Hist. Phil. Life Sci., 26 (2004), 5-28

Genes: Philosophical Analyses Put to the Test

Karola Stotz[1] and Paul Griffiths[2]

(1) Department of History and Philosophy of Science
1017 Cathedral of Learning
Pittsburgh, PA 15260, USA

(2) *ARC Federation Fellow in Biohumanities,*
University of Queensland
Brisbane, Queensland 4072, Australia

ABSTRACT - This paper describes one complete and one ongoing empirical study in which philosophical analyses of the concept of the gene were operationalized and tested against questionnaire data obtained from working biologists to determine whether and when biologists conceive genes in the ways suggested. These studies throw light on how different gene concepts contribute to biological research. Their aim is not to arrive at one or more correct 'definitions' of the gene, but rather to map out the variation in the gene concept and to explore its causes and its effects.

KEYWORDS: gene concept, conceptual change, gene expression, gene-D, gene-P, genomics

1. Introduction: Empirical Philosophy Tracks Shift in a Scientific Concept

This is a particularly exciting time to be studying molecular bioscience because of the extraordinary rate of change in basic concepts. Discoveries that would constitute a 'scientific revolution' in many disciplines are regular occurrences. The ways in which bioscientists conceptualize DNA and related molecules are thus not only fascinating in their own right, but also an important case study for the history and philosophy of science – a case study of conceptual change and its role in science.

Empirical science is a powerhouse of conceptual innovation. Scientists use and reuse their terminology in a way that Hans-Jörg Rheinberger has accurately characterized as 'exuberant' (Rheinberger 2000). Examples for this are the use of '*cis*' and '*trans*' without reference to the cis-trans test, or the use of the term 'exon' to refer to a stretch of DNA which is included in the mature mRNA transcript after various forms of post-transcriptional processing but is not translated into

protein. Both these usages contradict all but the most recent textbook definitions and, speaking from experience, seem irritating and wrongheaded to some biologists. Yet both usages are common in certain research communities.[1] For the practitioner concepts are tools that classify experience shaped by experimentalists to meet their specific needs and reshaped in the light of new empirical findings and as those needs change. This attitude is sometimes made explicit: '…we must sharpen our conceptual tools as best we can and have faith that in using them to untangle the complexity we shall see how to fashion better ones' (Hinde 1985, 990). The same attitude is implicit when scientists describe a statement as a 'definition' and yet regard it as hostage to future empirical findings, as they commonly do. If scientific concepts are evolving tools, it should not be the aim of philosophers of science to identify the one correct conception associated with a word or phrase, or even to eliminate slippage of meaning caused by the presence of several different conceptions in a single community. Historical analysis has shown how slippage of meaning was essential to the rapid progress of genetics in the first half of this century (Falk 1986; Rheinberger 2000; Falk 2000). History of science thus allows us to take what James G. Lennox has called a 'phylogenetic' approach to the study of science. Like the history of a biological species, the history of a concept allows the reconstruction of a transformation series and provides evidence about the selective pressures that drove change and diversification.[2] If history of science can become *conceptual phylogenetics*, philosophy of science, so we believe, can and should take on the corresponding role of *conceptual ecology*. In the research reported here we are trying to describe the current diversification of the gene concept to meet the increasingly diverse goals of workers in the increasingly diverse set of fields we call 'molecular bioscience'. At the very least, we hope to discern some species boundaries or incipient speciation events, at best to throw some light on the particular epistemic pressures that have caused scientists to explore one *conceptual niche* or another.

In the next section we give an overview of some historical and philosophical analyses of the gene concept. Section three summarizes the objective, design and results of a small-scale survey of molecular biologists at Sydney University (Stotz, Griffiths, and Knight 2004) which acted as a preliminary study for the ongoing project described

[1] For a similar example, see (Falk In Press).

[2] It remains to be determined whether and to what extent conceptual phylogenies mirror sociological descent relations amongst scientists.

in section four. The methodologically innovative element of this ongoing research is the integration of empirical research methods with the traditional philosophical task of analyzing key concepts and elucidating their role in scientific reasoning. With the help of a group of experts in fields ranging from history and philosophy of science, bioethics and science communication, to practitioners of molecular bioscience we have administered an online survey to a wide range of working biologists in order to explore the prevalence of different conceptions of the gene in various fields.

2. The Concept of the Gene

Since the term 'gene' was introduced a century has passed during which its meaning has been transformed almost beyond recognition (Johannsen 1909; Keller 2000). Despite the lack of a stable definition the gene has proven to be an enormously fruitful scientific tool, and in addition has become a cultural icon and the carrier of multiple hopes and promises within science and medicine. Throughout its history there has been a tension between the current conception of the gene and the latest empirical results. Following the general acceptance of the view that genes can be structurally identified as sections of chromosomes, the function of those chromosome segments evolved from determining a unit character, to determining an enzyme, to determining a polypeptide. This dialectical development of the gene concept can be interpreted as reflecting a desire to keep the structural and functional definitions of the gene focused on a single entity. When the best structural definition turns out to create units with indeterminate function, structure and function can be brought back into step by using a more proximal description of function: rather than a gene having an indeterminate effect on the phenotype, it has a determinate effect on one of the structural elements that contributes to the phenotype (Griffiths and Neumann-Held 1999; Kitcher 1984). This process of conceptual evolution seemed to reach a stable resting point with the 'Classical Molecular Gene Concept' (Neumann-Held 1999). The classical molecular gene has a clear functional identity as a segment of DNA that codes for a single polypeptide chain.[3] It has a clear structural identity as an open reading frame with adjacent promoter. This structure and this function are taken to be tightly associated with one another.

[3] Little violence is done to the classical conception by extending it to cover genes that determine the sequence of a functional RNA rather than a polypeptide.

The reality of genome structure today challenges the classical picture of the molecular gene in the same way that the reality of particle physics challenges the traditional picture of matter. The 'particles' of the quantum world can lack such apparently essential features as having mass or being in some particular place. In the same way, just about any of the normal expectations we have when we hear the word 'gene' is violated by some important class of DNA sequences. Physicists changed their concept of a particle in response to the strange world that quantum physics revealed. Just so, in the 'post-genomic' world bioscientists continue to talk about 'genes' but often mean something quite at odds with the picture of the gene found in introductory textbooks.

