

What is a Gene?

Paul E. Griffiths

Biohumanities Project

301 Michie Building

University of Queensland

Brisbane, QLD 4072

Australia

Karola Stotz

Cognitive Science Program

Indiana University

810 Eigenmann

Bloomington, IN 47406

USA

Abstract

We outline three very different concepts of the gene - 'instrumental', 'nominal', and 'postgenomic'. The instrumental gene has a critical role in the construction and interpretation of experiments in which the relationship between genotype and phenotype is explored via hybridization between organisms or directly between nucleic acid molecules. It also plays an important theoretical role in the foundations of disciplines such as quantitative genetics and population genetics. The nominal gene is a critical practical tool, allowing stable communication between bioscientists in a wide range of fields grounded in well-defined sequences of nucleotides, but this concept does not embody major theoretical insights into genome structure or function. The post-genomic gene embodies the continuing project of understanding how genome structure supports genome function, but with a deflationary picture of the gene as a structural unit. This final concept of the gene poses a significant challenge to conventional assumptions about the relationship between genome structure and function, and between genotype and phenotype.

Keywords: Gene definition; gene concepts; postgenomics; history of genetics; conceptual change

Introduction

The term 'gene' has several different meanings in contemporary biology. Moreover, DNA sequences that are genes in one legitimate sense of the term may not be genes in another equally legitimate sense. In this article we will outline three important things that genes can be:

- Even today, many genes remain what they were in the early days of genetics, namely, factors in a model of the transmission of a heritable phenotype, or in a population genetic model of a changing population. The role of 'genes' in these models is akin to the role of centers of mass in calculating the effects of physical forces on massive bodies. We will refer to these as 'instrumental' genes (following Falk 1986).
- Most formal gene names, such as *sonic hedgehog (shh)*, refer to specific DNA sequences that are annotated as genes because of their similarity to the sequences that were the focus of study as biologists uncovered the functions of DNA from the mid-1950s to the 1970s. We refer to these DNA sequences as 'nominal' genes, following (following Burian 2004). Many but not all instrumental genes correspond to nominal molecular genes (and lend them their names), and many but not all nominal molecular genes correspond to instrumental genes.

- Finally, some genes are collections of DNA elements that play the role of the gene as envisaged in early molecular biology - acting as templates for the synthesis of gene products - but which are not 'nominal' genes, because the way in which DNA is used in the production of the relevant gene products does not fit the traditional stereotype. In many of these complex cases of genome transcription the DNA sequences involved can be annotated in different, equally legitimate ways, producing different numbers of gene or genes with different boundaries. Because the analysis of these complex cases has become easier in the 'postgenomic era' of massive amounts of sequence data and bioinformatic and other tools for analyzing that data, we call these 'postgenomic' molecular genes.

The use of the term 'gene' is so complex because it has evolved over time. Newly discovered phenomenon have necessitated new conceptions of the gene, but the new conceptions have not displaced earlier conceptions, which often remain the best way to deal with the classes of genetic phenomena that inspired them. As a result, multiple conceptions of the gene have come to coexist. In the next section we briefly outline some of this history.

In the third section of the paper we discuss some of the challenges which the postgenomic molecular gene poses to conventional views about what genes are and what they do. We argue that the most general conception of a molecular gene is one that recognizes that genes - and the 'genetic information' that they contain - are constituted during development, making the gene a flexible, context-dependent entity. Genes are

'things you can do with your genome'. Classical molecular genes are a special case of this more general conception.

Our final answer to the question 'what is a gene' consists of this general, postgenomic vision of the molecular gene, plus a reminder that we cannot do without the older, instrumental gene, and an acknowledgement of the practical value of the nominal gene.

The evolution of the gene concept

The gene of 'classical' genetics¹ had a dual identity (Falk 1986, 2000). The gene was a postulated physical unit of heredity. But genes were also intervening variables which allowed prediction of the phenotypes of offspring from the phenotypes of parents. It was this later conception of the gene with which geneticists were concerned in their actual scientific work. As Morgan wrote in his Nobel address, "There is not consensus of opinion amongst geneticists as to what genes are – whether they are real or purely fictitious – because at the level at which genetic experiments lie, it does not make the slightest difference whether the gene is a hypothetical unit, or whether the gene is a material particle" (1933, quoted in Falk 1986, 148). Recent authors have stressed that classical genetics should not be thought of as merely a theory of heredity embodied in Mendel's laws and their later refinements (Waters 2004; Falk In Press). Instead, this theory of heredity functioned as an investigative tool with which geneticists could explore broader biological questions. The 'genetic analysis' of particular phenotypes, identifying genetic loci related to those phenotypes, sets of alleles at each locus, relations

of linkage and epistasis between loci, and relations of dominance between alleles, provided data bearing on more general questions about the mechanisms of heredity, development and physiological function. The aim of genetic analysis was not to test the theory of the gene, but to answer other biological questions by assuming the theory of the gene and working out what else must be true to make that assumption consistent with the results of carefully chosen hybridizations (see the detailed reconstructions in Waters 2004). Research in classical genetics thus resembled Thomas Kuhn's famous characterization of 'normal science' as the activity of making the world fit the paradigm (Kuhn 1962). For example, Raphael Falk (1986, 141-145) has discussed how early results calling into question the 'purity of the gametes' (the doctrine that an allele is not modified by the allele with which it shares a locus) were reinterpreted to render them consistent with that important doctrine.

