Welcome To All

Introduction to Genomics

Dr.B.Vaseeharan,
Associate Professor &Head
Department of Animal Health and Management
Alagappa University
Karaikudi-630003

Why was this subject proposed?

Genomics is the study of genome structure and function. This is a new and exciting area that has recently witnessed many conceptual and technical advances. This information is vital to our day-to-day living in this century.

What will you learn in this class?

- Genes, expression and characters
- What is genomics?
- Central dogma: DNA to RNA to Proteins
- Features of human genome
- How is genome sequenced and studied?
- Applications of genomics in Agriculture, Medicine and Bioenergy

- **We are in the midst of a "Golden Era" of biology**

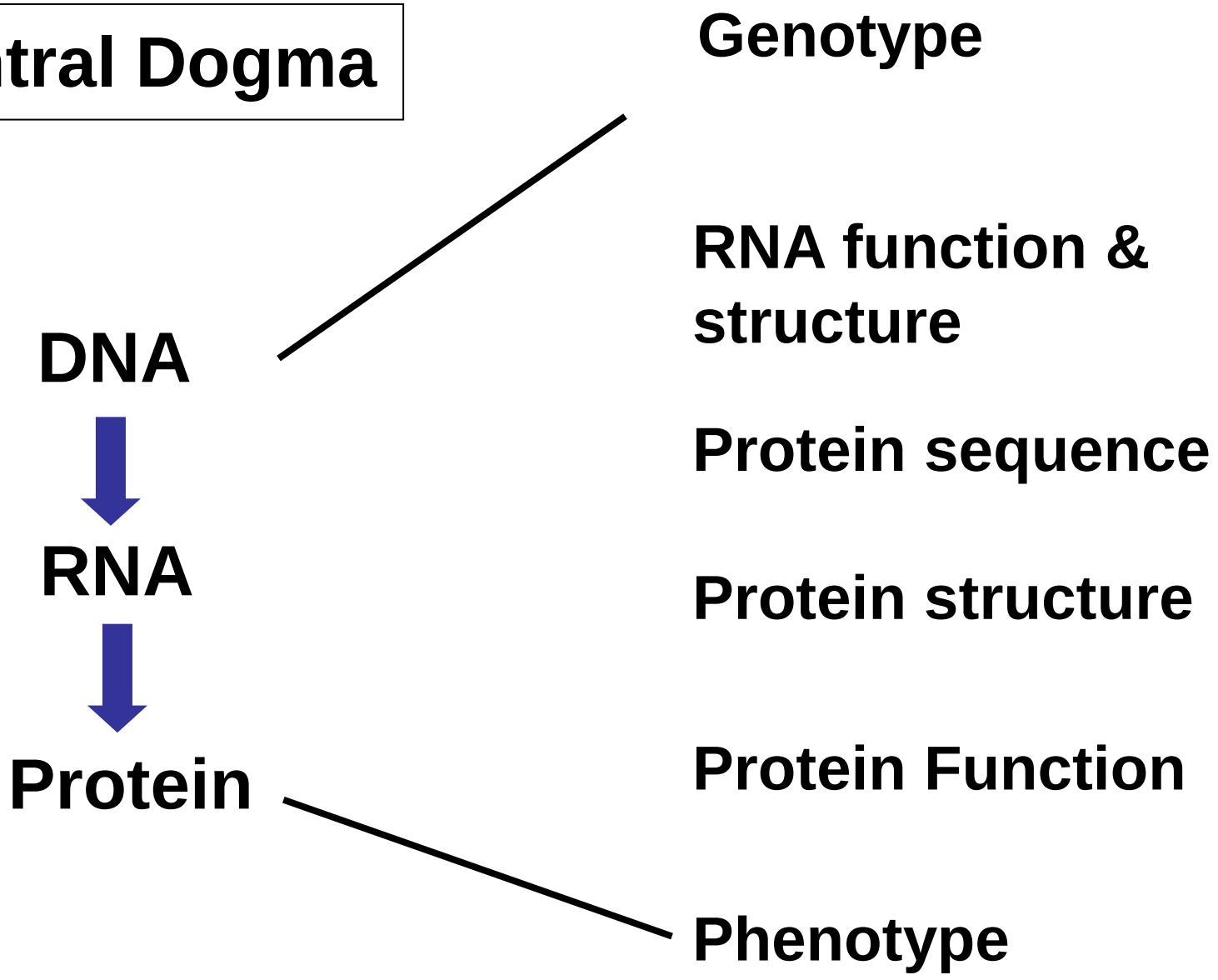
- **The revolution is mostly about treating biology as an information science, not only specific biochemical technologies.**

Genomics

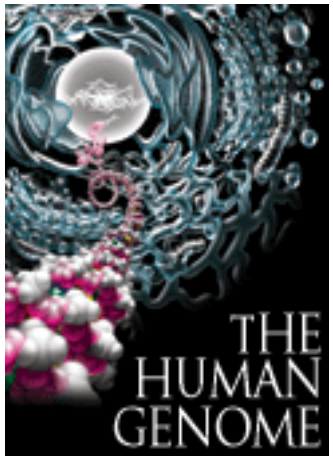
Study of sequences, gene organization & mutations at the DNA level

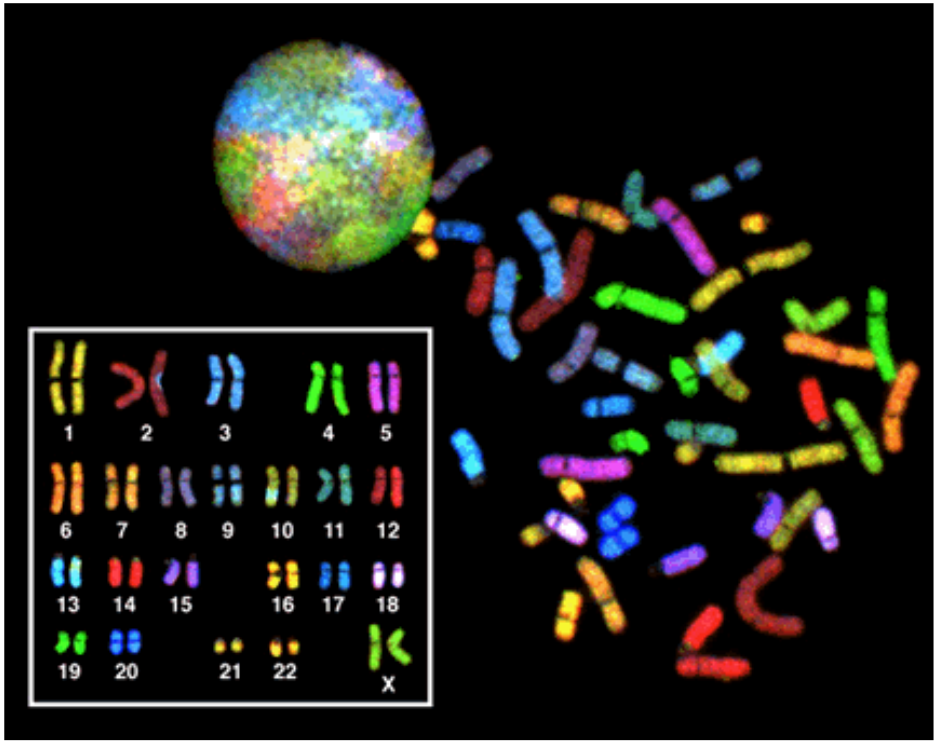
the study of information flow within a cell

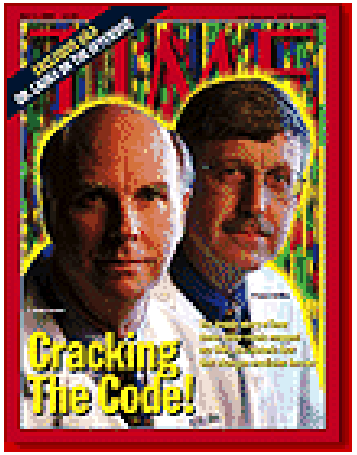
Central Dogma



The Human Genome Project







3 billion bases **30,000 genes**

http://www.genome.gov/

- **Would it be important to know your personal DNA sequence?**

- **Would you want to know if you were susceptible to a disease? Why or why not?**

Impact of Genomics on Medicine

- How to characterize new diseases?
- What new treatments can be discovered?
- How do we treat individual patients? Tailoring treatments?

Implications for Biomedicine

- Physicians will use genetic information to diagnose and treat disease.
    - Virtually all medical conditions have a genetic component
- Faster drug development research: (pharmacogenomics)
    - Individualized drugs
- All Biologists/Doctors will use gene sequence information in their daily work

URFs and ORFs — definitions

URF: Unidentified Reading Frame — any potential string of amino acids encoded by a stretch of DNA. Any given stretch of DNA has potential URFs on any combination of six potential reading frames, three forward and three backward.

ORF: Open Reading Frame — by definition any continuous reading frame that starts with a start codon and stops with a stop codon. Not usually relevant to discussions of genomic eukaryotic DNA, but very relevant when dealing with mRNA/cDNA or prokaryotic DNA.

Signal Searching:

locating transcription and translation affecter sites.

One strategy — One-Dimensional Signal Recognition.

Start Sites:

Prokaryote promoter ‘Pribnow Box,’
TTGACwx{15,21}TAtAaT;

Eukaryote transcription factor site database,
TFSites.Dat;

Shine-Dalgarno site, **(AGG,GAG,GGA)x{6,9}ATG**, in prokaryotes;

Kozak eukaryote start consensus, **cc(A,g)ccAUGg**;

AUG start codon in about 90% of genomes, exceptions in some prokaryotes and organelles.

How many types of genomes are there in this world?

Prokaryotic genomes

Eukaryotic Genomes

Nuclear Genomes

Mitochondrial genomes

Choloroplast genomes

Why should we study genomes?

- **Each and everyone is a unique creation!**
- **Life's little book of instructions**
- **DNA blue print of life!**
- **Human body has 10^{13} cells and each cell has 6 billion base pairs (A, C, G, T)**
- **A hidden language/code determines which proteins should be made and when**
- **This language is common to all organisms**

What can genome sequence tell us?

- Everything about the organism's life
- Its developmental program
- Disease resistance or susceptibility
- History
- Where you are going?

How is human genome organized?

- 3% coding and rest of it junk (repetitive DNA).
- Nuclear and mitochondrial
- You are 99.99% similar to your neighbor

Why human genome?

- We want to know about ourselves
- How do we develop?
- How do we struggle, survive and die?
- Where are we going and where we came from?
- How similar are we to apes, trees, and yeast?

