

Tracking the Shift to ‘Postgenomics’

Running Head: Tracking the Shift

Karola C. Stotz

History and Philosophy of Science

1017 Cathedral of Learning

University of Pittsburgh

Pittsburgh, PA 15260

+1 (412) 624-9177

+1 (412) 624 6825 (Fax)

kstotz@pitt.edu

Adam Bostanci

ESRC Centre for Genomics in Society

University of Exeter

United Kingdom

+44 (1392) 262 054

+44 (1392) 263305 (Fax)

a.w.s.bostanci@ex.ac.uk

Paul E. Griffiths

Biohumanities Project

School of History, Philosophy, Religion and Classics

University of Queensland,

Queensland 4072, Australia

+61 (07) 3365-2646

+61 (07) 3365-1968 (Fax)

paul.griffiths@uq.edu.au

Keywords: gene, genomics, conceptual change, paradigm shift, gene expression

Abstract

Current knowledge about the variety and complexity of the processes that allow regulated gene expression in living organisms calls for a new understanding of genes. A ‘postgenomic’ understanding of genes as entities constituted during *genome* expression is outlined and illustrated with specific examples that formed part of a survey research instrument developed by two of the authors for an ongoing empirical study of conceptual change in contemporary biology.

1. Introduction

Philosophers and historians of biology have emphasized that 'the gene', though defined differently in different sub-disciplines, has nevertheless been the unifying theme of 20th century biology as well as a powerful cultural icon [1-3]. Given this prominence of 'the gene' in contemporary science and society, several commentators have begun to assess the implications of the on-going molecular genomics revolution with respect to gene concepts [4, 5]. Two of the present authors have been involved for some years in an attempt to put this work on a more empirical footing using survey research of working biologists. The enterprise of trying to operationalise contemporary biological understandings of the gene has led us to form the view that what is now known about the variety and complexity of the processes that allow regulated gene expression in living organisms calls for a new understanding of genes. This ‘postgenomic’ understanding of genes as entities constituted during *genome* expression will be developed and illustrated throughout this paper; some of the broader implications of this conceptual shift will be considered in the conclusion.

The survey instrument developed for the Representing Genes Project (<http://www.pitt.edu/~kstotz/genes/genes.html>) contains diagrams and descriptions of real genome transcription events that challenge various aspects of the 'classical molecular gene concept' [6]. Biologists completing the online survey are asked to assess whether the events in question involve one or several genes. However, since common definitions of the gene are insufficient for making this decision, the survey is designed to reveal the implicit criteria biologists draw upon when judging whether a DNA structure corresponds to one or several genes. The survey builds on previous research by two of the present authors, which has established that it is possible to operationalise questions about conceptual variation in a questionnaire format and that claims about the prevalence of particular gene concepts in different biological disciplines can be tested by statistical analysis of this questionnaire data [7]. The study currently under way aims to extend that earlier work [8]. Once the data are analyzed the investigators hope that the project will reveal which disciplines are leading the field in modernizing concepts-in-use by broadening current models of genomic elements that constitute 'genes', and which retain the traditional concept of a stable, particulate gene, a concept that seems unlikely to be able to accommodate the genomic complexities that are being discovered on a daily basis. It is also hoped that the results will be of interest to those concerned with the broader significance of current research in human genomics, such as bioethicists, medical sociologists, science communicators, and philosophers of biology.

2. From Genetics to Genomics – and beyond

Even before the announcement in February 2001 of the first drafts of the human genome sequence, the principles and technologies that had enabled this impressive achievement were being turned to the study of other areas of cell biology. The burgeoning fields of genomics and proteomics are spawning multiple ‘-omic’ subdisciplines and related areas. This suffix generally refers to the study of a complete system of biomolecules, such as a genome, containing all of an organism's ‘genes’, such as its proteome, containing all of its proteins, its transcriptome, containing all coding RNA transcripts, its rnome, containing all of its - increasingly acknowledged - non-coding but functional RNA transcripts, and its metabolome, containing all of its metabolites. Phenomics finally attempts to integrate these areas of study into a holistic picture of the complete organism – its phenotype.

