Medical Genomics

Primer on Medical Genomics
Part VII: The Evolving Concept of the Gene

ERIC D. WIEBEN, PHD

The draft sequence of the human genome was reported 2 years ago, and the task of filling gaps and polishing the sequence is nearing completion. However, despite this remarkable achievement, there is still no definitive assessment of the number of genes contained in the human genome. In part, this uncertainty reflects our growing understanding of the complexity and diversity of gene structure. Examples of complex gene structure are considered in the context of a discussion about the evolution of our understanding of gene structure and function.

***Mayo Clin Proc*. 2003;78:580-587**

ARF = alternative reading frame; hnRNA = heterogeneous nuclear RNA; INK = inhibitor of cyclin dependent kinase; mRNA = messenger RNA; snoRNA = small nucleolar RNA

What is a gene?

This is the first question posed to medical students on the first day of classes at Mayo Medical School, Rochester, Minn. That timing is not accidental because an understanding of gene structure and function is becoming a fundamental component of training in modern medicine. The answers from students tend to be in 3 categories—definitions that focus on phenotype ("something that determines a particular trait"), those that focus on structure ("a segment of DNA that codes for a protein"), and those that focus on heritability ("a unit of inheritance"). All are correct, and all are lacking.

Almost the entire human genome sequence is now in "finished" form (with an accuracy of fewer than 1 error per 10,000 bases of DNA); however, bets are still being taken (literally—see Gene Sweepstake[1]) about how many genes it takes to make a human. If the sequence is known, why is there difficulty in determining the final number of genes?

Although we now have a solid understanding of the order of nucleotides across most areas of the genome, our increasing knowledge of gene structure and function has revealed new complexities that complicate our attempts to describe succinctly the structure and function of human genes.

In practice, experts do little better than beginning medical students in trying to arrive at a universal definition of a gene.

From the Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, Minn. Dr Wieben is a member of the Mayo Clinic Genomics Education Steering Committee.

Individual reprints of this article are not available. The entire Primer on Medical Genomics will be available for purchase from the Proceedings Editorial Office at a later date.

The Human Gene Nomenclature Committee focuses on phenotype and defines a gene as "a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription or homology."[2]

The Gene Sweepstake Web site concentrates more on the expression pattern of a DNA sequence and defines a gene as follows:

> A gene is a set of connected transcripts. A transcript is a set of exons via transcription followed (optionally) by pre-mRNA [messenger RNA] splicing. Two transcripts are connected if they share at least part of one exon in the genomic coordinates. At least one transcript must be expressed outside of the nucleus and one transcript must encode a protein.[1]

An article in a Web issue of *ISUMA: Canadian Journal of Policy Research* presents a thoughtful analysis of the issues in defining a gene and concludes that "recent studies have made it impossible to give a functional definition of the gene."[3]

Thus, it is instructive to review how our understanding of genes has changed over time.

EARLY CONCEPTS

Aristotle[4] noted that "children often inherit anything that is peculiar in their parents," but an understanding of how parents can pass traits on to their offspring is relatively recent. The careful work of Mendel[5] on peas was not noticed or appreciated at the time it was first reported in 1865; however, it was rediscovered at the turn of the century, and the central conclusions were applied to several other early studies in what William Bateson termed *genetics*. In the preface to a translation of some of Mendel's work in 1902, Bateson summed up what was

then known about the physical basis of heredity in this memorable passage:

> We have no glimmering of an idea as to what constitutes the essential process by which the likeness of the parent is transmitted to the offspring.…The process is as utterly mysterious to us as a flash of lightning is to a savage. We do not know what is the essential agent in the transmission of parental characters, not even whether it is a material agent or not. Not only is our ignorance complete, but no one has the remotest idea how to set to work on that part of the problem.[6]

Bateson's enthusiasm for Mendel's conclusions on the "science of heredity" led him to propose further studies to define the "precise definition of their scope and limitations." Despite increasing interest in these issues at that time, Bateson's grant application to the Carnegie Institution of Washington was not funded.[7] Perhaps he had been a little too convincing about the state of ignorance in the field at that time.

The word *gene* is attributed to Johannsen, who introduced the term in 1909 with an imprecision that continues to this day. According to Shull,[8] Johannsen used the term *gene* "to denote an internal something or condition upon whose presence an elementary morphological or physiological characteristic depends." Thus, the first definition of the gene focused on a cellular component that directs some aspect of phenotype. This concept is central to our understanding of genes and remains a key element of any definition of a gene.

There was little progress in determining the physical nature of the gene for several decades after the term was coined. Early experiments with *Drosophila* by Morgan[9] led to the conclusion that genes were actual physical entities that are linked together on chromosomes. An article in 1933 by Demerec[10] suggested that genes were "single complex organic molecules" and used thymus nucleic acid (DNA) as an illustrative example.

