Single Cell Proteomics in Biomedicine: High-dimensional Data Acquisition, Visualization and Analysis

Yapeng Su[1,2], Qihui Shi[3]*, and Wei Wei[1,4]*

[1]NanoSystems Biology Cancer Center;

[2]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, 91125, USA;

[3]Key Laboratory of Systems Biomedicine (Ministry of Education), School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China.

[4]Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA;

*Correspondence and requests for materials should be addressed to W.W. (weiwei@mednet.ucla.edu) or Q.H.S. (qihuishi@sjtu.edu.cn)

Keywords:

Information theoretical approaches/ Mass cytometry / Single cell barcode chip / Single cell data analysis / Single cell proteomics

Abbreviations:

ELISPOT, Enzyme-Linked ImmunoSpot; **SiMoA**, single molecule array; **SCBC**, single cell barcode chip; **scWesterns**, single cell western blot; **FFC**, fluorescence flow cytometry; **CyTOF**, cytometry by time-

Received: 31/10/2016; Revised: 20/01/2017; Accepted: 20/01/2017
This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201600267.

of-flight; **ELISA**, enzyme-linked immunosorbent assay; **SPADE**, spanning tree progression of density normalized events; **CLARA**, clustering for large applications; Citrus, cluster identification, characterization and regression; **mTORC1**, mechanistic target of rapamycin complex 1; **viSNE**, visualization of t-distributed stochastic neighbor embedding; **One-SENSE**, one-dimensional soli-expression by nonlinear stochastic embedding; **ACCENSE**, automatic classification of cellular expression by nonlinear stochastic embedding; **SCUBA**, single-cell clustering using bifurcation analysis; **HIF-1α**, hypoxia-inducible factor; **GBM**, glioblastoma; **DREMI**, conditional-density resampled estimate of mutual information**; DREVI**, conditional-density rescaled visualization; **GFP**, green fluorescent protein;

Abstract

New insights on cellular heterogeneity in the last decade provoke the development of a variety of single cell omics tools at a lightning pace. The resultant high-dimensional single cell data generated by these tools require new theoretical approaches and analytical algorithms for effective visualization and interpretation. In this review, we briefly survey the state-of-the-art single cell proteomic tools with a particular focus on data acquisition and quantification, followed by an elaboration of a number of statistical and computational approaches developed to date for dissecting the high-dimensional single cell data. The underlying assumptions, unique features and limitations of the analytical methods with the designated biological questions they seek to answer will be discussed. Particular attention will be given to those information theoretical approaches that are anchored in a set of first principles of physics and can yield detailed (and often surprising) predictions.

1. Introduction

The flourish of single cell technology in the last decade has led to increased recognition of cellular heterogeneity as a universal feature of any cell population [1]. The improved understanding of the cause and consequence of such heterogeneity further drives the research community to develop analytic approaches by which multiple molecular landscapes of cellular processes can be measured simultaneously at single cell resolution, inaugurating the multi-omics age of single cell biology and allowing researchers to ask questions from perspectives previously unattainable. In principle, one wants to know, for each single cell, the molecular code of the cell (the genome), the functionality of that cell (the proteome and metabolome), and the connection between the two – the transcriptome. This requires single cell discovery science that extends from genomics to biological function. Recent technological advances have brought a suite of single cell toolkits that permit robust and high-throughput quantitation of the genome, transcriptome, proteome and metabolome at single cell level [2]. Tools for integrated measurements of multiple classes of biomolecules simultaneously from the same single cells have also been demonstrated. Such measurements offer unprecedented resolution to the diversity of cellular states in a given tissue and enable detailed investigations of cellular lineage, intracellular signaling network, cellular function, and the role of significant cellular subpopulations or rare cell types. The simultaneous profiling of a profusion of cellular processes provides a wholly different kind of insight, revolutionizing our holistic view on the complex cellular system.

As key executors of biological processes – the functional proteins connect genomic information to biological functions [3]. A variety of single cell proteomic tools have been developed for assaying different types of functional proteins, including cytokines, growth factors, signaling phosphoproteins, transcriptional factors, etc., with increasing multiplexing and throughput. Features and technical details of these tools been extensively reviewed elsewhere [2-7]. However, the methods for in-depth analysis of high-dimensional single cell proteomic data are less matured than the experimental platforms. Many analytic approaches for visualizing and understanding these large

datasets are still subjective, labor intensive and varying across research groups, which poses a major challenging for effectively gleaning useful biological insights from the measurements. In this review, we briefly survey the single cell proteomics tools and their applications in biomedicine, with a focus on data acquisition and the degree of quantification of each tool. With the technical foundation established, we'll turn our attention to the data analysis and elaborate a number of statistical and computational approaches developed to date for visualizing and analyzing the high-dimensional single cell data. The underlying principles, unique features and limitations of the approaches with the designated biological questions they seek to answer will be covered. We will pay particular attention to the information theoretical approaches anchored in a set of first principles of physics as they can yield detailed (and often surprising) predictions.

2. Single cell proteomic tools with varying degrees of quantification

Single cell proteomic tools can be categorized into two complementary types: measuring a large number of parameters across thousands of single cells at a given time point (snapshot), or monitoring a handful of parameters in the same cells over time [8]. Remarkable advances have been made for the tools in the first category with the emergence of highly multiplex mass cytometry and microchip-based platforms in the last few years. Therefore, we start with a brief review on the high-dimensional population snapshot tools below.

The characteristics of a single-cell proteomic assay include multiplexing capacity, throughput, sensitivity and dynamic range. Multiplexing capacity determines the number of proteins assayed in a single cell measurement and throughput dictates the number of cells analyzed in parallel. Based upon the nature of the reported results, the single-cell proteomics assays are assorted into three classes, qualitative methods that qualitatively identify cells that express a given proteins (*e.g.,* ELISPOT), semi-quantitative methods that measure protein abundance in relative units (*e.g.*, flow/mass cytometry, image cytometry, single cell western blot) and quantitative methods in which

calibration curves can be established to translate analytical signals into protein concentration or even copy numbers (*e.g.*, SiMoA, Microengraving chip, SCBC).

2.1 Qualitative methods

Qualitative methods are utilized to differentiate positive and negative expression of target proteins in cells. The enzyme-linked immunospot (ELISPOT) assay is a qualitative assay for detecting cytokine secretion at single cell level. Typically, immune cells are localized on an antibody-coated surface, followed by cytokine secretion upon stimulation. Secreted proteins are captured by the immobilized antibodies in the vicinity of individual cells, and then detected by secondary antibody with enzyme amplification for signal readout. The numbers of spots are measured to evaluate the frequency of cytokine-secreting cells for monitoring immune system activation. ELISPOT is highly sensitive for detection of secreted proteins, but is colorimetrically limited to detect only 1-3 cytokines simultaneously [9].

2.2 Semi-quantitative methods

Semi-quantitative methods measure protein abundance in relative units. Fluorescence flow cytometry (FFC) is the most established method for single cell protein analysis. With fluorophore-labeled antibodies, it can analyze primarily, membrane and cytoplasmic proteins associated with signaling pathways underlying many diseases in millions of single cells at a moderate level of multiplexing (<15 proteins) [10-13]. Mass cytometry (CyTOF) extends the concept of flow cytometry to assay more than 30 proteins through the use of antibodies that are tagged with transitional metal mass labels rather than fluorophore labels (Fig. 1A) [14]. For measuring secreted cytokines, both FFC and CyTOF require first blocking protein secretion and then fixing and making permeable the cells to allow for perfusion of dye-labeled antibodies. Blocking cytokine secretion constitutes a significant perturbation to the cells and the level of 'secrete-able' cytokines may not faithfully recapitulate the functional measurement of cellular secretion. Droplet-based microfluidic flow cytometry alleviates

this concern via encapsulating single cells and cytokine-capture beads in droplets, enabling the measurement of proteins secreted by single cells [15, 16]. The protein levels measured by FFC or CyTOF are in arbitrary unit, depending on the instrument settings used in taking the measurement. Because each instrument has a characteristic efficiency prolife, comparison of data across instruments is challenging and requires sophisticated normalization [17]. Even when consistent settings are used, variability in instrument performance makes comparison between datasets acquired on different days uncertain unless the instrument is calibrated. The calibration can be performed by running a sample of calibration beads to normalize multiple datasets for comparison. In addition, calibration beads coated with known and increasing numbers of IgG are utilized to mimic the binding of specific monoclonal antibodies to surface proteins. These beads allow generation of a calibration curve relating mean fluorescence intensity to the number of target proteins assayed [18]. However, such quantitation might not be reliable due to the differences between cells and beads. Meanwhile, cytoplasmic proteins require intracellular staining and thus fail to be calibrated to determine the number of protein molecules expressed per cell due to a lack of calibration method.

Image cytometry based on cell staining typically assay 3-4 membrane or intracellular proteins per cell because of the spectral overlap of fluorophore-labeled antibodies. Multiple cycles of staining and de-staining enable measurement of more than 20 proteins simultaneously [19]. Similar to flow cytometry, image cytometry also has the difficulty in calibration of membrane and cytoplasmic proteins and fails to relate fluorescence intensities to protein copy number. A variant of image cytometry is to label antibodies with photocleavable DNA barcodes in replace of fluorophores. Each antibody has a unique sequence label [20]. After antibody binding to the proteins within the cells, the photocleavable linkers are broken upon UV radiation and release the unique DNA barcodes that are detected by hybridizing to fluorescent complementary array for quantification [21]. However, antibodies in this detection scheme are not conjugated to a fixed number of DNA molecules. The

efficiency of DNA barcode release is also subject to variation. These factors attribute to difficulty in a reliable calibration.