So what exactly are the problems that confront the classical molecular picture of a gene? Most of them arise from complications that have been added to the classical picture of gene expression. Classically, gene expression proceeds via the transcription of the continuous open reading frame of a classical molecular gene into a single messenger RNA and the translation of that mRNA into a polypeptide, both processes starting at a determinate point on the nucleic acid molecule which is being processed. An immediate complication arises in eukaryotes (fungi, plants and animals), since in eukaryote cells a primary or pre-messenger RNA is transcribed from the DNA sequence and the final mRNA transcript is derived from this by cutting out non-coding sequences (introns), and splicing together of the remaining coding sequences (exons). This alternative (*cis*-) splicing means that one classical gene can correspond to more than one polypeptide, often far more than one. The resultant 'one-many' relationship between stretches of DNA and gene products is further complicated in cases of 'overlapping genes' or 'gene sharing' (Burian, this volume). Genes are not lined up on a chromosome like pearls on a string, but instead one gene can start within another gene. The similarity of the products depends on the proportion of shared sequences, and on whether these shared sequences are read in the same frame. If transcription of the second gene does not begin at the beginning of a codon of the first gene, then the reading frame is altered and the two transcripts might share much of their sequence whilst coding for quite different products.[4] An even more surprising

[4] 'Frameshift' in reading English would turn 'The old man can run' into the meaningless' (T) heo ldm anc anr un'. In the genetic code, however, every sequence of three 'letters' (bases) is guaranteed to be a meaningful 'word' (codon). A good example of such a case is (Sharpless and DePinho 1999).

source of 'one-many' relationships arises from the discovery that the same DNA can be read in both directions. A single stranded DNA can only be read in one direction, from its 3' end to its 5' end, but the complimentary strand can also be read from *its* 3' end to *its* 5' end. The discovery of this process of 'antisense transcription' has demonstrated that the very same stretch of DNA can encode gene products as radically different as, in one case, a polypeptide and a functional RNA (Coelho *et al.* 2002).

The relationship between DNA and product can be 'many-one' as well as 'one-many'. For example, adjacent genes can be co-transcribed to produce a single 'fusion transcript'.[5] Or, in the phenomenon of *trans*-splicing, a final mRNA transcript is produced by splicing together from more than one independently transcribed pre mRNA. Like *cis*-splicing, *trans*-splicing can occur in alternate forms to make several products from the same collection of DNA elements (Finta and Zaphiropoulos 2000b; Pirrotta 2002). The 'many-many' relationship between DNA and gene products that has been revealed by the discovery of these processes leaves one critical feature of the classical conception untouched. The linear order of amino acids in the gene product still corresponds to the linear order of DNA bases in some set of sequences of DNA read in some frame or another. But this too turns out not to be a universal feature of gene expression. Exons can be repeated, they can be put together in a new order (exon scrambling) and they can be inverted in the final transcript (antisense *trans*-splicing), so that the linear order of codons no longer corresponds to that of bases in the DNA from which they are derived.[6] Moreover, the complete mRNA transcript can be edited one base at a time before translation. This process of 'mRNA editing' converts C bases into U bases, which can have such radical effects as truncating translation by introducing a novel stop codon. Last but not least, the recently discovered process of 'protein splicing' changes the final product once more, but in this case by splicing 'inteins' in and out of the actual polypeptide (Liu 2000).

In the face of these complications, the statement that a gene is a DNA segment which determines a polypeptide is an insufficient basis on which to answer such apparently simple questions as: 'How many genes are contained in a given genome?'; 'Where does one gene ends and

[5] See for example (Magrangeas *et al*. 1998; Finta and Zaphiropoulos 2000; Communi *et al.* 2001).
[6] For a range of examples see (Takahara *et al.* 2002; Caudevilla *et al.* 1998; Flouriot *et al.* 2002).

another begin?'; and 'To which gene does this particular segment of DNA belong?'. The recent literature both in the philosophy of biology and in biology itself contains many proposals that aim to answer these questions in a principled and useful manner. Our work is in part an attempt to test whether the claims made by these authors reflect how biologists in various fields have chosen to handle these issues.

Some Alternative Conceptions of the Gene

The geneticist and historian of genetics Raphael Falk distinguishes four ways in which genes might be conceived (Falk 2000). First, the gene can be understood abstractly as something which figures in certain calculations, as in population genetics and possibly in predictive medicine. This conception of the gene resembles what Lenny Moss has termed 'Gene-P' (see below) and, more generally, a 'top-down' approach to genes that identifies them via their effects. Second, genes can be primarily conceived as material, structural entities, perhaps consistently associated with certain functions. Seymour Benzer's interpretation of H.J. Muller's particulate gene falls into this category, which Falk thinks is a good candidate of the genetic engineer's ideal but which he doubts is viable as an approach to the whole genome and its significance. A third approach conceives genes as functional, biological entities, whose structural identity is secondary, allowing for multiple realizability of 'the same gene' in different DNA. The fourth approach conceives genes as generic operational entities, with the term 'gene' merely shorthand for whatever class of DNA elements is currently of interest. Falk thinks this pragmatic approach to the concept has been adopted by many molecular biologists.

The biologist Thomas Fogle has suggested that biologists use what he has termed a 'consensus gene' concept: a collection of flexibly applied parameters of features of well-defined genes (Fogle 2000). A gene is a sequence that has 'enough' of the features of the gene stereotype (e.g. has an RNA transcript, has a TATA box, contains an ORF, etc., etc.). Fogle argues that by combining structural and functional features in a single stereotype the consensus concept hides both the diversity of structure that can perform the same function and the diverse functional roles of the same structures. We would argue further that as is the case with stereotypes more generally, even when people have been exposed to cases that violate the stereotype, they tend to forget the problematic cases and revert to the stereotype in future work. It is perhaps for this reason that many of the forms of gene expression introduced in the last

section were initially treated as exceptions that happen either only in 'low' animals (prokaryotes; trypanosomes; nematodes), or in very rare instances (in certain cells at very particular developmental stages), or in organelle genomes (mitochondria; chloroplast).

