THE UNIVERSITY OF CHICAGO

SIMULTANEOUS QUANTIFICATION OF GENE EXPRESSION, PROTEIN AND PROTEIN COMPLEX IN SINGLE CELLS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE PRITZKER SCHOOL OF MOLECULAR ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
HOANG VAN PHAN

CHICAGO, ILLINOIS
MARCH 2022

To my parents, without whose sacrifices I would not be here.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Savaş Tay, for his guidance in my research. He has allowed me great freedom in coming up with my own ideas, and provided me with the best resources so that I could pursue them.

I want to thank Dr Luke Vistain, whom I have worked with the most. You have provided me useful advice, on projects that we work on together and those that we do not. I want to thank Dr Nir Drayman for his help and guidance. Your advice was crucial to the completion of one of my main projects.

I want to thank all the members in the Tay Lab for being important collaborators, helpful colleagues, and supportive friends. You are always willing to help me with whatever problems I have in my research. Our discussion has made me a better scientist. Your friendliness and companionship have made Chicago a fun experience for me.

Most importantly, I want to thank my father and my mother. You have made so many sacrifices so that I could pursue and complete my PhD study. You have always been there for me, provided emotional support to me, and given me invaluable words of encouragement when I needed them most.

ABSTRACT

Single-cell analysis has become an increasingly important tool in cell biology. One of the most popular analysis methods is single-cell ribonucleic acid sequencing (scRNA-seq), which enables quantification of gene expression in single cells. This type of information has led to many important discoveries, from existence of new cell types to tumor heterogeneity. However, it is proteins, not RNAs, that are the main effector molecules in biological processes. Cells recognize environmental signals using protein receptors on the cell surface. The environmental signals are then transmitted through protein interactions inside the cells as part of a signaling cascade. Finally, the appropriate genes are transcribed, allowing the cells to respond to the environmental signals. These protein activities cannot be inferred from gene expression level alone, because many of them occur prior to transcription. In this dissertation, we introduce Proximity sequencing (Prox-seq), a novel single-cell multiomic method that bridges the gap between gene expression and protein-protein interactions. Prox-seq combines proximity ligation assay (PLA) with scRNA-seq to quantify gene expression level, protein abundance, and most importantly, protein complexes in the same single cell. First, we develop Prox-seq for surface receptors, and show that we can indeed obtain these three types of information from the same single cells. By applying Prox-seq to peripheral blood mononuclear cells (PBMCs), we find a putative interaction between CD8 and CD9 receptors on CD8 T cells. Then, we develop a high-throughput, droplet-based scRNA-seq for fixed and permeablized cells, called FD-seq. FD-seq can serve as a platform to extend Prox-seq to intracellular proteins. Using FD-seq, we identify host genes that are associated with herpesvirus reactivation, and show that following exposure to coronavirus, only a minority of the cells express a high level of viral genes. Finally, we develop computational frameworks for simulating Prox-seq data, and for prediction of protein complexes in single cells from Prox-seq data.

CHAPTER 1

INTRODUCTION

Many biological processes show important heterogeneity at the single-cell level [1–4]. Such observations could not be obtained from the population-averaged measurements of traditional bulk assays. One of the most important and popular single-cell methods is single-cell RNA sequencing (scRNA-seq) [5]. By providing information of gene expression levels in single cells, scRNA-seq has led to many important biological insights, from discovering rare cell types to unraveling heterogeneity in the immune responses and human brain tumors [6–9].

However, there are biological questions that cannot fully be explained by gene expression information alone [10–12]. First, cells perform their functions and communicate with one another through proteins. Second, many single-cell studies have found weak correlations between messenger RNA (mRNA) and protein levels [13–16]. Third, many protein interactions occur much earlier than RNA transcription, such as receptor clustering, protein phosphorylation and ubiquitination. As a result, gene expression levels are not necessarily representative of protein activity and interaction.

Despite this, scRNA-seq is still the most widely used method of single-cell analysis. One reason is that it can reliably measure 4,000–9,000 genes simultaneously in thousands of single cells in one experiment [17, 18]. In contrast, current single-cell proteomics methods can only achieve either a high number of proteins, or a high number of cells, not both. For instance, single-cell mass spectrometry can detect thousands of proteins but has a low throughput due to limited single-cell barcoding capacity [19, 20]. On the other hand, while flow-based methods can measure hundreds to thousands of cells per second, they can only detect of tens of proteins at the same time: flow cytometry's multiplexing capacity is limited to 40 protein targets due to spectral overlapping, and mass cytometry (CyTOF) is limited to 50 targets due to the availability of the metal tags.

The state-of-the-art technology for single-cell protein quantification is a family of methods that utilizes DNA-barcoded antibodies (Abs) and next-generation sequencing (NGS) readout

[21–23]. The most prominent example of such methods is CITE-seq [21]. By designing the DNA barcodes to be compatible with existing scRNA-seq methods, these methods not only measure a high number of proteins (more than 100 proteins) at high throughput (thousands of single cells in the same experiment), but also provide additional types of information like gene expression and chromatin accessibility [24].

The ability to obtain different types of information in the same single cells is called single-cell multiomics, or multimodal, analysis [25–27]. This extremely powerful capability has enabled a more comprehensive characterization of biological processes [28–31]. Furthermore, having multiple types of information from the same single cells can help overcome the sparsity of single-cell data, i.e., a molecule might be detected in some cells but not the others. Single-cell multiomics methods can be broadly categorized into four components: transcriptome (gene expression), proteome (protein expression), genome (genomic sequence), and epigenome (chromatin accessibility and histone modifications) [26].

An important proteomic component of cellular functions is protein-protein interactions (PPIs). Many signaling pathways are mediated by receptor clustering and dimerization [32–35]. An example is T cell receptor (TCR) signaling. TCR recognition of a foreign antigen presented by the major histocompatibility complex (MHC) on an antigen-presenting cell requires binding of the co-receptor CD4 or CD8 [36]. Another example is the NF-κB signaling pathway [3, 37]. When the pathogen-recognition receptors (PRRs) on the cell surface encounter a foreign ligand, a series of protein interactions lead to the translocation of the NF-κB transcription factor from the cell cytoplasm to the nucleus, and finally activation of downstream gene expression [38]. Nuclear translocation of NF-κB only occurs after the protein IκBα has dissociated from NF-κB. Since IκBα is one of the downstream genes of NF-κB, dissociation of the complex eventually leads to an increase in expression of IκBα protein, which then associates with free NF-κB molecules and prevents further translocation in a negative feedback loop.

Given the importance of protein-protein interactions, there is a lack of powerful methods

for characterization of protein complexes in single cells, however. On the one hand, current bulk PPI assays do not have sufficient sensitivity for single-cell resolution. On the other hand, the few existing single-cell PPI methods can only detect very few PPIs per experiment, have throughput, and cannot be integrated with multiomics analysis.

In this dissertation, we present a novel single-cell multiomic sequencing method called Prox-seq (Proximity sequencing). Prox-seq combines scRNA-seq with proximity ligation assay (PLA) to simultaneously provide information on gene expression, protein expression, and protein complex abundance in single cells. Prox-seq utilizes DNA oligonucleotide-conjugated antibodies to detect the proteins of interest on the cell surface [21, 22]. In contrast to other antibody-based assays such as flow cytometry and CITE-seq, Prox-seq does not measure the signal from single antibodies. Instead, a protein signal in Prox-seq is produced from two antibodies. More specifically, when two antibodies are in proximity to each other (e.g., when they bind to the subunits of a protein complex), the DNA oligonucleotides (oligos) on the antibodies can be ligated, and only the ligated oligos can be measured. This ligation requirement enables detection of protein proximity and protein interactions. Finally, by processing the cells with scRNA-seq, we obtain two types of sequencing data, one for gene expression level, and one for the count of PLA products that are produced from the ligated DNA oligos. From the count of these PLA products, we can estimate the protein expression and, using a statistical method that we develop, predict the abundance of protein complexes on single cells. To demonstrate its possible biological applications, we use Prox-seq to characterize surface protein complexes on single immune cells, and changes in receptor interactions following Toll-like receptor stimulation.

Next, we present two new tools that can help extend Prox-seq to intracellular proteins and intracellular protein complexes. First, we present FD-seq (Fixed Droplet RNA sequencing), a high-throughput, droplet-based method for scRNA-seq of paraformaldehyde-fixed single cells. Such a method is necessary, because detection of intracellular proteins necessitates fixation and permeabilization, which can strongly affect the quality of RNA from single

cells. Using FD-seq, we study herpesvirus reactivation and coronavirus infection at the single-cell level.

Second, we present computational frameworks for simulating PLA product counts obtained in Prox-seq, and for predicting protein complexes from Prox-seq data. We show that the background amount produced by random, non-specific protein proximity is more significant for highly expressed proteins. In addition, we propose two independent methods for prediction of protein complex abundances from Prox-seq data, and show that the two methods generally produce similar results in experimental data.

The dissertation is organized as following. Chapter 2 summarizes the state-of-the-art methods for scRNA-seq, protein quantification, and protein-protein interaction characterization. The working principle of PLA, the assay underlying Prox-seq, is also described. Chapter 3 describes the working principle of Prox-seq, and its application in characterization of surface proteins and protein interactions. Chapter 4 introduces FD-seq, a droplet-based scRNA-seq method for paraformaldehyde-fixed and permeabilized single cells. Chapter 5 presents computational frameworks for simulation of Prox-seq data, and for prediction of protein-protein interactions. Finally, Chapter 6 summarizes the dissertation work, the current limitations of Prox-seq, and its future directions.

CHAPTER 2

BACKGROUND

2.1 Methods for Single-Cell RNA Sequencing

scRNA-seq is an extension of bulk RNA sequencing to single cells. There are many different scRNA-seq protocols, but they all consist of similar steps. First, single cells are isolated and lysed. Next, the mRNAs are reverse transcribed into complementary DNAs (cDNAs). Usually, the cDNAs are also barcoded with the single-cell identity during this step. Third, the cDNAs are amplified using PCR to increase the amount of input materials. This step is also known as whole transcriptome amplification. Finally, sequencing libraries are constructed from the amplified cDNAs, and the libraries are then analyzed using next-generation sequencing (NGS) [39]. From the sequencing data, we obtain the expression levels of thousands of genes in single cells [17].

scRNA-seq methods have improved significantly over the past decade. The number of cells analyzed per experiment has increased from fewer than 10 cells in 2009 to more than 100,000 cells in 2021 [40], easily beating the Moore's law on the number of transistors [41, 42]. To improve scRNA-seq, scientist have had to overcome two major challenges: efficient sample processing of low-input RNA material, and scalable single-cell isolation and barcoding methods. To solve the first challenge, different approaches have been proposed to improve the reverse transcription and whole transcriptome amplification steps, such as using template-switching reverse transcriptase [43], *in vitro* transcription [44, 45], optimal buffer conditions [46, 47], molecular crowding [48], and second-strand synthesis [49].

Improvements in single-cell isolation and barcoding methods are the most important factor behind the rapid improvement of scRNA-seq throughput (Figure 2.1). Early scRNA-seq studies relied on manual pipetting of individual cells [50]. Then, single cells were obtained by fluorescence-activated cell sorting (FACS) [46, 51], or by single-cell isolation in a microfluidic device, such as the Fluidigm C1 system [52, 53], bringing the throughput to approximately

100 cells per experiment. Currently, the highest throughput single-cell barcoding methods are droplet encapsulation (also using a microfluidic device) [54–56], combinatorial indexing (also known as split-pool indexing) [57, 58], or a combination of both [59]. Thanks to these advances, scRNA-seq has been successfully commercialized and now widely used by biologists and medical scientists.

High-throughput isolation of cells can be achieved with droplet microfluidics [60], which makes use of microscale fluid dynamics to generate thousands of aqueous droplets in oil per second (Figure 2.1d). When the cell loading concentration is sufficiently low, only a small fraction of the droplets contain one cell per droplet, with the majority of the droplets containing no cell at all, and a negligible fraction containing more than one cells. The number of cells in a droplet can be well approximated by the Poisson distribution [54, 56]. To mark the mRNA molecules that come from a single cell, barcoded beads are also loaded at a low concentration such that most droplets contain at most one bead. Inside the droplets that co-encapsulate a single cell and a bead, the cells are lysed, and the released mRNAs are captured by the barcoded beads. Through reverse transcription, the barcode on the bead is incorporated into the cDNAs so that each cDNA molecule is labeled with the single cell's barcode. The most popular droplet-based scRNA-seq methods are Drop-seq [54] and the commercial platform 10x Chromium [56]. They differ in the type of barcoded beads used (ceramic beads and hydrogel beads, respectively), and the chemistry of sample processing. As a result, 10x requires fewer input cells, and generally produces higher quality data than Drop-seq [17, 61].

Droplet-based scRNA-seq methods often require sophisticated and expensive equipment. In contrast, combinatorial indexing (or split-pool indexing) can be performed with readily available equipment, such as a FACS sorter. Cells are first split into multiple pools, and all the cells in each pool are barcoded with a pool-specific index (Figure 2.1e). The pools are then pooled together, and split the second time into multiple pools for a second round of barcoding. Each combinatorial indexing experiment involves at least two such rounds of split-

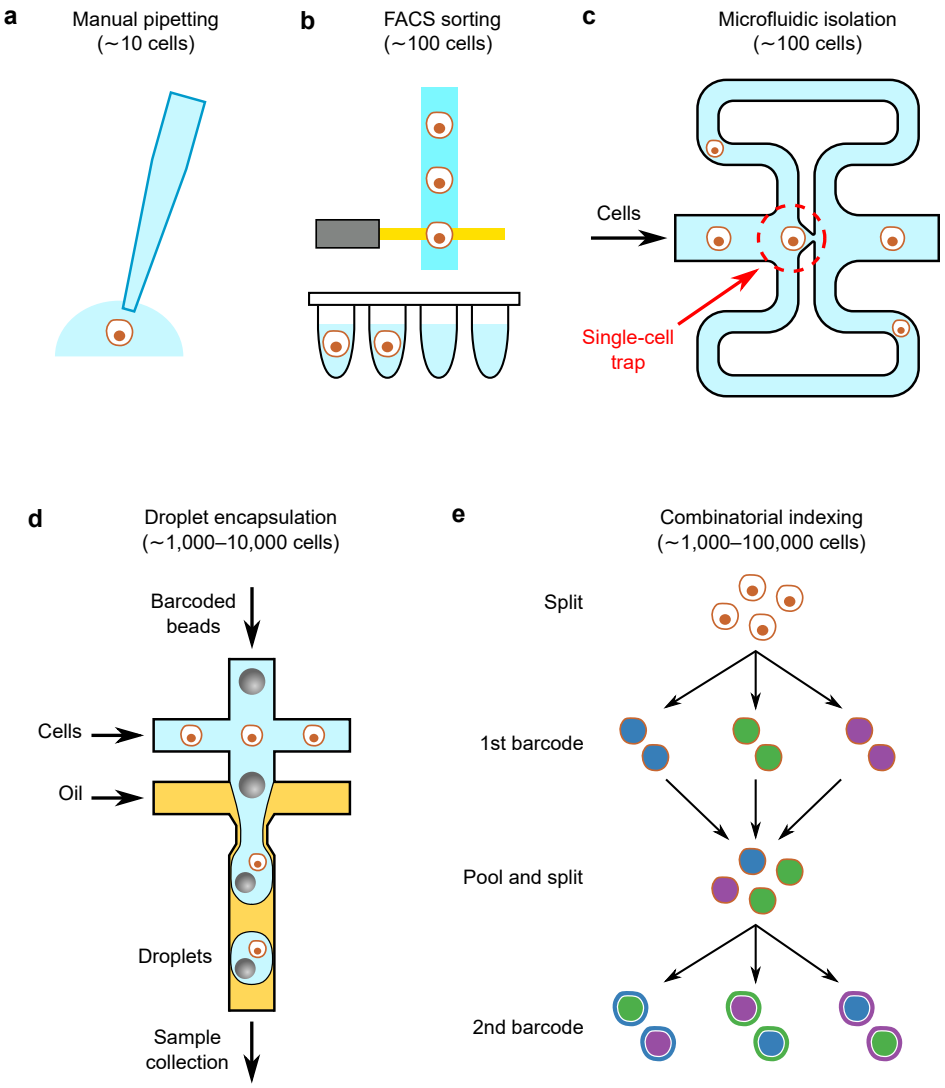

Figure 2.1: Common single-cell isolation and barcoding methods. (a) Single-cell isolation by manual pipetting. (b) Single-cell isolation by FACS sorting. Each single cell is manually pipetted into a separate container for lysis. (c) Single-cell isolation by Fluidigm's C1 microfluidic system. A single-cell is captured by a hydrodynamic trap (dashed circle). (d) Single-cell isolation and barcoding through microfluidic droplet encapsulation. Thousands of droplets are produced per second in the microfluidic device, and each droplet contains either 0 or 1 cell. (e) Single-cell barcoding through combinatorial indexing (or split-pool indexing). In this example, 2 rounds of barcoding are done with 3 indexes per round, resulting in $3^2 = 9$ distinct single-cell barcodes. By aiming for a high number of possible single-cell barcodes relatively to the number of cells, each cell is likely to have a unique barcode. In (a-e), the number in parentheses indicates the approximate number of single cells obtained per experiment.

pool barcoding. For combinatorial indexing scRNA-seq, the first barcode is incorporated during reverse transcription with an indexing primer, and subsequent barcodes are ligated

to the reverse transcription primer. During each barcoding round, multiple cells are barcoded with the same index. The ingenuity of this simple approach is that, with a sufficiently high number of rounds and/or indexes per round, each single cell has a very high probability of having a unique barcode combination. If r is the number of barcoding rounds, and i is the number of indexes used in each round, the number of unique index combinations is i^r. As an example, by using 2 rounds of barcoding with 96 indexes per round, the maximum number of single cells that can be uniquely barcoded is $96^2 = 9,216$. In practice, a smaller number of cells are used to further minimize the probability of barcode duplication. A disadvantage of combinatorial indexing is that it requires laborious and time-consuming processing.

2.2 Methods for Quantification of Proteins

Methods for protein detection and quantification can be broadly categorized into two groups, targeted and non-targeted. Targeted methods make use of a label, usually an antibody (Ab), to detect the proteins of interest. Non-targeted methods identify a protein directly without any labeling. The targeted approach is simpler to use, but it is restricted by the antibody availability for the proteins of interest. On the other hand, the non-targeted approach can detect a much higher number of proteins at the same time, and enable hypothesis-free proteomic studies.

The most widely used method for protein measurement in single cells is a targeted approach called flow cytometry [62]. It uses fluorescent labeled Abs that bind to the protein targets, and the protein abundance is measured from the fluorescence intensity. Flow cytometry has an extremely high throughput, because it can analyze thousands of single cells per second. Different protein targets can be measured from the same cells using multicolor flow cytometry. However, the number of measurable protein targets (i.e., the number of colors) is limited by spectral overlap [63].

Spectral overlap occurs when different fluorescence signals can be detected by the same detector, and therefore could not be distinguished from one another. To correct for it, careful

control samples and complicated color compensation methods are needed. This is probably why, even though 48-color flow cytometers are commercially available [64], researchers have only demonstrated a 28-color panel [65]. The highest number of colors achieved so far is 40, using a more advanced version of flow cytometry called full spectrum flow cytometry [66, 67]. While conventional flow cytometry measures a small portion of the emission spectrum of each fluorescence signal, full spectrum flow cytometry records its full emission spectrum. This allows the latter to better distinguish the fluorescence signals with similar emission, indirectly overcoming spectral overlap.

Another commonly used targeted method is western blot. First, the protein samples are denatured and separated by size using gel electrophoresis. Next, the proteins are transferred to a blotting membrane. Finally, the proteins are detected and quantified with Ab staining.

While western blot is traditionally used for bulk samples and requires a large amount of input material, Hughes et al. [68] was able to optimize it for single-cell quantification. This single-cell western blotting (scWesterns) method has two major disadvantages however. First, the number of protein targets that can be measured from the same single cell is limited, because it relies on stripping and reprobing to achieve protein multiplexing. In stripping and reprobing, after the Ab staining step for the first protein, the Ab is stripped away chemically, and a different Ab is added to stain (reprobe) the second protein. The stripping and reprobing process is repeated for each of the remaining protein targets. Because chemical stripping can damage the proteins, and the Abs from previous rounds might not be stripped away completely, stripping and reprobing progressively reduce the signal-to-background. Second, scWesterns cannot be easily integrated with other types of information, such as gene expression level.

Mass spectrometry (MS) is the gold standard method for non-targeted, label-free protein quantification [69, 70]. An MS-based experiment typically starts with digestion of a protein sample into peptides, followed by fragmentation, ionization, and finally measurement with a MS machine. The measurement output is called the mass spectra, which contain information

about the mass-to-charge ratios of the detected ions. Computational methods are then used to identify and quantify the proteins in the sample from these mass spectra [70–73]. MS is a powerful proteomic method, as it can provide absolute quantification of up to 10,000 proteins [74, 75].

Budnik et al. [19] recently developed a single-cell version of MS, called SCoPE-MS. Its improved version, SCoPE2 [20], can quantify more than 3,000 proteins in more than 1,000 single cells across multiple experiments. SCoPE-MS and SCoPE2 barcode the protein sample from each single cell with a tandem mass tag (TMT), and use carrier cells to enhance peptide identification of the single-cell proteins, which would otherwise be difficult due to the low abundance of single-cell materials. While SCoPE2 can detect significantly more proteins than Ab-based methods, it is limited in throughput: current TMT panels only allow barcoding of up to 16–18 samples per experiment [76].

Mass cytometry, also known as cytometry by time-of-flight (CyTOF), leverages the advantages of both flow cytometry and MS to provide highly multiplexed protein quantification in single cells at high throughput [77–80]. The Abs are labeled with heavy metal isotopes instead of fluorophores, so that cells can be stained like in a flow cytometry experiment, while the heavy metals can be detected based on their mass-to-charge ratios with time-of-flight MS. The amount of heavy metal tags per protein corresponds to its expression level. Currently, 40–50 proteins can be measured simultaneously with CyTOF [81, 82]. Unfortunately, CyTOF's multiplexing capacity is limited by the availability of the heavy metal tags [83].

The most advanced approach for quantification of proteins in single cells is arguably DNA-barcoded antibody-based assay. The earliest methods of this approach are CITE-seq [21], REAP-seq [22] and Abseq [23]. Among them, CITE-seq is commercially available as TotalSeq from BioLegend, and is therefore the most widely used and most advanced method [84]. These methods use DNA oligonucleotide-conjugated Abs to detect the protein targets, similar to flow and mass cytometry, but the signal is measured using NGS instead. This confers two major advantages over the alternative single-cell proteomic methodss. First,

because the protein signal is converted to DNA signal, CITE-seq and others can be coupled with other single-cell sequencing assays to simultaneously provide many different types of measurement, such as gene expression level [21], chromatin accessibility [24], or gene perturbation information [85]. This capability is called multiomic analysis. Second, because the protein identity is encoded in the sequence of the DNA oligonucleotide, these methods can measure a much higher number of protein targets from the same sample. At the time of introduction, REAP-seq could measure up to 80 protein targets simultaneously [22]. At present, TotalSeq antibody cocktails for at least 130 protein targets are commercially available [86], with SCITO-seq demonstrating quantification of 165 protein targets simultaneously [87].

2.3 Methods for Analysis of Protein-Protein Interactions

Detection of protein-protein interactions is more complicated than detection of individual proteins, because both a protein and its protein interaction partners have to be identified. The traditional methods for detection of protein complexes are the pull-down and co-immunoprecipitation (co-IP) assay. In a pull-down assay, a bait protein is first immobilized on a solid surface through an affinity tag [88]. Then, a solution of putative interacting proteins is added so that they can bind to the bait. After affinity purification, all complex partners that bind to the bait proteins can be detected with western blot. With co-IP, the protein of interest is captured by an Ab instead, and its complex partners are isolated and identified [89].

A powerful technique for large-scale investigation of binary protein-protein interactions (PPIs) is the yeast two-hybrid assay (Y2H) [90–92]. It uses two genetically engineered yeast clones, called the bait clone and the prey clone. The bait clone contains a protein fusion between the first interacting partner and a DNA-binding domain specific to a reporter gene. The prey clone contains a protein fusion between the second complex partner and an activation domain specific to the bait clone's reporter gene. After systematically mating the bait clones with the prey clones, the offspring generation will contain both protein

fusions from the parents. Only offspring with interacting bait-prey combinations will result in expression of the reporter gene.

Y2H assay has several important drawbacks. First, it involves many complicated and time-consuming steps, from generating the fusion proteins and the yeast clones to mating them. Second, the interactions between the bait and prey might not be symmetric (i.e., whether a protein in the interaction is a bait or a prey matters). Third, because the assay is performed in yeast cells, the PPIs might not occur in their natural cellular environment. Last, and most importantly, the assay is not intended for quantification purpose, and is not suitable for single-cell analysis.

Two MS-based methods are available for characterization of *in situ* PPIs. The first and most established one is affinity purification MS (AP-MS), which uses affinity tagging to isolate the proteins that interact with a bait protein [93, 94]. In this method, the bait protein is first fused to an affinity tag, and the fusion protein is transfected into the cells. Then, the cells are lysed, and the bait protein and its protein interaction partners are enriched using an anti-AP tag antibody. Finally, the interacting proteins are eluted and quantified with MS. The second method is co-fractionation MS (CF-MS), in which cell lysates are subjected to size-exclusion or ion-exchange chromatography [95–97]. This results in separation of interacting proteins into different fractions. Next, MS is used to obtain the co-fractionation profiles of the proteins, which are then used to infer PPIs [98].

AP-MS has three main limitations: it necessitates genetic manipulation, can only detect the interaction partners of one protein at a time, and lacks single-cell resolution. While CF-MS circumvents the first two limitations, it is also not suitable for single-cell measurement, in addition to being prone to false positive results [99], and having a lower sensitivity [100].

Two other MS-based methods utilize proximity labeling to identify *in situ* PPIs. They are BioID [101] and APEX [102], which fuse an engineered protein (a biotin ligase and an ascorbate peroxidase, respectively) to the protein of interest to biotinylate its interacting partners. The biotinylated proteins can then be purified and quantified with MS. BioID can

biotinylate proteins within 10 nm of the protein of interest [103], and APEX within 20 nm [102]. An important difference between BioID and APEX is time: APEX's labeling can be done within minutes, compared to 24 hours for BioID's [102]. Like AP-MS, these methods require genetic manipulation, and lacks single-cell resolution.

Single-cell detection of protein complexes can be achieved with Förster resonance energy transfer (FRET) [104–106]. In FRET, two proteins are labeled with an acceptor fluorophore and a donor fluorophore, called a FRET pair. The acceptor fluorophore produces a fluorescence signal only when it is within 10 nm of the donor fluorophore [107]. This also means that protein interactions can be measured in real-time with FRET. However, FRET can only detect one protein-protein interaction at a time, and FRET pairs are limited, with only 19 or so pairs available at the moment [108]. FRET data is also difficult to analyze, limiting its use to specialized labs [109, 110].

2.4 Proximity Ligation Assay

Proximity ligation assay (PLA) is a versatile Ab-based proteomic assay. It has been used for detection of protein-protein interactions [111, 112], cytokine measurement [113–115] and single-cell protein quantification [16, 116].

For a protein molecule or a protein complex to be detected in PLA, it has to simultaneously bind two separate DNA oligonucleotide- (oligo-) conjugated Abs. These pairs of Abs are called PLA probe A and PLA probe B (Figure 2.2). When this occurs, the DNA oligos are sufficiently close to each other to be ligated with the help of a third oligonucleotide called the connector. The resulting ligated products can then be quantified by rolling circle amplification (RCA), real-time polymerase chain reaction (qPCR), droplet digital PCR (ddPCR) or NGS. For protein quantification, the dual Ab binding event could significantly reduce the background signal compared to single Ab binding [114].

A closely related assay is the proximity extension assay (PEA) [117]. PEA was first developed to circumvent the ligase inhibitors in human samples, which would interfere with

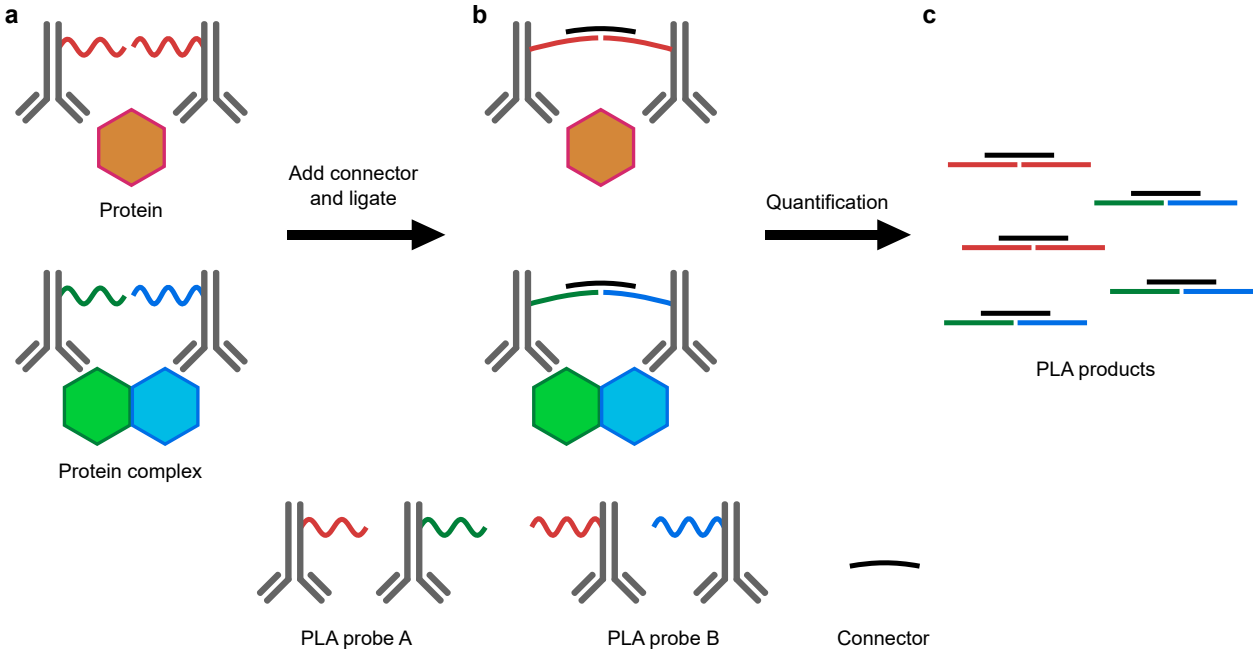

Figure 2.2: Working principle of proximity ligation assay (PLA). (a) A protein molecule or protein complex is bound by two DNA-conjugated antibodies, called PLA probes A and B. (b) The DNA oligonucleotides on the probes are ligated with the help of a connector DNA oligo. (c) The ligated PLA products can be quantified using any DNA quantification method.

PLA's performance. In PEA, when two DNA-conjugated Abs bind to the same protein molecule, or the components of a protein complex, the DNA oligos on the Abs hybridize with each other. A DNA polymerase is then used to extend one of the DNA oligos, thus converting protein signal to DNA signal. PEA has been used to quantify protein abundance in single cells [118, 119].

An untapped advantage of PLA is that the use of two Abs enables combinatorial detection of binary PPIs. The DNA oligos on the Abs can be designed such that any PLA probe A can be ligated with any PLA probe B. PLA can thus screen for binary PPIs like Y2H assay, without the latter's need for genetic manipulation. Because PLA uses Abs, it is also suitable for detection of *in situ* protein complexes. Moreover, the number of detectable PPIs scales quadratically with the number of protein targets. For example, an antibody panel of 10 protein targets can measure up to 55 pairwise protein interactions, including homodimers, and a 40-target panel can measure up to 820 protein interactions. Demultiplexing such a

large number of measurements requires a sequencing readout, which is naturally compatible with the DNA products that PLA produces. Together with the advances in scRNA-seq, PLA is a prime candidate to bring forth the ability to quantify protein complexes at the single-cell level.

CHAPTER 3

QUANTIFICATION OF SURFACE PROTEIN, PROTEIN COMPLEX, AND MRNA IN SINGLE CELLS BY PROXIMITY SEQUENCING

In this chapter, we present Proximity sequencing (Prox-seq), a novel single-cell sequencing technology for simultaneous quantification of gene expression, proteins and protein complexes. Part of this chapter is reproduced from a manuscript that is under preparation for journal publication.

3.1 Summary

We introduce Proximity sequencing (Prox-seq) for simultaneous measurement of proteins, protein complexes and mRNAs in thousands of single cells. Prox-seq enables highly multiplexed single-cell quantification in a wide range of applications in phenotyping, signaling, and drug studies. Prox-seq combines proximity ligation assay with single-cell RNA sequencing to measure proteins and their complexes from all pairwise combinations of targeted proteins, providing quadratically-scaled multiplexing potential. First, we validate Prox-seq by analyzing a mixture of human T cells and B cells, and find that it accurately identifies these cells and detects well-known complexes, such as the CD3:CD3 homodimer. Next, by studying human peripheral blood mononuclear cells (PBMCs) with joint protein, protein complex and mRNA measurements, we discover that naive CD8 T cells display a potential protein complex CD8:CD9. Finally, we use Prox-seq to study the dynamics of surface receptor interactions during toll-like receptor (TLR) signaling in human macrophages. We observe formation and dissociation of protein complexes after stimulation with pathogen signals, observe CD36 co-receptor activity in TLR signaling, and show that TLR signaling is additive under LPS (TLR4) and Pam2CSK4 (TLR2) stimulation. Using logistic regression, we find that immune signaling inputs received by macrophages can be identified from Prox-seq data

despite single-cell stochasticity. Prox-seq provides access to an untapped, powerful measurement modality for single-cell phenotyping, namely the quantification of protein complexes, and can discover novel protein interactions in different cell types and signaling studies.

3.2 Introduction

Singe-cell measurements have expanded our understanding of many aspects of the cells, such as enabling identification of rare cell subsets, tracking transient cellular states, and incorporating noise and variability into our understanding of cellular phenotypes [3, 120, 121]. These phenotypes are emergent properties of both the biological molecules in the cell and the interaction between these molecules. Many biological functions such as signaling, differentiation, development, and cellular decision making are driven in large part by changes in the interactions of proteins. In particular, signaling processes are primarily mediated by the formation and dissociation of protein complexes, and thus cannot be studied from RNA expression or protein expression alone. Therefore, the ability to measure individual proteins and their complexes at the single-cell level would be among the most informative measurements for understanding cellular functions. Despite the apparent value, there are major hurdles in making highly multiplexed measurements of single-cell proteins and their complexes because the number of pairwise complexes a protein can form scales quadratically with the number of proteins. This necessitates a method that can encode a large number of outputs, as each measurement must enable identification of both the protein of interest and its interacting proteins, along with their transcripts.

Motivated by this unmet need in multiomic analysis in single cells, we developed a single-cell sequencing assay called Proximity sequencing (Prox-seq), and demonstrated an end-to-end experimental and computational pipeline for proteomic analysis of single cells (Figure 3.1a). Prox-seq simultaneously measures surface proteins, their protein complexes, and mRNAs in thousands of single cells by combining commonly used scRNA-seq methods with with proximity ligation assay (PLA) [111]. Prox-seq uses pairs of DNA-conjugated

antibodies, called Prox-seq probes, that are designed such that, upon being in proximity (i.e., both antibodies bind to epitopes that are spatially proximal, such as two interacting proteins), the DNA oligonucleotides (oligos) on the antibodies can be ligated. This process yields a ligated PLA product that can then be tagged with single-cell barcodes by any commonly used scRNA-seq method and read out with next-generation sequencing. From the counts of these PLA products, we can calculate the protein expression (similar to CITE-seq and REAP-seq), and most uniquely, the protein complex abundance in single cells (Figure 3.1b). For example, an anti-CD3 probe A can be ligated with an anti-CD4 probe B to create a PLA product called CD3:CD4 (Figure 3.1c). The count of this PLA product is then used to calculate the expression of CD3 and CD4 proteins, and to predict the count of the protein complex CD3:CD4.

Prox-seq has immense potential for highly-multiplexed proteomic analysis, because the number of possible protein complexes scales quadratically with the number of targeted proteins (Figure 3.1b). As an example, a panel of Prox-seq probes targeting 100 proteins can measure more than 5,000 unique pairwise protein-protein interactions. The DNA oligos and library construction were designed such that only ligated products can be measured by sequencing, preventing unpaired probes from contributing to the measured signal (Figure 3.1d). In addition, Prox-seq can readily quantify gene expression in single cells, enabling multimodal analysis of single cells.

For each protein target in our Prox-seq probe panel, we generate a pair of Prox-seq probes (called probe A and B), each of which is an antibody conjugated to a single-stranded DNA oligo (Figure 3.1a). The ratio of DNA oligo-to-antibody is selected to ensure that probes retain their ability to bind their targets (Supplementary Figure 3.8). The DNA oligos are designed such that each member of probe A can ligate with any member of probe B through a universal connector region. After probe binding and ligation, the cells can be processed through scRNA-seq methods that utilize poly-A capture, including Smart-seq2, Drop-seq, and 10x Chromium (10x) [46, 54, 56], to retrieve both ligated PLA products

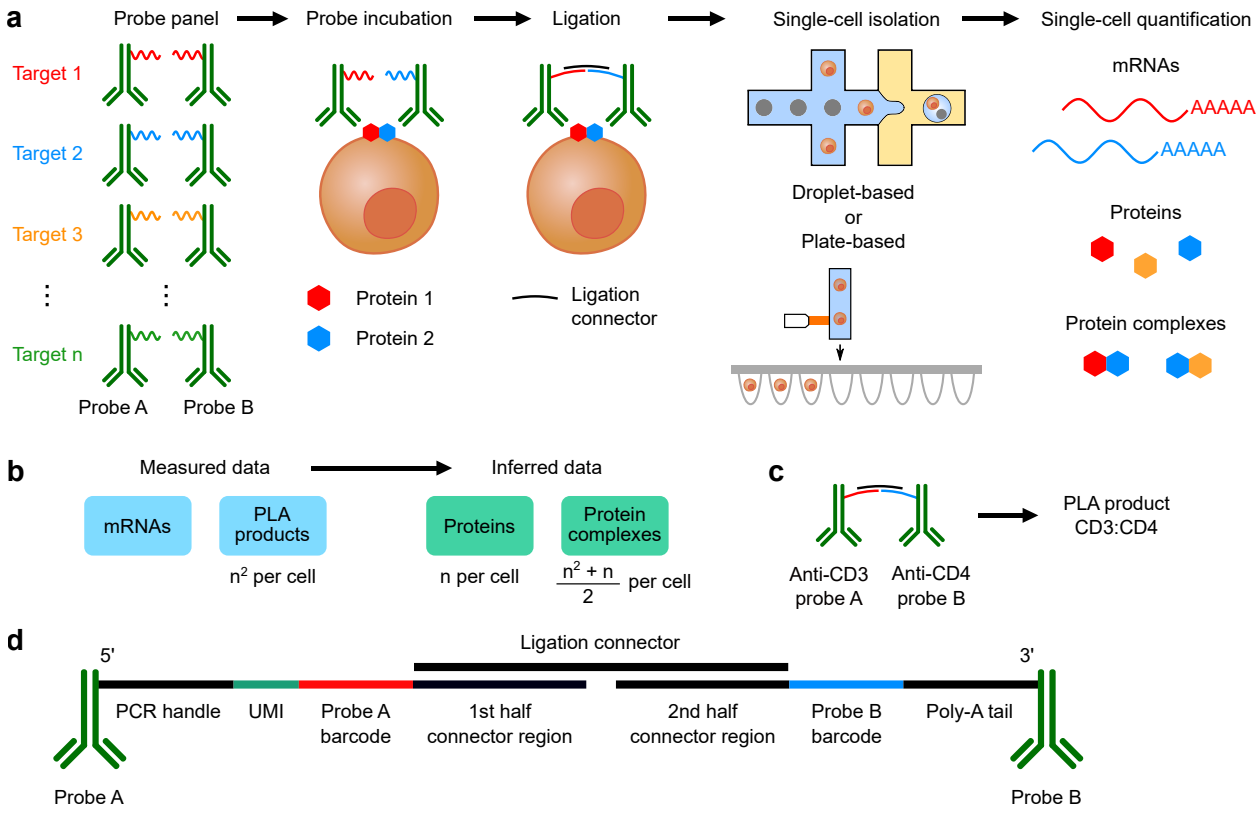

Figure 3.1: Overview of Prox-seq assay for multiomics analysis of mRNA, protein and protein complex in single cells. (a) Prox-seq workflow: cells are stained with a panel of Prox-seq probe pairs (Prox-seq probe A and B), ligated, and processed using a droplet-based or plate-based scRNA-seq protocol. Each Prox-seq probe is an antibody conjugated to a DNA oligo barcode. (b) The measurement output of Prox-seq is the transcript count for each single cell, and the count of n^2 PLA products for each single cell. The count of PLA products is then used to calculate the abundance of n proteins and $(n^2+n)/2$ possible protein complexes. (c) Naming convention of a PLA product. For example, ligation of an anti-CD3 probe A and an anti-CD4 probe B produces the PLA product CD3:CD4. (d) The layout of a PLA product. Probe A and probe B are conjugated to the DNA oligos at the 5' and 3' end, respectively. Because ligation requires both an A probe and a B probe, only ligated products can be measured by Prox-seq. Unligated Prox-seq probes are automatically discarded during the library construction step.

and mRNAs. We have shown that Prox-seq is compatible with all three of these popular single-cell sequencing readouts, demonstrating its wide-scale applicability. Thus, Prox-seq can readily be integrated into existing research workflows.

The DNA oligos used to form PLA products include several features that facilitate integration into these single-cell sequencing methods (Figure 3.1c, d). A complete PLA product includes a unique molecular identifier (UMI) region for PCR bias correction, two barcode

regions to indicate the identity of the A and B antibodies, a 3' poly-A tail for capturing of PLA products by scRNA-seq protocols, and a primer binding site for PCR. The universal connector regions enable proximity ligation, and only ligated products can be PCR amplified and sequenced. Prox-seq expands on the state-of-the-art in several ways. Recent technologies like CITE-seq and REAP-seq have displayed the capacity to simultaneously quantify extracellular proteins and mRNAs in single cells [21, 22]. However, a major advantage of Prox-seq is its ability to detect protein complexes. The generation of a PLA product indicates that the two targeted proteins were spatially close enough to ligate the two 60-base long DNA strands. The complete PLA product spans 119 bases, with 20 of those bases hybridized to a connector. Based on estimates of ssDNA and dsDNA length, we expect that Prox-seq has a range of 53.8–73.7 nm [122]. At this length scale, PLA products can span the entire length of typical protein complexes. For example, the T Cell Receptor (TCR) bound to peptide-Major Histocompatibility Complex (pMHC) and anti-CD3 Fab spans a length of 25.6–42.6 nm [123].

PLA, and the closely related Proximity Extension Assay (PEA), have been separately applied to make single-cell measurements [16, 116, 118, 119], and to use a sequencing readout to measure proteins [114]. However, these two functionalities have never been combined into the same single cell assay. Furthermore, while PLA has been used to detect protein complexes in situ, it has never been paired with a sequencing output to enable the measurement of many protein complexes [111]. No currently available technologies are able to capture so many desirable types of information into a single workflow.

3.3 Results

3.3.1 Validation of Prox-seq for Multiplexed Protein Quantification From Single Cells

We first sought to show that PLA products can be measured using scRNA-seq, and that the PLA data display cell type-specific markers. 11 protein targets were selected corresponding to various markers of T cells and B cells (Figure 3.2a, Supplementary table 3.1). Prox-seq probes were made for these targets along with two isotype controls. This panel was applied to a mixture of T cells (Jurkat cells) and B cells (Raji cells), which was then analyzed using the Drop-seq pipeline [54] (Supplementary Figure 3.9a). Drop-seq is a high-throughput, droplet-based single-cell sequencing method that can analyze up to thousands of cells in a single experiment.

Prox-seq measurements showed that cells could be accurately clustered using either mRNA, protein abundance, or all PLA products (Figure 3.2b–d). The protein abundance of a target is estimated by taking the total number of times the protein target's DNA barcode is detected in PLA products, from either Prox-seq probe A or B (see "Methods"). We found that clustering of cells by mRNA or protein identified the same cell types (Figure 3.2c, e). Similarly, cells could be clustered using all 169 PLA products directly, which includes both protein proximity information and protein abundance (Figure 3.2d). Regardless of the data type used to cluster the cells, Prox-seq displayed good concordance between gene expression and the protein abundance for a given cluster (Figure 3.2f).

Similar to other studies, we found that the correlation between mRNA and protein for individual cells varied greatly between genes, and was typically modest (Supplementary Figure 3.10) [21, 22]. We also found that PD1:CD3 and CD3:CD3 PLA products were two of the most significantly enriched PLA products in the Jurkat cluster (Benjamini-Hochberg-adjusted P-value = 2.2×10^{-55} and 5.1×10^{-53}, respectively) (Figure 3.2g). This is expected, as CD3 is a known T cell (Jurkat) marker, and flow cytometry confirmed CD3

a

Protein target	Clone	Probe barcode	Jurkat expression	Raji expression
CD3	UCHT1	ATAGCGTC	+	−
PD1	NAT105	TAAGGCTC	+	−
HLADR	L243	AGAAGAGG	−	+
ICAM1	HA58	ATACGTGC	−	+

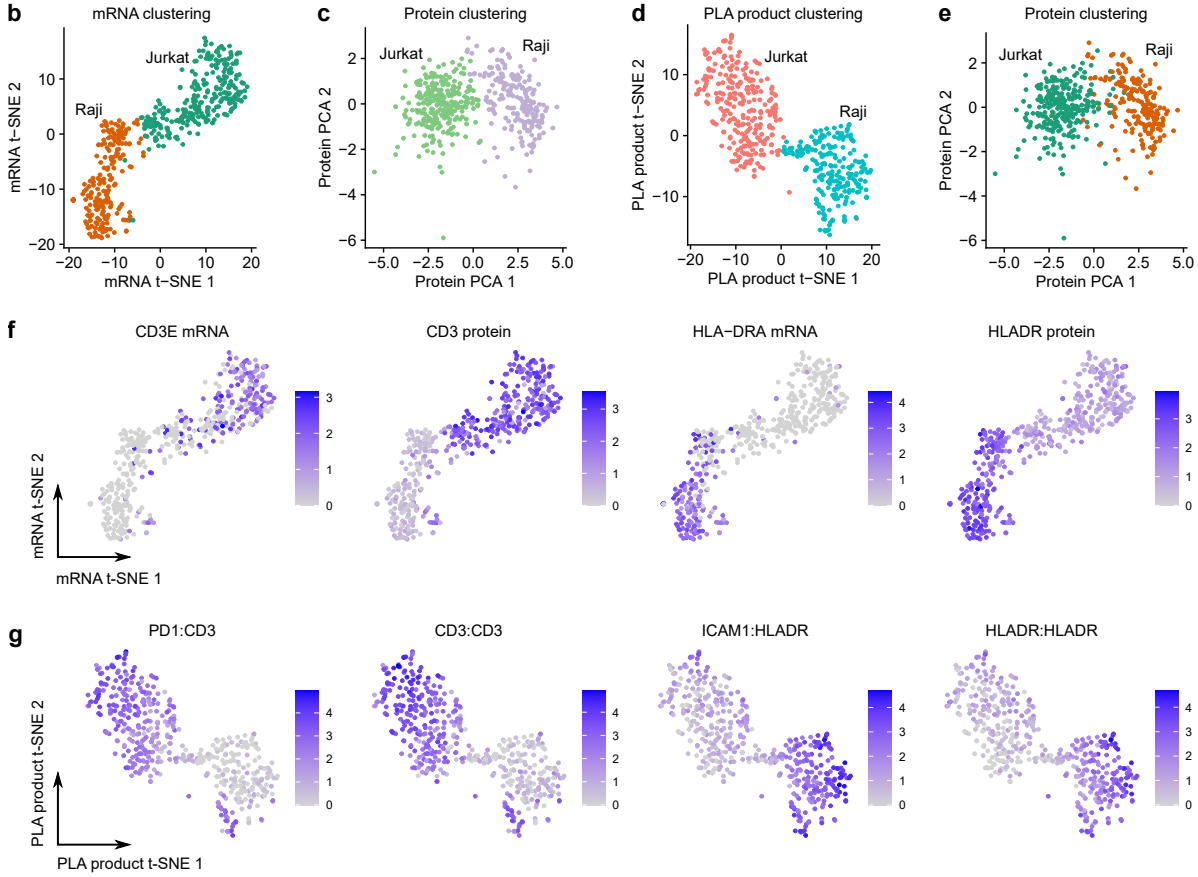

Figure 3.2: Prox-seq measurements identify cell types and accurately reflect protein abundance in single-cells. A mixture of Jurkat cells (T cells) and Raji cells (B cells) were measured with Drop-seq-based Prox-seq. Protein, protein complex and whole transcriptome mRNA were measured simultaneously in the same single cell. (a) Table showing two example markers for Jurkat and Raji cells, and the associated probe DNA barcode. (b) t-SNE plot showing single cells clustered with mRNA data. (c) PCA plot showing single cells clustered with protein abundance data (which is calculated from PLA product data). (d) t-SNE plot showing single cells clustered with PLA product data. (e) Concordance between cell type clusters is displayed using the same PCA plot as in (c), but with cluster labels obtained from mRNA data as in (b). (f) Correlation between protein and mRNA levels is shown for the *CD3D* gene, CD3 protein, *HLA-DRA* gene, and HLADR protein. (g) Plots showing the expression level of two of the most significant PLA product markers for each cell type (P-values $< 10^{-40}$, two-sided Wilcoxon Rank Sum test with Benjamini-Hochber correction).

and PD1 proteins to be Jurkat-specific (Supplementary Figure 3.11). For the Raji cluster, ICAM1:HLADR and HLADR:HLADR were two of the most significantly enriched PLA

products (Benjamini-Hochberg-adjusted P-value = 2.7×10^{-54} and 2.4×10^{-46}, respectively) (Figure 3.2g). Both ICAM1 and HLADR proteins are known to be expressed highly on Raji (B cells), and flow cytometry confirmed that both ICAM1 and HLADR are indeed uniquely expressed on Raji cells (Supplementary Figure 3.12).

3.3.2 Quantification of Proteins and Protein Complexes in Single Cells

We next sought to show that Prox-seq can quantify protein expression levels in single cells. To this end, we treated Jurkat and Raji cells with a panel of 13 Prox-seq probes and analyzed the PLA products using the plate-based Smart-seq2 pipeline [46] (Supplementary table 3.1). A plate-based method was chosen because such methods typically yield more UMIs and features per cell [17] (Supplementary Figure 3.9c). This panel allowed us to measure up to 91 potential pairwise protein complexes (Figure 3.1b). We analyzed the amount of non-specific antibody binding in Prox-seq, and observed that each cell-type specific protein displayed a difference of 1–3 orders of magnitude between expressing and non-expressing cell types (Supplementary Figure 3.13). Comparison between flow cytometry data and Prox-seq data showed a strong agreement in protein quantification (Spearman correlation coefficient $\rho = 0.87$, Supplementary Figure 3.14). These results demonstrated that Prox-seq accurately characterizes protein species in single cells, recapitulates the protein quantification feature of other assays such as REAP-seq and CITE-seq, and is compatible with two commonly used scRNA-seq technologies, Drop-seq and Smart-seq2.

A unique feature of Prox-seq, and a major advantage over existing single-cell multiomics techniques, is that it can reveal pairwise protein interaction information for each of the proteins that are targeted. However, the existence of a PLA product does not necessarily mean that the corresponding proteins are indeed interacting [124]. Therefore, we sought to specifically identify the PLA products that represent a stable protein complex using a probability model. We reasoned that, in the absence of stable protein complexes, the probability of a PLA product forming by random proximity is determined by the concentration of its

corresponding probe A and B on the surface of the cell. Using this assumption, we then calculated an expected random count for each PLA product based on the Prox-seq probe abundance, which reflected the number of PLA products we would expect to see if none of the targeted proteins were in complex with one another (see "Methods" for more details). When we compared the expected random counts to our experimental data, we found many PLA products that were present at a much higher abundance than expected, indicating the presence of true protein interactions (Figure 3.3a–b, Supplementary Figures 3.15, 3.15).

To quantify the abundance of the protein complexes from PLA data, we developed a Protein Complex Estimation Algorithm, which leverages the differences between the observed and expected random counts of PLA products (see "Methods"). First, the algorithm calculates the expected random count of each PLA product. It then compares the observed (i.e., measured) count of each PLA product for each cell type to the expected random count. If no statistically significant difference is detected by a one-sided t-test, the algorithm predicts that the PLA product does not correspond to a stable protein complex and assigns a complex count of 0. If the t-test returns a statistically significant results, then algorithm predicts that the PLA product correspond to a stable protein complex. Furthermore, the count of said protein complex is equal to the difference between the observed and expected random count of the PLA product. This method was applied to protein complex quantification results this study, and we refer a PLA product as a protein complex only if it is predicted by the algorithm. It is important to note that, due to the spatial nature of Prox-seq, it cannot distinguish between direct protein-protein contact and proximity induced by an additional unmeasured factor. Therefore, our identification of a protein complex should not be interpreted as a direct protein-protein contact between those interactors.

