Lecture 1
Genome Projects: Organization & Objectives

薛 佑 玲 Yow-Ling Shiue
Institute of Biomedical Science
National Sun Yat-sen University
✉ ylshiue@mail.nsysu.edu.tw

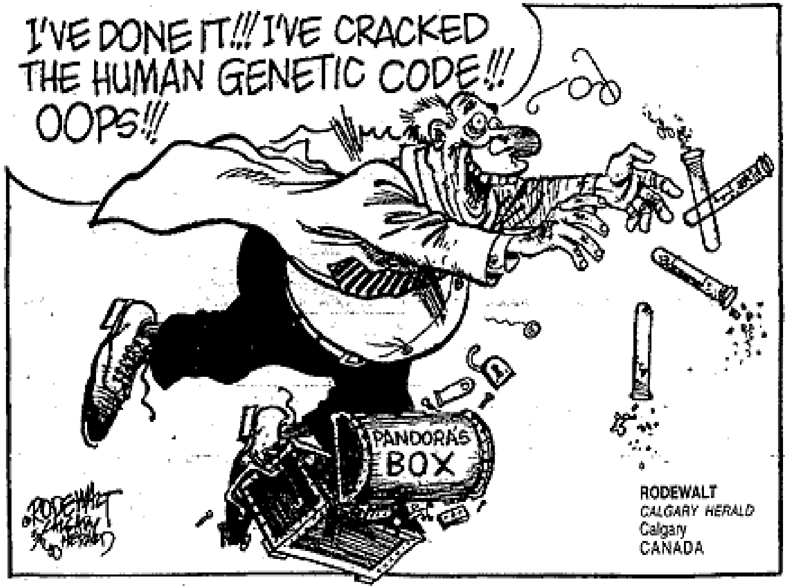
Course Outline

- **Genome Projects: Organization & Objectives**

- Genome Sequencing & Annotation
- Gene Expression & the Transcriptome
- Proteomics & Functional Genomics
- Integrative Genomics & Bioinformatics Tools

Textbooks

- G Gibson & SV Muse (2002) A primer of Genome Science. Sinauer Associates, Inc. Publishers.
  - Chapter 1: Genome Projects: Organization & Objectives
- K. Davis (2001) 基因組圖譜解密。潘震澤譯。Cracking the Genome (Inside the Race to Unlock Human DNA) 。時報出版社。Taiwan。



"All the News That's Fit to Print"

The New York Times

National Edition

Southern California: Mostly sunny with light winds. Highs ranging from the 70's along the beaches to over 100 in the deserts. Tonight, mainly clear, low 65-70. Weather map, Page A24.

VOL. CXLIX . . . . No. 51,432 Copyright © 2000 The New York Times TUESDAY, JUNE 27, 2000 Printed in California ONE DOLLAR

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part of our national culture."

The court said in an opinion by Chief Justice William H. Rehnquist that because the 1966 Miranda decision "announced a constitutional rule," a statute by which Congress had sought to overrule the decision was itself unconstitutional.

Miranda had appeared to be in jeopardy, both because of that long-ignored but recently rediscovered law, by which Congress had tried to overrule Miranda 32 years ago, and because of the court's perceived hostility to the original decision.

The chief justice said, though, that the 1968 law, which replaced the Miranda warnings with a case-by-case test of whether a confession was voluntary, could be upheld only if the Supreme Court decided to overturn Miranda. But with Miranda having "become embedded in routine police practice" without causing any measurable difficulty for prosecutors, there was no justification for doing so, he said. [Excerpts, Page A18.]

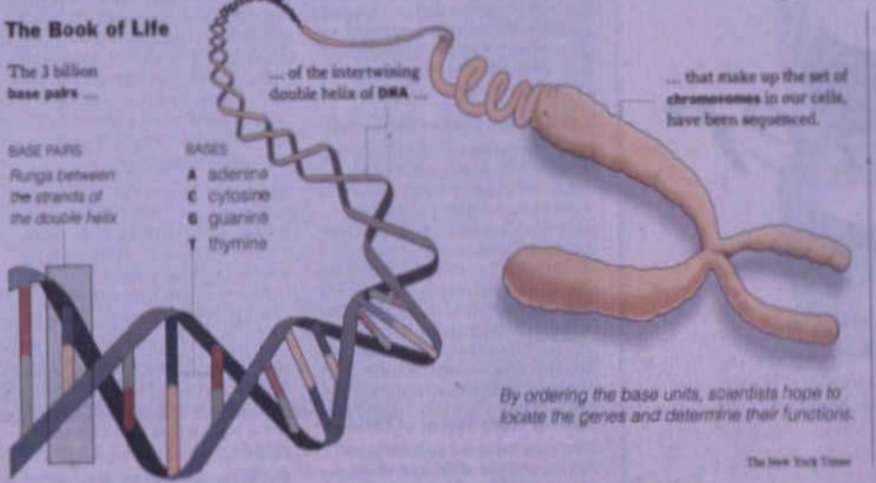
Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presumption that a confession was not voluntary unless preceded by the warnings was not required by the Constitution.

The decision today — only 14 pages long, in Chief Justice Rehnquist's typically spare style — brought an abrupt end to one of the odder episodes in the court's recent history, an intense and strangely delayed refighting of a previous generation's battle over the rights of criminal suspects. Miranda v. Arizona was a hallmark of the Warren Court, and Chief Justice Rehnquist, despite his record as an early and tenacious critic of the decision, evidently did not want its repudiation to be an imprint of his own tenure.

There was considerable drama in the courtroom today as the chief justice announced that he would deliver the decision in the case, Dickerson v. United States, No. 99-5525. The announcement meant that he was the majority opinion's author. Given his statements over more than 25 years about Miranda's lack of constitutional foundation, there was the

The Book of Life


The 3 billion **base pairs** ...
... of the intertwining double helix of **DNA** ...
... that make up the set of **chromosomes** in our cells, have been sequenced.
BASE PAIRS
Rungs between the strands of the double helix
BASES
A adenine
C cytosine
G guanine
T thymine

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

Science Times

A special issue

- Putting the genome to work.
- Some information has already paid research dividends.
- Two research methods, two results
- More articles, charts and photos of the genome effort.
- From Mendel to helix to genome.

Section D

Francis S. Collins, head of the Human Genome Project, right, with J. Craig Venter, head of Celera Genomics, after the announcement yesterday that they had finished the first survey of the human genome.

Paul Hosefros/The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions that defines the human organism.

"Today we are learning the language in which God created life," President Clinton said at a White House ceremony attended by members of the two teams and, via satellite, Prime Minister Tony Blair of England. [Excerpts, Page D8.]

The teams' leaders, Dr. J. Craig Venter, president of Celera Genomics, and Dr. Francis S. Collins, director of the National Human Genome Research Institute, praised each other's contributions and signaled a spirit of cooperation from now on, even though the two efforts will remain firmly independent.

The human genome, the ancient script that has now been deciphered, consists of two sets of 23 giant DNA molecules, or chromosomes, with each set — one inherited from each parent — containing more than three billion chemical units.

The successful deciphering of this vast genetic archive attests to the extraordinary pace of biology's advance since 1953, when the structure of DNA was first discovered and presages an era of even brisker

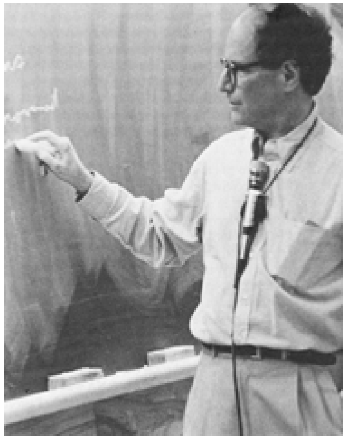
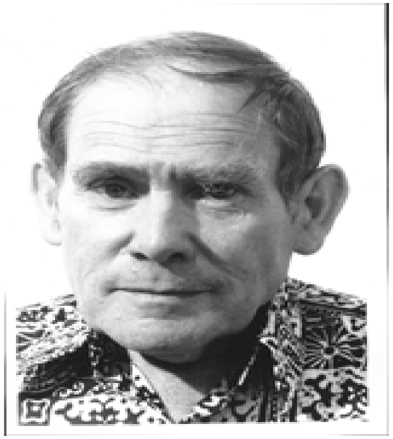
A Pearl and a Hodgepodge: Human DNA

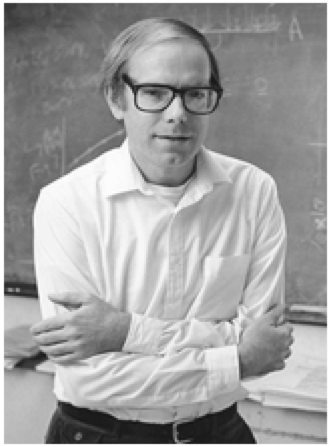
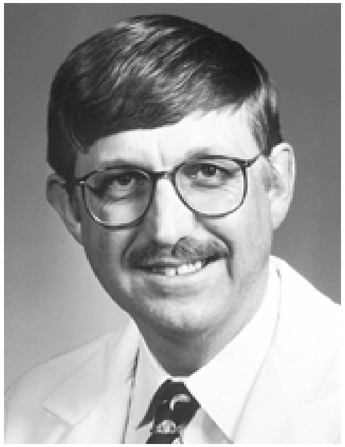
By NATALIE ANGIER

Collins, director of the National Human Genome Research Institute. "We only have to do this once, read-

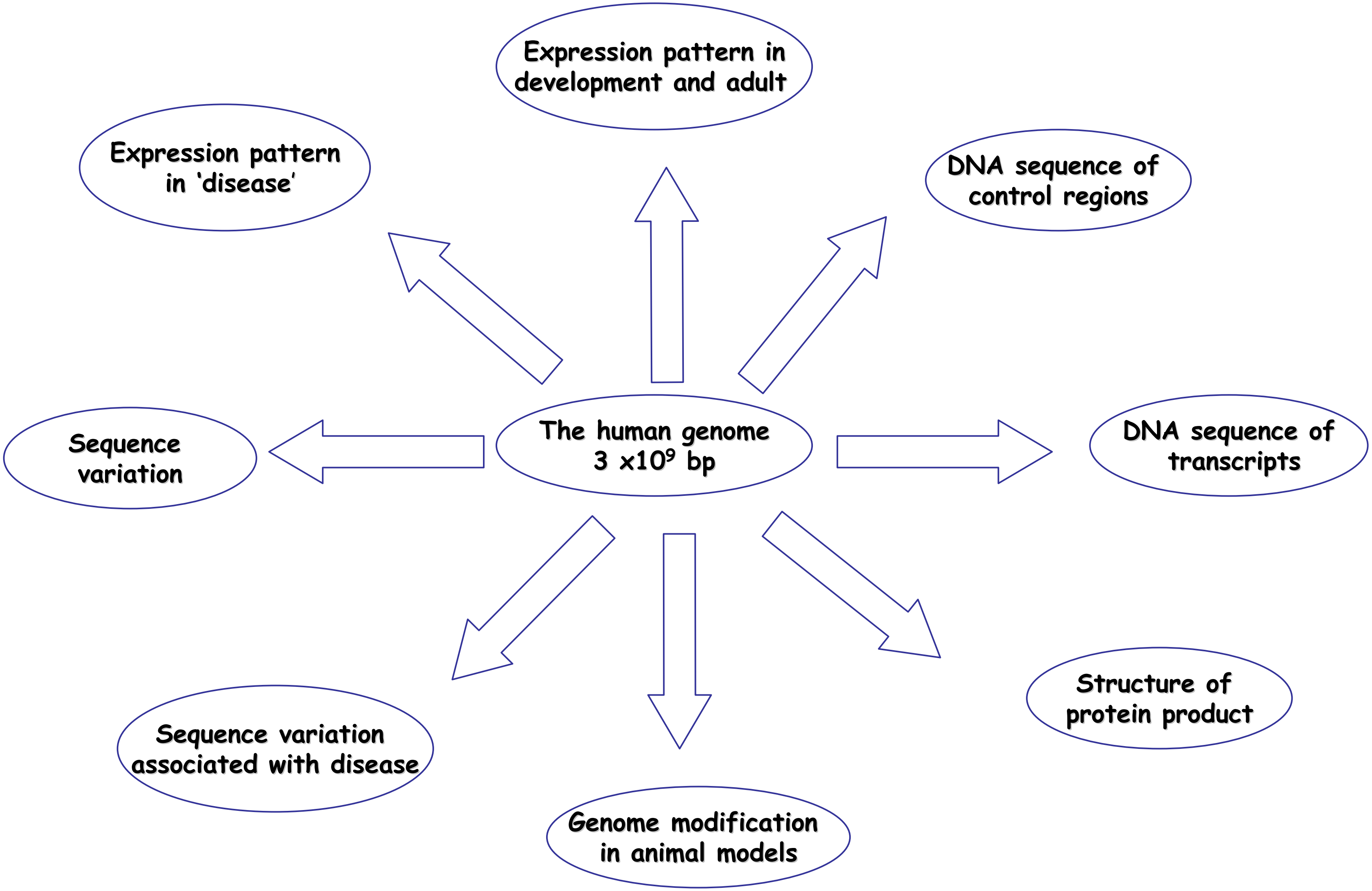
Though scientists underscore the importance of their accomplishment by calling the genome a "portrait of

The Genome Crackers

- Walter Gilbert: A crucial early proponent, he later tried to set up a company to produce and sell genome data
- Sydney Brenner: Joked that sequencing was so boring it should be done by prisoners.
- Charles DeLisi: An early advocate, he launched the Human Genome Initiative within the **Department of Energy** in 1986.
- Maynard Olson: Helped pave the way with work on mapping the yeast genome.
- Francis S. Collins: Favored a deliberate, methodical approach to mapping and sequencing.
- J. Craig Venter: Threw down the gauntlet with his commercial plan to shotgun sequence the human genome.



The human genome 3 x10^9 bp

Expression pattern in development and adult

DNA sequence of control regions

DNA sequence of transcripts

Structure of protein product

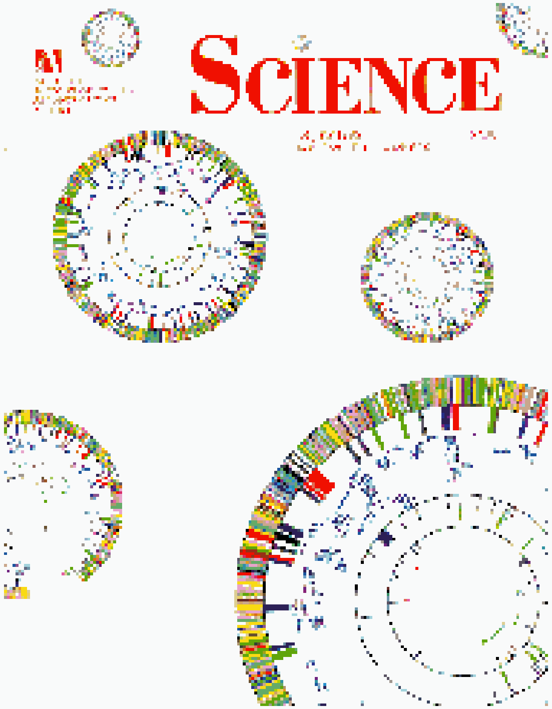
Genome modification in animal models

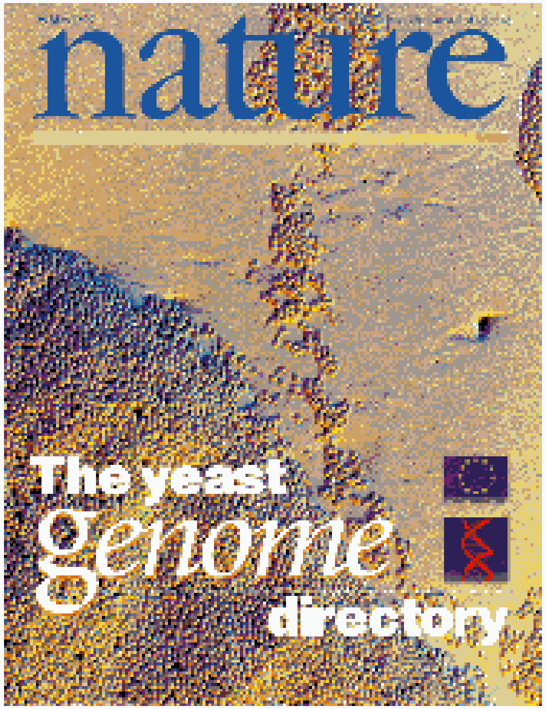
Sequence variation associated with disease

Sequence variation

Expression pattern in ‘disease’

Genomes highlight the **Finiteness** of the World of Sequences





1995

Bacteria, 1.6 Mb, ~1600 genes [*Science* 269: 496]

1997

Eukaryote, 13 Mb, ~6K genes [Nature 387: 1]

1998

Animal, ~100 Mb, ~20K genes [*Science* 282: 1945]

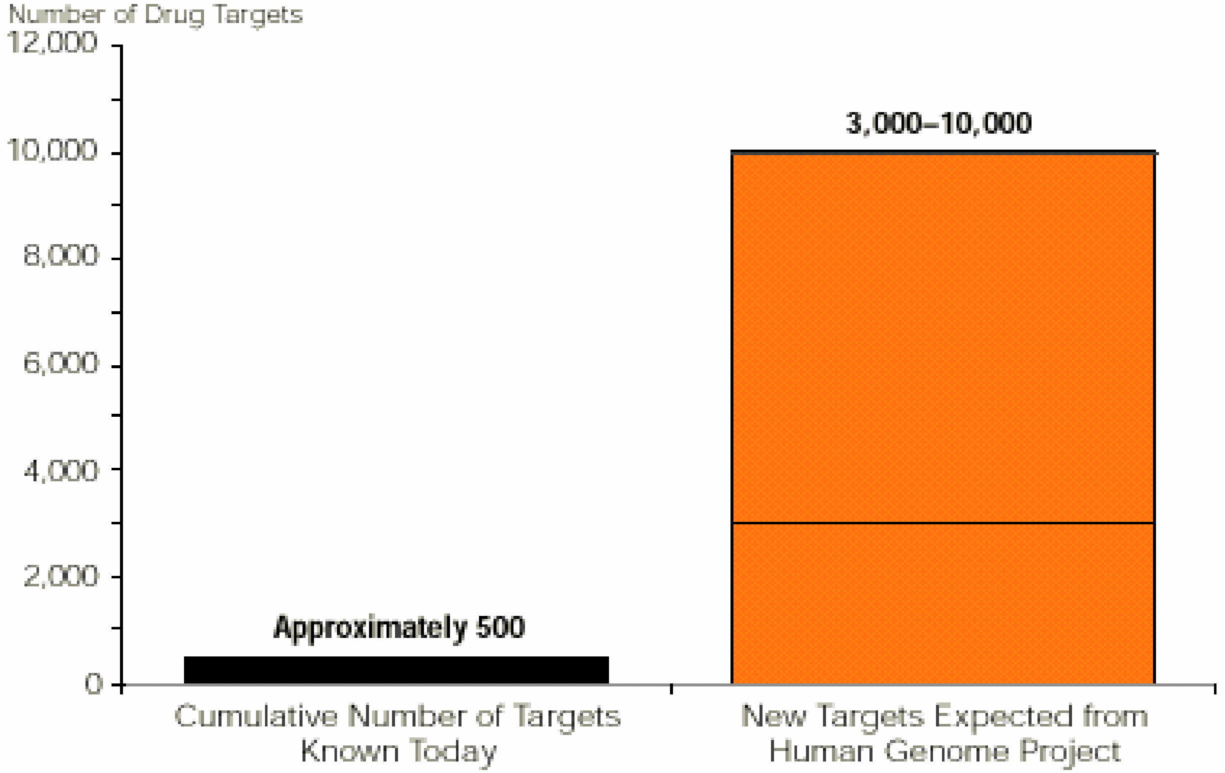
2000?

Human, ~3 Gb, ~100K genes [???]

Genomics Revolution

The Opportunity & the Hope: New Targets, New Therapies

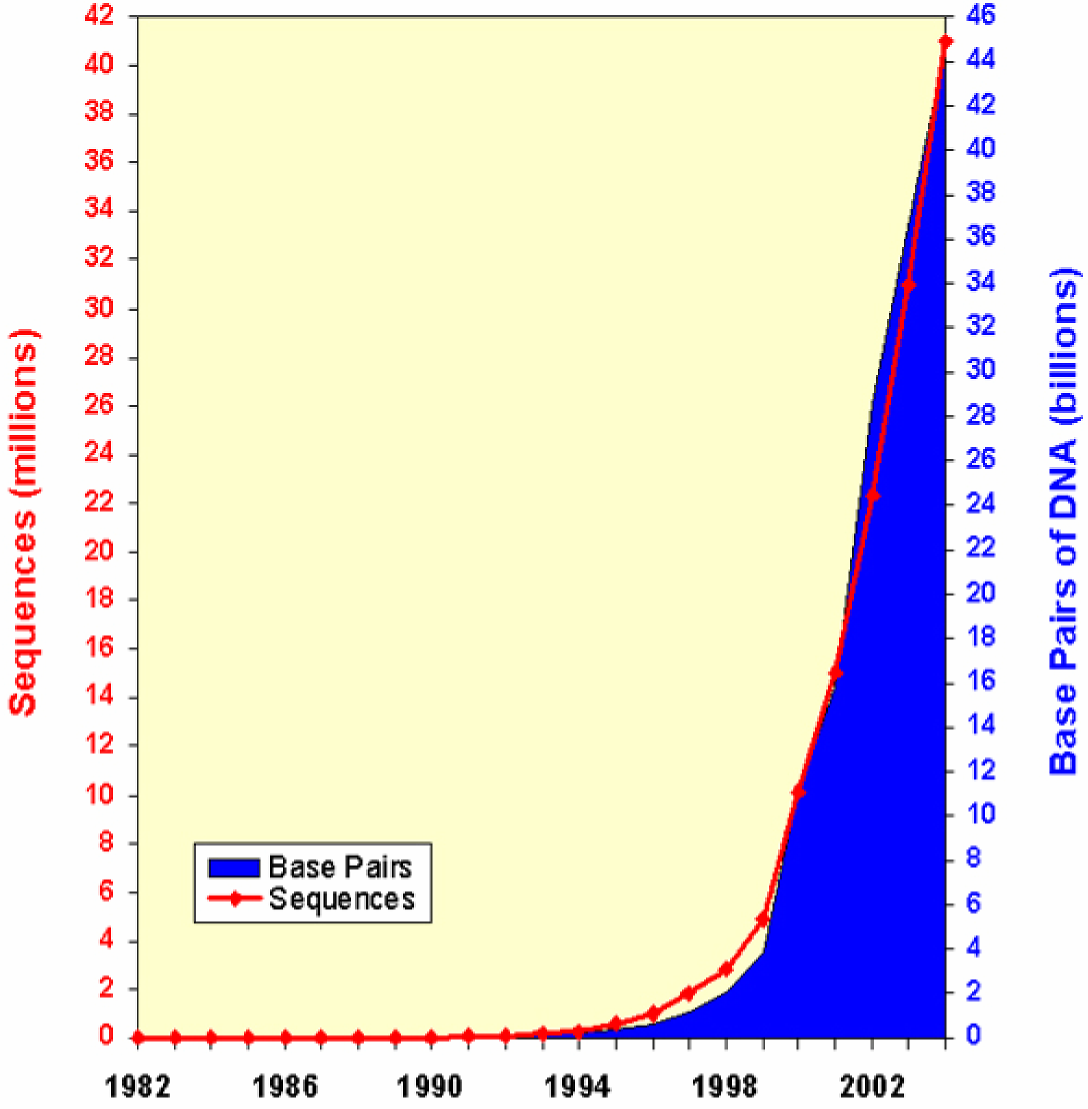
HUMAN GENOME PROJECT TO SPARK EXPONENTIAL GROWTH IN NUMBER OF TARGETS FOR DRUG INNOVATION



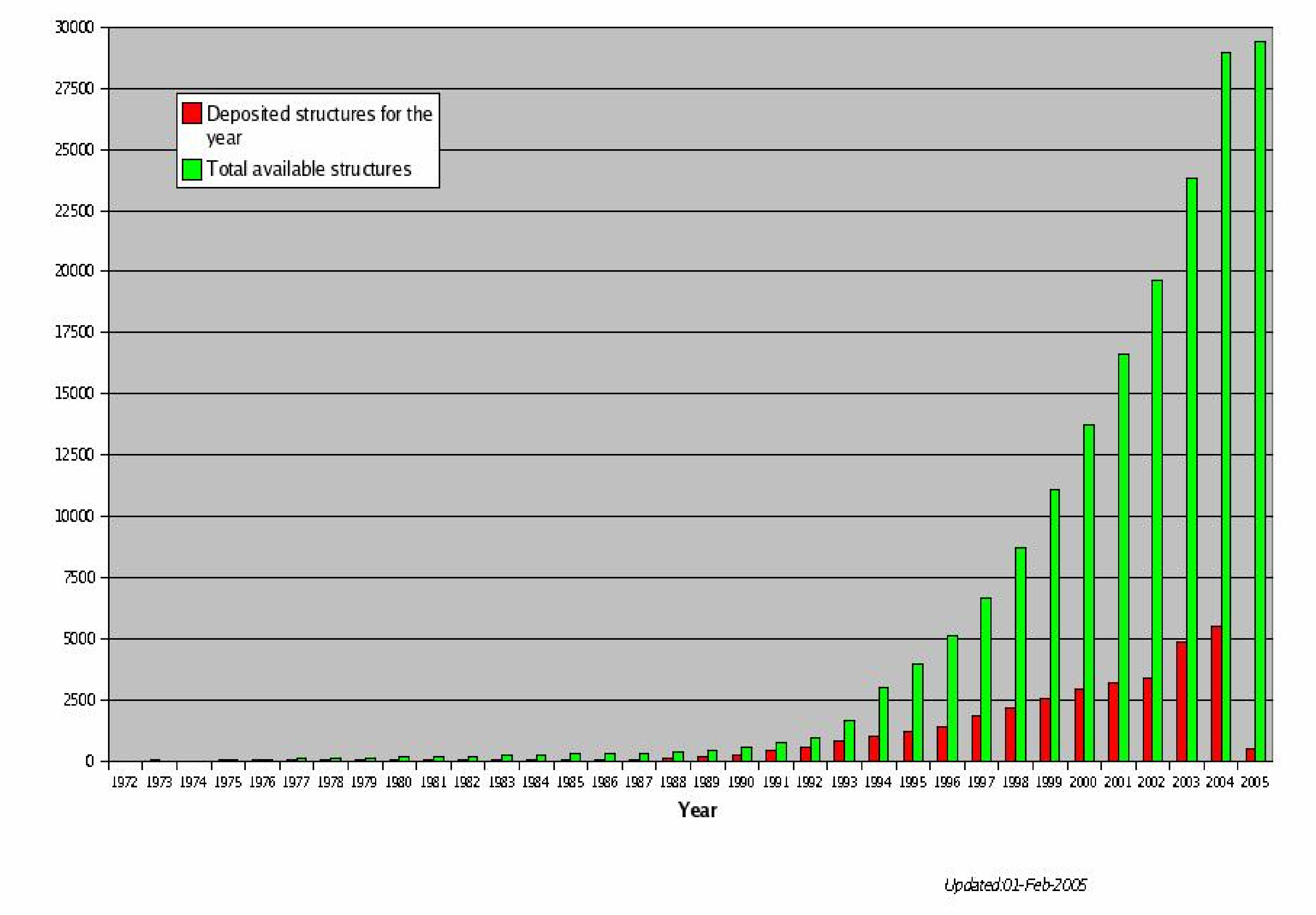
Source: Drews, Jurgen, M.D., "Genomic Sciences and the Medicine of Tomorrow: Commentary on Drug Development," Nature Biotechnology, Vol. 14, November 1996.

