EXPANDING THE CAPABILITIES OF MASS CYTOMETRY DATA ACQUISITION AND ANALYSIS

SUBMITTED TO THE DEPARTMENT OF CANCER BIOLOGY
AND THE COMMITTEE OF GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Tyler J. Burns

August 2017

Table of Contents

Chapter 1: Introduction

1.1 The advent of high parameter high throughput single cell analysis

Given that life as we know it is the product of an increasingly complex genetic algorithm iterating for roughly one quarter the age of the universe, it should not be surprising that rich biodiversity is observed at nearly every level of analysis from organisms within a species to single cells within multicellular life. The unmet need to document the latter is rooted not only in our existential pursuit to understand our origins, but also to explain and treat various medical conditions that arise and act at the single cell level.

This has motivated a variety of single cell technologies to meet these ends. Originally, such methods were in two categories, one in which hundreds of thousands of cells could be captured with usually no more than 10 features, and a newer one in which very few cells could be captured but with tens of thousands of features. Only recently have methods been able to merge these two categories, allowing for 50-100 or more features along with hundreds of thousands of cells.

Such methods have required the development of a completely new single cell analysis toolkit and vocabulary. It is no longer sufficient, for example, to conclude that a tumor is "heterogeneous." It is fruitful to test the hypothesis that single cells vary in any given biological study because of an inherent superstructure of the system.

Here, we focus on mass cytometry, a method that emerged from our lab in 2011 providing datasets of 100,000+ cells by 45 parameters. Mass cytometry data is clean and initial studies have covered well-studied biological systems. This has allowed for the development of statistical tools to analyze this particularly novel biological data structure. As such, those studying mass cytometry at this time have the responsibility to set forth the research and statistical paradigms for the remainder of the emerging high-throughput high-parameter single cell analysis methods.

1.2 The niche that mass cytometry fills

1.2.1 The mass cytometry instrument

Mass cytometry is a flow cytometry-based technology that uniquely uses antibodies conjugated to isotope mass reports (typically of the lanthanide series) rather than fluorophores. As a result, the spectral overlap limiting the maximum number of parameters of the latter is no longer a problem. The instrument is specifically named Cytometry by Time-Of-Flight (CyTOF) (1). In brief, single cells stained with these isotope-antibody conjugates are vaporized into ion clouds by inductively coupled argon plasma with a temperature comparable to the surface of the sun (7500K). These ion clouds, one per cell, are directed into a time-of-flight mass spectrometer. The amount of a given protein an antibody was directed against is measured by the amount of the specific metal isotopes within the ion cloud by mass-to-charge ratio (m/z). Mass cytometry is currently capable of detecting over 45 parameters per cell. The theoretical limit of detection is the number of isotopes within the current instrument's detection range of 75m/z to 209m/z, at 134 parameters. CyTOF acquires up to 1000 cells per second, with individual samples therefore able to exceed 10^5 or 10^6 cells with relative ease(2).

1.2.2 Additional acquisition parameters, aside from antibody count

Given that the instrument does not use fluorescence, scatter properties are not possible to obtain. Rather, cells are distinguished from debris using two parameters: a rough DNA count feature using iridium-conjugated DNA intercalators, and a "cell length" feature is given as a function of the size of each ion cloud(2). These intercalators are not sufficiently sensitive to detect stage of the cell cycle as other DNA binding agents like DAPI would, but our lab has done additional work engineering cell cycle specific antibody panels(3,4). An additional computational pipeline that includes palladium cell "barcoding" channels can assist in distinguishing singlets from doublets

(5,6). To account for observed subtle fluctuations in the instrument's performance over the course of a long experiment (upwards of several hours of acquisition), bead standards are used to normalize input data(7). To distinguish live versus dead cells, a one-minute treatment with cisplatin can be used prior to cell processing. The mass instrument detects the internalization of platinum accordingly (8).

1.2.3 Antibody panel design

A panel is defined as the set of all antibodies and other detection reagents to be used in a mass cytometry experiment. Initial studies have focused on the immune system and related cancers. As such, cell surface markers are common input parameters for demarcating specific cell subsets. Expression of transcription factors have served a similar purpose in more recent studies(9,10). For the immune system in particular, there are many pre-existing antibody panels that include surface markers and phosphorylated residues of signaling markers. These can be either purchased from Fluidigm, or obtained by any mass cytometry publication.

The difficulty in panel design comes when one is studying a completely different model system (eg. tissue or tumor). The panel development is iterative. For each antibody, one must determine if the antibody reliably detects the macromolecule of interest. Not every antibody on the market is rigorously tested, and sometimes the only antibodies available for a specific macromolecule were tested against denatured proteins in a western blot. Thus, one needs negative and positive controls (eg unstimulated and stimulated cells for a phospho-protein). Furthermore, one may need to check for proper subcellular localization using immunofluorescence microscopy. We have found, for example, an antibody against Histone H3 with erroneous surface staining in human monocytes (Nolan lab, unpublished). When it is found that the antibody is detecting what it is supposed to, it must be titrated to determine the optimum concentration. Following this step, it is ready for conjugation to an available lanthanide channel.

1.2.4 The "phospho-flow" paradigm applied to mass cytometry

Mass cytometry has been able to utilize a paradigm brought forth from our lab's initial "phospho-flow" studies, in which one measures the change in the phosphorylation state of a particular signaling protein in response to *ex vivo* perturbation(11). In the simplest experimental scenario, one tube of cells is left untreated and one tube is treated with a chemical or environmental agent. These tubes are both stained with the same cocktail of antibodies against phospho-proteins expected to change as a result of the stimulation. After running the cells through a flow cytometer, the levels of a given phospho-protein are compared between the untreated and treated data. In the case of mass cytometry, one can use many surface markers to demarcate different cell populations, such that this "fold-change" analysis can be performed across each individual population within a complex system (eg. blood). With mass cytometry's ability to deeply subset diverse populations, this phospho-flow paradigm has expanded our understanding of how very particular cell subsets respond to the environment.

1.2.5 Expanding what mass cytometry is capable of measuring

Along with expanding the number of features that can be detected, significant effort is being spent expanding the types of features that can be observed per cell. Initially, features were limited to proteins for which reliable antibodies had been developed. These antibodies were further divided into those that bound to a specific protein, and those that bound to a post-translationally modified protein (eg. a phosphorylation at a specific residue). An initial push in our laboratory was toward the measurement of protein-protein interactions, specific nucleic acid sequences, and subcellular localization using the mass cytometry platform. A method initially developed for immunofluorescence microscopy, the *in situ* proximity ligation assay (PLA), was able to unify these three pursuits(12). This method produces a detectable signal if two specifically modified antibodies or transcripts are spatially within 40nm of each other.

Our lab adapted PLA to CyTOF in a collaboration with the method's inventors, Ola Söderberg, Ulf Lendrigan and colleagues at Uppsala University in Sweden. We initially used this method to quantify the abundance of specific RNA transcripts simultaneously with traditional protein-targeted antibodies through mass cytometry(13). Given that mass cytometry with simple antibody staining cannot provide positional information of a given protein within a cell, we further utilized the method to achieve this unmet need, a method we aptly named Subcellular Localization Assay (SLA)(14). This allowed us to obtain direct readouts of cell signaling pathway activation through measuring levels of nuclear import of a transcription factor rather than its change in phosphorylation levels of a specific residue, with the expanded throughput and parameters that mass cytometry has to offer as compared to immunofluoresence microscopy. Thus, the mass cytometry "phospho-flow" paradigm now has a more direct readout of cell signaling activity. Additional efforts are being made to use a hybrid of the RNA detection system and SLA to measure epigenetic marks of specific DNA sequences (Nolan lab, unpublished).

1.3 How to analyze high parameter high throughput single cell data

1.3.1 Initial processing

The output of a mass cytometry experiment is an fcs file, readable by standard flow cytometry analysis platforms including FlowJo, Cytobank, CYT in Matlab, and FlowCore in R. Following aforementioned normalization and de-barcoding of these fcs files, the data is by convention transformed with the inverse hyperbolic sine function using a cofactor of five, as this transformation is able to account for raw values that are less than zero(2).

Assuming a given scientist chose input parameters that are individually well studied and relevant to the biological questions of interest, a 45-parameter dataset can provide both unprecedented systems-level intuition and predictive power for a given biological system. Historically, flow cytometry data has been measured through the use of biaxial plots, and manual gating of the cell populations therein. For a simple 4-parameter flow cytometry experiment, the number of biaxial plots to be analyzed is 6. For a 45-parameter mass cytometry experiment, the number of biaxial plots to be analyzed is 990. Analyzing 990 biaxial plots is not temporally efficient for a human, no matter how well trained. Thus, there initially was and still is an unmet need for computational methods to analyze these increasingly complex datasets. We will focus on methods that subset cells for the purpose of characterizing the populations that exist within a biological system.

1.3.2 Clustering and cluster visualizations

One major goal of single cell analysis is to characterize the specific cell subsets that exist within a diverse population. For high-parameter datasets, this can be achieved computationally through clustering. Clustering is an unsupervised machine learning method that categorizes uncategorized data points (in our case, cells) into two or more specific groups. Prior to CyTOF, clustering had been used extensively with the analysis of microarray data (15,16). There are many different algorithms that do this, and an exhaustive comparison for CyTOF analysis is beyond the scope of this dissertation. I therefore point the reader toward the following Flow-CAP project, which discusses a number of them, their evaluation metrics, and which are more appropriate for which datasets (17).

In general, each parameter in the mass cytometry dataset is treated as a spatial dimension. Just as 2-parameter data can be treated as coordinates (x, y) denoting where a cell lies on a two dimensional plane, a 45-parameter data can be treated as 45 coordinates (x, y, z, … n) per cell denoting where a given cell lies in 45 dimensional space. A given clustering algorithm groups cells that are similar to each other, in our case, in the original high dimensional space. Of note, clustering can be performed on lower dimensional embeddings of the original high-dimensional manifold as well (see *single cell visualizations*)(18). However, there is no current consensus as to when lower dimensional embeddings are more appropriate from the original manifold to use for clustering analysis, and this is an area of active research.

Once an experimenter has obtained a set of clusters from the data, additional statistical analyses can be performed per cluster, with a prominent case being the previously described fold change of a phosphoprotein after *ex vivo* stimulation. The next step is to visualize these clusters in the two dimensions allowed for the figures of a publication or three dimensions allowed for the human brain. One example of this is the minimum spanning tree, which we originally used for mass cytometry analysis in a method called SPADE(19). For model systems where the cells therein are expected to be structured like a branching tree, like bone marrow, minimum spanning trees provide an intuitive "bird's eye view" of the structure of a given dataset. Additional cluster visualization tools utilize force-directed graphs as the final output, removing some of the bias a minimum spanning tree would put onto the system. For example, a method developed in our lab called Flow-Map is typically used in time-course data, giving clusters attractive and repulsive forces based on similarity to each other in high-dimensional space(10). Another method developed in our lab called Scaffold also uses force-directed graphing but utilizes user-generated “landmark” populations based on manual gating. These populations are placed on a two-dimensional map. Here, physical forces are applied to the clusters as a function of the similarity to any given landmark population and to each other. The end result is a graph that can generate a reference map across multiple populations to determine cell subset abundance changes (among other things) in an unbiased manner across a multitude of comparisons (9).

Of note, these clusters can be used to make statistical inference. One method out of our lab, called Citrus, performs multivariate regression for each cluster within CyTOF data to make clinical predictions(20,21). This type of statistical inference has been added to the aforementioned Scaffold paradigm as well(22).

1.3.3 Single-cell visualizations

A two or three-dimensional embedding of high dimensional data is often done in with single cells. This allows the observer to make his or her own conclusions about the topology of the data without being subject to the bias introduced by clustering. As with clustering, there are many ways to do this.

One computationally efficient dimensionality reduction method developed well over a century ago is Principal Components Analysis (PCA)(2). PCA is an eigenvector-based multivariate analysis that finds the first n vectors that explain the most variance in the data matrix. For visualization, the first two principal components are typically used, and cells are plotted as a function of those two vectors. However, PCA assumes a linear relationship between all parameters, which is not always the case in biology.

Furthermore, the first two principal components often only explain a small percentage of the variance of the data matrix, resulting in substantial loss of information. Nonetheless, PCA as a visualization strategy is a very quick way to get preliminary intuition around one's data.

A popular method particularly for mass cytometry that emerged more recently is called t-distributed stochastic neighborhood embedding (t-SNE) (23). This method when applied to mass cytometry was named visualization of t-SNE, or viSNE, in collaboration between our lab and the lab of Dana Pe'er (24). t-SNE converts the high dimensional single cell data into a probability distribution, and iteratively produces a probability distribution in lower dimensional space that minimizes a particular cost function. The use of a t distribution has the property of "clumping" the two-dimensional embedding, which accentuates the existence and separation of cell subsets. A side-effect of this property is that the relative cell density in t-SNE space does not necessarily recapitulate the relative cell density of high dimensional space, though the separation of cell subsets as described on the map is accurate. Unlike PCA, t-SNE is computationally expensive, making datasets of greater than 10^5 cells difficult due to runtime. Furthermore, datasets of this size are not as well visually separated than smaller datasets on the order of 10^4 cells (unpublished). Nonetheless, t-SNE remains popular among mass cytometry users because of its accessibility and visually appealing intuitive readouts.

As previously introduced with clustering, force-directed graphs can be used to visualize single cell data. Here, single cells rather than clusters are used as the data points for which attractive and repulsive forces will be applied as a function of their coordinates in high dimensional space. The algorithm typically runs until the cells reach stasis in the two dimensional map (or when the user stops it manually). When using force-directed graphs on mass cytometry data of mouse bone marrow, the algorithm converted upon a branched layout expected from prior knowledge about the model system's structure(25).

For datasets assumed to follow a uniform trajectory over time without branching, an algorithm was developed in a collaboration between our lab and the lab of Dana Pe'er that finds each cell's position in the trajectory. This algorithm, called Wanderlust, requires user input to define the starting point of the trajectory, and then iteratively identifies the shortest path through the dataset's phenotypic space. These values allowed for a clear picture of exactly which markers change along the arrow of pseudo-time(26). Marker levels of interest can be plotted as a function of pseudo-time, or aforementioned two-dimensional single cell visualizations can be applied and colored by Wanderlust value to make inferences about the flow of information (stasis, bottlenecks) through pseudo-time.

1.3.4 Merging clustering and single-cell visualizations with a new paradigm

A fundamental problem with the single cell visualization methods above is that one cannot make biological comparisons across concatenated datasets (eg. fold-change of a phospho-protein after stimulation). This is because such comparisons require binning of the dataset. However, visualization of such comparisons in a single cell manifold embedding, such as t-SNE, would provide an intuitive readout about these comparisons that could in turn inform how one should best partition the single cells. This would be especially useful for many of the Proximity Ligation Assay-derived parameters, wherein there is often signal in baseline state that one wants to compare with a stimulated state with an expected increase in signal (eg. nuclear localization of NFkB). As such, we developed Smooth Comparison Over Nearest Neighborhoods (SCONE), which builds exhaustive bins around each cell consisting of its k-nearest neighbors in the original space of markers not expected to change between conditions (typically surface markers). From here, one can represent the value of each cell in the dataset as the biological comparison of interest made between its k-nearest neighbors in a concatenated dataset. This can range from statistical tests to cell-abundance changes between multiple biological conditions. These values can then be encoded as colors on a t-SNE map or other dimension-reduction map, allowing one to visualize both continuous and discrete chagnes across biological conditions. The computational section of this dissertation revolves around re-analysis of several landmark mass cytometry datasets within this paradigm.

1.4 Emerging methods and the future of high parameter high-throughput single cell analysis

Two emerging single cell method categories are on the heels of mass cytometry to provide similar throughput and parameters. The first category is imaging, wherein two methods, MIBI(27) and CODEX (Nolan lab, unpublished), allow immunofluorescence and immunohistochemistry data with parameters exceeding 50-100. The second category is single-cell sequencing, where only recently the number of cells able to be sequenced at any given time has exceeded 10,000, which makes mass cytometry analysis pipelines amenable to this type of data(28).

With respect to imaging, there is a rich abundance of spatial features that mass cytometry and single cell sequencing cannot provide, such as categorizing the physical neighbors of a given cell (Nolan lab, unpublished). If these new features can be represented as numerical features on the data matrix (and our lab has found that they are), then these technologies can utilize mass cytometry pipelines established for clustering (eg. SPADE, Scaffold, Citrus) and single-cell visualizations (eg. t-SNE, SCONE).

Single cell sequencing has the unique challenge of having a data matrix with 1-2 x 10^4 features rather than the 30-100 used for mass cytometry, MIBI, and CODEX. Increasing dimensionality leads to increased sparsity of the dataset, often referred to as the “curse of dimensionality”. Here, t-SNE has become popular for visualizing such datasets. However, one must perform t-SNE on the first ~50 principal components rather than the original high-dimensional manifold to combat the curse of dimensionality. As such, the final visualizations are somewhat harder to interpret than the t-SNE output from a mass cytometry dataset. Nonetheless, single cell sequencing will only further improve both in throughput and quality, and it would be of great benefit to adapt established mass cytometry analyses to this type of data. As such, rigorous study needs to be done regarding dimensionality reduction methods to optimally preserve the information content of the original manifold for this specific data type. These methods will likely become relevant to mass cytometry and the high parameter imaging methods as the respective maximum number of parameters per cell continues to increase.

These new methods, along with mass cytometry, will allow researchers to tackle fundamental questions about the diversity of both development and disease. Given the similarity of the data structures across these platforms, we expect them to benefit from cross-pollination within these computational analysis paradigms.

Chapter 2: High-throughput precision measurement of subcellular localization in single cells

2.1 Abstract

To quantify visual and spatial information in single cells with a throughput of thousands of cells per second, we developed SLA (Subcellular Localization Assay). This adaptation of Proximity Ligation Assay expands the capabilities of flow cytometry to include data relating to localization of proteins to and within organelles. We used SLA to detect the nuclear import of transcription factors across cell subsets in complex samples. We further measured intranuclear re-localization of target proteins across the cell cycle and upon DNA damage induction. SLA combines multiple single-cell methods to bring about a new dimension of inquiry and analysis in complex cell populations.

2.2 Introduction

Cells can efficiently respond to a dynamic environment by re-localizing proteins both between and within intracellular compartments. Thus, quantifying localization of proteins to specific intracellular structures is fundamental for understanding cell behavior, both in normal and diseased conditions.

Immunofluorescence microscopy (IFM) is often used for obtaining such information. With IFM one can visually estimate co-localization of a protein with intracellular structures provided there exist antibodies or dyes for each, though throughput is typically low. In addition, imaging flow cytometry(29) has been a useful addition to the field with a much higher throughput than IFM (up to 5000 cells per second), but currently limited availability.

It would be optimal to obtain such information with traditional flow cytometry, a well-entrenched technology with throughput of more than 10,000 cells per second, and a much wider availability in both hospitals and laboratories (30). Within flow cytometry, one can use phosphorylation of specific proteins as an approximation for their activation (e.g. phosphorylation of a transcription factor associated with nuclear localization), provided one has an antibody for a phosphorylation site sufficiently studied to make such an assumption (11). Thus, a flow cytometric readout of a single phosphorylation site of a specific protein may only provide limited information about its varied activation states, and is therefore not always a suitable proxy for localization. Moreover, protein localization to “protein neighborhoods” within cells is important. The presence of two or more proteins in a given locale is often an indicator of a series of mechanistically determined events whose consummation is the goal of the machine being built.

Other than nanoscale imaging, proxies have been developed to indicate locale. One approach used for over 30 years is fluorescence resonance energy transfer, either with chemical or genetically encoded fluorophores (31). This requires previous tagging of the molecules in question and might interfere with their supposed functions. Another method involves “splitting” of enzymatic functions into separate sub-proteins (32). Again, this involves most often the creation of genetic fusion events. In each of these cases the co-localization of proteins in the cell creates an “event” that can thus be read. Finally, a method recently described allows for rough approximation of protein location within a given cell by analyzing pulse width and height of a fluorescently labeled protein with flow cytometry(33). However, one may want a quantitative readout based not on location within a cell or phosphorylation, but rather from molecularly tagging these said events at specific subcellular structures of interest by relative proximity without creating genetic fusion events.

Here, we utilized the Proximity Ligation Assay(12) method to measure proximal co-localization of specific proteins to specific subcellular compartments with flow cytometry. This adaptation, herein termed Subcellular Localization Assay (SLA), is extensible to IFM, CyTOF, MIBI, and other detection systems. SLA quantifies localization of any given protein, for which there exists a representative antibody, and a second molecular tag for the subcellular structure with which the first protein interacts. The system is compatible with simultaneous detection of additional cell surface or intracellular markers in primary cells. Importantly, the method does not require instrument modifications or new analysis software.