We applied this algorithm to our plate-based Jurkat and Raji data set, and found that four proteins were calculated to have more than 50% of their PLA product counts attributed to protein complexes: CD3 and CD28 homodimers in Jurkat cells, and PDL1 and HLADR homodimers in Raji cells (Figure 3.3c–d). It is important to note that, apart from CD147 and

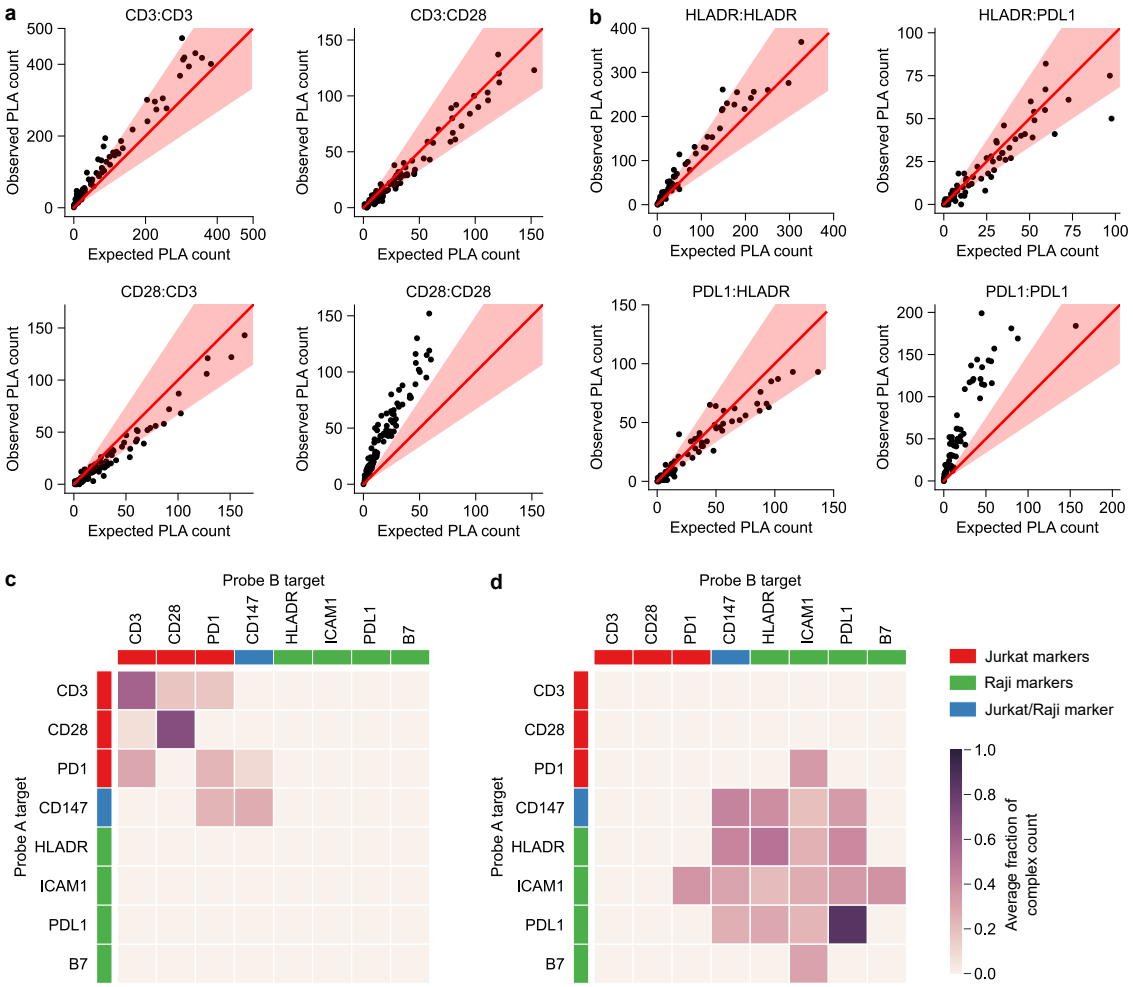

Figure 3.3: Quantification of stable protein complexes using Prox-seq. (a) Stable protein complexes are identified in single Jurkat cells by determining which PLA products show more counts than expected from random ligation of individual proteins. In the scatter plots, the X-axis indicate the expected amount of PLA products from random ligation, and Y-axis indicate actual measured PLA product counts for that dimer. The red solid lines indicate the line y = x, and the red regions indicate 1.5-fold change from the expected random PLA count. (b) Stable protein complexes can be identified in single Raji cells using the same method in (a). Scatter plots showing observed count against the expected random count of 4 example PLA products in Raji cells. (c, d) Heatmaps showing the estimated protein complex abundance as a fraction of observed PLA count, averaged across (c) single Jurkat cells and (d) single Raji cells.

the isotype controls, all antibodies in our panel are monoclonal, thus enabling quantification of both heterodimers and homodimers.

Accurate identification of the CD3 and CD28 homodimers in T cells is noteworthy because they serve as positive controls in our panel. The CD3 Prox-seq probes target the CD3e

protein, two of which are part of the TCR complex [125]. CD28 is known to form a stable homodimer on the cell surface through a disulfide bridge [126]. While it is unclear from previous studies if PDL1 forms a homodimer on the cell surface, all known crystal structures of PDL1 feature a homodimer [127], in agreement with our measurements. HLADR is thought to exist in an equilibrium between monomers and homodimers on the B cell surface [128, 129]. Therefore, our Protein Complex Estimation Algorithm correctly identified the presence of four known protein complexes in these cell types.

However, B7 and ICAM1 are both thought to undergo some degree of homodimerization [130, 131]. ICAM1 does indeed have the highest number of PLA products attributed to homodimers but, due to its very high expression level, the homodimer represents a small percentage of ICAM1 UMIs (approximately 27%, Figure 3.3d). The absence of B7 homodimers raises the possibility that the monoclonal antibody in this panel is unable to bind to its homodimerized form. In summary, the proposed algorithm allowed us to determine the PLA products that correspond to stable protein complexes, and provided us with a statistical framework to identify and quantify these complexes in our Prox-seq data.

3.3.3 Analysis of Human PBMCs With Prox-seq Identifies Known Protein Complexes and Shows a Potential Protein Complex CD9:CD8 in CD8 T Cells

Next, we explored the potential of Prox-seq to measure a large number of protein complexes on primary human immune cells. To achieve this, we generated a panel of Prox-seq probe pairs targeting 38 immune cell markers, with a primary focus on T cell markers (Supplementary table 3.2). This probe panel can measure up to 741 unique protein complexes. We applied this panel to human peripheral blood mononuclear cells (PBMCs) and analyzed the sample using two different methods: a plate-based method to maximize our ability to measure potentially rare protein complexes, and a droplet-based 10x method to simultaneously

measure mRNA and PLA products in a high-throughput manner (Supplementary Figure 3.9b, d).

The plate-based protein measurements clearly identified the expected cell types: CD8 T cells, CD4 T cells, and non-T cells (Figure 3.4a). Globally, our complex detection algorithm was able to identify 20 protein complexes present in these cells (Figure 3.4b). As before, we identified several known homodimers including the CD3 homodimer, the CD28 homodimer, and the CD9 homodimer [125, 126, 132] (Figure 3.4b). In addition, we identified the existence of both the CD3:CD8 and CD3:CD4 protein complexes (Figure 3.4b). Formation of both complexes is consistent with stimulation of T cells with the anti-CD3 antibody in our probe panel [133, 134]. Single cell heatmaps of an example CD4 T cell (Figure 3.4c) and CD8 T cell (Figure 3.4d) show clear differences, both in terms of detected PLA products and detected protein complexes.

Beyond these known protein complexes, we also identified a potentially novel interaction between CD9 and CD8 in single PBMCs. For CD8 T cells, we observed that cells could be split into two clear subpopulations based on PLA product abundance. In one subpopulation, CD9-related PLA products were primarily identified as paired with itself (CD9:CD9). The other subpopulation displayed CD9-related PLA products that were primarily paired with proteins other than CD9 (Figure 3.4e). It is noteworthy that this manner of cell state identification is only possible with the protein proximity data provided by Prox-seq.

We then sought to identify which protein was interacting with CD9 when the CD9:CD9 PLA product was disfavored. Interestingly, analysis of the CD9:CD9 PLA product-low subpopulation identified the existence of CD9:CD8 protein complex (Figure 3.4f). This is not a previously known complex. However, CD9 is known to participate in immune synapse formation, colocalize with CD3, and co-precipitate with CD3 [135, 136]. The appearance of this protein complex is not clearly attributable to changes in protein expression levels, as CD3, CD8, and CD9 are all similarly expressed in both cell populations (Figure 3.4g). While CD4 T cells also displayed these two subpopulations (also based on the level of CD9:CD9

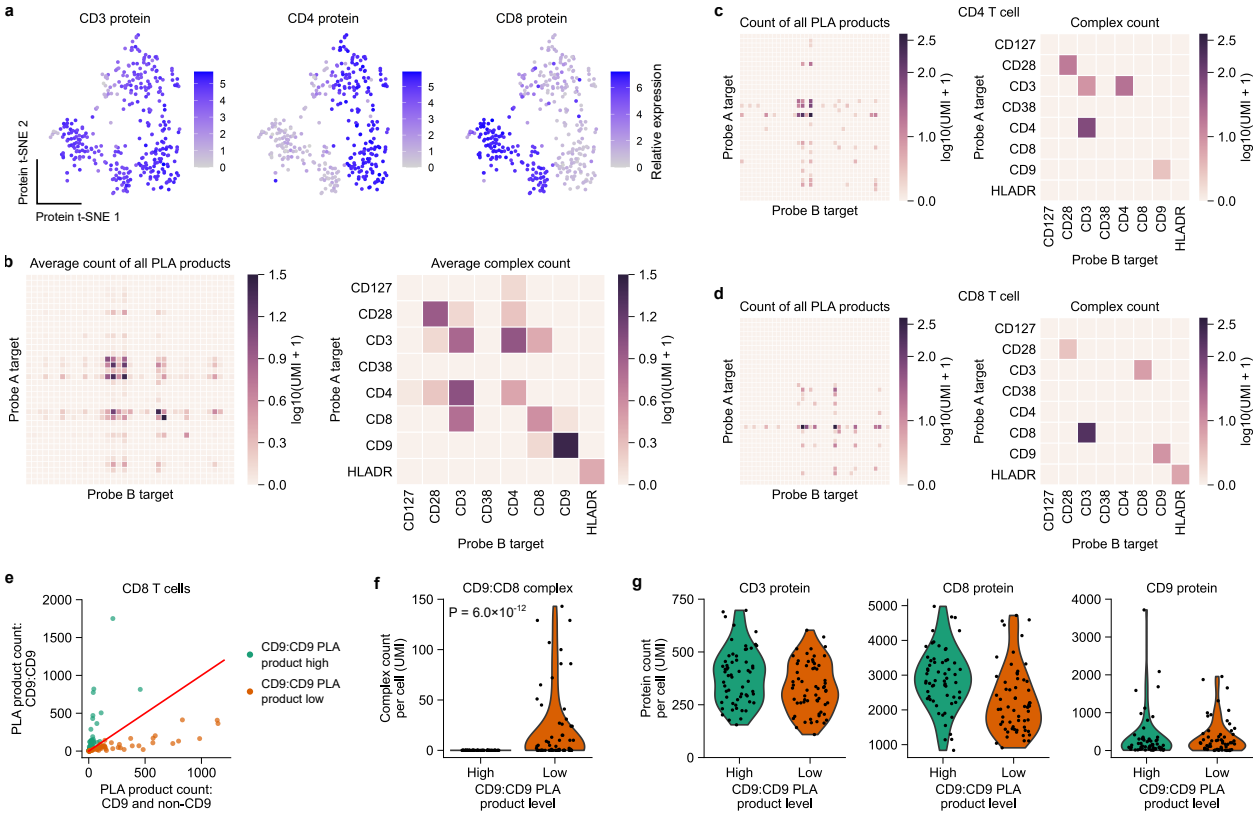

Figure 3.4: Prox-seq reveals a novel CD9:CD8 interaction in peripheral blood mononuclear cells (PBMCs). (a) t-SNE plots, made from protein data, showing that Prox-seq can identify CD8 and CD4 T cells based on normalized expression of CD3, CD4 and CD8 proteins (centered log-ratio transformation). The single cells are colored by relative expression of CD3, CD4 and CD8 proteins. (b) Heatmaps showing the average count of PLA products (left) and complexes predicted from the complex detection algorithm (right) across all single cells. The counts were log-transformed before the average is calculated. (c, d) Heatmaps showing the count of PLA products (left) and detected complexes (right) of an example CD4 T cell (c) and CD8 T cells (d). (e) The presence of two CD9 dimerization states can be seen from scatter plot of CD9 homodimer counts compared to CD9 heterodimer counts in CD8 T cells. Cells were divided into two groups based on the red line, which indicates y = x. (f) Violin plot showing the distribution of the protein complex CD9:CD8 in the two subpopulations of CD8 T cells (P-value is calculated with one-sided Wilcoxon rank-sum test). (g) Violin plots showing the distribution of proteins CD3, CD8 and CD9 in the two subpopulations of CD8 T cells.

PLA product) to a lesser degree, CD4:CD9 protein complex was not detected in these cells (Supplementary Figure 3.17).

3.3.4 Prox-seq Shows Modest Correlation Between Transcripts and Proteins, and Identifies CD9:CD8 as a Protein Complex in Naive CD8 T Cells

To more thoroughly explore the interplay between protein complexes and mRNA, and to identify the biological functions of the two CD8 T cell subpopulations, we performed a matching experiment using the 10x pipeline. This experiment yielded simultaneous measurements of mRNA, protein complexes, and protein levels for over 8,700 single cells. We were able to cluster cell types based on their mRNA levels, and dimensional reduction showed that the PLA product information correlates well with the cell types identified by the mRNA information (Figure 3.5a,b).

Next, we investigated the correlation between mRNA and protein levels for each of our targets. We once again found that mRNA and protein are correlated on the level of clusters, but only modestly correlated on the single cell level (Figure 3.5c, Supplementary Figure 3.18). Furthermore, pairwise correlation among PLA products was lower than that among proteins (Supplementary Figure 3.19). Nevertheless, PLA products reflected the protein expression levels of various clusters (Figure 3.5c, d).

In total, we identify 37 unique protein complexes in this sample, which largely overlaps with the 20 complexes identified in the plate-based data set (Supplementary Figure 3.20). Of those 37 protein complexes, 21 of them are supported in the literature or the IntAct protein complex database [137]. Prox-seq failed to identify 8 protein complexes found in the IntAct database. This is likely because each of these complexes included a protein with a median expression of less than 5 UMIs per cell.

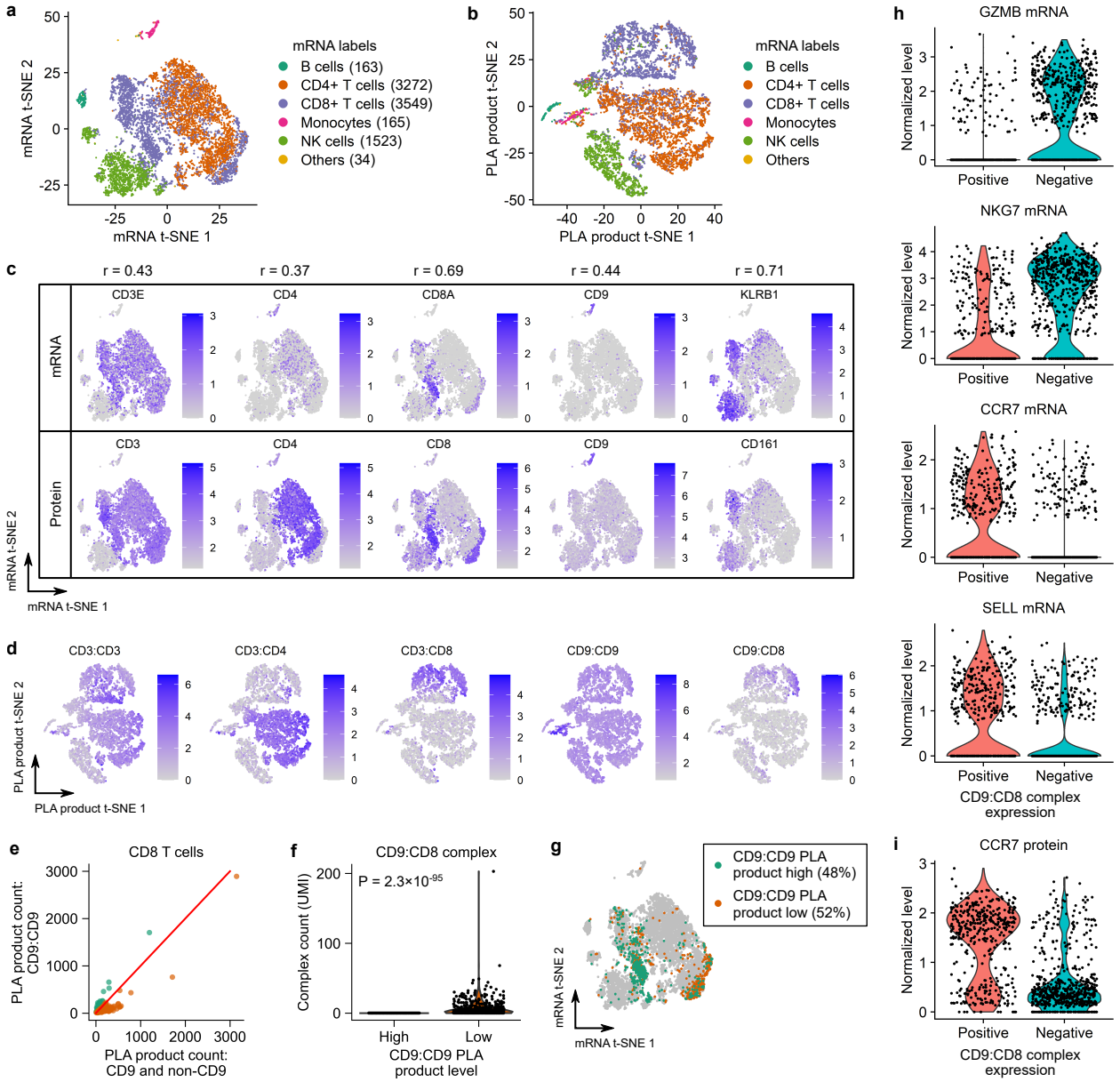

Figure 3.5: Prox-seq enables multiomics analysis of 8,700 single PBMCs. (a, b) t-SNE plots of single cells clustered on (a) mRNAs or (b) PLA products. In (a, b), the cells are labeled using mRNA data. (c) t-SNE plots showing correlation between mRNA and protein level. The plots also show the Pearson's correlation coefficient, r, for each mRNA-protein pair. (d) t-SNE plots showing the normalized expression of of select PLA products. (e) Scatter plot showing two subpopulations of CD8 T cells, according to CD9 homodimer level. Cells were divided into two groups based on the red line, which indicates y = x. (f) Violin plot showing that the CD9:CD8 protein complex is present in the CD9 homodimer-low subpopulation of CD8 T cells (one-sided Wilcoxon rank-sum test). (g) t-SNE plot showing the location of the two subpopulations of CD8 T cells.

Figure 3.5: (continued) (h) Violin plots showing that the subpopulation of CD8 T cells expressing CD9:CD8 protein complex are downregulated in activation markers (*GZMB* and *NKG7* mRNAs) and upregulated in naive T cell markers (*CCR7* and *SELL* mRNAs). (i) Violin plot showing that the subpopulation of cells expressing CD9:CD8 complex also upregulates CCR7 protein.

Our PBMC antibody panel enabled measurement of up to 741 protein complexes. Of those 741 potential complexes, we identified 37 as being present, which largely overlapped with the 20 complexes identified by plate-based methods (Supplementary Figure 3.20, Supplementary table 3.3). Of those 37 protein complexes, 21 of them were supported in the literature or the IntAct protein complex database [137] (Supplementary table 3.3). Prox-seq failed to identify 8 protein complexes found in the IntAct database. Each of these complexes included a protein with a median expression of less than 5 UMIs per cell, which suggested that Prox-seq was not effective at detecting interactions of low-abundance proteins.

Nonetheless, even with the reduced sensitivity of droplet-based method, we could reproduce the findings from the plate-based method that CD8 T cells could be separated into two subpopulations based on the level of CD9:CD9 PLA product (Figure 3.5e). In the cell subpopulation with low CD9:CD9 PLA product, CD9 was again found to be in a protein complex with CD8 (Figure 3.5f). In the cell subpopulation with low CD9:CD9 PLA product, CD9 was found to be interact with CD8 (Figure 3.5f).

By leveraging mRNA data, we found that these two CD 8 T cell subpopulations displayed very different transcriptional profiles (Figure 3.5g). Cells without the CD9:CD8 protein complex showed upregulation of *GZMB* and *NKG7* genes (Figure 3.5h). These genes are markers of activated lymphocytes [138]. Conversely, cells with the CD9:CD8 protein complex displayed upregulation of *SELL* and *CCR7* genes, both of which are markers for naive T cells (Figure 3.5h). Furthermore, the upregulation of *CCR7* was apparent at the protein level as well (Figure 3.5i).

Taken together, these data suggest that the presence of the CD9:CD8 protein complex is a marker of naive T cells. Moreover, we showed how Prox-seq could enable identification of

cell types using protein proximity information, and how the transcriptomic and proteomic information could be used to investigate their biological functions. We note that it was unlikely that the activation status displayed by some cells was a response to our Prox-seq panel. While the probe panel did include stimulatory antibodies, the entire time course of antibody exposure was 30 minutes, far less than the time typically required to activate T cells [139].

3.3.5 Prox-seq Reveals Receptor Interaction Dynamics During TLR Signaling in Single Macrophages

Prox-seq is particularly well-suited for studying cell signaling. Such processes are primarily mediated by the formation and dissociation of protein complexes, and cannot be studied by measuring RNA expression or protein expression alone, especially at the earlier time points. To explore this potential, we developed a panel of Prox-seq probes targeting 15 proteins known to be involved in the NF-κB pathway, a central mediator of innate immunity [3] (Supplementary table 3.4). Some of these proteins are activators of NF-κB (such as TLR2 and TLR4 receptors), and others are under regulation of the NF-κB pathway. First, primary human macrophages in culture were exposed to bacterial signaling ligands that activate the NF-κB pathway in the form of lipopolysaccharide (LPS), Pam2CSK4 (PAM), or both. For each ligand condition, cells were stimulated for 5 minutes, 2 hours, or 12 hours (Figure 3.6a). Then, the cells were harvested, fixed with paraformaldehyde (PFA), and processed with the plate-based Prox-seq pipeline (Supplementary Figure 3.9e). We chose to fix the cells cells were fixed with PFA after ligand stimulation, and before staining with Prox-seq probes, to prevent the antibodies from inducing signaling activity.

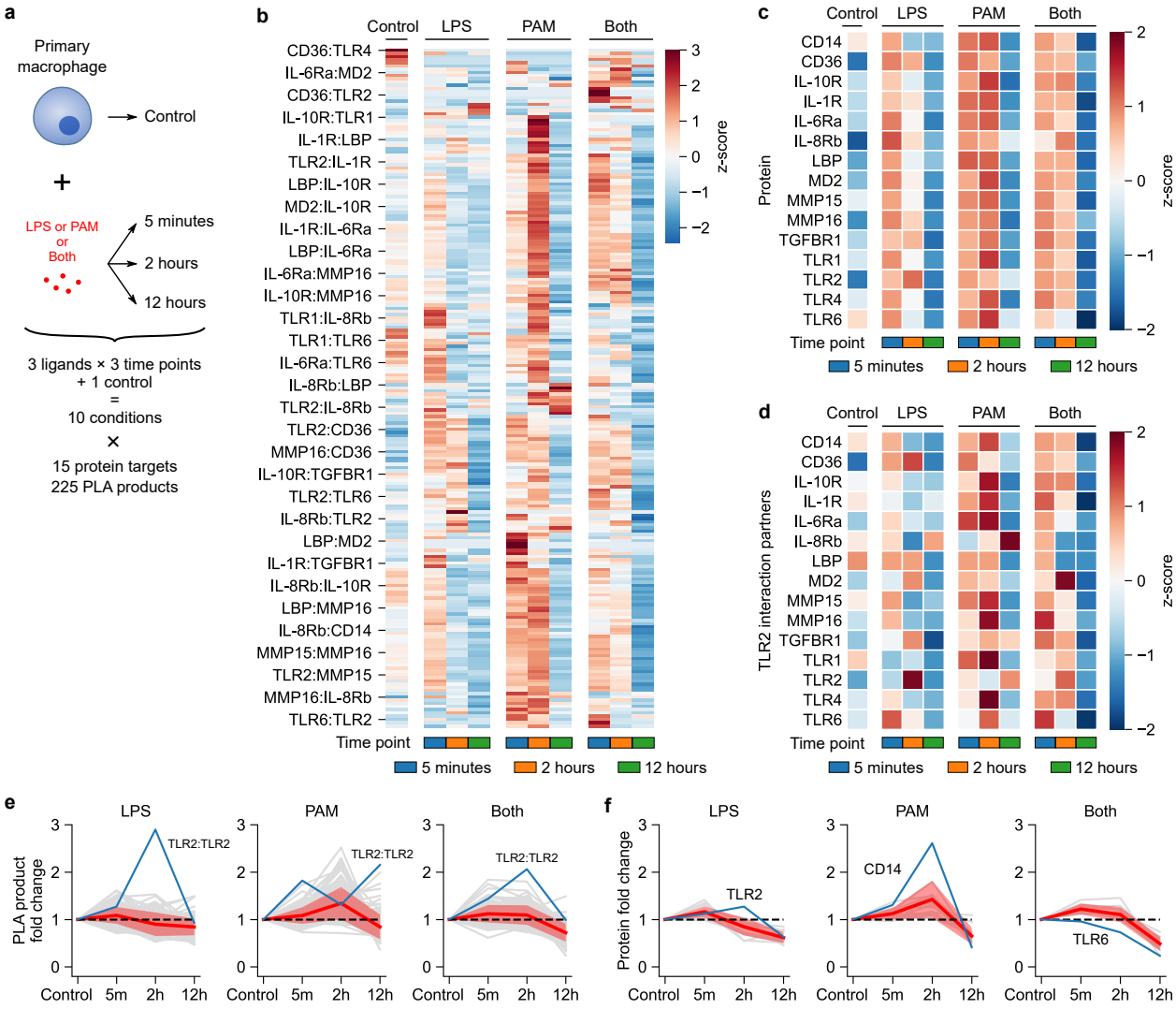

Figure 3.6: Prox-seq enables the study of receptor interaction's dynamics under combined TLR stimulation in macrophages. (a) Schematic of the experiment. Primary human macrophages are treated with either LPS, Pam2CSK4 (PAM), or both LPS and PAM for 5 minutes, 2 hours or 12 hours. The cells are then fixed and processed with Prox-seq. LPS activates TLR4, and PAM activates TLR2, and both receptors signal to the NF-κB pathway. Untreated cells are included as the control. Cells are then processed with plate-based Prox-seq, resulting in detection of 15 protein targets and up to 225 PLA products. (b) The general time course of stimulation response can be seen from a heatmap showing the average expression of all PLA products across the 10 conditions. For visualization purpose, the rows are clustered with hierarchical clustering (Euclidean distance metric and complete linkage), and the dendrogram is hidden. (c) Heatmap showing the average expression of all proteins across the 10 conditions. (d) The binding partners for TLR2 differ significantly from the average protein expression values, as displayed in a heatmap showing the average of all binding partners with TLR2. In (b–d), the UMI counts are log-transformed, then averaged by condition and standardized to calculate the row-wise z-score.

Figure 3.6: (continued) (e, f) LPS displays a fast response, PAM displays a delayed response, and both displays a sustained response as measured by the average fold change of (e) all PLA products, and (f) all proteins for each type of ligand. A pseudocount of one UMI was added to the numerator and denominator for fold change calculations (see "Methods"). The grey lines indicate the fold change of individual PLA products or proteins, the red lines indicate the average of all PLA products or proteins, and the red bands indicate the standard deviation. In (e), the blue lines indicate the fold change of TLR2:TLR2 PLA product for each stimulation condition.

Overall, there was a clear trend of increasing levels of PLA products following stimulation through 2 hours, and a sharp decline at 12 hours (Figure 3.6b). However, this trend was not universal, with some PLA products rising though the entire time course or appearing only at 12 hours for some treatment conditions (Figure 3.6b). In contrast, protein levels were consistently lower at 12 hours (Figure 3.6c). We found that the tendency for a protein to produce a pair is not strictly a result of protein expression levels. This can be clearly seen in the case of TLR2, which displays major changes in its preferred PLA product partners, depending on both time and stimulant, that do not track with the protein levels for these partners (Figure 3.6d). Consistent with previous single live-cell imaging studies on NF-κB dynamics, LPS stimulation displayed a faster response with PLA products peaking at 5 minutes, whereas PAM displayed a slower response that peaked at 2 hours [140] (Figure 3.6e). When cells were stimulated with both signaling molecules LPS and PAM, the change in PLA products showed traits of both stimuli in an additive manner across time, with a broad peak that was sustained until finally dropping at 12 hours (Figure 3.6e). Proteins showed a similar trend as PLA products (Figure 3.6f). This simple additivity suggested that for the proteins that we measured, these two signaling molecules were operating independently, without synergy or antagonism.

3.3.6 Prox-seq Enables Identification of Immune Signaling Inputs in Macrophages

The dynamics of NF-κB nuclear translocation could be used to predict whether a cell was stimulated with LPS or PAM using live-cell microscopy [140]. We reasoned that the changes in receptor organization measured through Prox-seq should also enable us to predict the stimulating ligand. To explore this possibility, we trained a logistic regression classifier using PLA product count data at each time point. Our classifier was able to identify the response of cells as PAM-like or LPS-like best at the 2-hour time point (Figure 3.7a). The single largest coefficient for this classification was the presence of the PLA product TLR2:TLR2, which was highly elevated in the LPS-like responder cells (Figure 3.7b). We then applied this classifier to the sample of cells stimulated with both LPS and PAM for 2 hours, and classified them into LPS-like, PAM-like, or mixed response cells (Supplementary Figure 3.21).

In addition to identifying the stimulating ligand, several PLA products were determined to indicate the presence of a protein complex. Each of these complexes showed dynamic regulation and some, but not all, could distinguish between stimulants (Figure 3.7c). Consistent with the logistic regression classifier's results, the TLR2:TLR2 protein complex appeared 2 hours after LPS treatment in macrophages, then disappeared at 12 hours (Figure 3.7c). In contrast, under PAM stimulation, this protein complex was absent at early time points (before 2 hours), and appeared only after 12 hours (Figure 3.7c). While the TLR2 homodimer is known to exist, it is not previously believed to participate in either LPS or PAM signaling [141, 142].

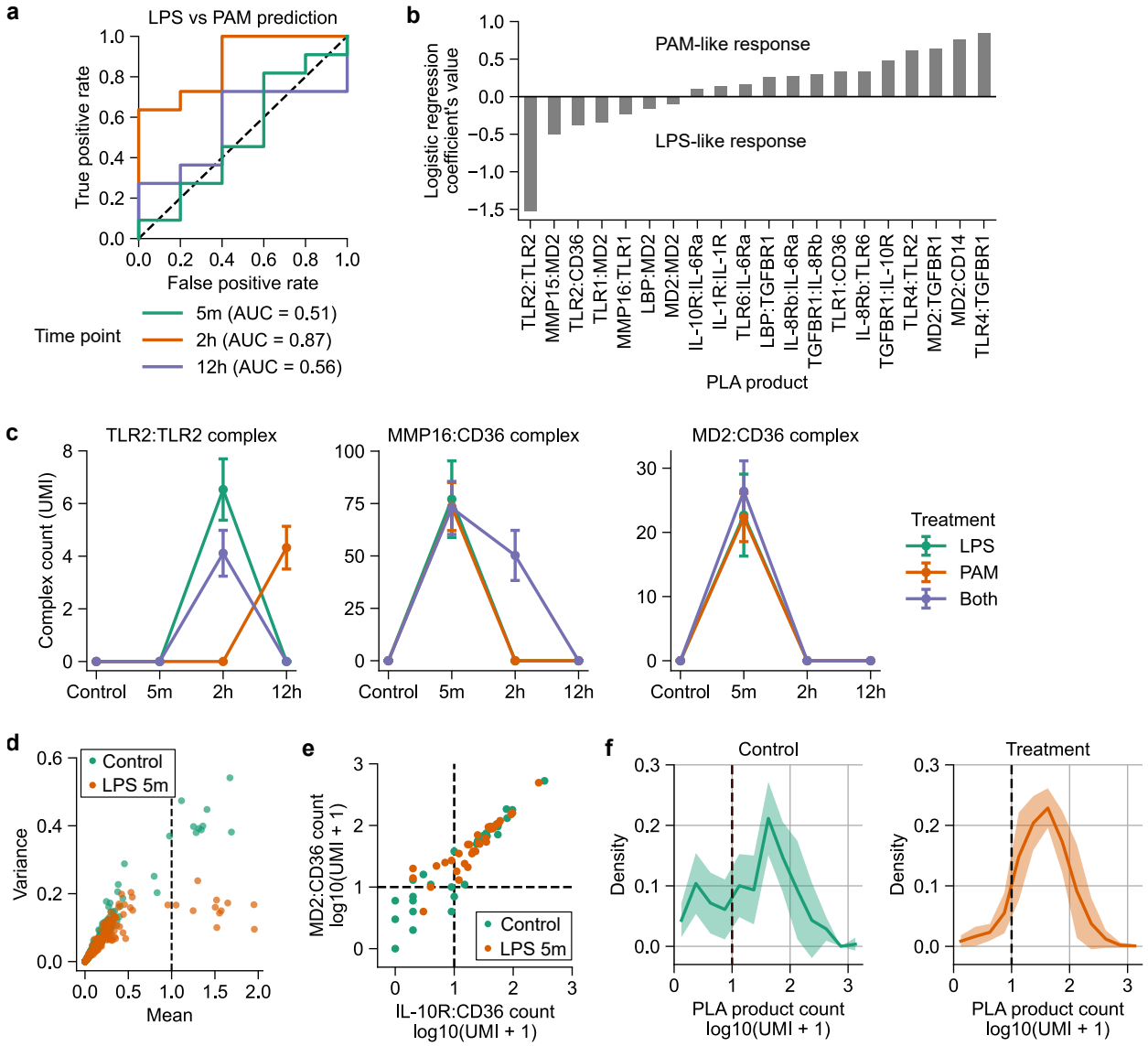

Figure 3.7: Prox-seq reveals single-cell stochasticity in TLR signaling and enables prediction of signaling stimulants. (a) Receiver operating characteristic curves of a logistic regression classifier using PLA product count from different time points. The classifier is trained to predict whether a single cell was stimulated with LPS or PAM. (b) Bar plot showing the PLA products that contribute to the LPS vs PAM prediction. Only products with absolute coefficients above 0.1 are shown. A positive value indicates that the PLA product is higher in PAM-treated cells; a negative value indicates that the PLA product is higher in LPS-treated cells. (c) Plots showing the dynamics of 3 example protein complexes. Data is presented as mean ± s.e.m. (d) Scatter plot showing the relationship between the mean and variance of log-transformed PLA count of control sample, and sample treated with LPS for 5 minutes. The PLA products with mean greater than or equal to 1 are CD36-related. (e) Scatter plot showing relationship between IL-10R:CD36 and MD2:CD36 PLA products in control group and 5-minute LPS treatment group.

Figure 3.7: (continued) (f) Plot showing the distribution of PLA product counts of the nine CD36-replated PLA products in control (left) and treatment (right) sample. The solid lines indicate the mean, and the ribbons indicate the standard deviation.

3.3.7 Prox-seq Reveals Stochasticity in TLR Signaling

Finally, we explored the stochasticity in single-cell signaling responses in our PLA product data. When we compared the mean and variance among all PLA products, we observed a sharp decrease in variance in all stimulation conditions compared to the control group (Figure 3.7d, Supplementary Figure 3.22). We observed that the low-variance PLA products all contained a CD36 Prox-seq probe (Figure 7e, Supplementary Figure 3.23). Histograms of all PLA products containing CD36 showed two modes in control cells, separated by the number of UMIs (Figure 3.7f). LPS treatment caused cells to shift to the higher expression mode (Figure 3.7f). It should be noted that, because this change is already occurring at the 5-minute time point, the increase in PLA products is unlikely to be a result of increased protein expression. Rather, CD36 is involved in interactions and rearrangement of the other proteins targeted by our probe panel.

This is further supported by the appearance of new CD36 protein complexes at 5 minutes (Figure 3.7c). CD36 is a scavenger receptor that recognizes a variety of bacterial lipid and lipoprotein molecules [143]. It has also been shown act as a co-receptor for both TLR2 and TLR4, both of which are stimulated in response to the ligands in our study [144, 145]. Furthermore, stimulation with oxidized LDL can induce CD36 to form a protein complex with the TLR4 and TLR6 [146]. Overall, this shows that Prox-seq can identify rearrangement of surface receptors and stochasticity during TLR signaling.

3.4 Discussion

In summary, we presented Proximity sequencing (Prox-seq) as a practical and broadly applicable method for multiomics measurements of surface proteins, protein complexes and mRNAs in single-cells. We also demonstrated its application in several cell types and under different biological contexts. Prox-seq will be a valuable tool for understanding cellular functions, and for identification of cell types and cell states. Prox-seq is particularly useful for studying many biological functions such as signaling, differentiation, development and cellular decision making, which are largely driven by changes in protein interactions. A broader insight into these interactions can assist in the aspiring goal of constructing mathematical models that can more accurately predict cellular behavior from knowledge of their constituent parts.

We showed that this technology is compatible with commonly used single-cell sequencing methods like droplet-based (Drop-seq and 10x Chromium) and plate-based methods (Smart-seq2). This allows easy adoption by many laboratories and researchers. Most importantly, Prox-seq can identify members of pairwise protein complexes, providing an entirely new modality to single-cell analysis. While we only showed detection of surface proteins in intact single cells in this study, Prox-seq can theoretically be applied to intracellular proteins as well as cell lysates [16, 116]. Development of intracellular Prox-seq and demonstration of these potential applications are currently underway in our group.

There are some limitations inherent to the design of Prox-seq. Several of these stem from the limitations antibodies, and therefore antibody selection should be carefully considered. Monoclonal antibodies were primarily used in this study because they enabled detection of homodimers. However, monoclonal antibodies are likely to suffer from a higher false negative rate than polyclonal antibodies. This is because it is typically difficult to know if the epitope that a monoclonal antibody bind to is positioned in such a way to allow both antibody binding and protein complex formation. Indeed, we believe this property might have hindered our ability to measure the B7 homodimer. Polyclonal antibodies, by

virtue of having multiple epitopes, should ameliorate some false negative concerns at the cost of losing the ability to detect homodimers specifically. As with any other antibody-based assay, antibodies should be validated for compatibility with Prox-seq. In addition, antibody assays are typically stimulatory when the antibody is directed at a receptor. When this is an undesirable property, cells should be fixed prior to Prox-seq analysis, like our study of TLR signaling in macrophages (see "Methods"). Most single-cell sequencing methods are compatible with fixation, however there is usually some loss of data quality [147, 148] (see Chapter 4).

Despite these limitations, Prox-seq was still able to display accurate identification and clustering of T cells and B cells, measure known and potentially new protein complexes in human PBMCs, and study receptor rearrangement and complex formation during TLR signaling in human macrophages. We were able to detect known protein complexes such as CD3 homodimer and CD28 homodimer in single T cells. We also identified a novel receptor interaction between CD8 and CD9 in a subset of human primary CD8 T cells. Lastly, we observed different temporal changes in receptor arrangements under LPS and PAM stimulation in macrophages, and showed additive integration of TLR signals. This is the first time such receptor dynamics has been observed in single cells in a high-throughput manner, proving the utility of Prox-seq for signaling and immunology studies.

Recent advances in single-cell sequencing technology have enabled comprehensive characterization of the transcriptome, genome and epigenome at the single-cell level [54, 149, 150]. Several methods have expanded these approaches to incorporate antibody-based protein measurements [21–23]. Furthermore, the field of single-cell mass-spectrometry has been advancing rapidly [20, 151]. However, the measurement of protein complexes on the single-cell levels has been lagging compared to other modes of measurement. Prox-seq addresses this weakness by providing quadratically-scaled multiplexing capability to greatly increase the number of protein complexes that can be measured in a single cell. Currently, in order to make highly multiplexed measurements of protein complexes, one is limited to bulk sam-

ples. Methods such as yeast-two-hybrid assays and affinity purification/mass spectrometry (AP-MS) can measure several thousands of interactions, but cannot yet be adapted to single human cells [152, 153]. Single-cell methods are very limited in their multiplexing capacity, typically measuring fewer than 10 complexes [111, 154]. With Prox-seq we have demonstrated the ability to survey 741 possible protein complexes in single PBMCs. Furthermore, Prox-seq incorporates single-cell RNA sequencing, thereby providing multiple single-cell data types simultaneously. Finally, future studies will expand Prox-seq functionality to intracellular proteins. This would enable characterization of post-translation modifications, and interactions between proteins and histone modifications. In short, Prox-seq provides a unique mode of proteomic information, and greatly enhances multiomic analysis capability in single cells.

3.5 Supplementary Figures

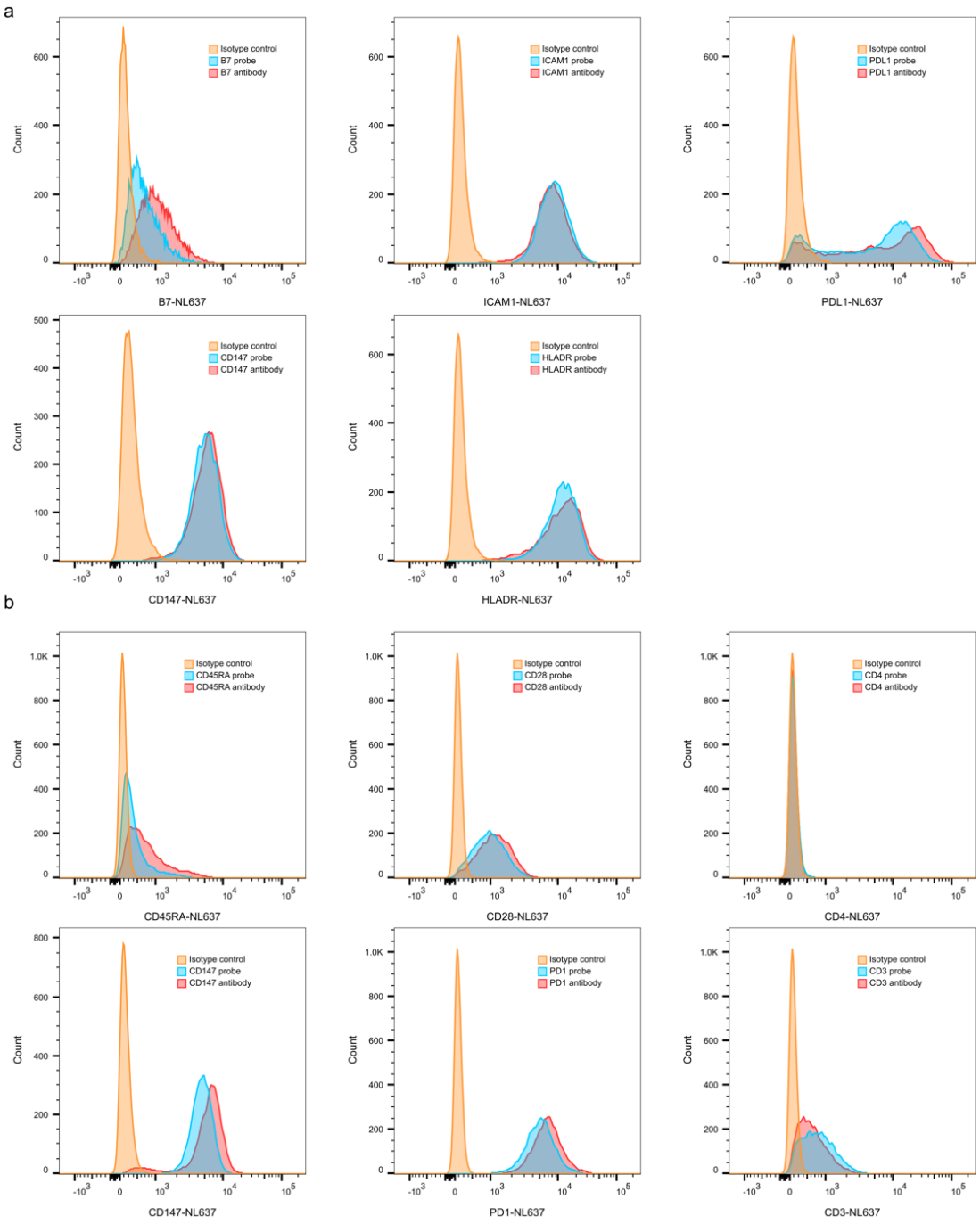

Figure 3.8: Flow cytometry data showing changes in binding before and after DNA oligo conjugation. Raji (a) and Jurkat (b) cells are stained with isotype control (orange), antibody (red) or Prox-seq probe (blue). In each case, Prox-seq probes retain their binding affinity.

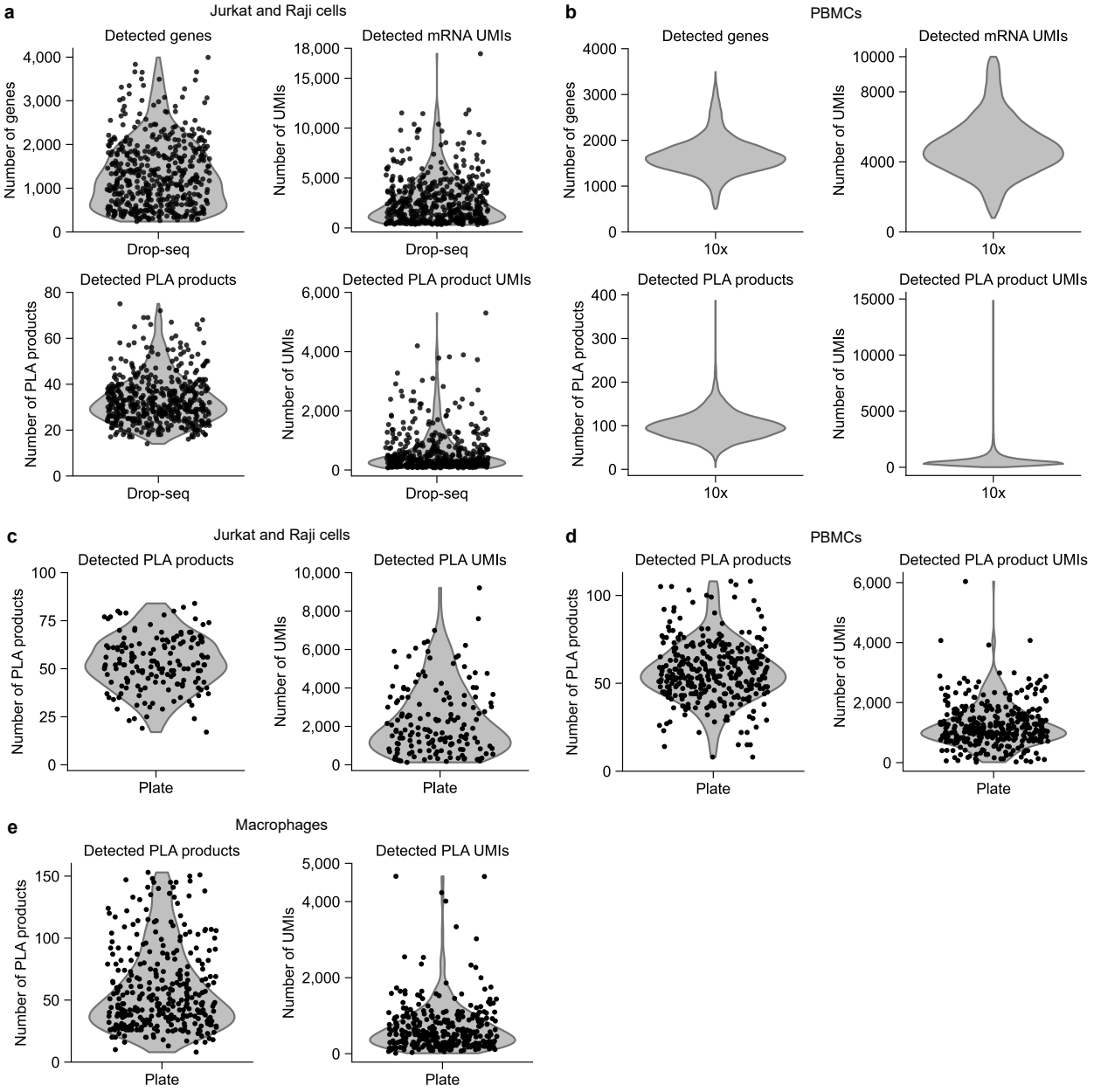

Figure 3.9: Distribution of the number of detected features and UMIs in Prox-seq data from different pipelines and samples. (a) Drop-seq-based Jurkat and Raji sample. n = 475 single cells. The total number of detected PLA products across all single cells is 163. (b) 10x-based PBMC sample. n = 8,706 single cells. The total number of detected PLA products is 1,333. (c) Plate-based Jurkat and Raji sample. n = 158 single cells. The total number of detected PLA products is 142. (d) Plate-based PBMC sample. n = 352 single cells. The total number of detected PLA products is 679. (e) Plate-based macrophage sample. n = 328 single cells. The total number of detected PLA products is 213.

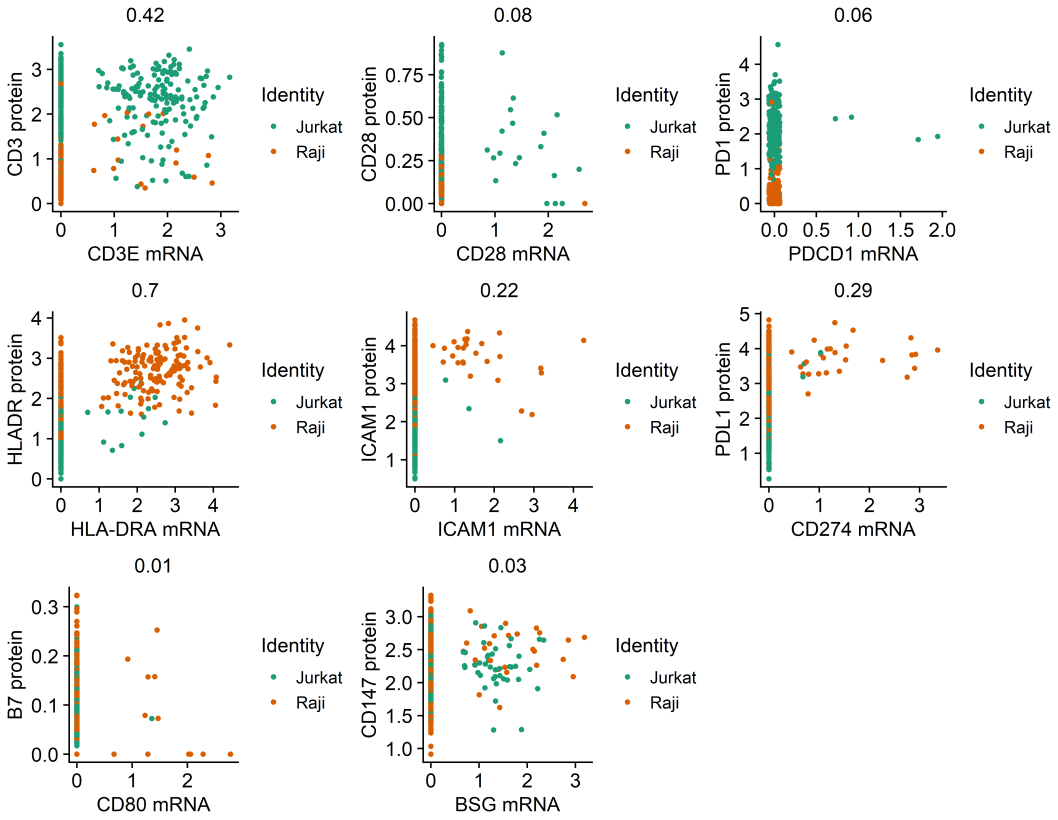

Figure 3.10: Correlation between mRNA and protein levels in Jurkat and Raji cells. Scatter plots showing the normalized level of mRNAs and proteins for 3 Jurkat markers (CD3, CD28, PD1), 4 Raji markers (HLADR, ICAM1, PDL1, B7), and CD147 protein (expressed by both Jurkat and Raji cells). The numbers above the figures indicate the Pearson's correlation coefficients.

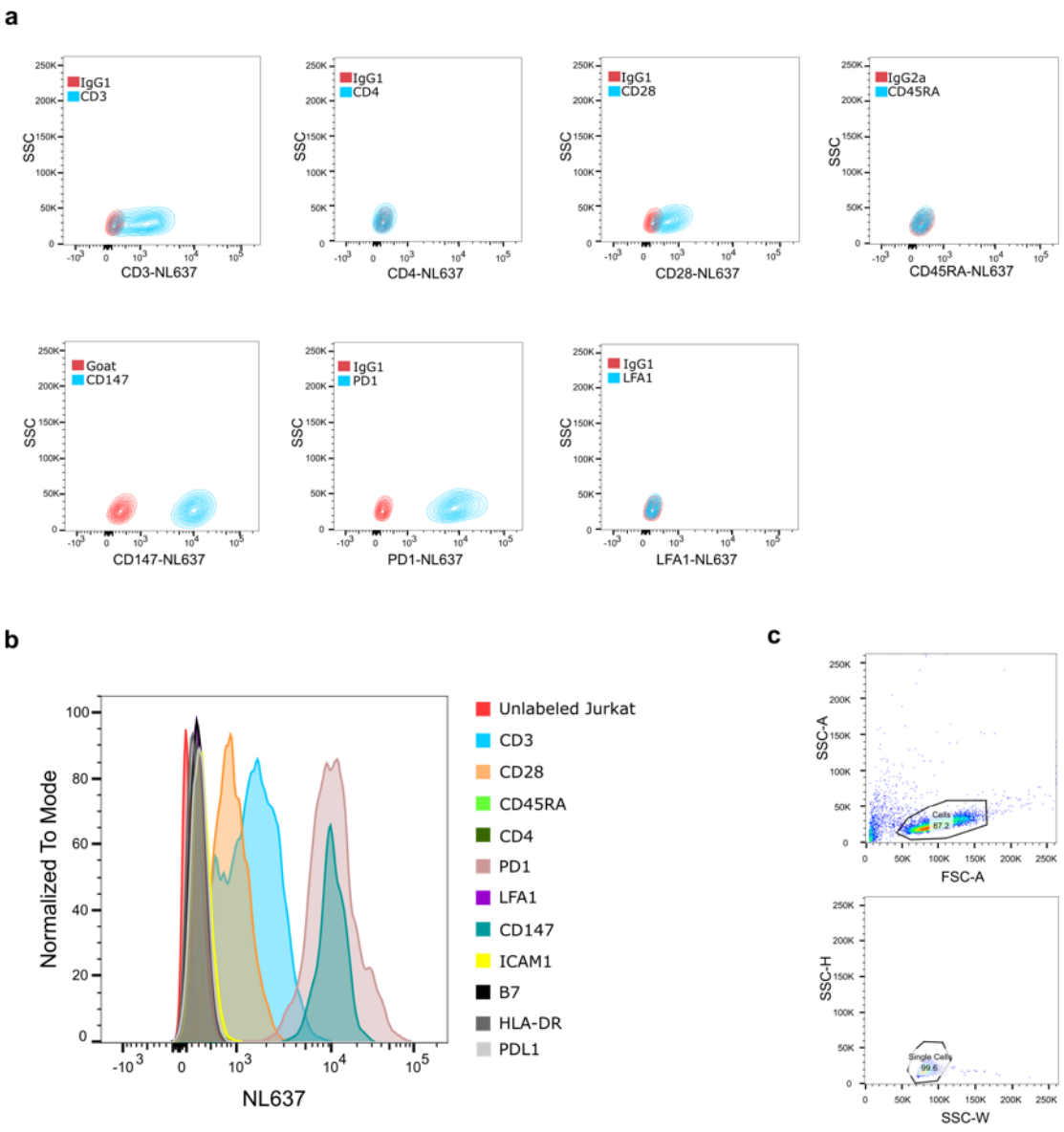

Figure 3.11: Flow cytometry data showing Prox-seq probe binding on Jurkat cells. (a) Each T-cell marker in the panel compared to its corresponding isotype control. (b) The amount of binding for all Prox-seq probes on Jurkat cells. (c) The gating strategy to identify single Jurkat cells.

