PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Computational prediction of protein interactions in single cells by proximity sequencing

Junjie Xia[1☯], Hoang Van Phan[1☯¤a], Luke Vistain[1☯¤b], Mengjie Chen[2,3], Aly A. Khan[4], Savaş Tay[1]*

1 Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois, United States of America, **2** Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, Illinois, United States of America, **3** Department Human Genetics, The University of Chicago, Chicago, Illinois, United States of America, **4** Department of Pathology, The University of Chicago, Chicago, Illinois, United States of America

☯ These authors contributed equally to this work.
¤a Current Address: Present address: Division of Infectious Disease, University of California, San Francisco, California, United States of America
¤b Current Address: Present address: Lymphocyte Biology Section, Laboratory of Immune Systems Biology, NIAID, NIH, Bethesda, Maryland, United States of America
* tays@uchicago.edu

Abstract

Proximity sequencing (Prox-seq) simultaneously measures gene expression, protein expression and protein complexes on single cells. Using information from dual-antibody binding events, Prox-seq infers surface protein dimers at the single-cell level. Prox-seq provides multi-dimensional phenotyping of single cells in high throughput, and was recently used to track the formation of receptor complexes during cell signaling and discovered a novel interaction between CD9 and CD8 in naïve T cells. The distribution of protein abundance can affect identification of protein complexes in a complicated manner in dual-binding assays like Prox-seq. These effects are difficult to explore with experiments, yet important for accurate quantification of protein complexes. Here, we introduce a physical model of Prox-seq and computationally evaluate several different methods for reducing background noise when quantifying protein complexes. Furthermore, we developed an improved method for analysis of Prox-seq data, which resulted in more accurate and robust quantification of protein complexes. Finally, our Prox-seq model offers a simple way to investigate the behavior of Prox-seq data under various biological conditions and guide users toward selecting the best analysis method for their data.

Author summary

We introduce a physical model for protein complexes at the cell membrane and report a systematic study of statistical and computational methods for their measurements using proximity sequencing.

OPEN ACCESS

Citation: Xia J, Phan HV, Vistain L, Chen M, Khan AA, Tay S (2024) Computational prediction of protein interactions in single cells by proximity sequencing. PLoS Comput Biol 20(3): e1011915. https://doi.org/10.1371/journal.pcbi.1011915

Editor: Shihua Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Science, CHINA

Received: August 3, 2023

Accepted: February 13, 2024

Published: March 14, 2024

Data Availability Statement: All code is implemented in Python3/Anaconda3 (v4.10.3). The code is deposited 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).

Funding: S.T. was awarded the NIH grants GM127527, R35GM148231, and a Paul G. Allen Distinguished Investigator Award, which supported this work. M.C. was awarded NIH grants

GM126553 and HG011883, and NSF grants 2016307 and 2016307, which supported this work. This work was supported in part by the Intramural Research program of NIAID, NIH (L.V.). L.V. received salary from the NIH. The sponsors did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Advances in single cell sequencing have enabled unprecedented analyses of cellular heterogeneity in complex biological systems [1,2]. Single-cell RNA sequencing [3] (scRNA-seq) is among the most widely used methods. However, because proteins are the effector molecules for the majority of biological functions, RNA data alone is not sufficient to investigate these protein functions thoroughly. Signaling events, for example, typically begin with receptor clustering, protein phosphorylation, and other protein-protein interactions, all of which occur prior to transcription.

To investigate the roles of protein interactions in greater depth, we recently developed a method called proximity sequencing (Prox-seq) for simultaneous quantification of mRNA, surface proteins and protein complexes at the single-cell level [4] Prox-seq captures protein complex information in barcoded DNA oligonucleotides (oligos) using a proximity ligation assay [5,6] (PLA). Each protein in Prox-seq is targeted by two DNA-conjugated antibodies, called Prox-seq probes A and B (Fig 1a). The DNA oligos on probes A and B are ligated only if two protein molecules are sufficiently close to each other. The result of this ligation is referred to as a "PLA product." The ligation distance is expected to be 50-70nm [7]. In order to generate a PLA product, the oligo belonging to a Prox-seq probe A must ligate to the oligo belonging to a Prox-seq probe B. Importantly, unligated probes do not contribute to the signal because both library preparation and sequence alignment require barcodes from both the A and B probe. Upon sequencing, the number of PLA products can be determined by counting the number of unique molecular identifiers (UMIs). Because of this design, the number of PLA products measured for a protein is a reflection of both the abundance of that protein and the availability of nearby Prox-seq probes. By combining Prox-seq with scRNA-seq, these PLA products can be sequenced alongside complementary DNA (cDNA) libraries, providing information on gene expression, protein abundances, and protein complex formation from single cells [4].

Prox-seq protein data contains a unique source of background noise, namely the ligation of two protein molecules that do not functionally interact but are nevertheless sufficiently close to each other by random chance. We call this effect "proximity noise" (Fig 1b). Proximity noise exists because the average distance between probes on the cell surface decreases with increasing protein abundance (see Methods). A previous study showed that proximity noise

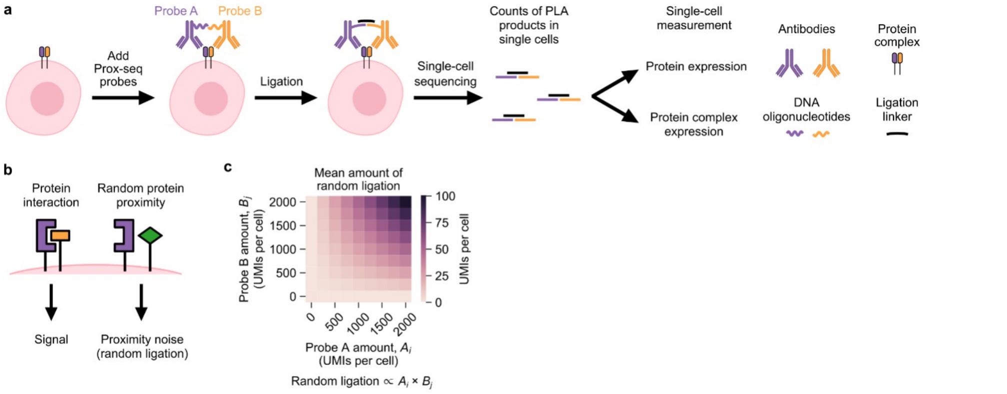

Fig 1. Working principle of Prox-seq and identification of proximity noise. (a) Schematic showing the main steps of Prox-seq. (b) Schematic showing the background in Prox-seq that is caused by proximity noise (random ligation of non-interacting protein molecules). (c) Heatmap showing the expected amount of proximity noise created from simulations of two protein molecules at varying expression levels. By modeling the mean amount of proximity noise with a binomial distribution (see Methods), we found that it was proportional to the product of the abundances of the two protein molecules.

https://doi.org/10.1371/journal.pcbi.1011915.g001

led to false positive detection of protein interactions for in situ PLA [8]. A theoretical model showed that the mean amount of proximity noise is proportional to the product of the expression levels of the two proteins that made up the PLA product (Fig 1c). In short, the presence of PLA products for a specific pair of proteins does not guarantee that the two proteins functionally interact and form stable complexes.

To account for PLA products generated by proximity noise, we previously proposed and used a statistical method, termed the iterative method, to differentiate protein complexes from random ligation in PLA product counts [4]. Initially, this method establishes an "expected value" for each PLA product, representing the number of PLA products that would exist if Prox-seq probes were randomly distributed across the cell surface. Subsequently, the method subtracts the expected background from the PLA product counts. If a PLA product's count exceeds its expected value, the difference between observed and expected PLA products is attributed to non-random protein complexes. This procedure is iteratively executed for every type of PLA product in each individual cell. Although this method successfully recovered positive controls of known protein complexes, assessing its performance on experimental data is challenging, as Prox-seq datasets lack comprehensive knowledge of the entire set of protein complexes and their expression levels.

