What Does It Mean to Be 75% Pumpkin? The Units of Comparative Genomics

Monika Piotrowska[†][‡]

Comparative genomicists seem to be convinced that the unit of measurement employed in their studies is a gene that drives the function of cells and ultimately organisms. As a result, they have come to some substantive conclusions about how similar humans are to other organisms based on the percentage of genetic makeup they share. I argue that the actual unit of measurement employed in the studies corresponds to a structural rather than a functional gene concept, thus rendering many of the implications drawn from comparative genomic studies largely unwarranted, if not completely mistaken.

1. Introduction. In 1999 the genome of a pumpkin was sequenced, and a BBC science reporter confidently announced that "75% of our genetic make-up is the same as a pumpkin" (Durrani 1999). Six years later, the genome of a chimpanzee was sequenced, and the headlines reported that humans and chimpanzees are 98% similar. This announcement has since been used as an additional argument in the ongoing debate regarding chimpanzee research restrictions, the ethical concern being that if chimpanzees are so similar to humans, then perhaps the kind of research that is conducted on them ought to be restricted to the kind that is conducted on us. The pumpkin and chimpanzee genomes comprise merely two of the 472 genomes that have been sequenced as of January 25, 2007 (Sivanshankari and Shanmughavel 2007, 376). Only a select few genomes

[†] To contact the author, please write to: Department of Philosophy, University of Utah, 215 South Central Campus Drive, Carolyn Tanner Irish Humanities Building, 4th Floor, Salt Lake City, UT 84112; e-mail: monika.piotrowska@utah.edu.

[‡] Special thanks to Steve Downes for valuable comments, criticism, and encouragement. I am also grateful to Bryan Benham, Matt Haber, Anya Plutynski, and Jim Tabery for helpful feedback, and most of all to Matt Mosdell, whose healthy skepticism of statistics inspired me to write this article. Earlier versions of this work were presented at ISHPSSB 2007 in Exeter and PSA 2008 in Pittsburgh. Conversations with various audience members helped clarify my ideas on the topic.

Philosophy of Science, 76 (December 2009) pp. 838–850. 0031-8248/2009/7605-0026$10.00

make the headlines upon their completion (e.g., mouse, rat, dog, opossum, etc.), but each time an article is written on the topic, a claim is made about the percentage of genes humans share with the organism of interest. Of course, the media are famous for skewing scientific results to attract more attention, but in this case, similar claims are being made in the scientific literature (cf. Rubin et al. 2000; Waterston et al. 2002; Patis 2007).

Undoubtedly, a number of questions come to mind when one is confronted with claims of genetic similarity between humans and other organisms, which is why the first goal of this article is to identify who is making these claims, how they are coming up with their statistics, and to what end such claims are being used in the scientific community. The second goal is to argue that comparative genomic studies do not successfully compare genetic makeup, because what they in fact measure does not match what they claim to be measuring.

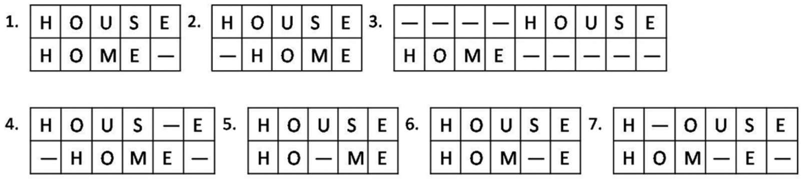
The structure of the article is as follows: in the next section, I will introduce the field of comparative genomics and explain why genetic comparisons are conducted at the very first level of gene expression. In Sections 3 and 4, I will discuss the methods used to compare genes and supply three different motives for using genetic similarity claims in scientific research. In Section 5, I will begin my critical analysis by arguing that genetic similarity claims are not entitled to the substantive implications attributed to them because the structural gene unit employed in comparative genomic studies does not map onto the functional gene unit that comparative genomicists claim to be using (cf. Keller 2000). I will conclude with a summary of my argument.

