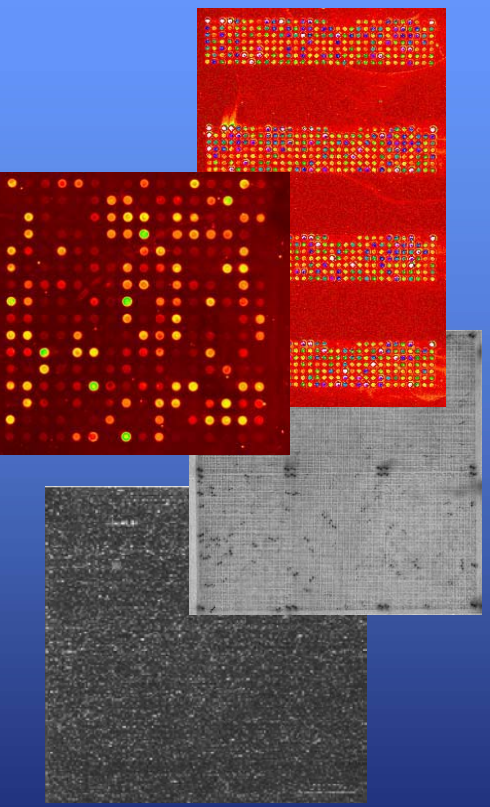


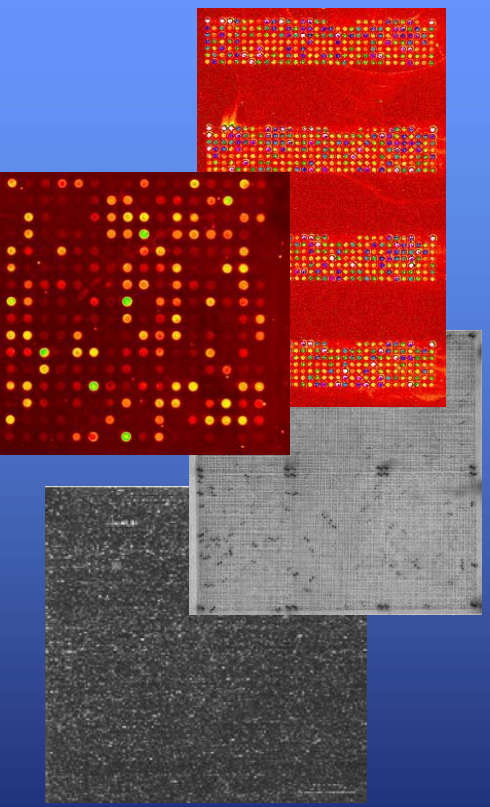
From Algorithms to Scientific Application: An Age-Old Issue Updated for the Genomics Era

Robert Nadon

McGill University
Department of Human Genetics

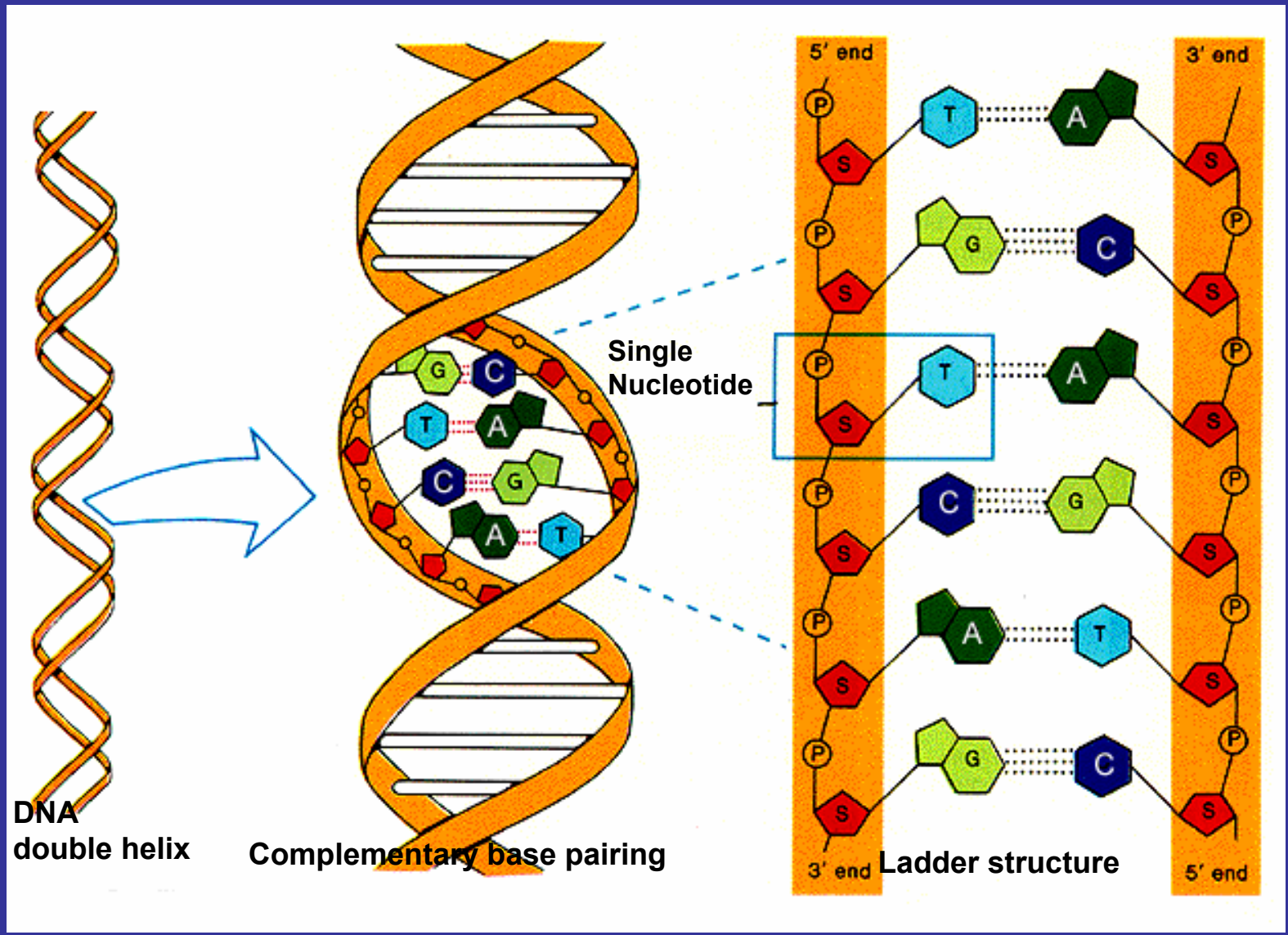
McGill University and Genome Quebec
Innovation Centre

