Multilayered Analysis of HCS Data:

An Integrated Approach

Amit Bahl, Wendy Bailey, Bonnie Howell, Ed Keough, Irene Pak, Ansu Bagchi

Merck & Co., Inc.

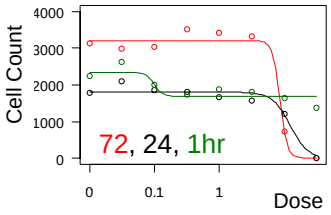
January 13, 2010

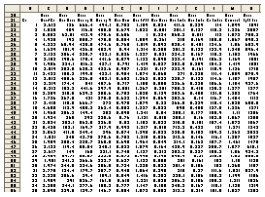
Outline

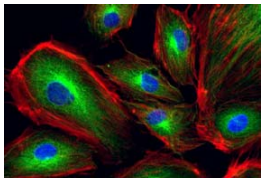
- Data Analysis
  - Data from multi-factorial experiments
  - Multivariable data – cannot ignore correlations
  - Univariate vs. Multivariate strategies
- Assay Quality
  - What goes in...
- Data / Information Management
  - Storage / Retrieval / Navigation
- Putting it all together

Data Analysis

HCS Workflow







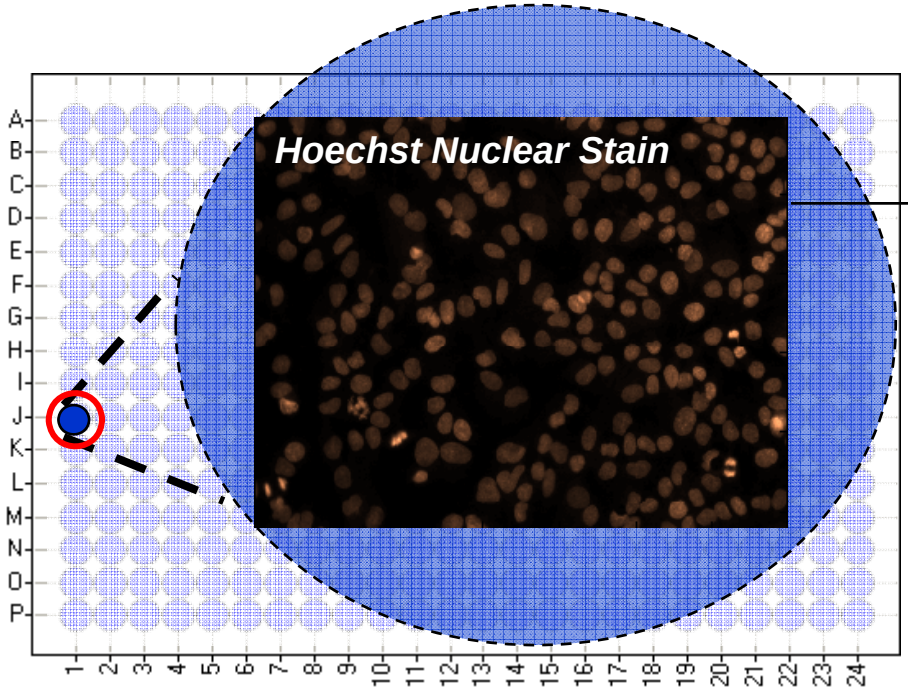
Scientific Insight!

Statistical Analysis / Data Mining

Image Analysis

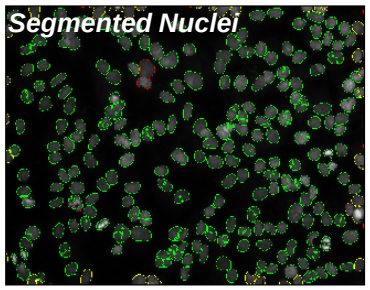
HCScreen

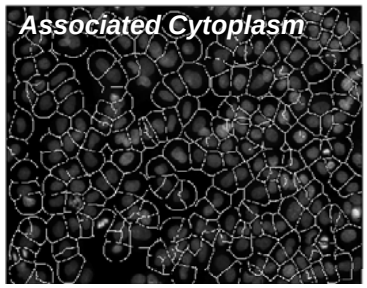
High Content Screens



Assay and Image Acquisition

Automated Image Analysis





cell #	1	2 ...
Cell area	40.1	42.5
Cell perimeter	35.3	38.9
Cell aspect ratio	1.56	2.01
Actin content	4510	4939
Actin texture	16.8	17.2
Cell solidity	0.99	1.03
Cell extent	0.68	0.35
Nuclear area	10.9	11.1
Nuclear perimeter	1.0	1.1
Nuclear aspect ratio	1.56	2.01

Cell-level Data

Systematic titration study of compounds, or other therapeutic agents, across multiple readouts, cell lines, or several time points

