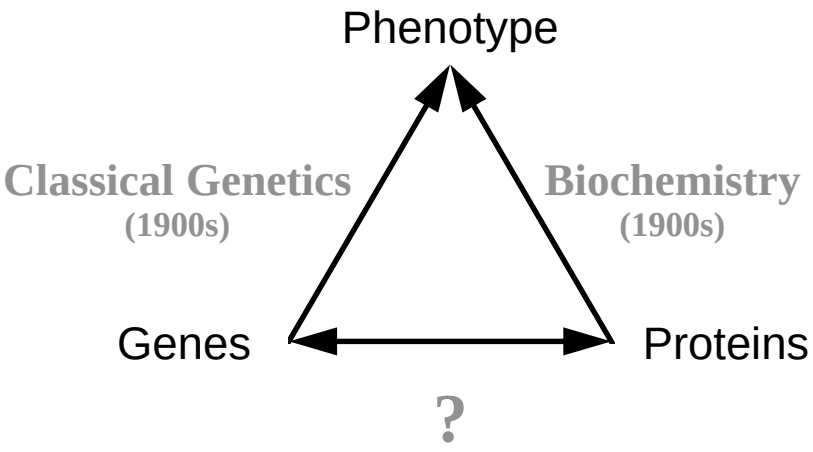
Molecular Biology Fundamentals

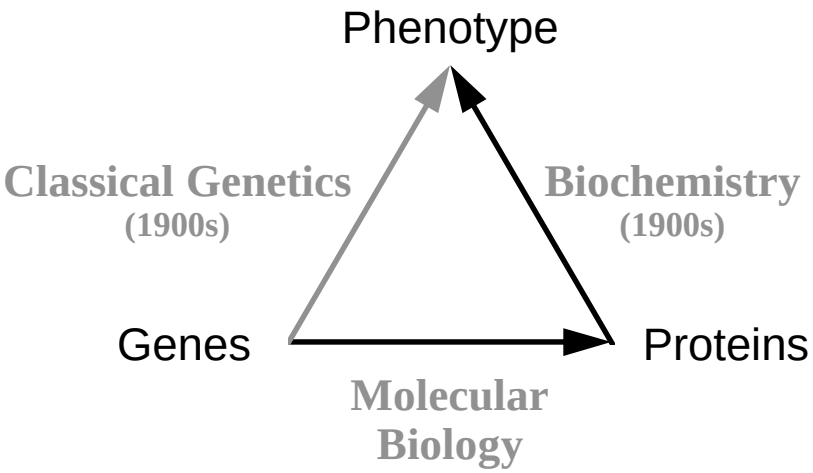
Robert J. Robbins

Johns Hopkins University
rrobbins@gdb.org

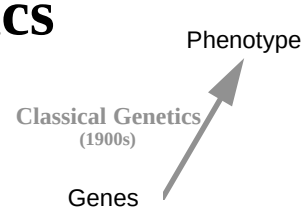
Origins of Molecular Biology

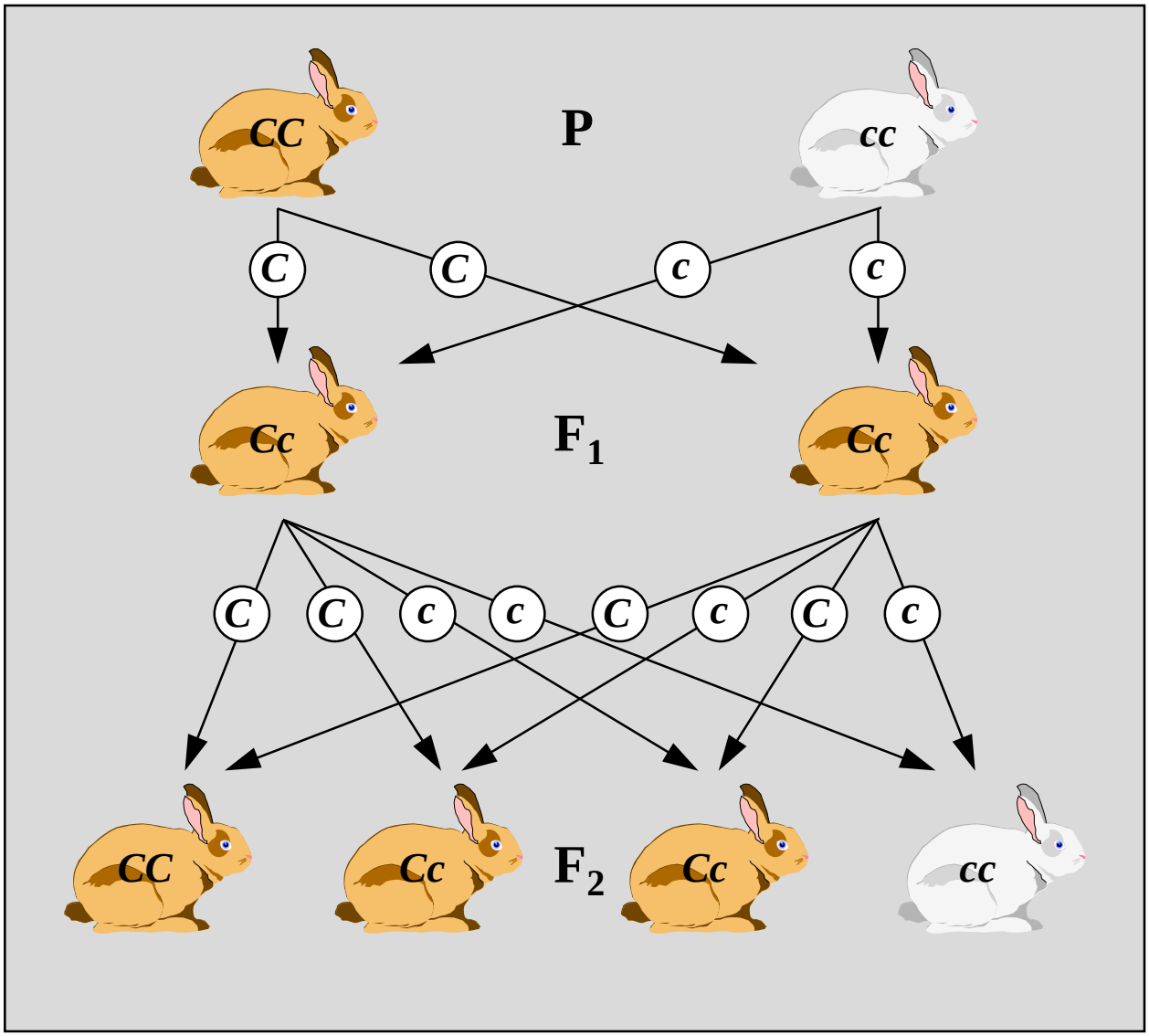


The ***phenotype*** of an organism denotes its external appearance (size, color, intelligence, etc.). ***Classical genetics*** showed that genes control the transmission of phenotype from one generation to the next. ***Biochemistry*** showed that within one generation, *proteins* had a determining effect on phenotype. For many years, however, the relationship between genes and proteins was a mystery. Then, it was found that genes contain digitally encoded instructions that direct the synthesis of proteins. The crucial insight of ***molecular biology*** is that hereditary information is passed between generations in a form that is truly, not metaphorically, digital. Understanding how that digital code directs the creation of life is the goal of molecular biology.



Classical Genetics





Regular numerical patterns of inheritance showed that the passage of traits from one generation to the next could be explained with the assumption that hypothetical particles, or ***genes***, were carried in pairs in adults, but transmitted individually to progeny.