In this study, we present a simulation model for single-cell proteomic data in proximity sequencing experiments and use it to computationally benchmark the performance of several new and existing protein complex prediction methods. After calibrating the model with experimental data, the simulation model allowed us to quantitatively analyze the proximity noise and its effects on the measured PLA product counts. We compare the performance of three methods: the iterative method, a new linear regression-based method, and a new ensemble method that combines the two. We find that, while both iterative and linear regression-based methods perform well in several different scenarios, combining them into the ensemble method yielded the most accurate and robust quantification of protein complexes. These results shed insight onto how the co-localization of surface proteins translate into Prox-seq data and provides guidelines for use of Prox-seq and related dual-binding technologies for multi-omic analysis of single cells.

Results

Overview of the simulation model

Based on a physical model of how PLA products are formed in each single cell, we created a simulation model of PLA product count data. We reasoned that proximity alone would determine if a Prox-seq probe A and a Prox-seq probe B ligate and produce a PLA product. We constructed the simulation model in a way that allowed us to simulate probes that bind to non-interacting protein molecules (proteins that are not part of a complex) separately from probes that bind to interacting molecules (proteins that are part of a complex). This procedure enabled us to independently tune the abundance of proteins and protein complexes in the simulation, and to observe how these properties affected Prox-seq data.

First, we generated the non-interacting Prox-seq probes A as random points on a sphere (Fig 2a). These points indicated that the protein molecules exist as monomers, that their complex partners were not targeted by the Prox-seq probe panel, or that they were caused by non-specific antibody binding. Further, we assumed such protein monomers were distributed randomly on the cell surface. Then, we repeated the process to generate the non-interacting Prox-seq probes B signal. Second, we generated the interacting Prox-seq probes A and B by generating a sphere of random points. These points corresponded to detectable protein complexes. Because these two probes A and B both bound to the same protein complexes, the Prox-seq

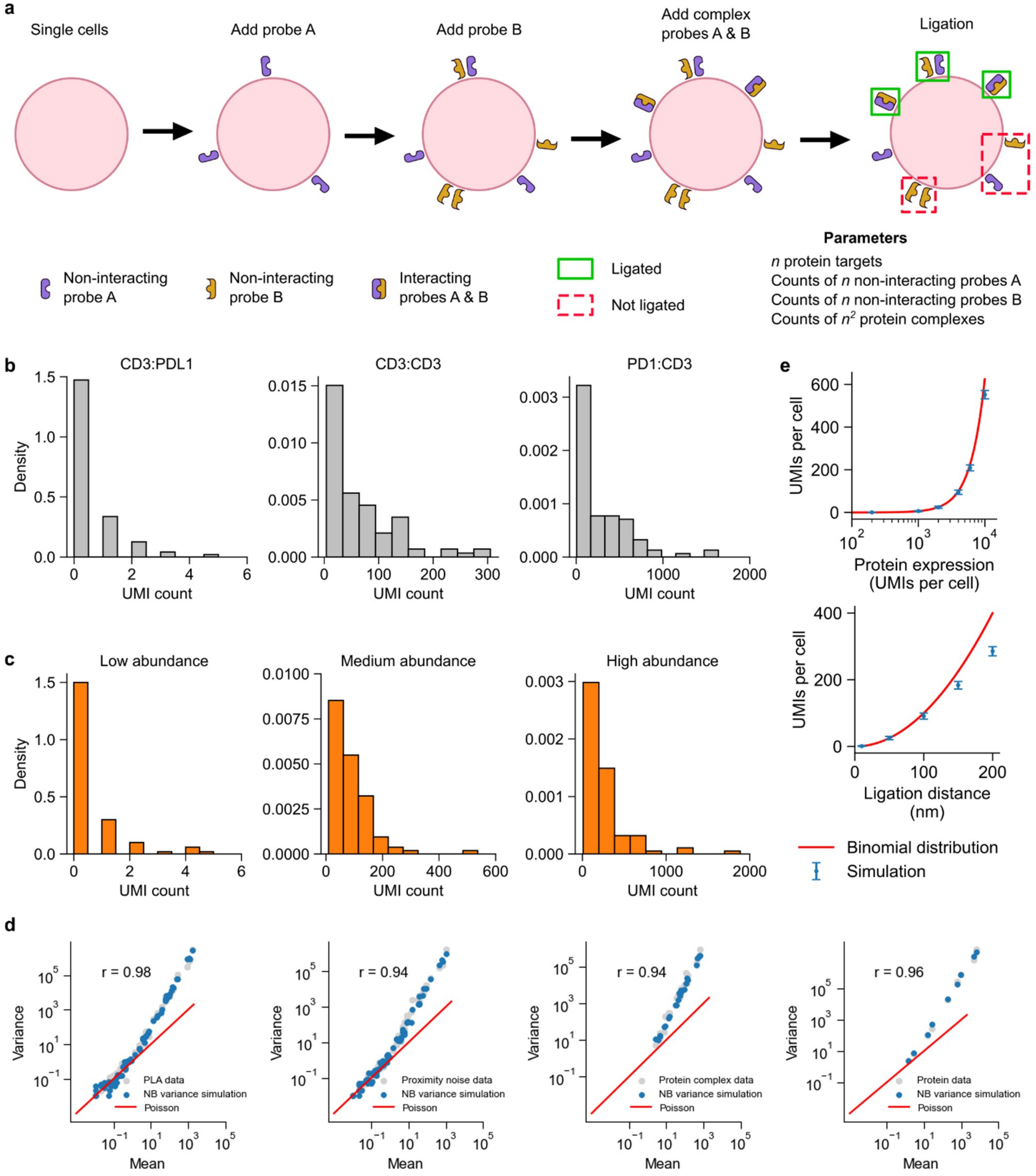

Fig 2. Overview and calibration of simulated Prox-seq data. (a) Schematic for the simulation model of PLA products. The simulation was separately performed on a cell-by-cell basis. First, a number of non-interacting probes A and non-interacting probes B were added as random points on a sphere. Next, a number of protein complexes were added as random points on a sphere. These points corresponded to probes A and B that bound to interacting protein molecules. Finally, probes A and B that had a Euclidean distance lower than the ligation distance were ligated, thus creating 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 with NB variance. (d) Scatter plots of mean-variance relationship show how negative binomial variance captures overdispersion in PLA data, proximity noise data, protein complex data, and protein data. (e) The relationship between proximity noise (measured as UMI counts) and protein abundance (top) or ligation distance (bottom). Please refer to the Methods section for derivation of the binomial distribution approximation.

https://doi.org/10.1371/journal.pcbi.1011915.g002

probe A points would necessarily be in proximity with their corresponding Prox-seq probe B points. Finally, any pairs of probe A and B with Euclidean distances less than the ligation distance were considered ligated and produced PLA products (see Methods). If a probe A was within the ligation distance with more than one probes B, one such probe B was chosen at random to ligate with said probe A.

We next compared the simulated Prox-seq data to the experimental data. We analyzed T cells (Jurkat cell line) and B cells (Raji cell line) with a panel of Prox-seq probes that targeted both T cell and B cell markers from a previously reported study[4]. Simulated counts of PLA product and protein expression followed the Poisson distribution, whereas the experimental data exhibited overdispersion (S1a and S2a Figs). We found that adding variance in the form of a negative binomial distribution (NB) for non-interacting probes and protein complexes was sufficient to capture the overdispersion of the real data (NB variance, see Methods). With the added NB variance, the simulated data, like the experimental data, had a right-skewed distribution across different PLA product abundances (Fig 2b and 2c). Notably, the simulation model with added variance captured the positive correlation between observed PLA product count and non-proximal probe count in real data (S1b–S1g Fig). The simulation model with no variance, however, showed a negative correlation between PLA product count and non-proximal probe count (S1d and S1e Fig). The NB variance model also produced non-proximal probe counts with similar distributions to those observed in experimental data (S2 Fig).

