Hist. Phil. Life Sci., 26 (2004), 59-80

Molecular Epigenesis, Molecular Pleiotropy, and Molecular Gene Definitions[1]

Richard M. Burian

Virginia Polytechnic Institute and State University
Department of Science and Technology in Society
Blacksburg, VA 24061-0126, USA

ABSTRACT - Recent work on gene concepts has been influenced by recognition of the extent to which RNA transcripts from a given DNA sequence yield different products in different cellular environments. These transcripts are altered in many ways and yield many products based, somehow, on the sequence of nucleotides in the DNA. I focus on alternative splicing of RNA transcripts (which often yields distinct proteins from the same raw transcript) and on 'gene sharing', in which a single gene produces distinct proteins with the exact same amino acid sequence. These are instances of *molecular pleiotropy*, in which distinct molecules are derived from a single putative gene. In such cases the cellular and external environments play major roles in determining which protein is produced. Where there is molecular pleiotropy, alternative gene concepts are naturally deployed; *molecular epigenesis* (revision of sequence-based information by altering molecular conformations or by action of non-informational molecules) plays a major role in orderly development. These results show that gene concepts in molecular biology do, and should, have both structural and functional components. They also show the need for a plurality of gene concepts and reveal fundamental difficulties in stabilizing gene concepts solely by reference to nucleotide sequence.

KEYWORDS: alternative splicing, gene concepts, gene identification, gene sharing, genotype-phenotype mapping, molecular epigenesis, molecular pleiotropy

> 'Genes are not autonomous entities. Rather, they interact with other genes and gene products to make pathways and networks.'
> (Gilbert 2003, 691)

Introduction

There has been considerable debate recently about the status of molecular concepts of the gene (e.g. Beurton, Falk and Rheinberger 2000; Falk 1986, 1995, 2001; Griffiths 2002; Griffiths and Neumann-Held 1999; Hall 2001; Morange 2001; Moss 2003; Neumann-Held

[1] The original version of this paper was presented at the International Society for History, Philosophy, and Social Studies of Biology in a symposium organized by Karola Stotz entitled 'Representing Genes: Testing Competing Philosophical Analyses of the Gene Concept in Contemporary Molecular Biology'. I thank Dr. Stotz, my fellow symposiasts, and the audience for constructive comments. Thanks also to Scott Gilbert for helpful discussions and to Rafi Falk, Joram Piatigorsky, Bob Richardson, and Karola Stotz for helpful comments on a late draft.

2001; Portin 2002; Snyder and Gerstein 2003; Waters 2000). A good number of authorities hold that no exact molecular definition of the gene or molecular criteria for delimiting genes can serve the needs of molecular biology in general, let alone the various disciplines with which molecular biology is allied.[2] I share this view because I am committed to the idea that the criteria used to identify or delimit genes usually combine structural features of the genetic material with phenotypic or functional effects of those materials across cellular or organismal generations (The genetic material has an impact on what is inherited, and genes cannot be properly understood independently of this impact). This view has the consequence that the structure of the relevant nucleic acids alone is not sufficient for specifying or delimiting genes, a position that, I think, is now fairly widely held.[3] If this view is correct, scientists are often clear about what they mean when, in context, they talk about particular genes and they have fairly intuitive and natural ways of communicating across disciplinary barriers without confusion. But it is nonetheless not possible to specify *the* structure of genes in terms of nucleic acid sequence alone. Furthermore, attention to the phenotypes and functions in terms of which genes are delimited helps to explain some of the ways in which scientists and popularizers fall into traps when they conflate, for example, evolutionary gene concepts with traditional molecular gene concepts referring to segments of DNA that, supposedly at least, contain the information specifying the amino acid sequence of a polypeptide as produced on the ribosomes.

If delimitation of genes depends on identifying phenotypes or functions of interest, then it is necessary to achieve a clear understanding of how phenotypes and the relevant biological contexts are delimited before one can hope to work out a determinate account of gene structure. Phenotype delimitation depends, at least in part, on the problems, traits, or functions of interest to different scientists and different disciplines and on the available background knowledge. Furthermore, if genotypes and phenotypes are to be brought into

[2] 'Today the gene is not *the* material unit or *the* instrumental unit of inheritance, but rather *a* unit, *a* segment that corresponds to *a* unit-function as defined by the individual experimentalist's needs' (Falk 1986, 169). See also Rheinberger (2000).

[3] Lenny Moss's term 'gene D' purports to treat nucleic acid sequences independent of their functions. If gene D is intended to refer to any arbitrary nucleotide sequence, fair enough. But if it really is meant to pick out a nucleotide sequence that is, in some sense, available as a developmental resource (which seems to be how he uses the term at least some of the time), then I think that there are some issues to work out about whether even Moss's use of the concept of a gene D is completely free of commitment to functional criteria. For more on this topic see the appendix to Burian (2005 b).

relation with each other, phenotype delimitation also turns on the specifics of the biology involved – and those specifics are typically extremely complex. In this paper, using only examples from eukaryotes, I deal mainly with a subset of molecular phenotypes connected to synthesis of polypeptide chains and the roles those chains play in different contexts (including genes for transcription factors). This leaves aside major cases of interest, such as genes for ribosomal and transfer RNAs, many genes for regulatory controls of gene expression (though not genes for transcription factors), and many genes delimited in work on development, evolution, or medical genetics. There are plenty of biological complexities in the restricted group of cases of interest here, including, for example, alternative splicing of sequence-identical RNAs different physiological circumstances in the nuclei of cells of different cell types, and trans-splicing of materials from different 'raw' transcripts, yielding one mature mRNA.[4] These phenomena, and many more (which add endless complexity to the full story), are described in great detail in up-to-date textbooks of molecular biology or genetics. Yet more complexity arises from additional processing that occurs after an mRNA has been exported from the nucleus. This includes post-translational splitting of polypeptide chains, splicing of polypeptide chains from different sources, and alteration of the conformation and composition of chains. Yet further, chaperoning and the blocking of chaperoning (Rutherford and Lindquist 1998) affect the conformation of polypeptide chains, enable such chains to perform different functions in different physiological circumstances (Li and Lindquist 2000), and mark some cells containing chains with 'aberrant' conformation for destruction. But we have enough on our plate without considering these additional complexities in any detail.

