GET READY FOR JUMP-CP: ANALYSIS OF A PILOT DATA SET REVEALS CRITICAL INSIGHTS FOR PLATFORM DEVELOPMENT

Victor S.C. Wong, Wienand Omta, David A. Egan
Core Life Analytics B.V., 's Hertogenbosch, The Netherlands.

Introduction

There is a growing interest in adopting automated high content phenotypic profiling method for target and drug discovery pipelines. The Broad Institute has recently established a JUMP (Joint Undertaking in Morphological Profiling) Cell Painting consortium to generate a large public reference Cell Painting dataset. The aim is to present the 'ground truth' for the study of phenotypic relationships between various biological perturbations that target the same genes in cells. Access to such a dataset has the potential to accelerate drug discovery projects, but the large volume of data also presents challenges for scientists. We have analyzed a preliminary subset of the JUMP-CP dataset in our StratoMineR™ analytics platform. We show that our web-based data analytics platform, StratoMineR™, can detect and generate distinct plate maps, select relevant features, and perform dimensionality reduction and unbiased hit picking. We can make several comparisons to examine differences in phenotypic outcomes between cell lines, time points and conditions. Moreover, we further identified compounds that give robust and distinct phenotypes across two cell lines and time points. These compounds have utility for assay development and validation. Our analysis from StratoMineR™ provides valuable insights that can break the data analytics barrier for scientists who wish to develop their own Cell Painting platforms and take advantage of the full JUMP-CP dataset when it becomes available.

Method

The JUMP-CP is a high content (HC) screening dataset of microscopic images and image-based phenotypic profiles from two cell lines (A549 and U2OS) treated with chemical compounds and genetic perturbations. The cells were fixed and the standard Cell Painting assay protocol with six fluorescent dyes[1] were used to label various components of the cell. Diverse morphological features were segmented and extracted from CellProfiler. We used the publicly available preliminary JUMP-CP dataset which can be found in the Github repository[2].

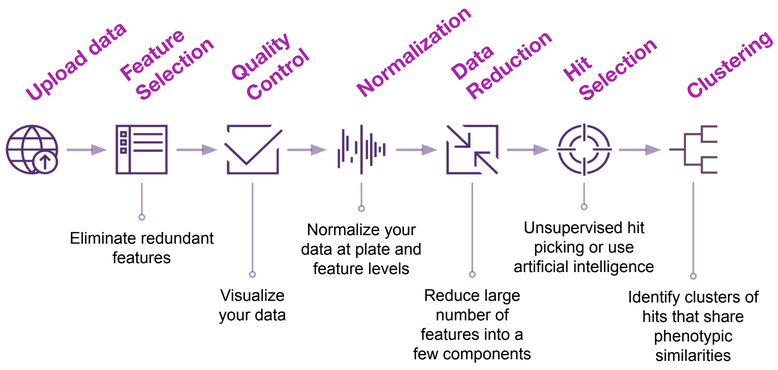
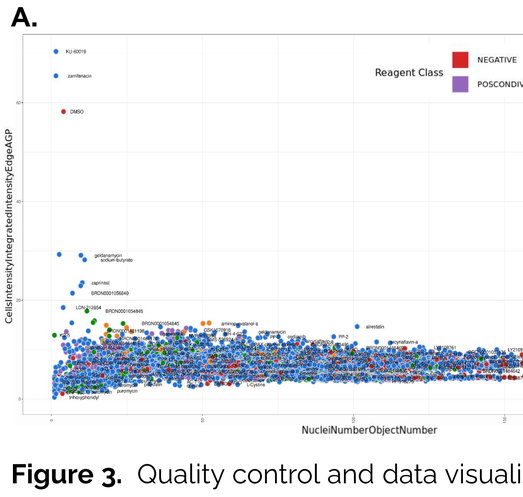


Figure 1. StratoMineR™ is a web-based platform which guides users through a typical workflow in analysis of high content multi-parametric data[3]. Starting with data upload in .csv or .txt formats, subsequent steps in StratoMineR™ allow rapid data mining and analyses.

Results

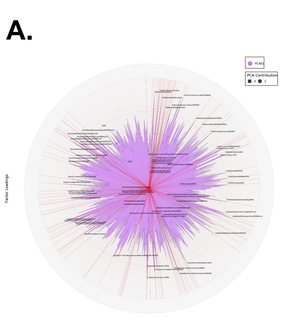
We began our approach to analyze the JUMP-CP dataset by eliminating redundant measurements (**Figure 2**). Critical annotations (e.g., cell line, reagent classes, compound names, gene targets, time) were added into the analysis using the Merge Metadata step in StratoMineR™. We could then generate three distinct plate maps based on the reagent classes (not shown). All data could be labeled with relevant annotations, and separated by cell line using the Visual Data Mining (VDM) interactive data visualization tool in Quality Control (**Figure 3**). Further downstream analysis was performed such as plate normalization, data transformation, and feature scaling to normalize the range of independent measurements (data not shown).



