ABSTRACT

Title of Dissertation:	MULTI-DIMENSIONAL ANALYSIS APPROACHES FOR HETEROGENEOUS SINGLE-CELL DATA
	Yang Shen, Doctor of Philosophy, 2018
Dissertation directed by:	Professor Wolfgang Losert, Department of Physics

Improvements in experimental techniques have led to an explosion of information in biology research. The increasing number of measurements comes with challenges in analyzing resulting data, as well as opportunities to obtain deeper insights of biological systems. Conventional average based methods are unfit to analyze high dimensional datasets since they fail to take full advantage of such rich information. More importantly, they are not able to capture the heterogeneity that is prevalent in biological systems. Sophisticated algorithms that are able to utilize all available measurements simultaneously are hence emerging rapidly. These algorithms excel at making full use of information within datasets and revealing detailed heterogeneity.

However, there are several important disadvantages of existing algorithms. First, specific knowledge in statistics or machine learning is required to appropriately interpret and tune parameters in these algorithms for future use. This may result in misusage and misinterpretation. Second, using all measurements with equal weighting

runs the risk of noise contamination. In addition, information overload has become more common in biology research, with a large volume of irrelevant measurements. Third, regardless of the quality of measurements, analysis methods that simultaneously use a large number of measurements need to avoid the “curse of dimensionality”, which warns that distance estimation and nearest neighbor estimation are not meaningful in high dimensional space. However, most current sophisticated algorithms involve distance estimation and/or nearest neighbor estimation.

In this dissertation, my goal is to build analysis methods that are complex enough to capture heterogeneity and at the same time output results in a format that is easy to interpret and familiar to biologists and medical researchers. I tackle the dimension reduction problem by finding not the best subspace but dividing them into multiple subspaces and examine them one by one. I demonstrate my methods with three types of datasets: image-based high-throughput screening data, flow cytometry data, and mass cytometry data. From each dataset, I was able to discover new biological insights as well as re-validate well-established findings with my methods.

MULTI-DIMENSIONAL ANALYSIS APPROACHES FOR HETEROGENEOUS SINGLE-CELL DATA

by

Yang Shen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Wolfgang Losert, Chair
Professor Hector Bravo
Dr. Benjamin Chaigne-Delalande
Professor Christopher Jarzynski
Dr. Richard Lee
Professor Arpita Upadhyaya

Acknowledgments

I would like to extend thanks to the many people, who generously contributed to work presented in this dissertation and who kindly supported me through my journey to Ph.D.

Special thanks go to my great advisor Professor Wolfgang Losert, who has encouraged and trained me towards an independent thinker and researcher. My journey to Ph.D. has been bumpy at times, but Prof. Losert has always been very supportive and responsive, making me feel secure and free to explore my research interests.

Similarly, profound gratitude goes to Dr. Benjamin Chaigne-Delalande and Dr. Richard Lee. They have been truly dedicated mentors, always available when I need a discussion, giving me very useful feedbacks from biological and clinical points of view. What’s more, they are always excited about my work, making me feel encouraged and motivated. A special mention goes to Dr. Richard Lee, who flew halfway around the world to attend my defense.

I am also grateful to Dr. Robert Nussenblatt, who had been a very enthusiastic mentor and very patiently guided me to a better appreciation of clinical research.

Special mention goes to Leonard Campanello, Julian Candia, Desu Chen, Sima Hirani, Nard Kubben, Rachel Lee, Zhixin Lu, Sebastian Schmidt, Maggie Wei, Haiqing Zhao, and many more people for all the useful and interesting discussions. And to Rachel Caspi, Michael Coplan, Debbie Jenkins, Zhiyu Li, Baoying Liu, Tom Misteli, Alexandra Morozov and Nida Sen for going far beyond the call of duty.

Last but by no means least, thanks go to mom, dad, my fiancé, Hsiang-Ling Hsiao, Ying-Shan Su and Yurong He for their unbelievable emotional and moral support. They have made this journey more enjoyable and extra meaningful for me, and I am truly grateful.

Table of Contents

List of Tables

List of figures

List of Abbreviations

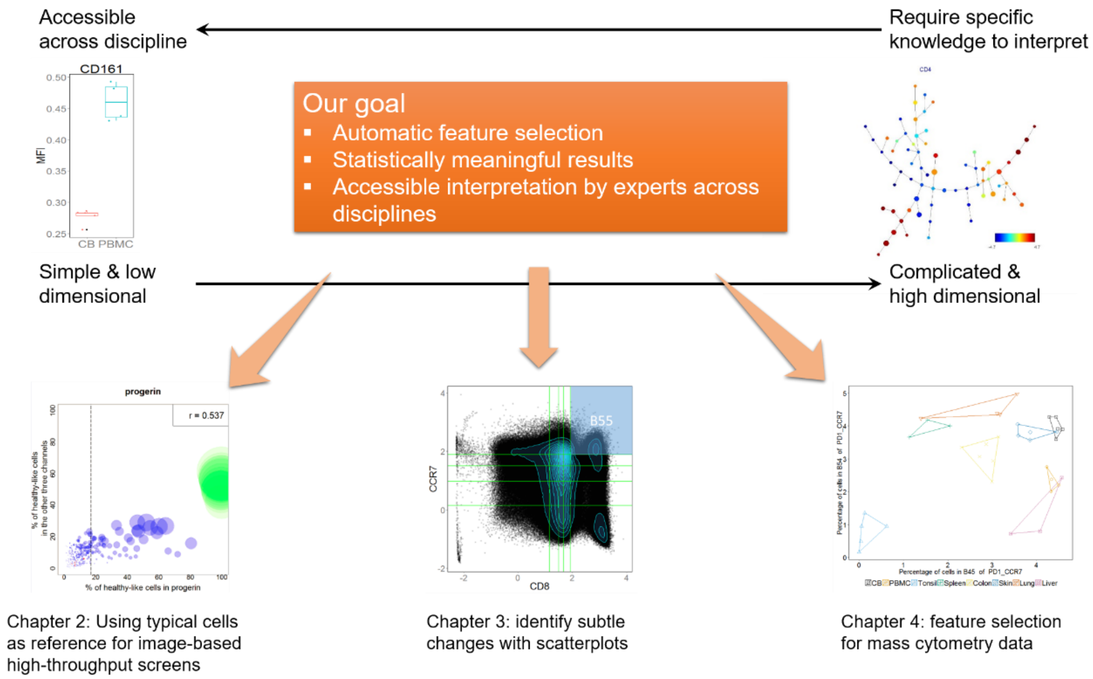
HCS	High Content Screening
HTS	High Throughput Screening
CyTOF	Cytometry by Time Of Flight
SVM	Support Vector Machine
EM	Expectation Maximization
FSC	Forward Scattering Center
SSC	Side Scattering Center
FCS	Flow Cytometry Standard
ROI	Region Of Interest
MST	Minimum Spanning Tree
KNN	k Nearest Neighbors
t-SNE	t distribution based stochastic neighorbood embedding
HGPS	Hutchinson-Gilford progeria syndrome
GMM	Gaussian Mixture Models
GFP	Green Fluorescent Protein
MI	Mutual Information
FPR	False Positive Rate
MFI	Mean Fluorescent Intensity
PBMC	Peripheral Blood Mononuclear Cell

Chapter 1 Introduction

1.1 Overview

Biological systems are complex and cope with a complex and ever-changing environment. Hence, emerging multi-parametric measurement techniques with single-cell resolution reveal increasing complexity among and within cell types. For example, one of the most widely used single-cell technique – flow cytometry – is responsible for the discovery of the diverse types and functionality of immune cells [1]. Novel cell (sub)-types are continually being identified with the increasing number of cellular markers that can be simultaneously measured with flow cytometry (e.g., Th17 cells, follicle T helper cells, and stem memory T cells) [2-4]. Today, sophisticated experimental techniques enable in-depth, high-content and high-throughput single cell measurements in almost every field of biological research. As an example, the number of parameters that can be measured by flow cytometry increased from the original 2 ~ 3 to about 20, and will likely reach 50 in the near future [5]. Mass cytometry, which is a combinatorial technique based on mass spectrometry and flow cytometry further expands this number to 100 [6]. In addition, people have combined high-throughput screening techniques with modern microscopes to get cell images at single cell or even sub-cellular resolution from which hundreds of measurements can be made [7]. On the other hand, some techniques which are capable of very high-dimensional measurements (e.g., RNA sequencing) have increased their resolution to single cell level, enabling tens of thousands measurements on several thousand of single cells [8]. The increasing ability to generate a large amount of high-dimensional single cell data demands new analysis methods suitable for this type of data.

The challenges of analyzing high-dimensional single cell data originate from two main factors: heterogeneity and high dimensionality. These two factors are in fact intertwined: to understand heterogeneous system requires more measurements and more measurements, in turn, reveal higher heterogeneity. For the first challenge, not all heterogeneities have biological significance [9]. Since the local environment for each cell is slightly different from another, the properties of cells, even from the same cellular type/state, would vary from cell to cell. There have been speculations that cellular status is not a collection of distinct states but a continuous change in response to outer stimulations [9]. However, it is not clear under what circumstances this heterogeneity is meaningful for the biological functionality of the cell or organism. In addition, experimental measurements are not accurate, random noise and systematic errors are ubiquitous, and all contribute to the heterogeneous signal we observe in experiments. As a result, the appropriate analysis methods should be able to capture the biologically meaningful heterogeneity of single cells while at the same time reduce the impact of random noise in experiments. For the second challenge, it has been shown that in high dimensional space the distance between two points “may not even be qualitatively meaningful” [10] – a phenomenon termed as “curse of dimensionality”, rendering any distance-based analysis (e.g., clustering) problematic. In addition, information overload is ubiquitous in biological datasets, with many redundant measurements or measurements that are irrelevant to the question of interest (or of unknown relevance). This adds to the already high noise level in biological datasets and further hinders the finding of real signals in data. As a result, dimension reduction methods are required for analyzing high dimensional data. Since most dimension reduction methods aim to identify a single subspace best capture the content of original data, the results of reduced dimensions are often linear or nonlinear combinations of all original measurements which impedes straightforward interpretation of data. What's more, reducing high dimensional datasets to one definite subspace may result in overlook of small cell groups or subtle changes. Hence designing and applying dimension reduction methods to biological data should take interpretability and importance of subtle changes into consideration. Currently, most algorithms developed to deal with high-dimensional single cell data largely ignore information overload and “curse of dimensionality”. Instead, they fully embrace the large number of measurements and try to make full use of them (except algorithms to analyze single-cell RNA sequencing data where the problem of high dimensionality is extremely prominent). However, there has been a study showing that algorithms like this would pick up noise in data and lead to incorrect results [11]. In addition, these algorithms often produce results that are hard to interpret and hence difficult to experimentally validate. Thus, there is a need to build analysis methods between the two extremes of conventional methods, which tend to be oversimplified, average-based, and involve manual selection, and novel methods, which are complex and sensitive to noise and the “curse of dimensionality”. Here I introduce methods that are complex enough to capture some level of heterogeneity, but at the same time provide outputs suitable for clear interpretation. The methods I introduce also incorporate feature selection to avoid the “curse of dimensionality” and deal with information overload.



Accessible across discipline
Require specific knowledge to interpret
CD161
MFI
CB PBMC
Our goal
▪ Automatic feature selection
▪ Statistically meaningful results
▪ Accessible interpretation by experts across disciplines
Simple & low dimensional
Complicated & high dimensional
progerin
r = 0.537
% of healthy-like cells in the other three channels
% of healthy-like cells in progerin
CCR7
CD8
B55
CB PBMC Tonsil Spleen Colon Skin Lung Liver
Chapter 2: Using typical cells as reference for image-based high-throughput screens
Chapter 3: identify subtle changes with scatterplots
Chapter 4: feature selection for mass cytometry data

Figure 1.1 Summary of this dissertation. Conventional average based, single variate analysis methods are not fitted to analyze high-dimensional single-cell data generated by current experimental techniques. Novel sophisticated methods developed specifically to incorporate high dimensionality can be too complicated for experts outside the field of bioinformatics/statistics to interpret and require specific knowledge of the model for them to appropriately tune the parameters for applying to their own data. In addition, these methods are sensitive to noise and may suffer from a more fundamental problem: the curse of dimensionality. The goal of this dissertation is to build analysis methods that are complicated enough to capture heterogeneity inside single-cell data, but not so complicated that impedes experts from different disciplines to understand and apply them to their own data.

In this dissertation, I am going to address the challenges in analyzing high-dimensional single cell data in two parts. The first part deals with the functionally non-meaningful heterogeneity and aims to develop a method that captures similarity within cells that are various but don't form functionally distinct subsets. The second part handles high dimensionality in flow and mass cytometry data by dividing the data into small subspaces and then combining information in useful subspaces. In all, this dissertation aims to develop methods that are able to automatically select important features, generate statistically meaningful results that are accessible to experts across disciplines for analyzing high-dimensional single-cell data.

1.1 High-dimensional single-cell measurement techniques

1.1.1 Image-based high-throughput screening/ High Content Screening (HCS)

High-throughput screening (HTS) is a powerful technique routinely used in drug discovery, systematic analysis of cellular functions, and exploration of gene regulation pathways [12-14]. The use of robotics instead of manual operation in HTS greatly reduces experimental variability, yet cellular heterogeneities persist. Combined with modern automated microscopes, image-based HTS (also known as high content screening, or HCS) allows for routine imaging of thousands of cells in multiple fluorescence channels [15-17]. HCS is one of the fastest growing techniques in biological and pharmaceutical research and is responsible for countless discoveries [18-21].

1.1.1.1 Experimental procedure

The typical experimental procedure of high content screening requires the following steps: sample preparation, image acquisition, and data handling [16]. Every step comes with its own challenges and requires proper planning with the overall goal of the experiment in mind. For sample preparation, it is important to figure out what type of cells is suitable for the experiment, how many samples (perturbations) should be tested, what perturbation reagents and/or fluorescent stain should be used, etc. [15]. Small-scale pilot screens are recommended before running the much larger whole screen. This step helps clear additional issues, test validity of the screen, optimize parameters in the experiment, and sometimes serves as a proof of concept for the experimental design. Sample preparation usually takes several months, much longer than the actual image acquisition process. The quality of image acquisition directly affects the overall quality of the screen. Thus it is crucial to make sure that parameters for image acquisition are set appropriately. For example, the resolution of image acquisition should be high enough to capture all required features, but the imaging should also cover a large enough area to capture a sufficient number of cells for robust statistics [15]. Downstream data analysis after image acquisition is also an important part of image-based high throughput screening, and usually takes more time than the experimental process.

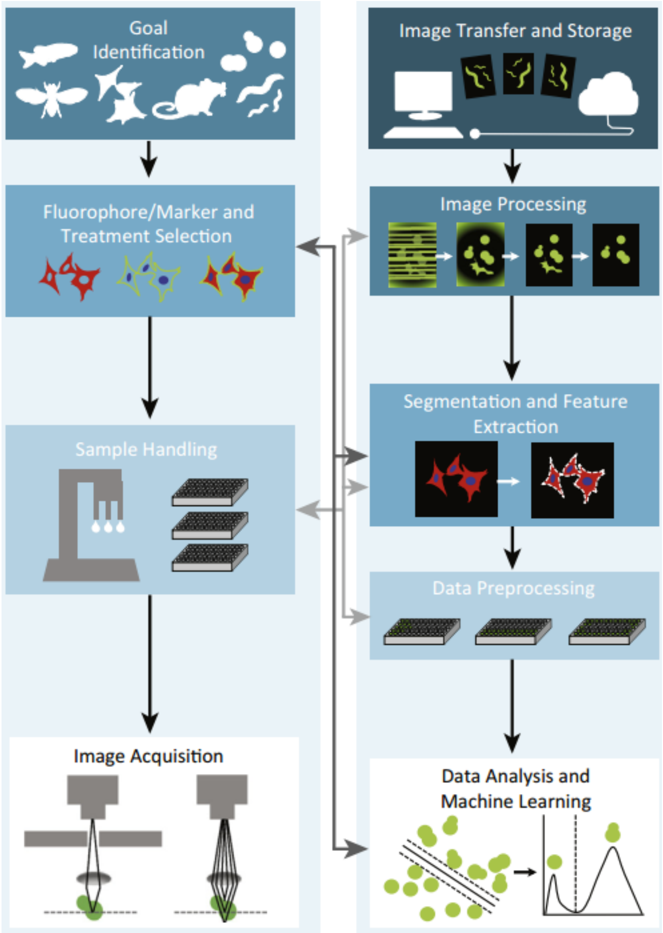

Goal Identification
Fluorophore/Marker and Treatment Selection
Sample Handling
Image Acquisition
Image Transfer and Storage
Image Processing
Segmentation and Feature Extraction
Data Preprocessing
Data Analysis and Machine Learning

Figure 1.2 Workflow of image-based high throughput screening (adapted from [16]). Image-based high throughput screening produces a large volume of single cell resolution image data. A great number of cellular features can be extracted from images. Due to noise in both the imaging and image analysis process, and due to natural variability, features extracted from images may have non-ignorable variation even for cells with

the same biological functionality. Hence, in order to understand the behavior of a certain cellular state, a method to reduce noise level and identify a typical description of the cellular state is needed. This is the problem I'm going to address in Chapter 2.

1.1.1.2 Data analysis

Data analysis for HCS can be divided into two parts: image analysis and downstream statistical analysis. Until recently manual scoring of images dominated the analysis of low- and medium throughput image screening studies, and it is still widely used today [16]. Building on the active and fast-growing field of computer vision, many algorithms have been developed for bio-image analysis. Some general purpose image analysis algorithms are included in widely used image processing programs such as ImageJ [22, 23], Fiji [24], and CellProfiler [25-28]. In addition to the general purpose image analysis platforms, a large number of algorithms have been developed for tackling specific image analysis problems in a specialized field of biology, especially in neuroscience. In fact, there are many specialized algorithms that can be found in several online listings, for example, image analysis tools for neuro-images (http://nitrc.org/) [29].

After images are analyzed, and features are measured, the second part of HCS data analysis utilizes statistical tools or more recently machine learning tools to phenotype the measured cells and reveal biological insights. Unfortunately, compared to the sophisticated image analysis tools, downstream data analysis methods are still "not high content" enough [30]. Although hundreds of features can be routinely extracted in image analysis steps, only a few of these features can be used in conventional downstream analysis procedures [16]. Conventional methods further oversimplify the selected few features by feature averaging, ignore rich heterogeneity among cells. This situation has improved recently, with the surge of analysis methods that incorporates supervised and unsupervised machine learning analysis [31]. Some of the more recent approaches are starting to use multiple metrics via classifiers such as support vector machine (SVM) [32], hierarchical clustering [33], EM clustering with Gaussian mixture models to identify multiple cell subtypes to quantify cellular heterogeneity [34], or deep learning [35]. One caveat to keep in mind is that although the number of features that can be measured from images is large, some of these features can be redundant or irrelevant or both. So it is important to carry out feature selection along with machine learning.

1.1.2 Flow cytometry

Flow cytometry is a widely used technique that measures multiple parameters of fluorescently labeled single cells that flow in a stream through a photodetector system. Flow cytometry is routinely used in both basic biological and clinical research, due to its fast speed (50,000 cells/s [1]), relatively low cost, and amenability to standardization [36]. The past 50 years have yielded great improvements in flow cytometry techniques, in particular, the recent increase from 2 to 30 in the number of markers that can be measured simultaneously. The impact of this increase is far-reaching for our understanding of biological systems (including the immune system, cancer biology, cell signaling, etc.) and for medical applications and drug development [37]. Even with the emergence of other, more powerful, single-cell measurement techniques, flow cytometry is still the most widely used experimental technique.

1.1.2.1 Experimental Procedure

Flow cytometers can be roughly divided into three parts: fluidics, optics, and detectors. The fluidics is a cell distribution system that directs single cells in the sample to the light source. It contains two parts: sheath fluid and pressured line. A pressured airline pump suspends cells from the sample tube into a flow chamber which is filled with sheath fluid.

Since the pressure of sheath fluid is always lower than the sample flow, this process ensures cells align as a coaxial flow as they pass through several light sources (Fig 1.3a). The optics system is where measurements of the fluorescence intensity of each cell take place simultaneously or in rapid succession for different fluorescence wavelengths. In addition to fluorescent signal, two types of scattered light are also detected: forward scattering (FSC) and side scattering (SSC) (Fig 1.3b). The intensity of FSC is proportional to the size of the cell, a valuable piece of information often used in Immunophenotyping. SSC is the result of diffraction of light perpendicular to the incoming laser beam. SSC is proportional to cell granularity, i.e., the inner complexity of the cell. Together, FSC and SSC provide information to coarsely divide cells (mainly white blood cells) into basic phenotypes (differentiating lymphocytes from macrophages for example). The detectors detect the fluorescent signal and convert it to a digital data that is proportional to light intensity. For surface marker detection, a logarithmic amplifier is used to capture the wide range of surface marker expression levels. [38, 39] The resulting data is stored in a standard manner according to flow cytometry standard (FCS), and the file format is hence called .fcs [40, 41].

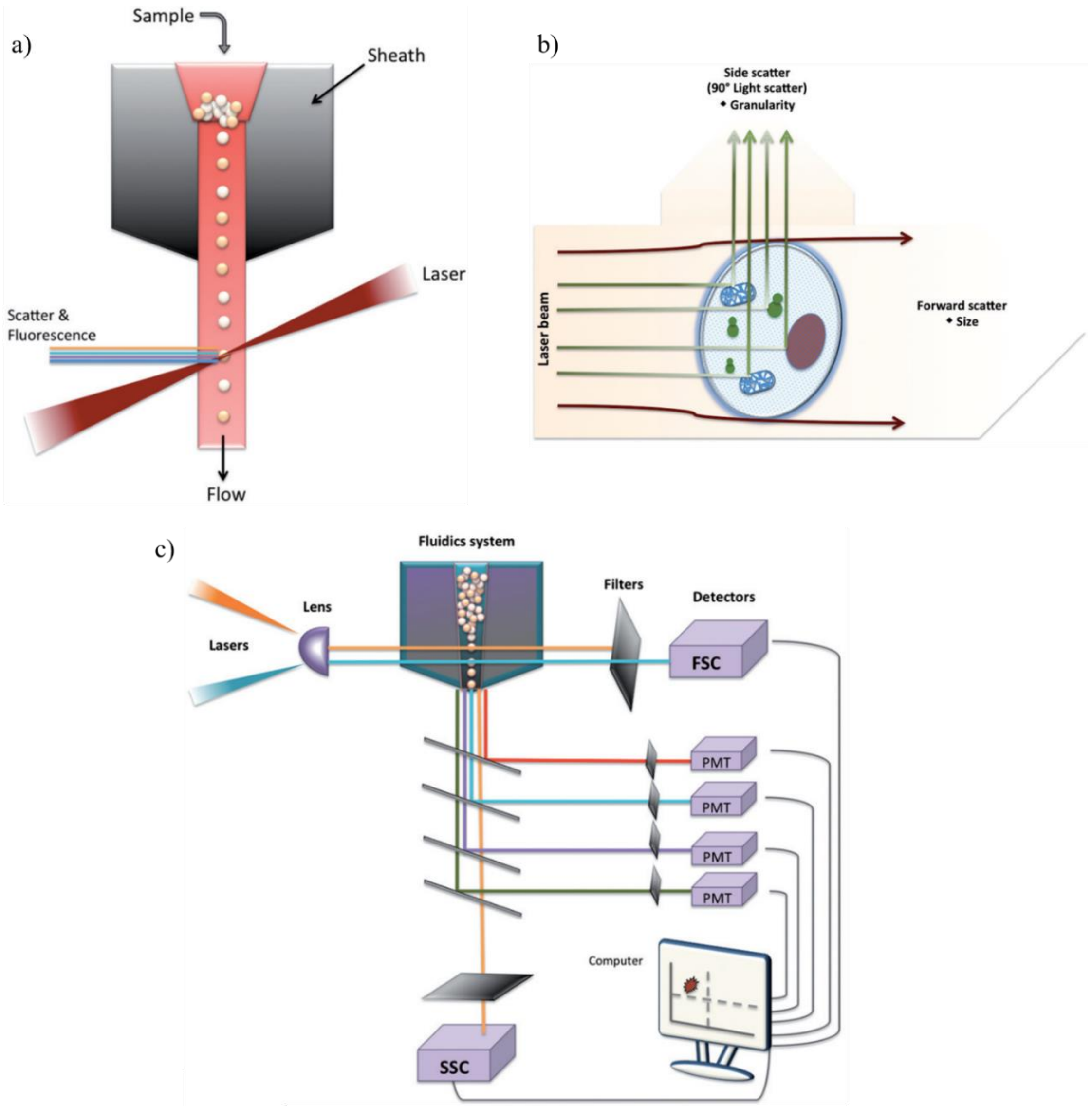


Figure 1.3 Fluorescent-based high throughput multi-dimensional quantification of protein expression (adapted from [38]). a) A cartoon for the fluidic system in a flow cytometer. The pressure difference between the sample chamber and sheath enables the flow of single cells past the light source(s), which enables high-throughput single cell measurements. b) An illustration for forward scatter (FSC) and side scatter (SSC). c) A cartoon showing how fluorescent measurements are made. With the development of brighter dyes and the improvement in differentiating light spectrum, more and more fluorescent markers can be measured simultaneously with flow cytometry.

1.1.2.2 Data processing

Because of the amplification process, raw data in FCS files has a wide range (~0 – 10^5) with a dominant peak near zero. Hence proper transformation algorithms are needed to make sense of the data [42]. Traditionally, flow cytometry data are presented on a logarithmic scale, and the so-called “negative populations” are near zero. There have been lots of studies and debate about the most appropriate way to transform and visualize flow cytometry data. Currently, the widely accepted and used method is an algorithm called “logicle transformation” [43]. In essence, the logicle transformation is a generalization of the hyperbolic sine function with tunable parameters to better suit flow cytometry data. These parameters can be determined based on the range of specific datasets. Another important aspect of pre-processing flow cytometry data is called compensation. The need for compensation arises from the fact that the wavelength of fluorescent light is a distribution instead of a single value. Hence there will be overlap in the light spectrum one of fluorochromes. Overlap in the fluorescent spectrum can result in false signals and false correlations between measurements, and it is a major source of error in flow cytometry. The process to detangle the spectrum overlap is called compensation. The currently used compensation method operates under the assumption that overlap of spectrum results in a linear combination of various light spectrums. Thus detangling the overlap equals solving a set of linear equations. Coefficients in these equations are stored in a matrix called compensation matrix and can be obtained via experiments. In addition to transformation and compensation, batch effects can be a real problem for flow cytometry data as well. Especially when combining samples tested in different experimental cores. Thus, it is important for the experimentalist to follow standard protocols, though algorithms designed to reduce batch effects exist [44]. These pre-processing steps are usually done manually in a commercial software platform like flowJo, though a package in R [45] that provides commands for automatic batch processing recently became available.

1.1.2.3 Data analysis

The traditional analysis method for flow cytometry data, called “gating”, involves manually drawing regions of interest (ROI) through a sequence of 2D scatter plots (Fig 1.4). Manual gating has long been criticized as being “subjective and biased”, “lack of reproducibility”, “inefficient and time consuming”, and has been considered one of the major source of variability in flow cytometry analyses [5, 46-48]. With the increasing number of measurements and the large volume of data to be analyzed, unbiased, automated, and reproducible algorithms are in great need for flow cytometry data analysis. Building such data mining tools has been called the “real challenge” in flow cytometry application [49, 50]. The last decade has seen a surge in new data analysis methods for flow cytometry [36, 51-53]. One important goal in flow cytometry data analyzing is identifying structures / cell clusters in the high-dimensional space span by all measurements. There are algorithms trying to achieve this by combining density based thresholding in 1D [54, 55] hence exhaustively identifying all possible cell subsets. Other algorithms apply clustering algorithms for cell subset finding, e.g., k-means clustering [56, 57] or mixture modeling [58-61], etc. Tree or map-based algorithms have also been developed to mine the structure of flow cytometry data [62-64]. Finally, binning methods that aim to resolve the position of density peak have been proposed as well [65-67]. There has also been an effort to combine multiple algorithms in a workflow so that researchers can pick the method that best fits their needs [68]. With a large number of cell subtypes discovered by automated algorithms, methods to compare the biological importance of cell types have been developed [69, 70], and so are methods to automatically label the ontology of cell subtypes [71]. In addition to cell type identification, visualization is also an important aspect of interpreting flow cytometry data, especially when the number of measurements is large.

Multiple visualization tools have been developed for this purpose [72-74]. Most clustering-based methods that allow novel cell type discovery aim to identify regions with high cell density in multi-dimensional space [56, 58-60, 63, 66, 70, 72, 73, 75]. This assumes cells form distinct phenotypes and that only cells inside those relative high-density areas (peaks) are of importance. However, cells that are in between two high-density clusters (valleys) may also have potential biological significance [76]. Another limitation of clustering based methods is that batch effects from merging of multiple samples (which is a widely used strategy [62, 75]) can be problematic. Batch effects are even less tractable for cross institutes datasets (which are common in clinical trials). In addition, these clustering-based methods require calculation of distance in high-dimensional space, which suffers from the “curse of dimensionality” and may lead to misleading results [77]. As a result, some groups have been calling for the return to lower dimensional methods such as gating based on 2D scatterplots [11].

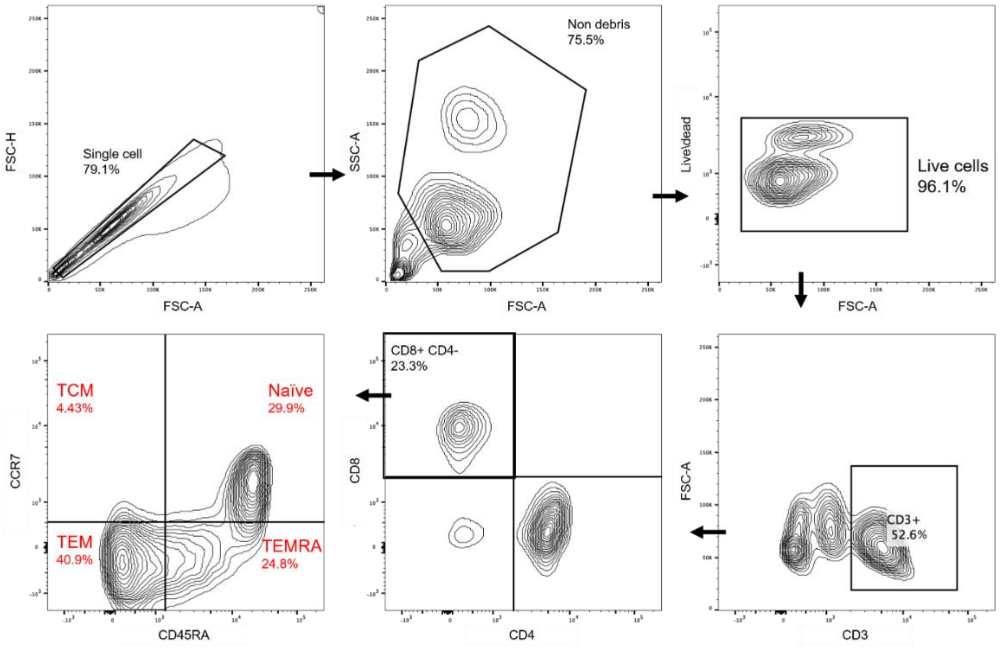


Figure 1.4 Manual gating is a process of sequentially drawing a region of interest (ROI) manually with the hierarchy of marker pairs pre-defined by domain knowledge/ experience. The percentage of cells remaining

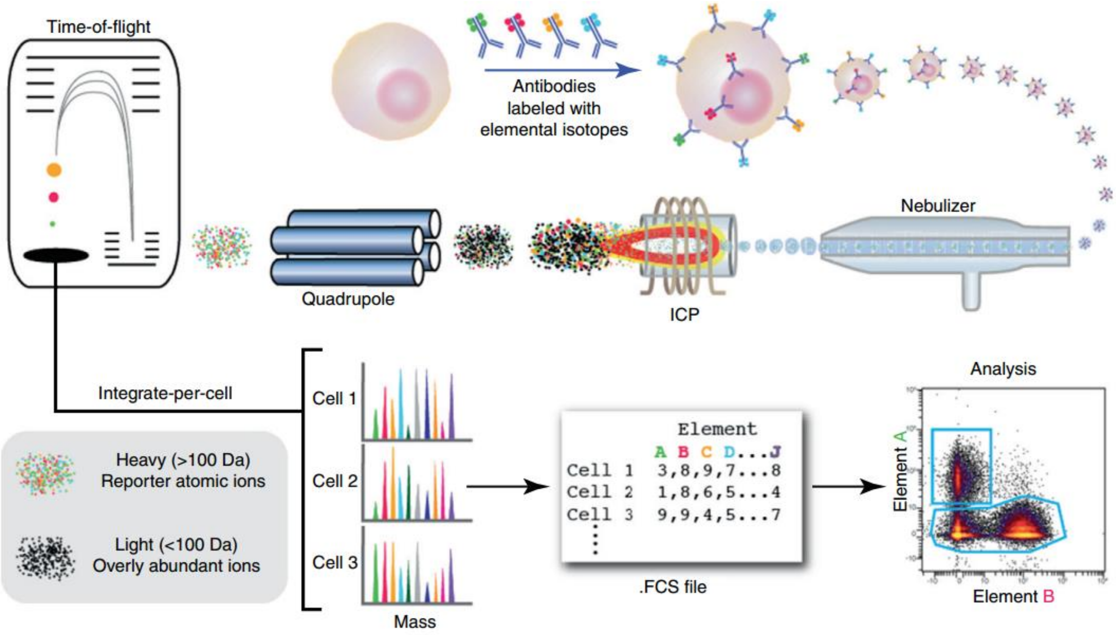
for further analysis decreases in each step of manual gating. Cells outside the ROI are discarded from downstream analysis, and the information they contain is lost. In addition, manual gating is a subjective process with low reproducibility, as ROIs are drawn manually. With the increasing number of measurements, manual gating is also becoming too labor intensive for large studies. However, due to its intuitive nature, manual gating is still the most widely used and understood method for analyzing flow cytometry data. Incorporating the increasing number of measurements, avoiding curse of dimensionality, while at the same time connecting the large volume of information biologists obtain from 2D scatterplots (manual gating) is the goal of Chapter 3.

1.1.3 Mass cytometry / Cytometry by time of flight (CyTOF)

Mass cytometry is a relatively new technique designed to overcome the bottleneck faced by flow cytometry to further expand the number of simultaneous measurements to about 100 in theory [78]. It combines two well-established techniques: mass spectrometry and flow cytometry. Currently, the most widely used mass cytometer can measure more than 40 cellular markers at the speed of 1000 cells/s [79]. Although “low throughput” compared to flow cytometry, the promise to measure more than 50 parameters is “yet unparalleled” [1] and is starting to get more and more attention and usage in both basic science research and clinical research.

1.1.3.1 Experimental Procedure

Mass cytometry uses rare earth metal isotopes instead of fluorescent stains as tags for cellular markers and uses mass spectrometry to detect signals instead of photodetectors. These rare earth metal elements do not naturally exist in cells hence are capable of being used as tags. In addition, the up to 100 distinguishable isotopes ensure an equal number of distinct measurements for each cell [6]. Fig 1.5 shows the experimental workflow of mass cytometry. Single cells are acquired and incubated with isotope tagged antibodies against proteins of interest. Cells are then nebulized into droplets as they are directed into the mass cytometer. Ions in the droplets are liberated through an inductively coupled argon plasma (ICP). Then the ion cloud is filtered according to their mass, and light elements that are abundant in cells are discarded where the heavier elements used as tags remain to be measured by a time-of-flight mass spectrometry [79]. Data collected in mass cytometry experiment is also stored in flow cytometry standard (FCS) format, and hence can be analyzed like flow cytometry data [80].


Time-of-flight
Antibodies labeled with elemental isotopes
Nebulizer
Quadrupole
ICP
Integrate-per-cell
Heavy (>100 Da) Reporter atomic ions
Light (<100 Da) Overly abundant ions
Cell 1
Cell 2
Cell 3
Mass
.FCS file
Analysis
Element A
Element B

```
        Element
        A B C D...J
Cell 1  3,8,9,7...8
Cell 2  1,8,6,5...4
Cell 3  9,9,4,5...7
  ⋮
```

TRENDS in Immunology

Figure 1.5 Experimental workflow for mass cytometry (adapted from [81]). By using rare earth metal elements that are usually not present in cells as indicators, mass cytometry further increases the number of measurements to about 100. With such a large number of measurements, redundant and irrelevant measurements become a real problem for analyzing mass cytometry data. It has been shown that the intrinsic dimension of mass cytometry data is much smaller than the number of total measurements. However, feature selection has been difficult for mass cytometry data since the value of a single measurement is not meaningful. In Chapter 4, I address the problem of feature selection in mass cytometry data by defining features as point patterns formed by two single measurements.

1.1.3.2 Data processing/cleaning

Since mass cytometry collects signal via mass spectrometry instead of fluorescent light intensity, there is no overlap of signals and hence no need for compensation. In addition, there is no light reflection and auto-fluorescence to take into consideration, so the background noise level is lower in mass cytometry data comparing to flow cytometry. However, due to the wide range of surface marker expression levels, a transformation step such as logicle transformation is still required to visualize and analyze mass cytometry data.

1.1.3.3 Data analysis

Mass cytometry shares the data storage format with flow cytometry, so the methods developed for analyzing flow cytometry data could be directly applied to mass cytometry data. However, due to the high dimensionality of mass cytometry data, analysis methods specifically designed for it are in need. Over the years, great progress has been made in developing methods that are customized for mass cytometry dataset [5]. One of the first methods specifically aimed to analyze mass cytometry dataset is SPADE [75]. SPADE utilizes an agglomerative clustering method to group cells into different subtypes. A minimum spanning tree (MST) based algorithm is then used for data visualization (Fig 1.6). SPADE has been successfully applied to several mass cytometry datasets [82-84]. Another method called Citrus uses hierarchical clustering for cell type discovery and utilizes properties of these cell groups to determine a list of cell types that behave differently between two sample groups [85]. More recently, an algorithm called PhenoGraph has applied kNN (k nearest neighbors) to detect cell types in a given sample [86]. In addition, methods designed for flow cytometry, like FLAME [58], flowSOM [62], etc. have also been successfully applied to mass cytometry data [87, 88]. Most recently, people have managed to apply a high-dimensional data visualization method – t-distribution based stochastic neighborhood embedding (t-SNE) (Fig 1.7) [89] – to help with mass cytometry data analysis, e.g., viSNE [73] and one-SENSE [90]. In summary, these algorithms try to make full use of the large number of measurements in mass cytometry data, and they are successful in revealing the high level of complexity in biological systems. However, these methods ignore the existence of redundant and irrelevant measurements in mass cytometry data, which could result in high level of noise and false signal, and the curse of dimensionality. As a result, feature selection methods are needed as a pre-processing for mass cytometry data. However, feature selection for mass cytometry data is a tricky endeavor, since the expression level of single markers (that are usually meaningful features in other types of data) is not meaningful in mass cytometry. Instead, the meaningful information exists in the combined expression pattern of multiple markers (e.g., 2D scatterplots of two markers). Hence new definition is needed for features in mass cytometry data as well as feature selection method.

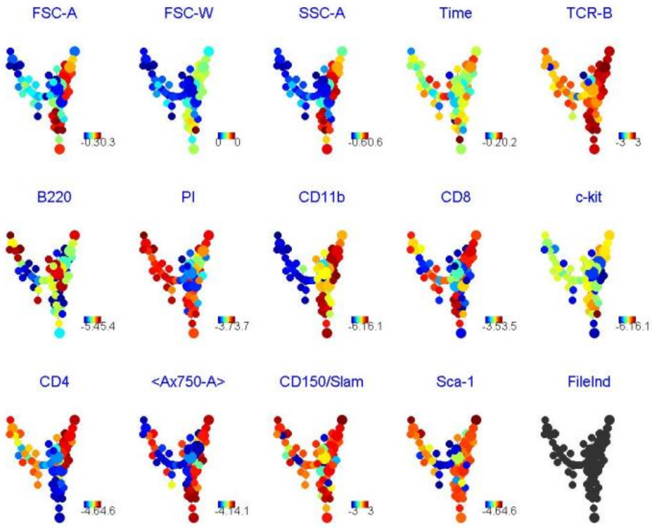


Figure 1.6 An example of results given by SPADE (generated based on dataset and guidelines provided in [75]). SPADE is one of the first methods developed specifically to analyze mass cytometry data. It uses an

agglomerative clustering methods to identify cellular subtypes in mass cytometry data with predetermined estimation of cluster numbers (each circle in the figure is a cluster). Each cluster is then represented by the mean expression level of markers (color coding in the figure shows the mean expression level for each marker). SPADE then uses a minimum spanning tree (MST) to visualize the clustering results as we see above. Though sophisticated, the results generated by SPADE lack clear biological interpretation, and require a good understanding of the underlying algorithms to extract robust information. Also, since SPADE utilizes all measurements simultaneously with equal weights, it suffers from the curse of dimensionality and extra noise buried in redundant and irrelevant measurements.

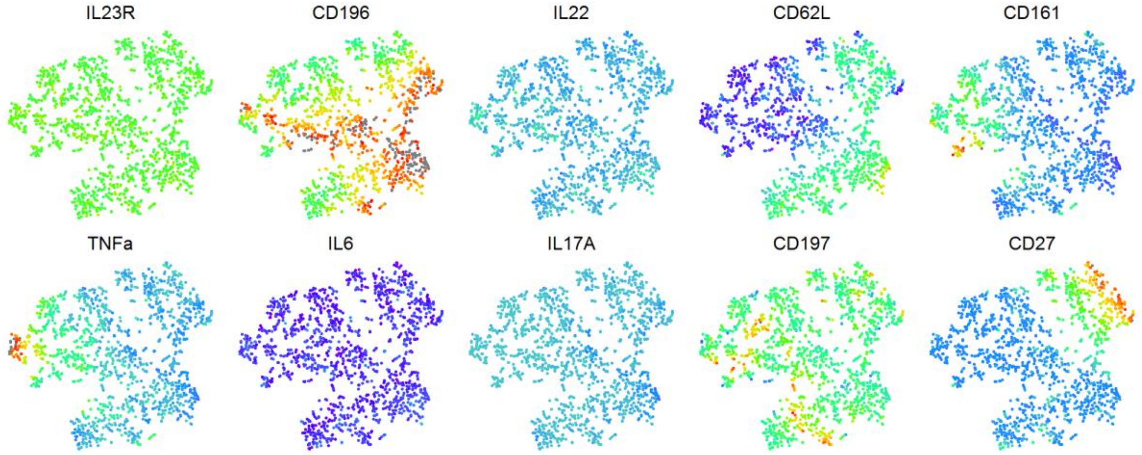


Figure 1.7 An example of t-SNE plots (colored by the expression level of individual markers). t-SNE (t-distribution based stochastic neighborhood embedding) is a dimension reduction algorithm introduced to embed high dimensional data into lower dimensional space while preserving similarities between data points. In other words, high cell density area in original high dimensional space should also be a high cell density area in the reduced dimension presentation. t-SNE is mainly used as a visualization tool for mass cytometry data, however, due to its similarity preserving property, there have been attempts to do manual gating based on dimension reduced t-SNE plots. t-SNE is a great method to visualize high dimensional data, nevertheless since the resulting dimensions are nonlinear combinations of all measurements, interpreting the results can be difficult. In addition, since t-SNE involves random process, the different realization of t-SNE may result in different patterns (with similar features).

Challenges in high-dimensional single cell analysis

In this dissertation, I will focus on two major challenges in high-dimensional single cell analysis: heterogeneity and high dimensionality. These two challenges are inter-correlated, to fully understand heterogeneity one needs more measurements, more measurements (higher dimensionality) leads to higher level of heterogeneity as well as difficulties in analysis.

1.1.4 Heterogeneity

Heterogeneity is ubiquitous in biological systems, and its significance is starting to be recognized. However, the full extent of heterogeneity and its biological implications are still largely unknown. Taking tumor cells as an example, the heterogeneity of tumor cells has been identified more than 40 years ago [91], but how heterogeneity in genetic and epigenetic features of tumor cells affect the progress of cancer is still not fully understood [92].

Simple methods that summarize cell population by its average behavior are not suitable for heterogeneous datasets since they oversimplify the data. The challenge of applying unsupervised methods to heterogeneous data lies in the fact that signals in these data come from a mixture of various cell phenotypes, but the exact number of phenotypes and best set of features to capture these phenotypes are both unknown. On the other hand, not all cell-to-cell variations result in meaningful cell phenotypes. Responses to the slightly different local cellular environments can also lead to variation which can be considered as noise [9]. However, when the difference in local environment is large enough, it could also induce cells to exhibit meaningful biological differences. As live cells constantly interact and respond to their surroundings, it is hard to determine when the noise stops and meaningful heterogeneity starts. Thus when average based methods oversimplify data, sophisticated multi-variate unsupervised methods may result in unnecessary complexity. As a result, a method in between that is able to capture some degree of heterogeneity but does not add more complexity is desired.

1.1.5 High dimensionality

The increasingly large number of features that can be simultaneously measured experimentally leads to two problems: information overload and high-dimensional analysis (e.g., curse of dimensionality). Information overload refers to the phenomenon that not all features measured are biologically important, and there exists redundancy in biologically meaningful measurements [93]. Both unrelated and redundant measurements contribute to noise in data analysis and can result in misleading interpretation. To address this problem, dimension reduction or feature selection methods are required to capture useful information in datasets. The most widely used dimension reduction methods include principal component analysis (PCA), t-distribution based stochastic neighborhood embedding (t-SNE), and independent component analysis. Problems with dimension reduction methods for biological datasets are that: (1) the reduced dimensions are a combination of all original measurements. Thus it is difficult to generate direct biological interpretation. (2) These methods try to identify only one subspace that best captures information content in original datasets, and some subtle but meaningful patterns may be ignored. The problem with simpler feature selection methods is that the selection is usually done one feature at a time, and ignores the correlation between features. On the other hand, analysis of high-dimensional data comes with its own problem – the curse of dimensionality - commonly used distance measurements (L2 distance, for example) loses their meaning in a high dimensional space as in high dimensional space every point is pushed towards its surface and are equal-distant from the center [77]. This further affects the estimation of point density and identification of nearest neighbors [11, 77] which is problematic for density-based algorithms and clustering algorithms. Although lots of efforts have been made on understanding effects of dimensionality on single distribution data, the study of the curse of dimensionality on heterogeneously distributed datasets (as are most biological datasets) has just begun [94].

1.2 Outline of this dissertation

In this dissertation, I will tackle the problem of heterogeneity and high-dimensionality in single-cell data analysis with concrete examples.

Chapter 2 will discuss the first example which concerns cellular heterogeneity in an image-based high throughput screening dataset. This dataset records nuclear images of cultured cells with experimentally induced progeria (a premature ageing disease) in response to different RNAi perturbations. Cellular responses to the same RNAi perturbation can already be various, with some cells resemble healthy control, some cells resemble diseased control, and some cells may resemble neither controls. On the other hand, controls cells have a large cell-to-cell variation as well. The overall goal of this dataset is to identify RNAi hits that help progeria cells restore to a healthy status. To achieve this goal, I propose a notion called “typical cells” which are defined as cells around the center of multi-variate feature distribution. Under the assumption that variation in control cells are mainly caused by unimportant local fluctuations, I extract typical control cells and use them to build a stable classifier between healthy and diseased cells. This classifier is then used to assess the health status of every perturbed cell. Moreover, the importance of each RNAi perturbation is determined by the percentage of healthy-like cells after being perturbed by that RNAi.

Chapter 3 deals with finding subtle differences in high-dimensional flow cytometry data. The dataset I use is a publicly available dataset that tries to clarify differences of immune cell composition in peripheral blood between old and young healthy donors. The most common workflow in dealing with this type of question is – first identify all possible (exhaustedly identified via automated methods) or related (judged by professional experience) cell subtypes in the data, and the compare their frequency in old and young groups one by one with statistical tests. Problems with this workflow are: (1) due to "curse of dimensionality", clustering methods that use all measurements simultaneously can lead to false cell subtypes [11]. (2) There has been hypothesis that cell states belong to a continuum instead of distinct stages. Hence not only high cell density regions but also relatively low-density region that may be transitioning between two cell stages can be biologically meaningful. As a result, methods that only focus on high cell density regions may miss potentially important signal. Here I propose a method called "CytoBinning" that quantifies 2D point patterns. With the help of CytoBinning, I am able to compare point patterns between patients without calculation of cell density in data space, thus avoid focusing only on high-density regions. In addition, by analyzing patterns in 2D at a time, "curse of dimensionality" could be avoided as well. So instead of identifying all possible cell subtypes before comparison, I first identify cell regions/patterns that are different between old and young groups and clarify types of cells in these regions later.

