

## ECS 289A

### Lecture 2: The Genome and BioTechnologies

---

---

---

---

---

---

---

### Today:

- Organization of Information in the Central Dogma
  - Genomes
  - Transcriptomes
  - Proteome
- Observing the Central Dogma
  - PCR
  - DNA Sequencing
  - Microarrays
  - ChIP

---

---

---

---

---

---

---

## Genomes

### Organization and complexity

- Genomes are the union of all DNA in an organism (there are different types of DNA: nuclear and mitochondrial)
- Only small % (2%) of the human genome is genes. The rest contains various promoter regions and "junk" (>50%)
- Genome sizes vary among organisms, shortest for Phages and Viruses, longest for mammals and some plants (figure from Baldi)

---

---

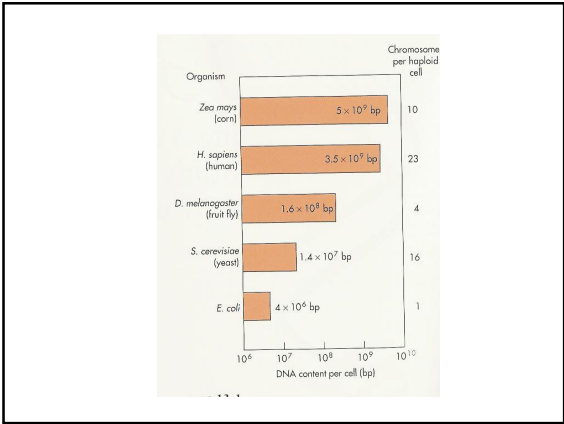
---

---

---

---

---




---

---

---

---

---

---

---

---

**Evolution**

- Changes in the genomes
- Mutations: changes in genome driven by random or particular events. Can be single base change or larger events.
- Recombination: mixing of genomes to produce a new one
- Natural selection: beneficial changes are passed on

---

---

---

---

---

---

---

---

**Similarity of genomes (i.e. organisms)**

- Evolution implies that different organisms would have common ancestors
- Thus similarity comparisons (homology searches) provide clues to evolutionary ancestry (mention phylogeny)

---

---

---

---

---

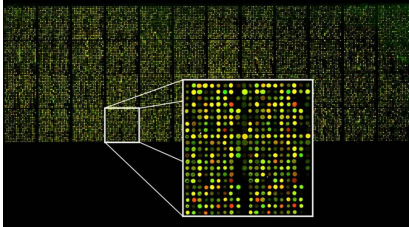
---

---

---

## Transcriptome

All possible gene expressions in the organism



---

---

---

---

---

---

---

---

### Organization and Complexity

- Transcriptome is the measurable level of all different mRNA's in an organism
- One DNA template multiple mRNAs: alternative splicing
- DNA to mRNA: one way street because of alternative splicing
- The "when and where" of mRNA concentration is coded in the promoter regions, and possibly elsewhere

### Evolution

- Evolution of gene expression under emergent properties like network organization

### Similarity of Organisms

- Comparison of gene expression from a "system's perspective"

---

---

---

---

---

---

---

---

## Proteome

Localization, abundance, and interaction of all proteins in an organism

- Structure: Amino acid sequence, 3D crystal structure
- Structure => Function?
- Sequence homology not always good indicator of functional similarity
- Study of protein expression

---

---

---

---

---

---

---

---

## BioTechnologies

- Observing the Central Dogma: sequence, gene and protein expression, DNA-protein and protein-protein interactions
- PCR, DNA Sequencing, DNA Microarrays, Chromatin ImmunoPrecipitation
- Large-Scale Technologies:
  - Thousands of measured variables
  - Require computational processing

---

---

---

---

---

---

---

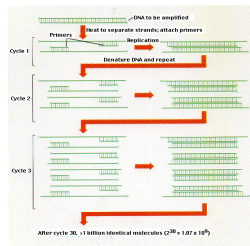
---

## Seeing the minute: PCR

Producing multiple copies of given DNA fragment (amplification)

Start: double stranded DNA molecule

1. Separate strands into templates by heating the mixture
2. Cool to allow "primers" to attach to single strands
3. The primers identify the starting points for DNA synthesis
4. DNA synthesis of strands complementary to the templates
5. Repeat 1.