Compared with other single-cell proteomic methods, single cell western blotting (scWestern) developed by the Herr group [22, 23] overcomes the antibody cross-reactivity because proteins are first separated by molecular mass (via electrophoresis) before the antibody probing step, thereby enabling clear discrimination between on-target and off-target signals. In scWesterns, a photoactive polyacrylamide gel is coated on a microscope slide and aligned with an array of open-microwells for cell lysis *in situ*, gel electrophoresis, photoinitiated blotting to immobilize proteins and antibody probes.[10] scWestern has been reported to exhibit a linear dynamic range of 1.3-2.2 orders and detection thresholds of ~27,000 molecules. However, it is subject to many variables in the assay and therefore difficult to be calibrated for quantitative measurements.

2.3 Quantitative methods

Quantitative methods report copy number of target proteins by directly counting or establishing calibration curves. Enzyme-linked immunosorbent assay (ELISA) is the most widely used quantitative protein assay in clinic, relying on calibration curves to transform fluorescence signals to the protein concentration. With 'spectral addresses' defined by distinct proportions of red and near-infrared fluorophores in the microbeads, Luminex xMAP (Multi-Analyte Profiling) platform utilizes a bead-based ELISA-like assay to significantly increase the multiplex level of protein detection in a very small sample volume, yet not to the resolution of single cells.

A handful of microfluidics-based single cell proteomics tools, as exemplified by single cell barcode chip (SCBC) [24-26] developed by the Heath group (Fig. 1B) and microengraved chip [27, 28] developed by the Love group, miniaturize an array of ELISA to surface-based immunoassays in microchip devices, leading to quantitative, multiplexed protein detection in single cells. SCBCs isolate single cells, or defined number of cells, into microchambers that each contains a many-element

antibody array (the barcode). Depending on the application, a few hundred to ten thousand [29, 30] individual microchambers with volumes between 0.1 and 2nL are included within a single chip. Spatially encoded antibody barcodes [31] in SCBCs enable simultaneous quantitation of more than 40 secreted, intracellular and membrane proteins from single cells [32]. An on-chip calibration curve with standard proteins transforms fluorescence readouts to the protein concentration, leading to the absolute quantitation in copy number of molecules detected based on the known volume of assaying microchambers. A caveat for such calibration is that recombinant standards may not always be commercially available or may be modified from the corresponding protein produced within the cells (Fig. 1C). Reporting copy number of target proteins in SCBCs enables direct comparison across platforms, cell types, time points, clinical samples, and so on, allowing clinical studies or investigations in which statistical cell behaviors are compared across a perturbation series [25, 33]. Such calibrations are tough to do using cytometry-based approaches.

Single molecule array (SiMoA) detects proteins with single molecule resolution and thereby leads to absolute quantification. SiMoA employs a large number of antibody-coated beads to capture small amount of proteins, which results in single molecules captured on the beads. Sandwich-type immunoassay with enzyme amplification is utilized for signal readout of single molecules. Serum and other biofluids have been investigated by SiMoA to demonstrate ultra-low detection limits and a large dynamic range compared to traditional ELISA [34]. The variation of prostate specific antigen across single prostate cancer cells have been interrogated with SiMoA to reveal the expression shifts with genetic drift measured [35]. However, SiMoA is limited by low multiplexing capacity, low throughput and high cost for single cell measurement.

A recent publication raised serious concerns about the quality of the commercial antibodies [36, 37]. In their experiments, only 452 antibodies out of the 1124 tested recognized their intended antigen in HEK293 cell lysate. Given this large caveat, the use of antibodies for staining (as with FFC or CyTOF) is very different from their use in fluorescent sandwich immunoassays in microfluidic single cell

chips. For the latter one, each individual protein assay provides two separate measurements per cell (since two antibodies per protein are used) to ensure the specificity. Each individual assay can also be compared against every other assay in the panel for eliminating cross-reactivity. Importantly, a careful analysis based on experiments and stimulation was conducted to evaluate the technical error of the SCBC which is around 5-10%, enabling determination of contributions from biological variation versus technical error [24, 38].

3. Descriptive statistical approaches for visualizing and analyzing single cell proteomic data

Rapid progress in single cell proteomic technologies empowers people to measure more and more parameters from each individual cells. In principle, with more measurements from each single cell, we should be able to gain more comprehensive understanding of the heterogeneous system that we are interested in. However, the power of those advanced technologies are, often times, not yet fully exploited. This is, to a great extent, due to the so-called "curse of dimensionality" [6]: visualizing and understanding these large, high-dimensional datasets poses a major analytical challenge. Various approaches have come into being with the purpose of assisting us to identify the subpopulations, discern the overall data structure and resolve the dynamic changes (Table 1). This in turn helps us to obtain deeper understanding of the high-dimensional dataset for making precise and testable predictions regarding how the heterogeneous system behaves.

3.1 Clustering-based analytical methods for identifying biologically meaningful subsets.

One of the main reasons for using single cell technology to study biological systems is because of their heterogeneous nature. The existence of multiple phenotypic and functional subpopulations is common in many cellular systems, despite the fact that cells have identical genomic sequences. For example, the human peripheral blood mononuclear cells (PBMCs) contain a diverse array of lymphocytes (T cells, B cells, NK cells, *etc.*) and monocytes where different cell types behave

differently from one to another. Partitioning the high-dimensional single cell data into biologically meaningful subpopulations is the major task for clustering-based algorithms.

The traditional method for looking at cell subtypes is manual gating [39]. A region of interest in a biaxial plot of two protein markers is used to select desired subpopulations for further analysis of other markers. The entire process of gating is carried out through a series of biaxial plots, which renders it extremely burdensome when a large number of proteins are measured simultaneously for each cell. In addition, manual gating requires extensive prior knowledge of cellular system under study. Therefore, it is mostly used in analyzing immune cell phenotypes with known surface marker combinations.

To efficiently analyze single cell proteomics data with increasing dimensionality, a cohort of unsupervised data-driven clustering methods have emerged recently [40-44]. Among them, SPADE (spanning tree progression of density normalized events), as a popular one, utilizes density-based algorithm to define cellular clusters and displays the underlying phenotypic hierarchy in a tree-like structure (Fig. 2A) [40]. It is especially useful for cellular hierarchy inference among subpopulations of similar cells. SPADE first performs a density-dependent down sampling followed with agglomerative clustering to group similar cells into subgroups. Each subgroup is represented as a node with a designated size that is promotional to its density. Then the algorithm connects subgroups together in a minimum-spanning tree where each node is connected to its two nearest neighbors while minimizing the total edge length. Finally, an up-sampling is performed to recapture the original density [40]. SPADE therefore enables visualization of the high-dimensional single cell data in a branched tree structure in one planar image without a predefined cellular ordering. While the stochastic nature of the density-based down sampling prevents the graph from being deterministic, this scheme prohibits the dominant cell types from dominating the statistics, therefore allowing people to identify both known cell types and rare/unexpected cell populations. This method has been applied to visualize human bone marrow datasets for recapitulating the entire

hematopoietic system [14, 40]. The data representation in SPADE comes with a tradeoff that single cell resolution is lost in the tree plot after clustering phenotypically similar cells together. The algorithm requires pre-specification of the number of clusters desired while the number of subpopulations is often unknown a priori.

One of the limitations of SPADE is that it does not permit incorporating prior knowledge into the final tree structure. This limit has been resolved by Scaffold algorithm [43] via including manually gated known populations as landmarks in the final layout, which facilitates the interpretation process. More specifically, in Scaffold, cells are first clustered using CLARA (clustering for large applications) algorithm [45] and then spatialized in a 2D plane using force directed layout [46]. Therefore, because of the overlaid known cell type, an advantage of Scaffold map is that it enables rapid comparison of the global data structure with an existing reference. Moreover, it also supports comparison between samples collected from different organs by simply getting rid of the manually identified landmarks and overlaying them with different colors on the force-directed layout, which is useful to reveal detailed local structure of cell subsets (Fig. 2B).

In addition to resolving the global data structure and identifying cellular subgroups, algorithms have been developed to correlate biological features of cell subsets with desired outcomes. Using a regularized regression-based method, Citrus (cluster identification, characterization and regression), a method that takes advantage of both traditional hierarchical clustering and machine learning approaches, helps investigators perform a correlation-based data mining within high-dimensional datasets, calculate the significant features of each cell subset and identify cell populations predictive of a clinical outcome [41]. For example, by applying Citrus to the single cell mass cytometry dataset taken from circulating immune cells from patients undergoing hip replacement, STAT3, CREB and NFκB signaling in subsets of CD14+ monocytes was found to be strongly correlated with clinical parameters of surgical recovery [47].

3.2 Dimensionality reduction algorithms for visualizing overall population structure

While clustering-based analysis can group cells into subpopulations, which facilitates interrogating the differences between subpopulations, yet in a lot of other cases, the cellular heterogeneity is more continuous instead of discrete. In those systems, it might be challenging to set a hard boundary to partition the cells into clusters. Instead, people seek to keep the single cell resolution of the data points and meanwhile reduce the dimensionality without losing too much information so one can directly look at the overall high-dimensional data structure in 2D or 3D space. Dimensionality reduction algorithms that help visualize the data but do not explicitly identify and partition cells into subpopulations would serve for this purpose.

Many dimensionality reduction approaches used for analyzing single cell proteomics data are derived from extensive analytical repertoire in the field of statistics and/or machine learning. Principle component analysis (PCA) is an old but representative method for this purpose [48]. It applies linear combinations of original measured parameters to create new principle variables that retain the most variance of the dataset. Usually, first couple principle components (PCs) are able to capture the main information of the dataset. When coupling first few PCs with cellular functions, the algorithm permits making accurate predictions regarding how a specific perturbation (*e.g.* drug) will disrupt the cellular signaling machinery as demonstrated by Wei *et al.*, in a human brain tumor model of mechanistic target of rapamycin (mTOR) kinase inhibitor resistance [49], and how to rewire the oncogenic signaling pathways to reactivate an extrinsic apoptotic pathway for improved cytotoxic chemotherapy in breast cancer cell lines [50]. However, one caveat of PCA is that a linear projection may be too restrictive to yield accurate representation as often times the biological datasets are nonlinear. Additionally, the representation by the first 2 or 3 PCs (which are easy to visualize) might not be useful for the questions we seek to answer, as the interesting biological differences are often subtle ones covered by the last few compound variables [6, 51].