Like many others, I have argued (Burian 1985, 1993a, 1993b, 1995, 1997, 2000) that biochemists, evolutionists, developmental biologists, molecular biologists of various stripes, pharmacogeneticists, regulatory geneticists, etc. deploy different means of delimiting genes that often do not map well on one another. One reason for this is that the scientists in question study different phenotypes and functions by different experimental modalities (see also Falk 2000; Rheinberger 2000). And, again like many others (e.g., Fogle 2000; Griffiths and

[4] For a technical description of some of the many types of alternative splicing found in the human genome and some of the tools used to study alternative splicing, see Croft, Schandorff *et al.* (2000). A related website, http://isis.bit.uq.edu.au/a_splicers.html, supplies a fuller account of the many variant classes of splicing.

Neumann-Held 1999; Moss 2003; Neumann-Held 2001; Portin 2002), I have argued that the complex and tangled pathways by means of which polypeptide chains are manufactured already make a clear general structural definition of genes impossible.

Complexities on the path from DNA to Polypeptide Chains

In this section I explore a new line of argument to support the claim that a clear general structural definition of genes in terms of nucleotide sequences alone cannot handle the complex and tangled pathways by means of which polypeptide chains are manufactured. My starting point is drawn from the epigraph of this paper: 'Genes are not autonomous entities. Rather, they interact with other genes and gene products to make pathways and networks' (Gilbert 2003, 691). This quotation serves as a marker for the clear-cut victory of epigenesis in developmental biology. There are, importantly, different, albeit interrelated, senses of epigenesis here,[5] one rather vague, but belonging to a deep tradition, another explicitly molecular. The traditional meaning may be formulated roughly as follows: epigenetic processes, which are required for development, are not determined by the contents of the fertilized egg alone, but are due to interactions among genes, gene products, and contingent cellular and environmental features that affect the determination, differentiation, or formation of cells, tissues, and organs, or specify the identities of cells or of major features (e.g., secondary sexual characters) of organisms.[6] Very often such changes (for example, changes in brain development or secondary sexual characters) are not reversible, but are highly canalized in Waddington's sense.[7]

In one of the recently-elaborated molecular senses of the term, epigenetic changes are changes of DNA or DNA packaging (such as methylation of DNA and alteration of histone or chromatin structure) that are stabilized in cells or cell lineages and that systematically affect

[5] For a more detailed and refined account of four senses of 'epigenesis' and related terms, see Müller and Olsson (2003). For a volume that presents a major review of historical and contemporary issues about epigenesis, see Van Speybroeck, Van de Vijver, and de Waele (2002).

[6] Wilhelm Roux (1885, 427) wrote that 'If…development occurs essentially by interaction between many or all parts [of the fertilized egg], the fertilized egg needs to consist of only a few different parts, which by mutual interactions gradually generate a great complexity. Development in this case essentially is *production of complexity, epigenesis* in our sense' (as translated in Sander 1991, 3). Granting that the interactions of the parts depend also on environmental inputs and conditions, this is a useful statement, revealing the roots of the molecular meanings of *epigenesis* in the traditional concept.

[7] This usage conforms to Waddington's introduction of the term 'epigenetics' in Waddington (1940). It traces back, of course, to the much older embryological notion that new structures in an embryo develop from an originally undifferentiated mass of living matter.

gene expression *without any change in nucleotide sequence*. Such changes are, in principle reversible, although they often have irreversible effects on the development of a particular organism or lineage of organisms.[8] Molecular epigenesis seems firmly established at this point. Even for the clearest examples of molecular genes such as those traditionally thought to specify polypeptide sequence, epigenetic change ensures that nucleotide sequence alone is not, in general, sufficient to predict whether a polypeptide product will be produced or, if it is, what the resulting sequence of amino acids will be.

The complexities already listed, such as alternative splicing, systematic silencing of DNA by methylation and various modifications of histones, have thoroughly disrupted the notion that the DNA encodes information or contains a program that can be read out in any simple way. A cellular context is required for DNA to function, and different cellular contexts extract different information from the same DNA sequence. Furthermore, the pathways and networks into which nucleic acids and their products enter are multi-leveled and are replete with feedback loops that cross multiple levels. Yet further, the physiological and nutritional states of cells, exogenous signals from the extracellular matrix, other cells, or the external environment (such as heatshock[9] or endocrine disruption[10]) alter the networks and can have stable molecularly epigenetic effects with dramatic lifelong consequences for the organism's morphology or physiology (elaborated in one direction by Newman and Müller 2000). Signal transduction modules work as packages, but what they do – what gets transduced to what by a given signal transduction module – is affected by evanescent signals, physiological states, chromatin packaging, timing, temperature, integration of the signal into a variety of larger modules, and much else. Thus the networks have components of strikingly different sorts, and their behavior is affected by intra-cellular, extra-cellular, and external environmental conditions. Nonetheless, *in context*, they determine when a particular segment of

[8] This articulation draws on Reik and Dean (2002) as well as Müller and Olsson (2003). See Jablonka and Lamb (1995) for an important early review of molecular epigenesis and Newman and Müller (2000) for an important article, which helped me think through this paper, that sets the narrow sense of epigenesis into a wider conceptual context and into a discussion of morphogenesis, character origination, and evolutionary change. Note the connection between the importance of timing of epigenetic change and Waddington's well-known concept of canalization: at appropriate stages, a small change, however triggered, may send an organism down a different development channel or pathway than would otherwise have been expected. As is argued in Newman and Müller (2000), epigenetic changes of this sort may be an important means of achieving developmental and evolutionary innovations.

[9] See, for example, Rutherford and Lindquist (1998); Wagner, Chiu, and Hansen (1999).

[10] For a classic exposition of permanent organismal effects of transient exposure to particular endocrine disruptors, see Vom Saal, Cooke *et al.* (1998).

DNA is transcribed, where the raw transcript of that segment begins and ends, how it is spliced – in short, what is made of it even at the level of nuclear RNA. To repeat a key point: although most network components are found within the cell, some are external to the cell (for example, hormones and other active compounds circulated within the body) and others come from outside the organism (for example, ambient temperature, the circadian light cycle, and exogenous endocrine disruptors taken in by the mother). And thanks to the importance of timing, large developmental effects may result from epigenetic changes such as those due to seasonal changes in timing of ecological events or some form of external stimulus or input.

All of this fulfills (in then-unimaginable ways) an old pre-Mendelian vision that haunted the dialectic between genetics and embryology for much of the last century.[11] Here, for instance, is an articulation by Hans Driesch in 1894:

> Insofar as it carries a nucleus, every cell, during ontogenesis, carries the totality of all primordia; insofar as it contains a specific cytoplasmic cell body, it is specifically enabled by this to respond to specific effects only... When nuclear material is activated, then, under its guidance, the cytoplasm of the cell that had first influenced the nucleus is in turn itself changed, and thus the basis is established for a new elementary process, which itself is not only a result but also a cause (Driesch 1894, my translation).

