Intercellular Signaling Networks of a Tissue from Single-Cell Transcriptomics

by

Brendan Innes

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Molecular Genetics
University of Toronto

Intercellular Signaling Networks of a Tissue from Single-Cell Transcriptomics

Brendan Innes

Doctor of Philosophy

Molecular Genetics
University of Toronto

2023

Abstract

Single-cell RNA sequencing promises to revolutionize our understanding of tissues at the molecular level. Single-cell transcriptomics should allow us to identify the various cell types of a tissue by their transcriptomes, as well as indicating how they coordinate to carry out tissue function by the ligands and receptors they present. To fulfill the promise of molecular models of tissue function from scRNAseq data, we must be able to: collect single-cell transcriptomes at scale; identify all contributing cells of a tissue from their transcriptome; and predict how these cell types are communicating based on their expressed ligands and receptors. In this work I address some of the challenges that have arisen in our efforts to realize this potential of scRNAseq technology.

The software tool scClustViz is my contribution to the workflow required to cluster single-cell sequencing libraries into cell type transcriptomes. It proposes a novel method to determine appropriate cluster resolution by differential gene expression and fixes a common error in differential expression magnitude due to pseudocounts. Its web-based interactive interface simplifies both the annotation of clusters and the sharing of results.

A systematic review of methods using scRNAseq to predict intercellular communication between cell types revealed that most, including my CCInx tool, infer communication solely based on ligand and cognate receptor expression. Specificity of these predictions may be improved by using evidence of ligand perturbation in the receptor cell transcriptome, as implemented by a handful of methods. These methods commonly expect that receptor activation yields a characteristic downstream gene expression signature. I test that assumption using both high-throughput and individual ligand perturbation assays and show that transcriptional response to ligand perturbation is cell type specific. If transcriptional response to ligand interaction can be inferred per cell type, this barrier to improving ligand-receptor inference methods can be overcome.

Acknowledgments

I am indebted to so many people for guiding me along this amazing yet convoluted academic journey. Stuart Berger got me started, and even encouraged me, the ridiculous undergrad, when I started trying to artificially generate amyloid plaques from brain tissue in his rheumatoid arthritis lab. He once advised me to pursue systems biology – I got there eventually, Dr. Berger!

Dave Litchfield taught me the nitty-gritty of biochemistry while making sure I was always thinking about the big question. And he let me go a little crazy learning Perl, which got me started on this whole bioinformatics thing.

Big thanks to Paul Boutros for helping me navigate the most terrifying period of my academic journey, and Fritz Roth for throwing me a rope when I needed it most.

Freda Miller and Trevor Pugh have made for a wonderful academic committee, filled with support and fun discussions. Big thanks as well to the Miller-Kaplan lab for welcoming me into their lab meetings and helping me navigate the frontal cortex.

Best for last of course is Gary Bader. Gary has created so many cool opportunities for me during my PhD – awesome collaborations, meetings with other cool scientists. Most importantly though, he's helped me understand how to get the best from myself.

This thesis owes a debt to all the trainees I've had the pleasure of working with in Toronto, but especially the Toronto single-cell troubleshooting group. Big shoutout to the founding crew of Danielle Croucher, Laura Richards, Erin Stewart, and Neke Ibeh, as well as thanks to Javier Díaz-Mejía and now Shamini Ayyadhury for keeping the group going.

Speaking of trainees, my MoGen cohort is a bunch of superstars and I'm so lucky to be a part of the gang. A special nod to drunken journal clubs chez Fraser – what other band of nerds sees the end of a mandatory academic course as simply an opportunity to change venues?

Finally, I owe it all to my biggest supporters, my family. Mom, I love when you ask me "dumb" questions of my work, they're always the most interesting. Dad, look at me doing quant stuff! Scott, your guidance beyond the lab means so much. And Dr. Gracie Pio, thank you for cheering on my journey while on your own arduous path to excellence. Love you guys.

Table of Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

AUPR: Area Under the Precision-Recall curve

BMP: Bone Morphogenic Protein

cDNA: complementary DNA

CITE-Seq: Cellular Indexing of Transcriptomes and Epitopes by Sequencing

dDR: difference in Detection Rate

DE: Differentially Expressed

DLRP: Database of Ligand-Receptor Partners

DR: Detection Rate

E17.5: Embryonic day 17.5

FACS: Fluorescence-Activated Cell Sorting

FDR: False Discovery Rate

FISH: Fluorescence *In Situ* Hybridization

FRET: Förster resonance energy transfer

GE: Ganglionic Eminence

GO: Gene Ontology

GUI: Graphical User Interface

HPMR: Human Plasma Membrane Receptor database

HPRD: Human Protein Reference Database

ICA: Independent Component Analysis

IDE: Integrated Development Environment

IL-1: Interleukin-1

IUPHAR: International Union of Basic and Clinical Pharmacology

logGER: log2 Gene Expression Ratio

MA plot: modified Bland-Altman plot comparing log-ratio (M) to average (A) gene expression

MAD: Median Absolute Deviation

MAST: Model-based Analysis of Single-cell Transcriptomics

MDGE: Mean Detected Gene Expression

MGE: Mean Gene Expression

NSC: Neural Stem Cell

PC: Principal Component

PCA: Principal Component Analysis

PPI: Protein-Protein Interaction

RP: Radial Precursor

scATACseq: single-cell Assay for Transposase-Accessible Chromatin using Sequencing

SCC: Spearman Correlation Coefficient

scRNAseq: single-cell RNA sequencing

SVM: Support Vector Machine

TF: Transcription Factor

TPM: Transcripts Per Million

tSNE: t-distributed Stochastic Neighbour Embedding

UMAP: Uniform Manifold Approximation and Projection

UMI: Unique Molecular Identifier

ZINB-WaVE: Zero-Inflated Negative Binomial-based Wanted Variation Extraction

Chapter 1
Introduction

1 Introduction

Single-cell RNA sequencing (scRNAseq) promises to revolutionize our understanding of tissues at the molecular level. Where previous technologies were limited to either quantifying the product of a handful of genes on a per-cell basis (i.e. flow cytometry) or required many cells worth of input to capture the full transcriptome (RNAseq), we now have the benefits of both in a single experiment. Single-cell transcriptomics should allow us to identify the various cell types of a tissue by their transcriptomes, as well as indicating how they coordinate to carry out tissue function by the ligands and receptors they present. To fulfill the promise of molecular models of tissue function from scRNAseq data, we must be able to: collect single-cell transcriptomes at scale; identify all contributing cells of a tissue from their transcriptome; and predict how these cell types are communicating based on their expressed ligands and receptors. In this work I address some of the challenges that have arisen in our efforts to realize this potential of scRNAseq technology.

1.1 High-throughput single-cell transcriptomics

To collect the transcriptomes of single cells, they must first be isolated prior to lysis and cDNA library generation. There are a few ways of doing this, but to capture a sufficient number of cells to represent the entire tissue the most common method is to encapsulate cells in droplets (Klein et al., 2015; Macosko et al., 2015). The cells of the tissue are suspended in an aqueous solution, which is then emulsified in oil using a microfluidic device with the aim of capturing a single cell per droplet (**Figure 1-1**). Each cell is tracked by the inclusion of a droplet-specific genetic barcode in the primer. Notably, inDrop (Klein et al., 2015) and the now widespread 10X Chromium platform (Zheng et al., 2017) use soft gel beads as a primer delivery method to ensure that each oil droplet contains barcoded primers, thereby avoiding situations where a droplet captured a cell but lacks the reagents to capture its mRNA. After emulsification, cell lysis and reverse transcription proceeds in each droplet. The reagent beads are in a solution containing lysis buffer, so that when they and a cell are emulsified together the cell is lysed. Unlike in traditional RNAseq, there is no fragmentation step; instead the mRNA is captured by primers in the reagent bead by its polyadenylated tail. After reverse transcription each cDNA molecule contains a barcode unique to its contributing reagent bead, which is therefore common to all cDNA molecules in the droplet, as well as a random octamer unique to each cDNA, known as its

Unique Molecular Identifier (UMI). At this point the emulsification can be broken so that amplification and preparation of the sequencing library can take place in a pooled fashion. After sequencing, barcodes are used to demultiplex the library into transcripts contributed by each droplet, and UMIs are used to remove redundant copies of the same cDNA transcript. The resulting single-cell libraries can be aligned to a reference genome to attain relative counts of genes expressed per cell.

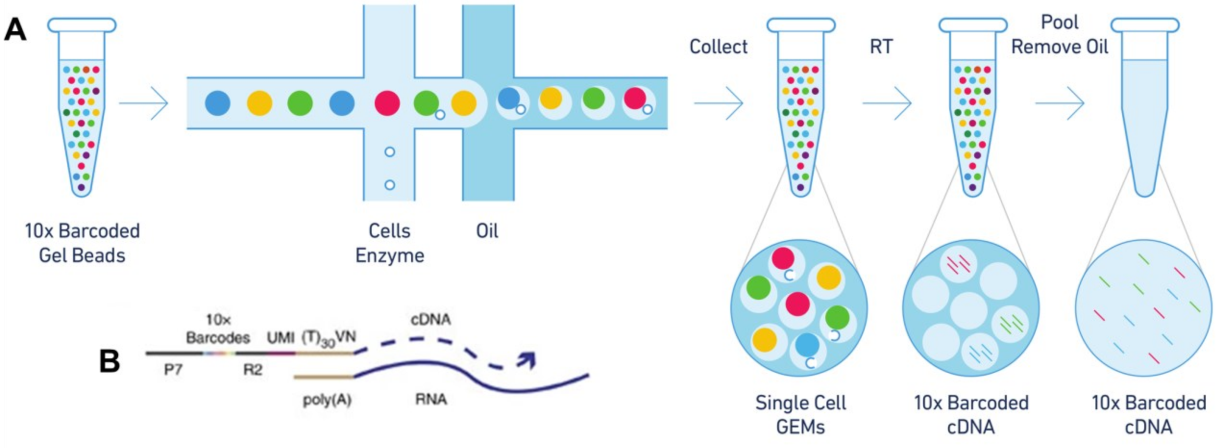


Figure 1-1. Single-cell cDNA library preparation in droplets.
A. Gel beads loaded with barcoded primers are mixed with cells and reagents before being encapsulated in oil, one bead per droplet. Cell lysis and reverse transcription takes place in oil droplets, then emulsion is broken for cDNA amplification and library construction to proceed.
B. Primers in gel beads contain both a barcode unique to each bead, a random octamer which serves as the unique molecular identifier (UMI) for that cDNA product, and a poly-T tail to capture polyadenylated mRNA. Figure adapted from Zheng et al. 2017.

The massive scalability of droplet-based scRNAseq does have unique drawbacks, however. Encapsulating a single cell per droplet is not a trivial task. There is always the possibility of getting more than one cell per droplet (a “doublet), and since droplets are transient, unlike wells of a plate, they cannot be easily imaged to determine aberrant loading. The number of cells per droplet can be modelled as a Poisson distribution, with rate being determined by the concentration of cells in the input suspension. By reducing the input concentration, you improve the likelihood of getting less than two cells per droplet. The number of transcripts detected in each library (“library size”) was used as a filter for doublets, but there is not sufficient separation to use this as a filter (**Figure 1-2**). More recently, computational methods such as DoubletFinder

(McGinnis et al., 2019) perform better at this task by predicting doublet transcriptomes from the various cell type transcriptomes present in the experiment. This relies on doublets being a rare occurrence, which means the cells being loaded must be well suspending and in low concentration.

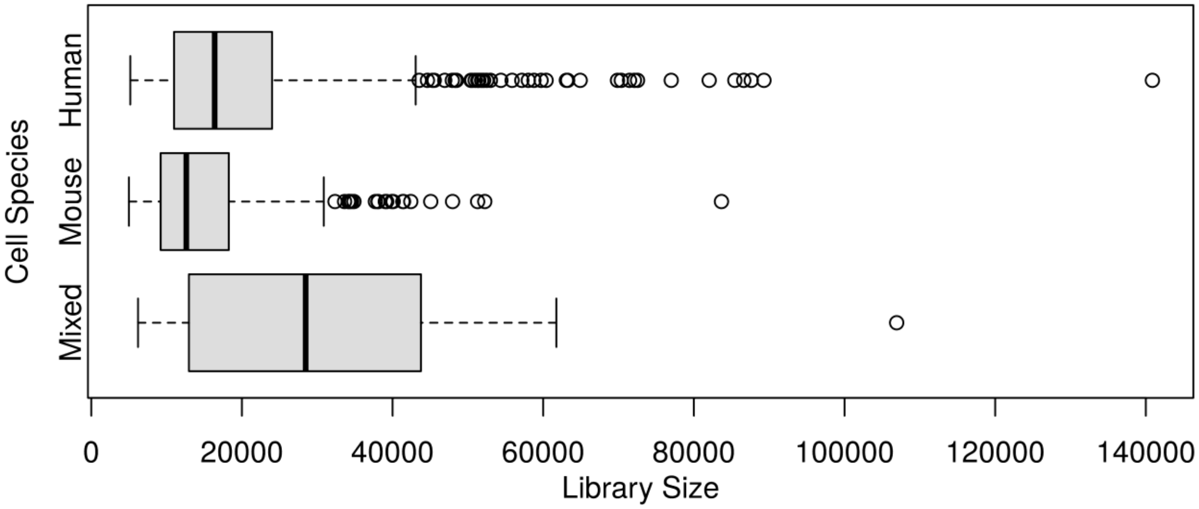


Figure 1-2. Transcript counts are not sufficient to distinguish single cell libraries from doublets. Mouse and Human cells were mixed in suspension at 50 cells/µL prior to single-cell cDNA library preparation by DropSeq (Macosko et al., 2015). cDNA transcripts were mapped to human and mouse reference genomes, and libraries containing transcripts mapping to both were used to determine rate at which two or more cells were captured in the same library ("Mixed"). While doublet libraries did have significantly more transcripts than single cells, library size alone is not sufficient to identify doublets in an scRNAseq analysis pipeline.

A low concentration of cells in the input suspension means most droplets will be empty. They will, however, contain primer and reverse transcriptase, and thus any mRNA encapsulated in the droplet may contribute transcripts to the final sequencing library. Thus, identification of these empty droplets is also an important, albeit relatively simple task. Thresholding by library size is generally sufficient to identify empty droplets.

The fact that empty droplets contribute reads suggests that cell-free mRNA exists in the cell suspension, which brings us to another important consideration in data interpretation from single-cell RNAseq. This is exemplified by the original DropSeq study of the mouse retina, where nearly 70% of cells were rod cells (Macosko et al., 2015). Specific marker genes of rod cells, especially their photoreceptor rhodopsin, were found in all cell transcriptomes in the data. Since rod cells highly express rhodopsin, are relatively delicate, and make up such a high proportion of the mouse retina, it follows that these cells would contribute the majority of mRNA from damaged and dying cells contaminating the input cell suspension. Most tissues contain a more even mixture of cell types, making the contribution of contaminant transcripts less obvious. To address this, methods such as SoupX use the transcripts sequenced from empty droplets to predict and correct for contaminant transcripts (Young and Behjati, 2020).

Some of those damaged cells are inevitably encapsulated and sequenced. Libraries sequenced from these droplets can be identified by the enrichment of RNA species prevalent in membrane-enclosed subcellular compartments relative to cytoplasmic mRNA (Ilicic et al., 2016; Muskovic and Powell, 2021). These species include mitochondrial gene products transcribed and translated solely in the mitochondria, as well as RNA that has yet to be spliced into mRNA and exported from the nucleus. Because these molecules are sequestered within an organelle, the assumption is that they are less likely to be lost if the cell's membrane is damaged when compared to cytoplasmic mRNA.