Data Scope

Plate level	Replicate level	Well level	Field level	Cell level

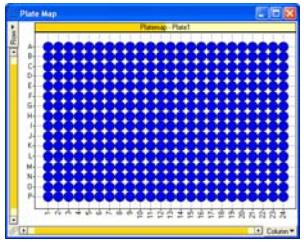
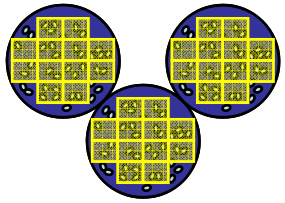
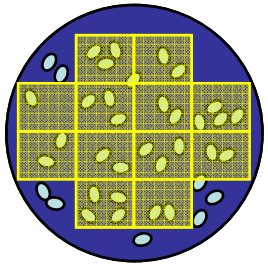
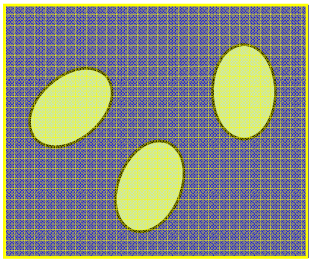
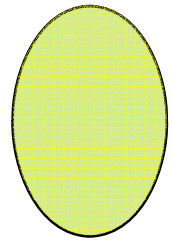

				

Plate setup ▪ Treatment type ▪ Treatment ID ▪ Concentration ▪ Reagents	Replicate summary (basic statistics on all the cells pooled together from one or more experimentally-identical wells) for feature of interest generated by the user.	Well summary (basic statistics on all the cells imaged in this well) for each feature measured by the instrument or generated by the user.	Field summary (basic statistics on the cells imaged in a single field) for each feature measured by the instrument or generated by the user.	Numerical data (a.k.a. “descriptors”) for all features measured for a single cell by the instrument.

Extracted from Platemap File

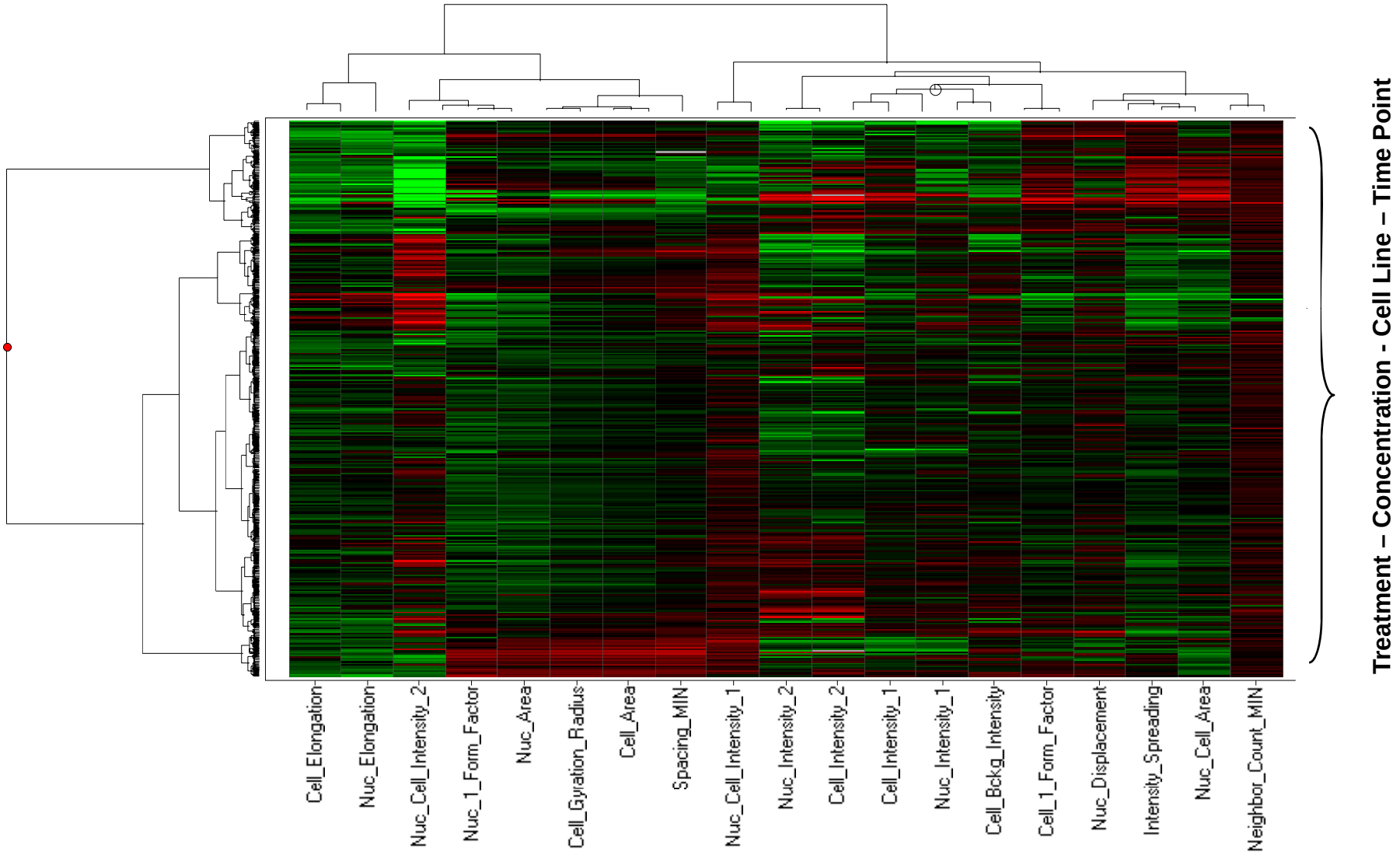
Extracted or derived from Instrument Data

