Computing the Genome: Efforts to Reverse Engineer Humans

Robert J. Robbins

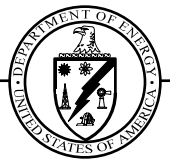
Johns Hopkins University

&

Department of Energy

rrobbins@gdb.org

robbins@er.doe.gov

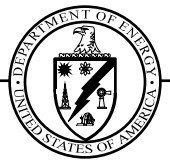


Human Genome Project

Overall Goals:

- **construction of a high-resolution genetic map of the human genome;**
- **production of a variety of physical maps of all human chromosomes and of the DNA of selected model organisms;**
- **determination of the complete sequence of human DNA and of the DNA of selected model organisms;**
- **development of capabilities for collecting, storing, distributing, and analyzing the data produced;**
- **creation of appropriate technologies necessary to achieve these objectives.**

USDOE. 1990. Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years.



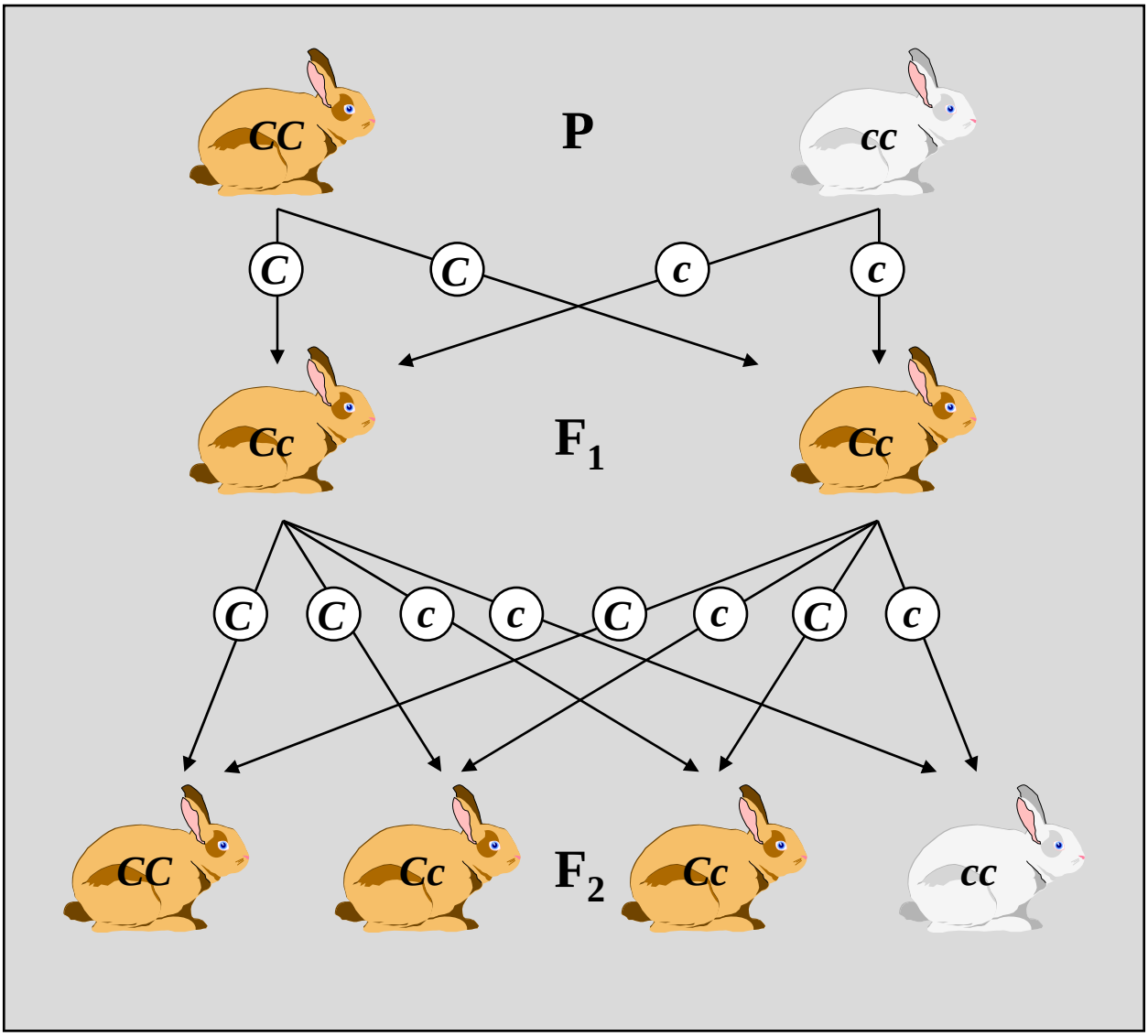
Human Genome Project

Informatics:

- **Develop effective software and data-base designs to support large-scale mapping and sequencing projects.**
- **Create database tools that provide easy access to up-to-date physical mapping, genetic mapping, chromo-some mapping, and sequencing information and allow ready com-parison of the data in these several data sets.**
- **Develop algorithms and analytical tools to interpret genomic information.**

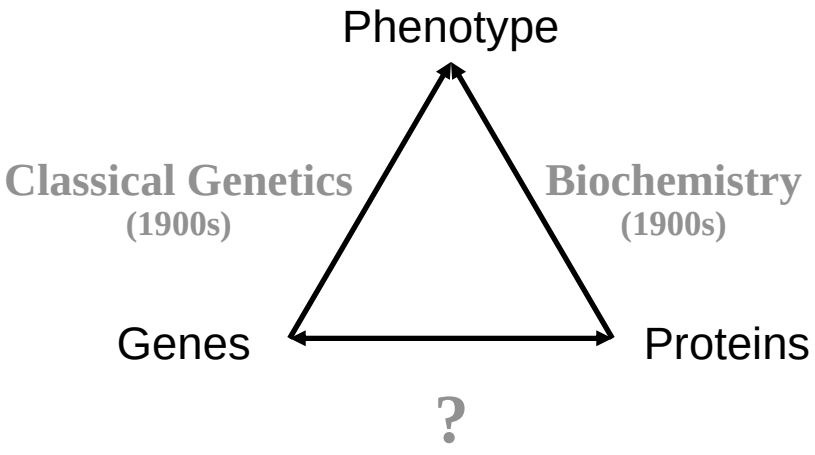
USDOE. 1990. Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years.

Classical Genetics

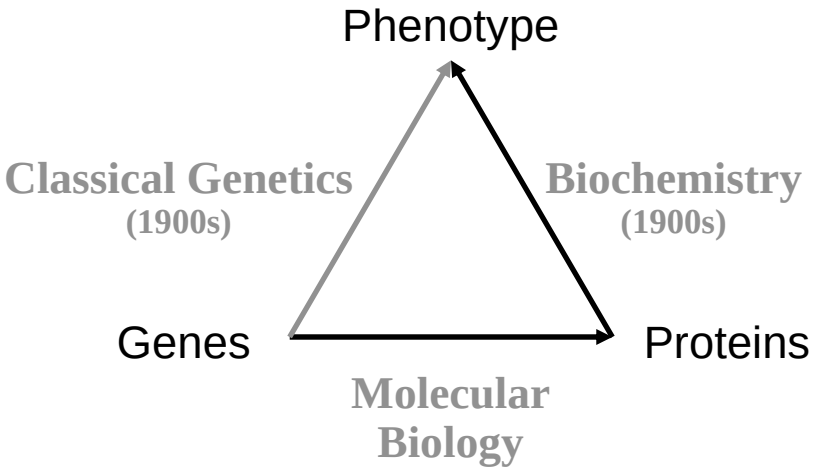


Regular numerical patterns of inheritance showed that the passage of traits from one generation to the next could be explained with the assumption that hypothetical particles, or ***genes***, were carried in pairs in adults, but transmitted individually to progeny.

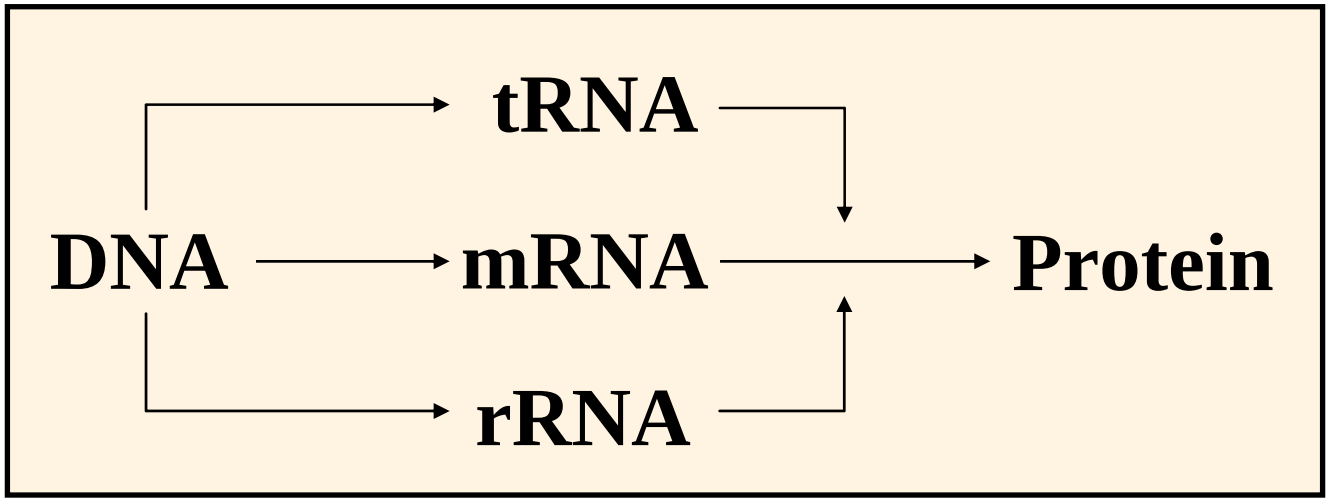
Origins of Molecular Biology



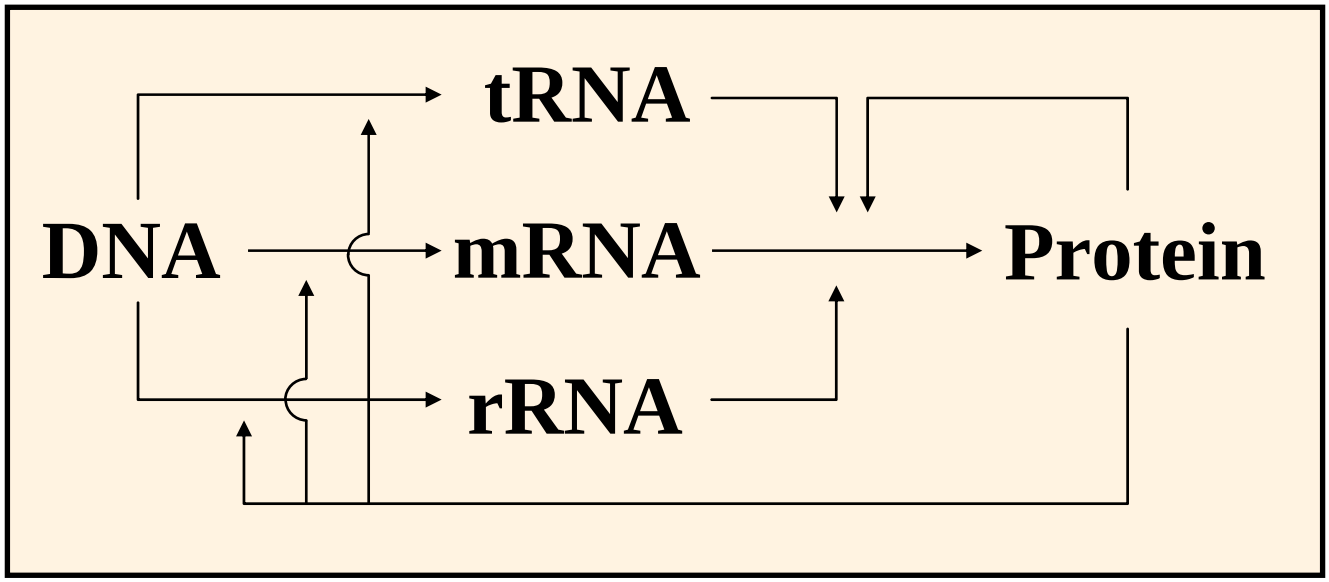
The ***phenotype*** of an organism denotes its external appearance (size, color, intelligence, etc.). ***Classical genetics*** showed that genes control the transmission of phenotype from one generation to the next. ***Biochemistry*** showed that within one generation, ***proteins*** had a determining 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.



The Fundamental Dogma

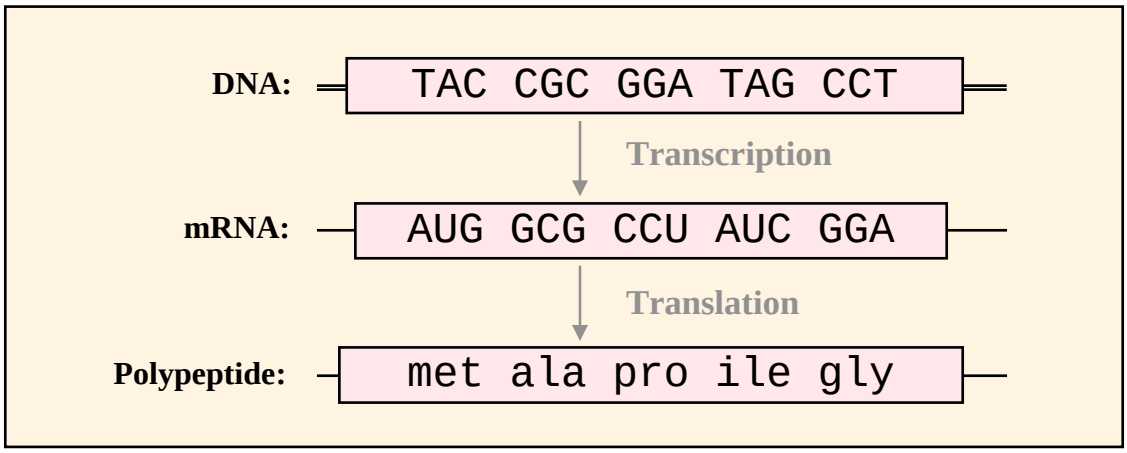


Genes are made of ***deoxyribonucleic acid***, or DNA. DNA is a linear string of four different kinds of subunits called ***nucleotides***. ***Proteins*** are linear strings of 20 different subunits called ***amino acids***. DNA controls the synthesis of proteins in a cell by digitially encoding their amino-acid sequences in the four-letter alphabet of DNA. This information is passed indirectly through intermediate molecules known as ***ribonucleic acids***, or RNA.