We first measured nuclear import with flow cytometry by detecting proximity of antibodies against transcription factors to a previously validated antibody against double-stranded DNA(27). We next measured DNA repair by re-localization of nuclear proteins to sites of DNA damage. This was achieved by detecting proximity of antibodies against DNA repair protein BRCA1 to antibodies against DNA damage marker γH2AX. Changes in localization were quantified by the increase or decrease in the observed SLA signal. Here, we performed SLA simultaneously with a DAPI stain for cell cycle analysis. Combining quantitative and high throughput measurements of subcellular localization with protein function in primary cells provide opportunities for understanding basic cellular mechanisms with implications in health and disease.

2.2 Methods and Materials

2.2.1 Cell lines and samples

All cell lines described below were of human origin. Non-adherent cell lines (U-937, THP-1, Jurkat) were purchased from ATCC (Manassas, VA, United States). U-937 and THP-1 cell lines are monocytic, and the Jurkat cell line is T-lymphocytic. These cell lines were cultured in Dulbecco's RPMI-1640 (Life Technologies, Carlsbad, CA, United States), with 10% FBS (Thermo Fisher Scientific, Waltham, MA, United States), 1% Penicillin/Streptomycin (Life Technologies) and 1% Glutamine (Life Technologies) added, maintaining a density on average between 500K and 1M cells per mL. The TYK-nu cell line was derived from an ovary with undifferentiated carcinoma. It was obtained from GCRB (Glasgow, Scotland) for our use. TYK-nu cells were cultured in Eagle's Minimum Essential Medium (ATCC) with 10% FBS and 1% Penicillin/Streptomycin added. All aforementioned cell lines were cultured at 37°C in a 5% CO_2 atmosphere.

Human peripheral blood was obtained from the Stanford University Blood center from anonymous healthy human donors. Collection procedure followed a Stanford University Institutional Review Board-approved protocol. SLA experiments used peripheral blood mononuclear cells (PBMCs) isolated using Ficoll Plaque Plus (Thermo Fisher Scientific). For these experiments, PBMCs were used fresh.

2.2.2 Cell stimulation, treatment, and processing

U937 and THP-1 human monocytic cells were stimulated with 10 ng/mL recombinant human TNFα (R&D Systems, Minneapolis, MN, United States) for 15 minutes, and Jurkat cells as well as human PBMCs were stimulated with 250 nM PMA (Sigma-Aldrich) and 1 μM ionomycin (Sigma-Aldrich, St. Louis, MO, United States) for 60 minutes in complete RPMI. PBMCs were stimulated with 50 ng/mL TNFα or 5 μg/mL ultrapure LPS (InvivoGen, San Diego, CA, United States) for times specified in complete RPMI. Cells were incubated gentle shaking at 37 °C. TYK-nu cells were treated with 10Gy of γ radiation using Cesium 137 at a dose of 8gy/min. Following irradiation, cells were incubated at 37 degrees for 6 hours before they were processed.

Following pathway stimulation treatments described above, cell lines were fixed at a density of 1 x 10^6 cells/mL, and PBMCs were fixed at a density of 5 x 10^6 cells/mL. Fixation occurred in 1.6% paraformaldehyde (Electron Microscopy Services, Hatfield, PA, United States) for 10 minutes at room temperature. Of note, all paraformaldehyde solutions described in this manuscript came from 16% stock samples, diluted into relevant cell culture media. Following fixation, cells were permeabilized in 100% methanol (Thermo Fischer Scientific) on ice for 15 minutes. These fixation and permeabilization conditions are a standard for our lab, as previously described (2).

For DNA damage experiments, irradiated cells were pre-extracted with 0.5% NP-40 (Abcam, Cambridge, England) in PBS (Life Technologies) for 5 minutes on ice, and fixed in 4% paraformaldehyde (Electron Microscopy Services) for an additional 20 minutes at room temperature as described previously (34). The higher percentage of paraformaldehyde was used to counteract the increase in cell loss observed in pre-extracted cells during each wash step. Pre-extracted cells were not permeabilized with methanol.

2.2.3 Preparation of proximity probes

Donkey anti-mouse and anti-rabbit secondary antibodies (Jackson Immunoresearch Labs, West Grove, PA, United States) were conjugated to a heterobifunctional sulfo-SMCC linker (Thermo Fischer Scientific). The linker was added at a molar excess of 25 to 1 with an antibody concentration of 1.3 mg/mL. Samples were incubated at room temperature for 45 minutes. Unconjugated SMCC linker was removed using 50-kDa Centricon filters (Thermo Fisher Scientific). Samples were filtered twice with addition of PBS after a spin at 12,000 g for 10 minutes. Samples were resuspended in 50 µL in PBS.

Oligonucleotides with a C6-thol modifier conjugated on the 5' end (Stanford Protein and Nucleic Acid facility, Stanford, CA, United States) were conjugated to the respective antibody-SMCC linker conjugates. In parallel to the antibody-SMCC incubation step, oligonucleotides were deprotected using 5 mM TCEP (Thermo Fischer Scientific) in 0.5X PBS for 20 minutes at 37 °C. TCEP was removed by precipitation from 0.3M sodium acetate and 70% ethanol. Oligonucleotides were resuspended in PBS with 1 M NaCl. Oligonucleotides were added at a 5 to 1 molar excess to 1.3 mg/mL SMCC-conjugated antibody and incubated overnight at 4 °C. Samples were then purified through Centricon filters as described above. PBS-based antibody stabilizer (Boca Scientific, Boca Raton, FL, United States) was added until the antibody concentration was approximately 1 mg/mL. In total, this procedure takes approximately one and a half hours of physical labor on day one, an overnight incubation, and another 30 minutes for the purification step on day two. Efficacy of antibody-oligonucleotide conjugation was evaluated by SDS-PAGE using Simply Blue Safe Stain (Life Technologies) for protein detection and SYBR-Gold (Life Technologies) for DNA detection. Each conjugate was further evaluated by SLA for the dsDNA:histone H3 interaction over a range of concentrations to optimize signal-to-noise ratio.

2.2.4 SLA protocol

For all experiments with PBMCs, cells were barcoded according to treatment condition with different concentrations of Pacific Orange NHS fluorophores (Thermo Fischer Scientific) as previously described (35) . Cells were then placed into PCR tubes for antibody staining and subsequent steps at a density of 1 million cells per 100µL. The following primary antibodies were used in this study: CD45 (Biolegend, clone H130, San Diego, CA, United States) 2 µg/mL, dsDNA (Abcam, clone 35I9) 0.5 µg/mL, NF-κB (Abcam, polyclonal) 2 µg/mL, histone H3 (Abcam, polyclonal), 2 µg/mL, NFAT (Cell Signaling Technology, Danvers, MA, United States), 1:200 dilution – mass not specified, Cytochrome C (Cell Signaling Technology), 1:500 – mass not specified, COXIV (Cell Signaling Technology), 1:500 – mass not specified, H2AX, pSer139 (Millipore, clone JBW301, Darmstadt, Germany), BRCA1 (Santa Cruz Biotechnology, clone C20, Dallas, TX, United States). Barcoded samples were incubated with the antibodies at 4°C overnight (approximately 15 hours) in PBS with 5mg/mL BSA (Santa Cruz Biotechnology) and 0.02% sodium azide (Sigma-Aldrich). Cells were subsequently washed three times with PBS containing 5mg/mL BSA and 0.02% sodium azide, and secondary antibodies conjugated to proximity probes were added. The master mix contained 100 μg/mL sheared salmon sperm DNA (Life Technologies) and 3 μg/mL proximity probes (anti-mouse and anti-rabbit) in the aforementioned wash buffer. Cells were incubated for 1 hour at room temperature on an inverter. For subsequent steps to the end, cells were washed in PBS with 0.1% Tween. Following secondary antibody incubation, 100 nM backbone and 100 nM insert oligonucleotides with 10 μg/mL of salmon sperm DNA were added in PBS with 0.1% tween. Cells were incubated for 30 minutes at 37 °C. Next, the oligonucleotides were ligated in 1x ligation buffer (Thermo Fisher Scientific), T4 DNA Ligase 10 U/mL, and 10 μg/mL sheared salmon sperm DNA. Following this, rolling circle amplification of the circularized oligonucleotide product was performed in 1x phi29 polymerase buffer (Thermo Fisher Scientific), 125 U/mL phi29 polymerase (Thermo Fischer Scientific), and 250 μM each dNTP (Thermo Fisher Scientific). Samples were incubated for 90 minutes at 37°C. We determined that longer amplification times (as long as overnight) could be used for interactions that produce low signals, After three washes with PBST, detection oligonucleotides labeled with Alexa 647 were added for a final concentration of 200 nM (Olink Biosciences, Uppsala, Sweden). It was also determined that this surface/intracellular staining step could occur following the addition of the secondary oligonucleotide conjugated probes in the beginning of the procedure, but the strength of signal was no different (data not shown), and the simultaneous staining with the detection reagents saved time. For experiments containing PBMCs, the following fluorescently conjugated antibodies were added for additional surface and intracellular staining: IκBα Alexa 488 (Cell Signaling Technology, clone L35A5), CD3 PE (BD Biosciences, Clone HIT3A, San Jose, CA, United States), CD7 Alexa 700 (BD Biosciences, Clone M-T701). Samples were incubated at 37 °C for 30 minutes. For DNA damage experiments, one subsequent step was added, wherein cells were incubated with 0.25ug/mL DAPI (Sigma-Aldrich) for 30min. In total, the SLA procedure described above takes approximately one hour for cell prep and primary antibody staining on day one, followed by an overnight incubation, followed by six hours of on the second day for the remaining steps and flow cytometry.

2.2.5 Data acquisition and analysis

Following SLA, cells were analyzed on an LSRII flow cytometer (BD Biosciences), equipped with 405, 488, and 633nm lasers. All flow cytometry data was subsequently analyzed using Cytobank software (Mountain View, CA, United States). For PBMCs, compensation was performed using Protein A/G bead standards for all antibodies used. For Pacific Orange dye, a mixture PBMCs with and without the dye was used. A compensation matrix was made within Cytobank. All images were acquired with a Marianas Spinning Disk Confocal microscope (Zeiss, Oberkochen, Germany), using the aforementioned primary antibody clones followed by incubation with fluorescent secondary antibodies conjugated to Alexa-fluor 488 and Alexa-fluor 647 fluorophores (Invitrogen, Carlsbad, CA, United States) for 30 minutes at room temperature. Following this, cells were counterstained with Hoechst (Life Technologies) for 5 minutes for cell nuclei. Cell samples (10 μL) were pipetted into wells of a Lab Tek chamber slide (Thermo Fischer Scientific), incubated in the dark for 10 minutes to allow cells to sink to the bottom of the slide, and imaged using the SlideBook 6.0

software (3i, Denver, CO, United States). Images were further processed using ImageJ software (National Institutes of Health, Bathesda, MD, United States). Bar plots were constructed using the ggplot2 R package.

2.2.6 Statistics

Statistical tests were performed using the stats R package. Specifically, the Welch Two Sample t-test was used, and all tests were two-tailed. For time-course experiments shown in Figure 3, data were transformed by the inverse hyperbolic sine (arcsinh), and therefore compared in arcsinh space. This is similar to a log transformation done on flow cytometry data to make relationships between expression levels and biological conditions more clear. The arcsinh transformation and its comparison to log transformation is described in previous work from our lab (2), and additional work comparing different data transformations commonly used in flow cytometry (36).

2.3 Results

2.3.1 The SLA approach

SLA uses the Proximity Ligation Assay(12,37),(38),

(13,39-41) to bring protein subcellular localization, normally accounted for by visual and spatial observation, to the traditionally non-visual flow cytometer. Proximity ligation assays traditionally measure protein-protein interactions via antibodies against the respective proteins of interest. Here, this method is adapted such that one antibody is against an abundant macromolecule marking an organelle or cell structure of interest. These antibodies are bound by oligonucleotide-conjugated secondary antibodies used for proximity detection (Figure 1A, left panel). If the two secondary antibodies are within 40 nanometers of each other(12), the reaction proceeds and the product can be measured either by microscopy or flow cytometry with fluorophore-conjugated detection oligonucleotides (Figure 1A, middle and right panel).

2.3.2 Measuring transcription factor localization to organelles

In initial experiments, SLA was used to measure the nuclear import, and presumed DNA binding, of the p65/RelA subunit of transcription factor NF-κB by quantifying its interaction with an antibody against double-stranded DNA (dsDNA) used previously(27). In these experiments, proximity probes for the interaction between CD45 and Histone H3 (Figure S1A) (not expected to interact), and proximity probes for the expected interaction between dsDNA and Histone H3 (Figure 2) were used as controls, where the signal was not expected to change between untreated and treated conditions. Nuclear import of p65/RelA (which will be referred to as NF-κB) was induced by treatment with NF-κB pathway activator tumor necrosis factor alpha (TNFα) (Figure 2B, Figure S1A). SLA detected an increase in nuclear NF-κB for both cell lines and gated monocytes from human peripheral blood mononuclear cells (PBMCs), consistent with what was observed by IFM of traditional fluorescent antibody staining (Figure 2A).

Additionally, SLA was used to measure nuclear import of transcription factor NFAT in the Jurkat T cell line upon combined treatment of NFAT pathway activators phorbol-12-myristate-13-acetate (PMA) and Ionomycin (Figure S2). Here, treatment with PMA and Ionomycin led to a strong increase in nuclear NFAT SLA signal both in Jurkat cell lines and lymphocytes gated out of healthy human PBMCs. Of note, $CD3^+$ T cells appeared to have a unanimous increase in SLA signal, suggesting that the nuclear NFAT translocation behavior is sufficiently similar among the numerous T cell subsets therein that they cannot be distinguished by these conditions. On the contrary, two populations were observed in $CD3^-$ lymphocytes after PMA and Ionomycin, suggesting diversity of nuclear NFAT translocation behavior among the remaining cell subsets (eg. B cells, Natural Killer cells) under these conditions.

To test the efficacy of SLA on other macromolecular structures, mitochondrial localization of cytochrome C was measured using a pair of antibodies against cytochrome C and mitochondrial protein COXIV (Figure 1B, Figure S1B). The NF-κB interaction with Cytochrome C was used as a negative control. Here, a small subset of cells had an increased SLA signal for this interaction (though still very low). Given the weakness of signal, this is likely due to experimental background noise, but it could also be due to NF-κB protein existing in or near the mitochondria of these cells, which has been reported previously (42).

IFM with fluorescent staining of the same primary antibodies validated the results for the aforementioned localizations (Figure S1-S4). Taken together, these experiments demonstrated that SLA measures protein localization to multiple intracellular locales by flow cytometry.

2.3.3 Profiling nuclear localization across cell subsets in primary samples

Given SLA was validated above in cell lines, we leveraged the throughput of the method to interrogate transcription factor nuclear localization across multiple cell subsets in complex primary samples. SLA was adapted for use with primary PBMCs from healthy human donors. Light forward and side scatter properties were maintained by following the SLA protocol, which allowed for singlet and myeloid/lymphoid cell gating. The protocol was adapted to include staining with antibodies that had been previously selected to delineate specific immune cell subsets (Figure S3). In addition to TNFα, bacterial lipopolysaccharide (LPS) was used as a NF-κB pathway activator to induce nuclear import of NF-κB exclusively in monocytes(43).

SLA revealed differences in NF-κB nuclear translocation across cell subsets and between stimulation conditions. While TNFα led to nuclear import of NF-κB in both myeloid and lymphoid cell subsets, LPS led to nuclear import exclusively in monocytes (Figure 3A, 3B). By comparing SLA activity across multiple time points in arcsinh space(2), we observed that NF-κB response kinetics in PBMCs differed across pathway activation conditions and cell types. For example, in both TNFα and LPS treated monocytes, we observed an initial increase in nuclear NF-κB levels in the first 15 minutes after treatment. TNFα treated cells had only a 25% additional increase in nuclear import between 15 minutes and 30 minutes (Figure 3A, Table S1), whereas

LPS-treated monocytes had an additional 70% increase in nuclear NF-κB levels between 15 minutes and 30 minutes (Figure 3B, Table S1). These results suggest that nuclear import of NF-κB in monocytes is more gradual when induced by LPS, as opposed to TNFα. These observations are consistent with and build upon previous nuclear NF-κB kinetics studies done *in vitro* (44). Such differences in NF-κB response kinetics were observed across cell types even within the same pathway activation conditions. Between TNFα-treated cell subsets, myeloid and NK cells exhibited a more rapid response than T cells. Following an initial increase in nuclear NF-κB levels in the first 15 minutes of treatment, NK cells had only a 19% increase in nuclear NF-κB levels between 15 minutes and 30 minutes. In contrast, T cells had an additional 66% increase in nuclear NF-κB levels between 15 minutes and 30 minutes (Figure 3A, Table S1). Taken together, SLA revealed that the kinetics of NF-κB nuclear translocation differ across multiple cell subsets, multiple conditions, and multiple time points in complex primary samples.

The NF-κB pathway is negatively regulated by IκBα, which sequesters NF-κB in the cytoplasm until pathway activation leads to the degradation of NF-κB (45,46). To determine if IκBα showed the expected kinetics relative to p65/RelA release into the nucleus, we simultaneously performed SLA with intracellular staining of total IκBα. We confirmed the inverse relationship between nuclear NF-κB and total IκBα at the single-cell level, as the median fluorescence intensities of the former increased and the latter decreased upon treatment with TNFα or LPS (Figure 3C). These results demonstrate SLA's ability to interrogate nuclear localization simultaneously with upstream regulators of a cell-signaling pathway.

SLA also quantified NFAT nuclear translocation in PBMCs treated with PMA and Ionomycin, with nuclear import of NFAT being detected in T-cells but not monocytes (Figure S2B). These results taken together with those for nuclear NF-κB suggest SLA can be a versatile determinant of regional localization and complex formation in primary samples.

2.3.4 Measuring intranuclear relocalization to damaged DNA

While transcription factor binding is a global event that occurs across multiple target loci, it was important to determine if SLA could be used to assay for other, less frequent, cellular events. We therefore sought to quantify DNA repair in terms of specific proteins localized to damaged DNA (Figure 1B). Traditional identification and quantification of DNA lesions is accomplished by assaying DNA repair foci with microscopy (Figure S4)(47,48), which is not (under most circumstances) considered a high throughput regime. We focused on the tumor suppressor BRCA1, which forms intranuclear foci both in S-phase and upon DNA damage induction(47) (49,50). BRCA1 is essential for the end-resection step of DNA double-stranded break repair by homologous recombination (51). DNA double-strand breaks are marked by phosphorylation of nearby Histone H2AX proteins at Serine 139 (γH2AX)[3]. As such, when DNA is being repaired by homologous recombination, BRCA1 will localize to the DNA damage site in proximity to γH2AX. Therefore, SLA provides a convenient means to measure this specific DNA repair mechanism and others like it by measuring proximity of specific DNA repair proteins (in this case, BRCA1) to γH2AX at the single-cell level (Figure 4).

It is known that the DNA double-stranded break repair mechanisms are regulated differently at distinct phases of the cell cycle(52). We therefore added a simultaneous DAPI stain (DNA content per cell) which allows for visualization of the cell cycle (Figure 4A)(53). To detect exclusively chromatin-bound BRCA1 in proximity to γH2AX, we utilized a detergent pre-extraction protocol which removes proteins from the nucleus which are not chromatin-bound (34).

We observed a BRCA1-γH2AX interaction signal that was significantly higher in S and G2 phases of untreated cells (Figure 4B, Figure S6, Table S2). These results recapitulated previous IFM observations from foci counting(47). SLA was able to quantify this interaction in 20,000 cells in under a minute. Of note, there appeared to be two peaks in the SLA signal for untreated cells in G1, suggesting that this particular interaction (though relatively low) may vary in a discrete manner across G1 (Figure 4B). Dot plots with SLA signal and DAPI provided a more detailed interpretation of the relationship between the BRCA1-γH2AX interaction by visualizing the cell cycle as a continuum rather than a series of gates (Figure 4C). We further confirmed that the levels of the BRCA1-γH2AX interaction (SLA signal) differ from the individual protein levels of BRCA1 and γH2AX across the cell cycle. This highlights the additional layer of information one can obtain from measuring interactions in this manner (Figure 4, Figure S5).

To induce DNA double-strand breaks and subsequent repair, we treated cells with ionizing radiation (IR) (54). In these irradiated cells, the G1 specific BRCA1-γH2AX interaction was significantly higher than that of untreated cells (Figure 4B, Figure S6, Table S2). This was an unexpected result, given that BRCA1 co-localization with γH2AX as viewed with microscopy is typically observed in S/G2 phase and not G1(47). These data suggest that BRCA1 may be playing a role in IR-specific DNA repair in G1 as well. Furthermore, these data suggest that SLA has sufficient resolution to identify interactions that are either novel or difficult to detect by microscopy.

