6

JUST HOW MESSY IS THE WORLD?

JANELLA K. BAXTER

A VIEW THAT IS GAINING IN POPULARITY in the philosophy of science is that the world is a mess (Waters 2019b; Havstad 2017; McConwell 2017; Dupré 1993; Cartwright 1999). That is, the world that science describes is characterized by many distinct structures. Philosophers of genetics have reached this conclusion by arguing that classical genetics and contemporary molecular genetics are distinct, theoretical, and investigative frameworks that biologists employ for different purposes (Waters 1994; 2004; 2006; Weber 2024). What is remarkable is that despite the thoroughgoing pluralism that these authors embrace regarding classical and molecular genetics, they are nevertheless monistic when it comes to the explanatory and investigative significance of contemporary molecular genetics.

I argue that the pluralism that characterizes molecular genetics is actually more radical than what authors have acknowledged. In fact, the world of genetics is messier in (at least) two ways. One way has to do with the number and relation of gene concepts at work in contemporary molecular biology. While Waters and Weber focus primarily on a conception of the contemporary molecular gene that omits *cis*-regulatory regions, several authors have clarified and defended a number of alternative molecular gene concepts that treat *cis*-regulatory regions as proper parts (Portin 2009; Griffiths and Neumann-Held 1999; Stotz 2004; Griffiths and Stotz 2013; Baetu 2012a; 2012b). I argue further that some genomic databases employ yet another distinct molecular gene concept—what I call the GenBank gene—that individuates regulatory sequences as distinct molecular genes on their own. With a fuller picture of the number of different molecular gene concepts at play in contemporary biology, it becomes apparent that different

gene concepts can overlay and crosscut each other. That is, the same nucleic acid sequence can be classified in a variety of different ways for different purposes. The other way pluralism in contemporary molecular genetics has been mischaracterized has to do with the scope of explanations that appeal to molecular coding genes. I show that the explanatory scope of molecular coding genes can extend beyond the linear sequences of gene products to include observable effects and (sometimes) phenotypic traits.

What my argument shows is that the picture of the world that genetics characterizes is like Cartwright's (1999) idea of a dappled world. The world genetics describes is a patchwork of structures whose boundaries form irregular shapes that can overlay and crosscut each other. Furthermore, the structures that characterize genetics change over scientific history as scientists develop new technologies and practices for managing genomic data. Indeed, the world may be so messy that one might be justified in questioning the usefulness of gene concepts.

1. INTRODUCTION

The world is a mess. Or, at least, this is a position defended by a number of authors in philosophy of science (Waters 2019b; Weber 2023; Havstad 2016; 2017; McConwell 2017; Dupré 1993; Cartwright 1999). This is understood as a metaphysical thesis about the structure of the world. The world that scientists investigate and explain is such that there are a plurality of compatible and distinct frameworks that scientists use to understand it. This is due in part to the way the world is and in part to pragmatic strategies that scientists have for achieving their ends. For example, Joyce Havstad (2016; 2017) has argued that there are multiple justified schemes of protein classification. Proteins have numerous properties and capabilities. Some properties and capabilities are especially useful to scientists for tracking one kind of relation, while others are useful for tracking another kind. For many authors advocating scientific pluralism, pluralism is not a marker of an immature science awaiting replacement by a more mature, fundamental, and unifying theory. Rather, scientific pluralism is "here to stay" (Havstad 2016).

When it comes to the part of the world that genetics describes, just how messy is the world, and what is the nature of the mess? A common view in the philosophy of biology is that there are primarily two distinct, compatible, yet successful frameworks that have characterized the investigative and explanatory pursuits of geneticists since the twentieth century—namely, the

classical and molecular (protein and RNA) coding gene frameworks (Waters 1994; 2004; 2006; Weber 2023). The two frameworks are related not by any theoretical or inter-level reduction but instead by a common investigative and explanatory approach. This approach involves using classical or molecular coding genes as tools for producing differences in life processes for the purposes of (primarily) investigating and explaining biological phenomena. The two differ in terms of conceptual structure. The classical gene concept referred to segments of chromosomes whose internal makeup was not known and that are inherited by future progeny according to a set of (relatively) reliable principles and helped scientists explain phenotypic differences in model organism populations. By contrast, the molecular coding gene refers to nucleic acid sequences that encode information about the linear sequences of RNA and proteins and are used to investigate and explain a different set of life processes. Both continue to be successful frameworks in modern biology. Waters and Weber emphasize that the success of these frameworks lies primarily in their experimental purposes. In fact, they maintain that the explanatory scope—or the phenomena explained by the explanans—of these frameworks is quite modest. Gene-centered explanations formulated by Thomas Hunt Morgan and other classical geneticists were often limited to the model organism populations with which they performed experiments. As for the explanatory scope of molecular coding genes, both authors maintain that they only explain the linear sequences of RNA and proteins.

Despite Waters's and Weber's commitment to scientific pluralism when it comes to genetics, they are surprisingly monistic in their attitudes concerning contemporary molecular genetics. In this chapter, I will argue that contemporary molecular genetics is messier than these authors have acknowledged. I do this by defending two theses. The first thesis is that contemporary molecular genetics itself employs a plurality of gene concepts, one of which—the GenBank gene concept—differs importantly from the molecular coding gene concept that focuses on how nucleic acid sequences determine the linear structure of gene products. In defense of this thesis, I draw from major institutional efforts to annotate, curate, and disseminate genome sequence data to communities of scientists. I argue that the classification of *cis*-regulatory gene sequences as genes in the National Center for Biotechnology Information GenBank database has been an effective method for achieving its aims. The second thesis is that the investigative and explanatory scope of molecular coding genes is much more heterogeneous

than Waters and Weber claim. I argue that differences in molecular coding genes are often employed in experimental conditions to produce distinctive, observable effects. Not only is this of immense investigative significance, but it is also of explanatory significance. Furthermore, differences in molecular coding genes are also of explanatory significance when it comes to some types of phenotypic traits—like Huntington's disease.

What the arguments of this chapter demonstrate is that, indeed, the world that genetics studies is a mess. But it is messier than what some philosophers of genetics have claimed. The picture of reality that is generated from my arguments is more like Nancy Cartwright's (1999) dappled world notion. The world of genetics consists of more than two major conceptual and investigative structures. It is a world characterized by patches of regularity and structure but whose scope and boundaries are not uniform. The scope and boundaries of these patches are likely to change as scientists develop new techniques and tools to interact with the world. Finally, gene concepts can apply to one and the same sequence of nucleic acid bases. The same sequence can be classified under multiple gene concepts. In light of this image of genetics, it is becoming increasingly hard to say anything more general about the structures that characterize this part of reality.

The structure of this chapter is as follows. In section 2, I outline Waters's and Weber's views about scientific pluralism in genetics. Section 3 demonstrates that their views are not pluralist enough when it comes to the number of gene concepts at play in contemporary molecular biology. In section 4, I argue that their views are not general enough when it comes to the explanatory and investigative scope of the molecular coding gene framework. Section 5 concludes with a discussion of the dappled world of genetics.

