The Concept of the Gene in Development and Evolution

Historical and Epistemological Perspectives

Edited by

PETER J. BEURTON
Max Planck Institute for the History of Science, Berlin

RAPHAEL FALK
Hebrew University, Jerusalem

HANS-JÖRG RHEINBERGER
Max Planck Institute for the History of Science, Berlin

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain

First published 2000

Printed in the United States of America

Typeface Palatino 10/13 pt. *System* MagnaType™ [AG]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication Data

The concept of the gene in development and evolution: historical and epistemological
perspectives / edited by Peter J. Beurton, Raphael Falk, Hans-Jörg Rheinberger.
p. cm. – (Cambridge studies in philosophy and biology)
Includes bibliographical references and index.
ISBN 0-521-77187-0
1. Genes. 2. Genetics – Philosophy. 3. Developmental genetics. 4. Evolutionary
genetics. I. Series. II. Beurton, Peter, J. III. Falk, Raphael. IV. Rheinberger, Hans-Jörg.
QH447 .C66 2000
572.8′38 21 – dc21 99-042106

ISBN 0 521 77187 0 hardback

Contents

PART ONE. GENES AND TRAITS

PART TWO. EXTRACTING THE UNITS OF HEREDITY

PART THREE. GENETIC PROGRAMS AND DEVELOPMENTAL GENES

PART FOUR. CONCEPTUAL PERSPECTIVES

1

The Dissolution of Protein Coding Genes in Molecular Biology

THOMAS FOGLE

ABSTRACT

The consensus gene, a methodological outcome of the rapid growth in molecular biology, is a collection of flexibly applied parameters derived from features of well-characterized genes. Broad flexibility unites research programs under one umbrella and simultaneously promotes the false impression that the molecular gene concept is an internally coherent universal. This suggests limitations for genomic interpretations of information content in biological systems and for explanatory models that use genes as a manipulative. Genomic referencing, the development of systemic relationships among DNA domains, will more fully interconnect molecular genetics to biology than the molecular gene alone. Advances in understanding regulation and expression of DNA and the current interest in large-scale sequencing will necessarily supervene on much of the attention currently bestowed on molecular genes.

INTRODUCTION

The gene concept, long regarded as a unit of inheritance, undergoes continuous transformation to accommodate novel structures and modes of action. A little more than a decade after the rediscovery of Mendel's work in 1900, new analytical strategies emerged for mapping genes as loci in a linear array on a chromosome. During the 1940s, the one-gene – one-enzyme model revealed that genes act to generate specific cellular products, a precursor to the science of molecular genetics. In the years that followed, the gene underwent further change. First, the double helix model of DNA made famous by Watson and Crick revealed the physical structure for particulate inheritance. Later efforts clarified the biochemistry of gene expression.

Today, in the era of genomic sequencing and intense effort to identify sites of expression, the declared goal is to search for genes, entities assumed to have physical integrity. Ironically, the sharper resolving power of modern investigative tools make less clear what, exactly, is meant by a molecular gene, and therefore, how this goal will be realized and what it will mean.

The legacies of particulate inheritance, localization through mapping and the Central Dogma, shape current perceptions of the gene. Although the empirical details are elaborated today, molecular genes retain an imprint from the past. In a previous paper (Fogle 1990) I analyzed the difficulty with continued attempts to bridge the gap between the Mendelian gene as a "unit of inheritance" and molecular genetics. Text-style definitions strain to find coherence when they incorporate language from both eras. Generic definitions, and hence what I termed "generic" genes, lack internal consistency.

Here, I view the problem through a different lens. The identification of a molecular gene does not stem from definitions. It is a methodological process. Genes are recognized by formally or informally comparing elements of structure, expression, and function to those previously documented. Properties and physical elements for the molecular gene concept have broad social acceptance in the community of molecular biologists. For example, detection of an RNA product serves as strong evidence that there exists a site of transcription, a gene, that acts to generate the RNA. RNA is one component from a collection of consensus features found commonly among well-described genes.

The criteria necessary to anoint new genes require research programs to adopt a community structure that places value on particular chemical states, events, and conditions while accepting considerable flexibility on how to apply them. Flexibility is essential because the large (and growing) array of molecular conjunctions prevents a strict application of rules for the molecular characterization of a gene. The need to bring a set of empirical results in line with other claims for genes forces research programs to emphasize different features in different situations or for different purposes. Molecular genes, then, are best understood as a general pattern of biochemical architecture and process at regions that actively transcribe the product of an ongoing development of consensus building in the face of rapidly changing empirical evidence. Hence, I term this shared interpretation to be a "consensus" gene.