Molecular Epigenesis and Molecular Pleiotropy

I now explore a more radical consequence of the way in which this old vision has been filled in, drawing specifically on what we have learned about the difficulty of getting from DNA sequence to amino acid sequence in eukaryotes. I begin with what seems to be a simple terminological point. As we will see, it clears away a crucial expository difficulty. The use of databases containing nucleotide sequences is well established. As part of this process, a particular use of gene concepts is codified on the basis of which one can identify various genes and count the number of genes in a given genome. This usage is important and legitimate, but, as I will argue, it employs an impoverished gene concept that cannot serve many of the purposes that gene concepts are supposed to serve. One symptom of the impoverishment of the sequence-based notion of a gene underlying these sorts of gene counts

[11] For references and an elaboration of this point, see Burian (2005a).

is that there are a good number of instances in the literature of scientists who agree that they are talking about the same gene in this sequence-based sense and disagree about whether the gene in question ('really') is a gene or a pseudo-gene. I argue below that we need to work with a plurality of gene concepts and that many legitimate gene concepts would recognize multiple genes within the particular genes picked out by use of the databases. Accordingly, I shall speak of genes as identified by sequence data alone as 'nominal genes'. A good way of parsing the conclusion of my argument is that nominal genes are a useful device for ensuring that our discourse is anchored in nucleotide sequences, but that nominal genes do not, and probably can not, pick out all, only, or exactly the genes that are intended in many other parts of genetic work.

The argument rests on the recognition that there is a trade-off at the molecular level between the criteria for identifying genes in various contexts and the extent to which genes are considered pleiotropic. (The restriction to the molecular level means that to count as pleiotropic a given gene must make more than one molecular product – or, to use a more slippery phrase, make products with more than one molecular phenotype – in different contexts.) My account of how genes ought to be identified is controversial and deserves wider examination than is feasible here. At root, the problem is this: when a nucleotide sequence is considered as a gene, i.e., as a functional entity, its identity *as a gene* is sufficiently sensitive to cellular-context, network embedment, and delimitation of functions that, in typical cases, we should think of the identity of the gene as context-dependent. For clarity I will use two moderately familiar examples to sharpen the point, focusing on molecular pleiotropy in order to illustrate what I mean by multiple contexts.

Take a standard and fairly typical nominal gene in vertebrates, say, the rat α-tropomyosin gene (see Figure 1).[12] The DNA contained in this gene, however its exact boundaries are drawn, includes nine exons and has three distinct terminal repeating units. The RNA transcripts of this nominal gene are known to be spliced in at least seven distinct ways, regulated by various promotors and enhancers that respond to signals found in different physiological, tissue, and cellular contexts. Here is a brief synopsis of the alternative splicing involved.

[12] This synopsis derives from Figure 5.28 of Gilbert (2003), which is the source of Figure 1; the basic information stems from Breitbart, Andreadis, and Nadal-Ginard (1987).

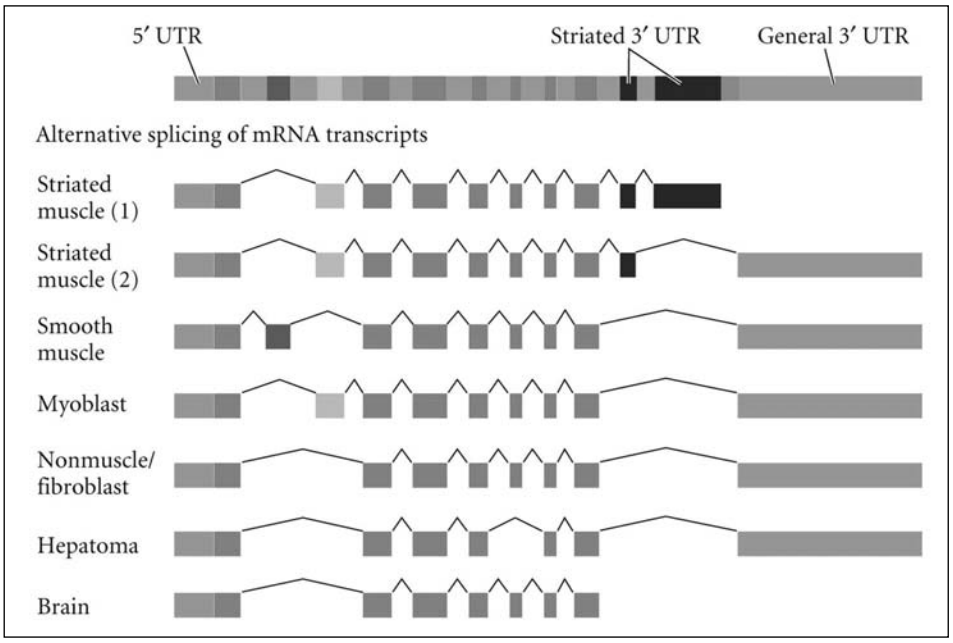

FIGURE 1. Alternative RNA splicing to form a family of rat α-tropomyosin proteins. The DNA sequence is represented at the top. Thin lines represent sequences that become introns and are spliced out in forming mature mRNA. Redrawn from Gilbert (2000, fig. 5.28), after Breitbart, Andreadis, and Nadal-Ginard (1987).

Most of the exons are included in all seven of the common variants of the mRNA produced from this gene. However, the second exon is included in only one of those mRNAs, an mRNA whose protein product is a component of smooth muscle tissue. The third exon is included in only three of those mRNAs (all distinct from the one that includes the second exon). Two of these mRNAs encode products normally occurring in striated muscle and the third in myoblasts. The mRNAs whose products occur in striated muscle have distinct terminal repeating sections, unique to striated muscle. With one exception, all of the other mRNAs use only the so-called general terminal repeating section. The mRNA expressed in myoblasts has the same exons as the striated muscle mRNA, but neither of the terminal repeats distinctive of mRNAs expressed in striated muscle cells. The remaining three common mRNAs include neither the second nor the third exon. The mRNA that yields a product standardly found in nonmuscle fibroblasts is identical to the mRNA found in myoblasts and in smooth muscle except that it includes neither the second nor the third exon. The remaining two mRNAs are produced in hepatomas

(i.e., a particular kind of liver cancer) and brain cells respectively. The hepatoma mRNA is like non-muscle fibroblast mRNA except that it deletes the seventh exon (and nothing else) and the mRNA expressed in brain cells is like non-muscle fibroblast mRNA except that it deletes only the terminal repeat unit.