One of the best known attempts to go beyond the classical conception of the gene is Kenneth C. Waters' 'fundamental' molecular gene concept. The fundamental element of the molecular gene concept according to Waters is the preservation of the linear sequence of the original DNA sequence in the product, however proximally or distally defined. As Rob D. Knight has usefully expressed it, the gene is the 'image in the DNA' of a gene product (personal communication). Within this guiding conception, the uses of the term 'gene' in molecular biology corresponds to research interests along a continuum of more or less distal stages of gene expression (Waters 1994; 2000). Waters' view expresses a critical insight into the conceptual structure of classical molecular biology, but it is unclear that the resulting conception of the gene is adequate to describe transcription events that include phenomenon such as mRNA editing or antisense *trans*-splicing (see above).

Another important recent analysis is that of Lenny Moss, who has argued that there are two distinct conceptions of gene in play in current scientific and clinical thought (Moss 2002). These are 'Gene-D' and 'Gene-P', each heir to one of two major historical schools of embryological thought, preformationism and epigenesis.[7] The Gene P conception treats genes as statistically valid predictors of phenotypes and abstracts away from the molecular nature of the DNA elements that underly these statistical patterns. In the simplest case, different genetic lesions that impair one or more of the functions of a DNA element to the same degree are treated as 'the' allele for the impaired phenotype(s). In contrast, the Gene D conception focuses on the intrinsic capacity of a given sequence to template for RNAs. Gene-D thus bears some resemblance to a suggested 'Contemporary Molecular Gene Concept' (Knight and Griffiths 1999) according to which a gene is a DNA sequence that is expressed as a particular range of molecular products across a range of cellular condition. In other words, one gene equals one molecular norm of reaction (see for similar conceptions Alberts *et al.* 2002; Falk 2001).

[7] For the idea that conceptions of the gene are fundamentally *embryological* in nature, see (Griesemer 2000)

Like Fogle, Moss thinks that current conceptions of the gene can hinder as well as facilitate research. Each of the two conceptions has a valuable role in the research contexts in which it arose, but their conflation into a single 'informational gene' whose intrinsic molecular nature is strongly linked to its ultimate phenotypic effect leads to a simplistic and unhelpful conception of genetic causation

3. How Scientists Conceptualize Genes: An Empirical Study

A survey of some of the analyses of the gene concept described in section two led us to advance some hypothesis about how genes are conceptualized differently in different fields of biology. In this section we summarize the objectives, methods and results of a questionnaire study of 81 biological scientists at the University of Sydney, Australia, in 2000 designed to test those hypotheses. We cannot give a detailed account of the study here for reasons of space. For full details see (Stotz, Griffiths, and Knight 2004) and supporting online materials cited therein. The results provide tentative support for our three hypotheses:

1. Hypothesis One: Molecular Versus Evolutionary Biologists. We expected molecular biologists to emphasize the investigation of the intrinsic, structural nature of the gene and to be reluctant to identify a gene only by its contributions to relatively distant levels of gene expression. Conversely, evolutionary biologists should be more interested in genes as markers of phenotypic effects and reluctant to treat two similar DNA sequences as the same gene when they lead to different outcomes for the larger system in which they are embedded.

2. Hypothesis Two: Developmental Versus Evolutionary Biologists. A second expectation was that developmental biologists would emphasize the intrinsic nature of the gene as a molecular object and contextual effects on gene expression, whereas evolutionary biologists would emphasize the predictive relationship between genes and phenotypes. Consequently, there should be stronger support for the informational concept of the gene from evolutionists.

3. Hypothesis Three: Molecular Versus Developmental Biologists. We expected developmental biologists to be less attracted to Moss's Gene-P and to the informational conception of the gene than (other) molecular biologists. We expected developmental biologists to be attracted to conceptions that emphasize contingency and context

dependency, such as Moss's Gene-D and various developmentally-oriented conceptions of the gene canvassed in the literature on evolutionary developmental biology.

In addition to these specific hypotheses, we saw this as an exploratory study and were interested in what the responses suggest about the general state of the gene concept in contemporary biology. We also examined the effects of age and gender.

The questionnaire had three sections, the first part designed to determine the subject's research field, the second asking them direct questions about the gene concept and the third asking them to apply the gene concept to specific cases. The first section of the questionnaire gathered data on the professional training, research experience and current research field of subjects, along with age and gender. The second section of the questionnaire contained direct questions about the definition of the gene, the function of the gene and the methodological value of the gene concept. The answer alternatives for each question were designed to capture the various conceptions of the gene discussed in the literature. We used a number of different formulations of each conception to avoid superficial effects, such as antipathy to particular words or phrases. The actual wordings of many of the answer alternatives were taken from the literature and from genomics websites.[8] Each question had an 'Other' alternative in which subjects could supply their own answer, but no useful data was obtained by this means. This section of the questionnaire contained both 'free choice' and 'forced choice' tasks. The former required subjects to indicate for each question all the answer alternatives to which they could agree. The latter required subjects to choose the single best answer amongst the alternatives offered. The third section of the questionnaire was based on the design of an informal study conducted by Rob D. Knight in New Zealand with 10 respondents. This section used 'indirect' questions that asked subjects to apply their conception of the gene, rather than to answer questions about it. Subjects were given twenty-two examples of specific ways in which two DNA sequences could differ from one another and asked whether, in each case, two such DNA sequences would be two copies of the same gene.

[8] An annotated version of the questionnaire indicating these sources is available in the online documents cited in Stotz, Griffiths and Knight 2004, documents which are located on the Philosophy of Science Association preprint server (http://philsci-archive.pitt.edu).

To test our hypotheses we identified those answers that, if the hypothesis were correct, should be more attractive to one group than another. For example, to test our hypothesis one using the free choice responses to section 2, we identified those answers to the five questions in section 2 that we expected would be more attractive to the molecular group than to the evolutionary group and, conversely, those answers that we expected would be more attractive to the evolutionary group than to the molecular group. We then tested for significant differences between the responses of the two groups in the predicted directions. As an example, the results for hypothesis three are given in Table 1.

1.6	D 50%, M 21%, ns	1.1	M 43%, D 14%, ns
2.4-7	D 60%, M 53%, ns	2.1	**M 41%, D 0%, .353/.041**
4.3	D 88%, M 75%, ns	2.2	M 72%, E 60%, ns
4.4v6	D 60%, M 55%, ns	4.1	M 73%, D 60%, ns
5.5	**D 40%, M 13%, .282/.048**	5.4	M 20%, D 0%, ns

TABLE 1. Test of hypothesis 3 with data from the free choice task. Left-hand column shows answer alternatives for which we predicted agreement by the developmental group (D), right-hand column those for which we predicted agreement by the molecular group (M). Result cells: the numbers behind the characters show percentage of yes answers among the respective group (D, M), the following fractions indicate strength (from 0 to 1) and significance (0 – 1) of association. Results marked ns were not significant (>10% or .100). **Bold** results indicate high significance (< 5% or .050), italic results show associations in the reverse direction to that predicted.