OVERVIEW

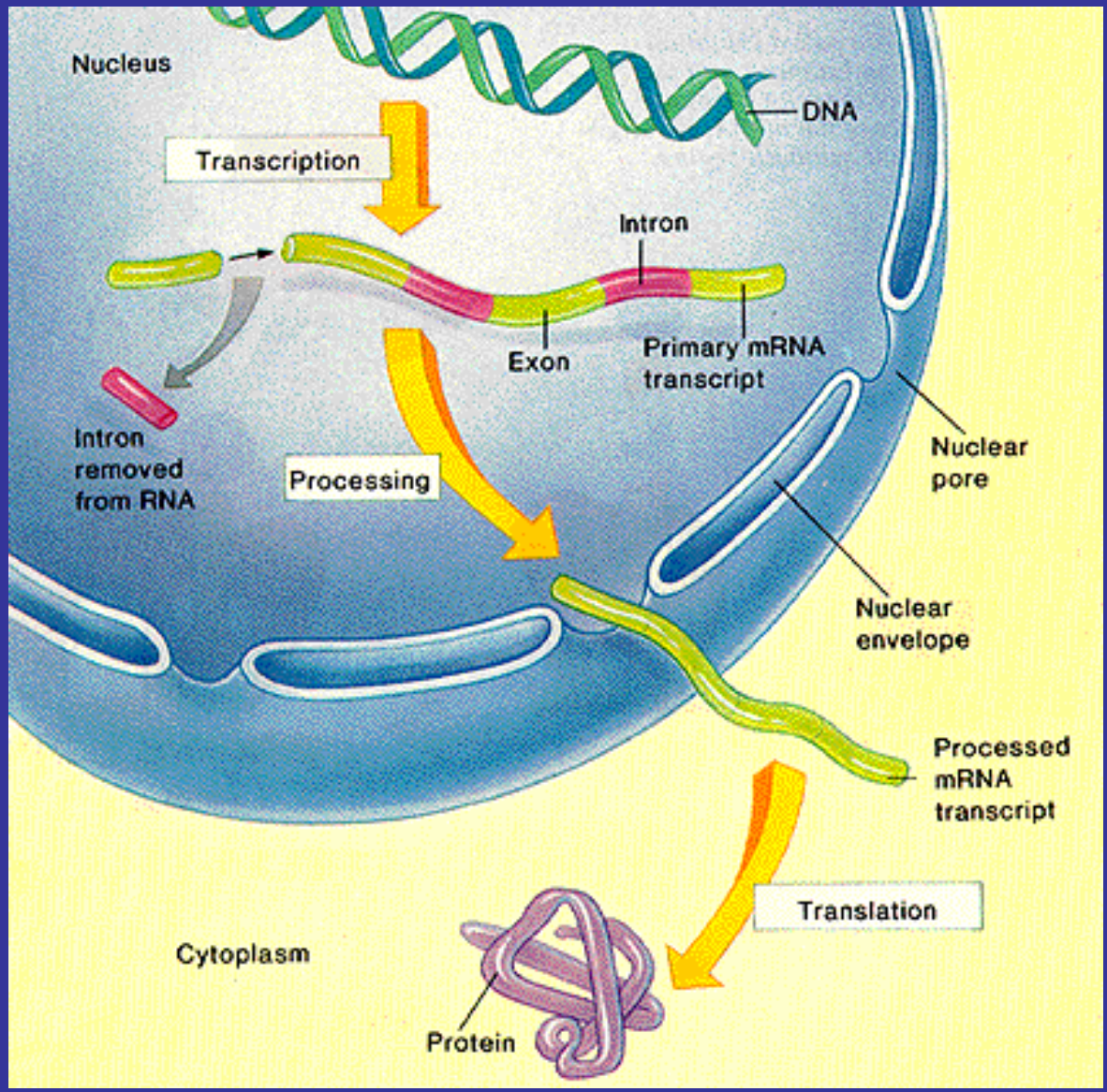


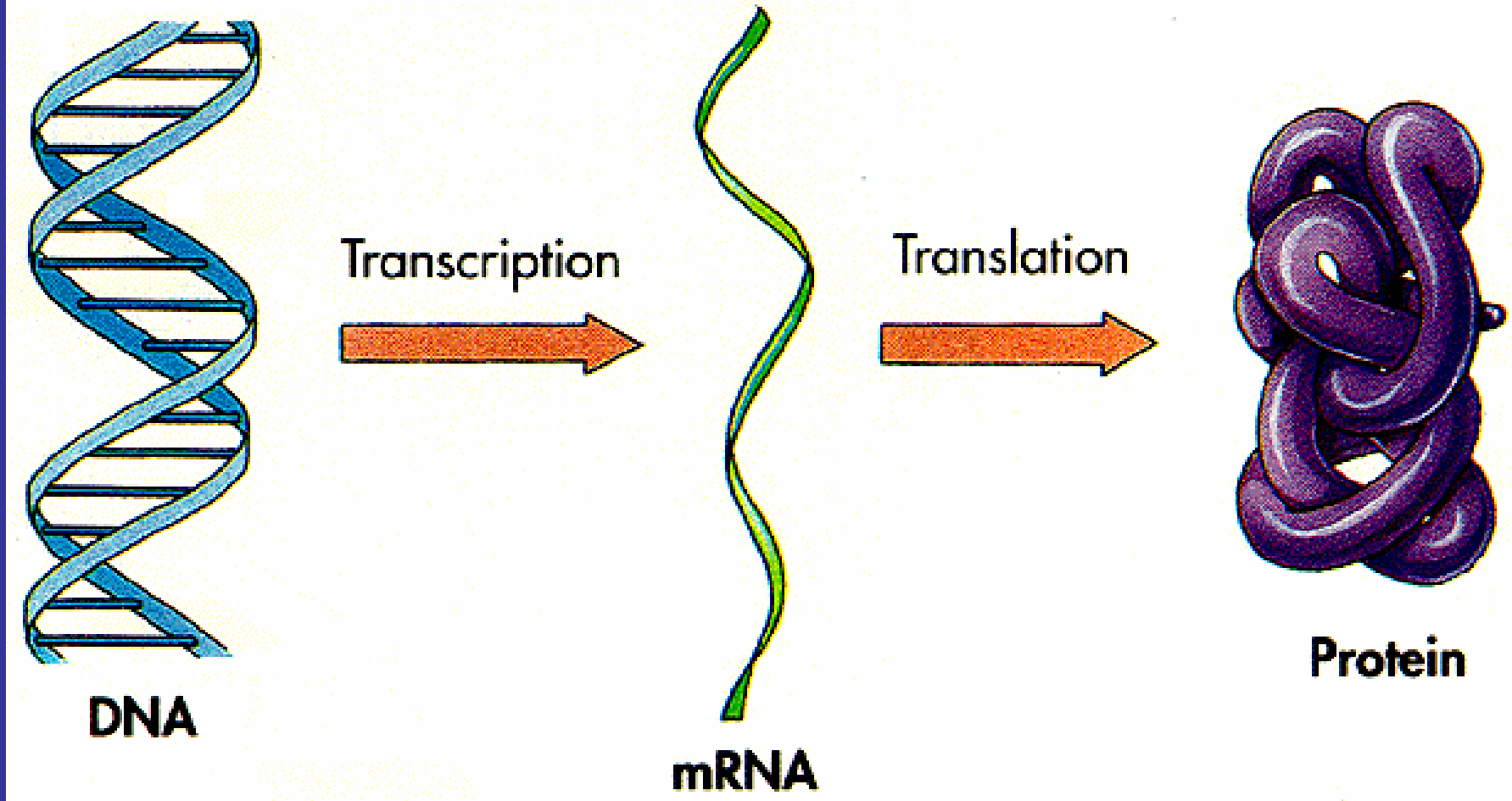
- What Is Measured in Microarrays
- Microarray Analysis
- Market of Ideas: Progress & Confusion
- Collaborator Model

What Is DNA?

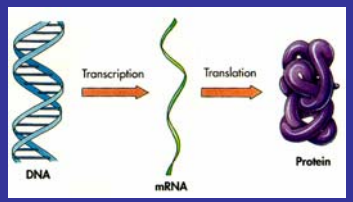

5′ end
3′ end
Single
Nucleotide
DNA
double helix
Complementary base pairing
3′ end
Ladder structure
5′ end

What Is Gene Expression?