We generated replicated datasets by sampling from the fitted model for posterior predictive checks (PPCs) [9]. We then assessed how well these data samplings maintained the properties of the observed data with two metrics. First, we measured the similarity between the coefficient of variation per PLA product, proximity noise, protein complex and protein. This comparison enables evaluation of how well the mean-variance relationship of real data is preserved (Fig 2d & S3a Fig). Second, we perform Mann-Whitney U-test statistic to measure the extent to which the replicated data and raw data come from the same distribution (S3b Fig). Finally, we characterized the amount of proximity noise in the most basic scenario when there were no protein complexes detectable by the Prox-seq probe panel. The simulation demonstrated that the amount of PLA product produced by random ligation scales quadratically with both protein abundance and ligation distance (Fig 2e). These results show that our model and simulations faithfully capture key aspects of real Prox-seq data in single cells and reiterates the importance of identifying and removing proximity noise, which can especially be large for highly expressed proteins.

Simulation of non-specific antibody binding in Prox-seq and heterogenous cell clustering by simulated PLA data

Nonspecific antibody binding occurs when an antibody binds to a cell that does not have an epitope for that antibody. This is a potential problem encountered in every antibody-based proteomics technology. The challenge of nonspecific staining becomes more complicated in Prox-seq due to its reliance on dual-binding events. For a PLA product, it can be categorized into three possible binding cases: nonspecific binding (both binding events of probe A and probe B are nonspecific), one-specific binding (only one probe is bound to its target), and both-specific binding (both probes are bound to their target). Clearly, we only desire both-specific binding PLA data for downstream analysis. To estimate nonspecific antibody binding within experiments, we include isotype control antibodies with oligonucleotide conjugation in both probe panels. This allows us to directly define the nonspecific binding distribution from the observed data. Given a probability of nonspecific binding for each antibody, we can use the simulation model to generate PLA data that recapitulates the properties of three binding cases in real data (S4 Fig). We found that the issue of nonspecific antibody binding is negligible in current extracellular Prox-seq assays and is much less important than proximity noise. For example, for a highly abundant PLA product CD147:CD147 in Jurkat cells, nonspecific PLA counts constitute less than one percent of both-specific binding PLA counts (S4 Fig). However, we anticipate that the significance of nonspecific antibody binding may escalate in intracellular Prox-seq experiments [10], which necessitate the analysis of denoising nonspecific binding before reliably predicting protein interactions. While our current extracellular assays are not heavily impacted by nonspecific binding, the potential challenges posed in intracellular experiments underscore the importance of refining and validating denoising methods for comprehensive and accurate analysis in future studies. Given the limited data availability of PLA datasets, our simulation model and synthetic datasets can be served as crucial tools to benchmark background staining denoise models for Prox-seq.

PLA data from Prox-seq introduces a new modality to single-cell omics research. In our simulation, we generated three distinct PLA datasets, each representing a unique cell type characterized by same protein expressions but different protein complex expressions. All three cell types express same abundance of proteins 1, 2, and 3. In specific details, cell type 1 exhibits the presence of the protein 1:2 heterodimer and the protein 1 homodimer. Conversely, cell type 2 exclusively expresses the protein 2:3 heterodimer. Lastly, cell type 3 features the protein 1:3 heterodimer along with the protein 1 homodimer. Through unsupervised clustering based on PLA features, we observed a high correlation with the known characteristics of the cell types (S5 Fig). This simple simulation study underscores the potential of utilizing PLA data to identify cell types by their protein complex arrangement.

Iterative prediction of protein complex abundance

An iterative method was used to previously identify the existence of stable protein complexes in Prox-seq measurements. This method 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 (see Methods). This calculation resulted in an expected random count for PLA products that represents the PLA count caused by proximity noise. 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 non-random protein interaction. To quantify the protein complexes on each single cell, we calculated the difference between the observed and expected random PLA product count (Fig 3a). This method was called the iterative method, because it involved solving a system of quadratic equations (describing all possible protein dimers) iteratively (see Methods) [4]. This method relied on the fact that Prox-seq can measure protein abundance, similar to flow cytometry and CITE-seq [11]. The abundance of a protein was the amount of protein molecules that were present on the cell surface, and therefore included both molecules in monomeric and complex forms. In our previous study [4], we proposed that the protein abundance could be estimated from Prox-seq data by summing the appearances of each protein across its associated PLA products (see Methods). Here, we find by using our simulated data that such an estimate is a good approximation of the true protein abundance, as they are strongly correlated (S6 Fig).

To further examine the assumptions underlying the iterative method, we now construct the following simulation scenario: The simulation had three protein targets, called protein 1, protein 2 and protein 3. These proteins did not interact with themselves, nor with any other proteins. Furthermore, protein 3 had a lower non-interacting probe count (mean of 100 UMIs/cell compared to 1000 UMIs/cell for proteins 1 and 2, S1 Table). Simulated data showed that our assumptions behind the iterative method were correct. When there were no interactions

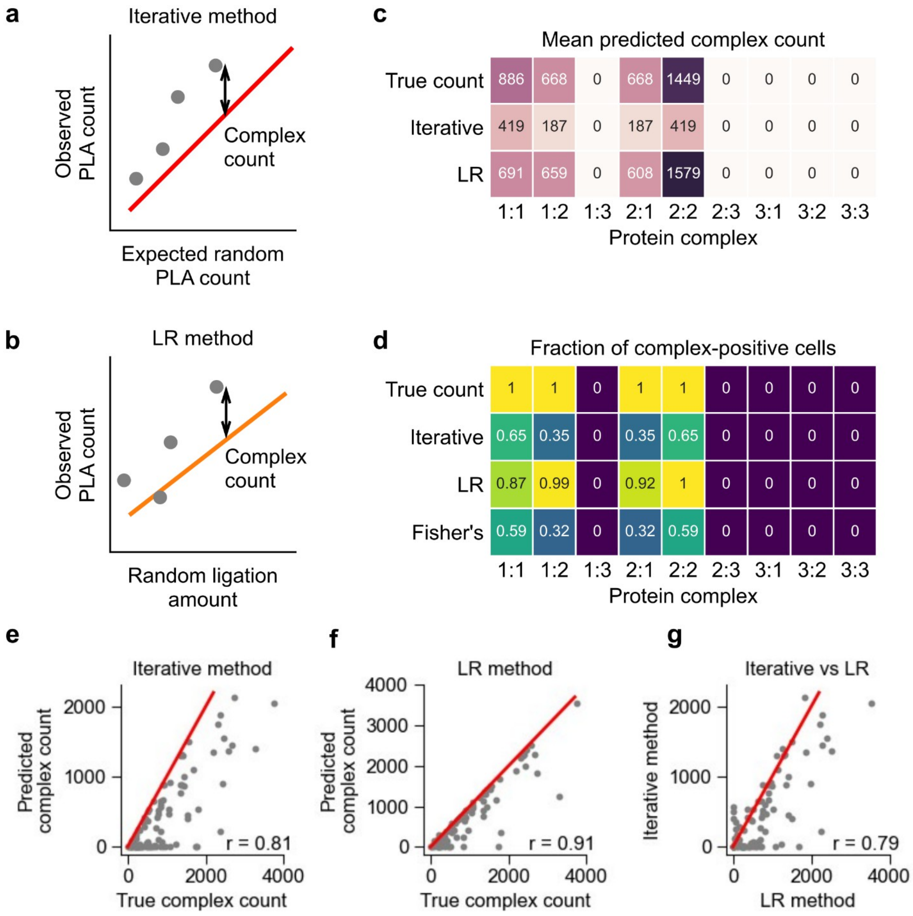

Fig 3. Comparison between the iterative and linear regression (LR) methods for protein complex prediction in simulated data. (a, b) Schematics showing the working principle of (a) the iterative method and (b) the LR method. In the iterative method, the protein complex count is the difference between the observed and expected PLA product count. In the LR method, the protein complex count is the difference between the observed PLA product count and its expected amount of random ligation, which is calculated from the non-proximal 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 represents the ground truth of 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.

https://doi.org/10.1371/journal.pcbi.1011915.g003

between the proteins, the observed PLA product counts were similar to the expected random count (S7a Fig). When we introduced the protein complex 1:1 to the simulation while keeping the other parameters the same, the observed counts of the PLA product 1:1 was higher than its expected random count (S7b Fig).