The purpose of the forced choice task was to reveal differences hidden by the free choice task, in which minimally acceptable options would not be distinguished from highly preferred options. Just as with the free choice task we predicted the answers that we expected from each group for each question. Because this was a forced choice task in which each subject chose only one option; in cases where more than one answer option seemed likely to be preferred by a particular group we coded a strong disjunction of these answers as a single answer. Table 2 shows the result of our grouping and recoding exercise for the forced choice answers according to hypothesis three.

Predictions for Developmental Group	Predictions for Molecular Group
1.3, 1.4, 1.6	1.1, 1.5
2.4, 2.5, 2.6, 2.7, 2.8	2.1, 2.2, 2.3
4.3, 4.4, 4.6	4.1
5.5, 5.7, 5.9	5.2, 5.3, 5.4, 5.6

TABLE 2. Grouped force-choice prediction for Hypothesis 3. The answer alternatives in each cell were combined by strong disjunction on the grounds of their expected appeal to one group.

Once again, we tested out hypotheses by looking for significant differences between our groups in the predicted directions. As an example, Figure 1 shows the results for hypothesis three using data from section 2, question 1.

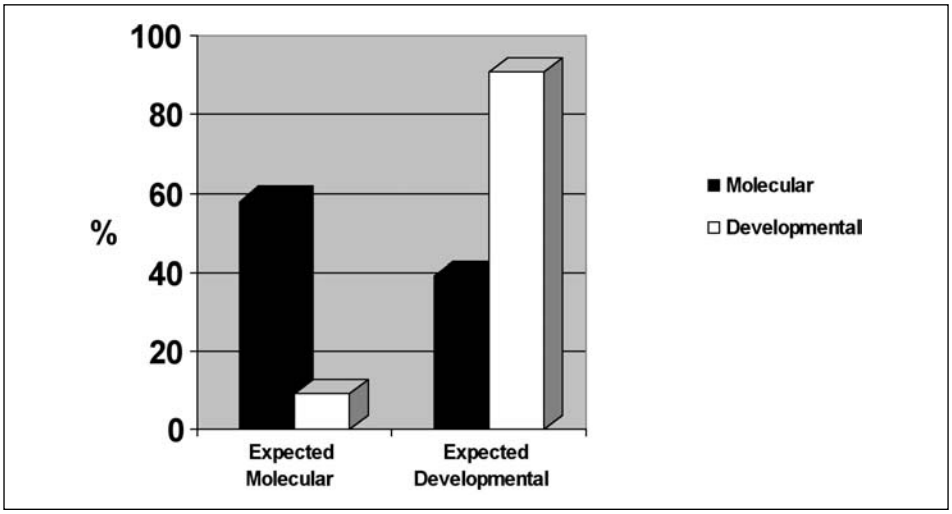

FIGURE 1. Results for Hypothesis 3 for the forced choice task on Section 2, Question 1 of the questionnaire. Paired columns show percentage of 'molecular' and 'developmental' respondents who actually gave the predicted answers for molecular and for developmental respondents. Association 0.430, significance 0.011.

The results from the analysis of the free and the forced choice task of section 2 of the questionnaire did not confirm our hypotheses one

and two, both involving evolutionary biologists as the comparison group. The results from this section, in which we asked direct questions, suggest that biologists with an evolutionary focus in their research do not conceptualize genes in terms of their phenotypic effects in any way that distinguishes them from biologists with a purely molecular or a developmental research focus. A very different picture emerged, however, from responses to the *indirect* questions in Section 3, in which the pattern of answers showed the evolutionary group responding significantly more strongly to changes in distal function than the molecular group.

Overall, the results provided tentative support for our three hypotheses. Hypothesis three seems most strongly supported. Biologists whose research focus is in developmental biology seem to conceptualize genes in a distinctive way, a way that appears to reflect their use of the gene concept to investigate the complex, developmental pathways through which genes are expressed. Hypotheses one and two, which suggest, in broad terms, that biologists whose research focus is in evolutionary biology, conceptualize genes primarily via their effects on phenotypes, are supported in some tests but not others. The fact that the hypotheses are supported when indirect questions are used, but not when direct questions are used, led us to advance an intriguing further hypothesis. We proposed that these biologists may have an explicit belief that genes are molecular entities and should be defined and investigated at that level, whilst deploying in their actual thinking about genetic problems a conception of the gene that abstracts away from differences at the molecular level and focuses on phenotypic effects. We hope to investigate this hypothesis in future research.

Our general results for the whole subject population were consistent with Fogle's suggestion that the classical molecular gene concept continues to function as something like a stereotype for biologists, despite the many cases in which that conception does not give a principled answer to the question of whether a particular sequence is a gene (Fogle 2000). Given the extensive psychological literature on prototype-based categorisation and on the reasoning processes it supports, this also suggests productive lines of future inquiry. Given the small number of subjects in this study and the simple criteria used to group them, we were encouraged by the ability of the study to discern differences between the groups. In ongoing research in the United States (see next section) we have been able to increase the number of subjects by half an order of magnitude (from

81 to 500) by recruiting subjects via the mailing lists of scientific societies and have used more sensitive measures to define our groups.

4. The Representing Genes Project[9]

Our earlier results indicate the importance of distinguishing between explicit and implicit ideas about the gene. In our ongoing research we asked subjects to engage in tasks such as ranking research proposals, or annotating transcription events, which were designed to reveal their implicit understanding. These tasks have the added advantage of providing numerical in addition to categorical data, allowing a wider range of statistical procedures to be employed. The study is not yet complete, and here we will only sketch its design and discuss its potential significance.

The research instrument was the outcome of a workshop at which we and our collaborators[10] agreed on a range of research questions we wanted to see addressed and on strategies for operationalizating those questions. This resulted in an instrument with four sections, three of which will be briefly summarized here. These were a task in 'abstract sequence annotation', a 'review' of competing proposals to research complex genetic diseases, and a set of questions designed to distinguish biologists working in different fields for comparison with one another. The fourth section asked some qualitative (open-ended) questions about the biologists' understanding of and attitude towards some prevalent genetic metaphors, including 'genetic information', 'gene for', 'genetic program', and 'developmental program'.

