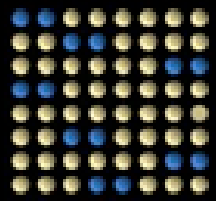
Data By Design

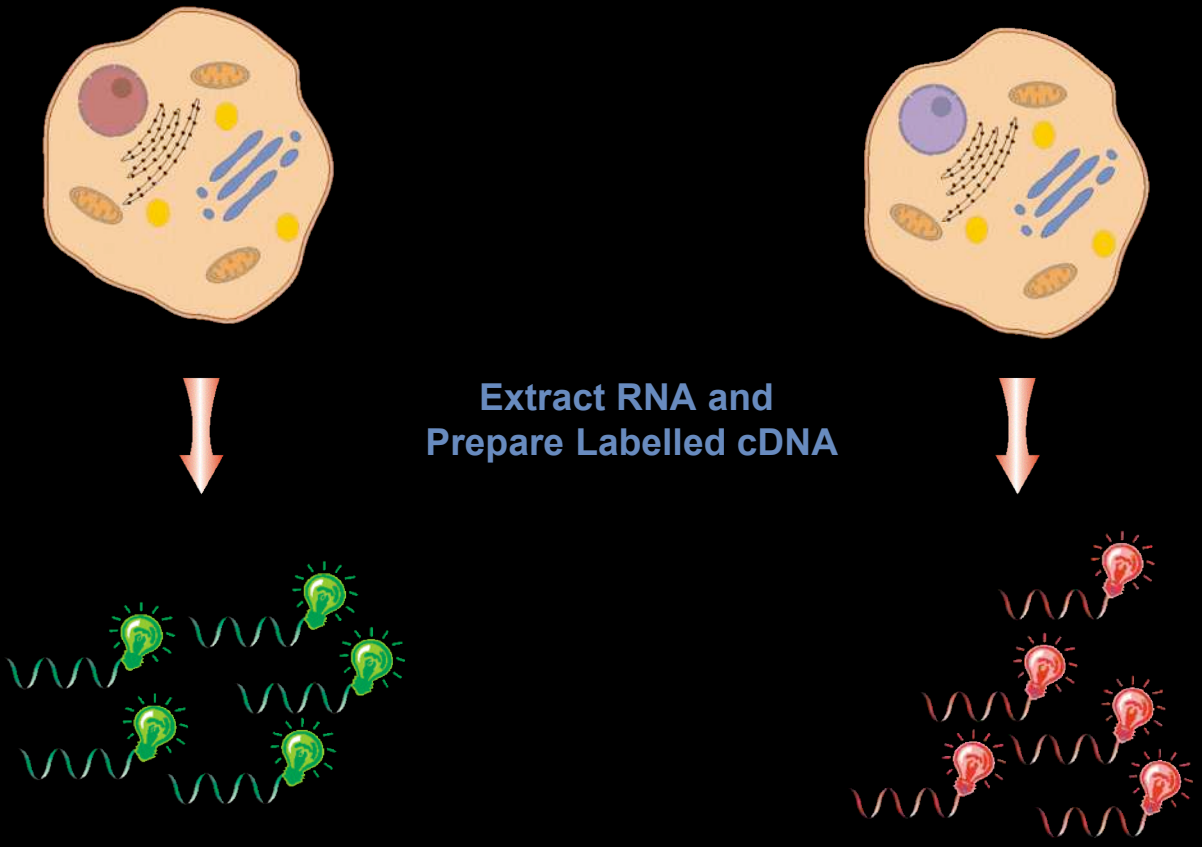
DNA Microarray Workshop

VIDO, University of Saskatchewan

February 26, 2004

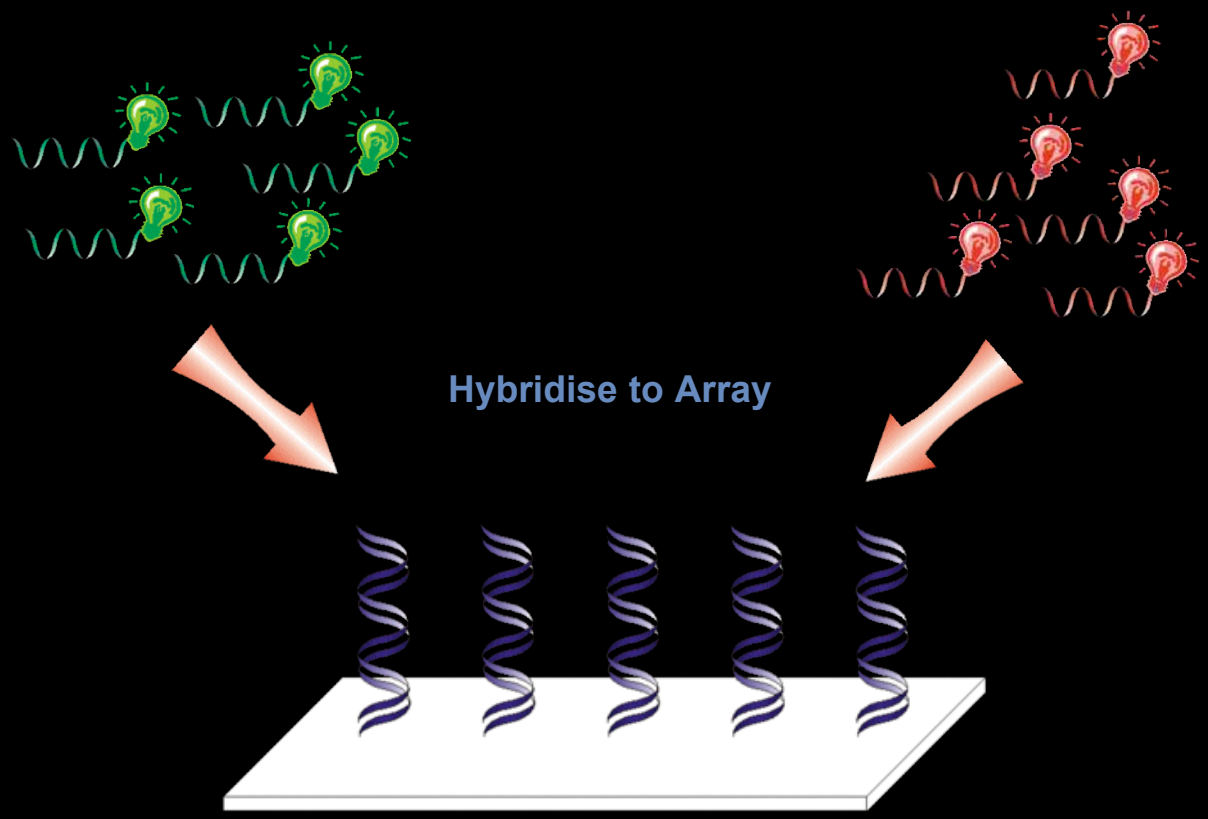


UHN Microarray Centre



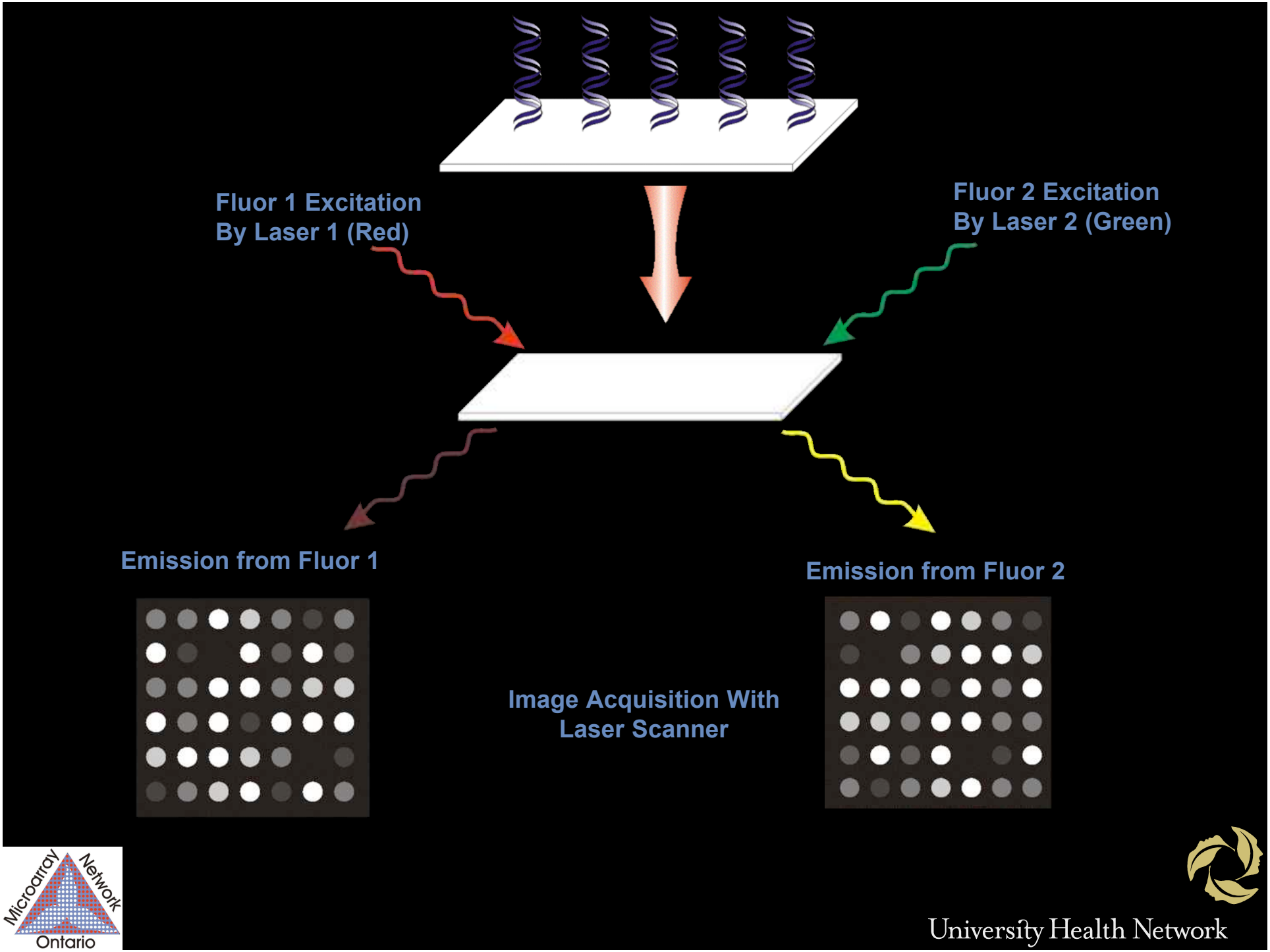
Microarray Network Ontario

University Health Network

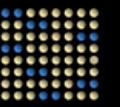


Microarray Network Ontario

University Health Network


Fluor 1 Excitation
By Laser 1 (Red)
Fluor 2 Excitation
By Laser 2 (Green)
Emission from Fluor 1
Emission from Fluor 2
Image Acquisition With
Laser Scanner
Microarray Network
Ontario
University Health Network

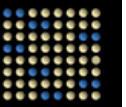
Microarrays have created a paradigm shift - causing experiments to become hypothesis-generating rather than hypothesis-driven.



UHN Microarray Centre

Ask yourself:

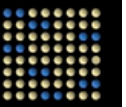
- What is the question I want to answer with this experiment?
- How much RNA do I have access to for each sample?
- How many samples do I need to run?
- Do I currently have all the samples in my possession or is will more be added later?
- How many arrays can I afford to run?