Quantitative characters, like height and weight, which vary continuously between individuals, posed a significant problem for early genetics, since only a character with discrete values can appear in Mendelian ratios in offspring. However, as early as 1918 R.A. Fisher had shown that statistical procedures for studying correlations between phenotypes could be interpreted in Mendelian terms. Quantitative traits are treated as if they were the effect of a large number of genes each of which makes an equal contribution to variation in the character. The attitude of the geneticist to these postulated genes is transparently instrumental.

Michel Morange has written that “Molecular biology was born when geneticists, no longer satisfied with a quasi-abstract view of the role of genes, focused on the problem of the nature of genes and their mechanism of action” (Morange 1998, 2). Herman J. Muller gave a particularly clear statement of the nature of the postulated physical gene. It must be capable of autocatalysis (self-replication) in order to explain heredity. It must be capable of heterocatalysis – producing something different in structure from itself – in order to explain the manifestation of genetic differences in different phenotypes. Finally, it must be mutable – able to change its structure – so as to create heritable variation upon which natural selection can act. Investigations of the physical reality of the gene – beginning with Muller and others' use of X-ray mutagenesis to estimate the size of the genetic target – mark a significant epistemological development in genetics. Because the aim of genetic analysis was to analyse phenotypes in terms of underlying genes and their properties, questioning the core assumptions of the theory of the gene could only result in intellectual paralysis. Patterns of inheritance which appeared to violate basic Mendelian assumptions must either be made to fit those assumptions by further genetic analysis, or accommodated with fudge-factors like 'penetrance' and 'expressivity', or put aside as anomalies. In contrast, the investigation of genes as physical entities opened up the possibility of obtaining robust, independent evidence against even the most fundamental Mendelian assumptions. Features of the gene which were previously treated as definitional became features which could be tested and potentially rejected.

Before the 1950s the most direct access to the gene as a physical entity was provided by studies of genetic linkage. It is an obvious implication of the chromosomal theory of

heredity that genes located close together on a chromosome are unlikely to be separated by crossing over between chromosomes in meiosis, and are thus likely to be inherited together. The linkage coefficients between genetic loci can be calculated by observing the results of breeding experiments, allowing the creation of linkage maps. The discovery of giant, polytenic chromosomes in the salivary glands of *Drosophila* allowed these linkage maps to be correlated with the observable patterns of banding in these chromosomes. Changes in linkage relationships were thus convincingly interpreted as the results of inversions and translocations of segments of chromosomes. The epistemological result of this was that a gene could be identified by two different criteria – via the phenotypic difference it makes and as something located on a specific segment of a chromosome as revealed by linkage analysis. This made it possible to discover 'position effects', in which changing the location of a gene on the chromosome changes its effect on the phenotype. In the hands of the eminent geneticist Richard Goldschmidt position effects became ammunition against the 'theory of the gene' itself.

Classical geneticists distinguished between mutations, in which a gene changes its intrinsic nature, and position effects in which an identical gene has a different effect because of its location on the chromosome. While these were operationally distinct – mutations are not accompanied by observable change in the chromosome or by changes in linkage relationships – Goldschmidt denied that this operational distinction corresponded to a significant biological distinction. Mutations, he argued, were small changes in the structure of the chromosome, whilst position effects resulted from larger changes in the structure of the chromosome. The data did not directly support the idea

that the chromosome is made up of discrete units called genes or that there is a fundamental difference between a change in one of those units and a larger change in a chromosome segment. Goldschmidt's call to replace discrete genes by a continuous chromosome with a hierarchical physical structure was not endorsed by other geneticists (Dietrich 2000, 2000). It demonstrates, however, the way in which the existence of multiple 'epistemic pathways' to the gene made it possible to use results derived by one investigative technique to challenge the assumptions underlying another investigative technique.

The work of Seymour Benzer which led to the establishment of the 'neo-classical' (Portin 1993) or 'classical molecular' (Neumann-Held 1998) concept of the gene made more successful use of a new pathway to the gene to overthrow some assumptions of the classical theory. The *cis-trans* or 'complementation' test is a classical technique for distinguishing mutations in a single gene from mutations in two different genes. Most mutations are recessive in the heterozygote. Hence, if an offspring receives a mutant allele of one gene from one parent and a mutant allele of a second gene from the second parent, it will also receive a normal allele of the second gene from the first parent and a normal allele of the first gene from the second parent. As a result, that offspring will be phenotypically normal. If both mutations are in a single gene, however, the offspring will receive two mutant copies of the gene, one from each parent, and one containing each mutation. That offspring will therefore be a phenotypic mutant. The *cis-trans* test assumes that recombination - the association of alleles from two homologous chromosomes of a parent on a single chromosome in the offspring as a result of crossing

over during meiosis - is a process that recombines whole genes. However, if recombination can occur *within* a gene, so that part of the gene on one chromosome comes to be united with part of the same gene from the other homologous chromosome, then it is possible for the *cis-trans* test to fail. Intragenic recombination can patch together a normal copy of a gene from two mutant copies. Obviously, this will happen in only a very small proportion of cases. During the late 1950s, and using the bacteriophage (bacterial viruses) that were an important model organism in early molecular biology, Benzer was able to create a high-resolution analogue of the *cis-trans* test (see e.g. Holmes 2000 for details) and to systematically detect intragenic recombination. This work showed that a single classical gene is a linear series of sites at which independent mutation and recombination events can occur. Benzer proposed replacing the traditional term 'gene' with three more specific terms, 'muton', 'recon', and 'cistron' – denoting the units of mutation, of recombination, and of genetic function as defined by the *cis-trans* test.

