THE EVOLVING CONCEPT OF THE GENE

■ RAFAEL VICUÑA

> *… where the meaning of most four-letter words is all too clear, that of gene is not. The more expert scientists become in molecular genetics, the less easy it is to be sure about what, if anything, a gene actually is.*[1]

One of the disciplines of biology that showed more progress than any other during the 20th century was genetics, with advances in the area of molecular genetics being perhaps the most outstanding. Statements such as: *The gene concept has certainly been one of the landmarks in the history of science in the 20th century*;[2] *There can be little doubt that the idea of the gene has been the central organizing theme of twentieth century biology*;[3] *During the twentieth century, the gene has emerged as the major driving force of biology*,[4] reflect the relevance of the gene as a focal subject of study during the last century. One might think that an accurate concept of the gene was pivotal for this outcome. However, paradoxically, this is not the case, since defining the gene has always proven to be a difficult task, especially at the present time. In spite of the latter, geneticists have been able to thoroughly study how traits are passed down to progeny and how gene variation constitutes the basis of evolution.

This essay does not pretend to summarize the history of genetics, nor of the gene itself. Rather, its purpose is to highlight the landmarks of the evolving concept of the gene and to depict some recent findings that are making understanding the gene even more difficult. For a comprehensive collection of essays dealing with historical and epistemological perspectives of the concept of the gene, a recent book published by Cambridge University Press is highly recommended.[5]

---

[1] Pearson, H. What is a gene? *Nature* 441, 399-401, 2006.

[2] El-Hani, C.B. Between the cross and the sword: The crisis of the gene concept. *Genet. Mol. Biol.* 30, 297–307, 2007.

[3] Moss, L. *What genes can't do*. Cambridge, The MIT Press, 2003.

[4] Rédei, G.P., Koncz, C. & Phillips, J.D. Changing images of the gene. *Adv. Genetics* 56, 53–100, 2006.

[5] *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, 2000.

The birth of the gene concept

Heredity began to be studied as a scientific discipline only in the 20th century. In the early days, breeders were more concerned with the statistical analysis of inherited traits than with the material causation of them. The Austrian monk Gregor Mendel, after his discovery of the laws of segregation and of independent assortment published in 1865, was the first to suggest that reproductive cells were endowed with elements, characters or factors that had the potential for a trait. However, he did not make any distinction between whatever was transmitted in the seeds and the traits themselves. Three years later, most probably unaware of Mendel's work but certainly inspired by the teachings of Hippocrates, Charles Darwin presented what he dubbed a 'provisional hypothesis of pangenesis', according to which body cells produce minute particles called gemmules that travel to the reproductive cells, where they congregate prior to fertilization.[6] Although this hypothesis was not supported by observation, it allowed Darwin to explain phenomena such as the intermediate nature of hybrids and the heredity of acquired characters. Interestingly, Darwin's half cousin Francis Galton committed himself to prove the validity of this hypothesis by transfusing blood between dissimilar breeds of rabbits. After examining the features of their offspring, he concluded that there was no evidence of transmission of gemmules by this body fluid.

In 1889, reflecting his disagreement with heredity of acquired characters, Hugo de Vries proposed the theory of 'intracellular pangenesis', according to which animals and plants possess independent characters that are correspondingly associated with distinct particles called pangenes. These particles are located in the nucleus, where they replicate when the cell divides. Daughter cells each receive a complete set of pangenes, including the precursors of reproductive cells during their differentiation. Thus, pangenes do not leave the cells and their travel along the body never takes place. A few years earlier, the German zoologist August Weismann had advanced the similar theory of 'The continuity of the germplasm'. In his own words: *…heredity is brought about by the transference from one generation to another of a substance with a definite chemical, and above all, molecular constitution. I have called this substance 'germ-plasm' and have assumed that it possesses a complex structure, conferring upon it the power of developing into a complex organism.*[7] Weismann

---

[6] Darwin, Ch. Provisional hypothesis of pangenesis. In: *Animals and plants under domestication*, vol. 2. Orange Judd, New York, pp. 428-483, 1868.

[7] Weismann, A. The continuity of the germ-plasm as the foundation of the theory of heredity. In: Poulton E.B., Schonland S., Shipley A.I.E. (eds) *Essays upon heredity and kindred biological problems by Dr. August Weismann*. Clarendon Press, Oxford, 1885.

thought that the germplasm was only present in the cell lineage that leads to the reproductive cells, whereas somatic cell lineages had received from their progenitors only the material required for the development of the respective organs and tissues. In contrast, studies conducted in plants had rightly convinced de Vries that each nucleus of the body contains the same complete set of pangenes.