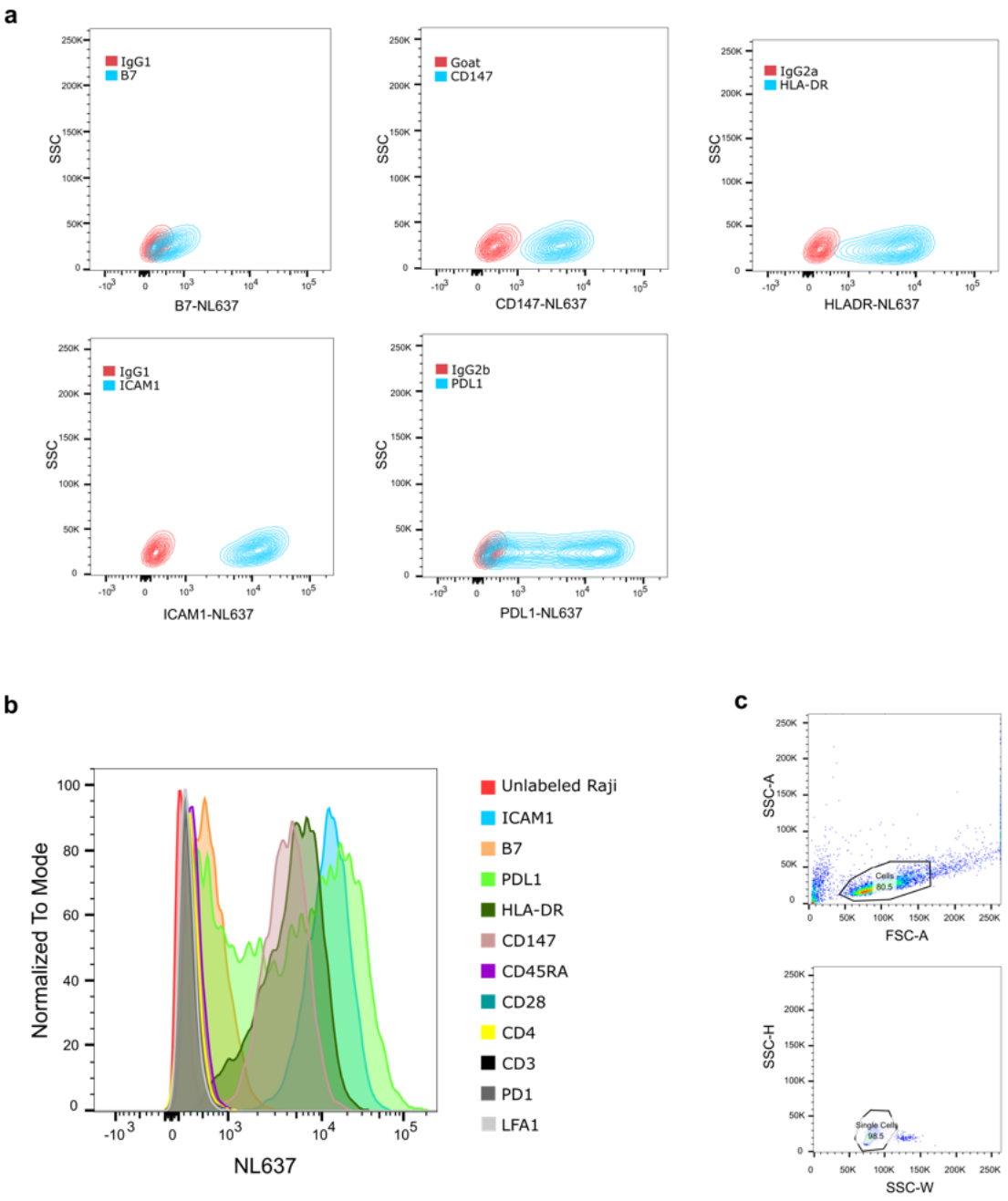

Figure 3.12: Flow cytometry data showing Prox-seq probe binding on Raji cells. (a) Each B-cell marker in the panel compared to its corresponding isotype control. (b) The amount of binding for all Prox-seq probes on Raji cells. (c) The gating strategy to identify single Raji cells.

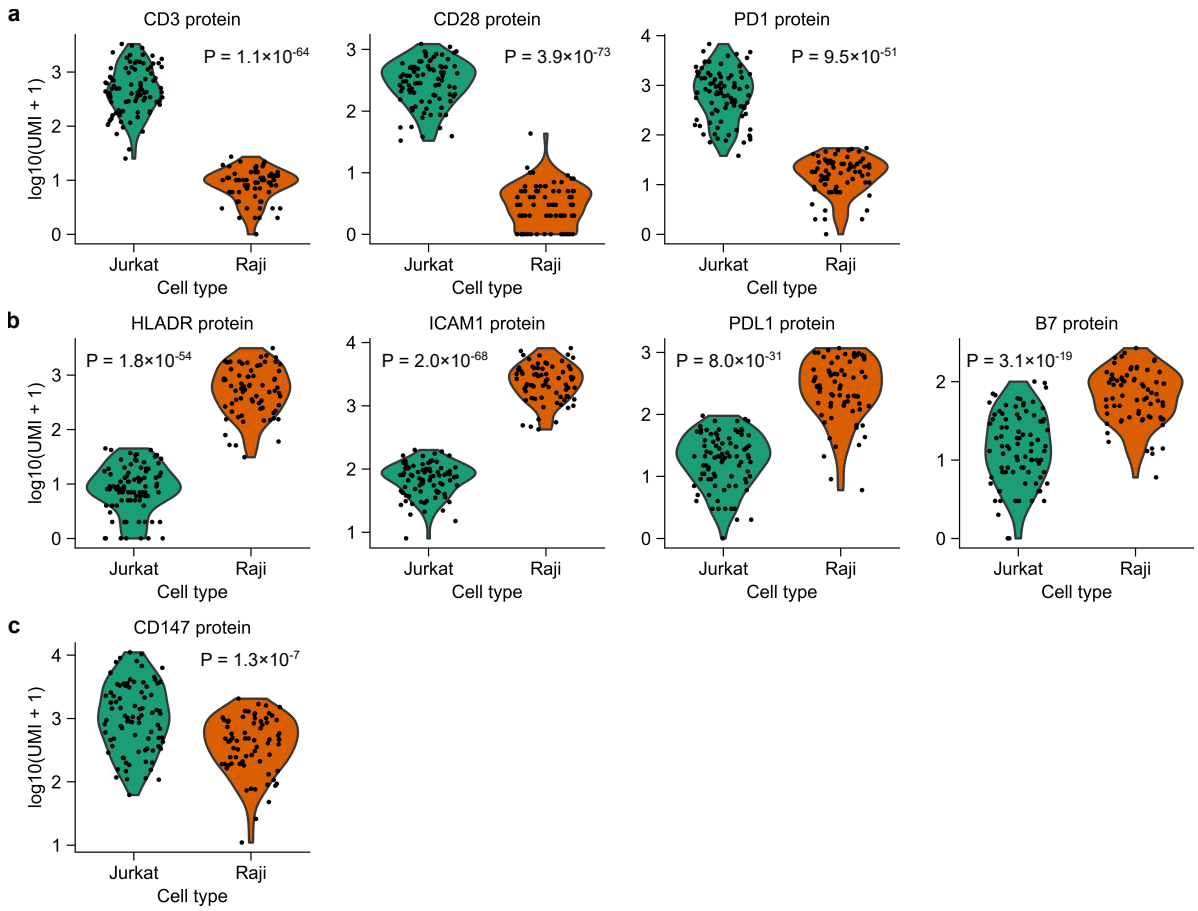

Figure 3.13: Characterization of background signal due to Prox-seq probe non-specific binding. (a–c) Violin plots showing the distribution of protein abundance of (a) Jurkat markers, (b) Raji markers, and (c) CD147 protein, which is expressed in both Jurkat and Raji cells. The protein abundance was log-transformed before plotting. P-values were calculated using Welch's t-test with Benjamini-Hochberg correction.

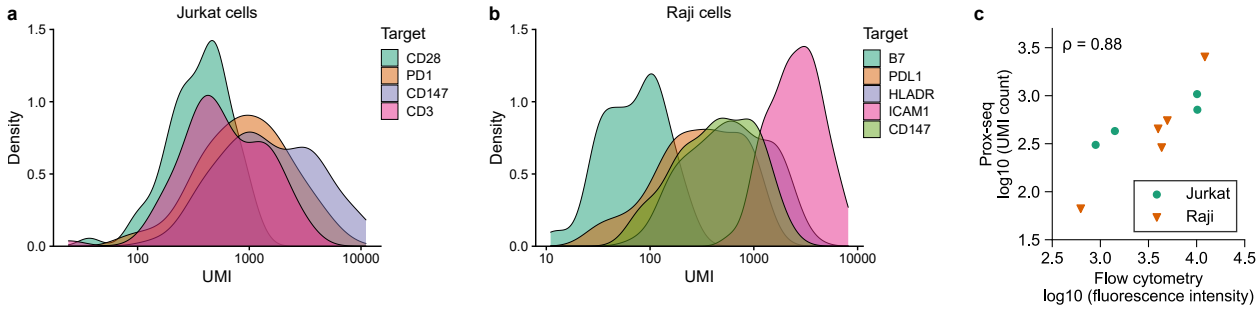

Figure 3.14: Comparison of protein quantification between Prox-seq and flow cytometry. (a, b) Distribution of the single-cell UMI counts of (a) Jurkat and (b) Raji cells' protein markers. (c) Scatter plot showing the median protein abundance as measured by flow cytometry or Prox-seq. Each point indicates a protein. The plot also shows the Spearman's correlation coefficient, ρ, between Prox-seq and flow cytometry measurements.

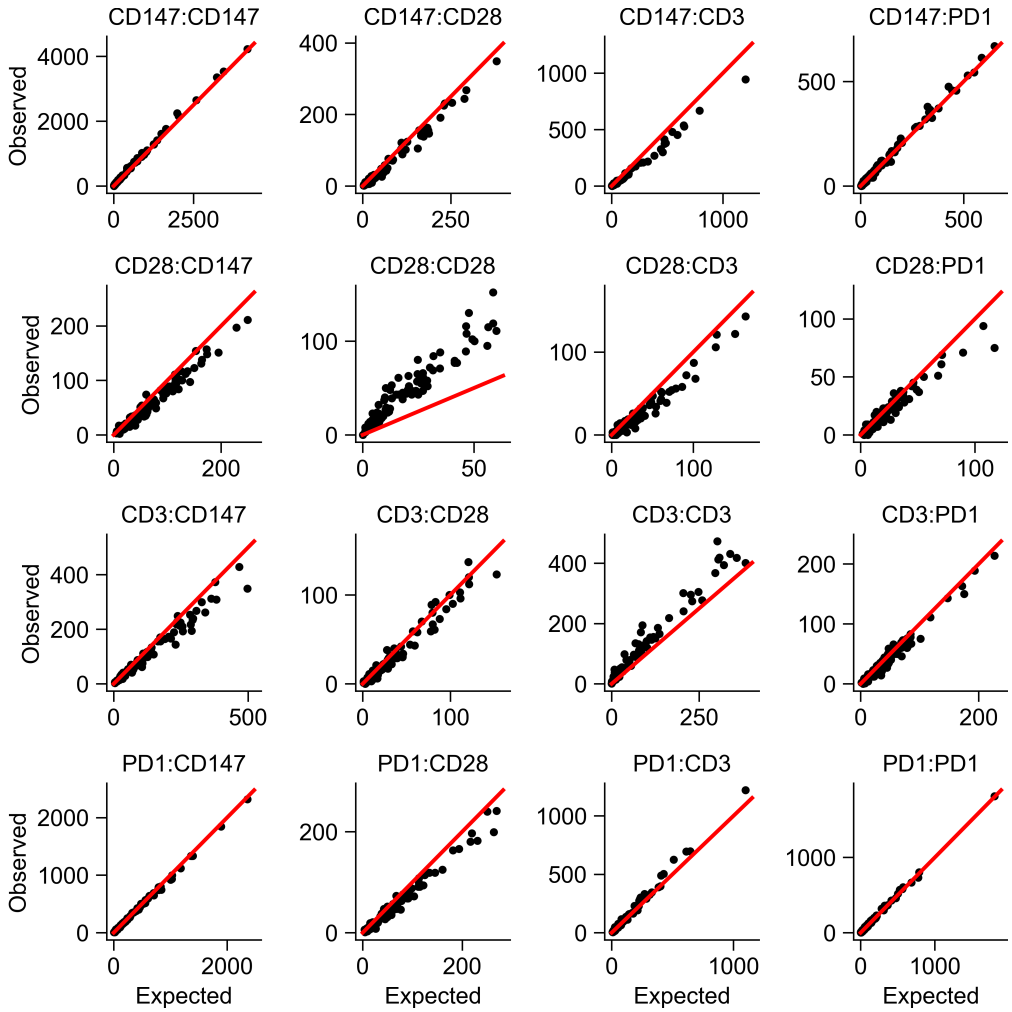

Figure 3.15: Plot of the observed and expected random PLA product counts for T cell markers among Jurkat cells. The observation that some PLA products have higher observed counts than expected indicate the existence of protein complexes in Jurkat cells. The red lines indicate y = x. This figure is related to Figure 3.3a.

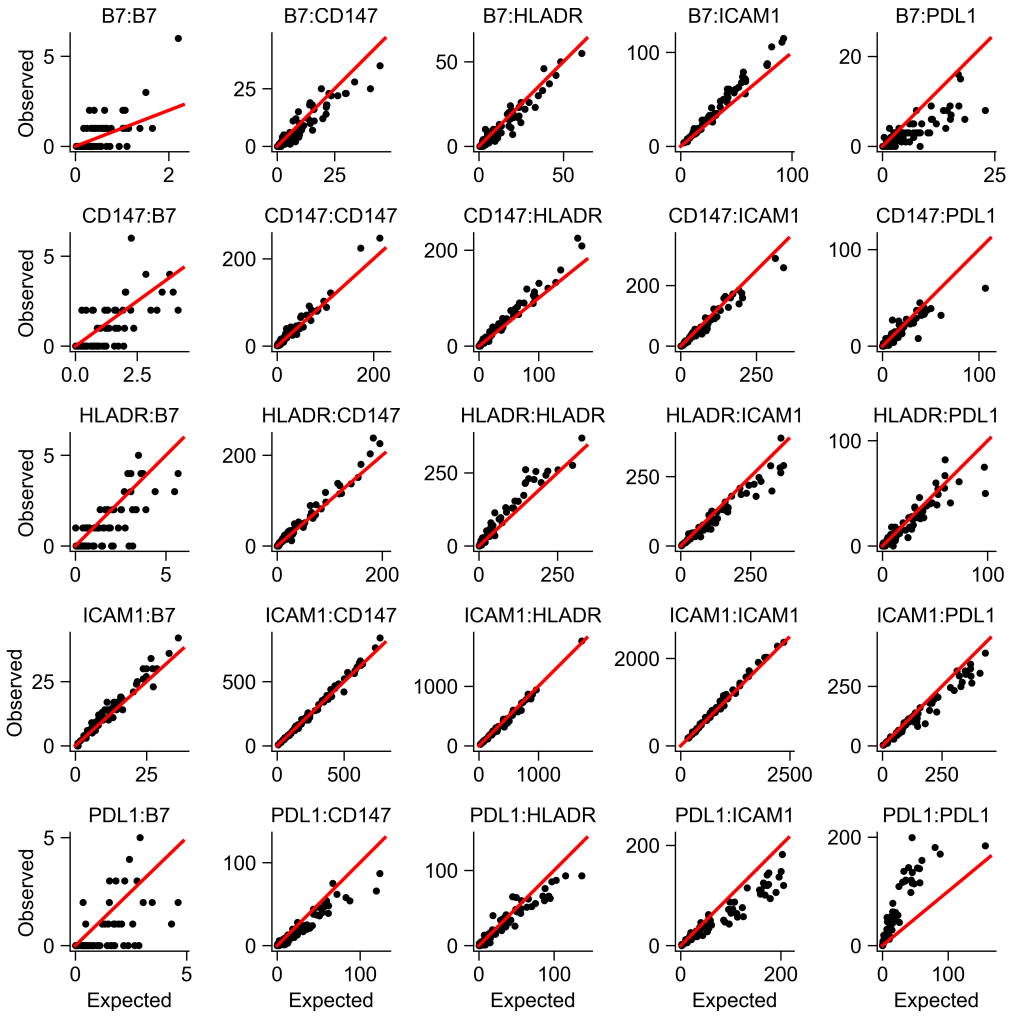

Figure 3.16: Plot of the observed and expected random PLA product counts for B cell markers among Raji cells. The observation that some PLA products have higher observed counts than expected indicate the existence of protein complexes in Raji cells. The red lines indicate y = x. This figure is related to Figure 3.3b.

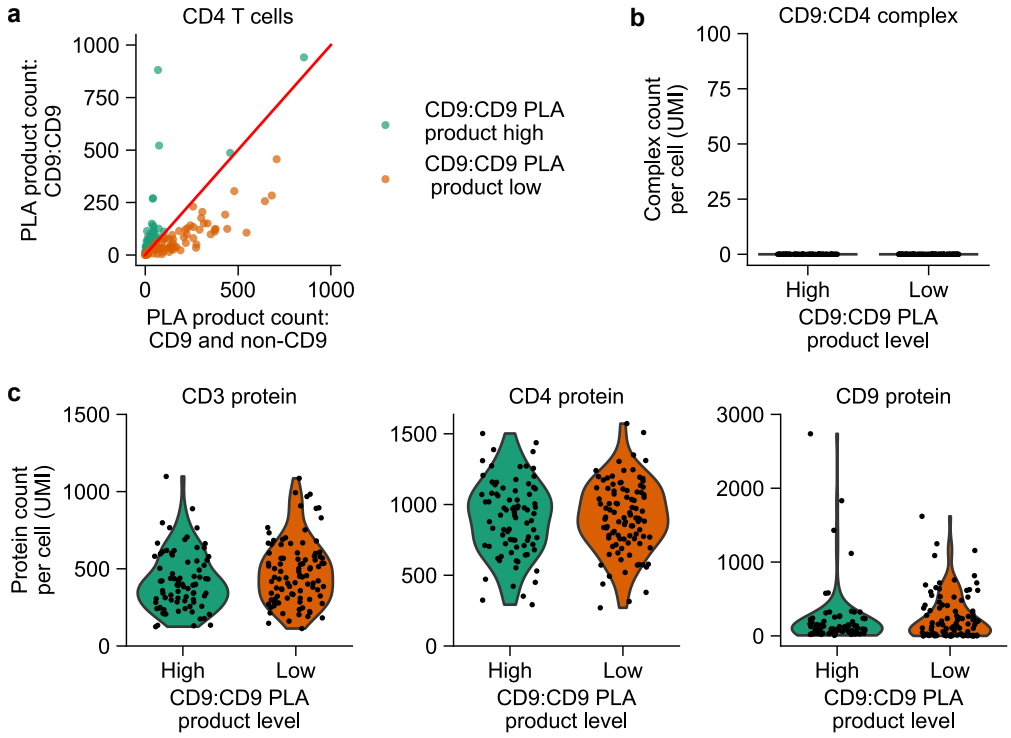

Figure 3.17: Characterization of CD4 T cells in human PBMCs. (a) Scatter plot showing two subpopulations of CD4 T cells, according to CD9-related PLA products level. (b) Violin plot showing that, unlike CD8 T cells, both subpopulations of CD4 T cells do not express the protein complex CD9:CD8. (c) Violin plots showing the distribution of proteins CD3, CD4 and CD9 in the two subpopulations of CD4 T cells.

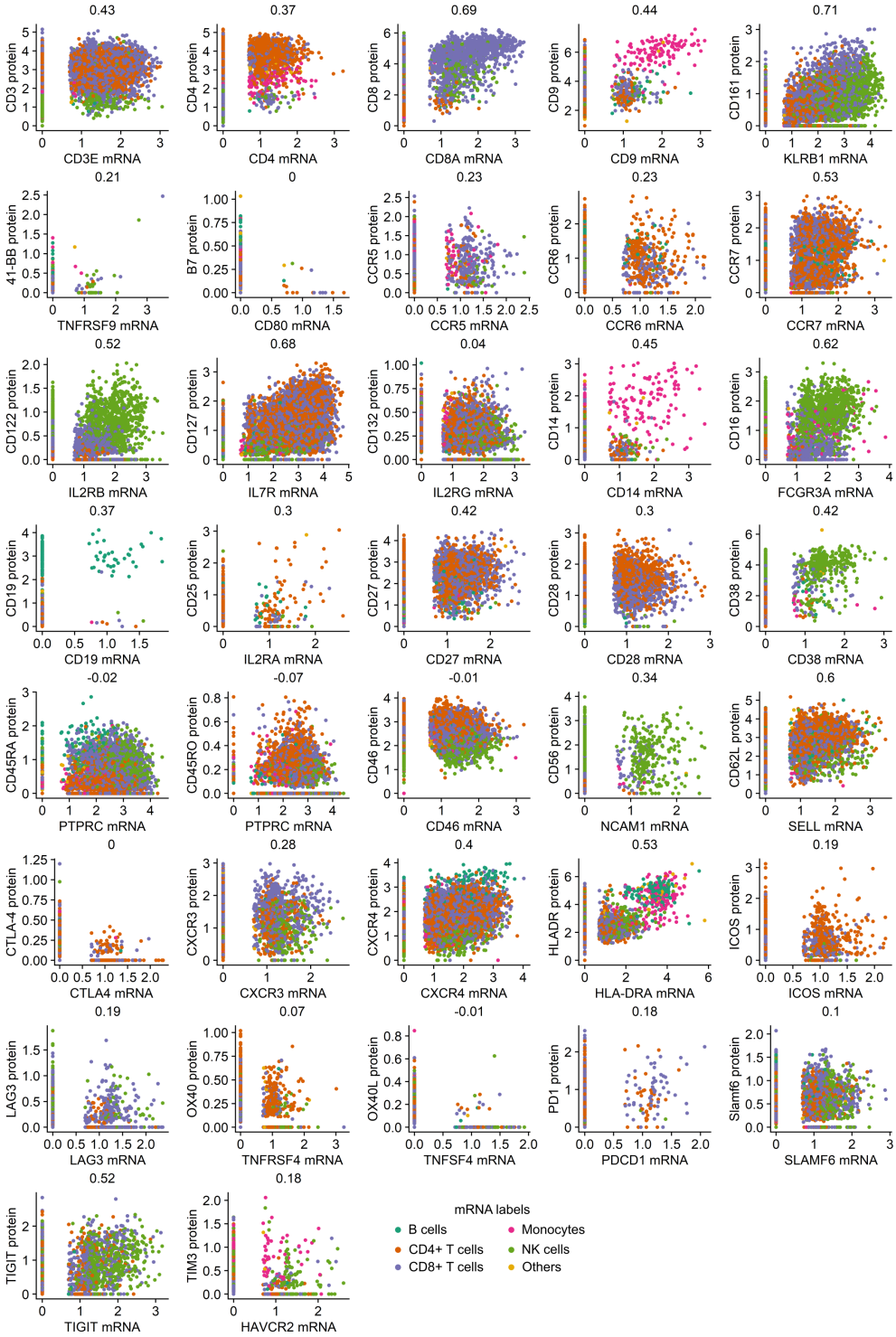

Figure 3.18: Relationship between mRNA and protein levels in PBMCs. Scatter plots showing the relationship between mRNA and protein levels in 37 proteins. The Pearson's correlation coefficients are shown above each plot. The T-cell receptor γ/δ was not plotted because its mRNA (*TRG* gene) was not detected.

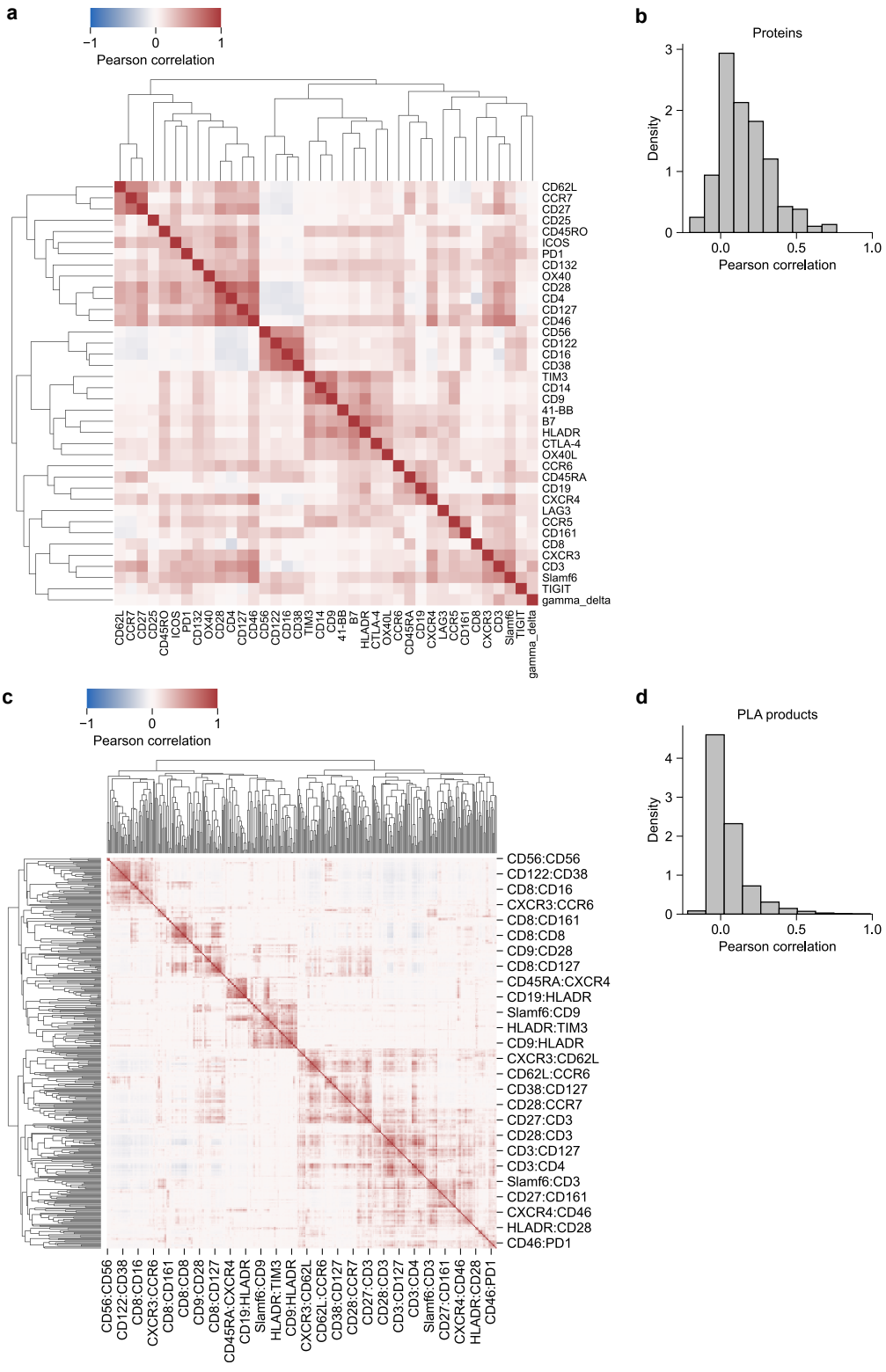

Figure 3.19: Correlation among proteins and among PLA products in PBMCs. (a) Heatmap showing the clustered pairwise Pearson's correlation coefficients of proteins. (b) Histogram showing the distribution of proteins' pairwise correlation.

Figure 3.19: (continued) (c) Heatmap showing the clustered pairwise Pearson's correlation coefficients of PLA products. (d) Histogram showing the distribution of PLA products' pairwise correlation. In (a) and (c), the proteins and PLA products were clustered using the distance metric 1 − Pearson's correlation coefficient, and the "complete" method. In (c) and (d), only PLA products detected in at least 5% of the cells were considered.

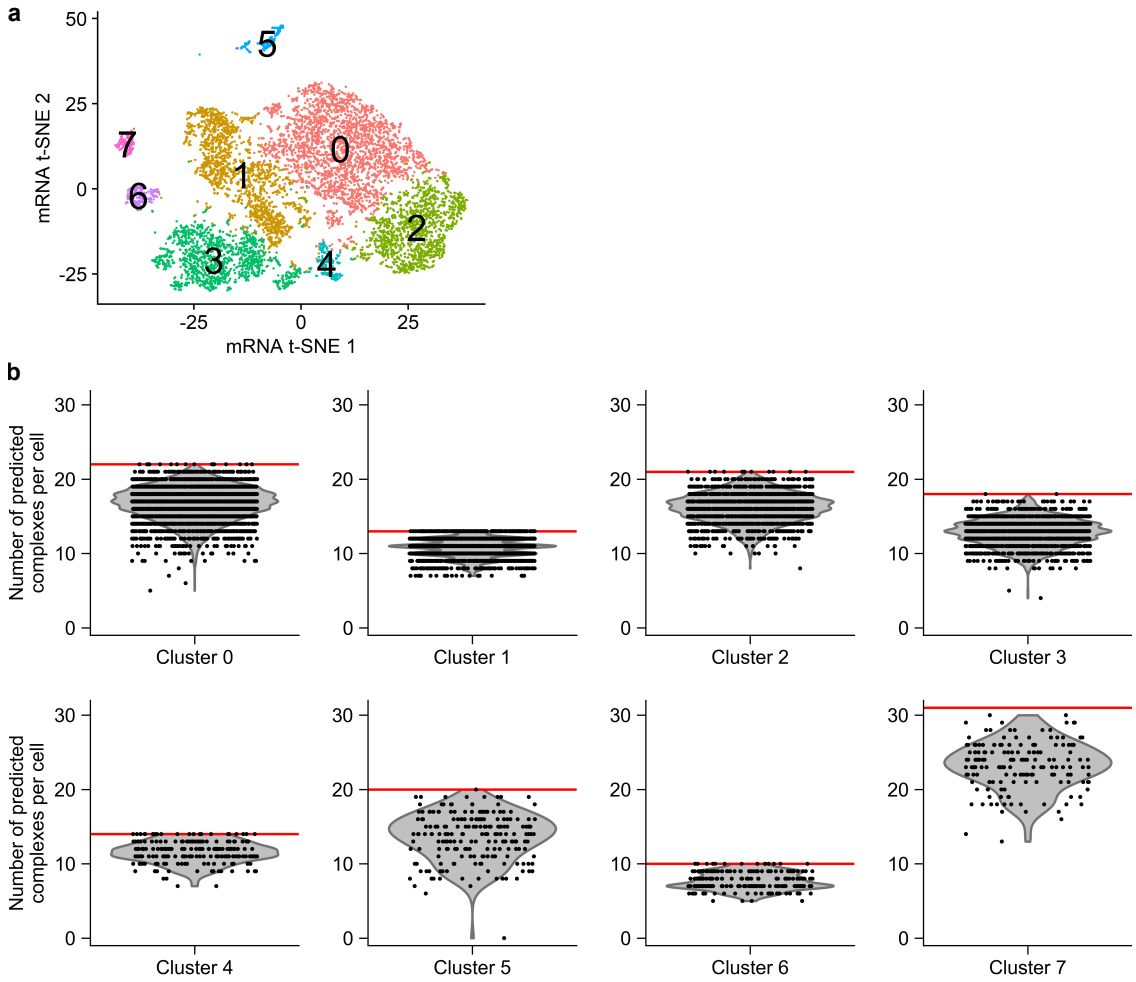

Figure 3.20: Number of predicted protein complexes across cell types. (a) t-SNE plot of PBMC clusters, obtained using mRNA expression profile. (b) Violin plots showing the number of predicted protein complexes per single cell, for each of the 8 clusters identified using mRNA data. The red horizontal lines indicate the total number of predicted protein complexes for each cluster. In total, 61 protein complexes were detected across all 8 clusters, of which 37 complexes are unique.

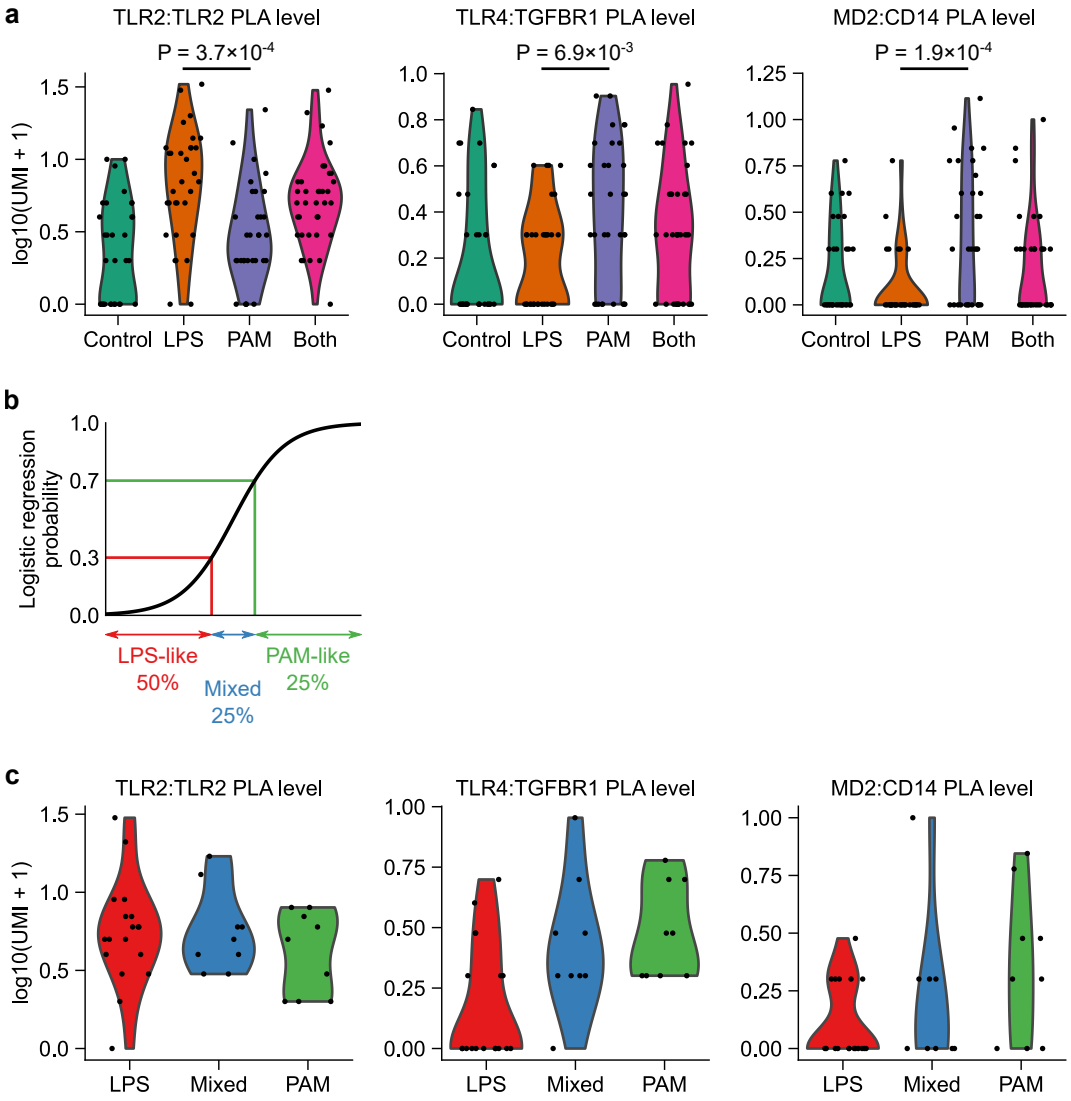

Figure 3.21: Top three PLA product markers in the logistic regression classifier. (a) Violin plots showing the log-transformed count of the top three PLA products. P-values are calculated using two-sided Welch's t-test. (b) Schematic showing how the logistic regression classifier is used to predict response in cells treated with both LPS and PAM. A single cell is classified as LPS-like (PAM-like) if the regression probability of LPS (PAM) is at least 70%. Otherwise, a cell is classified as mixed response. (c) Violin plots showing the log-transformed count of the top three PLA products in the predicted response.

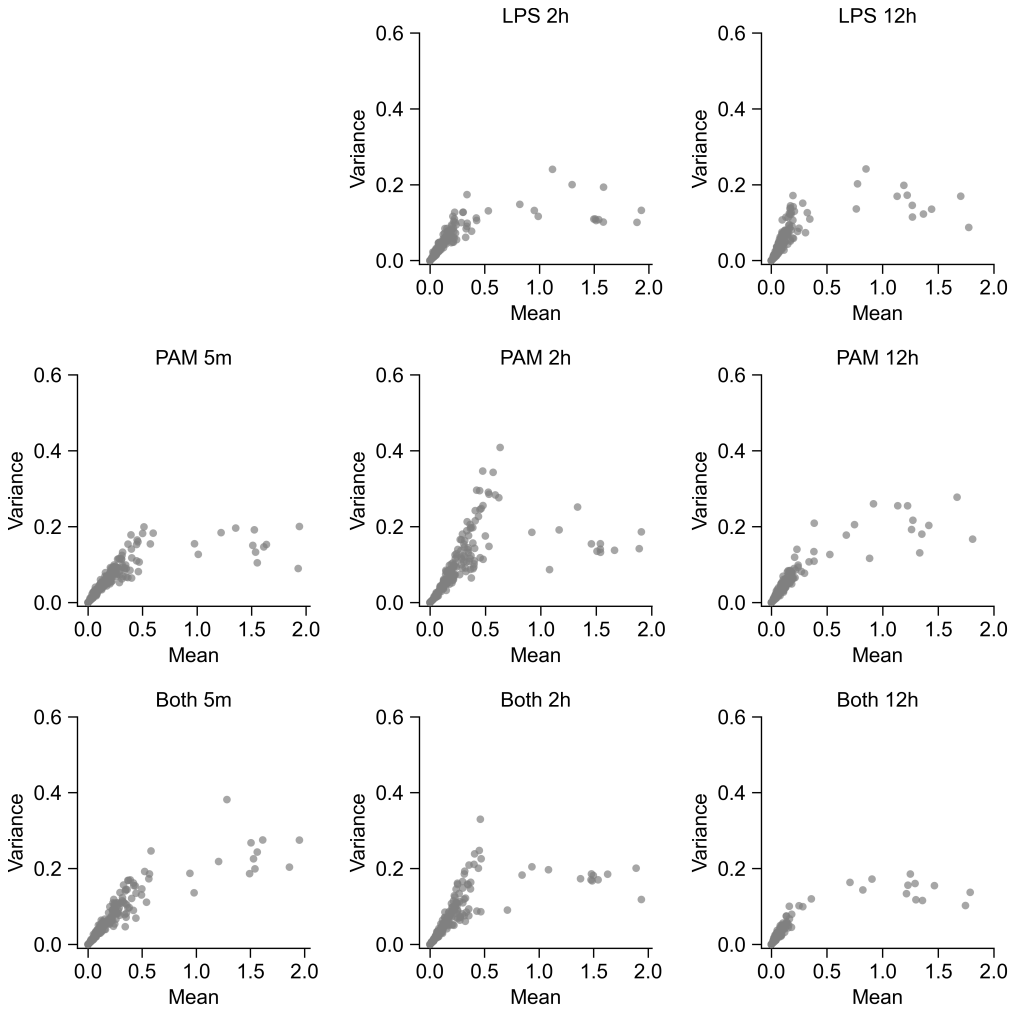

Figure 3.22: The mean-variance relationship in PLA products across treatments conditions in macrophages. The counts of PLA products are log-transformed before calculating the mean and variance. Each dot represents a PLA product. This shows that the reduction in variance across highly abundant PLA products is observed across all stimulants and time points.

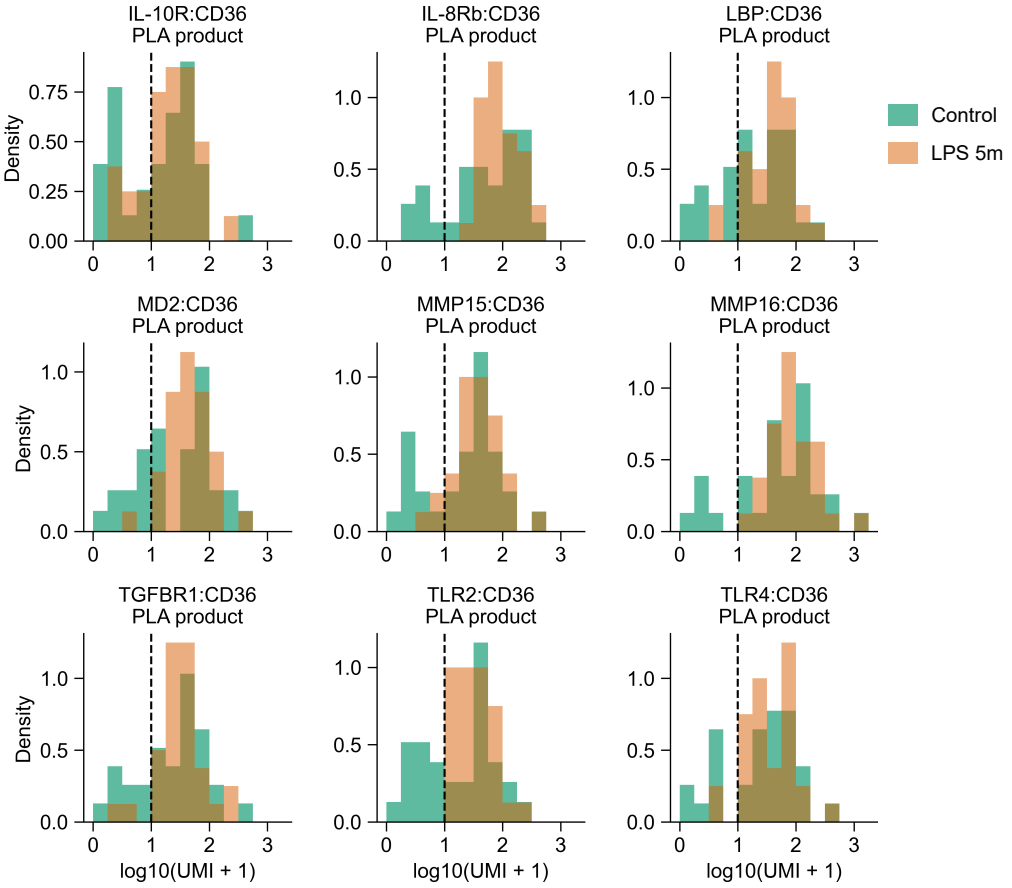

Figure 3.23: Distribution of CD36-related PLA products in macrophages. Histograms showing the distribution of the log-transformed counts of 9 CD36-related PLA products in control and LPS 5-minute treatment groups. These 9 PLA products are those that have log-transformed means greater than or equal to 1.

3.6 Methods

Cell Culture

The PD1+ Jurkat and and PDL1+ Raji cell lines were a generous gift from Jun Huang [155]. Both were maintained at 37°C with 5% CO2 in RPMI (Gibco) supplemented with 10% Fetal Bovine Serum (FBS, Hyclone).

Frozen PBMSs and macrophages were purchased from STEMCELL Technologies. They were quickly thawed at 37°C and washed three times by suspension in 10 mL RPMI + 10% FBS and centrifugation at 300 × g for 3 min. Cell were then allowed to rest overnight at 37°C with 5% CO2 in RPMI + 10% FBS.

Prox-seq Probe Preparation

Antibodies were DNA-conjugated using previously published methods [156]. Briefly, antibodies were concentrated and buffer exchanged into phosphate buffered saline (PBS) prior to conjugation using a 50K molecular weight cutoff (MWCO) concentrator (EMD Millipore). The antibodies were then reacted with dibenzocyclooctyne-PEG4-*N*-hydroxysuccinimidyl ester (DBCO, Sigma, catalog #764019) in dimethylsulfoxide (DMSO, Sigma Aldrich). This was done by combining the antibody solution with a 2 mM DBCO solution at a 10:1 volume-by-volume ratio, and incubated on ice for 1–2 hours. After incubation, the DBCO-conjugated antibodies were purified using a using a 50K MWCO concentrator and the antibody-to-DBCO ratio was measured via UV-Vis (Nanodrop). 1–2 μg DBCO-conjugated antibodies (at 13–3 μM) were combined with an equal volume of 80 μM azide-functionalized PLA oligomer (IDT) dissolved in PBS, and allowed to react overnight at 4°C. The sequences of the DNA oligos are given in Supplementary tables 3.5 and 3.6. For probes that were stored long-term, the reaction mixture was then brought to 50% glycerol in PBS (Sigma Aldrich).

Flow Cytometry

Jurkat and Raji cells were plated in a 96-well plate (Corning) at 100,000 cells/well. Cells were centrifuged at 500 × g for 5 min, media was removed and replaced with 30 μL 5 nM Prox-seq probes (2.5 nM probe A and 2.5 nM probe B) in Probe Binding Buffer (PBS, 0.1% BSA, 0.1 mg/ml sonicated salmon sperm DNA (Invitrogen), 6.7 nM of each of the antibodies' isotype controls). Cells were incubated with probes for 30 min at 37°C. Cells were then washed three times by centrifuging at 500 × g for 5 min and resuspending in 100 μL 1% BSA in PBS. Cells were then resuspended in a 1 to 100 dilution of secondary antibody in 1% BSA/PBS, and incubated at room temperature for 20 min. Cells were centrifuged and washed two times as before. Finally, they were analyzed using a Fortessa 4-15 (BD Biosciences) with a High Throughput Screening module.

Jurkat and Raji Sample Processing: Droplet-Based Prox-seq

150,000 Jurkat and 150,000 Raji cells were counted and spun at 500 × g for 3 min, washed once with 1% BSA/PBS, and spun again. Cells were then combined and plated in 96-well U-bottom plates. Cells were then centrifuged at 300 g for 3 min and fixed with 4 mM 3,3'-dithiobis(sulfosuccinimidyl propionate) (DTSSP, Life Technologies) in PBS at 37°C for 30 min. Cells were washed once with 1 mL 1% BSA/PBS and 90 μL probes were added at 5 nM each (2.5 nM probe A + 2.5 nM probe B) in Probe Binding Buffer. Cells were then incubated at 37°C for 60 min, washed twice with 1 mL 1% BSA/PBS, and ligated with 300 μL Ligation Solution (50mM HEPES pH 7.5, 10 mM $MgCl_2$, 1mM rATP (NEB), 9.5 nM connector oligo (TTTCACGACACGACACGATTTAGGTC, IDT), 130 U/ml T4 Ligase (NEB)). Finally, cells were unfixed with 30 mM DTT at 37°C for 30 min, centrifuged at 300 × g for 3 min, and resuspended in 0.1% BSA/PBS. The cells were then processed with Drop-seq (see below).

Jurkat and Raji Sample Processing: Plate-Based Prox-seq)

50,000 Jurkat cells and 50,000 Raji cells were collected, centrifuged at 500 × g for 3 min, and washed once in 1% BSA/PBS. Jurkat cells were resuspended in 5 μM carboxyfluorescein diacetate succinimidyl ester (CFSE, Biolegend) in PBS for 20 min at room temperature in order to specifically label Jurkat cells for cell sorting. The cells were then resuspended in 30 μL Prox-seq probes in Probe Binding Buffer. Each protein target has 2.5 nM probe A and 2.5 nM probe B. Cells were incubated at 37°C for 60 min. They were then centrifuged and washed three times by centrifuging and resuspending in 1% BSA/PBS as before. Cells were then resuspended in 100 μL Ligase Solution and rotated for 3 hr at 37°C. With 30 min remaining in the incubation, propidium iodide (Invitrogen) was added to the solution to a final concentration of 1 μg/ml. Cell were then centrifuged at 500 × g for 3 min, resuspended in 1% BSA/PBS and 1:500 propidium iodide (PI). Single PI-negative cells were sorted into each well of two 96-well plates, one for each cell line.

PBMC Sample Processing

One million rested human PBMCs (STEMCELL Technologies) were collected and pelleted. All centrifugation steps were performed at 300 × g for 3 min. The cells were then resuspended in 1 mL 1% BSA/PBS and pelleted. The cells were then resuspended in Fc blocker solution composed of 95 μL 1% BSA/PBS and 5 μL TruStain FcX (Biolegend) and incubated at room temperature for 5 min. Following this step, all buffers were supplemented with 1:1000 RNase inhibitor (NEB). The cells were then pelleted and resuspended in 300 μL 5 nM Prox-seq probes in Probe Binding Buffer as above. Due to its size, the probe panel was divided into three parts and administered in series. For each portion, cells were allowed incubate at 37°C for 15 min, then pelleted and resuspended in the next portion. Cells were then spin/washed 3 times with 1 ml 1% BSA/PBS. They were then resuspended in Ligation Solution, as described above, and incubated at 37°C for 30 min. For plate-based experiments, cells were spin/washed in 1% BSA/PBS and resuspended in 500 μL 1% BSA/PBS + 1:500 propidium

iodide (PI, Invitrogen) and incubated at room temperature for 10 min. Finally, cells were pelleted, resuspended in 500 μL 1% BSA/PBS, and live cells were sorted into plates.

For the 10x experiment, after ligation, the cells were washed twice with 1 ml 1% BSA/PBS and centrifuged at 300 × g for 5 min. Dead cells were removed using a Dead Cell Removal Kit (Miltenyi Biotec, catalog #130-090-101) following the manufacturer's protocol. Cells were then counted and diluted to 1,000 cells/μL. This sample was then processed using a modified 10x protocol (see below for more details).

Primary Macrophage Sample Processing

Thawed human peripheral blood macrophages (STEMCELL Technologies) were distributed into 10 wells of a non-tissue culture treated 24 well plate. They were then allowed to rest overnight in RPMI + 10% FBS. While in the plate, cells were centrifuged for 300 g for 3 min and stimulated for 12 hours, 2 hours, or 5 min with 100 ng/mL LPS, 40 ng/ml Pam2CSK4 (PAM), or both in RPMI + 10% FBS. All future centrifugation steps occurred at 300 × g for 3 min. After stimulation, cells were dissociated with TryplE (gibco), pelleted, and resuspended in 4% paraformaldehyde (PFA, Santa Cruz Biotechnology) for 15 min at room temperature. Cells were then spin/washed one time with 1 mL 1% BSA/PBS and resuspended in 95 μL 1% BSA/PBS + 5 μL TruStain FcX. After 5 min at room temperature, cells were pelleted and each sample was resuspended in 54 μL 5 nM prox-seq probes in Probe Binding Buffer (as above). The cells were incubated for 30 min at 37°C, then pelleted and washed with 1 mL 1% BSA/PBS three times. Next, the cells were resuspended in Ligation Solution as above and incubated for 30 min at 37°C. Finally, the cells were pelleted, resuspended in 500 μL 1% BSA/PBS + 1:500 PI (Invitrogen), and live cells were sorted into plates.

Droplet-Based Prox-seq

After the PLA process, cells were resuspended in 0.01% BSA at 100,000 cells/ml, injected into a 3 ml syringe and kept on ice until ready to use. Drop-seq beads (ChemGenes, catalog

Macosko-2011-10(V+)) were washed once with TE-TW (10 mM Tris pH 8.0, 1 mM EDTA and 0.01% Tween-20), resuspended in Drop-seq lysis buffer at 120,000 beads/ml, injected into a 3 ml syringe along with a magnetic disc (V&P Scientific, catalog #772DP-N42-5-2) and kept on ice until ready to use. Droplet oil (Bio-Rad, catalog #1864006) was injected into a 10 ml syringe and kept at room temperature until ready to use.

The Drop-seq chip was made from PDMS using the design from Macosko et al. [54]. The chip was mounted on an inverted microscope to monitor droplet formation. Then, the cell, bead and oil syringes were connected to the appropriate chip inlets via plastic tubes (Scientific Commodities Inc, catalog #BB31695PE/2). A magnetic stirrer (V&P Scientific, catalog #710D2) was used to keep the beads suspended during the experiment. To generate droplets, syringe pumps were used to inject cells at 3 ml/hour flow rate, beads at 3 ml/hour, and oil at 12 ml/hour. The generated droplets were collected in a 50 ml tube via a plastic tube that was inserted into the chip's outlet.

After droplet generation, the droplets were broken as follows. Oil at the bottom of the collection tube was removed as much as possible, without disturbing the droplets. Then, 30 ml of 6X Saline-Sodium Citrate solution (SSC, Fisher BioReagents) was added to the droplets, followed by 1ml of 1H,1H,2H,2H-Perfluorooctan-1-ol (Synquest Laboratories). Next, the tube was shaken strongly by hand 4 times, and centrifuged at 1000 $\times$ g for 1 minute. After this, the supernatant was transferred to another 50 ml tube (tube 1), and 30 ml of 6X SSC was added to the collection tube, and then the supernatant was transferred to another 50 ml tube (tube 2). Tubes 1 and 2 were centrifuged at 1000 $\times$ g for 3 min to pellet the beads, then the supernatant in both tubes was removed, and the beads in both tubes were transferred to a 1.5 ml microfuge tube. The beads were washed twice with 1 ml of 6X SSC, and once with 300 μL of 5X Maxima H minus reverse transcription (RT) buffer (Thermo Scientific).