Another difference between droplet-based scRNAseq and traditional transcriptomics is the location along each gene where transcripts can be sequenced. Because input mRNA is not fragmented and the primer used for cDNA generation is a poly-T sequence, only the approximately 50 base pairs at the 3' end of an mRNA are sequenced. Genotyping, isoform identification, and detection of preprocessed RNA are thus limited to events occurring near the 3' end of each gene. This generally limits their use to applications such as demultiplexing of samples (Xu et al., 2019) or predicting expression dynamics (La Manno et al., 2018). On the other hand, the one-to-one relationship between captured mRNA molecules and their uniquely barcoded cDNAs is possible thanks to this lack of RNA fragmentation step. The addition of the UMI eliminates concerns of quantification error due to PCR amplification bias, where transcripts with lower G+C content are preferentially amplified (Kivioja et al., 2011; Islam et al., 2014).

Finally, the most apparent downside of droplet scRNAseq over plate-based or bulk methods is sensitivity. In a direct comparison of FACS, microfluidics plate, and droplet based methods (Ziegenhain et al., 2017), droplet-based methods detected both fewer transcripts and genes per cell than their contemporary alternative methods (**Figure 1-3**). One might assume this is due to the reduction in sequencing depth per cell necessitated by having more cells to sequence, but increased sequencing does not necessarily capture more unique transcripts, nor genes (**Figure 1-4**). For most applications, this is not a limiting factor, as the reduction in sensitivity can be mitigated by the increase in number of cells sequenced. Since the probability of detecting a specific gene expressed by a cell is dependent on both its expression rate and the sensitivity of the assay, increasing the number of cells of that type sequenced increases the probability of detecting that gene in at least one of those sequenced cells. As a result, single-cell sequencing libraries should be considered a random sample of each cell's transcriptome, whereby pooling libraries from the same cell type will give a more accurate representation of that cell type's transcriptome. For identification tasks (what genes are expressed?) this is sufficient, though quantitative tasks (how much did expression change) may suffer from poor dynamic range due to fewer transcripts.

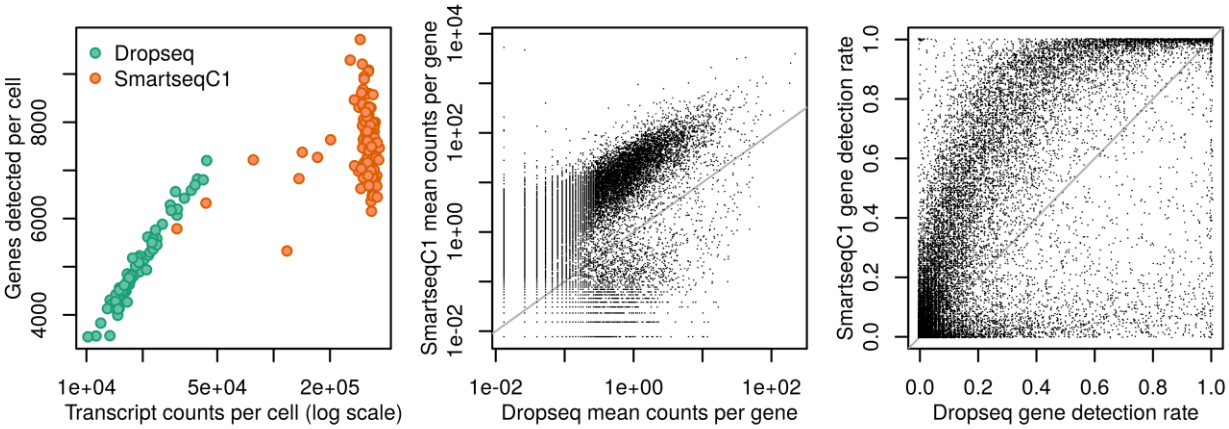


Figure 1-3. Comparison of droplet versus plate-based scRNAseq data.
500 mouse embryonic stem cells were sequenced using contemporary droplet-based (Dropseq) or microfluidic plate-based (SmartseqC1) technologies. Data from (Ziegenhain et al., 2017). Dropseq (green) captures fewer mRNA molecules (and consequently, genes) per cell than SmartseqC1 (orange). On a per-gene basis, Dropseq (x-axis) detects each gene fewer times and in a lower proportion of cells than SmartseqC1 (y-axis).

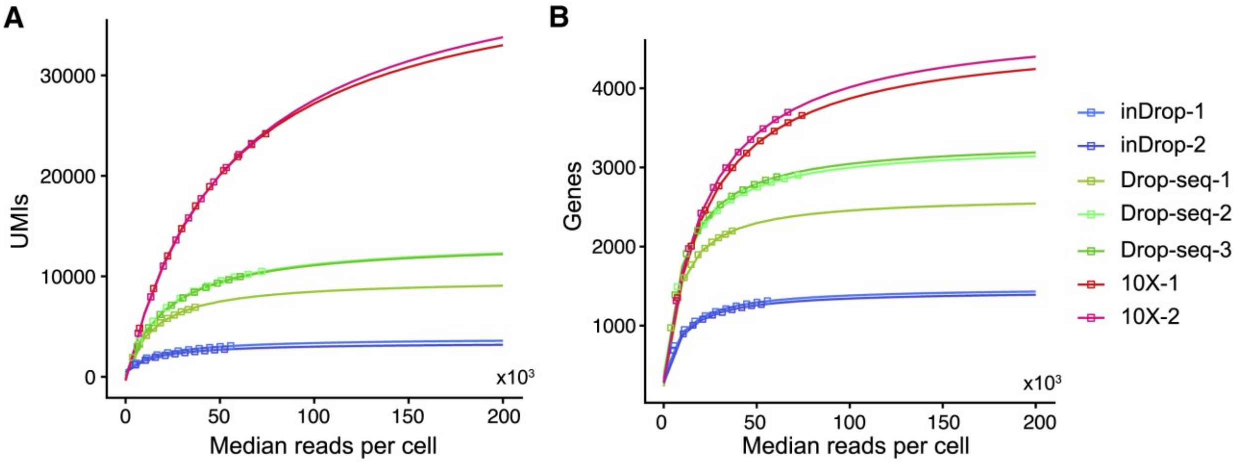


Figure 1-4. Low sequencing depth is sufficient to capture transcriptome complexity. The relationship between sequencing depth and per-cell sensitivity metrics for various droplet-based technologies. Drop-seq and inDrop are contemporary methods, while 10X (Chromium) is a modern commercial platform. Adapted from (Zhang et al., 2019b).

Making single-cell RNAseq scalable to whole tissues has trade-offs relative to traditional transcriptomics, but these can be mitigated with the appropriate adjustments to analysis. Without the increase in scale, efforts such as the first human whole-organ cell atlas would not be possible (MacParland et al., 2018).

1.2 Cell type identification from single-cell transcriptomics

1.2.1 What is a cell type?

Generating sequencing libraries from thousands of single cells is an important first step in building a molecular map of a tissue. The next challenge is determining what each of those libraries represents. As discussed above, scRNAseq is not sufficiently sensitive to treat each of those libraries as a comprehensive representation of an individual cell's transcriptome, so to take advantage of the quantity of cells assayed, the libraries must be classified and amalgamated. Because transcriptional information is all we have, the cells are necessarily grouped by transcriptional similarity and then identified as a cell type by prior knowledge. This definition of cell type is controversial, partly in response to questions raised by scRNAseq (Clevers et al., 2017).

Some argue that while a cell can be at least partly defined by its protein content, mRNA expression does not sufficiently correlate with protein content for this to be a satisfactory corollary. Correlation of protein and mRNA abundance in eukaryotes has been reported as low as 0.35 and high as 0.85, a range that dips low enough to cause alarm (Maier et al., 2009; Vogel and Marcotte, 2012; Li et al., 2014; Battle et al., 2015). However, this correlation between protein and mRNA abundance improves with increased expression; the most abundant proteins in a cell are most likely to be the products of the most highly expressed genes (**Figure 1-5**). Thus, assuming one is comfortable defining a cell by its most abundant proteins, the transcriptome should serve as an acceptable substitute.

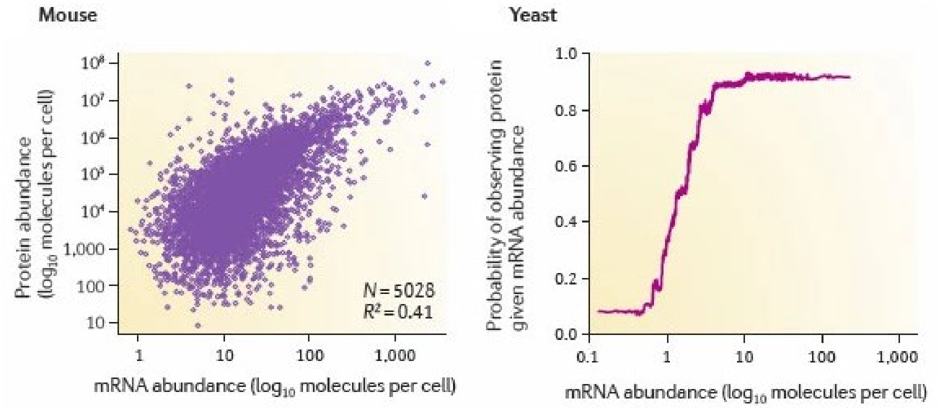


Figure 1-5. Highly abundant protein and mRNA are well correlated.
Association between protein and mRNA abundance in both multicellular and unicellular eukaryotes. Adapted from (Vogel and Marcotte, 2012).

A more compelling debate around the definition of cell types is concerned with whether it is an error to classify cells discretely when they may be more appropriately considered as positions in a manifold of cell states, akin to Waddington's epigenetic landscape (Waddington, 1957). Proponents of this theory argue that even terminally-differentiated cells can change their gene expression and function in response to external stimulus, essentially switching cell types (Pesaresi et al., 2019). Moreover, this switch is not discrete, and cells appear in between terminal points on this manifold in gene expression space when sequenced with scRNAseq. Since in gene expression space these cell types are not discrete, it is argued that defining cells by their position in gene expression space is more appropriate than using discrete cell type labels. Some use the term "cell state" to distinguish this view from the discrete "cell type". While this concept of a manifold of cell states is important to bear in mind, the admittedly artificial concept of the cell type as a tool for classifying and inferring cell function still has utility. For one thing, mapping scRNAseq data to prior knowledge would not be possible without using existing definitions of cell type. Furthermore, transcriptomes derived by scRNAseq are by definition a collection of Poisson distributions for each gene's expression in each cell; even if cell types were discrete positions in gene expression space, they would appear as probability densities when measured using scRNAseq due to stochasticity in both gene expression and its measurement. To incorporate scRNAseq data into our existing models of tissue function, the definition of cell types as sets of cells that share transcriptional similarity remains useful.

Finally, some argue that the transcriptome alone is not sufficient to differentiate cell types that have largely been defined by anatomical position or function. The most extreme example of this may be in the brain, where individual neurons may be transcriptional similar but functionally different by virtue of their role in a neural circuit. This is a valid concern that reinforces the need for a heuristic with which to classify scRNAseq libraries and map them to prior knowledge – in other words, the concept of cell types. By mapping single-cell transcriptomes to an existing classification based on histology and functional assays, we allow our model of a tissue to represent both functional and molecular definitions of a cell.

1.2.2 Clustering single-cell transcriptomes into cell types

To group thousands of single-cell sequencing libraries by their contributing cell types requires some data processing and unsupervised machine learning. First, libraries must be normalized both so that counts of each gene are meaningful relative to each other, and so that the assumptions of downstream analysis tasks are met. Transcriptomes are high dimensional with many correlated features, so feature selection and dimensionality reduction are the next important tasks. Finally, cell libraries are clustered to generate groups of transcriptionally similar cells that may be mapped to known cell types. While common practices have been established by consensus for most of these steps, they are still active areas of research. Here I outline established workflows and highlight evidence in support of best practices or alternatives where available.

Each single-cell library is a separate RNA sequencing experiment, where approximately 50 thousand cDNA molecules per cell are sequenced (see **Figure 1-4**). Since the number of reads per cell is not fixed, we must account for differences in sequence depth when comparing gene expression between cells. Scaling each single-cell library relative to its size (i.e. Reads Per Million mapped or RPM) is an intuitive solution, and the default in popular scRNAseq analysis software (Butler et al., 2018). However, because there are a finite number of reads available in a sequencing experiment, the composition of RNA being sequenced affects the relationship between actual abundance and proportion of reads for each gene product (Robinson and Oshlack, 2010; Quinn et al., 2019). This effect becomes more pronounced with asymmetric differences in expression – where one cell has more highly abundant genes than another. Calculating this asymmetry can be done by inferring the ratio of global RNA abundances between samples under the assumption that the majority of genes are not differentially expressed (Anders and Huber, 2010; Robinson and Oshlack, 2010). Unfortunately, while calculating these ratios works well in bulk RNAseq where cDNA counts are highly abundant, scRNAseq counts are sparse and counts of zero complicate ratio calculation. Pooling cells together ameliorates issues of sparsity, allowing a scaling factor to be calculated for the pool of cells by normalizing against the average of all cell libraries (Lun et al., 2016a). By iteratively generating pools, a system of linear equations can be generated to deconvolute cell-specific scaling factors. This pool and deconvolute strategy for normalization can be made robust to large-scale differential expression between groups of cells by clustering the libraries prior to normalization, and then normalizing each cluster separately before rescaling the cluster-specific scaling factors relative to each other. While there have been many methods proposed to address the challenges associated with depth normalization, systematic benchmarking studies have demonstrated that the challenge of compositional effects is more prevalent in single-cell transcriptomics, and thus normalization methods explicitly designed to address them are most appropriate for this data (Vieth et al., 2019). The pool and deconvolute strategy implemented in *scran* shows robust performance in benchmarking, even in the context of abundant and asymmetric differential expression.

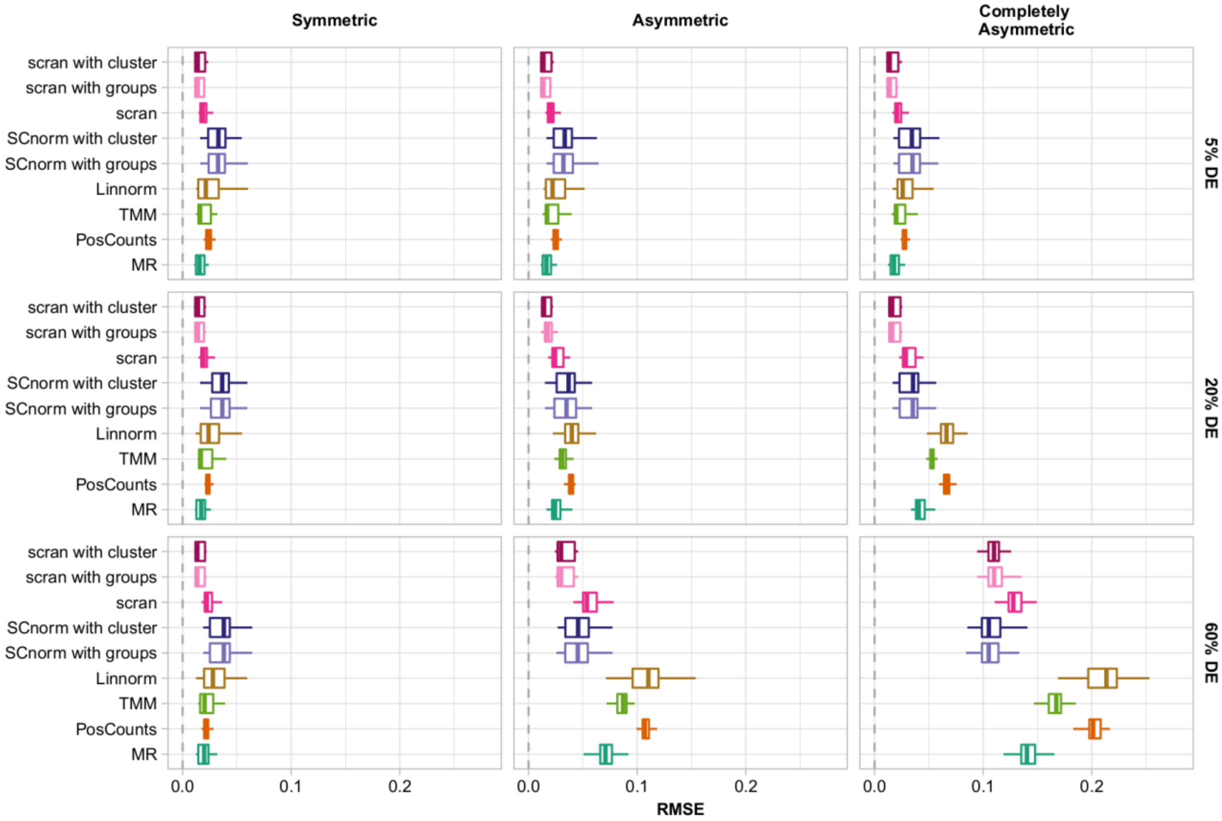


Figure 1-6. Normalization must be robust to asymmetric differential expression. Deviance (Root Mean Squared Error, RMSE) between estimated and simulated library scaling factors. The expression of ten thousand genes over 768 cells (384 cells per group) were simulated with different proportions of differentially expressed (DE) genes following increasingly asymmetric narrow gamma distributions. *TMM* is a normalization based on the ratio of global RNA abundances between samples used in traditional RNAseq (Robinson and Oshlack, 2010). The pool and deconvolute strategy is implemented in *scran* (Lun et al., 2016a). Adapted from (Vieth et al., 2019).