Visualization of t-Distributed Stochastic Neighbor Embedding (viSNE), a nonlinear dimensionality reduction method, showed great promise in preserving the geometry and nonlinearity of the original

high-dimensional dataset in a 2D or 3D space [52]. viSNE is a variant of stochastic neighbor embedding using the Student-t distribution where pairwise similarity between single cell measurements is quantified to randomly generate a scatter plot in low-dimensional space. The scatter plot is further optimized by gradient descent of the Kullback-Leibler divergence [53], leading to a final optimized low-dimensional embedding that retains local distances and thus arranges neighboring data points in the original space still nearby in the low-dimensional plot (Fig. 2C).

Unlike SPADE, the viSNE algorithm reserves single cell resolution rather being compromised by clustering. It has facilitated a number of high-dimensional single cell studies, including the analysis of the human bone marrow samples [52] and murine myeloid cell system [54]. The algorithm termed one-dimensional soli-expression by nonlinear stochastic embedding (One-SENSE) further assigns a manually predefined category (annotation) with specific biological meaning to each t-SNE axis, allowing testing hypotheses about relationships between different categories of cellular diversity [55]. However, t-SNE based approaches are computationally demanding and thus better suited for analyzing small datasets. When dealing with dataset with a large number of cells, in addition to the computational cost, not all subpopulations are visually distinct in the 2D t-SNE plane and rare cell subpopulations could be obscured by subpopulations with larger number of cells.

To resolve this issue, density-based approaches have been used in conjunction with the neighbor-embedding algorithm for reducing the dimensionality and partitioning the single cell observations into subpopulations. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE) combines t-SNE with density-based partitioning to identify local maxima from viSNE plots for automatic classification of putative cell subpopulations from high-dimensional protein expression data [56]. Alternatively, Phenograph algorithmically extract phenotypically distinct subpopulations from original high-dimensional dataset via a nearest neighbor-based community detection scheme [57]. It denotes phenotypes as communities of densely interconnected nodes and thus reduces the possibility of obscuring important rare cell subpopulations. With this

advantage, Phenograph has been successfully applied to investigate the functional heterogeneity and surface phenotypes of acute myeloid leukemia (Fig. 2D) [57].

3.3 Seriation-based analysis for visualizing cellular transition and progression

Cellular transition between cell states is a fundamental process in biology. High-dimensional single cell proteomic analysis could help resolve the progression trajectory in continuous cellular transition processes, such as immune cell development, based upon the ergodic hypothesis where a snapshot picture of an ensemble of individual cells can inform us about the behavior of an individual over time. A handful of methods have been developed for this purpose using either single cell proteomic or transcriptomic datasets as input [58-62].

As a representative example, Wanderlust converts high-dimensional single cell data into a nearest neighbor graph wherein cells that have similar expression profiles are connected [59]. The algorithm then selects random waypoint cells and assigns each cell's position based upon its relative distance from nearby waypoint cells. A repetitive randomized shortest path algorithm is applied to assign an average position to each cell until the cell's position converges (Fig. 2E). Wanderlust has been used to successfully recapitulate the developmental path of human B cell de novo where a lot more regulatory information has been revealed along the dynamic process [28].

However, a significant limitation of Wanderlust is its underlying assumption that the developmental process is composed of a series of consecutive stages, with no branching. In other words, bifurcating developmental trajectories cannot be handled by Wanderlust. To address this limitation, a couple of algorithms, including diffusion maps [63], Monocle [58], SCUBA [60] and Wishbone [62] have been proposed for pseudo-temporal ordering of cells along the differentiation path with the capability to identify the branch points. The SCUBA method is based on bifurcation analysis as in dynamic systems, while Wishbone roots on Wanderlust algorithm but with an additional module to calculate the mutual disagreement matrix for portraying bifurcation trajectories. Wishbone showed great

promise to model bifurcated T cell development toward CD4+ and CD8+ T cells [62]. Another Wanderlust-based algorithm, termed Cycler, is used to construct continuous trajectories of cell cycle progression from images of fixed cells, and thus allows handling heterogeneous microenvironments (Fig. 2F) [61]. These seriation-based algorithms provide a general analytical platform for interrogating the important continuous cellular transition (progression) in biology, such as lymphopoiesis and carcinogenesis, with multiplex single cell proteomic tools.

4. Biophysical or information theoretical approaches for understanding single cell proteomic data with predictive capacity

Descriptive statistical tools depict the global data structure through mapping individual cells in the high-dimensional space onto an interpretable low (2D or 3D) dimensional space with minimal loss of information. They identify static or pseudo-temporal phenotypic subpopulations via a cohort of clustering strategies. However, single cell proteomic assays, if applied to a statistical number of cells, provide distributions of measured variables for the entire population (Fig. 1C). Such distributions (also termed fluctuations of the variables), originated from stochastic gene expression and epigenetic regulations [64-66], offer an unprecedented wealth of information about the dynamics of cell states, far beyond cell–cell variability and higher statistical moments. As shown in §4.4, the high-amplitude fluctuation at the single cell level but stability across a population is a very common feature for a cell population at steady state [67, 68]. In other words, the population is stable exactly because it is heterogeneous (consider, for example, the robust nature of a diverse economy). Driven by various cellular and environmental cues, single cell population can go beyond the steady state. As in the case of a cellular transition discussed in §4.2 [69], protein fluctuations show long and uneven tails with increased heterogeneity, implying a instability when cells approach the transition point. A further query is that how the inherent heterogeneity of a single cell population contributes to the diverse responses to these perturbations, and how to understand such diversity from a system view of a stable cell population instead of enumerating phenotypic or functional subpopulations. As we

will discuss below, single cell proteomics can provide a conduit to the predictive world of information theoretical approaches underlain by physicochemical principles.

4.1 Quantitative Le Chatelier's principle for predicting cellular responses to weak perturbations

Le Chatelier's principle is extensively used in chemical systems to predict how an equilibrium system responses to an external perturbation. This principle has been generalized and adapted in several fields, including pharmacology and economics. In the theoretical framework of maximum entropy formalism, a quantitative version of Le Chatelier's principle has recently been derived to relate the change in functional protein levels to the change in external conditions [38, 51].

Equilibrium, as an axiomatic concept of statistical physics, denotes no net macroscopic flow of matter or energy within a system or between systems. As a result, cells are non-equilibrium open systems, since they actively maintain concentration gradients, and exchange energy and materials with the environment. However, many experimental results have indicated that a cell population can have a stable steady state within the limits of the measurements, and such steady state enables inferring cellular responses to external cues with physicochemical principles. A stable steady state is one in which the inputs and outputs of a cellular system are balanced. Stable means that, whenever slightly perturbed, the system will recover its original state following release of the perturbation. A prerequisite for using the Le Chatelier's principle is that the perturbation exerted onto the system should be small. This is because that, under a strong perturbation, the system, in our case the cells, may be displaced to a new stable state that is very different from its original unperturbed state [38], as what happened in a cellular transition.

The Le Chatelier's principle is summarized by the matrix equation $\Delta N = \beta \boldsymbol{\Sigma} \Delta \mu$, where ΔN is a column vector with P components representing the change in average protein levels of the P assayed proteins; β is $1/k_B T$, where k_B is Boltzmann's constant and T is temperature; $\boldsymbol{\Sigma}$ is a $P \times P$ matrix where each element is the experimentally measured covariance of a specific protein P_i with

another protein P_j; and $\Delta\mu$ is a column vector whose *P* components account for the change in chemical potentials of the *P* proteins, due to a change in external conditions (the perturbation). For a weak perturbation, the protein copy number changes following perturbation can be predicted by the equation above. However, the equation does not hold for strong perturbations.

Shin et al., coupled multiplex single cell proteomic measurement with this theoretical tool to investigate how the secretome of lipopolysaccharide-stimulated macrophage cells responded to neutralizing antibody perturbations [38]. They correctly predicted how specific cytokine levels would vary with the perturbation based solely on the protein copy numbers measured in unperturbed cells (Fig. 3A). Beyond weak perturbations, the theoretical tool could also infer when a cellular system experiences strong perturbation. In a human glioblastoma (GBM) tumor model, Wei et al. interrogated how the mTORC1 and hypoxia-inducible factor (HIF-1α) signaling axes respond to the changing oxygen partial pressure (pO_2) from normoxia to hypoxia [51]. The theory could correctly predict the change in relevant protein effectors associated mTORC1 above 2% pO_2 or below 1.5% pO_2. However, between 2% and 1.5% pO_2, the prediction did not hold, implying the existence of a strong perturbation (a switch) between two different stable states (Fig. 3B). Such switch renders mTOR unresponsive to external perturbations (such as inhibitors) within this narrow window of pO_2. These surprising predictions were found to be correct in both GBM cell lines and neurosphere models.