2. Who Is in the Business of Comparing Genes? The researchers supplying the curious claims reported by the media work in a burgeoning field known as *comparative genomics*, which, as the name suggests, is a field dedicated to genetic comparisons between different species. The aim of comparative genomics is "to decipher how genes function and provide an understanding of the link between genotype and phenotype" (Clark 2000, 1). Although the way in which genes function is still largely unknown, what most comparative genomicists take for granted is that "DNA makes RNA, RNA makes protein, and proteins make us" (Keller 2000, 54)—the last part is where things get fuzzy. That being said, most biologists agree that the pathway of expression most likely moves through a hierarchy of interactions; via metabolism, the pathway reaches cells and tissues until it finally arrives at the level of the organism (Auffray et al. 2003, 1130). Comparative genomic studies focus their comparisons at the very first level of gene expression—the DNA level, composed of strings of base pairs: adenine (A), thymine (T), guanine (G), and cytosine (C)—which may not seem like an ideal place to start if the aim is to say something meaningful about the way in which genes *function*. However, while the transcriptome or the proteome levels appear to be better candidates for a functional comparison, there are some drawbacks to comparing these so-called expressed genes.[1] For example, the fact that different cells show different patterns of gene expression makes it difficult to compare expressed genes between species because it is never the case that all the genes of an organism are expressed in one cell at a particular time. In addition, RNA transcripts are called into being only as needed, generally have short lifetimes, and reside outside of the chromosome. In some cases, they "might not even be found in the nucleus—that is, the final version of the transcript may be put together only after the original transcript has entered the cytoplast" (Keller 2000, 63–64). To be fair, comparing genes at the DNA level has its drawbacks as well (hence the discussion in this article), but there seems to be a certain appeal to comparing the raw code at its most fundamental level (Pearson 2006).

3. How Are Genes Compared? Various methods may be used to compare DNA sequences, but, whenever the point of the comparison is to locate gene counterparts between species, at least one genome must first be scanned for the presence of genes in the sequence. Both the enormous amount of information stored in a genome and the difficulty of empirically verifying the functionality of located genes have made it necessary to use computational methods for locating genes. Consequently, there has been an increase in the number of computer scientists playing the role of bioinformaticians to assist molecular biologists in their research (Wong 2004). The job of the bioinformatician begins with the raw sequence data—a long string of As, Cs, Ts, and Gs—and ends with a secondary structure prediction (e.g., locating genes in the sequence). The goal is to move away from the raw, unprocessed information and move closer to a complete understanding of how the sequence acts as a functional unit in the organism.

Bioinformaticians rely on gene-finding algorithms to locate genes in a DNA sequence. These gene-finding algorithms are typically based on complex probabilistic models (e.g., the hidden Marvok Model; Karplus and Sjolander 1997), whose task is to anticipate the unknown parameters from observable ones. Observable parameters are usually extracted from a training set, that is, a set of genes that have been studied experimentally and are thereby well annotated (Gregory 2005, 525–536). However, in most organisms whose genomes have been sequenced, only a small number of genes have been studied experimentally. Consequently, the size of the training sets available for these species is fairly small. For example, the

[1] Analogous with 'genome', the 'transcriptome' is the complete set of transcripts expressed by a cell, and the 'proteome' is the complete set of proteins.



1.

H	O	U	S	E
H	O	M	E	—

2.

H	O	U	S	E
—	H	O	M	E

3.

—	—	—	—	H	O	U	S	E
H	O	M	E	—	—	—	—	—

4.

H	O	U	S	—	E
—	H	O	M	E	—

5.

H	O	U	S	E
H	O	—	M	E

6.

H	O	U	S	E
H	O	M	—	E

7.

H	—	O	U	S	E
H	O	M	—	E	—

Figure 1. Redrawn from Dwyer 2002, 32–33.

algorithm used in DOUBLESCAN—a program that predicts genes in mouse and human DNA—was trained on a set composed of only 36 genes (Miller et al. 2004).

