Topical Review

Understanding health and disease with multidimensional single-cell methods

Julián Candia\textsuperscript{1,2,3}, Jayanth R Banavar\textsuperscript{1} and Wolfgang Losert\textsuperscript{1}

\textsuperscript{1} Department of Physics, University of Maryland, College Park, MD 20742, USA
\textsuperscript{2} School of Medicine, University of Maryland, Baltimore, MD 21201, USA
\textsuperscript{3} IFLYSIB and CONICET, University of La Plata, 1900 La Plata, Argentina

E-mail: candia@umd.edu

Received 5 November 2013, revised 25 November 2013
Accepted for publication 28 November 2013
Published 21 January 2014

Abstract

Current efforts in the biomedical sciences and related interdisciplinary fields are focused on gaining a molecular understanding of health and disease, which is a problem of daunting complexity that spans many orders of magnitude in characteristic length scales, from small molecules that regulate cell function to cell ensembles that form tissues and organs working together as an organism. In order to uncover the molecular nature of the emergent properties of a cell, it is essential to measure multiple-cell components simultaneously in the same cell. In turn, cell heterogeneity requires multiple-cells to be measured in order to understand health and disease in the organism. This review summarizes current efforts towards a data-driven framework that leverages single-cell technologies to build robust signatures of healthy and diseased phenotypes. While some approaches focus on multicolor flow cytometry data and other methods are designed to analyze high-content image-based screens, we emphasize the so-called Supercell/SVM paradigm (recently developed by the authors of this review and collaborators) as a unified framework that captures mesoscopic-scale emergence to build reliable phenotypes. Beyond their specific contributions to basic and translational biomedical research, these efforts illustrate, from a larger perspective, the powerful synergy that might be achieved from bringing together methods and ideas from statistical physics, data mining, and mathematics to solve the most pressing problems currently facing the life sciences.

Keywords: single-cell biophysics, complexity, flow cytometry, cell imaging, supercells

(Some figures may appear in colour only in the online journal)

Contents

1. Introduction 1
2. From molecules to cells to organisms: complexity, emergence, and the cell heterogeneity challenge 2
3. Single-cell-based phenotyping using multicolor flow cytometry data 3
4. Single-cell-based phenotyping using microscopy images 5
5. Emergent phenotyping via the Supercell/SVM paradigm 6

6. Conclusions 9
Acknowledgments 10
References 10

1. Introduction

Single-cell heterogeneity poses a huge challenge in the development and improvement of strategies for the diagnosis and treatment of many diseases. Indeed, it is a well-established fact that cells from the same tissue display significant qualitative
and quantitative heterogeneities, even within samples obtained from a single individual. This inherent biological diversity has complicated efforts to capture the essence of health and disease in terms of characteristic behaviors at the single-cell level and has, therefore, limited our ability to fully take advantage of new single cell analysis approaches to improve the current practice of personalized medicine.

For instance, Beckman et al [1] have very recently assessed the impact of single-cell heterogeneity, as well as that of genetic instability, in the development of effective nonstandard strategies for personalized cancer treatment. Manifestations of cell heterogeneity in healthy and diseased cell samples have ubiquitously been reported in the growing field of single-cell biology, ranging from human pluripotent embryonic stem cell cultures [2–4] and apoptosis mechanisms in cancer cell lines [5], to reversible adaptive plasticity in tumors such as human neuroblastoma [6] and pressure-driven shape features of C. elegans embryonic cells [7]. For recent reviews of the impact of tumor heterogeneity at different levels (genetic, epigenetic, the tumor microenvironment, the immune response, and other factors such as diet and the microbiota), see [8–11].

The difficulties of pinpointing specific characteristics of different healthy and diseased cell subpopulations prompted the development and refinement of experimental techniques that allow multidimensional measurements on single cells, such as e.g. multicolor flow cytometry [12–14], high performance kinetic image cytometry [15], and the very recently introduced mass cytometry (CyTOF) technique [16–18]. Certainly, the improvement of these experimental methods allows one to probe single cells in increasingly high-dimensional parameter spaces, which in turn enhances the resolution to identify and focus on specific cell subpopulations. As the experimental techniques evolve, however, the pressing need for improving our ability to process and analyze ‘Big Data’ in the life sciences becomes increasingly manifest. In fact, we need unbiased, mathematically robust, scalable methods that allow us to identify the key parameters that consistently characterize cell subpopulations across different samples in order to build signatures of health and disease across length scales spanning many orders of magnitude [19].

In this review, we summarize current data-driven efforts that leverage single-cell technologies to build robust signatures of healthy and diseased phenotypes. We focus on two key types of single-cell datasets, namely multicolor flow cytometry, in which each cell is characterized by a set of up to 20 measurements corresponding to scattering and fluorescent emission of light upon stimulation by laser beams, and microscopy via high-content image-based screens, in which multiple parameters characterize the shape of each cell, often used in combination with biomarker intensity measurements. In section 2, we discuss the challenges arising from biological complexity, emergent phenomena, and cell heterogeneity. In section 3, we review efforts to build phenotypes based on flow cytometry data analysis techniques. In section 4, we summarize profiling methods for microscopy image-based screens. In section 5, we present the Supercell/SVM paradigm, which is a general approach for emergent phenotyping that can be applied to different kinds of single-cell datasets, including multicolor flow cytometry and cell imaging. Finally, in section 6, we present our concluding remarks.