At present, there is strong momentum to absorb new molecular revelations into the consensus gene rather than effect a more fine-grained description of molecular parts and processes. The problem is analogous to that of evaluating when a related group of organisms should be clustered into one taxonomic group or splintered into several. The outcome, sometimes contentious, rests on the analysis of shared characters in relation to established taxa. A desirable outcome is to achieve a widely accepted taxonomic solution for the purpose of efficiently characterizing the biology of that group. In taxonomy, lumping different elements into a single taxon may impede deeper biological and/or evolutionary insights. Similarly, forcing diverse molecular phenomena into a single Procrustean bed, i.e., the gene, implies a universal construction. Therefore, the gene as a molecular vehicle for causation is an ambiguous referent. I explore the difficulties arising from the embrace of the consensus gene and discuss heuristic limitations of the gene concept.

THE PROBLEM WITH MOLECULAR GENES

The consensus gene is an abstraction of molecular detail, a socially generated model for what a gene is supposed to be, formed through the expected parts and processes that empiricists associate with it. Genes are identified by seeking a fit, or at least a partial fit, using empirical evidence at hand against the backdrop of an idealized construct, a consensus gene. The process supports genic claims of different entities with shared properties.

The consensus gene, a summary of the cellular route for expression, acts through production of RNA products that may or may not be translated into polypeptides. Function and structure are inseparable. Even when genes are identified strictly from physical readouts of the DNA sequences, functional significance is inferred by analogy to more fully characterized molecular sites that have similar organizational motifs. For example, detection of a common promoter sequence known as a TATA box, a binding site for the enzyme necessary to initiate transcription, signifies a nearby site of expression. By inference, the presence of the TATA box indicates that neighboring

DNA harbors the potential to produce a transcript with functional significance for the cellular system. Hence, the TATA box is a structural component, a consensus feature,[1] contained within a gene. A consensus gene, in its stereotypical format, places importance on the localized segment of DNA that forms the transcribed region. Additional nucleotide strings (elements) may reside externally or internally with respect to the site of transcription. In addition to TATA boxes, a variety of domains are essential for gene activation and regulation. Among other roles, domains bind RNA polymerase, the enzyme that copies one of the two strands of DNA to form a complementary sequence of RNA. Eukaryotic cells can process newly formed RNA by cutting and removing internal sections known as introns. Most eukaryotic genes have introns, sometimes several dozen. Coding regions, termed exons, are spliced into a contiguous piece of mature RNA ready for translation into a polypeptide at a ribosome. Bordering the coding region, or open reading frame (ORF) of the mature RNA, is an untranslated leader sequence at one end and a trailer sequence at the other. Start and stop codons flank the coding message.

The consensus gene implies a high degree of uniformity among genes and seems, at first glance, to be an internally consistent description of parts and action. However, no simple description embodies the breadth of molecular genes claimed by empiricists (see also Carlson 1991; Falk 1986; Fogle 1990; Kitcher 1982; Portin 1993). Therefore, it is impossible to retreat to abstraction about genes without masking the diversity within. The consensus gene is a framework, not a full elaboration of biochemical detail. To what extent does an outline of its principal components and interactions generalize? I will show that consensus mode of molecular biology struggles uncomfortably to unite disparate phenomena under one banner, the gene.

GENES AND THEIR PRODUCTS

The consensus gene embraces multiple products from a single locus. One way this can occur is with sliding edges. Another is through combinatorial splicing of the transcript.

Some genes have two or more staggered promoter sites that form distinct transcripts encoding different polypeptides. The human dystrophin gene (D'Souza et al. 1995) has at least seven promoters; each regulates expression in a tissue-specific manner, leading to production of polypeptides that vary markedly in size. The many products are considered to arise from a single gene, not a set of different genes that share many parts.

In addition to sliding edges on the transcript, multiple polypeptide products can result from alternatively spliced RNA molecules. Many examples of combinatorial splicing among subsets of exons are known (see Hodges and Bernstein 1994).

In deference to the Mendelian tradition, there is resistance among geneticists to subdivide a region into multiple genes when the variant products share functional relatedness and occupy a single locus. By centering the genic claim around localization of a DNA site for expression and functional significance for the cellular system, fuzzy borders or multiple products can be tolerated.

Despite differences in form, loci that produce multiple products share much of their biochemistry for expression. The relationship between DNA coding and polypeptide formation occurs through a recognizable and common set of events. The continuity of pattern and mode binds production of many products under one linguistic construct, the gene. The embedded familiarity reinforces the central framework of the consensus gene.