How will we change in this century because of the Genomics?

- You will control the destiny of this planet
- Big changes in our own life
- Biotechnology: more products
- GMOs: More food-More problems?
- Our society will not be the same!
- Individualized medicine
- Gene therapy
- Immortality? Disease free life?

Are we playing GOD?

Central dogma in Molecular Biology

DNA sequence

- 1 gtcgacccac gcgtccgtct tgaaagaata tgaagttgta aagagctggt aaagtggtaa
- 61 taagcaag**at** gatggaatct ggggctccta tatgccatac ctgtggtgaa caggtggggc
- 121 atgatgcaaa tggggagcta tttgtggctt gccatgagtg tagctatccc atgtgcaagt
- 181 cttgtttcga gtttgaaatc aatgagggcc ggaaagtttg cttgcggtgt ggctcgccat
- 241 atgatgagaa cttgctggat gatgtagaaa agaaggggtc tggcaatcaa tccacaatgg
- 301 catctcacct caacgattct caggatgtcg gaatccatgc tagacatatc agtagtgtgt
- 361 ccactgtgga tagtgaaatg aatgatgaat atgggaatcc aatttggaag aatcgggtga
- 421 agagctgtaa ggataaagag aacaagaaga aaaagagaag tcctaaggct gaaactgaac

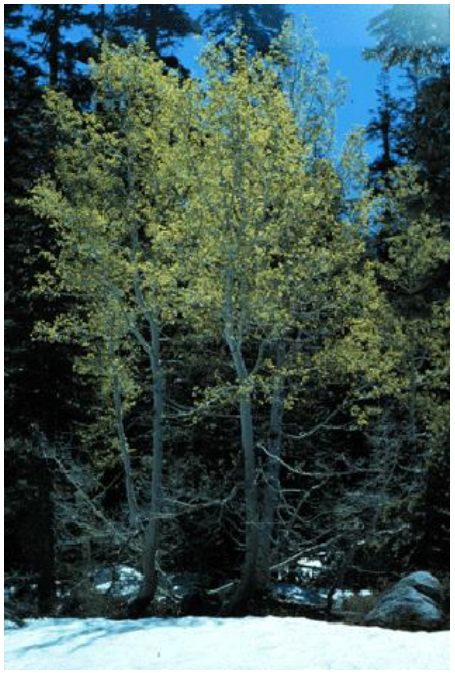
- **Protein coding regions of Genes begin with ATG and end with either TAG, TGA or TAA**

- **at**g atg gaa tct ggg gct cct… use genetic code..
- M M E S G A P ..*
- Study function of proteins and expression of genes in different organs and tissues

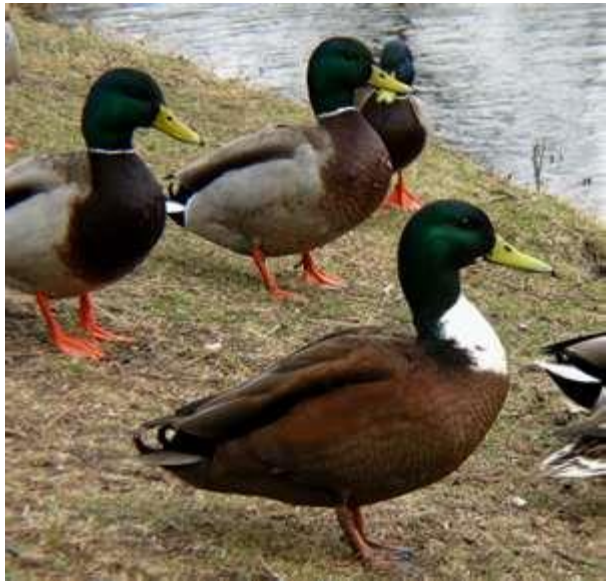
DNA → (transcription) **RNA** → (translation) **Proteins**

First Position	Second Position				Third Position
	T	C	A	G	
T	PHE	SER	TYR	CYS	T
	PHE	SER	TYR	CYS	C
	LEU	SER	stop	stop	A
	LEU	SER	stop	TRP	G
C	LEU	PRO	HIS	ARG	T
	LEU	PRO	HIS	ARG	C
	LEU	PRO	GLN	ARG	A
	LEU	PRO	GLN	ARG	G
A	ILE	THR	ASN	SER	T
	ILE	THR	ASN	SER	C
	ILE	THR	LYS	ARG	A
	MET	THR	LYS	ARG	G
G	VAL	ALA	ASP	GLY	T
	VAL	ALA	ASP	GLY	C
	VAL	ALA	GLU	GLY	A
	VAL	ALA	GLU	GLY	G

Wild type



What genomics has to do with me?



Why horse is a horse and duck is a duck?

- It is in their genes!
- DNA structure was discovered in 1953
- DNA replicates by making a copy of itself and passes to next generation of cells or organisms
- Purity of lineages maintained
- Biotechnology: fish genes in plants

Now look at your neighbor and say Hi!

- What do you see?
- Someone is different than you!
- Could be that your friend differs in his/her sex, looks, nature, smartness, or simply the way he/she dresses and talks
- How much similarity you think you share with your friend at the gene level?
- 99.9% so we could fix genes if we want

Now look at your own hands and legs

- Do they look similar? No!
- But they contain the same DNA in each of their cells
- DNA makes RNA makes proteins
- Different genes are expressed differently in different cells, tissues and organs of an organism
- Having a gene does not mean it will be expressed.

Genomcis is the study of all genes present in an organism

Origin of terminology

- The term genome was used by German botanist Hans Winker in 1920
- Collection of genes in haploid set of chromosomes
- Now it encompasses all DNA in a cell
- In 1986 mouse geneticist Thomas Roderick used Genomics for “mapping, sequencing and characterizing genomes”
- New terms: Functional genomics, transcriptomics, proteomics, metabolomics, phenomics (Omics)

What is genomics?

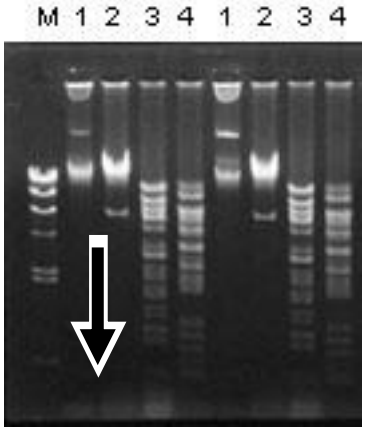
- A marriage of molecular biology, robotics, and computing
- Tools and techniques of recombinant DNA technology
  - e.g., DNA sequencing, making libraries and PCRs
- High-throughput technology
  - e.g., robotics for sequencing
- Computers are essential for processing and analyzing the large quantities of data generated

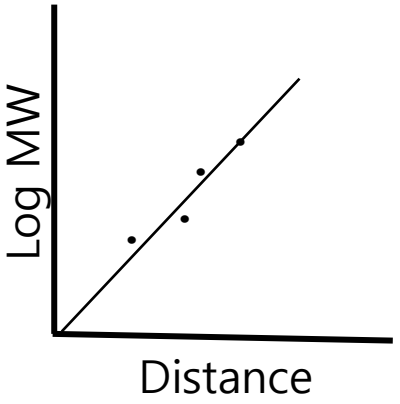
Origin of Genomics

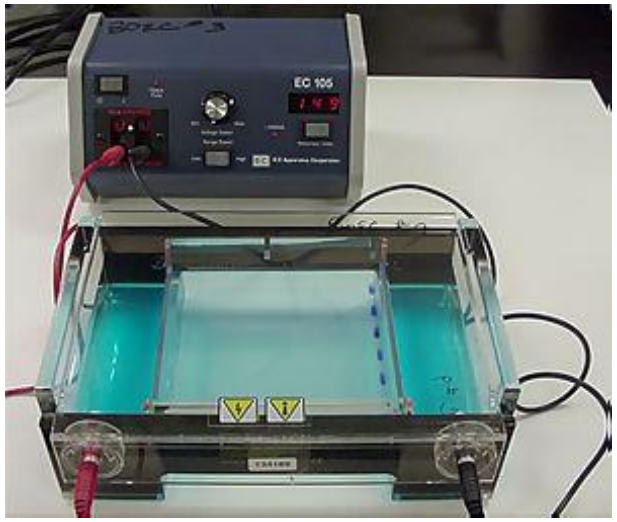
- Human Genome Project
  - Goal: sequence 3 billion base pairs
  - High-quality sequence (<1 error per 10 K bases) ACGT
- Immensity of task required new technologies
  - Automated sequencing
- Decision to sequence other genomes: yeast and bacteria
  - Beginnings of comparative genomics

Technical foundations of genomics

- Molecular biology: recombinant-DNA technology
- **DNA sequencing**
- Library construction
- PCR amplification
- Hybridization techniques

