Introduction to Bioinformatics and Gene Expression Technology

Utah State University – Spring 2014

STAT 5570: Statistical Bioinformatics

Notes 1.1

Vocabulary

- Gene: hereditary DNA sequence at a specific location on chromosome (that "does something")
- Genetics: study of heredity & variation in organisms
- Genome: an organism's total genetic content (full DNA sequence)
- Genomics: study of organisms in terms of their genome

Vocabulary

- Protein: sequence of amino acids that "does something"
- Proteomics: study of all of the proteins that can come from an organisms' genome
- Phylogeny: the evolutionary or historical development of an organism (or its DNA sequence)
- Phylogenetics: the study of an organism's phylogeny
- Phenotype: the physical characteristic of interest in each individual – for example, plant height, disease status, or embryo type

Vocabulary

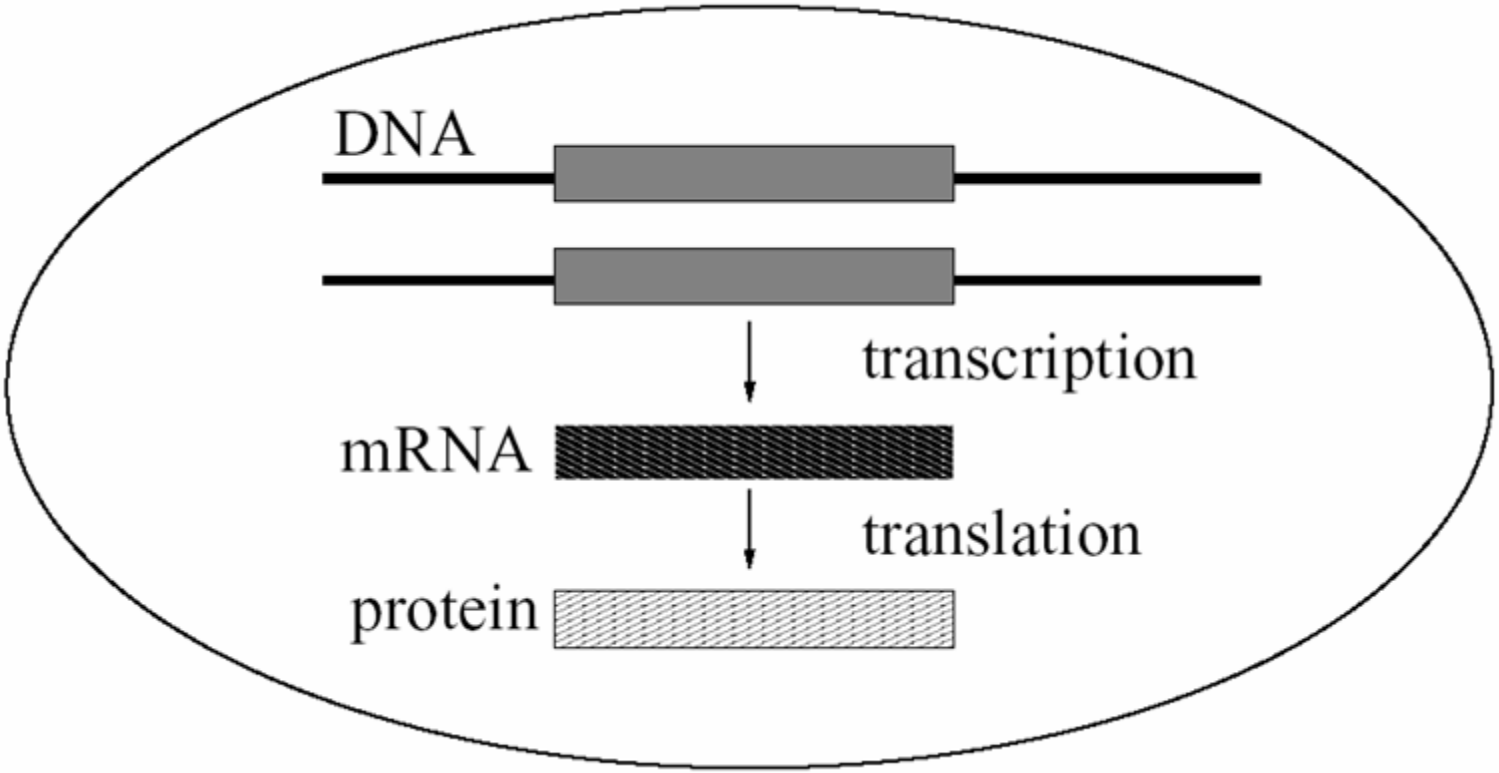
- Bioinformatics:

  the collection, organization, & analysis of large-scale, complex biological data

- Statistical Bioinformatics:

  the application of statistical approaches to bioinformatics, especially in identifying significant changes (in sequences, expression patterns, etc.) that are biologically relevant (especially in affecting the phenotype)