Scientific Goals

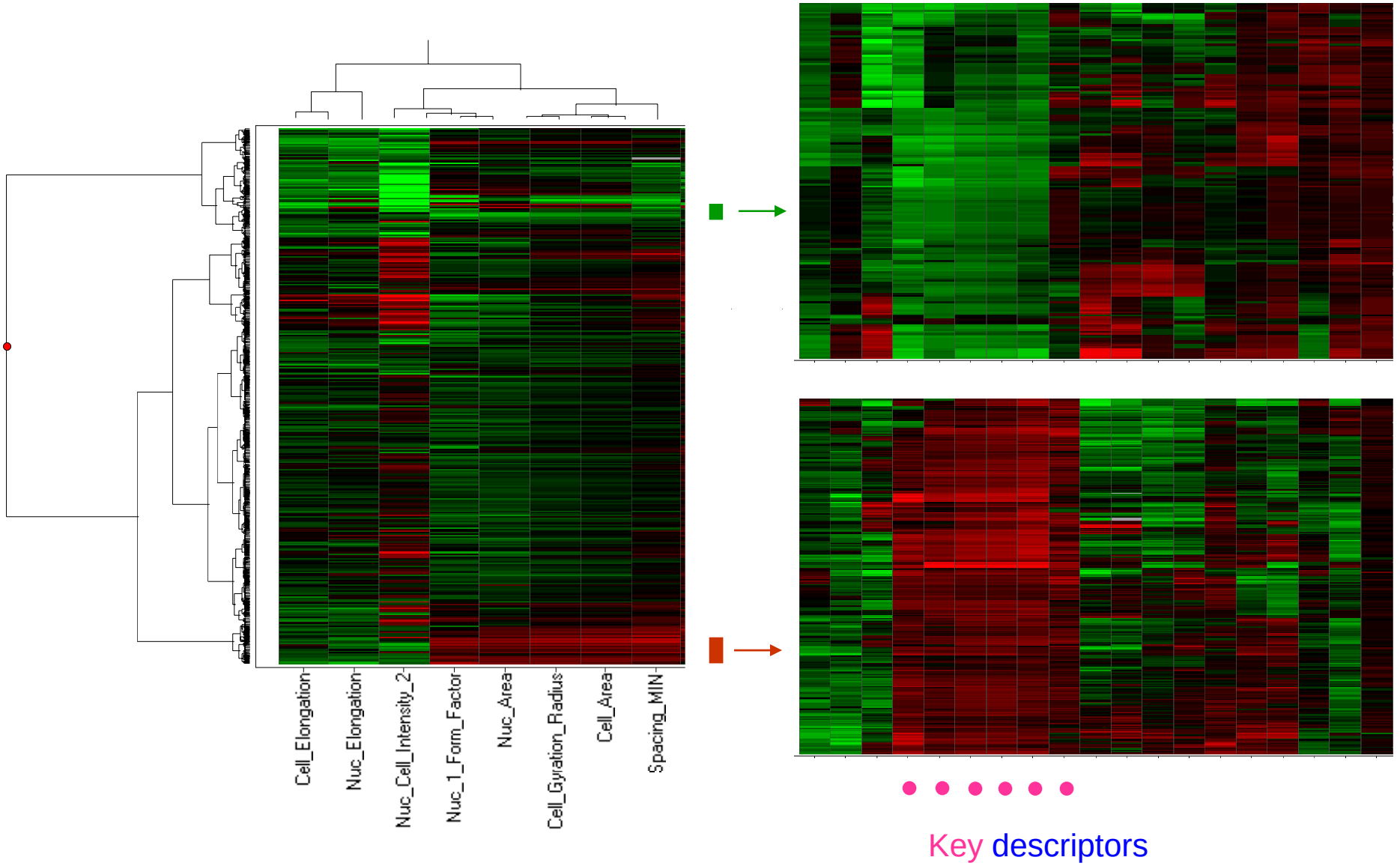
What questions are we interested in asking?

- Classification of therapeutic agents?
  - Into multiple pre-determined groups
- Differentiation between therapeutic agents?
  - Identify key phenotypic characteristics, and
  - The concentration of the test agent, at which the phenotype emerges
- To find items that are similar/dissimilar to positive/negative controls?
  - Controls with known bio-chemical characteristics (good/bad)
- Characterize MOA?