Growth of GenBank

(1982 - 2004)



National Center for Biotechnology Information (NCBI), USA



Updated:01-Feb-2005

Protein Data Bank (PDB, RCSB, USA)

Human Genome Project

Pharmaceuticals

Agriculture

Forensics

Anthropology

Bioremediation

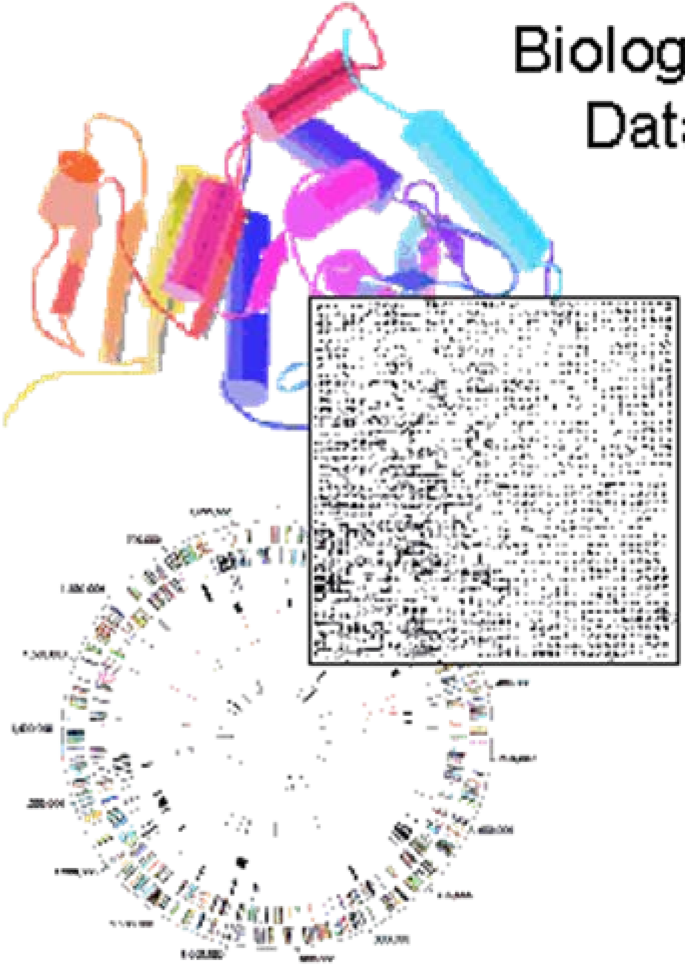
Industrial Processes

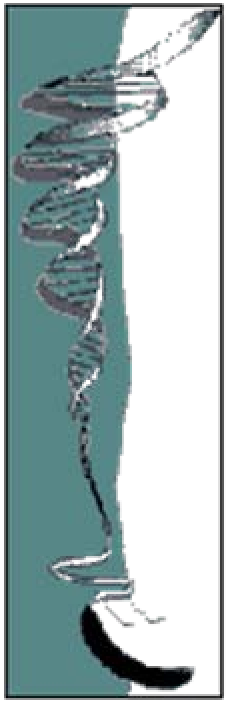
Biofuels

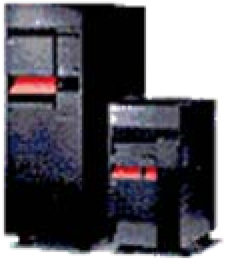
YGA 98-1453

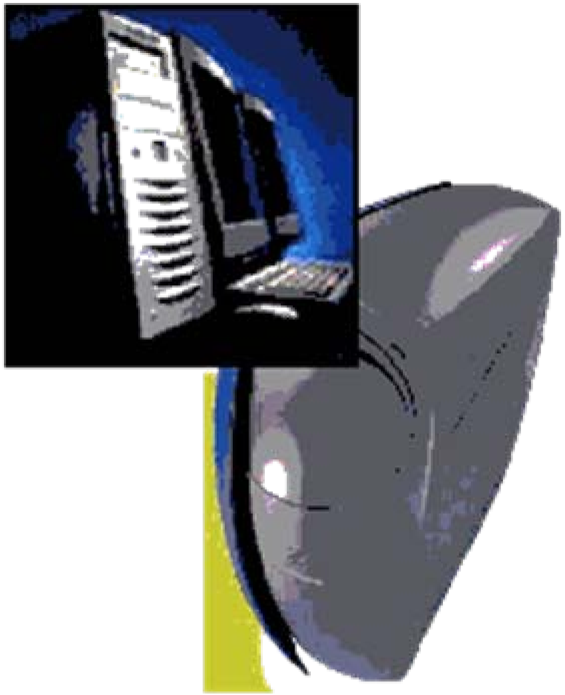
Bioinformatics

Biological Data + Computer Calculations









Consequences of the Human Genome Project (HGP)

- Complete sequencing of the Human Genome
- New branch of science and medicine
  - **Genomics**
  - Bioinformatics
  - Etc.

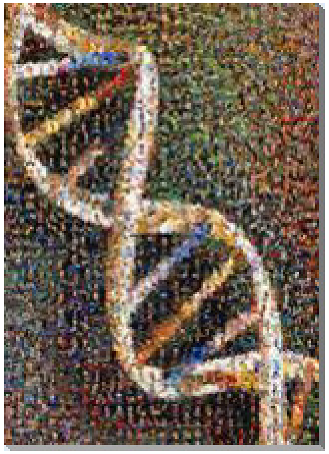
What is a Genome

- All of the DNA for an organism
  - One copy

- **Human genome**
  - N = 22 +XY
  - Nucleus
    - 3.2 billion base pairs packaged into **chromosomes**
  - Mitochondrion **(extra-nuclear)**
    - 16.5 Kb packaged into one circular chromosome

Goals of the Human Genome Project (HGP)

genome.gov
National Human Genome Research Institute



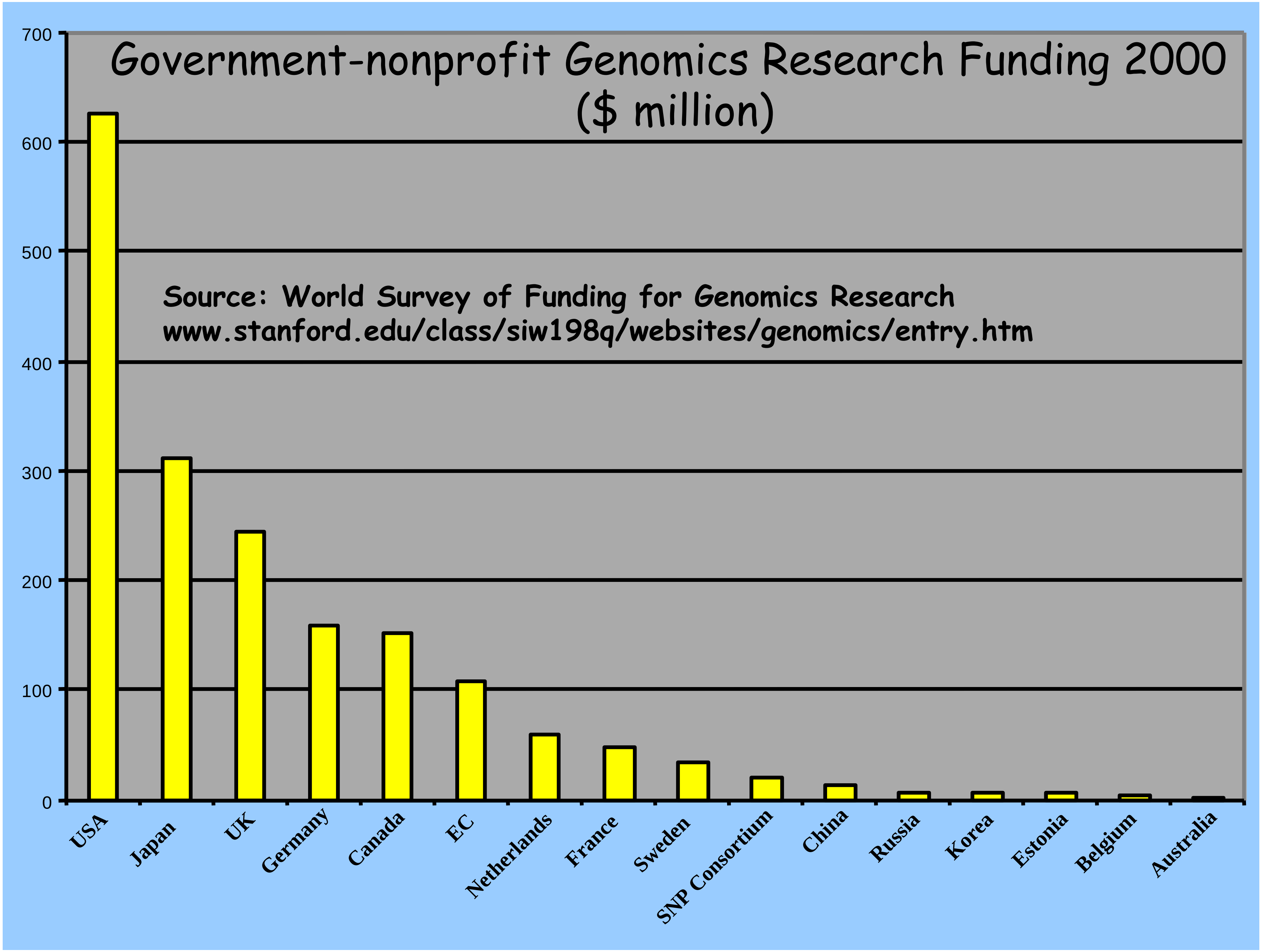
- http://www.genome.gov/page.cfm?pageID=10001694
- Identify all the ~30,000 genes in human DNA
- Determine the **sequences** of the 3 billion chemical bases that make up human DNA
- **Store this information in databases**
- **Develop tools for data analysis**

Bioinformatics

- Address the ethical, legal, and social issues (**ELSI**) that may arise from the project

Genomic Biology

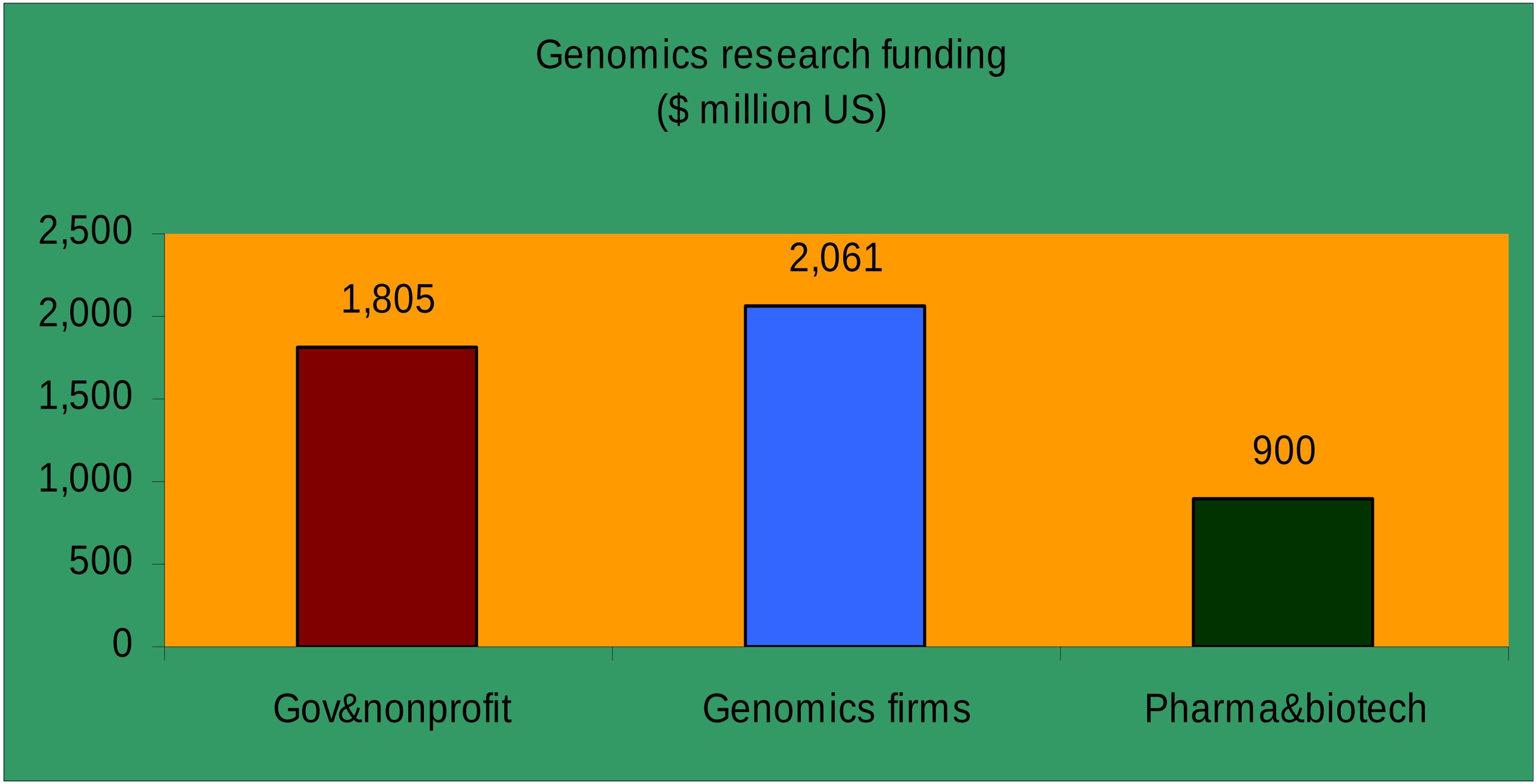
- Genomics is **changing our understanding of biology**
  - **Late 1980s**: the generation & **analysis** of information about genes & genomes
  - **Middle 1990s**: **functional genomics**
    - The generation & analysis of the information about **what genes do**
    - Genomics, proteomics, transcriptomics, metabolitmics etc.
    - [Broad sense] the generation of information about living things by **systematic approaches** that can be performed **on an industrial scale** (**high throughput**)



Government-nonprofit Genomics Research Funding 2000 ($ million)

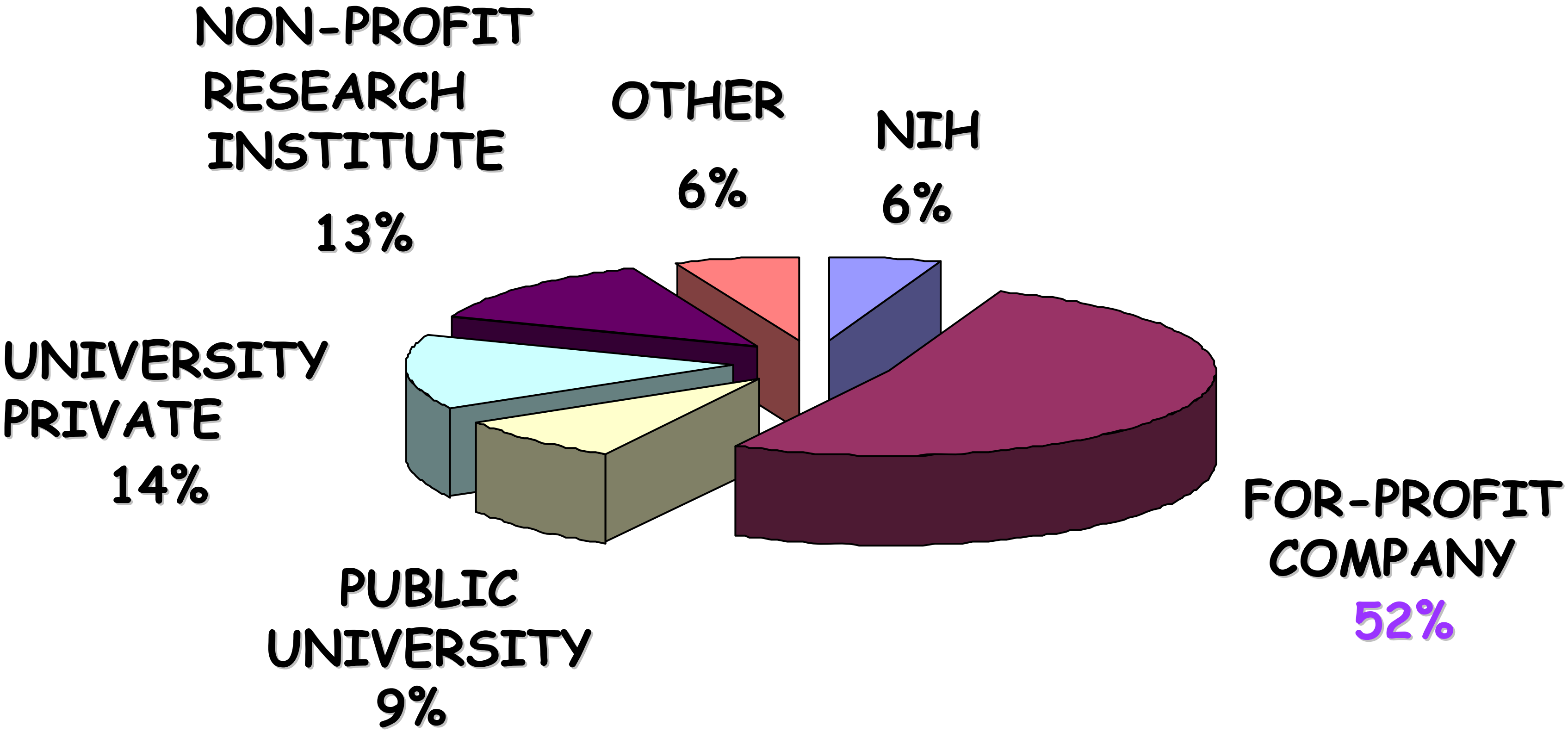
Source: World Survey of Funding for Genomics Research
www.stanford.edu/class/siw198q/websites/genomics/entry.htm

Funding: Private > Public (2000)



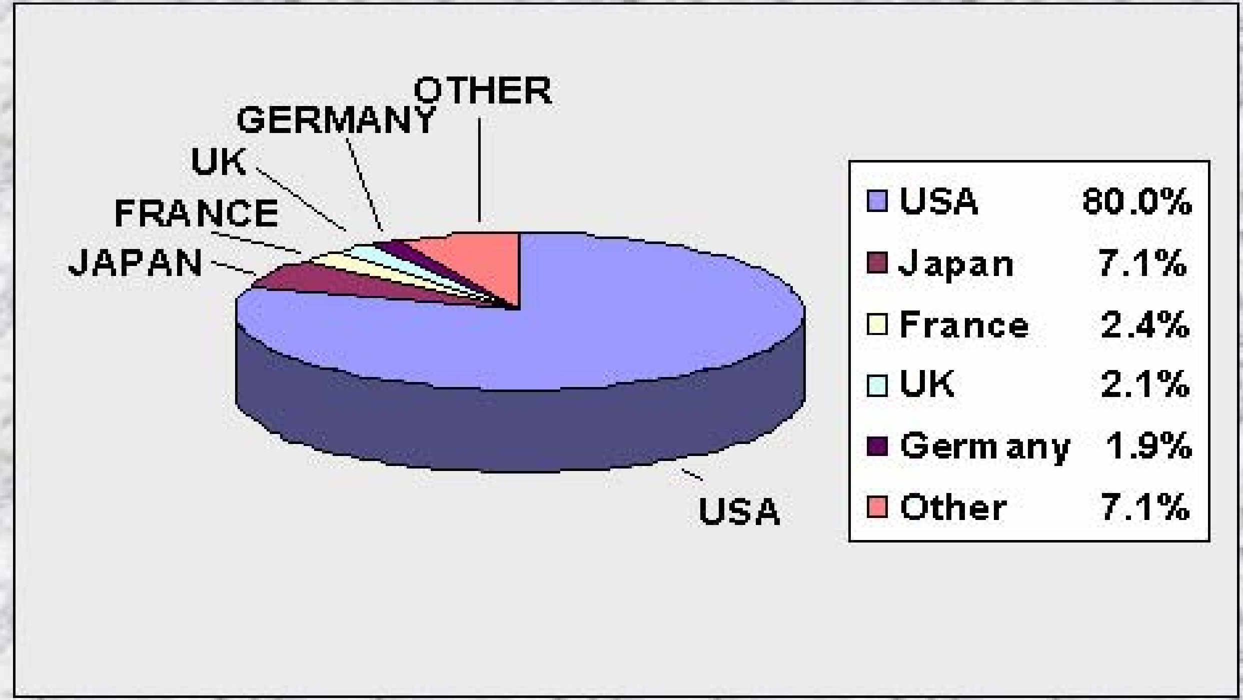
Source: World Survey of Funding for Genomics Research
Stanford in Washington Program
http://www.stanford.edu/class/siw198q/websites/genomics/entry.htm

Patent Assigned

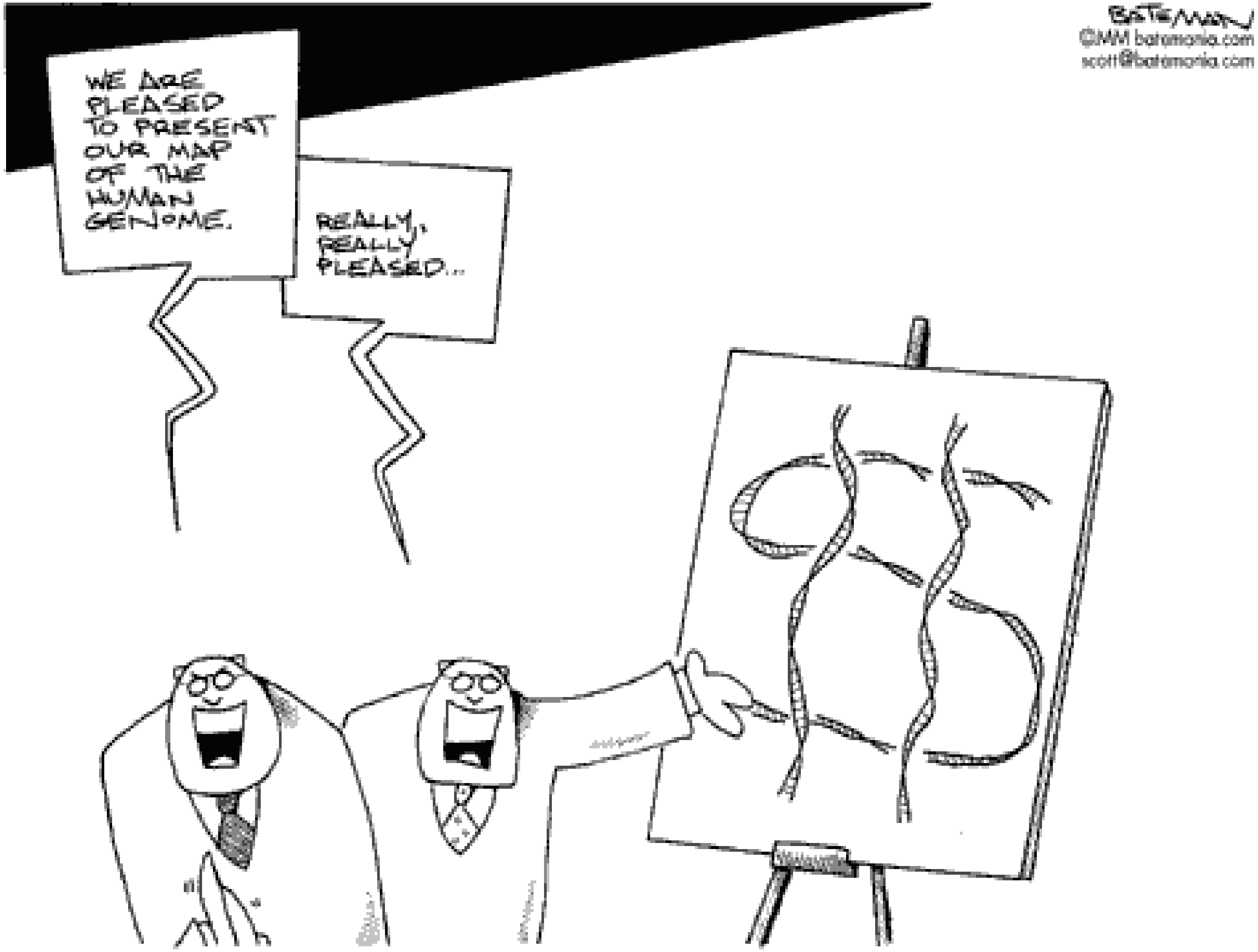


Source: Stephen McCormack and Robert Cook-Deegan
DNA Patent Database www.genomic.org