Benzer's work might perhaps have been interpreted as vindicating a Goldschmidt-like skepticism about dividing the chromosome into discrete genes. But instead, the cistron was identified with the gene and the other proposed genetic units were not taken up. One reason for this response was the availability of a chemical model of the gene, due to Francis Crick, James Watson and others (Olby 1974), which provided a natural interpretation for Benzer's results. The unit of mutation and recombination is the single nucleotide, whilst the cistron corresponds to a series of nucleotides involved in the synthesis of a single gene product through linear correspondences between DNA and

RNA and between RNA and protein. With the unraveling of the genetic code and of the basic processes of transcription and translation in the 1960s the instrumental and physical conceptions of the gene seemed to have converged neatly on a single, well-defined entity - the classical molecular gene. The functional role of the gene had been narrowed down to contributing to the phenotype through the (heterocatalytic) synthesis of a biomolecule. The autocatalytic synthesis of gene copies is more properly the function a whole DNA molecule (chromosome), and mutation and recombination are more properly the function of individual DNA nucleotides. This refined functional role was occupied by a specific physical structure – an 'open reading frame' (ORF) – a DNA sequence beginning with a start codon and ending with a stop codon. Today, it is primarily these sequences which are formally annotated as genes – sequences which are known or suspected to play the functional role of the molecular gene and which also have the characteristic structure of an ORF with adjacent regulatory elements such as the 'TATA box' - a binding site for transcription factors. Richard Burian has described these as 'nominal genes', a phrase intended to convey the following ideas, with which we find ourselves strongly in agreement: “The use of databases containing nucleotide sequences is well established. Codified as part of this process is a particular use of gene concepts on the basis of which one can identify various genes and count the number of genes in a given genome. ... I call genes, picked out in this way, nominal genes. A good way of parsing 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.” (Burian 2004, 64-5)

The gene of molecular biology is fundamentally the 'image of the gene product in the DNA'². The central epistemological role of linear correspondence between molecules in molecular biology was first emphasized by Kenneth C. Waters (1990; 1994; 2000). Linear correspondence between molecules is fundamental to biologists' ability to identify and manipulate those molecules, via a whole host of technologies such as cDNA libraries, microarrays and RNA interference, to take a random selection. Linear correspondence is thus at the heart of the molecular conception of what genes are. The molecular gene concept is a kind of schema which uses linear correspondence to elements in some molecule of interest to pick out a certain sequence of DNA nucleotides as the gene for that molecule. For example, overlapping genes are a common phenomenon in both prokaryote and eukaryote genomes. What distinguishes the set of nucleotides that make up one gene from those that make up the second is the relationship of linear correspondence between these two sets and two distinct gene products. When two very different products are produced, biologists annotate the sequence as two overlapping nominal genes. However, when the products are similar they are usually regarded as alternatively spliced products of a single nominal gene (Alberts et al. 2002, 438; 1994, 457). While this does not invalidate Waters' basic insight, it does suggest that things are more complicated than his formal model of the gene concept allows, as we discuss below.

Waters' insight into the epistemological structure of molecular biology illuminates other concepts as well as that of the gene. An 'exon' was traditionally defined a segment of a

eukaryote gene that is translated into protein. Increasingly, however, an exon is defined as a segment of a eukaryote gene that makes it through posttranscriptional processing to form part of the mature mRNA. In early 2005 our Google search for definitions of 'exon' yielded twenty-six examples, of which sixteen restricted exons to coding sequences³, five permitted them in untranslated regions of the gene (UTRs)⁴ and five were unclear on the point. The difference is significant. On the traditional definition it would make no sense to distinguish exons in the untranslated region at either end of a eukaryote gene, and in 2005 it was still possible to find biologists who regarded this as an inappropriate annotation. It would also make no sense to distinguish exons in sequences which are transcribed as RNA and never give rise to a protein. However, as biologists have discovered new classes of functional RNAs, and as interest in the regulatory role of alternative splicing of untranslated regions has grown, a shift has occurred in the meaning of 'exon'. If we looked at the exon concept in the spirit of Waters' analysis of the molecular gene concept, we would expect this change. Just as the molecular gene is the set of DNA nucleotides that corresponds to the gene product at whatever stage of interest is the focus of research, the exons of a gene are the sets of nucleotides that are spliced together to make the gene product that is the focus of research. As the proportion of the scientific community whose focus is on post-transcriptional processing of mRNAs rather than on the post-translational gene product (if any) has grown, the concept of an exon has been transformed in just the way this analysis would predict.