---

---

---

---

---

---

---

---

## PCR properties

- The primers can determine the amplified DNA fragment if chosen to flank that region
  - n steps of the above produce  $2^n$  copies of the intended DNA fragment
- $2^{30} \sim 10^9$

---

---

---

---

---

---

---

---

## Growing DNA: Synthesizers



ABI 3900 High-Throughput DNA Synthesizer

---

---

---

---

---

---

---

---

## DNA Sequencing

Reading the string: exact positions of the base pairs A, C, G, T

1. Break DNA into manageable fragments (500 – 700 bp)
2. Sequence the fragments

---

---

---

---

---

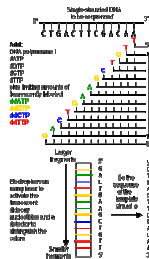
---

---

---

## Sequence Fragments

- Digest the DNA to be sequenced into small, 500 - 700 bp fragments
- Replicate sample (fragment) into four bins
- Each bin has a sufficient amount of all four bases and Polymerase
- Bin associated with base x has in addition a special version of x, a stopping version, which stops replication
- The stopping bases are also fluorescently labeled
- DNA replication creates fragments of different lengths in the bins, but all fragments in a bin end in the same labeled base
- Using Gel electrophoresis the fragments are separated by length, thus identifying the base at any given length.




---

---

---

---

---

---

---

---

Strand to be sequenced

Primers for replication

Prepare four reaction mixtures, include in each a different replication-terminating nucleotide

Primer DNA

Replication products of  $^{32}$ P reaction

Separate products by gel electrophoresis

Read sequence as emergence of bands containing labeled strands

ABI Prism 377, modern capillary sequencer

---

---

---

---

---

---

---

---

### Fragment assembly

Genome-Level Shotgun Sequencing

Unknown Sequence (Genome)

Sampling

Fragments

Sequence Reconstruction

Reconstructed Sequence (Genome)

A Fragment Assembly Program finds a Common Superstring.

---

---

---

---

---

---

---

---

- Gaps and overlaps
  - Lander-Waterman Equation

$$gaps = ne^{-n(l-1)/T}$$

- Coverage: ratio of sequenced length vs. genome length (figure)

$$coverage = nl / T$$


---

---

---

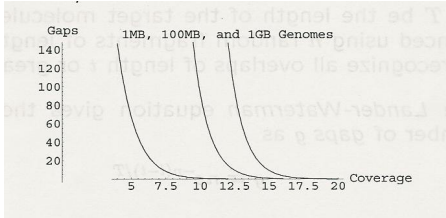
---

---

---

---

---




---

---

---

---

---

---

---

---

### Sequencing approaches

- Shotgun sequencing: "random" overlapping fragments (Celera)
- Mapped sequencing: shorter sequences are anchored (Human Genome Consortium)

---

---

---

---

---

---

---

---

### Microarrays

- Testing for the presence of a sequence fragment
  - De novo sequencing
  - Gene expression
- Hypothesis generation vs. promise of complete description on a large scale
- Possibility to do 100000 experiments at a time!

---

---

---

---

---

---

---

---

## What are Microarrays good for?

- Identifying differentially expressed genes
  - Genes that behave differently to treatments in same organisms
  - Different organisms
- Identifying naturally oscillating genes in the cell: example cell cycling genes in yeast
- Identifying SNPs
- Tumor vs normal cells

---

---

---

---

---

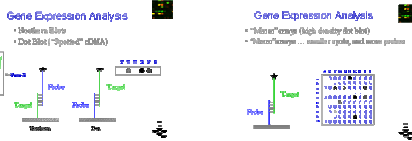
---

---

---

## How do they work?

(Source: SUNYSB microarray facility tour)



- Single stranded DNA/RNA molecules (probes) attached to a plate hybridize to their complement
- Probes attached in a square matrix typically
- Probes exposed to prepared solution (cellular extract) called a target
- They hybridize with their complements from the target
- Targets are labeled (usually by fluorescence)
- Reading the arrays: observing the color at each probe site
- Color indicates (relative) concentration of probe's complement

