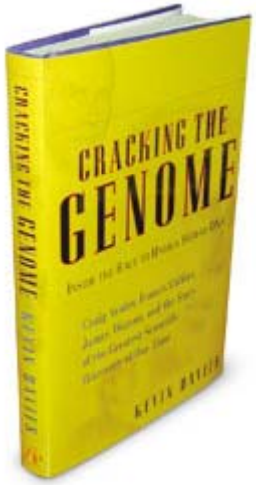
# 『基因體計畫 』與『生物資訊學』之崛起與現況

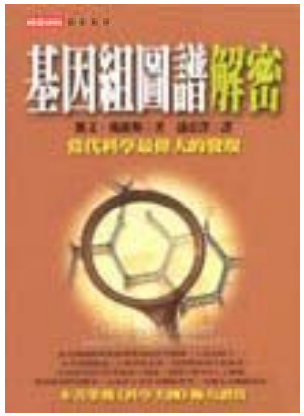
薛 佑 玲 Yow-Ling Shiue
國立中山大學生物醫學研究所
ylshiue@mail.nsysu.edu.tw

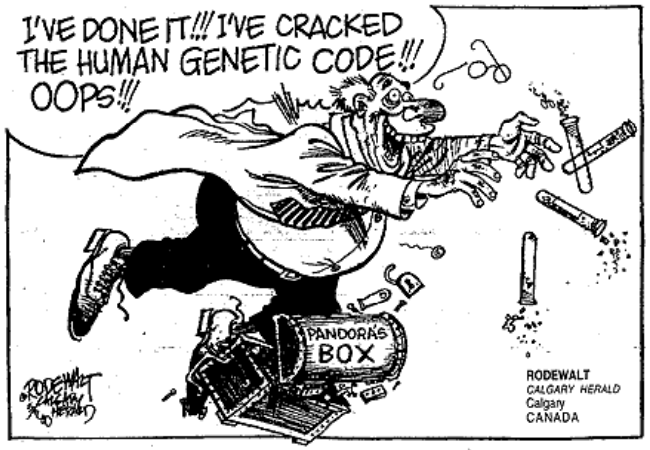
# Readings

- 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。

- G Gibson & SV Muse (2002) A primer of Genome Science. Sinauer Associates, Inc. Publishers.
  - Chapter 1: Genome Projects: Organization & Objectives



"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.]

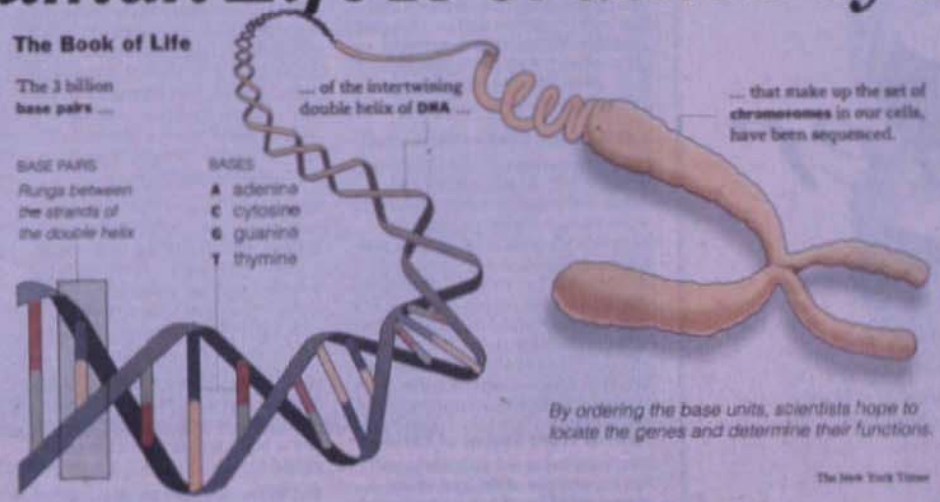
Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

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 his 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




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 Sakuma/The New York Times

## 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

## 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.

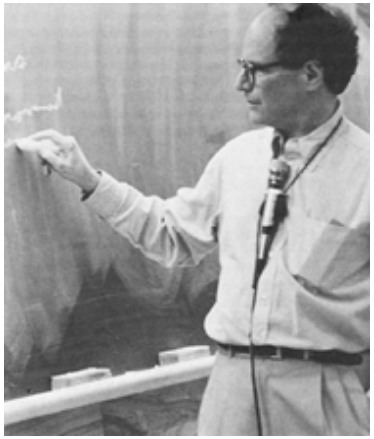
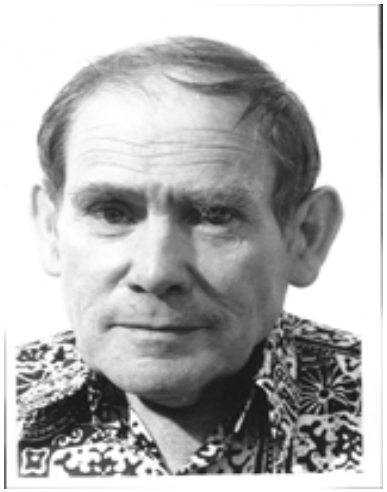
"Today we are learning the language in which God created life," President Clinton said at a White House ceremony attended by members of the two teams and, via satellite, Prime Minister Tony Blair of England. [Excerpts, Page D8.]

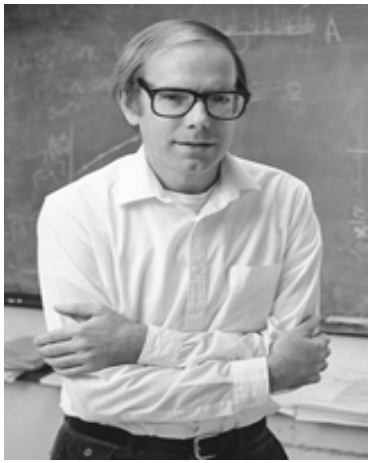
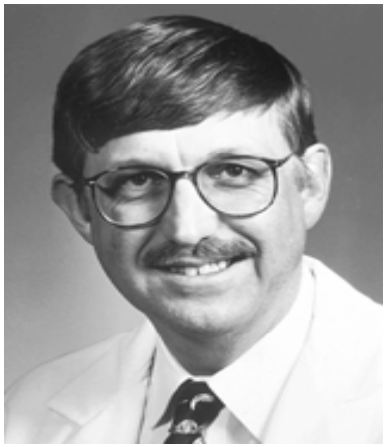
The teams' leaders, Dr. J. Craig Venter, president of Celera Genomics, and Dr. Francis S. Collins, director of the National Human Genome Research Institute, praised each other's contributions and signaled a spirit of cooperation from now on, even though the two efforts will remain firmly independent.

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

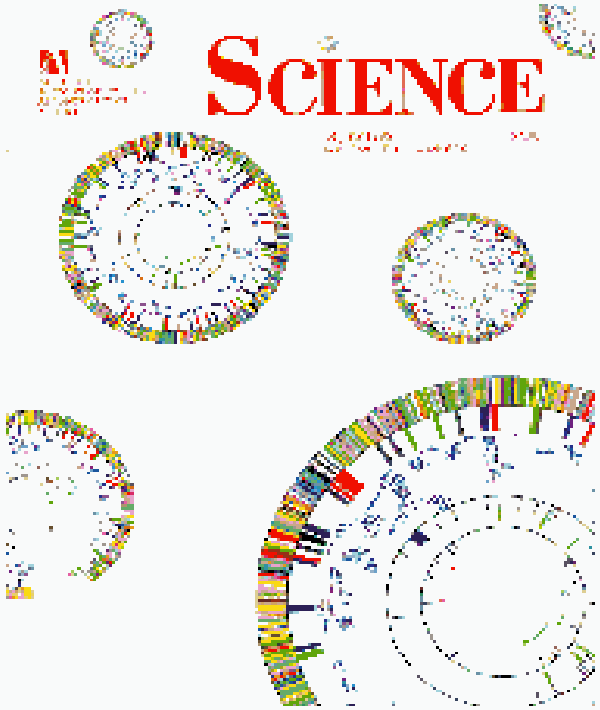
# The Genome Crackers

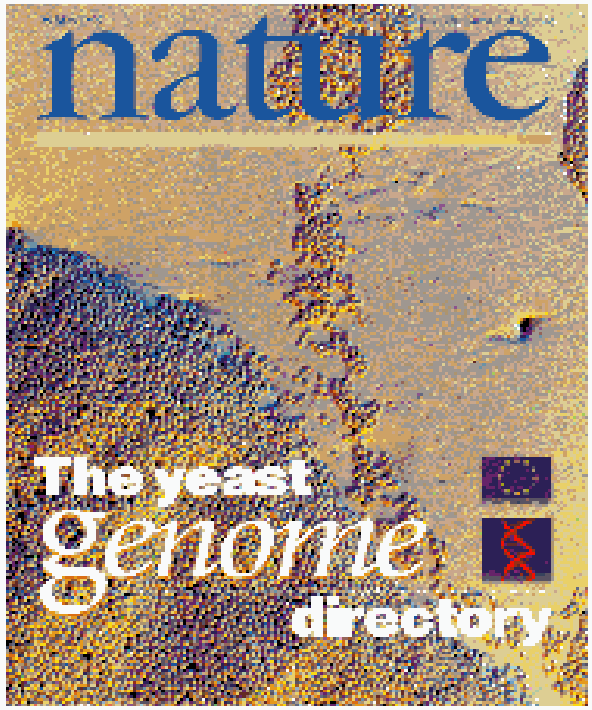




- **Walter Gilbert:** A crucial early proponent, he later tried to set up a company to produce and sell genome data
- **Sydney Brenner:** Joked that sequencing was so boring it should be done by prisoners
- **Charles DeLisi:** An early advocate, he launched the Human Genome Initiative within the **Department 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
- Threw down the gauntlet with **J. Craig Venter:** his commercial plan to shotgun sequence the human genome

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



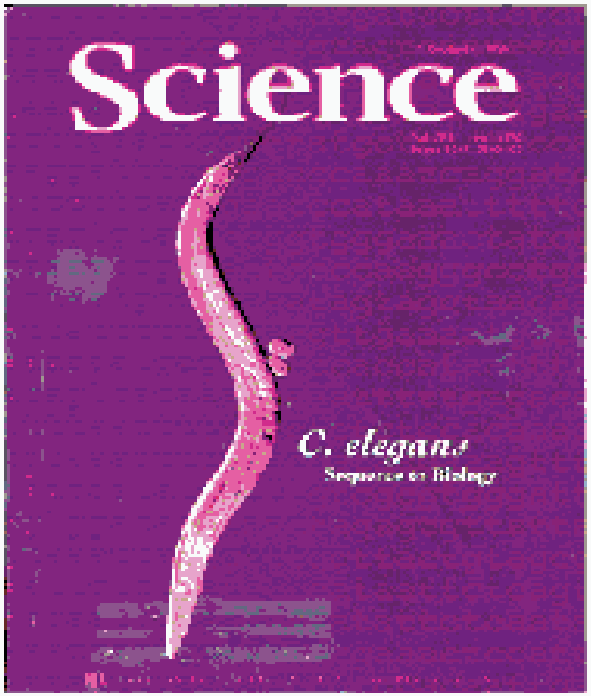



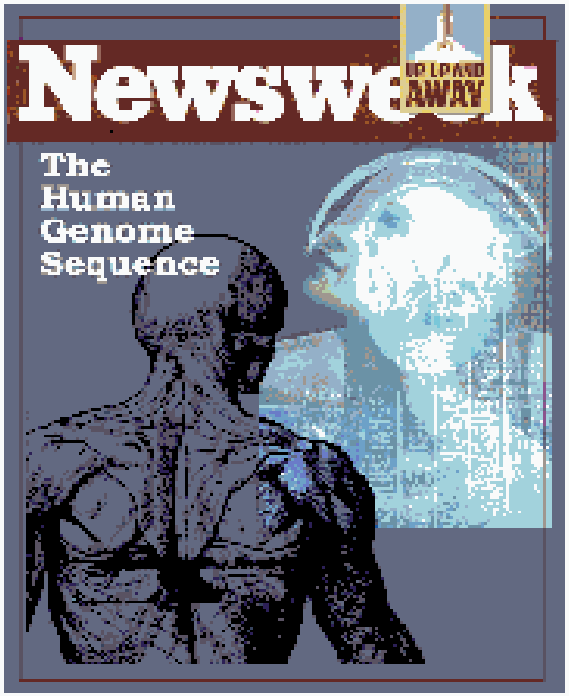

**1995**

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]

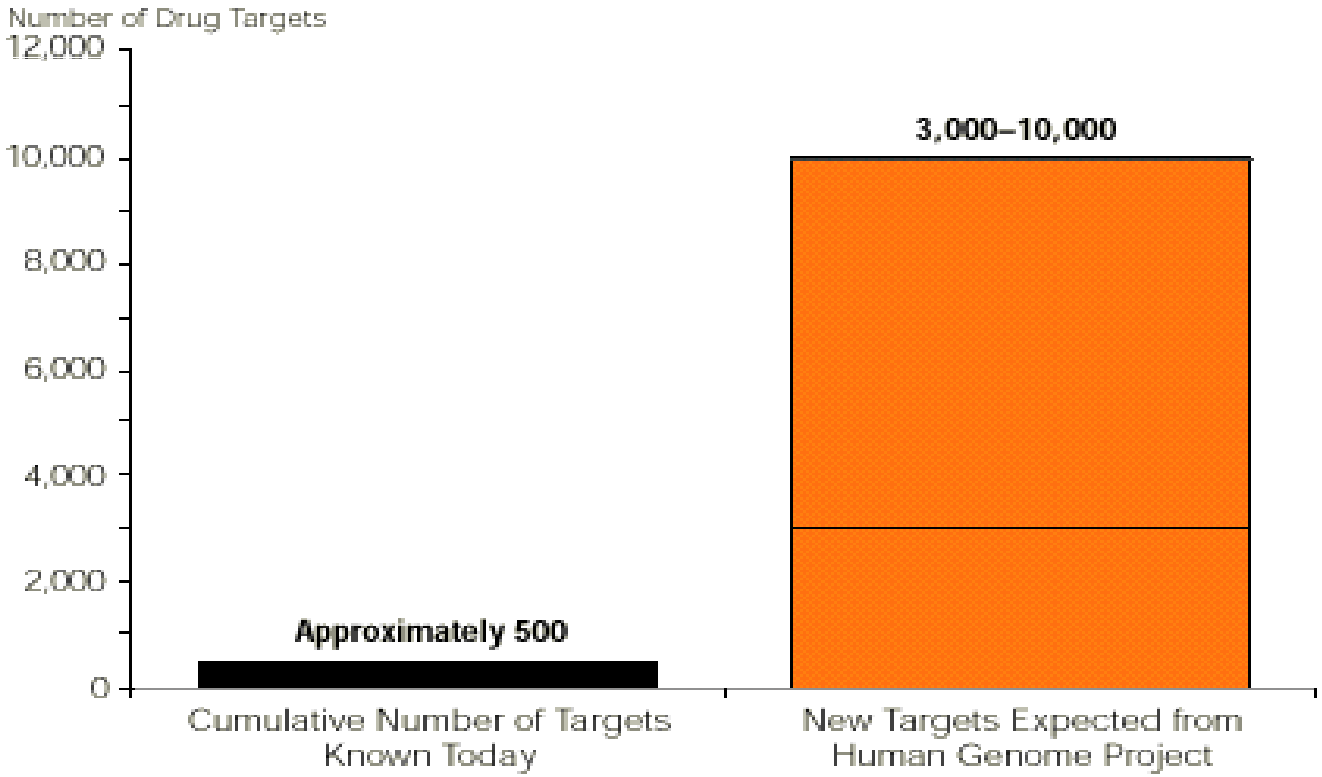
**2000?**

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**



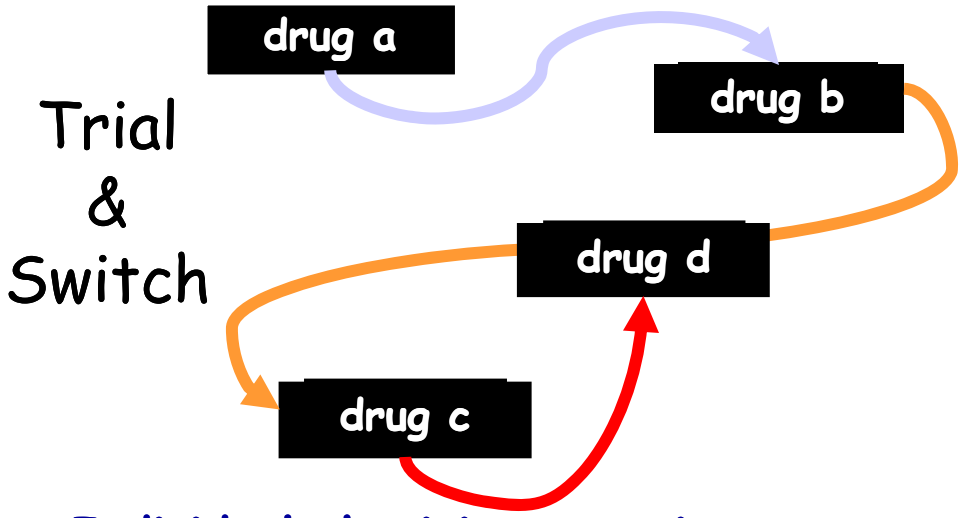

*Source: Drews, Jurgen, M.D., "Genomic Sciences and the Medicine of Tomorrow: Commentary on Drug Development,"* Nature Biotechnology, *Vol. 14, November 1996.*

# Targeted Prescription of Medicines: Applied Pharmacogenomics

**Today**

Empirical prescription

"**mass market**"



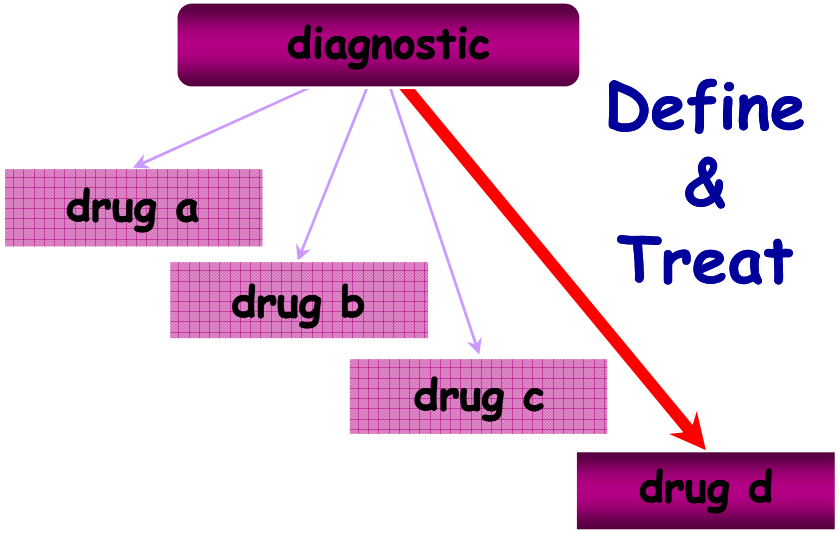

**Individual physician experience**

**Cost: time, money & well-being**

**Future**

Rational prescription

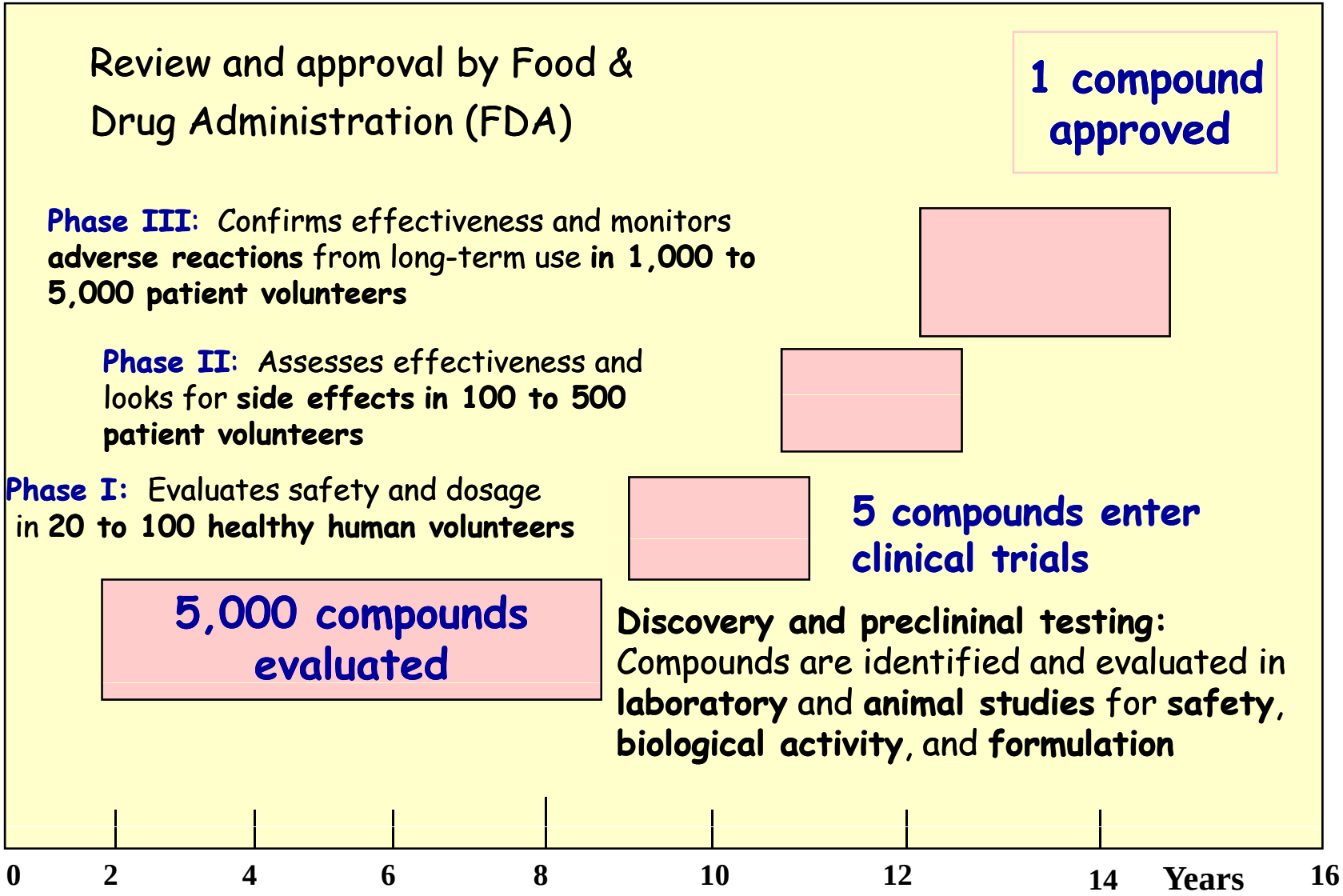
"**individualized**"




Informed physician diagnosis

**Savings**: time, money & illness

# Bringing a New Drug to Market




Source: Tufts Center for the Study of Drug Development

# Human Genome Project 1988

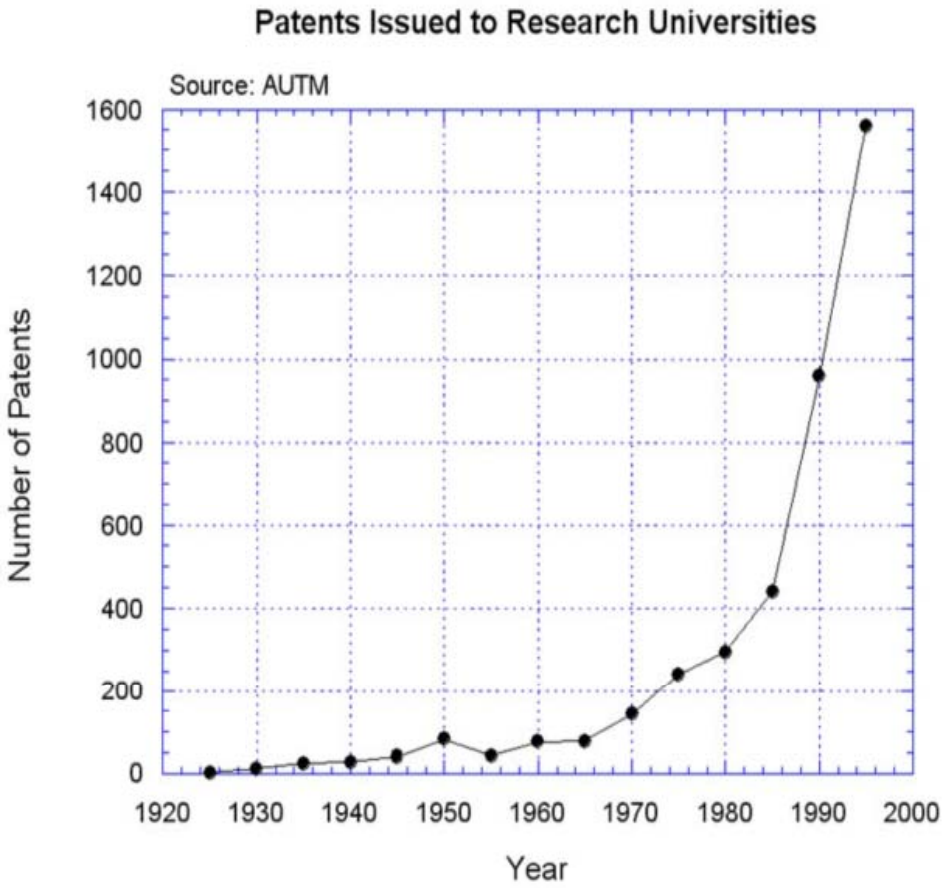
- Conceived as a resource for the **scientific community**
- Sharing of **genomic resources** and **IP** (Intellectual Property) rights a major concern
- HGP grounded on belief that **science is the best served by free access to genomic resources** such as DNA sequence
- Genome is a bounded set of fundamental information that should be **available to all**