Once the unknown parameters are extracted from the observable ones, an algorithm can provide a hypothesis on where the remaining genes in a given sequence are located. The nucleotides identified as "genes" are then entered into a database, and the genome of another organism may be scanned for positive matches. This next step is known as *comparative gene prediction* because it involves predicting the location of "matching" genes in the genome of another organism by identifying sequences that look similar to the sequences of the genes stored in the database. Possible matches are scored based on how well they align with the database genes. A sequence alignment typically involves a mapping of the nucleotides in one sequence onto the nucleotides in some other sequence, with gaps introduced into one or the other sequence to increase the number of positions with matching nucleotides (Hardison 2003, 157). Before the sequences can be scored, the letters of both strings must appear in their original order, and the lengths (counting gaps) of the strings must be equal. As an example of the potential complexity of this task, consider the various possible alignments of two simple words: HOUSE and HOME (see Figure 1).

Essentially, the purpose of these alignments is to show that two things that are not clearly related in fact have some relation. The goal is to align words, or genes, in such a way that a similarity between them becomes more apparent.

One way of choosing the best alignment is with the help of a scoring scheme. A scoring scheme assigns an optimality score to each alignment by counting the number of identical bases, different bases, and gaps. Consider the following scoring scheme as an example:

- If either character in a column is a gap, that column scores -2;
- If the letters are identical, the column scores $+1$;
- If the letters are different, the column scores -1;
- The score of the alignment is the sum of its columns' scores.

Following the rules listed above, the scores for the seven alignments pictured above are as follows:

1. $1 + 1 + (-1) + (-1) + (-2) = \mathbf{-2}$
2. $(-2) + (-1) + (-1) + (-1) + 1 = \mathbf{-4}$
3. $(-2) \times 9 = \mathbf{-18}$
4. $(-2) + (-1) + (-1) + (-1) + (-2) + (-2) = \mathbf{-9}$
5. $1 + 1 + (-2) + (-1) + 1 = \mathbf{0}$
6. $1 + 1 + (-1) + (-2) + 1 = \mathbf{0}$
7. $1 + (-2) + (-1) + (-2) + (-1) + (-2) = \mathbf{-7}$

The fifth and sixth alignments are the most optimal since they score the highest. However, the optimal alignment under a specific scoring scheme may not be the true one, if for no other reason than because the scoring scheme is itself a postulated norm that is in need of justification.

The point of these alignments is that if two sequences can be aligned correctly and the difference between them is small, then an assumption of a shared similar function is postulated. As an illustration of this point, consider a different (and more complex) example (see Figure 2). This example is, in fact, a real genetic alignment. More precisely, it is an example of three possible alignments between the genetic sequences of a human and an orangutan. As with the comparison between HOUSE and HOME, the three alignments would be scored using a scoring scheme, and the most optimal alignment could then be used as an inference for functional similarity. However, in practice, we may not always be entitled to make that inference because similarity between nucleotides does not guarantee functional similarity.

Currently, scientists believe that the most reliable way to infer function from similarity is to compare homologous genes, that is, genes that share a common evolutionary ancestor. Thus, if a newly sequenced gene turns out to be similar to a previously sequenced gene and a relevant evolutionary relationship can be inferred, then the function of the new gene is likely to be the same as, or at least similar to, the function of the known gene. The underlying assumption is that "common features of two or-

```
Human      CCTCCGCCGCGCCG    CTCCGC GCCGCCGGGCA           CGGCC
Orangutan  CC           GTCGCCTCCGCCACGCCGCGCCACCGGGCCGGGCCGGCCCGGCCC

Human            CCTCCGCCGCGCCGCT        CCGCGCCGCCGGGCACGGCCCCGC
Orangutan  CCGTCGCCTCCGCCACGCCGCGCCACCGGGCCGGGCCGGCCCGGCCCGCCCCGC

Human      CCTCCGCCGCGCCG    CTCCGCGCCGCCGGG CAC  GGCC
Orangutan             CCGTCGCCTCCGCCACGCCGCGCCACCGGGCCGGGCCGGCCCGGCCC
```

