PAUL E. GRIFFITHS AND KAROLA STOTZ

5 Gene

The historian Raphael Falk has described the gene as a 'concept in tension' (Falk 2000) – an idea pulled this way and that by the differing demands of different kinds of biological work. Several authors have suggested that in the light of contemporary molecular biology 'gene' is no more than a handy term that acquires a precise meaning only in some specific scientific context in which it is used. Hence the best way to answer the question 'What is a gene?', and the only way to provide a truly *philosophical* answer to that question is to outline the diversity of conceptions of the gene and the reasons for this diversity. In this essay we draw on the extensive literature in the history of biology to explain how the concept has changed over time in response to the changing demands of the biosciences. In this section we have drawn primarily on the work of Raphael Falk (1986, 1991, 1995, 2000, 2001, 2005, in press), Michael Dietrich (2000a, 2000b), Robert Olby (1974, 1985), Petter Portin (1993), and Michael Morange (1998). When our historical claims are commonplaces that can be found in several of these sources we do not cite specific works in their support. We have also chosen not to explain basic genetic terminology, as this would have occupied much of the chapter. More specialized terms are explained when they cannot be avoided. In the final part of the essay we outline some of the conceptions of the gene current today. The seeds of change are implicit in many of those current conceptions and the future of the gene concept appears set to be at as turbulent as its past.

THE INSTRUMENTAL GENE

In the first three decades of genetic research the gene had a dual identity (Falk 1986, 2005). Genes, or Mendelian factors, were

intervening variables defined by the Mendelian pattern of inheritance. From this perspective, the fact that some trait of an organism can be resolved into one or more Mendelian characters establishes definitively that there are genes for those characters. Indeed, it seems that at least some of the earliest Mendelians did not clearly distinguish between the Mendelian character itself and the Mendelian factor 'underlying' it. That distinction was made clear by Wilhelm Johannsen's introduction of the terms 'phenotype' and 'genotype' in 1909. But as well as intervening variables, genes were hypothetical material constituents of the cell whose physical transmission from parent to offspring causally explained the Mendelian pattern of inheritance. In his Nobel Prize acceptance speech Thomas Hunt Morgan, the father of classical genetics, noted, ''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). In our view, one of the clearest themes in the century-long evolution of the concept of the gene is the dialectic between these two conceptions of the gene, a structural conception anchored first in cytology and later in biochemistry, and a functional conception anchored in the observable results of hybridizations, at first between organisms and later directly between DNA molecules.

Recent scholarship has stressed the fact that 'classical genetics' was not merely a theory of heredity, but at least as importantly an experimental practice – 'genetic analysis' – in which the regularities postulated by the Mendelian theory of heredity were used to address other questions about the structure and function of living systems (Waters 2004; Falk in press). This experimental practice imposed strong constraints on the concept of the gene. In the earliest days of Mendelian genetics, William Castle's hybridization experiments with hooded rats challenged the discreteness and constancy of Mendelian factors. In those experiments alleles appeared to be 'contaminated' by the alleles they had shared a cell with in previous generations. The resulting debate exposed a circularity of argumentation: 'unit factors' (individual Mendelian genes) can only be identified by their effect on 'unit characters' (those that display a single, consistent Mendelian pattern of inheritance), but how

can a unit character that is supposed to stand for a unit factor be delimited? This circularity was resolved by definition: Mendelizing traits are determined by a single gene, and non-Mendelizing traits are controlled by more than one gene. The instrumental gene is by definition a Mendelizing unit – it is there to do a job that depends on this stipulation. The visible, heritable characters of organisms must be interpreted in such a way as to permit genetic analysis of those traits. If a character does not correspond to a gene then it must be decomposed into simpler characters that do (later described as 'primary characters'). In the same spirit, quantitative traits, which vary continuously between individuals and thus cannot occur in Mendelian ratios, were treated as the effect of many hypothetical genes, each of which makes an equal and inseparable contribution to the character, giving rise to the discipline of quantitative genetics.

