Core Life Analytics

ITERATIVE DATA MINING FOR HIGH CONTENT TOXICOLOGY WITH STRATOMINER™

APPLICATION NOTE

Rapid iterative data mining from cell painting data for toxicology.

This application note demonstrates how StratoMineR™ allows rapid iterative analysis of complex HCS data.

Introduction

One of the biggest nightmares of a drug development program leader is the emergence of unexpected toxicities in human trials, or worse, after a drug reaches the market. Many high profile drug withdrawals were due to the limitations of traditional toxicology assays, based on cytotoxic readouts, that were not sufficiently predictive of certain forms of toxicity[1].

This has led to a focus on measuring 'pre-toxic' events that can lead to an assay platform that is more concordant with what is seen in human trials. High content screening, (HCS), methods have been successfully used for these assay platforms, as they allow for the multiplexing of multiple readouts in an *in vitro* format that is amenable to laboratory automation. Such assays have been described for all the commonly investigated forms of toxicity, including hepatotoxicity[2], genotoxicity[3], and nephrotoxicity[4].

HCS yields multiparametric data that can provide critical insights into mechanisms of toxicity. The outputs, however, are large and complex datasets, that are often analytically overwhelming. This makes HCS assays challenging for biologists to develop and validate with little coding or data science experience. Here we demonstrate how StratoMineR™, our online data analytics platform[5], when combined with Iterative Data Mining, can be used to rapidly gain greater insights into the performance of a HCS toxicology platform.

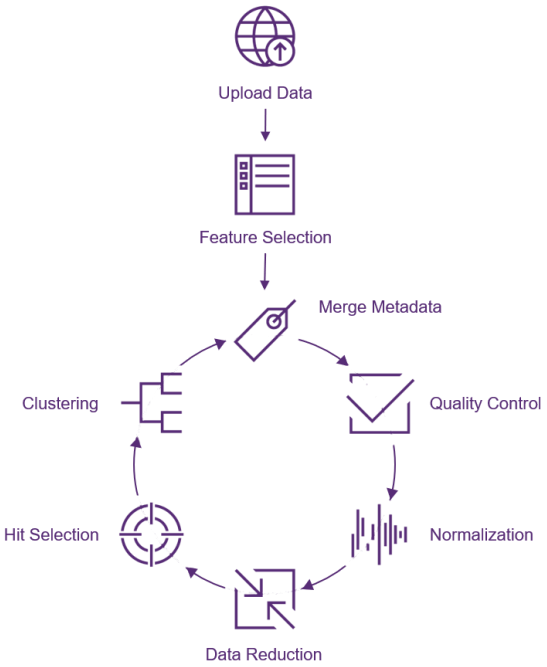

Figure 1: Overview of the StratoMineR™ workflow and an iterative data mining loop.

StratoMineR™, Iterative Data Mining and a Cell Painting Dataset

StratoMineR™ is an intuitive web-based data analytics platform that allows users to process and mine large numeric data sets from HCS experiments. The workflow is highly flexible and one of its unique functions allows users to engage an iterative data mining process (**Figure 1**). An initial analysis of a data set can be used to generate knowledge that can then be used as metadata. This metadata can be subsequently merged with the original data set and used in additional rounds of analyses.

We demonstrate this approach in the analysis of a Cell Painting data set, the LINCS data set that has recently been made publicly available[6]. Here, we use the Iterative Data Mining method to evaluate the utility of Cell Painting[7] for the profiling of 'pre-cytotoxic' phenotypes of known hepatotoxic compounds.

Iteration 1: Eliminating low cell counts

Our interest is in 'pre-cytotoxic' phenotypes so we first ran an analysis to identify wells with low cell count. We used a simple Hit Picking method based on the Cell Count feature to determine that there were 1029 cell count hits at p-value < 0.05. Wells at the p <0.05 had a medican cell count of 2210. This decreased to 761 wells with a median cell count of 1691 cells at the p-value < 0.0001 cut-off. The median cell count of Negative control wells was 2488.