2. GENE PLURALISM AND LOCAL EXPLANATIONS

Genes have been the focus of much theorizing and investigation in the life sciences. Yet do not be deceived. The ubiquity of gene-focused science should not be taken to indicate that all mention and use of genes in biology appeals to the same concept. Since the mid-twentieth century, biologists have employed primarily two distinct yet useful gene concepts—the classical and the molecular gene (Waters 2004; 2007; Griffiths and Stotz 2013). For some authors, this has been taken as evidence of scientific pluralism—the thesis that there is no single, fundamental, and comprehensive theory for explaining a given scientific domain of inquiry (Waters 2006; Weber 2023). Waters

and (to some extent) Weber have taken this to mean that gene pluralism suggests (at least) two distinct theoretical frameworks for explaining patterns of inheritance and biochemical sequences. Remarkably, notwithstanding their friendly approach to gene pluralism, they have occasionally described contemporary molecular genetics in a surprisingly monistic way.

Since at least the twentieth century, genes have enjoyed a special status in biology. Biologists often conceptualize them as significant difference makers. As the following discussion will reveal, the sense in which they are significant difference makers varies both synchronically and diachronically. That is, at a given point in scientific history (synchronically), there are many ways genes are significant difference makers. Diachronically—or over scientific history—the way genes are significant difference makers can change as well. By "significant difference maker," I mean that a gene or a few genes are singled out as having a causal property that sets them apart from all other relevant causal variables. Genes possess a variety of causal properties that set them apart from all other relevant causes. They can actually (as opposed to potentially) control an outcome of interest, they can determine the fine-grained structure of other biomolecules, they can turn biological processes "on" or "off," and so on. What is relevant for the purposes of this chapter is that the causal property that sets significant difference makers apart from other causal variables is that the property is conceptualized as having a high degree of control over the outcome of interest.[1] Furthermore, the singling out of a significant difference-making variable can also take various forms for different purposes. One way to single out a significant difference-making variable is conceptually. This may be achieved when, say, a model representing the field of all causal variables relevant to some effect of interest is idealized as being fixed or uniform with the exception of a few variables that are imagined as varying. The variables that vary in the model are the significant difference makers. Another way variables are singled out as significant difference makers is experimentally. This is achieved when the field of relevant causal variables are (for the most part) engineered to be fixed or uniform with the exception of the few variables allowed to vary. The two may overlap but needn't always do so. Philosophers of science have done some work to identify the ways in which genes are (and have been) significant difference makers. The purpose of this chapter is to show that contemporary molecular genes are significant difference makers in more ways than have previously been acknowledged. Once the various ways genes are significant difference makers have been parceled out, it will become easier to see what a mess the world of molecular genetics is.

Classical genetics, at least for the first part of the twentieth century in the United States, employed a gene concept that was suited to the investigation and explanation of patterns of inheritance observed in experimental populations.[2] For researchers like Thomas Hunt Morgan, the gene referred to linear units on chromosomes, whose internal structure was unknown (Waters 2004; Kohler 1994). Although the internal structure of the gene was not known, classical genes behave in relatively stable ways according to a handful of principles of inheritance—such as independent assortment, segregation, and recombination (Waters 1994; 2004; 2007). Morgan and his team devised carefully controlled breeding regimens to exploit these principles to generate observable differences in populations of fruit flies that they could attribute to differences in genetics (Bridges and Morgan 1919). Fruit flies with known phenotypic traits, like red eye color, were interbred to generate whole populations that were relatively genetically uniform. This enabled Morgan's group to then breed two distinct genetically uniform populations to generate phenotypic differences. So, for example, a population of red-eye flies might be bred with a population of purple-eye flies. This would generate populations that are heterozygous for the red-eye gene and the purple-eye gene—that is, populations with individuals carrying one copy of the red-eye gene and one copy of the purple-eye gene at the same chromosomal locus. Purple eyes were thought to be a recessive trait, meaning that an individual needs to be homozygous (carrying two copies) for the purple gene to express purple eyes. Another round of interbreeding would generate some individuals who were homozygous for the red gene, some homozygous for the purple gene, and some heterozygous for purple and red genes. Since red eyes were thought to be dominant, only one copy of the red gene is needed for the red eye trait to be expressed. So red eyes were attributed to the presence of red-eye genes, whereas purple eyes were attributed to the presence of two purple-eye genes.

C. Kenneth Waters has emphasized that the success of classical genetics was not its explanatory scope but rather its investigative approach (Waters 2004; 2006; 2010). The explanations Morgan and his team formulated were often modest in scope. When they highlighted the causal significance of classical genes on phenotypic differences, their explanations were partial. Appeals to classical genes only explained the immediate causes of phenotypic differences in experimental populations. Explanations from classical genetics fall incredibly short of all the things one might wish to explain about the

origin and maintenance of a trait. Furthermore, the explanations Morgan's group formulated were local. Because the populations used to infer the genetic causes of phenotypic traits were raised and bred in controlled laboratory settings, there were serious questions about how much one is justified in extrapolating findings from Morgan's experiments to wild-type fly populations. Thus (at least on Waters's account), many of the explanations classical geneticists formulated were relative to the experimental populations in which inheritance patterns were observed. If classical genetics was not successful in arriving at highly general explanations—explanations that hold across a wide range of populations, species, genetic and environmental backgrounds, and so on—then why was it successful? Waters's answer: the set of techniques and practices for carefully generating observable patterns in populations whose causes may be reliably inferred is where the success lies. For the purposes of Morgan's group, intervening on genes was an especially effective way to manipulate and control phenotypic traits. Indeed, the gene-centered approach whereby Morgan's group investigated the genotype's causal relationship on phenotype continues to be an indispensable method for much of contemporary biology.

Since the molecular discovery of DNA and the decoding of the genetic code, modern biology has developed another crucial gene concept—the molecular coding gene. Molecular coding genes are sequences of nucleic acid bases—adenine (A), uracil (T), guanine (G), and cytosine (C)—in DNA that encode information for the sequence of a gene product—be it an RNA or protein.[3] Nearby segments of DNA consist of regulatory modules that bind to transcription factors to control the expression of protein coding sequences. Transcription is the process by which a copy of a molecular coding gene (called a messenger RNA or mRNA) is produced according to Watson-Crick base pair rules. When molecular coding genes determine the linear sequence of proteins, they do so during a process called translation whereby mRNA are "read" in units of triplets (or codons) by protein synthesis molecules called ribosomes. Protein synthesis machinery chains together amino acids in a polypeptide according to the (nearly) universal genetic code that associates nearly each codon with one of the twenty canonical amino acids.[4] The genetic code is nearly universal—aside from a few subtle variations, the same triplet associates with the same amino acid across all living species. For example, the UGG codon always "codes" for the amino acid tryptophan and nothing else.

Much like the classical gene, the molecular coding gene also has difference-making capabilities. Molecular coding genes make fine-grained differences to the amino acid sequences of proteins (Waters 2007; Woodward 2010; Weber 2006; 2013; 2017; Griffiths and Stotz 2013). Any given molecular coding gene can have a large number of alternative nucleic acid sequences. Many of the possible alternative sequences a molecular coding gene can take will make a difference to the amino acid sequence of a protein.[5] There is some redundancy in the genetic code, meaning that more than one codon specifies the same amino acid, as in the case of ACU, ACC, ACA, and ACG all of which specify threonine. Furthermore, three codons function as "stop codons"—codons that do not "code" for any amino acid at all but instead carry the information to the protein-synthesis machinery to stop production. Despite these qualifications, a large number of alternative nucleic acid sequences that a molecular coding gene can take will associate with a unique amino acid sequence in a protein. In this way, molecular coding genes can instantiate a large number of differences in the nucleic acid sequences that constitute them, and many of these differences can make many specific differences to the linear sequence of a protein.