Classical Genetics

During the first half of this century, classical investigation of the gene established that theoretical objects called genes were the fundamental units of heredity. According to the classical model of the gene:

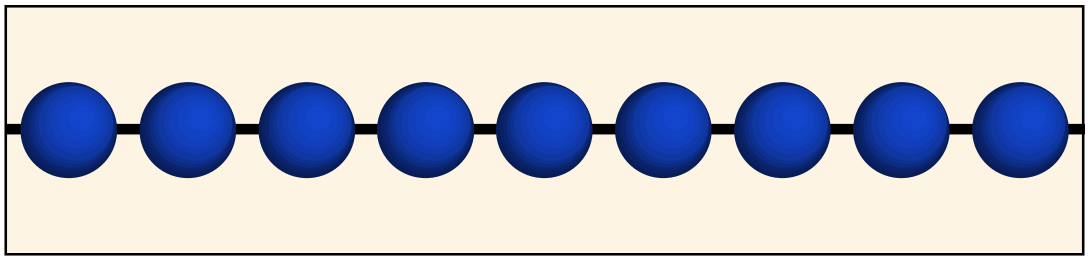
Genes behave in inheritance as independent particles.

Genes are carried in a linear arrangement in the chromosome, where they occupy stable positions.

Genes recombine as discrete units.

Genes can mutate to stable new forms.

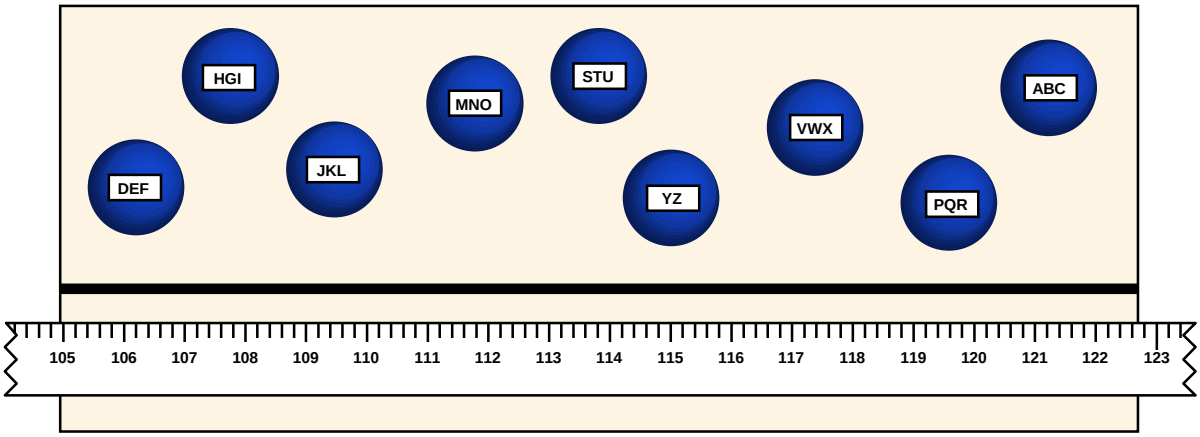
Basically, genes seemed to be particulate objects, arranged on the chromosome like “beads on a string.”



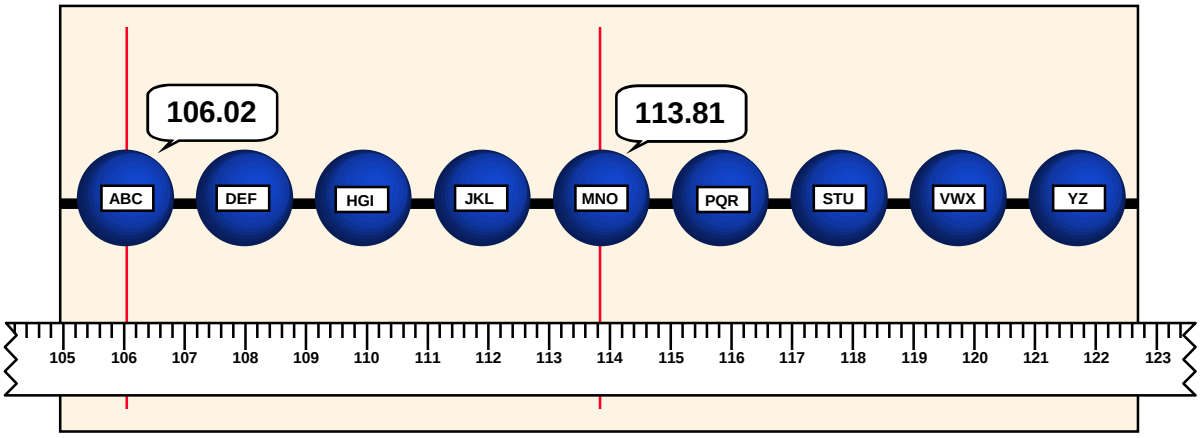
The genes are arranged in a manner similar to beads strung on a loose string.

Sturtevant, A.H., and Beadle, G.W., 1939. *An Introduction to Genetics*. W. B. Saunders Company, Philadelphia, p. 94.

Classical Genetics

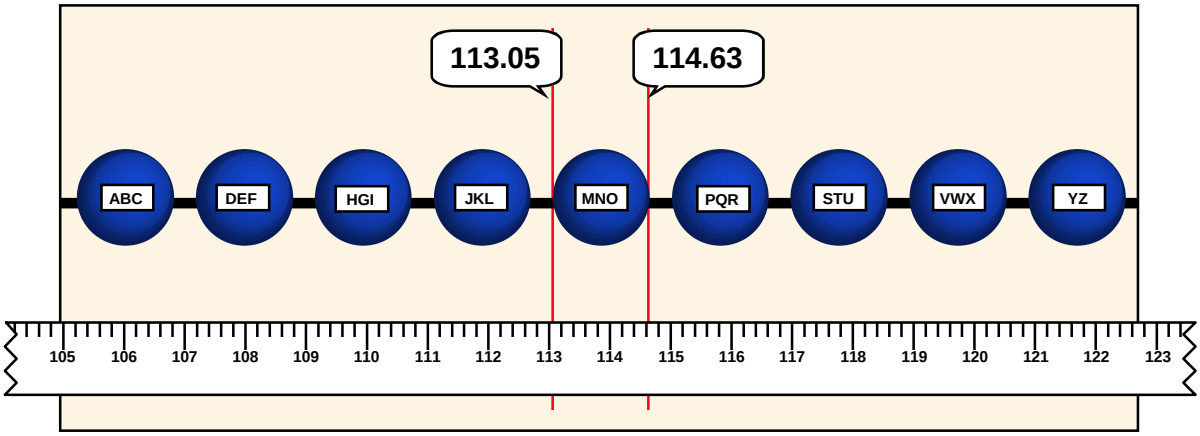


The beads can be conceptually separated from the string, which has "addresses" that are independent of the beads.

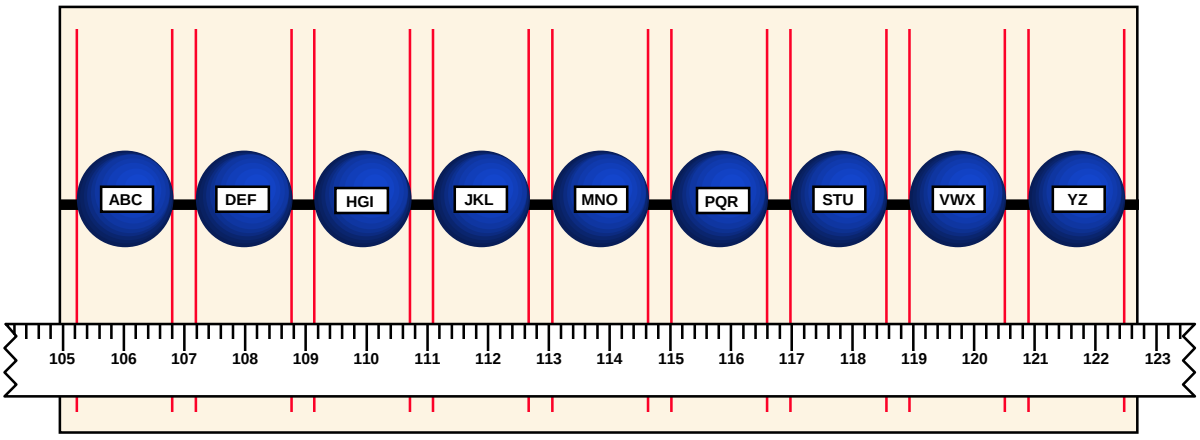


Mapping involves placing the beads in the correct order and assigning a correct address to each bead. The address assigned to a bead is its locus.