THE DNA ERA

The next real progress in the understanding of the physical nature of heredity did not occur until the 1940s, when work by Beadle and Tatum on the bread mold *Neurospora* led to what was later called the "one gene-one enzyme hypothesis."[11] In 1944, Avery et al[12] established that DNA alone was capable of passing on a heritable trait, the encapsulation of pneumococci. Nevertheless, 9 years passed before Watson and Crick[13] published the structure of DNA in 1953. The first experiments supporting the idea that information flows from DNA to protein via a triplet, nonoverlapping, degenerate genetic code were not published until 1961.[14] Twenty years after the experiments on pneumococci, Yanofsky et al[15] and Sarabhai et al[16] independently demonstrated the colinearity between gene structure and protein structure, cementing the concept of a gene as a linear segment of nucleotides in DNA that code for a particular protein.

GENES IN 1975

By the mid-1970s there was pronounced confidence that we had a mature understanding of gene structure and a solid understanding of the mechanics of gene expression. Estimates of the number of genes in humans based almost solely on genetic considerations suggested that there were between 4×10^4 and 10^5 genes in the genome, accounting for less than 5% of the total amount of DNA in the nucleus of a human cell. It was known that the proportion of the genome that was transcribed exceeded that which was represented in messenger RNA (mRNA), but there was no clear consensus about the identity or possible functions of the "extra" transcribed sequences that never made it to the final mRNA.[17] Remember, DNA cloning methods were still in their infancy and techniques for rapid and reliable DNA sequencing had not yet been developed. Thus, all the information on gene structure was indirect and had to be inferred by careful analysis of genetic experiments in model organisms. Further understanding of the structure of genomes came from studies that separated the 2 strands of DNA and then observed what happened when the strands were allowed to reanneal. Given these limitations, the prevailing thoughts about gene structure based on genetic analysis were surprisingly accurate.

The concept of a "transcription unit" at that time consisted of a single regulatory region adjacent to a region that codes for protein. The primary products of transcription (collectively referred to as heterogeneous nuclear RNA [hnRNA]) were thought to be longer than the final mRNAs in many cases, necessitating some sort of cleavage reaction to remove the extra sequences before a functional mRNA could be produced. Since both hnRNA and mRNA molecules were known to have poly(A) tails at their 3′ ends, the prevailing wisdom was that the coding portion of the transcript was localized to a contiguous stretch of nucleotides near the 3′ end of the transcribed sequence, while the 5′ end was simply degraded in the nucleus (Figure 1) shortly after transcription was completed. This model accounted for the observed higher turnover rates for the precursor hnRNAs and provided a logical path for information flow from DNA to mRNA,[17] but the models of gene structure and function at that time also acknowledged some uncertainties. It was known that not every mRNA became polyadenylated; thus, there was some thought that a given hnRNA could give rise to more than 1 mRNA. Additionally, it was not clear that every hnRNA transcript was capable of giving rise to functional mRNA. It was thought

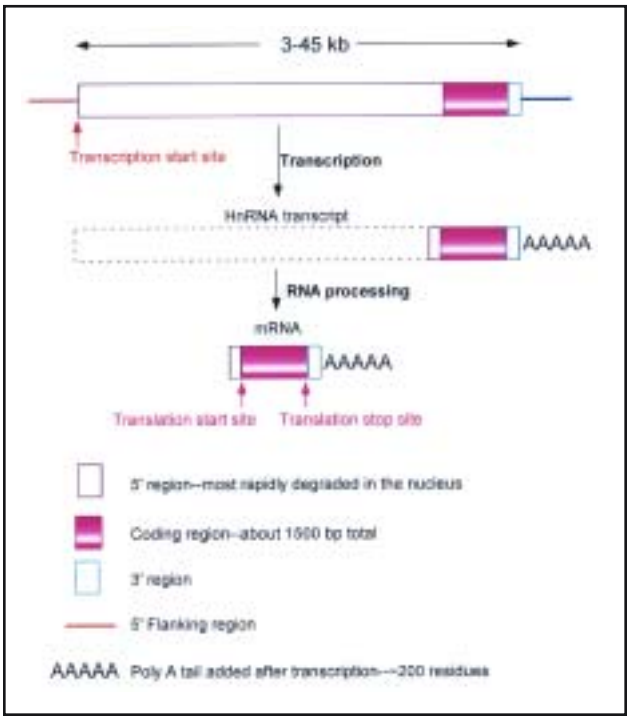

Figure 1. Gene structure and function as understood before the discovery of introns. By 1975, there was good reason to believe that the primary transcripts of genes were longer than the final messenger RNAs (mRNAs). Since poly(A) tails were known to be present on both heterogeneous nuclear RNA (HnRNA) transcripts and mRNA, and the poly(A) segments were localized at the 3′ end of transcripts, it was logical to hypothesize that extra 5′ untranslated sequences were removed during processing in the nucleus. bp = base pair; kb = kilobase.

that some of the observed instability of hnRNA might derive from a set of transcripts that were simply made and rapidly degraded.

