REPRESENTING GENOMIC MAPS IN A RELATIONAL DATABASE[1]

ROBERT J. ROBBINS[2]

Johns Hopkins University

rrobbins@gdb.org

Original version published in S. Suhai (ed). 1994. *Computational Methods in Genome Research*. New York: Plenum Publishing. pp 85-96.

[1] Presentation made at the International Symposium on Computational Methods in Genome Research, held 1–4 July 1992 in Heidelberg, Germany.

[2] Current address: Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, WA. Email: rrobbins@fhcrc.org

TABLE OF CONTENTS

Original published in S. Suhai (ed). 1994. *Computational Methods in Genome Research*. New York: Plenum Publishing. pp 85-96.

REPRESENTING GENOMIC MAPS IN A RELATIONAL DATABASE[1]

ROBERT J. ROBBINS[2]

INTRODUCTION

The goals of the Human Genome Project are: (1) construction of a high-resolution genetic map of the human genome, (2) production of a variety of physical maps of all human chromosomes and of the DNA of selected model organisms, (3) determination of the complete sequence of human DNA and of the DNA of selected model organisms, (4) development of capabilities for collecting, storing, distributing, and analyzing the data produced, and (5) creation of appropriate technologies necessary to achieve these objectives.

Given the amount of data that will be generated as progress toward these goals is made, it is imperative that electronic means for storing and manipulating the data be available. Databases must be built to describe map objects and mapping reagents, and to accommodate genetic and physical genomic maps as they are produced.

As our understanding of the human genome grows, the concepts that must be represented in these databases will increase in complexity and subtlety. Since these databases are expected to become a new scientific literature through which electronic data publishing will occur (cf. Cinkosky et al., 1991; Courteau, 1991; Pearson and Söll, 1991), they must be designed to handle the changing concepts of "gene" and "genomic map" without requiring major redesign each time a new finding occurs. The data models used must be sufficiently complex and abstract to represent all of our present concepts of genes and maps, as well as to evolve gracefully with the findings on genomic anatomy.

---

[1] Presentation made at the International Symposium on Computational Methods in Genome Reseach, held 1–4 July 1992 in Heidelberg, Germany.

[2] Current address: Fred Hutchinson Cancer Research Center, 1100 Fairview Ave, N., Seattle, WA. Email: rrobbins@fhcrc.org

Although the word “gene” may be the most frequently used word in biology, it has proven remarkably difficult to define. Entire books have been written describing the early history of the gene concept (Carlson, 1966), and many eminent biologists addressed the question during the classical period of genetics (Demerec, 1933; Demerec, 1955; Muller, 1945; Stadler, 1954). In the modern era, major textbooks on molecular and cellular biology all devote significant efforts to defining the gene (e.g., Alberts, et al., 1983; Darnell, et al., 1986), with one recent work (Singer and Berg, 1992) simply claiming that no single definition of the gene exists :

> The unexpected features of eukaryotic genes have stimulated discussion about how a gene ... should be defined. Several different possible definitions are plausible, but no single one is entirely satisfactory or appropriate for every gene.

If genes cannot be defined, then how is one to design a data model to represent them? And, without a definition for genes, how possibly could we represent genomic maps? It is a truism in information science that an adequate data model cannot be developed without an understanding of the thing being modeled. Therefore, to build good databases we must turn our attention to the notion of the gene.

WHAT IS A GENE?

The gene, originally described as the hypothetical fundamental unit of heredity, is now known to consist of instructions encoded in the nucleotide sequence of a DNA molecule. To see why advances in our understanding of gene structure have complicated our notion of what a gene is, let us consider briefly the history of the gene concept.

Insights from Classical Genetics

The first notion of the gene came from Mendel’s breeding studies on peas, in which he showed that patterns of inheritance were consistent with the assumption that the traits of each individual were controlled by a pair of independent factors, with one received from each parent. After 1900, Mendel’s findings were rediscovered and extended, especially by the *Drosophila* group at Columbia under T. H. Morgan.

Very early, Morgan hypothesized and Sturtevant showed (Figure 1) that the inheritance patterns associated with genes believed to be carried in the same linkage group could be explained by assuming that linkage groups were chromosomes and that genes were carried on chromosomes in a regular, linear order.

