Design Issues in Omics Studies

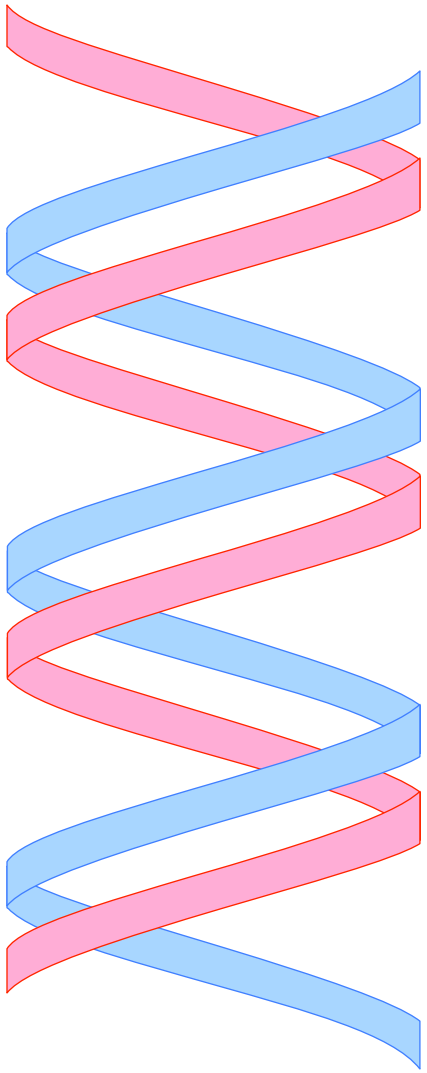
Patty Solomon
University of Adelaide

BioInfoSummer 30 November 2010
Walter and Eliza Hall Institute, Melbourne

Why design matters in omics studies

- We still have little intuition about "what makes sense" in high dimensions.
- So, if we are to use "genomic signatures" as biomarkers, we need to know that the **underlying data and analyses are correct, and checkable.**
- **The role of design** is to ensure the study is **feasible**, that the **questions of interest can be answered**, and to avoid disaster.
- Statistical efficiency comes a long way behind these.