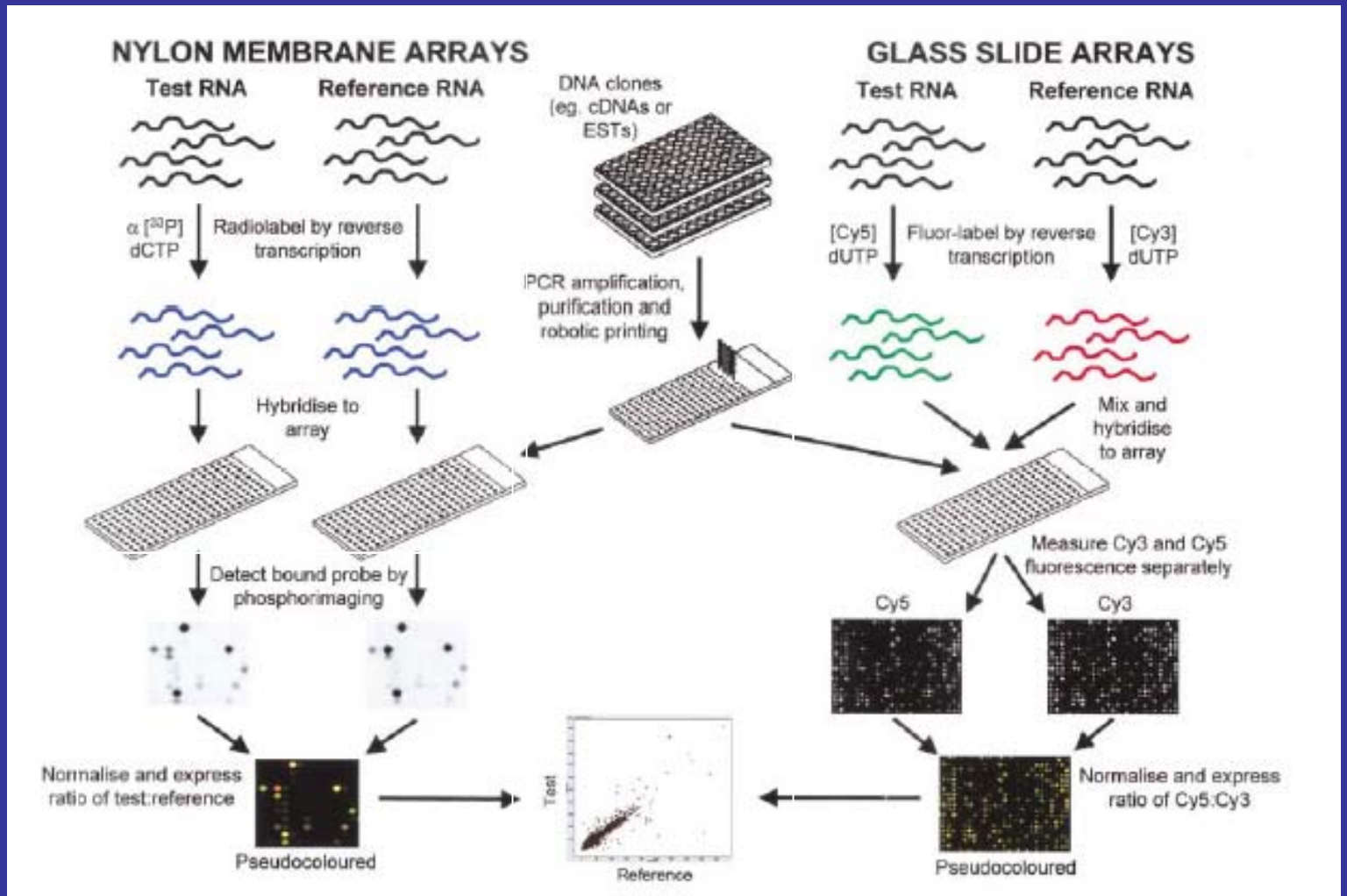



Transcription
Translation
DNA
mRNA
Protein

cDNA Arrays

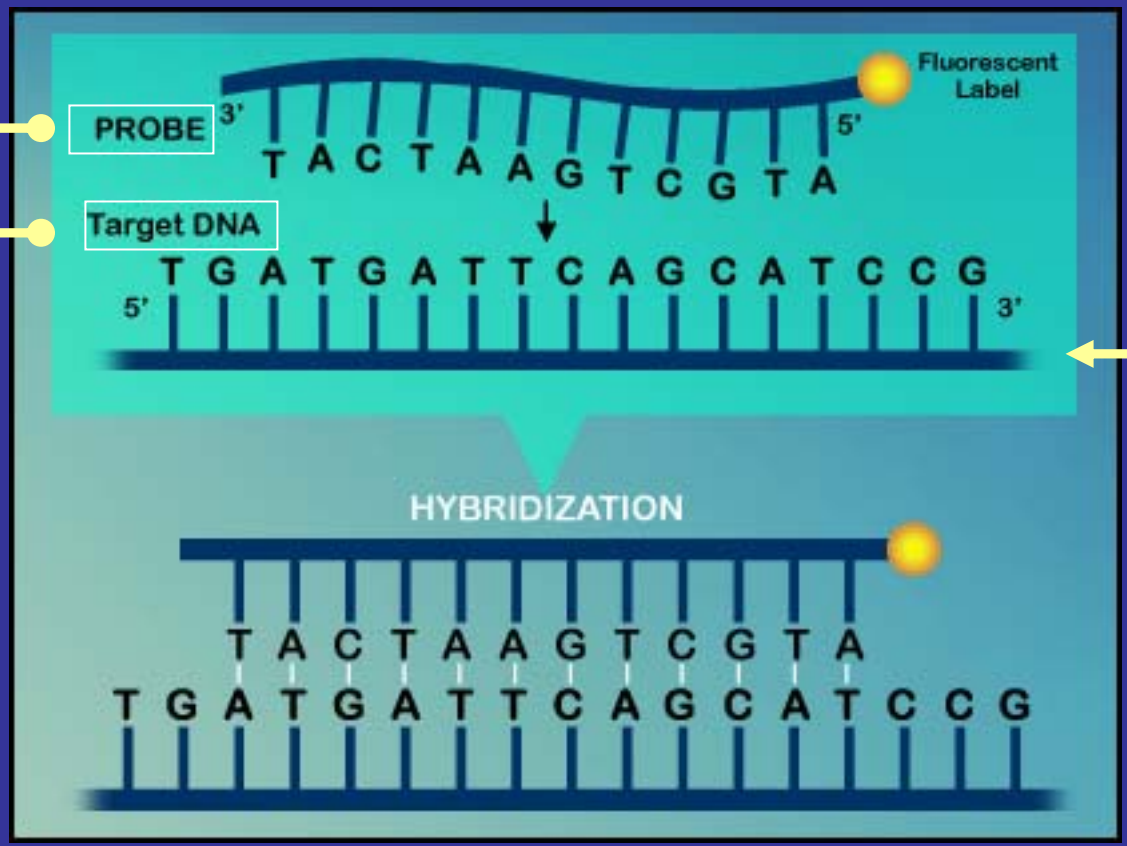


Reverse Transcription



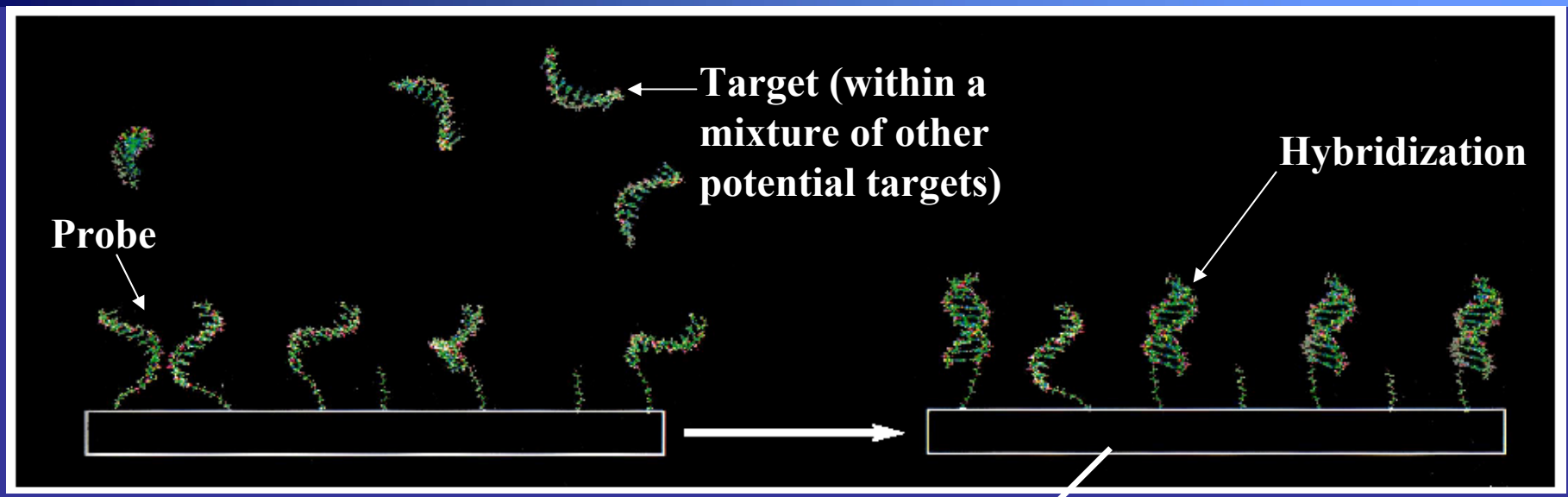