Although these disciplinary shifts are driven by the availability of improved technology, their consequences are not merely technological. In fact, these new fields produce new kinds of data that undermine the classical molecular gene concept. For example, genome sequencing projects produce an avalanche of DNA sequence data that calls for 'annotation' and thus confronts working biologists with many of the conceptual problems simulated in the Representing Genes survey instrument. Improved analytic methods also make it easier to study gene expression patterns in all their molecular complexity. The genomics revolution means that biologists are presented with these complexities – some of which have in principle been known for a long time – on a daily basis. As a more appropriate term and to allow for a

generalized understanding of biological events such as those represented in the survey, two biologists have recently also coined the term “genome transcription” [9].

The classical molecular gene was defined *functionally* as a segment of DNA that codes for a single protein and *structurally* as an open reading frame (ORF) that specifies the amino acid sequence of that protein. The basic stages in the formation of a protein are called 'transcription' and 'translation'. During transcription, one DNA chain functions as a template for the production of a specific messenger-RNA (mRNA). The mRNA molecule then serves as a template for the synthesis of a polypeptide. During translation triplets of mRNA sequence known as 'codons' specify which amino acid is added to the growing polypeptide chain and, in the case of start and stop codons, where translation is initiated and terminated. A stereotypical ORF comprises a start codon, a stop codon, and the coding nucleic acid sequence in between. In the stereotypical case, these processes ensure that there is a linear correspondence between the original DNA sequence of the ORF, the sequence of the mRNA, and the amino acid sequence of the resulting protein. The functional and structural aspects of the definition of the classical molecular gene are therefore generally assumed to be consistent with one another [10].

However, many additional processes complicate the basic stages of gene expression to the extent that the classical molecular gene concept has become increasingly untenable. In eukaryotes (fungi, plants and animals) a primary or pre-messenger RNA may be transcribed from the DNA sequence, from which the final RNA transcript is processed by cutting out non-coding sequences, the *introns*, and splicing together of

the remaining coding sequences, the *exons*. When more than one mature mRNA transcript results from these processes of cutting and joining alternative exons, biologists speak of alternative *cis*-splicing¹. Splicing can also occur between adjacent genes that are sometimes cotranscribed, that is, transcribed together to produce a *single* pre-mRNA that is then spliced [11-13], or between a gene and an adjacent 'pseudo-gene' that is normally incapable of producing a product [11, 14]. 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) which 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 [15, 16]. The similarity of the products derived from these overlapping genes depends on the extent of the overlap, and on whether these shared sequences are read in the same reading frame. Which codons a DNA sequence contains depends on the precise nucleotide at which reading begins. Starting at a different nucleotide is called 'frameshift', and in an English sentence the phenomenon of frameshift would look like this: 'A gene is a flexible entity' becomes 'Age nei saf lex ibl een tit y'. Unlike any human language, however, whatever frame a DNA sequence is read in it will always be made up of meaningful 'three-letter-words' (codons that specify an amino acid during translation).

In the process of *trans*-splicing a final mRNA transcript is processed from two or more independently transcribed pre-mRNAs. The prefix *trans* suggests that these pre-mRNAs are from DNA sequences far apart from one other, but this is not always the

¹ In 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). See below for *trans*-splicing phenomena.

case. In fact, two copies of the very same sequence can be spliced together this way, as can alternative exons of what would normally be regarded as a 'normal' gene [16, 17]. The main reason for assuming *trans*-splicing, even when the exons are on the same strand of DNA, is that all alternative exons feature their own promoters [18, 19]. Moreover, it was once thought that only one strand of DNA is read, but since then unique anti-sense transcripts have been found. In other words, DNA can be read both forwards and backwards by the cellular machinery, producing different products in each case [20]. RNA editing is another mechanism of modification of the original transcribed sequence, which 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 four known mechanisms: the insertion or deletion of U or the insertion of G nucleotides, the substitution of C to U or the substitution of A to I nucleotides. Although we will not describe them here, other processes may occur before the final mRNA transcript is translated into a protein sequence. The relationship between DNA and protein is indirect and mediated to an extent that was never anticipated when the basic mechanisms of transcription and translation were clarified in the 1960s.