2. From molecules to cells to organisms: complexity, emergence, and the cell heterogeneity challenge

Certainly, gaining a molecular understanding of disease is the holy grail of current biomedical research and related interdisciplinary fields. The complexity of this problem, however, is daunting: from a physical perspective, this complexity arises from the plethora of scales, the great diversity of system components, and the sheer size of biological systems. Indeed, a human being is a hierarchically organized, multi-scale system that spans about 10 orders of magnitude in relevant length scales, from water molecules (0.2 nm), DNA molecules (2.5 nm) and proteins (typically 2–10 nm) to cells (10–100 µm), tissues, organs, and organ systems. Besides the multi-scale nature of the problem, complexity arises from the high-dimensional variety of subtypes (e.g. about 300 human cell types and 25 000 protein-coding genes) and the astonishingly large system size (e.g. 3 × 10^9 DNA base pairs contained in each one of the 10–100 trillion cells of a human being).

Physics is familiar with the concept of emergence, that is, the collective, large-scale generic behavior of a system that manifests itself independently of small-scale details [20]. In macroscopic systems, thermodynamics and statistical mechanics provide the theoretical framework to understand macroscopic emergence, which explain large-scale phenomena such as ferromagnetism and irreversibility. At mesoscopic scales, however, other kinds of emergent behavior may arise out of as yet unknown physical laws [21]. Collisions of macroscopic particles in low concentration granular suspensions, for instance, tend to disrupt the inherent reversibility of low Reynolds number fluid flows [22], and the onset of irreversibility decreases approximately as the square of particle density. In dense granular systems under cyclic shear, rearrangements in dense granular suspensions, the collective, large-scale generic behavior of a system that has, therefore, limited our ability to fully take advantage of new single cell analysis approaches to improve the current practice of personalized medicine.

In section 2, we discuss the challenges arising from biological complexity, emergent phenomena, and cell heterogeneity. In section 3, we review efforts to build phenotypes based on flow cytometry data analysis techniques. In section 4, we summarize profiling methods for microscopy image-based screens. In section 5, we present the Supercell/SVM paradigm, which is a general approach for emergent phenotyping that can be applied to different kinds of single-cell datasets, including multicolor flow cytometry and cell imaging. Finally, in section 6, we present our concluding remarks.

2. From molecules to cells to organisms: complexity, emergence, and the cell heterogeneity challenge

Certainly, gaining a molecular understanding of disease is the holy grail of current biomedical research and related interdisciplinary fields. The complexity of this problem, however, is daunting: from a physical perspective, this complexity arises from the plethora of scales, the great diversity of system components, and the sheer size of biological systems. Indeed, a human being is a hierarchically organized, multi-scale system that spans about 10 orders of magnitude in relevant length scales, from water molecules (0.2 nm), DNA molecules (2.5 nm) and proteins (typically 2–10 nm) to cells (10–100 µm), tissues, organs, and organ systems. Besides the multi-scale nature of the problem, complexity arises from the high-dimensional variety of subtypes (e.g. about 300 human cell types and 25 000 protein-coding genes) and the astonishingly large system size (e.g. 3 × 10^9 DNA base pairs contained in each one of the 10–100 trillion cells of a human being).

Physics is familiar with the concept of emergence, that is, the collective, large-scale generic behavior of a system that manifests itself independently of small-scale details [20]. In macroscopic systems, thermodynamics and statistical mechanics provide the theoretical framework to understand macroscopic emergence, which explain large-scale phenomena such as ferromagnetism and irreversibility. At mesoscopic scales, however, other kinds of emergent behavior may arise out of as yet unknown physical laws [21]. Collisions of macroscopic particles in low concentration granular suspensions, for instance, tend to disrupt the inherent reversibility of low Reynolds number fluid flows [22], and the onset of irreversibility decreases approximately as the square of particle density. In dense granular systems under cyclic shear, rearrangements in dense granular suspensions, the collective, large-scale generic behavior of a system that has, therefore, limited our ability to fully take advantage of new single cell analysis approaches to improve the current practice of personalized medicine.

In section 2, we discuss the challenges arising from biological complexity, emergent phenomena, and cell heterogeneity. In section 3, we review efforts to build phenotypes based on flow cytometry data analysis techniques. In section 4, we summarize profiling methods for microscopy image-based screens. In section 5, we present the Supercell/SVM paradigm, which is a general approach for emergent phenotyping that can be applied to different kinds of single-cell datasets, including multicolor flow cytometry and cell imaging. Finally, in section 6, we present our concluding remarks.