# IP Rights to Federally Funded Research Results (1)

- Bayh-Dole Act 1980
  - Grantee/contractor retains rights to **inventions**
  - **Universities & non-for-profit institutes**
  - Enacted into law in 1984
  - **Exception**
    - **Declaration of Exceptional Circumstances (DEC)** invoked by government **to prevent patents** by grantees/contractors
  - March-in rights if invention not developed appropriately

# IP Rights to Federally Funded Research Results (2)

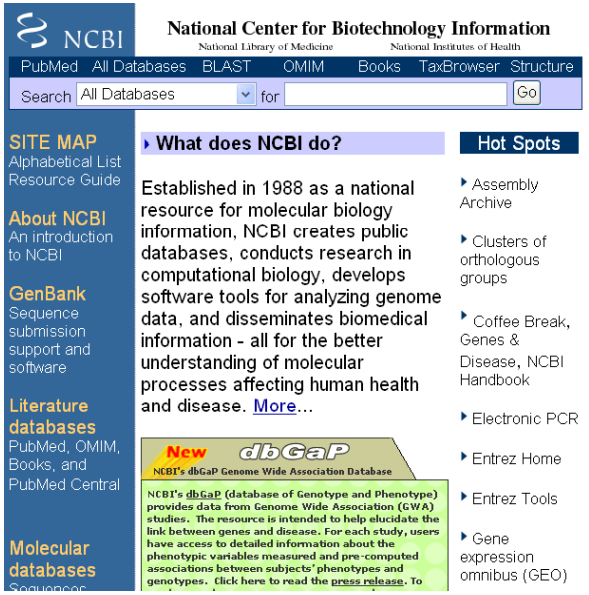
- Bayh-Dole Act 1980 (cont.)
  - **Benefits**
    - Encourages interactions between **academia** and **industry**
    - **Inventions** developed rapidly
    - Biotech industry has blossomed
  - **Problems**
    - Constraints on **availability** of some basic tools/resource
      - Reach – through rights
    - Stifling of innovation if there are **problems licensing underlying technology**

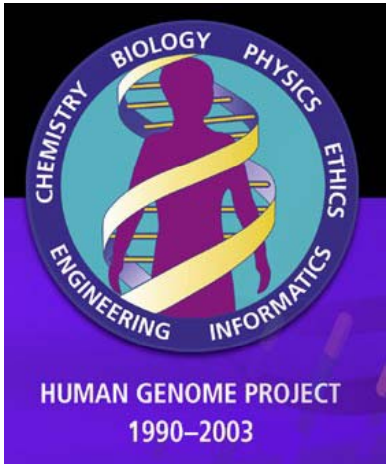


# IP Rights to Federally Funded Research Results (3)

- Results should be published
- Data & materials should be **shared at time of publication**
- Deposit in **public databases** or repositories when available

- Problems with respect to HGP
  - Difficult to **enforce**
  - Many data **not published** but **still useful**
  - Sharing at time of publication is **too late** for genome resource



# Basic NHGRI Sharing Policy

genome.gov
National Human Genome Research Institute
*Advancing human health through genetic research*

- Release of all data and materials **within six months** of generation
- **Applicants** asked to state their plans for sharing
- Awards made only **if plans acceptable**
- Plans for sharing become **condition of award**

- **NHGRI** = National Human Genome Research Institute

# NHGRI Policy on DNA Sequence

- Early genome products were **maps**, **markers** & **DNA clones**
- Later DNA sequence predominated
- New policy needed because DNA sequence
  - **Can be produced rapidly** in very large amount
  - Is **immediately** useful in raw form

http://www.genome.gov/PolicyEthics/LegDatabase/pubMapSearch.cfm

# Bermuda Agreement

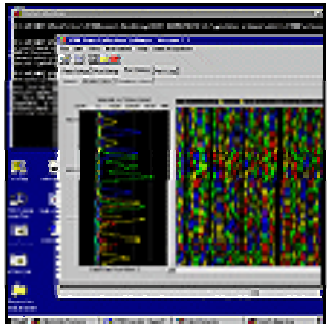
- 1996 1st International **Strategy Meeting** on Human Genome Sequencing
  - Principles enunciated
    - Sequence assemblies **greater than 1Kb** should be released automatically **on a daily basis**
    - http://www.genome.gov/Pages/Education/Kit/main.cfm?pageid=61
- 1997 2nd Meeting
  - Principles reaffirmed
- 1998 3rd Meeting
  - Principles extended to **all genomic sequence**
- NHGRI requires all grantees funded for production sequencing **to abide by these principles**


2000

1996


1997

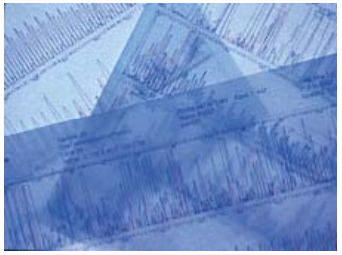
1998



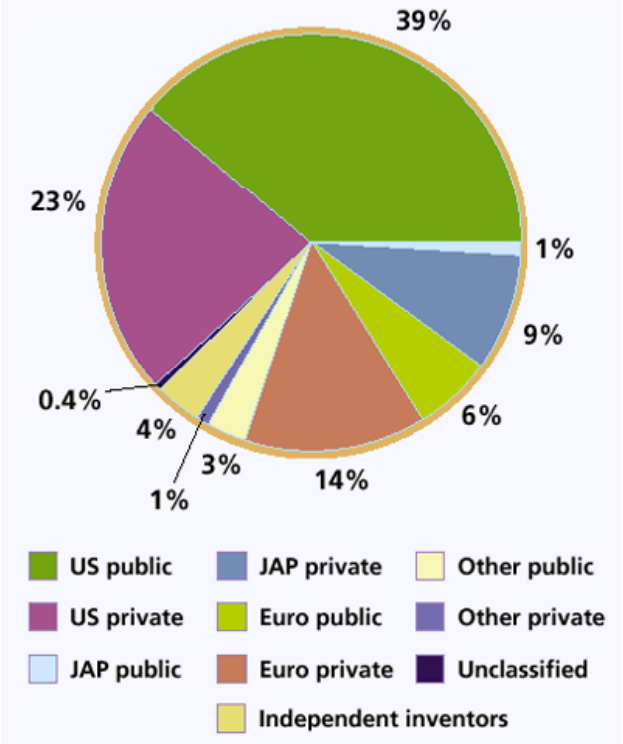
1999

# NHGRI Policy on Patenting DNA Sequence

- 1996 NHGRI announces policy on **patenting** human genomic sequence
  - Raw genomic sequence **in the absence of additional demonstrated biological information lacks utility** is **not** appropriate for patient filing
  - NHGRI requires **rapid release** of raw sequence & will **monitor** patenting of large blocks of primary sequence
  - If NHGRI determines there is a problem, a **DEC** may be considered
    - **DEC = Declaration of Exceptional Circumstance**
      - To prevent patents by grantees/contractors



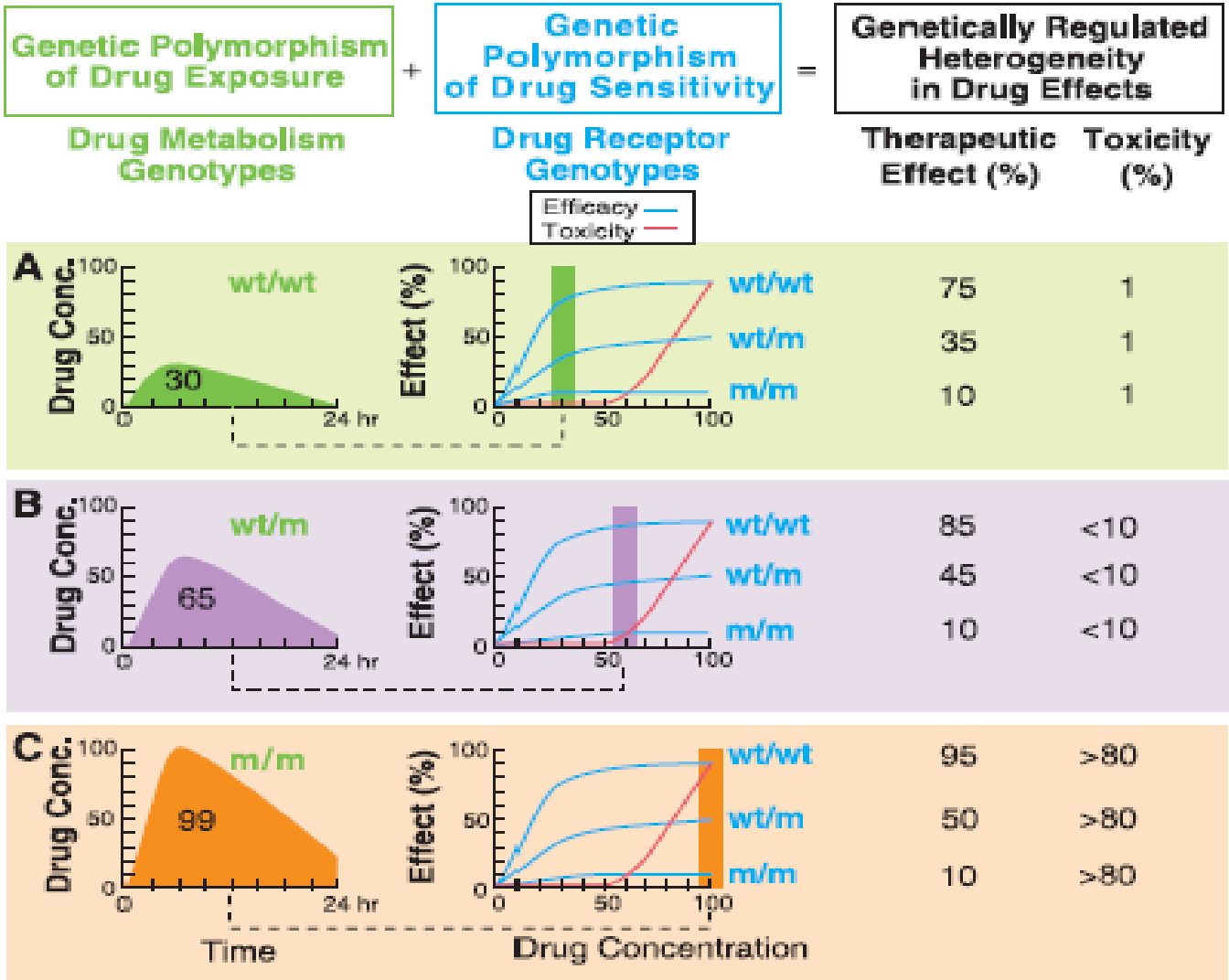
**NIH opposes plans for patenting 'similar' gene sequences**
**David Dickson**
*Nature* **405**, 3 (2000).




**Patents claiming DNA sequence filed between 1996 and 1999 by country and sector (Nature Biotechnology 2002, 20:1185-1188**

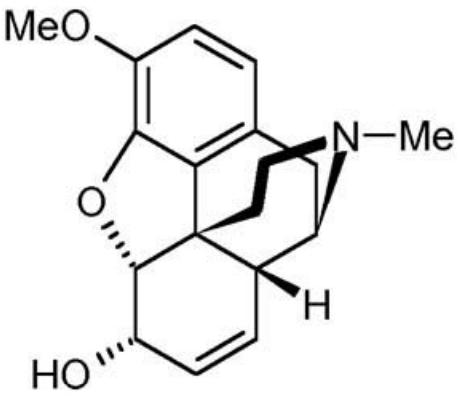
# Polygenic Determinants of Drug Effects

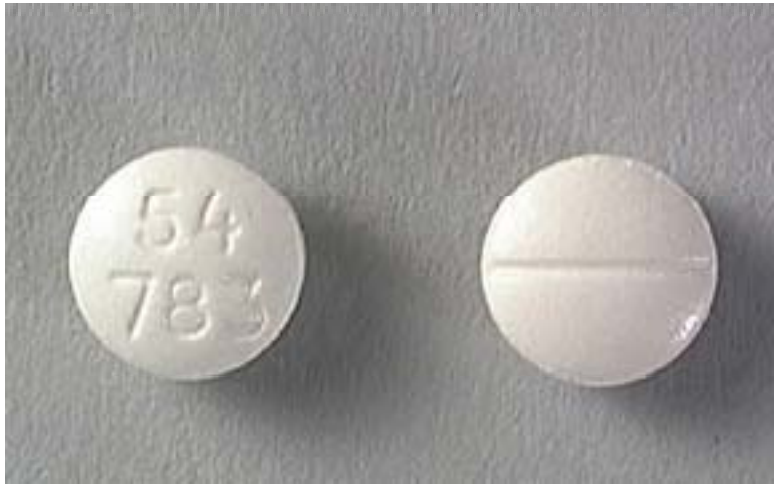
(Evans & Relling, 1999; Science 286, 487)



# SNPs in Drug-metabolizing Enzymes

- *CYP2D6*
  - Mutant alleles: responsible for individual variability in pain relief by opioid analgesics (止痛劑)
    - *E.g.*, **Codeine** (可待因)
      - Require **activation by *CYP2D6***
    - Individuals with non-functional *CYP2D6* **mutant alleles** → resistant to the effects of opioid analgesics
    - Several mutant alleles of the *CYP2D6* gene coding for **debrisoquine 4-hydroxyase** predispose to **toxicity** with
      - Metaprolol, timolol, nortriptyline, perhexeline, propafenone and **codeine**
  - Genetic tests (**genotyping**): to prevent potential toxicity by lowering dosages or **not** prescribing certain drugs → selection of optimal drug therapy



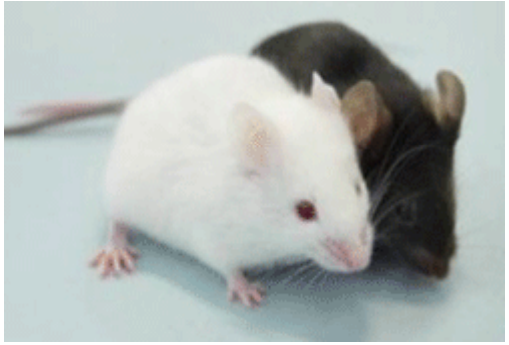



# NHGRI Policy on SNPs

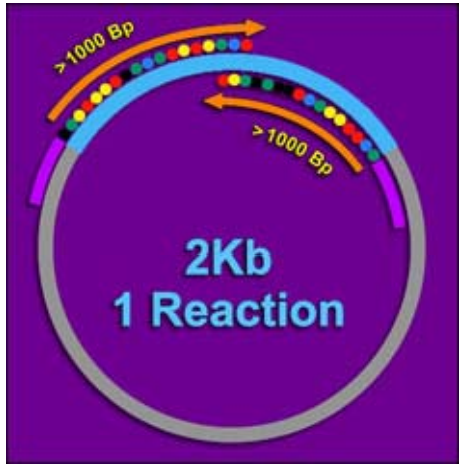
- Single Nucleotide Polymorphisms
- **SNPs** are a new kind of DNA markers with great utility for mapping genes
- Large sets covering entire genome are needed
- Important to have such sets **publicly available** to stimulate research on **genes involved in complex diseases** & other phenotypes
  - *E.g.*, Pharmacogenomics
- NHGRI SNP production **grantees must agree not** to seek patents on SNPs lacking demonstrated **functional utility specific to SNP(s)**
- The SNP Consortium (TSC)
  - Consortium of **pharmaceutical companies** that is also investing in production of a public SNP collect

# Examples of NIH Use of DEC for Genomic Research Tools

- Mouse mutagenesis and phenotyping centers
  - **Mutant mice** **may not** be patented
  - Other inventions such as new technology are not affected



- **Mammalian** Gene Collection
  - Full length cDNA clones & sequences
    - Clones & their sequences **may not** be patented
    - Other inventions not affected



# Recent Developments

- NHGRI policies on **research tools** being adopted across NIH
  - National Institute of Health (USA)
- New NIH policy statement issued in 1999
  - Outgrowth of recommendations of Working Group on Research Tools
    - 1988

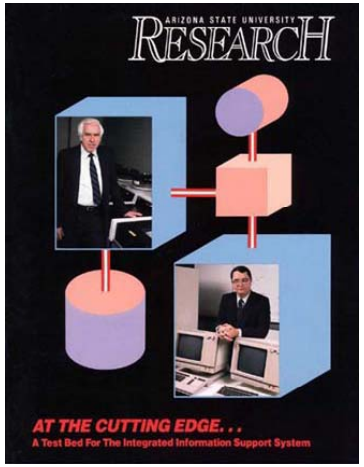
# New NIH Policy Statement, 1999 (1)

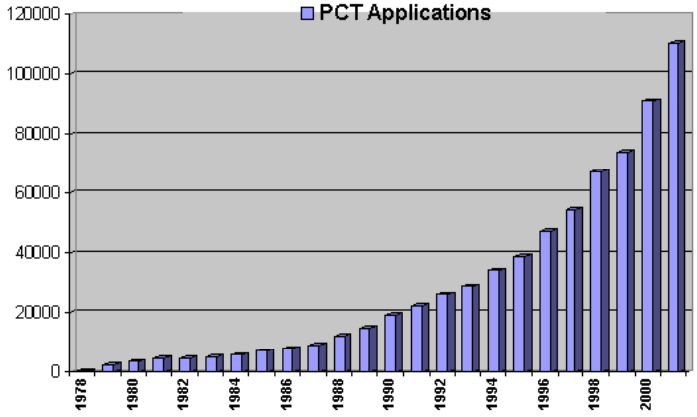
- Sharing **Biomedical Research Resources**
  - http://www.nih.gov/science/models/sharing.html

- **Principles**

  1. **Academic Freedom & Publication**
     - Institutions have a obligation to preserve academic freedom and ensure timely disclosure of research results

  2. **Appropriate Implementation of Bayh-Dole Act**
     - Intent of Act is to **promote** utilization of inventions & public availability
     - Use of patents and exclusive licenses is not always the best way to assure this



# New NIH Policy Statement, 1999 (2)

3. **Minimizing Administrative Impediments to Research**
   - Streamline process for **transferring tools to others**
   - Develop **clear policies** on acceptable conditions for acquiring tools from others

4. **Dissemination of Resources founded by NIH**
   - Progress in science depends on prompt access to **new research resources**
   - Unique resources developed with NIH funds are to be made available to **the research community**

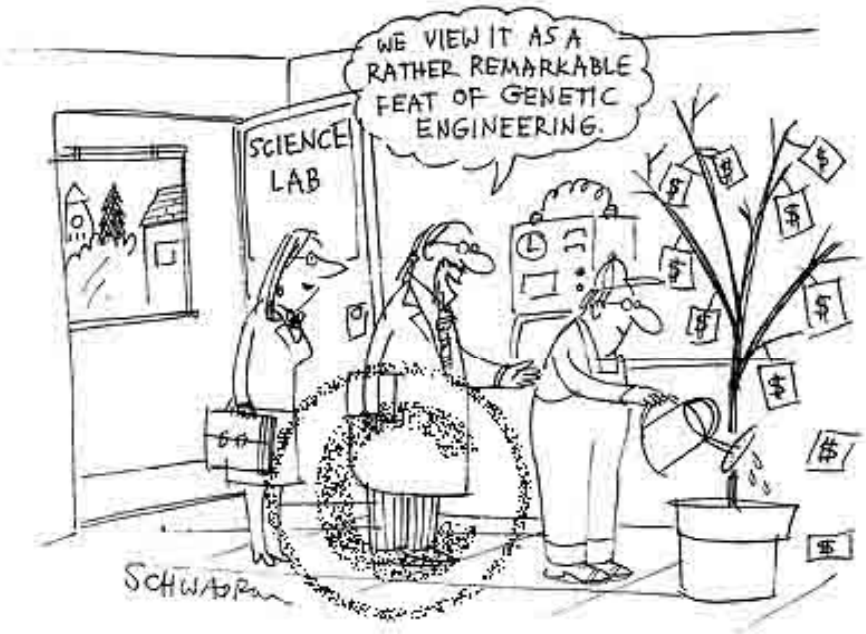
   - **Web address – full document** http://www.nih.gov/od/ott/Rtguide_final.htm
     - Ott= Office of Technology Transfer

# Gene Patents

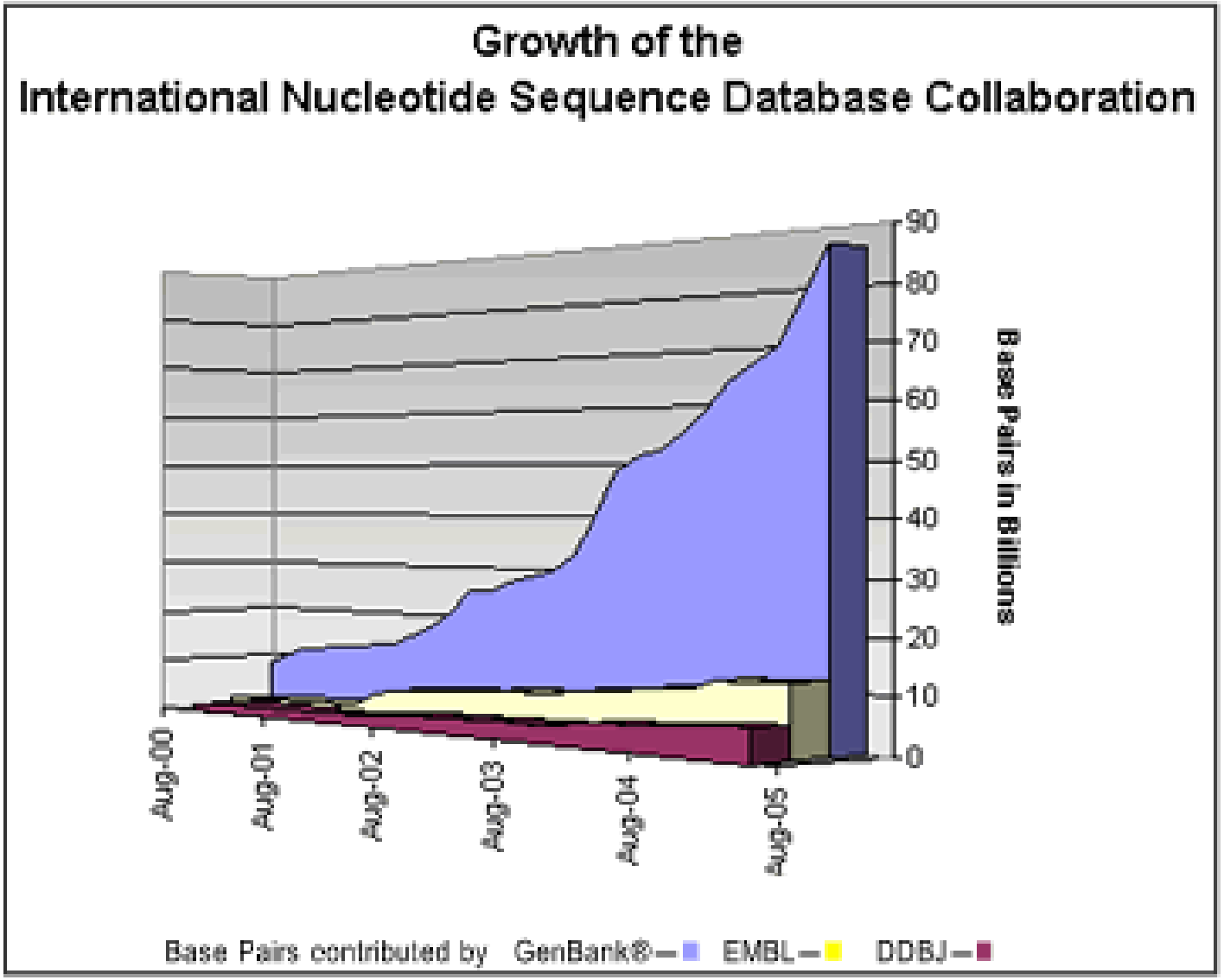
- Human genome international effort to sequence all the genes (1990)

