Guttinger, S. and A.C. Love. Forthcoming. modENCODE and the elaboration of functional genomic methodology. In: C. Donohue and A.C. Love (eds.), *Perspectives on the Human Genome Project and Genomics*. Minnesota Studies in Philosophy of Science. Minneapolis: University of Minnesota Press.

modENCODE and the elaboration of functional genomic methodology[†]

Stephan Guttinger

Center for Philosophy of Natural and Social Science, London School of Economics, London, UK.

E-mail: s.m.guettinger@lse.ac.uk

Alan C. Love

Department of Philosophy & Minnesota Center for Philosophy of Science, University of Minnesota, Minneapolis, MN, USA.

E-mail: aclove@umn.edu

1. Functional Genomics: Controversy and Confusion

As the Human Genome Project (HGP) approached completion in the early 2000s, there was a growing interest in moving from structural to functional genomics (Guttinger 2019). This new approach was initially defined as the study of gene function at a global genomic scale (Hieter and Boguski 1997). Later characterizations focused on the attempt to understand how 'genetic information' gives rise to an integrated array of organismal phenotypes (Celniker et al. 2009).[1] This more general concept of genetic information is often cashed out in terms of different kinds of "functional elements" in the genome, such as promoters or enhancers, as well as chromatin states (e.g., euchromatin versus heterochromatin). The main challenge that functional genomics faces is how to reliably identify, characterize, and validate these functional elements.

[†] **Acknowledgments**: We are grateful for the assistance and feedback we have received from a number of different individuals. Brenna Urmanski provided research assistance, looking into unpublished material in the archival database from the History of Genomics Program at the NHGRI. Chris Donohue helped direct us to fruitful framings and needed resources at an early stage and commented helpfully on the manuscript at a later stage. In December 2019, one of us (ACL) was able to interview Elise Feingold and Michael Pazin, the Program Directors for modENCODE at the NHGRI; their patient answers to a variety of historical and philosophical questions were much appreciated and usefully provocative.

[1] "The development and application of global (genome-wide or system-wide) experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics" (Hieter and Boguski, 1997, 601; cf. Fields et al. 1999); "how the information encoded in a genome can produce a complex multicellular organism" (Celniker et al. 2009, 927).

How can researchers know what kind of elements the genome contains ("identify")? How can they discover what these elements contribute to the functioning of the genome and cells ("characterize")? And how can these findings be confirmed ("validate")? We refer to this as the FICV problem (for Functional Identification, Characterization, and Validation).

Within the National Human Genome Research Institute (NHGRI), these challenges were grappled with by establishing a large-scale functional genomics initiative: the ENCyclopedia Of DNA Elements (ENCODE). Touted as a follow-up to the HGP, ENCODE was launched in 2003 and continues to this day, with the findings from phase 3 recently published (e.g., ENCODE Consortium 2020) and phase 4 already under way (ENCODE 4: https://www.genome.gov/Funded-Programs-Projects/ENCODE-Project-ENCyclopedia-Of-DNA-Elements). Although ENCODE was and is a systematic attempt to address the FICV problem and thereby make functional genomics a reality, it also is a site where this problem was transformed in unanticipated ways. However, to uncover these dynamic aspects of functional genomics, we need to look past the most prominent strand of the ENCODE project and controversy that engulfed it in 2012. On the surface, this controversy was about how much of the human genome was defined as 'functional' by ENCODE.[2] On a deeper level, the dispute was about methodology and concepts: critics held that ENCODE was too liberal in the way it

[2] The claim that triggered the controversy—"[t]he vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type" (ENCODE Consortium 2012, 57)—was interpreted by many as showing that *most* of the human genome is functional. This appeared to directly contradict a long-held consensus that a large fraction of the human genome was "junk" DNA. Accompanying news reports and press releases heightened this interpretation, with some referring to ENCODE results as a eulogy for junk DNA (Pennisi 2012). NHGRI Director, Eric Green, claimed in an associated press release: "ENCODE has revealed that most of the human genome is involved in the complex molecular choreography required for converting genetic information into living cells and organisms" (https://www.genome.gov/27549810/2012-release-encode-data-describes-function-of-human-genome). The ENCODE consortium paper's abstract also encouraged this kind of interpretation in claiming that they were able "to assign biochemical functions for 80% of the genome." There are interesting questions about what counts as a "genome" and how this has changed over time, but we set those aside herein (see Keller 2015).

assigned functions to genomic elements because it utilized the wrong concept of function (Doolittle 2013; Graur et al. 2013; see Guttinger and Dupre 2016).[3] This top-level theoretical mistake was seen as having pernicious downstream effects: the methods used to accomplish FICV were too permissive.[4] For the critics of ENCODE, the challenge of getting functional genomics to work—solving the FICV problem—was first and foremost an issue of finding the correct theory of function; this would have led to a different, more appropriate methodology.[5]

Here we directly challenge this assumption about ENCODE and functional genomics more generally. We do so by arguing for three clusters of claims:

(1) Methods do not flow directly from theory. Theories of biological function are insufficient guides to experimental practice. Finding 'the' correct theory of biological function does not provide a trustworthy guide to solve the FICV problem.

(2) Focusing on methodological practices, we discover a complex landscape of proxy-measures being used in functional genomics. There is a toolkit of proxies that are applied to assess how genomes work. These proxies are dynamic and cannot be classified neatly along the lines of different functional concepts.

(3) These proxies have a life of their own, recombining in novel and unexpected ways. This expansion and reassortment of the proxy toolkit in experimental practice is

[3] "[ENCODE] authors focus on reconciling "the strengths and limitations of biochemical, evolutionary, and genetic approaches for defining functional DNA segments" but avoid dealing with the central conceptual issue, which is the problematic nature of "function" itself" (Doolittle et al. 2018, 1236).

[4] "[ENCODE] researchers have adopted an extremely liberal criterion for ascribing causal role functions to genetic elements […] such permissive criteria will identify a great many genetic elements regardless of whether they have been under selective pressure or contribute to any meaningful organism-level capacities" (Elliott et al. 2014, 16). Some critics went further: "ENCODE not only uses the wrong concept of functionality, it uses it wrongly and inconsistently" (Graur et al. 2013, 580).

[5] "The main advantage of the selected-effect function definition is that it suggests a clear and conservative method of inference for function in DNA sequences" (Graur et al. 2013, 579).

observable within ENCODE and gives birth to new forms of data and theoretical tensions that have largely been overlooked in the ENCODE controversy.

In short, our analysis shows that the ENCODE controversy has, in many ways, missed the mark. To understand ENCODE methodology, its output, and what functional genomics could or should be, we need to focus on the diverse and dynamic landscape of these proxy measures rather than conceptual disputes. This becomes increasingly clear once we move away from the prominent ENCODE papers published in 2012 and look to the sprawling project as a whole, especially efforts devoted to genomic comparisons across taxa, such as 'model organism ENCODE' (modENCODE). There we see a novel elaboration of functional genomic methodology and its empirical consequences. Framing the debate as a 'controversy' about percentages and concepts distracts attention away from important methodological questions about functional genomics.[6]

We begin by analyzing flaws in the *from-theory-to-practice* assumption that dominates the debate about functional analysis in biology (and therefore the ENCODE controversy). This shows that the two main accounts of function in biology—causal role (CR) and selected effects (SE)—are insufficient guides for experimental practice oriented toward solving the FICV problem in genomics. Then we turn our attention to three proxy strategies that dominate functional genomics: genetic, evolutionary, and biochemical. These form the core of a dynamic methodological toolkit that researchers mobilize; the assays used in these strategies can be mixed and matched creatively, thereby leading to new empirical patterns as well as theoretical tensions. This is precisely what we see in modENCODE, which ran from 2007 to 2014 (with funding ending in 2012). Our analysis of this project strand indicates that the FICV problem was

[6] The ENCODE controversy rumbles on to this day. There is still active debate about what 'the' correct account of function should be in genomics (see, e.g., Doolittle 2018; Brzović and Šustar 2020; Linquist et al. 2020).

transformed in novel ways, eventually resulting in a combination of evolutionary and biochemical strategies that ENCODE (or modENCODE) researchers could not have anticipated. This entailed a shift away from the analysis of structures (i.e., sequence-based elements) to an analysis of general regulatory principles of the genome (i.e., abstract functional rules), something that FICV of the human genome alone would not have uncovered. However, the methodological maneuver and resulting empirical findings gave rise to a theoretical tension: the notion of conservation or homology that applies straightforwardly to sequence-based functional elements (i.e., material traits) applies less clearly to quantitative functional relationships or associations that obtain between diverse cellular processes—control principles or rules of the entire genome that underlie its functioning.

We conclude that the dynamic proxy landscape is not only interesting in its own right but illustrates the diversity of analytic approaches to FICV pursued within ENCODE as a whole. This yields a more balanced appreciation of the power and limitations of functional genomics and moves us beyond the ENCODE function controversy. Instead of fixating on a purported conceptual error, our analysis of modENCODE practices demonstrates how the methodology of functional genomics was elaborated creatively and discovered something unexpected—not just a catalogue of what and where the functional elements are, but principles for what counts as genome functionality more systemically.

