Tackling the JUMP-CP Data: Best Practices to Navigate High Content Multiparametric Data

Maria Roa Oyaga[1], Maaike Stuart[1], David Egan[1], Wienand Omta[1], Victor Wong[1]

[1]Core Life Analytics BV, 5211 DA 's-Hertogenbosch, The Netherlands

Introduction

High content image-based approaches yield rich phenotypic data that can reveal important insights into a candidate drug's mechanism of action or toxicity. Cell Painting is an example of such a phenotypic assay and is quickly gaining traction. Led by the Broad Institute, the JUMP (Joint Undertaking in Morphological Profiling) Cell Painting (CP) consortium has generated a reference dataset using ~140,000 different genetic and small molecule perturbations[1]. This unprecedented dataset was made public at the end of last year, and has the potential to be an outstanding resource for drug discovery research. Its size and complexity, however, pose a significant barrier to leveraging it. Here, we demonstrate how cloud computing can help overcome some of these challenges. Specifically, we will present a robust and iterative data analytics workflow that will allow users to mine it for information relevant to their drug discovery projects.

Methods

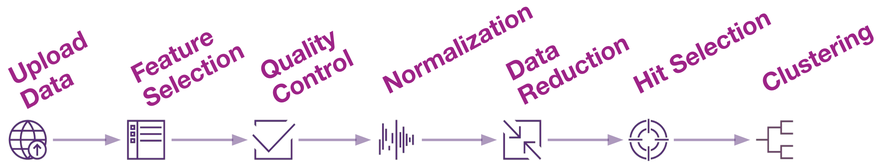

Figure 1: The StratoMineR™ workflow. StratoMineR™ is a web-based platform which guides users through a typical workflow in analysis of high content multi-parametric data.

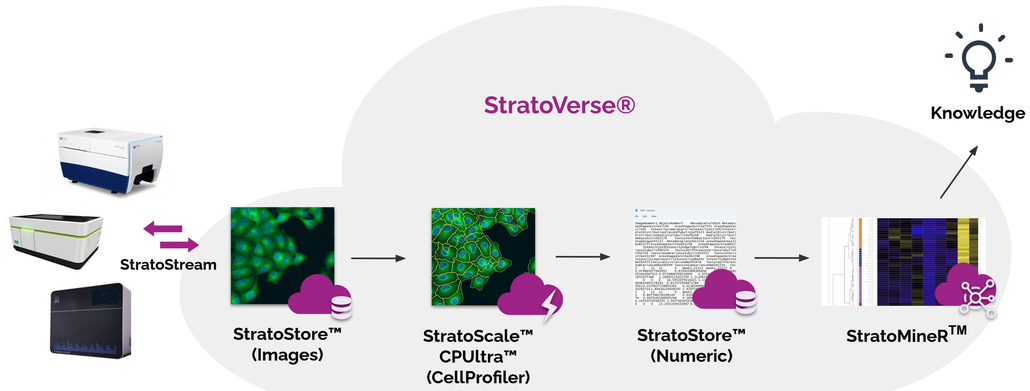

Figure 2: StratoVerse® is a complete, cloud-based platform for high content analysis. It features apps for data storage (StratoStore™), high powered image analysis (StratoScale™ and CPUltra™) and intuitive data analytics (StratoMineR™).