By the dawn of the new century, geneticists were mostly using the term *unit-character* for the entities responsible for specific traits that behaved as indivisible units of Mendelian inheritance. But the Danish botanist Wilhelm Johanssen liked de Vries's theory of pangenes and in 1909 coined the term gene for the *special conditions, foundations and determiners which are present in unique, separate and thereby independent ways (by which) many characteristics of the organisms are specified*.[8] The new designation came to replace that of unit-character. Although by inference, genes had to be present in the germ cells, their physical constitution was unknown and the concept proved useful to account for the transmission of traits from one generation to the next. In other words, it was a concept mainly based on a function that could be identified by genetic methods. Johanssen also introduced the terms genotype and phenotype, thus clearly distinguishing between the potential for a trait and the trait itself.[9] From 1910 to 1915, studying segregation of mutations in the fruit fly *Drosophila melanogaster*, Thomas Morgan and his group showed that genes reside in chromosomes and that they occupy specific locations in them, as beads on a string. He figured that the ability of genes to recombine was proportional to their distance in the chromosome. Studies on X-linked inheritance in the same organism allowed him to assign genes to the X chromosome. All together, contemporary work gave rise to the perception of the gene as a unit of function (one trait), a unit of mutation and a unit of recombination, a vision that prevailed until the early 1940s.

However, genes were still considered mainly as entities having the potential for a trait and whose effects could be inferred from them. In other words, efforts were focused more in traits as manifestations of genes rather than in their material counterparts. Morgan himself made this clear during his Nobel Prize Lecture in 1933: *Now that we locate [the genes] in the chromosomes are we justified in regarding them as material units; as chemical bodies of a*

[8] Johannsen, W. *Elemente der Exakten Erblichkeitslehre*. Gustav Fisher, Jena, 1909. Cited by Hall, B.K. The gene is not dead, merely orphaned and seeking a home. *Evol. Develop.* 3(4), 225-228, 2001.

[9] Falk, R. What is a gene. *Stud. Hist. Phil. Sci.* 17, 133-173, 1986.

higher order than molecules? Frankly, these are questions with which the working geneticist has not much concerned himself, except now and then to speculate as to the nature of postulated elements. There is not consensus of opinion amongst geneticists as to what genes are – whether they are real or purely fictitious – because at the level at which the genetic experiments lie, it does not make the slightest difference whether the gene is a hypothetical unit, or whether the gene is a material particle.[10] This reductionist approach, still not constrained to a specific material counterpart, led Raphael Falk to coin the term *instrumental gene*, to imply a hypothetical construct that was accepted as if it was a real entity.[11]

But there were also manifestations of a more material conceptualization of the gene. The very fact that a gene could be mutated[12] or recombined was certainly a consequence of its physical identity. Perhaps this evidence may have influenced Herman J. Muller, a member of Morgan's group, to support the notion that genes are 'ultra microscopic particles' found in the chromosomes rather than a 'pure idealistic concept divorced from real things'.[13] Another inclination towards a material nature of the gene was the genomere model proposed by Eyster to interpret gene instability expressed in variegated traits in fruit flies and spotting in corn kernels. This model stated that genes are composed of different particles that are unequally distributed during mitotic divisions.[14] Investigators such as Correns, Anderson and Demerec favored the genomere hypothesis, until it was disproven few years later by Muller.[15] And there were also the results obtained in 1928 by Griffith, showing that some substance originally present in killed virulent *Pneumococcus* cells was able to transform a non-virulent live strain into a virulent one.[16]

In the early 40s, George W. Beadle and Edward L. Tatum were studying metabolism in *Neurospora* and showed that certain mutations in genes caused errors in specific steps in metabolic pathways. This observation gave rise to

[10] Thomas H. Morgan, The relation of genetics to physiology and medicine. Nobel Lecture, Stockholm, June 1933; cited by R. Falk in: What is a gene? *Stud. Hist. Phil. Sci.* 17, 133-173, 1986.

[11] Falk, R. The gene in search of an identity. *Hum. Genet.* 68, 195-204, 1984.

[12] Muller, H.J. Artificial transmutation of the gene. *Science* 46, 84-87, 1927.

[13] Falk, R. What is a gene? *Stud. Hist. Philos. Sci.* 17, 133-173, 1986.

[14] Eyster, W.H. A genetic analysis of variegation. *Genetics* 9, 372-404, 1924.

[15] Muller, H.J. The problem of genic modification. Proceedings of the Fifth International Congress of Genetics, Berlin, 1927. *Z Induktive Abstammungs Vererbungslehre* [Suppl 1]: 234-260.