2. Functional Analysis and FICV in Genomics

Functional genomics faces both theoretical problems and practical issues. Theoretical problems arise because there are different ways of defining biological function. Practical issues arise because there are different ways of identifying, characterizing, and validating functional elements in a complex system. As might be expected, the two issues are connected: different practices can presuppose different definitions of function, and different definitions of function can be operationalized distinctly for FICV. In debates about functional genomics, and the ENCODE controversy in particular, many assume that theory precedes practice: an account of 'function' gives researchers a clear methodological guide for functional analysis. Solving the theoretical problem supposedly secures resources for addressing the practical issues. Get the theory wrong, and the methodology will go astray also. Here we discuss two definitional orientations for functional analysis in biology and their putative methodological implications. Contrary to the assumption, theory is an insufficient guide for FICV in genomics.

2.1. Functional Analysis: Different Concepts and their Methodological Prescriptions

On the causal-role (CR) definition of function, a functional analysis aims to identify the operational parts of a complex system and the roles they play within that system to bring about some outcome, typically a system capacity, such as the heart being able to pump blood (Cummins 1975). This yields three distinct procedural steps of identification, characterization, and validation: (1) identify the working parts; (2) characterize their operating behaviors; and, (3) validate their contributing roles to the complex system. Overall, the methodological strategy is reductionist: decompose the complex system into parts and localize the roles to one or more parts that contribute to the functional outcome (Bechtel and Richardson 1993). It is assumed that this decomposition and localization strategy will yield simpler and more tractable items of study compared with the complex system itself. This encourages further decomposition and localization of system parts to identify and individuate simpler and more tractable items (where possible and fruitful). Importantly, functional analysis, according to the CR definition of function, focuses on parts, activities, and systems in the present and largely neglects the histories of these different items.

The CR definition of function appears to provide methodological directives. In fact, it seems primed to solve the FICV problem. To begin, the investigator has to identify a property or outcome of the system they want to understand (e.g., kidney filtration). Then the system has to be decomposed into different kinds of working parts (renal lobes, nephrons, renal tubules, etc.). An aspect of this decomposition involves detailing how the parts operate. Finally, investigators must determine the function of any of these parts in relation to the property of interest and do so in a validated manner. Two common experimental interventions to achieve this are blocking the operation of the part or removing it altogether and then checking whether (and how) this intervention affects the system behavior (e.g., decreasing waste filtration). In some instances, the part can be isolated and put into a different system to test if it serves the particular role of interest in a more generic context.

Within biology, there is an important alternative concept of function: the 'selected effects' (SE) account (Wright 1973; Neander 1991). As the name implies, this definitional orientation assigns a function to a trait, entity, or process only if it made a positive contribution to an organism's fitness in the past (i.e., on the basis of a "selected effect"). Contrary to the CR definition, this approach is historical in that present contributions of parts to a system behavior are not genuine functions unless there is a history of selection for the contribution that differentially affected fitness. Rather than asking what a part is doing presently in a system, the SE definition is concerned with why a part is present in the system (Millikan 1989).[7]

[7] There is a rich and complex literature on SE functions that we cannot review here that includes 'forward-looking' accounts (see Garson 2016 for a critical overview).

This difference in the question asked also has methodological implications. Answering the question of why a part is present in a system requires distinctive forms of historical evidence. Importantly, experimental interventions, such as those used for the CR account, may not speak to why the part is present in the system. The current behavior and contribution of working parts may not be the same as it was in the past; a part may have originally come to be in the system for a different reason than why it is maintained presently. Evolutionary processes involve shifts in the function of parts, sometimes as a consequence of being co-opted from one context into another (Piatigorsky 2007). Tracing history becomes important and with it the need to restrict the possibility space. Using reverse engineering, one can start with a complex functional trait and ask what environmental problem would be solved by it. For example, vertebrate kidneys solve the problem of blood filtration and water reabsorption. The claim of their suitedness depends on a close study of how kidney structure and physiological operation accomplish this task distinctively—kidneys appear "designed" by evolution to accomplish this function. Additionally, comparative methods look at the same trait in different taxa and assess whether it is tightly correlated with a particular environment. The repeated evolution of ecomorphs with functional traits distinctively suited to a specific habitat from a common ancestor is another form of historical evidence for SE effect functions (Stroud and Losos 2016).

2.2. Functional Analysis: Difficulties When Applied to Genomes

From what we have described, it might appear that an assumption of *from-theory-to-practice* to address FICV is justified. There is a connection between accounts of biological function and methods of functional analysis. However, a connection is not the same as a clear methodological directive. For example, even though a CR account encourages researchers to identify working parts, it provides little to no guidance in how to do so, which is poignant in the context of genomics where it is not always clear what 'the' parts should be (i.e., identification). This question has at least three components. First, it is not always clear *where* the parts are (*location*). Although genomes contain genes, promoters, and other elements, it is not so easy to decompose the sequence into parts like we would for a human-made machine (e.g., a bicycle). This affects how FICV is approached: which sequences should the investigator concentrate on? What should be removed or isolated for functional analysis? Second, even though researchers have a growing list of different *types* of elements to look for, they do not know whether this list is exhaustive (*completeness*). There could be unknown genomic elements that are biologically relevant; researchers not only have to find the different tokens of known types, but also identify, characterize, and validate novel types of elements. Third, definitions of known types of functional elements are often vague, and new findings force researchers to re-think their categories, such as what counts as a gene or the boundaries of a promoter region (*vagueness*). Large-scale genomics projects have played a role in this process. The concept of a 'gene'—arguably the most prominent and best-characterized type of functional element—has undergone significant changes that were triggered in part by insights arising from the HGP and ENCODE (see, e.g., Keller 2000; Gerstein et al. 2007). Something similar applies to how findings from ENCODE have challenged existing definitions of genome function more generally (Kellis et al. 2014; see below, Section 3).

The issues of location, completeness, and vagueness illustrate that functional genomics is not simply an inventory project. It is also a global discovery process that aims to eventually annotate what is not yet known. And the CR definition alone does not provide sufficient methodological resources for tackling this task. Arguably, one could address some of these difficulties by brute force, going through the genome one "step" at a time. However, testing each nucleotide or all combinations of nucleotides systematically with genetic assays was not only largely impossible in the early 2000s at the birth of functional genomics but would be an extremely time-consuming and expensive task even today. Apart from practical worries, there is a deeper issue connected to the vagueness of definitions for functional elements. This vagueness is not just due to a lack of data but part of the nature of the genomic system. There is every indication that genomic elements, both in terms of their boundaries and their activity or effects, are highly context dependent (Stamatoyannopoulos 2012). Deleting a DNA sequence might not give the researcher any indication of the role played by that functional element because living systems often contain redundant elements and pathways; another part of the system can compensate for the lost element. Alternatively, an experimental approach that relies on putting the sequence of interest into a test assay can produce false positives and negatives because it might not behave as it would in its native context. These issues indicate that the CR account of function alone is not sufficient to provide clear methodological answers to the FICV problem.[8]

The SE definition of function can provide different and helpful resources, but it cannot solve the problem either (*pace* Graur et al. 2013). As noted, there is an inherent difficulty with evidential ambiguity, and this has at least three components beyond the gap already described between present functional operation and past functional origination (*past-present ambiguity*). First, many traits are multi-functional (*functional ambiguity*), which compounds the origination problem. Vertebrate kidneys may solve the problem of blood filtration and water reabsorption, but they also secrete hormones and regulate blood pressure. Deciding what evolution originally "designed" kidneys or genes for (if only one thing) is difficult (if not impossible). Comparative

[8] This problematizes extant accounts of genomic parthood based on the CR definition (e.g., Kaiser 2018).

methods can help restrict the space of possibilities, but not often in a way that singles out only one function for an element, genomic or otherwise. Second, not only can a function change over time but it can be lost. Thus, a part can be present but no longer exhibit any behavior that can be characterized and validated (*functional absence*), such as inactive transposable elements in the form of short and long interspersed nuclear elements. Third, a part may exhibit a behavior that is a SE effect function with respect to a different evolutionary lineage. Active transposable elements have been selected to replicate and intersperse copies of themselves in a genomic environment but not to the fitness benefit of the host organism (*functional parasitism*). To identify, characterize, and validate a genomic element of this type would not help answer "how the information encoded in a genome can produce a complex multicellular organism" (Celniker et al. 2009, 927). Additionally, the most prominently used strategy for detecting functional elements being maintained by natural selection—comparative sequence analysis to identify conservation—is beset by these same issues (see below, Section 3.2). Using the SE definition of function alone does not solve the core methodological problem of FICV in genomics.

FICV in genomics thus remains a challenge no matter what account of function is adopted. There is not a clear line that leads from theory to practice despite their interconnections. Arguments over which account of function to apply are generally important but specifically irrelevant for solving the methodological problem of functional genomics. This suggests we should redirect our attention from theoretical discussions surrounding the ENCODE controversy to scrutinize how FICV is dealt with in practice, successfully or not, using a complex toolkit of proxy measures that form the core of functional analysis in genomics, especially for ENCODE.