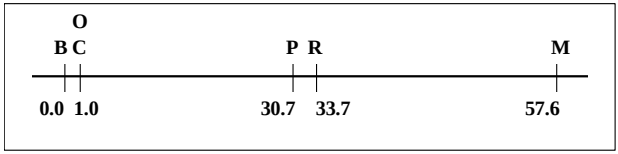

Figure 1. The first genomic map, as derived by Sturtevant (1913) from recombination data in Drosophila. “B” = yellow body, “C” = white eye, “O” = eosin eye, “P” = vermilion eye, “R” = rudimentary wing, “M” = miniature wing. The distances are given as percent recombinants.

When recombinational mapping proved generally applicable to all organisms, the resulting classical concept held that genes are (1) the fundamental unit of heredity, (2) subject to rare mutation, (3) stable across generations, (4) carried on chromosomes, and (5) capable of recombining during meiosis. According to these properties, two genetic traits that could not be separated by recombination were thought to involve mutations to the same gene, whereas traits that could be separated were held to involve two different genes.

The view of genes as essentially indivisible fundamental units occupying fixed chromosomal positions was summarized by Sturtevant and Beadle in 1939:

> The relative constancy of crossover values and the constant order of genes in chromosomes imply that every gene occupies a fixed position in a chromosome, and its allele a corresponding position in a homologous chromosome. ... Such a position is known as a *locus*. ... [By asserting] the linear arrangement of genes in chromosomes[, w]e do not, of course, imply by linear arrangement a straight line, but rather that the genes are arranged in a manner similar to beads strung on a loose string.

This “beads on a string” metaphor, so characteristic to the classical view of the gene, carried with it several corollary notions that have provided impediments to clear thinking about gene mapping, even to the present day. These include the idea of the indivisible gene and of the existence of a “string” — a scaffolding molecule that could provide a coordinate space on which genes might be placed independently of the presence or absence of other genes. The coordinate-space problem will be discussed in more detail later.

Modeling the Classical Gene in a Relational Database

The classical model of the genome lends itself to a very straightforward data model for representing genes and maps. Genes become nodes and pair-wise orderings and distances become arcs in a directed acyclic graph (DAG). DAGs are fine data structures for representing partial ordering among defined objects, and DAGs are relatively easy to implement in a relational database. One or more

entity tables can be designed so that each tuple contains the information describing an individual gene or other map object (a node). Tuples in a separate relationship table can represent arcs, by containing the identifiers denoting two map objects joined by an arc. Additional attributes of the arc, like measured distances) are easily added to the relationship table. Although no standard SQL commands currently exist for executing the necessary transitive closure queries over the DAG, methods for implementing such queries are well known.

The Discovery of Pseudoalleles Challenged the Classical View

The classical notion of the gene was shaken when the first instance of apparent intragenic recombination was observed. The process was said to define "pseudoalleles," since by definition, true alleles could not recombine (review by Carlson, 1959). In 1955, Bentley Glass wrote, "Fifty years from now it seems very likely that the most significant development of genetics in the current decade (1945-1955) will stand out as being the discovery of pseudoallelism." Although this claim now seems wildly inaccurate (especially since that is the same decade in which the structure of DNA was first established and bacterial genetics was founded), the strength of the sentiment shows just how firmly held was the notion of the indivisible gene.

The interest in pseudoalleles proved short-lived, as the DNA-sequence concept of the gene made intragenic recombination seem inescapable, not implausible. When Benzer's development of the cis-trans complementation test rendered obsolete the old recombinational test for gene individuality (Benzer, 1955), the storm over pseudoalleles faded away. The complementation test establishes whether two genetic traits involve the same or different genes by testing whether or not two chromosomes, each carrying a different defective gene, complement each other's deficit and restore normal function. If complementation occurs, the defects are held to occur in different functional units. If not, the defects are presumed to occur in the same functional unit. Benzer coined the term "cistron" to describe the functional units so identified.

The concept of the cis-trans test as defining genetic functional units became so widespread that many began to equate the cistron with the gene, as in the following textbook definitions:

> **Cistron** A nucleotide sequence in DNA specifying a single genetic function as defined by the complementation test; a nucleotide sequence coding for a single polypeptide; a gene. (Ayala and Kiger, 1984)

> **Cistron** Originally defined as a functional genetic unit within which two mutations cannot complement. Now equated with the term gene, as the region of DNA that encodes a single polypeptide (or functional RNA molecule such as tRNA or rRNA). (Suzuki, et al., 1986)

The Early Molecular Perspective

The molecular notion of the gene originated from biochemical studies, first on eye color in *Drosophila* and later on nutrient requirements in *Neurospora*, showing that individual genes seemed to be associated with the presence of individual functional enzymes. From this, the famous *one-gene, one-enzyme* hypothesis was proposed, then modified to *one-gene, one-polypeptide*.