Results Numeric Data

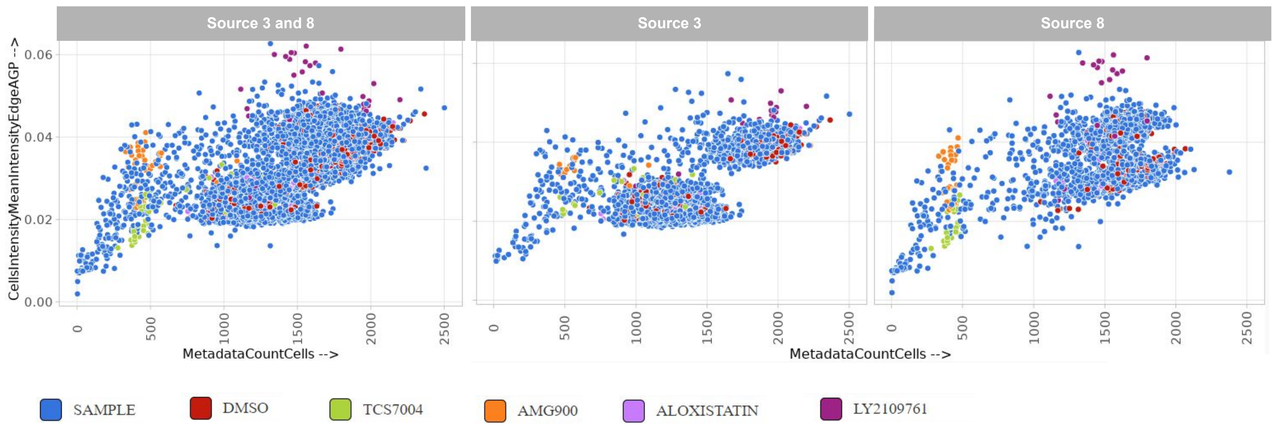

Figure 3: Raw data from two JUMP-CP sources. We analyzed 8 replicate compound and 1 Target2 plates from source 3 and 8 (n = 18 plates). We used the StratoMineR™ Quality Control interactive data visualization module to get an overview of the data. We used the Merge Metadata module to combine an annotation file with the raw data, this allows to include details about the experiment (compound names, reagent classes, etc) which results in more plotting options. Plotted here are scatter plots of two features out of the 5000+ features present.The data points are colored based on some positive and negative controls and the data is tiled based on source.

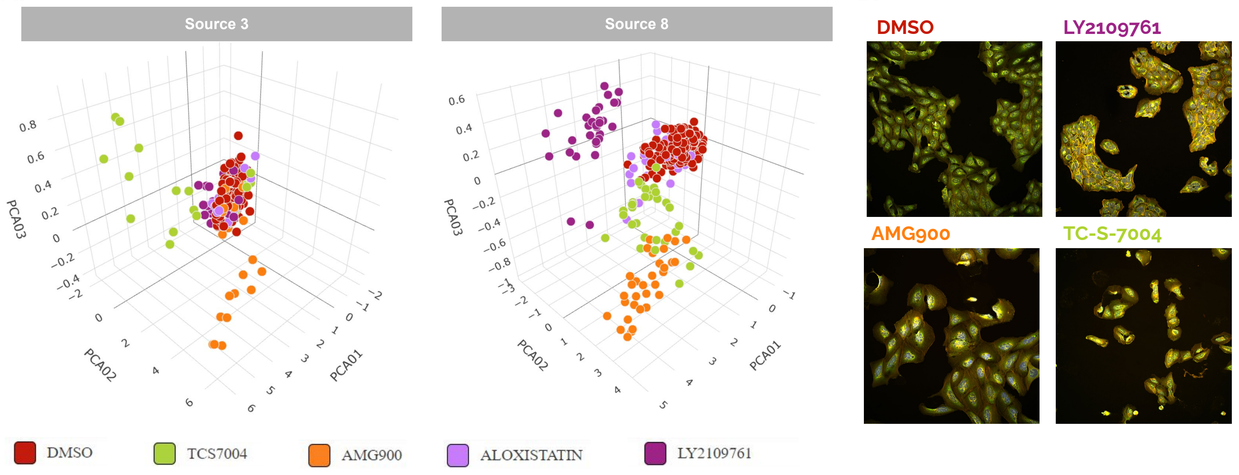

Figure 4: Dimensionality Reduction, Principal Component Analysis. Due to the large number of features in the JUMP-CP dataset, we performed PCA to reduce the computational load and redundancy and reveal the biology behind the data. **A**) Differences between sources in the first 3 components; note that some compounds such as LY2109761 show a more distinct phenotype compared to the negative control in Source 8. **B**) Images taken from the Cell Painting repository from source 8, wells treated with compounds shown in A.

