Statistics in high-content biology

Rebecca Walls

Advanced Science & Technology Laboratory

ASTL
advanced science and technology lab

AstraZeneca
life inspiring ideas

Outline

- Introduction and aim of high-content biology
- Predicting liver toxicity *in vivo*
- Distinguishing distinct modes of compound action

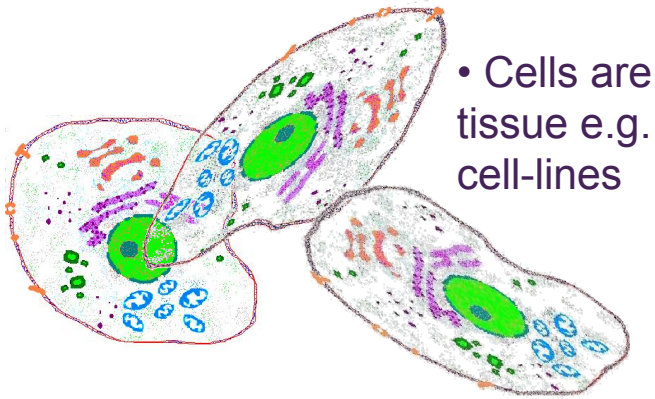
Current issues facing the pharmaceutical industry

- All pharmaceutical companies face high attrition of compounds through the discovery and development process
- Two key issues that face project progression are
  - Safety and toxicity
  - Efficacy in disease process
- Need to know more about the mechanism of action and toxicity of our compounds at an earlier stage in the discovery process
- More information enables front-loading of risk, early go/no-go decisions and improvements in toxicological attrition

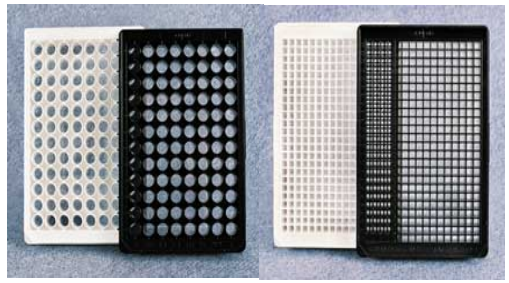
High-content biological assays

- Attempt to use *in vitro* cell models to mimic the complexity of an *in vivo* situation
- Advanced imaging techniques used to generate large, complex datasets describing the response of a population of cells to a compound
- Aim is to build predictive models or 'fingerprints' from the multiparametric assay data for well-characterised compounds that elicit known responses
- Fingerprints applied to new drugs to predict biological mechanism of action and its toxicity

Cell culture

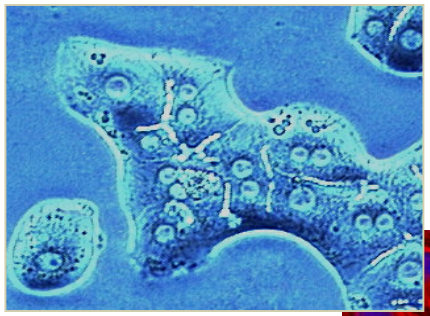


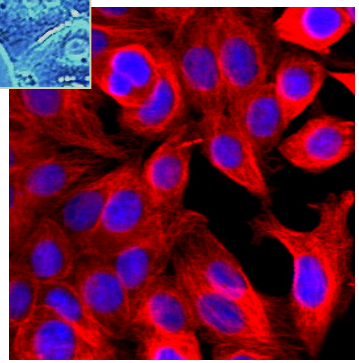
• Cells are extracted from some source tissue e.g. rat hepatocytes, tumour derived cell-lines



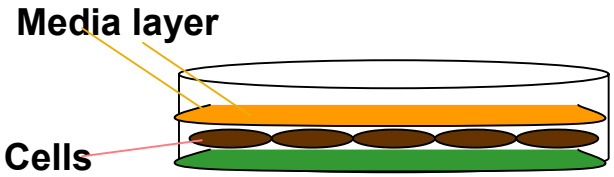
• Cells are plated into multi-well plates, typically hundred or thousands of cells per well