Ownership (assignee country) of 1028 DNA-based patents 1980-1993

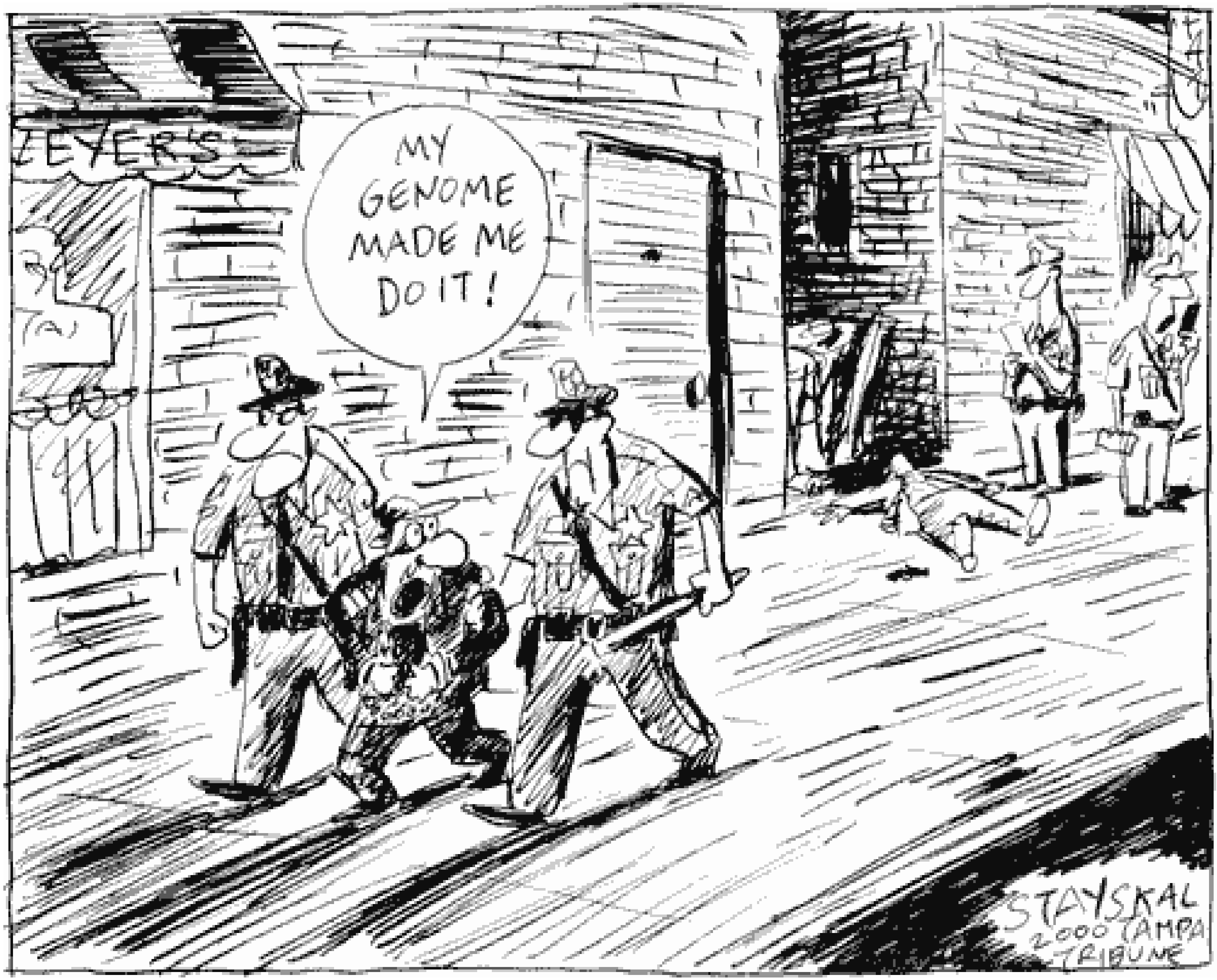


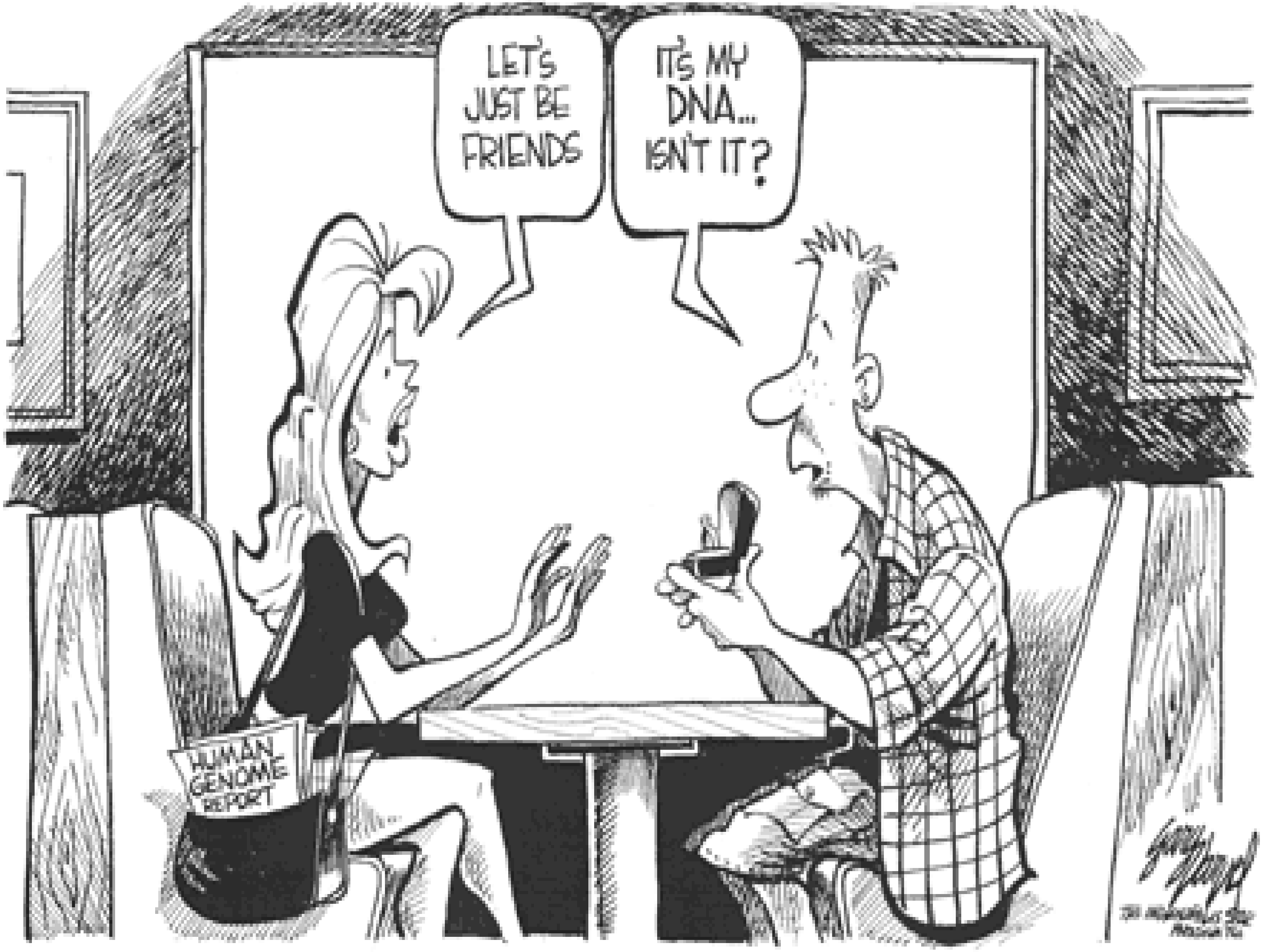
Source: Stephen McCormack and Robert Cook-Deegan DNA Patent Database, August 1999, www.genomic.org

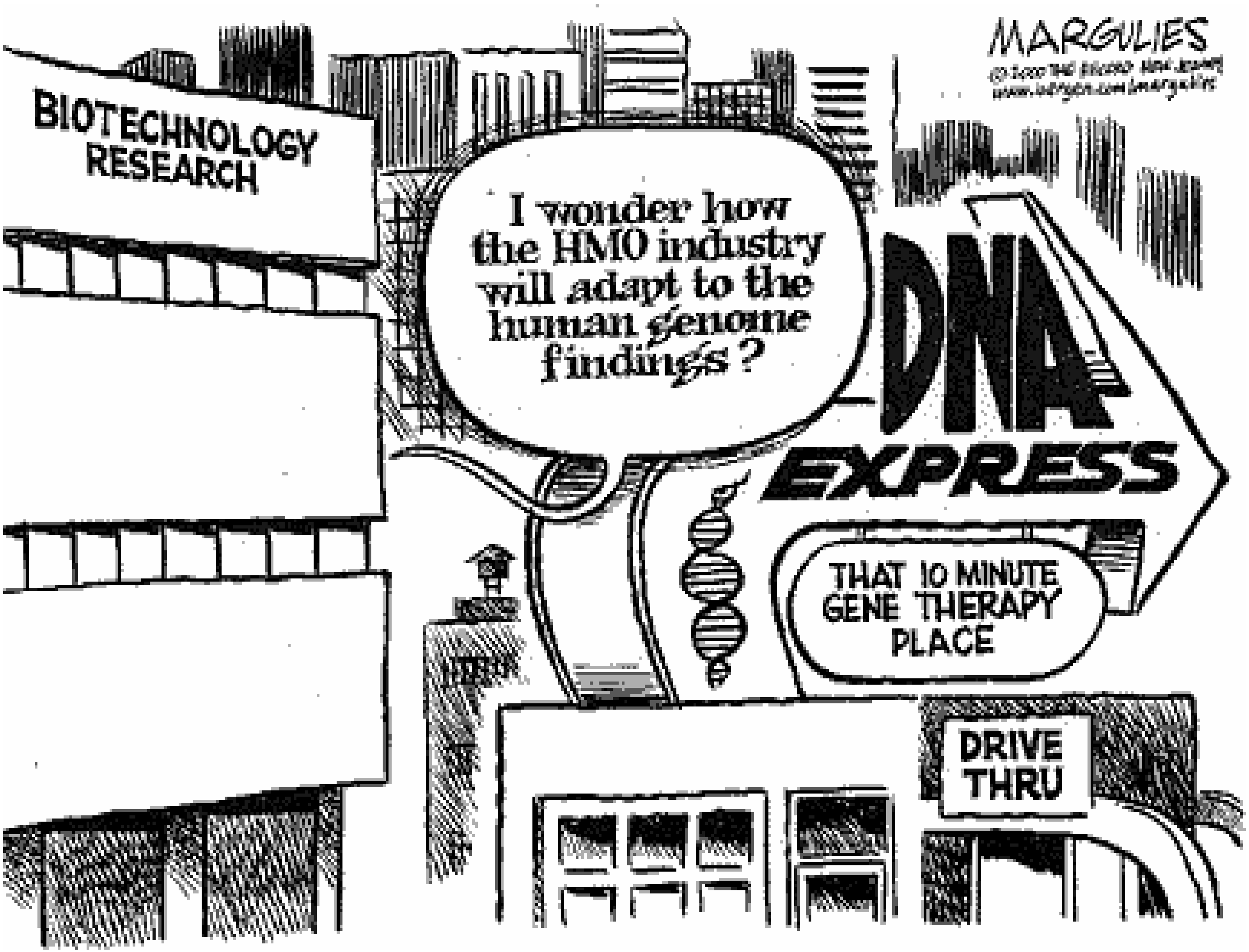

BATEMAN
©MM batemania.com
scott@batemania.com
WE ARE PLEASED TO PRESENT OUR MAP OF THE HUMAN GENOME.
REALLY, REALLY PLEASED...

Timetable of HGP

- Begun formally in 1990
- The project originally was planned to last 15 years
- Rapid technological advances have accelerated the expected completion date to 2003
- Celera announces a 3-year plan to complete the project early
- First draft: June 28th, 2000
  - Sequencing completed first: chromosome 22 (Dec. 2nd 1999, Nature)
- Feb. 2001
  - June 2002 (**TIGR**): 7,801 genes' functions identified
  - International Human Genome Sequencing Consortium: http://www.nature.com (Nature)
  - The Celera database: http://www.sciencemag.org (Science)


EYERS
MY GENOME MADE ME DO IT!
STAYSKAL 2000 TAMPA TRIBUNE


LET'S JUST BE FRIENDS
IT'S MY DNA... ISN'T IT?
HUMAN GENOME REPORT


MARGULIES
BIOTECHNOLOGY RESEARCH
I wonder how the HMO industry will adapt to the human genome findings?
DNA EXPRESS
THAT 10 MINUTE GENE THERAPY PLACE
DRIVE THRU

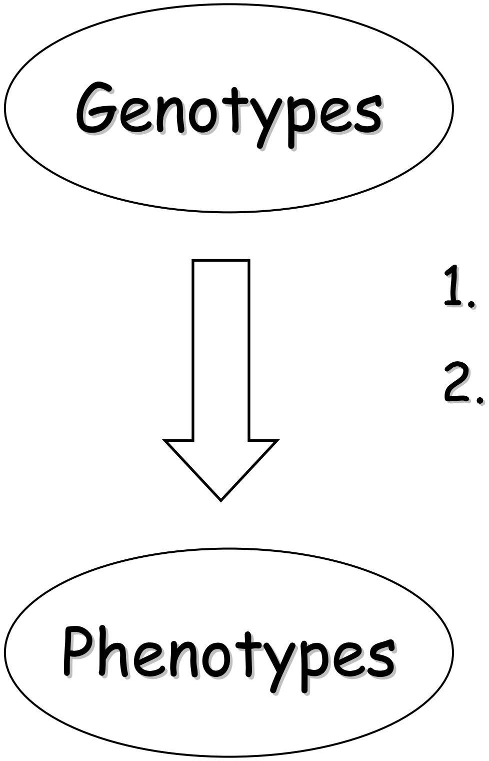
The Core Aims of Genome Sciences (1)

- To establish an integrated Web-based database & research interface
  - Most sites are now build on state-of-the-art **relational databases** & include **innovative software** for **data searches** and **online analysis**

- To assemble physical & genetic maps of the genome
  - For putting together **phenotypic** and **genetic data**
    - Particularly when mapping **disease loci**

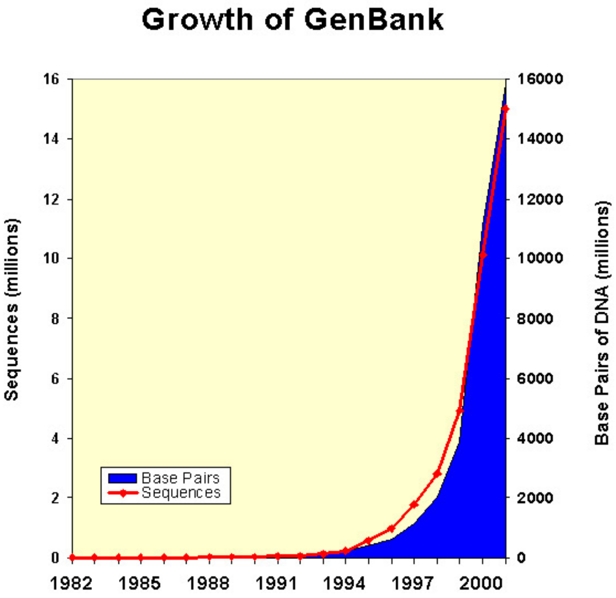
Genotypes vs. Phenotypes

- Genomic DNA: has **almost** all the information about life



1. Environments
2. Interactions and regulations among genes

Sequence



Proteome

Phenotype

The Core Aims of Genome Sciences (2)

- **To generate & order genomic and expressed gene sequences**
  - “Top-down” vs. “shotgun” (next lecture)
  - cDNA (from mRNA, complementary)
  - ESTs (Expressed Sequence Tags)
    - Only one end of a cDNA need be sequenced to identify a clone, fragments
    - A good first approximation of the diversity of genes expressed in a tissue

The Core Aims of Genome Sciences (3)

- **To identify & annotate the complete set of genes encoded within a genome**
  - Using a combination of **experimental** & **bioinformatics strategies**
    - Aligning cDNA & genomic sequences
    - Looking for sequences that are similar to those already identified in other genome, *e.g.*, **BLAST**
    - Applying **gene-finding software** that recognizes DNA features that associated with genes, *e.g.*, open reading frames (ORFs), transcription start and termination sites, **exon/intron boundaries**

Gene Annotation

- Entitles linking its **sequence** to genetic data about the **function**, **expression**, and **mutant phenotypes** of the **protein** associated with the **locus**, as well as to **comparative data** from homologous proteins in **other species**

The Core Aims of Genome Sciences (4)

- **To compile atlases of gene expression**
  - Analyzing profiles of transcription & protein synthesis
    - **Traditional methods**
      - Northern blotting, *in situ* hybridization, Western blotting, immunohistochemistry
    - **Genomic methods**
      - EST sequencing, SAGE, differential display
      - Microarray, gene chips
    - **Bioinformatic methods**
      - Analyzing patterns of covariation in gene expression provides information about the regulation of gene expression, and can yield clues to unknown gene function as a result of **"guilt by association"**

The Core Aims of Genome Sciences (5)

- To accumulate functional data, including biochemical & phenotypic properties of genes
  - **Functional genomics**
    - A panoply of approaches under development to ascertain the **biochemical**, **cellular**, and/or **physiological** properties of each and every gene product
      - Near-saturation mutagenesis
      - High-throughput reverse genetics
      - **Proteomics**
        - Detecting protein expression
        - Detecting protein-protein interactions
  - **Structural genomics**
    - To elucidate **the tertiary structure** of each class of protein found in cells

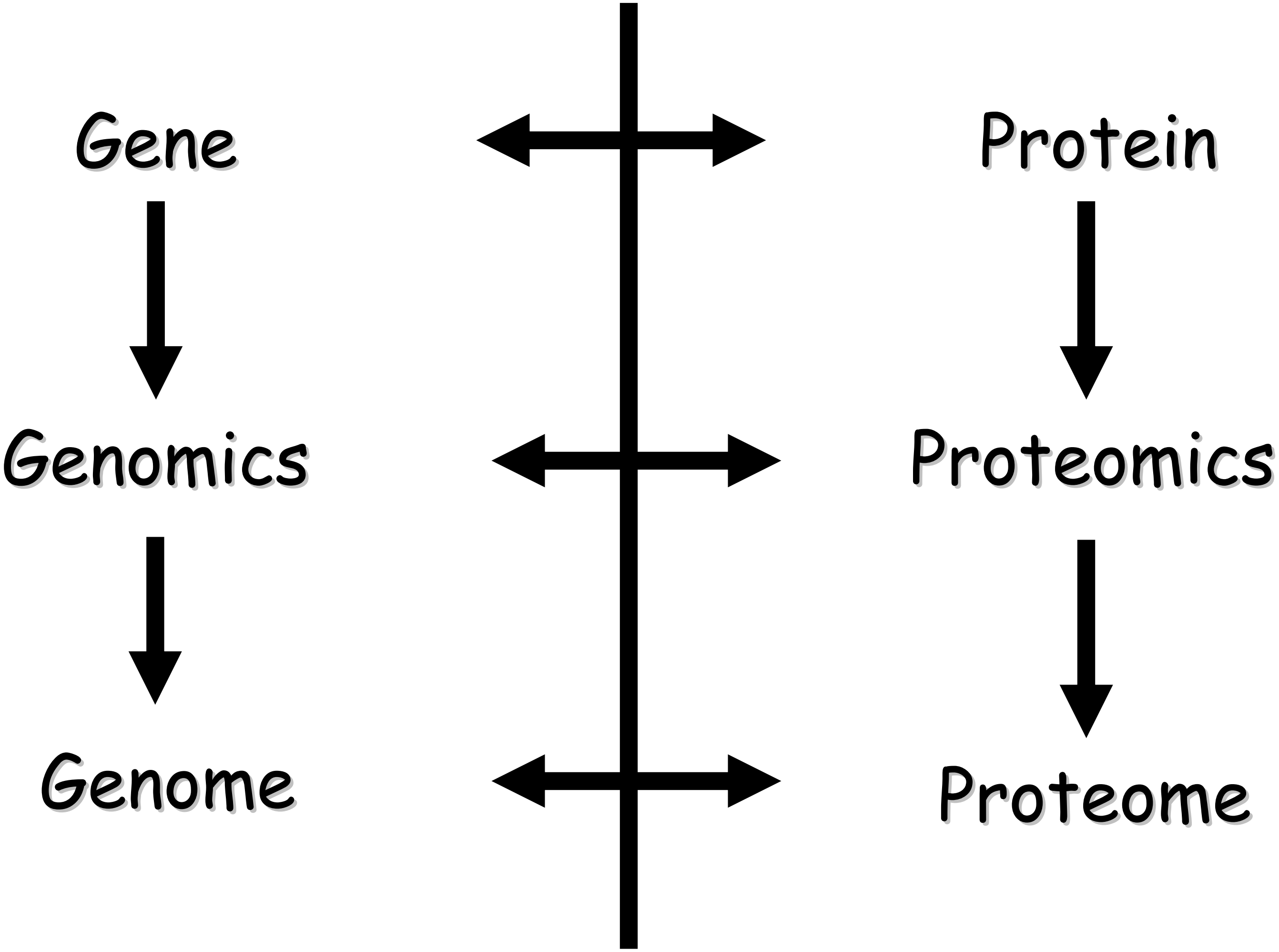
Genomics

- **Genetic Markers**
  - Blood group, allozyme, RFLPs, STRs, EST, STS & SNP
- **Gene Location (**Mapping)
  - Physical mapping (pseudogenetics & cytogenetics)
  - Linkage mapping
- **QTL Mapping**
  - **Complex human diseases**
- **Genomic Glossary**
  - http://www.geocities.com/bioinformaticsweb/genomicglossary.html

Proteomics

- The study of **gene expression** at the **protein** level, by the identification and characterization of proteins present in **a biological sample**
- Glossary
  - http://www.genomicglossaries.com/content/proteomics.asp

Linguistic Analogy



Gene

Genomics

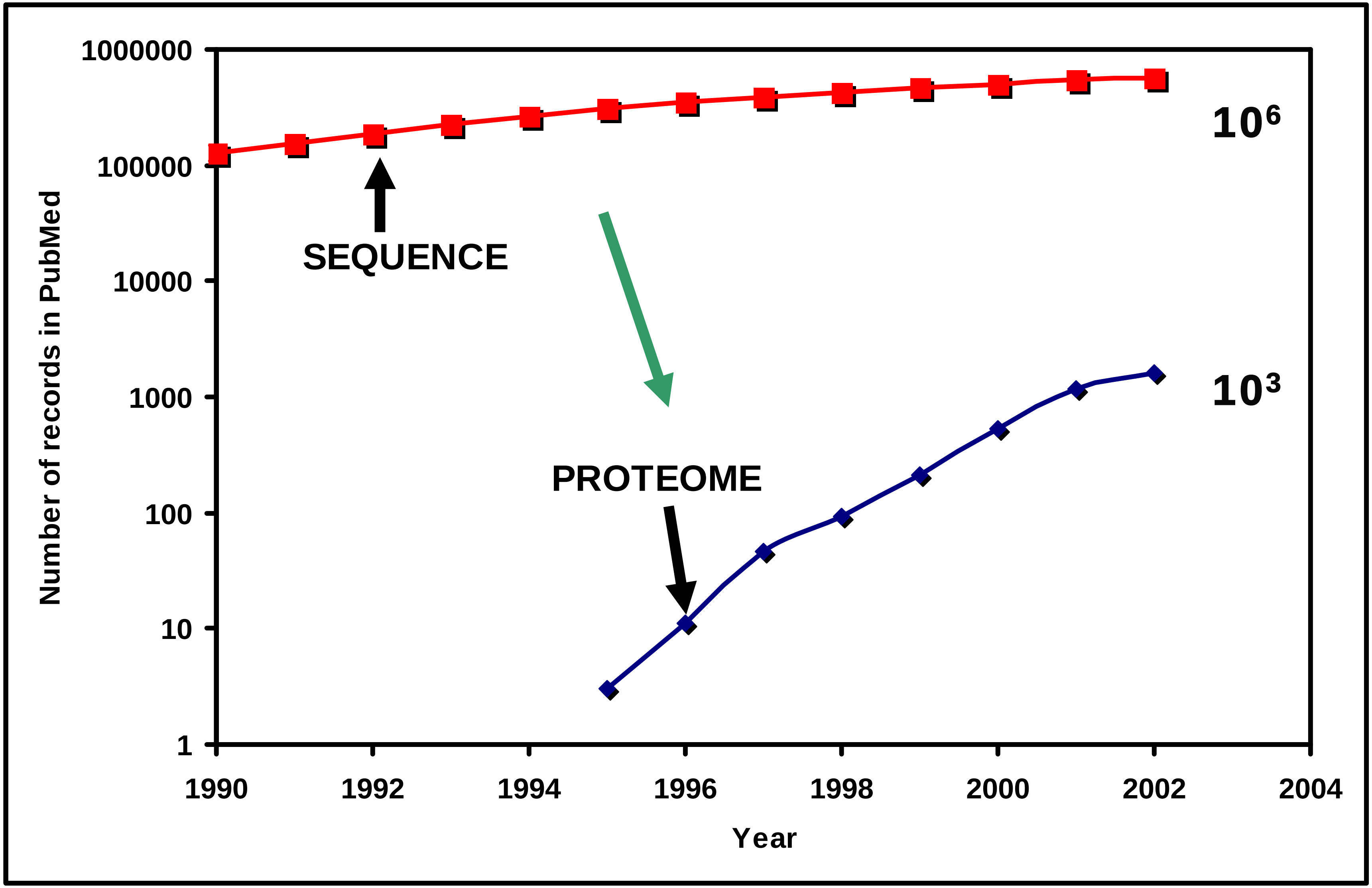
Genome

Protein

Proteomics

Proteome

From Sequences to Proteome



The Core Aims of Genome Sciences (6)

- To accumulate functional data, including biochemical & phenotypic properties of genes
  - **Pharmacogenomics**
    - Comprises the study of variations in targets or target pathways, variation in **metabolizing enzymes** (**pharmacogenetics**) or, in the case of infectious organisms, genetic variations in the pathogen
    - http://www.genomicglossaries.com/content/pharmacogenomics.asp

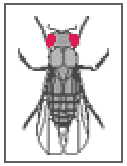
The Core Aims of Genome Sciences (7)

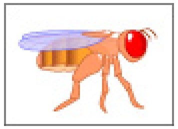
- To characterize DNA sequence diversity
    - All genomes are full of polymorphisms
        - Two or more variants are found in natural populations
        - Single-nucleotide polymorphisms (SNPs)
            - Most quantitative genetic variation
            - Size, shape, yield, and **disease susceptibility** should be traceable to SNPs or to insertion/deletion polymorphisms
        - The level of **linkage disequilibrium (LD)**
            - Nonrandom associations between sites
        - **Disease locus** mapping now generally utilizes detailed knowledge of LD
            - SNPs
            - Microsatellites

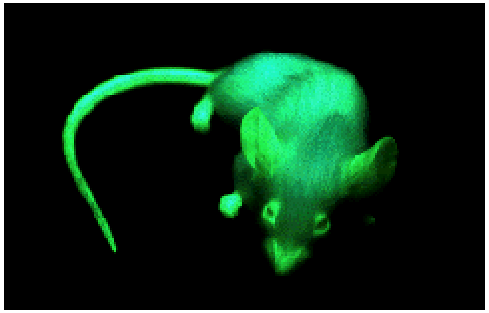
The Core Aims of Genome Sciences (8)

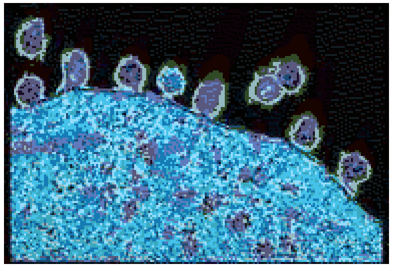
- **To provide the resources for comparison with other genomes**
  - "Nothing in biology makes sense except in the light of evolution" ⇨ "Nothing in genomics makes sense except in the light of **comparative data**"
  - **Synteny**
    - **Local gene order** along a chromosome tends to be **conserved** over millions of years
      - Comparative maps allow **genetic data** from **one species** to be used in the analysis of another
  - The conservation of **gene function**