Classical Genetics

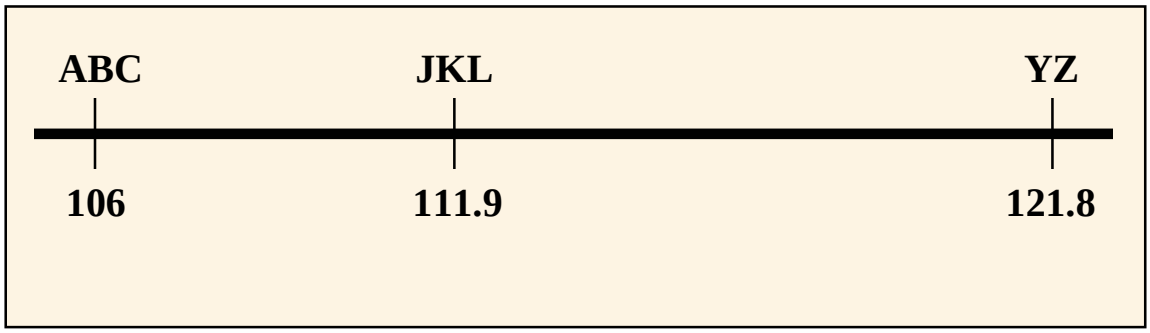


Recognizing that the beads have width, mapping could be extended to assigning a pair of numbers to each bead so that a locus is defined as a region, not a point.



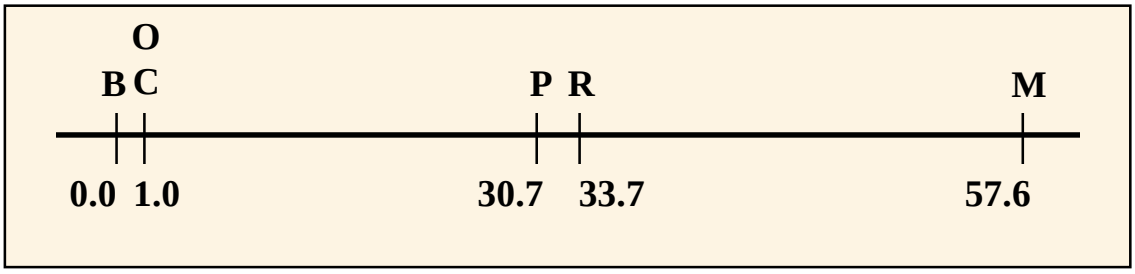
In this model, genes are independent, mutually exclusive, non-overlapping entities, each with its own absolute address.

Classical Genetics



In principle, maps of a few genes might be represented by showing the gene names in order, with their relative positions indicated.

Drosophila melanogaster



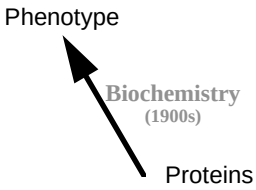
B = yellow body
C = white eye
O = eosin eye

P = vermilion eye
R = rudimentary wing

M = miniature wing

And, in fact, the first genetic map ever published was of just that type. Sturtevant, A.H., 1913, The linear arrangement of six sex-linked factors in *Drosophila* as shown by their mode of association, *Journal of Experimental Zoology*, 14:43-59.

Biochemistry



The aim of modern biology is to interpret the properties of the organism by the structure of its constituent molecules.

Jacob, F. 1973. *The Logic of Life.* New York: Pantheon Books.

Understanding the molecular basis of life had its beginnings with the advent of biochemistry. Early in the nineteenth century, it was discovered that preparations of fibrous material could be obtained from cell extracts of plants and animals. Mulder concluded in 1838 that this material was:

> without doubt the most important of the known components of living matter, and it would appear that without life would not be possible. This substance has been named *protein*.

Later, many wondered whether chemical processes in living systems obeyed the same laws as did chemistry elsewhere. Complex carbon-based compounds were readily synthesized in cells, but seemed impossible to construct in the laboratory.

By the beginning of the twentieth century, chemists had been able to synthesize a few organic compounds, and, more importantly, to demonstrate that complex organic reactions could be accomplished in non-living cellular extracts. These reactions were found to be catalyzed by a class of proteins called *enzymes.*

Early biochemistry, then, was characterized by (1) efforts to understand the structure and chemistry of proteins themselves, and (2) efforts to discover, catalog, and understand enzymatically catalayzed biochemical reactions.

Genetic Fallacies

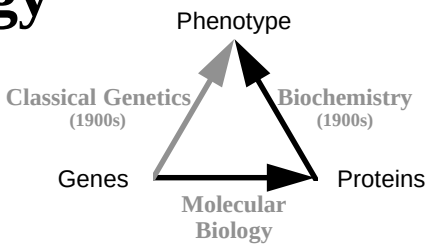
Before molecular biology began, biochemists believed that DNA was composed of a monotonous rotation of four basic components, the nucleotides adenine, cytosine, guanine, and thymine. Since a repeating polymer consisting of four subunits could not encode information, it was widely held that DNA provided only a structural role in chromosomes and that genetic information was stored in protein.

If the genes are conceived as chemical substances, only one class of compounds need be given to which they can be reckoned as belonging, and that is the proteins in the wider sense, on account of the inexhaustible possibilities for variation which they offer. ... Such being the case, the most likely role for the nucleic acids seems to be that of the structure-determining supporting substance.

T. Caspersson. 1936. Über den chemischen Aufbau der Strukturen des Zellkernes. *Acta Med. Skand*., 73, Suppl. 8, 1-151.

At any given time in a particular science, there will be beliefs that are held so strongly that they are considered beyond challenge, yet they will prove to be wildly wrong. This poses a great challenge for the design of scientific databases, which must reflect current beliefs in the field, yet be robust in the face of changes in fundamental concepts or practices.

Molecular Biology



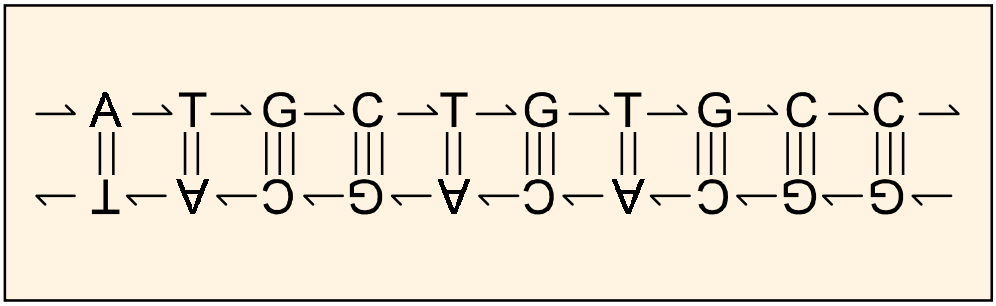
Key Discoveries:

1928 Heritable changes can be transmitted from bacterium to bacterium through a chemical extract (the ***transforming factor***) taken from other bacteria.

1944 The transforming factor appears to be DNA.

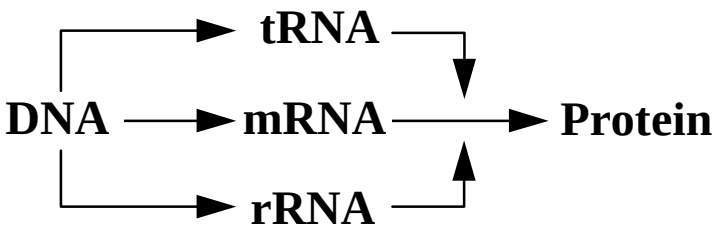
1950 The tetranucleotide hypothesis of DNA structure is overthrown.

1953 The structure of DNA is established to be a double helix.



DNA is constructed as a double-stranded molecule, with absolutely no constraints upon the liner order of subcomponents along each strand, but with the pairing between strands totally constrained according to complementarity rules: A always pairs with T and C always pairs with G.