Central Dogma of Molecular Biology



A road map to bioinformatics

Central Dogma		Technology		Genomic Hypothesis		Type of Study or Analysis
Gene	→	Genome Sequencing	→	Genotype	←→	QTL
↓				↓		
mRNA transcript	→	Transcript Profiling	→	Transcriptome	←→	Microarrays or Next-Gen Sequencing
↓				↓	←→	(Epigenetics / methylation)
Protein	→	Protein quantification and function	→	Proteome	←→	Protein Microarrays or Proteomics
		Phenotype				

(From introductory lecture by RW Doerge at 2013 Joint Statistical Meetings)

“Alphabets”

- DNA sequences defined by nucleotides (4)
- DNA sequence $\longrightarrow$ mRNA sequence $\longrightarrow$ Protein sequence
- Protein sequences defined by amino acids (20)

General assumption of microarray technology

- Use mRNA transcript abundance level as a measure of the level of “expression” for the corresponding gene

- Proportional to degree of gene expression

How to measure mRNA abundance?

- Several different approaches with similar themes:
  - Affymetrix GeneChip
  - Nimblegen array

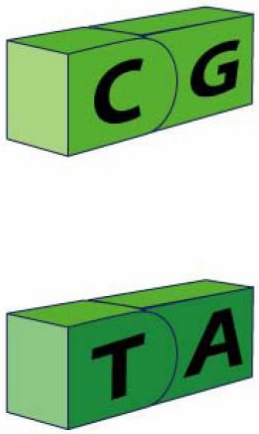
  } oligonucleotide arrays
  - Two-color cDNA array
  - More, including next-gen sequencing (later)
- Representation of genes on slide
  - Small portion of gene
  - Larger sequence of gene

A short word on bioinformatic technologies

- "Never marry a technology, because it will always leave you."
  – Scott Tingey,
  Director of Genetic Discovery at DuPont
  (shared in RW Doerge 2013 introductory overview lecture at 2013 JSM)

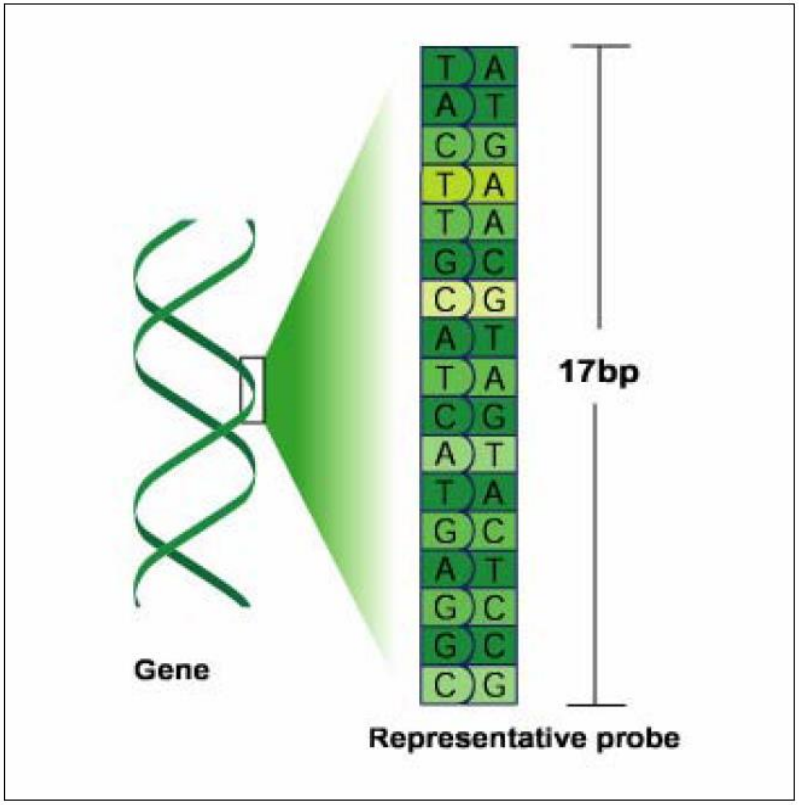
- In this class, we will discuss several technologies, emphasizing their recurring statistical issues
  - These are perpetual (and compounding)