Next, RT was performed by adding 200 μL of RT mixture was added to the beads. The RT mixture was prepared as follows: 1$\times$ RT buffer, 4% Ficoll PM-400 (Sigma-Aldrich), 1 mM each dNTP (NEB), 1,000 units/ml RNase inhibitor, murine (NEB), 2.5 μM TSO_RNAhybrid

(IDT), and 10 units/μL Maxima H minus reverse transcriptase (Thermo Scientific). The beads were incubated at room temperature for 30 min with rotation, followed by incubation at 42°C for 90 min with rotation.

After RT, the beads were centrifuged at 1000 × g for 1 minute, the supernatant was discarded, and the beads were washed once with 1 ml of TE-SDS (10 mM Tris pH 8.0, 1 mM EDTA and 0.5% SDS), twice with 1 ml of TE-TW, and once with 1 ml of Tris pH 8.0. Then, 200 μL of exonuclease I mixture (1,000 units/ml exonuclease I and 1× exonuclease I buffer, NEB, catalog #M0293) was added to the beads, and the beads were incubated at 37°C for 45 min with rotation.

After exonuclease I treatment, the beads were centrifuged at 1000 × g for 1 minute, washed once with 1 ml of TE-SDS, twice with 1 ml of TE-TW, and once with 1 ml of nuclease-free water. Next, the cDNA and PLA products on the beads were amplified using PCR. The beads were distributed into a 96-well plate at approximately 5,000 beads per well. Then, a 50 μL PCR reaction was set up in each well, which contained 1× KAPA HiFi HotStart Readymix (Roche), 0.8 μM TSO_PCR primer, and 0.8 μM U_fwd primer. The thermal cycle program was: 95°C 3 min; 4 cycles of 98°C 20 s, 65°C 45 s, 72°C 3 min; 10 cycles of 98°C 20 s, 67°C 20 s, 72°C 3 min; 72°C 5 min; 4°C hold.

After PCR, 10 μL of nuclease-free water was added to each well, the plate was centrifuged at 1000 × g for 1 minute, and 50 μL of supernatant from each well was transferred to another 96-well plate (plate 1). Then, 30 μL of AMPure XP beads (Beckman Coulter) was added to each well (i.e., 0.6× AMPure bead concentration), the new plate was incubated at room temperature for 5 min, and left on a magnetic rack for 5 min. Next, the supernatant, which contained PLA products, was transferred into another 96-well plate (plate 2). Each well in plate 1 was washed four times with 200 μL of 80% ethanol, and left to air dry for 2 min. 11.5 μL of nuclease-free water was added to each well to elute the cDNA products from the AMPure beads, the plate was incubated at room temperature for 5 min, left on a magnetic rack for 3 min, and 10 μL of the supernatant from each well was collected and pooled into

a 1.5 ml microfuge tube. To recover the PLA products from the supernatant stored in plate 2, 60 μL of AMPure beads was added to each well (i.e., 1.8× AMPure bead concentration), and the same protocol for plate 1 was followed afterwards.

cDNA products were quantified using TapeStation High Sensitivity dsDNA kit (Agilent). They have a broad length distribution between 400 to more than 3000 base pair, with a peak at approximately 1000 bp. Library preparation of cDNA products was performed using Nextera XT DNA Library Preparation Kit (Illumina, catalog #FC-131-1024). 450 pg of cDNA products were combined with 10 μL of TD buffer and 5 μL of enzyme mix ATM. Next, the mixture was incubated at 55°C for 5 min, after which 5 μL of NT buffer was added, and the mixture was incubated at room temperature for 5 min. After this, the following components were added to the mixture to set up a PCR reaction: 8 μL of nuclease-free water, 1 μL of 10 μM P5_TSO primer, 1 μL of 10 μM P7_N70X_Custom2 primer, and 15 μL of Nextera PCR Master Mix. The thermal cycle program was: 95°C 30 s; 12 cycles of 95°C 10 s, 55°C 30 s, 72°C 30 s; 72°C 5 min; 4°C hold. Finally, the PCR products were cleaned up using 0.6× AMPure beads as above.

Library preparation of PLA products was performed as following. 1 μL of the pooled PLA products was used to set up a 20 μL PCR reaction (1× KAPA HiFi HotStart Readymix, 0.2 μM P5_TSO primer and 0.2 μM P7_N7XX_Custom2 primer). The thermal cycling program was: 95°C 3 min; 12 cycles of 98°C 20 s, 67°C 15 s, 72°C 20 s; 72°C 5 min; 4°C hold. Then, the PCR products were cleaned up with 1.8× AMPure beads as above.

Both cDNA and PLA product libraries were sequenced together on a NextSeq 550 machine. The cDNA and PLA libraries each received 20% of the total reads. PhiX control was spiked in at 40% concentration according to Illumina's instruction, because of the low diversity of the PLA product libraries. Custom read 1 sequencing primer (Read1CustomSeqB) was spiked in with Illumina's read 1 primer, custom read 2 primer (DropPLA_Read2) and custom i7 index read primer (DropPLA_i7Read) were used according to Illumina's instruction. Read distribution is 20 bases for read 1, 85 bases for read 2, and 8 bases for i7 index

read.

The sequences of the primers are given in Supplementary table 3.7.

Plate-Based Prox-seq

For non-PFA-fixed cells, 96-well plates were first prepared by adding 4 μL of lysis buffer per well (0.1% Triton X-100, 1 units/μL RNase inhibitor, murine (NEB), 2.5 mM dNTPs (NEB), 2.5 uM SmartSeq2_oligodTVN, 2.5 uM SmartSeq2_oligodTGT in water), as per Smart-seq2 protocol [46]. After single-cell sorting, the plates were centrifuged at 700 × g for at least 10 s, and kept at −80°C for storage. When ready to use, the plates were thawed on ice, incubated at 72C for 3 min, and centrifuged at 700 × g for at least 10 s.

For PFA-fixed cells, the 96-well plates were prepared by adding 6 μL of modified lysis buffer per well (0.1% Triton X-100, 1000 units/mL RNase inhibitor, murine (NEB), 20 units/mL proteinase K (NEB), 2.5 mM dNTPs (NEB), 2.5 uM SmartSeq2_oligodTVN, 2.5 uM SmartSeq2_oligodTGT in TE buffer). After cell sorting, the plates were briefly centrifuged, and incubated at 56°C for 1 hour, 95°C for 10 min, 4°C for at least 5 min, and stored at −80°C.

Before library preparation, the plates were thawed on ice and briefly centrifuged. Then, 2–4 μL of the cell lysate was used as input for pre-amplification (the lysate volume depends on how much lysate is needed to prepare the mRNA library, which can be skipped if one is only interested in the PLA products of Prox-seq). Each pre-amplification PCR reaction contained 1× KAPA HiFi HotStart Readymix, 0.1 μM SmartSeq2_oligodTGT primer and 0.1 μM U_fwd primer, for a total volume of 25 μL. The thermal cycle program was: 98°C 3 min; 5 cycles of 98°C 20 s, 55°C 15 s, 72°C 1 min; 17 cycles of 98°C 20 s, 67°C 15 s, 72°C 1 min; 72°C 5 min; 4°C hold. After PCR, the products were cleaned up using 1.8× AMPure beads.

Then, library preparation was performed to attach Illumina sequencing adapters, and to barcode the PLA products from each single cell with dual indices. To do this, 2 μL

of PLA products from each well was used as input to a 25 µL PCR reaction, which contained 1× KAPA HiFi HotStart Readymix, 1 µM SmartPLA_P5_S5XX primer or SmartPLA_P5_HXXX primer, and 1 µM SmartPLA_P7_N7XX primer. The thermal cycle program was: 98°C 3 min; 12 cycles of 98°C 20 s, 67°C 15 s, 72°C 1 min; 72°C 5 min; 4°C hold.

The library preparation procedure was performed either manually or automatically with an automatic liquid handling system (PerkinElmer Janus G3 and Tecan Freedom Evo 200).

After library preparation, the barcoded PLA products were cleaned up using 1.8× AMPure beads as above, and quantified with Qubit 1× dsDNA HS Assay Kit (Invitrogen, catalog #Q33231) or Fragment Analyzer DNA High Sensitivity kit. Then, the library from each well was normalized and pooled at equimolar, and up to four 96-well plates were sequenced together with a mid-output NextSeq kit v2.5. PhiX control was spiked in at 40% concentration. Custom read 1 sequencing primer (SmartPLA_Read1) was spiked in with Illumina's read 1 primer, custom i5 index read primer (SmartPLA_i5Read) and custom i7 index read primer (SmartPLA_i7Read) were used according to Illumina's instruction. Read distribution was 75 bases for read 1, 8 bases for i7 index read, and 8 bases for i5 index read.

The sequences of the primers are given in Supplementary table 3.8–3.10.

10x-Based Prox-seq

PBMCs were loaded onto a single cell Chromium Next GEM chip at a concentration of 1000 cells/µL. Using the 10x Genomics single-cell 3' V3.1 protocol (10x Genomics, CG000315 RevB), GEM was generated. For cDNA amplification, along with the standard cDNA primers, 1 µL of 2 µM U_fwd primer was added to the reaction. Following amplification, PLA products were separated from cDNA using solid phase reversible immobilization (SPRI)-based size selection step by incubating the cDNA amplification product in 0.6× SPRIselect (Beckman Coulter, catalog #B23317) for 5 min at room temperature. The cDNA fragments remain on the SPRIselect beads while the PLA products are contained in the supernatant, which is stored for PLA library construction (10x Genomics, CG000317 RevB). The SPRI-

select beads were washed with 80% ethanol and cDNA was eluted into 40 μL elution buffer. Gene expression libraries were constructed from the purified cDNA (10 μL) using 10x's enzymatic fragmentation, adaptor ligation and sample index attachment and then eluted in 35 μL elution buffer according to the standard 10x Genomics single-cell 3' V3.1 protocol.

For PLA library construction, PLA products were mixed with 2.1× SPRIselect to undergo the cleanup process and eluted into 40 μL elution buffer. Afterward, 10 μL eluted solution was used to set up a 50 μL PCR reaction (1× KAPA HiFi HotStart Readymix, 0.2 μM 10X_SI_PCR primer and 0.2 μM P7_N7XX_Custom2 primer). The thermal cycling program was: 98°C 2 min; 12 cycles of 98°C 20 s, 60°C 30 s, 72°C 20 s; 72°C 5 min; 4°C hold. The PCR products were cleaned up using 1.2× SPRIselect and then eluted into 38 μL elution buffer.

The samples (cDNA and PLA product) were quantified with Qubit 1× dsDNA HS Assay Kit (Invitrogen, catalog #Q33231) and High Sensitivity DNA TapeStation (Agilent) to determine the library size. The average library sizes were ∼570 bp for cDNA and ∼266 bp for PLA product. The libraries were sequenced on the NextSeq 550 platform. For cDNA library, a high output 150-cycle kit was used (read 1, 28 cycles; i7 index, 10 cycles; i5 index, 10 cycles; read 2, 90 cycles) with 1% PhiX. For PLA product library, a mid output 150-cycle kit was used (read 1, 28 cycles; i7 index, 8 cycles; read 2, 75 cycles) with 40% PhiX. For PLA product library only, we spiked in custom read 2 primer (DropPLA_Read2) to Illumina read 2 primer and custom i7 read primer (DropPLA_i7Read).

The sequences of the primers are given in Supplementary table 3.7.

Alignment of mRNA Sequencing Data

Drop-seq mRNA sequencing data was aligned using Drop-seq tools v2.3.0. First, the first 19 bases of read 2 were trimmed using the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) in order to remove the mosaic sequence from the Nextera transposase. Then, the read 1 and processed read 2 fastq files were used as input for Drop-seq tools v2.3.0 following

the online instruction (https://github.com/broadinstitute/Drop-seq/blob/v2.3.0/doc/Drop-seq_Alignment_Cookbook.pdf). GRCh38 was used as the reference human genome. To extract the digital gene expression matrix, we chose a list of cell barcodes according to the knee plot. Note that this list of cell barcodes will later be used for alignment of PLA product libraries.

10x mRNA sequencing data was aligned using Cell Ranger v6.1.1 and the human reference genome GRCh38 version 2020-A from the 10x Genomics (https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/build).

Alignment of PLA Product Sequencing Data

To convert raw sequencing reads from PLA product libraries to PLA product count matrices, we developed a custom Java program (https://github.com/tay-lab/Prox-seq) that has three modes of alignment, depending on whether the samples were processed with Drop-seq, 10x Chromium or plate-based pipeline.

First, the program performs alignment by extracting the cell barcode and the UMI from the sequencing reads. For Drop-seq and 10x modes, the program will extract the cell barcode and UMI from read 1, and the sequence of the PLA product from read 2. For plate mode, the cell barcode is demultiplexed from the dual index by Illumina's software, and the UMI is extracted from the Prox-seq probe A. The program filters out reads that have at least 3 N bases in the UMI, or at least 6 repetitions (for droplet mode) or 9 repetitions (for plate mode) of the same bases in the UMI. For all three modes, the program discards reads that have more than 1 bases with quality score below 10.

Next, the program looks for the connector region's sequence in the reads within a ± 1 base window around the expected location of the connector region. A read is considered a valid PLA read if the found connector region is within 2 Levenshtein distance of the expected sequence (TCGTGTCGTGTCGTGTCTAAAG). Invalid reads are discarded. Next, we look for the barcodes of antibodies A and B, and match them up against a text file containing

the list of reference antibody barcodes and their protein barcodes. An antibody barcode is considered a match if it exactly matches a reference barcode, or if it matches to exactly one reference barcode at 1 Hamming distance. Any reads with at least one antibody barcode that do not match are discarded.

After alignment, we perform cell barcode correction for droplet mode only. This step is done to ensure that the barcodes obtained from PLA alignment match with the barcodes obtained from the mRNA alignment. First, we import the list produced by Drop-seq tools v2.3.0 function BamTagHistogram, which contains the reference cell barcodes and their read counts, and kept the cell barcodes with at least 100 read counts. Next, we correct a found cell barcode to one of the reference cell barcodes, if the found cell barcode differs by 1 modified Hamming distance from exactly one reference cell barcode. The modified Hamming distance allows for substitution like the normal Hamming distance, but an N base is not considered in the calculation. Any reads with more than 1 N bases in the cell barcodes, or differ by 1 modified Hamming distance from more than one reference cell barcodes are discarded.

Next, we perform UMI merging for each single cell separately. For each unique combination of cell barcode-PLA product, we collect a list of unique UMIs. Then, we iterate through the list in increasing order of read count, and remove UMIs that are 1 Hamming distance away to at least one other UMI that received higher read counts. This merging step is the same for both droplet and plate modes. Lastly, we export the UMI merged reads into a count matrix as a text file. For Drop-seq mode, we choose a set of single cell barcodes to be exported, either by making use of the knee plot for PLA product read counts, or by using the same set of cell barcodes chosen for Drop-seq tools v2.3.0 function DigitalCount. For 10x mode, we use all cell barcodes in Cell Ranger's filtered count matrix. For plate mode, all single cell barcodes were exported.

Data Analysis for Jurkat and Raji Experiments

For the Drop-seq experiment, the mRNA and PLA product count data was processed using Seurat v4.0.4 [157]. We then removed cells with fewer than 200 detected genes, removed genes detected in fewer than 5 cells, and removed cells with more than 20% mitochondrial genes. Next, we only kept the cell barcodes that are present in both the mRNA count data and the PLA product count data.

The protein abundance was calculated from PLA data by calculating the number of UMI counts associated with each protein target, across both Prox-seq probes A and B. For example, if PLA products CD3:CD28 and CD3:CD3 have a UMI count of 10 and 20, respectively, then the abundance of CD3 was $10 \times 1 + 20 \times 2 = 50$.

The data was then normalized (log-normalization for mRNA data, and centered log-ratio (CLR) transformation for protein abundance and PLA product count), scaled, and clustered. We used resolution=0.1 for mRNA clustering and protein clustering, and resolution=0.3 for PLA product clustering). Finally, the data was visualized with principal component analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE). For clustering and t-SNE, we used the first 10 PCs.

For the plate-based experiment, the PLA product count data was processed with Python 3 and RStudio. We removed Jurkat and Raji cells with fewer than 100 detected UMIs. A Python script was used to estimate the complex abundance using the UMI data. More details on this is below.

Data Analysis for PBMC Experiments

For the plate-based experiment, the PLA product count data was analyzed with Python 3. We removed cells with fewer than 10 detected UMIs. CD4 T cells were defined as those that have more than 100 UMIs of CD3 protein, more than 100 UMIs of CD4 protein, and fewer than 100 UMIs of CD8 protein. CD8 T cells were defined as those who have more than 100 UMIs of CD3 protein, more than 100 UMIs of CD8 protein, and fewer than 100 UMIs of

CD4 protein. CD9:CD9 PLA product-high cells were defined as those that have a higher UMI count of CD9:CD9 PLA product than the total UMI count of all CD9 and non-CD9 PLA products.

For the 10x experiment, the mRNA and PLA product count data was analyzed using Seurat v4.0.4 and Python 3. First, we removed cells with more than 10,000 detected UMIs, removed cells with fewer than 500 detected genes, removed genes detected in fewer than 5 cells, and removed cells with more than 15% mitochondria genes. Next, we normalized the data (log-normalization for mRNA data, and center log ratio (CLR) transformation for protein abundance and PLA product count), scaled it, and visualized the data with t-SNE (using the first 15 PCs of mRNA data, or the first 20 PCs of PLA product data). To annotate the cell types from mRNA data, we used SingleR package (v1.6.1) and the Novershtern reference annotations from the celldex package (v1.2.0) [158, 159]. We also defined the CD4 T cells, CD8 T cells, and the CD9:CD9 PLA product-high cells similarly to the plate-based experiment.

Data Analysis for Primary Macrophage Experiment

The PLA product count data was analyzed using Python 3. We removed cells with fewer than or equal to 10 detected UMIs, and removed cells with at least 10,000 detected UMIs. To calculate the fold change in PLA products across different time points for each ligand, we divided the mean UMI count of a PLA product at each time point by its mean UMI count in the control group. To prevent lowly expressed PLA products from displaying unreasonably high fold changes, we only considered PLA products detected in at least 10% of the cells in the control group and the treatment group, and we add a pseudo count of 1 to both the numerator and denominator when calculating the fold change. To calculate the fold change in protein abundance across different time points for each ligand, we divided the mean UMI count of a protein by its mean UMI count in the control group. We also only considered proteins detected in at least 10% of the cells in the control group and the treatment group.

To predict ligand treatment from PLA product count, we trained a logistic regression classifier on the combined PLA product data from LPS treatment group and PAM treatment group. First, we log-transformed the data with one added pseudocount. Next, we used 75% of the data for training, standardized the data using the mean and standard deviation of the training data, and trained a logistic regression classifier (scikit-learn package v0.24.1) with L1 regularization and liblinear solver. The classifier performance was evaluated on the remaining 25% testing data. To predict response on the LPS and PAM treatment group, a 70% probability threshold is used: a cell is predicted to be a LPS-like responder only if the classifier predicts it to be LPS with at least 70% probability, and a PAM-like responder only if the classifier predicts it to be PAM with at least 70% probability. If a cell's probability does not meet either of these criteria, it is considered to be a mixed responder.

To quantify protein complexes, we separated the cells by ligand and time point, and used a convergence threshold of 5 (see below for more details).

Protein Complex Estimation Algorithm

The most unique feature of our assay is that it enables combinatorial screening of in situ protein complexes and protein interactions, at the single-cell level. Therefore, we developed a method for prediction and quantification of protein complexes from the count data of PLA products.

First, let us consider the case where no protein complexes are present on a cell. Then, two protein molecules can be close to each other by chance, allowing the Prox-seq probes to be ligated and generate PLA products (Supplementary Figure 3.24a). We call this event random ligation. The number of times random ligation occurs increases with the abundance of the proteins. Hence, the count of a PLA product $i{:}j$ is then proportional to the abundance of protein i and protein j. More specifically, the count of PLA product $i{:}j$ is proportional to the abundance of Prox-seq probe A targeting protein i, and Prox-seq probe B targeting protein j that are present on the cell.

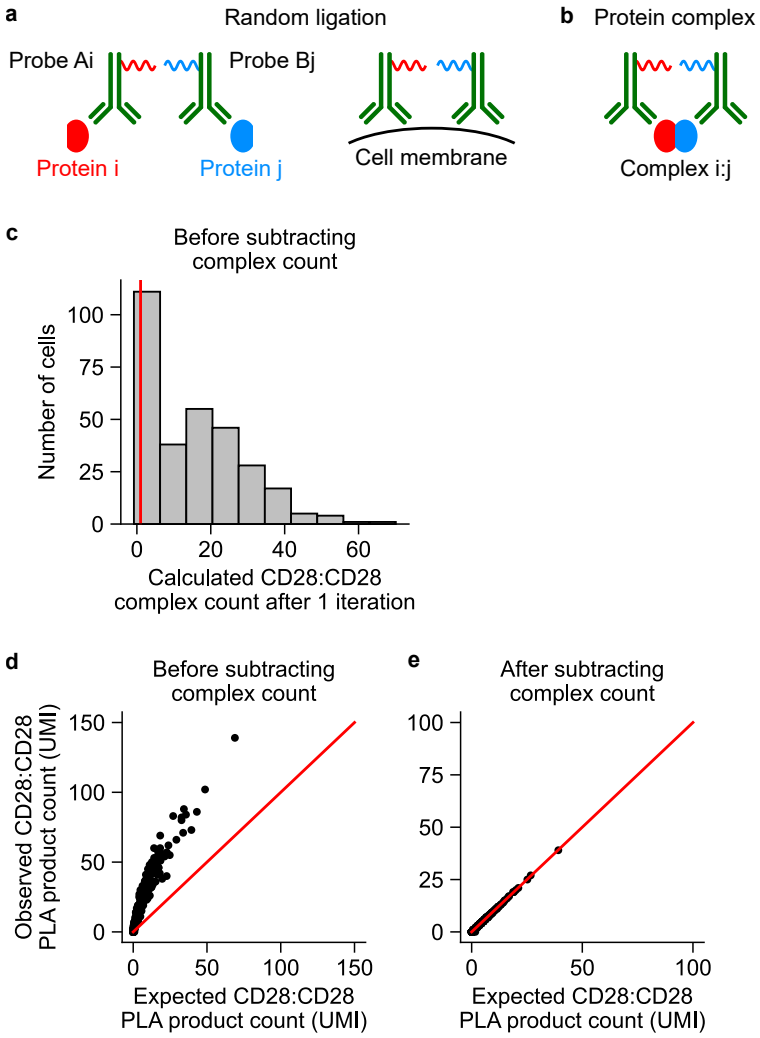

Figure 3.24: Quantification of protein complex from PLA products count data. (a, b) Count of PLA products are the sum of two components, (a) the random ligation component (background), and (b) the stable protein complex (signal). (c) Histogram of the calculated count of CD28:CD28 protein complex, after 1 iteration. The vertical red line indicates x = 1. (d, e) Scatter plot showing the expected random count and observed count of CD28:CD28 PLA product (d) before and (e) after subtracting the calculated complex count. The red lines indicate y = x.

If the cell has no protein complex among targeted proteins, then the binding of a probe A to protein i is independent of the binding of a probe B to protein j. Using the formula for calculating the probability of two independent events, the probability of a PLA product being i:j, $P(A = i, B = j)$, is equal to the product of the probability of probe A being i,

$P(A = i)$, and the probability of probe B being j, $P(B = j)$:

$$P(A = i, B = j) = P(A = i) \times P(B = j) \tag{3.1}$$

While we do not directly measure the binding of individual Prox-seq probes A and B, we can calculate them from the observed count of PLA products under the assumption of no protein complex:

$$\frac{X_{i,j}}{\sum_{k=1}^{n}\sum_{l=1}^{n}} = \frac{\sum_{l=1}^{n} X_{i,l}}{\sum_{k=1}^{n}\sum_{l=1}^{n} X_{k,l}} \times \frac{\sum_{k=1}^{n} X_{k,j}}{\sum_{k=1}^{n}\sum_{l=1}^{n} X_{k,l}} \tag{3.2}$$

where $X_{i,j}$ is the observed count of PLA product i:j, and n is the number of protein targets. Rearranging equation 3.2 provides the expected random count of a PLA product i:j, $E_{i,j}$, created by random ligation when there are no protein complexes on the cell:

$$E_{i,j} = \frac{\sum_{l=1}^{n} X_{i,l} \times \sum_{k=1}^{n} X_{k,j}}{\sum_{k=1}^{n}\sum_{l=1}^{n} X_{k,l}} \tag{3.3}$$

Therefore, if the observed count of PLA product i:j, $X_{i,j}$, is higher than its expected random count under the assumption of no protein complexes ($E_{i,j}$), then protein i interacts with protein j.

To quantify the abundance of protein complex i:j, we reason that the observed count of the PLA product i:j is the sum of two components, one from the protein complex abundance (signal), and one from the PLA products produced from random ligation (background) (Supplementary Figure 3.24b). If the signal component is subtracted from the observed count, we are left with the background component, which satisfies equation (3.3). Therefore, we have the following equation:

$$X_{i,j} - C_{i,j} = \frac{(\sum_{l=1}^{n} X_{i,l} - \sum_{l=1}^{n} C_{i,l}) \times (\sum_{k=1}^{n} X_{k,j} - \sum_{k=1}^{n} C_{k,j})}{\sum_{k=1}^{n}\sum_{l=1}^{n}(X_{k,l} - C_{k,l})} \tag{3.4}$$

where $C_{i,j}$ denotes the count of protein complex i:j. If $C_{i,j} = 0$, then protein i does not interact with protein j. Equation (3.4) is only valid for the PLA product count data of a single cell.

To quantify the protein complexes that are present on the cell, we have to solve a system of n^2 quadratic equations. We found that the system of equations could be solved iteratively:

$$C_{i,j}^{(m+1)} = X_{i,j} - \frac{(\sum_{l=1}^{n} X_{i,l} - \sum_{l=1}^{n} C_{i,l}^{(m)}) \times (\sum_{k=1}^{n} X_{k,j} - \sum_{k=1}^{n} C_{k,j}^{(m)})}{\sum_{k=1}^{n} \sum_{l=1}^{n} (X_{k,l} - C_{k,l}^{(m)})} \tag{3.5}$$

where $C_{i,j}^{(m)}$ is the count of protein complex i:j at the m^{th} iteration, and the initial values are $C_{i,j}^{(0)} = 0$ for all i,j. The iterative process is stopped either when the solutions converge, or when the maximum number of iterations is reached, whichever comes first. Convergence is defined as when the maximum absolute change in of all $C_{i,j}$ values from the previous iteration is below a threshold. In our study, we chose the threshold to be either 1 or 5, and maximum 200 iterations.

The count of a protein complex has to be non-negative. Therefore, during each iteration, we perform a one-tailed one sample t-test on $C_{i,j}$ for a particular complex i:j obtained from all single cells. The alternative hypothesis for the t-test is that the mean of $C_{i,j}$ is above the cutoff value. In our study, we choose the cutoff value to be 1, in order to reduce false positive protein complexes (Supplementary Figure 3.24c). Then, we perform Benjamini-Hochber multiple comparison correction for over all single-cell values of each i:j protein complex, and if the adjusted P-value is higher than 0.05, we conclude that the protein complex i:j does not exist for all single cells, and set $C_{i,j} = 0$ for that iteration. If the adjusted P-value is below 0.05, we round the calculated $C_{i,j}$ values, and set any negative single-cell $C_{i,j}$ values to 0.

We also enforce symmetry for the calculated protein complex count. At each iteration, if $C_{i,j}$ passes the t-test, but $C_{j,i}$ does not, the values of $C_{j,i}$ is set to be equal to $C_{i,j}$ multiplied

by a parameter called sym_weight:

$$C_{j,i}^{(m+1)} = \text{sym\_weight} \times C_{i,j}^{(m+1)} \tag{3.6}$$

In our study, we arbitrarily chose sym_weight = 0.25.

After solving the system of equations (3.5), we substituted the calculated protein complex counts to equation (3.4), and confirmed that the two sides are equal (Supplementary Figure 3.24d, e). Thus, after subtracting the complex count component, we are left with the background component generated by random ligation, and this background component can be calculated using equation (3.2).

Below is a summary of the algorithm at each iteration:

1. Subtract complex count (obtained from the previous iteration) from the observed PLA product count.

2. Calculate expected random count of PLA product i:j using the PLA product count from step 1, i.e., the right hand side of equation (3.5).

3. Find the difference between observed count (original, without any subtraction) and expected random count from step 2.

4. Perform a one-sided t-test on the difference. If the test is significant, store the difference as the count of complex i:j for the current iteration.

5. Check if the solutions have converged, or if the maximum number of iterations has been reached.

Data Availability

The sequencing data and count data are deposited in NCBI's Gene Expression Omnibus (accession number GSE149574). The codes for PLA product sequencing alignment and PLA product analysis are available at `https://github.com/tay-lab/Prox-seq`.

3.7 Supplementary Tables

Table 3.1: List of antibodies used in Jurkat and Raji experiments.

Target	Manufacturer	Catalog #
CD3	Biolegend	300437
CD28	Biolegend	302933
CD45RA	Biolegend	304102
CD4	Biolegend	317402
PD1	Biolegend	367402
LFA1	Biolegend	363402
ICAM1	Biolegend	353102
B7	Biolegend	305202

Target	Manufacturer	Catalog #
PDL1	Biolegend	329702
HLADR	Biolegend	307602
CD147	R&D Systems	AF972
Isotype control	EMD Millipore	PP54
Isotype control	R&D Systems	AB-108-C
Isotype control	Biolegend	401401
Isotype control	Biolegend	401501
Isotype control	Biolegend	402211

Table 3.2: List of antibodies used in PBMC experiments. These were all purchased from Biolegend.

Target	Catalog #
CD3	344802
PD1	367402
CD45RA	304102
CD28	302913
CD132	338602
CD122	339002
CD4	344602
CD45RO	304202
CCR7	353202
CD27	302802
CD127	351302
CD8	301002
CD38	356601

Target	Catalog #
CD25	302602
CCR6	353401
CXCR3	353702
CD161	339902
CD56	362502
γ/δ	331202
CXCR4	306502
CD62L	304802
ICOS	313502
OX40	350002
CD46	352404
CTLA4	369602
OX40L	326302

Target	Catalog #
TIM3	345019
TIGIT	372719
Slamf6	317202
LAG3	369302
41BB	309802
CD14	301802
CD16	360702
CD9	312102
CCR5	359102
CD19	302202
B7	305202
HLADR	307602

Table 3.3: List of protein complexes detectable with the PBMC Prox-seq probe panel. The protein complexes in the table are either identified by Prox-seq, present in the IntAct database [137], or found in the literature. The numbers in the "IntAct database" column indicate the number of evidences for protein interaction in the database. The types of literature evidences for protein interaction include colocalization, association, or physical association data. Complexes identified in the IntAct database were excluded if no cell clusters expressed both protein subunits.

Protein complex	Detected by Prox-seq	IntAct database	Literature
CD127:CD4	Yes	No	No
CD27:CD27	Yes	No	Borst et al. [160]
CD28:CD28	Yes	No	Esensten et al. [126]
CD38:CD38	Yes	No	Hara-Yokoyama et al. [161]
CD3:CD3	Yes	No	Merwe and Dushek [125]
CD3:CD4	Yes	No	Collins et al. [133]
CD46:CD4	Yes	No	No
CD46:CD46	Yes	No	Haralambieva et al. [162]
CD4:CD4	Yes	Yes[2]	Moldovan et al. [163]
CD4:CD62L	Yes	No	No
CD4:CXCR4	Yes	Yes[1]	Singer et al. [164]
CD62L:CD62L	Yes	No	No
CD8:CD8	Yes	No	Geng and Raghavan [165]
CD9:CD9	Yes	Yes[1]	Kovalenko et al. [132]
CD9:HLADR	Yes	No	Rubinstein et al. [166]
HLADR:HLADR	Yes	No	Cherry et al. [128]
CD3:CD8	Yes	No	Thome et al. [134]
CXCR4:CXCR4	Yes	Yes[2]	Wang et al. [167]
CD27:CD3	Yes	No	No
CD27:CD46	Yes	No	No
CD122:CD38	Yes	No	No
CD16:CD38	Yes	No	Deaglio et al. [168]
CD38:CD8	Yes	No	No
CD14:CD9	Yes	No	Suzuki et al. [169]
CD14:HLADR	Yes	No	No
CD38:CD9	Yes	No	Zilber et al. [170]
CD46:CD9	Yes	No	Kurita-Taniguchi et al. [171]
CD46:HLADR	Yes	No	No
CD56:CD62L	Yes	No	No
CCR6:HLADR	Yes	No	No
CD19:CD19	Yes	No	No
CD19:CXCR4	Yes	No	Becker et al. [172]
CD19:HLADR	Yes	No	Bobbitt and Justement [173]

Table 3.3: (continued)

Protein complex	Detected by Prox-seq	IntAct database	Literature
CD45RA:CXCR4	Yes	Yes[1]	Badr, Lefevre, and Mohany [174]
CD45RA:HLADR	Yes	No	No
CD62L:HLADR	Yes	No	No
CXCR4:HLADR	Yes	No	No
CD3e:CCR7	No	Yes[1]	Martín-Cófreces et al. [175]
CD4:CCR5	No	Yes[2]	Xiao et al. [176]
CXCR3:CXCR4	No	Yes[2]	Watts et al. [177]
CXCR3:CXCR3	No	Yes[2]	Vinet et al. [178]
Slamf6:Slamf6	No	Yes[1]	Yan et al. [179]
CD9:CD19	No	Yes[1]	Mazurov, Barbashova, and Filatov [180]
CCR5:CCR5	No	Yes[2]	Chakera et al. [181]

Table 3.4: List of antibodies used in the macrophage experiment.

Target	Manufacturer	Catalog #
TLR1	Biolegend	334502
TLR2	R&D Systems	AF2616
TLR4	R&D Systems	MAB6248
TLR6	Biolegend	334702
CD14	Abcam	ab182032
MD2	Abcam	ab24182
LBP	Biolegend	863801
CD36	Biolegend	336202

Target	Manufacturer	Catalog #
TGFBR1	Biolegend	399702
IL-1R	R&D Systems	AF269
IL-6Ra	Biolegend	352802
IL-8Rb	Biolegend	149611
IL-10R	Biolegend	112709
MMP15	R&D Systems	MAB9161
MMP16	R&D Systems	MAB1785

Table 3.5: Sequence of DNA oligonucleotides used to make Prox-seq probes.

Name	Sequence
Probe A	/5AzideN/CGCATTGCATCGTCTCGTGGGCTCGGCHHHH ACHHHHACHHHNGCAG[**barcode**]GATCGCTAAATCGTG
Probe B	/5Phos/TCGTGTCGTGTCTAAAGTCC[**barcode**] ACAAAAAAAAAAAAAAAAAAAAAAAAAAA/3AzideN/

Table 3.6: List of protein barcodes of Prox-seq probes. They are the sequences of the [barcode] regions of probes A and B DNA oligos.

Barcode	Sequence	Barcode	Sequence	Barcode	Sequence
1	AGATCTAT	18	AGAAGAGG	35	ATCAACTA
2	CGAGATTC	19	TAAGGCTC	36	TCCGAGAT
3	TTAGCCAG	20	TTACGGTG	37	CTCAAGGT
4	GAATCTCG	21	GCATGTAC	38	AGCTAGTC
5	GTACGCAT	22	CAAGGTCT	39	GTCAATCG
6	ACATGCGT	23	ATACGTGC	40	CGCTATGA
7	TCACAGCA	24	GCAGTATC	41	ACCGATTG
8	TGACCGAT	25	AGATTCCG	42	GTCTCAAG
9	ATAGCGTC	26	TCAGTCGA	43	AACACCGT
10	AGACCTTG	27	GAACTCTG	44	TTCTCCTC
11	CCACAATG	28	CGATTGAC	45	ATCTCGGA
12	TGAGACCT	29	ACAGTGCT	46	GACACTAC
13	GTAGCACT	30	TAACTGGC	47	CTCTCTCT
14	CAATCGGT	31	CCAGTTAG	48	TCCCCTTT
15	CTAGCTGA	32	GGCTAACT	49	CTCGGAAT
16	CTACGACA	33	CCCGAAGC	50	TCCAGAGT
17	GAAGGAAG	34	TGCTACAG	51	GACTGATC

Table 3.7: List of primers used in the droplet-based Prox-seq protocol.

Name	Sequence
TSO_RNAhybrid	AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG
TSO_PCR	AAGCAGTGGTATCAACGCAGAGT
U_fwd	GCATCGTCTCGTGGGCTC
P5_TSO	AATGATACGGCGACCACCGAGATCTACACGCCTG TCCGCGGAAGCAGTGGTATCAACGCAGAGT*A*C
P7_N7XX_Custom2	CAAGCAGAAGACGGCATACGAGAT[**i7**] AGATGCATCGTCTCGTGGGCTCGG
Read1CustomSeqB	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC
DropPLA_Read2	AGATGCATCGTCTCGTGGGCTCGG
DropPLA_i7Read	CCGAGCCCACGAGACGATGCATCT
10X_SI_PCR	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGC*T*C

Table 3.8: List of primers used in plate-based Prox-seq protocol.

Name	Sequence
SmartSeq2_oligodTVN	AAGCAGTGGTATCAACGCAGAGTAC TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SmartSeq2_oligodTGT	AAGCAGTGGTATCAACGCAGAGTAC TTTTTTTTTTTTTTTTTTTTTTTTTTTTGT
U_fwd	GCATCGTCTCGTGGGCTC
SmartPLA_P5_S5XX	AATGATACGGCGACCACCGAGATCTACAC[**i5**] TATTTGGCATCGTCTCGTGGGCTCGG
SmartPLA_P5_HXXX	AATGATACGGCGACCACCGAGATCTACAC[**i5**] TATTTGGCATCGTCTCGTGGGCTCGG
SmartPLA_P7_N7XX	CAAGCAGAAGACGGCATACGAGAT[**i7**] GCCGAAGCAGTGGTATCAACGCAGAGT
SmartPLA_Read1	TATTTGGCATCGTCTCGTGGGCTCGG
SmartPLA_i5Read	CCGAGCCCACGAGACGATGCCAAATA
SmartPLA_i7Read	ACTCTGCGTTGATACCACTGCTTCGGC

Table 3.9: List of i7 sequencing indices. These are the sequences of the [i7] region in P7_N7XX_Custom2 and SmartPLA_P7_N7XX indexing primers.

Index	Sequence
N701	TCGCCTTA
N702	CTAGTACG
N703	TTCTGCCT
N704	GCTCAGGA
N706	CATGCCTA
N707	GTAGAGAG
N708	CCTCTCTG
N709	AGCGTAGC
N710	CAGCCTCG
N711	TGCCTCTT
N712	TCCTCTAC
N715	CCTGAGAT

Index	Sequence
N716	TAGCGAGT
N718	GTAGCTCC
N719	TACTACGC
N720	AGGCTCCG
N721	GCAGCGTA
N722	CTGCGCAT
N723	GAGCGCTA
N724	CGCTCAGT
N726	GTCTTAGG
N727	ACTGATCG
N728	TAGCTGCA
N729	GACGTCGA

Table 3.10: List of i5 sequencing indices. These are the sequences of the [i5] region of SmartPLA_P5_S5XX and SmartPLA_P5_HXXX indexing primers.

Index	Sequence
S502	CTCTCTAT
S503	TATCCTCT
S505	GTAAGGAG
S506	ACTGCATA
S507	AAGGAGTA
S508	CTAAGCCT
S510	CGTCTAAT
S511	TCTCTCCG
S513	TCGACTAG
S515	TTCTAGCT
S516	CCTAGAGT
S517	GCGTAAGA
S518	CTATTAAG
S520	AAGGCTAT
S521	GAGCCTTA
S522	TTATGCGA

Index	Sequence
H001	CCACAATG
H002	TGAGACCT
H003	GCACACGC
H004	AGAGAGAG
H005	TCACAGCA
H006	CTATAGTA
H007	CGAGATTC
H008	GTAGCACT
H009	TTAGCCAG
H010	TGACCGAT
H011	CAATCGGT
H012	ATAGCGTC
H013	GAATCTCG
H014	CTAGCTGA
H015	AGACCTTG
H016	TAAGGCTC

CHAPTER 4

HIGH-THROUGHPUT RNA SEQUENCING OF PARAFORMALDEHYDE-FIXED SINGLE CELLS

In the previous chapter, we presented Prox-seq, a single-cell sequencing method for quantification of mRNA, and surface protein and protein complex. A natural extension of Prox-seq is to target intracellular proteins. Before this could happen, some technical challenges need to be solved. The first challenge is that detection of intracellular proteins necessitates fixation and permeabilization of the cells through chemical treatments, such as paraformaldehyde. These treatments significantly hinder the ability to measure the mRNA in single cells. In this chapter, we present FD-seq (Fixed Droplet RNA sequencing), a high-throughput scRNA-seq method for paraformaldehyde-fixed cells that can solve the first challenge in developing intracellular Prox-seq.

This chapter is reproduced in part from Phan et al. [147] under Open Access Creative Commons CC BY license.

4.1 Summary

Single-cell transcriptomic studies that require intracellular protein staining, rare cell sorting, or inactivation of infectious pathogens are severely limited. This is because current high-throughput single-cell RNA sequencing methods are either incompatible with or necessitate laborious sample preprocessing for paraformaldehyde treatment, a common tissue and cell fixation and preservation technique. Here we present FD-seq (Fixed Droplet RNA sequencing), a high-throughput method for droplet-based RNA sequencing of paraformaldehyde-fixed, permeabilized and sorted single cells. We show that FD-seq preserves the RNA integrity and relative gene expression levels after fixation and permeabilization. Furthermore, FD-seq can detect a higher number of genes and transcripts than methanol fixation. We first apply FD-seq to analyze a rare subpopulation of cells supporting lytic reactivation of the

human tumor virus KSHV, and identify *TMEM119* as a potential host factor that mediates viral reactivation. Second, we find that infection with the human betacoronavirus OC43 leads to upregulation of pro-inflammatory pathways in cells that are exposed to the virus but fail to express high levels of viral genes. FD-seq thus enables integrating phenotypic with transcriptomic information in rare cell subpopulations, and preserving and inactivating pathogenic samples.

4.2 Introduction

Single-cell RNA sequencing (scRNA-seq) has found many important biological applications, from the discovery of new cell types [9] to mapping the transcriptional landscape of human embryonic stem cells [182]. Droplet-based scRNA-seq methods, such as Drop-seq [54] and 10x Chromium [56], are particularly powerful due to their high throughput: thousands of single cells can be analyzed in a single experiment. However, even with these high-throughput techniques, analyzing rare cell subpopulations remains a challenging task, often requiring protein-based enrichment for the subpopulation of interest before scRNA-seq [183, 184].

Many cell types require intracellular protein staining to be enriched. For example, Foxp3 is an intracellular marker of regulatory T cells [185], and Oct4 and Nanog are intracellular reprogramming markers of induced pluripotent stem cells [186]. Intracellular protein staining requires cell fixation, which is most commonly achieved with paraformaldehyde (PFA) or methanol fixation. High-throughput techniques like Drop-seq and 10x Chromium have been shown to be compatible with methanol-fixed cells [187, 188]. In many applications, however, PFA is preferred over methanol fixation due to the improved signal-to-background ratio in intracellular staining [189, 190], better preservation of intracellular structures' integrity [191], or simply because methanol fixation does not produce a signal [192].

Important advances to scRNA-seq of PFA-fixed cells have recently been made. A well plate-based method [183] was shown to be compatible with PFA-fixed cells, but the relatively low throughput nature of this method excludes its applicability from a wide range of

problems that search for rare phenotypes in broad cellular populations. Most recently, a high-throughput scRNA-seq method that combines well plate-based combinatorial indexing and the 10x platform, called scifi-RNA-seq [59], has been shown to work with formaldehyde-fixed single cells and single nuclei. However, scifi-RNA-seq requires a separate reverse transcription step before droplet encapsulation, thus complicating the sample processing step. Another method called inCITE-seq [193] has been developed for sequencing formaldehyde-fixed single nuclei with 10x. In contrast to scifi-RNA-seq, inCITE-seq performs cross-link reversal and reverse transcription inside the droplets. Like scifi-RNA-seq, however, inCITE-seq requires laborious preprocessing of the samples. Because inCITE-seq has only been demonstrated on formaldehyde-fixed cell nuclei, most of the mature mRNA transcripts that reside in the cytoplasm cannot be measured. In summary, despite many technical advances, sequencing of PFA-fixed single cells continues be a complicated process with several shortcomings.

The ability to study PFA-fixed samples is important, particularly in virology. For example, Kaposi's sarcoma-associated herpesvirus (KSHV), also known as human herpesvirus type 8 (HHV-8), is a human gammaherpesvirus that causes a number of malignancies such as Kaposi's sarcoma, primary effusion lymphoma, and multicentric Castleman's disease [194, 195]. There is considerable interest in unraveling the molecular details of the host factors that modulate KSHV latency and reactivation, because both latency and low-level reactivation are known to contribute to viral tumorigenesis [196], and therapeutic induction of reactivation could sensitize latently infected cells to currently available anti-herpesvirus drugs20. However, studying KSHV reactivation is challenging because of the low level of reactivation: only a small proportion of latently infected cells typically undergo reactivation, even when treated with known chemical inducing agents such as sodium butyrate (NaBut) and tetradecanoyl phorbol acetate (TPA) [194]. Single-cell transcriptomic analysis of reactivated cells, therefore, requires enrichment beforehand, so that the majority of sequencing reads are not spent on non-reactivated cells. Such enrichment would involve PFA fixation and intracellular staining of a viral protein marker.

Another use of PFA fixation is to inactivate infected cells or patient-derived materials. This would greatly facilitate the study of highly pathogenic virus-infected cells outside high-containment BSL-3 facilities, which are often not readily available. The importance of such flexibility has become evident during the ongoing COVID-19 pandemic.

Here we describe FD-seq (Fixed Droplet RNA sequencing), an easy-to-use, droplet-based high-throughput method for single-cell RNA sequencing of PFA-fixed, permeabilized, stained, and sorted whole cells. We show that FD-seq preserves the RNA integrity and relative transcripts abundances compared to Drop-seq for live cells, and that FD-seq yields a higher number of detected genes and transcripts than the methanol fixation method. By applying FD-seq to studying KSHV reactivation, we find that *TMEM119*, a gene with unknown functions, potentially plays a role in promoting KSHV reactivation. We then utilize FD-seq to study OC43 infection in fixed cells. OC43 is a human betacoronavirus that causes the common cold, and it has successfully been used to discover drugs that can inhibit SARS-CoV-2 replication in vitro [197]. We find that after OC43 infection, a large subpopulation of cells express low levels of viral genes, but highly upregulate pro-inflammatory genes compared to cells with higher expression of viral genes. Our study shows that FD-seq is a valuable tool for studying rare cell subpopulations, and provides researchers with greater flexibility in choosing the cell fixation, permeabilization, or pathogen inactivation method.

4.3 Results

4.3.1 Development of FD-seq for Sequencing of PFA-Fixed Single Cells

To facilitate ease of adoption, we developed FD-seq based on Drop-seq [54]. In the standard Drop-seq protocol, single cells are partitioned with uniquely barcoded ceramic beads inside nanoliter droplets in oil, using a microfluidic device. The cells are individually lysed inside the droplets, and their mRNAs are captured by the oligonucleotides on the barcoded beads. Next, the droplets are broken to recover the beads, and the beads are extensively washed

to remove uncaptured mRNAs. After pooling the beads, the captured mRNAs undergo reverse transcription, exonuclease I digestion to remove the free oligonucleotides on the beads, and whole transcriptome amplification. Finally, the barcoded and amplified complementary DNAs (cDNAs) are tagmented and sequenced.

PFA fixation of cells induces cross-linking between nucleic acids and proteins. To extract RNA from PFA-fixed cells, a cross-link reversal step through heating is therefore required. We reasoned that for FD-seq to be readily used without additional significant processing steps, cross-link reversal must be performed inside the droplets following droplet encapsulation.

Therefore, we first tested different heating conditions for the cross-link reversal step at bulk level, followed by lysing the uncross-linked cells with Drop-seq lysis buffer, and total RNA extraction using a commercial kit (Figure 4.1a, Supplementary Figure 4.6, see "Methods"). We found that a 1-h incubation at 56°C in the standard Drop-seq lysis buffer efficiently reversed PFA cross-linking, in agreement with the previous literature [183]. The RNA yield could be improved further by adding proteinase K to the lysis buffer at an optimal concentration of 40 U/mL (Figure 4.1a and Supplementary Figure 4.6a). Whereas proteinase K treatment did not significantly affect the RNA quality at any of the tested concentrations, consistently resulting in high-quality total RNA as demonstrated by the high RNA integrity numbers (above 8.0, Supplementary Figure 4.6b, c), we observed that increasing the proteinase K concentration above 40 U/mL resulted in a reduction in the overall RNA yield and higher variability (Supplementary Figure 4.6a).

Next, we compared the single-cell capture efficiency, and the extent of cross-droplet RNA contamination between the standard Drop-seq method on live cells, and the FD-seq method on PFA-fixed cells by performing species-mixing experiments (Figure 4.1b). For Drop-seq, we analyzed a 1-to-1 mixture of live BC3 cells, a human primary effusion lymphoma (PEL) cell line, and mouse 3T3 cells. For FD-seq, we separately fixed BC3 and 3T3 cells with 4% PFA, permeabilized them with 0.1% Triton-X, and analyzed a 1-to-1 mixture of these

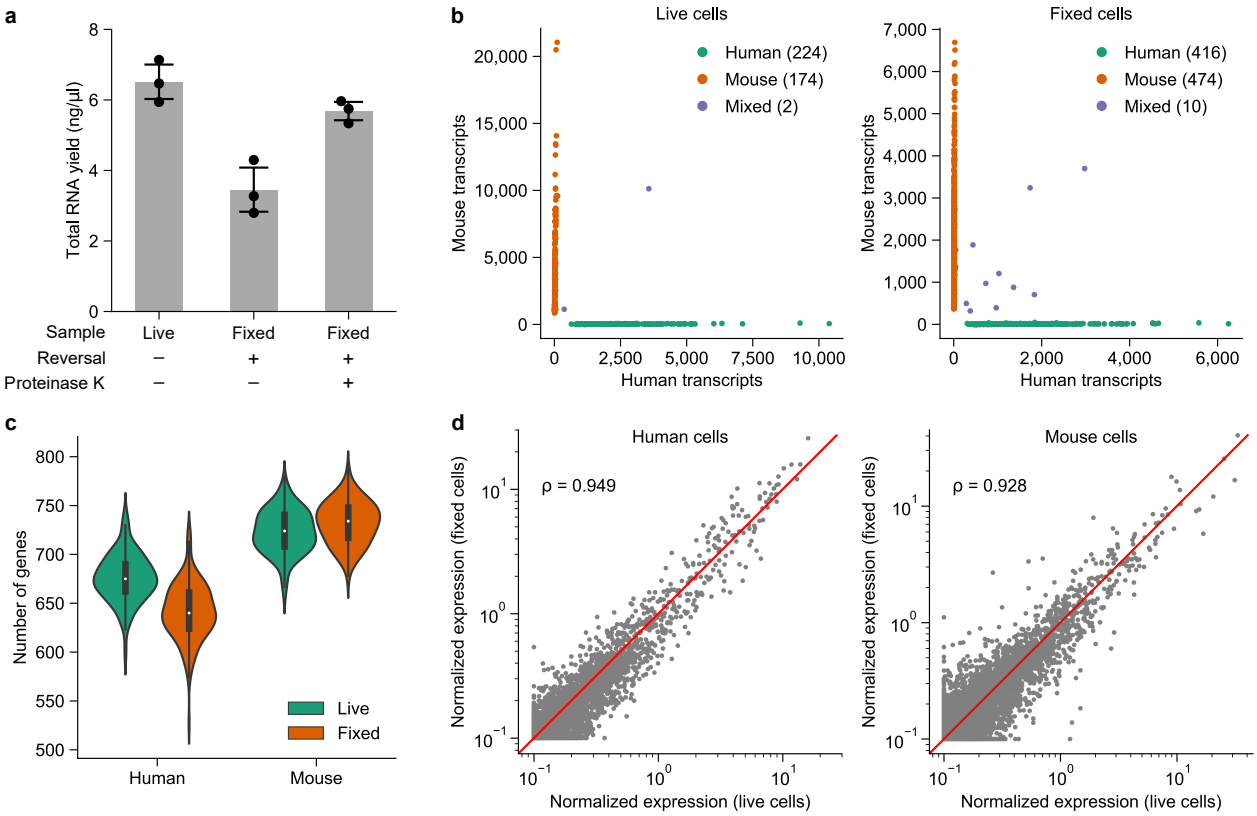

Figure 4.1: Benchmarking and validation of FD-seq. (a) Bar plots showing total RNA yield from bulk live cells, and bulk fixed cells that underwent cross-link heat reversal (1 h at 56°C) with or without 40 U/mL of proteinase K. Data are presented as mean standard deviation. n = 3 technical replicates. (b) Species-mixing plots showing the single-cell capture efficiency of Drop-seq and FD-seq. The multiplet rate for live cells and fixed cells were approximately 0.5% and 1%, respectively. Human BC3 cells were combined with mouse 3T3 cells at equal concentration, and processed with Drop-seq or FD-seq. See the "Methods" section for more details. (c) Violin plots and box plots showing the number of detected genes in live and fixed cells for each species. For this analysis, only cells with at least 1500 transcripts were considered, and 1,000 transcripts were randomly sampled from each single cell. The white dots inside the violin plots represent the median of the data, the black boxes represent the first and third quartiles, and the black lines represent the values 1.5× the interquartile range beyond the first and third quartiles. n = 157 and 164 single cells for live and fixed human samples. n = 150 and 267 single cells for live and fixed mouse samples. (d) Comparison of the normalized expression level of each gene between live and fixed cells for each species (see "Methods" section). Each dot represents the average expression level of a gene, and the red line indicates the line y = x. The plots also show the Pearson's correlation coefficient ρ of the log-normalized gene expression level between live and fixed cells for each species.

two cell lines. The sequencing results were aligned to a combined human-mouse reference genome, and the number of human and mouse transcripts per single-cell barcode was then

calculated. The rate of cell barcodes having both mouse and human transcripts, which indicates cross-droplet contamination and/or multiple cells captured in the same droplet, were very similar and minimally observed in both live and fixed cell samples (approximately 0.5% and 1% for live and fixed cells, respectively, Figure 4.1b). This indicates that FD-seq has similar single-cell capture efficiency as Drop-seq.

