Class Information

- http://www.biostat.jhsph.edu/~ririzarr/stanford
- Download and install R 2.3.1
- Download and install Bioconductor 1.8
- M and T we talk about general methods
- Th we introduce a problem and analyze related data using R
- If you can, please bring a laptop on Th
- If taking for grade final project requires data analysis otherwise literature review
- Class shaped as we go along

Introduction to Genome Biology and Microarray Technology

Lecture 1

Credit for some of today's materials:
Terry Speed, Sandrine Dudoit,
Victor Jongeneel, Giovanni Parmigiani

What we can learn

- Deal with *background noise*
- *Normalize* across arrays
- The probe effect
- Find differentially expressed genes
- Enrichment analysis
- The multiple comparison problem
- Experimental design
- Clustering and classification
- Time series experiments
- Annotation
- Using gene information
- New applications: SNP chips, tiling arrays, Epigenomics, etc…

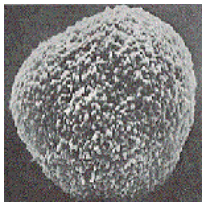
Today

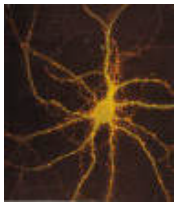
1. **Basics of Transcription**
2. **Basics of Hybridization Theory**
3. **How Microarrays Work**

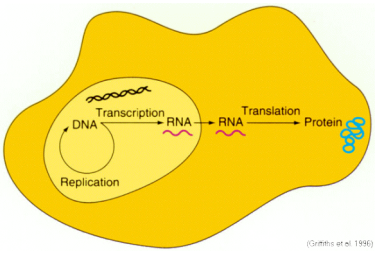
Cells and the genome

- **Each cell contains a complete copy of an organism's** genome**, or blueprint for all cellular structures and activities.**
- **The genome is distributed along** chromosomes**, which are made of compressed and entwined** DNA**.**
- **Cells are of many different types (e.g. blood, skin, nerve cells), but all can be traced back to a single cell, the fertilized egg.**

Why are cells different?





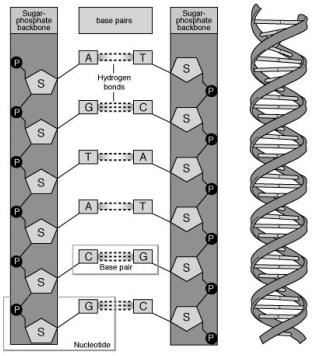


Gene expression experiments measure the amount of mRNA to see which genes are being expressed in (used by) the cell.

Measuring protein levels directly is also possible, but is currently harder.

DNA

- **A deoxyribonucleic acid or DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.**
- **Each nucleotide comprises a phosphate group, a deoxyribose sugar, and one of four nitrogen bases: adenine (A), guanine (G), cytosine (C), and thymine (T).**
- **The two chains are held together by hydrogen bonds between nitrogen bases.**
- **Base-pairing occurs according to the following rule: G pairs with C, and A pairs with T.**



Cells and the genome

- **A (protein-coding)** gene **is a segment of chromosomal** DNA **that directs the synthesis of a** protein
- **An intermediate step is the gene being transcribed or *expressed***
- **Most microarray experiments measure gene expression**

Transcription

From DNA to mRNA

Reverse transcription

Clone cDNA strands, complementary to the mRNA

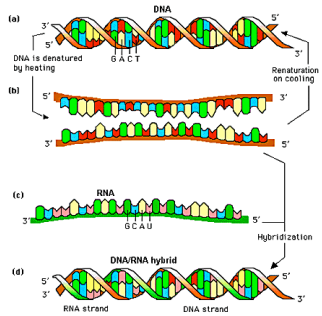
What are we measuring?