- How concerned about false-positives do I have to be?



UHN Microarray Centre

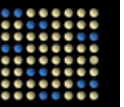
Sources of Variance

- Four main contributors to variance in a microarray experiment
  - Genes (or the probes spotted)
  - Arrays
  - Dyes
  - Treatments (aka varieties)
- Each contributor has a main effect - that which is solely due to that contributor and an interaction effect - that which is due to more than one contributor acting together

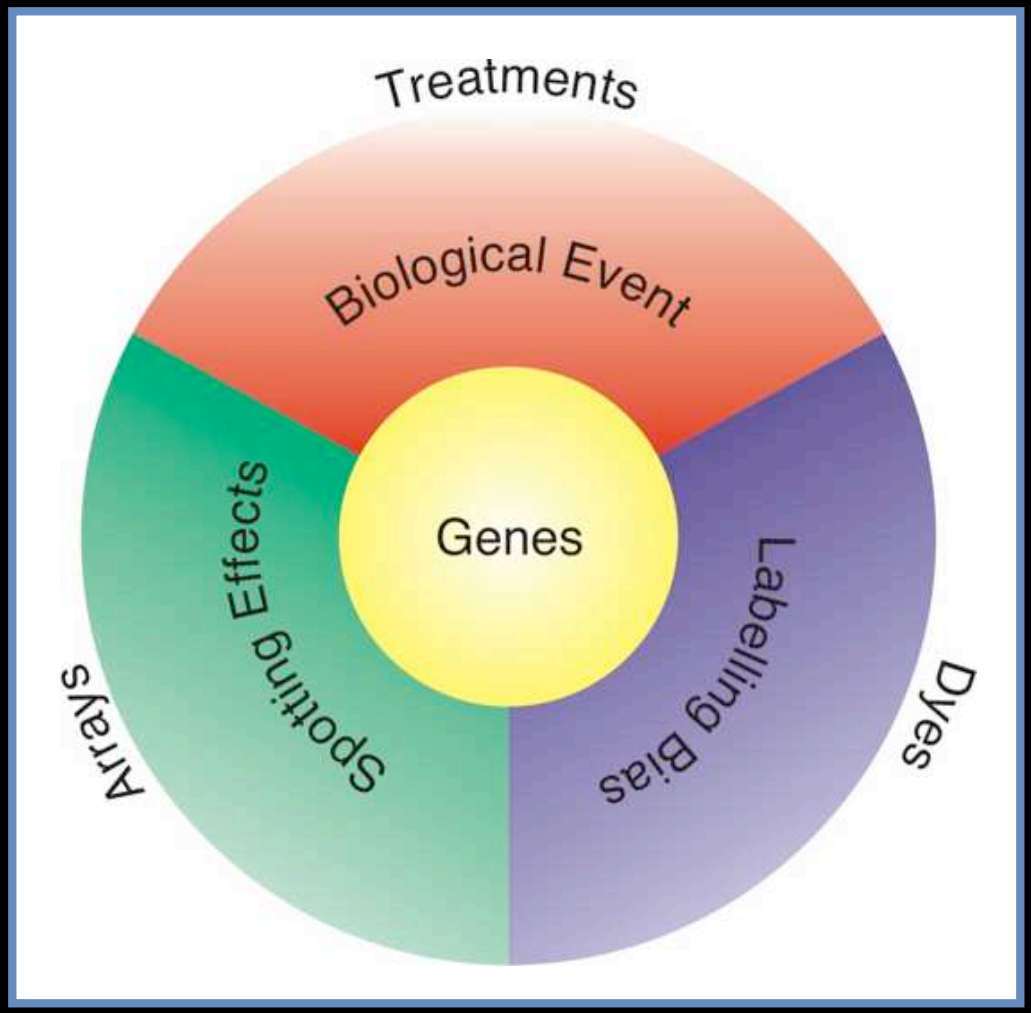


Sources of Variation - Main Effects

- Genes (Probes)
  - Each probe may differ in GC content, length
  - Each cDNA that gets labelled will differ in length, labelling efficiency, abundance
- Arrays
  - Each array is made separately even if in a single batch
  - Each array is on a separate piece of glass
  - Arrays may also be from different lots/batches
- Dyes
  - The dyes differ in photostability, brightness, detectability and incorporation efficiency
  - Dyes are also different from lot to lot/batch to batch
- Treatments
  - Each treatment is done individually and will have its own effects - and may be done at a different time or even by a different person