THE MATERIAL GENE

The Morgan school rapidly established the chromosomal theory of heredity, according to which genes are arranged in a linear fashion along the chromosomes that cytologists had observed in the cell nucleus. They were able to explain many deviations from the standard Mendelian pattern of inheritance in terms of the observable behavior of chromosomes. Most importantly, they were able to correlate closely the linkage maps generated by genetic analysis with observable changes in the structure of chromosomes, an achievement facilitated by the discovery of huge, polytenic chromosomes in the salivary glands of *Drosophila*. Linkage was thus both a (functional) measure of the probability that two genes would be inherited together and a (structural) fact about the relative position of visible bands on the salivary gland chromosomes. But despite these achievements, most members of the Morgan school did not concern themselves with the material nature of genes, both because this was not a question that could be pursued via genetic analysis and because the pursuit of genetic analysis did not require it to be answered.

''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). Foremost among these was Herman J. Muller, a student of

Morgan's not satisfied by the purely instrumental notion of the gene as an unknown physical entity localized on chromosomes. For Muller these particulate, atomic entities were the basis, the 'secret' of life, and the essential entities on which the Darwinian process of evolution rests. In order to fulfill these functions genes needed to have the properties of autocatalysis (self-replication) to make them units of heredity, heterocatalysis to allow them to contribute to the phenotype, and mutability to create heritable variation. Muller set up a research programme to study the material nature of the gene and reveal the physical basis of these properties. In 1927 Muller discovered the mutagenic effect of x-rays and used this to make the first estimates of the physical size of an individual gene.

For our purposes, Muller's emphasis on the material gene is important because of his commitment to finding an epistemic pathway to the gene that bypassed the observed effect of the gene of the phenotype. When this commitment started to bear fruit it became possible to advance a concept of the gene that abandoned some of the commitments required if genes were to be epistemically accessible via genetic analysis. Features of the gene that previously could not be meaningfully called into question – and that were thus treated as definitional – became features that could be tested and potentially rejected.

The material nature of the gene was progressively revealed by the new discipline of biochemistry, which came into being in the interwar years. One aim of this discipline was to understand the synthesis of the agents of organic *specificity* – organic molecules that interact only with a very narrow class of other molecules and thus allow the very precise chemistry required by living systems. From the mid-1930s it became increasingly clear that the specificity of organic molecules is explained by *conformation* and *weak interactions* between molecules. The conformation of a molecule is its three-dimensional shape, which determines whether specific sites on molecules can come together. The interactions between those sites are much weaker than the covalent bonds of standard inorganic chemistry, so that interactions between molecules and the conformation of individual molecules can be altered by relatively low energies. These principles turned out to underlie the structure and functioning of all forms of life (Morange 1998, 15). The concept of specificity rapidly began to be applied to the relationship between

genes and their products, as well as to the relationship between enzymes and their substrates.

If the activity of the cell is explained in terms of molecular specificity it is natural to suppose that the effects of genes on phenotypes are mediated by the production of biomolecules with appropriate specificity. Thus in 1941 the 'one gene–one enzyme' hypothesis, which helped to forge an experimental association between biochemistry and genetics, was born. George Beadle and Edward Tatum chose to attack the problem of gene action by genetic analysis of a known biochemical process. They produced and isolated mutant strains of the fungus *Neurospora* each unable to synthesize one of several chemicals involved in a single biosynthetic pathway. Genetic analysis of these mutants showed that each deficiency was the result of a mutation in a single gene. Only three years later Oswald T. Avery produced experimental evidence that genes were made of DNA. Looking back, his evidence seems compelling, but it needed another eight years and a different line of experiment for it to change the received 'protein model of the gene'. If the relationship between genes and enzymes was one of specificity, like the relationship between enzymes and their substrates, then it seemed unlikely that DNA could be responsible for 'genetic specificity'. The little that was known about DNA suggested it was an unspecific and monotonous molecule, perhaps with a structural role in the chromosome.

Historians have stressed the very substantial changes in approach produced by the influx of scientists trained in physics into biology during the 1940s. These changes moved genetics and biochemistry closer together and paved the way for the molecular conception of the gene that prevailed from the 1950s to the 1970s. One of these former physicists, Max Delbrück, was convinced that understanding the secret of life would require a physical approach and an organism as simple and pure as a bacterial virus – an organism so simple that it could be conceived as a naked gene. The bacteriophage appeared to have hardly more than the one key characteristic of life, self-replication, and was thus deemed perfect to study this property ''without opening the biochemical 'black box' '' (Morange 1998, 45). The 'phage group' around Delbrück, Salvador Luria, and Alfred Hershey helped to establish bacterial genetics and the prokaryotic age in genetic research.