Organism-specific Resources



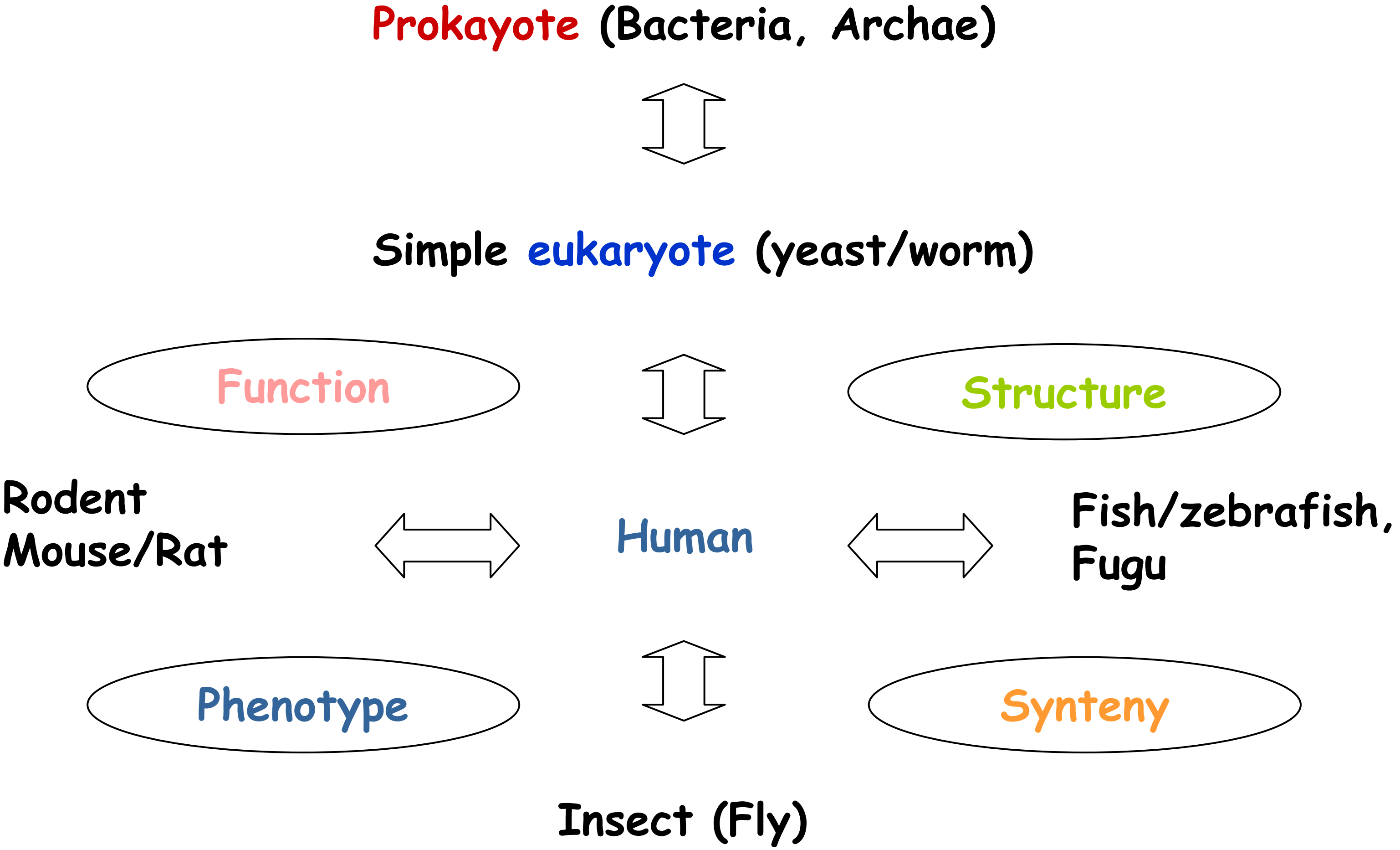






- *Human*
- *Drosophila*
- *Zebrafish*
- *Malaria parasite*
- *Microbial Genomes* (84 complete genomes, Aug. 2002)
- *Mouse*
- *Plant Genome Central*
- *Rat*
- *Retroviruses*

Model Organism Have a Fundamental Role in Assigning Function to Novel Genes (Rastan & Beeley 1997)



Prokayote (Bacteria, Archae)

Simple eukaryote (yeast/worm)

Function

Structure

Rodent Mouse/Rat

Human

Fish/zebrafish, Fugu

Phenotype

Synteny

Insect (Fly)

Definition of Bioinformatics (1)

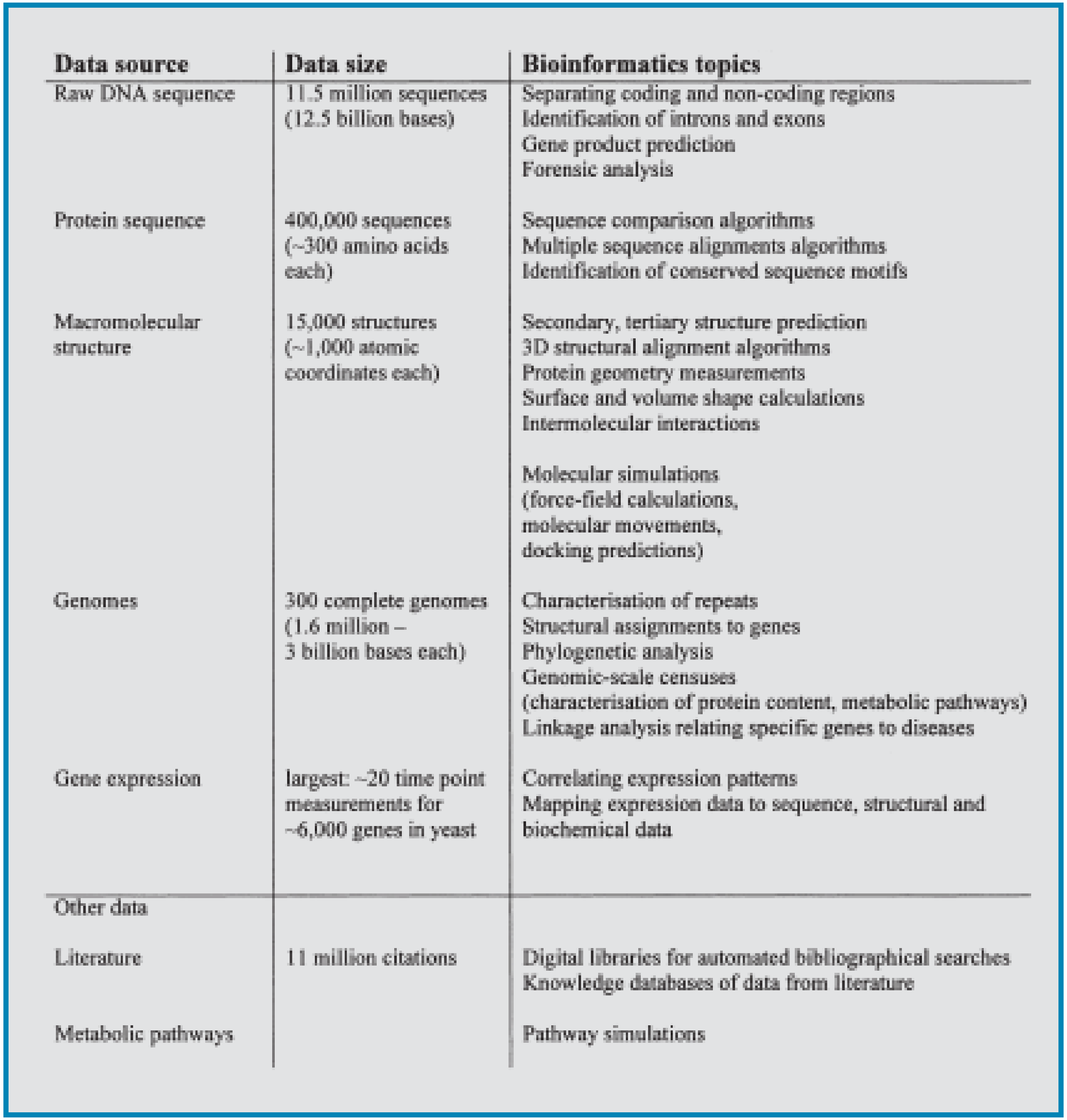
- Computational Biology
- Conceptualizing **biology** in terms of **molecules** (in the sense of physical-chemistry) and then applying "Informatics" techniques
  - Applied Math.
  - Computer Science
  - Statistics
  - Biology **(genomics)**
- To understand and organize the information associated with these molecules, **on a large-scale**

Definition of Bioinformatics (2)

- The "MIS" for **molecular biology** information
    - Management Information System (MIS)

- [Gibas C & Jambeck P 2001] A subset of the larger field of computational biology, the application of quantitative analytical techniques in modeling biological systems

Table 1 Sources of data used in bioinformatics, the quantity of each type of data that is currently (April 2001) available, and bioinformatics subject areas that utilize this data.



Data source	Data size	Bioinformatics topics
Raw DNA sequence	11.5 million sequences (12.5 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	400,000 sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs
Macromolecular structure	15,000 structures (~1,000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions Molecular simulations (force-field calculations, molecular movements, docking predictions)
Genomes	300 complete genomes (1.6 million – 3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterisation of protein content, metabolic pathways) Linkage analysis relating specific genes to diseases
Gene expression	largest: ~20 time point measurements for ~6,000 genes in yeast	Correlating expression patterns Mapping expression data to sequence, structural and biochemical data
Other data		
Literature	11 million citations	Digital libraries for automated bibliographical searches Knowledge databases of data from literature
Metabolic pathways		Pathway simulations

Luscombe *et al.* 2001

Contents & Goal

- Algorithms
- Databases
- User interfaces
- Statistical methodologies

- To identify **"potentially significant"** results

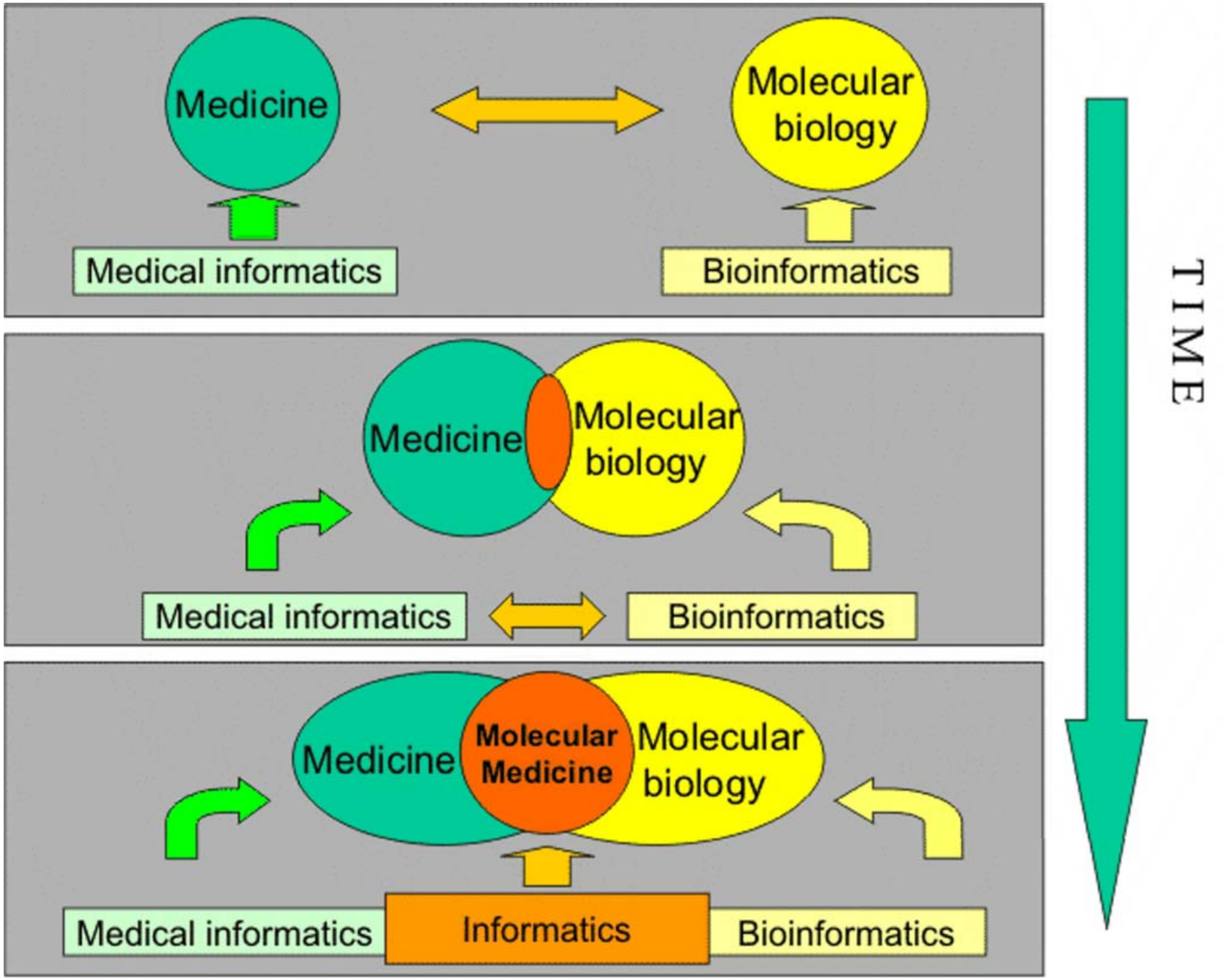
Bioinformatics – Origins & History

- http://www.geocities.com/bioinformaticsweb/his.html

Bioinformatics & Genomic Medicine – JH Kim (2002)

- 1960s
  - Extensive use of **computers** in the **medical sciences**
- 1974
  - Russian "**informatika**" = English **"medical informatics"**
- 1990s
  - **Modern bioinformatics**
    - The convergence of **bioinformatics** and **clinical informatics** (biochemistry a generation ago)

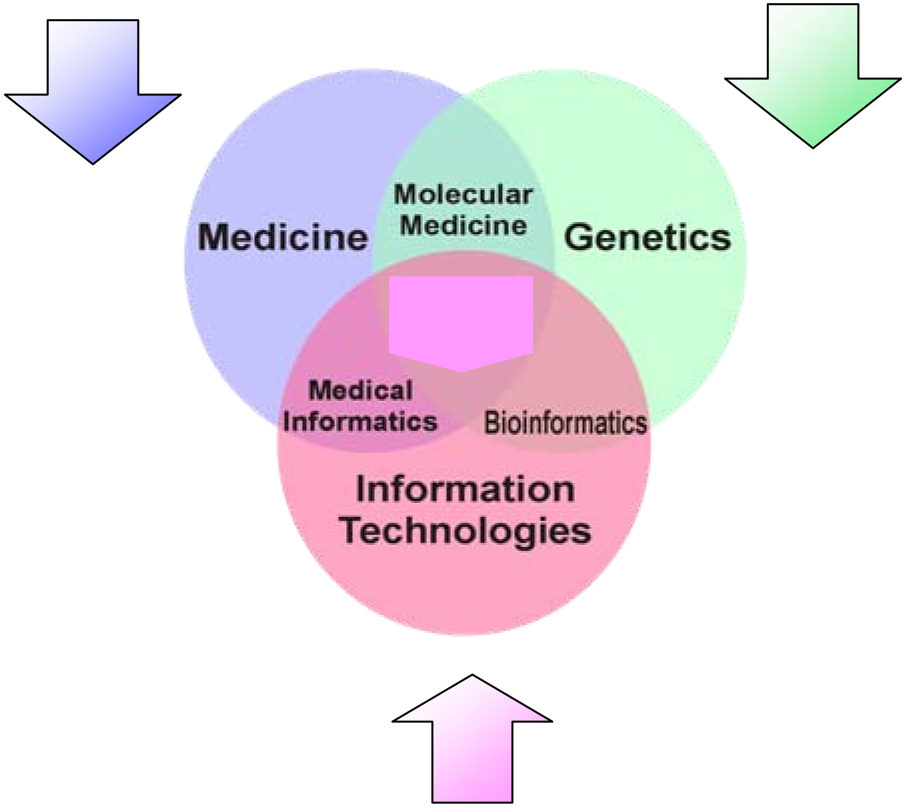
The Convergence between MI & BI



A Model to Study Interactions

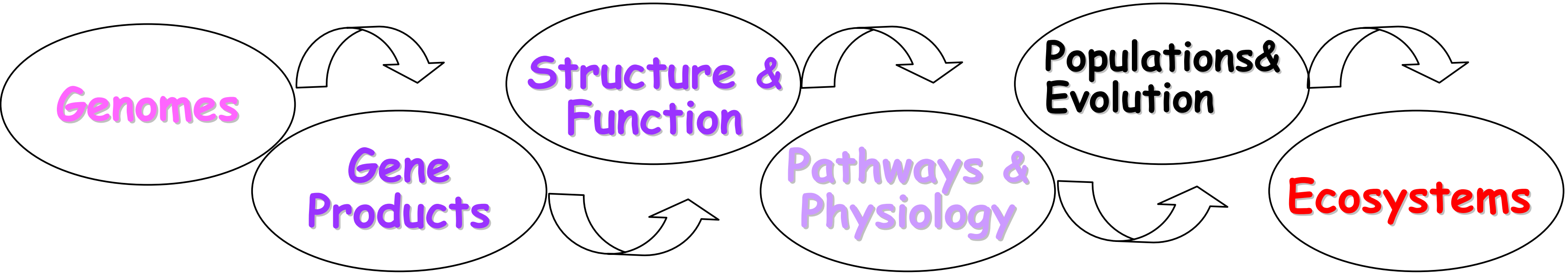
To foster the application of bioinformatics in health

To adapt medical informatics systems to the genetics paradigm



Apply IT to facilitate molecular medicine

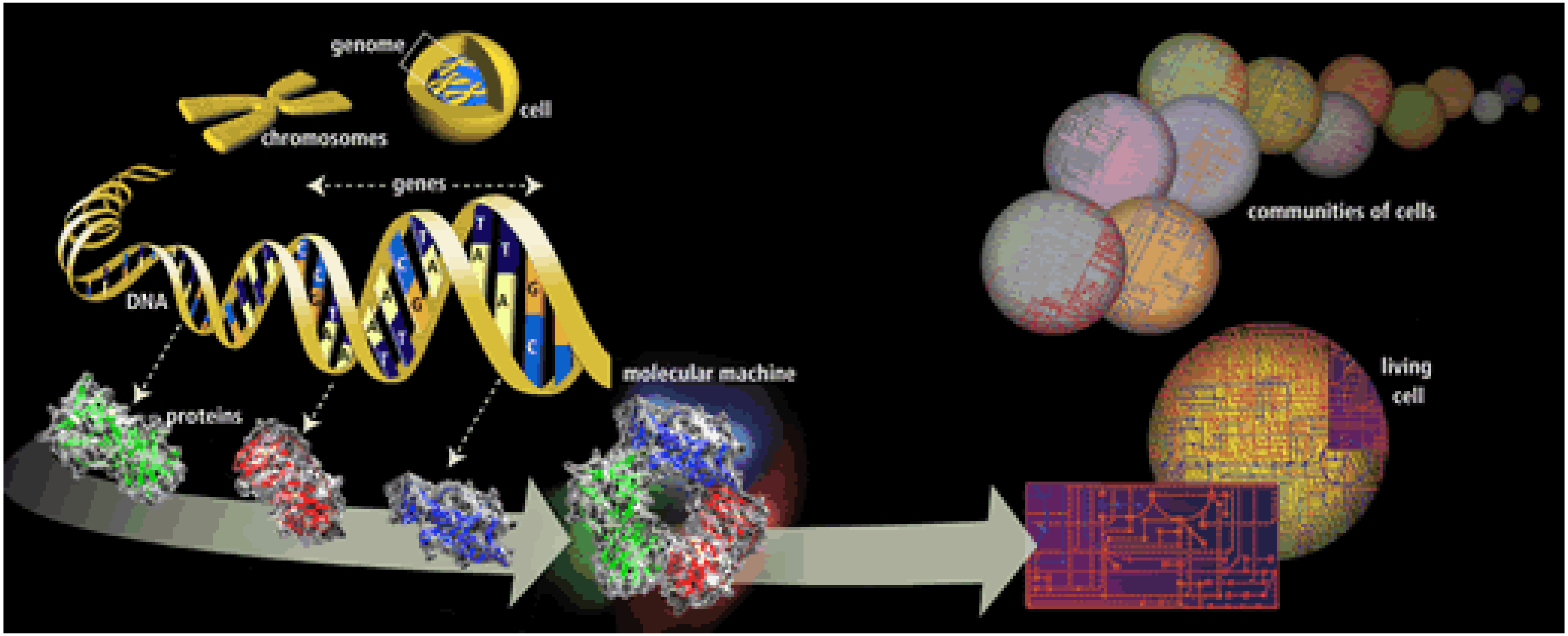
The Post-Genome Era



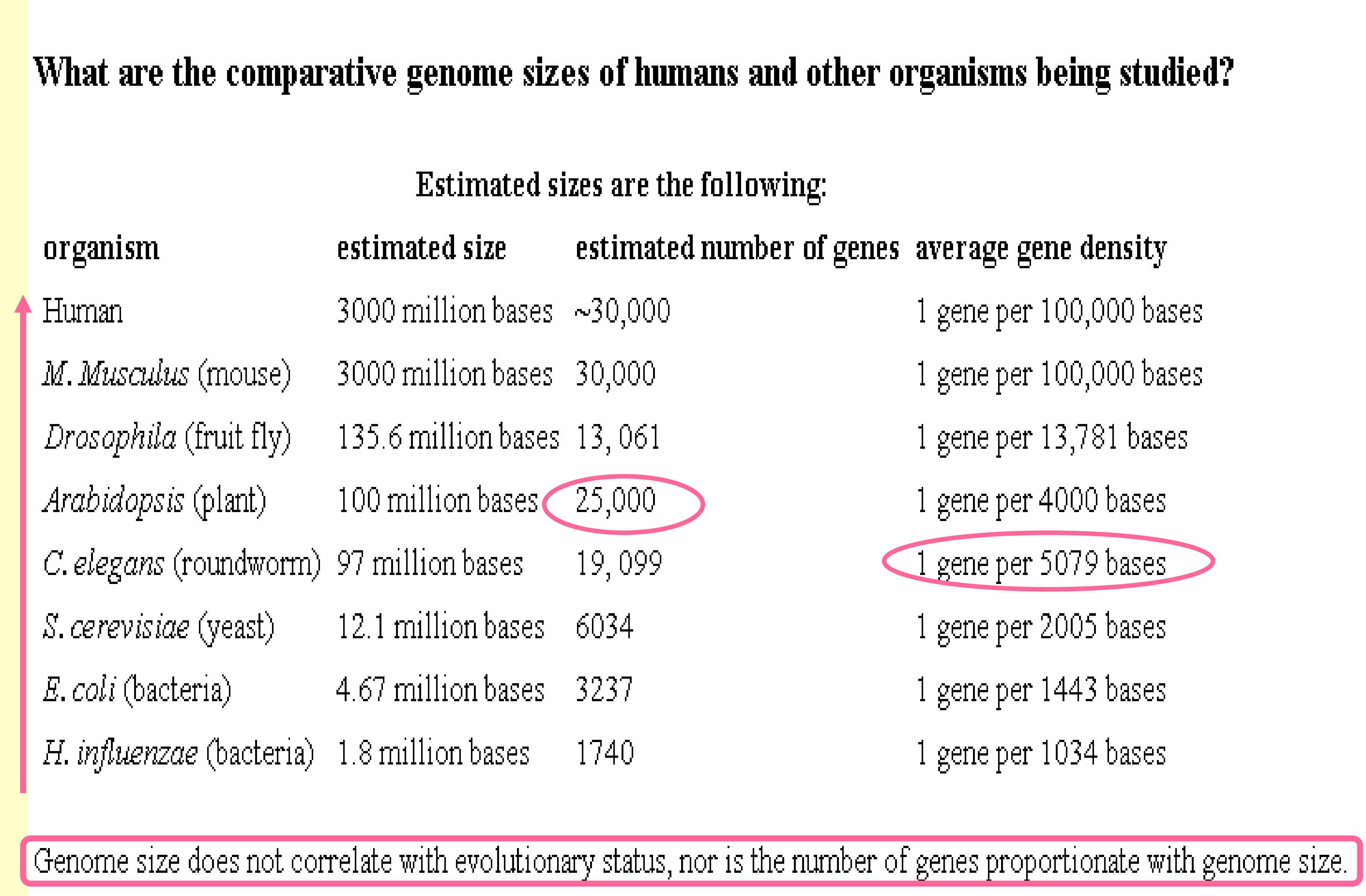
- Bioinformatics provides the tools
  - To **extract** and **combine** knowledge
  - From **isolated data** and results in biology into **meaningful working models** of **cells** and **organisms**
    - Their birth, life and death

Source: Shankar Subramaniam, UCSD

From DNA to Life



http://www.doegenomes.org/



What are the comparative genome sizes of humans and other organisms being studied?

Estimated sizes are the following:

organism	estimated size	estimated number of genes	average gene density
Human	3000 million bases	~30,000	1 gene per 100,000 bases
M. Musculus (mouse)	3000 million bases	30,000	1 gene per 100,000 bases
Drosophila (fruit fly)	135.6 million bases	13, 061	1 gene per 13,781 bases
Arabidopsis (plant)	100 million bases	25,000	1 gene per 4000 bases
C. elegans (roundworm)	97 million bases	19, 099	1 gene per 5079 bases
S. cerevisiae (yeast)	12.1 million bases	6034	1 gene per 2005 bases
E. coli (bacteria)	4.67 million bases	3237	1 gene per 1443 bases
H. influenzae (bacteria)	1.8 million bases	1740	1 gene per 1034 bases

Genome size does not correlate with evolutionary status, nor is the number of genes proportionate with genome size.