• Each well is like test tube where we can test a single prototype drug

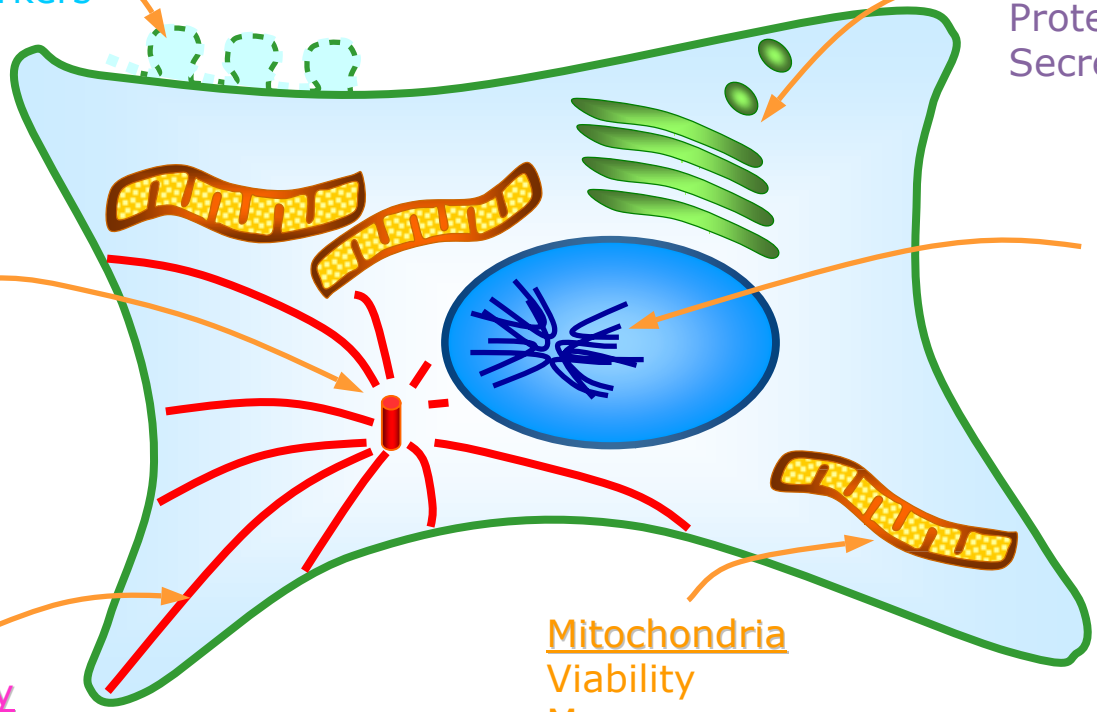




• Cells grown in the well can be labelled and imaged



HCB cellular profiling



Apoptosis
Membrane markers
Blebbing
Necrosis

ER/Golgi
Protein trafficking
Secretion

Cytoskeleton
Tubulin
Actin
Fibre content
Length
Mitotic arrest

Nucleus
DNA content
Size
Shape
Cell division
Fragmentation
Micronuclei

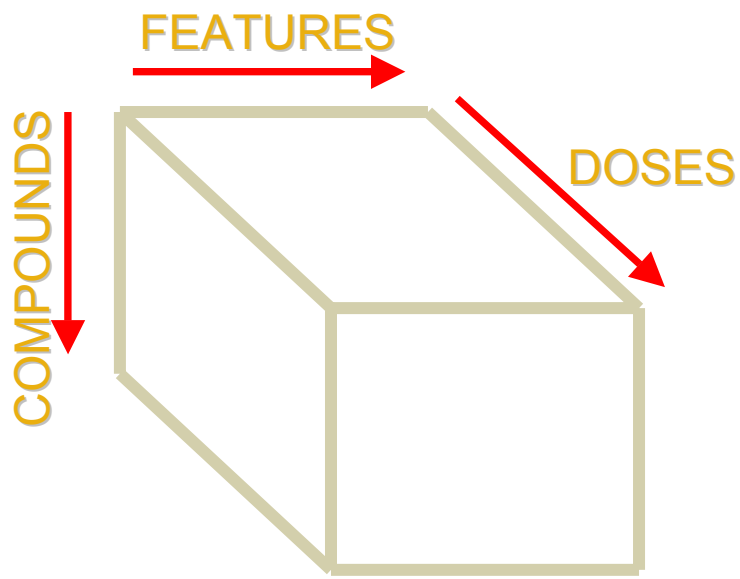
Cell Morphology
Count
Area
Form
Roundness
Length/Breadth
Perimeter