3. Functional Analysis: The Need for Proxies

3.1. Characterizing Proxy Approaches

Although there will be methodological problems for functional genomics regardless of what concept of function is used, this insufficiency points toward an important insight: function, regardless of its definition, is never directly measured. Instead, addressing FICV involves the extensive use of proxies: measurements of particular properties that stand in for or provide evidence of function. Proxies are inherently fallible and defeasible, by definition, as well as complicated in how they are situated within scientific reasoning.

Consider the difficulties faced by a CR definition of function: location, completeness, and vagueness (Section 2.2). Genomic analyses use multiple proxies for annotating putative functional elements, such as start codons for open reading frames or TATA boxes for promoter regions. One of the most common is a consensus sequence (e.g., ATG for start codon). However, these signature sequences are formulated in part through comparative analysis: the TATA box is a consensus sequence that is derived from comparing promoter regions of different genes within and across different organisms (TATAXAXY, where X can be either A or T, and Y is either A or G; Basehoar et al. 2004). This consensus signature only gives an indication of where a promoter could be located based on those independently identified in some subset of taxa. Even with a consensus sequence in hand, the identification process is not error-free. It does not provide a precise template for identification because particular promoter regions that contain TATA boxes will diverge in nucleotide sequence and many promoters do not have core elements (Roy and Singer 2015). Consensus sequence proxies are therefore fallible with respect to identifying something (*location*) and defeasible with respect to tracking boundaries (*vagueness*). And, since a proxy cannot confirm the existence of something not previously identified, at best sequence patterns that deviate from chance serve as a starting point (*completeness*).

Now consider the difficulties faced by a SE definition of function. Addressing past-present ambiguity requires specific kinds of comparative information about sequence conservation; alone it is only a rough proxy for function. Functional ambiguity is sometimes invisible until validation stages of analysis when pleiotropic effects of genomic elements can be detected. The absence of function is difficult to detect (absence of evidence is not evidence of absence), and doubly so given that conservation is not required for functionality. Activity on its own does not signal a function with some fitness effect on the organism because that activity might have fitness effects at another level of organization or for a different lineage (e.g., active transposable elements). In this respect, there is a parallel between functional parasitism from a SE theoretical orientation and noise from a CR theoretical orientation (i.e., working parts with characteristic behaviors but not contributing to a designated system outcome). Again, further characterization and validation studies are needed in each situation and this work typically cannot be done in a high-throughput fashion for the entire genome. Instead, a more piecemeal strategy is required, but this is precisely the opposite of what ENCODE was attempting (i.e., a global, genome-wide *encyclopedia* of functional DNA elements).

Although ENCODE researchers never directly addressed (at least in writing) whether they adopted a particular definition of function, they were quite concerned with these methodological challenges—what their proxies did and did not tell them—in relation to FICV. Researchers explicitly recognized three main proxy strategies of hunting for functional elements in the genome, none of them sufficient on their own (Kellis et al. 2014).

a) <u>Genetic</u>: identify a contribution of an individual nucleotide or group of nucleotides (i.e., genomic elements) through mutational analysis or assays and treat resulting effects as a proxy for (CR) functional relevance

b) <u>Evolutionary</u>: identify sequence conservation of genomic elements through cross-species comparison and treat conservation as a proxy for (SE) functional relevance

c) <u>Biochemical</u>: identify activity traces of genomic elements (chemical signatures such as methylation or histone modification) and treat these traces as a proxy for functional relevance

Here we largely ignore the genetic strategy. This is not because it is unimportant, but rather because it is (for the most part) not a strategy amenable to the systemic goal of creating a genome-wide encyclopedia of elements. It would be tremendously time-consuming and expensive to undertake on a global scale.[9] Often it is deployed tactically in combination with the other strategies, especially for functional validation.[10]

The remaining two strategies will be explored primarily through the lens of ENCODE's systemic cataloguing aims. An evolutionary strategy utilizes a comparative analysis that includes more than the human genome and can indicate functional relevance in terms of selected effects on the assumption that conservation is a proxy for maintenance due to fitness contributions (Section 3.2). However, as a matter of fact, the amount of conserved sequence (across mammals)

[9] This systemic orientation goes back to the origins of ENCODE. In 2002, a progenitor meeting was accurately described as the "Workshop on the Comprehensive Extraction of Biological Information from Genomic Sequence" (7108-019fcnotes). It also appears when announcing the modENCODE strand: "To interpret the genome accurately requires a *complete* list of functionally important elements and a description of their dynamic activities over time and across different cell types" (Celniker et al. 2009, 927, emphasis added).

[10] The development of so-called 'massively parallel reporter assays' (MPRAs) has the potential to overcome this problem to some degree (Melnikov et al. 2012). MPRAs are likely to become more important in functional genomics and currently are utilized by ENCODE 4. These assays, however, are still used in conjunction with other approaches, building on the libraries created by proxies from biochemical and evolutionary strategies.

cannot speak to a large portion of the genome. This points us toward the biochemical strategy (Section 3.3), which has the greatest potential for addressing FICV systemically—though fallibly and defeasibly—across the genome (Fig. 1). Thus, it unsurprisingly predominated within ENCODE, as well as in the early stages of modENCODE where it was elaborated in combination with the evolutionary strategy (Section 4).

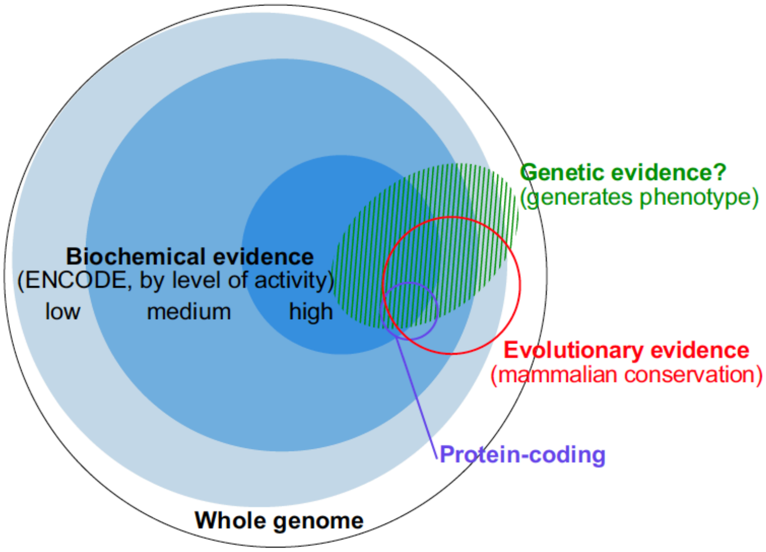

Figure 1
Different types of evidence for genomic function from three strategies (Kellis et al. 2014, 6132).

3.2. The Evolutionary Strategy

The central tool in comparative genomics is sequence alignment. This makes it possible to identify stretches of DNA that exhibit different degrees of similarity across both closely and distantly related organisms. Prior to full genome sequencing, much of this comparison was done on a gene-by-gene basis (Altschul et al. 1990; McGinnis and Madden 2004). As larger stretches of DNA became available, chromosomal alignments revealed substantial blocks of similarly organized sequence deemed “synteny” (e.g., Guiliano et al. 2002). However, the growing number of complete genomic sequences available from different taxa has made possible increasingly sophisticated bioinformatic comparisons of alignment across entire genomes (Lal and Verma 2017).

Sequence alignment is informative in a number of ways, but its use in the evolutionary strategy of functional genomics relates to the identification of "highly conserved" segments of DNA. Phylogenetic conservation of sequence is a proxy for the historical maintenance of this sequence due to fitness effects. Part of the justification for seeing conservation as evidence for function derives from successful empirical demonstrations. For example, the initial discovery of conservation across metazoans of the ~180 "homeobox" nucleotide sequence responsible for the ~60 amino acid "homeodomain" region of DNA-binding proteins (i.e., transcription factors) was completely unexpected (McGinnis et al. 1984; Scott and Weiner 1984). The conservation of this sequence related to the fact that *Hox* genes regulate anterior-poster axis specification early in development (McGinnis and Krumlauf 1992; Mallo and Alonso 2013; Pearson et al. 2005); too many changes in the homeodomain would prevent these transcription factors from properly coordinating a very basic feature of animals. Other investigations have uncovered a variety of conservation patterns, including ultra-conserved elements that are "perfectly" conserved across distantly related taxa within a clade (McCole et al. 2018). The identification of sequence conservation for genomic elements through different types of cross-species comparison is a proxy for SE functional relevance.

A key handicap for the evolutionary strategy within the context of ENCODE is coverage. Available evidence showed that the amount of conserved mammalian DNA sequence was estimated at around 5% and no more than 10% (Lindblad-Toh et al. 2011). Although one could argue that estimates of functionality based on evolutionary constraint increase when other processes are taken into account besides broad sequence conservation due to purifying selection, such as evolutionary turnover or lineage-specific constraint (Ward and Kellis 2012), this would not come close to addressing the aim of cataloguing functional elements in the genome on a global scale.[11] In addition to the handicap of coverage, those sequences that *are* conserved might not have biological relevance (Palazzo and Lee 2015). This means that sequence conservation is not sufficient to ascribe function to a genomic part. Although bioinformatic comparisons of alignment across entire genomes can be a defeasible discovery tool for functional elements (Lal and Verma 2017), they cannot distinguish present functional operation (maintenance) from past functional origination (Linquist et al. 2020), nor separate out multifunctionality: the method only says that something is functional, not what the function is. For similar reasons, functional parasitism (what lineage the function is for) cannot be teased apart easily. Finally, the evolutionary strategy cannot address function loss because areas of the genome that have biological relevance might only be poorly conserved at the sequence level (Fisher et al. 2006). Conservation *per se* is not necessary for biological relevance. And none of this determines where to set the threshold for how much conservation is needed to ascribe functionality to a sequence. "Highly" conserved regions are deemed more likely to have functional importance, but where is the cut-off to distinguish non-functional elements? Further measures and a more sophisticated approach to conservation are needed to make an informed decision about functional elements in a genome (Ponting 2017).

