks to answer a series of questions first in one model system and then in others:

- How do patterns emerge?
- How are specialized cells made?
- How do environmental inputs modify development?
- Is it possible to make plants that survive or mitigate their biology in response to changing climates?

The last question is a particularly practical one, she said, because people depend on plants for survival, and plants will inevitably be changed as the climate changes.

Like many of the other speakers in the workshop, Bergmann said, she is interested in how one moves from genotype to phenotype, but the usual picture of a single arrow pointing from a genotype at the level of DNA information to a phenotype of patterned cells and tissues should be thought of as having many different and non-linear steps.

To deal with that complexity, she said, researchers typically work with models, both model organisms and "models that extract features that are common to many developmental decisions, but do it in a very simple way." Her lab works with the model organism *Arabidopsis*, and focuses on a specialized cell type, the stomatal guard cell, a pair of which form a valve, which allows carbon dioxide into the plant, and water vapor and oxygen out. Stoma are arranged in patterns that are environmentally determined in some ways, but there are also some hardwired patterns.

"Based on lots of peoples' work over the last two decades," she said, "we have a pretty good molecular picture of what's required to make and to pattern these cells in *Arabidopsis*." In a young leaf the cells are all essentially equivalent, but then under the direction of some key regulators, the leaf cells proliferate and then differentiate. Much of the patterning of cells on the

leaf is decided by cells communicating with one another with secreted peptides and cell surface receptors, but there is also environmental and systemic information that is integrated into decisions on the development of the various cells. She noted the importance of being able to capture live how different parts interact in a dynamic system. That has been possible in *Arabidopsis* because of all of the tools and the deep knowledge available for that one system.

The question, though, is what happens when one moves away from *Arabidopsis*, Bergmann said, "because, frankly, no one eats *Arabidopsis*, and it doesn't contribute a whole lot to the global climate cycles." In the case of broad-leaf crop plants that might share much of their genome with *Arabidopsis*, she said, it might be simple to use the knowledge from *Arabidopsis* about what genes might be core features of environmental resilience, and then transfer them and their regulatory systems into the crop plant. However, in plants that are distantly related, including those with the highest commercial value, there may not be such a direct path. Bergmann's focus during the presentation was on this second issue.

Moving from *Arabidopsis*, Bergmann's group decided to look at the cereal crops because they are extremely important both economically and ecologically, and they also operate differently than *Arabidopsis* at the level of development. The cereals are a group within the grasses, and grass stomata are patterned differently on the leaves, for example. Another important consideration, she said, was that a great deal of work had already been done on grasses and a lot of tools were already in place.

In particular, *Brachypodium distachyon*, whose common name is "purple false brome," has become an important model for cereals. Its genome was published about 10 years ago, and the genome includes many homologs of genes that Bergmann had studied in *Arabidopsis*. But there were also key phenotypic differences between *Brachypodium* and *Arabidopsis*. In particular, the stomata on *Brachypodium*, like all grasses, were composed of four cells rather than two, as in *Arabidopsis*. The two extra cells line up on either side of the guard cells and are called "subsidiary cells." The grasses are more resilient during drought, and one reason appears to be the performance of their stomata. And this is something that studies in *Arabidopsis* can say nothing about, she commented.

Bergmann's team set out to understand this novel feature. The *Brachypodium* genome had already been worked out, but more tool optimization was still needed. They were able to optimize transformation conditions and create constructs with engineered mutations that allowed them to learn a lot about the system. However, to answer questions about novelty, they turned to genetics—in particular, forward genetics—to identify the factors required for innovations.

"We really were interested in this novelty, these great subsidiary cells that are so wonderful for making these functional complexes," Bergmann said, so they screened many plants with a microscope and finally found one that did not have them. "Great, we have a mutant," she said. "How do we find that gene?" As it turned out, they were able to find it fairly quickly. Multiple accessions had been sequenced, and these enabled rapid mapping. The mutation turned out to be a small deletion in a transcription factor.

Bergmann's group was surprised to find that the transcription factor was similar to one used in *Arabidopsis* to make a precursor to the stomatal guard cells. However, in *Brachypodium* it appears to switch its function to make the subsidiary cells. There was a change in the relationship between this one gene and the phenotype—that is, which type of cell it creates.

Ultimately, they discovered what was underlying that change in the relationship between the genotype and phenotype. It turns out that while the protein was being expressed in some cells, it was not expressed in those cells where the team believed it was required to carry out the

function. The explanation, Bergmann said, is that in plants transcription factors can move from one cell to another. So the innovation—the way that the gene has changed its activity—was not by being expressed in a different place because of changes in the promoter region or by having a different biochemical function, but by changing its ability to move from one cell to another.

Summing up, Bergmann said, "So what we found here by moving from one species to another was a rewiring," and the discovery was enabled by the fact that previous work had been done on the new organism they were interested in. "We had populations that we could screen, we had fairly good genomes, and then we created a number of tools because the genotype-to-phenotype connection needed those intermediate steps filled in." Their success was due to a number of factors, including long-term funding by the U.S. Department of Agriculture and the U.S. Department of Energy for the sequencing of many *Brachypodium* accessions, the building of molecular tools for transformation, and the creation of mutant libraries. Furthermore, she noted, there is an active and open *Brachypodium* community "that is driven to create things and share them."

In closing, Bergmann asked which tools are needed to find and understand innovations. In the case she described, what they actually identified was a rewiring. "Plants reused a gene that we knew a fair bit about," she said. But her goal is to look more broadly at diversity in animals and plants and see what can be learned. "We're looking for innovation, we're looking for novelty," she said. "How do we actually find that?"

Since 2012, she said, her team has identified eight mutations in their *Brachypodium* screen. Four of them were known genes with conserved functions in *Arabidopsis*, and two were known genes that had been rewired to have different functions, including the one she described, but two were novel genes. The novel genes have not been published, she said, and they will not be for a while. Understanding real novelty takes time.

To find and understand developmental organization, she concluded, "Choose the question, and then choose the organism." Once the organism is chosen, genome and transcriptome sequences and gene editing capabilities are prerequisites. There is an opportunity to revisit classical systems that may have been abandoned for a while but have already provided extensive descriptive data. Much of the foundational work to build new systems can be tedious. Furthermore, there is a question as to whether the lengthy amount of time it takes to develop a new system is compatible with various time lines, such as grant cycles, Ph.D. program times, or postdoctoral funding. Finally, she asked, "How can we make going into these leaps attractive to young PIs [principal investigators]?" They represent the future.

LOW-COST, HIGH-RESOLUTION CHROMATIN PROFILING

Unlike the first four speakers in the session who had focused on one or a few particular organisms, the final speaker, Steven Henikoff of the Fred Hutchinson Cancer Research Center, described some epigenomics tools his lab has developed that can be applied to a wide variety of organisms. In particular, he spoke about methods to perform low-cost, high-resolution chromatin profiling.

Most people who are interested in the genotype-to-phenotype connection will want to perform chromatin profiling at some point, Henikoff said, and the most common chromatin profiling approach is ChIP. He explained the ChIP process involves four basic steps: (1) the DNA and associated proteins on the chromatin are cross-linked, (2) the resulting complexes are broken up into pieces about 500 base pairs long, (3) an antibody is added to precipitate out the

protein of interest, and (4) the DNA is purified and the fragments are sequenced. "This part of the process really hasn't changed for the last 35 years," he said, but over the past decade or so, ChIP-seq has grown rapidly in popularity, and it has served as the basis for several large genome-scale projects, including the Encyclopedia of DNA Elements (ENCODE) project.

Because the ENCODE project relied so heavily on ChIP-seq, the group running the project developed a set of standards for using the technique. One of those standards was that there should be at least 10 million reads per replicate. Even with the cost of sequencing going down dramatically, Henikoff said, it can still be quite unwieldy and expensive to store and manipulate the larger and larger datasets generated with ChIP-seq, and so there is room for an alternative.

As it happens, there are other ways to perform chromatin profiling. In particular, Henikoff's lab modified a chromatin immunoclevage method to create what they call "CUT&RUN" (cleavage under targets and release using nuclease) (Skene and Henikoff, 2017). He described the technique in this way (see Figure 3-3): Live cells are mixed with Concanavalin A beads, which help the cells stick together. In the next step, the cells are made permeable and an antibody is added to the desired target, such as a transcription factor. The antibody diffuses into the cell and finds its target, and then a fusion protein of protein A complexed to micrococcal nuclease (MNase) enters the cell, where protein A binds to the antibody. Adding calcium activates the MNase, which cleaves the DNA on both sides of the target, at which point the cleaved DNA is extracted and sequenced. "From live cells to purified DNA, it takes about a day," Henikoff said.

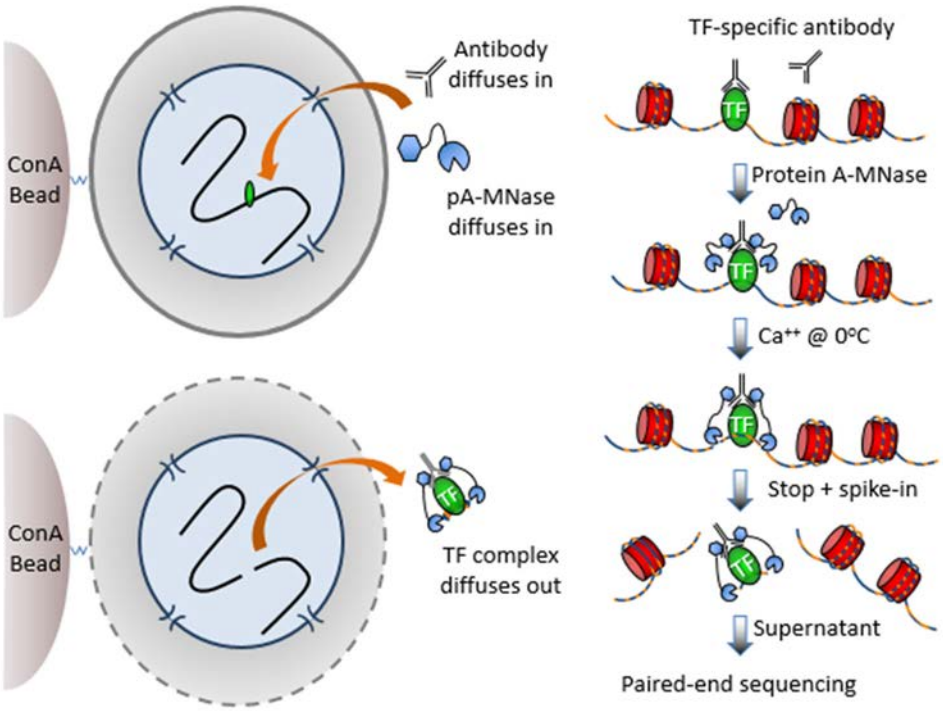

FIGURE 3-3 General explanation of how the CUT&RUN method works.
NOTE: Ca++ = calcium ion; ConA = Concanavalin A; pA-MNASE = protein A-micrococcal nuclease; TF = transcription factor.
SOURCE: Steven Henikoff presentation, slide 7.

An important advantage of CUT&RUN is that the backgrounds are very low. The reason for that, he explained, is that unlike ChIP, where one solubilizes the entire cellular contents and then grabs the antibody like pulling a needle out of a haystack, "with CUT&RUN we leave the haystack behind." The lower background means that fewer reads are needed to get a clear signal. To illustrate, he showed ChIP-seq data from ENCODE that was generated with 56 million reads and comparable data from CUT&RUN with only 7.5 million reads. When the ENCODE data were restricted to just 7.5 million reads, the background noise was much worse and obscured some of the peaks in the data (Skene et al., 2018). That background is what makes ChIP-seq so expensive, he commented. CUT&RUN is also sensitive enough that even with just 100 cells, the resulting data are of reasonable quality, Henikoff said, and it gives base-pair resolution.

Because of the sensitivity of CUT&RUN, Henikoff became interested in using it in the clinic. As a proof of concept, they set up a robot to carry out high-throughput CUT&RUN using patient-derived xenograft tissue (Janssens et al., 2018). They produced about 6 million mapped human fragments per sample for an in-house cost of about $50. This produces the entire epigenome, he said. "It's very informative data and really doesn't cost a whole lot." Since that initial proof of concept, Henikoff set up the robot for a number of his clinical colleagues to do chromatin profiling, and from there it has spread to become "quite popular" and has been one of the most viewed protocols on Protocols.io, a popular open-source website for laboratory protocols.

Nonethless, Henikoff said, he believes that there is an even better way to do chromatin profiling that was recently developed by a postdoc in his lab and is called CUT&Tag (cleavage under targets and tagmentation) (Kaya-Okur et al., 2019). It is similar to CUT&RUN but with the transposase Tn5 in place of MNase, and the activation is done with magnesium instead of calcium. A major difference is that the DNA segments acquire tags on their ends so that it is possible to create sequencing libraries from the samples. Again, the process is fast—about 1 day from live cells to sequencing-ready libraries.

The process also has the same advantage as CUT&RUN in terms of low background. As an example, Henikoff showed some data where CUT&Tag identified the same peaks as ChIP-seq but with 8 million reads instead of 27 million. A million probably would have been enough, he said. "The background's down and we don't have to sequence that deeply." The comparison with ATAC-seq was similar, with CUT&Tag identifying the same peaks as ATAC-seq with 6 million reads instead of 34 million reads or more, depending on the lab that had done the work.

And, like CUT&RUN, CUT&Tag is very efficient. In one test, Henikoff's lab used data from 6,000 cells (4.6 million reads), 600 cells (4 million reads), and even just 60 cells (2.8 million reads), and in each case the peaks were comparable to those from the ENCODE ChIP-based data that relied on 56 million reads. Taking this pattern to its extreme, Henikoff's lab has even carried out the technique on single cells. With CUT&Tag, he explained, "everything basically holds together in the cells, so we go from antibody binding all the way through to the tagmentation, and the cells are intact."

In conclusion, Henikoff offered some closing thoughts. First, he said, ChIP-seq is very wasteful; more than 90 percent of the sequence is noise. Both CUT&RUN and CUT&Tag offer more information with fewer reads. "Getting rid of the background is key," he said. It also seems that the two methods can be used to perform routine clinical profiling that is cost-effective. Also, the two techniques are highly reproducible, while ChIP-seq is inherently not reproducible because of the use of cross-linking and sonication to break the chromatin complexes into fragments.

Finally, he said, the online protocols allow almost anyone, even undergraduates, to use CUT&Tag. "The cheaper it gets, the easier it gets, the more people are going to do it themselves," he said, and this do-it-yourself approach"is the future of chromatin profiling.

DISCUSSION

Following the presentations in this session, Gary Churchill from The Jackson Laboratory opened the discussion period with a comment about Donaldson's presentation. "I admire your heroic work to create transgenic mice and transgenic voles," he said. Researchers at The Jackson Laboratory carry out transgenesis at an industrial scale—people can order customized transgenic mice and get them quickly—but it generally only works in C57 black 6 mice or another equally well studied strain. As soon as one moves beyond these basic strains, even staying with *Mus musculus domesticus*, everything falls apart. "A lot more work needs to be done to optimize transgenesis in genetically diverse mice, much less reaching out to voles and plants and everything else that we need transgenesis for," he said, and he asked Donaldson for her thoughts.

Donaldson agreed that bespoke methods are typically required to make germline transgenesis work in diverse species. Thus, she said, her preference is often just to use a workaround. For instance, viral vectors tend to work very well between species. But it seems unlikely, she said, that there will ever be germline transgenesis techniques that are not species specific to some degree.

Sweigart commented that this is also a problem in plants, "where many of the biological aspects of why transformation even works are mysterious to us." So, it is difficult for researchers in the area to understand genotype-specific reasons for why some things work and others do not. At least there is quite a bit of funding for mouse research, she said. "How do we deal with these sorts of problems in organisms for which there's never going to be that sort of funding?"

The same is true for bacteria, Dutton added. She and her colleagues have done much of their work in *E. coli* because of the availability of so many genetic tools that work well in that organism. They have been able to translate those tools easily to bacteria that are closely related to *E. coli*, but as soon as researchers move away from that species group to other less closely related bacteria, she said, "the implementation of preexisting genetic tools becomes really spotty."

Bergmann agreed with these speakers and went on to summarize ideas the previous speakers had touched on related to research that does not involve well-studied "model organisms." Although tools such as CRISPR are, in theory, available to all researchers in functional genomics, most researchers must still figure out how to adapt those tools to their own organisms, and this adaptation is the biggest bottleneck in the field.

Jason Rasgon of The Pennsylvania State University offered a more positive outlook. His lab has been "reasonably successful in developing new delivery methods working across a wide taxonomic diversity of arthropods, and we're even moving into non-arthropods, mollusks, and potentially some vertebrates," he said. In short, there are some novel ways to take some of these genomic technologies and move them into "non-model species."

One audience member asked the panelists for thoughts on how to start exploring genes of unknown function. That is a central question, Donaldson responded. She noted that many papers do a quantitative trait locus (QTL) analysis and find the gene they expected to find based on the literature. Frequently, these papers ignore the gene hits with no known function because the researchers do not know how to start experiments with these hits. She proposed a funding model

where a researcher could study these genes of unknown function, assuming they had a plan outlining a research plan. She hopes this would prevent researchers from immediately getting dismissed when trying to study a challenging but important area of research, such as gene hits with unknown function.

Bergmann agreed and said that there are two different issues at play. First is "the unknown that we know." Researchers do have various approaches to identifying the unknown. It just takes time and funding. She suggested that it would be a good idea if every grant asked "What are you proposing that will not have an answer in 3 years or 5 years but that will advance the field because it is really novel?" The second issue is "Are there other ways that we can explore gene function?" It will be difficult to discover these new ways without allowing researchers to, in a way, go back 50 years to when none of the current tools were available and "everyone had to do these fairly slow, painstaking ways of interrogating gene function." Bergmann reiterated that researchers will need time and funding to follow such a path.

Another audience member offered a different perspective on the unknown, saying, "even in model organisms, there are tens of thousands of transcripts that are expressed at high levels that also make proteins, and they're considered either LNCRNAs (long non-coding RNAs) or usually not even considered at all." Thus, the audience member concluded, even in a model organism there are many unknowns in terms of potential genes.

Sweigart expanded on that saying that what is known and unknown also influences the phenotypes that researchers study because it is much easier to study the phenotypes that have well-known candidate genes. On the other hand, much of the interesting unknown biology is much more complex and thus not as likely to be the subject of a research project.

Finally, an audience member asked the panelists about training. As a young principal investigator, he said, he is having difficulty convincing the graduate students he hires, who have molecular and cellular backgrounds, to tackle the difficult computational challenges the work requires. What can be done?

Dutton responded that her department has just redone its graduate curriculum so that the core classes include biostatistics and bioinformatics classes, and now some of the classes that were previously part of the core curriculum that were focused on molecular and cellular biology are electives. "So, we've had to do a lot of hiring in those areas to teach those on a wider scale," she said. Before that, her students were mainly self-taught through online tutorials. But there is indeed a gap between what students have traditionally been taught and what they need to know in the current functional genomics environment. Departments and universities are only now starting to address this gap, not just at the graduate level but at the undergraduate level as well. This is a shift that needs to happen, she said. (Further discussion on functional genomics education is highlighted in Chapter 9.)

4

Understanding the Contributions of Non-Protein-Coding DNA to Phenotype

The early view of the genome was that the important part was its genes, which carry the instructions for how to make proteins. The remainder of the genome—referred to as "non-coding" sequences—was thought to be less important or perhaps not important at all. However, it is now understood that this view was misguided and that non-coding elements may play crucial roles, such as regulating the transcription of genes into RNA (King and Wilson, 1975). But there is much less known about the non-protein-coding parts of the genome than about the genes themselves. Developing a clear understanding of how genotype leads to phenotype will require a greatly enhanced comprehension of the role of non-protein-coding DNA.

The three speakers in this panel explained their work to illuminate the issue of how these non-coding regions of the genome contribute to an organism's phenotype. The first speaker was Felicity Jones of the Friedrich Miescher Laboratories of the Max Planck Society, who described her studies of the role played by non-coding portions of the genome in adaptation by stickleback fish. Scott Edwards of Harvard University spoke about his research on the evolution of non-coding regulatory sequences in various taxa of birds that have lost the ability to fly through evolutionary changes. The third speaker was Francois Spitz, who described what is known about topologically associated domains, a feature of the genome that shapes interactions among various noncoding elements of the genome. A discussion with the panel members and the audience followed the three presentations.

FUNCTIONAL GENOMICS OF ADAPTATION IN STICKLEBACKS

Felicity Jones began her presentation by noting that many of the specialized functions that organisms develop through evolution are achieved not by modifications in genes, but rather by variations in the regulatory mechanisms that control the genes. "And in my lab," she said, "we're really interested in understanding how those regulatory mechanisms play a role in determining the specialized functions that enable organisms to adapt to the environment." Her lab asks such questions as:

- What are the molecular mechanisms involved in this environmental adaptation?
- How does the genome interact with the environment?
- How does genome function change as an organism adapts to a different environment?

To address these questions, Jones works with stickleback fish. They are particularly useful for this purpose because they underwent an adaptive radiation about 10,000 to 20,000 years ago in which the ancestral marine (i.e., ocean-living) stickleback colonized a large number of newly formed freshwater habitats across the northern hemisphere. As a result, she said, "as you go across the northern hemisphere, in any freshwater body you care to look at, you're very likely to

find sticklebacks and see much phenotypic, morphological, physiological, and behavioral diversity amongst these populations." One useful aspect of this biological system is that many traits evolved in parallel in different populations. For example, the freshwater sticklebacks have lost many or all of their bony lateral plates, which the marine ancestor still retains, and that adaptation in the freshwater fish occurred independently many different times. This makes it possible to study the mechanisms behind these parallel evolutionary changes.

Another useful feature about sticklebacks, Jones said, is that despite having different forms, the different ecotypes, a form of an organism occupying a specific habitat, exist within close spatial proximity. In some cases these ecotypes even overlap with one another in the lower regions of rivers throughout the world. In these areas of overlap, hybridization and gene flow are common. In other words, the different ecotypes are not reproductively isolated. "That gene flow is super useful from a genomic point of view," she said, because it helps identify the parts of the genome that are important to the divergence between different ecotypes. Natural selection helps maintain differences at the loci that underlie the divergence to their varying ecotypes and habitats, she explained. The resulting "divergence in the face of ongoing gene flow" provides a good way to differentiate between "signal" and "noise" in genomic differences.

Furthermore, she said, just as with zebrafish, there is an entire suite of developmental and transgenic tools available for working with sticklebacks. Researchers have been able to easily adapt many of the transgenic tools used with zebrafish to sticklebacks.

Finally, sticklebacks can be bred in the lab where their environment can be manipulated. Since it is also relatively straightforward to study them in the wild, Jones said, the species "bridges the gap between the lab and the field very well."

Although there are many "axes of ecological divergence" for sticklebacks, Jones said, her lab focuses mostly on the marine–freshwater axis. Many years ago her lab did some whole-genome sequencing on both marine and freshwater sticklebacks, looking for regions of the genome where the freshwater fish were consistently different from their marine ancestors (Jones et al., 2012). What they found were blocks of DNA sequence that differed consistently between the marine and freshwater species.

Her team learned a lot from this initial work. First, there are many places across the genome where freshwater and marine fish have evolved divergent blocks of DNA, such that "any freshwater fish collected in the wild will be carrying the freshwater allele in as many as 81 different locations around the genome." Furthermore, marine versus freshwater adaptation is highly polygenic. Finally, the adaptive loci are primarily intergenic—that is, they fall between genes and into the non-coding parts of the genome, which are likely to have regulatory elements that control gene expression.

Understanding how these non-coding parts of the stickleback genome work has been a major focus of her team's work. One of their first experiments was to determine whether the divergence in gene expression they saw between marine and freshwater sticklebacks was controlled by cis-regulatory mechanisms or trans-regulatory mechanisms. Cis-regulatory mechanisms, she explained, are stretches of DNA in proximity of the gene being regulated. Trans-regulatory mechanisms include genes in a different part of the genome that, for example, code for proteins that bind to a promoter to activate transcription of the gene being regulated.

"We were interested in knowing whether it is the mutations in the proximal cis elements, for example, versus the trans-elements that are driving the parallel expression divergence that we see," Jones said. What she and her team found was that in the four different pairs of marine and freshwater fish they have studied, the expression divergence in each pair is controlled by cis-

regulatory variance (Verta and Jones, 2019). This cis-regulatory variance is primarily responsible for the difference in gene expression between marine and freshwater sticklebacks. This does not mean that the trans-regulatory factors do not play a role, but the cis-regulatory elements seem to be particularly important in explaining the divergence between the marine and freshwater sticklebacks.

Now Jones and her team are working to identify the specific regulatory elements responsible for the divergence between the two types of sticklebacks. They have identified blocks of DNA in which the regulatory elements reside, but those are large blocks—as much as 40 kb in length—and up to 4 percent of the nucleotides are divergent, so the challenge is to determine exactly where the relevant regulatory elements are.

They start by cloning the entire blocks from a freshwater and a marine stickleback, combining the regulatory element with a green fluorescent protein (GFP) reporter, and inserting these DNA stretches into embryos. They then look to see if the GFP was expressed in the embryo and if the marine and freshwater constructs show different expression patterns.

If there were differences in GFP expression related to a specific block, they then want to find exactly where the relevant promoter is in the block. They accomplished this by mapping potential functional elements using StickleCODE, an ENCODE-style (Encyclopedia of DNA Elements project) approach to identifying functional regulatory elements across the stickleback genome. They have used a number of different assays, including ATAC-seq (assay for transposase-accessible chromatin using sequencing) profiling, ChIP-seq (chromatin immunoprecipitation sequencing) on histone modifications, and RNA-seq (RNA sequencing) to look at expression levels, and they have done that in three different tissues in two different sexes in two different ecotypes. “And,” she added, “we’ve been doing a whole bunch of transgenic assays to develop what we see.”

The result has been a tremendous amount of data to analyze to identify regions that differ between the marine and freshwater ecotypes and thus to identify putatively divergent regulatory elements. Once the candidates have been identified, Jones and her team perform functional assays to test whether the blocks of DNA actually contain regulatory elements.

Jones showed one example where the GFP reporter caused the livers of the larvae to glow green when the region from the freshwater fish was used but not the marine sequence, indicating that the sequence regulated gene expression in the liver. Looking more closely, Jones found that the marine sequence had a deletion that the freshwater sequence did not. When the same deletion was made to the freshwater sequence, the gene expression in the liver disappeared. Furthermore, if the relevant part of the freshwater sequence was inserted into the marine sequence, the gene was expressed.

Finally, Jones spoke briefly about work to examine the regulation of open chromatin. The various sticklebacks show differences in their chromatin. Her group has been doing an allele-specific chromatin assay to determine where the regulatory control of chromatin accessibility is located. What they found with allele-specific ATAC was a pattern that was similar to what they saw for gene expression. That is, the marine–freshwater divergence is mostly due to cis-regulatory changes.

Looking to the future, Jones said that the next step in her work will be to do similar experiments under varying environmental conditions to understand how the regulatory landscape gets rewired when sticklebacks are living in different conditions. Adding variable environmental conditions to the mix will sharply increase the amount of data that will need to be taken and analyzed. This will make the work even more challenging, she said.

PHYLOGENETICS OF FLIGHTLESS BIRDS

Scott Edwards of Harvard University spoke about a project involving paleognathous birds, a group of mostly flightless birds that includes ostriches, emus, cassowaries, rheas, and kiwis. Although flightless birds are rather unwieldy as models, Edwards said, they are powerful in what they allow people to learn about the genetics of adaptive evolution—and, in particular, about the convergent changes in non-coding regulatory sequences that occurred in various taxa of birds that independently lost the ability to fly.

Edwards and his team are interested in various phenotypes mainly involving changes related to loss of flight. One such phenotypic change, for instance, is the loss or reduction of skeletal elements such as the keel, an extension of the sternum to which the flight muscles attach (de Bakker et al., 2013). Flying species such as pigeons have a prominent keel, but the keels of flightless birds tend to degenerate. "We're also looking at big differences in body size," he said, "as well as variable loss of forelimb elements."

The first step of the project, was to develop a substantial comparative genomics dataset of the paleognathous birds, including a draft genome of an extinct member of this clade, the little bush moa. They also developed a phylogeny of these paleognathous birds.

"Our first goal here was to determine whether flight was convergently lost or not," Edwards said. The phylogeny they developed (see Figure 4-1), indicated that there were a number of separate convergent losses of flight (Liu et al., 2010; Sackton et al., 2019). "This is in contrast to the paradigm for many decades, namely that all the flightless lineages of this clade descended from a flightless colony ancestor," he noted. One of the biggest surprises was the position in the phylogeny of the tinamou, a flying bird. Some in the field had thought that the tinamou, because they were able to fly, would have fallen outside the group.

FIGURE 4-1 Images of different types of paleognathous birds.
SOURCE: Scott Edwards presentation, slide 5.

The fact that the ability to fly had been lost multiple times meant that Edwards could use this convergent evolution to identify loci in the genome related to loss of flight—much like the process that Jones described for her work with sticklebacks, albeit on a different scale. His group focused primarily on conserved non-coding elements, explaining that these elements are easy to identify in the genome and noting that they often act as enhancers whose role is to bring together the transcriptional machinery to drive gene expression (Janes et al., 2011).

"We first used the signature of rapid evolution, either adaptive evolution or release of constraint, as an indicator of which non-coding elements might be important for loss-of-flight phenotypes," Edwards said. As an example, he mentioned a 250-base-pair, non-coding element that was highly conserved throughout the lineages of birds that maintained the ability to fly but that was visibly increasing in rate of appearance among some of the flightless lineages. The identification of such elements can be formalized in a statistical model, he said, explaining that two of the key things needed in this approach to functional genomics are statistical models that link genotype to phenotype and other models that detect changes in rate—and potentially changes in function—in non-coding elements (Hu et al., 2019).

An element like the aforementioned 250-base-pair sequence, which is conserved in most of bird evolution but rapidly changed in flightless lineages, "is the kind of signature that we start with," Edwards said. In particular, they look for regions with a large number of these accelerated non-coding elements and then look for nearby genes. What they have found is that nearby genes are often important in limb development.

To find candidates for genes important in the loss of flight, the group combined the rate acceleration data with two other datasets. First they generated a lot of ATAC-seq data and also used ChIP-seq data from the literature (Sackton et al., 2019). By looking at the overlap between candidates from the three datasets, they identified 42 conserved non-exonic elements as primary candidates. These 42 loci were prioritized for further study

To test one of those elements, they compared a version from the rhea, a flightless bird, with versions from the chicken and the tinamou, which are both able to fly or glide. They found that the versions from the two flying species were able to successfully drive gene expression in the developing forelimb of chickens, while the version from the rhea was not. This showed that at least one of the elements identified by their method has functional consequences for gene expression.

"We are scaling this approach up in collaboration with Emma Farley to try to interrogate hundreds, if not thousands, of these enhancers using the chicken limb as a developmental model," Edwards said. (For more information on Emma Farley's work, see Chapter 7.)

In another study they compared differences in gene expression between the forelimb and hind limb in five species of birds, two of them volant (chicken, tinamou) and three flightless (emu, ostrich, rhea). They found that there were relatively few differences between the two volant birds and a substantially larger number of differences in the other three flightless species.

To more completely understand this comparative dataset in detail, Edwards said, it is necessary to interpret it in the context of phylogeny. There are some initial tools for doing this in the literature on multi-variant evolution of complex phenotypic traits. Dean Adams, in particular, has done a lot of work on how to control for the many kinds of correlations not only between species in a phylogeny, but also between genes or other aspects of a complex multi-variant trait (Adams and Collyer, 2018; Bolnick et al., 2018).

In analyzing the ATAC-seq data from the forelimb/hindlimb study, Edwards concluded that there is a clear signature of convergent evolution leading to the three flightless species he and his lab studied.

There are other ways to determine what might have been the ancestral ATAC-seq profile, including the use of statistical approaches on assays of extinct populations. The main takeaway, however, was the importance of treating comparative data in an appropriately phylogenetic context.

Finally, he showed a slide that illustrated what the ATAC-seq data look like, with five species and about 363,000 sites of open and closed chromatin, for which he commented on its complexity. "One of my hopes," he said, "is that we can develop models that will allow us to pull out of this complex dataset, different groups of loci of open and closed chromatin, which may predict, in essence, the patterns of phenotype that we're interested in at the tips of the phylogeny." For that dataset, he commented, the trait of interest is binary—volant or flightless—but one could also imagine having a continuous trait such as body size. Edwards called for a model that could one day predict the loci responsible for traits on either end of a phylogenetic tree, but noted that this is not currently possible. The existing models are simple, with most of them designed for a small number of loci or characters in the genome, and they can predict simple traits. What is needed now, Edwards said, is to develop a series of methods that are tailored to high-dimensional genomic data. To get his paradigm to work, he said, it will be necessary to analyze large numbers of species across potentially multiple developmental stages. That sort of data is appearing in the literature now, he said, but it is data for which there are no good analytical models.

In conclusion, Edwards said, the non-coding genome seems to be important in the convergent loss of flight in the paleognathous birds, but new comparative models are needed to link genotype to phenotype in a phylogenetic context. "I do believe that phylogenetics is one of the most powerful approaches for that majority of biodiversity which isn't amenable to laboratory analysis," he said, "and yet the tools . . . aren't there yet."

ROLE OF CHROMATIN FOLDING IN GENE EXPRESSION

The final speaker in the session was Francois Spitz of the University of Chicago, who discussed the role of chromatin folding in gene expression.

Spitz began by noting, as other speakers had, that gene expression is shaped to a substantial degree by cis-regulatory elements, DNA regions that are near genes but separate from the core promoter regions. What is seen in many animals, especially vertebrates, is that gene regulation is extremely modular, and is frequently related to a large series of regulatory elements around the gene. More than 80 percent of the genomic variants identified by genome-wide association studies are far from the genes they regulate. Furthermore, mutations of these distant genome regions are a frequent cause of human developmental disorders and cancer (Uslu et al., 2014).

Fortunately for those interested in understanding these issues, a large toolbox has been developed to identify regulatory elements and characterize them functionally. One piece of information learned from using these tools, Spitz said, is that the regulatory elements are not necessarily controlling the genes closest to them. Indeed, it often happens that one of these elements skips over a number of intervening genes and interacts with a gene that is 100,000 or even 1 million bases away.

Understanding the enhancer–promoter interactions that drive gene expression requires examining the three-dimensional folding of the chromatin, the protein–DNA complex in which the DNA strands are packaged. That folding determines the physical proximity between regulatory elements, such as enhancers and promoters, and shapes their function.

For a long time this folding process was mysterious, he said, but the development of new technologies has made possible a much better understanding of how the genome is folded in the nucleus during interphase, the cell's resting period between divisions. Those technologies make it possible to see which regions along a chromosome are close to one another in the folded structure. What researchers have found is that the genome is organized in distinct and nested structures of different sizes. The structure looks different at different scales, but if one examines the structure at the scale of a few hundred thousand bases, what appears are a series of self-interacting domains called topologically associated domains, or TADs. In essence, a TAD is a stretch of DNA, typically hundreds of thousands of base pairs to 1 million or more base pairs long, in which the various sequences are much more likely to interact with one another when the chromatin is folded than with sequences that lie beyond the boundaries of the TAD (Symmons et al., 2014; Lupiáñez et al. 2015).

TADs are interesting, Spitz said, because they seem to define a space along the genome that is a regulatory domain, with a set of related genes and regulatory elements that work together. Furthermore, the boundaries seem to play an important role because deletions of the boundaries lead to "enhancer re-allocation," in which enhancers within the TAD can act on new target genes outside of it (Symmons et al., 2014; Lupiáñez et al. 2015; Tsujimura et al., 2015; Franke et al., 2016). "This enhancer relocation—or enhancer "hijacking"—has been implicated in growing numbers of human diseases," Spitz noted.

In addition, abolishing a TAD will prevent efficient long-distance interactions between an enhancer and a gene even if the distance along the genome between them is unchanged. The interaction can be partially rescued by bringing the enhancer and gene closer together, he commented.

Research has shown that TADs subdivide the genome into regulatory domains that are relatively invariant from one cell type to another. TADs have two basic functions. They ensure the specificity of long-distance enhancer–promoter interactions by preventing the activation of genes in adjacent domains, and they promote efficiency of the long-distance interactions, enabling distant elements to exert a robust influence on gene expression.

The interactions in the TAD are organized by what Spitz described as "the two major actors," the proteins CTCF and cohesin, which both accumulate at the boundaries of the TAD. Knocking out either CTCF or cohesin eliminates the TAD. Loss of CTCF reduces contacts between the regions inside the TAD (Nora et al., 2017), while removing cohesin leads to increased contacts between the regions inside the TAD and those outside it (Schwarzer et al., 2017). Interestingly, he said, when the cohesin is removed, "you lose the TADs, but it doesn't mean that you lose all the structures." In particular, the compartmentalization of the genome into active and inactive chromatin is reinforced, with the compartmental signals becoming stronger and fine-scale structures appearing within the former TAD.

These experiments, Spitz said, demonstrate that two distinct processes contribute to the three-dimensional organization of the genome. The first is based on the underlying chromatin structure and segregates regions into active and inactive compartments. "That's a system which will ensure maintenance of activities," he said. The second is mediated by cohesin and leads to the TAD structure. "This is a dynamic process which is enabling genes to scan their neighbors

looking for partners," he said. It results in a "mix and match system" that is important for dynamic changes of gene expression.

Much remains to be understood about how the various elements within a TAD interact and change in response to genetic variation. For example, the responsiveness to enhancers is not homogeneously distributed within a TAD, Spitz said, but depends on where a gene is located. Determining the factors that define the properties of enhancer–promoter interactions within a TAD will be essential in understanding the consequences of genomic variations.

One goal is to predict how changes in the linear genome sequence will affect the folding and enhancer–promoter contacts within a TAD. Spitz said that his group has begun examining this in mice by engineering variants and examining the results. "So far it's just a beginning, and there is no simple explanation." For example, their studies found that the distance between an enhancer and a target gene could be increased significantly—in one case from 1.7 to 2.7 Mb—and no changes occured, whereas significantly decreasing the distance led to a sharp drop in the expression of the gene. The complex rules of folding are not yet fully understood.

Looking to the future, Spitz said that one of the challenges will be to use tools such as Cas9-mediated genome editing to modify the genome and then assay the chromatin in ways that give insights into how the three-dimensional folding is controlled. This is particularly important, because various diseases lead to changes in the three-dimensional organization of the genome.

It will also be important to characterize three-dimensional genome folding across different types of organisms. Currently few experiments have been done in a variety of organisms, he said. Other organisms do show signs of interaction domains, and future work will uncover whether these domain share the same rules of formation, or if the genomes organize differently, based on the species.

DISCUSSION

The discussion that followed these talks touched on topics around functional gene regulation. Charles Danko of Cornell University and Emma Farley of the University of California, San Diego, had comments and questions related to the debate around the roles of topological domains in gene regulation. Danko began by explaining that there are disagreements in the literature related to the degree by which the interactions are observed between enhancers and genes. This makes it challenging to understand the role of many topological domains. He noted that Spitz highlighted some good examples of how topological domains can influence which enhancers interact with which genes, but Danko commented that there are many other factors that could disrupt the results or make them inconclusive.