DISCOVERY OF INTRONS

Two articles published nearly simultaneously in 1977 radically changed our understanding of gene structure.[18,19] Direct electron microscope comparisons of the structures of the adenovirus DNA genome to the mRNAs made from the viral genes revealed that the gene sequences were interrupted by DNA that was not represented in the mRNA. The genes were split into several pieces. The immediate implication of this finding was that RNAs made from the gene would somehow need to be cut and rejoined (spliced) to produce functional mRNA. Not long after this finding, the scientific community realized that split genes were the norm rather than the exception in eukaryotic organisms (Figure 2).

The realization that most mRNAs need to be stitched together from segments of longer precursor RNAs started an immediate effort to define the cellular machinery responsible for this exacting task. Were there specific sequences in all transcripts that marked intron-exon junctions, or was intron excision directed by higher order structures of the transcripts? This task was facilitated by the nearly simultaneous development of rapid and efficient techniques for DNA sequencing, which allowed gene structures to be determined routinely at the nucleotide level for the first time. As more genes were sequenced, it became clear that there were short consensus sequences that surrounded intron-exon junctions (Figure 3). The first 2 nucleotides of almost all intron sequences in the DNA template were GT, the last 2 nucleotides were AG (the GT-AG rule), and lower levels of conservation were found at other regions near intron-exon borders.[22] Much of the specificity for the splicing process derives from the recognition of these short sequences by small nuclear ribonucleoprotein complexes and specific protein splicing factors. Specific small nuclear ribonucleoprotein complexes contain small RNAs that base pair with particular sequences in the precursor RNA, directing the precise excision of introns shortly after synthesis of the precursor in the cell nucleus.[23] (Emphasizing the importance of small nuclear RNAs in this process, recent work has identified a minor class of small nuclear RNAs that direct the excision of the small minority of introns that do not conform to the GT-AG rule.[24])

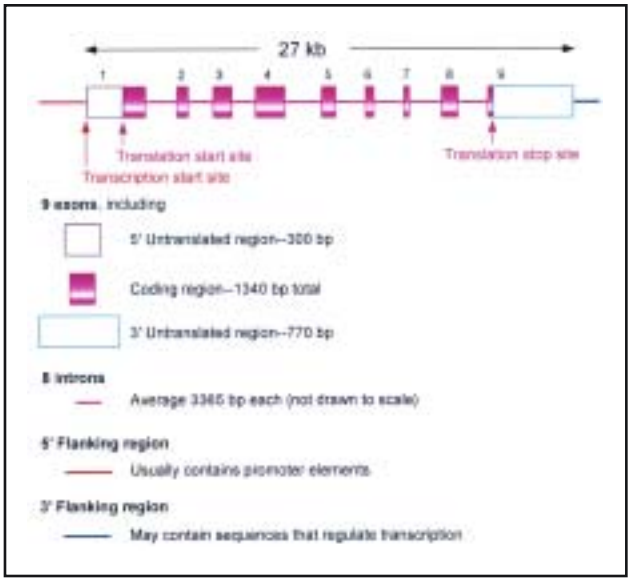

Figure 2. Structure of a “typical” gene in the era of the sequenced genome that conforms to the mean statistics for human genes as described at the time of the publication of the draft human genome sequence.[20] bp = base pair; kb = kilobase.

Why are genes divided into multiple segments? As DNA sequencing became routine and more genes and gene transcripts were sequenced, the extent to which splicing could influence gene expression began to be more completely appreciated. It became clear that the one gene-one enzyme idea was a vast oversimplification because the splicing process allowed a single gene to give rise to an entire family of related transcripts. In the simplest case, a cell can make a choice to include or exclude a particular exon from a given mRNA; thus, the resulting protein will either include or not include the amino acids coded by that exon. In other cases, genes are structured to create a mutually exclusive spectrum of exon choices that may be inserted into an mRNA. The degree of variability that splicing can bring to gene expression is not trivial. The *Drosophila DSCAM* (Down syndrome cell adhesive molecule) gene has 48 different alternatives for one exon, and an astounding 38,016 variant transcripts can be made from this one "gene" that is only 61.2 kilobase in length (Figure 4).[2]

The implications from pre-mRNA splicing extend well beyond the ability to create families of related proteins from relatively compact segments of DNA, and they further complicate efforts to develop a universal definition of a gene. Several human genes are structured in such a way that they have multiple choices for a first exon (not just different choices for internal exons). This variation on a theme creates significant possibilities for making the expression of that gene responsive to differing external stimuli. For example, the human dystrophin gene (which spans more than 2 million base pairs) has at least 7 different promoters, each of which directs expression of the gene in a particular cell type.[26] Interestingly, 4 of these promoters are located in the middle of the gene (upstream of exons 30, 45, 56, and 63), and 1 directs the synthesis of a protein that is less than one third the size of the longest dystrophin. Should these be considered separate proteins from separate genes?