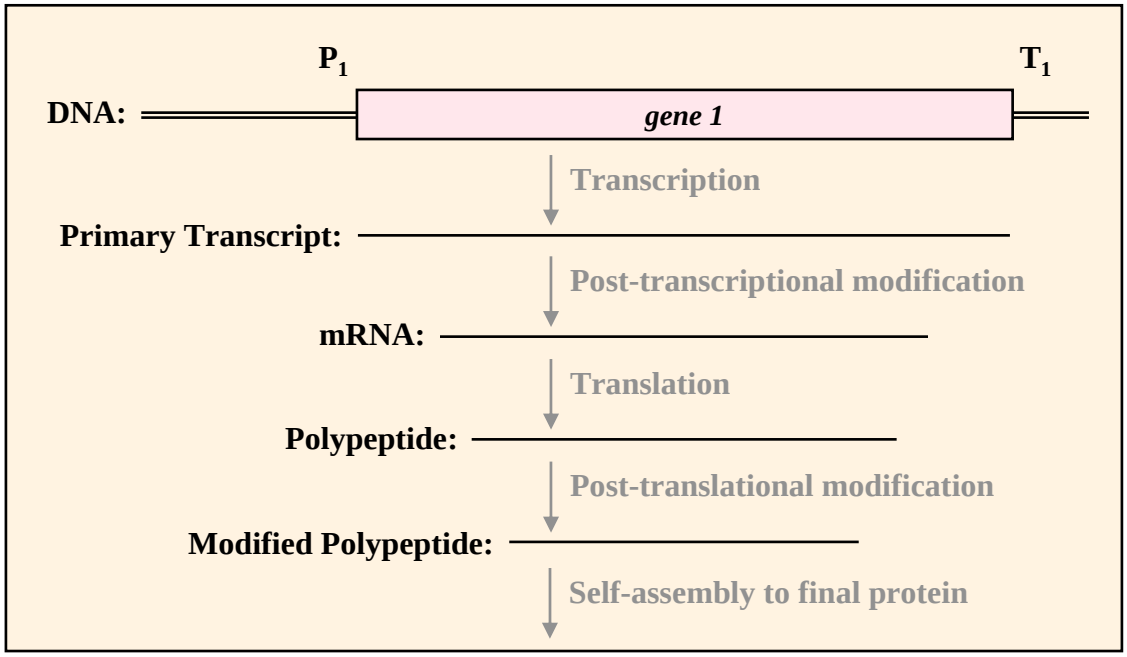


Proteins are responsible for most of the activities inside a living cell, including the synthesis of more DNA.

The Fundamental Dogma



DNA directs protein synthesis through a multi-step process. First, DNA is copied to mRNA. Then the mRNA is translated to produce a protein 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

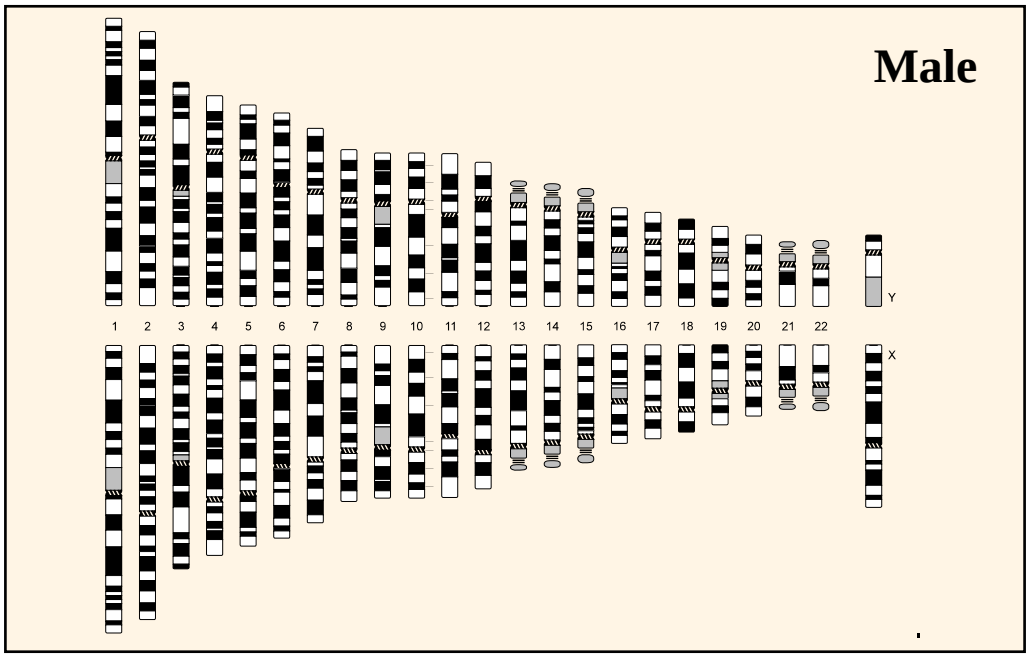
5´	U	C	A	G	3´
U	phe	ser	tyr	cys	U
	phe	ser	tyr	cys	C
	leu	ser	STOP	STOP	A
	leu	ser	STOP	trp	G
C	leu	pro	his	arg	U
	leu	pro	his	arg	C
	leu	pro	gln	arg	A
	leu	pro	gln	arg	G
A	ile	thr	asn	ser	U
	ile	thr	asn	ser	C
	ile	thr	lys	arg	A
	met	thr	lys	arg	G
G	val	ala	asp	gly	U
	val	ala	asp	gly	C
	val	ala	glu	gly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

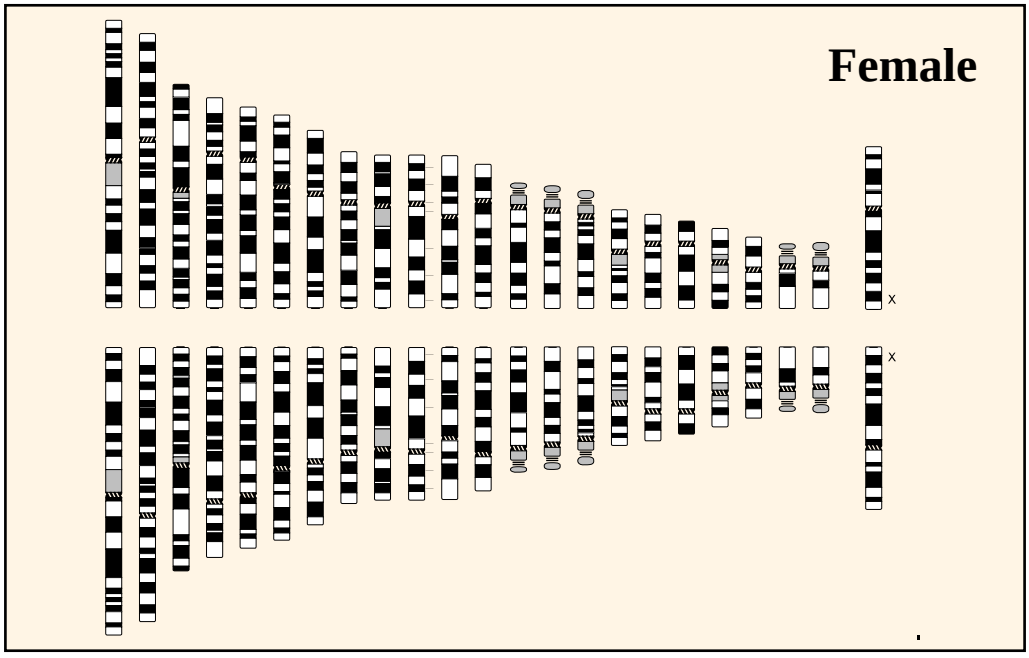
For example, the meaning of the codon 5'AUG3' is determined as follows:

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.

Human Chromosomes



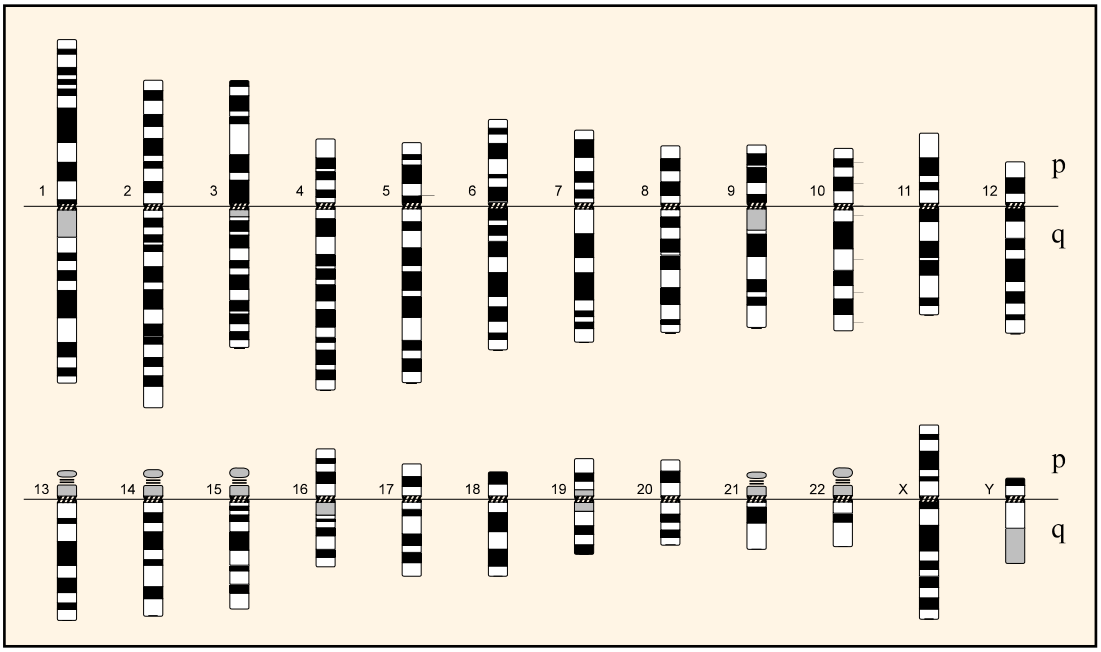
At conception, a normal human receives 23 chromosomes from each parent -- 22 ***autosomes*** and one ***sex chromosome***. The mother always contributes 22 autosomes and one ***X chromosome***. If the father also contributes an X chromosome, the child will be female. If the father contributes a ***Y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



Human Chromosomes

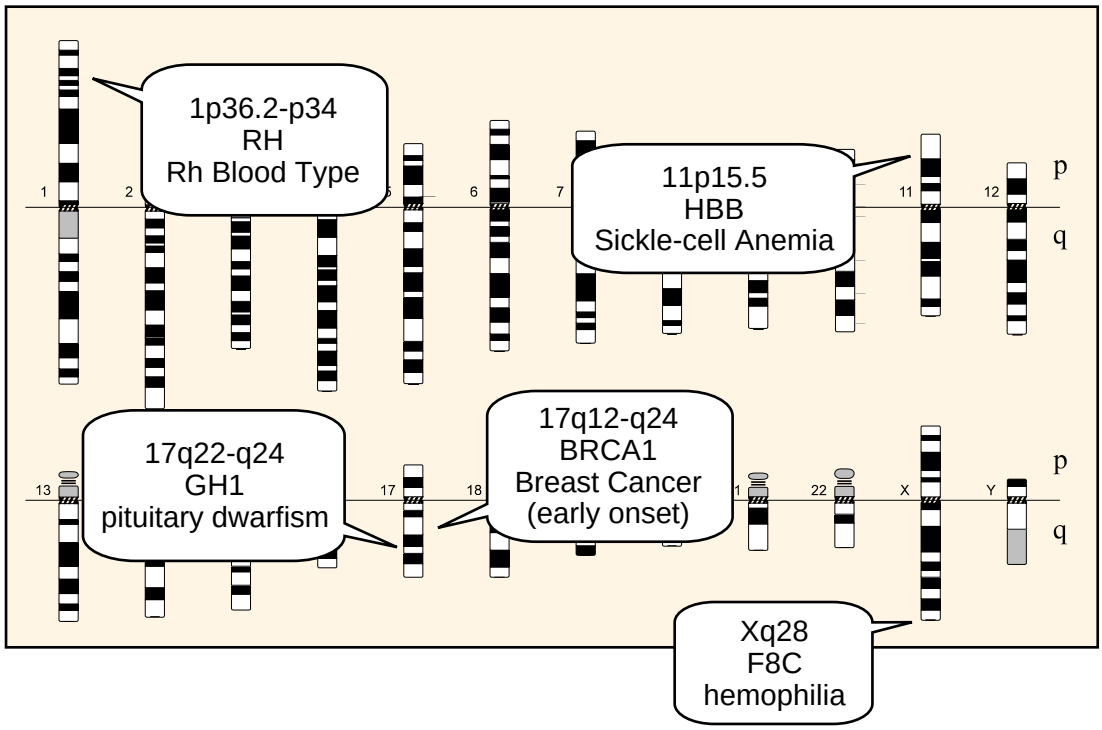
The human genome is believed to consist of 50,000 to 100,000 genes encoded in 3.3 billion base pairs of DNA, which are packaged into 23 chromosomes. The goal of the Human Genome Project (HGP) is learning the specific order of those 3.3 billion base pairs and of identifying and locating all of the genes 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 Genome 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. Mapping the genome is equivalent to obtaining a file allocation table for the device. Understanding the genome is equivalent to reverse engineering that unknown computer system all the way back to a full set of design and maintenance specifications.