The molecular gene concept is flexible and can be used to account for common genomic processes that occur during gene expression. In many eukaryotic organisms, nucleic acid sequences of DNA alone do not always fully determine the linear sequences of gene products (Griffiths and Stotz 2006; 2013; Stotz 2004; Falk 2010). A variety of biochemical processes often have fine-grained causal control over the linear sequences of gene products. An illustrative example of how non-DNA related biomolecules can have fine-grained causal control over the linear sequences of gene products is alternative splicing. In many eukaryotes, the coding sequences (exons) of many molecular coding genes are not continuous but are interrupted by noncoding sequences (introns). During transcription of the gene into mRNAs, introns are cut out and exons are reassembled to form a final mRNA product (called a mature mRNA). Alternative splicing can rearrange, swap, and even scramble the nucleic acid sequence that results in the final mRNA, which in turn determines the amino acid sequence of the protein that is translated. This is how a single protein coding gene sequence can produce more than one—sometimes many—alternative protein sequences.[6] What determines the arrangements of nucleic acids in an alternatively spliced mRNA are cellular environmental conditions. So long as molecular coding genes are iden-

tified as the collection of nucleic acid sequences in DNA or RNA that determine the linear sequence of other gene products, the molecular gene concept can accommodate alternative splicing (Waters 2007; Weber 2017).[7] An important consequence of processes like alternative splicing is that the number of molecular coding genes in a genome may be greater than the number of genes annotated by sequencing efforts (Stotz 2006; Burian 2004). Since alternative splicing means that the same annotated gene can give rise to a large number of alternative gene products, a number of molecular coding genes overlap.

As in the case of classical genetics, authors have stressed the modest explanatory significance of molecular coding genes. A common view among philosophers has been that at most molecular protein coding genes explain the linear sequence of other (so-called) information-bearing molecules, like DNA or RNA, and proteins. For example, Waters writes: "The significance of gene-based explanations is modest. Genes can be individuated to explain why particular molecules have the linear structure they do, but that by itself does not explain much" (2019a, 97). Marcel Weber (2023), adopting Waters's view, echoes a similar sentiment:

> As Ken Waters has shown, molecular biology provided a basic theory about how DNA as the genetic material is replicated and expressed. However, unlike so-called "fundamental" theories (as they are thought by some to exist in physics), this basic theory is not able nor does it aspire to explain all the phenomena in its domain. It really only explains how DNA molecules can be copied to produce new DNA molecules with the same or complementary nucleotide sequence (including repair mechanisms), how RNA molecules are processed after transcription, how proteins are synthesized, and how these processes are regulated.

Both authors maintain that although genes encoded in DNA can be appealed to in explanations of extremely limited scope, this is nevertheless a crucial element to the success of contemporary molecular genetics. Again, what is significant about protein coding genes is its use as a tool for intervening and manipulating life processes in an effort not to explain but to investigate the bespoke and complex nature of biological systems (Waters 2010). An important difference between the genetic approach employed by modern biologists and that of classical geneticists is that contemporary biologists have

gotten even better at intervening on genes. While classical geneticists had an array of relatively imprecise means of manipulating classical genes—through breeding regimens and crude processes like X-rays—contemporary molecular biologists can now use molecular genes to intervene on other molecular genes. This is best illustrated by the breakthrough gene-editing tool CRISPR-Cas9, whose RNA-guide sequence can be "programmed" to identify almost any nucleic acid sequence in DNA that researchers wish to target and edit with a Cas9 endonuclease (Jinek et al. 2012). Contemporary scientists can determine the precise nucleic acid sequence of a CRISPR-guide sequence, which can in turn aid in introducing precise changes to the nucleic acid sequence of a desired gene in DNA.

Both gene concepts, while related, continue to be useful distinct frameworks.[8] When biologists wish to explain or investigate the inheritance of phenotypic traits that follow regular patterns of independent assortment, segregation, recombination, dominance, and recessiveness, they can employ the classical gene concept. The classical gene concept can be especially useful on the grounds that the exact molecular structure of the gene needn't be known for biologists to utilize it (Griffiths and Stotz 2006; 2013; Waters 1994). By contrast, when biologists wish to explain or manipulate the biochemical processes of life, they can turn to the molecular protein coding gene concept. The older framework is not reducible to the newer one. The two concepts employ different theoretical and practical schemes for explaining and manipulating life processes. What distinct gene concepts show is that there is "no general structure" to genetic explanations and genetic approaches (Waters 2017; also see Griffiths and Stotz 2013). Instead, there is a plurality of frameworks that biologists often switch between fluidly for different purposes.

What is surprising about the authors I have discussed is that despite their thoroughgoing commitment to metaphysical pluralism when it comes to classical and molecular frameworks, their pluralism drops out when they discuss the explanatory and investigative significance of molecular protein coding genes. Of particular interest for my purposes is their exclusion of regulatory regions in DNA from the molecular gene concept. In what follows, I argue that examination of the explanatory and investigative strategies of contemporary molecular biologists unearths an even more radical diversity of structures within the domain of genetics. Not only are there more gene concepts than what has been articulated, but the explanatory scope of molecular genes is more heterogeneous than what has been acknowledged.

3. NOT PLURALIST ENOUGH: THE GENBANK GENE CONCEPT (AND MORE!)

Omitted from the molecular coding gene concept are regulatory regions. *Cis*-regulatory sequences don't have causal control over the linear sequences of gene products but instead control whether and how much molecular coding genes are transcribed. In what follows I trace different ways *cis*-regulatory sequences have been conceptualized in relation to coding sequences by different researchers and for different purposes. In doing so, I show how there have been a plurality of molecular gene concepts at play in contemporary biology. While the determination of linear sequences is an important element to many of the molecular gene concepts I discuss, one concept—notably, the GenBank gene concept—counts *cis*-regulatory sequences as distinct molecular genes. In this way, the GenBank gene concept represents the most notable departure from Waters's molecular coding gene.

Regulatory regions of DNA sequences were (perhaps) first conceptualized as distinct genomic elements by François Jacob and Jacques Monod, whose work in the 1960s characterized the famous lac operon model in prokaryotes (Jacob and Monod 1961a; 1961b; Schaffner 1993; Judson 1996; Keller 2002). The lac operon model is a paradigmatic molecular mechanism consisting of a set of genes (named a, y, z, o, and i) located next to each other in linear fashion along the same chromosomal region (Lac region) (Machamer, Darden, and Craver 2000; Baetu 2012b). In the absence of lactose in the cellular environment, a repressor (encoded by the i gene) binds specifically to the operator (o) gene, which prevents the DNA transcription machinery (RNA polymerase) from producing mRNA transcripts of the z and y genes. The process proceeds until lactose is introduced and binds to the repressor, initiating its release from the operator, thus allowing the transcription of the lactose-metabolizing enzyme β-galactosidase (z) and accompanying genes. Jacob and Monod distinguish between several types of genes, including the operator gene (Jacob and Monad 1961a, 318). The operator gene is a *cis*-regulatory region consisting of a nucleic acid sequence embedded in the same strand of DNA as the molecular coding genes whose transcription it controls. Differences in the nucleic acid sequence of the operator gene do not produce differences in the linear sequence of any gene product. Instead, differences in the sequence of the operator gene can alter whether the repressor protein binds to it, thereby turning transcription of the other molecular coding genes in the lac operon "on" or "off" (Jacob 1977; 1988).