We also observed that the number of genes detected in fixed cells was comparable to that of live cells across species (median of 640 genes in fixed cells compared to 675 genes in live cells for human cells, and median of 734 genes compared to 724 genes for mouse cells) (Figure 4.1c). The relative expression levels of the detected genes were well correlated between live and fixed cells (Figure 4.1d), even though fewer transcripts were detected in fixed cells than in live cells on average (Figure 4.1b and Supplementary Figure 4.7a). Finally, live and fixed cells showed a similar percentage of reads mapped to introns and exons (Supplementary Figure 4.7b).

We also assessed the technical repeatability of FD-seq by analyzing two technical replicates of fixed and permeabilized A549 cells, a human lung epithelial cell line, with FD-seq. We found that the two replicates show strong agreement in terms of the number of detected transcripts and genes, and the relative gene expression level (Supplementary Figure 4.8).

Taken together, these results demonstrate that FD-seq is a reliable method for the whole transcriptome analysis of PFA-fixed single cells. The performance of single-cell sequencing is maintained with fixed cells when FD-Seq is utilized, and FD-seq shows comparable performance to the standard Drop-seq protocol used for live, unfixed cells.

4.3.2 PFA Fixation Detects a Higher Number of Genes and Transcripts Compared to Methanol Fixation in Single Cells

We next investigated how PFA fixation in FD-seq compare to methanol (MeOH) fixation. Methanol fixation has recently been shown to be compatible with Drop-seq by introducing a suitable rehydration step before droplet generation [187]. Here we fixed A549 cells with

either PFA or methanol, and then analyzed the single-cell transcriptomes by RNA-seq using the FD-seq or Drop-seq protocol, respectively (see "Methods" section).

We found that PFA fixation resulted in a higher number of detected transcripts, or UMIs, (median 8,083 compared to 6,194 transcripts) and genes (median 3,049 compared to 2,856 genes), and a lower percentage of mitochondrial genes (median 6.2% compared to 11.9%) than methanol fixation (Figure 4.2a). To determine the effect of sequencing depth on the number of UMIs and genes, we randomly sub-sampled the sequencing reads and found that PFA fixation consistently returned a higher number of UMIs and genes across different read depths (Figure 4.2b). The effect of sequencing depth on the number of transcripts is much more pronounced: the difference between the two fixation techniques could reach almost 2000 UMIs. Compared to methanol fixation, FD-seq has a higher proportion of reads mapped to the coding and untranslated regions, and a lower proportion mapped to the intergenic and intronic regions (Figure 4.2c). Despite this, the average gene expression levels were well correlated between the two fixation methods (Pearson's correlation coefficient was approximately 0.90, Figure 4.2d). Lastly, we estimated the technical variance of each method by calculating the gene-level coefficient of variation (CV) [61]. As expected, for both fixation methods, gene abundance negatively correlates with gene variance (Figure 4.2e). In addition, both methods exhibited very similar technical variance. In summary, FD-seq was able to recover more genes and transcripts than methanol fixation, and both PFA-fixed cells and methanol-fixed cells showed similar average gene expression levels and technical variance.

4.3.3 FD-seq Reveals Heterogeneity in KSHV Reactivation Single Cells

Having established the validity of FD-seq for single-cell transcriptome analysis of PFA-fixed cells, we applied FD-seq to test a hypothesis, that the heterogeneity in specific host genes contributed to the restricted KSHV reactivation in human primary effusion lymphoma (PEL) cells BC3, and to identify those genes.

To identify lytically reactivated BC3 cells, we used flow cytometry to measure the expres-

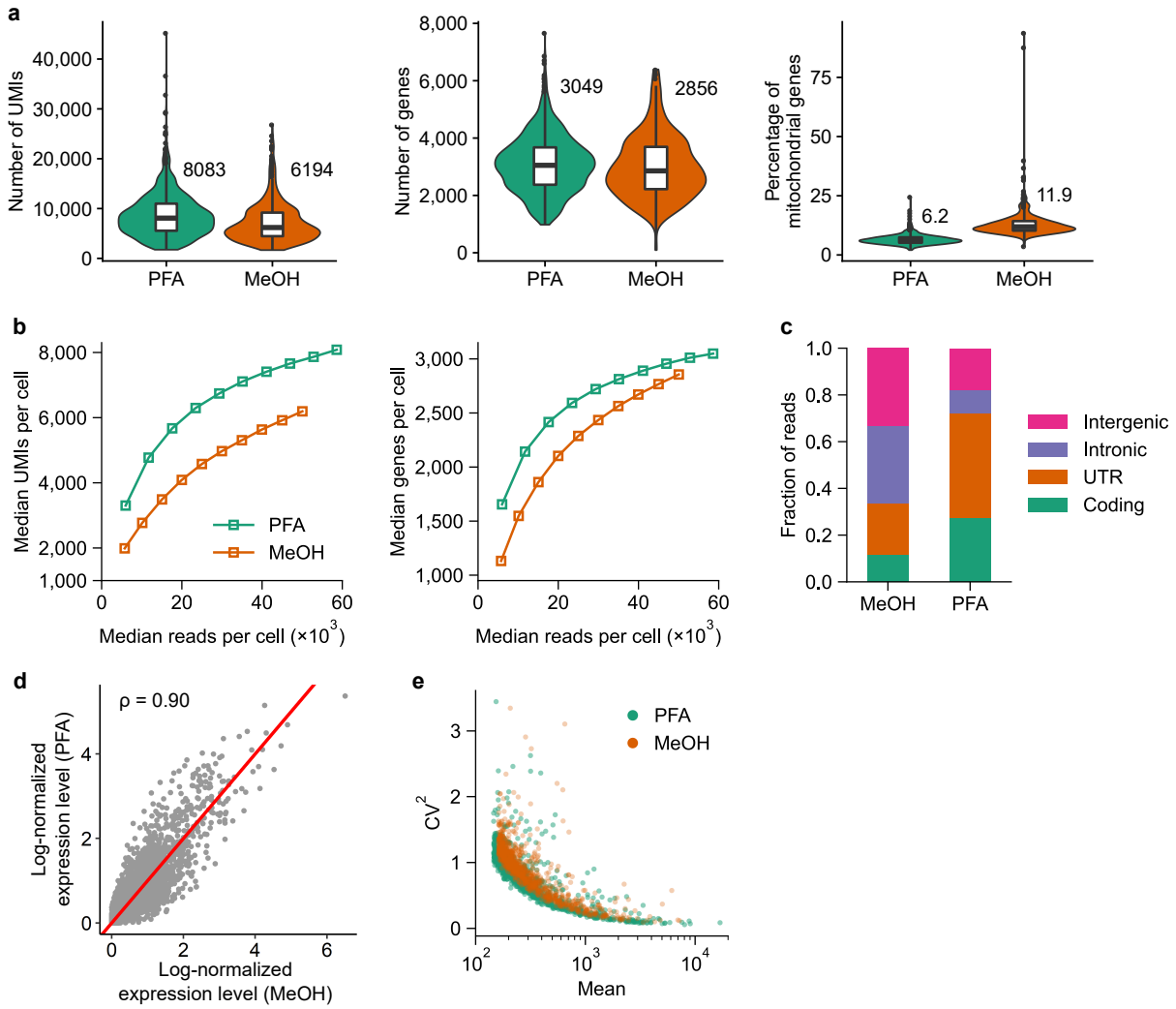

Figure 4.2: Comparison between PFA and methanol fixation shows higher gene and transcript recovery with FD-seq. (a) Violin and box plots of the number of UMIs, the number of genes, and the percentage of mitochondrial genes detected in single A549 cells for each fixation method. The PFA and methanol (MeOH) samples received approximately 58,600 and 50,000 median reads/cell, respectively. The numbers by the violin plot indicate the median values. The middle line inside the box indicates the median, the upper and lower edges of the box indicate the first and third quartiles, and the whiskers extend to 1.5× the interquartile range beyond the first and third quartiles. (b) The effects of sequencing depth on the number of detected UMIs and genes per cell. (c) The distribution of mapped reads to different genomic regions. (d) Correlation of the log-normalized average expression level of each gene between the two fixation methods (see "Methods" section). The plot also shows the Pearson's correlation coefficient ρ between the two methods. The red line indicates y = x. (e) Technical variance of each gene estimated by the gene's mean and squared coefficient of variation (CV^2). In a–e, n = 999 and 498 single cells for PFA and methanol samples, respectively.

sion of the intracellular viral glycoprotein K8.1, which is only expressed during lytic reactivation. As expected, we did not observe any appreciable K8.1 expression in untreated BC3 cells (Supplementary Figure 4.9a). Treatment with NaBut or TPA, which are known to induce KSHV reactivation [194], resulted in a small fraction of cells expressing K8.1 (∼8% compared to ∼2.5%, respectively, Supplementary Figure 4.9b). To enrich for reactivated cells, TPA-treated BC3 cells were fixed, permeabilized, stained for K8.1 viral protein, and sorted by fluorescence-activated cell sorting (FACS) based on the K8.1 expression level (2.2% of the cells were K8.1-positive, Supplementary Figure 4.10), followed by single-cell transcriptome analysis with FD-seq. We obtained high-quality data for 1035 K8.1-positive (reactivated) single cells and 286 K8.1-negative (latent, non-reactivated) cells. High dimensional clustering and visualization showed a clear separation between reactivated and non-reactivated cells (Figure 4.3a), and analysis of the *K8.1* mRNA level confirmed the enrichment of the K8.1+ cell subpopulation of interest (Figure 4.3b). Moreover, the high proportion of viral transcripts in the K8.1+ subpopulation compared to the K8.1− population (69% and 4% on average, respectively) confirmed that the sorted population was indeed mostly composed of reactivated cells (Figure 4.3c, d).

Herpesvirus reactivation involves the highly regulated, sequential expression of immediate-early, early, and late viral genes [198] that was recapitulated in our FD-seq results (Figure 4.3e–g). By ordering the cells by the percentage of viral genes relative to the total transcript content, we found that the relative abundance of immediate-early viral genes increased with total viral gene content, then plateaued after the total viral transcript abundance reached 50%. Early viral transcripts also increased early and monotonically with the abundance of total viral transcripts, without plateauing like the immediate-early genes. On the other hand, late viral transcripts were only detected in cells with more than 25% total viral transcripts, and increased strongly with higher total viral transcript abundance. Thus, FD-seq's results are in agreement with the expected kinetics of viral gene expression obtained from population-averaged measurements, and suggest that the percentage of viral transcript con-

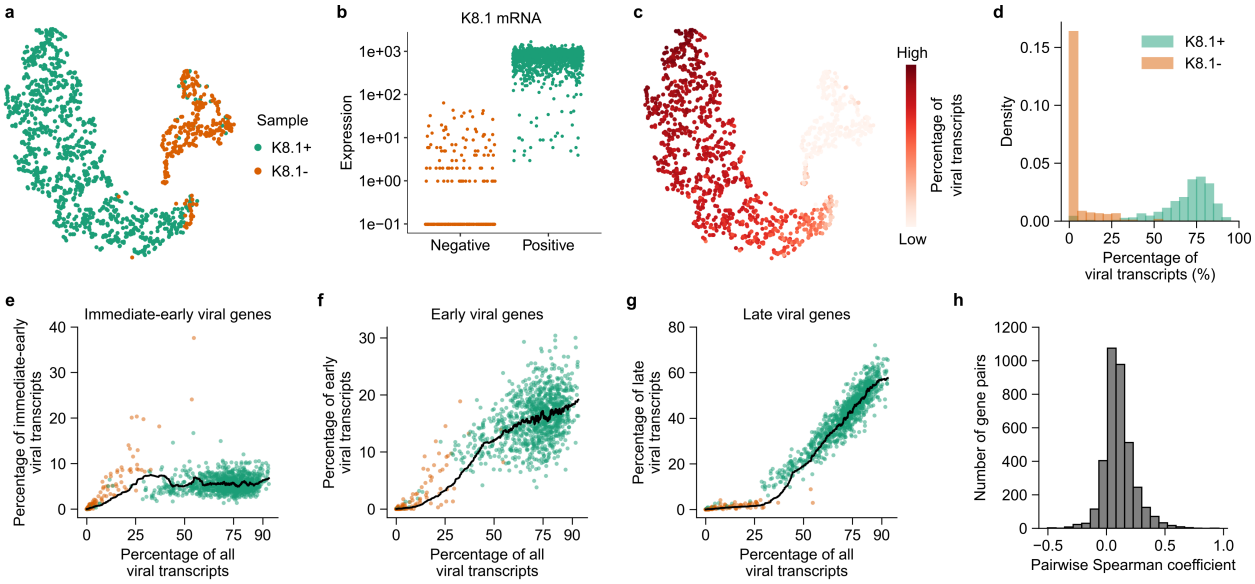

Figure 4.3: FD-seq reveals heterogeneity in KSHV viral reactivation. (a) t-SNE plot of K8.1+ (green, lytically reactivated) and K8.1− (orange, non-reactivated) BC3 cells as analyzed by FD-seq. Clustering was performed using both host and viral transcripts. (b) KSHV *K8.1* mRNA levels in the sorted K8.1− (non-reactivated) and K8.1+ (reactivated) subpopulations. (c) t-SNE plot of reactivated and non-reactivated cells as in (a), colored by the percentage of total viral transcripts. (d) Histogram showing the percentages of detected transcripts that are from KSHV in K8.1+ and K8.1− BC3 cells. (e–g) Change in the percentages of (e) intermediate-early, (f) early, and (g) late viral genes as a function of the percentage of total viral transcripts. The black lines indicate the moving averages (50-cell window). (h) Histogram of pairwise Spearman correlation coefficients between viral genes.

tent is a good indicator of the stage of KSHV reactivation.

Interestingly, we found that the K8.1+ population was highly heterogeneous in viral transcript expression: the proportion of viral transcripts among all detected transcripts varied from below 50% to over 90% (Figure 4.3d), and the correlation between the expression levels of viral genes was very low, even between viral genes that have the same kinetics (Figure 4.3h and Supplementary Figure 4.11a). This agrees with two recent studies that showed heterogeneity in host cell factor abundance at the single-cell level in herpes simplex virus type 1 (HSV-1) infection [199], and low correlation in expression of viral genes in a murine gammaherpesvirus infection model [200]. This may partially be caused by the possibility that the sorted K8.1+ cells were at different stages of reactivation. Indeed, flow cytometry analysis showed that K8.1 protein abundance within the positive population varied

over one order of magnitude (Supplementary Figure 4.10). However, this alone could not sufficiently explain the poor correlation between viral genes, because we also observed a low correlation between cells in the same stages of reactivation (i.e., cells with similar viral transcript abundance) (Supplementary Figure 4.11b).

In short, we successfully applied FD-seq to characterize KSHV-infected cells undergoing reactivation from latency, revealing the highly heterogeneous nature of this process.

4.3.4 FD-seq Shows That TMEM119 Facilitates KSHV Reactivation

Next, we sought to identify host genes that facilitate KSHV reactivation by looking for differentially expressed host genes that are positively correlated with viral transcript abundance. We only looked for positively correlated host genes in our analysis, because the majority of differentially expressed genes were negatively correlated with the relative abundance of viral transcripts (Supplementary Figure 4.12a). While this could suggest that their downregulation promotes KSHV reactivation in the cells [201], the lower expression level measured in these genes could be due to undersampling caused by the possibility that sequencing reads are mostly used by the highly abundant viral transcripts in the K8.1+ subpopulation (which can account for up to 96% of all detected transcripts, Figure 4.3d).

We found four host genes whose expression positively correlated with the level of KSHV transcripts: *ISCU, CDH1, CORO1C* and *TMEM119* (q-value $< 10^{-100}$ for all four genes, Figure 4.4a). We observed that the expression of *ISCU* was relatively low in non-reactivated cells, and increased significantly with the abundance of viral transcripts. On the other hand, *CDH1, CORO1C* and *TMEM119* were mostly undetectable in non-reactivated cells, and increased by 1–2 orders of magnitude in cells expressing high levels of viral transcripts. We also confirmed the upregulation of *ISCU, CORO1C,* and *TMEM119* following KSHV reactivation in bulk cell samples by qPCR (Supplementary Figure 4.12b, c).

To determine whether the strong correlation between these host transcripts and viral gene expression means that these genes modulate KSHV reactivation induction and/or efficiency,

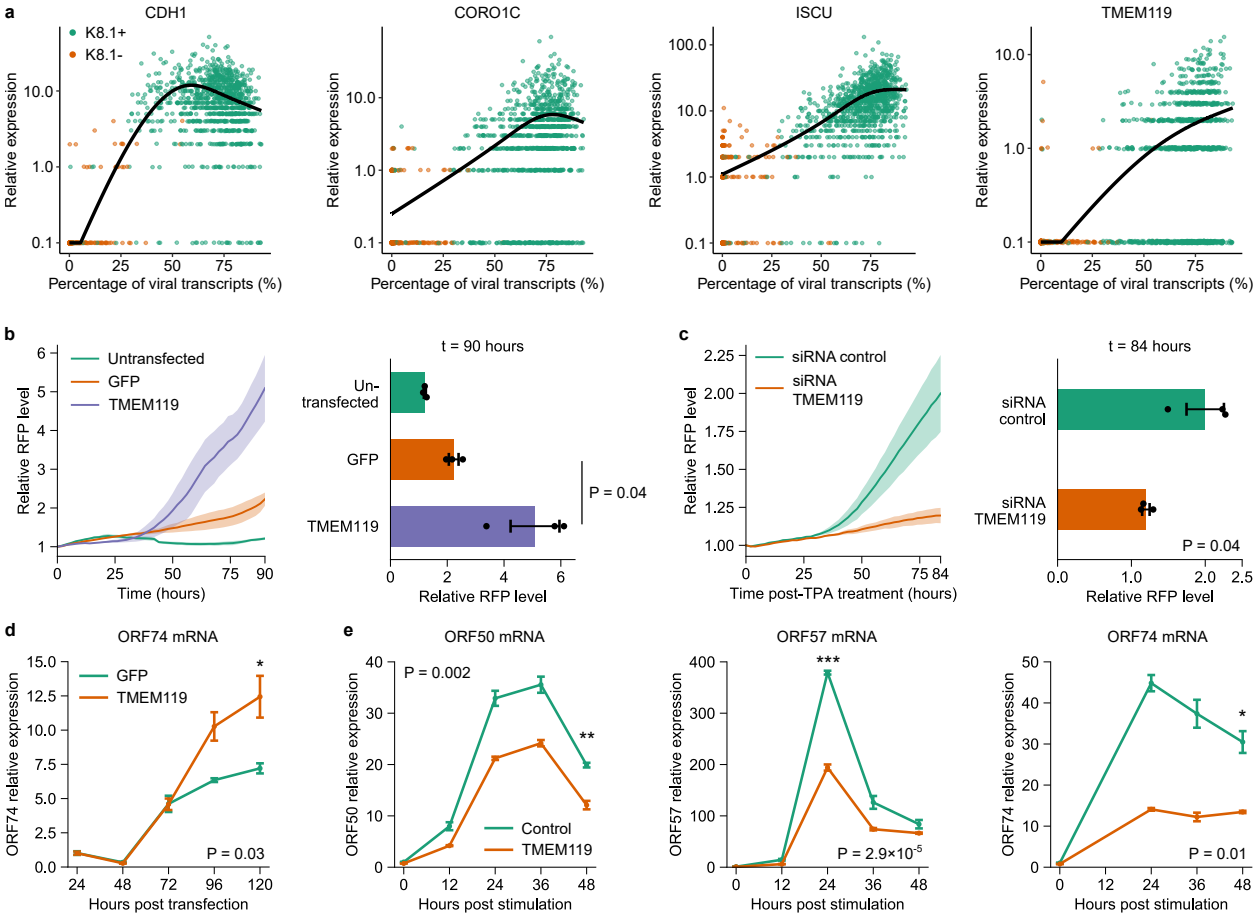

Figure 4.4: FD-seq identifies TMEM119 as a potential host factor that mediates KSHV reactivation. (a) Plots showing the relative expression of the four host genes *CDH1, CORO1C, ISCU*, and *TMEM119* (based on the genes' normalized counts) as a function of the percentage of KSHV transcripts. Each dot indicates a single cell. (b) Live-cell imaging analysis of RFP expression, which is indicative of KSHV reactivation level, in untransfected HEK293T.rKSHV219 cells or cells transfected with *TMEM119* or GFP control (left), and end-point quantification of RFP expression at 90 h (right). (c) Live-cell imaging analysis of RFP expression in HEK293T.rKSHV219 cells transfected with control or *TMEM119*-targeting siRNAs for 48 h, followed by treatment with 2 ng/mL TPA (left), and end-point quantification of RFP expression at 84 h (right). In (b, c) the lines indicate the mean, and the ribbons indicate the s.e.m. of the data. (d) Time-course RT-qPCR analysis of KSHV ORF74 abundance in HEK293T.rKSHV219 cells transfected with *TMEM119* or GFP control. (e) Time-course RT-qPCR analysis of KSHV genes *ORF50, ORF57*, and *ORF74* levels in HEK293T.rKSHV219 cells transfected with control or *TMEM119*-targeting siRNAs. (b–e) Data are represented as mean ± s.e.m. n = 3 biological replicates per sample. (d, e) $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (one-sided Welch's t-test).

we next tested the effect of their overexpression on KSHV reactivation. To this end, we used live-cell imaging to monitor KSHV reactivation in HEK293T.rKSHV219 cells following

exogenous expression of *CDH1, CORO1C, ISCU* or *TMEM119*. HEK293T.rKSHV219 are HEK293T cells that have been latently infected with the recombinant KSHV.219 virus strain, which constitutively expresses GFP, and further encodes RFP under control of the viral lytic PAN promoter [202]. Overexpression of *CDH1, CORO1C,* or *ISCU* had no significant effect on KSHV reactivation efficiency. On the other hand, we observed a clear increase in RFP expression, which is indicative of viral reactivation, upon exogenous expression of *TMEM119* relative to a GFP-encoding control vector (Figure 4.4b and Supplementary Figure 4.13a–c). Conversely, silencing of endogenous *TMEM119* significantly reduced the level of TPA-induced KSHV reactivation (Figure 4.4c and Supplementary Figure 4.13b, c). These results were corroborated by RT-qPCR analysis of KSHV lytic gene expression, which showed that *TMEM119* overexpression enhanced KSHV *ORF74* expression compared to the GFP control (Figure 4.4d), while *TMEM119* silencing led to a reduction in KSHV genes *ORF50, ORF57* and *ORF74* expression (Figure 4.4e). Together, these results showed that *TMEM119* positively modulates KSHV reactivation efficiency.

4.3.5 FD-seq Reveals Pro-Inflammatory Signatures in a subpopulation of OC43-Infected Cells

We next applied FD-seq to study the infection of the betacoronavirus OC43 in single human lung cells. OC43 is a human pathogen that causes the common cold, and is a close relative of SARS-CoV-2. A recent study showed that many drugs that inhibit the replication of OC43 also inhibit SARS-CoV-2 replication in vitro [197], suggesting that OC43 is a good model system for SARS-CoV-2 infection.

First, we infected A549 cells with OC43 at a multiplicity of infection (MOI) of 1, fixed the cells with PFA, and performed FD-seq on mock-infected and OC43-infected cells. We obtained 1167 and 1924 high-quality single cells from mock-infected and OC43-infected cells, respectively. After regressing out the effects of cell cycle variations (Supplementary Figure 4.14a, b), high dimensional clustering yielded three main clusters of cells (Figure 4.5a).

Cluster 0 mainly contained the mock-infected cell, while clusters 1 and 2 mainly contained OC43-infected cells (Figure 4.5b). At an MOI of 1, more than 70% of the cells showed some level of viral gene expression (Figure 4.5c). Interestingly, this infected population consisted of two subpopulations, clearly separated by the number of detected viral transcripts: a larger subpopulation expressing a low number of viral transcripts (below 100 viral transcripts, median 2 transcripts) and a smaller subpopulation expressing a much higher number of viral transcripts (above 100 viral transcripts, median 428 transcripts) (Figure 4.5d). This small subpopulation of highly infected cells corresponded to cluster 2 (Figure 4.5e and Supplementary Figure 4.14c, d).

Next, we investigated the expression profile of OC43 viral genes as a function of total viral gene level within cluster 2's cells. Gene *N* (which encodes the viral nucleocapsid protein [203]) was uniformly highly expressed (Figure 4.5f). The expression of the remaining viral genes was more heterogeneous, with *ORF1ab, M* (membrane gene), and *S* (spike gene) being more highly expressed overall. These patterns of viral gene expression level were also observed in all infected cells (Supplementary Figure 4.15).

On the other hand, cluster 1 consisted of cells that were exposed to the virus but failed to express high levels of viral genes (Figure 4.5d and Supplementary Figure 4.14c, d). Interestingly, this cluster was also enriched in pro-inflammatory genes, such as *CXCL1, CXCL5, CCL2* and *NFKBIA* (Figure 4.5g and Supplementary Figure 4.16). *CXCL1* and *CXCL5* encode the protein ligands of the CXC chemokine receptor 2 (CXCR2), and are crucial to neutrophil recruitment [204]. *CCL2* encodes a different chemokine that is responsible for monocyte recruitment from the bone marrow [205]. *NFKBIA* encodes the inhibitor IκBα to the transcription factor NF-κB, which is an important regulator of the immune response [3].

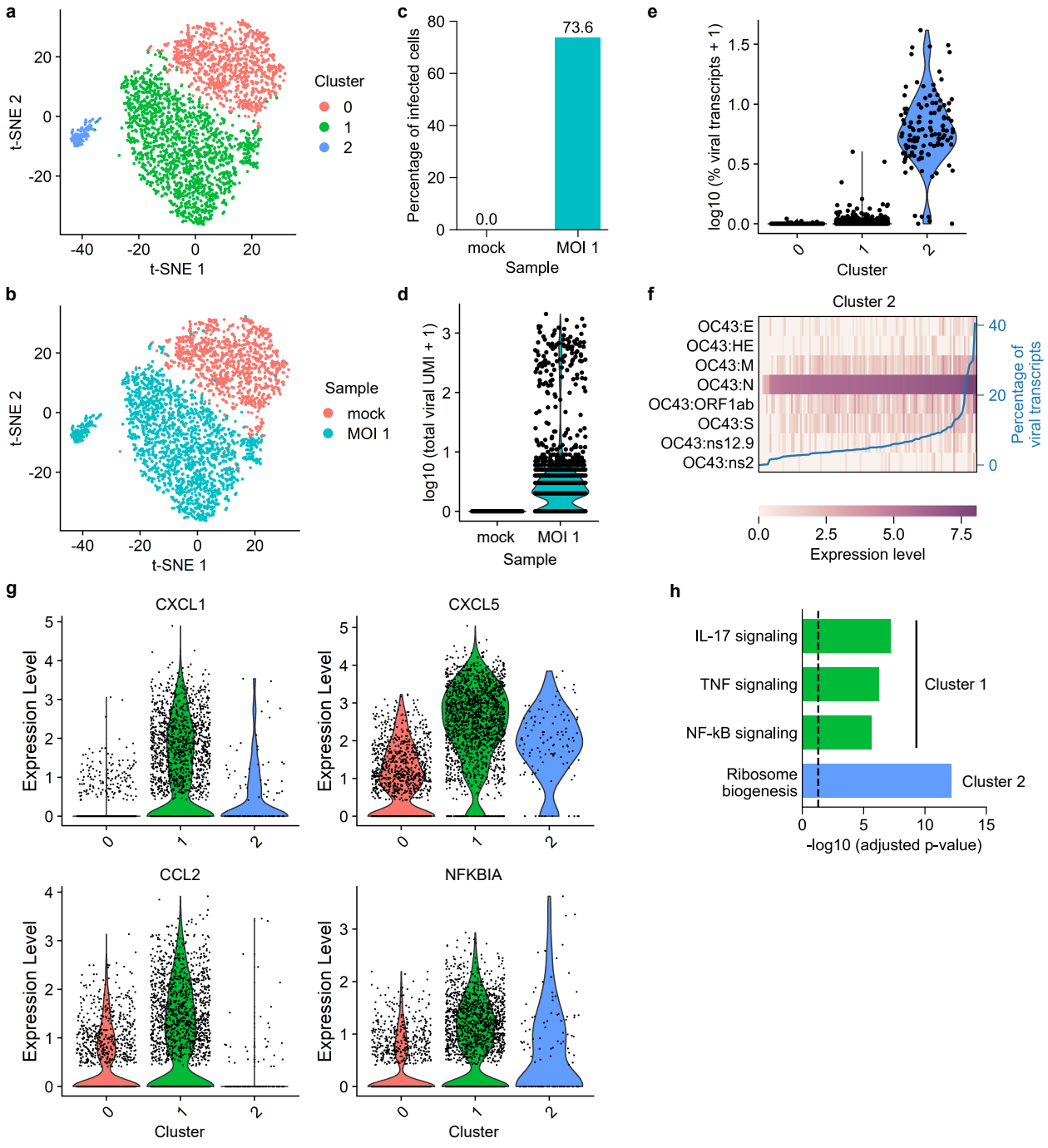

Figure 4.5: Single-cell heterogeneity and pro-inflammatory signatures after OC43 coronavirus infection. (a, b) t-SNE plots with cells colored by (a) cluster identity or (b) sample type. (c) Bar plot showing the percentage of infected cells (defined as cells that expressed at least 1 detected viral transcript) of mock-infected and MOI 1 sample. (d) Violin plot showing the distribution of total viral transcript counts. (e) Violin plot showing the distribution of the percentage of the total viral transcript by cluster identity. (f) Heatmap showing the relative expression level of each viral gene in each single cell in cluster 2. The blue line shows each cell's percentage of total viral transcripts.

Figure 4.5: (continued) (g) Violin plots showing the expression of four representative immune-related genes that are upregulated in cluster 1. The expression levels in (f, g) were log-normalized transcript counts. (h) Bar plot showing the multiple comparison corrected P-values (Fisher's one-tailed test with g:SCS correction [206]) of upregulated KEGG pathways in clusters 1 and 2. The vertical dashed line indicates P-value = 0.05.

KEGG pathway analysis [206] showed that cluster 1 was enriched for three main pro-inflammatory pathways: TNF, IL-17, and NF-κB signaling pathways (Figure 4.5h). Cluster 2 was enriched for the ribosome biogenesis pathway, suggesting an increased need for protein biogenesis during OC43 replication.

In short, FD-seq revealed that the majority of cells did not express a high level of viral genes after exposure to OC43 at MOI of 1, and that these cells upregulated pro-inflammatory genes. These findings are in agreement with a previous characterization of HSV-1 infection at the single-cell level by Drayman et al. [199], which showed that most HSV-1 infected cells did not express high levels of viral genes.

4.4 Discussion

Here we present FD-seq, a method for high-throughput droplet-based RNA sequencing of PFA-fixed, permeabilized, stained, and sorted single cells. FD-seq is particularly useful for sequencing rare cell subpopulations that require intracellular staining and FACS-enrichment, and for rendering infectious samples safe for handling. Although Drop-seq has been shown to work with methanol fixation [187], we demonstrated that FD-seq performs better than methanol fixation by detecting more genes and more transcripts. FD-seq will increase the flexibility for researchers in using high-throughput scRNA-seq, because PFA fixation has been shown to be better than methanol fixation for many applications [189–192], and because PFA is a very commonly used fixative.

One problem with RNA sequencing of PFA-fixed cells is the need for cross-link reversal, which is usually done through heating. To avoid cross-cell contamination, the reversal step

has to be done while cells are still inside the droplets. scifi-RNA-seq [59] circumvented this challenge of cross-link reversal by performing reverse transcription with permeabilized cells and nuclei on a 96-well or 384-well plate before droplet generation, at the cost of higher experiment complexity. inCITE-seq [193] also took an approach similar to FD-seq's, by performing both cross-link reversal and reverse transcription inside the droplets after 10x droplet encapsulation. However, inCITE-seq was specifically optimized for single nuclei samples, and therefore cytoplasmic antibody staining signal and mature mRNAs are not detectable. In addition, inCITE-seq cannot use proteinase K like FD-seq, because with the 10x platform, reverse transcription is performed inside the droplets, and therefore the reverse transcriptase would have been digested had the proteinase K been added to the lysis buffer.

We applied FD-seq to study the process of KSHV lytic reactivation at the single-cell level. We found that reactivation is a very heterogeneous process, and that the expression levels of viral genes correlated poorly with one another. Additionally, we found the expression level of four host factors, namely *CDH1, CORO1C, ISCU* and *TMEM119*, to be significantly positively correlated with viral reactivation. Using live-cell imaging and time-course study, we found *TMEM119* to have the strongest effect among these four genes in enhancing the degree of viral reactivation. Further studies are required to determine the mechanistic details behind the effects of *TMEM119* on KSHV reactivation.

Furthermore, we used FD-seq to study the host response of human lung cells to OC43 infection, which is a betacoronavirus, and the subsequent expression of viral genes at the single-cell level. First, we infected the cells with the virus, then fixed them with PFA, thereby inactivating the virus for safer sample handling, before analyzing the cells with FD-seq. We showed that most OC43-infected cells were unable to support a high level of viral gene expression, and instead upregulated pro-inflammatory signaling pathways. Because the cells were analyzed 4 days post-infection, it is not possible to determine whether these cells, which expressed a low level of viral genes, were abortively infected cells, or if they were infected through secondary infection. The investigation of the cellular mechanisms that lead

to this heterogeneous infection outcome, and the biological functions of the upregulated pro-inflammatory genes, require further studies.

Taken together, these applications of FD-seq show that it is a valuable tool in the field of virology, and any other biological applications that require PFA fixation. Moreover, FD-seq can also be used for integrating protein activity with transcriptome information. For example, transcription factor expression levels and phosphorylation can be integrated with whole transcriptome analysis by combining intracellular protein staining with FD-seq. Furthermore, FD-seq is compatible with RNA velocity analysis, which relies on the detection of unspliced mRNA molecules [207] (Figure 4.17). Finally, we anticipate that FD-seq could serve as a basis for the development of methods that allow sequencing of formalin-fixed paraffin-embedded (FFPE) tissues, due to the similarity between PFA fixation and FFPE, thereby enabling high-throughput single-cell sequencing of readily available samples in tissue banks.

4.5 Supplementary Figures

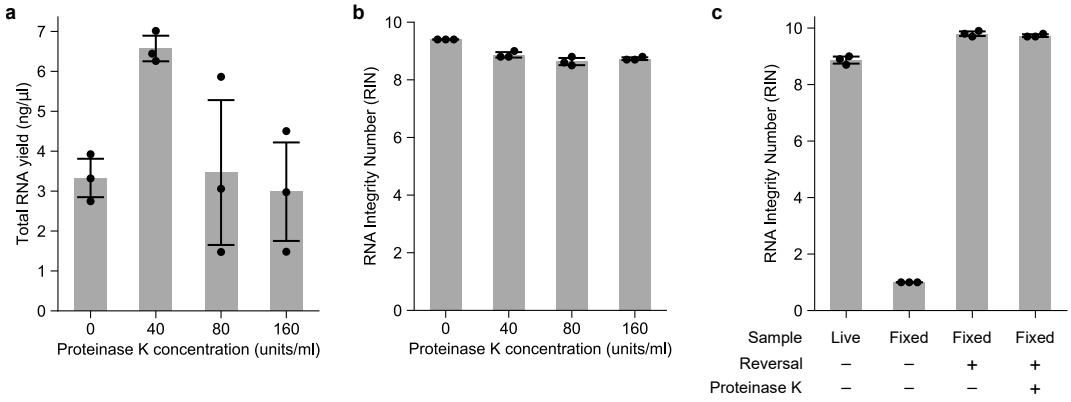

Figure 4.6: Optimization of total RNA extraction from bulk fixed cells. (a, b) Total RNA yield (a) and RNA integrity number (RIN) (b) at different proteinase K concentrations. (c) RINs of RNA extracted from fresh live cells, of RNA extracted from fixed cells without heat reversal and without using proteinase K, and of RNA extracted from fixed cells after heat reversal, and with and without using proteinase K. Data is presented as mean standard deviation. n = 3 technical replicates for all samples.

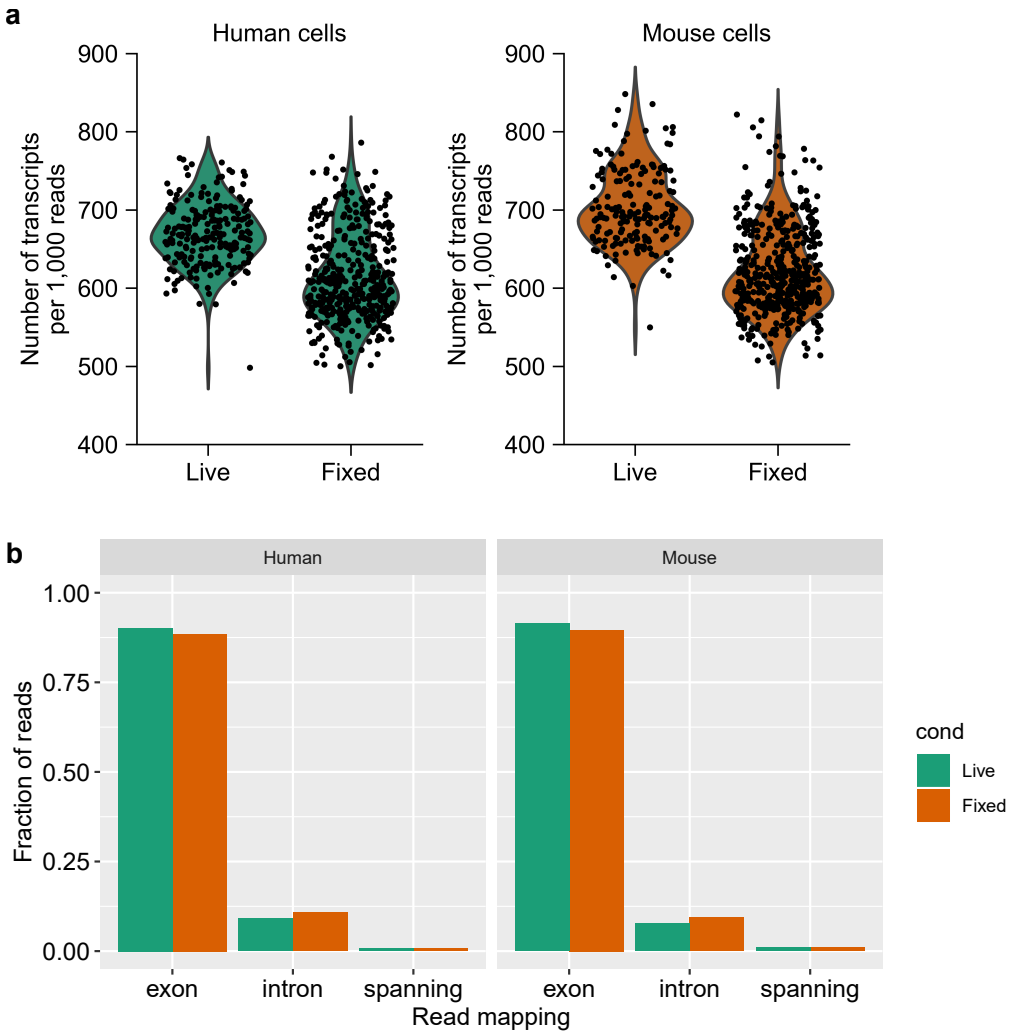

Figure 4.7: Comparison between number of transcripts discovered and exon/intron mapped reads in fresh live and fixed cells. (a) The number of transcripts per 1,000 reads detected. (b) The fraction of reads mapped to exon, intron, or exon/intron spanning region. The data is from the same species-mixing experiment as Figure 4.1.

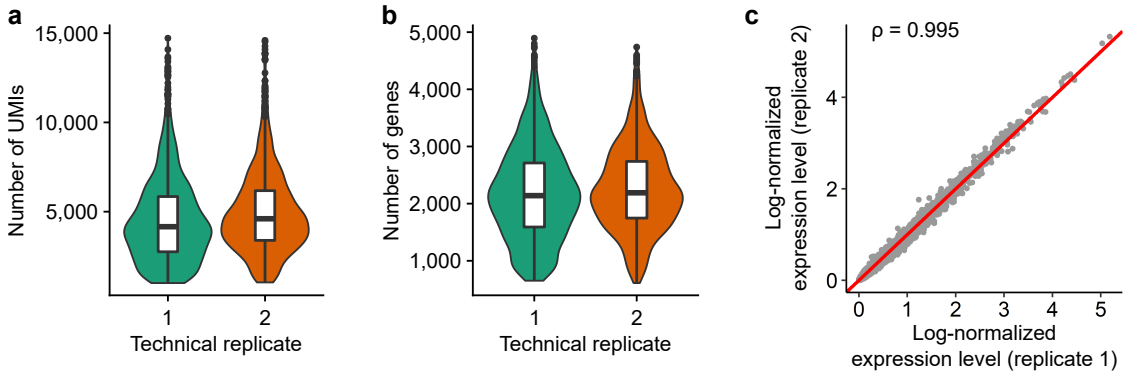

Figure 4.8: Technical replication of FD-seq. (a, b) Violin plots showing the distribution of the number of (a) detected UMIs and (b) detected genes in the two technical replicates. The middle line inside the box indicates the median, the upper and lower edges of the box indicate the first and third quartiles, and the whiskers extend to 1.5× the interquartile range beyond first and third quartiles. (c) Scatter plot showing the log-normalized expression level of two technical replicates. Each dot represents the average normalized expression level, and the red line indicates y = x. The plot also shows the Pearson's correlation coefficient ρ of the log-normalized expression levels between the two replicates. In (a–c), n = 1,483 and 1,338 single cells for replicates 1 and 2, respectively.

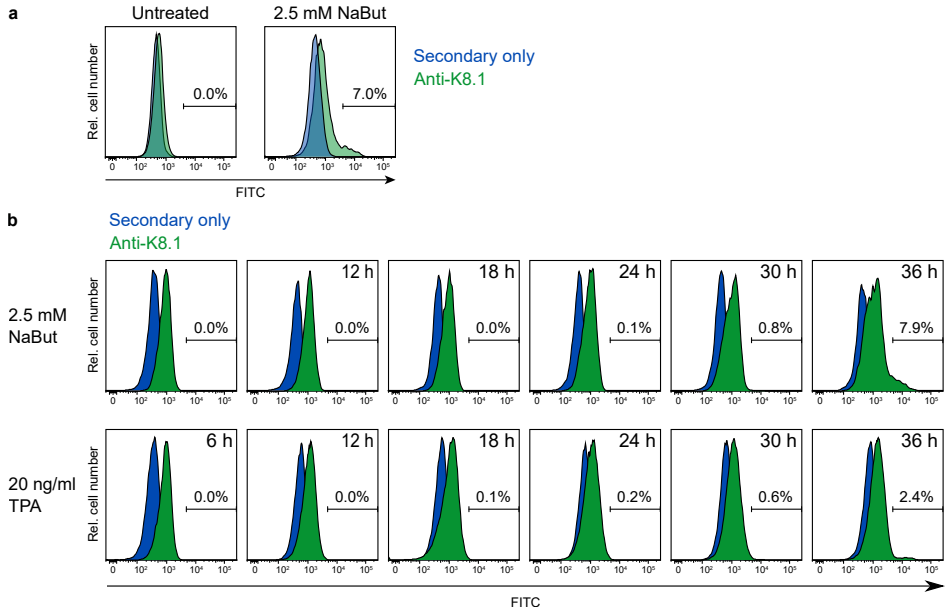

Figure 4.9: Optimization of K8.1 antibody staining and induction of reactivation. (a) Frequency of spontaneous and induced reactivation in BC3 cells. These cells were treated with 2.5 mM of NaBut for 48 hours. (b) Time course of reactivation induced by NaBut (top) or TPA (bottom) treatment.

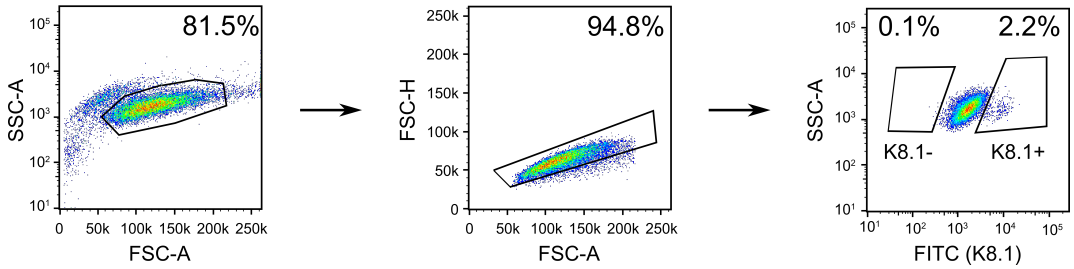

Figure 4.10: Gating strategy for K8.1-positive and K8.1-negative subpopulations. This gating strategy was used for FD-seq processing of reactivated and non-reactivated BC3 cells.

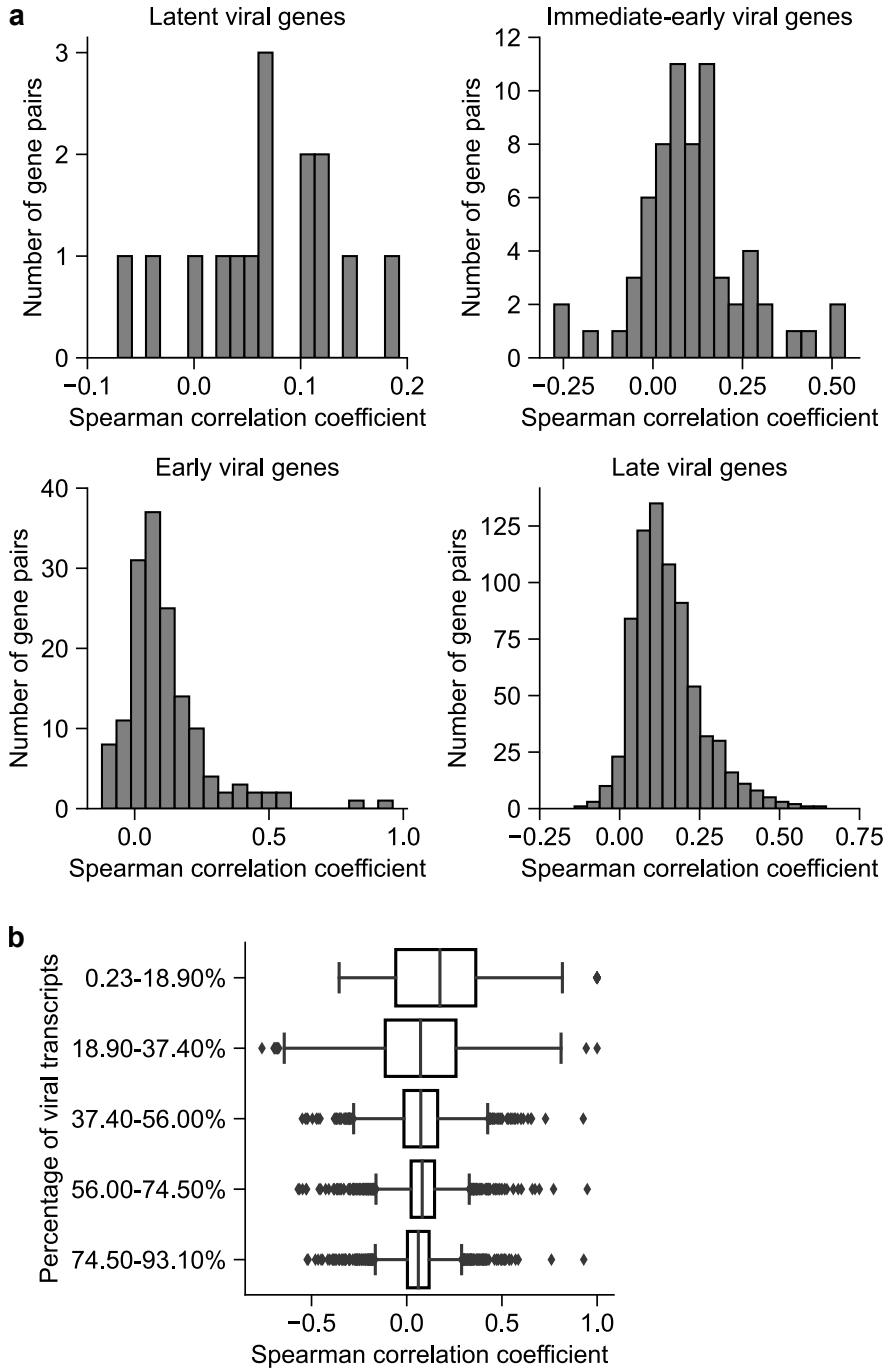

Figure 4.11: Pairwise correlation of viral genes by timing and viral transcript abundance. (a) Histograms showing the pairwise Spearman correlation coefficients of latent, immediate-early, early and late viral genes. (b) Box plot showing the pairwise Spearman correlation coefficient between viral transcripts binned by the percentage of viral transcripts. The middle line indicates the median of the data, the box edges indicate the first and third quartile, the whiskers indicate 1.5× the interquartile range beyond the first or third quartile, and the dots indicate the outliers. In (b), the bins have n = 1891, 3321, 3570, 3655 and 3655 coefficients, in increasing order of the bins' values.

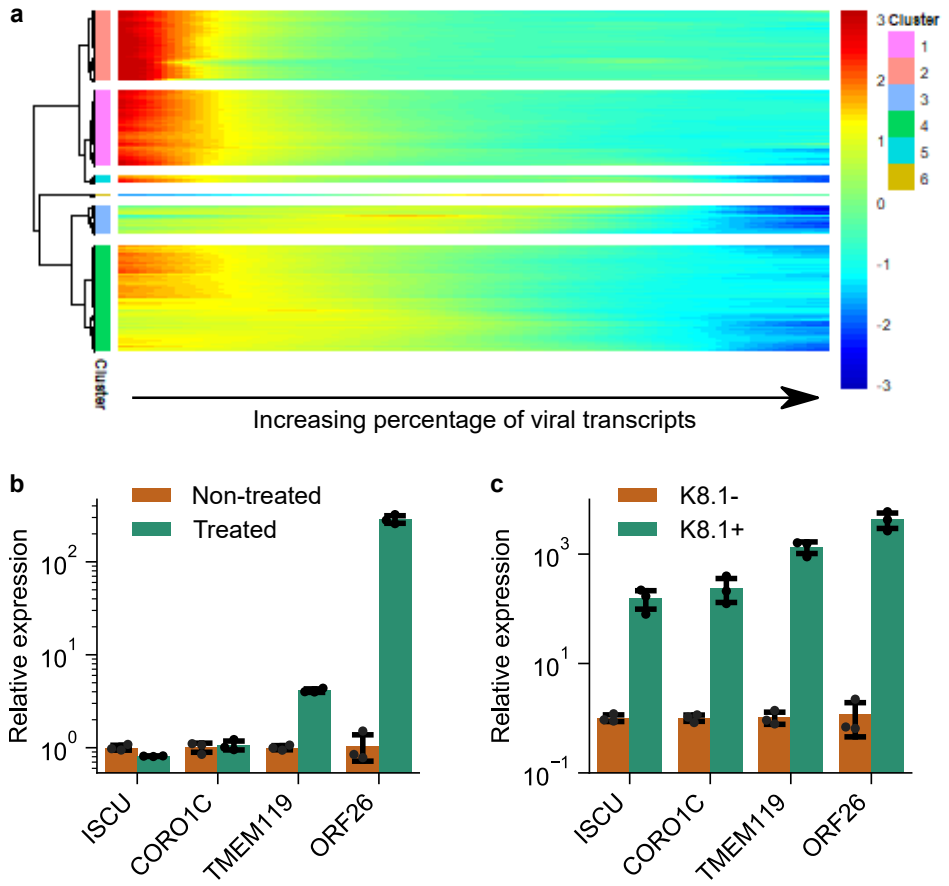

Figure 4.12: Expression of host genes as a function of the abundance of viral transcripts. (a) Heatmap of relative expression level of differentially expressed host genes. Each row shows the relative expression level of a host gene. The genes are clustered based on their expression profile, with only cluster 6 showing a positive correlation with the percentage of viral transcripts. (b, c) qPCR validation of the upregulated host genes in reactivated BC3 cells. The cells were treated or not treated with TPA (b), or the treated cells were sorted into K8.1-negative and K8.1-positive subpopulations (c). *ORF26* is a viral transcript, and serves as a positive control. (b, c) Data is presented as mean ± standard deviation. n = 3 biological replicates.

