

# Some Contributions of Statistics to the Genomics Revolution

Naomi Altman

Dept. of Statistics

Bioinformatics Consulting Center

PSU

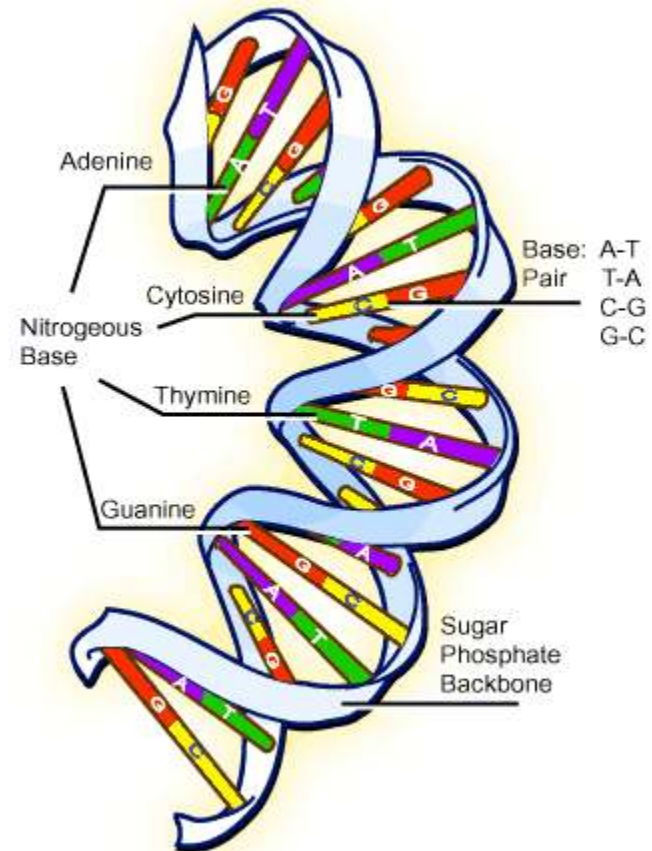
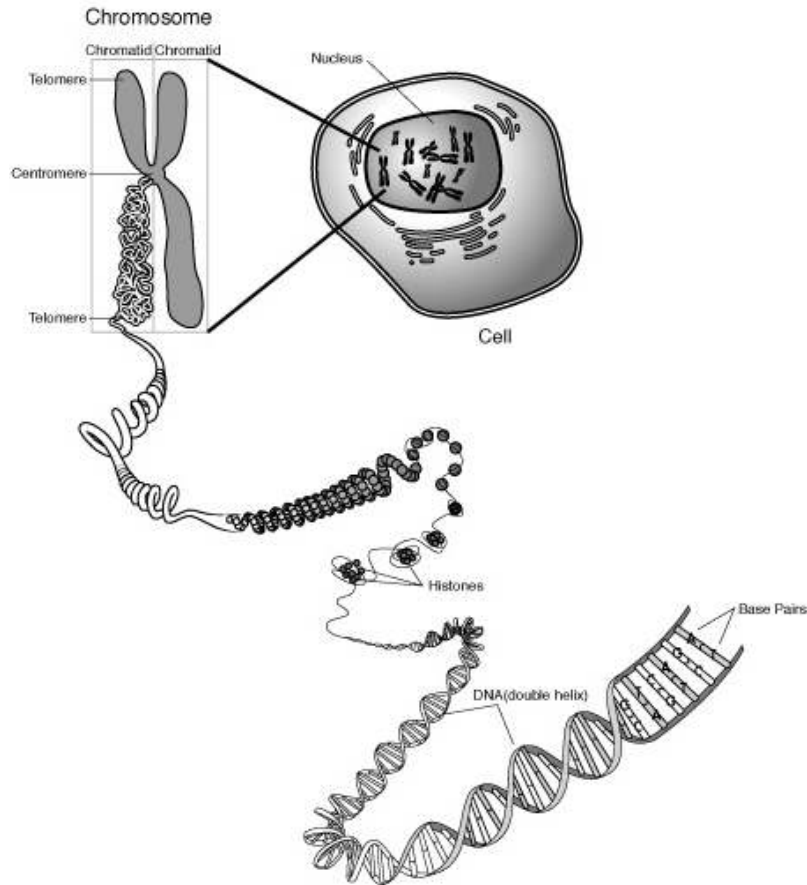
Oct. 18, 2004

Genome – the set of all genes in an organism (and the DNA that regulates it)