- Craig **Venter** (NIH) - 1991
  - Filed **315 ESTs**
  - **Initiative failed**

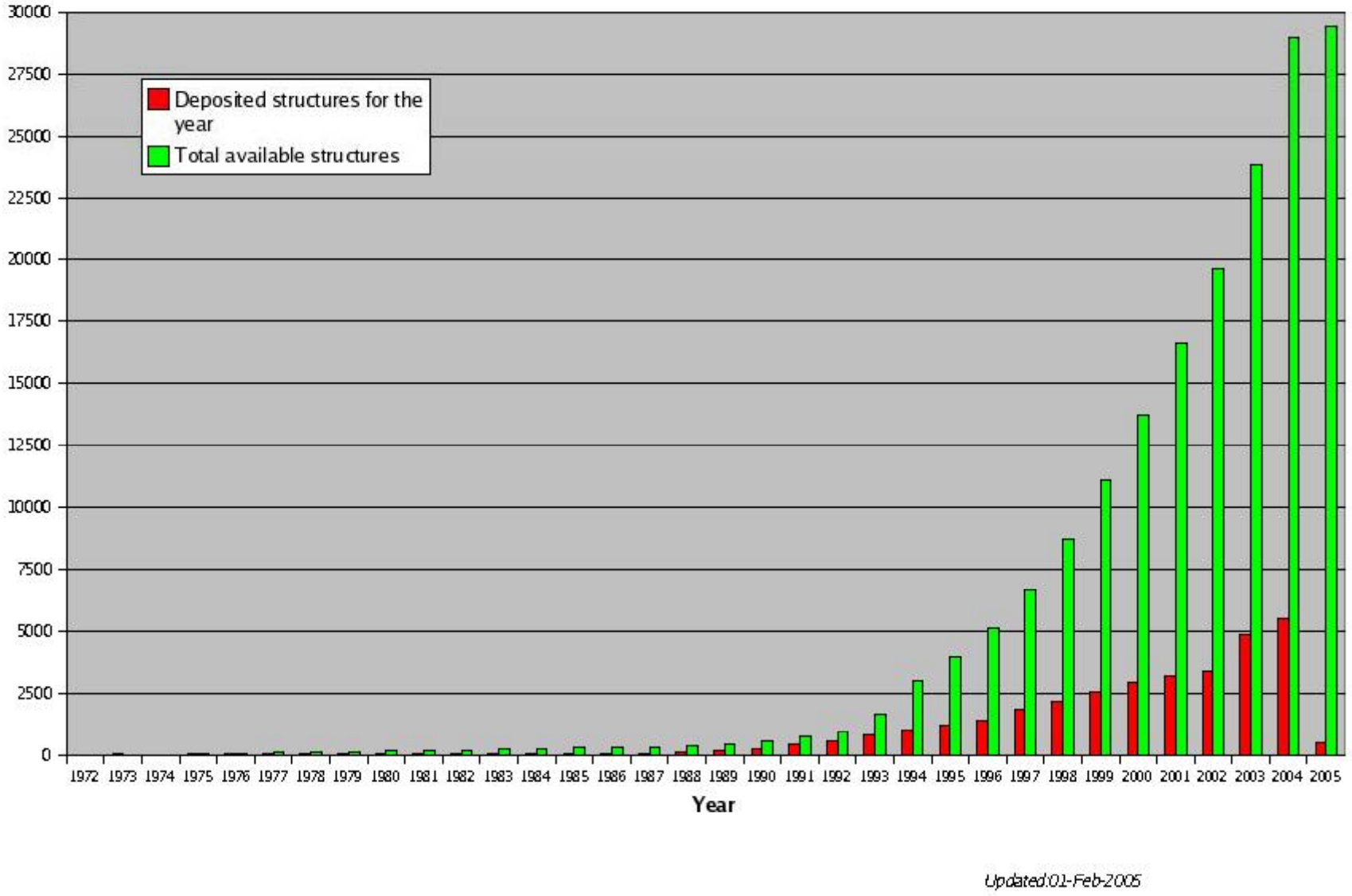
- Rejected on grounds of utilities (1992)
- NIH withdrew – could have appealed 1994
- Craig Venter from private sector filed many such gene patent applications




**DNA Patent Database**

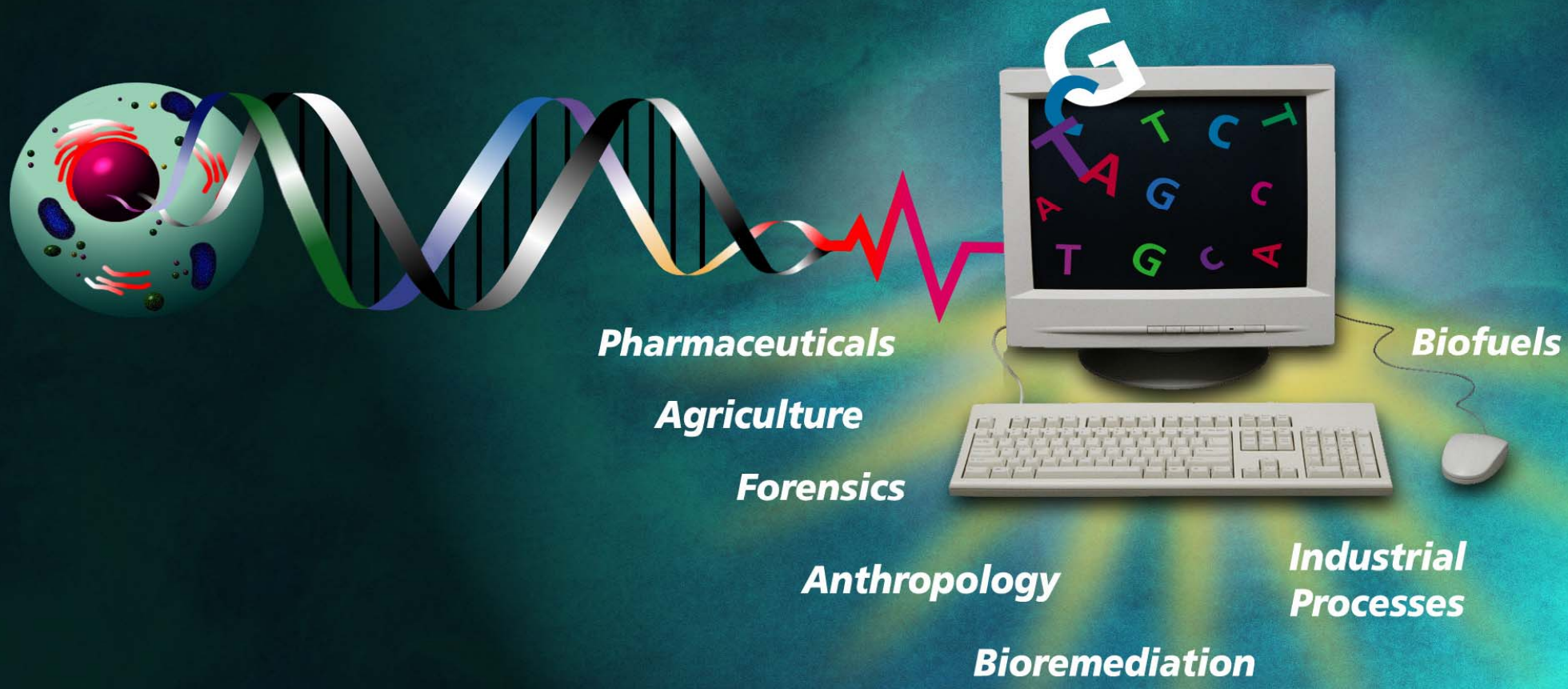



http://www.ncbi.nlm.nih.gov/Genbank/



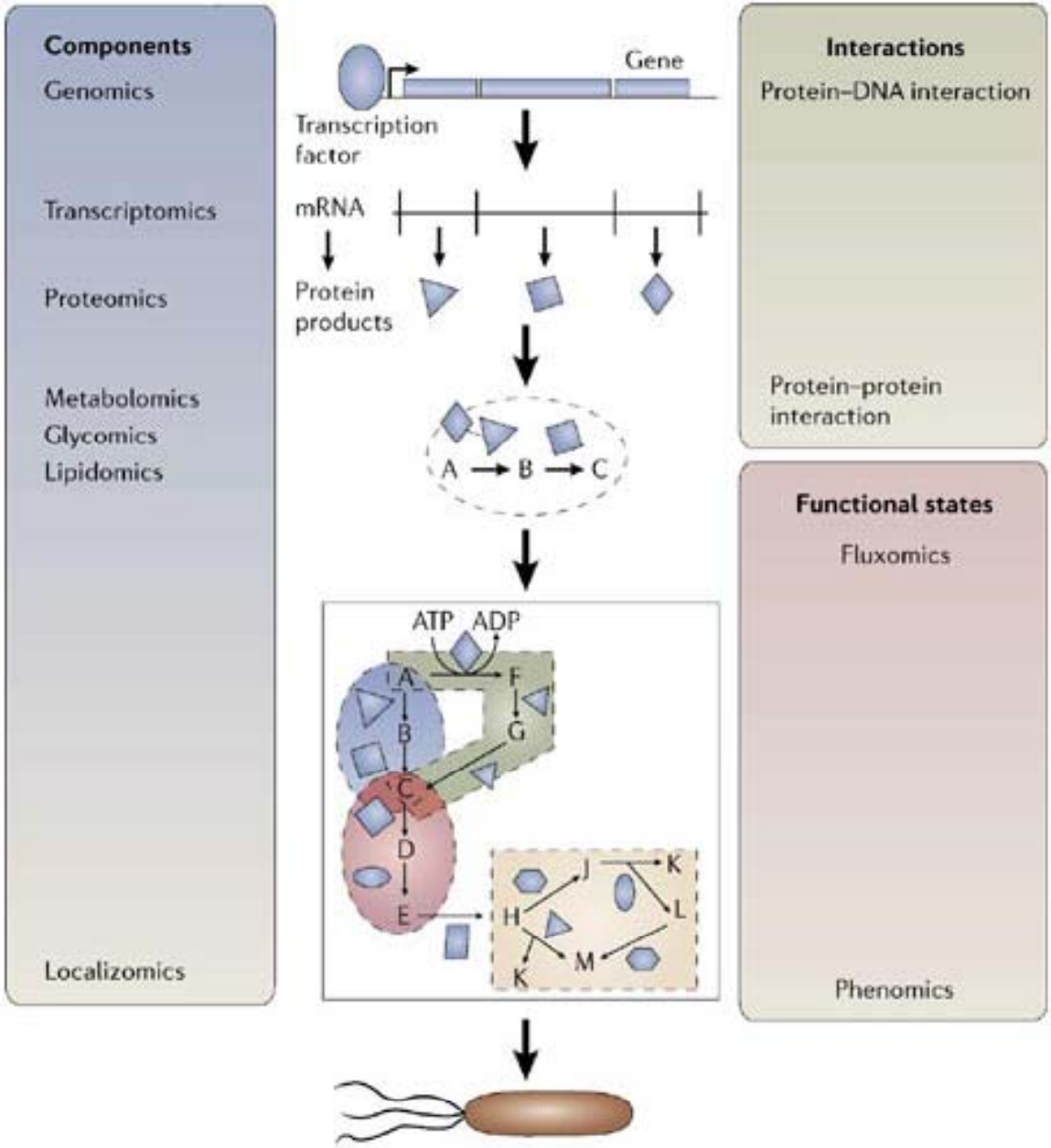

**Protein Data Bank (PDB, RCSB, USA)**

# Human Genome Project




YGA 98-1453

The cell or system




Copyright © 2006 Nature Publishing Group
Nature Reviews | Molecular Cell Biology

Links between specific molecular components

An integrated readout of all omics data types by revealing the overall cellular phenotype

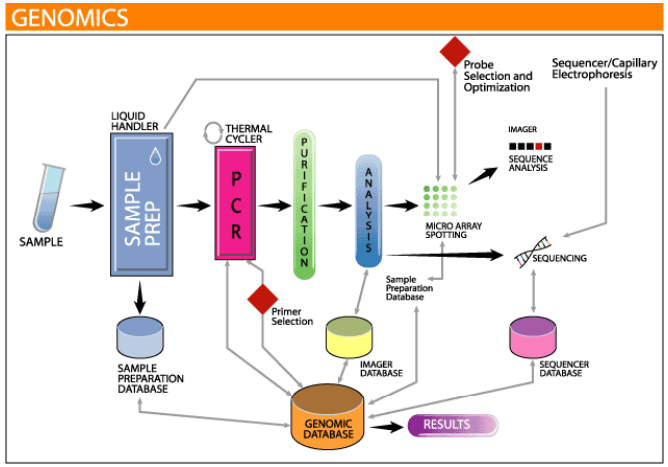
| Genomics | Transcriptomics | Proteomics | Metabolomics | Protein–DNA interactions | Protein–protein interactions | Fluxomics | Phenomics |
|---|---|---|---|---|---|---|---|
| Genomics (sequence annotation) | • ORF validation<br>• Regulatory element identification[74] | • SNP effect on protein activity or abundance | • Enzyme annotation | • Binding-site identification[75] | • Functional annotation[79] | • Functional annotation | • Functional annotation[71,103]<br>• Biomarkers[125] |
| | Transcriptomics (microarray, SAGE) | • Protein: transcript correlation[20] | • Enzyme annotation[109] | • Gene-regulatory networks[76] | • Functional annotation[99]<br>• Protein complex identification[82] | | • Functional annotation[102] |
| | | Proteomics (abundance, post-translational modification) | • Enzyme annotation[95] | • Regulatory complex identification | • Differential complex formation | • Enzyme capacity | • Functional annotation |
| | | | Metabolomics (metabolite abundance) | • Metabolic-transcriptional response | | • Metabolic pathway bottlenecks | • Metabolic flexibility<br>• Metabolic engineering[109] |
| | | | | Protein–DNA interactions (ChIP–chip) | • Signalling cascades[86,102] | | • Dynamic network responses[84] |
| | | | | | Protein–protein interactions (yeast 2H, coAP–MS) | | • Pathway identification activity[89] |
| | | | | | | Fluxomics (isotopic tracing) | • Metabolic engineering |
| | | | | | | | Phenomics (phenotype arrays, RNAi screens, synthetic lethals) |

Pairwise integration of omics data

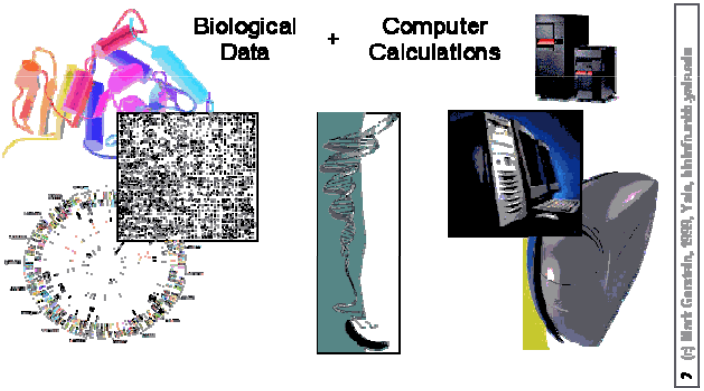
Copyright © 2006 Nature Publishing Group
Nature Reviews | Molecular Cell Biology

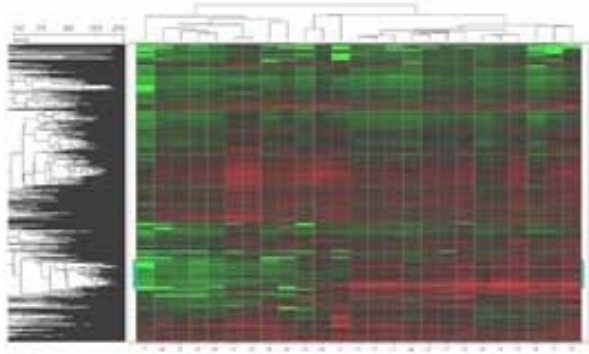
# Consequences of the Human Genome Project (HGP) (1)

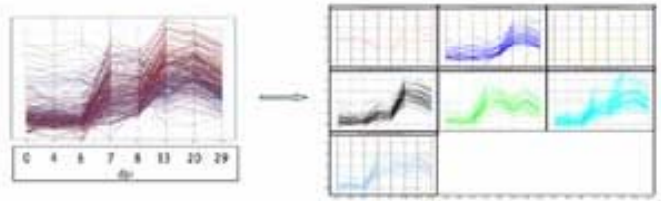
- Complete sequencing of the Human Genome
- New branch of science and medicine
  - Genomics
  - Bioinformatics
  - Transcriptomics




**Bioinformatics**

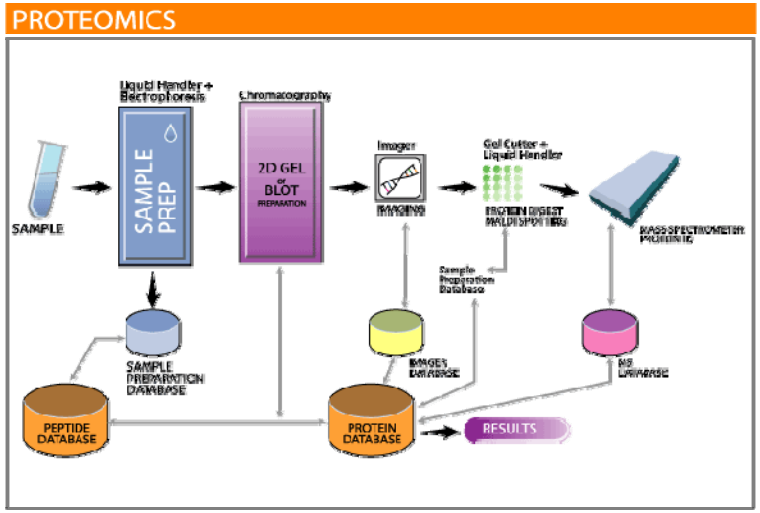


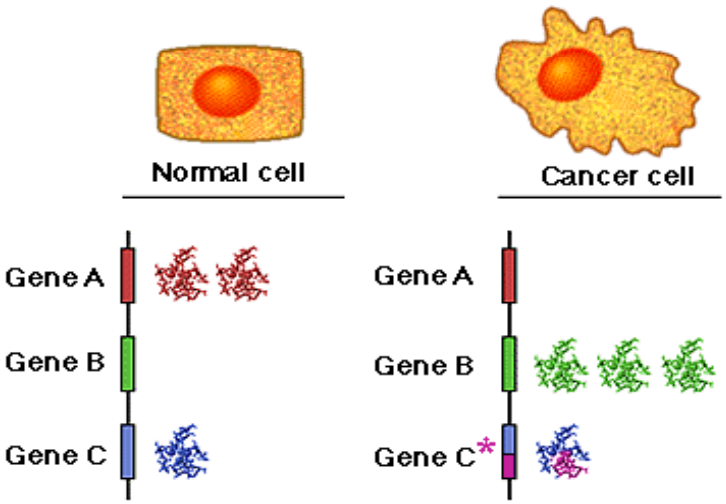



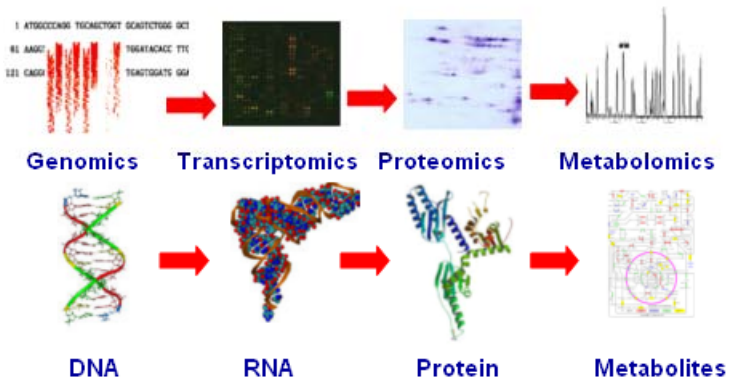


# Consequences of the Human Genome Project (HGP) (2)

- New branch of science and medicine
  - Proteomics
  - Cellomics
  - Metabolomics

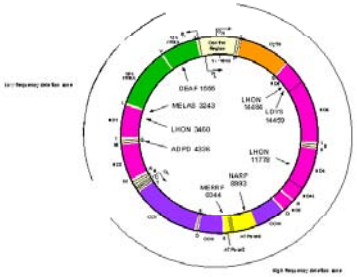


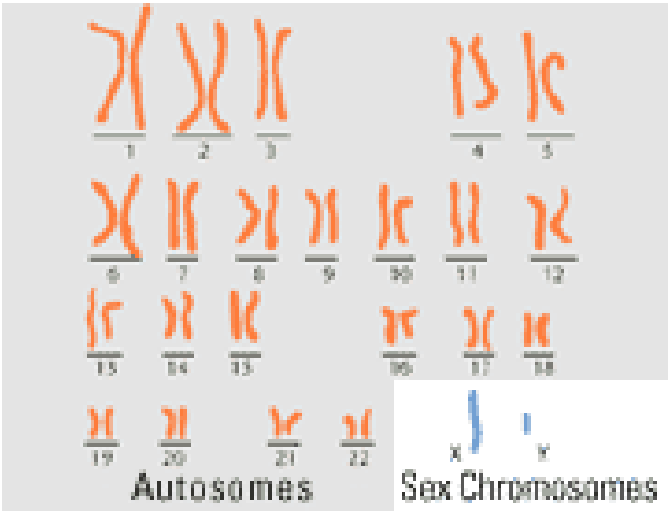



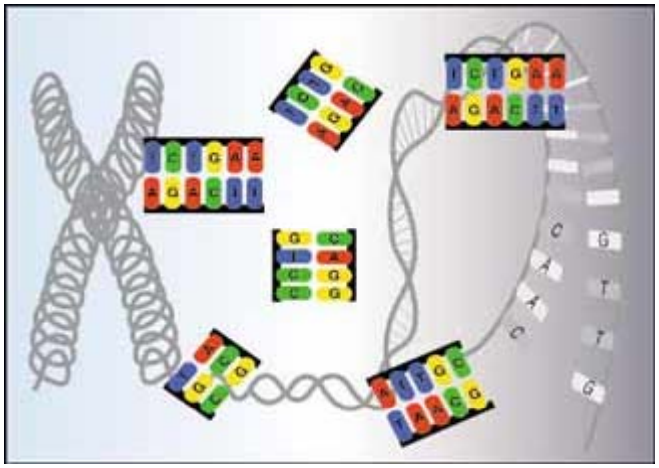



# What is a Genome

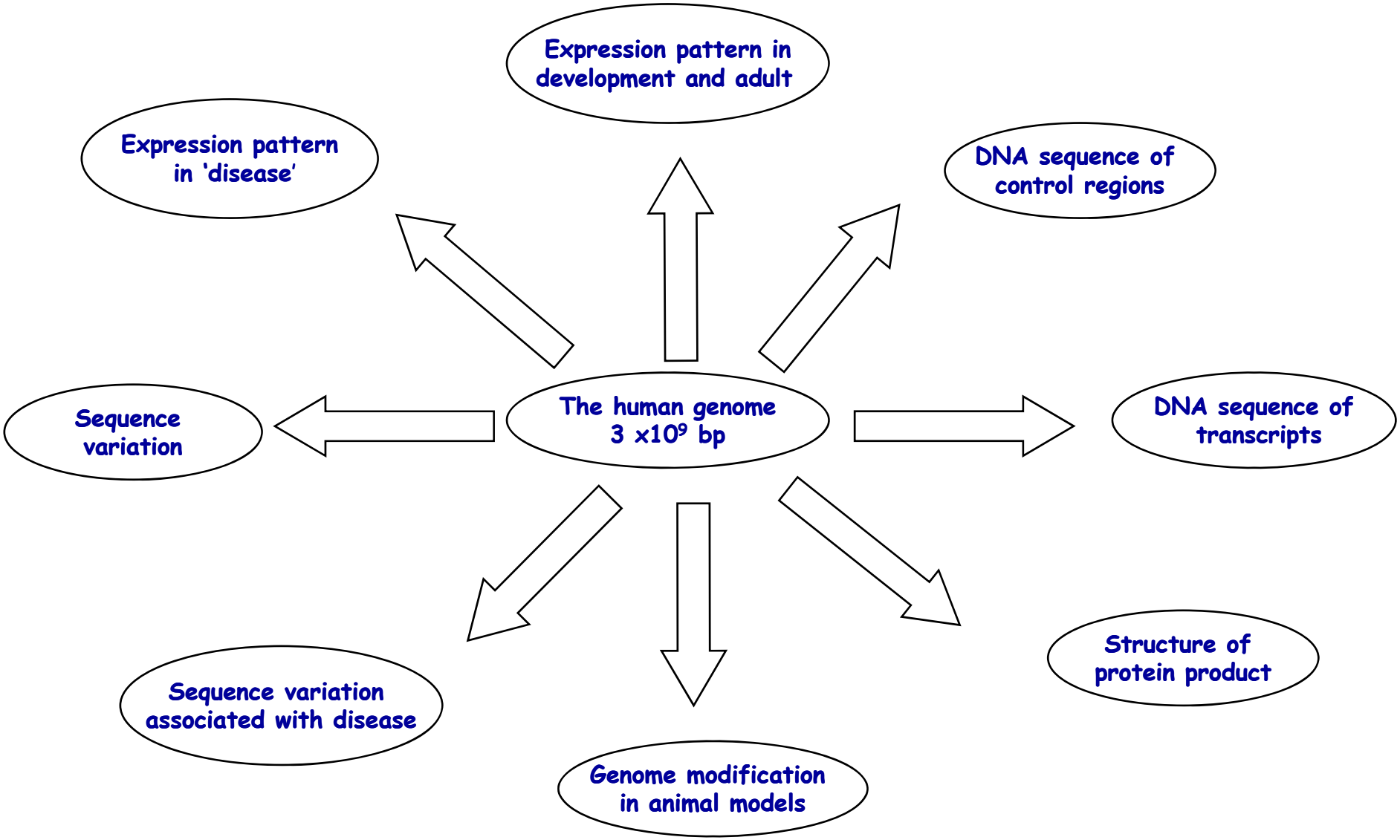
- All of the DNA for an organism
  - One copy
- Human genome
  - N = 22 +XY
  - Nucleus
    - 3.2 billion base pairs packaged into chromosomes
  - Mitochondrion
    - 16.5 Kb packaged into one circular chromosome