Feature	Closest Feature
NucleiNumberObjectNumber	CellsNumberObjectNumber
CellsNeighborsFirstClosestDistance5	CellsNeighborsFirstClosestDistanceAdjacent
CellsAreaShapeBoundingBoxMaximumY	CytoplasmAreaShapeBoundingBoxMaximumY
CellsAreaShapeBoundingBoxMaximumX	CytoplasmAreaShapeBoundingBoxMaximumX
CellsNeighborsAngleBetweenNeighbors5	CellsNeighborsAngleBetweenNeighborsAdjacent
CellsIntensityMinIntensityDNA	CytoplasmIntensityMinIntensityDNA
CellsIntensityMinIntensityEdgeRNA	CytoplasmIntensityMinIntensityEdgeRNA
CytoplasmCorrelationMandersBrightfieldDNA	CytoplasmCorrelationMandersLowZBFDNA
CellsCorrelationKMitoBrightfield	CellsCorrelationKMitoLowZBF
CellsIntensityMinIntensityAGP	CytoplasmIntensityMinIntensityAGP
CellsCorrelationMandersBrightfieldMito	CellsCorrelationMandersLowZBFMito
CellsCorrelationKAGPBrightfield	CellsCorrelationKAGPLowZBF
CellsCorrelationKERBrightfield	CellsCorrelationKERLowZBF
CellsCorrelationKBrightfieldMito	CellsCorrelationKLowZBFMito
NucleiIntensityMinIntensityDNA	NucleiIntensityMinIntensityEdgeDNA

Figure 2. Feature redundancy in the preliminary JUMP-CP dataset. There were a total of 5739 features extracted from CellProlifer software for the JUMP-CP consortium. Among those, 3486 features were highly correlated with another feature with correlation value of 0.99 or higher. The correlation matrix shows only 250 features, and the table lists top 15 features with the most variance, and their correlated features ('Closest Features') with the correlation values.

Figure 3. Quality control and data visualization. Using the StratoMineR™ VDM QC interactive data visualization module, we can quickly get an overview of the entire preliminary JUMP-CP dataset (**A**). We can also assign new reagent classes or labels (not shown). We used the merged metadata module in StratoMineR™ to combine annotation file with the raw data file to label samples, such as compound name or gene target. There is also a functionality in this step that allows for the tiling of data based on a metadata feature (example shown based on cell lines A549 (**B**) and U2OS (**C**)).



Feature	Loading
CellsCorrelationCorrelationDNAMito	0.76607147
NucleiRadialDistributionFracAtDER1of4	0.749011287
CellsAreaShapeZernike40	0.745050453
CellsAreaShapeEccentricity	-0.739707355
CellsAreaShapeZernike66	-0.730278203
CellsCorrelationCorrelationDNAER	0.725992801
CellsAreaShapeExtent	0.711136546
CellsAreaShapeZernike44	-0.709795077
CellsAreaShapeZernike75	-0.704073421
CellsAreaShapeZernike97	-0.685009443

