Comparative Genomics: A New Integrative Biology

Robert J. Robbins

Johns Hopkins University
&
Department of Energy

rrobbins@gdb.org

Integrative Approaches to Molecular Biology

The Challenge:

- **Do we need to begin integrating information in molecular biology?**
- **Do we need people with a specific interest in integrating information?**
- **Is the intuition and expertise of the experimentalist enough to provide the required integrative framework?**
- **What is the role of formalization in integrative molecular biology?**

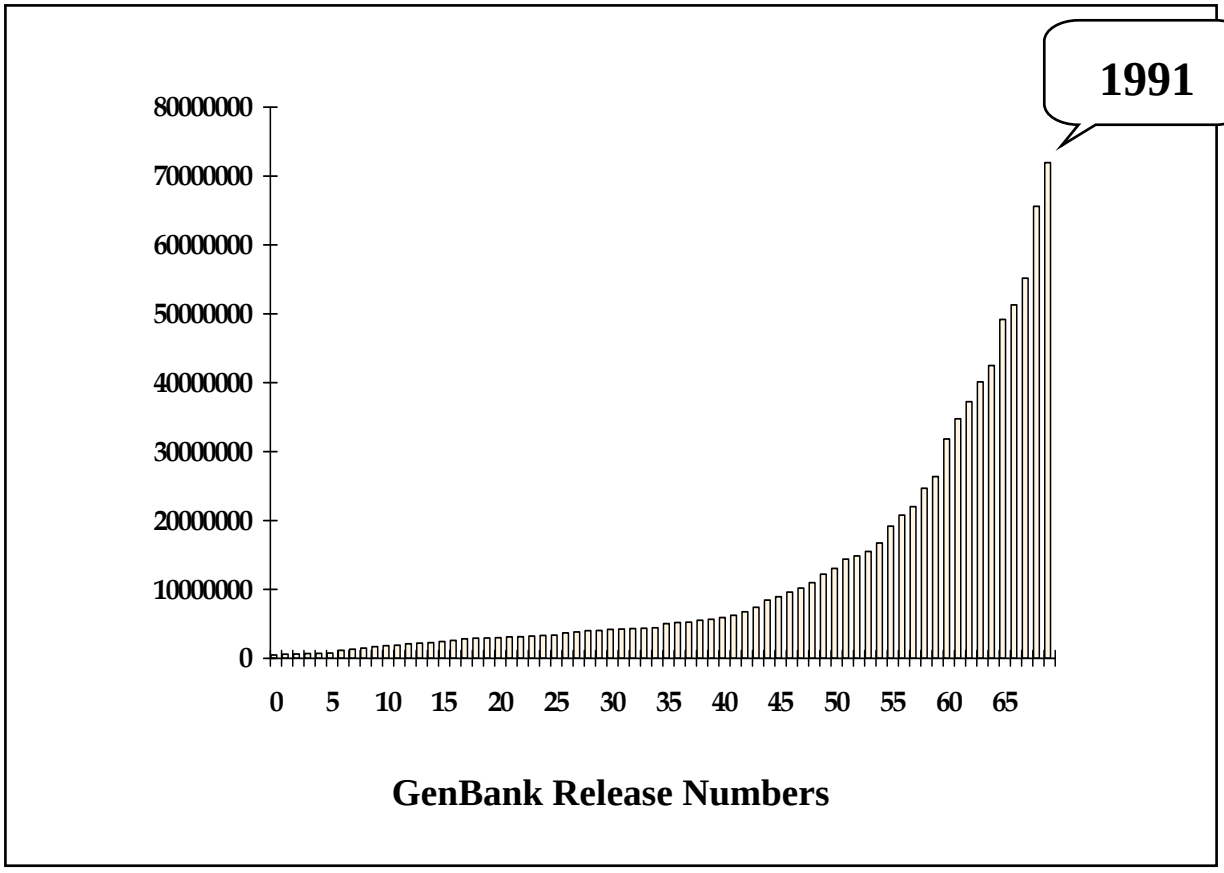
Human Genome Project

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.**

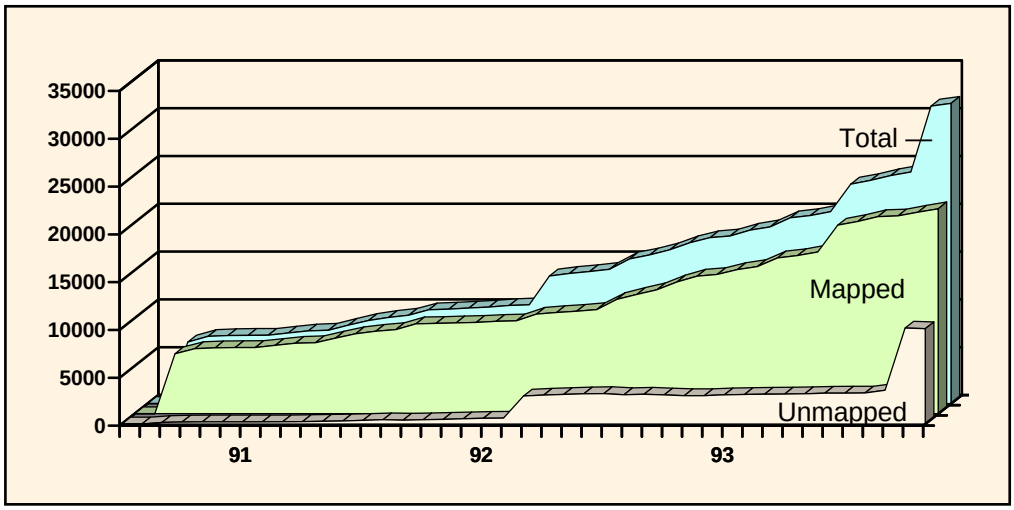
USDOE. 1990. Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years.

Cumulative Totals
GenBank Entries

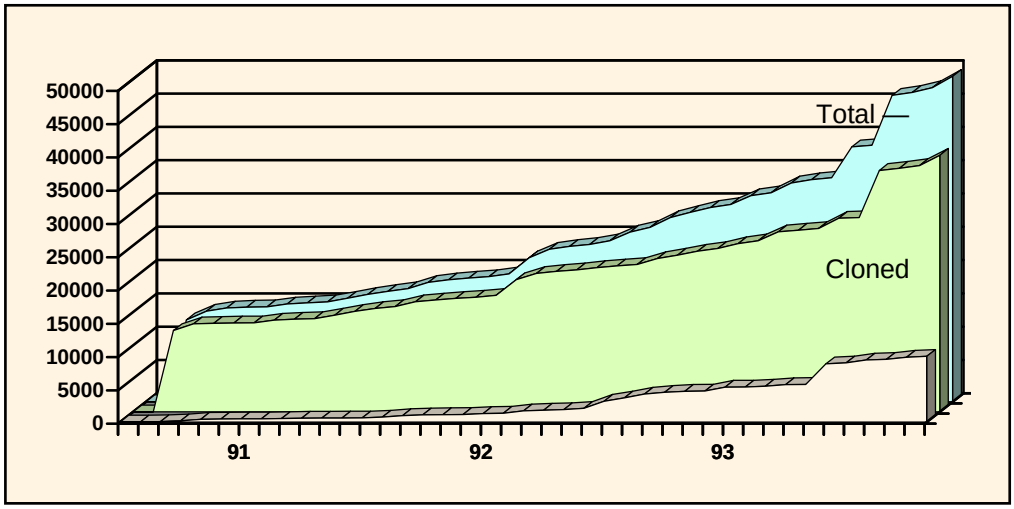


GDB Content

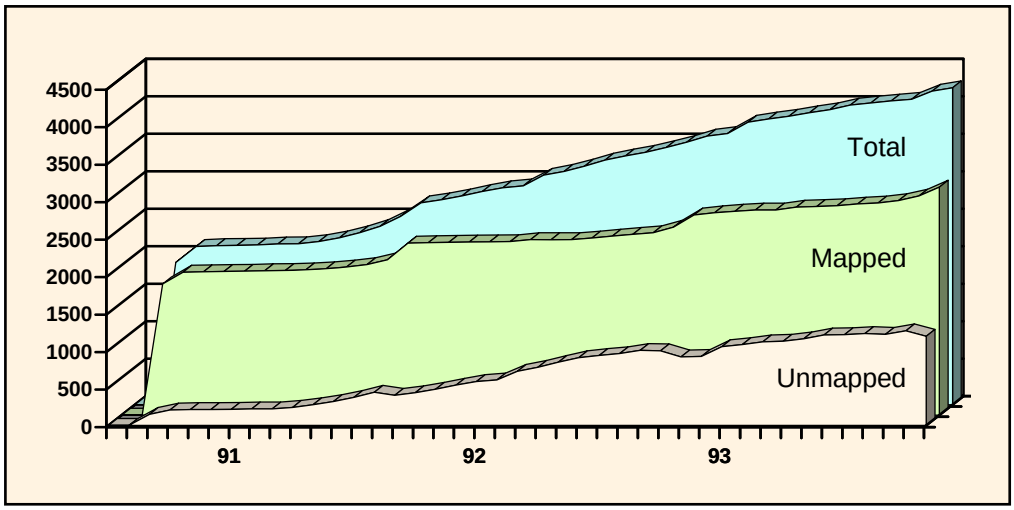
Loci



Probes



Genes



Getting the 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%

Assumptions:

- cost is direct cost per base of physical map elements
- average indirect cost rate is 50%
- overlap between physical map elements is 40%
- redundancy rate is 10%

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?"