Spitz responded by saying that there is no consensus in the community related to how accurately interactions between genes and enhancers can be measured, and researchers are just starting to address this question with new tools such as live imaging, high-resolution microscopy, and functional strategies. In relation to topological domains, he pointed out the fact that some loci provide a clear picture of what interaction is happening, while others show interactions that are diffuse and difficult to define.

Later in the discussion, Farley alluded to the varying functional significance of topological domain interactions. She asked for Spitz's thoughts on how to take these large datasets of interactions and discover those that have functional importance.

Spitz responded by first saying that because these ideas are new, he does not believe that the field has generated enough data to be able to fully explore these questions. He proposed a

bifurcated approach that looks at the problem on both a global and a focused level. The global approach involves a systematic exploration of the functional contribution of different ATAC-seq peaks. The focused part of the approach would dissect the interactions of model loci to understand how they operate at the levels of both sequence and structure. The plan would be to extrapolate general principles from this work and apply them to other loci.

Following this, Steven Henikoff asked Jones and Edwards. In both of their talks, he noted, they showed examples where there were large regions of conservation, but when they actually did the mapping, they found discrete sites of cis-regulatory elements. "I'm wondering why," he said, "because I wouldn't expect the whole region to be conserved, just the elements. And I'm wondering if it's something like TADs you might be looking at."

"We're not quite at TADs yet," Edwards answered. Instead, he believed that in his case what he was seeing was rate acceleration. "Those accelerated elements are accelerating for a variety of reasons. Sometimes it's simple point mutations. Other times it's clearly gene conversion events." So it was a variety of elements that were leading to the sequence differences.

For sticklebacks, Jones believed that it is related to the way the animals have evolved. They tend to evolve by making use of preexisting, or "standing," genetic variation, which repeatedly sweeps to fixation. The parts that get swept to fixation repeatedly accumulate mutations, which may be slightly beneficial. "So we believe that these blocks that we see are highly pleiotropic and have a lot of mutations that have been repeatedly pre-screened by selection in the course of evolution."

Sarah Kocher from Princeton University asked the panelists to compare how much enhancer variation contributes to the degradation of an existing phenotype versus the novelty of a new trait in the first place, and also what the relative contributions are of changes in master regulators versus downstream changes in non-coding sequences.

"I can't speak exactly to that," Jones said, "but we have some interesting data where we studied the degree of correlation in cis-regulation of gene expression in these marine–freshwater pairs." Upregulated genes tend to be strongly cis-regulated, suggesting that most upregulation is done through cis-regulatory mechanisms. That correlation does not appear, however, when the fish lose expression. "So there might be many ways to kill gene expression, but few ways to evolve or gain gene expression."

Edwards answered the first part of Kocher's question by saying that for non-coding elements it is often not clear whether adaptive evolution is driving changes or if it is just degradation of the element. Concerning the second part of the question related to master regulators, he said that it should be possible to combine something like ATAC-seq data and gene expression data to learn more about where the regulators are for gene expression. "There have been a few papers out on that using some really exciting approaches."

Gene Robinson from the University of Illinois asked the panel members to talk specifically about the sort of resources they need to do their functional genomics work. "What level of genomes do you need? Over what kind of diversity? . . . Just give us a flavor for what you see as where we need to be."

"For the experiments we are doing, we definitely need almost a chromosomal level assembly of the genome," Spitz answered. The genome can have some gaps, "but I think having high-quality genomes where you can see linear or 3-D order is essential." However, he added, the technology has now reached a point where the people in his lab can put together good genome assemblies themselves. The community can help by further annotating the genome, making sure the information is useful, and organizing it.

Jones echoed Spitz's reply. "With the technology changes we do our own de novo assemblies, and we've had . . . success with the linked read sequencing," she said. "It happens that the stickleback genome is small and well behaved." She also agreed with Spitz's comment about the community and the resources. "We have a number of different assembly versions kicking around for different ecotypes," she said. "I know that my academic brothers and sisters equally have that many assemblies. As a community, what we would like to have is more ability to put this together, including all the associated metadata that [go] with it." However, she said, as a principal investigator with limited resources, she finds it hard to attract someone to take on that job. She added that suggestions for how to get that kind of support from funding agencies would be much appreciated.

Edwards said that although he did not need particularly high-quality genomes for the research on flightless birds, better genomes could be valuable in other types of research. "I think that would be a great place for NSF [the National Science Foundation] to invest in," he said. "That is a great way to jump-start a lot of this—to get more genomes out there, with good annotations."

Spitz added that one thing that should be funded in parallel is the development of tools that would enable "even someone with limited knowledge of computational biology to navigate . . . genomic maps." It would be particularly valuable to provide ways to analyze things such as ATAC-seq maps or chromatin maps for different species in a comparative manner.

Edwards agreed, saying that visualization is important. "We've been fortunate enough to be able to produce a whole-genome alignment with a bunch of birds and to put it up into a browser, and it's available to the community. I think that can just lead to all kinds of discoveries. Just being able to view the data easily is a huge asset."

Lastly, Magnuson, the moderator, asked a question of all three panelists. Given that they were looking at folding, open chromatin, histone modifications, and the like, have they examined chromatin remodeling via specific chromatin remodeling complexes?

While Spitz had not, Jones reiterated that a number of trans-acting elements had been discovered during experiments where they are doing standard quantitative trait locus (QTL) mapping with chromatin profiles as a phenotype. "We're at the point of trying to identify whether they are these known complexes," she noted. "Stay tuned."

Edwards said that he really had not heard of the chromatin remodeling factors that Magnuson asked about, but he thought this illustrated an important point. Much of the ecological community gravitates toward assays that are straightforward and easily done on nonmodel systems. It would be useful to get molecular biologists and genome biologists together with people in the evolutionary community. "Just having interactions between folks for whom these factors are their daily bread versus folks like me who hadn't really heard of them before, that's going to catalyze a lot of important synergy," he said.

5

Advancing Research on the Environmental Regulation of Gene Function

After the session devoted to the contributions of non-coding DNA to phenotype, described in Chapter 4, the workshop moved on to the closely related topic of the environmental regulation of gene function. Because the environment exerts its effects on genes in large part through the non-coding elements of the genome, there was inevitably a great deal of overlap between the presentations in the two sessions. Research into how environmental factors affect gene function and, ultimately, an organism's phenotype requires studying how the genome's non-coding regions shape gene expression, and vice versa. Still, the two sessions were clearly differentiated by differing emphases. The presentations in Chapter 4 focused on non-coding elements and their interactions with the rest of the genome, whereas the talks described in this chapter were more focused on environmental factors and how they affect the regulation of gene function—and, ultimately, the organism's phenotype.

The main session described in this chapter was moderated by Philip Benfey of Duke University, and there were four speakers. Sarah Kocher of Princeton University described how the environment shapes social behavior in sweat bees, some of whom create highly organized societies like those of honeybees, while others are solitary. Joanna Kelley of Washington State University described how certain small fish have adapted to live in water that is highly sulfidic. Nathan Springer of the University of Minnesota discussed research aimed at understanding the interaction between genome and environment in corn plants, which has significant economic implications because of the importance of corn as a crop. And Trudy MacKay of Clemson University, who works with *Drosophila*, described how genotype, environment, and sex interact to determine an animal's phenotype. This chapter also includes a talk from David Page of the Whitehead Institute for Biomedical Research, whose talk on the phenotypic effects of sex differences fits the theme of this chapter, while his talk was given during a different session of the workshop.

FACTORS SHAPING VARIATION IN SOCIAL BEHAVIOR

Sarah Kocher's group at Princeton University studies social insects to gain insight into the factors that shape variation in social behavior. In particular, she said, they are interested in understanding what facilitates the transition between individuals that live and reproduce independently and individuals that come together to reproduce as a group.

Social insects, such as ants and honeybees, are excellent organisms in which to study social behavior because they are extreme examples of social living and have been helpful in understanding the molecular mechanisms behind social behaviors, such as caste differentiation. Using social insects, researchers who are interested in understanding what facilitates the

transition between solitary and social living must work in systems that include closely related species of both highly social animals and animals that are solitary or only weakly social.

With that in mind, Kocher and her group chose to work on a family of bees called the halictid bees, or the "sweat" bees. This group encompasses the full range of social behavior, she said. Some species are solitary and live and reproduce independently. Others are eusocial and have overlapping generations, cooperative brood care, and a reproductive queen or queens with the other members of a colony not reproducing. Still other species are socially polymorphic, so that in a single species there are some females that establish solitary nests and others that found eusocial nests. It is thought that eusociality evolved twice in the halictids independently and that it has been lost independently about a dozen different times. The result is a family of organisms that exhibits the entire spectrum of social behavior occurring both within species and between species.

When she started working on this system, she began with one of the socially polymorphic species, *Lasioglossum albipes*. Previous research had found that populations of this bee in the west of France are eusocial, while populations in the east of France are solitary (Plateaux-Quénu et al., 2000). Whether a particular population of these bees is eusocial or solitary appears to depend largely on environmental factors, Kocher said. In the solitary bees, females found nests independently and produce a single brood consisting of reproductives, both males and females. After the bees have overwintered, the fertilized females leave the nest to establish their own nests and start the cycle all over again. In a social nest, by contrast, the females produce workers first, followed by a reproductive brood. The queens make roughly twice as many reproductives in a social colony as in a solitary nest, and there is a huge fitness payoff to that, Kocher said, but because the queens in the social nests have to produce two broods, first workers and then reproductives, it takes them twice as long to reach reproductive maturity.

That difference in timing is related to the mean temperature of the area in which a nest is located, Kocher said. In those parts of France that are colder, on average, bees have fewer days when it is warm enough to forage during the day, and the populations are solitary. In the warmer areas where there are more days when it is warm enough to forage, the populations are eusocial. In short, she said, it seems that local adaptation is shaping the social/solitary divide.

Earlier research on these bees had shown that if eusocial and solitary populations are brought into the lab and reared under the same conditions, there is no plasticity in social behaviors. This showed that the social behavior of the populations is largely genetic (Plateaux-Quénu et al., 2000).

When she started working on this system, Kocher and her team had to build the genetic resources for the system from the ground up, so they started by generating a reference genome. Having done that and having figured out how to catch the bees in the wild, she brought 25 individuals back to the lab from each of six populations scattered across France, populations that were at different points on the eusocial/solitary spectrum. Her team then generated whole-genome sequences from the bees with the goal of identifying some of the genetic differences associated with the various social patterns.

The first thing they showed was that the populations were not simply incipient species (Kocher et al., 2018). Using a principal component analysis of the genomes showed that the populations did not cluster in the way that would be expected if they were speciating. Instead, Kocher said, "it seems like there have been repeated shifts in social behavior within the species, and this is a signature of local adaptation."

Next, she carried out a genome-wide association study (GWAS) looking for associations between genotype and social polymorphisms. Several regions across the genome show strong associations with social behavior, which is not surprising given that it is a complex behavior. But one of the most striking observations was a single window where there were seven single nucleotide polymorphisms (SNPs) located in non-coding regions of the gene syntaxin 1A (Munson, 2015). Because the SNPs were all non-coding variants, Kocher's first question was whether there were differences in the expression of syntaxin 1A in the natural populations that could be related to those SNPs. A quantitative polymerase chain reaction (qPCR) showed that social bees have higher expression levels of syntaxin 1A in their brains than solitary bees.

The next question was whether any of the seven SNPs could help to explain the observed variation in expression. She chose two SNPs that showed the greatest degree of differentiation between social and solitary populations and used luciferase assays to test for enhancer function. One of the SNPs lies within the first intron of syntaxin 1A, a sequence having enhancer activity, and that activity varies according to the allele that an individual carries. In particular, she found that the social allele drives higher levels of gene expression in the same direction seen in social populations of the bees.

Syntaxin is an interesting gene for a number of different reasons, Kocher said. It has been implicated in social behaviors in many different species, including other insect species (Chen et al., 2015) as well as in mice (Fujiwara et al., 2016), where decreasing its expression disrupts normal maternal care. Surprisingly, it has also been linked to autism in humans, in GWASs as well as gene expression studies

"And so," Kocher said, "our conclusion from this was that it seems like selection could be acting across these deep evolutionary timescales to shape variation in social behavior in both insects and vertebrate species like ourselves." For the past several years, she and her group have been trying to identify some of the core mechanisms that might shape variation in social behavior. In particular, working with postdocs Ben Rubin and Beryl Jones, Kocher's group has been carrying out comparative genomics on a collection of 19 different species of sweat bees, including *L. albipes*, in an effort to identify the genetic mechanisms shaping variation in their social behavior.

Among the genomic resources they have built are de novo assemblies for the 19 different sweat bee species with a combination of 10× sequencing and high-throughput chromosome conformation capture. The genome sizes average 414 megabases. "To improve annotation for each of these genomes," Kocher said, "we also did some tissue-specific RNA sequencing for as many species as we could get our hands on." They now have calculated a mean complete BUSCO, a metric assessing the completeness of a genome assembly, of 94 percent for the 19 different genomes, she said, adding that this is similar to what is seen in many of the other well-developed reference genomes in existence now.

With that dataset they began to investigate the natural selection that led to the different types of social behavior. In particular, in the evolutionary tree of the 19 species (see Figure 5-1), eusociality appeared twice, in two separate branches, and it also disappeared on a number of branches. In the two branches where eusociality developed, Kocher's team did tests for signatures of positive selection on each and then looked at the genes that intersected both of those branches. They found nine genes that showed positive selection on both branches, and because there were only nine, they did not see a strong signature of gene ontology (GO) enrichment at this level.

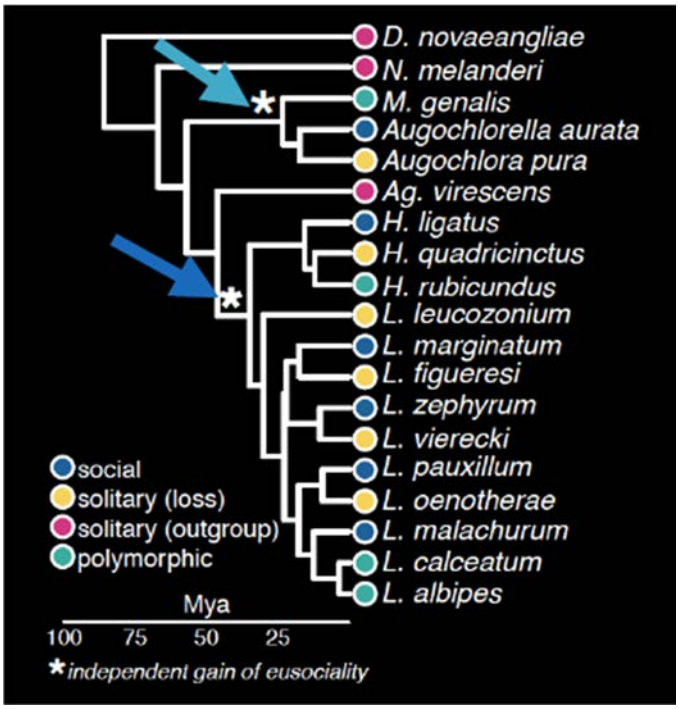

FIGURE 5-1 Evolutionary tree showing the different origins of eusocialtiy in the 19 different species of sweat bees studied by Kocher.
SOURCE: Sarah Kocher presentation, slide 26.

However, she said, most of the analytical power in the system is through the multiple losses of sociality. They tested for relaxed selection on each of the branches that represent losses of social behavior and compared those results with species that are eusocial and have maintained that behavior. This comparison identified 443 genes with relaxed selection associated with the loss of eusociality, Kocher said, and those genes were enriched for things such as chromosome condensation and DNA packaging, indicating an important role in chromatin accessibility.

One of her team's most exciting findings is that four of the nine genes that show the signatures of positive selection with the origins of social behavior also showed relaxation when social behavior was lost, thus identifying them as genes that play an important role in the evolution of social behavior. "I think that this represents a really powerful way . . . to identify some of the convergent mechanisms that shape the evolution of social behavior in the system," she said.

In the last part of her talk, Kocher turned to the role of gene regulation in the development of social behavior. A few studies over the past several years have suggested that changes in gene regulation are important in mediating social behavior and social interactions in many different species (Kapheim et al., 2015). When her team examined the proportion of transcription factor binding sites associated with either solitary or social genomes across the 19 different species, they found three times as many binding sites positively correlated with social genomes as binding sites associated with solitary genomes. This suggests, she said, that there has been some sort of expansion in the binding capacity of the social genomes—which in turn would point to this particular type of change in gene regulation as playing a role in mediating social behavior.

They also looked for signatures of positive selection on putative regulatory elements in the 19 genomes and found about 600 non-coding regions that could be aligned across all 19 different species and that show signatures of positive selection on the social lineages. These non-coding regions are enriched for "exactly the kinds of things that you would expect to find," she said. There is enrichment for neural functions, for things that are differentially expressed between social and solitary species, and for the genes linked to autism risk in humans. "So I think this is giving us some hint that maybe the continued development and maintenance of social behaviors may be fine-tuned by changes in gene regulation," she said.

The bigger take-home point, she said, is that both coding and non-coding sequences shape the evolution of social behavior. "If you're studying behavior, you have to acknowledge that there's a genetic component, but there are also environmental components that shape that variation."

Looking to the future, Kocher said that her team has been focusing on developing ways to generate comparable functional genomic datasets across the different species in order to examine how gene regulation might change with the gains and losses of social behavior. Instead of bringing 19 different species into the lab and doing a common garden—which seems really hard, Kocher said—"we've decided that we're going to take the lab to the bees." They have carried out pop-up common garden experiments where they use cages to observe nesting behaviors of bees in their natural environment. By comparing social and solitary bees in the same environment, they are able to look at how gene expression changes and how it varies across different species in the same environmental context.

In conclusion, Kocher listed some of her successes: using techniques from population and comparative genomics to provide insights into social evolution and building genomic tools that have allowed them to unmask some of the convergent mechanisms shaping social behavior. One of the challenges they face is generating comparative genomic datasets for 20 different species in a comparable way.

Finally, she said, it is important to have a strong community that can help build some of the necessary functional genomic tools. Because her lab is not large and there are only half a dozen labs in the entire country that work on sweat bees, they have turned to other systems, such as *Drosophila*, where genomes and genetic tools are readily available, and have adapted those tools to their own purposes.

HOW ENVIRONMENTAL FACTORS INFLUENCE A COMPLEX PHENOTYPE

In the next presentation, Joanna Kelley discussed how variation in natural systems can be used to understand complex phenotypes. She began by sketching out the usual frame for understanding the gene–phenotype connection: genetic variation leads to differences in gene expression and in protein abundance, which affect biochemical and physiological function, and, ultimately, the function of an organism and its fitness.

Environmental factors can affect every level of this hierarchy, she said, and the particular environmental factor she studies is the presence of a stressor, hydrogen sulfide, in aquatic environments. This stressor acts on organisms not only directly, but also through its effects on both the abiotic and biotic environment. These in turn influence all of the different hierarchical levels (Tobler et al., 2018; see Figure 5-2). "While we like to think about our beautiful linear path from genetic variation to phenotype," Kelley said, "in reality it's much, much more complex."

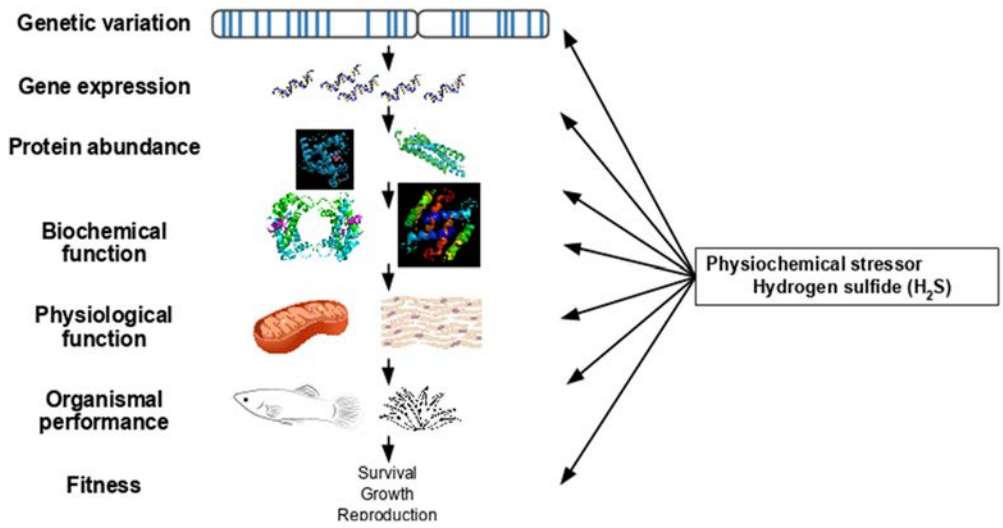

FIGURE 5-2 Representation of how an environmental stressor can influence various hierarchical levels of organism function.
SOURCES: Joanna Kelley presentation, slide 5; Tobler et al., 2018.

Hydrogen sulfide is both a toxicant and a signaling molecule (Tobler et al., 2006). In the regions in Mexico where Kelley carries out her research—and in various volcanic regions throughout the world—the concentrations of hydrogen sulfide (H_2S) in springs and other water bodies can be as high as 1 millimolar (Tobler et al., 2006). For organisms that are not adapted to these conditions, H_2S at 5 to 40 micromolar can be acutely toxic. It inhibits oxygen transport and cellular respiration and causes and aggravates hypoxia in aquatic environments.

There are, however, fish in many places that have adapted to high levels of H_2S, Kelley said. In particular, there are several different species in Mexico. Kelley studies those species adapted for living in a single location to better understand the connection between phenotype and genetic variation.

In particular, she studies three different sympatric lineages of poeciliids that live in divergent stream environments. Each of the three species has two populations, one that lives in a sulfidic environment and another that lives in a freshwater environment. The two environments are very close to one another—within about 200 meters. "So," Kelley said, "we're going to leverage this comparative contrast with having multiple different species in exactly the same environment to see how that phenotype of survival in hydrogen sulfide arises and connects to underlying genetic variation."

The three species that she studied are widely distributed across the large phylogeny of *Poeciliidae*. Because they are not closely related, their separate adaptations to H_2S are probably not due to adaptive genes moving from one species into another one. The three species, therefore, allow them to examine convergence and the origins of this phenotype.

Testing the tolerance to H_2S in the three species shows that all three of the freshwater populations die relatively quickly once the H_2S concentration reaches a certain level, while fish in the sulfidic populations can survive much longer. There are also consistent morphological differences between the sulfidic and freshwater populations, Kelley said, with similar changes in head and body shape across the three species, so there have been convergent shifts in body form as well.

Since much is known about how H_2S acts in organisms, she said, it is possible to make predictions about both the genetic and transcriptional changes as well as the physiological and biochemical changes that may be happening in the three species. For example, as H_2S comes into a system, it blocks cytochrome c oxidase (COX), which is a major target of hydrogen sulfide toxicity, so one prediction would be that COX would be modified somehow. There should also be up-regulation of the enzymes detoxifying hydrogen sulfide, down-regulation of enzymes producing hydrogen sulfide within the body, and differential regulation of other molecular targets.

Testing bears out these predictions, Kelley said. For instance, one of the major enzymes involved in the oxidation of hydrogen sulfide is sulfide–quinone reductase (SQR). Lab experiments with two populations of *Poecilia mexicana* showed that SQR activity in fishes from the sulfidic population increased as the H_2S concentration increased, whereas SQR activity in the freshwater population actually decreased with increasing hydrogen sulfide (Greenway et al., 2020).

Similarly, there was greater expression of H_2S-related genes in the sulfidic populations than in the freshwater populations, and this was true in all three species, showing convergent shifts in gene expression (Kelley et al., 2016). On the other hand, Kelley's team saw no evidence of down-regulation related to H_2S production.

Kelley pointed out that she was only discussing genes that she knew were related to H_2S somehow and that there is actually a huge list of genes that are expressed differently between the populations, but whose role in adapting to H_2S is unclear. "They don't fit into this nice framework of what I understand about their biology," she said, and are candidates for discovering other ways that these fish are adapting to their environments.

Since most of her group's experiments had been conducted on wild-caught individuals, it was impossible to tell whether the increased expression of certain genes in the sulfidic populations was because of evolved changes or whether these were responses to the fish being in water with high levels of H_2S. To determine which differences in gene expression between ecotypes were due to adaptation and which were due to plasticity, Kelley and her collaborators carried out a common garden experiment in which one of the species was brought into the lab where several generations were raised and then exposed to H_2S. This was done with the descendants of both sulfidic and freshwater populations, and there were also control groups where the fish were kept in water without H_2S. They found that many, but not all, of the genes whose expression differed between the two populations in the wild actually had evolved changes in gene expression rather than having expression plasticity (Passow et al., 2017).

Kelley noted that these evolved changes between the two populations, freshwater and sulfidic, in a single species are possible. Even though the two populations live only about 200 meters apart, there is little gene flow between the populations. There is some migration of individuals from sulfidic into non-sulfidic populations, so there are low levels of gene flow, but an ancestry analysis found that most of the non-sulfidic individuals had 100 percent non-sulfidic ancestry and all of the sulfidic individuals had 100 percent sulfidic ancestry.

Next Kelley described looking for highly differentiated regions in the genomes of these fish that might underlie the traits related to survival in a hydrogen sulfide environment. This was done by looking primarily for regions that differed in the same ways across all three poecilid. They found basically nothing. While the two populations from the three species have phenotypes that diverge in similar ways, the underlying genetic bases for those differences vary among the species. Kelley concluded that the populations took different genetic paths to get to the same place.

In closing, Kelley brought up two issues with which researchers in the field must grapple. First, with which phenotypes should they be working? Are the phenotypes of gene expression or even protein abundance really representative of organismal function? Perhaps, she said, but researchers must be careful because while some phenotypes are easier to measure than others, those are not always the best choice. "We may be most interested in physiological function, but those are hard phenotypes to assay," she said. "I think it is really important to think about which phenotypes we are really interested in and how much we can learn about organismal performance from gene expression or protein abundance."

The other issue, she said, is whether researchers should be limiting themselves to organisms that are tractable in the lab. "How tractable does an organism need to be? Are cells or tissues or cell culture sufficient to answer some of our questions?" The answers are not obvious, she said, but the question should be asked.

ENVIRONMENTAL REGULATION OF GENE FUNCTION IN AGRICULTURE

Moving from the first two presentations, which focused on insects and fish living in natural environments, Nathan Springer of the University of Minnesota spoke about crop plants, which have been carefully bred for optimized performance in farm fields. In particular, he focused on maize, which he described as "a species with wonderful diversity at many levels."

As a geneticist, Springer said, he is interested in learning about the molecular variants that underlie that diversity. Furthermore, he added, he is interested not just in how the variants directly influence crop traits, but also in how they combine to influence them. In corn, he explained, one frequently observes heterosis, or the combining of two different lines to generate a much different, generally superior line. That phenomenon requires an interaction between the variants present in the different parents.

Plants such as corn also offer a particularly good system in which to assess genotype-by-environment (GxE) effects, Springer said. There is a variety of highly inbred lines of corn that can be planted in different environments and monitored with ease because they stay in one place. Furthermore, he said, there is an economic incentive to learning more about corn. "We're trying to create GxE effects that would provide higher yield for these genotypes."

Introducing the key messages of his presentation, Springer said he would be focusing on three critical questions:

- What factors influence variation for gene expression patterns?
- How do genes acquire environmental responsiveness?
- How do we effectively link genomic variation to GxE effects?

Concerning differential gene expression, Springer said that like many of the previous speakers, he had studied how gene expression varies between closely related species or even between different individuals in a single species, and that he agreed with most of what had been said at the workshop. There are many genes that are expressed differently in different individuals, he said, and often variation in cis-regulation is the important factor. "One of the things that I think is important that I didn't appreciate for a long time," he said, "is that there are not stronger and weaker alleles."

To illustrate, he showed a dataset on allelic expression patterns in 23 different tissues taken from 3 different maize genotypes. There were more than 22,000 genes that were differentially

expressed in at least one tissue, which, he said, was not surprising. It is similar to the findings that other speakers had reported. What was surprising, however, was the consistency of patterns across tissues and developmental stages. Nearly 1 in 10 of the genes identified as being differentially expressed in at least one tissue were differentially expressed in more than 80 percent of the tissues in which they were expressed. More than 6 in 10 of the identified genes were differentially expressed in 20 to 80 percent of the tissues in which they were expressed. And about 3 in 10 of the genes were differentially expressed in less than 20 percent of the tissues in which they were expressed.

Furthermore, Springer added, it was not just that there was tissue specificity for differential expression, but the alleles often varied. Of all differentially expressed genes, 45 percent exhibit these so-called "mixed effects." He explained that, "it's not that one allele has a stronger promoter and just outcompetes the other; it's that they have different inputs into their tissue-specific expression levels."

One facet Springer is curious about is how plants generate regulatory novelty. In particular, he asks, how does a plant go from an ancestral state in which a particular allele does not respond to the environment, where it does not have cis-regulatory inputs, to a state where there is some difference in the cis-regulatory elements and the plant now responds to that environmental input? What happens to create this shift?

There may be some ways in which SNPs can do this, he said, but his suspicion is that it will be hard for single-base-pair changes to do that. So he brought up an idea that was put forward years ago by Barbara McClintock—that transposons, as they move throughout the genome, could shuffle or create novel regulatory elements (McClintock, 1950). As an example, he mentioned the Naiba family of elements, of which there are about 500 in the maize genome. More than 80 percent of genes that are near Naiba elements turn on in response to cold stress. Furthermore, alleles that have such an insertion often respond to cold. "So there's an association between the presence or absence of this transposon near a gene and the expression responsiveness," he said.

What he really wanted was to do this on a much larger scale. "I'm going to tell you about why this has been difficult, and why I think we're getting closer," he said.

For a decade or so, Springer said, corn researchers had a single reference genome, but within the past couple of years that has grown to about 40 fairly high-quality reference genomes, and that has revealed a great deal of additional complexity that must be taken into account. For instance, he described a study of the transposons in the genomes of two corn hybrids that revealed great transposon variation. The maize genome is about 2.5 Gb, and transposons accounted for about 1.35 Gb of that in each of the hybrid genomes. Of that 1.35 Gb, about 500 Mb of transposons—or about 20 percent of the entire genome—were not shared between the two genomes, or at least not at the same position, Springer said. Furthermore, when four hybrid genomes were analyzed, while a subset of transposons were shared across all four lines, more than 78 percent of the transposons varied across the four hybrids, and many of the transposons were present in only three, two, or one of the lines (Anderson et al., 2019).

This is important, he said, because in the four corn genomes they examined, more than 50 percent of the genes were located within 5 kb of a polymorphic transposon. So even if the transposons do not do anything themselves, the three-dimensional architecture, the local neighborhoods, and distances between genetic elements are radically different among these genotypes. That makes it difficult to align various maize genomes, he said. So the 40 reference genomes offer a wonderful opportunity to study and compare multiple genomes, but at the same

time they bring new challenges. "How do we even do an alignment of genomes that have highly variable content?"

Switching to the topic of chromatin, Springer said that it had been really heartening to hear all the various talks about using chromatic marks. "This gives us new insights into finding potential elements and functions in genomes," he said (Ricci et al., 2019). In maize, open chromatin makes up less than 1 percent of the genome, but nucleotide variation in such regions accounts for more than 40 percent of phenotypic variations (Rodgers-Melnick et al., 2016). Furthermore, more than 20 percent of all those open chromatin regions are found in transposons, which offers support for the idea that transposons might be carrying regulatory elements and influencing genes.

In the future, Springer said, generating genotypes will no longer be the expensive or difficult part of plant research. "Genotypes are relatively cheap," he said. "The hard part is measuring phenotype. What is the phenotypic outcome of a genome?" Plant researchers working for private-sector companies sometimes have access to large datasets with phenotypic data, but that is not common in the public sector. To address this issue, the Genomes to Fields initiative, a consortium of plant researchers, have been growing a set of 500 maize genotypes in about 20 to 30 locations over the past 5 summers. "We collect environmental data with the same brand of weather station in every field, monitoring core data on agronomic traits at this point," Springer said. "This is certainly not enough." Other phenotype-measuring technologies are needed, he said. For example, his lab flies drones over their field every 2 days gathering data on such things as how fast the plants are growing and canopy closure. They probably need additional sensors, he added, and they also need progress on how they store and share phenotypic data.

Summing up, Springer said, "We really need a better understanding of the phenotypic outcomes associated with our environmental data." As he pointed out, "in nature environments happen only once. You'll never get exactly the same environment the next year," he said. "So we're throwing away data right now that we should be gathering and keeping forever."

ENVIRONMENTAL FACTORS AFFECTING QUANTITATIVE TRAITS IN *DROSOPHILA*

Next, Trudy Mackay of Clemson University spoke about how environmental factors affect quantitative traits in *Drosophila*. Quantitative traits, she explained, are continuously distributed in natural populations and include such things as height, weight, and blood pressure. A century ago, Sir Ronald Fisher, a British statistician and geneticist, attributed this continuous variation to the presence of multiple genes that each contributed to the overall trait as well as non-genetic variation, which he referred to as "environmental variation" (Fisher, 1918). In Fisher's model, the genetic and environmental variation were separate factors that, when added together, produced the total phenotypic variation.

To illustrate, Mackay posited two genetically identical populations, one of which was in a hot environment and the other in a cold environment. The different temperature regimes caused the two populations to differ, on average, in a particular phenotype. This is *phenotypic plasticity*, Mackay said, and in her illustration the plasticity was such that the phenotype did not overlap between the two populations (see Figure 5-3).

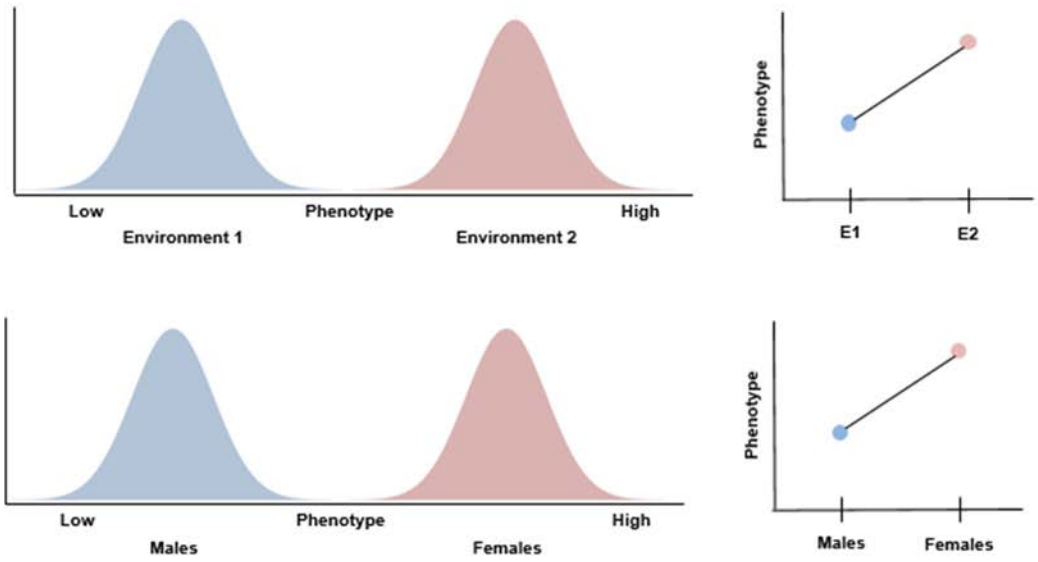

FIGURE 5-3 Display of variation in quantitative traits due to the environment (*top*) and sex (*bottom*). SOURCE: Trudy Mackay presentation, slide 3.

Organisms that have two sexes, male and female, can also exhibit sexual dimorphism, Mackay noted, and this can add extra variation to the phenotype (see Figure 4-3). In this situation, sex can be considered a genotype as well as an environment, she said. (See David Page's talk below for more discussion of the role of sex in shaping gene expression.)

Thus, in this model three different factors can lead to phenotypic variation: genetic and environmental variation and sexual dimorphism. In Fisher's view, these three factors would be additive—the total phenotypic variation would simply be the sum of the genetic and environmental variation plus the sexual dimorphism considered separately. However, the real world is more complicated and includes interactions among all three of these factors. For example, males and females might respond differently to increasing temperatures.

To illustrate how these varying interactions can contribute to phenotypic variation, Mackay discussed work that has been done in her lab with the fruit fly, *Drosophila melanogaster*. She began by describing the *D. melanogaster* Genetic Reference Panel (DGRP) that was developed in her lab (MacKay et al., 2012; Huang et al., 2014). The initial version had 205 inbred lines created by 20 generations of full brother–sister matings. The collection of different lines is valuable because the members of each line are largely genetically identical, but they are different from individuals from other lines, and they can be raised in multiple environments in order to study the interaction between genetics and the environment on phenotypes.

In one such experiment, Mackay exposed 72 males and 72 females from each of 186 lines to each of three different thermal environments—18°C, 25°C, and 28°C—and observed how long they lived (Huang et al., 2020). The ones kept at 18°C had the longest lifespans, followed by those at 25°C and 28°C. After collecting these data, Mackay and her team analyzed them to uncover the effects of genotype, environment, sex, and interactions among the three factors on the lifespans of the flies.

They found clear genetic variation within each temperature—some lines survived significantly longer than other lines—as well as sexual dimorphism and phenotypic plasticity (Huang et al., 2020). Importantly the interactions among the three factors were also significant.

There was, for example, genetic variation in sexual dimorphism, so that the differences between the sexes varied in different environments. There was also genetic variation in phenotypic plasticity, and the three-way interaction was highly significant. "What this means," Mackay said, "is that if I had information on the lifespan of a particular genotype female at 25 degrees, that tells me very little about that same genotype in other environments."

An additional complication is how variable the lifespans of the flies are within different lines. "This is obvious when you're measuring the flies," she said. "Sometimes the first one dies when it's 3 days old, and the last one lives to 100 days, whereas for some lines you come in one day, they're all alive, and the next day they're all dead." This characteristic is clearly genetic, she noted, but it also varies according to the environment—this "micro-environmental plasticity," as it is called, is different for the same line of flies in different environments (Huang et al., 2020). And while there is no overall sexual dimorphism, Mackay said, the amount of lifespan variation in each sex differs according to the environment. "So life is a little bit more complicated at the level of quantitative traits than one might think," she said.

Next Mackay described an experiment that examined how much of the genome exhibits variation in gene expression in different environments (Zhou et al., 2012). Her team worked with an outbred *Drosophila* population that was a mix of 40 DGRP lines—"all sorts of different genotypes, all mixed up," she said—rather than working with single inbred lines. The flies were exposed to 20 different environments: a control and 19 that involved distinct environmental stressors ranging from heat shock to social crowding. The researchers used microarrays to measure gene expression across the genome. Surprisingly, only about 10 percent of the transcripts show environmental plasticity. By contrast, she said, nearly half of the transcripts exhibited sexual dimorphism, and a few of them had sexual dimorphism that differed across environments. Their conclusion was that the majority of the transcriptome was robust to environmental fluctuations, at least in this outbred population where the effects were averaged over all the different genotypes that they observed.

To close, Mackay offered brief descriptions of two other studies of variation in transcription. In one, her group examined the response of fly brains to cocaine using single-cell expression techniques. They found 691 unique genes expressed differentially in males and 322 in females. The response was highly sexually dimorphic and cell-type specific.

