Genomics, "Discovery Science," Systems Biology, and Causal Explanation: What Really Works?

Eric H. Davidson

Perspectives in Biology and Medicine, Volume 58, Number 2, Spring 2015, pp. 165-181 (Article)

Published by Johns Hopkins University Press
DOI: 10.1353/pbm.2015.0025

Genomics, "Discovery Science," Systems Biology, and Causal Explanation

what really works?

Eric H. Davidson

ABSTRACT Diverse and non-coherent sets of epistemological principles currently inform research in the general area of functional genomics. Here, from the personal point of view of a scientist with over half a century of immersion in hypothesis driven scientific discovery, I compare and deconstruct the ideological bases of prominent recent alternatives, such as "discovery science," some productions of the ENCODE project, and aspects of large data set systems biology. The outputs of these types of scientific enterprise qualitatively reflect their radical definitions of scientific knowledge, and of its logical requirements. Their properties emerge in high relief when contrasted (as an example) to a recent, system-wide, predictive analysis of a developmental regulatory apparatus that was instead based directly on hypothesis-driven experimental tests of mechanism.

Explanation in Developmental Biology

In my field, animal developmental biology, and in what could be regarded as its "deep time derivative," the evolutionary biology of the animal body plan, there

Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125.

The concepts underlying GRN theory and practice have largely been developed in collaboration with Isabelle Peter of Caltech. NICHD has supported our GRN initiatives from their inception, and it is the author's pleasure to acknowledge this invaluable support, in particular grant HD05753, and the program headed by James Coulombe. The author is extremely grateful to Ellen Rothenberg of Caltech and to Isabelle Peter for their own marvelous perspicacity in their comments upon drafts of this essay.

Editor's note: The author died unexpectedly on September 1, 2015, shortly after completing the manuscript as it is printed below.

Perspectives in Biology and Medicine, volume 58, number 2 (spring 2015): 165–181.

exist two kinds of experimentally supported causal explanation. These can be described as "rooted" and "unrooted." Rooted causal explanation provides logical links to and from the genomic regulatory code, extending right into the genomic sequences that control regulatory gene expression. The genomic regulatory code ultimately determines the developmental process in a direct way, since subsequent developmental events all depend directly and specifically on what regulatory genes are being expressed and where in the developing organism they are being expressed (Peter and Davidson 2015). Here the term "regulatory genes" strictly denotes genes encoding the transcription factors that control each phase of developmental function by specifically mediating the expression or repression of genes. Transcription factors have the unique ability to read DNA regulatory sequence and so to specify the parts of the genetic apparatus to be deployed. (RNAs can also read DNA sequence, but turn out to play relatively minor roles in development.)

By unrooted explanations, I mean those in which the only causality is to be located within a process considered, for example within a synthesis pathway (without reference to why the enzymes are expressed where they are in the first place), or within a signaling event (without reference to why the signal is expressed in the sending cells, or what it does to gene activation in the receiving cells). The tangent to the curve presently describing publication of results in developmental biology would show that most newly generated literature in this field is oriented toward unrooted explanations (if there is any explanatory causality included at all). Nonetheless, it can be predicted that research programs confined to unrooted explanation of developmental processes are going to become endangered species. Unrooted research will inevitably become extinct, or be relegated to boutique applications, by the comparatively enormous explanatory power of rooted explanation. The prediction is therefore that for any scientific contributions in basic science aimed at mechanistic conceptual understanding, an increasingly stringent requirement will be explanation rooted directly in genomic regulatory sequence. As we see later in this essay, the evidence is already in that such explanations are experimentally accessible, and that they fit the bill in the sense of conveying intellectually satisfying chains of causality extending from genome to the observed processes of development.

Throughout my own career in bioscience, I have exploited the explanatory effectiveness of hypothesis-driven scientific logic to explore how animals attain their properties, i.e., express their genomes and develop. Now, new winds are blowing in the realm of scientific inquiry, as they are throughout our intellectual culture. It has been for me thrilling to have participated in the discovery, just over the last decade, that when applied to basic science, the new precepts of systems biology generate a lift that propels science into rooted causal explanation. But other winds blow in contrary directions, challenging some of the intrinsic epistemological aspects of experimental scientific inquiry as they have richly developed over the last 400 years, and most relevantly for us in the last half of the 20th century, when experimental

molecular biology rose to dominance. Alternative approaches to answering these same questions about genome function are on offer, bolstered by powerful institutional support. In the first parts of this essay I have sought to deconstruct some of the diverse precepts of doing science that now populate current literature, in order to reflect on the nature of the functional genomics to which these different scientific practices give rise. There then follows, by contrast, an illustration of the breadth and depth of the resolving power of the new precepts of systems biology, when synthesized with the old precepts of "traditional" scientific logic. Recently Isabelle Peter and I have (2015) shown that it is possible to encompass regulatory, developmental, and evolutionary biology in the single conceptual framework of rooted explanation based on this synthesis.