Figure 2. Redrawn from Marks 2002, 26.

ganisms will often be encoded within the DNA that is conserved between species" (Hardison 2003, 156). In other words, the DNA sequences, which encode the proteins responsible for functions conserved between common ancestors, will be preserved. Now, if the organisms being compared are related to a sufficiently high degree, such that the gene order between them has been conserved (this is known as "synteny"; Brown 2002, 214), then homologous genes may be identified based on their relation to other homologous genes. Otherwise, homologous genes may be identified based on sufficient sequence similarity. Some comparative genomic studies, for example, require that a sequence must show significant similarity to that of another species—over at least 80% of its length—before it can be considered its homolog (Rubin et al. 2000, 2205). In the end, neither sequence similarity nor synteny can guarantee homology—homology can only be inferred from each with more or less certainty.

Although there are problems with identifying homologous genes with certainty, homology usually serves as the basis for the frequently quoted statement that we share a certain percentage of our genes with other species. Thus, if the media report that humans share 75% of their genes with pumpkins, this translates to "there is 75% nucleotide similarity between homologous genes in humans and in pumpkins." The percentage quoted is usually an average score, since some homologous genes are almost identical, while others are hardly recognizable as closely related (Stubbs 1999).

4. How Are Comparative Genomic Claims Guiding Scientific Research?

Perhaps it is not a surprise that the media have taken an interest in reporting amusing facts about the percentage of genes that humans share with other animals, but one may wonder if, and to what end, such claims are used in scientific research. There are at least three motives for using similarity claims in research. First, knowing the overall genetic similarity between humans and other organisms helps in selecting model organisms. Second, identifying similarities between genetic sequences of different species can aid in gene annotation. Finally, comparing gene counts or the percentage of genes shared between species is more than a simple empirical exercise. If genes are functional units and there are thousands of functional units present, the expression of the phenotype is significantly affected by these thousands of units of genetic activity. To put this point slightly differently, "the set of these genes tightly mirrors what is meant by a genetic contribution to the phenotype" (Fogle 2001, 22). This underlying assumption is the reason why scientists find it puzzling that virtually all (99%) of the protein-coding genes in humans align with homologs in mice, yet only about 60% of the *C. elegans* genes encoding proteins have clear homologs in *C. briggsae* (Stein et al. 2003). The extensive conservation in protein-coding regions may be expected between humans and mice because they share many biochemical functions. However, the two worms (*C. elegans* and *C. briggsae*) share an equivalent amount of phylogenetic separation as humans and mice, they are difficult to distinguish morphologically, and they probably have similar patterns of development, yet "they achieve these similarities with some significant differences in the gene sets" (Hardison 2003, 158).

The discrepancy seen between the percentages of genes shared by animals with an equivalent amount of phylogenetic separation and a comparable amount of functional similarity—on the biochemical level—is a recent puzzle, but there have been others, some acquiring the status of "paradoxes," and all involving the relationship between genes and organismal complexity.[2]

For example, before the structure of DNA was discovered, Vendrely and Vendrely (1948) hypothesized that the size of the DNA complement (its C-value) correlated with the level of the complexity of an organism. However, their C-value hypothesis soon turned into a "C-value paradox" (Thomas 1971), when it became clear that the size of the human DNA complement was smaller than that of various amphibians, insects, and plants. Some hope of resolving the C-value paradox came with the discovery of "junk DNA" (i.e., noncoding DNA), especially when scientists realized that the functional elements (i.e., protein-coding genes) may occupy only as little as 1.5% of the entire DNA complement (Brown 2002). Unfortunately, once noncoding DNA was taken into account, the number of genes still failed to correlate with organismal complexity. The new puzzle, dealing with genes rather than the entire DNA complement, was named the "G-value paradox" by Hahn and Wray, who point out that "the published G values of the completely sequenced eukaryotes make it clear that we have not yet resolved the C-value paradox—it has merely given way to the G-value paradox" (2002, 73). The resulting G-value paradox is another puzzling reminder that differences in complexity between species do not seem to be related, in any simple way, to gene number.