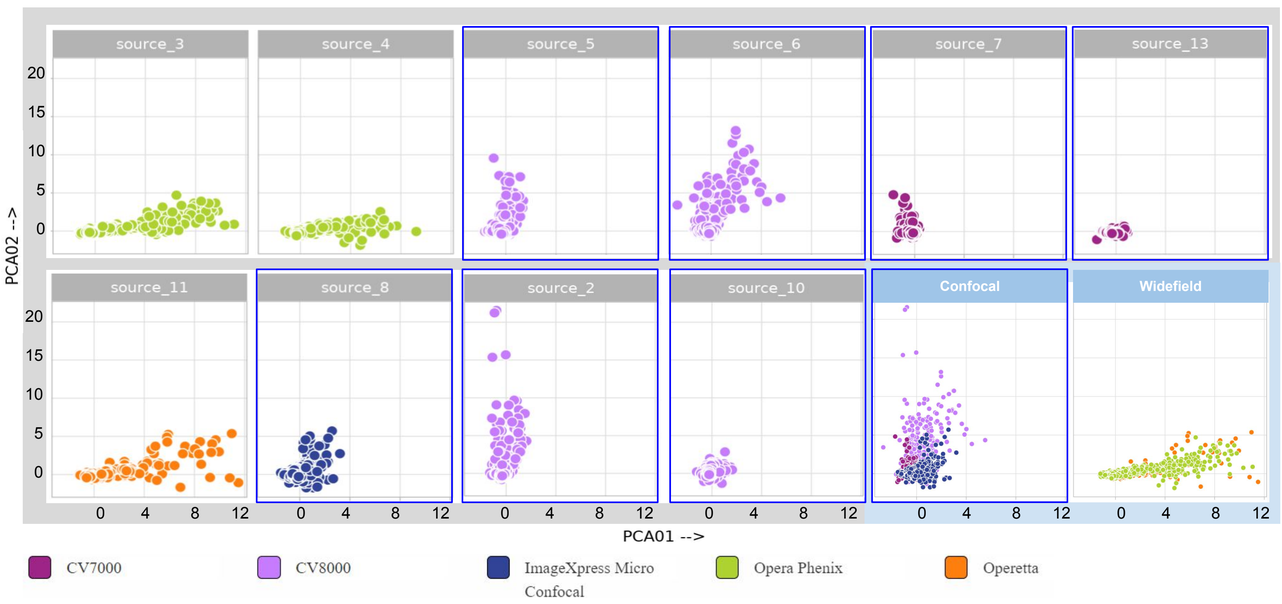

Figure 5: Target2 plate comparison based on microscope type. We selected 3 replicate Target2 plates from the sources shown above and performed PCA (n = 30 plates). Component 1 is shown on the X axis and component 2 on the Y axis. The data points are colored by microscope type. Note the patterns in the data depending on the microscopes used for image acquisition, when grouping the sources by confocal or widefield microscopes, the differences become more apparent.

Conclusions Numeric Data

We demonstrate that StratoMineR can be used to generate actionable knowledge from the JUMP-CP data. Our analysis indicates that the variability between the sources might be due to the type of microscope used for image acquisition. We will continue to explore the JUMP-CP numerical data to understand whether the variability observed here holds when analyzing bigger subsets of the data set.

Results Image Data

In order to validate the integration of Cell Profiler within our scalable cloud computing system (CPUltra) we downloaded raw images from Source 8 and run the JUMP-CP image analysis pipeline[2].The images from the 18 plates shown below were processed at a speed of 7 min/plate. We then analyzed the resulting numerical data in StratoMineR™ following the steps in Figure 1 and compared it with the original numerical data in the Cell Painting repository.

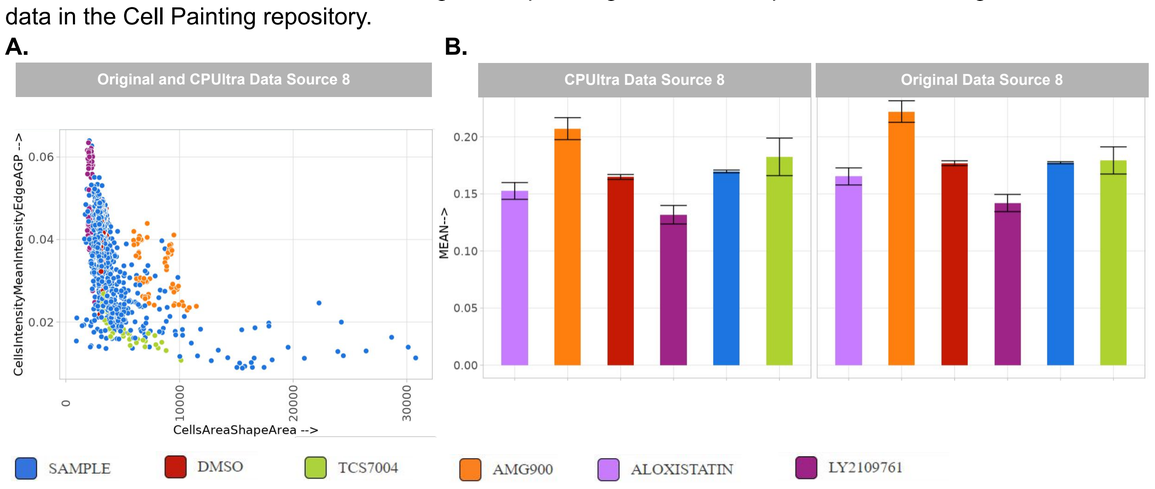

Figure 6: Raw data from Source 8: original and extracted using CPUltra. A) Scatter plot with both data origins. B) Bar plots grouped by origin, feature shown: CellsTextureInverseDifferenceMomentRNA1000256.

