1

Genes and Traits

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