SPECIAL CASES

Although most genes that have been studied are transcribed to RNAs that are processed and translated in a fairly straightforward manner, numerous genes have a more complex structure involving alternative splicing or alternative promoters. However, even the more extreme examples of alternative exon usage previously cited are straightforward compared with some of the truly exotic gene arrangements found in nature.

In trypanosomes and some other microorganisms, "trans-splicing" can occur between the products of 2 different transcription units (2 different "genes"?) to create a novel mRNA that combines the sequences of both transcription units (Figure 5).[27] Should this be interpreted as 1 gene with 2 distinct transcription units or 2 separate genes combining to make a single product?

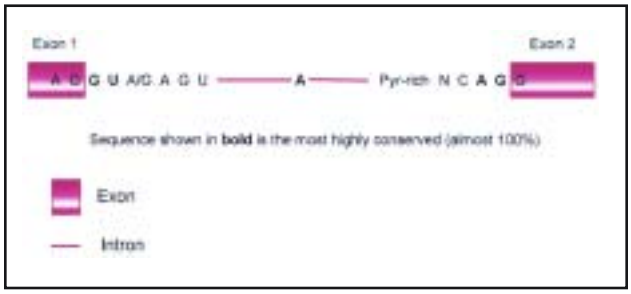

Figure 3. Conserved sequences at splice junctions. Only the sequences at both ends of the intron (GT at the 5′ end and AG at the 3′ end [bold]) are found in almost all human introns. The other residues shown are the most common at each position relative to intron-exon junctions, but there is considerably more variability at other positions.[21]

For genes that undergo RNA editing, nucleotides are added to or subtracted from the middle of an initial transcript under the direction of so-called guide RNAs[28] (Figure 6). These guide RNAs are the products of entirely separate segments of DNA located in other parts of the genome, and yet they are absolutely required for the production of a single functional mRNA. In one of the more radical cases (cytochrome oxidase III in trypanosome mitochondria), more than half of the nucleotides in the final functional version of the mRNA are added posttranscriptionally under the direction of a series of guide RNAs.[30] In this setting, can the segment of DNA that contributes less than half of the mRNA sequence really be called "the gene" for cytochrome oxidase III?

Not all the special cases are limited to microorganisms. Despite the fact that exons for protein-coding genes make

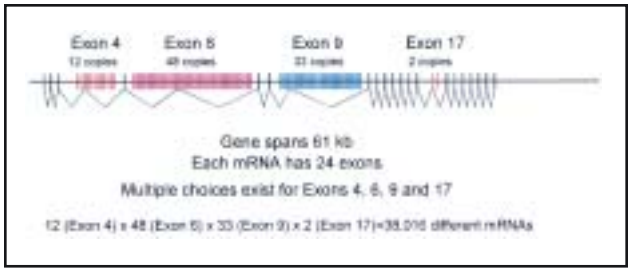

Figure 4. Possibilities for alternative splicing in the *Drosophila DSCAM* (Down syndrome cell adhesion molecule) gene. Exons 4, 6, 9, and 17 of this gene exist in multiple, similar (but not identical) copies. During RNA processing, only 1 copy of each of these exons is included in each messenger RNA (mRNA). This creates the theoretical possibility of producing 38,016 distinct mRNAs from this single gene.[25] kb = kilobase.

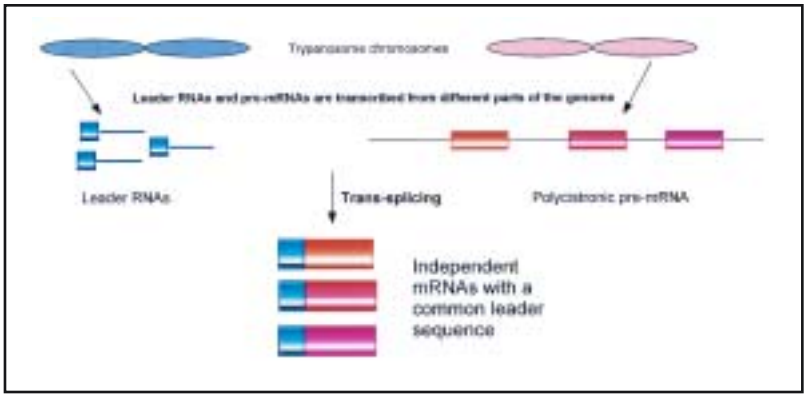

Figure 5. Mechanism of trans-splicing in trypanosomes. A common 5′ leader sequence (blue solid) is added to separate coding exons (shown as colored solids), which can be synthesized as part of a single polycistronic transcript. Thus, the final functional messenger RNAs (mRNAs) are hybrids of sequences from 2 independent transcription units.

up only about 1% to 2% of the human genome, there are actually several instances where genes overlap. In one of the most bizarre instances uncovered to date, 2 human genes with medical importance (called *p16INK4A* and *p14ARF* [INK = inhibitor of cyclin dependent kinase; ARF = alternative reading frame]) actually use common exons to code for completely different proteins.