Target

Probe



Substrate

Nucleic Acid Hybridization

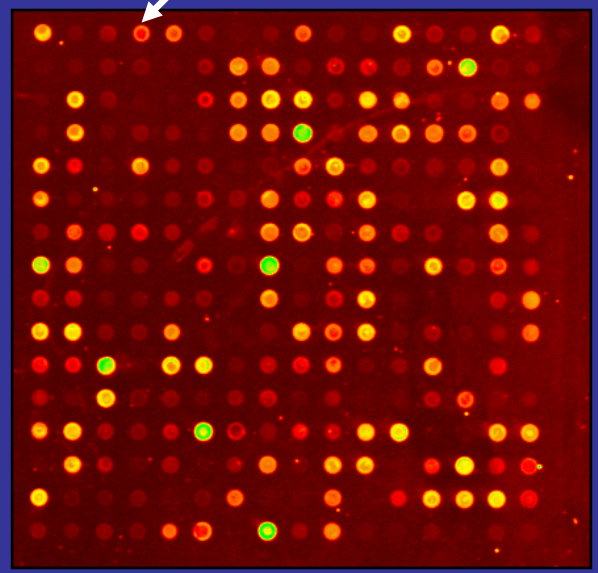


Pseudo-Colour

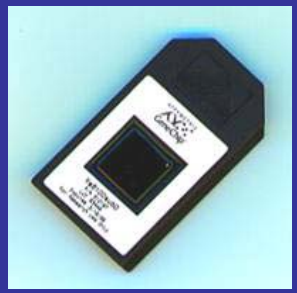
Red: Cy5 Channel Up-regulated