---

---

---

---

---

---

---

---

## Microarray Formats

### Spotted Microarrays

- Ed Southern 25 years ago, Patrick Brown recently
- Glass slide DNA arrays
- 100,000 sites per 1cm<sup>2</sup>

---

---

---

---

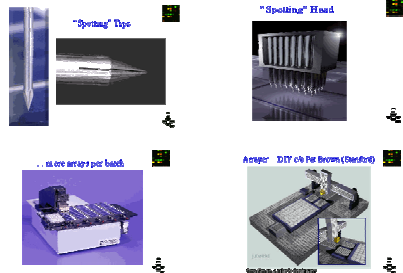
---

---

---

---

## Printing a Microarray



---

---

---

---

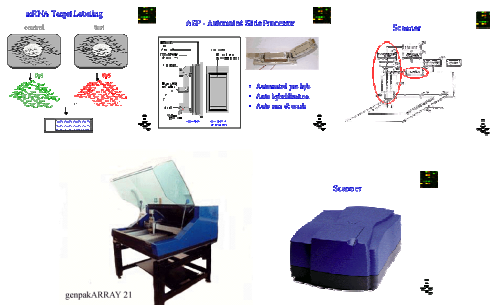
---

---

---

---

## Experiment and Reading



---

---

---

---

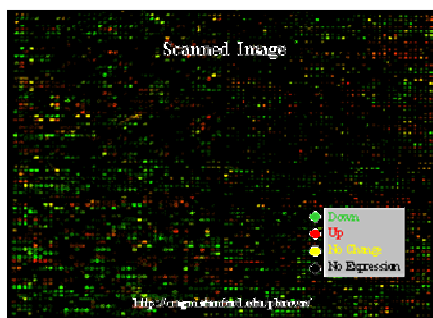
---

---

---

---

## Results



---

---

---

---

---

---

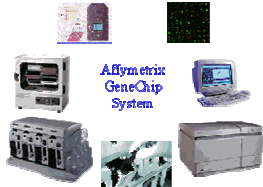
---

---

## Gene Chips

### Oligonucleotide arrays

- Photolithographic method (Affymetrix Inc.) just like computer chips
- Masks used to synthesize oligonucleotides to a chip
- 1,000,000 sites per 1cm<sup>2</sup>.



---

---

---

---

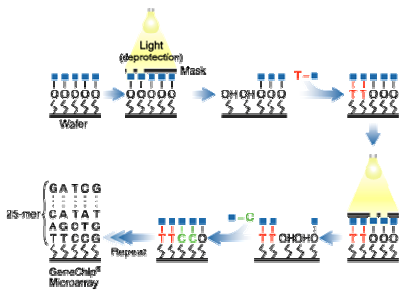
---

---

---

---

## Photolithography



---

---

---

---

---

---

---

---

## Other Microarray Technologies

- Ink jet (Agilent),
- Addressable beads (Lynx),
- etc.

---

---

---

---

---

---

---

---

## Sources of Error in Microarrays

- Length of probes
- Cross and self hybridization
- Environmental conditions

---

---

---

---

---

---

---

## Algorithmic Problems

- Probe design
- Plate design
- Data Analysis:
  - classification,
  - clustering,
  - regulation inference,
  - gene networks

---

---

---

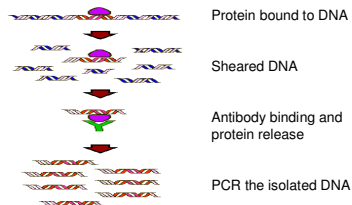
---

---

---

---

## ChIP



**ChIP:** Chromatin Immuno-Precipitation  
DNA-Protein Interactions

---

---

---

---

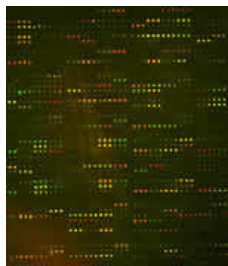
---

---

---

## Protein Expression Arrays

- Abundance of peptides and polypeptides
- Much more difficult to work with, especially analyze



---

---

---

---

---

---

---