One weakness of the iterative method is complexity of hyper-parameter tuning, which can result in sub-optimal convergence. They key parameter is the initialization setting, which are the initial estimates of protein complex abundances. By default, the algorithm assigns and initial value of 0 to all protein complexes. However, different initialization settings will influence iterative behaviors to convergence, as well as tolerance (S8a Fig). Unsensible initialization tends to generate nonsensical predictive outputs (S8b Fig). This led us to consider more robust methods for protein complex quantification.

Prediction of protein complex abundance using linear regression

To address the limitations of the iterative method we developed a new approach (the linear regression—LR method). This method uses an experimentally modified Prox-seq procedure that enables direct measurement of Prox-seq probes that were not ligated because they were not proximal to another Prox-seq probe (we refer to these as non-proximal probes) [4]. The proximity noise for a PLA product i:j should be proportional to the product of the non-proximal probe A targeting protein i, and the non-proximal probe B targeting protein j. We reasoned that if linear regression is used to model the observed PLA product count onto the estimated proximity noise amount, true protein complexes would have positive intercepts (see Methods). The slope was then used to estimate the amount of proximity noise, and the count of a protein complex was calculated by subtracting the estimated proximity noise from the observed PLA product count (Fig 3b). Experimentally, we observed strong heteroscedasticity in the PLA product count when regressed on to the proximity noise amount (S9 Fig). Therefore, we performed linear regression using weighted least squares instead of ordinary least squares (see Methods).

We created a new simulation to directly compare the iterative and LR methods. The simulation's parameters were set to approximate the experimental data. More specifically, the simulation had three protein targets: protein 1, protein 2 and protein 3. Proteins 1 and 2 interacted both with themselves and each other (Fig 3c, S1 Table). 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 for proteins 1 and 2, respectively). We found that the iterative method correctly identified protein complexes 1:1, 1:2, 2:1 and 2:2 (Fig 3c).

To determine if we can statistically infer the enrichment of PLA products, we performed a one-sided Fisher's exact test on the counts of PLA products (Fig 3d, see Methods). This analysis correctly identified the four protein complexes present in the sample, independently confirming that the generated protein complexes occur at a higher frequency than random and can be statistically inferred (Fig 3d, see Methods). With regards to quantification of protein complexes on single cells, we observed that the iterative method consistently underestimated the true protein complex count (Fig 3c and 3e). Conversely, the LR method not only correctly identified the four true protein complexes (complexes 1:1, 1:2, 2:1 and 2:2), but also produced much more accurate counts for them (Fig 3c, 3d and 3f). Overall, the results of the two methods were correlated on the single-cell level (Fig 3g).

Ensemble method that combines both the LR and iterative methods for analysis of Prox-seq data

We next chose to explore a method that had the potential to outperform both the LR and iterative methods. As shown previously, the major weakness of the iterative method is its sensitivity to initialization conditions. We reasoned that the output from the LR method could be used as a sensible initialization for the iterative method (Fig 4a). Starting the iteration close to the correct result would make it less likely that the method would fall into a spurious local optimization. The performance of all three methods was compared in two simulations: one in which a high percentage of proteins were in complex with other proteins (high signal-to-noise) and one in which a low percentage of proteins were in complex (low signal-to-noise) (S1 Table).

The iterative method performed well when signal was high, but generated false positives when signal was low (Fig 4b and 4c). The LR method performed better in the low signal-to-noise simulation but suffered from false positives when noise was low (Fig 4c). This is not surprising because LR method depends on performing regression with product of non-

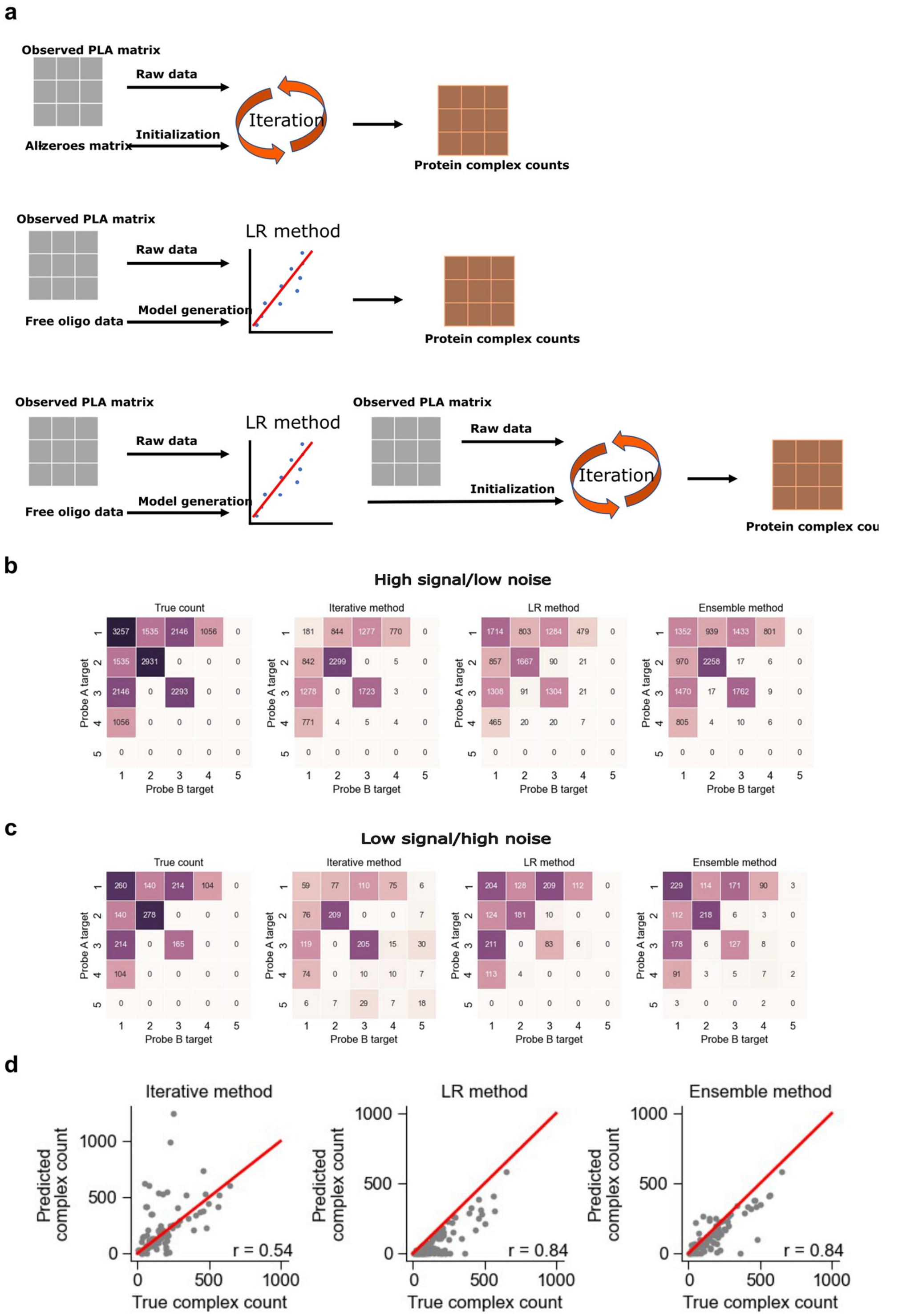

Fig 4. The ensemble method for improved analysis of Prox-seq data. We combine the iterative and LR methods for better prediction of protein complexes. (a) Schematic showing how all three methods arrive at protein complex estimation. The iterative method combines raw data and an initialization with an all-zeroes matrix to quantify protein complexes. The LR method uses raw data and free-oligo data to construct a linear regression model that quantifies protein complexes. The ensemble method begins with applying the LR workflow and uses the output of it to initialize the iterative method. (b) Comparison of all three methods in a regime of high signal and low noise, compared to the true counts. (c) Comparison of all three methods in a regime of low signal and high noise, compared to the true counts. (d) The Pearson's correlation between true counts and the outputs for each method across single cells. Each example shows complex 3:3 from the low signal/high noise regime.

https://doi.org/10.1371/journal.pcbi.1011915.g004

proximal probes as the explanatory variable, and the LR method will become unstable if there are few non-proximal probes across all single cells (or low noise in our simulation). Both methods consistently underestimated the abundance of protein complexes. For the iterative method, this is partly because expected PLA count we assumed is the maximal proximity noise it might have. The slope we use to quantify protein complexes from LR method tends to be larger than correct one because counts of non-proximal probes we can measure are inevitably lower than real counts both in experiment and simulation, which would give us a smaller positive intercept and protein complex count. In contrast, the ensemble method was able to maintain strong performance in both scenarios. It was less likely to produce a false positive, assigned fewer reads to false positives than other methods, and was closer to the true count for most of the protein complexes (Fig 4b and 4c). Finally, for a given PLA product, the ensemble method was more accurate in quantifying the abundance of true-positive complexes in single cells (Fig 4d).