Taken together, these results demonstrate SLA can provide a high-throughput and quantitative readout of co-localization that can compliment classical lower throughput methods such as IFM-based foci counting. Furthermore, SLA can be enhanced with DAPI staining for cell cycle and detergent pre-extraction for detecting only chromatin-bound nuclear proteins.

2.4 Discussion

SLA enables measurements of spatial localization with a resolution of 40nm and a throughput of thousands of cells per second. SLA can be performed simultaneously with surface and intracellular antibody staining, allowing for interrogation of subcellular localization across multiple subpopulations in complex samples, like human PBMCs.

SLA allowed for the interrogation of pathway activation in terms of transcription factor nuclear localization across tens of thousands of cells. In this study, we identified differences in NF-κB signaling kinetics across cell subsets of human PBMCs stimulated by TNFα or LPS (Figure 2). Furthermore, SLA allows for one to study the relationship between nuclear localization of a transcription factor and activation of upstream regulatory proteins in a signaling pathway, as we investigated with NF-κB and IκBα (Figure 2).

The combination of SLA with surface antibodies allows for this method to be expanded to complex primary samples without the need for cell sorting. Given SLA was optimized in this report in healthy human PBMCs, this method should be readily expandable to study immune signaling dysregulation in disease. Signaling in tissue specimens may be studied with SLA as well, though one must optimize single-cell suspension to retain cell surface markers of interest.

We further used SLA to study the DNA damage response through the proximity of DNA repair protein BRCA1 and DNA damage marker γH2AX across the cell cycle in the TYK-nu ovarian cancer cell line. We showed the cells in S/G2 have higher levels of BRCA1 localized to γH2AX, as compared to cells in G1. This interaction was expected given that BRCA1 plays a role in the end-resection step of homologous recombination repair in S/G2 (55,56). Furthermore, IR treatment led to increased localization of BRCA1 to γH2AX in G1 as well. Our data suggest that BRCA1 could be playing a role in G1-specific DNA repair, such as non-homologous end joining (NHEJ) (56), in ovarian cancer cells.

The protocol modifications specific to studying the DNA damage response have potential for studying additional biological phenomena. SLA was adapted for a simultaneous DAPI stain for cell cycle analysis without the need for cell cycle-specific markers or thymidine analog (eg. BrdU) treatment. This modification allows for study of cell cycle-specific mechanisms, like the shuttling of cyclins in and out of the nucleus
(57,58).

Furthermore, SLA was optimized for compatibility with detergent pre-extraction of cells to study exclusively chromatin-bound nuclear proteins. Thus, one can robustly interrogate complexes and structures across various contexts and across the cell cycle. These readouts have strong potential in clinical settings, where reliance on low-throughput methods such as foci-counting with microscopy is the current gold-standard for measuring DNA repair mechanisms important for targeted cancer therapy (52,59,60).

As needed, SLA readouts of multiple simultaneous interactions could be achieved by using unique backbone and insert sequences for each antibody pair of interest that will bind detection oligonucleotides with different fluorophores, as demonstrated in recent work from our lab (13). Taken together, by adapting Proximity Ligation Assay to study subcellular localization with flow cytometry, one can interrogate a variety of biological phenomena with quantitative single-cell resolution and high throughput, including but not limited to transcription factor dynamics and DNA repair.

2.5 Acknowledgements

We would like to acknowledge Alex Loktev in the lab of Peter Jackson at Stanford University for his assistance with the confocal microscopy in this manuscript, and the laboratory of Peter Jackson for allowing us to use their confocal microscope accordingly. We would like to acknowledge Alan Ashworth at UCSF, for his expert feedback throughout the DNA damage section of this project and Markus Covert for his expert feedback on the NF-κB signaling work.

2.6 Author Contributions

T.J.B. conceived and designed the method, drove the project's direction, and wrote the manuscript. A.P.F. and P.F.G. adapted the general Proximity Ligation method to flow cytometry for SLA. F.A.B. designed PBMC time-course experiments. J.E.B.

conceived and validated the idea to study the relationship between nuclear NF-κB and IκBα. Y.Y adapted SLA for use in PBMCs. J.M.Y. validated idea to study the relationship between nuclear NF-κB and IκBα. A.R.G validated cell stimulatory conditions for foundational SLA experiments. S.C.K. performed time-course SLA experiments in PBMCs. V.D.G designed experiments and validated antibodies for SLA directed at DNA repair. W.J.F designed experiments for SLA directed at DNA repair and edited the manuscript. G.P.N supervised the work and wrote the manuscript.

2.7 Figures

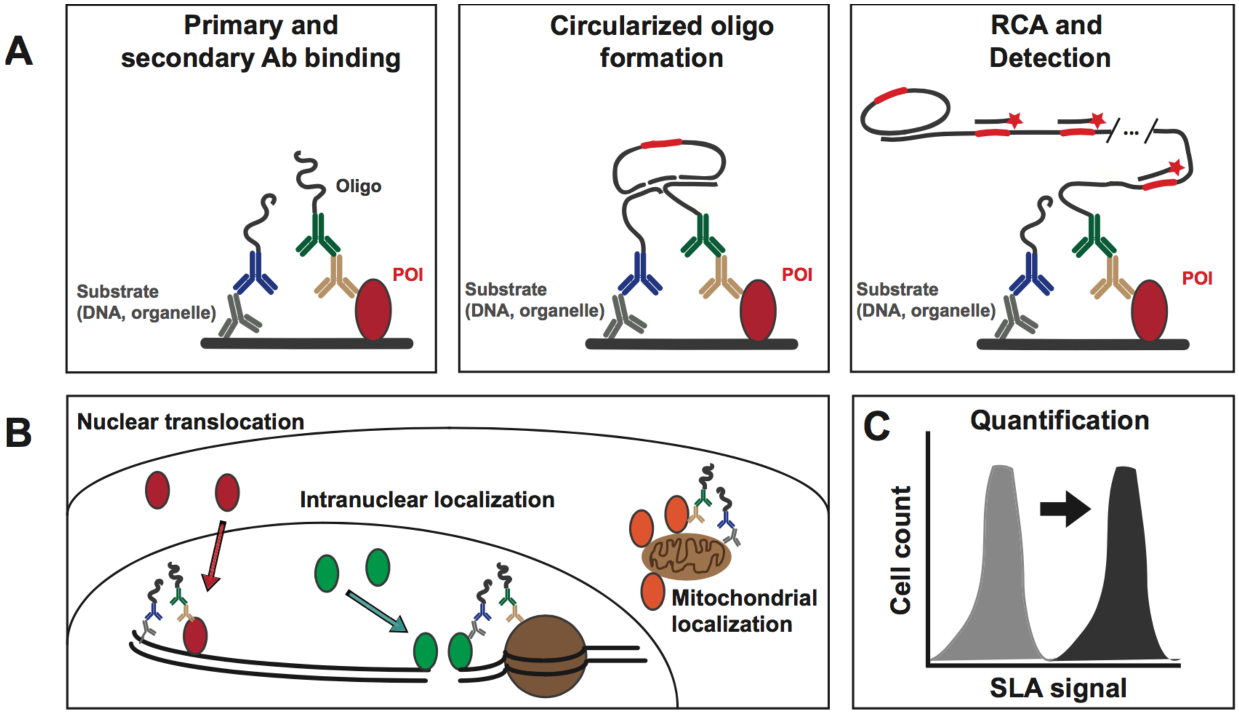

A
Primary and secondary Ab binding
Oligo
POI
Substrate (DNA, organelle)
Circularized oligo formation
POI
Substrate (DNA, organelle)
RCA and Detection
POI
Substrate (DNA, organelle)
B
Nuclear translocation
Intranuclear localization
Mitochondrial localization
C
Quantification
Cell count
SLA signal

Figure 1: SLA detects localization to distinct subcellular structures in single cells

(**A**) Working principle. Antibodies (Abs) to protein of interest (POI) and abundant material in organelle of interest are bound by oligonucleotide (oligo)-conjugated secondary antibodies. When secondary antibodies are in proximity, oligonucleotides can be circularized, amplified, and detected by fluorophore-conjugated probes. Dotted line indicates the length of the amplified region is much greater than depicted. (**B**) SLA has been optimized to detect nuclear localization, localization to specific regions in the nucleus (damaged DNA), and mitochondrial localization, (**C**) leading to flow cytometric readouts of these aspects of subcellular localization.

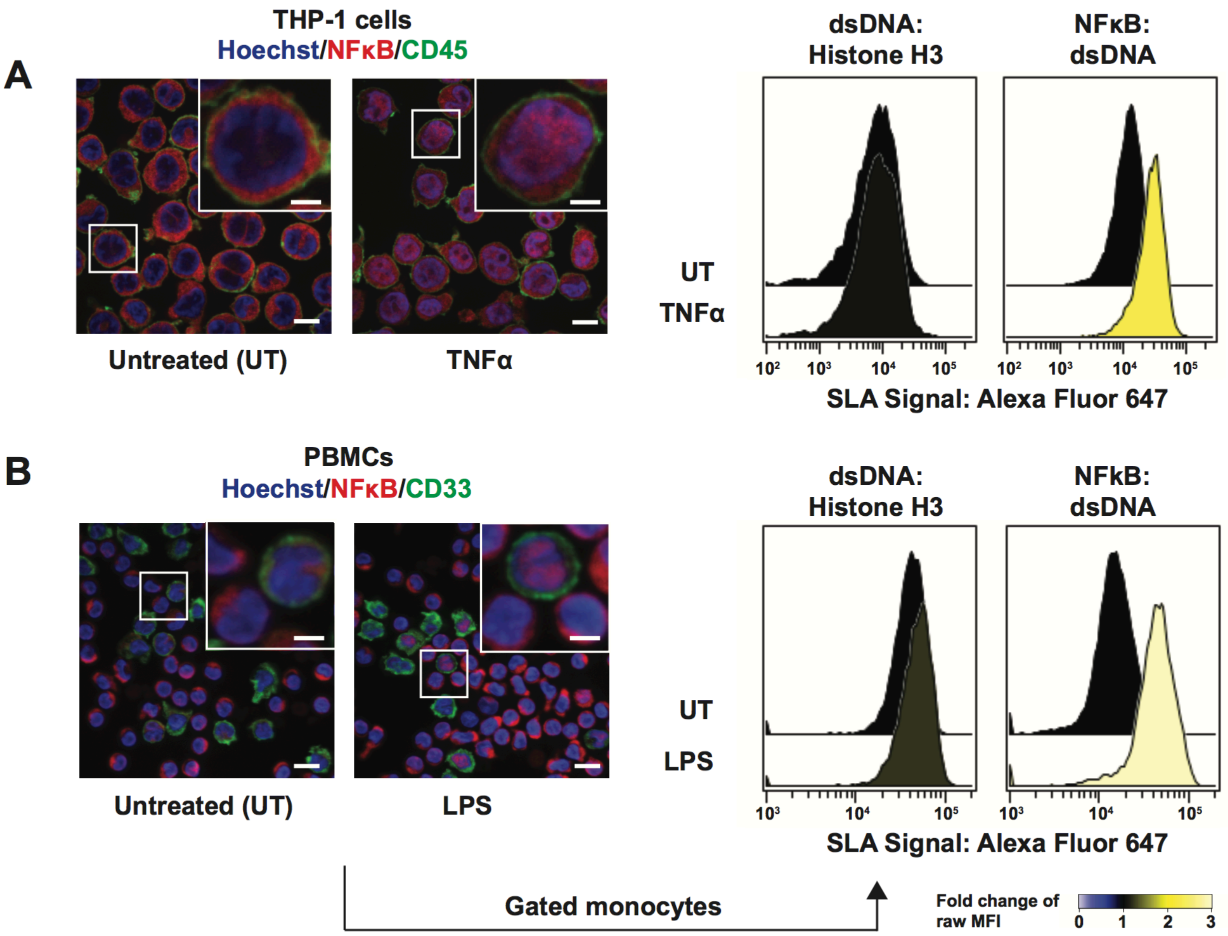


Figure 2: SLA detects nuclear import of NF-κB in cell lines and primary samples.

(**A**) Confocal microscopy (left) and SLA (right) in THP-1 monocytic cell line. Magnified image in upper right is of boxed cells. DNA was stained with Hoechst (blue), and CD45 (green) was used as a cell surface marker. SLA readouts are for the double-stranded DNA (dsDNA)-histone H3 interaction (not expected to be affected by TNFα), and NF-κB and dsDNA interaction in THP-1 cells. (**B**) Confocal microscopy (left) and SLA (right) for NF-κB in PBMCs, illustrated as in (**A**) with CD33 is used as a myeloid cell marker. Scale bars represent 10μm (main images), 4 μm (insets).

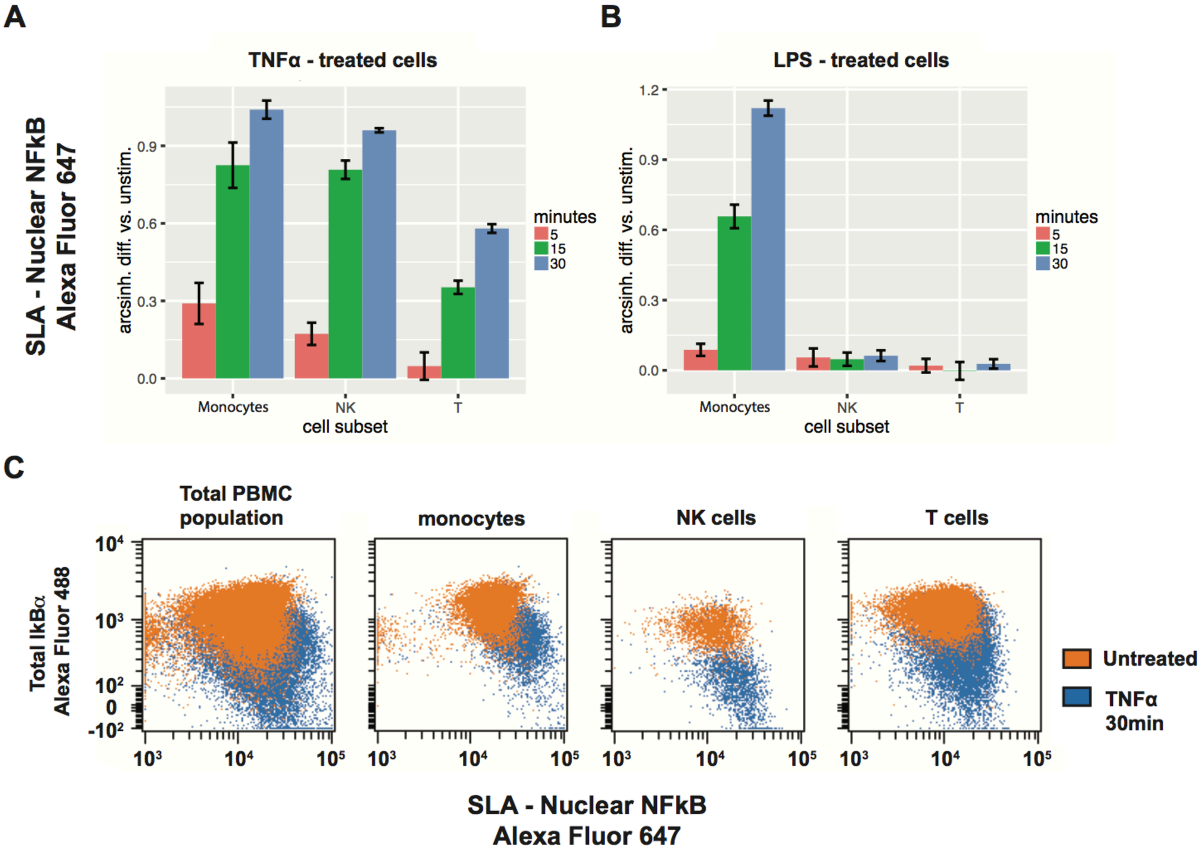


Figure 3: NF-κB nuclear import kinetics and its relationship with total IκBα in single cells across multiple cell subsets in PBMCs.

(**A**) Time course experiment in which PBMCs were treated with TNFα for 5, 15, or 30 minutes. SLA for the NF-κB and dsDNA interaction is indicated as “nuclear NF-κB”.

NF-κB nuclear translocation was calculated as the difference of inverse hyperbolic sine (arcsinh) medians of the indicated timepoint post-treatment compared to that timepoint's untreated control. Bars represent the mean ± SEM (n = 4) (**B**) Same experimental setup as (**A**), but using LPS as the pathway activator. (**C**) SLA for nuclear NF-κB with simultaneous intracellular antibody staining for negative regulator IκBα reveals their relationship at the single-cell level across multiple cell subsets.

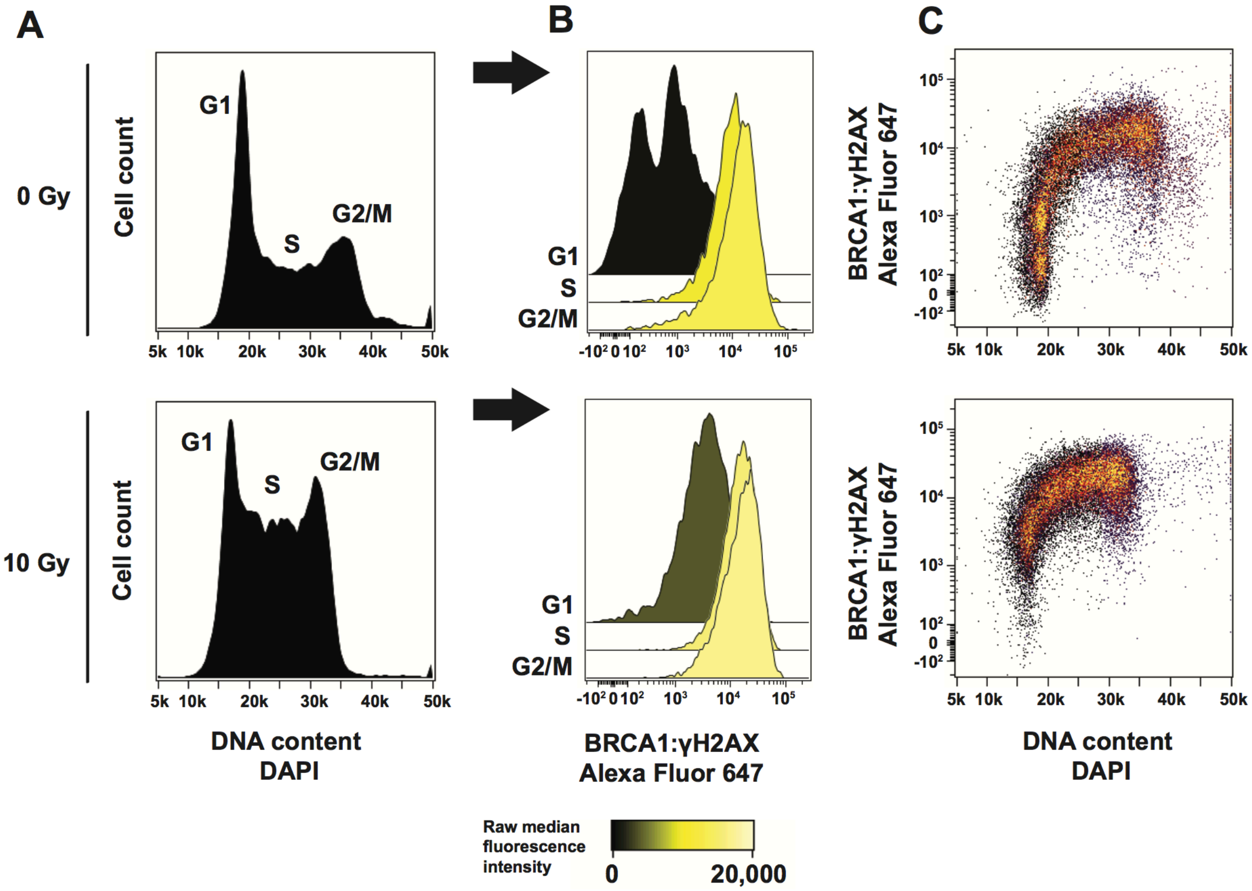


Figure 4: SLA quantifies BRCA1 localization to DNA damage sites.

(**A**) Cells were either untreated (top) or irradiated (bottom), and stained with DAPI for DNA content during the SLA procedure to gate between G1, S, and G2/M phase. (**B**) Gating reveals differences in the BRCA1-γH2AX interaction between cell cycle phases. (**C**) Dot plots reveal single-cell topology of the BRCA1-γH2AX interaction as a function of the cell cycle.