2. From molecules to cells to organisms: complexity, emergence, and the cell heterogeneity challenge

Certainly, gaining a molecular understanding of disease is the holy grail of current biomedical research and related interdisciplinary fields. The complexity of this problem, however, is daunting: from a physical perspective, this complexity arises from the plethora of scales, the great diversity of system components, and the sheer size of biological systems. Indeed, a human being is a hierarchically organized, multi-scale system that spans about 10 orders of magnitude in relevant length scales, from water molecules (0.2 nm), DNA molecules (2.5 nm) and proteins (typically 2–10 nm) to cells (10–100 µm), tissues, organs, and organ systems. Besides the multi-scale nature of the problem, complexity arises from the high-dimensional variety of subtypes (e.g. about 300 human cell types and 25 000 protein-coding genes) and the astonishingly large system size (e.g. 3 × 10^9 DNA base pairs contained in each one of the 10–100 trillion cells of a human being).

Physics is familiar with the concept of emergence, that is, the collective, large-scale generic behavior of a system that manifests itself independently of small-scale details [20]. In macroscopic systems, thermodynamics and statistical mechanics provide the theoretical framework to understand macroscopic emergence, which explain large-scale phenomena such as ferromagnetism and irreversibility. At mesoscopic scales, however, other kinds of emergent behavior may arise out of as yet unknown physical laws [21]. Collisions of macroscopic particles in low concentration granular suspensions, for instance, tend to disrupt the inherent reversibility of low Reynolds number fluid flows [22], and the onset of irreversibility decreases approximately as the square of particle density. In dense granular systems under cyclic shear, rearrangements in dense granular suspensions, the collective, large-scale generic behavior of a system that has, therefore, limited our ability to fully take advantage of new single cell analysis approaches to improve the current practice of personalized medicine.
Figure 1. (a) 2D scatter plot of CD8$^+$ T-cell flow cytometry samples from a cohort of seven patients diagnosed with sarcoidosis and six patients diagnosed with Behçet’s disease, two autoimmune diseases with very similar clinical manifestations. Highly overlapping cell populations (shown as a function of the intensity of fluorochrome stains for markers CD3 and IL22) are the hallmark for the phenomenon of cell heterogeneity. (b) Nuclear shapes of diseased ((i) and (ii)) and healthy ((iii) and (iv)) cells can be classified as either blebbed ((i) and (iii)) or non-blebbed ((ii) and (iv)). The diseased nuclei are from Hutchinson–Gilford progeria syndrome (HGPS) cell lines. Scale bar: 10 µm. Adapted from [30]. CCBY.

clinical manifestations. The cells have been stained with 14 fluorochromes; together with forward- and side-scattering, 16 measures have thus been performed on each cell. Despite single-cell flow cytometry data being high dimensional, it is common practice to perform sequential manual partitioning (so-called ‘gating’) of cell events through visual inspection of scatter plots in 2D. Despite the fact that the two markers chosen, CD3 and IL22, are the top markers in a multi-parametric phenotype designed to optimally separate these two diseases (see figure 6(b)), these cell populations are highly overlapping. Similar overlaps are also observed for other cell types and marker pairs. In figure 1(b), images of cell nuclei from a cell line classified as Hutchinson–Gilford progeria syndrome (HGPS), a premature aging disorder, are compared to healthy cells. HGPS is known to be caused by mutant lamin A, which affects the nuclear scaffolding; the phenotypic hallmark of HGPS is nuclear blebbing [31]. Due to cell heterogeneity, however, we observe both blebbed (i) and non-blebbed (ii) diseased cells, and similarly, blebbed (iii) and non-blebbed (iv) normal cells.

There are multiple potential mechanisms for the heterogeneity in cell shape. In addition to the heterogeneity in protein expression levels, molecular mechanisms for intracellular transport are also known to be highly heterogeneous [32, 33]. These transport mechanisms can be broadly classified into two categories: passive diffusion and motor-driven active transport. By performing extensive single-tracking microscopy of endogenous lipid granules in living fission yeast cells, it has been shown that the granules exhibit a crossover behavior from continuous time random walk subdiffusion at short time scales to subdiffusive fractional Brownian motion at longer times [34]. The associated phenomenon of weak ergodicity breaking leads to the observed crossover in the measured mean squared displacements [34]. Very recently, other scenarios based on heterogeneous diffusion, in which the diffusion coefficient depends on the particle position within the cell, have been explored as well showing that they lead to sub- and super-diffusive regimes with weak ergodicity breaking [35–37]. In the case of active transport, a variety of stochastic approaches have been proposed to explain in vivo intracellular observations of tracked biomolecules and microbeads including random walk models, quasi-steady-state reduction methods, exclusion processes, random intermittent search processes, Brownian ratchets, along with mean-field approximations (see [32, 33] and references therein).

The interplay of complexity, emergence, and cell heterogeneity poses a huge challenge to build reliable molecular phenotypes of disease. In the following sections, we briefly summarize current data-driven efforts based on flow cytometry and cell imaging, although some of those approaches may also be applied to other kinds of state-of-the-art and forthcoming single-cell techniques, such as mass cytometry (CyTOF), single-cell gene expression, and single-cell full genome sequencing technologies.