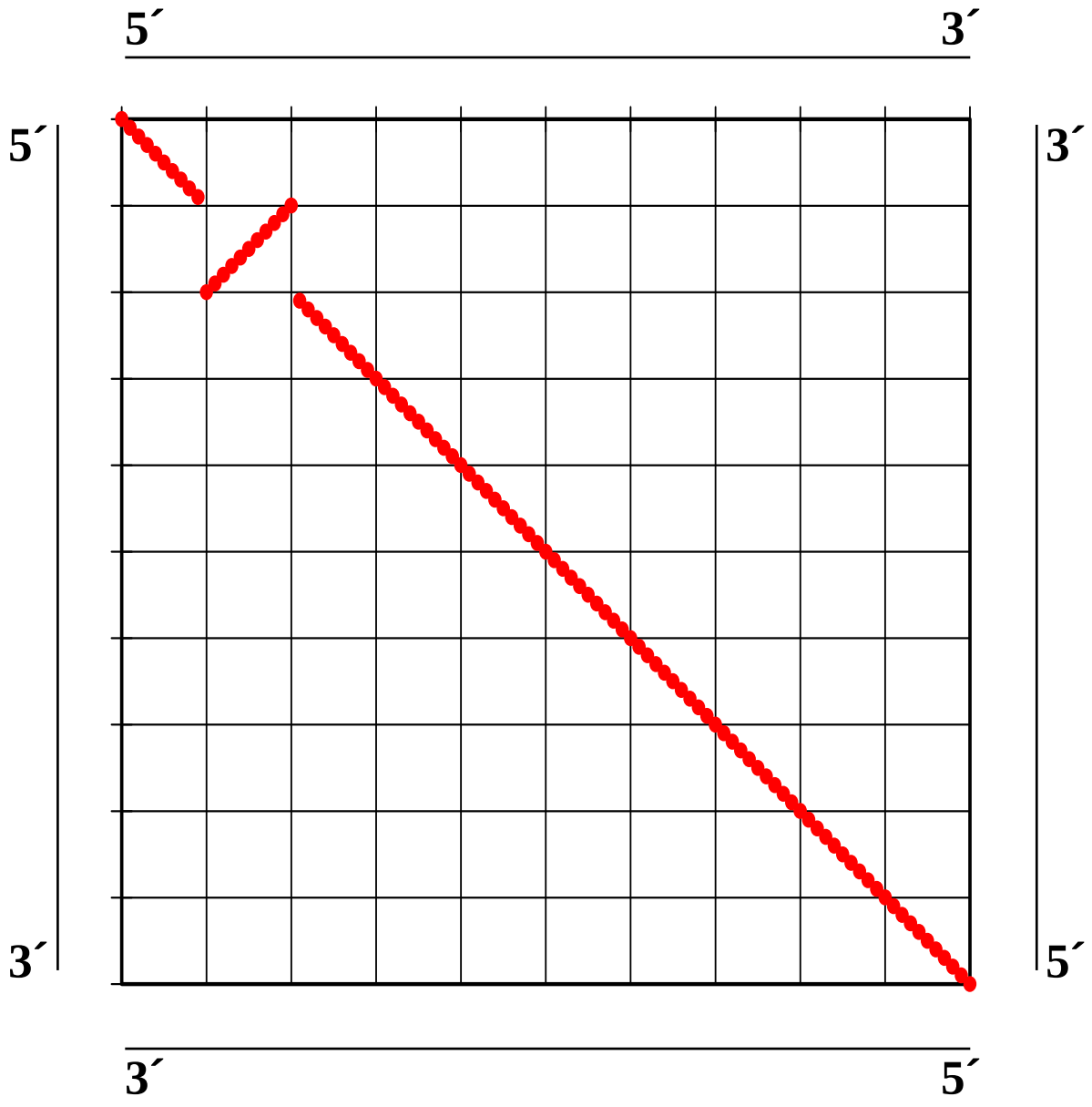
Sample Integrated Genome Queries

- **Return all sequences which map 'close' to marker M on human chromosome 19, are putative members of the olfactory receptor family, and have been mapped on a contig map of the region; return also the contig descriptions. (This is nominally a link between GenBank, GDB, and LLNL's databases.)**
- **Return all genomic sequences for which alu elements are located internal to a gene domain.**
- **Return the map location, where known, of all alu elements having homology greater than "h" with the alu sequence "S".**
- **Return all human gene sequences, with annotation information, for which a putative functional homologue has been identified in a non vertebrate organism; return also the GenBank accession number of the homologue sequence where available.**
- **Return all mammalian gene sequences for proteins identified as being involved in intra cellular signal transduction; return annotation information and literature citations.**
- **Return any annotation added to my sequence number #### since I last updated it.**
- **Return the genes for zinc-finger proteins on chromosome 19 that have been sequenced. (Note that answering this requires either query by sequence similarity or uniformity of nomenclature.)**
- **Return the number and a list of the distinct human genes that have been sequenced.**
- **Return all the human contigs greater than 150 kb.**
- **Return all sequences, for which at least two sequence variants are known, from regions of the genome within +/ one chromosome band of DS14###.**
- **Return all publications from the last two years about my favorite gene, accession number X####.**
- **Return all G1/S serine/threonine kinase genes (and their translated proteins) that are known (experimentally) or are thought (by similarity) also to exhibit tyrosine phosphorylation activity. Keep clear the distinction in the output.**

Comparative Genomic Techniques

Map Dot Plot

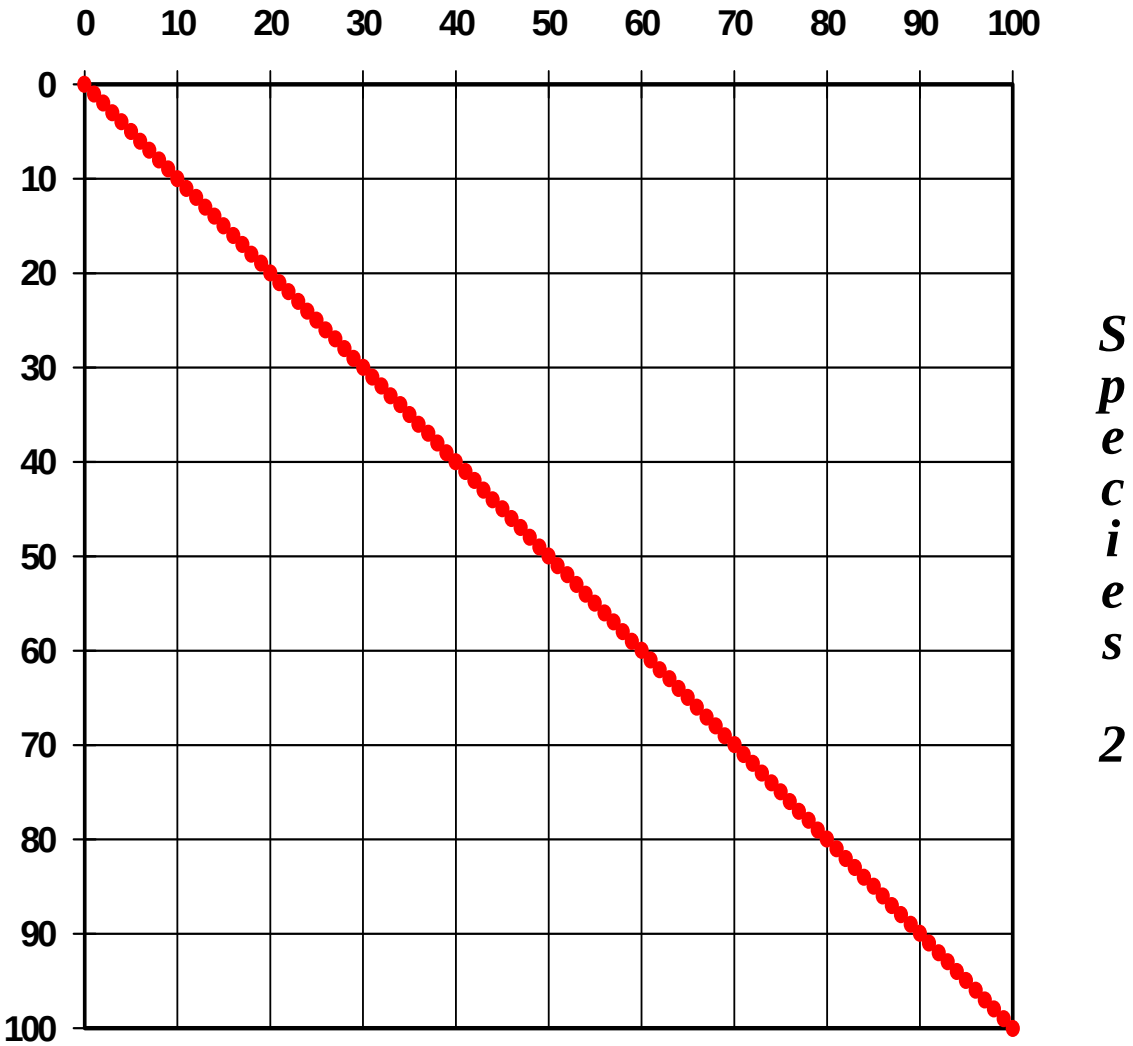
Bidirectional DNA Sequence Similarity Comparison