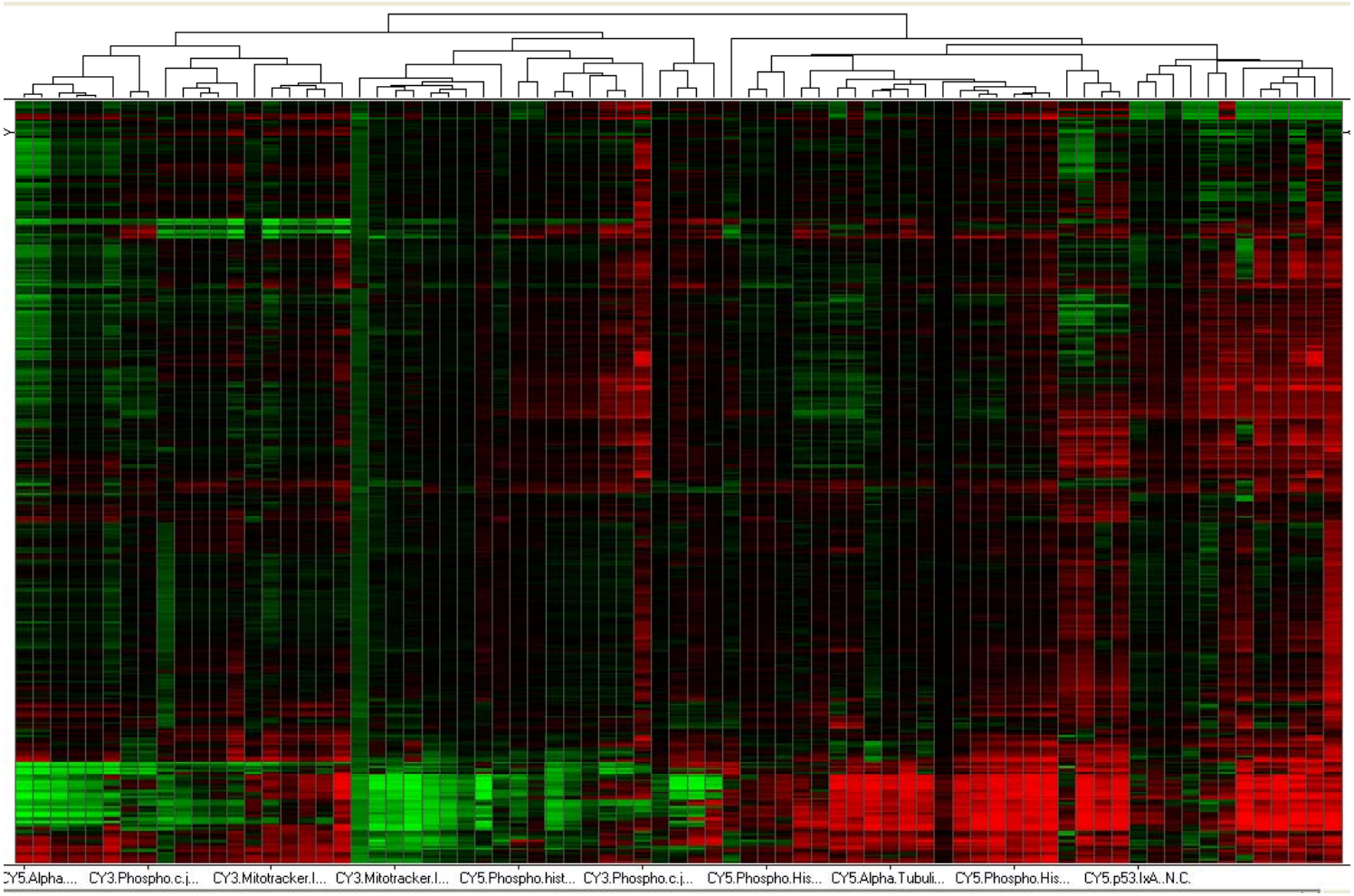
Univariate: Clustering



Univariate: Differentiation

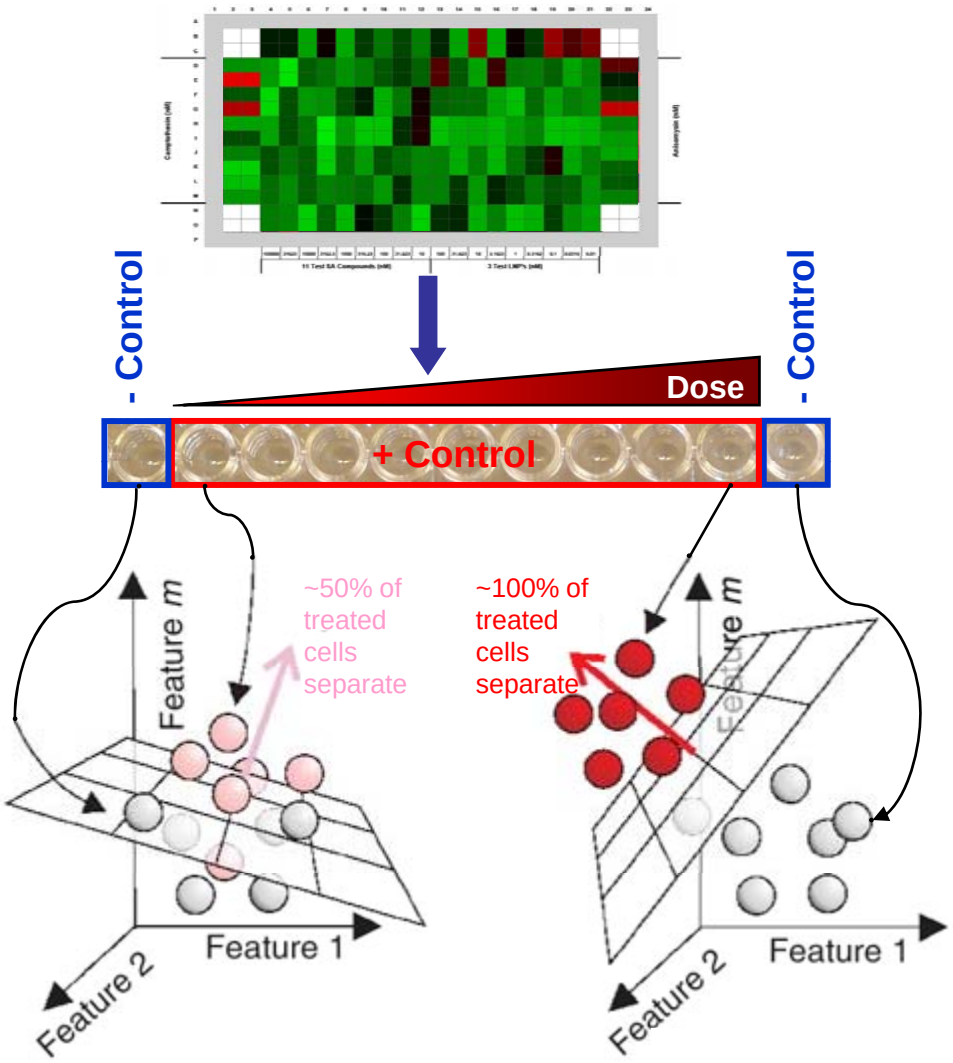