[16] Griffith, F. The significance of pneumococcal types. *J. Hyg.* (London) 27, 113-159, 1928.

the 'one gene-one enzyme' hypothesis, supporting the view that genes carried information related to the metabolic processes taking place inside the cells and more specifically, that each individual gene is responsible for the synthesis of a single enzyme.[17] Chemistry also had a role to play. Vernon Ingram showed that changes in two abnormal hemoglobins due to mutations were in each case confined to a single amino acid residue of the globin polypeptide. Since there could be no doubt that genes determine the amino acid residues of polypeptide chains, the expression 'one gene-one enzyme' was modified to 'one gene-one polypeptide'.[18]

The nature of the genetic material became even more tangible when Oswald Avery[19] and collaborators showed that the substance causing transformation in experiments that followed the protocols of Griffith's was DNA. Unambiguous confirmation of the DNA theory of inheritance was obtained few years later by Alfred D. Hershey and Martha Chase.[20] The structure of DNA proposed by Watson and Crick in 1953 gave the definite stroke to the instrumentalist view of the gene in favor of the realistic one, initiating the *classical molecular gene concept*. This states that a gene is a stretch of DNA that encodes a functional product, a single polypeptide chain or RNA molecule. Implicit in it is the idea that this genome unit performs one single function. At last, then, structure and function were blended in the same concept.

The newly revealed structure of DNA also encouraged speculation about the still prevailing idea of the gene as a unity of function, mutation and recombination. Prior to 1955, several investigators had already obtained the first hints that the unit of function might not be indivisible, since not only more than one mutation could be mapped to the same gene but also intragenic recombination had been detected in *D. melanogaster* and the fungus *Aspergillus nidulans*[21] (see also references therein). Who most clearly confirmed this was Seymour Benzer. The so-called *cis-trans* complementation test led him to coin the word *cistron* to imply the unit of genetic function.

[17] Beadle, G.W. and Tatum, E.L. Genetic control of biochemical reactions in Neurospora. *Proc. Natl. Acad. Sci*. 27**:** 499-506, 1941.

[18] Dunn, L.C. Old and New in Genetics. *Bull. New York Acad. Med.* 40(5): 325-333, 329, 1964.

[19] Avery, O.T., MacLeod, C.M., and McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med*. 79**:** 137-158, 1944.

[20] Hershey, A.D., and Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36, 39-56, 1952.

[21] Portin, P. The origin, development and present status of the concept of the gene: A short historical account of the discoveries. *Current Genomics* 1, 29-40, 2000.

Essentially, a cistron is a contiguous or uninterrupted piece of DNA encoding a particular protein. Cistrons turned out to be much larger than the units of mutation and recombination, thus confirming the possibility of multiple mutations within a single gene and of intragenic recombination.[22] Up to the present time, cistron is considered to be a synonym of gene, although its use is rather unfrequent.

New findings set hurdles to the gene concept

The development of DNA sequencing and of gene manipulation techniques in the 1970s allowed the rapid unveiling of the structure of genes and the detailed mechanisms involved in the regulation of their expression. Furthermore, the sequencing of large stretches of DNA and even of whole genomes led to the concept of *nominal gene*, to denote a sequence of DNA whose features allow the prediction of a protein sequence or of a known RNA product.[23] However, several novel situations related to the structure and function of genes that could not have been previously envisaged made evident that the classical molecular gene concept was completely improper. All at once, the gene seemed to have lost its identity both as a structural and as a functional unit. First of all, genomes from eukaryotes do not only consist of genes. They contain a significant fraction of non-protein coding and even highly repetitive sequences. Although originally labeled junk DNA due to its apparent lack of function, it has recently been shown that a large fraction of this DNA is transcribed (see later). Meaningless sequences in the genome did not necessarily challenge the classical molecular gene concept implying a stretch of DNA encoding a functional protein or RNA macromolecule. Neither did the fact that there are certain tandem repetitions of meaningful sequences, such as those encoding histones and ribosomal RNA. But other features widespread in genomes certainly did, especially some that entailed sequence elements external to the coding region, as well as dissection of the gene into smaller units. Some of the new findings contesting the classical gene concept (a stretch of DNA encoding a functional product) are the following:

a) Regulatory elements: already in 1961, F. Jacob and J. Monod had introduced the term promoter to describe a sequence located upstream of

[22] Benzer S. The elementary units of heredity. In: McElroy, W. and Glass, B. (eds) *The Chemical Basis of Heredity*. John Hopkins Press, Baltimore, pp. 70-93, 1957.

[23] Burian, RM. Molecular epigénesis, molecular pleiotropy and molecular gene definitions. *Hist. Phil. Sci.* 26, 59-80, 2004.

the protein coding sequence that was responsible for controlling gene expression in bacteria. Later findings showed that all genes, both in prokaryotes and eukaryotes, require promoters for being transcribed into RNA. Since promoters can readily be recognized by their typical nucleotide sequences, they facilitate the identification of protein coding sequences in a genome. Thus, the term open reading frame (ORF) is commonly used to imply a DNA sequence that allegedly encodes a protein because it is flanked by a promoter (next to an initiation codon) and a stop codon. Often, promoters in eukaryotes are more effectively used when they are stimulated by cis-acting sequence elements called enhancers. These can be located either upstream or downstream of the promoter, sometimes thousands of base pairs away. Besides transcription, translation can also be regulated by sequence elements, which in this case are present in the transcript. A six nucleotide sequence located upstream of the initiating codon in bacterial mRNA, known as the Shine-Dalgarno element, contributes to positioning the initiating codon in the proper site of the ribosome. In mature eukaryotic mRNA, untranslated regions (UTRs) before the start codon (5' UTR) and after the stop codon (3' UTR) influence mRNA stability, mRNA localization and translational efficiency through proteins that specifically bind either one or the other, depending on the aspect to be regulated.