5. Problems. The story of the G-value paradox serves as a simple illustration of how gene concepts can influence scientific results. Given that the G-value of an organism depends entirely on the definition of the gene used, what we see is that, if we want to avoid problematic scientific results, we have to ensure that our gene concepts are mapping onto what we take them to be measuring. This is deceptively simple advice, considering Rich-

[2] Complexity is a moving target. It is hard to say what the word 'complexity' does, or ought to, refer to. At present, the best-known indicator of complexity, in this context, is the diversity of the proteome of an organism (cf. Moss 2003).

ard Burian's observation that "there is a single fact of the matter about the structure of DNA, but there is no single fact of the matter about what the gene is" (1985, 37). In other words, ensuring that our gene concepts map onto what we take them to be measuring is not easy because it is not always clear to what the term 'gene' does, or ought to, refer. The gene concept has grown and evolved in complexity, and there is still no consensus among philosophers and historians of biology as to what the gene actually is (cf. Kitcher 1992; Waters 1994; Burton, Falk, and Rheinberger 2000; Keller 2000; Neumann-Held 2001; Moss 2003; Sarkar 2005; Stotz 2006).

Luckily, we do not need to know the *true* nature of the gene in order to assess whether comparative genomic studies generate problematic results. Instead, as mentioned above, the assessment consists of checking if the gene concept employed in the studies maps onto what comparative genomicists suppose their studies to be measuring. To begin this assessment, let us first investigate whether the gene concept employed in the studies maps onto a *structural* or a *functional* conception of the gene. If the gene employed as the unit of measurement corresponds to a structural gene concept, then comparative genomicists must view genes as ordered sequences of DNA (Keller 2000, 71–72; Moss 2003, 46). If this is the case, then the claim that "75% of human genetic make-up is the same as a pumpkin" is merely a claim about the percentage of nucleic acid we share with pumpkins. However, if the gene employed in the relevant study maps onto a functional gene concept, then comparative genomicists must view genes as entities that give rise to phenotypes by first impressing themselves upon the cell through the dynamic interaction among many players (Keller 2000, 71–72). On this interpretation, the above claim is now substantive in the sense that it tells us how similar we are to pumpkins overall.

If we look back on the information I have presented to this point—regarding the *what*, *how*, and *why* of comparative genomics—it should be clear that comparative genomicists think their studies are measuring something substantive. Various pieces of evidence point to this conclusion. First, recall that the aim of comparative genomics is to decipher how genes function and to understand the relationship between genotype and phenotype. Second, comparative genomic findings have been used to pick out model organisms, assign functions to unannotated genes, and explain the relationship between genes and organismal complexity. Finally, the theoretical assumption guiding the field of comparative genomics is that common features of organisms will often be encoded within the DNA that is conserved between the species. Notice that in each case the assumed unit of measurement is a gene that drives the function of cells and ultimately organisms. If the assumed unit of measurement is based on a functional gene concept, then comparative genomicists must be under the impression that their studies are measuring something substantive. However, I will now argue that the *actual* unit of measurement employed in comparative genomic studies corresponds to a *structural* gene concept. If correct, my argument will render many of the implications drawn from comparative genomic studies largely unwarranted, if not completely mistaken.