4.2 Surprisal Analysis for resolving the steady state and driving constraints in biological system.

Surprisal analysis was first formulated in 1972 by Levine and coworkers under the information theoretical framework of maximum entropy for understanding the dynamics of non-equilibrium systems, particularly of small systems [70, 71]. Later, it has been adapted to various disciplines including engineering, physics, chemistry and recently in cellular systems at both population level [72-74] and single cell level [69, 73, 75]. The basis for applying Surprisal analysis to the single cell system is the concept termed single cell ensemble where relevant molecular distributions are

measured from many independent replicas of a compartment containing a single cell in a nutrient solution at thermal equilibrium [38]. The core ideas of surprisal analysis involve resolving the common steady state and identifying how weak and strong perturbations on the cells are manifested as constraints via the following matrix equation [69]:

$$\underbrace{X_i(\text{cell}, v)}_{\substack{\text{experimental} \\ \text{level of analyte } i}} = \underbrace{X_i^0(\text{cell}, v)}_{\substack{\text{level of analyte} \\ \text{in the steady state}}} \exp\underbrace{\left(-\sum\nolimits_{\alpha=1} G_{i\alpha}\lambda_a(\text{cell}, v)\right)}_{\substack{\text{changes of the free energy} \\ \text{due to the constraints } \alpha=1,2...}}$$

Here, $X_i(\text{cell}, v)$ is the experimentally measured copy number of analyte *i* in a given cell as a function of a parameter *v* (time, drug, etc.) and $X_i^0(\text{cell}, v)$ is the analyte expression level at the steady state. Surprisal analysis is flexible to experimental inputs, and the analytes can be transcript, protein or even metabolite levels. The index *α* refers to a given constraint and $\lambda_a(\text{cell}, v)$ is the weight of that constraint as a function of *v*. $G_{i\alpha}$ is the influence of that constraint on analyte *i.* In practice, 10^4-10^6 data points are integrated into the application of this equation to resolve the steady state, plus any constraints. In the presence of a perturbation, the resolved constraints usually have amplitude of a few percent of the steady state.

Surprisal analysis has been applied towards understanding of early stage carcinogenesis [72], cellular homeostasis [76] and cellular transitions [74]. Recently, it has been extended to investigate cell-cell interactions for predicting GBM cellular architectures. In this work, Kravchenko-Balasha et al. analyzed interactions of tumor cells through measuring the abundance of cytokines and phosphoproteins in isolated pairs of GBM cells at varying separation distances [77]. Surprisal analysis was applied towards determining the most balanced state of the two cells across a range of cell-cell separation distance. The steady-state separation distance was identified between two cells for a couple of GBM cell lines. The prediction was found to be consistent with the most probable distance of those cells in bulk culture (Fig. 3C) [77]. Using the same approach, Kravchenko-Balasha et al. also

made the surprising prediction of cell migration towards the steady state via analyzing secreted proteins from hundreds of isolated GBM cell pairs [75].

In another example, Poovathingal et al. explored the chemically-induced carcinogenesis in MCF-10A human mammary epithelial cells by exposing those cells, *in vitro*, to benzo[a]pyrene (B[a]P) up to 3 months [69]. They analyzed a panel of functional proteins and transcriptional factors associated with this process at single cell level. Surprisal analysis successfully identified two steady states before and after the treatment, and a bifurcation of the cellular populations around 2-4 weeks after the treatment, pointing to a phase coexistence that is reminiscent of a phase transition in physical system before traditional biomarkers of carcinogenic transformation becoming detectable in experiments (Fig. 3D) [69].

Maximum entropy-based information theoretical approaches, coupled with single cell proteomics, have been employed to predict a cohort of biologically complicated cellular behaviors, which can be experimentally validated. One limitation of these tools is that they normally require, as input, absolute quantification in protein copy number or some measure that is linearly proportional to copy number. Such requirement limits their utilities in analyzing some proteomic datasets with low abundant proteins or excessively high abundant house-keeping proteins whose measured intensities are beyond the linear dynamic ranges and not stoichiometrically related to the protein concentrations.

4.3 Mutual Information-based analytic method for studying parameter dependency

Mutual information is a measurement of the mutual dependence between two random variables. In other words, it quantifies how much information we know about one variable when we already know the other variable. It has been experimentally confirmed that gene expression is a stochastic process [78]. Mutual information approach provides an objective means to quantify the influence of this stochasticity [79]. However, a potential challenge of using this approach in analyzing single cell

datasets is that mutual information is difficult to compute on continuous data at single cell resolution. While adaptive partitioning [80] has been adapted [81, 82], yet it is biased toward the dense region of the data points, which may fail to cover the important information in the sparse edge regions. To address this challenge, Krishnaswamy et al. developed an algorithm termed DREMI (conditional-Density Resampled estimate of Mutual Information) to quantify the dependency across all populated regions of a dynamic range regardless of the original distribution through computing the mutual information on the conditional density estimate of the data rather than the raw data [83]. It thus allows ascertaining the effect of one protein's activity on that of another by considering one protein as a stochastic function of another and quantifying the strength of underlying relationships between the proteins.

To be specific, DREMI utilizes conditional density estimation and rescaling to evenly resample the data across the entire range. Mutual information based upon the conditional probability is computed on the resampled data to recover the dependency. Worth noting, DREVI (conditional-Density Rescaled Visualization) method is applied to visualize the dependency relationship as a rescaled heatmap. Applying the method to analyze dynamics of singling response to TCR activation, they showed the strength of signal transfer peaks in canonical pathway order. They found naïve and antigen exposed CD4+ T cells have differential information transfer rates along the signaling cascade. Naive cells had more information transferred than did antigen-exposed cells. This prediction was successfully validated in mouse models (Fig. 3E) [83].

4.4 Potential landscape model for predicting cell state dynamics

The cell state (phenotype) distribution and transition within a cell population have been investigated using potential landscape model. The ideas originate from Waddington’s landscape [84, 85] that is one of the most famous metaphors for depicting how stem cells differentiate into discrete, robust cell states as a marble rolls down a hilly landscape towards a number of valleys. Marbles eventually settle into one of valley, which is similar to that cells differentiate into one matured state [7].

In a modern version of Waddington's landscape model, multiple attractors have been proposed, in which cells reside [86]. The spread of the cluster around an attractor state is a measure of heterogeneity of this specific cell type [7]. The phenotypic transition induced by an external cue can be viewed as transition between two attractors, where the barrier height in between governs the probability and direction of the transition. The landscape geometry determines the dynamic trajectory of each individual cell and also the equilibrium phenotypic distribution of cells in the high dimensional parameter space. Therefore, steady state phenotypic distribution as well as the dynamics of cell state transition could be used to infer the potential landscape. Sisan et al. demonstrated the utility of an 1D quantitative potential landscape by investigating the GFP expression levels of a fibroblast cell line under the control of the promoter for tenascin-C [67]. They used a Langevin-type stochastic differential equation to quantify the steady state distribution of GFP expression to calculate the potential. The diffusion coefficient obtained from time-lapse microscopy was used to characterize the GFP fluctuation. This approach accurately predicted the rates at which segregated subpopulations relaxed back to the steady state (Fig. 3F).

A similar approach was used to analyze expression fluctuation of a stem cell surface marker Sca1 in mouse hematopoietic progenitor cells [87]. The results indicated that the dynamics lay close to a critical state where the trade-off between maximizing cell-cell variability and maintaining the capacity to respond rapidly to environment changes was well balanced.

So far, most of these landscape-modeling approaches are working on low-dimensional single cell data due to the computational cost. A rational incorporation of these theoretical tools with appropriate dimensionality reduction algorithms to dissect the high-dimensional single cell data would lead to a more comprehensive picture of the landscape geometry and cell state dynamics.

5. Challenges, limitations, and outlook

The remarkable advances of single cell proteomics provide powerful toolkits capable of assaying up to 50 proteins simultaneously in thousands of individual cells, opening new opportunities for interrogating a wealth of biological inquiries that were disguised by traditional population measurements. The majority of technical platforms and analytical methods discussed here were not existed or just beginning to emerge 5 years ago. They are now becoming routine biological tools in many laboratories. From a technical perspective, a major bottleneck of currently available tools is the level of multiplexing. Both cytometry and microchip tools sample around 20–40 proteins in regular practice and likely reach a limit between 50 and 100 proteins, which represents sampling only a tiny part of the proteome. The limit arises because of the reliance on antibody-based detection schemes. It poses biased analyses, as extensive prior knowledge is normally required for selecting protein panel that is most relevant to the problem under study, and thus precludes the utility in discovery level studies where prior knowledge is limited. This challenge might be overcome through the development of highly sensitive mass spectrometry-based tools at single cell resolution. Targeted proteomics using mass spectrometry has already evolved to the extent of analyzing small cell numbers. Protein processing with immobilized enzymes [88] or novel column chromatography methods [89] may eventually allow mass spectrometry to be a single cell proteomics discovery tool.

A fusion of two or more detection schemes for simultaneous assaying, at single cell level, multiple classes of biomolecules in a single test is a natural step forward in the technical development. Such assays would allow a suite of cellular processes to be portrayed from the same single cells, revealing the behavior of individual cells in a more holistic manner. Primitive efforts have already been made towards co-profiling functional proteins and genomes [90], transcripts [91-93] or metabolites [26,

Received: 31/10/2016; Revised: 20/01/2017; Accepted: 20/01/2017
This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201600267.

94] from the same single cells. Given the rapid evolvement of this field, high throughput single cell technologies for integrated analysis of several classes of biomolecules may be on the horizon.

Matching cellular heterogeneity with biological context remains a grand challenge for most single cell proteomic tools as the assays typically involve removal of the cells from their native context by dissociating the tissue samples into single cell suspension before analysis. Thus, while cellular heterogeneity is resolved, the context of that heterogeneity is lost. The spatial distribution of antigens in tissue context is important in certain scenarios, such as resolving the presence and distribution of CD4+ and CD8+ tumor infiltrating T cells in the tumor tissue section for immunotherapy design. The recent development of multiplexed ion bead imaging [95] and imaging mass cytometry [96] enables the analysis of single cells *in situ* within formalin-fixed, paraffin-embedded tissue section, with a level of multiplexing that significantly exceeds traditional immunohistochemistry. The integration of molecular barcoding methods [97] with expansion microscopy [98] might provide an alternative approach towards analyzing the molecular profiles of the single cells within intact tissue samples. While the proteomic analysis on fixed tissues limits resolving the activities or dynamics of the protein signaling, we expect further advances in these *in situ* multiplexed single cell proteomic approaches will provide messages complementary to other single cell tools and transform our understanding of the cellular heterogeneity in the unperturbed tissue context.