b) Intervening sequences: most eukaryotic genes are interrupted by non-protein coding sequences called introns, which are transcribed into RNA and thereafter removed prior to translation. Removal of introns and joining of the coding sequences (exons) is called splicing. Intron sequences largely exceed exons sequences with values of 20% versus 1.5%, respectively, in the human genome. On the other hand, many eukaryotic mRNAs can undergo alternative splicing, a process by which some exons are left out of the final transcript. In this case, a particular DNA segment in the genome can give rise to several variant proteins, thus expanding the coding capacity of the genome. It is estimated that 75% of the human genes are processed by alternative splicing. There is also the phenomenon of transplicing, mainly in lower eukaryotes, in which separate transcripts that may derive even from separate chromosomes are ligated to produce one mature mRNA. Splicing does not only occur at the RNA level. Intervening sequences in proteins (inteins) can be removed from a precursor protein and the flanking segments (exteins) can be ligated to generate a mature protein.

c) Transcripts including several genes: in bacteria it is widespread that genes involved in a particular biochemical pathway are clustered on the chromosome and transcribed together in a single polycistronic RNA. The gene cluster plus its single promoter is called an operon. Distribution of

genes in operon allows an efficient control of gene expression. Also, in bacteria, genes encoding 16S, 23S and 5S ribosomal RNAs are transcribed in a single pre-ribosomal RNA (30S), which is thereafter processed into its mature products. In turn, eukaryotes produce a 45S pre-ribosomal RNA that gives rise to 18S, 28S and 5.8S RNA. Studies on the human genome have also revealed the phenomenon called tandem chimerism, where two consecutive genes are transcribed into a single RNA. Differential splicing of this RNA can give rise to a fused protein containing domains encoded in both genes.[24,25]

d) Polyproteins: in this case, a transcript is translated into a protein that is subsequently cleaved to generate proteins with different activities. For example, transcription of retroviral DNA engenders one transcript comprising the *gag*, *pol* and *env* coding sequences. There are no intergenic sequences between them. The transcript is translated into a polyprotein corresponding to the *gag* and *pol* sequences that is cleaved into a total of six proteins: three viral structural proteins, an integrase, a protease and reverse transcriptase. On the other hand, splicing of the primary transcript gives rise to an mRNA encoding mainly the *env* gene. This is translated into another polyprotein that is processed to produce the viral envelope proteins.

e) Overlapping genes: in bacteria and viruses, as well in eukaryotes, genes sometimes overlap. Different proteins may be read from the same strand although in different reading frames. Reading frames may also be convergent or divergent, in which cases both DNA strands carry genetic information. When an entire coding sequence lies within the start and stop codon of another gene, typically in an intron, one speaks of a nested gene. There are also genes nested opposite the coding sequences of their host genes.

f) Genome rearrangements: immunoglobulins consist of two heavy and two light polypeptide chains. In turn, there are two types of light chains: kappa and lambda. Each of these chains, namely the heavy kappa and lambda chains, has a constant and a variable region. In all cases, the variable domain is encoded in a few hundred different gene sequences. Recombination of the latter with the sequences encoding the corresponding constant regions produces a wide diversity of light and heavy chains, which can in

[24] Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A. and Sorek, R. Transcription-mediated gene fusion in the human genome. *Genome Res.* 16, 30–36, 2006.

[25] Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E. and Guigó, R. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.* 16, 37–44, 2006.

turn associate in all combinations. These DNA rearrangements explain how a mammal genome can literally produce millions of immunoglobulins.

g) Gene sharing: occurs when a single protein performs multiple functions. The most well-known example of gene sharing is eye lens crystallins. When expressed at low levels, the protein in many tissues functions as a housekeeping enzyme, but when expressed at high levels in eye tissue, it becomes densely packed and forms lenses.

h) Transcript editing: in eukaryotes, tRNAs, rRNAs and mRNAs may undergo chemical modifications that alter the information originally present at the DNA level. RNA editing mechanisms include cytidine to uridine and adenosine to inosine deaminations, as well as nucleotide additions and insertions. For example, in RNA transcripts coding for proteins in the mitochondria of trypanosomes, uridine nucleotides are inserted with the help of guide RNAs that hybridize to the transcript and direct the endonucleolytic cleavage of the RNA, the insertion of uridine nucleotides by uridylyl transferase and the subsequent ligation of the transcript, which now has both an altered sequence and reading frame. It is estimated that about one thousand human genes have an adenosine to inosine deamination. Editing is at odds with the classical gene concept because the RNA requires retrieving information from other genes to configure its final message.