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 GGCAGGttgg tatcaaggtt acaagacagg tttaaggaga ccaatagaaa ctgggcatgt
 301 ggagacagag aagactcttg ggtttctgat aggcactgac tctctctgcc tattggtcta
 361 ttttcccacc cttaggCTGC TGGTGGTCTA CCCTTGGACC CAGAGGTTCT TTGAGTCCTT
 421 TGGGGATCTG TCCACTCCTG ATGCTGTTAT GGGCAACCCT AAGGTGAAGG CTCATGGCAA
 481 GAAAGTGCTC GGTGCCTTTA GTGATGGCCT GGCTCACCTG GACAACCTCA AGGGCACCTT
 541 TGCCACACTG AGTGAGCTGC ACTGTGACAA GCTGCACGTG GATCCTGAGA ACTTCAGGgt
 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 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
```

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

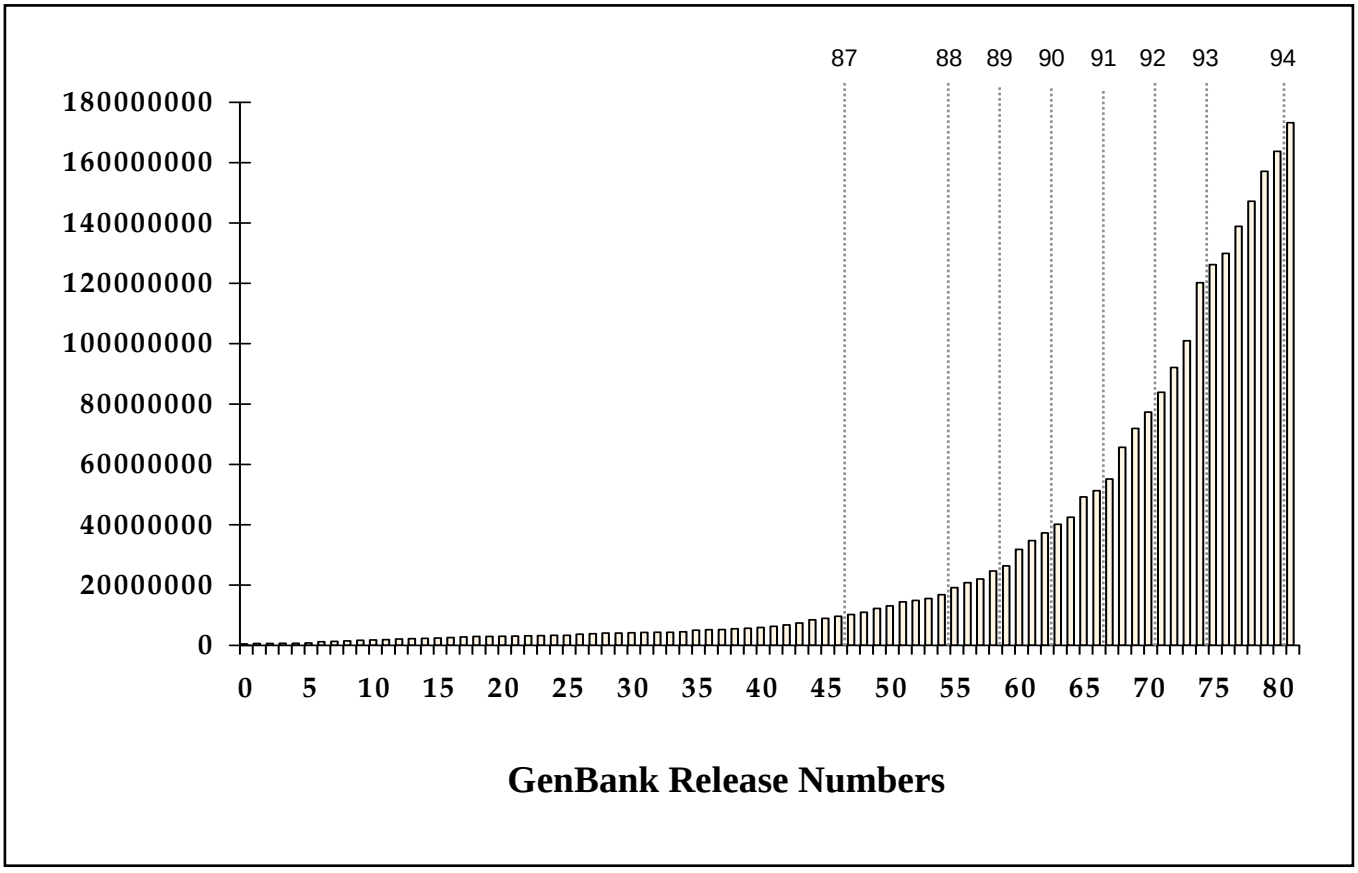
```
   1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
  61 ccagggctgg gcataaaagt cagggcagag ccatctattg ctt acatttg cttctgacac
 121 aactgtgttc actagcaacc tcaaacagac accATGGTGC ACCTGACTCC TGAGGAGAAG
 181 TCTGCCGTTA CTGCCCTGTG GGGCAAGGTG AACGTGGATG AAGTTGGTGG TGAGGCCCTG
 241 GGCAGGttgg tatcaaggtt acaagacagg tttaaggaga ccaatagc  ctgggcatgt
 301 ggagacagag aagactctt                                  gtcta
 361 ttttcccacc cttaggCT                                  CCTT
 421 TGGGGATCTG TCCACTCC                                  GCAA
 481 GAAAGTGCTC GGTGCCTT                                  CCTT
 541 TGCCACACTG AGTGAGCT                                  AGGgt
 601 gagtctatgg gacccttg                                  tgtca
 661 taggaagggg agaagtaa                                  catc
 721 agtgtggaag tctcaggat                                 gtttt
 781 cttttgttta attcttgctt tctttttttt tcttctccac 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.

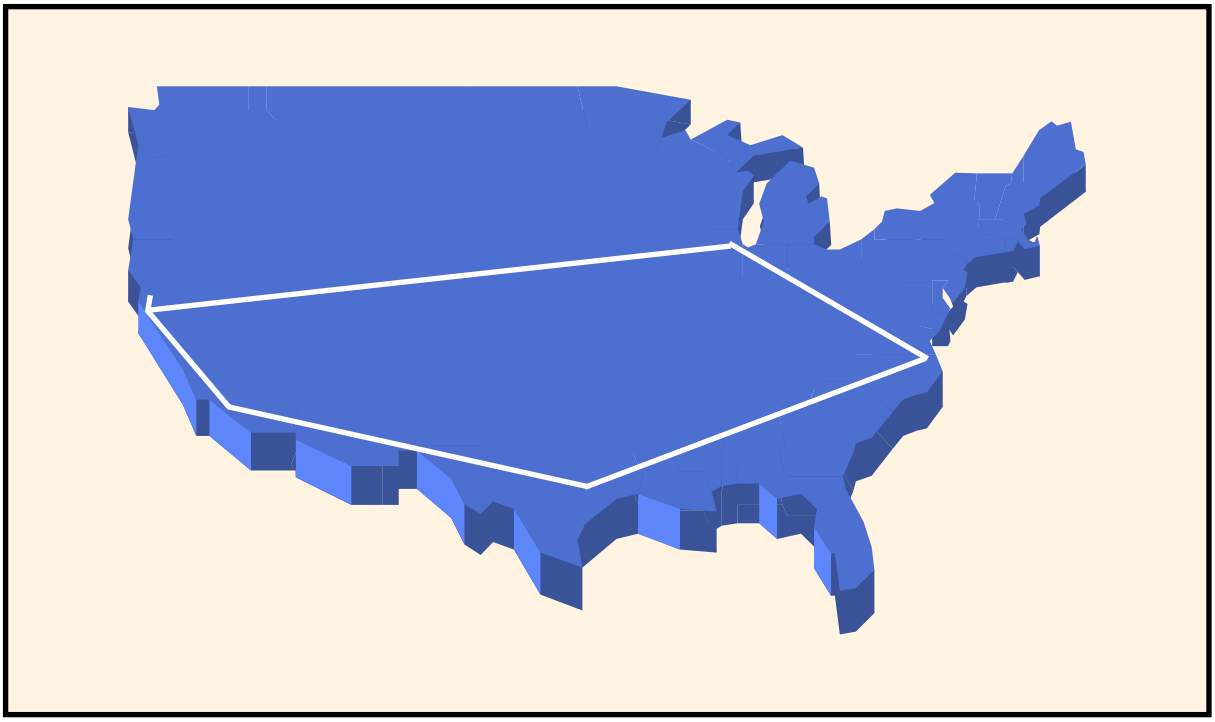
A change in this nucleic acid from an A to T causes glutamic acid to be replaced with valine. This produces the sickle-cell allele.

```
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
```

Cumulative Totals
GenBank Entries in Base Pairs

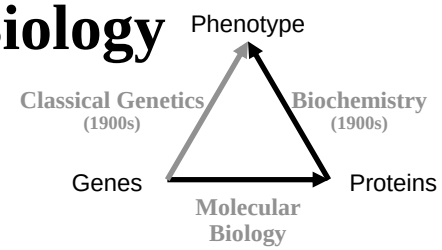


One Human Sequence



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%

Origins of Molecular Biology



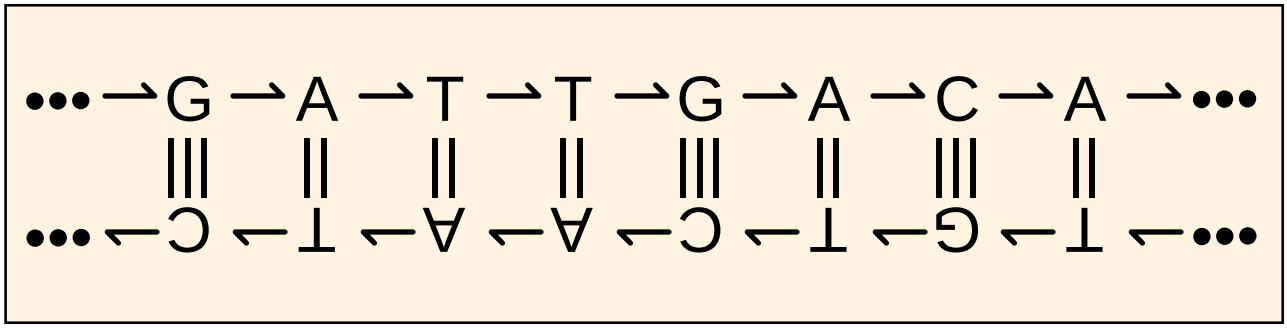
Key Discoveries:

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.

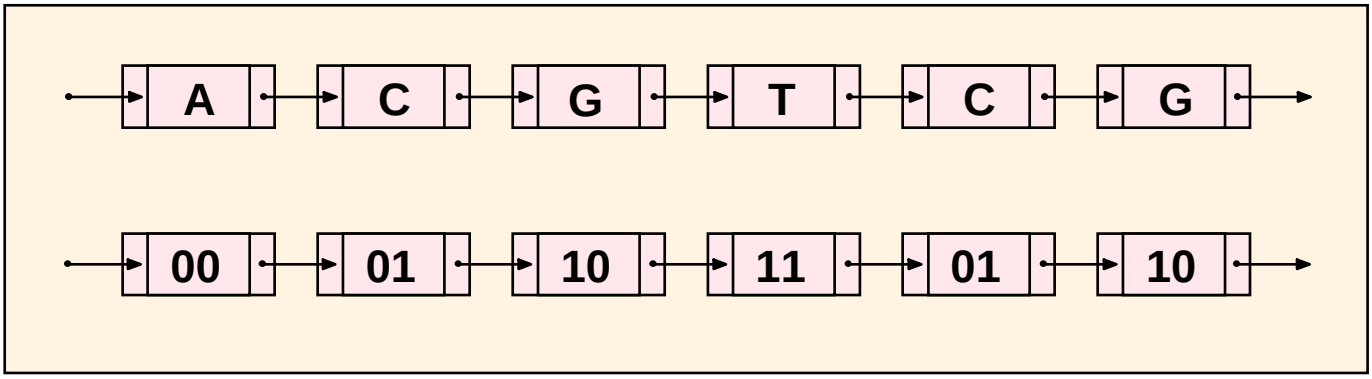
1950 The tetranucleotide hypothesis of DNA structure is overthrown.