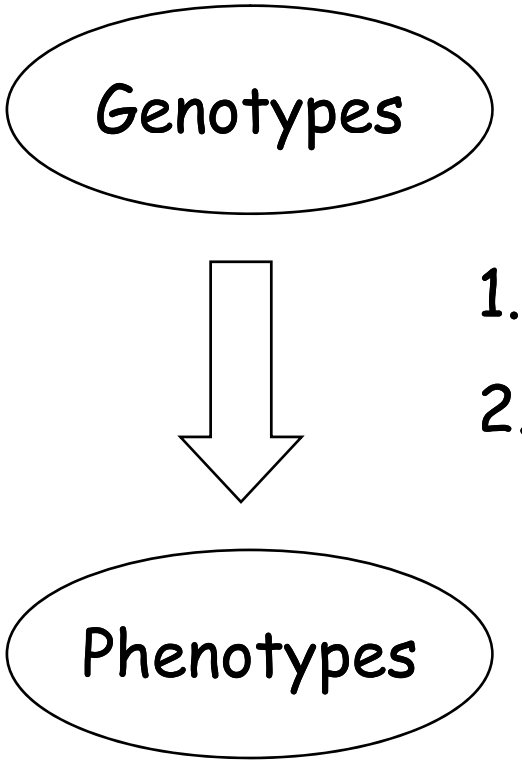



$3.2 \times 10^9$ base pairs


Expression pattern in development and adult
Expression pattern in 'disease'
DNA sequence of control regions
Sequence variation
The human genome 3 x10⁹ bp
DNA sequence of transcripts
Sequence variation associated with disease
Genome modification in animal models
Structure of protein product

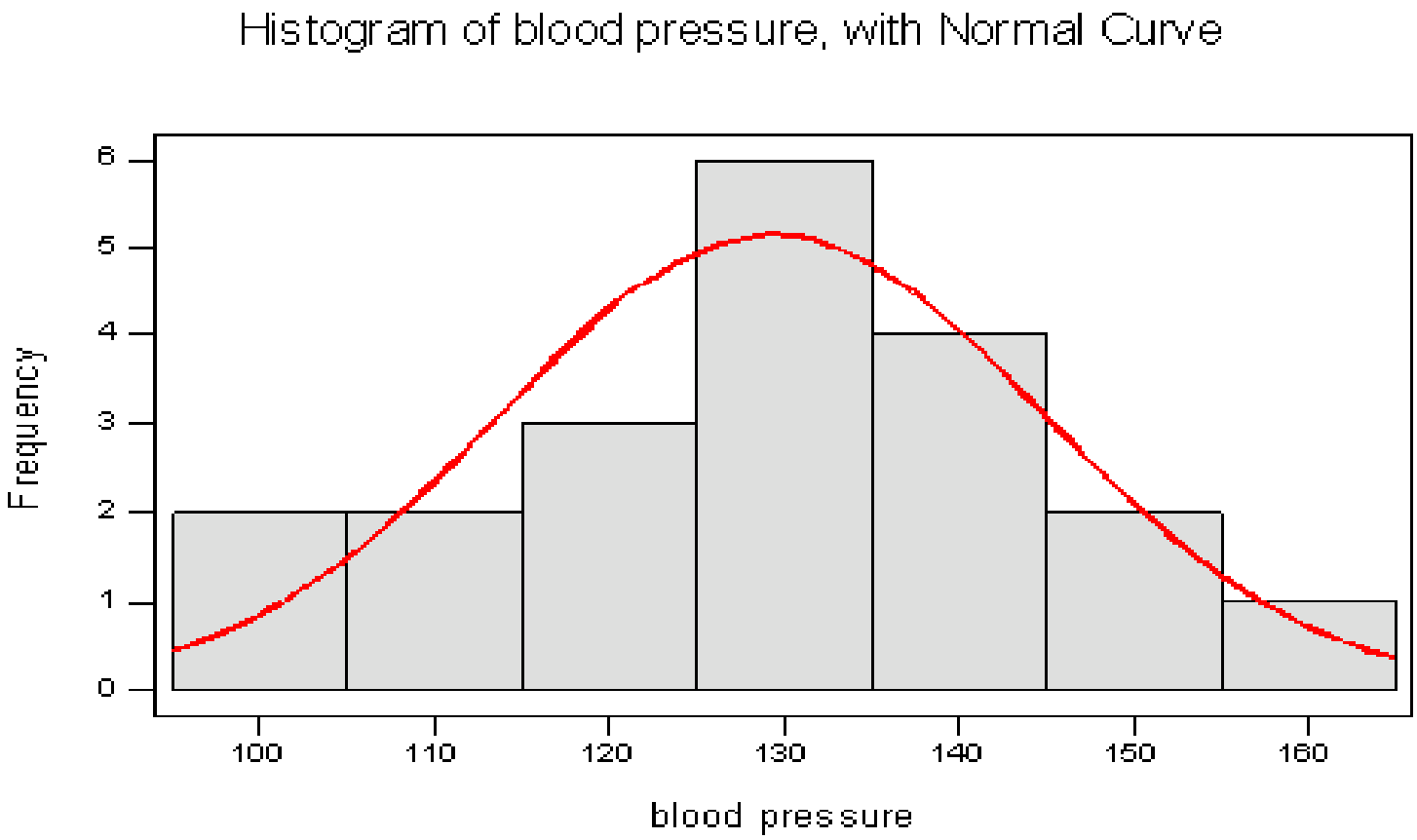
# Why Genome Projects?

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

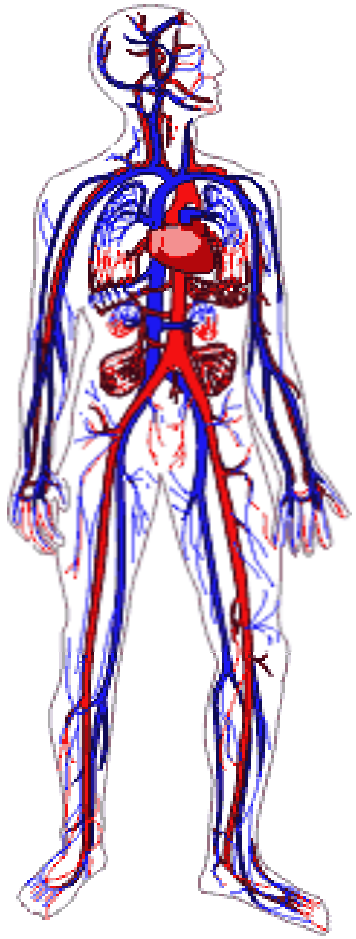



1. Environments
2. Interactions and regulations among genes

# Normal Distribution in Phenotype of Common Complex Disease



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



Stroke

Breast cancer

Diabetes

Obesity

Inflammatory Bowel Disease

Manic-depression

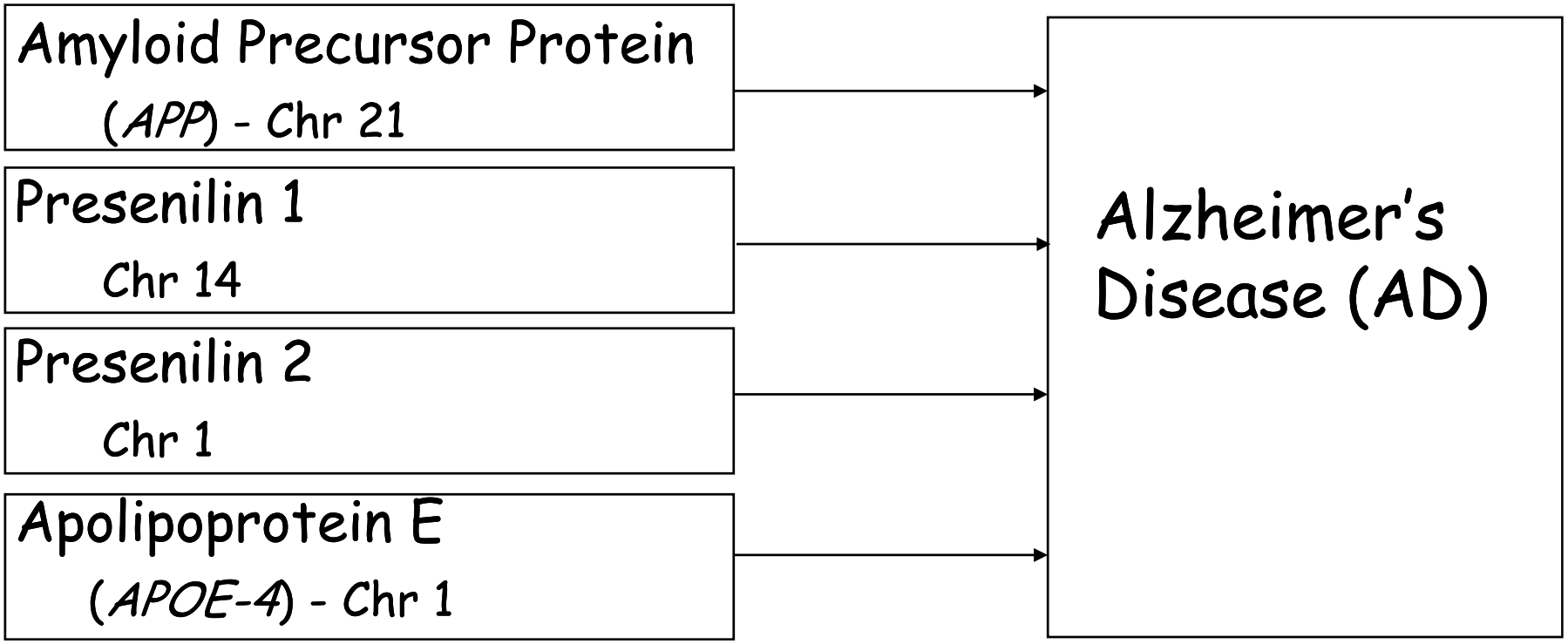
Myocardial Infarction

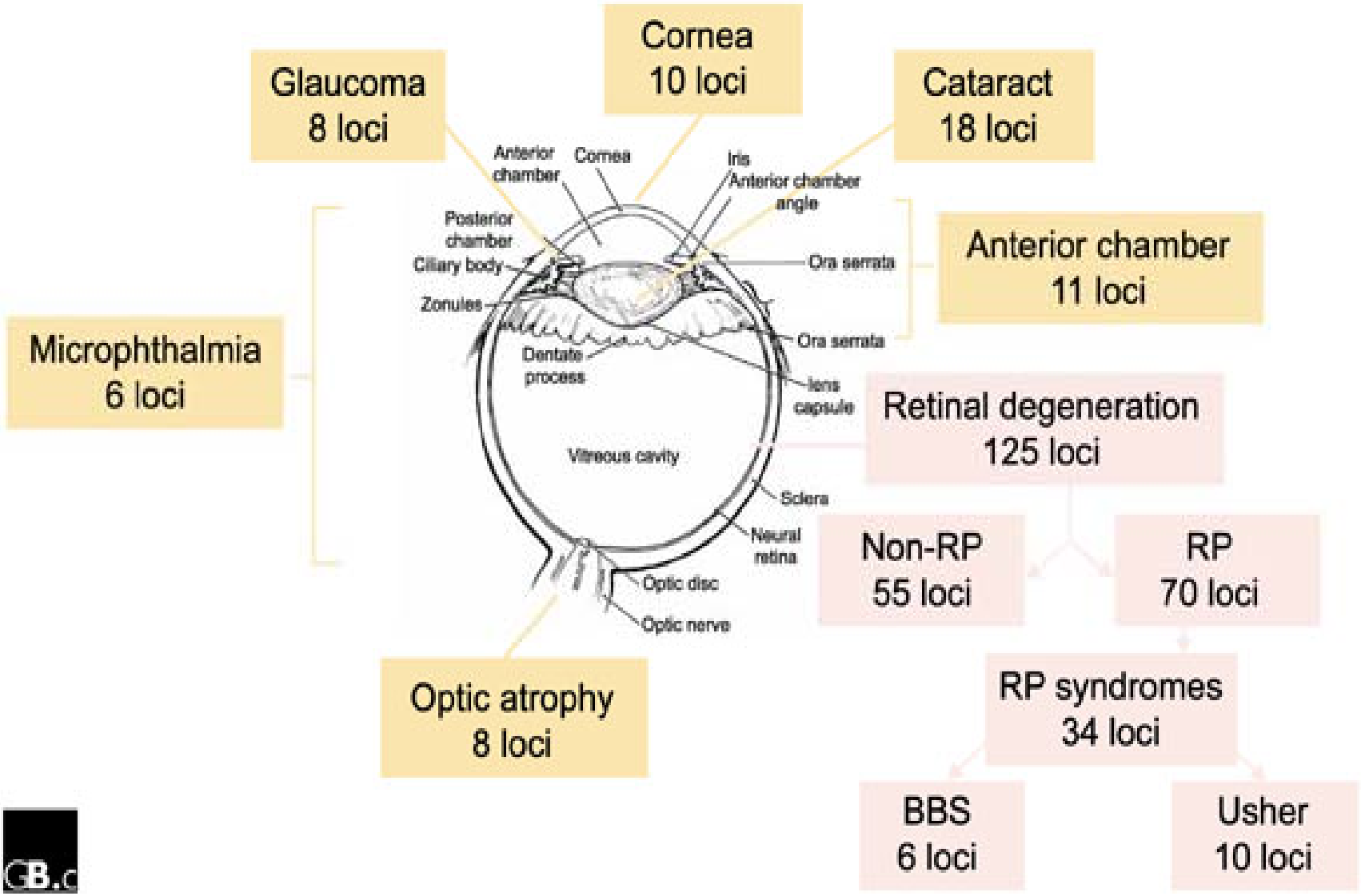
Hypertension

High Cholesterol

Schizophrenia

# Locus Heterogeneity in Alzheimer's Disease

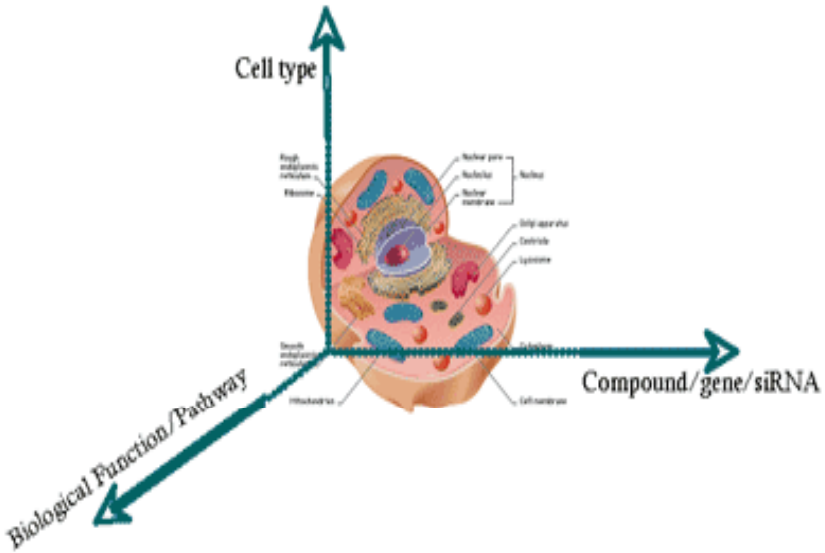




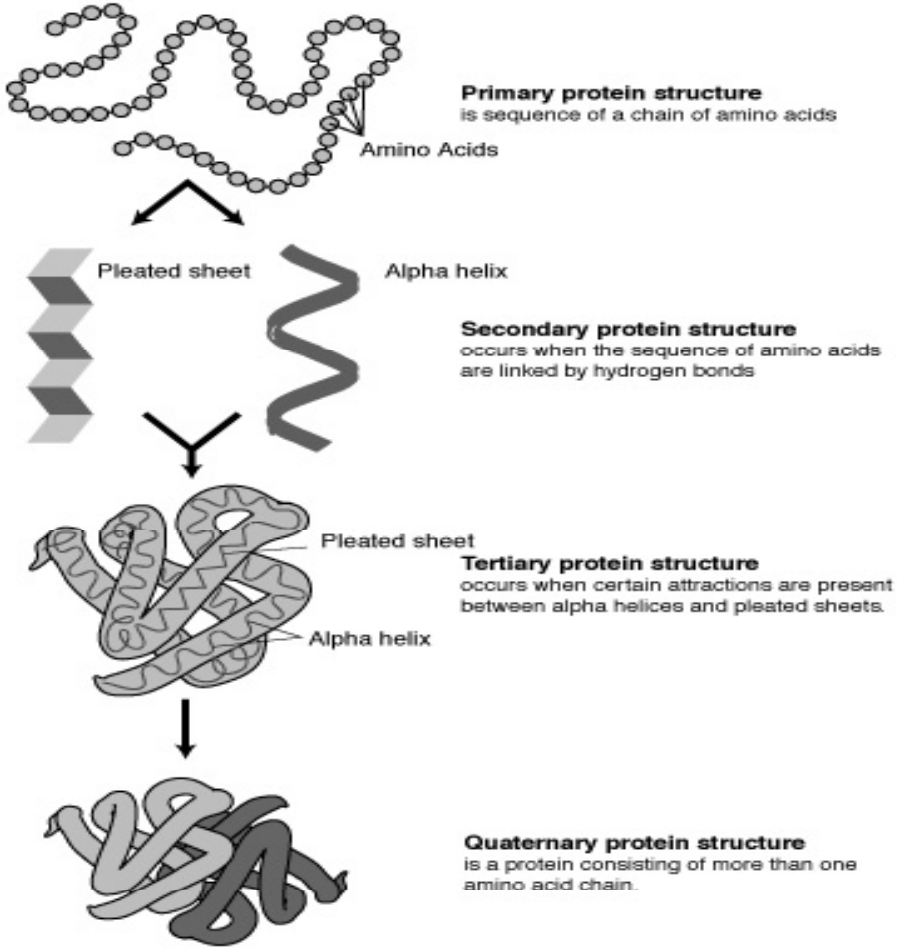

Wright *et al. Genome Biology* 2001
2:comment2007.1 doi:10.1186/gb-2001-2-8-comment2007

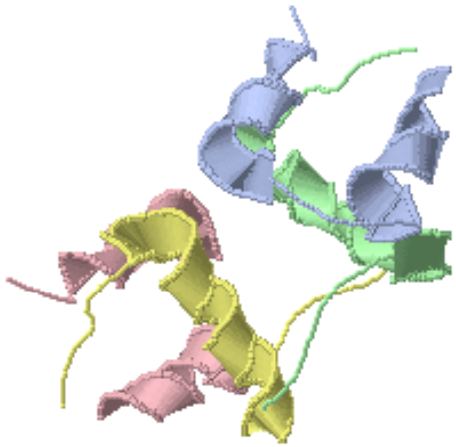
# Genomic Biology

- 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**)

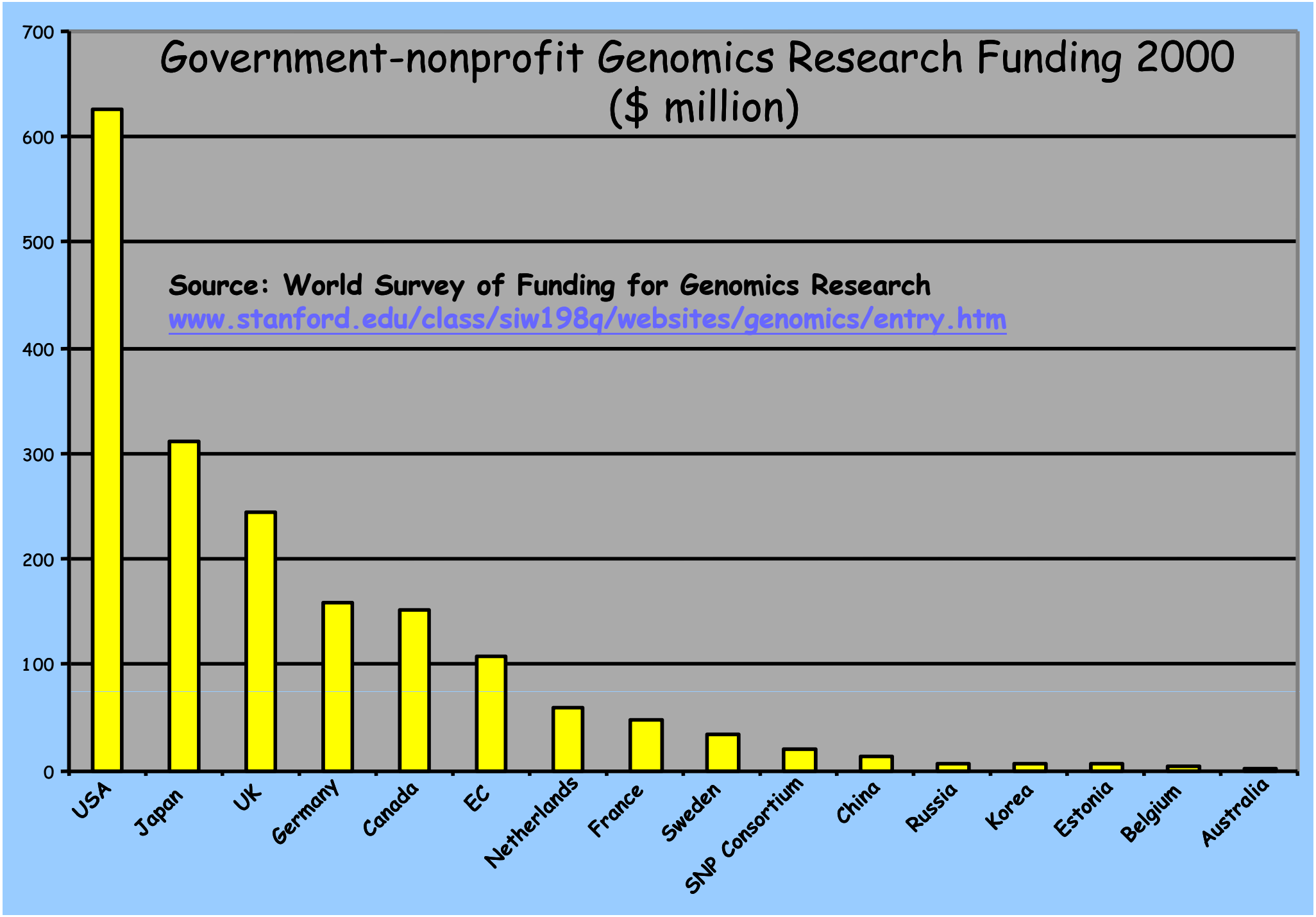


# The Central Paradigm of Molecular Biology

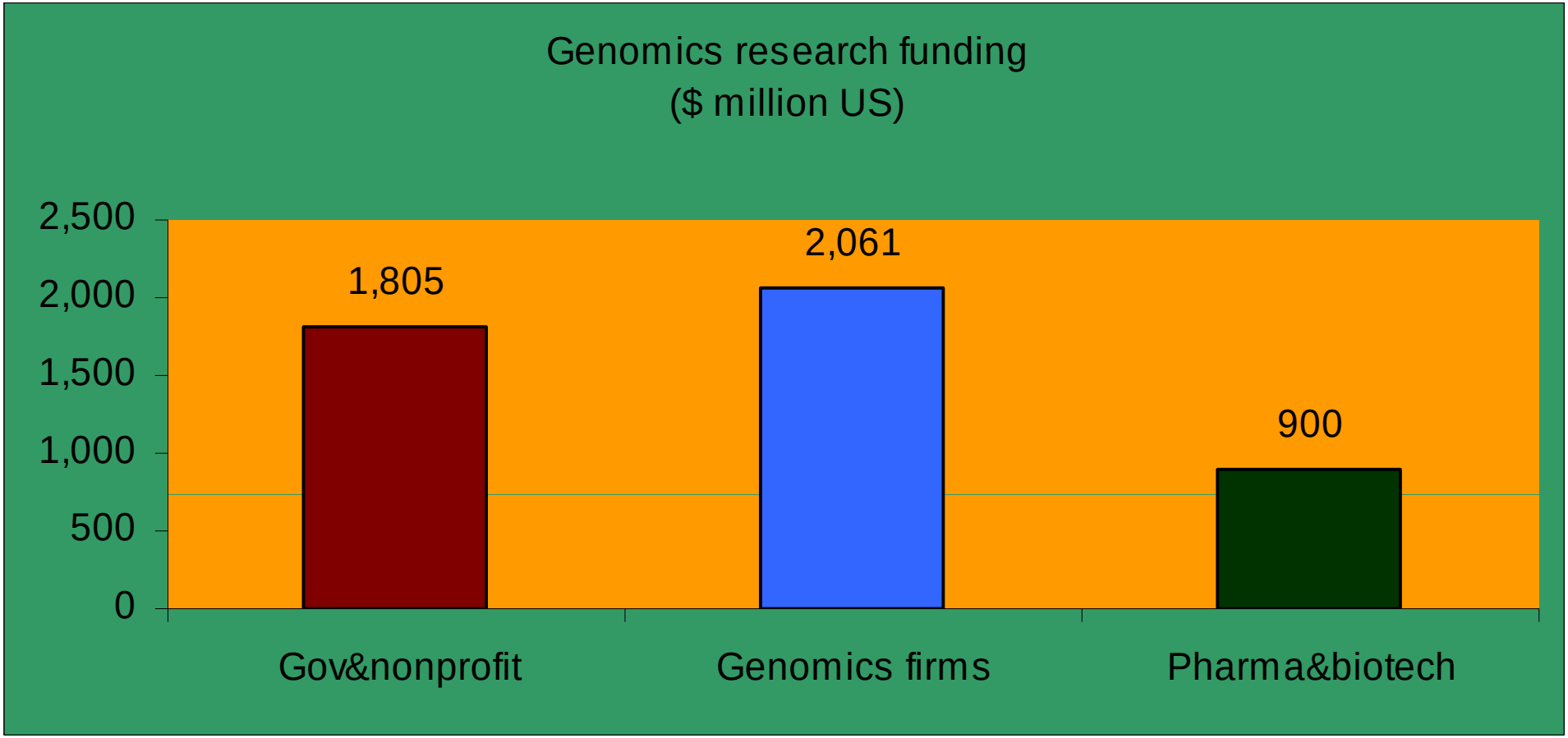


Protein Databank

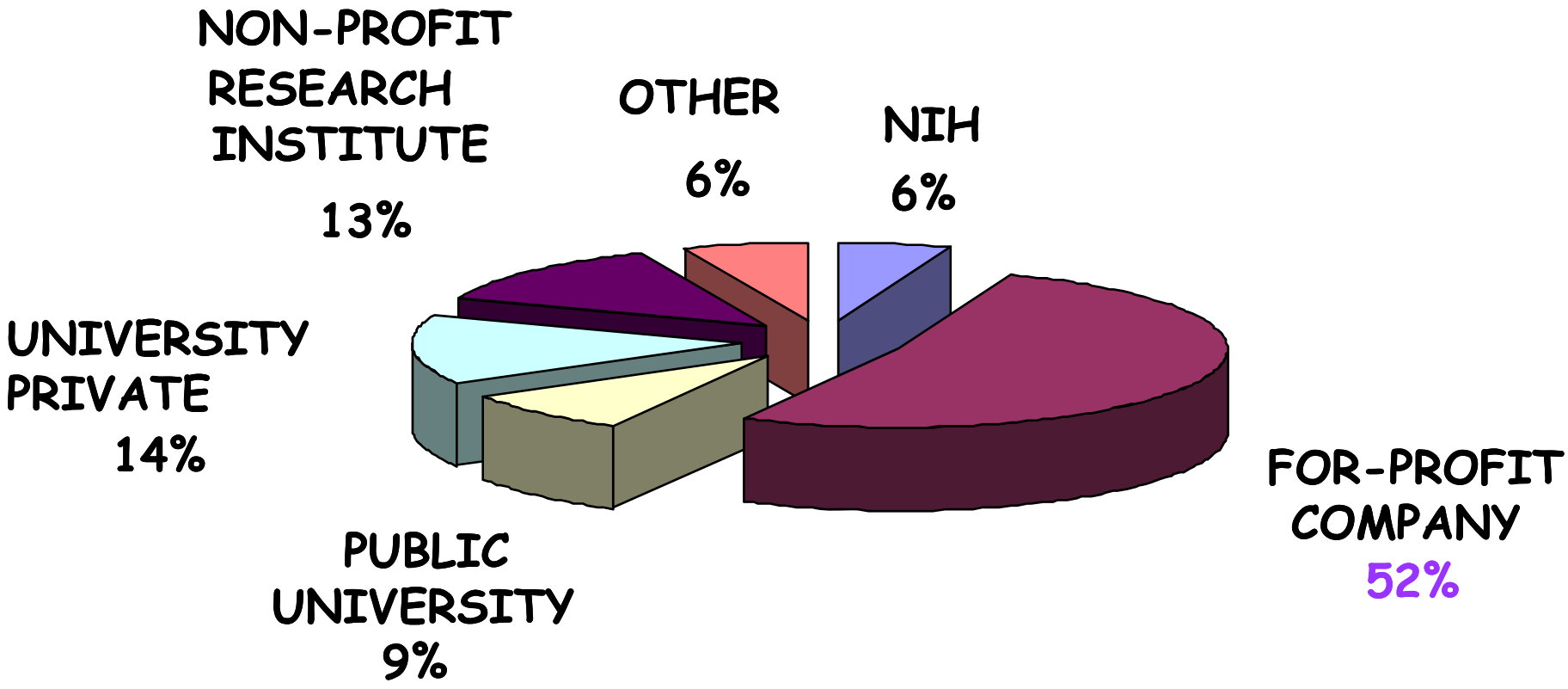

Government-nonprofit Genomics Research Funding 2000
($ million)
Source: World Survey of Funding for Genomics Research
www.stanford.edu/class/siw198q/websites/genomics/entry.htm
700
600
500
400
300
200
100
0
USA
Japan
UK
Germany
Canada
EC
Netherlands
France
Sweden
SNP Consortium
China
Russia
Korea
Estonia
Belgium
Australia