Additional regulatory regions were later discovered in other prokaryotes that display different types of control over transcription, such as the L-arabinose operon, which involves a regulatory region with positive control over transcription of the molecular coding genes when bound to a regulatory protein (Englesberg and Wilcox 1974).

In eukaryotes, *cis*-regulatory regions are more diverse and complex than what the operon model suggests. *Cis* and *trans* in this context refer to where a factor lies in relation to the coding sequences with which it interacts. *Cis*-acting regions are located on the same DNA strand as the molecular coding genes they regulate. While the regulatory regions of prokaryotic operons are commonly located nearby and upstream of the molecular coding sequences they regulate, *cis*-regulatory regions in eukaryotes can be found at distal locations either up- or downstream of the molecular coding genes they control (Wittkopp and Kalay 2012). A nucleic acid sequence that is contained within a molecular coding sequence (as either an intron or exon) may, in some contexts, function as a noncoding regulatory region (Gerstein et al. 2007). It is also common for multiple noncoding regulatory regions to be involved in transcriptional control. Different regulatory regions—promoters, enhancers, silencers, insulators, and so on—bind to different types of biomolecules (called transcription factors). Different combinations of transcriptional factors binding to regulatory regions control not only whether a molecular coding sequence is transcribed but the rate and duration of transcription (Griffiths and Stotz 2013).

The Human Genome Project's annotation of regulatory regions raised conceptual questions to how biologists individuate molecular genes. In its guidelines for gene annotation, the Human Genome Project employed a rather inclusive concept of the molecular gene as "a DNA segment that contributes to phenotype/function. In the absence of a demonstrated function a gene may be characterized by sequence, transcription, or homology" (Wain et al. 2002, 464). This conception counts nucleic acid sequences with a wide variety of functional roles—of which the linear sequence-determining capacity of molecular coding genes is only one—as molecular genes. Since differences in noncoding regulatory regions can produce differences in phenotypic traits, *cis*-regulatory regions count as molecular genes on this conception. The Human Genome Project's inclusive gene concept prompted biologists to reevaluate the molecular gene definition (Baetu 2011). Are *cis*-regulatory regions parts of molecular genes—even when distally located from the coding sequence whose transcription they controlled? There is no uni-

vocal answer. A great variety of gene concepts have been defended in the philosophical and historical literature; in what follows, I mention only a few.[9] Some scientists advocated for the molecular coding gene concept that forms the heart of Waters's and Weber's concept (Gerstein et al. 2007). Others have argued for individuating molecular genes by the collection of nucleic acid sequences (including noncoding sequences)—however distally related to coding sequences—that determine not only the linear sequence of gene products but the expression as well (Singer and Berg 1991; Piatigorsky 2007).[10] On this conception, noncoding regulatory regions are a proper part of a single molecular gene that includes the nucleic acid sequence that determines the linear sequence of other biomolecules. Yet more radically, some authors have defended what Paul Griffiths and Karola Stotz (2006) have described as the postgenomic gene concept. On this conception, molecular genes are not entities but processes involved in regulating transcription, splicing, editing, and translating coding sequences (Portin 2009; Griffiths and Neumann-Held 1999; Stotz 2004; Griffiths and Stotz 2013; Baetu 2012a; 2012b).

While the various gene concepts discussed thus far concern how scientists think about molecular genes in experimental and theoretical contexts, curators of genome databases play a crucial role in individuating and annotating important genomic elements. Major genome databases—such as GenBank, European Molecular Biology Laboratory (EMBL), and the DNA Databank of Japan (DDBJ)—serve as central hubs for the most updated and comprehensive genome information. Individual research labs deposit nucleic acid sequences to genome databases, where the information is checked for redundancy and accuracy. When novel sequences are deposited, database curators process the data in ways that make it portable across scientific, institutional, and database boundaries (Leonelli 2016). GenBank, EMBL, and DDBJ collect and disseminate genomic information for several crucial purposes. One is to make genome information accessible to facilitate research of labs located all around the world (Benson et al. 2007). Another is to inform the structure and content of other more specialized databases (such as Protein Data Bank and TrEMBL) (Gutierrez-Preciado, Peimbert, and Merino 2009). Managing genomic information requires a delicate balancing act between processing data in ways that are informative but not too informative for users (Kanehisa, Fickett, and Goad 1984; Leonelli 2016). Research labs that use genome databases represent very diverse epistemic communities with different methodological approaches

and scientific aims. The success of a genome database depends on how well it makes genomic information accessible to diverse users. Labeling and organizing genomic information is crucial for making entries searchable in a myriad of ways. At the same time, the way curators label and organize data risks biasing future research away from some questions that may deserve attention. Database curators seek to manage this balancing act by employing simple conceptual schemes (Kanehisa, Fickett, and Goad 1984).

Implicit to GenBank's individuation and annotation practices is a simple conceptual scheme that treats *cis*-regulatory regions as distinct molecular genes.[11] On the GenBank gene concept, any nucleic acid sequence that has a confirmed difference-making capacity with respect to some phenotype counts as a distinct molecular gene. As far as the GenBank gene concept is concerned, coding and noncoding sequences are named as "genes." That the GenBank gene concept is distinct from other molecular gene concepts is not lost on maintainers of the database. Instructors running the National Institute of Health's training sessions on how to submit genome sequence data highlight this fact. For example, in Part 1 of "A Submitter's Guide to GenBank," Bonnie L. Maidak states:

> The last set of questions regard the biological meaning of the sequence: and that is what gene does this represent? When I say "gene," sometimes you actually have a sequence not as officially recognized gene, but a genomic region which encompasses a specific genomic marker. You still need to tell us the biological meaning of the sequence even if the sequence that you determine might be an intron of a gene and not the coding region, or it might be just a genomic region but we still need to know why the sequence is important and what the value is of it and what the meaning is of it. ("Webinar" 2014)

GenBank's distinctive gene concept is manifested in the way diverse genome sequences are annotated in the database. For example, the ZPA regulatory sequence (ZRS) has its own entry in GenBank where it has a gene symbol, gene description, gene type, and gene ID number.[12] ZRS is an intron embedded in another molecular coding gene (limb development protein 1 (LMBR1)) that regulates the expression of the sonic hedgehog signaling molecule (SHH) gene (Lettice et al. 2003). Differences in either the nucleic acid sequence or the transcriptional factors that bind to ZRS can associate with significant differences in phenotype like polydactyly in humans (Wu et al. 2016). Of particular concern to the data curators of GenBank are nu-

cleic acid sequences that contribute causally in some way to human health and disease (Sayers et al. 2020). Coding and noncoding sequences alike have this capacity. So, for the purposes of GenBank, a molecular gene is any nucleic acid sequence (coding and noncoding) differences that have a causal effect on a phenotypic trait relevant to biomedicine.