The increasing complexity of the high-dimensional single cell datasets requires continuous progress in the development of new analytical strategies and computational tools for gleaning useful biological insights from these data. While significant efforts have been made so far, the development of computational tools is still lagging behind the advances in experimental technologies. A major goal of high-dimensional analysis of single cells is not only to understand the relationships among various conceptual aspects of a cell population, but also to generate testable hypotheses regarding how the heterogeneous population would respond and adapt to various cellular and environmental

cues. However, the majority of algorithms for dissecting single cell proteomic datasets center on discerning the global data structure via data visualization, and dimensionality reduction, as well as identifying significant patterns (either pseudo-temporal order of cellular progression or static clusters). Very few algorithms are designated for making statistical inference or identifying cellular features that correlate with a desired outcome. This may in part clarify why most investigations using these powerful single cell proteomic tools are explorative and descriptive in nature rather than hypothesis driven. Now the time is ripe to move beyond these descriptive computational analyses. The idea that single cell functional proteomics can provide a conduit to the predictive world of statistical physics is exciting. Preliminary explorations of this idea have been encouraging, but the benefits (and limitations) of this type of thinking are largely untapped. It is certain, however, as the single cell tools continue to improve in multiplexing capacity, throughput, sensitivity and quantification, an overarching analytical framework that connects biological questions, experimental designs, to data analysis will eventually transform the practice of biomedical research as well as our understanding in single cell biology.

Acknowledgement

The authors acknowledge the following funding agencies and grants for support some of the work described in this Review: NIH/NCI 1U54 CA199090-01 (W.W.); 5U54 CA151819 (W.W.); the Phelps Family Foundation (W.W.); Youth Program of the National 1000 Talents Project (Q.H.S.).

Competing interests statement

The authors have declared no conflict of interest

6. References

[1] Altschuler, S. J., Wu, L. F., Cellular heterogeneity: do differences make a difference? *Cell* 2010, *141*, 559-563.

[2] Heath, J. R., Ribas, A., Mischel, P. S., Single-cell analysis tools for drug discovery and development. *Nat Rev Drug Discov* 2016, *15*, 204-216.

[3] Wei, W., Shin, Y. S., Ma, C., Wang, J.*, et al.*, Microchip platforms for multiplex single-cell functional proteomics with applications to immunology and cancer research. *Genome Med* 2013, *5*, 75.

[4] Heath, J. R., Nanotechnologies for biomedical science and translational medicine. *Proc Natl Acad Sci U S A* 2015, *112*, 14436-14443.

[5] Yu, J., Zhou, J., Sutherland, A., Wei, W.*, et al.*, Microfluidics-based single-cell functional proteomics for fundamental and applied biomedical applications. *Annu Rev Anal Chem (Palo Alto Calif)* 2014, *7*, 275-295.

[6] Spitzer, M. H., Nolan, G. P., Mass Cytometry: Single Cells, Many Features. *Cell* 2016, *165*, 780-791.

[7] Wang, J., Yang, F., Emerging single-cell technologies for functional proteomics in oncology. *Expert Rev Proteomics* 2016, *13*, 805-815.

[8] Marr, C., Zhou, J. X., Huang, S., Single-cell gene expression profiling and cell state dynamics: collecting data, correlating data points and connecting the dots. *Curr Opin Biotechnol* 2016, *39*, 207-214.

[9] Czerkinsky, C. C., Nilsson, L. A., Nygren, H., Ouchterlony, O., Tarkowski, A., A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells. *J Immunol Methods* 1983, *65*, 109-121.

Received: 31/10/2016; Revised: 20/01/2017; Accepted: 20/01/2017
This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201600267.

[10] Herzenberg, L. A., Parks, D., Sahaf, B., Perez, O.*, et al.*, The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. *Clin Chem* 2002, *48*, 1819-1827.

[11] De Rosa, S. C., Herzenberg, L. A., Herzenberg, L. A., Roederer, M., 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nat Med* 2001, *7*, 245-248.

[12] Perfetto, S. P., Chattopadhyay, P. K., Roederer, M., Seventeen-colour flow cytometry: unravelling the immune system. *Nat Rev Immunol* 2004, *4*, 648-655.

[13] Perez, O. D., Nolan, G. P., Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nat Biotechnol* 2002, *20*, 155-162.

[14] Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E. A. D.*, et al.*, Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science* 2011, *332*, 687-696.

[15] Huebner, A., Srisa-Art, M., Holt, D., Abell, C.*, et al.*, Quantitative detection of protein expression in single cells using droplet microfluidics. *Chem Commun* 2007, 1218-1220.

[16] Mazutis, L., Gilbert, J., Ung, W. L., Weitz, D. A.*, et al.*, Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols* 2013, *8*, 870-891.

[17] Tricot, S., Meyrand, M., Sammicheli, C., Elhmouzi-Younes, J.*, et al.*, Evaluating the Efficiency of Isotope Transmission for Improved Panel Design and a Comparison of the Detection Sensitivities of Mass Cytometer Instruments. *Cytom Part A* 2015, *87a*, 357-368.

[18] Finck, R., Simonds, E. F., Jager, A., Krishnaswamy, S.*, et al.*, Normalization of mass cytometry data with bead standards. *Cytom Part A* 2013, *83a*, 483-494.

[19] Zrazhevskiy, P., Gao, X. H., Quantum dot imaging platform for single-cell molecular profiling. *Nat Commun* 2013, *4*, 1619.

[20] Agasti, S. S., Liong, M., Peterson, V. M., Lee, H., Weissleder, R., Photocleavable DNA Barcode-Antibody Conjugates Allow Sensitive and Multiplexed Protein Analysis in Single Cells. *J Am Chem Soc* 2012, *134*, 18499-18502.

[21] Ullal, A. V., Peterson, V., Agasti, S. S., Tuang, S.*, et al.*, Cancer Cell Profiling by Barcoding Allows Multiplexed Protein Analysis in Fine-Needle Aspirates. *Sci Transl Med* 2014, *6*, 219ra9.

[22] Hughes, A. J., Spelke, D. P., Xu, Z., Kang, C. C.*, et al.*, Single-cell western blotting. *Nat Methods* 2014, *11*, 749-755.

[23] Kang, C. C., Yamauchi, K. A., Vlassakis, J., Sinkala, E.*, et al.*, Single cell-resolution western blotting. *Nat Protoc* 2016, *11*, 1508-1530.

[24] Shi, Q. H., Qin, L. D., Wei, W., Geng, F.*, et al.*, Single-cell proteomic chip for profiling intracellular signaling pathways in single tumor cells. *P Natl Acad Sci USA* 2012, *109*, 419-424.

[25] Ma, C., Fan, R., Ahmad, H., Shi, Q. H.*, et al.*, A clinical microchip for evaluation of single immune cells reveals high functional heterogeneity in phenotypically similar T cells. *Nature Medicine* 2011, *17*, 738-U133.

[26] Xue, M., Wei, W., Su, Y., Kim, J.*, et al.*, Chemical methods for the simultaneous quantitation of metabolites and proteins from single cells. *J Am Chem Soc* 2015, *137*, 4066-4069.

[27] Love, J. C., Ronan, J. L., Grotenbreg, G. M., van der Veen, A. G., Ploegh, H. L., A microengraving method for rapid selection of single cells producing antigen-specific antibodies. *Nature Biotechnology* 2006, *24*, 703-707.

[28] Torres, A. J., Contento, R. L., Gordo, S., Wucherpfennig, K. W., Love, J. C., Functional single-cell analysis of T-cell activation by supported lipid bilayer-tethered ligands on arrays of nanowells. *Lab Chip* 2013, *13*, 90-99.

[29] Wang, J., Tham, D., Wei, W., Shin, Y. S.*, et al.*, Quantitating Cell-Cell Interaction Functions with Applications to Glioblastoma Multiforme Cancer Cells. *Nano Lett* 2012, *12*, 6101-6106.

[30] Yang, L., Wang, Z., Deng, Y., Li, Y.*, et al.*, Single-Cell, Multiplexed Protein Detection of Rare Tumor Cells Based on a Beads-on-Barcode Antibody Microarray. *Anal Chem* 2016, *88*, 11077–11083.

[31] Shin, Y. S., Ahmad, H., Shi, Q., Kim, H.*, et al.*, Chemistries for patterning robust DNA microbarcodes enable multiplex assays of cytoplasm proteins from single cancer cells. *Chemphyschem* 2010, *11*, 3063-3069.

[32] Lu, Y., Xue, Q., Eisele, M. R., Sulistijo, E. S*., et al.*, Highly multiplexed profiling of single-cell effector functions reveals deep functional heterogeneity in response to pathogenic ligands. *Proc Natl Acad Sci U S A* 2015, *112*, E607-615.

[33] Ma, C., Cheung, A. F., Chodon, T., Koya, R. C*., et al.*, Multifunctional T-cell analyses to study response and progression in adoptive cell transfer immunotherapy. *Cancer Discov* 2013, *3*, 418-429.

[34] Rissin, D. M., Kan, C. W., Campbell, T. G., Howes, S. C*., et al.*, Single-molecule enzyme-linked immunosorbent assay detects serum proteins at subfemtomolar concentrations. *Nature Biotechnology* 2010, *28*, 595-599.

[35] Schubert, S. M., Walter, S. R., Manesse, M., Walt, D. R., Protein Counting in Single Cancer Cells. *Analytical Chemistry* 2016, *88*, 2952-2957.

[36] Baker, M., Reproducibility crisis: Blame it on the antibodies. *Nature* 2015, *521*, 274-276.

[37] Marcon, E., Jain, H., Bhattacharya, A., Guo, H*., et al.*, Assessment of a method to characterize antibody selectivity and specificity for use in immunoprecipitation. *Nature methods* 2015, *12*, 725-731.