With the discovery of the structure of DNA and the elucidation of the triplet code, the *one-gene, one-enzyme* concept was extended to *one-gene, one-macromolecule* and a gene became identified as that stretch of DNA responsible for determining a particular polypeptide or RNA. Such a definition is still commonly encountered today:

> **Gene** A hereditary unit that, in the classical sense, occupies a specific position (locus) with the genome or chromosome; a unit that has one or more specific effects upon the phenotype of the organism; a unit that can mutate to various allelic forms; a unit that codes for a single protein or functional RNA molecule. (Committee on Mapping and Sequencing the Human Genome, 1988)

The early molecular model of the gene can be summarized in either an operational, functional definition (the cistron detected with a complementation test) or a structural *one-gene, one-product* definition (the DNA sequence encoding a functional macromolecule).

Modeling the Early Molecular Gene in a Relational Database

In either of these molecular definitions, the gene is no longer indivisible, but it is still discrete and contiguous. And, it still has a well defined unitary function — the specification of its protein or RNA product. By extending the notion of the gene so that it has linear extent, the data model developed for the classical gene can easily be used to represent the early molecular concept of the gene.

Current Molecular Perspectives on the Gene

Continuing work on the regulation of gene expression began to show that regions of DNA not directly involved in specifying the sequence of a protein were nonetheless essential in determining its level of production. Although the discovery in the late 1970s that some genes contained introns upset the simplistic notion that genes and their proteins products were perfectly colinear, the early molecular model of the gene required only minor modifications to accommodate new findings so long as a continuous region of DNA produced one continuous polypeptide or RNA.

However, additional discoveries have begun to disturb the early molecular model of the gene just as surely as the discovery of pseudoalleles refuted the

classical definition. Complex regulatory units, with very subtle regulation has been described. Some regulatory units, such as the SOS and heat-shock regulons in *E. coli*, were found to control the expression of non-contiguous genes.

The discovery of overlapping coding regions has undercut (and the finding of nested genes within the introns of other genes destroyed) the idea of the gene as a necessarily discrete unit. Even more complex structures, such as nested gene families call into question the definitiveness of the complementation test. The occurrence of complex post-translational processing (whereby one transcript yields one polypeptide, which is then processed into multiple different functional peptides) leads to counter-intuitive results when combined with some definitions of the gene. Most challenging of all, the discovery of RNA editing (where the transcript of one DNA segment is enzymatically modified, under the direction of an RNA transcript from another piece of DNA, to yield a functional mRNA) undermines the notion of the gene as a contiguous region of DNA.

Complex Regulation.

The macromolecular synthesis (MMS) operon (Figure 2) in *E. coli* could also be called the "fundamental dogma" operon, since its three protein products are involved in DNA replication, transcription, and translation. Given the divergent times at which these processes occur, it is difficult to imagine how all three proteins could be effectively regulated in a single simple operon.

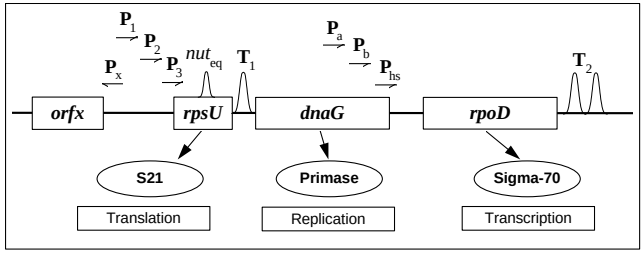

Figure 2. The complex macromolecular synthesis operon in *E. coli*, as described by Lupski and Godson (1989).

However, in addition to normal operon control, the MMS operon contains a maze of complex, overlapping control mechanisms. The operon has six promoters (seven, if the P_x promoter for "orf_x," an open reading frame of unknown function is included). The "P_1," "P_2," and "P_3," promoters control transcription initiation for the operon as a whole. Two other promoters, "P_a" and "P_b", also affect the *rpoD* locus, and these additional promoters are embedded in the coding region of the *dnaG* locus. Another promoter, "P_{hs}", is a heat-shock

promoter. "T_2" indicates the main transcription termination signal, and "T_1" indicates an alternate terminator.

With all of these overlapping control and coding regions, the definition of precise boundaries and extents for the genes in this operon are subject to reasonable debate among competent biologists.