Thus, of the nine exons, 1, 4, 5, 6, 8, and 9 are expressed in all of the α-tropomyosin proteins, 7 is expressed in all but the hepatoma protein, and the terminal repeats (which do not ultimately appear in protein) are part of what distinguishes how the products are processed and used in the cells in which they occur. At least four of the proteins produced from these mRNAs have distinct amino acid sequences. If genes are identified by the polypeptide sequences they produce, 'the' α-tropomyosin gene should count as at least four genes, producing four products in the same family of proteins. To count it as a single gene is to count it as molecularly pleiotropic, i.e., as yielding distinct molecules in distinct locations. This is not a trivially semantic point; the system is tightly segulated in such a way that the potential impact of mutations in particular exons differs from exon to exon. The differences in the range of cell types in which different exons are expressed places limits on the range of tissues initially affected by a mutation and on the specific direct effects of different variants of the gene product. Indeed, since the second exon is expressed only in smooth muscle cells and the third exon is expressed only in striated muscle cells and myoblasts, mutations in one of these exons should have medically quite distinct classes of effects from mutations in the other. This *might*, in practice, prove to be important for understanding differences between, and potential treatments for, different muscle-wasting diseases and thus is a potentially serious ground for handling the matter in terms of distinct genes. Parallel claims seem justified in most cases in which highly regulated alternative splicing in different cellular or tissue contexts causes nominal genes to produce distinct proteins. The number of cases of this sort is quite large: a standard textbook claim is that approximately 40% of human (nominal) genes are alternatively spliced in different contexts!

The fact that the same raw RNA transcript yields different molecules in different cellular contexts is the result of specific (molecular) regulatory controls that greatly alter the probability of alternative ways of splicing the RNA transcript from one context to another. Almost always, the resultant mRNA is, therefore, context-specific. The result is a form of molecular epigenesis: the different molecular environments encountered by the nominal gene alter the ways in which the DNA sequence is processed (or block it from being

processed), thus producing a product with a different amino acid sequence (or no product). If we knew enough of the contextual details, we could predict the probability of one vs. another of the amino acid sequences being produced. But we could not do so solely from knowledge of the nominal gene; we would need to know the details of the molecular environment. We could also make such predictions from cruder information, not about the molecules, but about the cellular context. If a cell's fate has already been determined, if it has been determined to be, say, a myoblast precursor, by far the most likely α-tropomyosin family member to be generated is the one that includes introns 1, 3, 4, 5, 6, 7, 8, and 9, and it will hook up with other proteins in a way that is typical for myoblasts and not for striated muscle. Thus if we wish to take amino acid sequence or protein product as a molecular phenotype, the nominal gene from which it is produced is, in general, not sufficient to predict the phenotype. *Molecular epigenesis is thus responsible for the molecular pleiotropy of the nominal gene and for the insufficiency of complete DNA sequence information for predicting which product(s) of that gene will be produced in which contexts.*

Narrow Cases of Molecular Pleiotropy

A second class of examples strengthens the idea that it is sensible to think in terms of trade-offs between molecular pleiotropy and criteria for identifying genes. It illustrates how powerfully cellular context determines the function of a polypeptide sequence and the highly distinctive uses to which a single polypeptide product can be put. The issue involves what Joram Piatigorsky (one of the leading experimentalists studying lens crystallins) and his colleagues call 'gene sharing' (reviewed in Piatigorsky 1998; Piatigorsky 2003).[13] I don't yet know enough about the actual biology (which is quite difficult) to answer some obvious questions, but I hope that the importance of the questions *will* be obvious. Gene sharing occurs when 'two distinct protein phenotypes are produced by the same transcriptional unit' (Piatigorsky and Wistow 1991, 1078). In nearly all vertebrates, and many invertebrates, lens crystallins are composed largely of proteins that belong to one or another family of proteins (in the sense in which

[13] The phenomenon of gene sharing, also known as 'moonlighting proteins' turns out to be quite widespread, with many cases now known that do not involve lens crystallins. For a recent review, which I discovered while writing the penultimate draft of this article, see Jeffery (2003).

the α-tropomyosins are a family of proteins). Typically (at least in vertebrates), the lens is composed of several lens crystallins, some of them produced at distinct developmental stages. In invertebrates the proteins belong mainly to families of metabolic enzymes; in vertebrates many crystallins belong to small heat shock protein families and the rest mainly to such metabolic enzyme families as lactate dehydrogenase or cytoplasmic aldehyde dehydrogenase, enzymes produced in the liver. In some species, two or more nominal genes, evolutionary duplicates belonging to the same gene family, are involved. But in many cases organisms of a given species have only one copy of a gene from the family in question. In such cases, there are normally no amino acid sequence differences between the enzyme produced from the gene (e.g., in the liver) and the corresponding lens crystallin in the eye.

In the most extreme cases of gene sharing, the amino acid sequences of two proteins are identical, but the proteins are distinct. This is what I mean by narrow molecular pleiotropy: two distinct proteins (e.g., an enzyme and a lens crystalline) with identical amino acid sequences are derived from one nominal gene. In vertebrates, a significant percent (10%-50%) of the lens crystallins are commonly produced by this kind of gene sharing (Piatigorsky 2003; Piatigorsky and Wistow 1991). The enzymes that share genes with a lens crystallin have been found to be quite diverse; gene sharing of this sort occurs in a wide variety of protein families. Many instances of this sort involve taxon-specific gene sharing, which is to say that in each taxon a distinctive enzyme and a lens crystallin share a gene. The same phenomenon is found in invertebrates, although it is not yet as well studied.[14]

Gene sharing is helpful for the present argument because it provides clear illustrations of a way in which identification of genes is, in practice, dependent upon the functions or phenotypes considered and the specific criteria employed for identifying genes. Assume (as is almost certain) that the conformations of, say, lactate dehydrogenase