The consensus gene readily absorbs convoluted twists on the traditional route to production of a functional product, as demonstrated by "inside out genes" (Tycowski, Mei-Di, and Steltz 1996). Usually spliced exons contain coded information and introns are nonfunctional. The transcript of the U22 snoRNA host gene is processed as usual to remove the introns (nine in the human form and ten in the mouse form) and splice the exons into a segment of mature RNA. The spliced RNA, however, lacks coding ability whereas the introns form RNA constituents of the nucleolus, a nuclear structure that participates in the assembly of the ribosome. Unlike all other genes studied to date, processed introns are functional and spliced exons are not.

The consensus gene of molecular biology embraces the "inside out gene" as new in form, not new in kind. It retains nearly all the structural and biochemical activities of protein coding genes except translation, and except the many types of functional RNA that are processed. The "inside out gene" widens the biochemical modes of expression attributed to the molecular gene. As the consensus gene accommodates new molecular events like the "inside out gene," it must incorporate more contingencies into its fold.

SOLUTIONS TO THE ONE-LOCUS – MULTIPLE-PRODUCT DILEMMA

The molecular revelations from multiple products and biochemical novelties suggest two alternative solutions. Either enlarge the constellation of biochemistry for the gene or propose narrower guidelines for genic ascription. Even prior to the discovery of inside out genes, there was no agreement in the literature on whether multiple functional products from a localized segment of DNA should be considered more than one gene. Lewin (1994) argues that we can reverse the usual statement "one-gene – one-polypeptide" to "one-polypeptide – one–gene." He is emphatic in stating that these are "overlapping" or "alternative" genes.

Lewin's claim is a re-evaluation of the meaning of the gene, yet he is uncommitted to pursuing its implications or upsetting the current paradigm. The implications to the molecular genetics community are substantial. Taken at face value, Lewin's proposal would require a revision of the nomenclature system for thousands of loci as a consequence of his call for a more refined relationship between functionality and a gene. It would also profoundly influence estimates of gene number for humans and most other eukaryotes. Lewin does not discuss either the methodological or ontological consequences. He is clearly ill at ease with the consensus gene that readily accepts multiple products. I suspect that he is applying a Band-Aid to a problem, one that he considers worthy of further reflection, but not one that he takes too seriously.

A more widely held perspective is that polypeptide "isoforms," proteins with nearly the same amino acid structure derived from one expression site, originate from a single gene (for example, Strachen and Read 1996). Here, similarity in structure and function of the products suggest a natural grouping into one causal unit. For those cases in which polypeptides are very different, an indicator of functional divergence, some authors recommend subdividing a site of expression into separate genes (Alberts et al. 1994). How different do the polypeptides have to be to split the locus into more than one gene? Molecular biologists do not quantitatively evaluate polypeptide divergence for this purpose. Like Lewin's call for gene splitting of alternatively spliced RNA products, the recommendation to discriminate types using the polypeptide and/or function is an ad hoc solution to situations that do not fit a one-gene – one-product model. The solution is offered more as a helpful suggestion than as a committed proposal to redefine the gene.

Defining "genes" by working backward from the polypeptide is a slippery venture. Many polypeptides undergo post-translational modifications into a functional form. Conventionally, genic identity correlates with the primary product of translation. Post-translational changes in structure are secondary effects of cytoplasmic interactions with the polypeptide. If function becomes a dominant criterion for the task of mapping the locus, as Alberts et al. recommend, then translation no longer serves as a boundary condition. This is not the intended consequence of the proposal. Their hope is to clarify parameters for a gene. Instead, they expand possible interpretations.

Several examples will show how problems arise with their proposal. A variety of post-translational modifications have been documented. After translation, some polypeptides, particularly neuropeptides or hormones, subdivide by proteolytic cleavage. Polyproteins are consistently regarded as products of one gene, whether or not they cleave into identical or divergent forms. For example, the DNA locus for the alpha factor regulating mating behavior in yeast (Fuller, Brake, and Thorner 1986) encodes a translated polypeptide clipped into four identical peptides. In contrast, an ascribed gene in silkworms produces five functionally distinct products (a diapause hormone, pheromone biosynthesis activating neuropeptide, and three other neuropeptides) cleaved from a 192 amino acid precursor (Xu et al. 1995), each an independent functional unit.