Affymetrix Probes



DNA Base Pairing Rules

A Double Stranded Piece of DNA



A Segment of a Gene Used to Make a Probe

(Images courtesy Affymetrix, www.affymetrix.com)

Affymetrix Approach

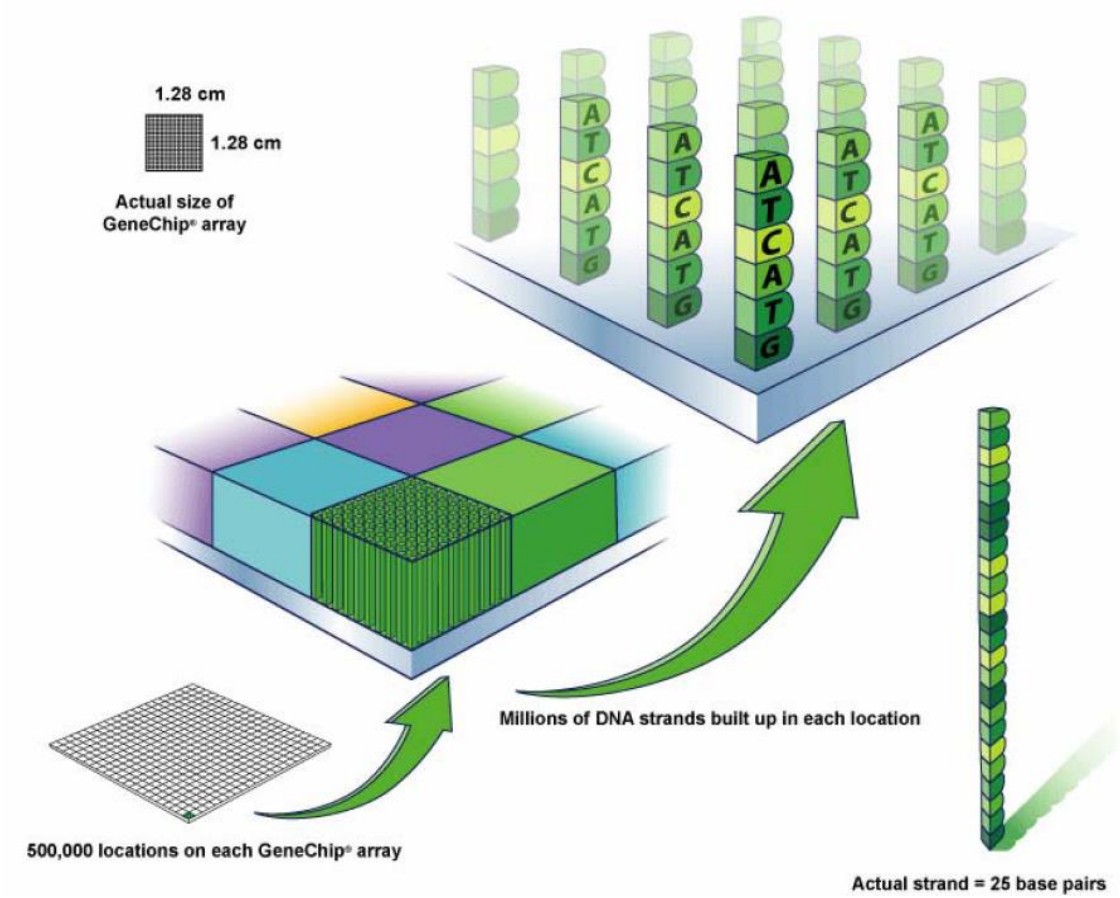
- Gene represented by a set of G probe pairs
  - PM – perfect match
  - MM – mismatch