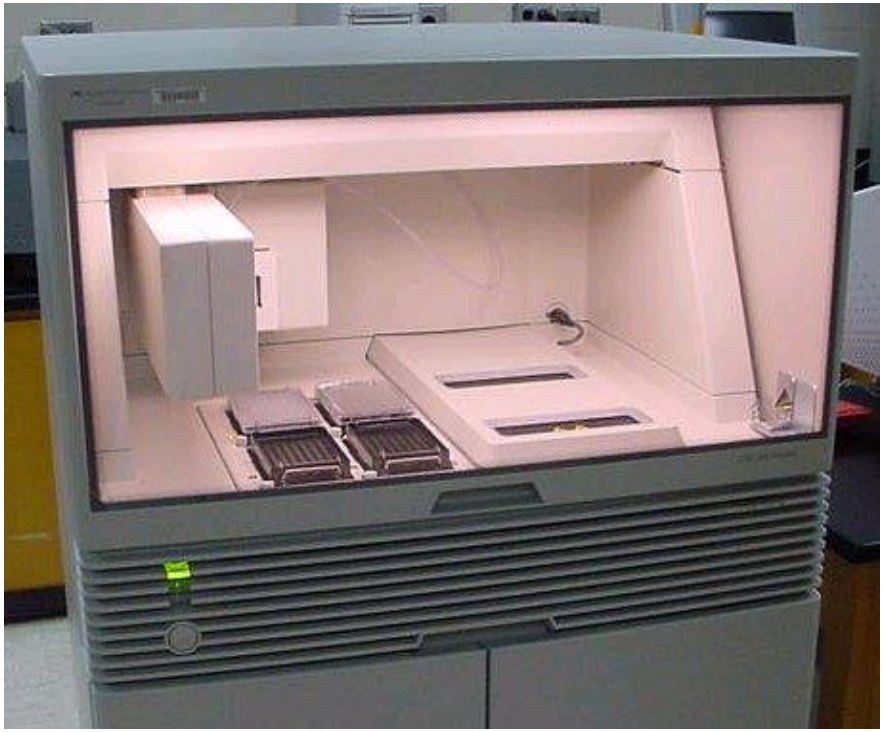






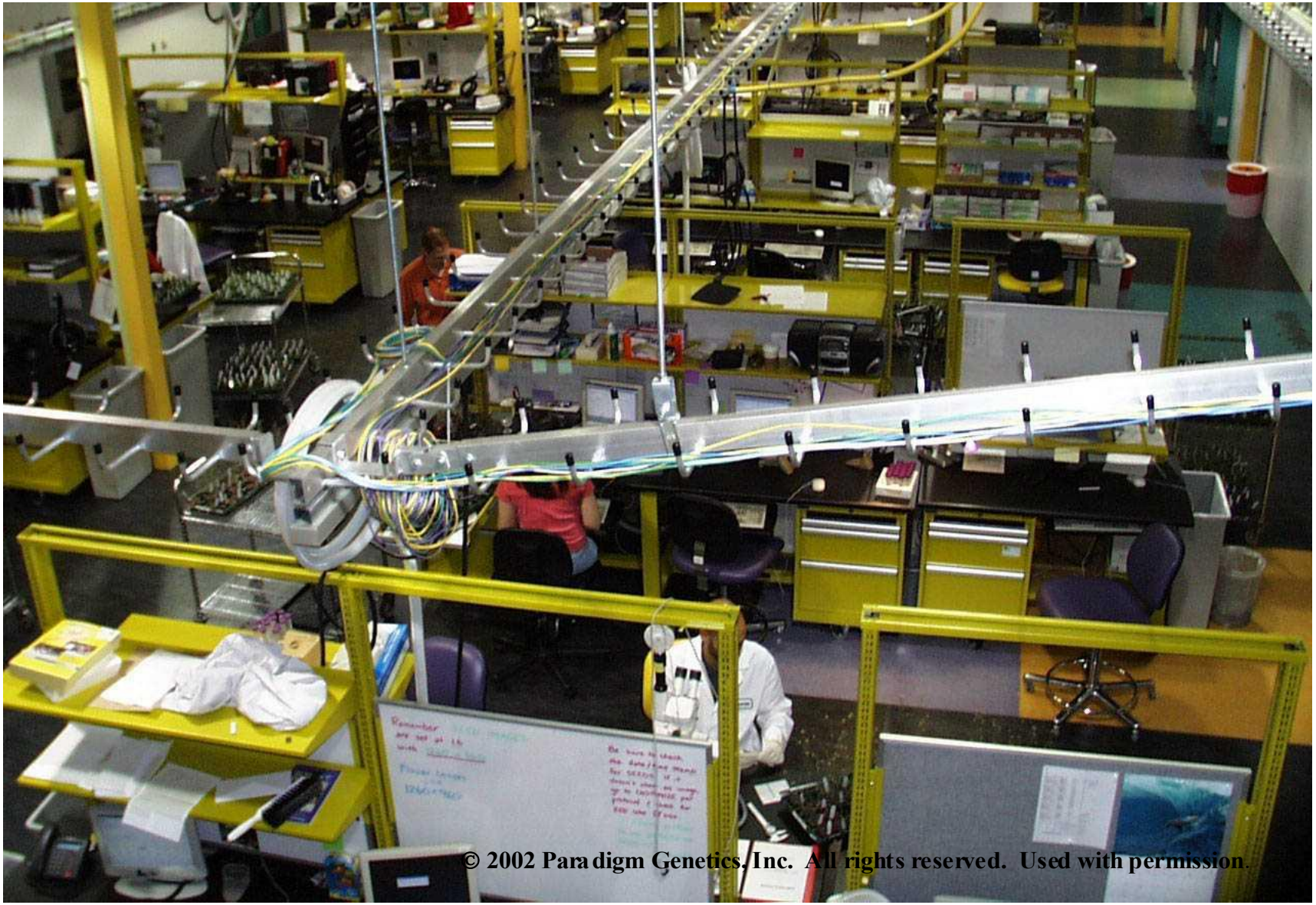
Genomics relies on high-throughput technologies

- 200 Automated sequencers
- Fluorescent dyes
- Robotics
  - Microarray spotters
  - Colony pickers
- High-throughput genetics



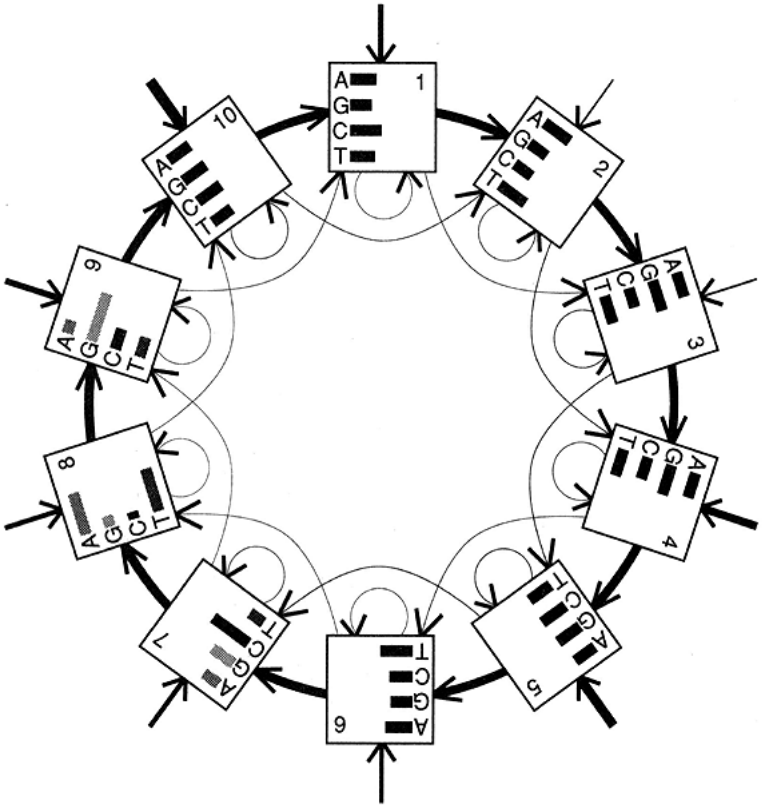
ABI3700

Industrial-scale genomics laboratory



Bioinformatics: computational analysis of genomics data

- Uses computational approaches to solve genomics problems
  - Sequence analysis
  - Gene prediction
  - Modeling of biological processes/network

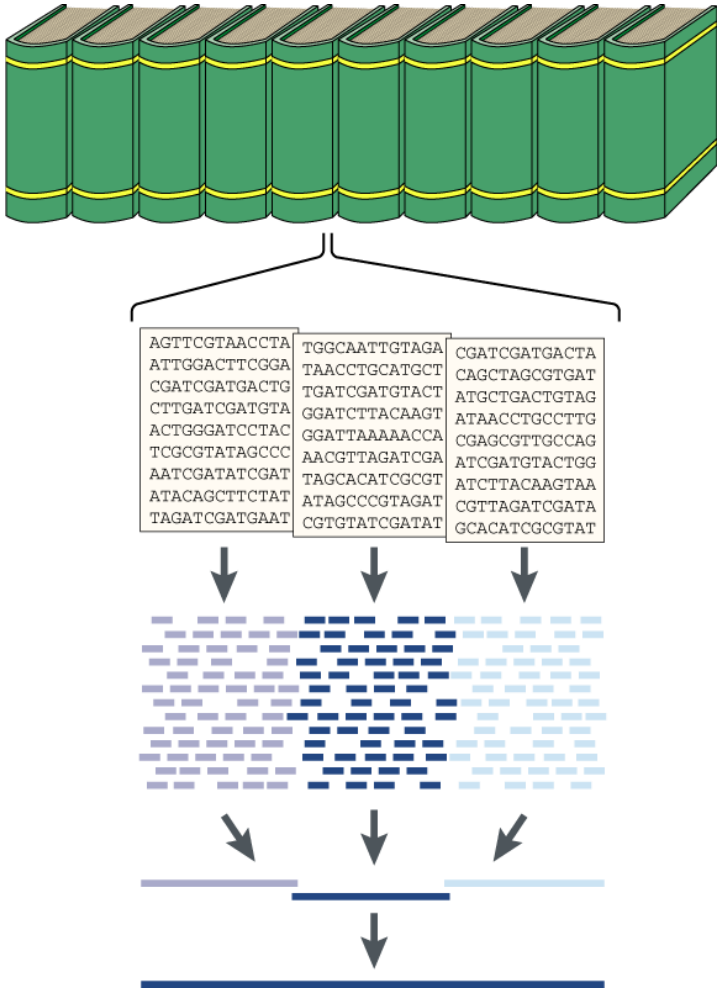


Genome sequencing

- Analogy: Complete works of an author
  - in partially understood language
- Two approaches
  - Page by page
  - All at once

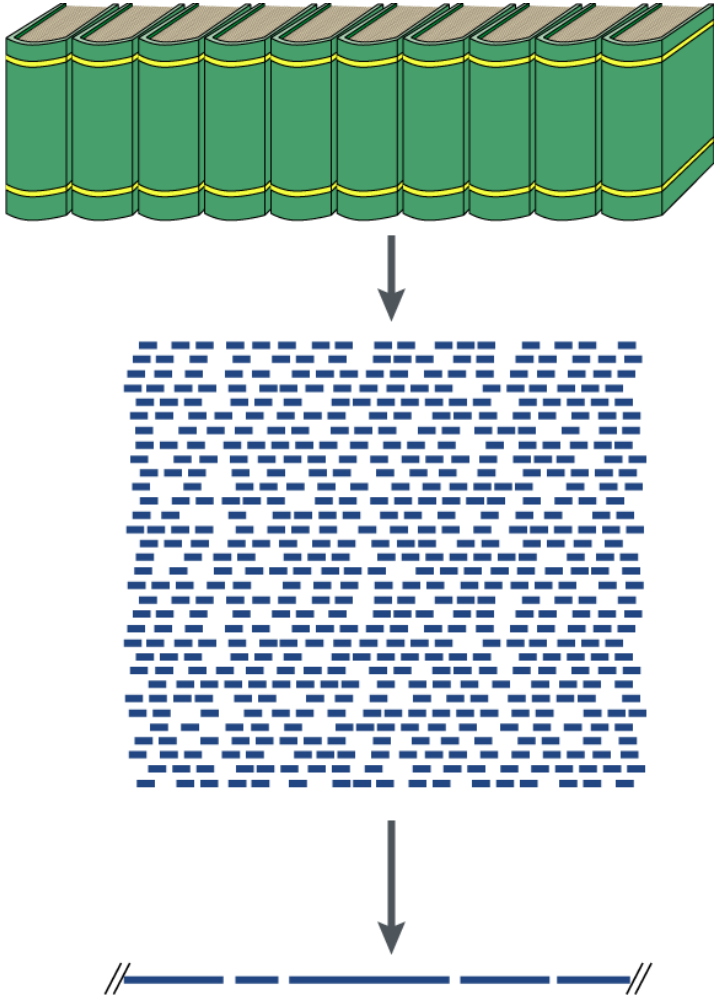
Page-by-page sequencing strategy

- Sequence = determining the letters of each word on each piece of paper
- Assembly = fitting the words back together in the correct order