[11] A press release about the findings of ENCODE 1 glimpsed this possibility and offered a speculative model of why so little of the sequence was conserved. "According to ENCODE researchers, this lack of evolutionary constraint may indicate that many species' genomes contain a pool of functional elements, including RNA transcripts, that provide no specific benefits in terms of survival or reproduction. As this pool turns over during evolutionary time, researchers speculate it may serve as a "warehouse for natural selection" by acting as a source of functional elements unique to each species and of elements that perform the similar functions among species despite having sequences that appear dissimilar" (https://www.nih.gov/news-events/news-releases/new-findings-challenge-established-views-human-genome). To our knowledge, this idea of a functional 'reservoir' or 'pool' was not developed further, though it was severely criticized (Doolittle 2013).

3.3. The Biochemical Strategy

Although both the genetic and evolutionary strategies can assist in addressing the problem of FICV in genomics, something else was needed to secure a comprehensive vantage point. The biochemical strategy starts from the premise that there are certain activities in the genome that leave behind traces which are correlated with well-established functional patterns of genomic DNA. These activity patterns can be divided roughly into two classes. On the one hand, there are *products* of genome activity, such as coding and non-coding RNA molecules transcribed from DNA. Detailed analyses of these transcripts indicate active areas of the genome and often give researchers clues to the location of other functional elements, such as transcription start or splice sites. On the other hand, there are *modifications* of chromosomes that are characteristic of active regions, such as the methylation of DNA, histone modifications, transcription factor binding, or the physical accessibility of the DNA. These modifications are often associated with known, non-coding functional elements (e.g., promoters or enhancers). Tracking such biochemical signatures allows researchers to assess entire genomes for functional elements because the relevant techniques, such as RNA sequencing and DNA-binding assays, can be scaled up for whole-genome analysis.

Within ENCODE, the biochemical strategy was implemented using about two dozen different types of assays. These included the sequencing of RNA to identify coding and non-coding genes; Chromatin Immuno-Precipitation (ChIP) coupled to high-throughput sequencing (ChIP-seq) to identify binding sites for transcription factors and other proteins; DNase I hypersensitivity assays to identify regulatory regions that are more accessible; and, reduced representation bisulfite sequencing to identify methylated sites in the genome that are often associated with regions of regulatory relevance. These and other assays were performed using a wide range of cell types (over 100 different cell lines) to ensure that cell-type specific modifications or behaviors could be detected (ENCODE Consortium 2012). This made it possible to apply proxies of the biochemical strategy in order to ascertain functional elements systemically, while taking into account the highly context-dependent workings of the genome.

Although the biochemical strategy is powerful in its scope, it is not without limitations. One limitation frequently highlighted by critics of ENCODE is that the existence of biochemical traces does not prove the functional relevance of the DNA sequences involved (Doolittle 2013; Eddy 2013; Graur et al. 2013; Niu and Jiang 2013). The observed product or modification could be due to the noisy nature of cellular operations. Some transcription of genomic DNA can be spurious and give rise to non-coding RNAs that have no function (Struhl 2007; Chen et al. 2014; Palazzo and Lee 2015; Quinn and Chang 2016). Additionally, biochemical traces can be experimental artefacts. This difficulty is visible in the large-scale analysis of DNA-binding events via ChIP-seq, a method that relies on the use of target-specific antibodies, which can give rise to false positives due to issues with antibody specificity (Wardle and Tan 2015). ‘Hyper-ChIPable’ regions indicate a high accumulation of DNA-binding factors that potentially have no biological relevance (Teytelman et al. 2013; Park et al. 2013; Wreczycka et al. 2019). As should be expected from proxies, not every biochemical trace reliably signals a functionally relevant activity. At most, they provide a list of *putative* functional elements that must be validated in a second step, such as by using genetic or evolutionary approaches (Germain et al. 2014).

Working with defeasible and fallible proxies is the main *modus operandi* for functional genomics. However, the fact that researchers rely on imperfect measures should not be interpreted negatively. Not only are they necessary but these methods foster agility and flexibility in the research process. Precisely because each of them is individually insufficient, researchers routinely mix and mingle strategies in different ways to find new tactics for addressing FICV. Importantly, these complicated dynamics in the use of proxies has played out in different ways across strands of the ENCODE project. To achieve an understanding of these dynamics, it is necessary to look at more than one strand of ENCODE and at more than one particular phase (e.g., the 2012 release of data by ENCODE 2). We therefore intentionally move away from the ENCODE controversy to focus on the modENCODE strand that has received little attention in the literature. It helps display the dynamic marshaling of proxies that unfolds in the context of understanding and dealing with the FICV problem.

4. Doing Things Differently: modENCODE

4.1. The road to modENCODE

For the researchers involved in ENCODE, it was clear from the beginning that in order to achieve their goal of functionally analyzing the genome at a global scale they would have to tackle multiple methodological issues. This shaped how the project was set up and how it developed over time. ENCODE began with a pilot phase ('ENCODE 1') that focused on only 1% of the human genome and could serve as a test bed for methods that would be suitable for whole-genome functional analysis (ENCODE Consortium 2007). These methods were largely assays from the biochemical strategy. This pilot phase was followed by two strands: the well-documented main production phase (ENCODE 2) and a less publicized model organism strand (modENCODE) (see Fig. 2 for an overview of modENCODE and how it aligns with the different phases of ENCODE).

modENCODE had several investigative goals (http://www.modencode.org/).[12] First, it aimed to cross-catalogue all the functional elements from the genomes of two powerhouse model organisms: *Caenorhabditis elegans* and *Drosophila melanogaster*. Beyond deepening the understanding of these model organisms, researchers hoped that findings from these systems might add further insights into how the human genome works, as well as how genomes function more generally. Second, scientists were pursuing increased genomic 'literacy' by looking at more tractable genomes that could inform how functional genomics should be performed.[13] Tractability pertained not only to smaller genome size but also to the tools available for undertaking functional validation using the genetic strategy.[14]

Altogether, investigating model organisms was presented as an effective way of achieving what the human strand of ENCODE could not in the mid-2000s.[15] However, if we look more closely at what came out of modENCODE, the genetic approach was not used in a significant manner. Publications were dominated by methods associated with the biochemical strategy. Even more intriguing is that modENCODE researchers combined the biochemical and evolutionary strategies in new ways, using them not only to validate functional annotations but

[12] https://www.genome.gov/26524507/the-modencode-project-model-organism-encyclopedia-of-dna-elements-modencode.

[13] "The tractable genomes of model organisms will help us develop full literacy in "reading" genomes. Genome literacy in turn will provide a framework for understanding how complex biological systems work" (Drosophila ENCODE white paper, 2005; D_ENCODE_WP).

[14] This was explicitly noted by researchers from the fly and worm communities when they pitched the idea of a model organism ENCODE to the NHGRI. In a May 2005 letter sent to Elise Feingold and Peter Good, ten researchers from these communities emphasized that: "[T]he ability to test the *in vivo* function of the discovered elements systematically [using genetic tools] is the single most important benefit of carrying out ENCODE projects in these species. We expect that the technology devised and the lessons learned from study of both *D. melanogaster* and *C. elegans* will transfer rapidly to human ENCODE projects" (Joint_Letter_050405; ID #784). It later was stated in the public announcement of modENCODE: "The genomes are small enough to be investigated comprehensively with current technologies and findings can be validated *in vivo*. The research communities that study these two organisms will rapidly make use of the modENCODE results, deploying powerful experimental approaches that are often not possible or practical in mammals, including genetic, genomic, transgenic, biochemical and RNAi assays. modENCODE, with its potential for biological validation, will add value to the human ENCODE effort by illuminating the relationship between molecular and biological events" (Celniker et al. 2009, 927).

[15] There also was a mouse ENCODE strand, the findings of which were published in 2014 (Yue et al. 2014).

also to isolate novel conserved principles of genome operation. While the idea of finding general principles underlying genome functioning was a goal of many ENCODE researchers,[16] the modENCODE project enabled researchers to push this idea further, uncovering principles that not only direct the human genome but metazoan genomes more generally.