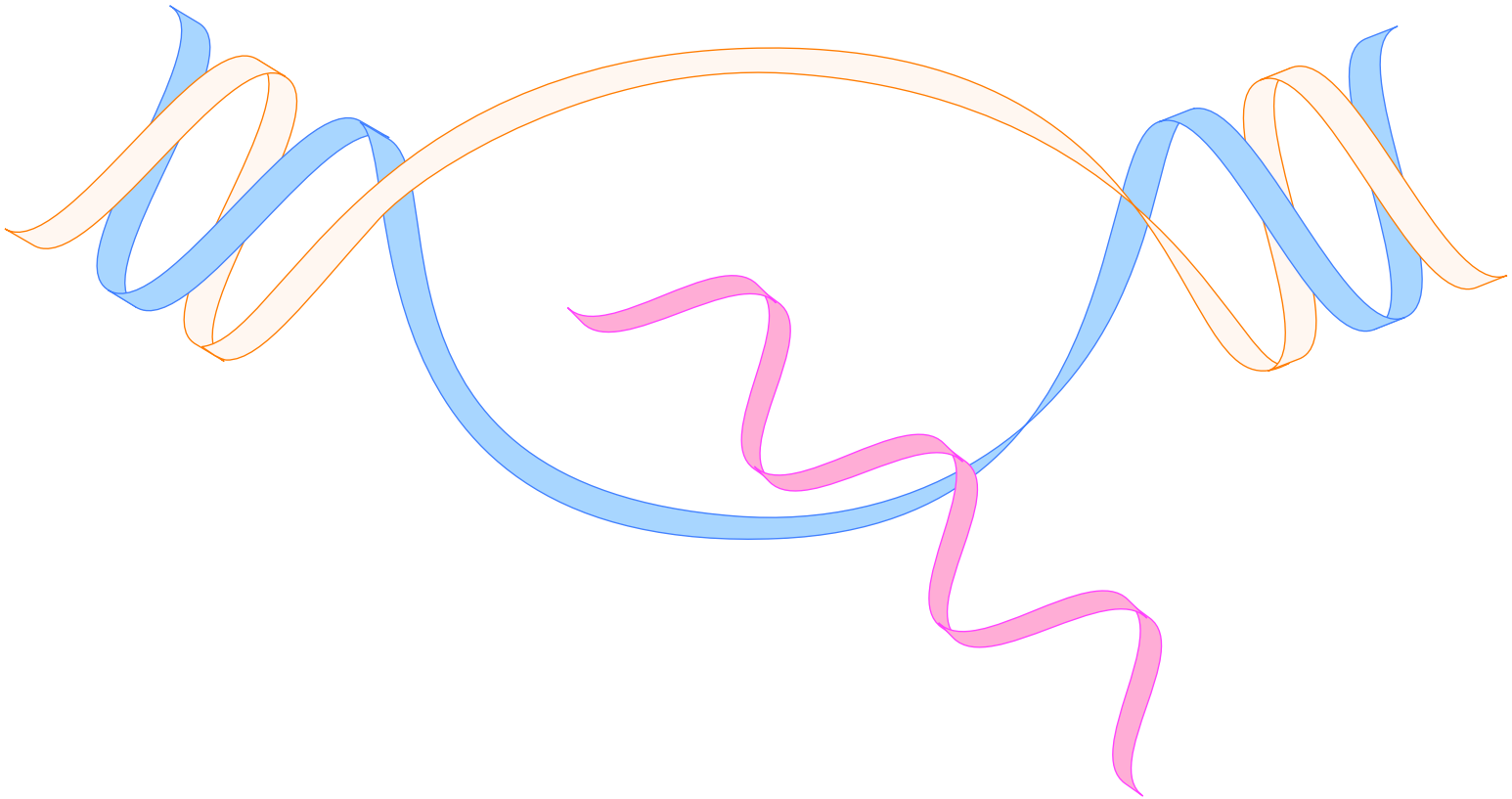
Genome = DNA double helix



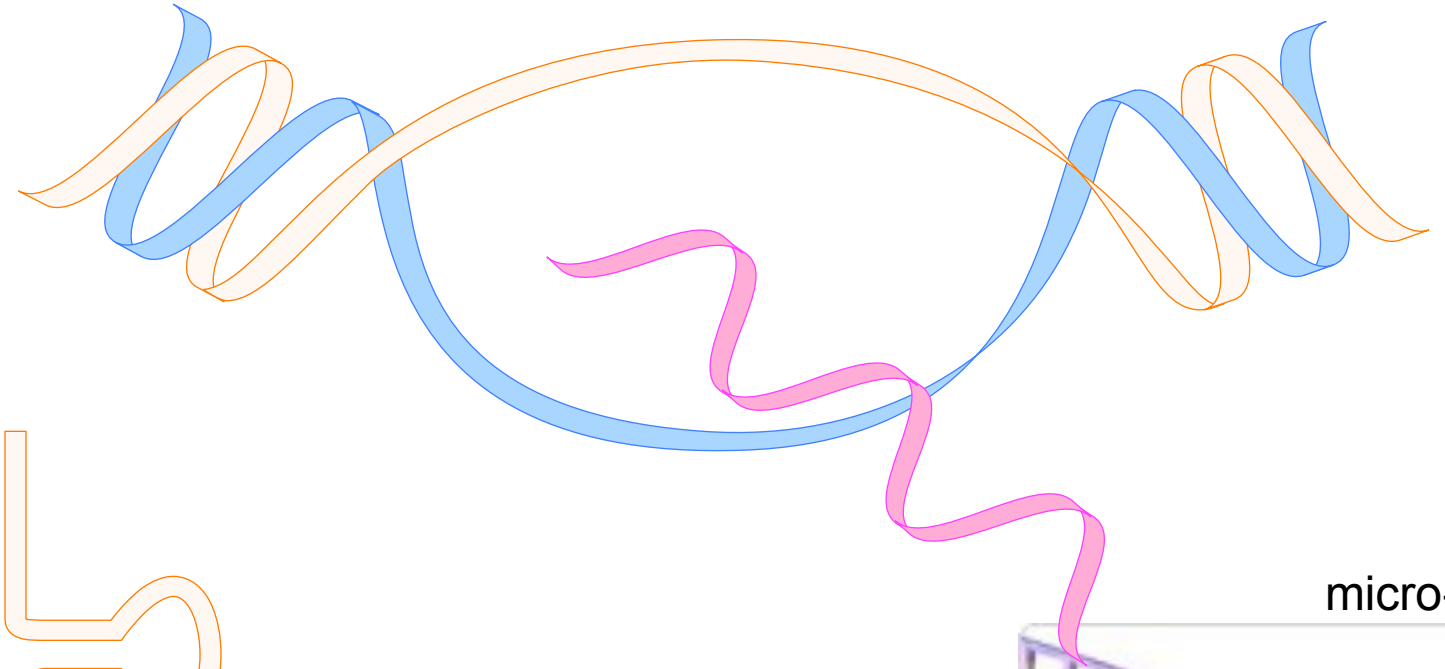
DNA makes RNA makes protein

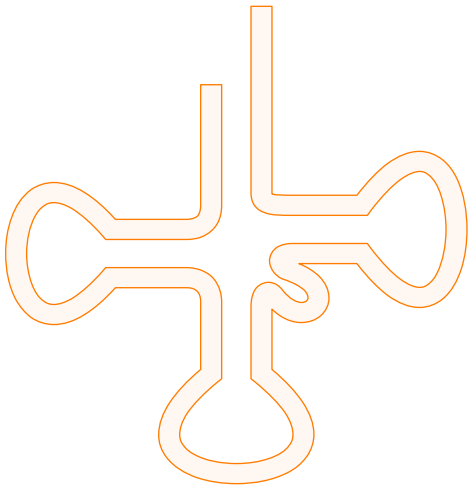
Transcriptome I : coding RNA

mRNA

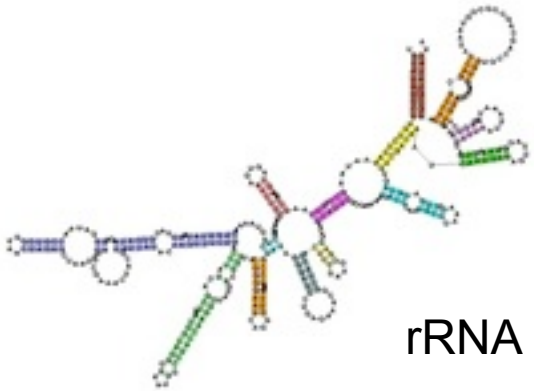


Transcriptome II : non-coding RNA



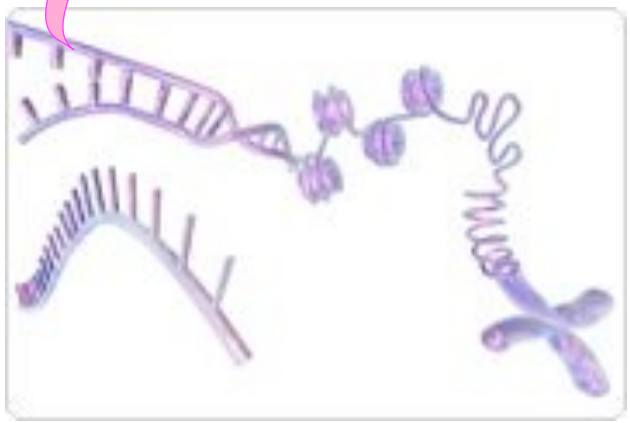


tRNA



rRNA

micro-RNA



Types of omics studies I

- Most are **comparative** e.g., mutant vs wild-type, tumour vs normal, drug-treated vs control cells, ...
- *If the assigned 'treatments' are under the control of the investigator, the study is an 'experiment'*
  e.g., response surface designs.
- Most are **observational** *i.e., the investigator determines which units are studied and the observational process*
  e.g., match serrated (case) with conventional (control) polyps in colon cancer study.
- There is rarely a single objective or hypothesis. Often, the study is a 'screen' where the aim is to identify genes associated with a condition or outcome.

Types of omics studies II

- Classification: of samples into groups given *a priori.*
- Association: of gene expression with e.g. survival time.
- Exploratory: seeking sets of genes sharing observed *patterns,* or sets of samples which *cluster in a meaningful way.*

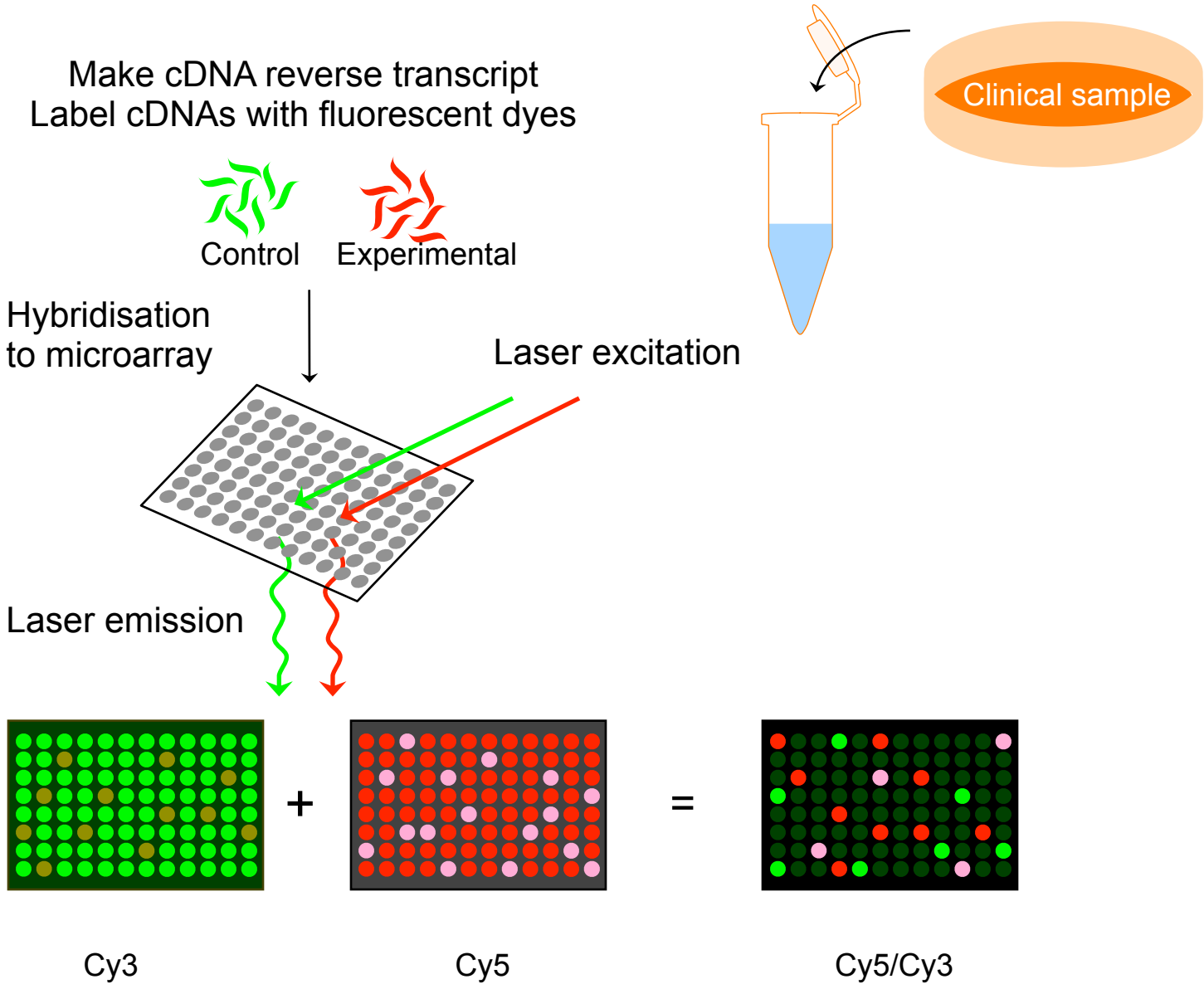
All these studies have characteristics in common:

* lots of *complex measurement processes* in wet lab.