4.1 Task in 'Abstract Sequence Annotation'

We showed in section two that the classical molecular gene concept leaves open many decisions researchers have to make when annotating genomic sequences. In the post-genomic world it is not at all obvious which sequences to count as genes. The simplest product of a genome 'annotation' is an overall figure for the number of genes in a genome, such

[9] Project website: http://www.pitt.edu/~kstotz/genes/genes.html

[10] The following people have actively collaborated with us on the Representing Genes Project: Richard Burian, Sharyn Clough, Raphael Falk, Thomas Fogle, Scott Gilbert, James Griesemer, Jonathan Kaplan, Evelyn Fox Keller, Rob Knight, Brendan Larson, Lenny Moss, Hans-Jörg Rheinberger, Jason Scott Robert, Sahotra Sarkar, Kenneth Schaffner, Kenneth Waters, and from the University of Pittsburgh Ingo Brigandt, Megan Delahanty, James Lennox, Alan Love, Sandy Mitchell, Robert Olby, Lisa Parker, Jeff Schwartz, and James Tabery.

as the strikingly low figures produced in the immediate aftermath of the sequencing of the human genome, ranging from a conservative 26 000 to a generous 40 000. But simple gene counts disguise the highly problematic nature of most initial annotations. In our current state of knowledge many different approaches to identifying genes are defensible, and different research groups often produce only partially overlapping lists. Celera Genomics and the public Human Genome Consortium, for example, counted in their drafts a similar *number* of potential genes, but significantly often these were not *the same* potential genes (Hogenesch *et al.* 2001).

The decisions made in annotating a sequence or a transcription event reveals how the respondent is conceptualizing genes and other genetic elements. We set out to extract some of the information implicit in these decisions by asking which factors distinguish cases that are treated as one gene with several products from those that are treated as multiple genes. This part of the survey contains graphical representations of fourteen known transcription events, two of which will be discussed in detail below.[11] There are several potential 'axes of difference' that separate the cases from each other and which might influence the decision whether to recognise one or more than one gene. The cases were chosen to allow as far as possible pairwise comparisons with respect to just one potential axis of difference at a time. We give two examples here to indicate the nature of the questionnaire materials and to anchor the ensuing discussion of the framework in which we intend to analyse this data.

Example 1. Overlapping Genes with Shared Sequences in Alternative Reading Frames

The first example involves a primary RNA transcript that is processed into two mRNA transcripts by alternative splicing, and thereby gives rise to two structurally divergent protein products (Figure 2). Both proteins play important, though different roles in cell growth. The two transcripts differ in their first coding exons (1 or 2) but share the coding sequences of the remaining exons (3 and 4). However, the presence of the different first exon (1 or 2) in the two cases results in exons 3 and 4 being read in alternative reading frames (ARF) in the two transcripts. As a consequence, there is hardly any amino acid identity between the resulting proteins.[12]

[11] For a more detailed description of 6 of the 14 cases see (Stotz and Bostanci In press). The entire questionnaire with all 14 cases is available at the Representing Genes website: http://www.pitt.edu/~kstotz/genes/genes.html

[12] See the human *INK4A/ARF* tumor suppressor region (Sharpless and DePinho 1999).

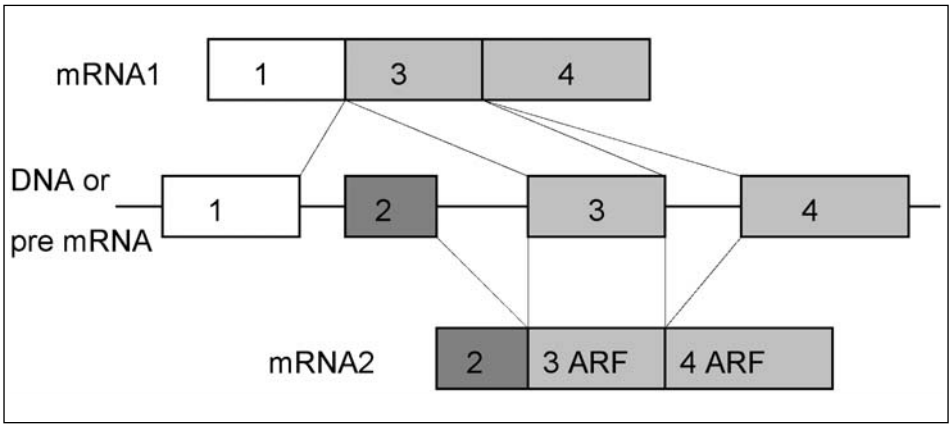

FIGURE 2. Overlapping genes with shared sequences in alternative reading frames

Biologists may label such cases of alternative splicing either as one or as several genes, and they may endorse a given annotation with different levels of confidence. The questions asked about the case were designed to capture those decisions (Table 3).

Question:				
a. Would you describe this case as one in which one or more than one gene is involved in generating the final transcript/s and/or the polypeptide/s that result from the process described?				
Clearly only ☐ one gene	Probably only ☐ one gene	Unclear ☐	Probably more than ☐ one gene	Clearly more than ☐ one gene
b. How appropriate are the following descriptions of this case?				
One gene: 1 to 4	appropriate ☐	neutral ☐	inappropriate ☐	
Two genes: 1+3+4 and 2 to 4	appropriate ☐	neutral ☐	inappropriate ☐	
Three genes: 1 to 4; 1+3+4 and 2 to 4	appropriate ☐	neutral ☐	inappropriate ☐	
Other: ……..				
c. Are there any other specific names you would use for any of the regions of the sequence in this case?				
d. If the case description does not provide you with the information you need to reply, please indicate what else you would need to know.				

TABLE 3. Follow-up questions concerning example 1.