A quantitative scoring strategy to comprehensively evaluate prediction methods

To evaluate the predictive performance of these methods more comprehensively, we further propose a quantitative scoring strategy to assign a prediction score for every prediction (S10a Fig). We simulate different biological scenarios with our model and score the overall prediction performance of each method by considering sum of absolute deviation between mean true counts and predicted counts ($\sum Mean_{deviation}$), sum of Pearson correlation coefficient ($\sum Pearson$) across singles cells (S10b Fig), and sum of ratios of false positive prediction ($\sum FPrate$) across single cells (S10c Fig) (see Methods). Comparing the methods across all scenarios showed that the ensemble method had the highest average prediction score and the lowest variance (S10d Fig & S2 Table). To further benchmark the performance of the three methods, we expanded our test cases to eight hundred. These tests are categorized into eight biological scenarios, mirroring the structure of the previous simulation. The scenarios include cases of only heterodimer, only homodimer, one overabundant protein, and multiple protein dimer situations. Each scenario is further divided into binary cases, featuring both high signal-to-noise ratios (complex abundance to monomer abundance is 10:1) and low signal-to-noise ratios (complex abundance to monomer abundance is 1:10). The input for the simulation is randomly sampled from a generator under a specific distribution, repeated 100 times within each scenario (S11e and S11f Fig). The results strongly support our earlier conclusion that the LR method excels in low signal-to-noise situations, whereas the iterative method exhibits unstable performance. The iterative method performs better in predicting only heterodimer or homodimer situations, while LR demonstrates greater robustness in handling more complex scenarios involving multiple protein dimers or overabundant proteins. While the iterative and LR methods each had regimes where they underperformed, the ensemble method consistently performs well across each scenario, making it a reliable choice for typical situations in which the true biological conditions are uncertain (S11a, S11b, S11c and S11d Fig).

Comparison of all three analytical methods to real data and performance evaluations

Next, we evaluated the concordance between all three methods on experimental data from single Jurkat and Raji cells. Overall, we found that each method largely agreed on which PLA products were predicted to be protein complexes (Fig 5a–5f). While the bulk measurements of protein complexes showed good agreement between methods, the three methods had varying

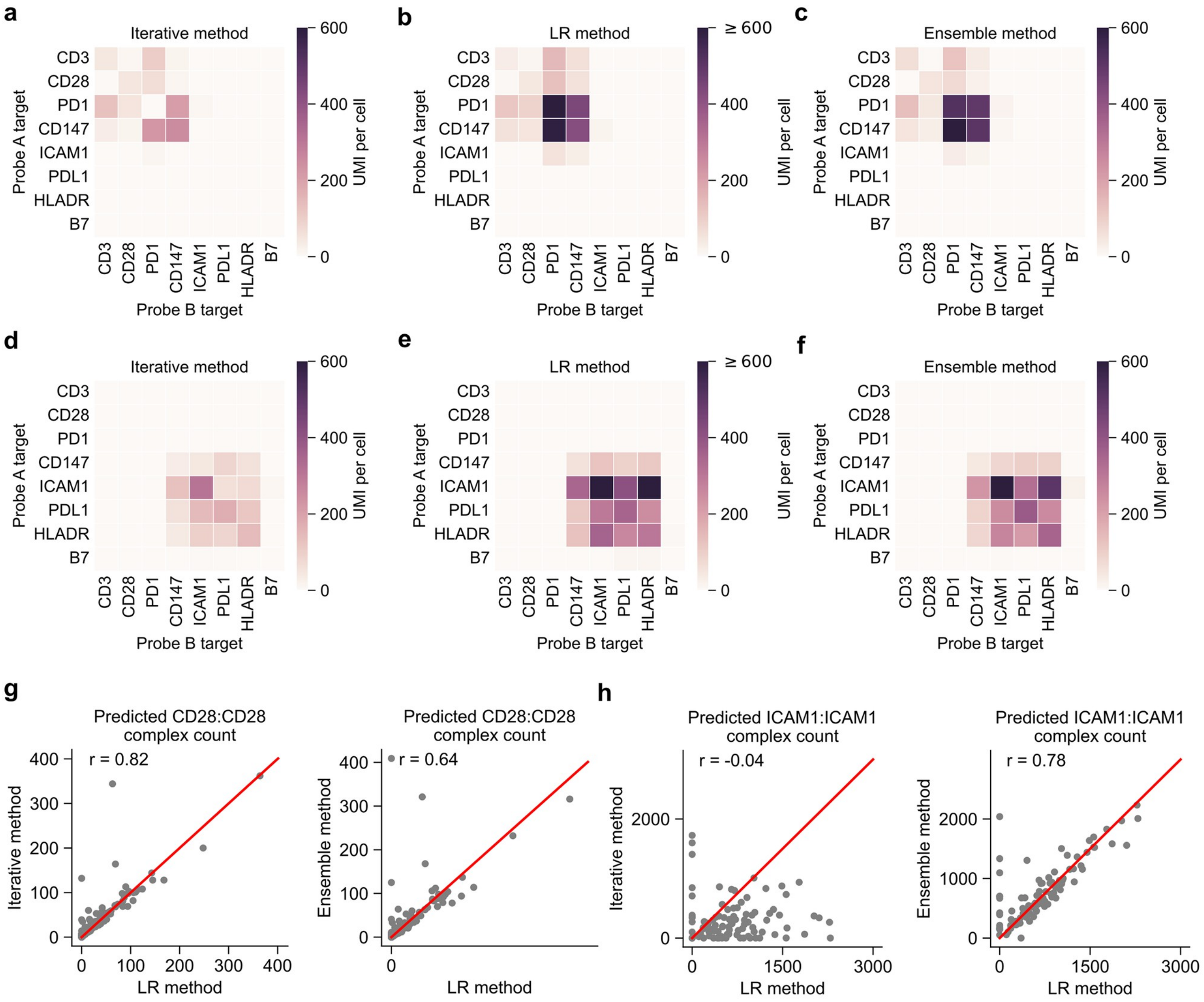

Fig 5. Comparison between the iterative and LR methods on experimental data. (a-c) Heatmaps showing the average of protein complex count, predicted by (a) the iterative method, (b) the LR method, and (c) the ensemble method in Jurkat cells. (d-f) Heatmaps showing the average of protein complex count, predicted by (a) the iterative method, (b) the LR method, and (c) the ensemble method in Raji cells. (g) Comparison of methods for predicting counts of protein complexes of CD28:CD28 and in Jurkat cells. (h) Comparison of methods for predicting counts of protein complexes of ICAM1:ICAM1 and in Raji cells. In (g, h), the red lines indicate y = x, and r indicates the Pearson's correlation coefficient.

https://doi.org/10.1371/journal.pcbi.1011915.g005

levels of correlation for single cells (Fig 5g and 5h). In addition, we observed all three methods, along with the Fisher's Exact test, largely identified the same protein complexes (S9 Fig).

All methods predicted protein complexes CD3:CD3 and CD28:CD28 in Jurkat cells, both of which are known protein complexes[12,13]. All three methods also predicted protein complex ICAM1:ICAM1 in Raji cells, which was shown to dimerize on the cell surface [14]. We also evaluated our methods against a simulation designed to more closely represent the experimental data. Protein expression levels were estimated from the experimental data and used to create simulation models for Jurkat and Raji cells (S1 Table). Then, protein complexes corresponding to CD3:CD3, CD28:CD28, and CD3:CD28 were added to Jurkat cells, whereas HLADA:HLADR and PDL1:PDL1 were added to Raji cells. Once again, we observe largely similar performance for all methods (S10 Fig).