The conclusion that must be drawn from these newly discovered complexities is that the classical molecular gene concept does not succeed in keeping structure and function tightly united. We find a gap between genetic 'information' and its biological 'meaning', because local DNA sequences only contain partial information about the functional product or products that will be derived from them. Hence, the classical

molecular definition of the gene leaves open many decisions about the boundaries of genes that have to be made when annotating genome sequences.

In recognition of the difficulties of the classical molecular gene concept, it has been suggested that working biologists employ a kind of *consensus gene* concept that is a stereotype combining features from a number of exemplary cases. The consensus gene concept is based on a collection of flexibly applied features of well-established genes. One might say that a stretch of DNA is considered to be a gene, if it has 'enough' of these features, e.g. it contains an open reading frame, a TATA box,² and is transcribed into an RNA molecule. The originator of this view, Thomas Fogle, has argued that by combining structural and functional features into a single stereotype, the consensus gene concept hides both the diversity of DNA sequences that can perform the same function and the diverse functions of particular DNA sequences. In other words, the consensus gene concept inherently distracts from conceptually problematic cases [21]. As with stereotypes more generally, even when biologists have been exposed to cases that violate the rule, they tend to revert to the stereotype in future work. Complex genetic elements are consequently often presented as spectacular but isolated discoveries, rather than as grounds for questioning received views of the gene.

3. Genome annotation

The survey instrument of the Representing Genes project was designed to explore several of the issues posed by the existence of alternative gene concepts. This paper

² The TATA box is a sequence found in the promoter region of most genes transcribed by eukaryotic RNA polymerase II and has the consensus sequence: 5'-TATAAAA-3'.

focuses on only one of these, namely, the criteria which lead biologists to annotate a particular DNA sequence as either *one gene with several gene products* or as *several genes*, or conversely, to annotate a particular DNA sequence as either a *single gene* or as *several genes which contribute to a single functional product*. The survey instrument contains graphical representations of fourteen real biological cases, two of which will be discussed below. All cases were chosen to allow pair-wise comparisons highlighting differences that may influence the judgments on this issue during annotation. The examples are chosen to illustrate the flexibility and variability of 'genome expression', and the difficulty of pinning down what genes are. Below we list several *axes of difference* between cases of genome expression that might influence whether working biologists annotate complex genomic elements as involving one or more than one gene.

One potential axis of difference that may influence judgments during annotation is the *number of promoters* involved in the transcription of a chromosomal region in question, for the presence of a promoter is a standard criterion of 'genehood'. Another criterion may be whether the DNA elements in question have a known *function in gene expression*, even if they do not code for a product by themselves. For example, DNA elements without function in gene expression are often labeled as 'junk DNA', or as 'pseudo-genes' even when they are structurally very similar to known genes. Whether the DNA elements involved in a transcription event are able to *function independently* from one another may be a third axis of difference that influences whether such elements are treated as cooperating genes or as parts of a single gene. Biologists may also put some weight on the *relative position* of genetic elements (Are they in the same chromosomal region or on the same chromosome? Are they

transcribed in the same direction? Are they *cis* or *trans* located?). Whether a complex genetic element is annotated as either a case of *alternative splicing of a single gene* or as a case of *overlapping genes* may also be due to the amount of shared DNA sequence, whether the shared DNA sequence is coding DNA, and whether it is read in an alternative reading frame. Last but not least, biologists annotating a genetic structure might want to know whether the final product has an actual function in the cell. This question can be hard to answer and biologists might seek further evidence before annotating the underlying DNA sequence.