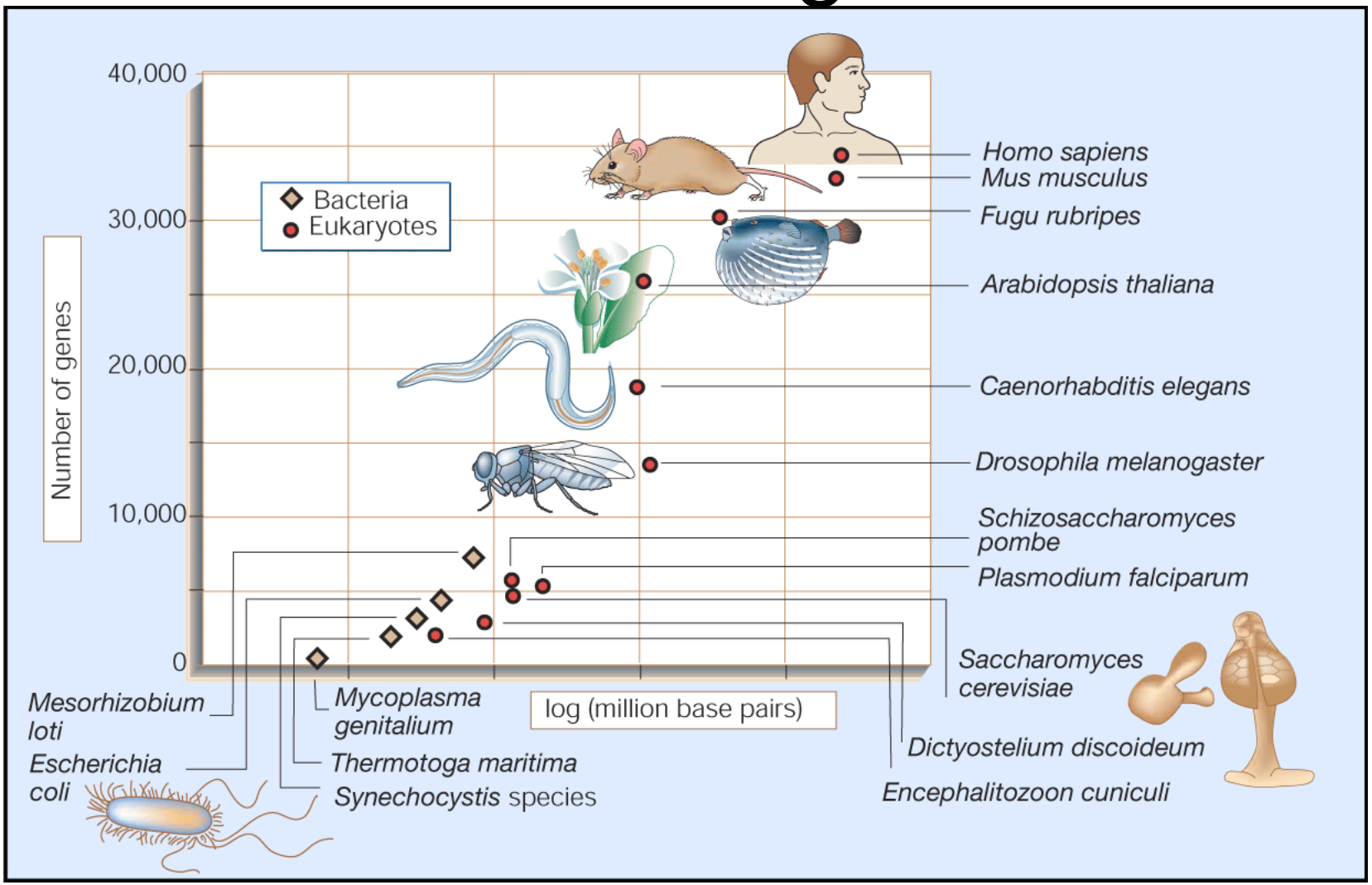


All-at-once sequencing strategy

- Find small pieces of paper
- Decipher the words on each fragment
- Look for overlaps to assemble



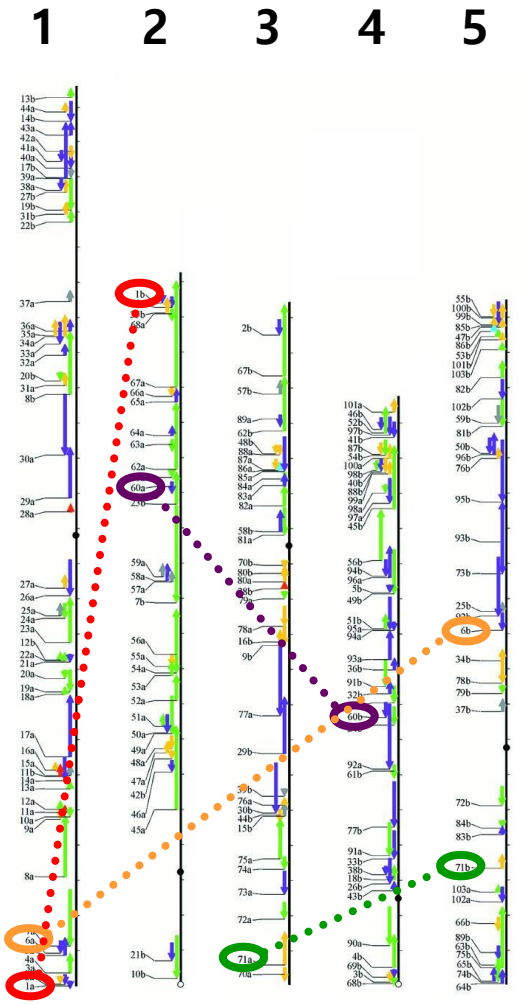
Genome size and gene number



Amoeba dubia: 670 billion base pairs

Lessons from sequencing

- Variability of genome structure: **junk**
  - Duplication events
  - Transposons
  - Microsatellites
  - Repetitive DNAs



Functional genomics

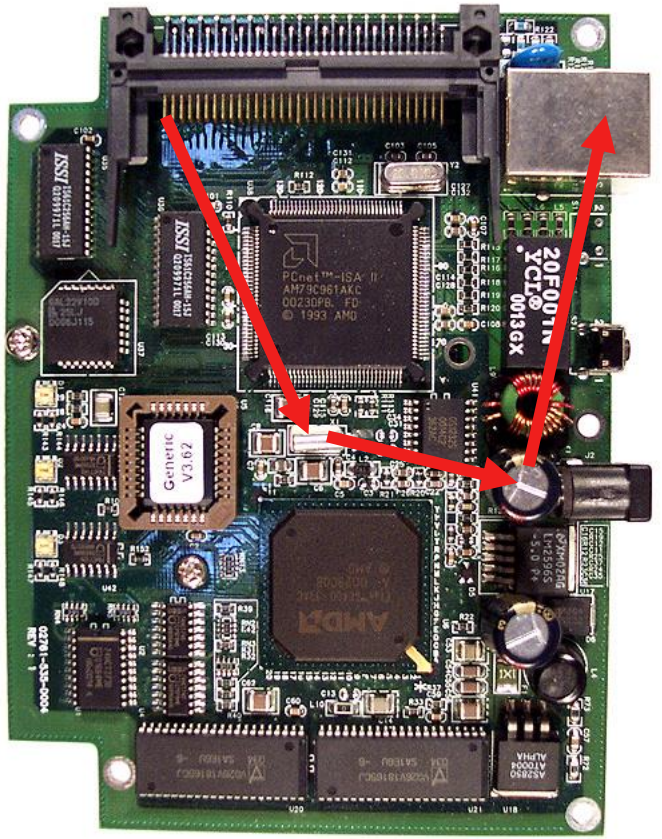
- Once we know the sequence of genes, we want to know the **function**
- The genome is the same in all cells of an individual, except for random mutations
- However, in each cell, only a subset of the genes is expressed
  - The portion of the genome that is used in each cell correlates with the cell's differentiated state

Analogy for gene expression

- Genome is a hard drive of a computer
  - Contains all the programs
- Gene expression
  - What's loaded into RAM (short-term memory)
  - Subset of genome used in each cell

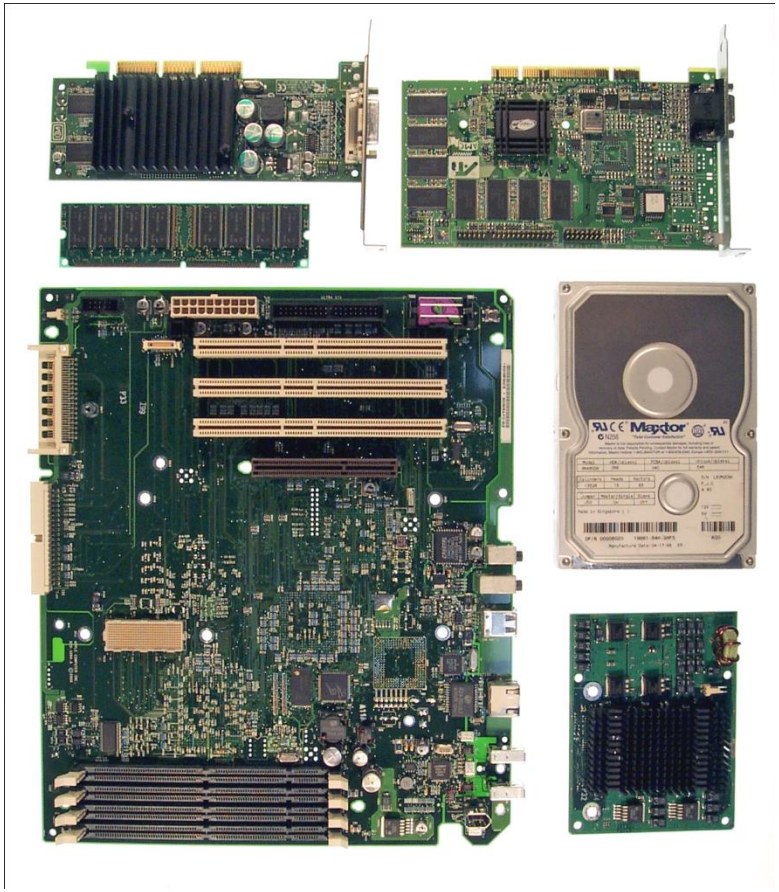
Gene-by-gene approach to understand biological processes