inversion

Map Dot Plot

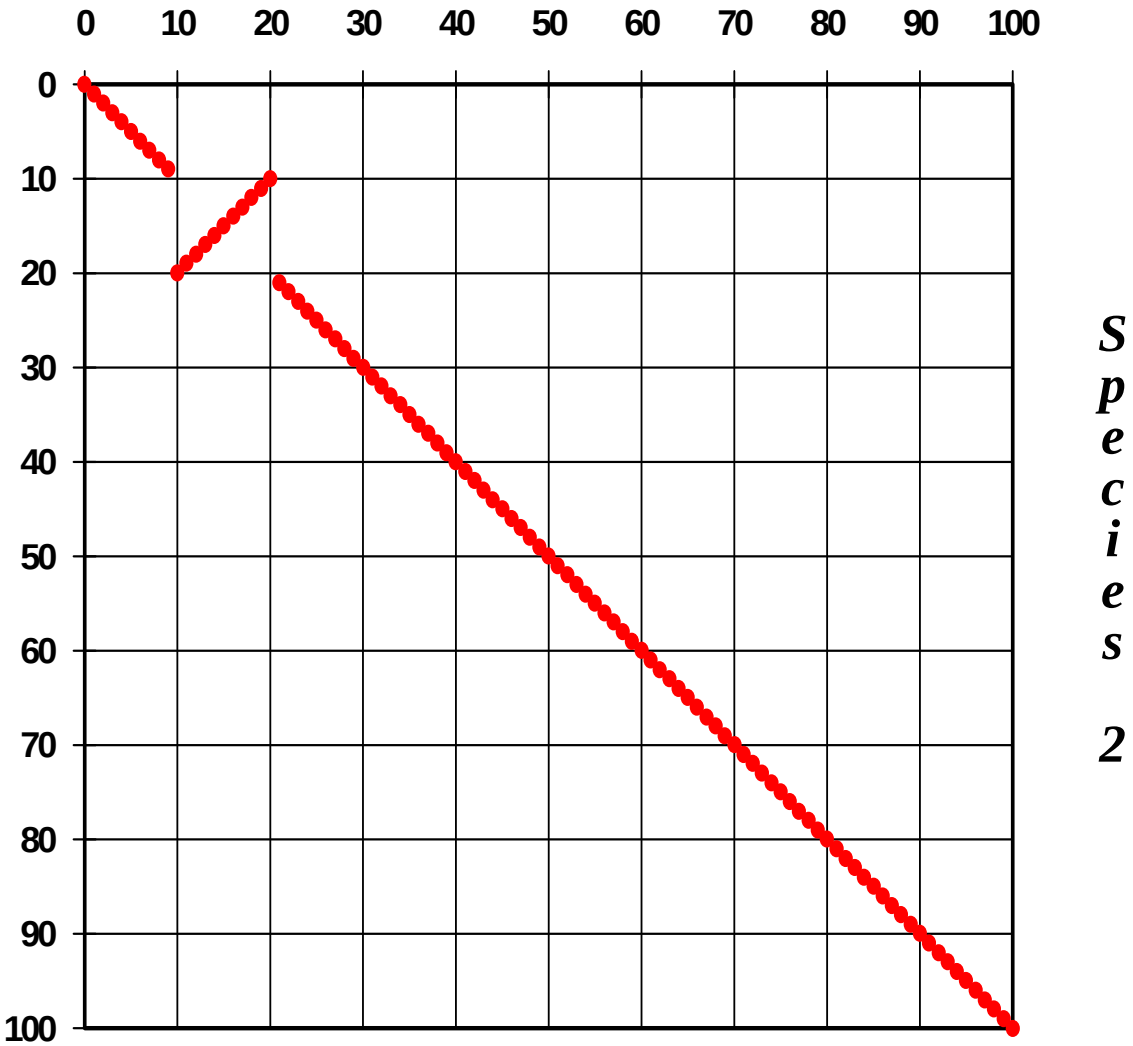
Homologous-Locus Position Comparison



perfect congruence

Map Dot Plot

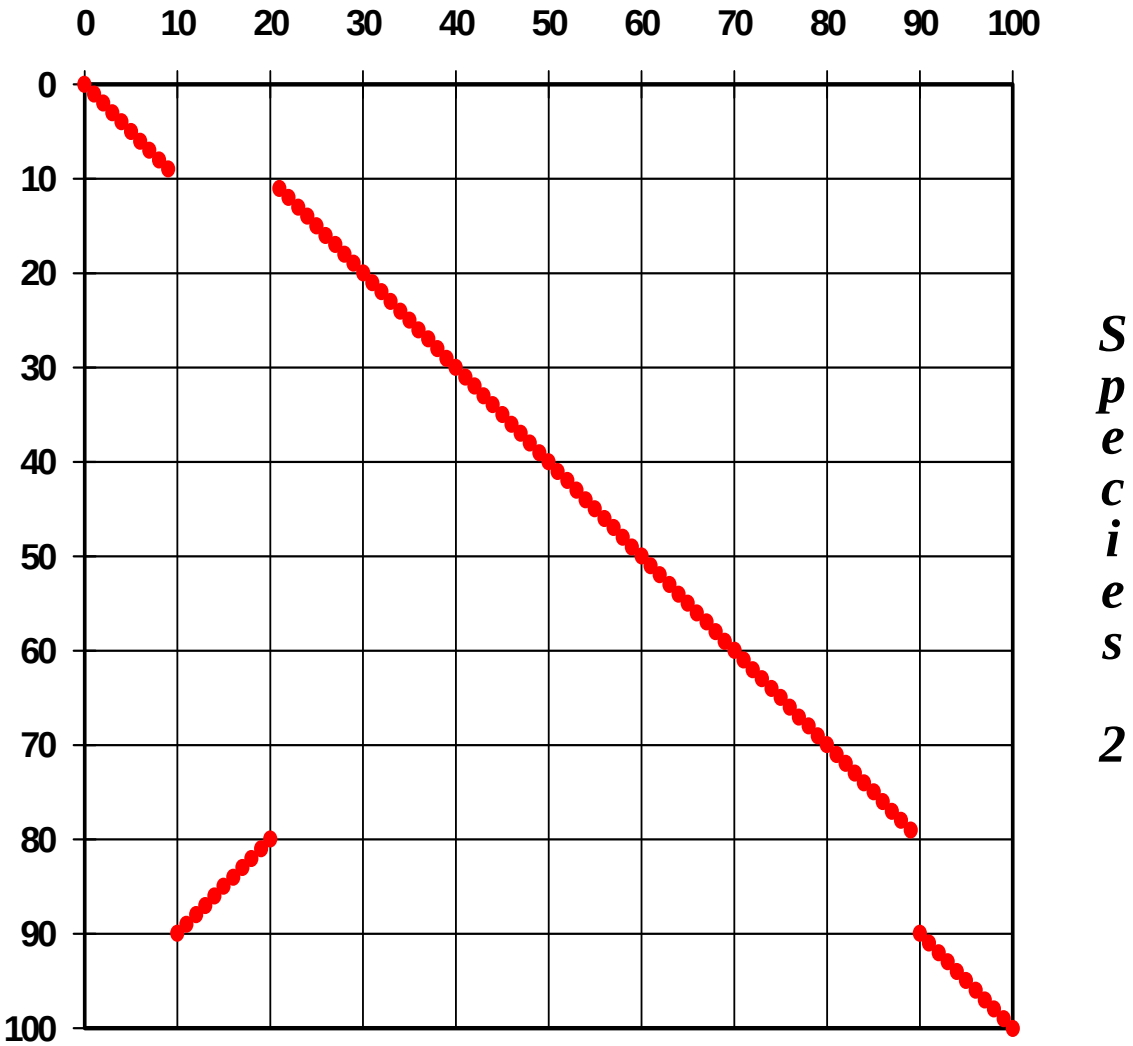
Homologous-Locus Position Comparison



inversion

Map Dot Plot

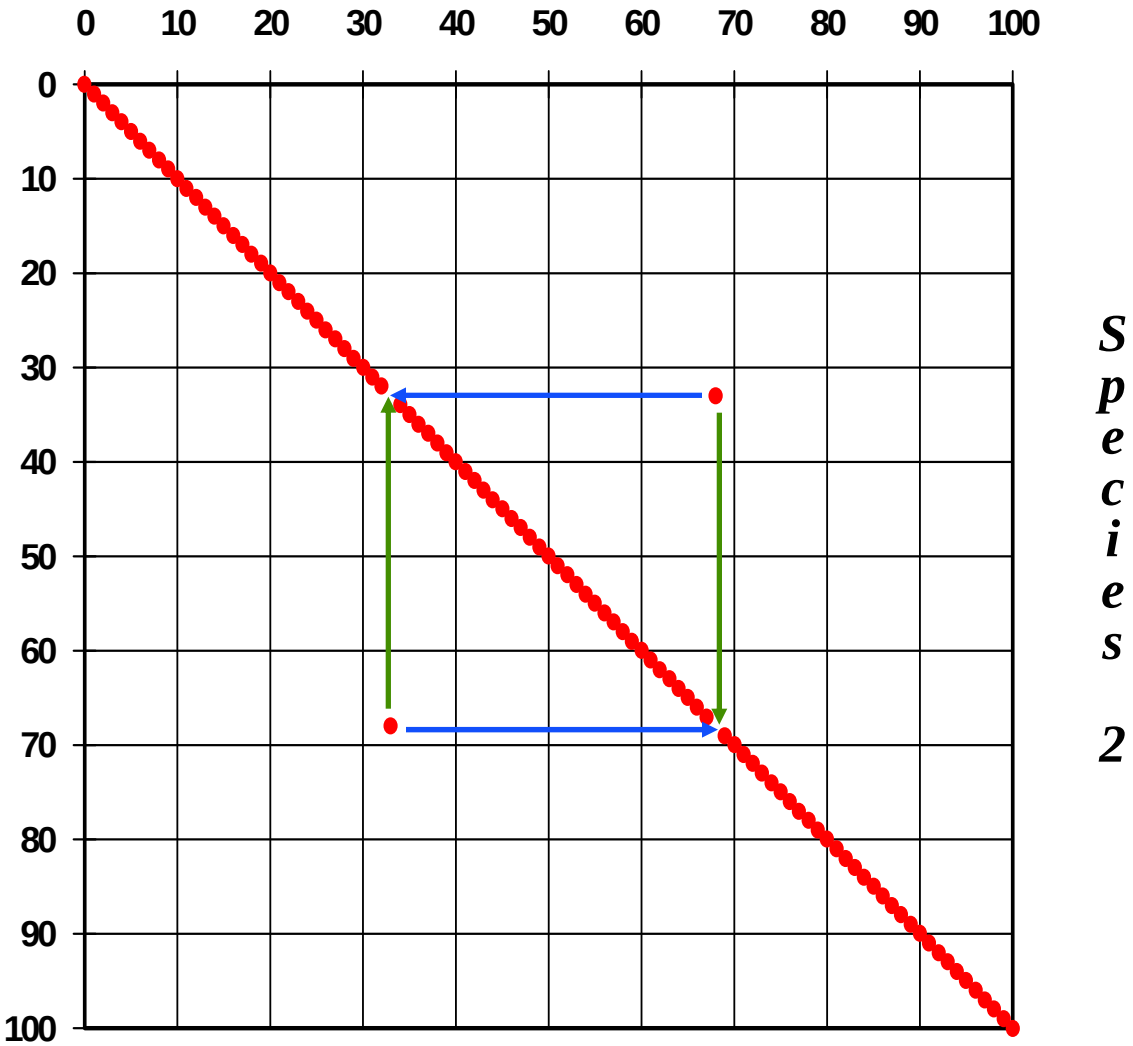
Homologous-Locus Position Comparison



translocation and inversion

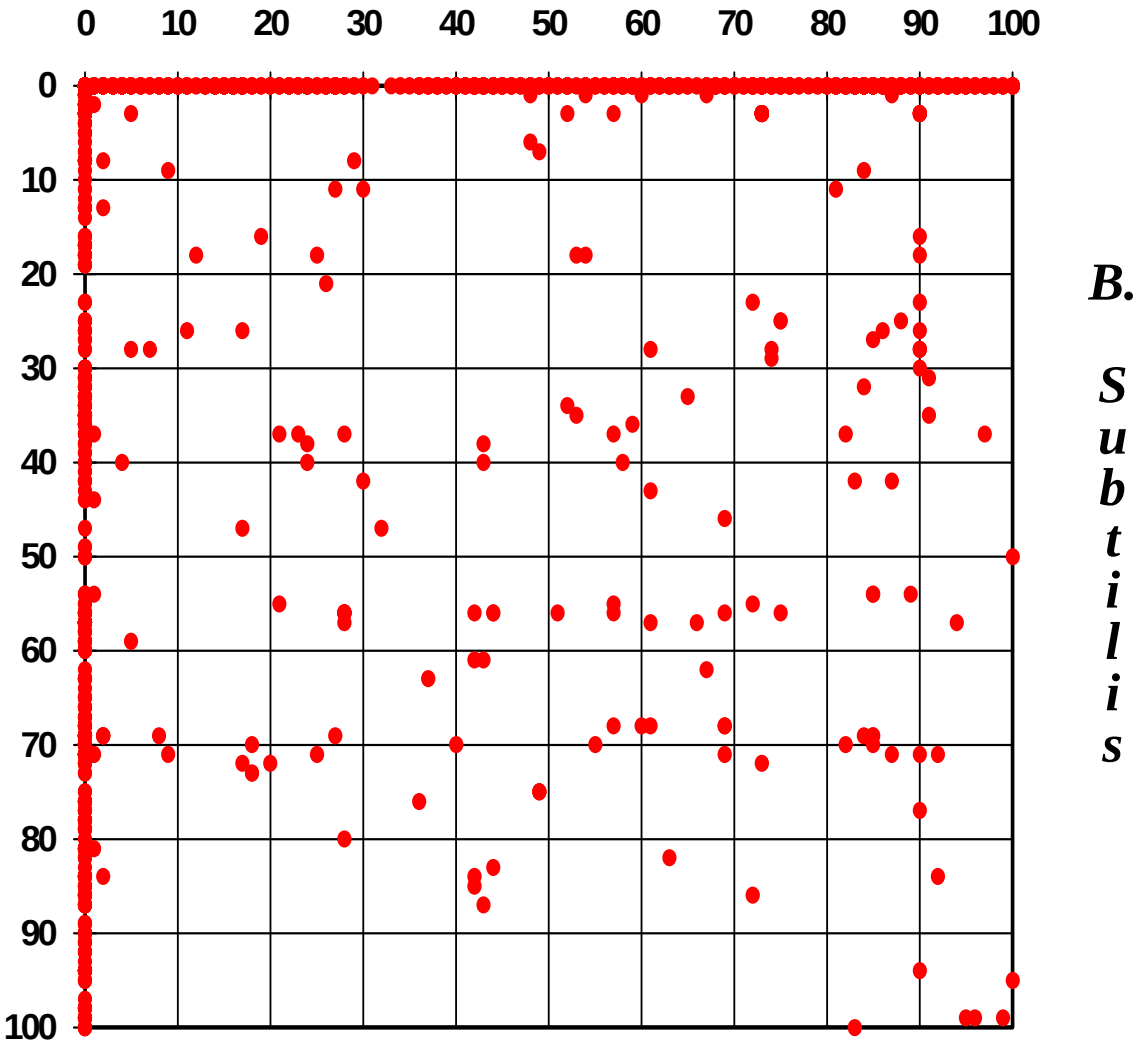
Map Dot Plot

Homologous-Locus Position Comparison