* followed by the application of a piece of *industrial-strength biotechnology equipment*

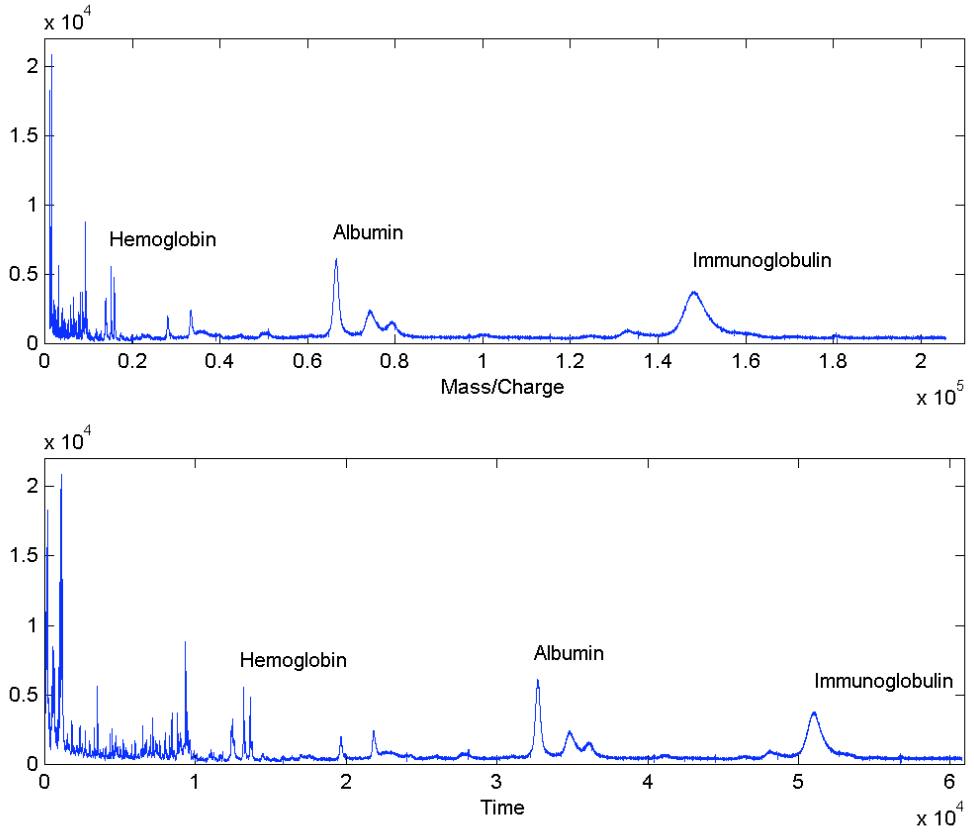
* which produces **lots** of measurements.

Biotechnology I: cDNA microarray assay for gene expression



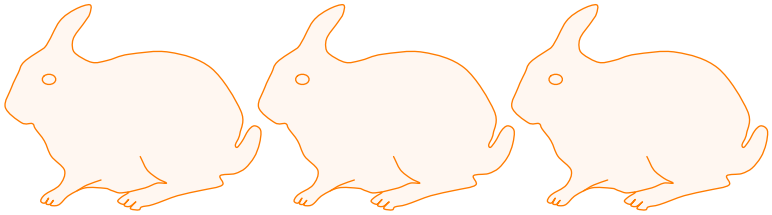
Biotechnology II

- Custom arrays, high-density oligonucleotide arrays, e tiling arrays, ...
- High-throughput PCR
- **Deep-sequencing**
- **Proteomic mass spectrometry**



More characteristics of omics studies:

The notion of **replicate** can be subtle.



Biological replicates usually clear.

But technical replicates arise at a number of levels, depending on how much of the process they share:

Always: # replicates << # measurements

Common characteristics concluded:

Biological variability is reasonably well understood.

But after that, things get more complicated. There are variability hierarchies, depending on context.

For example: variability between organisms/organs

>

variability between tissues

>

variability between cell samples

>

variability between single cells

So, in fitting models to your data, "+ σ^2" may be fantasy.

‘Design’ encompasses all the structural and material aspects of a study or experiment.

The key design principles are:

- Randomisation
- Control (or blocking)
- Replication
- Blinding

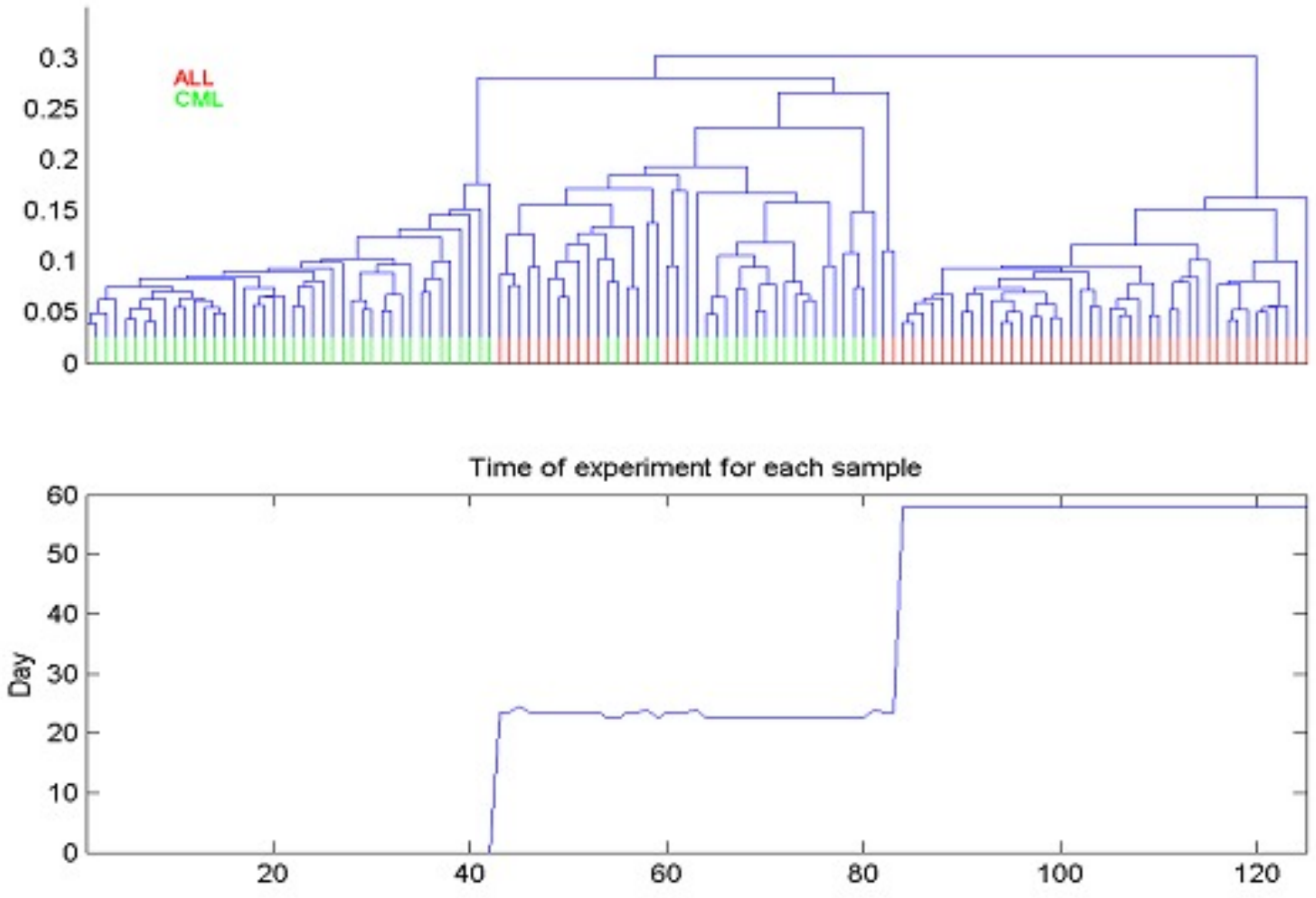
Their purpose is to avoid bias and confounding, among other things.