- Analogous to understanding circuitry by following wires
- Choose one wire
- Follow circuit to transistor
- Follow from transistor to capacitor
- Follow from capacitor to power source
- Do again



Genomics provides a parts list

- Provides list of all parts
- Parts list in itself doesn't say how the genome works
- Can use to get global picture
  - e.g., RNA expression

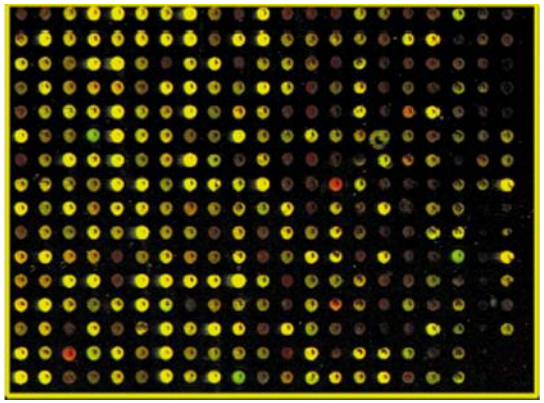


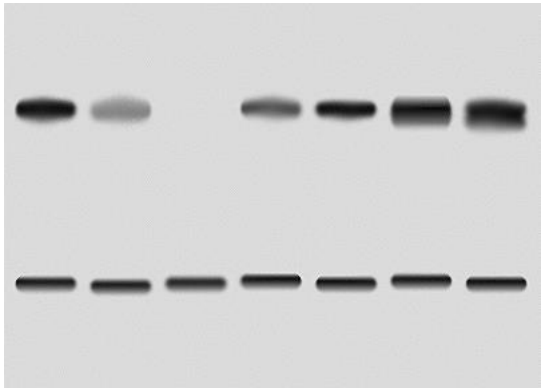
Genomics applications to biology

- Cellular function
  - Microarrays: RNA
  - Proteomics: proteins
  - Cellular networks: Metabolites
- Evolutionary mechanisms
  - Comparative genomics

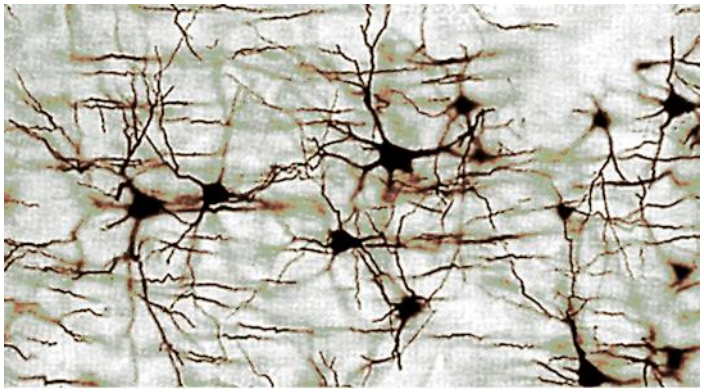
Expression microarrays

- Global expression analysis
- RNA levels of every gene in the genome analyzed in parallel
- Compare with Northern blot
  - Microarrays contain more information by many orders of magnitude

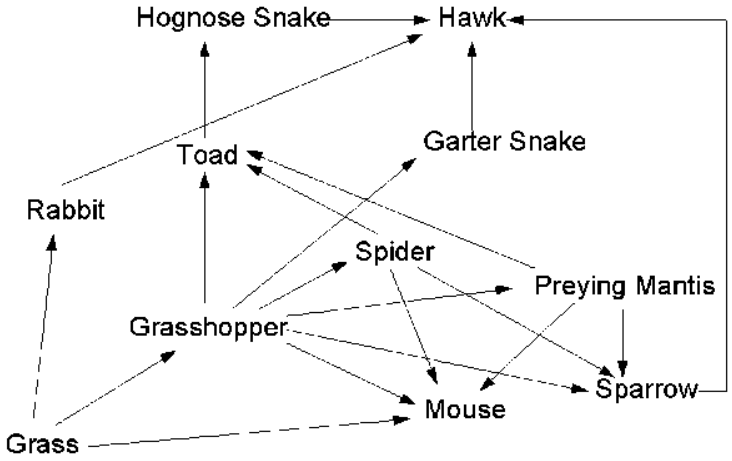




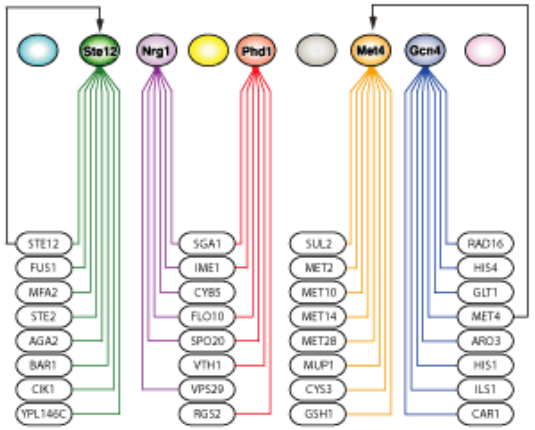
Biological networks: Systems Biology



Neuronal network



Food chain

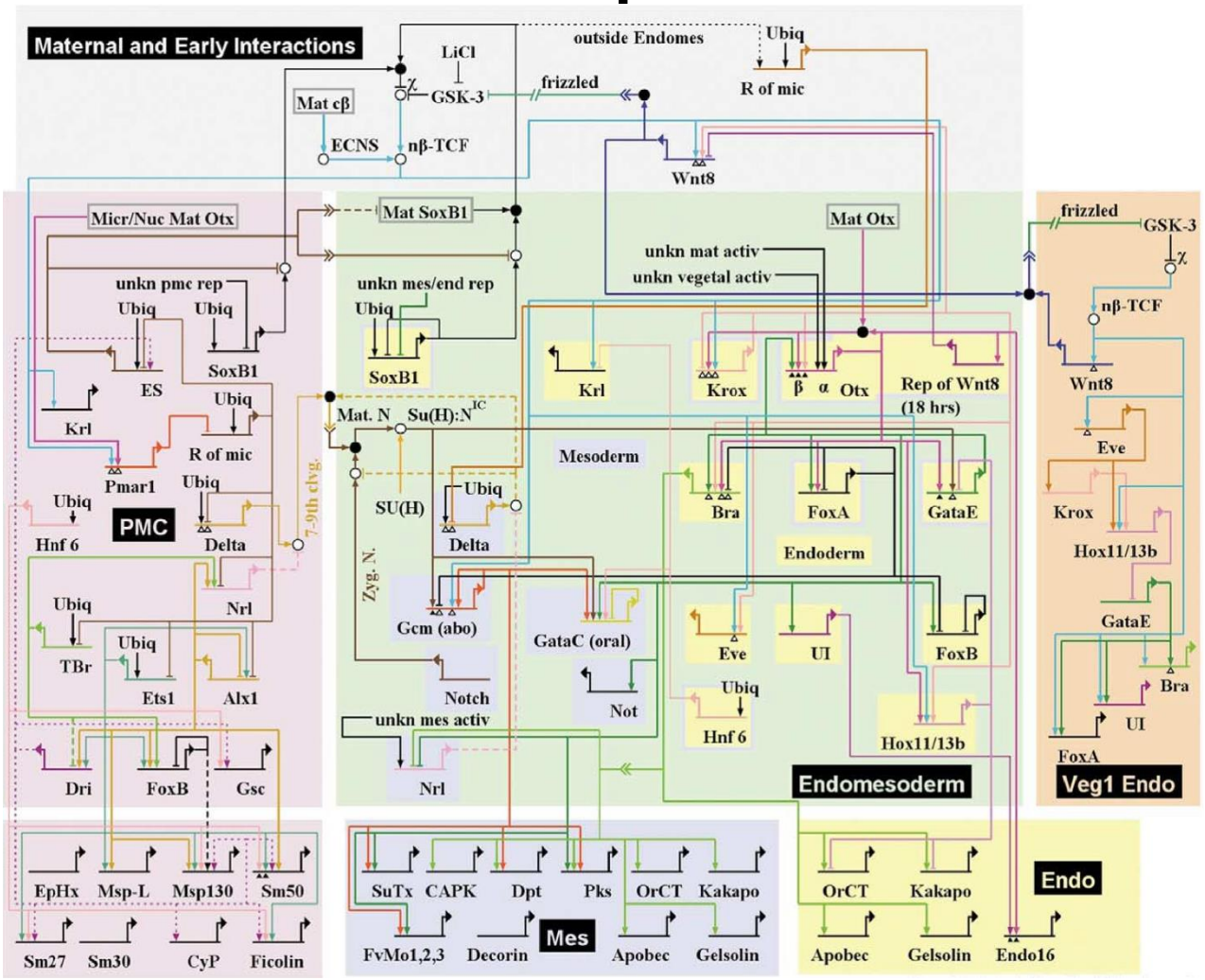


Transcriptional network

From parts to systems

- Parts list + interactions = road map
- Properties = traffic patterns
- Want to understand properties
  - Why certain traffic patterns emerge
- **Perturb** system and see how it responds
  - Place traffic light at intersection

Regulatory network of sea urchin development

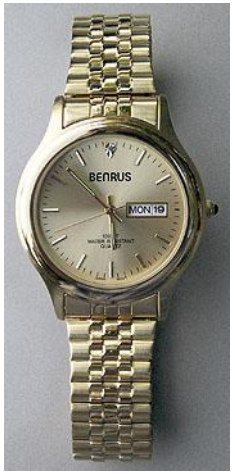


Comparative genomics

- Mechanisms of evolution
- What is conserved between species?
  - Genes for basic processes
- What makes closely related species different?
  - Their adaptive traits