Alberts et al. do not distinguish between subdivided polyproteins and polypeptides generated by alternative splicing, yet both can give rise to more than one functional form. The consensus gene is their salvation. By advancing the importance of function, imposing it as a tool for evaluating the expression site when needed, they can sidestep the problems that result if one hardens the rules and applies them to every case. They build a molecular case for a gene using a select cluster of consensus components with structural and transcription and/or translation elements. Alternative spliced variation takes place after transcription but prior to translation, two tightly entrenched processes for the protein coding model of the gene. In contrast, post-translationally formed polyproteins lie beyond the physiological boundary of gene-associated biochemistry. For DNA loci encoding polyproteins, translation is a boundary condition that makes functional distinctions unnecessary. Both lumpers and splitters of genes draw the same sharp line in the sand. Polypeptides formed directly from translation are qualitatively different from polypeptides that undergo post-translational modifications. Both views cling tightly to biochemical mechanisms to locate the gene. Function is not a universally important criterion; it gains or loses importance in a particular case against the backdrop of other consensus elements (structural and/or biochemical).

The comparison between genes encoding polyproteins and alternative spliced products suggests a set of parameters for interpreting well-characterized sites of the consensus gene. Three properties with variable weight designate a molecular gene: localization to a transcript-generating segment of DNA, physiological boundaries located at pre-translational (alternative splicing) compared to post-translational (polyprotein cleavage) activity, and an investigator-based assessment of functional divergence among products. Whether a DNA site constitutes a gene depends both on empirical evidence for that case and subjective emphasis of the parameters.

A closer look at expressed sites indicates that this appealing triangulation of conditions provides little help toward rigorously articulating molecular properties for the gene. Translation does not always neatly divide the origin for variation between pre- and post-translation, creating an additional source for a many-to-one relationship between the molecular phenotype and a locus in DNA. The mammalian gene governing S-adenosylmethionine decarboxylase (AdoMetDC) has two ORFs. The short form codes for a hexapeptide within the leader sequence of the larger AdoMetDC coding section (Hill and Morris 1993). The hexapeptide down-regulates AdoMetDC translation in a tissue specific fashion. The investigators avoid confusion about which is the "real" gene by subordinating the smaller ORF as a regulatory element of AdoMetDC. Once again, the chosen rhetoric is consistent with the consensus view that readily accommodates novelties of form and process in the gene concept. In this instance it also forces the investigators to choose referents to identify the gene. That is, the AdoMetDC coding region could be semantically repositioned as the trailer sequence for the production of the hexapeptide. Hill and Morris select the AdoMetDC coding region as the referent for the gene because they place greater functional importance upon it.

Both polypeptides of AdoMetDC are primary products of translation and could be viewed as separate genes. The investigators seem to unpack the consensus gene as follows. The two polypeptide products originate from a common site of transcription and a common transcript that, until translation, unifies their biochemistry of structure and action. The multiple polypeptide products, through their combined presence, effect one functional end. The smaller polypeptide is viewed by the investigators as a product of a regulatory region for the larger polypeptide. The larger polypeptide is, therefore, accorded the role of the principal gene product (a decarboxylase enzyme) through prior recognition of its importance to cellular physiology.

In this instance, the genic claim is, in a classic sense, a unit of function. The mechanics of transcription and translation are sufficiently similar to other claims for genes to warrant support at the molecular level. Consistent with the consensus gene model, Hill and Morris reconfigure the coding site for two polypeptides into one gene. Many other similar cases could be described.

The assignment of one gene for the CCAAT/enhancer binding protein (C/EBP) of vertebrates (Calkhoven et al. 1994) entails an even more careful choreography of semantics. The messenger RNA contains three ORFs, each starting at a different point along the transcript. The product of the shortest ORF regulates the ratio of product from the two longer, overlapping coding regions that have different start sites for translation. When Calkhoven et al. discuss the nucleotide sequence specific for one of the two large ORFs they choose the term *cistron,* a unit of function. This allows them to avoid attaching a gene label to the three coding regions, eliminating the need to declare whether they are working with three genes or one. Once again, unity of function, against the backdrop of a common molecular biology, provides a serviceable means for representing C/EBP as one gene. The interpretative rendering by Calkhoven et al. and others demonstrates how context, a normative mode for the consensus gene, impacts what is reported to be a gene and how difficult it would be to develop an internally consistent systematic taxonomy for genes.

In this section I have attempted to show how multiple products from one locus conspire to force arbitrary decisions about whether one or more genes are represented. The real problem is that there has been a steady creep of new genetic twists that must either be accepted as part of the structure and biology of the gene or abandoned in favor of an alternative description of molecular events. The reluctance to abandon the molecular gene, and instead, work around problems as they arise, erodes coherence. One might ask when told of a newly discovered molecular gene, "what kind? – one that produces a single product? multiple products? multiple products that have very different functions? functional isoforms? multiple products formed during transcription? or processing? or translation?"

