RESEARCH ARTICLE

PopulationProfiler: A Tool for Population Analysis and Visualization of Image-Based Cell Screening Data

Damian J. Matuszewski[1,3]*, Carolina Wählby[1,3], Jordi Carreras Puigvert[2,4☯], Ida-Maria Sintorn[1,3☯]

1 Science for Life Laboratory, Uppsala, Sweden, **2** Science for Life Laboratory, Stockholm, Sweden, **3** Centre for Image Analysis, Department of Information Technology, Uppsala University, Uppsala, Sweden, **4** Division of Translational Medicine and Chemical Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden

☯ These authors contributed equally to this work.
* damian.matuszewski@it.uu.se

Citation: Matuszewski DJ, Wählby C, Puigvert JC, Sintorn I-M (2016) PopulationProfiler: A Tool for Population Analysis and Visualization of Image-Based Cell Screening Data. PLoS ONE 11(3): e0151554. doi:10.1371/journal.pone.0151554

Editor: Michael Polymenis, Texas A&M University, UNITED STATES

Received: December 17, 2015

Accepted: March 1, 2016

Published: March 17, 2016

Data Availability Statement: Data are available at the following DOI: 10.5281/zenodo.47050.

Funding: This work was financially supported by Science for Life Laboratory Sweden and Swedish Research Council (Vetenskapsrådet http://www.vr.se/) grants 2012-4968 to CW and 2014-6075 to IMS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Image-based screening typically produces quantitative measurements of cell appearance. Large-scale screens involving tens of thousands of images, each containing hundreds of cells described by hundreds of measurements, result in overwhelming amounts of data. Reducing per-cell measurements to the averages across the image(s) for each treatment leads to loss of potentially valuable information on population variability. We present PopulationProfiler—a new software tool that reduces per-cell measurements to population statistics. The software imports measurements from a simple text file, visualizes population distributions in a compact and comprehensive way, and can create gates for subpopulation classes based on control samples. We validate the tool by showing how PopulationProfiler can be used to analyze the effect of drugs that disturb the cell cycle, and compare the results to those obtained with flow cytometry.

Introduction

Automated image-based high-content microscopy provides a platform for phenotypic screening of complex compound libraries and drug combination sets [1]. Image processing and analysis tools enable automated extraction of large numbers of quantitative measurements describing the phenotype on a single cell basis [2]. Predicting and characterizing the mechanism of action of each compound in a large library typically requires careful analysis of this multidimensional data. However, many studies reduce per-cell measurements to population means, leading to loss of potentially valuable information about population heterogeneity [3, 4]. Such an approach is not very surprising considering the complexity of handling hundreds of measurements from hundreds of cells per treatment, in assays often spanning libraries of thousands of compound-dose combinations.

There are commercially available software that allow definition and quantification of sub-populations such as Screener by GeneData, SpotFire by TIBCO, IN Cell Investigator Software by GE Healthcare, and Harmony by PerkinElmer. Additionally, the machine-learning tools within CellProfiler Analyst [5] and other software [6] can be trained to identify and count cells belonging to different sub-populations. However, to our knowledge, no simple, free and open source tools for full-plate visualization of per-cell measurement distributions has previously been presented.

We present PopulationProfiler, software that allows visualization of histograms and sub-population distribution of high-content screening data stored in the common csv text file format. The main idea is to reduce per-cell measurements to per-well distributions, each represented by a histogram, and optionally further reduce the histograms to sub-type counts based on gating (setting bin ranges) of known control distributions and local adjustments to histogram shape. Such analysis is necessary in a wide variety of applications, e.g. DNA damage assessment using foci intensity distributions, assessment of cell type specific markers, and cell cycle analysis. We show how PopulationProfiler can be used for cell cycle perturbation, protein translocation, and EdU incorporation analysis.

PopulationProfiler is written in Python which makes it platform independent. The source code, sample dataset and an executable program (for Windows only) are freely available at http://cb.uu.se/~damian/PopulationProfiler.html.