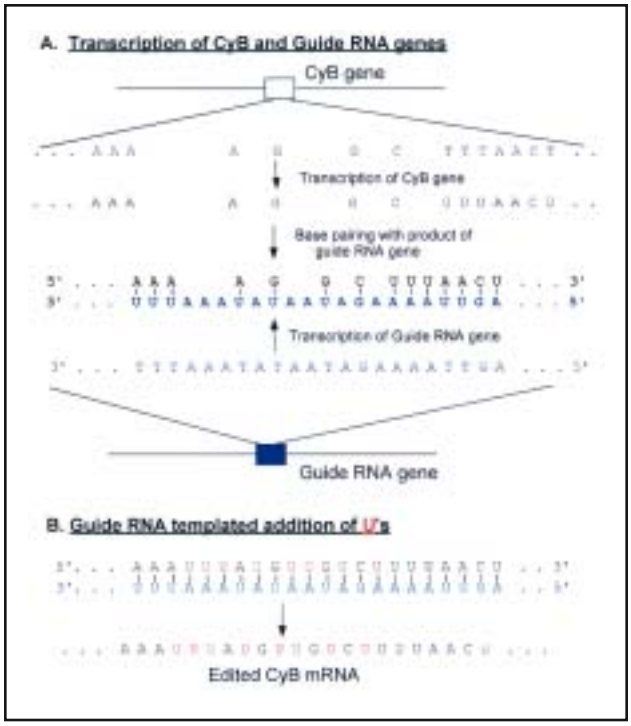

Figure 6. Editing of the cytochrome B (CyB) RNA from *Leishmania tarentolae* requires sequence information from a guide RNA that base pairs to part of the primary transcript. This guide RNA is transcribed from a separate gene (blue rectangle) and is used as a template for the addition of multiple uridines to the initial RNA transcript. The final edited RNA shown in B contains a number of uridine residues (red) that are not coded for in the *CyB* gene.[29]

One of the tumor suppressors implicated in some melanomas is p16INK4A, a small protein that inhibits the phosphorylation of the retinoblastoma gene product and therefore is a negative regulator of the cell cycle.[31] The 3-exon gene for p16INK4A has been characterized and maps to chromosome 9p21. Another tumor suppressor implicated in some melanomas is p14ARF, which promotes the activity of p53 by inhibiting the activity of mdm2.[32] Although the *p14ARF* gene also has 3 exons and maps to chromosome 9p21, the sequences of the p16INK4A and p14ARF gene products are completely unrelated at the protein level. However, analysis of mRNAs for these 2 tumor suppressors revealed that these 2 genes share common second and third exons (Figure 7).[33] The 2 separate "genes" really differ only in the sequence of their first exons, which code only for the first few amino acids of each protein.

How can 2 different proteins be synthesized from essentially the same mRNA sequence? In this case, the answer is that the common portions of the mRNA sequence are translated by the protein synthesis machinery in 2 different reading frames. In the normal process of protein synthesis, the reading frame that will be used to decode the information in the mRNA is established solely by the position of the initiating AUG codon. After the process is initiated, the ribosome decodes the mRNA sequence 3 nucleotides at a time, without examining the possible coding potential of ARFs. Since *p16INK4A* and *p14ARF* have separate and distinct first exons, the mRNAs made from these genes initiate protein synthesis in different reading frames, and