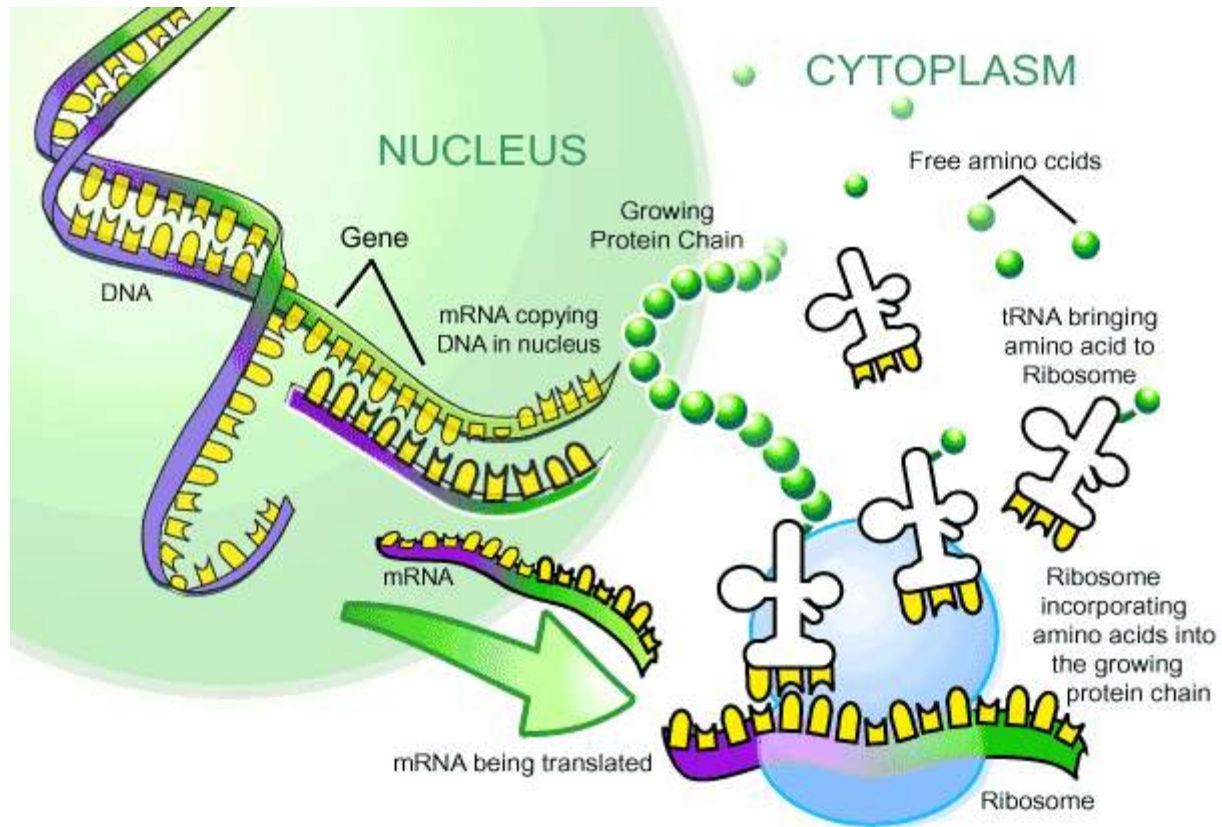
Proteome – the set of all proteins in an organism

High throughput data – data that are generated by techniques that involve a large number of genes, DNA, RNA, proteins or other biological components

# The Structure of DNA



# From DNA to Protein



# What is the Genomics Revolution?

A set of biological tools that enable biologists to

- Produce high throughput data such as
  - The genetic sequence of genes
  - The amino acid sequence of proteins
  - The amount of expression of each gene in a tissue
- Manipulate the genome in various ways

# The genomics wish-list

- What is the function of the DNA?
- Which DNA comprises genes?
- How does a cell regulate protein production?
- What is the function of each gene?
- What is the difference between a normal and diseased cell (especially cancer)?
- How did organisms evolve?

# Bioinformatics

The application of computational tools to genomic and proteomic information.

This includes data storage, summary, visualization and linking, as well as statistical inference.