Noting that the first study was done with just one genotype, Mackay described a second with young adult flies from 200 different DGRP lines. RNA sequencing was used to examine gene expression in both sexes of all these lines (Everett et al., 2020). They found that much of the transcriptome was highly sexually dimorphic and that there was enormous genetic variation and sexual dimorphism at the level of transcription—echoing what they had seen at the level of quantitative traits.

In concluding her presentation, Mackay offered her thoughts on the roadblocks facing this sort of work. Generally speaking, she agreed with what the previous speakers had said. "I think the roadblocks are not intellectual," she said, "but they're matters of scale and the money needed to do these kinds of experiments." The factorial experimental design needed to estimate and map the genetic base of environmental interactions across multiple environments, sexes, genotypes, tissues, developmental stages, exposure time courses, and multiple -omic levels is expensive to carry out.

"For example," she said, "we would want to look at bulk RNA, single cells, micro RNAs, epigenetic marks, metabolites, proteins, nuclear architecture. That would give us a very full picture of the effect of environmental variation on transcription and how that would relate to

quantitative trait." That list of items mirrors a list of items from the *Drosophila* Encyclopedia of DNA Elements (ENCODE) project, she said. "I think there would be great value in actually doing a similar large community-based project with genetic variation overlaid."

Sex Differences in Comparative Functional Genomics

Sex differences should be front and center in comparative functional genomics research. That was the thesis that David Page set forth at the beginning of his talk, and he spent the rest of his presentation expanding on and supporting that suggestion.

Although the X and Y chromosomes are considered to be the sex chromosomes, Page said, "I want to argue . . . that every chromosome is a 'sex chromosome' and that the sex chromosomes have no monopoly, and not even any statistically discernible specialization, in the matters of sex." In particular, he suggested that the study of sex differences should be moved to the entirety of the organism and should be at the center of functional genomics research.

The first challenge, Page said, "is to decide what we even mean by sex differences." According to the standard textbook account of mammalian sex determination—which can be generalized to other organisms—during the first 6 weeks of human development the XX and XY fetuses are anatomically indistinguishable. Then, around week 7, the bipotential gonad begins to take on the microscopic appearance of the testis or ovary. The remainder of the classic strict binary of sex determination follows this first appearance of sex organs.

The problem is, however, that not all sex differences are strictly binary, although the textbooks tend to focus on those traits that are. To illustrate, he spoke about height and size. "Males tend to be a little larger than females," he said. "This is actually true of most mammals, not all, but most," with the males tending to be 4 to 8 percent larger (Crow, 1997).

There are many other such traits in humans that show sex differences. "For every man who has lupus, there are six women," Page said. "Or if we flip it around to autism spectrum disorders, for every girl who's diagnosed with autism, there are four boys." However, he added, models of sex determination and sex differentiation, at least in mammals, do not generally accommodate these sorts of non-binary differences.

He has been working with others to develop a new model of sex determination and differentiation. The model will be fluid and dynamic, instead of binary. Furthermore, he said, the model will not be restricted to a few cell types in the reproductive tract but rather will encompass all cell types across the entire body and focus on the entire genome.

The model is based on three key conjectures. The first is that autosomes are read differently in males and females. That is, it is not just the X and Y chromosomes that have gene expression that differs between males and females, but in fact all of the chromosomes are read out a bit differently in the two sexes.

The second conjecture is that this has been true throughout human evolutionary history, going back 600 million years, or since before there were sex chromosomes. It was the appearance of structurally distinct gametes—eggs and sperm—that was the defining feature of the origin of males and females, Page said, and this happened about 600 million years ago.

The final conjecture is that in mammals the autosomes are read differently in every tissue, in every cell type, and at all developmental stages. "I don't have all the evidence to support that, of course," Page said, "but it's a sort of a guiding hypothesis, a speculative form."

To illustrate the sorts of research that can be motivated by these conjectures, Page described some work that was motivated by the general question of how autosomal gene expression differs

between males and females. His team sought to answer three specific questions: How does autosomal gene expression differ in organs that are outside the reproductive tract, that is, in organs that are "the same" in male and females? Are sex biases in gene expression conserved between humans and other mammals used in basic research and in pharmaceutical trials? And do sex differences in gene expression across the genome contribute to sex differences in a trait?

To address the second question, the team surveyed sex differences in gene expression across humans and four other animals—rhesus macaque, mouse, rat, and dog—which all had a common ancestor around 100 million years ago (Naqvi et al., 2019). They looked at 12 tissues present in both sexes that would typically be thought of as being the same in males and females: tissue from the brain, pituitary gland, thyroid gland, heart, lung, liver, spleen, adrenal gland, colon, skin, skeletal muscle, and adipose tissue. For humans they reanalyzed data from the Genotype-Tissue Expression (GTEx) Consortium, while they generated new RNA-sequencing (RNA-seq) data for the other four species.

The team clustered 72 human and 277 non-human mammalian RNA-seq libraries based on nearly 13,000 genes that were one-to-one orthologs across the species. Unsurprisingly, the samples clustered first by tissue and then by species. But after that they clustered—most of the time—by sex. Page notes "that sex is a subtle third-order determinant of each tissue's transcriptome. We can identify just about any tissue in any of these species as being male or female based on its transcriptome, but it is subtle." He noted the major contrast between the subtle sex differences one sees in gene expression and the sometimes large sex differences that appear in phenotypes.

The next question, he said, is whether such sex differences in autosomal gene expression actually contribute to sex differences in a trait. As an answer, he described a study that his lab did on sex differences in height. This is an exhaustively studied trait in humans, Page noted. There are at least 700 genes that have been implicated as making minuscule contributions to height variation in both men and women, and it is the same 700 genes for both sexes (Rawlik et al., 2016; Sanjak et al., 2018; Sidorenko et al., 2019). In theory, there could have been somewhat different sets of genes responsible for height differences in males and females, but that is not what has been found. Despite the enormous literature on this topic, no one has come up with an explanation for the average 5-inch height difference between men and women.

His lab has made some progress in that area, Page said, and they started by describing work done on a single gene—the gene for a transcription factor called LCORL. This transcription factor has been genetically associated with increased body size in humans, cattle, and horses, with increased expression of this gene being associated with a decrease in size. What Page's lab noticed is that LCORL is expressed at slightly higher levels in females than in males—a difference that would tend to result in a slightly smaller body size.

Furthermore, they found that this was generally true among genes that influence height. Other autosomal genes that are expressed at higher levels in females tend to decrease height, while, conversely, autosomal genes with a conserved male bias in expression tend to increase height. "It's not invariably true," Page said, "but there was a tendency." When his team summed the effects on height across the hundreds of implicated genes, they found that the result of the differences in expression explained 12 percent of the observed difference in the average heights between females and males.

The next steps, Page said, will be to collect data from more tissues, more cell types, more developmental stages, and more species, and to explore whether sex biases in gene expression explain sex differences in traits other than height. Furthermore, he would like to understand the

roots of this sex bias in gene expression. Is it a consequence of sex hormones, of sex chromosomes, of both, or is it perhaps something else?

DISCUSSION

The session moderator, Philip Benfey, opened the discussion by referring to the keynote talk by Aviv Regev, who had suggested that researchers in functional genomics think "in terms of a random sampling across a large data population versus a more defined factorial approach," as Benfey phrased it. He asked the panelists which approach would be most effective in their experimental systems. Mackay and Kelley responded that it would be difficult for them to use random design in large natural systems to answer the questions they are interested in, while Springer said he could see value in both approaches and which one best depended on what he wanted to learn. Kocher agreed with Springer. "I think a lot of it really depends on the type of question that we're asking," she said.

Scott Edwards asked about measuring gene expression in the wild versus in the lab. What is the value of studying gene expression in the wild, he asked, and is it only valuable if one can also do a common garden experiment? Kelley said that there is great value to doing it in the wild, but the limitation is that it can be impossible to tell if patterns of expression in the wild are due to adaptation or plasticity. Springer agreed about the value of doing experiments in the wild but added that the greatest value comes when he can also do controlled experiments to get additional information that cannot be gleaned in the wild. Kocher added that it is important to go into the wild and sample individuals there because it provides a baseline for things that are studied in the lab.

Gene Robinson of the University of Illinois, following up on a comment by Paul Katz of the University of Massachusetts Amherst, spoke about Kocher's finding that regulation of the levels of syntaxin 1A affected sociality in the sweat bees. This finding illustrates, Robinson said, that phenotypic differences in natural populations are often due to differences in expression levels rather than to the presence or absence of a gene. Therefore, functional genomics tools need to be able to dial up or dial down expression rather than just knocking out or adding genes to the genome. Kelley agreed, saying that the ability to turn expression up or down is hugely important and that many of the genes that she examines are essential for survival.

Benfey made an observation and asked a question. "I've been struck over the last 2 days that most of the discussion has been around transcriptional responses," he said. Is the current focus on transcriptional responses truly a reflection of where the most valuable investigations lie? Mackay suggested that the current popularity of studying transcriptional responses is due to the fact that this is what is easiest to measure at this point in time, but that in the future more attention should be paid to networks. Kocher added that one can learn a lot from studying sequence variation rather than just transcriptional variation. If researchers can start to identify transcription factors that have been under selection or changes in binding motifs across genome-wide scales, then that can give us something to grab onto when thinking about how these networks might change over time.

Gary Churchill from The Jackson Laboratory described observations he has made about GxE interactions seen with experiments in mice. Noting that, as Mackay had pointed out, sex can be thought of as an environmental variable, mice fed different diets were experiencing another environmental variable, and mice that had been allowed to age were experiencing a third variable. What he and his colleagues have seen is that GxE interactions for sex are largely local and are mediated through RNA to protein; GxE effects of diet are a mix of local interactions

mediated through RNA and distal, direct protein-to-protein interactions. The age effects are all distal. "My interpretation of this," he said, "is that our genomes are exquisitely programmed to be one sex or the other. They're pretty well programmed to respond to diets. But they're not designed to age at all." Then he asked the panel what their thoughts were.

Mackay responded that the sort of complexity that Churchill and others were describing is only going to be understood "when we can stop going from one variant to one transcript and to one trait, and start using our data to infer transcriptional genetic networks based on naturally occurring genetic variation and how those entire regulatory networks translate to organismal phenotypes." With enough data, she said, it should be possible to map expression quantitative trait loci (eQTLs) that have cis effects, eQTLs that have trans effects, figure out which eQTLs have both, and then start to infer a cis–trans-regulatory network. "We need to start to embrace network theory in order to understand how complex gene regulatory networks are going to affect complex traits," she said.

Kelley took a different angle in discussing how to deal with complexity. "To answer some of these complex questions or get at these complex phenotypes," she said, "we really need integrative biology. We need to get mathematicians in the room who are going to develop these network models, we need physiologists, we need behavioral biologists." In short, she said, it will be important to get all sorts of different disciplines into the same room to address the challenge of complexity.

Katz added that another problem is simply identifying what the bigger questions are. "In other words," he said, "you have to be exposed to a genome researcher to even understand what your questions are."

6

Predicting Current and Future Sources of Variation in Quantitative Traits

Early on the workshop's second day, Patricia Wittkopp of the University of Michigan delivered the workshop's second keynote speech. It was a fitting transition from the first half of the workshop to the second because Wittkopp's talk included both a look at the current state of functional genomics, including what has led up to it, and a discussion of what will be required to move effectively into the future and what that future may offer as functional genomics relies more and more on large-scale data and tools. Wittkopp's talk focused on functional validation and the testing of hypotheses derived from large-scale experiments. The talk acted as a complement to the keynote from Regev (see Chapter 2), who emphasized large-scale studies, computational inference, and pattern identification. The methodologies and research behind identifying patterns in datasets from Regev and the functional validation of genomic information from Wittkopp are both important parts of functional genomics research.

Wittkopp started by explaining the two pieces of her title. "Predicting current sources of variation in quantitative traits" refers to the ability to predict where the genetic basis of a particular variation or trait will lie within the genome—within a certain type of gene, for instance, or inside non-coding sequences. And "predicting future variation" refers to making predictive statements about what types of changes should arise under particular evolutionary scenarios, certain types of selection pressure, and so forth.

"Our way forward toward these things involves three key steps," she said. The first, which has been happening in the field for quite a while, is carrying out case studies. "I don't think we're done with case studies," she said. "I think they have a really critical role in generating the information we need to accomplish those broader goals."

Once enough case studies have been accumulated, the next step will be to develop a framework to integrate them. This will involve modeling that is based on both empirical observations and a theoretical understanding of how the evolutionary process, development, and gene regulation work.

The final step will be testing the predictions that the model makes. Such tests could take many forms, Wittkopp said. They could entail making predictions about where variations should be and then mapping those variations, for instance, or they might involve carrying out experimental evolution studies.

Over the past 10 to 15 years there has been a tremendous increase in the number of genetic changes that have been associated with trait variations, she said, mentioning in particular the website gephebase.org, which compiles genotype–phenotype relationships from the scientific literature (Martin and Orgogozo, 2013). It is important, Wittkopp said, that funding be available not only for generating data on genotype–phenotype relationships, but also for projects like GePheBase that collect the data in one place.

Furthermore, she reiterated that there is still a need for more case studies because only with a large enough collection of case studies is it possible to ask such questions as whether the

genetic basis differs for traits at different timescales or trait functions. In deciding which case studies to carry out, she added, it is important to be strategic about figuring out which traits deserve greater attention and also to be mindful of ascertainment bias, which arises when data are collected not from a random sampling of a population but from a population that has been biased by choices made in the sampling. In this case, existing datasets are heavily biased toward candidate genes and coding sequences, among other things, and so it is important to be cautious when synthesizing those data.

DISCOVERING THE GENETIC BASIS OF A CHANGE: AN EXAMPLE

Having finished her introductory remarks, Wittkopp turned to the first part of her talk in which she described an example from her own lab of looking for the genetic basis of a change in an organism. The story she told echoed many of the themes that emerged in the sessions in the first half of the workshop (i.e., the sessions described in Chapters 2, 3, and 4).

The trait that Wittkopp studied was a pigmentation difference between two members of the *Drosophila* virlis group, *Drosophila americana* and *Drosophila novamexicana*. *D. americana* has a brown body, while *D. novamexicana* has a yellow body. These are sister species, and they can be interbred and create fertile offspring. The F1 hybrids are closer in color to *D. americana*, Wittkopp said, so the dark pigmentation is largely dominant (see Figure 6-1). When the F1 hybrids are crossed back with *D. novamexicana*, it creates a population of recombinant individuals that each has a shuffled combination of the two species' genomes and then one set of chromosomes from *D. novamexicana*. The distribution of pigmentation phenotypes in this back-crossed population is not continuous, Wittkopp said, but instead the phenotypes fell into about five distinct classes, suggesting that only two, three, or possibly four genes are contributing to the pigmentation patterns (Wittkopp et al., 2003).

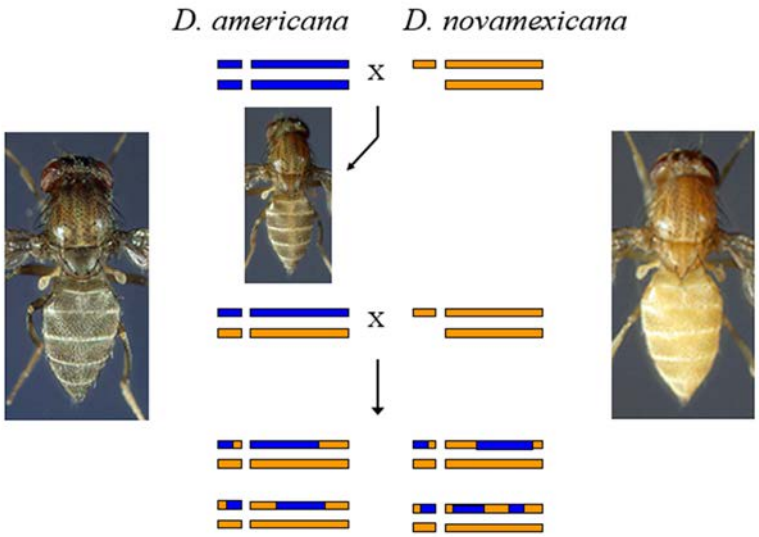

FIGURE 6-1 Breeding of *D. americana* with *D. novamexicana*, two different species of *Drosophila*, to look at the outcome of the coloration phenotypes.
SOURCE: Patricia Wittkopp presentation, slide 7.

When her lab started working on this trait in 2005, Wittkopp said, they did not have genomes for those species, so they did a very sparse quantitative trait locus (QTL) map that was focused on particular candidate genes. They found that two of those genes, *tan* and *ebony*, were strongly associated with pigmentation—that is, when the region of the genome that contained *tan* or *ebony* was inherited from *D. americana*, which is the dark species, the flies were darker, and when that region was inherited from *D. novamexicana*, they were lighter (Wittkopp et al., 2009).

The two genes, *tan* and *ebony*, encode enzymes that are involved in the pigment synthesis pathway, Wittkopp said. The *tan* and *ebony* genes catalyze a reversible reaction that controls the balance between yellow and dark black and brown pigments (Cooley et al., 2012). Noting that the gene names in flies are backwards because they are named for the mutant phenotype, Wittkopp explained that the function of the *ebony* gene is to make yellow pigment and the function of the *yellow* gene is to make black pigment.

Wittkopp also examined the sources of the differences in gene expression using an allele-specific expression trick that she had developed as a postdoc. By working in the F1 hybrid, she explained, she "put the two different species' cis-regulatory alleles in the same actual cells so they're exposed to the same trans-regulatory factors, and any differences in expression between the two alleles must come from cis-regulatory differences between those alleles." She found that most (in the case of *tan*) or all (in the case of *ebony*) of the expression differences between the species were due to cis-regulatory changes (Wittkopp et al., 2004).

At that point she also wanted to locate the nucleotide changes that cause the difference in gene expression, which presumably cause the pigmentation differences. "And this has been a mix of success and failure," she said, "so that's what I'll share here."

One of the first steps in the effort to find those nucleotide changes was to use a transgenic approach. Her lab cloned the *tan* genes from *D. americana* and *D. novamexicana* along with the flanking noncoding sequences—about 14 kb of genome—and transformed them into *D. melanogaster*, the fly that has served as a model organism for decades and thus is most convenient to work with. Working with a *D. melanogaster* with no *tan* gene and no *yellow* gene, they added a *tan* gene from *D. novamexicana*, the yellow species. As a result they produced the *tan* phenotype and got some pigmentation. When they added the *D. americana tan* gene, including its flanking non-coding sequences, they got darker pigmentation which was consistent with what they had seen in the two species. They repeated the work but put the transgenes in *D. novamexicana* and saw the same difference in pigmentation. Unfortunately, Wittkopp said, they were not able to do the same thing with the *ebony* gene because it is too large. "We couldn't get it to clone, we couldn't get it to transform in [*D. melanogaster*]," she said. "That's a technical/biological limitation that we hit in this work."

To figure out where within *tan* the functional changes are, Wittkopp said, she next used a fine-scale genetic mapping approach. Her lab had created introgression lines with flies that were mostly *D. novamexicana* but contained a small region of the genome, including *tan*, from *D. americana*. They crossed these with *D. novamexicana* and screened 10,000 flies, categorizing each one as having either light or dark pigmentation and then genotyping each fly at markers near *tan* to try to find recombinants within *tan*. The result was two "very informative recombinants," Wittkopp said.

They isolated the first intron. There are no sequence differences between species in the first exon, so the first intron seems to be where the functional differences between the two species are located. Flies with the *D. novamexicana* version have lighter pigmentation, while those with the *D. americana* version have darker pigmentation.

They actually got lucky in being able to locate the source of the functional difference, Wittkopp said, because recombinants that are informative for fine-scale mapping are rare. Getting enough recombinants and being able to efficiently choose the recombinants that are informative for the trait of interest are major bottlenecks that could benefit from some technological advances, she said. It would be useful, for example, to be able to target recombination to a particular region of the genome. "We have opportunities here to speed up this part of the process with technology," she said.

With the 14-kb transgene in hand, Wittkopp and her group set out to swap the intron and parts of the intron between the different flies. At the time—about 10 years ago—they had to use restriction enzyme–based cloning, which was difficult. Wittkopp commented that with the technological advances that have occurred in the meantime, however, the process is not as challenging now. Unfortunately, Wittkopp said, by the time they were ready to put these transgenes into the flies, the original line they had been working with had died. Those flies had landing sites that could be targeted so that the transgenes would go in the same spot in the genome, but with a different line of flies the transgenes ended up in another landing site—and no pigmentation difference was produced. "So this was obviously a little heartbreaking and depressing," she said. This points to another technical barrier, Wittkopp said—that even if a researcher can insert transgenes into a species of interest, it matters where they land, and the exact landing site can affect the results (John et al., 2016).

In response to that result, Wittkopp said, she and her team decided to put the transgenes into a number of landing sites and see what happened in different parts of the genome. In three of the other four sites they tried, they did see the expected differences in pigmentation between the two species, which showed that the relevant factor was the expression level of the transgenes. Some regions were just better able to express the transgene at high levels than others, she said.

When looking more closely, however, they found a pattern they could not explain. A transgene that should have led to darker pigmentation led to lighter, for instance. And the effect depended on how long the transplanted DNA was. "Here we have *americana*, we put in a little bit of *novamexicana*, it should lighten it, it does," she said. "We put in a bigger chunk of *novamexicana*. Nope, darkens it again." They stopped pursuing this line, she said, but she believes there are important messages in this result: there must be multiple sites that are affecting pigmentation, they are probably closely linked, and they have effects in opposite directions. Thus, this region of the genome presents an extremely complex situation.

Other researchers have reported similar findings, she said. When one is doing a broader-scale QTL map, the areas with opposite effects can cancel each other out, which can make it easy to miss them (Bernstein et al., 2019). Wittkopp's lab has observed the same sort of thing in three different strains of yeast (Metzger and Wittkopp, 2019). In all three cases, she said, the number of regions that increased gene expression were approximately balanced out by those that decreased gene expression. "So we have a lot more variants within the genome that are affecting expression of one gene than we knew about," she said. "When we just look at the expression level of these strains they're actually not that different from each other, yet there's still lots of genetic variation that's present."

The presence of these multiple linked variants with opposing effects is, therefore, a major biological challenge in understanding how phenotypic variations arise. She also added a couple of other biological challenges to the list. One is the sheer number of loci that can harbor variation, and another is the fact that many of these variants are relatively rare but collectively can play a major role. For example, she said, a group led by Leonid Kruglyak of the University

of California, Los Angeles, examined trait variation in yeast and concluded that rare variants explain much of the variation (Bloom et al., 2019). This is important, Wittkopp said, because such rare variants are missed by genome-wide association studies (GWASs). Other challenges are the presence of epistatic interactions among genes that affect the measurement of allelic differences and trait variation, as well as the fact that many phenotypes are sensitive to fluctuations in the environment. These are parts of biology that make it much harder to understand the relationships between genotypes and phenotypes, especially with the technologies that are available today.

Wrapping up the first part of her presentation, Wittkopp offered the take-home message that the challenges in mapping quantitative trait nucleotides do not end with clearing the transgenic hurdle. There are many more technical and, especially, biological challenges to overcome. And the grand challenge, she said, is how to scale mapping—to speed up the process of finding the nucleotides that matter. Ultimately, she said, the gold standard is functional testing with allele swaps where only one nucleotide in the genome is changed, and her lab has developed a strategy to do that. "We basically swap in a marker, swap it back out, and in the process of the swaps restore the scars that were created the first time in," she said (Lamb et al., 2017).

EXPLORING THE REGULATION OF GENE EXPRESSION

In the second part of her talk, Wittkopp spoke about the work she has been doing to understand how the regulation of gene expression evolves. In studying that evolution, she said, she has used several approaches, including single-gene strategies, genomic strategies, and single-mutation strategies. Her ultimate goal, she said, is to study the universal rules of life that apply to multiple organisms.

Generally speaking, the people who study the evolution of phenotypic differences believe that these differences are less likely to be caused by changes in coding sequences than by changes in expression. This may be because creating changes in gene expression makes it possible to alter a gene's function in one part of the organism but not others, while a change in the gene's coding sequence changes that gene everywhere.

Wittkopp began by discussing some of the single-gene strategies she has used to study the evolution of gene regulation, particularly her studies of enhancers. Much comparative work on enhancers starts by using sequence conservation to find orthologous enhancers (Wittkopp and Kalay, 2011). This works well in certain invertebrates, although not so well in flies.

Researchers in the field take it for granted that if enhancers are in a particular order for one species, they will follow that order in another species, and that assumption is often validated. "People have taken these pieces of DNA from other species and shown that they drive the same patterns," she said. But when she took a systematic approach to looking at orthologous enhancers, she "got a bit of a surprise."

They were looking at enhancers of the *yellow* gene, which had been pretty well characterized, and they found that, depending on which species they were examining, gene activity was driven by different stretches; sometimes the activity was driven primarily by the intron, for instance, while in other cases it was driven mostly by the 5′ region. When they looked at the sequences in the different species, however, they found no evidence that bits of the sequence were jumping between the intron and the 5′ region. "Rather," she said, "the data suggest it's a gradual gain and loss of transcription factor binding sites that is changing which

parts of this gene's cis-regulatory region are driving expression in the same tissue between the same species" (Kalay and Wittkopp, 2010).

There was another surprise when they cut up the regions to try to localize the relevant sequences. Instead they found expression in the same tissue coming from multiple fragments. That is, there was tremendous redundancy in the enhancers. "We also see what I'm considering cryptic or latent enhancer activities, where there's a fragment of DNA from one species that drives expression in a pattern that doesn't exist in that species, but exists in other species," Wittkopp said. The enhancers seem to have "a lot of potential for future evolution already in place" (Kalay et al., 2019).

Consequently, there is still much to learn about the structure–function relationships for enhancer activity and their evolution. In the future, in vivo functional assays for enhancer activity will remain an essential tool. Wittkopp noted that various strategies, such as barcoding and single-cell RNA-sequencing (RNA-seq), could provide advances that would accelerate progress in this area.

Next Wittkopp turned to a discussion of the second approach—genomic strategies—for studying the evolution of gene regulation. With the development of microarrays and, more recently, of RNA-seq, researchers can carry out genome-wide surveys of gene expression. This research has shown that in virtually every system studied, levels of gene expression often vary among individuals and are divergent among species. Some of this variation is expected to contribute to trait differences, she said, but it is likely that much of it does not. This means that a major challenge for understanding the role of regulatory variation in evolution is determining how much of the variation in gene expression is neutral and how much is not.

Wittkopp said that her lab has done some of this work with RNA-seq in flies and yeast, while others have generated similar datasets comparing genome-wide expression between species in mammals. They were surprised to see that the rate and pattern with which expression is changing at a genomic scale are remarkably similar among widely divergent species. "So it suggests to us," she said, "that there may be some general properties about how regulatory systems are evolving."

Her team also saw a second, somewhat mysterious, pattern that seems to hold in divergent species. When one characterizes the expression differences between species in terms of whether they are cis- or trans-regulatory changes, over evolutionary timescales the proportion of the expression differences that are cis-regulatory changes, steadily increases (Wittkopp et al., 2008; McManus et al., 2010; Coolon et al., 2014; Metzger et al., 2017). This has been seen in yeast and in flies, and there are hints of the pattern in available data on mice. Thus, it seems that this sort of proportional accumulation of cis changes is a general pattern in regulatory evolution. The question, then, is why this pattern or other patterns in regulatory evolution exist.

Finally, Wittkopp turned to the third approach, single-mutation strategies, which she is using to gain an insight into how evolutionary selection acts on gene regulation. She began with an observation: "If you are going to make sense of the variation you see within a species or between species and you are going to say something about the role of selection in generating that variation, it works best if you have an understanding of what happens in the absence of selection." This is not a new idea, she said. Indeed, it underlies a classic paradigm concerning the selection of genes: if one understands how genes change due to neutral processes alone—that is, mutations and genetic drift—then that can be compared with the changes seen in natural populations, which also have a component due to natural selection, and one can infer the impact of selection.

A similar process is used to understand the role of natural selection in the regulation of gene expression. However, unlike the case with DNA sequences, for which information on how they change over time in the absence of selection has been collected for decades, models of neutral regulatory evolution have been based on many assumptions, but little empirical data.

This is beginning to change, she said. "We're starting to get ways to actually survey the mutational space" and generate the empirical data necessary to gain an insight into the role of selection in gene regulation.

Wittkopp explained the basic idea in this way: "If we have the mutational distribution and mutational effects for a quantitative trait and we can compare those to the effects of variants we see in natural populations, which have been subject to the same mutational processes but also selection, then we can separate the roles of neutral and non-neutral processes." If polymorphisms turn out to be a random subset of the mutational distribution, there is no need to invoke selection—the variation that is seen may be explained by the mutation process alone. If there is a difference, one can assume that this difference is due to selection.

However, Wittkopp continued, getting the empirical data needed to understand what the mutational distribution looks like has been challenging for quantitative traits. Fortunately, she said, there are now new tools for introducing mutations and surveying their effects on a large scale that have made it possible to greatly accelerate the process over the past 3 years.

Wittkopp described work that her lab has done to learn about the effects of mutations in gene promoters in yeast. Working with the TDH3 promoter, they systematically changed 236 of the 241 Gs and Cs in that promoter to As and Ts, respectively, and then measured the effect of each mutation on the gene's expression (Metzger et al., 2015). "This allowed us to describe the distribution of mutational effects for this promoter," she said. Next, they examined the natural variation in TDH3 promoters in strains of yeast that had been isolated from a variety of locations. They introduced the natural variant sequences into the same study system in order to have an apples-to-apples comparison, and then compared the effects of the mutation and the polymorphisms to infer whether selection had occurred.

What they found was that the polymorphisms looked like a random sample of the mutations. This was surprising, Wittkopp said. Even though this is an important gene in yeast—it is one of the most highly expressed genes and is involved in metabolism—variations in the promoter that affect the expression level of the gene do not seem to be acted on by selection. What they found on further inspection was that few of the mutations changed the gene expression very much—and not enough to have a measurable impact on the distribution of effects. "That's why we get the signature of neutral evolution," she said.

They uncovered another layer of complexity as well. Mutations do not just change the average expression of a gene, Wittkopp said, but also change the consistency in that gene's expression among genetically identical cells in the same environment. This is a property known as expression noise, she said. As an example, she said that her skin cells are all genetically identical and live in about the same environment, but if she used single-cell techniques, she would find slight differences in expression among these cells. This is expression noise, and it is genetically controlled.

In her study of the effects of mutations on gene expression in yeast, Wittkopp found that mutations tended to increase expression noise, while the polymorphisms they found in the natural population did not, which indicated that selection had acted to maintain a particular degree of noise (Metzger et al., 2015). That raised the question of why expression noise matters. Why might there be selection for or against noisier genotypes? And how big are the fitness

effects of variation in expression noise? It is challenging to answer that question, because most mutations change both the mean expression level and expression noise.

To look into the role of expression noise more deeply, Wittkopp showed the results of a fitness analysis performed on two genotypes that produce the same mean expression but differ in the noisiness of that expression among cells. The "noisier" genotype had a broader range of expression, with more cells with higher or lower levels of expression.

On average, when the mean expression is close to the optimum, the noisier genotype generates more cells with suboptimal fitness—because they are spread out further away from the optimum—and thus the average fitness of the noisier genotype is lower. However, when the analysis is done with the mean expression far from the optimum, the noisier genotype generates more cells with expression closer to the fitness optimum, which results in the noisier genotype having greater fitness than the less noisy genotype (Duveau et al., 2018).

Why might the mean expression be far from the optimum? One possibility is that the allele was fixed by genetic drift. Another is that there was a change in environment that altered the fitness curve.

For her purposes, Wittkopp said, different environments for yeast are those with different carbon sources, and her group has found that the relationship between the expression of the TDH3 promoter and fitness differs across environments. For example, if a cell that expresses the optimal level of the gene in a glucose environment is put into a galactose environment and expresses the gene at the same level, that expression will be suboptimal.

If a cell regularly experiences both of two environments, Wittkopp said, there is a better solution than noise, and that is phenotypic plasticity. Indeed, this phenotypic plasticity can be seen in the wild-type allele in that when a cell is moved from glucose to galactose, the expression of the gene goes down. In this case, she said, plasticity seems to be adaptive, although in the other environments they have tested it does not (Duveau et al., 2017). "So, I don't think we should assume that all the plasticity we see is adaptive," she said, "but I think it's an open question about how much of it is or isn't."

CONCLUSIONS AND NEXT STEPS

Summing up her results, Wittkopp said that her work has shown that the relationship between genotype and phenotype for gene expression is not just about the average expression level, but also involves the plasticity in that expression level, the noisiness of that expression level, and various other pieces, all of which shape the variation that exists within a species.

In closing, Wittkopp reiterated the technical challenges (gene sizes, hosts for transgenics, landing sites for transgenes, generating recombinants, scaling mapping, and allele swaps) and the biological challenges (many loci, small effects, linked loci, rare variants, epistasis, plasticity, and genome structure, in particular, inversions) that face those who seek to study the genetic basis of trait differences.

Looking forward, she said that a number of technical advances promise to increase and speed up capabilities in this area, and she mentioned specific advances in genotyping and phenotyping, cloning tools, reciprocal hemizygosity testing, and CRISPR/Cas-based tools. "The ability to perturb with targeted mutagenesis on a large scale is also coming on board," she added. "I'm optimistic about the future of these directions."

7

Interpreting and Validating Results from High-Throughput Screening Approaches

Today's sequencing and "-omics" technologies are so powerful that scientists can collect huge amounts of data in relatively little time. This is one of the reasons for the tremendous promise of functional genomics. However, to have confidence in these data, it is important to be able to validate the results, which can be challenging.

The speakers in this session spoke of their experiences in validating different types of high-throughput screening (HTS) approaches, but one commonality emerged: validation generally requires carefully thought-out hard work. The session was moderated by Trudy Mackay of Clemson University. The presenters were Emma Farley of the University of California, San Diego; Philip Benfey of Duke University; Grace Anderson of Octant; and Gary Churchill of The Jackson Laboratory. A panel discussion period followed the presentations.

LESSONS ON DESIGN AND VALIDATION FROM A CRISPR LOSS-OF-FUNCTION SCREEN ON *KRAS*-MUTANT CANCERS

Hindsight, when used correctly, can offer some valuable lessons for the future, said Grace Anderson of Octant as she described her own experience on a project from her first year of graduate school which aimed to study acquired resistance in *KRAS*-mutant cancers. They mentioned that "blind spots" of the past can help identify "blind spots" in current technological approaches.

Cancer, Anderson explained, is an umbrella term used to describe more than 200 unique diseases that are highly diverse. In fact, the only thing cancers have in common is the hallmark of uncontrolled cell growth. Over the past 15 years or so, researchers have come to appreciate the diversity of mutations that occur in cancer, and they have identified commonly mutated driver oncogenes that are similar in many of these diverse cancer types. Researchers have done a good job in identifying targets and developing inhibitor molecules to block the protein products of these commonly mutated oncogenes. However, these molecular targeted therapies have a huge problem. Virtually every one of them faces intrinsic and acquired resistance from the cancer cells it is attacking.

The best example of an acquired resistance, Anderson said, is what happens when BRAF-mutant melanoma is treated with vemurafenib. *BRAF* is an activating mutation in about 50 percent of all melanomas. Testing has shown that not all patients with a *BRAF* mutation will respond in the same way to vemurafenib. Some exhibit little to no response; this is referred to as having intrinsic resistance to the drug. Others, even those with metastatic melanoma lesions all over their body, seem to respond very well at first, with the lesions completely disappearing. However, every patient that exhibits this complete response eventually relapses, with the drug no longer working; this is acquired resistance (Chapman et al., 2011; Wagle et al., 2011; Sosman et al., 2012).

As a first-year graduate student, Anderson set out with a colleague, Peter Winter, to study this phenomenon in cancers driven by mutations in another gene, *KRAS*. It is challenging to target *KRAS* directly due to the exceptionally high binding affinity *KRAS* has to its endogenous ligand, GTP. With this in mind, Anderson said, they chose to look into downstream methods of targeting the associated pathway. Looking back, Anderson recognizes that they would design the study differently, knowing what they know now. These lessons come in the form of changing library design and validation.

Anderson was interested in *KRAS* mutations in cancer for a couple of reasons. The most obvious was that *KRAS* is mutated in about one-fifth of all cancers, and nearly ubiquitously in some, such as pancreatic cancer (Cox et al., 2014). Furthermore, *KRAS* mutations lead to constitutive signaling through its associated mitogen-activated protein kinase (MAPK) pathway, which is involved in cell cycle progression, among other functions. However, monotherapies blocking downstream signaling proteins, such as MEK or ERK, seem to be insufficient to stop the cancer, which is likely due to the complexity of the pathway as well as its built-in compensatory mechanisms. To overcome this, it would be necessary to look toward combination treatment strategies.

There were some open questions that they thought a CRISPR loss-of-function screening approach would help to answer, Anderson said: What are the pathways that, when inhibited, can sensitize a patient to ERK inhibition in *KRAS*-driven disease? How do these pathways vary with respect to the specific tissue in which the *KRAS* mutation is present? And, what accounts for the diversity of responses to seemingly appropriate therapies?

To start, they selected about 400 interesting cancer-related genes, which, Anderson noted, was a heavily biased list. They screened 12 cell lines and 4 tissue types, for a total of about 94 screens. The CRISPR loss-of-function screens used drug treatment to inhibit either MEK or ERK. Genes that "dropped out" in drug-treated samples were candidate "sensitizers" to that drug (Anderson et al., 2017).

They first noted that very few of the hits spanned more than two tissue types. In fact, only one hit spanned all of them, which was a known pathway reactivation gene, *REF*. Thus, Anderson noted, most of the boundaries of sensitizers are at the tissue level, an important finding for clinical trial design because most clinical trials were being defined in terms of genotype. This type of testing was, "setting the trial up for failure," Anderson said, if the drug was not going to work in several tissue types.

More generally, Anderson said, even though they were able to answer some questions, there was not a lot of novel basic science that came out of the screens. Reflecting on this, Anderson noted that they may have made more discoveries by not biasing their library so much.

Moving to the issue of validation, Anderson said that they validated 44 of their 46 hits with two separate assays. The two gene hits that did not validate were damaged genes, with the damage being an artifact of the CRISPR cutting. There were a variety of problems with the validation, however. For one, they did not know the true effect sizes of their phenotypes because they did not include any true lethal genes, such as ribosomal genes, a standard practice today. Perhaps the biggest issue was that, because all of the hits worked well in both validation assays, they had no way of prioritizing which of the 44 hits to test in vivo.

Later they developed a long-term in vitro assay that allowed them to differentiate among the hits. "What we found," Anderson said, "was for these 22 combinations in colorectal cancer, all of which looked amazing on short-term validation, there was crazy diversity in how well they

could actually suppress resistance long term." This is now the standard for how researchers validate their lethality screens in the targeted therapy space.

In conclusion, Anderson said, heavily biasing the library was probably a pretty big mistake. In designing a library, one has to find the proper balance between genome-scale and a small-scale targeted library. In retrospect, Anderson said, they should have come up with some way of scaling down from the entire genome without handpicking genes that are important in cancer. Last, when validating, it is important to choose the proper assays for the phenotype. In their case, they were interested in long-term durability, and so any shorter assays were not appropriate.

USING FUNCTIONAL GENOMICS TO UNDERSTAND DEVELOPMENT

In the next talk Philip Benfey described applying functional genomics tools to study development in a plant—in particular, the development of the *Arabidopsis* root. "When I say development, our primary focus is on how you go from an undifferentiated stem cell to a fully differentiated cell and how that cell can function to make an organ that has some real purpose in the world. And the organ we focus on is the root."