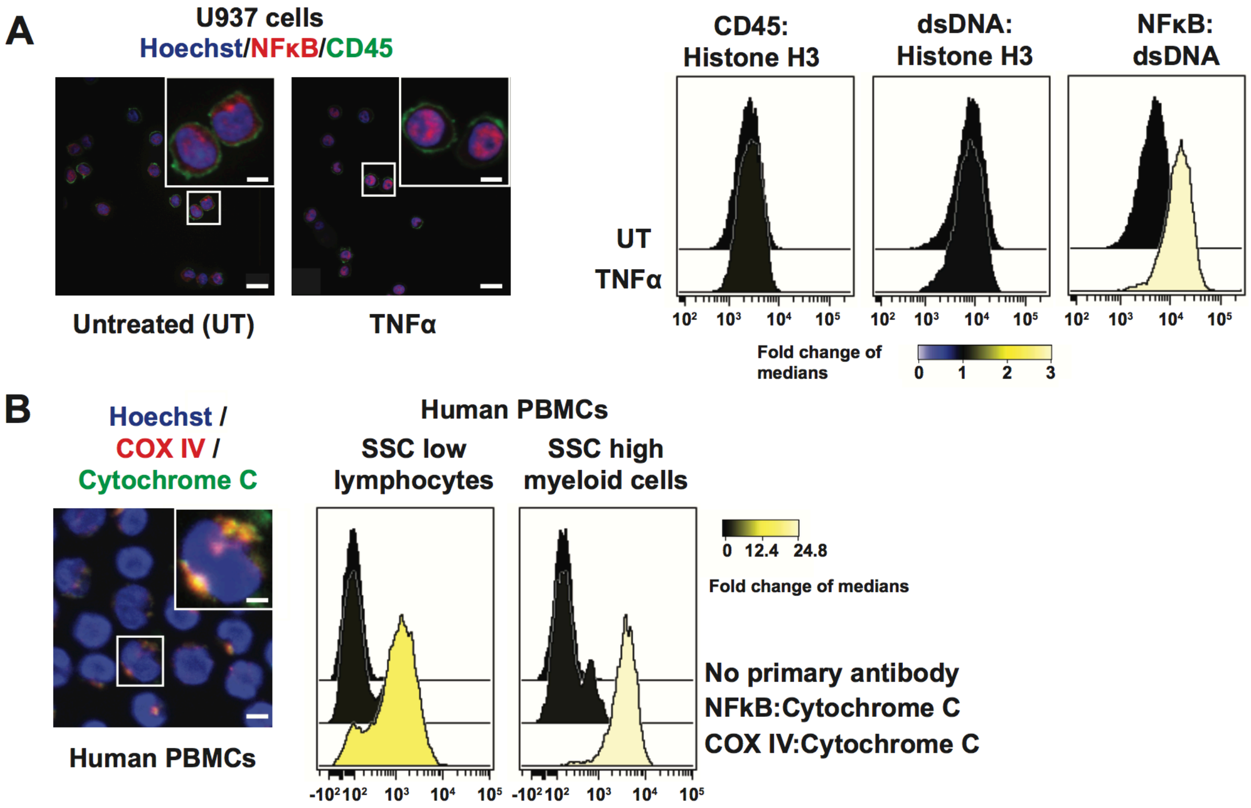


Figure S1: SLA was originally optimized for cell lines, and can detect mitochondrial localization. (**A**) Confocal microscopy (left) and SLA (right) for NF-κB interacting with dsDNA in the U-937 monocytic cell line (**B**) Confocal microscopy (left) and SLA (right) for mitochondria-specific interactions in human PBMCs. For SLA, myeloid and lymphoid cells were gated out by side scatter (SSC). Scale bars represent 10μm (main images in (**A**)), 3 μm (inset in (**A**)), and 5μm (main images in (**B**)), and 2.5μm (inset in (**B**)).

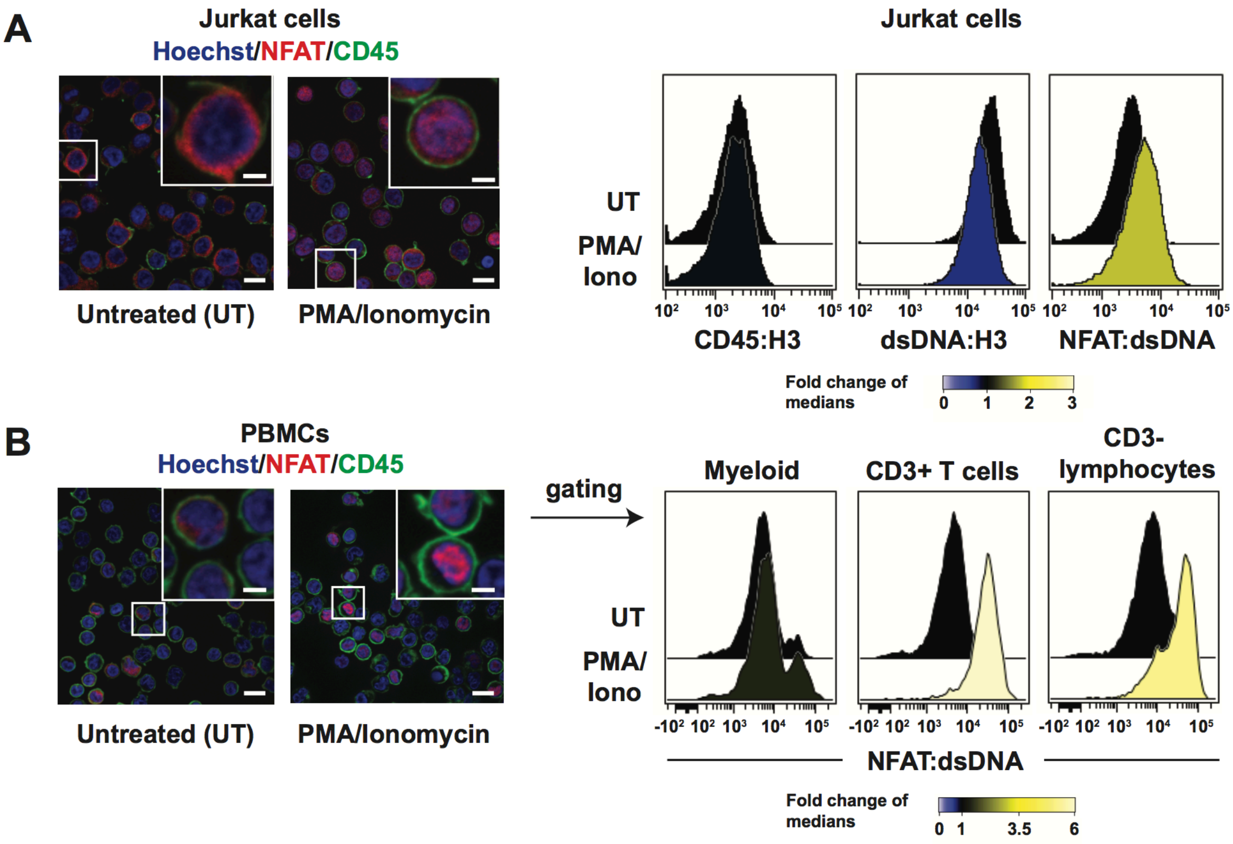


Figure S2: SLA detects nuclear localization of NFAT in cell lines and primary samples.

(**A**) Confocal microscopy (left) and SLA (right) for NFAT nuclear localization in the Jurkat T-cell line, as detected by the interaction between NFAT and a dsDNA antibody. The CD45:Histone H3 and dsDNA:Histone H3 interactions represented negative and positive signals respectively, and were not expected to be affected by PMA/Ionomycin treatment (**B**) Confocal microscopy (left) of NFAT in PBMCs, with SLA (right) performed in cell subsets gated out by surface markers. Scale bars represent 10µm (main images), and 3µm (insets).

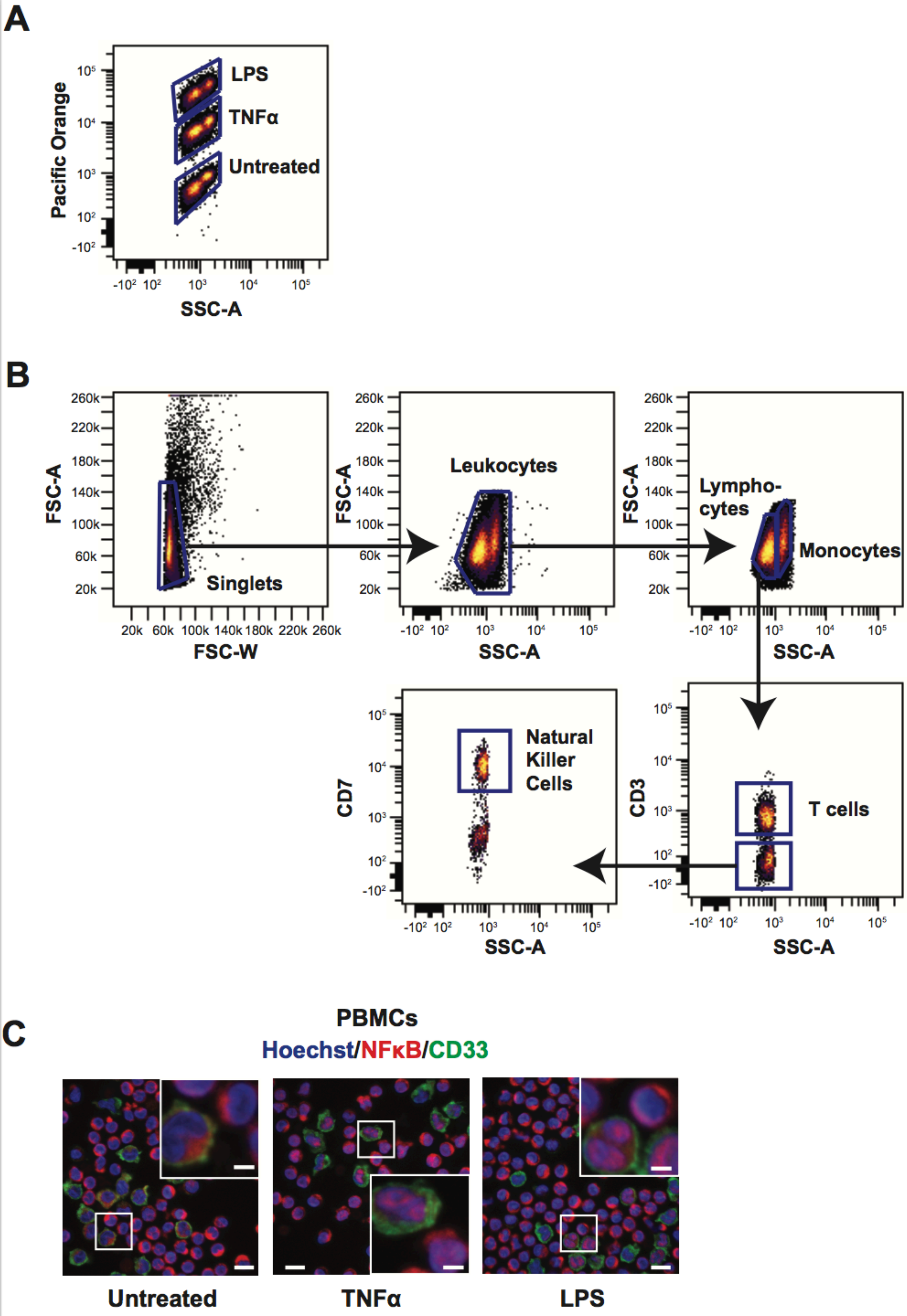

A
Pacific Orange
SSC-A
LPS
TNFα
Untreated
B
FSC-A
FSC-W
Singlets
Leukocytes
SSC-A
Lympho-cytes
Monocytes
CD7
Natural Killer Cells
CD3
T cells
C
PBMCs
Hoechst/NFκB/CD33
Untreated
TNFα
LPS

Figure S3: PBMC characterization by gating and confocal microscopy. (**A**) De-barcoding treatment conditions by strength of Pacific Orange signal. (**B**) Gating strategy. Singlets were gated and debris were cleared using FSC and SSC. Monocytes were gated by side scatter. Lymphocytes were gated by side scatter were further gated on CD3 and CD7, wherein CD3+ cells were labeled as T cells, and CD3- CD7+ cells were labeled as natural killer cells. (**C**) Confocal microscopy images taken show cytoplasmic localization of NF-κB in untreated conditions (left panel), both monocyte and lymphocyte nuclear translocation of NF-κB upon 15 minutes of TNFα treatment, and monocyte only nuclear translocation of NF-κB upon 15 minutes of LPS treatment. Scale bars represent 10μm (main images), and 3μm (insets).

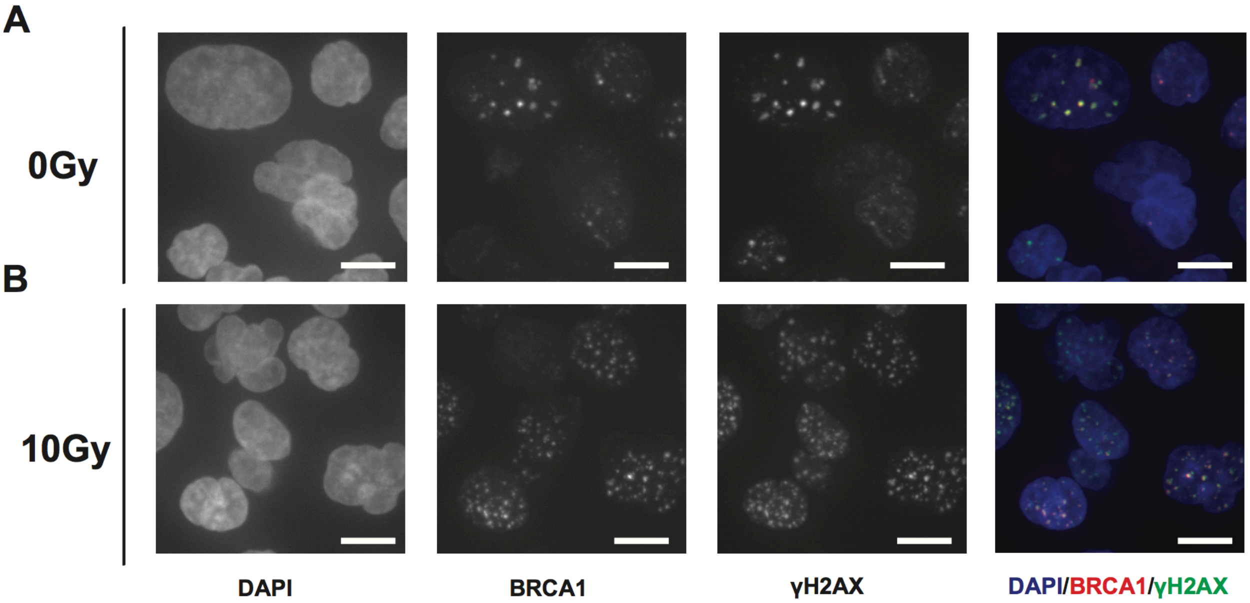


Figure S4: Confocal microscopy reveals the BRCA1 co-localization with γH2AX that SLA is able to quantify. IFM for BRCA1 and γH2AX with the same antibodies used for SLA in TYK-nu cells either (**A**) untreated and (**B**) treated with 10Gy of ionizing

radiation reveal BRCA1 and γH2AX foci that co-localize within the nucleus, as delineated by DAPI. Scale bars represent 10μm.

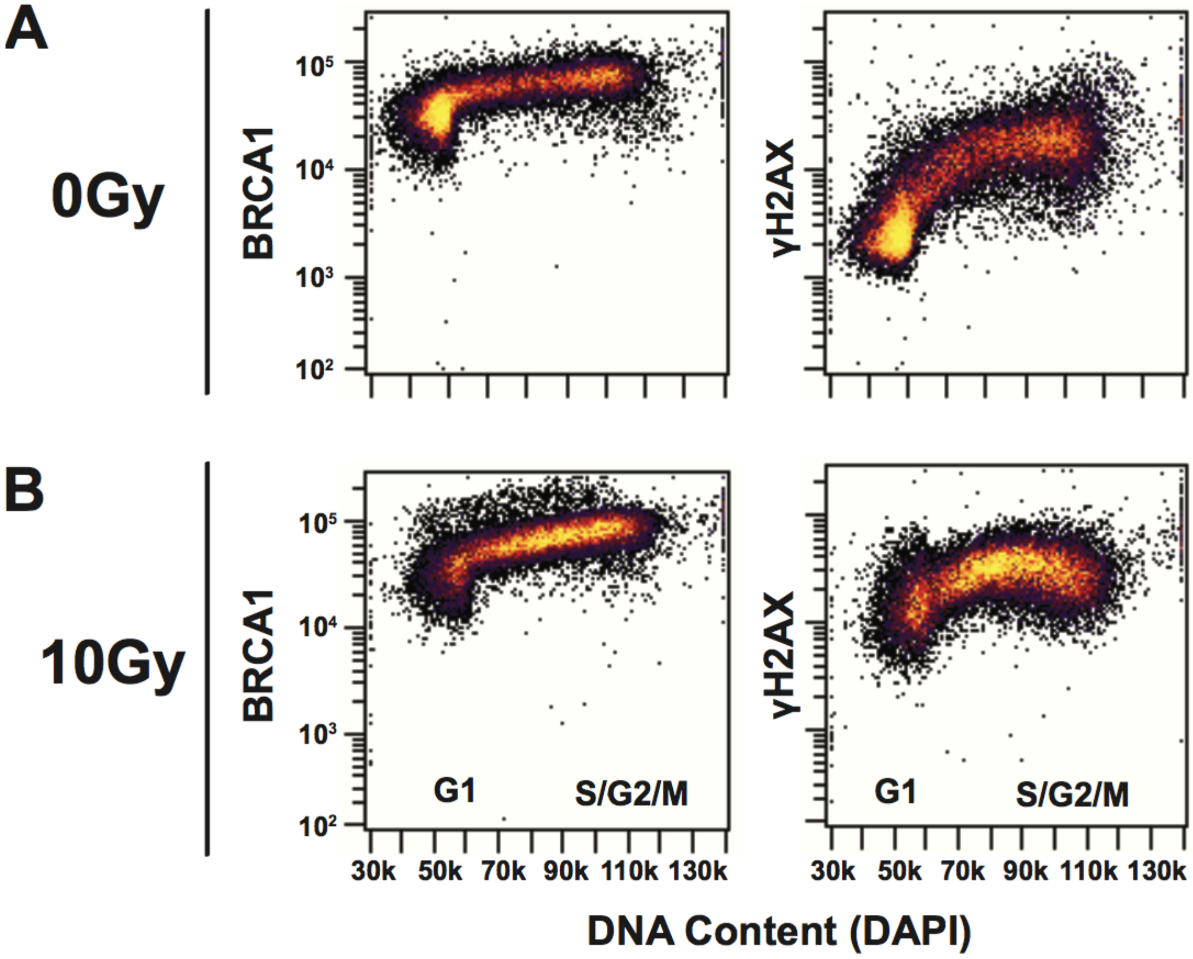


Figure S5: Dot plots of immunostaining for the γH2AX and BRCA1 antibodies used in SLA to delineate raw levels of the respective proteins as a function of the cell cycle in (**A**) untreated and (**B**) irradiated cells.