Discussion

Here, we presented a comprehensive computational framework for simulating Prox-seq data, and for predicting protein complex count from Prox-seq data. We studied how the quantification of protein complexes was affected by proximity noise, which is caused by proteins that are not functionally interacting but are sufficiently close to each other by random chance to produce valid ligation products. Our simulation model showed that the amount of proximity noise is strongly depended on the protein abundance. Similar results have been observed in commercial *in situ* PLA [8].

We showed that with respect to protein complex prediction, the iterative method, LR method, and ensemble method largely agree on real experimental data. Therefore, we propose that each of these methods could be used for protein complex detection and quantification, and any protein complexes that were predicted by these methods were highly likely to be true protein complexes. However, in head-to-head comparisons using simulated data, the ensemble method performed well over a larger range of data types than the other methods (Table 1).

Our simulation model had some limitations. First, it did not consider interactions higher than dimers, diffusion of the protein molecules, their physical sizes, and the technical variability of the Prox-seq assay. It is important to note that our model will consider two proteins to have interaction if they are a part of a higher-order protein complex, even in the absence of direct physical contact, since they need only be within the designated interaction range determined by the proximity ligation distance. Second, the simulation model requires the user to independently select the abundance of a protein complex and its constituents' non-interacting counterpart. In real cells, these abundances are likely highly correlated. Finally, it assumed that the protein complexes and the non-interacting proteins were uniformly distributed on the cell surface. Despite these limitations, we showed that the overall structure of simulated Prox-seq data is very similar to real Prox-seq data.

Currently, application of each method requires a relatively homogeneous population of single cells. In practice, this requires that simultaneously acquired mRNA data is first used to cluster cell types, and then either method can be applied to individual clusters. This requirement is

Table 1. Comparison of features between three predictive methods.

	Iterative method	LR method	Ensemble method
Mechanism	Approximate the count of protein complexes by iteratively solving multiple quadratic equations. Protein complex count is the difference between observed PLA counts and expected PLA counts.	Construct a weighted least square model with non-proximal probes as independent variables and observed PLA count as response. Extrapolate the protein complex count based on the difference between observed PLA count and predicted proximity noise.	Bridge the iterative method and LR method by transferring the output of LR as the initial values into iterative method. This can make iteration go in a sensible part of the space that is likely to produce a good solution.
Features	Expected PLA count is calculated by multiplying the joint probability of simultaneously observing specific antibodies from two probes with total observed PLA counts. No additional experiments and information needed.	Predicted proximity noise or random ligation counts of PLA is calculated by multiplying the fitted slope coefficient with the product of non-proximal probe counts. Free oligo modification [4] to experiment is required to measure non-proximal probe count.	A combination of iterative method and LR method. LR method should be applied in advance to perform ensemble method. Free oligo modification [4] to experiment is required. The ensemble approach effectively enhances the generalization of the iterative method and LR method across different biological scenarios.
Applicable situations	Simpler biological scenarios where there are only homodimers or only heterodimers.	Low signal-to-noise situation and more complicated scenarios where there are multiple protein dimers or overabundant proteins.	Robust and consistent across various biological scenarios.
Limitations	Can be very unstable when applied to complicated scenarios where there are multiple protein dimers or overabundant proteins.	Not optimal for scenarios of high signal-to-noise, only homodimers, and only heterodimers. Additional experimental procedure is required to measure non-proximal probe count.	Additional experimental procedure is required to measure non-proximal probe count.

https://doi.org/10.1371/journal.pcbi.1011915.t001 because each method relied on a statistic of the whole population (the difference between observed and expected random PLA product count for the iterative method, and the linear regression's intercept and slope coefficient for the LR method) and having different complex expression levels would lower the power the methods. Further study is required to extend these methods to a population of heterogeneous cell types without the use of mRNA data.

We envision that the Ensemble method will be particularly 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 even more important source of noise given common macromolecular crowding effect within cells. 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 previously [4], proposed two additional independent methods for protein complex prediction, and introduced a model for simulating Prox-seq data.

Methods

Theoretical calculation of proximity noise

Suppose there are A_i probes A and B_j probes B on the cell surface. Assume that the probes are random points on a spherical surface, and proteins i and j do not interact. Because the ligation distance is significantly shorter than the cell's radius, we assume that a probe A and a probe B can be ligated if and only if the Euclidean distance between them, L, is less than or equal to the ligation distance, $d_{ligation}$. The Euclidean distance L between any pair of random points has the following probability distribution[15]:

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

where R is the cell radius.

Then, the probability of ligation between two random points on the cell surface is:

$$P\left(L \leq d_{ligation}\right) = \frac{d_{ligation}^2}{4R^2}$$

Assume that each probe could be ligated as many times as possible, the mean counts of ligated PLA product i:j, $X_{i,j}$, follow a binomial distribution:

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

The expected count of PLA product that is created from random ligation of non-interacting probes is:

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

Note that this approximation assumes that each probe can be ligated many times, while the simulation model assumes that each probe can only be ligated at most once. Experimentally, each probe can only be ligated 3–7 times, depending on the number of DNA oligomers per probe. As a result, this estimate represents the upper limit of the random ligation amount.

Simulation model

Assume that each protein molecule and the Prox-seq probe that binds to it are point particles. Let there be n protein targets. Let A_1, A_2,. . ., A_n be the simulation parameters that represent the count of probe A that targets proteins 1, 2,. . ., n. Let B_1, B_2,. . ., B_n be the simulation parameters that represent the count of probe B that targets 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 represent the counts of protein complexes 1:1, 1:2,. . ., 1:n, 2:1, 2:2,. . ., n:n.

The simulation is performed separately on each single cell. For the single cell t, we first generate $A_i^{(t)}$ number of random points on a sphere surface, which correspond to the number of detected probe A that targets protein i on cell t. The coordinates of each point are [16]:

$$x = R\sqrt{1-u^2}\cos\theta$$
$$y = R\sqrt{1-u^2}\sin\theta$$
$$z = Ru$$

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

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

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

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 provides the probability of getting $A_i^{(t)}$ failures, given n_{NB} successes and p_{NB} is the probability of success. n_{NB} is used to control the variance of the probe count, and p_{NB} is calculated such that the mean of $A_i^{(t)}$ is equal to A_i. The choice of n_{NB} value here is based on observation of experimental data. While different PLA products have different best-fitted n_{NB} values, we take the mean value of n_{NB} fitted for various PLA products on either Jurakt T cells or Raji B cells for our standard simulation (S3c Fig). n_{NB} is a flexible parameter to change in the simulation.

Second, we randomly generate $B_i^{(t)}$ number of points on a surface of a sphere, which correspond to the number of detected probe B that targets protein i on cell t. The coordinates of each point are generated identically to above.

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

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

This is to ensure that the counts of detected probe A and probe B that target the same protein are proportional to each other.

Third, we randomly generate $c_{i,j}^{(t)}$ number of points on a surface of a sphere, which correspond to the count of protein complex i:j on cell t. Then, these $c_{i,j}^{(t)}$ points are 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 variance:

$$c_{i,j}^{(t)} \sim NegativeBinomial(n_{NB}, p_{NB})$$

where n_{NB} = 1.5 in our study, and $p_{NB} = \left(1 + \frac{c_{i,j}}{n_{NB}}\right)^{-1}$.

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

The number of probe A and probe B points that are not ligated are returned as the simulated non-proximal probe count that is measured by the free oligo modification.

The simulation is repeated 100 times to simulate PLA product counts of 100 single cells. The parameters for all simulations are listed in S1 Table. All simulations include negative binomial variance, unless stated otherwise.

Calculation of protein count and expected PLA product count

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

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

where $X_{i,l}$ and $X_{k,i}$ indicate the observed (i.e., measured) counts of PLA products i:l and k:i, respectively. The PLA product i:i is counted twice to the protein count to account for the fact that two molecules are present in a homodimer.