The second goal of normalization is to make the data suitable for downstream analysis. Many statistical tests and analyses are based on linear models, and a fundamental assumption of these models is homoscedasticity – equal variance of all variables. Sampling gene products from a cell is a Poisson process where the likelihood of capturing a particular gene's transcript is dependent on the abundance of that mRNA in the transcriptome. Unlike the normal distribution, in the Poisson distribution the variance increases linearly with the mean and is thus heteroscedastic. To further complicate matters, single-cell RNA sequencing represents a whole collection of Poisson distributions with different rates for each gene in each cell. The gamma-Poisson or negative binomial distribution can be interpreted as a mixture of Poisson distributions with the rate parameter sampled from a gamma distribution specified by the dispersion term. This distribution models scRNAseq data well (**Figure 1-7**). It is worth noting at this point that the rate of zeroes present in droplet scRNAseq data (referred to as dropout, overdispersion, or zero-inflation in the literature) fit the negative binomial model (Svensson, 2020). To render data sampled from a negative binomial distribution homoscedastic, one can apply a variance-stabilizing transformation (Anscombe, 1948). Anscombe's transformation has been simplified slightly in the field of transcriptomics to the binary logarithm of the normalized counts, plus a pseudocount of one: $\log_2(x+1)$. In Anscombe's transformation the pseudocount is related to the dispersion parameter, but has been simplified to a constant to both eliminate the need to estimate parameters and to maintain zero counts as zero after transformation. The choice of pseudocount does have ramifications, as explored further in **Figure 2-1**. Despite simplifying the pseudocount, this log-transformation does render the data homoscedastic, allowing for downstream analyses based on linear models.

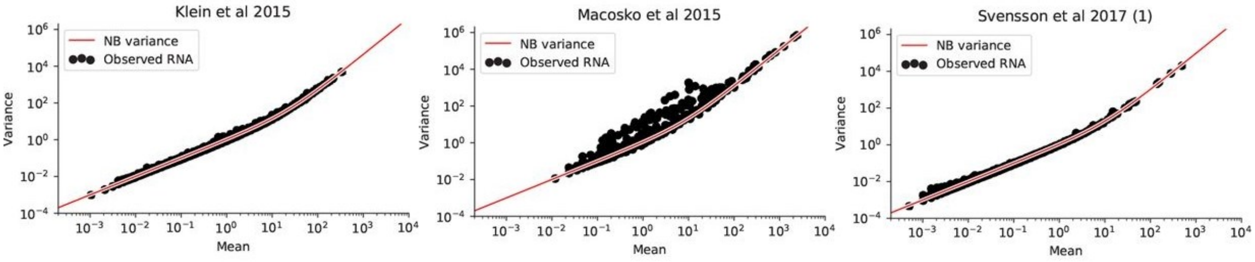


Figure 1-7. Mean-variance relationships in droplet scRNAseq data.
Data from inDrop (Klein et al., 2015), DropSeq (Macosko et al., 2015), and 10X Chromium (Svensson et al., 2017). Points are genes in negative control datasets from each technology. Red line is the expected variance relationship when sampling from the negative binomial distribution. Adapted from (Svensson, 2019a).

There are nearly twenty thousand protein coding genes in the human reference genome (GRCh38). Since in any cell library most of these gene products will be missing, and as discussed above correlation between transcriptome and proteome improves when considering highly expressed genes, some filtering of genes may improve the calculation of cell similarities necessary for clustering. A common choice is to select genes that are the most highly variable, with the assumption that this eliminates those genes whose changes in the dataset are due to noise or stochasticity in transcription. This can be performed by modeling the mean-variance relationship and selecting genes that show higher than expected variance (Lun et al., 2016b; Stuart et al., 2019). More advanced methods of feature selection have also been developed, including NBDrop (Andrews and Hemberg, 2019), which identifies genes whose expression is missing in more cells than expected under the assumption that they are more likely to reflect differences in cell type, and DUBStepR (Ranjan et al., 2021), which looks for a minimally redundant set of genes with the maximal range of gene-gene correlations with the assumption these are involved in cell type specific biological pathways. An incomplete benchmark of feature selection methods (missing both NBDrop and DUBStepR) was performed by the authors of an iterative feature selection method FEAST (Su et al., 2021) which found that after their novel method, identifying highly variable features based on the mean-variance relationship performed reasonably well in subsequent clustering. It is notable, however, that they used only feature selection to shrink feature space prior to clustering, skipping a further dimensionality reduction step such as PCA.

Dimensionality reduction methods are not limited to selecting meaningful features. Since gene expression is a structured process, where genes in the same biological pathway are more likely to be coexpressed, the features in a transcriptomic dataset can be highly correlated. We can take advantage of this by summarizing correlated features into a set of factors using linear algebra methods such as Principal Component Analysis (PCA). PCA transforms the feature space into an ordered set of orthogonal components where each linear component aims to explain the maximum amount of remaining variance in sample space. By selecting the top principal components (PCs), one attains a small set of features that explain much of the variance in the data. Note that though PCs are ordered by the proportion of variance they explain in the original data, their loadings are not scaled to reflect their relative importance. Performing this scaling is a necessary step to avoid overemphasizing minor sources of variance in the original data. Dimensionality reduction methods were benchmarked in the context of both clustering and trajectory inference, and PCA was one of the best performing methods in both contexts (Sun et al., 2019). Given that most feature selection methods (with the exception of DUBStepR) prioritize the removal of "noisy" features over the reduction of the redundancy inherent in gene coexpression, a common workflow for shrinking feature space prior to clustering is to filter for highly variable genes in an attempt to reduce technical noise, then using PCA to transform the redundant feature space into a small number of uncorrelated components (Amezquita et al., 2020).

Clustering of single-cell libraries to identify contributing cell types is performed on this matrix of scaled PC loadings per sample. Segregating samples into groups by feature similarity is a common unsupervised learning task, and many existing algorithms have been adapted for scRNAseq analysis. Comprehensive benchmarking of these methods (**Figure 1-8**) revealed that the *PhenoGraph* clustering method adapted for the popular *Seurat* workflow was the top performer, narrowly surpassing the consensus clustering method *SC3* (Kiselev et al., 2017; Butler et al., 2018; Duò et al., 2018). The *PhenoGraph* algorithm, developed for analysis of CyTOF data, clusters by maximizing modularity in a shared nearest neighbours graph (Levine et al., 2015). First they construct a K-nearest neighbours graph based on Euclidean distance the scaled PCs of each cell. Edge weights are calculated by Jaccard similarity of each node's 1st degree neighbours, and then low-weight edges are pruned to yield a network where cells are connected if their expression is similar to each other and their respective similar cells. Communities are defined in this network using the Louvain method for modularity optimization (Blondel et al., 2008). *SC3* approaches scRNAseq analysis by consensus – it applies multiple distance metrics, dimensionality reduction techniques, and number of features retained. It then performs k-means clustering on each, summarizes those results as an average of binary similarity matrices, and finally performs hierarchical clustering on the consensus matrix. This is robust but computationally expensive, so it is notable that *PhenoGraph* performs similarly while being much more scalable.

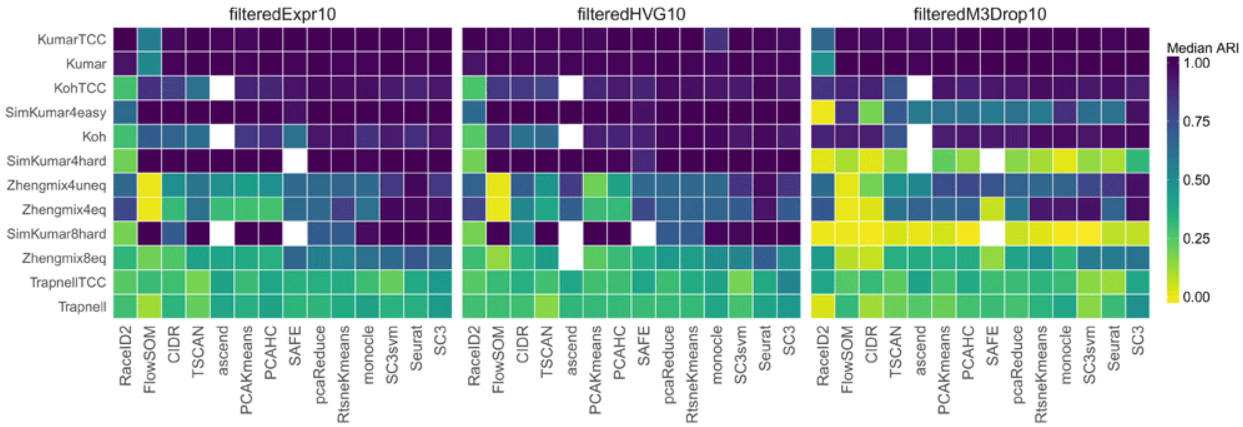


Figure 1-8. Benchmarking scRNAseq clustering methods.

Performance of clustering methods (columns) measured by Adjusted Rand Index (ARI) at capturing known structure in simulated or published scRNAseq datasets (rows). Input features were selected by expression, variance, or dropout rate (Duò et al., 2018).

The outstanding challenge in clustering remains determining how many cell types exist in an scRNAseq dataset. This is an explicit or implicit input parameter of all clustering methods; k-means sets a specified number of clusters, while input parameters can affect the size of communities detected by the Louvain algorithm (Lambiotte et al., 2008). There are methods to determine the appropriate number of clusters prior to clustering. *SC3* implements a method based on random matrix theory, using the Tracy-Widom distribution to identify significant PCs by their eigenvalues (Tracy and Widom, 1994; Patterson et al., 2006). More common are post-hoc methods for assessing the quality of various cluster solutions, though these increase computational burden by requiring multiple clustering solutions to be calculated. For example, if there is stochasticity in the method used to cluster, one can repeat clustering at different resolutions and assess the stability of the resulting clusters, or enforce stochasticity by subsampling as implemented in *scclusteval* (Tang et al., 2021). A variation of this approach is to embed the cluster solutions into a tree structure, where each level of the tree has a different cluster solution with nodes as clusters and edges between levels representing shared cells (Zappia and Oshlack, 2018). This allows researchers to visualize cluster stability and the effect of changing the number of clusters on each putative grouping of cells. Finally, one can assess the quality of an individual cluster solution using metrics such as the silhouette, which calculates cluster cohesion in Euclidean space (Rousseeuw, 1987). In Chapter 2 I propose a novel, biologically motivated method for evaluating clustering of scRNAseq data.

Finally, once single-cell libraries are clustered into putative cell types, these cell types can be identified. This has been covered in detail in our recently published protocol (Clarke et al., 2021), but I will briefly summarize some approaches to the problem. Prior to the explosion of available single-cell data, annotation of single-cell transcriptomes required extensive domain knowledge of the assayed tissue. Thanks to the popularity of cell purification methods such as fluorescence-activated cell sorting (FACS) that isolate a cell type by its uniquely expressed surface proteins, many cell types have known sets of marker genes. Databases such as *MSigDB* now include lists of cell type gene sets (Liberzon et al., 2015). Given this knowledge, one can simply compare known marker genes to those genes most differentially expressed in each cluster. Identifying these marker genes per cluster is further discussed in Chapter 2, although this marker-based cluster annotation process can be automated using *GSVA* (Hänzelmann et al., 2013). However, some cell types are better defined histologically than molecularly, and may not have known marker genes. To annotate cell libraries from such tissues, one must be able to map the scRNAseq data back to its spatial position in the tissue. While spatial transcriptomics is rapidly improving and becoming more available, one could combine scRNAseq and fluorescence *in situ* hybridization (FISH) to accomplish this task. First, identify sets of genes expressed in each cluster, the combination of which is unique to each cluster (Dumitrascu et al., 2021). These gene sets can then be used in a multiplexed FISH assay to link each cluster to its histological information, and thus annotate the cell transcriptome. With the increasing availability of published scRNAseq data, these methods may no longer be needed, at least for major, commonly observed cell types. Instead, a classifier such as a Support Vector Machine (SVM) can be trained using a reference scRNAseq data from your tissue of interest (Tabula Muris Consortium et al., 2018; Tabula Sapiens Consortium et al., 2022). One of the top performing methods for this task is *scmap*, which is notable for its ability to distinguish unknown cell types from those present in the reference dataset (Kiselev et al., 2018; Abdelaal et al., 2019).

This section has outlined the steps of a robust workflow for the analysis of scRNAseq data with the aim of clustering single-cell transcriptomes into annotated cell types. This workflow is the product of a many scientists in the single-cell community, contributing not only novel methods, but helping develop a consensus on best practices, and creating toolsets that enabled the wider community to use the methods and data generated. This workflow is useful for identifying the transcriptomes of all cell types present in an assayed tissue, though continued research is expected to refine and extend it over time. From here we can begin to predict how these cells may be communicating to coordinate tissue function.

1.3 Intercellular signaling from single-cell transcriptomics

1.3.1 Modelling signaling between cell types of a tissue

To coordinate tissue function, evolution has yielded a variety of biochemical methods by which cells communicate. We generally classify these signaling methods by the distance between sender and receiver cells. Autocrine signaling refers to situations where cells use intercellular signaling mechanisms to modulate their own behaviour. As cells necrose in an injured tissue, they secrete the ligand interleukin-1 (IL-1), which is picked up by the IL-1 receptor on tissue-resident macrophages. The macrophages respond by producing more IL-1 as well as other pro-inflammatory cytokines in an autocrine signaling positive feedback loop that initiates the inflammatory response (Kaneko et al., 2019). Cells often signal to their neighbours to coordinate tissue function, especially pattern formation during development. This is juxtacrine signaling, unique in that both ligand and receptor are membrane bound. Interactions between the delta ligand and notch receptor in spatial organization of *Drosphila* nervous system are a classic example of juxtacrine signaling (Alberts et al., 2007). Paracrine signaling involves secreted ligands diffusing through a tissue to confer information to other cells. For example, the morphogen sonic hedgehog is secreted from the notochord and diffuses through the epithelial cells of the developing neural tube, interacting with its transmembrane receptor patched1 to provide ventral fate cues (Garcia et al., 2018). Finally, signaling that passes between tissues via the circulatory system is termed endocrine signaling. Endocrine signaling ligands can be either peptide hormones such as insulin, or steroid hormones such as estrogen (Alberts et al., 2007).

Our lab has used the transcriptome to predict how different cells of a tissue may interact for the past decade, beginning with the transcriptomes of isolated hematopoietic cell types collected by microarray (Kirouac et al., 2010; Qiao et al., 2014). Inspired by earlier work that sought to apply the graph theory analyses made popular by the rise of online social networks to the interactions of the immune system (Frankenstein et al., 2006), they added transcriptional information as a way of associating nodes of the dense cytokine interaction network with their contributing cell types. This allowed them to study how the cell-cell interaction networks changed with hematopoiesis and develop hypotheses about how individual signals guide cell fate.