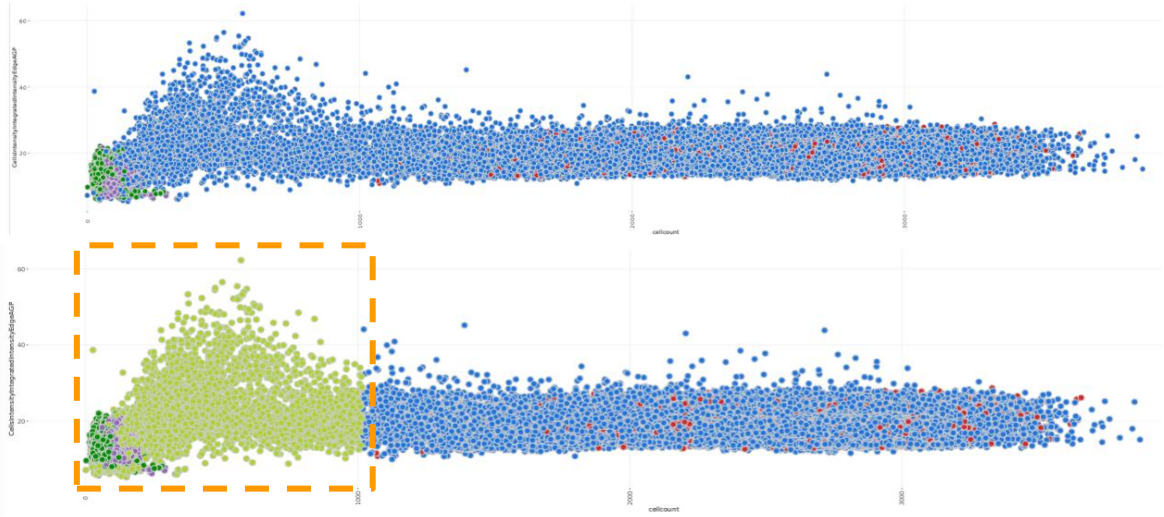

Figure 2: Using the VDM QC interactive data visualization module we defined one class of cytotoxic wells (orange dashed box).

Iteration 2: Phenotyping DILI Drugs

Our partner Cytochroma assisted in the annotation of all the samples in the LINCS data set as to whether they had been previously shown to cause Drug Induced Liver Injury (DILI). This led to the annotation of 202 compounds which were split into DILI Low and DILI High reagent classes. This annotation was added to the experiment using the Merge Metadata module in StratoMineR™. We could then label all the DILI c[illegible]

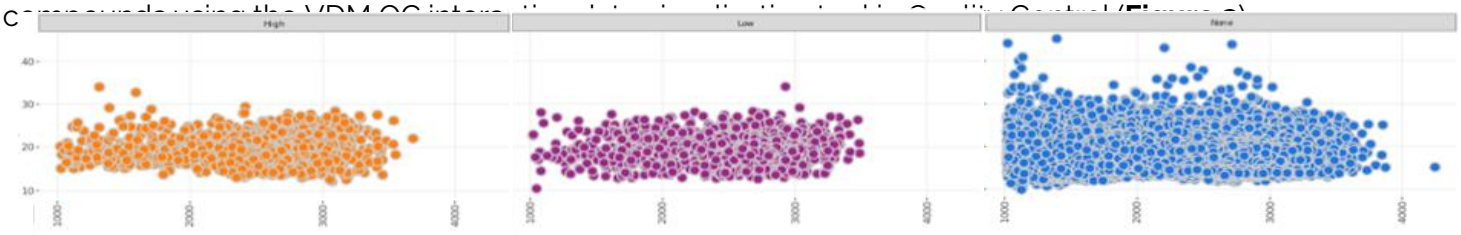

Figure 3: In StratoMineR™ there is a functionality in VDM QC that allows for the tiling of data based on a metadata feature. We used the merged metadata module that included DILI annotation to label samples as DILI High or DILI Low.

To determine whether the DILI compounds showed 'pre-cytotoxic' phenotypes in Cell Painting, we determined how many still gave significant phenotypes in the absence of wells with low cell counts. Without the removal of low cell count wells 144/202 gave a significant phenotype (**Table 1**), and for many, a phenotypic dose response could be shown (**Figure 4**). The number dropped off to 134/202, 111/202 and 77/202 when wells with cell counts < 500, < 1000 and < 1500 were eliminated, respectively.