The Fundamental Dogma



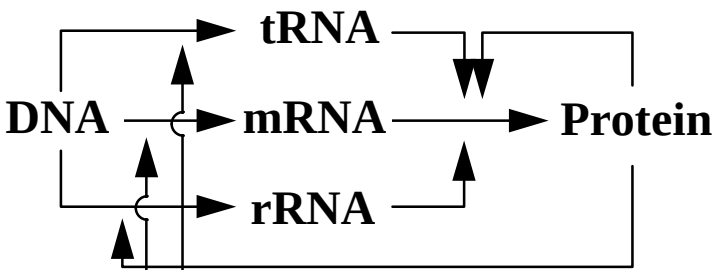
Information coded in **DNA** (deoxyribonucleic acid) directs the synthesis of different **RNA** (ribonucleic acid) molecules. RNA molecules fall into several different categories:

rRNA: ***ribosomal RNA*** that is required for building ribosomes, which are structures necessary for protein synthesis.

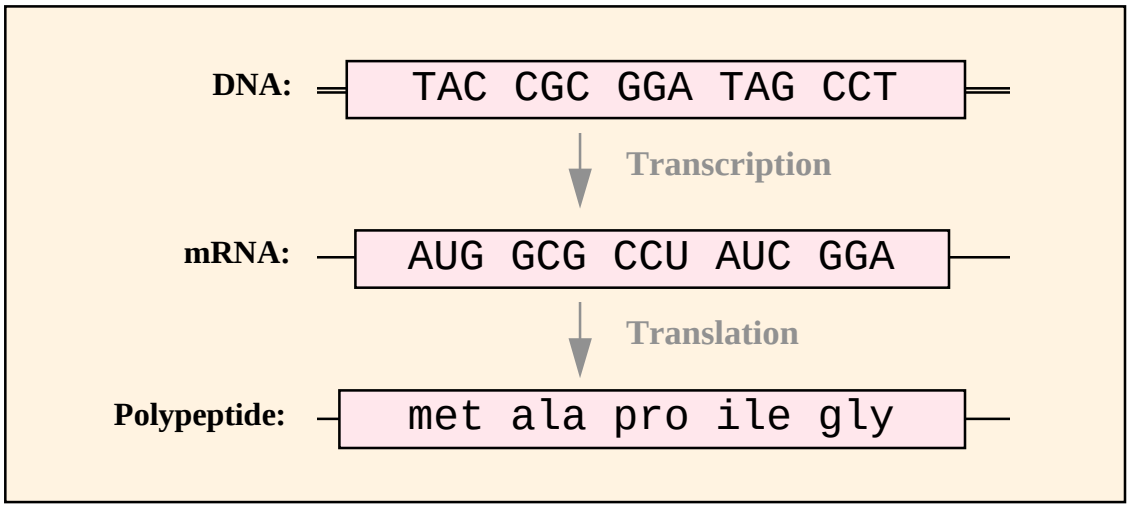
tRNA: ***transfer RNA*** that serves to transfer individual amino acid molecules from the general cytoplasm to their appropriate location in a growing polypeptide during protein synthesis.

mRNA: ***messenger RNA*** that carries the specific instructions for building a specific protein.

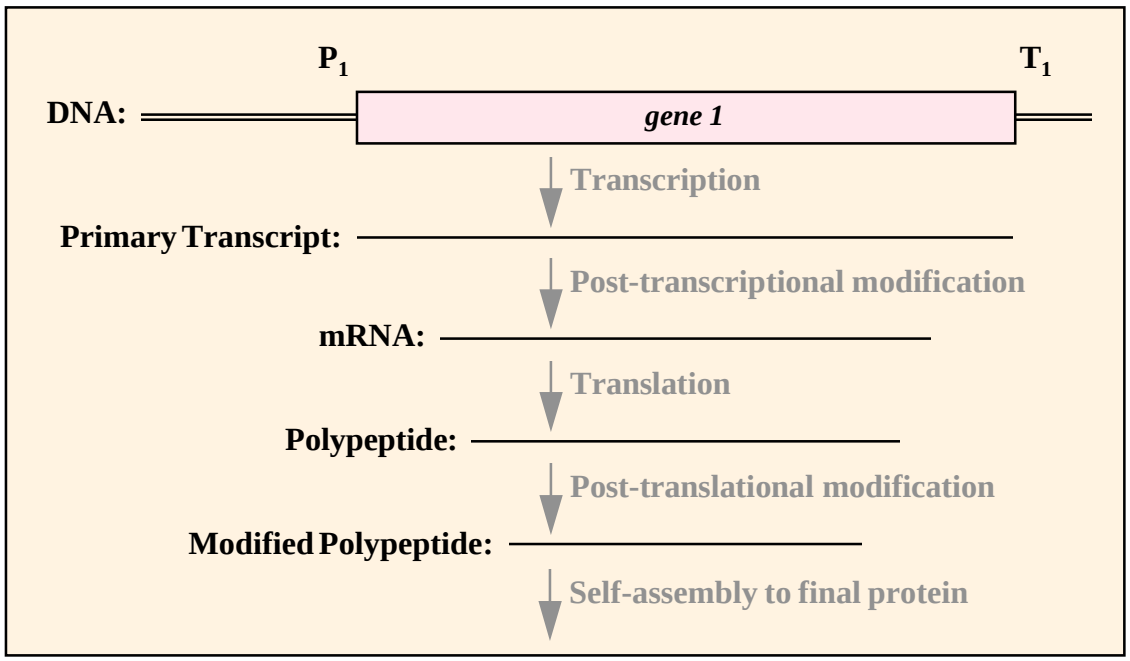
Both rRNA and tRNA are generic groups of molecules in that all types of rRNA and all types of tRNA are involved in the synthesis of every type of protein. However, mRNA is specific in that a different type of mRNA is required for every different type of protein.



The whole system is recursive, in that certain proteins are required for the synthesis of RNAs, as well as for the synthesis of DNA itself.



DNA directs protein synthesis through a multi-step process. First, DNA is copied to mRNA through the process of transcription. The rules governing transcription are the same as the rules govering the interstrand constraint in DNA. Then translation produces a polypeptide with an amino-acid sequence that is completely specified by the sequence of nucleotides in the RNA. A simple code, the same for all living things on this planet, governs the synthesis of protein from mRNA instructions.



Some post-transcriptional processing of the immediate RNA transcript is necessary to produce a finished RNA, and post-translational processing of polypeptides can be needed to produce a final protein.

mRNA to Amino Acid Dictionary

5´	U	C	A	G	3´
U	phe	ser	tyr	cys	U
	phe	ser	tyr	cys	C
	leu	ser	STOP	STOP	A
	leu	ser	STOP	trp	G
C	leu	pro	his	arg	U
	leu	pro	his	arg	C
	leu	pro	gln	arg	A
	leu	pro	gln	arg	G
A	ile	thr	asn	ser	U
	ile	thr	asn	ser	C
	ile	thr	lys	arg	A
	met	thr	lys	arg	G
G	val	ala	asp	gly	U
	val	ala	asp	gly	C
	val	ala	glu	gly	A
	val	ala	glu	gly	G

This dictionary gives the sixty four different mRNA codons and the amino acids (or stop signals) for which they code. The 5' nucleotides are given along the left hand border, the middle nucleotides are given across the top, and the 3' nucleotides are given along the right hand border. The decoded meaning of a particular codon is given by the entry in the table.

For example, the meaning of the codon 5'AUG3' is determined as follows:

1. Examine the entries along the left hand side of the table to locate the horizontal block corresponding to the sixteen codons that have A in the 5' position.
2. Examine the entries along the top of the table to locate the vertical block corresponding to the sixteen codons that have U in the middle position.
3. Find the intersection of these two blocks. This intersection represents the four codons that have A in the 5' position and U in the middle position.
4. Examine the entries along the right hand side of the table to find the entry for the one codon that has A in the 5' position, U in the middle position, and G in the 3' position. The “met” indicates that the decoded meaning of the codon 5'AUG3' is methionine. That is, the codon 5'AUG3' codes for the amino acid methionine.

What is a Gene?

neoClassical Sequence Definitions (AB)

Gene (cistron) the fundamental unit of genetic function.

Gene (muton) the fundamental unit of genetic mutation.

Gene (recon) the fundamental unit of genetic recombination.

Gene (codon) the fundamental unit of genetic coding.