Some of the above axes of difference can be broken down into criteria that further distinguish gene expression cases from each other and may influence annotation decisions. For example, a host of cellular processes subvert the idealized one-to-one relationship between DNA sequence and product into a one-to-many or many-to-one relationship. Moreover, these processes may intervene at different stages of gene expression, and the specific stage at which these "branching points" occur may influence decisions during annotation. For example, at the DNA level alternative promoters may define alternative transcripts. At the pre-mRNA level, splicing mechanisms may change both the number and nature of *exons* and *introns* as well as the number and nature of the products derived. At the mRNA level, mRNA editing, and other processes may once again give rise to several products from a single mRNA. Finally at the protein level, so-called 'inteins' may be spliced out of proto-protein to leave only the 'exeins', or protein *trans*-splicing may splice two separate polypeptide strands into one single protein product.

Analysis of responses to the cases in the survey according to the schema laid out above should suggest which of these considerations lead members of particular groups of biologist to annotate a genomic structure in one way rather than another. For, as Oliver and Leblanc put it, annotation always involves an "executive decision about the relevancy, accuracy, and quality of the evidence, and by definition exposes the curator's point of view" [22]. Practitioners thus stress that annotation is an open-ended process that depends on future evidence and subjective judgments. However, the argument underlying the design of this survey instrument goes further. No amount of evidence can settle whether the cases graphically represented below *should* be annotated as one or several genes, because the classical molecular gene concept and other common gene concepts are inherently insufficient for making this decision. Annotation will, so we hope, expose particular subjects' conceptions of the gene.

4. Complex cases of genome expression

In this section we describe two of the cases that featured in the Representing Genes survey instrument.

1. Overlapping Genes with Shared Sequences in Alternative Reading Frames

The first example from our survey 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. Both proteins play important, though different roles in cell growth. The two transcripts differ in their first coding exons (exons 1 or 2) but share the coding sequences of exon 3 and 4. However, the presence of exon 1 or exon 2 respectively, results in exons 3 and 4 being read in alternative

reading frames (ARF) in the two transcripts. Consequently, there is hardly any amino acid identity between the resulting proteins [23]

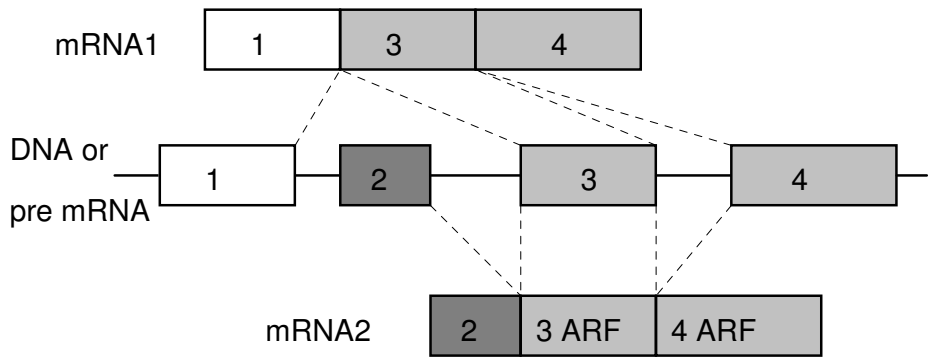


Figure 1. An abstract representation of transcription of the human *INK4A/ARF* tumor suppressor region.

Biologists may label such cases of alternative splicing as either one or as several genes, and they may endorse a given annotation with different levels of confidence. The following questions, which accompany each of the fourteen cases included in the survey instrument, were designed to explore these preferences.

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 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 1. Questions about the transcription events illustrated in Figure 1 (above) from the Representing Genes research instrument.

2. Mitochondria *Trans*-Splicing and RNA editing

Subunit 1 of the respiratory chain NADH dehydrogenase is encoded by the gene *nad1*, which in the mitochondrial genomes of flowering plants is fragmented into five coding segments that are scattered over at least 40kb of DNA sequence and interspersed with other unrelated coding sequences. In wheat, for example, 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 [24]. 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.