Methodology

PopulationProfiler's simple graphical user interface (GUI) imports data from image-based screening measurements; it allows selection of multiple csv files containing information on treatment and position (well) within a multi-well plate. Each file is considered as an independent experiment with rows representing individual cell measurements. One type of measurement is processed at a time and cells are grouped (aggregated) based on well labels. The labels for cell aggregation and the measurement are selected by the user from a drop-down list created from the csv file header (first row). The GUI also allows selection of control wells based on the treatment labels (there can be more than one well per treatment). If such labels are not available, the user can select control wells manually. The corresponding data is pooled and stored as a separate record in the output csv file. PopulationProfiler thereafter calculates and displays the distribution of the selected measurement as a histogram for each well (Fig 1a). A vector representation of each well's histogram is saved in the output file, and can be used as input for e.g., cluster analysis, elsewhere. The cell count for each well is also saved as a measure of statistical relevance of population effects. A very low cell count usually indicates cell death, and morphological measurements are then less likely to convey useful information.

Case study—cell cycle analysis

A commonly studied treatment response is disruption of the cell cycle. We therefore added functionality specialized for analysis of relative per-cell DNA content, measured as log2 of the integrated intensity of a DNA stain such as DAPI, Hoechst or PI [7]. For an unperturbed cell population, a histogram of the DNA content typically consists of two peaks, as shown in Fig 1a (DMSO). The higher peak to the left (2N) corresponds to the larger part of the cell population with a single copy of the genome, whereas the smaller peak on the right (4N), corresponds to the sub-population that has doubled the amount of DNA. Before exploring the effect of treatments that potentially perturb the cell cycle, PopulationProfiler allows the user to set bin ranges (subpopulation gates) using data from untreated control wells. Values corresponding to the centers of the 2N and 4N sub-populations are defined as the largest and second largest