[38] Shin, Y. S., Remacle, F., Fan, R., Hwang, K*., et al.*, Protein signaling networks from single cell fluctuations and information theory profiling. *Biophys J* 2011, *100*, 2378-2386.

[39] Alvarez, D. F., Helm, K., Degregori, J., Roederer, M., Majka, S., Publishing flow cytometry data. *Am J Physiol Lung Cell Mol Physiol* 2010, *298*, L127-130.

[40] Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Jr*., et al.*, Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol* 2011, *29*, 886-891.

[41] Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J., Nolan, G. P., Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci U S A* 2014, *111*, E2770-2777.

[42] Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L., Nolan, G. P., Automated mapping of phenotype space with single-cell data. *Nat Methods* 2016, *13*, 493-496.

[43] Spitzer, M. H., Gherardini, P. F., Fragiadakis, G. K., Bhattacharya, N.*, et al.*, IMMUNOLOGY. An interactive reference framework for modeling a dynamic immune system. *Science* 2015, *349*, 1259425.

[44] Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N.*, et al.*, FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* 2015, *87*, 636-645.

[45] Kaufman, L., Rousseeuw, P. J., *Finding Groups in Data*, John Wiley & Sons, Inc. 2008, pp. 126-163.

[46] Dutkowski, J., Kramer, M., Surma, M. A., Balakrishnan, R.*, et al.*, A gene ontology inferred from molecular networks. *Nat Biotechnol* 2013, *31*, 38-45.

[47] Gaudilliere, B., Fragiadakis, G. K., Bruggner, R. V., Nicolau, M.*, et al.*, Clinical recovery from surgery correlates with single-cell immune signatures. *Sci Transl Med* 2014, *6*, 255ra131.

[48] Kholodenko, B., Yaffe, M. B., Kolch, W., Computational approaches for analyzing information flow in biological networks. *Sci Signal* 2012, *5*, re1.

[49] Wei, W., Shin, Y. S., Xue, M., Matsutani, T.*, et al.*, Single-Cell Phosphoproteomics Resolves Adaptive Signaling Dynamics and Informs Targeted Combination Therapy in Glioblastoma. *Cancer Cell* 2016, *29*, 563-573.

[50] Lee, M. J., Ye, A. S., Gardino, A. K., Heijink, A. M.*, et al.*, Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* 2012, *149*, 780-794.

[51] Wei, W., Shi, Q. H., Remacle, F., Qin, L. D.*, et al.*, Hypoxia induces a phase transition within a kinase signaling network in cancer cells. *P Natl Acad Sci USA* 2013, *110*, E1352-E1360.

[52] Amir el, A. D., Davis, K. L., Tadmor, M. D., Simonds, E. F.*, et al.*, viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 2013, *31*, 545-552.

[53] Kullback, S., Leibler, R. A., On Information and Sufficiency. *Ann Math Stat* 1951, *22*, 79-86.

[54] Becher, B., Schlitzer, A., Chen, J., Mair, F.*, et al.*, High-dimensional analysis of the murine myeloid cell system. *Nat Immunol* 2014, *15*, 1181-1189.

[55] Cheng, Y., Wong, M. T., van der Maaten, L., Newell, E. W., Categorical Analysis of Human T Cell Heterogeneity with One-Dimensional Soli-Expression by Nonlinear Stochastic Embedding. *J Immunol* 2016, *196*, 924-932.

[56] Shekhar, K., Brodin, P., Davis, M. M., Chakraborty, A. K., Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proc Natl Acad Sci U S A* 2014, *111*, 202-207.

[57] Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L.*, et al.*, Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 2015, *162*, 184-197.

[58] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P.*, et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014, *32*, 381-386.

[59] Bendall, S. C., Davis, K. L., Amir el, A. D., Tadmor, M. D.*, et al.*, Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 2014, *157*, 714-725.

[60] Marco, E., Karp, R. L., Guo, G., Robson, P.*, et al.*, Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci U S A* 2014, *111*, E5643-5650.

[61] Gut, G., Tadmor, M. D., Pe'er, D., Pelkmans, L., Liberali, P., Trajectories of cell-cycle progression from fixed cell populations. *Nat Methods* 2015, *12*, 951-954.

[62] Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O.*, et al.*, Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* 2016, *34*, 637-645.

[63] Haghverdi, L., Buettner, F., Theis, F. J., Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 2015, *31*, 2989-2998.

[64] Mettetal, J. T., Muzzey, D., Pedraza, J. M., Ozbudak, E. M., van Oudenaarden, A., Predicting stochastic gene expression dynamics in single cells. *Proc Natl Acad Sci U S A* 2006, *103*, 7304-7309.

[65] Blake, W. J., M, K. A., Cantor, C. R., Collins, J. J., Noise in eukaryotic gene expression. *Nature* 2003, *422*, 633-637.

[66] Bajic, D., Poyatos, J. F., Balancing noise and plasticity in eukaryotic gene expression. *BMC Genomics* 2012, *13*, 343.

[67] Sisan, D. R., Halter, M., Hubbard, J. B., Plant, A. L., Predicting rates of cell state change caused by stochastic fluctuations using a data-driven landscape model. *P Natl Acad Sci USA* 2012, *109*, 19262-19267.

[68] Ridden, S. J., Chang, H. H., Zygalakis, K. C., MacArthur, B. D., Entropy, Ergodicity, and Stem Cell Multipotency. *Phys Rev Lett* 2015, *115*, 208103.

[69] Poovathingal, S. K., Kravchenko-Balasha, N., Shin, Y. S., Levine, R. D., Heath, J. R., Critical Points in Tumorigenesis: A Carcinogen-Initiated Phase Transition Analyzed via Single-Cell Proteomics. *Small* 2016, *12*, 1425-1431.

[70] Agmon, N., Alhassid, Y., Levine, R. D., Algorithm for Finding the Distribution of Maximal Entropy. *J Comput Phys* 1979, *30*, 250-258.

[71] Levine, R. D., *Molecular reaction dynamics*, Cambridge University Press, Cambridge, UK ; New York 2005.

[72] Remacle, F., Kravchenko-Balasha, N., Levitzki, A., Levine, R. D., Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *Proc Natl Acad Sci U S A* 2010, *107*, 10324-10329.

[73] Kravchenko-Balasha, N., Johnson, H., White, F. M., Heath, J. R., Levine, R. D., A Thermodynamic-Based Interpretation of Protein Expression Heterogeneity in Different Glioblastoma Multiforme Tumors Identifies Tumor-Specific Unbalanced Processes. *J Phys Chem B* 2016, *120*, 5990-5997.

[74] Zadran, S., Arumugam, R., Herschman, H., Phelps, M. E., Levine, R. D., Surprisal analysis characterizes the free energy time course of cancer cells undergoing epithelial-to-mesenchymal transition. *Proc Natl Acad Sci U S A* 2014, *111*, 13235-13240.

[75] Kravchenko-Balasha, N., Shin, Y. S., Sutherland, A., Levine, R. D., Heath, J. R., Intercellular signaling through secreted proteins induces free-energy gradient-directed cell movement. *P Natl Acad Sci USA* 2016, *113*, 5520-5525.

[76] Kravchenko-Balasha, N., Levitzki, A., Goldstein, A., Rotter, V.*, et al.*, On a fundamental structure of gene networks in living cells. *Proc Natl Acad Sci U S A* 2012, *109*, 4702-4707.

[77] Kravchenko-Balasha, N., Wang, J., Remacle, F., Levine, R. D., Heath, J. R., Glioblastoma cellular architectures are predicted through the characterization of two-cell interactions. *P Natl Acad Sci USA* 2014, *111*, 6521-6526.

[78] Raj, A., van Oudenaarden, A., Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 2008, *135*, 216-226.

[79] Bowsher, C. G., Swain, P. S., Environmental sensing, information transfer, and cellular decision-making. *Curr Opin Biotechnol* 2014, *28*, 149-155.

[80] Darbellay, G. A., Vajda, I., Estimation of the information by an adaptive partitioning of the observation space. *Ieee T Inform Theory* 1999, *45*, 1315-1321.

[81] Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M.*, et al.*, Reverse engineering cellular networks. *Nat Protoc* 2006, *1*, 662-671.

[82] Jang, I. S., Margolin, A., Califano, A., hARACNe: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests. *Interface Focus* 2013, *3*, 20130011.

[83] Krishnaswamy, S., Spitzer, M. H., Mingueneau, M., Bendall, S. C.*, et al.*, Conditional density-based analysis of T cell signaling in single-cell data. *Science* 2014, *346*, 1250689.

[84] Waddington, C. H., *The strategy of the genes; a discussion of some aspects of theoretical biology*, Allen & Unwin, London, 1957.

[85] Yamanaka, S., Elite and stochastic models for induced pluripotent stem cell generation. *Nature* 2009, *460*, 49-52.

[86] Wang, J., Xu, L., Wang, E., Huang, S., The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophys J* 2010, *99*, 29-39.

[87] Ridden, S. J., Chang, H. H., Zygalakis, K. C., MacArthur, B. D., Entropy, Ergodicity, and Stem Cell Multipotency. *Phys Rev Lett* 2015, *115*, 208103.

[88] Tian, R. J., Wang, S. A., Elisma, F., Li, L.*, et al.*, Rare Cell Proteomic Reactor Applied to Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)-based Quantitative Proteomics Study of Human Embryonic Stem Cell Differentiation. *Mol Cell Proteomics* 2011, *10*, M110.000679.

[89] Thakur, D., Rejtar, T., Wang, D. D., Bones, J.*, et al.*, Microproteomic analysis of 10,000 laser captured microdissected breast tumor cells using short-range sodium dodecyl sulfate-polyacrylamide gel electrophoresis and porous layer open tubular liquid chromatography tandem mass spectrometry. *J Chromatogr A* 2011, *1218*, 8168-8174.