C-value paradox

http://www.ornl.gov/hgmis/faq/compgen.html

Comparative Genomics (1)

- Life histories for all living things
- The **Human** – diseases control
  - Non-human vertebrate model organisms
    - Models of **human genetic diseases**

Comparative Genomics (2)

- **Animals & Plants**
  - Comparative gene mapping **& breeding**
    - **Map-rich genomes ⇨ map-poor genomes**
    - Marker-aid-selection (MAS) for economics trait loci (ETLs)
    - Limited choice of suitable transgenes
      - Regulatory elements
      - Transgenes

Comparative Genomics (3)

- Microbes
  - Host & pests/parasites relationships, prevention & treatments
    - Malaria genomics
    - Tuberculosis (TB)

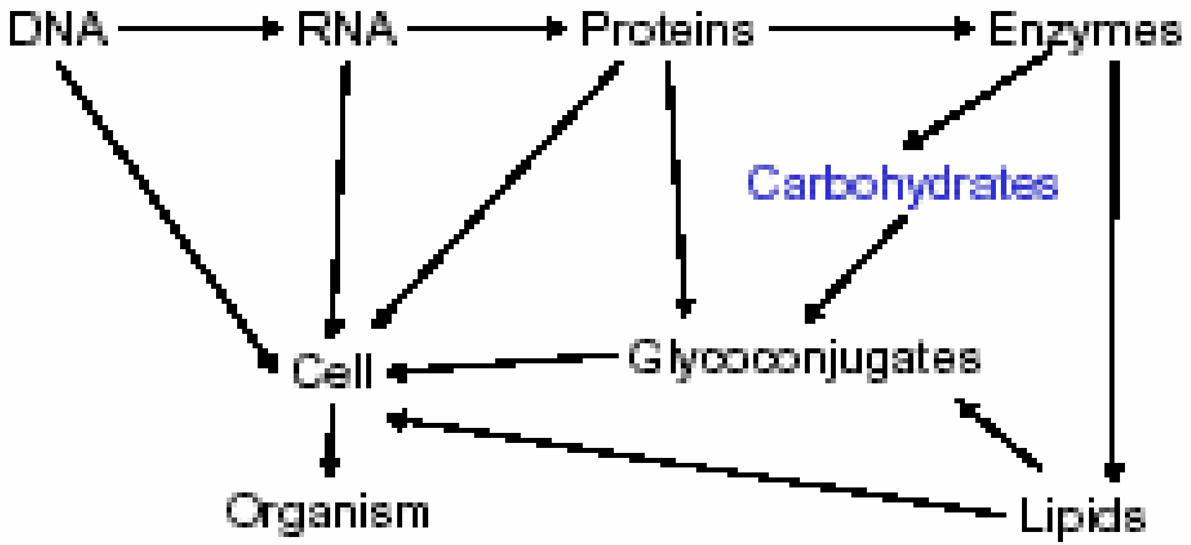
Central paradigm of molecular biology

Flow of information from DNA to RNA to proteins to cells

DNA → RNA → Protein → Cell → Organism

Extended implications

Lipids, carbohydrates and glycoconjugates are necessary to make a cell



DNA → RNA → Proteins → Enzymes

Carbohydrates

Cell

Glycoconjugates

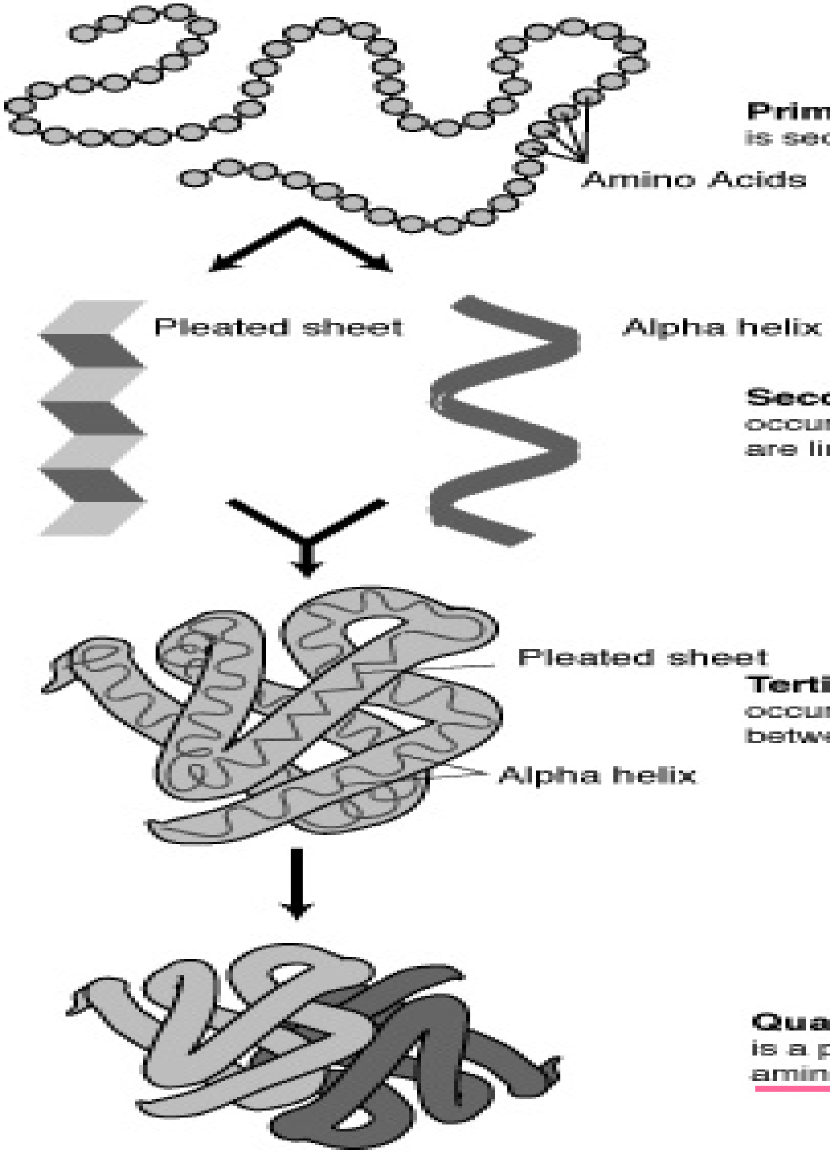
Organism

Lipids

From *Essentials of Glycobiology*, 1999, Varki et al, Cold Spring Harbor Press

Central Paradigm of Bioinformatics

- Central dogma of molecular biology
  - [DNA → RNA → protein]→ phenotype
- **Central paradigm of Bioinformatics (molecular levels)**
  - **Sequence → structure → function**
    - Most cellular functions are performed or facilitated by **proteins**
      - Primary biocatalyst, co-factor transport/storage, mechanical motion/support, immune protection, control of growth/differentiation
- **Genomic sequence information**
  - mRNA → protein **sequence** → protein **structure** → protein **function** → **phenotype**
  - [Comparative genomics] To understand evolutionary relationships in terms of the expression of protein function



Primary protein structure
is sequence of a chain of amino acids

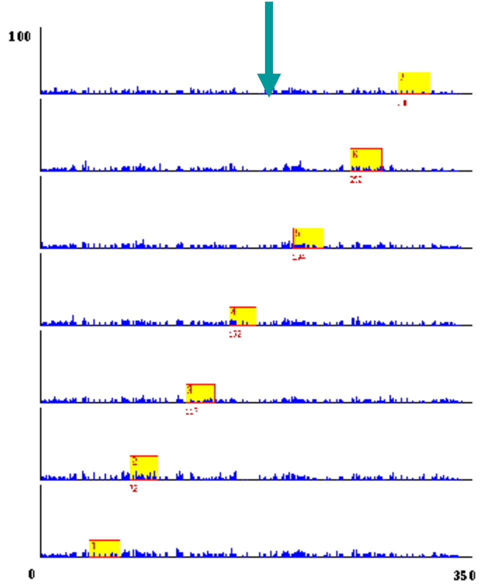
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds

Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.

Quaternary protein structure
is a protein consisting of more than one amino acid chain.

The Reality of Sequence Analysis

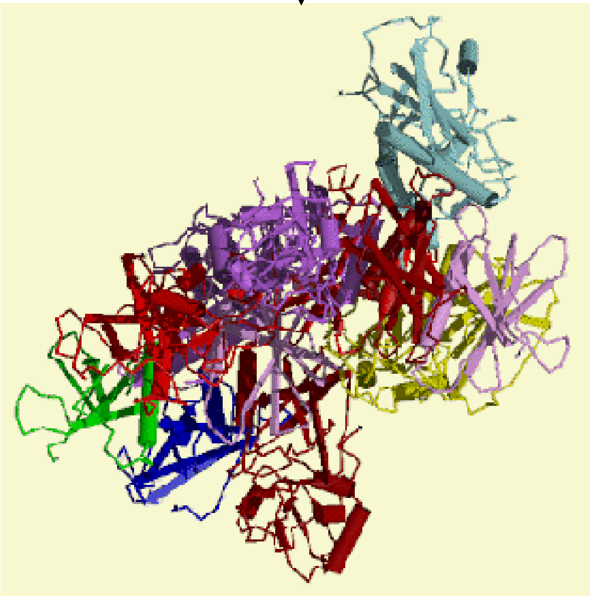
```
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN
```

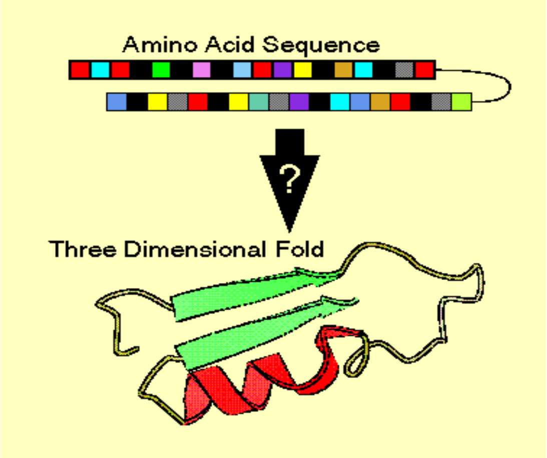


...isn't so glamorous....but means we can recognize **words** that form **characteristic patterns**, even if we don't know the precise syntax to build complete protein sentences

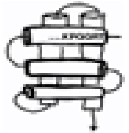
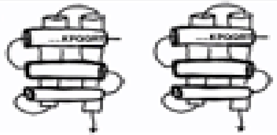
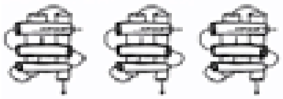
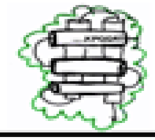
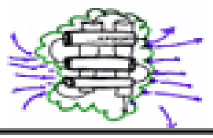
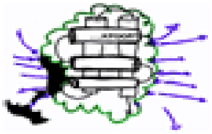
The Holy Grail of Bioinformatics

```
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN
```





...to be able to understand the words in a sequence sentence that form a particular **protein structure**

		Breadth: Homologs, Large-scale Surveys, Informatics→			
			pairwise comparison, sequence & structure alignment	multiple alignment, patterns, templates, trees	databases, scoring schemes, censuses
		1	2	3-100	100+
	Genome Sequence	atcgatcgatatttgggatttggga	atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga	atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga	atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga atcgatcgatatttgggatttggga
gene finding	↓				
	Protein Sequence	ALMNAKKKPQQRT	ALMNAKKKPQQRT ALMNAKKKPQQRT	ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT	ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT ALMNAKKKPQQRT
structure prediction	↓				
	Protein Structure				
geometry calculation	↓				
	Protein Surface				
molecular simulation	↓				
	Force Field				
structure docking	↓				
	Ligand Complex				

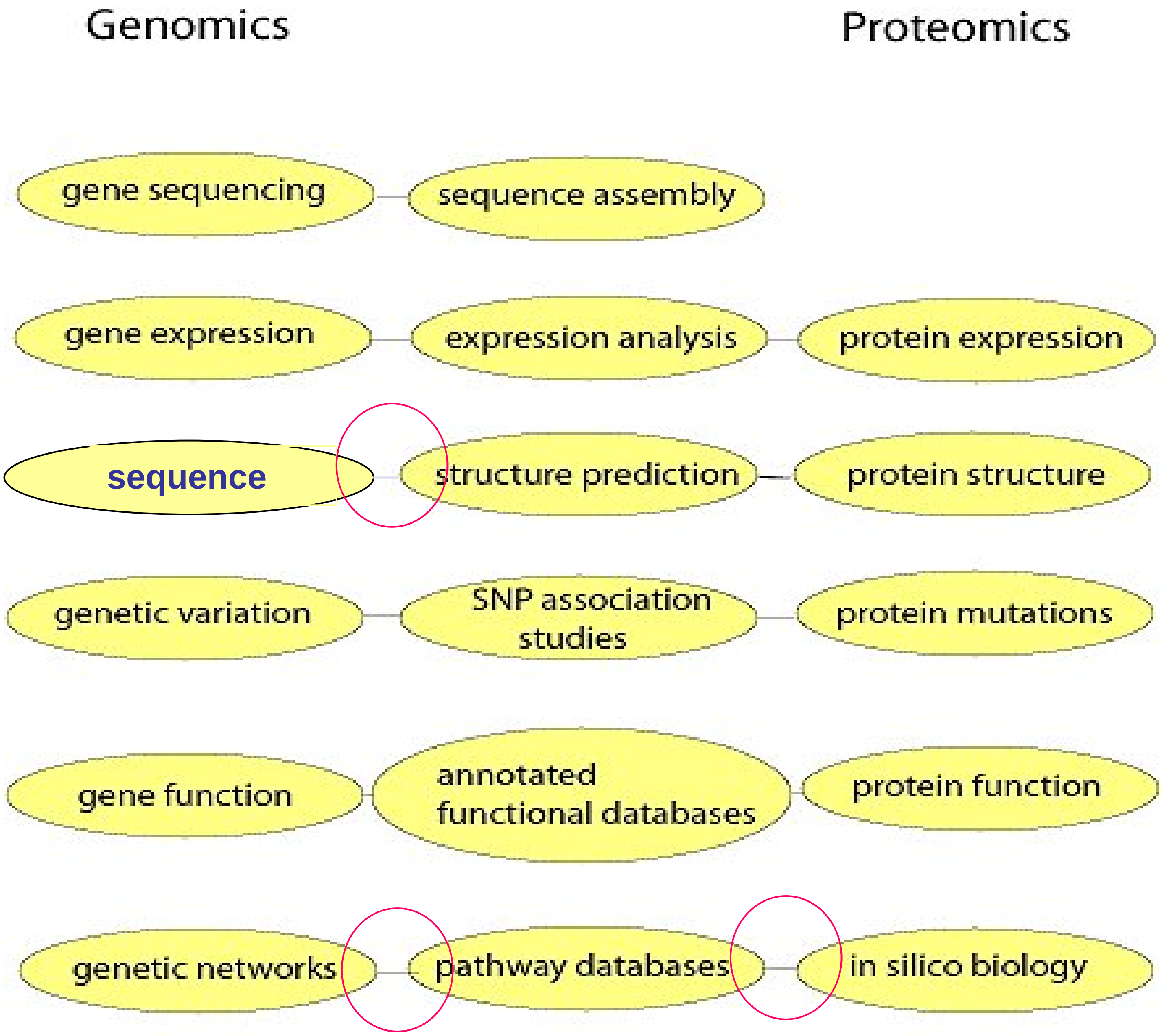
Depth: Rational Drug Design (physics)→

http://bioinfo.mbb.yale.edu/what-is-it

The Most Useful Tools so Far

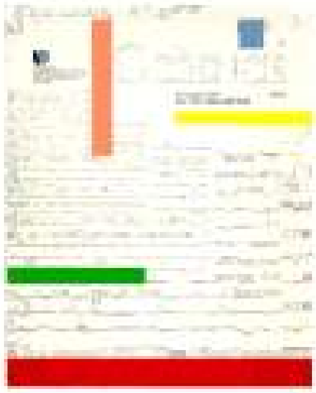
- **Sequence comparison**
  - To compare an **un-characterize DNA sequence** to the entire publicly held **collection of DNA** sequences
    - BLAST
    - FASTA

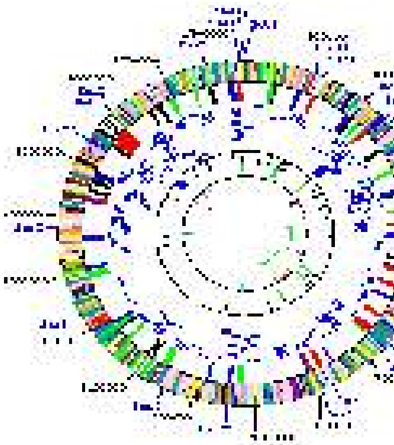
Bioinformatics



Genomics		Proteomics
gene sequencing	sequence assembly	
gene expression	expression analysis	protein expression
sequence	structure prediction	protein structure
genetic variation	SNP association studies	protein mutations
gene function	annotated functional databases	protein function
genetic networks	pathway databases	in silico biology

Molecular Biology Information: Whole Genomes





- **The Revolution Driving Everything**

  Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae rd." *Science* 269: 496-512.

  (Picture adapted from TIGR website, http://www.tigr.org)

- **Integrative Data**

  1995, HI (bacteria): 1.6 Mb & 1600 genes done

  1997, yeast: 13 Mb & ~6000 genes for yeast

  1998, worm: ~100Mb with 19 K genes

  1999: >30 completed genomes!

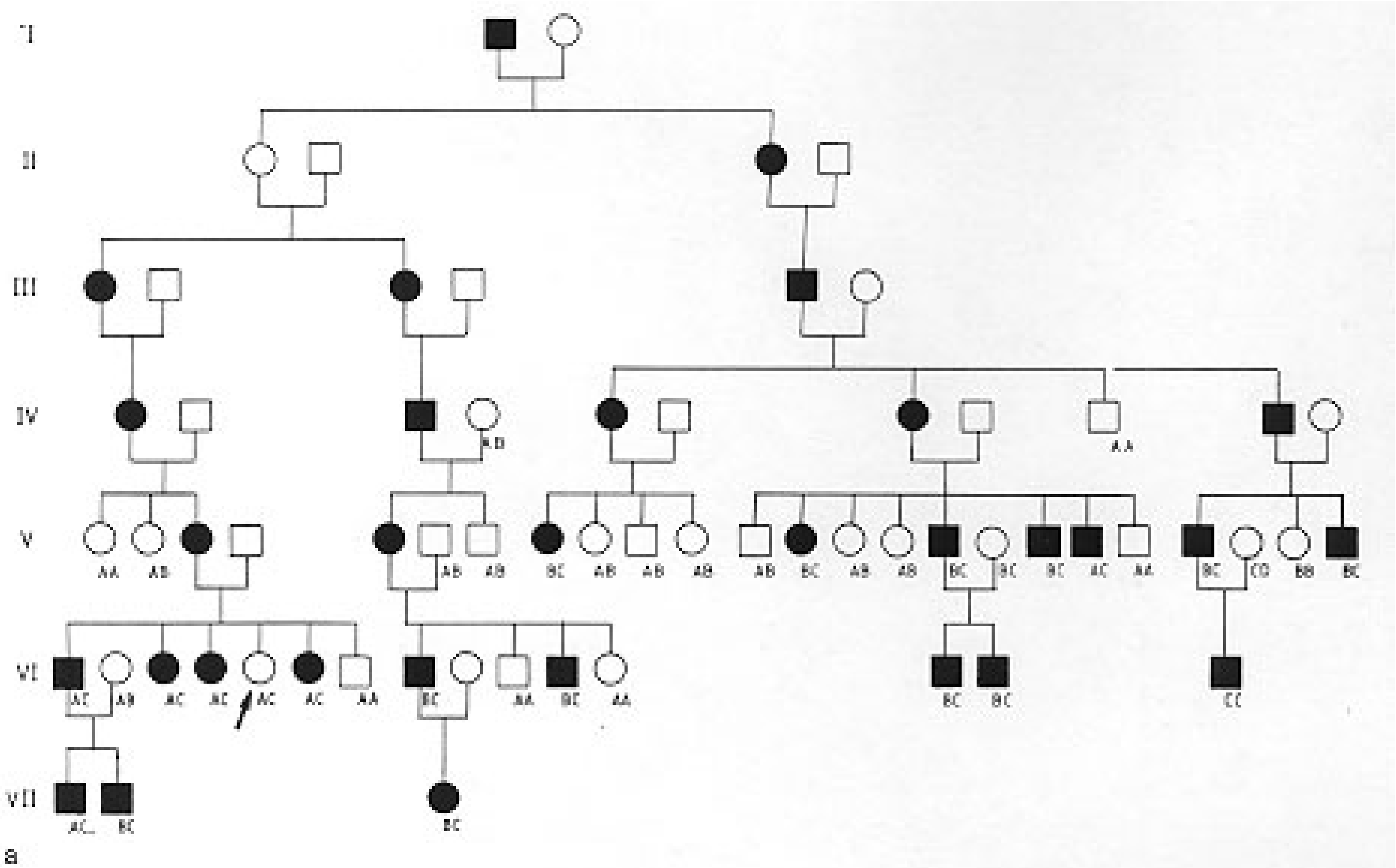
  2003, human: 3 Gb & 100 K genes...

> Genome sequence now accumulate so quickly that, in less than a week, a single laboratory can produce more bits of data than Shakespeare managed in a lifetime, although the latter make better reading.
>
> – G A Pekso, *Nature* **401**: 115-116 (1999)

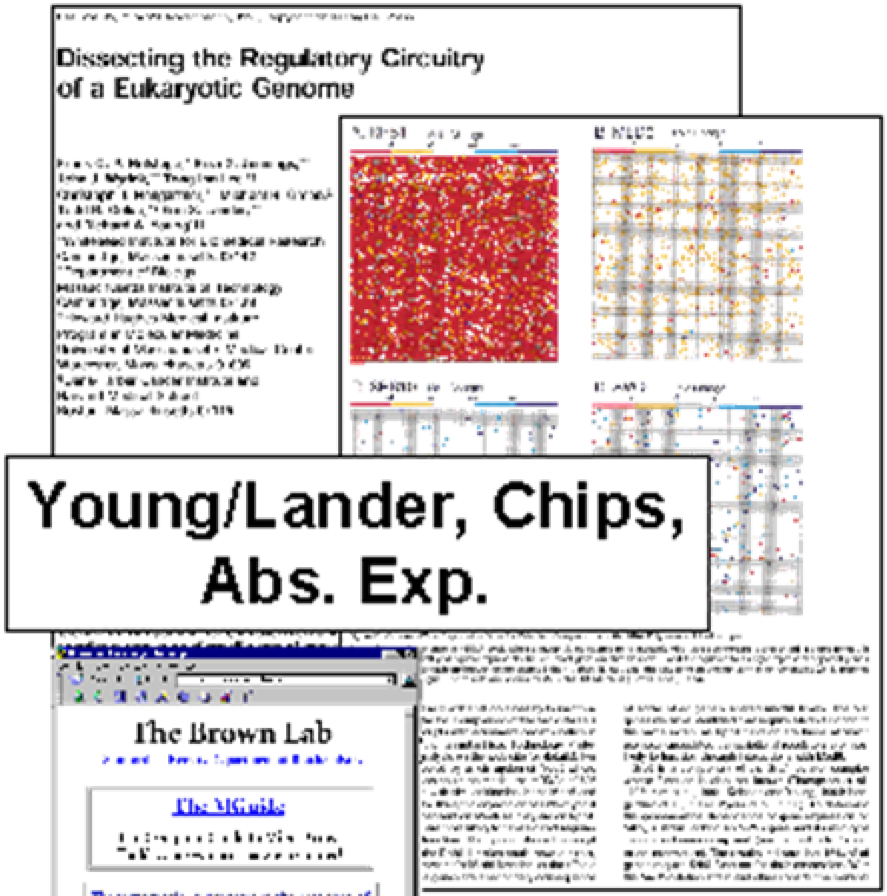
Information from Gene Mapping & Sequencing (1)

- **Linkage information** ⇨ DNA/chromosome walking/landing/jumping
  - Huntington's disease (Bender *et al.* 1983)
- **Genome organization**
  - Sequences, promoter, exons & introns etc.
- **Protein complement**
  - Genomic DNAs, ESTs & full-length cDNA ⇨ increasing complete lists of encodes effector molecules
    - Algorithms: Local vs. global, BLAST, FASTA etc.

Pedigree of Huntington Disease



Gene Expression Datasets: the Transcriptosome



Dissecting the Regulatory Circuitry of a Eukaryotic Genome

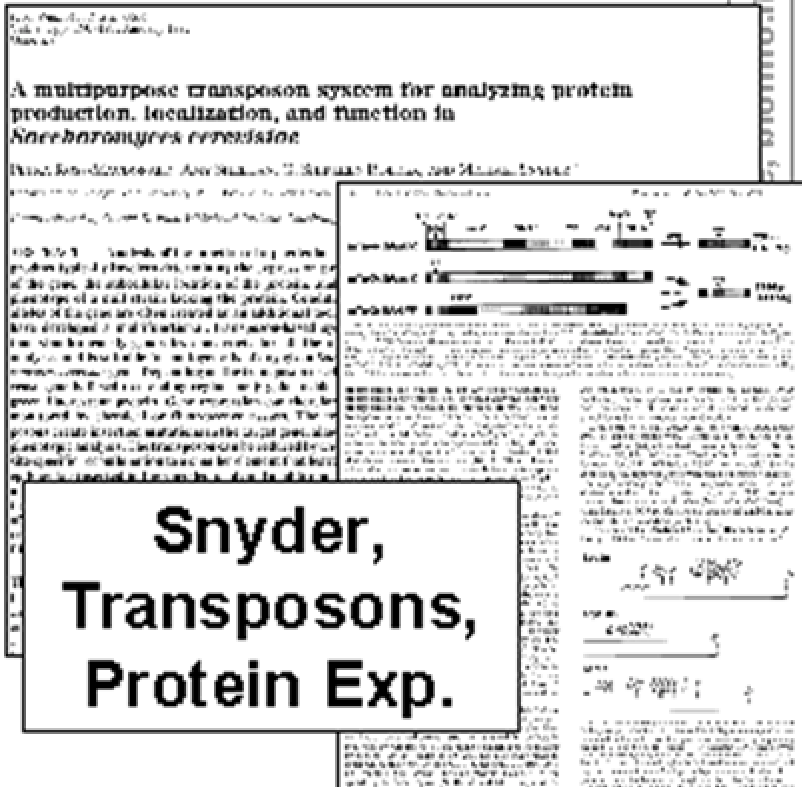
Young/Lander, Chips, Abs. Exp.

The Brown Lab

The MGuide

Brown, µarray, Rel. Exp. over Timecourse

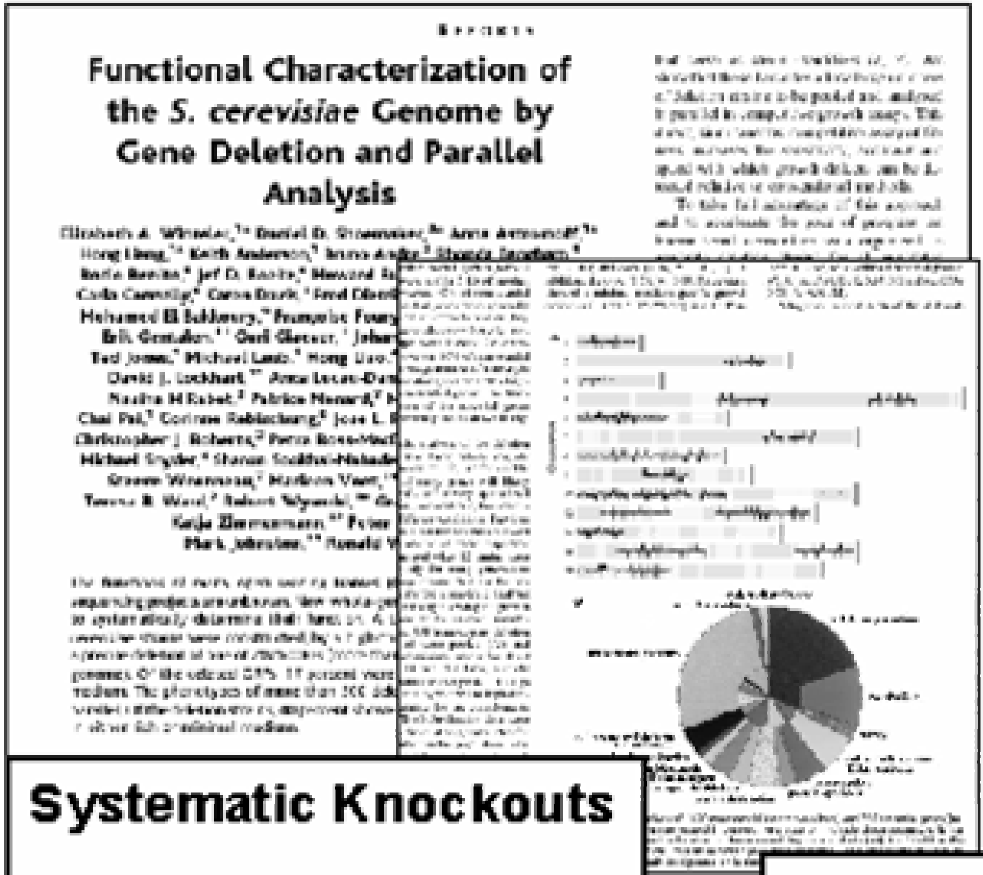
Also: SAGE; Samson and Church, Chips; Aebersold, Protein Expression



A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*

Snyder, Transposons, Protein Exp.

