Methods 112 (2017) 201–210

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

An open-source solution for advanced imaging flow cytometry data analysis using machine learning

Holger Hennig [a,b,c], Paul Rees [a,c], Thomas Blasi [d], Lee Kamentsky [a,1], Jane Hung [a], David Dao [a], Anne E. Carpenter [a], Andrew Filby [e,*]

[a] Imaging Platform at the Broad Institute of Harvard and MIT, 415 Main St, Cambridge, MA 02142, USA
[b] Dept. of Systems Biology & Bioinformatics, University of Rostock, 18051 Rostock, Germany
[c] College of Engineering, Swansea University, Singleton Park, Swansea SA2 8PP, UK
[d] Helmholtz Zentrum München — German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany
[e] Flow Cytometry Core Facility, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

ARTICLE INFO

Article history:
Received 4 July 2016
Received in revised form 18 August 2016
Accepted 31 August 2016
Available online 2 September 2016

Keywords:
Imaging flow cytometry
Machine learning
Open-source software
High-throughput
Feature selection
Profiling

ABSTRACT

Imaging flow cytometry (IFC) enables the high throughput collection of morphological and spatial information from hundreds of thousands of single cells. This high content, information rich image data can in theory resolve important biological differences among complex, often heterogeneous biological samples. However, data analysis is often performed in a highly manual and subjective manner using very limited image analysis techniques in combination with conventional flow cytometry gating strategies. This approach is not scalable to the hundreds of available image-based features per cell and thus makes use of only a fraction of the spatial and morphometric information. As a result, the quality, reproducibility and rigour of results are limited by the skill, experience and ingenuity of the data analyst. Here, we describe a pipeline using open-source software that leverages the rich information in digital imagery using machine learning algorithms. Compensated and corrected raw image files (.rif) data files from an imaging flow cytometer (the proprietary .cif file format) are imported into the open-source software CellProfiler, where an image processing pipeline identifies cells and subcellular compartments allowing hundreds of morphological features to be measured. This high-dimensional data can then be analysed using cutting-edge machine learning and clustering approaches using "user-friendly" platforms such as CellProfiler Analyst. Researchers can train an automated cell classifier to recognize different cell types, cell cycle phases, drug treatment/control conditions, etc., using supervised machine learning. This workflow should enable the scientific community to leverage the full analytical power of IFC-derived data sets. It will help to reveal otherwise unappreciated populations of cells based on features that may be hidden to the human eye that include subtle measured differences in label free detection channels such as bright-field and dark-field imagery.

1. Introduction

It is now widely accepted that cellular and molecular heterogeneity pervades all biological systems [1,2]. This creates a complex set of challenges for understanding how individual cells within heterogeneous communities interact with one another in order to determine the phenotype and function of higher organisms with respect to both healthy and disease states. Our ability to appreciate biological heterogeneity is limited by the resolving power of the analytical approaches at our disposal. At the methodological level, there is currently a massive paradigm shift away from so called "bulk" analysis techniques toward single cell-focused approaches that are able to cope far better with the challenges posed by heterogeneity [3]. "Cytometry" translates in literal terms to mean "cell measurement" and can best be described as the derivation of numbers from the measurement of large populations of single cells. While extremely powerful, it is a significant challenge to derive meaningful, objective conclusions from the high parameter output inherent to nearly all cytometric approaches. While cytometric technologies such as fluorescence-based flow and mass cytometry

Abbreviation: IFC, imaging flow cytometry.
* Corresponding author.
E-mail address: Andrew.Filby@newcastle.ac.uk (A. Filby).
[1] Current address: Visual Computing Group, SEAS, Harvard University, Cambridge, MA 02138, USA.

http://dx.doi.org/10.1016/j.ymeth.2016.08.018
1046-2023/

can currently measure 30–40 parameters per cell [4], the parameter output from image-based cytometry systems can be almost infinite and often continuous (non-discrete) in nature.

One very powerful image-based cytometric technology is imaging flow cytometry (IFC, Fig. 1). It combines the high-throughput, multi-parameter capabilities of conventional flow cytometry with the current capability to capture up to 12 spatially registered multi-spectral images for each cell as it passes through the system [5]. These imaging channels can capture label-dependent fluorescence signals (currently up to 10) as well as transmitted bright-field and laser side scatter (dark-field) information, the latter of which do not require any introduced fluorescence (label-free). IFC very much fits the paradigm of "image cytometry" as it produces quantifiable image data in a high throughput, multi-parameter format. This ensures that a fair, unbiased comparison can be made between the output images so that any measured differences can be considered biological rather than an artefact of variable imaging conditions.

In many cases the unique capabilities of IFC to deliver high-throughput, multispectral, spatially registered imagery has been essential to the development of new assays to ask novel, cutting edge biological questions. Most IFC-based assays take advantage of the technologies' inherent ability to measure fluorescence signals with spatial context. Such assays include measuring nuclear translocation [7], mitochondrial localisation [8], co-localization assays using "similarity" features [9], calcium signalling at the organelle level [10], organelle inheritance during mitosis [11], cell cycle phases [12], receptor activity [13], asymmetric cell division [14–16], fission yeast cell cycle [17], dendritic cell morphology [18], autophagy [19,20], detection of DNA damage foci [21,22] and modelling intracellular infection [23]. Most, if not all, of these assays would not be possible using traditional flow cytometry (lack of spatial information) or conventional imaging techniques (low throughput and poor quantitation).