**BIOLOGISTS ARE DROWNING IN DATA!!**

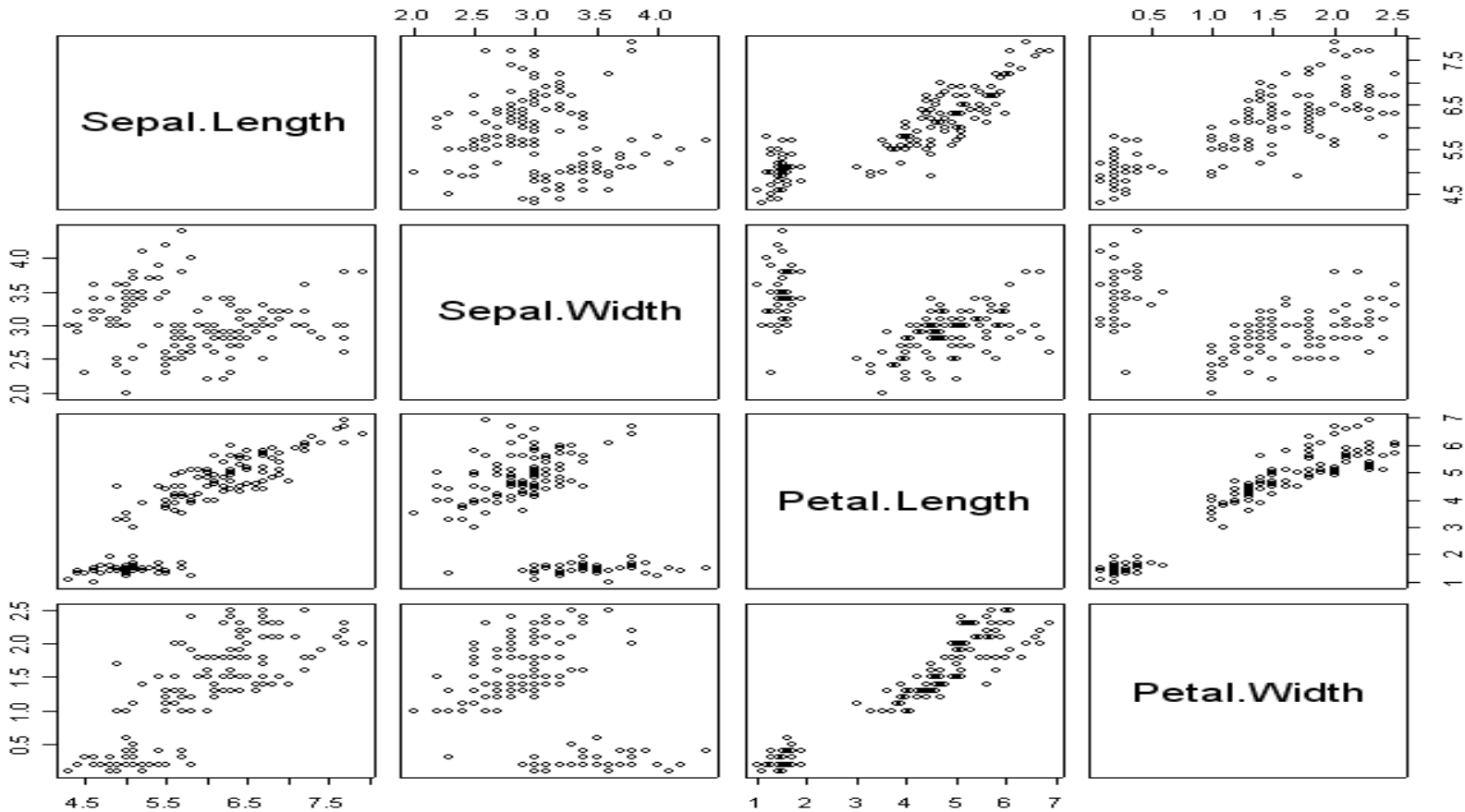
# The data explosion

“It took a year to isolate and sequence one gene from a red sponge when he began in the early 1980s...Within a few years, he was able to sequence 10 to 15 genes a year. Today, he can do 1000 overnight.”

From: This is your ancestor,  
*Discover Magazine*, Nov.  
2004.