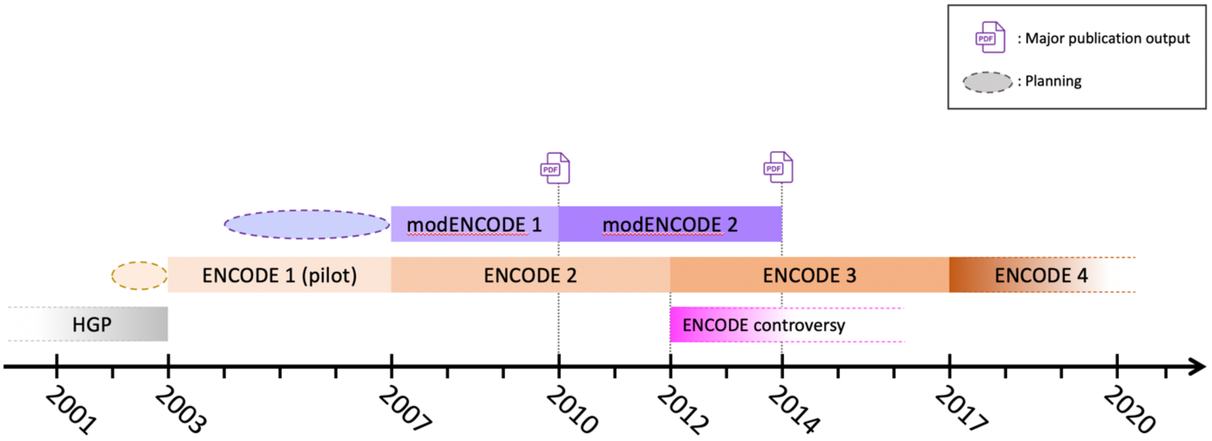

Figure 2
A timeline of modENCODE that illustrates how its different phases and key outputs align with the main ENCODE strand.

4.2. modENCODE: "Phase 1" (Biochemical Traces)

The initial findings of modENCODE for both fly and worm were published in 2010 (Fig. 2). These publications can be thought of as the output of a first phase (2007–2010), which was broadly similar to ENCODE 2: biochemical traces collected to ascertain the identity and location of different functional elements. Researchers in the modENCODE consortium looked at proxy measures of gene expression, histone modifications, transcription-factor binding sites, and chromatin structures to create a list of functional elements for fly and worm genomes (Gerstein

[16] The first report from the pilot phase in 2007 concluded: "The scale of the pilot phase of the ENCODE Project was also sufficiently large and unbiased to reveal important principles about the organization of functional elements in the human genome" (ENCODE Consortium 2007, 812).

et al. 2010; Roy et al. 2010). They implemented the biochemical strategy using a wide range of methods, from RNA-sequencing (using both poly(A)$^+$ and total RNA as a source) to Chromatin-Immuno-Precipitation (ChIP) coupled to microarray analysis (ChIP-chip).[17] With these and other methods, the *Drosophila* strand of modENCODE found almost 2,000 new genes that had not been annotated previously (Roy et al. 2010). Additionally, they identified a bevy of small regulatory RNAs (e.g., micro-RNAs or piwi-interacting RNAs), mapped the chromatin structure of these organisms (discovering significant changes in chromatin organization between different cell types), and uncovered key patterns in histone modifications that were enriched in regions close to transcription start sites (thereby potentially relevant for gene expression regulation).

In this first phase, modENCODE researchers mainly talked of ‘candidate functions’ being identified using these assays; they were aware that the biochemical strategy involved proxy measures that did not directly solve the FICV problem.[18] Significantly, these early modENCODE papers were published *before* the ENCODE controversy broke out in 2012. As we saw, a key claim of critics was that ENCODE researchers were not careful in their claims and did not use the correct methodology to identify genuine functional elements. The early modENCODE output shows that ENCODE researchers approached functional ascriptions in a more differentiated manner. Although some interpreted the output from ENCODE 2 as supporting strong claims about the total amount of genome functionality, modENCODE shows that other researchers were more cautious. The ENCODE controversy is not representative of how different ENCODE project participants approached the FICV problem.

[17] The latter was used to identify sequences where specific proteins or modified proteins could be found on DNA. ChIP-chip was quickly replaced by ChIP coupled to high-throughput sequencing (ChIP-seq), which gives higher resolution and better coverage than ChIP-chip.

[18] “We assigned candidate functions to the fraction of the nonrepetitive genome covered by the data sets, excluding large blocks of repeats and low-complexity sequences” (Roy et al. 2010, 1792).

4.3. Moving Towards Validation

After publishing their initial findings in 2010, a significant part of modENCODE research focused on further characterizing the fly transcriptome.[19] The goal was not only to refine the existing annotations of fly transcripts, but to *validate* the transcripts that had been identified. This aligns with the cautious formulations of modENCODE researchers regarding putative functional elements. What additional test could validate these elements as genuinely functional? Would the assays from the genetic strategy, distinctly available in model organisms, be employed (a key motivation for setting up a modENCODE in the first place)?[20] Interestingly, not so much.[21] What became crucial was tracing the dynamics of the model organism genomes and identifying novel patterns. For instance, it was possible to run time-courses at short intervals to follow how the composition of the transcriptome changed over developmental time in *Drosophila* (Graveley et al. 2011). Such analyses were not possible in humans.[22] The data from these analyses were then used to draw conclusions about the functional significance of the differentially expressed transcripts. Researchers knew that different parts of the genome are activated differently across developmental stages, but if the transcripts identified showed specific

[19] There also was an effort to understand transcriptional *regulation* in worm as part of the modERN project (model organism Encyclopedia of Regulatory Networks) (Kudron et al. 2018).

[20] "The biological significance of the genomic features identified will be tested in experiments designed to evaluate the accuracy and functionality of subsets of the structural and regulatory annotations" (Celniker et al. 2009, 929).

[21] Some modENCODE studies did use the genetic strategy for validation of elements identified with the biochemical strategy. Nègre et al. (2011) used a reporter assay in transgenic *Drosophila* to validate a small sample of putative enhancer and insulator elements identified using CBP-binding as a biochemical marker. Schwartz et al. (2012) used another reporter assay to validate putative insulator protein binding sites. Brooks et al. (2011) used minigene reporter assays to validate putative exonic and intronic splicing enhancers they identified in *Drosophila*. However, the genetic strategy was not employed systematically. For example, Nègre et al. (2011) only tested 0.2% of the putative elements they identified in their genome-wide screen for CBP-binding sites. Brooks et al. (2011) only validated 7% of the exonic splicing enhancers they identified and 1.3% of the intronic splicing enhancers.

[22] "The use of model organisms enabled a variety of experimental approaches not tenable in humans: conducting developmental time courses in isogenic populations, performing environmental perturbations, and validating findings in the most powerful metazoan genetic systems yet developed" (Brown and Celniker 2015, 32).

patterns of developmental variation in expression, then these patterns would indicate functional relevance; that is, they helped to validate the proxy findings (i.e., activity traces) from the biochemical strategy. They also raised the possibility that similar patterns could be detected across species, pointing to general principles of how genomes function.

This more elaborate use of data from proxy measures is not free of potential error. For example, Palazzo and Lee (2015) argue that context-specific expression is not necessarily a sign of functional relevance.[23] Such expression patterns could arise for sequences that were randomly translated due to transcriptional noise because noisy expression is still driven by specific transcription factors. Individual transcription factors or clusters might have their own pattern of noise—a set of sequences for which it is more likely to trigger background expression. If that is the case, then certain sequences could display context-specific expression noise because transcription factors are typically expressed in a highly regulated manner. Their pattern of noise would also manifest in a tissue- or development-specific manner. If true, then even tissue- or stage-specific expression of an RNA would not provide functional validation. Overall, this serves as an important warning against jumping to conclusions about solving the FICV problem.

The validation of transcriptome data in fly was advanced further in a study that not only looked at the expression of RNA in different developmental stages but also at the conservation of expression of different transcripts across different fruit fly species (Chen et al. 2014). Sequence conservation and conservation of expression were combined into a 'conservation index' that could then be compared to what a neutral model would predict for a randomly generated transcript (i.e., noise). Transcripts that had a conservation index above neutral levels were treated as functionally relevant. This analysis showed that most of the identified transcripts in

[23] See also Eddy (2013) and his proposal of a 'Random Genome Project' for an incisive discussion of this point.

Drosophila, which cover a majority of the genome, should be counted as conserved and thus are 'functional.' Putative functional elements are attributed biological relevance through a novel measure of conservation (i.e., validation by the evolutionary strategy). However, whether this applies more broadly is unclear because other analyses of the human transcriptome suggest that most of the expressed sequences reflect noise rather than biologically relevant transcription events (Pertea et al. 2018).

Regardless of the status of diverging empirical claims, these types of validation studies deviate from the linear narrative that is still being used for the human ENCODE project: first use the biochemical strategy and then validate with the genetic strategy. In the case of modENCODE, the biochemical strategy was combined with conservation proxies from the evolutionary strategy to validate a list of putative functional elements. This type of combination was only possible because of opportunities arising from model organisms and led to an approach that no longer followed the narrative that was originally used to sell the project idea to the NHGRI (and which was largely followed for human ENCODE). Independent of a presumed concept of function, the proxy toolkit deployed to address FICV developed dynamically within modENCODE and generated novel insights about genomic function. This is even more evident in the last batch of papers that came out of modENCODE in 2014.