We call what we want to measure the target

- The amount of RNA transcripts
  - Expression arrays
  - RT-PCR
- The existence or abundance of a DNA sequence
  - SNP chips, Tiling arrays
- Yeast mutant representation
  - With TAG arrays

Notice all of them are Nucleic Acid molecules uniquely defined by a sequence of bases

Nucleic acid hybridization

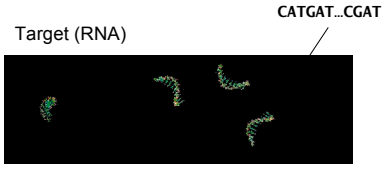


Nucleic Acid Hybridization

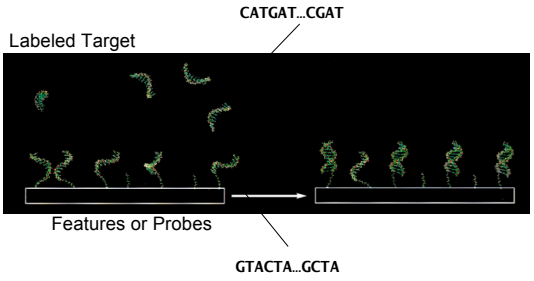
Microarrays: the game plan

- **Use hybridization to measure abundance of *target molecule***
- **Fix *probes* to a solid support and create *features***
- **Hybridize labeled target to probes and wash to get rid of non-hybridized material**
- **Use labels to measure feature intensity**

Hybridization



Hybridization

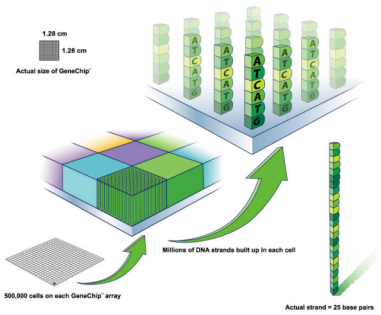


Technology Overview

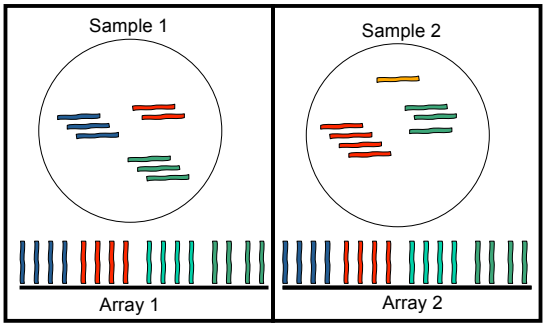
Various platforms:

- **Probes can be sequenced or cloned**
- **Features can be high-density or circles in a grid**
- **One or two samples hybridized to array**

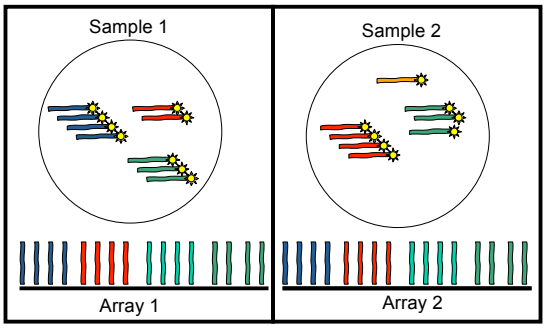
Sequenced (High density)