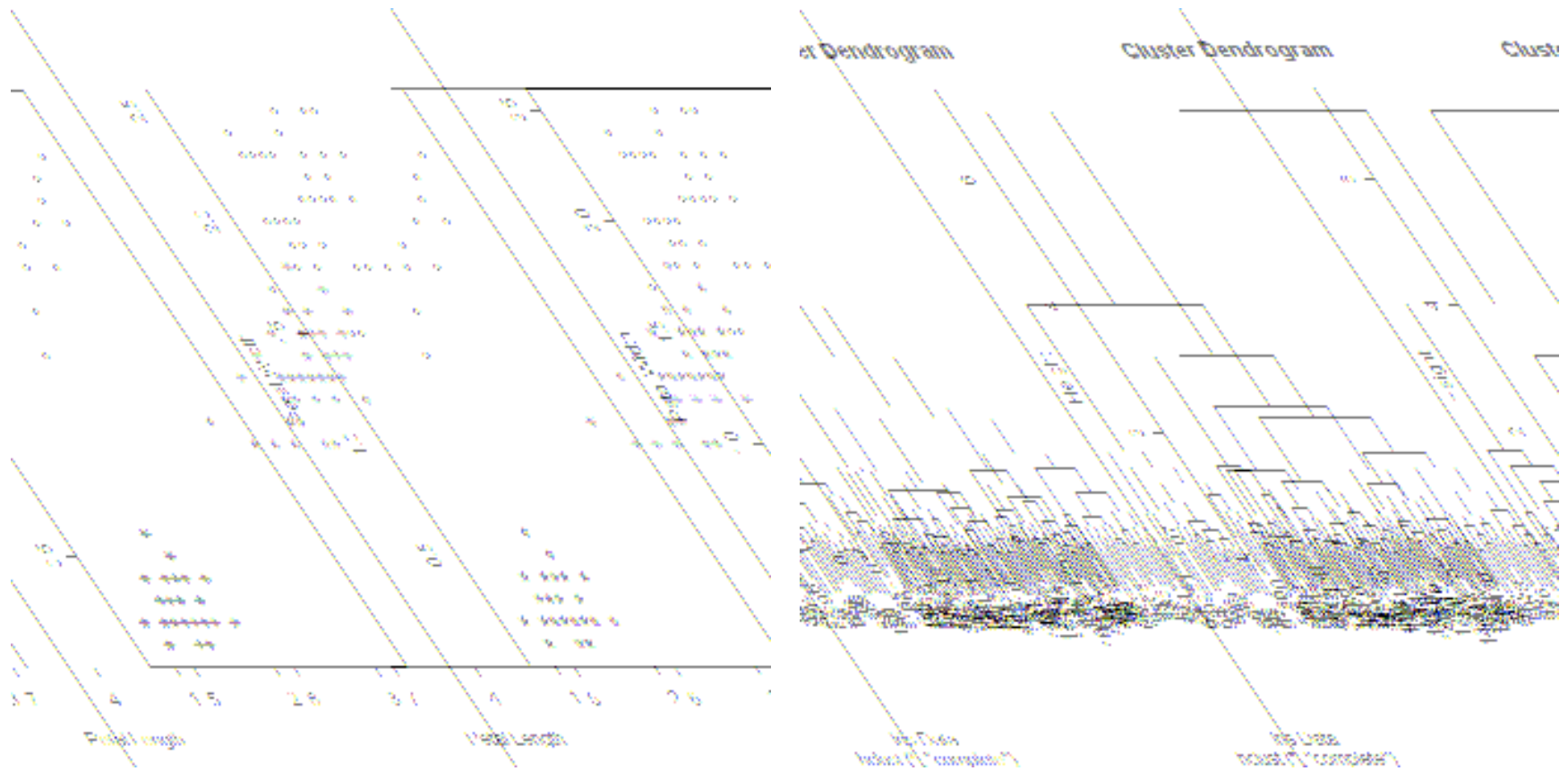


# Clustering – a useful statistical idea



Fisher Iris Data – measurements on 3 species of iris

# The 3 species



# Some Uses of Clustering In Genomics

## Genomic (Proteomic) Sequence

- Organisms which are more closely related have genes which are more similar
- Proteins which have similar function are more similar

## Gene Expression

- Tissues which have similar gene expression are more similar
- Genes which have similar expression patterns over differing conditions work together

# Similarity of Genomic Sequence

FFHPLECEPTLQMGFHSQIS - VAA - - - AGPS - - VNNN - - -  
FFHPLDCGPTLQMGYPSDSLTAEEAASVAGPS - - C - - S - - -  
FFHPLECEPTLQIGYQPDPIT - VAA - - - AGPS - - VN - NYMP  
FFHPIECEPTLQMGYQQDQIT - VAAA - - AGPSMTMN - S - - -  
FFQHIECEPTLHIGYQPDQIT - VAA - - - AGPS - - MN - NYMQ  
FFHPLECEPTLQIGYQHDQIT - IAA - - - PGPS - - VS - NYMP

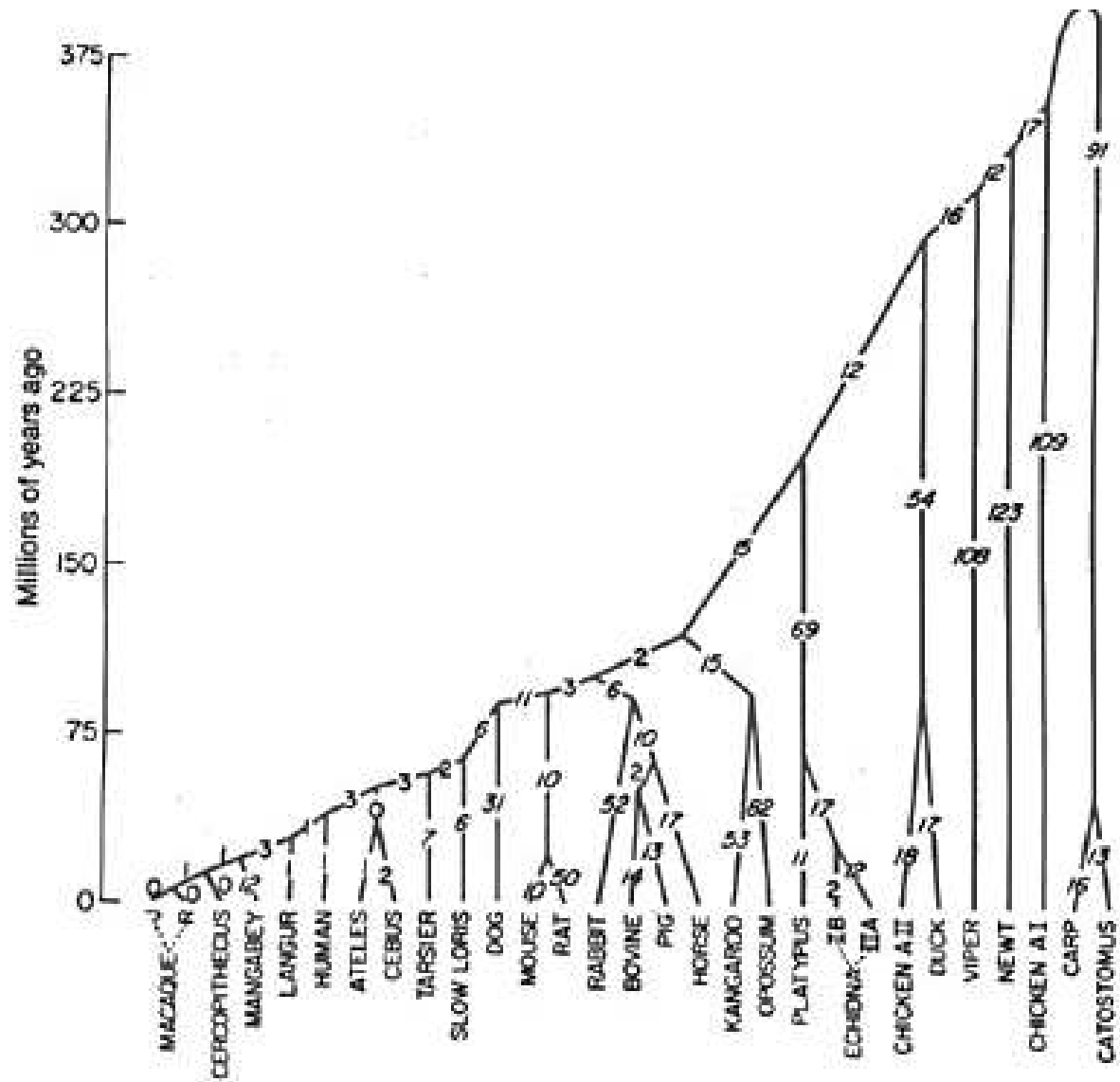
Each row represents a protein from a different species.