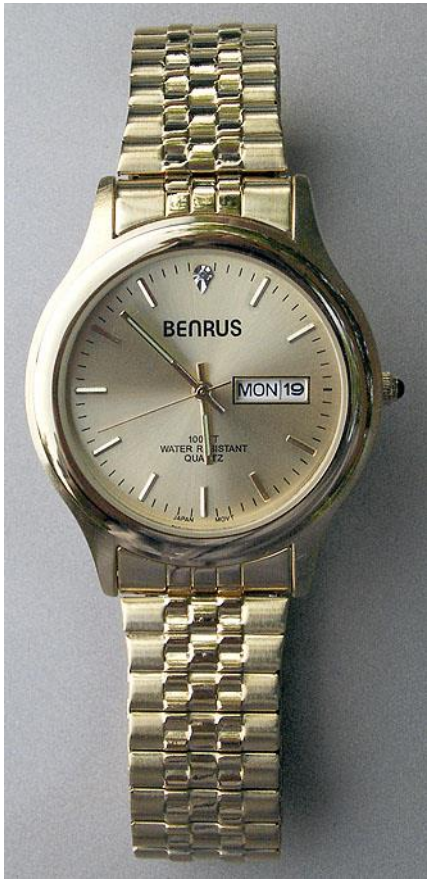
Conservation between species identifies important components

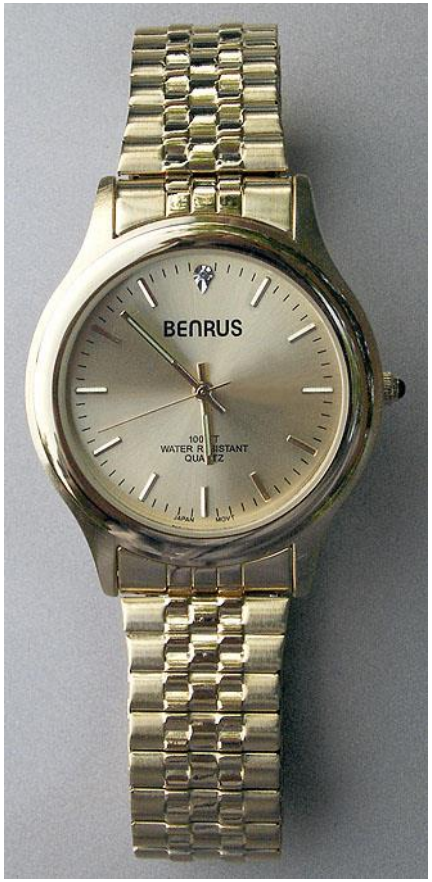
- Compare parts lists
  - Mantle clock
  - Pocket watch
  - Wristwatch
- Identify essential elements of timekeeping
  - Gears, hands, etc.
- Superfluous parts
  - Wristband



Identifying adaptive traits

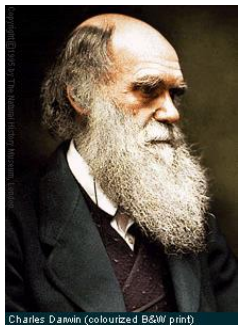
- Compare parts lists
  - Two watches from same maker: one with date, other without
- Reveals parts likely to function in date mechanism





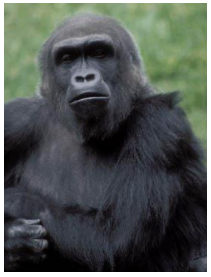
Humans and their ancestors

- All great apes have high level of cognitive ability
- But very different social behaviors



human

chimpanzee



gorilla

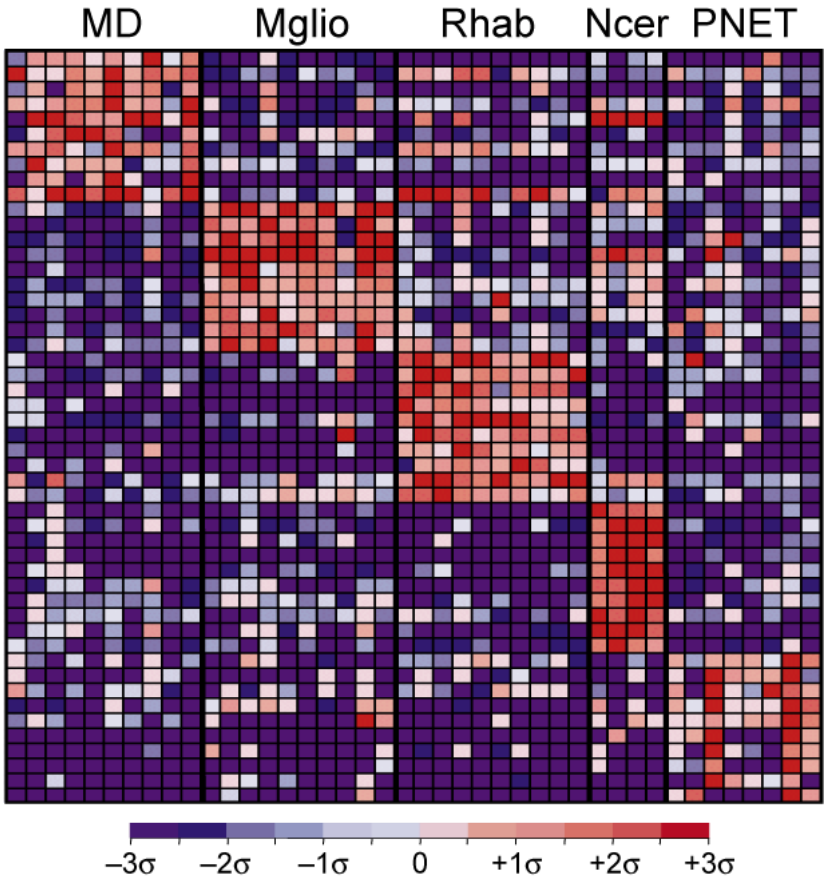
orangutan

Applications of genomics to medicine

- Genes for disease susceptibility
- Improved diagnosis
- Pharmacogenomics

Improved disease diagnostics from genomics

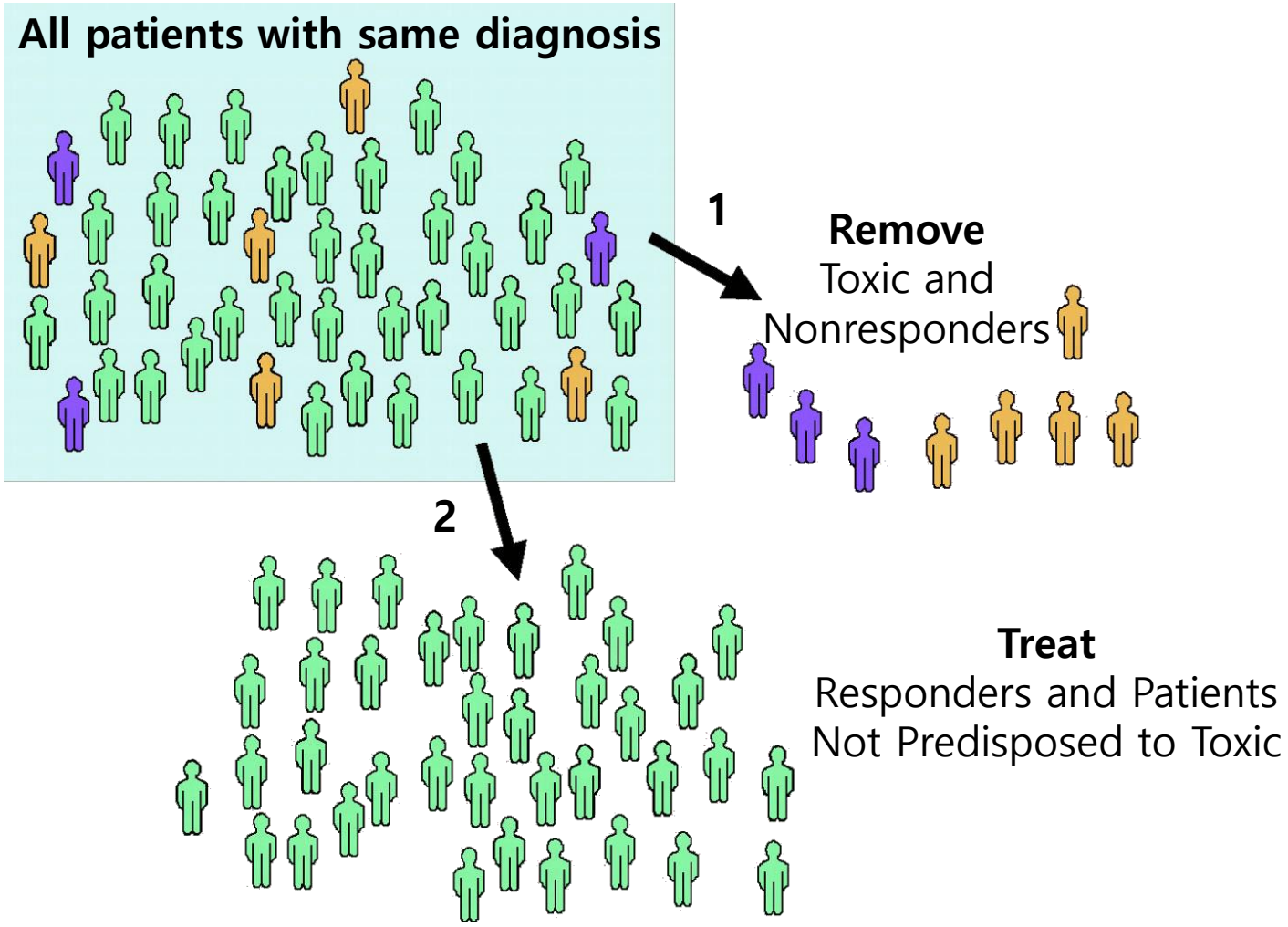
- Microarray analysis of gene expression from four different types of tumors
- Grouping of gene expression patterns shows very clear differences among the tumors
- Used to tailor therapy to individuals



Pharmacogenomics: drug therapies tailored to individuals

- Design therapies based on the individual's genome
- Subtle, but important, differences in genomes
  - Cause differences in how one responds to drugs
- Identify those who will suffer harmful side effects from particular drugs