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



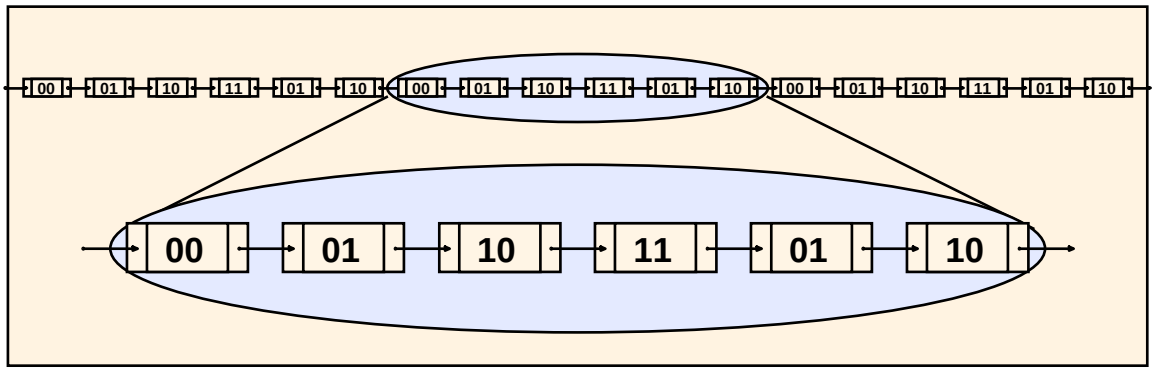
DNA is constructed as a double-stranded molecule, with absolutely no constraints upon the liner order of subcomponents along each strand, but with the pairing between strands totally constrained according to complementarity rules: A always pairs with T and C always pairs with G.

DNA As a Mass-storage Device



DNA is a linked list of nucleotides. Therefore, it may be represented as a linked list of bit patterns.

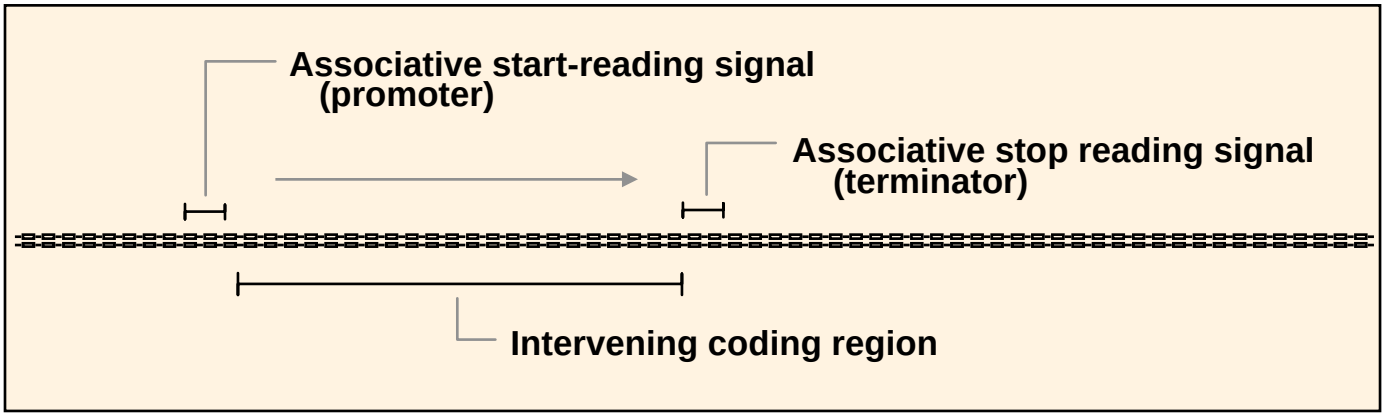
DNA As a Mass-storage Device



Mass Storage System:

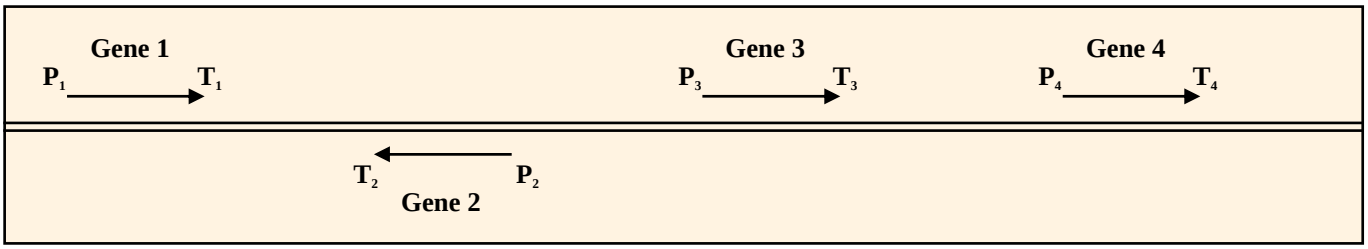
- Underlying primitive structure is linked list, not physical medium.
- List has polarity.
- Addressing is associative, not physical.
- No content restriction on linear order of data.

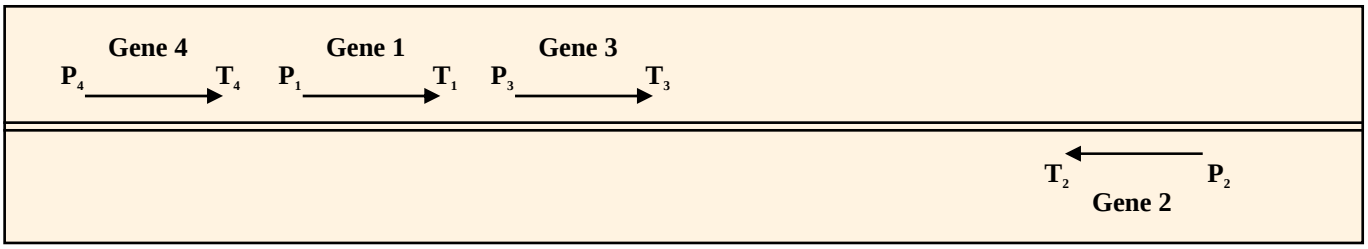
DNA As a Mass-storage Device



Individual coded objects in DNA are identified and “executed” via their associative start and stop signals, not through their actual physical location in the genome.

DNA As a Mass-storage Device





Because individual coded objects in DNA are located associatively, these two versions of the genome would be functionally identical.

Reverse Engineering Codes

WARMBOOT:

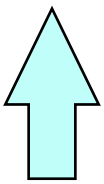
```
BA 40 00 8E DA BB 72 00 C7 07 00 00 EA 00 00 FF FF
```

COLDBOOT:

```
BA 40 00 8E DA BB 72 00 C7 07 34 12 EA 00 00 FF FF
```

ALIGNMENT:

```
BA 40 00 8E DA BB 72 00 C7 07 00 00 EA 00 00 FF FF
BA 40 00 8E DA BB 72 00 C7 07 34 12 EA 00 00 FF FF
```



Reverse Engineering Codes

Assume that you have the executables for four short programs, each of which causes a short message to be written to the screen:

1 = Hello world

2 = Hi world

3 = Goodbye world

4 = Hello

```
EB 0D 90 48 65 6C 6C 6F 20 77 6F 72 6C 64 24 B4 00 B4 09 BA 03 01 CD 21 C3
```

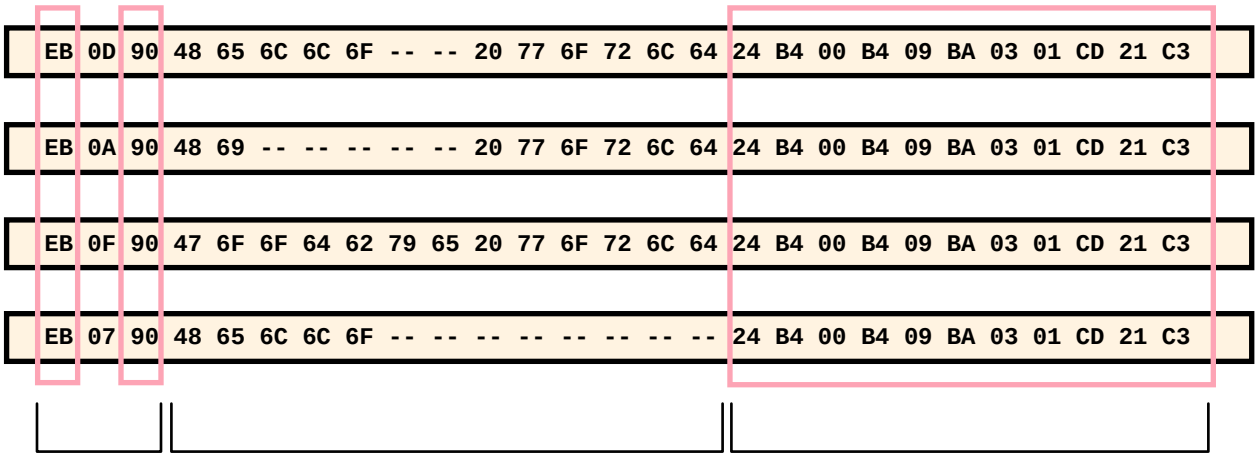
```
EB 0A 90 48 69 20 77 6F 72 6C 64 24 B4 00 B4 09 BA 03 01 CD 21 C3
```

```
EB 0F 90 47 6F 6F 64 62 79 65 20 77 6F 72 6C 64 24 B4 00 B4 09 BA 03 01 CD 21 C3
```

```
EB 07 90 48 65 6C 6C 6F 24 B4 00 B4 09 BA 03 01 CD 21 C3
```

Reverse Engineering Codes

Aligning the sequences (inserting blanks where necessary) allows the detection of common features and even permits functional hypotheses to be developed.



Reverse Engineering Codes

Now, suppose you locate a fifth program, that writes the same "Hello world" message as did the first program, but which has different binaries. At first, the sequences appear fairly different:

```
EB 0D 90 48 65 6C 6C 6F 20 77 6F 72 6C 64 24 B4 00 B4 09 BA 03 01 CD 21 C3
```

```
EB 01 90 B4 00 B4 09 BA 0F 01 CD 21 EB 0D 90 48 65 6C 6C 64 20 77 6F 72 6C 6C 24 C3
```

Again, aligned sequence similarities provide the clues...

```
-- -- -- EB 0D 90 48 65 6C 6C 6F 20 77 6F 72 6C 64 24 B4 00 B4 09 BA 03 01 CD 21 C3
```

```
EB 01 90 B4 00 B4 09 BA 0F 01 CD 21 EB 0D 90 48 65 6C 6C 64 20 77 6F 72 6C 6C 24 C3
```

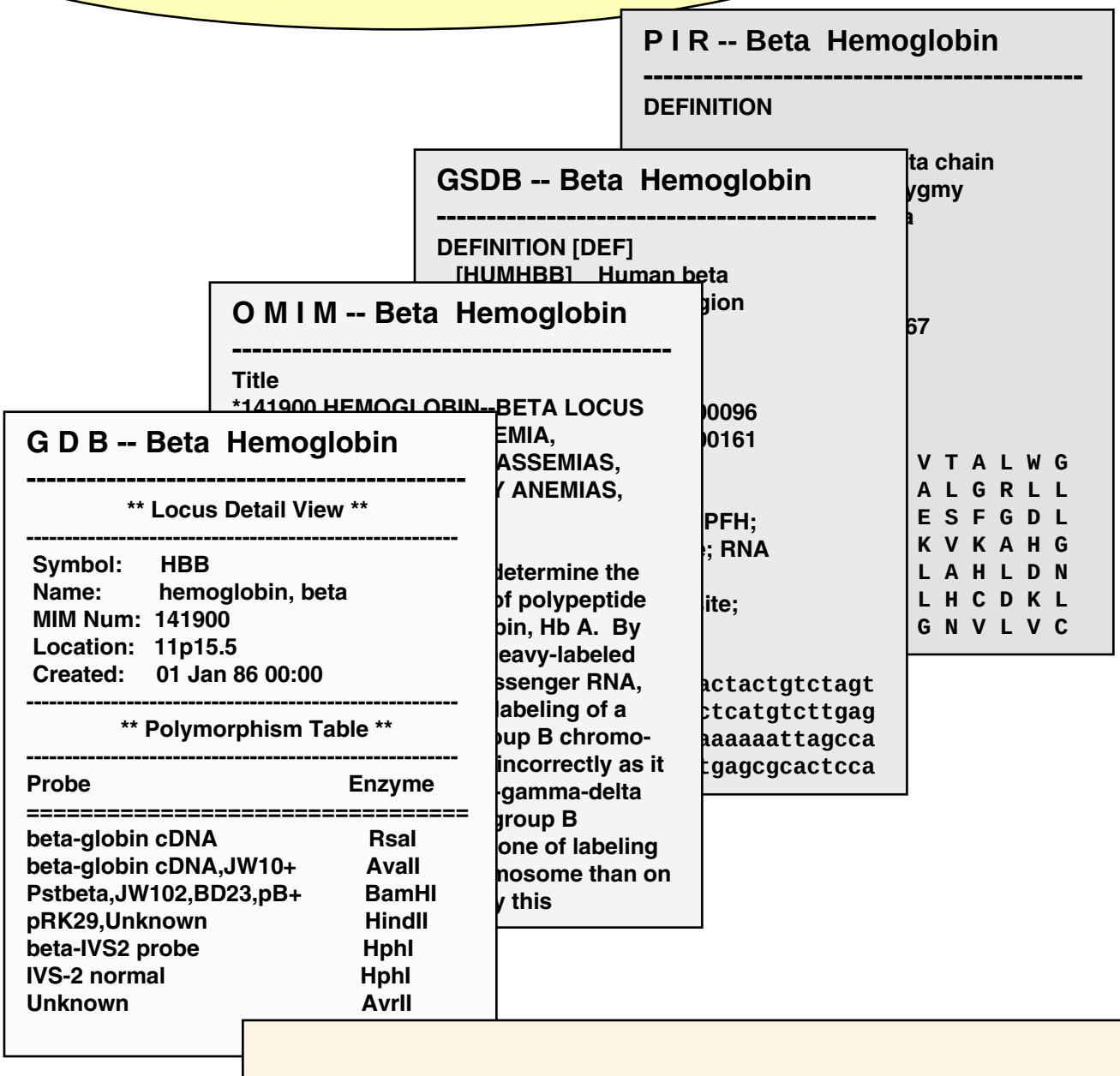
Reverse Engineering Codes

gene name	DNA sequence near transcription initiation site				
lacZ	---ccaggc	TTtACA	ctttatgcttccggctcg-	TATgtT	--gtgtgga -
malT	--tcatcgc	TTGcat	tagaaaggtttctggcc--	gAcctT	--ataacca -
araC	--atccatg	TgGACt	tttctgccgtgattata--	gAcAcT	tttgttacg -
galP1	----catgt	cacACt	tttcgcatctttgttatgc	TATggT	--tatttca -
deoP2	----gtgta	TcGAag	tgtgttgcggagtagatgt	TAgAAT	--actaaca -
cat	gatcggcac	gtaAgA	ggttccaactttcac----	cATAAT	-gaaataag -
tnaA	-ttcagaa	TaGACA	aaaactctgagtgtaa---	TAatgT	--agcctcg -
araE	----ccgac	cTGACA	cctgcgtgagttgttcacg	TATttT	ttcactatg -
consensus	---------	TTGACA	------------------	TATAAT	---------- -

-35 -10

Alignments allow functional analysis of trans-cription-initiation sites (i.e., associative start-read signals). Biological op-codes prove to be probabilistic, not deterministic.