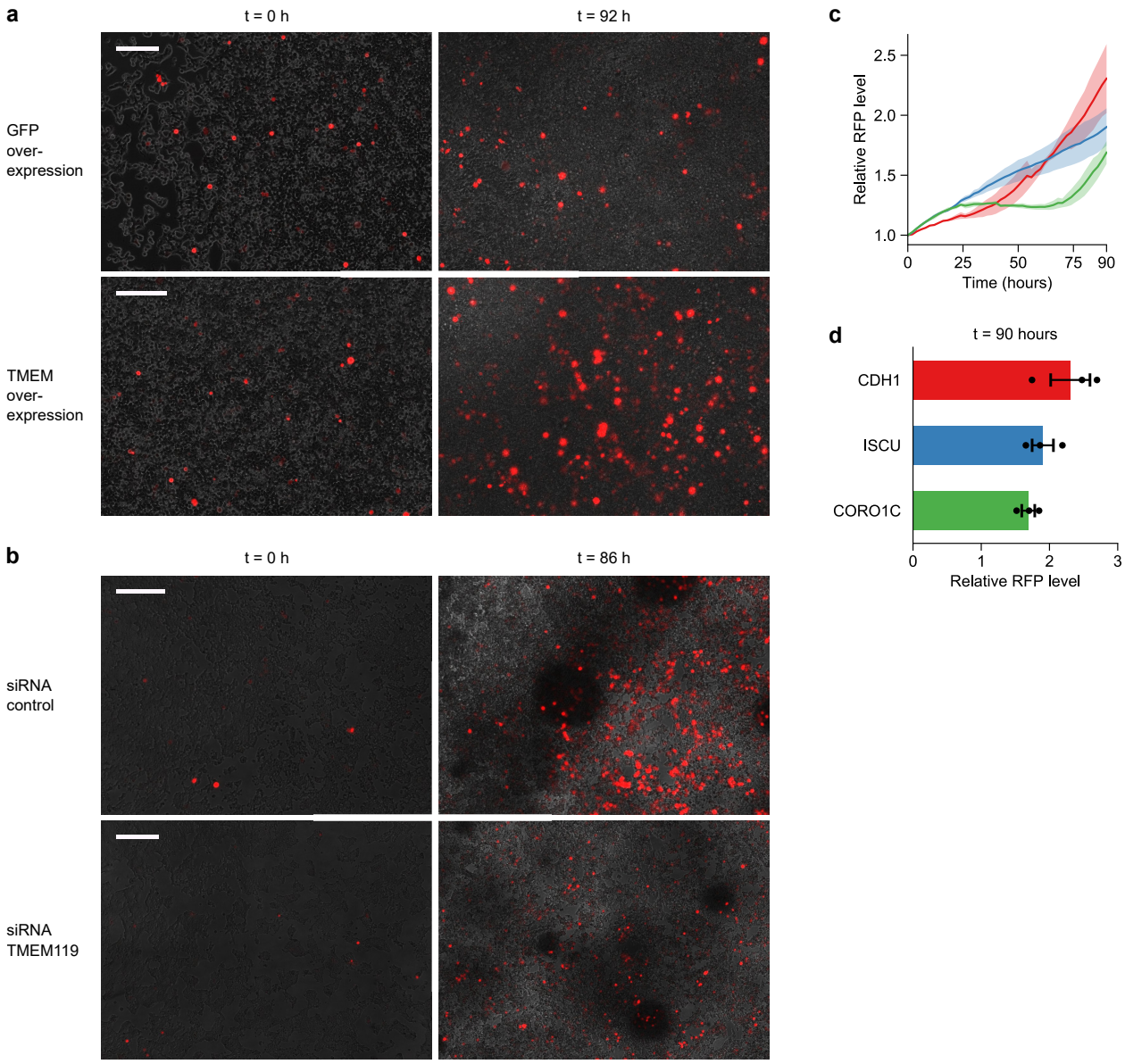

Figure 4.13: Live-cell imaging experiment of KSHV reactivation. (a) Representative images of RFP level, which indicates KSHV reactivation level, in TPA-treated HEK293T.rKSHV219 cells transfected with GFP control (top) or *TMEM119* (bottom) at 0 and 92 hours. (b) Representative images of RFP level in HEK293T.rKSHV219 cells transfected with control siRNA (top) or *TMEM119* siRNA (bottom) at 0 and 86 hours. The scale bars in (a) and (b) represent approximately 100 μm. (c) Time course of RFP level in HEK293T.rKSHV219 cells transfected with *CDH1, ISCU* or *CORO1C* genes. (d) Endpoint RFP level at t = 90 h of the time course data in (c). The ribbons in (c) and the error bars in (d) indicate s.e.m. In (a–d), n = 3 biological replicates. All live imaging experiment replicates show similar results as the representative images in (a, b).

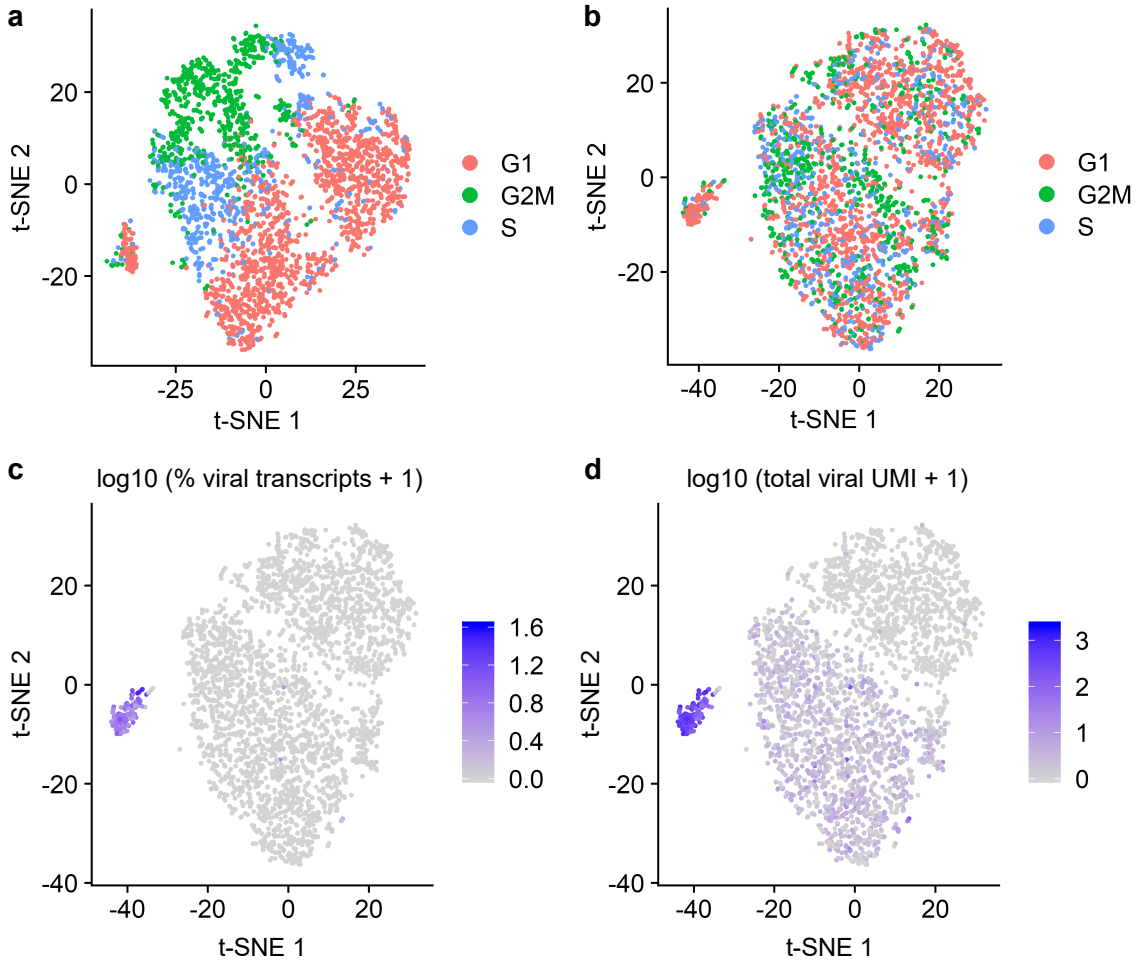

Figure 4.14: Cell cycle effects and expression of viral transcripts in OC43-infected A549 cells. (a, b) t-SNE plots of cells (a) before and (b) after removing cell cycle effects. The colors indicate the cell cycle scores. (c, d) t-SNE plots showing (c) the percentage of viral transcripts and (d) the total UMI of viral transcripts of each single cell.

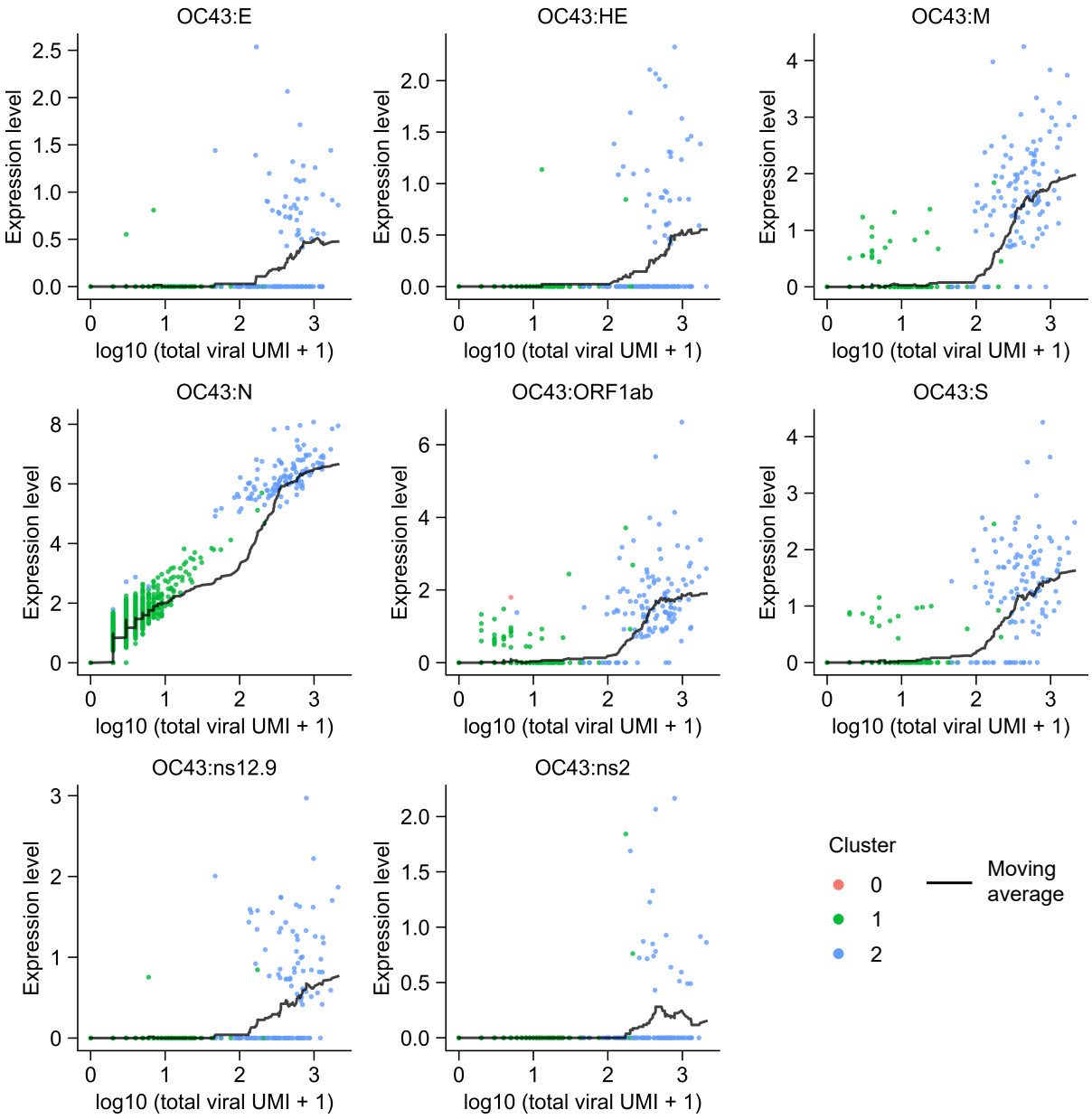

Figure 4.15: Expression profiles of all OC43 viral genes in MOI 1 infected cells. Scatter plots showing the expression level of each gene against the total abundance of viral genes. The black lines show the 50-cell moving average of the normalized expression level.

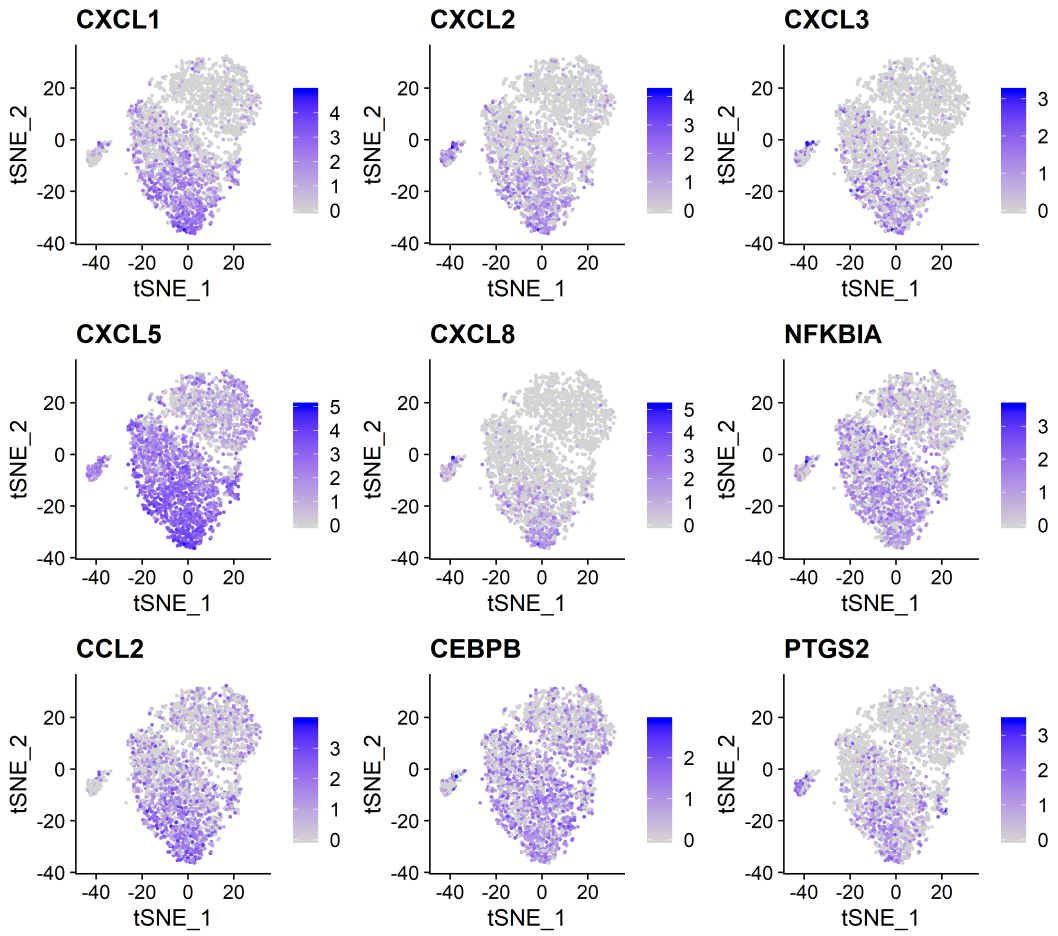

Figure 4.16: Expression level of 9 representative immune-related genes in OC43-infected cells. t-SNE plots showing the log-normalized expression level of the 9 genes.

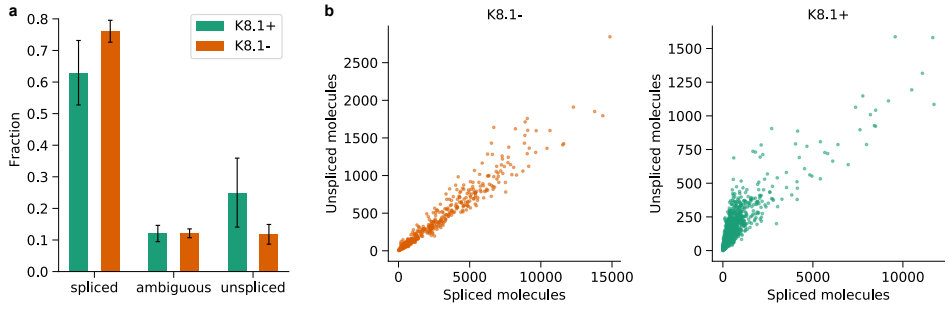

Figure 4.17: Detection of unspliced host mRNAs from fixed BC3 cells. (a) Fractions of spliced, spliced and unassigned mRNA molecules. Data is presented as mean ± standard deviation. (b) The number of spliced and unspliced molecules in non-reactivated (left) and reactivated cells (right). In (a, b), n = 1,035 and n = 286 single cells for K8.1+ and K8.1− sample, respectively.

4.6 Methods

Microfluidic Device Fabrication

The microfluidic device was fabricated based on the AutoCAD file provided in Macosko et al. [54]. The mold was made from SU8-3050 photoresist of approximately 110 μm in height. Then, polydimethylsiloxane (PDMS, Momentive RTV615) was cast on the mold to form the microfluidic devices. Plasma bonding was used to bond the PDMS device to a glass slide, and the microfluidic channels were coated with Aquapel.

Cell Culture

HEK293T.rKSHV219 cells were maintained in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% (v/v) fetal bovine serum (FBS), 2 mM GlutaMAX (Gibco), 1% (v/v) penicillin–streptomycin (P/S, Gibco), and 1 μg/mL puromycin. HEK293T.rKSHV219 cells were generated by infecting HEK293T cells with rKSHV.219 [202], and selecting with 1 μg/mL puromycin. The KSHV-positive human PEL cell line BC3 (ATCC) was cultured in Roswell Park Memorial Institute (RPMI) supplemented with 20% (v/v) FBS, 2 mM GlutaMAX, and 1% (v/v) P/S. A549 cells (ATCC) or A549 cells with H2B-Ruby fusion21 were maintained in DMEM supplemented with 10% FBS, without antibiotics. The 3T3 cells ($p65^{-/-}$ 3T3 mouse embryonic fibroblast cells expressing p65-DsRed and H2B-GFP nucleus marker [3]) were cultured in DMEM supplemented with 10% (v/v) fetal bovine calf serum (HyClone), 1% (v/v) GlutaMAX, and 1× P/S.

Optimization of RNA Extraction Condition for Fixed and Permeabilized cells

BC3 cells were harvested, centrifuged at 300–400 × g for 3 min to remove the cell media, and washed once with 1 mL of 1% BSA (molecular biology grade, Gemini Bio-Products)

in PBS. Cells were fixed by adding 1 mL of 4% PFA in PBS (Santa Cruz Biotechnology) and incubated at room temperature for 15 min. Next, cells were centrifuged at 400 × g for 3 min, the paraformaldehyde was discarded, and cells were washed once with 1 mL of wash buffer (PBS with 1% BSA and 40 U/mL of RNase inhibitor, murine (NEB)). Cells were permeabilized by adding 500 µL of 0.1% Triton X-100 (molecular biology grade, Acros Organics) diluted in the wash buffer, and incubated at room temperature for 15 min. Then, 1 mL of wash buffer was added directly to the cells, cells were pelleted, the supernatant was discarded, and the cells were washed once more with 1 mL of wash buffer. To mimic the condition of antibody staining during optimization experiments, cells were resuspended in wash buffer and incubated on ice for 1 h, washed twice with the wash buffer, and finally kept on ice in the wash buffer before RNA extraction.

To serve as a positive control, RNA from bulk live cells was extracted using the RNeasy Plus Mini Kit (Qiagen). Cells were first suspended in a combination of the standard Drop-seq lysis buffer with PBS at 1-to-1 volume ratio (to make the lysis buffer's final concentration the same as that in the droplets), incubated at room temperature for 10 min, and placed on ice for at least 5 min. Next, the cell lysate was combined with 350 µL of RLT plus buffer from the RNeasy kit, and processed according to the manufacturer's protocol.

To extract RNA from bulk fixed cells, fixed and permeabilized cells (as described above) were suspended in a 1-to-1 mixture of PBS and Drop-seq lysis buffer, the latter of which was spiked in with various concentrations of proteinase K (NEB). Next, the sample was at 56°C for 1 h, left at room temperature for 10 min, and placed on ice for at least 5 min. Finally, the cell lysate was combined with 350 µL of RLT plus buffer, and processed according to the RNeasy Plus Mini Kit's protocol. The extracted RNA is then quantified using BioAnalyzer.

Treatment of BC3 Cells, Flow Cytometry, and FACS-Sorting of Reactivated Cells

BC3 cells were mock-treated or treated with a final concentration of 20 ng/mL tetradecanoyl phorbol acetate (TPA, Millipore Sigma) for 36 h to induce KSHV reactivation. Cells were pelleted by centrifugation at 300 × g for 3 min and fixed with 4% PFA in PBS for 15 min at room temperature, followed by permeabilization with 0.1% (v/v) Triton X-100 in PBS with 1% (w/v) BSA for 15 min at room temperature. After washing and blocking the cells in PBS supplemented with 4% BSA, staining was performed with an anti-KSHV K8.1 antibody (sc-65446, Santa Cruz, 1:50 dilution) and an Alexa Fluor 488-conjugated goat-anti-mouse secondary antibody (A10667, Life Technologies, 1:4000 dilution) in PBS with 1% BSA and 40 U/mL RNase inhibitor, murine (NEB). K8.1+ and K8.1− cell subpopulations were analyzed on an LSRFortessa flow cytometer (BD Biosciences) and/or sorted on a FACSAriaIIIu cell sorter (BD Biosciences). FACS Diva and FlowJo software was used to analyze the flow cytometry data.

Infection of A549 Cells

OC43 (ATCC) were grown and titrated on A549 cells. For FD-seq experiments, A549 cells were seeded in 6-well plates and infected with OC43 at an MOI of 1 the following day. Cells were incubated at 33°C for 4 days and were subsequently harvested for FD-seq.

The MOI was calculated as follows. A549 cells were infected with 10-fold serial dilutions of OC43. The virus was allowed to adsorb for 1 h at 33°C, after which the inoculum was removed and the cells were overlaid with DMEM, 2% FCS, and 2% Cellulose. Infected cells were incubated at 33°C for 2 weeks and subsequently fixed and stained with a 1:1 dilution of methanol:crystal violet. Plaques were counted and the virus stock concentration was determined in units of PFU/mL. The MOI reflects the ratio of PFU to the number of infected cells.

FD-seq Protocol

A step-by-step protocol for FD-seq is available on Protocol Exchange [208]. The sequences of the primers used in Drop-seq and FD-seq are listed in Supplementary table 4.1.

The protocol for FD-seq is based on the McCaroll lab's online Drop-seq protocol3. In summary, there are only two significant differences: 40 U/mL proteinase K was added to the Drop-seq lysis, and a heating step was added between the droplet generation and droplet breakage steps. First, the barcoded beads (ChemGenes Corporation, catalog number Macosko-2011-10(V+)) were suspended at 120,000 beads/mL in a modified Drop-seq lysis buffer: 200 mM Tris pH 7.5, 6% Ficoll type 400, 0.2% Sarkosyl, 20 mM EDTA, 50 mM DTT and 40 U/mL proteinase K (NEB). Cells were suspended in PBS with 0.01% BSA at 100,000 cells/mL. For droplet generation, the cells, beads, and oil (Bio-Rad, catalog number 1864005) were injected at flow rates of 3, 3 and 12 mL/h, respectively, using syringe pumps. After droplet generation, the droplets were heated on a heat block at 56°C for 1 h to reverse the PFA cross-links, then incubated at room temperature for 10 min and kept on ice for at least 5 min. After this, the droplets were broken, the beads were collected, washed, and subjected to reverse transcription (using TSO_RNAhybrid primer) and exonuclease I digestion as per standard Drop-seq protocol.

For whole transcriptome amplification (WTA), the beads were distributed into a 96-well plate such that each well contained 5,000 beads. A 50 µL PCR reaction was set up in each well using 1× KAPA HiFi Hotstart Readymix, and 0.8 µM TSO_PCR primer. The following thermal cycle program was used: 95°C 3 min; 4 cycles of 98°C 20 s, 65°C 45 s, 72°C 3 min; 10 cycles of 98°C 20 s, 67°C 20 s, 72°C 3 min; 72°C 5 min; 4°C hold. After PCR, the WTA products were purified with 0.6× Ampure XP beads (Beckman Coulter), pooled and quantified with TapeStation (Agilent) or Qubit (Invitrogen).

To prepare the sequencing library, 450 pg of WTA products were incubated with the tagment buffer and tagment enzyme (Nextera XT DNA Library Prep Kit, Illumina) at 55°C for 5 min. Next, the neutralization buffer was added, and the sample was incubated at room

temperature for 5 min. After this, the Nextera PCR Master Mix and PCR primers P5-TSO_Hybrid and Nextera_N70X were added, and the sample was thermocycled as following: 95°C 30 s; 12 cycles of 95°C 10 s, 55°C 30 s, 72°C 30 s; 72°C 5 min; 4°C hold. Finally, the tagmentation products were purified with 0.6× Ampure XP beads, and quantified with TapeStation, BioAnalyzer, or Fragment Analyzer.

Drop-seq Protocol

Because FD-seq was developed based on Drop-seq, the two protocols were very similar. The only two differences were that the lysis buffer did not include proteinase K, and the droplets were not subjected to the heating step between droplet generation and droplet breaking. For more details on the standard Drop-seq protocol, see Chapter 3.

Species-Mixing Experiment

Species-mixing experiments of fresh live samples were performed using human BC3 and mouse 3T3 cells, combined at 50,000 cells/mL/cell type. The samples were then processed with the standard Drop-seq protocol. For the species-mixing experiment of fixed samples, human BC3 and mouse 3T3 cells were separately fixed with PFA and permeabilized as described above. Then, the two cell types were combined at 50,000 cells/mL/cell type, and processed with the FD-seq protocol.

Technical Replication Experiment

Two identical samples of A549-H2B-Ruby cells were separately fixed with PFA and permeabilized as above, and then processed with FD-seq on the same day.

Methanol Fixation Experiment

A549 cells were harvested and split into two samples of approximately 1 million cells each, one for PFA fixation and FD-seq, and one for methanol fixation and Drop-seq. PFA fixation was performed as above. For methanol fixation [187], cells were washed once with 1 mL of 1% BSA, then resuspended in 200 µL of ice-cold PBS. Next, ice-cold 100% methanol was added dropwise into the cells while gently vortexing the cells, and the cells were incubated on ice for 15 min. To rehydrate the cells, they were centrifuged at 1,000 × g for 5 min, methanol was discarded, and cells were washed twice with 1 mL of 0.01% BSA in PBS. The rehydrated cells were then counted, resuspended at 100,000 cells/mL in 0.01% BSA in PBS, and processed according to the standard Drop-seq protocol.

Plasmids and Transfection

Expression plasmids encoding human *CDH1*, *TMEM119*, *ISCU* and *CORO1C*, as well as KSHV ORF50 and nls-eGFP under control of the human EF1α promoter in the pLV backbone were ordered from VectorBuilder. All sequences were verified by Sanger sequencing. HEK293T.rKSHV219 cells were reverse transfected with 0.8 µg plasmid DNA using Lipofectamine2000 (Life Technologies) following the manufacturer's instructions, and seeded in 12- or 24-well plates for analysis by live-cell imaging or RT-qPCR as indicated. For silencing experiments, HEK293T.rKSHV219 cells were reverse transfected with 40 nM siRNA targeting *TMEM119* (Dharmacon siGenome SMARTpool M-018636-01-0005), *CDH1* (M-003877-02-0005), *CORO1C* (M-017331-00-0005), *ISCU* (M-012837-03-0005), or non-targeting control (D-001206-14-05) in 12- or 24-well plates using Lipofectamine RNAiMAX reagent (Life Technologies) following the manufacturer's protocol. After 48 h, the cells were treated with 2 ng/mL TPA as indicated, and KSHV reactivation efficiency was assessed by live-cell imaging or RT-qPCR.

Reverse Transcription and Real-Time PCR (RT-qPCR)

Total RNA was extracted by using the HP Total RNA Kit (OMEGA Bio-Tek) using the manufacturer's instructions. RT-qPCR was performed using either a one-step or a two-step protocol. For the one-step protocol, RT-qPCR was performed with equal amounts (25–500 ng) of RNA using the SuperScript III Platinum One-Step qRT-PCR kit with ROX (Thermo Fisher Scientific) on a 7500 Fast Real-Time PCR Machine (Applied Biosystems). Premixed master mixes containing TaqMan primers and probes for each individual gene were purchased from IDT (*GAPDH, TMEM119, CORO1C, ISCU* and *CDH1*) or Applied Biosystems (18S). For the two-step protocol, reverse transcription was performed using the ProtoScript II First Strand cDNA Synthesis Kit (NEB) according to the manufacturer's protocol. The kit's oligo-dT primer was used as the primer in this step. Next, the RT products were diluted by 20-fold in water, and used as the input for qPCR (Luna Universal qPCR Master Mix, NEB). The primers' sequences are given in Supplementary table 4.2. The relative expression level of each target gene was calculated by normalizing for *GAPDH* or 18S levels using the Comparative Ct Method ($\Delta\Delta$Ct Method), and presented relative to the control sample.

Live-Cell Imaging Experiments

HEK293T.rKSHV219 cells were seeded in 24-well plates and transfected with plasmids for overexpression or siRNA for knock-down experiments. Cells were then imaged on a Nikon Ti-Eclipse microscope containing an environmental chamber (37°C, 100% humidity, 5% CO2). Images were acquired every 4 h for 4 days.

Sequencing and Alignment

Drop-seq and FD-seq libraries were sequenced on a NextSeq 550 machine with the following read distribution: 20 bp for read 1, 60 bp for read 2, and 8 bp for index 1 read. The custom read 1 sequencing primer Read1CustomSeqB was used in place of Illumina's read

1 primer (Supplementary table 4.1). Alignment was performed using Picard version 2.21.8 (https://github.com/ broadinstitute/picard), Drop-seq tools version 1.13 or version 2.3, and STAR aligner version 2.6.1b [209]. The valid cell barcodes were chosen using the knee plot method. The species-mixing sequencing results were aligned to the combined human-mouse reference genomes (accession number GSE63269). Alignment was performed to a concatenated version of the human genome with KSHV genome GQ994935.1 [210], or the human genome GRCh38 and OC43 genome NC_006213.1 [203].

Read Subsampling and Read Mapping

In the FD-seq and methanol fixation comparison experiment, we used samtools view command [211] to subsample the output BAM file from Drop-seq tools version 2.3 DetectBeadSynthesisErrors command. To calculate the proportion of reads mapped to different genomic regions, we used Picard CollectRnaSeqMetrics command on the output BAM file directly from the STAR aligner. The refFlat file for the GRCh38 reference genome was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/.

scRNA-seq Data Analysis

Data analysis was done using Python 3, RStudio, and R packages Monocle 2 [8, 212, 213] and Seurat 3 [214, 215]. Data visualization was done using R and Python.

For species-mixing data, a cell is classified as human or mouse if more than 90% of the detected transcripts are mapped to human or mouse, respectively. Otherwise, the cell is considered mixed. dropEst pipeline (v0.8.5) [216] was used with argument -L eiEIBA to annotate reads mapping to exon, intron, or exon/intron spanning regions. To compare the mean expression level between live and fixed samples of the same species, we normalized the single-cell UMI counts using Monocle's estimateSizeFactors function, then calculated the average of each gene in each sample. Any average value below 0.1 is set to 0.1 for plotting.

For the technical replication experiment, we first removed cells with fewer than 1,000

UMIs or more than 15,000 UMIs, and discarded genes detected in fewer than 5 cells. We then imported the count data into Seurat 3, normalized, and scaled the data with the default settings. To compare the gene expression level between the two technical replicates, we calculated the average normalized expression using Seurat's AverageExpression function, then plot the natural logarithm of the average expression with one added pseudocount.

For the FD-seq and methanol fixation comparison experiment, we first discarded cells with lower than 1,000 UMIs and genes detected in fewer than 5 cells. We then used Seurat 3 to log-normalize and scale the data. To compare the two fixation methods, we performed similar calculations as for the technical replication experiment above. To calculate the technical variance [61], we normalized the UMI count by converting it to UMIs per million, then randomly chose 400 cells from each fixation method and used the 1,000 most highly expressed non-mitochondrial genes to calculate the mean and squared coefficient of variation.

For the KSHV reactivation experiment, we first discarded cells with more than 3,000 genes detected, and discarded cells with fewer than 1,000 UMIs or more than 10,000 UMIs. Then we discarded genes detected in fewer than 5 cells. To find the relative expression of each gene, we normalized the transcript counts using estimateSizeFactors and estimateDispersions functions in Monocle 2. To cluster the cells, we first performed dimension reduction with t-SNE using the first 4 principal components, then clustered the cells by setting rho_threshold = 17 and delta_threshold = 11. We then set the percentage of viral transcripts as the pseudotime parameter, and performed a pseudotime analysis in order to find genes correlated with the abundance of viral transcripts.

For the OC43 infection experiment, we discarded cells with fewer than 1,000 UMIs or more than 20,000 UMIs, and discarded genes detected in fewer than 5 cells. We then used Seurat 3 to remove the cell cycle effects (using CellCycleScoring function), log-normalize and scale the data. The log-normalized counts are used to show gene expression level. To cluster the cells, we used the first 10 principal components for the FindNeighbors and RunTSNE functions, and resolution = 0.2 for the FindClusters function. To find the clus-

ter markers, we used min.pct = 0.25 and logfc.threshold = 0.25 settings for the FindAllMarkers function. To find enriched KEGG pathways in each cluster, we used g:Profiler (https://biit.cs.ut.ee/gprofiler/gost) [206] with the default settings on the list of genes upregulated in each cluster as found by Seurat's FindAllMarkers function.

RNA Velocity Analysis

The output files of Drop-seq tools DetectBeadSynthesisErrors function were processed with the dropEst pipeline (v0.8.5) [216] to tag spliced and unspliced transcripts, and the results were analyzed with the Python velocyto package (v0.17.17) [207].

Data Availability

The raw sequencing data and gene count data are deposited in NBCI's Gene Expression Omnibus (accession number GSE156988). The codes used for sequencing alignment and data analysis are available at `https://github.com/tay-lab/FD-seq`.

4.7 Supplementary Tables

Table 4.1: List of primers used in FD-seq.

Name	Sequence
TSO_RNAhybrid	AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG
TSO_PCR	AAGCAGTGGTATCAACGCAGAGT
P5-TSO_Hybrid	AATGATACGGCGACCACCGAGATCTACACGCCTG TCCGCGGAAGCAGTGGTATCAACGCAGAGT*A*C
Nextera_N701	CAAGCAGAAGACGGCATACGAGA TTCGCCTTAGTCTCGTGGGCTCGG
Nextera_N702	CAAGCAGAAGACGGCATACGAGA TCTAGTACGGTCTCGTGGGCTCGG
Nextera_N703	CAAGCAGAAGACGGCATACGAGA TTTCTGCCTGTCTCGTGGGCTCGG
Nextera_N704	CAAGCAGAAGACGGCATACGAGA TGCTCAGGAGTCTCGTGGGCTCGG
Read1CustomSeqB	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC

Table 4.2: List of primers used in RT-qPCR. The ORF genes are KSHV viral genes.

Gene	Forward primer	Reverse primer	Probe
GAPDH	GAACATCATCC CTGCCTCTACTG	CAGTGAGCTT CCCGTTCAGC	
ISCU	CTGCACTG CTCCATGCT	CTCATTTCTTC TCTGCCTCTCC	
CORO1C	GTCCACTACCT CAACACATTCA	TGAAGAATCTG GCAATCTCACA	
TMEM119	CCTGGCGTGA AGCAGTATTT	GCACAGGCAG AATGACACTAA	
ORF26	GCTAGCAGTG CTACCCCCATT	GGTCAAATCC GTTGGATTCG	
ORF50	CACAAAAATG GCGCAAGATGA	TGGTAGAGTTG GGCCTTCAGTT	AGAAGCTTC GGCGGTCCTG
ORF57	TGGACATTATG AAGGGCATCCTA	CGGGTTCGG ACAATTGCT	TGACGAATCGA GGGACGACGAGA
ORF74	GTTCCCCTGA TATACTCCTGC	GGACATGAAA GACTGCCTGAG	AGGATGTACGGT CTCTTCCAAAGCC

CHAPTER 5

COMPUTATIONAL FRAMEWORKS FOR PREDICTING PROTEIN INTERACTIONS VIA SINGLE-CELL PROXIMITY SEQUENCING

In this chapter, we introduce a simulation model to characterize the effects of background signal on Prox-seq data, and compare two independent methods for prediction of protein complexes from both simulated and experimental Prox-seq data. Part of this chapter is reproduced from a manuscript that is under preparation for journal submission.

5.1 Summary

Proximity sequencing (Prox-seq) is a method that can simultaneously measure gene expression, protein and protein complex in single cells. By making use of proximity information from proteins, Prox-seq can infer protein complex information on the surface of a cell. However, proximity does not necessarily correspond to direct interaction between two proteins, and is significantly influenced by the abundance of the proteins of interest. In Chapter 3, we proposed a statistical method for predicting protein complexes from Prox-seq data. Unfortunately, the method was difficult to fully validate, because there are no comparable alternatives to Prox-seq that could be used to benchmark the method's predictions. Here we introduce computational frameworks to assist in investigation of protein-protein interactions from Prox-seq data. First, we propose a physical model to simulate Prox-seq data, with the protein interactions and the protein abundance as the model's inputs. The simulated data are then used to validate the statistical method that we proposed in Chapter 3, and to understand how the background is produced from non-specific protein proximity, and how it influences protein complex prediction. Second, we propose a modification to the Prox-seq protocol, called the free oligo modification, and use the new measurements enabled by the modification to propose a second method for predicting protein complexes. We then compare

the two protein complex prediction methods on experimental data, and demonstrate that they generally agree with each other. Thus, our computational frameworks provide a simple way to investigate the behavior of Prox-seq data under various protein interaction scenarios, and independent methods for predicting and quantifying protein complexes in single cells.

5.2 Introduction

One potential source of background in Prox-seq was the ligation of two Prox-seq probes that happen to be in proximity with each other by random chance [124]. This non-specific interaction reflected the chance encounter between two proteins that were not functionally interacting with each other. We called this effect the random ligation of non-interacting proteins/probes, or random ligation in short. As a result, the presence of PLA products for a specific pair of proteins did not necessarily imply that the two proteins interact. In Chapter 3, we proposed a statistical method, called the iterative method henceforth, to predict the true protein complexes from PLA product count data. However, it was not possible to experimentally validate the predicted protein complex abundance, because of a lack of alternative assays for protein complex quantification in single cells.

A simpler way to validate the protein complex predictions was to use positive and negative protein complex control. In the experiments in Chapter 3, CD3:CD3 and CD28:CD28 protein complexes were the positive controls, and were indeed predicted to be protein complexes by the algorithm.

In contrast, negative controls were much more difficult to implement. Consider the PLA assay with rolling circle amplification readout (PLA-RCA)[217]. In a PLA-RCA experiment, first the interacting proteins were detected by primary antibodies. Then, secondary antibodies were used to detect the primary antibodies, and the DNA oligos on the secondary antibodies were amplified with RCA to produce a fluorescence signal. The usual negative control for PLA-RCA was to not include any primary antibodies [217–219]. If the protein interaction of interest produced higher signal than the no primary antibody control, then the

interaction is considered to be positive. However, such a control was not suitable for when the amount of random ligation is high. This is because when no primary antibodies were used, the signal produced by the secondary antibodies would intrinsically be low, because secondary antibodies often had low non-specific binding. Consequently, the no primary antibody control would always produce an artificially lower signal than that of the protein interaction of interest, regardless of the existence of protein interaction.

A better negative control would be to use two non-interacting proteins with similar expression levels to the two interacting proteins that we were trying to detect. This is the main reason why negative controls for Prox-seq were difficult to implement: it was already challenging enough to find two proteins that we were certain to not interact, let alone finding two that were of equal abundance to the proteins of interest.

In this chapter, we presented computational frameworks for denoising Prox-seq data. First, we proposed a physical model for simulating the count data of PLA products. The simulation model allowed us to quantitatively analyze random ligation and its effects on the measured PLA product counts. Second, we proposed the free oligo modification to Prox-seq, and leveraged it to develop an additional independent method, called the linear regression (LR) method, for protein complex prediction. We then compared the LR method and the iterative method proposed in Chapter 3 on both simulated and experimental data. We found that both methods largely agreed with each other, and the new LR method was better at quantifying the found protein complexes in some scenarios. Thus, our computational frameworks suggested that, for extracellular protein targets, we could reliably detect and quantify protein complexes at the single-cell level from Prox-seq data.

5.3 Results

5.3.1 Terminology

In this chapter, we used the term non-interacting protein (or probe) to indicate the protein molecule (or probe) that did not interact with any other targeted proteins, PLA product count to indicate the UMI count of a PLA product (which, together with gene count, constituted the output of Prox-seq, Figure 3.1c), protein count to indicate the UMI count of a protein that is calculated from the PLA product count, unligated probe count to indicate the UMI count of probes that bind to non-interacting protein molecules, and protein complex count to indicate the UMI count of a protein complex predicted from the PLA product count.

5.3.2 Free Oligo Modification

Because a PLA product was only produced upon a proximal dual-binding event of two Prox-seq probes (Figure 5.1a), we expected that there were Prox-seq probes that remain unligated due to a lack of suitable proximal probes. This could be because these probes bound to monomeric protein molecules, or to protein molecules that were interacting with another protein that was not targeted by the Prox-seq probe panel. We reasoned that the amount of unligated probes should be a good estimate for the amount of non-interacting probes. Therefore, the count of unligated probes could be used to estimate the amount of background from random ligation.

To enable measurement of these unligated probes, we modified Prox-seq by adding free DNA oligonucleotides (oligos) to the ligation reaction after the probe ligation step (see "Methods"), so that after all proximal Prox-seq probe pairs had been ligated, the remaining unligated probes would be ligated to the added free oligos (Figure 5.1b).

We found that the unligated probe count in Jurkat and Raji cells matched the expected expression of the proteins in each cell type (Supplementary Figure 5.6). The unligated probes

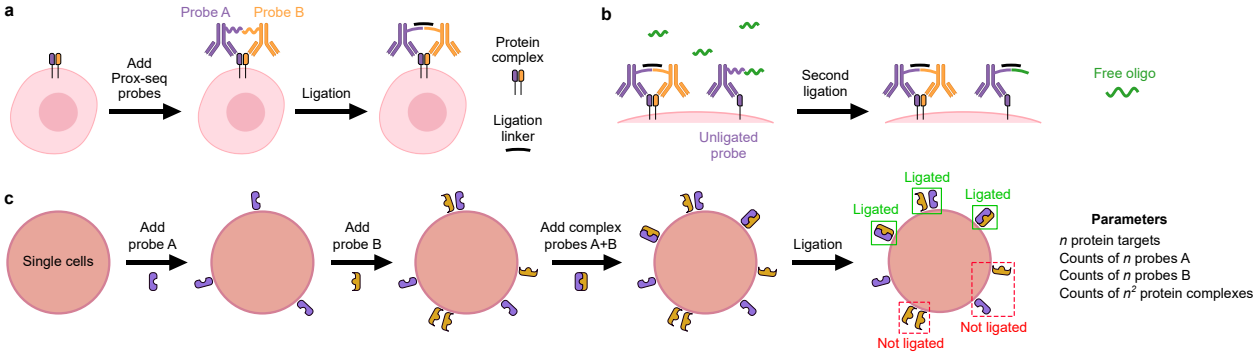

Figure 5.1: Schematics of proximity ligation assay, the free oligo modification, and the simulation model. (a) In Prox-seq, cells were incubated with DNA-conjugated antibodies, called Prox-seq probes. Each protein target could be bound by either Prox-seq probe A or Prox-seq probe B. After the probes bound to their protein targets, ligation occurred between probes A and B that were sufficiently closed, with the help of a DNA oligo ligation linker. Finally, the ligated DNA products could be quantified with scRNA-seq methods. (b) Measurement of unligated Prox-seq probes. After the ligation step in the standard Prox-seq protocol, free oligos were added so that they can be ligated to free Prox-seq probes. (c) Schematic for the simulation model of PLA products. The simulation was separately performed on a cell-by-cell basis. First, a number of non-interacting probes A and non-interacting probes B were added as random points on a sphere surface. Next, a number of protein complexes were added in the same manner. Finally, probes A and B that had a Euclidean distance lower than the ligation distance were ligated, thus creating PLA products.

of Jurkat markers are at least 1 order of magnitude higher than those of Raji markers on Jurkat cells, and vice versa. At most, the total unligated probe count (i.e., sum of unligated probe A and unligated probed B) was approximately 300 UMIs per single cell (Supplementary Figure 5.6). Unligated PD1 probe was the most abundant in Jurkat cells, while unligated ICAM1 and PDL1 probes were the most abundant in Raji cells.

5.3.3 Overview of the Simulation Model

Based on a physical model of how PLA products were formed in each single cell, we created a simulation model of PLA product count data. We reasoned that if a Prox-seq probe A and a Prox-seq probe B were sufficiently close to each other, they would be ligated and produce a PLA product, regardless of whether the protein molecules to which the probes bound functionally interacted with each other. As a result, the simulation model involved

simulating probes that bound to non-interacting protein molecules separately from probes that bound to interacting molecules.

First, we generated the non-interacting Prox-seq probes A as a sphere of random points (Figure 5.1c). These points corresponded to the protein molecules that existed as monomers, or that their complex partners were not targeted by the Prox-seq probe panel, or that they were caused by antibody non-specific binding. (We experimentally observed that the amount of non-specific binding in the case of surface Ab staining was low.) We assumed that these protein molecules did not aggregate, and were distributed uniformly on the cell surface. Then, we repeated the process to generate the non-interacting Prox-seq probes B. Second, we repeated the process a third time to generate the complex Prox-seq probes A and B. These points corresponded to detectable protein complexes. Because these two probes A and B bound to the same protein complexes, the Prox-seq probe A points would be in the same location as their corresponding Prox-seq probe B points in this step. Finally, any pairs of probe A and B with Euclidean distances less than the ligation distance were considered ligated and produced PLA products. In our simulation, we chose the ligation distance to be 50 nm. If a probe A could ligate with more than one probes B, one such probe B was chosen at random to ligate with said probe A. Furthermore, each probe A and each probe B could only be ligated once.

In the absence of additional variance, the simulated PLA product count followed the Poisson distribution, whereas the experimental data exhibited overdispersion (Figure 5.2a, Supplementary Figure 5.7a). We discovered that using a negative binomial distribution for non-interacting proteins and protein complexes was sufficient to capture the overdispersion of experimental data (NB variance, Figure 5.2a, see "Methods"). Second, the simulated data, like the experimental data, had a right-skewed distribution across different PLA product abundances (Figure 5.2b, c). Third, the pairwise Spearman's correlation coefficients between PLA products were mostly positive and could be close to 1 (Figure 5.2d). In contrast, the simulation model with no variance showed Spearman correlation coefficients centered at zero,

whereas the model with NB variance produced a correlation coefficient distribution similar to experimental data. The lower proportion of moderate correlation coefficients (between 0 and 0.5) in simulated data compared to experimental data was most likely to caused by the presence of many non-specific Prox-seq probes in the experiment.

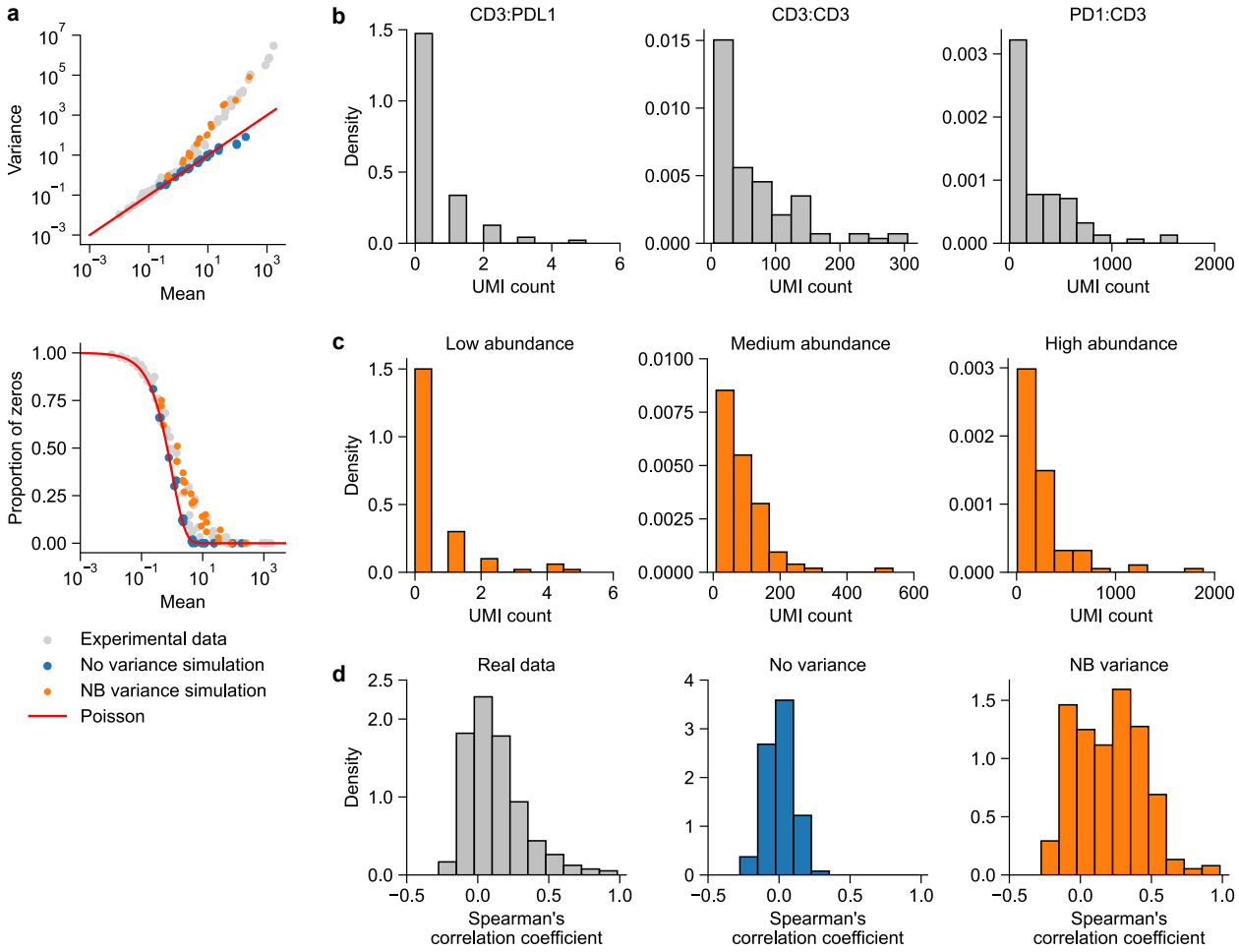

Figure 5.2: Comparison between simulated and real Prox-seq data. (a) Scatter plots showing the relationship between mean and variance (top), and between mean and proportion of zeros (bottom) of all PLA products. (b) Histograms showing the UMI counts of three example PLA products in single Jurkat cells. (c) Histograms showing the UMI counts of three example simulated PLA products. (d) Histograms showing the pairwise Spearman correlation coefficients between PLA products in real single Jurkat cell data (left), simulated data without variance (center) and simulated data with added variance in protein expression (right). The simulated data in (b–d) simulates 25 PLA products from 5 protein targets.

Notably, the simulation model with added NB variance captured the positive correlation between observed PLA product count and unligated probe count in experimental data

(Supplementary Figure 5.7b–g). The simulation model with no variance, however, showed a negative correlation between PLA product count and unligated probe count. The NB variance model also produced unligated probe counts with similar distributions to those observed in experimental data (Supplementary Figure 5.8). More specifically, experimental unligated counts displayed overdispersion, and a right-skewed distribution for both specific probe binding (e.g., CD3 and PD1 probe for Jurkat cells) and non-specific probe binding (e.g., PDL1 probe for Jurkat cells). In short, our simulation model accurately reproduced the behaviors of Prox-seq experimental count data.

5.3.4 Characterization of Prox-seq Data

Having established a simulation model, we next characterized the amount of random ligation in the most basic scenario when there is no protein interaction captured by Prox-seq probes. First, the simulation demonstrated that the amount of a PLA product produced by random ligation scaled quadratically with protein abundance and the ligation distance (Figure 5.3). Therefore, it is important to identify and remove the noise caused by random ligation, especially for highly expressed proteins.

A binomial distribution could well approximate the relationships of random ligation with protein abundance and ligation distance (Figure 5.3, see "Methods"). For the simulation with varying ligation distance, the amount of random ligation was lower than the binomial approximation. This is because the simulation assumed each probe could only be ligated at most once, while the binomial approximation imposed no such limit. As the ligation distance increased, the number of probes that could be ligated more than once increased, leading to a lower amount of random ligation in the simulation compared to the binomial distribution.

In Chapter 3, we proposed that the protein abundance, which could be measured by flow cytometry or CITE-seq, could be estimated from Prox-seq data by summing the appearances of each protein across its associated PLA products (see "Methods"). According to the simulated data, such an estimate was a good approximation of the true protein abundance,

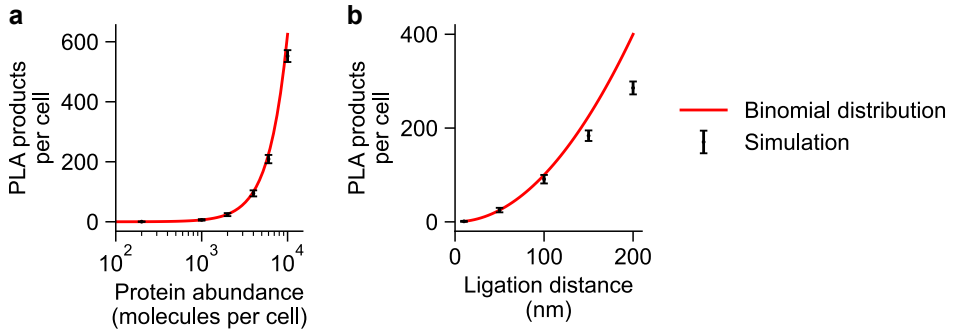

Figure 5.3: The amount of random ligation follows a binomial distribution. (a, b) Plots showing the count of PLA products created from random ligation as a function of (a) total non-interacting protein abundance, and (b) ligation distance. In (a), the ligation distance was 50 nm, and the abundances of non-interacting probe A and B were equal to half the indicated total protein abundance. In (b), the abundances of probe A and B were equal to 1,000 molecules per cell. In (a, b), no variance was included in the simulation. The simulated amount from random ligation was always lower than that predicted by the binomial distribution because in the simulation, each probe A and B could only be ligated at most once, whereas the binomial distribution did not have this limit. Data is presented as mean ± standard deviation.

as they are strongly correlated (Supplementary Figure 5.9). However, we observed that the calculated protein abundance always underestimated the true abundance. This is because there were non-interacting proteins that did not form PLA products, and therefore were not measured by Prox-seq.

5.3.5 Iterative Prediction of Protein Complex Abundance

Quantification of protein complexes is the most important characteristic of Prox-seq. In Chapter 3, we proposed that when there were no protein complexes, the observed count of a PLA product i:j could be calculated from the abundance of the probe A targeting protein i, and the probe B targeting protein j. This calculation resulted in an expected random count for PLA products that were produced only from random ligation. We reasoned that if the observed count of PLA product i:j was higher than the calculated expected random count, then i:j indicated a true protein interaction.

To quantify the protein complexes on each single cell, we calculated the difference between the observed and expected random PLA product count (Figure 5.4a). Here, we called

this method the iterative method, because it involved solving a system of quadratic equations iteratively (see Chapter 3). This was to ensure that after subtracting the protein complex count from the observed PLA product count, the remaining count displayed the characteristics of random ligation.