homology error

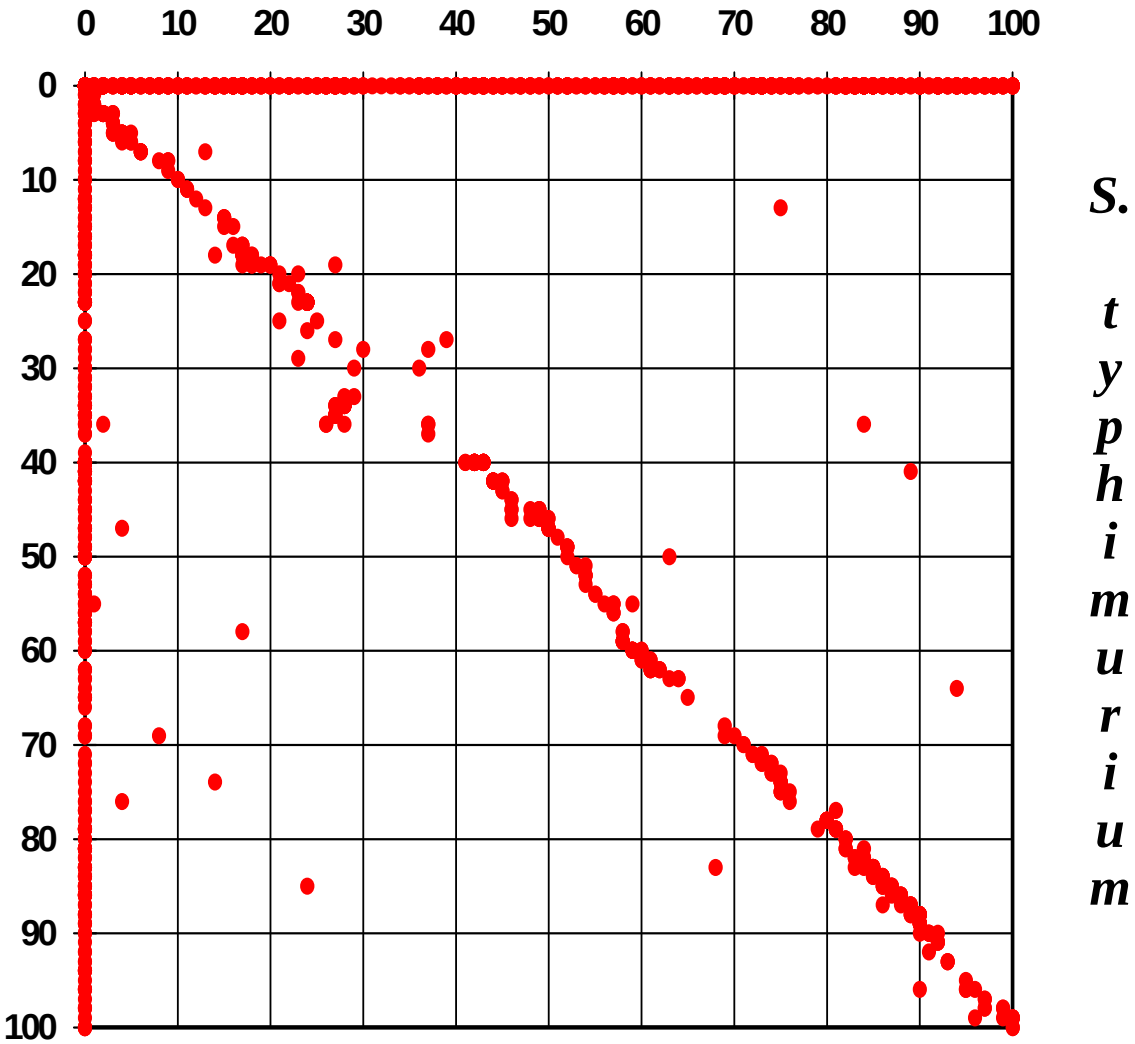
Map Dot Plot



E. coli* vs *B. subtilis

Map Dot Plot

Homologous-Locus Position Comparison



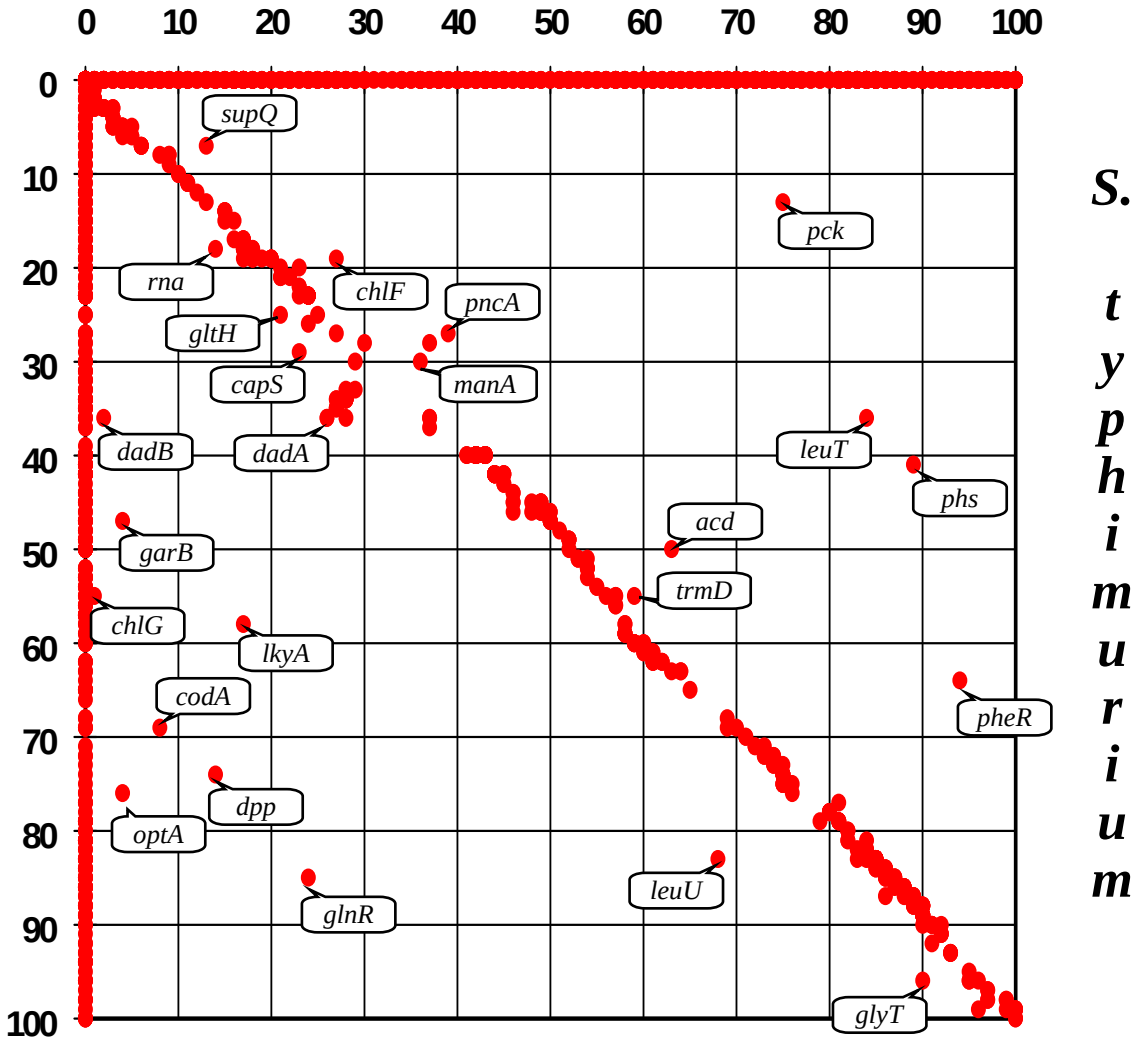
E. coli* vs *S. typhimurium

Data from:

Abel, Y., and Cedergren, R. 1990. The Normalized Gene Designation Database

Map Dot Plot

Homologous-Locus Position Comparison



E. coli* vs *S. typhimurium

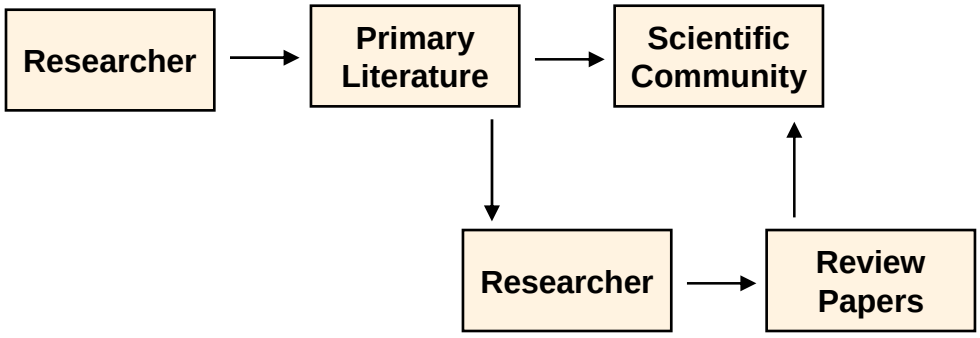
Data from:

Abel, Y., and Cedergren, R. 1990. The Normalized Gene Designation Database

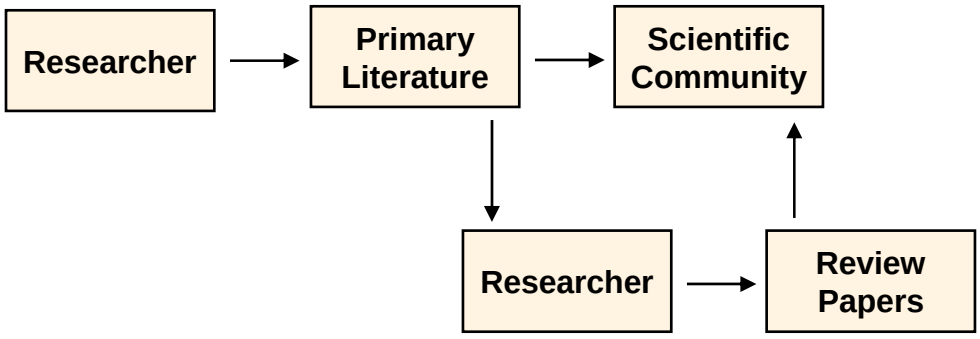
Electronic Data Publishing

Databases as Publishing

Traditional Publishing

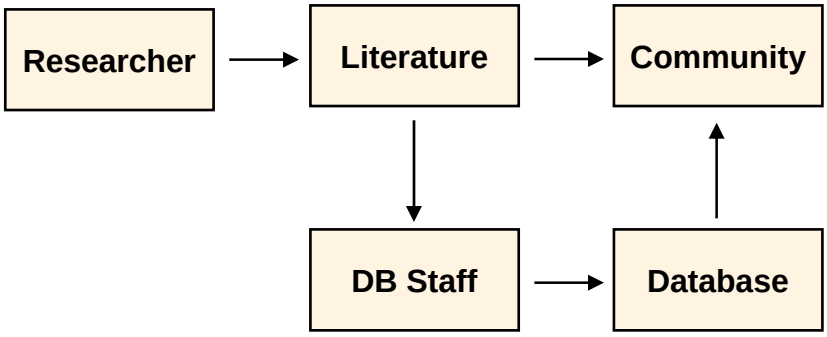


Early Database Development

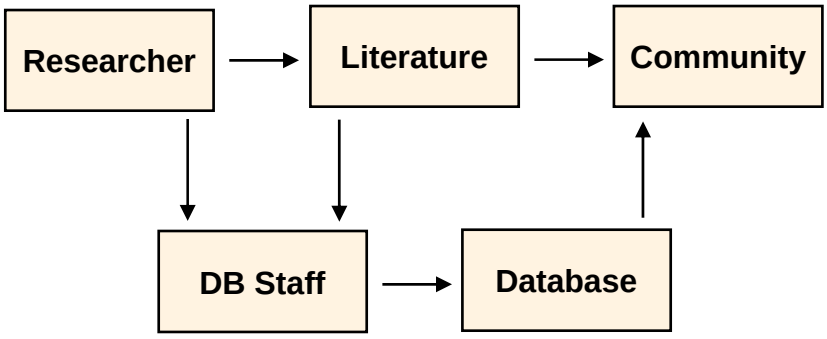


Electronic Data Publishing

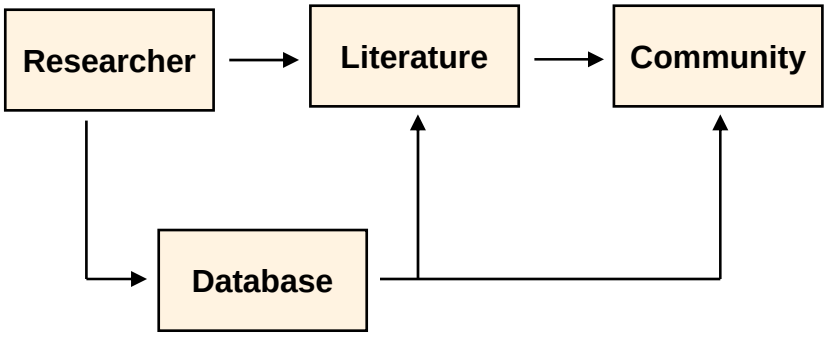
Standard Database Development



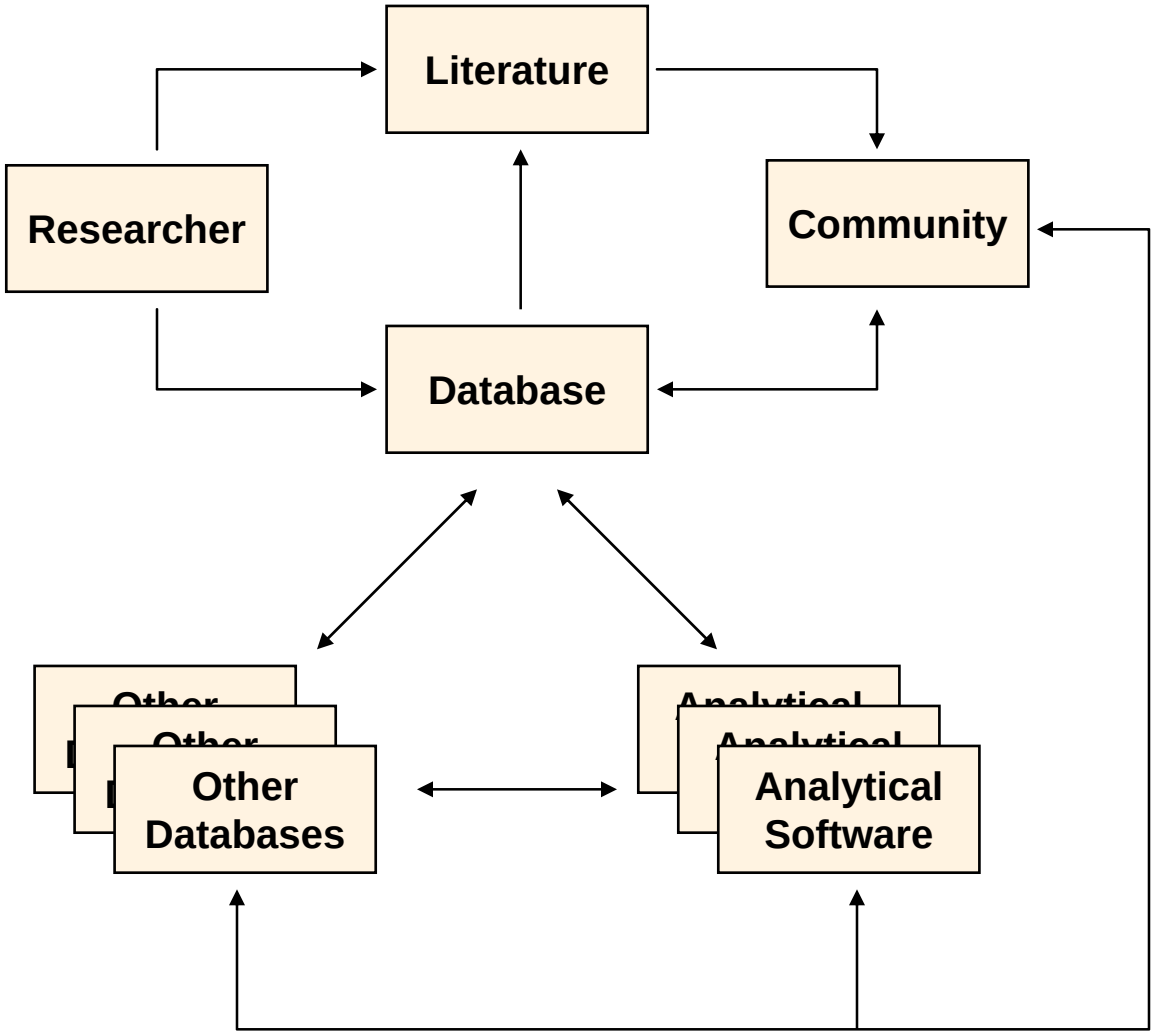
Early Electronic Data Publishing



Electronic Data Publishing

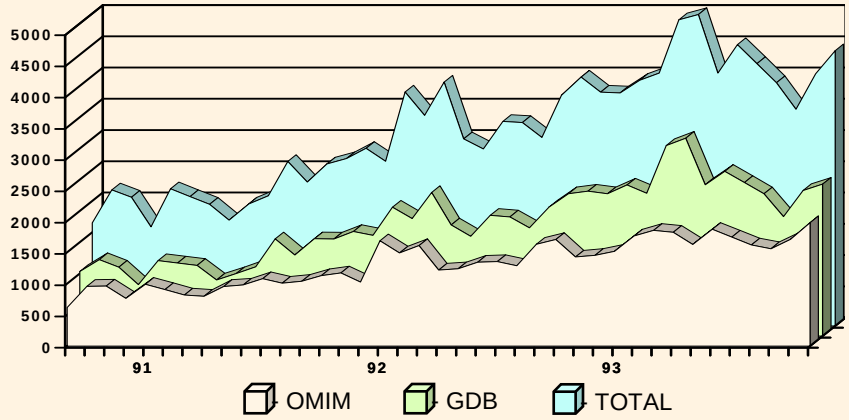


Electronic Data Publishing and Integrated Analysis



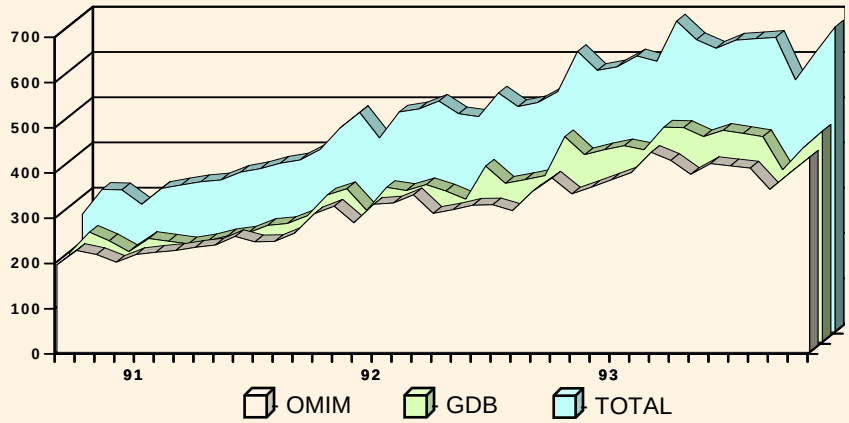
GDB Usage

Sessions



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

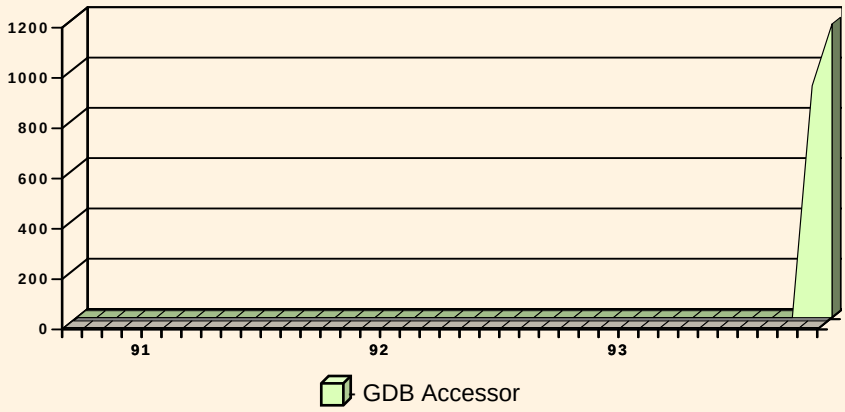
Unique Users



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.

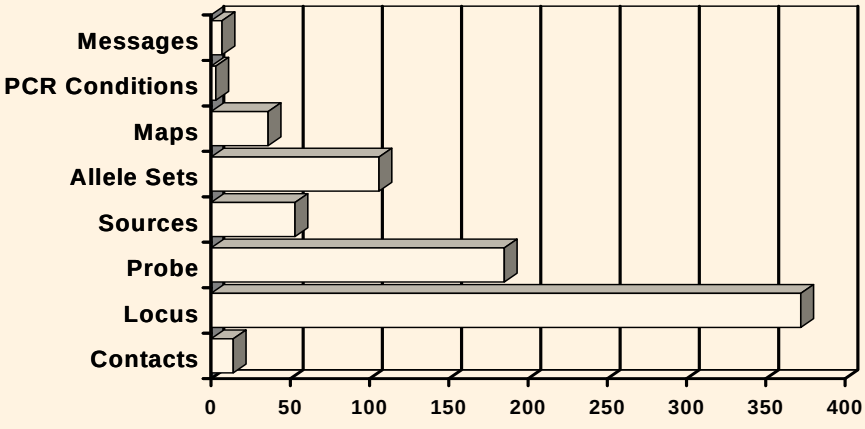
GDB Usage

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

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.

Genome Informatics
Requirements and Challenges

Requirements	Challenges
Data Acquisition	**Design**
Data Analysis	**Design**
Data Exchange	**Design**
Data Publication	**Design**
Data Access	**Design**
Data Visualization	**Design**

It is well known that 80% of a data processing budget is spent on maintenance. ... this cost is largely the result of inadequate database design.

Finkelstein, R. 1988. Database design: The road seldom taken. *Database Programming & Design*. 1(4):11-15.