The method used to model these cell-cell interactions is relatively simple. First, a reference network of ligand-receptor interactions must be curated from available data. They began by using COPE, a database of cytokines curated as a passion project by one man from the early days of the web (Ibelgaufts, 1997). Cell type specific transcriptomes are then used to associate ligands and receptors with each cell type. To attain transcriptomes for each cell type, they isolated cells by FACS prior to collecting RNA for transcriptome determination by microarray hybridization. Ligands and receptors were associated with each cell type if their gene expression was above a certain threshold. Using this approach, the resulting cell type specific ligand-receptor interaction network models both autocrine, juxtacrine, and paracrine interactions. Autocrine interactions are predicted when a ligand and its cognate receptor are present in the same cell type, though this does not necessarily indicate that an individual cell is modulating its own function using that interaction. When a pair of cell types each present a ligand or its cognate receptor, a paracrine interaction is predicted between those cell types. Juxtacrine interactions can be predicted in the same way, though the physical association between cells required for juxtacrine signaling is not captured in these transcriptional assays. Finally, endocrine signaling is outside the scope of this assay, both literally in that the sender and receiver cells can reside in separate tissues with the signal transported via the circulatory system, and because at least for steroid hormones the ligands are not gene products.

This model makes considerable assumptions to predict protein-based intercellular signaling from transcriptomes, as summarized in **Figure 1-9**. Essentially, if ligand and receptor genes are transcribed in their respective cells, the model assumes that they are able to interact. In order for an interaction to occur between these gene products, a number of events must occur (Alberts et al., 2007). Both mRNA must be translated by the ribosome, a step under post-transcriptional regulation (Zahr et al., 2018). Ligand peptides must undergo the requisite post-translational modification and be packaged for secretion by exocytosis. They must then diffuse through the extracellular matrix, presuming no impermeable membrane blocks their path to the receptor-expressing cell. Meanwhile the receptor must be post-translationally modified, integrated into a phospholipid bilayer, and presented at the cell surface. It must also have the requisite co-receptors available. None of this information is readily available from a transcriptomic assay, so there is a high likelihood of false-positive predictions from this model of intercellular signaling. However, the advent of single-cell RNAseq has made this model very attractive by eliminating the need to isolate cell types *in vitro*. Many adaptations and expansions on this model have now been developed for scRNAseq data, as catalogued in **Appendix A**. The following sections will review these advances, both in the generation of ligand-receptor interaction networks and modeling these interactions in a tissue from single-cell transcriptomes.

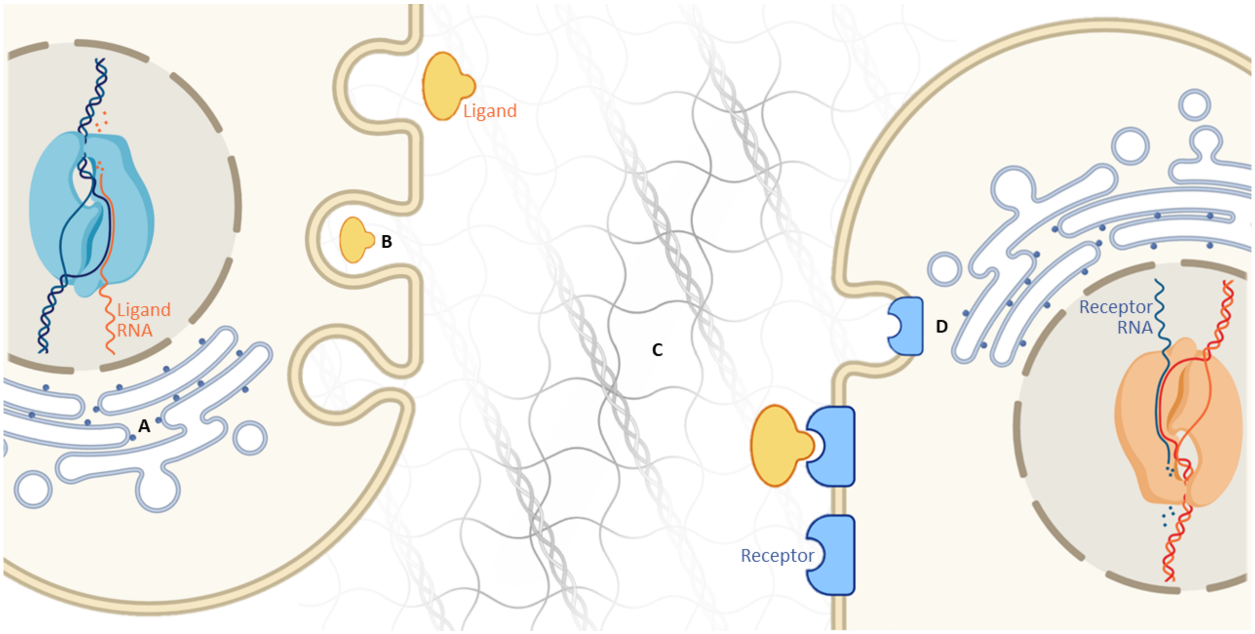


Figure 1-9. Assumptions inherent in predicting intercellular signaling from transcriptomics. Many things must happen for an interaction between a ligand from one cell (orange, left) and its cognate receptor on another cell (blue, right) to occur. **A**. Both gene products must be translated by the ribosome, and packaged for transport in the golgi. **B**. Ligands must be exocytosed and **C**. diffuse through the extracellular matrix to the receptor-expressing cell. **D**. Receptors must be appropriately post-translationally modified and presented on the cell membrane. Created with BioRender.com

1.3.2 Curating ligand-receptor interaction networks

To model cell-cell interactions from the transcriptome, we must first identify ligand- and receptor-encoding genes. Traditionally our understanding of ligand-receptor interactions stems from hypothesis-driven science, where through a combination of genetics and biochemistry a phenotype has been determined to be the result of an interaction between these two gene products. Mutagenesis may be used to systematically knock out genes until the phenotype is abrogated, thereby identifying a gene involved in the phenotype. Cloning the gene in conjunction with a fluorophore allows its subcellular location of its protein product to be determined, potentially identifying it as a secreted ligand or membrane-associated receptor. Protein-protein interaction assays such as affinity chromatography or two-hybrid screening may then identify interacting proteins, and further genetic experiments could determine which interactions are necessary and sufficient for induction of the phenotype in question. This is not a high-throughput process, and thus early databases of ligands and receptors relied entirely on expert curation of primary literature.

While experimental evidence of a functional interaction between ligand and receptor remains the gold standard, high throughput assays and computational predictions have expedited the cataloguing of potential intercellular signaling proteins. For example, a cell’s secretome is readily accessible to proteomics assays thanks to the solubility of secreted proteins (Tjalsma et al., 2004). Secreted proteins can also be predicted from sequence, since they canonically require a signal peptide to be processed by a cell’s secretory machinery, and recent efforts have built models to identify those secreted through non-canonical methods (Voet and Voet, 2010; Wang et al., 2022). Similarly, membrane associated proteins are identifiable by the unique biochemistry of their transmembrane domains, a feature used by many structure prediction algorithms to infer membrane association (Fagerberg et al., 2010; Baxevanis et al., 2020). Not all secreted or membrane-associated proteins are ligands or receptors. But inherent in broadening the scope of effort to catalogue ligands and receptors is the assumption that if a secreted protein is shown to physically interact with the extracellular domain of a membrane associated protein, they may be involved in an intercellular signaling process. Thus we must also have high-throughput methods for identifying protein-protein interactions. Two-hybrid systems such as yeast-2-hybrid have served this purpose, with extracellular domains of putative receptors being expressed as single domains to improve their solubility in the yeast cytoplasm (Brückner et al., 2009). Juxtacrine interactions are overlooked by this model, however, as both the ligand and receptor are membrane-associated. Novel interaction screening methods have been developed to capture these interactions, including a method that uses CRISPR activation to systematically overexpress individual putative receptors to ensure high avidity, allowing the cells expressing the correct receptor to be isolated by a bait consisting of the extracellular domain of a membrane-bound ligand (Chong et al., 2018). Thanks to such high-throughput methods, databases now exist to catalogue the subcellular location and physical interactions for most proteins (Razick et al., 2008; Rodchenkov et al., 2020; Oughtred et al., 2021; UniProt Consortium, 2023).

As the biochemical databases have become more widespread, the curation of ligand-receptor databases has become more of a database-parsing exercise, as exemplified by the workflow used to build the widely used FANTOM5 database shown in **Figure 1-10** (Ramilowski et al., 2015). They began with three existing databases: the Database of Ligand-Receptor Partners (DLRP), curated manually to support one of the first attempts at cell-cell interaction prediction from the transcriptional data (Graeber and Eisenberg, 2001); the Human Plasma Membrane Receptor (HPMR) database, a collection of receptors for the purpose of sequence phylogeny analysis (Ben-Shlomo et al., 2003); and the International Union of Basic and Clinical Pharmacology (IUPHAR) database, a regularly maintained database of druggable receptors and their endogenous and pharmacologic interactors (Harmar et al., 2009; Harding et al., 2022). Since only IUPHAR's database is still maintained, and HPMR contained both ligands and receptors lacking interactors ("orphans"), they then aimed to expand these lists by inferring ligand-receptor interactions. They first used subcellular location information from UniProt, a general protein knowledgebase maintained by the European and Swiss bioinformatics institutes (UniProt Consortium, 2023), and the Human Protein Reference Database (HPRD), a database of experimentally determined protein information (Keshava Prasad et al., 2009). Proteins known or predicted to be present at the plasma membrane were predicted to be receptors, unless already classified as known ligands, and proteins known or predicted to be secreted were inferred to be ligands. To link both orphan and predicted ligands to their putative receptors, physical protein-protein interaction evidence was sourced from HPRD and STRING, a regularly maintained database of known and predicted physical and functional protein interactions (Szklarczyk et al., 2015). Finally, an attempt was made to associate citations with each interaction, though interactions lacking primary literature evidence were only discarded if one of the partners was inaccurately annotated as a ligand or receptor. This workflow was highlighted as an example that touches on many of the considerations involved in the production of a ligand-receptor database, but is also noteworthy as the FANTOM5 database has been used as the exclusive ligand-receptor resource for one third of the 56 novel cell-cell interaction prediction methods developed since its publication in 2015 (**Appendix A**).

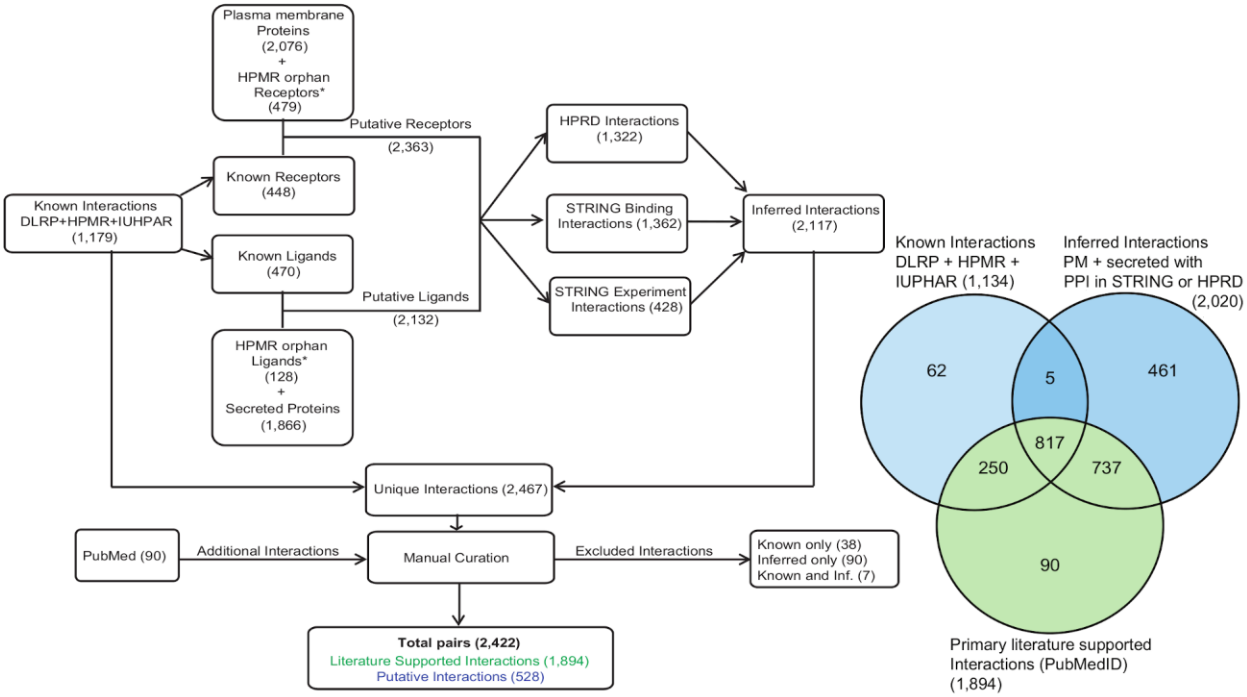

Plasma membrane Proteins (2,076) + HPMR orphan Receptors* (479)
Putative Receptors (2,363)
Known Receptors (448)
Known Interactions DLRP+HPMR+IUHPAR (1,179)
Known Ligands (470)
Putative Ligands (2,132)
HPMR orphan Ligands* (128) + Secreted Proteins (1,866)
HPRD Interactions (1,322)
STRING Binding Interactions (1,362)
STRING Experiment Interactions (428)
Inferred Interactions (2,117)
Unique Interactions (2,467)
PubMed (90)
Additional Interactions
Manual Curation
Excluded Interactions
Known only (38)
Inferred only (90)
Known and Inf. (7)
Total pairs (2,422)
Literature Supported Interactions (1,894)
Putative Interactions (528)
Known Interactions DLRP + HPMR + IUPHAR (1,134)
Inferred Interactions PM + secreted with PPI in STRING or HPRD (2,020)
62
5
461
817
250
737
90
Primary literature supported Interactions (PubMedID) (1,894)

Figure 1-10. Example of ligand-receptor database curation.
The *FANTOM5* database curation workflow. Adapted from (Ramilowski et al., 2015).

Many of the cell-cell interaction prediction methods that don't use the FANTOM5 database curate their own using a similar workflow, albeit curating from different databases, including most of the databases compared in **Figure 1-11**. That is reflected in the 334 ligand/receptor nodes present in all but the smallest database compared. That smallest database is ICELLNET (Noël et al., 2021), which focused solely on immune signaling. The three databases with the more unique nodes took a notably different approach to database curation. CellChat (Jin et al., 2021) and LIANA (Dimitrov et al., 2022) focused more heavily on the ligands and receptors involved in intracellular signaling pathways, using KEGG (Kanehisa et al., 2016) and OmniPath (Türei et al., 2021) respectively. The largest database is our contribution, CCInx (Ximerakis et al., 2019), which leaned heavily on the large Gene Ontology (GO) classification of protein subcellular location and function to assign ligands and receptors. Another notable difference in the curation of ligand-receptor databases is explicit consideration of receptor complexes, first introduced by CellPhoneDB (Vento-Tormo et al., 2018; Efremova et al., 2020) and expanded upon in CellChat (Jin et al., 2021), where agonist and antagonist cofactors have also been curated so that their function can be modeled in interaction scoring.