DOING WITHOUT GENES?

The clash between the leading geneticist Richard Goldschmidt and his contemporaries in the 1940s and early 1950s provides further insight into the classical gene concept. The successes of the Morgan school in determining the linear order of genes on chromosomes allowed the discovery of 'position effects' in which a change in the relative position of genes on the chromosome is associated with a change in their phenotypic effects. This in turn raises questions concerning the nature of mutation. Today we define a mutation as any heritable change in the nucleotide sequence of a chromosome, which may occur either by the substitution of one nucleotide for another or by the translocation or inversion of a chromosome segment. In classical genetics, however, mutation was necessarily defined as a change in the *intrinsic* nature of an individual gene manifest in a heritable difference in phenotype. Mutations were thus distinguished from position effects, in which an intrinsically identical gene has a different effect because it has changed its location. Goldschmidt challenged this distinction. As there was no direct evidence that chromosomes have distinctive structural parts corresponding to individual genes, he suggested that 'mutations' and 'position effects' were simply smaller and larger changes in the structure of the chromosome. Because chromosomal changes on very different scales were known to have phenotypic effects, Goldschmidt argued that chromosomes probably contained a hierarchy of units of function. Famously, he denied that 'genes' exist, by which he meant that no unique structural unit corresponded to the unit of function of classical genetics. Although ''Goldschmidt's efforts from 1940 to 1958 stand out as one of the first attempts to develop a theory which integrated models of genetic structure, genetic action, developmental processes and evolutionary dynamics'' (Dietrich 2000a, 738), his views were completely unacceptable to most of his contemporaries. Effectively, Goldschmidt was insisting that both aspects of the dual identity of the classical gene converge on a single unit – the material gene must correspond to the instrumental unit of genetic analysis. Evidence to the contrary is thus evidence that there are no genes in the classical sense. Goldschmidt's contemporaries perhaps differed in that they were more hopeful that future discoveries would reveal a unique unit of genetic function at the

molecular level. They certainly differed in their commitment to continuing existing lines of research and unwillingness to undertake the radical reorientation that Goldschmidt was suggesting.

'neo-classical' genetics and the molecular gene

By the mid-1950s DNA was established as the genetic material, its structure had been analyzed by James Watson and Francis Crick (1953), and Crick had stated the 'Central Dogma' of molecular biology and its related 'sequence hypothesis' (1958): the linear sequence of nucleotides in a segment of a DNA molecule determines the linear sequence of nucleotides in an RNA molecule, which in turn determines the sequence of amino acids in a protein by 'informational specificity', that is, via the genetic code whose details were to be elucidated in the early 1960s. The same period saw a sea change in the gene concept itself, one that Petter Portin has labeled the transition from the 'classical' to the 'neo-classical' gene (Portin 1993). It may appear slightly confusing that the latter conception has also been labeled the 'classical molecular gene' (Neumann-Held 1998), but as Portin's 'neo-classicism' is precisely a molecularized classicism, the two names are complementary.

The new, molecular concept of the gene was the result of technical developments that allowed much more detailed maps of the chromosome ('fine structure mapping') and the interpretation of the results of this enhanced form of genetic analysis in the light of the new understanding of the material gene. The new conception departed from the classical in recognizing that the gene is not the fundamental unit of mutation or of genetic recombination. Recombination in classical genetics was the process in which alleles from two copies of a chromosome were combined on a single copy as a result of crossing over between homologous chromosome pairs during meiosis. Recombination was thus recombination of an allele of one gene with an allele of another gene, so that genes themselves were the minimal unit of recombination. Working with bacteriophage from 1954 to 1961 Seymour Benzer was able to increase the resolution of the '*cis-trans*' or 'complementation' test so as to map out in detail the location of different mutations within the same gene and demonstrated conclusively that recombination