[I]nadequate analysis and design, more than any other factor, deny programmers the chance to perform well. Programming without a precise statement of requirements or a plan for achieving those requirements is like trying to reach an unknown destination in the dark without a map.

Page-Jones, M. 1988. *The Practical Guide to Structured Systems Design*. Englewood Cliffs, New Jersey: Yourdon Press.

Building Databases - An Example

Problem:

Design a database to keep track of the buildings owned and occupied by the federal government.

How tough could it be?

The General Services Administration said last week it will discontinue development of Stride, a complex computer system to automate the Public Buildings Service. GSA has spent $100 million since 1983 trying to make the system work -- a figure that does not include $ 78 million spent of the system Stride was intended to replace.

GSA officials said Stride fell apart largely because of a failure to create a workable systems-integration plan for it. The most glaring problem with Stride was that it lacked an integration design. Stride went directly from the functional design to work packages without the intermediate step of a detailed design to show how all of the packages would fit together.

Levine, Arnold S. 1988. GSA razes $100M PBS data project. *Federal Computer Week*, 2(29):1,53.

Cost of Software R & D

vendor	software revenues			amt revs to R&D		
	1987	1988	1989	1987	1988	1989
Computer Associates	649	925	1,290	84	120	168
Microsoft	301	625	691	33	75	97
Lotus	396	469	556	59	84	106
Dun & Bradstreet	200	200	435	50	50	NA
Oracle	196	425	418	22	47	46
Software AG of N America	185	221	294	37	44	59
Novell	126	200	282	9	14	28
WordPerfect	100	179	281	NA	NA	28
Ashton-Tate	267	307	265	35	52	69
Pansophic Systems	87	122	232	7	10	19
SAS Institute	135	170	205	61	77	92

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.

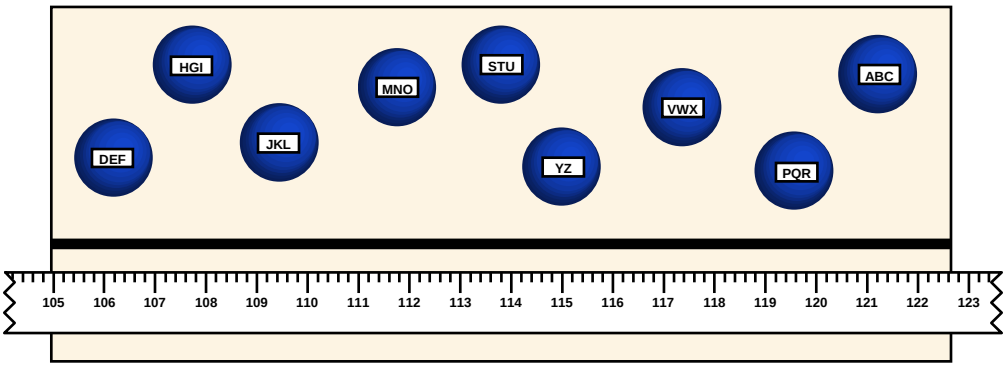
Fifty years from now it seems very likely that the most significant development of genetics in the current decade (1945-1955) will stand out as being the discovery of pseudoallelism.

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

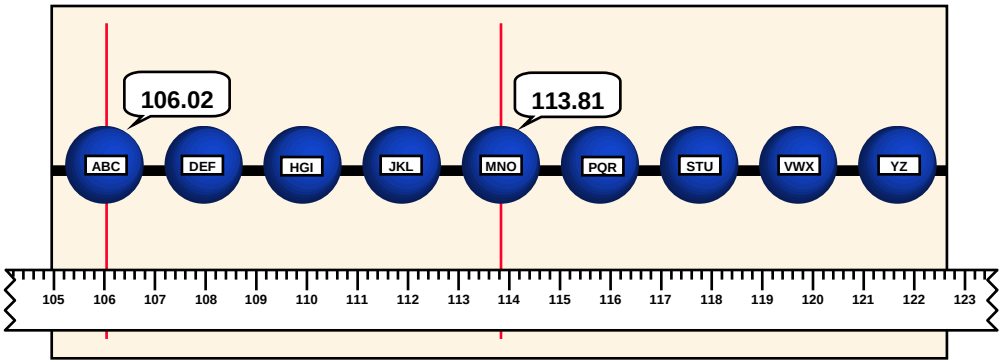
The ultimate ... map [will be] the complete DNA sequence of the human genome.

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.

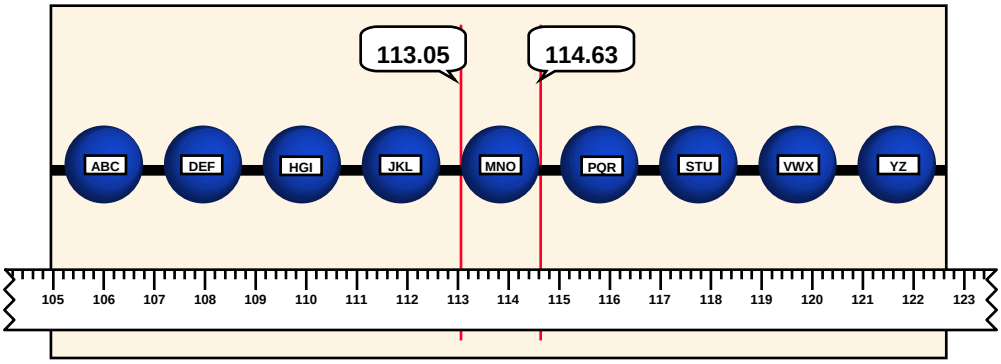


Mapping involves placing the beads in the correct order and assigning a correct address to each bead. The address assigned to a bead is its locus.

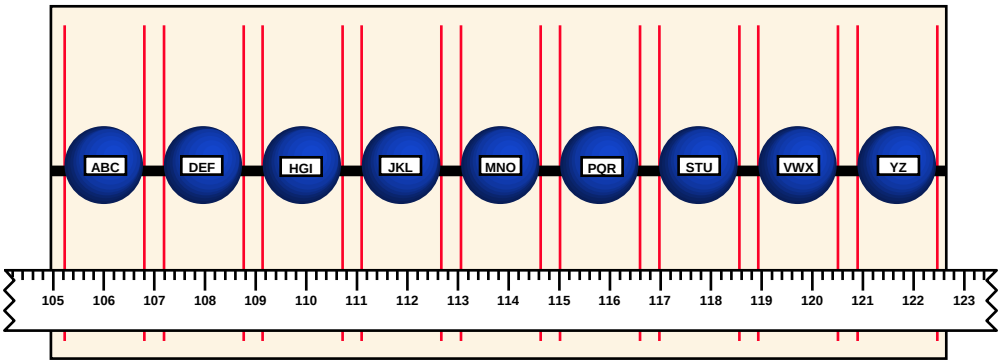
The genes are arranged in a manner similar to beads strung on a loose string.

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.

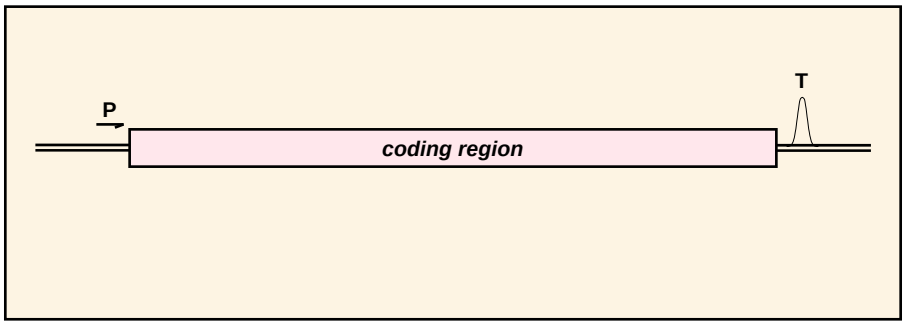
Lewin, Benjamin. 1990. *Genes IV*. Oxford University Press, New York.

What is a Gene?

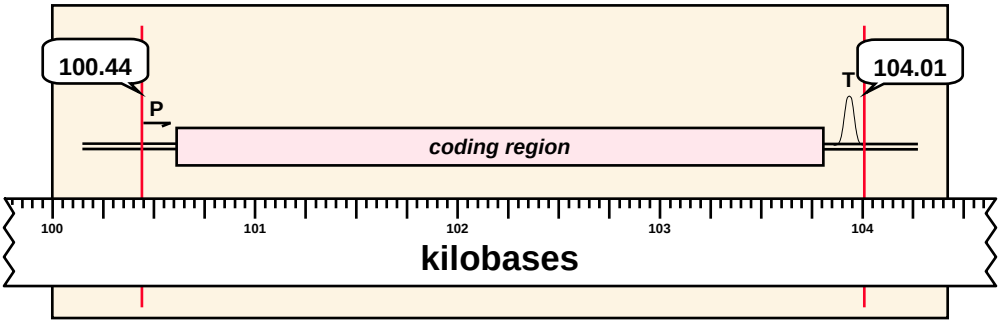
The unexpected features of eukaryotic genes have stimulated discussion about how a gene, a single unit of hereditary information, should be defined. Several different possible definitions are plausible, but no single one is entirely satisfactory or appropriate for every gene.

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

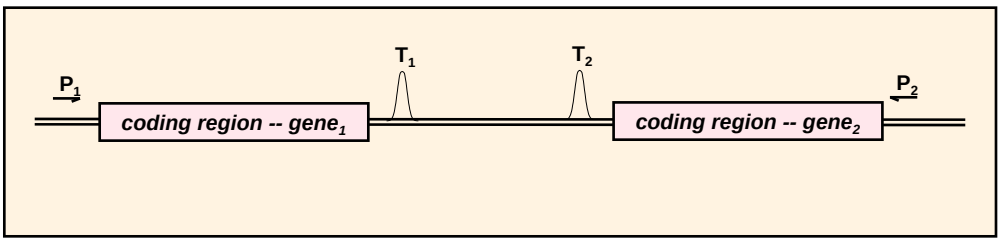
The Simplistic View of a Genome



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

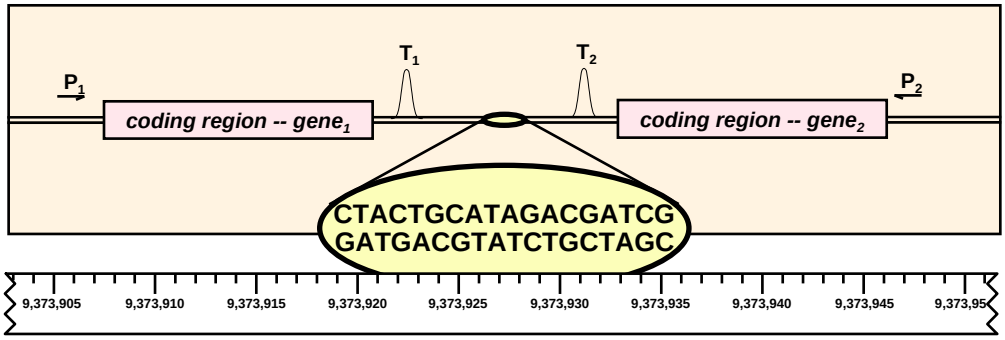


The location of a gene can be designated by specifying the base-pair location of its beginning and end.



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.