In Chapter 4, I try to tackle feature selection problem in mass cytometry data with a publically available dataset that aims to differentiate tissue-specific immune cell signatures in eight types of human tissues. Although the problem of information overload in mass cytometry data has started to being recognized in mass cytometry data, not many methods have been proposed to perform feature selection for it. This is in part due to the fact that meaningful features in mass cytometry data lie not in single measurements, but the interplay between two or more measurements (point patterns). In addition, the only quantifiable measurements of the interplay between two variables are correlation coefficients and mutual information which oversimplify the pattern and provide no details. Here I use CytoBinning as a quantified, detail compatible method to capture the interplay between two variables that enables direct, the quantified comparison between two point patterns. I then define features in mass cytometry data as point patterns formed by two measurements at a time and am able to identify tissue-specific immune cell signatures in eight types of tissues with only 5 features (6 unique measurements).

Finally, in Chapter 5 I will make short summaries for each the above examples, discuss the potential applications of the methods I developed, and talk about future directions with some preliminary results.

Chapter 2 RefCell: Using typical cells as reference for multi-dimensional analysis of image-based high-throughput screens

2.1 Overview

In this chapter, I try to tackle cell-to-cell variation in image-based data using "typical cells," a method I proposed to capture typical features of single cells within mono-state cell samples. The experiments were performed at NCI/NIH by Dr. Nard Kubben in the laboratory of Dr. Tom Misteli. This chapter is adapted from a manuscript submitting to BMC Bioinformatics.

2.2 Abstract

Image-based high-throughput screening (HTS) reveals a high level of heterogeneity in single cells and multiple cellular states may be observed within a single population. However, in biomedical and clinical practice most image-based HTS analysis is based on average behavior. While complex analysis methods that illustrate and capture this heterogeneity are under development, their reliability and predictive power are still uncertain. Here we introduce RefCell, a multi-dimensional analysis pipeline for image-based HTS that focuses on reproducible capturing of the cell heterogeneity and automated, systematic reduction of the multidimensional HTS information into biomedically actionable figures. Instead of averaging, RefCell selects single cells for which all cellular measurements are typical. RefCell is based on both these "typical cells" and quantitative assessment of the heterogeneous deviations from this typical behavior. We apply this pipeline to the analysis of data from a high-throughput imaging screen of a library of 320 ubiquitin protein targeted siRNAs selected to gain insights into the mechanisms of premature aging.

2.3 Introduction

High-throughput screening (HTS) is a powerful technique routinely used in drug discovery, systematic analysis of cellular functions, and exploration of gene regulation pathways [14, 95-97]. With modern automated microscopes, image-based HTS allows for routine imaging of thousands of cells in multiple fluorescence channels. Due to the volume and complexity of imaging data, building analysis methods has become a big challenge.

During the last decade, powerful new automated image analysis tools [13, 33, 98, 99] that reproducibly parametrize each cell started to emerge, as well as methods for analyzing high-dimensional data specifically applicable to image-based HTS [25, 32, 34, 100-107]. To identify multiple cell subtypes and quantify cellular heterogeneity, machine learning methods such as support vector machines (SVM) [32], hierarchical clustering [33], or clustering with Gaussian mixture models [34] have been introduced. While these methods are very successful in revealing cellular heterogeneity and identifying subpopulations, the "curse of dimensionality" dictates that for high dimensional systems clustering becomes ambiguous. Due to this curse of dimensionality as well as redundant and irrelevant measurements, these sophisticated analysis approaches may produce misleading results when applied to high dimensional data [11]. Furthermore, the outputs of advanced high dimensional analysis methods such as t-SNE or SPADE are not yet standardized, and biomedical and clinical researchers have little experience in how to derive actionable insights from the output graphs.

Due to these challenges, recent publications are suggesting that conventional average-based methods are sufficient for analyzing data collected from cell populations [9]. However, one major disadvantage of average-based analysis is that averaging is generally done for each dimension separately. Thus any mutual information between multiple measurements for each cell are lost in the analysis. In addition, averaging tends to oversimplify the data by smoothing out potentially important cellular variations.

Here we introduce a new method that incorporates multiple measurements simultaneously and captures similarities of cells in a single state population. Our flexible pipeline is focused on analysis of image-based HTS experiments of cellular phenotypes. Our approach captures the typical features of a single state cell population with single cell resolution. This is achieved by the introducing of "typical cells".

We introduce our approach in the context of an RNAi screen to identify cellular factors involved in the premature aging disease progeria. The starting point of the analysis is a set of single-cell metrics obtained through standard image-processing tools (e.g. [25, 108]). The main output of the analysis is identification of the most significant morphological features that together provide a holistic view of the disease phenotype, and a list of significant siRNA perturbations that partially rescue the disease phenotype, which we call "hits". We compare our pipeline to one of the more complex methods for characterizing heterogenous cellular response [34] and found that our pipeline yields similar hits, yet is simpler, faster, and yields output graphs directly interpretable by biomedical researchers.

2.4 Results

We demonstrate our pipeline using datasets from an image-based high-throughput siRNA screen designed to investigate cellular factors that contribute to the disease mechanism in the premature aging disorder Hutchinson-Gilford progeria syndrome (HGPS), or progeria [109], a rare, fatal disease which affects one in 4 to 8 million live births [110]. HGPS is caused by a point mutation in the *LMNA* gene encoding the nuclear structural proteins lamin A and C [111]. The HGPS mutation creates an alternative splice donor site that results in a shorter mRNA which is later translated into the progerin protein – a mutant isoform of the wild-type lamin A protein [110, 111]. HGPS is thought to be relevant to normal physiological aging as well [112-117] since low levels of the progerin protein have been found in blood vessels, skin and skin fibroblasts of normally aged individuals [115]. The progerin protein is thought to associate with the nuclear membrane and cause membrane bulging [118]. In addition to nuclear shape abnormalities and progerin expression, two additional features that have been associated with progeria are the accumulation of DNA damage inside the nucleus [119], as well as reduced and mislocalized expression of lamin B1, another lamin that functions together with lamin A [114].

These cellular hallmarks of progeria are evident at the single-cell level (Fig 2.1a; Fig A1). Typical nuclei from healthy skin fibroblasts with no progerin expression exhibit round nuclear shape, homogeneous lamin B1 expression along the nuclear boundary, and little evidence of DNA damage (Fig A1, top). In contrast, typical nuclei from HGPS patient skin fibroblasts show aberrant nuclear shape, reduced lamin B levels, and increased DNA damage (Fig A1, bottom). For a controlled RNAi screening experiment, a previously described hTERT immortalized skin fibroblast cell line was used in which GFP-progerin expression can be induced by exposure to doxycycline, causing the various defects observed in HGPS patient fibroblasts [120]. RNAi screening controls consisted of fibroblasts in which GFP-progerin expression was induced by doxycycline treatment, in the presence of 1) a non-targeting control siRNA, which allowed for full expression of GFP-progerin and formation of a progeria-like cellular phenotype in most cells, and from here on will be referred to as the GFP-progerin expressing control, or 2) a GFP-targeting siRNA, which eliminated GFP-progerin, restored a healthy-like phenotype, and from here on will be referred to as the GFP-progerin repressed control. Progerin-induced cells were plated in 384-well plates and screened against a library of 320 ubiquitin family targeted siRNAs. In addition, 12 GFP-progerin expressing controls and 12 GFP-progerin repressed controls were prepared on each imaging plate which enables estimation of control variability. Four fluorescent channels were analyzed (DAPI to visualize DNA, far red: the nuclear architectural protein lamin B1, green: progerin, red: γH2AX as a marker of DNA damage). Images were taken at 6 different locations in each well, and each plate was imaged 4 times under the same conditions; the whole imaging procedure was applied to 4 replicate plates with identical setup (see Methods). Details of the screening process are reported in [120].

2.4.1 Definition of stable classification boundaries based on typical cells

Single cell heterogeneity is prevalent in our screen (Fig 2.1). While typical progerin expressing cells exhibit reduced and inhomogeneous lamin B1 expression, pronounced DNA damage, high expression of progerin, and a blebbed cell shape, some cells in this population look like a typical healthy cell, with normal levels of homogeneously distributed lamin B1, little or no DNA damage, little to no expression of progerin, and round nuclear shape (Fig 2.1). Conversely, the cellular population of GFP-progerin repressed controls consists mostly of healthy-looking cells. However, a small fraction of cells in this population display features characteristic of progeria (Fig 2.1a). This heterogeneity is a well-established feature of HGPS patient cells [114].

Quantification of single cell features shows the distribution of the mean intensity for all nuclei (progerin channel), the distribution of standard deviations of curvature (Lamin B1 channel), the distribution of fluorescence intensities found along the nuclear boundary (boundary intensities; Lamin B1 channel), and the standard deviation of intensities inside nucleus (γH2AX channel) (Fig 2.1b). These metrics were extracted via automated image analysis tools (see Methods) from all images in all control samples. For each of the four channels imaged, we show the metric that best separates GFP-progerin expressing controls (red) from GFP-progerin repressed controls (green). Except for the intensity of progerin, distributions overlap significantly, highlighting substantial heterogeneity among nuclei within each control group. The heterogeneity is largest for γH2AX, followed by nuclear shape and lamin B1.

Despite heterogenous cellular expression, the average behavior of GFP-progerin expressing and repressed control cells are significantly different. Since the goal of this screen (as many other screens for identifying potential drugs) is to identify important perturbations that reverse the states of diseased cells to healthy-like, we focus on similarities of cells in each type of controls.

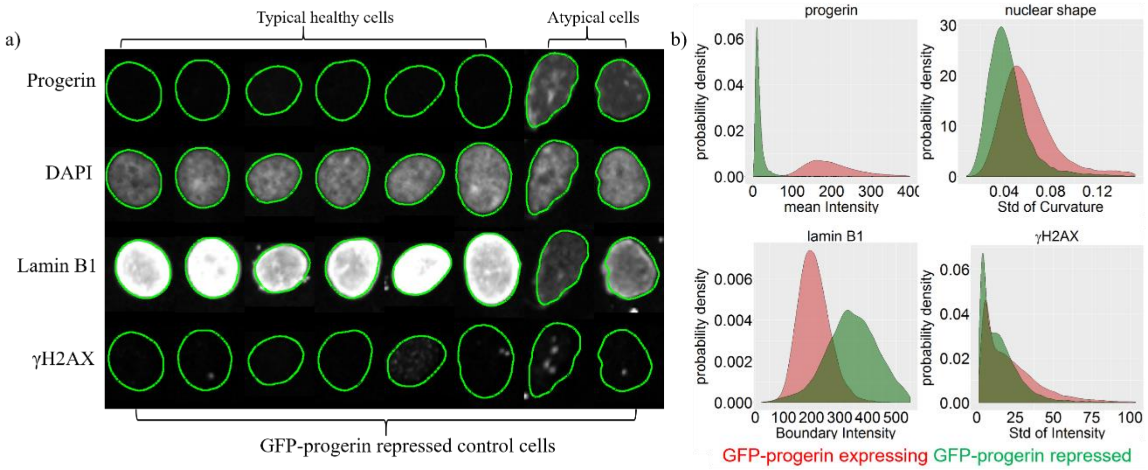


Figure 2.1 Single cell heterogeneity leads to overlapping cell populations. a) Each row corresponds to one fluorescent marker; columns show different nuclei selected from GFP-progerin repressed controls. Nuclear shapes (green contours) were extracted from the DAPI channel and mapped onto the other channels. Typical

healthy cells (first six columns), exhibiting normal lamin B1 expression, little DNA damage, no expression of progerin, and round nuclear shape, as expected for GFP-progerin repressed controls. Atypical cells (two rightmost columns) exhibit characteristics of progeria, namely reduced lamin B1 expression, increased DNA damage in the γH2AX channel, expression of progerin, and blebbed nuclear shape. b) Distribution of the metric that best separates the two types of controls in each channel, based on all cells in the control samples (green: GFP-progerin repressed cells, red: GFP-progerin expressing cells). Note that the contours obtained from the DAPI channel appear slightly smaller and misaligned with the images obtained in the lamin B1 channel (see Fig S2 for the analysis of cross-channel discrepancies). The scale bar is 5 µm.

Classification of individual cells based on such overlapping distributions is challenging, as indicated by the fact that the analysis of multiple sets of 300 randomly selected cells of each of the two reference types via a Support Vector Machine (SVM) approach (see Methods) does not result in a stable classification boundary (Fig 2.2). To illustrate this limitation, we use 200 bootstrap samplings to identify a classification boundary using all metric dimensions simultaneously. We then extract the variability of the classification boundary in each channel (Fig 2.2b). We observe that classification boundaries rotated on average by more than 10 degrees between trials in the progerin channel, and by somewhat smaller amounts in the other channels.

Note that the angle of the classification boundary determines the relative weight of the two metrics shown in the scatter plot: for example, a vertical classification boundary indicates that the metric plotted along the vertical axis is not important for classification. Thus uncertainty about the orientation of the classification boundary implies uncertainty about the relative weight of the metrics in distinguishing both controls. To provide a reliable weighting of metrics and to find reproducible classification boundaries, we use typical cells, defined as cells close to the center of distribution of given cell population in a given channel (see Methods). Typical cells lead to stable classification boundaries with variations of less than 5 degrees in all channels (Fig 2.2b).

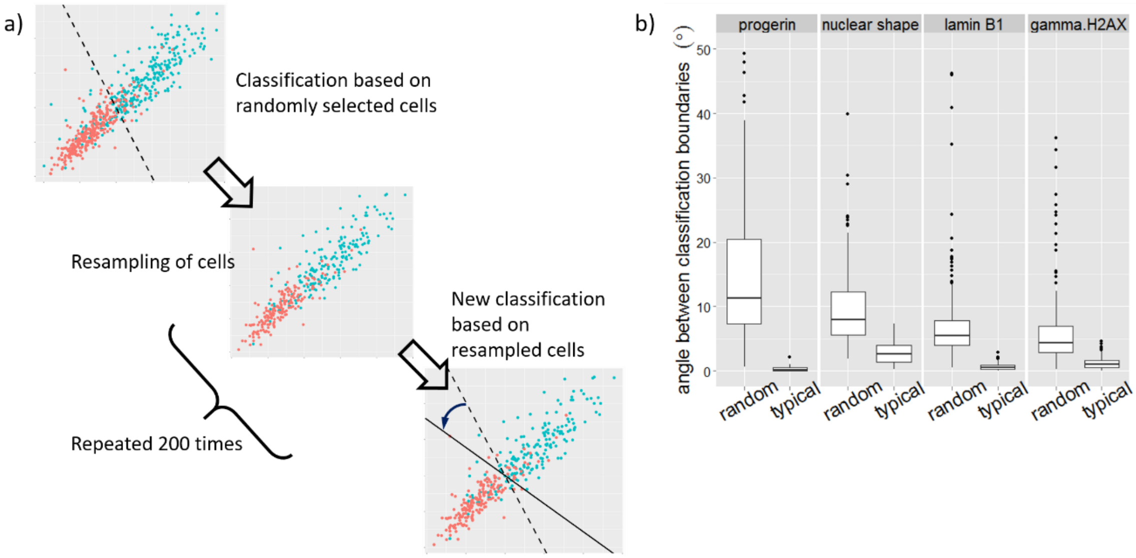


Figure 2.2 “Typical” cells yield robust metrics weighting and stable classification. a) A cartoon showing 300 randomly selected cells for each of the two control populations and a putative classification boundary. The variability in angle for 200 repeats is shown in (b). The range of angles is substantially smaller when “typical” cells are used.

2.4.2 Stable classification boundary enables identification of potential siRNA hits based on the fraction of healthy-like cells

Once a stable classification boundary is drawn based on typical healthy-like (GFP-progerin repressed control) and progeria-like (GFP-progerin expressed control) samples, all cells in all samples can be analyzed using the classification boundary. Specifically, we measured the percentage of healthy-like cells in every sample (Fig 2.3). We define significant siRNA perturbations, or “hits”, based on the ability of the siRNA perturbation to significantly increase the percentage of healthy-like cells (see Methods).

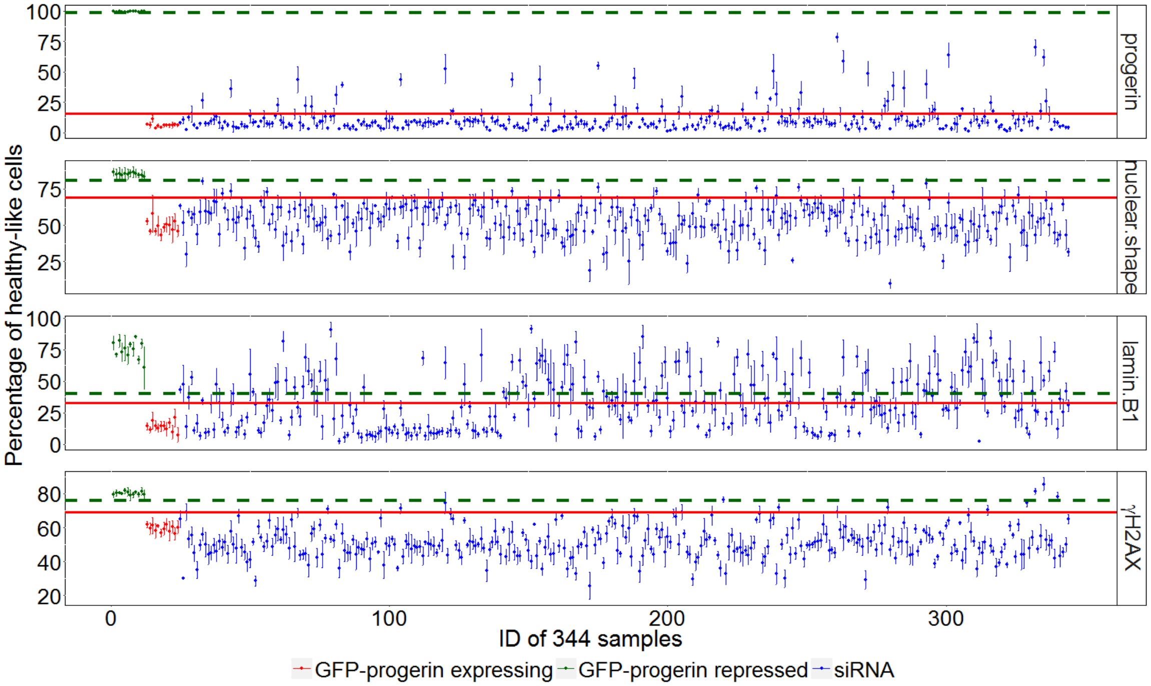


Figure 2.3 Identifying hits from the percentage of cells classified as healthy-like. A visual representation of the entire screen (320 siRNA samples, 12 GFP-progerin repressed control samples, and 12 GFP-progerin expressed control samples). Each dot represents a sample (green: GFP-progerin repressed control, red: GFP-progerin expressing control, blue: siRNA samples), with the vertical axis showing the average percentage and the error bar showing standard deviation of healthy-like cells computed from the 4 independent replicates. False positive rate (FPR) for each siRNA is estimated from this standard deviation. The red horizontal line marks the upper boundary for GFP-progerin expressing control samples used to identify hits (5 standard deviations from the mean of all GFP-progerin expressing controls). Only siRNAs above this line with FPR < 0.05 are considered as hits. The green dashed horizontal line marks the lower boundary for GFP-progerin repressed control samples (5 standard deviations from the mean of all GFP-progerin repressed controls).

In all channels, GFP-progerin expressing and repressed controls are well separated, with the healthy-like phenotype boundary (green dashed line in Fig 2.3) above the hit selection threshold (red solid line in Fig 2.3). The separation between GPF-progerin expressing and repressed controls are the largest in the progerin channel, as expected since GFP-progerin repressed controls are derived from GFP-progerin expressing controls via GFP siRNA modulation. According to our criteria for the selection of siRNA hits (see Methods), the lamin B1 has the largest number of hits (75) and followed by progerin (31), nuclear shape (8), and γH2AX (5) (see details in A7).

The fraction of healthy-like cells in each well in the screen constitutes a metric not yet widely used in screen analysis. This metric highlights the ability of the siRNA to significantly alter some of the cells, but not all, whereas the more traditional metrics – which were also used in the original analysis of this dataset in Ref. [120] – emphasize shifts in the overall behavior. To compare the two metrics, we determine the Z-scores of the shifts in average properties (Fig 2.4a). Both types of Z-scores are determined based on GFP-progerin expressing control samples. For the traditional metric, the threshold is held at Z-score of 2, while our threshold is at Z-score of 5 (by Chebyshev's inequality the probability that the hit is spurious is less than 0.04). Note that if we increase the Z-score threshold for traditional metrics to 5, there will be no hits identified. These two thresholds (gray lines) separate each panel of Figure 2.4a into four quadrants: perturbations identified as hits by both methods (upper right), hits identified only by traditional metrics (lower right), hits identified only by the fraction of healthy-like cells (upper left), and perturbations not identified as hits by either method (lower left). The bottom right quadrant is empty except for two siRNAs in the γH2AX channel, suggesting that our method captured nearly all hits determined by the traditional metric. On the other hand, points in the top left quadrant represent siRNA hits identified only by our approach, suggesting that our metric is more sensitive in the sense of identifying additional possible hits.

In addition, we benchmarked our method against one of the existing multi-dimensional analysis approaches that is also based on the difference in cell type fractions [34]. The method in Ref [34] is based on more complex clustering of all cells into multiple cell types (Fig 2.4b). Using the method of Ref [34], we first identified multiple clusters (9 clusters in progerin and γH2AX channels, and 8 clusters in lamin B1 channel) in 10,000 combined controls cells (5,000 for each control type). We then calculated the profile of cell distribution in each cluster for all siRNA samples and compared with GFP-progerin repressed controls (healthy-like). Since the original workflows of Ref [34] did not include hits selection, we adapted the workflow of Ref [34] and introduce the inverse distance between each siRNA sample and GFP-progerin repressed controls as the metric for hits selection. Figure 2.4 shows a strong correlation between the metric derived from this benchmarking test (horizontal axis) and the RefCell analysis pipeline (vertical axis) (see details in Appendix A.8).

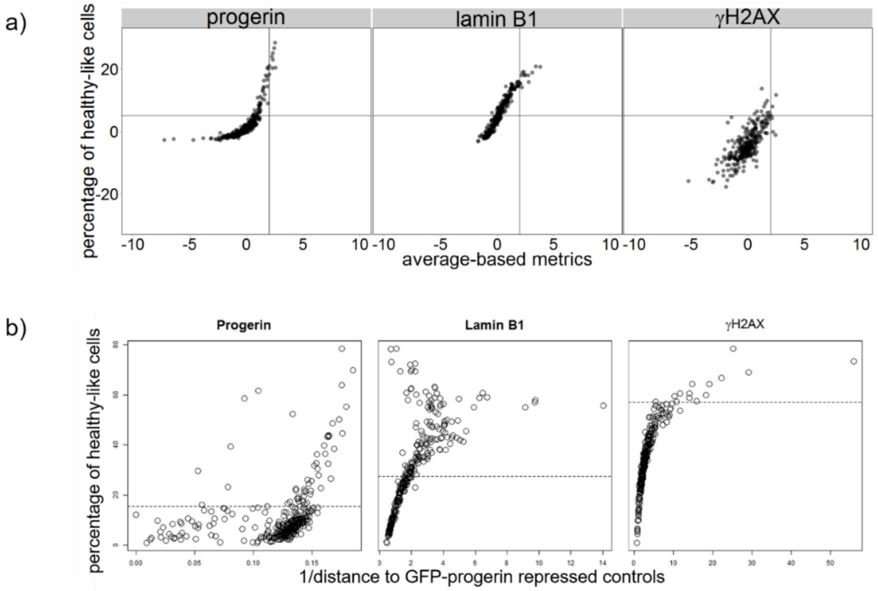


Figure 2.4 Comparing the percentage of healthy-like cells with traditional average-based metrics and another multi-dimensional analysis approach. a) Each panel depicts one channel (nuclear shape (DAPI channel) is not taken into consideration in Ref. [120], hence it is not included here). Every point represents a siRNA sample; the value shows Z-scores calculated based on the distance from the mean of all GFP-progerin expressing control samples using the traditional average-based metric calculated by directly averaging

intensity measurements for all cells in a sample (x-axis) and our metric (y-axis). Gray lines indicate hit thresholds for the corresponding metrics. Note that our metric identifies every hit found by the traditional method (except for the two hits in the γH2AX channel). In addition, our metric selects additional potential hits (siRNAs in the upper left corner) missed by the traditional metric. b) Similar as in a) each panel shows one channel in the screen. Each circle depicts a siRNA sample. The horizontal axis shows inverse of the distance to GFP-progerin repressed (healthy-like) controls, the larger this value, the similar the siRNA to GFP-progerin repressed controls. The vertical axis shows the percentage of healthy-like cells, and the dashed lines are thresholds for hits in respective channels.

2.4.3 Classification boundary and metric weighting obtained via typical cells is useful for characterization of all perturbations

As explained above, we assess the phenotype for each perturbation in our high-throughput screen relative to two types of controls. Thus, the weighting of metrics given by the SVM classification boundary is based on both control phenotypes (Fig 2.2). In Figure 2.3, we had focused on subsets of cells that cross the classification boundary, i.e., that exhibit a shift in property perpendicular to the classification boundary.

In our next step, we characterize shifts of the phenotype both perpendicular and parallel to the SVM classification boundary (Fig 2.5a). We find that most perturbations shift cell properties perpendicular to the classification boundary. This indicates that the imaging metrics which are most important to distinguish typical cells in the two control phenotypes are also the imaging metrics that change most for the siRNA perturbations. However, when the classification metrics are computed from randomly selected cells, - the blue points in Figure 2.5b – we observe shifts both parallel and perpendicular to the classification boundary. (Fig 2.5b). One notable exception is the progerin channel in which the two control cases are very well separated (Fig 2.1b). The analysis above indicates that the 320 siRNA screen focuses on perturbations that primarily affect the progeria phenotype.

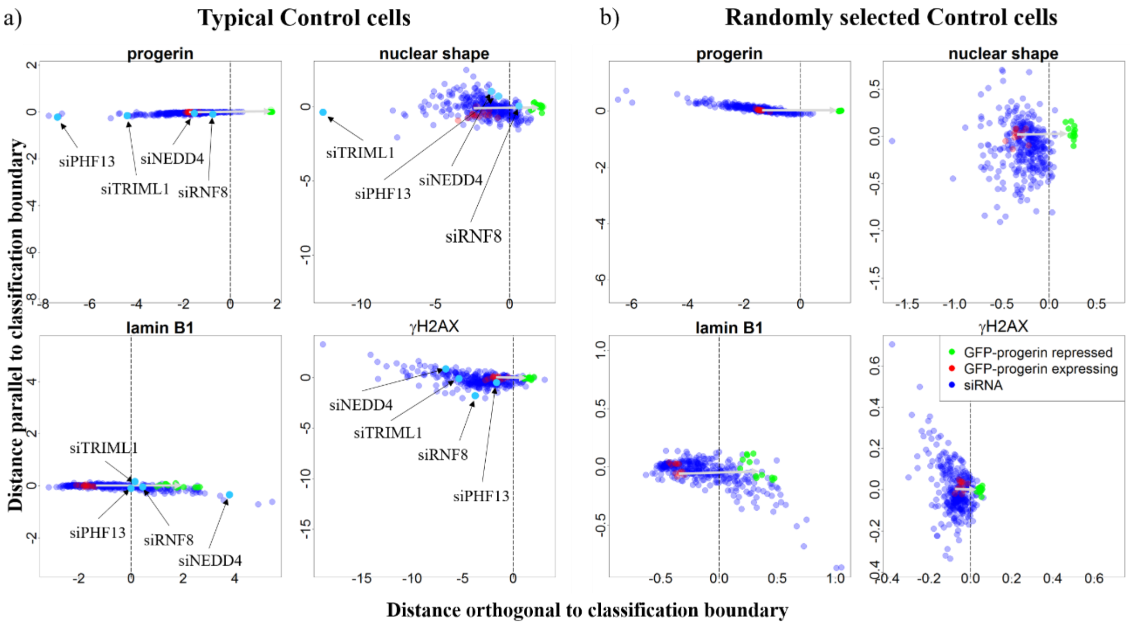


Figure 2.5 The shift of mean cell properties by siRNA perturbations for classification boundaries computed from (a) typical cells and (b) randomly selected cells. Each green and red point represents the mean of all cells in one GFP-progerin repressed (healthy-like) or GFP-progerin expressing (progeria-like) control sample, respectively. There are 12 samples for each control type. Each blue point represents the mean of all cells for one siRNA perturbation. The classification boundary is shown as a vertical dotted black line. Four siRNA samples that deviate significantly from both controls in each of the four channels are labeled (siPHF13 for progerin; siNEDD4 for lamin B1; siTRIML1 for DAPI (nuclear shape), and siRNF8 for γH2AX). Note that the range of the x-axis is the same as the range of the y-axis in all panels. a) Most points are preferentially shifted perpendicular to the classification boundary. Variation parallel to the classification boundary is small compared to the variation perpendicular to it. b) siRNA perturbations are shifted both parallel and perpendicular to the classification boundary when the classification boundary is computed from randomly selected cells.

Figure 2.5a also identifies siRNA perturbations that yield unusual changes in phenotype. Four examples of such siRNAs are highlighted here, one for each channel: siPHF13 for the progerin channel, siNEDD4 for the lamin B1 channel, siTRIML1 for the DAPI channel, and siRNF8 for the γH2AX channel. From each of these siRNA samples, four typical cells (picked using the same method as typical control cells; see Methods) are shown below in

Figure 2.6 (a, b, d, and e). For comparison, four typical cells in both progeria-like and healthy-like controls are also selected (Fig 2.6c and f). siPHF13 treated cells (Fig 2.6a) express even higher levels of progerin than cells in progeria-like controls and progerin aggregates in the nucleus. Upon examining lamin B1 levels expressed by cells treated with siNEDD4 (Fig 2.6b), we find that lamin B1 no longer localizes only to the nuclear boundary, but spreads throughout the nucleus in an inhomogeneous way. In addition, in this case, lamin B1 expression co-localizes with progerin expression. siTRIML1 is an outlier in both the progerin and nuclear shape channel, with overexpression of progerin similar to that observed in cells treated with siPHF13. Furthermore, cells treated with siTRIML1 have nuclear shapes that are even less regular than progeria controls'. Finally, for cells treated with siRNF8 DNA damage is more substantial but also more localized (isolated bright dots in the γH2AX channel) than in progeria-like controls. These results suggest that a classification boundary built from typical cells in controls is valuable to analyze the full perturbation screen and that outliers identified in this classification point to perturbations that yield unusual properties.

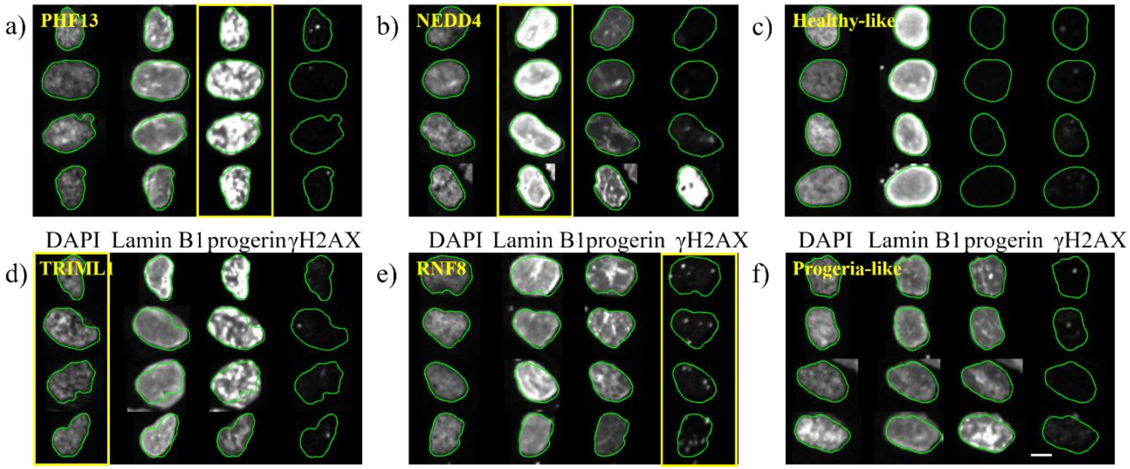

a)
PHF13
b)
NEDD4
c)
Healthy-like
DAPI Lamin B1 progerin γH2AX
DAPI Lamin B1 progerin γH2AX
DAPI Lamin B1 progerin γH2AX
d)
TRIML1
e)
RNF8
f)
Progeria-like

Figure 2.6 Typical cells in siRNA perturbations identified as different from both controls. a) siPHF13 is an outlier in the progerin channel: cells treated with siPHF13 express more progerin than the progeria-like

control cells (f), and the expressed progerin appears to be distributed differently from the progeria control. b) siNEDD4 is an outlier in the lamin B1 channel; cells treated by siNEDD4 express more lamin B1 than the healthy-like control cells (c), and the expression is less homogenous. In addition, the expression of lamin B1 is spatially co-localized with the expression of progerin in siNEDD4-treated cells. d) siTRIML1 is an outlier in both DAPI (nuclear shape) and progerin channels. Cells treated by siTRIML1 tend to have elongated nuclei compared to the healthy-like and the progeria-like controls. Also, clusters and increased progerin expression (compared to the progeria-like control (f)) can be observed. e) siRNF8 is an outlier in the γH2AX (DNA damage) channel. Note that the contours obtained from the DAPI channel appear slightly smaller and misaligned with the images obtained in the lamin B1 channel (see Fig A2 for the analysis of cross-channel discrepancies). The scale bar is 5 µm.

2.4.4 Integrating information from multiple channels increases hit detection accuracy

So far we have considered multiple metrics separately for each channel. This means that we may have labeled a cell as healthy-like based on one channel, but progeria-like when it is analyzed in another channel. This approach reflects uncertainty regarding the progeria phenotype at the single cell level: although it is known that progeria is caused by the expression of the lamin A-mutant progerin, it remains unknown how progerin expression changes other features, such as blebbed nuclear envelope, DNA damage accumulation, and mislocalized lamin B1 expression at the single-cell level, and how these different features correlate with one another. For example, in one study progeria and healthy cells were distinguished using only nuclear shape measurements [121], implying that nuclear shape is a dominant criterion in detecting progeria. However, another study found that nuclear shape could change independently from DNA damage accumulation inside the nucleus [119].

Thus, as a final analysis step, we study the relationships among the four features associated with progeria at the single-cell level. RefCell integrates single cell information from multiple channels in two different ways. First, we display the percentage of healthy-like cells for a primary marker vs. the percentage of cells identified as healthy-like according to the other three markers (Fig 2.7). The diameter of the circle represents the fraction of cells identified as healthy-like according to all four markers. As expected, GFP-progerin repressed controls (i.e., healthy-like controls, green circles) show a larger percentage of cells identified as healthy-like for all four markers than any of the 320 perturbations (blue circles). Figure 2.7 shows that the percentage of healthy-like cells according to a given marker is correlated with the percentage identified as healthy-like according to the other three markers are correlated, although the correlation is weak in all channels except progerin.

Second, we integrated image metrics from all channels together and applied our method on combined metrics. We found that the three metrics related to progerin (mean intensity, standard deviation of intensity and boundary intensity) are the most important metrics in separating GFP-progerin expressing and repressed controls, contributing more than 60% in the direction of classification boundary. Lamin B1 is next, contributing about 20%. In addition, we found that 99% siRNA hits identified by combining all channels are also identified by detecting hits separately for each channels; however, the combined analysis allows us to hone in on a subset of 61% of all hits (based on separate analysis of each channel).

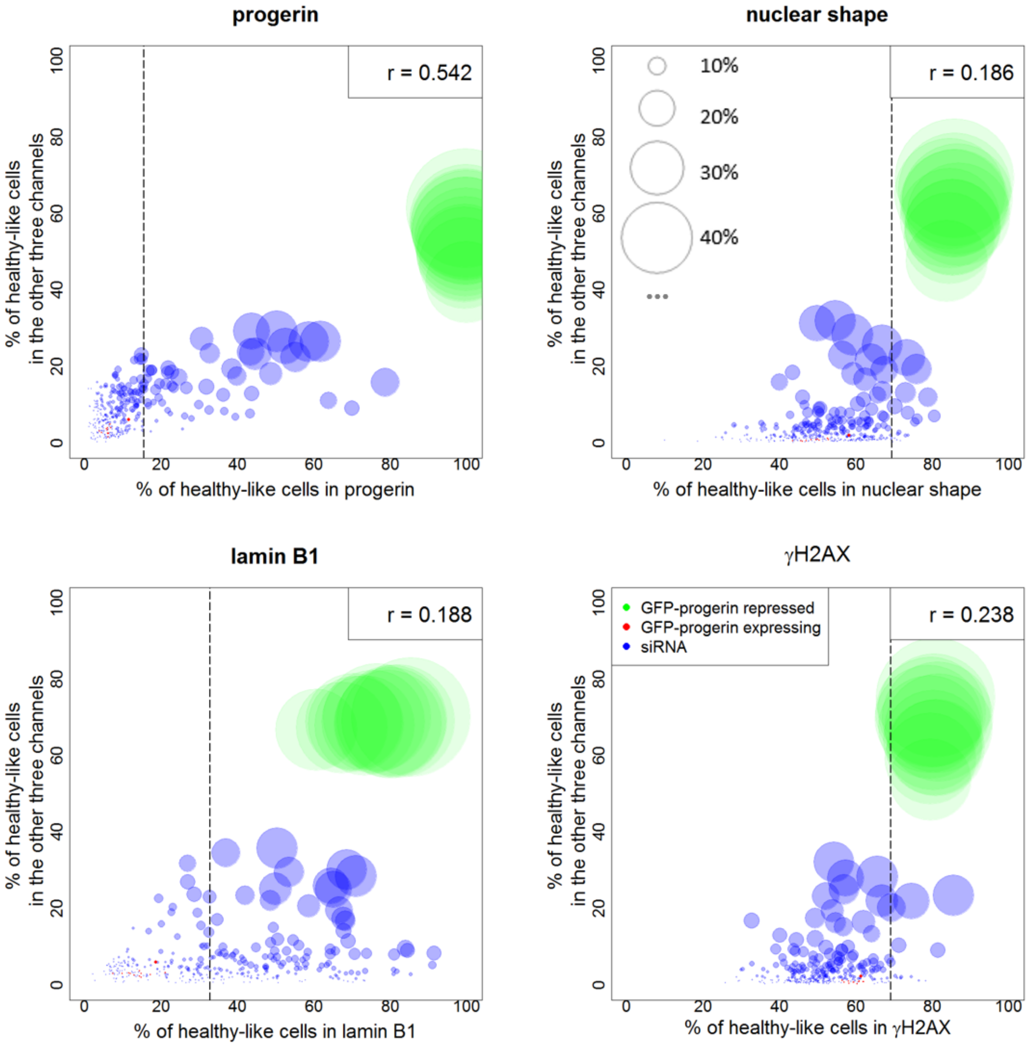


Figure 2.7 Integrating information from all channels: Percentage of healthy-like cells in one channel vs. percentage of cells classified as healthy-like in the other three channels. Each circle stands for a sample (green: GFP-progerin repressed, red: GFP-progerin expressing, blue: siRNA). The size of the circle is proportional to the percentage of cells that are classified as healthy-like in all four channels (scales are shown in top-right panel). The dashed vertical lines are thresholds for hit selection in the corresponding channel. Shown in the upper right corner of each panel is the linear correlation coefficient (note that p < 0.01 after Bonferroni correction in all cases).

2.5 Discussion

One of the major usages of image-based high-throughput screening (HTS) experiments is to identify important RNAi perturbations for pathway identification or drug discovery. A major strength of image based HTS is that measurements of multiple parameters are carried out on each cell, thus promising insights into mutual information and correlations among parameters at the single cell level. However, newly developed analysis methods yield complex and hard to interpret end results, and may actually misrepresent the data due to the curse of dimensionality [11]. In fact, most current applications of image based HTS still rely on simple averaging of each parameter. Here we introduce RefCell, a method that fills the gap between statistically sound average-based methods and statistically challenging high-dimensional methods. The underlying assumptions of RefCell are that the properties of typical cells are useful reference points for the biological or clinical question of interest and that the best approach to identify hits is to measure changes along a straight path (in high dimensions) between the references points.

The first step in RefCell is the selection of two sets of controls and the choice of "typical" cells within these controls. Here we choose typical cells as cells that are average in all aspects of their phenotype, i.e., all the metrics are close to the mean. In our dataset, one control represents cell nuclei of a model for progeria which show several defects, and the other control approximates healthy cell nuclei. Since image-based metrics are heterogeneous, the corresponding distributions of measured values overlap significantly at the single-cell level (Fig 2.1). Selecting typical cells yields distributions that are well separated, enabling stable classification boundaries between healthy-like and progeria-like cells. The classification boundary reveals both the value of each metric that marks this transition and the relative weight of each metric (Fig 2.2).

For the HTS used in this investigation, we find that surprisingly the metrics we identified as important are also the metrics that change most for all perturbations. A graphical representation of this observation is shown in Figure 2.5a, where the two controls (green and red dots) lay out a straight path between a progeria-like phenotype and a healthy-like phenotype. All siRNA perturbations (blue dots in Fig 2.5a) fall along this straight path indicating that the metrics that were identified as important are the ones that are changing most in the 320 siRNA perturbations. On the other hand, if all cells rather than typical cells are used for classification and weighting, classification boundaries are less stable (Fig 2.2), and the 320 siRNA perturbations do not change the highly weighted metrics more than other metrics (the blue dots in Fig 2.5b form a cloud). This indicates that the screen does not involve random perturbations but perturbations targeted specifically to progeria.

With these weights and a stable classification boundary, we were able to quantify the heterogeneity of all cells in all samples. This analysis yields a simple parameter: The fraction of cells identified as healthy-like in each sample. The fraction of normal cells had been identified in other studies as a useful parameter [122]. In RefCell, this parameter is used in multiple steps and is first determined separately for each channel to identify potential "hits" in the siRNA perturbation screen (Fig 2.3).

Our new parameter, the fraction of healthy-like cells represents an ensemble metric for a cell population and correlates with the average-based metrics traditionally used in most HTS. However, the classification boundary location and weight of metrics are tuned to the specific classification challenge by identifying two control populations. RefCell identifies all hits selected by an average-based method, but with higher statistical significance (Z-score). In addition, RefCell identifies additional hits even when the cutoff Z-score for hits is increased to 5 for in RefCell, compared to the Z-score cutoff of 2 for the average-based m.

Furthermore, RefCells focus on the fraction of healthy-like cells means that any perturbation that makes a substantial fraction of cell nuclei appear healthy-like is included as a possible hit, even if the average cell properties do not change. This allows us to include all perturbations that are capable of making at least a subset of cells appear healthy-like, even if the same perturbation is ineffective or detrimental for other cells.

The final step in RefCell focuses on integrating information from multiple imaging channels (Fig 2.7). When considering all siRNA perturbations and all channels simultaneously, our analysis confirms that the progerin level is the most important feature in progeria disease, and that decreasing progerin expression levels is the most efficient way of removing all four principal phenotypes associated with progeria. However, we also note significant variability in how effectively a given perturbation leads to healthy-like phenotypes in each channel. This information helps prioritize hits that have been identified separately in each channel.

In addition, we compared RefCell with a published method that aims to characterize heterogeneity in cells using EM clustering with Gaussian mixture models (GMM) [34]. Since the published method did not provide a metric for hits selection, we used inverse distance to GFP-progerin repressed controls. This distance is calculated using symmetrized KL divergence as used in [34]. The higher the inversed distance, the more important the perturbation. We show that in both progerin and lamin B1 channel, our metric agrees well with the other method (see Appendix A.8) with Spearman correlation coefficient 0.98 for γH2AX channel and 0.91 for lamin B1 channel (p value << 0.05 in both cases). However, the complex clustering approach does not allow us to integrate information from all channels, since complex clustering cannot be used for the analysis of cell morphology (the dimensionality of metrics is too large for meaningful clustering with “only” thousands of cell images in each sample).

In summary, RefCell represents a simple but useful computational approach for analyzing image-based HTS datasets. RefCell is broadly applicable to single-cell-based high-throughput screens that focus on perturbing cells from one distinct phenotype to another. RefCell uses image processing and machine learning algorithms to identify hits that substantially increase the fraction of cells that regain one of the two reference phenotypes. RefCell can be used to analyze each fluorescent channel separately, and also to integrate the single-cell information from all channels. Applied to a progeria HTS dataset, RefCell reveals a complex interplay among the four standard indicators of proteria (measured in four independent fluorescence channels), revealing that the list of hits depends strongly on the choice of indicator. RefCell analysis further revealed that the screen was not unbiased, but focused on a pathway with known links to the disease.

2.6 Methods

2.6.1 Experimental procedure

hTert immortalized doxycycline GFP-progerin inducible human skin fibroblasts (P1 cells) were generated and induced (96 hr). Reverse siRNA transfections were carried out in quadruplicate in a 384-well format (Perkin Elmer Cell carrier plates) in the presence of doxycycline (1 mg/ml) with pooled siRNA oligos (50nM; 4 siRNAs/target) from the Dharmacon siGENOMESMARTpool siRNA Human Ubiquitin Conjugation subset 1 and 2 libraries. Positive and negative controls consisted of GFP-targeting and non-targeting siRNA (50nM; Ambion, #AM4626, #AM4611G), respectively. Transfected cells were incubated overnight, after which 60 ml of antibiotic and doxycycline (1 mg/ml) containing medium was added, and cells were incubated for another 3 days (37 ^{o}C, 5% CO2). Details of the experiments are reported in [109]. A full list of screened siRNAs can be found in Appendix A Section 9.

2.6.2 Image analysis

While metrics similar to the one used in this study could be obtained with commercial software, we used a custom image analysis method modified from methods in [123]. Details are described in Appendix A Section 3. A list of measurements and short descriptions are shown in Table 1.

Table 2-1 Image measurements used in this study.

	Name of measurement	Description
Nuclear shape	Area	Area of nucleus
	Circularity	Ratio of perimeter to area, normalized so that a circle would have ratio 1
	Eccentricity	Eccentricity of nucleus
	Invaginations	Number of invaginations along the nuclear boundary
	Major Axis Length	Major axis length of the best fit ellipse to the nuclear boundary
	Mean Curvature	Mean curvature along the nuclear boundary
	Mean Negative Curvature	Average of only negative curvatures along the nuclear boundary
	Minor Axis Length	Minor axis length of the best fit ellipse
	Perimeter	Perimeter of nucleus
	Solidity	Percentage of pixels inside the convex hull that are inside the boundary
	Std of Curvature	The standard deviation of curvature
	Tortuosity	Tortuosity of nuclear boundary
Intensity	BP Intensity	Mean intensity of points along the nuclear boundary
	Mean Intensity	Mean intensity inside the nucleus
	Std of Intensity	Standard deviation of intensity inside nucleus

2.6.3 Analysis of the two control groups

Selection of typical control cells. Within each control population, typical cells were defined as a core of n=300 cells closest to the mean based on the L1 (Manhattan) distance, calculated separately for each channel. We pooled all control samples together for typical control cell selection. On average there are about 20,000 cells in each type of controls. Typical progeria-like cells, selected out of the population of GFP-progerin expressed controls, show HGPS characteristic nuclear defects (increased progerin expression, misshapen nuclei, reduced lamin B1 protein levels, and increased DNA damage shown by expression of γH2AX). Typical healthy-like cells, selected from GFP-progerin repressed controls, show no sign of HGPS nuclear defects. This selection procedure was carried out independently for each replicate plate. Additional details are provided in Appendix A Section S4.

Classification using Support Vector Machines (SVM). The sets of typical cells were used to classify healthy- and progeria-like phenotypes via SVM, an efficient and robust supervised machine learning algorithm for classification [124]. Using a linear kernel, SVM finds the optimal linear boundary in instance space (straight line in 2D, planes in higher-dimensional spaces) that separates two classes of instance data points, while maximizing the margin of class separation. We performed SVM using the ksvm() function in kernlab package in R (version 3.1.1). After rescaling all nucleus metrics to zero mean and unit variance, a classification boundary was obtained between typical healthy and typical progeria cells. The distance from each nucleus to the classification boundary, which is a linear combination of all the measurements, can be used as a score to classify the proximity of that cell to each phenotype (healthy- or progeria-like). In order to distinguish between the two sides of the classification boundary, we define positive distances as associated with healthy-like cells, and negative distances with progeria-like cells. The SVM analysis also yields the relative importance of each metric in distinguishing between the two phenotypes as shown in Appendix A Section 5.