[90] Zhang, Y., Tang, Y., Sun, S., Wang, Z. H.*, et al.*, Single-Cell Codetection of Metabolic Activity, Intracellular Functional Proteins, and Genetic Mutations from Rare Circulating Tumor Cells. *Analytical Chemistry* 2015, *87*, 9761-9768.

[91] George, J., Wang, J., Assay of Genome-Wide Transcriptome and Secreted Proteins on the Same Single Immune Cells by Microfluidics and RNA Sequencing. *Anal Chem* 2016, *88*, 10309-10315.

[92] Frei, A. P., Bava, F. A., Zunder, E. R., Hsieh, E. W. Y.*, et al.*, Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nature Methods* 2016, *13*, 269-275.

[93] Albayrak, C., Jordi, C. A., Zechner, C., Lin, J.*, et al.*, Digital Quantification of Proteins and mRNA in Single Mammalian Cells. *Mol Cell* 2016, *61*, 914-924.

[94] Xue, M., Wei, W., Su, Y. P., Johnson, D., Heath, J. R., Supramolecular Probes for Assessing Glutamine Uptake Enable Semi-Quantitative Metabolic Models in Single Cells. *J Am Chem Soc* 2016, *138*, 3085-3093.

[95] Angelo, M., Bendall, S. C., Finck, R., Hale, M. B.*, et al.*, Multiplexed ion beam imaging of human breast tumors. *Nature Medicine* 2014, *20*, 436-442.

[96] Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N.*, et al.*, Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods* 2014, *11*, 417-422.

[97] Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., Cai, L., Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods* 2014, *11*, 360-361.

[98] Chozinski, T. J., Halpern, A. R., Okawa, H., Kim, H. J.*, et al.*, Expansion microscopy with conventional antibodies and fluorescent proteins. *Nature Methods* 2016, *13*, 485-488.

Figure legends:

Figure 1. Representative multiplex single cell proteomic platforms and datasets. **(A)** Illustration of the workflow of a mass cytometry experiment. Cells labeled with mass-tagged antibodies are nebulized into droplets, ionized and atomized by argon plasma. The resulting ion cloud passes through a mass filter where transition metal reporters are quantified by a time-of-flight mass spectrometry. **(B)** Illustration of a microfluidics-based single cell barcode chip. Single cells are loaded into microchambers equipped with miniaturized antibody microarray. Cytokines secreted from cells as well as cytoplasmic and membrane proteins released upon cell lysis are captured by the designated antibody barcodes. Protein assays are developed using fluorophore-labeled detection antibodies and the signals are digitized by a microarray scanner. **(C)** A typical single cell proteomic dataset can be formularized as a table (left) where each row denotes a single cell measurements and each column denotes a measured protein level across the single cells. The distribution of a protein level as tabulated across many single cells is termed fluctuation of that protein (middle) that reveals the inherent heterogeneity of the cell population. The biaxial plot of two proteins (right) can be used to identify specific subpopulations or extract protein-protein correlations.

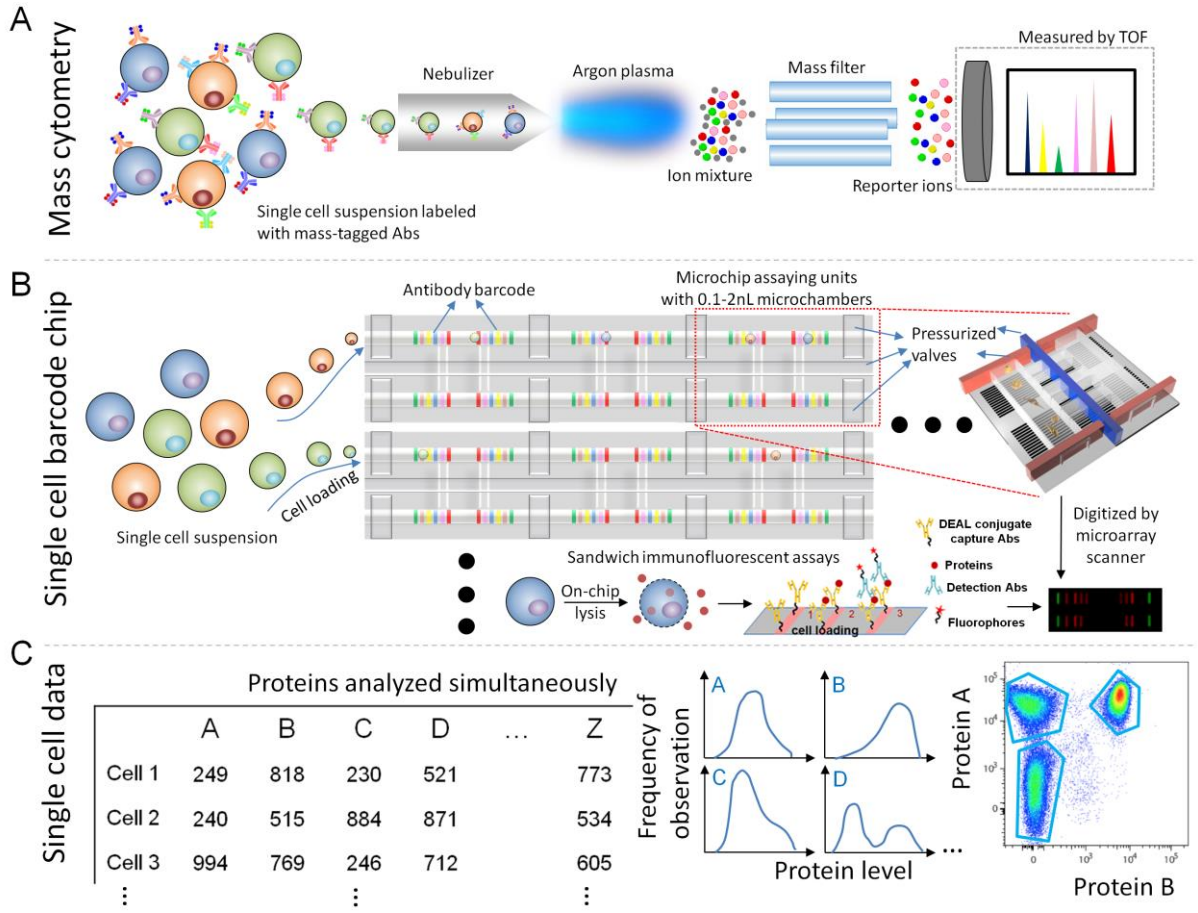

Figure 2. Representative analytical tools for high-dimensional data visualization and pattern identification. **(A)** SPADE visualization of immunophenotypic progression in healthy human bone marrow. **(B)** Scaffold map visualization of bone marrow sample taken from C57BL/6 mice with canonical subpopulation labeled as landmark on map. Respective cell type were manually gated, subjected to unsupervised clustering, and laid out in an unsupervised force-directed graph with the tissue of origin color-coded. **(C)** viSNE visualization of healthy human bone marrow sample with all cell type being automatically separated. Different colors represent different cell types. **(D)** Phenograph clusters visualized on a t-SNE plane color-coded with sample ID (upper left) or average marker expression. Each cluster is represented by a single point scaled to its sample proportion. **(E)** Wanderlust trace visualization of signal intensities for four marker proteins across a five-point drug dose–response (0–1 μM) profile of a cancer cell line. The color-box on the bottom represents the cell

densities while the color scale is showed on the side. **(F)** Cycler visualization of cell-cycle progression from image-based dataset where cells and nuclei are segmented, and features are extracted for the construction of trajectory. Cells are then ordered along the trajectory with fractions of cells in the cell-cycle stage phases overlaid. Panels A-F are adapted with permission from Reference [40], [43], [52], [57], [59], [61], respectively.