But: Many features are correlated



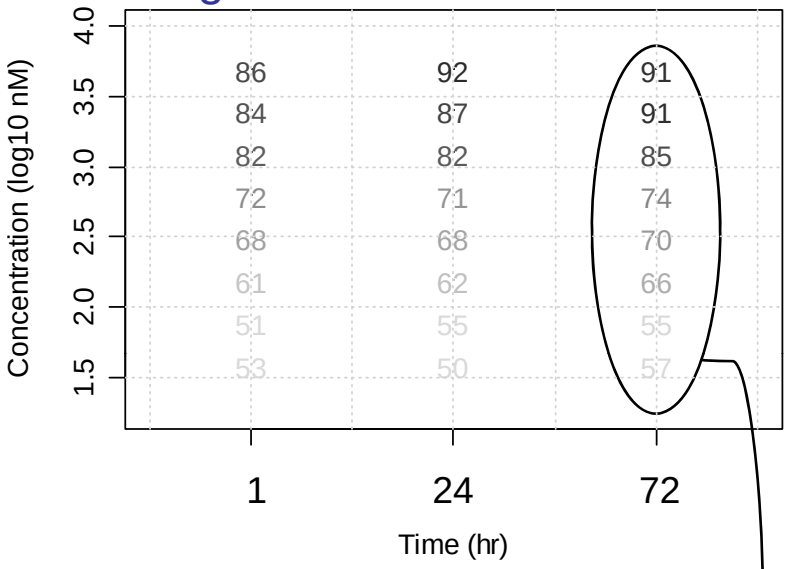
Consider multivariate methods: Support Vector Machine/Random Forest