GENES AND CODING

The translational assembly interprets the genetic code. After transcriptional processing removes introns and splices exons, RNA is read in tandem triplets of codons. Each codon specifies an amino acid in the growing polypeptide. For some RNAs there are other mechanisms for readout (Gesteland, Weiss, and Atkins 1992). In some cases, the ribosomal assembly skips from one to fifty nucleotides in the RNA, shifting the reading frame before continuing. In other instances, the meaning of the code changes to read, for example, a stop codon, a polypeptide termination signal, in place of an amino acid. A particular physiological state, not just the transcript itself, causes the translational change. Either form of recoding partially shifts informational specificity for the product into the cytoplasmic space, removed from its usual habitat in the sequence of nucleotides of DNA.

The translational machinery is often likened to a computer reading software, an ungainly metaphor with only superficial similarities. The DNA, imagined as a bit stream of computer code read by hardware, sends a template copy of RNA to the site of translation and threads the sequence through the ribosome to read nucleotides in consecutive blocks of three. With cellular recoding, the cytoplasm is rewriting the software program produced by the DNA. The cellular architecture itself contains an information coding ability that becomes apparent during translation.

The coding regions are not the only portion of the transcript that direct the form and function of the product. Leader and trailer sequences that border the ORF of a transcript are crucial in RNA recruitment for translation (Sonenberg 1994) and also regulate activation or rates of translation. For example, the insulin-like growth factor gene (IGF-II) forms two mature transcripts with an identical coding region and trailer sequence but leader sequences of different lengths (Nielsen and Christiansen 1995). The shorter transcript participates in protein synthesis while the longer form complexes with protein to become a ribonucleoprotein particle, a component of a ribosome. One functional product is translated, the other not.

The IGF-II gene produces qualitatively different functional products, an RNA and a polypeptide, that share a common transcript and DNA locus. One might expect that functional divergence of the IGF-II gene products would lead to a claim for two genes since cases of functional relatedness (AdometDC and C/EBP genes) led to a claim for one gene. The IGF-II locus demonstrates a different set of problems and a similar approach to a solution. To unite radically different end products as components of one system requires that other consensus properties compensate. IGF-II becomes a normative gene by ignoring conflicts with the standard outline for translational events and placing emphasis on pre-translational events.

The cellular biochemistry of gene structure and expression consists of a set of contingent statements substantially larger than molecular biologists, such as Lewin or Strachen and Read, or Alberts et al. seem to admit. Gene-splitters run the risk that any post-transcriptional modifications of RNA altering the polypeptide product, any novel variation of translation, any post-translational chemical modifications of a polypeptide, map to a separate gene. Equally difficult to justify are conservative renditions for genic enumeration that de-emphasize function and read different physiological constructs of RNA or polypeptides as members of one gene. The genes of research programs, as opposed to generic descriptions in texts, form a continuum of material forms and processes. There are no discrete functional packets or molecular mechanisms in the protoplasm to serve as guides for delimiting a gene.

From this unsettling outcome, a molecular gene lacks demarcation without at once specifying the temporal and spatial cyto-complex of the system. Accordingly, the dynamics of the system, more than just the sequence and a vague notion about function, characterize a gene. The route geneticists choose to move past this roadblock, as we have seen, is to craft ad hoc solutions to subsets of problems (such as Lewin's one-polypeptide – one-gene proposal). In this way, flexibility is maintained and genic definitions should be read only as statements about common gene patterns, e.g. most genes have introns that are nonfunctional and most have exons that are functional.

One purpose for a flexible gene concept is to link molecular phenomena to Mendelian genes. Of course, Mendelian analysis does not depend on knowledge from molecular biology. During the expansion of genetics as a discipline, when genes lacked physical description of the sort assigned today, one widely held viewpoint was to liken them to beads on a string. The very success of molecular biology in the 1950s and beyond solidified perspectives about the gene to a physical reality of one type – a site on the DNA. Much of the history of molecular biology reified that the Mendelian gene can, in principle and in practice, be described in molecular terms. As details poured forth from an expanding research enterprise in molecular biology, the molecular gene acquired greater contingency without formally abandoning the Mendelian gene.

GENES AND THEIR BORDERS

Although the same outline of molecular biology has been used to argue for a different number of genes, the most common approach is to claim that multiple products due to transcription processing, translational reading, or post-translational interactions are endpoints of a single agent, a gene whose physical origin lies in a section of

DNA. If so, its residence should have property lines. One approach is to fix the position for the genic site through the generation of a transcript. The other is to locate the gene using both the site of transcription and regions that regulate it.