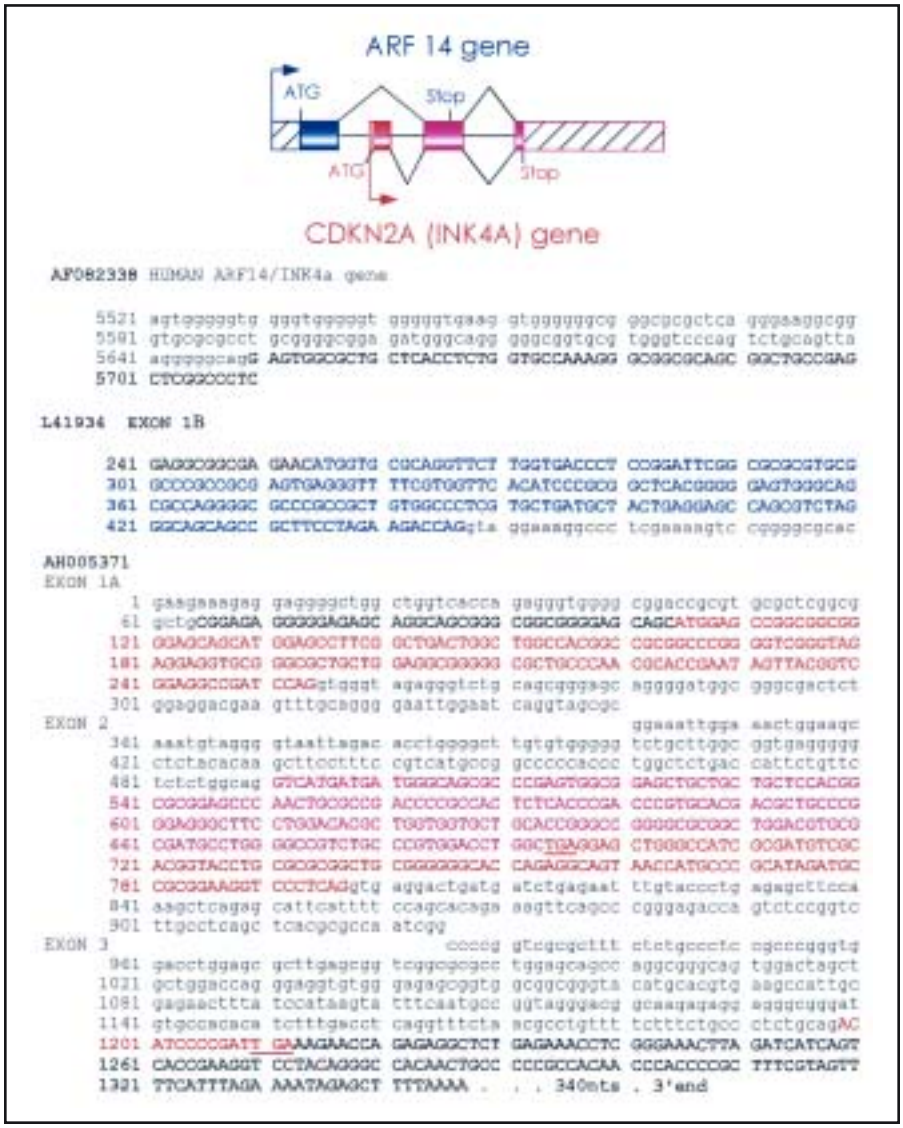

Figure 7. Structure and sequence of the composite *INK4A/ARF14* gene. Top, Sequences unique to the ARF (alternative reading frame) transcript are shown in blue. Sequences unique to the INK4A transcript are shown in red. Sequences that are common to both transcripts are shown in pink. Start sites for transcription are depicted with bent arrows, and start sites for translation are marked with translation initiation codons (ATG). The location of the respective translation stop codons is also marked (stop). Untranslated regions of the exons are crosshatched. Bottom, Sequences surrounding the exons of the *INK4A/ARF14* gene are shown, and GenBank accession numbers are given. Exon sequences are shown in all capital letters, whereas flanking sequences and introns are shown in lowercase letters. Coding regions unique to the ARF14 transcript are shown in blue, whereas coding regions unique to the INK4A transcript are shown in red. Coding sequences shared by both transcripts are shown in pink. Translation stop codons are underlined.

the remainder of the sequence from the common exons is decoded by the ribosome to give proteins with different amino acid sequences (Figure 8).

Note that this overlapping arrangement of exons creates the possibility that a single point mutation can change the structure of 2 proteins (this does occur). Nevertheless, there are point mutations in the common exons that are silent in the context of one gene but lead to amino acid changes in the other protein (for example a T→C mutation at codon 62 of the INK4A mRNA leads to substitution of proline for