2.6.4 Identification of significant perturbations

Determination of the fraction of healthy-like cells. Having obtained a classifier boundary based on typical control cells, we then applied it to all samples (including all control samples and siRNA perturbations samples). For this, we first normalize all cells to be classified using the z-score transformation determined from typical control cells (i.e., subtracting the mean of typical control cells and dividing by their standard deviation). Next, we calculate the distance from each cell to the classification boundary and use the sign of the distance to classify individual cells as either healthy- or progeria-like. Finally, we calculate the percentage of healthy-like cells in each sample. This percentage is obtained separately for each replicate plate. This allows us to report the mean percentage (averaged over all replicate plates) and its estimated uncertainty (resulting from the variance over multiple replicates). The number of cells in each perturbation sample ranges from 500 to 2000. For more details, see Appendix A Section 6.

Identification of siRNAs that generate significant healthy-like perturbations ("hits"). We repeated the screen 4 times (yielding 4 independent replicates), and the analysis described above was done separately for each plate (i.e., given a sample, there are 4 independent estimates for each parameter). To carry out the hit selection process, we first averaged each parameter over the 4 replicates. Then we excluded potentially cytotoxic siRNA samples, by excluding those that contain less than 50% of cells compared to GFP-progerin repressed samples (the number of cells is similar in each sample at the start of the experiment). Next, a siRNA hit was selected based on the following two criteria: 1) the fraction of healthy-like cells is above a threshold (a mean and standard deviation were computed based on the percentage of healthy-like cells in each of the 12 GFP-progerin expressing control samples, the threshold was set to 5 standard deviations higher than the mean); 2) the false positive rate (FPR) based on the variation among the 4 replicates is less than 0.05.

Chapter 3 CytoBinning: immunological insights from multi-dimensional data

3.1 *Overview*

In this chapter, I try to reconcile the conflict between difficulty in interpreting sophisticated high dimensional analysis methods and the impression of the low information content of 2D scatterplots by introducing a binning method I termed “CytoBinning”. The increasingly large number of measurements that can now be made simultaneously using cytometry platforms have created the impression that 2D scatter plots, which used to be the center stage of cytometry data analysis, don’t contain enough information. However, sophisticated methods that fully embrace large numbers of measurements are hampered by the difficulties of interpreting high-dimensional datasets, and this limits their practical utility. CytoBinning fills the gap of complexity between conventional manual analysis and complex automated analysis to extract deep content in scatter plots which can be later cascaded into more complicated clustering or classification algorithms to obtain novel biological insights. The experimental data were obtained from the online repository: https://flowrepository.org/. This chapter is adapted from a manuscript submitted to PLOS Computational Biology.

3.2 *Abstract*

New cytometric techniques continue to push the boundaries of multi-parameter quantitative data acquisition at the single-cell level, particularly in immunology and medicine. Sophisticated analysis methods for such ever higher dimensional datasets are rapidly emerging, with advanced data representations and dimensional reduction approaches. However, these are not yet standardized, and clinical scientists and cell biologists are not yet experienced in their interpretation. More fundamentally their range

of statistical validity is not yet fully established. We, therefore, propose a new method for the automated and unbiased analysis of high-dimensional single-cell datasets that is simple and robust, with the goal of reducing this complex information into a familiar 2D scatter plot representation that is of immediate utility to a range of biomedical and clinical settings. Using publicly available flow cytometry and mass cytometry datasets, we demonstrate that this method (termed CytoBinning), recapitulates the results of traditional manual cytometric analyses and leads to new and testable hypotheses.

3.3 Introduction

Cytometry is a multi-parameter single-cell measurement technique that is widely used in biological and clinical studies [37, 125-129]. One of the main uses of flow cytometry, which has had a major impact across the fields of immunology and medicine, is to differentiate immune cells compositions among cell types or patients. Modern flow cytometers can routinely measure 15-20 cellular markers on millions of cells from dozens of samples in one experiment and can sort cells into subpopulations based on those markers. Recently mass cytometry has expanded the number of markers that can be measured simultaneously to 100, though the technique is destructive to cells and does not allow for sorting. The conventional way of analyzing flow cytometry data uses a gating strategy which requires the manual selection of regions of interest (ROI) on sequential 2D scatterplots. This type of analysis is very labor intensive and inefficient for such large datasets and also suffers from subjectivity in both the sequence of 2D scatterplots and selection of thresholds (ROI) [5, 37, 45, 51, 52, 127]. Therefore, as both the number of cells analyzed and the number of markers quantified for each cell have grown over the past decade, novel, automated and unbiased analysis methods for flow cytometry data are emerging [1].

These novel analysis methods can be divided into two categories based on the problem they address: 1) methods trying to mimic and automatize the process of manual gating [54, 55, 61, 64, 69, 130, 131]; and 2) methods trying to identify cell populations using all markers simultaneously without prior biological knowledge [58, 59, 132, 133]. Some cutting-edge approaches to automating manual gating, such as flowDensity [54], are very successful in re-identifying cell subsets that match with manually gated subsets in an automatic, reproducible way. However, gating (both manual and automatic) relies heavily on prior experience to inform the sequence of markers to gate. Furthermore, in gating, researchers must define the cell phenotypes to look for in advance of their analysis, hence hindering discovery of novel cell types and not tapping into the full potential of the acquired data. Gating methods also only explore a very limited portion of the total data space, though unsupervised methods have been published that enhance the efficiency of data usage, with the potential to reveal otherwise hidden differences between datasets [70]. Most unsupervised methods that allow novel cell type discovery aim to identify regions with high cell density in multi-dimensional space [56, 58-60, 63, 66, 70, 72, 73, 75]. This assumes cells form distinct phenotypes and that only cells inside those relative high-density areas (peaks) are of importance. However, cells that are in between two high-density clusters (valleys) may also have potential biological significance [76]. Another limitation of clustering based methods is that concatenating different samples (which is a widely used strategy [62, 75]) with potential batch effects can be problematic, hence limiting the meaningful combination datasets across institutions (which is very common in clinical trials). In addition, these clustering-based methods require estimation of nearest neighbors in high-dimensional space which suffers from "curse of dimensionality" and may lead to misleading results [77]. As a result, people have been calling for the use of lower dimensional methods such as gating based on 2D scatterplots [11].

In this paper, we present a new method for analyzing cytometry data that utilizes such 2D scatter plots. Instead of gating, we dig deeper into the scatter plots mining the information that is largely bypassed by other methods. This method is useful for the majority of comparative studies that aim to elucidate the difference between two groups of samples. Our method, which we term CytoBinning, identifies the most information-rich 2D scatter plots and extracts biological insights from them. We show that biologically relevant differences can be discovered from the pairs of markers identified with this approach. First, we introduce CytoBinning with a synthetic dataset, and then apply it to two public high-dimensional single-cell datasets, a flow cytometry dataset comparing composition in immune cells between old and young healthy human donors [59], and a mass cytometry dataset analyzing the immune signature of eight types of human tissues [134].

3.4 Results

We synthesized two point-patterns based on the expression of two virtual markers: maker A and marker B. Ten samples were generated for each point-pattern. The first point-pattern, called pattern A, consists of three point-clusters. Two large clusters each contain 5,000 points and a third relatively small cluster contains about 2,000 points. The three clusters are randomly sampled from Gaussian distributions that centered at point (0, 4), (0, -4) and (4, 0) with standard deviation 2, 2, and 1 respectively. The second point-pattern, called pattern B, also consists of three point-clusters. The two large clusters are generated in the same way as point-pattern A. However, the third smaller point-pattern only contains 200 to 500 points, sampled from a Gaussian distribution centered at point (-4, 6) with standard deviation 1 (Fig. B1).

3.4.1 Percentile-based binning is a coarse-grained representation of point patterns

An example of percentile-based binning is shown in Figure 3.1 using one synthetic sample with point-pattern A. Points inside the point-pattern were first binned into 3 levels based on the expression of marker A and B independently, each level containing one third of points. The 3 levels for marker A and B were then combined on a 2D scatter plot to form 9 sub-regions (these sub-regions are called boxes). The percentage of points in each box changes depending on the point-pattern. This binning method has been used as an alternative method to calculate mutual information (MI) in a robust and computationally efficient way [135]. MI is a measure of dependence between two random variables widely used in gene network inference [136] as a general measure of interdependency between genes. In our method, instead of summarizing the binning information into one number (MI), we used percentage of points in each box as a coarse-grained representation of point-patterns to obtain detailed information of point-patterns.

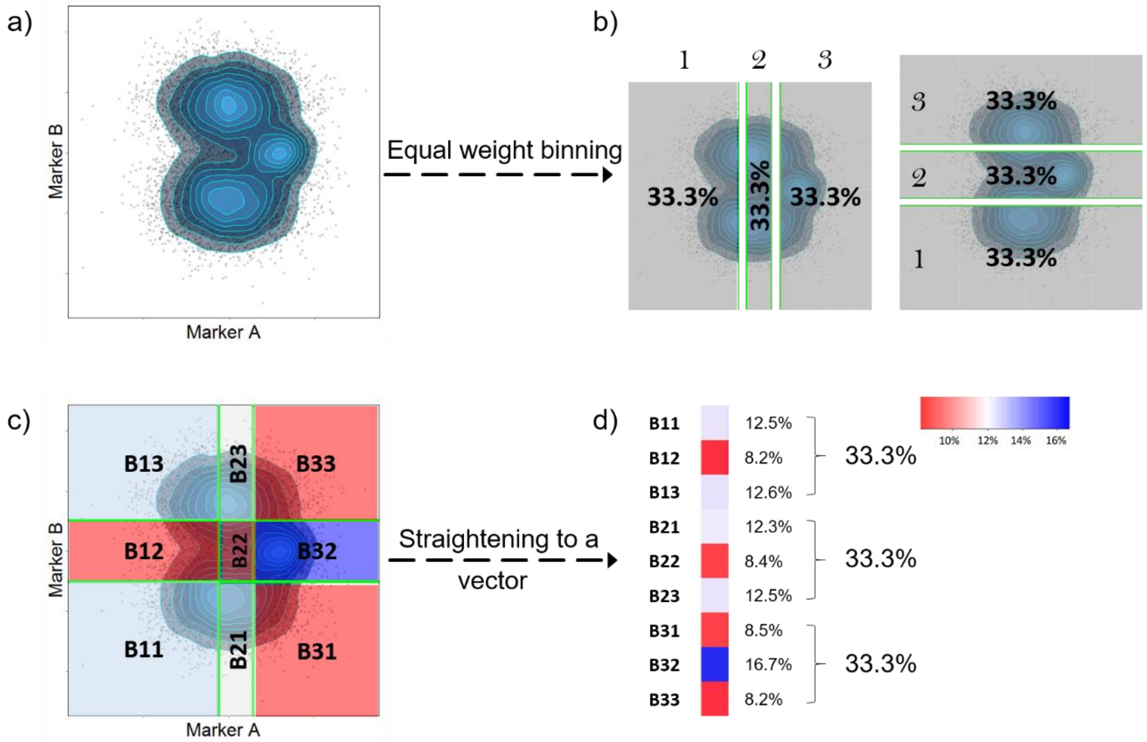


Figure 3.1 An example of percentile-based binning as a representation of 2D point-pattern (number of bins = 3). (a) Synthesized point-pattern formed by expression of marker A and marker B. (b) Points are binned into 3 bins each containing 1/3 (33.3%) of the total points. The bins are labeled numerically based on the expression level of the related marker, with 1 the lowest and 3 the highest (similarly if points are divided into 5 bins then the highest level is 5). This binning is done independently for marker A and marker B based on their expression. (c) Bins obtained in (b) are combined so that 9 sub-regions referred as boxes are formed. The percentage of points inside each box is calculated, and the matrix of percentages is straightened to a vector so that the 2D point-pattern shown in panel (a) is coarse-grained to the vector in (d).

3.4.2 Applying percentile-based binning to multiple samples enables meaningful classification

After we demonstrate how to represent point-patterns with percentile-based binning, next we show that this representation is able to capture real differences in point-patterns. Figure 3.2a shows two examples of the synthetic point-patterns. In total, 10 samples were generated for each point-pattern, and each sample was analyzed using percentile-based binning to generate the row vectors shown in Figure 3.2b. Using 3 bins, our method is able to cluster the two point-patterns into distinct groups, and correctly identifies the most significant difference (Fig. 3.2b & c). Boxplots of cell percentage in each box show that the box with most distinct difference between the two point-patterns is box B32 which contains the third cluster of point-pattern A (lower panel of Fig. 3.2b). This is the most significant difference between these two point-patterns, and it was captured without referring to density distribution of points. Minor differences between these two point-patterns (the small cluster located at the top left corner) were not spotted, since the percentage of points in box B13 is similar in both point patterns (Fig. 3.2b). However, this third cluster in pattern B (~ 2% to 5%), was identified when the number of bins was increased to 6 (Fig. B2). Hence, the depth of analysis depends on the number of bins.

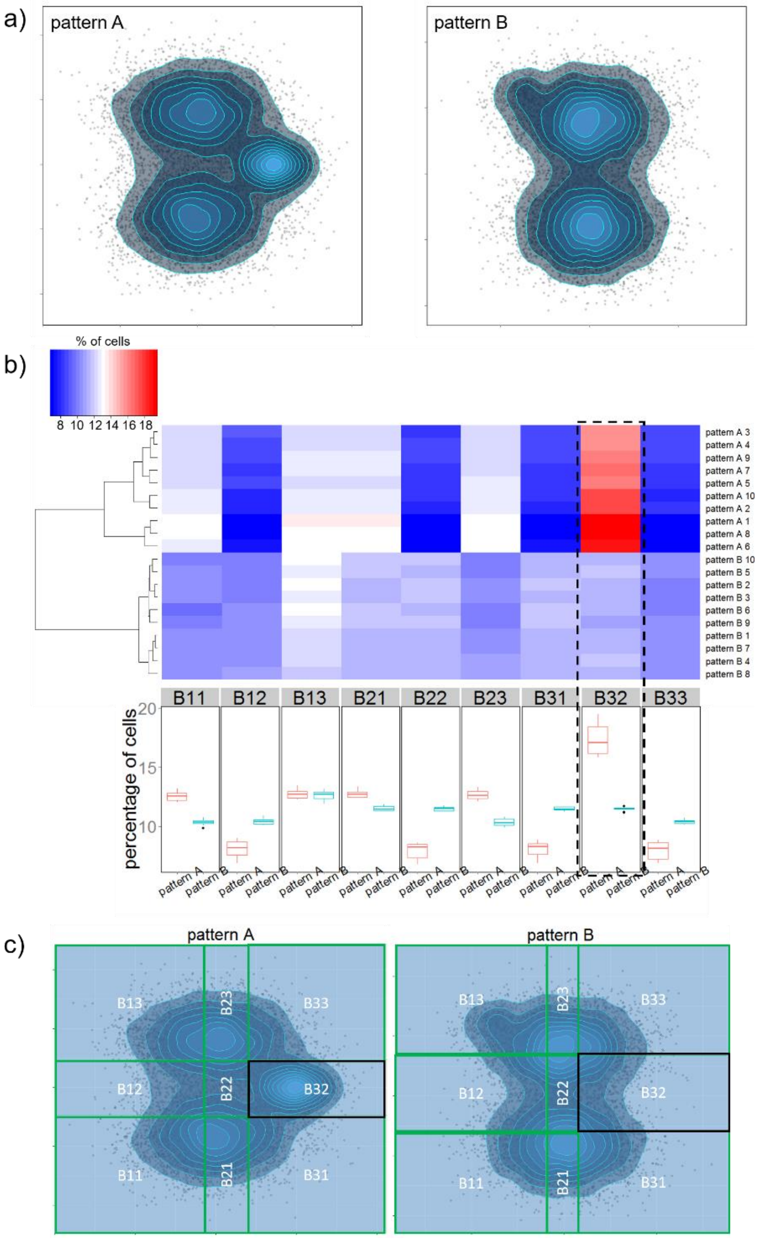


Figure 3.2 Percentile based binning is able to detect real differences between point-patterns. (a) Example of the two synthesized point-patterns A and B. The two large clusters in pattern A and B contain same number of cells and were generated with the same distribution. Pattern A contains a relatively large third cluster (10% to 20% of cells) at center right of the pattern and pattern B includes a smaller third cluster (2% to 5% of cells) on the top left corner. (b) Upper panel shows a heatmap of point percentage in each box for all samples, and lower panel shows boxplots of point percentage in each box between the groups of point-patterns. (c) Labels of each box. Highlighted is box B32.

3.4.3 The maximum number of bins for binning depends on the number of samples (patients)

We’ve seen in the previous section that the depth of analysis depends on the number of bins used. And here we are going to show that the maximum number of bins we could use depends on the total number of samples (patients), for using a large number of bins to classify a small set of samples would cause overfitting. We see that false positive rate (FPR) increases with the numbers of bins used for binning (Fig. 3.3a). However, the maximum number of bins with tolerable FPR (FPR < 0.05) increased when we increase the number of samples from 20 to 60 (Fig. 3.3a). While with 20 samples we can only use as many as 3 bins to keep FPR under 0.05, with 60 samples this number increased to 6. And using 6 bins, our method is able to identify both of the differences we artificially generated between point-pattern A and B (Fig. B3). To get a general picture of how the maximum number of bins relates to number of samples, we calculated the maximum number of bins with FPR = 0 (we use this stringent condition because i) Synthetic data is easier to classify; ii) Real dataset contains more than 2 markers, and multiple tests correction should be taken into consideration) for various number of samples. We found that when the two groups to be classified contain the same number of samples (patients), the maximum number of bins is around the square root of half the sample size (Fig. 3.3b). In reality, the number of samples (patients) in different groups is rarely equal. However, we can overcome this inequality by assigning different number of samples to cross validation set for different groups so that in training dataset each group will have the same number of samples. Thus, once we know the number of samples in training dataset, we get a reasonable estimate for the number of bins to use.

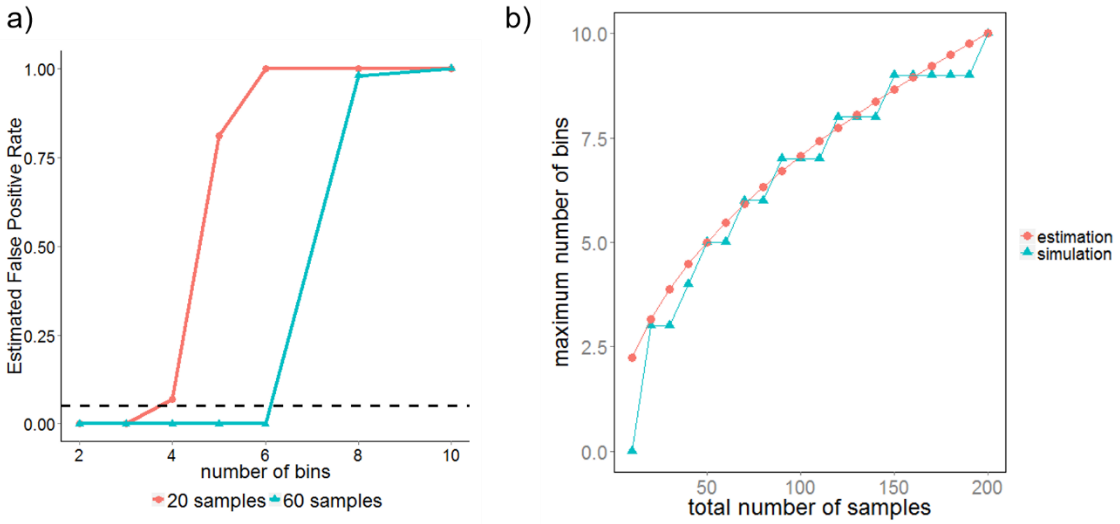


Figure 3.3 The maximum number of bins depends on the number of samples (patients). (a) Estimated false positive rate (FPR) vs. number of bins for 20 samples (red) and 60 samples (blue). The dotted black line represents FPR = 0.05. The number of bins that leads to a high FPR (>0.05) is considered overfitting the dataset. (b) The maximum number of bins with FPR = 0 vs. total number of samples used in the dataset. Blue dots show simulation results with our synthetic datasets and red dots shows the estimated number of bins using the rule of thumb: $maximum\ number\ of\ bins =\ round(\sqrt{number\ of\ training\ samples/2})$.

3.4.3.1 Application to two real human cytometry datasets

Next, we applied our method to two real flow cytometry datasets. Both datasets aim to identify differences between two biologically different patients/donor groups. In general, in order to get rid of debris and dead cells, some pre-processing steps should be taken before applying our method (e.g. manual/automatic gating to get live cells). In addition, depends on the question of interest, further gating can be applied to get more focused cell types, e.g. T cells, CD4^{+} T cells, etc. The pre-processed datasets are then the input for our method. We first determine the appropriate number of bins to use based on the number of samples in a dataset. Next, we apply the binning method showing in Figure 1 to the pre-processed dataset. Unlike the simulated dataset showing above which only contains two markers, real cytometry datasets usually measure much more markers which results in even more marker pairs. The binning method is applied to every possible pairs of markers. Then, in order to identify the important marker pairs, we separated the dataset into training and testing subsets. Using a classification algorithm called support vector machine (SVM) [124], we define important marker pairs as the ones that are able to achieve 100% classification accuracy in both training and testing subsets. Once these marker pairs were determined, we move on to identify which regions formed by these marker pairs (boxes) are significantly different between the two groups.

3.4.3.2 Old versus young

The first dataset we analyzed aims to find differences in the composition of immune cell types between old and young healthy donors [137]. Peripheral blood mononuclear cell (PBMC) samples from 34 healthy old donors (ages 60 and above) and 22 healthy young donors (ages 19 to 35) were taken, and their cellular composition were quantified by flow cytometry. In total, 16 markers were measured: Ki67, CD95, CD127, CD57, CD3, CD45RA, CD8, CD14, CCR4, CD27, CD11b, PD-1, CD4, CD28, CCR7, and a viability dye (live/dead). We first manually gated for the live cells (Fig. B4) which were used as input for our method. At this stage, about 20% of samples (4 young samples and 6 old samples) were randomly chosen as a cross validation set. We determined the optimal number of bins in remaining training dataset to be 5 (as the total number of samples in training set is 46, Fig. 3.3b), then we applied SVM classification based on the binning results of all possible pairs of markers. In total, we identified two pairs of markers (CD8 - CCR7, CD3 - CD4) that are able to classify old and young donors with 100% accuracy both in training and testing dataset (Fig. B5). Scatterplots of CD8 vs. CCR7 and CD3 vs. CD4 for randomly selected old and young donors are shown in Figure B6 and B7. Boxes whose cell percentages are significantly different between old and young donors are also identified. We selected the two boxes that are most different between old and young donors for demonstration below; remaining results can be found in supplementary information (Fig. B8 – B10).

3.4.3.3 Naïve CD8$^+$ T cells are found significantly decreased in elderly donors using only CD8 and CCR7 expression.

We first look at box B55 which contains cells whose expression of both CD8 and CCR7 are in the top 20% (i.e., CD8^{high} CCR7^{high}, Fig. 3.4a). We find that percentage of cells inside box B55 decrease significantly in old donors (Fig. 3.4b). On the other hand, mean fluorescence intensity (MFI) of cells inside box B55 are similar among donors for all markers, indicating cells inside box B55 are homogeneous across all samples (Fig. 3.4c). Notice that CD3 and CD45RA MFI levels are high for all samples, and since cells inside box B55 already express highest 20% of both CD8 and CCR7, one possibility is that cells inside B55 are naïve CD8$^+$ T cells. Indeed, cells in B55 agrees well with manually gated naïve CD8 cells (Fig. B11 & B12a) on single cell level. In addition, when comparing the expression of CD45RA and CCR7 between cells in B55 and manually gated CD8 naïve and memory cell types, we find that cells in B55 match well with naïve cells for young donors with slightly higher variation on CD45RA (Fig. 3.4d). Cells in B55 express higher variation in CD45RA for older donors, which is expected since box B55 was selected without expression information of CD45RA (Fig. 3.4e). Together, these results suggest that cells inside box B55 resemble naïve CD8$^+$ T cells. Decreasing of naïve CD8$^+$ T cells with ageing is a well-known observation in immunology [138] and is also identified in this dataset (Fig. B12b). In addition, we found that the abundancy of effector memory (TEM) and effector memory RA$^+$ (TEMRA) CD8$^+$ T cells are increased in old donors, as suggested by the increased percentage of cells in B51 (CD8^{high} CCR7^{low}) (Fig. B13).

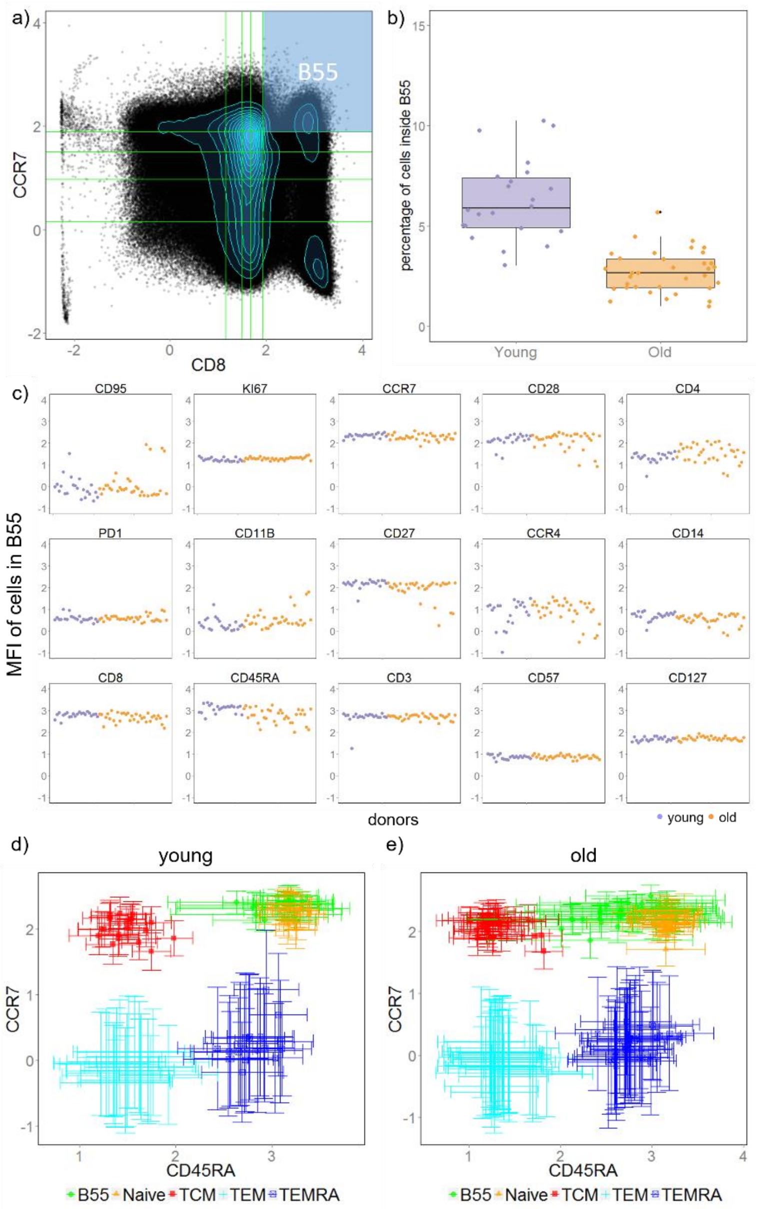


Figure 3.4 Naïve CD8$^+$ T cells were identified by our method as significantly decreased in old donors using only two markers: CD8 and CCR7. (a) An example showing scatter plot of CD8 vs. CCR7 with box B55 highlighted. (b) Boxplot of cell percentage inside box B55 between the two groups of donors. Each dot is a donor. (c) Scatter plot of mean fluorescent intensity (MFI) of each donor, each point shows a donor (purple: young, orange: old). (d) & (e) MFI of CD45RA vs. MFI of CCR7 for cells in B51, naïve and memory CD8

T cells. Each symbol shows a donor (young donors in d and old donors in e), vertical and horizontal error bars show standard deviation of CCR7 and CD45RA intensity respectively.

3.4.3.4 Distinction between naïve and memory CD8$^+$ T cells is blurred in old donors

Next, we analyzed cells inside box B52 (CD8^{high}CCR7$^{intermediate/low}$). The percentage of cells inside box B52 (Fig. 3.5a) was found to be increased in the old group (Fig. 3.5b). Similar to box B55, the MFI of cells in box B52 for all samples were at similar levels for most markers, indicating a homogeneous cell subset is identified among all donors (Fig. 3.5c). Notice that box B52 lies in between two peaks (Fig. 3.5a) which is a region often neglected or assigned to one of the peaks by manual gating, and we have shown above that cells in peak above B52 (i.e., B55) resemble naïve CD8 T cells and cells in peak below B52 (i.e., B51) resemble memory CD8 T cells (TEM and TEMRA). We hence infer that cells in B52 are transition cells between naïve and memory cells which increases with ageing. Figure 3.5d & e show how cells in B52 locate relative to manually gated naïve and memory CD8 T cells.

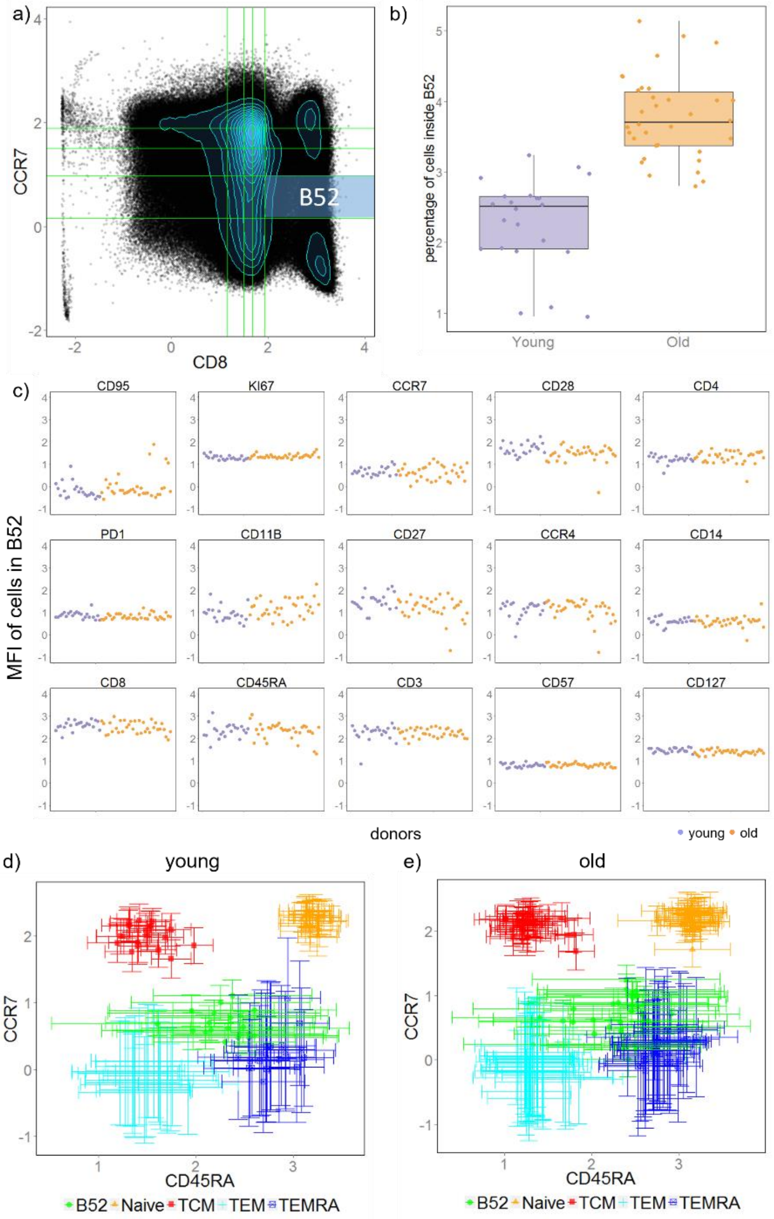


Figure 3.5 An intermediate cell region which is often neglected by gating methods is identified as significantly increased in old donors. (a) An example of scatter plot of CD8 vs. CCR7 with box B52 highlighted. (b) Boxplot of cell percentage inside box B52 between the two groups of donors. Each dot is a donor. (c) Scatter plot of mean fluorescent intensity (MFI) of each donor, each point shows a donor (purple: young, orange: old). (d) & (e) MFI of CD45RA vs. MFI of CCR7 for cells in B52, naïve, and memory CD8 T cells. Each symbol shows a donor (young donors in D and old donors in E), vertical and horizontal error bars show standard deviation of CCR7 and CD45RA intensity respectively.

3.4.3.5 CD4 versus CD8

Next, we applied our method to a mass cytometry dataset that originally aims to identify immune signatures among 8 types of human tissues: cord blood, PBMC, liver, spleen, skin, lung, tonsil, and colon [134]. There are in total 35 samples, 3 to 6 samples for each type of tissue (see Methods). The marker panel used for mass cytometry contains 41 markers with a focus on the function (cytokine expression) of T cells (a full list of all 41 markers can be found in supplementary information and [134]). Instead of differentiating the 8 types of tissues, here we tried to classify CD4$^+$ cells from CD8$^+$ cells in all types of tissues. This is a good test for our method since there exists great within-group variance (different tissues) in the two groups we're comparing, and we aim to find patterns that are consistent / similar across all types of tissues but are significantly different between CD4$^+$ and CD8$^+$ cells. Like the previous dataset, we divide these tissue samples into training and testing sets as well. From the 35 CD4 samples, 5 samples are randomly selected to be cross validation set; and the same was done for the 35 CD8 samples separately. Since there are in total 60 samples in training set (30 for each cell type), the number of bins to use is 5 (Fig. 3.3b). We identified 7 pairs of markers that were able to classify CD4$^+$ and CD8$^+$ cells with 100% accuracy for both training and cross validation datasets. Only 1 marker pair (CCR10 vs. CCR9) out of the 7 contains purely trafficking markers. This indicates that CD4$^+$ and CD8$^+$ T cells can be more easily differentiated by their function and lineage markers than trafficking markers, which is consistent with the results in the original paper [134]. We selected one of the seven marker pairs: Interleukin (IL)-2 vs. CD25 to show in Figure 3.6. The pattern formed by CD4 cells is distinct from CD8 cells in that CD4 cells express significantly more IL-2 and slightly more CD25 in all types of tissues, which agrees with previous findings based on circulating immune cells [139]. In addition, we found that percentage of cells in box B13 (red shaded region in Fig 3.6, IL-2 low and CD25 intermediate) is significantly higher in CD8 cells, which is a subtle difference that would be missed by algorithms based on a peak finding.

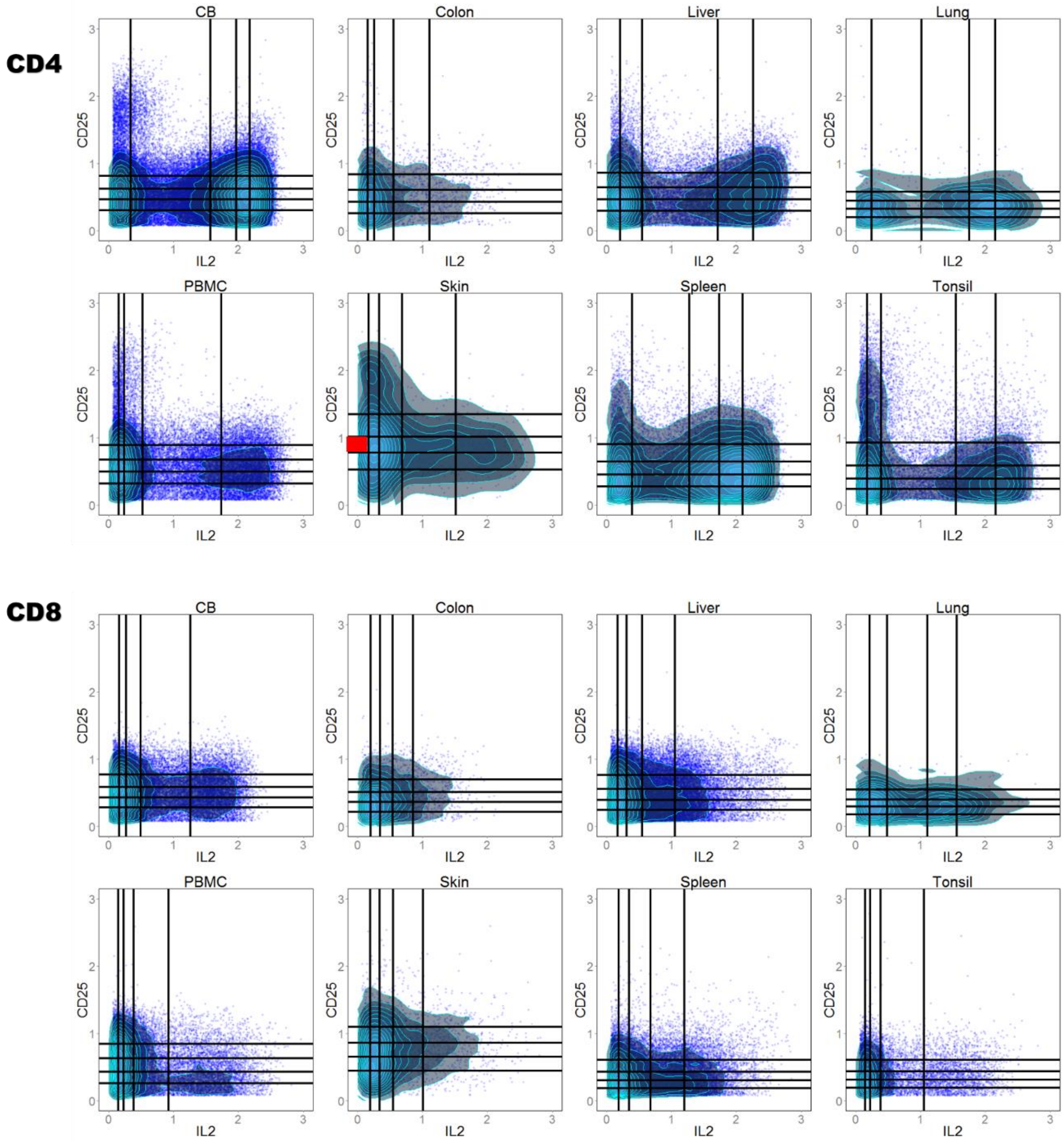


Figure 3.6 Patterns formed by IL-2 vs. CD25 is distinct between CD4 and CD8 cells. One randomly chosen sample for each tissue is shown. The same sample for each type of tissue is chosen to illustrate both CD4 and CD8 cells. Percentage of cells in the red shaded box (B13: IL-2 negative and CD25 intermediate) is

significantly higher in CD8 cells comparing to CD4 cells. Cells inside box B13 also express CD45RA, TNFα, and CD127 (Fig. B12).

3.5 Discussion

The complexity of cytometry data has increased significantly in the last few years due to the advancement in experimental techniques that enable measurements of dozens of parameters on each cell for millions of cells [5]. Novel analysis algorithms are being introduced at a rapid pace to deal with this data deluge that identifies clusters of cells and project the high dimensional information graphically in innovative ways. However, these graphics are not directly interpretable and translatable into hypotheses and actions by biomedical researchers and clinicians. There is also the flaw that nearest neighbors are not meaningful in high dimensions, which is a phenomenon referred to as the "curse of dimensionality" [11, 77]. Here we introduce a simpler, alternative approach we term CytoBinning. Our analysis approach combines automation of a more traditional workflow (as advocated in [11]) and machine learning which links the high dimensional data back to two biomarkers which can be represented as 2D scatter plots. The 2D scatter plot outputs are designed to be directly interpretable by biomedical researchers and clinicians, who have an established intuition for the meaning of these graphics. Thus, we are able to leverage their existing expertise in interpreting these kinds of scatterplots. When the differences in phenotype are small, CytoBinning is able to further focus the researcher or clinician's attention by identifying, which specific regions of the scatter plot exhibits the most notable differences between two groups of donors, allowing subtle shifts in the immune phenotype to be highlighted.

In contrast to automated gating methods that focus on the exact position of density peaks or the number of groups formed by cells, CytoBinning doesn't estimate the probability density distribution of cells, and thus its findings are not limited to regions with high cell density or sensitive to shifts in calibration. Instead, it extracts the pattern of 2D dot-plots and represents it with a sequence of cell percentages. This enables the comparison across samples measured in different experiments (given the markers are the same and they are measured in the same channel respectively). In addition, CytoBinning does not require any *a priori* biological understanding to guide the path of analysis. Conversely, it provides a list of important marker pairs and related important cell sub-regions for biological researchers to subsequently interrogate.

In the first public dataset we analyzed, which compares lymphocyte populations in old and young healthy donors, CytoBinning automatically discovered a decrease of naïve CD8$^+$ T cells in the elderly, a well-known yet subtle phenotype. In addition, CytoBinning identified a region in the scatterplot of relatively low cell density between two well-established cell clusters which is increased with ageing as a new area of interest for the biological researcher. Two markers (CD8 and CCR7) are sufficient to pinpoint this subset of cells which resides between naïve and memory CD8$^+$ T cells and is not associated with a local peak in cell density in the scatterplot. Such an area would be missed by both manual gating and density-based algorithms, or by focusing exclusively on peaks in density.

The second public dataset we analyzed was even higher dimensional, based on mass cytometry from eight types of human tissues. CytoBinning analysis of CD4$^+$ vs. CD8$^+$ T cells automatically discovered higher expression of IL-2 in CD4$^+$ T cells as we would expect [139], and shows that this overexpression is consistent throughout all eight types of human tissues studied. In addition, CytoBinning correctly identified that CD25 is also more highly expressed in CD4$^+$ T cells [139]. This difference in CD25 and IL-2 was consistent among all types of tissues, which is known and therefore obvious to a biological researcher. However, it also demonstrates the power of our method as this marker pair was re-discovered without prior knowledge from a heterogeneous dataset incorporating 35 samples from 8 different tissues, each labelled with 41 markers. Hence, in addition to avoiding the pitfalls of density-based approaches, when applied to very high-dimensional datasets CytoBinning is able to select the salient markers which discriminate between groups of samples.

In summary, CytoBinning as a robust, automated approach to analyze high throughput cytometry data presented in familiar and interpretable 2D scatter plots. While simultaneous assessment of all markers is an important vision and challenge, in the interim there is a need to facilitate interpretation of high-dimensional data given the evident gap between our technological ability to acquire this information and our ability to understand it. CytoBinning fills the void between conventional manual analysis and complex automated analysis to extract deep content in scatterplots which can be later cascaded into more complicated clustering or classification algorithms to obtain novel biological insights. This has particular potential value in clinical and biological research settings where high-dimensional data is increasingly available and commonly not fully understood. CytoBinning is able to identify the most important markers, while also highlighting novel cell populations that distinguish comparator datasets even if these are to be found in areas of low cell density. Hence, it is a practical analysis approach with potential to fill the complexity gap in the interpretation of high-dimensional data in a wide range of biomedical and clinical settings.

3.6 Methods

3.6.1 Binning

The binning we used in our method has been previously proposed to estimate mutual information (MI) [135]. Given bin number *b*, equally populated bins are drawn based on single cell expression of marker A and marker B independently. These bins are then overlaid on each other so that a grid is formed with b^2 regions (boxes). Percentage of cells inside each box is then an estimation of the joint probability $P(A_i, B_j)$, where i and j are the corresponding bins this box locates at. For a random distribution where marker A and marker B is not correlated in any way, $P(A_i, B_j)$ should be approximately the same in for every box. This is not true if marker A and marker B is related in any way (i.e., their mutual information is not zero, this relationship can be both linear and nonlinear). We use all $P(A_i, B_j)$s as a coarse-grained representation of the point pattern between single cell expression of marker A and marker B. (Fig. 3.1) In our method this binning is done for every pair of markers.

3.6.2 Determine appropriate number of bins

We deduced a relationship between the maximum number of bins with zero false positive rate (FPR) and the number of samples used in classification using our synthetic data. The relationship we found is:

$$maximum\ number\ of\ bins = \mathrm{round}\left(\sqrt{number\ of\ training\ samples/2}\right)$$

Thus, for a given dataset, an estimation of the number of bins to be used is achieved. In addition, we estimate FPR as follows:

1. For a given number of bins, apply the afore-mentioned binning method to one pair of markers. Each sample is now represented by the vector of $P(A_i, B_j)$.
2. Randomly divide all samples into two groups.
3. Apply SVM classification (ksvm function in R package ks, with linear kernel and C=10) on the randomly divided groups.
4. Repeat step 2 & 3 for 100 iterations, record the frequency when classification accuracy achieved 100% in step 3.
5. Repeat step 1 to 4 for all marker pairs, calculate the mean frequency of one pair achieving 100% accuracy. This frequency is used as an estimation of FPR.
6. Repeat steps above for all numbers of bins.

3.6.3 Log ratio transformation

The percentages of cells in each box obtained with CytoBinning is compositional as they add up to 100. To get rid of this dependency, we divide the percentages by their median before taking log with base 2 for every sample and every marker pair.

3.6.4 Selecting important marker pairs

Once the number of bins is determined, we divide all samples into a training set (about 80% of total samples) and testing set (the remaining 20% of all samples). SVM is applied to the training set and classification boundary obtained for every pair of markers. We use the obtained classification boundary to predict the cross-validation set. Pairs that reached 100% accuracy for both training and cross-validation datasets are chosen as important marker pairs.

3.6.5 Selecting important boxes

We combined boxes formed by all selected marker pairs and applied statistical test (Wilcox) for the percentage of cells in each box. We then corrected the p values for multiple comparisons with Bonferroni correction, and boxes with p-value <0.001 after correction are selected as important boxes. Important marker pairs selected above without any important boxes are eliminated from the important marker pair list.

3.6.6 Dataset 1: Comparing old and young healthy PBMCs

Overview of samples. This dataset is published in reference [59] and downloaded at Flow Repository (http://flowrepository.org) website [140]. These samples were processed in two experiments, with 19 samples from young donors and 20 samples from old donors processed in the first experiment, and the remaining samples processed in the second experiment. The panel of markers was kept the same for both experiments. In total, 16 markers are measured: Ki67, CD95, CD127, CD57, CD3, CD45RA, CD8, CD14, CCR4, CD27, CD11b, PD-1, CD4, CD28, CCR7 and a viability dye (live/dead). Details of sample storage and processing can be found in [59].

Pre-processing. Downloaded FACS files were first compensated based on the spill matrix in the FCS files, and then manually gated to get live cells (Fig. S4). Logicle transformation was performed with w=0.5, t=262144, and m=4.5 using logicleTransform function in flowCore package with R.

3.6.7 Dataset 2: Comparing CD4 and CD8 T cells in various types of tissues

The dataset used for the demonstration was first published in [134] and downloaded from flow repository website (https://flowrepository.org/) [140]. Tissue types, number of samples, and the reason for surgery are listed in Table 1. Immune cells were isolated from collected tissues and cryopreserved. They were then thawed and washed for mass cytometry experiment. Two panels of antibodies were used for staining, each containing 41 markers. The two panels were named as “Function” and “Traffic” according to the antibodies included in it. We only used function panel in this paper. Details of experimental process and the lists of antibodies can be found in [134]. The downloaded samples from flow repository are FACS files, pre-gated to major immune types (e.g. CD4, CD8, NKT, etc.). We used only CD4 and CD8 cells. We performed logicle transformation using logicleTransform in R package flowCore, with parameters w = 0.25, t= 16409, m =4.5, and a=0 according to [134]. The logicle transformed data were then saved as text files for further analysis.

Table 3-1 Summary of sample information

Tissue type	**Number of samples**	**Reason for surgery**
Cord Blood	5	Healthy donation at neonate
PBMC	4	Healthy donation
Tonsil	5	Tonsillar Hypertrophy
Spleen	3	Splenectomy (Due to Distal Pancreatectomy)
Colon	6	Routine Colonoscopy
Skin	5	Abdominoplasty or Mastectomy - Invasive Ductal carcinoma
Lung	4	Lung cancer resection
Liver	3	Liver transplantation

Chapter 4 Feature selection with CytoBinning in mass cytometry data

4.1 Overview

In this chapter, I try to tackle feature selection problem in mass cytometry data. Since the informative feature in cytometry data reside not in single measurements but the point patterns formed by multiple measurements, I first define features for cytometry data as the patterns in 2D scatterplots between two measurements. Using CytoBinning, I developed and introduced in Chapter 3; I was able to quantify these point patterns and select features accordingly. The experimental data were obtained from the online repository: https://flowrepository.org/.