Mitochondria
Viability
Mass
Activity
Cellular distibution
Pre-Apoptotic indicators

General imaging indicators

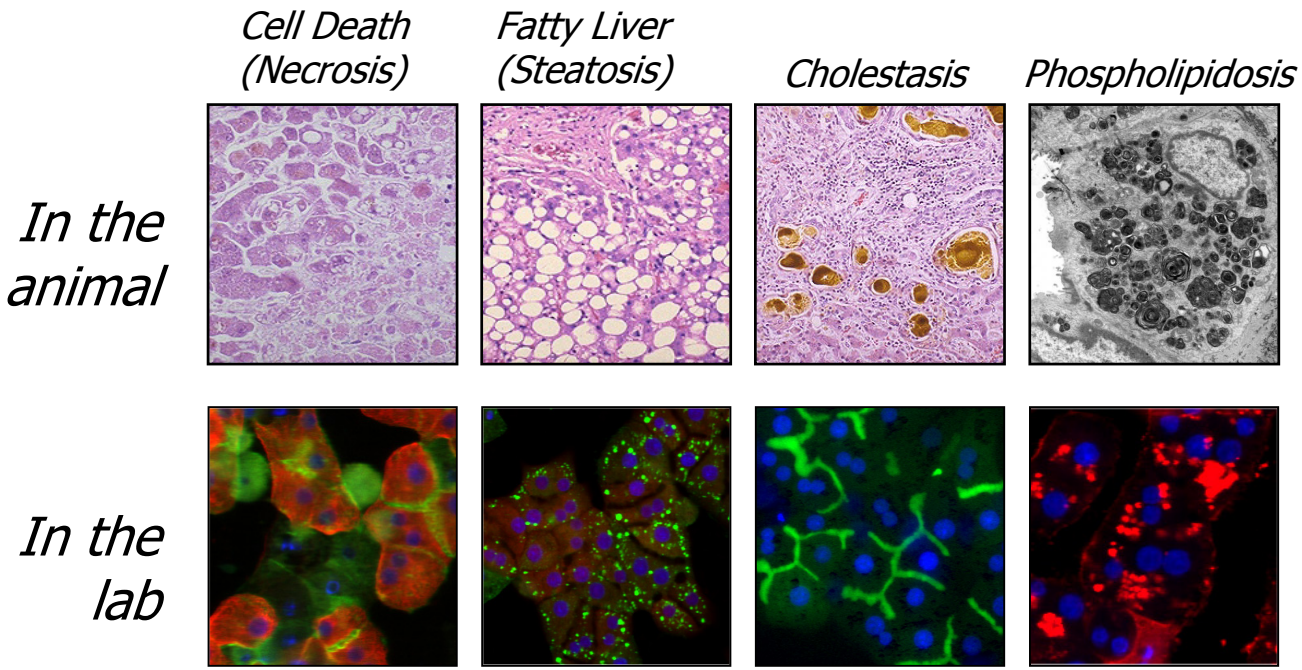
Statistical challenges

- Information captured for each feature is a dynamic response to the compound over an 8-point dose-range
- Datasets possess three-dimensional cube-like structure
- Traditional multivariate approaches are difficult to apply to this type of data directly