PM: ATAAGCCAGGGACTGACTACCTTAA

MM: ATAAGCCAGGGAGTGACTACCTTAA

homomeric substitution

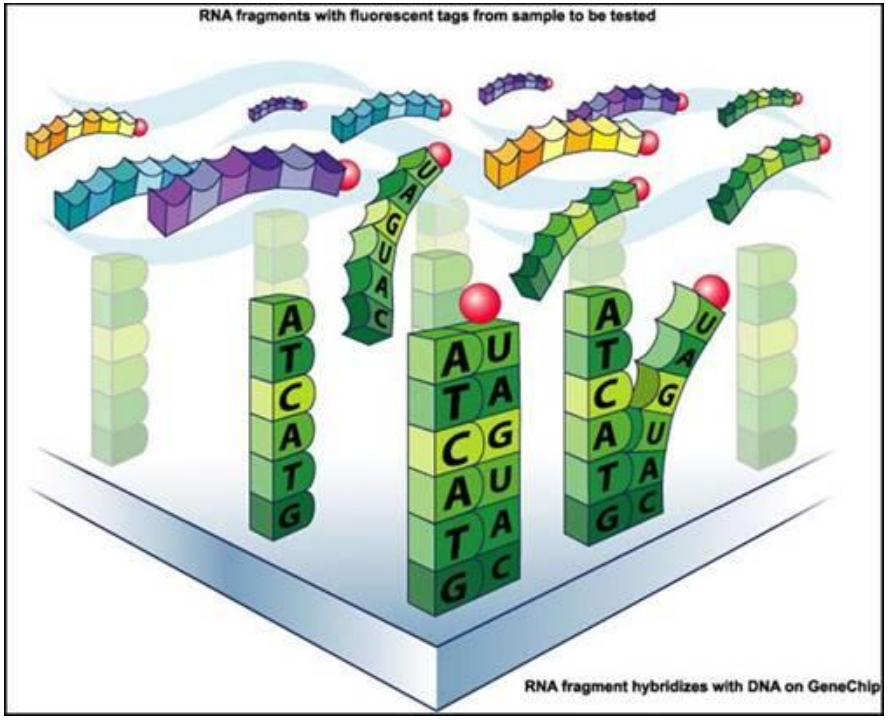
Affymetrix Technology – GeneChip

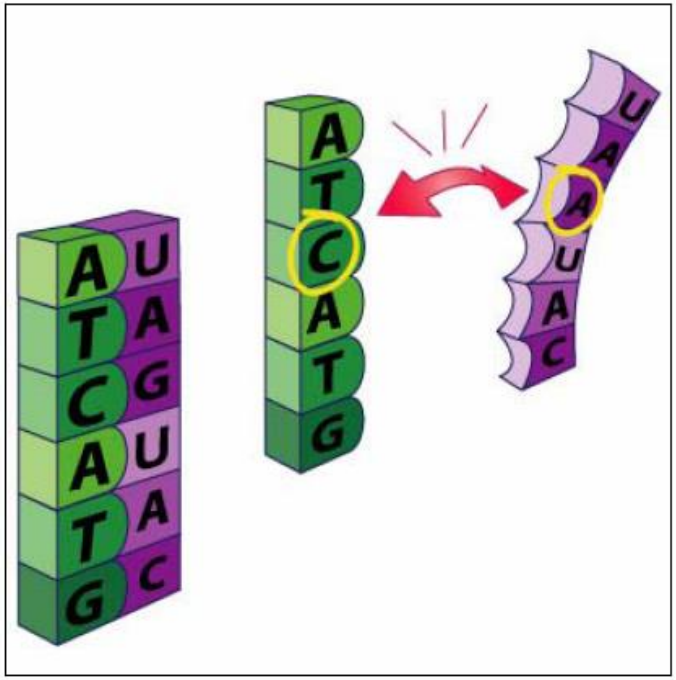


- Each gene is represented by a unique set of probe pairs (usually 12-20 probe pairs per probe set)
- Each spot on array represents a single probe (with millions of copies)
- These probes are fixed to the array

(Image courtesy Affymetrix, www.affymetrix.com)

Affymetrix Technology – Expression





A tissue sample is prepared so that its mRNA has fluorescent tags; wait for hybridization

(Images courtesy Affymetrix, www.affymetrix.com)

Affymetrix GeneChip

Image courtesy Affymetrix, www.affymetrix.com

Cartoon Representations (originally from Affymetrix outreach)

- Animation 1: GeneChip structure
  (1 min.)