Green: Cy3 Channel Up-regulated

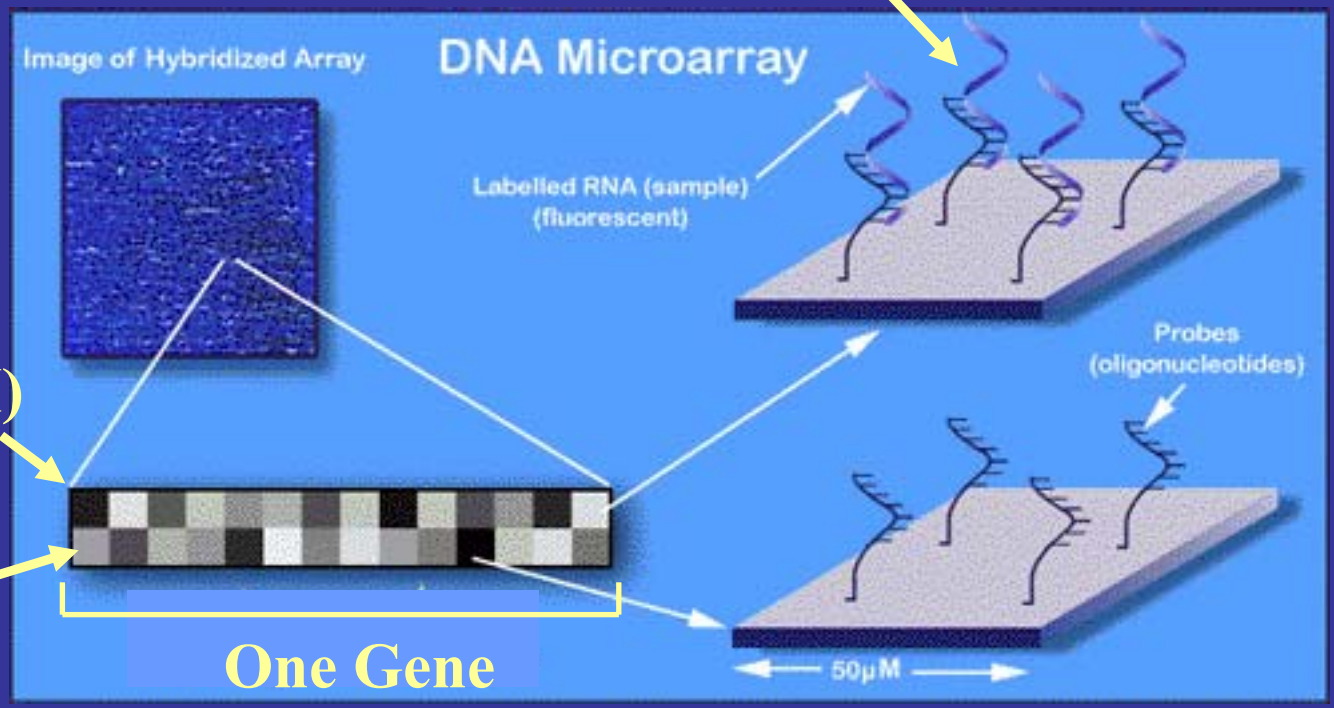
Yellow: No change



Affymetrix Arrays



Oligonucleotides (Short sequences of RNA)



Perfect Match (PM)