In molecular biology, the term *expression* denotes active production of an RNA transcript and is an indicator for the presence of a gene. But to map the genic site through its DNA complement in RNA is to ignore either post-transcriptional cytochemistry which modifies the transcript prior to translation or pre-transcriptional sites and events that proscribe what becomes the transcript. For example, post-transcriptional changes caused by the cytoenvironment can change a noncoding intron into a coding exon through a change in the splicing pattern, preempting a simple means for crafting a molecular referent for the physical structure of the gene through a primary transcript. Pre-transcriptional influences directed by neighboring DNA elements, such as multiple promoters discussed earlier, extend this concern. The informational locus on the DNA dictating the transcript formed, or its ability to form, resides in both neighboring elements and in physiological conditions at the time of expression.

Many types of elements have been described (for example, silencers and enhancers); each is a short length of DNA that affects the timing or rate of transcription. Their position and number per gene is highly variable. The activation of protein coding genes is a stepwise series of interactions between protein and multiple DNA elements upstream from the start of the transcribed region. An assembly of more than a dozen globular proteins attracts RNA polymerase to the promoter to initiate transcription. The interplay of multiple DNA elements and multiple proteins is a key regulatory mechanism for gene expression. Therefore, many recent descriptions for a molecular gene include DNA elements within their borders, even at the expense of clarity about limits and boundaries. Lodish et al. (1995) state that a gene is the "entire DNA sequence necessary for the synthesis of a functional polypeptide or RNA molecule" (p. G-8). Similarly, Alberts et al. (1994) consider the gene to include the "entire functional unit, encompassing coding DNA sequences, noncoding regulatory DNA sequences, and introns" (p. G-10). Note the juxtaposition of the Mendelian language of "functional unit" and the molecular language of "DNA sequences." The consensus gene results from a struggle to hold on to the past and represent the present. Methodologically, domains are treated in much the same way as other aspects of the consensus model. Despite clear proposals either for or against inclusion of elements as part of the gene, they are included or excluded to fulfill specific needs.

A gene concept that includes all DNA domains connected to the ability to express is not to be taken literally. Surely the inclusiveness does not mean that a substantial fraction of the genome is a domain of each gene because *in vivo* activity of every gene is interdependent on the products of many others. The claim is much more restricted and localized; a molecular gene produces a transcript together with other regional domains. Yet detected interactions at even the local level suggest complex relationships among domains. Individual or joint synergistic effects of elements on expression can act like a rheostat dialed to the lowest active setting to produce transcriptional effects barely above a detectable level or magnify the rate of expression. And empirical limits complicate the proposal; it is unrealistic to experimentally reference every regional genomic domain with respect to all others.

We are left with a sketchy framework for determining when a DNA segment is part of a gene. Elements are often judged by whether they affect expression and act in a local manner. If one looks more broadly at a larger swath of the genome, the problems associated with elements are further evident.

The sharing of regulatory elements contributes to the problem of finding physical borders for genes also. The beta globin gene cluster in humans produces five related polypolypeptides that form part of the hemoglobin protein. The locus control region (LCR), is positioned at one end of the gene cluster and regulates their expression in a developmental-specific manner (Wood 1996). The LCR orchestrates the timing of transcription activation and rate. Embryonic tissue produces high levels of epsilon globin and low levels of beta, A-gamma, and G-gamma chains. Fetal cells have large quantities of A-gamma and G-gamma globin and small quantities of beta globin. By adulthood, a small amount of delta globin can be detected and beta globin production predominates. A genic model that includes regulatory sequences cannot deny the LCR as part of its structure. It is clear from the literature, however, that this is not the case. The LCR is presented as a separate domain, neither a component of any molecular genes nor a gene itself. For multiple local transcripts, like the beta globin cluster, regulatory elements are attributed responsibility for functional coordination of globin production. Isolating the LCR from any one gene more accurately conveys the functional relationships among the domains comprising the cluster. It also contradicts definitions which embed domains within genes and reveals that the physical dimensions of genes once again depend on methodological need.

Neither the edges of the gene, its relationship to function, nor its biochemistry of expression are constants that can aid the formulation of a finely characterized molecular gene. However, that genes do localize is an important part of the genic claim.

GENES, PSEUDOGENES, AND GENIC STATES

Additional modalities of molecular variation further erode prevailing views of the molecular gene as an integrated localized structure. Expression in trypanosomes, protozoan parasites, and nematode worms commonly requires trans-splicing a short and a long RNA transcribed from different regions of the genome. The spliced leader is less than a few dozen nucleotides long and not part of the coding region for a polypeptide. Maroney et al. (1995) find that trans-splicing in nematodes is essential for "translational efficiency," subordinating the smaller entity as a contributor to the effectiveness of the larger, coding RNA. The smaller locus forming the trans-spliced transcript does not have protein coding function, although this is not unusual. Many genes transcribe RNA that does not translate. The DNA site that transcribes the leader has the hallmarks of gene structure and expression without the title. It is treated as a buttress for the integrity of the larger RNA companion.