Studying *Arabidopsis* roots simplifies that problem of development because of the way the root grows in one direction along one dimension and because it has radial symmetry—that is, the different cell types essentially grow in concentric tubes, one inside the other, surrounding an inner cylinder of vascular tissue. Thus, the position of a cell can be specified by only two variables: how far from the tip it is and how far from the centerline it is (see Figure 7-1). The stem cells, which are the source of all the other cells, are found at the tip of the root, and because the cells do not move in relation to one another during development, the youngest cells will always be found at the root tip and cell age increases with distance from the tip, meaning that one sees a developmental time line as you move up the root. Cell development resembles an assembly line where one starts with fairly undifferentiated materials and ends with specialized materials. These two variables of the *Arabidopsis* root, distance from centerline and distance from the tip, are sufficient to specify cell type and development stage.

About 20 years ago, Benfey said, he and his team studied expression patterns in the different cell types at different times in development. He used 19 different markers of cell type and sliced individual roots into 12 sections, each a certain distance from the tip and thus at a particular developmental stage. When he examined the data from each of the 19 markers at the 12 developmental stages, he found clear expression patterns related to both cell type and developmental stage (Brady et al., 2007; Dinneny et al., 2008). What was surprising at the time—although it would not be surprising today, he said—was how much cell-specific expression they found.

"We went on to show that that cell-type-specific expression was also responsive to different environmental stresses," he said, explaining that this was an unexpected result. It seemed unusual that there would be cell-type specific responses to the same stressor.

There were, however, a couple of weaknesses with this approach, Benfey said. First, it required the marker lines, and developing them for a new species would have required a lot of work. Second, it was difficult to monitor responses over time, again because of the amount of work involved. "As a high-throughput approach, it left a lot to be desired."

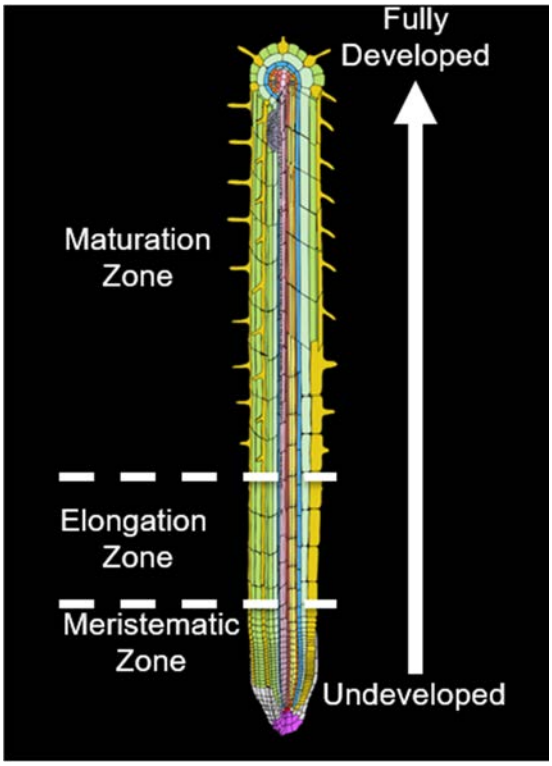

FIGURE 7-1 Growth in the *Arabidopsis* root showing that the tip of the root has the youngest cells, while those cells farther away from the tip are the most fully developed.
SOURCE: Philip Benfey presentation, slide 2.

A newer alternative approach is to use single-cell RNA-sequencing (RNA-seq), which generates essentially the same data but from individual cells rather than from cell collections all of a single type. In the past year, Benfey said, there have been five different publications of single cell RNA-seq studies done on the *Arabidopsis* root, all of which relied heavily on the annotations that his group had previously developed (Denyer et al., 2019; Jean-Baptiste et al., 2019; Ryu et al., 2019; Shulse et al., 2019; Zhang et al., 2019). However, these studies were missing a lot of the fine detail about changing expression patterns as the root developed. His team decided to approach this by combining some of the published datasets with their own datasets, ultimately using data from 80,000 cells to create a "reference transcriptome" that included information on transcription separated both by cell type and age. Ultimately, the goal with this reference transcriptome would be to use it in a way that is similar to how a reference genome is used.

At that point, Benfey turned to a separate, if related, topic—his team's efforts to understand what is happening in the initial divisions of a stem cell in the *Arabidopsis* root. To do that, he said, they take a reductionist approach and examine what happens in a single stem cell (see Figure 7-2).

The cell first divides in the direction of growth to regenerate itself. Then the upper of the two resulting cells—the one farther from the root tip—divides along the longitudinal axis to produce the first two cells of what will become two very different lineages. "This is an asymmetric cell division in the sense that the two cells will have different fates," Benfey said.

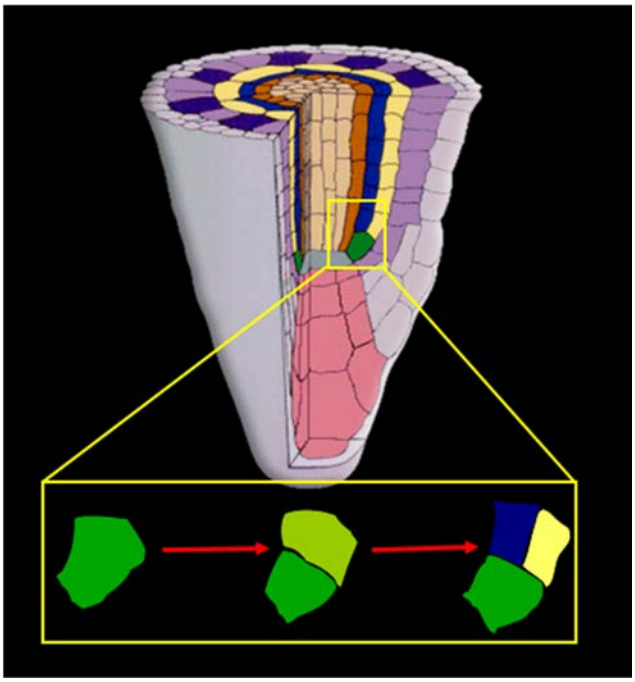

FIGURE 7-2 Display of asymmetric cell divisions in the *Arabidopsis* root, where the mature cell divides along the longitudinal axis into two cells with different fates.
SOURCE: Philip Benfey presentation, slide 19.

In particular, in wild-type (WT) *Arabidopsis*, the result of that asymmetric division is the creation of two inner layers, the cortex and the endodermis, which are situated between the outer layer (the epidermis) and the inner cylinder of vascular tissue. But in a *shortroot* (SHR) mutant variety of *Arabidopsis*, named "shortroot," there is only a single layer between the epidermis and the vascular tissue that only has attributes of the cortex, the outer of the two asymmetric layers (Benfey et al., 1993).

A second variety called "scarecrow" also has just a single layer there, but that single layer has attributes of both the endodermis and the cortex (Benfery et al., 1993). In other words, it never makes that initial asymmetric division, but instead the resulting cells have characteristics of both cell types.

After cloning the genes responsible for shortroot and scarecrow, Benfey's team was surprised to find that shortroot was not expressed in either of the two lineages from the asymmetric cell division. The SHR protein, a transcription factor, is made internally and then moves out to the adjacent cell layer (Nakajima et al., 2001). As discussed in Bergmann's talk (see Chapter 3), plants can signal by moving proteins from one cell to another. As it moves, SHR comes into contact and interacts with the SCARECROW protein, which is another member of this plant-specific transcription factor family, and SCARECROW prevents SHR from continuing to move to the outer layers (Di Laurenzio et al., 1996; Helariutta et al., 2000; Nakajima et al., 2001; Cui et al., 2007).

When SHR interacts with SCARECROW, it binds to the SCARECROW promoter, producing a positive feedback loop (Cruz-Ramírez et al., 2012). The protein complex also binds to the promoters of a highly specialized member of the cell cycle machinery that is critical to the asymmetric cell division. To understand the dynamics of the various interactions, Benfey's

group, in collaboration with the Scheres group at Wageningen University, created a mathematical model that predicted a bi-stable behavior in which SCARECROW is at either a low or high level. "The naïve prediction from that," he said, "is that if it's a low level, the switch is off, and if it's a high level, the switch is on. And when the switch is on, you get the cell division."

Next they studied the dynamics of the process by studying the behavior of SHR, SCARECROW, and the subsequent asymmetric cell division. To do this, they used light sheet microscopy, which allowed them to examine the levels and kinetics of both SHR and SCARECROW. They could perform live imaging to watch the levels of both proteins in the different cells of the growing roots and even see the asymmetric cell divisions. The behavior they observed did not completely match up with the prediction of bi-stability from the mathematical model, Benfey said, but perhaps the more important fact is that they were actually able to test the predictions of their mathematical model by watching the expression levels in real time.

Finally, Benfey mentioned two other techniques that his team is working with. One is using single-cell RNA-seq to follow spatial and temporal expression in the root. The second is building synthetic circuits with transcription factors and other pieces, embedding them in a larger network, and then modifying the circuits to see what effect it has on that larger network.

In conclusion, he spoke briefly about the challenges his team faced. Time courses are critical, not only for what happens after the induction of shortroot and other key regulators, but also for what happens in response to environmental perturbations. The challenge, Benfey said, is how to analyze those time courses. It is not obvious how one can use standard time-course algorithms to study something where there are 10,000 cells, each with sparse data and a different time course. Synthetic networks offer a way of validating one's models in the deepest way, Benfey said. As engineers will say, "You don't really understand it until you can make it yourself." Finally, he said, integrating functional genomic approaches with time-lapse imaging is also a great challenge, but one with a significant potential payoff.

VALIDATING RESULTS FROM HIGH-THROUGHPUT ENHANCER SCREENS

Enhancers are short stretches of DNA that, when bound to by proteins called transcription factors, control the temporal and spatial expression of a gene. These enhancers provide instructions for tissue-specific gene expression. Emma Farley of the University of California, San Diego, spoke about issues related to the HTS of enhancers.

"I think of [enhancers] as switches in the genome that control when and where genes are expressed," she said. The major shortcoming of screens for enhancers, is the lack of functional validation. She mentioned that many studies use correlative methods to predict putative enhancers in the genome, or test how sequence changes in enhancers affect phenotype. These studies often identify tens of thousands of putative enhancers and yet only a handful are validated functionally with experimental assays. This is typical and points to a weakness in the current use of HTS—the lack of true functional validation of the large datasets. Such validation is imperative to help researchers learn how much of what they are inferring is actually true.

To provide further insight, Farley described some of her own work and the issues that arise concerning validating results from HTS. Referring to Figure 7-3, she offered a simple example of the sort of approach she takes to understand how the sequence of an enhancer encodes tissue-specific gene expression and what changes in an enhancer affect gene expression and phenotype.

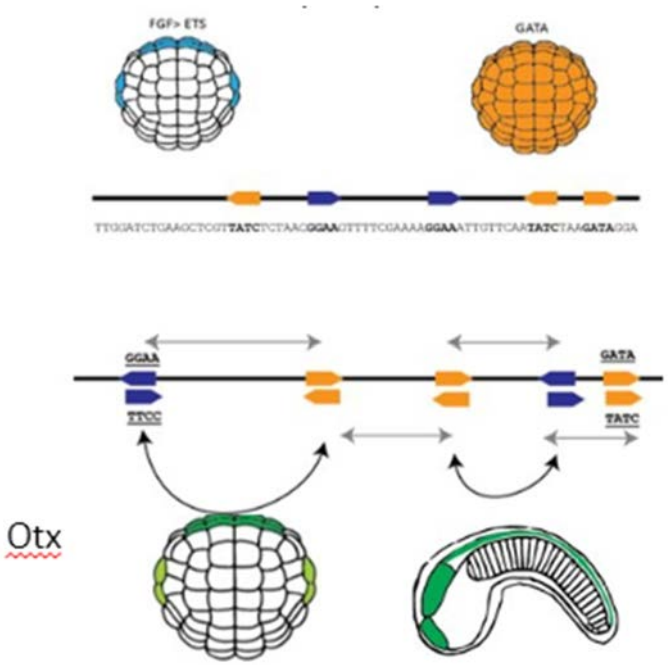

FIGURE 7-3 Display of how the sequences of the enhancers ETS and GATA encode for tissue-specific expression.
SOURCE: Emma Farley presentation, slide 3.

The simple enhancer illustrated in the figure contains two types of binding sites, one shown in orange and one in blue. As Farley explained, the ball-shaped object is a gastrula-stage embryo, and the cells on the visible side of that embryo could give rise to either skin or nervous system. The embryo in the top left shows the signaling of fibroblast growth factor (FGF), which activates the transcription factor ETS, and the embryo on the top right shows the cells on the visible side of the embryo expressing the transcription factor GATA. The binding of ETS to the blue sites in the enhancer and of GATA to the orange sites in the enhancer act to turn on the gene *otx*, which is one of the first genes to specify nervous tissue in an embryo. The bottom right of the figure shows a later-stage embryo, where the head, tail, and nervous system can be seen.

A major issue in functional genomics is understanding which regions of the genome encode enhancers, where and when these enhancers are active, what genes they control, and which SNPs impact enhancer function. To illustrate this problem, Farley said there are more than 100,000 clusters of ETS and GATA binding sites in the genome, some are non-functional, some act as enhancers that turn on in the nervous system, and some turn on in the gut and the heart. Farley went on to ask why this particular (Otx-a enhancer) sequence is turning on in the nervous system?

In an ideal world, one would take the sequence and then change various nucleotides and see how that affects gene expression, while keeping constant those nucleotides known to be important in order to make it more likely that the resulting variants are actually functional. However, the complexity of the required analysis increases rapidly with the number of nucleotides, which, Farley said, is one of the reasons she chose to work with this enhancer—because it is very small. But even in a small enhancer, making every sequence combination would still produce 10^{30} different synthetic enhancer variants, an amount that is currently impossible to test and analyze. The other reason for choosing this enhancer is that it has a

common logic, as the enhancer is activated by a signaling pathway (FGF>ETS; see Figure 7-3) in combination with a tissue-specific determinant (GATA). Farley noted that she did not just want to understand how this enhancer works but how enhancers regulated by ETS and GATA, and indeed how enhancers regulated by this common logic of signaling factor and tissue determinant, generally work. She mentioned that her lab not only looks at how sequence encodes gene expression, but also how the role of organization of bindings sites (the order, orientation, and spacing of these sites) encode gene expression.

She works with embryos of the sea squirt *Ciona intestinalis*. In doing her experiments she takes an enhancer, attaches it to a promoter and a green fluorescent protein (GFP) reporter, and electroporates the construct into fertilized eggs. Then, wherever the enhancer turns on in the resulting embryo, it can be detected using GFP.

"We can electroporate hundreds of thousands or millions of fertilized embryos in a single hour-long experiment," Farley said, "and the idea is to make all of these different variants and then test them in whole embryos to see how changes in sequence . . . impact gene expression." The core of the binding site is kept constant, and the rest of the sequence is randomized. Each enhancer sequence is attached to a promoter, GFP, and a unique barcode. If an enhancer turns on transcription, it will make a messenger RNA (mRNA) of the barcode, which can be detected, measured, and associated with the appropriate enhancer sequence.

After testing millions of enhancer sequences, Farley identified 20,000 that were active at or above WT levels of *Otx* gene expression. To partially validate the data, she re-tested 100 of the enhancers, some of which were active and some inactive according to their data, using the same general assay, but looking for GFP expression under a microscope instead of at pooled sequencing data. Those results lined up with her original findings, and this gave her faith in the sequencing data.

Next, she examined her set of 20,000 functional enhancers for common features and found two motifs that were high-affinity sites for ETS and GATA. Although the high-throughput data suggested that high-affinity sites were enriched in the functional enhancers from the HTS, when she measured the affinity of the sites within the WT Otx enhancer, she found that there were some high- but also some very-low-affinity sites. To test the hypothesis that only the high-affinity sites were needed, she took enhancer variants from the library that were not functional and sought to make them functional.

"I think that it's key when you're doing these high-throughput screens," she said, "that you test your hypotheses in the most challenging way you can think of. We thought trying to make inert DNA into a functional enhancer was a real challenge because it's incredibly hard to build tissue-specific enhancers."

When they tried this by just mutating the sequence flanking the core of the binding site to make high-affinity binding sites, they found that they could indeed turn an inert piece of DNA into a functional enhancer, but they had lost tissue specificity. With further experimentation, she said, they discovered that the lower-affinity sites seen in the WT enhancer were needed for tissue specificity.

Exploring further, the next step was to examine how the organization of the binding sites affects gene expression. They tried manipulating the spacing between the binding sites and found that changing the spacing between the binding sites affected the levels of gene expression. Interestingly, the spacing seen in the WT enhancer is not optimized to give the highest level of expression. Because the WT enhancer has both binding sites that are of suboptimal affinity and a spacing of the binding sites that is suboptimal for transcriptional output, she tested what would

happen if both the affinity and the organization of the binding sites were optimized. The answer was that tissue specificity completely disappeared with optimization (Farley et al., 2015).

To test whether this was a general principle, Farley was successful with work on similar enhancers and other tissues. Such validation, Farley said, is important if one wants to "translate principles and make claims about the rules of life."

After that, she said, she ran into a roadblock working with the sea squirt because it is difficult in that species to replace one enhancer with another, so she switched to mice. It was a blessing in disguise, she said, because to test if the principles she discovered are truly generalizable, it was important to test them in a completely different organism.

She worked with the well-known enhancer ZRS, which drives expression of the *sonic hedgehog* gene in developing limb buds. As she had seen with the sea squirt, the ETS binding sites on the enhancer for sonic hedgehog were low affinity. As it happens, she said, there is a human mutation that causes polydactyly, and no one had understood why that mutation has this particular impact, but she noticed that the mutation tripled the affinity of a binding site for ETS (Albuisson et al., 2011). Her work in the sea squirts led to the hypothesis that this mutation could lead to a less precise expression of the *sonic hedgehog* gene and an extra digit. They tested this in mice and found that optimizing the affinity of the binding site did indeed produce extra digits, demonstrating that this principle of enhancers translates across species.

Wrapping up, Farley said that the key to using HTS is reducing complexity as much as possible. "When we first started," she said, "all we did was change the sequence outside the core of the binding sites, and, by doing that, we were able to understand what was necessary and sufficient to drive expression." This led to work in which they explored changing order, orientation and spacing of the binding sites, to investigate further principles behind enhancer sequences.

Farley noted that the key aspects of HTSs are

- Experimental design:
  - Biological context
  - Experimental tractability
  - Reducing variability and confounding factors
- Validating at multiple levels:
  - Validating the primary data
  - Rigorous testing of any hypotheses obtained from the HTS analysis
  - Making and testing predictions to further challenge your hypothesis

Validating one's findings in many contexts shows the generalizability of any principles identified.

The major roadblocks facing people in this field, she said, include finding a way to reduce the complexity of a problem, the fact that validation of high-throughput screens is very slow, and the fact that no organism provides all the tools needed to answer these questions, so one must decide whether to develop the necessary tools or change organisms.

Looking to the future, Farley said, "I think we need a culture shift. I think functional genomics used to be about correlation, but now I think it should be about causality at scale." The field needs to develop a framework for validation that relies heavily on testing of predictions, she said. Funding initiatives are needed that focus on innovations to validate large-scale screens more efficiently, whether it is through automated imaging, the integration of other -omics, transgenic tools, or phenotyping at scale.

HARNESSING GENETIC DIVERSITY TO UNDERSTAND MAINTENANCE OF PLURIPOTENCY IN EMBRYONIC STEM CELLS

The session's final speaker, Gary Churchill of The Jackson Laboratory, began his presentation by announcing that he was going to lead off with his conclusion. Touching on some of the themes from the previous talks, he said that if functional genomics is to make the transition from a correlational approach to a causal approach, genetic variation holds great promise as an experimental perturbation. Under certain circumstances one can assume that genotypes precede phenotypes. However, he continued, this is an assertion, and it requires some really strong assumptions that can be somewhat alleviated by the construction of artificially structured genetic populations. Mediation analysis is often used to explore the mechanisms of how independent variables, such as a genotypes, influence dependent variables, such as phenotypes, but there are some problems with this approach that cannot be entirely eliminated. Furthermore, functional genomics researchers are faced with the challenging situation of having an overwhelming number of results coming from these analyses but a limited bandwidth with which to do validation.

"So my question is," Churchill said, "when is validation needed and when are correlational results sufficient to just move on with our lives?"

With that, he began his presentation by describing the mouse populations that are bred at Jackson Lab. There are eight inbred founder strains that are genetically very different from one another, and there are both inbred and outbred populations. The diversity outbreds have a balanced population structure, with more than 400 recombinations per animal, and a high heterogeneity, with the disadvantage that each mouse is unique. The inbreds, or collaborative crosses, have reproducible genomes and a high genetic diversity, but fewer recombinations per line.

Next, Churchill introduced expressional quantitative trait locus (eQTL) and protein quantitative trait locus (pQTL) mapping. These are locations in the genome that are linked to variation in either gene expression levels, that is, mRNA levels (eQTLs) or protein levels (pQTLs). Churchill's group carried out a study comparing eQTLs and pQTLs in liver tissue in mice, and the overlap between the two "was so close to random that it was shocking." But a closer look revealed that the overlap between the local eQTLs and pQTLs—that is, those near the corresponding genes—was almost perfect, while the overlap between distal eQTLs and pQTLs was almost entirely absent. "That tells us a lot about how proteins and RNAs are regulated," he said (Chick et al., 2016).

With those data the team applied mediation analysis in an effort to reveal some of the details of the relationship between the RNAs and the proteins. Such an analysis requires starting with a hypothesis, so the team hypothesized that variation in the DNA causes variation in the RNA, which in turn causes variation in the protein. They use a standard strategy called the causal steps method, which involves four logical statements (see Figure 7-4). "We asked that the protein have a QTL, that the RNA have a QTL, and that these both co-localize"—(i) and (ii) in the figure. "We asked that if we calculate the partial correlation by regressing out the protein, the RNA should still have a QTL" (iii). "And we asked that if we regress out the RNA, the protein no longer has the QTL" (iv).

The fourth statement is the stumbling block, Churchill said. "That's the one that we can't prove. And it's the one that causes us all the headaches and all the focus."

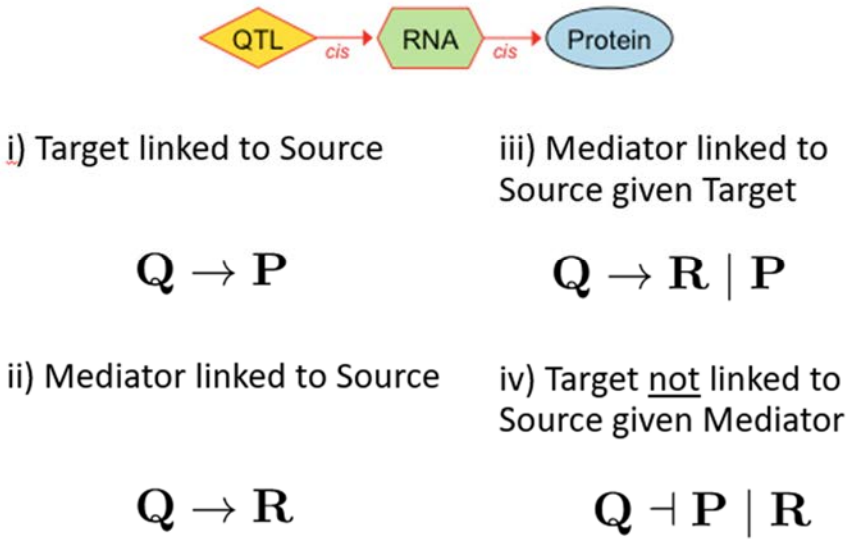

FIGURE 7-4 Mediation analysis, called the "steps method," which contains four logical statements. SOURCE: Gary Churchill presentation, slide 4.

With that background, Churchill moved to the main subject of his talk—how he and his team used this approach to understand the maintenance of ground-state pluripotency in mouse embryonic stem cell lines. Ground-state pluripotency is the ability of cells to make unlimited copies of themselves and also, under the right conditions, to differentiate into any sort of cell in the body, which is, in a sense, the most primitive and unlimited form of pluripotency. To study how this pluripotency is maintained, Churchill's group derived a large set of embryonic stem cells from individual DO mice and performed RNA-seq and assay for transposase-accessible chromatin using sequencing (ATAC-seq) to get data on RNA levels and on chromatin accessibility across the genome, respectively.

When they mapped out their data, they found characteristic patterns where there were strong signals of local regulation with a significant amount of distal variation. And that distal variation sometimes seemed to cluster in bands where a single genetic locus appeared to affect the expression of many genes or a single genetic locus seemed to affect the openness of many ATAC peaks in the genome, corresponding to regions in the genome where the chromatin was accessible (see Figure 7-5).

Interpreting this according to their prior hypotheses, the data indicated that there was a genetic variant that affected the chromatin in multiple places and that, in turn, affected the expression of multiple genes. They found that there were many more ATAC peaks than there were genes, Churchill said, but "for virtually every gene where we saw variation and expression, there was one or more ATAC peaks which were clear mediators, and they followed the right allele effects and everything was cool."

The more challenging part of the study was to understand what was going on with the "hotspots," the collections of distal regulatory elements that appeared as bands in the maps. One hypothesis was that there is a genetic variant that by some mechanism affects a distal ATAC-seq defined open chromtin region, which then regulates the expression of the gene.

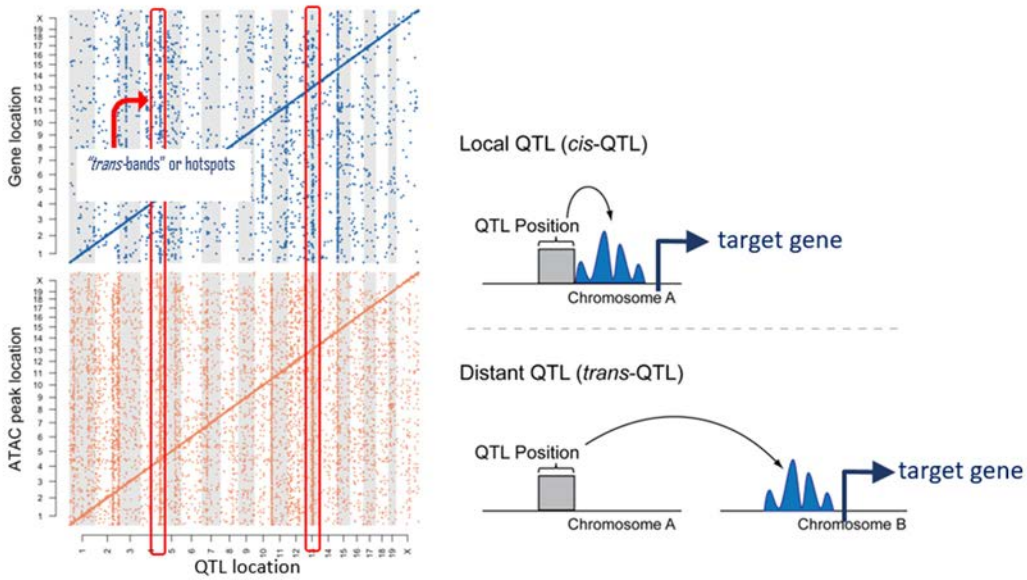

FIGURE 7-5 Data of distal variation clustering in bands (*left*) and representation of local QTL effects versus distal QTL effects (*right*).
SOURCE: Gary Churchill presentation, slide 13.

There were many hotspots throughout the genome that Churchill identified in this work, but for the purposes of the presentation he focused on a hotspot on chromosome 15. His team found a particular gene in the region that seemed to mediate a lot of the downstream target genes, and that was the gene *lifr* for the leukemia inhibitory factor (LIF) receptor. This gene is involved in regulating the differentiation of embryonic stem cells and pluripotency (Graf et al., 2011).

When Churchill examined the situation more closely, he found greater complexity than was apparent at first. Looking at the genetic pattern at that locus among the eight founder mouse strains at The Jackson Laboratory, he found that of the downstream target genes, many of which are involved in pluripotency, there was high expression in four of the strains and low expression in the other four. Furthermore, the four strains with low expression were "recalcitrant" strains for which the researchers found it difficult to maintain embryonic stem cell lines. By contrast, the four with high expression were "permissive" strains for which the embryonic stem cell lines were easy to maintain. "The downstream targets do the right thing," Churchill said. "The LIF receptor is doing the right thing."

Furthermore, they were able to identify a single-nucleotide polymorphism (SNP) correlated with the difference in *lifr* expression among the four recalcitrant strains and the four permissive strains. It is about 10,000 base pairs upstream of the gene, and so it is likely to be some sort of enhancer.

Finally, the SNP is just seven base pairs upstream of a known binding side for the transcription factor Nr5a2, and the evidence suggests that Nr5a2 regulates *lifr*. "Indeed, in the case of the high-expression allele, Nr5a2 is probably the key transcription factor," he said. "It just binds." He continued, "if you change the context a little bit, all of a sudden the effect of Nr5a2 becomes quantitative."

To validate the role of the SNP, Churchill swapped out the two versions of the SNP between a high-*lifr* strain of mice and a low-*lifr* strain, so that the high-*lifr* strain now had the

SNP from the low-*lifr* strain, and vice versa When he did this, the response to LIF of embryonic stem cells from the two different strains of mice flipped, confirming the key role of the SNP.

Summing up, Churchill said, "we mapped an important regulator of pluripotency maintenance in these embryonic stem cells. We identified a causal gene, and we identified a single causal SNP, and we went to great lengths to validate it."

DISCUSSION

Mackay, the moderator, opened the discussion with her own question. Any functional genomic screen will produce statistically significant results, depending on what has been screened for, she noted, and many of the implicated genes will have unknown functions. Also, most validation strategies are one hit at a time. So, first, how should these results be interpreted in terms of networks, "bearing in mind that most known networks are based on loss-of-function mutations"? Second, how could one approach validating entire networks rather than one node at a time?

Churchill responded that it is important to keep in mind that when one performs such a screen and has thousands of hits, there will be an error rate—something that many people do not recognize. To determine what that error rate is, it is necessary to do some sort of validation experiment, perhaps with a complementary experiment that does not share the same biases and pitfalls as the original experiment.

Next, Farley suggested, "If you have thousands of regions that you think are having an effect on some phenotype, [you] would test them." For example, if a study produces a set of ATAC-seq peaks, test as many as possible, if not all. The difficulty will vary depending on the type of data, she said, but with transcriptional regulation studies, for instance, "you could go about asking those questions if you design your experiments right and you design massively parallel reporter assays in the correct way."

Benfey had a different take on the question. Much of what researchers do is find something interesting and then go deep into it, learning everything they want to learn. And that is a valuable approach. "I'm not that convinced that having totally comprehensive analysis validation is the way we need to go," he said.

Nathan Springer of the University of Minnesota pushed back on a theme that he had heard in a couple of the presentations—that researchers need to move beyond correlation to causation. There is value in that, he acknowledged, but "I want to be careful placing all of the value there." As an example, he pointed to the dairy cattle industry, which has made tremendous progress in genomic prediction without spending much time on mechanisms. These researchers care about predicting a trait, not in exploring the genes underlying that trait.

Farley responded that she did not doubt that there is great value to many of the studies that do not touch on causation, but she believes that "we should learn which assumptions we're making are accurate and which ones are not." She was not arguing that researchers should functionally validate everything they can, but she does believe that researchers should start doing some studies that examine causality because such studies have been underrepresented in the work on how genomes encode phenotypes to date. Understanding causation could end up making correlative studies even more powerful.

Churchill suggested that whether a study should look at correlation or causation depends on its goal. Understanding causation can be particularly important in biomedical research where the ultimate goal is to learn enough to develop effective therapies.

Joanna Kelley of Washington State University, playing devil's advocate, asked why, if validation is so important, shouldn't the field put all its effort into "studying the role of transcription factors, ATAC-seq peaks, all of these things, in one organism, to really understand deeply how all of these different functional genomic things that we're measuring work"? In short, why not focus on one organism in depth rather than researchers each trying to validate in their own systems, which presents so many challenges?

"We were asked to talk about how to validate high-throughput screens," Farley said, "and so I was using my research as an example of how I think we should at the moment validate high-throughput screens to understand the underlying patterns in the information that we're trying to understand, to get to how changes in our genome actually impact phenotype." Everyone is asking different questions, and whatever the question is, it is important to test one's assumptions.

Marc Halfon of the University at Buffalo suggested that how one validates often depends on the question one is trying to answer and what one hopes to accomplish. Also, researchers need to be aware of "validation creep," which Halfon described as "when you forget that the stuff that you haven't validated may still not be right or a certain percentage of that may not be right." If research succumbs to this validation creep and builds on data that have not been validated, it can be very dangerous, he said. On the other hand, digging into the few findings that have been validated can be very valuable. "You can do a lot with that," he said. "The risk is forgetting that you haven't validated all the rest." Or if one wishes to build on the remaining 19,990, it is important to have more extensive validation.

Farley agreed. "I think the issue is just what you said. We do these genomic-scale studies, and then often there are general statements made based on those genome-wide studies" rather than on the 10 things that have been validated. The other issue is how researchers should approach this issue over the long term. "We can't possibly measure every parameter in every cell in every time point," she said, "so what are the ways that we can find the structure or the patterns embedded in DNA sequence?"

8

Large Databases and Consortia

Much of the power of today's functional genomics depends on large databases that can hold the huge amounts of data that are being generated with a rapidly growing suite of technologies. However, there are a variety of challenges associated with developing and integrating such databases. One approach to creating and operating these large databases is through multi-institution consortia, which have their own challenges.

The workshop sessions described in this chapter examined some of the infrastructure challenges facing the field of functional genomics. In particular, the chapter combines two closely related sessions from the second day, one on the challenges and successes of integrating large datasets, and the other on the pros and cons of consortia and large databases.

The moderator of the first session, on large datasets, was Norbert Tavares of the Chan Zuckerberg Initiative. The speakers were Charles Danko of Cornell University, Alexis Battle of Johns Hopkins University, Rahul Satija of the New York Genome Center, Saurabh Sinha of the University of Illinois at Urbana-Champaign, and Genevieve Haliburton of the Chan Zuckerberg Initiative. The second session, which focused more on consortia, had no presenters but instead was made up entirely of a panel discussion.

CHRO-SEQ: A NEW TECHNIQUE FOR INTERPRETING GENOME SEQUENCE

The large datasets that are at the heart of an increasingly large portion of functional genomics are populated by a variety of techniques that produce an array of different types of genomic data. In his presentation, Charles Danko described chromatin run-on and sequencing (ChRO-seq), a new technique that promises to provide an important window into the relationship between genotype and phenotype.

ChRO-seq is a way of measuring the genome-wide location and orientation of RNA polymerase, Danko explained, and he provided an illustration of what the resulting data look like (Chu et al., 2018; see Figure 8-1). The red bars pointing up denote the amount of RNA polymerase at each position in the genome going from left to right, while the blue bars pointing down denote transcription from right to left. These are the raw signal data, and a rich source of information about genome function, he said. "We can read off the location of enhancers and promoters, polyadenylation cleavage sites, and even these gene bodies just from the ChRO-seq data." This is a good single assay for maximizing the information gained about the genome.

One way to think about the value of ChRO-seq, is related to the fact that the amount of RNA polymerase loaded onto a gene corresponds extremely well with the levels of mRNA that are produced from that gene. "This goes along hand in hand with the idea that the majority [of the] variation in gene expression is regulated at the level of transcription," Danko noted (Blumberg et al., 2019).

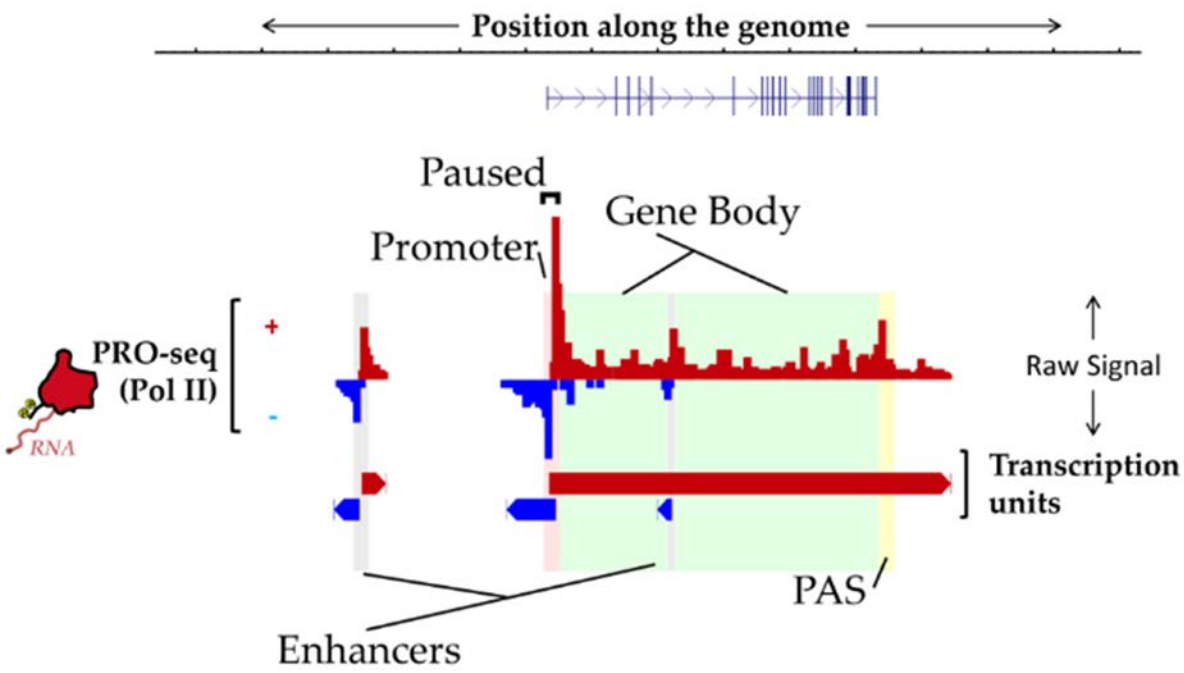

FIGURE 8-1 Representation of ChRO-seq data and interpretation of those data.
SOURCE: Charles Danko presentation, slide 2.

Additionally, the ChRO-seq data have features that correspond strongly with certain features of chromatin modifications. One of the major goals of his lab has been to "deconvolve what these chromatin states might look like."

To accomplish this, the Danko lab took the ChRO-seq data and wrote a program to recognize the patterns of a number of different histone modifications. To this point, Danko said, they have trained models for 11 of the histone modifications, including H3K27 acetylation and H3K4 trimethylation, which are markers for promoters and enhancer regions; H3K36 trimethylation, which is a marker for gene bodies; and two markers for the repressed state. The only one of those that does not work at all is H3K9 trimethylation.

Using data from a stretch of chromosome 21 that was left out during the training of the model, Danko was able to test how well the model would impute distal marks along the chromosome by comparing the imputed values from the model against measured values from the Encyclopedia of DNA Elements (ENCODE) project. He displayed a couple of slides showing that the model closely reproduced both the broad patterns and the fine structure for most of the histone modifications.

One important question, Danko said, is how strong the relationship is between transcription and histone marks. "We've known for 60 years that transcription and histone modifications correlate with each other, but how strong is that correlation?" If there is a lot of unexplained variation, there could be multiple ways to encode a particular functional element in the chromatin state, and there are claims in the literature that this is the case. He asked, "Can we use this tool to evaluate whether those models are plausible or not?"