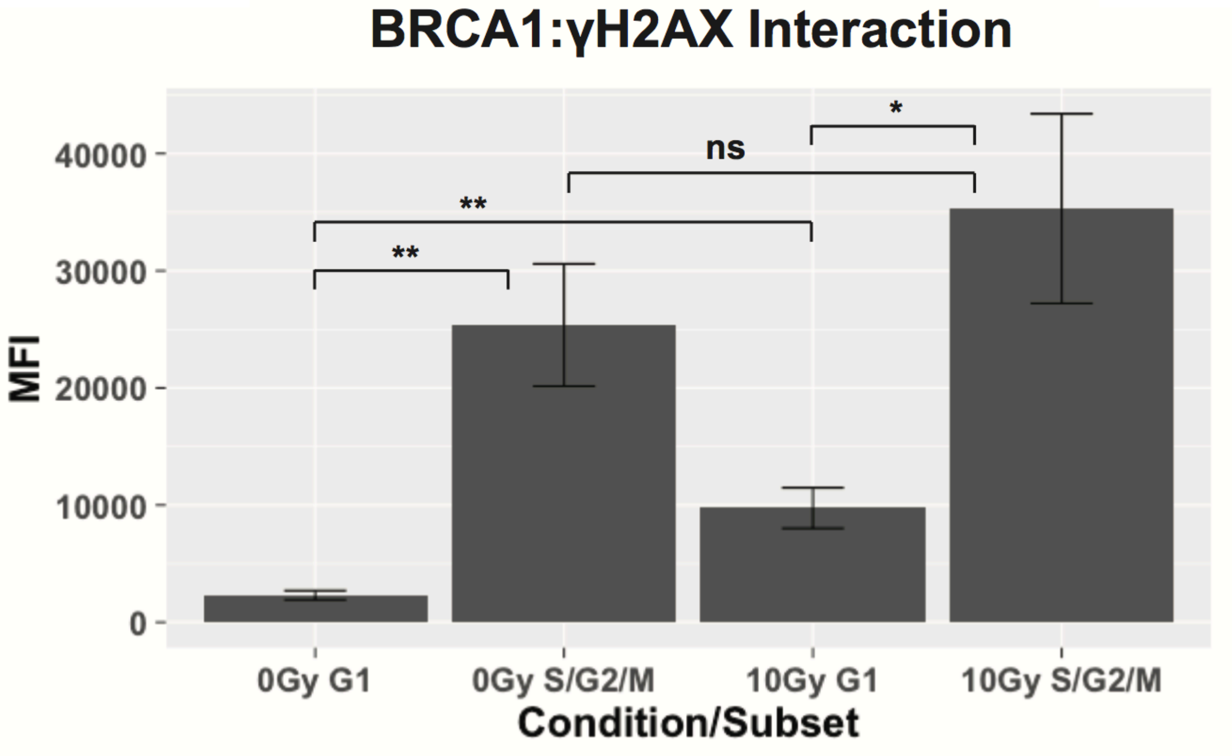
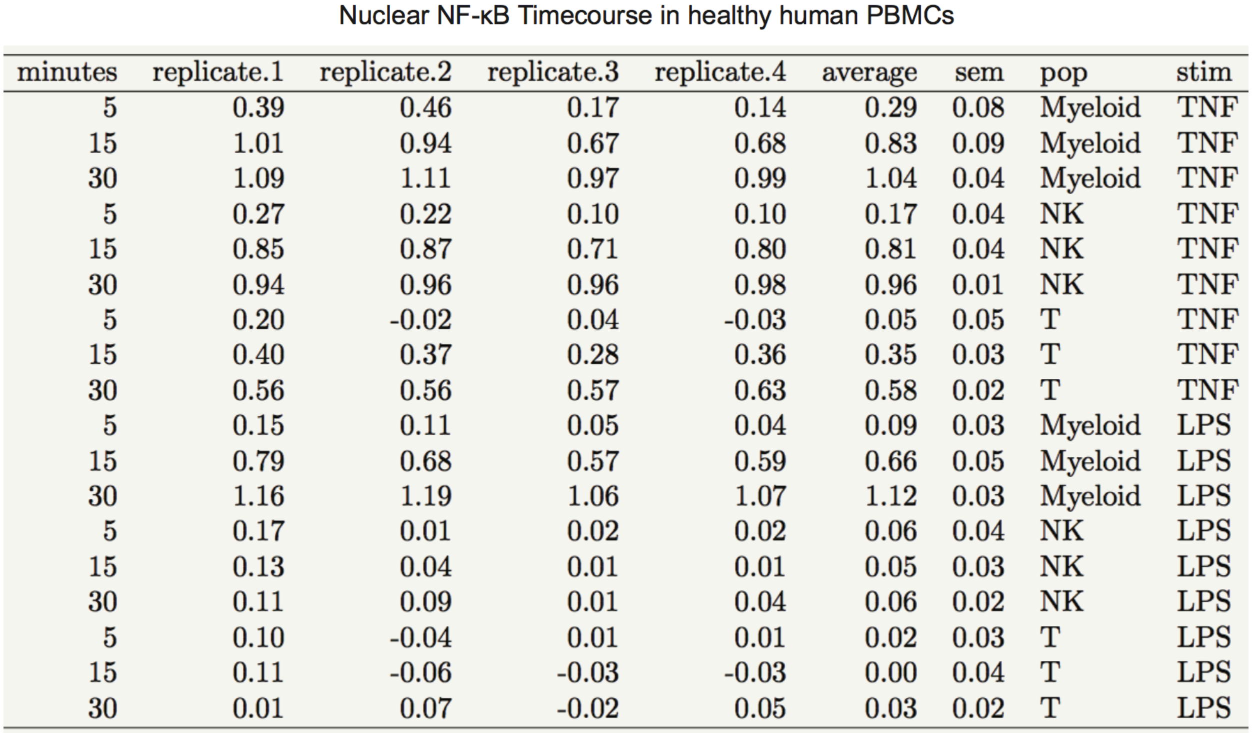


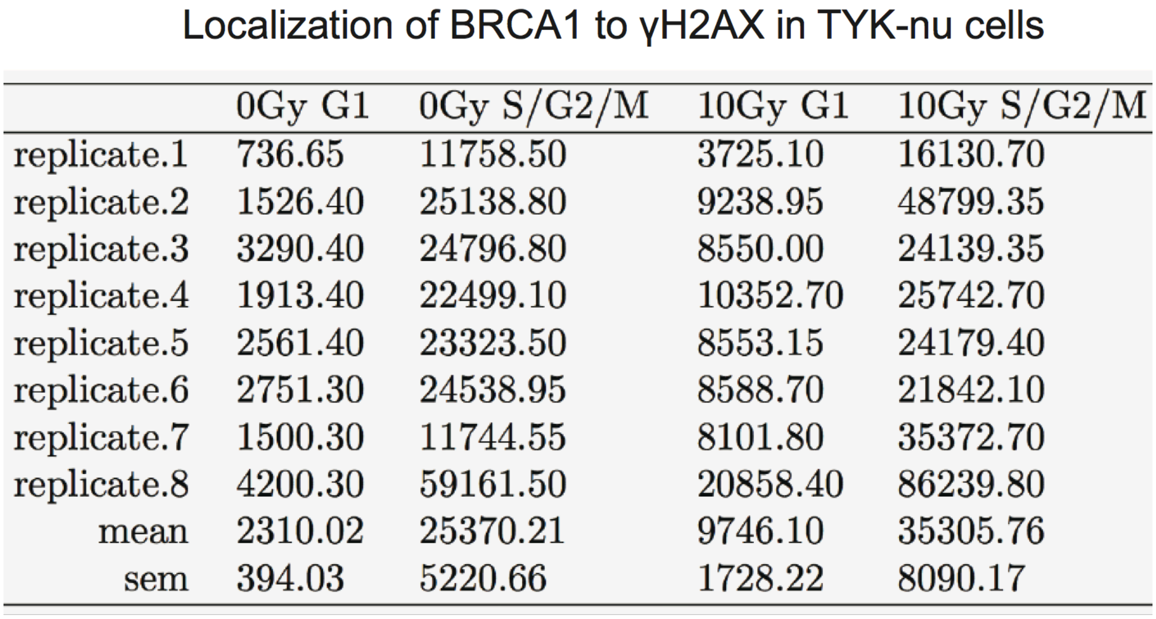
Figure S6: Ionizing radiation induces a G1-specific increase in BRCA1-γH2AX interaction in TYK-nu ovarian cancer cells. Bar plot for the experimental setup shown in Figure 4. Bars represent the mean ± SEM (n = 8). Statistical analysis was performed by two-tailed welch two sample t-test for the comparisons indicated. **p<0.01, *p<0.05, ns, not significant. T tests: column 1 v 2: t = -4.4046, df = 7.0798, p-value = 0.003055. Column 1 v 3: t = -4.1951, df = 7.7258, p-value = 0.00326. Column 2 v 4: t = -1.0319, df = 11.968, p-value = 0.3225. Column 3 v 4: t = -3.0896, df = 7.6375, p-value = 0.01577.



Nuclear NF-κB Timecourse in healthy human PBMCs

minutes	replicate.1	replicate.2	replicate.3	replicate.4	average	sem	pop	stim
5	0.39	0.46	0.17	0.14	0.29	0.08	Myeloid	TNF
15	1.01	0.94	0.67	0.68	0.83	0.09	Myeloid	TNF
30	1.09	1.11	0.97	0.99	1.04	0.04	Myeloid	TNF
5	0.27	0.22	0.10	0.10	0.17	0.04	NK	TNF
15	0.85	0.87	0.71	0.80	0.81	0.04	NK	TNF
30	0.94	0.96	0.96	0.98	0.96	0.01	NK	TNF
5	0.20	-0.02	0.04	-0.03	0.05	0.05	T	TNF
15	0.40	0.37	0.28	0.36	0.35	0.03	T	TNF
30	0.56	0.56	0.57	0.63	0.58	0.02	T	TNF
5	0.15	0.11	0.05	0.04	0.09	0.03	Myeloid	LPS
15	0.79	0.68	0.57	0.59	0.66	0.05	Myeloid	LPS
30	1.16	1.19	1.06	1.07	1.12	0.03	Myeloid	LPS
5	0.17	0.01	0.02	0.02	0.06	0.04	NK	LPS
15	0.13	0.04	0.01	0.01	0.05	0.03	NK	LPS
30	0.11	0.09	0.01	0.04	0.06	0.02	NK	LPS
5	0.10	-0.04	0.01	0.01	0.02	0.03	T	LPS
15	0.11	-0.06	-0.03	-0.03	0.00	0.04	T	LPS
30	0.01	0.07	-0.02	0.05	0.03	0.02	T	LPS

Table S1: Arcsinh transformed nuclear NF-κB MFI values in PBMCs. Raw data used to make the bar graphs in Figure 3A and 3B. sem, standard error of the mean. pop, population. stim, cytokine treatment (see Methods). Numbers are arcsinh transformed differences between the SLA signal during the given timepoint, and the untreated control tube for the given timepoint. MFI, median fluorescence intensity.



Localization of BRCA1 to γH2AX in TYK-nu cells

	0Gy G1	0Gy S/G2/M	10Gy G1	10Gy S/G2/M
replicate.1	736.65	11758.50	3725.10	16130.70
replicate.2	1526.40	25138.80	9238.95	48799.35
replicate.3	3290.40	24796.80	8550.00	24139.35
replicate.4	1913.40	22499.10	10352.70	25742.70
replicate.5	2561.40	23323.50	8553.15	24179.40
replicate.6	2751.30	24538.95	8588.70	21842.10
replicate.7	1500.30	11744.55	8101.80	35372.70
replicate.8	4200.30	59161.50	20858.40	86239.80
mean	2310.02	25370.21	9746.10	35305.76
sem	394.03	5220.66	1728.22	8090.17

Table S2: Raw Median Fluorescence Intensity of BRCA1-γH2AX SLA in TYK-nu cells. Data used to construct Figure S6. Mean and standard error of the mean (SEM) were used as bars and error bars respectively in said figure. Values are median fluorescence intensity (MFI).

Chapter 3: Continuous visualization of differences between biological conditions in single-cell data

3.1 Abstract

In high dimensional single cell data, comparison of functional responses across biological conditions typically requires partitioning of cell populations (clustering or gating). To address this, we developed **S**mooth **C**omparison **o**f **NE**ighbors (SCONE), an algorithm that performs statistical comparisons of functional response markers in continuously overlapping phenotypic neighborhoods. SCONE output is intended to be used for coloring lower-dimensional embeddings (eg. t-SNE maps), heatmaps or biaxial plots, thus directly visualizing local changes within immune cell subsets across biological conditions. We applied SCONE to a B cell precursors dataset from human bone marrow to reveal the coordination between surface marker expression changes and IL-7 responsiveness through pSTAT5 in the B-cell developmental trajectory. SCONE allows for a direct visualization and analysis of changes and trends at the single cell level across multiple biological conditions in a wide variety of model systems. This will provide useful information that could reveal direct biological insights and help inform downstream analysis.

3.2 Introduction

Novel technologies have emerged providing high parameter information from single cells, providing many opportunities to study the diversity of complex biological systems. These technologies include CyTOF (2), MIBI (27), and single-cell sequencing (61). Manifold embedding algorithms have been adapted to mass cytometry such as force-directed graphs (62) (25), t-SNE (24,63), and principal component analysis (PCA) (64) to visually represent the distribution of cells in high-dimensional space near each other in two dimensions. This provides an intuitive readout of the diversity of a given dataset and provides insight into a system's superstructure. Despite providing a visual and intuitive manner to explore single-cell data, dimensionality reduction plots are based on single cells and therefore to date have not allowed for direct visualization and quantification of differences between biological samples.

To perform such comparisons, researchers routinely resort to first partitioning the concatenated dataset into disjoint subsets (clusters or gates) based on surface markers that are not expected to change between conditions, and then for each subset performing sample-to-sample comparison of markers that are expected to change (functional markers). This approach allows visualization of discrete changes such as signaling differences between subsets(9,18-20,22,65,66). However, we deemed that it would be useful to be able to visualize the patterns of functional marker change in a continuous manner independently of partitioning (clustering/gating). One could immediately see the patterns of signaling response and use them to locate and extract the cell populations where the functional changes are localized.

Recently, a method was developed that tests for visualizing differential abundance of cells across biological conditions using per-cell overlapping hyperspheres in high dimensional data(67). Here, we present a computational method that focuses on changes in marker expression across biological conditions. Our method is called SCONE, which stands for **S**mooth **C**omparison **o**f **Ne**ighbors. SCONE makes statistical comparisons in overlapping k-nearest neighborhoods (KNN) rather than mutually exclusive clusters. Thus, each cell represents the information contained within itself and its KNN (which contains all biological conditions of interest). By applying normalization and choosing the markers that do not change across biological conditions for KNN calculations, the conditions being compared can be assumed to be distributed evenly in the marker space such that each cell and its KNN will contain data points from all biological conditions in question. We provide an objective statistical test that checks if this assumption is true and thus provides feedback on the tube-to-tube marker variation, and efficacy of normalizations. For each k-nearest neighborhood, the algorithm quantifies the differences in expression of each functional marker between cells belonging to two respective biological conditions. Given that k-nearest neighbor sets of adjacent cells are partially overlapping, SCONE can visualize and locate the boundary and shape of functional marker change patterns across the manifold, and view these changes at the single cell level prior to partitioning the data.

We used SCONE in conjunction with a tSNE based output, to study the changes in cell subsets occurring after an *ex vivo* perturbation. In human B cell precursors, we show the precise location of an IL-7 sensitive population in the developmental trajectory.

3.3 Methods

3.3.1 Mass cytometry experiments

Mass cytometry data used in this manuscript, together with the information regarding cell preparation, data acquisition, and processing was obtained from the original publications

(5,26,68). Through the rest of the manuscript, we will call these respective datasets by their first and/or co-first authors: Bodenmiller/Zunder/Finck, Bendall/Davis/Amir, and Fragiadakis.

3.3.2 Data input

A schematic of the algorithm's workflow is provided (Figure 1A). In brief, cells from a basal condition and one or more stimulatory conditions are concatenated into a single matrix of cells by features, with an additional column denoting condition. The software produces this matrix from a list of fcs files the user provides as input. These cells are then subject to an appropriate normalizing transformation (e.g. arcsinh-transformed with a cofactor of five, which became a *de facto* standard for CyTOF analysis(2)). The user then has the option to do per-marker quantile normalization (69) and/or z-score transformation, as a means to correct batch effects and reduce sample-to-sample variability (see **Correcting for technical artifacts in the data** for efficacy analysis). Each cell's k-nearest neighbors are determined with user-chosen input markers (in our case, surface markers) using Euclidean distance in the Fast Nearest Neighbors (FNN) R package.

3.3.3 Per-cell comparisons

For each k-nearest neighborhood, for each functional marker of interest, two values are calculated. The first is the raw change between the two biological conditions, which is defined as the difference between the median value of one condition and the condition the user defines as "basal." The second is a p-value output from a user-chosen statistical test (currently, Mann-Whitney U test (70) or T test (71)) between the distributions of marker values for each biological condition. Accordingly, the p-value is adjusted for false discovery rate using p.adjust function within R. We therefore call the statistical test output q-values. Following this, we give the user the option to threshold the raw changes by q-value. In other words, if the q-value for a given k-nearest neighborhood is lower than a user-defined cutoff (e.g. 0.05), then the raw change will be reported. Otherwise, the raw change will take on the value zero.

3.3.4 Per-replicate comparisons

In experimental setups with replicates, the user has an option to make per-replicate comparisons across multiple biological conditions. For each k-nearest neighborhood, for each marker of interest, the user-chosen median or mean values of expression are computed for each replicate designated as “control” condition and each replicate designated as an “experimental” condition (e.g. stimulated cells). These values are then used as input in a t-test comparing basal versus experimental per-replicate expression for the given marker. The resulting output is the replicate comparison p-values . Like the per-cell p-values, they are corrected for false discovery rate using the p.adjust function within R. They are therefore also referred to as q-values, as shown in basal versus LPS-treated healthy human whole blood (68) (Figure S1).

3.3.5 Structure of SCONE output

These per-replicate and/or per-cell statistics are appended column by column to the end of the original single cell expression matrix (the content of the FCS file). Each new column is either a q-value or a raw change of a single functional marker and therefore can be parsed as one would with normal cytometry data. The user then has the option to run t-SNE within our software and add the t-SNE embedding coordinates as columns to the end of the data matrix. We recommend the user choose the same input markers as those that went into the KNN calculation. We show an example of our output using these t-SNE maps of basal versus IL-10 treated whole blood from a previously published mass cytometry dataset (68), comparing raw pSTAT3 expression values with their q-value and raw change values calculated by our algorithm within k-nearest neighbors, with surface marker expression shown to reference specific cell subsets (Figure 1B, C). While visually it’s hard to call out the differences between the basal and the IL-10 stimulation conditions, the SCONE algorithm effectively highlighted the responsive cell populations as well as subtle patterns of responsiveness within those populations. For instance, the CD14+ CD33+ monocyte population exhibited a stronger response to IL-10 stimulation through pSTAT3 than the adjacent CD33- CD16+ population (Figure 1C, right panel).

3.3.6 Selection of the number of nearest neighbors

We chose the optimal number of nearest neighbors (k) by solving the functional marker imputation problem, i.e. by determining how well the functional response markers could be predicted from the respective values of the k-nearest neighbors. Specifically, for each cell we computed median values of the functional variables from its k-nearest neighborhood. We then optimized the *k* value (neighborhood size), using a dataset median of the Euclidean distance between the actual functional variable vector for a given cell and the imputed median vector as a loss function (72). We evaluated this optimization on a previously published mass cytometry dataset consisting of human PBMCs across a variety of stimulatory conditions on the Bodenmiller/Zunder/Finck and Bendall/Davis/Amir datasets. We observed that across a range of *k* values ranging from n/5 to n/1000, the relationship between $\log_{10}$(k) and the imputation loss was parabolic, with a clearly defined minimum. As such, one could find the value of k that effectively minimized the imputation loss (Figure S2. We found that for n = 10,000 PBMCs in these data, the ideal k was determined to range between n/20 to n/100 depending on stimulatory condition. We expect this value to differ depending on datasets and biological conditions being used. If the calculated ideal k value differs between biological conditions being compared, we recommend choosing a value k that is the mean of ideal k values found across the biological conditions. Within our software, we provide the user the ability to use this metric for the dataset in question prior to executing the SCONE workflow, and we recommend this be done before each new analysis.

We next sought to determine if sparse regions of the data would require a different value of k to minimize the imputation error. To test this, we used our same *k* titration from the Bodenmiller/Zunder/Finck PBMC dataset to examine the value of *k* that minimizes the per-cell imputation error, rather than the global imputation error (which was the median of the per-cell imputation error). As we expected, different cells had different ideal *k* values. However, in all conditions tested, the ideal *k* per cell did not correlate with the density of our data (Figure S3). Thus, we provide the user with software to test and find the ideal global k for the dataset being analyzed, but we do not need to adjust the per-cell k as a function of the data's density.

3.3.7 Correcting for technical artifacts in the data

There may be shifts in marker expression levels between biological conditions attributed entirely to technical artifact. This would affect the accuracy of the statistics within each cell's KNN. To address this, we provide the user with the ability to perform quantile normalization, and/or z-score transformation of each feature going into the KNN generation. We see this as an important pre-processing step primarily with datasets in which multiple donors are combined. Accordingly, our software automatically outputs the per-KNN fraction of cells belonging to each non-baseline biological condition compared to the user-designated baseline for each KNN. If we let

x_b equal the user designated baseline condition and x_i equal the ith non-baseline condition, then we let α_n be a function that takes in two conditions as input (here x_i and x_b), and outputs the fraction of the cells belonging to a condition x_i divided by the number of cells belonging to both x_i and x_b, in the *k*-nearest neighborhood of nth cell.

$$\alpha_n(x_i, x_b) = \frac{count_n(x_i)}{count_n(x_i) + count_n(x_b)}$$

We then let vector $\alpha(x_i, x_b)$ contain all values outputted by $\alpha_n(x_i, x_b)$ for all cells within the dataset. A different $\alpha(x_i, x_b)$ is calculated for each non-baseline x_i condition in the dataset (as it compares to the user-designated baseline condition x_b).

$$\alpha(x_i, x_b) = \{\alpha_1(x_i, x_b), \alpha_2(x_i, x_b), \alpha_3(x_i, x_b), \alpha_4(x_i, x_b), \dots, \alpha_n(x_i, x_b)\}$$

Using $\alpha(x_i, x_b)$ for a given non-baseline condition, we sought to quantify the "overlap" of data between conditions for the expression markers not expected to change, and further test efficacy of normalization and scaling procedures in terms of this overlap. We used the Bodenmiller/Zunder/Finck dataset along with the newer Bendall/Davis/Amir dataset (26), and we assessed the α of cells treated with GM-CSF or IL-7, respectively. Given that we performed KNN on surface markers that are not expected to change with a 15-30-minute *ex vivo* perturbation, we expected each α to be near 0.5 if there were minimal technical artifacts. Our primary focus was to test the distribution of this KNN fraction across datasets with and without per-marker quantile normalization and scaling of parameters with a z-score transformation prior to generating KNN. As an evaluation metric, we calculated the standard deviation of the α distributions with each normalization/scaling method, and visually displayed this output as a histogram.