can occur between different parts of a single gene. Two mutations are said to be in *cis*-position when they are on the same copy of a chromosome. They are in *trans*-position when one is on each of two homologous chromosomes. The logic of the *cis-trans* test depends on the fact that most mutations are recessive in the heterozygote. Hence, if an offspring derives a mutant allele of one gene from one parent and a mutant allele of another gene from the other parent, it should also receive a mutation-free, functional copy of each gene from the other parent and appear phenotypically normal. If, however, an offspring receives a different mutation from each parent, but they are in the same gene, then it will have no mutation-free copy of that gene and will be a phenotypic mutant. Thus, crossing two mutant lines to produce offspring with the two mutations in *trans*-position tests whether they are in the same gene. If, however, genetic recombination can occur within a single gene, then a small proportion of the offspring of a cross between carriers of two different mutant alleles of the same gene will receive a copy of the gene that recombines the undamaged portion from one mutant allele with the undamaged portion from the other mutant allele and is thus restored to normal function. Benzer used an analogue of the *cis-trans* test in bacteriophage to demonstrate that the gene as a functional unit defined by the *cis-trans* test (the 'cistron') can be represented as a linear recombination map of mutated sites. This acknowledgment led him to distinguish between units of recombination, 'recons', mutation, 'mutons', and genetic function, the 'cistron'.

Benzer's work could have been seen as a vindication of Goldschmidt and other skepticism about the unified, particulate gene (Holmes 2000; Falk 2005). But this was not how it was viewed by his contemporaries. Instead, the cistron was more or less immediately identified with the gene. From this followed the conventional gene concept of molecular biology. One reason the results were interpreted in this way was that the physical structure of the DNA molecule was now known and offered a natural interpretation for Benzer's findings. The unit of recombination and mutation is the single nucleotide, whilst the unit of genetic function (heterocatalysis) is the sequence of nucleotides from which a single RNA is transcribed, corresponding to a single protein, and thus vindicating the existing doctrine of 'one gene–one enzyme'.

CHALLENGES TO THE CLASSICAL MOLECULAR GENE CONCEPT

By the mid-1960s many scientists thought that the major problems of molecular genetics had been solved and were inclined to leave other investigators ''to iron out the details'' (Stent 1968). But the claim that 'what is true for *E. coli* is true for the elephant' turned out to be premature, and it seems unlikely that molecular geneticists will find themselves out of work anytime soon. According to the classical molecular conception a gene is a series of contiguous nucleotides whose sequence corresponds to the sequence of amino acids in a single polypeptide chain (one or more of which makes up a protein). It was soon realized that some genes code for functional RNAs that are not translated to a protein, but this fact is easily accommodated by the classical conception. As C. Kenneth Waters has stressed, the fundamental molecular gene concept is that of a DNA sequence that determines the structure of some *gene product* by linear correspondence (Waters 1994, 2000). The molecular gene is the 'image in the DNA' of the molecule whose biological activity is of interest to the experimenter (Rob D. Knight, pers. comm.). The classical molecular gene seemed to unite the two identities of the classical gene in a single natural unit. The functional definition of the gene that underlay genetic analysis and the structural definition of the material gene had turned out to be two ways to pick out the very same thing. Looked at more closely, however, the functional definition had been significantly revised so as to take account of findings about the material gene. In Muller's original vision genes reproduce themselves (autocatalysis), influence the phenotype (heterocatalysis), and mutate. The classical molecular gene, however, is not the unit of replication, which is the whole DNA molecule of which it is a part. Nor is it the unit of mutation. The only function with respect to which the molecular gene is the unit of function is that of contributing to the phenotype (Muller's heterocatalysis). So the functional role of the gene was revised to fit the molecular reality that had been uncovered. Furthermore, the concept of the gene was restricted to sequences that fulfilled this new functional role: not all segments of chromosomes that behave as Mendelian factors count as genes under the new conception. Untranscribed regulatory regions not immediately adjacent to the

coding sequences they regulate can segregate independently of those coding sequences, and so can function as separate Mendelian factors, but they are not separate molecular genes. Nevertheless, the classical molecular gene was a highly successful example of the research strategy of identifying a functional role, searching for the mechanism that fulfills that role at a lower level of analysis, and using knowledge of that mechanism to refine understanding of function at the original (in this case phenotypic) level of analysis.