Like other gene concepts, the GenBank gene concept is useful for facilitating investigation and explanation of biological processes. A notable difference between the GenBank gene concept and the molecular coding gene concept is that the former facilitates investigation and explanation of phenotypic differences rather than the linear sequences of gene products. Each entry in the GenBank database provides what Sabina Leonelli (2016) calls a classificatory theory—a concise representation of what the scientific community takes itself to know about the genome sequence rather than novel hypotheses. Each GenBank gene entry includes the full genomic sequence, the sequence's genomic context, gene products (if any), known variants in the population, phylogenetic relationships and organisms carrying orthologues, a brief summary and explanation of the sequence's causal relationship to a biomedically relevant phenotype, alternative names found in the scientific literature, and so on. Also included for each entry is bibliographic information about the sources from which many of the knowledge claims made by the entry come. The GenBank gene concept makes diverse genomic elements searchable within the same database and provides a unifying structure for genomic information that would otherwise be widely dispersed throughout the scientific literature. By making genomic information searchable, the GenBank database helps guide researchers from diverse disciplinary and epistemic cultures to bibliographic and conceptual resources that further their inquiries.

Importantly, the GenBank gene concept—like other gene concepts—is a basic feature of local, partial theories and investigative strategies. Although users of GenBank may explicitly adopt the GenBank gene concept when perusing the database, the concept is implicitly adopted outside the context of database usage to the extent that the GenBank classificatory scheme might organize and direct further inquiry and management of genomic concepts. Furthermore, each GenBank gene will only have causal control over a phenotype for some restricted range of background conditions. Some genes will have causal control over a given phenotype in very restricted genetic and extra-organismal environmental contexts; in other cases, genes will have causal control over a phenotype for a much broader range of background

conditions. Even in cases whether the investigative and explanatory scope of a gene is broad, there will nevertheless be much the gene can't explain. For example, differences in the ZRS gene or differences in the transcription factors that bind to may account for polydactyly in humans; however, such differences still won't be able to fully account for all differences in limb differences in humans. What this shows is that explanations featuring GenBank genes have a varied scope.

The world of contemporary molecular genetics is in fact much more pluralistic than what has been suggested by Waters. Molecular genes are in fact individuated by scientists in a variety of ways and for a myriad of purposes. For the purposes of explaining stark differences in a gene product's phenotype, it may be useful to employ a molecular gene concept that incorporates regulatory sequences and processes. The GenBank gene concept serves to signal to diverse scientific communities a common structure shared by incredibly diverse genomic elements for the purpose of directing further research. While GenBank genes differ in terms of the type of causal property that sets them apart from each other, they share the common feature of being heritable genomic units.

A messier kind of pluralism emerges from this picture of gene concepts in contemporary molecular biology than what Waters and Weber have articulated. For one thing, the number of distinct, irreducible concepts is greater than what both authors have indicated. For another, distinct gene concepts don't simply pick out distinct phenomena. That is, contemporary molecular gene concepts can overlay each other. One and the same nucleic acid sequence can be classified under a variety of gene concepts. For example, the limb development protein 1 molecular gene is simultaneously a protein coding gene and a regulatory gene. The same logic Waters employs to reach the metaphysical conclusion that the world is messy and characterized by a plurality of structures should prompt us to accept that the world is in fact much messier than has previously been appreciated.[13]

4. NOT GENERAL ENOUGH: THE EXPLANATORY SCOPE OF PROTEIN CODING GENES

Another way Waters and (in adopting Waters's view) Weber have mischaracterized the significance of gene concepts has to do with the explanatory scope of protein coding genes. Waters and Weber emphasize that protein coding genes are significant difference makers with respect to other bio-

molecules owing to their fine-grained causal control over the linear sequences of RNA and proteins. I argue that they have understated the explanatory scope of protein coding genes. At least in some cases, protein coding genes can have fine-grained control over phenomena that extend beyond the linear sequences of some biomolecules.

Being a significant difference maker—however "significant" is conceptualized—has explanatory import for Waters and Weber. Difference makers provide answers to *what if things had been different* or counterfactual questions—a property many take to be relevant to explanation. Now, of course, nearly every effect has a great number of difference making variables; however, not all are cited in scientific explanations. Instead, scientists tend to single out some (significant) difference makers as being genuinely explanatory. For this practice to be principled and nonarbitrary, the singled-out difference makers must have some property that sets them apart from all other relevant difference makers.[14] Waters and Weber maintain that the property that sets protein coding genes apart from all other difference makers is fine-grained control over the molecular sequences of other biomolecules (RNA, proteins, etc.). In what follows, I argue the protein coding genes can have this property with respect to phenomena that extend beyond the molecular sequences of biomolecules.

Importantly, Waters and Weber emphasize protein coding genes as tools for manipulating and controlling biological processes. Much of the history of molecular biology has been about making protein coding genes into experimental tools that make observable otherwise unobservable phenomena. Fluorescent proteins are an illustrative example. These proteins are crucial observable technologies employed by biologists in laboratories globally. The molecular protein coding genes for fluorescent proteins used in biology have been modified in various ways. The gene for green fluorescent protein has been modified to have spectral properties that work best with laboratory microscopes (Chalfie et al. 1994). Other genes for fluorescent proteins have been modified to encode proteins that fluoresce new colors (Shen et al. 2017). It is becoming more common for structural biologists to study the three-dimensional structure of proteins by manipulating the nucleic acid sequences of protein coding genes to alter the amino acid sequences of proteins (Spencer and Nowick 2015; Neumann-Staubitz and Neumann 2016). The atoms of some amino acids generate distinctive X-ray scattering and magnetic resonance patterns. Structural biologists take many X-ray images or nuclear magnetic resonance (NMR) readings of proteins whose

linear sequences differ by just a few amino acids (Mitchell and Gronenborn 2017). Comparing X-ray diffraction patterns produced by slightly different proteins helps scientists overcome experimental challenges like the phase problem in X-ray crystallography (Barwich 2017).[15]

So far, my discussion has shown that protein coding genes are useful for engineering novel biomolecules and producing distinctive observable effects, but are they explanatory? Biologists often do single out protein coding genes to explain the effects they produce in experimental situations (Baxter 2019). Differences in the nucleic acid sequences of fluorescent proteins and proteins used in X-ray crystallography or NMR spectroscopy can (at least sometimes) cause differences in an observable effect—like the color of fluorescence, intensity of X-ray scattering or NMR readings. Biologists often rely on observable effects to make inferences about otherwise unobservable phenomena. When biologists are called upon to justify the inferences they make about a target phenomenon, they appeal to differences in the nucleic acid sequences of protein coding genes to explain how the observable effect was produced. At least in experimental contexts, biologists care very much about explaining phenomena that may not have evolved by natural means or even exist outside the laboratory. In this way, the explanations biologists formulate that appeal to the causal significance of protein coding genes is narrow in scope in the populations and situations in which the explanation applies.