[14] Jeffery makes a point about gene sharing parallel to one I made earlier about α-tropomyosin. When a specific polypeptide chain has multiple functions in distinct locations, 'knowing one function of a protein, for example its enzymatic activity, might not fully describe the function of a protein in the cell or organism'. She goes on to add that this impacts rational drug design because 'correcting only one function of a multifunctional protein might not be sufficient to effectively treat a disease' (Jeffery 2003, 33). I add that one could allocate the two functions to one gene or allocate the distinct functions (e.g. for drug discovery purposes) to 'the gene for protein$_1$' and 'the gene for protein$_2$', both contained in the same stretch of DNA and yielding the same polypeptide sequence serving distinct functions in different locations. Which strategy is appropriate might well depend on the task and tools at hand or on the phenotypes under consideration. The differences between a change in protein conformation or function and a change in amino acid sequence is obviously important here, but its relevance to the 'protein phenotype' (Piatigorsky and Wistow's term) and to the delimitation of genes may not be as straightforward as it at first appears.

and the corresponding lens crystallin produced by the same nominal gene are different.[15] The proteins in question are, thus, easily distinguished in spite of having the same amino acid sequence. Furthermore, before the 1950s, even if it had been shown that the amino acid formulae of the enzyme and crystalline were identical, they would have counted as distinct proteins. For it was only in the 1950s (with the work of Sanger and others) that amino acid *sequence* was recognized as a key to identifying proteins. Well into the 1950s (until the work of Sanger and others on amino acid sequences was widely appreciated) the orthodox view was that genes (or cytoplasmic derivatives of genes) serve as templates and that a single substrate could yield several distinct proteins by this template mechanism. Thus, at the time, it was expected that distinct proteins would have the same biochemical formula; such a finding in a particular case would not have led geneticists or biochemists to suppose that the same gene produced the two proteins. If, by some lucky chance, there were appropriate mutations available so that a 1:1 correlation between a mutation in the enzyme and the lens crystallin were observed, that would raise the question whether, surprisingly, the same gene made (or controlled) both proteins and/or whether the proteins were distinct. The chromosomal theory of the gene, but not the pre-chromosomal theories of Bateson or Johannsen (see Burian 2000), would provide a clear-cut test to help resolve this question – the linkage test. If it the two mutations could not be separated in linkage tests, the chromosomal theory would indicate, at least provisionally, that just one gene was involved. But it is quite difficult (using classical techniques) to carry out linkage experiments sufficiently powerful to distinguish very closely linked genes, as is illustrated by the classical example of Sturtevant's inability to separate *eosin* and *white* in Drosophila in an experiment using 150,000 flies (Burian 1985, 32-33; Carlson 1996, 64 and chap. 8).

For this reason, Mendelian genetics would have faced conflicting criteria for delimiting the genes in question in cases of gene sharing. The existence of two distinct proteins in different locations would suggest (insofar as it was thought that genes somehow determine protein structure) that there were two different genes, one producing (for example) a particular lens crystallin, the other lactate dehydrogenase. On the other hand, if the materials were available to carry out a linkage test,

[15] As I understand it, lens crystallins must be very nearly linear and positioned orthogonally to the nearest surface of the lens if light is to be transmitted through the lens to the retina. In order to facilitate the specific reactions that they do, enzymes require specifically shaped pockets. They thus almost certainly have a different conformation than the sequence-similar lens crystallins.

that test would have suggested that a single gene at a single chromosomal location determined the two proteins. Bringing things up to date, one might argue that in cases in which the exact same transcript is produced from a single nominal gene, molecular genetics trumps Mendelian genetics. The identity of the polypeptide sequence, built by standard processes from the same DNA but utilized differently in different cellular contexts, provides a powerful argument for saying that we have one gene, with one product, deployed differently in different contexts. I would argue, to the contrary, that whether we count this as one gene or two can still be couched in terms of how we divide up phenotypes – by amino acid sequence or by functional protein. And it is also probable (although I do not know whether a clear example has been found) that there are instances in which two distinct nominal genes from the same gene family produce sequence-identical polypeptides, for in many species gene duplication and corresponding specialization of function has occurred.

Furthermore, even when only one nominal gene is involved, it is not always clear whether all of the polypeptide sequences of the 'different' products of a given gene have the identical amino acid sequence. It will be very difficult to rule out the possibility that different polypeptide products might result from alternative splicing or other differences in processing. A *prima facie* candidate of this sort is an interesting lens crystallin studied by Piatigorsky's group (Piatigorsky, Norman *et al.* 2001). This is the J3-crystallin of the eye of a particular species of jellyfish (*Tripedalia cystophora*), which appears to be produced from a unique nominal gene (with seven exons) in this species. The gene belongs to an evolutionarily widespread and diverse family of genes that produce proteins called saposins. In *T. cystophora*, the J3-crystallin contains amino acids corresponding only to exons 2-4 of the gene and is thought to be produced by cleaving the amino acids corresponding to these three exons from a longer polypeptide in the eye. The gene is also expressed elsewhere in the jellyfish, e.g., in the tentacles. The literature I have found so far does not resolve the question whether the same three exons (only) are expressed in the tentacles or where else these or other exons from this gene are expressed.[16]

The general point, however, is established, whatever the outcome of

[16] In response to a draft version of this paper, Joram Piatigorsky kindly informed me (pers. commun.) that his laboratory has not yet been able to firmly establish the structure of the J3 crystallin mRNA, whether the crystallin is in fact produced (as is anticipated) by cleavage of a larger polypeptide, or whether the product of the nominal gene is different in the different locations in which the gene is expressed. An additional question that arises in this case is whether some other tag, such as a different leader or terminal sequence, is employed as a marker that determines the way in which the transcript or the subsequently-produced polypeptide chain is processed.

this case. There are many subtle ways in which 'the same' gene yields subtly different molecules in different contexts. The cases of pleiotropy, as Karola Stotz pointed out to me, form a rough gradient of increasingly dramatic pleiotropy. Gene sharing is the mildest form of pleiotropy, at least when the distinct products share a common amino acid sequence. Pleiotropy is more clearly established where there are distinct products with different amino acid sequences. Other cases, not discussed here, are yet more dramatic. Among the examples are instances of overlapping genes, including instances in which a single transcription unit yields very different polypeptide chains because of RNA editing or ribosomal frameshifting (Alberts *et al.*, 2002, 438). Epigenetic factors of very diverse sorts act at every level, even at very narrow molecular levels.

For present purposes, important as it is to understand the different mechanisms of producing distinct proteins, the precise basis for the differences in protein phenotypes produced from a specific RNA transcript does not matter much. Even where we have a clear account of the causes of the protein phenotype differences in particular cases, the problem of molecular pleiotropy remains important. The issue as to whether to count the proteins in question as the products of one pleiotropic gene or of two distinct genes located in the same initial segment of DNA depends on how we classify the molecular phenotypes and what weight we give to the common source of the nucleotide sequences that yield systematically different products in different contexts. This claim does not depend on whether or not there is a difference in the amino acid sequences of the 'different' proteins in the different locations, nor on how the systematic differences in different contexts were brought about.