Recently, as a result of the ENCODE project, new surprises complicated even further our understanding of the organization of the genome and of the gene concept itself. The *enc*yclopedia *o*f *D*NA *e*lements (ENCODE) consortium is an initiative launched by the National Human Genome Research Institute of the National Institute of Health (USA). It started with a pilot project aimed at thoroughly scrutinizing 30 mega bases (one percent) of the human genome, distributed in 44 genomic regions, with the goal to identify and map all the functional genetic elements.[26] Conducted between 2003 and 2007 by 35 groups from 80 organizations around the world, the ENCODE project confirmed what the Human Genome Project had anticipated, namely, that a genome entails much more than a mere collection of protein coding genes. One of the major findings of the ENCODE project was the realization that the majority (>90%) of the DNA is transcribed into primary transcripts that give rise to RNAs of various sizes. Most of them correspond to novel non-protein coding transcripts, some of which overlap protein coding sequences,

---

[26] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799-816, 2007.

whereas others derive from regions outside the previously annotated genes. Furthermore, there is abundant (>60%) overlapping between sense–antisense transcription across the genome. On average, 5.4 transcripts were found per annotated gene, showing some strand overlapping as well. Alongside, several alternative transcription start sites were identified for each protein coding sequence. About two thirds of the transcripts possess a tissue-specific 5 distal sequence, which can be located >100 kb upstream of the canonical start site.[27] Occasionally, transcripts span more than one annotated gene, as revealed by the presence of exons deriving from them. It is unclear how these fusion transcripts are generated. Possible mechanisms include trans–splicing and simply extended transcription. Neither is it known whether these fusion transcripts are translated. In addition, contrary to traditional views, regulatory regions were found both upstream and downstream transcription starting sites. In a subject which is beyond that of the gene concept, the ENCODE project also revealed that functional sequences do not seem to be evolutionary constrained, as shown by comparison with 28 other mammals. Moreover, about 40% of the constrained regions do not seem to play any functional role. In summary, the pilot step of the ENCODE project showed that the genome is a far more complex system than originally envisaged, with a variety of interconnected elements whose functionality we are only beginning to unravel.

Multiple efforts aiming at a consensus notion of the gene

As mentioned above, the gene concept was initially instrumental. It then turned into a material one, when the DNA macromolecule was identified as the carrier of the genetic message. The unit character of Mendelian genetics became a sequence in the DNA encoding a functional product, either a protein or RNA. But from the time this classical gene concept proved to be inadequate in the light of the complexity of the genome, there have been various attempts to improve it. For example, Fogle has proposed that as opposed to a unit, a gene is a construct resulting from the assemblage of embedded, tandem and overlapping domains in the DNA, a domain being a sequence that can be distinguished by virtue of its structural properties (exon, promoter, enhancer, etc).[28] Thus, although two organisms may have

---

[27] Denoeud, F., Krapanov, P. *et al.*: Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* 17, 746-759, 2007.

[28] Fogle, T. Are genes units of inheritance? *Biol. and Philos.* 5, 349-371, 1990.

a similar number of 'classical' genes, they may differ to a great extent in the way the domains combine to constitute *domain sets for active transcription* (DSATs, in Fogle's nomenclature). In turn, Brosius and Gould offer a new terminology for better understanding genome structure, function and evolution.[29] They recommend the term *nuon* for denoting any stretch of nucleic acid sequence that may be identifiable by some criterion. A nuon can be a protein-coding sequence, intergenic region, exon, intron, promoter, enhancer, terminator, pseudogene, telomere, etc. A sequence undergoing an adaptive change should be called *aptonuon*. On the other hand, it is well known that duplicated genes have the potential to give rise to a novel function, after passing through a silent stage. During this period, they should be called potogenes or *potonuons* (*pot*ential nuons). If these sequences appear to have become obliterated as genomic noise, they should be termed *naptonuons* (*n*on*apt*ive nuons). If, in contrast, potonuons have been coopted for another function, they should be called *xaptonuons* or xaptogenes, since they constitute exaptation events.[30] This proposition by Brosius and Gould has not prevailed in the scientific community. A couple of years later, Waters proposed that the fundamental concept of the gene is that of a linear sequence in a product at some stage of genetic expression.[31] Thus, an intron is part of the gene if the focal point is the process of transcription itself, but it is not a gene if the focus of interest is the function of the protein encoded in it. Strange as it may seem, this concept allows a single gene at the DNA level to encode for several genes at the mRNA level (alternative splicing). Considering that this definition varies during different stages of the expression process, it does not contribute to clarification in language use.