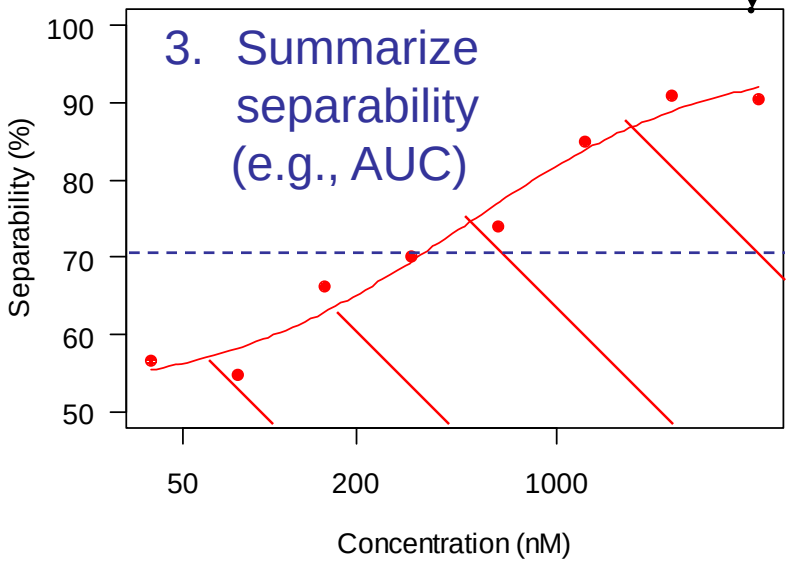
Multivariate: Evaluate Separability



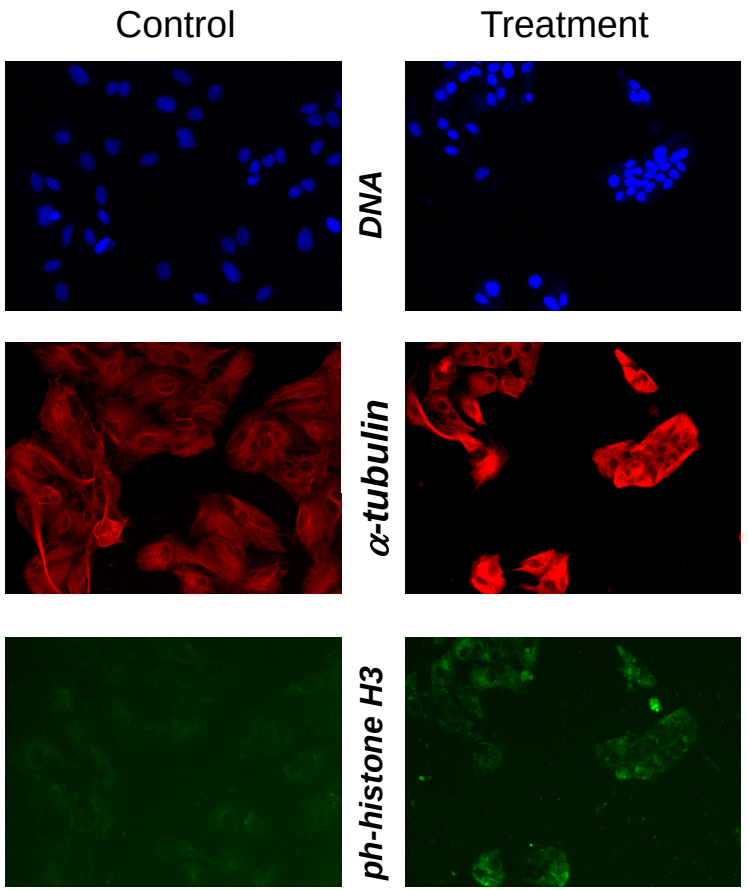
1. For each sample well, a SVM is constructed relative to control wells.

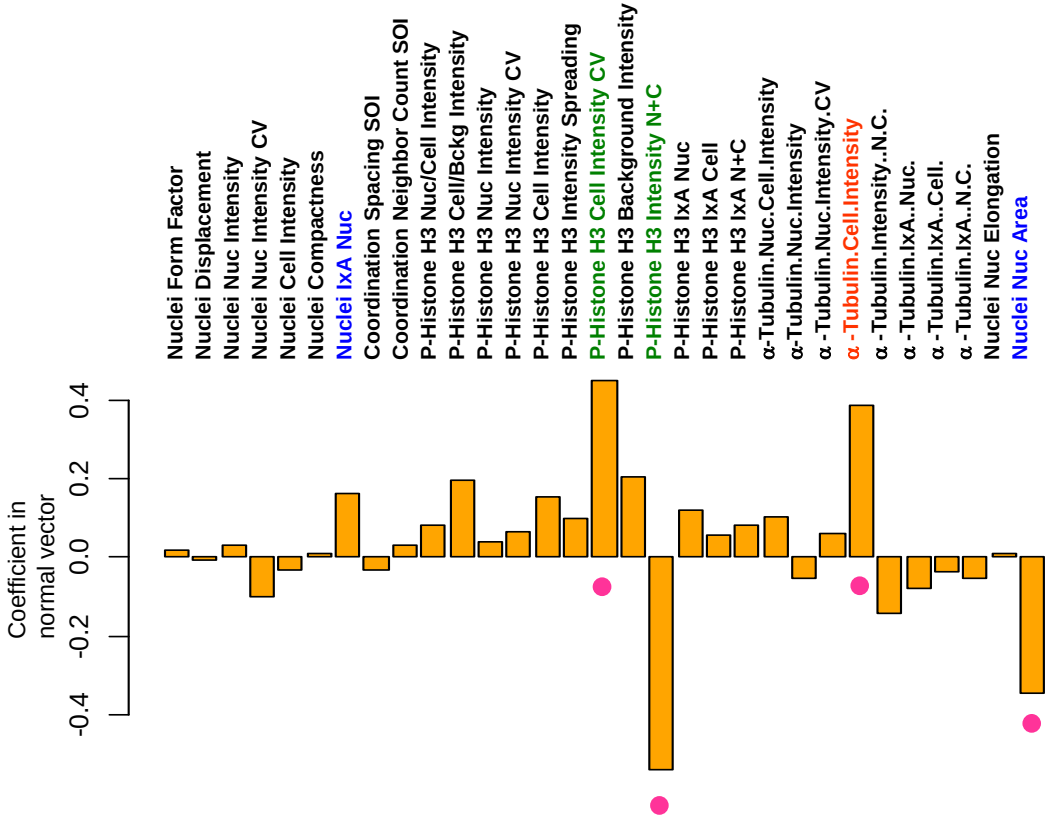
2. Resulting *separability* is a surrogate for differential behavior





Extract meaningful distinguishing parameters





Images captured on INCell 1000 (GE Healthcare) - 20x objective

Univariate vs. Multivariate

Well Summary (KS)

Pros	***Cons***
Very fast, space/memory efficient	**No subpopulation analysis**
Focuses on large-scale population shifts	**Single summary for interleaving CDFs may be inadequate**
Can combine plates/probe-sets	**One variable at a time**
Existing capabilities in SF	**Multiple measures of separability**