Semantic issues like this will often arise where there is significant molecular pleiotropy. This is true even in the rather standard case of rat α-tropomyosins. It is an open question whether the seven α-tropomyosins, each produced in specific tissues, are best considered to be produced by one gene or by distinct genes. For some purposes it may be simpler or more sensible to delimit these genes one way, and for other purposes to delimit them in other ways. If this 'easy' case is unpersuasive because the terminology of 'the' α-tropomyosin gene is deeply entrenched, there are plenty of hard cases in which no easy resolution is feasible. The possibility of divergent ways of delimiting genes arises whenever (as is common!) a nominal gene produces molecularly distinct products that perform different functions.

For the sake of clarity, perhaps I should add that I am not yet

arguing about downstream effects, at least not directly. Changes in striated muscle α-tropomyosin can have effects in virtually every part of the body. Thus any mutation that affects the functioning of this protein will have (conventional) pleiotropic effects. Rather, my case is built on *molecular* pleiotropy, the production of distinct molecular products based on the nucleotide sequences found within a putatively single gene (or distinct protein conformations of a particular polypeptide sequence built by readout from the same sequence of nucleotides). The difficulty arises *because the concept of a gene is at least partly delimited in terms of function*. The need to parse the functions of concern opens up the need to introduce conceptual alternatives at just this juncture.

These considerations feed back onto an evaluation of the status of nominal genes. Note that the actual procedures for counting genes covertly take structures into account *because they have known effects connected to gene expression*. We cannot treat *any* particular nucleotide sequence as a gene. Can there be any genes that are one nucleotide long? Three nucleotides long? That consist of a group of 300 repeats near the middle of a highly repetitive short sequence? Criteria that are strictly intrinsic to the DNA, i.e., that are based on structural features of nucleotide sequences alone and do not take into account (perhaps tacitly!) how those features of DNA affect the organism in various contexts offer virtually no prospect of providing widely usable criteria for delimiting genes. For example, requiring that a gene begin with a specific sort of sequence tag, e.g., an open reading frame ('ORF'), does not work. Not only is this structural feature not present in all nominal genes,[17] but ORFs are tacitly chosen as a significant structural feature because they have significant impact on the probability that transcription might be initiated near them in appropriate intracellular molecular contexts. Indeed, while there is no single correct way to delimit genes, the gene concept will lose all value if there are no principles by means of which to answer when to count two stretches of DNA as belonging to different genes and when to count them as belonging to no gene at all. In general, those principles depend (often tacitly) on the phenotypes and functions under study.

[17] There are many instances known in which multiple nominal genes are expressed in one initial raw transcript, starting from one ORF. In many cases the raw transcript is separated into three (or more) separate transcripts or polypeptide chains at some later stage of the processes of transcription, editing, translation, and post-translational processing. The division into three separate genes in such cases is based on knowledge not of the structural features of the DNA downstream from the ORF, but on knowledge of subsequent events and, often, of the existence of the corresponding genetic material in other organisms in separate nominal genes with separate ORFs.

The nucleotide sequences of nominal genes sometimes match closely and sometimes match poorly with those that are relevant to the phenotypes in question. For practical purposes, the boundaries of the nucleotide sequences relevant to a given function must often be revised or contextualized after considering the contextually relevant molecular and supra-molecular matters that determine when and where which portions of the nucleotide sequence become causally relevant to the phenotypes of concern. For reasons like these, the semantic issue as to whether to count a stretch of DNA as containing multiple genes or containing one gene with highly pleiotropic effects will not be easily resolved. Indeed, no *general* or *all-purpose* resolution to this issue is available. The issue is best treated as a pragmatic one, to be answered for the convenience of the parties who are interested in it and need to communicate about it *according to what particular phenotypes or functions are of focal interest in their discussion, their experiments, or their disciplines*.

Conclusion

This brief examination of two illustrative examples (rat α-tropomyosin and the various enzyme and heat shock protein genes that produce lens crystallins by gene sharing) establishes that molecular pleiotropy raises serious problems for gene delimitation. In these examples, what looks like pleiotropy can often be equally well interpreted as activation of distinct genes even though those genes happen to pass through a stage in which they yield sequence-identical raw transcripts. There are, after all, a good number of other cases of overlapping genes at a single locus, such as the viral genes that can be read out properly from frameshifted starting points and the cases in which the introns for one gene contain exons for another gene or material that is converted into small nuclear RNAs and other distinctive functional products. In our cases, the molecular networks that control gene expression have triggered the transcription of the DNA from which, in specifically different cellular circumstances, distinctly different proteins are eventually derived. In the α-tropomyosin case, the cellular conditions in question cause the raw transcript produced from the nominal α-tropomyosin gene to yield different polypeptide strings in spite of starting from 'the same' – i.e., sequence-identical – raw transcript. But, for *some* purposes, isn't that to say that the DNA in question in *these* cells functions as one gene, and in *those* cells as another? And should – or shouldn't – the same claim apply to the nominal gene that produces a lens crystallin in the

eye and lactate dehydrogenase in the liver? If the phenotype of concern is the amino acid sequence, it shouldn't, at least not unless it is produced by starting from two numerically distinct nominal genes. But if the phenotype is the protein, the claim that the genes are distinct is equally supportable.

The stability of the distinct contexts – myoblast vs. smooth muscle cells, liver vs. lenticular cells, etc. – is exactly the sort of stability that is needed to determine what will be produced from a given RNA transcript, and hence from the DNA sequence that, *in that context*, produces *that* raw transcript.[18] And the stability of context arises because of the stability of *epigenetic* developmental processes. This, then, constitutes an argument that the concept of a gene that specifies a polypeptide sequence is incomplete without specification of the cellular context (or the relevant features of the cellular context). Thus, it is not sufficient to list the relevant nucleotide sequence of the DNA that is mirrored in the mRNA that is actually translated into an amino acid sequence or to list the nucleotide sequence containing all of the relevant exons and associated introns. We are forced to admit that one nominal gene yelds distinct products because of the impact of non-sequence based information or causes. We are also forced to admit that, in general, we cannot asses what polypeptide chain or protein product will be produced from nucleotide sequence alone. Thus, even such limited phenotypes as polypeptide sequence or protein product cannot be specified solely by information about nucleotide sequences. This means that insofar as we wish to develop analyses of genes as units that determine amino acid sequences or protein formation, we need to interpret genes with greater latitude than is provided by the nominal genes from which gene counts are obtained. Putting the point more generally, no system of specifying or delimiting genes by reference only to nucleotide sequences (with no further information about context) can satisfactorily specify the gene(s) that yield a particular protein.