Example 2. Overlapping Genes Without Shared Coding Sequences

Alternative splicing creates two mature mRNAs from a single pre-mRNA (Figure 3). The mature mRNAs share a noncoding exon as a common translation start site. However, the entire coding region of the first transcript is found within the first intron of the second transcript. The two transcripts consequently have no overlapping coding sequences and encode structurally unrelated proteins.[13]

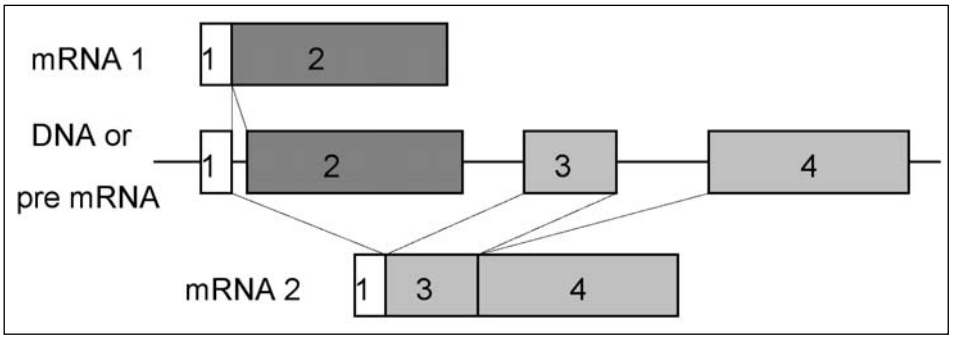

FIGURE 3. Overlapping genes without shared coding sequences

In contrast to the previous question, which involves overlapping genes with a significant amount of shared coding sequence, albeit in a different reading frame, this event highlights the possibility of overlapping genes that only share a non-coding sequence. In both cases the two products result from a single primary transcript.

We aim to identify axes of difference between cases such as the two above which seem to influence preferences in annotating the cases, either in the whole group of subjects or in subgroups with particular research interests. A first axis of difference is the *number of promoters* involved in the transcription of the DNA segments in question. Possession of a separate promoter is a standard criterion used in real-life genome annotation. A second axis of difference is whether the DNA elements in question have a known biological *function in the gene expression process*, even if they are alone not able to code for a product (e.g. non-coding region, regulatory binding site, etc). Without

[13] Example: the DNA complex *IP259/Dub80* in *D. melanogaster* (Mottus *et al.* 1997; Blumenthal 1998).

such a function DNA elements are often dismissed as 'junk DNA' or 'pseudo-genes'. Whether the DNA elements involved in a transcription event are able to *function independently* of one another is a third axis of difference. Can they code for a product without working together with the other DNA element present in this specific transcription event? This seems likely to affect whether the elements are treated as several cooperating genes or as parts of a single gene. A fourth potential axis of difference is the *relative position* or distance of these elements to each other in the genome. Are they *cis* or *trans* located,[14] are they in the same or distal chromosomal region or at least at the same chromosome, and are they transcribed in the same direction? A fifth consideration that might influence the decision between alternative splicing of one gene and overlapping sequences of two genes is the amount of shared sequence, whether the shared sequence is a coding sequence, and whether it is read in an alternative reading frame. We treat facts of this kind as contributing to a general axis of *sequence similarity*. The two cases given as examples above differ along this last axis.

Other potential axes of difference concern the *coding relationship* between DNA and product. Given the importance normally accorded to the template capacity of genes one obvious axis of difference is whether the *linear order* of nucleotides in the open reading frame is preserved in the linear sequence of amino acids in the gene product. Another point that has been the focus of much discussion of the gene concept is the *numerical relation* between DNA segments and 'gene products'. This relation may be one-to-many, many-to-one or many-to-many, as outlined in Section 2. If the ratio is either one-to-many or many-to-one another question follows: how *proximal or distal* from the DNA sequence is the initial branching point in the process of DNA expression? Last but not least one might want to know about the *functionality of the final product*. This can be hard to answer, as sometimes the result of a known transcription event is a polypeptide or RNA structure with unknown function. In this case it helps, for example, to show that the particular case of gene expression is regulated by the cell (shows a typical spatial or temporal distribution), or otherwise shows biological activity, in order to distinguish a case of functional gene expression from so-called 'transcriptional noise'.

[14] In one current usage, *cis*-elements are transcribed as part of a single unprocessed mRNA whereas *trans*-elements are transcribed separately and united at some stage of post-transcriptional processing (*trans*-splicing).

4.2 Review of Research Strategies for Complex Genetic Diseases

A second part of the questionnaire asked subjects to undertake a minimal version of another real-world task, in this case refereeing research proposals for a funding agency. This section was designed to document the use of a top-down gene concept, the first option in Falk's list of potential conceptions of the gene and one whose importance has been suggested elsewhere in the literature. We also hoped to find indications of two conceptual schemes corresponding to Moss's Gene D and Gene P.

We approached these issues by constructing four different research strategies to investigate the molecular bases of certain 'genetic' diseases. Each strategy was intended to appear compelling in the light of a particular conception of the gene. The nature of the genetic disease was recognized as another factor that might affect the assessment of research strategies. Two axes of difference were expected to influence responses: physiological versus behavioral/psychological, and human versus animal. We therefore chose to look at four types of complex genetic pathology:

- Human behavioral: Novelty Seeking
- Human psycho-physical: Frontotemporal dementia (similar to but not as well-known as Alzheimers)
- Human physiological: Malignant Hypothermia
- Animal physiological: Porcine Stress Syndrome (a.k.a. Porcine Malignant Hypothermia)

As an example we will summarize our treatment of Frontotemporal dementia (FTD). The initial question (Table 4) was followed by a short description of the genetic disease highlighting some of the scientific knowledge to date including probable genes and possible epigenetic factors affecting the onset and course of the disease (Table 5).

Question
You have been asked to referee proposals for research on the genetic basis of FTD. Assess the following proposals, assuming that all the lead investigators have equally impressive track-records, institutional backing, and so forth. Please remember, as you would when faced with a real proposal, that resources are limited.

TABLE 4. Question asked in the questionnaire to the case example FTD.

Case: Frontotemporal Dementia
Frontotemporal dementia (FTD) is a common, non-Alzheimers form of neurodegenerative disease. FTD displays considerable genetic heterogeneity. In some families it is associated with several alleles of the tau gene on chromosome 17, while in others it is associated with a less well-characterized polymorphism on chromosome 3. At least one allele of the APOE gene on chromosome 21 is a risk factor for FTD in individuals who lack both the chromosome 17 and chromosome 3 variants associated with the disease.

TABLE 5. Description of the case FTD.

This information was followed by short descriptions of four research strategies which were presented in random order. Respondents were asked to assess the potential value of pursuing each of the strategies on a five-point scale similar to that used by some real funding agencies.