- Animation 2: Measuring gene expression
  (2.5 min)

Affymetrix – Steps of the Expression Assay

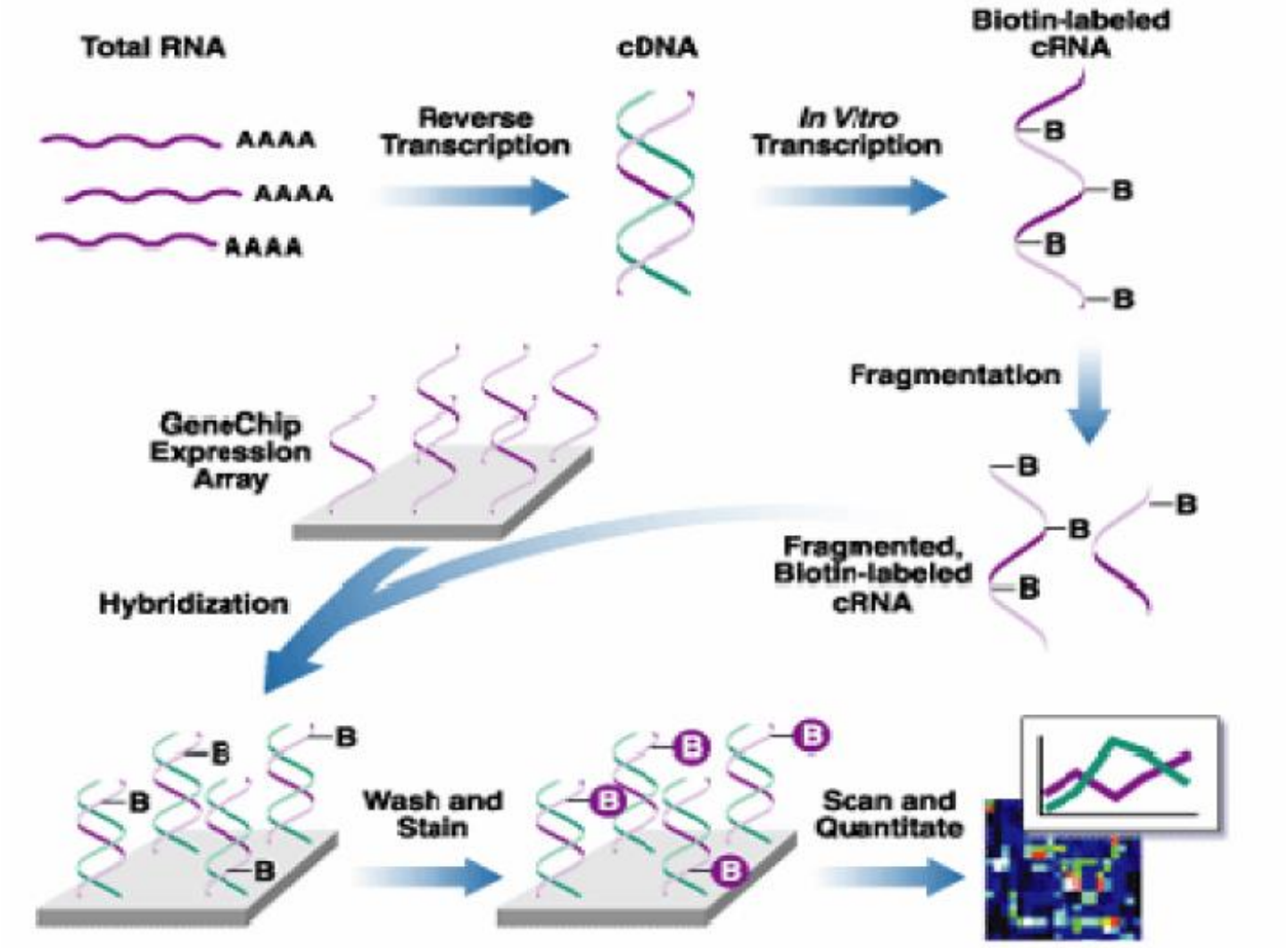
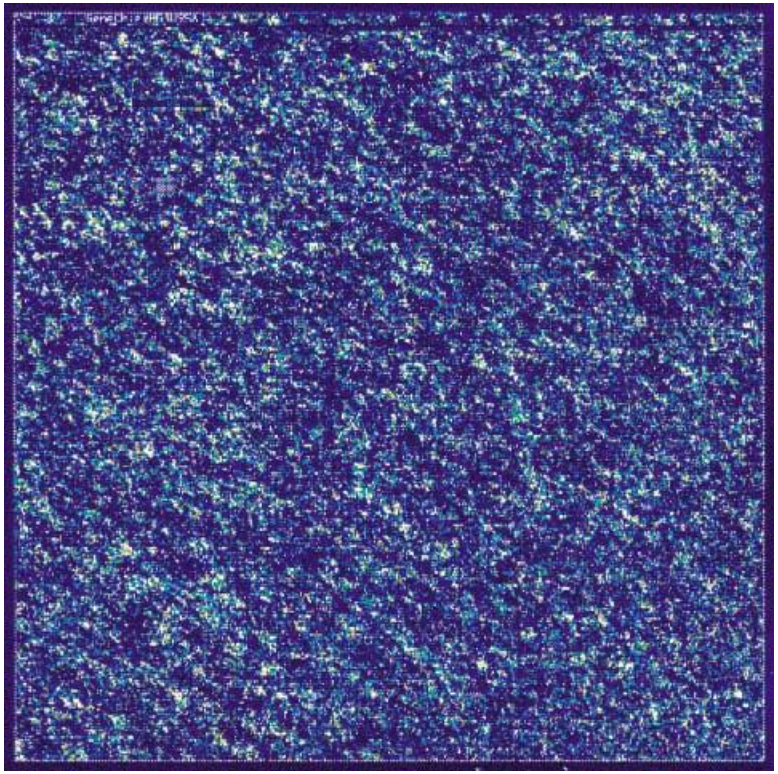
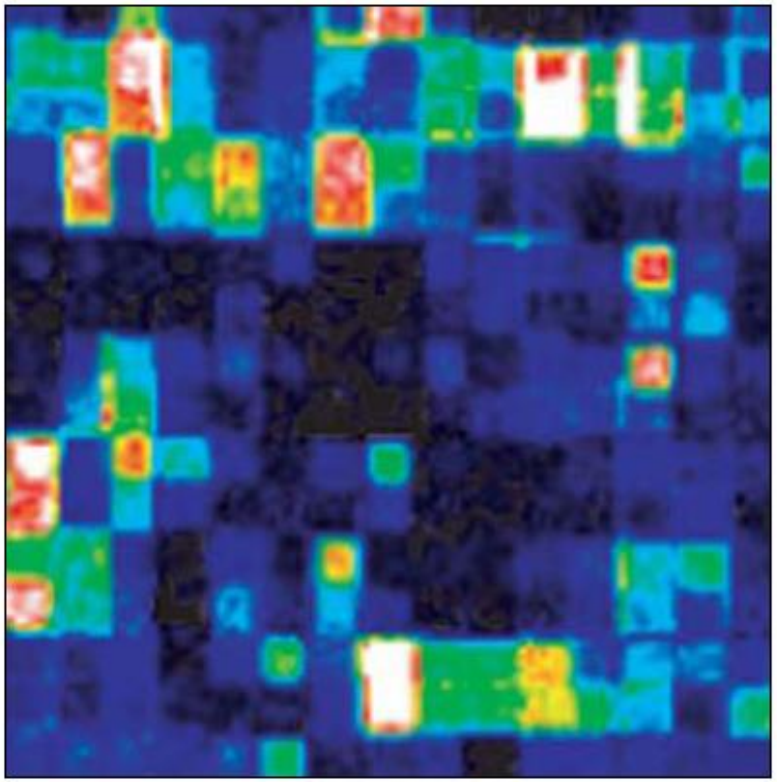