Other Whole-Genome Experiments

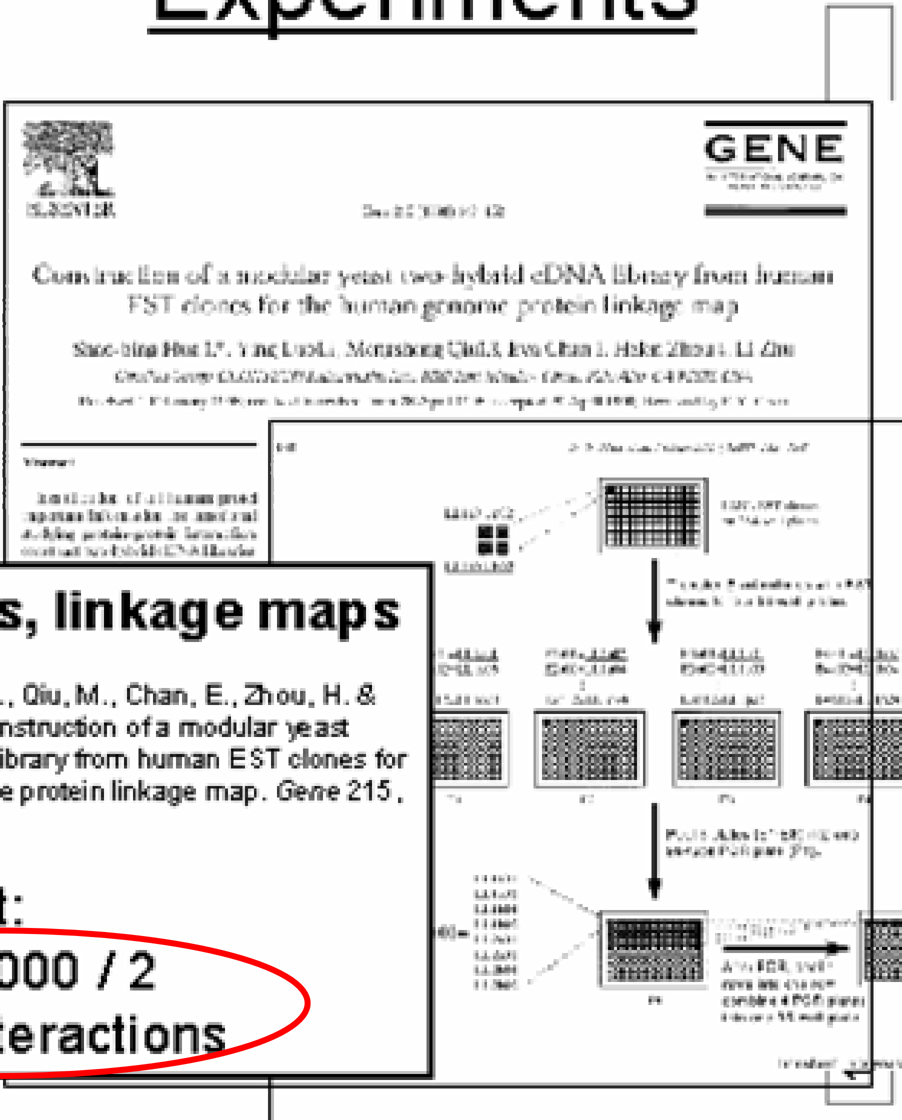


REPORTS

Functional Characterization of the *S. cerevisiae* Genome by Gene Deletion and Parallel Analysis

Systematic Knockouts

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., Davis, R. W. & et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6



GENE

ELSEVIER

Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map

2 hybrids, linkage maps

Hua, S. B., Luo, Y., Qiu, M., Chan, E., Zhou, H. & Zhu, L. (1998). Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map. *Gene* 215, 143-52

For yeast:

6000 x 6000 / 2

~ 18M interactions

Information from Gene Mapping & Sequencing (2)

- Protein complement (cont.)
  - The function of between **15- and 40%** of the proteins encoded by any genome is not apparent from their **sequences**
    - Absence sequence similarity to known protein
      - The biochemical function & the higher order function (*e.g.*, transcriptional controls)
- **Gene regulation**
  - Large-scale identification of sites of regulatory protein action
    - Comparison of **sequence near coding regions** in somewhat diverged organisms ⇨ **functional sites**
      - *C. elegans* vs. *C. bergerac*
      - *D. melanogaster* vs. *D. virilis*
      - *Mus musculus* vs. *Fugu rubripes*

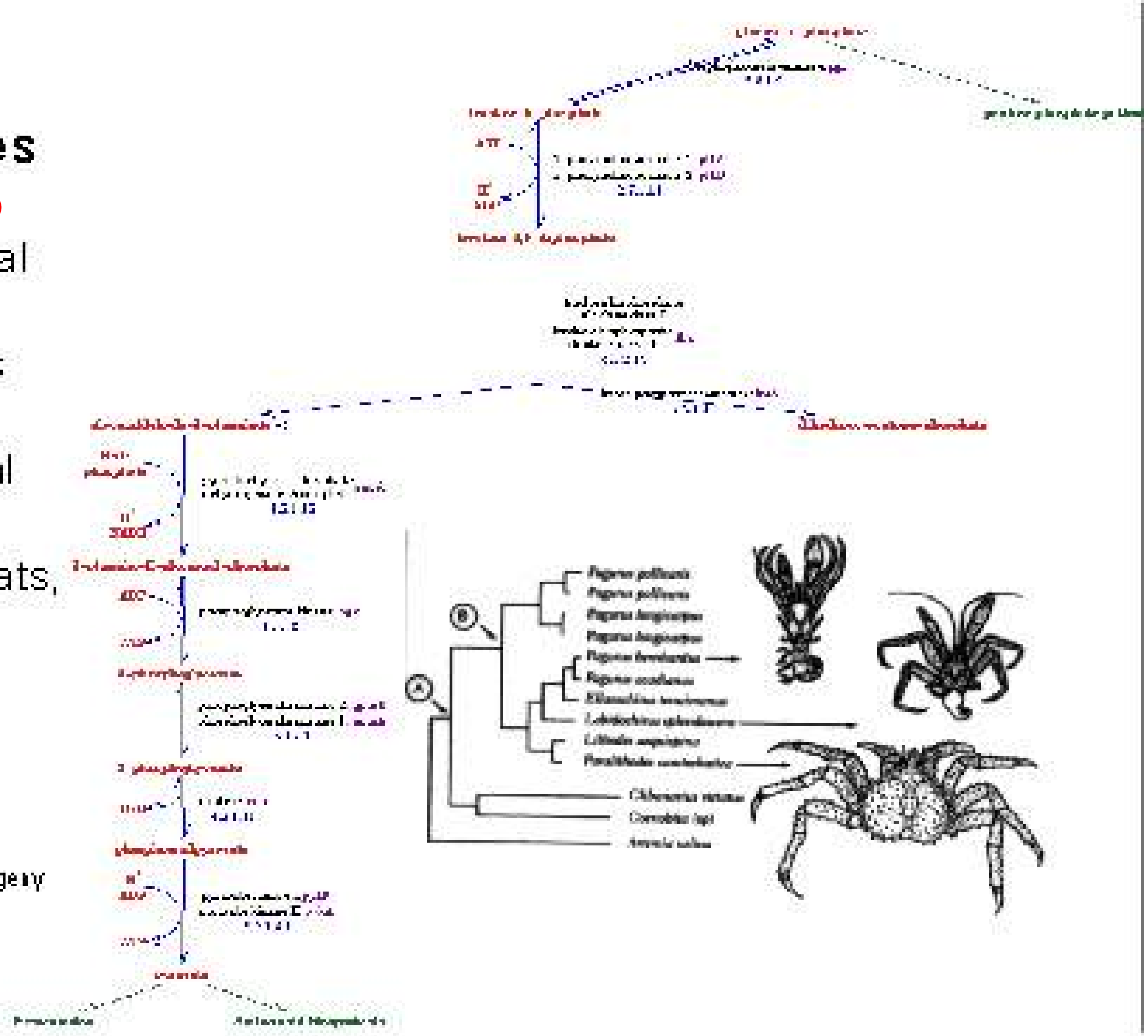
Information from Gene Mapping & Sequencing (3)

- Information about **phylogeny** & **evolution**
  - The changes that have led to speciation & **existing phylogeny**
    - DNA sequencing revealed a number of genomic rearrangements
      - Duplication events (*e.g.*, yeast)
      - **Synteny rearrangements** for many *phyla*: *e.g.*, vertebrate genomes may represent a quadruplication of ancestral metazoan genome that also give rise to worm & flies
      - Molecular evolution vs. morphological or paleontological information
      - DNA sequence demonstrates numerous individual instances of **horizontal gene transfer** among prokaryotic species (Jain *et al.* 1999) – transformation, conjugation, transduction

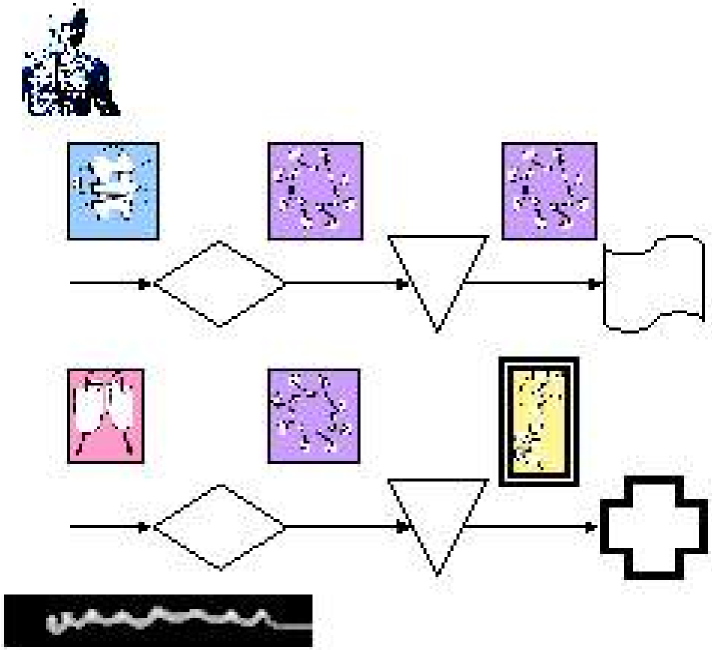
Molecular Biology Information: Other Integrative Data

- **Information to understand genomes**
  - Metabolic Pathways (glycolysis), traditional biochemistry
  - Regulatory Networks
  - Whole Organisms Phylogeny, traditional zoology
  - Environments, Habitats, ecology
  - The Literature (MEDLINE)
- **The Future....**

(Pathway drawing from P Karp's EcoCyc, Phylogeny from S J Gould, Dinosaur in a Haystack)



The Character of Molecular Biology Information: Redundancy and Multiplicity



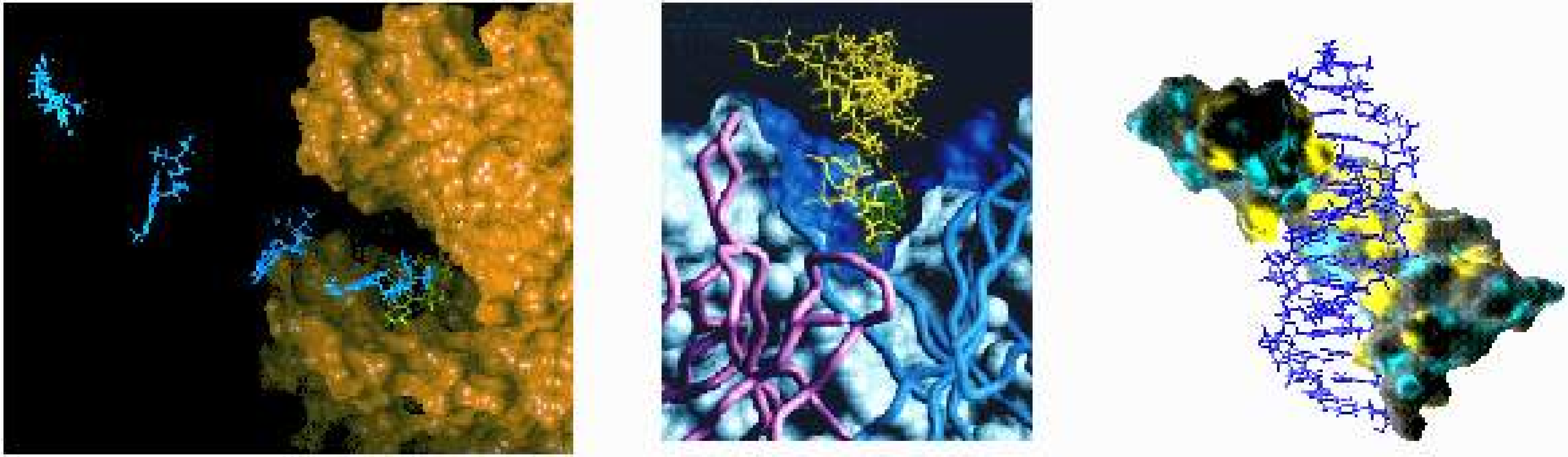
- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions **Pleiotrophic**
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code
- **How do we find the similarities?**

Integrative Genomics -
genes ↔ structures ↔
functions ↔ pathways ↔
expression levels ↔
regulatory systems ↔ ....

Major Application I: Designing Drugs

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

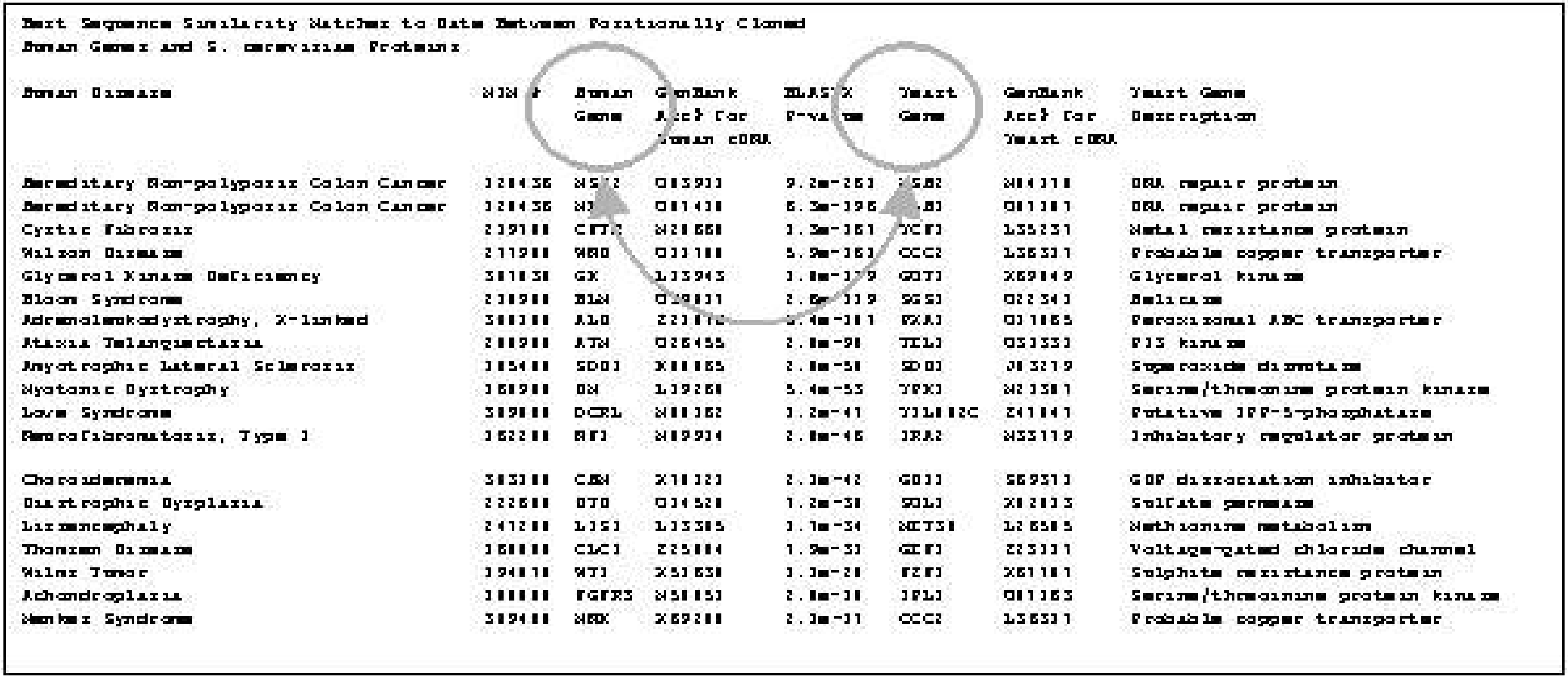
(From left to right, figures adapted from Olson Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).



Major Application II: Finding Homologues

- Find Similar Ones in Different Organisms
- Human vs. Mouse vs. Yeast
  - ◊ Easier to do Expts. on latter!

(Section from NCBI Disease Genes Database Reproduced Below)



Best Sequence Similarity Matches to Date Between Positionally Cloned Human Genes and S. cerevisiae Proteins

Human Disease	MIM #	Human Gene	GenBank Acc# for Human cDNA	BLASTX P-value	Yeast Gene	GenBank Acc# for Yeast cDNA	Yeast Gene Description
Hereditary Non-polyposis Colon Cancer	120436	MSH2	U03911	9.2e-261	MSH2	M84170	DNA repair protein
Hereditary Non-polyposis Colon Cancer	120436	MLH1	U07418	6.3e-196	MLH1	U07187	DNA repair protein
Cystic Fibrosis	219700	CFTR	M28668	1.3e-167	YCF1	L35237	Metal resistance protein
Wilson Disease	277900	WND	U11700	5.9e-161	CCC2	L36317	Probable copper transporter
Glycerol Kinase Deficiency	307030	GK	L13943	1.8e-129	GUT1	X69049	Glycerol kinase
Bloom Syndrome	210900	BLM	U39817	2.6e-119	SGS1	U22341	Helicase
Adrenoleukodystrophy, X-linked	300100	ALD	Z21876	3.4e-107	PXA1	U17065	Peroxisomal ABC transporter
Ataxia Telangiectasia	208900	ATM	U26455	2.8e-90	TEL1	U31331	PI3 kinase
Amyotrophic Lateral Sclerosis	105400	SOD1	K00065	2.0e-58	SOD1	J03279	Superoxide dismutase
Myotonic Dystrophy	160900	DM	L19268	5.4e-53	YPK1	M21307	Serine/threonine protein kinase
Lowe Syndrome	309000	OCRL	M88162	1.2e-47	YIL002C	Z47047	Putative IPP-5-phosphatase
Neurofibromatosis, Type 1	162200	NF1	M89914	2.0e-46	IRA2	M33779	Inhibitory regulator protein
Choroideremia	303100	CHM	X78121	2.1e-42	GDI1	S69371	GDP dissociation inhibitor
Diastrophic Dysplasia	222600	DTD	U14528	7.2e-38	SUL1	X82013	Sulfate permease
Lissencephaly	247200	LIS1	L13385	1.7e-34	MET30	L26505	Methionine metabolism
Thomsen Disease	160800	CLC1	Z25884	7.9e-31	GEF1	Z23117	Voltage-gated chloride channel
Wilms Tumor	194070	WT1	X51630	1.1e-20	FZF1	X67787	Sulphite resistance protein
Achondroplasia	100800	FGFR3	M58051	2.0e-18	IPL1	U07163	Serine/threonine protein kinase
Menkes Syndrome	309400	MNK	X69208	2.1e-17	CCC2	L36317	Probable copper transporter

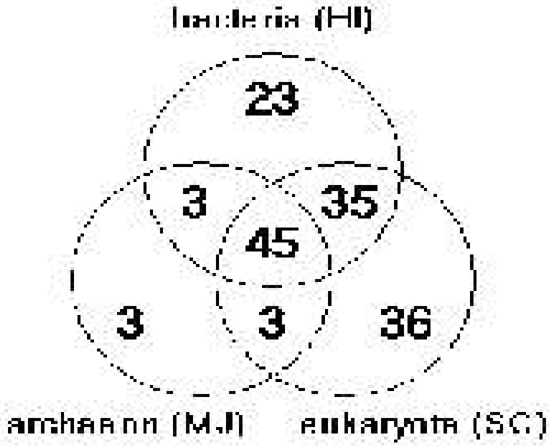
Major Application I|I: Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics

(Clock figures, yeast v. Synechocystis, adapted from GeneQuiz Web Page, Sander Group, EBI)

Homology levels
Computed by GeneQuiz

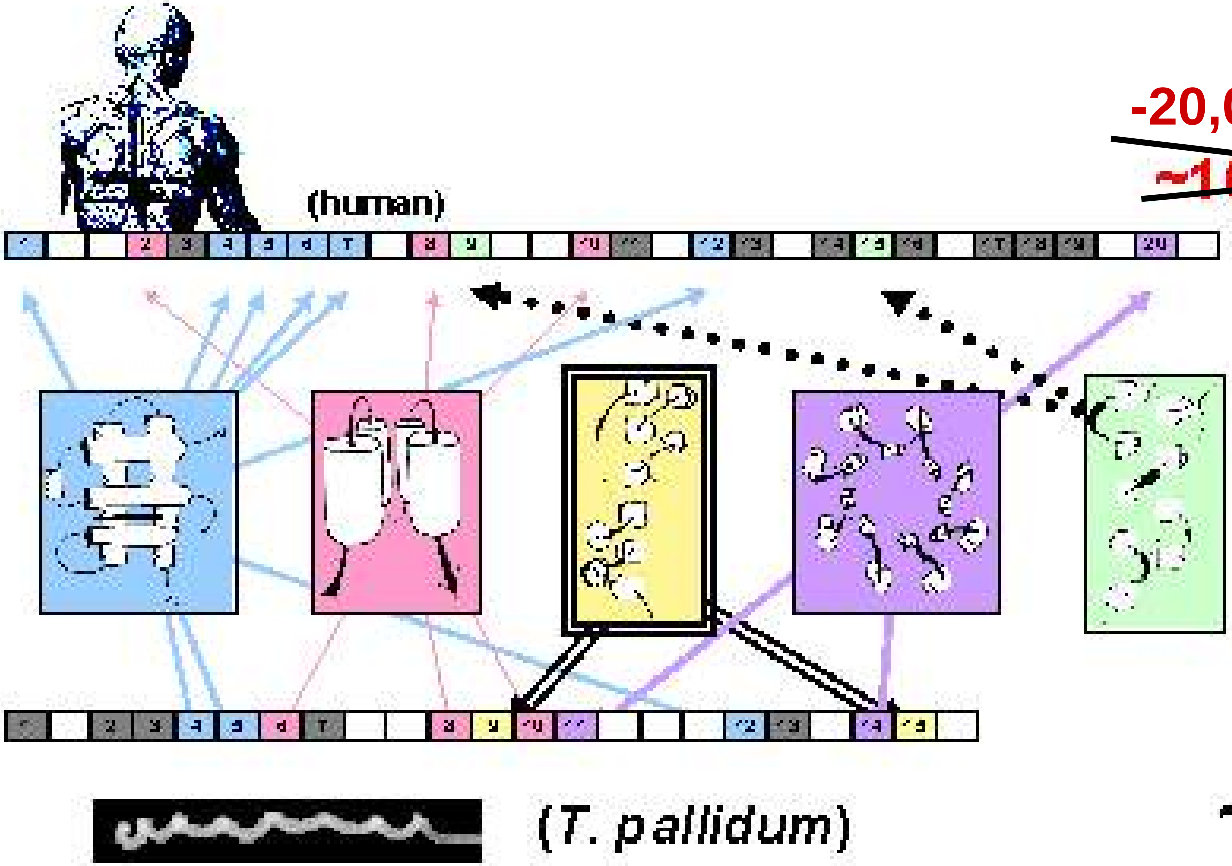
Homology levels
Computed by GeneQuiz



Yeast Protein Functions

Regulatory	45	1.05%
Cell structure	182	4.24
Transposons,etc	87	2.03
Transport & binding	281	6.55
Putative transport	146	3.40
Replication, repair	115	2.68
Transcription	55	1.28
Translation	182	4.24
Enzymes	251	5.85
Unknown	1632	38.06

Simplfying Genomes with Folds, Pathways, &c

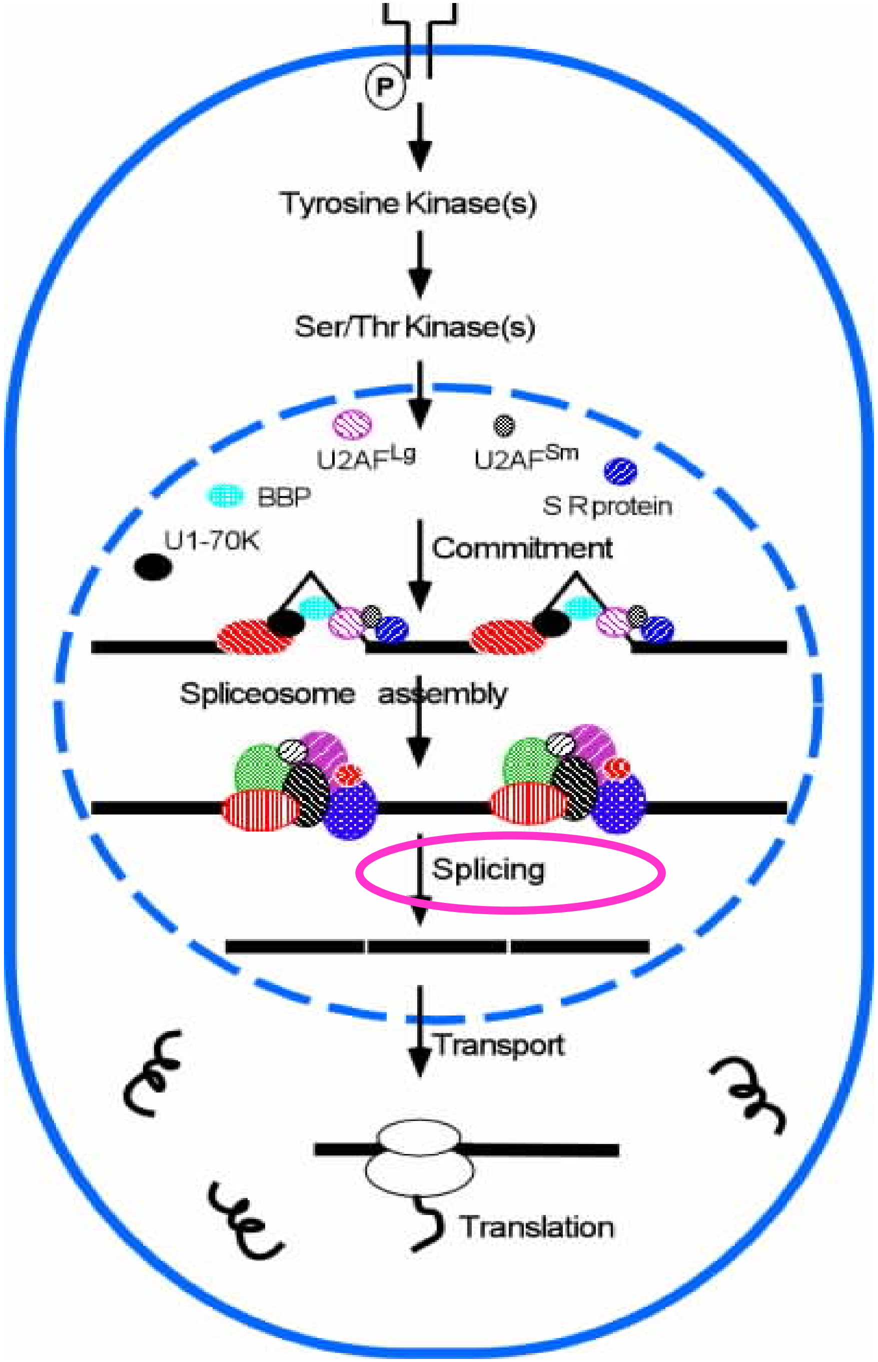


(human)