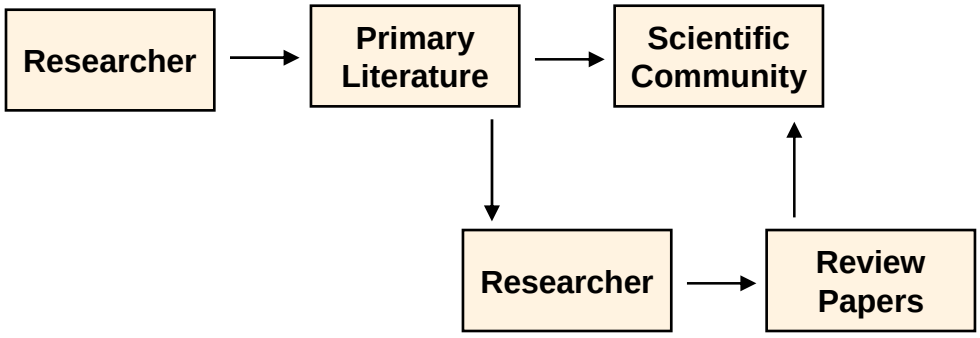
Human Genome



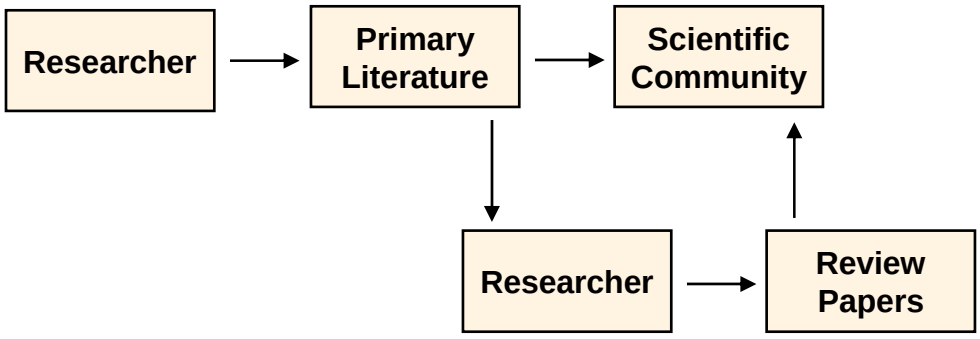
Knowledge Management through Electronic Data Publishing