The classical molecular gene concept was the product of a highly successful attempt to identify the physical basis of the instrumental gene. However, it was not able to simply

replace the instrumental gene, because that concept is embedded in biological theory, and in biological practice, in ways that would be artificially and unhelpfully restricted by replacing the instrumental with the molecular concept. A particularly clear example is provided by the 'evolutionary gene concept', a generalization of the instrumental gene concept famously espoused by George C. Williams (1966). The population genetics at the heart of neo-Darwinian evolutionary theory assumes that the phenotypic differences upon which selection acts result from individuals having different alleles at various Mendelian loci, and that changes in the composition of populations over time is fully reflected in changes in the ratio of different alleles at each locus (there is clearly no difficulty in generalising the theory to cover other genetic systems, such as the maternal inheritance of mitochondrial genes or haplo-diploid systems in which males and females differ in chromosome number, but for simplicity we will ignore this here). For the purposes of population genetics and evolutionary theory, therefore, a gene is anything which causes a phenotypic difference and which behaves like a Mendelian allele. Hence, Williams writes, "I use the term gene to mean "that which segregates and recombines with appreciable frequency."" (1966, 24) The critical property of an evolutionary gene is not that it codes for a protein, but that it is a unit of recombination – a segment of chromosome which regularly recombines with other segments in meiosis and which is short enough to survive enough episodes of meiosis for selection to act upon it as a unit (see the careful elaboration in Dawkins 1982, 86-91). Since the only truly indivisible unit of recombination is the single nucleotide, the unity of the evolutionary gene is a matter of degree, but this is no impediment to the use of the evolutionary gene concept in population biology.

Many chromosome segments which behave as Mendelian alleles, and thus evolutionary genes, are not nominal genes. Untranscribed regulatory regions, such as 'enhancer' and 'silencer' regions that bind transcription factors acting on genes located thousands of base pairs away, behave as separate Mendelian alleles and are open to the action of natural selection. Even 'insulator' regions, which affect gene expression by physically separating genes and regulatory elements from one another, are potential evolutionary genes. An adequate evolutionary genetics must deal with all DNA-based heritable differences in fitness. Restricting the units of evolutionary genetics to coding sequences, or to indivisible units consisting of coding sequences and their complement of regulatory regions, would make the theory inadequate to explain evolutionary change. The evolutionary genes that are not simultaneously nominal molecular genes have their own 'selfish' (Dawkins 1976) evolutionary dynamics and respond to selection on that basis. Richard Dawkins made this point in an exchange with the molecular biologist Gunther Stent, who had objected to Williams definition of the gene (Stent 1977). It is surely unquestionable that we understand genetics better in the light of molecular biology, so how can we ignore that knowledge in defining the gene? We can do so, argues Dawkins, because the original Mendelian gene played a number of different roles in biological theory (as we have already seen in Benzer's distinguishing the units of mutation, recombination and function) and the growth of biological knowledge has revealed that these roles are not always filled by the same units (1982, 85-86). The unit of genetic function is not always the unit of genetic evolution⁵.

It is not necessary to go to population genetics to find 'genes' that are not nominal genes. Geneticists continue to make use of classical genetic techniques to identify regions of chromosome in which nominal genes may be located. Even when the explicit aim of this work is to identify nominal genes, the conceptualization of the gene that is actually used to do the work is the classical, instrumental conception. This is shown by the fact that well-conducted work of this kind, free from any experimental error or errors of reasoning may locate a candidate 'gene' that does not correspond to a molecular gene, but to some other functional DNA element, such as an untranscribed regulatory region. As Marcel Weber concludes, after an insightful comparison of Mendelian and molecular analyses of *Drosophila* loci, "even though the classical gene concept had long been abandoned at the theoretical level, it continues to function in experimental practice up to the present." (Weber 2004, 223). Consider, for example, a report of the localization of a 'gene for' a psychiatric disorder to some chromosomal region. Clearly, one way to interpret such a report is as a prediction that a sequence straightforwardly encoding a protein or a functional RNA – a nominal gene - will be found at that locus. But it is equally legitimate to interpret the report as evidence that something about that locus makes a heritable difference to the disease phenotype. Nor would the fact that the eventual annotation of the sequence at that locus does not identify a nominal gene necessarily be a criticism of the earlier work. The instrumental gene concept is alive and well.

The epigenesis of genetic information

We argued in the last section that Waters' analysis of the classical molecular gene concept rests on a deep insight into the epistemology of molecular biology. But we do not believe that it provides a fully adequate analysis of the contemporary molecular gene. According to Waters there is a clear and uniform way to understand genes at the molecular level, namely as “a gene g for linear sequence l in product p synthesized in cellular context c ” (2000, 544). He recognizes cellular conditions as playing a role in the process of gene expression but singles out the DNA sequence as having “special determining role because the differences in the linear structures among different polypeptides synthesized in a cell or cell structures results from actual differences in the linear structures of the DNA segment expressed, not from differences in the many other causal agents essential for the production of the polypeptide” (2000, 543). In this section, we argue *contra* Waters that the linear sequence of a gene product in eukaryotes is rarely specified or determined by its DNA sequence alone. The very high number of alternatively spliced forms expressed by the human genome argues against Waters' view⁶ as do all the other distinct novel transcripts that cannot be accounted for by canonical forms of transcription. Since diverse sequences in gene products are derived from a single DNA sequence, mechanisms for the regulation of genome expression must provide additional sequence information. The main actors of these mechanisms, proteins and functional RNAs, relay environmental information to the genome with important consequences for sequence selection, processing and, in extreme cases, sequence creation. Since these selection and creation mechanisms determine if a given DNA sequence is able to produce a gene product,

arguably the very status of a DNA sequence as a ‘gene’ is dependent on its cellular and broader context.