4.2 Abstract

Mass cytometry provides the potential to measure up to 100 different protein expressions simultaneously on a single cell level. To keep up with the high dimensionality of measurements, most analysis methods aim to utilize all measurements simultaneously by engaging sophisticated high dimensional algorithms. However, the real number of measurements that are directly related to the question of interest is much lower than the total number of measurements, i.e. a large number of measurements are either irrelevant or redundant, resulting in information overload and noise. Hence, feature selection methods are much needed for mass cytometry data. The difficulty in selecting features for mass cytometry data is that the valuable information resides not in single marker expression but the point pattern formed by expression of multiple markers which is hard to quantify. Here, I apply the binning method – CytoBinning, introduced in Chapter 3 – to quantify information in point patterns hence enable feature selection for mass cytometry data. I demonstrate with a publicly available mass cytometry dataset that CytoBinning is able to

select meaningful features (i.e. marker pairs) within the dataset that are able to classify eight types of human tissues simultaneously. In addition, novel tissue-specific profiles are identified when combining information contained in markers from selected pairs.

4.3 Introduction

Mass cytometry is a relatively new technique that enables simultaneous measurements of more than 40 cellular parameters at a single-cell level. It outputs data in a format similar to flow cytometry but offers the potential of much higher dimensionality. In contrast to flow cytometry, it does not suffer from between channel fluorescence leaking that plagues flow cytometry data and hence doesn't need compensation. On the other hand, it is able to measure far more cells (on the magnitude of millions) comparing to other high-content single cell measurement techniques like cellular imaging and single-cell RNA sequencing. These features make mass cytometry a perfect tool to investigate cellular heterogeneity as we often see in the immune system, tumor cells, etc. [79, 80, 141]. The rapidly increasing number of measurements, on the other hand, poses great challenges in analyzing mass cytometry data and also increases the probability of information overload.

In recent years, analysis methods specifically designed for mass cytometry dataset are rapidly emerging [5]. Most of these methods aim to fully embrace the high dimensionality provided by mass cytometry by utilizing all measurements simultaneously with the same weights. For example, SPADE [75], as one of the most well-known methods for mass cytometry analysis, uses an agglomerative clustering algorithm to divide cells into different groups and then applies a minimum spanning tree (MST) based algorithm to visualize the clustering results. SPADE has been successfully applied to several mass cytometry datasets [82-84]. Another example is the method called Citrus which utilizes hierarchy clustering for cell type discovery and then calculates properties of these cell subtypes and use them to determine a list of cell subtypes with significantly different frequencies between two sample groups [85]. There is also an algorithm called PhenoGraph that applies k nearest neighbors (k-NN) to identify cell subtypes in a given sample [86]. On the other hand, clustering based methods that are developed for flow cytometry datasets, like FLAME [58], flowSOM [62], etc. have also been successfully applied to mass cytometry datasets [87, 88]. One problem of using all measurements simultaneously without weighting is that the definition of distance and density become less meaningful with the increase of dimension due to "curse of dimensionality" [77, 142]. In addition, in real datasets, the number of measurements that are directly related to the question of interest is often much smaller than the total number of measurements due to redundant measurements (e.g., markers that have similar biological functions) and irrelevant measurements (e.g., not known before experiment). Hence feature selection methods are much in need to pre-process mass cytometry data before distance, and density-based methods can be applied. Nevertheless, since the valuable information in mass cytometry data exists not in single measurement but the content of point patterns formed by multiple measurements, it is hard to quantify meaningful features for mass cytometry data.

Here we define features of mass cytometry data as the 2D point pattern formed by two measurements. We use CytoBinning – the binning method we proposed in Chapter 3 – to quantify content in a point pattern into a vector of cell percentages. In this way, direct comparisons between point patterns are made possible, and important features are selected accordingly. We demonstrate our method with a publicly available mass cytometry dataset that measures immune cells collected from eight types of human tissues [134]. Without any prior biological information about immune signatures in different types of tissues, we are able to select five pairs of markers whose point patterns are distinct among different types of tissues. In addition, we show tissue-specific trafficking features for immune cells in different types of tissues with a combination of only six markers.

4.4 *Results*

We demonstrate our method using a publicly available mass cytometry dataset first published in [134]. In this dataset, a total of 35 samples collected from 8 types of human tissues were measured using mass cytometry with two marker panels. Both panels contain 41 cellular markers; the first panel includes mostly trafficking and surface markers while the second panel contains both surface markers and intracellular cytokines (see Methods). The first panel is called "trafficking panel" while the second is called "function panel" according to [134]. Table 1 summarizes the types of tissues, number of samples for each tissue type and the source of tissue samples. For illustration purpose, here we focus on CD4$^+$ T cells that were manually gated by the authors of reference [134] with the traffic panel, and our goal is to identify tissue-specific trafficking profiles within CD4$^+$ T cells. Using CytoBinning, we show that we are able to simultaneously classify the 8 types of tissues with only 2 markers (we will refer to them as marker pairs). In total, we identified 5 such marker pairs that enable simultaneous classification. There are only 6 distinct markers in these 5 marker pairs. t-SNE analysis based on these 6 markers reveals point patterns with tissue-specific signatures.

4.4.1 An overview of CytoBinning

The binning method we use in this Chapter – CytoBinning – is first proposed in Chapter 3 for analyzing flow cytometry data. In summary, CytoBinning quantifies information in 2D scatterplots with percentile-based binning. To illustrate this with mass cytometry data, we selected two patients with similar but distinct patterns between CD45RO and CCR7 expression (Fig 4.1a and 4.1c). The point pattern shown in Figure 4.1a contains three clusters, one big cluster on the upper left corner and two relatively smaller clusters on the upper right and lower right corner. While the pattern in Figure 4.1c is similar to 4.1a, they are different in that there are only two clusters in 4.1c. Instead of two distinct clusters, the lower right cluster has a thick and long tail that extends upwardly. As illustrated in Figure 4.1a, CytoBinning first divides cells into a pre-defined number of bins (3 bins in Fig 4.1) based on the expression profile of the two markers independently; every bin contains the same number of cells. The bins are then combined so that the scatterplot is further divided into smaller boxes, and the percentage of cells in each box is calculated. These percentages change with the patterns in a scatterplot. We use these percentages as a coarse-grained representation of the point patterns in a scatterplot. CytoBinning is able to capture the subtle difference between different patterns as well as prominent features of a pattern. As shown in Figure 4.1, the distinct upper left and lower right clusters in both patient samples are captured by the two boxes with largest cell percentage (upper left and lower right box, Fig 4.1b, 4.1d). For pattern in Figure 4.1a, the third cluster (upper right) is cut into multiple regions with majority of the cluster inside the middle and middle right box, which reflects the shift between the center of the upper left and upper right cluster (Fig 4.1b). Since the patterns are different between Figure 4.1a and 4.1c, the CytoBinning representation is also different, with cell percentages in the four lower right boxes more homogeneous in Figure 4.1d comparing to Figure 4.1b, corresponding to the wider cluster in the lower right region of Figure 4.1c.

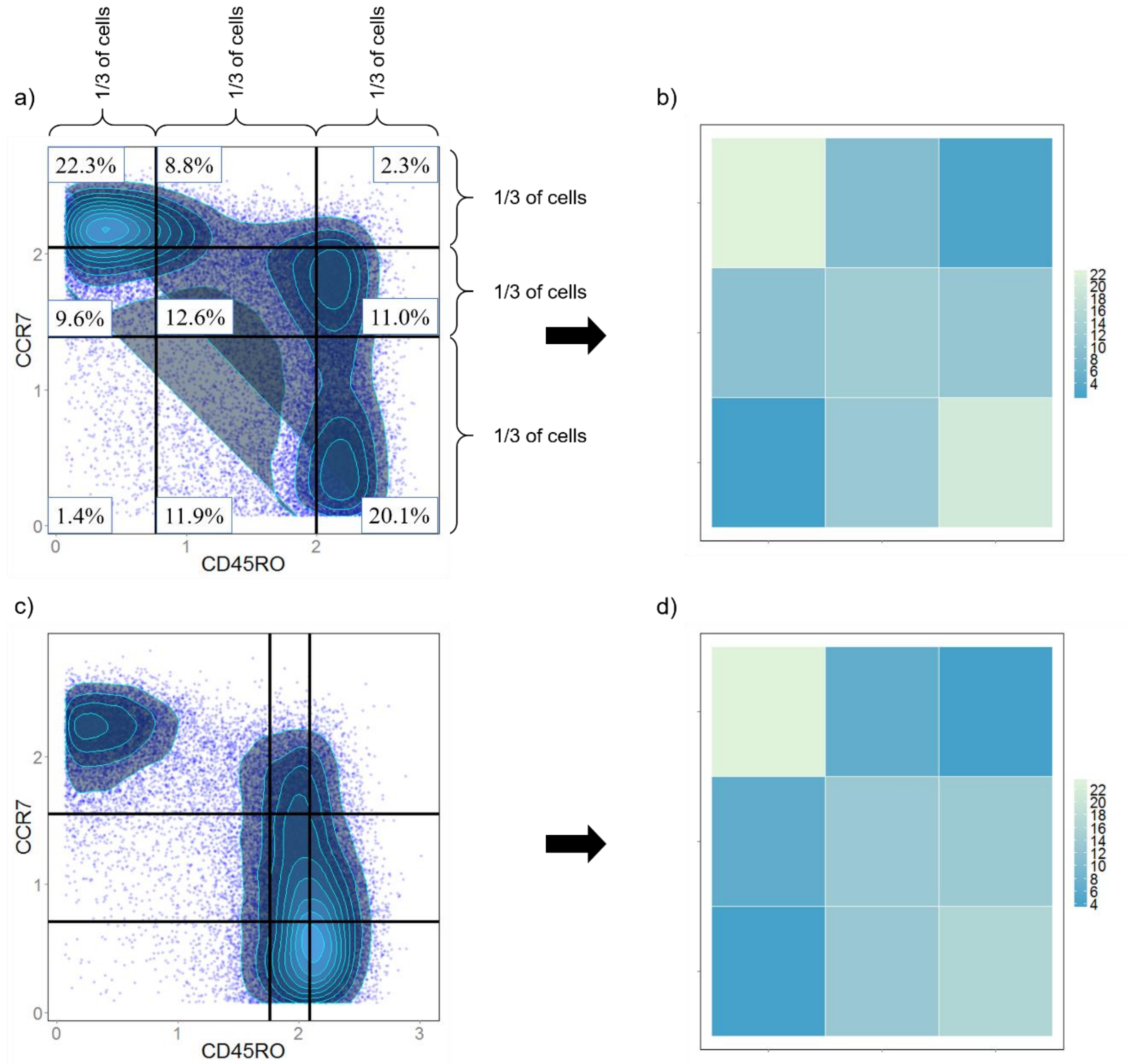


Figure 4.1 An illustration of CytoBinning and how it can represent the subtle difference between two similar patterns. a) An example of percentile based binning as a representation of 2D point-pattern (number of bins = 3). b) The corresponding representation is shown in heatmap for a). c) A similar but distinct point pattern to a). d) The corresponding representation is shown in heatmap for c).

4.4.2 CytoBinning is able to identify key differences among CD4 T cells in eight types of human tissues

Our analysis revealed 5 marker pairs (features) that are able to simultaneously classify the 8 types of human tissues with less than one error on average (see Methods and Fig C1): PD-1 & CCR7, CCR7 & CD161, CCR5 & CD27, CCR7 & CD27, and CXCR6 & CCR5. Each marker pair formed distinct expression patterns among different types of tissues.

Figure 4.2a, figure C4, figure C7, figure C10 and figure C11 illustrate point patterns formed by each marker pair respectively. We select the marker pair with the lowest number of average error – PD-1 & CCR7 – to discuss in detail here. Results for the other four marker pairs can be found in Appendix C.

The scatterplots in Figure 4.2a shows randomly selected examples of PD-1 and CCR7 expression pattern in different types of tissues. Indeed, these patterns appear different from tissue to tissue. Even though patterns of cord blood, liver, PBMC samples look similar to each other at first glance, they are indeed distinct from each other. The PBMC sample has a longer tail along CCR7 axis – comparing to cord blood and liver samples. Moreover, the liver sample has a wider spread in the lower right corner which has high expression of PD-1 and low expression of CCR7. In particular, we notice that compared to other tissues, tonsil has a unique cell subtype that is PD-1^{high} and CCR7^{-} (highlighted in orange circle in Fig 4.2a). To validate this subtype, we manually gated for it (Fig C2a), calculated percentage of PD-1^{high} CCR7^{-} cells in each sample, and compared their percentage across tissue types. The results show that this subtype is indeed highly expressed in tonsil samples, but are almost not seen in other tissue types except lung and skin (Fig C2b). To better understand the PD-1^{high} CCR7^{-} subtype, we calculated their mean expression level of other markers expressed by cells in this subtype and identified that they also express CXCR4, CXCR5, CD27, CD69, ICOS and CD95 (Fig C3). Even though the simultaneous tissue type classification is achieved with cell percentages in all the 25 boxes, we show that with cell percentage in the two most important boxes in classification (B15, shaded in blue in Fig 4.2a; and B54, shaded in red in Fig 4.2a) one can already get a reasonably good separation of the eight types of tissues (Fig 4.2b).

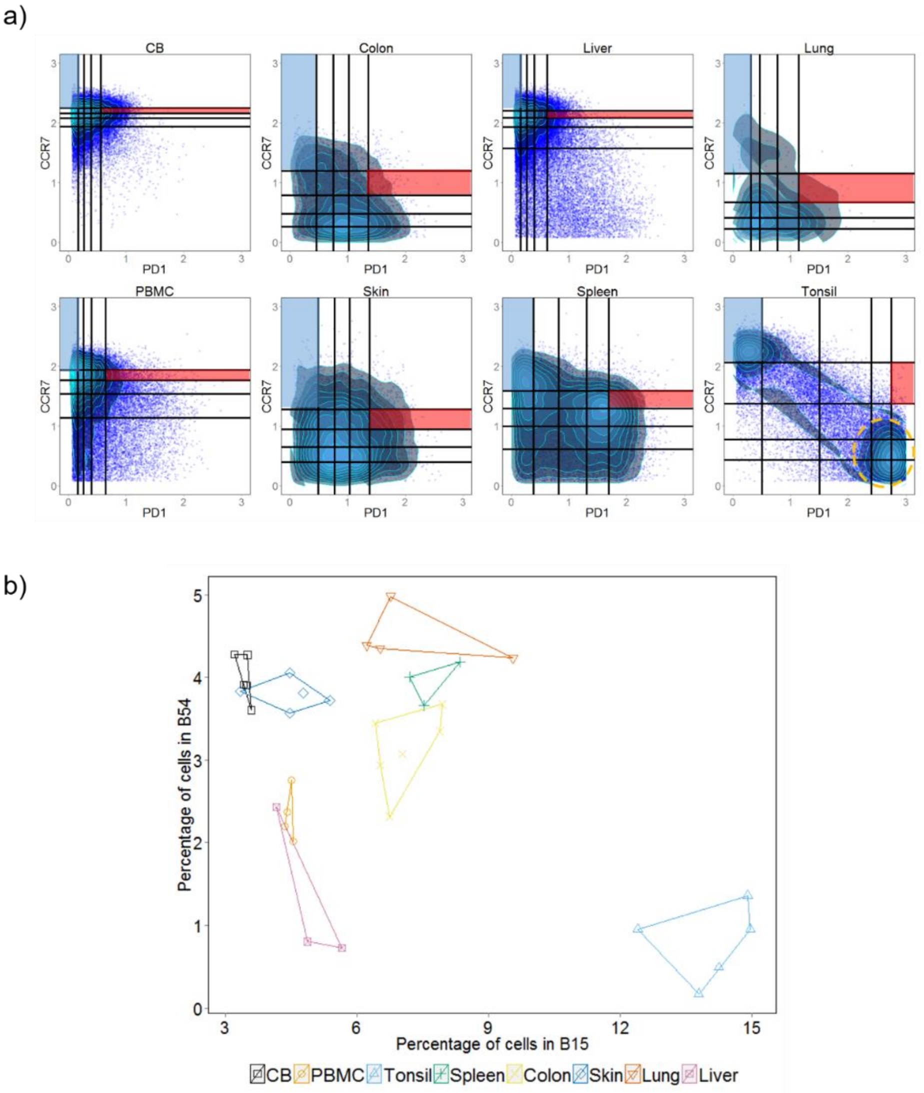


Figure 4.2 Distinct patterns across eight types of human tissues are observed with marker pair PD-1 & CCR7. a) Scatterplot showing single cell expression pattern formed by CCR7 and PD-1. Solid black lines are thresholds determined by CytoBinning. The shaded areas are the two most important boxes in classifying tissue types (blue: B15, red: B54). The region labeled with the orange circle is a cell subset express a high level of PD-1 and express little CCR7. This cell subset is not observed in other tissue types except tonsil. b) Plot of cell percentage in box B54 (shaded in red) and box B15 (shaded in blue). Each color and shape denotes one type of tissues.

Similarly, results can be seen from the other four marker pairs (Appendix C). For example, we found an increased frequency of CD161$^+$CCR7$^+$ cells in colon tissues (Fig C4 – C6) and a higher level of CD27$^+$CCR5$^+$ cells in lung tissues (Fig C7 – C9) and validated it with manual gating. However, deeper biological insights cannot be generated from these observations since not all samples are collected from healthy donors.

4.4.3 New insights can be identified by combining selected marker pairs

After selection of marker pairs whose patterns are distinct across different tissue types, we combined these marker pairs to gain new insights of the dataset. In particular, we combined the 6 distinct markers in the 5 marker pairs we selected and performed t-SNE analysis on all samples combined (the samples are down-sampled before combining, with 250 cells randomly selected from each sample). We found that, as shown in Figure 4.3, the resulting 2D scatterplot shows distinct tissue-related clusters. CD4$^+$ T cells in cord blood, PBMC, and liver tissues are mixed and form a big cluster, cells in lymphoid tissues (spleen and tonsil) stay close together with cells in tonsil form a distinct small cluster. Cells in skin, lung, and colon are close together with some mixtures. In addition, we notice that cells in colon samples are divided into two clusters (Fig 4.3a). Both clusters express CD161 (Fig 4.3c). However, the expression level of CCR7 are differentiated between these two clusters, with one cluster express higher level of CCR7 than the other (Fig 4.3b). Also, the highest level of CCR7 expression is observed in blood samples (cord blood and PBMC) and liver samples. This suggests there are more naïve CD4 T cells in circulation than in resident tissues. Fig 4.3b also shows the transition of CCR7$^+$ to CCR7$^-$ from left to right, i.e., from circulating immune cells to immune cells resident in a distant tissue.

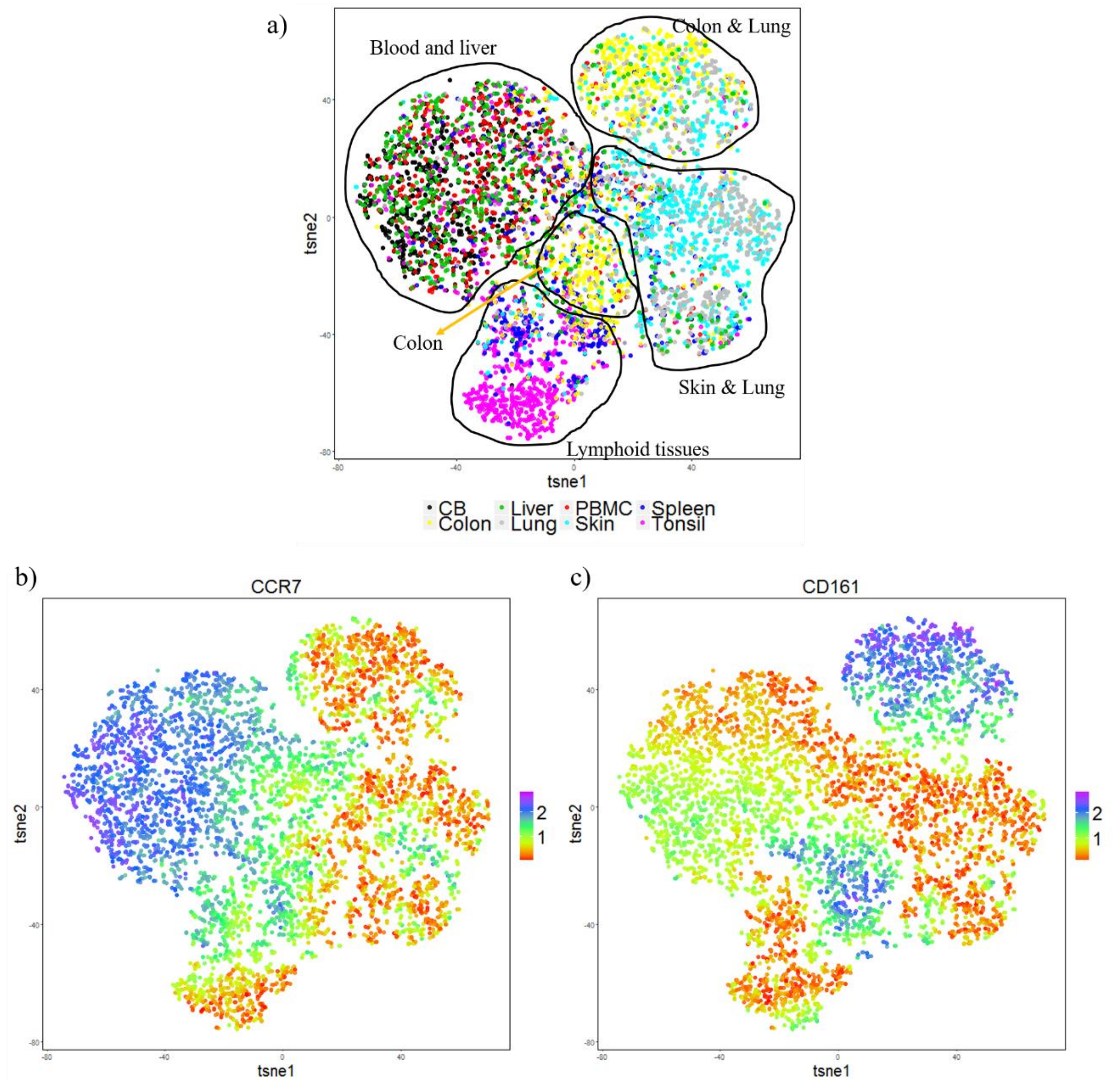


Figure 4.3 Separation of tissue types can be observed in t-SNE plot generated by the expression level of the 6 selected markers (PD-1, CCR7, CCR5, CD27, CD161, and CXCR6). a) The contours are manually drawn to show the related clusters of cells. Each point is a single cell, and different color indicates cell in different types of tissues. b) The same scatter plot as in a) with color coding of CCR7 expression level. c) Similar to b), with the expression level of CD161.

4.4.4 Distinct cytokine secretion patterns across human tissues can be illustrated with 8 markers

Next, we applied similar analysis to CD4$^+$ T cells measured with function panel. Similarly, we identified 5 pairs of markers that are able to simultaneously classify the eight types of human tissues with less than 1.5 errors on average (but larger than 1 error) – CXCR6 & CCR9, CXCR6 & CD161, CXCR6 & CXCR3, PD-1 & CCR5, and CCR2 & CCR6. Note that they are all surface markers. We combined the unique markers from these 5 pairs and again performed t-SNE analysis based on all samples combined (the samples are down-sampled before combining, with 250 cells randomly selected from each sample). As with the traffic panel, the t-SNE plot can be coarsely divided into 4 big regions: blood (cord blood and PBMC) and liver –right, secondary lymphoid tissues (tonsil and spleen) – center top, skin and lung – middle left, and colon – top left and bottom center (Fig 4.4a). In addition, we showed cytokine expression levels in t-SNE scatter plot (Fig 4.4b) divided into 3 levels: 0 to 1, 1 to 2 and 2 to 3. The expression level of different cytokines varies significantly, with IFNγ and Granulocyte-Monocyte Colony-Stimulating Factor (GM-CSF) most highly expressed and the expression of other cytokines much lower and sparse. In addition, cytokine secretions are localized related to tissue types. For example, both IFNγ and GM-CSF are mainly expressed by CD4 cells in colon, lung, and skin. Moreover, both IL-17A and IL-22 are mainly expressed by CD4 cells in colon and skin. Similar results were reported in reference [134] with the use of all 41 markers simultaneously. In comparison, our method is able to achieve better tissue distinctions. As discussed above, CD4 cells in colon tend to form two groups where one group express mid-level of CCR7 and the other group do not express CCR7. We observed two groups of CD4 cells in colon with function panel as well, where one group (bottom center) doesn't express any cytokines comparing to the other group (upper left). It is possible that the group of cells with low to no cytokine expression are the same as the CCR7 expressing group (naïve) identified with trafficking panel (Fig 4.3). In addition, other regions with low cytokine expression (blood samples and lymphoid samples) are also the regions with higher CCR7 expression observed in Figure 4.3. However, since the function panel does not contain CCR7, we are not able to verify this.

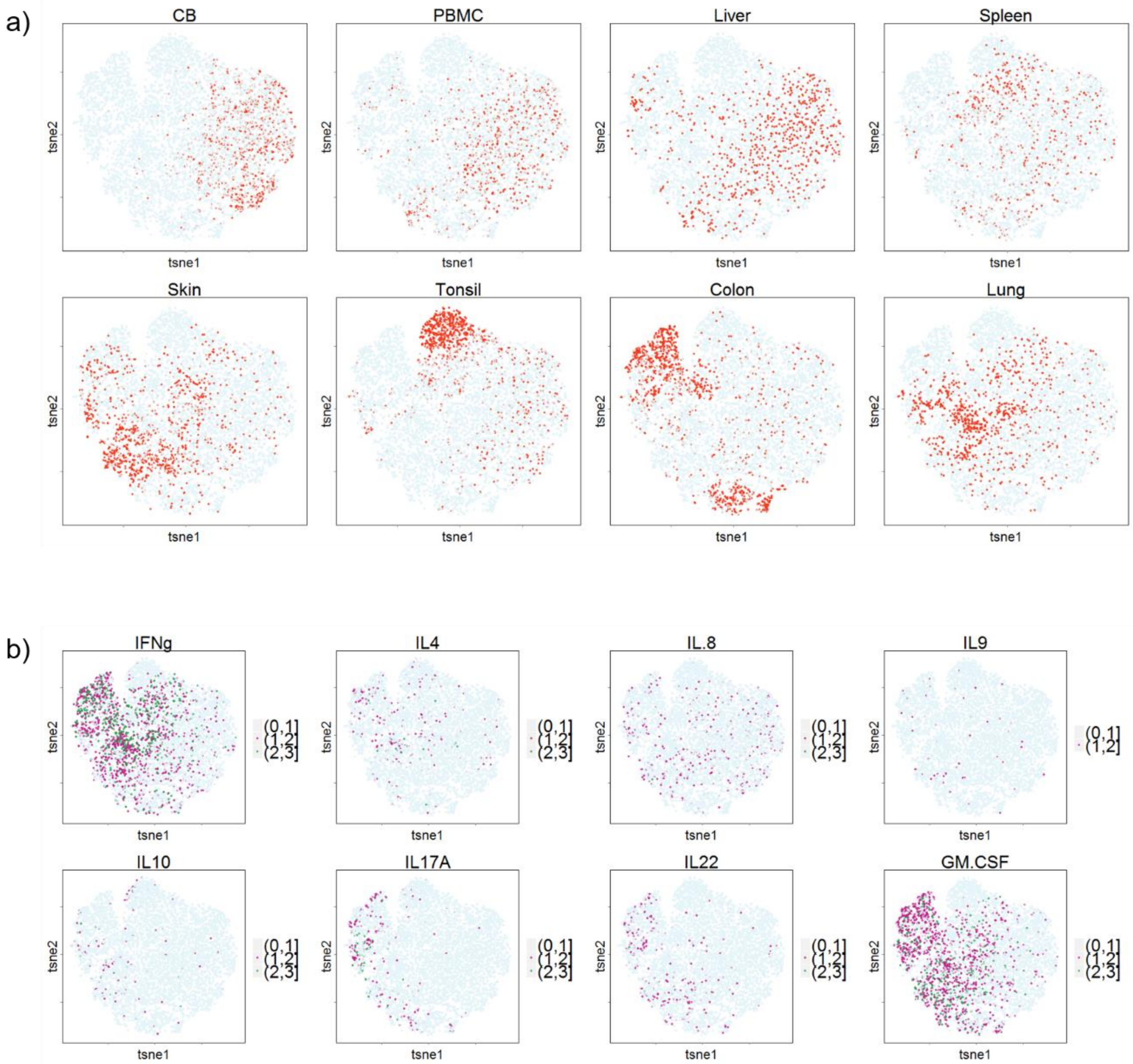


Figure 4.4 Tissue-specific profiles can be observed in t-SNE plot generated by the expression level of the 8 selected markers (CXCR6, CCR9, CD161, CXCR3, PD-1, CCR5, CCR2, and CCR6) in function panel. a) t-SNE plots labeled in red by cells in each tissue type. b) Expression of cytokines. The expression level is divided into 3 categories: (0, 1) light blue; (1, 2) purple; and (2, 3) dark green.

4.5 Discussion

Mass cytometry expands the horizon of cytometry by enabling simultaneous measurements of a large number of markers and hence provides an opportunity to discover new features and cell types. However, the large number of measurements can also raise problems because of information overload and "curse of dimension". Currently, the state-of-art methods incorporate all markers simultaneously via sophisticated algorithms with the assumption that all markers are equally important, even though the importance of markers varies based on the biological question and some markers could be completely irrelevant. The lack of dimension reduction/feature selection methods designed for mass cytometry data is in part due to the fact that unlike other single-cell data, meaningful information within mass cytometry data comes from the point patterns formed by multiple measurements which is difficult to quantify. Here, we aim to tackle the feature selection problem by define features as point pattern in pairwise scatterplots, and we propose a binning method – CytoBinning – to quantify point patterns and enable direct comparison of patterns.

With a publicly available dataset that tries to identify the immune signature in different types of human tissues, we demonstrated the use of CytoBinning for feature selection in mass cytometry data. We identified five marker pairs that are able to simultaneously classify all eight types of tissues. In particular, we discovered a cell subset that is unique to CD4 cells in tonsil samples: PD-1^{high}CCR7^{-}. This subset also expresses CXCR4, CXCR5, CD27, CD69, ICOS, and CD95. However, since the tonsil samples are collected from tonsil hypertrophy patients, we do not know if this unique cell subset is due to the disease or a signature of CD4 cells in tonsil. Similar discoveries are made for the other marker pairs, such as the high expression of CD161 and CCR7 in CD4 cells in colon. In addition, we identified a cell subset CD27$^+$CCR5$^+$ that is high in CD4 cells from lung samples. This cell subset also expresses CD29, CXCR4, and CD45RO. Again, because these lung tissues are collected from lung cancer resection, it is impossible to tell whether this cell subset is unique to CD4 cells in lung or is induced by tumor as CCR5 is related to cancer therapy.

Combining the markers selected by CytoBinning, we were able to recapitulate the uneven distribution of helper T cells in different human tissues with only 8 markers instead of all 41 markers as was done in the original publication and we achieved better tissue separation with the selected 8 markers. Also, we also illustrated the diversity of memory cells across different human tissues. What's more, these analysis results were achieved without any prior biological background and hence is unbiased. In all, our simple method is able to obtain comparable analysis results as the more sophisticated high-dimensional analysis method used in the original paper, and also because the straightforwardness of our method, we are able to provide easy to interpret and test results (such as the PD-1^{high}CCR7$^-$ subset that is unique to tonsil samples).

We have shown that CytoBinning is able to detect subtle changes in patterns. However since it is designed to detect differences in overall patterns, it is indifferent to the position and scale of patterns. Hence, for example, if one sample contains only PD-1$^+$CCR7$^+$ cells and the other contains only PD-1$^-$CCR7$^-$ cells, and they both form the same pattern (e.g. a single cluster), CytoBinning is not able to detect that. Nevertheless, if there exist small differences (for example a tail or thick outliers, etc.) CytoBinning will be able to identify them. In addition, we argue that if two markers form the same pattern (e.g. one large cluster) in different patient samples with very different expression level the difference will show in patterns when these markers are pairing with other markers and can then be identified with CytoBinning.

All in all, we defined informative features in mass cytometry dataset as point patter in a 2D scatterplot and proposed a binning method – CytoBinning – to quantify contents in scatterplots. We demonstrated with a publicly available dataset that our method is able to recapitulate, without any prior biological background, analysis results obtained with more sophisticated high-dimensional analysis guided by expertise in immunology. In addition, our method has the advantage of easy interpretation and the ability to generate testable hypotheses. With moderate modification, our method can be applied to other comparative studies (two or multiple groups), or serve as a fast screening of datasets to quickly identify driving differences without diving into more complicated and time-consuming algorithms.

4.6 Methods

4.6.1 Experimental procedures and pre-processing

The dataset used for the demonstration was first published in [134] and downloaded from flow repository website (https://flowrepository.org/) [140]. Tissue types, number of samples, and the reason for surgery is listed in Table 1. Immune cells were isolated from collected tissues and cryopreserved. They were then thawed and washed for mass cytometry experiment. Two panels of antibodies were used for staining, each containing 41 markers. The two panels were named as “Function” and “Traffic” according to the antibodies included in it. Details of experimental process and the lists of antibodies can be found in [134]. The downloaded samples from flow repository are FACS files, pre-gated to major immune types (e.g. CD4, CD8, NKT, etc.). We performed logicle transformation using logicleTransform in R package flowCore, with parameters w = 0.25, t= 16409, m

=4.5, and a=0 according to [134]. The logicle transformed data were then saved as text files for further analysis.

Table 4-1 Summary of sample information

Tissue type	Number of samples	Reason for surgery
Cord Blood	5	Healthy donation at neonate
PBMC	4	Healthy donation
Tonsil	5	Tonsillar Hypertrophy
Spleen	3	Splenectomy (Due to Distal Pancreatectomy)
Colon	6	Routine Colonoscopy
Skin	5	Abdominoplasty or Mastectomy - Invasive Ductal carcinoma
Lung	4	Lung cancer resection
Liver	3	Liver transplantation

Markers in traffic panel:

CD45, CD14, CD57, TCRγδ, CD3, HLA-DR, CD29, CD38, CD69, CD62L, CD8, CD45RO, CLA, CD4, CD103, CCR4, CD25, CD49a, CCR10, CXCR6, CD19, CD27, CD56, ICOS, PD-1, CD161, CCR9, CXCR3, CD95, CD31, CXCR5, CD49d, CCR2, Intergrinβ7, CCR5, CCR6, CD45RA, CCR7, CX3CR1, CXCR4, CD127

Markers in function panel:

CD45, CD14, CD57, TCRγδ, IFNγ, TNFα, IL-8, Granzyme B, IL-17F, CD45RA, CLA, CTLA.4, IL-2, CD25, CD103, CCR10, CXCR6, IL-5, CD19, CD56, IntegrinB7, PD-1, IL-9, CCR9, CXCR3, CD127, Mip1b, CXCR5, CD161, CCR2, IL-4, IL-10, CCR6, GM-CSF, CCR4, IL-22, CCR5, IL-17A

4.6.2 CytoBinning

In this paper, we used CytoBinning with 5 bins. For a given marker pair A & B, equally populated bins are drawn based on single cell expression of marker A and marker B independently. These bins are then overlaid on each other so that a grid is formed with 25 (5^2) regions (boxes). Percentage of cells inside each box is then an estimation of the joint probability $P(A_i, B_j)$, where i and j are the corresponding bins this box locates at ($1 \leq i, j \leq 5$). For a random distribution where marker A and marker B is not correlated in any way, $P(A_i, B_j)$ should be approximately the same in for every box. This is not true if marker A and marker B is related in any way (i.e., their mutual information is not zero, this relationship can be both linear and nonlinear). We use all $P(A_i, B_j)$s as a coarse-grained representation of the point pattern between single cell expression of marker A and marker B. This binning process is repeated for every possible marker pairs in the dataset.

4.6.3 Marker pair selection

For each marker pair, multi-class SVM is applied to the binning results in R. We used svm function in package e1071 with linear kernel, cost = 1 and gamma = 0.01. We randomly divided the 35 samples into 7 groups and performed 7-fold cross-validation based on the dividing. The error rate of the classification was calculated as the average cross-validation error rate. Marker pairs with error rate < 0.2 (less than 1 error in average) were selected as important marker pairs.

4.6.4 t-SNE plots

After marker pair selection, we picked out unique markers within these pairs and performed tsne in R with tsne function in package tsne. We randomly selected 250 cells from each sample and concatenated them together before performing tsne (the number of cells for each tissue ranges from 750 for spleen and liver to 1,500 for colon).

Chapter 5 Conclusion and Discussion

5.1 Overview

Advancements in experimental technologies have changed the landscape of research in biology. As the data becoming increasingly high content and throughput, traditional manual-based analysis regime has passed, and computer-based automated analysis methods that are able to handle high-dimensional, single-cell data are on the rise. However, these methods often use all measurements simultaneously without weighting. This can result in problems such as misinterpretation of datasets due to noisy and irrelevant measurements, and less meaningful results due to the "curse of dimensionality". In this dissertation, my goal is to develop analysis methods that are complex and multi-dimensional to automatically extract meaningful information out of biological datasets, but not so high-dimensional and complicated so that straightforward interpretation can still be achieved.

5.1 Conclusion

5.1.1 Identify stable classification boundaries based on typical cells

In Chapter 2, I presented a pipeline for analyzing high-throughput imaging data that tackles cell-to-cell variability in a mono-state cell population by capturing the typical features of cells. I demonstrated this approach using a screening dataset that aims to identify potential drug (siRNA) hits for a premature ageing disease (progeria) [109]. The pipeline starts with the selection of typical control cells (which I defined as cells close to the mean of a single peak multi-variate distribution, see section in Appendix A4), followed by the computation of a stable classification boundary between two reference conditions (e.g. healthy and diseased control, Fig 2.2) at the opposite ends of a continuous spectrum of cellular phenotypes. Each cell in all siRNA perturbations is then compared to this classification boundary. Cells on the same side with typical healthy control cells were considered as “healthy-like”, and the fraction of “healthy-like” cells was used as a new metric for analysis of the screen. Moreover, potential drug hits were identified as siRNA perturbations that significantly increase the fraction of healthy-like cells in the population comparing to diseased controls (Fig 2.3). The classification boundary built with typical cells captured the direction where progeria cells are most significantly different from healthy cells. The underlying biological assumption is that – by muting one of the ubiquitin proteins, progeria cells would become healthier or worse (due to the loss of house-keeping protein). However, siRNA perturbation can also drive cells to a state that is completely different from healthy or progeria controls. I tested that by calculating the position of each cell in the direction parallel to the classification boundary. Moreover, I found that siRNA perturbed samples fell nicely into a line along the direction of classification boundary (Fig 2.5).

5.1.2 CytoBinning as a new method to identify meaningful differences in 2D point patterns

In Chapter 3, I introduced a binning method – CytoBinning – that can quantify information in 2D point patterns (Fig 3.1). The binning was done based on density distribution of single measurements. The bins of two measurements were then combined to get a grid that maps the point pattern formed by these two measurements. Percentage of cells inside each box of the grid was then calculated. Since these percentages change with the point pattern, they can be used as a numeric presentation of the pattern. I showed with synthetic datasets that CytoBinning can help identify both significant (Fig 3.2) and subtle differences (Fig B3) of point patterns, and the same can be achieved in real biological datasets (Fig 4.1). For classification problem, the appropriate number of bins to use is dependent on the number of samples to be classified. If the two groups contain an equal number of samples, a good estimation of the bin number can be obtained by this formula: $maximum\ number\ of\ bins = round(\sqrt{number\ of\ training\ samples/2})$.

In addition, I used a publicly available dataset that aims to find the difference of immune signature in peripheral blood mononuclear cells (PBMC) between old and young healthy donors [59] to demonstrate how to use CytoBinning for comparative studies in Chapter 3. I showed that CytoBinning was able to re-identify the decrease of naïve CD8 T cells in elderlies (Fig 3.4). In addition, CytoBinning found a new cell sub-region whose cell frequency increases with ageing. This sub-region resides in the valley between two well-defined peaks – naïve CD8 T cells and memory CD8 T cells – which may be a transition cell state from naïve to memory.

5.1.3 CytoBinning can be used for feature selection in high-dimensional data

In Chapter 4, I demonstrated how to use CytoBinning for feature selection with a publicly available mass cytometry dataset that measures immune cells in eight types of human tissues [134]. The original publication managed to clarify tissue-specific features of immune cell composition with all 41 markers that were measured in the experiment. Since the expression level of individual markers is not meaningful in cytometry data, I defined features of mass cytometry data as the point patterns formed by two markers. With CytoBinning, I was able to select 5 important features (i.e., 5 marker pairs) that provide tissue-specific information of immune cell composition. There are only 6 unique markers in the 5 marker pairs, and I showed that with only the 6 markers tissue-specific features as identified in the original publication could be re-identified. In addition, I found that cells form distinct patterns among the eight types of tissues for each of the marker pair identified by CytoBinning (Fig 4.2, Fig C4, Fig C7, Fig C10-11), and cell phenotypes that are unique to or overexpressed in particular tissue types can be found. For example, a cell phenotype – PD-1^{high} CCR7^{low} – was found to be unique to immune cells in tonsil tissues (Fig C2).

5.2 Discussion

5.2.1 Heterogeneity: important feature or byproduct of complex systems?

We know that single cells are heterogeneous, not two cells are the same. As the measurement of more and more single cell features are made available by experiments, cellular heterogeneity becomes more and more apparent. However this is a paradox since random events are prevalent, differences are guaranteed to be found when examined closely enough. Hence there has been the question: “Do differences make a difference?” [9]. Moreover, it has been suggested that only heterogeneity that has biologically functional differences should be considered meaningful. Nevertheless, biologically functional differences are abundant as well. Since living organisms deal with environmental stress constantly, cells with the same genomic and epigenetic features could develop different features in response to slightly different environmental cues. Understanding heterogeneity is important, it helps us to get deeper insights into how living systems respond to its surroundings by regulating itself. For example, our immune cells are famously heterogeneous. There are at least tens of immune phenotypes that have been validated experimentally, and they each have distinct functions – e.g., Th1 cells are in charge of defense over intracellular bacteria, Th2 cells help fight against the extracellular parasite, Th17 cells are responsible for defense against fungi infection. Moreover, all these subsets can be involved in autoimmunity. In addition, our immune system protects us from various antigens by being able to produce antigen-specific cells. Each antigen attack would leave a trace in our immune system in the form of “memory cells”, and as we age our immune system collects a history of antigen attacks in our body, hence no two people have the same repertoire of immune cell types. The immune system is a great example of how heterogeneous single cells can be, and how the heterogeneity can result from a dynamical response to the environment. Moreover, a better understanding of heterogeneity in immune cells can help us understand how our immune system work and how to restore it when it is not working properly.

On the other hand, even though cellular heterogeneity can be meaningful and ubiquitous, sometimes it is the similarity shared by cells in a single phenotype that’s more important, especially in cultured cell lines where macroscopic and genetic variables are carefully controlled. One argument for the importance of similarity is that there are countless reasons for the cells to be heterogeneous that have nothing to do with the biological process of interest (e.g., fluctuation in the local distribution of chemicals, local cell density, etc.). Since it is nearly impossible to account for all possible reasons, focusing on heterogeneity could result in misleading results. Hence the typical behavior that can be observed in the majority of cells in response to the targeted perturbation should be of real importance. For example, in RNAi perturbation studies (Chapter 2), researchers often add RNAi to a group of cells that belong to the same phenotype (cultured from the same progenitor in the same macroscopic condition) and studies how these cells respond to the perturbation. That is where the notion of “typical cells” comes in. The definition of “typical cells” (i.e. typical behaviors) enables the characterization of typical behaviors shared by the majority of cells, which should be a more direct representation of cellular response. In addition, typical cells can help build a more stable classification boundary between two cell groups.

5.2.2 High dimensional data: divide and conquer

Biological research today produce high-content, high-throughput data in large volume. One particular challenge in analyzing these datasets is the high-dimensionality. To make full use of the information-rich data, biological society has been embracing high-dimensionality with more and more algorithms that can handle all the parameters simultaneously have been developed. However, fully embracing high-dimensionality can be problematic. On the one hand, the definition of distance and density is not meaningful in high-dimensional space which could lead to inaccurate results for distance/density based algorithms [11]. In addition, cell states may be a continuum instead of distinct stages and cells that are in between two high-density regions (i.e., the clusters found by most clustering algorithms) also have potential biological significance [76]. On the other hand, information overload is prevalent as much more parameters are measured experimentally than the intrinsic dimension of the question of interest. Dimension reduction and feature selection methods have been recruited to tackle this problem. However, some of these methods make strong assumptions about the datasets (e.g., PCA assumes linearity of the data) that are not always satisfied; and almost all methods aim to find one particular subspace that best captures the content of the whole dataset which will possibly undermine subtle signals. Dimension reduction methods often mix all parameters in finding optimal subspace, and the resulting dimension reduced space is a combination of all parameters. This is a problem for biological data analysis because the explicit interpretation of resulting features which is lost in parameter combination is crucial to the understanding of biological questions. My approach to tackling this problem is to divide high dimensional space into not one but multiple lower dimensional subspaces. These subspaces add up to the whole original space and may overlap so that there’s no missing information. For example in

Chapter 2, instead of combining all available measurements, I divided them into four subspaces based on the fluorescent staining. In so doing, the minor signal in DNA damage (γH2AX) could be kept from being overshadowed by the strong signal in progerin expression. Another commonly used method to deal with information overload in high dimensional data is feature selection. Instead of finding a subspace by combining features, feature selection methods select meaningful features separately. However, for some datasets (like cytometry data), instead of individual parameters, the correlation between two parameters is where the true information lies. Feature selection methods then rely on correlation measurements to assess the importance of parameters. However, to-date the state-of-art measurement of the correlation between two parameters are still correlation coefficients and mutual information which give only a gloss estimate of how much the two parameters are related instead of how the two parameters are related. CytoBinning fills the gap by providing a more detailed measurement that can quantify how two parameters are related (Fig 3.1, Fig 4.1). In addition, by dividing high-dimensional datasets into 2D facets and considering them one by one, it is possible to select all meaningful signals, both significant and subtle (as shown in Chapter 3). Also as CytoBinning only mix two parameters at a time, interpretation of its results is more straightforward comparing to the more sophisticated and complex algorithms. The usage of CytoBinning is not limited to cytometry data, in fact, it can be applied to any single cell data where the correlation between two parameters is of interest (e.g., single cell RNA sequencing, or even point patterns in images).

5.2.3 Next step: Application to clinical datasets

CytoBinning could be applied to large clinical cytometry datasets since it is computationally efficient and parallelizable. The overall goal is to further validate the method and mine novel biological insights based on these datasets. There are two large clinical flow cytometry datasets that are made available to our lab. The first clinical trial attempts to generate hypotheses of immune mechanisms in an inflammatory eye disease (uveitis). With the help of CytoBinning, my collaborators hope to identify novel cell subtypes that play a role in the development of this disease. The second dataset contains blood samples from over 1,000 healthy donors, and the goal is to use CytoBinning to map the trace of ageing in peripheral immune cells [143].

5.2.4 Generalization: CytoBinning as a tool for point pattern analysis

In addition to analyzing cytometry data, CytoBinning can potentially be used as a general method to quantify how two parameters are correlated. Comparing to conventional measurements, CytoBinning outputs a vector that contains information about the patterns of point distribution, whereas correlation coefficients and mutual information (MI) only gives one number indicating the degree of relationship between two parameters. As a result, CytoBinning can be used to better estimate the similarity between two relations (e.g. similarity between the expression pattern of gene A and gene B and the expression pattern of gene A and gene C). This is important because two parameters can form different point patterns but still have the same mutual information (Fig 5.1). Moreover, in examples like reconstructing gene networks, two genes are considered connected/ similar if they are both highly correlated with the same gene. However, this should not be the case if the correlation patterns are different. Since neither correlation coefficients nor MI provides information about the pattern, the distinction cannot be made using these measurements alone.