Databases as Publishing

Traditional Publishing

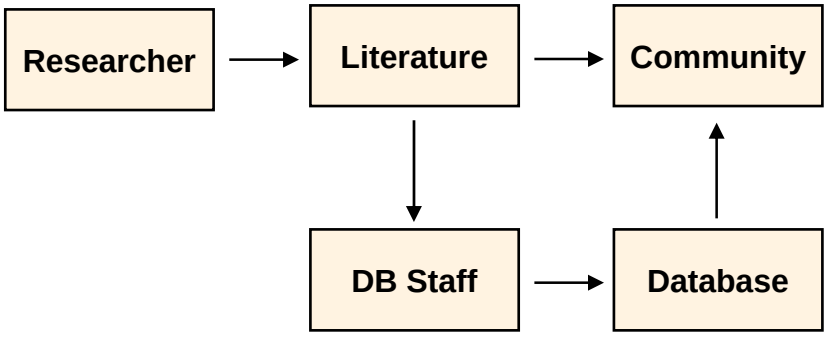


Early Database Development

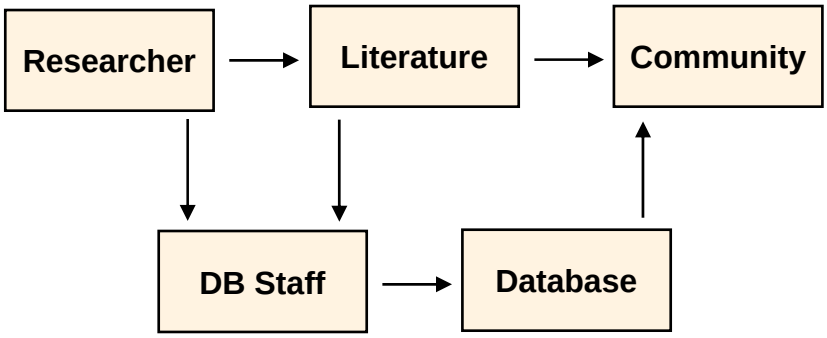


Electronic Data Publishing

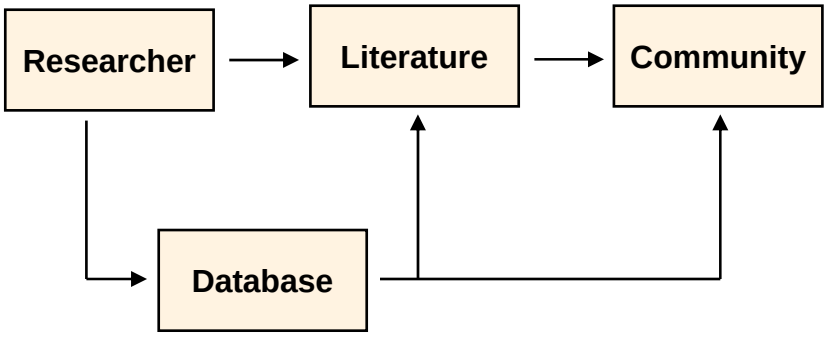
Standard Database Development



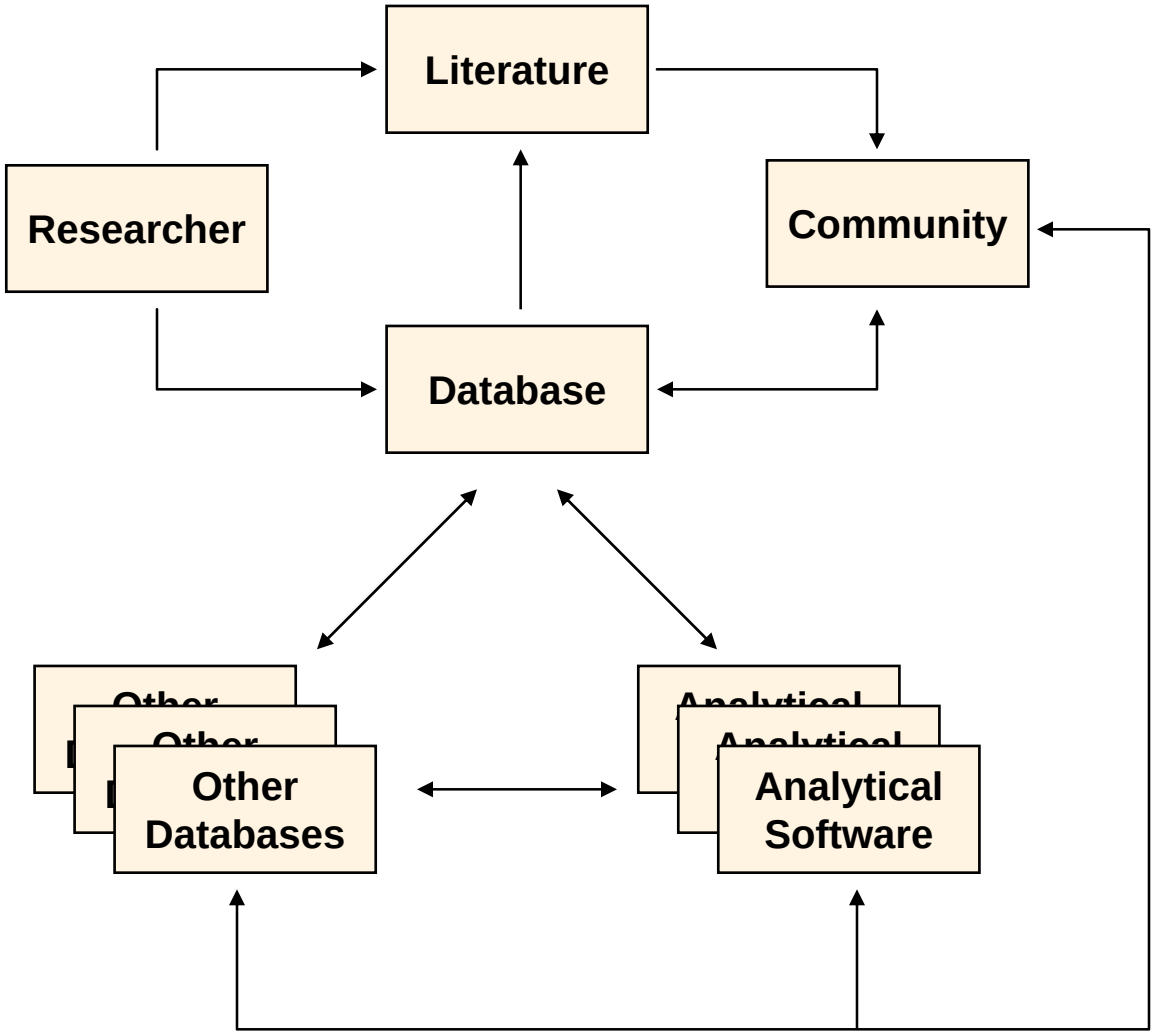
Early Electronic Data Publishing



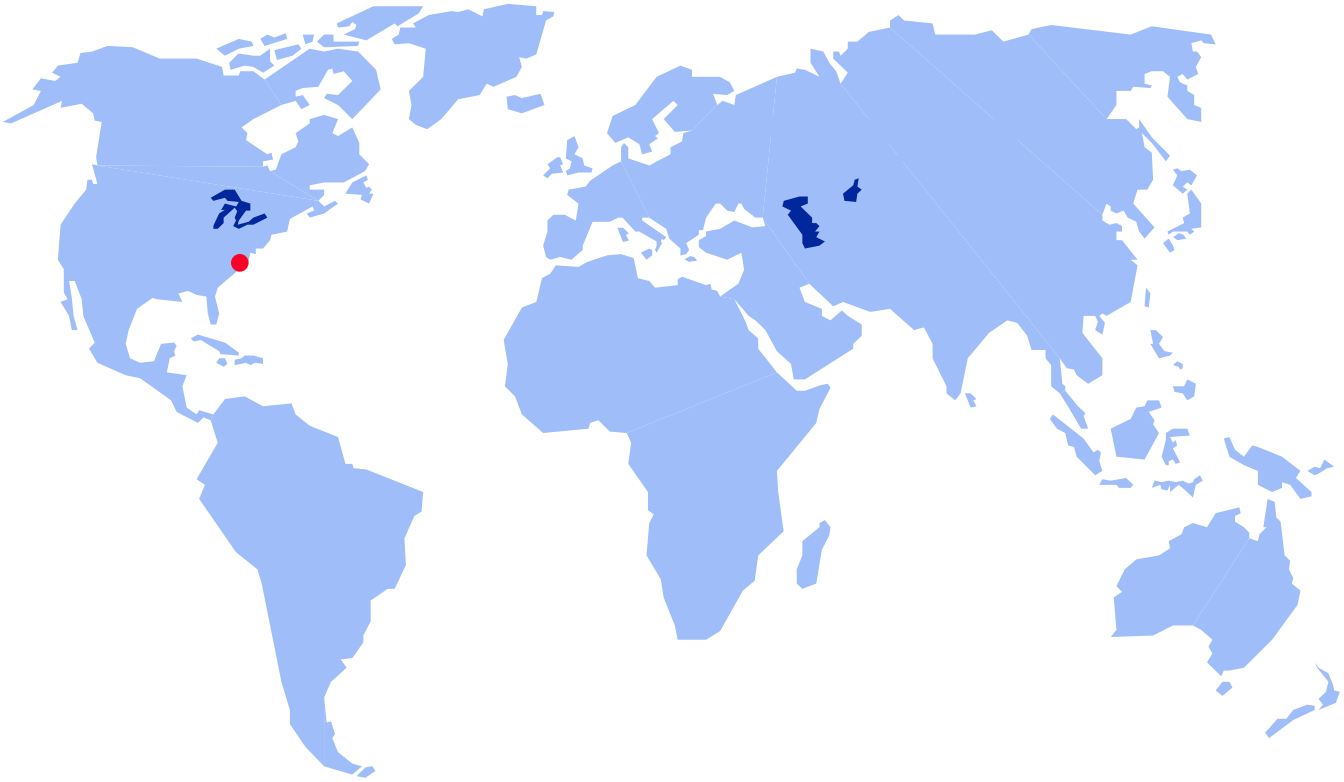
Electronic Data Publishing



Electronic Data Publishing and Integrated Analysis



Genome Data Base



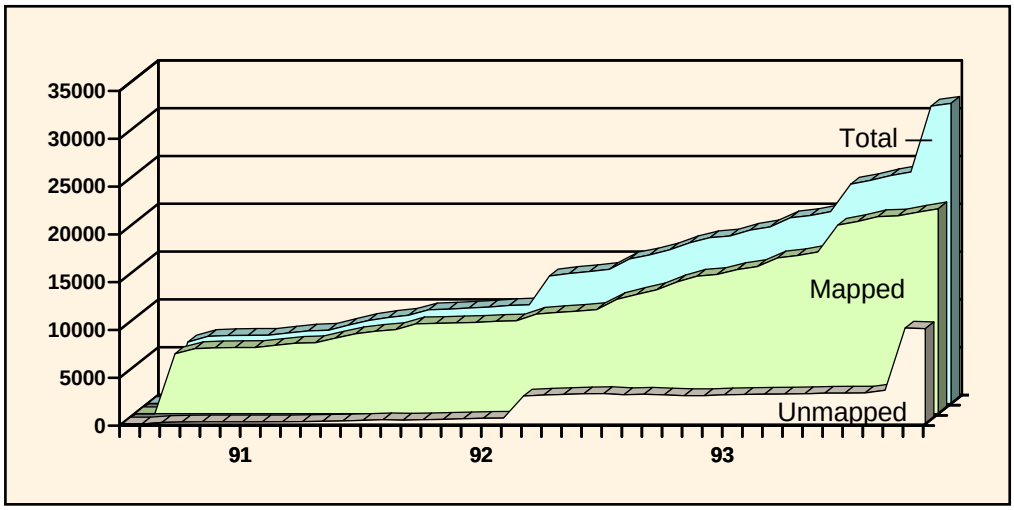
The Genome Data Base (GDB) contains information about human genes and genetic maps. GDB itself resides on a computer system at Johns Hopkins University in the United States. Scientists access the database on the Internet or by using dial-up connections.

Genome Data Base

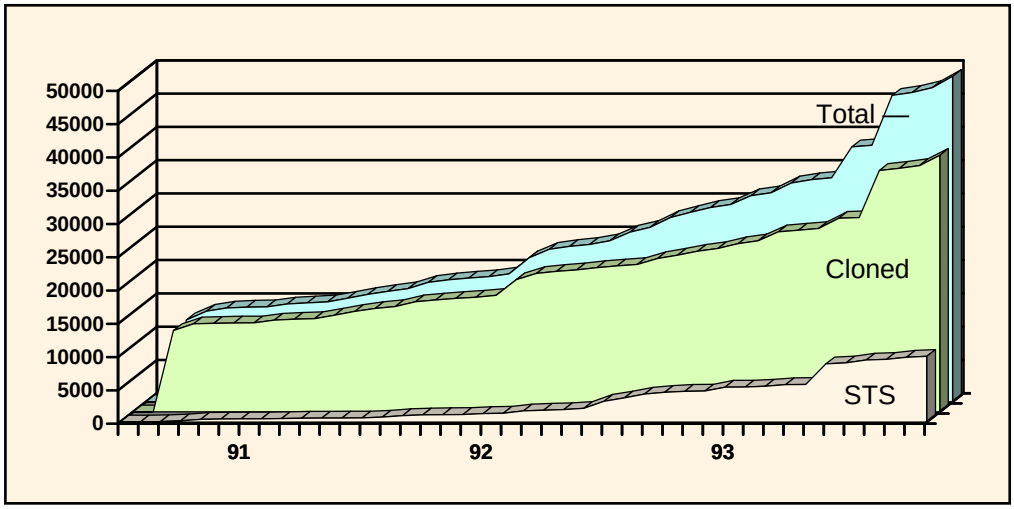
The establishment of multiple distribution sites provides scientists around the world much easier access to GDB data. However, for this to work well, every such site must offer an official, published copy of the master database in Baltimore.

GDB Content

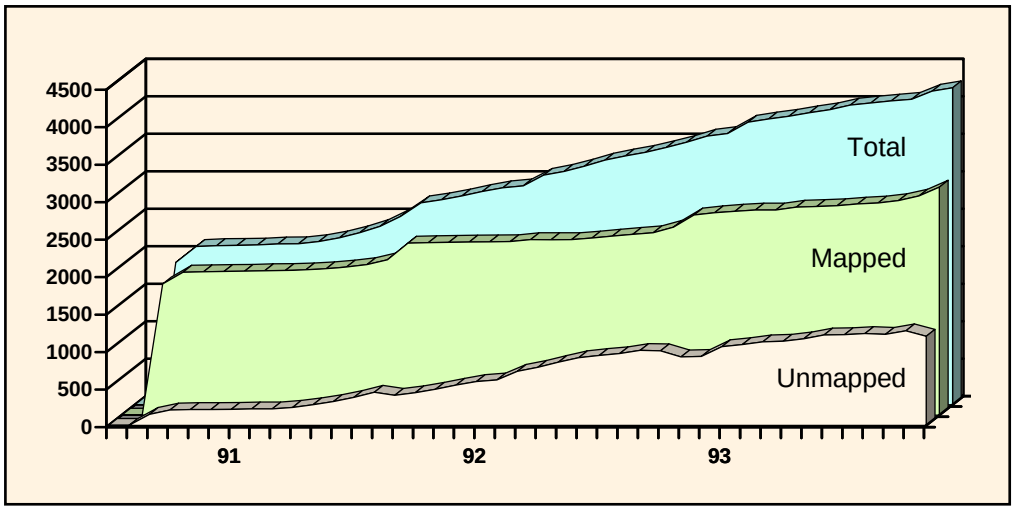
Loci



Probes

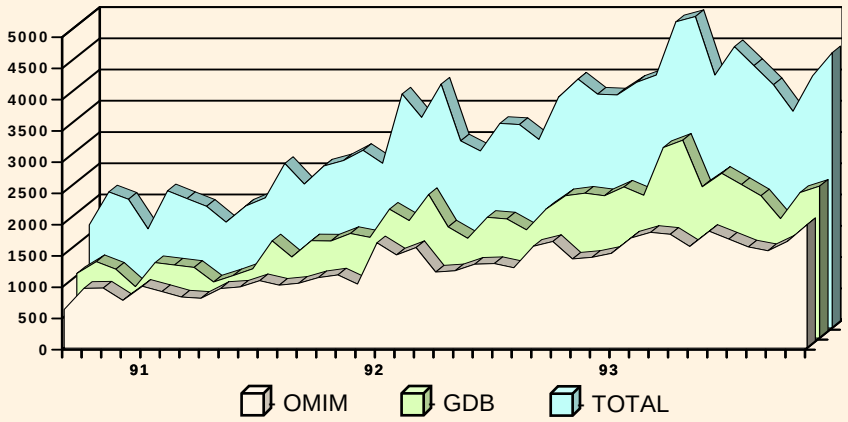


Genes



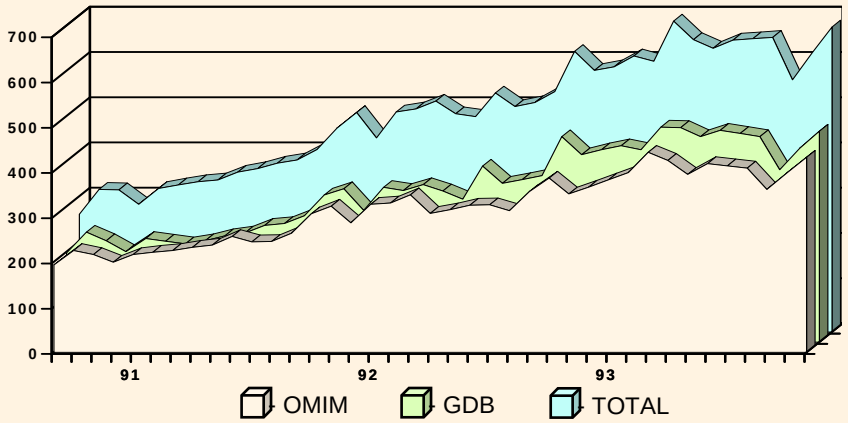
GDB Usage

Sessions



GDB/OMIM monthly usage statistics for sessions, where a “session” is defined as one login session of any duration.

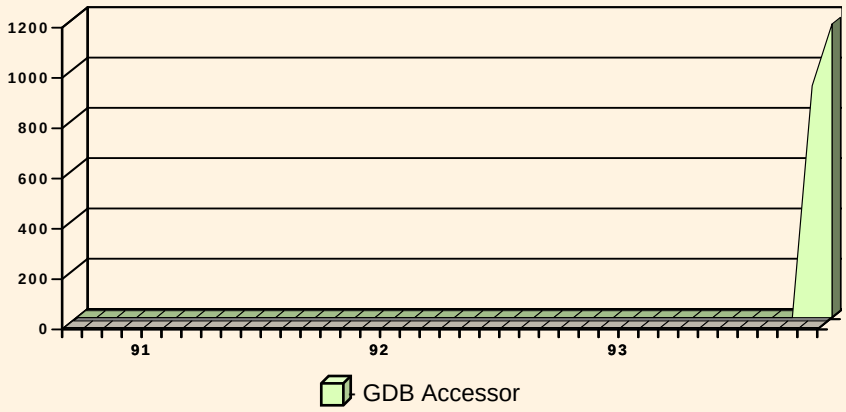
Unique Users



GDB/OMIM monthly usage statistics for unique users, where a “unique user” is defined as at least one login during the month for an individual user.

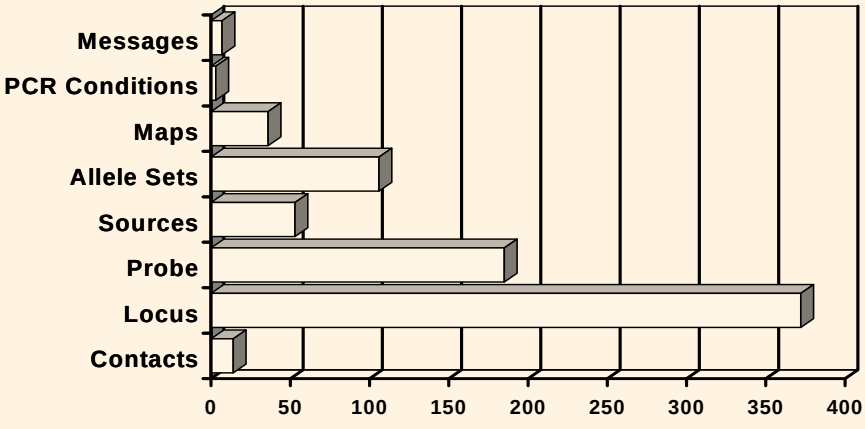
GDB Usage

Sessions (third-party software)



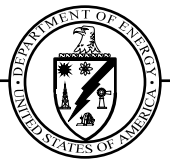
GDB/OMIM monthly usage statistics for sessions, where a “session” is defined as one login session of any duration.

Publishing on Demand



Over 750 custom reports were generated during the month of October, 1993.

Technical Impediments



Genome Informatics Summit Report

The success of the genome project will increasingly depend on the ease with which accurate and timely answers to interesting questions about genomic data can be obtained.

If repeating experiments becomes easier than locating previous results, genome informatics will have failed.

All extant community databases have serious deficiencies and fall short of meeting community needs.

An embarrassment to the Human Genome Project is our inability to answer simple questions such as, “How many genes on the long arm of chromosome 21 have been sequenced?"

Genome Informatics Summit Report

We must think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces.

Each database should be designed as a component of a larger information infrastructure for computational biology.

Adding a new database to the federation should be no more difficult than adding another computer to the Internet.

Successful HGP data management requires the development of a federated information infrastructure, with data flowing electronically over networks from producers to databases to users.

Genome Informatics Summit Report

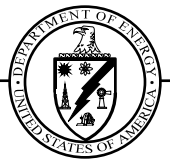
Any biologist should be able to submit research results to multiple appropriate databases with a single electronic transaction.

Professional data curators should be supported for community databases and, in addition, tools for direct author curation should be developed.

True, loss free data exchange can occur only if participating databases first achieve some kind of semantic parity.

When research advances change our perception of the real world, our databases must track the change or become inadequate.

Conceptual Impediments



Significant Errors

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.

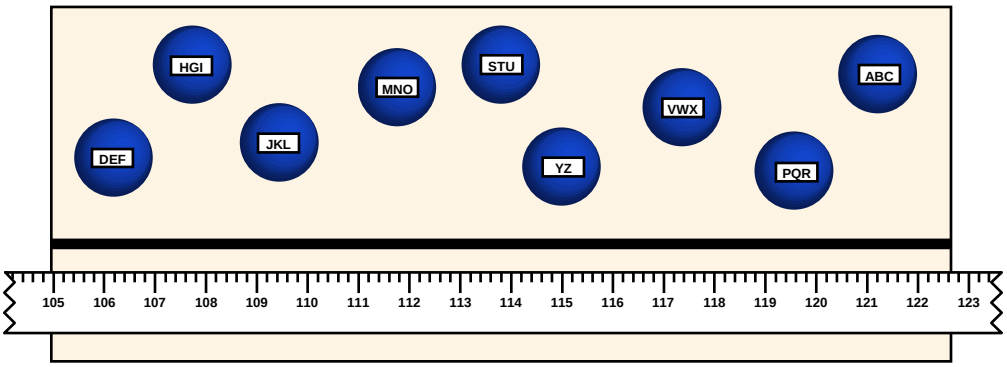
Fifty years from now it seems very likely that the most significant development of genetics in the current decade (1945-1955) will stand out as being the discovery of pseudoallelism.