The SCALE of experimentation in bioinformatics is highlighting the presence of batch effects.

Case Study 1: Cancer subtypes

- Researchers at MD Anderson Texas, conducted an experiment on serum samples from patients with Chronic myelogenous or Acute lymphoblastic leukaemia.
- SELDI-TOF-MS was applied for *protein profiling* of the serum samples.
- **Aim***: to identify the protein peaks that uniquely defined a given leukaemia subtype (CML or ALL).*
- The raw spectra were pre-processed using 'in-house' routines for SPDBC, followed by normalisation to total ion current.
- Performed hierarchical clustering of all samples to evaluate the ability of the peaks to discriminate between cancer subtypes.
- Surprisingly ...

Hierarchical clustering of samples



Run date effects can be larger than biological effects

Case Study 1: Cancer subtypes cont.

- Spectra from QC material run concurrently showed the same clustering pattern as the biological samples.
- Attempts to apply simple additive shifts to align the QC samples to fix the problem failed (Hu *et al*, 2005).

What to do ?

- *Avoid running samples 'as they come in'.*
- Include some members from each contrasting sample in each laboratory-run group.
- If the run groups are large, randomise the run order.
- Record all relevant and clinical information.

Case Study 2: Try to avoid disaster ...

Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

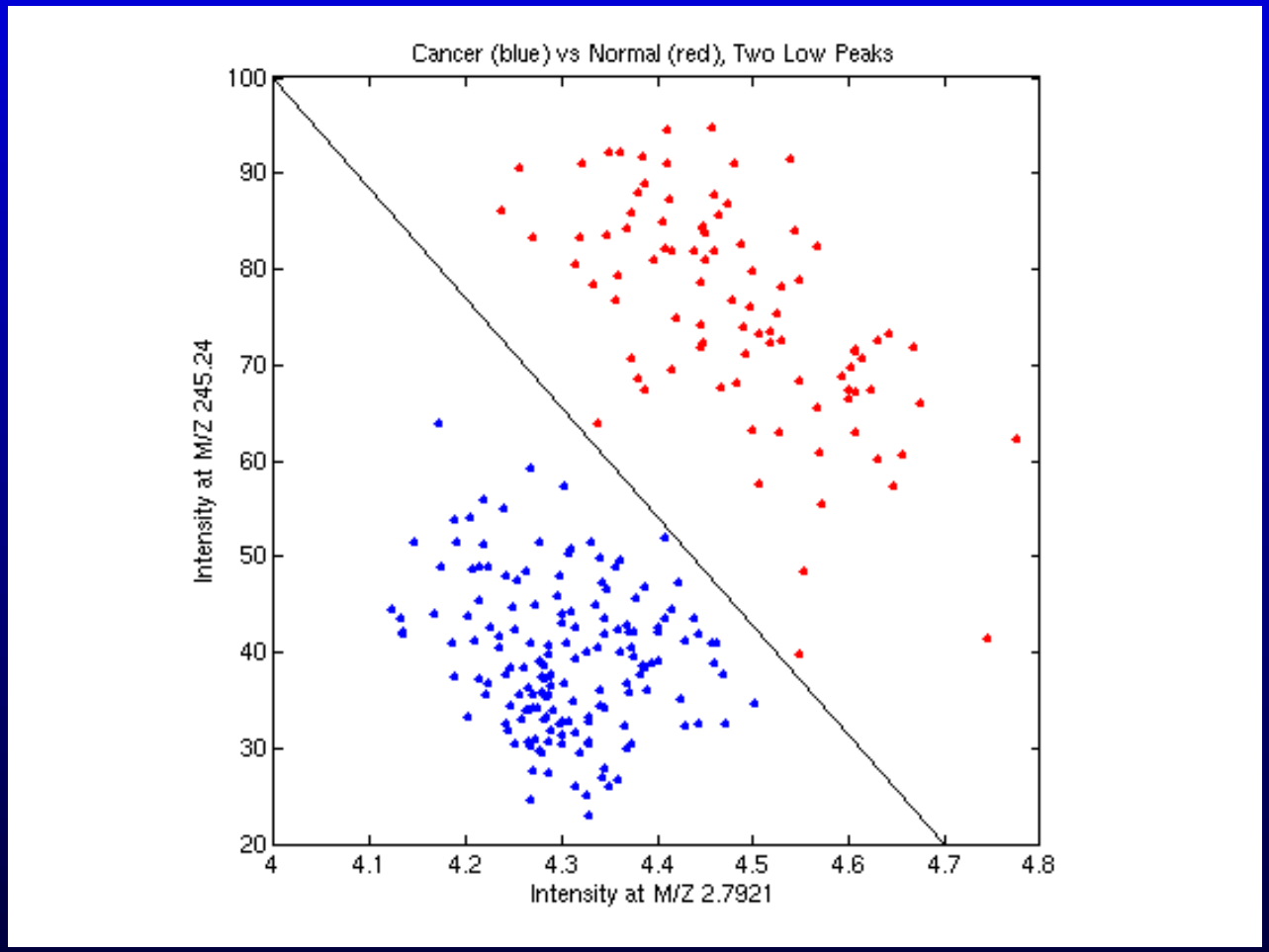
- 100 ovarian cancer patients; 100 normal controls; 16 patients with 'benign' disease.
- Used 50 cancer and 50 normal spectra to train a classifier and tested it on the rest.
- Correctly classified 50/50 of the cancer cases; 46/50 of the controls, and 16/16 of the benign disease as 'other'.

Case Study 2: Almost immediately, various questions about oddities in the data are raised by Keith Baggerly and others...

- The results are not reproducible from the 'same' data.
- There was an apparent change of protocol near the end of the dataset.
- No time-m/z calibration.
- No evidence that the order of processing was randomised.
- *Perfect classification of peaks is achieved in the "noise" region of the data (see next slide) ...*

Case Study 2: Disaster unfolding

Another Bivariate Plot: M/Z = (2.79,245.2)



Perfect Separation, using a completely different pair. Further, look at the masses: this is the noise region.