Genomics

The Prejudice Against Explanation as the Objective of Scientific Research

One might think that in that branch of biology the explicit aim of which is to understand what the genome means and what it does, exemplary relations between explanation, causality, and observation might already prevail. Modern genomics began as an enormous and enormously successful effort to obtain DNA sequence on an organism-wide scale and to create accessible databases thereof, an effort that has most certainly and most irreversibly transformed biology. But unfortunately, as the larger genomics enterprises, mainly government agencies, have expanded their brief to include research into how the genome works and what it means, we see the objectives of making databases being mistaken for the objectives of doing science. The scientific products of this effort are so far devoid of rooted explanations of mechanism, or unrooted explanations of mechanism, and indeed are not actually aimed at any explanations at all. The purview of "institutional" functional genomics of this genre is best summarized in its own words, the objectives overview of the ENCODE project:

> The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. . . . ENCODE investigators employ a variety of assays and methods to identify functional elements. The discovery and annotation of gene elements is accomplished primarily by sequencing a diverse range of RNA sources, comparative genomics, integrative bioinformatic methods, and human curation. Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation (IP) of proteins that interact with DNA and RNA, i.e., modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing. (ENCODE 2015)

Parts List sans Predictability

It is indeed essential to have a parts list. But note that the list of approved approaches excludes any perturbations, such as changing a sequence experimentally to test a functional prediction, or test of regulatory activity by insertion of synthetic expression constructs, which is of course how everything functional that we have learned about animal gene regulation over the last 40 years has been discovered (Peter and Davidson 2015). The methods, attitudes, significance, and interpretation of ENCODE as a scientific enterprise, and as an actual scientific guide to how the genome works, have become controversial and are recently the subject of critical discourse (Doolittle 2013; Eddy 2013; Graur and Zheng 2015; Graur et al. 2011; Kellis et al. 2014; Niu and Jiang 2013), though almost universally the databases generated by the ENCODE consortia are regarded as useful. The general and deep controversies surrounding ENCODE, however, are aside from the fish I have in mind to fry in this commentary. What I am interested in here is the curious epistemological consequences to which the ENCODE enterprise has given rise: witness the following example from a paper entitled "Identification of Functional Elements and Regulatory Circuits by *Drosophila* modENCODE" (2010). Figure 7 of this paper presents an enormous predicted gene regulatory network (GRN) for *Drosophila* development. This "physical" GRN is based on chromatin immunoprecipitation data (ChIP) data obtained with 76 transcription factors for which antibodies were available, in which linkages to about 12,500 target genes were logged, and on computational analysis of conserved "promoter" target site motifs. The paper includes an extensive account of the data acquisition and data reduction procedures underlying the "physical" GRN, and the statistics of the results upon which this GRN is based are to be found *in extenso* in the Supplemental files. A sentence in Discussion notes that the "physical" GRN performed about the same as did a totally randomized GRN in predicting actual network linkages, and deep within the Supplemental data one can learn that the actual ratio of "enrichment," i.e., correct predictions of the "physical" GRN of Figure 7 to those of the randomized GRN, is, to be exact, 1.04. Of course the basic problems are that binding in ChIP does not equal function, and that motif identification does not equal (functional) binding. But imagine the fate of an experimental molecular biologist who would attempt to publish the result of such an enormous effort as a 4% improvement in predictability over random. The point this example illustrates is that in the world of ENCODE genomics, the analysis is to be published on the basis that the measurements, which fall in the approved category, were made and analyzed by sophisticated mathematical statistics; whether the result has any power of predictability is not relevant. In science, conceptual predictability accruing from an analysis is the golden criterion of value, of progress in understanding. Predictability is the lynchpin of scientific knowledge and acts as its ratchet of progress as the tests become more exacting, due to greater and greater available knowledge. Extrapolate this small example from the modENCODE report, and what we see is a rival

enterprise that is technologically (more-or-less) scientific, but is conceptually incompatible with real science. The epistemological criterion that guides science, predictability, is erased.

That "predictability is the lynchpin of scientific knowledge" is one of those metaphors, unusually for metaphors, that expands with perfect accuracy in this context. A dictionary definition of *lynchpin* is "something that holds the various elements of a complicated structure together," so that when it is withdrawn the pieces of the structure lie about in no particular order and display no functional interrelationships. So when predictability is removed as the central criterion of scientific success, the remaining pieces, such as knowledge, relevance to the subject, reality, significance, conceptual import, falsifiability, all lose the roles and relations that constitute the scientific edifice, and they all soon disappear as well. What we are left with is pure observation; as in the 1950s radio/TV detective drama *Dragnet*, "All we want is the facts, ma'am, nothing but the facts."

In responding to some of the issues raised by critics, the principals of the ENCODE project have produced an epistemological argument that theirs is a "biochemical" measure of functionality in the genome, as opposed to "genetic" measures or "evolutionary" measures (by which is meant reference to sequence-dependent conservation across species; Kellis et al. 2014). This argument is doubly fallacious, in that these are not simply innocently alternative ways of looking at the world, as the article proclaims. Genetic and evolutionary criteria of functional significance are directly predictive, because of the mechanistic consequences of mutation and selective resistance to change, respectively. But even the most detailed biochemical measurements on normal (i.e., tissue culture) cells, like *Dragnet*'s "facts," in themselves predict nothing unique or causal. Furthermore, not even mentioned is the main relevant alternative to assessing genomic functionality only by making an approved set of "biochemical" measurements. That alternative is determining the consequences of predictively conceived experimental perturbations of the control state. With respect to regulatory functions of the genome, this is of course how we know everything we know about causes in transcription molecular biology. Imagine trying to learn how this all works by looking only at the controls!