Alternate Splicing and Nested Genes

The discovery in the early 1980s that some regions of DNA can produce more than one polypeptide through the alternate splicing of its transcript product began to undermine the generality of the *one-gene, one-product* notion of the gene. The discovery of genes nested within the introns of other genes further eroded the concept of genes as continuous, discrete regions of DNA.

The *Gart/Lcp* loci in *Drosophila* (Figure 3) illustrate both of these situations. The transcript of the *Gart* locus can undergo alternate splicing to yield two different gene products, and the *Lcp* locus resides entirely within the large intron of the common region of the *Gart* splice options. Given that the two loci are encoded on opposite strands of the DNA, it is arguable whether it is better to consider these two loci as being nested, or whether it is better to consider the outer locus to be discontinuous. One might even contend that, in general, loci should be assigned to regions of a particular strand of DNA, in which case the *Gart/Lcp* locus is unexceptional.

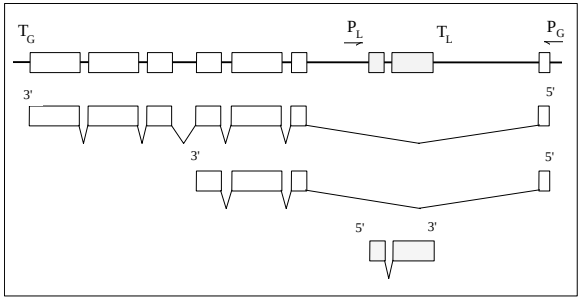

Figure 3. The exon map of the *Gart* locus in *Drosophila*, as described by Henikoff et al. (1986). Although the larval cuticle protein gene (*Lcp*) is fully contained within an intron of the *Gart* locus, it gene is coded on the opposite DNA strand. *Lcp* contains an intron of its own. P_G and T_G are the promoter and terminator for *Gart*, P_L and T_L are those for *Lcp*.

Nested Gene Families

A recent study on the human UDP-glucuronosyltransferase locus on chromosome 2 has found a complex region, with six promoters, six duplicated

(and diverged) alternate first exons, and four common trailing exons (Figure 4). Since each alternate first exon has its own promoter, one might suggest that this is not a case of alternate RNA splicing, but rather a case of a nested gene family, where only part of the gene has been duplicated, but where divergent evolution has occurred nonetheless.

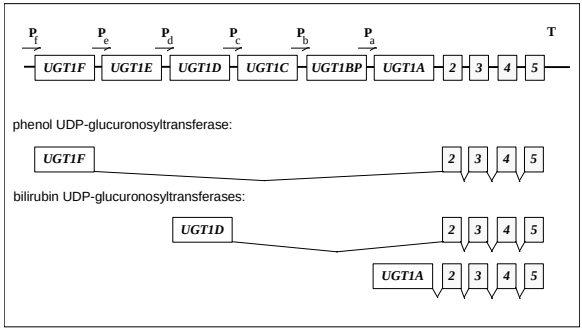

Figure 4. The complex *UGT1* locus of humans (Ritter et al., 1992) contains six promoter sites and one terminator. Each promoter is associated with a separate first exon that is spliced with exons 2–4 to make final mRNAs for translation. The exon labeled *UGT1BP* carries a frame-shift mutation that produces a premature stop codon and is considered a pseudogene.

This nested structure challenges the cistron notion of the gene, since mutations in different alternate first exons should complement, whereas mutations in the other four exons would not. Under either the cistron definition of the gene or the *one-gene, one-product* definition, this region must be considered to be five genes and a pseudogene.

Complex Post-translational Processing

The human POMC locus produces one large polypeptide from its mRNA. This polypeptide is then processed differentially in different tissue types to give a variety of neuropeptide and hormonal products. In the anterior lobe of the pituitary, the protein is cleaved once, cutting ß-lipotropin free from the C-terminal end. The remaining fragment is cleaved again, releasing corticotropin (ACTH) and CLIP (corticotropin like intermediate-lobe peptide). In the intermediate lobe, corticotropin is cleaved again, releasing -melanocyte stimulating hormone (-MSH). The ß-lipotropin is also cleaved, yielding ß-endorphin. Additional processing yielding additional products also occurs.

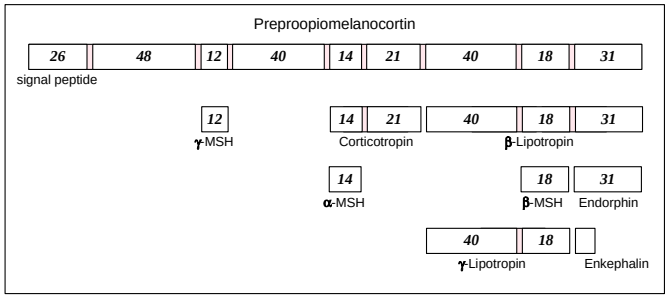

Figure 5. The human POMC locus produces a polypeptide that is processed differentially in different tissues to yield many different functional peptides.