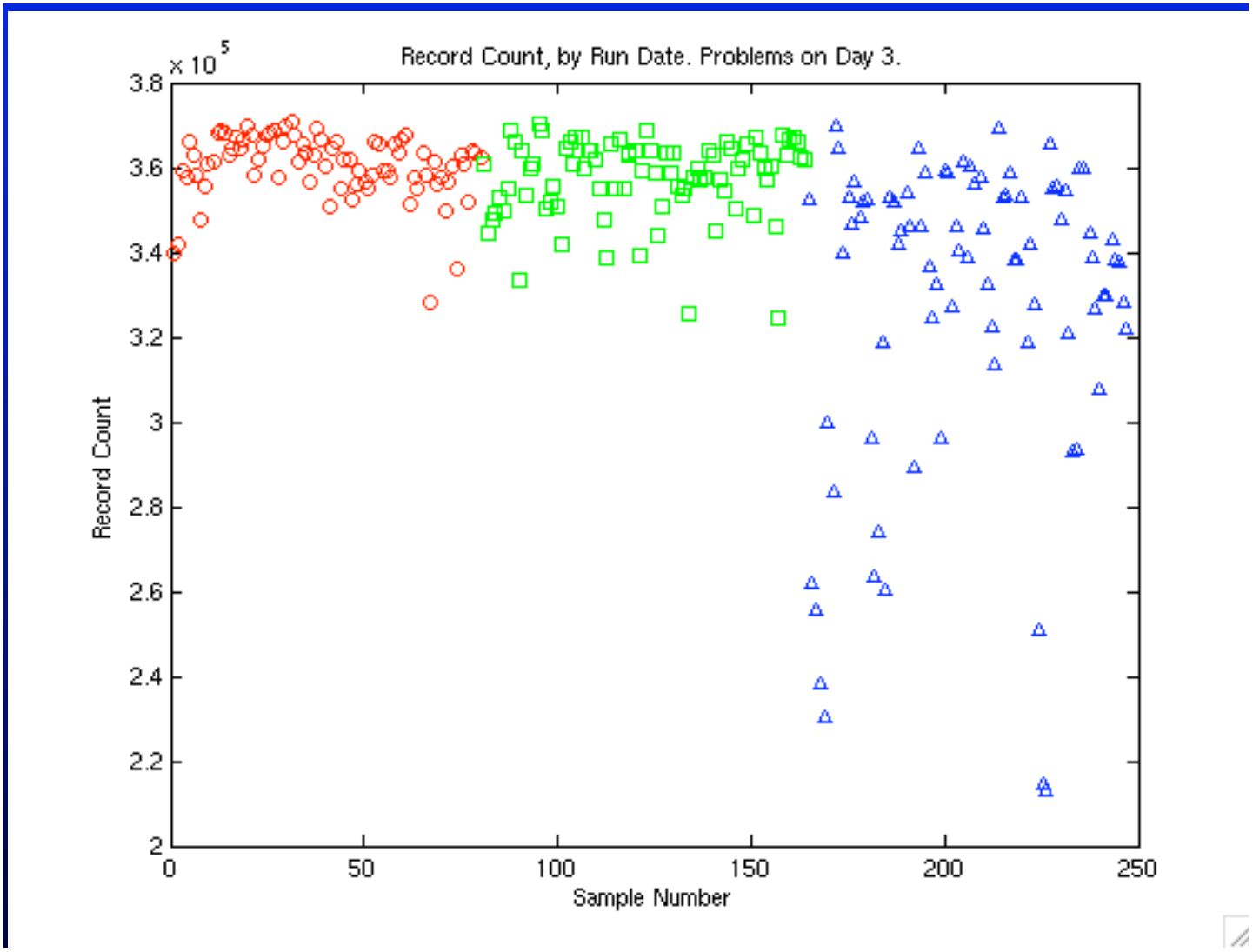
Case Study 2: The abortive followup

- In January 2004, three companies announced plans to offer a "home brew" diagnostic assay called **Ovacheck**.
- In July 2004, Conrads *et al** processed samples with their original SELDI technology and also with a higher resolution instrument called the QqTOF. *They added some QA/QC steps to remove bad spectra.*
- *They demonstrated 100% sensitivity and 100% specificity for identifying cancer from normal,* and stated that this "**emerging paradigm**" is ready to go to a full-scale clinical study.

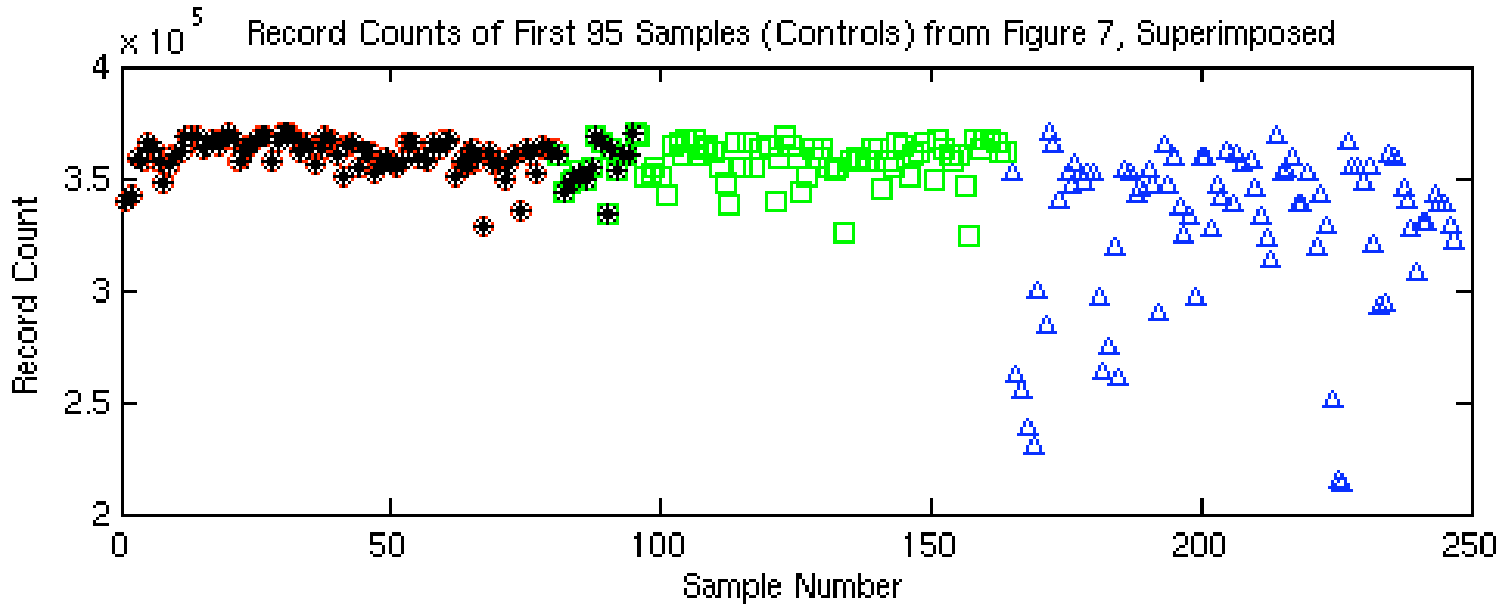
What was going on?