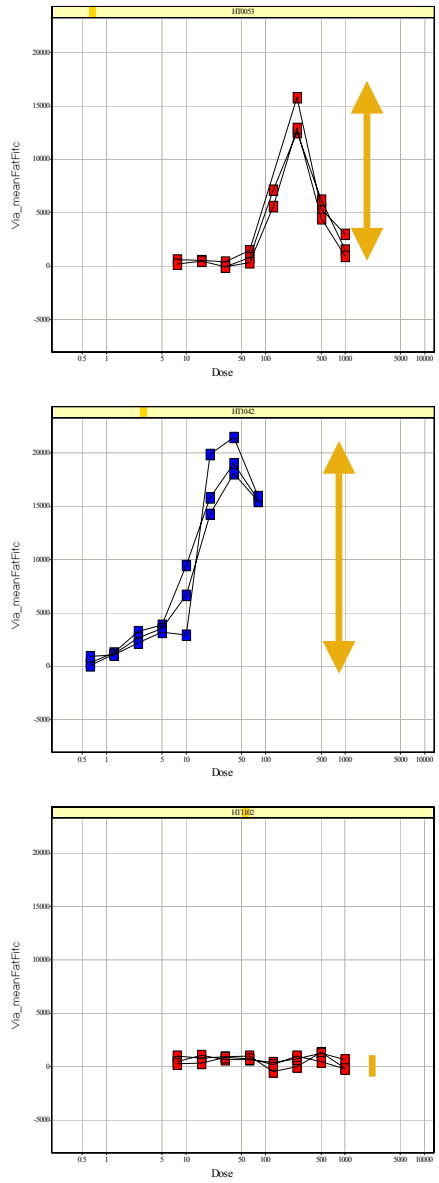
Case study 1: Predicting liver toxicity *in vivo*

- Drug-induced liver toxicity is one of the most common causes of drug non-approval
- Early *in vitro* identification of compounds with hepatotoxic risk would allow their de-selection early in the drug development process



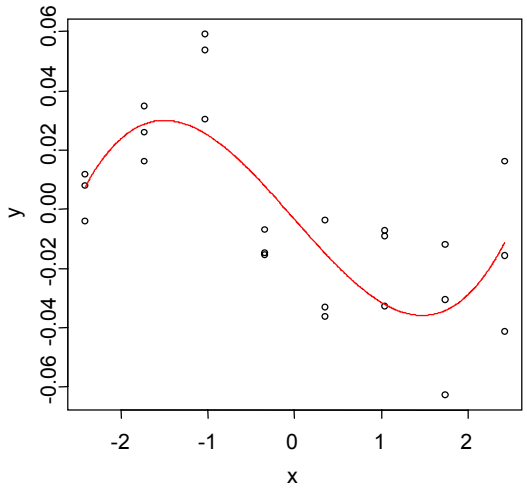
Predicting steatosis - data

- Primary rat hepatotcytes treated with 60 compound set at a range of doses, consisting of known steatotics and non-steatotics
- Bespoke algorithms designed to quantify differences in localisation and morphology of lipid droplets in the cells
- Generates 32 different continuous measurements per cell
- Averaged over cell population to give well-level measurements for each compound and dose combination
- Use partial least squares modelling (stepwise) with the steatotic annotation as a binary response