Table 1

DILI	All	p < 0.05	P < 0.05 Remove Low Count 500	P < 0.05 Remove Low Count 1000	P < 0.05 Remove Low Count 1500
High	101	84	76	66	43
Low	101	60	58	45	34
Total	202	144	134	111	77

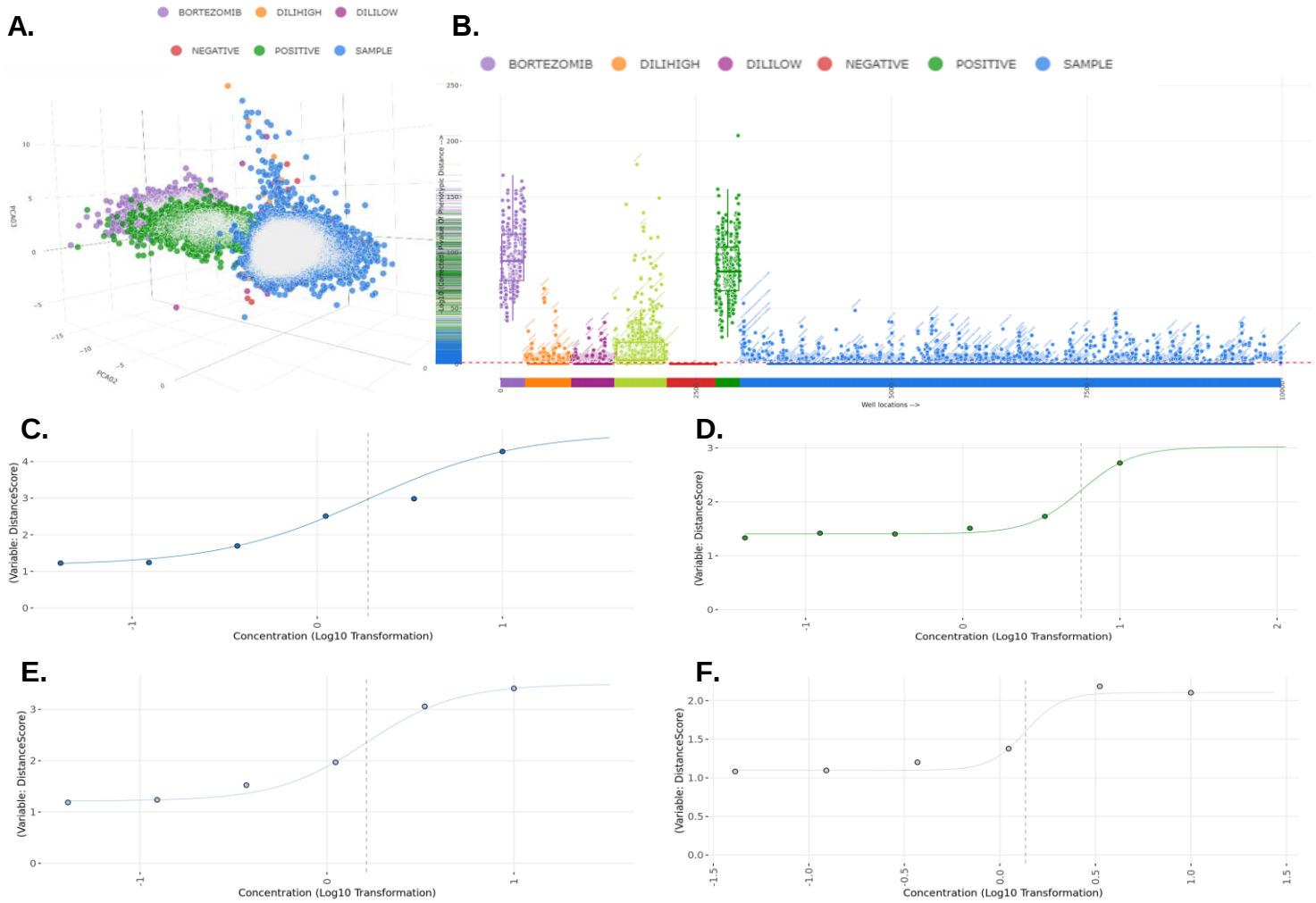

Figure 4: Data after the first iteration can be further visualized in 3D space with respect to the first three components (**A**). We can quickly identify distinct clusters based on phenotypic signatures. Distance scores from all wells to the negative controls re-calculated and in **B** phenotypic outliers with a p-value < 0.05 can be identified. Distance scores can be used to generate dose response curves for all the hit compounds. Dose response curves of a few hit compounds are shown here: ciglitazone (DILI low; **C**), aripiprazole (DILI low; **D**), dronedarone (DILI high; **E**), and dantrolene (DILI high; **E**).