4.4. modENCODE: "Phase 2" (Abstract Principles of Genomic Function)

In section 4.1, we implied that modENCODE can be split into two phases: Phase 1 (~2007-2010) and Phase 2 (~2010-2014). Distinguishing phases of modENCODE is somewhat arbitrary because there was no official separation (as was the case for the human ENCODE strand, Fig. 2). Yet drawing such a distinction can help to identify and isolate particular tendencies in modENCODE. The separation we propose aligns with key data releases in publications from the project, first in 2010 and then in 2014 (even though other publications and review articles were released in the interim). These two waves of publication also mark an observable shift in the way proxy measures were used from the biochemical and evolutionary strategies. Phase 1 of modENCODE parallels the methodology of human ENCODE with both strands primarily using proxy measures from the biochemical strategy. However, subsequently, we see shifts in modENCODE, such as in Chen et al. (2014) where transcript analysis became combined with new ways of using conservation as a proxy. This combinatorial methodology is prominent in a number of publications from 2014. Instead of using proxy measures from one or the other of the three strategies in varying combinations, Phase 2 of modENCODE often fused the biochemical and evolutionary strategies to isolate shared relational functional properties—not elements or structures—of metazoan genomes. The somewhat vague initial promise and hazily glimpsed prospect of isolating "fundamental principles of fly and worm genome biology" and "emergent properties" of genomes finally began to come into focus.[24]

Examples of this novel methodological fusion can be found in three papers published in 2014 (Boyle et al. 2014; Gerstein et al. 2014; Ho et al. 2014). These papers addressed gene regulatory patterns, the transcriptome, and chromatin organization in metazoans, respectively. Gerstein and co-workers focused their analysis on the metazoan transcriptome, paying particular attention to the transcription of orthologous genes. Comparing the levels and the timing of expression of these gene, they were able to identify sixteen modules of co-expressed sequences that display significant conservation across humans, fly, and worm (Gerstein et al. 2014). Twelve

[24] "Integrative analysis of these data across the different types of functional element will be used to reveal fundamental principles of fly and worm genome biology and to begin to uncover the emergent properties of these complex genomes. …We will search global patterns identified in the regulatory programs for emerging principles of gene regulation within and across species" (Celniker et al. 2009, 928-9).

of these conserved modules were enriched for 'hourglass genes.' Investigations of the maximally conserved "phylotypic stage" in different taxa have accumulated transcriptomic and genomic data in support of constrained genomic regulation at this time (Kalinka et al. 2010; Piasecka et al. 2013). The identification of enrichment correlated with this developmental stage suggests the existence of what Gerstein and co-workers call "conserved biological principles" across metazoan genomes. Furthermore, they used chromatin modifications at the promoters across the three species to quantitatively predict gene-expression levels. This finding suggests "quantitative preservation of the biological impact of chromatin modifications on transcription across phyla" (Brown and Celniker 2015, 38).

Similar findings were made for chromatin organization (Ho et al. 2014) and gene regulatory circuits (Boyle et al. 2014). Comparing the levels and distributions of biochemical marks (e.g. histone methylation) across humans, fly and worm, several conserved features of chromatin structure were uncovered, such as the organization of lamina-associated domains or particular patterns of histone modifications (Ho et al. 2014).[25] An analysis of the genome-wide binding sites of over 300 transcription factors (165 in human, 52 in fly, and 93 in worm), revealed that general regulatory principles, such as motif recognition and network structure, are conserved across metazoans. For instance, looking at orthologous transcription factor families, Boyle and co-workers found that many factors within these families display similar sequence specificity, indicating metazoan conservation of binding motif use. They also discovered that binding sites for regulatory factors are not randomly distributed, with about half of binding

[25] "Combinatorial patterns of histone marks were recovered in both species, many with a striking resemblance to those previously reported in humans; marks appear to function in globally similar fashions, … Patterns of histone modifications at promoters were highly similar and deeply conserved in all three species. … [there is] global similarity of average patterns of chromatin modifications at promoters and enhancers across 600 million years of evolutionary divergence" (Brown and Celniker 2015, 34).

events in all three species occurring in high-occupancy target (HOT) regions. Furthermore, general network motifs, such as feed-forward loops, were found to be conserved in all three species. Overall, these findings point to conserved principles that underlie transcriptional regulation in metazoan genomes (Boyle et al. 2014).

4.5. A Theoretical Tension Manifests: Conservation of Functional Principles?

The methodological developments in Phase 2 changed what was being measured and assessed. Rather than only looking at structural features of the genome, what modENCODE showed was that "quantitative relationships among chromatin state, transcription, and cotranscriptional RNA processing are deeply conserved" (Brown and Celniker 2015, 31). This, however, introduced a theoretical tension: the notion of conservation applies straightforwardly to sequence-based functional elements (i.e., structures), but less clearly to quantitative functional relationships ("average patterns") or prerequisites of genome operation. A particular trait, such as a stretch of DNA, is homologous when it is judged to be the "same" under every variety of form and function through a process of common descent (i.e., specific sequence details and contribution to organismal working can vary). However, as the phrase "every variety of form and function" implies, and classical examples of homology such as the tetrapod limb indicate, the function of a trait typically does not contribute to whether the trait is homologous. Homologues can and do often vary dramatically in what functions they perform, whether it be grasping hands of humans, swimming fins of whales, digging claws of moles, or hoofed forelegs of horses. Talk of "functional homology" has long been flagged as suspect (Abouheif et al. 1997).

Some philosophical analyses have offered strategies for recovering a legitimate notion of "homology of function" based on more nuanced distinctions about function, such as distinguishing use versus activity (Love 2007) or appealing to processes as units (Gilbert and Bolker 2001). There is also a corresponding notion of "conserved mechanism" that applies to developmental signaling pathways which have been maintained throughout metazoan evolution (e.g., Wnt signaling) by anchoring the activity of these mechanisms in molecules of special quality (Love 2018). Yet these "surprisingly deep similarities in the mechanisms underlying developmental processes across a wide range of bilaterally symmetric metazoans" (Bier and McGinnis 2003, 25) tend not to depend on genomic organization but rather shared gene regulatory networks and signaling pathways: "the spatial control of Hox gene expression seems to be mostly independent of genomic arrangement" (Mallo and Alonso 2013, 3953). How can a quantitative relationship among transcriptional states be homologous?

This tension can be brought into sharper relief by noting that other quantitative functional relationships described by evolutionary biologists are not often considered homologous. For example, the biomechanics of swimming demands that organisms reduce the resistance of their bodies to movement in terms of morphological profile (form) and its encounter with the fluid medium (friction) (Alexander 2003; Vogel 2003). The relationship between these two aspects of resistance is represented by the Reynolds number, which is calculated from an equation that relates variables of the organism (e.g., cross-sectional area) to variables of the medium (e.g., fluid density and viscosity). In short, organisms operating at higher Reynolds numbers experience more resistance. To be clear, this is a quantitative functional relationship. However, the conformity of organismal morphology to the Reynolds number, such as body streamlining in larger animals that swim (e.g., dolphins, sharks, or tuna), is not described as a deeply conserved principle. This is because it is a physico-chemical principle that applies to both living and non-living systems. Populations of swimming organisms are shaped evolutionarily by the fitness effects that arise from greater or lesser degrees of locomotory ability. This is convergence on a functional optimum (constrained by various trade-offs), not homology.

We can now state the theoretical tension more explicitly. Should we consider the quantitative relationship among chromatin or transcriptional states as more like the Reynolds number in biomechanics or the tetrapod limb? If the former, then claims about conservation appear misplaced. These quantitative relationships among states in the genome correspond to optima in yet-to-be-formulated mathematical equations that relate the relevant variables. Their appearance across metazoans is because that is how genomes work relatively efficiently (depending on trade-offs). If the latter, then the increased abstraction involved in formulating these quantitative relationships makes it unclear how to understand them as "structure" that could be homologous. The notion of conservation applies to sequence-based functional elements (i.e., structures), but not necessarily quantitative functional relationships (e.g., average patterns of chromatin modification). Importantly, this is a tension rather than a paradox. How it gets resolved, if at all, is an open question. What it highlights starkly is that the innovative combination of proxy methods from the biochemical and evolutionary approaches used by modENCODE led to discoveries whose categorization is not so clear. modENCODE may have identified physicochemical rules (abstract, quantitative relationships) rather than mechanistic structure. Although our knowledge of how genomes operate was advanced, it may not have been with respect to the translation of genomic form into organismal complexity.[26] The original research question was transformed in the process of inquiry (Yanai and Lercher 2019), from how

[26] "It might be that the genome tells us no more about how an organism builds and sustains itself than a dictionary does about how a story unfolds. New methods, rather than finally answering old questions, could merely beggar them, shifting the goalposts entirely" (Ball 2019, 31).

genomic functional elements produce an organism to what properties or rules of the genome make it possible to function.

5. Conclusion: Methods have a Life of their Own

The developments described above were not anticipated. ENCODE had focused on functional elements of the genome (i.e., structures), but modENCODE began to characterize general regulatory principles across genomes (i.e., functional rules). These rules involve a reasoning strategy of abstraction, moving away from a catalog of functional elements to global relationships of genome operation that apply to all of these elements simultaneously. This also modulated the methodological difficulties inherent in FICV when applied to genomes (Section 2.2). Location is no longer an issue. These functional rules are not uncovered by decomposing the genome into working parts, localizing their characteristic behaviors, and validating their biological contribution. The issue of completeness is similarly transformed. modENCODE was not only trying to catalog or list genomic elements. Instead of creating exhaustive lists of functional part-types, they were exploring the existence of properties that characterize systemic genomic functioning in a way not previously conceived. As a consequence, vagueness is mitigated; the issue is not clarifying the criteria of membership for categories of functional elements but re-characterizing what it means to talk about genome function *per se*.