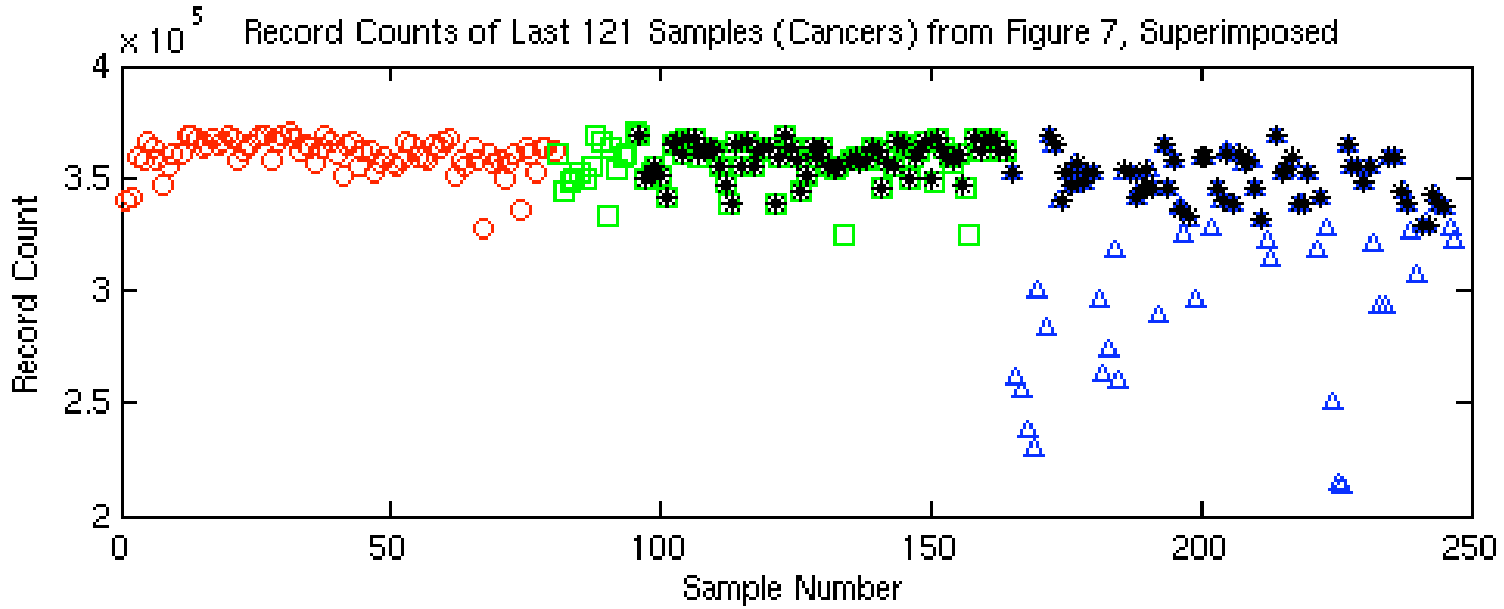
* Endocrine Related Cancer 11 163-178, 2004

Here is their Figure 6A



Day 1 Day 2 Day 3





All of the controls were run before *all* of the cancers

Case Study 2: Concluded

- A better machine will not save you if the study design is poor!
- Obviously, there is no way a woman should be told she needs an oophorectomy based on this test.
- *In June 2004, the US FDA blocked its use "pending further validation" ...*
- and we are still waiting for such.

Case Study 3: 1000 Genomes Project

OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

NATURE REVIEWS | GENETICS

Their definition of batch effects includes laboratory conditions, reagent lots and personnel.

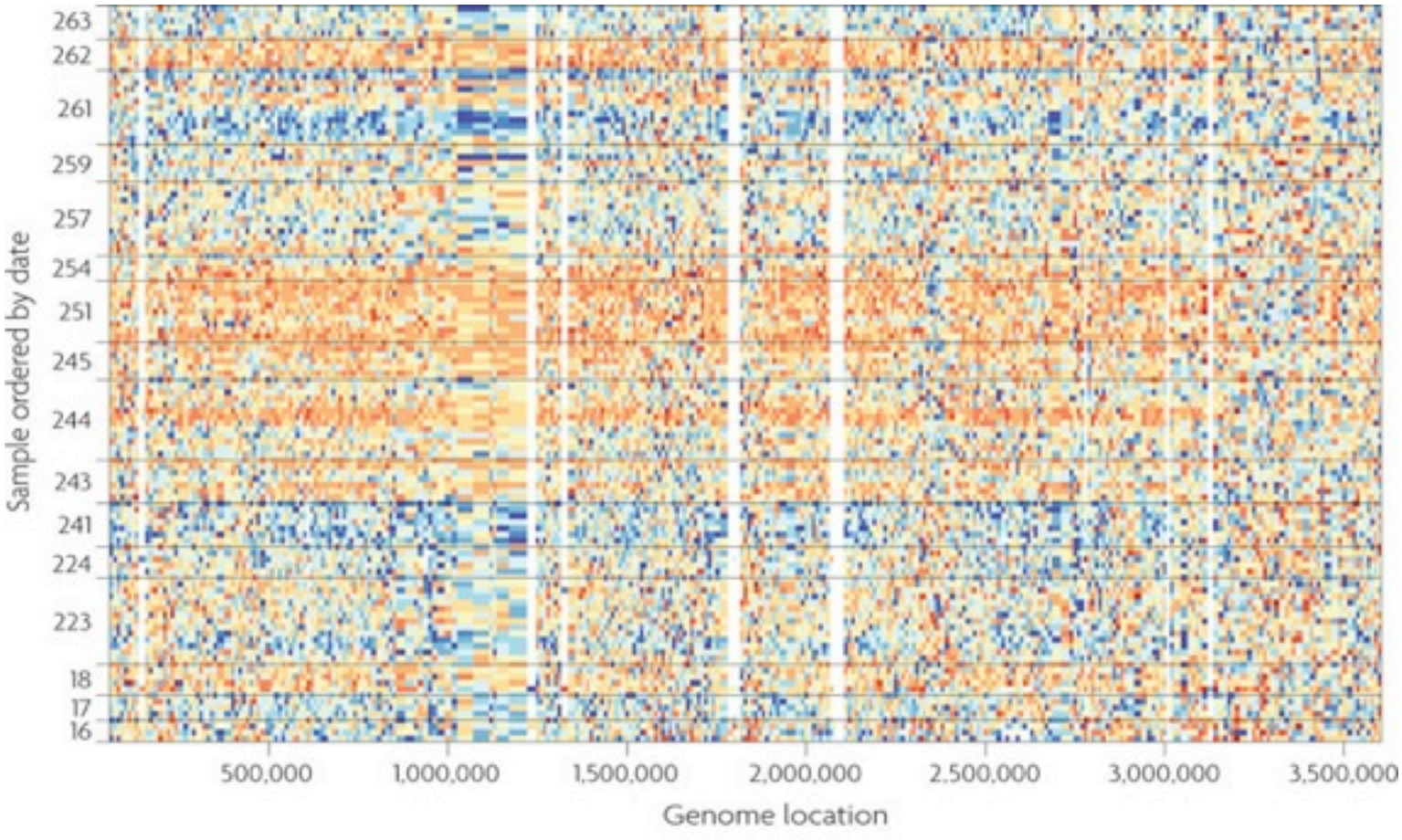
When these effects are correlated or (worse) confounded with the biological outcome of interest, we get wrong answers.

Case Study 3: 1000 Genomes Project

- We consider the analysis of second-generation sequencing data from 131 individuals in 6 Hapmap populations,
- and a 3.5 Mb region from chromosome 16.
- Chromosome 16 was binned into 10Kb regions and *total number of reads aligned to each bin for each individual* was the *statistic* (counts were then quantile normalised).
- Feature data were standardised across samples: blue 3 s.d. < average and orange 3 s.d. > average.
- Samples are ordered by processing date ...

Batch effects for second-generation sequencing data

Each row is a different individual, all from same processing group



Nature Reviews | Genetics

The largest batch effect occurs between days 243 and 251

Case Study 3: Continued

- Further analysis showed that 32% of features were associated with processing date.
- But up to 73% were associated with the **second Principal Component** (ranked in order of variability explained) - *in a situation where the PCs cannot be explained by the biology.*
- *This strongly suggests other (unknown) sources of batch variability are present.*

What to do?

Case Study 3: Concluded

- Use clinical trial design principles, including *distributing biological contrast groups equally across centres in a multi-centre study; use identical protocols.*
- Randomise run order within centres, etc, as before.
- Record all relevant information and clinical variables so that potential *confounding effects and unwanted structure* can be checked by the statistician (e.g., using PCA or clustering).
- Use statistical analysis solutions as well, and if possible.