Yet protein coding genes are also of explanatory significance when it comes to some phenotypic traits as well. Consider Huntington's disease. Huntington's disease is a neurodegenerative monogenetic disorder caused by specific mutations in a single dominant protein coding gene. The huntingtin mutation consists of an expanded number of unstable CAG (cytosine-adenosine-guanine) repeats. Individuals carrying a single huntingtin gene with about forty or more CAG repeats will have nearly a 100 percent chance of developing the disease in all cases (Ross 2023; Arévalo, Wojcieszek, and Conneally 2001). Moreover, the number of CAG repeats is thought to inversely associate with the age of onset, with higher numbers of repeats causing earlier onsets. When it comes to the occurrence/nonoccurrence of Huntington's disease, it is the presence/absence of the huntingtin mutation that makes the difference. That is, the presence or absence of an expanded number of CAG repeats in the huntingtin gene. In fact, the huntingtin mutation may even be a fine-grained difference maker when it comes to the age of onset. Typically, more than seventy CAG repeats associates with juvenile onset, whereas between forty and seventy repeats associates with

adult onset. What is noteworthy about this case is that the explanatory scope that the huntingtin gene has is broader than the explanatory scope of the protein coding genes previously mentioned. Huntington's disease is a highly penetrant disease—every individual carrying the huntingtin mutation develops the disease. This means that differences in the huntingtin gene account for differences in the occurrence/nonoccurrence of the disease for nearly all human populations. Importantly, explanations of Huntington's disease that appeal to huntingtin mutations are only partial. Such explanations hardly explain everything we might want to know about Huntington's disease—like how the disease originated and how it has been maintained in human populations. Moreover, diseases like Huntington's are very much an exception. Most genetic diseases have multiple—sometimes hundreds—of genetic determinants, each of which has a very low penetrance. Nevertheless, it remains that protein coding genes can (at least sometimes) explain phenotypic traits.

The discussion of this section demonstrates that the explanatory scope of protein coding genes is not uniform, as Waters and Weber have claimed. The metaphysical picture that falls out of their claim that protein coding genes only explain the linear sequences of RNA and proteins is one where the scope of each protein coding gene's explanatory and investigative significance is the same. I have argued in this section that things are not this neat. The explanatory and investigative significance of protein coding genes extends significantly beyond the linear sequences of RNA and proteins. They can be used for a very heterogeneous set of things—ranging from the manipulating the color of fluorescence, X-ray diffraction and NMR reading patterns, and even some phenotypic traits. Importantly, the scope of investigative and explanatory significance that one protein coding gene has will differ from another. For example, the huntingtin gene may have a broader explanatory and investigative significance than, say, a protein coding gene that encodes unnatural amino acids that is used for X-ray crystallography (Liu and Schultz 2010).

5. THE DAPPLED WORLD OF GENETICS—CONCLUSION

The world of genetics is a mess. It is even messier than philosophers have appreciated. Waters has argued that two main investigative and theoretical frameworks have characterized genetics in the past few centuries—the classical and molecular protein coding gene frameworks. Waters and Weber

argue further that these frameworks offer partial, local explanations. In this chapter, I have argued that this view is neither sufficiently pluralist nor sufficiently general to fully capture the mess that constitutes the world that geneticists investigate and explain. It is not sufficiently pluralist because there are many molecular gene concepts at play in addition to the molecular coding gene concept. In addition to the molecular coding gene concept, there are molecular gene concepts that count regulatory gene sequences as a proper part of the gene, while others treat *cis*-regulatory sequences as distinct genes. In contrast to the molecular coding gene concept, these other gene concepts account for different types of phenomena and serve different investigative purposes. It is not sufficiently general, because protein coding genes can (sometimes) explain differences in phenomena that extend beyond the linear sequences of RNA and proteins. Protein coding genes are amenable to direct manipulation by researchers. Researchers continue to get better and better at precisely determining the nucleic acid sequences of protein coding genes. This has made protein coding genes especially useful "handles" by which to manipulate and control life processes for experimental and technological purposes. Beyond the artificial confines of the laboratory, differences in protein coding genes can also account for differences in some phenotypic traits.

Much of the metaphysical view that Waters and Weber advance is left unchallenged by this argument; however, what the arguments of this chapter do is complicate how we should characterize the messy nature of genetics. Explanations that appeal to regulatory or protein coding genes are partial and limited in scope. Genes aren't a fundamental entity of the world but especially useful tools for manipulating and controlling life processes. However, Waters and Weber are drawing (1) too few boundaries and (2) boundaries in the wrong places. They are drawing too few boundaries because they have overlooked the explanatory and investigative significance of alternative molecular gene concepts. There are in fact many more ways to "carve up" the world of contemporary genetics than just in terms of the molecular coding and classical gene concepts. They are also drawing their boundaries in the wrong places when it comes to the investigative and explanatory scope of protein coding genes. They suggest that the explanatory scope extends only to the linear sequences of RNA and protein coding genes. This is mistaken. The explanatory significance of protein coding genes often extends significantly beyond the linear sequences of RNA and protein coding genes. However, there is no simple way to characterize the explana-

tory scope of protein coding genes. Different protein coding genes will have different explanatory scope. This is because the explanatory scope of a gene (protein coding or otherwise) depends on a myriad of factors, such as the investigative interests of scientists at a point in history, the penetrance of a gene, genetic diversity in a population, the investigative techniques and tools that are available to scientists at a time, and more. Notice that this list of factors includes both objective features of the world and epistemic features of scientists. At least some of these factors (epistemic features of scientists) will change over history as new techniques and tools are developed. In turn, this can change the explanatory scope of a given protein coding gene. For example, the explanatory scope of the gene for green fluorescent protein was narrower than it is today before it was developed into an observational tool. This means that the boundaries that demarcate the explanatory and investigative significance will differ for each gene and will likely be irregularly shaped and sized.

The picture of reality that emerges from this discussion is very much in the image of Nancy Cartwright's (1999) dappled world notion. Cartwright's view is aimed specifically at physics and economics, but it is applicable to the area of genetics as well. In her view, the empirical success of our best theories of physics and economics suggest their truth but not their universality—far from it. The empirical successes of physics and economics are extremely limited in scope and confined to very specific experimental conditions—"arranged *just so*" (Cartwright 1999, 2). The dappled world notion describes patches of order and regularity. The patches are not uniform; they take irregular shapes and sizes. Much the same can be said about the world of genetics. In many cases, with enormous labor and thought scientists can arrange living systems in just the right configuration so that genes can exert causal regular control over a host of processes. This is true in experimental situations as well as in the management of genomic databases. It takes a lot of painstaking work to prepare proteins for X-ray crystallography, and it has taken even more labor to develop synthetic technologies that help crystallographers solve the phase problem. Thus, the explanations crystallographers formulate to justify the inferences they make about a protein's structure are restricted to the experimental conditions they engineer. It takes a different kind of labor to facilitate manipulation and control of life processes by managing genomic databases. NCBI's GenBank employs its distinctive molecular gene concept to organize and structure classificatory theories in such a way that facilitates inquiry. Just as the explanations in

protein crystallography are restricted to specific experimental conditions, classificatory theories in a genomic database are often restricted to a limited range of conceptual conditions embedded in the repository. Occasionally, some gene is a significant difference maker with respect to some life process beyond the walls of a laboratory, but this is likely to be an exception to the rule (as with monogenetic diseases). Either way the point is that the explanatory and investigative scope of genes constitute patches of regularity and order. Where the boundaries of each patch lie is likely to differ for different genes (types and tokens). An adequate understanding of where the boundaries lie not only enriches our picture of reality, it also acts as a guide to effective interaction with the world.

NOTES

1. Of course, it is quite common for many—sometimes hundreds of—genes to be singled out as significant difference makers for explanatory purposes (of, say, a phenotypic outcome or disease). In this case, each gene only has a very small degree of causal control over the outcome of interest and, thus, lies beyond the scope of this chapter.

2. The conceptual, theoretical, and investigative strategies of geneticists in other parts of the world at the time were importantly different from classical genetics in the United States. (For examples, see Goldschmidt 1928 and Harwood 1993.)