The first strategy was based on the assumption that increased understanding of the genetic basis of a disease requires identifying a stable association between a specific mutation and a specific phenotype (Table 6). The existence of multiple independent correlations in complex genetic diseases threatens the viability of this approach and one strategy to recover a simpler mutation-disease relationship is to divide the original phenotype into subtypes, each corresponding to its own genotype ('splitting the trait'). This approach to the role of genes in disease would suggest that a researcher is employing a top-down conception of what genes are, or something like Moss's Gene-P conception.

Strategy 1.
Team A presumes cases correlated with variation at different loci will reveal phenotypic differences on further investigation ('splitting the trait'). The team proposes to refine the diagnostic criteria for FTD by identifying differences in age of onset, course of the disease and other symptomatic factors and to use these refined phenotypes to further investigate the genetic basis of the disease. The expectation is that these refined phenotypes will prove more genetically uniform.

TABLE 6. First research proposal (research strategy) to be assessed by the respondents of the questionnaire.

A weaker top-down or Gene-P conception might give up on the idea that a full-fledged phenotypic disease can be directly and unambiguously linked to a mutation and choose instead to bring the phenotype closer to the genotype by substituting some biochemical or cellular state (endophenotype) for an organism level phenotype (Table 7).

Strategy 2.
Team B suggests that the phenotype defined by clinical presentation or by the neuropathology associated with the full-blown disorder is unsuitable for genetic analysis. The team proposes to characterize a number of intermediate phenotypes (endophenotypes), using physiological challenges, biochemical assays and physiological measures to obtain primary indicators of disease pathology. These endophenotypes would form the basis of further investigations of the genetic basis of the disease. The expectation is that the endophenotypes will prove more genetically uniform.

TABLE 7. Second research proposal to be assessed by the respondents of the questionnaire.

The third strategy reflects the idea that genes primarily represent sources of multi-potent gene products which interact with many other such products in a biochemical pathway (Table 8). This is consistent with a Gene-D conception because the gene is primarily identified via its biochemical template capacity and, conversely, the disease phenotype is not associated directly with the presence of one or more genes, but instead with a biochemical pathway to which different genes can contribute on different occasions.

Strategy 3.
Team C proposes to locate a common pathway to which variations at different loci can contribute in different sub-populations of affected individuals. They plan to combine microarray and other gene expression data with protein annotation and new advances in the understanding of protein-protein interactions to identify the role of the several polymorphisms in the developmental pathway leading to FTD, using a series of transgenic mice. The expectation is that the disease can only be properly understood at the genetic level by understanding how multiple genes interact.

TABLE 8. Third research proposal to be assessed by the respondents of the questionnaire.

The last strategy emphasizes the potential importance of epigenetic factors that confer specific significance on a gene that otherwise has no privileged link to a specific phenotype (Table 9). This is a full-blown Gene-D conception, one that would emphasize, for example, the fact that oncogenic mutations may be found in non-cancerous cells of breast (sometime the other breast) of women with cancerous cells that have the same mutation. The strategy embodies the belief that genotype-phenotype links are held in place by a rich context of other developmental factors.

Strategy 4.
Team D proposes studying asymptomatic individuals with known mutations and symptomatic individuals with no known mutations, so as to identify epigenetic factors affecting the development of FTD, including maternal effects, environmental factors (such as disease states, head trauma in earlier life), variation in inflammatory and immune responses, and modifier genes. The expectation is that FTD can be explained as a developmental phenomenon, with known risk factors making a contribution to this developmental outcome in a suitable context.

TABLE 9. Fourth research proposal to be assessed by the respondents of the questionnaire.

4.3 Kinds of Biologists: Independent Variable

In our earlier survey we identified different groups of biologists in the simplest possible way, by asking subjects about the disciplines in which they had been educated and in which they presently worked. In our ongoing study we have chosen to study a series of potentially subtler indicators. In each case, subjects were presented with a list of options and also given the opportunity to provide 'Other' responses. We asked the following questions:

- In which scientific journals would they most like to see their work published?
- Which scientific journals do they make most effort to read?
- Which professional societies do they belong to?
- Which levels of biological organization do they see as the focus of their research?
- How would they describe their own field of research?
- Which laboratory and other techniques do they use?
- Which model organisms do they use?

– Do they see their work as part of medical science?
– Do they see their work as fundamentally comparative?

In addition to their value in analysing responses to the other parts of the survey, we anticipate that the analysis of correlations between these several indicators will provide new insights into the structure of contemporary biology, similar to those obtained in the new discipline of 'knowledge domain mapping' (Shiffrin and Börner 2004).

5. Conclusion

Philosophical, historical and experimental research on conceptualizations of the gene and of other DNA elements, and on related ideas about heredity and development are important because these concepts play roles both in scientific discourse and in a much larger set of overlapping discourses in bioethics and public policy, in popular science and, ultimately, in contemporary understanding of what it is to be human. One might thus expect that different representations of the same or similar genomic elements, as a consequence of different conceptualizations of the elements and their action, may result in significantly different understandings of these biological processes or 'genes' on the part of wider audiences. The work presented in this paper may therefore lead not only to a better understanding of how various gene concepts contribute to biological research, but also to a better understanding of how they figure in the dissemination of genomic knowledge to other audiences. The study of conceptual change and conceptual diversity in genomics is thus relevant to the work of bioethicists, medical sociologists, and science communicators, as well as to philosophers and historians of biology

Acknowledgments

We are grateful to Richard M. Burian for comments on an earlier draft of this paper.

This research was supported by ARC Large Grant A-59906145 and NSF grant #0217567 and by the University of Pittsburgh.

References

Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson J.D., 2002, *The Molecular Biology of the Cell*, (4 ed.), New York / London: Garland.

Blumenthal T., 1998, 'Gene Clusters and Polycistronic Transcription in Eukaryotes', *BioEssays*, 20 (6):480-487.

Caudevilla C., Serra D., Miliar A., Codony C., Asins G., Bach M., Hegardt F.G., 1998, 'Natural Trans-Splicing in Carnitine Octanoyltransferase Pre-mRNAs in Rat Liver', *Proceedings of the National Academy of Sciences of the United States of America*, 95 (21):12185-12190.

Coelho P.S.R., Bryan A.C., Kumar A., Shadel G.S., SnyderM., 2002, 'A Novel Mitochondrial Protein, Tar1p, Is Encoded on the Antisense Strand of the Nuclear 25S rDNA', *Genes and Development*, 16:2755-2760.