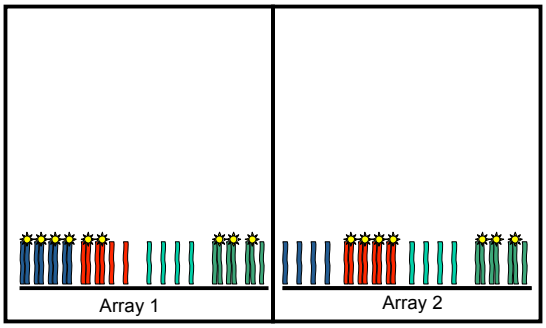
Before Labeling



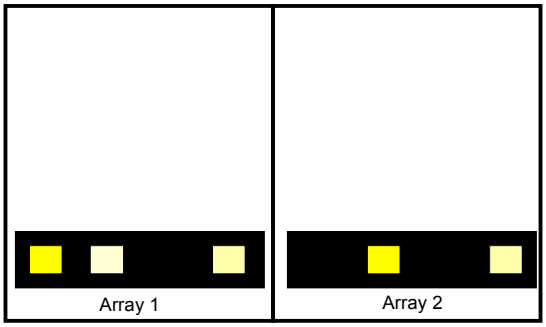
Before Hybridization: One Channel



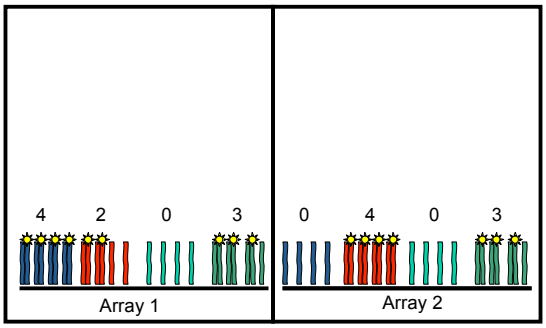
After Hybridization



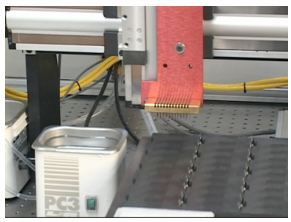
Scanner Image



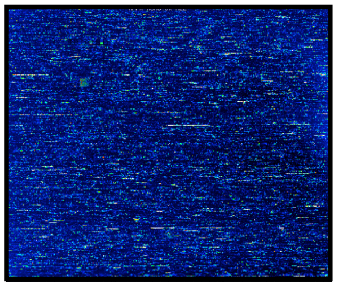
Quantification



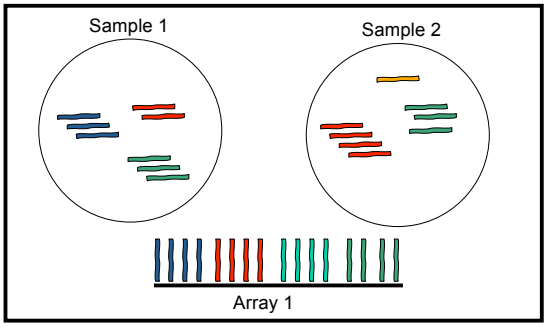
Two color



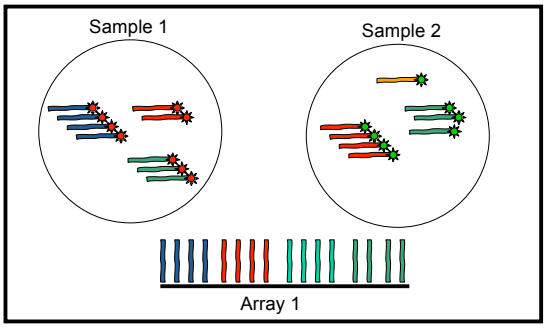
Microarray Image



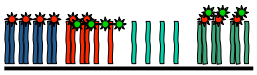
Before Labeling: Two Channel



Before Hybridization



After Hybridization

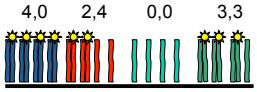


Array 1

Scanner Image

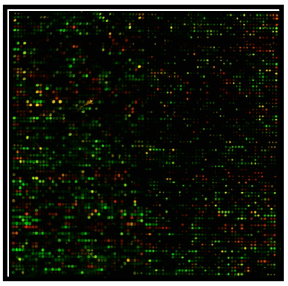
Array 1

Quantification



Array 1

Microarray Image



More on Spotted Arrays

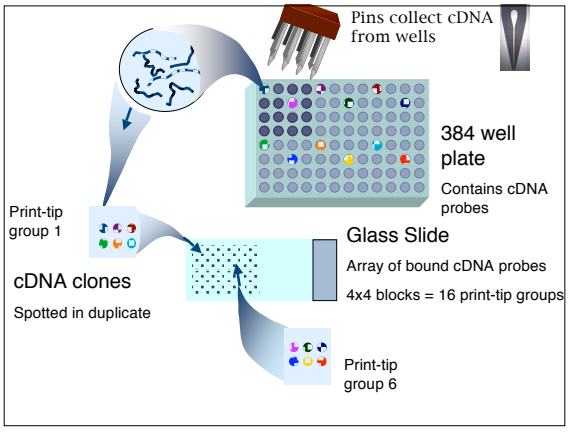


Image analysis

- With the images in place, we have data for first time
- First step is image analysis: determine which pixels are part of features and which are not
- We leave this to the company engineers although some academics have attacked the problems

Nomenclature Review

- Target - what we want to measure. Can be RNA, trated RNA, DNA, treated DNA, DNA Barcodes
- Probes - Molecules used to measure target. Can be synthesized or cloned
- Features - contiguous region on the array with same probe. We usually obtain one intensity reading from each feature

Feature Level Data

- Image analysis software produces feature level data
- This is where we starts
- First step is to get a hold of the files with this data and parse them
- Currently most files are CEL (Affymetrix), XYS (Nimblegen), and GPR (Two color platforms read with genepix scanner). But others exists!
- We also need to match each feature with a target molecule of interest. This is sometimes done in another file.