Interestingly, Griffiths and Neumann-Held think that a univocal definition of the gene may not be necessary or even desirable, since different gene concepts may be useful in different areas of biology.[32] However, their opinion is that it is critical to be aware of the differences among the various concepts in order to use them properly in their corresponding domains. These authors

---

[29] Brosius, J. and Gould, S.J. On 'genomenclature': A comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proc. Natl. Acad. Sci. USA* 89, 10706, 10710, 1992.

[30] Gould and Vrba coined the term exaptation for designating functional features of the phenotype that were not built by natural selection as adaptations of the original function but were rather coopted from structures arising from adaptations for other functions.

[31] Waters, C.K. Genes made molecular. *Phil. Sci.* 61, 163-185, 1994.

[32] Griffiths, P.E. and Neumann-Held, E.M. The many faces of the gene. *BioScience* 49, 656-662, 1999.

are particularly concerned about the distinction between the molecular gene and the evolutionary gene. According to them, the difficulties with the classical molecular gene concept arise because it is centered in structure rather than in function (a stretch of DNA encoding a protein or RNA product). Therefore, they suggest replacing it by the *molecular process gene concept*, in which gene denotes the recurring process that leads to the temporally and spatially regulated expression of a particular polypeptide product. In this new definition, although the gene still implies a DNA segment encoding a polypeptide, the emphasis is placed on the process that allows this sequence to be expressed. Thus, if a transcript of a certain DNA segment undergoes differential splicing or editing that vary depending on the tissue or stage of development, such segment fits the proposed definition. Moreover, the latter takes into consideration other functions that participate in causing the sequence to generate its product. On the other hand, there is the *evolutionary gene concept*, first introduced by Williams[33] and then elaborated by Dawkins to denote any stretch of DNA that can be replaced by an alternative (alelomorphic) sequence in future generations.[34] Griffiths and Neumann-Held agree with Dawkins in that evolutionary genes need not necessarily be molecular genes, *e.g.* often do not correspond to specific stretches of DNA. However, as opposed to Dawkins, they lay emphasis on the fact that rather than being loosely defined segments in the DNA, evolutionary genes have particular roles in the expression of phenotypic traits. The evolutionary gene concept has also been worked by P. Beurton, who claims that the gene is the smallest collection of genetic elements that underlies a single adaptive difference and is thus a target of natural selection.[35] In this case, a collection refers to the fact that a phenotypic trait may involve several genetic elements, each of which is a target of selection. Another approach that is focused to function rather than to structure is the *developmental gene concept*, as advanced by Gilbert[36] and

[33] Williams, G.C. *Adaptation and natural selection*. Princeton NJ. Princeton University Press, 1966.

[34] Dawkins, R. *The extended phenotype*. Oxford: W.H. Freeman, 1982.

[35] Beurton, P.J. A unified view of the gene, or how to overcome reductionism, in *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, pp. 286–316, 2000.

[36] Gilbert, S.F. Genes classical and genes developmental: The different use of genes in evolutionary syntheses, in *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, pp. 178–192, 2000.

Morange.[37] They assign this name to genetic elements that play a leading role in shaping the phenotype through the development of parts or segments of organisms. Although these genes are widely distributed among metazoans, the concept of developmental gene is rather restricted to this particular aspect of biological function.

In 2003 Snyder and Gerstein defined the gene as a complete chromosomal segment responsible for making a functional product.[38] This definition encompasses the inclusion of both regulatory and coding regions, the expression of a gene product and the requirement that it be functional. Criteria to be used in order to identify genes in the DNA sequence of a genome include the identification of ORFs, specific sequence features (codon bias, splicing sites), sequence conservation among organisms, evidence for transcription and gene inactivation, the latter being aimed at ascertaining gene's function.

A different approach has been undertaken by Lenny Moss.[39,40] This author argues that there are two markedly distinctive meanings or senses of the gene. Although both are associated to the phenotype, neither indicates that the phenotype can be decomposed down to a compilation of genes. First there is *gene-P*, which has a predictable relationship with some feature of the phenotype. One speaks of the gene for muscular dystrophy, obesity or premature aging. In other words, every time we use the expression a 'gene for' a certain trait, we are referring to a gene-P. This concept is indeterminate with respect to the material gene, *i.e.* to the specific sequence of DNA. So indeterminate is this concept with respect to the DNA sequence that in common language one often speaks of the gene for a certain trait when such trait (a disease for example) is expressed due to the absence of the wild type or normal sequence. The P in the gene-P concept stands for 'preformationism', because it evokes the idea that all the traits are determined at the moment of birth. In contrast, there is the concept of *gene-D*,

---

[37] Morange, M. The developmental gene concept: History and limits, in *The concept of the gene in development and evolution: Historical and epistemological perspectives.* Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, pp. 193–218, 2000.