# Funding: Private > Public (2000)



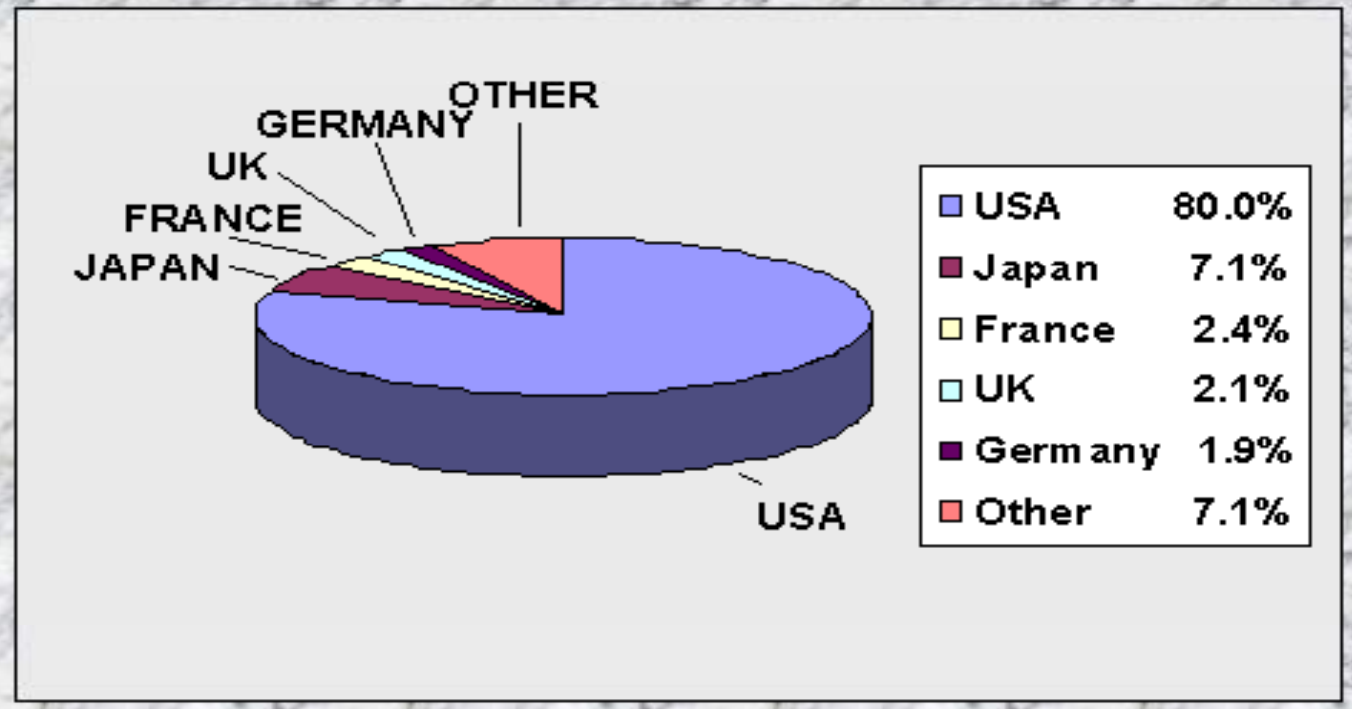

**Source**: World Survey of Funding for Genomics Research
Stanford in Washington Program
http://www.stanford.edu/class/siw198q/websites/genomics/entry.htm

# Patent Assigned

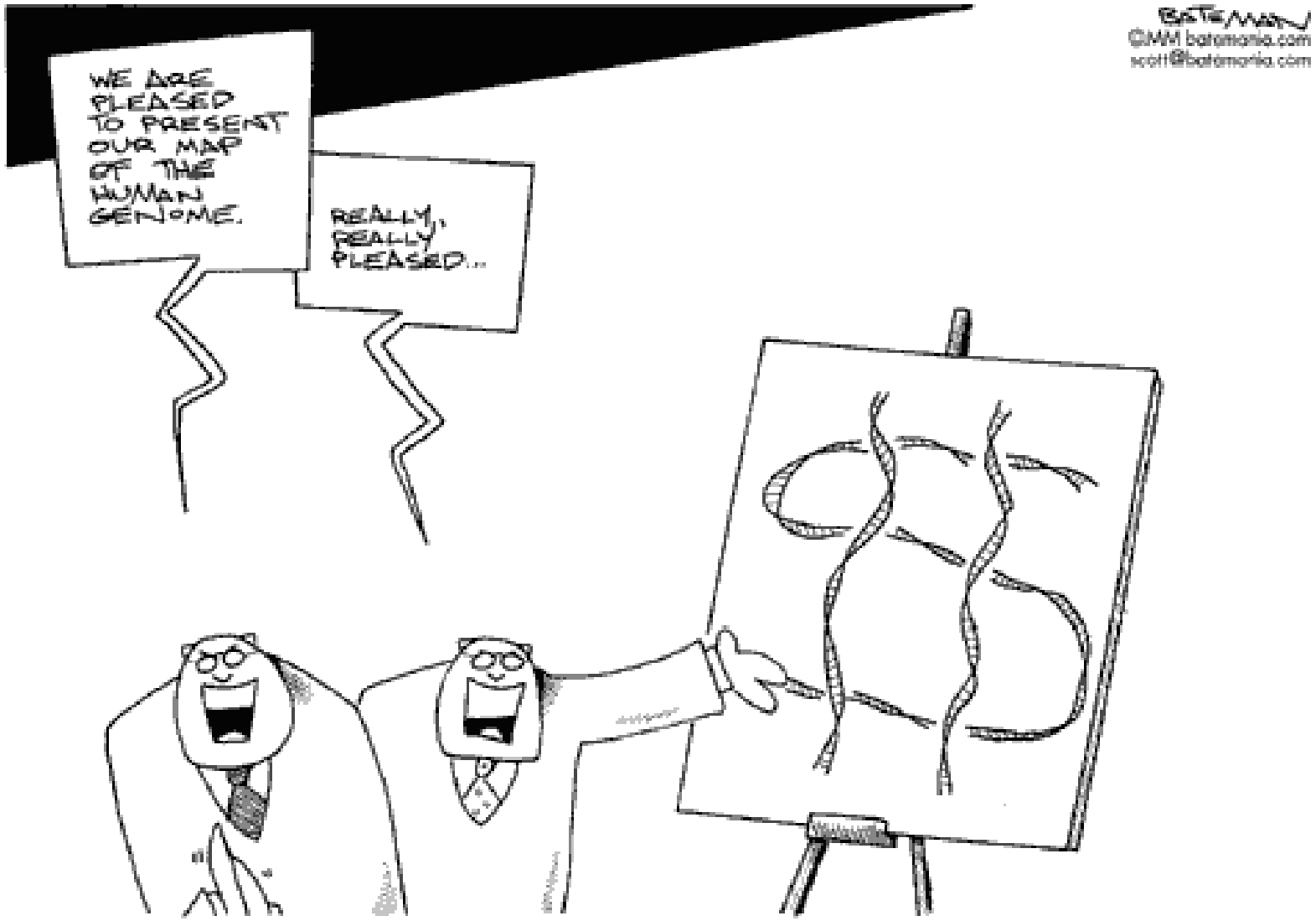



Source: Stephen McCormack and Robert Cook-Deegan
DNA Patent Database www.genomic.org

# 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


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

# Rational Human/Plant/Animal Improvements

**Functions** of all genes

+

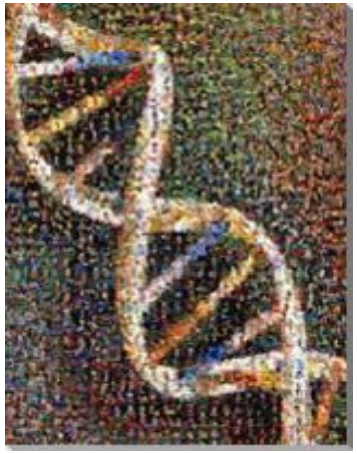
**Tools** for modifying genome

↓

Development a robust Genetic engineering discipline

# Goals of the Human Genome Project (HGP)

- *identify* all the approximately **20,000-25,000 genes** in human DNA
- *determine* the sequences of the **3 billion chemical base pairs** that make up human DNA
- *store* this information in **databases**
- *improve* tools for **data analysis**
- *transfer* related technologies **to the private sector**, and
- *address* the ethical, legal, and social issues (ELSI) that may arise from the project.



# 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)

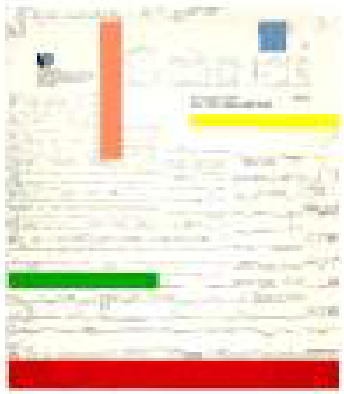
- Feb. 2001
  - June 2002 (**TIGR**): 7,801 genes' functions identified
  - International Human Genome Sequencing Consortium: http://www.nature.com (Nature)
  - The Celera database: http://www.sciencemag.org (Science)

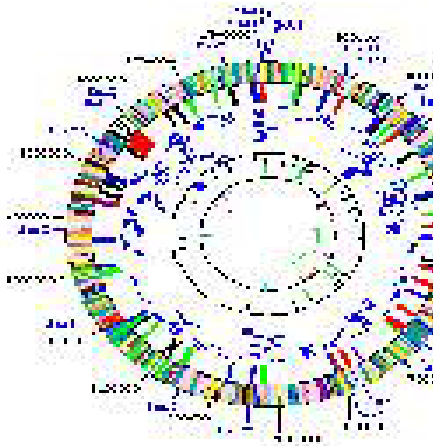

LET'S JUST BE FRIENDS
IT'S MY DNA... ISN'T IT?
HUMAN GENOME REPORT


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


IT'S A BOY! ...
... SMOKER, HOMOSEXUAL, AND INSURANCE AGENT, WITH A 75% ALZHEIMER-PROBABILITY!
GENOME
CHAPATTE

# 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.

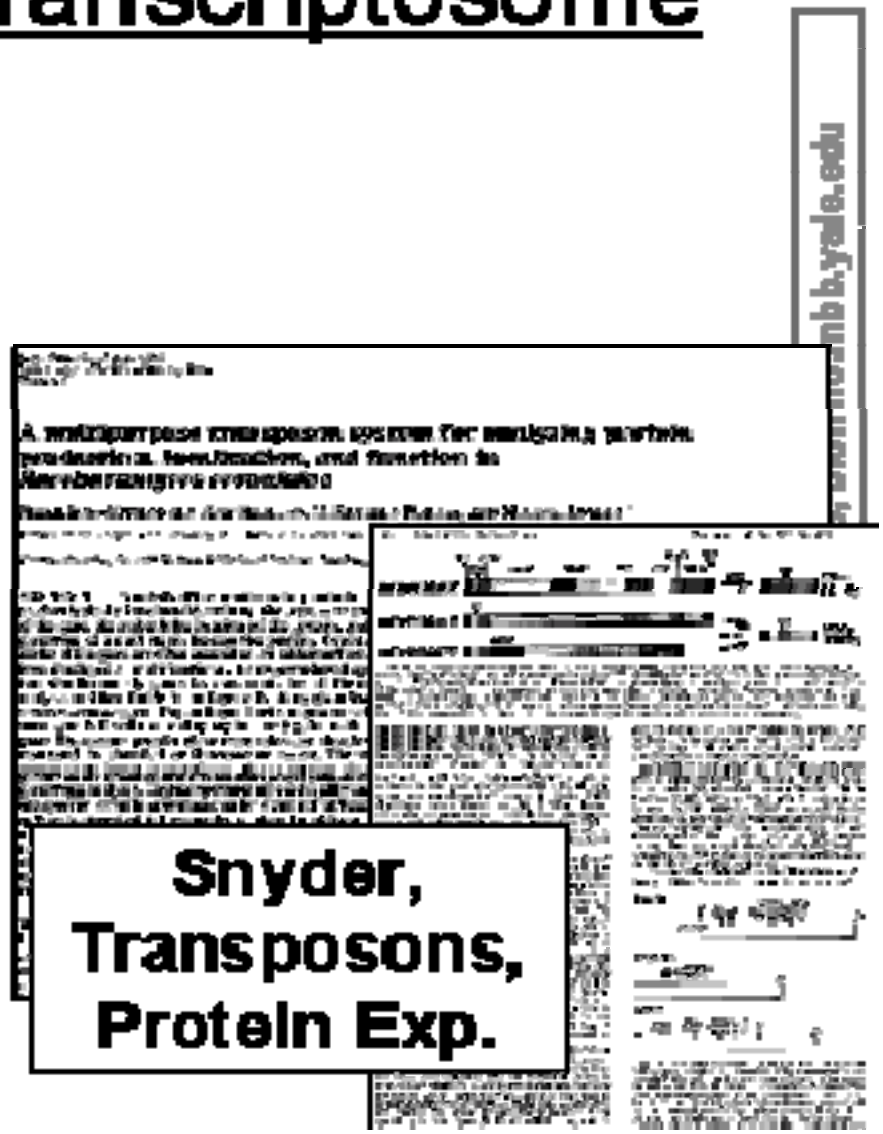
– G A Pekso, *Nature* **401**: 115-116 (1999)

# Gene Expression Datasets: the Transcriptosome


**Young/Lander, Chips, Abs. Exp.**

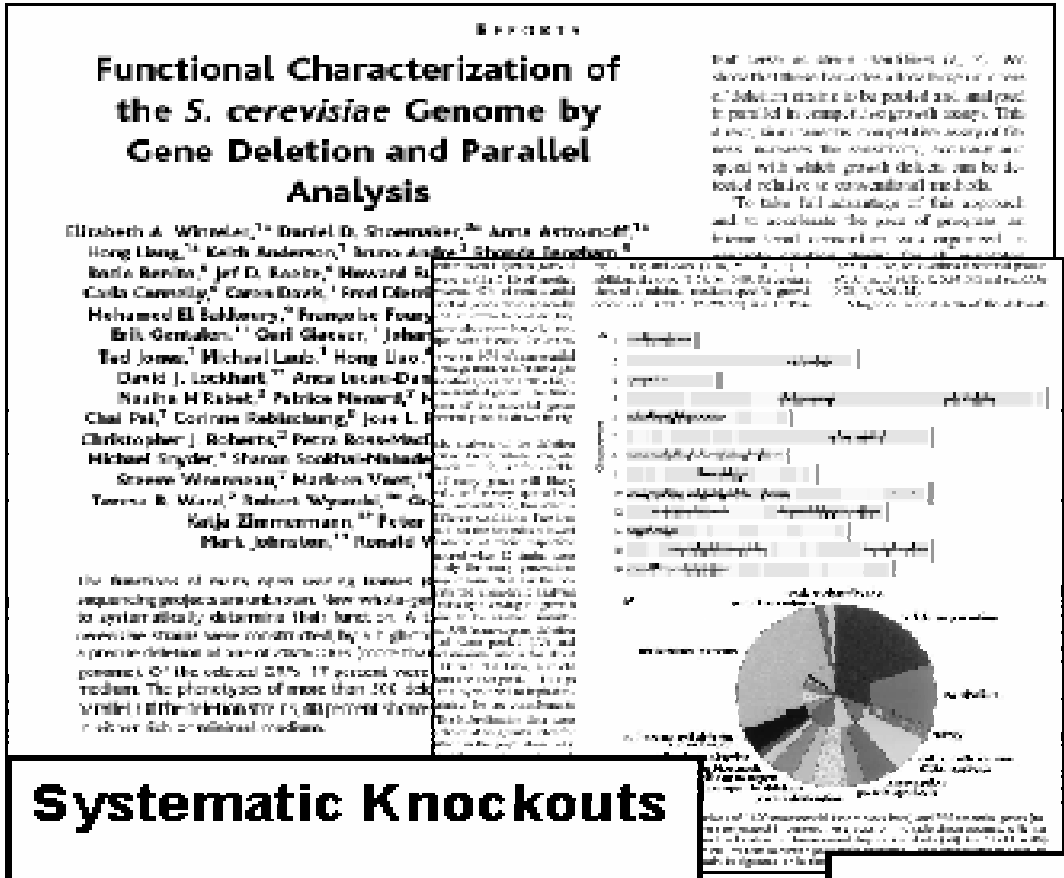

**Brown, µarray, Rel. Exp. over Timecourse**

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



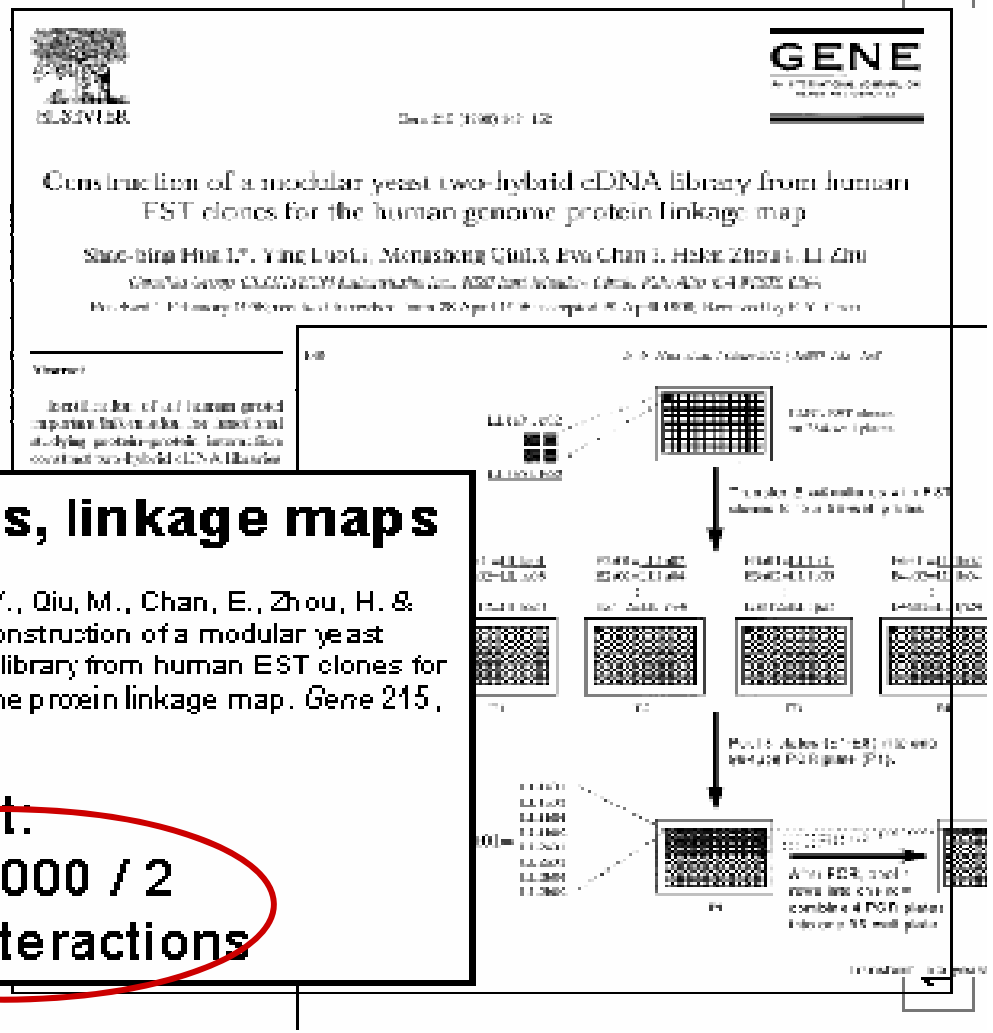
**Snyder, Transposons, Protein Exp.**

# Other Whole-Genome Experiments




## 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




## 2 hybrids, linkage maps

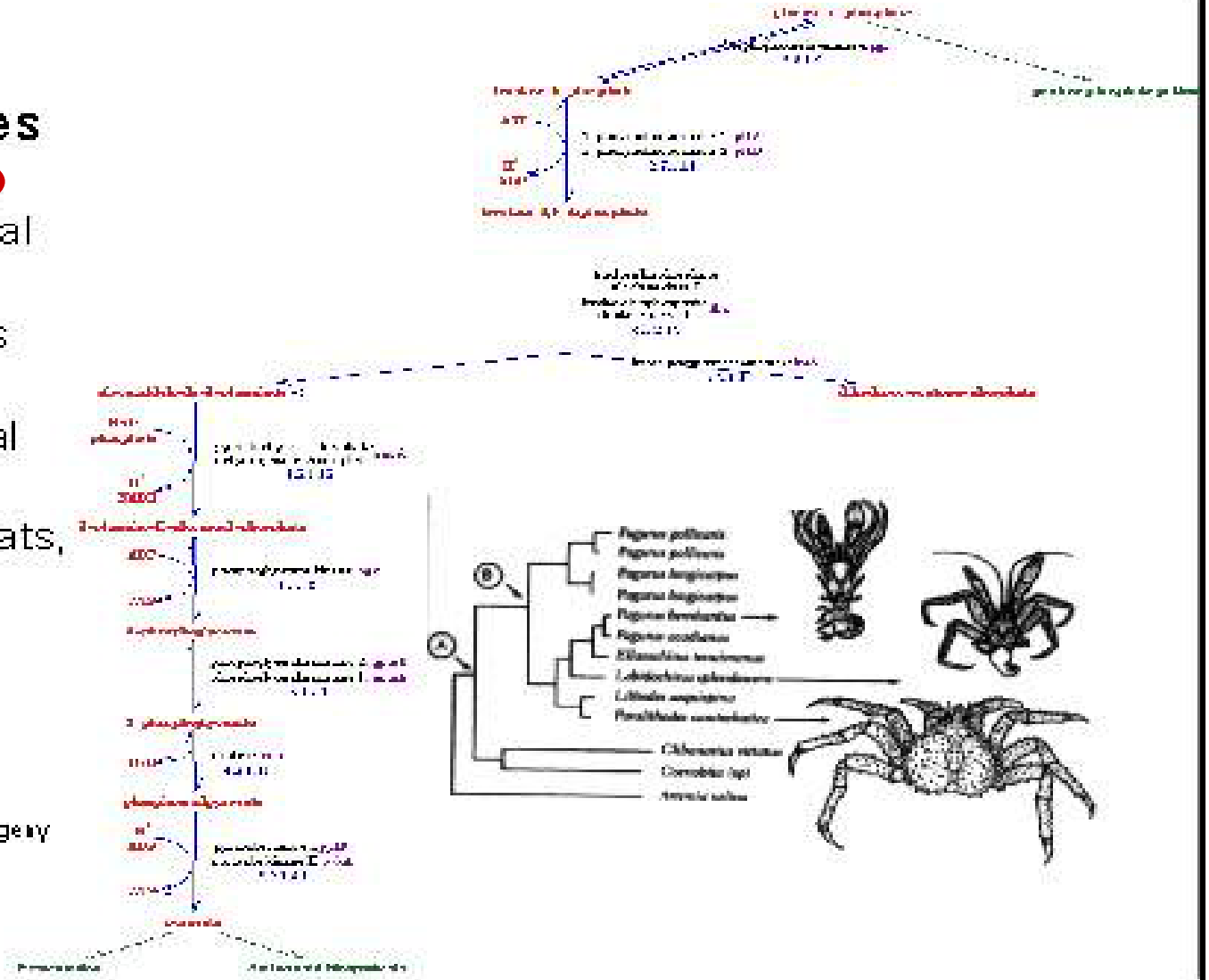
Hua, S. B., Luo, Y., Qiu, M., Chan, E., Zhou, H. & Zhu, L. (1998). Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map. *Gene* 215, 143-52