Since the 1970s, however, further investigation of the underlying structural unit has tended to undermine the idea that the revised functional role of the gene – determining the structure of a gene product – is filled by natural units of structure at the level of the DNA. The structures in the genome that play a genelike role need not be physically distinct: they can overlap one another or occur inside one another (in the same direction on the DNA molecule or in reverse). The relationship between structural genes and genelike functions is not one to one but many to many: some gene products are made from more than one structural gene and individual structural genes make multiple products. Finally, the sequence of elements in the gene product depends on much more than the sequence of nucleotides in the structural gene: different sequence elements can be repeated, scrambled, and reversed in the product, and the precise sequence of a gene product can reflect posttranscriptional and translational processing as well as the original DNA sequence. To put flesh on these bones we will briefly describe some of these mechanisms and give an example (Figure 5.1). (For more examples, details, and references, see Stotz and Griffiths 2004; Stotz, Bostanci, and Griffiths 2006; Stotz 2006.)

In eukaryotes (organisms whose cells have a nucleus and organelles, including fungi, plants, and animals) the DNA sequence is transcribed into a premessenger RNA (pre-mRNA) from which the final RNA transcript is processed by cutting out large noncoding sequences, called *introns*, and splicing together the remaining coding sequences, the *exons*. Biologists speak of alternative *cis*-splicing[1] when more than one mature mRNA transcript results from these processes through the cutting and joining of alternative exons. Adjacent genes are sometimes cotranscribed, that is, transcribed together to produce a *single* pre-mRNA that is then spliced. Splicing may also occur between a gene and an adjacent 'pseudogene' that

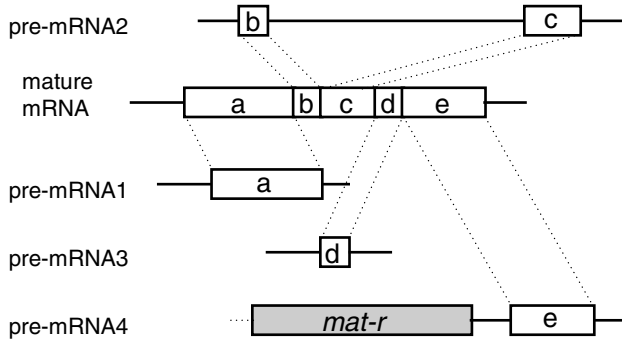

Figure 5.1 A contemporary molecular gene. Lines denote introns; boxes denote exons. Subunit 1 of the respiratory chain NADH dehydrogenase is encoded by the gene *nad1*, which in the mitochondrial genomes of flowering plants is fragmented into five coding segments that are scattered over at least 40 kb 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). For more examples, visit http://representinggenes.org.

would be incapable of producing a product on its own. Alternative gene products may also be derived from so-called overlapping genes. In these cases, the 'genes', in the sense of the 'open reading frames' (ORFs) that are transcribed into RNA, are not lined up like so many pearls on a string, but instead may overlap one another or even be completely contained one within another. While some cases of alternative splicing produce a range of proteins that are structurally related to one another, in other cases the products are quite different from each other (in which case they are often described as products of overlapping genes, rather than alternative splicing of the same gene). The degree of difference between the products depends on the extent of overlap between their exons, 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 in an English sentence: 'A gene is a flexible entity' becomes 'Age nei saf lex ibl een tit y'. But unlike any human language, a DNA sequence is always made up of meaningful 'three-letter words' (codons that specify an amino acid during translation) no matter where 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 noncoding RNAs (such as microRNAs) from the intronic regions of the premature transcript, which may be involved in the regulation of coding transcript of the same gene.

In the process of *trans*-splicing a final mRNA transcript is processed from two or more independently transcribed pre-mRNAs. Whilst the prefix *trans* might suggest that these pre-mRNAs are derived from DNA sequences far apart from each other, this is by no means always the case. In fact, two copies of the very same sequence can be spliced together this way, as can alternative exons in what would at first glance look like a 'normal' case of *cis*-splicing. Moreover, 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, will result in an antisense transcript with likely regulatory function, possibly through silencing its complementary 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 posttranscriptional modifications of mRNA (capping, polyadenylation, and *cis*-splicing) retain the *correspondence* of the primary structure of coding sequence and gene product, RNA editing disturbs this correspondence by changing the primary sequence of mRNA after its transcription. The 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 varieties of RNA editing as there are organisms, all belong to

one of three known mechanisms: 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). Although we will not describe them here, other processes may occur before 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, RNA processing, and translation were clarified in the 1960s.