3. Single-cell-based phenotyping using multicolor flow cytometry data

Flow cytometry is used routinely both in research labs to investigate cell structure and function as well as in clinical labs to diagnose disease, assess its progression and its response to therapy. The flow cytometry technique is based on fluorochem-conjugated antibodies that attach to cell surface proteins and intracellular molecules. Laser beams and detectors are used to measure the forward-scattered light (which provides a measurement of the size of the cell), the side-scattered light (which is used to determine the internal structures within the cell) and the emitted light from each type of fluorochrome (which is proportional to the corresponding antigen density). By using compensation techniques to resolve the overlaps of absorption and emission spectra, up to 20 different properties can be determined on a cell-by-cell basis [38, 39].

Although single-cell flow cytometry data is high dimensional, practitioners usually perform sequential manual gating of cell events through visual inspection of scatter plots in
2D, a procedure aided by a number of analysis software programs such as FlowJo [40] and FCS Express [41]. Over the past few years, new computational methods have been developed in order to overcome the serious limitations in manual gating-based analysis [42–49]. All these methods share a common rationale, which is first to define a procedure to extract characteristic features from each sample, and then to perform machine-learning-based classification of different sample types. A comparative assessment of the performance and accuracy of different methods is summarized in [50]. Typically, machine learning instance space is defined by statistical population measures (such as mean, standard deviation and higher moments), histogram classification (through binning into low-dimensional measurement subsets), clustering of events (e.g. k-means or spectral clustering) or mixture models (Gaussian, nonparametric Bayesian, skew-t, etc). Since flow cytometry is high-throughput, with samples typically consisting of hundreds of thousands and even millions of cells, some methods incorporate down-sampling preprocessing steps. In particular, density-based down-sampling methods are designed to filter out denser regions, so that small subpopulations of cells are more significantly represented in the down-sampled datasets than in the original data.

For the purpose of illustration, we summarize here one particular approach that has successfully been applied to classifying acute myeloid leukemia (AML) positive patients and healthy donors using flow cytometry data [46], which is based on the so-called SPADE algorithm [51] that has also proved useful in cell subpopulation analysis of mass-spectroscopy-based cytometry [17]. First, SPADE is applied to perform the extraction of features based on the relative size of subpopulations. SPADE consists of four computational modules: density-dependent down-sampling, agglomerative clustering, minimum-spanning tree construction, and up-sampling. Using data from all samples (both diseased patients and healthy donors), a tree formed by 150 clusters was obtained. It is assumed that groups of clusters roughly correspond to different cell subpopulations. Second, the distribution of cells across the 150 clusters is computed for each sample. Figure 2(a) shows the cell distribution across the SPADE tree for a healthy donor, while figure 2(b) displays the cell distribution for a diseased patient. Third, the pairwise Earth mover’s distance (EMD) between cell distributions is calculated. EMD measures the minimum effort needed to make one distribution the same as the other by moving cells, which can be calculated as
Figure 3. Multiparameter automated feature extraction from nuclei imaging. In this example, after variations in illumination and staining are corrected, nuclei are identified by thresholding and then used as seeds to identify cell edges. Finally, markers are used to identify DNA-damage foci. Multiparametric quantitative single-cell features may then result from transformations of the image-based measurements. From [57].

4. Single-cell-based phenotyping using microscopy images

The field of biological image analysis has seen rapid progress with the advent of automated image acquisition systems that enable new types of microscopy experiments. A variety of imaging techniques yield high-content, high-throughput images from fluorescence microscopy, electron microscopy, bright-field microscopy, differential-interference-contrast microscopy, as well as from multi-spectral and multi-dimensional microscopy. Yet, the bottleneck for progress in this field continues by extracting image content descriptors, see figure 4. Feature extraction algorithms read image pixels from foreground ROIs and output a variety of numerical image features that characterize the assay, including the number of objects per image (e.g. the number of DNA-damage-induced foci per cell in the example from figure 3), the intensity of a marker as reflected by pixel intensity in a given channel (e.g. the total intensity of DNA marker in the nucleus can be used to identify cell-cycle phases), texture descriptors (such as statistical, structural, and spectral texture measures), and shape descriptors that enable morphological phenotyping (e.g. Zernicke shape features [74] and Hu moments [75]). In some applications, image features are extracted from arbitrarily defined tiles that do not necessarily correlate with actual cellular or subcellular features, which is the kind of approach used in algorithms such as PhenoRipper [56] and WND-CHARM [76].

Finally, phenotyping healthy versus diseased samples, or chemically and genetically perturbed samples, is usually achieved by implementing a machine learning classifier. As
shown schematically in Figure 4, this process is iterative. The initial set of features usually ranges between few tens and one hundred, although it is assumed that many of them are irrelevant or redundant. Two general approaches to feature selection are known as filters and wrappers [77]. Filtering methods are based on statistical measures before the classifiers are applied. For instance, Fisher scores, which measure inter-class variance relative to intra-class variance, can be used to weight features according to their classification discriminative power [76]. Another filtering approach, which aims at avoiding the bias due to highly correlated features, is the so-called minimum redundancy maximum relevance (mRMR) algorithm [78]. Other filtering methods are based on remapping the original feature space into lower dimensional ones, including linear approaches such as singular value decomposition [79–81] and principal component analysis [82, 83], as well as non-linear ones such as self-organizing maps [84, 85], isomap [86], local linear embedding [87] and multidimensional scaling [88, 89]. Wrapping approaches, in turn, select features based on their actual performance in the classifier (see e.g. [77, 90, 91]); they oftentimes provide better feature selection and greater classification accuracy than filtering, although in most cases wrapping is significantly slower because it relies on running many iterations with different feature sets (as indicated by the dashed upwards arrow in Figure 4).