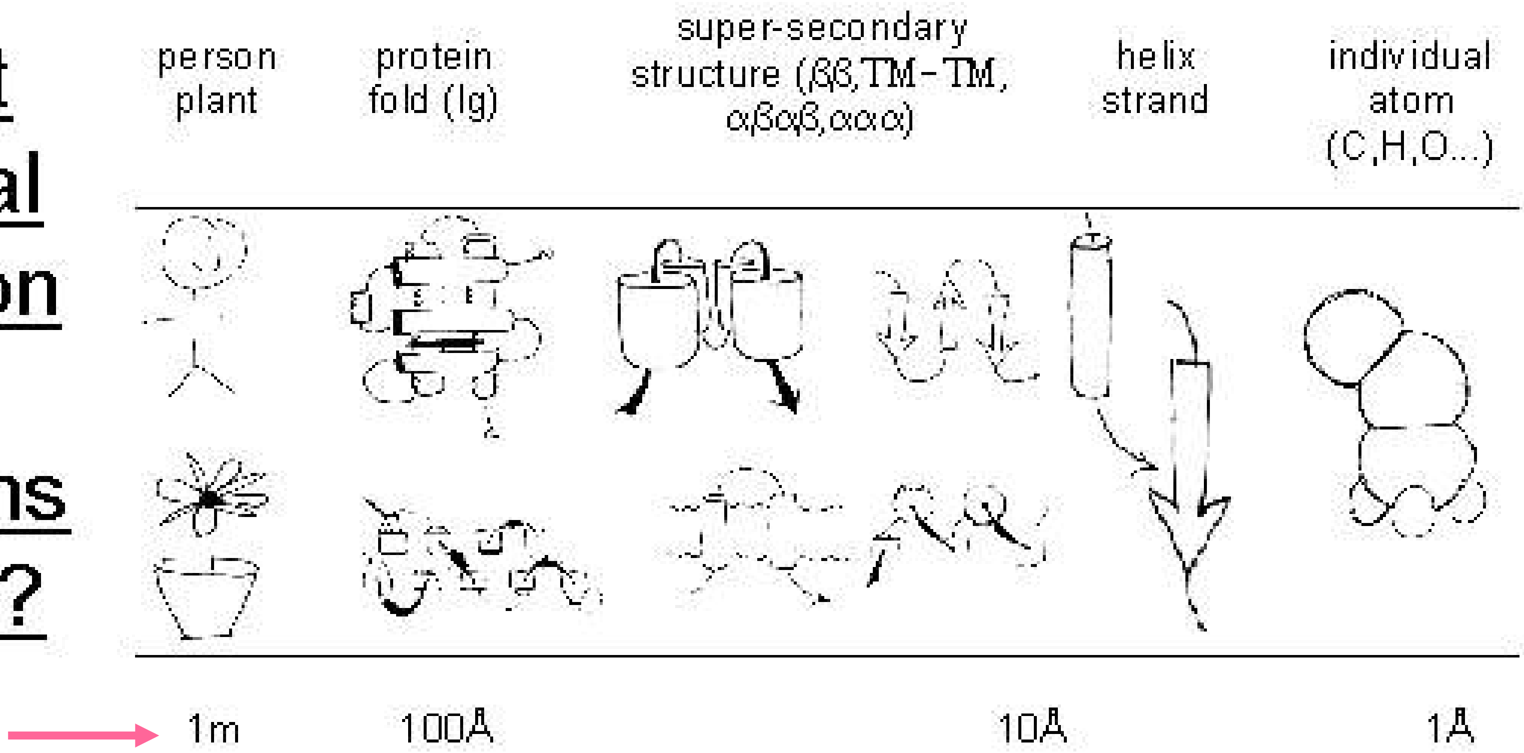
-20,000~25,000

~~~100000 genes~~

~1000 folds

(*T. pallidum*)

~1000 genes
~~~

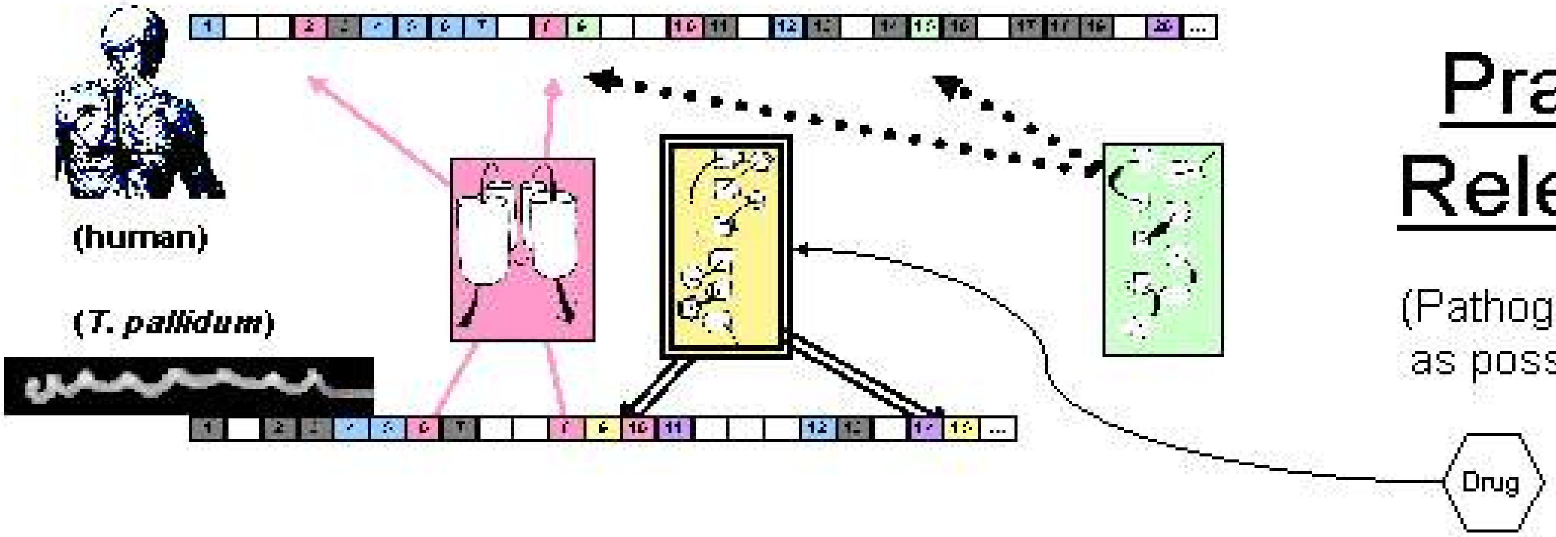
- The **mechanism of splicing** is not well understood



At What Structural Resolution Are Organisms Different?



Practical Relevance

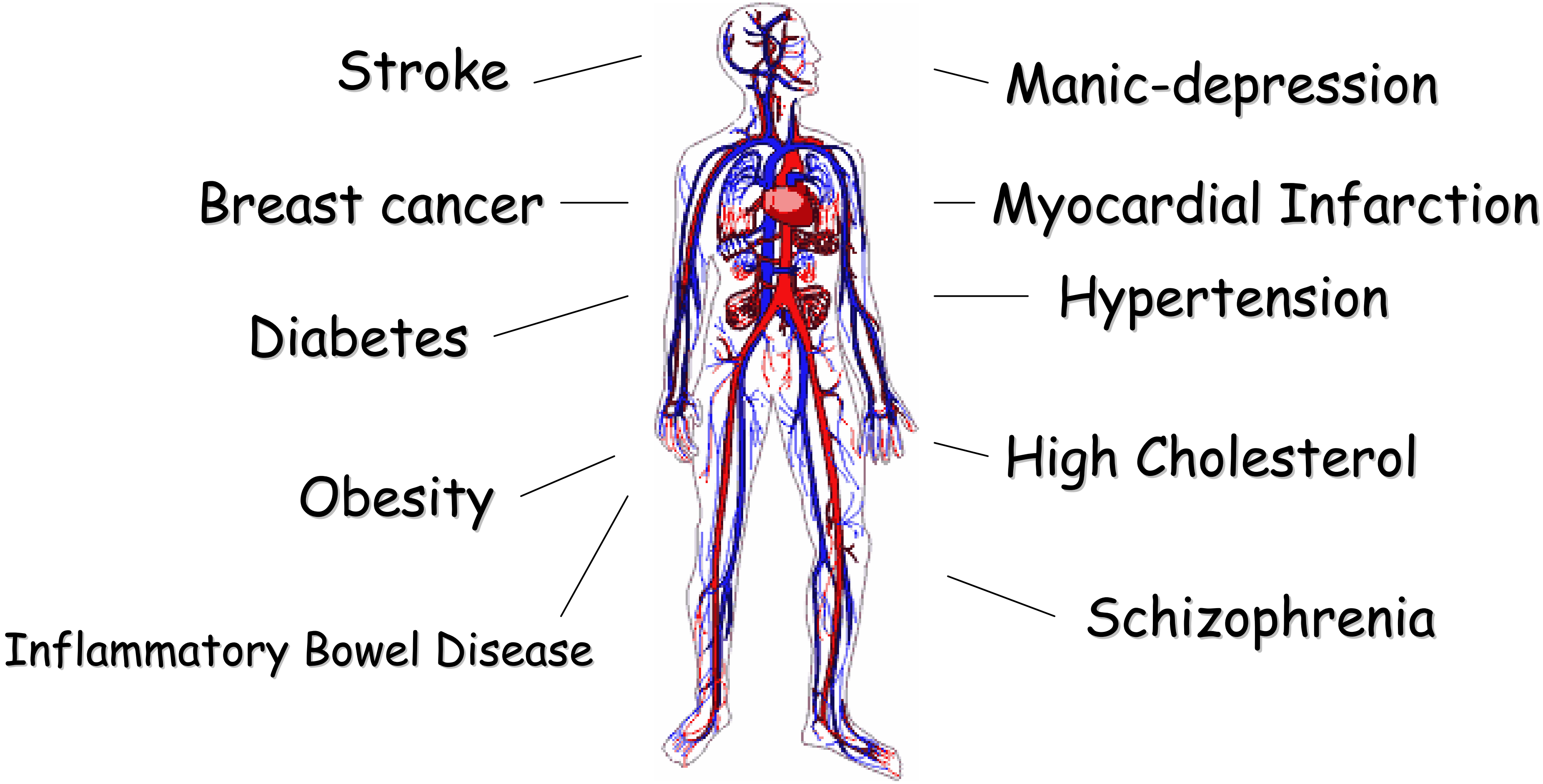


(Pathogen only folds as possible targets)

Data

- **Data** is crucial to the success of analysis
  - "Garbage in → garbage out"
  - Understand your data set and its surrounding **metadata**

Most Common Diseases are Caused by a Combination of Genes and Environment



Stroke

Breast cancer

Diabetes

Obesity

Inflammatory Bowel Disease

Manic-depression

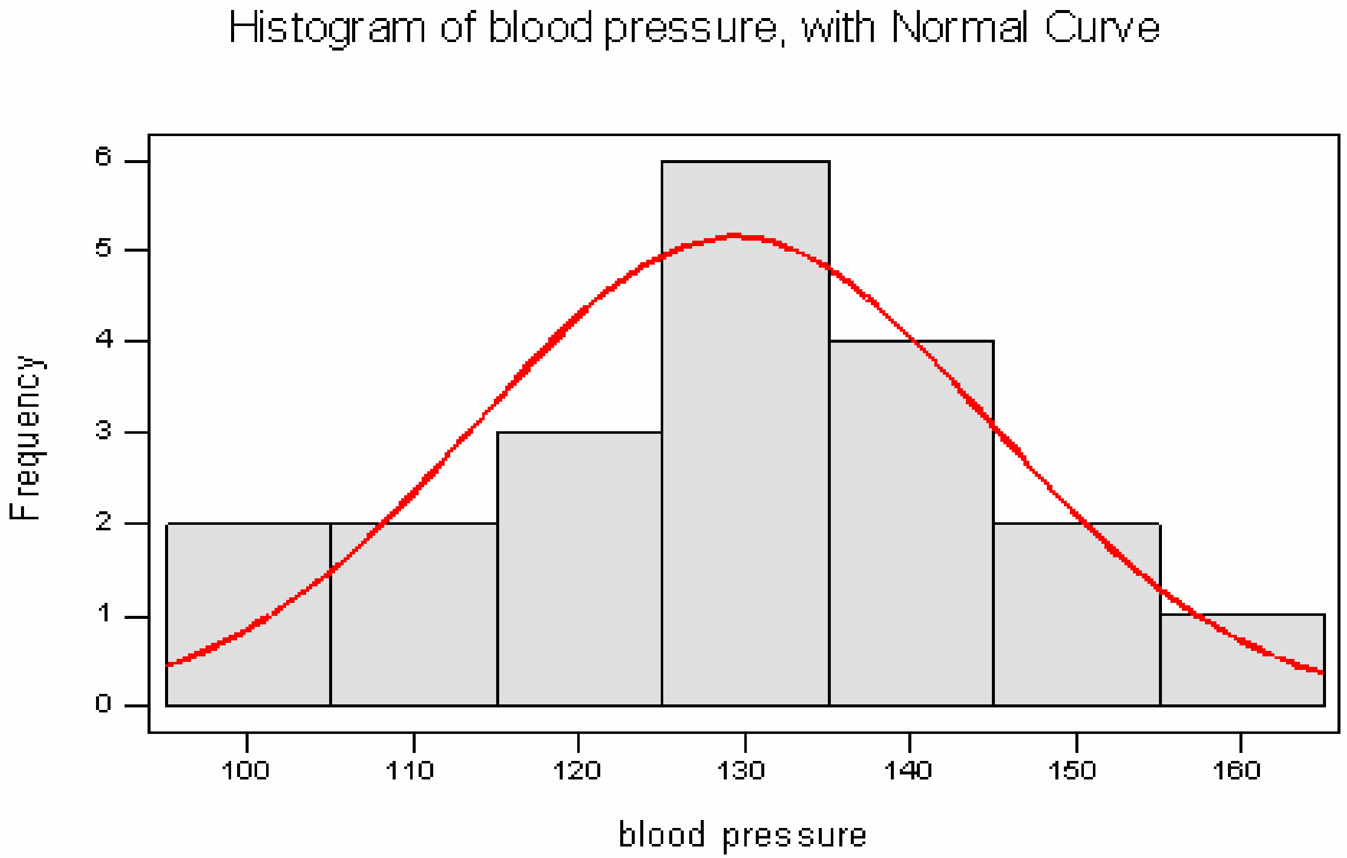
Myocardial Infarction

Hypertension

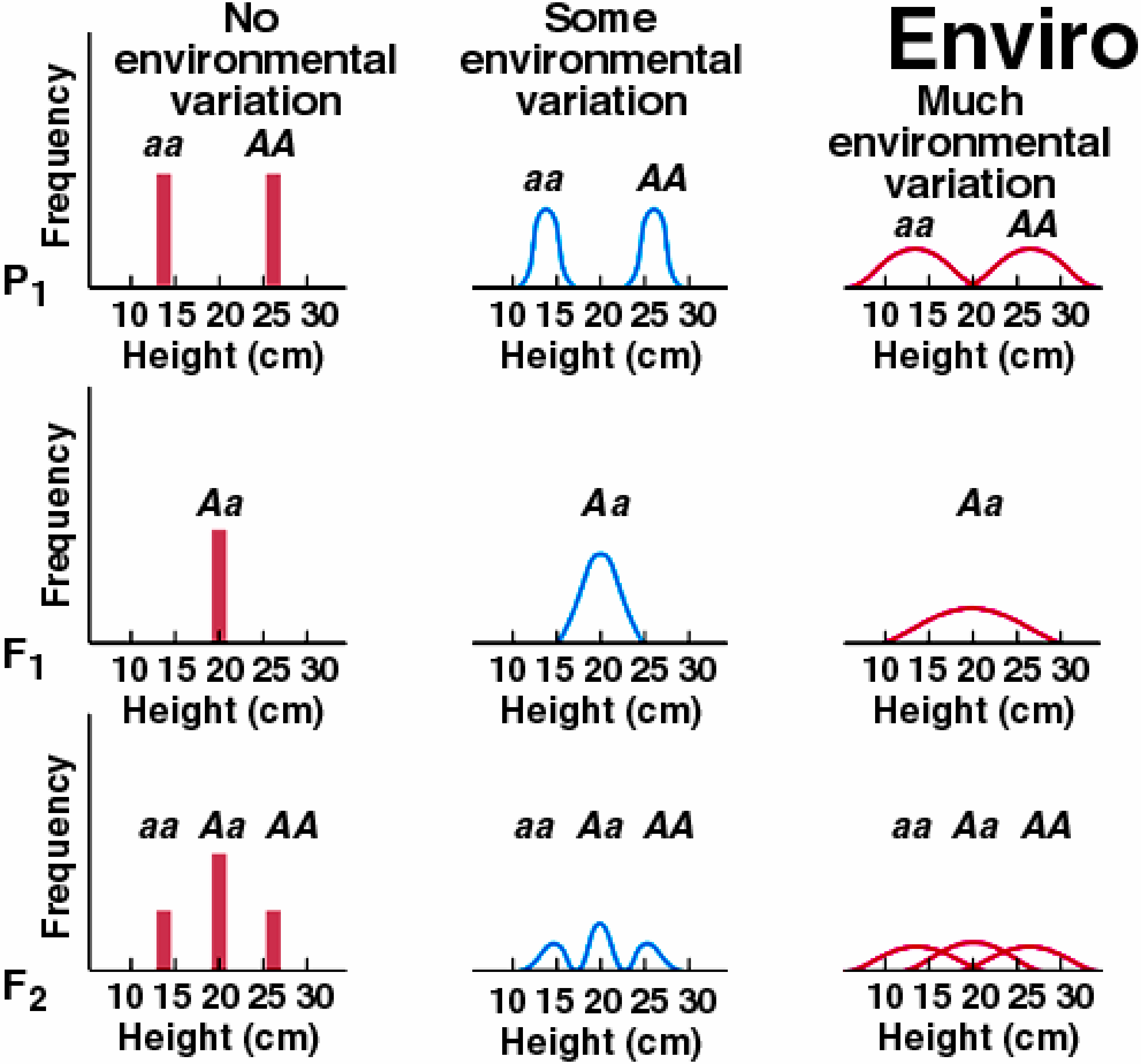
High Cholesterol

Schizophrenia

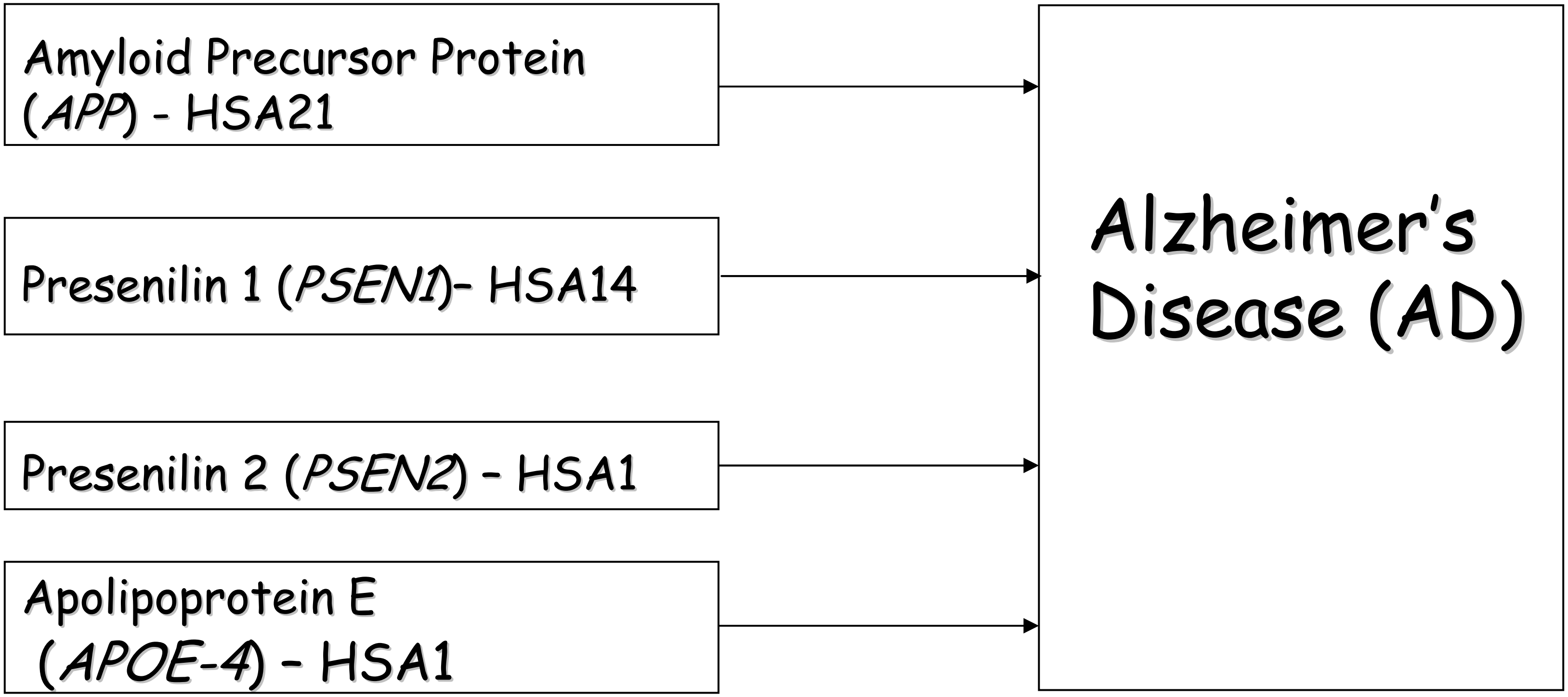
Normal Distribution in Phenotype of Complex Disease



Environmental effects



Locus Heterogeneity in Alzheimer's Disease



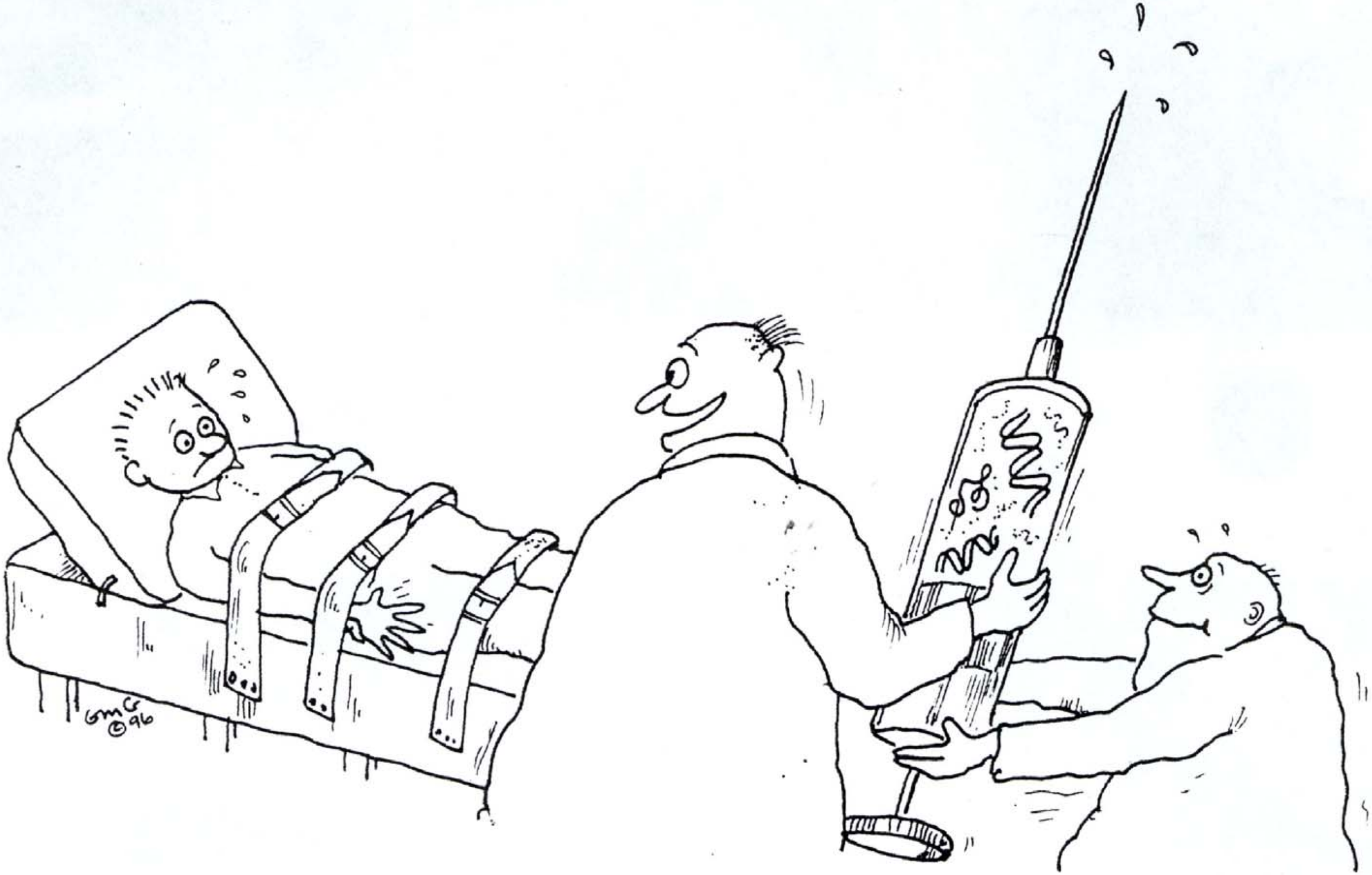
Amyloid Precursor Protein (*APP*) - HSA21

Presenilin 1 (*PSEN1*)– HSA14

Presenilin 2 (*PSEN2*) – HSA1

Apolipoprotein E (*APOE-4*) – HSA1

Alzheimer's Disease (AD)



'I'm afraid that whole-genome studies are an important precursor to developing small-molecule therapeutics…'

Genomics: Derivative Disciplines (1)

- Transcriptomics
  - Transcript is an RNA copy of a gene
  - Transcriptome is all RNA gene copies in a cell, tissue or individual
- Proteomics
  - Proteome is all proteins in a cell, tissue or individual

RNA Genomics

- High-throughput monitor gene expression
  - Array-based: expensive
    - Oligonucleotides vs. PCR products (cDNA)
      - *E.g.,* human fibroblasts, genes involved in **wound healing** are expressed when **starved fibroblasts** are induced to proliferate by serum
        - Wound healing is a normal function of proliferating fibroblasts
      - *E.g.,* tumor vs. non-tumor tissues

Protein Genomics

- Large-scale surveys of protein content in samples using two-dimensional gels (O'Farrell **1975**, **Proteomics**)

- **Mutagenesis**: insertional mutagenesis (Ross-MacDonald *et al.* 1999)

- **Yeast two-hybrid**
  - The mass testing of interactions among binary protein pair
    - <10^{-6} M

Genomics: Derivative Disciplines (2)

- Metabolomics
  - All of the small molecule components of a **cell**, **tissue** or **individual** that are produced by the proteins of the proteome

Functional Genomics

- **What to know**
  - Gene Expression
  - Gene Regulation
  - Genome-wide Mutagenesis

- **How to do**
  - Microarray analysis
  - Transposon targeting
  - Transgenics
  - RNAi

Genomic Information on the Horizon – Next 10 Years (1)

- Structural genomics & bioinformatics
  - Prototype protein ⇨ accurate modeling by homology of proteins
    - Related by sequences
    - But how many?
  - Protein mass spectrometry (MS) & **bioinformatics**
    - Genome sequences ⇨ prediction of their mass ⇨ compare with mass spectrometry measured (databases)

Genomic Information on the Horizon – Next 10 Years (2)

- Difficulties
  - Genome data do not immediately address the question about
    - Regulation
    - Mechanism
  - Genome data are prone to errors (due to high-throughput pressure)
  - Bioinformatic prediction ⇨ lab experimentation confirmation

Genomic Information on the Horizon – Next 10 Years (3)

- Limitations of bioinformatics nowadays
  - **"Guilt by association"**
    - A gene whose transcription behavior resembles that of **a known gene** may function in the same process as the known gene
  - **"Post-hoc" (**因果關係）
    - A gene whose transcription is **induced before** transcription of a group of another genes may regulate transcription of that group of genes