Summary Definitions

Classical Definition: fundamental unit of heredity, mutation, and recombination (beads on a string).

Physiological Definition: fundamental unit of function (one gene, one enzyme).

Cistronic Definition: fundamental unit of expression (cis-trans test).

Sequence Definition: the smallest segment of the gene-string consistently associated with the occurrence of a specific genetic effect.

Current Definition: ???

What is a Gene?
Current Textbook Definitions

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

Singer, M., and Berg, P. 1991. *Genes & Genomes*. University Science Books, Mill Valley, California.

Gene (cistron) is the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

Allele is one of several alternative forms of a gene occupying a given locus on a chromosome.

Locus is the position on a chromosome at which the gene for a particular trait resides; locus may be occupied by any one of the alleles for the gene.

Lewin, Benjamin. 1990. *Genes IV*. Oxford University Press, New York.

What is a Gene?
Current Textbook Definitions

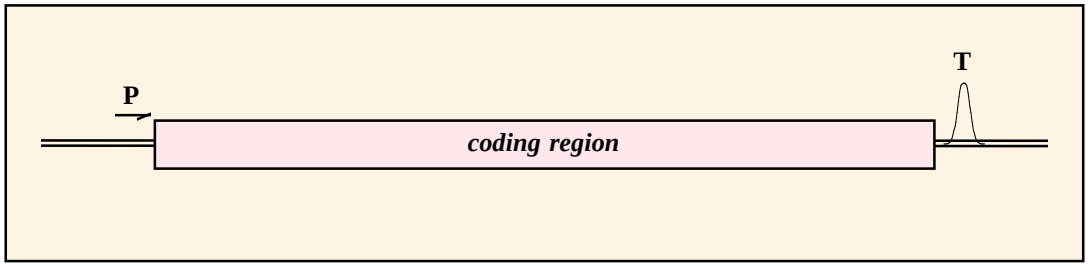
DNA molecules (chromosomes) should thus be functionally regarded as linear collections of discrete transcriptional units, each designed for the synthesis of a specific RNA molecule. Whether such "transcriptional units" should now be redefined as genes, or whether the term *gene* should be restricted to the smaller segments that directly code for individual mature rRNA or tRNA molecules or for individual peptide chains is now an open question.

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M. 1992. *Molecular Biology of the Gene*. Benjamin/Cummins Publishing Company: Menlo Park, California. p. 233.

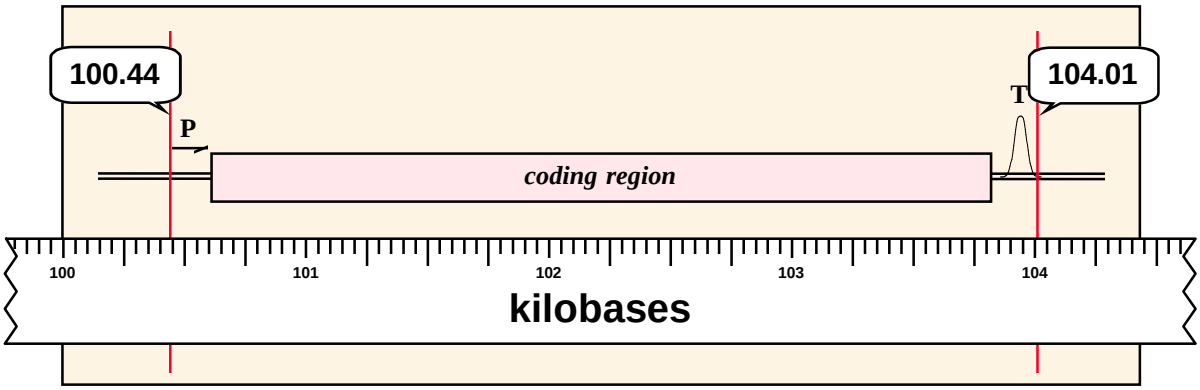
For the purposes of this book, we have adopted a molecular definition. A eukaryotic gene is a combination of DNA segments that together constitute an expressible unit, expression leading to the formation of one or more specific functional gene products that may be either RNA molecules or polypeptides.

Singer, M., and Berg, P. 1991. *Genes & Genomes*. University Science Books, Mill Valley, California.

The Simplistic View of a Gene as Sequence

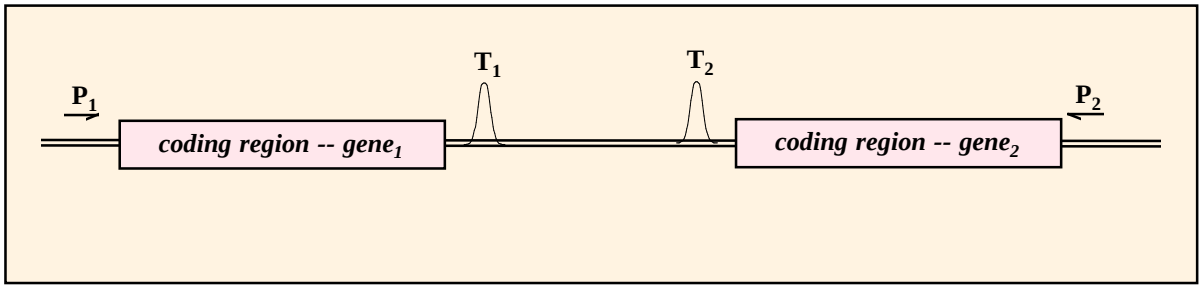


A gene is a transcribed region of DNA, flanked by upstream start regulatory sequences and downstream stop regulatory sequences.

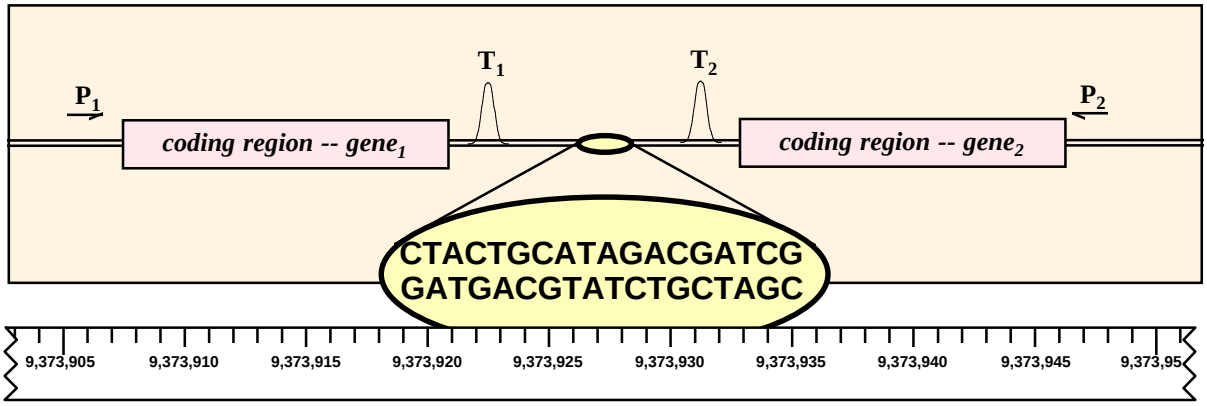


The location of a gene can be designated by specifying the base-pair location of its beginning and end.