The expected count of a PLA product i:j is:

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

Protein complex prediction: iterative method

The count of protein complex i:j is calculated iteratively using the following equation:

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

where $Y_{i,j}^{(m)}$ is the predicted count of protein complex i:j at the m$^{\text{th}}$ iteration. The initial values for all protein complexes are 0.

The second term of the right hand side represents the count of PLA product i:j that is caused by random ligation.

After each iteration, a one-sided t-test is performed on the values of $Y_{i,j}^{(m+1)}$ across all single cells. The alternative hypothesis is that the mean of $Y_{i,j}^{(m+1)}$ is 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 were determined to not represent true protein interactions.

There is also a symmetry condition, such that if i:j is a protein complex, then j:i should also be a protein complex, even if $Y_{j,i}^{(m+1)}$ fails the t-test. This is done by setting $Y_{j,i}^{(m+1)}$ as a fraction of $Y_{i,jj}^{(m+1)}$:

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

where sym_weight is arbitrarily chosen to be 1 in our study.

Protein complex prediction: linear regression (LR) method

For each PLA product i:j, its observed count is regressed onto the product of its corresponding non-proximal probe A count and non-proximal probe B count, using weighted least squares:

$$X_{i,j} \sim \beta_0 + \beta_1 A_i' B_j'$$

where $A_i^{'}$ and $B_j^{'}$ are the count of non-proximal probe A targeting protein i, and non-proximal probe B targeting protein j, respectively. The weight for a sample (ie, a single cell) p is:

$$w_p = \frac{1}{A_i' B_j'}$$

For simulated data, we also scale the interaction term by 10^6 whenever necessary, such that it is close to the orders of magnitude of $X_{i,j}$. $A_i^{'}$ and $B_j^{'}$ are obtained from PLA products that contain the added free oligos. For example, the count of non-proximal CD3 probe A is equal to the count of PLA product CD3:free_oligo_B, and the count of non-proximal CD28 probe B is 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 is that $\beta_0 > \beta_{cutoff}$. For simulated data, $\beta_{cutoff} = 1$. For experimental data $\beta_{cutoff} = 10$. All PLA products with Benjamini-Hochberg-corrected P-values below 0.05 are considered to be true protein complexes. The protein complex count, $Y_{i,j}$, is calculated as the difference between the observed PLA product count and the interaction term:

$$Y_{i,j} = X_{i,j} - \beta_1 A_i' B_j'$$

The LR method is related to the binomial approximation of the random ligation signal above. If the counts of non-proximal probes are perfect proxies for the count of non-interacting probes, then we have the following relationship:

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

Protein complex prediction: Ensemble method

The ensemble method relies on solving same quadratic equations as iterative method to approximate counts of protein complex. The only difference is that it takes protein complex matrix calculated from LR method as initial values. There is an argument called df_guess embedded in predictive function which is set to be all zeros by default. Note that LR method should be applied in advance in order to perform ensemble method.

Protein complex prediction: Fisher's exact test

A one-sided Fisher's exact test is conducted on the table below (Table 2). The alternative hypothesis being tested is whether $X_{i,j}$ (the observed value) is significantly greater than what would be expected by chance. Following this, Benjamini-Hochberg correction is applied to the P-values obtained from all PLA products, for each individual cell. We assume there is a protein-protein interaction on a given cell if the corrected P-value falls below the threshold of 0.05.

Table 2. A 2 by 2 contingency table is first constructed for each PLA product i:j.

	Probe B = j	Probe B $\neq$ j
Probe A = i	$X_{i,j}$	$\sum_{\substack{l=1 \\ l \neq j}}^{n} X_{i,l}$
Probe A $\neq$ i	$\sum_{\substack{k=1 \\ k \neq i}}^{n} X_{k,j}$	$\sum_{\substack{k=1 \\ k \neq i}}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{n} X_{k,l}$

https://doi.org/10.1371/journal.pcbi.1011915.t002

Prediction score mechanism

$$Score = w_1 * \sum Pearson - w_2 * \sum Mean_{deviation} - w_3 * \sum FPrate$$

where w_1, w_2, w_3 are chosen to be 0.5, 0.4, 0.1 in our study. $\sum Pearson$ equals to the sum of Pearson correlation coefficients for every real protein complex between true complex counts and predicted complex counts across all singles cells. $\sum Mean_{deviation}$ equals to the sum of absolute difference between mean true counts and predicted counts for every PLA product:

$$Mean_{deviation} = \frac{|Mean_{pred} - Mean_{true}|}{Mean_{true}}$$

$\sum FPrate$ equals to the sum of ratios of false positive prediction across single cells for every non-existing PLA product.

For quantification accuracy evaluation where there are true protein complex counts, we consider parameters $\sum Pearson$ and $\sum Mean_{deviation}$. Pearson correlation coefficient takes single cells into consideration while means counts can give us information about bulk abundance of different PLA products. We found that poor prediction of PLA counts in single cells might still contribute to seemingly good mean counts estimation, which shed lights on us that Pearson correlation should be a more important and robust parameter than mean counts. For $\sum FPrate$ evaluation where there is no true complex, we use fraction of complex-positive cells to represent how many ratios of single cells are wrongly assigned at least a complex count. According to our multiple tests, each method tends to assign only few false positive reads, mostly only one in some single cells to PLA products. So that we assume false positive rate a minor metric to be considered in our scoring strategy. In conclusion, we arbitrarily choose effector weight for each parameter given relative importance discussed above.

Supporting information

S1 Fig. Comparison of real and simulated data for protein count and non-proximal probe count. (a) Scatter plot showing the mean-variance relationship in real and simulated protein count. (b, c) Scatter plots showing the relationship between observed CD3:CD3 PLA product and (b) non-proximal CD3 probe A or (c) non-proximal CD3 probe B in Jurkat cells. (d, e) Scatter plots showing the relationship between observed 1:1 PLA product and (d) non-proximal protein 1 probe A or (e) non-proximal protein 1 probe B in simulated data without variance. (f, g) Scatter plots showing the relationship between observed 1:1 PLA product and (f) non-proximal protein 1 probe A or (g) non-proximal protein 1 probe B in simulated data with negative binomial variance.
(TIF)

S2 Fig. Distribution of non-proximal probe counts in Jurkat cells and simulated data. (a) Scatter plot showing the mean-variance relationship in experimental (Jurkat cells) and simulated non-proximal probe count. (b, c) Violin plots showing the experimental and simulated count of (b) non-proximal probe A and (c) non-proximal probe B for CD3 protein. (d, e) Violin plots showing the experimental and simulated count of (d) non-proximal probe A and (e) non-proximal probe B for PD1 protein. (f, g) Violin plots showing the experimental and simulated count of (f) non-proximal probe A and (g) non-proximal probe B for PDL1 protein. non-proximal probe counts for CD3, PD1 and PDL1 proteins were simulated by using the mean non-proximal 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.
(TIF)

S3 Fig. Negative binomial variance causes the simulation to produce distributions that closely match real data (a) The simulation with NB variance outperforms Poisson variance for both the mean-variance relationship (top) and proportion of zeroes (bottom). (b) For each observed PLA, protein complex, proximity noise and protein, the Mann-Whitney U statistic between posterior predictive samples and observed data averaged over samples. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. Higher is better. (c) Scatter plot of fitted nNB value versus mean value of PLA products on Jurkat and Raji cells. PLA product count is fit with a negative binomial distribution model across single cells. Different PLA products have different best-fit nNB values. As shown by the dashed blue horizontal line, the mean value of nNB is close to 1.5 for PLA products of different mean values. Thus, nNB = 1.5 is used as default value in the simulation model.
(TIF)

S4 Fig. Histogram showing the distribution of three binding cases of PLA product. (a) Simulation data where nonspecific binding probability of antibody 1, 2, 3 is set to 0.2, 0.1, and 0.05, respectively. (b) Experimental data where isotype control antibody is used to estimate the probability of nonspecific binding.
(TIF)

S5 Fig. Heterogeneous cell clustering based on simulated PLA data. (a) Single cell uniform manifold approximation and projection (UMAP) plot of three cell types. Clustering of cells is computationally determined by unsupervised learning. Cell types are ground truth for comparison. (b) Dot plot showing that differential expression analysis of each cluster caputures the featured PLA product of each cell type. (c) UMAP plot showing the relative expression (log2FC) of all PLA products in each cell cluster.
(TIF)