Iteration 3; Investigating the Diversity of DILI Phenotypes

In order to get further insights into the diversity of Cell Painting phenotypes generated by DILI compounds, we were able to use the Clustering function of StratoMineR™ to carry our hierarchical clustering of the outliers that were generated in the 1029 cell count cut-off scenario. The clustering was carried out using the nine principal component scores that contributed to the distance score from the negatives (**Figure 5**).

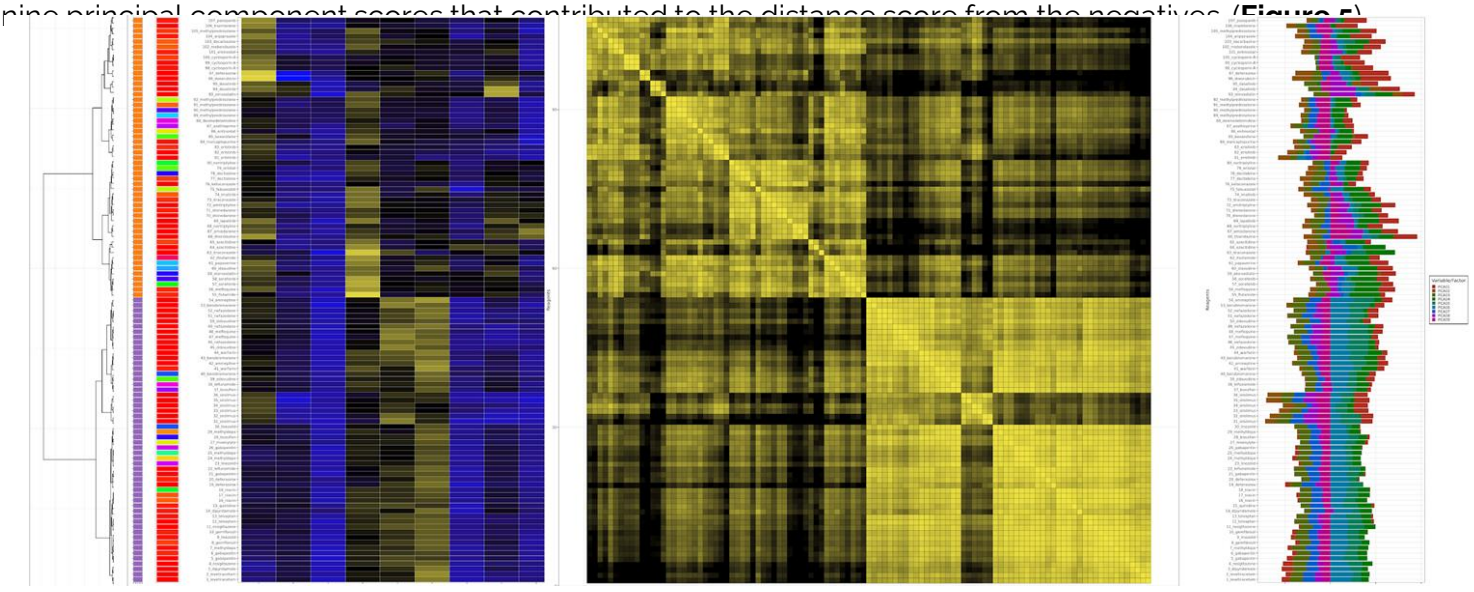

Figure 5. Hierarchical clustering analysis of hits reveals two 'super-clusters' based on nine principal component scores..

The cluster analysis identified two very obvious large clusters. Using the VDM Explorer module we were able to label these two clusters as Cluster 1 and 2. And then look at these in 3D phenotypic space using three components in a 3D scatter plot (**Figure 6**).