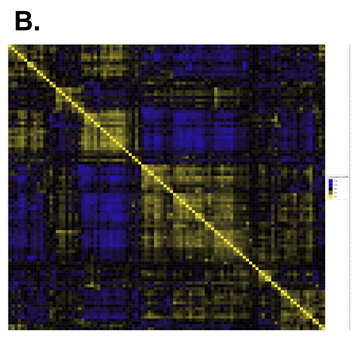
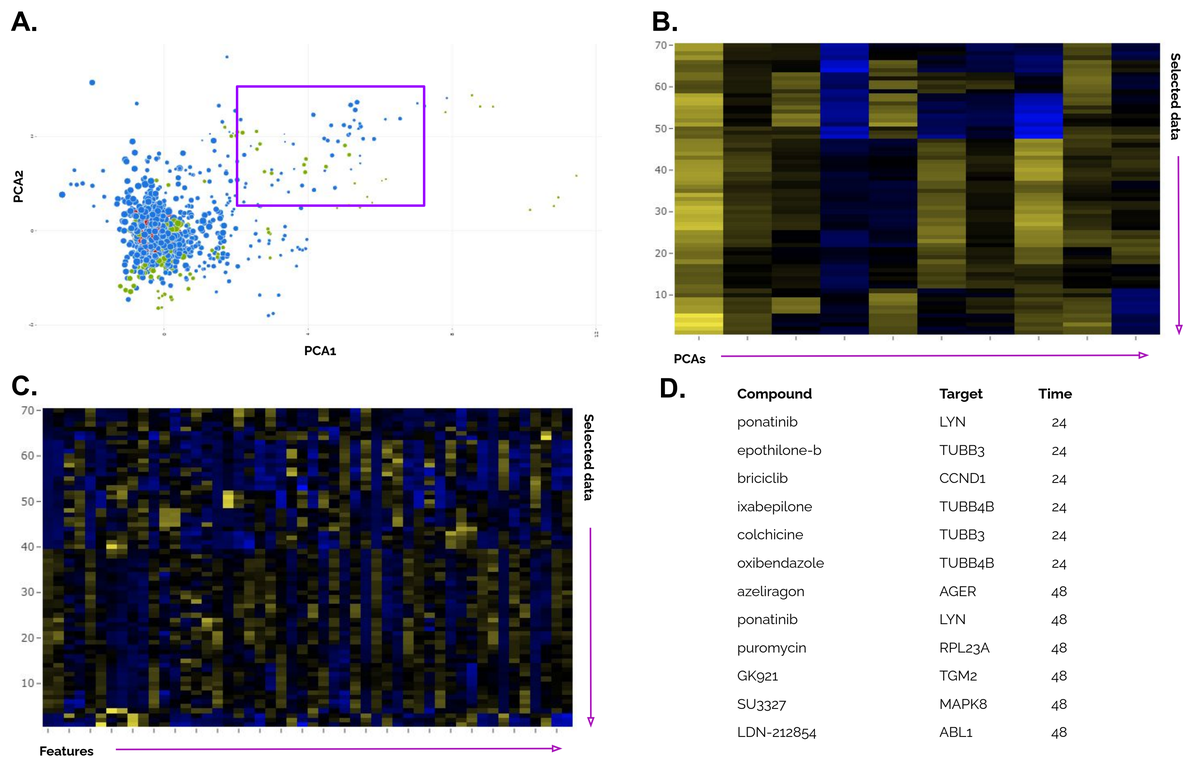


Figure 5. Principal component analysis. **A.** A rose plot showing one of the principal components (PCAs; total of 10 PCAs were generated based on a scree plot (not shown) that was generated from StratoMineR™), with top 10 features listed in the table with significant loading scores. **B.** Correlation matrix of the top 850 features. Additionally, PCA was also performed to capture phenotypic changes in compound-treated A549 and U2OS cell lines independently (not shown).

Results (Continued)



Compound	Target	Time
ponatinib	LYN	24
epothilone-b	TUBB3	24
briciclib	CCND1	24
ixabepilone	TUBB4B	24
colchicine	TUBB3	24
oxibendazole	TUBB4B	24
azeliragon	AGER	48
ponatinib	LYN	48
puromycin	RPL23A	48
GK921	TGM2	48
SU3327	MAPK8	48
LDN-212854	ABL1	48

Figure 6. Hit selection and clustering identified compounds with related and unrelated targets. Unsupervised hit selection using Euclidean distance scoring was used in StratoMineR™ for a subset of data from compound-treated A549 experiment, and distance scores were calculated from the median of the negative controls with $p < 0.05$. This approach identified 57 compound hits that were phenotypically distinct from the negative controls. Shown here is a hit selection scatter plot for A549 cell line (**A**) and selected hit compounds (inset) can be clustered based on 10 PCAs (**B**) or across 50 features (**C**). List of selected hits from compound-treated A549 cells reveals groups of related and unrelated targets (**D**).