Each letter represents an amino acid.

Each – represents a space which is missing in this sequence but has something in it in a different species

In closely related species, the distance between genes is the number of mismatches.

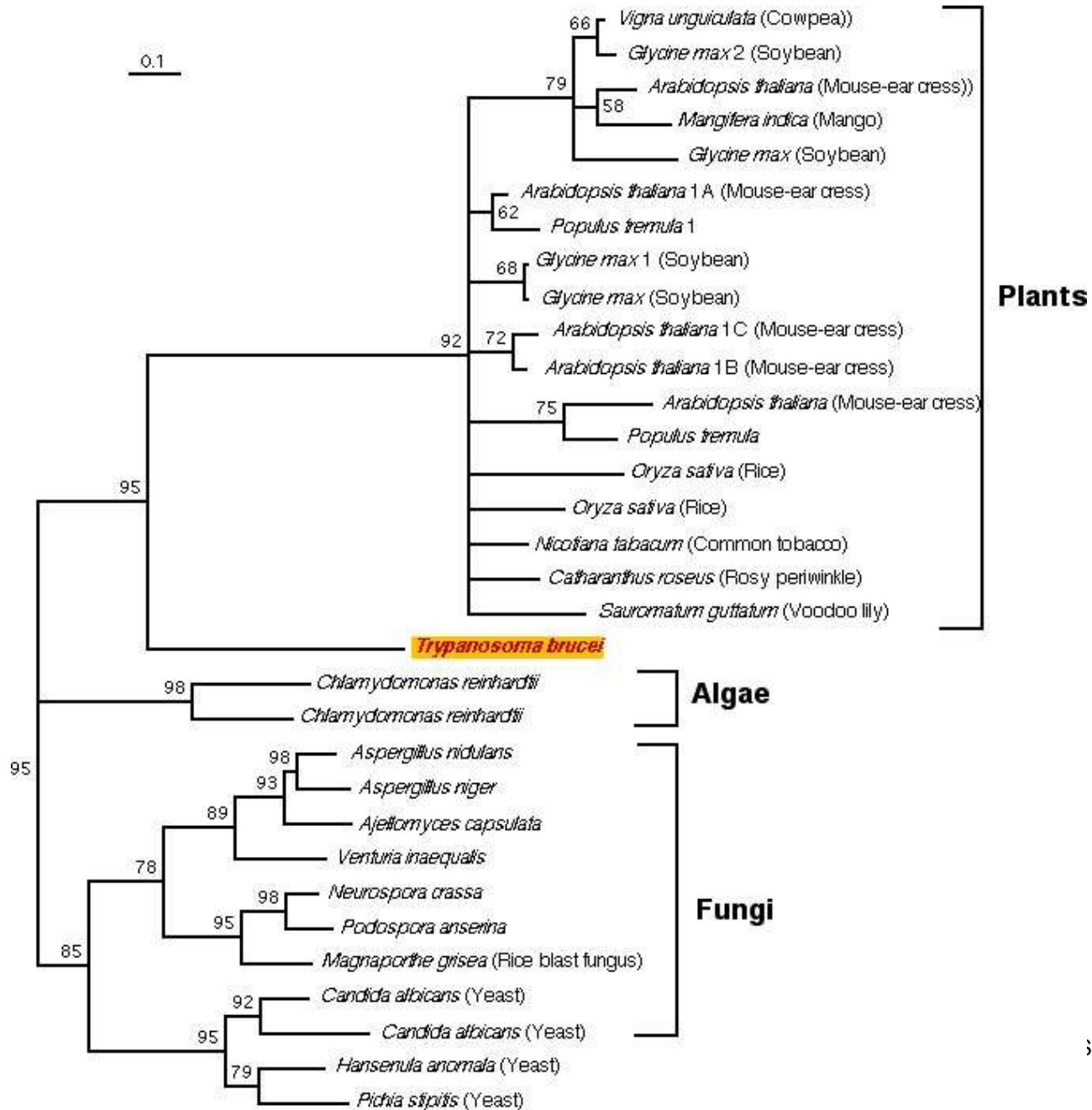
In distantly related species, the sequences are given a score – often the probability that a random sequence matches as well.