We will suggest that in contemporary molecular bioscience genes are not straightforward structurally defined entities, or even the mixed functional-structural entities defined by Water's schema given above. Instead, genes are ‘things an organism can do with its genome’: they are ways in which cells utilize available template resources to create biomolecules that are needed in a specific place at a specific time. The same DNA sequence potentially leads to a large number of different gene products and the need for a rare product calls for the assembly of novel mRNA sequences. Hence the information for a product is not simply encoded in the DNA sequence but has to be *read into* that sequence by mechanisms that go beyond the sequence itself. Certain coding sequences, plus regulatory and intronic sequences, are targeted by transcription, splicing and editing factors (proteins and functional RNAs), which in turn are cued by specific environmental signals. Regulatory mechanisms determine not only whether a sequence is transcribed, but where transcription starts and ends, how much of the sequence will be transcribed, which coding and noncoding regions will be spliced out, how and in which order the remaining coding sequences will be reassembled, which nucleotides will be substituted, deleted or inserted, and if and how the remaining sequence will be translated. Many of these mechanisms do not simply produce alternative protein-coding transcripts. A sequence may be transcribed into several parallel, coding and noncoding transcripts. The factors that interactively regulate genomic expression are far from mere background condition or supportive environment; rather they are on a par with genetic information

since they *co-specify* the linear sequence of the gene product together with the target DNA sequence. Networks of genome regulation, including several different kinds of gene products and instructional environmental resources, specify a range of products from a gene through the selective use of nucleotide sequence information and, more radically, the creation of nucleotide sequence information.

To exemplify these general claims we will briefly describe some of these mechanisms and give examples⁷. In eukaryotes, the DNA sequence is transcribed into a pre-messenger RNA from which the final RNA transcript is processed by cutting out large non-coding sequences, called *introns*, and splicing together the remaining, mostly but not always, coding sequences, called *exons*. Biologists speak of alternative *cis*-splicing when more than one mature mRNA transcript results from these processes through the cutting and splicing of alternative exons⁸. Beside these canonical splice variants the genome produces a large variety of transcripts that cannot be so easily attributed to a single nominal gene. Some transcripts are made of exons from adjacent nominal genes that are 'co-transcribed' to produce a *single* pre-mRNA . Co-transcription may also occur between a gene and an adjacent 'pseudo-gene' rendering the latter capable of providing part of the coding sequence . Alternative gene products may also be derived from so-called 'overlapping genes'. Until very recently it was thought that only one strand of DNA is transcribed, but in fact DNA can be read both forwards and backwards by the cellular machinery, producing either different or matching (complementary) products. The latter case, in which exactly the same sequence is read in reverse, may result in an antisense transcript with regulatory function, possibly through silencing its complementary transcript. If two

proteins are produced from overlapping genes, their degree of difference depends on the extent of overlap of coding sequences, and on whether these shared sequences are read in the same reading frame. It is the precise nucleotide at which reading begins that determines which codons a DNA sequence contains. Starting at a different nucleotide is called 'frameshift', a phenomenon that would look like this if applied to an English sentence: 'A gene is a flexible entity' becomes 'Age nei saf lex ibl een tit y'. But unlike a sequence of letters a DNA sequence will always be made up of meaningful 'words' (codons) wherever reading begins. This means that very different products can be read from the same sequence merely by frameshifting by one nucleotide. As well as alternative transcripts from a DNA sequence, multiple simultaneous transcripts can occur, as is the case of the parallel processing of functional non-coding RNAs (such as microRNAs) from the intronic regions of the initial transcript. These RNAs may be involved in the regulation of coding transcript of the same gene. In all of the above instances one can argue that the selective use of nucleotide sequences through a range of transcriptional and post-transcriptional mechanisms specify or at least co-determine the linear sequence of the final product.

In the following cases the linear sequence is not mirrored in the DNA sequence at all but must be created through a variety of post-transcriptional processes. Biologists speak of *trans*-splicing when a final mRNA transcript is processed from two or more independently transcribed pre-mRNAs. While the prefix *trans* might suggest that these pre-mRNAs are derived from DNA sequences far apart from each other, this is not always the case. In fact, two copies of the very same sequence can be spliced together

this way to include multiple copies of the same exons or reverse the order of several exons in the final transcript. In some cases alternative exons each feature their own promoter to allow their individual selection for the final transcript. RNA editing is another mechanism of modification that can significantly diversify the 'transcriptome' or 'proteome' (the total complement of final transcripts or proteins in the cells of an organism). Whereas most other forms of post-transcriptional modification of mRNA (capping, polyadenylation and *cis*-splicing) can be said to retain the *correspondence* of coding sequence and gene product (even though certain coding and noncoding regions have been cut out), RNA editing disturbs this correspondence to a large extent by changing the primary sequence of mRNA during or after its transcription. This creation of 'cryptogenes' via RNA editing can potentially have radical effects on the final product, depending on whether editing changes the sense of the codon in which it occurs. While there are likely as many mechanisms of RNA editing as there are organisms, all belong to one of three principle kinds: the site-specific *insertion* or *deletion* of one or several nucleotides, or nucleotide *substitution* (cytidine-to-uridine and adenosine-to-inosine deamination, uridine-to-cytidine transamination). Another rather common mechanism able to disrupt the colinearity between DNA sequence and final product is the nonstandard translation of mRNA. The three different ways through which the translational machinery is able to recode the message are frameshifting, programmed slippage or bypassing, and codon redefinition. Although we will not describe them here, other processes may occur before, during or after the final mRNA transcript is translated into a protein sequence or processed into a functional RNA. The relationship between DNA and gene product is indirect and mediated to an extent that was never anticipated

when the basic mechanisms of transcription and translation were clarified in the 1960s.