Prescreening based on genomes



Genomics applied to agriculture

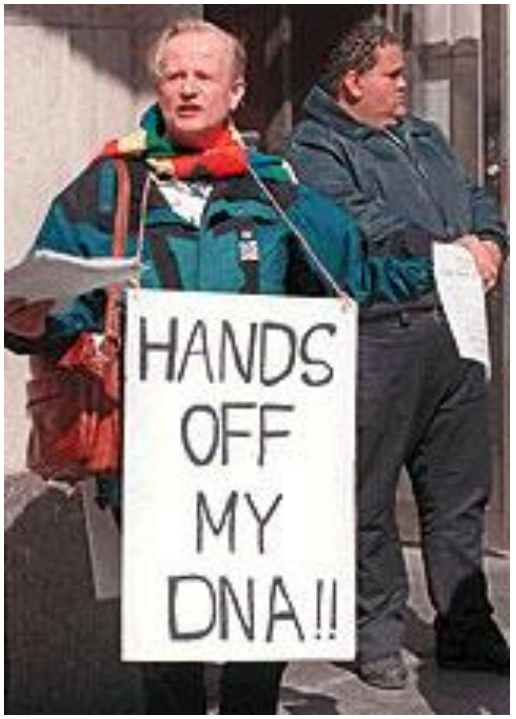
- Sequencing of crop-plant genomes
- Gene discovery for useful traits
- Genomewide regulatory networks to improve traits

Farm-animal genomics

- Genome sequencing of pigs, cows, sheep, and poultry
- EST sequencing
- Agricultural pathogens
  - Potential bioterrorism agents

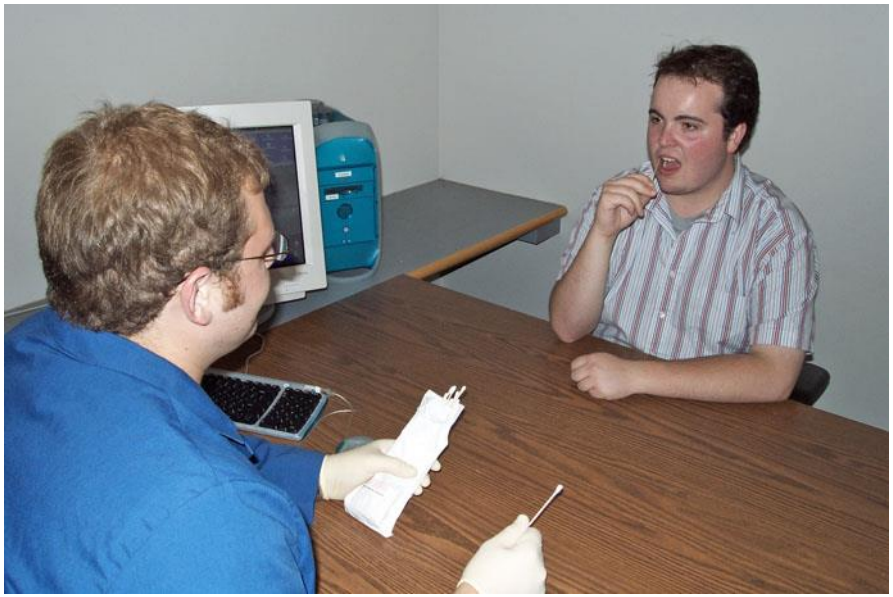
Ethical issues raised by genomics (ELSI) (Ethical legal, societal implications)

- Individual's genome holds key to disease susceptibility
- Potential for misuse recognized by founders of Human Genome Project



Genetic testing in the workplace

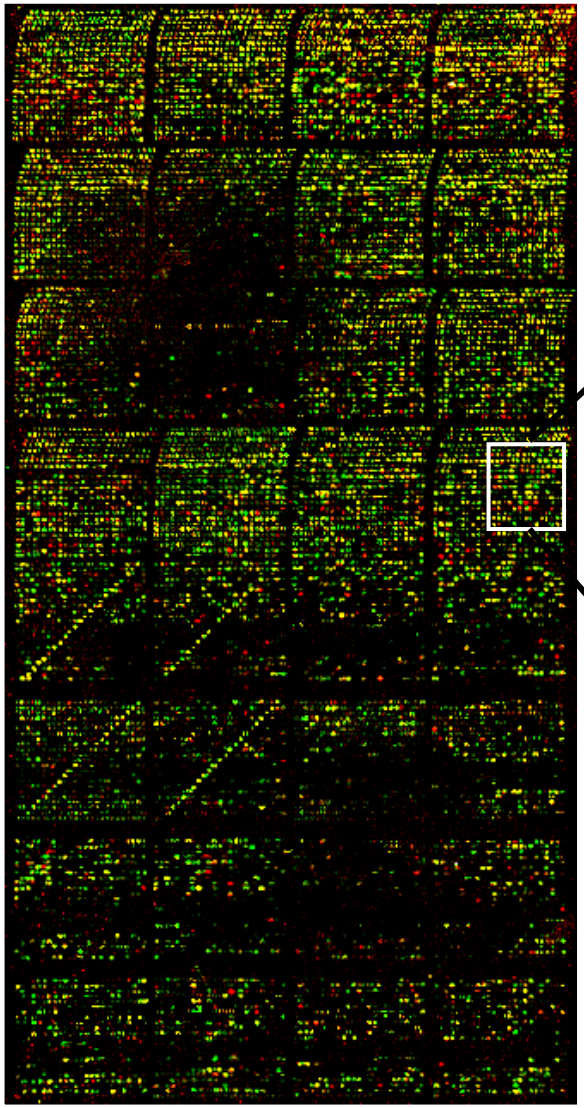
- Major railroad company decided to perform DNA tests on employees
- Wanted to identify susceptibility to carpal tunnel syndrome
- Equal Employment Opportunity Commission filed suit to block action

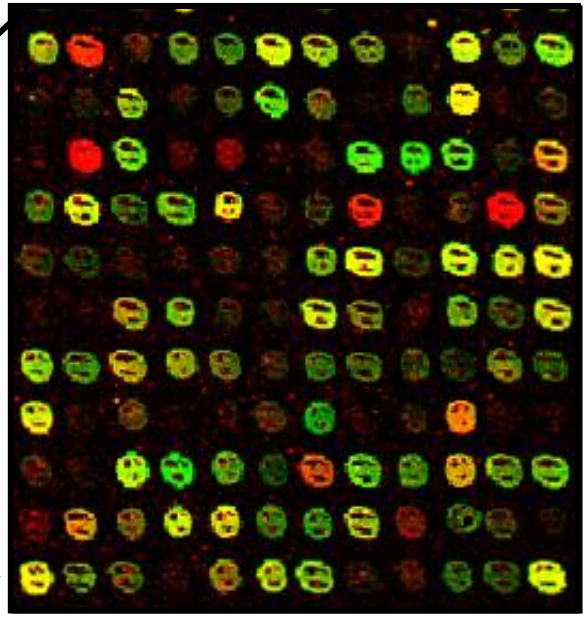


Genetic modification of humans

- Once we know the genes responsible for particular diseases, should we “cure” the diseases?
- Should we also modify genes responsible for traits such as height or beauty?
- Should we allow the cloning of human beings?