The bacterial genetic system presents a complementary example. To degrade an abnormal protein formed from a faulty coding sequence, two unconnected transcripts expressed at different sites on DNA will form one translated product. The transcript that encodes the ability to degrade the abnormal transcript joins the ribosomal complex, remains unbound with the incomplete message, and attaches a linker amino acid together with an additional ten amino acids coded by its nucleotide sequence (Keiler, Waller, and Sauer 1996). The eleven amino acid tag signals the cell to dispose of the polypeptide. The authors consider each locus, independently transcribed yet cotranslating a polypeptide, to be a separate gene.

The organizational and functional construction of the bacterial and trypanosome loci do not reveal why the former should be a two gene system and the latter a one gene system. Both transpliced RNA and the bacterial degradation system bring together products from two loci into the translational machinery. The bacterial degrading system and trypanosome loci coding each contain a structural and regulatory domain on the transcript. Each could be interpreted as one gene or two. The genic systems of bacteria and trypanosomes lead the investigators to opposite interpretations to give explanatory flow to the empirical evidence. Trans-spliced sections take on a parochial role as a regulator of translation; the larger transcript of the two becomes the central object of inquiry, promoting its functional importance. On the other hand, the polypeptide degrading system of bacteria serves a global function for the cellular system. Therefore both the RNA that does the degrading and the RNA product that requires degradation, are functionally independent, the products of two genes.

The perceived significance of a DNA locus to the cell can be critical to the case by case interpretation of the presumptive gene. When function is unknown, molecular biologists sometimes postulate a contribution to the cellular system from contextual cues. Functional effect operates through expression, the formation of a transcript. Pseudogenes are categorized separately from true genes because they have low rates of transcription. *Alu*-sequences, as one example, are short pseudogenes found in large copy number in the human genome. They have the signature of genes, many capable of transcribing small amounts of RNA with ORFs that do not undergo translation. The dividing lines between whether something is or is not a gene can be thin. Pseudogenes point to a minimum level of expression as necessary but not sufficient.

In addition to variable levels of gene expression, the chemical structure of DNA can act like a toggle switch, alternating between two states that influence gene expression. A string of nucleotides that acquires methyl groups on the cytosine bases can activate or, more commonly, repress transcription. In some cases the methylated pattern, known as imprinting, is preferentially inherited through one sex, resulting in the maternal or paternal expression of specific genes. In others, the pattern of methylation changes between tissues or stages of development. Bartolomei et al. (1993) report a domain of methylation surrounding the transcribed region and promoter of the H19 gene, beyond which they did not detect repression. Imprinting of the H19 gene functions similarly to a regulatory element, a domain often positioned within a gene. Methylation patterns differ from regulatory elements in two ways. They are not sequence specific sites and they can spread over the entire face of the locus, as in the case of the H19 region.

Therefore, at least two classes of transcription regulating sites are present – temporal domains of methylation on nucleotides and heritable DNA strings of nucleotides (enhancers, promoters, and the like). Despite their similarities on physiological effect, their differences point to a key property of genes. Alterations in the methylation patterns differ from mutational changes that take place in DNA strings. Mutations have generational stability, reproduced during the process of replication and transmitted to descendants. Methylation patterns change when specific physiological conditions occur that are not well understood. Therefore, methylation can not be absorbed within the consensus gene as another novelty uncovered through molecular biology.

DNA action and function become meaningful in the context of a cellular system. Coding information in the DNA is necessary but insufficient for the operation of living systems. The mutual dependency of DNA and protoplasmic interactions bedevils a simplistic labeling scheme for expressed segments of hereditary information. The more molecular biology that is unpacked, the greater the need to acknowledge the mutuality of the component parts, forcing arbitrary choices about the physical edges or the physiological properties of the gene. The consensus gene devalues mutual dependency in order to locate hereditary units from a loose and changing confederation of molecular constituents. As a result, the research enterprise can successfully search for genes so long as there is no demand for a rigorous underpinning for their specification.

SEARCHING FOR GENES

Much of molecular genetics research focuses on a search for expression units that do not depend on tight matching to a universal construct. Liberal applications of the gene concept weld research programs into a community dedicated to a common mission. For this purpose, a molecular gene is a useful instrumental tool.

There are, however, consequences for vague notions about molecular genes. Talk of genes plays a major role in the intellectual advancement of evolutionary biology and organismal development. If genes are contextually dependent for structural and functional evaluation, then it is unclear how a fully realized, or at least a richly detailed, theoretical presentation would be possible using genes as an explanatory manipulative.