How many genes are involved here? Many geneticists seem to prefer the notion that the POMC locus is a single gene. However, under the *one-gene, one-product* definition, this must be considered multiple overlapping loci that happen to share the same transcriptional apparatus. Although some geneticists might argue that all this is splitting hairs and that the POMC peptide is the product of one and only one gene, others disagree. Alberts et al. (1983) note:

> But it is now known that some DNA sequences ... participate in the production of at least two different mRNA molecules and therefore at least two different proteins with distinct biological roles. How then is a gene to be defined? At present, it seem best to retain the one-gene-one-polypeptide-chain definition. This means that in those cases where more than one polypeptide is specified by the same DNA sequence, two or more genes are considered to overlap on the chromosome.

By this definition, the POMC region must be considered at least eight overlapping loci.

Guide RNAs

The most intriguing recent discovery is that some primary transcripts require editing before they become fully functional mRNA molecules. In the mitochondria of trypanosomes, a primary transcript from one transcription unit is modified enzymatically, under the control of guide information contained in another RNA molecule that has been transcribed from a different region of the genome. The resulting mRNA contains information from both original transcripts in a merged form that can then be translated to yield a functional protein. Here the *one-gene, one-product* concept requires that the two separate transcription units be considered one gene.

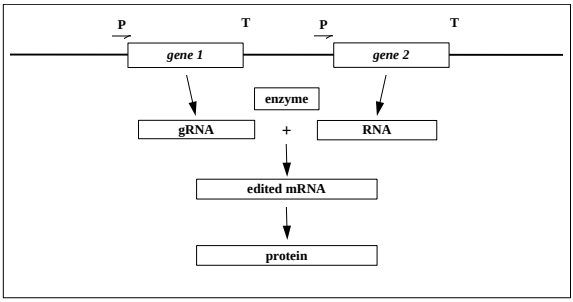

Figure 6. In many mitochondrial systems, it has been demonstrated that RNA editing, under the control of information stored in another RNA molecule (gRNA), is required for the production of a functional mRNA. In this case a single mRNA and a single resulting polypeptide can truly be said to derive from two transcriptional units.

VARIATION IS THE KEY

The study of biology must be the study of variation. Despite the fact that much has been written about sequencing *the* human genome and about obtaining *the* human map, it is generally recognized that considerable genomic variation exists from individual to individual, with estimates that place the amount of sequence difference between individuals at about one nucleotide in three hundred being widely quoted. In addition to base-substitution differences, there are also significant differences in the size of chromosomes, and thus in the actual amount of DNA present. Some have estimated, for example, that human chromosome 1 may show up to 10% differences in length among normal individuals. With that 10% amounting to 30 million bases pairs, the size variance of just one human chromosome may equal ten entire *E. coli* genomes.

With this kind of variation occurring, it is difficult to see how one might really conceive of *the* human sequence and *the* human map, much less represent it as a singularity in a database. Other aspects of normal variation, such as multi-copy genes, also pose challenges for database design.

Multi-copy Genes

By their nature, databases keep track of information relating to individual objects of interest. Databases are not really appropriate for characterizing arbitrary collections of things without individual distinction. Thus, if all genes have equivalent status and if the ultimate genome database is to represent all the genes in the human genome, it will be necessary to identify, name, and describe

all the genes. This could be complicated by the occurrence of genes that have clearly recognized function, reasonably straightforward structure, but which occur in the genome as variable-number, multiple copies.

Consider a simple case in *E. coli* where two identical tRNA genes occur as a tandem repeat. It would be easy to give both genes individual names, according to their relative position on the chromosome: $tRNA_1$ and $tRNA_2$. But what if we must represent the genome of a strain of *E. coli* that has lost one of these genes? Which locus are we to say is missing and which one remains, given that there is now only one present and that the two are not otherwise distinguishable, except by their positions relative to each other? Although this may seem a forced example, what if there were hundreds of copies at different locations in the genome? Such is the case with human rRNA loci.