Stages of a study/experiment

1. Planning and preparation

The statistician should be involved from the outset to discuss:

- the **aims or hypotheses** (if you have any)
- the **study population** (people, organs, cadavers, cell lines, ...) especially *sources of variability*, potential *replication*, choice of *control;*
- the **type of study** (observational, time course, ...)
- the **choice of technology**/platform
- **important clinical and epidemiological variables** (e.g., age and tumour stage at diagnosis, date of death).

2. Samples and cells

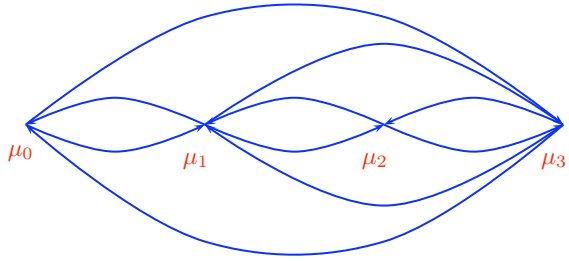
- This stage is about **obtaining the cell samples** *without selection bias* from the original organisms/organs.
- Depending on context, this involves **choosing the samples**, the **number of samples**, **extracting** tissue/blood/cell samples from these. *Using randomisation wherever possible. And blinding.*
- This may take days/weeks/months, so **time** may be an issue (temporal trends, mode of **storage**, time spent in storage, **degradation** of samples, ...).
- Records should be kept - *dates, changes in lab. personnel, how the lab. technician performs the extractions, protocols, ...*

3. Molecular samples

- This stage is about starting with cells samples and ending up with **molecular samples** for measuring.
- That is: **extraction, amplification, pooled vs unpooled**, **probe labelling**, etc.
- And, deciding on the *nature and amount of replication*.
- Amplification is a major source of variation, but there are others, such as quality of RNA.
- Record everything, including times, reagents and equipment used, protocols, operators, ...

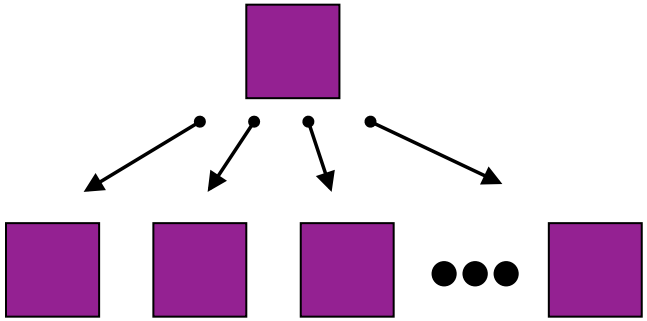
4. Assay design

- This stage is about **assigning** the molecular samples to components of the technology.



- For example: pairing(dye-swaps) in two-channel microarrays (*direct comparisons versus reference design*); allocating samples to runs in single-channel (Affymetrix) microarrays.

Day 0 3 6 9



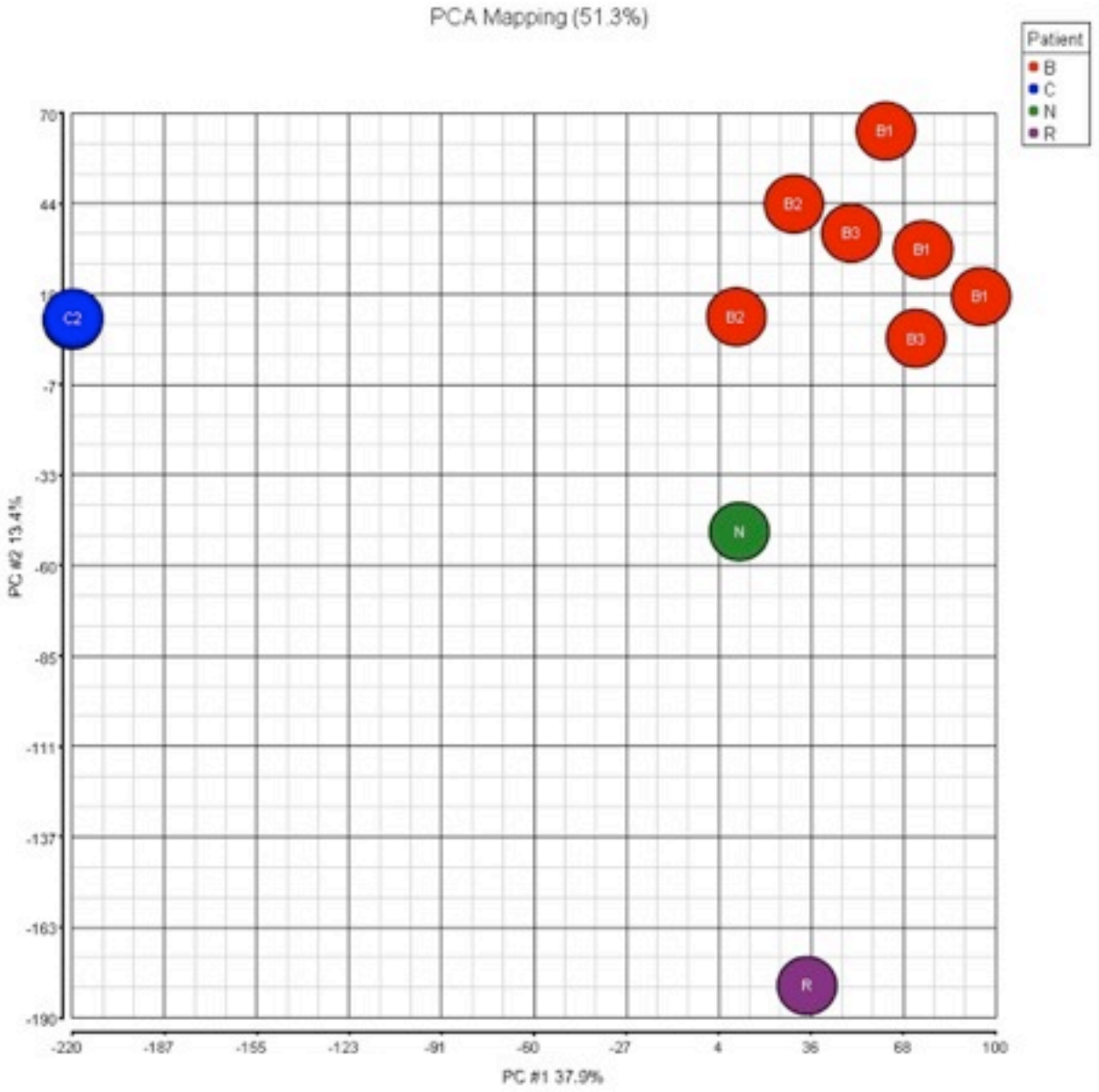
- Issues to watch out for: *choice of reference* (esp. in common reference design) or control.