[38] Snyder, M. and Gerstein, M. Defining genes in the genomics era. *Science* 300, 258–260, 2003.

[39] Moss, L. The question of questions: What is a gene? Comments on Rolston and Griffiths&Stotz. *Theor. Med. Bioethics* 27, 523–534, 2006.

[40] Moss, L. The meanings of the gene and the future of phenotype. *Genet. Soc. Policy* 4, 38–57, 2008.

which is specifically associated with a particular DNA sequence that can give rise to an RNA transcript. Gene-D is not determined with respect to phenotype, because it is unable to predict the appearance of a particular trait. Most often this is due to the fact that, as shown by studies at the molecular level, each DNA sequence contributes or is involved in the manifestation of several phenotypic outcomes, the resulting one depending on contextual factors. The D in gene-D stands for 'developmental resource', having therefore a more holistic scope that gene-P. According to Moss, a typical gene-D is NCAM (neural cell adhesion molecule), which in the fly *Drosophila* can give rise to about 38,000 proteins by differential splicing of a gene that possesses 19 exons. The different domains encoded in each of these exons will determine the cellular activity of the protein and hence the resulting phenotype.

On the other hand, Scherrer and Jost have proposed to preserve the concept of the gene as a basis of a function, that is to say, the sequence encoding a polypeptide within an mRNA, even though in most cases such sequence is not present at the DNA level as an uninterrupted sequence.[41] The gene in the mRNA is flanked by untranslated regions (5'- and 3'- UTRs). Superimposed onto the coding sequence is the *genon*, a program of oligomotifs that are eventual binding sites for regulatory proteins or small RNAs. At a higher level there is the *transgenon*, constituted by all the factors that influence gene expression by binding to the motifs in the genon. These factors are selected from the holotransgenon, which comprises all the factors (polypeptides and small RNAs) influencing gene expression in the cell. These concepts also apply when the gene product is RNA instead of a protein. A different approach is taken by Keller and Harel, which, according to these authors, is better grounded in biological findings than the gene has proven to be.[42] They define a *dene* as a DNA sequence plus all the elements that in a dynamic fashion make it functional (regulatory proteins and RNAs, epigenetic modifications, etc). The *bene* is the behavior of the organisms with which the dene is associated. In turn, the *genetic functor* or *genitor* is the logical relation that says whenever the organism's DNA is seen to satisfy the property expressed by the dene, its behavior satisfies the property expressed by the bene. This nomenclature offered by Keller and Harel is intended to em-

[41] Scherrer, K. and Jost, J. The gene and the genon concept: a functional and information-theoretic analysis. *Molec. Syst. Biol.* 3, 1-11, 2007.

[42] Keller, E.F. and Harel, D. Beyond the gene. PLoS ONE 2(11):e1231. doi:10.1371/journal.pone.0001231.

phasize the distinction between what an organism statically is (what it inherits) and what it dynamically does (its functionality and behavior).

After assessing the novel findings of the ENCODE project, Gerstein *et al.*[43] suggested five criteria to update the definition of the gene, namely: 1) the new description should comprise the former meaning of a gene; 2) it should be valid for any living organism; 3) it should be simple; 4) it should be straightforward, so anybody could distinguish the number of genes in a particular genome and 5) it should be compatible with other biological nomenclature. In addition, the new definition must take into account that the gene is a genomic sequence encoding a functional protein or RNA, it must consider the union of overlapping sequences when there are several functional products and it must be coherent in the sense that the union must be done separately for protein and RNA products, not being necessary that all the products share a common sequence. Gerstein *et al.* further put forward a new definition of the gene as a union of genomic sequences encoding a coherent set of potentially overlapping functional products. If there are no introns or no overlapping products, the new definition coincides with the classical one. Since this new definition covers only coding sequences, it does not include regulatory regions and untranslated regions (5' and 3' UTRs) in the RNA. In addition, it does not cover RNA editing.

There are two other recent attempts to define the gene that deserve to be mentioned because they represent collective efforts. One is that of the Human Genome Nomenclature Organization, which states that a gene is a DNA segment that contributes to phenotype/function. In the absence of a demonstrated function, a gene may be characterized by sequence, transcription or homology.[44] The other one, adopted by the Sequence Ontology consortium, was elaborated by 25 scientists and required two days to reach a consensus: a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions. The latter definition seems quite ample enough to accommodate most of the features challenging the classical gene concept, although it does not seem to accommodate well phenomena such as transplicing and gene rearrangements.

[43] Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S. and Snyder, M. What is a gene, post ENCODE? History and updated definition. *Genome Research* 17, 669-681, 2007.

[44] Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W. and Povey, S. Guidelines for human gene nomenclature. *Genomics* 79, 464-470, 2002.