As I explained in Section 3, comparative genomic studies compare genes as if they were words; that is, they assume that genes "match" if the type and order of their letters match. What does it mean for genes to match? It means that they are functionally similar, insofar as the matching sequences of DNA encode a similar functional product. However, it turns out that genes cannot be accurately compared in the same way as words because the functional products of genes do not necessarily match when the type and order of the nucleotides match. In fact, there are at least two ways for regulatory mechanisms of gene expression to construct functional products, that is, proteins, which are *not mirrored* in any linear DNA sequence (Stotz 2006, 908). One is through alternative splicing—when the original DNA sequence is reshuffled. The other is through mRNA editing—when the original sequence is modified. What I will turn to now is an overview of these two frequently overlooked features of cell biology.

In higher eukaryotes, most of the DNA sequences describing proteins have a modular arrangement in which exons—the protein-coding regions—are interspersed with noncoding introns. Human genes have, on average, nine exons per gene, though there is substantial variation in the number of exons per gene (Brown 2002, 19). During gene expression, the initial RNA that is synthesized is a copy of the entire structural gene, including the introns and the exons, but before the messenger RNA (mRNA) is transported from the nucleus to the cytoplasm (where it directs synthesis of the protein), the introns are spliced out from the pre-mRNA, and the exons join together to form mRNA. Initially, it was thought that splicing was a straightforward process that led to the joining of neighboring exons. Now, however, it is thought that many pre-mRNAs undergo *alternative splicing*, during which some exons are skipped and the rest are joined together, creating messages that can code for different proteins (Brett et al. 2001; Downes 2004). For example, in the egg-laying hormone of *Aplysia*, one and the same stretch of DNA gives rise to 11 protein products involved in the reproductive behavior of the snail (Rheinberger and Müller-Wille 2007). While the exact number of genes that undergo alternative splicing is not known, the estimates for the human genome are currently at 60%, with some of them having up to 100 different splice forms (Leipzig, Pevzner, and Heber 2004).

Besides alternative splicing, another gene regulatory mechanism that can significantly diversify the proteome is *mRNA editing*. Whereas most other forms of posttranscriptional modifications (e.g., alternative splicing) retain the correspondence of the primary structure of exon and gene product, mRNA editing disturbs this correspondence by changing the primary sequence after its transcription. During mRNA editing, the nucleotide sequence is systematically altered by enzymes that excise old and insert new nucleotides, giving rise to protein products not reflected in the original DNA sequence (Sarkar 2005, 193; Stotz 2006, 909). mRNA editing is thus a very extreme mechanism of "*genomic information modification*, which can be rather extensive with up to several hundred modified nucleotides" (Stotz 2006, 909). The resultant editing events can have profound effects on the function of transmembrane receptors and ion channels in mammalian neural tissues, inflammation, cardiovascular disease, as well as cancer (Stotz 2006, 909).

The two above-mentioned genetic features demonstrate that eukaryotic DNA alone does not lead us directly to the primary sequence of a protein, let alone the protein's tertiary structure and its role in complex phenotypic traits (Stotz 2006, 908). However, because comparative genomic studies test neither the DNA sequences experimentally to see what combination of exons will join together nor which nucleotides will be modified after transcription, *in silico* they cannot account for the various protein structures that may result from this high percentage of genes that undergo alternative splicing and mRNA editing. In order to appreciate the drastic consequences of all this, it might be helpful to consider an example of the *paralytic* (*para*) gene, which encodes the major voltage-gated action potential sodium channel in *Drosophila*. This gene contains 13 alternative exons, so it can potentially synthesize 1,536 different RNAs through alternative splicing. If we include mRNA editing, 1,032,192 different *para* transcripts can theoretically be synthesized from this single gene (Graveley 2001). Although the extended functional consequences of alternative splicing and mRNA editing are not obvious, it seems that ignoring the complex way in which these genes are expressed can result only in misleading genetic comparisons.