**For yeast:**
**6000 x 6000 / 2**
**~ 18M interactions**

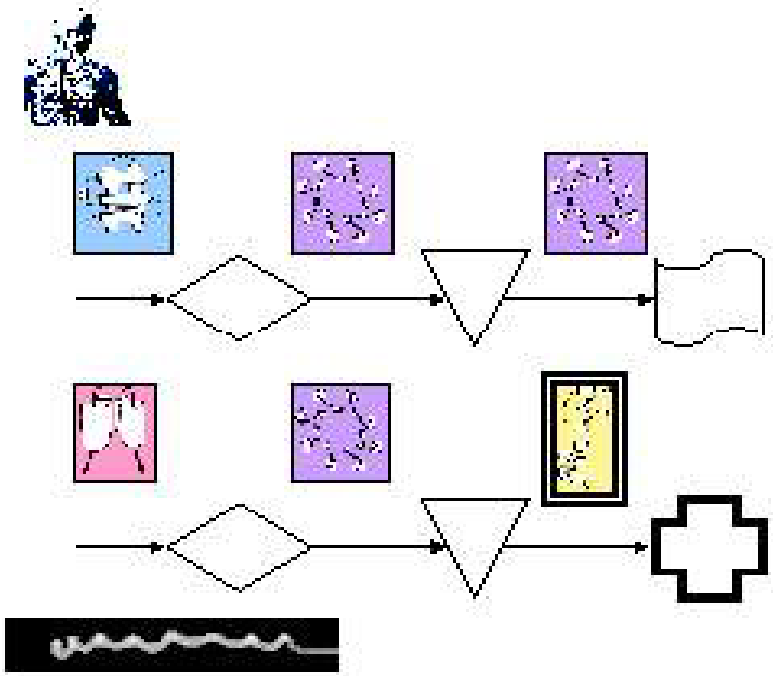
# Molecular Biology Information: Other Integrative Data

- **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



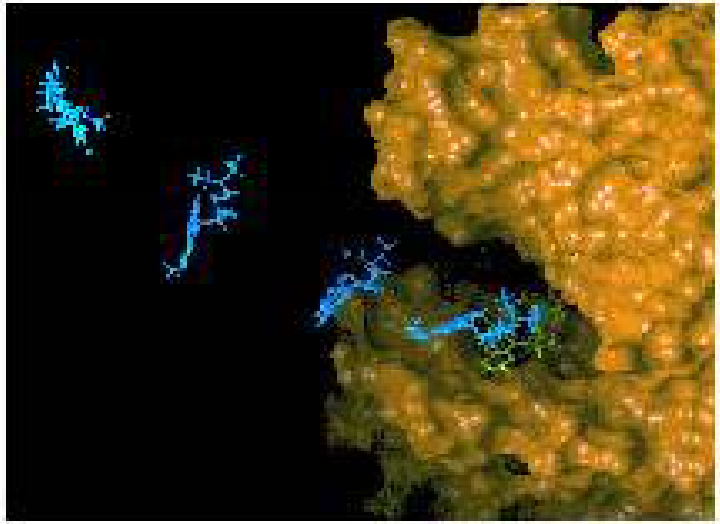
- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions Pleiotrophic
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code
- **How do we find the similarities?**

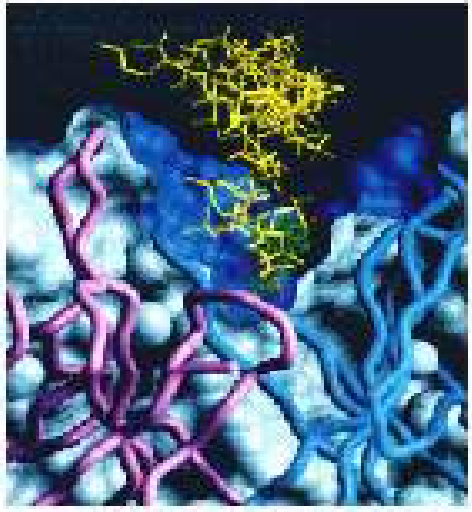
**Integrative** Genomics -
genes ↔ structures ↔
**functions** ↔ **pathways** ↔
expression levels ↔
regulatory systems ↔ ....

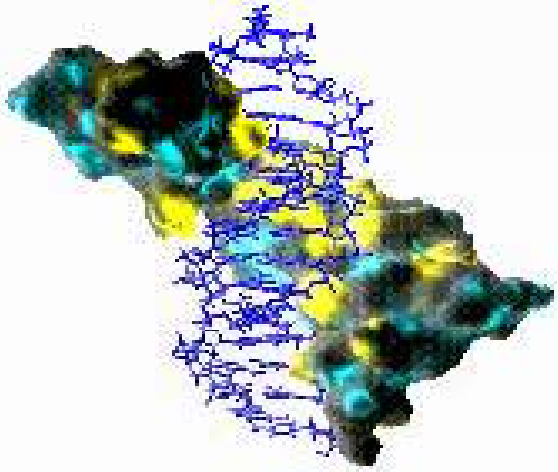
# Major Application I: Designing Drugs

- 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!

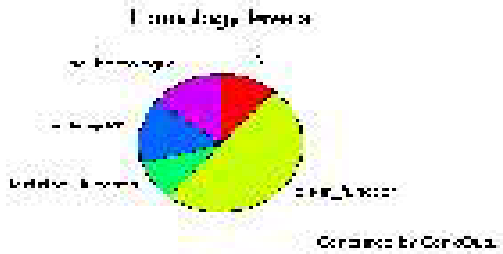
(Section from NCBI Disease Genes Database Reproduced Below)

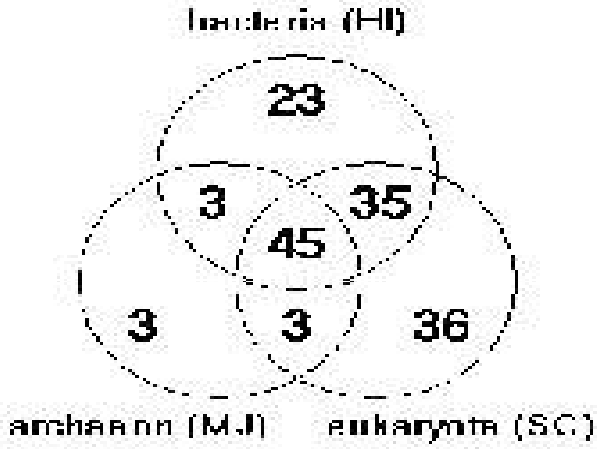
Best Sequence Similarity Matches to Data Between Positionally Cloned Human Genes and S. cerevisiae Proteins

| 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 | M84318 | DNA repair protein |
| Hereditary Non-polyposis Colon Cancer | 120436 | MLH1 | U07418 | 6.3e-196 | MLH1 | U07181 | DNA repair protein |
| Cystic Fibrosis | 219700 | CFTR | M28668 | 1.3e-161 | YCF1 | L35271 | Metal resistance protein |
| Wilson Disease | 277900 | WND | U33188 | 5.9e-161 | CCC2 | L36531 | Probable copper transporter |
| Glycerol Kinase Deficiency | 307030 | GK | L33943 | 1.8e-129 | GUT1 | X69949 | Glycerol kinase |
| Bloom Syndrome | 210900 | BLM | U29831 | 2.6e-119 | SGS1 | U22341 | Helicase |
| Adrenoleukodystrophy, X-linked | 300100 | ALD | Z23116 | 9.4e-101 | PXA1 | U11865 | Peroxisomal ABC transporter |
| Ataxia Telangiectasia | 208900 | ATM | U26455 | 2.8e-98 | TEL1 | U53531 | PIK kinase |
| Amyotrophic Lateral Sclerosis | 105400 | SOD1 | X08165 | 2.8e-58 | SOD1 | J03219 | Superoxide dismutase |
| Myotonic Dystrophy | 160900 | DM | L39268 | 5.4e-55 | YPK1 | M21581 | Serine/threonine protein kinase |
| Lowe Syndrome | 309000 | OCRL | M88162 | 3.2e-41 | YIL012C | Z41841 | Putative IPP-5-phosphatase |
| Neurofibromatosis, Type 1 | 162200 | NF1 | M89914 | 2.8e-46 | IRA2 | M33119 | Inhibitory regulator protein |
| Choroideremia | 303100 | CHM | X16323 | 2.1e-42 | GDI1 | S69511 | GDP dissociation inhibitor |
| Diastrophic Dysplasia | 222600 | DTD | U14528 | 1.2e-38 | SUL1 | X12813 | Sulfate permease |
| Lissencephaly | 247200 | LIS1 | L33385 | 3.1e-34 | MET30 | L26585 | Methionine metabolism |
| Thomsen Disease | 160800 | CLCN1 | Z25884 | 1.9e-33 | GEF1 | Z23331 | Voltage-gated chloride channel |
| Wilms Tumor | 194070 | WT1 | X51630 | 3.1e-28 | FZF1 | X61181 | Sulphite resistance protein |
| Achondroplasia | 100800 | FGFR3 | M58053 | 2.8e-28 | IPL1 | U11363 | Serine/threonine protein kinase |
| Menkes Syndrome | 309400 | MNK | X69208 | 2.1e-21 | CCC2 | L36531 | Probable copper transporter |

# Major Application III: Overall Genome Characterization

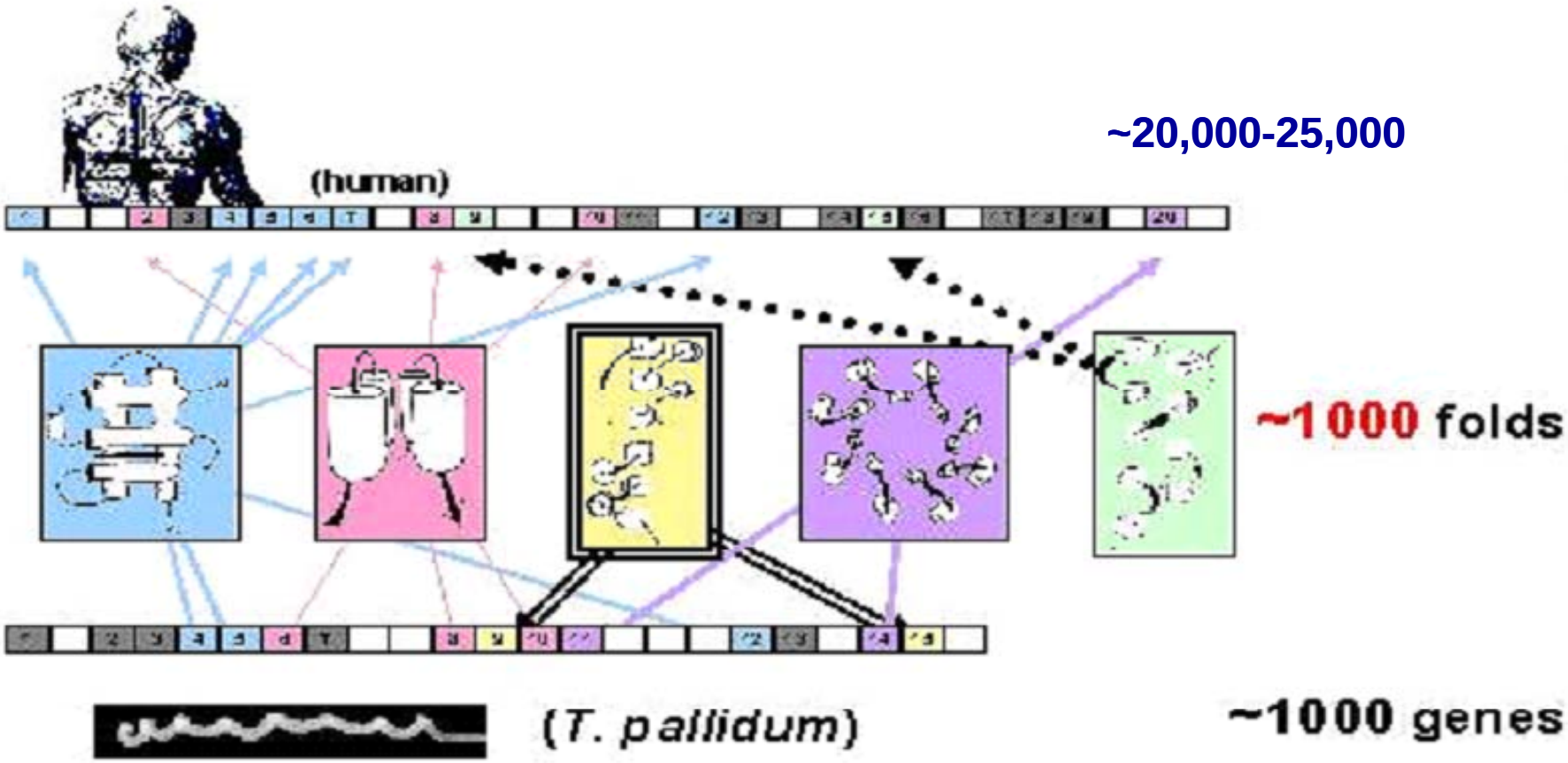
- Overall Occurrence of a Certain Feature in the Genome
  - e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics



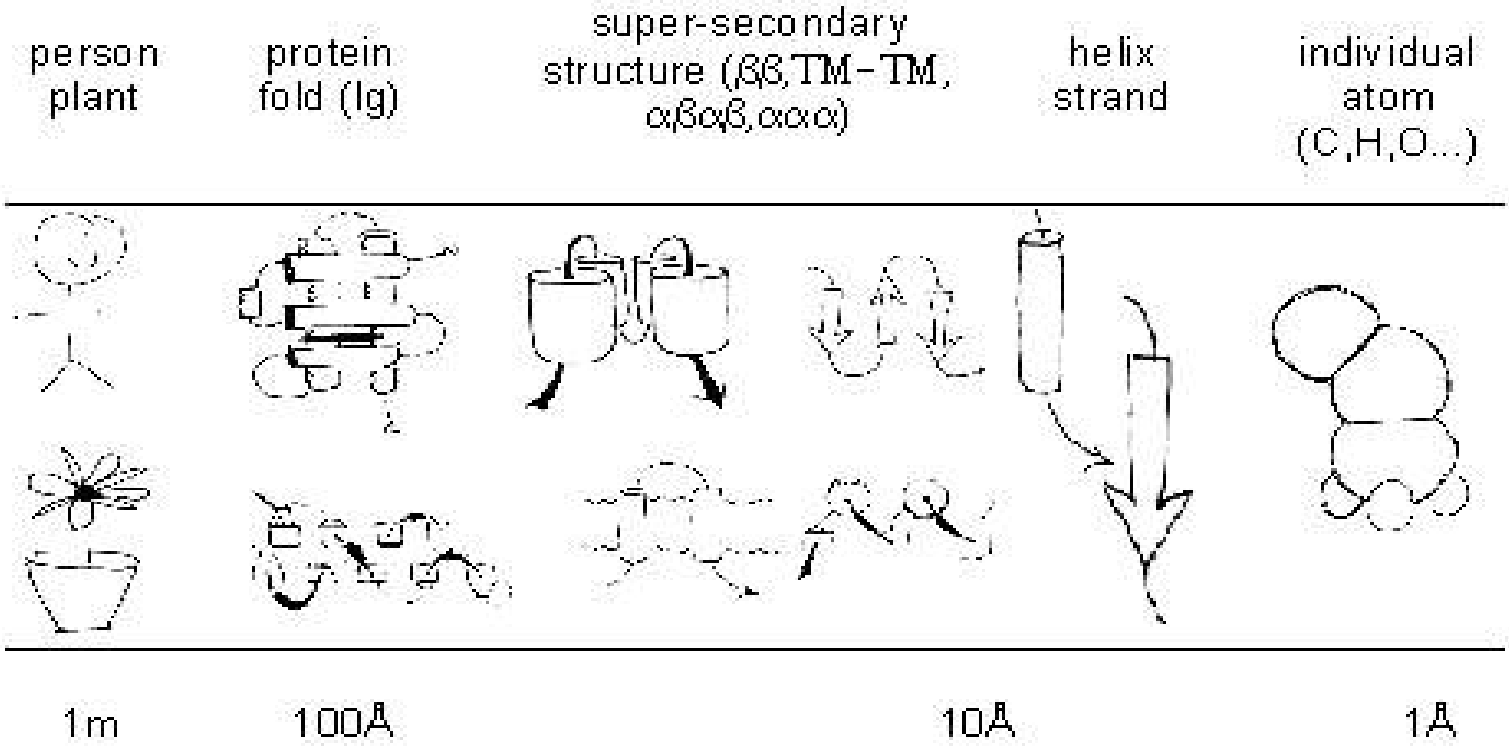


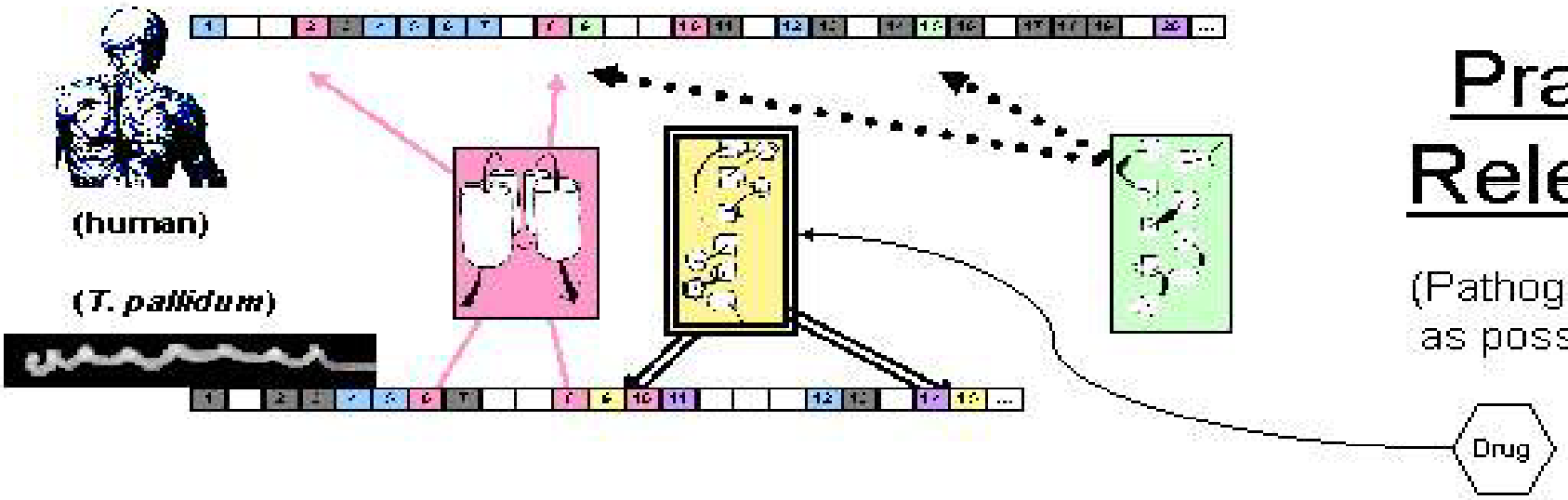

(Clock figures, yeast u. Synechocystis, adapted from GeneOntologyWeb Page, Sander Group, EBI)

# Simplfying Genomes with Folds, Pathways, &c



# At What Structural Resolution Are Organisms Different?



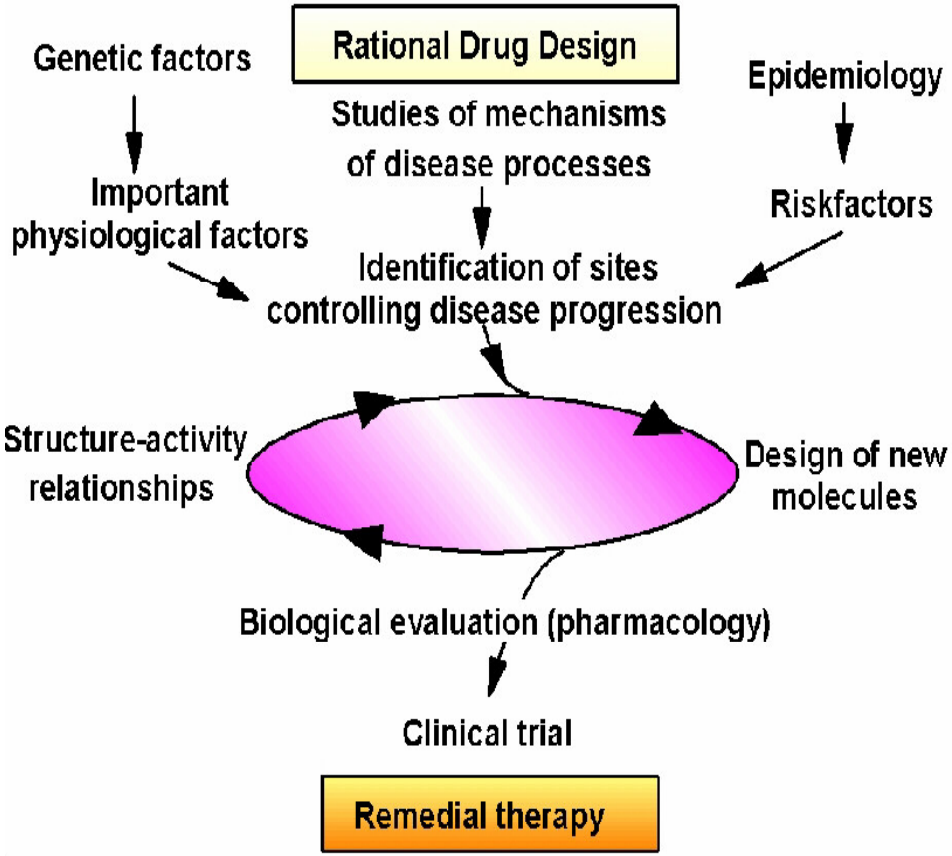

# Practical Relevance

(Pathogen only folds as possible targets)

# New Medical Applications (1)

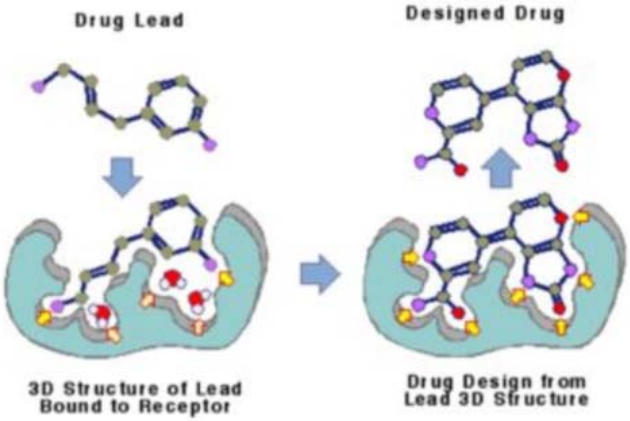
- Diseases control
  - Diagnosis
  - Monitoring
  - Treatments

- Rational drug design
  - New classes of medicines based on **a reasoned approach**
    - Gene sequence, protein structure, function information vs. trial-and-error methods