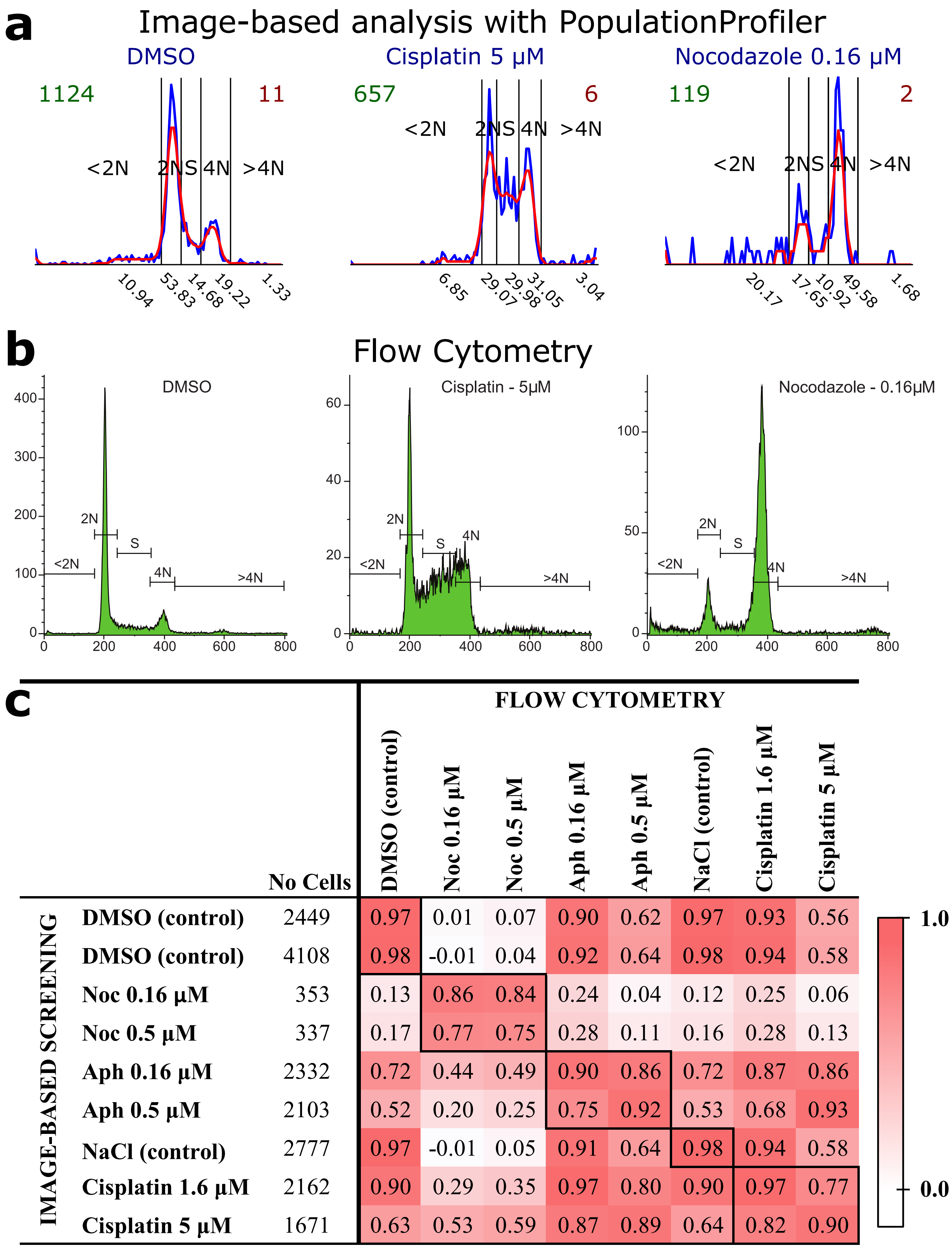

IMAGE-BASED SCREENING	No Cells	DMSO (control)	Noc 0.16 μM	Noc 0.5 μM	Aph 0.16 μM	Aph 0.5 μM	NaCl (control)	Cisplatin 1.6 μM	Cisplatin 5 μM
		FLOW CYTOMETRY							
DMSO (control)	2449	0.97	0.01	0.07	0.90	0.62	0.97	0.93	0.56
DMSO (control)	4108	0.98	-0.01	0.04	0.92	0.64	0.98	0.94	0.58
Noc 0.16 μM	353	0.13	0.86	0.84	0.24	0.04	0.12	0.25	0.06
Noc 0.5 μM	337	0.17	0.77	0.75	0.28	0.11	0.16	0.28	0.13
Aph 0.16 μM	2332	0.72	0.44	0.49	0.90	0.86	0.72	0.87	0.86
Aph 0.5 μM	2103	0.52	0.20	0.25	0.75	0.92	0.53	0.68	0.93
NaCl (control)	2777	0.97	-0.01	0.05	0.91	0.64	0.98	0.94	0.58
Cisplatin 1.6 μM	2162	0.90	0.29	0.35	0.97	0.80	0.90	0.97	0.77
Cisplatin 5 μM	1671	0.63	0.53	0.59	0.87	0.89	0.64	0.82	0.90

Fig 1. Image-based cell cycle analysis of cell line A549 with PopulationProfiler and its comparison to flow cytometry. a) DNA content histograms created with PopulationProfiler. The blue and red lines show data before and after smoothing, respectively. The numbers under the x-axis present the percentage contribution of each cell cycle sub-population. b) The corresponding cell cycle analysis with flow cytometry. c) A comparison of the results (the

contributions of the 5 cell cycle sub-populations) reveals high correlation. The respective total cell counts used by PopulationProfiler and flow cytometry are 18292 and 102751.

doi:10.1371/journal.pone.0151554.g001

maximum respectively, and all DNA intensity measurements are normalized such that the maximum of the 2N peak corresponds to 1 and the center of the 4N peak corresponds to 2. In order to avoid multiple peaks at 2N and 4N locations the histograms are smoothed with a Gaussian filter ($\sigma = 1.5$). Individual cells are thereafter assigned to five classes named <2N, 2N, S, 4N, and >4N based on thresholds at 0.75, 1.25, 1.75 and 2.5 respectively, in accordance with [7]. During analysis of treated wells, thresholds are automatically adjusted to the shape of each well's histogram within limits defined by the 2N and 4N peaks of the untreated wells. This adaptive gating allows a comparison of cell cycle effects decoupled from changes in, e.g., cell size or uptake of DNA stain. The alternative, i.e. using the gates found for the negative controls for all the other samples, is also possible with the PopulationProfiler. In addition, the tool allows setting manual customized gates which gives more analysis possibilities (arbitrary number and range of non-overlapping gates) to the user.