His team compared their own methods with chromatin immunoprecipitation sequencing (ChIP-seq) data that were downloaded from ENCODE. These comparisons were done on two widely used cell lines, K562 and GM12878. The imputations from their own data showed results similar to biological replicates of the downloaded ChIP-seq data.

However, Danko's team found a number of cases where the imputations failed. In one, the imputation predicted a lot of H3K27 acetylation and there was no evidence for the same amount of acetylation in the ChIP-seq data from ENCODE. One possible explanation for this was the possibility of a great deal of biological variation between the K562 cells analyzed in ENCODE and those analyzed in his lab. Indeed, when they produced their own ChIP-seq data, that is what they found. There was clearly H3K27 acetylation there. This was similarly true for every one of the several hundred cases where their imputations differed substantially from the experimental ChIP-seq data. Because of this, Danko noted that "transcription is nearly identical to the histone modification for the marks they were studying."

Another important question, said Danko, is whether the relationship between transcription and each histone modification varies with cell type. They compared models trained on K562 cells with data from a variety of cell types including HeLa, T-cells, liver cells, mouse embryonic stem cells, and others. In every case except the mouse embryonic stem cells, "the active histone modifications were recovered with the same fidelity in the training cell type as in the other cell types, which indicates that the there's extensive conservation in the relationship between these marks and transcription." In the case of the one exception, Danko said he suspects that there is some fundamental shift in the association between the chromatin mark and transcription so that the relationship differs between embryonic stem cells and a fully differentiated cell.

Next, Danko asked, "Is transcription informative about complex chromatin states consisting of multiple different histones?" The classic example of such a complex state is a bivalent histone mark where H3K27 trimethylation and H3K4 trimethylation are found on the same nucleosome simultaneously (Young et al., 2017). Running their analysis in the region surrounding the bivalent *prox1* gene, they found that they could recover the H3K4 trimethylation signal, although it did not work as well for H3K27 trimethylation. However, the model recognized that *prox1* has "a promoter that should have H3K27 trimethylation on it." So the answer to the question is yes: transcription contains substantial information about even such complex chromatin states.

Danko offered one more question that ChRO-seq can address: Does transcription initiate at all open regions? Danko showed examples explaining that it is not true that any open chromatin region can initiate transcription, but rather it is confined to specific regions.

To sum up, Danko said that the imputation works well and offers opportunities in genome annotation. He mentioned earlier speakers who talked about the need to sample a lot of different tissues in multiple experimental conditions in order to obtain functional elements in a species. "ChRO-seq is one strategy where you can use a single assay to extrapolate that kind of information on chromatin state and get gene expression patterns as well," he said. "We're also using this to better understand what the histone modifications actually do. Now that we have a model, we can interpret it by perturbing the modification or transcription and asking how the system changes?"

He closed with a series of challenges for this work: First, cell types that people think of as static actually vary a lot. This can create problems when researchers do not take the presence of biological variation into account. Second, differences in genome structure between humans and other organisms lead to technical challenges. In *Drosophila*, for example, the genes are more tightly packed. The implication is that machine learning models will have to be trained

separately for these species. Finally, it is not yet clear whether transcription or histone modifications are informative in all species. "It's plausible that we could have types of enhancers that are not marked by anything," he said. "I think we need to understand the basic mechanisms by which gene regulation works in order to extrapolate across species."

USING GENE EXPRESSION TO UNDERSTAND THE GENETICS OF DISEASE

At Johns Hopkins University, Alexis Battle has been turning to gene expression data in her quest to understand the genetic variation that is associated with disease. "The focus of my lab," she said, "has basically been to identify the effects of non-coding and regulatory variation on the cell and ultimately on disease." These effects can be complex and depend on the specific content and specific tissues in which they take place. To deal with that complexity she develops computational models to work with large-scale gene expression data. "We do a lot of work in novel methods development and machine learning and statistical modeling to try to make sense of these data."

Most disease-associated variants in the human genome are non-coding, Battle noted, which makes it difficult to use genome-wide association studies (GWASs) to understand the disease mechanism or to design interventions. Thus, she typically uses large-scale expression quantitative trait locus (eQTL) studies to examine the association between the genotype at a genetic locus and the RNA expression levels for a specific gene.

In theory, she said, eQTLs can help in the interpretation of genetic variation in complex disease. The idea is that if there is a non-coding variant and it is not known what gene it affects, eQTL data can point to a gene, we can design a drug, and everything is solved. Of course, it is more complicated than that. Many factors can alter the effects of a genetic sequence variant, such as environment, sex, and cell type, which can affect such things as transcription factor abundance and epigenetic changes.

To address these issues Battle has become involved in the Genotype-Tissue Expression (GTEx) project, which is a database focused on tissue-specific gene expression data (GTEx Consortium, 2017; Aguet et al., 2019). There have been 948 donors, with whole-genome sequencing done for each. They have also provided up to 54 different types of tissues—not everyone is able to donate every tissue, Battle noted—on which RNA sequencing has been performed. The resulting dataset has been used to look at GWAS variance and to interpret disease-associated variation in terms of what genes and what tissues those variants seem to primarily affect.

Unfortunately, Battle said, data from GTEx and other large studies such as eQTLGen only solve half of the problem because only about half of the GWAS hits co-localize, that is, share a causal signal with eQTLs in some tissue. However, there are still useful insights to be gained. For instance, of the half of the GWAS hits that co-localize with an eQTL, about half of those co-localize with a gene that is not the nearest gene.

For the half of GWAS hits that are not informed by the eQTL studies, Battle asked, "What are we missing?" One thing that is not adequately considered, she said, is that gene expression is a dynamic process. "The effects of genotype on your cells change throughout your lifetime, during development, during disease progression, during aging," she said. "Your genetic variants are acting differently at these different stages." But almost all eQTL datasets, especially the large cohorts, are done at a single point in time, usually in healthy adults. Static healthy adult data may not reflect genetic effects in diverse, disease-relevant contexts.

To examine the issue of dynamic conditions directly, they began with induced pluripotent stem cells (IPSCs) from 19 genetically diverse individuals. These cells were grown in the lab, and then caused to differentiate into cardiomyocytes. They collected RNA-sequencing (RNA-seq) at 16 different points in time to produce a total of about 300 RNA sequencing samples along with genotype data for each cell line. Analysis of those data showed that gene expression is affected by genetic variation in a dynamic way. They observed, for instance, that many eQTLs that were present in the stem cell state disappeared as the cells differentiated to cardiomyocytes, and vice versa. But perhaps most intriguing, she said, are cases where a genetic variant has an effect on gene expression at intermediate stages of differentiation. Even though this was a study with a small sample size on a single cell type, she said, "we're starting to see explanation of some GWAS hits that are not explained by static tissue."

Next, Battle turned to the issue of the effects of rare variants, which, she said, are completely missed by the sorts of studies she had been describing with GWAS hits and eQTLs. Each individual has about 50,000 variants in its genome that appear with a minor allele frequency (MAF) of 0.01 or less, as well as a few thousand variants not seen in any existing study and not found in any datasets (Li et al., 2014, 2017). The effects of most of these rare variants are unknown. What is known, is that about 8 percent of Americans have a rare genetic disorder, and for about half of those, the variant has not been identified. And, Battle added, "I think 8 percent is a wild underestimate." This is because we only sequence the people who have severe conditions and miss the impact of most rare variants, even of large effect, on many phenotypes.

To address this, Battle begins with the assumption that a variant having a functional impact on an individual's health should be disrupting something at a cellular level. So, it should be possible to look for individuals whose expression for a particular gene is far outside the standard population distribution for that gene. She described a method she has developed to find such outliers that is called "splicing outliers." While gene expression is basically a one-dimensional measurement—either up or down—splicing can vary in a multiple-dimensional environmental space across all the different combinations of splice junctions.

Working from the GTEx dataset, she selected whole-genome sequencing and RNA-seq data across multiple tissues for 714 individuals of European ancestry. The unfortunate current limitation to people with European ancestry, she explained, was because the patterns of rare variation differ across different populations. Then she carried out an analysis of which sorts of variants are near genes for which someone is an extreme expresser or near genes for which someone is an extreme under-expresser, and they found that certain types of variants, such as duplication and some interesting splicing variants, are associated with over-expression, whereas others, such as deletions and frame shifts, are associated with under-expression.

Finally, Battle briefly described a machine learning method her lab developed to predict which variants are likely to have a functional effect on genes in particular individuals. This method, which they call "Watershed," is trained in a completely unsupervised manner, she said. "It really vastly improves performance over using whole-genome sequencing alone," she said. By adding RNA from personal transcriptomic data, it is possible to take any state-of-the-art method and make it "better at identifying variants that actually look functional."

Watershed can be used to inform disease association for rare variants in large studies. The bottom line, she concluded, is that "gene expression is very helpful for interpreting disease variants."

AN ATLAS OF ATLASES

To begin his presentation, Rahul Satija noted that in recent years the quantity and types of data available to researchers in functional genomics have grown dramatically. Technologies for single-cell RNA-seq have been massively scaled, for example, going from tens or hundreds of cells in a single experiment to hundreds of thousands of cells in a single day. And a variety of profiling technologies have joined RNA-seq, such as single-cell assay for transposase-accessible chromatin using sequencing (ATAC-seq) or cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq), which can measure both RNA and protein simultaneously in single cells. There is also spatially-resolved transcript amplicon readout mapping (STARmap), which can make spatially resolved transcriptomic measurements in tissue. Each of these technologies offers a unique perspective on cellular identity (Wang et al., 2018). "If you want to know which transcription factors are particular drivers of cellular maintenance or cell state, you may want to look at the chromatin. If you want to identify surface markers for enrichment, you might want to use CITE-seq (Stoeckius et al., 2017). And if you want to understand how a cell's neighbors and environment influence behavior, you might need spatially resolved measures."

While individually powerful, what might happen if these methods could all be used together to inform each other? "The goal of my lab," Satija said, "is to think about a way that we might be able to form an integrated analysis within and across all of these different datasets, so that instead of looking at these aspects of cellular identity one at a time, we can combine them together and perhaps come up with a more holistic view for what a cell is doing." The experimental and computational methods they have developed for this task were the topic of Satija's talk.

A few years ago, four different groups profiled the human pancreas with single-cell RNA-seq with the goal of discovering cell types in the pancreas. There were four separate experiments done with four different single-cell RNA-seq technologies at four different labs. "We thought it would be interesting if we could integrate and pull these datasets together," Satija said, but when they tried a straightforward type of meta-analysis, the result was a mess. The cells ended up grouping both by their underlying biological states and by the technology used to measure them.

To avoid this problem, Satija looked for a way to align the datasets so that cells in the same biological state grouped together across experiments. He developed a process he called "single-cell RNA-seq data alignment."

The datasets from the four different experiments were very different, and it was difficult to integrate them or compare the cells in one set with those in another. To deal with that issue, Satija's idea was to look for shared sources of variation across the datasets. For example, the correlation between two particular genes was conserved from one dataset to another. That was not by accident, he said. "These genes are both markers of alpha cells, so as a result they're biologically co-regulated, so this correlation should show up regardless of what technology we look at." His group started identifying such shared sources of variation, and then used those shared sources of variation as a scaffold for alignment.

The method works as follows, Satija said, and to keep it simple he used two datasets in his example (see Figure 8-2). The data in the datasets need to be from a single cell, but they do not have to be single-cell RNA-seq. The datasets should share some of the same underlying biological cell populations, but there can be some populations that are present in one dataset but not in the other. The first step is to use a method called canonical correlation analysis to project cells into a common cellular space. The next step is to look for "mutual nearest neighbors." For

each cell in the dataset, one looks for the most similar cell in the other dataset, and then the process is repeated. The lines or connections between mutual nearest neighbors in the two datasets are called anchors (Stuart et al., 2019).

"The accuracy of these anchors is key," Satija said, "because when we're drawing an anchor, we're making a biological statement that we think that two cells share some underlying biological similarity." He referred audience members to his paper (Stuart et al., 2019) for the details on how to ensure that the anchors are correct, which he said is a "key part of the robustness of our method." Once one has the anchors, it is relatively straightforward to pool the datasets into a harmonized reference set or to transfer information from one dataset to another.

Satija next described some of the ways this technique can be used. In one case his group went to the literature and found every single-cell RNA-seq set that had been published since the onset of the technology. Focusing on pancreatic islet cells, they integrated data from 25,000 cells collected with six different technologies from human and mouse subjects. The clusters that emerged were sharp and distinct. "By boosting our statistical power, we dramatically improve our ability to define and discover cellular phenotypes," he said. One particularly impressive aspect of the work, he said, is the way that the datasets aligned across species. That opens a lot of possibilities.

In another example, researchers from the Allen Institute for Brain Science were interested in learning more about von Economo neurons, or spindle neurons, which are vulnerable in certain types of human diseases. Their function is poorly understood, and it is difficult to get healthy brain tissue from humans to study them. The researchers used Satija's tool to align a single-nucleus RNA-seq dataset that they had generated from human brain tissue with a single-cell RNA-seq set they had previously generated in mouse. They found an exact match between the von Economo neurons and a particular cell type in mouse brains. Because there was something known about what that cell type does in mice, it offered some insights into the human von Economo neurons and opened the way to a series of experiments with which to learn more (Hodge et al., 2020). Satija noted, "his is an exciting opportunity in the same way that comparative genomics was essential for being able to interpret the genome."

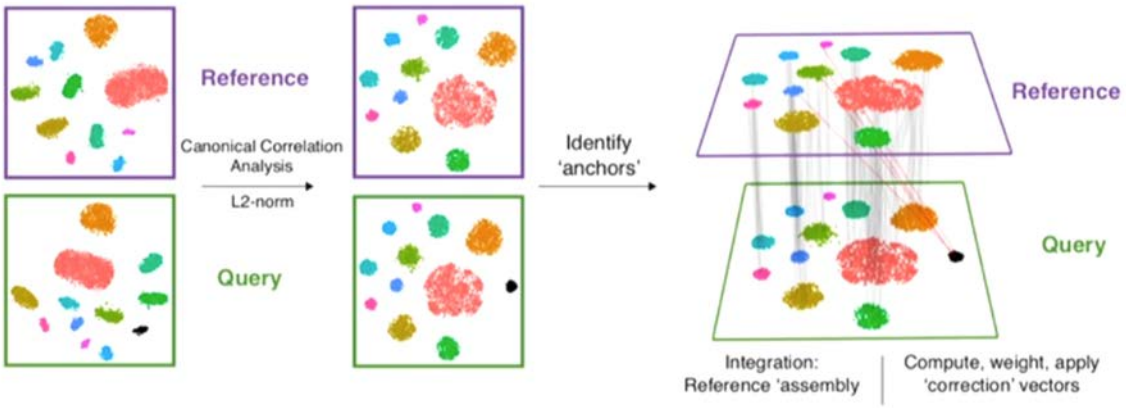

FIGURE 8-2 Representation of how to identify "anchors" across datasets.
SOURCES: Rahul Satija presentation, slide 6; Stuart et al., 2019.

A more ambitious use of the alignment technique is to integrate datasets not just across an organ but across an entire organism. Satija's group has done this with Tabula Muris datasets from the Chan Zuckerberg Biohub. The datasets were generated from 100,000 mouse cells from 26 different tissue types using two completely different technologies. Satija's group was able to align the datasets, and through this process they were able to identify some "exquisitely rare populations" that made up only 0.001 percent of the entire dataset. These populations would have been undiscoverable if they had worked with the datasets independently.

There is no reason to be restricted to just single-cell RNA-seq data, Satija said. The same technique can be extended to single-cell ATAC-seq data and single-cell protein data, for example. He described a case using a new technology that provides a spatial context as well. This new technology, STARmap, measures single-cell gene expression while keeping track of where each cell is located and, in fact, where each molecule is located, in three dimensions. The only downside to STARmap is that the limit is about 1,000 genes in a single experiment, he said, "which is pretty incredible, but it's not transcriptome-wide like single-cell RNA-seq."

His team set out to integrate STARmap (Wang et al., 2018) with single-cell RNA-seq (Tasic et al., 2018). By using the 1,000 genes from the STARmap spatial dataset to find anchors between it and the single-cell RNA-seq, they were able to effectively transfer information on all the rest of the genes that were not in the STARmap dataset and impute spatial gene expression from the single-cell RNA-seq work. After the analysis and validation, they were left with a transcriptome-wide spatial atlas of the mouse brain with "this beautiful gene expression measurement representing different layers in the cortex."

Satija concluded by discussing two key uses for the integration technique that he has developed. One is to assemble a reference atlas. "If you have lots of different experiments produced across different labs, across different consortia, across different technologies . . . you want to build a single cellular atlas. These types of technologies can be used to build those references." The assembly of such an atlas should be unsupervised, he said, and it will dramatically boost the statistical power to define rare and subtle cellular phenotypes.

The second is using the reference atlas to interpret new data. "Aligning datasets to that reference can help you to interpret information from other modalities, whether they're coming from spatial measurements, or epigenomic measurements, or protein measurements," he said. This alignment can also facilitate comparisons between different experimental models, such as aligning human data to mouse data, or to data from other species to perform comparative genomics.

A CLOUD-BASED PLATFORM FOR GENOMICS DATA MINING

The rate at which genomics data are generated is rapidly increasing each year, said Saurabh Sinha of the University of Illinois at Urbana-Champaign (UIUC), to the point that by the year 2025 there will be an astronomical amount of data to deal with. Because genomics is starting to generate much more data than astronomy, perhaps the adjective should be "genomical."

Sinha mentioned that the National Institutes of Health's (NIH's) Genomic Data Commons project is one effort to deal with this issue of huge amounts of data. In particular, he highlighted two goals of that project:

1. Developing and testing cloud-based platforms to store, manage, and interact with biomedical data and tools.

2. Harnessing and further developing community-based tools and services that support interoperability between existing biomedical data and tool repositories as well as portability between service providers.

Those ambitions were the motivation for the establishment of KnowEnG, a center of excellence in big data computing, where he is a co-director and in charge of the data science research arm (Blatti et al., 2020). It is funded by NIH and involves researchers from UIUC and the Mayo Clinic. "Our goal was to build a cloud-based platform for genomics data mining," he said, adding that software sharing and accessibility were two of the main foci.

Their idea at the highest level was that data analysis often involves a researcher having a spreadsheet of -omics data, which can be quite variable, and various forms of analysis can be applied to those data. The KnowEnG platform allows you to upload a spreadsheet and ask questions with machine learning and data mining tools.

In building the system, he and his colleagues thought about the important questions that people would want to address with the types of genomics data the portal was designed to support. The tools they included carry out tasks such as clustering of samples and gene set characterization, prioritization of the most important genes related to a phenotype, phenotype prediction from gene expression, and gene regulatory network reconstruction.

"One of the things we wanted these tools to do was to incorporate prior knowledge," Sinha said. There are many public databases that provide extremely valuable information about genes, proteins, and their properties and relationships, and the people developing KnowEnG wanted the information in these databases to be available to inform the analysis of a user's data. To do that, they captured all that information in a massive heterogeneous network, where the nodes were genes, proteins, and their properties, and the edges represented the relationships that these knowledge bases captured between nodes, such as protein–protein interactions, gene ontology information, pathways, and so on.

With the knowledge network established, the next task was to incorporate a user's analysis. Suppose, for example, Sinha said, that a user has a spreadsheet of gene expression data and wants to do a classification task, perhaps identifying important genes. The knowledge network will use all its information about genes in carrying out the analysis of the information in the spreadsheet. "That's our take on integrating large datasets," Sinha said.

This approach posed two broad categories of challenges. The first set consisted of cyber infrastructure challenges. How can such a platform be made available for everyone to use? The solution they settled on was to put both the knowledge network and the tools to work with it on the cloud, where they would be in contact with each other. This had the advantages of being scalable and does not require any one lab to have the resources to support the knowledge base or the tools.

Building this cyber infrastructure involved "wrapping all the tools in containers so they are easily portable to a variety of different platforms and connecting those tools to storage buckets," The end result, he said, is a cloud-based platform that has a Web portal where users can select a tool from the platform's large collection, upload their spreadsheets, and have their data analyzed in the light of the large knowledge network.

One of the advantages of this approach, Sinha said, is that it is possible to link up with other resources on the cloud. For example, there is the Cancer Genomics Cloud, hosted by the same service, which has its own knowledge set and tools. KnowEnG and this platform are

interoperable so that, for instance, a researcher working with data on the Cancer Genomics Cloud can also use tools on the KnowEnG knowledge network.

The other broad category of challenges facing KnowEnG, he said, relates to the algorithmic issue of how to incorporate the knowledge network into analysis. The major approach used by KnowEnG is based in network diffusion-based approaches. It is very popular in data mining, Sinha said, and is used by Google in its search engine operations.

As an example of how it works, Sinha discussed how the knowledge network had been used to guide the prioritization of genes determining drug response (Emad et al., 2017). The input was gene expression data from multiple cell lines, and the knowledge network was used to "smooth the gene expression data so that each gene's expression not only reflects its measured expression, but also the activity or expression of its network neighbors." The smoothed expression profiles of each of the cell lines were then correlated with drug response data on those cell lines to identify genes whose expression was most predictive of cytotoxicity in response to a particular drug. The smoothed expression profiles can be more powerful than the original expression profile because the smoothed profiles incorporate information from the prior knowledge network.

Sinha offered other examples of this knowledge-network-guided analysis, including a clustering of somatic mutation profiles of cancer patients. This can be difficult because collections of such mutation profiles are generally sparse. Using the knowledge network to smooth the profiles, he said, "can lead to much better detection of similarity among individuals and patients and much better clustering."

In the future, Sinha said, there are still major unsolved problems, such as how to best determine which parts of the knowledge network to use in any given analysis, and how to deal with errors in the knowledge network. Another issue is that while network diffusion-based approaches have been shown to improve accuracy in many cases, they do not necessarily help reveal a mechanism. So, there is the question of how to use existing knowledge bases to make predictive models more mechanistic.

Sinha predicted that there will be more and more large and specialized databases in the cloud, and there will be a growing push for those databases to work together. There will also be a movement toward more automatic uses of knowledge bases by algorithms to analyze users' data. The grand challenge, he said, is to take this to multi-level, mechanism-driven analysis of multi-omics datasets representing multiple scales of biological organization, informed by knowledge bases of molecular properties and interactions, and to be able to do it all on the cloud in a user-friendly way.

SUPPORTING DEVELOPMENT OF METHODS AND TOOLS

Genevieve Haliburton of the Chan Zuckerberg Initiative (CZI) described what the organization is doing to support biological research. Haliburton noted that she would not be presenting any original research because the initiative does not do research. Instead, she said, it attempts to find ways to support the development of methods and tools that can enable researchers such as those present at the workshop to do discovery research.

CZI funds three separate programs, one in science, one in education, and one in justice and opportunity. Because CZI is new, she had few outcomes to describe; instead, she spoke about what sorts of research and development are being funded.

The organization has two arms, a grant-making foundation and a technology organization, which have some overlap. The funding arm provides grants to individual researchers, groups of researchers, and existing standalone partners. One of its program areas is single-cell biology, which is where the organization engages with the functional genomics community. They fund a number of other project types, including open science, which is meant to foster healthy open research communities that really drive innovation. They support open access publication, open source software development, protocol sharing, and free and easy data sharing and reuse.

On the technology side the initiative has an in-house organization with designers, engineers, computational biologists, and others who build open-source software and tools aimed at enabling a wide range of sciences. Separate from the initiative is the Chan Zuckerberg Biohub, a medical science research center.

Haliburton listed several challenges that the initiative has heard about from the functional genomics community and provided details on each. The challenges were

- protocol sharing and standardization,
- integrating data both within a modality and in a multimodal situation, and
- visualizing data for interpretation and collaboration.

The issue of protocol sharing and standardization arises both in the area of experimental methods and in computational tool development. Haliburton pointed to Protocols.io, which Steven Henikoff had mentioned in his talk the day before, as a platform for sharing protocols and as an example of methods sharing that the initiative would like to see more of. Open access to protocols allows for reproducible methods development, she said. In particular, the initiative has been funding the Human Cell Atlas (HCA) community, which now has hundreds of members, 125 publications, and more than 350 ongoing protocol discussions. These discussions are important for understanding how to perform a protocol. They also help to enable a global community of researchers.

The initiative has also funded grants for experimental and computational tool development related to the HCA work and, more recently, has funded collaborative networks in single-cell biology, most of them focused on single organs. The goal is to integrate and build a single network involving mathematicians, geneticists, tissue biologists, and clinicians all working together to provide a more comprehensive view similar to what is being done with the HCA.

The second challenge in functional genomics that Haliburton identified was data integration. “The way that I’m defining it right now is simply combining multiple studies for biological interpretation,” she said. This requires normalization, batch correction, and various other techniques, Haliburton said, and there are many people developing various algorithms in attempts to make this happen. Unfortunately, the algorithms are usually custom-made for a specific function. They are trying to find a more generalizable way to use the large amount of data that exists.

The technology side of CZI is helping develop the data coordination platform for the HCA, which is intended to be robust and scalable with a high-quality user interface, and that will house all of the data that are being generated. The goal is to build a platform that people can interact with on multiple levels, she said, “not necessarily just computational biologists but also people who have a hypothesis and want to go look at it.” Currently, the platform allows only downloads of individual study data, but there are plans to include standardized multi-study, multimodal integration.

Much of this learning is done through community working meetings, Haliburton said. These meetings are called "jams." Algorithm and method developers who have developed something with a specific purpose in mind but have encountered some challenges attend a jam, and others at the jam join in to help figure it out.

As an example, she described a "normjam," where methods developers gathered to discuss high-level questions regarding normalization. "When I say normalizing here," she explained, "I'm just talking about removing the technical variation from, say, a single-cell RNA-seq" in order to focus on the biological variation and not any variation due to differences in the equipment or methods.

Data integration methods inevitably involve trade-offs, Haliburton said. "You have to give up some information in order to align [different] various data types," she said. "The more different your data are, the more you potentially have to give up. And if you know the downstream research question you want to be answering, you have some good idea of what you're willing to give up and what you really need to hold on to." However, she continued, if the goal is to come up with a method that is very generalizable for use in something like the HCA or that is interoperable between different datasets, it is not clear what trade-offs one might want to make. That is an area of active investigation that the CZI is supporting by engaging the community.

Haliburton's last topic was visualization. "We know that a lot of our ability to interpret data and to infer things comes from looking at it from multiple perspectives," she said, and one of the tools that the initiative has developed in-house is "cellxgene" (read as "cell by gene"). "It allows you to visualize any expression matrix or numerical matrix and explore it by a lot of different lines of metadata that you have," she said. The metadata are a crucial tool, she added. For example, much of the work in the area is siloed, and in order to create bridges between these siloes, one needs good metadata describing the experiment in question. "All of the tools that we build are trying to deeply support a lot of different metadata."

DISCUSSION

After these five presentations there was a short discussion on the challenges and successes of integrating large databases.

Norbert Tavares, the manager of the single-cell biology program at CZI, started with a question. "As I'm managing this program, what do I need to look out for? What am I going to be blindsided by later?" More generally, what should funders be thinking about in this area?

Battle listed one of the challenges as the fact that the effects being studied are so highly context dependent. Collecting the data that will elucidate this context dependence will require huge sample sizes. It will be necessary to think ahead about the effects one is looking for, how big they are likely to be, and what sorts of data are needed to answer the questions.

Danko agreed and added that it is important to have people trained to be comfortable with both the biological and statistical sides of this work.

Haliburton added that it is important to think about what one is considering as "phenotypes." If one's phenotype is just an increase in expression or a differential expression, it is difficult to make the connection between that and the real biological questions of interest. So, one challenge is to figure out how to make strong ties to the biology in these large abstract datasets.

Going in a different direction, Satija said that there is often a desire to automate analyses and have them all be done through a standardized pipeline. While it would indeed "be wonderful

to just push a button and have all this magically happen, I feel like a lot of the steps that go into many of these approaches do require some biological supervision." So when one is writing software, the goal should not be to write one piece of code that will be guaranteed to work for everybody, but rather to provide the appropriate amount of flexibility so that users can explore their data and return results that actually make biological sense.

An audience member commented that when researchers make their data available, it would be useful to have it as raw data. Battle commented that this is something that funders have a great deal of control over, because they can require researchers to deposit their raw data.

Philip Benfey of Duke University asked about handling the sort of time-dependent data that are necessary to study the dynamics of the cell. "How do you analyze something that's essentially four-dimensional?"

Satija said he thought that one of the most exciting aspects of single-cell data is that if one is studying a developmental or transitioning process, even if the data seem to represent a static snapshot, there is still variation in those data in terms of exactly how far cells have moved along through the process. By combining a number of seemingly static snapshots, one can create datasets that include a fourth dimension, although that dimension is not time, per se, but rather "pseudo time" defined by the different points in the developmental progression. "We can order different events with extremely high precision and actually get to the point where we can almost understand causality or at least know that A comes before B, and therefore, B did not cause A."

One of the first papers from his lab included data from an adult mouse brain and from a developing mouse brain, Satija said, "and the data from the developing mouse brain looked kind of like a cloud. It looked like the cells hadn't differentiated yet." So using the anchoring technique he described in his talk, his team found anchors between the developing cells and the differentiated cells, which allowed them to use the adult dataset, where there was plenty of diversity, to guide their analysis of the early embryo. With that ability, they were able to see early hints of differentiation and ask, which were the first genes to change?

Gene Robinson of the University of Illinois asked those panel members working with big datasets about their thoughts on genomic security and privacy. "Are there safeguards that need to be put in place?"

Haliburton said that it is something she and her team have thought a lot about. "There are certain policy-driven limitations that are informing the way we're thinking about it," she said. In particular, they are building their platforms to support whatever sort of controlled access might be required by the consent forms that human subjects have signed. "Policy and science aren't always aligned," she said, "and the policy might not actually reflect what the science needs, so we're trying to think deeply and thoughtfully about that because these things could get out of hand, and that would be very bad."

Satija said that given how much personal information most people share on Facebook and other social media—information that can be much more sensitive than genetic information from a research study—it makes sense that when talking to the broader community about genetic privacy, we should not talk about it with fear. Putting genetic privacy in the context of other information is the goal we should be working toward.

Battle added that it is extremely important when getting consent from patients that they understand that "there are breaches of all of our sensitive data." Researchers should do their best to protect people's privacy, but they should also be clear that they cannot guarantee 100 percent certainty.

IMPORTANCE OF CONSORTIA AND LARGE DATABASES

The second part of the discussion focused on consortia and was moderated by Charles Danko of Cornell University. The discussion panel consisted of Felicity Jones of the Friedrich Miescher Laboratory of the Max Planck Society, Alexis Battle of Johns Hopkins University, Saurabh Sinja of UIUC, Rahul Satija of the New York Genome Center, and Sean Hanlon of the National Cancer Institute.

On the first day of the workshop, Aviv Regev offered her thoughts on the importance of large initiatives (see Chapter 2 for further description of Regev's keynote address). Included in this, was a brief overview of the HCA and other functional genomics initiatives.

People have always understood the value of maps, Regev said, so as more and more information was collected about human cells, the realization arose that it would be useful to have an atlas of human cells, and that realization led her and others to start the HCA initiative. The mission of that initiative, she said, was "to create a comprehensive reference map of the types and properties of all human cells, the fundamental unit of life, as a basis for understanding, diagnosing, monitoring, and treating health and disease." The originators of the initiative did not want to promise—or give people the impression that they were promising—that they would be curing all disease. "It's a basic biology problem with basic biology answers. But it's also useful, and that distinction needs to be made and repeated, especially in any public context."

Because humans are diverse and so are their cells, Regev and her colleagues realized that no one lab or institute or country should build the atlas. Instead it should be an international effort that is open to everyone regardless of the funding source. "There's over 1,700 members now in HCA from 71 countries and more than 1,000 institutes." Openness is a core value of the atlas. The data coordination platform, the data releases, the lab protocols—everything the atlas does—is shared and in the open.

The first draft of the atlas is expected to span at least 100 million cells, including most major tissues and systems from healthy donors of both sexes with geographic and ethnic diversity and some age diversity. Ultimately, the comprehensive atlas is expected to have up to 10 billion cells representing all tissues, organs, and systems as well as full organs, again from a diverse group of healthy donors but also with mini-cohorts representing various disease conditions.

A sister initiative to the HCA is the International Common Disease Alliance (ICDA), which is focused on the connection between human genetic function and disease. A recent meeting developed a series of recommendations for the ICDA, which Regev reviewed. The first was to build references atlases of tissues, diseases, and organisms.

A second line of recommendations involved performing massive high-content pooled screens for function. This will require continuing to develop new experimental modalities, as well as increasing efficiency and improving computational approaches.

A third set of recommendations centered on developing new module-level analytics for gene function that associate cell types with genes, use gene modules to understand biology and to increase signal detection, and detect interactions between single nucleotide polymorphisms both within and between gene modules.

Panel Discussion on the Pros and Cons of Consortia and Large Databases

On the second day of the workshop in the discussion period devoted to consortia such as the HCA, the participants talked about the advantages and the disadvantages of such consortia. The discussion was moderated by Charles Danko of Cornell University. The discussion panel consisted of Felicity Jones of the Friedrich Miescher Laboratory of the Max Planck Society, Alexis Battle of Johns Hopkins University, Saurabh Sinja of UIUC, Rahul Satija of the New York Genome Center, and Sean Hanlon of the National Cancer Institute. Danko, the moderator, opened the period by asking each of the panelists to describe whatever experience they had had working with consortia and what had worked and what had not.

Battle started by saying that the consortia she has primarily been involved with were the GTEx Project and also the HCA. She has mixed feelings about consortia, she said, but her work depends on very large datasets and those usually cannot be produced by a single lab. On the other hand, she said, about half of her lab's projects are not working with large data, and she also does small one-on-one collaborations. She noted that these efforts "are often more well thought out in terms of looking for a very specific effect that we think is important." So, both approaches—large consortia and smaller-scale research—can be valuable.

She added that her experiences with consortia have had some really good parts, and consortia helped launch her career. But there are also cases where people, especially students, get lost in the middle. One should also think about the effects that consortia have on training. Bottom line: there are pluses and minuses.

Sinha said that he had also seen positives and negatives from consortia. On the one hand they bring people together to share knowledge and learn from each other in unique ways. On the other hand, because consortium teams are larger, the chances of trainees getting lost are higher. Continuity can also be an issue. For example, if the consortium period ends, then it may be necessary to let go of the large software team that was assembled for the effort. For better or worse, the large datasets assembled by consortia can shape the direction of the science in strong ways.

Satija said that his best experiences with consortia have been in cases where there were small teams of people with complementary expertise working closely together. Some consortium grants explicitly require multiple principle investigators with complementary expertise. In his case there was a technology developer, an immunologist, and a computational biologist. On the flip side, it can be a problem in large consortia with groups that have complementary skill sets who are explicitly assigned different tasks. The groups may have been funded with separate applications and might be in very different parts of the country or world. If those groups are not tightly aligned, it can feel less efficient.

Jones said that her work with sticklebacks depends on having access to the datasets produced by large consortia, and that the stickleback community has received tremendous benefit from consortia, not just in the resources they make available but in terms of training as well. "We used to run a stickleback summer course where people from around the world would come and learn to do transgenics and bioinformatics, QTL [quantitative trait locus] mapping, and so on. . . . We have developed a really tight and friendly network of researchers in this area, and I think the stickleback community has really, really benefited from that."

Hanlon said that the benefits of consortia include not only the data and tools that they generate but also the policies that they have instituted, such as pushing for early data sharing. In his experience, he said, consortia have been positive for trainees. "Many of the programs I'm

involved in have some sort of junior investigator–associated meetings so we can have the trainees involved. Also, there's a leadership opportunity many times for trainees. We have trainees that lead a number of the working groups and really contribute."

An audience member suggested that with the costs of collecting genomics data dropping and the ease of collecting those data increasing, large consortia are likely to become less important.

Battle agreed, but said that there are still some areas, particularly in human genetics, where the necessary datasets are too large for a single lab to collect at this time, "and if we are waiting for it to get cheap enough for one lab to do that or even have the time to do it, we would be waiting quite a long time." Furthermore, a consortium like GTEx makes a major contribution by enforcing the uniformity of its data, which contributes to increased signal during processing.

Satija agreed with Battle and said that human genetics in particular is an area in which it is important to minimize technical variation in order to do extremely large-scale comparisons.

Battle commented that the uniformity of data collection, often an advantage, can also be a disadvantage for consortia. "If a consortium like that makes the wrong choice, then you have spent a giant chunk of change on a technology that's actually not the right one to use."

Danko noted that consortia can also have an impact by standardizing the methods of sample processing and what constitutes good data.

An audience member asked how consortia could standardize data collection and analysis in the face of multiple conditions, for example, when different members of a consortium are working under different environmental conditions.

"I think that's really challenging," Battle said. Particularly as researchers begin to focus more on the environmental effects on human health, standardizing those measurements will be essential. She mentioned a study in which hospitals were given long and detailed instructions on how to collect each sample—exactly where to make a cut, for instance. "They were also extremely careful about recording everything. . . . It is beyond anything I've seen with other projects, and even then we still have huge sources of technical and experimental variation." Central coordination can help, as well as standardizing protocols and sending training teams to help. "But regardless of how perfect you are, you're still going to have huge sources of variation between sites."

Hanlon commented that while it is often important to have standards across a consortium, it is a good idea to have some room for flexibility, and he offered an example of a study of tumors where it would have been impossible to gather all of the required samples if doctors had to adhere to the original instructions.

Gary Churchill said that he appreciates consortia, not only because they are powerful mechanisms for generating large standardized datasets, which are extremely valuable, but also for how they—or at least GTEx—have made it easy for him to get information. "I go to the website. I type in the name of the gene that I'm interested in. I can immediately see which tissues it is expressed in and whether it has a QTL."

Battle agreed and said that, more generally, it is valuable for consortia to have a user-friendly portal. When agencies fund large-scale data collection, she said, one requirement should be that the grantees make the data readily usable and easily accessible.

Hanlon added that the Human Tumor Atlas program was funded through the Cancer Moonshot Program, which has a goal of making data accessible to a broader community—not just biometricians, but also cancer biologists, and even patients and clinicians.

Marc Halfon of the University at Buffalo argued that the agencies who fund the development of databases should be ready to provide the necessary funding to maintain the databases over time. Technology can change, and data can become stale, so maintaining a database means more than just keeping it running as it has been for years.

Hanlon said that some federal agencies, including NIH, have programs aimed at maintaining databases. They are currently thinking about the broader issue of how to sustain them over time.

Maintaining a project may actually be more difficult for smaller-scale efforts, Battle said. "Sharing data and maintaining data and maintaining methods and software [are] really challenging when it's just your own lab." For instance, when lab members move on, it can be difficult to interest incoming students or postdocs in maintaining a previous lab member's software. "At least with consortia there's usually a funding institution that cares that it's maintained, and they will sometimes help," she said.

9

Big-Picture Challenges in Research, Education, and Training

At various points in the workshop, participants took time to step back and examine the larger picture for function genomics. In particular, there were three separate "big picture" discussions that took place: on education and training, on the definition and use of "model systems," and on the social and ethical implications of functional genomics research.

EDUCATION AND TRAINING

As the capacity of functional genomics grows, it is likely that more and more people will look to use its tools to answer questions of interest and develop new capabilities and applications. This in turn will require an increased ability to educate and train people in this field.