S6 Fig. Comparison between true and calculated protein expression. Scatter plots showing the correlation between true and calculated protein expression in simulated data. The protein expression is equal to the UMI count of each protein per single cell. Each panel also displays the corresponding Spearman's correlation coefficient, ρ.
(TIF)

S7 Fig. Comparison between observed and expected 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.
(TIF)

S8 Fig. The iterative method is sensitive to initial values. (a) The change in output for each iteration (tolerance) changes depending on the initial values given to the algorithm. (b) The resulting protein complex estimates for each initialization, compared to the true complex values (left panel).
(TIF)

S9 Fig. 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, and the corresponding weighted least squares (WLS) and ordinary least squares (OLS) regression lines. (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, and the corresponding WLS and OLS regression lines. (d) Residual plot of ordinary least squares regression for PLA product HLADR:HLADR in Raji cells.
(TIF)

S10 Fig. Overview of the prediction score. (a) Scheme of using simulation model to evaluate prediction algorithms with a quantitative scoring strategy. (b) The Pearson's correlation coefficient between true and predicted complex counts for each method. (c) The percent of cells called to have a protein complex for each method, along with the true counts. (d) The median prediction score of the ensemble method is higher than both the LR and iterative when comparing across all scenarios. Ensemble also displays a lower variance than LR.
(TIF)

S11 Fig. Benchmarking of three methods in eight hundred randomized input tests. (a-d) Violin plot showing the prediction score of three methods under different simulation scenario with either high signal-to-noise or low signal-to-noise cases. (e) Distribution of input from random generator under high signal-to-noise ratio. Single probe and homodimer pair follow a uniform distribution while heterodimer pair follows a triangular distribution. (f) Distribution of input from random generator under low signal-to-noise ratio. Single probe and homodimer pair follow uniform distribution while heterodimer pair follows triangular distribution.
(TIF)

S12 Fig. Comparison between the iterative method, the LR method, Ensemble, and Fisher's exact test for protein complex detection in real data. (a-d) Heatmaps showing the fraction of Jurkat cells that express a protein complex, as predicted by (a) the iterative method, (b) the LR method, (c) the Ensemble method, and (d) the Fisher's exact test. (e-h) Heatmaps showing the fraction of Raji cells that express a protein complex, as predicted by (e) the iterative method, (f) the LR method, (g) the Ensemble method, and (h) the Fisher's exact test.
(TIF)

S13 Fig. Simulation of T cells' and B cells' experimental data. (a-d) Heatmaps of (a) true protein complex counts, and protein complex counts predicted by (b) iterative method, (c) LR method, and (d) Ensemble method in T cell simulation. (e-h) Heatmaps of (e) true protein complex counts, and protein complex counts predicted by (f) iterative method (g) LR method, and (h) the Ensemble method in B cell simulation. Here, the simulation parameters were chosen such that the total protein abundance was similar to experimental data. T cells were simulated to only express the protein complexes CD3:CD3, CD28:CD28, CD3:CD28 and CD28: CD3. B cells were simulated to only express the protein complexes PDL1:PDL1 and HLADR: HLADR.
(TIF)

S1 Table. List of simulation parameters.
(DOCX)

S2 Table. Prediction score under different simulation scenarios.
(DOCX)

Author Contributions

Conceptualization: Junjie Xia, Hoang Van Phan, Luke Vistain, Savaş Tay.

Data curation: Junjie Xia, Hoang Van Phan, Luke Vistain.

Formal analysis: Junjie Xia, Hoang Van Phan, Mengjie Chen, Aly A. Khan, Savaş Tay.

Funding acquisition: Mengjie Chen, Aly A. Khan, Savaş Tay.

Investigation: Junjie Xia, Hoang Van Phan, Luke Vistain, Mengjie Chen, Aly A. Khan, Savaş Tay.

Methodology: Mengjie Chen.

Project administration: Savaş Tay.

Software: Junjie Xia, Hoang Van Phan.

Supervision: Savaş Tay.

Validation: Junjie Xia, Aly A. Khan.

Visualization: Junjie Xia, Hoang Van Phan, Luke Vistain, Savaş Tay.

Writing – original draft: Junjie Xia, Hoang Van Phan, Luke Vistain, Mengjie Chen, Aly A. Khan, Savaş Tay.

Writing – review & editing: Junjie Xia, Hoang Van Phan, Luke Vistain, Mengjie Chen, Aly A. Khan, Savaş Tay.

References

1. Tay S, Hughey J, Lee T, Lipniacki T, Quake S, Covert M. Single-cell NF-κB dynamics reveal digital activation and analogue information processing. Nature. 2010; 466:267–271.
2. Patel AP, Tirosh A, Trombetta J, Shalek A, Gillespie S, Cahill D, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344:1396–1401. https://doi.org/10.1126/science.1254257 PMID: 24925914
3. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell. 2017; 65:631–643.e4. https://doi.org/10.1016/j.molcel.2017.01.023 PMID: 28212749
4. Vistain L, Phan H, Keisham B, Jordi C, Chen M, Reddy S, et al. Quantification of extracellular proteins, protein complexes and mRNAs in single cells by proximity sequencing. Nat. Methods. 2022; 19:1578–1589. https://doi.org/10.1038/s41592-022-01684-z PMID: 36456784
5. Fredriksson S, Gullberg M, Jarvius J, Olsson C, Pietras K, Gústafsdóttir S, et al. Protein detection using proximity-dependent DNA ligation assays. Nat. Biotechnol. 2002; 20:473–477. https://doi.org/10.1038/nbt0502-473 PMID: 11981560
6. Söderberg O, Leuchowius K, Gullberg M, Jarvius M, Weibrecht I, Larsson L, et al. Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. Methods. 2008; 45:227–232. https://doi.org/10.1016/j.ymeth.2008.06.014 PMID: 18620061
7. Chi Q, Wang G, Jiang J. The persistence length and length per base of single-stranded DNA obtained from fluorescence correlation spectroscopy measurements using mean field theory. Phys. Stat. Mech. Its Appl. 2013; 392:1072–1079.
8. Alsemarz A, Lasko P, Fagotto F. *Limited significance of the in situ proximity ligation assay.* 411355 https://www.biorxiv.org/content/10.1101/411355v2 (2018).

9. Gayoso A, Steier Z, Lopez R, Regier J, Nazor K, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat. Methods. 2021; 18:272–282. https://doi.org/10.1038/s41592-020-01050-x PMID: 33589839

10. Chung H, Parkhurst C, Magee E, Phillips D, Habibi E, Chen F, et al. Joint single-cell measurements of nuclear proteins and RNA in vivo. Nat. Methods. 2021; 18:1204–1212. https://doi.org/10.1038/s41592-021-01278-1 PMID: 34608310

11. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay P, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods. 2017; 14:865–868. https://doi.org/10.1038/nmeth.4380 PMID: 28759029

12. van der Merwe PA, Dushek O. Mechanisms for T cell receptor triggering. Nat. Rev. Immunol. 2011; 11:47–55. https://doi.org/10.1038/nri2887 PMID: 21127503

13. Esensten JH, Helou YA, Chopra G, Weiss A, Bluestone JA. CD28 Costimulation: From Mechanism to Therapy. Immunity. 2016; 44:973–988. https://doi.org/10.1016/j.immuni.2016.04.020 PMID: 27192564

14. Miller J, Knorr R, Ferrone M, Houdei R, Carron C, Dustin M. Intercellular adhesion molecule-1 dimerization and its consequences for adhesion mediated by lymphocyte function associated-1. J. Exp. Med. 1995; 182:1231–1241. https://doi.org/10.1084/jem.182.5.1231 PMID: 7595194

15. Weisstein, E. W. Sphere Line Picking. *Wolfram MathWorld Sphere Line Picking* https://mathworld.wolfram.com/SphereLinePicking.html.

16. Weisstein, E. W. Sphere Point Picking. *Wolfram MathWorld Sphere Point Picking* https://mathworld.wolfram.com/.