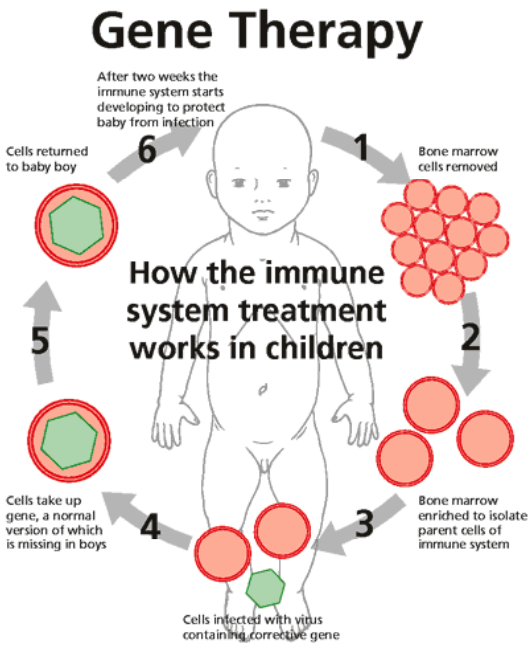


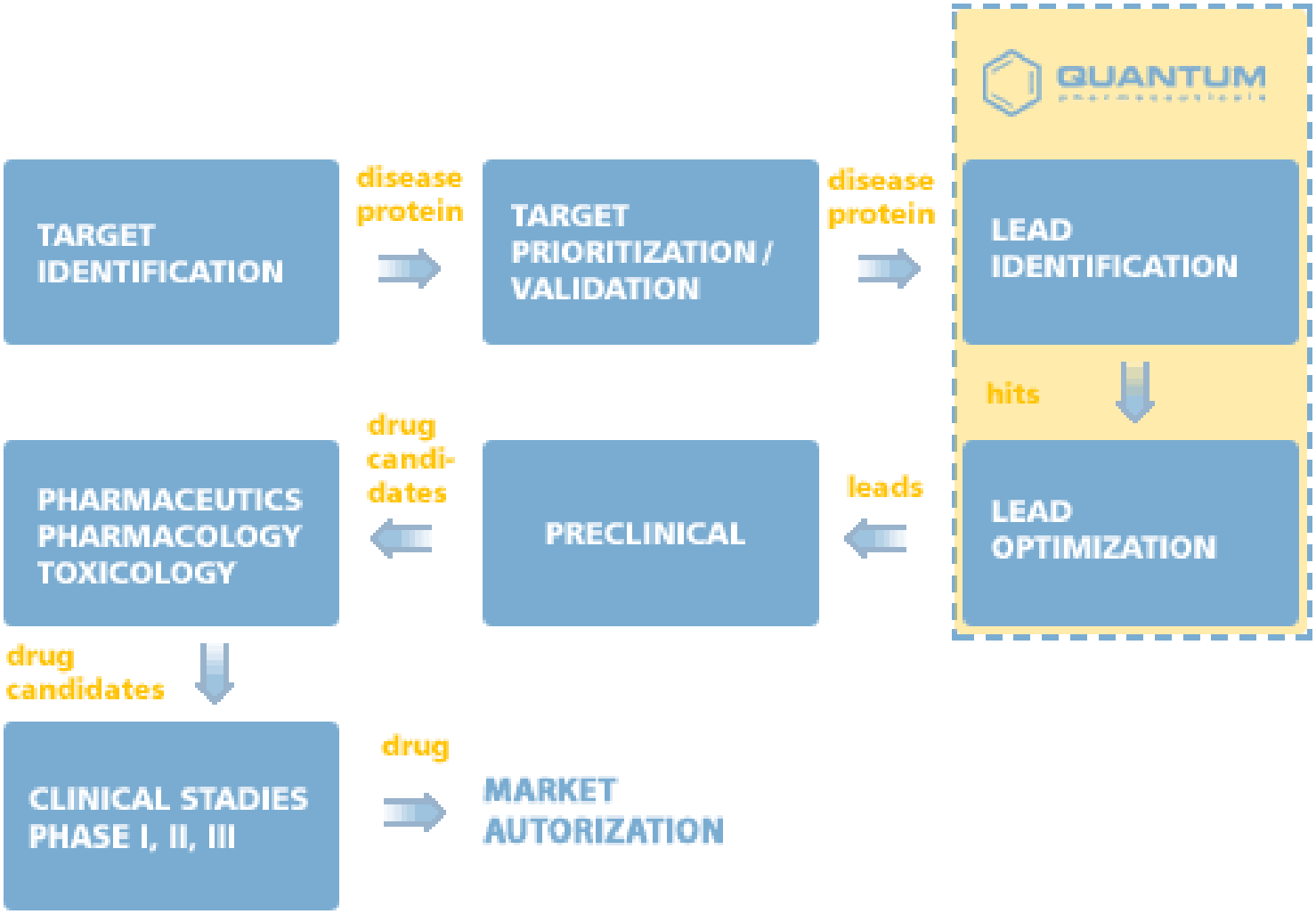
# New Medical Applications (2)

- **Rational drug design** (cont.)
  - These drugs, targeted to specific sites in the body, promise to have fewer side effects than many of today's medicines




- **Gene Therapy**
  - **Normal genes** replace or supplement a defective gene or to bolster immunity to disease
    - *E.g.,* by adding a gene that **suppresses tumor growth**

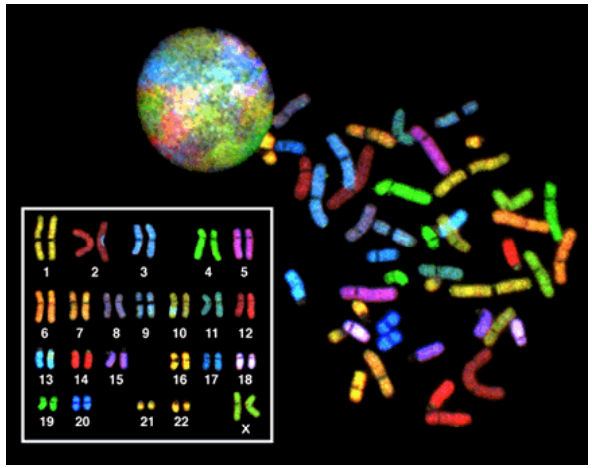




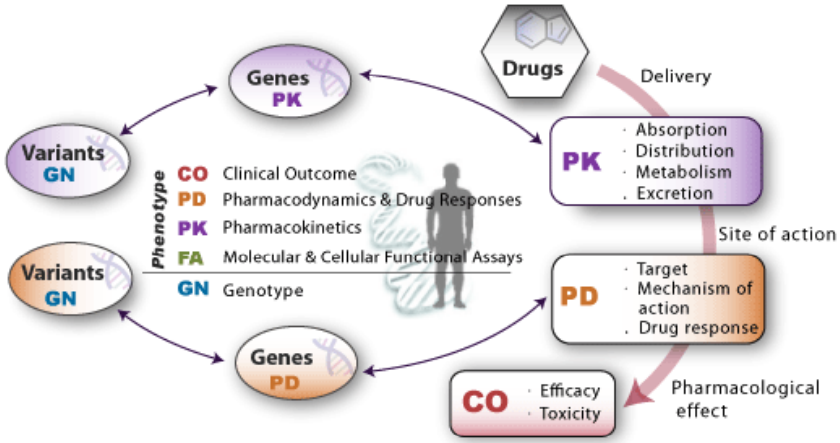

QUANTUM PHARMACEUTICALS' ROLE IN DRUG DISCOVERY

# New Medical Applications (3)

- Other Information
  - http://www.ornl.gov/hgmis/medicine/medicine.htm
  - Gene testing
    - Biochemical tests (enzymes & other protein)
    - Karyotyping
    - DNA level



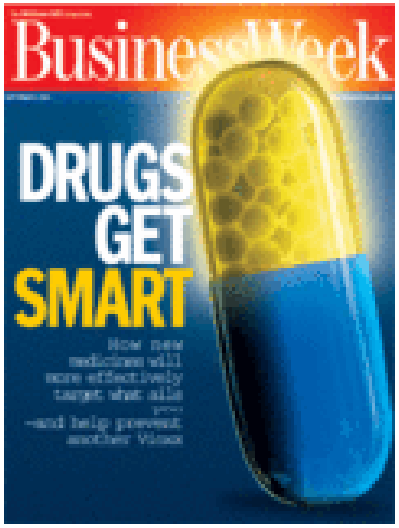

- Pharmacogenetics ⇨ **Pharmacogenomics**
  - The genetic basis of variable drug response in individual people
  - Genetically determined variation in effectiveness & side effects

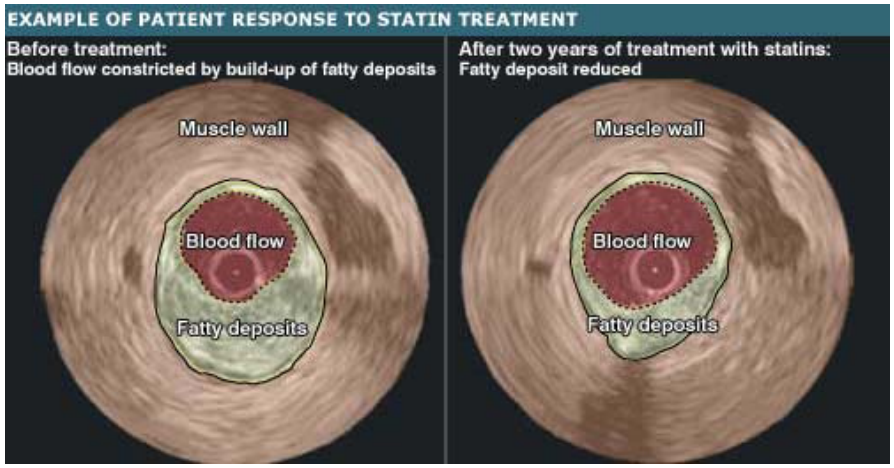



http://www.pharmgkb.org/

# New Medical Applications (4)

- Other Information
  - Tailored drugs vs. "one-size fits all"
    - ***B1B1* variant of *CETP* gene** ⇨ ***paravastatin*** is more effective **in lowering lipid levels** (than other people) ⇨ reduce the risk of cardiovascular disease
    - Drug ***tamoxifen*** prevents breast cancer among women with ***BRCA1* & *BRCA2* gene mutations**



# New Medical Applications (6)

- **Genetic counseling**
  - Genetic counselors are **health professionals** with specialized **graduate degrees** & experience in the areas of medical genetics & counseling

- **Disease specific information**
  - OMIM (Online Mendelian Inheritance in Man)
  - NSGC (National Society of Genetic Counselors)

**Components of the Genetic Counseling Process**

1. Information gathering
2. Diagnosis
3. Risk assessment
4. Information giving
5. Psychological assessment and counseling
6. Help with decision making
7. On-going client support

*[Modified from: AP Walker (1997)]*

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
網址(D) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=OMIM
Norton AntiVirus 網站小幫手 Dr.eye 繁 繁轉簡 繁簡轉換 繁 簡
輸入您想搜尋的文字 搜尋 書籤 我的 服務 股市 信箱 新聞 購物1 娛樂 行事曆 通訊錄 登出
搜尋 醒目提示 選項 封鎖快顯視窗 (1) Hotmail Messenger 我的 MSN

NCBI

OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University

My NCBI
[Sign In] [Register]

PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM

Search OMIM for oral cancer Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Titles Show: 20 Send to Text

All: 65

Entrez

OMIM
Search OMIM
Search Gene Map
Search Morbid Map

Help
OMIM Help
How to Link

FAQ
Numbering System
Symbols
How to Print
Citing OMIM
Download

OMIM Facts
Statistics
Update Log
Restrictions on Use

Allied Resources
Genetic Alliance
Databases
HGMD
Locus-Specific
Model Organisms
MitoMap
Phenotype
Davis
Human/Mouse
Homology Maps
Coriell
The Jackson Laboratory
Human Gene Nomenclature

Items 1 - 20 of 65

Page 1 of 4 Next

1: +113705 GeneTests, Links
BREAST CANCER 1 GENE; BRCA1
BREAST CANCER, TYPE 1, INCLUDED
Gene map locus 17q21

2: #176807 GeneTests, Links
PROSTATE CANCER
Gene map locus 1p36.1-p35, Xq27-q28, Xq11-q12, 22q12.1, 20q13, 1q42.2-q43, 19q, 17p11, 1q25, 13q12.3, 11p11.2, 10q25, 10q23.31, 8p22, 8p22, 7q11.23, 7p11-q21, 7p22

3: %148500 Links
TYLOSIS WITH ESOPHAGEAL CANCER; TOC
Gene map locus 17q25

4: *602198 Links
CDK2-ASSOCIATED PROTEIN 1; CDK2AP1
Gene map locus 12q24.31

5: *607224 Links
ORAL CANCER OVEREXPRESSED GENE 1; ORAOV1

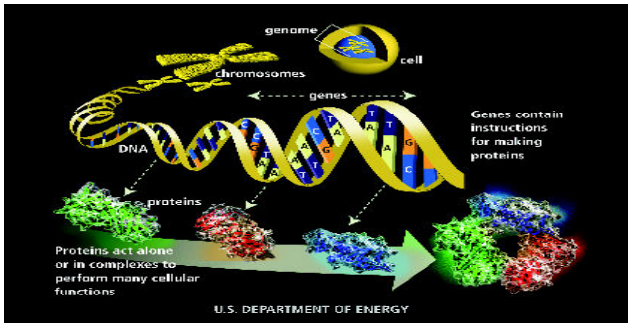
6: *601728 GeneTests, Links
PHOSPHATASE AND TENSIN HOMOLOG; PTEN
PTEN HAMARTOMA TUMOR SYNDROME WITH GRANULAR CELL TUMOR, INCLUDED
Gene map locus 10q23.31

7: #208900 GeneTests, Links
ATAXIA-TELANGIECTASIA; AT
AT, COMPLEMENTATION GROUP A, INCLUDED; ATA, INCLUDED

Human Genome Project Information

Home | Site Index | News | About HGP | Research | Education | Ethics | Medicine | Media

## Genome Glossary




## Genetics 101

Human Genome Project Information

Home | Site Index | News | About HGP | Research | Education | Ethics | Medicine | Media

## 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

Human Genome Project Information

Home | Site Index | News | About HGP | Research | Education | Ethics | Medicine | Media

Mapping | Sequencing | Technology | Bioinformatics | Gene Function | ELSI | Microbes

## DOE Human Genome Program Research in Progress

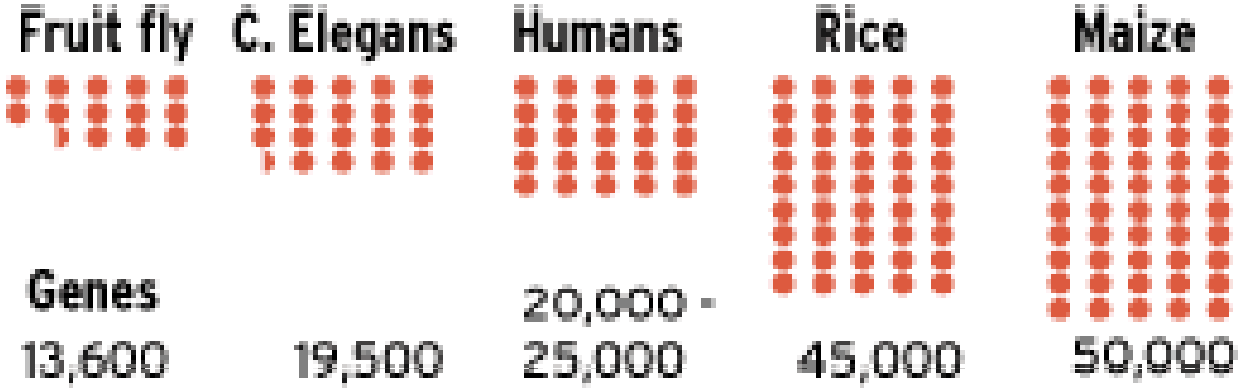
Human Genome Project Information

Home | Site Index | News | About HGP | Research | Education | Ethics | Medicine | Media

## Genome & Biotechnology Meetings Calendar

# Humans have fewer genes

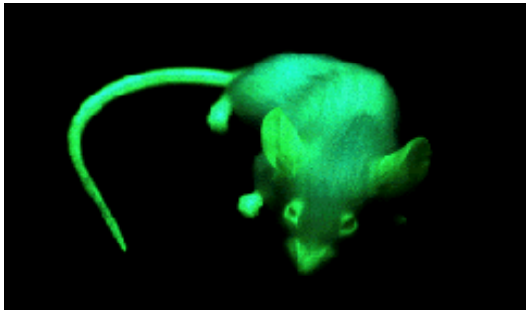
In Thursday's issue of the journal Nature, researchers who decoded the human genome concluded that people have only 20,000 to 25,000 genes, a drop from the 30,000 to 40,000 estimated in 2001.

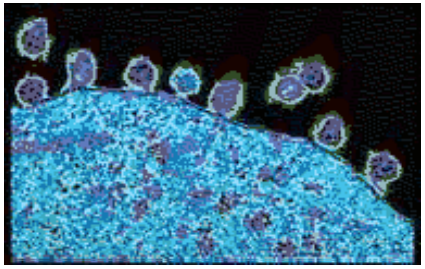



SOURCE: Nature

AP


# Organism-specific Resources





*Human*

*Drosophila*

*Zebrafish*

*Malaria parasite*

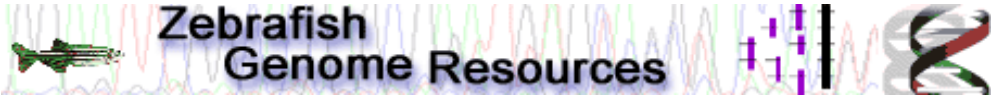
*Microbial Genomes* (84 complete genomes, Aug. 2002)

*Mouse*

*Plant Genome Central*

*Rat*

*Retroviruses*



# 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

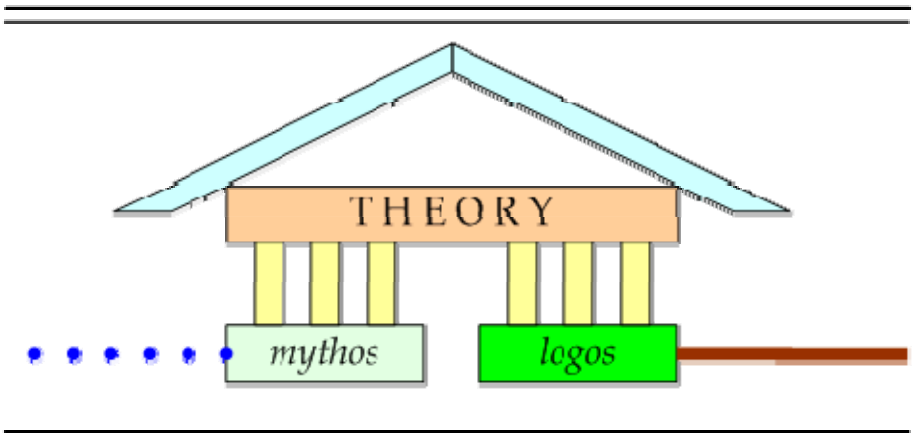
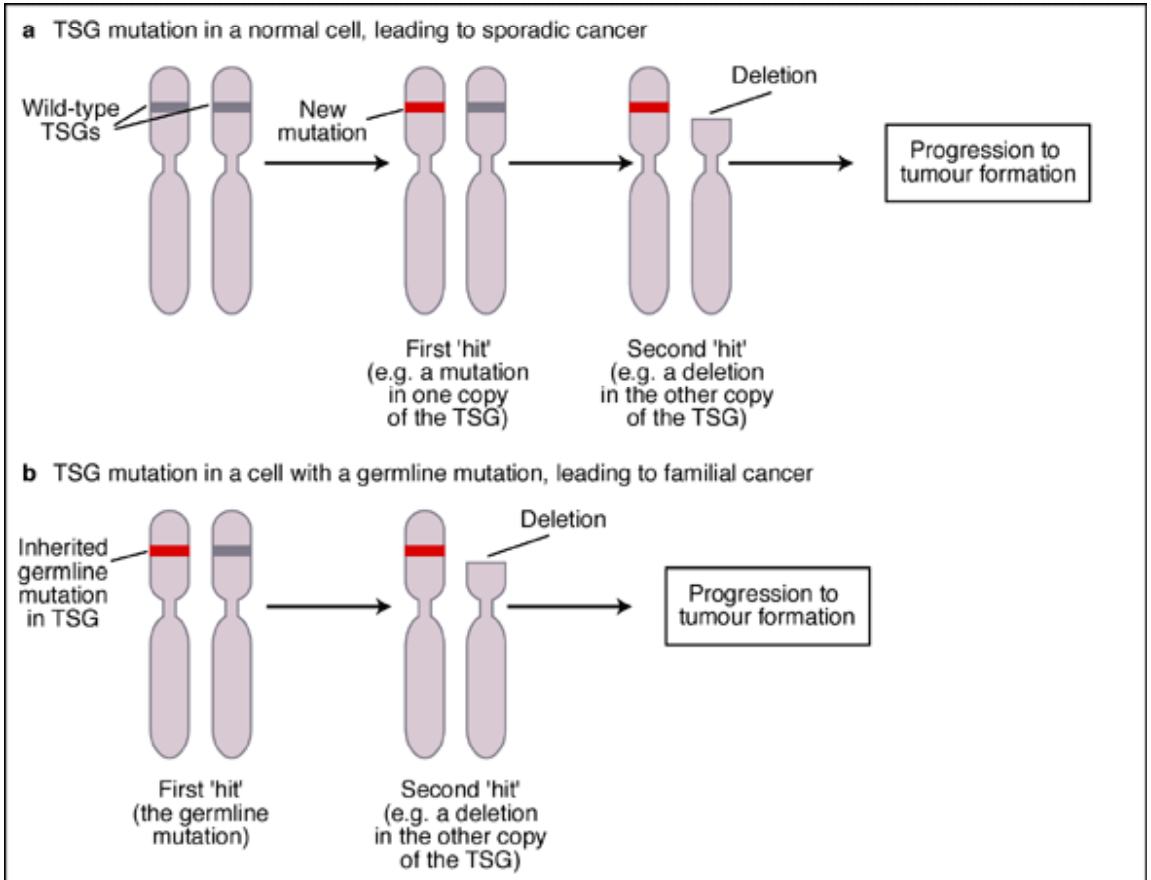



*Figure 3 – Building a Theory on Mythos and Logos*

http://userpages.burgoyne.com/bdespain/grammar/gram012.htm

# Two-hit Hypothesis for Tumorigenesis




Knudson's two-hit hypothesis for tumourigenesis involving a tumour suppressor gene (TSG)

Expert Reviews in Molecular Medicine ©2001 Cambridge University Press

# 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

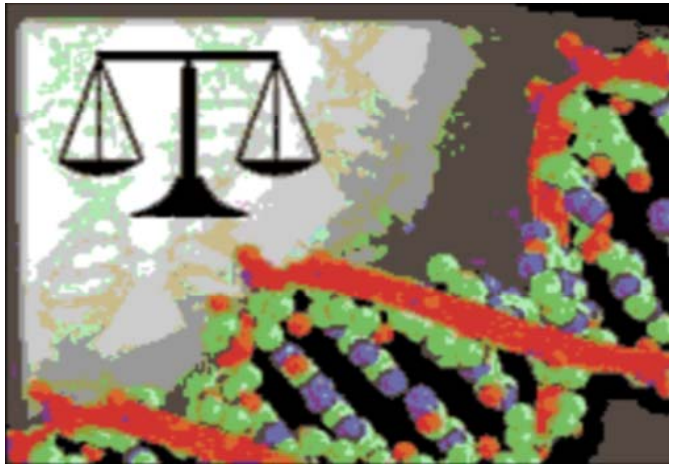

PHASE TWO : INTERPRETATION
SHERMAN
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME


ONCE YOU UNFOLD ONE OF THESE THINGS, ITS NEVER THE SAME.
MAP OF THE HUMAN GENOME
Ethical Questions
Legal Tangles
Privacy concerns
Moral Issues
Insurance Applications
Medical Dilemmas

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

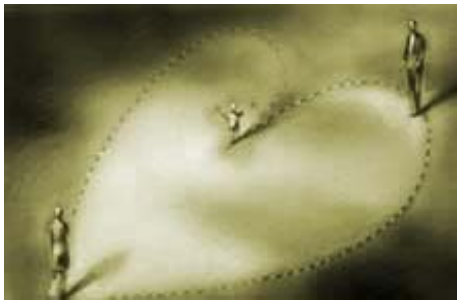
# Major Implications of the Genetic Revolution for the Legal Discipline (1)

- 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



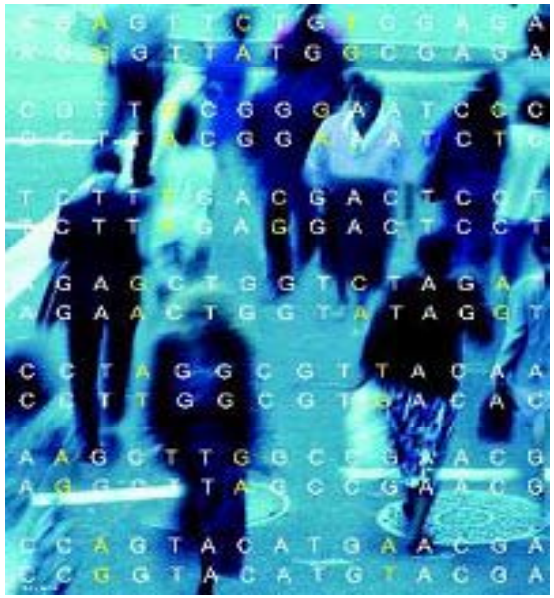

# Major Implications of the Genetic Revolution for the Legal Discipline (2)

- 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

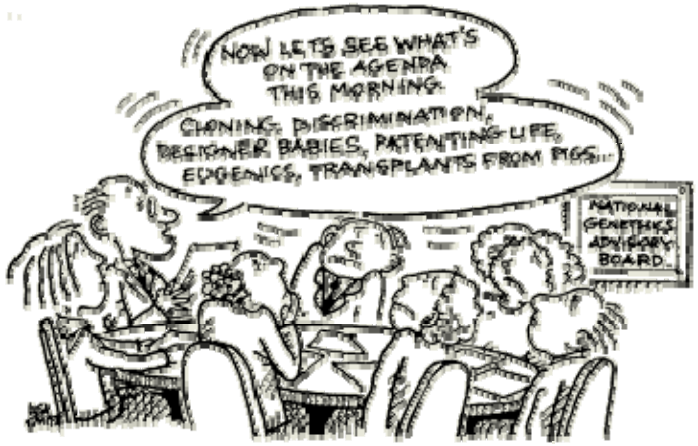


# Genetic Discrimination (1)

- 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



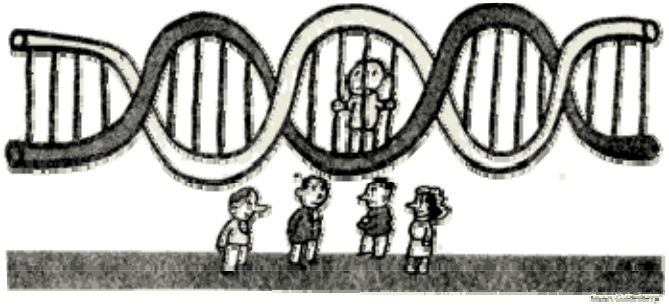
DOE, USA



# 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.*

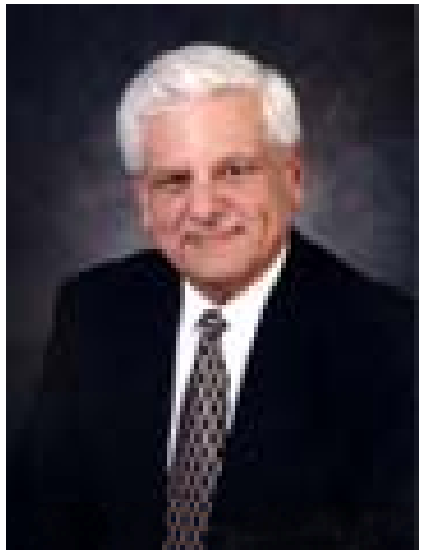


# Genetic Discrimination (3)

- 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
  - JHU: the Johns Hopkins University