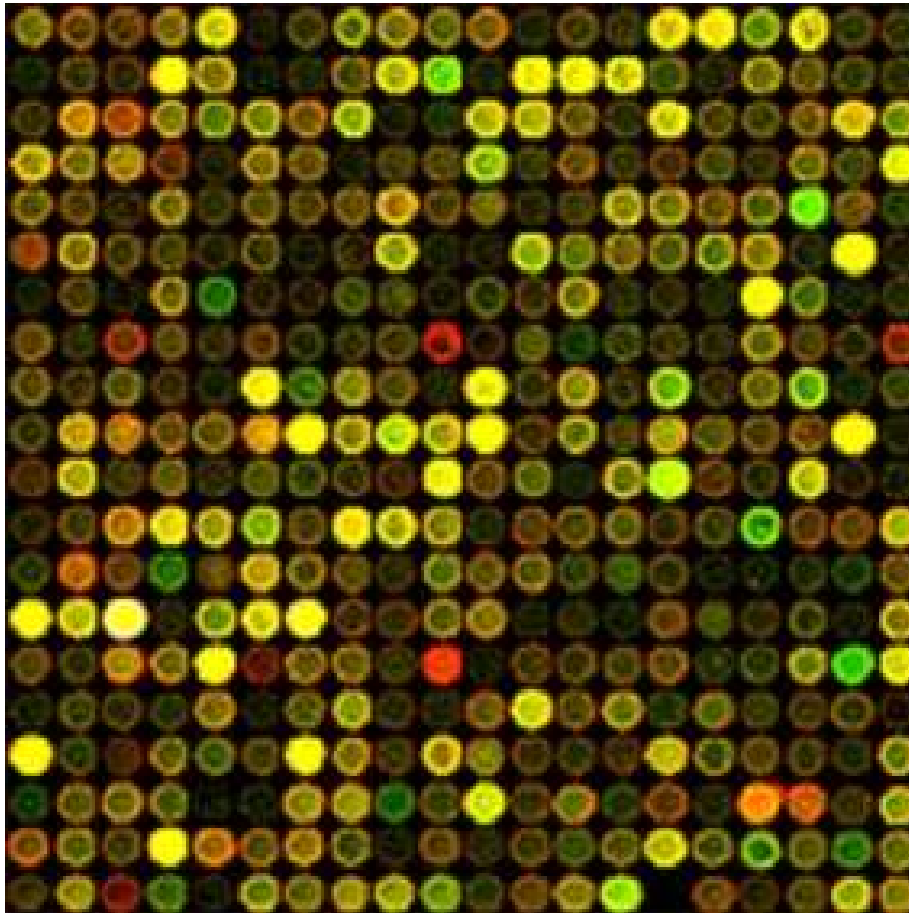
**Figure 12-6**

*A phylogeny of  $\alpha$ -hemoglobin chains in a variety of vertebrate species, determined by calculating the number of nucleotide replacements that account for the number of observed amino acid substitutions presumed to have occurred during each evolutionary interval. The vertical scale, given in millions of years, is based on paleontological estimates for the age of the common ancestor of each branch. (From Goodman 1976.)*

# Alternative oxidase



# Similarity of Expression



Each circle is a dot of a strand of DNA on the surface of a microscope slide from 1 gene.

2 RNA samples from different tissues are processed and allowed to attach (hybridize) to the spots.

1 sample is labeled with RED dye and the other with GREEN.