Next, we constructed a simulation scenario to validate the assumptions underlying the iterative method. The simulation has three protein targets, called protein 1, protein 2 and protein 3. There proteins did not interact with itself, nor with any other proteins. Furthermore, protein 3 had a lower non-interacting protein count (mean of 100 UMIs/cell compared to 1000 UMIs/cell for proteins 1 and 2, Supplementary table 5.2).

Simulated data showed that the assumptions behind the iterative method were indeed correct. The observed PLA product counts were similar to the expected random counts (Supplementary Figure 5.10a). When we allowed protein 1 to only interact with itself, the observed count of the PLA product 1:1 was higher than its expected random count (Supplementary Figure 5.10b).

5.4 Prediction of Protein Complex Abundance Using Linear Regression

We sought to leverage the measurement of unligated probes, which was enabled by the free oligo modification (Figure 5.1b), to propose another method for protein complex prediction. The random ligation amount for a PLA product i:j should be proportional to the product of the unligated probe A targeting protein i, and the unligated probe B targeting protein j. Next, we reasoned that if we used linear regression to model the observed PLA product count onto the estimated random ligation amount, true protein complexes would have positive intercepts (see "Methods"). The slope was then used to scale the estimated random ligation amount, and the abundance of a protein complex was calculated by subtracting the scaled random ligation from the observed PLA product count (Figure 5.4b). Thus, we called this

new method the linear regression (LR) method. Experimentally, we observed strong heteroscedasticity in the PLA product count when regressed on to the random ligation amount (Supplementary Figure 5.11). Therefore, we performed linear regression using weighted least squares instead of ordinary least squares (see "Methods").

As the LR method relied on the measurement of unligated probe counts, its prediction results were independent from those of the iterative method. Therefore, the LR method could act as an independent validation for the iterative method in quantifying protein complexes in experimental data.

To benchmark the protein complex prediction results of both the iterative method and the LR method, we performed another simulation whose parameters were set to approximate the experiments (Supplementary table 5.2). More specifically, the simulation had three protein targets: protein 1, protein 2 and protein 3. Proteins 1 and 2 interacted with each other and with itself (Figure 5.4c). Protein 3 did not interact with itself, nor with protein 1 or protein 2. Furthermore, protein 3 had very low non-interacting protein count (mean of 2 UMIs/cell compared to 20 and 15 UMIs/cell for proteins 1 and 2, respectively).

Both the iterative and LR methods were able to correctly predicted which protein complexes existed (Figure 5.4c). Quantitatively, the iterative method consistently underestimated the true protein complex count (Figure 5.4c–e). On the other hand, the LR method produced much more accurate quantification of the protein complexes (Figure 5.4f). We also performed a one-sided Fisher's exact test as another way to validate the two methods with respect to protein complex detection (Figure 5.4d).

Next, we evaluated the concordance between the iterative and the LR methods on experimental data from single Jurkat and Raji cells. First, we found that both methods largely agreed on which PLA products were predicted to be protein complexes (Figures 5.5a–f). The protein complex count predicted by the two methods showed higher correlation among Jurkat cells than Raji cells (Figure 5.5c, g). The comparatively lower correlation in predicted complex count in Raji cells was because the iterative method's predictions were generally

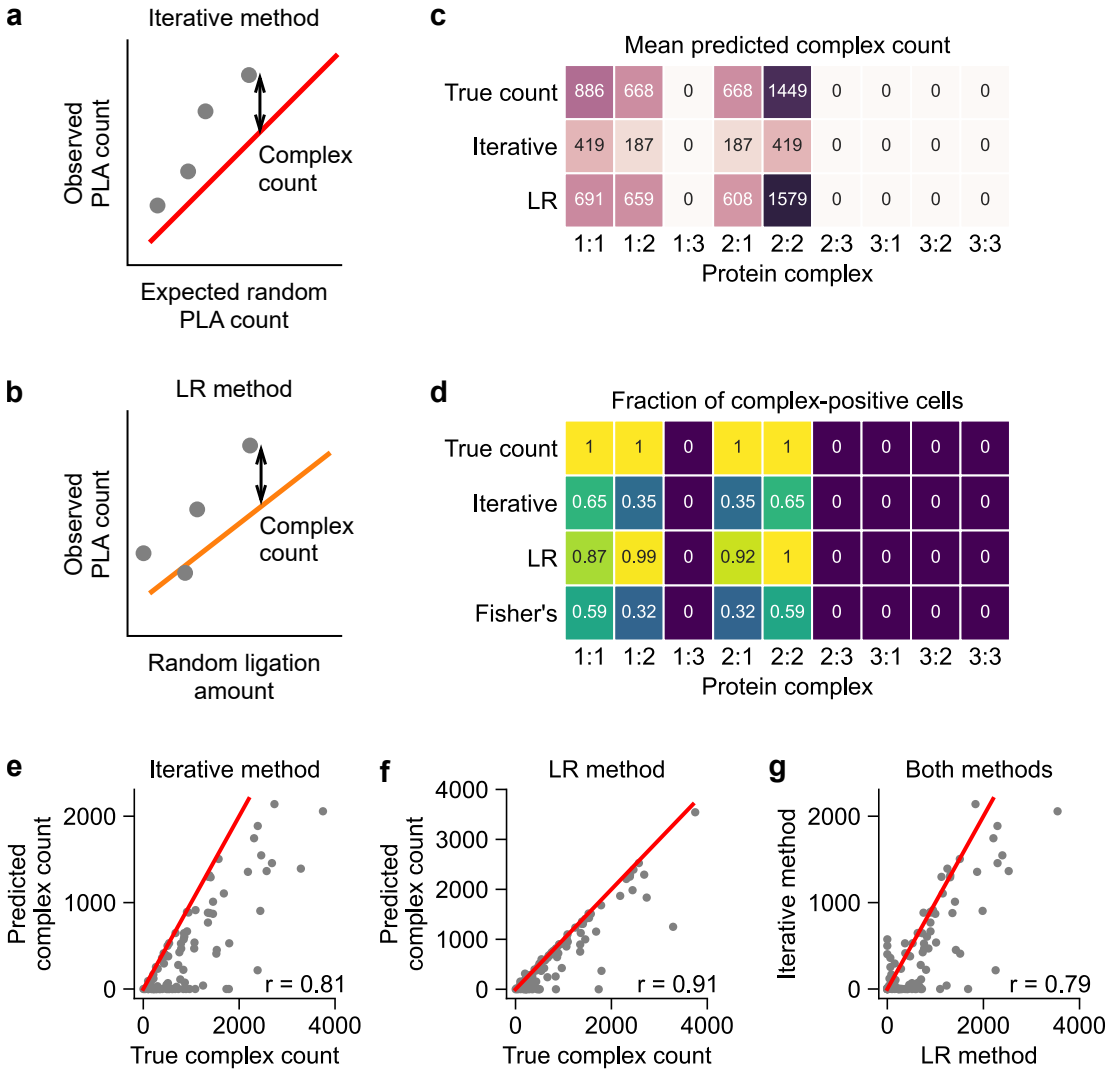

Figure 5.4: Comparison between the iterative and linear regression (LR) methods on simulated data. (a, b) Schematics showing the working principle of (a) the iterative method and (b) the LR method. In the iterative method, the protein complex count is the difference between the observed and expected random PLA product count. In the LR method, the protein complex count was equal to the difference between the observed PLA product count and its expected amount of random ligation. The random ligation amount was calculated from the unligated probe count. In (a), the red line indicates y = x. In (b), the orange line indicates the linear regression fit. (c) Heatmap showing the mean complex count of simulated data, and of the iterative and LR methods' prediction results. (d) Heatmap showing the fraction of cells expressing a protein complex, as predicted by the iterative method, the LR method, and Fisher's exact test. In (c, d), the true count represented the true protein complex count in the simulation. (e, f) Scatter plots showing the simulated and predicted count of protein complex 1:1 using (e) the iterative and (f) the LR method. (g) Scatter plot comparing the predicted count of protein complex 1:1 from the iterative and the LR methods. In (e–g), the red lines indicate y = x, and each dot represents a single cell.

lower than that of the LR method (Figure 5.5h). A similar underestimation of the iterative method was also observed in the simulated data, albeit to a much lesser degree (Figure 5.4g).

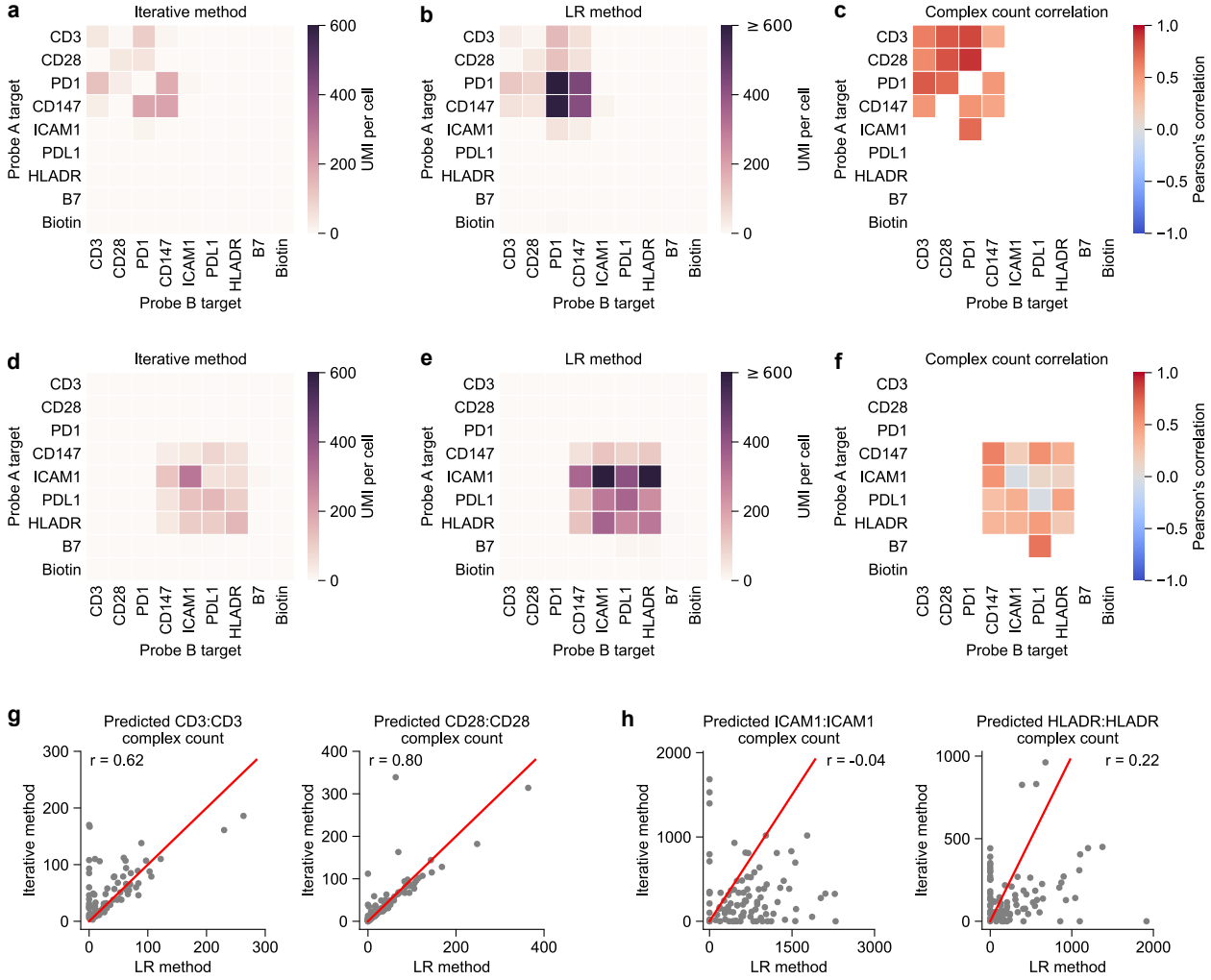

Figure 5.5: Comparison between the iterative and linear regression (LR) methods on experimental data. (a, b) Heatmaps showing the average of protein complex count, predicted by (a) the iterative method and (b) the LR method in Jurkat cells. (c) Heatmap showing the Pearson's correlation coefficients of predicted complex count between the iterative and LR methods in Jurkat cells. (d, e) Heatmaps showing the average of protein complex count, predicted by (d) the iterative method and (e) the LR method in Raji cells. (f) Heatmap showing the Pearson's correlation coefficients of predicted complex count between the iterative and LR methods in Raji cells. In (c, f), only protein complexes predicted by both iterative and LR methods are shown. (g) Comparison of predicted counts of protein complexes CD3:CD3 and CD28:CD28 in Jurkat cells. (h) Comparison of predicted counts of protein complexes ICAM1:ICAM1 and HLADR:HLADR in Raji cells. In (g, h), the red lines indicate y = x, and r indicates the Pearson's correlation coefficient.

In addition, we observed that the LR method predicted more protein complexes than

the iterative method, especially in Jurkat cells, and the iterative method itself predicted more protein complexes than the Fisher's exact test (Supplementary Figure 5.12). All three methods predicted protein complexes CD3:CD3 and CD28:CD28 in Jurkat cells, both of which are known to exist on T cells. All three methods also predicted protein complex ICAM1:ICAM1 in Raji cells, which has been shown to form on the cell surface [220].

We also biotinylated the cell surface (see "Methods") in order to introduce a negative control target that we expected not to interact with any other proteins. The iterative method, the LR method, and the Fisher's exact test did show that the surface biotin only interacted with a few proteins, and these complexes had very low abundances (Figure 5.5a–f, Supplementary Figure 5.12). For example, in Jurkat cells, the LR method predicted that biotin:PD1 protein complex was expressed at an average of 8 UMIs per cell, compared to an average of 34 and 43 UMIs per cell for CD3:CD3 and CD28:CD28 complexes, respectively (Figure 5.5b).

In short, we found that both the iterative and LR methods agreed well on which protein complexes were present in the experimental data, and that, in some situations, the LR method produced more accurate protein complex counts.

5.5 Discussion

Here we presented computational frameworks for simulating Prox-seq data, and for predicting protein complex abundance from Prox-seq data. First, our simulation model showed that the background of Prox-seq, which was caused by random ligation of non-interacting proteins that were in proximity to one another by random chance, strongly depended on the protein abundance. This calls into questions the choice of no primary antibody as a negative control that is recommended by commercial PLA kits [217]. In our study, the random ligation effects were not significant, because the number of unligated probes was low, and because non-specific antibody binding is minimal for extracellular proteins.

Second, we showed that with respect to protein complex prediction, the iterative method

(proposed in Chapter 3) agreed well with the LR method on real experimental data. Therefore, we recommended that both the iterative and LR methods should be used for protein complex detection, and any protein complexes that were predicted by both methods were highly likely to be true protein complexes. For a more conservative result, the Fisher's exact test could be used to determine which protein complexes were present. Simulation suggested that the LR method produced more accurate protein complex quantification.

Our simulation model had several limitations. First, it did not take into account interactions higher than dimers, diffusion of the protein molecules, their physical sizes, and the technical variability of the Prox-seq assay. Second, each antibody molecule was assumed to be conjugated to only one DNA oligonucleotide. Third, the simulation model assumed that the abundance of a protein complex was unrelated to the abundance of the non-interacting components of the complex's subunits . Finally, it assumed that the protein complexes and the non-interacting proteins were uniformly distributed on the cell surface.

In this study, we applied the iterative method and the LR method on a homogeneous population of single cells. The current implementation of these methods made them unsuitable to study a heterogeneous population of single cells. This is because both methods relied on a statistics of the whole population (the difference between observed and expected random PLA product count for the iterative method, and the linear regression's intercept coefficient for the LR method). Having different complex expression levels within the population would lower the power of these methods. Further study is required to extend these methods to a population of heterogeneous cell types, perhaps by leveraging mRNA data to identify the cell types prior to protein complex prediction.

We envision that the LR method will be particular useful when Prox-seq is extended to intracellular proteins. Indeed, since non-specific antibody binding is much more severe in intracellular staining than extracellular staining, random ligation is an all the more important source of background. The simulation model can also be further extended to model Prox-seq data of intracellular proteins. In short, we have validated the protein complex prediction

algorithm that was proposed in Chapter 3, proposed an additional independent method for protein complex prediction, and introduced a physical model for simulating Prox-seq data.

5.6 Supplementary Figures

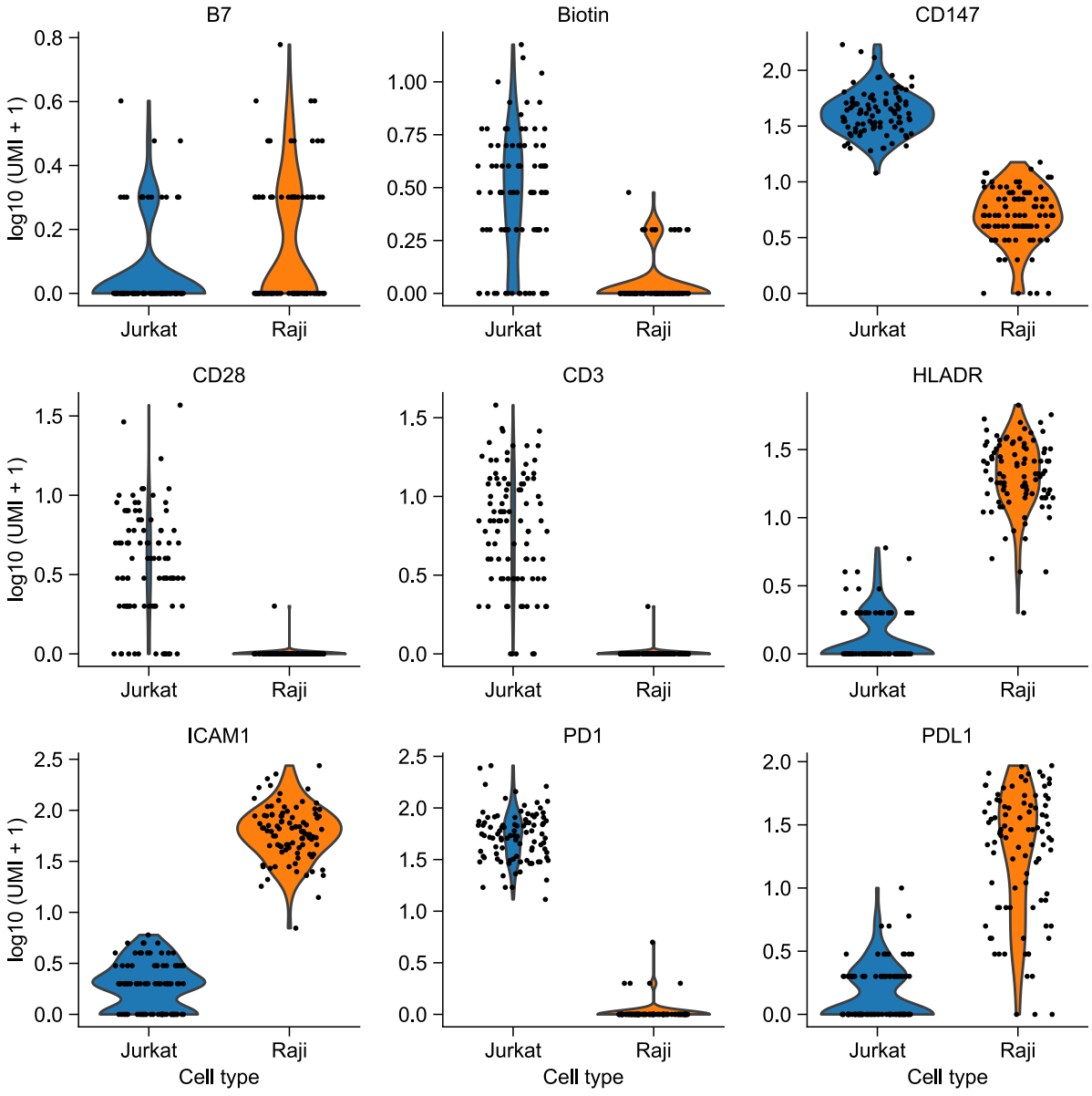

Figure 5.6: Measured unligated probe count in experimental data. Violin plots showing the log-transformed total unligated count (i.e., sum of unligated probe A and unligated probe B) in single Jurkat and Raji cells.

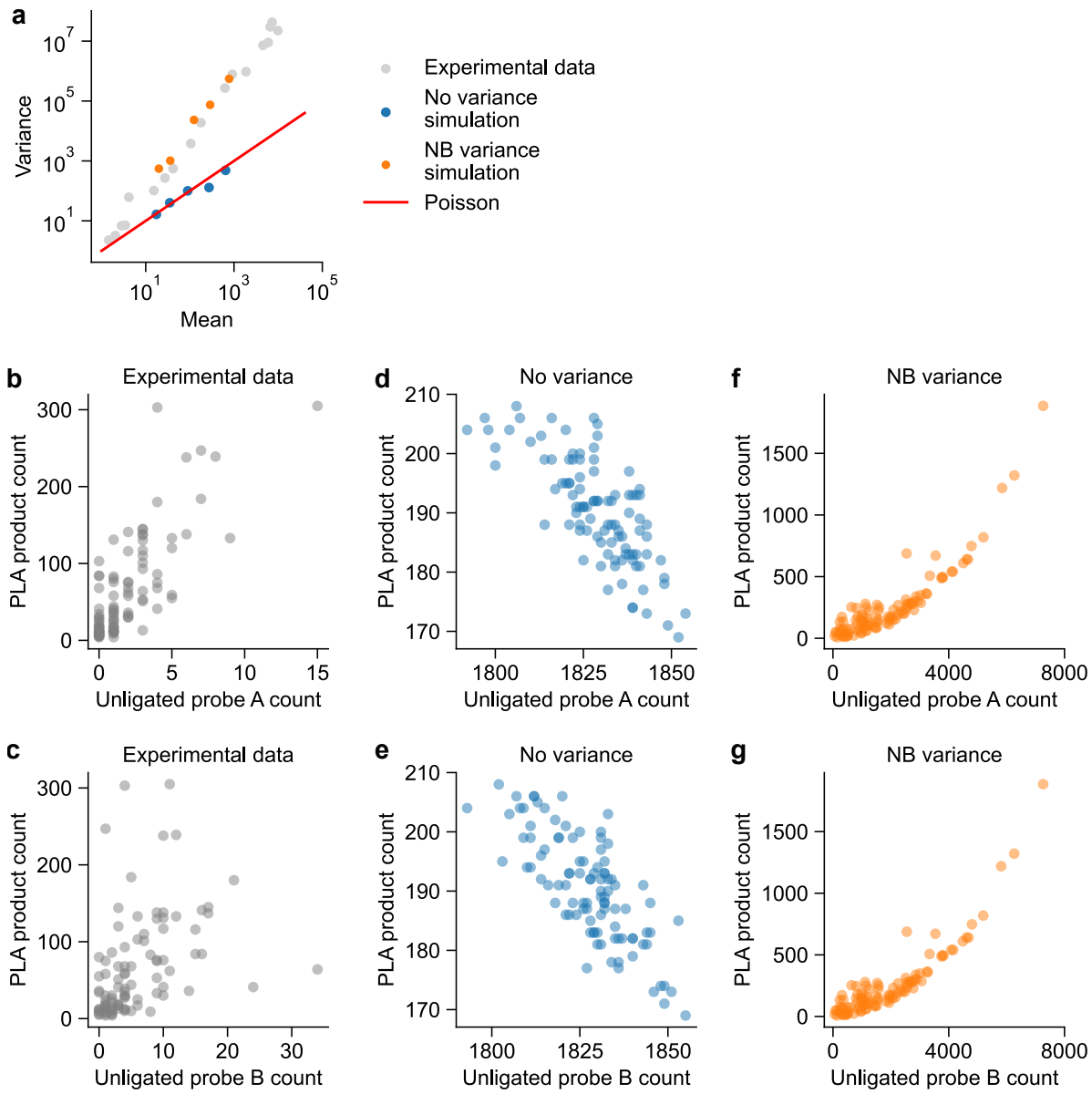

Figure 5.7: Comparison of experimental and simulated data for protein count and unligated probe count. (a) Scatter plot showing the mean-variance relationship in experimental and simulated protein count. (b, c) Scatter plots showing the relationship between observed CD3:CD3 PLA product and (b) unligated CD3 probe A or (c) unligated CD3 probe B in Jurkat cells. (d, e) Scatter plots showing the relationship between observed 1:1 PLA product and (d) unligated protein 1 probe A or (e) unligated protein 1 probe B in simulated data without variance. (f, g) Scatter plots showing the relationship between observed 1:1 PLA product and (f) unligated protein 1 probe A or (g) unligated protein 1 probe B in simulated data with negative binomial variance.

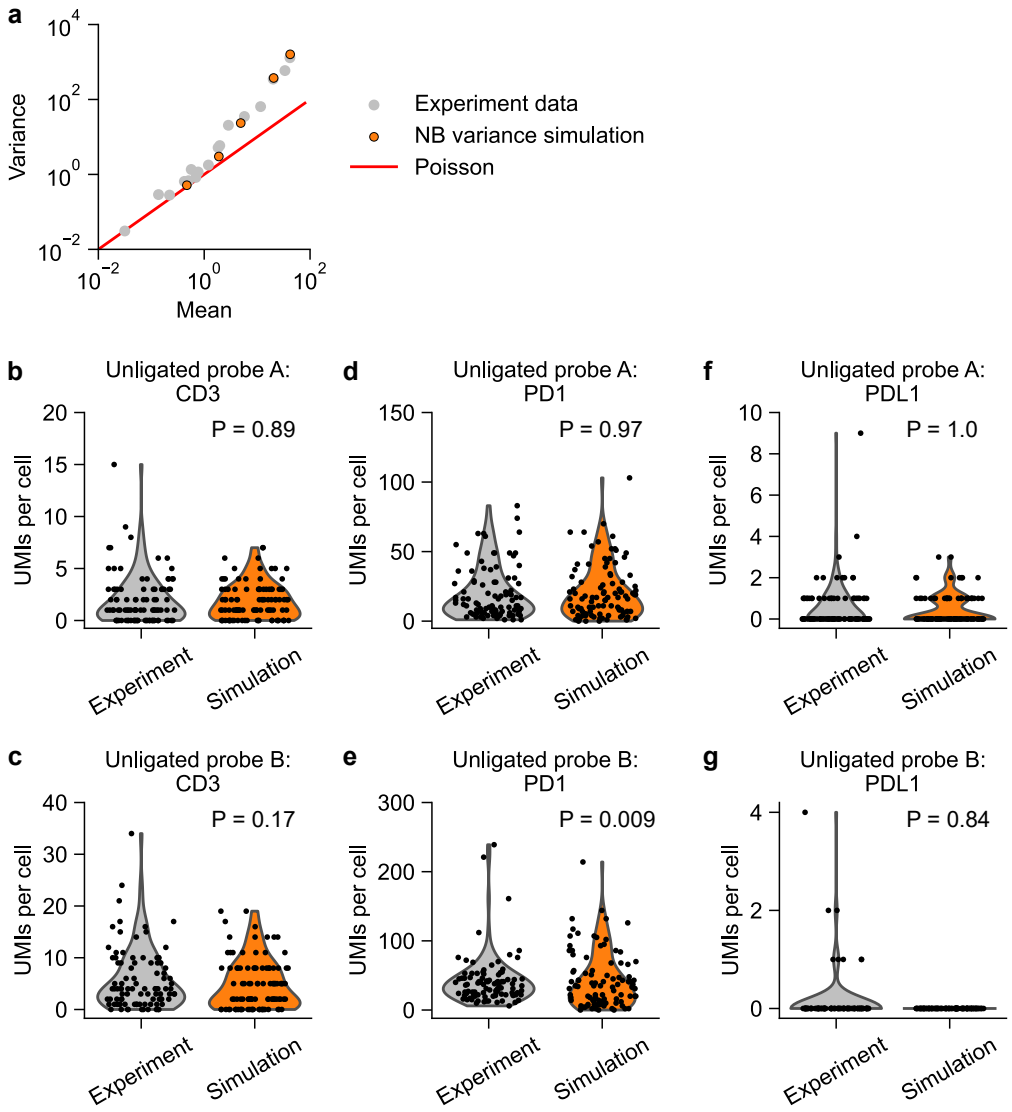

Figure 5.8: Distribution of unligated probe counts in Jurkat cells and simulated data. (a) Scatter plot showing the mean-variance relationship in experimental (Jurkat cells) and simulated unligated probe count. (b, c) Violin plots showing the experimental and simulated count of (b) unligated probe A and (c) unligated probe B for CD3 protein. (d, e) Violin plots showing the experimental and simulated count of (d) unligated probe A and (e) unligated probe B for PD1 protein. (f, g) Violin plots showing the experimental and simulated count of (f) unligated probe A and (g) unligated probe B for PDL1 protein. Unligated probe counts for CD3, PD1 and PDL1 proteins were simulated by using the mean unligated probe counts in experimental data. Note that Jurkat cells expressed CD3 and PD1 proteins, but not PDL1 protein. P-values are calculated using KS test.

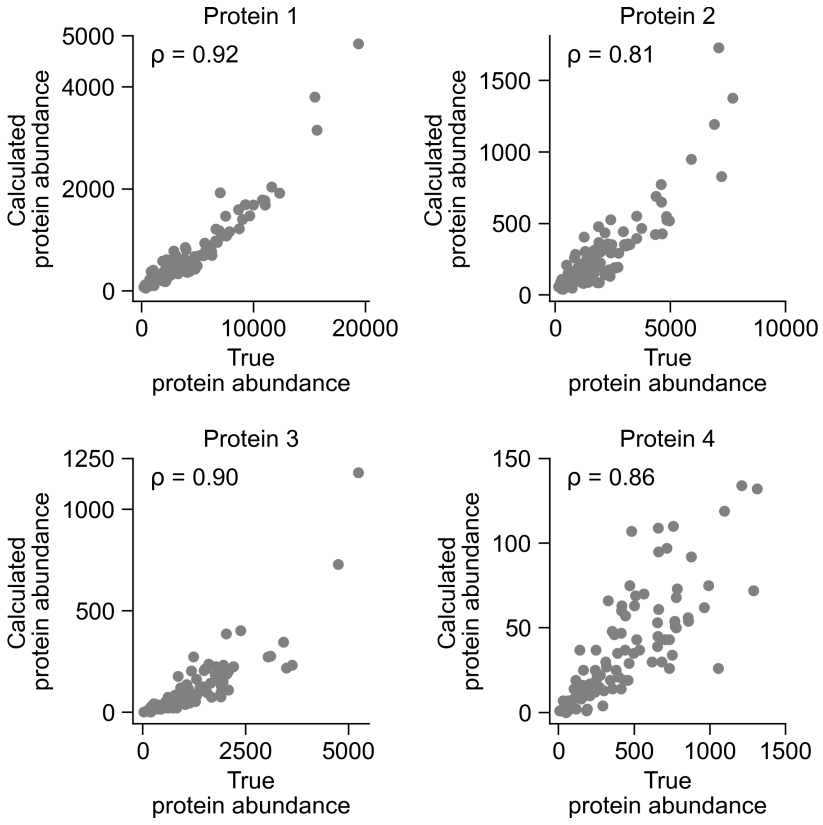

Figure 5.9: Comparison between true and calculated protein count. Each panel also displays the corresponding Spearman's correlation coefficient, ρ.

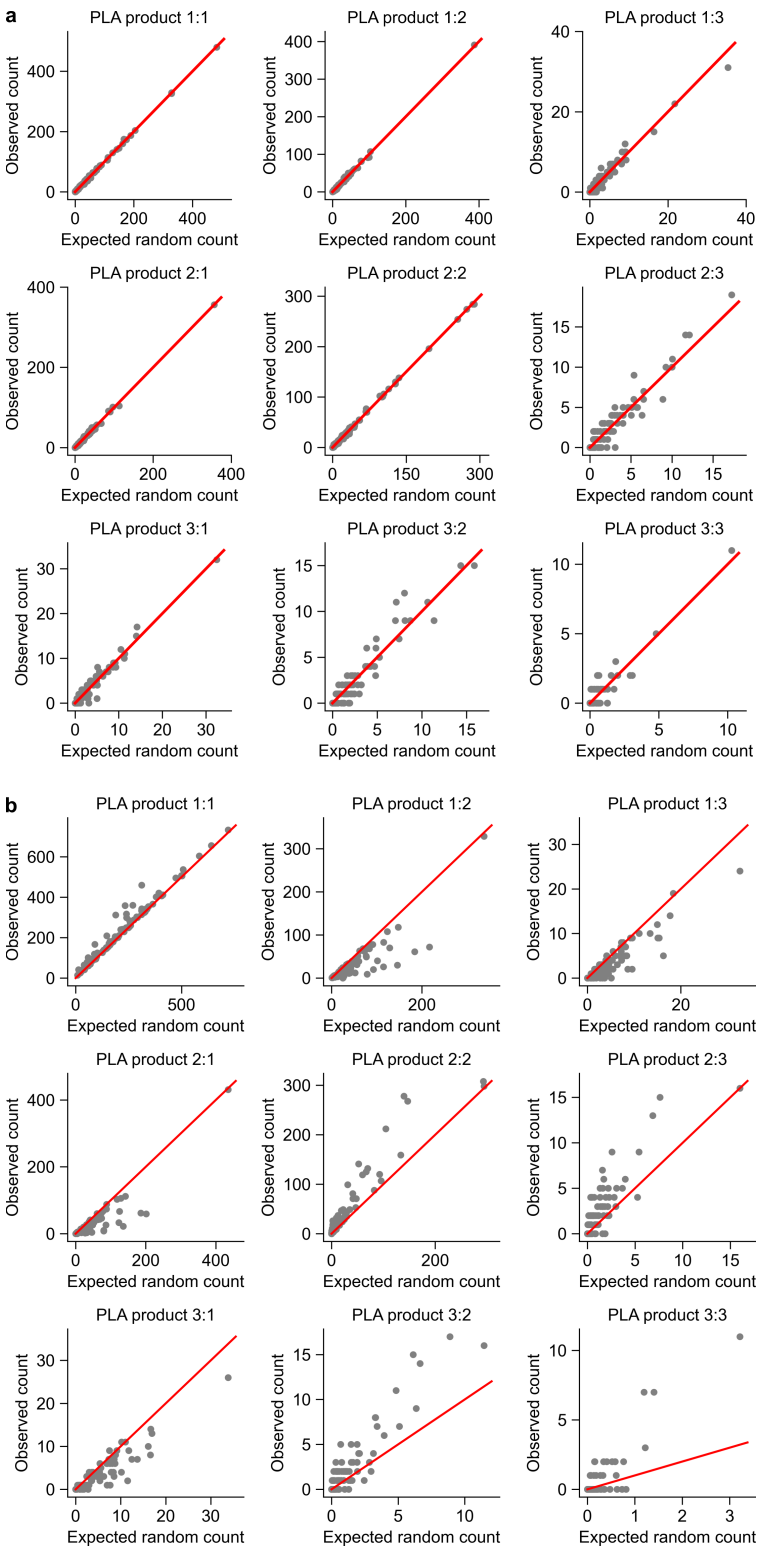

Figure 5.10: Comparison between observed and expected random PLA product count in simulated data. (a, b) Scatter plots showing the observed and expected random count of each PLA product in the scenario when (a) no protein complex, and when (b) 1:1 is the only protein complex. The red lines indicate y = x.

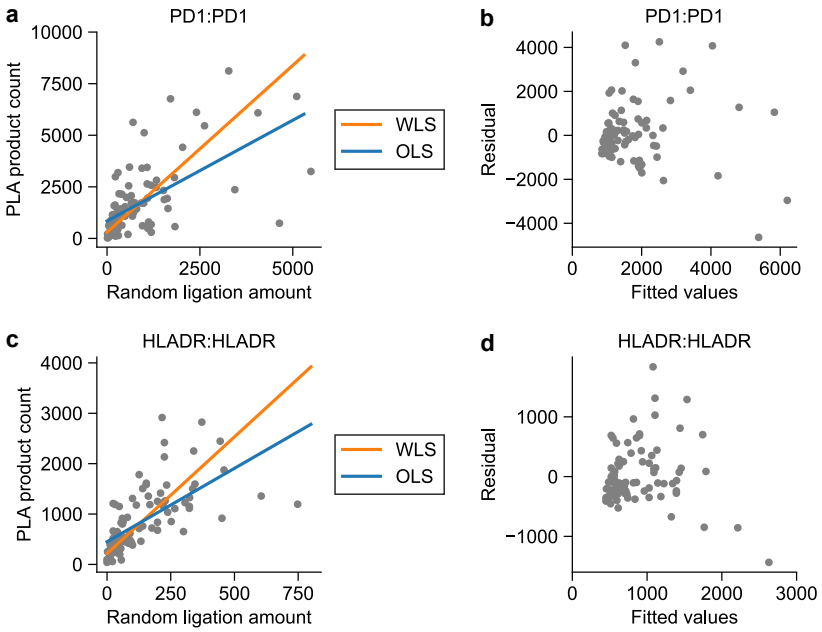

Figure 5.11: Heteroscedasticity in PLA product count. (a) Scatter plot showing the relationship between observed count of PLA product PD1:PD1 and the measured random ligation amount in Jurkat cells. (b) Residual plot of ordinary least squares regression for PLA product PD1:PD1 in Jurkat cells. (c) Scatter plot showing the relationship between observed count of PLA product HLADR:HLADR and the measured random ligation amount in Raji cells. (d) Residual plot of ordinary least squares regression for PLA product HLADR:HLADR in Raji cells. In (a, c), WLS and OLS stand for weighted least squares and ordinary least squares, respectively.

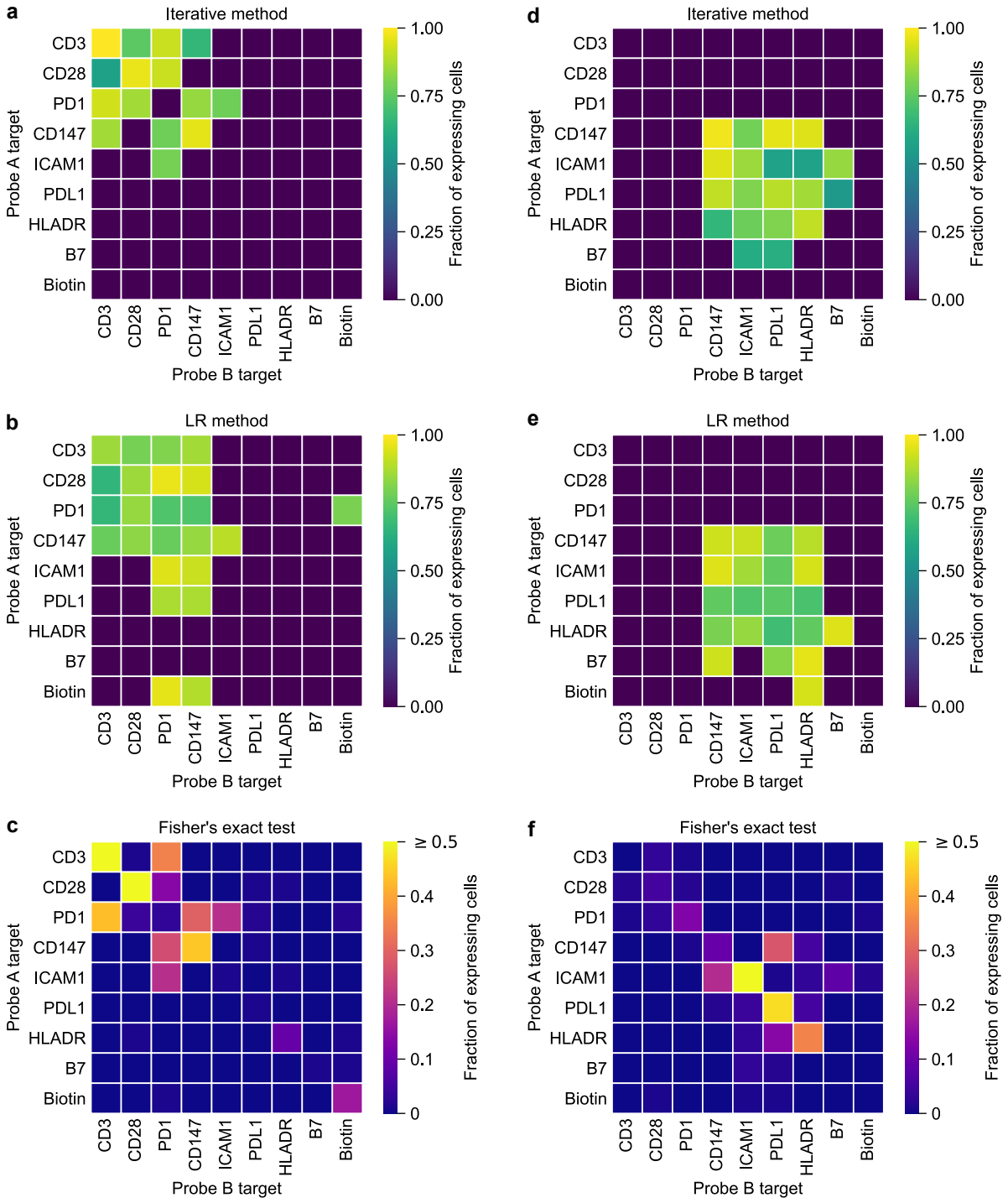

Figure 5.12: Comparison between the iterative method, the LR method, and Fisher's exact test for protein complex detection. (a–c) Heatmaps showing the fraction of Jurkat cells that express a protein complex, as predicted by (a) the iterative method, (b) the LR method, and (c) the Fisher's exact test. (d–f) Heatmaps showing the fraction of Raji cells that express a protein complex, as predicted by (d) the iterative method, (e) the LR method, and (f) the Fisher's exact test.

5.7 Methods

Cell Culture

Jurkat PD1 and Raji PDL1 cell lines were a generous gift from Jun Huang [155]. Both cell lines were maintained at 37°C with 5% CO2 in RPMI (Gibco, Thermo Scientific) supplemented with 10% Fetal Bovine Serum (FBS, Hyclone, Fischer Scientific).

Prox-seq Probe Conjugation

Prox-seq probes were generated using previously published methods [156]. First, antibodies were concentrated and buffer exchanged using a 50,000 molecular weight cutoff (MWCO) concentrator (EMD Millipore). This was accomplished by first wetting the concentrator membrane with 500 μL PBS followed by a 14,000 × g spin at 4°C for 5 min. Antibodies were then diluted to 500 μL with PBS, placed in the concentrator, and centrifuged a 14,000 × g at 4°C for 5 min. This procedure was repeated twice more, with the final spin time increased to 10 min. Concentrated antibody solutions were then combined in a 10:1 volume-to-volume ratio with 2 mM dibenzocyclooctyne-PEG4-N-hydroxysuccinimidyl ester (DBCO, Sigma) in dimethyl sulfoxide (DMSO, Sigma Aldrich) and allowed to react on ice for 1 hour. The unreacted DBCO was removed by repeating the buffer exchange step detailed above.

To produce finished Prox-seq probes, DBCO-conjugated antibodies were reacted with azide-functionalized DNA oligonucleotides (oligos) (Supplementary tables 3.5, 3.6). This was achieved by combining 2 μg DBCO-conjugated antibody with an equal volume of 80 μM oligo. The reaction was allowed to proceed overnight at 4°C before being brought to 1.3 μM with PBS or 50% glycerol in PBS.

The list of antibodies used in this chapter is given in Supplementary table 5.1.

Sample Processing

Jurkat and Raji cells (300,000 each) were collected and centrifuged at 300 × g for 5 min. These centrifuge settings were applied to all pelleting steps in this procedure. Cells were resuspended in 1 mL 1% BSA/PBS (PBS from Gibco, 10% BSA from Thermo Scientific) and pelleted. Cells were then resuspended in PBS and pelleted. At this stage, cells treated with biotin were resuspended in 200 μL 5 μM Biotin-NHS (Sigma) in PBS and allowed to incubate at 37°C for 15 min. Biotinylated cells were quenched by the addition of 1 mL 1% BSA/PBS and pelleted.

All cells were pelleted and resuspended in 90 μL Probe Binding Buffer (PBS, 0.1% BSA (Thermo Scientific), 0.1 mg/ml sonicated salmon sperm DNA (Invitrogen), 50 μg/ml of the following isotype antibodies: goat, mouse, rat, rabbit, Armenian hamster, and sheep) with Prox-seq probes, each probe at 2.5 nM. The cells were allowed to incubate for 37°C for 30 min. The cells were then pelleted and washed twice with 1 ml 1% BSA/PBS.

To ligate cells, they were first pelleted, resuspended in 100 μL Ligation Solution (50mM HEPES pH 7.5, 10 mM $MgCl_2$, 1 mM rATP (NEB), 9.5 nM Connector oligomer (TTTCACGACACGACACGATTTAGGTC, IDT), 130 U/ml T4 ligase (NEB)), and allowed to incubate at 37°C for 30 min. Free oligo A was then added to this solution to a final concentration of 50 nM and connector was added to a final concentration of 59.5 nM. After a 15-min incubation at 37°C, free oligo B was added to a final concentration of 60 nM along with additional connector to a final concentration of 109.5 nM. This solution was allowed to react for an additional 15 min at 37°C. Finally, 1 ml 1% BSA/PBS was added to the cells and they were pelleted. This procedure was repeated two times. The final addition of 1 ml 1% BSA/PBS included 1:500 propidium iodide (Invitrogen). Each single live cell was sorted into a well on a 96-well plate using a BD FACSAria Fusion sorter. Each well contained 2 μL of lysis buffer (0.2% Triton X and 2 units/μL RNase inhibitor, murine (NEB) in water), 1 μL of 10 μM SmartSeq2_oligodTVN and 10 μM SmartSeq2_oligodTGT, and 1 μL of 10 mM dNTPs (NEB).

Library Construction

All 4 μL of single-cell lysate was amplified in a 25-μL PCR reaction (1× Kapa Hifi Hotstart Readymix, 0.1 μM SmartSeq2_oligodTGT, 0.1 μM U_fwd). The PCR program was: 98°C 3 min; 5 cycles of 98°C 20 s, 55°C 15 s, 72°C 1 min; 17 cycles of 98°C 20 s, 67°C 15 s, 72°C 1 min; 72°C 5 min; 4°C hold. After PCR, the PLA products were purified using 1.8× AMPure XP beads.

To construct the sequencing library, 5 μL of the purified products were used to set up a 25-μL PCR reaction (1× Kapa Hifi Hotstart Readymix, 1 μM SmartPLA_P7_N7XX, 1 μM SmartPLA_P5_S5XX). The PCR program was: 98°C 3 min; 11 cycles of 98°C 20 s, 67°C 15 s, 72°C 1 min; 72°C 5 min; 4°C hold. After PCR, the libraries were purified using 1.8× AMPure beads, and quantified with Qubit 1× dsDNA High Sensitivity Assay Kit (Thermo Fisher) or Fragment Analyzer (Agilent).

The primer sequences are listed in Supplementary tables 3.8–3.10.

Next-Generation Sequencing, Sequencing Alignment, and Data Analysis

The single-cell libraries were pooled at equimolar, mixed with 40% PhiX control, and sequenced with a NextSeq 550 v2.5 mid-output kit. The read lengths are 75 bases for read 1, 8 bases for index 1 read, and 8 bases for index 2 read.

The sequencing data was aligned using a custom alignment program (`https://github.com/tay-lab/Prox-seq`, see Chapter 3). Sequencing reads with more than one bases with quality score below 20 are discarded.

For data analysis, we only kept single cells with more than 10 UMIs and fewer than 30,000 UMIs detected, and cells with more than 20 detected PLA products.

Simulation Model

Assume that each protein molecule and the Prox-seq probe that bound to it were point particles. Let there be n protein targets. Let $A_1, A_2, ..., A_n$ be the simulation parameters that represented the count of non-interacting (i.e., not part of a protein complex) probe A that targeted proteins $1, 2, ..., n$. Let $B_1, B_2, ..., B_n$ be the simulation parameters that represented the count of non-interacting probe B that targeted proteins $1, 2, ..., n$. Let $c_{1,1}$, $c_{1,2}, ..., c_{1,n}, c_{2,1}, c_{2,2}, ..., c_{n,n}$ be the simulation parameters that represented the count of protein complexes that corresponded to PLA products 1:1, 1:2, ..., 1:n, 2:1, 2:2, ..., n:n. If protein i did not interact with protein j, then $c_{i,j} = 0$.

The simulation was performed separately on each single cell. For the single cell t, we first randomly generated a sphere of $A_i^{(t)}$ points, which corresponded to the number of detected probe A that targeted protein i on cell t. The coordinates of each point were generated using the following equations [221]:

$$x = R\sqrt{1-u^2}\cos\theta \tag{5.1}$$

$$y = R\sqrt{1-u^2}\sin\theta \tag{5.2}$$

$$z = Ru \tag{5.3}$$

where R was the radius of the sphere (taken to be 5 μm, or 5000 units, in our study), u was uniformly distributed over $[-1, 1)$, and θ was uniformly distributed over $[0, 2\pi)$.

Without added variance, $A_i^{(t)} = A_i$. With added negative binomial (NB) variance:

$$A_i^{(t)} \sim \text{NegativeBinomial}(n_{NB}, p_{NB}) \tag{5.4}$$

where $n_{NB} = 1.5$ in our study, and $p_{NB} = \left(1 + \frac{A_i}{n_{NB}}\right)^{-1}$. The negative binomial distribution formulated this way provided the probability of getting $A_i^{(t)}$ failures, given n_{NB} successes and p_{NB} was the probability of success. n_{NB} was used to control the variance of

the probe count, and p_{NB} was calculated such that the mean of $A_i^{(t)}$ was equal to A_i.

Second, we randomly generated a sphere of $B_i^{(t)}$ points, which corresponded to the number of detected probe B that targeted protein i on cell t. The coordinates of each point were generated using equations (5.1)–(5.3).

Without added variance, $B_i^{(t)} = B_i$. With added NB variance:

$$B_i^{(t)} = \frac{B_i}{A_i} \times A_i^{(t)} \tag{5.5}$$

Equation (5.5) ensured that the counts of detected probe A and probe B that targeted the same protein were proportional to each other.

Third, we randomly generated a sphere of $c_{i,j}^{(t)}$ points, which corresponded to the count of protein complex i:j on cell t. Then, these $c_{i,j}^{(t)}$ points were added to the previously generated probe A points targeting protein i ($A_i^{(t)}$), and also to the previously generated probe B targeting protein j ($B_j^{(t)}$).

Without added variance, $c_{i,j}^{(t)} = c_{i,j}$. With added NB variance:

$$c_{i,j}^{(t)} \sim \text{NegativeBinomial}(n_{NB}, p_{NB}) \qquad \text{if } i \leq j \tag{5.6}$$

$$c_{i,j}^{(t)} = \frac{c_{i,j}}{c_{j,i}} c_{j,i}^{(t)} \qquad \text{if } i > j \tag{5.7}$$

where $n_{NB} = 1.5$ in our study, and $p_{NB} = \left(1 + \frac{c_{i,j}}{n_{NB}}\right)^{-1}$. Equation (5.7) ensured that the counts of protein complexes i:j and j:i were correlated.

Fourth, we calculated the pairwise Euclidean distances between all generated probe A points and all generated probe B points. Finally, we randomly iterated through the pairs of points that were within a ligation distance threshold (chosen to be 50 nm, or 50 units, in our study, unless stated otherwise), and added the corresponding PLA product to the simulated count matrix. Any probe A and probe B points that were chosen were excluded from future ligation. In other words, each probe A and each probe B could be ligated at most once.

The number of probe A and probe B points that were not ligated are returned as the unligated probe count. In an experiment, this unligated probe count was measured with the free oligo modification.

In this chapter, the simulation was repeated 100 times to simulate PLA product counts of 100 single cells. The simulation parameters of all simulated data sets are listed in Supplementary table 5.2. All simulations used the NB variance model, unless stated otherwise.

Binomial Approximation of Random Ligation Amount

Suppose there are A_i probes A and B_j probes B on the cell surface. Assume that the probes were uniformly distributed on the surface, and proteins i and j did not interact. Then, the Euclidean distance L between any pair of points had the following distribution [222]:

$$P(L) = \frac{L}{2R^2} \tag{5.8}$$

For a pair of points to be ligated and produce a PLA product, the distance L must be less than or equal to the ligation distance threshold, d_{ligation}:

$$P(L \leq d_{\text{ligation}}) = \frac{d_{\text{ligation}}^2}{4R^2} \tag{5.9}$$

Assume that each probe could be ligated as many times as possible, the count of the PLA product i:j followed a binomial distribution:

$$X_{i,j} \sim \text{Binomial}\left(n = A_i \times B_j, p = P(L \leq d_{\text{ligation}})\right) \tag{5.10}$$

The expected value of the count of a PLA product that was created from random ligation of non-interacting probes was:

$$E(X_{i,j}) = \frac{d_{\text{ligation}}^2}{4R^2} A_i B_j \tag{5.11}$$

Note that this approximation assumed that each probe can be ligated many times, while the simulation model assumed that each probe could only be ligated at most once.

Calculation of Protein Abundance and Expected Random PLA Product Count

The count of a protein i in a single cell was equal to the total number of detected PLA products that were related to the protein i:

$$\text{Protein}_i = \sum_{l=1}^{n} X_{i,l} + \sum_{k=1}^{n} X_{k,i} \tag{5.12}$$

Note that the PLA product $i{:}i$ was counted twice towards the protein count.

The expected random count of a PLA product $i{:}j$ was:

$$E_{i,j} = \frac{\sum_{l=1}^{n} X_{i,l} \times \sum_{k=1}^{n} X_{k,j}}{\sum_{k=1}^{n} \sum_{l=1}^{n} X_{k,l}} \tag{5.13}$$

Protein Complex Prediction: Iterative Method

The count of protein complex $i{:}j$ was calculated iteratively using the following equation:

$$Y_{i,j}^{(m+1)} = X_{i,j} - \frac{\sum_{l=1}^{n} \left(X_{i,l} - Y_{i,l}^{(m)} \right) \times \sum_{k=1}^{n} \left(X_{k,j} - Y_{k,j}^{(m)} \right)}{\sum_{k=1}^{n} \sum_{l=1}^{n} \left(X_{k,l} - Y_{k,l}^{(m)} \right)} \tag{5.14}$$

where $X_{i,j}$ is the observed (i.e., measured) count of PLA product $i{:}j$, and $Y_{i,j}^{(m)}$ is the predicted count of protein complex $i{:}j$ at the m^{th} iteration. The initial values for all protein complexes were 0.