3. Some molecular coding genes only determine the linear sequences of RNA molecules. Thus, I employ the term "molecular coding gene" to describe cases where a segment of DNA determines the linear sequence of either an RNA or protein molecule.

4. There are a few additional amino acids (i.e., selenocysteine and pyrrolysine) commonly found in living systems on earth. Both selenocysteine and pyrrolysine are encoded by nonstandard means.

5. An average gene of, say, 900 nucleic acid bases will have 4^{900} possible alternatives, since there are four different nucleic acid bases possible for each position in the sequence.

6. Other biochemical processes have this property as well, such as frameshifting during protein translation (Griffiths and Stotz 2013; Falk 2010).

7. Of course, molecular coding genes needn't be identified with only the collection of nucleic acid sequences that make a difference to the linear sequences of gene products. I take this up in the following section.

8. For good analyses of the relationship between classical and molecular genetics, see Vance 1996 and Baetu 2011.

9. A gene concept I don't discuss but that may be yet another way to individuate molecular genes is Lenny Moss's (2003) Gene-P and Gene-D distinction.

10. Piatigorsky calls this the open gene concept. Piatigorsky (2007, 52–3) recognizes that the open gene concept can easily give way to something more like the postgenomic gene concept.

11. This is what Griffiths and Stotz (2006) call a nominal gene.

12. GenBank no longer assigns a gene ID to new entries. This change was due partly to gene ID numbers being redundant information as accession numbers are also assigned to gene entries and partly to the accession number being more "human-readable" (Benson et al. 2016).

13. Of course, there are reasons to be skeptical about Waters's inference from the ontology of a scientific paradigm to metaphysical conclusions about the world beyond paradigms. See Bausman 2023 for a thoughtful exploration of this problem. Personally, I am disinclined to make the kind of inference Waters employs. Rather, I prefer restricting myself to something like perspectival realism—or at least realism relative to scientific paradigms. Nevertheless, the point of this chapter is to argue that Waters's logic should prompt us to conclude something more radical than he admits.

14. The logic of this claim is compatible with there being more than one variable with the relevant explanatory property. In such cases, it would be principled and justified to single out all causal variables possessing the relevant explanatory property. So, in a case where there are multiple variables with fine-grained causal control, it would be appropriate to single out all such variables in explanation.

15. To make inferences about the structure of a protein, structural biologists need to introduce heavy atoms into their specimens. This has commonly been achieved by soaking proteins in a bath of heavy atoms; however, this method is limited by a lack of precision. Synthetic tRNA molecules have been engineered to incorporate amino acids carrying heavy atoms at site specific locations by synthetic biologists (Liu and Schultz 2010).

REFERENCES

Arévalo, J., J. Wojcieszek, and P. M. Conneally. 2001. "Tracing Woody Guthrie and Huntington's Disease." *Seminars in Neurology* 21, no. 2: 209–23.

Baetu, T. M. 2011. "The Referential Convergence of Gene Concepts Based on Classical and Molecular Analyses." *International Studies in the Philosophy of Science* 24, no. 4: 411–27.

Baetu, T. M. 2012a. "Genes after the Human Genome Project." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 191–201.

Baetu, T. M. 2012b. "Genomic Programs as Mechanism Schemas: A Non-Reductionist Interpretation." *British Journal for the Philosophy of Science* 63, no. 3: 649–71.

Barwich, S. A. 2017. "Is Captain Kirk a Natural Blonde? Do X-Ray Crystallographers Dream of Electron Clouds? Comparing Model-Based Inferences in Science with Fiction." In *Thinking about Science, Reflecting on Art,* edited by O. Bueno, G. Darby, S. French, and D. Rickles. New York: Routledge.

Bausman, W. C. 2023. "How to Infer Metaphysics from Scientific Practice as a Biologist Might." In *From Biological Practice to Scientific Metaphysics,* edited by W. C. Bausman, J. K. Baxter, and O. M. Lean. Minneapolis: University of Minnesota Press.

Baxter, J. 2019. "How Biological Technology Should Inform the Causal Selection Debate." *Philosophy, Theory, and Practice in Biology* 11, no. 002: 1–17.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2007. "GenBank." *Nucleic Acid Research* 35: D25–D30.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2016. "GenBank." *Nucleic Acid Research* 45: D37–D42.

Bridges, C. B., and T. H. Morgan 1919. "The Second-Chromosome Group of Mutant Characters." In *Carnegie Institution of Washington Publication* 278: 123–304.

Burian, R. 2004. "Molecular Epigenesis, Molecular Pleiotropy, and Molecular Gene Definitions." *History and Philosophy of the Life Sciences* 26: 59–80.

Cartwright, N. 1999. *The Dappled World: A Study of the Boundaries of Science.* New York: Cambridge University Press.

Chalfie, M., Y. Tu, G. Euskirchen, W. Ward, and D. Prasher. 1994. "Green Fluorescent Protein as a Marker for Gene Expression." *Science* 263: 802–5.

Dupré, J. 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science.* Cambridge, Mass.: Harvard University Press.

Englesberg, E., and G. Wilcox. 1974. "Regulation: Positive Control." *Annual Review of Genetics* 8: 219–42.

Falk, R. 2010. "What Is a Gene? Revisited." *Studies in History and Philosophy of Biological and Biomedical Sciences* 41: 396–406.

Gerstein, M. B., C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. 2007. "What Is a Gene, Post-ENCODE? History and Updated Definition." *Genome Research* 17: 669–81.

Goldschmidt, R. 1928. "The Gene." *The Quarterly Review of Biology* 3, no. 3: 307–24.

Griffiths, P., and E. M. Neumann-Held. 1999. "The Many Faces of the Gene." *Bioscience* 49, no. 8: 656–62.

Griffiths, P., and K. Stotz. 2006. "Genes in the Postgenomic Era." *Theoretical Medicine and Bioethics* 27: 499–521.

Griffiths, P., and K. Stotz. 2013. *Genetics and Philosophy: An Introduction*. New York: Cambridge University Press.

Gutierrez-Preciado, A., M. Peimbert, and E. Merino. 2009. "Genome Sequence Database: Types of Data and Bioinformatic Tools." In *Encyclopedia of Microbiology,* edited by Moselio Schaechter, 211–36. Oxford: Elsevier.

Harwood, J. 1993. *Styles of Scientific Thought: The German Genetics Community 1900–1933*. Chicago: University of Chicago Press.

Havstad, J. C. 2016. "Proteins, Tokens, Types, and Taxa." In *Natural Kinds and Classification in Scientific Practice,* edited by C. Kendig. New York: Routledge.

Havstad, J. C. 2017. "Messy Chemical Kinds." *British Journal for the Philosophy of Science* 69, no. 3: 719–43.

Jacob, F. 1977. "Genetics of the Bacterial Cell." In *Nobel Lectures*, edited by D. Baltimore. New York: Elsevier North Holland.

Jacob, F. 1988. *The Statue Within: An Autobiography*. New York: Basic Books.

Jacob, F., and J. Monod. 1961a. "Genetic Regulatory Mechanisms in the Synthesis of Proteins." *Journal of Molecular Biology* 3: 318–56.

Jacob, F., and J. Monod. 1961b. "On the Regulation of Gene Activity." *Cold Spring Harbor Symposia on Quantitative Biology* 26: 193–211.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. 2012. "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science* 337, no. 6096: 816–21.