Human cells contain about 200 rRNA gene copies per haploid genome, distributed in clusters on the short arms of the acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22). Should each of these 200 or so copies be considered an individual gene in its own right? If each copy is a gene in its own right, do all 200 get separate names? The length of the rRNA-bearing chromosomal arms vary significantly among individuals, and thus so presumably do the number of copies of genes. If each gene gets its own name, exactly which of these several hundred named genes does a particular individual carry? And, exactly how many should be placed on *the* human map?

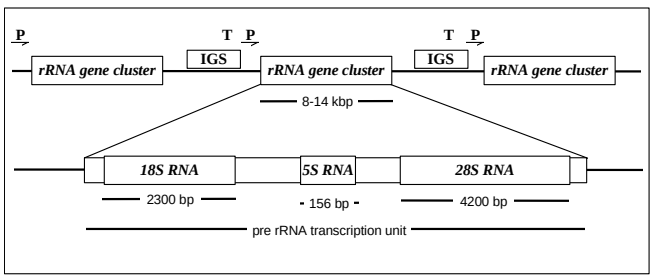

Figure 7. Human rRNA gene clusters occur as variable numbers of tandem repeats on the short arms of the acrocentric chromosomes. Each gene cluster is a transcriptional unit whose RNA transcript is ultimately processed to yield three ribosomal RNAs. “P” indicates a promoter, “T” a terminator site. “IGS” labels an intergenic spacer region. Overall, a haploid human genome carries about 200 copies of the rRNA gene cluster.

All of these rRNA genes produce equivalent RNA transcripts of about 13,000 nucleotides in length. Each primary transcript is processed in the nucleus to yield one copy each of three different ribosomal rRNAs: 28S rRNA, 18S rRNA, and 5.8S rRNA (Figure 7). According to the *one-gene, one-product*

definition of the gene (Alberts et al.), each of these 200 identical transcriptional units must really be considered to be three separate genes. Some believe that the primary rRNA gene transcript yields additional fragments that play a brief functional role during ribosome assembly. If that is the case, then each transcription unit contains more than three genes.

Other Repetitive Elements

Variable Nucleotide Tandem Repeat (VNTR) polymorphisms (Figure 8) have proven to be very useful as mapping reagents. They have also proven to constitute a serious challenge to the notion of a coordinate space on which genes can be placed in the absence of knowledge of what other alleles are present.

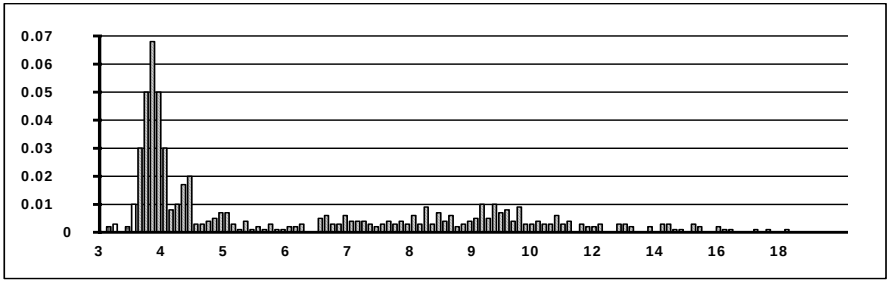

Figure 8. Frequency distribution of *Pst*I-DNA fragments in random individuals from a Caucasian population (Balazs, et al., 1992). Variable nucleotide tandem repeat (VNTR) loci occur in a highly polymorphic form. D14S1 represents the first VNTR studied.

It is a given in gene mapping that what one is mapping is the locus at which the alleles of a gene occur, not the particular alleles themselves. It has also been stated many times that "The ultimate, highest resolution map of the human genome is the nucleotide sequence, in which the identity and location of each of 3 billion nucleotide pairs is known." (Committee on Mapping and Sequencing the Human Genome, 1988) This implies that each gene, each functional segment along the DNA, can be identified by the actual address numbers of its first base pair and of its last base pair in the human sequence.

If VNTR alleles can vary in size by thousands of base pairs, and if the most common allele in a population may occur in only a few percent of individuals, then how could we meaningfully assign base-pair numbers as addresses to the *loci* of VNTR genes, or even to any genes on the other side of a VNTR locus? It cannot be done. There is no coordinate space on which we can pin the location of genes independent of the location of other genes. We must recognize that all genomic mapping must be as offsets relative to other genes known to be present.

We must also recognize that errors and uncertainties in the these base-pair offsets will increase as a function of measured distance, so that base-pair-level resolution can only have meaning over relatively short distances.

GENOMIC MAPS OR GENOMIC ANATOMIES?