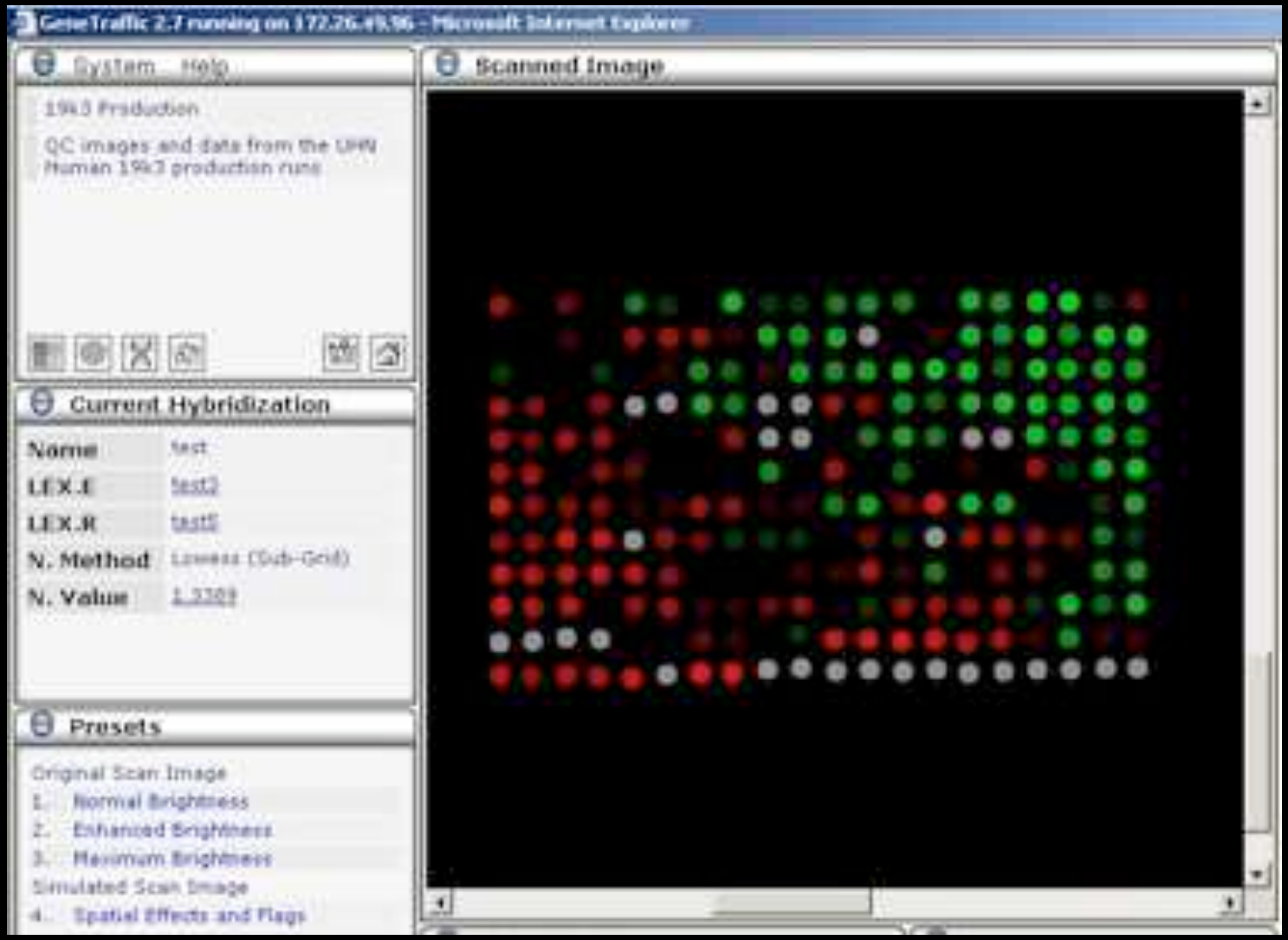


Sources of Variance - Interactions



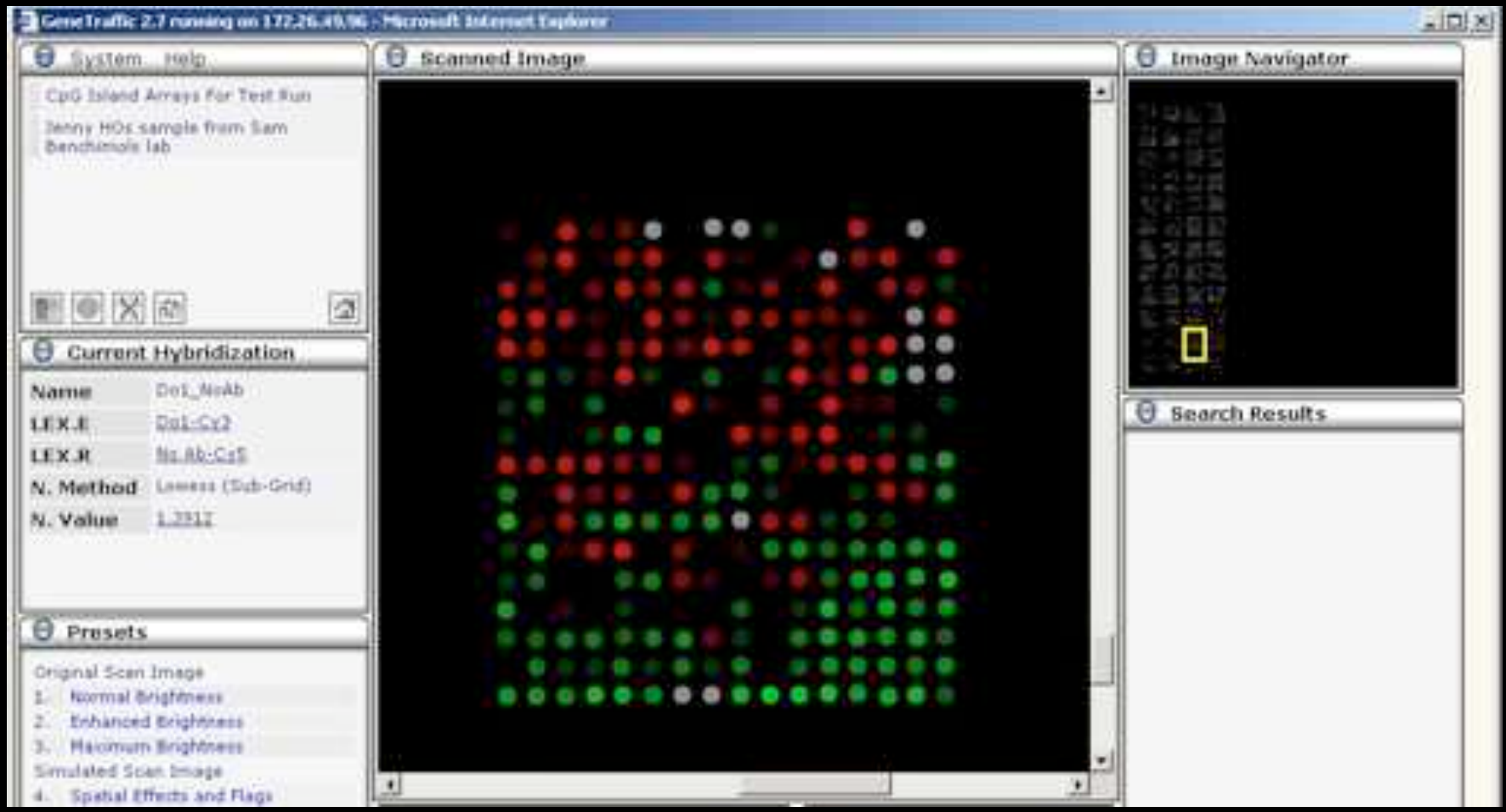
UHN Microarray Centre

Identification of Spatial (AG) Effects



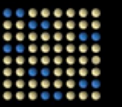
UHN Microarray Centre

Identification of Spatial Effects



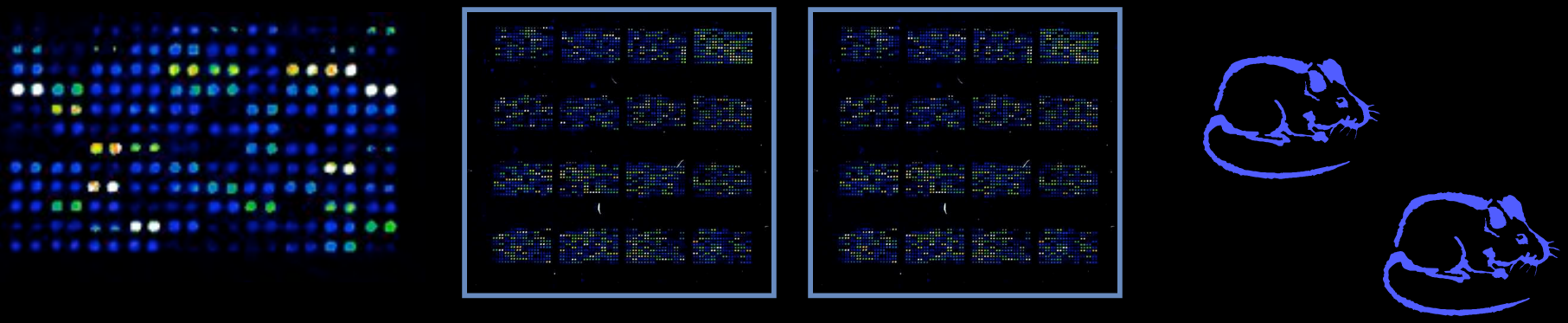
Types of Replication

- Technical Replicates
  - Replicate Spotting - putting more than one spot of the same gene on each array
  - Replicate Arrays - running more than one array for each sample assayed
- Biological Replicates
  - Replicate Samples - preparing multiple independent samples for each treatment



UHN Microarray Centre

Replication



Technical Replicates

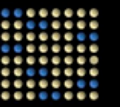
- Technical replicates allow for an assessment of measurement error
- Replicate spots help to identify spotting effects
- Replicate arrays help to identify array effects

Biological Replicates