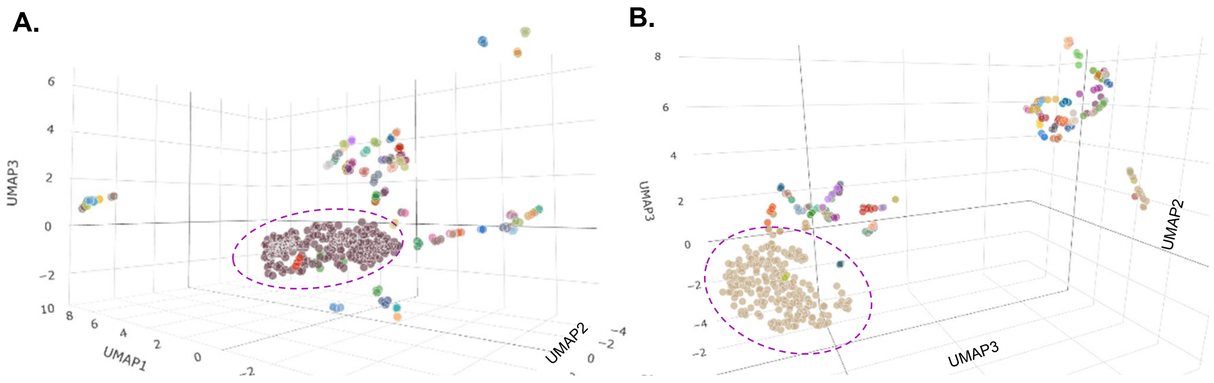


Figure 7. Uniform manifold approximation and projection (UMAP) analysis. In addition, we also cluster data from compound treated A549 (**A**) and U2OS (**B**) cell lines using 3D UMAP (only 48 hour data is shown). Euclidean distance metric was used with *k* neighbors = 15. StratoMineR™ platform allows viewing of data points based on compounds that are phenotypically distinct from the negative control (indicated by purple dashed line). Please contact us to find out more.

Compound	Target	Broad Sample ID	Pubchem ID	InChIKey	Target Radix
AVL292	BTK	BRD-K87782578-001-01-4	59174488	KXBDTLQSDKGAEB-UHFFFAOYSA-N	70
Briciclib	CCND1	BRD-K05531427-001-01-7	11248490	LXENKEWVEVKKGV-BQYQJAHWSA-N	129
Colchicine	TUBB3	BRD-K00259736-001-16-4	6167	IAKHMKGGTNLKSZ-INIZCTEOSA-N	18
Cycloheximide	RPL3	BRD-K36055864-001-20-0	6197	YPHMISFOHDHNIV-FSZOTQKASA-N	46
Delanzomib	CTSG	BRD-K59325863-001-04-4	24800541	SJFBTAPEPRWNKH-CCKFTAQKSA-N	137
Epothilone B	TUBB3	BRD-K68164687-001-01-6	129010071	QXRSDHAAWVKZLJ-AXCXJGFASA-N	18
Fludarabine Phosphate	DCK	BRD-K71106091-001-09-5	30751	GIUYCYHIANZCFB-FJFJXFQQSA-N	158
KU60019	ATM	BRD-K36016295-001-05-2	15953870	SCELLOWTHJGVIC-BGYRXZFFSA-N	30
NVP HSP990	HSP90AB1	BRD-K67915505-001-01-2	46216556	WSMQUUGTQYPVPD-OAHLLOKOSA-N	65
Oxibendazole	TUBB4B	BRD-K52075715-001-06-7	4622	RAOCRURYZCVHMG-UHFFFAOYSA-N	90
PF431396	PTK2B	BRD-K93788137-001-04-1	11598628	POJZIZBONPAWIV-UHFFFAOYSA-N	96
PP2	LYN	BRD-K95785537-001-22-3	4878	PBBRWFOVCUAONR-UHFFFAOYSA-N	86
Puromycin	RPL23A	BRD-K01824976-300-02-9	5702164	RXWNCPJZOCPEPQ-DYKMBDCPSA-N	94
TG02	CDK9	BRD-K14560436-001-01-4	16739650	VXBAJLGYBMTJCY-NSCUHMNNSA-N	79

Table 1. List of phenotypically unique JUMP-CP compounds. We have compiled a list of compounds that demonstrated robust and distinct phenotypes across A549 and U2OS cell lines and two time points (24 and 48 hr). These compounds are available from Specs (**www.specs.net**) for Cell Painting assay development purposes.

Conclusion

We demonstrate that StratoMineR™ can be used to generate actionable knowledge from the preliminary JUMP-CP dataset. Our analysis highlights substantial redundancy in the dataset. Our online data analytics platform gives the user the ability to dynamically label and tile data points using external annotations via Merge Metadata. Moreover, StratoMineR™ allows one to explore subsets of the JUMP-CP dataset to examine phenotypic changes based on cell line, time point, and experiment type. From the compound screen alone, substantial numbers of treatments gave diverse phenotypic outcomes and most of these compounds have related and unrelated targets. We will further apply the Artificial Intelligence functionality of StratoMineR™ for building specific phenotypic classes but this requires object-level data, which is currently under investigation. More JUMP-CP analyses are available: please come to our **booth R11**!

References

1. *Bray MA et al. Nat Protoc. 2016 Sep; 11(9): 1757-1774.*
2. *Chandrasekaran SN et al. bioRxiv https://doi.org/10.1101/2022.01.05.475090.*
3. *Omta W et al. Assay Drug Dev Technol. 2016; 14(8): 439-452.*

Core Life Analytics

Your Experiment. Your Data. Your Analysis.