A number of proprietary and open-source software tools are available to perform specific image analysis tasks such as segmentation, feature selection and classification. Moreover, some packages integrate all the tasks needed along the image analysis workflow. Software packages such as WND-CHARM [76], CellProfiler [92], ImageJ [93], PSLID [94], Ilastik [95], and BIOCAT [96] offer diverse capabilities in terms of graphic user interface, 3D-image analysis, machine learning classifiers, extensible algorithm plugins, ROI detection, automatic comparison among algorithms, commercial software requirements, and OS platform requirements.

5. Emergent phenotyping via the Supercell/SVM paradigm

Flow-cytometry-based phenotyping methods (see section 3 and references therein) usually rely on the fact that typical cytometry experiments are able to perform multidimensional single-cell measurements on tens and hundreds of thousands, if not even millions, of cells. Therefore, single-cell-level heterogeneities such as those described in section 2 are averaged out by means of large population averages. High-content image-based phenotyping methods (see section 4 and references therein), in turn, usually focus on pattern recognition and classification of regions of interest, but typically disregard cell-to-cell heterogeneities within a given phenotype.

Heterogeneity from cell to cell, however, is now recognized in many cases not as noise, but as an important feature of many living systems that enables adaptation to changing environmental conditions for a heterogeneous ensemble, from bacterial persistence under antibiotic treatment [97, 98] to plasticity and functional diversity in the human brain [99]. Single-cell heterogeneity is often encountered in biomedical basic and translational research, as well as in the clinical realm, and leads to particular challenges for studies that are based on a limited number of single cells. There are two overall constraints that limit the number of analyzed cells. In some diseases, the number of cells that can be collected is limited; for instance, diagnosis of intraocular lymphoma is...
Figure 5. Analysis of nuclear shapes from images obtained from two normal and two HGPS cell lines. Probability density distributions for the number of invaginations of the nuclear boundary corresponding to: (a) single cells and (b) supercells of size $N = 30$. Distances to the SVM boundary by applying machine learning to (c) single cells and (d) supercells of size $N = 30$, where 15 parameters per nucleus are used to characterize shape features and concentration levels of lamin A/C. Adapted from [30] CCBY.

performed through difficult specimen collection procedures that yield very low numbers of collected cells, typically in the range of ten to a hundred cells [100]. In the field of stem cell research, identifying stem cell phenotypes through flow cytometry cell sorting often relies on extremely rare subpopulations. For instance, long-term hematopoietic stem cells (LT-HSCs) identified via immunophenotypes such as Lin$^{-}$Kit$^{+}$Sca$^{+}$CD34$^{hi}$Flt3$^{-}$ [101] and SLAM [102] represent only about $0.0075\%$ of the cells from whole bone marrow specimens; thus, more than a million whole bone marrow cells need to be extracted, stained with multiple fluorochromes and sorted in order to yield about one hundred LT-HSCs. In early embryonic development studies, e.g. at the stage of mouse early development when the founder population of germline cells have just emerged, there are just about 30 primordial germ cells in the embryo [103]. The second constraint is due to the analysis approach, where more sophisticated analysis with a larger number of parameters (i.e. higher dimensional data) usually are accompanied by lower numbers of cells analyzed. Image-based screens are typically based on tens to hundreds of cells, with higher dimensional analyses (e.g. 3D cell shapes, multi-spectral images) often based on fewer cells than simpler lower dimensional datasets. State-of-the-art single-cell genomics technologies are an extreme example of this trend. It is now possible to measure the expression level of all $20\,000+$ genes in a single cell [104], but the number of cells for which all genes can be measured is limited by both cost and instrument capacity. For these new high-dimensional data with limited numbers of datapoints, data analysis methods that rely on high-dimensional clustering procedures, Gaussian mixture approximations, etc may be expected to fail.

Therefore, it is of paramount importance to address the problem of phenotypic classification when single cells are highly heterogeneous and the number of cells available is small. Within this context, we have recently proposed the so-called ‘Supercell/SVM Paradigm’ [30] as a general method for single-cell phenotyping that focuses on mesoscopic-level emergent properties of groups of cells. The key contribution of this method is to provide a quantitative assessment of the critical sample size and number of simultaneous single-cell measurements needed to identify a phenotype with strong predictive power.