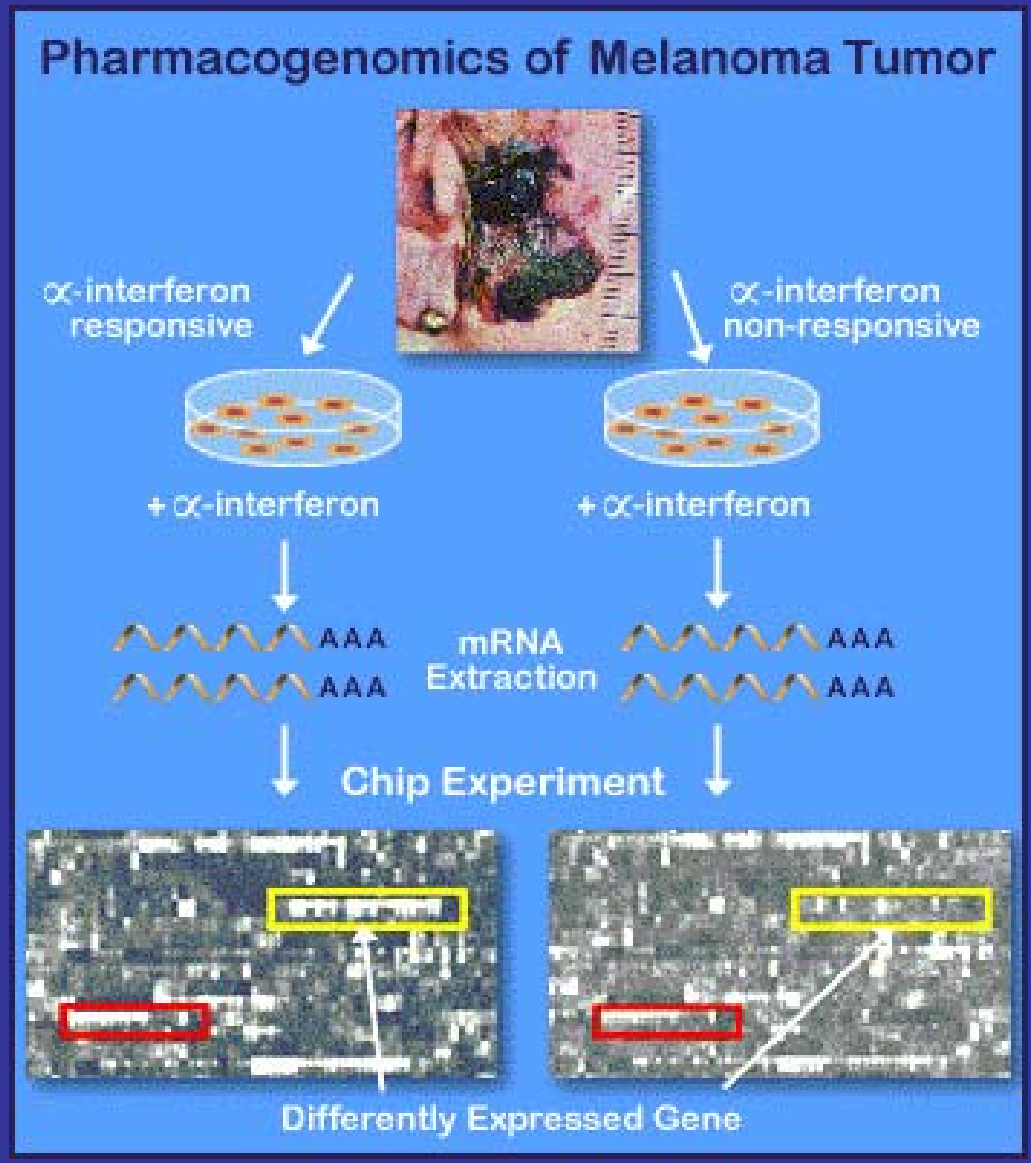
Mismatch (MM)

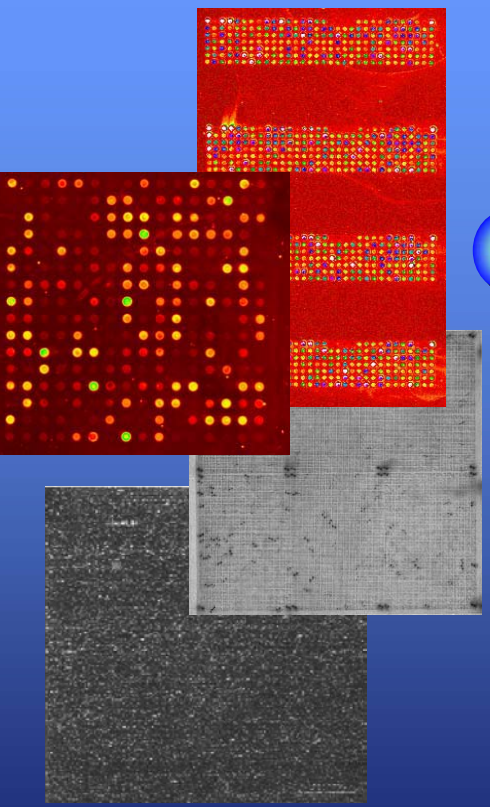
Comparative Study

Tumor Tissues

Same initial diagnosis (Phenotype) but different response to treatment

So investigate potential differences in underlying genotype





What Is Measured in Microarrays

- **Microarray Analysis**
- **Market of Ideas: Progress & Confusion**
- **Collaborator Model**

Evolution of Gene Expression Analysis

Visual Assessment

Quantification (Informal Inference Rules)