McKusick (1988) has suggested that perhaps we should think in terms of anatomies rather than maps: "The anatomic metaphor is appropriate since the linear arrangement of genes in our chromosomes is part of our anatomy. It is also useful for a logical discussion of the significance of the information: the morbid anatomy, the comparative anatomy and evolution, the functional anatomy, the developmental anatomy and the applied anatomy of the human genome."

The anatomy metaphor is also desirable for a more fundamental reason: it is simply better and leads to clearer thinking. A problem with the map analogy is shown in the following story from Richard Feynman's delightful memoir (Feynman and Leighton, 1985):

> After that I went around to the biology table at dinner time. I had always had some interest in biology, and the guys talked about some very interesting things. Some of them invited me to come to a course they were going to have. ... I had to report on papers along with everyone else [and one of the papers] selected for me ... kept talking about extensors and flexors, the gastrocnemius muscle, and so on. This and that muscle were named, but I hadn't the foggiest idea of where they were located in relation to the nerves or to the cat. So I went to the librarian in the biology section and asked her if she could find me a map of the cat. "A *map* of the *cat*, sir?" she asked, horrified. ... From then on there were rumors about some dumb biology graduate student who was looking for a "map of the cat."

Of course Feynman should have asked for an anatomy of the cat, not a map. But why is his mistake funny? Why is it so obviously crazy to ask for a map of a cat? The *Oxford English Dictionary* defines "map" as:

> **map** *n* A representation of the earth's surface or a part of it, its physical and political features, etc., or of the heavens, delineated on a flat surface of paper or other material, each point in the drawing corresponding to a geographical or celestial position according to a definite scale or projection.

"Mapping," in the mathematical sense, the *OED* defines as:

> **map** *vt* To place (a mathematical aggregate) in a one-to-one correspondence with an aggregate <a set is called denumerable if it can be *mapped* ... onto the set of all the natural numbers —A. H. Wallace.

Maps describe the specific subcomponents of *individual* objects: a particular city, a specific state. Anatomies, on the other hand, describe the average characteristics of *collections* of objects. The failure to recognize this fundamental difference makes Feynman's request for a cat map so ludicrous.

Genome researchers either laugh or are annoyed when lay persons ask, whose genome will you map? When asked why this is so annoying, they tend to answer, "Because it is so wrong-headed." Some geneticists have been known to respond, "Asking that question is like asking whose face is in *Gray's Anatomy*!" Gray's ***anatomy***, indeed.

If someone says that he is making a map of, say, a European country, it is natural to ask which country. The notion of mapping an unspecified singular thing is almost meaningless. Conversely, developing an anatomy based on one specimen is equally problematic. If I measure my dog and you measure your dog, how will we ever agree on canine anatomy?

Insisting that the concept of genomic anatomy is preferable to that of genomic map is not mere wordplay. With an anatomy, for example, we may know that some structures are very regular from individual to individual and these are represented very precisely in our anatomical description. Each regular part gets a name and is well described. But some structures may be equally well known to vary considerably from individual to individual. These structures get only generic names, and we take care to point out that much variation is to be expected. No anatomist gives individual names to each of the small venous anastomoses in the human forearm. They are simply too variable to warrant individual names. Similarly, no genome informaticist should be expected to keep track of the many copies of the human rRNA genes, they are simply too variable in number and location. In both cases we must accurately record the variation as perhaps the most important part of the observation.

CONCLUSIONS

What, then, is a gene? Given all of the complex units of regulation and of transcription and of translation that are now know to occur, we must abandon the early molecular concept of the gene as a discrete, contiguous region of DNA with definable function. Instead, we must recognize that a gene or other map object of interest may well consist of a set of not necessarily discrete and not necessarily contiguous regions along a DNA molecule. Only by defining a *set* of regions can we construct data models of sufficient complexity to represent reality. Darnell et al. (1986) have said as much:

> The concept of the gene as a biological entity remains intact: a gene is still considered a heritable function detected by observing the effect of the mutation. However, according to the current definition, a gene consists of all

> the DNA sequences necessary to produce a single peptide or RNA product. Thus, the gene is no longer thought of a single contiguous stretch of DNA.

These sets must be hierarchical in the sense that permits more complex map-object sets to be created from sets of less complex sets. The set-of-sets concept allows us to represent easily the many regulatory regions of the MMS operon and the participation of the MMS operon in the heat-shock regulon. It also allows us to represent the *UGT1* region as both five separate genes and a pseudogene and as a higher order locus consisting of that set of five genes and a pseudogene.