- A biological replicate is essential as it allows for an assessment of the natural biological variance associated with your treatment

The Use of Replicates

- Technical replicates are becoming less important as the technology matures
- Replicate spotting is no longer of much value to the researcher as the variance between replicate spots (especially side-by-side) is extremely low
- Replicate arrays tend to most useful in the case of reciprocal labellings

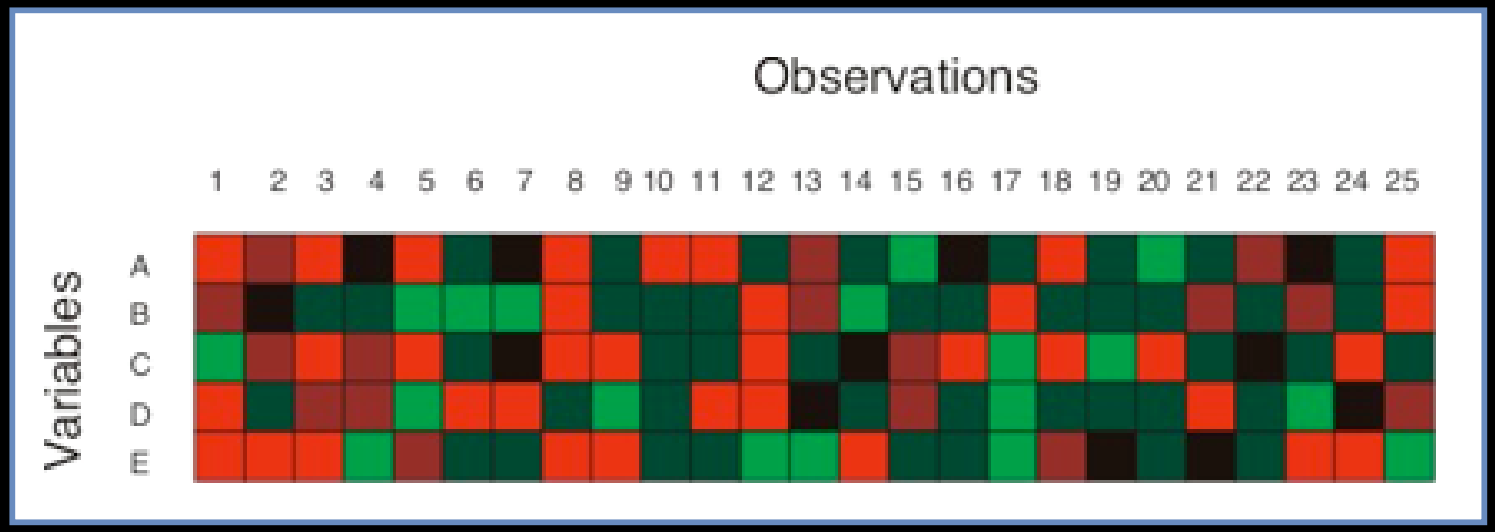


UHN Microarray Centre

About Replicate Spotting

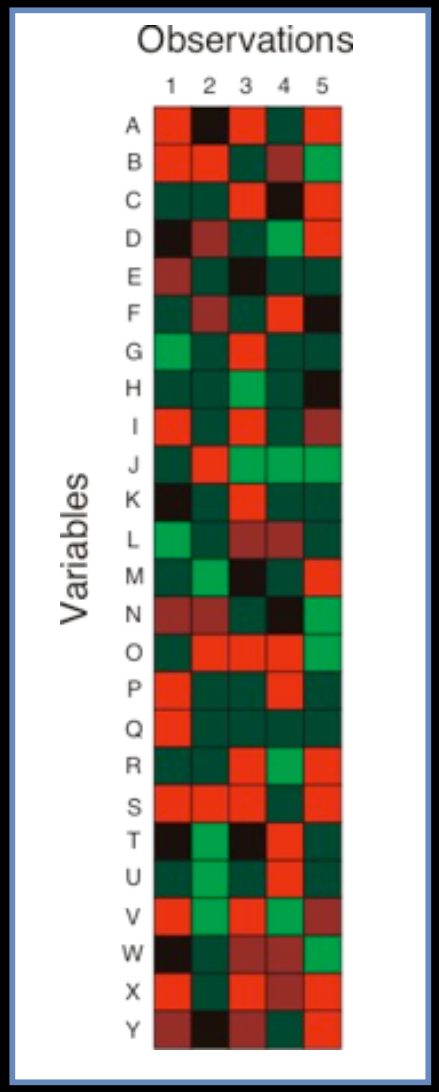
- Replicate spots are most commonly placed side-by-side
  - In doing so - variance is very small - local conditions are nearly identical
  - Also, it is possible for such a design to lower overall signal strength
- Placing replicate spots further apart you are able to measure positional variance
  - Making such arrays are difficult because of how robotic arrayers work
  - The ideal placement is random - but robots don't do random

What the Statistician Wants….