Coding information acts within a codependent cellular setting: localized sites of expression interacting among DNA domains and contingent upon genomic composition. Here, the term genome means more than the collective set of molecular genes of the organism; it refers to the rich tapestry of DNA domains that weave a pattern of expression. Genetic information is layered within ordered, structured chromosomes. Genomic analysis is expanding rapidly, and will unveil integration among domains positioned far apart, an outcome foreshadowed by trans-splicing. From this broader vantage point on the entirety of genetic information in an organism, domains of action and regulation connect locally and distantly positioned DNA loci into a functional network. As sophistication in the understanding of biological relationships of DNA domains increases, explanatory constructs will subordinate genes as instrumental constructs and increasingly emphasize systemic interactions to communicate insight into cellular processes.

There are many examples, too numerous to document here, that demonstrate positional and contextual integration of function at all levels of genomic organization: coordinated regulation of gene families, loops of chromatin that regulate clusters of genes, distinctive sequence patterns within chromosome banding patterns, and regional functions at the tips and centromeres of whole chromosomes. How these levels of organization cooperatively orchestrate information has yet to be explored, largely because of experimental limitations. Cellular activity requires this hierarchy of genomic information in addition to each locally acting hereditary site (a "gene," however defined). This is more than just a problem for the molecular taxonomy of genes; mutual sites of interaction interpenetrate the genome at many levels, much of which is left on the sidelines when gene number is equated with genomes and genomes with information content.

A molecular gene is a coarse parameter for genomic analysis, poorly suited for the future growth of empirical results. It may be possible to count the number of ORFs or the number of alternatively spliced products or the number of DNA sites producing primary transcripts, but it will not be possible to conduct an exact gene count, at least not without resorting to the consensus model.

The goal of the human genome project is to find the 60,000–80,000 genes, a number based on three methods of estimation. Each method plucks some parameter from the consensus gene as a tool for estimation. But, because the consensus gene is a fluid concept, the derived values are themselves a crude statement about the genome. With most of the DNA sequenced at the time of this writing, one can scan this sample of genetic information and search for the structural hallmarks of a gene (ORF, TATA box, etc.) and extrapolate from an average density of one gene per 20,000 bases to arrive at an estimate for the total number of such sites (70,000 genes). A second measure assesses the number of kinds of expressed RNAs (cloned as complementary copies of cDNA) in different tissues to determine gene number (65,000 genes). Since no one tissue expresses all genes and, as we have seen, alternative splicing and other mechanisms can produce more RNAs than local expression sites, this too is a rough estimate. The third method relies on counting the number of CpG islands, regions often surrounding promoters that have a higher density of neighboring CG bases than the rest of the genome. Slightly more than half of all expressed loci have CpG islands. The counting method offers no indication of the number of gene products or their function. Estimates for the number of genes (80,000) are consistent with the other methods. The three methods, collectively, suggest that somewhere between 65,000–80,000 loci in the human genome fit the standards of the consensus gene. Note that each method is successful for the purpose of approximation and cannot be applied to a fully resolved cellular system, because to do so would require many subjective evaluations of the sort discussed earlier. Even genomes with completed sequences (a nematode, yeast, and many species of bacteria) are evaluated for gene number from the DNA readout alone, leaving ambiguities from gene action unaddressed.

Reporting gene counts, particularly for the human genome, is more than an empirical exercise. It is intended as a scale of information content essential for normal function. The thinking goes like this: Genes are functional units; thousands of functional units are present; the expression of the phenotype is significantly impacted by these thousands of units of genetic activity; the set of these genes tightly mirrors what is meant by a genetic contribution to the phenotype. The failure to successfully proscribe universal genic borders or events for expression calls into question the significance of gene counts for higher organisms. What new insight would result from discovering that there are twice as many genes as thought, or half as many? And some genes have no detectable function. On the other hand, knowing how many domains of a particular type are present might be helpful (e.g., how many CpG islands), indicative of the cellularwide importance of a specific mode of molecular interaction.

Mosaic architecture and activity among claimed genes greatly limits meaningful inference about information content, molecular activity, and functional effect. Suppose, for the moment, a complete human DNA sequence were available. It would be possible to scan the genome through a computer search for the number copies of particular domains and collections of domains, some of which might match those DNA strings currently recognized as genes. It would require many hair-splitting choices using the consensus gene as a guide. And to what end? The real advantage to detailed genomic sequencing will be to make sense of the functional contribution from combinations of domains, not to label lots of loci as genes. As valuable as they are for reductionist evaluations of the genetic contribution to a trait, they limit the potential to integrate a conceptual framework for large-scale complexity within living systems. Genomic analysis will lead to further insight about the distribution of expres-