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

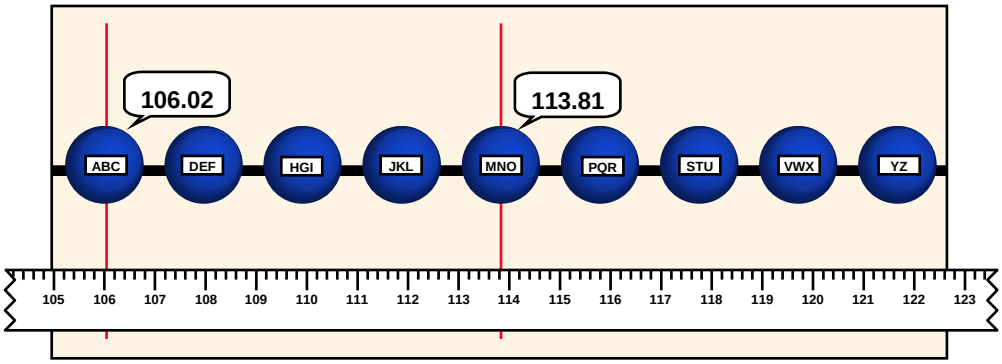
The ultimate ... map [will be] the complete DNA sequence of the human genome.

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

What is a Gene?



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

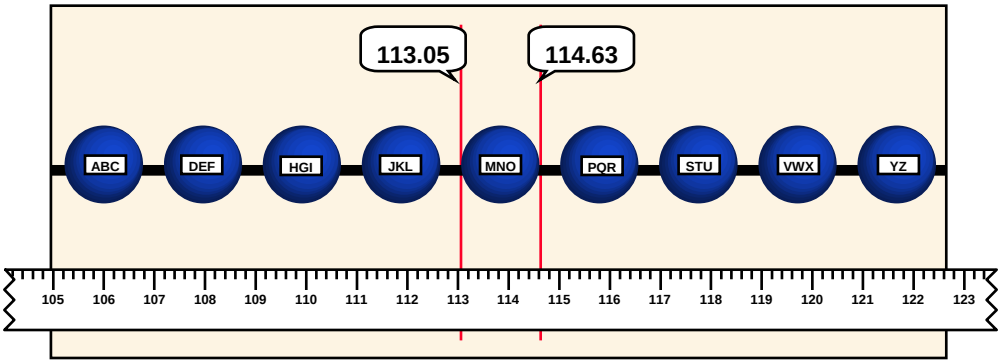


Mapping involves placing the beads in the correct order and assigning a correct address to each bead. The address assigned to a bead is its locus.

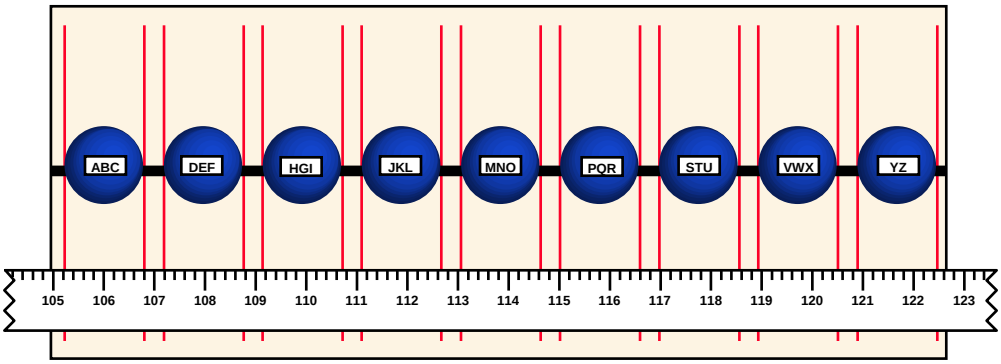
The genes are arranged in a manner similar to beads strung on a loose string.

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

What is a Gene?



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.

What is a Gene?

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.

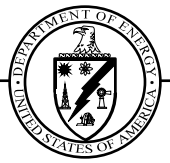
Current Definition: ???

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.

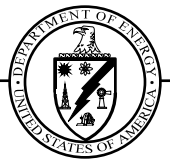
Lewin, Benjamin. 1990. *Genes IV*. Oxford University Press, New York.



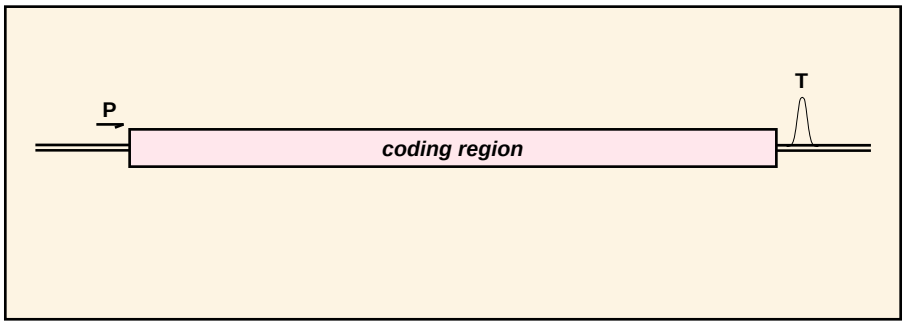
What is a Gene?

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.

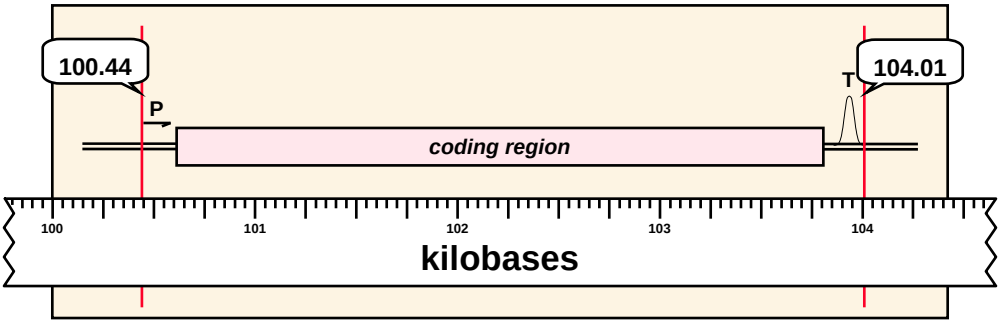
Singer, M., and Berg, P. *Genes & Genomes*. University Science Books, Mill Valley, California.



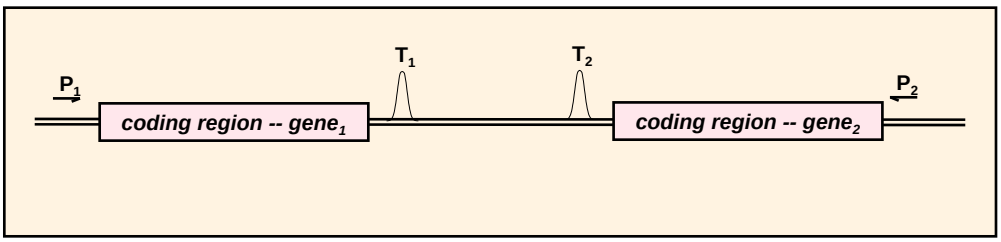
The Simplistic View of a Genome



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

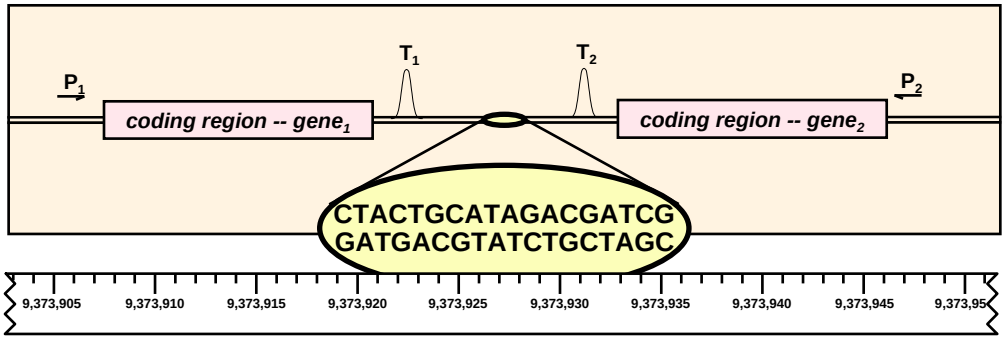


The location of a gene can be designated by specifying the base-pair location of its beginning and end.



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.

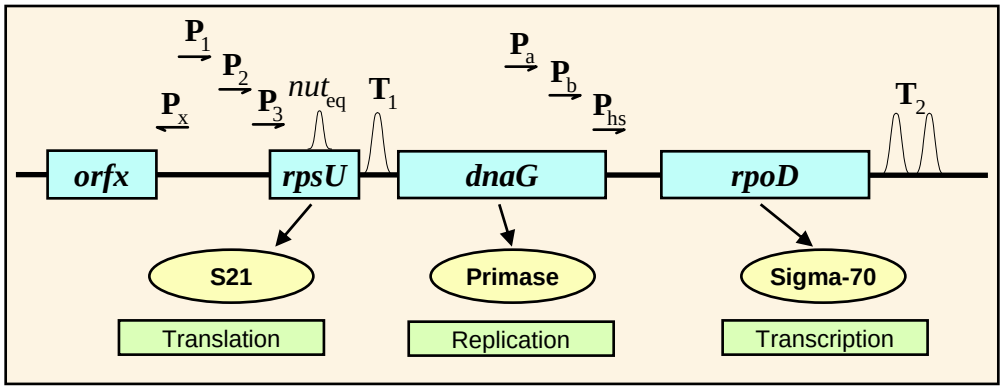
The Simplistic View of a Genome



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 human chromosomes may vary in length by tens of millions of nucleotides.

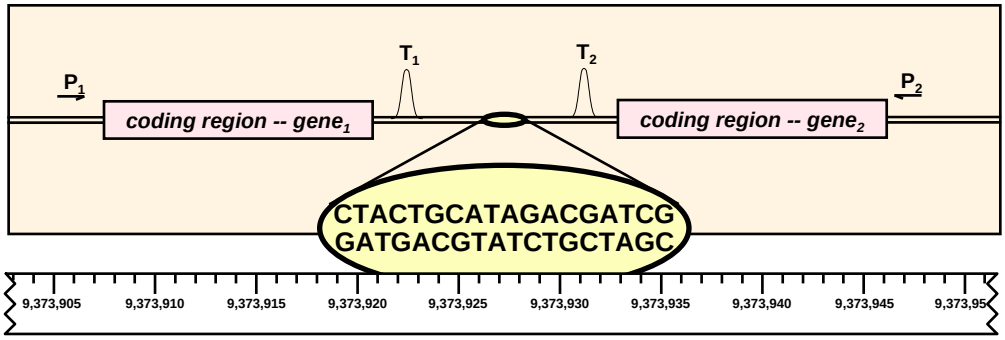
Complex Genomic Regions

Escherichia coli: the MMS Operon



Lupski, J.R., Godson, G.N., 1989, DNA→DNA, and DNA→RNA→Protein: Orchestration by a single complex operon, *BioEssays*, 10:152-157.

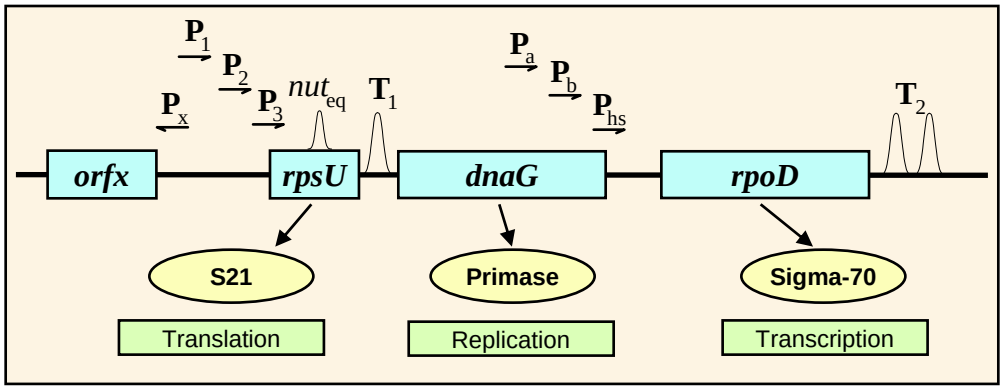
The Simplistic View of a Genome



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 human chromosomes may vary in length by tens of millions of nucleotides.