As for mapping, we must abandon the notion of creating a single correct human genomic *map* and begin thinking about developing an accurate genomic *anatomy* instead. This anatomy must be capable of representing regular regions with precision and variable regions with due attention to the variability and uncertainty. Only by recognizing the anatomical nature of the genome will we be able to develop the data models and database systems to carry us through the successful completion of the human genome project.

REFERENCES

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D., 1983, *Molecular Biology of the Cell*, Garland Publishing Company, New York.

Ayala, F.J., and Kiger, J.A., Jr., 1984, *Modern Genetics, Second Edition*, Benjamin/Cummings Publishing Company, Inc., Menlo Park, California.

Bachmann, B.J., 1990, Linkage map of *Escherichia coli* K-12, edition 8., *Microbiol. Rev.* 54:130-97.

Benzer, S., 1955, Fine structure of a genetic region in bacteriophage, *Proc. Nat. Acad. Sci.*, 41:344.

Carlson, E.A., 1959, Comparative genetics of complex loci, *Quarterly Review of Biology*, 34:33-67.

Carlson, E.A., 1966, *The Gene: A Critical History*, W. B. Saunders Company, Philadelphia.

Cinkosky, M.J., Fickett, J.W., Gilna, P., and Burks, C., 1991, Electronic Data Publishing and GenBank, *Science* 252:1273.

Committee on Mapping and Sequencing the Human Genome, Board on Basic Biology, Commission on Life Sciences, National Research Council, 1988, *Mapping and Sequencing the Human Genome*, National Academy Press, Washington, D.C.

Courteau, J., 1991, Genome databases, *Science* 254:201.

Darnell, J., Lodish, H., and Baltimore, D., 1986, *Molecular Cell Biology*, Scientific American Books, New York.

Demerec, M., 1933, What is a gene?, *J. Hered.*, 64:369.

Demerec, M., 1955, What is a gene—twenty years later?, *Amer. Nat.*, 89:5.

Feynman, R.P., and Leighton, R., 1985, *Surely You're Joking, Mr. Feynman!* W. W. Norton and Company, Inc, New York.

Glass, B., 1955, Pseudoalleles, *Science*, 122:233.

Henikoff, S., Keene, M.A., Fechtel, K., and Fristrom, J.W., 1986, Gene within a gene: Nested *Drosophila* genes encode unrelated proteins on opposite DNA strands, *Cell* 44:33.

Kohara, Y., Akiyama, K., Isono, K., 1987, The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library, *Cell* 50:495.

Lupski, J.R., and Godson, G.N., 1989, DNA ➔ DNA, and DNA ➔ RNA ➔ protein: Orchestration by a single complex operon, *BioEssays* 10:152

McKusick, V.A., 1988, *The Morbid Anatomy of the Human Genome*, Howard Hughes Medical Institute, Bethesda, Maryland.

McKusick, V.A., and Ruddle, F.H., 1987, Toward a complete map of the human genome, *Genomics* 1:103-6.

Muller, H.J., 1945., The gene, *Proc. Roy. Soc. Biol.*, 134:1.

Pearson, M.L., and Söll, D., 1991, The human genome project: a paradigm for information management in the life sciences, *The FASEB Journal*, 5:35.

Ritter, J.K.,Chen, F., Sheen, Y.Y., Tran, H.M., Kimura, S., Yeatman, M.T., and Owens, I.S., 1992, A novel complex locus *UGT1* encodes human bilirubin, phenol, and other UDP-glucuronosyltransferases isozymes with identical carboxyl termini, *J. Biol. Chem.* 267:3257.

Singer, M., and Berg, P., 1992, *Genes & Genomes*, University Science Books, Mill Valley, California.

Stadler, L.J., 1954, The gene, *Science*, 120:811.

Sturtevant, A.H., 1913, The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association, *J. Exp. Zool.* 14:43, *reprinted in:* "Conceptual Foundations of Genetics," H.O. Corwin and J. B. Jenkins, eds., Houghton Mifflin Company, Boston.

Sturtevant, A.H., and Beadle, G.W., 1939, *An Introduction to Genetics*, W. B. Saunders Company, Philadelphia.

Suzuki, D.T., Griffiths, A.J.F., Miller, J.M., and Lewontin, R.C., 1986, *An Introduction to Genetics*, W. H. Freeman and Company, San Francisco.

U.S. Department of Health and Human Services and U.S. Department of Energy, 1990, *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years*, National Technical Information Service, U.S. Department of Commerce, Washington, D.C.