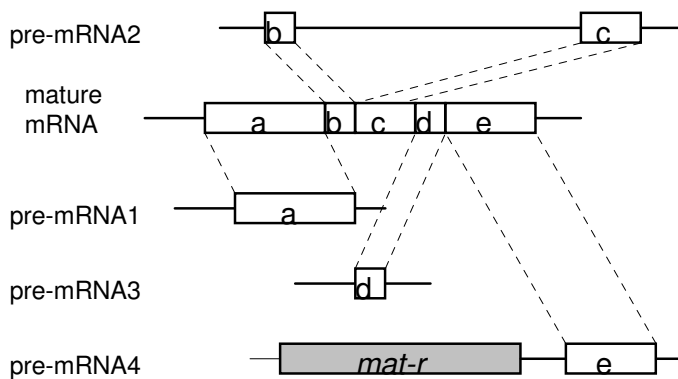


Figure 2. An abstract representation the cis- and trans-splicing events in the processing of the mature *nad1* mRNA in Wheat. A maturase related ORF (*mat-r*) is encoded in the intron upstream of the exon e (in some other plant species *mat-r* is encoded by an intron between the exons d and e in a single pre mRNA).

Conclusion: From the Active Gene to the Reactive Genome

The biological processes represented in these and many similar examples suggest that the common perception that genes are well-defined chromosomal segments embedded within non-functional sequence ('junk DNA') is overly simplistic. The fact that exons of what might otherwise be regarded as a single 'gene' can undergo alternative splicing, that different 'genes' can be co-transcribed by way of bypassing transcriptional terminations, as well as the use of alternative promoters, intra- and intergenic *trans*-splicing of otherwise unrelated pre-mRNA molecules, and even rare cases of 'split' genes translated into their functional product by protein *trans*-splicing – all these phenomena imply that the term 'gene' may no longer usefully refer to a stable, structural unit of genomic information. Moreover, recent findings suggest that the formation of 'chimeric' mRNAs and *trans*-splicing of pre-mRNA molecules may represent additional general splicing pathways that further deepen the splicing capacity and complexity of the organism [9]. In fact, it has been suggested that the most meaningful measure of an organism's complexity is given by its repertoire of unique transcripts – by how much it can do with its genome rather than by how many genes it possesses.

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 of 'genome transcription' to allow for the incorporation of RNA transcripts that contain sequences outside the border of

canonical genes. This view does not sit easily with a more traditional idea of gene transcription events, which from the new perspective seem like “statistical peaks within a wider pattern of genome expression” [9]. Against the background of the recent biotechnological ‘gene frenzy’ and the popular concept of the gene as the incarnation of pure and ultimate nature on which “nurture can never stick” [25: 318], we can now place a picture that delivers itself the arguments against genetic determinism: the view of the whole organism as the center of agency with *genes as things the organisms can do with its genome* [26]. According to this ‘post-genomic’ understanding, ‘genes’ are as much acted upon as they are actors. What a ‘gene’ is and what it does depends on the cellular environment, on interactions with other genomic elements, gene products and other factors present in the cell. Even at the level of the ‘gene’, we may find that wholes determine parts as well as parts determining wholes [27]. The diverse mechanisms of *trans*-splicing allow for the cooperation of sequences that are far apart in the genome. Several molecular biologists have theorized such functional units or ‘genes’ as akin to a “dynamic information processor” [28, 29].

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. Drawing on the contrast between the more traditional, particulate understandings of the gene and the view of genes as constituted through genome expression, one might thus expect that presentations of the same genomic elements from these two very different

perspectives would result in significantly different understandings of 'genes' on the part of wider audiences.

Acknowledgment

The work discussed in this paper was supported by National Science Foundation, grant #0217567, with supplemental funding from the University of Pittsburgh.