The issues that a SE approach struggles with also become transformed in the mixed strategy employed by Phase 2 of modENCODE. Past-present ambiguity is mitigated, and multi-functionality becomes difficult to interpret. Is it appropriate to say that these abstract control principles solve the problem of how genomes work? This is not about "how the information encoded in a genome can produce a complex multicellular organism" (Celniker et al. 2009, 927), but rather how a genome as a unit operates. modENCODE engaged in a different kind of global discovery process. Instead of comprehensively extracting genomic information, the task shifted to identifying comprehensive operating rules for the genome—not how information in the genome is used but how it is possible for a genome to use information.

Even more poignant is how functional validation is transformed. How do you validate quantitative functional relationships that apply to the entire genome simultaneously? Notice that the genetic strategy is of no help. Experimental interventions such as blocking the operation of a part or removing it altogether to ascertain if and how this intervention affects the system behavior will not work. We are no longer talking about parts at all; the focus is on the whole. Instead, validation is accomplished through tactical applications of the evolutionary strategy to the whole genome to confirm that these quantitative relationships are conserved. These applications include establishing conservation of these relationships on the widest phylogenetic scale possible (hence the ascription of "deeply conserved"),[27] as well as fine-grained evaluation of conservation across clades, whether different species of *Drosophila* (Chen et al. 2014) or the entire arthropod phylum (Lewis et al. 2020).[28]

Proxies have a life of their own, developing in ways that cannot always be predicted, in part as a consequence of new technologies (e.g., new RNA sequencing power; Wang et al. 2009) and in part as a result of more data (e.g., more genomes sequenced). Thus, it is not surprising that a theoretical tension crept into reasoning about genomics (Section 4.5). As in the case of the transcriptome analysis (Section 4.3), the 2014 papers (Section 4.4) illustrate a novel methodological orientation to the FICV problem, which grows out of the use of proxies in the

[27] "[A]pparently pan-metazoan aspects of transcriptome organization and dynamics … The exon painting structure of this mark is conserved across all metazoans surveyed" (Brown and Celniker 2015, 38).

[28] "RNA and chromatin data enabled the interrogation of several long-standing phylum-specific phenomena, such as genome-wide profiling of trans-splicing in worms" (Brown and Celniker 2015, 38).

biochemical strategy alongside the comparative possibilities for the evolutionary strategy that model organism research offers. Conservation, the proxy signal of the evolutionary strategy, becomes central but is no longer focused on sequence information but instead on average patterns in biochemical traces. The proxy data from the biochemical strategy was increasingly available, and this nurtured novel extensions and combinations not envisioned at the outset. This flexibility and open-ended nature of the proxy toolkit, which doesn't flow directly from a particular function concept, fosters dynamic experimental possibilities. This moves the researchers into a new space where they can ask new questions.[29]

Although one can trace the origin of these new questions from those that formed the starting point of the project, their emergence was not a foregone conclusion and is accompanied by other features. For example, the formulation of quantitative functional relationships in Phase 2 of modENCODE involves increasingly abstract reasoning. This abstract conceptualization of general principles that underlie genome function was fostered by the new proxy approach that fused biochemical and evolutionary considerations. In the process, we end up with a different understanding of what it means to do functional genomics. The associated theoretical tension that arises out of these maneuvers becomes its own distinct challenge. Only time will tell if a homology perspective ("conservation") is the most empirically appropriate way to understand these quantitative functional relationships, as opposed to seeing these control principles from an analogy perspective where they arise from convergent evolution.

From the beginning, ENCODE was seen as a methodological challenge. Making functional genomics work was not about getting the theory of function right, but about figuring out how to use a wide range of proxies to tackle FICV. Theories of biological function, whether

[29] The use and importance of proxies applies to genomics more generally. For example, the International HapMap Project relied on proxies to identify loci of significance for health and disease.

the CR or SE account, do not provide sufficient methodological guidance (Section 2).[30] Functional analysis is a messy business that works with a complex toolkit of proxy approaches that are only loosely aligned with the different function concepts (Section 3). The alignment is not precise nor exclusive and, as a result, the proxies that each approach uses exhibit degrees of freedom that permit combination and development in different and unexpected ways. This openness, however, does not mean 'anything goes.' The biochemical approach became dominant in human ENCODE because of the goal of comprehensive information extraction. This approach also was dominant in Phase 1 of modENCODE (Section 4) and for similar reasons: it was the most powerful approach for scanning whole genomes in a range of backgrounds and gaining insights into their workings. However, the collection of proxies started to develop further in response to data and was driven by practical opportunities that model organisms offered. Eventually, this led researchers to fuse the biochemical and evolutionary approaches through novel conservation proxies. Sequence-based elements moved into the background, whereas more abstract, quantitative principles underlying genomic functioning came into focus and yielded new challenges, such as how the concept of 'conservation' can be applied to these quantitative relationships.

Ultimately, modENCODE moved into a space that was not anticipated at the beginning of the project. It moved away from the linear two-step narrative that dominated the human ENCODE and its own Phase 1. The idea was no longer that the biochemical proxy would deliver putative elements to be subsequently validated using the genetic approach. The FICV problem took on a different shape. The question that was addressed changed. In all of this, the discussion

[30] Other accounts of function would not overcome this problem. It is not that the two dominant accounts are not sufficiently developed or powerful. The problem lies with the flawed idea of a linear, asymmetrical path that leads from theory to practice.

about concepts of function that dominated the ENCODE controversy in 2012 did not play a role. It was the dynamics of proxy measures in modENCODE, together with new methodological developments, that drove the research forward and led ENCODE scientists to novel concepts, abstractions, and other epistemological innovations. Our analysis reveals this more variegated understanding of what functional genomics is, especially the decoupling of theoretical concepts from methodological practices, and moves the conversation away from the ENCODE controversy to the real challenge of addressing FICV in all of its manifestations.

References

Abouheif, E., M. Akam, W.J. Dickinson, P.W.H. Holland, A. Meyer, N.H. Patel, R.A. Raff, V.L. Roth and G.A. Wray. 1997. Homology and developmental genes. *Trends in Genetics* **13**:432–433.

Alexander, R.M. 2003. *Principles of Animal Locomotion*. Princeton: Princeton University Press.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410.

Ball, P. 2019. Science must move with the times. *Nature* 575:29–31.

Basehoar, A.D., S.J. Zanton, and B.F. Pugh. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**:699–709.

Bechtel, W. and R. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Bier, E. and W. McGinnis. 2003. Model organisms in the study of development and disease. In: C.J. Epstein, R.P. Erickson, and A. Wynshaw-Boris (eds), *Molecular Basis of Inborn Errors of Development*. New York: Oxford University Press, 25–45.

Boyle, A.P., C.L. Araya, C. Brdlik, P. Cayting, C. Cheng, et al. 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**:453–456.

Brooks, A.N., J.L. Aspden, A.I. Podgornaia, D.C. Rio, and S.E. Brenner. 2011. Identification and experimental validation of splicing regulatory elements in *Drosophila melanogaster* reveals functionally conserved splicing enhancers in metazoans. *RNA* **17**:1884–1894.

Brown, J.B. and S.E. Celniker. 2015. Lessons from modENCODE. *Annual Review of Genomics and Human Genetics* **16**:31–53.

Brzović, Z., and P. Šustar. 2020. Postgenomics function monism. *Studies in History and Philosophy of Biological and Biomedical Sciences* **80**:101243.

Celniker, S. E., L.A.L. Dillon, M.B. Gerstein, K.C. Gunsalus, S. Henikoff, et al. 2009. Unlocking the secrets of the genome. *Nature* **459**:927–930.

Chen, Z.X., D. Sturgill, J. Qu, H. Jiang, S. Park, et al. 2014. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Research* **24**:1209–1223.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* **72**:741–765.

Doolittle, W.F. 2013. Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences USA* **110**:5294–5300.

Doolittle, W.F. 2018. We simply cannot go on being so vague about 'function'. *Genome Biology* **19**:1–3.

Doolittle, W.F., T.D. Brunet, S. Linquist, and T.R. Gregory. 2014. Distinguishing between "function" and "effect" in genome biology. *Genome Biology and Evolution* **6**:1234–1237.

Eddy, S.R. 2013. The ENCODE project: missteps overshadowing a success. *Current Biology* **23**:R259–R261.

Elliott, T.A., S. Linquist, and T.R. Gregory. 2014. Conceptual and empirical challenges of ascribing functions to transposable elements. *The American Naturalist* **184**:14–24.

ENCODE Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**:799–816.

ENCODE Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74.

ENCODE Consortium. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**:699–710.

Fields, S., Y. Kohara, and D.J. Lockhart. 1999. Functional genomics. *Proceedings of the National Academy of Sciences USA* **96**:8825–8826.

Fisher, S., E.A. Grice, R.M. Vinton, S.L. Bessling, and A.S. McCallion. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**:276–279.

Garson, J. 2016. *A Critical Overview of Biological Functions*. Berlin: Springer.