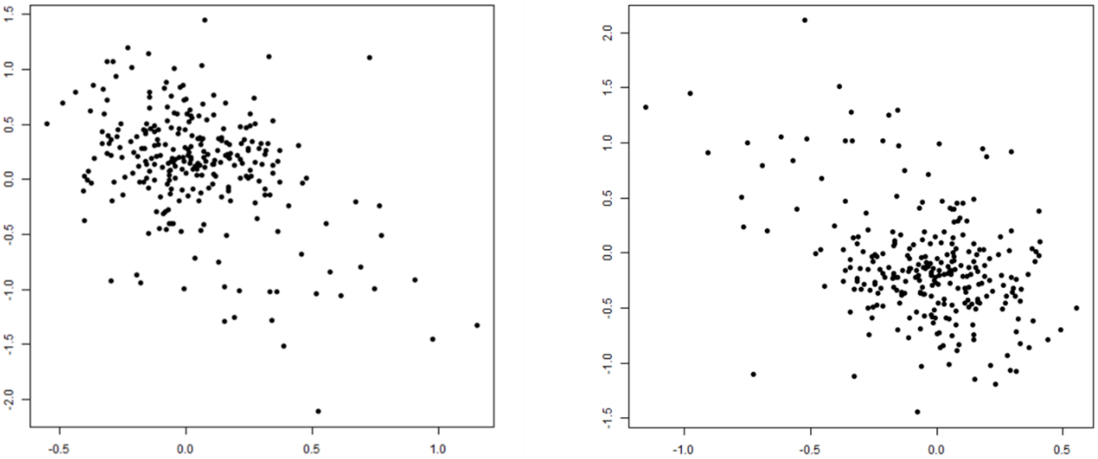


Figure 5.1 Two different point patterns with the same value of mutual information (0.23).

To illustrate how CytoBinning can be used to distinguish point patterns, I generated eight synthetic patterns (Fig 5.2, ten samples/ replicates for each pattern, 1,000 points in each sample) and applied CytoBinning to cluster them. The results show that CytoBinning is able to correctly cluster five out of the eight patterns, but the other three patterns (circle, random, two circles) are mixed (Fig 5.3).

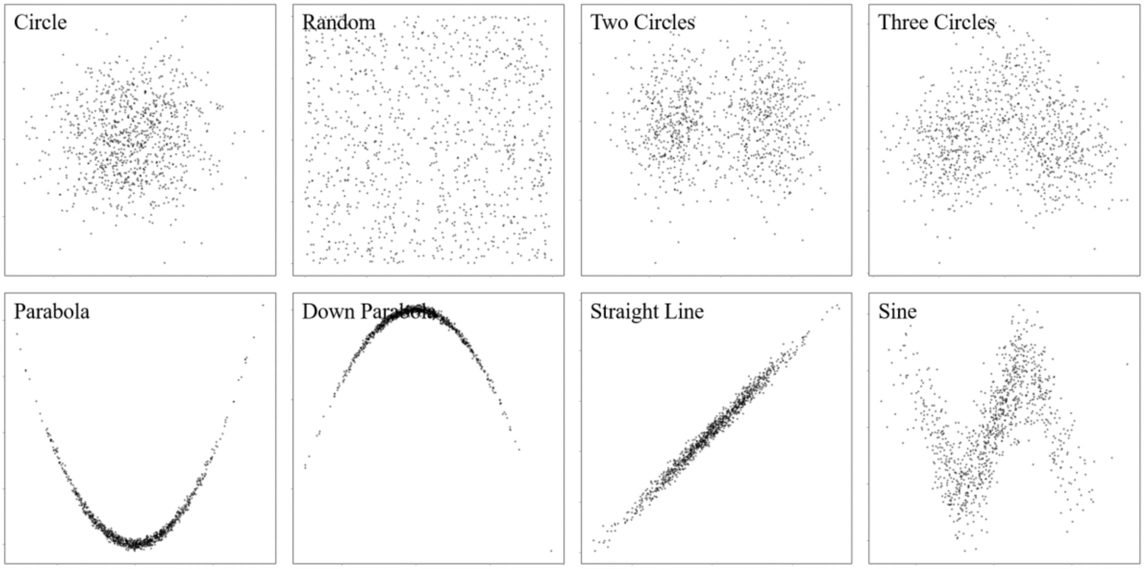


Figure 5.2 Examples of the eight simulated patterns. Circle: x, y-axes are both randomly sampled from distribution *N*(0, 2). Random: x, y-axes are randomly sampled from a uniform distribution ranging from 0 to 2. Two Circles: x is sampled from two normal distributions *N*(-2, 1) and *N*(2, 1) each contains 500 points,

and y is sampled from $N(0, 1)$. Three Circles: the bottom two clusters (400 points in each cluster) are sampled from a normal distribution centered around (-2, 0), (2, 0) respectively with standard deviation equals 1. And the top cluster is sampled from a normal distribution centered (0, 1) with standard deviation 1. Parabola: x values are sampled from $N(0, 2)$, y = x^2 + noise where noise is sampled from $N(0, 0.5)$. Down Parabola: the same as parabola, except y = -x^2 + noise. Straight Line: $x \in N(0, 2)$, y = x + 1 + noise, $noise \in N(0, 0.25)$. Sine: $x \in N(0, 2)$, y = sin(x) + noise, $noise \in N(0, 0.5)$

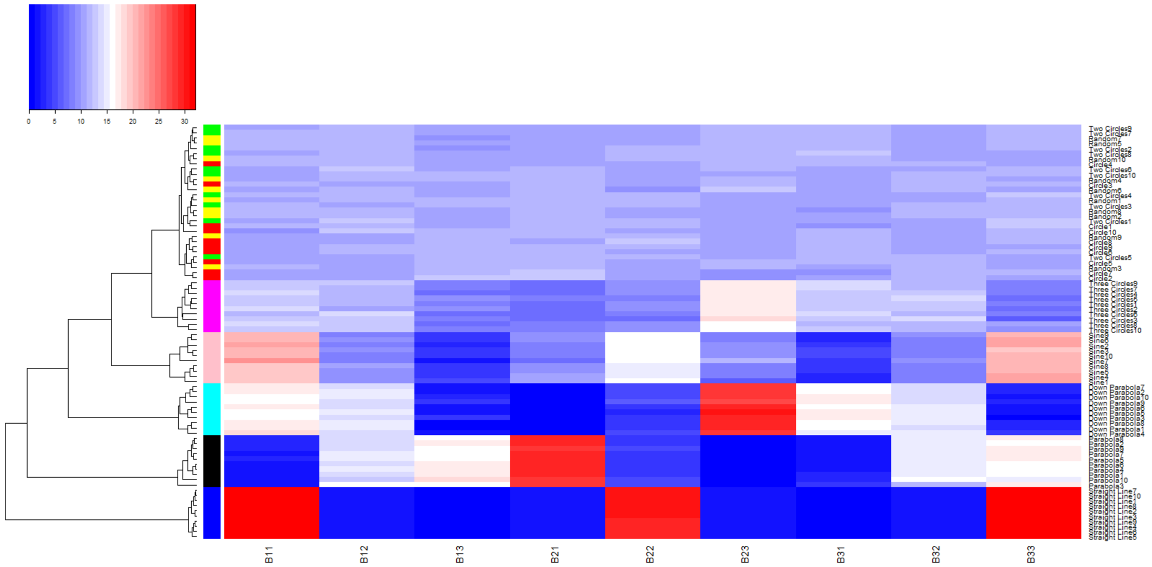


Figure 5.3 Clustering results of the eight patterns based on CytoBinning results (3 bins). Patterns are labeled by different colors (circle: red, random: yellow, two circles: green, three circles: magenta, parabola: black, down parabola: cyan, straight line: blue, and sine: cyan).

The possible reason CytoBinning fails to distinguish "circle", "random" and "two circles" could be that these three patterns are topologically similar (i.e., they are all symmetric to both axes). To test this hypothesis, I created subtle asymmetric features in these patterns (Fig 5.4. A small cluster is added to the upper right region of Circle pattern, and lower left region of the Random pattern. And the right cluster in Two Circles pattern is shifted upwards. These changes should break similarity shared by the three types of patterns.), and clustered the patterns with added asymmetry. CytoBinning is able to differentiate these three patterns now the symmetry is broken (Fig 5.5). As a comparison, if instead of adding a small cluster in the upper right region of Circle pattern, the small cluster was added to the lower left region (similar to what I did to Random pattern, Fig 5.6), then Circle pattern and Random pattern still hold topological similarity and cannot be distinguished by CytoBinning (Fig 5.7). Hence we infer that it is the intrinsic similarities between patterns instead of symmetry of pattern itself that impedes correct clustering of pattern “Circle”, “Random”, and “Tow Circles”. In other words, CytoBinning would fail to recognize different point patterns only when they share global similarities.

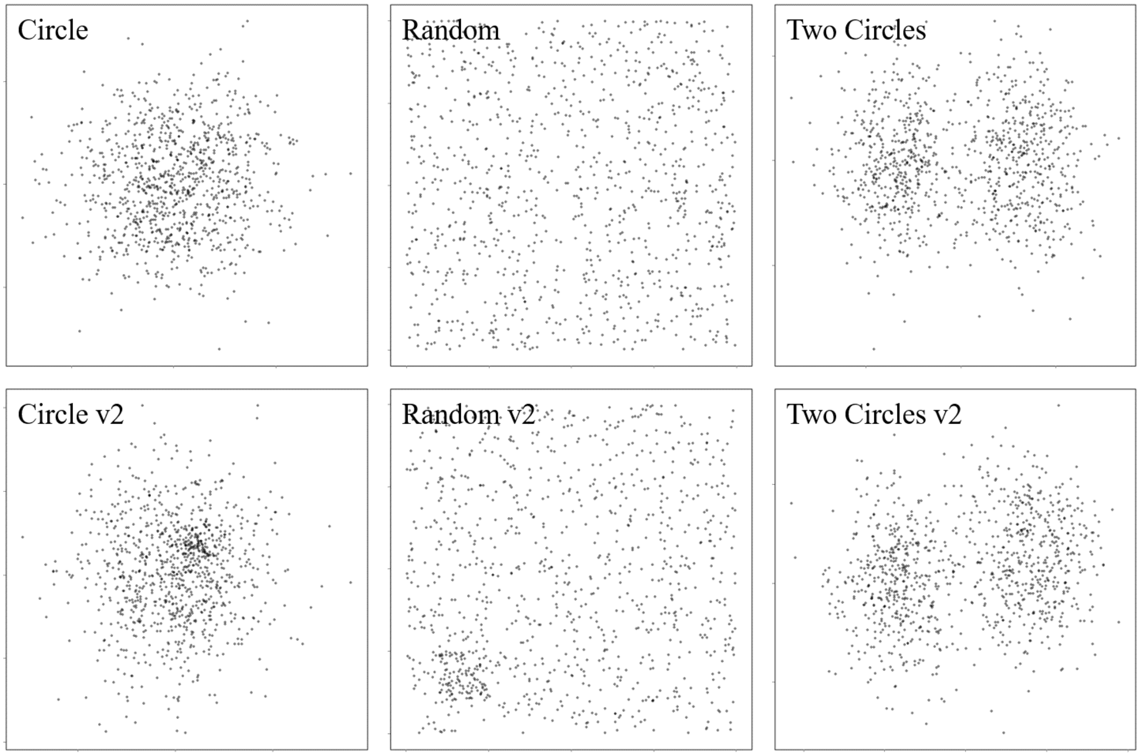


Figure 5.4 Examples of patterns with added features (lower row) comparing to original patterns (upper row). Circle v2: 1,000 points in total, for 900 points $x, y \in N(0, 2)$. And for the rest 100 points $x, y \in N(1, 0.5)$,. Random v2: 1,000 points in total, for 900 points, $x, y \in uniform(0, 2)$; and for the rest 100 points $x, y \in uniform(0.2, 0.5)$. Tow Circles v2: same as Two Circles, except the center of the right cluster is shifted up by 0.5.

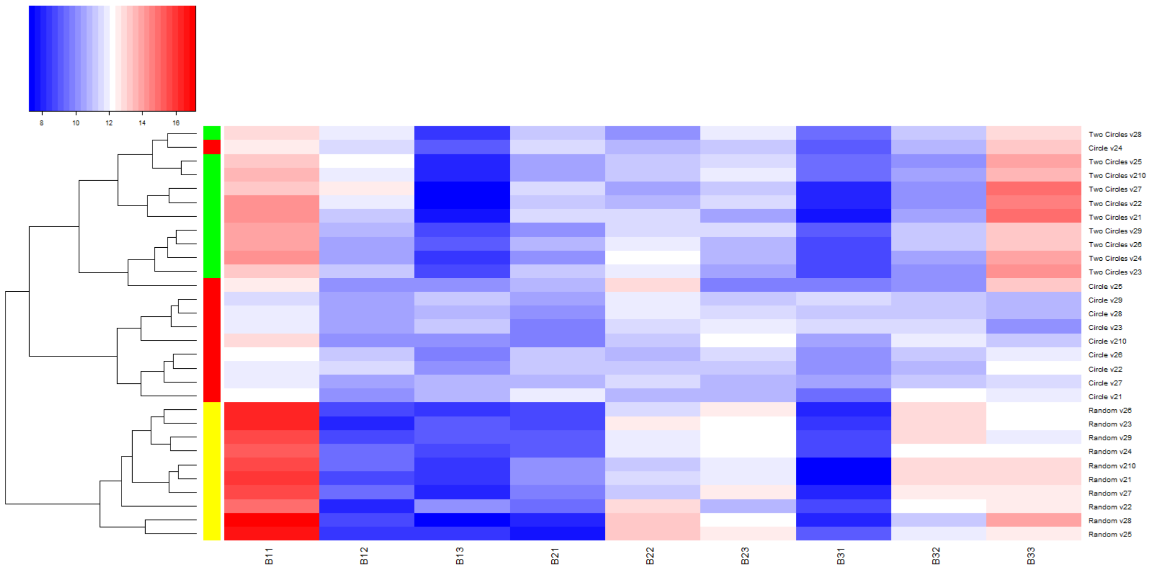


Figure 5.5 Clustering results of patterns with added asymmetrical features. The three patterns are largely separated (Circle: red, Two Circles: green, Random: yellow) with only one mistake.

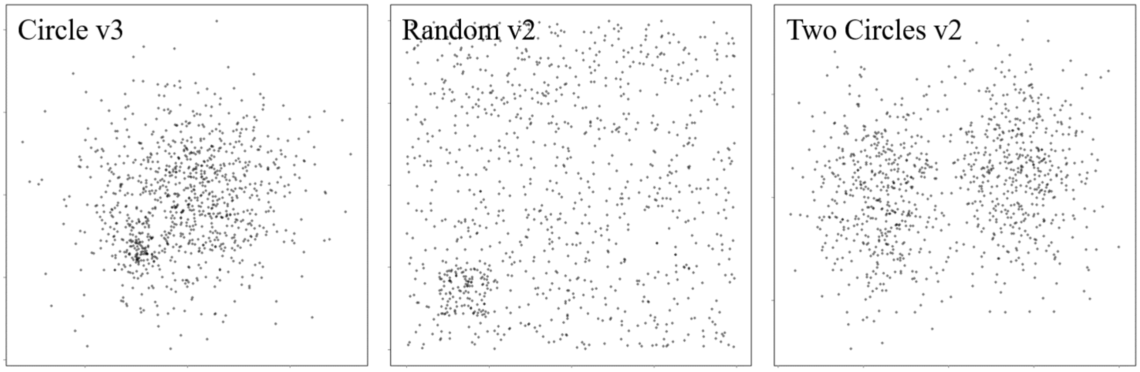


Figure 5.6 Example of point patterns with added features. Random v2 and Two Circles v2 are generated in the same way described in Fig 4. For Circle v3, similar to Circle v2, except the small cluster is now centered around (-2, -2) with standard deviation 0.5 (lower left region).

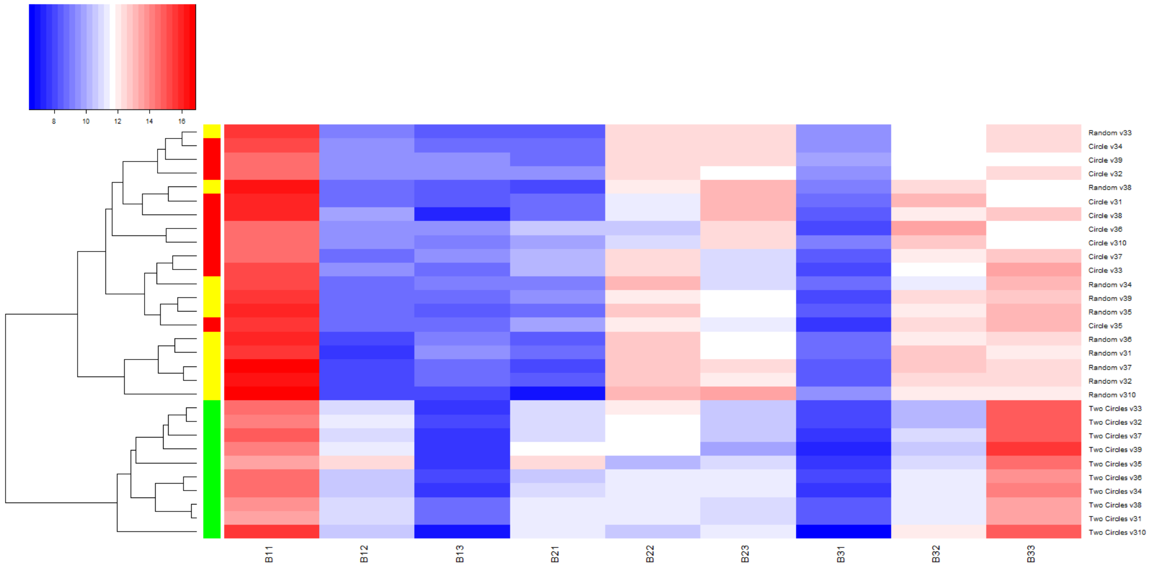


Figure 5.7 Clustering results with the feature added in Circle pattern moved to lower left region (Circle: red, Random: yellow, Two Circles: green).

One additional advantage of CytoBinning is that it is invariant under order-preserving transformations of parameters. For example, cytometry data always requires a careful transformation step before data analysis because the expression of surface markers tends to have large ranges. Figure 5.8a and Figure 5.8b show the same scatterplot before and after logicle transformation. The patterns look dramatically different, and algorithms based on the exact value of parameters would be completely confused. However, CytoBinning is able to give the same analysis results for both scatterplots (Fig 5.8c and 5.8d), showing that CytoBinning does not depend on the exact value of each parameter to measure the patterns, only relative position of each point.

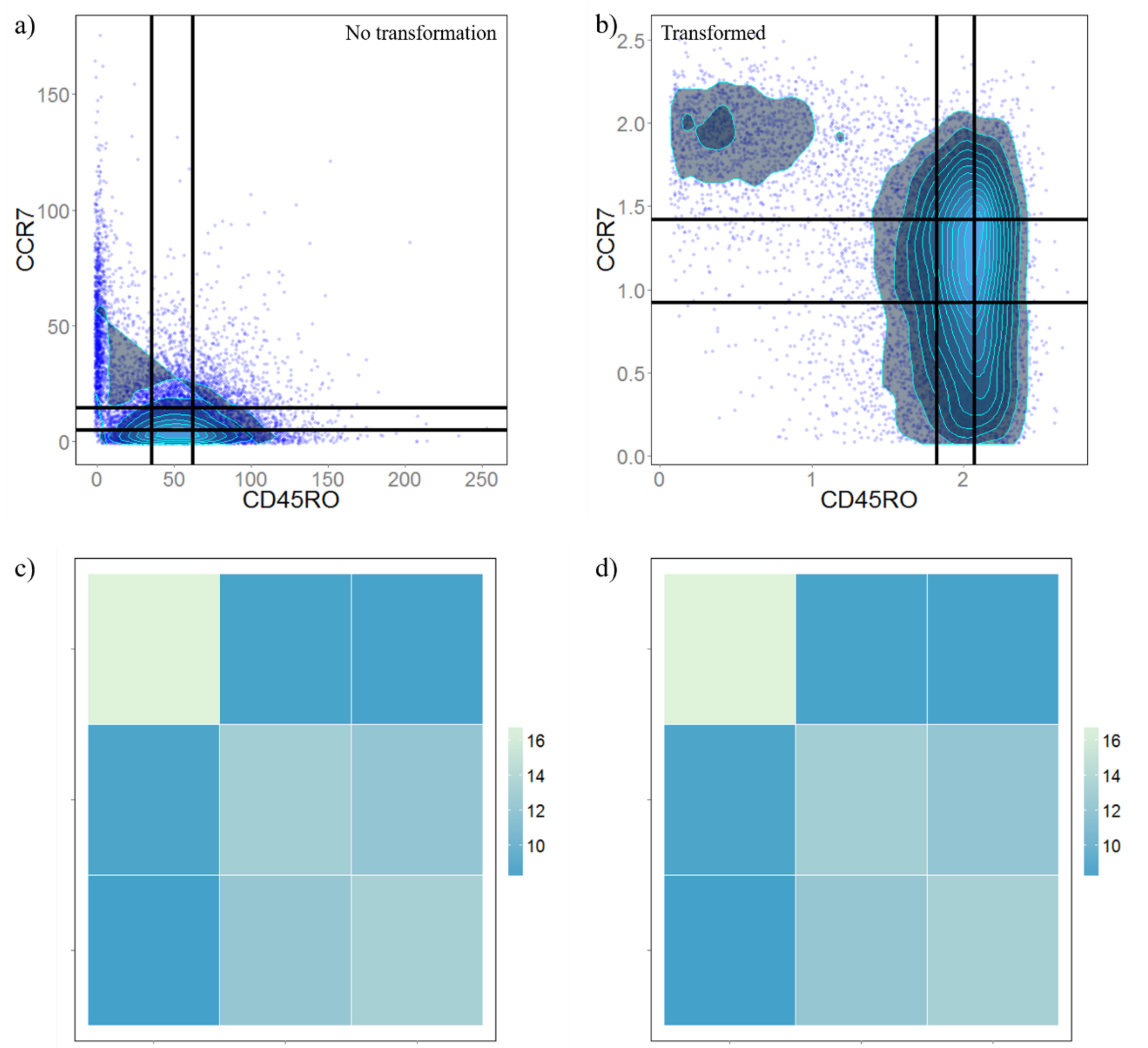


Figure 5.8 The results of CytoBinning is invariant under order-preserve transformations. a) and b) scatterplot of the expression level between CCR7 and CD45RO of the SAME sample (taken from the dataset used in Chapter 4). c) and d) heatmap showing CytoBinning results of the corresponding point patterns.

Thanks to this invariant feature, CytoBinning can be applied to not only data analysis but also image analysis. For example, in neural imaging experiments, researchers often record neurons (points) that fire (light up) together under certain stimulations. Clarify the patterns of neurons firing together help the understanding of how neural circuits are connected to each other. CytoBinning can help with this problem by providing quantifiable measurements of these point patterns.

Appendix A Supplementary Information for Chapter 2

Typical healthy and progeria cell nuclei

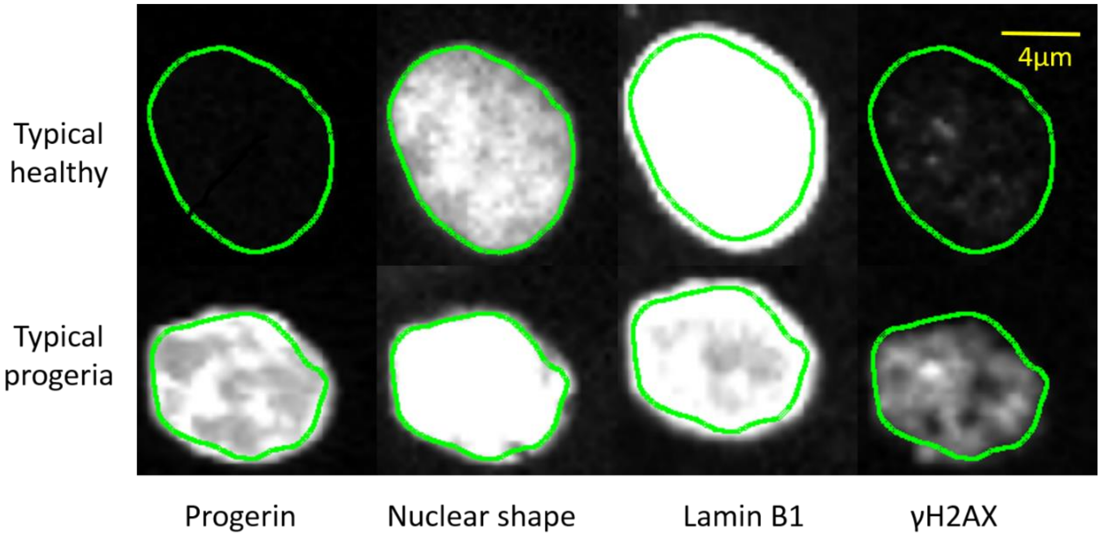


Figure A1. Representative images of typical healthy (top row) and progeria (bottom row) nuclei. Images are shown in four different channels: progerin, DAPI, lamin B1, and γH2AX. The outline of each nucleus (shown in green) was first extracted from the DAPI channel (nuclear shape) and then mapped onto the other three channels. As the images show, typical progeria nuclei have pronounced progerin expression, blebbed nuclear outlines, decreased lamin B1 expression, and high levels of DNA damage (γH2AX).

A.1 Sorting out the mismatch of images across different channels

In Figure 2.1 (a), Figure 2.6 and Figure A1, we noticed that nucleus in Lamin B1 channel tends to be larger than its counterpart in DAPI channel. In addition, there is also a shift in the image when we tried to overlay the outline segmented from DAPI channel directly onto Lamin B1 channel. Since our intensity measurements (measurements for Lamin B1, Progerin and γH2AX channel) are based on the outlines segmented in DAPI channel (we did this because cells used in our experiment are perturbed in lamin B1, progerin, and γH2AX expression, causing them not reliable for outline segmentation), these two problems can lead to serious mistake. In order to sort out these two issues, we compared measurements between their values calculated based on outline segmented from DAPI channel and outline segmented in lamin B1 channel. A random sample of nuclei in GFP-progerin repressed control are used for this comparison, and four measurements (square root of area, boundary point intensity, mean intensity and standard deviation of intensity) are compared. Plotted in Figure A2 are the results. Each dot in Figure A2 represents a nucleus. As shown in Figure A2 (a), the size of nucleus measured in Lamin B1 channel is constantly larger than DAPI channel, with the average nuclear radius calculated in Lamin B1 channel ~ 600 nm longer than in DAPI channel. This difference may be due to the fact that DAPI attaches to DNA while Lamin B1 stain directly attaches to lamina that supports nuclear membrane. As to the shifting, since different cameras are used to capture images in different channels, even though the alignment of cameras was auto-corrected, there can still be slight shifts of direction. As shown in Figure A2 (b) – (c), the three intensity measurements are highly correlated between measurements calculated based on DAPI outline or Lamin B1 outline with correlation coefficients almost 1 (p-value close to 0) for mean intensity and boundary point intensity. These results suggest the shift in direction is small and intensity measurement based on DAPI outline is a reliable replacement for direct measurement based on outlines segmented in lamin B1 channel. Since Lamin B1 (as well as progerin and γH2AX) is perturbed as its expression level decreases when progerin exists leading some nuclei invisible in lamin B1 channel, we performed lamin B1 (as well as progerin and γH2AX) measurements based on DAPI outline.

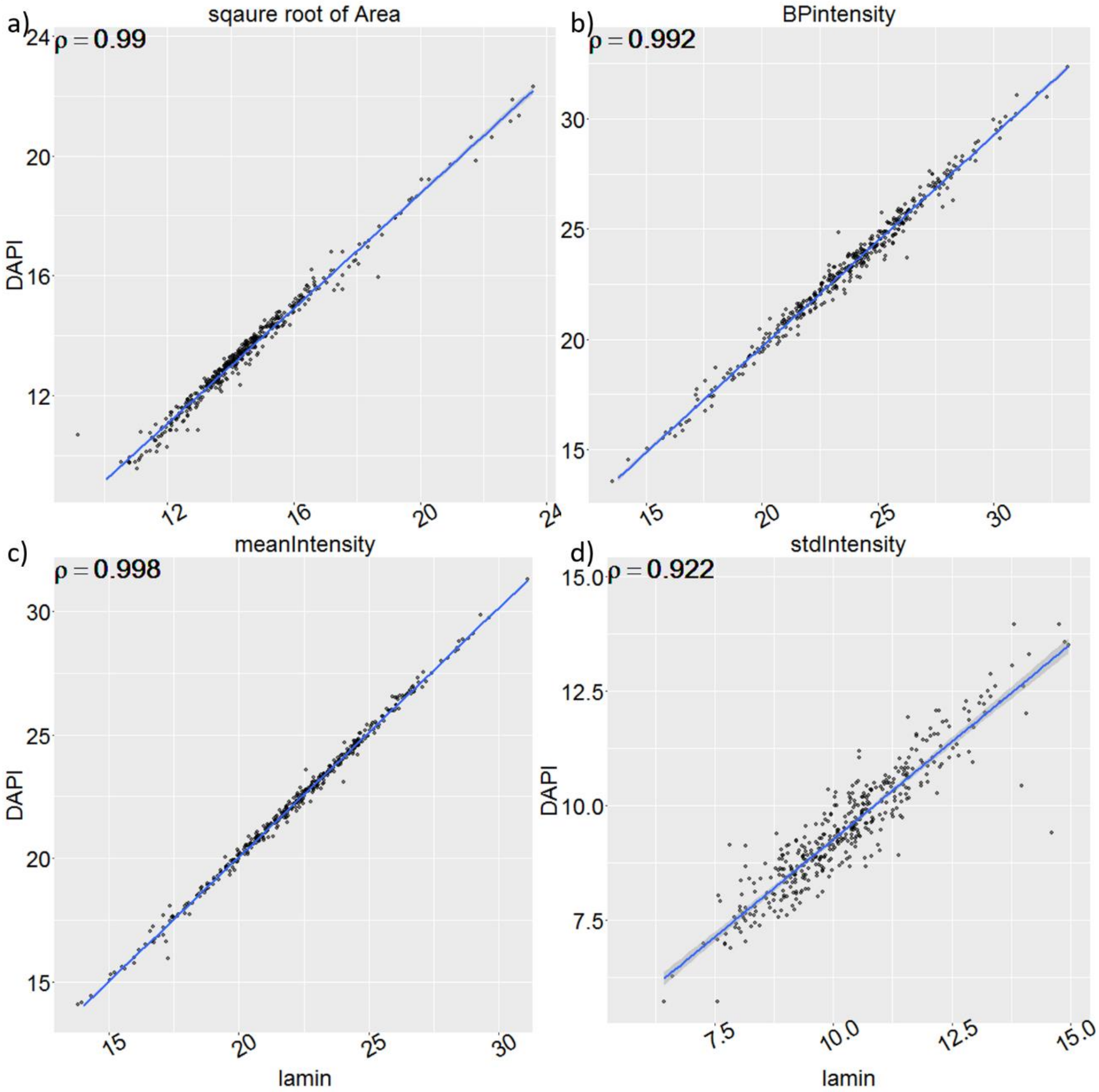


Figure A2. Comparison between measurements obtained from outlines segmented in DAPI vs. Lamin B1 channels. Corresponding correlation coefficients are shown in the top left corner of each panel. BPintensity stands for boundary point intensity, which measures mean fluorescent intensity along the nuclear boundary; meanIntensity stands for mean intensity inside nucleus, and stdIntensity is the standard deviation of fluorescent intensity inside nucleus boundary.

A.2 Image processing and feature selection

The raw data for image-based HTS consists of fluorescence microscopy images. We studied cell nuclei imaged in four channels (see Fig A1). From the DAPI channel, we first analyzed nuclear shapes extracted with an active contour algorithm [123], from which we determined 12 shape metrics. Out of these, 5 measurements are *global metrics*, i.e. parameters that describe the overall shape of the boundary (area, perimeter, eccentricity, major and minor axis length), while the remaining 7 measurements are *local metrics*, sensitive to the local features of the shape (number of invaginations, standard deviation of the curvature, mean curvature, solidity, mean negative curvature, circularity, and tortuosity). For analysis, an online open source package (http://downloads.openmicroscopy.org/bio-formats/5.1.2/) was incorporated into the nuclear shape extraction algorithm [123] to read flex format images. Furthermore, to detect and remove overlapping nuclei from multiple cells, we set up outlier detection thresholds for area and solidity at two standard deviations from the mean and discarded segmented nuclear outlines with area or solidity beyond the thresholds (about 6% of nuclei were discarded at this stage). The nuclear outlines extracted in the DAPI channel were mapped to the other 3 channels for analysis of fluorescence intensity, as shown by the green boundaries in Figure A1. In each of the 3 other channels, we determined 3 metrics as basic characteristics of the intensity distribution in each nucleus: mean intensity, the standard deviation of intensity, and mean intensity along the boundary. Therefore, in total, we obtained 21 metrics for each nucleus, of which 12 represent shape features and 9 represent 3 metrics for each of the 3 channels label lamin B1, progerin, and DNA damage (γH2AX).

As a starting point, to select meaningful metrics that differ between GFP-progerin expressing and repressed controls, we analyzed each metric separately using F-scores [144], defined as

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \qquad (1)$$

where $x_i^{(+/-)}$ is the mean of the *i*th measurement for GFP-progerin expressing and repressed controls, respectively; $\bar{x}_i$ is the mean of the *i*th measurement for both controls combined, and $n_{+/-}$ is the number of cells in GPF-progerin expressing and repressed controls. Using this procedure, we removed 6 shape metrics (eccentricity, minor axis length, major axis length, mean curvature, area, and perimeter), which yielded very low F-scores (<0.003), from further analysis. The F-scores for the included shape metrics ranged from 0.5 to 17.

A.3 Determination of parameters used in the selection of typical control cells

Typical cells for control cell type A are defined as cells that are closest to the center of the multi-dimensional distribution of all cells in type A. Hence, there are two parameters to be determined before typical cell selection: the definition of center, and the number of typical cells to be selected. In this paper, we tested 3 definitions of center: the mean, the median, and the global peak of each distribution. The peak of distribution was calculated using .find_peaks() function in R package openCyto [68]. We also tested 8 different numbers of typical cells: 100, 200, 300, 500, 1000, 2000, 3000, and 5000. 1500 cells that are next closest to the center (compare to typical cells) were selected as cross-validation set (CV). The distance between each cell and the center of distribution was calculated using L1 Manhattan distance. Typical cells were selected independently in each channel, thus different cells may be selected as typical cells in different channels. We applied support vector machine (SVM) to classify typical healthy and typical progeria cells selected using all cells numbers and corresponding to all 3 definitions of center (in total 3 x 8=24 different conditions) for each replicate plate. After the classification, predictions were made for CV cells. The accuracy of the training set (dashed lines) and cross-validation set (solid line) for each condition in replicate plate 1 were plotted in Figure A3. We concluded from the results of all replicate plates that using mean as the center definition and choosing 300 typical cells gave the best accuracy for our data.

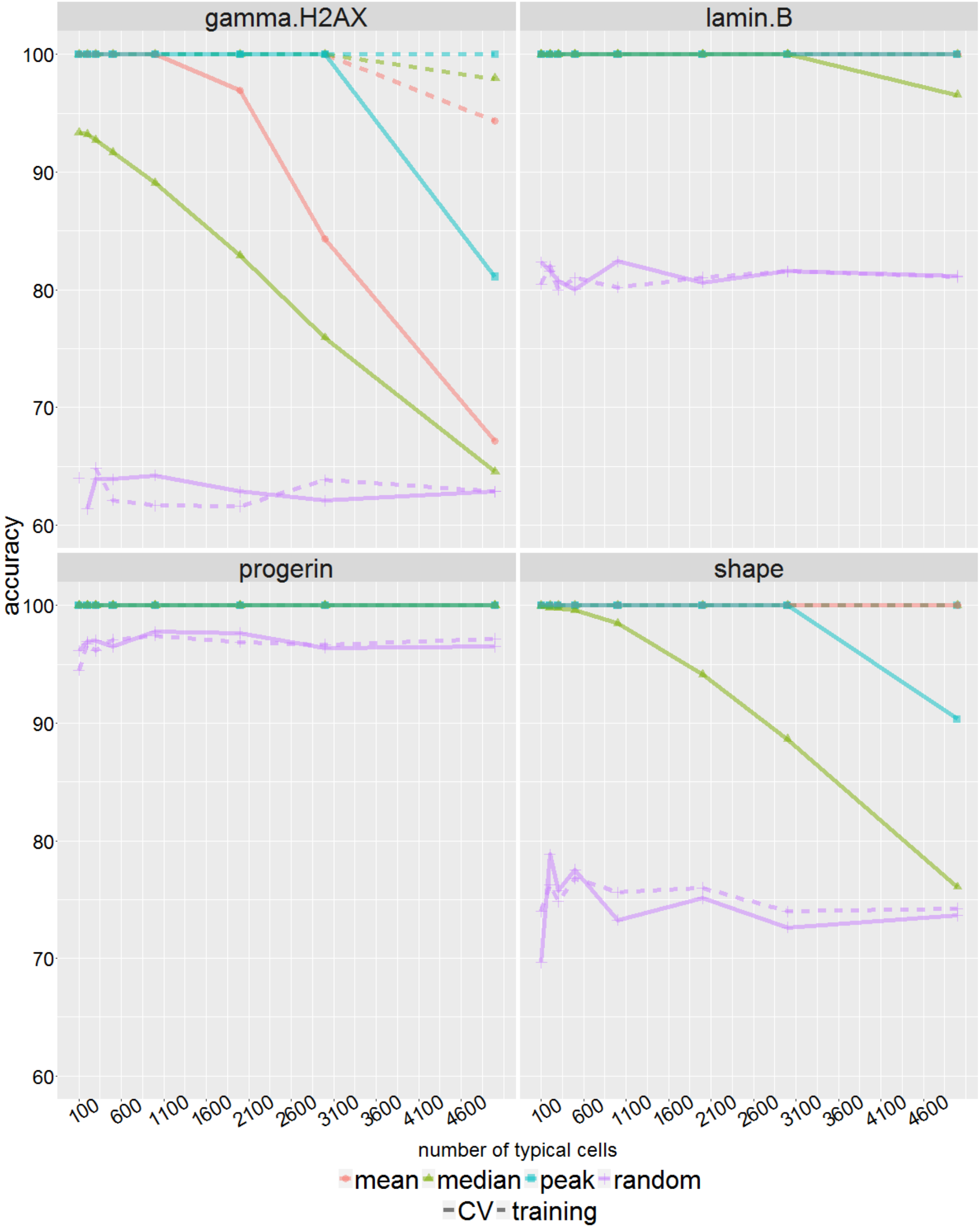


Figure SA3. Comparison of different center definition: mean (pink), median (green), peak (blue); and different typical cell numbers in plate 1. For reference, we also included randomly selected cells (purple). The vertical axis shows the accuracy of classification, and the horizontal axis is the selected typical cell number. Each subfigure plots results in one channel. The dashed line shows accuracy for the training set, and the solid line shows cross-validation set. Notice that randomly selected cells consistently behave worst, and for all channels, mean and peak behaved equally well in this replicate, but overall mean behaves the best.

A.4 Relative weights of each measurement in the 4 channels

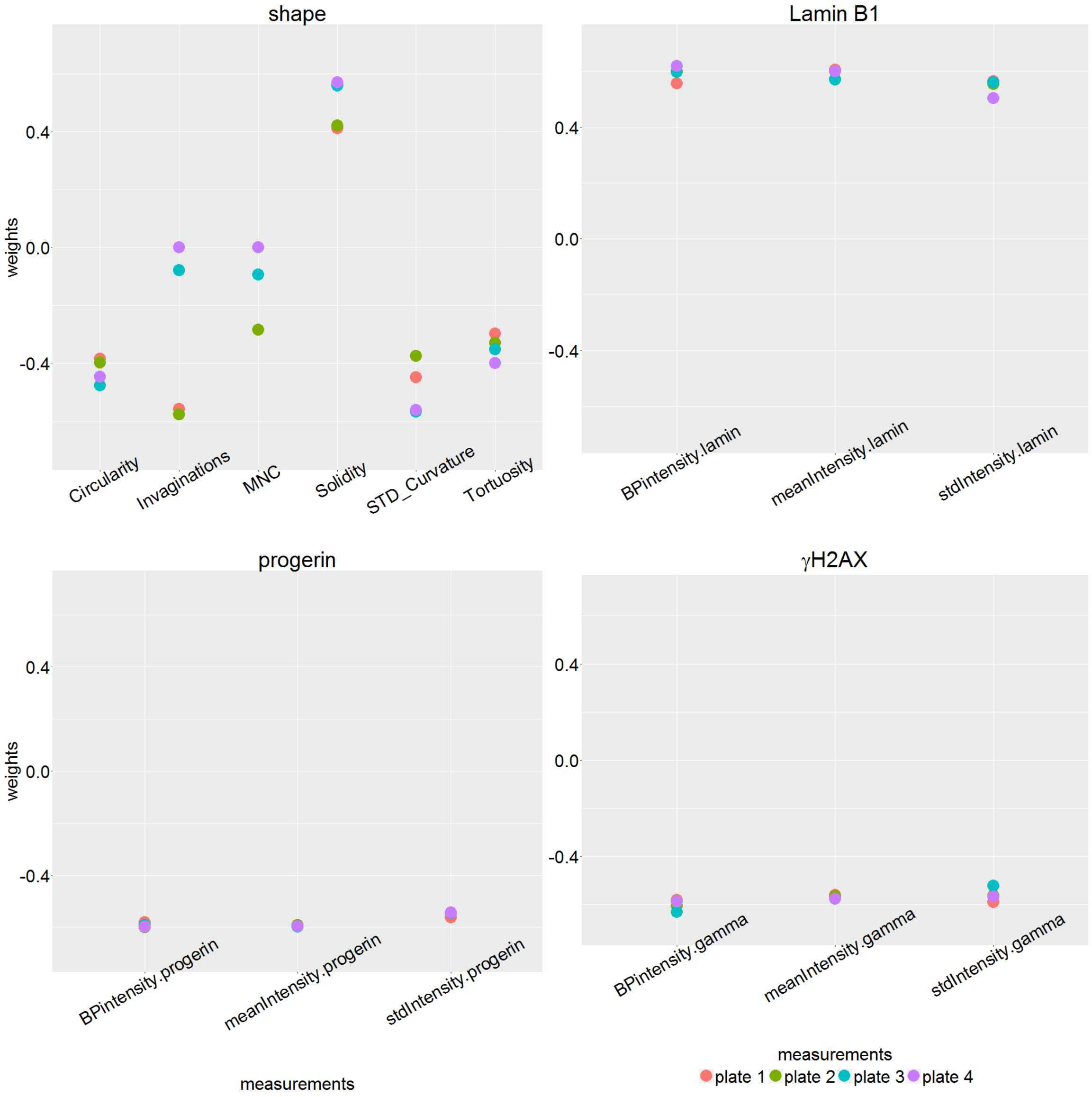


Figure A4. Weights of each measurement in all channels, plates are labeled with different colors. The absolute value of weights indicates the importance of the corresponding measurement in classifying typical

healthy vs. typical progeria cells. Sign of the weights indicates in which control type the very measurement is higher. The negative sign means it's higher in progeria controls, and vice versa. For example, solidity in shape channel has a positive sign, which means its value is higher in healthy control cells. And all measurements in progerin channel have a negative sign; this is to say all the 3 measurements have a higher value in progeria controls, which is what we expected. Solidity is the most important measurement in shape channel. Weights for intensity measurements are quite similar with only small differences and fluctuates over different plates.

A.5 Calculation of healthy-like cell percentage in each channel

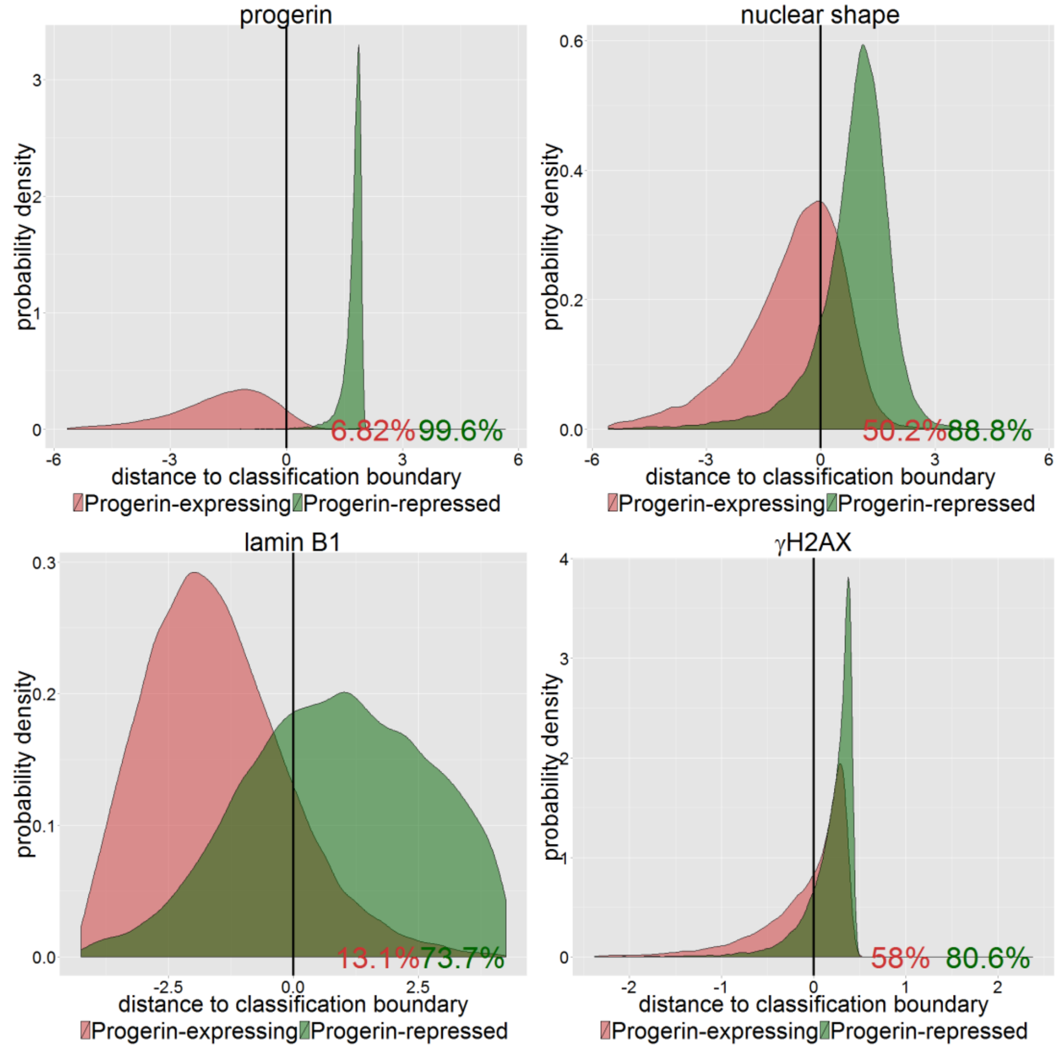


Figure SA5. Probability density distribution of all cells in GFP-progerin repressed (green) and GFP-progerin expressing (red) controls to the classification boundary in replicate plate 1. In each panel, the vertical line indicates the classification boundary located at x=0. We show here the combination of cells in 12 GFP-progerin repressed control samples (green) and 12 GFP-progerin expressing control samples (red). As expected, the progerin channel is the most distinctive between GFP-progerin repressed and expressing controls. On the bottom right of each panel, we show the average percentage of healthy-like (green) and progeria-like (red) cells in each channel, calculated from the 4 replicate plates.