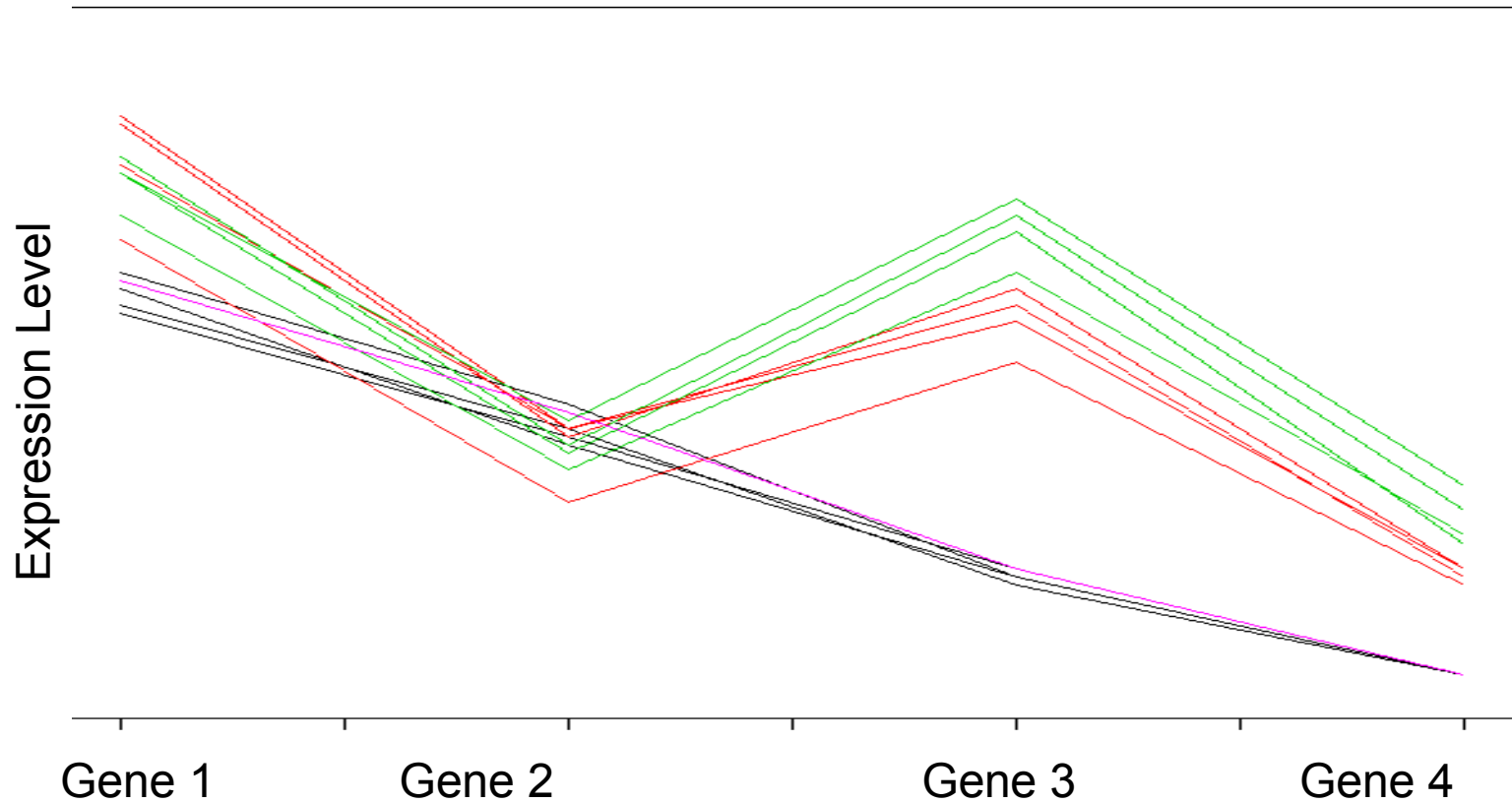
The dye intensity is proportional to the amount of RNA in the sample.

RED implies the gene is expressing more highly from the RED sample.

Yellow implies about the same expression in both samples.

Green implies the gene is expressing more highly from the GREEN sample.

# Similarity of Tissues



Each line represents RNA from 1 tissue sample. Pink line is unknown tissue. Distance is Euclidean distance, or 1-correlation, or ...