Judson, H. F. 1996. *The Eight Day of Creation: Makers of the Revolution in Biology*. New York: Cold Spring Harbor Laboratory Press.

Kanehisa, M., J. W. Fickett, and W. B. Goad. 1984. "A Relational Database System for the Maintenance and Verification of the Los Alamos Sequence Library." *Nucleic Acids Research* 12, no. 1.

Keller, E. F. 2002. *The Century of the Gene*. Cambridge, Mass.: Harvard University Press.

Kohler, R. E. 1994. *Lords of the Fly: Drosophila Genetics and the Experimental Life.* Chicago: University of Chicago Press.

Leonelli, S. 2016. *Data-Centric Biology: A Philosophical Study.* Chicago: University of Chicago Press.

Lettice, A. L., S. J. H. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff. 2003. "A Long-Range Shh Enhancer Regulates Expression in the Developing Limb and Fin and Is Associated with Preaxial Polydactyly." *Human Molecular Genetics* 12, no. 14: 1725–35.

Liu, C. C., and P. G. Schultz. 2010. "Adding New Chemistries to the Genetic Code." *Annual Review of Biochemistry* 79: 413–44.

Machamer, P. K., L. Darden, and C. F. Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67: 1–25.

McConwell, A. K. 2017. "Contingency and Individuality: A Plurality of Evolutionary Individuality Types." *Philosophy of Science* 84, no. 5: 1104–16.

Mitchell, S. D., and A. M. Gronenborn. 2017. "After Fifty Years, Why Are Protein X-Ray Crystallographers Still in Business?" *British Journal for the Philosophy of Science* 68: 703–23.

Moss, L. 2003. *What Genes Can't Do.* Cambridge, Mass.: MIT Press.

Neumann-Staubitz, P., and H. Neumann. 2016. "The Use of Unnatural Amino Acids to Study and Engineer Protein Function." *Current Opinion in Structural Biology* 38.

Piatigorsky, J. 2007. *Gene Sharing and Evolution.* Cambridge, Mass.: Harvard University Press.

Portin, P. 2009. "The Elusive Concept of the Gene." *Hereditas* 146: 112–17.

Ross, L. N. 2023. "Explanation in Contexts of Causal Complexity: Lessons from Psychiatric Genetics." In *From Biological Practice to Scientific Metaphysics*, edited by W. C. Bausman, J. K. Baxter, and O. M. Lean. Minneapolis: University of Minnesota Press.

Sayers, E. W., J. Beck, J. R. Brister, E. E. Bolton, K. Canese, D.C. Comeau, K. Funk, et al. 2020. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 48: D9–D16.

Schaffner, K. F. 1993. *Discovery and Explanation in Biology and Medicine.* Chicago: University of Chicago Press.

Shen, Y., Y. Chen, J. Wu, N. C. Shaner, and R. E. Campbell. 2017. "Engineering of mCherry Variants with Long Stokes Shift, Red-Shifted Fluorescence, and Low Cytotoxicity." *PLOS One* 12, no. 2: 1–14.

Singer, M., and P. Berg. 1991. *Genes and Genomes: A Changing Perspective.* Mill Valley, Calif.: University Science Books.

Spencer, R. K., and J. S. Nowick. 2015. "A Newcomer's Guide to Peptide Crystallography." *Israel Journal of Chemistry* 55: 698–710.

Stotz, K. 2004. "With 'Genes' Like That, Who Needs an Environment? Postgenomic's Argument for the 'Ontogeny of Information.'" *Philosophy of Science* 73, no. 5: 905–17.

Stotz, K. 2006. "Molecular Epigenesis: Distributed Specificity as a Break in the Central Dogma." *History and Philosophy of the Life Sciences* 28: 527–44.

Vance, R. E. 1996. "Heroic Antireductionism and Genetics: A Tale of One Science." *Philosophy of Science* 63: S36–45.

Wain, H. M., E. A. Bruford, R. C. Lovering, M. J. Lush, M. W. Wright, and S. Povey. 2002. "Guidelines for Human Gene Nomenclature." *Genomics* 79, no. 4: 464–70.

Waters, C. K. 1994. "Genes Made Molecular." *Philosophy of Science* 61: 163–85.

Waters, C. K. 2004. "What Was Classical Genetics?" *Studies in History and Philosophy of Science* 35: 783–809.

Waters, C. K. 2006. "A Pluralist Interpretation of Gene-Centered Biology." In *Scientific Pluralism, Minnesota Studies in the Philosophy of Science, Volume XIX*, edited by S. H. Kellert, H. E. Longino, and C. K. Waters. Minneapolis: University of Minnesota Press.

Waters, C. K. 2007. "Causes That Make a Difference." *Journal of Philosophy* 104, no. 11: 551–79.

Waters, C. K. 2010. "Beyond Theoretical Reduction and Layer-Cake Antireduction: How DNA Retooled Genetics and Transformed Biological Practice." In *The Oxford Handbook of Biology*, edited by Michael Ruse. Oxford: Oxford University Press.

Waters, C. K. 2017. "No General Structure." In *Metaphysics and the Philosophy of Science: New Essays,* edited by M. H. Slater and Z. Yudell. New York: Oxford University Press.

Waters, C. K. 2019a. "Ask Not 'What *Is* an Individual?'" In *Individuation, Process, and Scientific Practices*, edited by O. Bueno, R.-L. Chen, and M. B. Fagan. Oxford: Oxford University Press.

Waters, C. K. 2019b. "Presidential Address, PSA 2016: An Epistemology of Scientific Practice." *Philosophy of Science* 86: 585–611.

Weber, M. 2006. "The Central Dogma as a Thesis of Causal Specificity." *History and Philosophy of Life Sciences* 28: 595–610.

Weber, M. 2013. "Causal Selection vs Causal Parity in Biology: Relevant Counterfactuals and Biologically Normal Interventions." In *Causation in Biology and Philosophy*, edited by C. K. Waters, M. Travisano, and J. Woodward. Minneapolis: University of Minnesota Press.

Weber, M. 2017. "Discussion Note: Which Kind of Causal Specificity Matters Biologically?" *Philosophy of Science* 84, no. 3: 574–85.

Weber, M. 2023. "The Reduction of Classical Experimental Embryology to Molecular Developmental Biology: A Tale of Three Sciences." In *From Biological Practice to Scientific Metaphysics*, edited by W. C. Bausman, J. K. Baxter, and O. M. Lean. Minneapolis: University of Minnesota Press.

"Webinar: A Submitter's Guide to GenBank, Part 1: Using BankIt for Small-Scale Nucleotide Sequence Submission." 2014. https://youtu.be/OZxxsRm0pP4.

Wittkopp, P. J., and G. Kalay. 2012. "*Cis*-Regulatory Elements: Molecular Mechanisms and Evolutionary Processes Underlying Divergence." *Nature Reviews Genetics* 13: 59–69.

Woodward, J. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology and Philosophy* 25: 287–318.

Wu, P. F., S. Guo, X. F. Fan, L. L. Fan, J. Y. Jin, J. Y. Tang, and R. Xiang. 2016. "A Novel ZRS Mutation in a Chinese Patient with Preaxial Polydactyly and Triphalangeal Thumb." *Cytogenetic and Genome Research* 149, no. 3: 171–75.