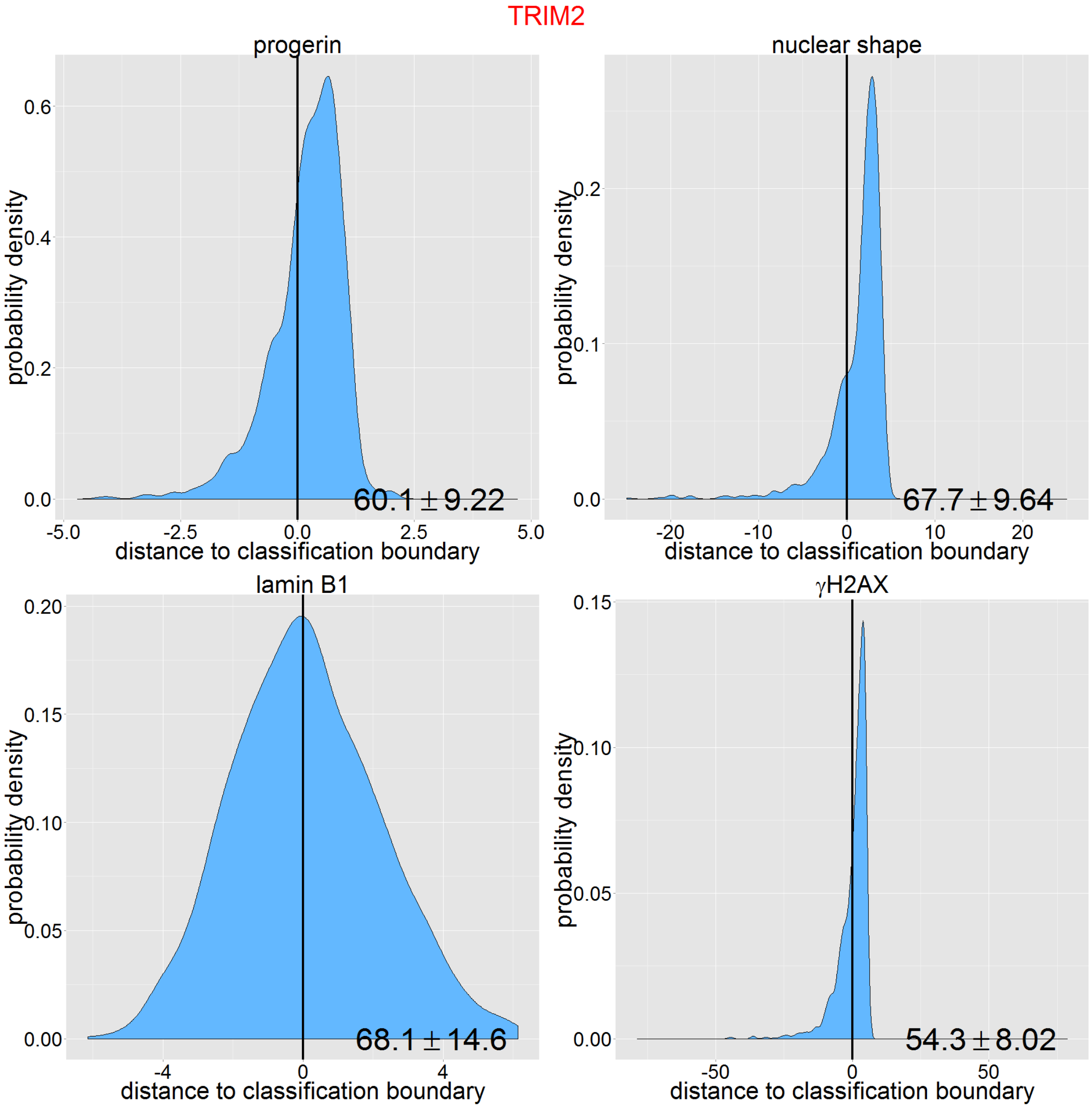


Figure A6. An example of healthy-like cell percentage calculation for one siRNA: TRIM2. Plotted here are the probability density distributions of all cells in TRIM2 sample in plate 1. Vertical lines are classification boundary, and numbers showed the average percentage of cells have positive (right, healthy-like) distance to the classification boundary with the standard deviations calculated from the 4 replicate plates. This average was later used for siRNA hits identification.

A.6 Table of selected siRNA hits

siRNA hits in each channel. Numbers in parenthesis show the percentage of healthy-like cells averaged over 4 independent replicate plates per siRNA well. Percentage of healthy-like cells of each channel in progeria controls are listed under channel name.

Channel	siRNA hits
Nuclear Shape (Progeria control baseline: 49%)	ASB12 (80%) UBE2D2 (79%) SKP2 (76%) LOC554251 (76%) RNF150 (75%) PCGF1 (74%) FLJ25076 (73%) FBXL10 (71%)
Lamin B1 (Progeria control baseline: 14%)	UBE2T (84%) CDC34 (82%) KUA-UEV (81%) UBE2O (81%) UBE2L6 (73%) PHF21B (73%) RNF39 (71%) FBXO8 (71%) HERC4 (70%) CUL5 (69%) RNF122 (69%) TRIM2 (68%) FBXO28 (68%) SMURF1 (68%) UBE3B (68%) FBXL11 (68%) RNF44 (68%) HERC3 (67%) HECTD1 (67%) UBE1L (66%) HERC5 (66%) PHF20 (66%) PHF11 (65%) FLJ25076 (65%) PHF17 (65%) FBXO38 (64%) NDP52 (64%) UBE2U (64%) WWP1 (64%) HERC2 (64%) ITCH (63%) RNF180 (62%) UBE2N (61%) TRIM55 (59%) ZMYND11 (58%) CUL7 (58%) DCUN1D4 (58%) VPS41 (58%) PRICKLE1 (57%) UBE2M (57%) TRIM52 (56%) SOCS2 (56%) UBE2D3 (56%) UBE2E2 (56%) RNF12 (56%) PDZRN3 (54%) LOC554251 (54%) LMO6 (53%) ARIH1 (53%) UBE2Q2 (52%) HERC6 (51%) UBE3A (51%) DCUN1D1 (50%) UBE2E3 (50%) WWP2 (50%) UBE2V2 (50%) UBE3C (50%) CBL (49%) TRIP12 (48%) RNF8 (48%) HECTD3 (46%) CUL4B (45%) INTS12 (45%) NEDD4L (45%) CUL2 (45%) DTX4 (44%) 39876 (43%) UBE2D1 (43%) RFPL2 (43%) ZNRF2 (42%) BMI1 (41%) FBXW8 (41%) MGRN1 (41%) HACE1 (41%) UBE2D4 (39%)
Progerin (Progeria control baseline: 6%)	TIP120A (79%) WSB1 (70%) UBE2G2 (64%) WWP2 (62%) TRIM2 (59%) LOC554251 (55%) FBXO38 (52%) RNF39 (50%) TRIM55 (49%) MLLT6 (45%) FBXO17 (44%) FLJ25076 (44%) HERC3 (43%) CUL3 (43%) UBE2D2 (40%) FBXL13 (39%) TRIP12 (38%) UBE1C (37%) ASB5 (36%) SMURF1 (33%) RNF150 (33%) RNF44 (32%) FBXL11 (31%) PHF20L1 (30%) RNF32 (28%) ASB12 (27%) TRIM8 (26%) UBE2T (25%) HIP2 (23%) CBLC (23%) MYCBP2 (20%)
γH2AX (Progeria control baseline: 60%)	WWP2 (86%) WSB1 (81%) ZNF330 (78%) RFPL4B (77%) WDR24 (75%)

A.7 Comparison with another method

We compared the results of our method with another multi-dimensional analysis method proposed in Ref [34]. We chose two channels: progerin and Lamin B1 for this comparison.

To analyze our data using the method in [34], we first randomly selected 5,000 cells (about 25%) from both GFP-progerin expressing and repressed controls respectively. We pooled these 10,000 cells together and clustered them with GMM as described in [34]. In total, we identified 9 (8) clusters in progerin and γH2AX (lamin B1) channel; these were used as reference models for siRNA perturbation samples. We then computed the probability of each cell belonging to one of these 9 (8) clusters. The expectation of the proportion of cells inside each cluster for each perturbation is then calculated based on these probabilities. Using an expected fraction of cells in each cluster as a vector, the distance between each perturbation to GFP-progerin repressed (healthy-like) controls are calculated using KL divergence. The inverse of this divergence is then used as the metric to select important perturbations, the larger the metric, the more similar to the healthy-like controls, hence the more important the perturbation. We then compared our metric (percentage of healthy-like cells) with this metric, and we found that they correlate well with each other.

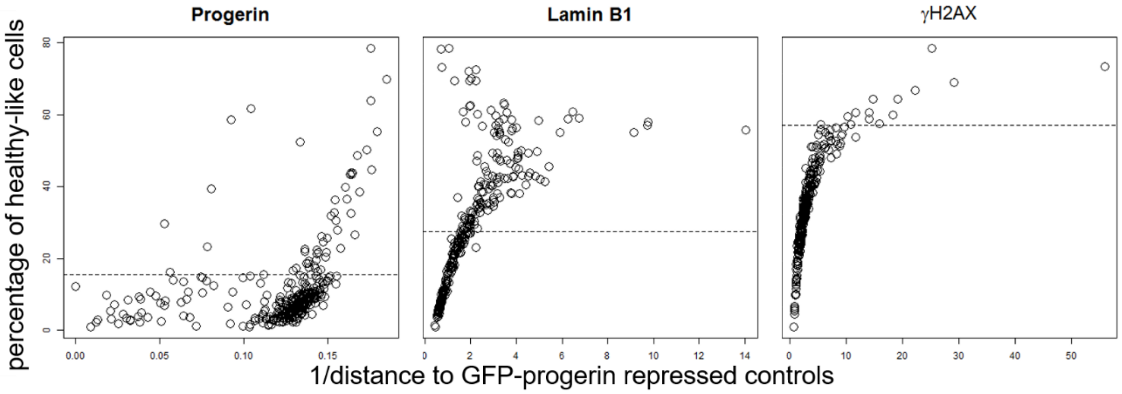


Figure A7. Comparison between our metric (healthy-like cell percentage, y-axis) and metric derived using the method proposed in [34] (x-axis) in lamin B1, progerin and γH2AX channel. The two metrics correlate well in all channels, with Spearman correlation coefficient 0.98 for γH2AX channel, 0.91 for lamin B1 channel, 0.58 for progerin channel (p-value << 0.05 in both cases).

A.8 List of screened siRNAs

UBE2C SMURF1 HERC3 UBE2W DCUN1D3 BIRC6 CUL3
ITCH DCUN1D2 CUL1 UBE2O UBE2V1 UBE1C UBE2E3
UBE2U KUA-UEV EDD1 HIP2 UBE2V2 DCUN1D5 UBE2J2
HECW1 HUWE1 CAND2 UBE3B UBE2A UBE2R2
AKTIP UBE2D1 UBE3A TIP120A KIAA0317
HECTD1 UBE2T HERC2 UBE2N UBE1L UBE1
UBE2D3 NEDD4 UBE2E2 HECTD3 UBE2E1 UBE3C
TRIP12 HECTD2 UBE2D4 DCUN1D4 CDC34 FLJ34154
UBE2Z UBE2L3 HERC1 HACE1 UBE1DC1 UBE2M UBE2I
UBE2Q2 TSG101 UBE2G2 CUL4A HERC4 CUL4B
UBE2L6 DCUN1D1 CUL2 HERC5 UBE2NL CUL7 UBE2S
UEVLD UBE2F HERC6 WWP1SMURF2 NEDD4L HECW2
UBE2B UBE2J1 FLJ25076 UBE2Q1 UBE1L2 UBE2G1 FBXL15
UBE2H FBXO18 CUL5 LOC554251 ARIH1FBXL19 UBE2D2 WSB1
WWP2 FBXL7 C10ORF46 ASB18
FBXL16 FBXO27 FBXW10 FBXO42 FBXO21 FBXO40
FBXO30 FBXO17 FBXL10 SKP2 SOCS7 ASB12 FBXL11
WSB2 FBXL20 SOCS2 FBXO43 LOC200933 ASB13SPSB1
SOCS6 FBXO6 FBXL8 ASB16LRRC29 FBXW8
ASB9 FBXO4 FBXO16 ASB14FBXW11 ASB6
RAB40C FBXW2 SPSB3FBXO31 FBXO9 FBXL3P SPSB2
WDR71 FBXO41 SOCS5 FBXL13 ASB11FBXO5
FBXL17 FBXO28 FBXO46
LOC342897 FBXL18 SOCS3 FBXO15 FBXO3 FBXO22
FBXO11 FBXL14 FBXO7 FBXO25 RAB40A
FBXL2 ASB5 CISH ASB8 FBXW9
LOC652759 TULP4 NLRC5 FBXO8 CCNF FBXO44
LL0XNC01-237H1.1 ASB2 FBXO24 LOC440456 FBXO2
FBXO38 BTRC ASB7 FBXL12 FBXL3A
ASB4 FBXO33 LGR6 FBXL6 NEURL2 ASB15FBXO36
FBXW5 SOCS1 FLJ10916 FBXW12 FBXL4
FBXO32 SPSB4FBXW7 FBXO39
SHFM3 ASB10FBXL5 ASB3 FBXO10 FBXL22 FBXO47
RAB40B FBXO34 MDM2 SOCS4 PRPF19 ASB17
TRIM6-TRIM34 ASB1 ZMYND11
JARID1B RNF123 RKHD1 TRIM67 TRIM75 PHF17 OIT3
MGRN1 PHF7 TRIM39 RBX1 LOC653111 LOC642678
3/8/2009 BIRC3 MIB2
PHF11 TRIM60 LOC644006 PHF6 TRIM42 BAHD1 RNF7
WDSUB1 TRIM41 RNF133 HRC MYCBP2 PHF20L1
RNF152 TRIM62 RNF125
TRIM8 RNF122 TRIM63 RFPL4B WDR24 DTX4
PCGF1 TRIM3 KIAA1718 RNF32PRICKLE1 CBL
RFWD2 RSPRY1 BMI1 PHF20

RNF39 RNF12 PDZRN3 C6ORF49 TRIM26 PHF21B ZNF645
RNF5 INTS12 ZNF592 CHD5 RNF180 UNKL MID2
ZNF313 RNF185

RNF135 ZNRF2 PHF5A C20ORF18 3/4/2009 ANKIB1
PHF15 BRCA1 LOC92312 ZFAND6 PHF21A HR
MLLT6 TRIM14 ZNRF3 NDP52

LOC643904 TRIM40 LOC399937 TRIM43 LMO6 TRIM52
RNF144 LONRF1 SH3RF2 RNF150 PHF23 RNF25
RUFY1 ZNF330 UBR2 TRIM2

RFPL2 PHF16 ZNF179 RAD18 CBLC RNF44 TRIM55 BRPF3
PCGF3 RNF8 PHF13 DTX3L RNF148 VPS41 RNF103
TRIML1

Appendix B Supplementary Information for Chapter 3

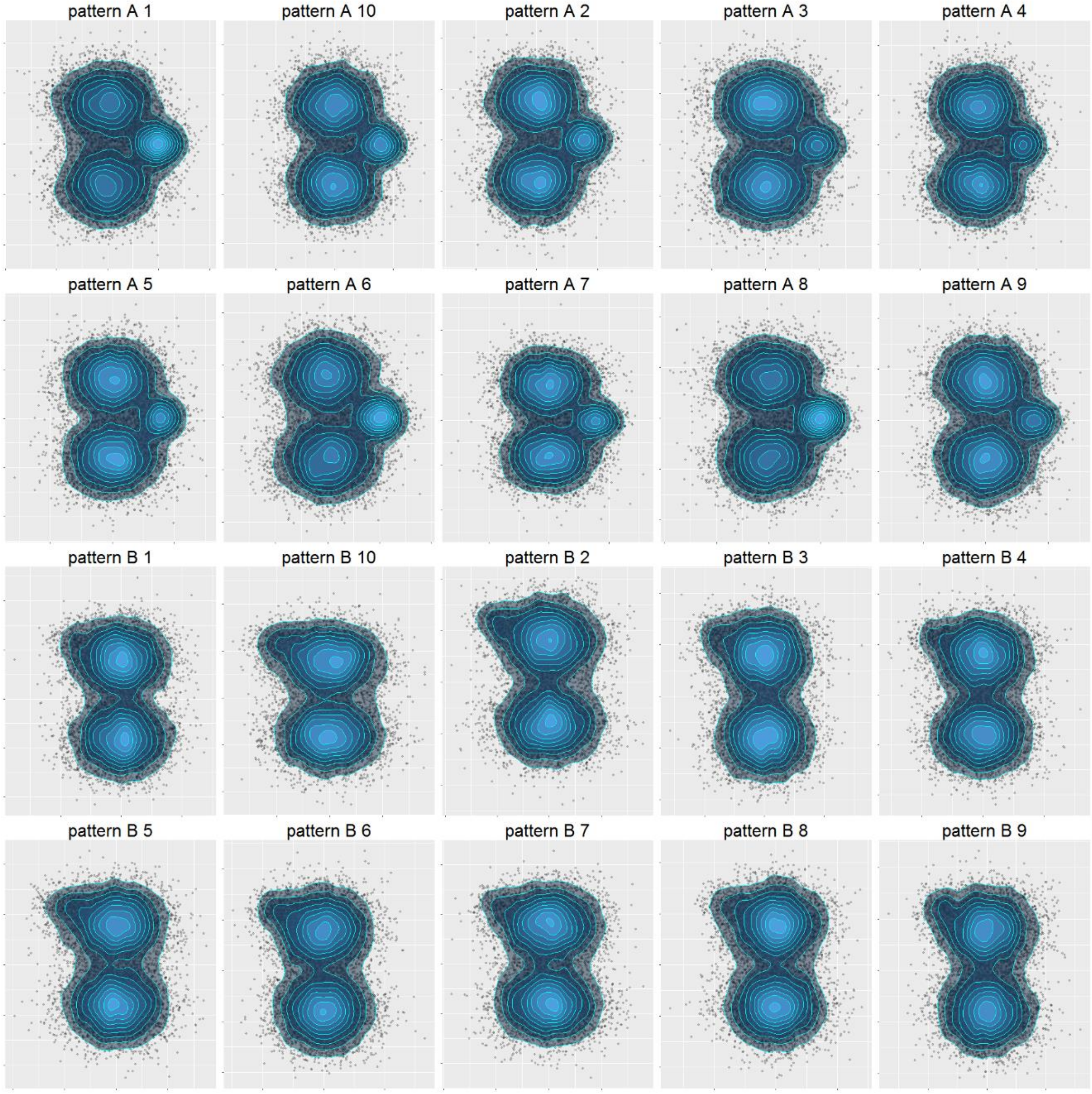


Figure B1. Scatter plots of simulated point patterns. First two rows show point pattern A; the lower two rows show point pattern B. Two major clusters in both point pattern A and B are generated from the same distributions. The third cluster of point pattern A, located on the center right, consists of about 10 to 20% of total cells. The third cluster of point pattern B, located at upper left of all points, contains only 2 to 5% of all cells.

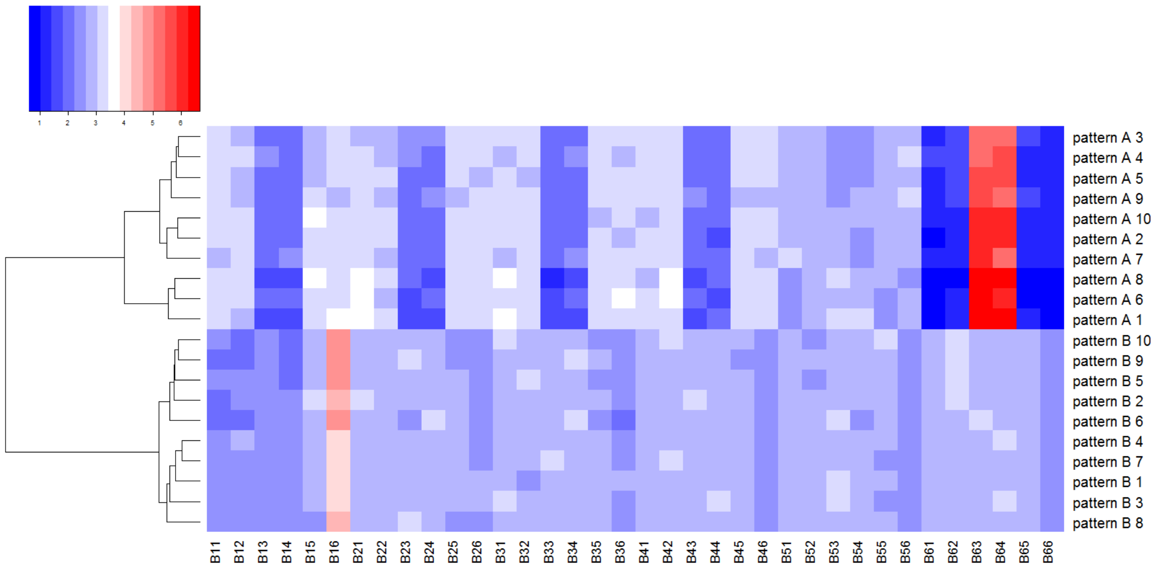


Figure B2. Heatmap for a percentage of cells inside each box with 6 bins. Percentage of cells in box B16 (which corresponds to the third cluster in point pattern B) is significantly different between these two point patterns. This is not seen with only 3 bins. However, with 20 samples, analysis results using 6 bins is not reliable. Hence, in order to identify the fine difference, more samples are needed.

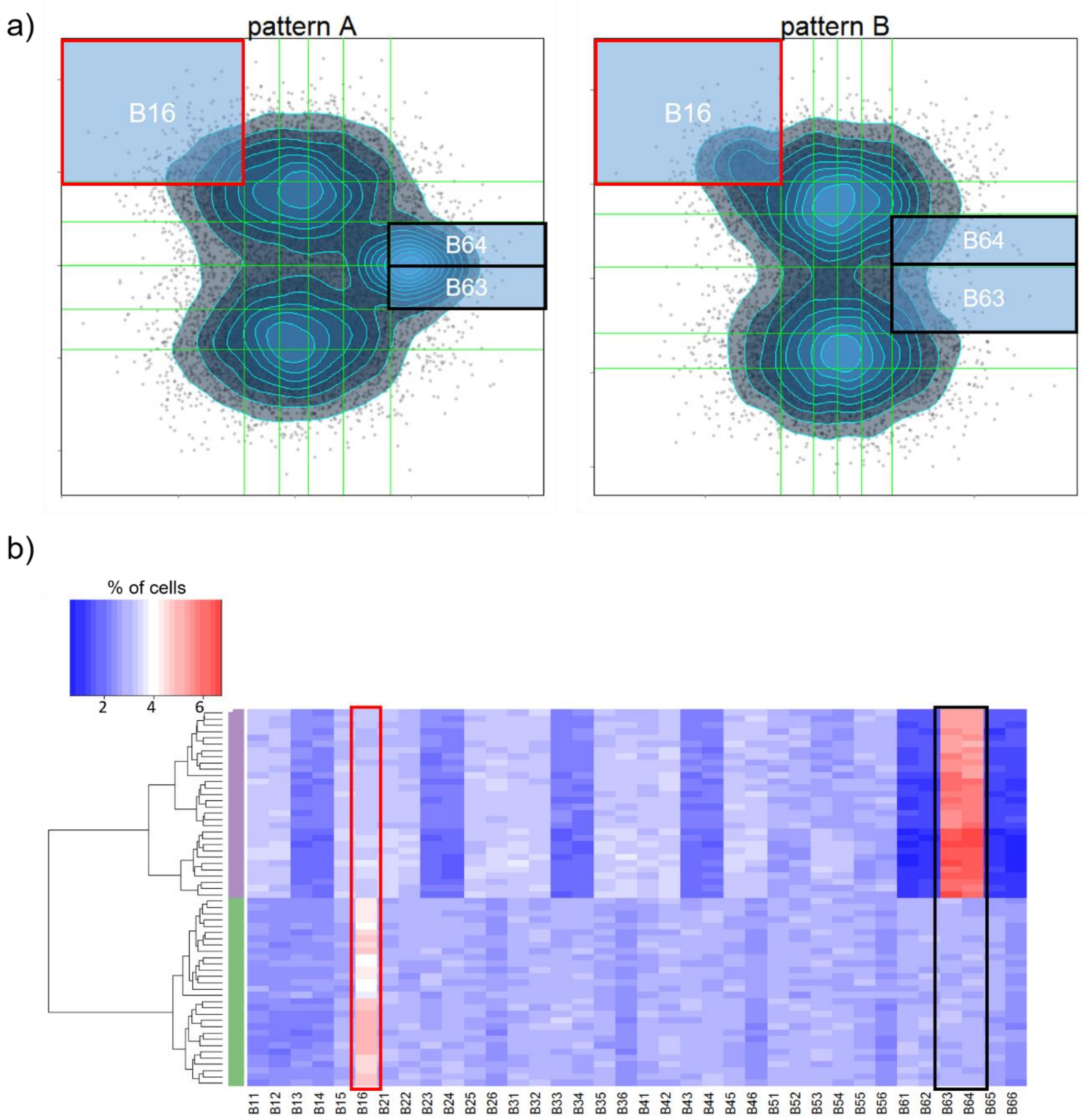


Figure B3. With 6 bins, both different between pattern A and pattern B can be found by CytoBinning.

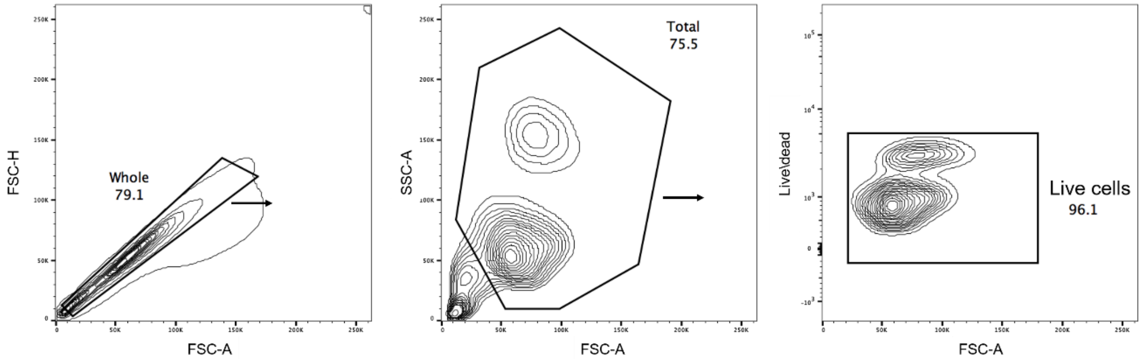


Figure B4. Illustration of manual gating strategy to get live cells.

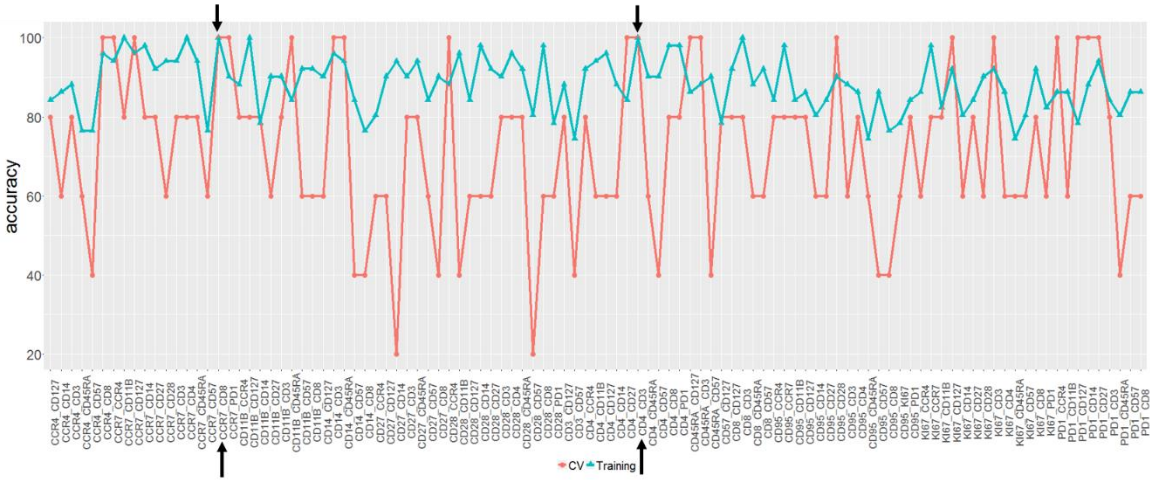


Figure B5. Select important marker pairs for the first dataset (old vs. young). Ten samples are randomly selected as cross-validation dataset (4 in young group and 6 in the old group). SVM classification was used to separate old and young samples with binning results for each marker pair separately. Two marker pairs are able to achieve 100% classification accuracy for both training and cross-validation dataset (CD4 vs. CD3 and CD8 vs. CCR7).

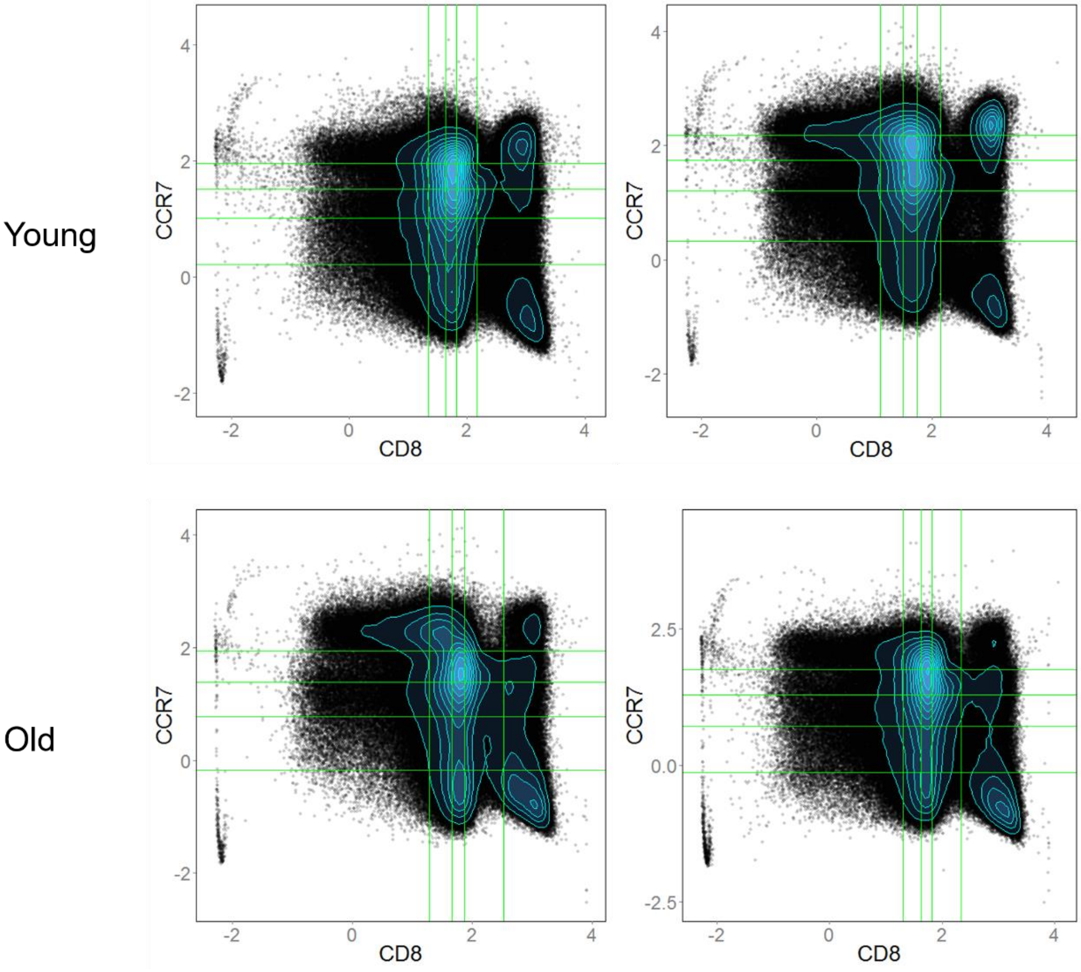


Figure B6. Scatterplots of CD8 vs. CCR7 for two randomly selected young donors (up) and two randomly selected old donors (down).

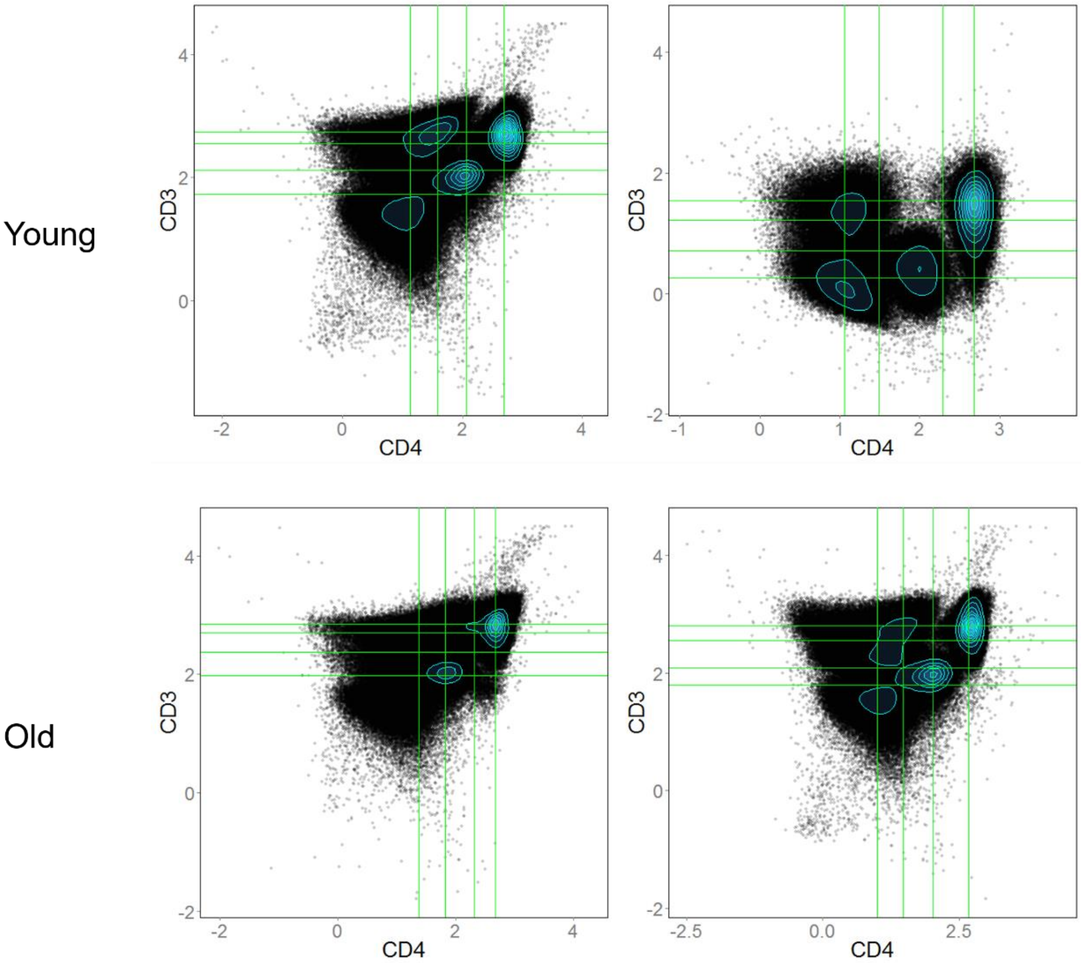


Figure B7. Scatterplots of CD4 vs. CD3 for two randomly selected young donors (up) and two randomly selected old donors (down).

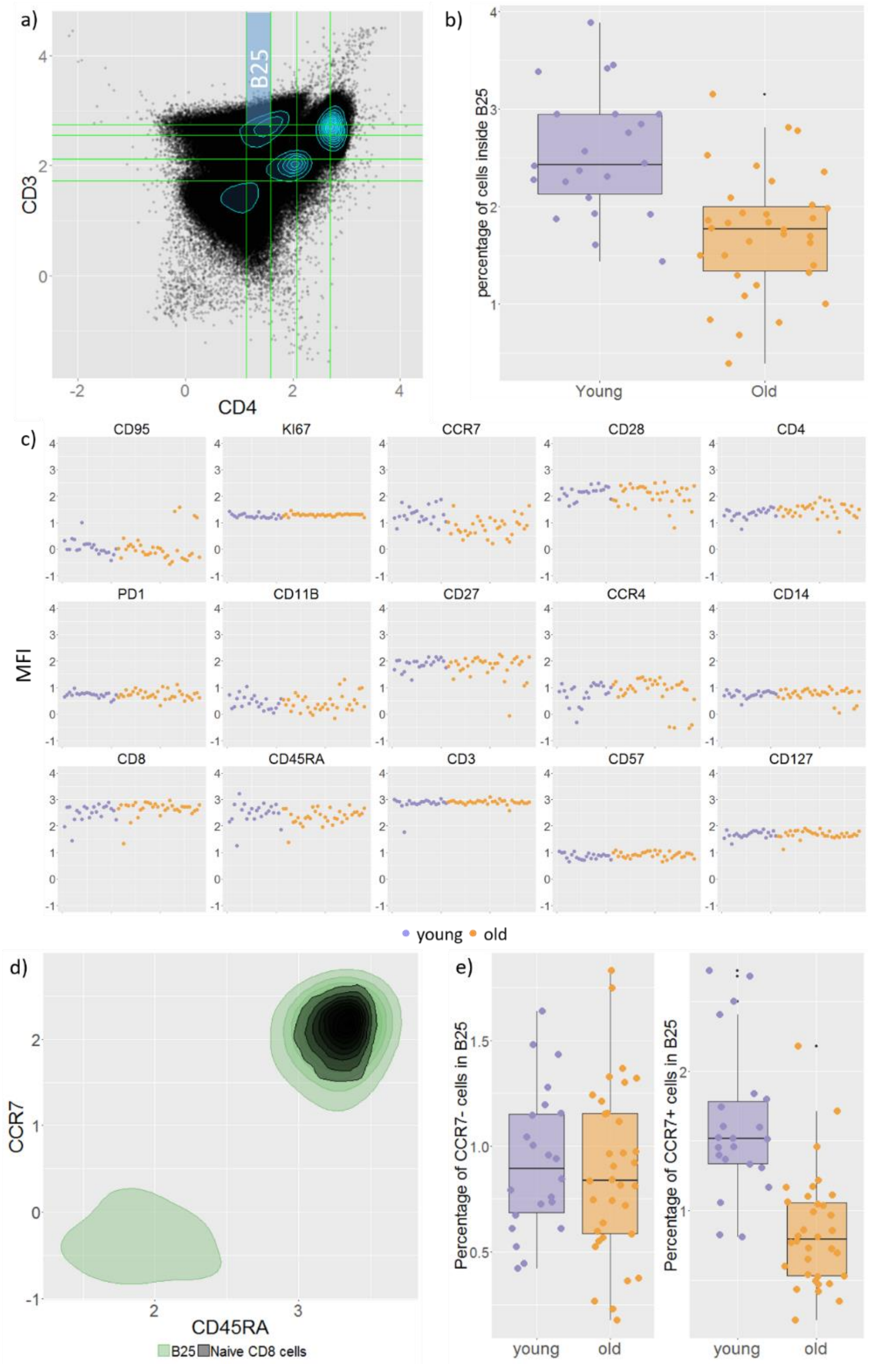


Figure B8. Illustration of box B25 formed by CD4 and CD3. a) The position of box B25. b) Percentage of cells in B25 is higher in young donors. c) Scatter plot of mean fluorescent intensity (MFI) for all donors and all markers. This suggests cells in B25 are CD3+, CD8+, and CD45RA+. d) An example showing how cells in B25 (green) compare to manually gated naïve CD8 cells. e) Cells in B25 are divided into two groups: CCR7+ (expression of CCR7>1) and CCR7- (expression of CCR7<1). The boxplots show that difference of cell percentage between old and young donors in B25 is driven by CCR7+ cells.

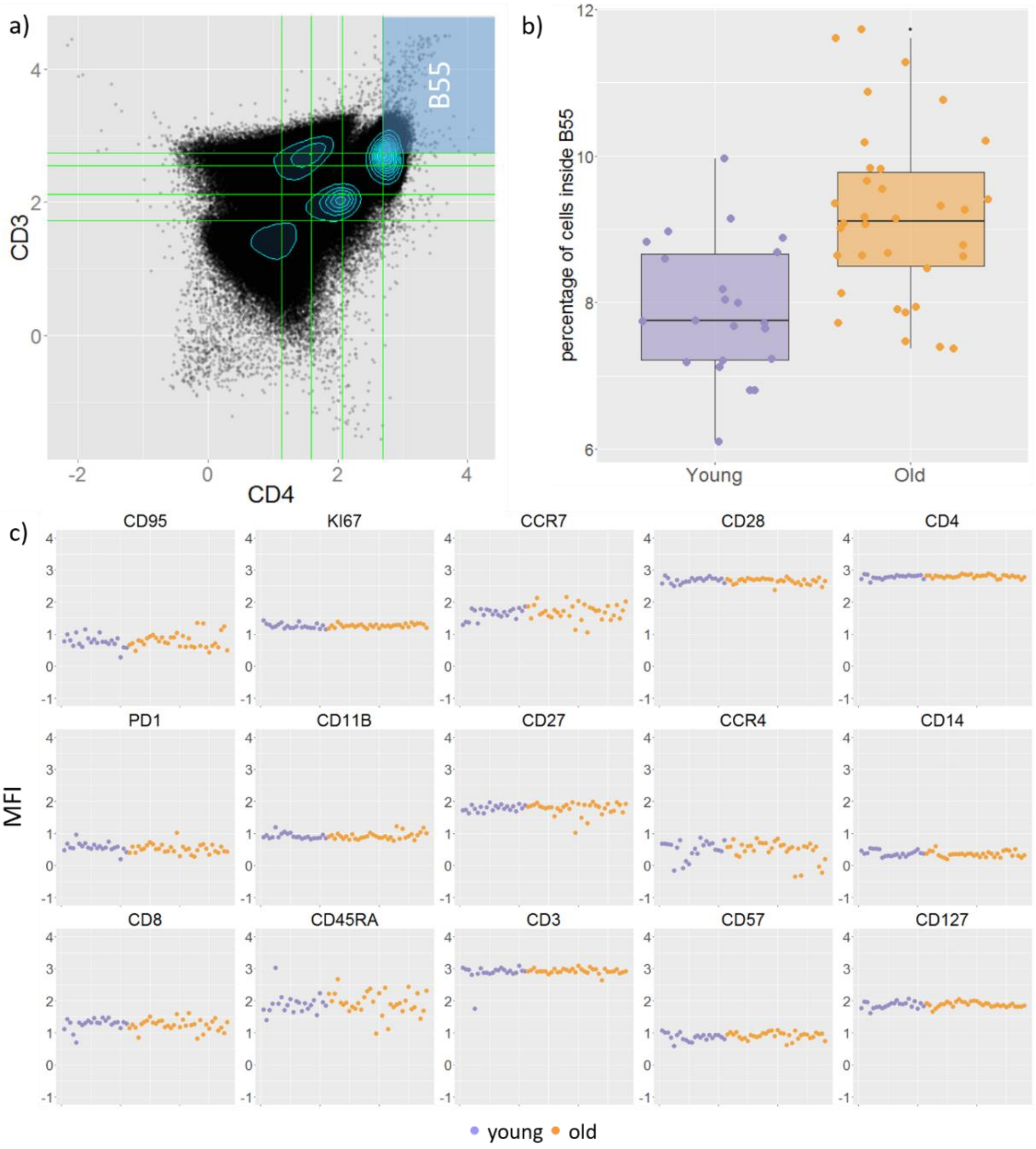


Figure B9. Illustration of box B55 formed by CD4 and CD3. a) The position of box B55. Cells in B55 express the highest 20% of both CD3 and CD4. Hence they might be CD4 T cells. b) Percentage of cells in B55 is higher in old donors. c) Scatter plot of mean fluorescent intensity (MFI) for all donors and all markers. It suggests cells in B55 might be CD8-, CCR7+, and CD45RA+.

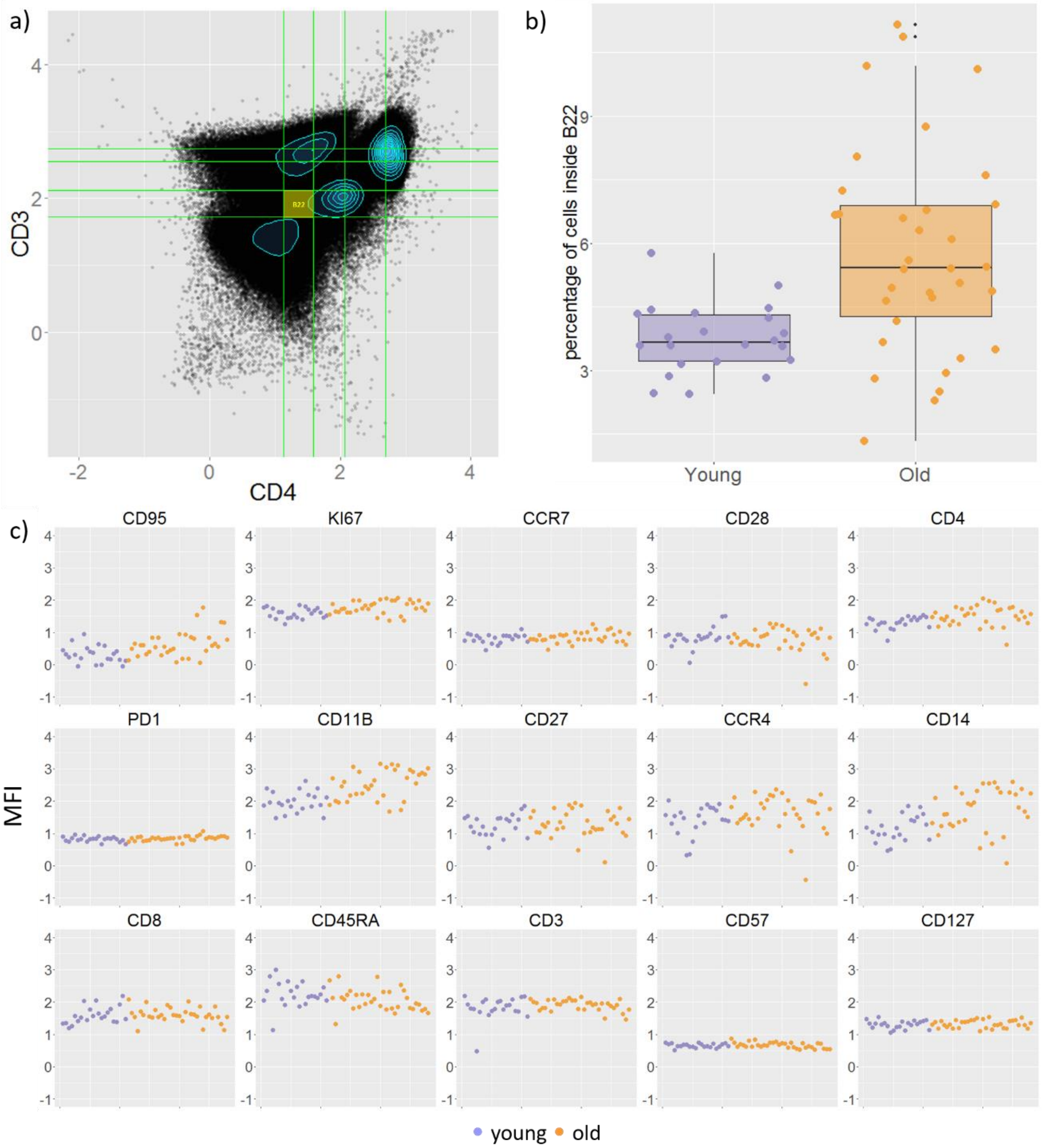


Figure B10. Illustration of box B22 formed by CD4 and CD3. a) The position of box B22. b) Percentage of cells in B55 is higher in old donors. c) Scatter plot of mean fluorescent intensity (MFI) for all donors and all markers. It suggests cells in B22 might be CD11b+, CD14+, and CD45RA+.

List of markers measured in Function panel.

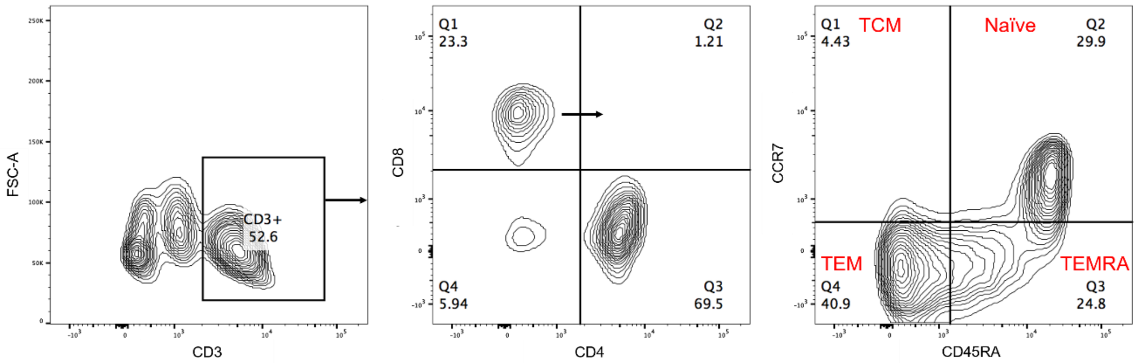


Figure B11. Illustration of manual gating strategy for naïve and memory CD8 T cells.