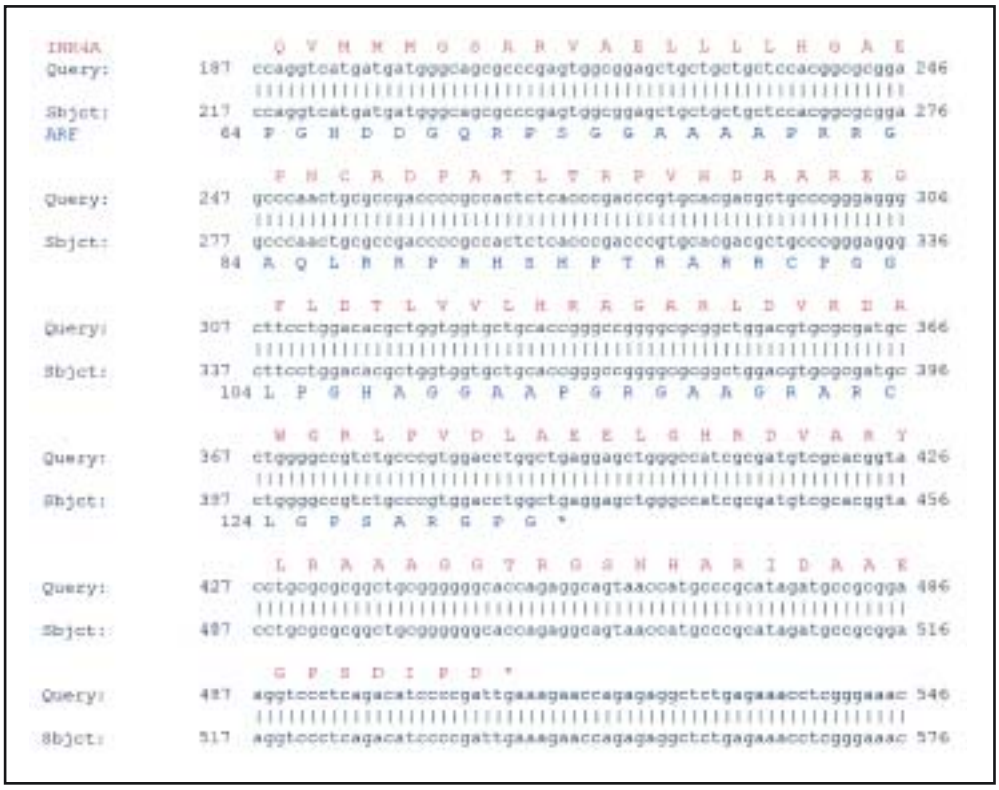

Figure 8. BLAST alignment of the INK4A/ARF14 messenger RNAs (mRNAs) showing differences in the protein sequence arising from translation in 2 different reading frames. RNA sequences of the p16INK4A mRNA (top, in black lower-case letters) and the p14ARF mRNAs (bottom, in black lower-case letters) are identical over the region shown. Protein sequences, shown in the single-letter amino acid code (p16INK4A in red, p14ARF in blue) are different because the mRNAs are translated in different reading frames. BLAST = Basic Local Alignment Search Tool; Sbjct = subject.

leucine, but this is a silent substitution in the context of p14ARF coding potential).[34]

Based on the definition of a gene by the Human Gene Nomenclature Committee, *INK* and *ARF* are clearly different genes because they make 2 different contributions to function—2 different proteins are produced. However, based on the definition of a gene by the Gene Sweepstake group, the "set of connected transcripts" that code for both proteins defines the structure of a single *INK/ARF* gene.[1] Clearly, having a molecular understanding of a gene does not always provide a simple answer to the semantic issues.

Other overlapping sets of genes illustrate further complications. In many organisms, some small nucleolar RNAs (snoRNAs) that function in ribosomal RNA processing reactions are cleaved from the introns of transcripts that code for otherwise unrelated proteins. For example, murine U14 snoRNA is encoded within an intron of the *hsc70* gene and relies on transcription of that gene for its synthesis.[35] In this case, one "set of connected transcripts" produced under the direction of a single regulatory region codes for both a protein product and a small RNA product, the U14 snoRNA. Can the U14 coding sequence be considered a free-standing gene when it does not have its own promoter? This theme can be carried to further extremes. Some transcripts actually encode multiple, different mammalian snoRNAs,[36] yielding several RNAs that function independently but all come from a single "set of connected transcripts."

WHAT IS A GENE?

Gelbart[37] considers some of the difficulties mentioned herein and concludes that "we may well have come to the point where the use of the term 'gene' is of limited value and might in fact be a hindrance to our understanding of the genome." Given the widespread use of the term and its demonstrated ability to evolve with changing times, that view is perhaps a bit unrealistic. However, it is important to recognize that our increasing understanding of the genome in molecular terms has broadened our thinking to the point in which the term *gene* is sometimes as ambiguous as it was

when it was first coined. From a structural standpoint, it makes sense to define a gene in terms of the product(s) it produces. A useful definition observing this spirit is found in the latest edition of the *Molecular Cell Biology* text by Lodish et al. A *gene* is a "physical and functional unit of heredity, which carries information from one generation to the next. In molecular terms, it is the entire DNA sequence—including exons, introns, and noncoding transcription control regions—necessary for production of an functional protein or RNA."[21]