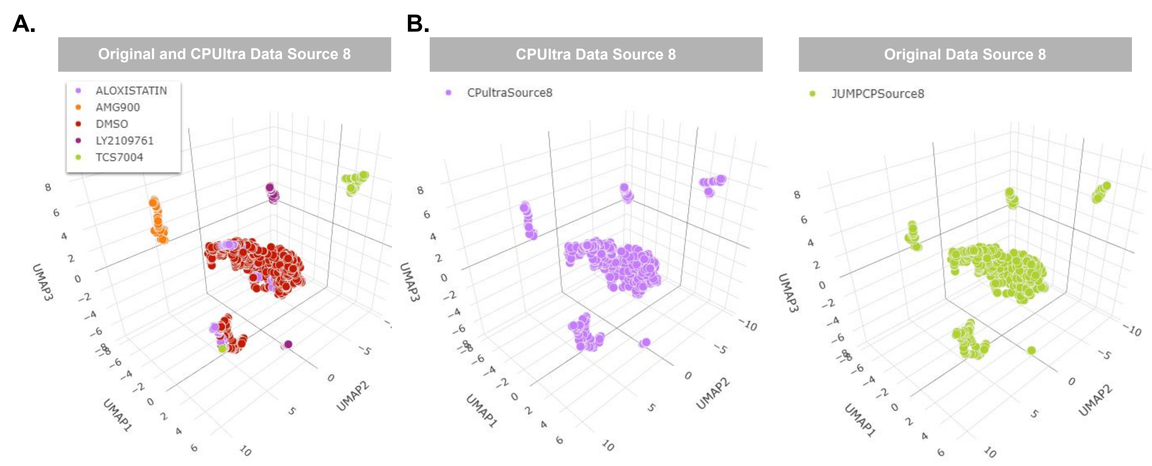

Figure 7: Uniform manifold approximation and projection (UMAP) analysis. This method is commonly used for dimensionality reduction and clustering purposes, we applied Euclidean distance and k neighbors = 15. **A**) Colored by reagent class. Note the most of the positive control compounds cluster together and are phenotypically distinct from the negative controls. **B**) Colored by data origin. Note the consistency between the original data and the CPUltra data.

Well Location	Compound	Data Origin	Euclidean Distance Score	p value
a24	LY2109761	JUMPCPSource8	12.77	0.00000000
a24	LY2109761	CPultraSource8	12.73	0.00000000
b01	AMG900	JUMPCPSource8	11.88	0.00000000
b01	AMG900	CPultraSource8	11.01	0.00000000
d24	TCS7004	JUMPCPSource8	8.84	0.00000000
d24	TCS7004	CPultraSource8	7.34	0.00000052
e01	LY2109761	JUMPCPSource8	9.41	0.00000000
e01	LY2109761	CPultraSource8	9.19	0.00000000
f24	AMG900	JUMPCPSource8	11.16	0.00000000
f24	AMG900	CPultraSource8	10.69	0.00000000
h01	TCS7004	JUMPCPSource8	9.01	0.00000000
h01	TCS7004	CPultraSource8	6.99	0.00000271

Table 1: Hit selection list. We used Euclidean metric to calculate the distance from the median of the negative controls to all compounds and defined $p < 0.05$. Shown here is the top part of the hit selection list, that coincides with the positives controls, which is to be expected. Note that the same hits on the same well locations were identified in both data origins.

Conclusions Image Data

We show that CPUltra can be utilized to perform image analysis in a reliable, reproducible and rapid way. This could help alleviate the computational power limitations when running CellProfiler in standard CPUs and, overall, accelerate the drug discovery process.

1. We used the dataset cpg0016 (Chandrasekaran et al., 2022), available from the Cell Painting Gallery on the Registry of Open Data on AWS (https://registry.opendata.aws/cellpainting-gallery/).
2. https://github.com/broadinstitute/imaging-platform-pipelines/blob/master/JUMP_production/JUMP_analysis_v3.cppipe