UHN Microarray Centre

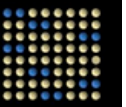
….What the Statistician Gets



UHN Microarray Centre

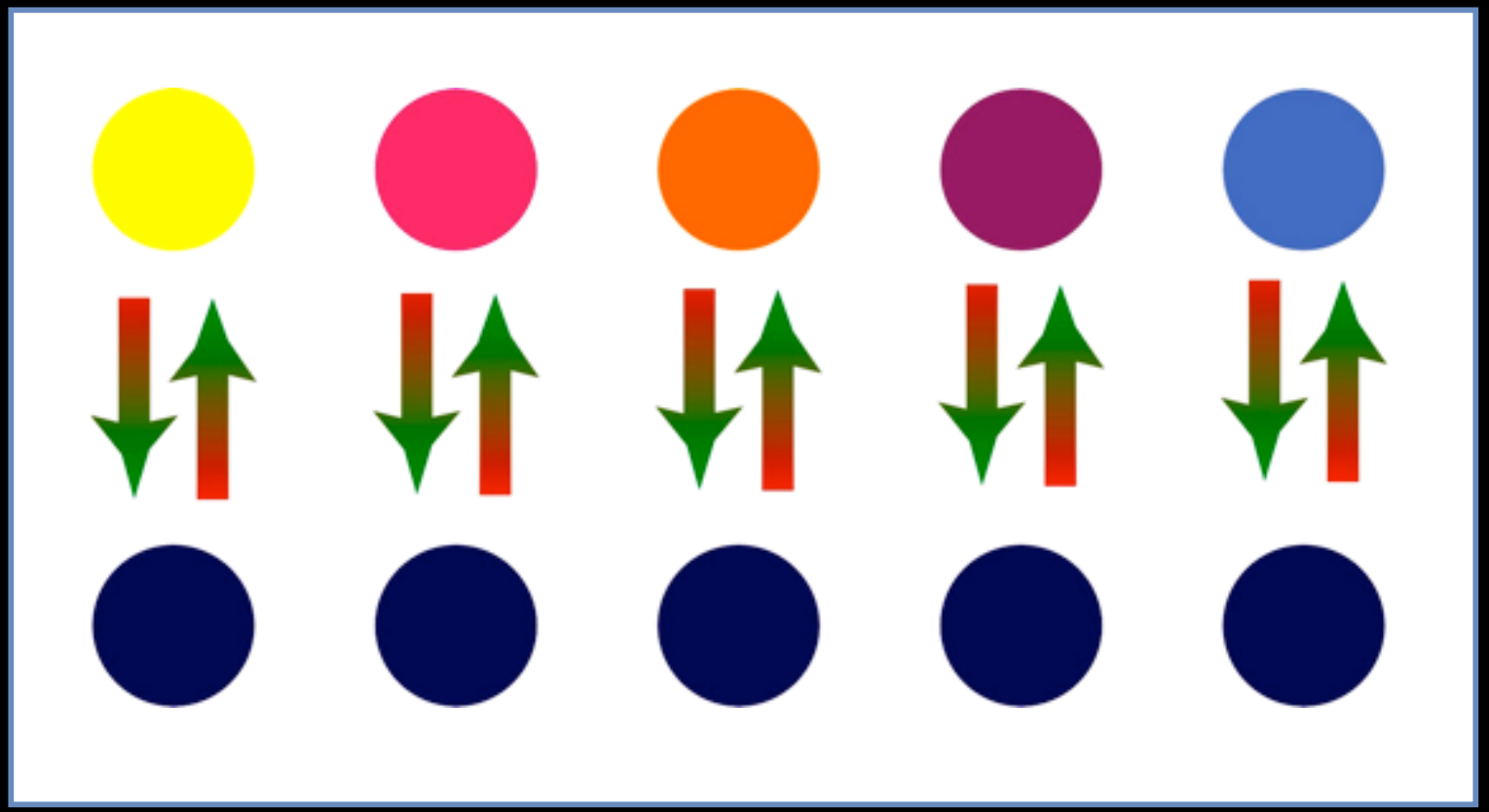
The Reference Design

- This is still probably the most popular experimental design
- It is easy to set up, analyse and interpret, and virtually every microarray analysis package can handle this design



UHN Microarray Centre

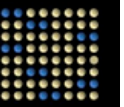
The Reference Design



UHN Microarray Centre

Reference Design Issues

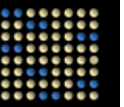
- In a single directional reference design, AT effects (the biological effects) are completely confounded by the GD (dye bias) effects.
- The reference RNA, which is actually of very little interest, is actually the sample for which the most data is accumulated.



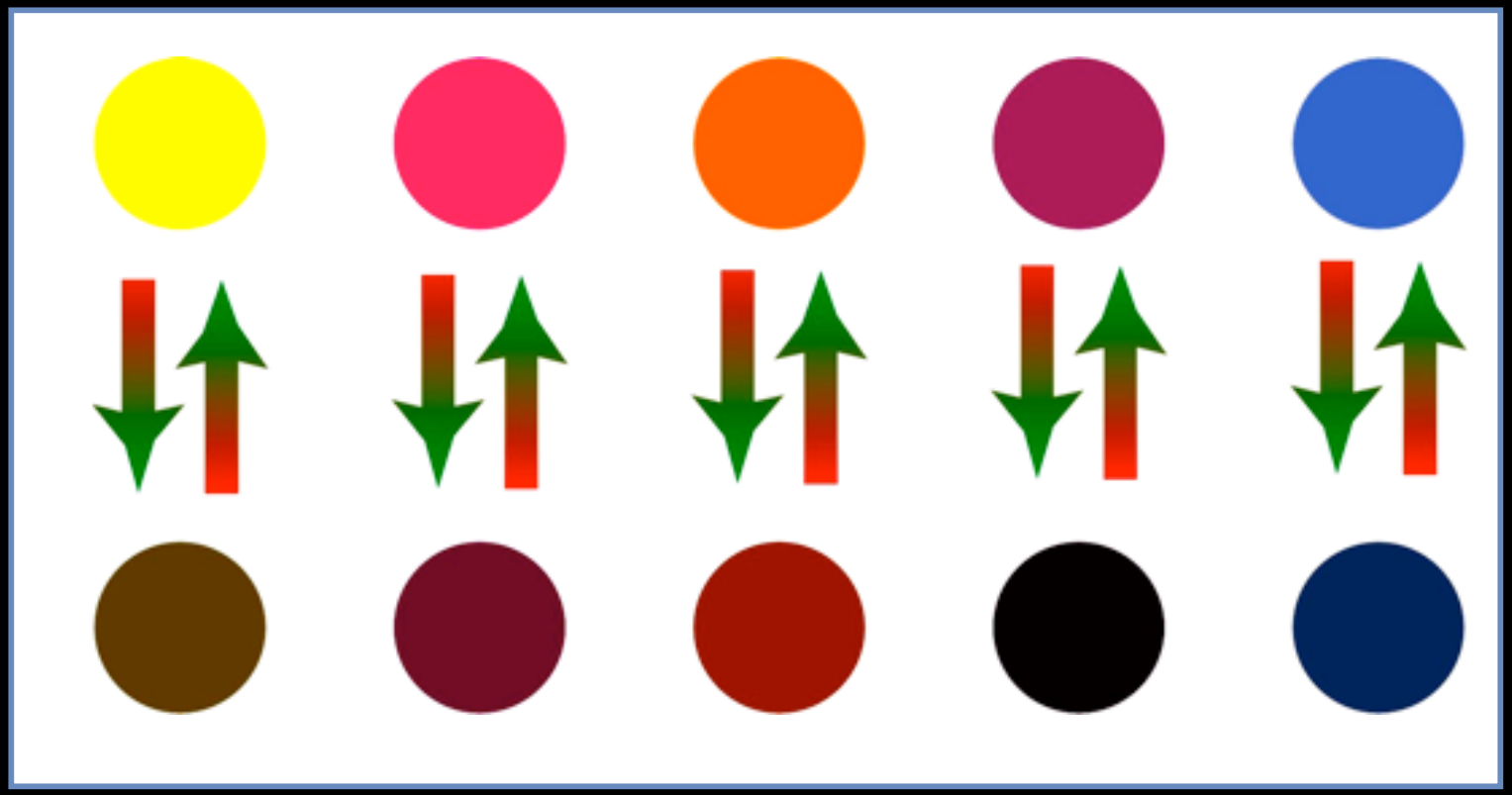
UHN Microarray Centre

The Paired Control Design

- A modification of the reference design
- Rather than using one universal reference RNA, each experimental or treated sample, has a matched control sample
- Particularly useful when looking at tumour samples - compare tumour with matched normal
- Also be useful in time course experiments
- Helps balance biological variance