Low Level Statistical Analysis

Pattern Recognition/ Visualization

Information Management

Evolution of Gene Expression Analysis

Low Level
Statistical Analysis

Early Data Analysis Schematic

$n = 1$

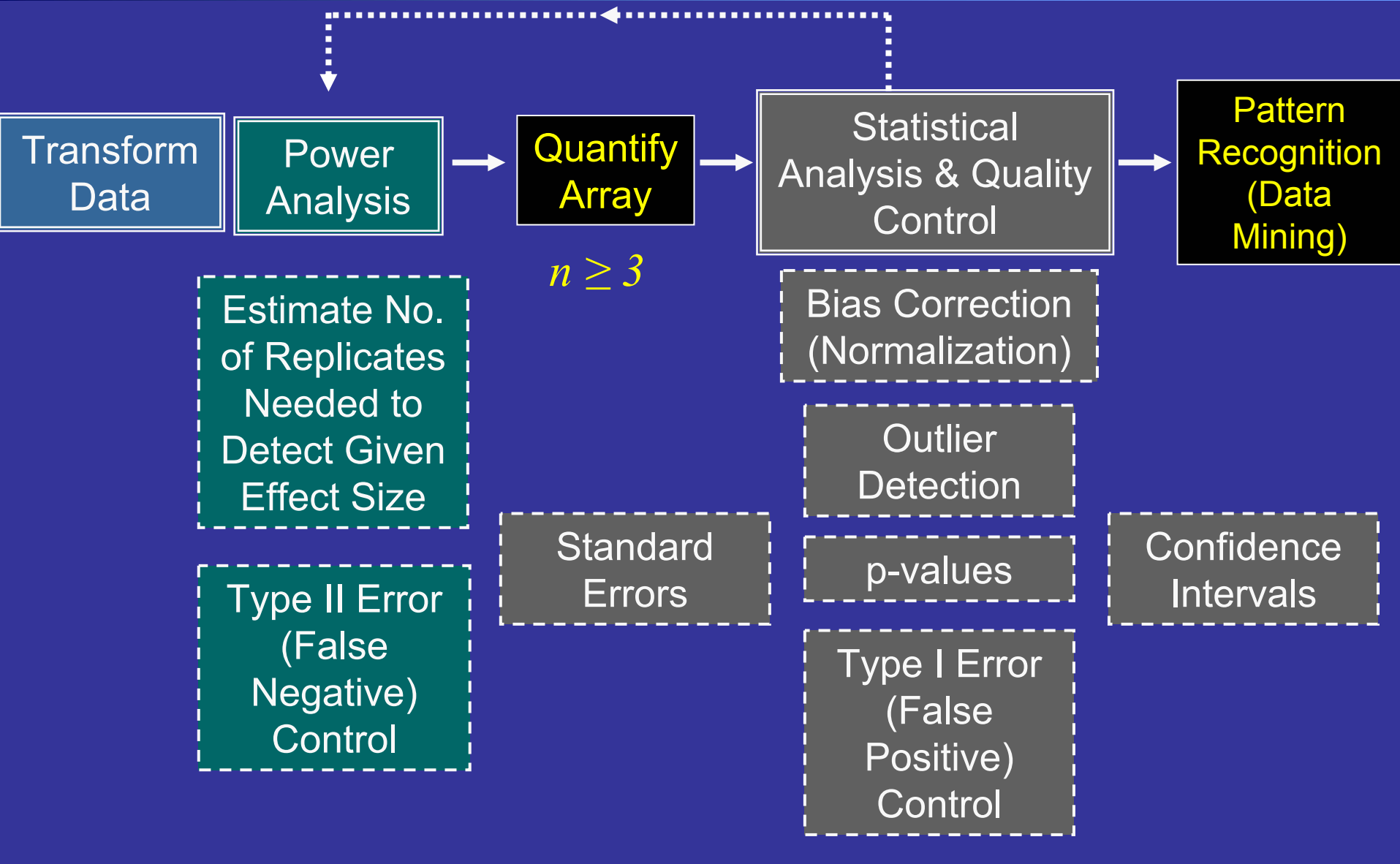
Microarray Technology Oversold (or perhaps overbought)

Reliability and precision are less than initially thought

Partly due to

- Technology itself
- Poor study designs
- Neglect of low level preprocessing and analysis issues
  - Fundamental questions were not asked

A Data Analysis Schematic

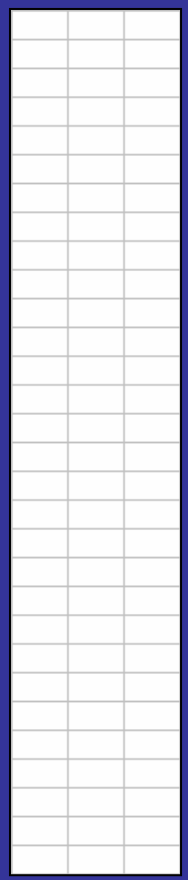


Dimensionality Problem

"Standard" Experiment

3 Variables

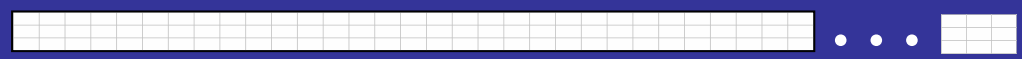
30 Observations



Microarray Experiment

1-3 Observations

10,000+ "Gene Experiments"



Important Statistical Issues for Low Level Analysis

Bias correction
- Background
- Normalization

Outliers
- Missing data

Data transformation

Random error estimation

Statistical testing of differential expression
- Multiple test problem

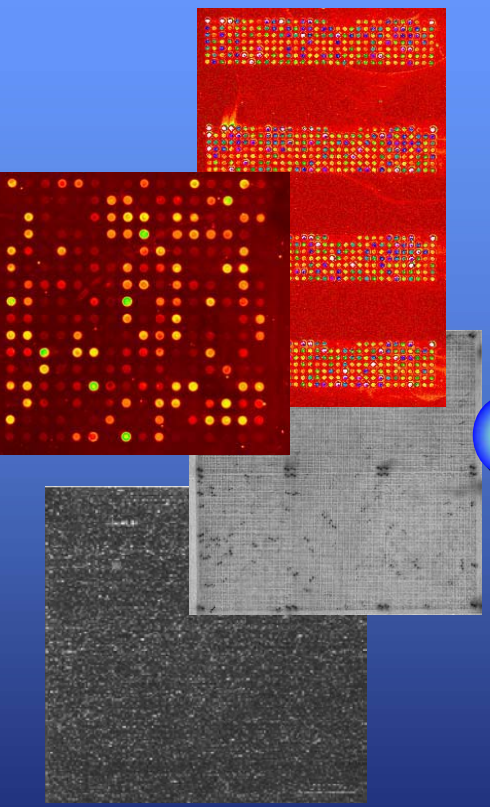
Nontrivial Problems

Only small number of replicates typically available

Bottleneck occurs at data analysis stage

Hybridization of probes to targets is not sufficiently understood

- Problem of particular importance for Affymetrix technology



- What Is Measured in Microarrays
- Microarray Analysis
- **Market of Ideas: Progress & Confusion**
- **Collaborator Model**