While questions around this topic came up throughout the workshop, there was also a session devoted specifically to education and training. This session was moderated by Patricia Wittkopp of the University of Michigan. To begin, participants divided themselves into smaller groups to discuss three questions related to the future of functional genomics education. The three questions were:

- What are the training needs for future genotype-to-phenotype researchers?
- What subjects and topics are important to advance students toward research in functional genomics?
- What are the best strategies for attracting students to this research?

After about 30 minutes of group discussions, the entire group reconvened in plenary, and one participant from each group reported on the main points that the group members discussed. Following these report, there was also a panel discussion. The panel included Terry Magnuson of the University of North Carolina at Chapel Hill, Arnaud Martin of The George Washington University, Lauren O'Connell of Stanford University, Grace Anderson of Octant, and Rebecca Walker of the University of North Carolina at Chapel Hill.

Education and Training Needs

Much of the conversation in the breakout groups and in the discussion that followed the group reports was focused on what students need to learn in order to pursue careers in functional genomics. For example, Trudy Mackay of Clemson University, who reported for one group, said that the participants had talked about the need for biology majors to be exposed to computer science as well as to other cross-disciplinary topics such as statistics, chemistry, and engineering. MacKay noted that her group found themselves "going down a rabbit hole" of naming topics that would be good for students to understand. In addition to those above, they talked about the basic disciplines of genetics, genomics, molecular biology, and then the more population-level subjects such as quantitative population and evolutionary genetics "because that is at the heart of the

genotype-to-phenotype map." Furthermore, students need to learn how to think in evolutionary terms and be exposed to taxonomically diverse systems, all the way up to human biology.

The group also thought it was important to train students to understand the pressing questions in the field, but they were not certain how to accomplish this. "The best we could come up with is that maybe that's up to the advisor," Mackay said, "but I wonder if there are more formal ways of teaching this."

In addition to topics of classes that students should take, the second group also discussed the importance of creating environments where close collaboration can take place among students, such as between computational biologists and experimentalists.

The reporter for group three said that much of their discussion was focused on what the end goal of training should be. "Are we trying to find specialists or generalists? Should everyone be some sort of new hybrid genomic scientist who incorporates all of these different disciplines?" Some in the group were concerned that this could lead to a lack of depth in the various individual disciplines. They came to an agreement that you need researchers who specialize in fields such as experimental biology and computer science, but there still needs to be a "basic level of literacy across these disciplines that all of the trainees need to be getting . . . early."

Marla Sokolowski of the University of Toronto, another group reporter, said her group also grappled with the interdisciplinary issue and the "push–pull" between giving students a broad integrative understanding and providing them with the deep knowledge needed to earn a Ph.D. It is crucial, she said, for students to learn the vocabularies and concepts of the various areas, and one way to do that is to have graduate students work in interdisciplinary groups where they discuss papers on diverse multi-disciplinary topics. Another approach is to offer short courses to introduce students to various topics. There are some areas that all students should be familiar with, Sokolowski said, including strong training in evolution, and be comfortable thinking about issues from an evolutionary perspective. They should all also have some basic familiarity with ethical issues.

These points led to a further discussion of how to break down the silos around scientific disciplines that exist in universities. One way, Sokolowski mentioned, would be through funding that encouraged interdisciplinary education, such as if graduate fellowships made their funding contingent on cross-disciplinary training.

Terry Magnuson described a slightly different approach toward the same end. "We have two curriculums among many: one is called computational biology and the other one is called genetics and molecular biology." The former is focused on dry-lab work, while the latter involves a lot of "wet" bench work. Students are paired up, one from each curriculum, to work together on a project, so that they "cross-train" each other. Magnuson spoke to the success of the program in getting students to start working on tasks, such as "pipeting," or writing in code, in which they are not experts.

Lauren O'Connell added that students should be trained in skills other than analyzing DNA sequences. That is not all that bioinformatics is, she noted, and "often image analysis and phenotyping are more difficult than running something through an RNA-seq pipeline."

Another participant added that students should also be exposed to regulatory issues. Researchers are not trained to navigate regulatory hurdles, such as compliance with protocols for accessing and using samples. Another useful topic for students is safety, not just of the traditional type, but also biosecurity and other issues that could be involved with functional genomics.

Teaching Creativity

Although it was not explicitly contained in any of the questions posed to the breakout groups, the topic of how to teach individual initiative and creativity to students received a great deal of attention. Wittkopp began the discussion with a question: Granted that students must be taught such things as genetics and bioinformatics, how can they learn about scientific inquiry? For those who will be staying in academia, their primary job will be coming up with the next research question.

O'Connell suggested that one issue making it difficult for students to identify scientific gaps is simply that they cannot or do not read scientific papers. Teaching graduate students how to read scientific papers and getting them to read more broadly would be helpful, she said, although she acknowledged that she did not know how to do that.

Wittkopp responded that one tactic she uses is that when a new student comes into the lab, she "asks them to read about 20 to 25 papers and keep a journal." She asks the students to record their thoughts—what the student was wondering about when reading an article. "Then we review that together, and it helps me see where their core interests are even if they can't articulate them."

"One of the things that often gets lost in undergraduate science education is creativity and trying to get people to understand that science as a practice really is creative," said Jeff Dudycha of the University of South Carolina. He reported that he has instituted a number of practices in undergraduate and graduate courses that are intended to encourage creativity. For example, in undergraduate classes he has students write poetry to encourage them to think about creative ways to express scientific ideas. "For grad students, we have free association times sometimes in lab meetings where we try to figure out two different things that are going on in the lab or with another lab or what is our connection to another lab in our department that we may not have an obvious link to." At the undergraduate level he is team-teaching a biology and music course with a composer. In answer to a question from Arnaud Martin about what the students do in the biology and music course, Dudycha said that the class is half biology students and half students in the composing program, and they form teams to build musical simulations of genetic processes. The point is not to make the music sound good, but to make it represent the biology.

Following up on the creativity question, Scott Edwards from Harvard University commented, "At some level you can't really teach creativity, although I think you can expose students to how to identify the limits of knowledge and what's an exciting question." Conversely, creativity can be stifled in cases where an advisor is too heavy handed. There are cases in which a student has been told what to do for the entirety of their Ph.D., and cannot generate a scientific question of their own during their postdoc. Advisors should give more thought to what is in the best interests of their students and less about getting out another high-profile paper, he said.

Attracting Students to the Field

Complimentary to educational needs, there was also a great deal said about how to attract more students to the field, as many participants felt that with the growing power of functional genomics, there will be many more research opportunities to exploit if there are enough researchers to exploit them.

A theme that emerged from the breakout group reports and the ensuing discussions was the importance of reaching students early and exposing them to genomics—either somewhat early in the undergraduate years or, when possible, while they are still in high school. For example, Sokolowski's group reported that they thought it was important for functional genomics to be introduced as a topic in basic biology courses.

Mackay, reporting for her group, suggested that capstone projects, summer internships, and exposure to the research environment are all ways to attract students to the field. She mentioned the importance of giving students hands-on functional genomics experience. "You really want students to be in the wet lab, extracting DNA, getting their results, analyzing them statistically, and being guided through that process."

O'Connell's group concluded that it is important to have some genomics tools, such as rapid sequencing tools, in undergraduate classrooms so that students can use them ad get a sense of what genomics researchers do. "We also talked about integrating bioinformatics into courses on genetics and evolution and each of these core classes rather than having a bioinformatics class on its own," she said.

Another breakout panel reporter, Arnaud Martin, spoke about his own experiences in providing undergraduates with practice in genome editing. "I gather 12 to 16 students in the classroom. We have microscopes, we have little micro manipulators. I have heard of very cheap micro-injectors as well. . . . We acquire Cas9 for very cheap. And we ask the student to run experiments across the semester and actually do loss-of-function assays." They do the work in frogs and butterflies and can observe the effects of removing a gene. "So, the undergrads get beautiful mutant butterflies, and they're super excited," he said. "I think it gets them hooked."

At one point the discussion turned to attracting students from underrepresented groups into science and, particularly, into functional genomics. "There is an entire population of people out there who are at non-research institutions and underrepresented groups at minority colleges and so forth," said one workshop participant. This person had a National Science Foundation (NSF)-funded program to bring students from tribal colleges in New Mexico into the lab and give them research experience.

O'Connell agreed with the idea that these programs are important, saying that she runs a program that brings in community college students. "Most underrepresented groups start off their education at community colleges." Thus, it is important, she said, for researchers and research universities "to also take students from outside your bubble."

Having said that, O'Connell added that it can be difficult to get funding for such endeavors. She has been paying for it with her career grant but is not sure what will happen when that grant ends. "It's hard to get long-term funding for community-based projects," she said.

Edwards said that he had been running a program to bring diverse undergraduate students to the evolution meetings every year and that he, too, had found it can be difficult to secure long-term funding. "My guess is that a number of program officers at NSF would be excited about it, especially if you can show you've had a track record with it." These sorts of activities don't require huge budgets, he added, so that should make the opportunities more attractive for funding agencies.

DETERMINING AND DEFINING "MODEL" SYSTEMS

Another overarching theme of the workshop was the use of model organisms. Related to this, a session on the afternoon of the first day was devoted to a discussion concerning the

concept of "model" organisms. What is a model system in today's biology? Does it even make sense to talk about model systems in a world where it is cheap and easy to sequence any species and analyze its genome? To start, Paul Katz of the University of Massachusetts Amherst offered an introduction to the topic. This was followed by a facilitated discussion among Katz, Lauren O'Connell of Stanford University, Zoe Donaldson of the University of Colorado Boulder, and Dominique Bergman of Stanford University, with further contributions from audience members.

Katz began with two questions: "What do we mean by model organisms? And is it problematic to classify some organisms as models?" To offer some context, he said that when he examined PubMed for the use of terms such as "model organism," "model animal," and "model species," he found that the use of these terms has gone up exponentially. In 1990, the term was still relatively rare, with only a few uses a year, but by 2000 there were nearly 700 examples of such terms used in papers listed in PubMed, and the number had climbed to nearly 1,400 by 2016.

While the term has become increasingly popular, he wanted to know exactly how people have been using it. To answer this question Katz began by examining how various federal science agencies define "model organism." The National Institutes of Health (NIH) website says, "The term 'model organism' includes mammalian models, such as the mouse and rat, and non-mammalian models, such as budding yeast, social amoeba, roundworm, *Arabidopsis*, fruit fly, zebrafish, and frog." Katz noted that the website did not really define the term and that the examples it offered differed in "level of granularity," from a genus name to a "round worm" and the "frog." "It's really not clear what they mean by model organism," he concluded.

Moving to NSF, he noted that different divisions of NSF have contradictory definitions. Some define it as a "traditional laboratory model species," whereas others say that they want to identify new model organisms as they come into use, indicating that they view model organisms not as "traditional laboratory model species" but rather as any organisms used to model specific biological processes, even if that use is new.

Switching to a more basic question, Katz asked, what is a model? There are various definitions, he noted, but science has a specific meaning for the term. "The science definition is that a model is a systematic description of an object or phenomenon that shares important characteristics with the object or phenomenon," he said. In other words, he explained, a model is not the object or phenomenon itself, but rather a model that shares characteristics with the object or phenomenon being studied. Scientific models can be material, visual, mathematical, or computational, Katz said, but they all share one key feature: they explain how something works. As an example, he pointed to the physical model of a DNA molecule that James Watson and Francis Crick built. Not only did the model show what the molecule looked like, but it also made clear how a DNA molecule could be copied to create additional identical molecules.

So, he asked, are model organisms actually scientific models? Do they explain how something works? "In and of themselves," he said, "I don't think so." But because the term has become so common, people tend to conflate model organisms with models, and "this has had really important repercussions."

As an example, Katz pointed to a paper published in the *Proceedings of the National Academy of Sciences* in 2013 (Seok et al., 2013) showing that mouse models of inflammation had almost nothing in common with inflammation in humans. In *The New York Times* article describing the research (Kolata, 2013), one of the study's lead authors, Ronald Davis of Stanford University, explained that they began studying the mouse model of inflammation when they submitted findings from a 10-year study on inflammatory responses in humans and the findings

were rejected because they hadn't validated them in a mouse model. The underlying issue, Katz explained, was that many scientists assumed that inflammation worked the same in humans as in mice because the mouse, being a "model organism," was thought of as a model for how inflammation worked in humans. Yet, when the researchers explored inflammation in mice, they found almost no overlap between how it worked in mice and how it worked in humans.

The idea of using animals as models, Katz said, can be traced back to some extent to August Krogh, a Danish physiologist. In a 1929 article he wrote, "For a large number of problems there will be some animal or a choice of a few such animals, on which it can be most conveniently studied" (Krogh, 1929). This came to be known as "Krogh's principle."

Another issue with models, Katz continued, is that there is a bias in favor of homology instead of convergence for determining general principles. This explains, he said, why NIH prefers mammalian models for studying issues that are important in humans over other invertebrate species.

"Similarity due to homology at one level of organization," he continued, "does not guarantee similarity of mechanism." To illustrate, he described findings from nudibranchs, organisms that he works with. Two different nudibranch species, *Dendronotus iris* and *Melibe leonine*, both swim by flexing from side to side, movements that Katz showed in a video. This behavior is homologous, he said, because all of the species in the clade to which these two species belong swim this way. Furthermore, the two species also have homologous neurons. "We can go from species to species of nudibranchs and find the same individual neurons based on neuroanatomy, and neurochemistry," Katz said. However, closer inspection shows that the neurons are connected differently in the different species, and the neurological mechanisms underlying the swimming are fundamentally different (Sakurai and Katz, 2017).

Finally, Katz offered several concluding points for "seeding a conversation" with the panel members. First, similarity due to homology does not guarantee similarity of mechanism. Second, general principles can be found only by comparing examples of evolutionary convergence. Finally, he said, it is important that people not allow language to corrupt their thinking. In particular, overvaluing "model organisms" can lead to an undervaluing of comparative science, and models are often considered the norms, with other species being thought of as the exceptions. There are several examples where this has happened, he said. "Certainly it happened in *Drosophila* regarding the period genes where they were thought to be the norm, and other insects were weirdos. It turns out *Drosophila* was the weirdo."

Discussion

To begin the discussion, panel member Lauren O'Connell of Stanford University commented that one of the reasons the planning committee chose to include a panel on model organisms was because "we were kind of arguing about the language regarding a model system versus non-model systems and whether that was a useful framework in the first place." At some level, she said, all of the organisms that people study are modeling something fundamental about biology, or else they would be the subject of study, but the use of the term "model organism" can have harmful effects on the way researchers approach scientific questions and the way that reviewers see their grants. "So," she said, "I think language has corrupted our thought a little bit in this process." But, she added that she is heartened by a shift that she is seeing in how people are talking about "model organisms." For instance, the National Institute of General Medical

Sciences now uses the phrase "research organism" instead of "model system" or "model organism," which O'Connell called "a step in the right direction:"

Zoe Donaldson of the University of Colorado Boulder suggested that instead of talking about model organisms, researchers should speak of a research species as a model for studying something, such as a model for studying social behavior. Katz responded that it makes more sense to simply speak of studying social behavior rather than of studying a model of social behavior. "When you say you're studying a model of something," he said, "you're not studying the thing itself. You're studying a representation of that thing, and I think it denigrates our own work to call what we're doing models of something else." Furthermore, he said, using the language in this way gets away from the idea that the behavior in the animal is being studied as a model of human behavior, which is the implication when one speaks of studying "models."

Donaldson agreed. "If you're studying an organism for something that the organism naturally does, you are studying what the organism does and not a model of that behavior."

Dominique Bergmann of Stanford University said that she appreciated Katz's comments about homology versus convergence because it emphasizes how studying plants can provide a powerful tool for finding general principles. For example, if one finds "a set of transcription factors that work in a certain way with certain partners that is absolutely the same" in a plant species and an animal species, that can only be explained by convergence and explains something about the developmental events that drive this convergence.

One speaker noted that discussions around work using non-traditional model organisms, such as *Ciona intestinalis*, end up devolving into conversations about whether the organism is a "model" or not, and are not necessarily about the science being represented. In adopting the term "research organism," she noted that "it's a question of how you differentiate between the level of tool development or the amount of knowledge in each of these different research organisms."

Bergmann responded that it is valuable to talk about different organisms and the tools and investment needed for each. In particular, she said, it is important to keep in mind that there are some pairs of species in which it is useful to study their differences because they are closely related and other pairs in which it is useful to study their differences because they are far apart.

Responding to Bergmann, the audience member said that it is natural when a researcher has a question and wants to pick the best system to address that question, but many times the research is limited by the available tools. "If you don't have the tools to work on your question in that system, you have to go to a system where those tools exist, even if it's suboptimal." Thus, a lot of the discussion about models versus non-models might go away if more generalized tools can be developed.

Katz agreed, and explained that it would democratize the biological world if researchers were able to use many different tools on a wide variety of species. It is important to avoid studying things simply because they can be studied. "If we want to understand the generalities of biology," he said, "we want to spread our net diversely." One must acknowledge that it has been valuable to focus on a small number of species, he added. Classically science has advanced by first going deep on a small number of things and then broadening the focus—and now is the point in genomics when researchers should be looking more broadly, he said.

An audience member commented that one problem, at least in the mouse world, is that researchers have been seduced by their models. "We forget that they're representations, and we forget that they're intended for a specific purpose, to help us learn something about X, and that they may not generalize to everything." Still, it is important to keep in mind that the models are seductive because they do offer many insights into humans.

"What worries me at the end of the day," Donaldson said, "is that what I'm studying in my one species is not generalizable." What she has found in looking at the convergent evolution of specific traits is that while a behavior or phenotype may be similar across species, the underlying processes or mechanisms are different.

Continuing that line of thought, Bergmann said that assuming research is funded to improve lives, one argument for studying diverse organisms is because it may reveal multiple solutions to a problem. That led to a brief discussion among multiple participants about balancing research on single organisms versus research across broad systems in such a way that the research across systems can be used to determine if principles discovered in a single organism represent general patterns or are specific only to that organism.

SOCIETAL AND ETHICAL IMPLICATIONS OF FUNCTIONAL GENOMICS RESEARCH

As the discussions outlined in Chapter 8 about consortia and large databases make clear, functional genomics research can be expected to grow vastly in scope and power in coming years. This makes it particularly important that much thought be given in advance to the sorts of societal and ethical issues raised by this research. Thus, on the afternoon of the first day of the workshop, a panel discussion moderated by Zoe Donaldson of the University of Colorado Boulder was devoted to the societal and ethical implications of functional genomics research. The three panelists—Rebecca Walker of the University of North Carolina at Chapel Hill, Scott Jackson of Bayer Crop Science, and Ronald Sandler of Northeastern University—each offered some initial comments on the topic, followed by a discussion with the audience.

Rebecca Walker spoke first and began by discussing what she referred to as the "ethical dimensions" of research organism choice (see Figure 9-1). Displaying a series of research organisms—*C. elegans*, zebrafish, mouse, dog, chimpanzees—as a linear array, she suggested that there are a number of dimensions along which research organisms can be organized.

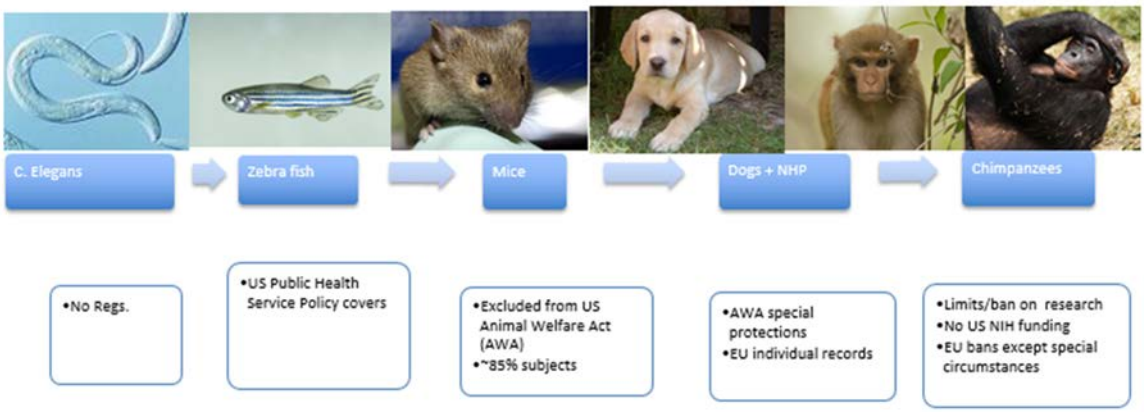

FIGURE 9-1 Ethical dimensions of consideration behind choosing a research organism.
NOTE: EU = European Union; NIH = National Institutes of Health.
SOURCE: Rebecca Walker presentation, slide 1.

One such dimension reflects how research on the different organisms is regulated. There are some for which there are few or no regulations, such as *Caenorhabditis elegans*, and others, such as chimpanzees, on which NIH no longer funds research.

Selecting an organism for research, she said, involves various scientific and pragmatic choices, but also ethical choices that require certain considerations. One such consideration, Walker said, is the capacity for pleasure and pain—which creatures have it, which creatures do not, and what sort of capacity they have. Cognitive capacity is another consideration, she said. How much cognitive capacity does a creature have, and how should that be taken into account? Yet another factor is sociability with its own kind and with other organisms.

How the animals are studied involves another set of ethical considerations, she said. Are they studied in their own environment with non-invasive methods? Are they kept in a laboratory? If so, what things are put into place to make sure that they are capable of flourishing in the environment provided?

Finally, genome editing raises another whole set of issues. As researchers develop an increasingly better understanding of functional genomics and better editing tools, there will be an increasing capability for carrying out genome editing. What sorts of ethical limits should there be on this editing, and what sorts of ethical requirements should there be? One consideration, Walker suggested, should be the purpose of the editing. Editing aimed at preventing disease, for instance, should perhaps be treated differently than editing for enhancement.

Scott Jackson spoke about the use of functional genomics tools in agriculture. Biotechnology is just one of many tools used to improve crops, he said, with the others including such things as breeding, mutagenesis, polyploidy, and interspecies crossing. When the implications for the plants are taken into account, he said, functional genomics is "much less scary than what we've been doing for a thousand years." For example, mutagenesis "really screws up a genome much more than any of the biotech approaches that we talk about." Thus, he said, the ethical implications of using functional genomics tools may be less worrisome than for some types of breeding that are an accepted part of agriculture.

Furthermore, he said, the biotechnology tools used or under consideration for use in agriculture are built on things that occur in nature. The gene transfer technique used in agriculture for decades relies on a bacterium that causes tumors in plants. CRISPR/Cas comes from adaptive immunity in bacteria, which has been used for decades in dairy products.

Jackson noted that gene editing is the most precise breeding method yet, making it possible to edit single genes versus changing hundreds or thousands of genes, as is done in traditional approaches.

"From an ethical perspective," Jackson said, "over the past 30, 40 years, cost has been a major hindrance for [the] public sector and for small companies to get involved in biotechnology." Engineering a trait and taking that trait all the way to the field, including getting through the entire regulatory process, can cost upward of $130 million. Most universities cannot afford to do that, nor can small companies or nonprofits, which means that only a few very large multi-nationals are able to create new crops using biotechnology. It may turn out to cost less with gene editing techniques. The regulatory system in the United States and Canada will also be much easier to navigate than the one in Europe.

The final short presentation was by Ronald Sandler, a professor of philosophy and the director of the Ethics Institute at Northeastern University, who discussed the ethics of biotechnology in conservation. There is already a robust discussion going on among ethicists

about the ethics of using biotechnology in a conservation context, he said, and his goal was to offer a quick overview of that discourse.

He began by listing some of the prominent cases being discussed. There is, for example, talk of using synthetic biology and conservation cloning to increase the genetic diversity of populations that have been through a genetic bottleneck, such as the black-footed ferret. Some have suggested using gene drives to suppress or eliminate populations of invasive species, such as rodents on islands. There has also been discussion about using genetic modification to help certain plant species, such as modifying the American chestnut to make it resistant to chestnut blight.

Much of the discussion about the ethics of these technologies concerns how to use them responsibly in a risk–benefit sense, Sandler said, and the conversation involves such factors as risk assessment and analysis, risk management, cost–benefit analysis, and opportunity cost. In other words, he said, the ethics discussion sees these new technologies as tools, and "then the question becomes how you use these in a way to get the benefits and to avoid the risk and unintended consequences."

However, he continued, ethics is much richer than that. There are many other ethical issues associated with biotechnology in a conservation context. The goal of the standard conservation approach, he said, is to eliminate anthropogenic impacts wherever possible, and the strategies for this typically have to do with limiting human activities in and around a space. Examples include carrying out captive breeding programs and performing ecological restorations.

However, that standard paradigm is increasingly insufficient, he said, in large part because of climate change. Macro-scale, high-rate, high-magnitude ecological change is putting a great deal of stress on the standard conservation paradigm in a couple of ways. First, he said, when the reasons that species are at risk are grounded in climatic processes, local strategies do not work. Second, in more and more cases, it is not possible to undo the human impacts. "So, one of the reasons people are attracted to biotechnologies in conservation is because they offer new strategies for these really hard problems."

The standard conservation model, he explained, is all about limiting human activities or undoing human impact. "But when you start to talk about using genetic tools. . . . It's no longer thinking about how we have to change ourselves in order to accommodate other populations; it's more about thinking how we change these populations so they're better suited to our world." That is a radical change in conservation philosophy. Furthermore, the values underlying the standard conservation paradigm do not support—and, indeed, are in tension with—biotechnological interventions, and so it becomes necessary to make trade-offs.

The bottom line is that not only are the scientific issues involved with functional genomics complex, but so are the ethical issues. The risk–benefit ethical considerations are still there, and they are still extremely important in oversight and public engagement, Sandler said, but when functional genomics interacts with conservation, many other values come into play as well, and the way they intersect is complex. "The ethical issues become not just can you do this safely," he said, "but does this preserve other sorts of values that we care about?" And there are other questions, such as what should be the goals of conservation, and how do we see our relation with the natural world if we are starting to design it and modify it in a more intensive way?

DISCUSSION

Following those short presentations, Donaldson opened the session up to discussion among the panelists and audience members. To get the conversation started she asked a question of the panel: "What are some of the best strategies to avoid the hubris of previous generations? So many invasive species exist because they were seen as a solution to a problem."

Sandler offered his thoughts about one aspect of that problem—trying to restore ecosystems to their previous state after they have been affected by the introduction of invasive species or other sorts of damage. For a decade or so, restorationists have realized that "historical reference conditions aren't going to be as good a proxy for future ecological integrity as they have been in the past." As a result, they are faced with the question, "To what extent can we continue to call something restoration as it becomes more forward-looking, when we're not using those reference conditions?" One of the reasons that using historical reference conditions was attractive was that those conditions served as a check on hubris because people were not trying to design the systems but only return them to a previous state. In a world where the future climate will be different from that of the past, however, "the question becomes, what are you designing things for? And it becomes more of an open question that I don't have an answer for."

In response, Walker suggested that it is important to foster intellectual humility in such a way that people do not believe that they know more than they actually do about the effects certain actions might have.

Sandler commented that people in the ethics field do not think in terms of yes or no, do or do not. Instead, when thinking about a technology, ethicists will ask how to maximize the social and ecological benefits of the technology while avoiding negative consequences. Their goal is to help inform the design and implementation of the technology in ways that maximize the good.

In response to an audience question about genetically modifying animals, Walker said that people need to think carefully about whether they are modifying the animals in ways that would undermine or increase their welfare.

Chris Peterson from the U.S. Department of Agriculture noted that the public is often leery of, if not completely opposed to, genetically modified organisms despite various efforts to help gain their acceptance. "What needs to change in our messaging so that we can make some progress in communicating what these technologies have to offer?"

Sandler answered that from the point of view of consumers, there seems to be no benefit to a genetically modified organism such as Bt corn—it still looks the same and tastes the same and has the same nutritional content—but now there is something new in it that they do not fully understand. Perhaps it is good for the people producing and selling the product, but consumers may not see the benefit to them, so Sandler thinks it is a deeper problem than just finding the right public education campaign. Furthermore, he said, there is plenty of literature showing that people's resistance to genetically modified foods is not so much an information deficit issue as it is a trust issue.

In response to a participant's comment that research animals seem to recieve much more ethical consideration than farm animals, Walker said that it seems to be a form of "research exceptionalism." When humans take part in research, for instance, a high level of informed consent is required even for actions that other people may just choose to do. Something similar may be taking place in regard to animals.

Steven Moss from the National Academies of Sciences, Engineering, and Medicine asked about the ethics of editing species in nature. He mentioned specifically a study on corals which brought up the possibility of using genomic editing tools in corals as well as the possibility of using genomic tools to deal with invasive species. At what point could these things be reasonably considered?

Walker commented that there is a major distinction between genetically modified organisms that are isolated from the environment and those that are put into the environment. So, in the cases where the genetically modified organisms will be released into the environment, it will be important to be much more careful in thinking about potential consequences.

Sandler agreed. "I would say the earlier the better" because there are so many considerations to take into account. "We want to run these value analyses to understand all the different ways in which proposed interventions could intersect with cultural values or ethnic values or symbolic values. . . . I'm thinking particularly about how the non-human world is valued by different groups of people that you might not think of initially as being relevant." Furthermore, it is important to be aware of the danger of hubris in making these decisions. In particular, he addressed the tendency of people to overestimate their ability to predict and control complex biological and ecological systems. There is always going to be an element of uncertainty, so at what point does a willingness to accept a certain amount of uncertainty become hubris? The question of how much confidence to have versus how much precaution to take is not just for researchers to answer but also for people who might be affected by the answer.

Nathan Springer of the University of Minnesota pointed out that there is a massive policy lag in regulating many of these technologies—that the technologies offer various capabilities that the regulations were not designed to take into account. Today's policies relevant to functional genomics, for instance, almost all date to before the development of CRISPR, which has been a game changer. Walker added that the problem has been made worse by the fact that ethical policies concerning genome editing are almost entirely focused on humans, but that is not where most of the work is being done in the genomics area. There are about 63 different policy proposals floating around related to genome editing in humans, she said, but that is not where the most attention is focused.

Sandler responded that fast-moving technologies will always get out in front of policies and regulations. This is why ethics is an important topic for researchers to take seriously. "The research community and other folks who could potentially implement this stuff are going to be facing these hard questions before they have good guidance from regulatory bodies." If those researchers on the technological frontier think carefully through the issues, their practices can serve as the bases for the regulations or policies that come along afterward.

Jackson commented that this is what happened with genetically modified organisms in the 1980s and 1990s—that many of the policies eventually put into effect by regulatory agencies were first developed by the companies working on these organisms. Donaldson pointed out that something similar happened with recombinant DNA, as the scientists who were developing that technology spent a great deal of time thinking about the ethics before the use of the technology became widespread. Similarly, Sandler added, researchers have been proactive in the areas of cloning and various kinds of human modification, considering the ethics before the capabilities have been fully developed.

10

Future of Functional Genomics

The workshop's final, two-part session was devoted to looking at functional genomics going forward—where it is going, what obstacles might be encountered, and what might be done to help ensure the field's ability to advance? The first part was an interactive session in which the participants divided into breakout groups to discuss a series of questions, after which one person from each group reported on those discussions. The second part was a "town hall" in which everyone had the chance to address the points heard during the meeting or bring up issues that had not been discussed.

Moderator Emma Farley of the University of California, San Diego, opened the first part of the session by describing its format and providing a list of issues and questions for each breakout group to address:

- List 5 to 10 research and knowledge goals for the field of functional genomics. Categorize each as a short-, medium-, or long-term goal.
- What obstacles are preventing these research and knowledge goals from being realized?
- What specific pathways or strategies could be used to overcome these obstacles? Are there strategies that could be used to overcome more than one of the obstacles listed?

BREAKOUT GROUP DISCUSSIONS ON THE FUTURE OF FUNCTIONAL GENOMICS

The first group reporter, Gene Robinson of the University of Illinois at Urbana-Champaign, said that under the research and knowledge goals should be "networks, networks, and more networks." As a tool, networks can be used to study gene–gene interactions, protein–protein interactions, transcriptional regulatory interactions, chromatin, etc. One goal that came up during the workshop, noted Robinson, is to develop ways to integrate networks and be able to use them at different scales and across different species to be able to derive predictive information from those networks that could guide future work.

A second goal would be to develop tools to manipulate networks in graded ways to be able to exert minor perturbations to nodes, both for validation purposes and for prediction.

A third goal is to be able to distinguish the different forms of "functional." The word "functional" has different meanings, Robinson noted. Over the course of the workshop, "functional" was used mostly in the context of validation, but a deeper goal of functional genomics is to understand mechanism. So "functional" in the context of validation-type analyses should be differentiated from "functional" in the context of mechanistic analysis.

A fourth goal is "to understand how networks co-evolve within species and with lineage specificity."

A fifth would be to grapple with the issue of the right unit of analysis for different sorts of research. "Is it the gene, or is it the gene variant? Are we focused at the right level?"

Beyond that, researchers should develop standardized methods for providing phenotypic information, Robinson said. Databases are offering an increasing number of standardized ways to deposit genetic and genomic information and then to access that information. However, there are not yet comparable ways of integrating phenotypic information and making it accessible.

Finally, Robinson said, sequencing the genomes of more species would be a useful foundational effort. At present only 0.2 percent of the genomes of eukaryotic species have been sequenced.

Turning to obstacles, Robinson first mentioned funding for genomic infrastructure. An example of the value of such spending is the success of the National Science Foundation's (NSF's) Plant Genome Research Program, which has transformed that community.

A second obstacle is the fact that model genetic systems, which were previously well funded at the National Institutes of Health (NIH), are systematically being pushed out, specifically the historic model genetic systems such as *Arabidopsis*, *Escherichia coli*, *Drosophila*, and *Caenorhabditis elegans*. At NSF, these model systems are considered the province of NIH, and so proposed research on these species is less welcome at NSF. This leaves researchers in those model organism communities feeling that they do not have a home.

Concerning strategies to overcome obstacles, the group pointed to successful programs in integrating math and biology, which could be applied specifically to functional genomics.

Philip Benfey of Duke University reported for the second group. Viewing the exercise as a form of strategic planning, the group took the approach of splitting their goals into specific objectives, so that one can focus on individual objectives instead of mixing them together.

In the group, Benfey said, there was a strong consensus that functional genomics is a set of tools as opposed to an end in itself, and so the group's objectives were biological in the form of the following questions: How do genes work? How do regulatory networks function? How does multi-cellularity function? What makes organisms different among themselves and within a species? How do organisms adapt and speciate?

The group offered specific goals for these high-level objectives. One is the need for understanding regulatory networks, as Robinson had already emphasized. Another is the ability to perturb at scale and observe how the network, not just the individual genes, responds. A third is to be able to define all the components of a network, and not just its component genes. Finally, the group offered a number of specific goals concerning how genes work; how they are organized, both in two and three dimensions; what they do; and when and where they are expressed.

Among the obstacles the group listed, Benfey said, was "the impossibility of doing a totally comprehensive analysis of anything," which arises from the fact that even reasonably simply systems have far too many possible combinations to completely analyze every one. There were various suggestions for how to deal with that issue. One was to take a deep dive into an area that is of interest. Or else one could try, as Aviv Regev discussed in her keynote address, doing a targeted random sampling that might provide 80 to 90 percent understanding by looking at only 20 percent of the relevant parts. Another approach would be to bring together as much existing knowledge as possible in one place so that it can be easily queried.

Finally, Benfey said, the group spent some time on what is more a philosophical issue. "As humans, we need linear narratives. We think in straight lines." But functional genomics is anything but linear. It pulls together many different factors. So how should one think about them? How should they be written about or presented? How does one get funding for them?

"The problem comes back not to the quality or type of data, but to our inability to think of them except in linear narratives."

Switching to the topic of challenges, Benfey noted the fact that functional genomics is all about generating large datasets. What is the best way to ensure that the people who generate them get credit for their work? How should the people producing those datasets be trained?

Some of the answers might be found, Benfey said, by looking to the success of genome sequencing. When researchers started sequencing a lot of genomes three decades ago, it was done in several different ways. In the case of *E. coli*, the sequencing was distributed to a large number of different labs, taking them 15 years to complete. In other cases the sequencing was done in centralized locations. Examining the successes and failures of these efforts could be useful to understand how to train the next generation of functional genomics researchers.

Reporting for the third group, Lauren O'Connell of Stanford University commented that her group had covered much of the same ground as the previous ones. One challenge her group identified was that multiple timescales and feedback loops are important in biological systems, but research measurements tend to be static and linear. A second concern was the lack of focus on phenotypes. To date, genotypes have received more attention because they are easier to measure, but a change is due. "We need ways to measure phenotypic diversity that is quantitative so that we can understand how environmental inputs shape the phenotype." What makes some species more plastic than others? What are the genetic elements that govern plasticity? Her group thought these were major questions, O'Connell said.

Other challenges that O'Connell's group discussed included the lack of generalizable computational and genomics tools, the lack of annotation of existing genomes, and whether to embrace diversity in humans and animals. So far, she said, genomics researchers do not know a lot about the genomes that they have, so would more genomes help? The answer the group came up with was "Yes, but we need to be able to annotate them." Just having a sequence is not enough, she said. Annotation "is quite hard and a big bottleneck in our community."

Richard Dixon reported for the fourth group. The group came up with three goals. The first was more multi-omics work and, specifically, proteomics. "We thought that the proteins were more predictive, and we want to get more goals based around that." A second goal was adding an ecological context to provide more details for the genotype-by-environment interaction. Finally, more diverse models could be useful, moving beyond those such as the typical inbred mouse models.

Concerning obstacles, the group pointed toward the usefulness of interoperability of technologies across organisms and the fact that proteomics technology is not yet sufficiently mature or robust. Also, they reemphasized the importance of education being both wide and deep.

As for specific strategies to reach these goals and overcome these obstacles, the group identified better undergraduate training and, specifically, cross-training in various fields, not just in different areas of biology, but in math and computational methods as well.

COMMENTS FROM THE TOWN HALL DISCUSSION

For the final discussion, moderator Gene Robinson emphasized that anyone who wished to make sure that a particular point of view was included in the proceedings should take this opportunity to make a comment.

The first commenter was Eve Wurtele from Iowa State University who made a case for the importance of what she called the "dark transcriptome." In any organism there are genes that are unique to that organism. These are called species-specific genes or orphan genes. In *Arabidopsis thaliana*, for example, some of the genes that have not been seen in any other species are very young genes and protein coding genes.

The functions of some of these species-specific genes involve interacting with external organisms. Examples include attractants and the toxins of jellyfish. Another function involves interacting with an organism's own existing networks. An example of an orphan gene that does both is the *qqs* gene of *Arabidopsis*. Its product both mediates predator resistance and increases the plant's protein content. Because it interacts with internal networks, it can be taken from *Arabidopsis* and transferred into corn or soybean, where it also produces pathogen resistance and an increase in protein.

While some orphan genes are known to have functions, Wurtele said, scientists "don't know how many of them do anything because people don't usually study them." The problem is that when RNA-seq data and proteomics data are analyzed, the analysis generally only includes genes or proteins that align with known genes or proteins. The result is that many orphan genes, particularly those that are younger, do not get annotated. Of 1,000 or so known orphan genes in *Arabidopsis*, for example, only about 10 percent have been annotated, Wurtele added.