Genomic Information on the Horizon – Next 10 Years (4)

- To combine different data types & bioinformatics
  - mRNAs encoding those proteins are expressed **in the same cell** at the same **time** ⇨ strengthens the idea that the two proteins **interact**

Genomic Experimentation (1)

- Most of the strong conclusions will continue to come from directed experimentation
  - Bright researchers (IQ & EQ)
  - Trained for years
  - Expert in the system/organism in which the experiments are performed
  - Well-funded

Genomic Experimentation (2)

- [Bacon 1962] Science proceeds by the formulation & carefully **testing of hypotheses**
  - Observation-, obsession-, engineering-, or ‘what-if”-driven hypothesis play a small part
- Genomics de-emphasis of hypothesis-driven research
  - Valuable knowledge can be gained from the systematic production of simple kinds of biological information
  - Genomic research ⇨ observational

Genomic Experimentation (3)

- Stereotypical hypotheses
  - Transcription of genes in the kidney may be controlled by transcription regulatory proteins present in the kidney
  - Must be some mutations cause abnormality

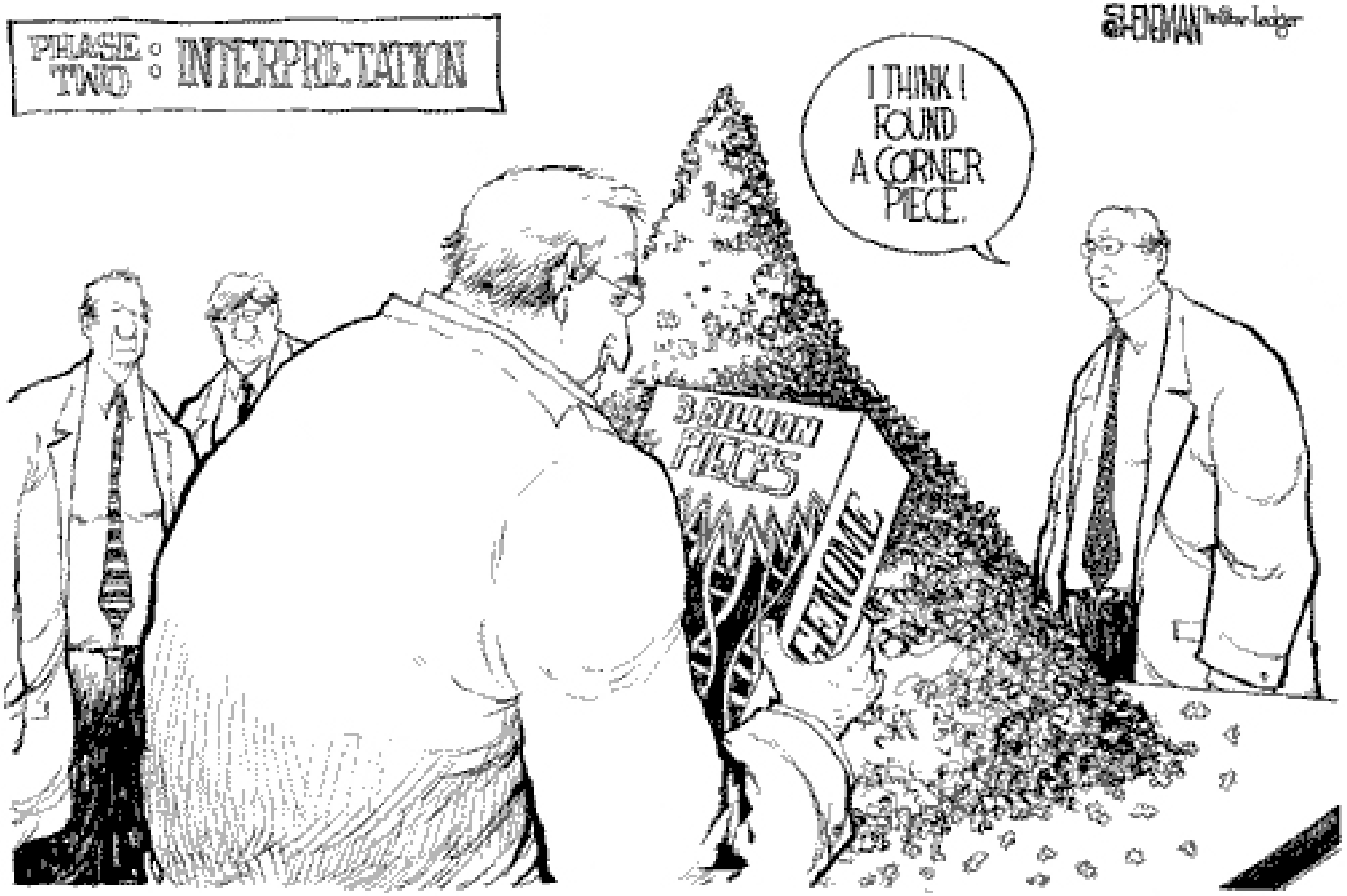
- Scientific standards have changed
  - 1988, the finding that a protein contains a **homeobox** ⇨ suggested DNA-binding & regulate expression
    - Have been tested **experimentally**
  - 2000, we would accept that claim without further experiment

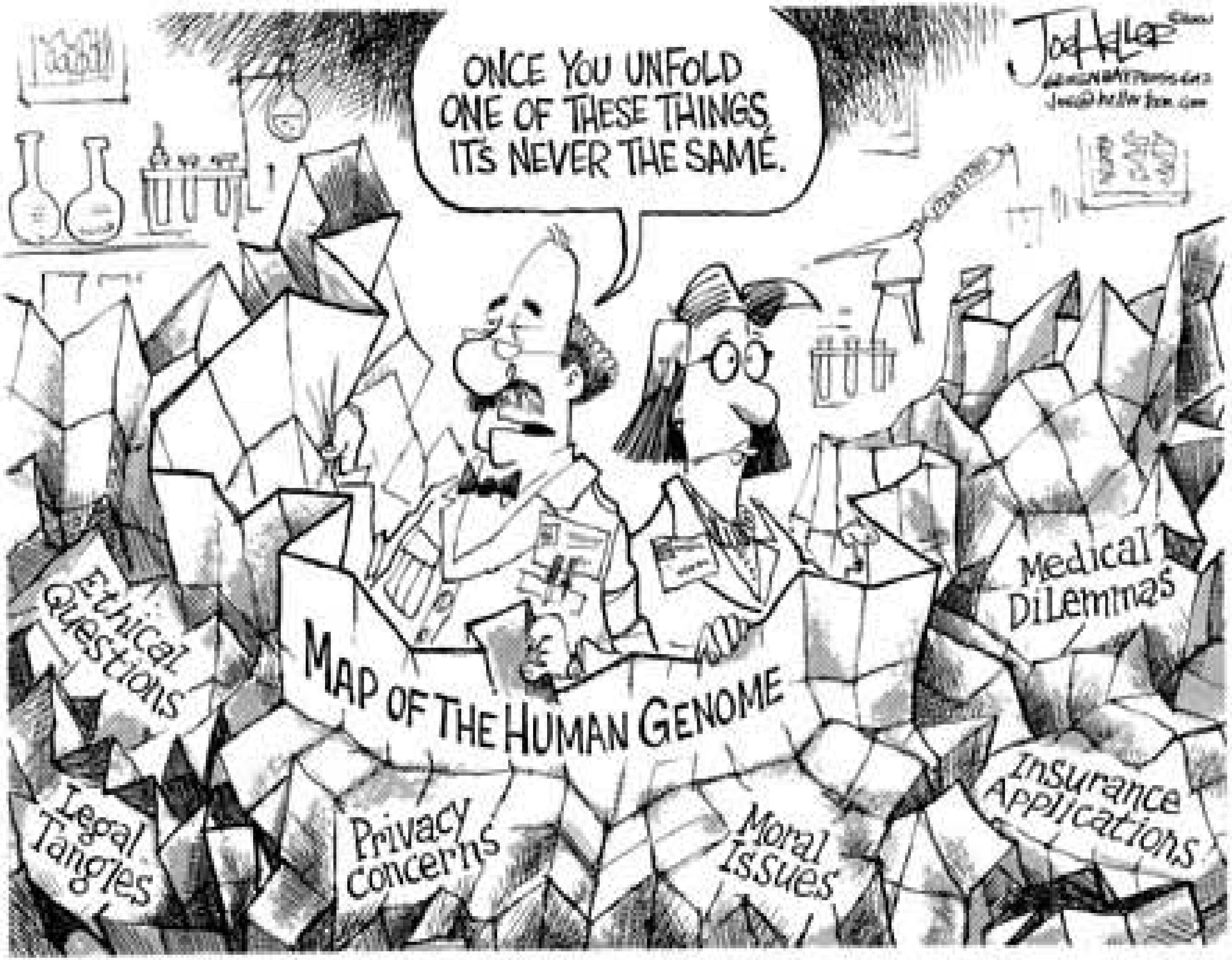
Post-Genomic Age

- **Mammalian genomes**
  - 25,000 – 30,000 genes
  - With ~8,000 known function
  - How long to solve the functions of all genes?

- **Structural Genomics**
  - Map-base gene discovery → sequence-based gene discovery

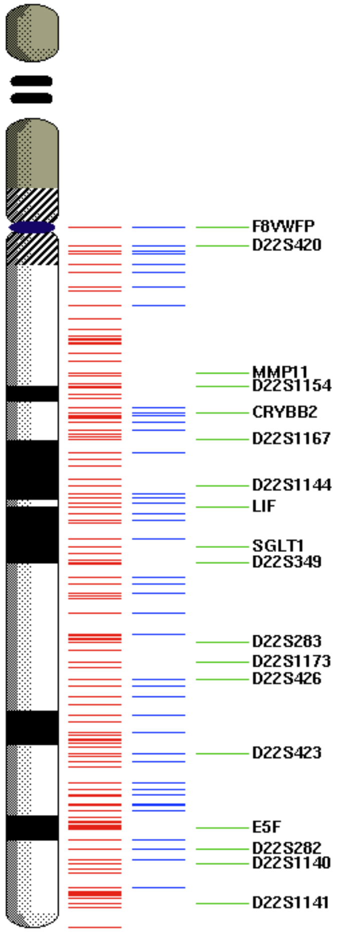
- **Functional Genomics** (mutation analysis)
  - Transgenic model organisms
  - ES cells knock-out
  - Transposition
  - PTGs (RNAi)


PHASE TWO : INTERPRETATION
SHENEMAN The Star-Ledger
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME


ONCE YOU UNFOLD ONE OF THESE THINGS, IT'S NEVER THE SAME.
Joe Heller
MAP OF THE HUMAN GENOME
Ethical Questions
Medical Dilemmas
Legal Tangles
Privacy Concerns
Moral Issues
Insurance Applications

Mapping the Genomes

- Map components
  - Markers or genes
  - Locations: mapping
    - **Linkage map**
      - cM (1 cM ~ 10^6 bp)
    - **Physical map**
      - Base pairs (bp)
  - HSA= Homo sapiens autosomal

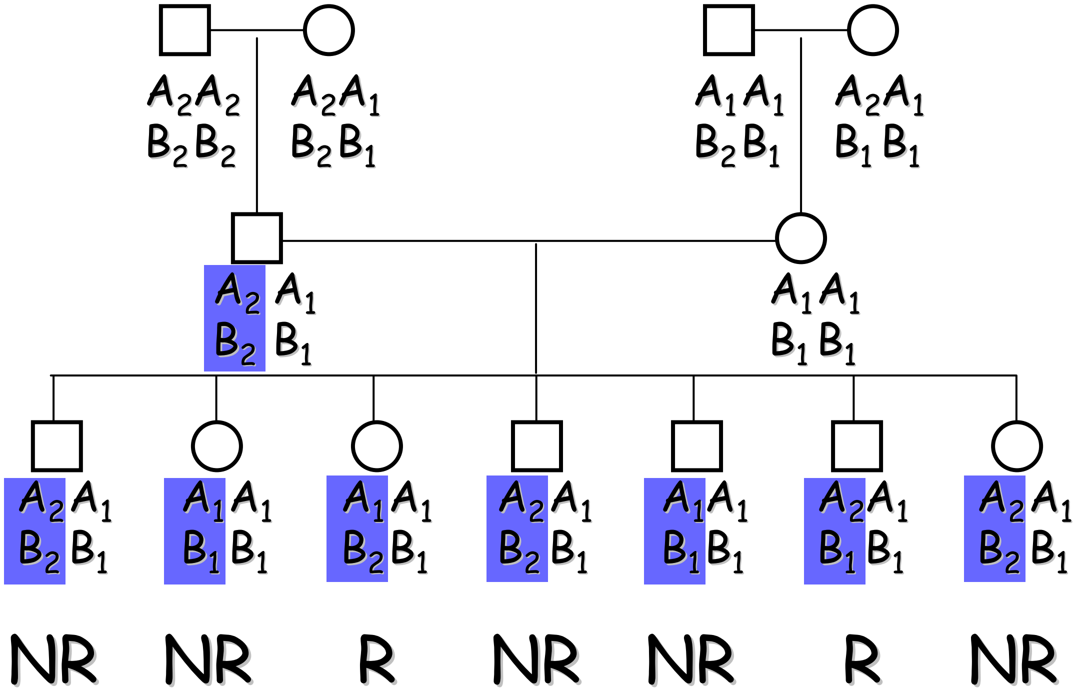


HSA22

Genetic Mapping (1)

- Requires informative **markers** – **polymorphic**
- A population with known relationships – **pedigree**
- Best if a measured **between** **"close" markers**
- Unit of distance in genetic maps = **centimorgans**, **cM**
  - 1 cM = **1% chance** of recombination between markers

Genetic Mapping (2)



θ = # recombinant / # total = 2/7 = 0.286

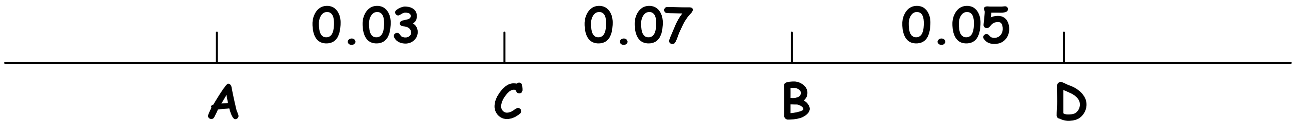
Example

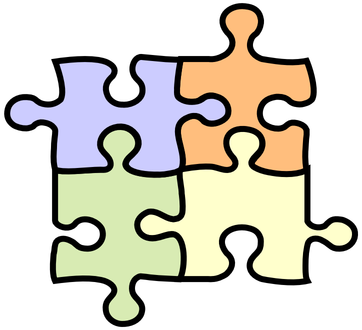
Table. Example Development of a Genetic Map using Four Linked Loci A, B, C & D, scored in 100 offspring[a]

Locus Scored	**Number of Recombinants**	**Frequency of Recombination**
A-B	10	0.10
A-C	3	0.03
A-D	15	0.15
B-C	7	0.07
B-D	5	0.05
C-D	12	0.12

[a]: no ***interference***

Loci Order & Recombination Fraction



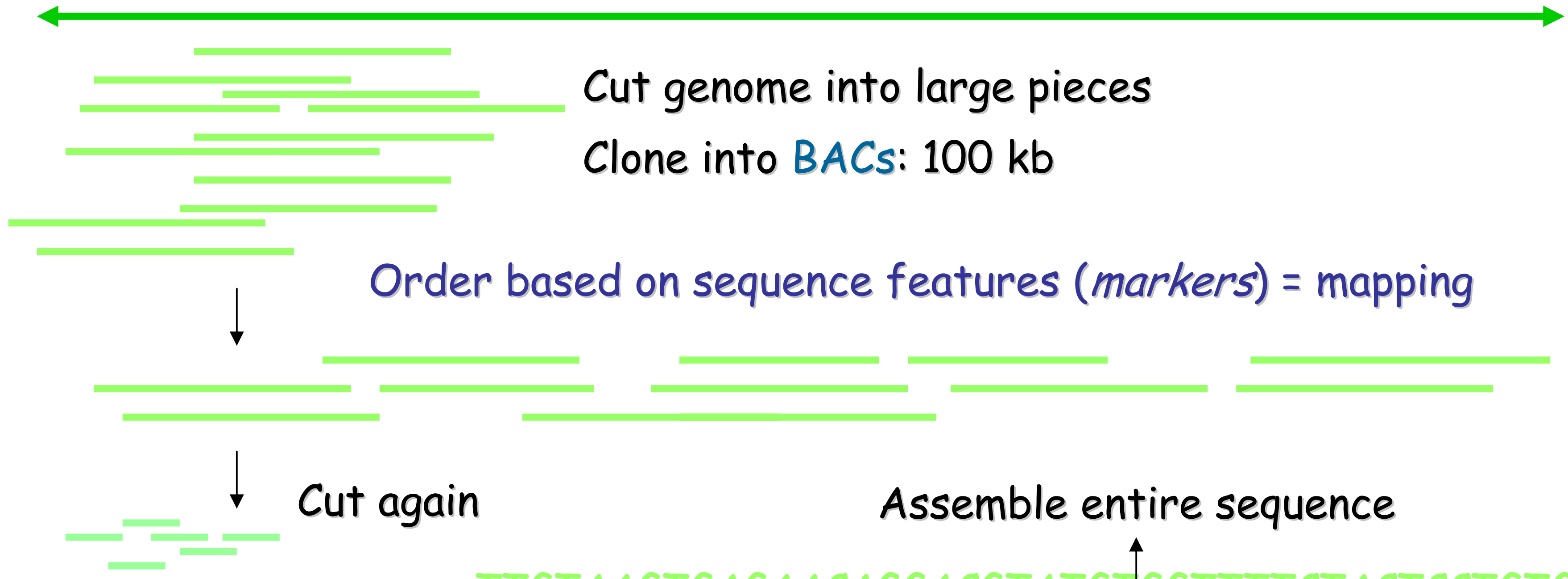


Two Strategies for Sequencing Genomes

- The Clone Contig Approach **(up-down)**
  - Relies on shotgun sequencing as well
    - But on **a smaller scale**
- The Shotgun Approach **(bottom-up)**
  - A length of DNA
    - A defined subset of the genome
    - A whole genome

Genome Sequencing

Genome: 3 Gb



Cut genome into large pieces

Clone into BACs: 100 kb

Order based on sequence features (*markers*) = mapping

Cut again

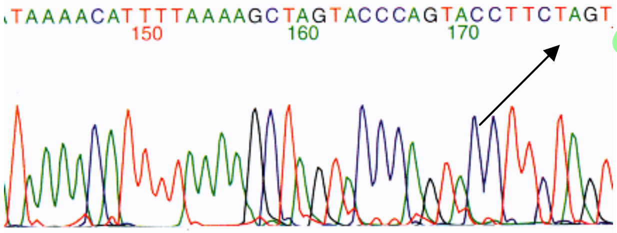
Assemble entire sequence

...TTGTAAGTGAGAACAGGACGTATGTGGTTTTCTACTCCTGTGTT...

Assemble each BAC

TTGTAAGTGAGAACA
AGAACAGGACGTATGTGGT
TGTGGTTTTCTACTCC
CTACTCCTGTGTT

Sequence



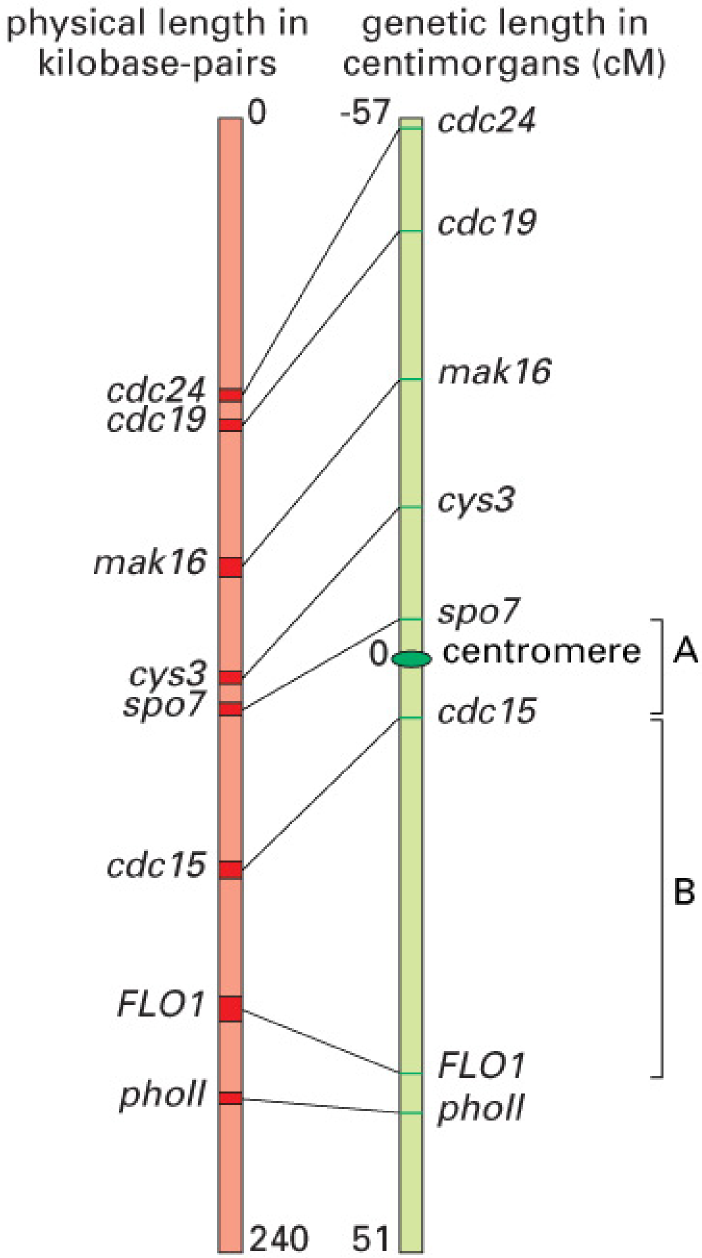


Figure 20–14. Molecular Biology of the Cell, 4th Edition.

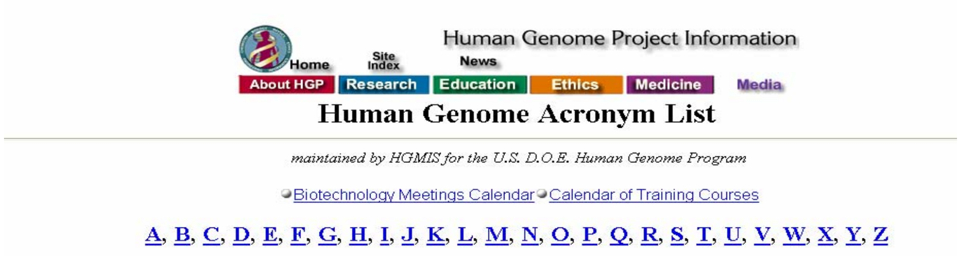
- By measuring the **reciprocal exchanges** in **meiosis**, a genetic map can be constructed
  - Genetic distance is roughly **correlated** with physical distance
- Genetic and physical maps help to identify **genes responsible for specific processes**

Genome Glossary

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION

NCBI Coffee Break

DOE Human Genome Program Research in Progress



Human Genome Acronym List

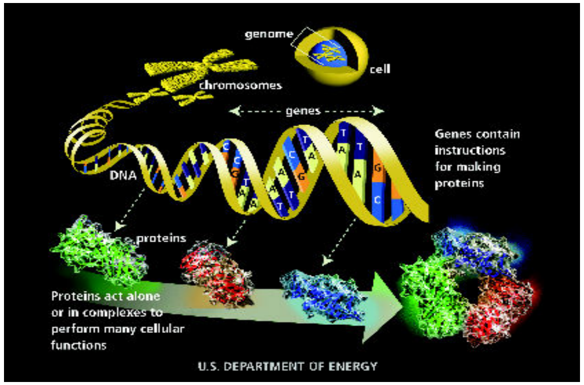
maintained by HGMIS for the U.S. D.O.E. Human Genome Program

Biotechnology Meetings Calendar Calendar of Training Courses

A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z

Genome & Biotechnology Meetings Calendar

Genetics 101


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

Coffee Break

- The Japanese eats little fat and suffer fewer heart attacks than the British or Americans
- The French eat a lot of fat and also suffer fewer heart attacks than the British or Americans
-
- The Italians drink a lot of red wine and also suffer fewer heart attacks than the British or Americans

- **Conclusion**: Eat and drink what you like. Speaking English is apparently what kills you.

- By Irwin Knopf
- Retyped by Zoey Chen

Major Implications of the Genetic Revolution for the Legal Discipline (1)

- How **regulation** will be possible in the fast moving genetic revolution
- What are its **implications** for **human dignity and human rights**
- Should the law condone **interventions** in the human genome which **alter the genetics of living persons** and future generations

Major Implications of the Genetic Revolution for the Legal Discipline (2)

- What will be the implications of these developments for **family law**

- What **consequences** will they present for insurance, given the potential of genetic data to remove entirely predictive doubts about an insured's likely health prognosis

- Will the criminal law need to be revised in so far as it posits the free will of the individual? If the conduct of some persons stems from their genes, should this be exculpation, a defence or at least mitigation

Genetic Discrimination (1)

- All disease has one or more **genetic components**
  - Therefore, we are all **at risk** for genetic diseases
  - If we accept these statements, then there is **no basis for genetic discrimination**, since we are all in the same risk pool
  - **But** the insurance industry is **based on** the ability to discriminate and assign risk

Genetic Discrimination (2)

- At this point in the evolution of our knowledge, we have the information to permit us to identify **predisposition** to certain relatively rare genetic diseases, *e.g.*,
  - CF, Huntington disease *etc.*

- The **burden of genetic disease**, however, **is among all of us** with predisposition to common, complex genetic disease, *e.g.*, cancer, cardiovascular disease, diabetes mellitus *etc.*

Genetic Discrimination (3)

- William Brody, JHU President, in a recent Wall Street Journal op-ed (opposite editorial page) piece, argued that **the loss of ability of health insurers to stratify populations by genetic risk** will lead ultimately to a single payer

Manhattan Project of Biology

- Al Carnesale , UCLA Chancellor
  - “We have just come through **the Manhattan project** of biology. **Let’s get it right this time**”
    - Ethical, Legal and Social Issue (ELSI) Program, NIH
    - US DHHS Secretary’s Advisory Committee on Genetic Testing (SACGT) and Secretary’s Advisory Committee on Genetics, Health and Society (SACGHS)
    - UCLA Center for Society, the Individual and Genetics

Small Business & Health Insurance (1)

- A patient who works for **a small self-insured company** has a positive family history for emphysema（肺氣腫）on both her mother’s and her father’s sides

- Her physician recommends that she have a number of tests performed, including one for α1-antitrypsin (α1AT)

- When the α1AT test is reported to be **abnormal**, he tells her that this may explain the emphysema in her family and **places her at very high risk** this lung disease

- Her physician reports the results of his evaluation **to her insurance company** as required

- Several days later she is called into the office of her employer and fired

Small Business & Health Insurance (2)

- Actual case
  - Patient had symptoms at time of testing
- Commissioner Paul Miller, EEOC, argued this case under ADA
  - EEOC = Equal Employment Opportunity Commission (美國)就業機會均等委員會
    - Settled in favor of employee
    - Remains to be determined whether an abnormal test result in absence of **physical signs** and **symptoms** would be covered by ADA
    - ADA: Americans with Disabilities ACT