The Essentially Unscientific Consequences

It is interesting how odd is the configuration of the functional genome that emerges from this philosophical starting point. The genome is a new wonderland. All manner of strange and marvelous things remind us of our innocence of what is really out there, as if we were medieval travelers gazing at the starry universe. For example, deep scientific evidence has gradually allowed us to realize that proximal gene order is extremely flexible in animal genomes, and this feature of genomes is subject to continuous scrambling in evolution, within very long-range conserved syntenic scaffolds (Holland et al. 2008), and genes expressed together are not located together (Shoguchi et al. 2011). On the several gene scale, gene order cannot in general be

a functionally important parameter, except for those specific clusters of paralogous genes where the cluster is mechanistically useful, as in *hox* gene clusters or olfactory receptor gene clusters, and animal genomes operate similarly with orthologous genes in differing linear orders. But just as if it were a functional parameter, gene order can be deduced, computed, biochemically measured, described. Because the goal of ENCODE was "systematically mapping functional elements in the human genome at high resolution," whatever is mapped must be presumed functional. Mysterious features emerge, for instance the colorfully named "gene deserts," enormous genomic regions lacking any protein-coding genes. What is interesting is of course just what is not mysterious, but rather what is scientifically illuminating in both functional and evolutionary terms: in this case, the fascinating evidence, which could only have been obtained experimentally, that such "deserts" are packed with *cis*-regulatory modules which act over very long distances (Montavon et al. 2011).

The most directly contentious wonderland outcome of ENCODE epistemology has probably been its highly publicized assertion that 70 or 80% of the human genome is transcribed, which without delay was translated into the direct implication that scientists cannot understand what most of the genome is or does. These assertions were based on mapping of RNA sequence fragments against the genome, eschewing the use of any significance cutoff with respect to prevalence. To its credit, the very same recent paper reminding us of the goal of ENCODE also provides an extremely careful and illuminating quantitative investigation of the biochemical transcriptome measurements on which this remarkable transcriptional coverage claim is based (Kellis et al. 2014; Marinov et al. 2014). This analysis shows that for the major class of RNA subject to quantitative measurement, poly(A)RNA, almost all the transcripts that account for the unexpectedly high genome sequence representation are expressed at levels so extremely low as to be almost certainly meaningless in respect to function, at least for most transcripts. It is the sum of the lengths of these extremely rare transcript fragments that so shockingly exceeds the complexity anyone would assume to be included in pre-mRNA of the productively transcribed 10^4 or so genes that the tissue culture cells are running. As pointed out cogently in this same discussion, there would be a large (and unnecessary) regulatory cost to absolute exclusion of very low-level "noisy" "off" states. Thus, background presence of .01 to <1 molecules of transcript per cell for most of the sequence represented in the 80% transcriptome statistic should not surprise, nor does it in any way suggest functional meaning, rather the opposite. Knowing what we do of transcription and translation rates and of the required functional concentrations of mRNAs encoding even classes of proteins that operate at low numbers of molecules per cell such as transcription factors (Peter and Davidson 2015), there is little likelihood that these reports of very high genome representation of very low frequency transcripts mean anything.

The origins of the decision in ENCODE epistemology to remove as a criterion the lynchpin of scientific predictability are partly internal to genomics, but also

partly more general. This is a complex historical issue. On the internal side, in its nascent period, the immensely expensive and challenging task of genome sequencing was driven forward by technologists, engineers, computational innovators, and organizational geniuses, mainly not by basic bioscientists, and the different purview of these founders of the field has certainly continued to influence the intellectual atmosphere within the genomics enterprise. Externally, as often discussed, the practical economic and political requirements of "big science" have powerfully torqued the coordinates of the guidance systems operating in most parts of government-supported genomics. Among the force vectors most removed from those of basic research have been the necessity of persuading nonscientists of the medical and other life-changing benefits to the human condition that will accrue from the huge government-supported genomics enterprise. But I note these factors only en passant, and here I wish to look in a different direction, in order to take note of a more widespread intellectual movement of which the foregoing can seem only a subcase (though it would be more accurate to describe it as a feedback partner of genomics): the rise of "discovery science."

Discovery Science

A Radical Attack on the Epistemological Backbone of Science

One reasonably representative summary of the precepts of "discovery science" can be found in Wikipedia:

> Discovery science (also known as discovery-based science) is a scientific methodology which emphasizes analysis of large volumes of experimental data with the goal of finding new patterns or correlations, leading to hypothesis formation and other scientific methodologies.
>
> Discovery-based methodologies are often viewed in contrast to traditional scientific practice, where hypotheses are formed before close examination of experimental data. . . .
>
> Data mining is the most common tool used in discovery science, and is applied to data from diverse fields of study such as DNA analysis, climate modeling, nuclear reaction modeling, and others.
>
> The use of data mining in discovery science follows a general trend of increasing use of computers and computational theory in all fields of science. Further following this trend, the cutting edge of data mining employs specialized machine learning algorithms for automated hypothesis forming and automated theorem proving.