Microarray Technology

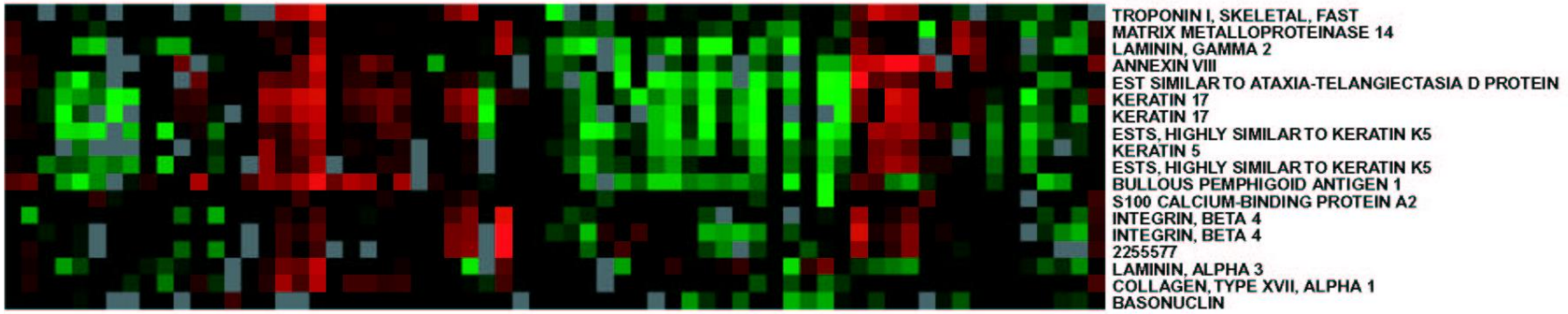




Biomarkers and gene expression

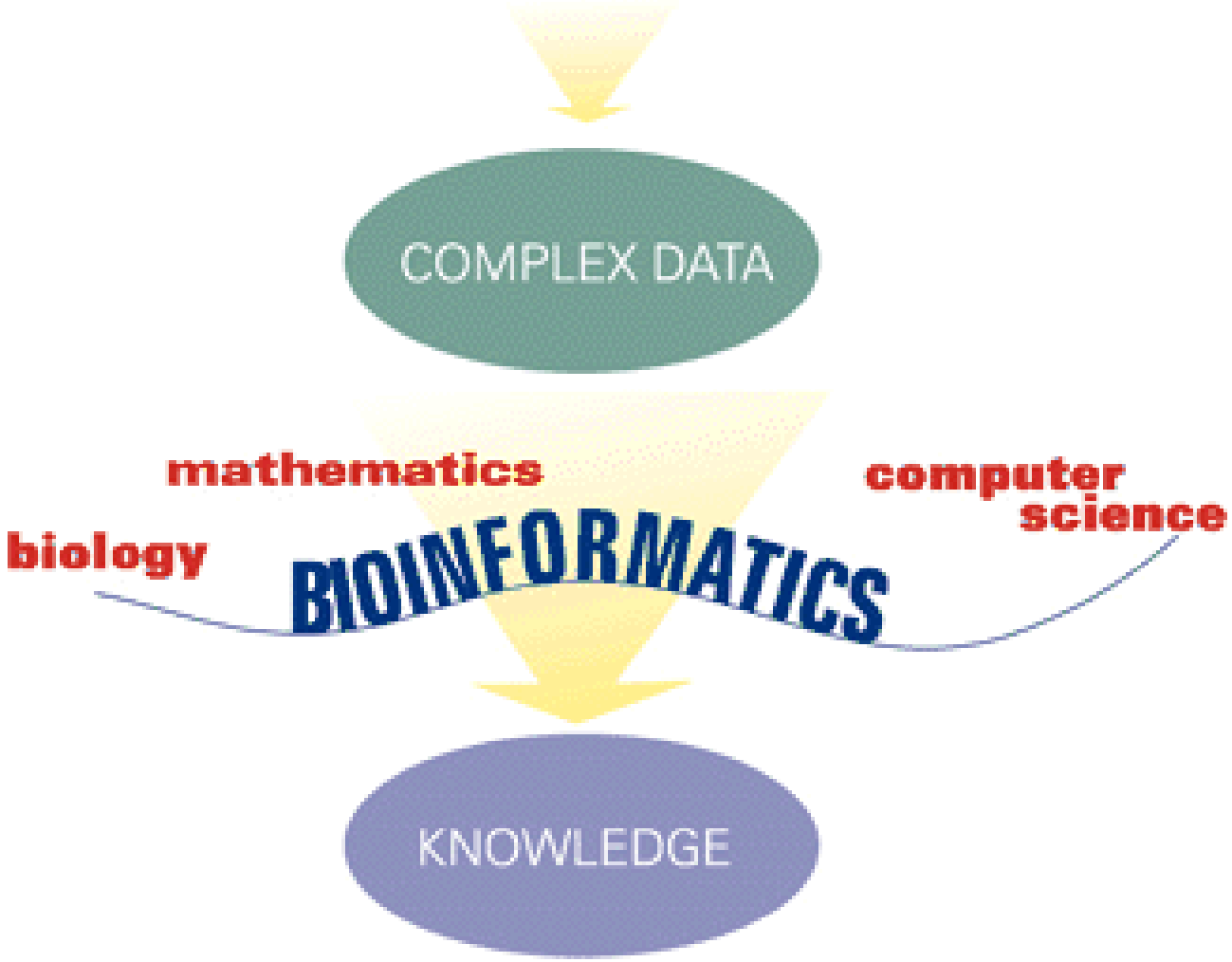
normal

malignant



Perou et al. Molecular Portraits of Breast Cancer, Nature, May 2000

What is Bioinformatics?



What is Bioinformatics?

Conceptualizing biology in terms of molecules and then applying "informatics" techniques from math, computer science, and statistics to understand and organize the information associated with these molecules on a large scale

How do we use Bioinformatics?

- Store/retrieve biological information (databases)
- Retrieve/compare gene sequences
- Predict function of unknown genes/proteins
- Search for previously known functions of a gene
- Compare data with other researchers
- Compile/distribute data for other researchers

Sequence retrieval:

National Center for Biotechnology Information

GenBank and other genome databases

Sequence comparison programs:

BLAST **GCG** **MacVector**

Protein Structure:

3D modeling programs – RasMol, Protein Explorer

1: L28171 **Simian immunodeficiency virus glycoprotein 120 (gp120), glycoprotein 41 (gp41 rev and nef genes, partial cds's**

```
LOCUS       SIVNEF       2206 bp    DNA             VRL       02-JUN-2000
DEFINITION  Simian immunodeficiency virus glycoprotein 120 (gp120),
            glycoprotein 41 (gp41), tat, rev and nef genes, partial cds's.
ACCESSION   L28171
VERSION     L28171.1  GI:506842
KEYWORDS    glycoprotein 120; glycoprotein 41; gp120 gene; gp41 gene; nef gene;
            replication enhancer; rev gene; tat gene.
SOURCE      Simian immunodeficiency virus.
  ORGANISM  Simian immunodeficiency virus
            Viruses; Retroid viruses; Retroviridae; Lentivirus; Primate
            lentivirus group.
REFERENCE   1  (bases 1 to 2206)
  AUTHORS   Kodama,T., Kawahara,T. and Desrosiers,R.C.
  TITLE     The Nef gene of neurotropic SIV functions as replicative enhancer
            in microglial cells
  JOURNAL   Unpublished
     gene             1004..1795
                      /gene="nef"
     CDS              1004..1795
                      /gene="nef"
                      /codon_start=1
                      /protein_id="AAA47715.1"
                      /db_xref="GI:506845"
                      /translation="MGGAISMRRSRPSGDLRQRLLRARGETYGRLLGEVEDGYSQSPG
                      GLDKGLSSLSCEGQKYNQGQYMNTPWRNPAEKREKLAYRKQNMDDIDEEDNDLVGVSV
                      WPRVPLRTMSYKLAIDMSHFIKEKGGLEGIYYSERRHRILDIYLEKEEGIIPDWQDYT
                      SGPGIRYPKTFGWLWKLVPVNVSDEAQEDEEHCLIHPAQTSQWDDPWGEVLAWKFDPT
                      LAYTHEAYVRYPEEFGSKSGLSEEEVRRRLTARGLLNMADKKETR"
     LTR              1389..2206
                      /note="3'LTR"
     LTR              1389..1905
                      /note="U3"
     LTR              1390..2082
                      /note="R"
     LTR              2083..2206
                      /note="U5"
BASE COUNT      650 a    461 c    588 g    507 t
ORIGIN
        1 cgattggaat tgggagatta taaattagta gagatcactc caattggctt ggcccccaca
       61 aatgtgaaga ggtacactac tggtggcacc tcaagaaata aaagaggggt ctttgtgcta
      121 gggttcttgg gttttctcgc aatggcaggt actcgaatgg gcgcggcgtc gttgacgctg
      181 accgctcagt cccggacttt attggctggg atagtgcagc aacagcaaca gctgttggac
      241 gtggtcaaga gacaacaaga attgttgcga ctgaccgtct ggggagtaaa gaacctccag
      301 actagggtca ctgccatcga gaagtactta aaggaccagg cgcagctaaa tgcttggggg
```

Similarity Search: BLAST

A tool for searching gene or protein sequence databases for related genes of interest

Alignments between the query sequence and any given database sequence, allowing for mismatches and gaps, indicate their degree of similarity

The structure, function, and evolution of a gene may be determined by such comparisons

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

% identity

```
CATTATGATA
   |||||||
GTTTATGATT
```

70%

```
MRCKTETGAR
|||  |||||
MRCGTETGAR
```

90%

Strengths:

- **Accessibility**
- **Growing rapidly**
- **User friendly**

Weaknesses:

- **Sometimes not up-to-date**
- **Limited possibilities**
- **Limited comparisons and information**
- **Not accurate**

Need for improved Bioinformatics

Genomics:

Human Genome Project
Gene array technology
Comparative genomics
Functional genomics

Proteomics:

Global view of protein function/interactions

Protein motifs

Structural databases

Data Mining

Handling enormous amounts of data

Sort through what is important and what is not

Manipulate and analyze data to find patterns and variations that correlate with biological function

Proteomics

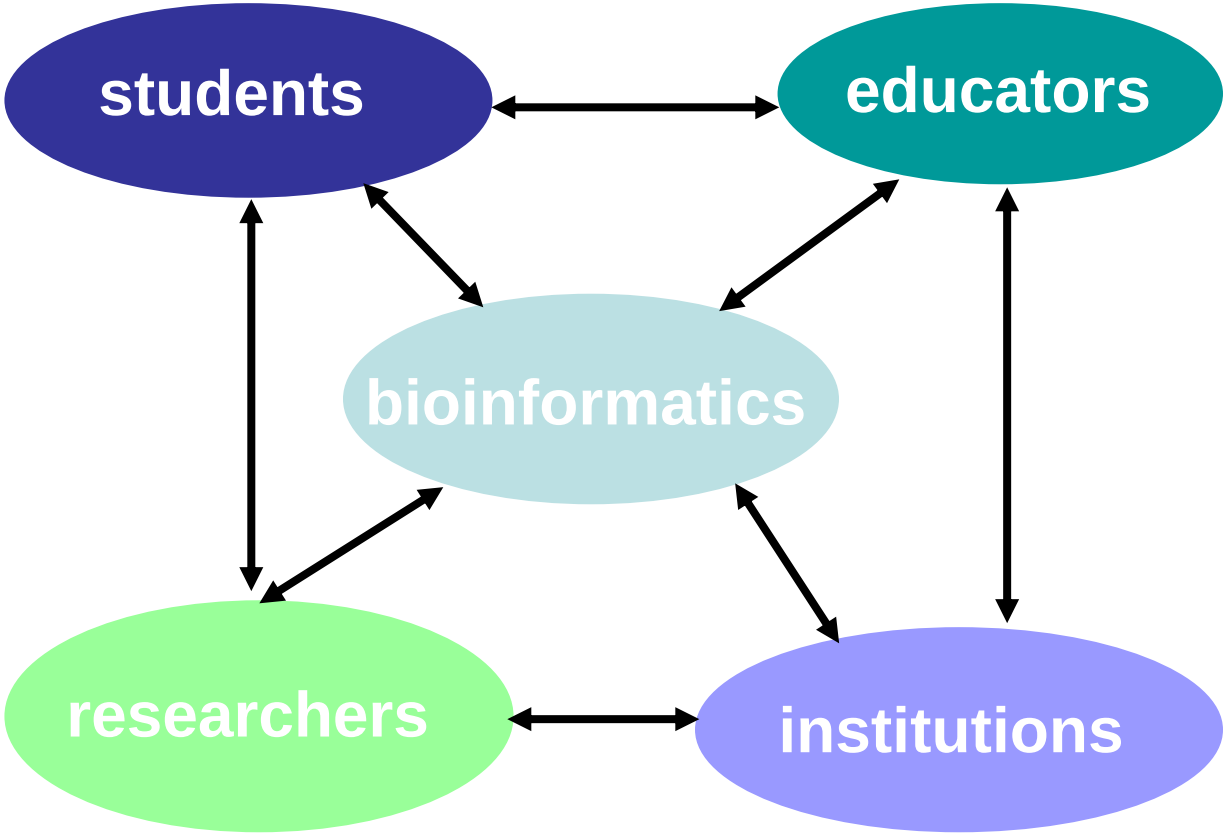
- **Uses information determined by biochemical/crystal structure methods**
- **Visualization of protein structure**
- **Make protein-protein comparisons**
- **Used to determine:**

  conformation/folding

  antibody binding sites

  protein-protein interactions

  computer aided drug design


students
educators
bioinformatics
researchers
institutions

Future of sequencing
We have the genome! What's next?
(post genome era)

- Sequencing costs
  - Dropping each year
  - Could go down to $1,000/genome
- Opens possibility of sequencing genomes of individuals
- Greatly facilitates comparative genomics