REFERENCES

1. Gene Sweepstake. Wellcome Trust, Sanger Institute. Project Ensembl. Available at: http://www.ensembl.org/Genesweep/. Accessibility verified March 18, 2003.
2. Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. *Genomics.* 2002;79:464-470.
3. Morange M. Century of the gene. *ISUMA Can J Policy Res.* 2001;2:22-27. Available at: www.isuma.net/v02n03/morange/morange_e.pdf. Accessibility verified March 20, 2003.
4. Aristotle. *The History of Animals, 350 BC*. Thompson DW, trans. London, England: John Bell; 1907.
5. Mendel G. Experiments in plant hybridization. *Verhandl Naturforsch Vereines Brunn.* 1865;3-47.
6. Bateson W (1902). Application for support of an experimental investigation of Mendel's principles of heredity in animals and plants. In: Bateson B. *William Bateson, F.R.S.: His Essays & Addresses, Together With a Short Account of His Life.* Cambridge, England: Cambridge University Press; 1928.
7. Robbins RJ. In the forward to Bateson's application for support of an experimental investigation of Mendel's principles of heredity in animals and plants. In: Bateson B. *William Bateson, F.R.S.: His Essays & Addresses, Together With a Short Account of His Life.* Cambridge, England: Cambridge University Press; 1928:iii-iv.
8. Shull GH. Genetic definitions in the New Standard Dictionary. *Am Nat.* 1915;49:52-59.
9. Morgan TH. Sex limited inheritance in *Drosophila*. *Science.* 1910; 32:120-122.
10. Demerec M. What is a gene? *J Hered.* 1933;24:368-378.
11. Beadle GW, Tatum EL. Genetic control of biochemical reactions in Neurospora. *Proc Natl Acad Sci U S A.* 1941;27:499-506. In: Joklik WK, et al, eds. *Microbiology: A Centenary Perspective*. Washington, DC: ASM Press; 1999:308.
12. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *J Exp Med*. 1944;79:137-157. In: Joklik WK, et al, eds. *Microbiology: A Centenary Perspective*. Washington, DC: ASM Press; 1999:116.
13. Watson JD, Crick FHC. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature.* 1953;171:737-738.
14. Crick FHC, Barnett L, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. *Nature*. 1961;192:1227-1232. In: Joklik WK, et al, eds. *Microbiology: A Centenary Perspective*. Washington, DC: ASM Press; 1999:384.
15. Yanofsky C, Carlton BC, Guest JR, Helinski DR, Henning U. On the colinearity of gene structure and protein structure. *Proc Natl Acad Sci U S A.* 1964;51:266-272. In: Joklik WK, et al, eds. *Microbiology: A Centenary Perspective*. Washington, DC: ASM Press; 1999:392.
16. Sarabhai AS, Stretton AOW, Brenner S, Bolle A. Co-linearity of the gene with the polypeptide chain. *Nature.* 1964;201:13-17.
17. Lewin B. *Gene Expression.* Vol 2. New York, NY: John Wiley & Sons; 1974.
18. Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell.* 1977;12:1-8. In: Joklik WK, et al, eds. *Microbiology: A Centenary Perspective*. Washington, DC: ASM Press; 1999: 74.
19. Berget SM, Moore C, Sharpe PA. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A.* 1977;74:3171-3175. In: Joklik WK, et al, eds. *Microbiology: A Centenary Perspective*. Washington, DC: ASM Press; 1999:568.
20. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome [published corrections appear in *Nature*. 2001;411:720 and *Nature*. 2001;412:565]. *Nature.* 2001;409:860-921.
21. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell JE. *Molecular Cell Biology*. 4th ed. New York, NY: WH Freeman and Co; 2000.
22. Breathnach R, Chambon P. Organization and expression of eucaryotic split genes coding for proteins. *Annu Rev Biochem*. 1981;50:349-383.
23. Lerner MR, Steitz JA. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc Natl Acad Sci U S A.* 1979;76:5495-5499.
24. Hall SL, Padgett RA. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science.* 1996;271:1716-1718.
25. Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell.* 2000;103:367-370.
26. O'Brien KF, Kunkel LM. Dystrophin and muscular dystrophy: past, present, and future. *Mol Genet Metab*. 2001;74:75-88.
27. Bruzik JP, Van Doren K, Hirsh D, Steitz JA. Trans splicing involves a novel form of small nuclear ribonucleoprotein particles. *Nature.* 1988;335:559-562.
28. Blum B, Bakalara N, Simpson L. A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell.* 1990;60:189-198.
29. Lewin B. *Genes VII.* New York, NY: Oxford University Press; 2000.
30. Feagin JE, Abraham JM, Stuart K. Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. *Cell.* 1988; 53:413-422.
31. Wainwright B. Familial melanoma and p16—a hung jury. *Nat Genet.* 1994;8:3-5.
32. Zhang Y, Xiong Y. Mutations in human ARF exon 2 disrupt its nucleolar localization and impair its ability to block nuclear export of MDM2 and p53. *Mol Cell.* 1999;3:579-591.
33. Stone S, Jiang P, Dayananth P, et al. Complex structure and regulation of the P16 (MTS1) locus. *Cancer Res.* 1995;55:2988-2994.
34. Soufir N, Avril MF, Chompret A, et al, French Familial Melanoma Study Group. Prevalence of p16 and CDK4 germline mutations in 48 melanoma-prone families in France [published correction appears in *Hum Mol Genet*. 1998;7:941]. *Hum Mol Genet*. 1998;7: 209-216.
35. Xia L, Watkins NJ, Maxwell ES. Identification of specific nucleotide sequences and structural elements required for intronic U14 snoRNA processing. *RNA.* 1997;3:17-26.
36. Leader DJ, Clark GP, Watters J, Beven AF, Shaw PJ, Brown JW. Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs. *EMBO J.* 1997;16:5742-5751.
37. Gelbart WM. Databases in genomic research. *Science.* 1998;282: 659-661.

Primer on Medical Genomics Part VIII will appear in the July issue.