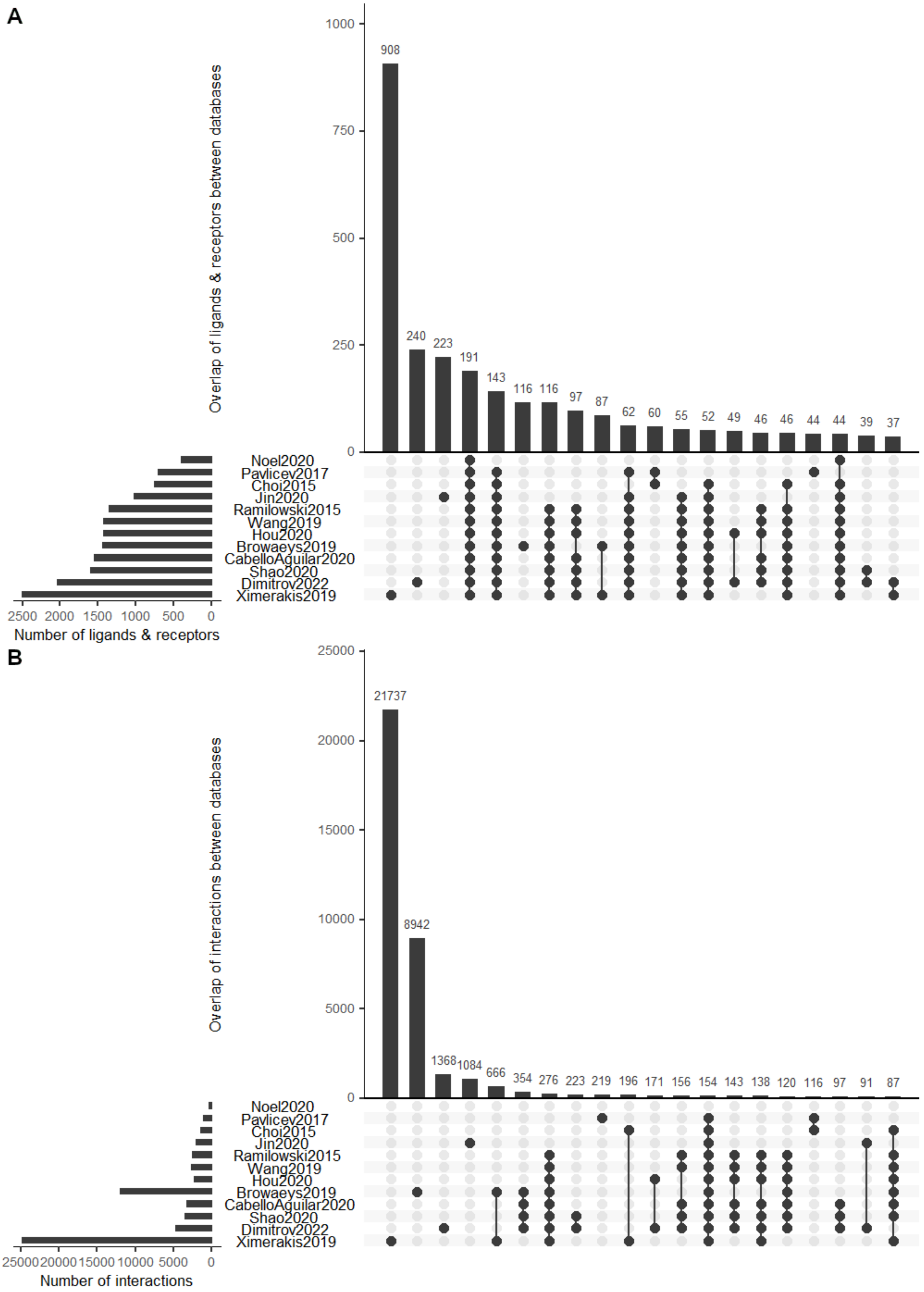

A
Overlap of ligands & receptors between databases
1000
750
500
250
0
908
240
223
191
143
116
116
97
87
62
60
55
52
49
46
46
44
44
39
37
Noel2020
Pavlicev2017
Choi2015
Jin2020
Ramilowski2015
Wang2019
Hou2020
Browaeys2019
CabelloAguilar2020
Shao2020
Dimitrov2022
Ximerakis2019
2500
2000
1500
1000
500
0
Number of ligands & receptors
B
Overlap of interactions between databases
25000
20000
15000
10000
5000
0
21737
8942
1368
1084
666
354
276
223
219
196
171
156
154
143
138
120
116
97
91
87
Noel2020
Pavlicev2017
Choi2015
Jin2020
Ramilowski2015
Wang2019
Hou2020
Browaeys2019
CabelloAguilar2020
Shao2020
Dimitrov2022
Ximerakis2019
25000
20000
15000
10000
5000
0
Number of interactions

Figure 1-11. Comparison of ligand-receptor databases used for cell-cell interaction prediction. **A.** An UpSet plot (Lex et al., 2014) showing overlap of ligands and receptors included in the backing databases of recent cell-cell interaction inference methods reviewed by (Armingol et al., 2021). **B.** An UpSet plot comparing overlap of interactions. Datasets are ordered as in A.

Despite being curated using similar curation workflows, ligand-receptor databases differ a surprising amount in their links between ligand and cognate receptor (**Figure 1-11b**). As noted in a recent benchmarking study, this contributes to the lack of consensus in findings from different methods (Dimitrov et al., 2022). This was a major motivation of our work on the database backing CCInx. By using a broad definition for ligands, receptors, and their interactions, we aimed to reduce the amount of bias introduced by our database, at the cost of decreasing the specificity of the resulting predictions (Isserlin et al., 2020). InterCellDB takes this concept even further, as it indexes all human and mouse proteins, providing annotations of location, function, and interactions where available, and allows the user to set their own filters to define their required interaction database (Jin et al., 2022). One method completely omits the use of a ligand-receptor network as a prior, building a network of cell-cell interactions purely based on coexpression of gene modules, though it lacks the resolution to infer individual ligand-receptor interactions as a result (Chen et al., 2022). The problem with reducing the reliance on prior knowledge of ligand-receptor interactions is that it puts more reliance on the method used to infer specific interactions from transcriptional data. As the following section outlines, there is no clear solution to that problem.

1.3.3 Methods to predict cell-cell interactions from single-cell RNAseq

With a database to map scRNAseq transcript counts to putative ligands and their cognate receptors, the final step is the prediction itself – how to weight the edge represented by each ligand-receptor pair expressed in a pair of cell types. Prior to cell type-specific resolution of transcriptomes, correlation of ligand and cognate receptor expression across cell type pairs was used to score interactions, though this makes the implicit assumption that receptor activation drives a positive feedback loop that increases receptor expression (Graeber and Eisenberg, 2001). That assumption was more valid when considering autocrine interactions in cancer, but is less generally relevant, which might explain why few modern methods employ ligand-receptor correlation scores. Instead, many of the methods outlined in **Appendix A** use either the arithmetic or geometric mean of normalized expression of ligand and receptor to score each interaction. The advantage of the geometric mean is that it preserves zeroes – if either interactor is not expressed, the interaction score will be zero. While this is convenient, most of the methods using the arithmetic mean also implement a filter requiring sufficient expression of both ligand and receptor, so in practice the difference between the different mean implementations is inconsequential. These techniques aggregate the transcript counts from clusters of single-cell libraries representing putative cell types which increases the number of genes detected, so the resulting ligand-receptor interaction networks tend to be very large. Various filtering strategies have been adopted, including requiring a minimum transcript count for inclusion of a ligand or receptor node in the network. While using a fixed count runs the risk of being arbitrary, there is some evidence that a threshold of 10 transcripts per million (TPM) maximizes the likelihood of the resulting gene product being present in that cell (Ramilowski et al., 2015). An alternative to fixed thresholds of expression is to use frequentist statistics to determine which nodes or edges are notable. Some methods use this as their node filter, considering only ligands or receptors that are significant "marker genes" for their cluster, genes significantly positively differentially expressed in the specific cell cluster relative to all cells. Others, notably CellPhoneDB, build the ligand-receptor network based on gene expression and then use a statistical model to test cell type pair specificity of each edge (Vento-Tormo et al., 2018). Caution must be used when interpreting these models, however, since they are no longer prioritizing edges solely on the likelihood that the interaction is occurring. In fact, none of the scoring methods discussed so far address the disconnect between ligand and receptor transcript expression, and the interaction of their protein products in a tissue. This motivated the development of my scoring method CCInx, published in our collaboration studying age-mediated changes in the murine brain transcriptome (Ximerakis et al., 2019). Using our lab's large ligand-receptor database to reducing literature bias as much as possible, CCInx weights nodes and edges by differential expression magnitude between tested states. While the resulting networks are large, the ranking highlights interactions most strongly associated with the tested hypothesis. I reasoned that if the scoring method is unlikely to improve accuracy, then it can at least help generate meaningful hypotheses associated with the question under investigation.

Beyond scoring ligand-receptor interactions, methods have also considered cell-cell interaction inference from the scale of whole cell types rather than their individual ligand-receptor pairs.

The simplest method used to weight a cell-cell edge is to simply sum the number of discrete ligand-receptor pairs predicted to interact between two cell types. If the ligand-receptor edge has an associated weight, often those weights are summed rather than treating each edge discretely. As with ligand-receptor interactions, a frequentist approach can be used to filter these cell-cell edges for statistical significance, either by comparing summed ligand-receptor edge weights to a null distribution (Farbehi et al., 2019), or testing for specificity of those interactions to the cell type pair (Smillie et al., 2019; Armingol et al., 2022b). An alternative to aggregating ligand-receptor edges to form a cell-cell edge is to consider both. Hypergraphs are networks where a cell type node can contain ligand and receptor nodes, and cell-cell edges can be composed of ligand-receptor edges (Tsuyuzaki et al., 2019). Tensors are multidimensional matrices, allowing for 2D matrices of ligand-receptor expression across all sender and receiver cells to be arranged in a 3D tensor containing all ligand-receptor pairs (Armingol et al., 2022a). A fourth dimension representing experimental context can be added, allowing tensor factorization methods to determine associations with experimental context at the ligand-receptor and cell-cell level concurrently. There are methods that focus solely on infering interactions at the cell type level, using gene correlation networks to build causal Bayesian networks predicting the degree to which a cell type impacts the phenotype of other cell types in the tissue (Chen et al., 2022; Yuan et al., 2022). Finally, recently there has been a shift in focus from cell-cell networks to cell-niche networks. Rather than considering both ligand- and receptor-expressing cell, the focus is on the receptors of a single cell type and considers the proportions of all available ligands (by weighting ligand expression by proportion of ligand-expressing cells) in an attempt to predict how the cell's phenotype is a product of the niche (Griffiths et al., 2022; Raredon et al., 2023).

Since the publication of my CCInx method, novel ligand-receptor scoring methods have been published that aim to have edge weight reflect the likelihood of interaction. CellChat uses a database containing ligands, receptors, and their agonist and antagonist cofactors, and uses the expressions of all contributing factors in lieu of concentrations when modeling each interaction using the law of mass action (Jin et al., 2021). This is a theoretical improvement on the more basic scoring methods as it accounts for the fact that ligand-receptor interactions are not independent events, and require or are modulated by various cofactors. Another method that considers confounding interactions is REMI, which uses partial correlation of expression between nodes in the ligand-receptor interaction network to build a graphical Bayesian model of conditional dependence amongst ligand-receptor interactions (Yu et al., 2022). The assumption here is that if cells are communicating through a ligand-receptor interaction, the expression of ligand and receptor will be tuned either in direct response to the interaction or through other axes of communication between the cell pair, and thus their expression is not independent. Finally, there are multilayer network models such as the popular NicheNet that use the transcriptome of the receptor-expressing cell to support ligand-receptor interaction inference (Browaeys et al., 2020). They do this by connecting ligand-receptor interactions to the downstream intracellular signaling pathway of each receptor, and then linking the terminating transcription factors of these signaling pathways to their target genes. This is heavily dependent on prior knowledge: databases such as OmniPath (Türei et al., 2016) and Pathway Commons (Cerami et al., 2011) contain intercellular signaling networks, and TRRUST (Han et al., 2015) and JASPAR (Mathelier et al., 2014) can associate transcription factors with the genes they may regulate. However, as the authors of NicheNet discovered, these networks can branch extensively resulting in many genes putatively being regulated by a single ligand-receptor interaction. Their solution was to curate experiments where the transcriptome was measured in response to ligand perturbation and use these to train weights for their multilayer network to improve its ability to predict transcriptional response to ligand-receptor interaction. But the question remains: do the edge weights returned by this method reflect the likelihood of the actual interaction occurring in a tissue?

Evaluating accuracy of ligand-receptor inference is very difficult, as there is not sufficient labeled data to test on. One may have high confidence that individual ligand-receptor interactions are occurring in their tissue of interest due to prior knowledge (e.g., growth factors patterning the developing *Drosphila* nervous system) but it would be challenging to make such an evaluation systematic and the reliance on known interactions encourages literature bias. No studies have linked cell type resolved transcriptomes of a tissue with interactions involving the cell surface proteome of the tissue-resident cell types, which would be required for an unbiased assessment of ligand-receptor inference methods. Two groups have made efforts to benchmark these methods, using various strategies in lieu of a testing dataset with known ligand-receptor interactions. One group constructed LIANA, an evaluation framework to decouple the ligand-receptor database from the interaction scoring method (Dimitrov et al., 2022). They found that despite using the same set of ligand-receptor interactions, there was very little consensus among scoring methods. This reinforces the importance of understanding what hypothesis each scoring method is testing when applying them as a researcher. It also gives credence to the notion that few if any ligand-receptor interaction scores reflect the likelihood of those interactions occurring *in situ*. To evaluate prediction accuracy, they compared inferred ligand-receptor interactions to predictions from CytoSig, a predictive model of cytokine activity based on systematic experimental transcriptomic assays (Jiang et al., 2021). The methods were evaluated on two breast cancer CITE-Seq datasets, where the receptor availability per cell was confirmed by antibody binding (Wu et al., 2021). To quantify their comparison, they calculated odds ratios between the top ranked ligand-receptor interactions and predicted active cytokines across a range of ranks (**Figure 1-12a)**. Results were inconsistent, with some methods performing well on one dataset and very poorly on the other, and other methods showing more consistently moderate performance. Another approach to evaluating ligand-receptor interaction inference assumes that spatially proximal cells will interact with each other more often. A benchmarking study based on this assumption assigned ligand-receptor interactions as short- or long-range on the basis of distribution of expressed genes in the spatial transcriptomics datasets used for testing (Liu et al., 2022). They then evaluated ligand-receptor interaction predictions for their ability to predict short-range interactions between proximal cell types and long-range interactions between distant cell types. By this evaluation criteria they found that CellChat was the top performer (**Figure 1-12b**), though it is unclear whether their usage of CellChat's ligand-receptor database in the construction of their evaluation criteria introduced bias in the results. The LIANA authors also used proximity of cell-cell pairs as an evaluation metric with notably different conclusions. The evaluation metrics used in these benchmarking efforts are meant to serve as a proxy for a (non-existent) transcriptomic dataset with known interactions, but without the ability to evaluate their relationship with that ground truth, these benchmarking studies suffer from the same limitations as the methods they are testing – it is difficult to be confident in their results.