Complex Genomic Regions

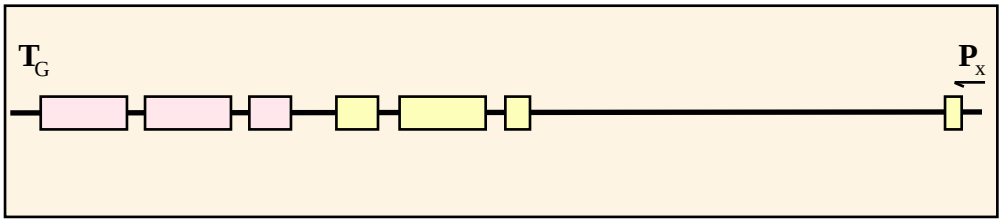
Escherichia coli: the MMS Operon



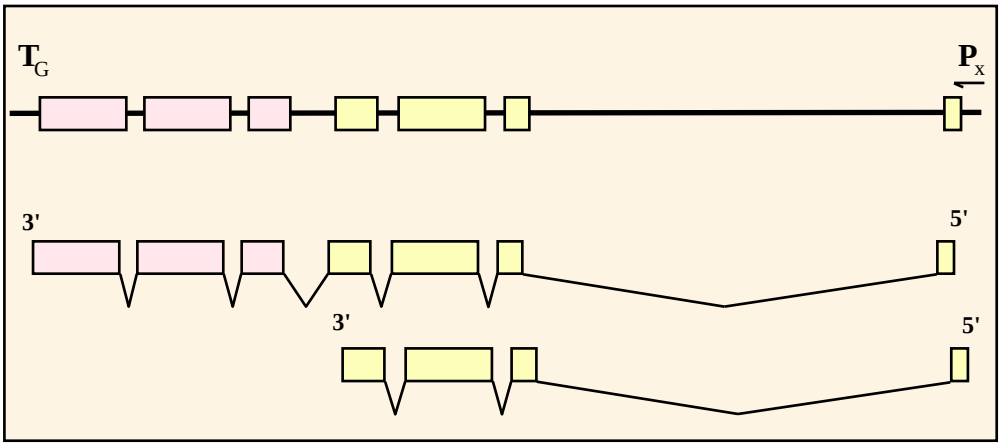
Lupski, J.R., Godson, G.N., 1989, DNA→DNA, and DNA→RNA→Protein: Orchestration by a single complex operon, *BioEssays*, 10:152-157.

Drosophila melanogaster: The Gart Locus

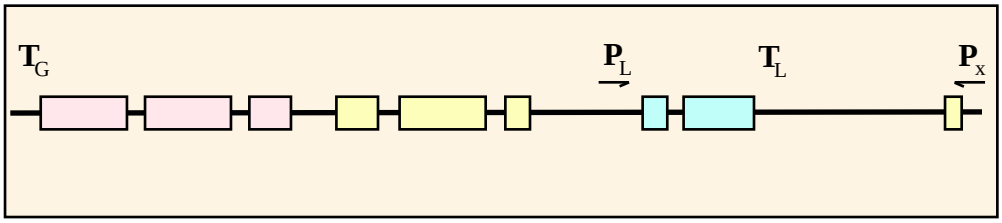
Fragmented Genes



Alternative Splicing



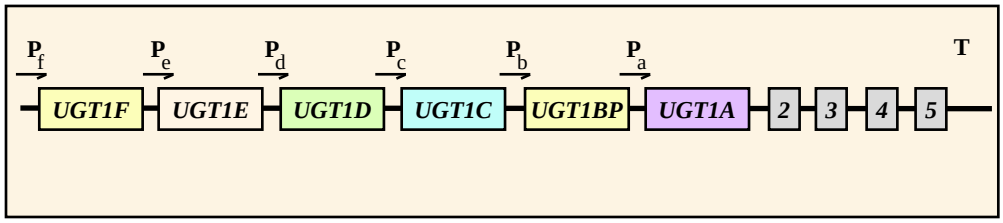
Nested Genes



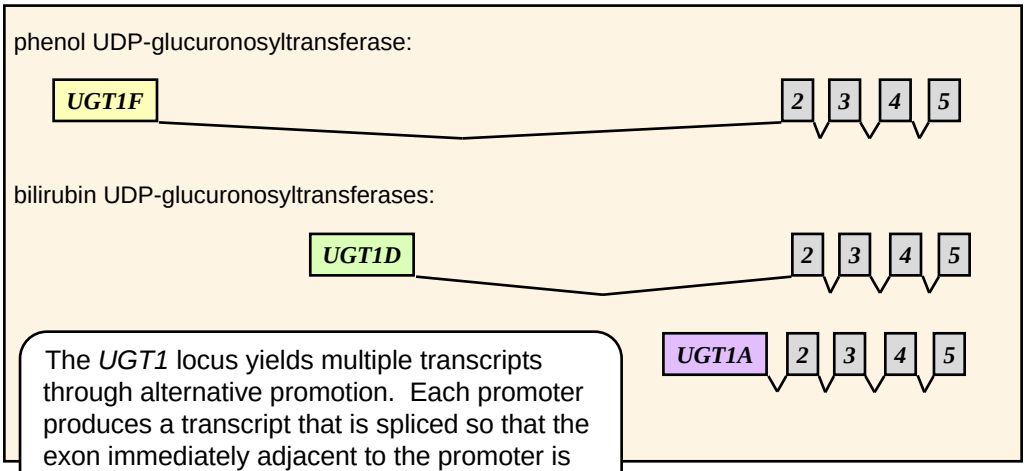
Henikoff, S., Keene, M.A., Fechtel, K., and Fristrom, J.W., 1986, Gene within a gene: Nested *Drosophila* genes encode unrelated proteins on opposite strands, *Cell* 44:33.

Nested Gene Families

Homo sapiens: The UGT1 Loci



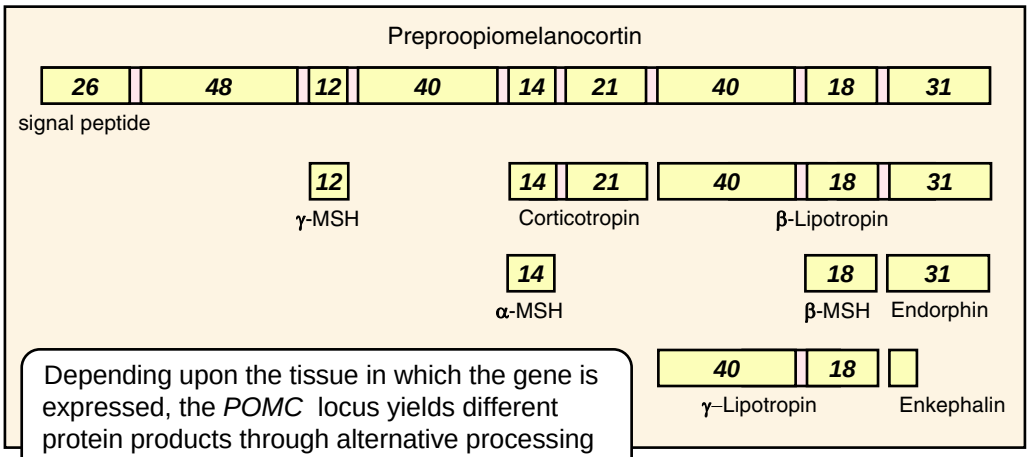
Ritter, J.K., Chen, F., et al., 1992, A novel complex locus *UGT1* encodes human bilirubin, phenol, and other UDP-glucuronosyltransferase isozymes with identical carboxyl termini, *J. Biol. Chem.* 267:3257.



The *UGT1* locus yields multiple transcripts through alternative promotion. Each promoter produces a transcript that is spliced so that the exon immediately adjacent to the promoter is joined with the four terminal exons shared by all of the transcripts.

Multiple Gene Products

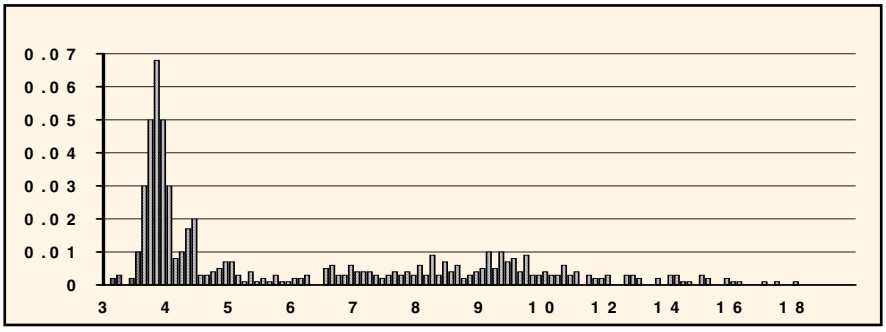
Homo sapiens: The POMC Locus



Depending upon the tissue in which the gene is expressed, the *POMC* locus yields different protein products through alternative processing of the resulting polypeptide.

VNTR Loci

D14S1: Frequency of PstI fragment sizes (kb)



Balazs, I., Neuweiler, J., Gunn, P., Kidd, J., Kidd, K.K., Kuhl, J., and Mingjun, L., 1992, Human population genetic studies using hypervariable loci, *Genetics*, 131:191-198.

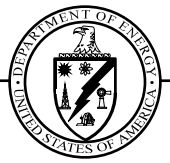
What is a Gene?

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.

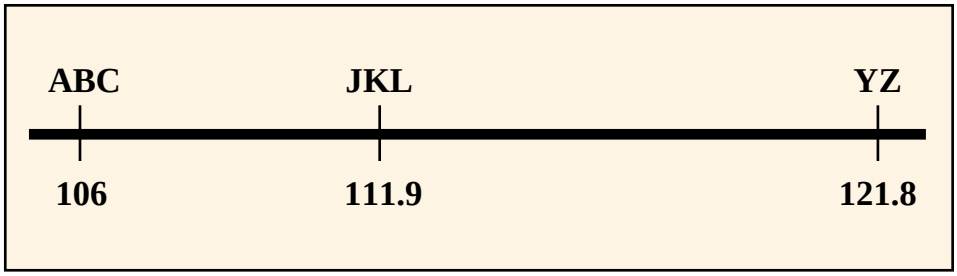
Singer, M., and Berg, P. *Genes & Genomes*. University Science Books, Mill Valley, California.

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.

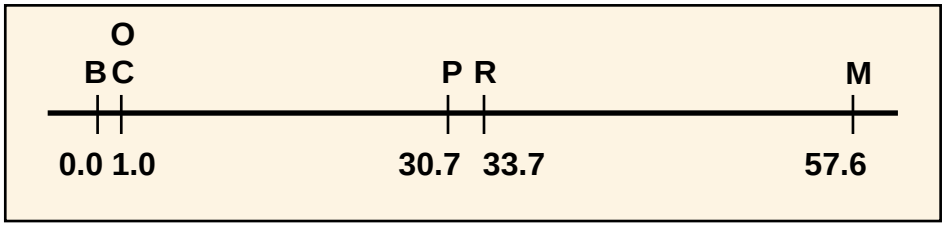
Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M. 1992. *Molecular Biology of the Gene*. Benjamin/Cummins Publishing Company: Menlo Park, California. p. 233.



What is a Map?



According to the beads on a string model, maps of a few genes might be represented by showing the gene names in order, with their relative positions indicated. And that is exactly the way the first genomic map was represented.

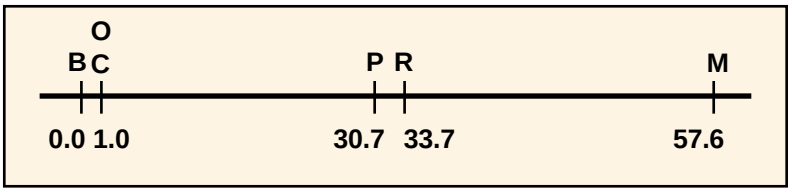


B = yellow body
C = white eye
O = eosin eye
P = vermilion eye
R = rudimentary wing
M = miniature wing

B, $\frac{O}{M}$, P, R, C

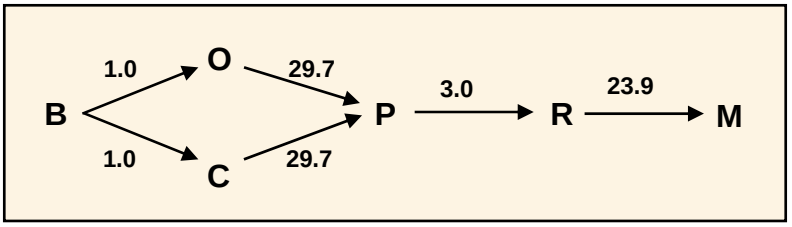
Sturtevant, A.H., 1913, The linear arrangement of six sex-linked factors in *Drosophila* as shown by their mode of association, *Journal of Experimental Zoology*, 14:43-59.

What is a Map?
Appropriate Data Structures



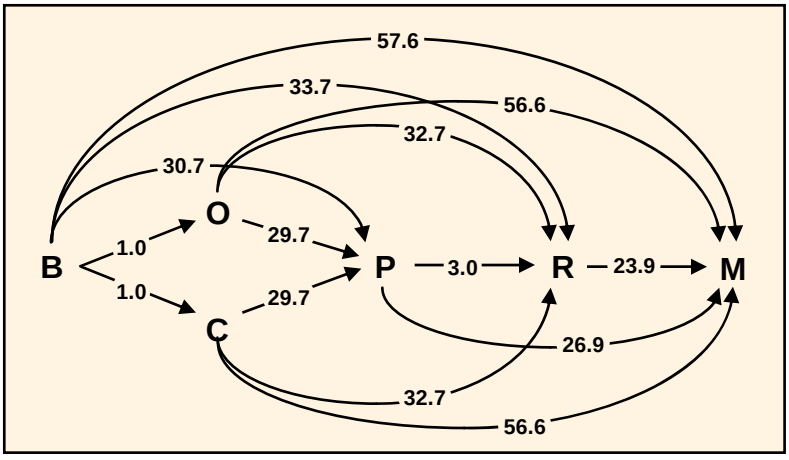
gene	locus
B	0.0
C	1.0
O	1.0
P	30.7
R	33.7
M	57.6

Many geneticists still think of maps as ordered lists, and ordered list representations are used in many genome databases..



arc	length
B, O	1.0
B, C	1.0
O, P	29.7
C, P	29.7
P, R	3.0
R, M	23.9

Directed graph data structures can be represented pictorially (above) or tabularly (right).



arc	length
B, O	1.0
B, C	1.0
B, P	30.7
B, R	33.7
B, M	57.6
O, P	29.7
O, R	32.7
O, M	56.6
C, P	29.7
C, R	32.7
C, M	56.6