Figure 1 shows one such 'postgenomic **gene**'.

Comment [pa1]: Insert Figure 1 about here

Focusing on the cutting-edge of contemporary genomics can induce an extremely deflationary view of the gene. The classical molecular gene concept was derived from work on a limited range of organisms: prokaryotes and bacteriophage. Further investigation of the manner in which a wider range of genomes generate a wider range of gene products has revealed that the functional role of genes can be filled by diverse, highly flexible mechanisms at the level of the DNA itself. Our increased comprehension of the structure and organization of the genetic material has “left us with a rather abstract, open and generalized concept of the gene” (Portin 1993, 173), or as Falk was already suggesting 20 years ago, “Today the gene is [...] *a* unit, *a* segment that corresponds to *a* unit-function, as defined by the individual experimentalist’s need. It is neither discrete [...] nor continuous [...], nor does it have a constant location [...], nor a clearcut function [...], not even constant sequences [...] nor definite borderlines.” (Falk 1986, 169) In the light of our current understanding of genome structure and function some have even conclude that Goldschmidt’s critique of the particulate gene has been vindicated: “The particular gene has shaped thinking in the biological sciences over the past century. But attempts to translate such a complex concept into a discrete physical structure with clearly defined boundaries were always likely to be problematic, and now seem doomed to failure. Instead, the gene has become a flexible entity with borders that are defined by a combination of spatial organization and location, the ability to respond specifically to a particular set of cellular signals, and the relationship between expression patterns and the

final phenotypic effect” (Dillon 2003, 457). Some molecular biologists, realizing that the concepts of gene transcription or gene expression may not suffice to capture the complex architecture of the transcriptome of many eukaryotes have proposed the more general term of 'genome transcription' to allow for the incorporation of RNA transcripts that contain sequences outside the border of canonical genes. This view does not sit easily with the classical molecular conception of genes, which from the new perspective seem like “statistical peaks within a wider pattern of genome expression” (Finta and Zaphiropoulos 2001, 160).

One pragmatic, technological reason that today’s biologists are prepared to consider such radical options is that the challenge of automated gene annotation has turned the apparently semantic issue of the definition of ‘gene’ into a pressing and practical one as the limitations of a purely structural, sequenced-based definition of the gene have become apparent. According to some, one “possible consideration stemming from the growing list of transcribed regions of the genome is the likelihood that the present efforts in estimating the total number of genes in the genome is misguided and at the very least miscalculated. These efforts are misguided given the discussion presented previously that a more useful entity to be counted is the number of transcripts. They are also miscalculated because such estimates are biased strongly in favor of protein- coding transcripts (Kampa et al. 2004, 341). Recent investigations of the complexities of the human transcriptome support this view. One study revealed a large number of transcriptional events, 60% of which involve novel transcriptional units outside annotated genic regions, and the rest of which involve newly discovered exons or exon isoforms of

known genes. This study detected overlapping transcription on the positive and negative strand in 60% of the surveyed loci, and a variety of intronic transcriptional fragments and intergenic transcription. If correct, these results have important implications for the definition of a gene, and for the relationship between genotype and phenotype. (Kapranov et al. 2005)

Conclusions

The gene began life as an intervening variable, defined functionally by the Mendelian pattern of heredity, and rapidly acquired a second identity as a hypothetical material unit. A productive dialectic between functional and structural conceptions of the gene concluded with the ‘classical molecular’ conception of the gene, which fused structure and function in a single definition. Further investigation of a wider range of genomes and a wider range of gene products suggests that the structural basis upon which gene-like functions are performed may be very broad. At this point in time, then, it is necessary to distinguish between (at least) three senses of ‘gene’:

1. The traditional, *instrumental* gene retains a critical role in the construction and interpretation of a range of experiments in which the relationship between genotype and phenotype is explored via hybridization between organisms or directly between nucleic acid molecules. It also retains an important theoretical role in the foundations of disciplines such as quantitative genetics and population genetics. While these areas of the biosciences are in a continuous and fruitful exchange with work that utilizes

molecular conceptions of the gene, to attempt to reduce instrumental genes to molecular genes is to misunderstand the epistemological role of the instrumental gene.⁹

2. The *nominal* molecular gene is a critical practical tool, allowing stable communication between bioscientists in a wide range of fields grounded in well-defined sequences of nucleotides. But this does not imply that the scientific community has a clear understanding of what makes a sequence a gene that needs only to be made explicit. Thomas Fogle has argued powerfully that this is not the case (Fogle 2001). The concept of the gene used in sequence annotation is something like a stereotype or prototype: a sequence is a gene if it has enough similarities to other genes, e.g. it contains an open reading frame, has one or more promoters, has one or more transcripts which are not too functionally diverse from one another, etc. This is more or less a description of automated 'gene discovery' methods, and Fogle's suggestion is that the working concept of the gene is no more principled or definition-like than this. The various 'gene-like' features are not weighted against one another in any principled, theory-driven way, but rather are weighted differently on different occasions in order to segment the DNA sequence into fairly traditional looking 'genes', sometimes giving up on structural criteria to save functional ones (as in cases of *trans*-splicing), at other times giving up on functional criteria to save structural ones (as in co-transcription of a gene and a 'pseudo-gene'). Thus, while the nominal gene is an important practical tool, the nominal gene concept does not constitute a major theoretical insight into genome structure or function.