References: Kerr & Churchill, Glonek & Solomon, Mukherjee, and others

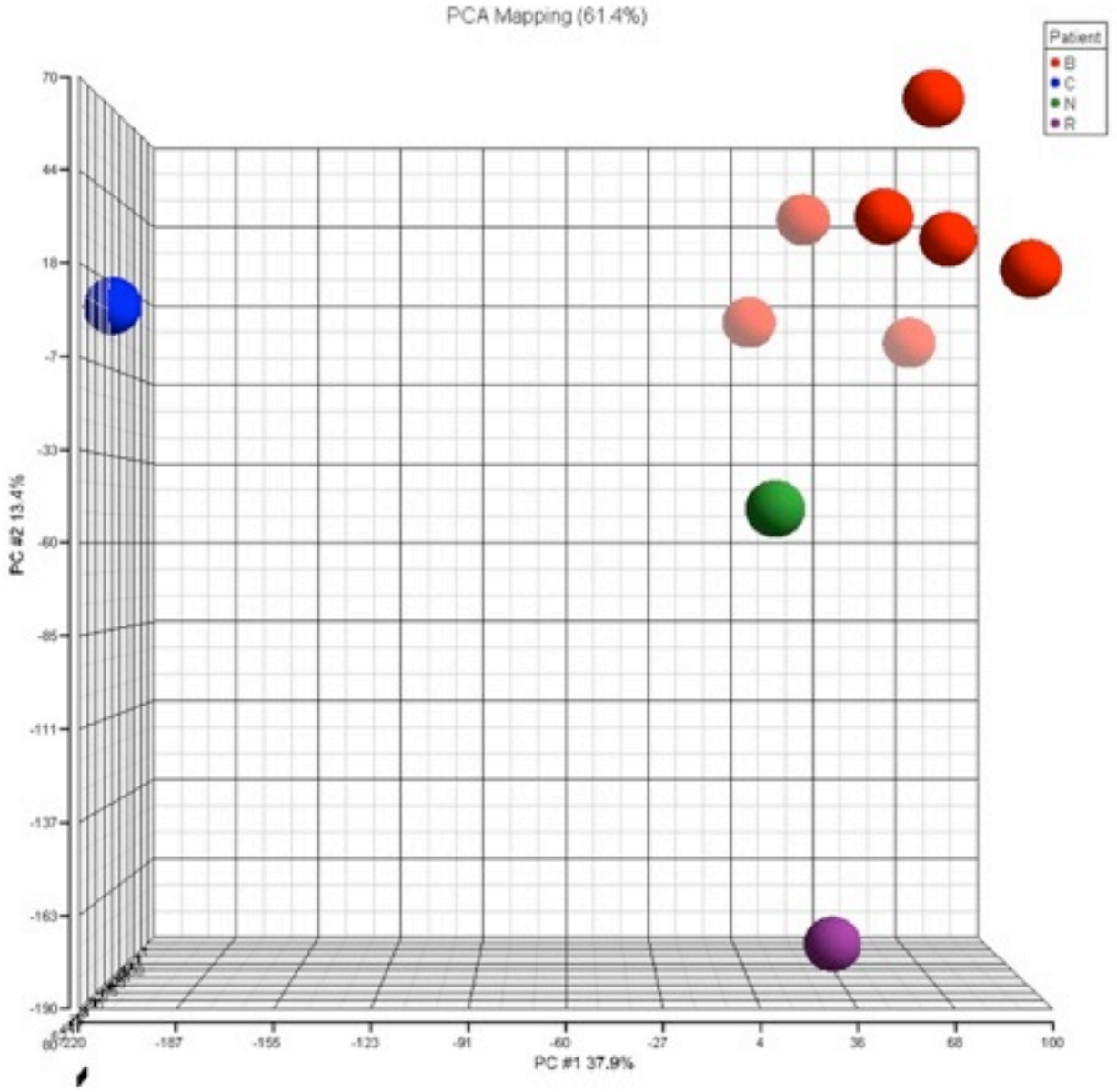
Case study 4: Cushing's Disease

- Is a rare disease of the pituitary gland which causes it to release too much ATCH. Symptoms in adults include obesity, 'moon face', and a large adrenal gland.
- Researchers in Adelaide approached the AMF to conduct a microarray experiment on **3 brothers** with Cushing's Disease; a Pedigree of the family was also being established.
- Tissue samples from the **brothers** and **4 controls** were hybridised using Human GeneChip Arrays.
- *Early on, the **choice of controls** was an issue, because 'control tissue' was scarce ...*

Cushing's Disease and PCA



More Cushing's Disease and PCA



Case Study 4: Cushing's Disease

- The 2 Clontech pooled control samples were from male and female cadavers, 15-61 years old, with ? tissues.
- Unlikely to help shed light on the genotype of Cushing's Disease.
- Controls like Bruce, matched on sex and age would be better, if they could be obtained.
- *PCA is a useful exploratory tool - we may know structure in the data exists, but it can tell us how to respond.*

5. Assay execution

- In this crucial stage of the study, the (now prepared) molecular samples are **"run"** on the equipment.
- This is probably the most important stage for the **statistician** to be directly involved, but they usually aren't.

Case Study 5: Biostatistics vs Lab Research

Here's how *not* to consult with your statistician ...


xtra
normal

A few more recommendations ...

- Your data will become publicly available ...
- documentation is often poor - make sure yours isn't, and
- ensure your results are **reproducible.**
- You **do not** want your research to feature as one of Keith Baggerly's *Case Studies in Forensic Bioinformatics*.

Acknowledgements

Chris Bagley, Hospira

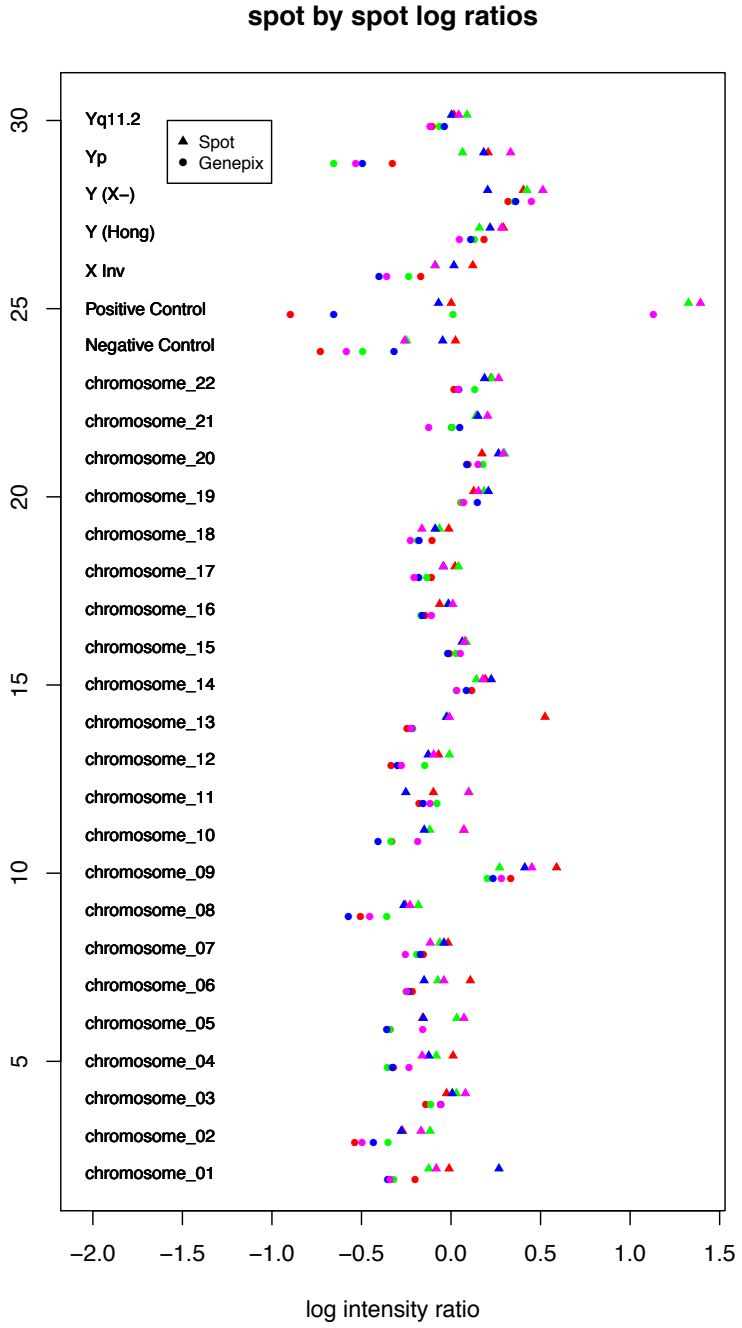
Keith Baggerly

Greg Goodall

Terry Speed

Custom array for chromosome screening

Spot
vs
Genepix



Custom array for chromosome screening