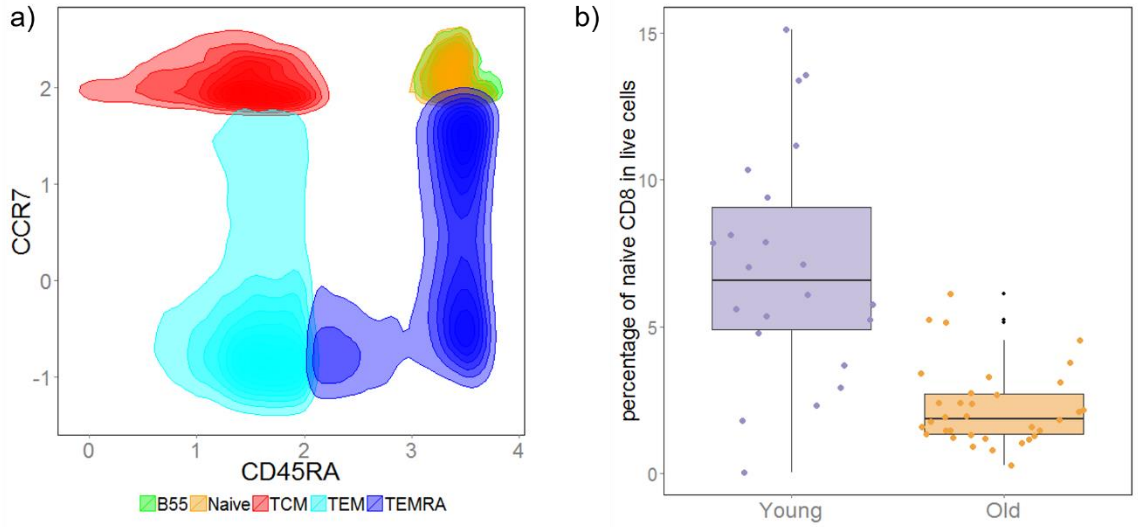


Figure B12. a) Overlay of cells in B55 on CD8 naïve and memory cell types for one donor. b) Boxplot of naïve CD8 cell percentage in live cells.

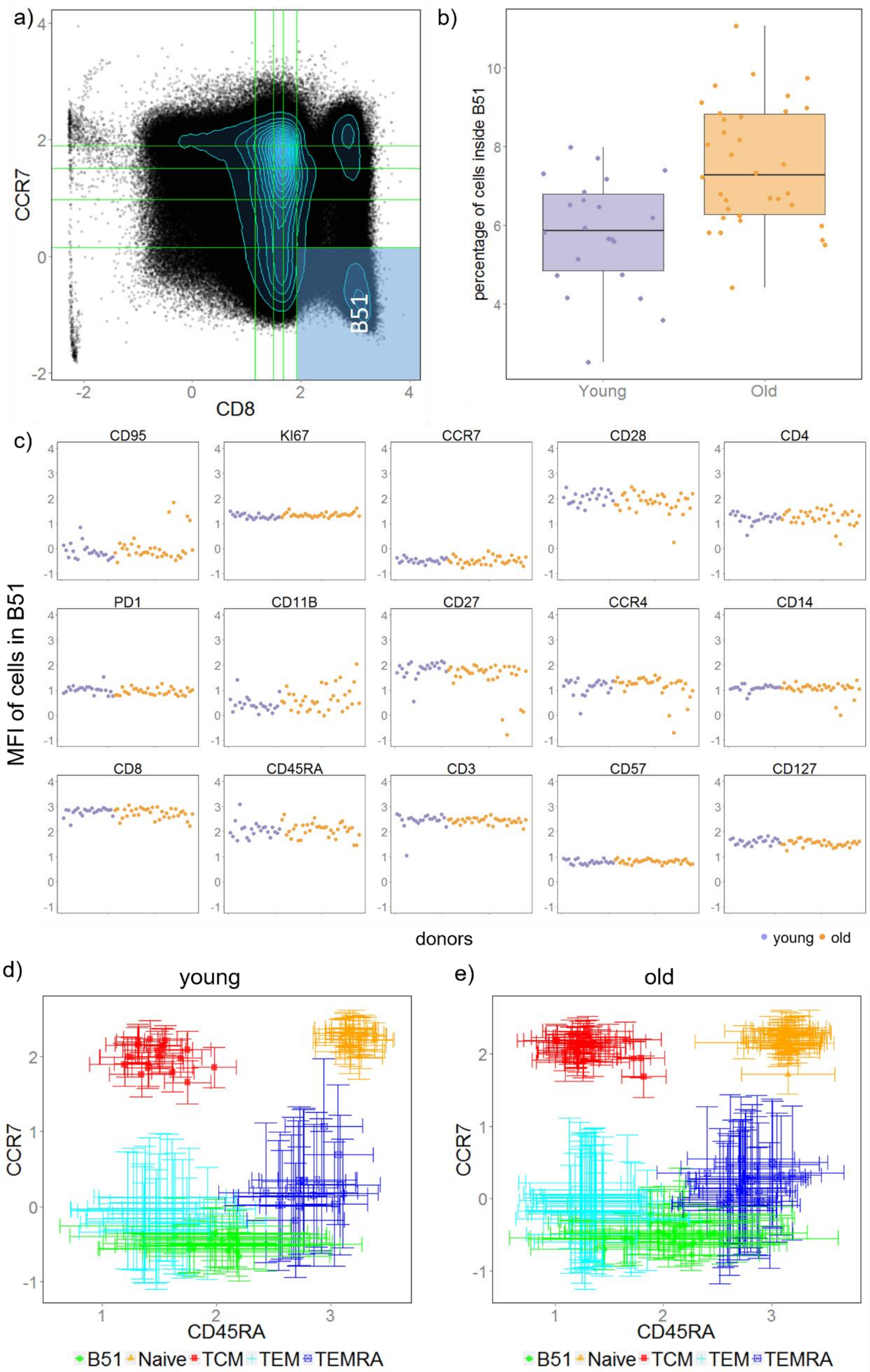

a)
CCR7
CD8
B51
b)
percentage of cells inside B51
Young
Old
c)
CD95
KI67
CCR7
CD28
CD4
PD1
CD11B
CD27
CCR4
CD14
CD8
CD45RA
CD3
CD57
CD127
MFI of cells in B51
donors
young
old
d)
young
e)
old
CCR7
CD45RA
B51
Naive
TCM
TEM
TEMRA

Figure B13. Illustration of box B51 formed by CD8 and CCR7. a) The position of box B51. b) Boxplot of cell percentage in B51 between young and old donors. c) Scatter plot of mean fluorescent intensity (MFI) for all donors and all markers. d & e) MFI of CD45RA vs. MFI of CCR7 for cells in B51, naïve and memory CD8 T cells. Each symbol shows a donor (young donors in d and old donors in e), vertical and horizontal errorbars show the standard deviation of CCR7 and CD45RA intensity respectively.

List of markers measured in CD4 vs CD8 dataset

CD45, CD14, CD57, TCRgD, IFNg, TNFa, IL.8, GranzymeB, IL17F, CD45RA, CLA, CTLA.4, IL2, CD25, CD103, CCR10, CXCR6, IL5, CD19, CD56, IntegrinB7, PD-1, IL9, CCR9, CXCR3, CD127, Mip1b, CXCR5, CD161, CCR2, IL4, IL10, CCR6, GM.CSF, CCR4, IL22, CCR5, IL17A

Appendix C Supplementary Information of Chapter 4

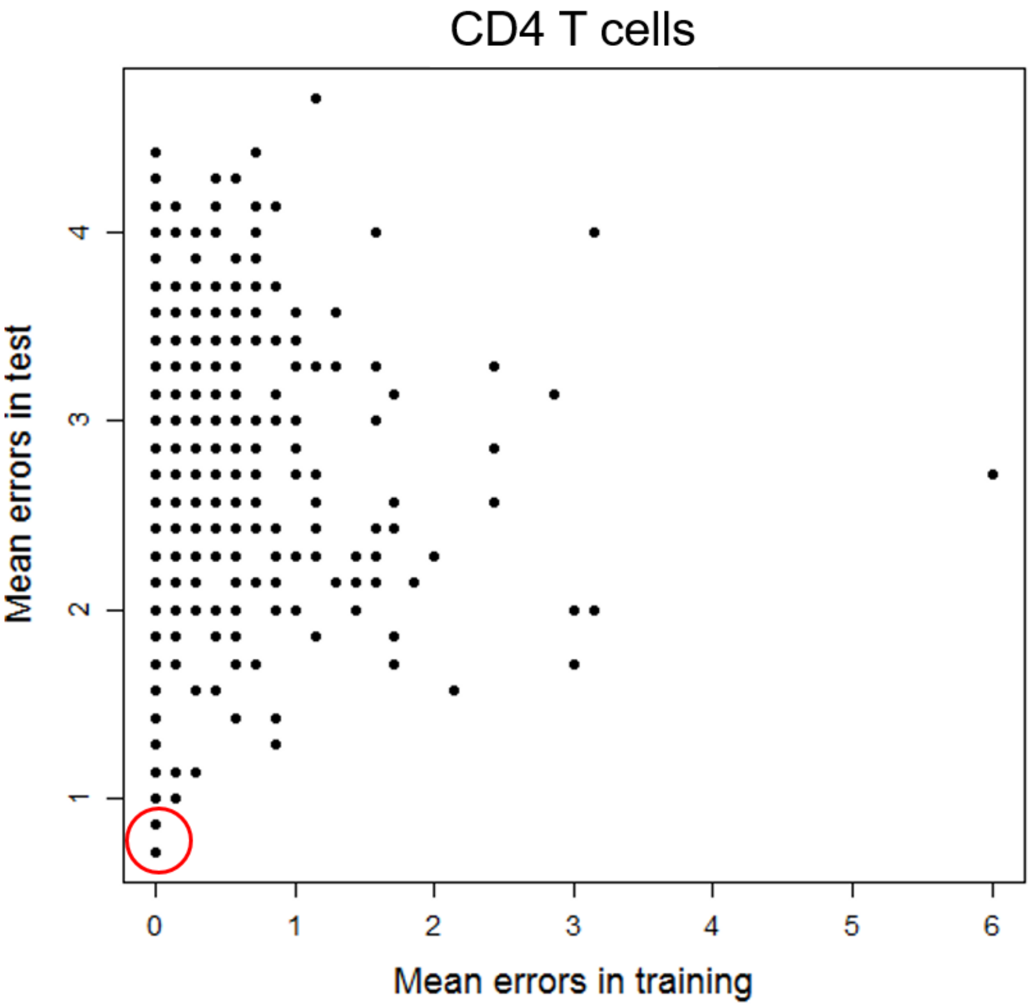


Figure C1. Important marker pairs are selected based on lowest training and testing errors. This example demonstrates the average training and testing error for a 7-fold cross validation based on multi-class Support Vector Machine (SVM) classification. Each training set contains 30 samples, and each testing set contains the remaining 5 samples.

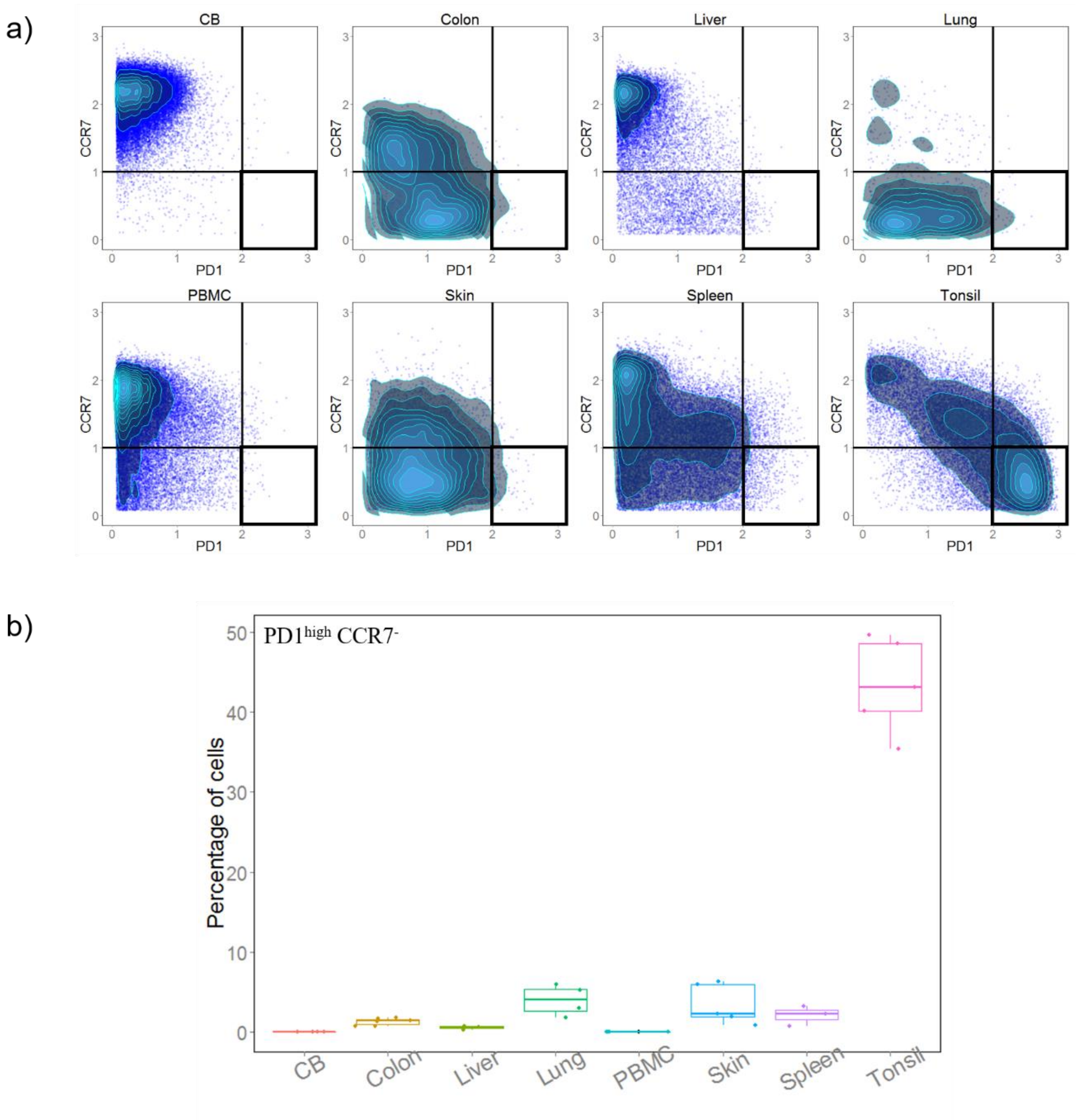


Figure C2. Manual gating validates the PD-1^{high} CCR7^{-} cell subset is highly overexpressed in tonsil samples. a) Example showing manual gating for cell subsets PD-1^{high} CCR7^{-} (highlighted in black boxes) in different types of tissues. b) Boxplot showing percentage of PD-1^{high} CCR7^{-} cells for each type of tissues.

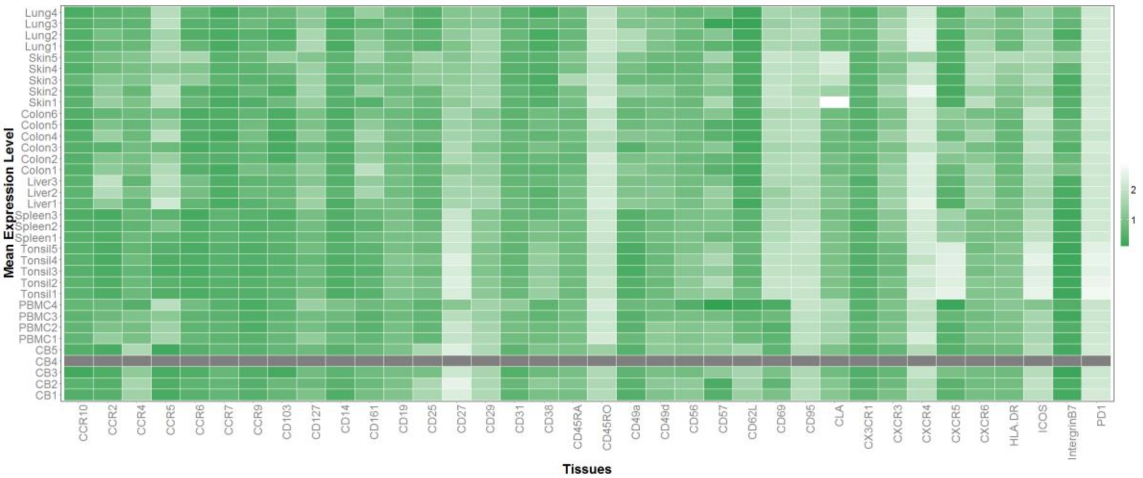


Figure C3. Mean marker expression of manually gated CCR7^{-}PD-1^{high} cells. The dark row of CB4 indicates that in this cord blood sample there are no CCR7^{-}PD-1^{high} cells.

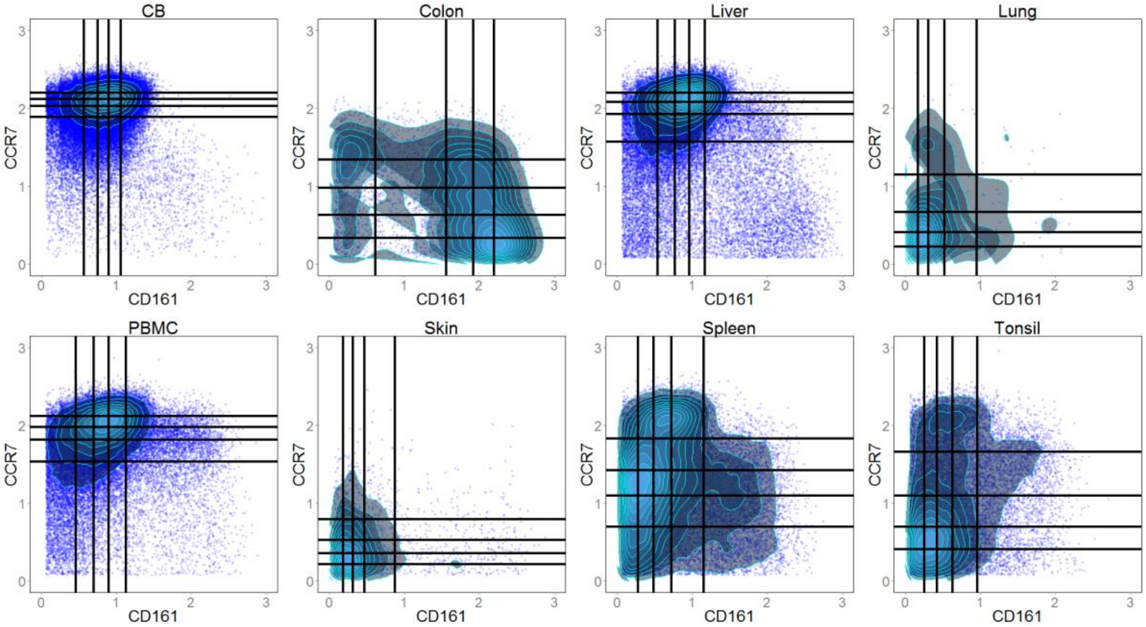


Figure C4. Patterns formed by CCR7 and CD161 expression are distinct among different types of tissues. The black lines are binning thresholds determined by CytoBinning. Five bins are used to divide these patterns. Cord blood, PBMC and liver samples look similar at first glance, however, if observe closely, PBMC express more CD161^{-}CCR7^{+} and CCR7^{+}CD161^{+} cells than cord blood, and liver express more CD161^{+}CCR7^{-} cells than both PBMC and cord blood. In addition, there's a cell phenotype: CD161^{+}CCR7^{+}, that is highly overexpressed in colon samples.

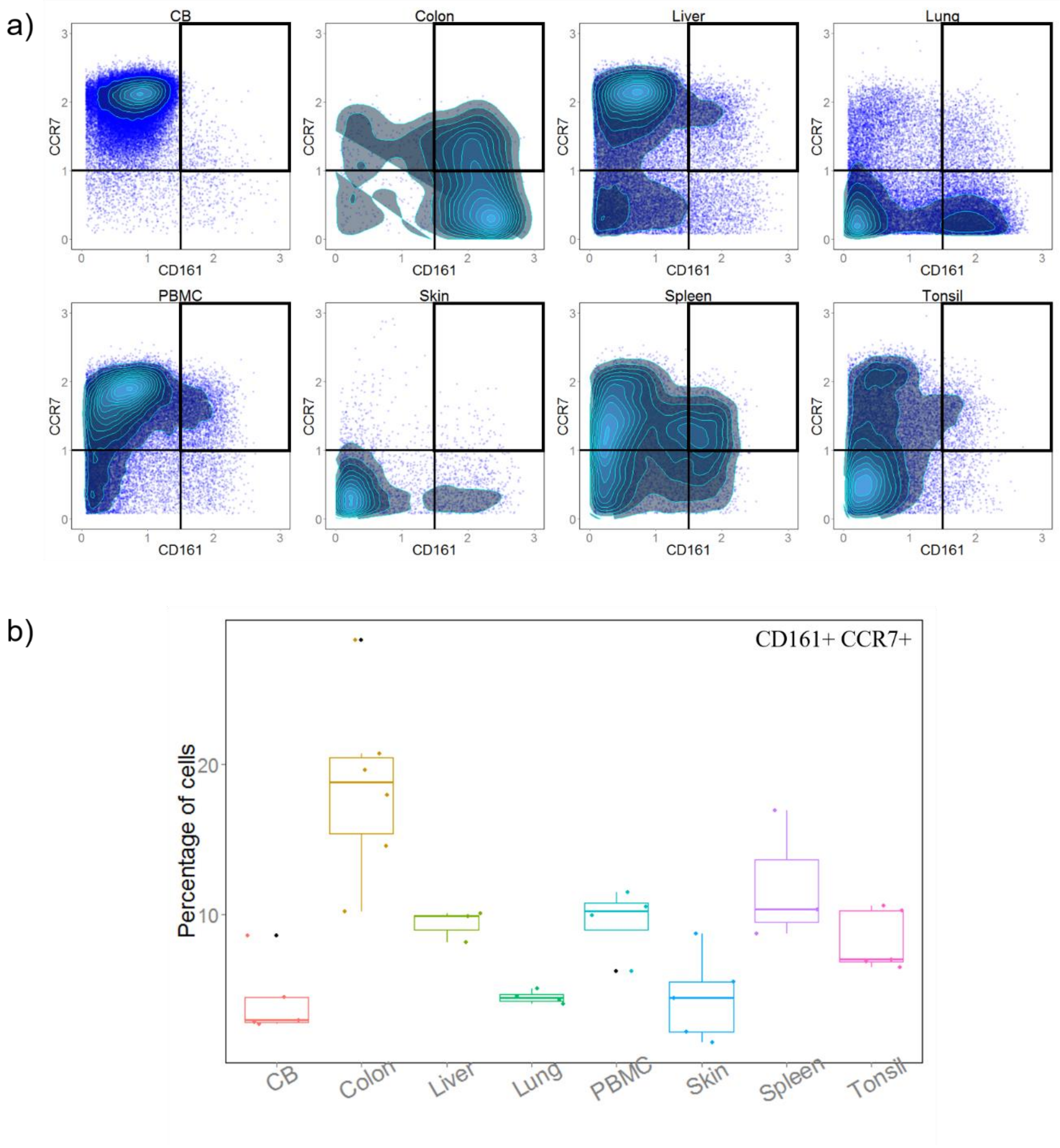


Figure C5. Manual gating validated that CD161$^+$ CCR7$^+$ cell subset is highly expressed in colon samples comparing to other types of tissues. a) Example showing manual gating for cell subsets CD161$^+$ CCR7$^+$ (highlighted in black boxes) in different types of tissues. b) Boxplot showing percentage of CD161$^+$ CCR7$^+$ cells for each type of tissues.

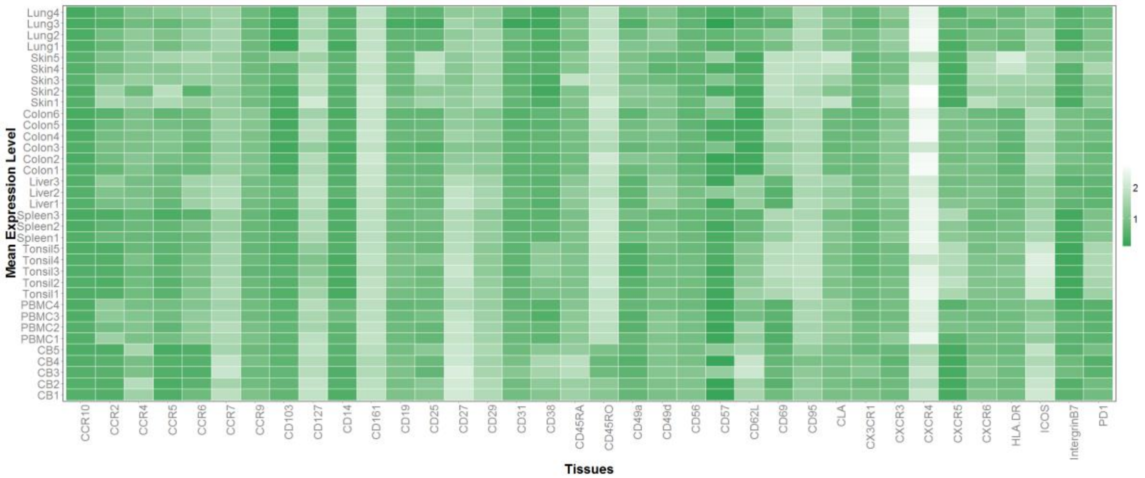


Figure C6. Mean marker expression level of CD161^{+}CCR7^{+} cells. In addition to CD161 and CCR7, these cells also express CD27, CXCR4, and CD45RO (except for cord blood samples).

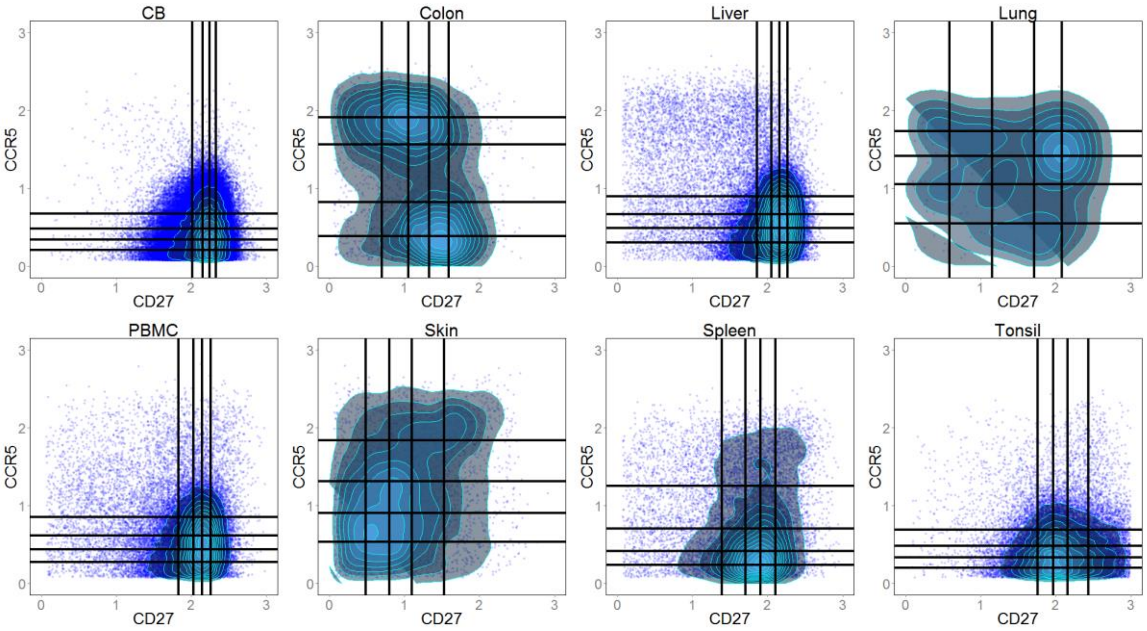


Figure C7. Patterns formed by CCR5 and CD27 expression are distinct among different types of tissues. The black lines are binning thresholds determined by CytoBinning. Five bins are used to divide these patterns. A cell phenotype CD27^{+}CCR5^{+} is identified to be highly overexpressed by CD4 T cells in lung. Since the lung samples in this dataset are collected from lung cancer resection, we're not sure if the overexpression is because of immune function of lung tissues or is due to the tumor mechanism.

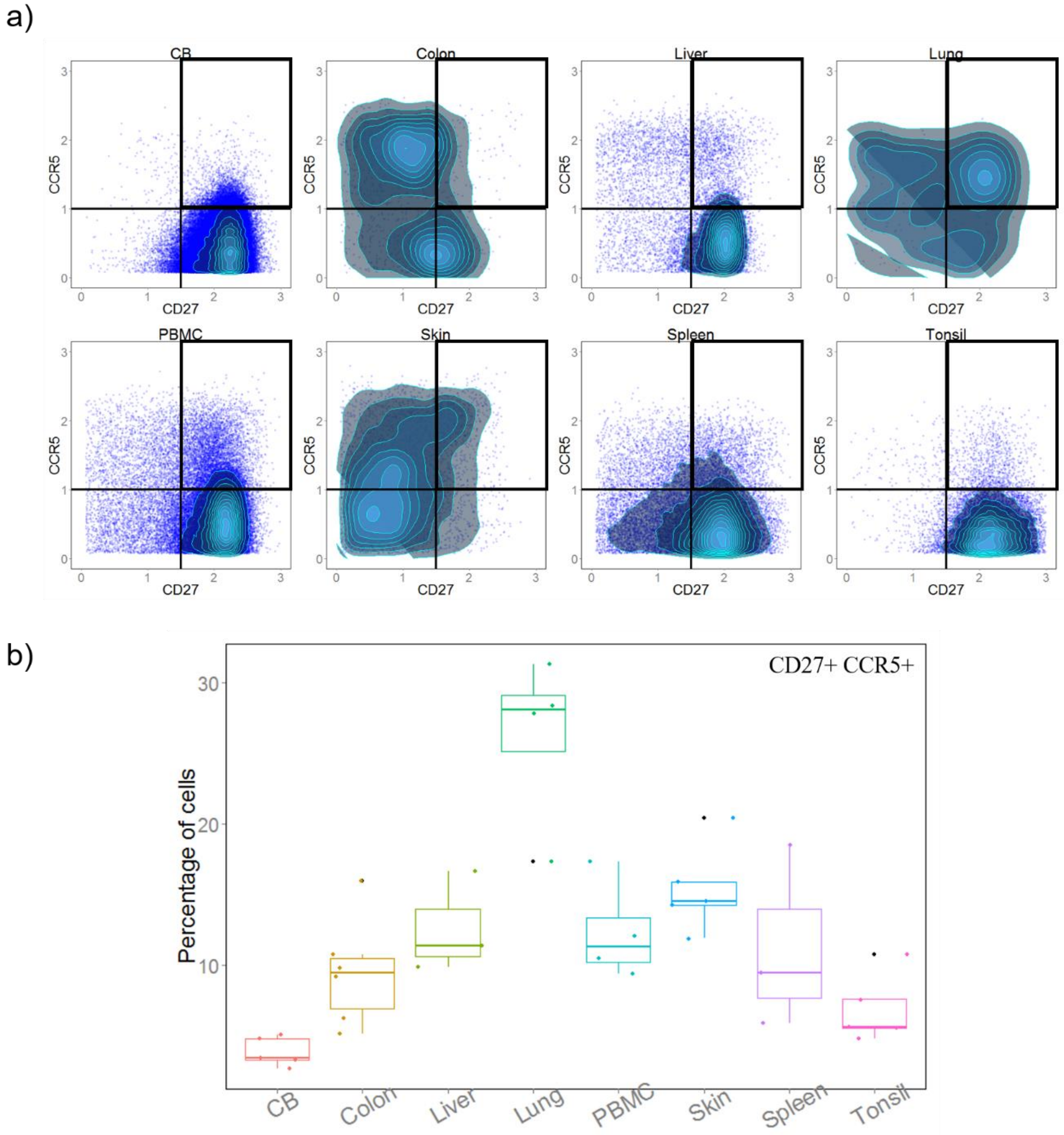


Figure C8. Manual gating validated that CD27$^+$CCR5$^+$ cell subset is highly expressed in lung samples comparing to other types of tissues. a) Example showing manual gating for cell subsets CD27$^+$CCR5$^+$ (highlighted in black boxes) in different types of tissues. b) Boxplot showing percentage of CD27$^+$ CCR5$^+$ cells for each type of tissues.

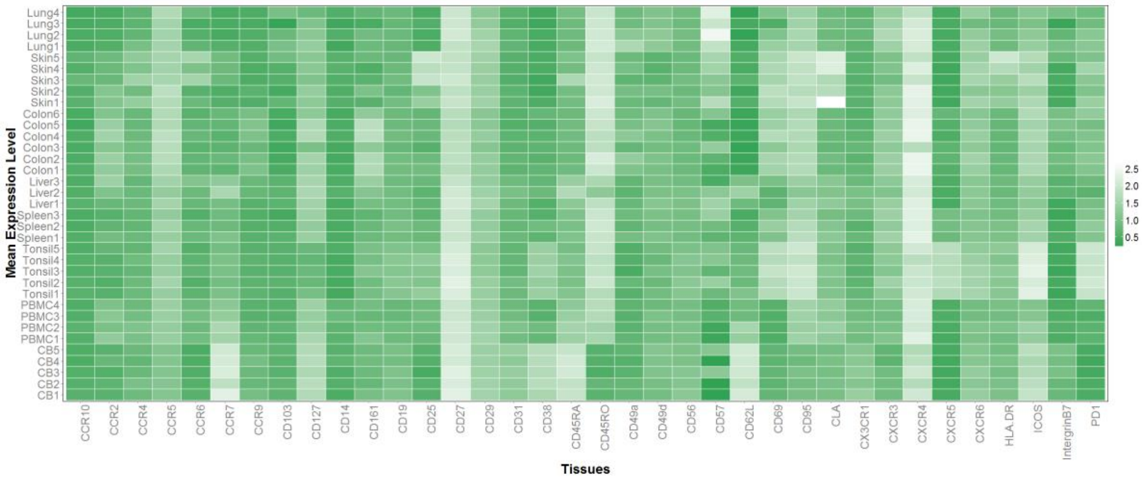


Figure C9. Mean marker expression level of CD27^{+}CCR5^{+} cells. In addition to CD27 and CCR5, these cells also express CD29, CXCR4, and CD45RO (except for cord blood samples).

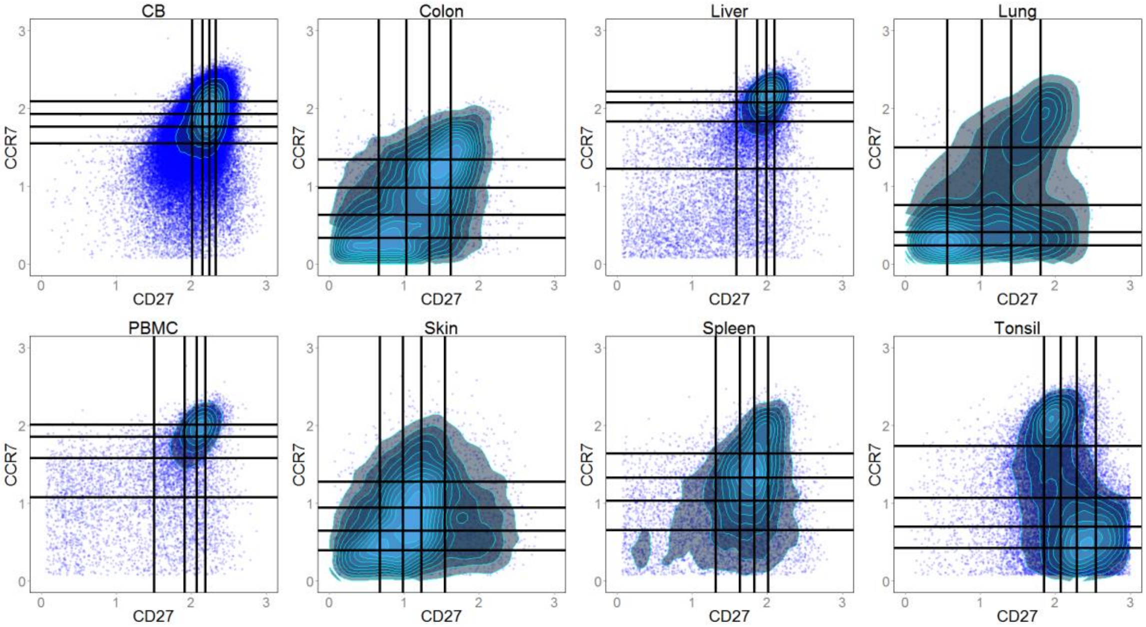


Figure C10. Patterns formed by CCR7 and CD27 expression are distinct among different types of tissues. The black lines are binning thresholds determined by CytoBinning. Five bins are used to divide these patterns.

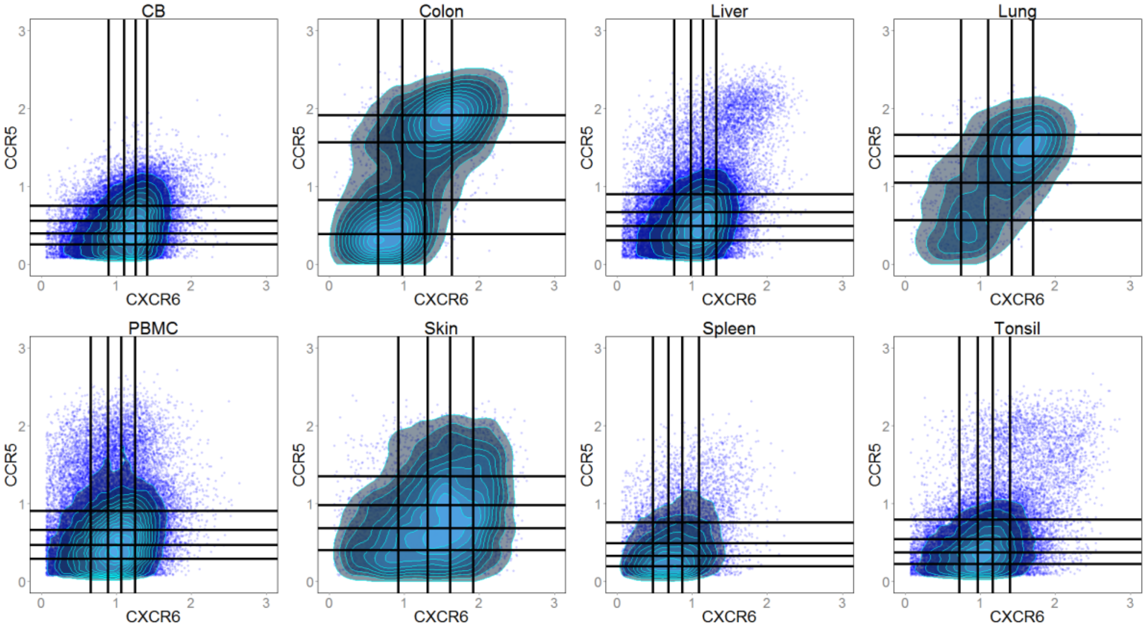


Figure C11. Patterns formed by CCR5 and CXCR6 expression are distinct among different types of tissues. The black lines are binning thresholds determined by CytoBinning. Five bins are used to divide these patterns.

Appendix D Package ‘RefCellScreening’

Type Package
Title Identify important siRNA hits using typical cells
Version 0.0.0.900
Date 2018-04-16
Author Yang Shen <shenyang@umd.edu>
Description This package contains functions for typical cell selection and batch processing of samples in high-throughput screening data
Imports data.table, kernlab, ks
Needs Compilation no

typical_cell_selection *selects typical cells*

Description
This function selects typical cells and returns selected typical cells as a matrix

Usage
```
typical_cell_selection (indata, 1000, features, method='peak')
```

Arguments

input	Input matrix that contains all cells
number	The number of typical cells to select
index	Indices of measurements based on which typical cells are selected, should be a vector of integers
method	Methods to define the center of typical cells, can choose from "mean" and "peak"

Values
A list contains two elements, the first element is a matrix contain typical cells, and the second element is a vector of index for typical cells

appends *append element to a list object*

Description
This function appends a new element to an existing list object at the end (the original list can be empty)

Usage
```
appends(list, to_add)
```

Arguments

lists	The original list

to_add The element to be added to the list

Value
A new list with the element added at the end

combine_data *combine all data files within a certain directory*

Description
This function combines all data files in a given folder

Usage

```
combine_data(directory, pattern = NULL)
```

Arguments

directory The directory of the folder that stores all the data files to be combined

pattern An optional regular expression. Only files match this expression with be combined. The default value is NULL.

Value
A data frame that contains all data

SVM *performs SVM to classify data points in two matrices*

Description
This function applies SVM in kernlab package to classify two groups of data points stored in two matrices and returns the accuracy of SVM classification and the direction of the classification boundary

Usage

```
SVM (healthy,disease,feature_index=c(1:ncol(healthy)))
```

Arguments
healthy Matrix of healthy cells to be classified

disease Matrix of diseased cells to be classified

feature_index A vector contains index of measurements used in classification, should be a vector of integers; length must be larger than 1

Values
accuracy Accuracy of the classification

weightnorm Normalized weights of each feature. Negative weights are higher in disease matrix

center The center (average) of all data points (healthy and disease combined). Can be used to normalize test datasets std The standard deviation of all data points (healthy and disease combined). Can be used to normalize test datasets

SVM_bn adjusted constant for calculating the distance between data point to classification boundary in test data points

SVM_cv *performs SVM classification and testing together*

Description

This function applies SVM in kernlab package to classify two groups of data points stored in two matrices, applies its classification boundary to the testing dataset and outputs the corresponding accuracy.

Usage

SVM_cv(controls,exper,feature_index,test.control,test.exp)

Arguments

controls Training matrix of healthy cells to be classified

exper Training matrix of diseased cells to be classified

feature_index A vector contains index of measurements used in classification, should be a vector of integers; length must be larger than 1

test.control Testing matrix of healthy cells

test.exp Testing matrix of diseased cells

Values

accuracy Classificaiton accuracy of the training dataset

test Classification accuracy of the test dataset

weight Normalized weights of the features

projection *project additional data points along the direction of the classification boundary*

Description

This function projects additional data points along the direction of classification boundary after normalizing it based on the center and std of classification data (used in SVM), calculates the length of projection and percentage of cells along the same direction of the classification boundary

Usage

projection(to_be_normed, center, std, weight)

Arguments

to_be_normed Matrix containing single cells information to be projected center Center of typical cells, output from SVM()

std Standard deviations of typical cells, output from SVM()

weight Direction of boundary plane given by SVM(), weight should be formatted as a n by 1 matrix, n is the number of dimensions

Values

projected A list of projected distance between each test data point to the classification boundary

percent percentage of data points in a to_be_normed matrix that has a positive distance to the classification boundary

---

batch_projection *batch projection of data in siRNA files to classification boundary*

Description

This function projects data in multiple data files to the direction of classification boundary identified in SVM()

Usage

batch_projection(directory,index,center,std,weight)

Arguments

directory Directory of the folder that contains all data files to be projected

index A vector of the index for features to be used for projection

center Center of typical cells, outputs of SVM()

std Standard deviations of typical cells, outputs of SVM()

weights Weights of each feature (i.e., the direction of the classification boundary), outputs of SVM()

Values

percentages A vector of the percentage of cells with a positive projection for each data file

cell.counts A vector of cell numbers in each data file

names A vector of names for each data file

---

hits_identification *Identify important siRNA hits*

Description

This function identifies important siRNA hits as the ones with the percentage of healthy-like cells above a threshold

Usage

```
hits_identification(disease_percent, siRNA_percent, std_siRNA_percent, siRNA_number,
siRNA_name, number_of_plates)
```

Arguments

disease_percent	A vector of healthy-like cell percentage in disease control samples
siRNA_percent	A vector of healthy-like cell percentage in siRNA samples
std_siRNA_percent	A vector of standard deviations of each siRNA sample across replicate experiments
siRNA_number	A vector of the ratio between the number of cells in each siRNA sample and mean cell numbers in healthy control samples
siRNA_name	A character vector contains the names of each siRNA
number_of_plates	The number of replicate experiments

Value

A character vector containing the names of siRNA hits

Appendix E Package ‘CytoBinning’

Type Package
Title CytoBinning
Version 0.0.0.900
Date 2018-04-16
Author Yang Shen <shenyang@umd.edu>
Description This package contains functions to perform CytoBinning
Imports flowCore, kernlab, ks
Needs Compilation no

preprocess *perform compensation and logicle transformation*

Description

This function performs logicle transformation on all FCS files in a folder and outputs transformed data in text format

Usage

```
preprocess = function(input.directory, output.directory, marker.index, marker.names)
```

Arguments

input.directory	Directory of the folder contains all fcs files
output.directory	Directory of the folder to store all transformed data in text format
marker.index	Indices of markers that need to be logicle transformed
marker.names	Names of markers in the same order as the indices

Value

NULL

CytoBinning *perform CytoBinning on all files in a folder*

Description

This function performs CytoBinning on all data files in a folder and output a matrix where each row is the binning result for a data file

Usage

```
CytoBinning(bin.number, marker.names, input.dir, output.dir)
```

Arguments

bin.number	A vector of the selected number of bins
marker.names	A vector contains names of markers to be used in binning
input.dir	Directory of the folder that contains pre-processed text data files

output.dir Directory of the folder that contains binning results, each file stores one marker pair of a given bin number

Value
NULL

SVM	*performs SVM to classify data points in two matrices*

Description
This function applies SVM in kernlab package to classify two groups of data points stored in two matrices and returns the accuracy of SVM classification and the direction of the classification boundary

Usage
```
SVM (healthy,disease,feature_index=c(1:ncol(healthy)))
```

Arguments
healthy Matrix of healthy cells to be classified

disease Matrix of diseased cells to be classified

feature_index A vector contains an index of measurements used in classification, should be a vector of integers; length must be larger than 1

Values
accuracy Accuracy of the classification

weightnorm Normalized weights of each feature. Negative weights are higher in disease matrix

center The center (average) of all data points (healthy and disease combined). Can be used to normalize test datasets

std The standard deviation of all data points (healthy and disease combined). Can be used to normalize test datasets

SVM_bn adjusted constant for calculating the distance between data point to classification boundary in test data points

Appendix F Between and within cell line heterogeneity in breast tumor: insights from tumor cell stiffness

F.1 Overview

The experiments and data pre-processing were performed by members in Professor Josef A. Käs's laboratory at Leipzig University in Germany. I designed downstream analysis approach, performed analysis and wrote the manuscript.

F.2 Abstract

Cellular heterogeneity in tumor cells is a well-established phenomenon. Genetic and phenotypic cell-to-cell variability has been observed in numerous studies both within the same type of cancer cells and across different types of cancers. Here we show that similar heterogeneity can also be seen in mechanical properties of cells both within and between breast tumor cell lines. In particular, we identified two clusters within MDA-231 cells, with cells in one cluster softer than the other. Since stiffness of tumor cells can be an indicator of malignancy potential, this result suggests that metastatic abilities could vary within the same tumor cell line. In addition, we show that MDA-231 and MDA-436 cells are more different from each other than from MCF-10A cells in their mechanical properties, suggesting the existence of different paths to metastasis and the possibility of differentiating and understanding these paths with mechanical properties of single cells.

F.3 Introduction

Recognized as early as 1958 [145], genetic heterogeneity is a well-established phenomenon in tumor cells, especially during metastatic stages [92, 146-148]. Studies have shown that cells from a single type of cancer typically contains multiple genetically distinct subgroups

[149]. Such high level of heterogeneity has been accused of being the reason why cancer is hard to cure [150-152]. However, to-date the reason and extent of tumor cell heterogeneity is still not well-understood [92]. Two main theories have been proposed to explain the origin of tumor cells heterogeneity: cancer stem cell [153] and clonal evolution [150]. These two theories try to explain the heterogeneity in ecological and evolutional aspects respectively, and there has been evidence for each theory [154]. Variations in gene expression lead to molecular variations which in turn affect cellular function and other properties. One of the most important impacts may be cellular stiffness. Changes in actin expression and cytoskeleton organization are related to the fact that metastatic tumor cells are softer than their benign counterparts as well as normal cells [155-157]. Since metastasis is responsible for more than 90% of cancer fatality [158], great effort has been made to study the properties of metastatic tumor cells and how mechanical properties of tumor cell affect its metastatic ability.