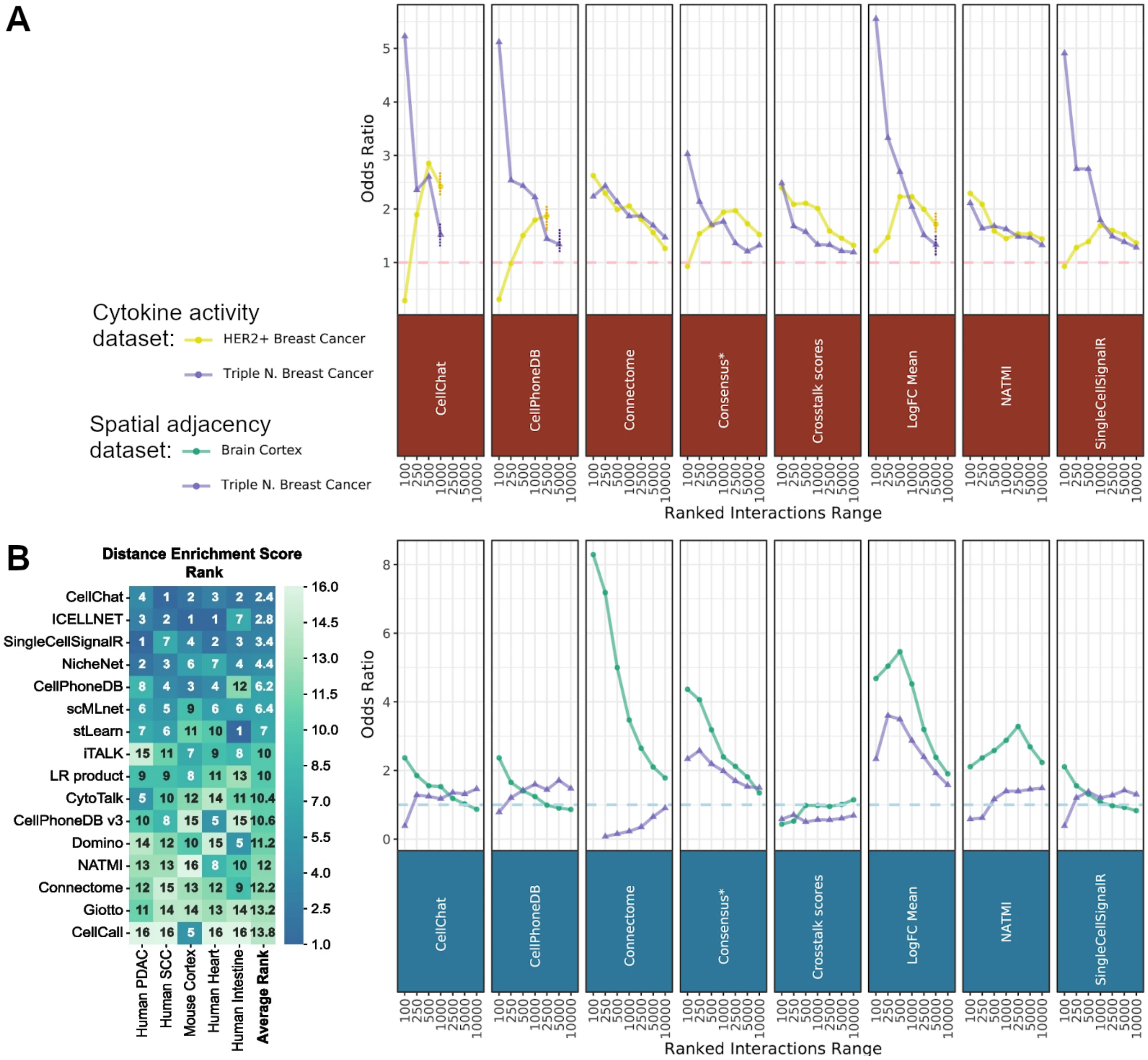


Figure 1-12. Benchmarking ligand-receptor inference methods.
A. Comparing CytoSig cytokine activity predictions with top ranked ligand-receptor interaction predictions from various methods (Dimitrov et al., 2022). **B** Left. Ranking ligand-receptor inference methods by their ability to associate expected short- and long-range interactions with nearby and distant cell pairs respectively (Liu et al., 2022). **B** Right. Comparing cell pairs involved in top ranked interaction predictions with spatially adjacent cell types (Dimitrov et al., 2022). Adapted from cited papers.

While the rise of high-throughput single-cell transcriptomics has led to a renewed interest in inferring intercellular signaling from the transcriptome, it is unclear whether any of these novel methods accurately reflect the realities of intercellular signaling in a tissue *in vivo*. Despite this they still have utility as a hypothesis-generating tool, in conjunction with follow-up assays. To make the most of these methods, its important to have a specific hypothesis in mind, and select a method whose assumptions and ranking criteria reflects the hypothesis being tested. Without this focus, the fundamental problems plaguing this field become apparent – poor specificity or filters that don't reflect the probability of the interaction occurring *in situ* means you either get biased results or are swamped with too many hypotheses to test.

1.4 Outline & Rationale

The advent of high-throughput single-cell transcriptomics has created an opportunity to systematically generate molecular models of tissue function. To create such a model, the cell types of the tissue must be identified, and communication between cell types mapped. Resolving scRNAseq libraries into cell type transcriptomes allows for both identification of cell types and inference of ligand-receptor interactions between them. This thesis aims to facilitate the generation of intercellular signaling models from scRNAseq data by aiding interpretation of clustering results and guiding improvements to ligand-receptor interaction inference.

In Chapter 2 I demonstrate the software tool scClustViz, which I built to help our collaborators investigate their scRNAseq data. It generates an interactive data report with visualizations tailored to help biologists identify and compare cell clusters. Along with its development, I propose a novel method for determining an appropriate number of clusters and note an error in the way the field represents magnitude of gene expression change caused by data transformations during normalization.

In Chapter 3 I test a fundamental assumption that underlies state-of-the-art multilayer ligand-receptor interaction inference methods, most notably the popular tool NicheNet (Browaeys et al., 2020). The promise of these methods is that they use the available transcriptomic information to identify evidence of ligand-receptor interactions. This requires knowledge of the transcriptomic signatures of ligand stimulus, which is often obtained by inference from published intracellular signaling pathways and gene regulatory networks. There is evidence that gene regulatory networks are cell type specific (Margolin et al., 2006; Chasman and Roy, 2017), which led me to ask whether transcriptional response to a ligand is also cell type specific. I systematically demonstrate that ligand response signatures are not consistent between cell types, but also show that the transcriptome does contain information that may allow these signatures to be inferred.

A model of intercellular signaling creates many opportunities to generate hypotheses around modulating tissue development and function by targeting relevant receptors. My collaboration with the Miller-Kaplan lab on the developing mammalian forebrain serves as a motivating example. In 2016 our labs published a study that inspired my work, identifying the intercellular signaling necessary for neurogenesis in the developing subventricular zone (SVZ), the forebrain region responsible for the generation of glutamatergic excitatory neurons (Yuzwa et al., 2016). The goal of this line of inquiry is to improve adult neurogenesis in response to disease or injury by understanding how neurogenesis is regulated in development. During embryonic neurogenesis, radial precursors of the SVZ make glutamatergic excitatory neurons and persist into adulthood to contribute to an adult neural stem cell (NSC) pool in the forebrain (Gauthier-Fisher and Miller, 2013; Fuentealba et al., 2015). However, in adulthood these NSCs make only GABAergic inhibitory neurons, leading to the prevailing view that they have restricted potency. First, using the software developed in Chapter 2 we established that the embryonic radial precursors of the SVZ share a core transcriptional identity with adult forebrain NSCs (Yuzwa et al., 2017). Then, using lineage tracing to differentiate adult NSCs derived from the SVZ (where their daughter cells make glutamatergic excitatory neurons in development) and the ganglionic eminence (GE, where GABAergic inhibitory neurons are made during development), we showed that NSCs derived from both regions share the previously defined transcriptional signature (Borrett et al., 2020). Furthermore, when these adult NSCs are activated, they become transcriptionally similar to the radial precursors seen in development (**Figure 1-13**). Finally, we showed that SVZ-derived adult NSCs, apparently fate-restricted to make glutamatergic neurons during development, are capable of making GABAergic neurons just like GE-derived adult NSCs. This suggests that fate specification is not fixed but is perhaps guided by intercellular signaling in the stem cell niche.

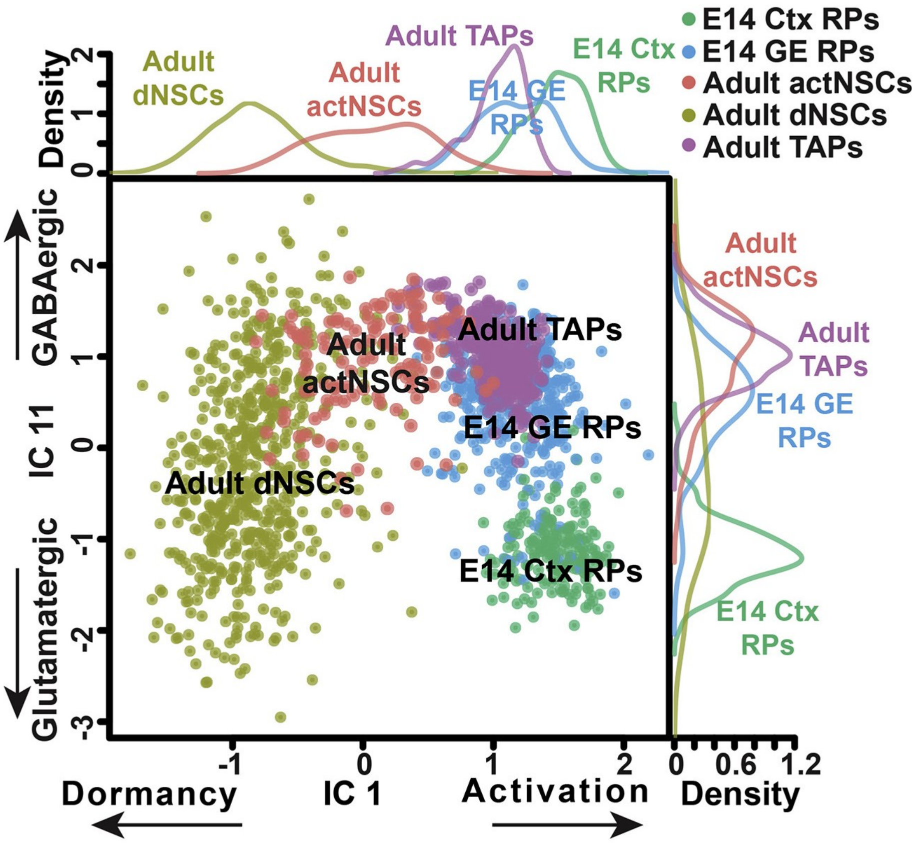


Figure 1-13. The transcriptional states of forebrain NSCs.
Independent Component Analysis (ICA) was used to identify gene expression components uniquely separating adult NSCs from embryonic radial precursors (RPs) and cortical (ctx) and ganglionic eminence (GE) derived RPs. IC1 (x-axis) was thus identified as the gene expression program defining activation versus dormancy, while IC11 (y-axis) separates glutamatergic versus GABAergic fated progeny. Adult activated NSCs (actNSCs) and transit-amplifying precursors (TAPs) were then projected onto this space based on the expression of genes involved in these components. This analysis showed that adult NSCs reacquire a transcriptional state similar to their embryonic counterparts when performing the same task (proliferation and differentiation to GABAergic neurons). Adapted from (Borrett et al., 2020).

There is another population of adult NSCs residing in the hippocampus that make glutamatergic excitatory neurons during adulthood. We investigated these two spatially-distinct adult NSC populations fated to make different types of neurons and found that the hippocampal NSCs were transcriptionally similar, following a dormancy-activation trajectory to the forebrain NSCs previously investigated, and identified the transcriptional programs defining GABAergic versus glutamatergic fate specification (Borrett et al., 2022). Ultimately these studies have led us to hypothesize that adult NSCs are functionally distinct by virtue of their environment rather than intrinsic limits, and thus with the appropriate perturbations to the ligand-receptor signaling guiding their fate specification, we may be able to improve their ability to regenerate the injured brain. These findings relied heavily on analysis facilitated by the software developed in Chapter 2, and my work in Chapter 3 will help guide future predictions of the ligand-receptor networks regulating adult NSC fate specification.

Chapter 2
scClustViz – Single-cell RNAseq cluster assessment and visualization

This work was published in *F1000Research* (Innes and Bader, 2018). I conceived of, designed, and conducted this project. Gary D. Bader supervised and advised on this project.

2 scClustViz – Single-cell RNAseq cluster assessment and visualization

Single-cell RNA sequencing (scRNAseq) represents a new kind of microscope that can measure the transcriptome profiles of thousands of individual cells from complex cellular mixtures, such as in a tissue, in a single experiment. This technology is particularly valuable for characterization of tissue heterogeneity because it can be used to identify and classify all cell types in a tissue. This is generally done by clustering the data, based on the assumption that cells of a particular type share similar transcriptomes, distinct from other cell types in the tissue. However, nearly all clustering algorithms have tunable parameters which affect the number of clusters they will identify in data.

The R Shiny software tool described here, scClustViz, provides a simple interactive graphical user interface for exploring scRNAseq data and assessing the biological relevance of clustering results. Given that cell types are expected to have distinct gene expression patterns, scClustViz uses differential gene expression between clusters as a metric for assessing the fit of a clustering result to the data at multiple cluster resolution levels. This helps select a clustering parameter for further analysis. scClustViz also provides interactive visualisation of: cluster-specific distributions of technical factors, such as predicted cell cycle stage and other metadata; cluster-wise gene expression statistics to simplify annotation of cell types and identification of cell type specific marker genes; and gene expression distributions over all cells and cell types.

scClustViz provides an interactive interface for visualisation, assessment, and biological interpretation of cell type classifications in scRNAseq experiments that can be easily added to existing analysis pipelines, enabling customization by bioinformaticians while enabling biologists to explore their results without the need for computational expertise. It is available at https://baderlab.github.io/scClustViz/.

2.1 Introduction

The development of high-throughput single-cell RNA sequencing (scRNAseq) methods, including droplet-based (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017) and multiplexed barcoding (Rosenberg et al., 2018) techniques, has led to a rapid increase in experiments aiming to map cell types within tissues and whole organisms (Ecker et al., 2017;

Regev et al., 2017; Han et al., 2018; Saunders et al., 2018). The most common initial analysis of such scRNAseq data is clustering and annotation of cells into cell types based on their transcriptomes. Many workflows have been built and published around this use case (Satija et al., 2015; Lun et al., 2016b; Risso et al., 2018b), and many clustering algorithms exist to find cell type-associated structure in scRNAseq data sets (Xu and Su, 2015; Ntranos et al., 2016; Žurauskienė and Yau, 2016; Li et al., 2017; Shao and Höfer, 2017). This paper focuses on how to interpret the results of a scRNAseq clustering analysis performed by existing methods, specifically when it comes to selecting parameters for the clustering algorithm used and analysis of the results. This is implemented as an R Shiny software tool called scClustViz, which provides an interactive, web-based graphical user interface (GUI) for exploring scRNAseq data and assessing the biological relevance of clustering results.

Nearly all unsupervised classification (clustering) algorithms take a parameter that affects the number of classes or clusters found in the data. Selection of the appropriate resolution of the classifier heavily impacts the interpretation of scRNAseq data. An inappropriate number of clusters may result in missing rare but distinct cell types, or aberrantly identifying novel cell types that result from overfitting of the data. While there are general machine-learning-based methods for preventing overfitting, we propose a biology-based cluster assessment method; namely whether you could identify a given cluster-defined cell type in situ using imaging techniques based on marker genes identified, such as single molecule RNA fluorescence in situ hybridization. To identify marker genes and quantify the measurable transcriptomic difference between putative cell types given a clustering solution, scClustViz uses a standard differential expression test between clusters. If there are few differentially expressed genes between two clusters, then those clusters should not be distinguished from each other and over-clustering is likely. The researcher can then select a cluster solution that has sufficiently fine granularity, while still maintaining statistically separable expression of genes between putative cell types.

Once cell types are defined using the clustering method and parameters of choice, the researcher must then go through several data interpretation steps to assess and annotate these clusters and identify marker genes for follow-up experimentation. Before a final clustering result is chosen, it is important to assess the impact of technical factors on clustering. While that may have been done as part of the upstream workflow, it is helpful to see the cluster-wise distribution of technical factors such as library size, gene detection rates, and proportion of transcripts from the mitochondrial genome (Ilicic et al., 2016). To annotate cell types identified by the classifier, it is helpful to see the genes uniquely upregulated per cluster, as well as assess the gene expression distribution of canonical marker genes for expected cell types in the data. Finally, novel marker genes may be identified for a cell population of interest, which requires identifying genes that are both upregulated in the cluster in question and detected sparingly or not at all in all other clusters in the experiment.