A close look at the assertions in this statement, which is quoted in its entirety except for one unnecessary sentence, reveals some quite curious attitudes. If it were not for their effect on public science policy, these attitudes would serve well as a cartoon illustration of belief-based decision making in public affairs.

What "Discovery Science" Really Leads To

First, consider the opening treatment of "hypothesis formation." In basic science, hypothesis formation is an inferential, inductive, creatively novel proposition of mechanism; hypotheses cannot be deduced, which leads to the conclusion that the authors of this statement have redefined "hypothesis" as deductive restatement of what is observed. No deductive restatement of what is observed can have mechanistically predictive value. Needless to say the corollary that "large volumes of experimental data" have the intrinsic property of "leading to hypothesis formation" has no meaning except after this redefinition. It is thus to be expected that here there is no mention of hypothesis testing by determination of the predictability engendered by the hypothesis, or of falsification or validation of the hypothesis. In other words, the object of the exercise, the "hypothesis" of this definition, has nothing to do with actually determining if an assertion has any reality or is just a deductive assertion. In fact, it has nothing to do with the real meaning of hypothesis in science.

Second, the definition states that in "traditional scientific practice" hypotheses are formed "before close examination" of the evidence, i.e., the "experimental data." Perhaps this is just a commentary on the unfamiliarity of the writer of the Wikipedia piece with scientific history in the last century of biological research: from Boveri to Morgan to Crick to Britten, just to leave present company aside, no one knowledgeable could imagine applying to these superb scientists the description that they failed at "close examination of experimental data."

Third, consider the corollary that progress in discovery science occurs by means of "data mining." Since data mining usually consists of application of previously canned databases such as GO Ontologies or off-the-shelf statistical packages, the truth now emerges that discovery science can generate no discoveries. Instead it generates further reorganization of received observation. Thus we see under the rug: discovery science is in essence the same scholasticism that was so painfully overthrown by the stunning predictive success of inductive, hypothesis-driven experimental science in the last four centuries, and in biology in only the last century and a half. Had these tools been available to analyze texts, the machine learning and automated theorem proving algorithms adduced as the way to go in discovery science would have been avidly utilized by late medieval professorial theologians and text redactors.

But nonsense aside, even helpful considerations of how to apply large datasets and how to accelerate true hypothesis testing will not take us where we need to go. There is indeed an epistemological revolution brewing in modern bioscience. The experimental biology of the 20th century amply provided us with rich evidentiary and mechanistic foundations. Only recently, however, has it come within reach to conceive executable blueprints for generating rooted, predictively successful scientific explanations at the necessary scale, in the sense in which this essay began.

Systems Biology

The reason "systems biology" means all things to all users is that neither of its component nouns at present denotes a precise concept. In place of a definition that clearly excludes certain types of bioscience while clearly including others, we have in "systems biology" more a slogan or advertisement than a specific precept or premise with specific corollaries. So for present purposes I propose to generate such a precept, as a framework for what follows. This leans heavily on an earlier discussion in the recently published *Handbook of Systems Biology* (Peter and Davidson 2013). To avoid an amorphous sociological or technological excursion, the following is confined to a particular domain of basic research, namely, processes of developmental biology that are amenable to system-level experimental analysis. In this domain, the specific objective of research is causal explanation of the developmental process, by relating what happens directly to encoded genomic regulatory information. That is, we return to where we began this essay, scientific research the objective of which is genomically rooted explanation.

Basic Corollaries of a Systems Developmental Biology That Explains

The fundamental premises of experimental systems developmental biology are that all processes that can be defined as observable episodes of development are generated through multiple interactions of multiple biologically active components, and that all these components, and all (or almost all) their interactions, must be included in an analysis in order to solve a mechanism that has sufficient predictive explanatory power. From this basic (and uncompromising) starting point, there follow sufficient corollaries to define unequivocally the practice and concepts of systems developmental biology (corollaries are "something that follows from another thing," here specific scientific conceptual relations that follow from the premises of system developmental biology).

Note insertion of the term *experimental* here. The purpose is to exclude from the following discussion purely computational simulations, or imaginary mathematical exercises that purport to represent developmental process in silico on the basis of a priori assumptions of mechanism (Peter and Davidson 2015). The object of scientific research is to arrive uniquely at an underlying explanatory mechanism, which object can obviously not be attained by assuming the mechanism a priori, whatever simulations or correlations the computation may throw up.

The first corollary of systems developmental biology is the separate proposition that if only a minor fraction of the components active in a process, and of their interactions, are encompassed in the analysis, it will be impossible to arrive at a solution of the mechanism. That is, except at infinity no addition of fragmentary unrooted explanations will ever sum to a complete rooted explanation. The basis of this conclusion is also the basis of the pessimistic comment above that unrooted research is sooner or later destined for extinction. The reason is the enormous

combinatorial number of degrees of freedom with which the interactive control systems of development can be, and evolutionarily have been, assembled.