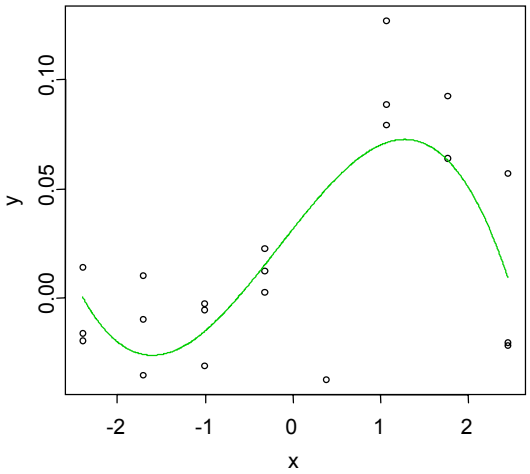


Polynomial model

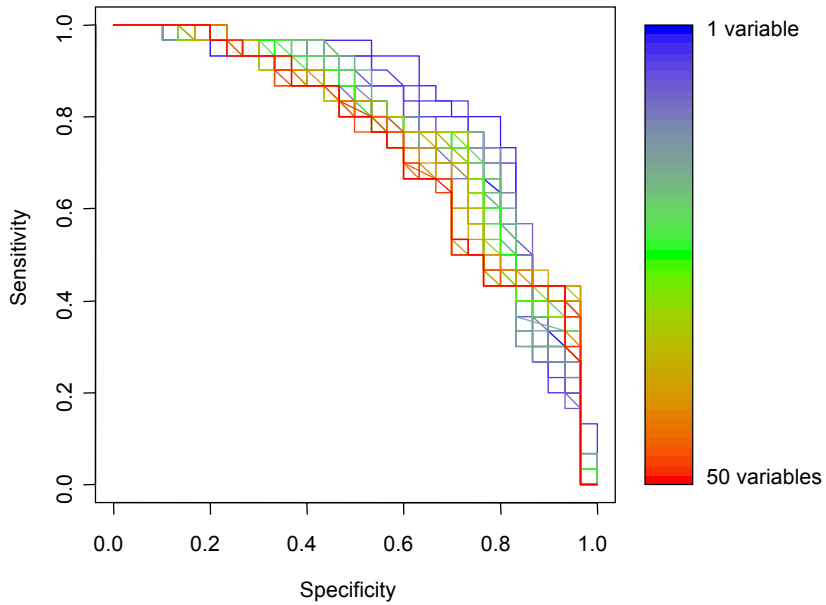
Proportion of edge fat – steatotic



Proportion of edge fat – non-steatotic

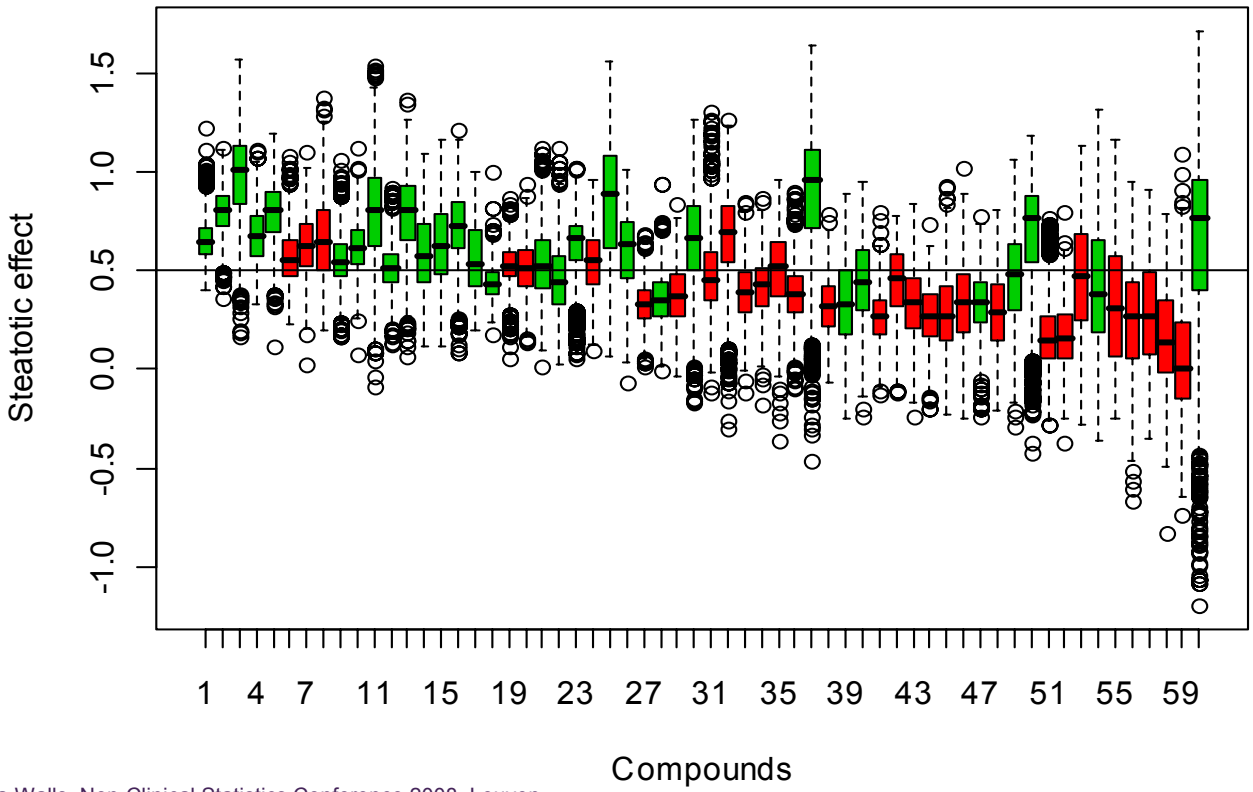


- Fit cubic polynomial to dose-response data for each feature
- *t*-statistics for each term in cubic form a new set of variables
- Only a small number of variables required to generate greatest predictivity
- After cross-validation, polynomial model is approximately 10% better than range model



Advantages of model

- Based on predictive scores, compounds can be ranked in order of steatotic effect
- Bootstrapping, incorporating random *x*-resampling, used to generate 95% confidence intervals for the predicted score