The Simplistic View of a Gene as Sequence

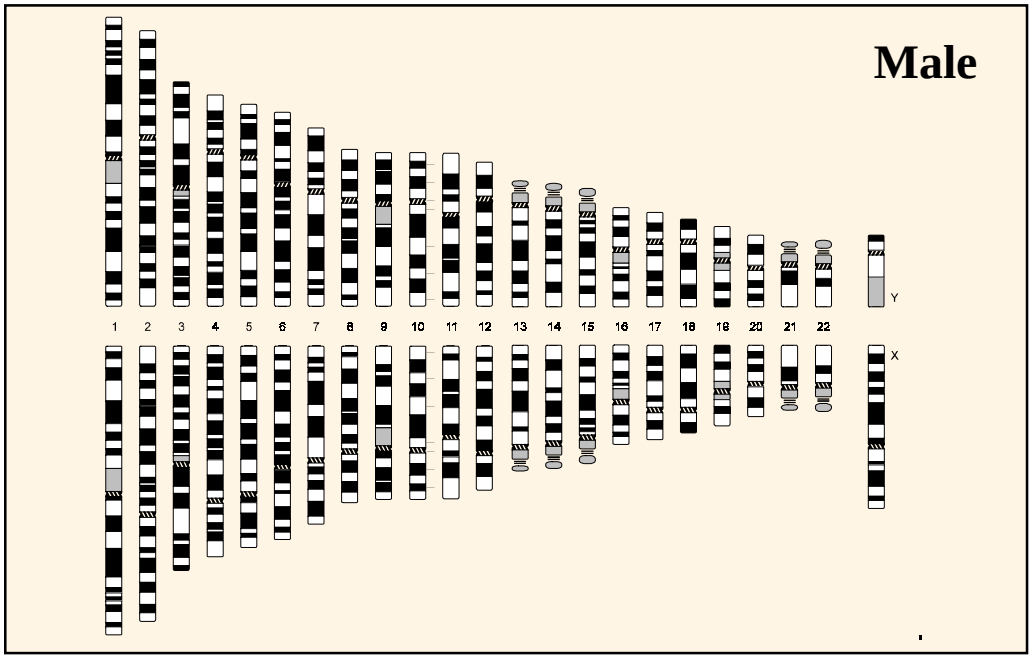


DNA may be transcribed in either direction. Therefore, fully specifying a gene's position requires noting its orientation as well as its start and stop positions.

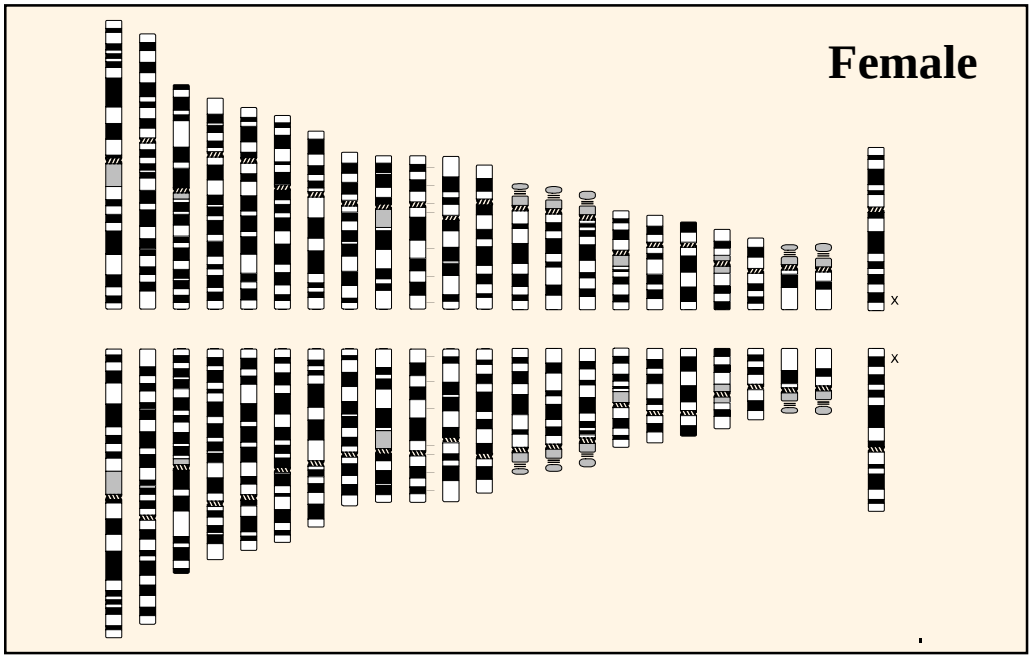


A naive view holds that a genome can be represented as a continuous linear string of nucleotides, with landmarks identified by the chromosome number followed by the offset number of the nucleotide at the beginning and end of the region of interest. This simplistic approach ignores the fact that chromosomes may vary in length by tens of millions of nucleotides.

The Human Genome Project

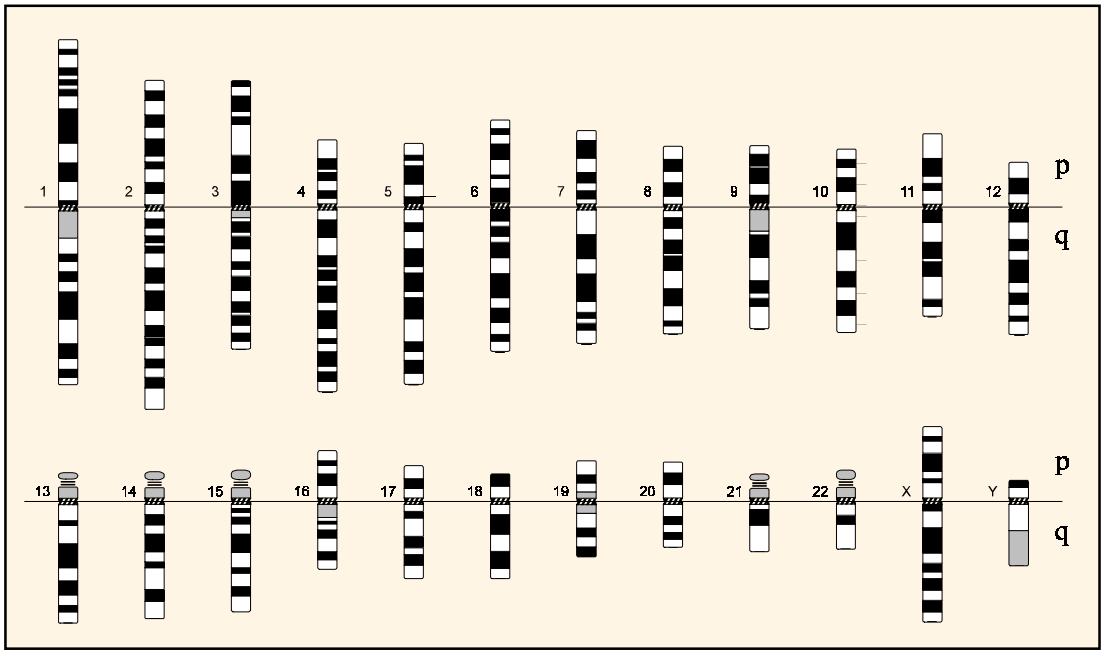


At conception, a normal human receives 23 chromosomes from each parent -- 22 **autosomes** and one **sex chromosome**. The mother always contributes 22 autosomes and one **X chromosome**. If the father also contributes an X chromosome, the child will be female. If the father contributes a **Y chromosome**, the child will be male.



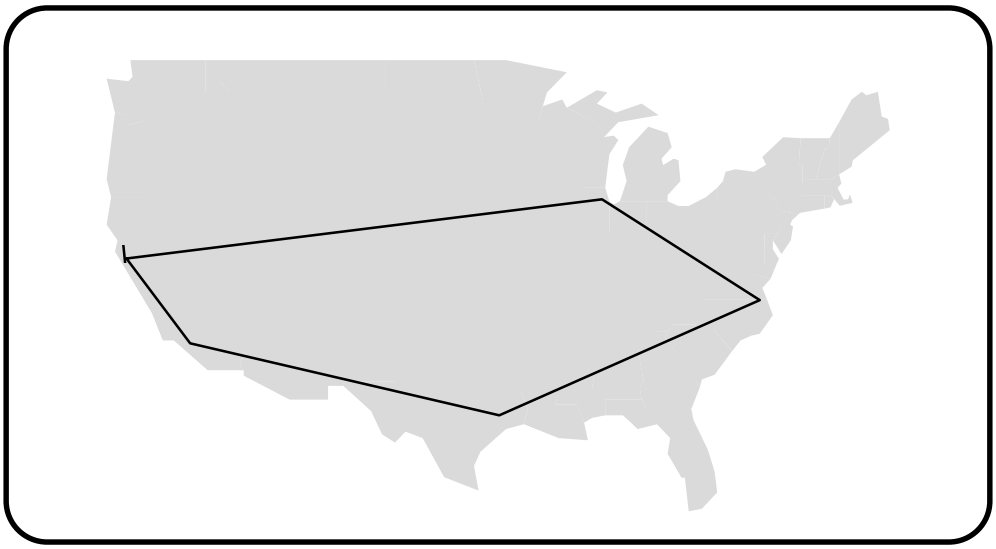
The human genome is believed to consist of 50,000 to 100,000 genes encoded in 3.3 billion base pairs of DNA, which are packaged into 23 chromosomes. The goal of the Human Genome Project (HGP) is learning the specific order of those 3.3 billion base pairs and of identifying and locating all of the genes encoded by that DNA. Databases must be developed to hold, manage, and distribute all of those findings

The HGP can be logically divided into two components: (1) obtaining the sequence, and (2) understanding the sequence, and neither of them involves a simple 3.3 gigabyte database with straightforward computational requirements.