Furthermore, evidence indicates that many of these orphan genes are transcribed and translated. "We do know that about 80 percent of these, at least in yeast and some other organisms, make proteins," Wurtele said. The bottom line, she said, is that there is a dark transcriptome that includes not only many orphan protein-coding genes but also non-coding genes as well, "and all this is probably intimately involved in functional genomics."

The next speaker touched on several topics including the observation that as sample sizes become increasingly large, it is inevitable that more and more genes will be found to have a connection with a particular phenotype. On the other hand, "we know . . . that not every gene is involved in every phenotype." Furthermore, of the many genes that might play some role in a particular phenotype, some will inevitably be more important than others. The question, then, is how we find those genes and gene networks that are more important?

Scott Jackson of Bayer Crop Science asked Donal Manahan from NSF if there is an interagency working group in genomics or functional genomics similar to one that had existed for plant genomes. Manahan answered that there is such a group involving NIH, the U.S. Department of Agriculture, and NSF. Its first big meeting took place in August 2019, with many of the same people at the workshop in attendance. "It isn't a formal interagency group," he said, "but there's certainly been a great deal of informal conversation over the last 6 months that we want to continue."

Gary Churchill of The Jackson Laboratory offered a comment about the importance of variation in functional genomics. Variation underlies much of functional genomics, he said. "It's the father of evolution. It's what makes us all unique." Variation makes possible many of the important approaches in functional genomics, and it is certainly important to understand how natural variability is distributed throughout populations. "But," he said, "if we're interested in function, natural genetic variation may not be ideally distributed in natural populations." And that is why it is important to have constructs like the *Drosophila* Genetic Reference Panel and various other panels such as those involving mice, corn, and *Arabidopsis* "where we can bring together genetic variation and use it as a tool."

Next Emma Farley of the University of California, San Diego, made a comment about what people mean by "functional genomics." "I'm wondering if maybe I don't understand what functional genomics is," she said. "Two people have said to me that functional genomics is a set of tools and functional genomics is generating large datasets. That's not how I think about functional genomics." To her, she said, functional genomics is about trying to understand how the genome encodes biological function, whether it is how development is encoded, or how changes in the genome lead to evolutionary adaptations, or how the genome interacts with the environment. The tools were developed to help answer these questions. "I think that's an important distinction."

Next, she addressed Robinson's comment from the first part of the session that one should keep in mind the distinction between functionality in relation to validation and functionality in regard to mechanism. "I would disagree that they're different functions," she said, "because what we're trying to do is understand this really complex problem and we can use sparse sampling to find structure, and that's looking for these patterns, but we need to equally validate that the patterns we're seeing are accurate and that these patterns and biological structure actually are signatures of mechanism." Thus, the two things cannot really be separated. "It's really important to have validation of the true functional data so that you can actually get at mechanism."

CLOSING REMARKS AND FINAL OVERVIEW

Donal Manahan of NSF took the microphone to make some closing comments. First, he described the attendees as a "fearless bunch" because of their willingness to work in a field where the magnitude of the data and of the possibilities are so large. "Let's start trying to make the case that one reason the science of biology is so dominant in the 21st century is that we're starting to get our heads around scales of numbers that were just completely unheard of 10, 20, 30 years ago," he said. Similarly, the group has no fear in moving from one area to the next—from molecular biology to cell biology, to organismal biology, to considerations of diversity in model systems, all put in an evolutionary context as the need arises. Furthermore, he said, "you had no fear at all of considering the massive infrastructure needs that are going to be needed to address this. You were honest, I felt, in pointing out the sometimes inadequacy of the way our current university systems and others are structured to be able to take on training for the next generation."

One of the key lessons he said he would take from the workshop was the need for a new type of training. "Most of us in this room are professors or have had teaching experience, so we have a sense of how to train, but I think we recognize that the training of the 20th century isn't really going to work for the biology of the 21st century when we take on these ginormous numbers as we think about the future of life sciences."

Robinson, the chair of the workshop's planning committee, closed out the session and the workshop with some brief comments. "I began a couple of days ago by saying that I thought we were at an interesting point in time with respect to genomics and the need to think about what the next steps are," he said. "I think our conversations the last couple of days have really borne that out. We've heard some amazing science and we've also heard some really passionate expressions of frustration and need for how to go to the next level."

Genomics is a very young science, he noted. It is only 40 years old. Furthermore, it is different from many other types of science. It is more of an enabling science that spans all of the classic subdisciplines in biology. "So we are pioneers," he said. "We're making it up here—how

to take the initial discoveries and turn them into deep understandings of biology." So, hopefully the conversations throughout the workshop have planted ideas in the minds of funders, especially NSF, about how to help the field of functional genomics go to the next level.

References

Adams, D. C., and M. L. Collyer. 2017. Multivariate phylogenetic comparative methods: Evaluations, comparisons, and recommendations. *Systematic Biology* 67(1):14-31.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287(5461):2185-2195.

Aguet, F., A. N. Barbeira, R. Bonazzola, A. Brown, S. E. Castel, B. Jo, S. Kasela, S. Kim-Hellmuth, Y. Liang, M. Oliva, P. E. Parsana, E. Flynn, L. Fresard, E. R. Gaamzon, A. R. Hamel, Y. He, F. Hormozdiari, P. Mohammadi, M. Muñoz-Aguirre, Y. Park, A. Saha, A. V. Segrć, B. J. Strober, X. Wen, V. Wucher, S. Das, D. Garrido-Martín, N. R. Gay, R. E. Handsaker, P. J. Hoffman, S. Kashin, A. Kwong, X. Li, D. MacArthur, J. M. Rouhana, M. Stephens, E. Todres, A. Viñuela, G. Wang, Y. Zou, C. D. Brown, N. Cox, E. Dermitzakis, B. E. Engelhardt, G. Getz, R. Guigo, S. B. Montgomery, B. E. Stranger, H. K. Im, A. Battle, K. G. Ardlie, and T. Lappalainen. 2019. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* 787903.

Albuisson, J., B. Isidor, M. Giraud, O. Pichon, T. Marsaud, A. David, C. Le Caignec, and S. Bezieau. 2011. Identification of two novel mutations in SHH long-range regulator associated with familial pre-axial polydactyly. *Clinical Genetics* 79(4):371-377.

Anderson, G. R., P. S. Winter, K. H. Lin, D. P. Nussbaum, M. Cakir, E. M. Stein, R. S. Soderquist, L. Crawford, J. C. Leeds, R. Newcomb, P. Stepp, C. Yip, S. E. Wardell, J. P. Tingley, M. Ali, M. Xu, M. Ryan, S. J. McCall, A. J. McRee, C. M. Counter, C. J. Der, and K. C. Wood. 2017. A landscape of therapeutic cooperativity in *KRAS* mutant cancers reveals principles for controlling tumor evolution. *Cell Reports* 20(4):999-1015.

Anderson, S. N., M. C. Stitzer, A. B. Brohammer, P. Zhou, J. M. Noshay, C. H. O'Connor, C. D. Hirsch, J. Ross-Ibarra, C. N. Hirsch, and N. M. Springer. 2019. Transposable elements contribute to dynamic genome content in maize. *The Plant Journal* 100(5):1052-1065.

Benfey, P. N., P. J. Linstead, K. Roberts, J. W. Schiefelbein, M. T. Hauser, and R. A. Aeschbacher. 1993. Root development in *Arabidopsis*: Four mutants with dramatically altered root morphogenesis. *Development* 119(1):57-70.

Bernstein, M. R., S. Zdraljevic, E. C. Andersen, and M. V. Rockman. 2019. Tightly linked antagonistic-effect loci underlie polygenic phenotypic variation in *C. elegans*. *Evolution Letters* 3(5):462-473.

Bielecki, P., S. J. Riesenfeld, M. S. Kowalczyk, M. C. Amezcua Vesely, L. Kroehling, P. Yaghoubi, D. Dionne, A. Jarret, H. R. Steach, H. M. McGee, C. B. M. Porter, P. Licona-Limon, W. Bailis, R. P. Jackson, N. Gagliani, R. M. Locksley, A. Regev, and R. A. Flavell. 2018. Skin inflammation driven by differentiation of quiescent tissue-resident ILCs into a spectrum of pathogenic effectors. *bioRxiv* 461228.

Blatti, C., III, A. Emad, M. J. Berry, L. Gatzke, M. Epstein, D. Lanier, P. Rizal, J. Ge, X. Liao, O. Sobh, M. Lambert, C. S. Post, J. Xiao, P. Groves, A. T. Epstein, X. Chen, S. Srinivasan, E. Lehnert, K. R. Kalari, L. Wang, R. M. Weinshilboum, J. S. Song, C. V. Jongeneel, J. Han, U. Ravaioli, N. Sobh, C. B. Bushell, and S. Sinha. 2020. Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform. *PLOS Biology* 18(1):e3000583.

Bloom, J. S., J. Boocock, S. Treusch, M. J. Sadhu, L. Day, H. Oates-Barker, and L. Kruglyak. 2019. Rare variants contribute disproportionately to quantitative trait variation in yeast. *eLife* 8:e49212.

Blumberg, A., Y. Zhao, Y.-F. Huang, N. Dukler, E. J. Rice, K. Krumholz, C. G. Danko, and A. Siepel. 2019. Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. *bioRxiv* 690644.

Bolnick, D. I., R. D. H. Barrett, K. B. Oke, D. J. Rennison, and Y. E. Stuart. 2018. (non)parallel evolution. *Annual Review of Ecology, Evolution, and Systematics* 49(1):303-330.

Brady, S. M., D. A. Orlando, J.-Y. Lee, J. Y. Wang, J. Koch, J. R. Dinneny, D. Mace, U. Ohler, and P. N. Benfey. 2007. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* 318(5851):801-806.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* 282(5396):2012-2018.

Chapman, P. B., A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S. J. O'Day, J. A. Sosman, J. M. Kirkwood, A. M. M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R. J. Lee, K. T. Flaherty, and G. A. McArthur. 2011. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine* 364(26):2507-2516.

Chen, Q., J. He, C. Ma, D. Yu, and L. Kang. 2015. *Syntaxin 1A* modulates the sexual maturity rate and progeny egg size related to phase changes in locusts. *Insect Biochemistry and Molecular Biology* 56:1-8.

Chick, J. M., S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L. Svenson, G. A. Churchill, and S. P. Gygi. 2016. Defining the consequences of genetic variation on a proteome-wide scale. *Nature* 534(7608):500-505.

Chinwalla, A. T., L. L. Cook, K. D. Delehaunty, G. A. Fewell, L. A. Fulton, R. S. Fulton, T. A. Graves, L. W. Hillier, E. R. Mardis, J. D. McPherson, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520-562.

Chu, T., E. J. Rice, G. T. Booth, H. H. Salamanca, Z. Wang, L. J. Core, S. L. Longo, R. J. Corona, L. S. Chin, J. T. Lis, H. Kwak, and C. G. Danko. 2018. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nature Genetics* 50(11):1553-1564.

Cleary, B., L. Cong, A. Cheung, E. S. Lander, and A. Regev. 2017. Efficient generation of transcriptomic profiles by random composite measurements. *Cell* 171(6):1424-1436.

Cooley, A. M., L. Shefner, W. N. McLaughlin, E. E. Stewart, and P. J. Wittkopp. 2012. The ontogeny of color: Developmental origins of divergent pigmentation in *Drosophila americana* and *D. novamexicana*. *Evolution & Development* 14(4):317-325.

Coolon, J. D., C. J. McManus, K. R. Stevenson, B. R. Graveley, and P. J. Wittkopp. 2014. Tempo and mode of regulatory evolution in *Drosophila*. *Genome Research* 24:797-808.

Cox, A. D., S. W. Fesik, A. C. Kimmelman, J. Luo, and C. J. Der. 2014. Drugging the undruggable RAS: Mission possible? *Nature Reviews Drug Discovery* 13(11):828-851.

Crow, J. F. 1997. Birth defects, Jimson weeds and bell curves. *Genetics* 147(1):1-6.

Cruz-Ramírez, A., S. Díaz-Triviño, I. Blilou, V. A. Grieneisen, R. Sozzani, C. Zamioudis, P. Miskolczi, J. Nieuwland, R. Benjamins, P. Dhonukshe, J. Caballero-Pérez, B. Horvath, Y. Long, A. P. Mähönen, H. Zhang, J. Xu, J. A. H. Murray, P. N. Benfey, L. Bako, A. F. M. Marée, and B. Scheres. 2012. A bistable circuit involving SCARECROW-RETINOBLASTOMA integrates cues to inform asymmetric stem cell division. *Cell* 150(5):1002-1015.

Cui, H., M. P. Levesque, T. Vernoux, J. W. Jung, A. J. Paquette, K. L. Gallagher, J. Y. Wang, I. Blilou, B. Scheres, and P. N. Benfey. 2007. An evolutionarily conserved mechanism delimiting SHR movement defines a single layer of endodermis in plants. *Science* 316(5823):421-425.

de Bakker, M. A. G., D. A. Fowler, K. den Oude, E. M. Dondorp, M. C. G. Navas, J. O. Horbanczuk, J.-Y. Sire, D. Szczerbińska, and M. K. Richardson. 2013. Digit loss in archosaur evolution and the interplay between selection and constraints. *Nature* 500(7463):445-448.

de Boer, C. G., E. D. Vaishnav, R. Sadeh, E. L. Abeyta, N. Friedman, and A. Regev. 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology* 38(1):56-65.

Denyer, T., X. Ma, S. Klesen, E. Scacchi, K. Nieselt, and M. C. P. Timmermans. 2019. Spatiotemporal developmental trajectories in the *Arabidopsis* root revealed using high-throughput single-cell RNA sequencing. *Developmental Cell* 48(6):840-852.

Di Laurenzio, L., J. Wysocka-Diller, J. E. Malamy, L. Pysh, Y. Helariutta, G. Freshour, M. G. Hahn, K. A. Feldmann, and P. N. Benfey. 1996. The *SCARECROW* gene regulates an asymmetric cell division that is essential for generating the radial organization of the *Arabidopsis* root. *Cell* 86(3):423-433.

Dinneny, J. R., T. A. Long, J. Y. Wang, J. W. Jung, D. Mace, S. Pointer, C. Barron, S. M. Brady, J. Schiefelbein, and P. N. Benfey. 2008. Cell identity mediates the response of *Arabidopsis* roots to abiotic stress. *Science* 320(5878):942-945.

Dixit, A., O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev. 2016. Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167(7):1853-1866.

Donaldson, Z. R., and L. J. Young. 2013. The relative contribution of proximal 5′ flanking sequence and microsatellite variation on brain vasopressin 1a receptor (*Avpr1a*) gene expression and behavior. *PLOS Genetics* 9(8):e1003729.

Donaldson, Z. R., S.-H. Yang, A. W. S. Chan, and L. J. Young. 2009. Production of germline transgenic prairie voles (*Microtus ochrogaster*) using lentiviral vectors. *Biology of Reproduction* 81(6):1189-1195.

Duveau, F., W. Toubiana, and P. J. Wittkopp. 2017. Fitness effects of cis-regulatory variants in the *Saccharomyces cerevisiae* TDH3 promoter. *Molecular Biology and Evolution* 34(11):2908-2912.

Duveau, F., A. Hodgins-Davis, B. P. H. Metzger, B. Yang, S. Tryban, E. A. Walker, T. Lybrook, and P. J. Wittkopp. 2018. Fitness effects of altering gene expression noise in *Saccharomyces cerevisiae*. *eLife* 7:e37272.

Emad, A., J. Cairns, K. R. Kalari, L. Wang, and S. Sinha. 2017. Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance. *Genome Biology* 18(1):153.

Everett, L. J., W. Huang, S. Zhou, M. A. Carbone, R. F. Lyman, G. H. Arya, M. S. Geisz, J. Ma, F. Morgante, G. St. Armour, L. Turlapati, R. R. H. Anholt, and T. F. C. Mackay. 2020. Gene expression networks in the *Drosophila* genetic reference panel. *Genome Research* 30(3):485-496.

Farley, E. K., K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar, and M. S. Levine. 2015. Suboptimization of developmental enhancers. *Science* 350(6258):325-328.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52(2):399-433.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496-512.

Franke, M., D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W.-L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos. 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* 538(7624):265-269.

Fujiwara, T., M. Sanada, T. Kofuji, and K. Akagawa. 2016. Unusual social behavior in HPC-1/syntaxin1A knockout mice is caused by disruption of the oxytocinergic neural system. *Journal of Neurochemistry* 138(1):117-123.

Function. 2000. 2000: A focus on function. *Nature Genetics* 25:243-244.

Graf, U., E. A. Casanova, and P. Cinelli. 2011. The role of the leukemia inhibitory factor (LIF)—pathway in derivation and maintenance of murine pluripotent stem cells. *Genes (Basel)* 2(1):280-297.

Greenway, R., N. Barts, C. Henpita, A. P. Brown, L. A. Rodriguez, C. M. Rodríguez Peña, S. Arndt, G. Y. Lau, M. P. Murphy, L. Wu, D. Lin, J. H. Shaw, J. L. Kelley, and M. Tobler. 2020. Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *bioRxiv* 2020.02.24.959916.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* 550(7675):204-213.

Hammock, E. A. D., and L. J. Young. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* 308(5728):1630-1634.

Helariutta, Y., H. Fukaki, J. Wysocka-Diller, K. Nakajima, J. Jung, G. Sena, M.-T. Hauser, and P. N. Benfey. 2000. The *SHORT-ROOT* gene controls radial patterning of the *Arabidopsis* root through radial signaling. *Cell* 101(5):555-567.

Hodge, R. D., J. A. Miller, M. Novotny, B. E. Kalmbach, J. T. Ting, T. E. Bakken, B. D. Aevermann, E. R. Barkan, M. L. Berkowitz-Cerasano, C. Cobbs, F. Diez-Fuertes, S.-L. Ding, J. McCorrison, N. J. Schork, S. I. Shehata, K. A. Smith, S. M. Sunkin, D. N. Tran, P. Venepally, A. M. Yanny, F. J. Steemers, J. W. Phillips, A. Bernard, C. Koch, R. S. Lasken, R. H. Scheuermann, and E. S. Lein. 2020. Transcriptomic evidence that von economo neurons are regionally specialized extratelencephalic-projecting excitatory neurons. *Nature Communications* 11(1):1172.

Hu, Z., T. B. Sackton, S. V. Edwards, and J. S. Liu. 2019. Bayesian detection of convergent rate changes of conserved noncoding elements on phylogenetic trees. *Molecular Biology and Evolution* 36(5):1086-1100.

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ràmia, A. M. Tarone, L. Turlapati, T. Zichner, D. Zhu, R. F. Lyman, M. M. Magwire, K. Blankenburg, M. A. Carbone, K. Chang, L. L. Ellis, S. Fernandez, Y. Han, G. Highnam, C. E. Hjelmen, J. R. Jack, M. Javaid, J. Jayaseelan, D. Kalra, S. Lee, L. Lewis, M. Munidasa, F. Ongeri, S. Patel, L. Perales, A. Perez, L. Pu, S. M. Rollmann, R. Ruth, N. Saada, C. Warner, A. Williams, Y.-Q. Wu, A. Yamamoto, Y. Zhang, Y. Zhu, R. R. H. Anholt, J. O. Korbel, D. Mittelman, D. M. Muzny, R. A. Gibbs, A. Barbadilla, J. S. Johnston, E. A. Stone, S. Richards, B. Deplancke, and T. F. C. Mackay. 2014. Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. *Genome Research* 24(7):1193-1208.

Huang, W., T. Campbell, M. A. Carbone, W. E. Jones, D. Unselt, R. R. H. Anholt, and T. F. C. Mackay. 2020. Context-dependent genetic architecture of *Drosophila* life span. *PLOS Biology* 18(3):e3000645.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431(7011):931-945.

Janes, D. E., C. Chapus, Y. Gondo, D. F. Clayton, S. Sinha, C. A. Blatti, C. L. Organ, M. K. Fujita, C. N. Balakrishnan, and S. V. Edwards. 2011. Reptiles and mammals have differentially retained long conserved noncoding sequences from the amniote ancestor. *Genome Biology and Evolution* 3:102-113.

Janssens, D. H., S. J. Wu, J. F. Sarthy, M. P. Meers, C. H. Myers, J. M. Olson, K. Ahmad, and S. Henikoff. 2018. Automated in situ chromatin profiling efficiently resolves cell types and gene regulatory programs. *Epigenetics & Chromatin* 11(1):74.

Jean-Baptiste, K., J. L. McFaline-Figueroa, C. M. Alexandre, M. W. Dorrity, L. Saunders, K. L. Bubb, C. Trapnell, S. Fields, C. Queitsch, and J. T. Cuperus. 2019. Dynamics of gene expression in single root cells of *Arabidopsis thaliana*. *The Plant Cell* 31(5):993.

Jin, X., S. K. Simmons, A. X. Guo, A. S. Shetty, M. Ko, L. Nguyen, E. Robinson, P. Oyler, N. Curry, G. Deangeli, S. Lodato, J. Z. Levin, A. Regev, F. Zhang, and P. Arlotta. 2019. *In vivo* Perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *bioRxiv* 791525.

John, A. V., L. L. Sramkoski, E. A. Walker, A. M. Cooley, and P. J. Wittkopp. 2016. Sensitivity of allelic divergence to genomic position: Lessons from the *Drosophila tan* gene. *G3: Genes|Genomes|Genetics* 6(9):2955-2962.

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, J. Baldwin, T. Bloom, D. B. Jaffe, R. Nicol, J. Wilkinson, Broad Institute Genome Sequencing Platform and Whole Genome Assembly Team, E. S. Lander, F. Di Palma, K. Lindblad-Toh, and D. M. Kingsley. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484(7392):55-61.

Kalay, G., and P. J. Wittkopp. 2010. Nomadic enhancers: Tissue-specific *cis*-regulatory elements of *yellow* have divergent genomic positions among *Drosophila* species. *PLOS Genetics* 6(11):e1001222.

Kalay, G., J. Lachowiec, U. Rosas, M. R. Dome, and P. J. Wittkopp. 2019. Redundant and cryptic enhancer activities of the *Drosophila yellow* gene. *Genetics* 212(1):343-360.

Kapheim, K. M., H. Pan, C. Li, S. L. Salzberg, D. Puiu, T. Magoc, H. M. Robertson, M. E. Hudson, A. Venkat, B. J. Fischman, A. Hernandez, M. Yandell, D. Ence, C. Holt, G. D. Yocum, W. P. Kemp, J. Bosch, R. M. Waterhouse, E. M. Zdobnov, E. Stolle, F. B. Kraus, S. Helbing, R. F. A. Moritz, K. M. Glastad, B. G. Hunt, M. A. D. Goodisman, F. Hauser, C. J. P. Grimmelikhuijzen, D. G. Pinheiro, F. M. F. Nunes, M. P. M. Soares, É. D. Tanaka, Z. L. P. Simões, K. Hartfelder, J. D. Evans, S. M. Barribeau, R. M. Johnson, J. H. Massey, B. R. Southey, M. Hasselmann, D. Hamacher, M. Biewer, C. F. Kent, A. Zayed, C. Blatti, S. Sinha, J. S. Johnston, S. J. Hanrahan, S. D. Kocher, J. Wang, G. E. Robinson, and G. Zhang. 2015. Genomic signatures of evolutionary transitions from solitary to group living. *Science* 348(6239):1139-1143.

Kaya-Okur, H. S., S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, and S. Henikoff. 2019. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications* 10(1):1930.

Kelley, J. L., L. Arias-Rodriguez, D. Patacsil Martin, M.-C. Yee, C. D. Bustamante, and M. Tobler. 2016. Mechanisms underlying adaptation to life in hydrogen sulfide–rich environments. *Molecular Biology and Evolution* 33(6):1419-1434.

Kerwin, R. E., and A. L. Sweigart. 2020. Rampant misexpression in a *Mimulus* (monkeyflower) introgression line caused by hybrid sterility, not regulatory divergence. *Molecular Biology and Evolution*. doi: 10.1093/molbev/msaa071.

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188(4184):107-116.

Kocher, S. D., R. Mallarino, B. E. R. Rubin, D. W. Yu, H. E. Hoekstra, and N. E. Pierce. 2018. The genetic basis of a social polymorphism in halictid bees. *Nature Communications* 9(1):4338.

Kolata, G. 2013. Mice fall short as test subjects for some of humans' deadly ills. *The New York Times*, February 11.

Krogh, A. 1929. The progress of physiology. *Science* 70(1809):200-204.

Lamb, A. M., E. A. Walker, and P. J. Wittkopp. 2017. Tools and strategies for scarless allele replacement in *Drosophila* using CRISPR/Cas9. *Fly* 11(1):53-64.

Li, X., A. Battle, K. J. Karczewski, Z. Zappala, D. A. Knowles, K. S. Smith, K. R. Kukurba, E. Wu, N. Simon, and S. B. Montgomery. 2014. Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *American Journal of Human Genetics* 95(3):245-256.

Li, X., Y. Kim, E. K. Tsang, J. R. Davis, F. N. Damani, C. Chiang, G. T. Hess, Z. Zappala, B. J. Strober, A. J. Scott, A. Li, A. Ganna, M. C. Bassik, J. D. Merker, GTEx Consortium, I. M. Hall, A. Battle, and S. B. Montgomery. 2017. The impact of rare variation on gene expression across tissues. *Nature* 550(7675):239-243.

Lim, M. M., Z. Wang, D. E. Olazábal, X. Ren, E. F. Terwilliger, and L. J. Young. 2004. Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene. *Nature* 429(6993):754-757.

Liu, L., L. Yu, and S. V. Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* 10(1):302.

Lupiáñez, D. G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161(5):1012-1025.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess,

K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482(7384):173-178.

Martin, A., and V. Orgogozo. 2013. The loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation. *Evolution* 67(5):1235-1250.

McClintock, B. 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America* 36:344-355.

McManus, C. J., J. D. Coolon, M. O. Duff, J. Eipper-Mains, B. R. Graveley, and P. J. Wittkopp. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Research* 20(6):816-825.

Metzger, B. P. H., and P. J. Wittkopp. 2019. Compensatory *trans*-regulatory alleles minimizing variation in *TDH3* expression are common within *Saccharomyces cerevisiae*. *Evolution Letters* 3(5):448-461.

Metzger, B. P. H., D. C. Yuan, J. D. Gruber, F. Duveau, and P. J. Wittkopp. 2015. Selection on noise constrains variation in a eukaryotic promoter. *Nature* 521(7552):344-347.

Metzger, B. P. H., P. J. Wittkopp, and J. D. Coolon. 2017. Evolutionary dynamics of regulatory changes underlying gene expression divergence among *Saccharomyces* species. *Genome Biology and Evolution* 9(4):843-854.

Morin, M., E. C. Pierce, and R. J. Dutton. 2018. Changes in the genetic requirements for microbial interactions with increasing community complexity. *eLife* 7:e37072.

Munson, M. 2015. Synaptic-vesicle fusion: A need for speed. *Nature Structural & Molecular Biology* 22(7):509-511.

Nakajima, K., G. Sena, T. Nawy, and P. N. Benfey. 2001. Intercellular movement of the putative transcription factor SHR in root patterning. *Nature* 413(6853):307-311.

Naqvi, S., A. K. Godfrey, J. F. Hughes, M. L. Goodheart, R. N. Mitchell, and D. C. Page. 2019. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science* 365(6450):eaaw7317.

Nora, E. P., A. Goloborodko, A.-L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny, and B. G. Bruneau. 2017. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* 169(5):930-944.

Passow, C. N., C. Henpita, J. H. Shaw, C. R. Quackenbush, W. C. Warren, M. Schartl, L. Arias-Rodriguez, J. L. Kelley, and M. Tobler. 2017. The roles of plasticity and evolutionary change in shaping gene expression variation in natural populations of extremophile fish. *Molecular Ecology* 26(22):6384-6399.

Phelps, S. M., and L. J. Young. 2003. Extraordinary diversity in vasopressin (V1a) receptor distributions among wild prairie voles (*Microtus ochrogaster*): Patterns of variation and covariation. *Journal of Comparative Neurology* 466(4):564-576.

Plateaux-Quénu, C., L. Plateaux, and L. Packer. 2000. Population-typical behaviours are retained when eusocial and non-eusocial forms of *Evylaeus albipes* (f.) (Hymenoptera, Halictidae) are reared simultaneously in the laboratory. *Insectes Sociaux* 47(3):263-270.

Rawlik, K., O. Canela-Xandri, and A. Tenesa. 2016. Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome Biology* 17(1):166.

Ricci, W. A., Z. Lu, L. Ji, A. P. Marand, C. L. Ethridge, N. G. Murphy, J. M. Noshay, M. Galli, M. K. Mejía-Guerra, M. Colomé-Tatché, F. Johannes, M. J. Rowley, V. G. Corces, J. Zhai, M. J. Scanlon, E. S. Buckler, A. Gallavotti, N. M. Springer, R. J. Schmitz, and X. Zhang. 2019. Widespread long-range *cis*-regulatory elements in the maize genome. *Nature Plants* 5(12):1237-1249.

Rodgers-Melnick, E., D. L. Vera, H. W. Bass, and E. S. Buckler. 2016. Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences of the United States of America* 113(22):E3177-E3184.

Ryu, K. H., L. Huang, H. M. Kang, and J. Schiefelbein. 2019. Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiology* 179(4):1444-1456.

Sackton, T. B., P. Grayson, A. Cloutier, Z. Hu, J. S. Liu, N. E. Wheeler, P. P. Gardner, J. A. Clarke, A. J. Baker, M. Clamp, and S. V. Edwards. 2019. Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* 364(6435):74-78.

Sakurai, A., and P. S. Katz. 2017. Artificial synaptic rewiring demonstrates that distinct neural circuit configurations underlie homologous behaviors. *Current Biology* 27(12):1721-1734.

Sanjak, J. S., J. Sidorenko, M. R. Robinson, K. R. Thornton, and P. M. Visscher. 2018. Evidence of directional and stabilizing selection in contemporary humans. *Proceedings of the National Academy of Sciences of the United States of America* 115(1):151-156.

Schwarzer, W., N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, N. A. Fonseca, W. Huber, C. H. Haering, L. Mirny, and F. Spitz. 2017. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* 551(7678):51-56.

Seok, J., H. S. Warren, A. G. Cuenca, M. N. Mindrinos, H. V. Baker, W. Xu, D. R. Richards, G. P. McDonald-Smith, H. Gao, L. Hennessy, C. C. Finnerty, C. M. López, S. Honari, E. E. Moore, J. P. Minei, J. Cuschieri, P. E. Bankey, J. L. Johnson, J. Sperry, A. B. Nathens, T. R. Billiar, M. A. West, M. G. Jeschke, M. B. Klein, R. L. Gamelli, N. S. Gibran, B. H. Brownstein, C. Miller-Graziano, S. E. Calvano, P. H. Mason, J. P. Cobb, L. G. Rahme, S. F. Lowry, R. V. Maier, L. L. Moldawer, D. N. Herndon, R. W. Davis, W. Xiao, and R. G. Tompkins. 2013. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America* 110(9):3507-3512.

Shulse, C. N., B. J. Cole, D. Ciobanu, J. Lin, Y. Yoshinaga, M. Gouran, G. M. Turco, Y. Zhu, R. C. O'Malley, S. M. Brady, and D. E. Dickel. 2019. High-throughput single-cell transcriptome profiling of plant cell types. *Cell Reports* 27(7):2241-2247.

Sidorenko, J., I. Kassam, K. E. Kemper, J. Zeng, L. R. Lloyd-Jones, G. W. Montgomery, G. Gibson, A. Metspalu, T. Esko, J. Yang, A. F. McRae, and P. M. Visscher. 2019. The effect of X-linked dosage compensation on complex trait variation. *Nature Communications* 10(1):3009.

Skene, P. J., and S. Henikoff. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* 6:e21856.

Skene, P. J., J. G. Henikoff, and S. Henikoff. 2018. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature Protocols* 13(5):1006-1019.

Smillie, C. S., M. Biton, J. Ordovas-Montanes, K. M. Sullivan, G. Burgin, D. B. Graham, R. H. Herbst, N. Rogel, M. Slyper, J. Waldman, M. Sud, E. Andrews, G. Velonias, A. L. Haber, K. Jagadeesh, S. Vickovic, J. Yao, C. Stevens, D. Dionne, L. T. Nguyen, A.-C. Villani, M. Hofree, E. A. Creasey, H. Huang, O. Rozenblatt-Rosen, J. J. Garber, H. Khalili, A. N. Desch, M. J. Daly, A. N. Ananthakrishnan, A. K. Shalek, R. J. Xavier, and A. Regev. 2019. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 178(3):714-730.

Sosman, J. A., K. B. Kim, L. Schuchter, R. Gonzalez, A. C. Pavlick, J. S. Weber, G. A. McArthur, T. E. Hutson, S. J. Moschos, K. T. Flaherty, P. Hersey, R. Kefford, D. Lawrence, I. Puzanov, K. D. Lewis, R. K. Amaravadi, B. Chmielowski, H. J. Lawrence, Y. Shyr, F. Ye, J. Li, K. B. Nolop, R. J. Lee, A. K. Joe, and A. Ribas. 2012. Survival in BRAF V600–mutant advanced melanoma treated with vemurafenib. *New England Journal of Medicine* 366(8):707-714.

Stoeckius, M., C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* 14(9):865-868.

Stuart, T., A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. 2019. Comprehensive integration of single-cell data. *Cell* 177(7):1888-1902.

Sweigart, A. L., and L. E. Flagel. 2015. Evidence of natural selection acting on a polymorphic hybrid incompatibility locus in *Mimulus*. *Genetics* 199(2):543-554.

Sweigart, A. L., L. Fishman, and J. H. Willis. 2006. A simple genetic incompatibility causes hybrid male sterility in *Mimulus*. *Genetics* 172(4):2465-2479.

Symmons, O., V. V. Uslu, T. Tsujimura, S. Ruf, S. Nassari, W. Schwarzer, L. Ettwiller, and F. Spitz. 2014. Functional and topological characteristics of mammalian regulatory domains. *Genome Research* 24(3):390-400.

Tasic, B., Z. Yao, L. T. Graybuck, K. A. Smith, T. N. Nguyen, D. Bertagnolli, J. Goldy, E. Garren, M. N. Economo, S. Viswanathan, O. Penn, T. Bakken, V. Menon, J. Miller, O. Fong, K. E. Hirokawa, K. Lathia, C. Rimorin, M. Tieu, R. Larsen, T. Casper, E. Barkan, M. Kroll, S. Parry, N. V. Shapovalova, D. Hirschstein, J. Pendergraft, H. A. Sullivan, T. K. Kim, A. Szafer, N. Dee, P. Groblewski, I. Wickersham, A. Cetin, J. A. Harris, B. P. Levi, S. M. Sunkin, L. Madisen, T. L. Daigle, L. Looger, A. Bernard, J. Phillips, E. Lein, M. Hawrylycz, K. Svoboda, A. R. Jones, C. Koch, and H. Zeng. 2018. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* 563(7729):72-78.

Tobler, M., I. Schlupp, K. U. Heubel, R. Riesch, F. J. G. de León, O. Giere, and M. Plath. 2006. Life on the edge: Hydrogen sulfide and the fish communities of a Mexican cave and surrounding waters. *Extremophiles* 10(6):577-585.

Tobler, M., J. L. Kelley, M. Plath, and R. Riesch. 2018. Extreme environments and the origins of biodiversity: Adaptation and speciation in sulphide spring fishes. *Molecular Ecology* 27(4):843-859.

Tsujimura, T., F. A. Klein, K. Langenfeld, J. Glaser, W. Huber, and F. Spitz. 2015. A discrete transition zone organizes the topological and regulatory autonomy of the adjacent TFAP2C and BMP7 genes. *PLOS Genetics* 11(1):e1004897.

Uslu, V. V., M. Petretich, S. Ruf, K. Langenfeld, N. A. Fonseca, J. C. Marioni, and F. Spitz. 2014. Long-range enhancers regulating MYC expression are required for normal facial morphogenesis. *Nature Genetics* 46(7):753-758.

Verta, J.-P., and F. C. Jones. 2019. Predominance of *cis*-regulatory changes in parallel expression divergence of sticklebacks. *eLife* 8:e43785.

Wagle, N., C. Emery, M. F. Berger, M. J. Davis, A. Sawyer, P. Pochanard, S. M. Kehoe, C. M. Johannessen, L. E. Macconaill, W. C. Hahn, M. Meyerson, and L. A. Garraway. 2011. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *Journal of Clinical Oncology* 29(22):3085-3096.

Wang, X., W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan, J. Liu, G. P. Nolan, F.-A. Bava, and K. Deisseroth. 2018. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361(6400):eaat5691.

Wetmore, K. M., M. N. Price, R. J. Waters, J. S. Lamson, J. He, C. A. Hoover, M. J. Blow, J. Bristow, G. Butland, A. P. Arkin, and A. Deutschbauer. 2015. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *mBio* 6(3):e00306-e00315.

Wittkopp, P. J., and G. Kalay. 2012. *Cis*-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13(1):59-69.

Wittkopp, P. J., B. L. Williams, J. E. Selegue, and S. B. Carroll. 2003. *Drosophila* pigmentation evolution: Divergent genotypes underlying convergent phenotypes. *Proceedings of the National Academy of Sciences of the United States of America* 100(4):1808-1813.

Wittkopp, P. J., B. K. Haerum, and A. G. Clark. 2004. Evolutionary changes in *cis* and *trans* gene regulation. *Nature* 430(6995):85-88.

Wittkopp, P. J., B. K. Haerum, and A. G. Clark. 2008. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics* 40(3):346-350.

Wittkopp, P. J., E. E. Stewart, L. L. Arnold, A. H. Neidert, B. K. Haerum, E. M. Thompson, S. Akhras, G. Smith-Winberry, and L. Shefner. 2009. Intraspecific polymorphism to interspecific divergence: Genetics of pigmentation in *Drosophila*. *Science* 326(5952):540-544.

Wolfe, B. E., J. E. Button, M. Santarelli, and R. J. Dutton. 2014. Cheese rind communities provide tractable systems for *in situ* and *in vitro* studies of microbial diversity. *Cell* 158(2):422-433.

Young, R. S., Y. Kumar, W. A. Bickmore, and M. S. Taylor. 2017. Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biology* 18(1):242.

Zhang, T.-Q., Z.-G. Xu, G.-D. Shang, and J.-W. Wang. 2019. A single-cell RNA sequencing profiles the developmental landscape of *Arabidopsis* root. *Molecular Plant* 12(5):648-660.

Zhou, S., T. G. Campbell, E. A. Stone, T. F. C. Mackay, and R. R. H. Anholt. 2012. Phenotypic plasticity of the *Drosophila* transcriptome. *PLOS Genetics* 8(3):e1002593.

Zuellig, M. P., and A. L. Sweigart. 2018. Gene duplicates cause hybrid lethality between sympatric species of *Mimulus. PLOS Genetics* 14(4):e1007130.