In order to capture emergent behavior, a ‘supercell of size $N$’ is defined as the average of the individual measurement vectors of a group of $N$ randomly chosen cells. By repeatedly taking different random subsets of $N$ cells, ‘supercell samples’ can be built out of the original single-cell datasets; on these samples, ‘supercell statistics’ can be readily computed. This procedure is illustrated in figures 5(a) and (b), where probability density distributions are shown
for one shape parameter (the number of invaginations of the nuclear boundary) obtained from images from two healthy and two HGPS cell lines. As mentioned above, HGPS is a premature aging disorder caused by mutant lamin A that affects the nuclear scaffolding; although HGPS is characterized by nuclear blebbing, phenotypes at the single-cell level are highly heterogeneous (recall figure 1(b)). The single-cell distributions in figure 5(a) are highly overlapping, reflecting the fact that, based on individual cells, one is not able to distinguish healthy cells from diseased ones. After applying the cell averaging procedure (using $N = 30$ randomly selected cells to generate each supercell), the resulting distributions do not show any significant overlap between healthy and diseased samples (see figure 5(b)). Indeed, the removal of distribution overlaps is a manifestation of the central limit theorem, by which distributions of supercells of size $N$ are expected to become narrower by $\sim 1/\sqrt{N}$. After cell averaging, machine learning is used to learn what combination of parameters best distinguishes healthy from diseased cells. The method implemented in [30] is a support vector machine (SVM) with a linear kernel, which is less prone to overfitting issues than non-linear mappings [105–107]. Moreover, the components of the vector normal to the boundary hyperplane can be straightforwardly interpreted as amplitudes that determine the relative significance of the measured parameters in achieving class separation. Healthy and HGPS nuclear shapes were characterized by 12 parameters including minor/major axis length, perimeter, area, tortuosity, mean and standard deviation of the curvature, and the number of invaginations of the nuclear boundary. Moreover, three fluorescence intensity measurements associated with the concentration of lamin A/C were measured on each nucleus, thus contributing three additional parameters. However, for single cells, even 15 parameters do not suffice to learn the distinction between healthy and diseased individual cells. Indeed, figure 5(c) shows the distance from each cell to the SVM boundary, where positive (negative) distances correspond to the boundary side identified with the healthy (diseased) class: it is observed that some cells from the healthy cell lines are classified as diseased, and vice versa. Instead, by applying machine learning to supercell distributions, 100% classification accuracy is achieved, as shown in figure 5(d).

In the analysis of HGPS through nuclear shape measurements, the Supercell/SVM method determined that about 30 cells are needed to classify samples as healthy or diseased, in agreement with usual lab procedures. Besides providing validation of current wet-bench practice, the supercell method lays out a straightforward framework towards determining how many cells should be measured in novel scenarios (e.g. drug testing, gene-knockout phenotypes, etc) where carrying over ‘tried and tested’ rule-of-thumb lab procedures may be inappropriate.

Let us consider next a more challenging problem, namely that of building flow-cytometry-based molecular phenotypes for two non-infectious uveitides (the ocular manifestations of sarcoidosis and Behçet’s disease), which are very difficult to diagnose. By performing two scattering and 14 fluorescent measurements on each cell, samples from seven sarcoidosis and six Behçet’s patients were measured. Since the cohort was small, prediction testing was carried out by a jackknife (leave-one-out) cross-validation procedure [108]. The SVM boundary allows one to rank-order the 16 measures from most to least significant, according to the components of the vector normal to the hyperplane that separates the two diseases. Thus, one can selectively remove the least significant measurements from the list and explore the minimal number of measures needed to correctly predict the class of all (or at least most of) the samples. Figure 6 shows jackknife results for supercells of size $N = 500$, where each patient is represented by a cloud of 100 supercells, as a function of the number of rank-ordered measures used. The list of the top 10 measures is shown to the right of each panel. Figure 6(a) shows jackknife results based on all cells, while figure 6(b) displays results based on CD8<sup>+</sup> T cells, a subpopulation that can be determined by manual gating (CD3<sup>+</sup>viab<sup>-</sup>CD8<sup>+</sup>CD4<sup>-</sup>) and typically represents about 5% of the peripheral blood sample. Since each patient is represented by a cloud of supercells, a prediction was made only when more than 95% of those supercells lie on any one side of the SVM boundary. Correct predictions are shown by green bars, incorrect predictions by red bars, while unclassified samples are shown in blue. While predictions based on all cells are very poor, for CD8<sup>+</sup> T cells no failed predictions are incurred when five or more measures are used. Therefore, the top five measures listed in figure 6(b),
if linearly combined as 29% IL22 + 19% CD3 + 18% viab + 17.5% CD8 - 16.5% CD62L, can be used on CD8+ T cells as molecular phenotypes that distinguish the two diseases.

In figure 7(a), the percentage of correctly classified CD8+ T supercells is shown as a function of the supercell size and the number of measures used. It turns out that, by using the top five markers, slightly less than 100 cells are sufficient for reliable prediction. Increasing the number of markers or averaging over more cells, however, do not significantly change the reliability of the classification. The Supercell/SVM framework, thus, allows one to determine which and how many measures, as well as how many cells, need to be measured. Because adding markers and measuring larger samples are expensive and time consuming (and, as mentioned above, might be strongly constrained by experimental and clinical considerations), the ability to determine the minimal setup required to build reliable phenotypes is of paramount importance. Finally, figure 7(b) presents a novel visualization based on the usual 2D scatter plots used in flow cytometry to find cell subpopulations (recall figure 1(a)). By allowing linear combinations of markers in the vertical axis, a clear separation between disease classes can be readily observed.