- High confidence, high steatotic effect compounds can be de-selected

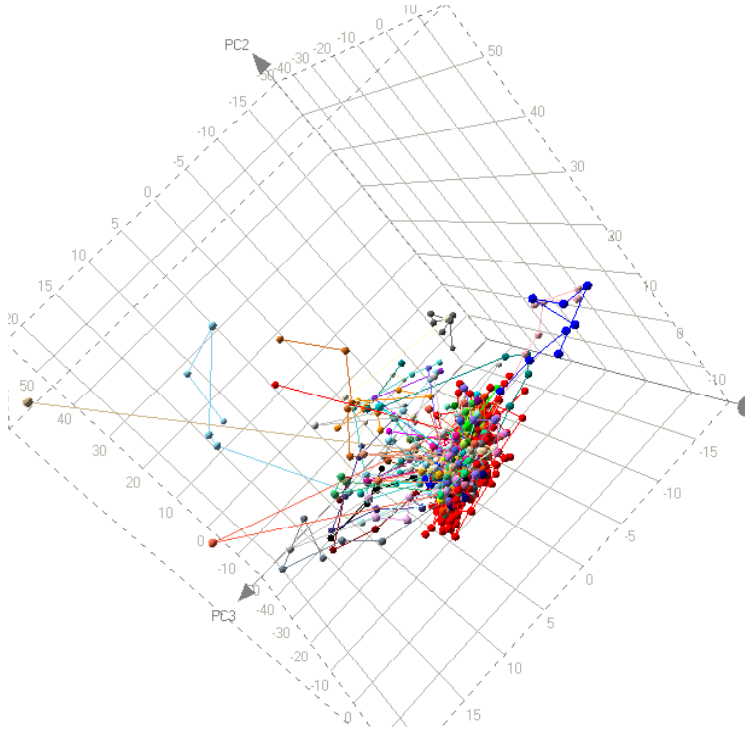


Case study 2: Identifying distinct modes of compound action

- Morphology high content assay developed specifically to examine microtubules and actin filaments as oncology targets –
- Describes how drugs influence entire complex cellular phenotype (i.e. multiple targets)
- 102 compounds screened through the morphology assay
- Primary aims are
  - Identify which compounds are active in the assay i.e. which are 'hits'?
  - Differentiate compound hits that have distinct morphological effects
  - Cluster hits together that have similar effects
- 138 features for each compound, tested over 8 doses
- 310 control wells