Cellular level (SVM/RF)

Pros	***Cons***
Range of ML methods	**Slower, memory intensive**
Multivariate	**Interpretability of results**
Subpopulation analysis	**Requires careful integration**
Single continuous measure of separability	**Relatively less familiar to biologists**

An optimal approach is often a combination of both

Assay Quality

Z'-factor

Integration of Multiple Readouts into the Z′ Factor for Assay Quality Assessment

ANNE KÜMMEL, HANSPETER GUBLER, PATRICIA GEHIN, MARTIN BEIBEL, DANIELA GABRIEL, and CHRISTIAN N. PARKER

Methods that monitor the quality of a biological assay (i.e., its ability to discriminate between positive and negative controls) are essential for the development of robust assays. In screening, the most commonly used parameter for monitoring assay quality is the Z′ factor, which is based on 1 selected readout. However, biological assays are able to monitor multiple readouts. For example, novel multiparametric screening technologies such as high-content screening provide information-rich data sets with multiple readouts on a compound's effect. Still, assay quality is commonly assessed by the Z′ factor based on a single selected readout. This report suggests an extension of the Z′ factor, which integrates multiple readouts for assay

Journal of Biomolecular Screening, November 25, 2009

- Z'-factors are standard measures of assay quality in high throughput screens.
  - Reduce "amplitude and variability" of each assay measurement to a single parameter
- HCS provides multiple readouts
  - Not optimal to calculate Z'-factors seperately
- Need for multivariate data analyses implies advanced methods to compute Z'.

Z'-factors for HCS

Probeset-specific Z'-factors

$$Z' = 1 - 3\left((\sigma^{+} + \sigma^{-}) \;/\; (|\mu^{+} + \mu^{-}|) \right)$$

where σ^{+}, μ^{+} (or, σ^{-}, μ^{-}) represent the standard deviation and mean, respectively, of the calculated biological measures (e.g., AUCs) relative to the positive (or, negative) control

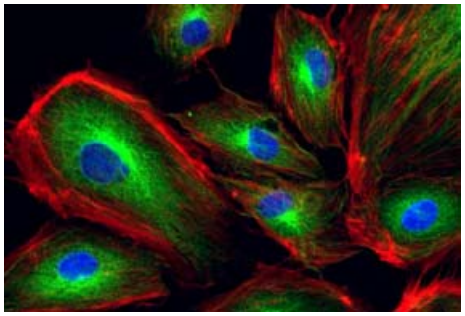
Time (hr)	MMP	Cell Cycle Arrest	Cyto-skeletal Integrity	Stress Kinase Pathway	Oxidative Stress	Nuclear Integrity	DNA Damage
1	0.89	0.84	0.88	0.92	0.86	0.83	0.92
24	0.90	0.90	0.88	0.84	0.88	0.92	0.93
72	0.89	0.94	0.84	0.91	0.94	0.92	0.96

- Z' measures using multivariate techniques suggest very high assay quality

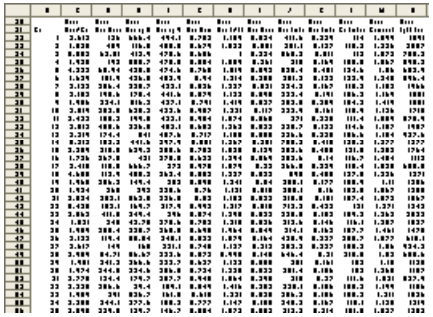
Data Management *and* Navigation/Analysis Tool