6. Conclusions

By complementing more traditional approaches based on computational (and, to a lesser extent, analytical) modeling, data-driven research is opening up new avenues of investigation fueled by the advent of powerful high-throughput biomedical technologies. Within that context, this review focuses on current efforts towards developing quantitative, unbiased, mathematically robust methods to learn healthy and diseased phenotypes from multidimensional, single-cell biomedical data.

In order to understand health and disease with multidimensional single-cell methods, one needs to overcome the challenges arising from the interplay of complexity, emergence, and cell heterogeneity. Analysis methods based on multicolor flow cytometry usually rely on rich datasets typically comprising hundreds of thousands (sometimes even millions) of cells. Under such high-throughput conditions, single-cell-level heterogeneities are often simply averaged out, thus allowing population-level signatures to emerge. A number of machine-learning-based classification methods have been developed recently; they define different kinds of instance spaces, on which machine learning classifiers are trained and then used to predict on new samples. For a comparative assessment of the performance and accuracy of different automated flow cytometry data analysis techniques, see [50].

High-content image-based phenotyping methods, in turn, usually focus on pattern recognition and classification of regions of interest. In some cases, however, image classification and phenotyping relies on pixel tiles that do not correlate with biologically relevant measures such as cell morphology and biomarker intensity. Moreover, pattern recognition methods are focused on pattern classification as opposed to phenotype classification, in the sense that they typically disregard cell-to-cell heterogeneities within a given phenotype. For reviews of image-based approaches, see [54, 57, 73].

In this review, we emphasize the so-called Supercell/SVM paradigm (recently developed by the authors of this review alongside collaborators, see [30]) as a unified framework that captures mesoscopic-scale emergence to build reliable phenotypes. Supercells are multidimensional objects that represent the collective behavior of groups of cells; within this approach, supercells represent the building blocks of healthy and diseased phenotypes. From a conceptual standpoint, this approach naturally incorporates emergent behavior and thus cell heterogeneity, usually regarded as a roadblock in the pursuit of characterizing single-cell-level behavior, becomes the fundamental conceptual unit to identify collective phenotypes. From a practical perspective, furthermore, the Supercell/SVM framework provides a quantitative assessment of the critical sample size and the number of simultaneous single-cell measurements needed to build a phenotype, which is a key piece of information given the fact that, in many single-cell applications, the number of measured cells and the number of measurements per cell are severely limited due to a variety of constraints, such as experimental costs, technological capabilities, specimen collection procedures, the availability of specialized personnel, and others.

Finally, let us point out the importance of interdisciplinary exchanges of methods and ideas such as these, in which tools and concepts from statistical physics, data mining, and mathematics are successfully applied to the most pressing problems in the life sciences. We hope that the powerful synergy of collaborative efforts across disciplines grows ever stronger and bears fruit by improving diagnosis, prognosis, and decision making in the clinical realm.
References