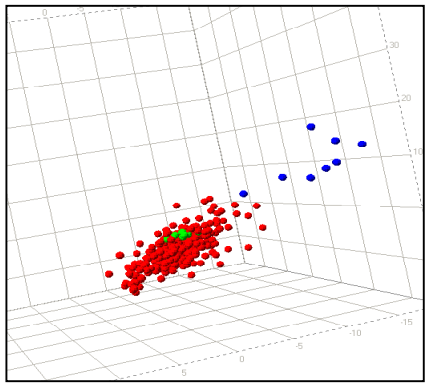
Principal components analysis

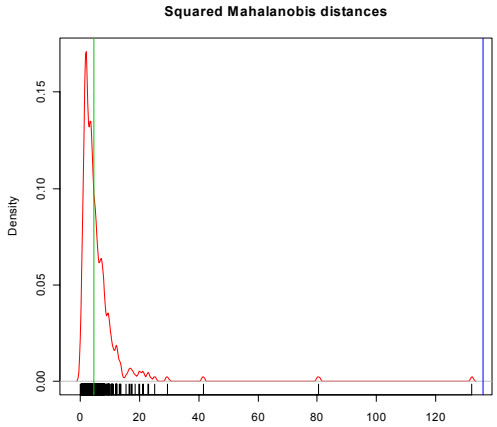
- PCA used in an attempt to reduce dimension of dataset, yielding 6 principal components which explain close to 80% of variation
- Mahalanobis distance is powerful means of determining how similar an unknown sample is to a known one
- Differs from Euclidean distance in that it takes into account the covariance between variables
- The Mahalanobis distance from a group of values with mean $\mu=(\mu_1, \mu_2, \ldots, \mu_p)^T$ and covariance matrix Σ for multivariate vector $x=(x_1, x_2, \ldots, x_p)^T$ is defined as

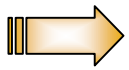


$$D_M(x) = \sqrt{(x-\mu)\Sigma^{-1}(x-\mu)}$$

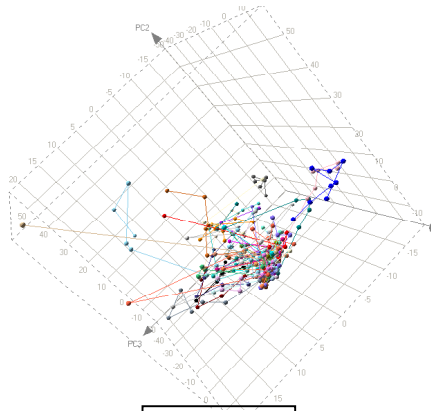
Using the Mahalanobis distance



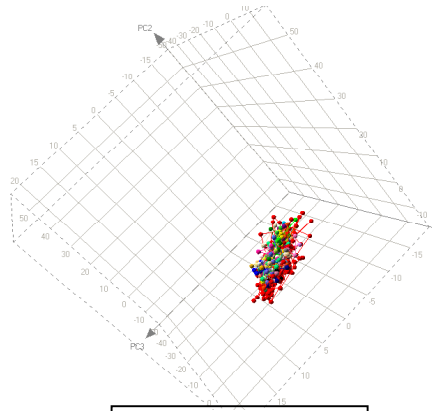




- Working on the PCA scores on the 6 principal components, the covariance matrix of the control cloud was calculated
- For each compound at every dose, the squared Mahalanobis distance to the centre of mass was calculated and compared to a chi-squared distribution with 6 degrees of freedom at some pre-chosen significance level, α.
- An adjustment was made to control the false discovery rate
- A compound with a significant result at ***at least*** one of the doses along its range was deemed to be an 'active hit'.