Statistics: From Nothing to "Nothing But"

Many statistical methods now available for all aspects of microarray studies

Usual discussion lines

- Bayesians versus frequentists
- "New" methods versus classical methods
- Nonparametric versus parametric
- Efficiency versus robustness
- Bias versus precision trade-off

Problems are difficult so not surprising that different solutions have emerged

But this has caused confusion among life scientists

Not simply a matter of the usual trade-offs and professional disagreements

Confusion reflects that

- No true gold standards exist in empirical data
- Not clear how to realistically simulate data

Most Important Problem

Life scientists want lists of differentially expressed genes

- Different methods produce different lists
- Screening versus data mining

Which of the many available methods work best under which circumstances?

- Not very exciting but is a practical imperative
- Reproducibility
- Specificity versus sensitivity

Large funding sources will dry up if these issues are not addressed

Statistics Oversold (or perhaps overbought)

Statisticians alone cannot provide the answers

Life scientists must be active participants in debate

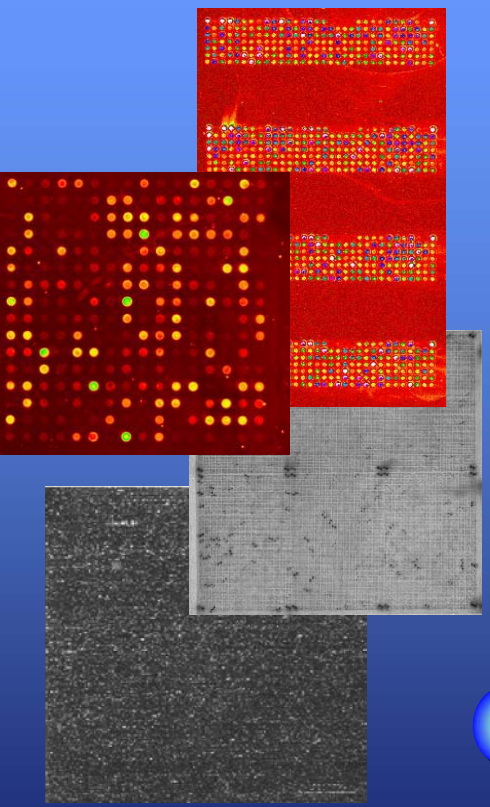
The Dreaded "J" Word

Task for both groups is to "know what we don't know"

Then learn as much as we need to learn… but not more

Most pressing problems are in data analysis - not data management

- Expression data & sequence are quite different



- What Is Measured in Microarrays
- Microarray Analysis
- Market of Ideas: Progress & Confusion
- **Collaborator Model**

“No offence but you’re really just a service provider”

Life scientists at times appear to believe that the role of data analysts is to provide service

Statisticians at times appear to believe that life scientists are incapable of correctly analyzing their data

Neither viewpoint is healthy

McGill University and Genome Quebec Innovation Centre

www.innovation.mcgill.ca

Collaborator Model

Only reasonable model for now

- But really quite inadequate

The Future

Freedom from statisticians for standard designs

Easy-to-use software

- Avoid pitfalls of other fields with button-click software

Textbooks

Still Not There Yet

Microarray analysis ≠ Clinical Trial analysis

How Do We Get There?

John Weinstein's (NCI) Bioinformatician Categories (*Revised*)

- Developmental
- Applied
- Experimental Design (Applied Statistics)

Acknowledgements

McGill University and Genome Quebec Innovation Centre

Tom Hudson
Rob Sladek
Vincent Ferretti
André Ponton
Yoshihiko Nagai
Jaroslav Novak
Saravanan Sundararajan
Tibor Van Rooij

My Lab

Sylvie Gervais
Mathieu Miron
Carl Murie
Owen Woody

Images

Roche Genetics Education Program (Issue 3.1) www.roche.com/home/science/science-gengen/science-gengen-cdrom.htm

Suzuki, D. T. (1986). An Introduction to genetic analysis (3rd ed.). New York: W.H. Freeman.

Clarke et al. (2001). Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. *Biochemical Pharmacology, 62*, 1311-1336

Question to a River Pilot

Q: “Do you know where all the rocks are?

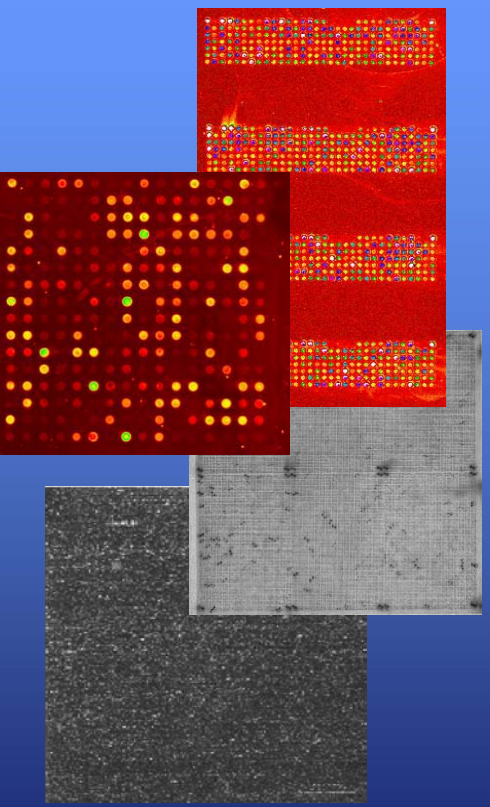
A: “No, but I sure know where they ain’t”.

For microarrays, we don’t yet have the knowledge of the river pilot… but we have spotted some of the rocks.

GenomeQuébec

Centre universitaire de santé McGill
McGill University Health Centre

McGill



Thank you

robert.nadon@mcgill.ca