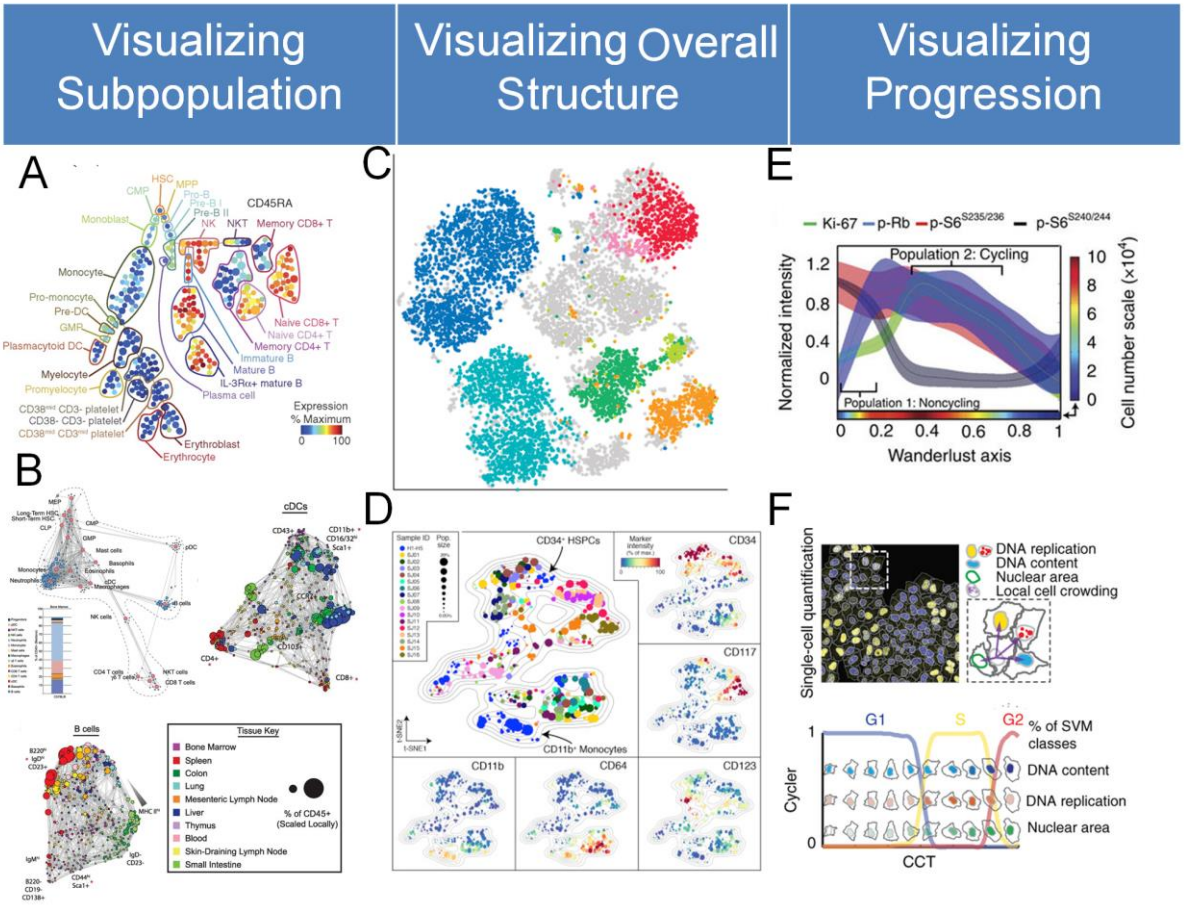

Figure 3. Representative biophysical or information theoretical approaches for analyzing single cell proteomic data. **(A)** Protein-protein interactions and the respective covariance matrix derived from the quantitative Le Chatelier's theorem is visualized by Heatmap representation (Top). The measured change in the mean copy number of eight proteins in response to the addition of a neutralizing antibody is compared against the predicted change computed by the theorem using the

unperturbed single cell data (Bottom). **(B)** Quantitative Le Chatelier's principle reveals an oxygen partial pressure (pO_2)-dependent phase transition in the mTORC1 signaling network within model GBM cells. Measured and predicted changes of the assayed proteins are compared as pO_2 varies between specified levels. The agreement between experiment and prediction for 21–3% and 1.5–1% implies that these pO_2 changes constitute only weak perturbations to the cellular system. The change from 3% to 2% pO_2 denotes stronger perturbation, whereas for the range 2–1.5% pO_2, a transition is implied by the qualitative disagreement between prediction and experiment. **(C)** The amplitudes of the top two constraints, as a function of separation distance are resolved from surprisal analysis of the single cell data. Note that both constraints are zero-valued near 90 micrometers (Top). Analysis of the model GBM cells in bulk culture (Bottom). The inset image is a digitized image used for calculating the radial distribution function (RDF) of the cells. The plot, which was extracted from the RDF, indicates that the most probable (and lowest free energy) cell-cell separation distance is around 90 micrometers, which is consistent with the theoretical predictions. **(D)** Number of cells *vs.* $G_{i0}\lambda_0$(cell, t), with different panels shown at different time points shown for the pS6K protein. The distribution at each time point was fitted to either unimodal or bimodal Gaussian distributions. Bimodal Gaussian distributions appear as the best fitting for days 12 and 28, implying a phase co-existence during the transition, whereas unimodal distribution is the best fitting for the control and day 96. **(E)** Dynamics of TCR signaling revealed by DREVI and DREMI analysis. Comparison of naïve T cells with antigen-exposed T cells is shown in the bar graph. Network representation shows signal transmission is sharper and more sustained in naïve cells. **(F)** Segregated subpopulations with differential GFP expression levels are cultured separately to relaxed back to the steady state distribution (Top). Estimated steady state distribution and respective potential landscape derived from the same distribution (Middle). Landscape model successfully predicts the dynamic of relaxation back to the original equilibrium from subpopulations sorted out from different regions in the original cell population (Bottom). Panels A-F are adapted with permission from Reference [38], [51], [77], [69], [83], [67], respectively.

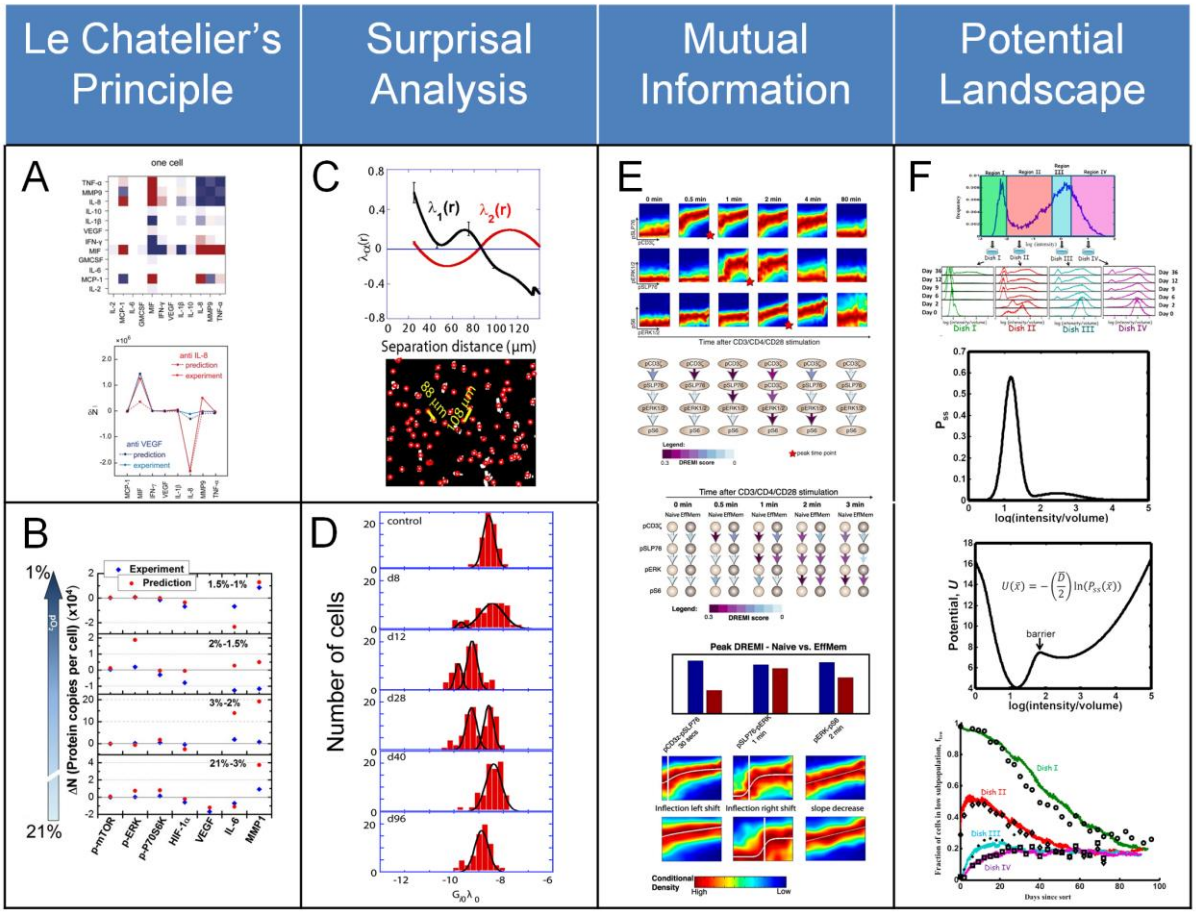

Table1. Descriptive statistical algorithms for visualizing and analyzing single cell proteomic data

Category	Type of question to address	Method name	Unique features	Limitations	References
Clustering-based analysis	Subpopulation/phenotype identification	Manual gating	Widely used in flow cytometry, easier to implement when prior knowledge is available.	Prior knowledge of the system is required; limited to low-dimensional dataset; gating is subjective	[39]
		SPADE	Unbiased density-based clustering; tree structure for visualizing subpopulation relationship.	Loss of single cell resolution; some algorithms require pre-specification of number of clusters	[40]
		Scaffold Map	Unbiased clustering of cell clusters; force directed layout for visualizing subpopulation relationship; prior knowledge is overlaid.		[43]
		Citrus	Identify cell subsets associated with an experimental endpoint of interest; allow correlating biological features with desired outcomes		[41]
		FlowSOM	Self-organized maps for data visualization; MST and t-SNE options are also provided.		[44]
		X-shift	Use fast k-nearest-neighbor estimation of cell event density for automated clustering; arrange populations by marker-based classification; high F-measure		[42]
Dimensionality reduction algorithm	Global data structure with single cell resolution	PCA	Linear combinations of original measured parameters to create new principle variables that retain the most variance	Do not account for nonlinear relationship between parameters	[48]
		viSNE	Nonlinear dimensionality reduction with single cell resolution resolved	Computationally demanding; rare cell subpopulations may be obscured	[52]
		One-SENSE	Assign a manually predefined category (annotation) with specific biological meaning to each t-SNE axis		[55]
Hybrid algorithm (clustering + dimensionality reduction)	Global data structure with partition of subpopulations	ACCENSE	Combine nonlinear dimensionality reduction with density-based partitioning, and displays multivariate cellular phenotypes on a 2D plot.	Loss of single cell resolution	[56]
		Phenograph	Clustering of cells from nearest neighbor graph generated from original high-dimensional space then present cell clusters in 2D t-SNE plot		[57]
Seriation-based analysis	Cell state progression with pseudo-temporal order	Wanderlust	Given a known starting point define most likely linear path	Incapable of dealing with bifurcating trajectory	[59]
		Wishbone	Position single cells along bifurcating developmental trajectories	Not directly applicable to time course dataset	[62]
		Cycler	Construct a continuous trajectory of cell-cycle progression from images of fixed cells		[61]

*Methods are grouped into these categories based upon the type questions they seek to answer. In practice, methods from multiple categories may be required to resolve the biological question.

Received: 31/10/2016; Revised: 20/01/2017; Accepted: 20/01/2017
This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201600267.