Despite IFC being available to the research community for over 10 years it is still often referred to as a "new and emerging technology". In reality this is no longer the case. While new applications of the technology continue to be developed on a regular basis, the data analysis methods for IFC have noticeably lagged behind, and certainly fall significantly short of their potential. In fact, it could be argued that data analysis presents the single biggest bottleneck/barrier to a more comprehensive adoption of IFC by the research and clinical diagnostic communities.

The most common approach to IFC data analysis is to use the proprietary analysis software called IDEAS (manufacturer supplied). This software is extremely powerful, allowing the user to explore a range of image features derived from each individual cell. There are a number of pre-calculated features that measure pixel-based parameters (including morphological, intensiometric and texture based features) using default cell segmentation algorithms that automatically generate masks for each available imaging channels on a per cell/object basis. It is also possible within the IDEAS software to derive novel user-defined features based on either the default channel masks or completely novel masks/segmentations that can be constructed using a powerful suite of adaptation algorithms. Briefly, the latter allows the user to adapt

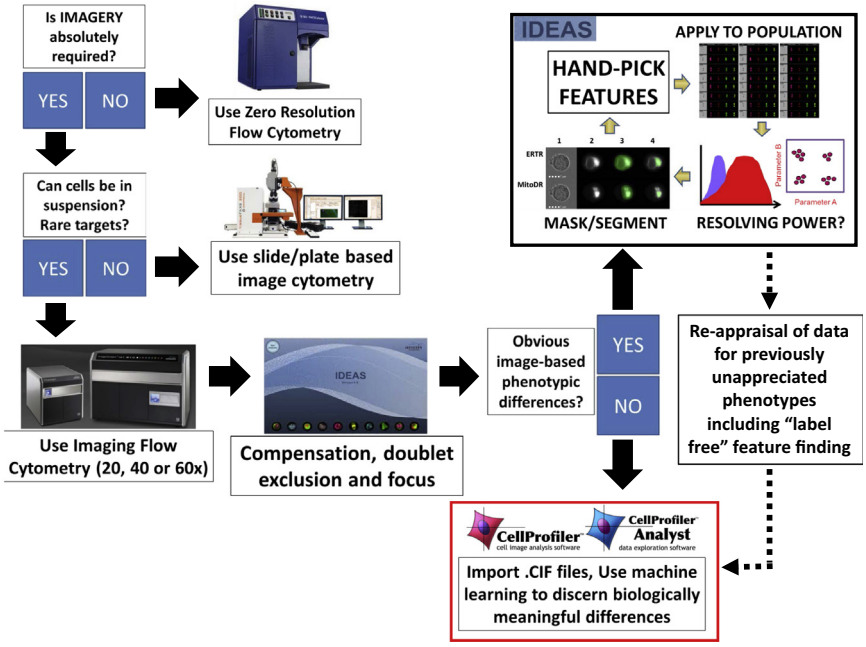

Fig. 1. Guidance on choosing cytometric method and analysis method. Any researcher who wants to use cytometry technology to ask a defined question should consider "what is the best approach" based on the question. For example if morphological/spatial information is not required then so-called "zero resolution flow cytometry" is best. If however the question absolutely requires imagery, then the sample type should next be considered, is it tissue? Can it be disaggregated? Could it be analysed in such a way that the spatial relationship of individual cells is lost? In our experience, IFC is best applied to situations where the cells biology can still be analysed when in suspension. This could still be disaggregated tissue or adherent cells and not just cells that exist in suspension. If the target cell population is rare, then suspension-based high throughput analysis is often necessary to collect sufficient events for statistical confidence. Once the IFC data is collected, several options can be chosen for data analysis. This figure summarises these options in light of our proposed solution. The historical option is to rely entirely on IDEAS software to perform a potentially subjective, iterative image analysis that involves adapting the masking/segmentation rules to best identify key pixels within an image channel and then to try and select the best feature calculated on these pixels with the aim of resolving different phenotypes from one another. This approach can be partially automated using the so-called "find the best feature" method. We propose however that a deeper analysis of features is more appropriate to IFC data sets. In this regard we have developed and validated a machine learning-based approach to analyse IFC data that has been corrected and compensated in IDEAS (.rif to .cif conversion). We then use the open source image analysis platforms CellProfiler and CellProfiler Analyst to better interrogate the imagery. Even in cases where the IDEAS-based iterative approach works very well, as is often the case when the outcome is well defined, there may be benefit to re-analysing these data using the approach presented here. It may uncover unappreciated features - in our own experience, this allowed us to perform a label-free classification of cell cycle stages, thus eliminating the need to add potentially confounding dyes to our cells [6].