The Challenge: Consider the DNA sequence of a human genome as equivalent to 3.3 gigabytes of files on the mass-storage device of some computer system of unknown design. Obtaining the sequence is equivalent to obtaining an image of the contents of that mass-storage device. Understanding the sequence is equivalent to reverse engineering that unknown computer system (both the hardware and the 3.3 gigabytes of software) all the way back to a full set of design and maintenance specifications.

Getting the Sequence

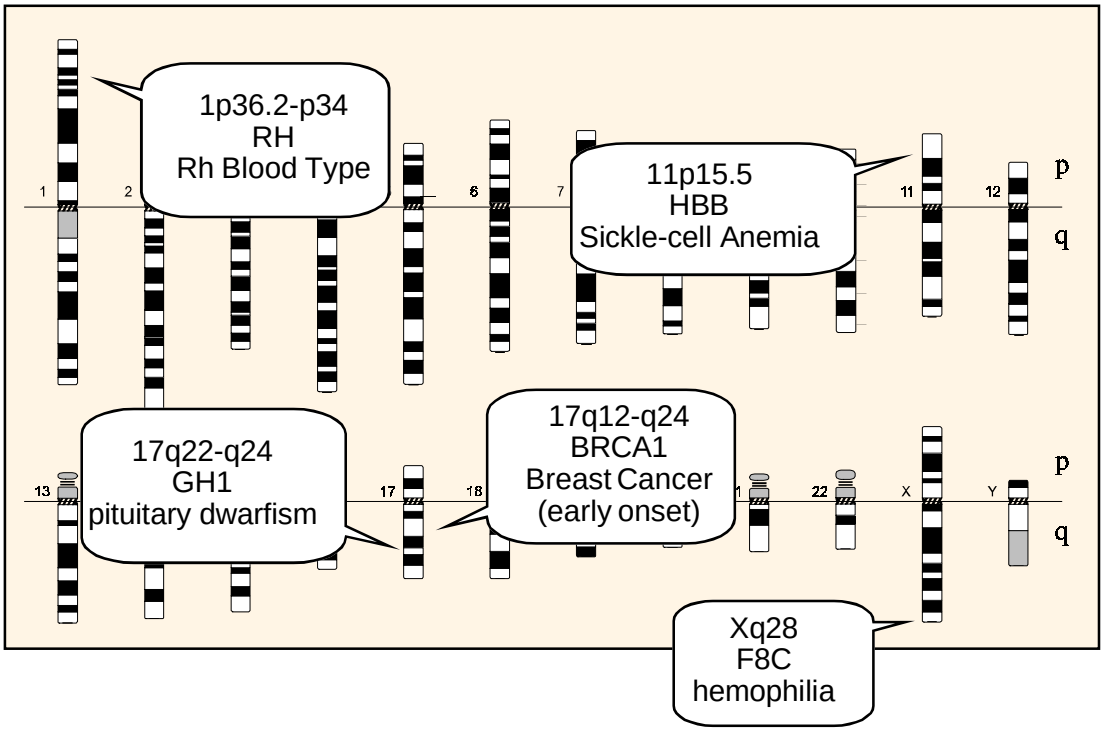


Obtaining one full human sequence will be a technical challenge. If the DNA sequence from a single human sperm cell were typed on a continuous ribbon in ten-pitch type, that ribbon could be stretched from San Francisco to Chicago to Washington to Houston to Los Angeles, and back to San Francisco, with about 60 miles of ribbon left over.

The amount of human sequence currently sequenced is equal to less than one-third of that left-over 60-mile fragment. We have a long way to go, and getting there will be expensive. Computers will play a crucial role in the entire process, from robotics to control experimental equipment to complex analytical methods for assembling sequence fragments.

year	per base cost	budget	year	cumulative	percent completed
1995	$0.50	16,000,000	10,774,411	10,774,411	0.33%
1996	$0.40	25,000,000	21,043,771	31,818,182	0.96%
1997	$0.30	35,000,000	39,281,706	71,099,888	2.15%
1998	$0.20	50,000,000	84,175,084	155,274,972	4.71%
1999	$0.15	75,000,000	168,350,168	323,625,140	9.81%
2000	$0.10	100,000,000	336,700,337	660,325,477	20.01%
2001	$0.05	100,000,000	673,400,673	1,333,726,150	40.42%
2002	$0.05	100,000,000	673,400,673	2,007,126,824	60.82%
2003	$0.05	100,000,000	673,400,673	2,680,527,497	81.23%
2004	$0.05	100,000,000	673,400,673	3,353,928,171	101.63%

Defective Genes Cause Disease



Many human diseases are known to associated with specific defects in particular genes. These defects are equivalent to coding errors in files on a mass storage system.

A defective copy of the gene for beta-hemoglobin (HBB) can lead to sickle-cell anemia.

Beta Hemoglobin

```
   1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
  61 ccagggctgg gcataaaagt cagggcagag ccatctattg ctt**acatttg cttctgacac**
 121 **aactgtgttc actagcaacc tcaaacagac accATGGTGC ACCTGACTCC TGAGGAGAAG**
 181 **TCTGCCGTTA CTGCCCTGTG GGGCAAGGTG AACGTGGATG AAGTTGGTGG TGAGGCCCTG**
 241 **GGCAGG**ttgg tatcaaggtt acaagacagg tttaaggaga ccaatagaaa ctgggcatgt
 301 ggagacagag aagactcttg ggtttctgat aggcactgac tctctctgcc tattggtcta
 361 ttttcccacc cttagg**CTGC TGGTGGTCTA CCCTTGGACC CAGAGGTTCT TTGAGTCCTT**
 421 **TGGGGATCTG TCCACTCCTG ATGCTGTTAT GGGCAACCCT AAGGTGAAGG CTCATGGCAA**
 481 **GAAAGTGCTC GGTGCCTTTA GTGATGGCCT GGCTCACCTG GACAACCTCA AGGGCACCTT**
 541 **TGCCACACTG AGTGAGCTGC ACTGTGACAA GCTGCACGTG GATCCTGAGA ACTTCAGG**gt
 601 gagtctatgg gacccttgat gttttctttc cccttctttt ctatggttaa gttcatgtca
 661 taggaagggg agaagtaaca gggtacagtt tagaatggga aacagacgaa tgattgcatc
 721 agtgtggaag tctcaggatc gttttagttt cttttatttg ctgttcataa caattgtttt
 781 cttttgttta attcttgctt tctttttttt tcttctccgc aatttttact attatactta
 841 atgccttaac attgtgtata acaaaaggaa atatctctga gatacattaa gtaacttaaa
 901 aaaaaacttt acacagtctg cctagtacat tactatttgg aatatatgtg tgcttatttg
 961 catattcata atctccctac tttatttttct tttatttta attgatacat aatcattata
1021 catatttatg ggttaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
1081 taattttgca tttgtaattt taaaaaatgc tttcttcttt taatatactt ttttgtttat
1141 cttatttcta atactttccc taatctcttt ctttcagggc aataatgata caatgtatca
1201 tgcctctttg caccattcta aagaataaca gtgataattt ctgggttaag gcaatagcaa
1261 tatttctgca tataaatatt tctgcatata aattgtaact gatgtaagag gtttcatatt
1321 gctaatagca gctacaatcc agctaccatt ctgcttttat tttatggttg ggataaggct
1381 ggattattct gagtccaagc taggcccttt tgctaatcat gttcatacct cttatcttcc
1441 tcccacag**CT CCTGGGCAAC GTGCTGGTCT GTGTGCTGGC CCATCACTTT GGCAAAGAAT**
1501 **TCACCCCACC AGTGCAGGCT GCCTATCAGA AAGTGGTGGC TGGTGTGGCT AATGCCCTGG**
1561 **CCCACAAGTA TCACTAAgct cgctttcttg ctgtccaatt tctattaaag gttcctttgt**
1621 **tccctaagtc caactactaa actgggggat attatgaagg gccttgagca tctggattct**
1681 **gcctaataaa aaacatttat tttcattgc**a atgatgtatt taaattattt ctgaatattt
1741 tactaaaaag ggaatgtggg aggtcagtgc atttaaaaca taaagaaatg atgagctgtt
1801 caaaccttgg gaaaatacac tatatcttaa actccatgaa agaaggtgag gctgcaacca
1861 gctaatgcac attggcaaca gcccctgatg cctatgcctt attcatccct cagaaaagga
1921 ttcttgtaga ggcttgattt gcaggttaaa gttttgctat gctgtatttt acattactta
1981 ttgttttagc tgtcctcatg aatgtctttt cactacccat ttgcttatcc tgcatctctc
2041 tcagccttga ct
```