A second corollary is that the usefulness or success or failure or validation or refutation of an explanation emerging from experimental systems developmental biology must be assessed system-wide. These tests must challenge the ability of the explanation to predict the behavior of the system as a whole, or the behavior of large sectors of it that include many individual components and their interactions.

A third corollary defines explanation and phenomenology in developmental biology in terms of operational scientific objectives. It states that phenomenology will inevitably result from research focused exclusively on a very small fraction of the components of a system and their interactions, and thus unrooted explanation inevitably produces phenomenological information. As I often commented in past years, 20th-century developmental biology continuously generated small islands of causality floating in a vast sea of phenomenology. System-wide rooted explanation inverts this relationship, so that when successful it gives rise to a framework of causality, within which are always to be found islands of not yet understood phenomenology.

A fourth corollary relates causality in systems developmental biology to the genome. Since development of the body plan is a species-specific, hardwired output of genomic regulatory information, and genomic information is physically resident in DNA sequence, explanation of a system-wide control mechanism for a developmental process must begin with recognition of regulatory genomic sequences. Hence explanation rooted in multiple elements of genomic regulatory sequence is what emerges from the systems biology of body plan development.

A fifth corollary states the essential roles of correctly observed phenomenology, and prior descriptive knowledge, in systems developmental biology. In the late 20th century descriptive observations of developmental systems were routinely derided as intrinsically inferior to "problem solving." But in the different conceptual coordinates of systems developmental biology, this is a dated opposition: "problem solving" within a tiny localized domain is a hopeless approach to system-wide explanation (see corollary three above). On the other hand, as in the first corollary, until the whole set of components and the output (behavior) of the whole interaction system is correctly measured/observed, it cannot be studied effectively by the precepts of systems developmental biology. Thus high-resolution quantitative and qualitative observation of transcriptional functions in time and space, and many other "descriptive" molecular, cell biological, and developmental aspects are of irreplaceable value as a starting point for a perturbation analysis, so long as they are system-wide.

A sixth corollary, general to all real science, states that only by deliberate experimental perturbation and predictive challenge of the system can the mechanisms by which it operates be revealed. However, applications to systems developmental biology produce specific complex constraints. System-level perturbations require

consideration of secondary and tertiary effects because of the functional interactions within the system, as well as of the effects of multiple inputs at each node of the system. Therefore perturbation analysis in systems developmental biology demands the intellectual guidance provided by the use of hypothesis at every step.

The Many Other Kinds of Activity Denoted by Systems Biology

It will be more than obvious that the vast majority of what is usually denoted "systems biology" has not very much in common with the foregoing. Most strikingly, as practiced in the United States and Europe, systems biology is almost always applied biology, not basic research. It is not supported by small, investigator-initiated research grants, as is most basic bioscience. Rather, it is typically institutionalized in very large organizational units, both academic and .org "institutes," supported on an institutional level by government funding instruments, and/or, by institutionally dedicated endowments and contracts. Various powerful external factors have contributed to this outcome. This is a proper subject for large-scale historical inquiry into current science funding policies in our societies, and their consequences. But there are also internal attributes that have irresistibly caused systems biology to drift in these same directions, among which are pressures for high-powered, expensive instrumentation and computation facilities, genuflection before very large-scale datasets, and the organizational constraints of large collaborative projects that involve a great many authors, many of whom are primarily technological contributors. The overall consequence is that the public trajectory of systems biology is ever more strongly tending toward applied science. Two self-perpetuating feedbacks have emerged, one economic and the other political. The evident economic feedback directly links the size and expense of running large institutions on the one hand, with the availability of enormous government and private support for targeted, applied research of perceived practical (and commercial) benefit on the other. A second kind of feedback mutually reinforces dedicated large group consortia that proprietarily capture targeted technological initiatives, and trendy mega-programs that attract political support outside of science.

Much of the mass of large-scale systems biology research is medically oriented, which could unexpectedly be having undesirable effects on medical research per se. Descriptive, large-scale correlative compilations of physiological and medical measurements are much in vogue for generation of diagnostic reference bases, and the technology required synergizes well with the predilections and capabilities unique to systems biology. We cannot do experiments on humans, and the type of high-powered correlation analysis that systems biology generates for resolving large-scale datasets provides an apparent substitute for experimental extraction of causal relations. Thus we see claims that computational inference from analyses of unperturbed human cell types of medical significance can reveal their underlying gene regulatory networks (Novershtern et al. 2011). In essence, though increasingly sophisticated (and increasingly large and expensive), such efforts are philosophically

akin to "discovery science," here powered by systems biology methodologies that operate under the constraints of direct medical relevance.

For fundamental science, however, all this is very, very far from the beneficial epistemological significance, and the real importance, of the precepts of systems biology. These precepts amount to a potential revolution in the conduct of basic science. In my field, systems developmental biology, the revolution is no longer potential. It is here.