With each dataset, we also randomly subsampled the user-designated baseline sample without replacement twice, treated each subsample as if it came from a separate condition, designating the subsamples as "conditions" s_1 and s_2. Our data quality reference was the standard deviation of $\alpha(s_1, s_2)$. Our final score was the quotient of the standard deviation of $\alpha(s_1, s_2)$ from the subsampled file and $\alpha(x_i, x_b)$ from the baseline condition and "stimulated" condition files. We call this value the manifold overlap score, or m.

$$m = \sigma\big(\alpha(s_1, s_2)\big)/\sigma\big(\alpha(x_i, x_b)\big)$$

In both cases, we found that per-marker quantile normalization followed by scaling of the data best minimized the standard deviation of $\alpha(x_i, x_b)$, and therefore maximized m (Figure S3A, B). Of note, the Bendall/Davis/Amir dataset is relatively newer, and therefore in all cases had a higher m than the Bodenmiller/Zunder/Finck dataset (Figure S3B). Given that the instruments and pre-processing are constantly being updated, we expect the newer datasets to have more marker expression overlap across runs.

We further validated the ability to detect change with SCONE after normalization and/or scaling in both datasets. With the Bodenmiller/Zunder/Finck dataset, we show that in all cases, pSTAT5 increase in CD33 positive cells could be detected after GM-CSF treatment (Figure S4A). With the Bendall/Davis/Amir dataset, we show that in all cases, a small population demarcated by pSTAT5 increase could be detected after IL-7 treatment (Figure S4B).

We recommend the user performing these same tests with the data of interest, as the need for normalization and scaling may vary with each dataset. We provide the software for this accordingly. Of note, while normalization methods can be of help in certain situations, one cannot possibly rely on a normalization method to “fix” data that has lots of technical artifacts, so special care must always go into the experimental design and data acquisition.

3.3.8 Visualization of output

For this manuscript, we use t-SNE as the primary mode of visualizing the single cell information that is obtained, coloring each cell on the map by the comparison values (raw change, q-value) between biological conditions for features of interest. Although t-SNE is an effective way to reduce high dimensional data into two dimensions, there are other methods that could just as effectively visualize the data that are beyond the scope of this manuscript, and other methods may be more optimal than t-SNE depending on the biological question and the number of cells being used as input. We nonetheless provide the user with the option of running t-SNE on the data, which will then add two columns to the data matrix containing the t-SNE1 and t-SNE2 features. This is done using the *Rtsne* R package, which in turn uses the accelerated Barnes-Hut implementation (73). We recommend the input features into this function to be the same as the input features that went into the k-nearest neighbor generation.

3.4 Results

3.4.1 Visualizing IL-7 responsiveness along the B cell developmental trajectory

The aforementioned Bendall/Davis/Amir dataset was from a study on B cell development using mass cytometry and a novel computational approach called Wanderlust to infer a developmental trajectory in static samples of B cell precursors manually gated from healthy human bone marrow (26). This previous study found a rare population effectively defined by responsiveness to IL-7 through pSTAT5. However, this population's responsiveness to IL-7 in comparison to other populations could only be interrogated by manual gating or clustering untreated and treated samples.

We first compared the identification of this population with SCONE to doing so with clustering. To this end, using data sub-sampled to 10,000 cells, we performed k-means clustering varying the number of clusters from 10 to 1000. We then performed SCONE on the same data with our optimized k of 100 (see **selection of the number of nearest neighbors**). We ran t-SNE on the concatenated SCONE and clustered data such that IL-7/pSTAT5 sensitive population would lie on the same region in t-SNE space in all cases. Statistics were computed either per-KNN or per-cluster. We then colored each cell by its per-KNN or per-cluster q-value derived from a FDR-adjusted Mann-Whitney U test between pSTAT5 expression in untreated versus IL-7 treated cells (Figure S5A), and per-KNN or per-cluster pSTAT5 raw change value between baseline and IL-7 (Figure S5B).

With k-means clustering, we found that this population was demarcated with statistical power comparable to SCONE only if the number of clusters was set to 500 (1/20 the number of cells in the dataset) or below. However, unlike the KNN neighborhoods that are all the same size by definition, clusters tended to be of an uneven size, which at high cluster numbers (along with the smaller cluster size here), led to increased noise in the output of pSTAT5 raw change values and declining statistical power. In addition, while the KNN-based output produced soft boundaries of this functional population, the k-means clustering was creating partitions with rigid boundaries, sometimes grouping together unrelated cell populations with varying stimulation response patterns. Taken together, this led to a tradeoff between bias and variance of statistical estimation, where the relevant IL7-pSTAT5 population was visible only depending on a specific cluster number setting.

To directly visualize the location of this subset was within the B cell developmental trajectory, we ran SCONE on these untreated and IL-7 treated samples. We first visualized the data with t-SNE to determine where this population was in relation to related cell surface markers along wanderlust values. Wanderlust values are the algorithmically-derived ordering of cells along a virtual developmental trajectory. Coloring the t-SNE maps by wanderlust values allowed for visualization of the inferred static and transition states within the populations (Figure 2A). This pSTAT5 responsive population was distinguishable easily through the SCONE values, as opposed to the general pSTAT5 levels of the untreated and IL-7 treated cells (Figure 2A bottom-right, Figure S6). The t-SNE maps of the SCONE values allowed for validation of a previously described B cell precursor subset with increased IL-7 responsiveness through pSTAT5. These pSTAT5 SCONE values showed visually that IL-7 responsiveness through pSTAT5 increases immediately upon increase in levels of specialized DNA polymerase TdT, and decreases immediately upon concurrent upregulation of CD24 (Figure 2A).

To identify this subset's relationship to coordination points in the developmental trajectory, we produced a heatmap colored by relative values of each marker ordered by cells occupying binned wanderlust values (Figure 2B). This heatmap revealed that the pSTAT5 increase is also coordinated with an increase in CD179a, CD38 along with TdT, and decrease is coordinated with an increase in CD10 along with CD24. Thus, two points of coordinated change in surface marker levels are connected by a small but distinct IL-7 responsive population. We found this architecture to be consistent across four healthy human donors (Figure S6).

3.5 Discussion

The SCONE approach presented here allows for statistical comparison of marker expression in multiple biological conditions by partitioning the cells into exhaustively overlapping subsets rather than disjoint subsets. We used t-SNE for visualization, but because the k-nearest neighborhoods are calculated in the original high dimensional manifold, any dimensionality reduction technique could be used for visualizing the results, as each cell will still report statistics the user chose to acquire within the KNN.

We produced and provide a method that allows the user to determine the optimal *k* within each individual dataset by minimizing the global imputation error for all functional markers of interest. This leads to an interesting fundamental question beyond the scope of this manuscript about how well a given set of markers can predict expression of another set of markers within a given CyTOF dataset.

We provide per-marker quantile normalization and z-score transformation methods for data pre-processing. Here, we use our k-nearest neighbors architecture to investigate the per-KNN overlap of cells across two biological conditions where the input features (surface markers) are not expected to change, which we designate as manifold overlap, or *m*. We then use this to evaluate per-tube variation across two datasets before and after the use of our normalization metrics. Outside of dataset quality and computational normalization metrics, the *m* score could be used to evaluate many experimental conditions, including new or existing blood preservation systems, barcoding systems, and automation systems. Furthermore, our $\alpha(x_{i,}\ x_b)$ vector construction could be used in model systems where cell subset abundance changes are expected, like the immune system in animal infection models. Here, t-SNE visualization of the per-cell $\alpha_n(x_i,\ x_b)$ output could provide preliminary information as to which cell subsets are changing prior to any partitioning steps.

When we compare the q-values for marker expression change overlapping k-nearest neighbors to k-means clustering, we show that the statistical power of such comparison decreases as the number of clusters increases (and therefore cluster size decreases and more variable). Given that clustering and gating are indispensable families of methods to identify cell subsets of interest, we see SCONE as a complimentary method that could be used initially to highlight functional changes that could then perhaps be used as input for downstream partitioning steps.

Using a previously published dataset consisting of B cells purified from healthy human bone marrow, we demarcated the boundaries of IL-7 responsiveness through pSTAT5 in relation to other relevant surface markers. We revealed that both the emergence and exit of IL-7 responsiveness through pSTAT5 across the developmental trajectory marked two distinct coordination points, where surface markers abruptly changed as well. Of note, the IL-7 responsive population was previously elucidated by mean of a manual gating analysis based on a combination of TdT (high) and CD24 (low) markers (26). We showed without gating or clustering that the IL-7 responsive population emerges upon increase in TdT, and decreases upon increase in CD24.

Given that TdT levels are high throughout the IgH gene rearrangement step of developing B cell precursors (74), the data suggest that cells are responsive to IL-7 through pSTAT5 during the early stages of this process. Our analysis explicitly shows that the subsequent IL-7 pSTAT5 pathway rewiring previously described (26) occurs concomitantly with changes in the expression of multiple surface markers.

SCONE can be used as a discovery and visualization tool for high-dimensional data, providing functional information prior to gating or clustering steps in data analysis pipelines. This approach is not limited to flow and mass cytometry. We expect our KNN-based approach to be of use with high-parameter imaging and single-cell sequencing data, which are approaching dimensionality and throughput levels sufficient for this type of analysis.

3.6 Author contributions

TJB wrote and implemented all code used in this manuscript, and wrote the manuscript. GPN edited the manuscript and provided direction and guidance. NS provided detailed direction and guidance with both the project and the code, wrote parts of the manuscript, and edited the manuscript.

3.7 Acknowledgements

We would like to acknowledge Kara Davis, and Sean Bendall, and Gabi Fragiadakis whose data was used for this manuscript, and whose insights guided our analysis. We would also like to acknowledge Alyssa Mike and Julie Yu, whose collaboration on a related problem inspired the SCONE approach.

This work was supported by the US National Institutes of Health (grants R33CA0183692, U19AI057229, 1U19AI100627, 7500108142, R01CA184968, 1R33 CA183654-01, 1R01GM10983601, 1R21CA183660, 1R01NS08953301, 5UH2AR067676, 1R01CA19665701, R01HL120724, N01-HV-00242 HHSN26820100034C, 41000411217, 5R01AI073724, U54-UCA149145A to G.P.N), US Department of Defense Congressionally Directed Medical Research Programs (grants OC110674, W81XWH-14-1-0180 to G.P.N), Bill and Melinda Gates Foundation (OPP1113682 to G.P.N), Pfizer, Inc (123214 to G.P.N), US Food and Drug Administration (grants BAA-15-00121, HHSF223201210194C to G.P.N), Novartis Pharmaceuticals Corp (CMEK162AUS06T to G.P.N), Juno Therapeutics Inc (122401 to G.P.N), California Institute for Regenerative Medicine (grant DR1-01477 to G.P.N), Gilead Sciences, Inc. Research Agreement (to G.P.N), NWCRA Entertainment Industry Foundation (to G.P.N), DiaTools (Grant 259796 to G.P.N) Cancer Biology Training Grant 2T32CA009302-36A1 to T.J.B, and the Rachford and Carlota A. Harris Endowed Chair (G.P.N.).

3.8 Figures

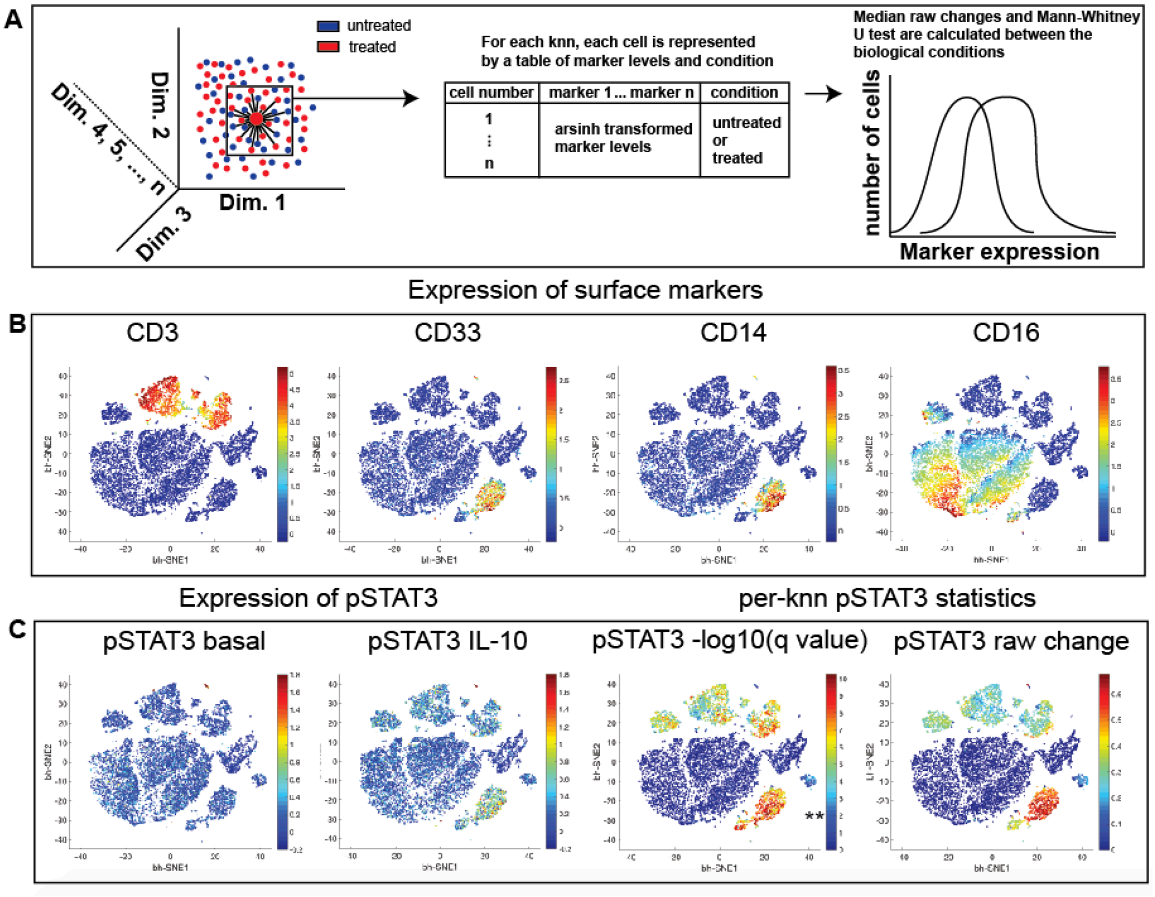


Figure 1: Schematic of the SCONE algorithm and its output. (A) (left) Cells from two or more biological conditions are used as input for k-nearest neighbors (KNN) generation, using user-defined features. (middle) Each KNN is a matrix of cells by features, which include functional markers hypothesized to change between conditions, along with an additional column demarcating which biological condition was used. (right) Statistical tests are performed between the distributions per feature, and Arsinh differences are thresholded by respective FDR-adjusted q values. (B) t-SNE map of whole blood colored by surface marker expression revealing specific cell

subsets of interest, to be used in the context of functional change analysis (C). Visualization of pSTAT3 expression in fresh blood in untreated and IL-10 treated cells, along with the SCONE-derived visualizations of the $-\log_{10}$ FDR-adjusted q values from a per-KNN Mann-Whitney U test, and raw pSTAT3 change thresholded on a q value of less than 0.05. **, q < 0.01.

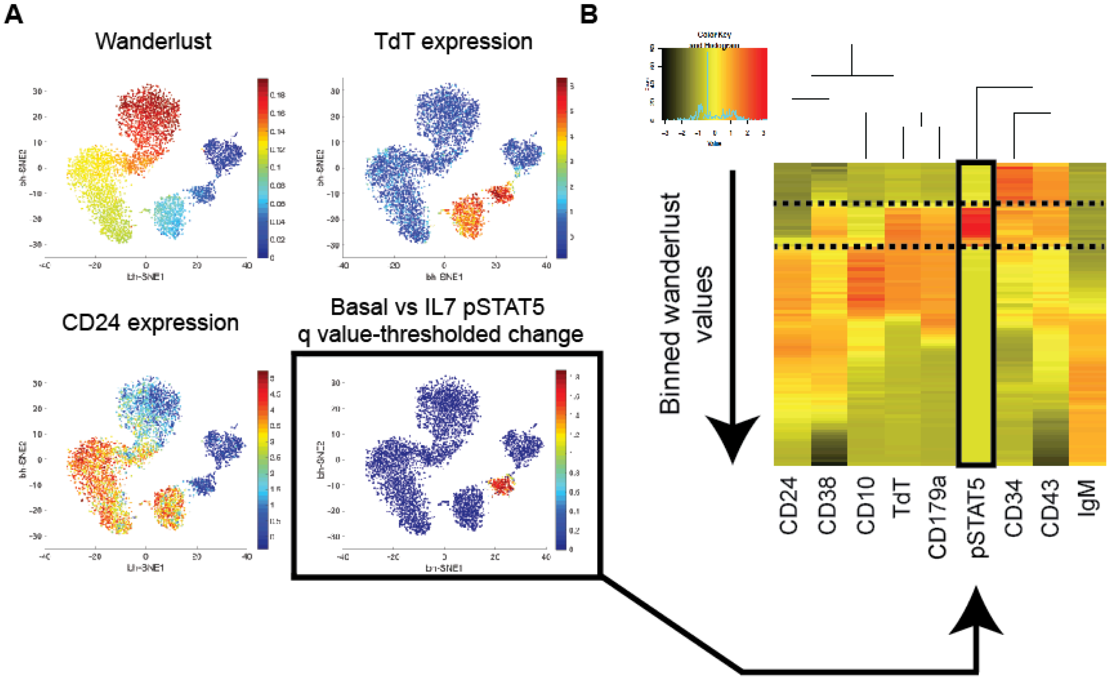
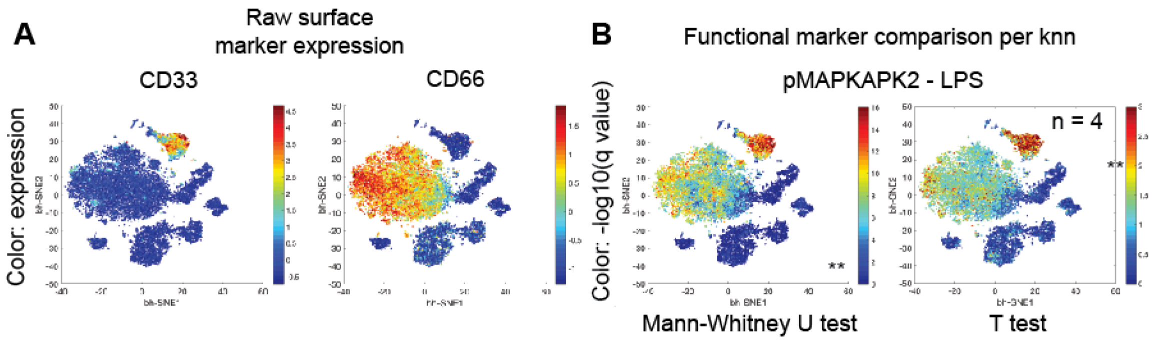


Figure 2: Adding functional statistics to single cell trajectory visualizations. (A) t-SNE map of B cell precursor data, colored by Wanderlust values and surface marker expression levels, and SCONE-derived q value thresholded change in pSTAT5 between untreated and IL-7. (B) Cells were binned and ordered by their Wanderlust values, visualized top to bottom on a heatmap. Marker expression values along with aforementioned change between untreated and IL-7 (solid black box) were merged

onto the same heatmap. Dashed lines indicate "coordination points," where expression values of many markers change simultaneously.



Dataset: Fragiadakis *et al, Anesthesiology* (2015)
Donors: healthy human
Cells: whole blood

Figure S1: Per-KNN statistical tests can reveal differences both across cells and across donors or replicates. (A) t-SNE maps colored by surface marker expression. (B) Per-cell Mann-Whitney U test or per-donor t test (n = 4) median values of pMAPKAPK2 expression between untreated and LPS-treated cells. **, $q < 0.01$.