# Clustering Tissues

e.g. Normal, benign tumors, cancerous tumors

Take RNA samples from each type of tissue

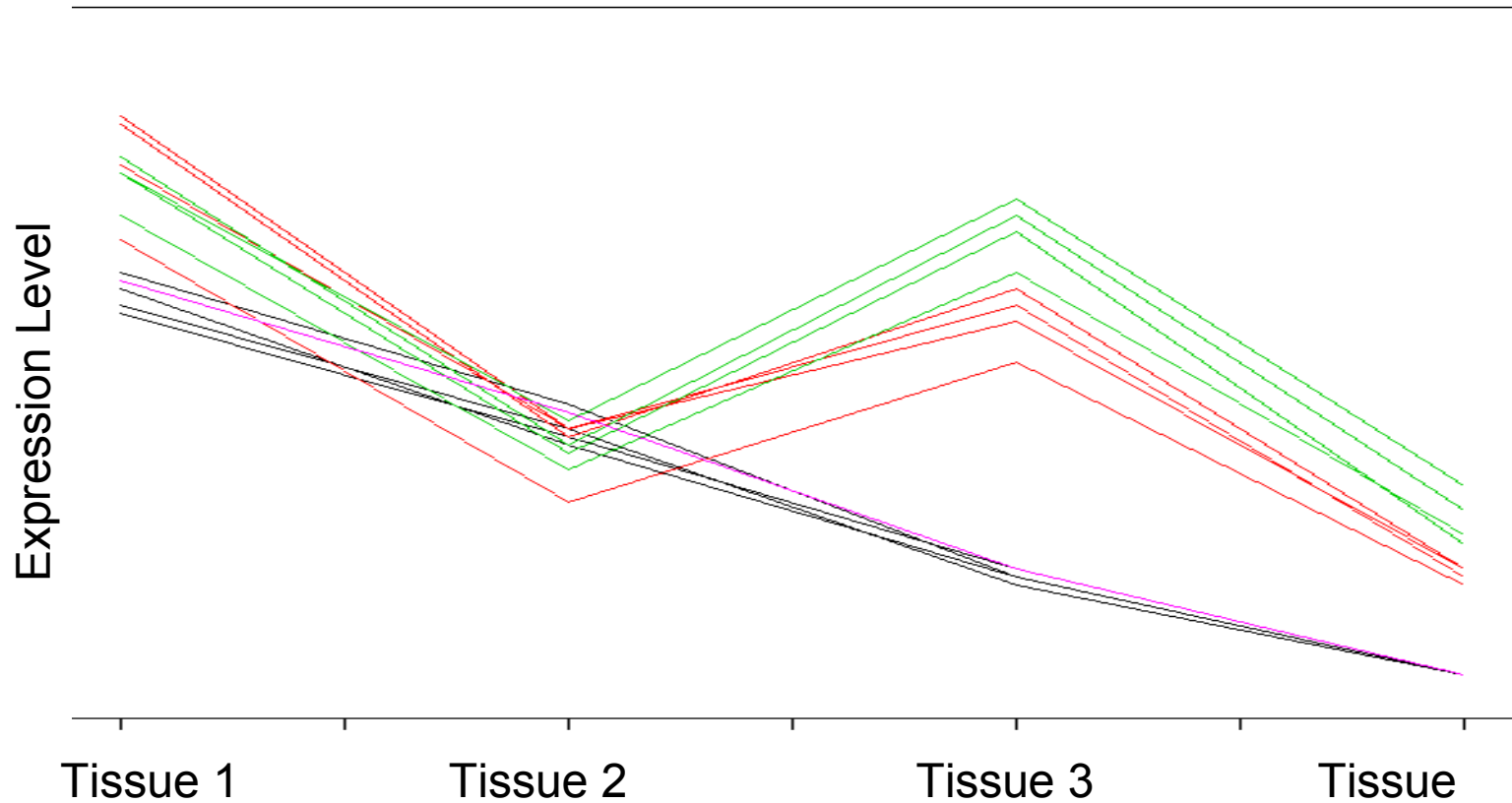
Determine intensity of color on each array

Tissues are similar if the same genes are high or low in the RNA sample.

Can be used to “discover” tumor types.

Or, start from a known samples to “train” the clustering algorithm to recognize tissue type. Then classify unknown tissue samples.

# Similarity of Genes



4 Each line represents RNA from gene. Pink line is gene with unknown function. Distance is Euclidean distance, or 1-correlation, or ...

# Clustering Genes

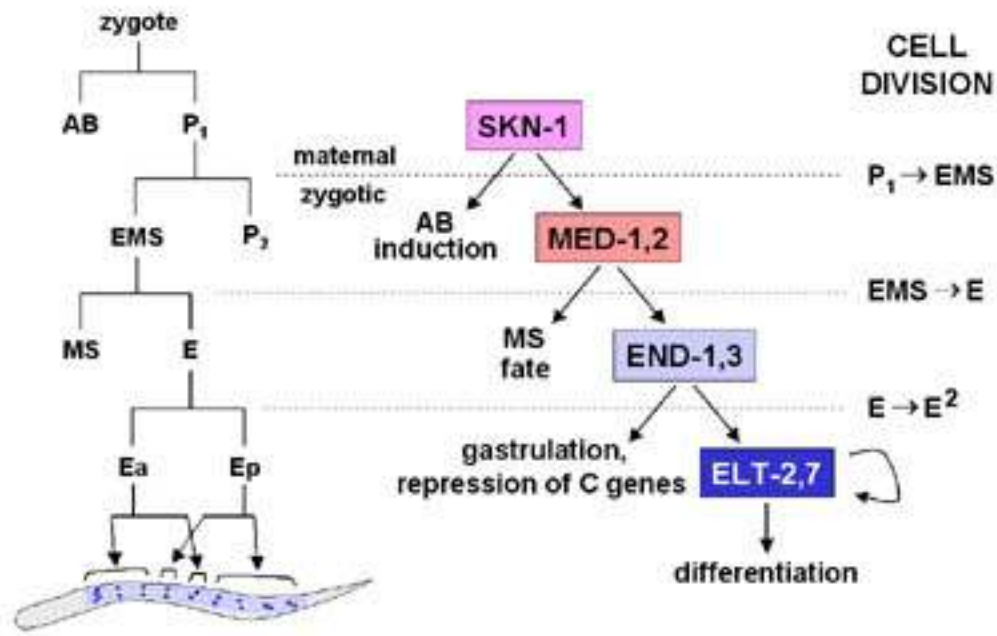
Very few gene functions are known.

How genes interact is unknown.

Hypothesis: Genes which have similar expression patterns (in different tissues or under different conditions) are related in function.

# A Gene Network

## The endoderm gene cascade



# What do you need to know?

Statistics (applied and theoretical)

Probability

Computer Science

Genetics or biochemistry

Currently: at least an MS needed

# Jobs

University

Genomics Research

Pharmaceutical Company

Medical Research

Anthropology

National Institutes of Health

Environmental Science

National Security

Agriculture

Forensic Science

And this is just the tip of the iceberg