References

1. Rheinberger H-J: Gene Concepts: Fragments from the Perspective of Molecular biology; in Beurton PJ, Falk R, Rheinberger H-J (eds): *The Concept of the Gene in Development and Evolution*. Cambridge, Cambridge University Press, 2000, pp 219-39.
2. Nelkin D, Lindee MS: *The DNA Mystique: The gene as a cultural icon*. New York, Freeman, 1995.
3. Keller EF: *The Century of the Gene*. Cambridge, Mass., MIT Press, 2000.
4. Moss L: *What Genes Can't Do*. Cambridge, Mass., MIT Press, 2002.
5. Beurton P, Falk R, Rheinberger H-J: *The Concept of the Gene in Development and Evolution*. Cambridge, Cambridge University Press, 2000.
6. Neumann-Held EM: 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, 1998, pp 105-37.
7. Stotz K, Griffiths PE, Knight R: How scientists conceptualise genes: An empirical study. *Studies in History & Philosophy of Biological and Biomedical Sciences* 2004;35(4):647-73.
8. Stotz K, Griffiths PE: Genes: Philosophical analyses put to the test. *History and Philosophy of the Life Sciences* 2004 (December).
9. Finta C, Zaphiropoulos PG: A statistical view of genome transcription. *Journal of Molecular Evolution* 2001;53:160-92.
10. Griffiths PE, Neumann-Held EM: The many faces of the gene. *BioScience* 1999;49(8):656-62.
11. Finta C, Zaphiropoulos PG: The human cytochrome P450 3A locus. Gene evolution by capture of downstream exons. *Gene* 2000;260(1-2):13-23.
12. Communi D, Suarez-Huerta N, Dussosoy D, Savi P, Boeynaems JM. Cotranscription and intergenic splicing of human P2Y(11) SSF1 genes. *Journal of Biological Chemistry* 2001;276(19):16561-6.
13. Magrangeas F, Pitiot G, Dubois S, et al.: 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* 1998;273(26):16005-10.

14. Finta C, Zaphiropoulos PG: Intergenic mRNA molecules resulting from trans-splicing. *Journal of Biological Chemistry* 2002;277(8):5882-90.
15. Mottus RC, Whitehead IP, Ogrady M, et al.: Unique gene organization: alternative splicing in *Drosophila* produces two structurally unrelated proteins. *Gene* 1997;198(1-2):229-36.
16. Blumenthal T: Gene clusters and polycistronic transcription in eukaryotes. *BioEssays* 1998;20(6):480-7.
17. Mottus RC, Whitehead IP, O'Grady M, et al.: Unique gene organization: alternative splicing in *Drosophila* produces two structurally unrelated proteins. *Gene* 1997;198:229-36.
18. Dorn R, Krauss V: The modifier of *mdg4* locus in *Drosophila*: functional complexity is resolved by trans splicing. *Genetica* 2003;117(2):165-77.
19. Pirrotta V: Trans-splicing in *Drosophila*. *Bioessays* 2002;24(11):988-91.
20. Coelho PSR, Bryan AC, Kumar A, Shadel GS, Snyder M: A novel mitochondrial protein, *Tar1p*, is encoded on the antisense strand of the nuclear 25S rDNA. *Genes and Development* 2002;16:2755 -60.
21. Fogle T: The Dissolution of Protein Coding Genes in *Molecular Biology*; in Beurton P, Falk R, Rheinberger H-J (eds): *The Concept of the Gene in Development and Evolution*. Cambridge, Cambridge University Press, 2001, pp 3-25.
22. Oliver B, Leblanc B: How many genes in a genome? *Genome Biology* 2003;5(1):204:1-3.
23. Sharpless NE, DePinho RA: The *INK4A/ARF* locus and tis two gene products. *Current Opinion in Genetics & Development* 1999;9:22-30.
24. Chapdelaine Y, Bonen L: The Wheat Mitochondrial Gene for Subunit I of the NADH Dehydrogenase Complex: A Trans-splicing Model for This Gene-in-Pieces. *Cell* 1991;65(3):465-72.
25. Falk R: 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, 2000:317-48.
26. Stotz K: With genes like that, who needs an environment? *Genomics'* argument against genetic determinism. *PSA Proceedings* (in preparation)
27. Gilbert SF, Sarkar S: Embracing complexity: Organicism for the Twenty-first Century. *Developmental Dynamics* 2000;219:1-9.
28. Dillon, N: Positions, please... *Nature* 2003;425:457.
29. Mattick, JS: Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* 2003;25(10):930-939.