The second term of the right hand side represented the count of PLA product $i{:}j$ that was caused by random ligation.

After each iteration, a one-sided t-test was performed on the values of $Y_{i,j}^{(m+1)}$ across all

single cells. The alternative hypothesis was that the mean of $Y_{i,j}^{(m+1)}$ was greater than 1. Next, any $Y_{i,j}^{(m+1)}$ with Benjamini-Hochberg-corrected P-values above 0.05 are set to 0. In other words, any such PLA products did not represent functional protein interactions.

We also enforced a symmetry condition, such that if i:j was a protein complex, then j:i should also be a protein complex. This was done by setting $Y_{j,i}^{(m+1)}$ as a fraction of $Y_{i,j}^{(m+1)}$ if $Y_{j,i}^{(m+1)}$ failed the t-test, but $Y_{i,j}^{(m+1)}$ passed:

$$Y_{j,i}^{(m+1)} = \text{sym\_weight} \times Y_{i,j}^{(m+1)} \tag{5.15}$$

where sym_weight = 0.25 for our study.

In our study, the iterative method was ran until the maximum change in predicted single-cell protein complex counts between successive iterations was below 1, or until reaching 200 iterations, whichever came first.

Protein Complex Prediction: Linear Regression (LR) Method

For each PLA product i:j, its observed count was linear regressed onto the product of its corresponding unligated probe A count and unligated probe B count:

$$X_{i,j} \sim \beta_0 + \beta_1 A_i^{(U)} B_j^{(U)} \tag{5.16}$$

where $A_i^{(U)}$ and $B_j^{(U)}$ are the count of unligated probe A targeting protein i, and unligated probe B targeting protein j, respectively. The coefficients were estimated using weighted least squares, with weights W equal to the reciprocal of the interaction term:

$$W = \frac{1}{A_i^{(U)} B_j^{(U)}} \tag{5.17}$$

To avoid division by zero, only cells with non-zero random ligation amount (i.e., the interaction term was higher than zero) were included in the weighted least squares model.

Moreover, only PLA products with at least three cells with non-zero random ligation amount were considered.

For simulated data, we also scaled the interaction term by 10^6, so that it was in a similar order of magnitude of $X_{i,j}$. $A_i^{(U)}$ and $B_j^{(U)}$ were obtained from PLA products that contained the added free oligos. For example, the count of unligated CD3 probe A, $A_{CD3}^{(U)}$, was equal to the count of PLA product CD3:free_oligo_B, and the count of unligated CD28 probe B, $B_{CD28}^{(U)}$, was equal to the count of PLA product free_oligo_A:CD28.

Next, we performed a one-sided t-test on the intercept coefficient, and the alternative hypothesis was that $\beta_0 > 1$. PLA products with Benjamini-Hochberg-corrected P-values below 0.05 were considered to be true protein complexes. The protein complex count, $Y_{i,j}$, was calculated as the difference between the observed PLA product count and the interaction term:

$$Y_{i,j} = X_{i,j} - \beta_1 A_i^{(U)} B_j^{(U)} \tag{5.18}$$

The LR method is related to the binomial approximation of the random ligation amount (equation (5.11)). If the counts of unligated probes were perfect proxies for the count of non-interacting probes, then we had the following relationship:

$$\beta_1 = \frac{d_{\text{ligation}}^2}{4R^2} \tag{5.19}$$

Protein Complex Prediction: Fisher's Exact Test

For each PLA product i:j, we constructed a 2x2 contingency table:

	Probe B target	
Probe A target	j	Not j
i	$X_{i,j}$	$\sum_{l=1,l\neq j}^{n} X_{i,l}$
Not i	$\sum_{k=1,k\neq i}^{n} X_{k,j}$	$\sum_{k=1,k\neq i}^{n} \sum_{l=1,l\neq j}^{n} X_{k,l}$

Next, we performed a one-sided Fisher's exact test on the table. The alternative hy-

pothesis was that $X_{i,j}$ was higher than expected. Then, for each single cell, we performed Benjamini-Hochberg correction on the P-values of all PLA products calculated for that particular cell. A protein complex was considered expressed by a cell if the corrected P-value was below 0.05.

Data Availability

The software was implemented in Python3/Anaconda3 (v4.10.3). The code is available at `https://github.com/tay-lab/Prox-seq_computation`. The raw sequencing data and processed PLA product count data are deposited in NCBI's Gene Expression Omnibus (accession number GSE196130).

5.8 Supplementary Tables

Table 5.1: List of antibodies used in this chapter.

Target	Manufacturer	Catalog #
CD3	Biolegend	300437
CD28	Biolegend	302933
PD1	Biolegend	367402
ICAM1	Biolegend	353102
B7	Biolegend	305202
PDL1	Biolegend	329702
HLADR	Biolegend	307602
CD147	R&D Systems	AF972
Biotin	Novus	NB120-6643

Table 5.2: Prox-seq simulation parameters.

Parameter	Values	Note
R	5000	Cell radius
d_{ligation}	50	Ligation distance
Figures 5.2a, 5.7, 5.9		
$c_{i,j}$	$\begin{bmatrix} 100 & 50 & 0 & 0 & 0 \\ 50 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	True protein complex count
A_i	$\begin{bmatrix} 2000 & 1000 & 500 & 200 & 100 \end{bmatrix}$	Non-interacting probe A count
B_j	$\begin{bmatrix} 2000 & 1000 & 500 & 200 & 100 \end{bmatrix}$	Non-interacting probe B count
Figure 5.4		
$c_{i,j}$	$\begin{bmatrix} 850 & 750 & 0 \\ 750 & 1400 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	True protein complex count
A_i	$\begin{bmatrix} 20 & 15 & 2 \end{bmatrix}$	Non-interacting probe A count
B_j	$\begin{bmatrix} 20 & 15 & 2 \end{bmatrix}$	Non-interacting probe B count
Figure 5.8		
$c_{i,j}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	True protein complex count
A_i	$\begin{bmatrix} 2 & 20.1 & 0.6 \end{bmatrix}$	Non-interacting probe A count
B_j	$\begin{bmatrix} 5.7 & 41.6 & 0.1 \end{bmatrix}$	Non-interacting probe B count
Figure 5.10a		
$c_{i,j}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	True protein complex count
A_i	$\begin{bmatrix} 1000 & 1000 & 100 \end{bmatrix}$	Non-interacting probe A count
B_j	$\begin{bmatrix} 1000 & 1000 & 100 \end{bmatrix}$	Non-interacting probe B count
Figure 5.10b		
$c_{i,j}$	$\begin{bmatrix} 200 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	True protein complex count
A_i	$\begin{bmatrix} 1000 & 1000 & 100 \end{bmatrix}$	Non-interacting probe A count
B_j	$\begin{bmatrix} 1000 & 1000 & 100 \end{bmatrix}$	Non-interacting probe B count

CHAPTER 6

CONCLUSION AND FUTURE OUTLOOK

In this dissertation, we have presented Prox-seq, a novel multiomic sequencing method for quantification of gene expression, protein and protein complex in single cells. By combining proximity ligation assay (PLA) with single-cell RNA-seq (scRNA-seq), Prox-seq produced information on mRNA expression and PLA product levels at the single-cell level, the latter of which was used to infer protein abundance and protein complex information. We then developed FD-seq, a scRNA-seq method for fixed cells, and computational frameworks for simulating and analyzing Prox-seq data. These two advances could contribute to the development of Prox-seq for intracellular proteins.

In Chapter 3, we described the working principle of Prox-seq, and proposed a statistical method, called the iterative method, for predicting protein complex from Prox-seq data. To demonstrate its potentials, we applied Prox-seq to characterized single human peripheral mononuclear blood cells. We showed that Prox-seq could be used to identify well-known immune cell types, such as CD4 and CD8 T cells, and characterized mRNA-protein correlation in single cells [16, 21, 22, 116]. Our iterative method showed that CD9:CD8 is a potential protein interaction on the surface of naive CD8 T cells. We then applied Prox-seq to study primary macrophages after lipopolysaccharide (LPS) and Pam2CSK4 (PAM) stimulation. We found that the effects of LPS and PAM on receptor interactions were additive. The stimulation resulted in lower variance in CD36-related PLA products, suggesting that the ligands resulted in CD36-mediated rearrangement of various surface receptors.

In Chapter 4, we presented FD-seq, a droplet-based scRNA-seq method for paraformaldehyde-fixed and permeabilized cells. FD-seq enabled integration of scRNA-seq with intracellular staining for analysis of rare cell subpopulations. Because intracellular staining was required for detection of intracellular proteins, FD-seq also served as a potential platform to develop Prox-seq for intracellular targets. We applied FD-seq to study KSHV reactivation, a process that occurs in lower than 10% of latently infected cells. We found that upregu-

lation of *TMEM119* was strongly associated with viral reactivation. We then used FD-seq to study OC43 infection, a human betacoronavirus, as a model for SARS-CoV-2 [197]. We showed that a large majority of virus-treated cells only expressed low levels of viral genes, but upregulated pro-inflammatory host genes.

In Chapter 5, we proposed computational frameworks for simulating Prox-seq data, and for protein complex prediction from Prox-seq data. We proposed the free oligo modification to Prox-seq in order to better measure the effect of random ligation, which represented a source of background in Prox-seq data. We then evaluated two methods for protein complex prediction, the iterative method and the linear regression (LR) method. The iterative method relied solely on PLA product counts to predict protein complexes, whereas the LR method utilized the random ligation amount that was enabled by the free oligo modification. We found that the two methods generally agreed with each other. Overall, the iterative method was more conservative in predicting which protein complexes were present in the data, and the LR method was more accurate at quantifying protein complexes in some situations. Therefore, we proposed using both methods for predicting and quantifying protein complexes from Prox-seq data.

Prox-seq had several disadvantages. First, like any antibody-based assays, it was limited by the availability of antibody. Because antibody binding was application-specific, an antibody had to be validated specifically for Prox-seq before usage [223]. On the other hand, protein identification by mass spectrometry (MS) was label-free, meaning that MS-based methods could analyze a significantly higher number of protein targets. However, MS currently had worse sensitivity than antibody-based sequencing assays, and was thus not suitable for characterization of protein-protein interactions in single cells. Second, a protein complex might obscure the epitope of the antibody, leading to false negative results. Third, proximity in Prox-seq might not necessarily imply direct physical interaction, because the ligation distance of Prox-seq was more than 50 nm. Relatively the size of a protein molecule, this is a large distance. In contrast, FRET had a range of 10 nm—the approximate size of an

antibody—and was therefore more suitable to detection of direct protein interactions [107, 224]. Forth, the Prox-seq protocol was harsh to the cells, so cell loss and other unintended harmful effects on the cells needed to be considered.

Currently, Prox-seq was only applied to surface proteins. The intracellular proteins represented a more intricate and interesting environment, with many intracellular protein interactions occurring with different dynamics following a signaling event [3]. The next natural step of this dissertation research is to extend Prox-seq to these intracelullar proteins.

Before that is achieved, we expected that Prox-seq needed further optimization for intracellular protein targets. First, the fixation and permeabilization used in FD-seq might not be optimal for Prox-seq probe binding and the ligation step. Therefore, a different fixation and/or permeabilization methods could be necessary. For example, inCITE-seq used complicated pre-processing steps with custom buffers to adapt CITE-seq to intracellular proteins [193].

Second, intracellular Prox-seq would undoubtedly suffer from the high background from non-specific antibody binding typical of intracellular staining, and the background caused by the DNA oligonucleotides (oligos) on the Prox-seq probes or the leftover unconjugated DNA oligos. Indeed, Srivatsan et al. [225] showed that DNA oligos preferentially bound to the nuclei of permeabilized cells. This background issue could be alleviated by purifying the Prox-seq probes after probe conjugation. Dextran sulfate has also been used to reduced background signal in REAP-seq and inCITE-seq [22, 31, 193]. While our preliminary experiments suggested that dextran sulfate inhibited the activity of the T4 ligase that is used in Prox-seq, this could be circumvented by changing the ligation buffer condition or switching to a different ligase.

Third, although the computational frameworks proposed in this dissertation could be readily modified for intracellular Prox-seq data, experimental validation and calibration are still needed. We also expected that the LR method would prove particularly useful for intracellular Prox-seq. This is because the increase in background would lead to a much

higher amount of random ligation, which is more effectively addressed by the LR method than the iterative method. Another issue was that during a signaling event, many different proteins can interact with one another at the same time, leading to formation of trimers or higher order interactions. This would complicate the simulation model, which at the moment only permitted dimer interaction, and would have unknown effects on the protein complex prediction performance of the iterative and LR methods.

A powerful biological application for intracellular Prox-seq is the NF-κB signaling pathway. This pathway is central to the innate immune response [37]. It involves many well-known protein-protein interactions, and displays many intriguing properties that can only be observed at the single-cell level [3, 226]. Prox-seq is thus well positioned to study these protein interactions, and their effects on the nuclear translocation dynamics of the NF-κB transcription factor. One possible way to biologically validate intracellular Prox-seq is to look for a decrease in the abundance of the protein complex RelA:IκBα upon stimulation (RelA is a subunit of the NF-κB transcription factor). Ideally, we should observe the oscillation of the protein complex that is approximately 180° out-of-phase of that of the nuclear RelA level. Such an oscillation pattern has been extensively characterized [3, 140, 227]. At a later time point following the stimulation, we should also see an increase in interaction between A20 and the proteins upstream of the IκB kinase [228]. Understanding the dynamics of IκBα interaction and A20 interaction, the two main negative feedback loops of NF-κB, is important to mathematical modeling of the NF-κB signaling pathway.

Prox-seq could also be extended to study interactions between proteins and other types of molecules by simply changing the antibody target. For example, one could prepare Prox-seq probes from anti-histone modification antibody (such as trimethylation of histone H3 lysine 4, or H3K4me3 [229]), anti-DNA modification antibody (such as 5-methylcytosine, or 5mC), anti-post-translation modification antibody, or anti-DNA antibody. Alternatively, one of the two Prox-seq probes could be changed to a purely DNA probe to allow hybridization to a DNA sequence of interest, enabling detection of protein interaction with a specific genomic

region.

Recent advances have given rise to single-cell multiomics sequencing, which simultaneously provides different types of information in addition to gene expression at the single-cell level [25–27]. These multimodal approaches enable a more comprehensive understanding of the cell biology, because different biological processes involve different molecules, and because the sparsity of single-cell data, incorporation of multiple data types leads to a deeper and more comprehensive understanding of the cell states [24, 31, 157, 230, 231]. Since Prox-seq was designed around existing scRNA-seq methods, it could readily integrate protein interaction information with existing single-cell multiomics technologies. Future studies could reveal how the additional information of protein interactions can help elucidate the regulatory relationships between different components of the cells, and identify novel and functionally important cell types and cell functions.

REFERENCES

[1] Michael B. Elowitz et al. "Stochastic gene expression in a single cell". *Science* 297.5584 (2002), pp. 1183–1186.

[2] Sabrina L. Spencer et al. "Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis". *Nature* 459.7245 (2009), pp. 428–432.

[3] Savaş Tay et al. "Single-cell NF-kB dynamics reveal digital activation and analogue information processing". *Nature* 466.7303 (2010), pp. 267–271.

[4] Jeremy E. Purvis and Galit Lahav. "Encoding and decoding cellular information through signaling dynamics". *Cell* 152.5 (2013), pp. 945–956.

[5] Antoine-Emmanuel Saliba et al. "Single-cell RNA-seq: advances and future challenges". *Nucleic acids research* 42.14 (2014), pp. 8845–8860.

[6] Alex K. Shalek et al. "Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells". *Nature* 498.7453 (2013), pp. 236–240.

[7] Anoop P. Patel et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". *Science* 344.6190 (2014), pp. 1396–1401.

[8] Cole Trapnell et al. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". *Nature Biotechnology* 32.4 (2014), pp. 381–386.

[9] Alexandra-Chloé Villani et al. "Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors". *Science* 356.6335 (2017), eaah4573.

[10] Akhilesh Pandey and Matthias Mann. "Proteomics to study genes and genomes". *Nature* 405.6788 (2000), pp. 837–846.

[11] Stanley Fields. "Proteomics in genomeland". *Science* 291.5507 (2001), pp. 1221–1224.

[12] Devon M. Cayer, Kristopher L. Nazor, and Nicholas J. Schork. “Mission critical: the need for proteomics in the era of next-generation sequencing and precision medicine”. *Human molecular genetics* 25.R2 (2016), R182–R189.

[13] Tobias Maier, Marc Güell, and Luis Serrano. “Correlation of mRNA and protein in complex biological samples”. *FEBS letters* 583.24 (2009), pp. 3966–3973.

[14] Yuichi Taniguchi et al. “Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells”. *Science* 329.5991 (2010), pp. 533–538.

[15] Gene-Wei Li and X. Sunney Xie. “Central dogma at the single-molecule level in living cells”. *Nature* 475.7356 (2011), pp. 308–315.

[16] Cem Albayrak et al. “Digital quantification of proteins and mRNA in single mammalian cells”. *Molecular cell* 61.6 (2016), pp. 914–924.

[17] Christoph Ziegenhain et al. “Comparative analysis of single-cell RNA sequencing methods”. *Molecular cell* 65.4 (2017), pp. 631–643.

[18] Jiarui Ding et al. “Systematic comparison of single-cell and single-nucleus RNA-sequencing methods”. *Nature Biotechnology* 38.6 (2020), pp. 737–746.

[19] Bogdan Budnik et al. “SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation”. *Genome biology* 19.1 (2018), pp. 1–12.

[20] Harrison Specht et al. “Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2”. *Genome biology* 22.1 (2021), pp. 1–27.

[21] Marlon Stoeckius et al. “Simultaneous epitope and transcriptome measurement in single cells”. *Nature Methods* 14.9 (2017), p. 865.

[22] Vanessa M. Peterson et al. “Multiplexed quantification of proteins and transcripts in single cells”. *Nature Biotechnology* 35.10 (2017), p. 936.

[23] Payam Shahi et al. “Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding”. *Scientific reports* 7.1 (2017), pp. 1–12.

[24] Eleni P. Mimitou et al. “Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells”. *Nature Biotechnology* (2021).

[25] Lia Chappell, Andrew J. C. Russell, and Thierry Voet. “Single-Cell (Multi)omics Technologies”. *Annual Review of Genomics and Human Genetics* 19.1 (2018), pp. 15–41.

[26] Jeongwoo Lee, Do Young Hyeon, and Daehee Hwang. “Single-cell multiomics: technologies and data analysis methods”. *Experimental & Molecular Medicine* 52.9 (2020), pp. 1428–1442.

[27] “Method of the Year 2019: Single-cell multimodal omics”. *Nature Methods* 17.1 (2020), pp. 1–1.

[28] Shuhui Bian et al. “Single-cell multiomics sequencing and analyses of human colorectal cancer”. *Science* 362.6418 (2018), pp. 1060–1063.

[29] Ricard Argelaguet et al. “Multi-omics profiling of mouse gastrulation at single-cell resolution”. *Nature* 576.7787 (2019), pp. 487–491.

[30] Tim Stuart et al. “Single-cell chromatin state analysis with Signac”. *Nature Methods* 18.11 (2021), pp. 1333–1341.

[31] Vivien Marx. “How single-cell multi-omics builds relationships”. *Nature Methods* (2022).

[32] Mario Mellado et al. “Chemokine signaling and functional responses: the role of receptor dimerization and TK pathway activation”. *Annual review of immunology* 19.1 (2001), pp. 397–421.

[33] Sonia Terrillon and Michel Bouvier. “Roles of G-protein-coupled receptor dimerization: From ontogeny to signalling regulation”. *EMBO reports* 5.1 (2004), pp. 30–34.

[34] Joshua N. Leonard et al. “The TLR3 signaling complex forms by cooperative receptor dimerization”. *Proceedings of the National Academy of Sciences* 105.1 (2008), pp. 258–263.

[35] Michael L. Dustin and Jay T. Groves. “Receptor signaling clusters in the immune synapse”. *Annual review of biophysics* 41 (2012), pp. 543–556.

[36] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland science, 2016.

[37] Qian Zhang, Michael J. Lenardo, and David Baltimore. “30 Years of NF-kB: a blossoming of relevance to human pathobiology”. *Cell* 168.1 (2017), pp. 37–57.

[38] Ting Liu et al. “NF-κB signaling in inflammation”. *Signal Transduction and Targeted Therapy* 2.1 (2017), p. 17023.

[39] Erwin L. Van Dijk et al. “Ten years of next-generation sequencing technology”. *Trends in genetics* 30.9 (2014), pp. 418–426.

[40] Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. “A curated database reveals trends in single-cell transcriptomics”. *Database* 2020.baaa073 (2020).

[41] Valentine Svensson, Roser Vento-Tormo, and Sarah A. Teichmann. “Exponential scaling of single-cell RNA-seq in the past decade”. *Nature protocols* 13.4 (2018), pp. 599–604.

[42] Peter V. Kharchenko. “The triumphs and limitations of computational methods for scRNA-seq”. *Nature Methods* (2021), pp. 1–10.

[43] Daniel Ramsköld et al. “Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells”. *Nature Biotechnology* 30.8 (2012), pp. 777–782.

[44] Tamar Hashimshony et al. “CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification”. *Cell reports* 2.3 (2012), pp. 666–673.

[45] Tamar Hashimshony et al. “CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq”. *Genome Biology* 17.1 (2016), p. 77.

[46] Simone Picelli et al. “Full-length RNA-seq from single cells using Smart-seq2”. *Nature protocols* 9.1 (2014), pp. 171–181.

[47] Michael Hagemann-Jensen et al. “Single-cell RNA counting at allele and isoform resolution using Smart-seq3”. *Nature Biotechnology* 38.6 (2020), pp. 708–714.

[48] Johannes W. Bagnoli et al. “Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq”. *Nature communications* 9.1 (2018), pp. 1–8.

[49] Travis K. Hughes et al. “Second-Strand Synthesis-Based Massively Parallel scRNA-Seq Reveals Cellular States and Molecular Features of Human Inflammatory Skin Pathologies”. *Immunity* 53.4 (2020), 878–894.e7.

[50] Fuchou Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. *Nature Methods* 6.5 (2009), pp. 377–382.

[51] John J. Trombetta et al. “Preparation of single-cell RNA-seq libraries for next generation sequencing”. *Current protocols in molecular biology* 107.1 (2014), pp. 4–22.

[52] Fluidigm. *C1 Instrument*. URL: `https://www.fluidigm.com/products-services/instruments/c1` (visited on 02/01/2022).

[53] Saiful Islam et al. “Quantitative single-cell RNA-seq with unique molecular identifiers”. *Nature Methods* 11.2 (2014), p. 163.

[54] Evan Z. Macosko et al. “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. *Cell* 161.5 (2015), pp. 1202–1214.

[55] Allon M. Klein et al. “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells”. *Cell* 161.5 (2015), pp. 1187–1201.

[56] Grace X. Y. Zheng et al. “Massively parallel digital transcriptional profiling of single cells”. *Nature Communications* 8.1 (2017), p. 14049.

[57] Sasan Amini et al. "Haplotype-resolved whole genome sequencing by contiguity preserving transposition and combinatorial indexing". *Nature Genetics* 46.12 (2014), pp. 1343–1349.

[58] Junyue Cao et al. "Comprehensive single-cell transcriptional profiling of a multicellular organism". *Science* 357.6352 (2017), pp. 661–667.

[59] Paul Datlinger et al. "Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing". *Nature Methods* 18.6 (2021), pp. 635–642.

[60] Shia-Yen Teh et al. "Droplet microfluidics". *Lab on a Chip* 8.2 (2008), pp. 198–220.

[61] Xiannian Zhang et al. "Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems". *Molecular cell* 73.1 (2019), pp. 130–142.

[62] Leonard A. Herzenberg et al. "The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford". *Clinical chemistry* 48.10 (2002), pp. 1819–1827.

[63] Mario Roederer. "Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats". *Cytometry Part A* 45.3 (2001), pp. 194–205.

[64] BD Biosciences. *BD FACSymphony^{TM} A5 Cell Analyzer Brochure*. URL: `https://www.bdbiosciences.com/content/dam/bdb/marketing-documents/BD-FACSymphony-Brochure.pdf` (visited on 07/12/2021).

[65] Kathryn Payne et al. "OMIP-063: 28-color flow cytometry panel for broad human Immunophenotyping". *Cytometry Part A* 97.8 (2020), pp. 777–781.

[66] Lily M. Park, Joanne Lannigan, and Maria C. Jaimes. "OMIP-069: forty-color full Spectrum flow cytometry panel for deep Immunophenotyping of major cell subsets in human peripheral blood". *Cytometry Part A* 97.10 (2020), pp. 1044–1051.

[67] Cytek Biosciences. *Cytek® Aurora Brochure*. URL: https://welcome.cytekbio.com/hubfs/Website%20Downloadable%20Content/Brochures/N9-20001_cytek_aurora_brochure.pdf (visited on 01/10/2022).

[68] Alex J. Hughes et al. "Single-cell western blotting". *Nature Methods* 11.7 (2014), pp. 749–755.

[69] Domon Bruno and Aebersold Ruedi. "Mass Spectrometry and Protein Analysis". *Science* 312.5771 (2006), pp. 212–217.

[70] Pieter C. Dorrestein et al. "Mass Spectrometry-Based Label-Free Quantitative Proteomics". *Journal of Biomedicine and Biotechnology* 2010 (2010), p. 840518.

[71] Dayin Lin, David L. Tabb, and John R. Yates. "Large-scale protein identification using mass spectrometry". *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1646.1 (2003), pp. 1–10.

[72] Peng Lu et al. "Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation". *Nature Biotechnology* 25.1 (2007), pp. 117–124.

[73] Johan Malmström et al. "Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans". *Nature* 460.7256 (2009), pp. 762–765.

[74] Björn Schwanhäusser et al. "Global quantification of mammalian gene expression control". *Nature* 473.7347 (2011), pp. 337–342.

[75] Jan C. Rieckmann et al. "Social network architecture of human immune cells unveiled by quantitative proteomics". *Nature Immunology* 18.5 (2017), pp. 583–593.

[76] Thermo Fisher Scientific. *TMTpro^{TM} 18-plex Label Reagent Set*. URL: https://www.thermofisher.com/order/catalog/product/A52045 (visited on 12/09/2021).

[77] Dmitry R. Bandura et al. "Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry". *Analytical Chemistry* 81.16 (2009), pp. 6813–6822.

[78] Sean C. Bendall et al. "Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum". *Science* 332.6030 (2011), pp. 687–696.

[79] Matthew H. Spitzer and Garry P. Nolan. "Mass cytometry: single cells, many features". *Cell* 165.4 (2016), pp. 780–791.

[80] Marcela Alcántara-Hernández et al. "High-dimensional phenotypic mapping of human dendritic cells reveals interindividual variation and tissue specialization". *Immunity* 47.6 (2017), pp. 1037–1050.

[81] Albert G. Tsai et al. "Multiplexed single-cell morphometry for hematopathology diagnostics". *Nature medicine* 26.3 (2020), pp. 408–417.

[82] Fluidigm. *Mass Cytometry Overview*. URL: `https://www.fluidigm.com/products-services/technologies/mass-cytometry` (visited on 01/10/2022).

[83] Guojun Han et al. "Metal-isotope-tagged monoclonal antibodies for high-dimensional mass cytometry". *Nature protocols* 13.10 (2018), pp. 2121–2148.

[84] BioLegend. *Comprehensive Solutions for Single-Cell and Bulk Multiomics*. URL: `https://www.biolegend.com/en-us/totalseq` (visited on 02/08/2022).

[85] Eleni P. Mimitou et al. "Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells". *Nature Methods* 16.5 (2019), pp. 409–412.

[86] BioLegend. *Single-Cell Protein and RNA Panels*. URL: `https://www.biolegend.com/en-us/totalseq/single-cell-rna` (visited on 07/19/2021).

[87] Byungjin Hwang et al. "SCITO-seq: single-cell combinatorial indexed cytometry sequencing". *Nature Methods* 18.8 (2021), pp. 903–911.

[88] Thermo Fisher Scientific. *Pull-Down Assays*. URL: https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/pull-down-assays.html (visited on 12/09/2021).

[89] Thermo Fisher Scientific. *Co-immunoprecipitation (Co-IP)*. URL: https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/co-immunoprecipitation-co-ip.html (visited on 12/09/2021).

[90] Takashi Ito et al. "Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins". *Proceedings of the National Academy of Sciences* 97.3 (2000), pp. 1143–1147.

[91] Takashi Ito et al. "A comprehensive two-hybrid analysis to explore the yeast protein interactome". *Proceedings of the National Academy of Sciences* 98.8 (2001), pp. 4569–4574.

[92] Anna Brückner et al. "Yeast two-hybrid, a powerful tool for systems biology". *International Journal of Molecular Sciences* 10.6 (2009), pp. 2763–2788.

[93] Anne-Claude Gingras et al. "Analysis of protein complexes using mass spectrometry". *Nature Reviews Molecular cell biology* 8.8 (2007), pp. 645–654.

[94] Marco Y. Hein et al. "A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances". *Cell* 163.3 (2015), pp. 712–723.

[95] Anders R. Kristensen, Joerg Gsponer, and Leonard J. Foster. "A high-throughput approach for measuring temporal changes in the interactome". *Nature Methods* 9.9 (2012), pp. 907–909.

[96] Pierre C. Havugimana et al. "A Census of Human Soluble Protein Complexes". *Cell* 150.5 (2012), pp. 1068–1081.

[97] Cuihong Wan et al. “Panorama of ancient metazoan macromolecular complexes”. *Nature* 525.7569 (2015), pp. 339–344.

[98] Lucas ZhongMing Hu et al. “EPIC: software toolkit for elution profile-based inference of protein complexes”. *Nature Methods* 16.8 (2019), pp. 737–742.

[99] Arne H. Smits and Michiel Vermeulen. “Characterizing protein–protein interactions using mass spectrometry: challenges and opportunities”. *Trends in biotechnology* 34.10 (2016), pp. 825–834.

[100] Fridtjof Lund-Johansen, Trung Tran, and Adi Mehta. “Towards reproducibility in large-scale analysis of protein-protein interactions”. *Nature Methods* (2021).

[101] Kyle J. Roux et al. “A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells”. *Journal of Cell Biology* 196.6 (2012), pp. 801–810.

[102] Hyun-Woo Rhee et al. “Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging”. *Science* 339.6125 (2013), pp. 1328–1331.

[103] Dae In Kim et al. “Probing nuclear pore complex architecture with proximity-dependent biotinylation”. *Proceedings of the National Academy of Sciences* 111.24 (2014), E2453–E2461.

[104] Paul R. Selvin. “The renaissance of fluorescence resonance energy transfer”. *Nature Structural and Molecular Biology* 7.9 (2000), p. 730.

[105] Damien Maurel et al. “Cell-surface protein-protein interaction analysis with time-resolved FRET and snap-tag technologies: application to GPCR oligomerization”. *Nature Methods* 5.6 (2008), pp. 561–567.

[106] W. Russ Algar et al. “FRET as a biomolecular research tool-understanding its potential while avoiding pitfalls”. *Nature Methods* 16.9 (2019), pp. 815–829.

[107] Rajesh Babu Sekar and Ammasi Periasamy. “Fluorescence resonance energy transfer (FRET) microscopy imaging of live cell protein localizations”. *Journal of Cell Biology* 160.5 (2003), pp. 629–633.

[108] Steven S. Vogel, B. Wieb van der Meer, and Paul S. Blank. “Estimating the distance separating fluorescent protein FRET pairs”. *Methods* 66.2 (2014), pp. 131–138.

[109] Yang Liu et al. “A Comparative Study of Multivariate and Univariate Hidden Markov Modelings in Time-Binned Single-Molecule FRET Data Analysis”. *The Journal of Physical Chemistry B* 114.16 (2010), pp. 5386–5403.

[110] Sergi Padilla-Parra and Marc Tramier. “FRET microscopy in the living cell: Different approaches, strengths and weaknesses”. *BioEssays* 34.5 (2012), pp. 369–376.

[111] Ola Söderberg et al. “Direct observation of individual endogenous protein complexes in situ by proximity ligation”. *Nature Methods* 3.12 (2006), pp. 995–1000.

[112] Ola Söderberg et al. “Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay”. *Methods* 45.3 (2008), pp. 227–232.

[113] Mats Gullberg et al. “Cytokine detection by antibody-based proximity ligation”. *Proceedings of the National Academy of Sciences* 101.22 (2004), pp. 8420–8424.

[114] Spyros Darmanis et al. “ProteinSeq: High-Performance Proteomic Analyses by Proximity Ligation and Next Generation Sequencing”. *PLOS ONE* 6.9 (2011), pp. 1–10.

[115] M. Fatih Abasıyanık et al. “Ultrasensitive digital quantification of cytokines and bacteria predicts septic shock outcomes”. *Nature communications* 11.1 (2020), pp. 1–12.

[116] Jing Lin et al. “Ultra-sensitive digital quantification of proteins and mRNA in single cells”. *Nature communications* 10.1 (2019), pp. 1–10.

[117] Martin Lundberg et al. "Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood". *Nucleic acids research* 39.15 (2011), e102.

[118] Spyros Darmanis et al. "Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells". *Cell Reports* 14.2 (2016), pp. 380–389.

[119] Alex S. Genshaft et al. "Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction". *Genome Biology* 17.1 (2016), p. 188.

[120] Dominic Grün et al. "Single-cell messenger RNA sequencing reveals rare intestinal cell types". *Nature* 525.7568 (2015), pp. 251–255.

[121] Abbas H. Rizvi et al. "Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development". *Nature Biotechnology* 35.6 (2017), pp. 551–560.

[122] Qingjia Chi, Guixue Wang, and Jiahuan Jiang. "The persistence length and length per base of single-stranded DNA obtained from fluorescence correlation spectroscopy measurements using mean field theory". *Physica A: Statistical Mechanics and its Applications* 392.5 (2013), pp. 1072–1079.

[123] Michael E. Birnbaum et al. "Molecular architecture of the $\alpha\beta$ T cell receptor–CD3 complex". *Proceedings of the National Academy of Sciences* 111.49 (2014), pp. 17576–17581.

[124] Azam Alsemarz, Paul Lasko, and François Fagotto. "Limited significance of the in situ proximity ligation assay". *bioRxiv* (2018).

[125] P. Anton van der Merwe and Omer Dushek. "Mechanisms for T cell receptor triggering". *Nature Reviews Immunology* 11.1 (2011), pp. 47–55.

[126] Jonathan H. Esensten et al. "CD28 Costimulation: From Mechanism to Therapy". *Immunity* 44.5 (2016), pp. 973–988.

[127] Krzysztof M. Zak et al. "Structural Biology of the Immune Checkpoint Receptor PD-1 and Its Ligands PD-L1/PD-L2". *Structure* 25.8 (2017), pp. 1163–1174.

[128] Richard J. Cherry et al. "Detection of Dimers of Dimers of Human Leukocyte Antigen (HLA)-DR on the Surface of Living Cells by Single-Particle Fluorescence Imaging". *Journal of Cell Biology* 140.1 (1998), pp. 71–79.

[129] Jennifer R. Cochran, Thomas O. Cameron, and Lawrence J. Stern. "The Relationship of MHC-Peptide Binding and T Cell Activation Probed Using Chemically Defined MHC Class II Oligomers". *Immunity* 12.3 (2000), pp. 241–250.

[130] P. L. Reilly et al. "The native structure of intercellular adhesion molecule-1 (ICAM-1) is a dimer: Correlation with binding to LFA-1". *The Journal of Immunology* 155.2 (1995), pp. 529–532.

[131] Sumeena Bhatia et al. "Different cell surface oligomeric states of B7-1 and B7-2: Implications for signaling". *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15569–15574.

[132] Oleg V. Kovalenko et al. "Evidence for specific tetraspanin homodimers: inhibition of palmitoylation makes cysteine residues available for cross-linking". *Biochemical Journal* 377.2 (2004), pp. 407–417.

[133] T. L. Collins et al. "p56lck association with CD4 is required for the interaction between CD4 and the TCR/CD3 complex and for optimal antigen stimulation." *The Journal of Immunology* 148.7 (1992), pp. 2159–2162.

[134] Margot Thome et al. "The $p56^{lck}$ SH2 domain mediates recruitment of $CD8/p56^{lck}$ to the activated T cell receptor/CD3/ζ complex". *European Journal of Immunology* 26.9 (1996), pp. 2093–2100.

[135] Kazuhito Toyo-oka et al. "Association of a tetraspanin CD9 with CD5 on the T cell surface: role of particular transmembrane domains in the association". *International Immunology* 11.12 (1999), pp. 2043–2052.

[136] Vera Rocha-Perugini et al. "Tetraspanins CD9 and CD151 at the immune synapse support T-cell integrin signaling". *European Journal of Immunology* 44.7 (2014), pp. 1967–1975.

[137] S. Kerrien et al. "IntAct–open source resource for molecular interaction data". *Nucleic Acids Research* 35.suppl_1 (2007), pp. D561–D565.

[138] Itay Tirosh et al. "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". *Science* 352.6282 (2016), pp. 189–196.

[139] Annette Trickett and Yiu Lam Kwan. "T cell stimulation and expansion using anti-CD3/CD28 beads". *Journal of Immunological Methods* 275.1 (2003), pp. 251–255.

[140] Ryan A. Kellogg et al. "Cellular decision making by non-integrative processing of TLR inputs". *Cell reports* 19.1 (2017), pp. 125–135.

[141] Adrian Ozinsky et al. "The repertoire for pattern recognition of pathogens by the innate immune system is defined by cooperation between Toll-like receptors". *Proceedings of the National Academy of Sciences* 97.25 (2000), p. 13766.

[142] Yuan Qiu et al. "Divergent Roles of Amino Acid Residues Inside and Outside the BB Loop Affect Human Toll-Like Receptor (TLR)2/2, TLR2/1 and TLR2/6 Responsiveness". *PLOS ONE* 8.4 (2013), e61508.

[143] Roy L. Silverstein and Maria Febbraio. "CD36, a Scavenger Receptor Involved in Immunity, Metabolism, Angiogenesis, and Behavior". *Science Signaling* 2.72 (2009), re3–re3.

[144] Tracie A. Seimon et al. "Atherogenic Lipids and Lipoproteins Trigger CD36-TLR2-Dependent Apoptosis in Macrophages Undergoing Endoplasmic Reticulum Stress". *Cell Metabolism* 12.5 (2010), pp. 467–482.

[145] Rafał Biedroń, Angelika Peruń, and Szczepan Józefowski. "CD36 Differently Regulates Macrophage Responses to Smooth and Rough Lipopolysaccharide". *PLOS ONE* 11.4 (2016), pp. 1–26.

[146] Cameron R. Stewart et al. "CD36 ligands promote sterile inflammation through assembly of a Toll-like receptor 4 and 6 heterodimer". *Nature Immunology* 11.2 (2010), pp. 155–161.

[147] Hoang Van Phan et al. "High-throughput RNA sequencing of paraformaldehyde-fixed single cells". *Nature Communications* 12.1 (2021), p. 5636.

[148] Elena Denisenko et al. "Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows". *Genome Biology* 21.1 (2020), p. 130.

[149] David van Dijk et al. "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion". *Cell* 174.3 (2018), 716–729.e27.

[150] Tim J. Stevens et al. "3D structures of individual mammalian genomes studied by single-cell Hi-C". *Nature* 544.7648 (2017), pp. 59–64.

[151] Ying Zhu et al. "Proteomic Analysis of Single Mammalian Cells Enabled by Microfluidic Nanodroplet Sample Preparation and Ultrasensitive NanoLC-MS". *Angewandte Chemie International Edition* 57.38 (2018), pp. 12370–12374.

[152] Jean-François Rual et al. "Towards a proteome-scale map of the human protein-protein interaction network". *Nature* 437.7062 (2005), pp. 1173–1178.

[153] Edward L. Huttlin et al. "The BioPlex Network: A Systematic Exploration of the Human Interactome". *Cell* 162.2 (2015), pp. 425–440.

[154] David M. Grant et al. "Multiplexed FRET to Image Multiple Signaling Events in Live Cells". *Biophysical Journal* 95.10 (2008), pp. L69–L71.

[155] Enfu Hui et al. "T cell costimulatory receptor CD28 is a primary target for PD-1-mediated inhibition". *Science* 355.6332 (2017), pp. 1428–1433.

[156] Haibiao Gong et al. “Simple method to prepare oligonucleotide-conjugated antibodies and its application in multiplex protein detection in single cells”. *Bioconjugate chemistry* 27.1 (2016), pp. 217–225.

[157] Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. *Cell* 184.13 (2021), 3573–3587.e29.

[158] Dvir Aran et al. “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage”. *Nature Immunology* 20.2 (2019), pp. 163–172.

[159] Noa Novershtern et al. “Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis”. *Cell* 144.2 (2011), pp. 296–309.

[160] Jannie Borst et al. “Alternative molecular form of human T cell-specific antigen CD27 expressed upon T cell activation”. *European Journal of Immunology* 19.2 (1989), pp. 357–364.

[161] Miki Hara-Yokoyama et al. “Tetrameric Interaction of the Ectoenzyme CD38 on the Cell Surface Enables Its Catalytic and Raft-Association Activities”. *Structure* 20.9 (2012), pp. 1585–1595.

[162] Iana H. Haralambieva et al. “Genome-wide associations of CD46 and IFI44L genetic variants with neutralizing antibody response to measles vaccine”. *Human Genetics* 136.4 (2017), pp. 421–435.

[163] Maria-Cristina Moldovan et al. “CD4 Dimers Constitute the Functional Component Required for T Cell Activation”. *The Journal of Immunology* 169.11 (2002), p. 6261.

[164] Irwin I. Singer et al. “CCR5, CXCR4, and CD4 Are Clustered and Closely Apposed on Microvilli of Human Macrophages and T Cells”. *Journal of Virology* 75.8 (2001), pp. 3779–3790.

[165] Jie Geng and Malini Raghavan. “CD8$\alpha\alpha$ homodimers function as a coreceptor for KIR3DL1”. *Proceedings of the National Academy of Sciences* 116.36 (2019), p. 17951.

[166] Eric Rubinstein et al. "CD9, CD63, CD81, and CD82 are components of a surface tetraspan network connected to HLA-DR and VLA integrins". *European Journal of Immunology* 26.11 (1996), pp. 2657–2665.

[167] Jinhai Wang et al. "Dimerization of CXCR4 in living malignant cells: control of cell migration by a synthetic peptide that reduces homologous CXCR4 interactions". *Molecular Cancer Therapeutics* 5.10 (2006), p. 2474.

[168] Silvia Deaglio et al. "Human CD38 and CD16 are functionally dependent and physically associated in natural killer cells". *Blood* 99.7 (2002), pp. 2490–2498.

[169] Mayumi Suzuki et al. "Tetraspanin CD9 Negatively Regulates Lipopolysaccharide-Induced Macrophage Activation and Lung Inflammation". *The Journal of Immunology* 182.10 (2009), p. 6485.

[170] Marie-Thérèse Zilber et al. "MHC class II/CD38/CD9: a lipid-raft-dependent signaling complex in human monocytes". *Blood* 106.9 (2005), pp. 3074–3081.

[171] Mitsue Kurita-Taniguchi et al. "Molecular assembly of CD46 with CD9, alpha3-beta1 integrin and protein tyrosine phosphatase SHP-1 in human macrophages through differentiation by GM-CSF". *Molecular Immunology* 38.9 (2002), pp. 689–700.

[172] Martin Becker et al. "CXCR4 signaling and function require the expression of the IgD-class B-cell antigen receptor". *Proceedings of the National Academy of Sciences* 114.20 (2017), p. 5231.

[173] Kevin R. Bobbitt and Louis B. Justement. "Regulation of MHC class II signal transduction by the B cell coreceptors CD19 and CD22". *The Journal of Immunology* 165.10 (2000), pp. 5588–5596.

[174] Gamal Badr, Eric A. Lefevre, and Mohamed Mohany. "Thymoquinone Inhibits the CXCL12-Induced Chemotaxis of Multiple Myeloma Cells and Increases Their Susceptibility to Fas-Mediated Apoptosis". *PLOS ONE* 6.9 (2011), e23741.

[175] Noa B. Martín-Cófreces et al. "End-binding protein 1 controls signal propagation from the T cell receptor". *The EMBO Journal* 31.21 (2012), pp. 4140–4152.

[176] Xiaodong Xiao et al. "Interactions of CCR5 and CXCR4 with CD4 and gp120 in Human Blood Monocyte-Derived Dendritic Cells". *Experimental and Molecular Pathology* 68.3 (2000), pp. 133–138.

[177] A. O. Watts et al. "Identification and profiling of CXCR3-CXCR4 chemokine receptor heteromer complexes". *British Journal of Pharmacology* 168.7 (2013), pp. 1662–1674.

[178] J. Vinet et al. "Inhibition of CXCR3-mediated chemotaxis by the human chemokine receptor-like protein CCX-CKR". *British Journal of Pharmacology* 168.6 (2013), pp. 1375–1387.

[179] Qingrong Yan et al. "Structure of CD84 provides insight into SLAM family function". *Proceedings of the National Academy of Sciences* 104.25 (2007), p. 10583.

[180] Dmitriy Mazurov, Lubov Barbashova, and Alexander Filatov. "Tetraspanin protein CD9 interacts with metalloprotease CD10 and enhances its release via exosomes". *The FEBS Journal* 280.5 (2013), pp. 1200–1213.

[181] Aron Chakera et al. "The Duffy Antigen/Receptor for Chemokines Exists in an Oligomeric Form in Living Cells and Functionally Antagonizes CCR5 Signaling through Hetero-Oligomerization". *Molecular Pharmacology* 73.5 (2008), p. 1362.

[182] Liying Yan et al. "Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells". *Nature Structural & Molecular Biology* 20.9 (2013), pp. 1131–1139.

[183] Elliot R. Thomsen et al. "Fixed single-cell transcriptomic characterization of human radial glial diversity". *Nature Methods* 13 (2015), p. 87.

[184] Yonatan Katzenelenbogen et al. "Coupled scRNA-Seq and Intracellular Protein Activity Reveal an Immunosuppressive Role of TREM2 in Cancer". *Cell* 182.4 (2020), 872–885.e19.

[185] Shohei Hori, Takashi Nomura, and Shimon Sakaguchi. "Control of regulatory T cell development by the transcription factor Foxp3". *Science* 299.5609 (2003), pp. 1057–1061.

[186] Tobias Brambrink et al. "Sequential Expression of Pluripotency Markers during Direct Reprogramming of Mouse Somatic Cells". *Cell Stem Cell* 2.2 (2008), pp. 151–159.

[187] Jonathan Alles et al. "Cell fixation and preservation for droplet-based single-cell transcriptomics". *BMC Biology* 15.1 (2017), p. 44.

[188] Jinguo Chen et al. "PBMC fixation and processing for Chromium single-cell RNA sequencing". *Journal of Translational Medicine* 16.1 (2018), p. 198.

[189] Agnese A. Pollice et al. "Sequential paraformaldehyde and methanol fixation for simultaneous flow cytometric analysis of DNA, cell surface proteins, and intracellular proteins". *Cytometry* 13.4 (1992), pp. 432–444.

[190] Peter O. Krutzik and Garry P. Nolan. "Intracellular phospho-protein staining techniques for flow cytometry: monitoring single cell signaling events". *Cytometry Part A: the journal of the International Society for Analytical Cytology* 55.2 (2003), pp. 61–70.

[191] Rob W. M. Hoetelmans et al. "Effects of Acetone, Methanol, or Paraformaldehyde on Cellular Structure, Visualized by Reflection Contrast Microscopy and Transmission and Scanning Electron Microscopy". *Applied Immunohistochemistry & Molecular Morphology* 9.4 (2001).

[192] Thomas Jung et al. "Detection of intracellular cytokines by flow cytometry". *Journal of Immunological Methods* 159.1 (1993), pp. 197–207.

[193] Hattie Chung et al. "Joint single-cell measurements of nuclear proteins and RNA in vivo". *Nature Methods* 18.10 (2021), pp. 1204–1212.

[194] Kwun Wah Wen and Blossom Damania. "Kaposi sarcoma-associated herpesvirus (KSHV): Molecular biology and oncogenesis". *Cancer Letters* 289.2 (2010), pp. 140–150.

[195] Ethel Cesarman. "Gammaherpesviruses and Lymphoproliferative Disorders". *Annual Review of Pathology: Mechanisms of Disease* 9.1 (2014), pp. 349–372.

[196] Oliver Manners et al. "Contribution of the KSHV and EBV lytic cycles to tumourigenesis". *Current Opinion in Virology* 32 (2018), pp. 60–70.

[197] Nir Drayman et al. "Masitinib is a broad coronavirus 3CL inhibitor that blocks replication of SARS-CoV-2". *Science* 373.6557 (2021), pp. 931–936.

[198] Carolina Arias et al. "KSHV 2.0: A Comprehensive Annotation of the Kaposi's Sarcoma-Associated Herpesvirus Genome Using Next-Generation Sequencing Reveals Novel Genomic and Functional Features". *PLOS Pathogens* 10.1 (2014), pp. 1–23.

[199] Nir Drayman et al. "HSV-1 single-cell analysis reveals the activation of anti-viral and developmental programs in distinct sub-populations". *eLife* 8 (2019), e46339.

[200] Jennifer N. Berger et al. "Redefining De Novo Gammaherpesvirus Infection Through High-Dimensional, Single-Cell Analysis of Virus and Host". *bioRxiv* (2020).

[201] Britt Glaunsinger and Don Ganem. "Lytic KSHV Infection Inhibits Host Gene Expression by Accelerating Global mRNA Turnover". *Molecular Cell* 13.5 (2004), pp. 713–723.

[202] Jeffrey Vieira and Patricia M. O'Hearn. "Use of the red fluorescent protein as a marker of Kaposi's sarcoma-associated herpesvirus lytic gene expression". *Virology* 325.2 (2004), pp. 225–240.

[203] Julien R. St-Jean et al. "Human Respiratory Coronavirus OC43: Genetic Stability and Neuroinvasion". *Journal of Virology* 78.16 (2004), pp. 8824–8834.

[204] Katia De Filippo et al. "Mast cell and macrophage chemokines CXCL1/CXCL2 control the early stage of neutrophil recruitment during tissue inflammation". *Blood* 121.24 (2013), pp. 4930–4937.

[205] Chao Shi and Eric G. Pamer. "Monocyte recruitment during infection and inflammation". *Nature Reviews Immunology* 11.11 (2011), pp. 762–774.

[206] Uku Raudvere et al. "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)". *Nucleic Acids Research* 47.W1 (2019), W191–W198.

[207] Gioele La Manno et al. "RNA velocity of single cells". *Nature* 560.7719 (2018), pp. 494–498.

[208] Hoang Van Phan et al. *Droplet-based single-cell RNA sequencing of paraformaldehyde-fixed cells*. Protocol Exchange. Oct. 7, 2021. URL: `https://doi.org/10.21203/rs.3.pex-1604/v1` (visited on 10/07/2021).

[209] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21.

[210] Kevin F. Brulois et al. "Construction and Manipulation of a New Kaposi's Sarcoma-Associated Herpesvirus Bacterial Artificial Chromosome Clone". *Journal of Virology* 86.18 (2012), pp. 9708–9720.

[211] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[212] Xiaojie Qiu et al. "Single-cell mRNA quantification and differential analysis with Census". *Nature Methods* 14.3 (2017), pp. 309–315.

[213] Xiaojie Qiu et al. "Reversed graph embedding resolves complex single-cell trajectories". *Nature Methods* 14.10 (2017), pp. 979–982.

[214] Andrew Butler et al. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". *Nature Biotechnology* 36.5 (2018), pp. 411–420.

[215] Tim Stuart et al. "Comprehensive Integration of Single-Cell Data". *Cell* 177.7 (2019), 1888–1902.e21.

[216] Viktor Petukhov et al. "dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments". *Genome Biology* 19.1 (2018), p. 78.

[217] Sigma. *Duolink PLA Fluorescence Protocol*. URL: `https://www.sigmaaldrich.com/US/en/technical-documents/protocol/protein-biology/protein-and-nucleic-acid-interactions/duolink-fluorescence-user-manual` (visited on 02/02/2022).

[218] Axel Klaesson et al. "Improved efficiency of in situ protein analysis by proximity ligation using UnFold probes". *Scientific Reports* 8.1 (2018), p. 5400.

[219] Kara L. Johnson et al. "Revealing protein-protein interactions at the transcriptome scale by sequencing". *Molecular Cell* 81.19 (2021), 4091–4103.e9.

[220] J. Miller et al. "Intercellular adhesion molecule-1 dimerization and its consequences for adhesion mediated by lymphocyte function associated-1". *Journal of Experimental Medicine* 182.5 (1995), pp. 1231–1241.

[221] Wolfram MathWorld. *Sphere Point Picking*. URL: `https://mathworld.wolfram.com/SpherePointPicking.html` (visited on 02/04/2022).

[222] Wolfram MathWorld. *Sphere Line Picking*. URL: `https://mathworld.wolfram.com/SphereLinePicking.html` (visited on 02/04/2022).

[223] Mathias Uhlen et al. "A proposal for validation of antibodies". *Nature Methods* 13.10 (2016), pp. 823–827.

[224] Michael Reth. "Matching cellular dimensions with molecular sizes". *Nature Immunology* 14.8 (2013), pp. 765–767.

[225] Sanjay R. Srivatsan et al. "Massively multiplex chemical transcriptomics at single-cell resolution". *Science* 367.6473 (2020), pp. 45–51.

[226] Ryan A. Kellogg and Savaş Tay. "Noise facilitates transcriptional control under dynamic inputs". *Cell* 160.3 (2015), pp. 381–392.

[227] Ryan A. Kellogg et al. "Digital signaling decouples activation probability and population heterogeneity". *elife* 4 (2015), e08931.

[228] Noula Shembade and Edward W. Harhaj. "Regulation of NF-κB signaling by the A20 deubiquitinase". *Cellular & Molecular Immunology* 9.2 (2012), pp. 123–130.

[229] Eva Bártová et al. "Histone Modifications and Nuclear Architecture: A Review". *Journal of Histochemistry & Cytochemistry* 56.8 (2008), pp. 711–721.

[230] Iain C. Macaulay, Chris P. Ponting, and Thierry Voet. "Single-Cell Multiomics: Multiple Measurements from Single Cells". *Trends in Genetics* 33.2 (2017), pp. 155–168.

[231] Adam Gayoso et al. "Joint probabilistic modeling of single-cell multi-omic data with totalVI". *Nature Methods* 18.3 (2021), pp. 272–282.