We describe scClustViz, an R package that aids this frequently encountered scRNAseq analysis workflow of identifying cell types and their marker genes from a heterogenous tissue sample. The package comprises two parts: a function to perform the differential gene expression testing between clusters for any set of clustering solutions generated by existing scRNAseq analysis workflows, and a R Shiny GUI that provides an interactive set of figures designed to help assess the clustering results, annotate cell types, and identify marker genes. The package was designed with transparency and modularity in mind to ease merging into existing workflows and sharing the results with collaborators and the public. This enables the tool to be of value to both experienced bioinformaticians developing workflows and bench scientists interpreting the results of a scRNAseq experiment.

2.2 Methods

2.2.1 Implementation

We propose a metric for assessing clustering solutions of scRNAseq data based on differential gene expression between clusters. We use the Wilcoxon rank-sum test to evaluate the statistical significance of differential gene expression between clusters (Wilcoxon, 1945). This test was selected based on the rigorous differential expression methodology review carried out by Soneson and Robinson (Soneson and Robinson, 2018). In their testing, the Wilcoxon test had accuracy on par with that of the majority of methods tested (most methods were adequately accurate) and identified sets of differentially expressed genes similar to MAST (Finak et al., 2015) and limma (Ritchie et al., 2015), two popular alternatives. What little bias the Wilcoxon rank-sum test does have tends to be towards genes detected at lower rates in the data (Soneson and Robinson, 2018), which can easily be corrected by using a detection rate filter prior to testing. In terms of power and control of type I error rate, the Wilcoxon test was less powerful than more advanced methods, with a false discovery rate (FDR) more conservative than expected. However, unlike some more complicated tests, the Wilcoxon test is compatible with parallel processing of testing calculations to increase computation speed. Ultimately, the simplicity of the Wilcoxon test made it appealing for default use in this tool, as it is understood by most users, is fast to compute and is available in base R. Alternatively, given the wide variety and constant growth of scRNAseq-specific differential gene expression tests, scClustViz can use the results of any test method that returns measures of effect size and statistical significance.

Two measures of effect size of differential gene abundance are reported by scClustViz: difference in detection rate (dDR) and gene expression ratio (logGER, log2 gene expression ratio). Detection rate refers to the proportion of cells from each cluster in which the gene in question was detected (per cluster gene detection rate). The concept of detection rate in scRNAseq data stems from the low per-cell sensitivity and minimal amplification noise of droplet-based assays. Since there is a correlation between gene expression magnitude and per cluster gene detection rate, the detection rate is a meaningful quantification of gene expression. Furthermore, it is suitable for identifying genes that uniquely “mark” certain cell populations, as such marker genes should be undetected outside of the cells they mark.

Log gene expression ratio (also known as log fold change) is a measure of effect size that considers both magnitude of gene expression as well as detection rate, as it is the ratio of mean gene abundance between two cell clusters. However, due to the sparsity of scRNAseq data, some clusters may not contain any cells in which a certain gene was detected. It is thus necessary to add a pseudocount to the logGER calculations to prevent divide-by-zero errors and the resulting logGER magnitudes of infinity. As exemplified in **Figure 2-1**, the choice of pseudocount impacts logGER results. A pseudocount of 1 is commonly used in the field of transcriptomics but creates two problems when used on the low abundance values common to droplet-based scRNAseq data. Since a value of 1 is a considerable fraction of small count data, adding 1 to all counts tends to compress the magnitude of the gene expression ratio in a manner that inversely correlates with the magnitude of abundances being compared (**Figure 2-1a**). As a result, not only is the calculated logGER less than true logGER, but this compression of true logGER is more pronounced when at least one side of the comparison has values near zero. Using a small pseudocount such as 10-99, on the other hand, results in logGER values being very close to their true value, rather than suffering from the compression caused by the pseudocount of 1 (**Figure 2-1b**). The problem with this is that comparisons with zero result in very high magnitude logGER values, well outside the range of the rest of the results. If zero counts of a transcript in a cell library truly represented that gene not being expressed at all in that cell (i.e. if high-throughput single-cell RNAseq experiments were exquisitely sensitive), then this wouldn't be a problem, since the true expression ratio would be infinitely large. However, zero counts are better interpreted to mean that transcripts for the gene in question were not detected in that cell. Given the relatively poor sensitivity of current high-throughput scRNAseq technology on a per cell basis, this does not necessarily mean that the gene was not expressed. Thus, it would be better if logGER values for comparisons with zero were reasonably close in magnitude to the rest of the results. To accomplish this, we use a pseudocount representing the smallest possible "step" in the count-based data, set to the reciprocal of the number of cells in the data. This is sufficiently small as to not compress logGER magnitudes, while keeping comparisons with zero reasonably close to the range of potential logGER values. In scClustViz, the reported logGER values are ratios of log-mean gene abundance calculated using the reciprocal of the number of cells in the data (the smallest possible "step" in the cDNA count) as the pseudocount.

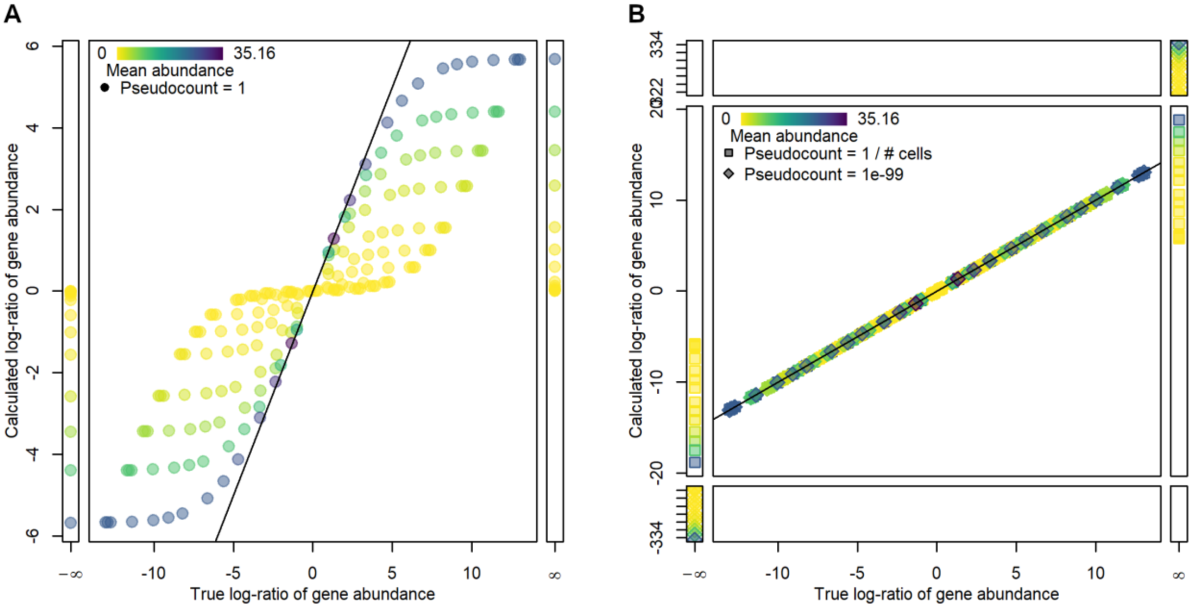


Figure 2-1. Mean and log gene expression ratio (logGER) calculations are affected by selection of the pseudocount used to prevent divide-by-zero errors.

This is demonstrated by calculating log gene expression ratios for pairwise comparisons from a simulated scRNAseq data set where the mean abundance of a single gene varied from zero to 50 across 15 clusters. Code to generate this figure is available in the scClustViz folder of the R library under paperFigs/Fig1.R **A.** A scatter plot comparing true logGER (x-axis) with logGER calculated with a pseudocount of 1 (y-axis). Points are coloured by the mean gene abundance of the comparison, with darker being larger. The black line denotes equality between x- and y- axes. With a pseudocount of 1, the magnitude of logGER is compressed at both ends relative to true logGER, and the magnitude of this compression is inversely correlated with gene abundance in the clusters being compared. **B.** Same plot comparing true logGER with logGER calculated with pseudocounts of 1e-99 (diamonds) and 1 / # of cells (squares). Calculated logGER are very close to true logGER when using smaller pseudocounts (as denoted by the black line). When using a very small pseudocount of 1e-99, the magnitude of logGER values are over 300 when comparing a cluster with zero gene abundance (division-by-zero resulting in a true logGER magnitude of infinity). This is far from the range of other logGER values. An alternative is to set the pseudocount to the smallest possible "step" in count-based data (1 / # of cells) to prevent magnitude compression of logGER calculations caused by using a pseudocount of 1, while keeping division-by-zero values within the range of the data.

Three different sets of differential gene expression results are reported by scClustViz. These are the results of two sets of hypothesis tests: each cluster versus the rest of the data combined (calculated by the function CalcDEvsRest), and all pairwise comparisons between clusters (calculated by the function CalcDEcombn). These comparisons are made using the Wilcoxon rank-sum test, with false discovery rate controlled using the method of Benjamini and Hochberg (Benjamini and Hochberg, 1995). Genes are included in the test if they pass a detection rate threshold (default is 10%) in at least one of the pair of clusters tested. In the case of both sets of tests, the results can be substituted with those of another statistical method by adding its results to the sCVdata object outlined below.

The first set of genes reported by scClustViz are those that are differentially expressed between each cluster and the rest of the data combined (referred to as DE vs Rest in the Shiny interface). This is not used to assess clustering results but may be visualized by the user to identify distinguishing genes for that cluster, although this will only be valuable if there is enough heterogeneity in the data to identify differential genes. Though this represents an unbalanced comparison, the non-parametric nature of the Wilcoxon rank-sum test makes it robust to such imbalances.

The second is referred to as marker genes. These are genes that are significantly positively differentially expressed in a cluster in pairwise comparisons with every other cluster (at a default FDR of 5%). This is taken from the results of the pairwise comparisons outlined above and returned by the function DEmarker. This method is one of the two sets of differential gene expression results used in scClustViz to quantify cluster granularity. It ensures that there are marker genes for all clusters that are unique to each cluster, given all other clusters in the data.

The other way cluster granularity is quantified by scClustViz is by comparing each cluster to its nearest neighbouring cluster. By ensuring there is at least one positively differentially expressed gene (default FDR of 5%) between each set of neighbouring clusters, this metric enforces the requirement for having statistically separable clusters, which is less restrictive than requiring unique marker genes per cluster. Nearest neighbours are clusters with the fewest differentially expressed genes between them, as calculated above. These are also taken from the results of the pairwise comparisons outlined above and returned by the function DEneighb.

To quickly compare multiple clustering solutions in the user interface, the above differential gene expression tests and other cluster-wise gene expression statistics are precomputed for each cluster solution. The results are stored as a named list containing entries for each cluster solution. The precomputed results for each cluster solution are stored as a novel S4 object class, sCVdata.

To support quick display of the various figures in the user interface, other cluster-wise gene statistics are calculated. Detection rate (DR) is the proportion of cells in a cluster in which a given gene has a non-zero expression value. Mean detected gene expression (MDGE) is the mean of the normalized transcript counts for a gene in the cells of the cluster in which that gene was detected. And mean gene expression (MGE) is the mean normalized transcript count for a gene for all cells in the cluster. These are stored as a named list of dataframes in a slot in sCVdata.

Both pairwise and one versus all differential expression test results are similarly stored in slots of sCVdata (DEvsRest and DEcombn). For the results of comparisons between a cluster and the rest of the data, each named list element contains a data frame with logGER for all genes, and p-value and FDR results for all tested genes. For pairwise comparisons between clusters, each named list element contains a data frame with logGER and dDR for all genes, and p-value and FDR results for all tested genes. List elements are named with cluster names, separated by a dash for pairwise comparisons.

The sCVdata object also stores the results of silhouette analysis, a metric for assessing the contribution of each cell to cluster cohesion and separation (Rousseeuw, 1987). This is included in the visualization as a complementary metric for cluster solution assessment. Finally, user-defined parameters pertaining to calculations on the input data are also stored as a slot in sCVdata, supporting replicability.

The package was built in R v3.5.0 (R Core Team, 2018). The R Shiny interactive web page generating tool (shiny v1.1.0) was used to generate the scClustViz user interface (Chang et al., 2018). Silhouette plots are generated using the R package cluster v2.0.7-1 (Maechler et al., 2018). Colour-split dots for plotting use code from the R package TeachingDemos v2.10 (Snow, 2016). Colour scales with transparency use the R packages scales v1.0.0, viridis v0.5.1, and RColorBrewer v1.1-2 (Neuwirth, 2014; Wickham, 2017; Garnier, 2018).

2.2.2 Operation

The scClustViz tool is available as an R package from GitHub, with usage details and example code available on the website. The typical usage requires one setup step prior to running the visualization to precompute and save the differential gene expression testing results. Once setup is complete, the user can quickly view and easily share the results of their analysis.

Setup is done using the function CalcAllSCV, which takes as input the user's scRNAseq data object and a data frame of cluster assignments where each variable refers to a different cluster solution. Currently scClustViz supports both the Bioconductor SingleCellExperiment class (Lun and Risso, 2017) and Seurat class (Satija et al., 2015; Butler et al., 2018). This function also takes optional arguments describing the state of the data and customizing testing thresholds. To calculate means of log-normalized data accurately, the function needs to know the log base and pseudocount used in the normalization. In most cases, gene expression data is transformed in log base 2, though Seurat uses the natural log. Most log-normalization methods add a pseudocount of 1 to avoid log-zero errors. As such, the function defaults to expecting log2-normalized data with a pseudocount of 1. The function also allows the user to set the gene detection rate threshold for inclusion in differential gene expression testing, defaulting to 10%.

Since this step may be time-consuming with many cluster solutions to test, the function includes an option to stop testing cluster solutions once differential gene expression between nearest neighbouring clusters has been lost. In order to do this, the function tests cluster solutions in order of increasing numbers of clusters and ensures that all nearest neighbouring cluster pairs (as determined by number of differentially expressed genes in pairwise tests) have at least one significant comparison. As such, the user may indicate the false discovery rate threshold for determining significance, which defaults to 5%.

Alternatively, the differential gene abundance testing and cluster overfitting determination can be incorporated into an existing analysis pipeline. This can be done by iteratively clustering with increasing resolution and calling CalcSCV after each clustering step. CalcSCV generates an sCVdata object for a single cluster resolution, and is called by CalcAllSCV to generate the list of sCVdata objects needed to run the Shiny interface. By checking for differential expression between nearest neighbouring clusters, this can be used to automatically stop generating cluster solutions once differential expression between clusters is lost.

The resulting list of sCVdata objects and input scRNAseq data object should be saved to disk as a single compressed RData file prior to viewing them in the GUI. This is done to ensure that setup is a one-time process, and to simplify sharing and reproducibility of analyses. The function runShiny launches the R Shiny instance with the interactive data figures in the R integrated development environment (IDE) or a web browser. It loads the data from a file and has optional arguments to specify the annotation database and marker genes for expected cell types. The annotation database is used to find gene names to improve clarity of some figures and expects a Bioconductor AnnotationDbi object such as org.Mm.eg.db for mouse or org.Hs.eg.db for human. Finally, if passed a named list of canonical marker genes for expected cell types in the data, scClustViz will automatically generate cluster annotations (labels). This is done by assigning each cluster to the cell type with the top aggregate rank of gene expression for its marker genes. More in-depth and unbiased methods for assigning cell type identities to clustering results exist (Crow et al., 2018; Kiselev et al., 2018), so this is meant more as a convenience option for labelling purposes than a definitive automatic cluster annotation method.

System requirements for this tool will depend heavily on the data set in question, since the data will have to be loaded into memory, and the memory footprint of scRNAseq data depends on the number of cells being analysed. However, in all tests loading objects from Seurat into scClustViz, the saved objects after the setup and differential expression testing steps were smaller than the original Seurat object. It is thus safe to assume that scClustViz will run on the computer on which the data set in question was analysed. For the data from the MouseCortex package, the largest data set (E15, containing nearly 3000 cells) uses less than 1.2GB of memory. Opening Shiny apps can be difficult in some computing environments, especially remote R sessions to servers without browsers or rendering capabilities. There are options in the Shiny runApp function to help troubleshoot these situations, and these are accessible from the runShiny function in scClustViz.