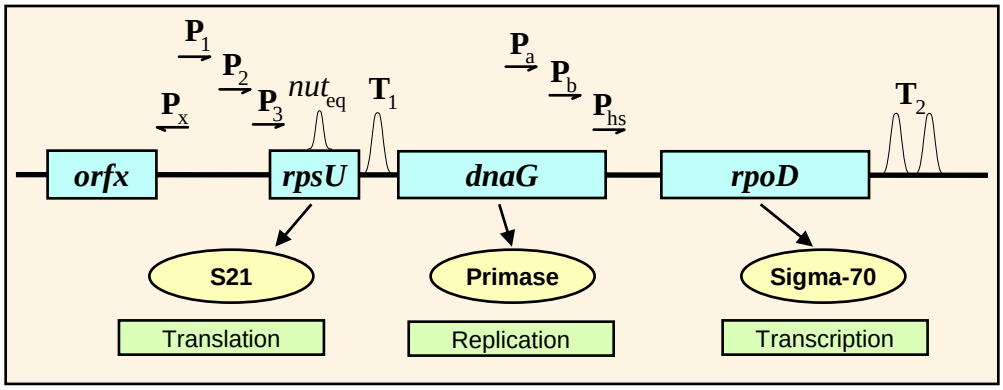
The Simplistic View of a Genome



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

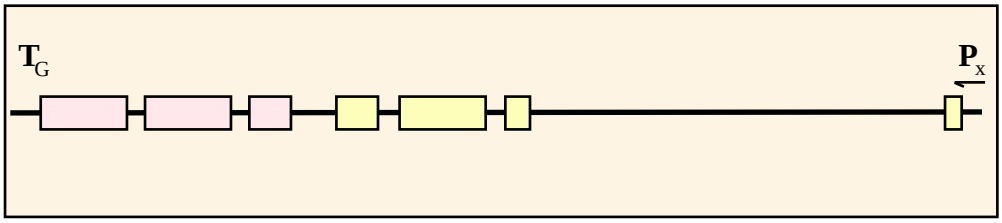
Escherichia coli: the MMS Operon



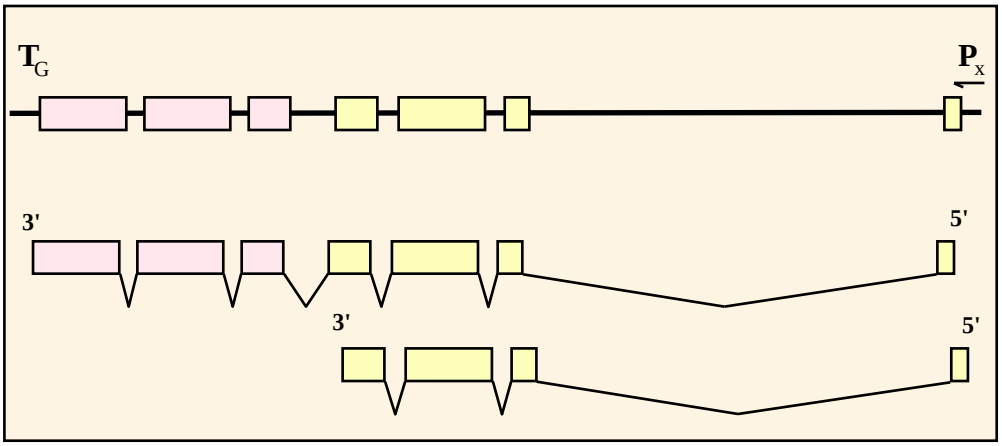
Lupski, J.R., Godson, G.N., 1989, DNA→DNA, and DNA→RNA→Protein: Orchestration by a single complex operon, *BioEssays*, 10:152-157.

Drosophila melanogaster: The Gart Locus

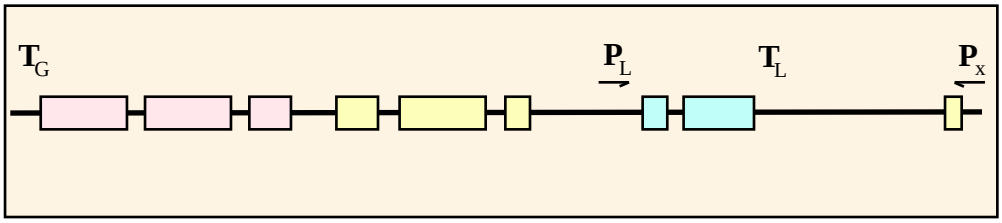
Fragmented Genes



Alternative Splicing



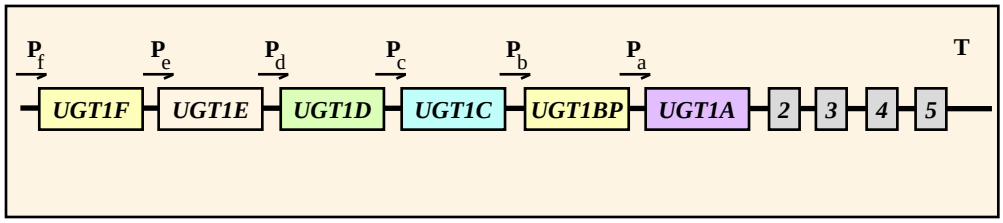
Nested Genes



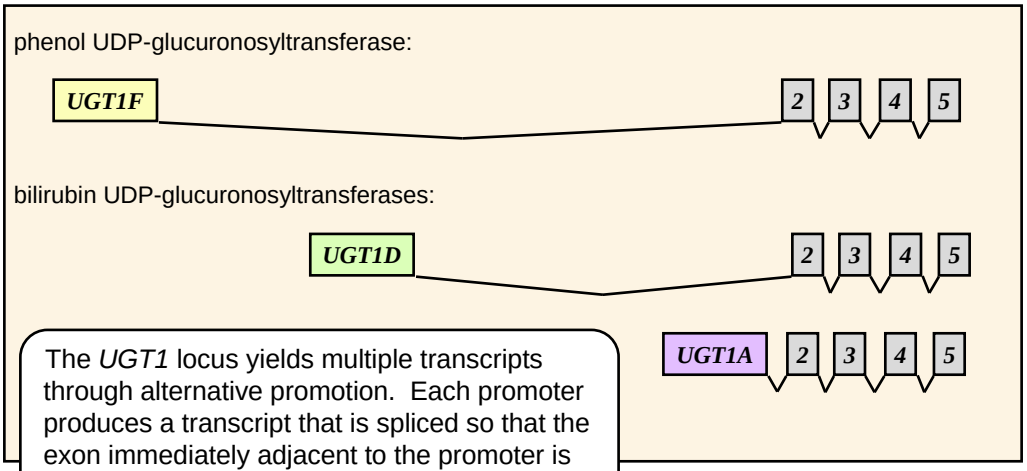
Henikoff, S., Keene, M.A., Fechtel, K., and Fristrom, J.W., 1986, Gene within a gene: Nested *Drosophila* genes encode unrelated proteins on opposite strands, *Cell* 44:33.

Nested Gene Families

Homo sapiens: The UGT1 Loci



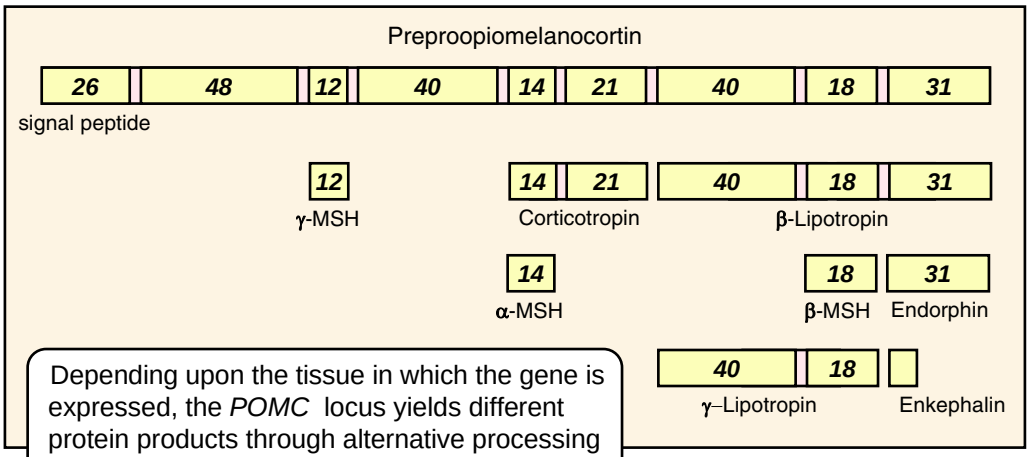
Ritter, J.K., Chen, F., et al., 1992, A novel complex locus *UGT1* encodes human bilirubin, phenol, and other UDP-glucuronosyltransferase isozymes with identical carboxyl termini, *J. Biol. Chem.* 267:3257.



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

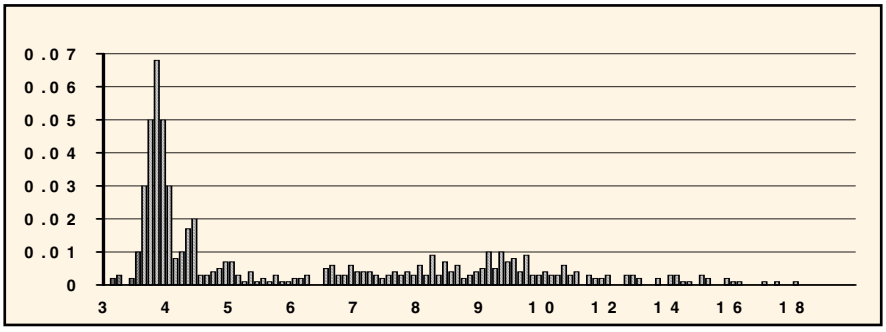
Homo sapiens: The POMC Locus



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?

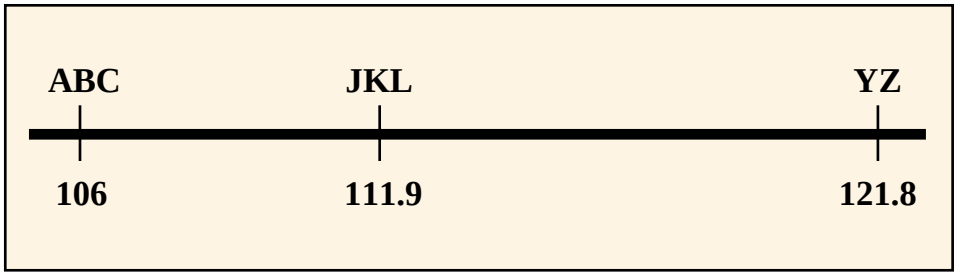
For the purposes of this book, we have adopted a molecular definition. A eukaryotic gene is a combination of DNA segments that together constitute an expressible unit, expression leading to the formation of one or more specific functional gene products that may be either RNA molecules or polypeptides.

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

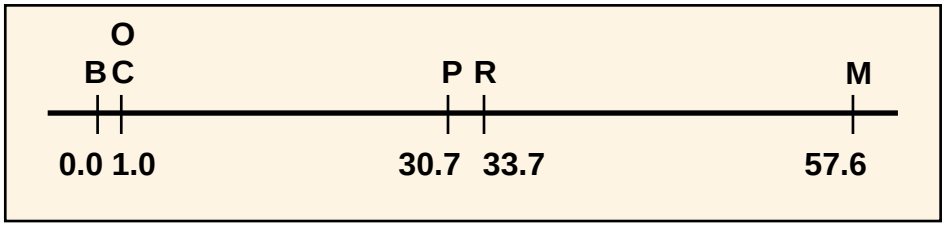
DNA molecules (chromosomes) should thus be functionally regarded as linear collections of discrete transcriptional units, each designed for the synthesis of a specific RNA molecule. Whether such "transcriptional units" should now be redefined as genes, or whether the term *gene* should be restricted 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.