CAUSAL SYSTEMS BIOLOGY THAT WORKS

Proof of Principle: The Long Road to Causal, Predictive, System-Level Explanation in Development

About the year 2000, my lab embarked on a journey never before taken except in imagination. Its objective was to discover and solve, and then try to understand as a system, the encoded genomic regulatory instructions directing a whole large phase of embryonic development. Unlike the transoceanic explorers of the 15th century, we had an approximately realistic idea of what had to lie on the other side if we could get there, but just as it was for those explorers, our means of getting there was incredibly primitive from the vantage point of current times (Wey-Gomez 2008). Theory initially formulated on almost purely logical grounds over 30 years earlier told us that what had to be on the other side would be encoded gene regulatory networks (GRNs, as now generally termed), physically resident in large sets of DNA sequences of regulatory function that could be uniquely recognized by sequence-specific gene regulatory macromolecules (Britten and Davidson 1969). The predicted control system had to have the form of a network, because there was no escape from the deduction that each regulatory macromolecule must control many target genes, but on the other hand each gene encoding regulatory macromolecules must also integrate multiple inputs combinatorially in order for novel regulatory functions to be established in development: hence, the configuration of a network. We thought in 1969 that the regulatory macromolecules would likely be RNAs, but two years later we realized the logic would be the same if they were proteins, as turned out mainly to be the case (Britten and Davidson 1971).

Gene Regulatory Networks

To fast forward, GRNs are now recognized as the canonical control apparatus of developing animal systems. They consist of genes encoding transcription factors and signaling ligands and receptors, plus the DNA regulatory sequences that control expression of these very genes in response to binding of transcription factors at their regulatory sequences. As so long ago inferred, genomic control systems are indeed networks, because of their combinatorial inputs at each regulatory gene and combinatorial outputs with respect to target genes. The function of GRNs is to generate sets of cellular "regulatory states," i.e., where "regulatory states" specifically

denote the sum total of active transcription factors that in each cell at each time in development determine specifically what genes will be expressed. Therefore GRNs determine what functions the cells can and will execute. Regulatory states are expressed dynamically in the embryonic space of the organism, determining its regional functional morphology, thus its development. The single most important attribute of GRNs is that they are hardwired into the hereditary genomic sequence that defines each species, and changes in this sequence define the evolutionary history of each species. GRNs are at the heart of the matter: they are the encoded devices that have the function of transforming the unchanging linear DNA sequence code into the regulatory molecular biology that ends up organizing development. Thus, about 15 years ago when we jumped off on this expedition, we knew that solving a developmental GRN, which had never been accomplished except at a toy level or in the imaginary world, would be an objective of enormous significance; but how to get an experimental grip on the control system for making an organism was anything but transparent.

Decades of experience had proved sea urchin embryos the most accessible to straight molecular biological regulatory analyses of any experimental embryonic system, and once again they did not disappoint. By about three years ago, we had pretty much in hand an experimentally established GRN controlling development of about half the embryo (its mesoderm and endoderm) up to gastrulation, including the interactions among about 50 regulatory genes over a 30-hour period (Davidson et al. 2002; Oliveri, Tu, and Davidson 2008; Peter and Davidson 2009, 2010, 2011; Peter, Faure, and Davidson 2012). We invented GRN theory as we went, and have recently formalized this whole area of systems developmental biology from our own and others' work (Peter and Davidson 2015). We also had to create the technology and algorithms needed for perturbation and regulatory state analyses on the scale required. The six corollaries above that for us underlie systems developmental biology, hammered out in this journey of discovery, now serve as canonical guides, as GRN analysis extends to more and more of the embryo, and more stages, and more embryonic systems.

How We Know It Works

There is no resting upon laurels in living science, and I end this essay with a brief retrospective on the outcome of the huge epistemological challenge with which experimental solution of the sea urchin embryo endomesoderm GRN then confronted us. For though we had gone to all lengths to build as complete a GRN as evidence warranted or suggested, that is scarcely a test of sufficiency of explanation. Therefore we determined to transform the GRN into a computational engine that according to the GRN structure would generate predictions of when and where in the embryonic endomesoderm every regulatory gene should be expressed or not expressed. Then we could compare these computational predictions to direct experimental observation: the comparison would show whether the GRN indeed

suffices to explain the spatial and temporal patterns of gene expression, or how much of these patterns the GRN explains. This is not the place to review in any detail the ways and means by which the necessary computational model was built, as this has been discussed amply in the literature (Faure, Peter, and Davidson 2013; Peter and Davidson 2015; Peter, Faure, and Davidson 2012). In a few words, we used the input/output results that in the solved GRN indicate what regulatory gene products are required for spatial and temporal expression of every regulatory gene included, and based on these requirements we built logic equations for each gene. These statements captured all we knew of the conditions upon which the encoded regulatory apparatus of each gene would generate expression or silence (in the case of repression). The expression state of each gene was computed every hour in in silico time, and expressed as a boolean 1 or 0, which could be directly compared with observation, as in parallel, the real world observed by in situ hybridization and quantitative transcript measurements showed genes to have "on" or "off" expression states in each region of the embryo.