Communi D., Suarez-Huerta N., Dussossoy D., Savi P., Boeynaems J.M., 2001, 'Cotranscription and Intergenic Splicing of Human P2Y(11) SSF1 Genes', *Journal of Biological Chemistry*, 276 (19):16561-16566.

Falk R., 1986, 'What Is a Gene?' *Studies in the History and Philosophy of Science*, 17:133-173.

Falk R., 2000, 'The Gene: A Concept in Tension'. In: Beurton P., Falk R., Rheinberger H.-J. (eds), *The Concept of the Gene in Development and Evolution*, Cambridge: Cambridge University Press, 317-348.

Falk R., 2001, 'Can the Norm of Reaction Save the Gene Concept?'. In: Singh R., Krimbas C., Paul D.B. and Beatty J. (eds), *Thinking About Evolution: Historical, Philosophical and Political Perspectives*, New York: Cambridge University Press, 317-343.

Falk R., 2003, 'How Many Chromosomes?', *Biology and Philosophy,* 18 (4): 619.

Finta C., Zaphiropoulos P.G., 2000, 'The Human CYP2C Locus: A Prototype for Intergenic and Exon Repetition Splicing Events', *Genomics*, 63 (3):433-438.

Finta C., Zaphiropoulos P.G., 2000b, 'The Human Cytochrome P450 3A Locus. Gene Evolution by Capture of Downstream Exons', *Gene*, 260 (1-2):13-23.

Flouriot G., Brand H., Seraphin B., Gannon F., 2002, 'Natural Trans-Spliced mRNAs Are Generated from the Human Estrogen Receptor-Alpha (hER Alpha) Gene', *Journal of Biological Chemistry*, 277 (29):26244-26251.

Fogle T., 2000, 'The Dissolution of Protein Coding Genes in Molecular Biology'. In: Beurton P., Falk R. and Rheinberger H.-J. (eds), *The Concept of the Gene in Development and Evolution*, Cambridge: Cambridge University Press, 3-39.

Griesemer J.R., 2000, 'Reproduction and the Reduction of Genetics'. In: Beurton P., Falk R. and Rheinberger H-J. (eds), *The Concept of the Gene in Development and Evolution*, Cambridge: Cambridge University Press, 240-285.

Griffiths P.E. and Neumann-Held E.M., 1999, 'The Many Faces of the Gene', *BioScience*, 49 (8):656-662.

Hinde R.A., 1985, 'Was "The Expression of Emotions" a Misleading Phrase?', *Animal Behaviour*, (33):985-992.

Hogenesch J.B., Ching K.A., Batalov S., Su A.I., Walker J.R., Zhou Y., Kay S.A., Schultz P.G., Cooke M.P., 2001, 'A Comparison of the Celera and Ensembl Predicted Gene Sets Reveals Little Overlap in Novel Genes', *Cell*, 106 (4):413-415.

Johannsen W., 1909, *Elemente der exakten Erblichkeitslehre*, Jena: Gustav Fischer.

Keller E.F., 2000, *The Century of the Gene*, Cambridge, MA: MIT Press.

Kitcher P., 1984, '1953 and All That: A Tale of Two Sciences ', *Philosophical Review*, 93:335-373.

Knight R.D., Griffiths P.E., 2001, *Selfish Genes: The Eunuchs of Selection* [Online preprint], Australasian Association for the History, Philosophy and Social Studies

of Science Preprint Series, 03/05/01 1999 [cited 06/09/01 2001]. Available from http://www.usyd.edu.au/hps/aahpsss/preprints.html.

Liu X.Q., 2000, 'Protein-Splicing Intein: Genetic Mobility, Origin, and Evolution', *Annual Review of Genetics*, 34:61-76.

Magrangeas F., Pitiot G., Dubois S., Bragado-Nilsson E., Cherel M., Jobert S., Lebeau B., Boisteau O., Lethe B., Mallet J., Jacques Y., Minvielle S., 1998, 'Cotranscription and Intergenic Splicing of Human Galactose-1-Phosphate Uridylyltransferase and Interleukin-11 Receptor Alpha-Chain Genes Generate a Fusion mRNA in Normal Cells - Implication for the Production of Multidomain Proteins During Evolution', *Journal of Biological Chemistry*, 273 (26):16005-16010.

Moss L., 2002, *What Genes Can't Do*, Cambridge, MA: MIT Press.

Mottus R.C., Whitehead I.P., Ogrady M., Sobel R.E., Burr R.H.L., Spiegelman G.B., Grigliatti T.A., 1997, 'Unique Gene Organization: Alternative Splicing in Drosophila Produces Two Structurally Unrelated Proteins', *Gene*, 198 (1-2):229-236.

Neumann-Held E.M., 1999, 'The Gene Is Dead - Long Live the Gene: Conceptualising the Gene the Constructionist Way'. In: Koslowski P. (ed.), *Sociobiology and Bioeconomics. The Theory of Evolution in Biological and Economic Theory*, Berlin: Springer-Verlag, 105-137.

Pirrotta V., 2002, 'Trans-Splicing in Drosophila', *BioEssays*, 24 (11):988-991.

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

Sharpless N.E., DePinho R.A., 1999, 'The INK4A/ARF Locus and Its Two Gene Products', *Current Opinion in Genetics & Development*, 9:22-30.

Shiffrin R., Börner K., 2004, 'Mapping Knowledge Domains', *PNAS* 101 (Suppl. 1):5183-5185.

Stotz K., Bostanci A., Griffiths P.E., (in press), 'Representing Genes Project: Tracking the Shift to Post-Genomics', *New Genetics in Society*.

Stotz K., Griffiths P.E., Knight R.D., 2004, 'How Scientists Conceptualize Genes: An Empirical Study', *Studies in History & Philosophy of Biological and Biomedical Sciences*.

Takahara T., Kasahara D., Mori D.,Yanagisawa S., Akanuma H., 2002, 'The Trans-spliced Variants of Sp1 mRNA in Rat', *Biochemical and Biophysical Research Communications*, 298 (1):156-162.

Waters C.K., 1994, 'Genes Made Molecular', *Philosophy of Science*, 61:163-185.

Waters C.K., 2000, 'Molecules Made Biological', *Revue internationale de philosophie*, 4 (214):539- 564.