THE MODERN GENE

The 'modern gene' as Portin (1993) has termed it represents a further stage in the dialectic of structure and function described. The classical gene, primarily defined by the functional role it played in heredity, became identified with the structural gene revealed by early molecular biology, primarily through the study of prokaryotes and bacteriophage. As a result, the functional role of the gene was redefined as the determination, by linear correspondence, of the structure of a gene product. Further investigation of the manner in which a wider range of genomes generate a wider range of gene products has revealed that this functional role can be filled by diverse, highly flexible mechanisms at the level of the DNA itself: ''We are currently left with a rather abstract, open and generalized concept of the gene, even though our comprehension of the structure and organization of the genetic material has greatly increased'' (Portin 1993, 173). Goldschmidt's critique of the particulate gene has been explicitly revived in the light of our new understanding of genome structure and function:

> The particulate 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)

In a prescient paper twenty years ago, Raphael Falk reviewed what were then newly emerging challenges to the classical molecular gene and concluded:

> Today the gene is not just *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 need. It is neither discrete – there are overlapping genes, nor continuous – there are introns within genes, nor does it have a constant location – there are transposons, nor a clearcut function – there are pseudogenes, not even constant sequences – there are consensus sequences, nor definite borderlines – there are variable sequences both 'upstream' and 'downstream'. (Falk 1986, 169)

Thus, as early as 1986 we were well on the way from the ''well defined material entity back to an abstraction, a hypothetical construct, if not an intervening variable, devised by scientist for their needs'' (Falk 1986, 160).

Focusing on the cutting edge of contemporary genomics can induce an extremely deflationary view of the gene. Some molecular biologists, realizing that the concepts of 'gene' transcription or 'gene' expression may not suffice to capture the variation in expressed genomic sequences, have proposed the more general term '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). 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. One influential recent review concludes that ''one solution for annotating genes in sequenced genomes may be to return to the original definition of a gene – a sequence encoding a functional product – and use functional genomics to identify them'' (Snyder and Gerstein 2003, 260).

The gene concept, however, plays a role in many other contexts besides the cutting edge of genomics (Stotz, Griffiths, and Knight 2004). We suggest, therefore, that there are at least three answers to

the question 'What is a gene?', none of which can be neglected if we hope to depict the state of contemporary biology accurately. These are the traditional, instrumental gene; the postgenomic molecular gene; and the 'nominal gene'.

THE TRADITIONAL GENE

Biologists can and do still use genetic analysis – the analysis of the phenomenon of heredity by the analysis of the results of hybridization, either between organisms or directly between DNA molecules (Waters 2004; Falk in press). Genetic analysis remains a key tool in addressing broader biological questions. For these purposes the gene remains an intervening variable, defined by the inheritance patterns that it enables us to follow, and the difficulties of providing a univocal account of its identity as a material unit can be put to one side. The traditional gene concept is retained in much the same way in population genetics. In an important recent analysis, Lenny D. Moss introduces the term 'Gene-P' for something very like our 'traditional' gene (Moss 2003). The *P* stands variously for 'phenotype', 'prediction', and 'preformation' since these genes are identified in terms of their phenotypic effects, are used to predict the phenotypic results of hybridization, and reflect what Moss terms 'instrumental preformationism' – a strategic neglect of the ways in which the gene-phene relationship depends upon other factors. Moss contrasts his Gene-P to a materialistic concept of the gene that he calls 'Gene-D' (for 'development'). Genes-D are defined by their intrinsic chemical capacity to template for gene products. Here, we wish to distinguish two importantly different ways to conceptualize genes that fall within the general area of conceptual space that Moss labels Gene-D.[2]