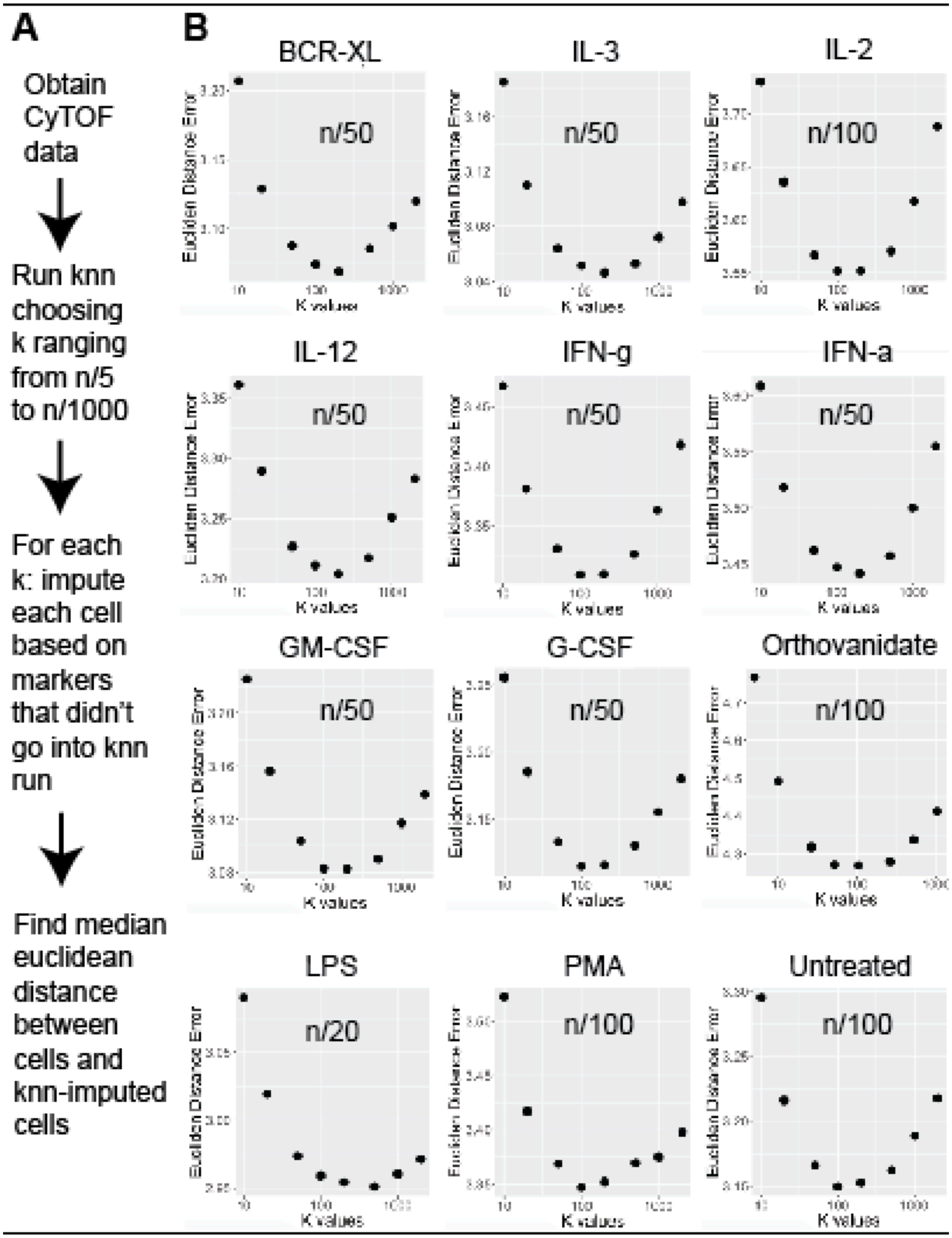
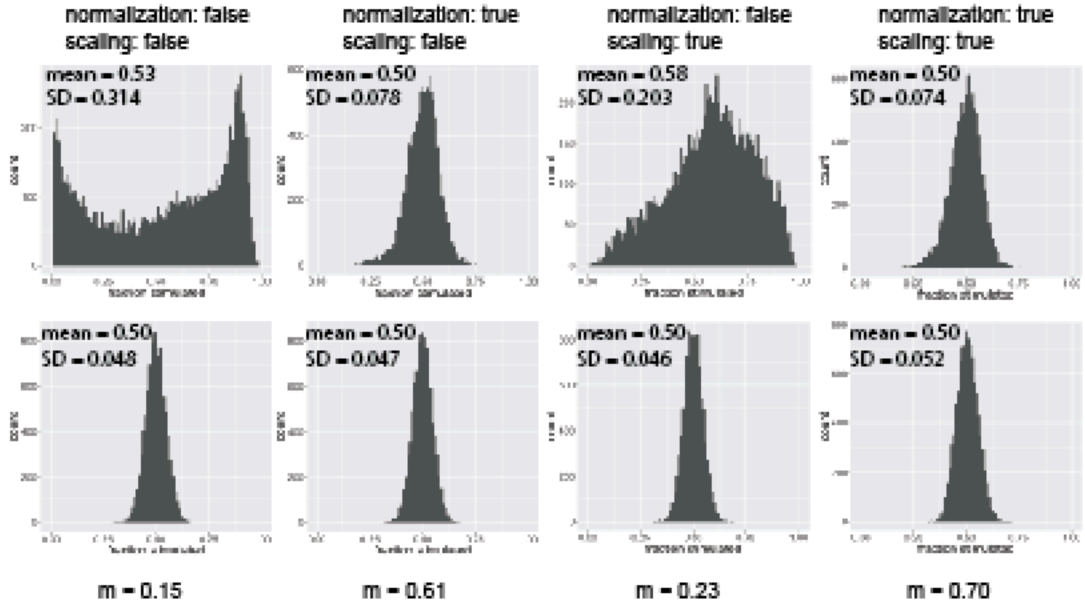


Figure S2: Computational strategy for selection of the number of nearest neighbors (A) Workflow for using various user-selected values of k to impute signaling marker

levels from surface markers, and minimizing the error between imputed “cell” and actual cell. (B) Parabolic relationship between selection of (k) and the imputation error in healthy human PBMCs untreated or treated with various immunomodulatory agents, with the minimum value per chart being displayed. n, number of cells used as input.

A. Dataset: Bodenmiller, Zunder et al, Nature Biotechnology 2012
Conditions: (top) untreated and GM-CSF, (bottom) sampled untreated without replacement



B. Dataset: Bendall, Davis, et al, Cell 2014
Conditions: (top) untreated and IL-7, (bottom) sampled untreated without replacement

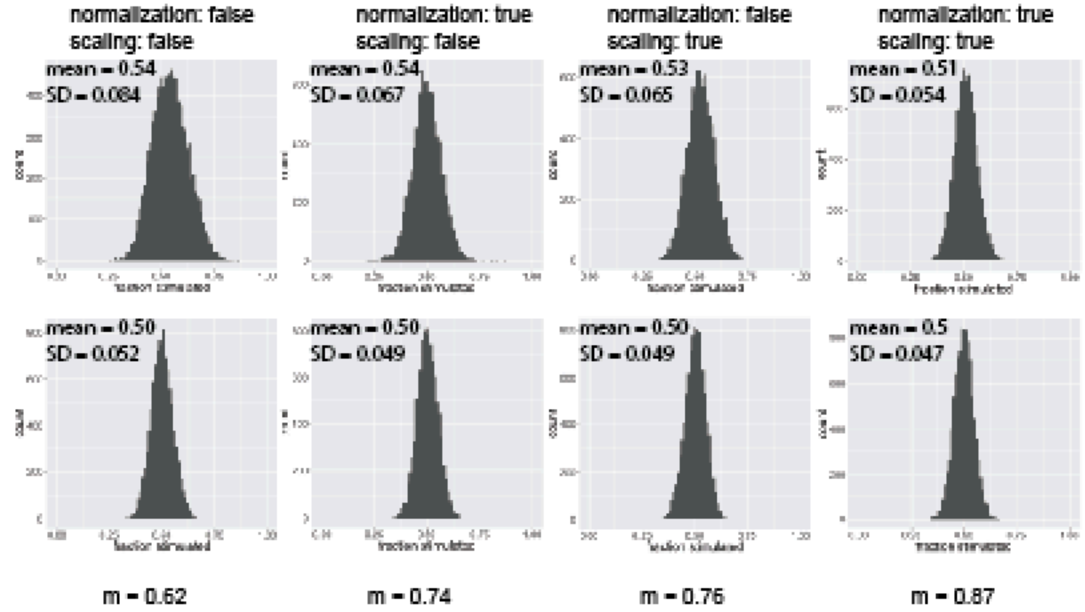


Figure S3: Per-marker quantile normalization and per-file z-score transformation (scaling) increase the overlap of two high dimensional manifolds being compared. (A) Human PBMC dataset from Bodenmiller/Zunder/Finck. (Top) Per-KNN ratio of cells belonging to condition 1 versus condition 2 is used as an evaluation metric, with mean and standard deviation of these values evaluated and histograms plotted. (Bottom) As a control, the untreated file is randomly sampled into two "conditions" for comparison, and the same workflow is performed. Manifold overlap score (*m*) is the quotient of the standard deviation of the twice-sampled untreated file and that of the comparison between two conditions (B) Human B cell precursor dataset from Bendall/Davis/Amir with the aforementioned workflow.

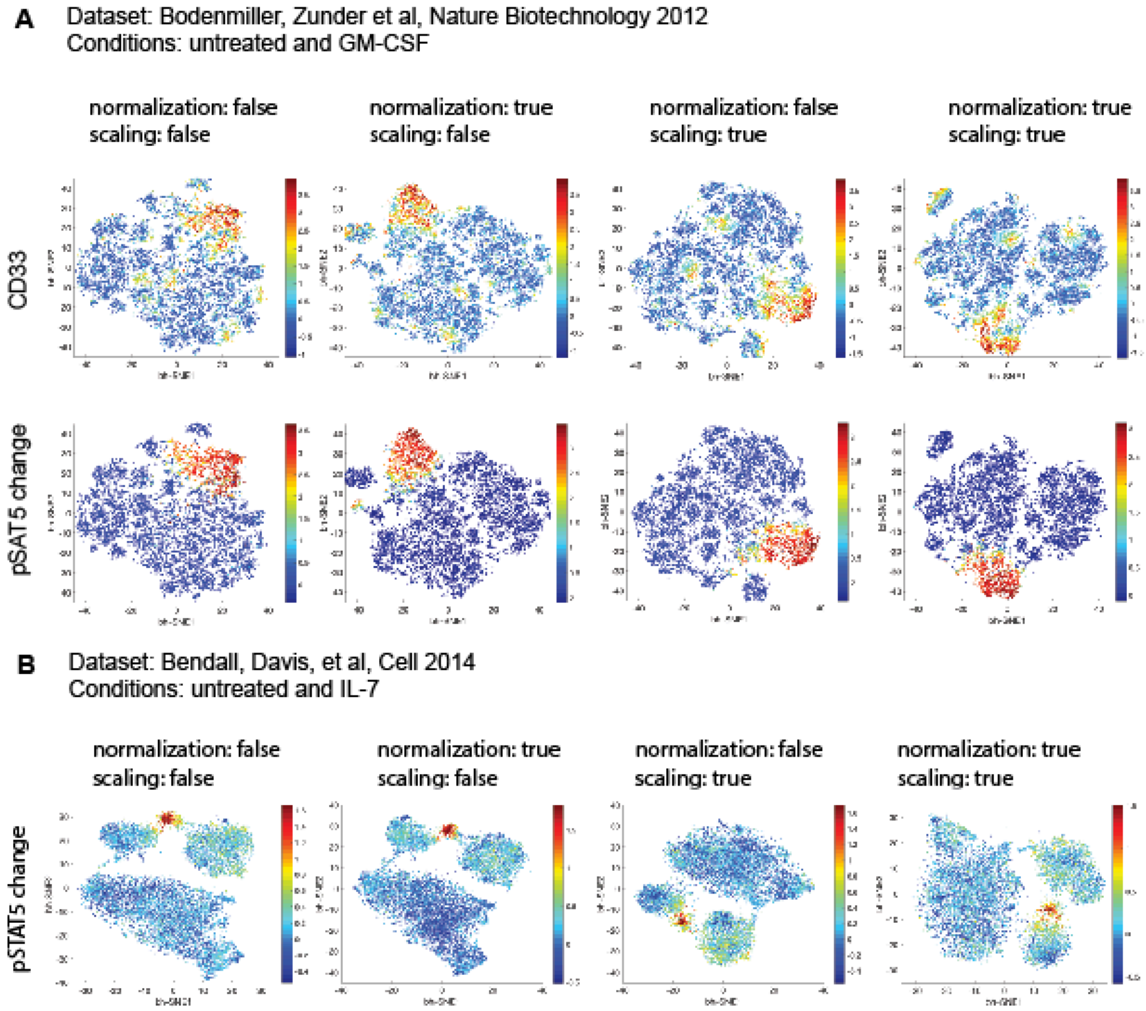


Figure S4: Quantile normalization and z score transformation (scaling) do not affect the visualization of functional populations. (A) t-SNE maps of Bodenmiller/Zunder/Finck human PBMC dataset. GM-CSF leads to a statistically significant increase in pSTAT5 (bottom row) in cells also expressing CD33 (top row). (B) t-SNE maps of Bendall/Davis/Amir B cell precursor dataset reveal a population with a statistically significant pSTAT5 increase after treatment with IL-7.

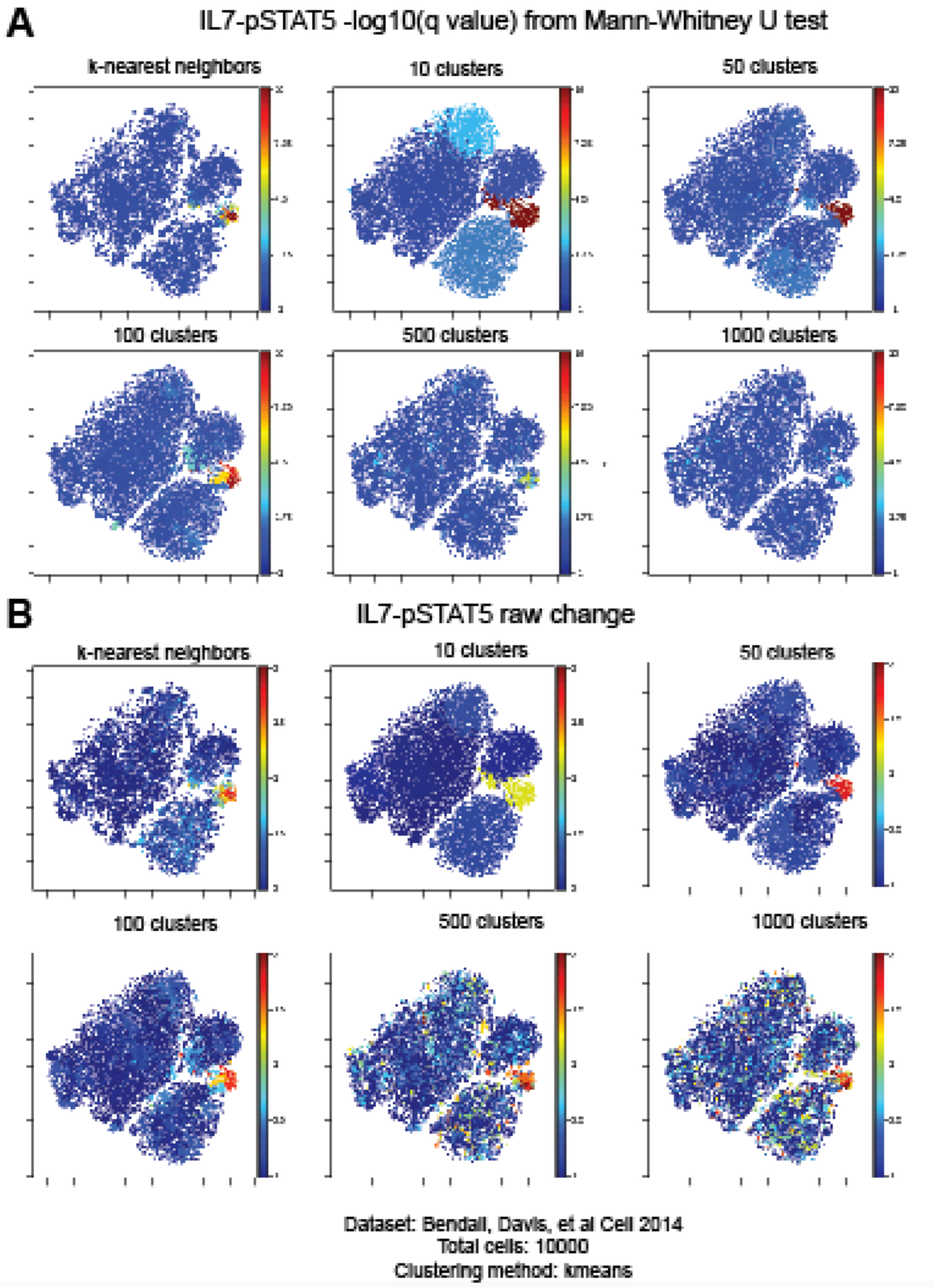


Figure S5: A comparison between k-nearest neighbors and k-means clustering for multiple biological condition analysis in the Bendall/Davis/Amir B cell precursor

dataset. (A) t-SNE maps colored by q values from the Mann-Whitney U test between each cell's k-nearest neighbors, or each cell's cluster membership for K-means clustering with a k set to values between 10 and 1000, for a 10,000 cell dataset. The k-nearest neighbors-identified IL7-pSTAT5 population is also revealed by k-means clustering, but lower number of clusters increase the apparent size of the population, and higher number of clusters lower the statistical significance of the observation. (B) t-SNE maps colored by change in pSTAT5 levels, with the same setup as in (A). A lower number of clusters reveal the k-nearest neighbors identified population, but leads to pSTAT5 change values lower than what was found with k-nearest neighbors across a larger number of cells. A higher number of clusters leads to an increase in noise across the dataset.

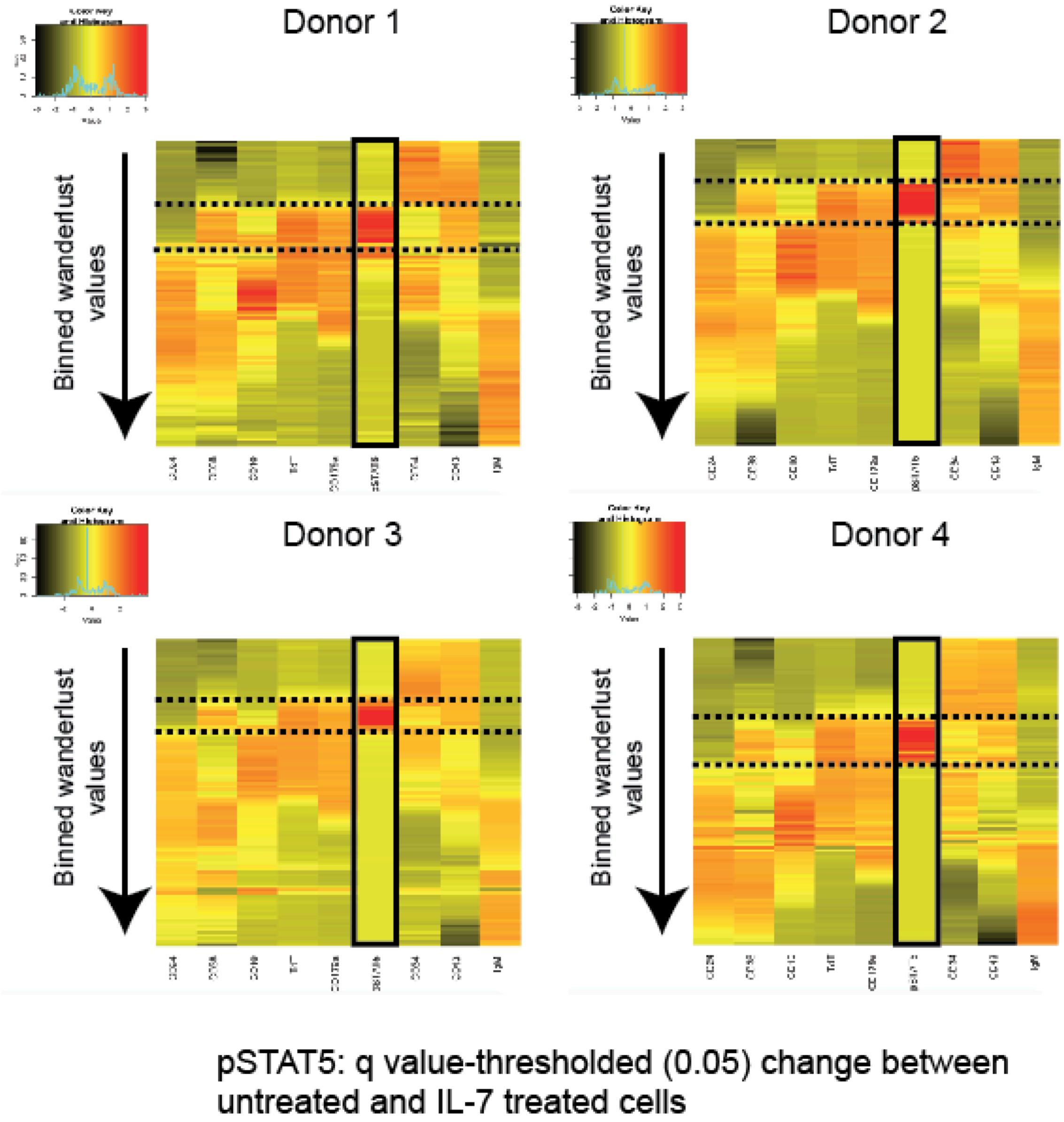


Figure S6: IL7 responsive population through pSTAT5 is consistent across four healthy human donors. Heatmaps for each donor, named donor 1-4, are shown. Wanderlust values are binned, and each bin contains the mean value of the given marker of interest for the cells in that bin. Black box indicates the change in pSTAT5 between untreated and IL7-treated cells for each cell's given k-nearest neighborhood.