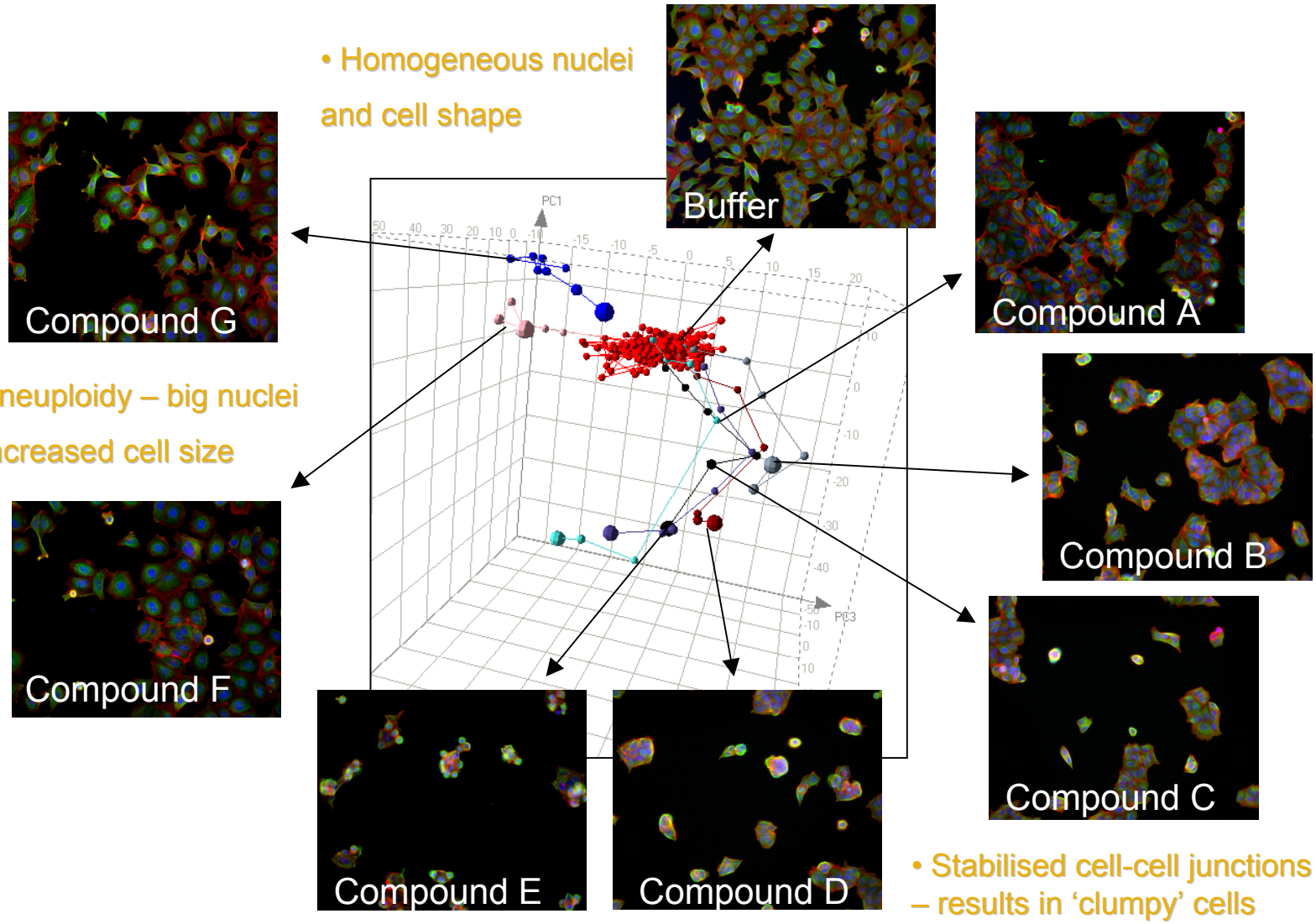


Hits



Non-hits

Distinguishing distinct phenotypes



Acknowledgements

- Discovery Statistics
  - Chris Harbron

- Advanced Science and Technology Laboratory
  - Ed Ainscow
  - Neil Carragher
  - Andy Hargreaves
  - Mike Sullivan
  - Helen Garside
  - James Pilling
  - Lisa Rice
  - Tom Houslay
  - Peter Caie
  - Alex Ingleston-Orme