Results

The experiments performed had two goals; to compare population data collected by the presented image-based cell cycle analysis approach using PopulationProfiler to population data collected by flow cytometry (one dimensional intensity measurements), and to compare their ability to detect treatments that disturb the cell population. We tested the two cell cycle analysis approaches on a cancer cell line (lung cancer, A549, known to be sensitive to cell cycle perturbants) and a slow replicating control cell line (insensitive to cell-cycle perturbations non-transformed colon epithelial, CCD841) exposed to five treatments (DMSO, Aphidicolin, Nocodazole, NaCl and Cisplatin) at one or two doses, as detailed in the supplementary material. Cell cycle histograms were compared visually (Fig 1a and 1b and Figs B and C in S1 Text), and by calculating Pearson's correlation coefficient of normalized cell cycle sub-population distribution vectors found using PopulationProfiler and Beckman Coulter Kaluza software for the flow cytometry data. Data from the drug sensitive cell line (A549) showed high Pearson's correlation coefficient for all equal drug-dose comparisons (Fig 1c and Fig D in S1 Text) and relatively high correlation for different doses of the same drug, while noticeably lower correlation between the effect of different drugs. A very similar pattern appears when comparing replicates of the flow cytometry experiments, while no effects were observed for the more stable CCD841 cell line (Fig E in S1 Text). More results, detailed experiment description and discussion can be found in S1 Text, together with example of protein translocation and EdU incorporation analysis.

Discussion

Rather than reducing per-cell measurements to population averages, PopulationProfiler allows data reduction while maintaining information on population heterogeneity. We show that PopulationProfiler keeps enough information to discriminate between drugs that perturb the cell cycle with similar detail as obtained by flow cytometry, but at significantly lower cell counts. Image based analysis allows efficient discrimination between true signals and artifacts and PopulationProfiler enables comparison of measurements of morphological features, such as sub-cellular signal localization and cytoskeletal patterns, not possible to observe by flow cytometry.

Supporting Information

S1 Text. PopulationProfiler: Supplementary Material. This file contains the user manual, additional application examples, and the detailed description and discussion of the presented cell cycle analysis comparison.
(PDF)

Acknowledgments

We want to thank Thomas Helleday and Sven Nelander for providing access to flow- and image data.

Author Contributions

Conceived and designed the experiments: CW JCP IS. Performed the experiments: JCP. Analyzed the data: DJM. Wrote the paper: DJM CW JCP IS.

References

1. Bickle M. The beautiful cell: high-content screening in drug discovery. Anal. Bioanal. Chem. 2010; 398: 219–226. doi: 10.1007/s00216-010-3788-3 PMID: 20577725
2. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006; 7. doi: 10.1186/gb-2006-7-10-r100 PMID: 17076895
3. Gough AH, Chen N, Shun TY, Lezon TR, Boltz RC, Reese CE et al. Identifying and quantifying heterogeneity in high content analysis: application of heterogeneity indices to drug discovery. PLoS ONE. 2014; 9. doi: 10.1371/journal.pone.0102678
4. Ljosa V, Caie PD, Horst R, Sokolnicki KL, Jenkins EL, Daya S et al. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. J. Biomol. Screen. 2013; 18: 1321–1329. doi: 10.1177/1087057113503553 PMID: 24045582
5. Jones TR, Kang IH, Wheeler DB, Lindquist RA, Papallo A, Sabatini DM et al. CellProfiler Analyst: data exploration and analysis software for complex image-based screens. BMC Bioinform. 2008; 9. doi: 10.1186/1471-2105-9-482
6. Sommer C, and Gerlich DW. Machine learning in cell biology–teaching computers to recognize phenotypes. J. Cell Sci. 2013; 126: 5529–5539. doi: 10.1242/jcs.123604 PMID: 24259662
7. Chan GKY, Kleinheinz TL, Peterson D, and Moffat JG. A simple high-content cell cycle assay reveals frequent discrepancies between cell number and ATP and MTS proliferation assays. PLoS ONE. 2013; 8.