This process can be repeated, splitting Cluster 1 into Cluster 1.1, Cluster 1.2 and so forth, until you get the desired resolution at which you wish to investigate your hits.

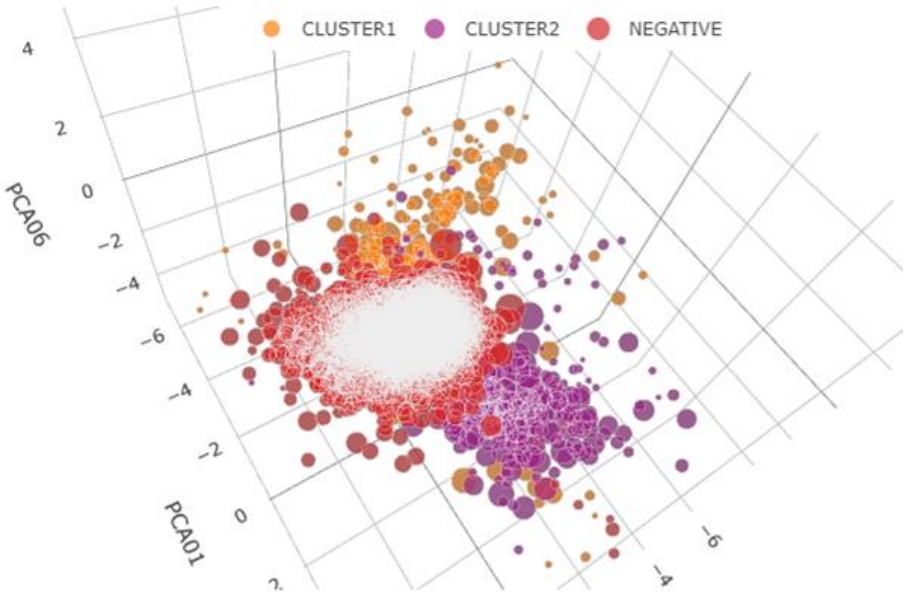

Figure 6. 3D scatter plot showing two super-clusters can be generated using the VDM Explorer module in StratoMineR™. Components 1, 6 and 4 are plotted on the axes and spot size is an indication of Cell Count. Clusters 1 and 2 are clearly distinct in this phenotypic space and it is clear that Cluster 1 wells are more likely to have a low Cell Count.

Conclusions & Discussion

Our analysis demonstrates how Iterative Data Mining in StratoMineR™ can be used to generate knowledge from very large and complex data sets. The Visual Data Mining functionality gives the user the ability to dynamically label data points using either external annotations via Merge Metadata or new knowledge that has been generated in Quality Control, Hit Selection, or Clustering.

Here, we have applied the iterative data mining to the analysis of DILI compound phenotypes in a Cell Painting data set. Our analysis shows that substantial numbers of DILI compounds give 'pre-cytotoxic' phenotypes in Cell Painting. The analysis also shows, however, that these phenotypes are quite diverse. This is not surprising as any contribution to phenotype that is specific for DILI is going to be overlapped with contributions that are related to their on-target mechanism of action.

We would further suggest that the Artificial Intelligence functionality of StratoMineR™ might be a good tool for building DILI-specific classes but this requires object-level data, which for this data set is not yet publicly available.

To find out more about StratoMineR™ for HCS image analysis, please visit **www.corelifeanalytics.com**. For more details about your data analytics needs or to schedule a live demonstration, please contact us at **sales@corelifeanalytics.com**.

References

1. Alavijeh MS et al. IDrugs. 2004 Aug; 7(8): 755-63.
2. Hussain F et al. Arch Toxicol. 2020 Aug; 94(8): 2749-2767.
3. Guo K et al. J Biomol Screen. 2014 Apr; 19(4): 538-46.
4. Su R et al. Arch Toxicol. 2016; 90(11): 2793-2808.
5. Omta W et al. Assay Drug Dev Technol. 2016; 14(8): 439-452.
6. Way GP et al. Mol Biol Cell. 2021 Feb 3:mbcE20120784.
7. Bray MA et al. Nat Protoc. 2016 Sep; 11(9): 1757-1774.