In the end, the fact that comparative genomic studies measure the similarity between species at the nucleic acid level (in the chromosomal DNA), without regard for the functional end product, means that the unit of measurement employed in these studies corresponds to a structural, rather than a functional, gene concept. The problem is that these studies go on to make substantive claims about the amount of similarity we share with other organisms—as though the unit of measurement were based on a functional gene concept.

6. Conclusion. In conclusion, let me first say that the discovery of alternative splicing and mRNA editing has made it clear that equating a gene with an uninterrupted stretch of DNA can no longer capture the complicated molecular-developmental details of gene expression. The fact that there is no linear flow from the DNA sequence to its product but, instead, the whole developmental system is involved in gene expression makes it very difficult to draw a clear boundary between gene and environment (Stotz 2006, 914). As a result, by focusing too much on stretches of DNA when comparing organisms, comparative genomicists have presupposed an overly simplistic connection between genotype and phenotype, a connection that, even at the narrowest molecular level, is far from straightforward.[3]

Finally, let me review what I have argued: I began by pointing out that the aim of comparative genomics is to say something about the way in which genes function, and 'function', at the bare minimum, means the production of proteins. Next, I explained that the fundamental assumption of homologous gene alignments employed in comparative genomic studies is the idea that genes are discrete entities—like words—whose functional end products can be inferred from the sequence. I then argued that the growing number of complexities involved in gene expression make it increasingly difficult to compare genes as if they were words because the functional products of genes do not necessarily match if the type and order of the nucleotides match. As a result, I showed that claims coming out of comparative genomics are not entitled to the substantive force they carry with them—in both scientific research as well as media reports—because what comparative genomicists in fact measure does not map onto what they claim to be measuring.

REFERENCES

Auffray, Charles, Sandrine Imbeaud, Magali Roux-Rouquie, and Leroy Hood (2003), "Self-Organized Living Systems: Conjunction of a Stable Organization with Chaotic Fluctuations in Biological Space-Time", *Philosophical Transactions of the Royal Society of London Series A* 361: 1125–1139.

Beurton, Peter J., Raphael Falk, and Hans-Jörg Rheinberger, eds. (2000), *The Concept of the Gene in Development and Evolution*. Cambridge: Cambridge University Press.

Brett, David, Heike Pospisil, Juan Valcárcel, Jens Reich, and Peer Bork (2001), "Alternative Splicing and Genome Complexity", *Nature Genetics* 30: 29–30.

Brown, Terence A. (2002), *Genomes*. New York: Wiley.

Burian, Richard M. (1985), "On Conceptual Change in Biology: The Case of the Gene", in David J. Depew and Bruce H. Weber (eds.), *Evolution at a Crossroads: The New Biology and the New Philosophy of Science*. Cambridge, MA: MIT Press, 21–42.

Clark, Melody, ed. (2000), *Comparative Genomics*. Norwell, MA: Kluwer.

Downes, Stephen M. (2004), "Alternative Splicing, the Gene Concept, and Evolution", *History and Philosophy of Life Science* 26: 91–104.

Durrani, Monise (1999), "Similarity in Diversity", *BBC News*, July 5, http://news.bbc.co.uk/2/hi/science/nature/386516.stm.

Dwyer, Rex A. (2002), *Genomic Perl: From Bioinformatics Basics to Working Code*. Cambridge: Cambridge University Press.

[3] See Lewontin 2002 and Moss 2003 for further discussion of the complex relationship between genotypes and phenotypes.

Fogle, Thomas (2001), "The Dissolution of Protein Coding Genes in Molecular Biology", in Beurton et al. 2000, 3–25.

Graveley, Brenton R. (2001), "Alternative Splicing: Increasing Diversity in the Proteomic World", *TRENDS in Genetics* 17: 100–107.

Gregory, Ryan T. (2005), *The Evolution of the Genome*. Burlington, MA: Elsevier.

Hahn, William M., and Gregory A. Wray (2002), "The G-value Paradox", *Evolution and Development* 4: 73–75.