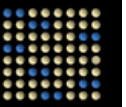
The Paired Control Design



UHN Microarray Centre

Paired Control Design Issues

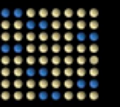
- As with the reference design - when performed unidirectionally dye effects confound biological effects
- Adds difficulty when wanting to compare between treatments (similarities)



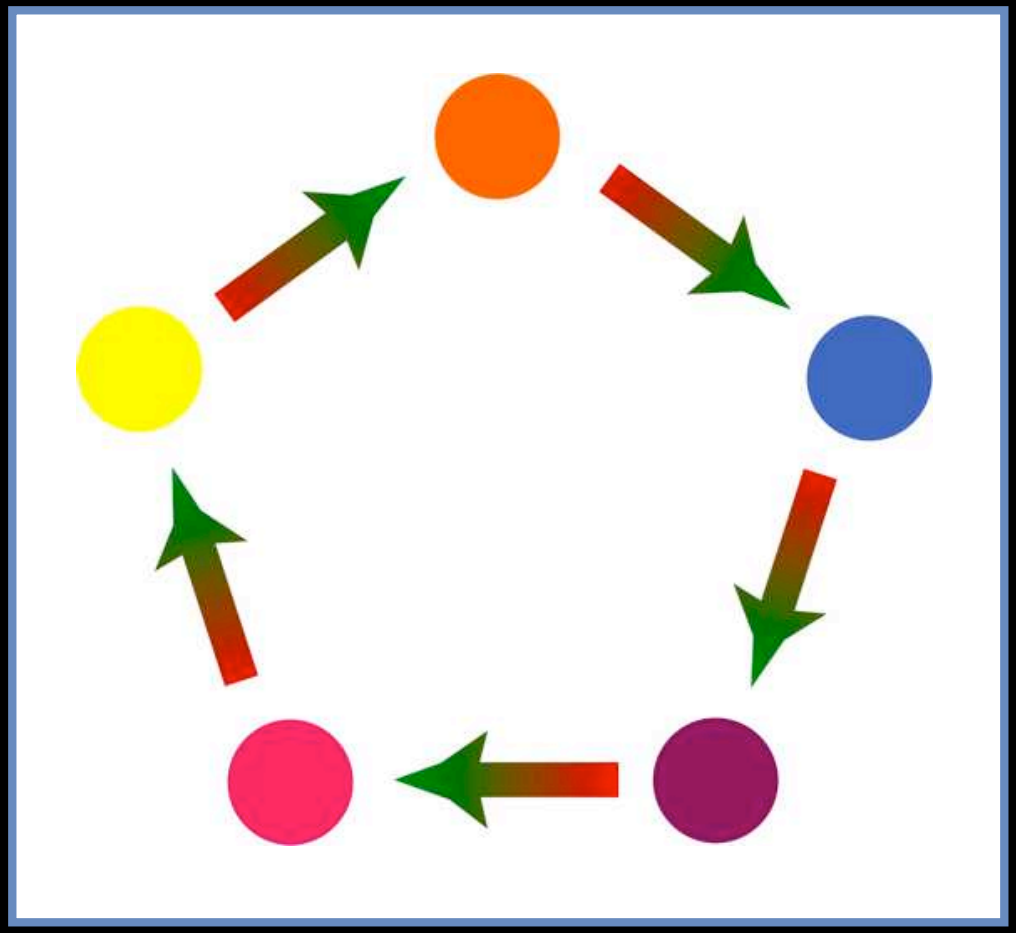
UHN Microarray Centre

The Loop Design

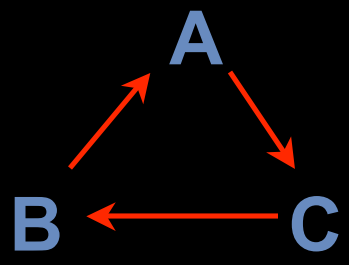
- An efficient design that does not require reference samples
- All data collected is for samples of immediate interest
- Allows for dye bias effects to be isolated from biological effects
- See Kerr and Churchill (2001) Biostatistics **2**:183-201



The Loop Design



Dissecting the Loop Model



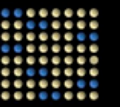
Changes in A vs B - Direct

log(A/B)

Changes in A vs B - Indirect

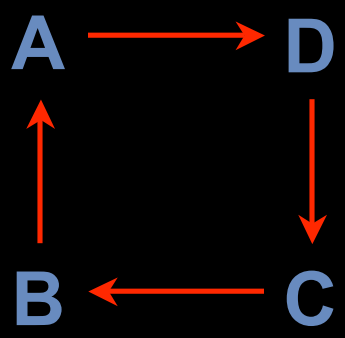
log(A/B) = log(A/C) - log(C/B)

- Two ways to assess changes in A in relation to B - direct or indirect
- Direct method is inherently more accurate



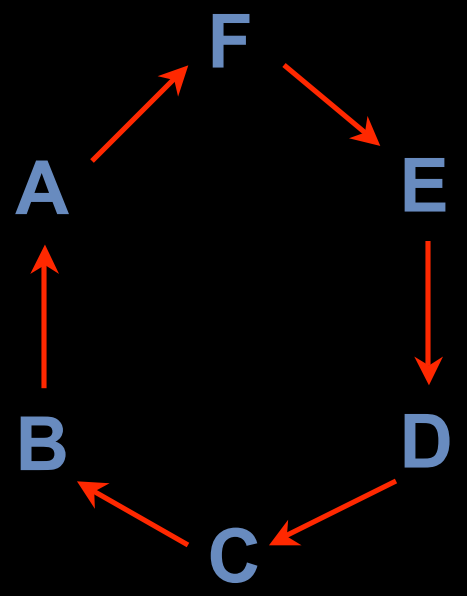
UHN Microarray Centre

Loop Designs - Forced Indirect Measurements



- Possible Direct Measurements:
  - A vs D
  - D vs C
  - C vs B
  - B vs A
- Possible Indirect Measurements (Distance = 2)
  - A vs C
  - B vs D

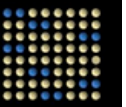
Increasing Samples = Increasing Distances



- 6 measurements can be made with distance = 1
- 5 measurements can only be made with distance = 2
- 3 measurement can only be made with distance = 3

Loop Design Issues

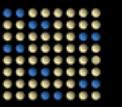
- Increasing numbers of samples lead to increased distances of measurements
- Many measurements must be indirect
- Very few software packages to assist the researcher deal with this design



UHN Microarray Centre

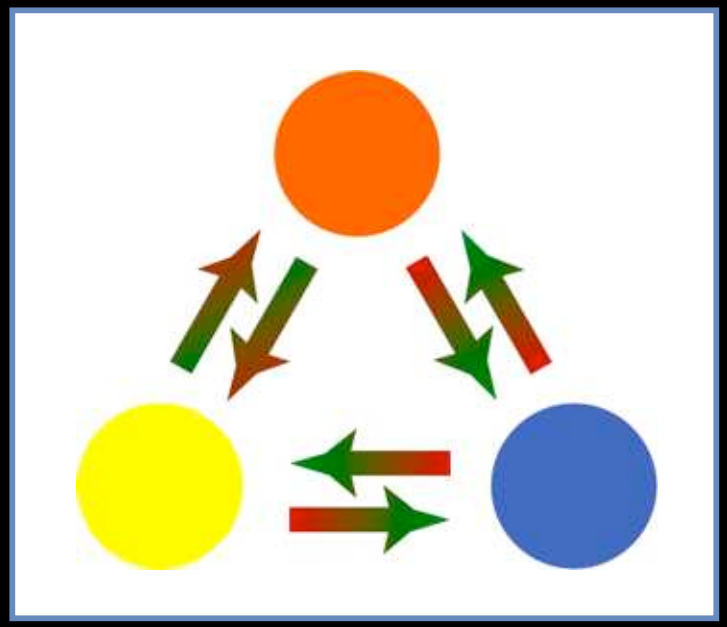
The Network (All Pairs) Design

- Reduces amount of variance in measurements
- Ensures that all measurements are done by the minimal distance
- Eliminates the need for reference samples thus all data collected is for samples of interest
- See Yang and Speed (2002) Nature Genetics, **3**:579-588