3. The *post-genomic* molecular gene is the 'image of the gene product in the DNA' no matter how fractured and distributed that image may be, and no matter how much supplementation the transcribed sequences requires to determine the sequence of elements in the product. This concept embodies the continuing project of understanding how genome structure supports genome function, but with a deflationary picture of the gene as a structural unit. The concept poses a significant challenge to conventional assumptions about the relationship between genome structure and function, and between genotype and phenotype. We have suggested that an adequate general conception of the molecular gene must acknowledge that genes are defined by the way DNA sequences are used in particular cellular and broader contexts, and not merely by their structure. Genes that can be recognized by their structure alone are a special case of this more general concept. Indeed, as shown in Figure 1, the very same gene (by the important criteria of descent from a common ancestral gene and conserved function) may be recognizable by traditional structural criteria alone in one organism and unrecognizable in another.

We also believe that the nature of the postgenomic gene supports the view that phenotypes are not simply expressions of genetic information but rather emerge from a 'developmental system' (Oyama, Griffiths, and Gray 2001) that encompasses many aspects of what would traditionally be regarded as the environment, but this is not the place to defend this broader view, and our claims about the gene concept do not depend upon it.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0217567. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Griffiths' work on this paper was supported by an Australian Research Council Federation Fellowship.

Notes

¹ The period roughly bounded by Thomas Hunt Morgan's 'The Theory of the Gene' (1917) and Seymour Benzer's work on fine-structure mapping in the late 1950s.

² A phrase we owe to Rob D. Knight (pers. comm.)

³ E.g. "The parts of the DNA sequence that code for proteins. Compare with intron"
Baylor College of Medicine www.bcm.edu/pa/genglossary.htm

⁴ E.g. "One of the parts of a gene whose sequence is present in the mature mRNA"
University College London <http://www.ucl.ac.uk/~ucbhjow/b241/glossary.html>

⁵ Dawkins goes too far, however, when he defines the evolutionary gene as *any* segment of chromosome: 'When I said 'arbitrarily chosen portion of chromosome', I really meant arbitrarily. The twenty-six codons I chose might well span the border between two

cistrons' (1982, 87) The objections to this extreme position are discussed by Sterelny and Griffiths (1999, 79-82)

⁶ “Based on data from Chromosomes 21 and 22, there is a distinct possibility that nearly all of the coding genes of the genome exhibit alternatively spliced forms.” (Kampa et al. 2004, 340)

⁷ (See also Stotz and Griffiths 2004; Stotz, Bostanci, and Griffiths In Press and our website <http://representinggenes.org>).

⁸ In contemporary usage, *cis*- elements are those transcribed together as parts of a single pre-mRNA whereas *trans*- elements are transcribed separately and united at some stage of post-transcriptional processing (*trans*-splicing). Thus *trans*- elements in the modern sense (*trans* on mRNA) may be *cis*- located in the older sense referred to in the second section of this paper (*cis* on the DNA).

⁹ For an important recent attempt to distinguish and analyse the complimentary epistemological roles of instrumental and physical genes, see (Moss 2003)

References

- Alberts, B, D Bray, J Lewis, M Raff, K Roberts, and J.D Watson. *The Molecular Biology of the Cell*. 3 ed. New York and London: Garland, 1994.
- . *Molecular Biology of the Cell*. 4 ed. New York: Garland, 2002.
- Burian, Richard M. "Molecular Epigenesis, Molecular Pleiotropy, and Molecular Gene Definitions." *History and Philosophy of the Life Sciences* 26, no. 1 (2004): 59-80.

- Chapdelaine, Y., and L. Bonen. "The Wheat Mitochondrial Gene for Subunit I of the Nadh Dehydrogenase Complex: A Trans-Splicing Model for This Gene-in-Pieces." *Cell* 65, no. 3 (1991): 465-72.
- Dawkins, Richard. *The Selfish Gene*. Oxford: Oxford University Press, 1976.
- Dawkins, Richard. *The Extended Phenotype: The Long Reach of the Gene*. San Francisco: Freeman, 1982.
- Dietrich, Michael R. "From Hopeful Monsters to Homeotic Effects: Richard Goldschmidt's Integration of Development, Evolution and Genetics." *American Zoologist* 40 (2000): 738-47.
- . "The Problem of the Gene." *Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie* 323, no. 12 (2000): 1139-46.
- Dillon, Niall. "Positions, Please..." *Nature* 425 (2003): 457.
- Falk, Raphael. "The Gene: A Concept in Tension." In *The Concept of the Gene in Development and Evolution*, edited by Peter Beurton, Raphael Falk and Hans-Jörg Rheinberger, 317-48. Cambridge: Cambridge University Press, 2000.
- . "Genetic Analysis." In *International Handbook of the Philosophy of Biology*, edited by Mohan Matthen and Chris Stephens, xxx-xxx: Elsevier, In Press.
- . "What Is a Gene?" *Studies in the History and Philosophy of Science* 17 (1986): 133-73.
- Finta, C., and P. G. Zaphiropoulos. "A Statistical View of Genome Transcription." *Journal of Molecular Evolution* 53 (2001): 160-62.
- Fogle, Thomas. "The Dissolution of Protein Coding Genes in Molecular Biology." In *The Concept of the Gene in Development and Evolution*, edited by Peter Beurton,

Raphael Falk and Hans-Jörg Rheinberger, 3-25. Cambridge: Cambridge University Press, 2001.