Hardison, Ross C. (2003), "Comparative Genomics", *PLoS Biology* 1: 156–160.

Karplus, Kevin, and Kimmen Sjolander (1997), "Predicting Protein Structure Using Hidden Markov Models", *Proteins: Structure, Function, and Genetics* 1: 134–139.

Keller, Evelyn F. (2000), *The Century of the Gene*. Cambridge, MA: Harvard University Press.

Kitcher, Philip (1992), "Gene: Current Usages", in Evelyn F. Keller and Elizabeth A. Lloyd (eds.), *Keywords in Evolutionary Biology*. Cambridge, MA: Harvard University Press, 128–131.

Leipzig, Jeremy, Pavel Pevzner, and Steffen Heber (2004), "The Alternative Splicing Gallery (ASG): Bridging the Gap between Genome and Transcriptome", *Nucleic Acids Research* 32: 3977–3983.

Lewontin, Richard (2002), *The Triple Helix*. Cambridge, MA: Harvard University Press.

Marks, Jonathan (2002), *What It Means to Be 98% Chimpanzee*. Berkeley: University of California Press.

Miller, Webb, Kateryna D. Markova, Anton Nekrutenko, and Ross C. Hardison (2004), "Comparative Genomics", *Annual Review of Genomics and Human Genetics* 5: 15–56.

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

Neumann-Held, Eva (2001), "Let's Talk about Genes: The Process Molecular Gene Concept and Its Context", in Susan Oyama, Paul E. Griffiths, and Russell D. Gray (eds.), *Cycles of Contingency*. Cambridge, MA: MIT Press.

Patis, Carrie (2007), "The First Marsupial Genome Sequence", *Nature Reviews Genetics* 8: 408–409.

Pearson, Helen (2006), "Codes and Enigmas", *Nature* 444: 259–261.

Rheinberger, Hans-Jörg, and Staffan Müller-Wille (2007), "Gene", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/fall2007/entries/gene/.

Rubin, Gerald M., Mark D. Yandell, Jennifer R. Wortman, George L. Gabor Miklos, Catherine R. Nelson, Iswar K. Hariharan, Mark E. Fortini, et al. (2000), "Comparative Genomics of the Eukaryotes", *Science* 287: 2204–2215.

Sarkar, Sahotra (2005), *Molecular Models of Life*. Cambridge, MA: MIT Press.

Sivanshankari, Selvarajan, and Piramanayagam Shanmughavel (2007), "Comparative Genomics—a Perspective", *Bioinformation* 1: 376–378.

Stein, Lincoln D., Zhirong Bao, Darin Blasiar, Thomas Blumenthal, Michael R. Brent, Nansheng Chen, Asif Chinwalla, et al. (2003), "The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics", *PLoS Biology* 1 (2): E45.

Stotz, Karola (2006), "With 'Genes' like That, Who Needs an Environment? Postgenomics's Argument for the 'Ontogeny of Information'", *Philosophy of Science* 73: 905–917.

Stubbs, Lisa (1999), "How Closely Related Are Mice and Humans? How Many Genes Are the Same?", in Functional and Comparative Genomics Fact Sheet, U.S. Department of Energy Office of Science, http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml.

Thomas, C. A. (1971), "The Genetic Organization of Chromosomes", *Annual Review of Genetics* 5: 237–256.

Vendrely, Roger, and Collette Vendrely (1948), "La teneur du noyau cellulaire en acide désoxyribonucléique à travers les organes, les individus et les espèces animales: Techniques et premiers résultats", *Experientia* 4: 434–436.

Waters, Kenneth C. (1994), "Genes Made Molecular", *Philosophy of Science* 61: 163–185.

Waterston, Robert H., Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, Richa Agarwala, et al. (2002), "Initial Sequencing and Comparative Analysis of the Mouse Genome", *Nature* 420: 520–562.

Wong, Limsoon (2004), *The Practical Bioinformatician*. Singapore: World Scientific.