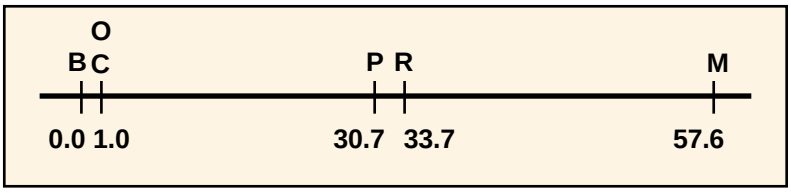


B = yellow body
C = white eye
O = eosin eye
P = vermilion eye
R = rudimentary wing
M = miniature wing

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

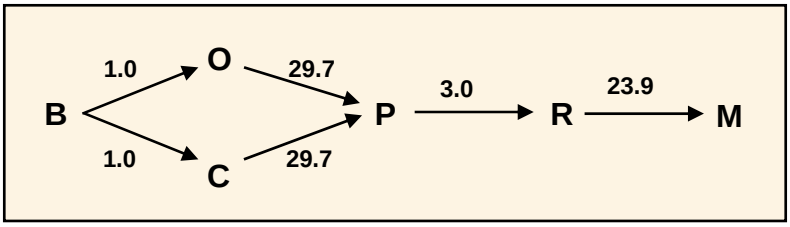
Sturtevant, A.H., 1913, The linear arrangement of six sex-linked factors in *Drosophila* as shown by their mode of association, *Journal of Experimental Zoology*, 14:43-59.

What is a Map?
Appropriate Data Structures



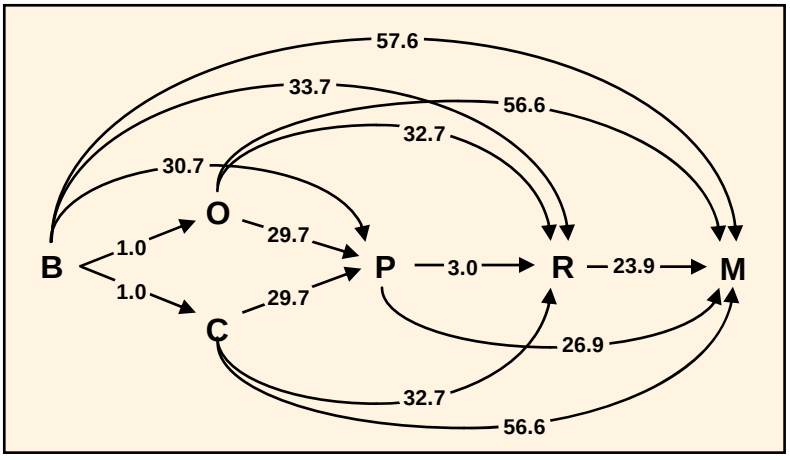
gene	locus
B	0.0
C	1.0
O	1.0
P	30.7
R	33.7
M	57.6

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

Sociological Impediments

Data Sharing -- Not a New Problem

Newton ... clashed with the Astronomer Royal, John Flamsteed, who had earlier provided Newton with much needed data for Principia, but was now withholding information that Newton wanted. Newton would not take no for an answer; he had himself appointed to the governing body of the Royal Observatory and then tried to force immediate publication of the data. Eventually he arranged for Flamsteed's work to be seized and prepared for publication by Flamsteed's mortal enemy, Edmond Halley. But Flamsteed took the case to court and, in the nick if time, won a court order preventing distribution of the stolen work. Newton was incensed and sought his revenge by systematically deleting all references to Flamsteed in later editions of Principia.

Hawking, Stephen W. 1988. *A Brief History of Time*. New York: Bantam Books. p. 181

Role for Formalisms

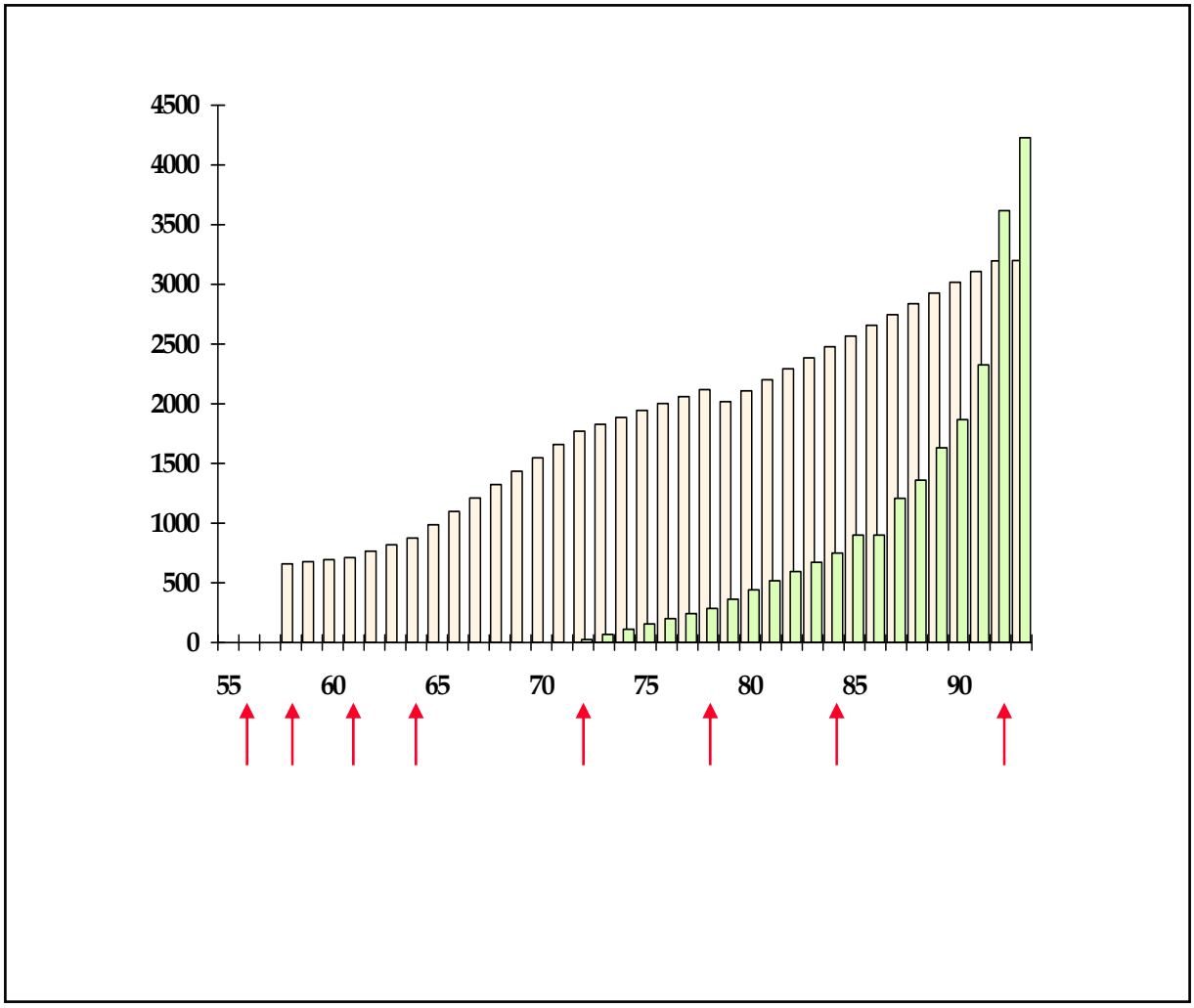
Limitations of Scientific Language

Geneticists, like all good scientists, proceed in the first instance intuitively and ... their intuition has vastly outstripped the possibilities of expression in the ordinary usages of natural languages. They know what they mean, but the current linguistic apparatus makes it very difficult for them to say what they mean. This apparatus conceals the complexity of the intuitions. It is part of the business of genetical methodology first to discover what geneticists mean and then to devise the simplest method of saying what they mean. If the result proves to be more complex than one would expect from the current expositions, that is because these devices are succeeding in making apparent a real complexity in the subject matter which the natural language conceals.

Woodger, J. H. 1952. *Biology and Language*. Cambridge: Cambridge University Press.

A Modest Proposal

Enzyme Commission History



Cumulative Totals
Known Enzymes
Human Genes

Why an Enzyme Commission

[M]any other workers were inventing names for new enzymes, and there was a widespread view that the results were chaotic and unsatisfactory. In many cases the enzymes became known by several different names, while conversely the same name was sometimes given to different enzymes. Many of the names conveyed little or no idea of the nature of the reactions catalyzed, and similar names were sometimes given to enzymes of quite different types.

Webb, E. C. 1993. Enzyme nomenclature: A personal retrospective. *The FASEB Journal*, 7:1192-1194.

Why an Enzyme Commission

A major part of this assignment was to see how the nomenclature of enzymes could best be brought into a satisfactory state and whether a code of systematicrules could be devised that would serve as a guide for the consistent naming of new enzymes in the future. ... [T]he overriding consideration was to reduce the confusion and to prevent further confusion from arising. This task could not have been accomplished without causing some inconvenience, for this was the inevitable result of not tackling the problem earlier." (p. 1193)

Webb, E. C. 1993. Enzyme nomenclature: A personal retrospective. *The FASEB Journal*, 7:1192-1194.

Estimates of Success

Strong Recommendations:

> In this context it is appropriate to express disapproval of a loose and misleading practice that is found in the biological literature. It consists in designation of a natural substance ... that cannot be described in terms of a definite chemical reaction, by the name of the phenomenon in conjugation with the suffix *-ase*... Some recent examples of such *phenomenase* nomenclature, which should be discouraged even if there are reasons to suppose that the particular agent may have enzymatic properties, are: *permease, translocase, replicase*, ... etc.
>
> Enzyme Nomenclature, 1984

Actual Results:

permease	**520**
translocase	**160**
replicase	**362**

Number of papers returned on a search of MedLine from 1986 - 1993.

A Modest Proposal

There will come a time when all the plant forms in existence will have been described; when herbaria will contain indubitable material of them; when botanists will have made, unmade, often remade, raised, or lowered, and above all modified several hundred thousand taxa ranging from classes to simple varieties, and when synonyms will have become much more numerous than accepted taxa. Then science will have need of some great renovation of its formulae. This nomenclature which we now strive to improve will then appear like an old scaffolding, laboriously patched together and surrounded and encumbered by the debris of rejected parts. The edifice of science will have been built but the rubbish incident to its construction not cleared away. Then perhaps there will
arise something wholly different from Linnaean nomenclature, something so designed as to give certain and definite names to certain and definite taxa.

That is the secret of the future, a future still very far off.

In the meantime, let us perfect the binomial system introduced by Linnaeus. Let us try to adapt it better to the continual, necessary changes in science ... drive out small abuses, the little negligences and, if possible, come to agreement on controversial points. thus
we shall prepare the way for the better progress of taxonomy.

Alphonse de Candolle, 1867, *Laws of Botanical Nomenclature*, quoted in Nicolson, D. H. 1991. A history of botanical nomenclature. *Annals of the Missouri Botanical Garden*, 78:33-56