Navigation through Multilayered Data

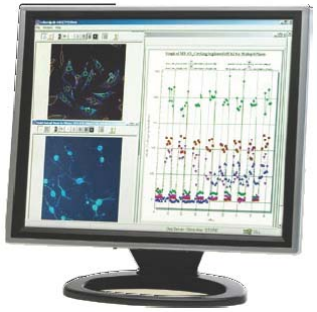
Set up and run experiment



Capture images of cells



Numeric descriptors



Visualization and Analysis Tool

HCS Repository

Images

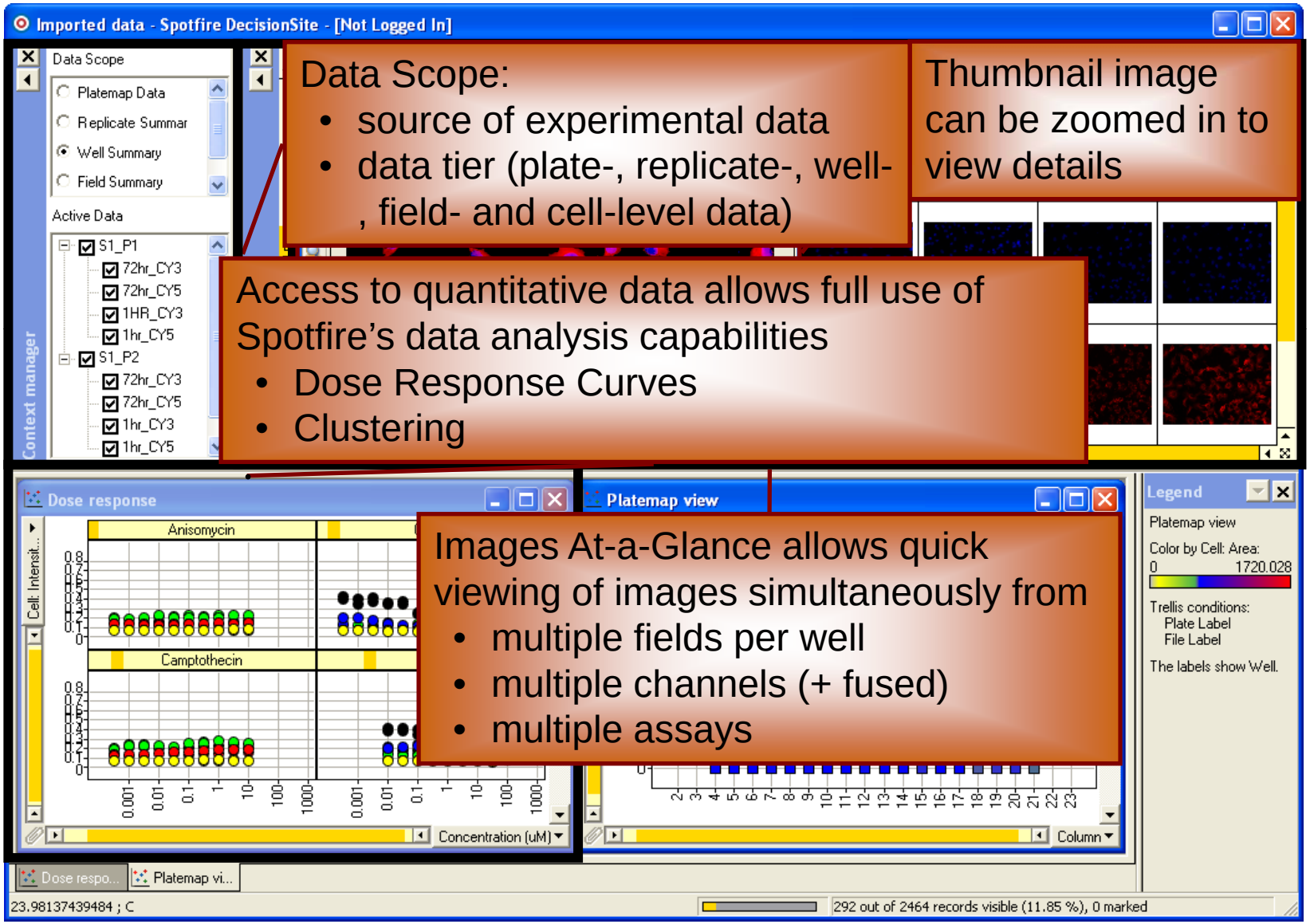
Numerical Data

Analysis Results

Tool Objectives

- To facilitate appropriate analyses of HCS experiments on a *single* platform:
  - navigation between raw data, processed data, and images
  - quality control of experimental data
  - visualization and analysis of data
  - Access to data mining tools
  - data export and sharing of analysis results

HCS Tool At-a-Glance



Desired Tool Characteristics

What you do *not* want

- Multiple applications used: Excel, Prism, Spotfire, IN CELL Developer
- Image and numerical data can only be referenced by well label, with no linkage to experimental details– there is no associated “context”
- Well image can only be brought up one field at a time, one assay at a time
- Additional effort required to generate fused image
- Analytics performed only at the well-summary level [no capability to examine data beyond well summary]
- User can only examine data from one assay at a time
- Clunky, laborious, and error-prone

HCS Tool that we would love to have

- Single platform (e.g., in Spotfire) to bring together numerical data and images
- Annotations are automatically applied to the data (via platemap) – images and numerical data can be examined “in context”
- Well images from multiple fields and assays can be viewed simultaneously
- Fused images are generated automatically
- Fast context switching between different data tier allowed: platemap, replicate, well-summary, field-summary, [cell data]
- User can examine multiple assays at a time (batch import allowed)
- Simple, easy, and error-proof

Summary

- Seek balance between
  - Data formats: standardized vs. customized
  - Mining tools: device specific vs. robust
  - Analytical strategy: univariate vs. multivariate
  - Efficiency: high vs. low throughput (data volume)
  - Required level of rigor: high vs. low resolution
  - Scientific Goals: signature identification or characterization vs. classification
- Ease of navigation is highly desired
  - Connect the dots ...
  - Through complex, layered, multi-type data

Acknowledgements

Amy Aslamkhan, **Amit Bahl**, **Wendy Bailey**, Randy Crawford, **Bonnie Howell**, **Ed Keough**, Adam Kozakov, Andrew Kraynak, Constantine Kreatsoulas, Steve Krotzer, Matt Kuhls, Kevin Leach, Liping Liu, Jim Monroe, **Irene Pak**, Francesca Santini, Joe Sina, Tom Skopek, Richard Storer, Dianne Umbenhauer, Jim Xu

Laura Sepp-Lorenzino **Frank Sistare** **Jeff Saltzman**