The overall result is shown in Figure 1, reproduced from this work (Peter, Faure, and Davidson 2012): here the regulatory genes appear along the top; the four colors represent four spatial/developmental fate domains of the embryo; and time proceeds in one hour intervals within each. The green bars under certain genes denote the maternally encoded roles of these gene products in setting the initial regulatory conditions. Where there is a grey cell, the computation and the observation agree the gene should be off in that time/space cell. Where there is a solid color cell, the computation and the observation agree the gene should be on in that time/space cell. Where there is a black symbol there is a discrepancy between what the GRN predicts and what is observed. As the figure shows, there are >2770 time/space cells, and the GRN successfully predicted the activity state in all but a tiny fraction of these cells. (The occasional open black symbols indicate possible discrepancies, which however are only temporal and are less than the resolution of the observations, so they can be ignored.) There were a few places where genes are observed to turn off (striped boxes), for which the GRN provided no explanation, and a few real temporal discrepancies, but no spatial discrepancies throughout.

What We Can Conclude

Major conclusions follow: First, this result provides proof of principle that systems developmental biology can generate a (nearly) compete causal explanation of observed gene expression patterns. Second, that proof of principle explicitly refers to explanation rooted in genomic regulatory sequence function, which is therefore where the causal program for development resides. Third, there is room for no other levels of explanation than what is encoded in the regulatory DNA of the regulatory genes of this system; i.e., other aspects of gene expression machinery (e.g., epigenetic) operate downstream of the DNA sequence recognition system. Fourth, the computational system runs as an automaton, in that it utilizes at each

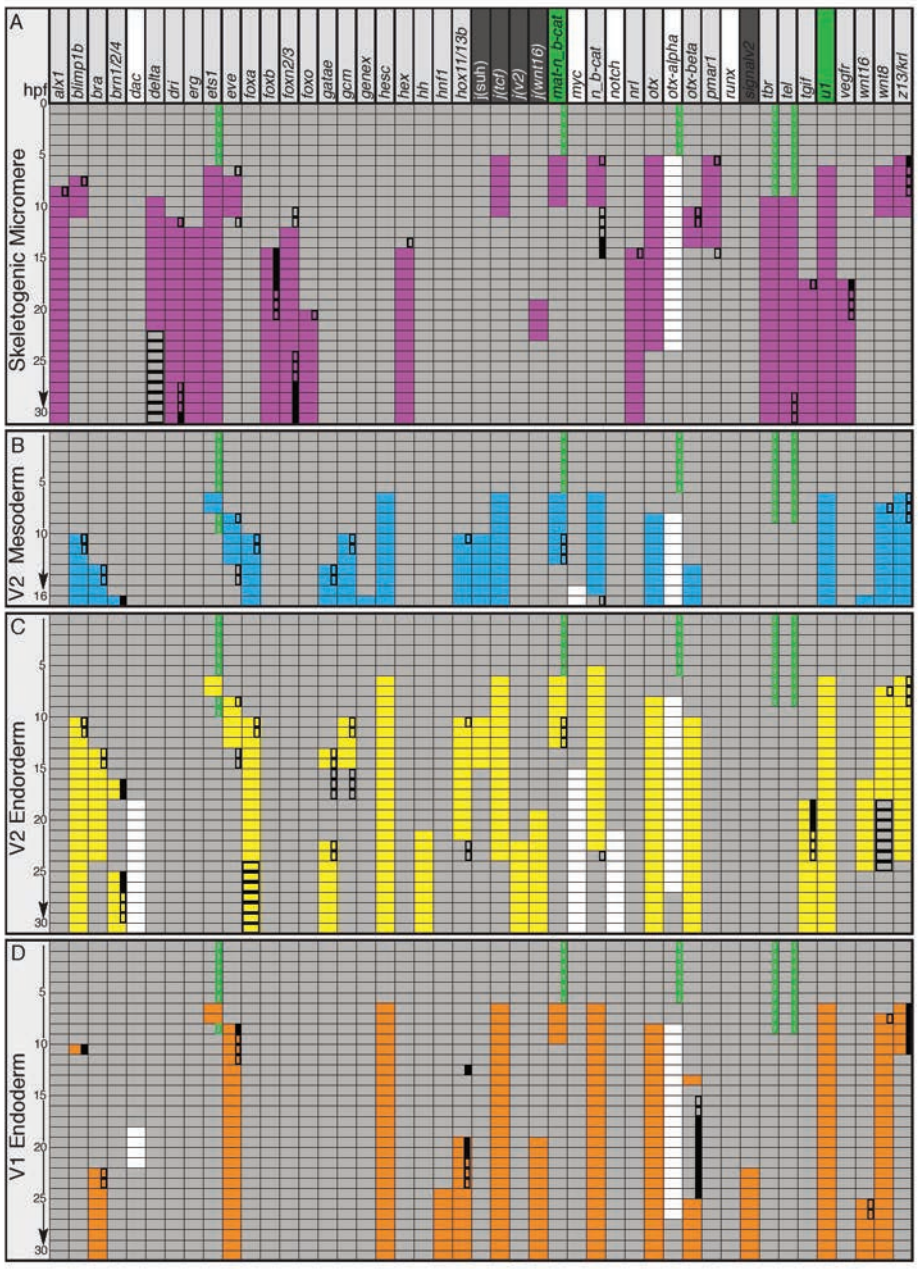

FIGURE 1

Computational test of GRN sufficiency (Peter, Faure, and Davidson 2012). Computation of hourly spatial expression versus data: out of 2,772 space/time expression domains, there are very few significant discrepancies.

interval the output at the previous interval, and once initiated operates progressively with no added external inputs. This is of course how development works in life, and also how computers work. It is worth a moment's reflection that the only configuration that could display the properties of an automaton is a network of regulatory genes: the requirement for an automaton is that the outputs of elements of the system (transcription factors) are also the inputs to the elements of the system.