Image courtesy Affymetrix, www.affymetrix.com

Images



Full Array Image



Close-up of Array Image

Images courtesy Affymetrix, www.affymetrix.com

What are the "data"?

- Usually spot intensities –

  from specific wavelengths

- Sometimes pixel-level intensities –

  but then getting a single measurement for the spot isn't always obvious

- Identifying the "spots" is not always obvious

- Quality control is an issue

How to analyze data meaningfully?

- Consider:
  - Data quality
  - Data distribution
  - Data format & organization
  - Appropriateness of measurement methods (& variance)
  - Sources of variability (and their types)
  - Appropriate models to account for sources of variability and address question of interest
  - Meaning of P-values and appropriate tests of significance
  - Statistical significance vs. biological relevance
  - Appropriate and useful representation of results

- Many useful tools available from Bioconductor

The Bioconductor Project

- “Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data”
- Not just for microarray data
- Like a living family of software packages, changing with needs
- Core team mainly at Fred Hutchinson Cancer Research, plus many other U.S. and international institutions

Source: www.bioconductor.org

Main Features of the Bioconductor Project

- Use of R
- Documentation and reproducible research
- Statistical and graphical methods
- Annotation
- Short courses
- Open source
- Open development

Source: www.bioconductor.org

What will we do?

- Learn basics of major Bioconductor tools
- Focus on statistical issues
- Discuss recent developments
- Learn to discuss all of this