All the aforementioned efforts to find a common norm to define the gene are undoubtedly worthwhile, but in some cases, perhaps overly elaborated. There have also been other straightforward propositions. For example, the highly respected textbook *Molecular Biology of the Cell,*[45] a bestseller among biology students throughout the world, defines a gene as a sequence that is transcribed as a single unit and encodes one set of closely related polypeptide chains. This definition has room for DNA segments that may give rise to various proteins due to alternative splicing, RNA editing, post-translational modifications, etc. Surprisingly, it doesn't make explicit the possibility that a gene may also generate a non protein coding RNA. Another definition, also with a molecular accent, is offered by Epp:[46] a gene is the nucleotide sequence that stores the information which specifies the order of the monomers in a final functional polypeptide or RNA molecule, or set of closely related isoforms. Epp stresses that regulatory sequences should not be considered part of a gene because there are too many types of them, they generally operate in complex combinations and often they influence the expression of several DNA segments. Besides, according to Epp, genes do not have to be expressed to be present.

Defining a gene remains an enduring endeavor

As it can be deduced from the aforementioned propositions for defining a gene, they are centered either in structure or in function. Interestingly, closely related definitions underlining the molecular quality of the gene, such as those offered by Gerstein *et al.*, the textbook *Molecular Biology of the Cell* and Epp, seem to be the most commonly accepted in the community of biological scientists. This is not only an impression based on subjective experience, but there are empirical signs that this is actually the case. For example, a few years ago, Stotz *et al.*[47] conducted a survey among Australian biologists from different areas (medicine, pharmacy, veterinary science, biochemistry, etc) to find out how they conceptualized the gene. Several questions were asked regarding the gene concept itself and the application of the gene concept to specific cases. The great majority of the responses ob-

---

[45] *Molecular Biology of the Cell*, 5th ed. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (eds) Garland Science, NY, 2008.

[46] Epp, C.D. Definition of a gene. *Nature* 389, 537, 1997.

[47] Stotz, K., Griffiths, P.E. and Knight, R. How biologists conceptualize genes: an empirical study. *Stud. Hist. Philos. & Biomed. Sci.* 35, 647-673, 2004.

tained favored the classical molecular concept, which is based more in structure than in function.

This way of conceptualization may reflect that in the era of genomics it is essential to be able to identify genes during the annotation of a newly sequenced genome. In addition, genetic engineering with academic or industrial purposes also requires a clear identification of the DNA segment that needs to be transferred in order to confer the desired phenotype to the recipient organism. These definitions assume the common criterion of leaving aside the concept of regulatory sequences controlling gene expression. However, they do not seem to encompass the phenomena of overlapping sequences, transplicing and RNA editing. Neither does Waters' more unified gene concept as a linear sequence in a product at some stage of genetic expression. The more comprehensive definitions to date appear to be those offered by ENCODE project and the Sequence Ontology Consortium, with the corresponding limitations already mentioned. Perhaps these could be overcome with a proper combination of both definitions. Thus, a gene could be defined as 'a union of genomic sequences encoding a coherent set of potentially overlapping functional products, these sequences being associated with regulatory regions, transcribed regions and/or other functional sequence regions'. However, this definition seems too intricate for everyday use. It is likely that once we get accustomed to the idea that a gene may comprise several segments dispersed throughout the genome and that it may also produce multiple transcripts that affect the same function, a definition such as the latter will prevail. In the meantime, other novel approaches may be worth considering. For example, taking into account the complex transcriptional organization of the genome, Gingeras contends that a simple operational unit linking a specific DNA sequence to phenotype/function is required.[48] According to this author, RNA transcripts are such fundamental operational units. Thus, each transcript could be catalogued according to the function it affects.

Raphael Falk, who has greatly contributed novel thoughts in the field, thinks that to arrive at a structural definition of the gene is a fruitless undertaking.[49] It may be even more difficult to merge the structural and func-

[48] Gingeras, T.R. Origin of phenotypes: Genes and transcripts. *Genome Res*. 17, 669-681, 2007.

[49] Falk, R. The gene – A concept in tension, in *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, pp. 317-348, 2000.

tional aspects in a single definition. Obviously, this state of affairs has not stopped scientists and philosophers to confront this task, simply because the gene concept represents a central issue in the biological sciences. Somewhere in the way, investigators have advanced reasons to declare the concept of the gene dead, to be thereafter refuted with arguments showing just the opposite.[50] Fortunately, finding a univocal definition of the gene persists as an ongoing intellectual challenge, because it gives us the opportunity to witness a fascinating display of thoughts and ideas at the boundary of knowledge. In the meantime, experimental molecular geneticists will continue to progress in the understanding of genome structure and expression. This situation evokes that of biology itself, in whose various branches scientists have been able to make paramount advances in spite of lacking a formal definition of living beings.

[50] Hall, B.K. The gene is not dead, merely orphaned and seeking a home. *Evol.&Dev.* 3:4, 225-228, 2001.