Perhaps this may be viewed as old news. I think it is not. By getting this close to molecular detail, we have seen why a DNA nucleotide sequence does not genuinely specify the molecules that are or can be produced from that DNA. These considerations help explain why, even at the early step of specifying proteins, the correlation between genotype and phenotype cannot be nailed down neatly and, *a forteriori*, why the genome does not specify the organism. They explain

[18] Note also that a given nominal gene often produces primary transcripts of different lengths in different physiological circumstances.

why and how there is a trade-off between molecular pleiotropy and gene identity and why identification of the phenotypes of concern is critical for identification of genes. But once phenotypes are specified in distinctive ways, we are required to recognize a plurality of gene concepts, even when the phenotypic distinctions of concern are as limited as the distinction between the amino acid sequence produced and the protein produced. Nucleotide sequence similarity plays an enormous practical role in identifying the various distinct embodiments of particular genes, such as 'the' α-tropomyosin gene (or genes) in various organisms. But without enormously deep and interesting evolutionary, functional, and phenotypic stories backing up our choices, the sequence similarities do not suffice to reidentify these distinct sequences as being instances of the same gene. Furthermore, we must recognize that even amino acid sequences of polypeptide products are sometimes misleading. Consider, for example, the cases in which multiple proteins, separated after translation, are produced from one transcript derived from one ORF, but are produced in related organisms from separate transcripts that start at different ORFs i.e., that belong to distinct (nominal) genes. The point, of course, is not to insist that there is one correct answer as to how to determine the number of genes involved in such cases. Rather, given the state of present knowledge and present means of individuating genes, there is no single correct answer about the number of genes involved. This is no great hindrance to clear communication. We simply have to be clear how we are building the unavoidable context-dependence into our gene concept.

Covertly or overtly, but unavoidably, the principles on which genes are identified and individuated are context dependent. To recur to the epigraph from Scott Gilbert, the sense in which genes are not autonomous entities is very strong indeed. Exactly what gene we are dealing with depends on the context and networks involved and how we take account of them. And those contexts and networks extend beyond the organism into the external environment. Given this, the muddiness of gene concepts, frustrating to many philosophers, but celebrated by a few, is inevitable. If this strong version of context dependence is correct, the continuing evolution of gene terminology is not going to stabilize at some new orthodoxy based on a strictly intrinsic characterization of genes. At the same time, this contextualist stance highlights the usefulness of gene concepts and the possibility of retaining reasonably good control of cross-disciplinary and cross-contextual use of gene terminology. As a bonus, it undermines the genetic determinism still found in the rhetoric of the human genome

program and still deployed by many geneticists. I hope, therefore, that I have taken a small step toward convincing readers who were puzzled by ongoing controversies over gene concepts that those concepts can be fairly readily understood once one recognizes their strong context dependence. Such understanding will help us straighten out the most serious difficulties that arise when people talk about genes across contexts. Handled with care, these difficulties, too, can be overcome.

References

Alberts B., Johnson A., Lewis J., Raff M., Roberts K., Walter P., 2002, *Molecular Biology of the Cell*, (fourth ed.), New York / London: Garland Science.

Beurton P., Falk R., Rheinberger H.-J. (eds), 2000, *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*, Cambridge / New York: Cambridge University Press.

Breitbart R.A., Andreadis A., Nadal-Ginard B., 1987, 'Alternative Splicing: A Ubiquitous Mechanism for the Generation of Multiple Protein Isoforms from Single Genes', *Annual Review of Biochemistry*, 56: 481-495.

Burian R.M., 1985, 'On Conceptual Change in Biology: The Case of the Gene', In: Depew D.J., Weber B.H. (eds), *Evolution at a Crossroads: The New Biology and the New Philosophy of Science*, Cambridge, MA: MIT Press, 21-42.

Burian R.M., 1993a, 'Technique, Task Definition, and the Transition from Genetics to Molecular Genetics: Aspects of the Work on Protein Synthesis in the Laboratories of J. Monod and P. Zamecnik', *Journal of the History of Biology*, 26: 387-407.

Burian R.M., 1993b, 'Unification and Coherence As Methodological Objectives in the Biological Sciences', *Biology and Philosophy*, 8: 301-318.

Burian R.M., 1995, 'Too Many Kinds of Genes? Some Problems Posed by Discontinuities in Gene Concepts and the Continuity of the Genetic Material', *Preprint 18: Gene Concepts and Evolution*, Berlin: Max Planck Institute for the History of Science, 43-51, revised reprint as chapter 9 of Burian 2005b.

Burian R.M., 1997, 'On Conflicts Between Genetic and Developmental Viewpoints - and Their Resolution in Molecular Biology'. In: Dalla Chiara M.L., Doets K., Mundici D., van Bentham J. (eds), *Structure and Norms in Science. Proceedings of the 10th International Congress of Logic, Methodology, and Philosophy of Science*, vol. 2, Dordrecht: Kluwer, 243-264.

Burian R.M., 2000, 'On the Internal Dynamics of Mendelian Genetics', *Comptes rendus de l'Académie des Sciences, Paris. Série III, Sciences de la Vie / Life Sciences*, 323: 1127-1137.

Burian R.M., (2005a) , 'Lillie's Paradox – or, Some Hazards of Cellular Geography', *Epistemological Essays on Development, Genetics, and Evolution*, New York: Cambridge University Press, 183-209.