Appendix A

Statement of Task

At the request of the National Science Foundation, the National Academies of Sciences, Engineering, and Medicine will appoint an ad hoc planning committee to organize and convene a workshop on research needs to advance the field of functional genomics over the next 10-20 years. The workshop will draw on information and ideas from experts in this new field and others. They will represent, primarily, the biological sciences, with additional input from genomically relevant environmental sciences, bioinformatics, and data sciences, and other pertinent communities. Speakers and attendees will be asked to discuss goals, strategies, and technical needs to allow functional genomics to contribute to the advancement of basic knowledge and its translation into applications that would benefit society (e.g., conservation, evolutionary studies, agriculture, energy, defense, human health, and other sectors). The workshop will delineate the current state of the science, what lessons have been learned thus far, what impediments there are to further progress, and areas where additional investments could help move the field forward. Presentations and discussions may explore the following topics and questions:

- Case studies of success and failure in functional genomics research on a variety of intensively studied model organisms, such as *E. coli* and *C. elegans*. What are the results of these projects? What obstacles did the investigators encounter and could they or could they not surmount them? What tools did the investigators use that produced high-quality results and what tools did they need that were not yet developed and readily available for use?
- Whether there are universal "rules of life" behind resilience, adaptation, and other emerging properties to guide the development of key baseline and comparative questions for research across the realms of microbes, animals, and plants.
- Ideas for short- and medium-term research and knowledge goals and potential strategies, pathways, and needs to achieve these goals.
- Research strategies to examine the interplay of genetic, epigenetic, and environmental factors to determine which factors or combinations of factors may be most influential for determining phenotype.
- Key considerations for selecting experimental systems (model organisms, "nonmodel" organisms, in vitro versus in vivo methods, computational models, etc.) and research approaches (e.g., convergence and research networks) to leverage the full range of disciplines that could contribute to research in future studies of functional genomics.
- The advantages and limitations of available research tools and databases, the potential for using emerging tools and databases that are not yet widely available, and the need for the development and dissemination of these new tools to the research community.
- The training needs for future genotype-to-phenotype research and how to attract the best research talent into the effort.

The workshop presentations and discussions will be documented in a workshop proceedings authored by rapporteurs in accordance with National Academies guidelines.

Appendix B

Workshop Agenda

Next Steps for Functional Genomics: A Workshop
February 10-12, 2020
National Academy of Sciences, Room 120
2101 Constitution Avenue, NW, Washington, DC 20418

DAY 1: MONDAY, FEBRUARY 10

8:45 **Welcome and Opening Remarks**
Steven Moss, National Academies of Sciences, Engineering, and Medicine
Donal Manahan, National Science Foundation
Gene Robinson, University of Illinois at Urbana-Champaign, Chair of the Steering Committee

9:00 **Keynote Speaker**
Introduced by ***Gene Robinson***, University of Illinois at Urbana-Champaign
Title: Design for Inference in Biology
Aviv Regev, Broad Institute and Massachusetts Institute of Technology

10:00-10:15 Coffee Break

10:15 **Talks and Facilitated Panel Discussion: Case Studies on Building Functional Genomics Tools in Diverse Systems**
15-minute talks from each person followed by panel guided questions and discussion
Moderator: ***Lauren O'Connell***, Stanford University
Andrea Sweigart, University of Georgia
Rachel Dutton, University of California, San Diego
Zoe Donaldson, University of Colorado Boulder
Dominique Bergmann, Stanford University
Steven Henikoff, Fred Hutchinson Cancer Research Center

12:00-1:00 Lunch

1:00 **Facilitated Panel Discussion: Determining and Defining "Model" Systems for Diverse Functional Genomics Applications**
10-minute introduction to topic from ***Paul Katz*** *followed by a facilitated discussion*
Moderator: ***Andrea Sweigart***, University of Georgia
Paul Katz, University of Massachusetts Amherst
Lauren O'Connell, Stanford University
Zoe Donaldson, University of Colorado Boulder
Dominique Bergmann, Stanford University

2:15 **Talks and Facilitated Panel Discussion: Understanding the Contributions of the Non-Protein-Coding DNA to Phenotype**
15-minute talks from each person followed by panel-guided questions and discussion
Moderator: ***Terry Magnuson***, University of North Carolina at Chapel Hill
Felicity Jones, Friedrich Miescher Laboratory of the Max Planck Society
Scott Edwards, Harvard University
Francois Spitz, University of Chicago

3:45-4:00 Coffee/Snack Break

4:00 **Facilitated Panel Discussion: Societal and Ethical Implications of Functional Genomics Research**
5 minutes of opening comments by each panelist, followed by a facilitated discussion
Moderator: ***Zoe Donaldson***, University of Colorado Boulder
Rebecca Walker, University of North Carolina at Chapel Hill
Scott Jackson, Bayer Crop Science
Ronald Sandler, Northeastern University

5:00 **Closing Remarks**
Philip Benfey, Duke University

5:00 **Adjourn**

DAY 2: TUESDAY, FEBRUARY 11

9:00 **Keynote Speaker**
Introduced by ***Trudy MacKay***, Clemson University
Title: Predicting Current and Future Sources of Variation in Quantitative Traits
Patricia Wittkopp, University of Michigan

10:00-10:15 Coffee Break

10:15 **Talks and Facilitated Panel Discussion: Advancement of Research on Environmental Regulation of Gene Function**
15-minute talks from each person followed by panel-guided questions and discussion
Moderator: ***Philip Benfey***, Duke University
Sarah Kocher, Princeton University
Joanna Kelley, Washington State University
Nathan Springer, University of Minnesota
Trudy MacKay, Clemson University

12:00-1:00 Lunch

1:00 **Talks and Facilitated Panel Discussion: Challenges and Successes of Integrating Large Datasets**
15-minute talks from each person followed by panel-guided questions and discussion
Moderator: ***Norbert Tavares***, Chan Zuckerberg Initiative
Alexis Battle, Johns Hopkins University
Saurabh Sinha, University of Illinois at Urbana-Champaign
Rahul Satija, New York Genome Center
Charles Danko, Cornell University
Genevieve Haliburton, Chan Zuckerberg Initiative

2:20 **Facilitated Panel Discussion: Pros and Cons of Consortia and Large Databases**
2-3 minutes of opening comments by each panelist, followed by a facilitated discussion
Moderator: ***Charles Danko***, Cornell University
Felicity Jones, Friedrich Miescher Laboratory of the Max Planck Society
Alexis Battle, Johns Hopkins University
Saurabh Sinha, University of Illinois at Urbana-Champaign
Rahul Satija, New York Genome Center
Sean Hanlon, National Cancer Institute, National Institutes of Health

3:15-3:30 Coffee/Snack Break

3:30 **Interactive Session and Panel Discussion: Addressing Challenges in Education and Training**
Groups of 6-8 will discuss prompt questions from 3:15-3:45. Each group will have 3-5 minutes to summarize its main points. We will then move into a panel discussion.
Moderator: ***Patricia Wittkopp***, University of Michigan
Terry Magnuson, University of North Carolina Chapel Hill
Arnaud Martin, The George Washington University
Lauren O'Connell, Stanford University
Grace Anderson, Octant
Rebecca Walker, University of North Carolina at Chapel Hill

5:00 **Closing Remarks**
Emma Farley, University of California, San Diego

5:05 **Adjourn**

DAY 3: WEDNESDAY, FEBRUARY 12

9:00 **Talks and Facilitated Panel Discussion: Interpreting and Validating Results from High-Throughput Screening Approaches**
15-minute talks from each person followed by panel-guided questions and discussion
Moderator: ***Trudy MacKay***, Clemson University
David C. Page, Whitehead Institute
Emma Farley, University of California, San Diego
Philip Benfey, Duke University
Grace Anderson, Octant
Gary Churchill, The Jackson Laboratory

10:45 **Interactive Session on the Future of Functional Genomics Research Goals**
Participants will get into groups of 6-8 to think about the prompt questions related to short- and medium-term research goals of functional genomics. These groups will discuss from 10:40-11:20, followed by 5 minutes of reporting from each group on the main points it came up with. This will end with a larger discussion by all attendees.
Moderator: ***Emma Farley***, University of California, San Diego

11:45 **Future of Functional Genomics Town Hall**
This is a final opportunity for any attendee to emphasize a point they heard during the meeting or bring up something that has not been addressed. Possible prompt questions will be provided.
Moderator: ***Gene Robinson***, University of Illinois at Urbana-Champaign, Chair of the Steering Committee

12:25 **Closing Remarks**
Gene Robinson, University of Illinois at Urbana-Champaign, Chair of the Steering Committee

12:30 **Meeting Adjourn**

Appendix C

Planning Committee Biographies

Gene E. Robinson (NAS, NAM), Ph.D. (*Chair*), is the director of the Carl R. Woese Institute for Genomic Biology. He holds a Swanlund Chair at the University of Illinois at Urbana-Champaign, where he has been since 1989 with a primary appointment in the Department of Entomology. He also holds affiliate appointments in the Department of Cell & Developmental Biology, the Program in Ecology, Evolution and Conservation Biology, and the Beckman Institute of Science and Technology. Dr. Robinson's research group uses genomics and systems biology to study the mechanisms and evolution of social life, using the western honeybee, *Apis mellifera*, as the principal model system along with other species of bees. The research is integrative, involving perspectives from evolutionary biology, behavior, neuroscience, molecular biology, and genomics. The goal is to explain the function and evolution of behavioral mechanisms that integrate the activity of individuals in a society, neural and neuroendocrine mechanisms that regulate behavior within the brain of the individual, and the genes that influence social behavior. Research focuses on division of labor, aggression, and the famous dance language, a system of symbolic communication. Dr. Robinson received his Ph.D. from Cornell University and was a National Science Foundation postdoctoral fellow at The Ohio State University.

Philip N. Benfey (NAS), Ph.D., graduated from the University of Paris and received his Ph.D. in cell and developmental biology from Harvard University under the guidance of Dr. Philip Leder. He did postdoctoral research at The Rockefeller University in the field of plant molecular biology with Dr. Nam-Hai Chua and was appointed assistant professor there in 1990. In 1991, he moved to New York University, where he became an associate professor in 1996 and full professor in 2001. He was the founding director of the Center for Comparative Functional Genomics at New York University. In 2002, he was named professor and chair of the Biology Department at Duke University and in 2003 was named a distinguished professor. Dr. Benfey is the recipient of a Natiional Science Foundation predoctoral fellowship and a Helen Hay Whitney postdoctoral fellowship. He was named a fellow of American Association for the Advancement of Science in 2004 and was elected to the National Academy of Sciences in 2010. In 2011, the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation named Dr. Benfey an investigator under an initiative to support fundamental plant science research. He currently serves on the editorial boards of *Proceedings of the National Academy of Sciences of the United States of America*, *Science*, *Developmental Cell*, and *BMC Plant Biology.* Dr. Benfey is also a pioneer in the cutting-edge technology of plant biology. His lab invented a device called the RootArray, which allows scientists to grow 60 to 120 seedlings at a time. With this device, it also is possible to observe the response of plants and tagged genes. In 2007, he formed a start-up company, GrassRoots Biotechnology, based on this technology that used systems biology approaches to develop new crop traits for the bioenergy, food, and industrial markets. The company was sold in 2013, after which he founded a second company, Hi Fidelity Genetics, which has developed a predictive breeding platform and invented a device, the RootTracker,

which is able to monitor root growth in the field over time. The company is breeding plants with enhanced root systems with the goal of producing resiliant crops in the face of climate change.

Charles Danko, Ph.D., is the Robert N. Noyce Assistant Professor at Cornell University. His primary research interest is to understand how DNA sequence encodes complex programs of gene expression. Dr. Danko uses genetic differences affecting various steps in the RNA polymerase II transcription cycle to understand the molecular basis of phenotypic changes between species. Much of his work is done using molecular and computational tools developed in the Danko lab. Dr. Danko has a longstanding interest in using nascent transcription as a rich source of information about multiple layers of mammalian genome function. He played a vital role in developing assays that map the location of RNA polymerase (PRO-seq, ChRO-seq), and has led to the development of computational tools that leverage this information to identify active functional elements (dREG).

Emma Farley, Ph.D., is an assistant professor at the University of California (UC), San Diego, in the Division on Biological Sciences and School of Medicine. She employs high-throughput functional approaches within developing embryos to decipher how the instructions for successful development are encoded in our genomes. She studies enhancers, which encode these instructions and act as genetic switches to control the timing and location of gene activity. Dr. Farley received a master's degree in biochemistry from Oxford University and a Ph.D. in developmental biology from the MRC London Institute of Medical Science. She worked as a postdoctoral researcher at UC Berkeley and Princeton University, where she exploited the sea squirt *Ciona intestinalis* as a model organism for functional genomics. She developed cost-effective and scalable methods to create and functionally test millions of enhancer variants in every cell of a developing embryo. Her research enabled the first high-throughput dissection of an enhancer within whole developing embryos and revealed the unexpected property that enhancer features and organization must be suboptimized to produce tissue-specific patterns of gene activity. Her lab at UC San Diego continues to investigate how enhancers encode the instructions for successful development and how mistakes in these instructions lead to disease.

Trudy F. C. Mackay (NAS), Ph.D., is the director of Clemson University's Center for Human Genetics located on the campus of the Greenwood (South Carolina) Genetic Center. She is recognized as one of the world's leading authorities on the genetics of complex traits. Dr. Mackay is also the Self Family Chair in Human Genetics and Professor of Genetics and Biochemistry at Clemson University and a member of the National Academy of Sciences (2010). Dr. Mackay received a bachelor of science degree in 1974 and master of science degree in 1976 in biology from Dalhousie University. She completed postgraduate study at the University of Edinburgh with a Ph.D. in genetics awarded in 1979 for research supervised by Alan Robertson. Dr. Mackay's research investigates the environmental and genetic factors that influence quantitative traits. These phenotypic traits include height or weight and are represented by continuous, rather than discrete, values. Her work is undertaken by studying the impact of natural variants and mutations on many behavioral, morphological, physiological, and life history traits in fruit flies, which she uses as a model organism.

Terry Magnuson (NAM), Ph.D., is the Sarah Graham Kenan Professor of Genetics and vice chancellor for research at the University of North Carolina at Chapel Hill. He was the founding chair of the Department of Genetics and the director of the Genome Science Center. He leads the Cancer Genetics Program in the Lineberger Comprehensive Cancer Center. Dr. Magnuson was appointed vice dean for research in the School of Medicine and then vice chancellor for research for the university. Dr. Magnuson served as the chair of the Jackson Laboratory Board of Scientific Overseers, a member of the Board of Directors for the Society for Developmental Biology and for the Genetics Society of America (GSA). He is currently the president of GSA. He was appointed by the National Academies to establish guidelines for human embryonic stem cells. He served as vice chair of an Institute of Medicine (IOM) committee that evaluated the California Institute for Regenerative Medicine, and as a member of the IOM committee reviewing the charge of the National Institutes of Health (NIH) Recombinant DNA Advisory Committee. Dr. Magnuson is a member of the NIH Council of Councils. He has been elected to the American Academy of Arts and Science and to the National Academy of Medicine and is a fellow of the American Association for the Advancement of Science. Dr. Magnuson's research focuses on the genome-wide dynamics of chromatin remodeling complexes. Dr. Magnuson received his Ph.D. from Cornell University and he was a postdoctoral fellow at the University of California, San Francisco.

Lauren O'Connell, Ph.D., is an assistant professor of biology at Stanford University. She received her Ph.D. in cellular and molecular biology from The University of Texas at Austin and her B.S. in biology, neurobiology, and behavior concentration at Cornell University. She holds interests in understanding how animals come up with new ways to face challenges and opportunities in their environment. She believes these evolutionary innovations in physiology and behavior can teach basic organismal biology, evolutionary mechanisms of adaptation, and how flexible organisms are to changing environments. Dr. O'Connell's lab uses amphibians as a model system for understanding the molecular and genomic contributions to biological diversity because they display tremendous variation in behavior and physiology. Members of her lab work on a variety of topics, but most of their work centers on investigating behavior and toxicity in poison frogs and they have worked to develop gene editing technologies in "non-model" amphibians to test these questions in the field and laboratory.

Andrea Sweigart, Ph.D., is an associate professor in the Department of Genetics at the University of Georgia. She received her Ph.D. from Duke University in 2006 with prime focus on genetics and evolutionary biology. Dr. Sweigart believes that a fundamental goal of evolutionary biology is to explain how populations become reproductively isolated species. Her research pursues this goal using *Mimulus* (monkeyflowers), a genus of closely related, ecologically diverse wildflowers that exhibit tremendous variation in reproductive isolation between populations and species. She uses a range of approaches—from field and greenhouse experiments to genetic mapping and bioinformatics—to investigate the genetic mechanisms and evolutionary dynamics of speciation.

Appendix D

Speaker Biographies

Grace Anderson, Ph.D., was originally from San Diego, California, but moved to North Carolina for college and graduated from the University of North Carolina at Greensboro in 2013 with a B.S. in biology (biotechnology concentration) with chemistry and anthropology minors. In 2018, they completed their Ph.D. in Kris Wood's lab at Duke University in the Molecular Cancer Biology program. Their work focused on using functional genomics approaches to uncover novel vulnerabilities in cancers with intrinsic or acquired resistance to anti-cancer therapies. This work resulted in several peer-reviewed publications in top journals (*Science Translational Medicine*, *Nature Communications*, and *Cell Reports*). They were fortunate enough in graduate school to be recognized as an accomplished young scientist and secured many fellowships and awards, most notably, the National Science Foundation Graduate Research Fellowship, the National Cancer Institute Pre- to Postdoctoral Transition Award (F99/K00), the Burroughs Wellcome Fund Graduate Diversity Enrichment Program, the Chancellor's Award for Research Excellence, and others. Following graduate school, they were a postdoctoral fellow at Stanford University in the Genetics Department where they worked closely with another postdoc to understand the genetic liabilities associated with three-dimensional growth in cancer spheroid models. This resulted in middle authorship on a manuscript currently in press at *Nature*. Currently, they are a scientist at a start-up that is mapping the universe of functional interactions between chemicals and human targets by linking biological pathways to digital outputs. The company uses DNA sequencing, gene synthesis, gene editing, and data science to engineer cells, the most sophisticated information processors on Earth, to be a data network. Dr. Anderson attributes much of their success to the mentorship they received in high school, undergraduate, and graduate training. As such, they have a strong passion for mentoring undergraduates, early graduate students, and research technicians.

Alexis Battle, Ph.D., is an associate professor of biomedical engineering at Johns Hopkins University and a 2016 Searle Scholar. She is also a 2020 Microsoft Investigator Fellow. Her research group focuses on understanding the impact of genetic variation on the human body, using machine learning and probabilistic methods to analyze large-scale genomic data. She is interested in applications to personal genomics, genetics of gene expression, and gene networks in disease, leveraging diverse data to infer more comprehensive models of genetic effects on the cell. She earned her Ph.D. in computer science in 2013 from Stanford University, where she also received her bachelor's degree in symbolic systems in 2003. Dr. Battle spent several years in industry as a manager and member of the technical staff at Google, Inc. She joined Johns Hopkins University in July 2014.

Dominique Bergmann (NAS), Ph.D., is a a professor of biology at Stanford University, an investigator of the Howard Hughes Medical Institute, and an adjunct staff member at the Carnegie Institution Department of Plant Biology. Dr. Bergmann's research group uses the development of plant stomata (the epidermal structures that regulate carbon dioxide and water

vapor exchange between the plant and atmosphere) as a model to understand how tissues integrate signals from a variety of sources into decisions about cell fate, cell signaling, and cell polarity. With anchors derived from detailed studies in the genetic reference plant, *Arabidopsis thaliana*, Dr. Bergmann's group identified conserved genetic modules that underlie stomatal cell identities and behaviors in a variety of plant species. Current work focuses on how these modules can be "re-wired" to produce the wide array of patterns seen in nature or engineered to improve plants' capacity to grow in limiting climates.

Gary Churchill, Ph.D., received his Ph.D. in biostatistics in 1988 from the University of Washington in Seattle under the direction of Dr. Elizabeth Thompson. He then worked as a postdoctoral associate with Dr. Michael Waterman in the Department of Mathematics at the University of Southern California in Los Angeles. In 1990, Dr. Churchill joined the faculty at Cornell University in Ithaca, New York, as an assistant and later an associate professor of statistics. In 1997, Dr. Churchill joined The Jackson Laboratory as a visiting investigator; in 1998 he was recruited to the faculty and was promoted to senior staff scientist in 2003. In 2016, he was awarded the Karl Gunnar Johannsen Chair in Computational Biology. Dr. Churchill has played a central role in the establishment of genetics resources including the Collaborative Cross and Diversity Outbred mouse populations. He has served on numerous editorial boards, including co-founding editor of *Statistical Applications in Genetics and Molecular Biology* and senior editor at *Genetics.* Dr. Churchill is the co-director of the JAX Center for Aging Research, and most recently was chosen as a 2019 fellow of the American Association for the Advancement of Science for his contributions to the field of science. His current research employs systems genetics approach to study aging and metabolic disease.

Zoe Donaldson, Ph.D., is an assistant professor of behavioral neuroscience at University of Colorado Boulder. Since arriving at Boulder in September 2016, Dr. Donaldson has focused her research on prairie voles (*Microtus ochrogaster*), small, furry rodents that—unlike mice, rats, and about 97 percent of all mammals—share a unique quality with humans: They tend to be monogamous. She joined the faculty after completing her Ph.D. in neuroscience from Emory University and persuing postdoctoral training at Columbia University. Dr. Donaldson's research is funded by the Whitehall Foundation, the National Science Foundation, and the National Institutes of Health. Her work has been highlighted in *The Economist*, and she routinely works with the National Academy of Sciences' Science & Entertainment Exchange to encourage the accurate representation of science in art and film. Her work with them includes an award for "Most Diabolical" movie murder plot from the Austin Film Festival.

Rachel Dutton, Ph.D., is an assistant professor at the University of California, San Diego. Her lab focuses on the use of microbial communities from cheese as models based on their simplicity, culturability, and experimental tractability. These communities show reproducible and dynamic patterns of community formation that depend on widespread interactions between species. The lab is now developing genetic, cell biological, and chemical approaches to studying species interactions in this model microbial community. As with any model system, their goal is to gain insight into the workings of more complex systems. Dr. Dutton received her Ph.D. from the Harvard Medical School in 2010. She is the recipient of many awards, including the National Institutes of Health's Director's New Innovator Award in 2018 and being named a Pew Scholar in 2017.

Scott V. Edwards (NAS), Ph.D., is the Alexander Agassiz Professor of Zoology and Curator of Ornithology in the Museum of Comparative Zoology at Harvard University. He joined Harvard in December 2003 after serving as faculty for 9 years in the Zoology Department and the Burke Museum at the University of Washington in Seattle. His research focuses on diverse aspects of avian biology, including evolutionary history and biogeography, disease ecology, population genetics, and comparative genomics. He has conducted fieldwork in phylogeography in Australia since 1987 and conducted some of the first phylogeographic analyses based on DNA sequencing. He did a postdoctoral fellowship in immunogenetics at the University of Florida and gained experience with studying the major histocompatibility complex (MHC) of birds, an important gene complex for interactions of birds and infectious diseases, pathogens, and mate choice. His work on the MHC led him to study the large-scale structure of the avian genome and informed his current interest in using comparative genomics to study the genetic basis of phenotypic innovation in birds. In the last 10 years, Dr. Edwards has helped develop novel methods for estimating phylogenetic trees from multi-locus DNA sequence data. His recent work uses comparative genomics in diverse contexts to study macroevolutionary patterns in birds, including the origin of feathers and the evolution of flightlessness.

Genevieve Haliburton, Ph.D., leads the computational biology team at the Chan Zuckerberg Initiative, where the team works to support grant-making and tech-building efforts across many areas including single-cell biology, infectious diseases, and open science. During her Ph.D. and postdoctoral training, she developed computational approaches to characterize transcriptional regulation using multimodal -omics data.

Sean Hanlon, Ph.D., is the acting deputy director of the National Cancer Institute's (NCI's) Center for Strategic Scientific Initiatives (CSSI) where he provides leadership in the planning, developing, and implementing initiatives with a focus on emerging areas of science with potential impact across the cancer research continuum. Dr. Hanlon is also the lead program director for NCI's Provocative Questions Initiative that aims to foster research in understudied areas. Additionally, he provides scientific leadership to collaborative transdisciplinary programs, including the National Institutes of Health's (NIH's) Common Fund's 4D Nucleome program and NCI's Human Tumor Atlas Network. Dr. Hanlon also serves as a representative on NCI, NIH, and inter-agency committees, including Cancer Moonshot Implementation teams, the trans-NCI Data Sharing working group, and the trans-NCI Artificial Intelligence working group. Prior to joining CSSI, Dr. Hanlon was a program director in the NCI Division of Cancer Biology where he managed a portfolio of grants focused on transcriptional and epigenetic regulation in cancer biology and served as director of the Physical Sciences–Oncology Network. He came to NCI in 2009 through the American Association for the Advancement of Science's Science & Technology Policy Fellowship program.

Steven Henikoff (NAS), Ph.D., is an assistant professor in genome sciences at the University of Washington School of Medicine, a member of the Division of Basic Sciences at the Fred Hutchinson Cancer Research Center, and an investigator with the Howard Hughes Medical Institute (HHMI). His field of study is chromatin-related transcriptional regulation. He earned his B.S. in chemistry at the University of Chicago and his Ph.D. in biochemistry and molecular biology from Harvard University in the lab of Matt Meselson in 1977. He did a postdoctoral fellowship at the University of Washington. His research has been funded by the National

Science Foundation, the National Institutes of Health, and HHMI. In 1992, Dr. Henikoff, together with his wife Jorja Henikoff, introduced the BLOSUM substitution matrices. The BLOSUM matrices are widely used for sequence alignment of proteins. In 2005, Dr. Henikoff was elected to the National Academy of Sciences.

Scott Jackson, Ph.D., is the Genetic Pipeline Design lead at Bayer Crop Science. He received his M.S. and Ph.D. from the University of Wisconsin–Madison followed by a fellowship at the University of Minnesota. He has held faculty positions at Purdue University (2001-2011) and the University of Georgia where he was the Georgia Research Alliance Eminent Scholar in Plant Functional Genomics (2011-2019). He is currently an adjunct professor at the University of Georgia. At the University of Georgia, he led the Center for Applied Genetic Technologies and was involved in many campus-wide activities. Dr. Jackson's research has focused on decoding plant genomes to better understand their evolutionary histories and to better engineer crop plants for the future. He led several genome sequencing efforts (e.g., soybean, peanut, and common bean) and has been involved in translating genome sequences into advances in understanding the structure and function of plant genomes with a focus on genome duplications common in plant histories.

Felicity Jones, Ph.D., is a Max Planck Research group leader at the Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany. Her research group studies genome function in natural vertebrate populations adapting to different environments. They leverage the comprehensive genetic and genomic toolkit of the threespine stickleback fish to bridge the gap between the lab and the wild. Using chromatin profiling, comparative epigenomics, transcriptomics, and transgenic approaches, one of their major projects focuses on how the regulatory genome contributes to adaptive evolution and speciation. Supported by the German and European Research Councils, her group also studies features of meiotic recombination and standing genetic variation that facilitate and constrain rates of adaptation in natural populations.

Paul Katz, Ph.D., is a professor of biology at the University of Massachusetts Amherst, where he is also the director of the Initiative on Neurosciences. He was previously a Regents Professor at Georgia State University. He has been studying the brains of nudibranch molluscs for more than 25 years, initially as a research assistant professor at the University of Texas Health Science Center in Houston. He is currently funded by a National Institutes of Health BRAIN award to determine the full brain connectome of the nudibranch *Berghia stephanieae*, which has only 4,000 neurons. His lab is characterizing behaviors and, in collaboration with four other institutions, they are performing large-scale optical recordings from identifiable neurons.

Joanna Kelley, Ph.D., is an associate professor in the School of Biological Sciences at Washington State University (WSU) in Pullman. She runs an evolutionary genomics laboratory that focuses on high-throughput genome sequencing and computational approaches to analyzing big data in genomics. Her research focuses on understanding the genomic basis for adaptation to extreme environments. Dr. Kelley received her B.A. in biology and mathematics, with honors in biology, from Brown University. She earned her Ph.D. in genome sciences from the University of Washington in Seattle. As a graduate student, Dr. Kelley spent a month at the McMurdo Station in Antarctica as part of a National Science Foundation International Graduate Training Course on Polar Biology. She received a National Institutes of Health Ruth L. Kirschstein

National Research Service Award as a postdoctoral researcher at the University of Chicago in human genetics. Dr. Kelley also conducted research as a postdoctoral researcher in the Department of Genetics at Stanford University, where she received the 2012 L'Oreal Fellowship for Women in Science. She is a faculty adviser from the College of Arts and Sciences to the WSU Spokane Genomics Core. She was named one of the GenomeWeb Young Investigators in 2013. In 2016, she was the inaugural recipient of the international Society for Molecular Biology and Evolution Junior Award for Independent Research Excellence. She received an Exceptional Professor Award from the Associated Students of Washington State University in 2018.

Sarah Kocher, Ph.D., is an assistant professor at Princeton University, jointly appointed in the Department of Ecology and Evolutionary Biology and the Lewis-Sigler Institute for Integrative Genomics. She received her Ph.D. in genetics from North Carolina State University, where she studied behavioral ecology and quantitative genetics under the mentorship of Drs. Christina Grozinger and Trudy Mackay. As a postdoctoral fellow with Drs. Naomi Pierce and Hopi Hoekstra in the Department of Organismic and Evolutionary Biology at Harvard, she developed halictid bees as a model system to examine how genetic and environmental factors shape variation in social behavior. Her lab's research bridges the gap between population and evolutionary genetics, neurobiology, and behavior.

Arnaud Martin, Ph.D., is an assistant professor at The George Washington University in the department of biological sciences. Prior to his role as assistant professor he was a postdoctoral researcher at the University of California (UC), Berkeley; Cornell University, and UC Irvine. He got his Ph.D. in 2012 in biological sciences from UC Irvine in the Evolutionary Genetics group with his thesis on "The Developmental Genetics of Color Pattern Evolution in Butterflies." His current research interest lies in the genetic and developmental mechanisms underlying pattern formation and structural coloration in butterfly wings.

David C. Page (NAS, NAM), M.D., is the director and the president of the Whitehead Institute for Biomedical Research, a professor of biology at the Massachusetts Institute of Technology (MIT), and a Howard Hughes Medical Institute investigator. A graduate of Swarthmore College, Dr. Page earned his M.D. from Harvard Medical School and the Harvard-MIT Health Sciences and Technology program. He joined the Whitehead Institute, as the first Whitehead Fellow, in 1984. Dr. Page's laboratory explores fundamental differences between males and females in health and disease, both within and beyond the reproductive tract. The Page lab recently discovered that XY and XX sex chromosomes account for subtle differences in the molecular biology of male and female cells throughout the body.

Aviv Regev (NAS), Ph.D., a computational and systems biologist, is a core member and the chair of the faculty at the Broad Institute, a professor of biology at the Massachusetts Institute of Technology, and a Howard Hughes Medical Institute investigator. Dr. Regev's research centers on combining experimental and computational approaches to decipher how complex molecular circuits function in cells and between cells in tissues. She is the founding director of the Klarman Cell Observatory and Cell Circuits Program at the Broad Institute, and the founding co-chair of the international initiative to build a Human Cell Atlas, whose mission is to create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease. Her lab has been a pioneer of single-cell genomics—inventing key experimental methods and computational

algorithms in the field and demonstrating how to apply it to understand cell taxonomies, histological organization, differentiation, and physiological processes, and how to infer the molecular and cellular circuits that control the function of cells and tissues in health and disease. Among her honors are the National Institutes of Health's Director's Pioneer Award, the Overton Prize, and the Innovator Award from the International Society for Computational Biology, the Earl and Thressa Stadtman Scholar Award from the American Society of Biochemistry and Molecular Biology, and the Paul Marks Prize. She is a fellow of the International Society of Computational Biology and a member of the National Academy of Sciences.

Ronald Sandler, Ph.D., is a professor of philosophy, the chair of the Department of Philosophy and Religion, and the director of the Ethics Institute at Northeastern University. His primary areas of research are environmental ethics, ethics and emerging technologies, and ethical theory. Dr. Sandler is the author of *Environmental Ethics* (Oxford University Press), *Food Ethics* (Routledge), *The Ethics of Species* (Cambridge University Press), and *Character and Environment* (Columbia University Press), as well as the editor or co-editor of *Ethics and Emerging Technologies* (Palgrave), *Environmental Justice and Environmentalism* (MIT Press), and *Environmental Virtue Ethics* (Rowman and Littlefield).

Rahul Satija, Ph.D., is a core member and an assistant investigator at the New York Genome Center (NYGC), with a joint appointment as an assistant professor at the Center for Genomics and Systems Biology at New York University. Dr. Satija's group focuses on developing computational and experimental methods to sequence and interpret the molecular contents of a single cell. His group applies single-cell genomics to understand the causes and consequences of cell-to-cell variation, with a particular focus on immune regulation and early development. His group has developed and maintained the R package Seurat for the analysis, exploration, and integration of single-cell data. Dr. Satija holds a B.S. in biology and music from Duke University and obtained his Ph.D. in statistics from Oxford University as a Rhodes Scholar. Prior to joining NYGC, he was a postdoctoral researcher at the Broad Institute of Harvard and Massachusetts Institute of Technology, where he developed new methods for single-cell analysis.

Saurabh Sinha, Ph.D., is a professor of computer science and the Carl R. Woese Institute of Genomic Biology at the University of Illinois at Urbana-Champaign (UIUC). Dr. Sinha's research is in the broad area of bioinformatics. He develops computational methods that decipher how properties of organisms are encoded in the DNA. His research has specifically focused on the parts of DNA that control the activities of genes—the so-called regulatory DNA. He uses techniques from a variety of disciplines including physics, statistics, and machine learning to model regulatory DNA and predict its biological function. Dr. Sinha has collaborated extensively with biologists to understand the role of regulatory DNA in several important biological processes, including embryonic development, social behavior, and cancer. He has also made important contributions to explaining how regulatory DNA evolves across species. More recently, Dr. Sinha served as the co-director of the National Institutes of Health's BD2K Center of Excellence at UIUC (2014-2018), and led a team of more than 30 researchers and programmers to build an entire suite of computational tools for genomics data analysis. Sinha has been recognized as a university scholar and is a fellow of the American Institute for Medical and Biological Engineering. He is an active member of his research community, is in the core organizing team for numerous conferences and serves on the Board of Directors of the International Society for Computational Biology.

Francois Spitz, Ph.D., is a professor at the University of Chicago and studies the genetic and epigenetic mechanisms that control gene expression during vertebrate development, with a specific emphasis on the interplay between the structural organization of the genome and gene regulation. Genes controlling embryonic development and organ formation are often regulated by genomic elements located far away from them. Dr. Spitz and others have shown that the influence of these distant elements is highly dependent on the folding of the chromosome in specific domains and loops. Dr. Spitz and others have developed a series of in vivo genomic engineering tools to unravel the genetic and molecular underpinnings of the associated processes. These experiments reveal key organizing principles of the architecture of the genome, how they may influence genome function and contribute to human phenotypic diversity and genetic susceptibility to diseases. Dr. Spitz aims to understand how changes in genome organisation (e.g., resulting from chromosomal rearrangements) or mutations of architectural elements may lead to developmental malformations or oncogenic transformations by perturbing normal enhancer-gene communications and identify new ways to correct such molecular dysfunctions.

Nathan Springer, Ph.D., is a McKnight Presidential Endowed Professor in the Department of Plant Biology at the University of Minnesota. He received a Ph.D. from the University of Minnesota in 2000 and his thesis research in Dr. Ron Phillips lab involved cloning of DNA methyltransferases from maize and analysis of anueploid-induced syndromes. Dr. Springer was a postdoctoral research in Shawn Kaeppler's group at the University of Wisconsin–Madison working on functional genomics of maize chromatin. He joined the faculty at the University of Minnesota in 2003. Members of the Springer lab use classical genetic, molecular genetic, and genomic approaches to study natural variation for gene expression and chromatin in maize. The Springer research group has also been involved in research on imprinting, heterosis, and structural genomic variation in maize.

Norbert Tavares, Ph.D., is a microbiologist and a biochemist by training and is a science program manager at the Chan Zuckerberg Initiative, where he manages Single-Cell Biology programs that support the international Human Cell Atlas Consortium. Previously, Dr. Tavares served at the National Cancer Institute, at the National Institutes of Health as an American Association for the Advancement of Science Science and Technology Policy Fellow. In his role, he helped to identify and support emerging and innovative solutions to cancer research problems and managed interdisciplinary trans-institute/agency grant programs. He completed his Ph.D. and postdoctoral work at the University of Georgia, investigating bacterial biosynthesis of coenzyme B_{12} and Sirtuin-dependent posttranslational modification.

Rebecca L. Walker, Ph.D., is a professor of social medicine, a professor of philosophy, and core faculty in the Center for Bioethics at the University of North Carolina at Chapel Hill. She has published widely on topics in animal and human research ethics, medical ethics, and methods of bioethics including virtue ethics. Her co-edited volumes include *Working Virtue: Virtue Ethics and Contemporary Moral Problems* (Oxford University Press, 2007); *Health Inequalities and Justice: New Conversations Across the Disciplines* (UNC Press, 2016) and the two volume series, *The Social Medicine Reader, 3rd edition* (Duke University Press, 2019). She is the co-principal investigator on a National Institutes of Health–funded project, Comparative Model Organism Research Ethics for Healthy Volunteers.

Patricia Wittkopp, Ph.D., is a biologist working at the interface of developmental, evolutionary, and computational biology. She earned her B.S. from the University of Michigan in 1997, studying interactions among genes; her Ph.D. from the University of Wisconsin in 2002, studying the evolution of development; and was a Damon Runyon Cancer Research Foundation postdoctoral fellow at Cornell University studying how gene expression evolves. In 2005, she returned to her alma matter, the University of Michigan, as an assistant professor. Dr. Wittkopp is now the Sally L. Allen and Arthur F. Thurnau Professor of Ecology and Evolutionary Biology as well as Molecular, Cellular, and Developmental Biology at the University of Michigan. She is also affiliated with the Center for Computational Medicine and Bioinformatics, Center for Statistical Genetics, and Program in the Biomedical Sciences. Dr. Wittkopp studies the genetic basis of trait differences within and between species, with an emphasis on the regulation of gene expression.

Appendix E

Acronyms and Abbreviations

ATAC-seq	assay for transposase-accessible chromatin using sequencing
ChIP-seq	chromatin immunoprecipitation sequencing
ChRO-seq	chromatin run-on and sequencing
CITE-seq	cellular indexing of transcriptomes and epitopes by sequencing
ConA	concanavalin A
CRISPR	clustered regularly interspaced short palindromic repeats
CUT&RUN	cleavage under targets and release using nuclease
CUT&Tag	cleavage under targets and tagmentation
CZI	Chan Zuckerberg Initiative
DGRP	*Drosophila Melanogaster* Genetic Reference Panel
EDGE	Enabling Discovery through Genomic Tools program
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait locus
GFP	green fluorescent protein
GO	gene ontology
GTEx	Genotype-Tissue Expression
GWAS	genome-wide association study
GxE	genotype-by-environment
HCA	Human Cell Atlas
Hi-C	high-throughput chromosome conformation capture
HTS	high-throughput screening
ICDA	International Common Disease Alliance
LPS	lipopolysaccharide
MNase	micrococcal nuclease
mRNA	messenger RNA
NIH	National Institutes of Health
NSF	National Science Foundation
pQTL	protein quantitative trait locus
qPCR	quantitative polymerase chain reaction
QTL	quantitative trait locus

RNA-seq	RNA sequencing
sc	single cell
shRNA	short-hairpin RNA
SNP	single nucleotide polymorphism
STARmap	spatially-resolved transcript amplicon readout mapping
TAD	topologically associated domain
UIUC	University of Illinois at Urbana-Champaign
WT	wild-type