Holmes, Frederic L. "Seymour Benzer and the Definition of the Gene." In *The Concept of the Gene in Development and Evolution*, edited by Peter Beurton, Raphael Falk and Hans-Jörg Rheinberger, 115-55. Cambridge: Cambridge University Press, 2000.

Kampa, D., J Cheng, P Kapranov, M Yamanaka, S. Brubaker, S Cawley, J Drenkow, A Piccolboni, S Bekiranov, G Helt, H Tammana, and T. R Gingeras. "Novel Rnas Identified from an in-Depth Analysis of the Transcriptome of Human Chromosomes 21 and 22." *Genome Research* 14, no. 331-342 (2004).

Kapranov, P , J Drenkow, J Cheng, J Long, H Gregg, S Dike, and T. R Gingeras. "Examples of the Complex Architecture of the Human Transcriptome Revealed by Race and High-Density Tiling Arrays." *Genome Research* 15 (2005): 987-97.

Kuhn, Thomas. *The Structure of Scientific Revolutions*. 1 ed. Chicago: University of Chicago Press, 1962.

Morange, Michel. *A History of Molecular Biology*. Cambridge, MA: Harvard University Press, 1998.

Morgan, Thomas Hunt. "The Theory of the Gene." *American Naturalist* 51 (1917): 513-44.

Moss, Lenny. *What Genes Can't Do*. Cambridge, MA: MIT Press, 2003.

Neumann-Held, E.M. "The Gene Is Dead - Long Live the Gene: Conceptualising the Gene the Constructionist Way." In *Sociobiology and Bioeconomics. The Theory*

of *Evolution in Biological and Economic Theory*, edited by P Koslowski, 105-37.

Berlin: Springer-Verlag, 1998.

Olby, Robert C. *The Path to the Double Helix*. Seattle: University of Washington Press, 1974.

Oyama, Susan, Paul E Griffiths, and Russell D Gray, eds. *Cycles of Contingency:*

Developmental Systems and Evolution. Cambridge, M.A: MIT Press, 2001.

Portin, Petter. "The Concept of the Gene: Short History and Present Status." *The Quarterly Review of Biology* 68 (2) (1993): 173-223.

Stent, G. "You Can Take the Ethics out of Altruism but You Can't Take the Altruism out of Ethics." *Hastings Center Report* 7, no. 6 (1977): 33-36.

Sterelny, Kim, and Paul E Griffiths. *Sex and Death: An Introduction to the Philosophy of Biology*. Chicago: University of Chicago Press, 1999.

Stotz, Karola, Adam Bostanci, and Paul E Griffiths. "Tracking the Shift to 'Postgenomics'." *Community Genetics* (In Press).

Stotz, Karola, and Paul E Griffiths. "Genes: Philosophical Analyses Put to the Test." *History and Philosophy of the Life Sciences* 26, no. 1 (2004): 5-28.

Waters, C. Kenneth. "Genes Made Molecular." *Philosophy of Science* 61 (1994): 163-85.

———. "Molecules Made Biological." *Rev. Int. de Philosophie* 4, no. 214 (2000): 539-64.

———. "What Was Classical Genetics?" *Studies in History and Philosophy of Science* 35, no. 4 (2004): 783-809.

———. "Why the Antireductionist Consensus Won't Survive the Case of Classical Mendelian Genetics." In *Proceedings of the Biennial Meeting of the Philosophy of*

Science Association, edited by Arthur Fine, Micky Forbes and Linda Wessells,
125-39: Philosophy of Science Association, 1990.

Weber, Marcel. *Philosophy of Experimental Biology*. Cambridge, New York: Cambridge
University Press, 2004.

Williams, George C. *Adaptation & Natural Selection*. Princeton: Princeton University
Press, 1966.

Figure 1

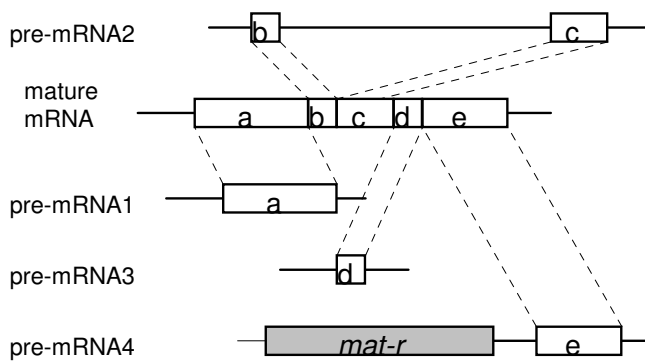


Figure Caption

Figure 1. An example of a 'postgenomic' gene (lines denote introns, boxes denote exons). Subunit 1 of the respiratory chain NADH dehydrogenase is encoded by the gene *nad1*, which in the mitochondrial genomes of flowering plants is fragmented into five coding segments that are scattered over at least 40kb of DNA sequence and interspersed with other unrelated coding sequences. In wheat (illustrated) the five exons that together encode the polypeptide of 325 amino acids, require one *cis*-splicing event (between the exons b/c) and three trans-splicing events (between exons a/b, c/d and d/e) for assembly of the open reading frame. In addition, RNA editing is required, including a C to U substitution to create the initiation codon for this ORF. In some mosses and in mammals the ORF for NAD1 is an uninterrupted stretch of nuclear genomic DNA. Finally, in wheat, a separate, ORF for a maturase enzyme (*mat-r*) is encoded in the intron upstream of exon e (Chapdelaine and Bonen 1991).