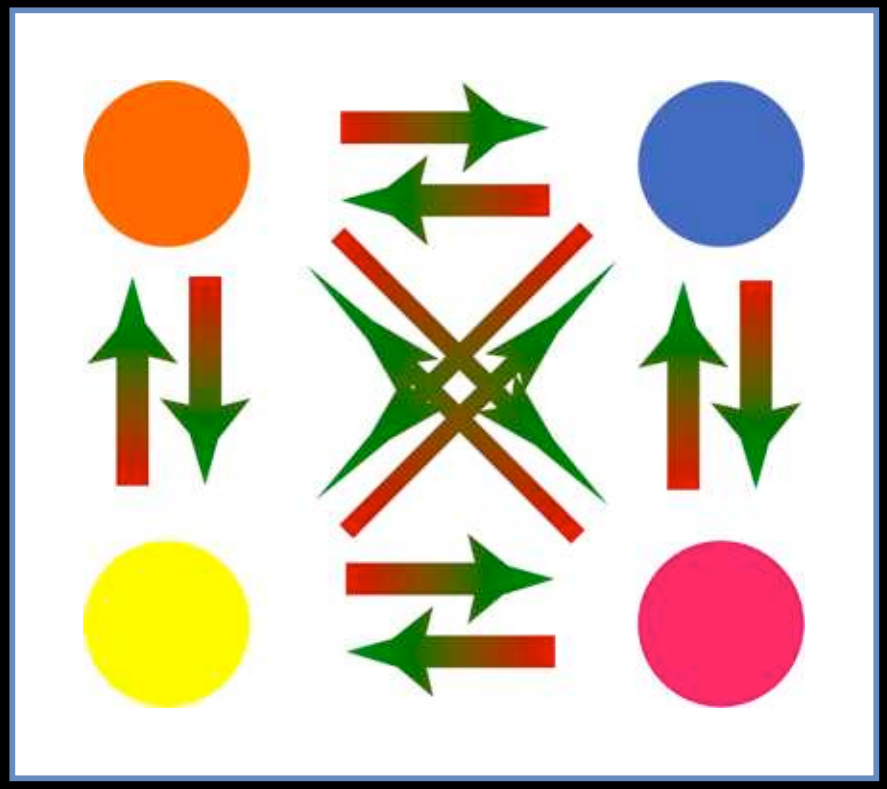
UHN Microarray Centre

The Network ("All Pairs") Design
3 samples



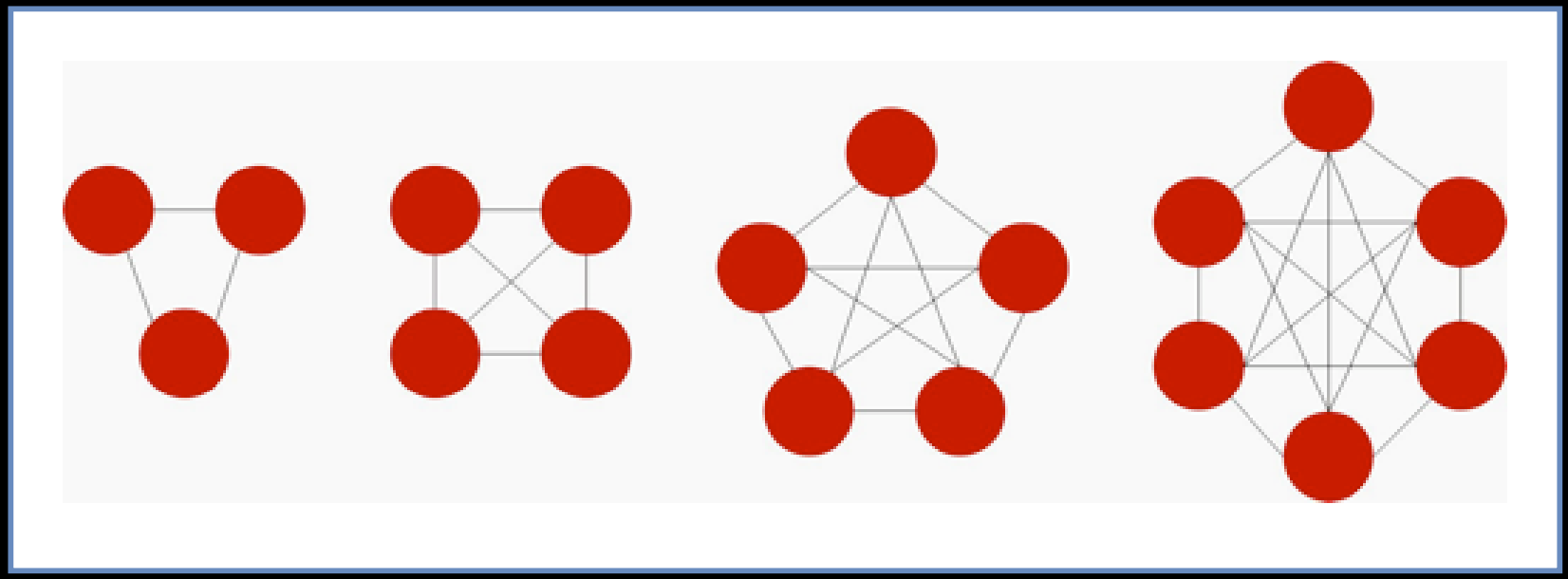
UHN Microarray Centre

The Network Design - 4 samples



UHN Microarray Centre

The Network Design - Scalability Issues



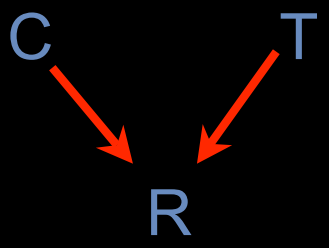
Number of samples	3	4	5	6
Number of arrays	3	6	10	15
RNA per sample	20 µg	30 µg	40 µg	50 µg

UHN Microarray Centre

Direct Analysis Reduces Variance

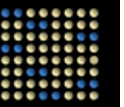
- Direct Analysis - Single Array
  - Variance is σ^2

- Direct Analysis - Two Arrays
  - Variance is $\sigma^2/2$



- Indirect Analysis - Two Arrays
  - Variance is $2\sigma^2$

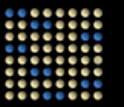
Yang and Speed (2002) Nat. Gen. 3:579-588



UHN Microarray Centre

Network Design Issues

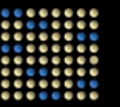
- Scalability is a large problem
- Very little software to assist the researcher



UHN Microarray Centre

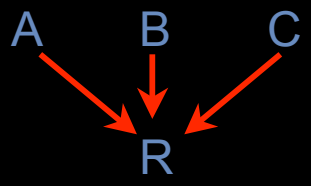
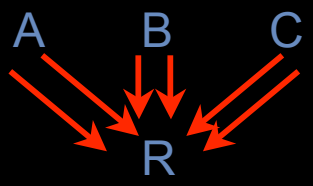
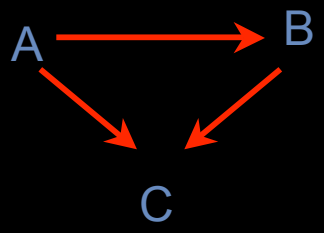
Choosing the Right Design