Germain, P.-L., E. Ratti, and F. Boem. 2014. Junk or functional DNA? ENCODE and the function controversy. *Biology & Philosophy* **29**:807–831.

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**: 669–681.

Gerstein, M.B., Z.J. Lu, E.L. Van Nostrand, C. Cheng, B.I. Arshinoff, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**:1775–1787.

Gerstein, M.B., J. Rozowsky, K.-K. Yan, D. Wang, C. Cheng, et al. 2014. Comparative analysis of the transcriptome across distant species. *Nature* **512**:445–448.

Gilbert, S.F. and J.A. Bolker. 2001. Homologies of process and modular elements of embryonic construction. In: G.P. Wagner (ed), *The Character Concept in Evolutionary Biology*. San Diego: Academic Press, 437–456.

Graur, D., Y. Zheng, N. Price, R.B. Azevedo, R.A. Zufall and E. Elhaik. 2013. On the immortality of television sets: “Function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution* **5**:578–590.

Graveley, B.R., A.N. Brooks, J.W. Carlson, M.O. Duff, J.M. Landolin, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**:473–479.

Guiliano, D.B., N. Hall, S.J. Jones, L.N. Clark, C.H. Corton, B.G. Barrell, and M.L. Blaxter. 2002. Conservation of long-range synteny and microsynteny between the genomes of two distantly related nematodes. *Genome Biology* **3**:1–0057.

Guttinger, S. 2019. Beyond the genome: the transformative power of functional genomics. In J. Lowe (ed), *Genomics in Context*: https://genomicsincontext.wordpress.com/beyond-the-genome-the-transformative-power-of-functional-genomics/ (last accessed 24 July 2020).

Guttinger, S. and J. Dupré. 2016. Genomics and Postgenomics. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/archives/win2016/entries/genomics

Hieter, P. and M. Boguski. 1997. Functional genomics: it's all how you read it. *Science* **278**:601–602.

Ho, J.W.K., Y.L. Jung, T. Liu, B.H. Alver, S. Lee, et al. 2014. Comparative analysis of metazoan chromatin organization. *Nature* **512**:449–452.

Kaiser, M.I. 2018. ENCODE and the parts of the human genome. *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 72:28–37.

Kalinka, A, K.M. Varga, D.T. Gerrard, S. Preibisch, D.L. Corcoran, J. Jarrells, U. Ohler, C.M. Bergman, and P. Tomancak. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 418:811–814.

Keller, E.F. 2000. *The Century of the Gene*. Cambridge, MA: MIT Press.

Keller, E.F. 2015. The postgenomic genome. In: S.S. Richardson and H. Stevens (eds), *Postgenomics: Perspectives on Biology after the Genome*. Durham: Duke University Press, 9–31.

Kellis, M., B. Wold, M.P. Snyder, B.E. Bernstein, A. Kundaje, et al. 2014. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences USA* **111**:6131–6138.

Kudron, M.M., A. Victorsen, L. Gevirtzman, L.W. Hillier, W.W. Fisher, D. Vafeados, M. Kirkey, A.S. Hammonds, J. Gersch, H. Ammouri, and M.L. Wall. 2018. The modERN resource: genome-wide binding profiles for hundreds of *Drosophila* and *Caenorhabditis elegans* transcription factors. *Genetics* **208**:937–949.

Lal, D. and M. Verma. 2017. Large-scale sequence comparison. *Methods in Molecular Biology* **1525**:191–224.

Lewis, S.H., L. Ross, S.A. Bain, E. Pahita, S.A. Smith, R. Cordaux, et al. 2020. Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. *PLoS Genetics* **16**:e1008864.

Lindblad-Toh, K., M. Garber, O. Zuk, M.F. Lin, B.J. Parker, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**:476–482.

Linquist, S., W.F. Doolittle, and A.F. Palazzo. 2020. Getting clear about the F-word in genomics. *PLoS Genetics* **16**:e1008702.

Love, A.C. 2007. Functional homology and homology of function: biological concepts and philosophical consequences. *Biology & Philosophy* **22**:691–708.

Love, A.C. 2018. Developmental mechanisms. In: S. Glennan and P. Illari (eds), *The Routledge Handbook of the Philosophy of Mechanisms and Mechanical Philosophy*. New York: Routledge, 332–347.

Mallo, M. and C.R. Alonso. 2013. The regulation of *Hox* gene expression during animal development. *Development* **140**:3951–3963.

McCole, R.B., J. Erceg, W. Saylor and C.-T. Wu. 2018. Ultraconserved elements occupy specific arenas of three-dimensional mammalian genome organization. *Cell Reports* **24**:479–488.

McGinnis, S. and T.L. Madden. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **32**:W20–25.

McGinnis, W., R.L. Garber, J. Wirz, A. Kuroiwa and W.J. Gehring. 1984. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* **37**:403–408.

McGinnis, W. and R. Krumlauf. 1992. Homeobox genes and axial patterning. *Cell* **68**:283–302.

Melnikov, A., A. Murugan, X. Zhang, T. Tesileanu, L. Wang, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30**:271–277.

Millikan, R. 1989. In defense of proper functions. *Philosophy of Science* **56**:288–302.

Neander, K. 1991. Functions as selected effects: the conceptual analyst's defense. *Philosophy of Science* 58:168–184.

Nègre, N., C.D. Brown, L. Ma, C.A. Bristow, S.W. Miller, et al. 2011. A *cis*-regulatory map of the *Drosophila* genome. *Nature* **471**:527–531.

Niu, D.K., and L. Jiang. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochemical and Biophysical Research Communications* **430**:1340–1343.

Pearson, J.C., D. Lemons and W. McGinnis. 2005. Modulating *Hox* gene functions during animal body patterning. *Nature Reviews Genetics* **6**:893–904.

Palazzo, A.F. and E.S. Lee. 2015. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics* **6**:2.

Park, D., Y. Lee, G. Bhupindersingh, and V.R. Iyer. 2013. Widespread misinterpretable ChIP-seq bias in yeast. *PLoS One* **8**:e83506.

Pennisi, E. 2012. Genomics. ENCODE project writes eulogy for junk DNA. *Science* **337**:1159–1161.

Pertea, M., A. Shumate, G. Pertea, A. Varabyou, F.P. Breitwieser, Y.C. Chang, A.K. Madugundu, A. Pandey, and S.L. Salzberg. 2018. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology* **19**:1–14.

Piasecka, B., P. Lichocki, S. Moretti, S. Bergmann, and M. Robinson-Rechavi. 2013. The hourglass and the early conservation models—co-existing patterns of developmental constraints in vertebrates. *PLoS Genetics* **9**:e1003476.

Piatigorsky, J. 2007. *Gene Sharing and Evolution: The Diversity of Protein Functions*. Cambridge: Harvard University Press.

Ponting, C.P. 2017. Biological function in the twilight zone of sequence conservation. *BMC Biology* **15**:1–9.

Quinn, J.J. and H.Y. Chang. 2016. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics* **17**:47–62.

Roy, A.L., and D.S. Singer. 2015. Core promoters in transcription: old problem, new insights. *Trends in Biochemical Sciences* **40**:165–171.

Roy, S., J. Ernst, P.V. Kharchenko, P. Kheradpour, N. Negre, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**:1787–1797.

Schwartz, Y.B., D. Linder-Basso, P.V. Kharchenko, M.Y. Tolstorukov, M. Kim, … and N.C. Riddle. 2012. Nature and function of insulator protein binding sites in the *Drosophila* genome. *Genome Research* **22**:2188–2198.

Scott, M.P. and A.J. Weiner. 1984. Structural relationships among genes that control development: Sequence homology between the *Antennapedia*, *Ultrabithorax*, and *fushi tarazu* loci of *Drosophila*. *Proceedings of the National Academy of Sciences USA* **81**:4115–4119.

Stamatoyannopoulos, J.A. 2012. What does our genome encode? *Genome Research* **22**:1602–1611.

Stroud, J.T. and J.B. Losos. 2016. Ecological opportunity and adaptive radiation. *Annual Review of Ecology, Evolution, and Systematics* **47**:507–532.

Struhl, K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology* **14**:103–105.

Teytelman, L., D.M. Thurtle, J. Rine, and A. van Oudenaarden. 2013. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences USA* **110**:18602–18607.

Vogel, S. 2003. *Comparative Biomechanics: Life's Physical World*. Princeton: Princeton University Press.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10**:57–63.

Ward, L.D. and M. Kellis. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**:1675–1678.

Wardle, F.C. and H. Tan. 2015. A ChIP on the shoulder? Chromatin immunoprecipitation and validation strategies for ChIP antibodies. *F1000Research* **4**:235.

Wreczycka, K., V. Franke, B. Uyar, R. Wurmus, S. Bulut, B. Tursun, and A. Akalin. 2019. HOT or not: examining the basis of high-occupancy target regions. *Nucleic Acids Research* **47**:5735–5745.

Wright, L. 1973. Functions. *Philosophical Review* **82**:139–168.

Yanai, I. and M. Lercher 2019. What is the question? *Genome Biology* **20**:289.

Yue, F., Y. Cheng, A. Breschi, J. Vierstra, W. Wu, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**:355–364.