[1] Beckman R A, Schemmann G S and Yeang C-H 2012 Proc. Natl Acad. Sci. USA 109 14586
[2] De Souza N 2012 Nature Methods 9 645
[3] Tang D G 2012 Cell Res. 22 457
[4] Drukker M et al 2012 Nature Biotechnol. 30 531
[5] Schmid J et al 2012 J. Biol. Chem. 287 41546
[6] Chakrabarti L et al 2012 Front. Oncol. 2 82
[7] Fujita M and Onami S 2012 PLoS ONE 7 e30224
[8] Meacham C E and Morrison S J 2013 Nature 501 328
[9] Burrell R A, McGranahan N, Bartek J and Swanton C 2013 Nature 501 338
[10] Juntila M R and de Sauvage F J 2013 Nature 501 346
[11] Bedard P L, Hansen A R, Ratain M J and Siu L L 2013 Nature 501 355
[12] Perfetto S P, Chattopadhyay P K and Roederer M 2004 Nat. Rev. Immunol. 4 648
[13] Sachs K et al 2005 Science 308 523
[14] Gonzalez V D et al 2008 J. Immunol. Methods 330 64
[15] Cerignoli F et al 2012 J. Pharmacol. Toxicol. Methods S1056-8719 00262
[16] Ornatsky O et al 2010 J. Immunol. Methods 361 1
[17] Bendall S C et al 2011 Science 332 867
[18] Benoist C and Hacohen N 2011 Science 332 677
[19] Candia J 2013 APS March Meeting http://meetings.aps.org/link/BAPS.2013.MAR.U44.8
[20] Anderson P W 1972 Science 177 393
[21] Laughlin R B et al 2000 Proc. Nat Acad. Sci. USA 97 32
[22] Corté L et al 2008 Nature Phys. 4 420
[23] Slotterback S et al 2012 Phys. Rev. E 85 021309
[24] Klein M L and Shinozawa W 2008 Science 321 798
[25] Shinoda W, Devane R and Klein M L 2012 Curr. Opin. Struct. Biol. 22 175
[26] Wilhelm J and Frey E 2003 Phys. Rev. Lett. 91 108103
[27] Bausch A R and Kroy K 2006 Nat. Phys. 2 231
[28] Levchenko A 2013 APS March Meeting http://meetings.aps.org/link/BAPS.2013.MAR.UF44.8
[29] Altschuler S J and Wu L F 2010 Cell 141 559
[30] Candia J et al 2013 PLoS Comput. Biol. 9 e1003215
[31] Driscoll M K et al 2012 Aging 4 119
[32] Höfling F and Franosch T 2013 Rep. Prog. Phys. 76 046602
[33] Bressloff P C and Newby J M 2013 Rev. Mod. Phys. 85 135
[34] Jean J-H, Tejedor V, Burov S, Barkai E, Selhuber-Unkel C, Berg-Sorensen K, Oddershede L and Metzler R 2011 Phys. Rev. Lett. 106 048103
[35] Cherstvy A G, Chechkin A V and Metzler R 2013 New J. Phys. 15 083039
[36] Cherstvy A G and Metzler R 2013 Phys. Chem. Chem. Phys. 15 20220
[37] Cherstvy A G, Chechkin A V and Metzler R 2013 (arXiv:1311.2810)
[38] Baumgarth N and Roederer M 2000 J. Immunol. Methods 243 77
[39] Lugli E, Roederer M and Cossarizza A 2010 Cytometry A 77 705
[40] Proprietary software available at www.flowjo.com
[41] Proprietary software available at www.denovosoftware.com
[42] Hartung F et al 2009 BMC Bioinformatics 10 106
[43] Naim I et al 2010 IEEE Acoustics Speech and Signal Processing p 509
[44] Qian Y et al 2010 Cytometry B 78 S69
[45] Aghaeiopour N et al 2012 Cytometry A 81A 1022
[46] Qiu P 2012 PLoS ONE 7 e37038
[47] Ge Y and Seafon S C 2012 Bioinformatics 28 2052
[48] Aghaeiopour N et al 2012 Bioinformatics 28 1009
[49] Zare H et al 2012 Am. J. Clin. Path. 137 75
[50] Aghaeiopour N et al 2013 Nature Methods 10 228
[51] Qiu P et al 2011 Nature Biotechnol. 29 886
[52] Rubner Y, Tomasi C and Guibas L 2000 Int. J. Comput. Vis. 40 99
[53] Kira K and Rendell L 1992 ML92: Proc. 9th Int. Workshop on Machine Learning (San Mateo, CA: Morgan Kaufmann) p 249
[54] Peng H 2008 Bioinformatics 24 1827
[55] Shamir L, Eckley D M, Delaney J, Orlov N and Goldberg I G 2009 IEEE NIH Life Sci. Syst. Appl. Workshop pp 96–100
[56] Rajaram S, Pavie B, Wu L F and Altschuler S J 2012 Nature Methods 9 635
[57] Li J, Pappalardo G and Garlaschelli D 2011 Bioinformatics 19 2341
[58] Peng H, Ruan Z, Long F, Simpson J H and Myers E W 2010 Nature Biotechnol. 28 348
[59] Otsu N 1979 IEEE Trans. Syst. Man Cybern. 9 62
[60] Sahoo P K, Soltani S, Wong A K and Chen Y C 1988 IEEE Trans. Syst. Man Cybern. 18 621
[61] Lindblad J and Bengtsson E 2001 Proc. 12th Scand. Conf. on Image Analysis pp 264–71
[62] de So´lizano C O et al 1999 J. Microsc. 193 212
[63] Bailey T, Arp C, Amin N and Linssen B 2006 Cytometry A 69A 652
[64] Meyer F and Beucher S 1990 J. Vis. Commun. Image Represent. 1 21
[65] Vincent L and Soille P 1991 IEEE Trans. Pattern Anal. Mach. Intell. 13 583
[66] Lindeberg T 1993 Int. J. Comput. Vis. 11 283
[67] Roerdink B T M and Meijster A 2001 Fundam. Inform. 41 187
[68] Cong G and Parvin B 2000 Pattern Recognit. 33 1383
[69] Duan G, Cao B, Lu H and Lu J 2008 Med. Imaging Inform. pp 80–8
[70] Vromen J and McCane B 2006 8th Proc. Ann. Colloq. Spatial Inform. Res. Ctr. pp 25–9
[71] Driscoll M K et al 2012 PLoS Comput. Biol. 8 e1002392
[72] Li G et al 2008 J. Microsc. 231 47
[73] Shamir L, Delaney J D, Orlov N, Eckley D M and Goldberg I G 2010 PLoS Comput. Biol. 6 e1000974
[74] Boland M V, Markey M K and Murphy R F 1998 Cytometry 3 366
[75] Hu M K 1962 IRE Trans. Inform. Theory IT-8 179–87
[76] Orlov N et al 2008 Pattern Recognit. Lett. 29 1684
[77] Guyon I, Gunn S, Nikravesh M and Zadeh L (ed) 2006 Feature Extraction, Foundations and Applications (New York: Springer)