- Choosing the correct design depends on:
  - The question being answered
  - The amount of RNA present
  - The number of arrays available
- A singular design may not be the best choice, but rather a combination of designs may be better



UHN Microarray Centre

Single Factor Experiments

Design Choice	Number of Arrays	Amount of RNA Required	Average Variance
Indirect Designs			
	3	10 µg	2.00
	6	20 µg	1.00
Direct Design			
	3	20 µg	0.67

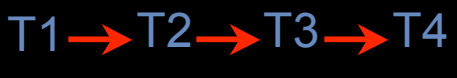
Yang and Speed (2002) Nat. Gen. 3:579-588

UHN Microarray Centre

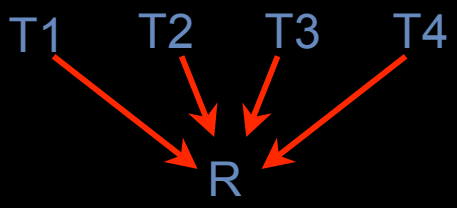
Time Course Experiments

T1 as a common reference

Direct sequential

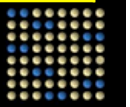


Common Reference



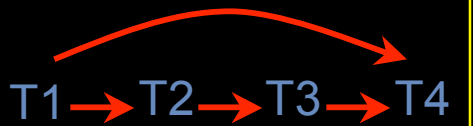
t vs t+1			t vs t+2		t vs t+3	Avg. Var.
T1/T2	T2/T3	T3/T4	T1/T3	T2/T4	T1/T4	
1.00	2.00	2.00	1.00	2.00	1.00	1.50
1.00	1.00	1.00	2.00	2.00	3.00	1.67
2.00	2.00	2.00	2.00	2.00	2.00	2.00

Yang and Speed (2002) Nat. Gen. 3:579-588



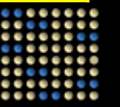
UHN Microarray Centre

Time Course Experiments Cont'd.



t vs t+1			t vs t+2		t vs t+3	Avg. Var.
T1/T2	T2/T3	T3/T4	T1/T3	T2/T4	T1/T4	
0.67	0.67	1.67	0.67	1.67	1.00	1.06
0.75	0.75	0.75	1.00	1.00	0.75	0.83
1.00	0.75	1.00	0.75	0.75	0.75	0.83

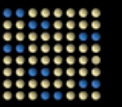
Yang and Speed (2002) Nat. Gen. 3:579-588



UHN Microarray Centre

Setting Up the Experiment

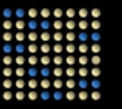
- Consider the following experiment:
  - A time course experiment
  - Human cell culture
  - A drug is added at time 0 and samples are taken at three time points (6 hours, 12 hours and 24 hours)
  - 3 Biological replicates are being performed
  - The effect of the drug over time is of interest



UHN Microarray Centre

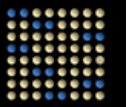
Time Course

- Seeding the cells
  - 3 replicates of the time course are to be performed
  - Seeding all of the plates necessary from one original plate reduces unnecessary variance
- Synchronising the cells
  - Call the cells be synchronised?
  - If not the biological variance will be large
- Picking the right control
  - If you choose to do an experimental design which uses a control sample - what is the right control?
  - Simply using time 0 will not account for differentiation and cell growth changes over time
  - Using untreated samples at each time point only works if you can first synchronise the cells.

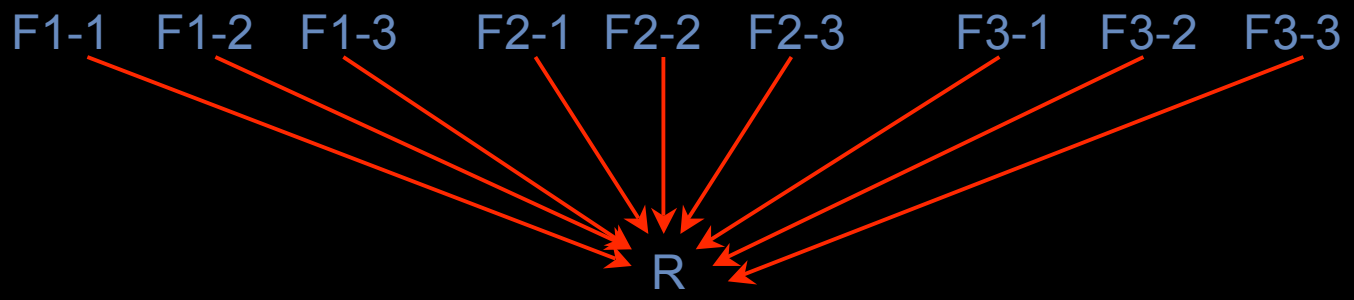


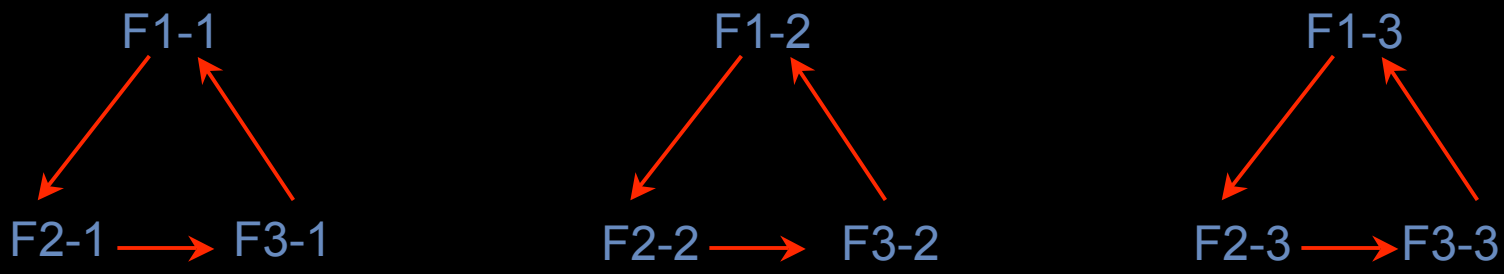
Setting Up the Experiment pt 2

- The effect of different soil conditions is to be assessed on a species of wheat crop
- The wheat is grown in three different fields
- 3 samples from each field are obtained to be analysed (9 samples total)



Possible Design Choices

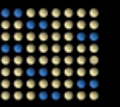




UHN Microarray Centre

A Word About Standards

- MGED – Microarray Gene Expression Data Working Group (www.mged.org)
  - An international group focusing on developing standards for sharing microarray data
- MIAME – Minimal Information About a Microarray Experiment
  - A Standard put forth by MGED to allow for researchers in one group to duplicate work done by another group
  - Standard set of ontologies or terms to describe a microarray experiment



The UHN Microarray Centre

Production
- Patrick Yau (Production Manager)
- Eric Ho
- Quyen Tran
- Tuyet Diep
- Robert Kardish
- Elizabeth Sakac
- Gurbaksh Basi (Tech Support)
- Stephanie Selders
- Joanne Walsh
- Christina Johnston
- Leslie Wyard

Quality Control
- Julissa Tsao

Engineering
- Sasan Raghibizadeh (Manager)
- Yasin Bismilla

Administration
- Susan Alexander

Research and Development
- Mark Takahashi (Research Scientist)
- Pascale Macgregor (Research Scientist)
- Carolyn Modi (Project Manager)
- Robert Clum (CANVAC Core Coordinator)
- Krista Attridge
- Kelly Jackson
- Shani Mintzberg
- Monika Sharma
- Natalie Stickle
- James Paris
- Monique Albert
- Tonya Martin

Informatics
- Carl Virtanen
- Zhibin Lu
- Elizabeth Tillier (Assistant Professor)
- Thomas Liu
- Andrew Smith
- Orianna Wong

www.microarrays.ca/MGED7.html

MGED 7

September 8-10, 2003

Toronto, ON, Canada

UHN Microarray Centre

Thank-You

Please visit us at: www.microarrays.ca