The genomic sequence for the beta-hemoglobin gene is given above. The letters in bold are the introns that are spliced together after initial transcription. The upper case letters are the actual coding region that specify the amino-acid sequence for beta-hemoglobin. The coding region is excerpted and given below.

```
ATG GTG CAC CTG ACT CCT GAG GAG AAG TCT GCC GTT ACT GCC CTG TGG GGC AAG GTG
AAC GTG GAT GAA GTT GGT GGT GAG GCC CTG GGC AGG CTG CTG GTG GTC TAC CCT TGG
ACC CAG AGG TTC TTT GAG TCC TTT GGG GAT CTG TCC ACT CCT GAT GCT GTT ATG GGC
AAC CCT AAG GTG AAG GCT CAT GGC AAG AAA GTG CTC GGT GCC TTT AGT GAT GGC CTG
GCT CAC CTG GAC AAC CTC AAG GGC ACC TTT GCC ACA CTG AGT GAG CTG CAC TGT GAC
AAG CTG CAC GTG GAT CCT GAG AAC TTC AGG CTC CTG GGC AAC GTG CTG GTC TGT GTG
CTG GCC CAT CAC TTT GGC AAA GAA TTC ACC CCA CCA GTG CAG GCT GCC TAT CAG AAA
GTG GTG GCT GGT GTG GCT AAT GCC CTG GCC CAC AAG TAT CAC TAA
```

Beta Hemoglobin

```
   1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
  61 ccagggctgg gcataaaagt cagggcagag ccatctattg ctt acatttg cttctgacac
 121 aactgtgttc actagcaacc tcaaacagac accATGGTGC ACCTGACTCC TGAGGAGAAG
 181 TCTGCCGTTA CTGCCCTGTG GGGCAAGGTG AACGTGGATG AAGTTGGTGG TGAGGCCCTG
 241 GGCAGGttgg tatcaaggtt acaagacagg tttaaggaga ccaatagc  ctgggcatgt
 301 ggagacagag aagactctt                                  gtcta
 361 ttttcccacc cttaggCT                                  CCTT
 421 TGGGGATCTG TCCACTCC                                  GCAA
 481 GAAAGTGCTC GGTGCCTT                                  CCTT
 541 TGCCACACTG AGTGAGCT                                  GGgt
 601 gagtctatgg gacccttg                                  gtca
 661 taggaagggg agaagtaa                                  catc
 721 agtgtggaag tctcaggat                                 gtttt
 781 cttttgttta attcttgctt tctttttttt tcctcctgga aatttttact attatactta
 841 atgccttaac attgtgtata acaaaaggaa atatctctga gatacattaa gtaacttaaa
 901 aaaaaacttt acacagtctg cctagtacat tactatttgg aatatatgtg tgcttatttg
 961 catattcata atctccctac tttattttct tttatttta attgatacat aatcattata
1021 catatttatg ggttaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
1081 taattttgca tttgtaattt taaaaaatgc tttcttcttt taatatactt ttttgtttat
1141 cttatttcta atactttccc taatctcttt ctttcagggc aataatgata caatgtatca
1201 tgcctctttg caccattcta aagaataaca gtgataattt ctgggttaag gcaatagcaa
1261 tatttctgca tataaatatt tctgcatata aattgtaact gatgtaagag gtttcatatt
1321 gctaatagca gctacaatcc agctaccatt ctgcttttat tttatggttg ggataaggct
1381 ggattattct gagtccaagc taggcccttt tgctaatcat gttcatacct cttatcttcc
1441 tcccacagCT CCTGGGCAAC GTGCTGGTCT GTGTGCTGGC CCATCACTTT GGCAAAGAAT
1501 TCACCCCACC AGTGCAGGCT GCCTATCAGA AAGTGGTGGC TGGTGTGGCT AATGCCCTGG
1561 CCCACAAGTA TCACTAAgct cgctttcttg ctgtccaatt tctattaaag gttcctttgt
1621 tccctaagtc caactactaa actgggggat attatgaagg gccttgagca tctggattct
1681 gcctaataaa aaacatttat tttcattgca atgatgtatt taaattattt ctgaatattt
1741 tactaaaaag ggaatgtggg aggtcagtgc atttaaaaca taaagaaatg atgagctgtt
1801 caaaccttgg gaaaatacac tatatcttaa actccatgaa agaaggtgag gctgcaacca
1861 gctaatgcac attggcaaca gcccctgatg cctatgcctt attcatccct cagaaaagga
1921 ttcttgtaga ggcttgattt gcaggttaaa gttttgctat gctgtatttt acattactta
1981 ttgttttagc tgtcctcatg aatgtctttt cactacccat ttgcttatcc tgcatctctc
2041 tcagccttga ct
```

Changing just one nucleotide out of 3,000,000,000 is enough to produce a lethal gene, just as one incorrect bit can crash an operating system.

A change in this nucleic acid from an A to T causes glutamic acid to be replaced with valine. This produces the sickle-cell allele.

```
ATG GTG CAC CTG ACT CCT GAG GAG AAG TCT GCC GTT ACT GCC CTG TGG GGC AAG GTG
AAC GTG GAT GAA GTT GGT GGT GAG GCC CTG GGC AGG CTG CTG GTG GTC TAC CCT TGG
ACC CAG AGG TTC TTT GAG TCC TTT GGG GAT CTG TCC ACT CCT GAT GCT GTT ATG GGC
AAC CCT AAG GTG AAG GCT CAT GGC AAG AAA GTG CTC GGT GCC TTT AGT GAT GGC CTG
GCT CAC CTG GAC AAC CTC AAG GGC ACC TTT GCC ACA CTG AGT GAG CTG CAC TGT GAC
AAG CTG CAC GTG GAT CCT GAG AAC TTC AGG CTC CTG GGC AAC GTG CTG GTC TGT GTG
CTG GCC CAT CAC TTT GGC AAA GAA TTC ACC CCA CCA GTG CAG GCT GCC TAT CAG AAA
GTG GTG GCT GGT GTG GCT AAT GCC CTG GCC CAC AAG TAT CAC TAA
```

Genomic Fallacies

Molecular Genetics:

> The ultimate ... map [will be] the complete DNA sequence of the human genome.

Committee on Mapping and Sequencing the Human Genome, 1988, *Mapping and Sequencing the Human Genome*. National Academy Press, Washington, D.C., p. 6.

The Ultimate Feature Table:

> As the Genome Project progresses, mapping and sequencing will converge. With the full human sequence available, it will be possible unambiguously to define every gene by the base-pair address of its functional subunits.

Genome Project as Database

When the Human Genome Project is finished, many of the innovative laboratory methods involved in its successful conclusion will begin to fade from memory. What will remain, as the project's enduring contribution, is a vast amount of computerized knowledge. Seen in this light, the Human Genome Project is nothing but the effort to create the most important database ever attempted—the database containing instructions for creating life.