THE POSTGENOMIC MOLECULAR GENE

We use the phrase 'postgenomic molecular gene' to refer to the entities that continue to play the functional role of the molecular gene – making gene products – in contemporary molecular biology. The postgenomic molecular gene 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. These genes are ''things you can do with your genome'' (Stotz, Bostanci, and Griffiths 2006): although the gene is still an 'image' in the DNA of the target molecule (the molecule whose activity we wish to understand) this image may be fragmented or distorted to such an extent that it cannot be discerned until functional genomics has revealed how these sequence elements are used in the broader genomic and cellular context. This conception of the gene remains a critical aspect of the epistemology of molecular bioscience simply because linear correspondence between molecules is fundamental to biologists' ability to identify and manipulate them, via technologies ranging from cDNA libraries to microarrays to RNA interference. But although it is important to know the 'gene for' some molecule in this sense, it does not matter very much whether that collection of sequence elements is a gene! To put it less paradoxically, the utility of knowing the DNA elements that underlie the production of the target molecule or its precursors does not at all depend on whether it is possible to give a univocal definition of the material gene. Finding the 'gene for' the molecule in this sense remains important even on the most deflationary, postgenomic view of the molecular gene.

NOMINAL GENE

> 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 2004b, 64–65)

It is hard to disagree that for many practical purposes genes are simply sequences that have been annotated as genes and whose annotation as such has been accepted by the scientific community. But, as Burian himself makes clear, 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

working concept of the gene, according to Fogle, is something like a stereotype or prototype: a sequence is a gene if it has enough similarities to other genes: for example, it contains an open reading frame, has one or more promoters, has one or more transcripts that are not too functionally diverse from one another, and so on. This is more or less a description of automated 'gene discovery' methods, and Fogle's suggestion is that the *concept* of the gene is no more principled or definitionlike than this. The various 'genelike' 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 the example in Figure 5.1), at other times giving up on functional criteria to save structural ones (as in cotranscription of a gene and a 'pseudogene').

Fogle is quite critical of this state of affairs, arguing that combining structural and functional features into a single stereotype, what he calls the 'consensus' gene, hides both the diversity of DNA sequences that can perform the same function and the diverse functions of particular DNA sequences. Burian takes a more positive view, emphasising the value of simply having a shared collection of named sequences known or suspected to be involved in the production of gene products.

CONCLUSION

The gene began life as an intervening variable, defined functionally in terms of the Mendelian pattern of heredity in observable phenotypic characters. It rapidly acquired a second identity as a hypothetical material unit. A productive dialectic between investigations of the gene that identified it in each of these two ways concluded with the 'neo-classical' or 'classical molecular' conception of the gene. The functional role of the gene was redefined to exclude mutation and recombination, which became properties of the DNA in its own right, rather than of individual genes. The function of the gene became the determination of the structure of gene products via linear correspondence between molecules. This functional role was played by a natural class of units at the molecular level – the structurally defined molecular gene. Further investigation of a wider

range of genomes and a wider range of gene products has thrown into doubt whether an adequate structural definition is possible – the structural basis upon which gene products are generated may be a very broad class of 'things you can do with your genome'. At this point it remains possible to think of genes in the traditional manner that dates back to the early twentieth century as intervening variable in the genetic analysis of phenotypes. It is also possible to think of them as the often complex collections of sequence elements that fill the functional role of the molecular gene ('postgenomic molecular genes'). Finally, it is possible to think of genes as simply those sequences whose similarity on various dimensions to stereotypical genes has led them to be annotated as genes and whose annotation as such has been accepted by the scientific community ('nominal genes').

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants 0217567 and 0323496 and supplemental funding from the University of Pittsburgh. Griffiths's work on this chapter was supported by an Australian Research Council Federation Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

NOTES

1. 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 posttranscriptional processing (*trans*-splicing). Thus *trans-* elements in the modern sense (*trans-* on mRNA) may be *cis-* located on the DNA.
2. Moss (pers. comm.) suggests that our 'postgenomic molecular' and 'nominal' material genes are perspectives on genes-D corresponding to what are, somewhat perversely, called 'forward' and 'reverse' genetics. The postgenomic molecular gene embodies the traditional, 'forward', strategy of locating the template resources corresponding to a known phenotype. The nominal gene is a template resource whose use we set out to understand.