Dashed lines indicate “coordination points,” where many surface markers were observed to change simultaneously.

Dataset: Fragiadakis *et al, Anesthesiology* 2015
Cells: healthy human whole blood
Conditions: Basal versus IL-10
Computations assayed in runtime analysis: Finding k-nearest neighbors, Mann-Whitney U test, q value thresholded fold change, t-SNE

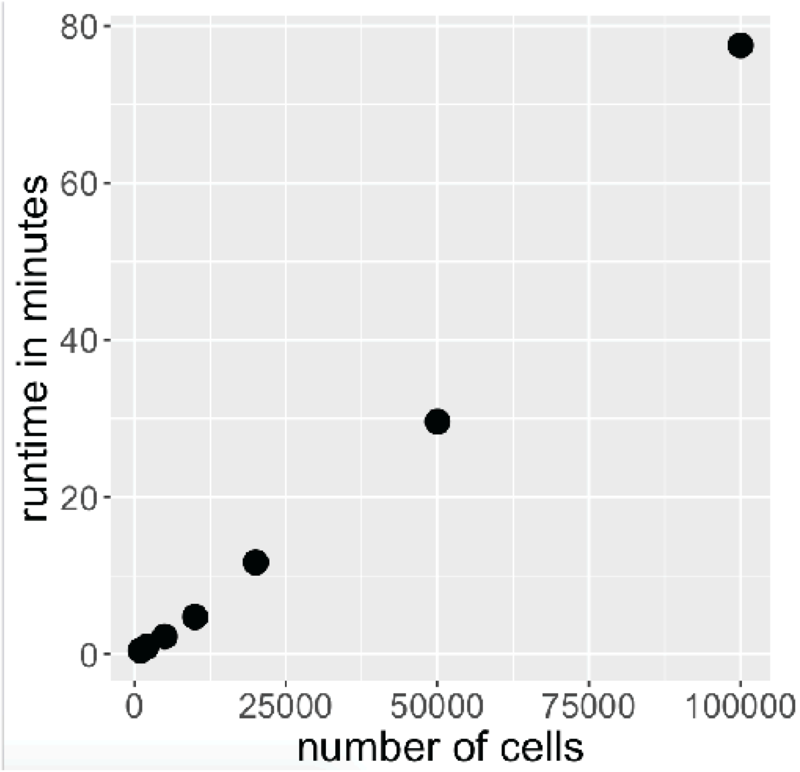


Figure S7: Runtime analysis of SCONE as a function of the number of cells. Total time in minutes includes finding KNN, performing statistical comparisons (Mann-Whitney U test, and q value thresholded change) within these nearest neighborhoods, and performing t-SNE on same features that were used as input for the KNN (in this case, surface markers).

Chapter 4: Conclusions and Perspectives

I began this thesis discussing organismal and single-cell biodiversity through the context of a longstanding genetic algorithm with a planet-sized multi-niche search space. Importantly, this gave rise not only to phenotypic expansion outward into these niches, but also expansion up the axis of complexity. At each level of organization from DNA strand to multicellular organism, new layers of emergent order arose over time as life continued to optimize itself to the environment. The work presented here centers primarily on uncovering the emergent order of the immune system, as it is sufficiently well studied that many biological facets can be validated as the new technologies are tested.

Importantly, as a cancer biologist by training who believes that it is a curable disease, I operate under the hypothesis that there exists an emergent order within cancer that is exploitable. In this context, high parameter high throughput single cell methods like CyTOF and associated algorithms like SCONE serve as powerful emergent order finders. Subsequent studies are already using CyTOF with SCONE to study leukemia and ovarian cancer. The application to cancer will bring about a new generation of cancer diagnostics and treatments that will hopefully bring us one step closer to the cure.

The body of work presented in this thesis is the result of numerous collaborations within our lab and between labs. I am grateful for the support of my advisor Garry Nolan, and the additional mentorship of Wendy Fantl throughout my graduate school trajectory.

I look forward to the future of high throughput high parameter single cell analysis. Many emerging technologies are entering this relatively new field, each filling out a specific niche. If we maintain a collaborative atmosphere moving forward, and set rigorous wet-lab and dry-lab paradigms along the bleeding edge, then the novel findings from these methods will effectively translate to the clinic. Beyond this, the emergent order we continue to find with these new technologies will enchant the minds of biologists for years to come.

Appendix A: Bibliography

1. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, Pavlov S, Vorobiev S, Dick JE, Tanner SD. Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. Anal. Chem. 2009;81:6813–6822.

2. Bendall SC, Simonds EF, Qiu P, Amir E-AD, Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI, Balderas RS, Plevritis SK, Sachs K, Pe'er D, Tanner SD, Nolan GP. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 2011;332:687–696.

3. Behbehani GK, Bendall SC, Clutter MR, Fantl WJ, Nolan GP. Single- cell mass cytometry adapted to measurements of the cell cycle. Cytometry Part A 2012;81A:552–566.

4. Behbehani GK, Samusik N, Bjornson ZB, Fantl WJ, Medeiros BC, Nolan GP. Mass cytometric functional profiling of acute myeloid leukemia defines cell cycle and immunophenotypic properties that correlate with known responses to therapy. Cancer Discov 2015.

5. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner RV, Simonds EF, Bendall SC, Sachs K, Krutzik PO, Nolan GP. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat. Biotechnol. 2012;30:858–867.

6. Zunder ER, Finck R, Behbehani GK, Amir E-AD, Krishnaswamy S, Gonzalez VD, Lorang CG, Bjornson Z, Spitzer MH, Bodenmiller B, Fantl WJ, Pe'er D, Nolan GP. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. Nature Protocols 2015;10:316–333.

7. Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, Fantl W, Pe'er D, Nolan GP, Bendall SC. Normalization of mass cytometry data with bead standards. Cytometry Part A 2013;83A:483–494.

8. Fienberg HG, Simonds EF, Fantl WJ, Nolan GP, Bodenmiller B. A platinum- based covalent viability reagent for single- cell mass cytometry. Cytometry Part A 2012;81A:467–475.

9. Spitzer MH, Gherardini PF, Fragiadakis GK, Bhattacharya N, Yuan RT, Hotson AN, Finck R, Carmi Y, Zunder ER, Fantl WJ, Bendall SC, Engleman EG, Nolan GP. An interactive reference framework for modeling a dynamic immune system. Science 2015;349:1259425.

10. Zunder ER, Lujan E, Goltsev Y, Wernig M, Nolan GP. A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry. Stem Cell 2015;16:323–337.

11. Irish JM, Hovland R, Krutzik PO, Perez OD, Bruserud O, Gjertsen BT, Nolan GP. Single cell profiling of potentiated phospho-protein networks in cancer cells. Cell 2004;118:217–228.

12. Söderberg O, Gullberg M, Jarvius M, Ridderstråle K, Leuchowius K-J, Jarvius J, Wester K, Hydbring P, Bahram F, Larsson L-G, Landegren U. Direct observation of individual endogenous protein complexes in situ by proximity ligation. Nat Meth 2006;3:995–1000.

13. Frei AP, Bava F-A, Zunder ER, Hsieh EWY, Chen S-Y, Nolan GP, Gherardini PF. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. Nat Meth 2016:1–9.

14. Burns TJ, Frei AP, Gherardini PF, Bava FA, Batchelder JE, Yoshiyasu Y, Yu JM, Groziak AR, Kimmey SC, Gonzalez VD, Fantl WJ, Nolan GP. High-throughput precision measurement of subcellular localization in single cells. Cytometry 2017:1–9.

15. Bjornson ZB, Nolan GP, Fantl WJ. ScienceDirectSingle-cell mass cytometry for analysis of immune system functional states. Current Opinion in Immunology 2013;25:484–494.

16. Jaskowiak PA, Campello RJGB, Costa IG. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics 2014;15 Suppl 2:S2–S2.

17. Aghaeepour N, Finak G, Dougall D, Khodabakhshi AH, Mah P, Obermoser G, Spidlen J, Taylor I, Wuensch SA, Bramson J, Eaves C, Weng AP, III ESF, Ho K, Kollmann T, Rogers W, De Rosa S, Dalal B, Azad A, Pothen A, Brandes A, Bretschneider H, Bruggner R, Finck R, Jia R, Zimmerman N, Linderman M, Dill D, Nolan G, Chan C, Khettabi FE, O'Neill K, Chikina M, Ge Y, Sealfon S, Sugár I,

Gupta A, Shooshtari P, Zare H, De Jager PL, Jiang M, Keilwagen J, Maisog JM, Luta G, Barbo AA, Májek P, Vilček J, Manninen T, Huttunen H, Ruusuvuori P, Nykter M, McLachlan GJ, Wang K, Naim I, Sharma G, Nikolic R, Pyne S, Qian Y, Qiu P, Quinn J, Roth A, Meyer P, Stolovitzky G, Saez-Rodriguez J, Norel R, Bhattacharjee M, Biehl M, Bucher P, Bunte K, Di Camillo B, Sambo F, Sanavia T, Trifoglio E, Toffolo G, Dimitrieva S, Dreos R, Ambrosini G, Grau J, Grosse I, Posch S, Guex N, Keilwagen J, Kursa M, Rudnicki W, Liu B, Maienschein-Cline M, Manninen T, Huttunen H, Ruusuvuori P, Nykter M, Schneider P, Seifert M, Strickert M, Vilar JMG, Hoos H, Mosmann TR, Brinkman R, Gottardo R, et al. Critical assessment of automated flow cytometry data analysis techniques. Nat Meth 2013;10:228–238.

18. Shekhar K, Brodin P, Davis MM, Chakraborty AK. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). PNAS 2014;111:202–207.

19. Qiu P, Simonds EF, Bendall SC, Gibbs KDJ, Bruggner RV, Linderman MD, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. 2011;29:886–891.

20. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. PNAS 2014;111:E2770–7.

21. Gaudillière B, Fragiadakis GK, Bruggner RV, Nicolau M, Finck R, Tingle M, Silva J, Ganio EA, Yeh CG, Maloney WJ, Huddleston JI, Goodman SB, Davis MM,

Bendall SC, Fantl WJ, Angst MS, Nolan GP. Clinical recovery from surgery correlates with single-cell immune signatures. Sci Transl Med 2014;6:255ra131–255ra131.

22. Spitzer MH, Carmi Y, Reticker-Flynn NE, Kwek SS, Madhireddy D, Martins MM, Gherardini PF, Prestwood TR, Chabon J, Bendall SC, Fong L, Nolan GP, Engleman EG. Systemic Immunity Is Required for Effective Cancer Immunotherapy. Cell 2017;168:487–502.e15.

23. Van DML. Visualizing data using t-SNE. Journal of Machine Learning Research 2008;9:2579–2625.

24. Amir E-AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. 2013;31:545–552.

25. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nat Meth 2016;13:493–496.

26. Bendall SC, Davis KL, Amir E-AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 2014;157:714–725.

27. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, Levenson RM, Lowe JB, Liu SD, Zhao S, Natkunam Y, Nolan GP. Multiplexed ion beam imaging of human breast tumors. Nat Med 2014.

28. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 2015;161:1202–1214.

29. Basiji DA, Ortyn WE, Liang L, Venkatachalam V, Morrissey P. Cellular Image Analysis and Imaging by Flow Cytometry. Clinics in Laboratory Medicine 2007;27:653–670.

30. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. Trends in Immunology 2012;33:323–332.

31. Heim R, Tsien RY. Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. Curr Biol 1996;6:178–182.

32. Rossi F, Charlton CA, Blau HM. Monitoring protein-protein interactions in intact eukaryotic cells by beta-galactosidase complementation. Proc Natl Acad Sci U S A 1997;94:8405–8410.

33. Ramdzan YM, Polling S, Chia CPZ, Ng IHW, Ormsby AR, Croft NP, Purcell AW, Bogoyevitch MA, Ng DCH, Gleeson PA, Hatters DM. Tracking protein aggregation and mislocalization in cells with flow cytometry. Nat Meth 2012;9:467–470.

34. Jackson SP, Forment JV. A flow cytometry&ndash;based method to simplify the analysis and quantification of protein association to chromatin in mammalian cells.

Nature Protocols 2015;10:1297–1307.

35. Krutzik PO, Clutter MR, Trejo A, Nolan GP. Fluorescent Cell Barcoding for Multiplex Flow Cytometry. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2001:1–22. 22 p.

36. Finak G, Perez J-M, Weng A, Gottardo R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinformatics 2010;11:546–13.

37. Leuchowius K-J, Clausson C-M, Grannas K, Erbilgin Y, Botling J, Zieba A, Landegren U, Söderberg O. Parallel visualization of multiple protein complexes in individual cells in tumor tissue. Mol. Cell Proteomics 2013;12:1563–1571.

38. Weibrecht I, Gavrilovic M, Lindbom L, Landegren U, Wählby C, Söderberg O. Visualising individual sequence-specific protein–DNA interactions in situ. New BIOTECHNOLOGY 2012;29:589–598.

39. Weibrecht I, Lundin E, Kiflemariam S, Mignardi M, Grundberg I, Larsson C, Koos BOR, Nilsson M, derberg OSO. In situ detection of individual mRNA molecules and protein complexes or post-translational modifications using padlock probes combined with the in situ proximity ligation assay. Nature Protocols 2013;8:355–372.

40. Leuchowius K-J, Weibrecht I, Landegren U, Gedda L, Söderberg O. Flow cytometric in situproximity ligation analyses of protein interactions and post-translational modification of the epidermal growth factor receptor family. Cytometry

2009;75A:833–839.

41. Jarvius M, Paulsson J, Weibrecht I, Leuchowius K-J, Andersson A-C, Wählby C, Gullberg M, Botling J, Sjöblom T, Markova B, Ostman A, Landegren U, Söderberg O. In situ detection of phosphorylated platelet-derived growth factor receptor beta using a generalized proximity ligation method. Mol. Cell Proteomics 2007;6:1500–1509.

42. Johnson RF, Perkins ND. Nuclear factor-κB, p53, and mitochondria: regulation of cellular metabolism and the Warburg effect. Trends in Biochemical Sciences 2012;37:317–324.

43. Sabroe I, Jones EC, Usher LR, Whyte MKB, Dower SK. Toll-like receptor (TLR)2 and TLR4 in human peripheral blood granulocytes: a critical role for monocytes in leukocyte lipopolysaccharide responses. J Immunol 2002;168:4701–4710.

44. Tay S, Hughey JJ, Lee TK, Lipniacki T, Quake SR, Covert MW. Single-cell NF. Nature 2010;466:267–271.

45. Nolan GP, Ghosh S, Liou H-C, Tempst P, Baltimore D. DNA binding and IκB inhibition of the cloned p65 subunit of NF-κB, a rel-related polypeptide. Cell 1991;64:961–969.

46. Scott ML, Fujita T, Liou HC, Nolan GP, Baltimore D. The p65 subunit of NF-kappa B regulates I kappa B by two distinct mechanisms. Genes & Development 1993;7:1266–1276.

47. Feng L, Li N, Li Y, Wang J, Gao M, Wang W, Chen J. Cell cycle-dependent

inhibition of 53BP1 signaling by BRCA1. Nature Publishing Group 2015:1–11.

48. Tarsounas M, Davies D, West SC. BRCA2-dependent and independent formation of RAD51 nuclear foci. Oncogene 2003;22:1115–1123.

49. Scully R, Chen J, Ochs RL, Keegan K, Hoekstra M, Feunteun J, Livingston DM. Dynamic changes of BRCA1 subnuclear location and phosphorylation state are initiated by DNA damage. Cell 1997;90:425–435.

50. Pageau GJ, Lawrence JB. BRCA1 foci in normal S-phase nuclei are linked to interphase centromeres and replication of pericentric heterochromatin. The Journal of Cell Biology 2006;175:693–701.

51. Lord CJ, Ashworth A. BRCAness revisited. Nature Publishing Group 2016:1–11.

52. Turner N, Tutt A, Ashworth A. Hallmarks of "BRCAness" in sporadic cancers. Nat Rev Cancer 2004;4:814–819.

53. Krishan A. Rapid flow cytofluorometric analysis of mammalian cell cycle by propidium iodide staining. The Journal of Cell Biology 1975;66:188–193.

54. Ward JF. DNA Damage Produced by Ionizing Radiation in Mammalian Cells: Identities, Mechanisms of Formation, and Reparability. In: Vol 35. Progress in Nucleic Acid Research and Molecular Biology. Elsevier; 1988. p 95–125.

55. Cruz-García A, López-Saavedra A, Huertas P. BRCA1 Accelerates CtIP-Mediated DNA-End Resection. CellReports 2014;9:451–459.

56. Bunting SF, CallEn E, Wong N, Chen H-T, Polato F, Gunn A, Bothmer A, Feldhahn N, Fernandez-Capetillo O, Cao L, Xu X, Deng CX, Finkel T, Nussenzweig M, Stark JM, Nussenzweig A. 53BP1 Inhibits Homologous Recombination in Brca1-Deficient Cells by Blocking Resection of DNA Breaks. Cell 2010;141:243–254.

57. Yang J, Bardes ES, Moore JD, Brennan J, Powers MA, Kornbluth S. Control of cyclin B1 localization through regulated binding of the nuclear export factor CRM1. Genes & Development 1998;12:2131–2143.

58. Shimura T, Kobayashi J, Komatsu K, Kunugita N. DNA damage signaling guards against perturbation of cyclin D1 expression triggered by low-dose long-term fractionated radiation. 2014;3:e132–8.

59. Graeser M, McCarthy A, Lord CJ, Savage K, Hills M, Salter J, Orr N, Parton M, Smith IE, Reis-Filho JS, Dowsett M, Ashworth A, Turner NC. A marker of homologous recombination predicts pathologic complete response to neoadjuvant chemotherapy in primary breast cancer. Clin Cancer Res 2010;16:6159–6168.

60. Farmer H, McCabe N, Lord CJ, Tutt ANJ, Johnson DA, Richardson TB, Santarosa M, Dillon KJ, Hickson I, Knights C, Martin NMB, Jackson SP, Smith GCM, Ashworth A. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 2005;434:917–921.

61. Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the

human mouth. Proc Natl Acad Sci U S A 2007;104:11889–11894.

62. Tutte WT. How to Draw a Graph. Proceedings of the London Mathematical Society 1963;s3-13:743–767.

63. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research 2008;9:2579–2625.

64. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6 1901;2:559–572.

65. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, Saeys Y. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data Brinkman RR, Aghaeepour N, Finak G, Gottardo R, Mosmann T, Scheuermann RH, editors. Cytometry 2015;87:636–645.

66. Arvaniti E, Claassen M. Sensitive detection of rare disease-associated cell subsets via representation learning. Nat Commun 2017;8:1–10.

67. Lun ATL, Richard AC, Marioni JC. Testing for differential abundance in mass cytometry data. Nature Publishing Group 2017:1–5.

68. Fragiadakis GK, Gaudillière B, Ganio EA, Aghaeepour N, Tingle M, Nolan GP, Angst MS. Patient-specific Immune States before Surgery Are Strong Correlates of Surgical Recovery. Anesthesiology 2015;123:1241–1255.

69. Qiu X, Wu H, Hu R. The impact of quantile and rank normalization procedures on

the testing power of gene differential expression analysis. BMC Bioinformatics 2013;14:124.

70. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statistics 1947;18:50–60.

71. STUDENT. The probable error of a mean. Biometrika 1908;6:1–25.

72. Bak N, Hansen LK. Data Driven Estimation of Imputation Error?A Strategy for Imputation with a Reject Option Zhang Z, editor. PLoS One 2016;11:e0164464–13.

73. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms. Journal of Machine Learning Research 2015;15:3221–3245.

74. de Villartay J-P, Fischer A, Durandy A. The mechanisms of immune diversification and their disorders. Nat Rev Immunol 2003;3:962–972.