Burian R.M., (2005b), 'Too Many Kinds of Genes? Some Problems Posed by Discontinuities in Gene Concepts and the Continuity of the Genetic Material',

Epistemological Essays on Development, Genetics, and Evolution, New York: Cambridge University Press, 166-178.
Carlson E.A., 1966, *The Gene: A Critical History*, Philadelphia / London: W.B. Saunders.
Croft L., Schandorff S., Clark F., Burrage K., Arctander P., Mattick J., 2000. 'ISIS, the Intron Information System Reveals the High Frequency of Alternative Splicing in the Human Genome', *Nature Genetics*, 24: 340-341.
Driesch H., 1894, *Analytische Theorie der organischen Entwicklung*, Leipzig: Wilhelm Engelmann.
Falk R., 1986, 'What Is a Gene?', *Studies in History and Philosophy of Science*, 17: 133-173.
Falk R., 1995, 'The Gene: from an Abstract to a Material Entity and Back', *Preprint 18: Gene Concepts and Evolution*, Berlin: Max Planck Institute for the History of Science, 21-30.
Falk R., 2000, 'The Gene - a Concept in Tension'. In: Beurton P., Falk R., Rheinberger H.-J. (eds), *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*, Cambridge / New York: Cambridge University Press, 317-348.
Falk R., 2001, 'Can the Norm of Reaction Save the Gene Concept?'. In: Singh R., Krimbas C., Paul D.B., Beatty J. (eds), *Thinking About Evolution: 'Historical, Philosophical and Political Perspectives*, New York / Cambridge: Cambridge University Press, 119-140.
Fogle T., 2000, 'The Dissolution of Protein Coding Genes in Molecular Biology'. In: Beurton P., Falk R., Rheinberger H.-J. (eds), *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*, Cambridge / New York: Cambridge University Press, 3-25.
Gilbert S.F., 2000, *Developmental Biology*, (sixth ed.), Sunderland, MA: Sinauer.
Gilbert S.F., 2003, *Developmental Biology*, (seventh ed.), Sunderland, MA: Sinauer.
Gilbert S.F., Burian R.M., 2003, 'Developmental Genetics'. In: Hall B.K., Olson W.M. (eds), *Keywords and Concepts in Evolutionary Developmental Biology*, Cambridge, MA: Harvard University Press, 68-74.
Griffiths P.E., 2002, 'Lost: One Gene Concept, Reward to Finder', *Biology and Philosophy*, 17: 271-283.
Griffiths P.E., Neumann-Held E.M., 1999, 'The Many Faces of the Gene', *BioScience*, 49: 656-662.
Hall B.K., 2001, 'The Gene Is Not Dead, Merely Orphaned and Seeking a Home', *Evolution and Development*, 3: 225-228.
Jablonka E., Lamb M.J., 1995, *Epigenetic Inheritance and Evolution: The Lamarckian Dimension*, Oxford / New York: Oxford University Press.
Jeffery C.J., 2003, 'Multifunctional Proteins: Examples of Gene Sharing', *Annals of Medicine*, 35: 28-35.
Li L., Lindquist S., 2000, 'Creating a Protein-Based Element of Inheritance', *Science*, 287: 661-664.
Morange M., 2001, *The Misunderstood Gene*, Cambridge, MA: Harvard University Press.
Moss L., 2003, *What Genes Can't Do*, Cambridge, MA: MIT.
Müller G.B., Olsson L., 2003, 'Epigenesis and Epigenetics'. In: Hall B.K., Olson

W.M. (eds), *Keywords and Concepts in Evolutionary Developmental Biology*, Cambridge, MA: Harvard University Press, 114-123.

Neumann-Held E.M., 2001, 'Let's Talk About Genes: The Process Molecular Gene Concept and Its Context'. In: Oyama S., Griffith P.E., Gray R.D. (eds), *Cycles of Contingency: Developmental Systems and Evolution*, Cambridge, MA: MIT Press, 69-84.

Newman S.A., Müller G.B., 2000, 'Epigenetic Mechanisms of Character Origination', *Journal of Experimental Zoology (Molecular and Developmental Evolution)*, 288: 304-317.

Piatigorsky J., 1998, 'Multifunctional Lens Crystallins and Corneal Enzymes. More Than Meets the Eye', *Annals of the New York Academy of Sciences*, 842: 7-15.

Piatigorsky J., 2003, 'Gene Sharing, Lens Crystallins and Speculations on an Eye/Ear Evolutionary Relationship', *Integrative and Comparative Biology*, 43: 492-499.

Piatigorsky J., Norman B., Dishaw L.J., Kos L., Horwitz J., Steinbach P.J., Kozmik Z., 2001, 'J3-Crystallin of the Jellyfish Lens: Similarity to Saposins', *Proceedings of the National Academy of Sciences, USA*, 98: 12362-12367.

Piatigorsky J., Wistow G., 1991, 'The Recruitment of Crystallins: New Functions Precede Gene Duplication', *Science*, 252: 1078-1079.

Portin P., 2002, 'Historical Development of the Concept of the Gene', *Journal of Medicine and Philosophy*, 27: 257-286.

Reik W., Dean W., 2002, 'Back to the Beginning: Epigenetic Reprogramming', *Nature*, 420: 127.

Rheinberger H.-J., 2000, 'Gene Concepts: Fragments from the Perspective of Molecular Biology'. In: Beurton P., Falk R., Rheinberger H.-J. (eds), *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*, Cambridge / New York: Cambridge University Press, 219-239.

Roux W., 1885, 'Beiträge zur Entwicklungsmechanik des Embryo. Nr. 1.', *Zeitschrift für Biologie* 21: 411-524.

Rutherford S.L., Lindquist S.L., 1998, 'Hsp90 As a Capacitor for Morphological Evolution', *Nature*, 396: 336 - 342.

Sander K., 1991, 'Wilhelm Roux and His Programme for Developmental Biology', *Roux's Archives of Developmental Biology*, 200: 1-3.

Snyder M., Gerstein M., 2003, 'Defining Genes in the Genomics Era', *Science*, 300: 258-260.

Van Speybroeck L., Van de Vijver G., de Waele D. (eds), 2002, *From Epigenesis to Epigenetics: The Genome in Context*, (Annals of the New York Academy of Sciences), New York: New York Academy of Sciences.

Vom Saal F.S., Cooke P.S., Buchanan D.L., Palanza P., Thayer K.A., Nagel S.C., Parmigiani S., Welshons W.V., 1998, 'A Physiologically Based Approach to the Study of Bisphenol A and Other Estrogenic Chemicals on the Size of Reproductive Organs, Daily Sperm Production, and Behavior', *Toxicology and Industrial Health*, 14: 239-260.

Waddington C.H., 1940, 'The Genetic Control of Wing Development in *Drosophila*', *Journal of Genetics*, 41: 75-139.

Wagner G.P., Chiu C.-H., Hansen T.F., 1999, 'Is Hsp90 a Regulator of

Evolvability?', *Journal of Experimental Zoology (Molecular and Developmental Evolution)*, 285B: 116-118.
Waters C.K., 2000, 'Molecules Made Biological', *Revue Internationale de Philosophie*, 54: 539-564.