Multiple experimental tools have been applied to study cell stiffness, such as atomic force microscope (AFM) [159], quantitative deformability cytometry (q-DC) [160], microfluidic optical cell stretcher [161], etc. Using these tools, a large volume of evidence has been identified linking lower stiffness of tumor cells to higher metastatic ability [155]. In addition, studies have suggested the potential of using mechanical properties as a biomarker of metastasis [162] and for cancer diagnosis [163]. Even though single-cell level measurements can be achieved using these techniques, most studies still perform average based analysis on these data, causing possible oversimplification of the results.

In this paper, we use a microfluidic optical cell stretcher to measure mechanical properties of single cells from three cell lines: MCF-10A, MDA-231, and MDA-436. With single-cell data analysis, we show that heterogeneity of cellular stiffness exists both within and between cell lines. In particular, we observed two groups of MDA-231 cells. Cells in one of the groups are significantly softer than the other. In addition, we found that although MDA-231 and MDA-436 are both triple-negative cell lines with the metastatic tendency, they are quite distinct: more different from each other than from nonmalignant cell line MCF-10A.

F.4 *Results*

We used a Microfluidic Optical Cell Stretcher to mechanically stretch individual tumor cells and measure their stiffness [164, 165]. Suspended single cells were trapped for 1 second and subsequently stretched for 2 seconds and then relaxed in trapping condition for another 2 seconds (Fig F1). Images of cells were taken at the rate of 30 frames per second, and the length of long axis was measured in each frame for each individual cell. In this paper, we use only two mechanical features calculated from these measurements: 1. Normalized long axis deformation at the end of stretch (Deformation EOS), and 2. Normalized long axis deformation during the relaxation period (Relaxation EOE) (Fig F1). The value of EOS is reversely proportional to Young's modulus, where higher EOS value indicates lower Young's modulus (easier to stretch). On the other hand, EOE is a measurement of the ability of a cell to restore its shape, where the higher absolute value of EOE suggests greater ability to restore original shape. Together, these two features give a good representation of the stiffness of a single cell.

Using this technique, we measured cells from three different breast cancer cell lines: MCF-10A, MDA-231, and MDA-436. MCF-10A is a non-malignant and non-metastatic cell line which is often used as a control cell line for cancer cell studies. MDA-231 and MDA-436 are both triple negative breast cancer cell lines (i.e. they do not express estrogen receptors, progesterone receptors nor HER2), they both have metastatic potential with MDA-231 considered more malignant than MDA-436 [166].

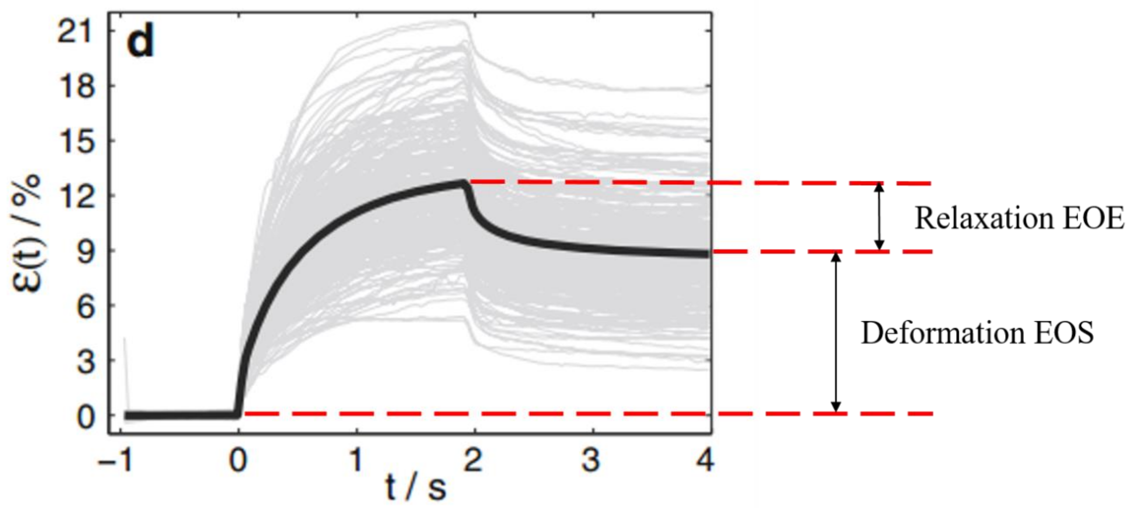


Figure F1. An illustration of the data and features calculated from the data (adapted from [164]).

F.4.1 Two subgroups observed in MDA-231 cells

We first identified two subgroups within MDA-231 cells, one subgroup (cluster 2, Fig F2) have higher deformation at the end of stretch (EOS) and higher absolute value of relaxation at the end of the experiment (EOE) than cluster 1 (Fig F2). Higher absolute values of both EOS and EOE indicates cells in cluster 2 are more elastic (easier to stretch and easier to restore original shape) comparing to cluster 1 which overlaps with MDA-436 and MCF-10A cells (Fig F3a).

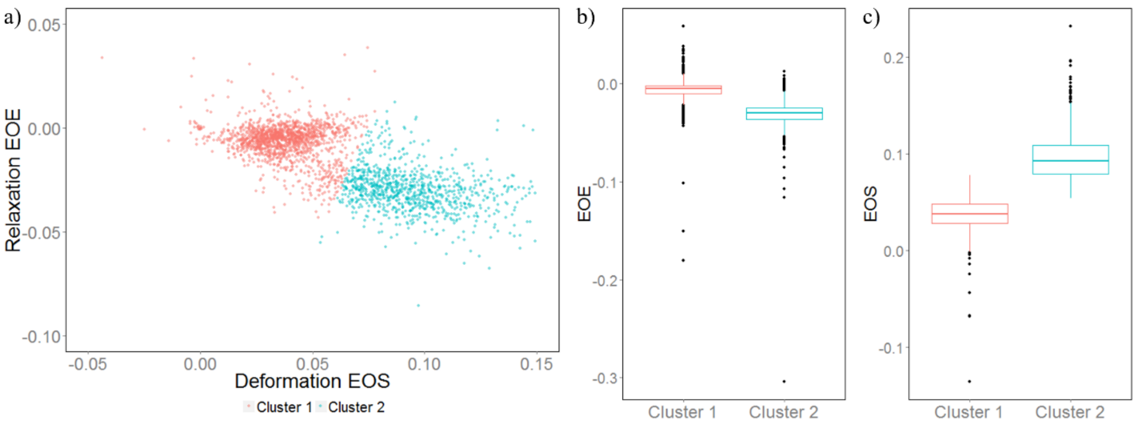


Figure F2. Two clusters of MDA-231 cells are observed. a) Scatterplot of Relaxation EOE vs. Deformation EOS for MDA-231 cells. The two subgroups are labeled by different colors (red: cluster 1, blue: cluster 2). b) Boxplot comparing relaxation at the end of experiment between cluster 1 and cluster 2 of MDA-231 cells. c) Boxplot comparing deformation at the end of stretch between the two subgroups of MDA-231 cells.

F.4.2 The more elastic group does not exist in E-cadherin labeled MDA-231 cells

E-cadherin is an important cellular marker for cell-cell adherence. Metastatic tumor cells like MDA-231 cells express limited E-cadherin and hence are less adhesive. In our experiment, we also measured mechanical properties for MDA-231 cells that are labeled with E-cadherin antibodies. Surprisingly, different properties are observed between the E-cadherin labeled and original MDA-231 cells. In particular, E-cadherin labeled MDA-231 cells only formed one cluster instead of two clusters as is observed in unlabeled MDA-231 cells. The labeled 231 cells overlap with cluster 1 of MDA-231 cells which is the less elastic subgroup (Fig. F3b).

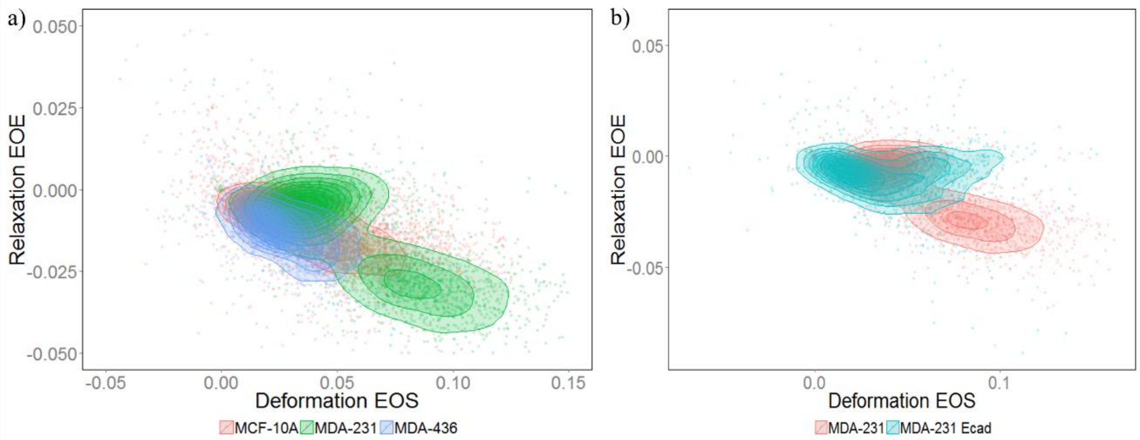


Figure F3. MCF-10A, MDA-436, and E-cadherin labeled MDA-231 cells all overlap with cluster 1 (the less elastic group) in unlabeled MDA-231 cells. a) Scatterplot of Relaxation EOE vs Deformation EOE for MCF-10A (red), MDA-231 (green) and MDA-436 (blue) cells. b) Scatterplot of Relaxation EOE vs. Deformation EOE for E-cadherin labeled (blue) and unlabeled (red) MDA-231 cells.

F.4.3 MDA-231 and MDA-436 cells are more different from each other than from MCF-10A cells

We showed above that there is a subgroup in MDA-231 cells that greatly overlap with MCF-10A and MDA-436 cells. We then move to ask the question that are these overlapping phenotypes similar to each other or are they separable on single cell level? To answer this question, we applied k nearest neighbors (k-NN) algorithm for a pairwise classification of the three phenotypes. We first divided the cells into two groups: train and test. Phenotype labels are provided for cells in train group but not for the test group. Then, given the position of a single cell in the test group, k-NN identifies its nearest k neighbors within the training group. The k neighbors then take a vote with their phenotype, and the cell from test group is assigned to the phenotype that has highest votes. After classification, we calculated the sensitivity (true positive rate) and specificity (true negative rate) for each pair of classification. We found that classification between MCF-10A and MDA-436 cells has the lowest sensitivity and specificity regardless of the value of k (Fig F4a and F4b).

This indicates that the distinction between MCF-10A and MDA-436 cells are less clear than the other two pairs. Indeed, F1 score (a measure of classification result, the higher the score the better the classification, the maximum F1 score is 1) is much lower for MCF-10A and MDA-436 cells for all tested values of k (Fig F4c). This result is expected given that MDA-436 is only a mildly malignant cell line. Since both MDA-436 and MDA-231 cell lines are malignant, it is reasonable to expect they would be similar as well. However, this is not what we observed. Based on the classification results, classification between cluster 1 of MDA-231 and MDA-436 cells has the highest level of F1 scores (Fig F4c) – which is even higher than the classification between MCF-10A and MDA-231 cells. Similar results are obtained with a different classification algorithm (SVM), where the classification between MDA-231 and MDA-436 cells also has the highest F1 value (Table F1). SVM takes a different approach in classification and aims to find the linear plane that best separates two groups to classify. In addition, when all the four phenotypes are classified simultaneously, MDA-436 cells are less likely to be miss-classified as MDA-231 cells than as MCF-10A cells and vice versa (Fig F5). Together, these results suggest that cells in cluster 1 of MDA-231 are more different from MDA-436 cells than from MCF-10A cells.

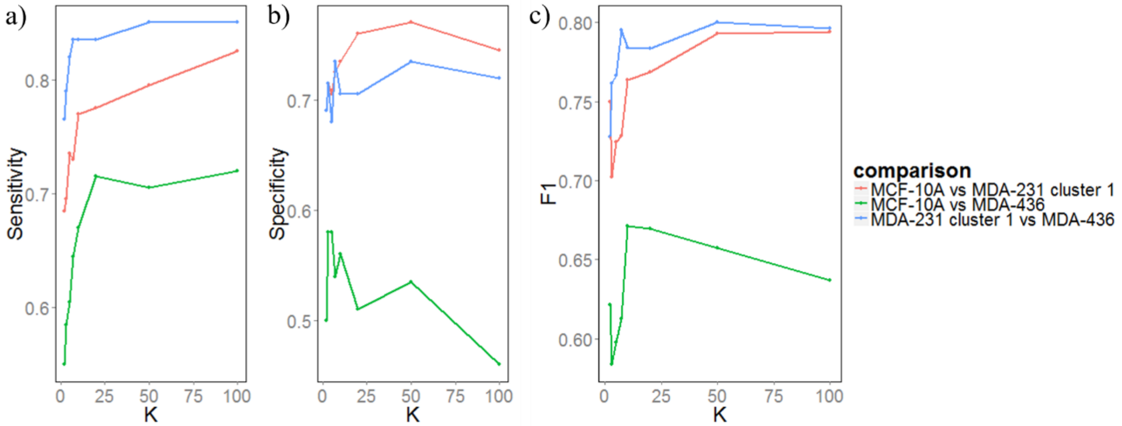


Figure F4. Pairwise k-NN classification results show that MDA-231 and MDA-436 cells are more different from each other than from MCF-10A cells. a) Sensitivity (true positive rate) for the three

comparisons versus different values of k. b) Specificity (true negative rate) for the three comparisons versus different values of k. c) F1 score for the three comparisons versus different values of k.

Table F1. Pairwise classification results by support vector machine (SVM).

	Sensitivity	Specificity	F1
MCF-10A (positive) vs MDA-231 cluster 1 (negative)	0.69	0.68	0.68
MCF-10A (positive) vs MDA-436 (negative)	0.63	0.61	0.60
MDA-231 cluster 1 (positive) vs MDA-436 (negative)	0.76	0.74	0.74

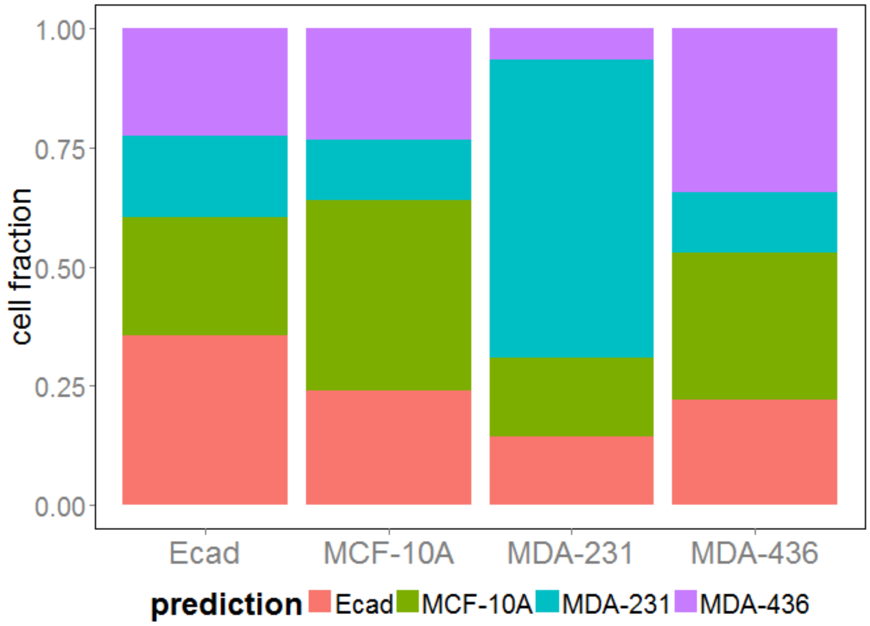


Figure F5. k-NN classification results of E-cadherin added MDA-231 cells (Ecad), MCF-10A cells, cluster 1 in MDA-231 cells and MDA-436 cells, with k = 10. MDA-231 and MDA-436 cells are less likely to be miss-classified as each other than as MCF-10A cells.

F.5 *Discussion*

Mechanical properties of tumors cells are important indicators of their malignancy. Studies have shown that metastatic tumor cells are on average softer than non-metastatic ones. In this paper, we illustrated the heterogeneity of tumor cell stiffness both within and between cell lines. With only mechanical properties, we show that there are two distinct clusters within MDA-231 cells. One of these two clusters greatly overlap with MCF-10A and

MDA-436 cells, and cells in the other cluster are softer (easier to deform and restore to original shape) than the other. In addition, we found that the two malignant cell lines: MDA-231 and MDA-436 are more different from each other than from the nonmalignant MCF-10A cell line.

Heterogeneity within MDA-231 cell line has been shown before based on their molecular expressions. For example, it has been shown that there are two distinct subgroups of MDA-231 cells which differ significantly in the cell surface density of various cytokine receptors (CCR5, CXCR3, CXCR1) [167]. In particular, CXCR3 was found to be overexpressed in metastatic tumor cells, and drugs targeting CXCR3 decreased tumor cell migration [168]. Hence our results agree with previous findings. Additional experiments can be performed to validate the two mechanically distinct subtypes of MDA-231 cells and whether this separation agrees with their cytokine receptor expression.

We also identified heterogeneity between triple negative breast cancer (TNBC) cell lines, i.e., we found that MDA-231 and MDA-436 cells are quite distinct from each other, even more so than from the nonmalignant MCF-10A cell line. This finding is feasible from the perspective of the classical clonal evolution model. Assuming the genetic and (more importantly) the phenotypic characteristics of normal breast tissue are similar among all women. Thus, both patients from which the MDA-231 and MDA-436 cell lines are derived had initially breast tissue which is very similar to the MCF-10A tissue. From this healthy starting population of cells, different paths can be taken to reach a metastatic phenotype. It seems actually unlikely that two completely different patients accumulate the exact same cancer cell phenotype with the same optical stretching characteristics. What’s more, our findings may have important clinical implications. Patients with triple-negative breast cancer are currently considered to have a very poor prognosis. However, there has been an emerging trend to regard TNBC as a heterogeneous group of patients with some actually having not such a bad prognosis. Furthermore, TNBCs can have very different molecular characteristics, potentially rendering some tumors more suitable to targeted therapies. It is of paramount clinical importance to identify those patients. The present data is intriguing in that it shows, that two TNBC cell lines (which would be put into one prognostic basket clinically) are indeed very different. It is interesting to speculate whether optical stretching analysis could be used to differentiate those TNBC cases with a better prognosis (i.e., a lower rate of relapse and distant metastasis) from those with a worse prognosis.

In addition, our findings on the between cell line heterogeneity is an indication that average based analysis method would oversimplify tumor cell data. If one only looks at the scatterplot as in figure 3a, one would conclude that MCF-10A, MDA-436 and cluster 1 of MDA-231 cells are similar to each other with the probably minor difference in the average values. However, when classified with a more sophisticated algorithm like k-NN, reasonably good classification accuracy was achieved. That is to say, even though cells from the three cell lines overlap on at a higher level, locally cells from a certain cell line are more close to cells from the same cell line than from other cell lines. This cannot be found if averaging methods were used.

Lastly, we found that E-cadherin labeled MDA-231 cells have different phenotype profiles comparing to the unlabeled cells. We reason that this is because binding of the antibody to the E-cadherin receptor simulates cell-cell binding, which causes cadherin clustering and stimulates the actin cortex bound to cadherin. This is a good example of how antibody labeling would change the properties of cells be labeled, and a label-free measuring mechanism would give better / more accurate results.

In all, we illustrated within and between cell line tumor cell heterogeneity with only cellular mechanical properties. Our results have great implications for future studies on how a change in chemokine receptor expression correlates with tumor cell stiffness and how differentiating mechanical properties of cancer cells could help identify triple negative breast cancer patients with better prognosis.

F.6 Methods

F.6.1 Experimental procedures

Suspended single cells were trapped for 1 second and subsequently stretched for 2 seconds and then relaxed in trapping condition for another 2 seconds (Fig F1). Images of cells were taken at the rate of 30 frames per second, and the length of long axis was measured in each frame for each individual cell.

F.6.2 Data analysis

The two clusters of MDA-231 cells were clustered using kmeans() function in R (version 3.0.3) with 2 centers, 1,000 iterations, and 50 random initial conditions. For kNN classification, 1,200 cells were first randomly selected from each cell line. From the 1,200 cells, 200 were randomly selected as a testing set, and the remaining 1,000 were used as a training set for each cell line. The classification was done separately for each pair of cell line using knn() function in R with 8 different values of k (2, 3, 5, 7, 10, 20, 50, 100). Similarly, simultaneous classification of the four cell lines was done. After classification, false positive rate was calculated as FPR = (false positives) / (false positives + true positives). And false negative rate (FNR) = (false negatives) / (false negatives + true negatives). Finally, pairwise SVM classifications were done based on all the 1,200 randomly selected cells using ksvm() function with linear kernel and C=10 in R package kernlab. All plots were made with the ggplot2 package in R.

Bibliography

1. Chattopadhyay, P.K. and M. Roederer, *A mine is a terrible thing to waste: high content, single cell technologies for comprehensive immune analysis.* Am J Transplant, 2015. **15**(5): p. 1155-61.
2. Gattinoni, L., et al., *A human memory T cell subset with stem cell-like properties.* Nat Med, 2011. **17**(10): p. 1290-7.
3. Johnston, R.J., et al., *Bcl6 and Blimp-1 are reciprocal and antagonistic regulators of T follicular helper cell differentiation.* Science, 2009. **325**(5943): p. 1006-10.
4. Park, H., et al., *A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17.* Nat Immunol, 2005. **6**(11): p. 1133-41.
5. Saeys, Y., S.V. Gassen, and B.N. Lambrecht, *Computational flow cytometry: helping to make sense of high-dimensional immunology data.* Nat Rev Immunol, 2016. **16**(7): p. 449-62.
6. Tanner, S.D., et al., *An introduction to mass cytometry: fundamentals and applications.* Cancer Immunol Immunother, 2013. **62**(5): p. 955-65.
7. Genovesio, A., et al., *Automated genome-wide visual profiling of cellular proteins involved in HIV infection.* J Biomol Screen, 2011. **16**(9): p. 945-58.
8. Kalisky, T., et al., *A brief review of single-cell transcriptomic technologies.* Brief Funct Genomics, 2018. **17**(1): p. 64-76.
9. Altschuler, S.J. and L.F. Wu, *Cellular heterogeneity: do differences make a difference?* Cell, 2010. **141**(4): p. 559-63.
10. Aggarwal, C.C., A. Hinneburg, and D.A. Keim. *On the Surprising Behavior of Distance Metrics in High Dimensional Space*. 2001. Berlin, Heidelberg %@ 978-3-540-44503-6: Springer Berlin Heidelberg.
11. Orlova, D.Y., L.A. Herzenberg, and G. Walther, *Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets.* Nature Reviews Immunology, 2018. **18**(1): p. 77 %@ 1474-1741.
12. Macarron, R., et al., *Impact of high-throughput screening in biomedical research.* Nat Rev Drug Discov, 2011. **10**(3): p. 188-95.
13. Inglese, J., C.E. Shamu, and R.K. Guy, *Reporting data from high-throughput screening of small-molecule libraries.* Nat Chem Biol, 2007. **3**(8): p. 438-41.
14. Mohr, S., C. Bakal, and N. Perrimon, *Genomic screening with RNAi: results and challenges.* Annu Rev Biochem, 2010. **79**: p. 37-64.
15. Boutros, M., F. Heigwer, and C. Laufer, *Microscopy-Based High-Content Screening.* Cell, 2015. **163**(6): p. 1314-25.
16. Mattiazzi Usaj, M., et al., *High-Content Screening for Quantitative Cell Biology.* Trends Cell Biol, 2016. **26**(8): p. 598-611.
17. Zanella, F., J.B. Lorens, and W. Link, *High content screening: seeing is believing.* Trends Biotechnol, 2010. **28**(5): p. 237-245 %@ 0167-7799.
18. Motti, D., et al., *High Content Screening of Mammalian Primary Cortical Neurons.* Methods Mol Biol, 2018. **1683**: p. 293-304.
19. Zhou, T., et al., *High-Content Screening in hPSC-Neural Progenitors Identifies Drug Candidates that Inhibit Zika Virus Infection in Fetal-like Organoids and Adult Brain.* Cell Stem Cell, 2017. **21**(2): p. 274-283 e5.

20. Johnston, R.L., et al., *High content screening application for cell-type specific behaviour in heterogeneous primary breast epithelial subpopulations.* Breast Cancer Res, 2016. **18**(1): p. 18.
21. Tolosa, L., M.J. Gomez-Lechon, and M.T. Donato, *High-content screening technology for studying drug-induced hepatotoxicity in cell models.* Arch Toxicol, 2015. **89**(7): p. 1007-22.
22. Collins, T.J., *ImageJ for microscopy.* Biotechniques, 2007. **43**(1 Suppl): p. 25-30.
23. Schneider, C.A., W.S. Rasband, and K.W. Eliceiri, *NIH Image to ImageJ: 25 years of image analysis.* Nature Methods, 2012. **9**(7): p. 671-5.
24. Schindelin, J., et al., *Fiji: an open-source platform for biological-image analysis.* Nature Methods, 2012. **9**(7): p. 676-82.
25. Carpenter, A.E., et al., *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome Biol, 2006. **7**(10): p. R100.
26. Lamprecht, M.R., D.M. Sabatini, and A.E. Carpenter, *CellProfiler: free, versatile software for automated biological image analysis.* Biotechniques, 2007. **42**(1): p. 71-5.
27. Soliman, K., *CellProfiler: Novel Automated Image Segmentation Procedure for Super-Resolution Microscopy.* Biol Proced Online, 2015. **17**: p. 11.
28. Kamentsky, L., et al., *Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software.* Bioinformatics, 2011. **27**(8): p. 1179-80.
29. Eliceiri, K.W., et al., *Biological imaging software tools.* Nature Methods, 2012. **9**(7): p. 697-710.
30. Singh, S., A.E. Carpenter, and A. Genovesio, *Increasing the Content of High-Content Screening: An Overview.* J Biomol Screen, 2014. **19**(5): p. 640-50.
31. Grys, B.T., et al., *Machine learning and computer vision approaches for phenotypic profiling.* J Cell Biol, 2017. **216**(1): p. 65-71.
32. Loo, L.H., L.F. Wu, and S.J. Altschuler, *Image-based multivariate profiling of drug responses from single cells.* Nature Methods, 2007. **4**(5): p. 445-453.
33. Shariff, A., et al., *Automated image analysis for high-content screening and analysis.* J Biomol Screen, 2010. **15**(7): p. 726-34.
34. Slack, M.D., et al., *Characterizing heterogeneous cellular responses to perturbations.* Proc Natl Acad Sci U S A, 2008. **105**(49): p. 19306-11.
35. Kraus, O.Z., et al., *Automated analysis of high-content microscopy data with deep learning.* Mol Syst Biol, 2017. **13**(4): p. 924.
36. Aghaeepour, N. and R. Brinkman, *Computational analysis of high-dimensional flow cytometric data for diagnosis and discovery.* Curr Top Microbiol Immunol, 2014. **377**: p. 159-75.
37. Chattopadhyay, P.K. and M. Roederer, *Cytometry: today's technology and tomorrow's horizons.* Methods, 2012. **57**(3): p. 251-8.
38. Adan, A., et al., *Flow cytometry: basic principles and applications.* Crit Rev Biotechnol, 2017. **37**(2): p. 163-176.
39. Shapiro, H.M., *Practical Flow Cytometry*2005: John Wiley & Sons, Inc.
40. Spidlen, J., et al., *Data File Standard for Flow Cytometry, version FCS 3.1.* Cytometry A, 2010. **77**(1): p. 97-100.

41. Seamer, L.C., et al., *Proposed new data file standard for flow cytometry, version FCS 3.0.* Cytometry, 1997. **28**(2): p. 118-22.
42. Herzenberg, L.A., et al., *Interpreting flow cytometry data: a guide for the perplexed.* Nat Immunol, 2006. **7**(7): p. 681-5.
43. Parks, D.R., M. Roederer, and W.A. Moore, *A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data.* Cytometry A, 2006. **69**(6): p. 541-51.
44. Hahne, F., et al., *Per-channel basis normalization methods for flow cytometry data.* Cytometry A, 2010. **77**(2): p. 121-31.
45. Hahne, F., et al., *flowCore: a Bioconductor package for high throughput flow cytometry.* BMC Bioinformatics, 2009. **10**: p. 106.
46. Maecker, H.T., J.P. McCoy, and R. Nussenblatt, *Standardizing immunophenotyping for the Human Immunology Project.* Nat Rev Immunol, 2012. **12**(3): p. 191-200.
47. Pachon, G., I. Caragol, and J. Petriz, *Subjectivity and flow cytometric variability.* Nat Rev Immunol, 2012. **12**(5): p. 396; author reply 396.
48. Gouttefangeas, C., et al., *Data analysis as a source of variability of the HLA-peptide multimer assay: from manual gating to automated recognition of cell clusters.* Cancer Immunol Immunother, 2015. **64**(5): p. 585-98.
49. Pedreira, C.E., et al., *Overview of clinical flow cytometry data analysis: recent advances and future challenges.* Trends Biotechnol, 2013. **31**(7): p. 415-25.
50. Lugli, E., M. Roederer, and A. Cossarizza, *Data analysis in flow cytometry: the future just started.* Cytometry A, 2010. **77**(7): p. 705-13.
51. Kvistborg, P., et al., *Thinking outside the gate: single-cell assessments in multiple dimensions.* Immunity, 2015. **42**(4): p. 591-2.
52. Bashashati, A. and R.R. Brinkman, *A survey of flow cytometry data analysis methods.* Adv Bioinformatics, 2009: p. 584603.
53. Aghaeepour, N., et al., *Critical assessment of automated flow cytometry data analysis techniques.* Nature Methods, 2013. **10**(3): p. 228-38.
54. Malek, M., et al., *flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification.* Bioinformatics, 2015. **31**(4): p. 606-7.
55. Aghaeepour, N., et al., *Early immunologic correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays.* Bioinformatics, 2012. **28**(7): p. 1009-16.
56. Ge, Y. and S.C. Sealfon, *flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding.* Bioinformatics, 2012. **28**(15): p. 2052-8.
57. Aghaeepour, N., et al., *Rapid cell population identification in flow cytometry data.* Cytometry A, 2011. **79**(1): p. 6-13.
58. Pyne, S., et al., *Automated high-dimensional flow cytometric data analysis.* Proc Natl Acad Sci U S A, 2009. **106**(21): p. 8519-24.
59. Rebhahn, J.A., et al., *Competitive SWIFT cluster templates enhance detection of aging changes.* Cytometry A, 2016. **89**(1): p. 59-70.
60. Lo, K., et al., *flowClust: a Bioconductor package for automated gating of flow cytometry data.* BMC Bioinformatics, 2009. **10**: p. 145.

61. Lo, K., R.R. Brinkman, and R. Gottardo, *Automated gating of flow cytometry data via robust model-based clustering.* Cytometry A, 2008. **73**(4): p. 321-32.

62. Van Gassen, S., et al., *FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data.* Cytometry A, 2015. **87**(7): p. 636-45.

63. Van Gassen, S., et al., *FloReMi: Flow density survival regression using minimal feature redundancy.* Cytometry A, 2016. **89**(1): p. 22-9.

64. Anchang, B., et al., *CCAST: a model-based gating strategy to isolate homogeneous subpopulations in a heterogeneous population of single cells.* PLoS Comput Biol, 2014. **10**(7): p. e1003664.

65. O'Neill, K., et al., *Deep profiling of multitube flow cytometry data.* Bioinformatics, 2015. **31**(10): p. 1623-31.

66. Roederer, M., et al., *Probability binning comparison: a metric for quantitating multivariate distribution differences.* Cytometry, 2001. **45**(1): p. 47-55.

67. Roederer, M. and R.R. Hardy, *Frequency difference gating: a multivariate method for identifying subsets that differ between samples.* Cytometry, 2001. **45**(1): p. 56-64.

68. Finak, G., et al., *OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis.* PLoS Comput Biol, 2014. **10**(8).

69. O'Neill, K., et al., *Enhanced flowType/RchyOptimyx: a BioConductor pipeline for discovery in high-dimensional cytometry data.* Bioinformatics, 2014. **30**(9): p. 1329-30.

70. Aghaeepour, N., et al., *RchyOptimyx: cellular hierarchy optimization for flow cytometry.* Cytometry A, 2012. **81**(12): p. 1022-30.

71. Courtot, M., et al., *flowCL: ontology-based cell population labelling in flow cytometry.* Bioinformatics, 2015. **31**(8): p. 1337-9.

72. Lin, L., et al., *Identification and visualization of multidimensional antigen-specific T-cell populations in polychromatic cytometry data.* Cytometry A, 2015. **87**(7): p. 675-82.

73. Amir el, A.D., et al., *viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia.* Nat Biotechnol, 2013. **31**(6): p. 545-52.

74. Carter, K.M., et al., *Information preserving component analysis: Data projections for flow cytometry analysis.* IEEE Journal of Selected Topics in Signal Processing, 2009. **3**(1): p. 148-158 %@ 1932-4553.

75. Qiu, P., et al., *Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE.* Nat Biotechnol, 2011. **29**(10): p. 886-91.

76. Nicol, B., et al., *An intermediate level of CD161 expression defines a novel activated, inflammatory, and pathogenic subset of CD8(+) T cells involved in multiple sclerosis.* J Autoimmun, 2017.

77. Beyer, K.G., J.; Ramakrishnan, R.; Shaft, U., *When is "nearest neighbor" meaningful?* Lecture Notes in Computer Science, 1999. **1540**: p. 217-235.

78. Bandura, D.R., et al., *Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry.* Anal Chem, 2009. **81**(16): p. 6813-22.

79. Spitzer, M.H. and G.P. Nolan, *Mass Cytometry: Single Cells, Many Features.* Cell, 2016. **165**(4): p. 780-91.

80. Ornatsky, O., et al., *Highly multiparametric analysis by mass cytometry.* J Immunol Methods, 2010. **361**(1-2): p. 1-20.

81. Bendall, S.C., et al., *A deep profiler's guide to cytometry.* Trends Immunol, 2012. **33**(7): p. 323-32.

82. Bendall, S.C., et al., *Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum.* Science, 2011. **332**(6030): p. 687-96.

83. Horowitz, A., et al., *Genetic and environmental determinants of human NK cell diversity revealed by mass cytometry.* Sci Transl Med, 2013. **5**(208): p. 208ra145.

84. Lujan, E., et al., *Early reprogramming regulators identified by prospective isolation and mass cytometry.* Nature, 2015. **521**(7552): p. 352-6.

85. Bruggner, R.V., et al., *Automated identification of stratifying signatures in cellular subpopulations.* Proc Natl Acad Sci U S A, 2014. **111**(26): p. E2770-7.

86. Levine, J.H., et al., *Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis.* Cell, 2015. **162**(1): p. 184-97.

87. Abraham, Y., et al., *Exploring Glucocorticoid Receptor Agonists Mechanism of Action Through Mass Cytometry and Radial Visualizations.* Cytometry B Clin Cytom, 2017. **92**(1): p. 42-56.

88. Kunicki, M.A., et al., *Identity and Diversity of Human Peripheral Th and T Regulatory Cells Defined by Single-Cell Mass Cytometry.* J Immunol, 2018. **200**(1): p. 336-346.

89. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.

90. Cheng, Y., et al., *Categorical Analysis of Human T Cell Heterogeneity with One-Dimensional Soli-Expression by Nonlinear Stochastic Embedding.* J Immunol, 2016. **196**(2): p. 924-32.

91. Dexter, D.L., et al., *Heterogeneity of tumor cells from a single mouse mammary tumor.* Cancer Res, 1978. **38**(10): p. 3174-81.

92. Alizadeh, A.A., et al., *Toward understanding and exploiting tumor heterogeneity.* Nat Med, 2015. **21**(8): p. 846-53.

93. Fodor, I.K., *A survey of dimension reduction techniques*, 2002, Lawrence Livermore National Lab., CA (US).

94. Houle, M.E., et al. *Can shared-neighbor distances defeat the curse of dimensionality?* in *International Conference on Scientific and Statistical Database Management*. 2010. Springer.

95. Kiefer, J., et al., *High-throughput siRNA screening as a method of perturbation of biological systems and identification of targeted pathways coupled with compound screening.* Methods Mol Biol, 2009. **563**: p. 275-87.

96. Varma, H., D.C. Lo, and B.R. Stockwell, *High-Throughput and High-Content Screening for Huntington's Disease Therapeutics*, in *Neurobiology of Huntington's Disease: Applications to Drug Discovery*, D.C. Lo and R.E. Hughes, Editors. 2011: Boca Raton (FL).

97. Liberali, P., B. Snijder, and L. Pelkmans, *Single-cell and multivariate approaches in genetic perturbation screens.* Nat Rev Genet, 2015. **16**(1): p. 18-32.

98. Kozak, K., et al., *Data mining techniques in high content screening: a survey.* J Comput Sci Syst Biol, 2009. **2**(04): p. 219-39.

99. Meijering, E., et al., *Imagining the future of bioimage analysis.* Nat Biotechnol, 2016. **34**(12): p. 1250-1255.

100. Jones, T.R., et al., *Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning.* Proc Natl Acad Sci U S A, 2009. **106**(6): p. 1826-31.

101. Ramo, P., et al., *CellClassifier: supervised learning of cellular phenotypes.* Bioinformatics, 2009. **25**(22): p. 3028-30.

102. Horvath, P., et al., *Machine Learning Improves the Precision and Robustness of High-Content Screens: Using Nonlinear Multiparametric Methods to Analyze Screening Results.* J Biomol Screen, 2011. **16**(9): p. 1059-1067.

103. Zhong, R., et al., *iScreen: Image-Based High-Content RNAi Screening Analysis Tools.* J Biomol Screen, 2015. **20**(8): p. 998-1002.

104. Perlman, Z.E., et al., *Multidimensional drug profiling by automated microscopy.* Science, 2004. **306**(5699): p. 1194-1198.

105. Jones, T.R., et al. *Methods for high-content, high-throughput image-based cell screening*. in *Proceedings of the Workshop on Microscopic Image Analysis with Applications in Biology*. 2006.

106. Birmingham, A., et al., *Statistical methods for analysis of high-throughput RNA interference screens.* Nature Methods, 2009. **6**(8): p. 569-75.

107. Kummel, A., et al., *Comparison of multivariate data analysis strategies for high-content screening.* J Biomol Screen, 2011. **16**(3): p. 338-47.

108. Verschuuren, M., et al., *Accurate Detection of Dysmorphic Nuclei Using Dynamic Programming and Supervised Classification.* PLoS One, 2017. **12**(1).

109. Kubben, N., et al., *Repression of the Antioxidant NRF2 Pathway in Premature Aging.* Cell, 2016. **165**(6): p. 1361-1374.

110. Capell, B.C. and F.S. Collins, *Human laminopathies: nuclei gone genetically awry.* Nat Rev Genet, 2006. **7**(12): p. 940-52.

111. Capell, B.C., et al., *Inhibiting farnesylation of progerin prevents the characteristic nuclear blebbing of Hutchinson-Gilford progeria syndrome.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12879-84.

112. Kudlow, B.A., B.K. Kennedy, and R.J. Monnat Jr, *Werner and Hutchinson–Gilford progeria syndromes: mechanistic basis of human progeroid diseases.* Nature Reviews Molecular Cell Biology, 2007. **8**(5): p. 394.

113. Brassard, J.A., et al., *Hutchinson–Gilford progeria syndrome as a model for vascular aging.* Biogerontology, 2016. **17**(1): p. 129-145.

114. Scaffidi, P. and T. Misteli, *Reversal of the cellular phenotype in the premature aging disease Hutchinson-Gilford progeria syndrome.* Molecular Biology of the Cell, 2004. **15**: p. 120a-120a.

115. Zwerger, M., C.Y. Ho, and J. Lammerding, *Nuclear Mechanics in Disease.* Annual Review of Biomedical Engineering, Vol 13, 2011. **13**: p. 397-428.

116. Allsopp, R.C., et al., *Telomere Length Predicts Replicative Capacity of Human Fibroblasts.* Proc Natl Acad Sci U S A, 1992. **89**(21): p. 10114-10118.

117. Cao, K., et al., *Progerin and telomere dysfunction collaborate to trigger cellular senescence in normal human fibroblasts.* Journal of Clinical Investigation, 2011. **121**(7): p. 2833-2844.

118. Goldman, R.D., et al., *Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson–Gilford progeria syndrome.* Proc Natl Acad Sci U S A, 2004. **101**(24): p. 8963-8968.

119. Liu, Y.Y., et al., *DNA damage responses in progeroid syndromes arise from defective maturation of prelamin A.* Journal of Cell Science, 2006. **119**(22): p. 4644-4649.

120. Kubben, N., et al., *A high-content imaging-based screening pipeline for the systematic identification of anti-progeroid compounds.* Methods, 2016. **96**: p. 46-58.

121. Candia, J., et al., *From Cellular Characteristics to Disease Diagnosis: Uncovering Phenotypes with Supercells.* PLoS Comput Biol, 2013. **9**(9).

122. Goransson, H., et al., *Quantification of normal cell fraction and copy number neutral LOH in clinical lung cancer samples using SNP array data.* PLoS One, 2009. **4**(6): p. e6057.

123. Driscoll, M.K., et al., *Automated image analysis of nuclear shape: what can we learn from a prematurely aged cell?* Aging (Albany NY), 2012. **4**(2): p. 119-32.

124. Burges, C.J.C., *A tutorial on Support Vector Machines for pattern recognition.* Data Mining and Knowledge Discovery, 1998. **2**(2): p. 121-167.

125. Tung, J.W., et al., *Modern flow cytometry: a practical approach.* Clin Lab Med, 2007. **27**(3): p. 453-68, v.

126. Baumgarth, N. and M. Roederer, *A practical approach to multicolor flow cytometry for immunophenotyping.* J Immunol Methods, 2000. **243**(1-2): p. 77-97.

127. Finak, G., et al., *Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium.* Sci Rep, 2016. **6**: p. 20686.

128. Santegoets, S.J., et al., *Monitoring regulatory T cells in clinical samples: consensus on an essential marker set and gating strategy for regulatory T cell analysis by flow cytometry.* Cancer Immunol Immunother, 2015. **64**(10): p. 1271-86.

129. Pitoiset, F., et al., *A standardized flow cytometry procedure for the monitoring of regulatory T cells in clinical trials.* Cytometry Part B: Clinical Cytometry %@ 1552-4957, 2018.

130. Finak, G., et al., *OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis.* PLoS Comput Biol, 2014. **10**(8): p. e1003806.

131. Spidlen, J., et al., *Gating-ML: XML-based gating descriptions in flow cytometry.* Cytometry A, 2008. **73A**(12): p. 1151-7.

132. Mosmann, T.R., et al., *SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 2: biological evaluation.* Cytometry A, 2014. **85**(5): p. 422-33.

133. Qian, Y., et al., *Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data.* Cytometry B Clin Cytom, 2010. **78 Suppl 1**: p. S69-82.

134. Wong, M.T., et al., *A High-Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine Signatures.* Immunity, 2016. **45**(2): p. 442-56.

135. Bialek, N.S.G.S.A.G.T.W., *Estimating mutual information and multi–information in large networks.* ArXiv preprint, 2005.

136. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S7.

137. Naim, I., et al., *SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: algorithm design.* Cytometry A, 2014. **85**(5): p. 408-21.

138. Weiskopf, D., B. Weinberger, and B. Grubeck-Loebenstein, *The aging of the immune system.* Transpl Int, 2009. **22**(11): p. 1041-50.

139. Nelson, B.H., *IL-2, regulatory T cells, and tolerance.* J Immunol, 2004. **172**(7): p. 3983-8.

140. Spidlen, J., et al., *FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications.* Cytometry A, 2012. **81**(9): p. 727-31.

141. Di Palma, S. and B. Bodenmiller, *Unraveling cell populations in tumors by single-cell mass cytometry.* Curr Opin Biotechnol, 2015. **31**: p. 122-9.

142. Orlova, D.Y., L.A. Herzenberg, and G. Walther, *Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets.* Nature Reviews Immunology, 2017. **18**: p. 77.

143. Patin, E., et al., *Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors.* Nat Immunol, 2018. **19**(3): p. 302.

144. Chen, Y.-W. and C.-J. Lin, *Combining SVMs with various feature selection strategies*, in *Feature extraction*2006, Springer. p. 315-324.

145. Huxley, J., *Biological aspects of cancer*1958: Harcourt, Brace.

146. Torres, L., et al., *Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases.* Breast Cancer Res Treat, 2007. **102**(2): p. 143-55.

147. Park, S.Y., et al., *Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype.* Journal of Clinical Investigation, 2010. **120**(2): p. 636-44.

148. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.

149. Cleary, A.S., et al., *Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers.* Nature, 2014. **508**(7494): p. 113-7.

150. McGranahan, N. and C. Swanton, *Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future.* Cell, 2017. **168**(4): p. 613-628.

151. Mann, K.M., et al., *Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq.* Nat Biotechnol, 2016. **34**(9): p. 962-72.

152. Koren, S. and M. Bentires-Alj, *Breast Tumor Heterogeneity: Source of Fitness, Hurdle for Therapy.* Mol Cell, 2015. **60**(4): p. 537-46.

153. Magee, J.A., E. Piskounova, and S.J. Morrison, *Cancer stem cells: impact, heterogeneity, and uncertainty.* Cancer Cell, 2012. **21**(3): p. 283-96.

154. Shackleton, M., et al., *Heterogeneity in cancer: cancer stem cells versus clonal evolution.* Cell, 2009. **138**(5): p. 822-9.

155. Lekka, M., et al., *Cancer cell recognition--mechanical phenotype.* Micron, 2012. **43**(12): p. 1259-66.

156. Plodinec, M., et al., *The nanomechanical signature of breast cancer.* Nat Nanotechnol, 2012. **7**(11): p. 757-65.

157. Swaminathan, V., et al., *Mechanical stiffness grades metastatic potential in patient tumor cells and in cancer cell lines.* Cancer Res, 2011. **71**(15): p. 5075-80.

158. Wirtz, D., K. Konstantopoulos, and P.C. Searson, *The physics of cancer: the role of physical interactions and mechanical forces in metastasis.* Nat Rev Cancer, 2011. **11**(7): p. 512-22.

159. Hayashi, K. and M. Iwata, *Stiffness of cancer cells measured with an AFM indentation method.* J Mech Behav Biomed Mater, 2015. **49**: p. 105-11.

160. Nyberg, K.D., et al., *Quantitative Deformability Cytometry: Rapid, Calibrated Measurements of Cell Mechanical Properties.* Biophysical Journal, 2017. **113**(7): p. 1574-1584.

161. Farzbod, A. and H. Moon, *Integration of reconfigurable potentiometric electrochemical sensors into a digital microfluidic platform.* Biosens Bioelectron, 2018. **106**: p. 37-42.

162. Xu, W., et al., *Cell stiffness is a biomarker of the metastatic potential of ovarian cancer cells.* PLoS One, 2012. **7**(10): p. e46609.

163. Remmerbach, T.W., et al., *Oral cancer diagnosis by mechanical phenotyping.* Cancer Res, 2009. **69**(5): p. 1728-32.

164. Kiessling, T.R., et al., *Analysis of multiple physical parameters for mechanical phenotyping of living cells.* Eur Biophys J, 2013. **42**(5): p. 383-94.

165. Lincoln, B., et al., *Reconfigurable microfluidic integration of a dual-beam laser trap with biomedical applications.* Biomed Microdevices, 2007. **9**(5): p. 703-10.

166. Bianchini, G., et al., *Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease.* Nat Rev Clin Oncol, 2016. **13**(11): p. 674-690.

167. Norton, K.A., A.S. Popel, and N.B. Pandey, *Heterogeneity of chemokine cell-surface receptor expression in triple-negative breast cancer.* Am J Cancer Res, 2015. **5**(4): p. 1295-307.

168. Zhu, G., et al., *CXCR3 as a molecular target in breast cancer metastasis: inhibition of tumor cell migration and promotion of host anti-tumor immunity.* Oncotarget, 2015. **6**(41): p. 43408.