# 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



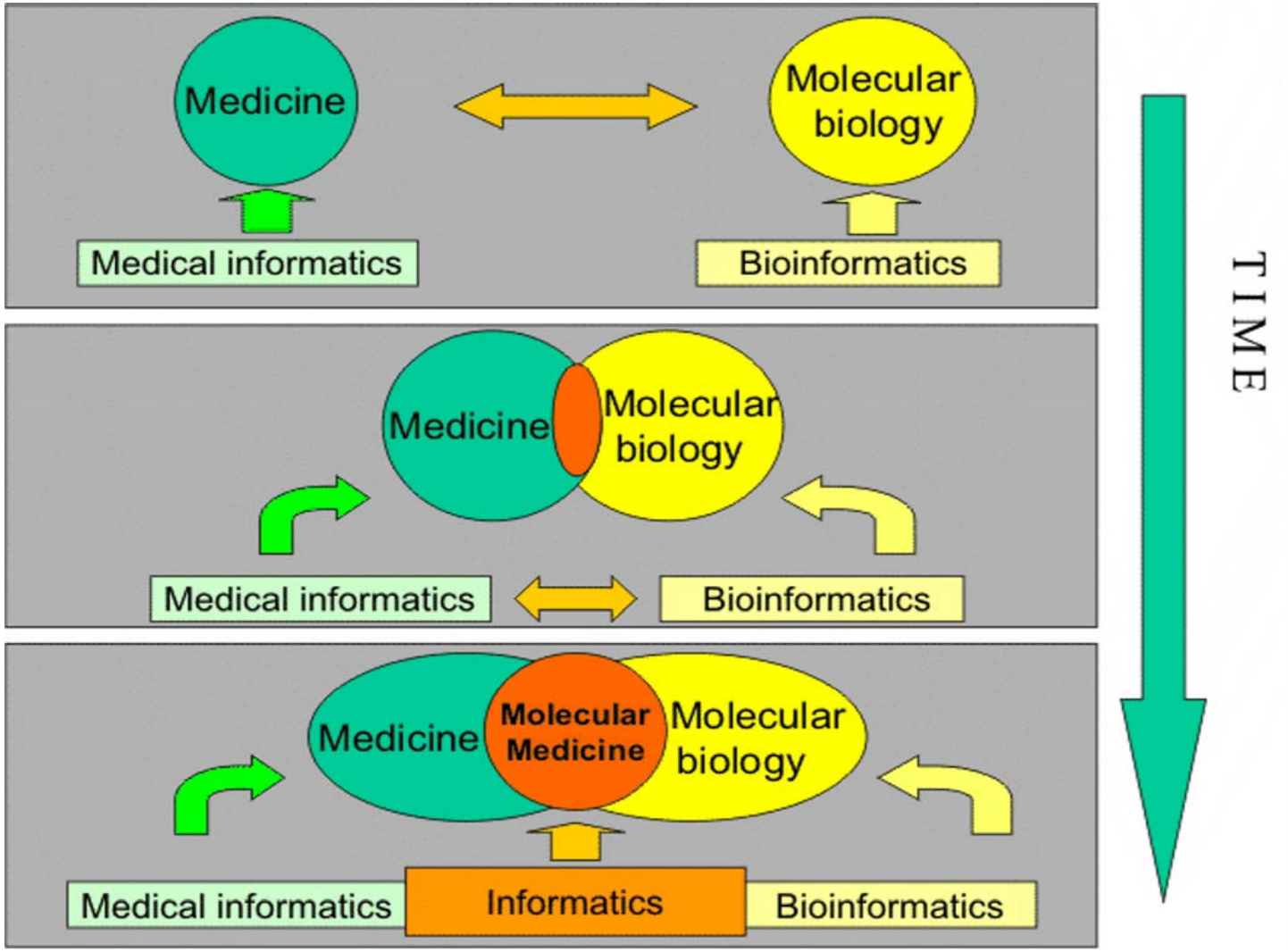
# Small Business & Health Insurance (1)

- 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

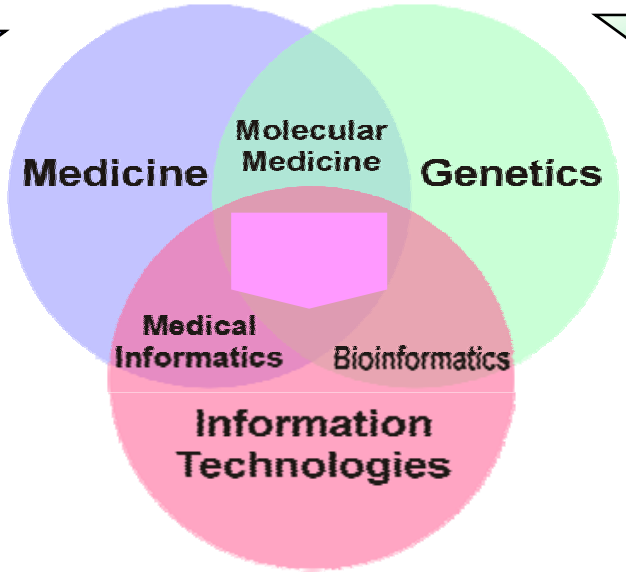

# The Convergence between MI & BI



# 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

# Definition (1)

- **Bioinformatics**
  - 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 (2)

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

# Central Paradigm of Bioinformatics

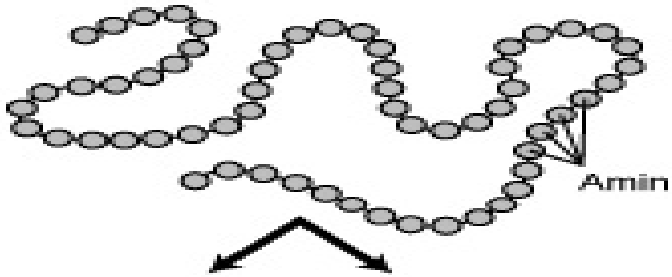
- **Central dogma of molecular biology**
  - [DNA → RNA → protein ]→ **phenotype**
- **Molecules**
  - **Sequence → structure → function**
    - Most cellular functions are performed or facilitated by **proteins**
      - Primary biocatalyst, co-factor **transport/storage**, mechanical motion/support,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

# Glossary of Bioinformatics

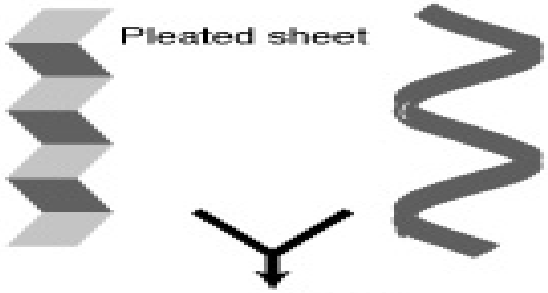
- Cambridge Health Institute
  - http://www.genomicglossaries.com/content/Bioinformatics_gloss.asp
- 2-can Glossary
  - http://www.ebi.ac.uk/2can/glossary/index.php
- **Contents**
  - Databases
  - Methodologies

# Contents – Databases (1)

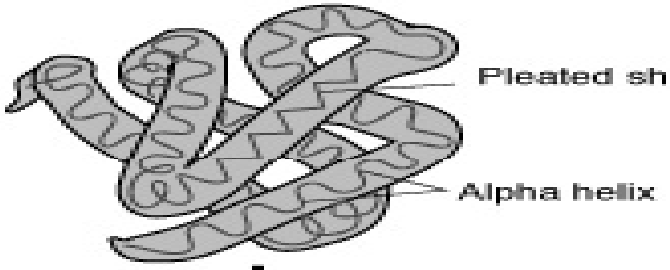
- *Nucleic Acid Research* (NAR) Jan. (every year)
  - http://nar.oupjournals.org/
- **Protein information resources**
  - **Primary (linear)**
    - PIR, MIPS, SWISSPROT, PDB
    - **Composite** protein sequence databases
  - **Secondary (motif)**
    - Prosite, Profiles, PRINTS, Pfam, Block, IDENTIFY
  - **Tertiary/Structure (domain, module)**
    - SCOP, CATH, PDBsum




**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



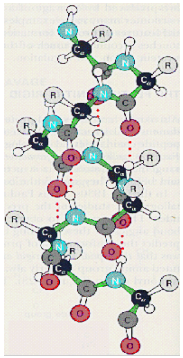

**Tertiary protein structure**
occurs when certain attractions are present between alpha helices and pleated sheets

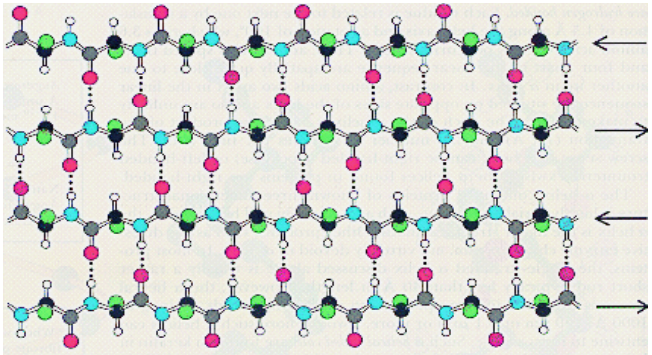
**Quaternary protein structure**
is a protein consisting of more than one amino acid chain.

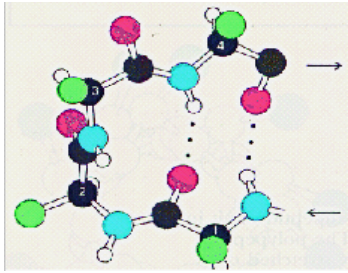
- **Primary structure**
  - The linear **sequence** of **amino acids in a protein**

```
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN
```

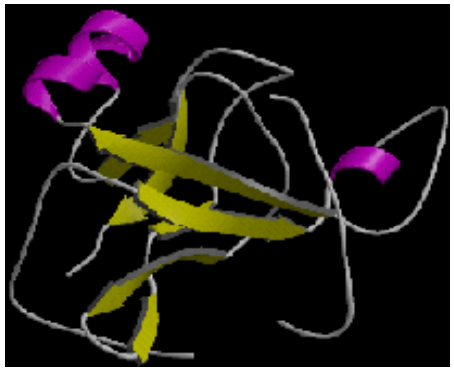
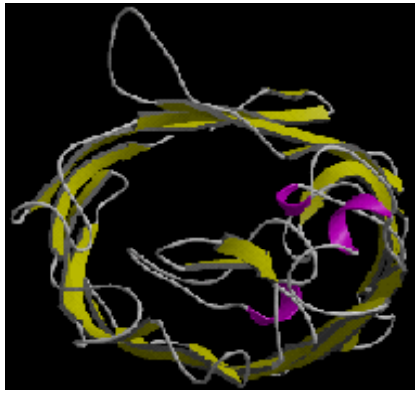
- **Secondary structure**
  - Regions of **local regularity**
    - *i.e.,* alpha-helices, beta-strands, beta-sheets & beta-turns



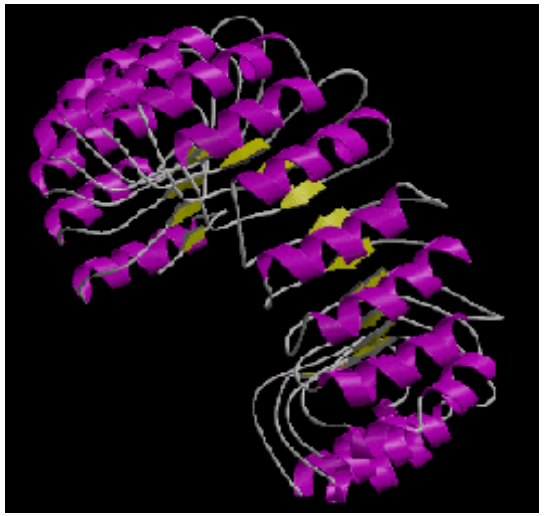




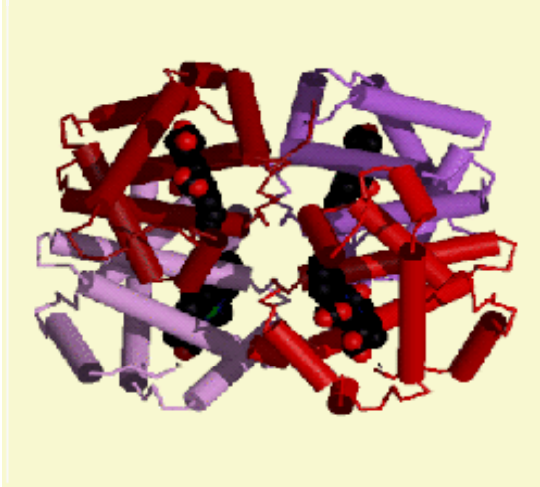
- **Super-secondary structure**
  - The packing of secondary structure elements into stable units **(motifs, modules)**
    - *e.g.*, β-barrels, βαβ units, Greek keys, *etc.*

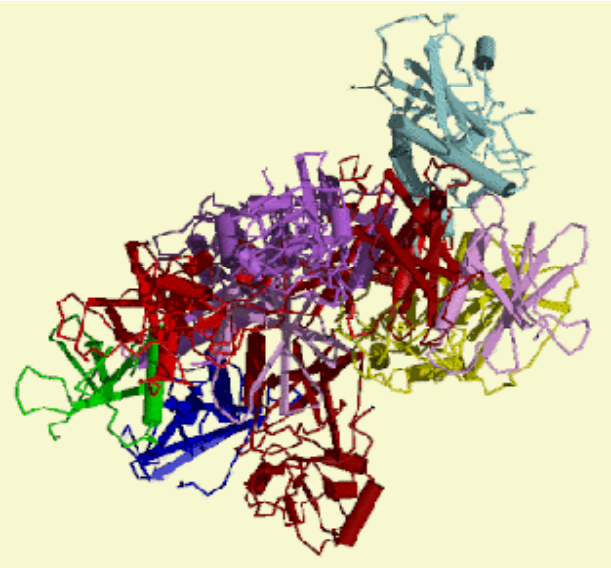




- **Tertiary structure**
  - **The overall chain fold** that results from packing of **secondary** structure elements

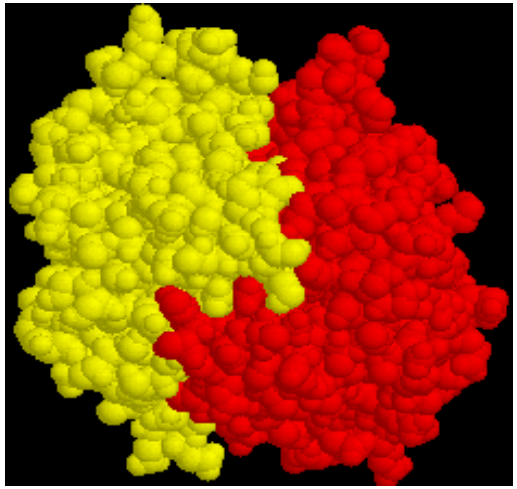


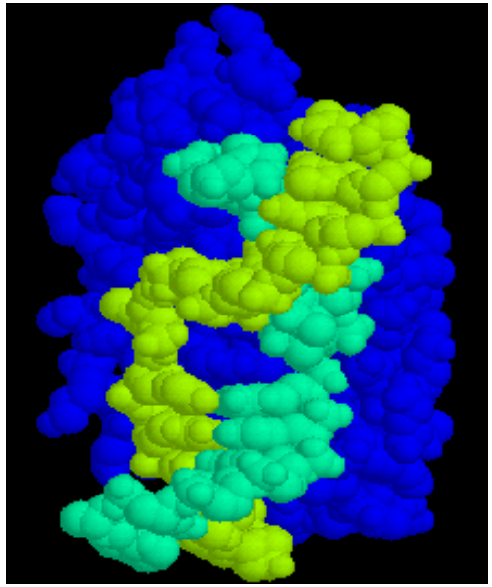
- **Quaternary structure**
  - The arrangement of **separate chains** within a protein that has **more than one subunit**
    - *e.g.*, hemoglobin





- **Quinternary structure**
  - The arrangement of separate molecules, such as in protein-protein or protein-nucleic acid interactions





# Contents – Databases (2)

- **Genome information resources**
  - **DNA sequence databases**
    - EMBL, DDBJ, GenBank
    - dbEST, dbSTS, dbSNP *etc.*
  - **Specialized genomic resources**
    - SGD (the *Saccharomyces* Genome Database)
    - Unigene (NCBI, USA)
  - **TDB (the TIGR database)**
    - **A suite of databases** containing **DNA** & **protein** sequences, gene expression, cellular role, and protein family information, taxonomic data for microbes, plants, humans
- Intermolecular **interactions** & biological pathways

# Contents – Methodologies (1)

- **Algorithms**
  - The logical sequence of **steps** by which a task can be performed
- **Comparison**
  - **Pairwise alignment**
    - Local vs. global alignment
  - **Multiple alignment**
    - PSI-BLAST (position-specific iterated – BLAST)
      - Automatic
- **Database searching**
  - Reference searching (by keyword, text)
  - Sequence-based
- **Editing**
  - Single or multiple sequence editing
    - *E.g.*, plasmid removal
- **Evolution**
  - Phylogenetic relatedness

# Contents – Methodologies (2)

- **Fragment assembly**
  - *E.g.*, **contig** assembly
- Gene finding and **pattern recognition**
  - Protein-coding regions
  - Protein-binding motifs
  - Repeats
  - CpG islands & promoter regions
- Importing/Exporting
  - Entering sequence data and converting the data between the **various sequence** file formats, *e.g.*, GCG, Staden, EMBL, GenBank, IntelliGenetics, PIR, and FASTA *etc.*

# Contents – Methodologies (3)

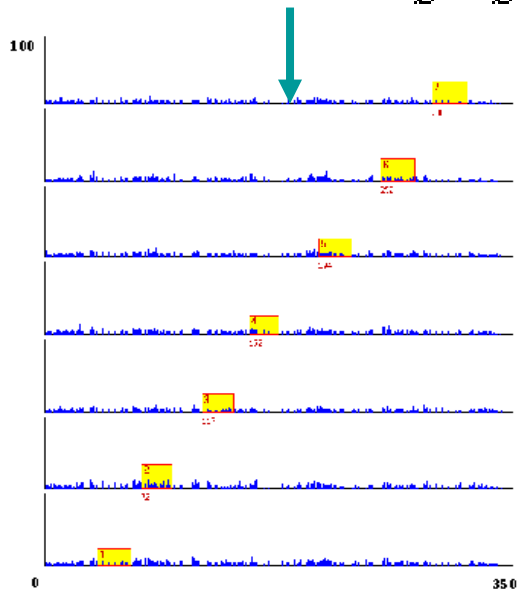
- **Mapping**
  - Restriction maps
  - ORF maps
  - Peptide digestions maps
  - Plasmid maps, *etc.*
- **Primer designs**
- **Protein analysis**
  - Determining information about protein & amino acid sequences
    - Plotting the isoelectric point
    - Location of **functional motifs**
    - Predictions of **secondary structure**
      - **Epitope** & antigenicity
      - **Secretory signals**
      - Nuclear localized signals (NLS)
      - Transmembrane proteins
- **Protein structure prediction & analysis**
- **Computational approaches in comparative genomics**
- **Using DNA microarrays to assay gene expression**
- **Proteomics & protein identification**

# Contents – Methodologies (4)

- **RNA secondary structure**
  - *E.g.*, inverted repeat sequences
- **Translation**
  - Translation nucleotide sequences into **peptide sequence** or *vice versa*
- **Other utilities**
  - Sequences management
  - Databasing
  - Printing/plotting
- **Internet connection**

# The Reality of Sequence Analysis

MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN

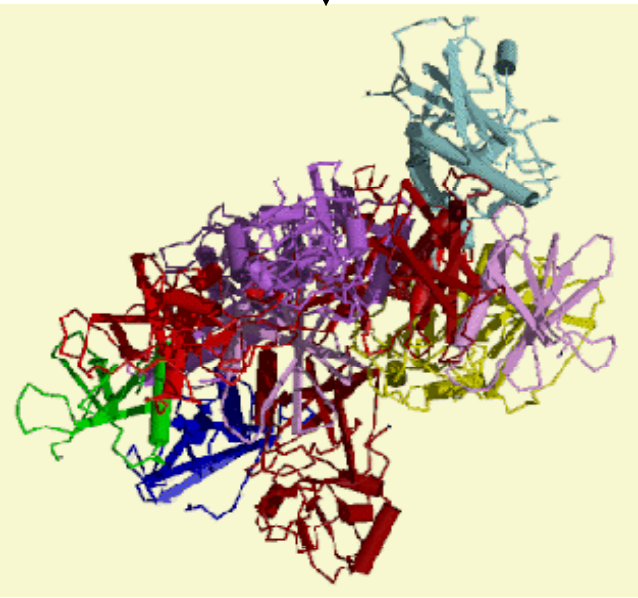


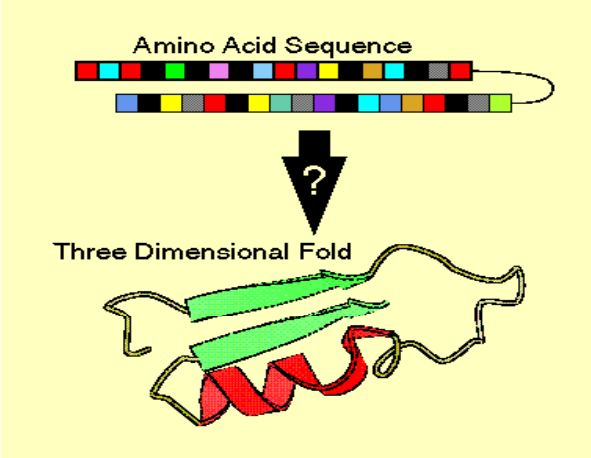

...isn't so glamorous....but means we **can** recognize words that form **characteristic patterns**, even if we don't know the precise syntax to build complete protein sentences

(from Attwood & Parry-Smith 1999)

# The Holy Grail of Bioinformatics

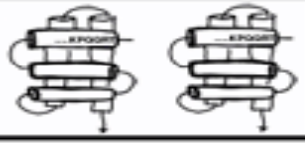
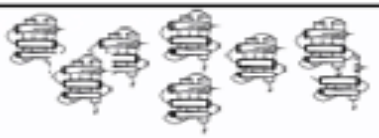
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVL
GFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN






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

(from Attwood & Parry-Smith 1999)

| Depth: Rational Drug Design (physics)→ | | 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 | atcgatcgatatttgggatttgggga | atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga | atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga | atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga<br>atcgatcgatatttgggatttgggga |
| gene finding | ↓ | | | | |
| | Protein Sequence | ALMNAKKKPQQRT | ALMNAKKKPQQRT<br>ALMNAKKKPQQRT | ALMNAKKKPQQRT<br>ALMNAKKKPQQRT<br>ALMNAKKKPQQRT<br>ALMNAKKKPQQRT | ALMNAKKKPQQRT<br>ALMNAKKKPQQRT<br>ALMNAKKKPQQRT<br>ALMNAKKKPQQRT<br>ALMNAKKKPQQRT<br>ALMNAKKKPQQRT<br>ALMNAKKKPQQRT<br>ALMNAKKKPQQRT |
| structure prediction | ↓ | | | | |
| | Protein Structure |  |  |  |  |
| geometry calculation | ↓ | | | | |
| | Protein Surface |  | | | |
| molecular simulation | ↓ | | | | |
| | Force Field |  | | | |
| structure docking | ↓ | | | | |
| | Ligand Complex |  | | | |

http://www.cs.ucsb.edu/~ambuj/Courses/bioinformatics/definition.pdf

# Data

- Data is **crucial** to the success of analysis
  - "Garbage in & garbage out"
    - ***A little*** *garbage in* ⇨ ***A lot of*** *garbage out*
  - Understand your data set and its surrounding **metadata**
    - Whole picture

"Don't just sit there! If you've processed all the data there is, go out and find <u>more</u> data!"

Reproduced in R.L. Weber, "*A random walk in science*", IOP Publishing, 1973