So in sum, this most fundamental functional output of the animal genome, development, can indeed be experimentally accessed and solved at a system-wide causal level. This is new, and indeed it constitutes a revolutionary change in our ability to generate comprehensive scientific knowledge. But what is not new, and what still lies uniquely at the heart of real science, including basic science done at the system level, is the inductive use of predictive hypothesis.

References

Britten, R. J., and Eric H. Davidson. 1969. "Gene Regulation for Higher Cells: A Theory." *Science* 165 (3891): 349–57.

Britten, R. J., and Eric H. Davidson. 1971. "Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty." *Q Rev Biol* 46 (2): 111–38.

Davidson, Eric H., et al. 2002. "A Genomic Regulatory Network for Development." *Science* 295 (5560): 1669–78.

Doolittle, W. F. 2013. "Is Junk DNA Bunk? A critique of ENCODE." *Proc Natl Acad Sci USA* 110 (14): 5294–5300.

Eddy, S. R. 2013. "The ENCODE Project: Missteps Overshadowing a Success." *Curr Biol* 23 (7): R259–61.

ENCODE. 2015. *ENCODE: Encyclopedia of DNA Elements*. https://www.encodeproject.org/.

Faure, E., Isabelle S. Peter, and Eric H. Davidson. 2013. "A New Software Package for Predictive Gene Regulatory Network Modeling and Redesign." *J Comput Biol* 20 (6): 419–23.

Graur, D., et al. 2011. "On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of ENCODE." *Genome Biol Evol* 5 (3): 578–90.

Graur, D., et al. 2015. "An Evolutionary Classification of Genomic Function." *Genome Biol Evol* 7 (3): 642–45.

Holland, L. Z., et al. 2008. "The Amphioxus Genome Illuminates Vertebrate Origins and Cephalochordate Biology." *Genome Res* 18 (7): 1100–11.

Kellis, M., et al. 2014. "Defining Functional DNA Elements in the Human Genome." *Proc Natl Acad Sci USA* 111 (17): 6131–38.

Marinov, G. K., et al. 2014. "From Single-Cell to Cell-Pool Transcriptomes: Stochasticity in Gene Expression and RNA Splicing." *Genome Res* 24 (3): 496–510.

modENCODE Consortium. 2010. "Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE." *Science* 330 (6012): 1787–97.

Montavon, T., et al. 2011. "A Regulatory Archipelago Controls Hox Genes Transcription in Digits." *Cell* 147 (5): 1132–45.

Niu, D. K., and L. Jiang. 2013. "Can ENCODE Tell Us How Much Junk DNA We Carry in Our Genome?" *Biochem Biophys Res Commun* 430 (4): 1340–43.

Novershtern, N., et al. 2011. "Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis." *Cell* 144 (2): 296–309.

Oliveri, P., Q. Tu, and Eric H. Davidson. 2008. "Global Regulatory Logic for Specification of an Embryonic Cell Lineage." *Proc Natl Acad Sci USA* 105 (16): 5955–62.

Peter, Isabelle S., and Eric H. Davidson. 2009. "Modularity and Design Principles in the Sea Urchin Embryo Gene Regulatory Network." *FEBS Lett* 583 (24): 3948–58.

Peter, Isabelle S., and Eric H. Davidson. 2010. "The Endoderm Gene Regulatory Network in Sea Urchin Embryos Up to Mid-Blastula Stage." *Dev Biol* 340 (2): 188–99.

Peter, Isabelle S., and Eric H. Davidson. 2011. "A Gene Regulatory Network Controlling the Embryonic Specification of Endoderm." *Nature* 474 (7353): 635–39.

Peter, Isabelle S., and Eric H. Davidson. 2013. "Transcriptional Network Logic: The Systems Biology of Development." In *Handbook of Systems Biology,* ed. Marian Walhout, Marc Vidal, and Job Dekker, 211–28. Amsterdam: Elsevier.

Peter, Isabelle S., and Eric H. Davidson. 2015. *Genomic Control Process: Development and Evolution*. London: Academic Press.

Peter, Isabelle S., E. Faure, and Eric H. Davidson. 2012. Feature Article: "Predictive Computation of Genomic Logic Processing Functions in Embryonic Development." *Proc Natl Acad Sci USA* 109 (41): 16434–42.

Shoguchi, E., et al. 2011. "Direct Examination of Chromosomal Clustering of Organ-Specific Genes in the Chordate Ciona Intestinalis." *Genesis* 49: 662–72.

Wey-Gomez, N. 2008. *The Tropics of Empire: Why Columbus Sailed South to the Indies*. Cambridge: MIT Press.