2.3 Use cases

To demonstrate the convenience of sharing analysed data with scClustViz, the MouseCortex package was built with data from a recent publication exploring the development of the mouse cerebral cortex using scRNAseq (Yuzwa et al., 2017). A tutorial for building similar R data packages calling scClustViz as the visualization tool can be found on the scClustViz website.

The MouseCortex package contains the four data sets published in the paper, and a wrapper function for runShiny that loads each data set with the appropriate arguments. The embryonic day 17.5 data set (opened by the command viewMouseCortex(“e17”)) will be used to demonstrate the purpose of the various figures in scClustViz and highlight its role in identifying a core gene set expressed in the neurogenic stem cell population of the cerebral cortex in the next sections. All figures from this point on were generated in the scClustViz Shiny app and saved using the “Save as PDF” buttons.

2.3.1 Clustering solution selection

The first step in the post-clustering workflow is to assess the results of the various clustering parameterizations used. scClustViz uses a combination of differential gene expression between clusters and silhouette analysis for this. Differential gene expression is used as a metric in two ways: the number of positively differentially expressed genes between a cluster and its nearest neighbour, and the number of marker genes (positively differentially expressed vs. all other clusters in pairwise tests) per cluster. In **Figure 2-2a** differential expression to the nearest neighbouring cluster is represented as a series of boxplots per cluster resolution, arranged on the x-axis to indicate the number of clusters in each boxplot. The highlighted boxplot indicates the currently selected cluster from the pulldown menu in the user interface. Both differential expression-based metrics can be visualized this way by switching the metric used, via the interface.

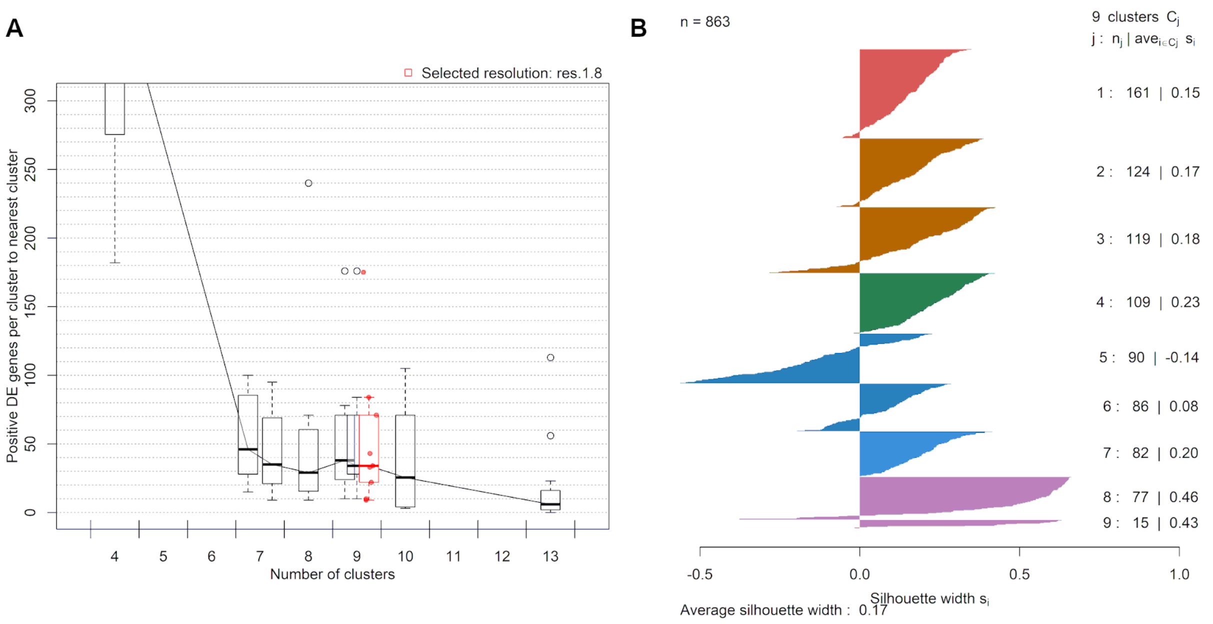


Figure 2-2. Interactive figures to assess clustering solutions.
A. Boxplots representing number of differentially expressed genes between neighbouring clusters for each cluster resolution. For each cluster at a specific resolution, the number of positively differentially expressed genes to its nearest neighbouring cluster is counted, and those counts are represented as a boxplot. The boxplots are arranged along the x-axis to reflect the number of clusters found at that resolution. Highlighted in red is the cluster resolution currently selected in the interface. This figure has been zoomed using the interactive interface to make it clear that at the selected resolution there is more than one differentially expressed gene between neighbouring clusters. The number of marker genes per cluster and average silhouette widths can be similarly viewed with the scClustViz interface. **B.** Silhouette plot for the selected cluster resolution. A horizontal bar plot where each bar is a cell, grouped by cluster. Silhouette width represents the difference between mean distance to other cells within the cluster and mean distance to cells in the neighbouring cluster. Distance is Euclidean in reduced dimensional (generally PCA) space. Positive values indicate that the cell is closer to cells within its cluster.

When a cluster resolution is selected, its silhouette plot is rendered to add another method of cluster assessment (**Figure 2-2b**). A silhouette plot is a horizontal bar plot where each bar is a cell, grouped by cluster. The width of each bar, referred to as silhouette width, represents the difference between mean distance to other cells within the cluster and mean distance to cells in the neighbouring cluster. Distance is Euclidean in the reduced dimensional space used in clustering (this is generally PCA space and is pulled from the input data object based on a user-defined parameter). Positive values indicate that the cell is closer to cells within its cluster. It is worth noting that the dimensions returned by methods such as PCA are not equally meaningful, since each explains a different proportion of the variance in the data, while Euclidean distance treats them all equally. This can be addressed by weighting the PCs by variance explained, a method implemented in newer versions of Seurat (Butler et al., 2018). To prevent unexpected results caused by assuming a PC weighting option in upstream analysis, the silhouette plot in scClustViz does not reweight PCs, so users are encouraged to consider this when interpreting this plot.

Once the user has chosen the appropriate cluster solution, they can click the "View clusters at this resolution" button to proceed to in-depth exploration and visualization of the results. They can also save this as the default resolution for future sessions. If a cluster resolution is saved as default, a file specifying the saved resolution will be generated in the same directory as the input data (or an optional output directory). Specifying a separate output directory is useful when the input data is part of a package, as in MouseCortex. If the same output directory is specified the next time the command is run, all saved data in that directory will be reloaded in the app.

2.3.2 Data set and cluster metadata inspection

In this section, the user can explore the data set as a whole. The first panel, **Figure 2-3a** shows a two-dimensional representation of cells in gene expression space. This is generally a tSNE or UMAP plot, and is pulled from the input data object based on a user-defined parameter (Maaten and Hinton, 2008; McInnes and Healy, 2018). The cells are coloured by cluster and can be labelled by cluster number or automatically annotated with a predicted cell type based on known marker genes for expected cell types passed to runShiny. The user can select any cluster for downstream exploration by clicking on a cell from that cluster in this plot. This will highlight the cluster in other plots in the interface. Since we are interested in identifying marker genes for the precursor cell population, we may click on cluster 8 (purple) to select it for downstream analysis.

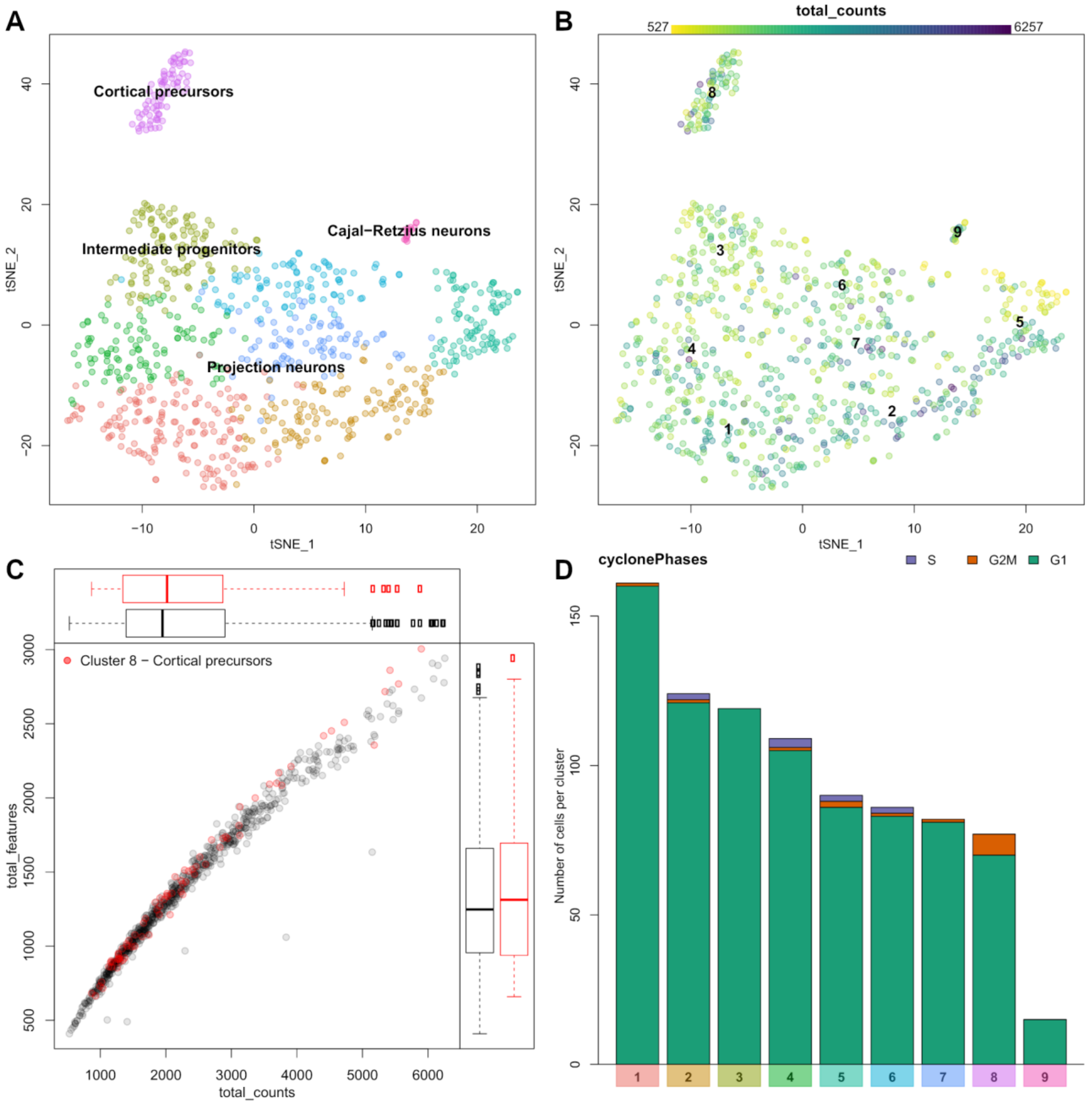


Figure 2-3. Visualizations of the data and its metadata.

A. A 2D projection of cells in gene expression space (frequently a tSNE plot) is coloured by cluster. Clusters can be labelled by number, or automatically annotated as seen here. **B.** An example of a metadata overlay on the tSNE plot. The library size (number of transcripts detected) per cell is represented by colour scale, where darker cells have larger library sizes. **C.** Metadata can be represented as a scatter plot. The relationship between number of unique genes detected (total features – y-axis) and library size (total counts – x-axis) is shown here. The cells from the selected cluster (cluster 8, cortical precursors) are highlighted in red. **D.** Categorical metadata is represented as a stacked bar plot showing the number of cells contributing to each category per cluster. This plot shows predicted cell cycle state, with G1 phase in green, G2/M in orange, and S phase in purple.

The distribution of various cellular metadata can be visualized in **Figure 2-3b**. Metadata is selected from a pulldown menu and is represented as colours on the cells in the 2D projection. In this manner the user can inspect the impact of technical artefacts such as gene detection rate, library size, or cell cycle stage on the clustering results. Numeric metadata can also be assessed as a scatter plot, where the axes can be defined by selecting from pulldown menus. **Figure 2-3c** shows the relationship between number of genes detected and library size per cell for both the data set as a whole and the selected cluster. The cells from cluster 8, a cortical precursor cluster, were selected in the previous plot and are thus highlighted in red here. The cluster 8 cells are similar to other cells in the data, thus do not seem to be biased by the measures visualized in this plot. If this was not the case, we may want to consider investigating confounding variables in the normalization process. For example, many authors have noted that gene detection rate is often strongly correlated with the first few principal components and can unduly influence clustering results (Finak et al., 2015; Risso et al., 2018a). There are a few ways to handle this, from simply excluding those principal components, or explicitly normalizing for those factors when scaling the data (as implemented in Seurat), to including the offending technical variables as covariates in more complex dimensionality reduction (i.e. ZINB-WaVE) or differential expression testing (i.e. MAST) models. While those specialized analyses are outside the scope of this tool, it is important to be able to visualize these technical factors in the analysed data to assess the efficacy of the chosen correction method.

Categorical metadata is represented as a stacked bar plot in **Figure 2-3d**, as either absolute counts or relative proportions. Here we see that by E17.5 the cortical precursors of cluster 8 are not predicted to be actively in the cell cycle using the cyclone method (Scialdone et al., 2015). This fits expectations from known developmental biology, since neurogenesis is nearly complete by this stage, and the stem cell population that persists into adulthood is thought to enter quiescence around E15.5 (Fuentealba et al., 2015). For demonstration purposes, we will continue to focus on cluster 8, which is predicted to form the adult neurogenic stem cell population in the cerebral cortex. We will aim to identify marker genes for these cells.

2.3.3 Differentially expressed genes per cluster

Once the user is satisfied that their cluster solution is appropriate and unaffected by technical factors, the next step in data interpretation is to determine the cellular identity of each cluster by its gene expression profile. The differential expression tests done prior to running the visualization assist with this by highlighting the most informative genes in the data set. In a sufficiently heterogeneous data set, differential expression between a cluster and the rest of the data can be useful for identifying genes that uniquely define a cluster's cellular identity. A more conservative form of this is the identification of marker genes – those genes that are significantly positively differentially expressed in all pairwise tests between a cluster and all other clusters. This highlights genes expected to be found at a significantly higher expression in this cluster than anywhere else in the data. Finally, there is the testing between each cluster and its nearest neighbour to highlight local differences in expression. Each of these sets of differentially expressed genes can be presented as a dot plot comparing clusters, as seen in **Figure 2-4**. A dot plot is a modified heatmap where each dot encodes both detection rate (by dot diameter) and average gene expression in detected cells (by dot colour) for a gene in a cluster. Here up to the top ten marker genes per cluster are shown, but both the type of differential expression test used to generate the gene set and the number of differentially expressed genes contributed per cluster can be adjusted using the interactive interface. At this point in the analysis, it is also possible to download any of these differential gene expression results as tab-separated value files for further analysis, by selecting the cluster of interest and differential expression type and clicking "Download gene list". This may be of value if the user is using this platform to share the data online, or with those who would prefer not to use R for further analysis. In this dot plot, we can see the top 10 marker genes for our putatively quiescent cortical precursor cell population (cluster 8) include known marker genes for cortical radial precursors (*Fabp7*, *Slc1a3*, *Ptprz1*, and *Vim*), a known marker for adult neural stem cells (*Dbi*), as well as novel marker genes for this population (*Mfge8*, *Ttyh1*, *Pea15a*, and *Ednrb*) (Yuzwa et al., 2017). The dot plot format also shows us that while *Ckb* and *Gpmgb* are significantly positively differentially expressed in cluster 8 relative to all other clusters, they are still detected in high proportions in all clusters, and thus would not be optimal marker genes.

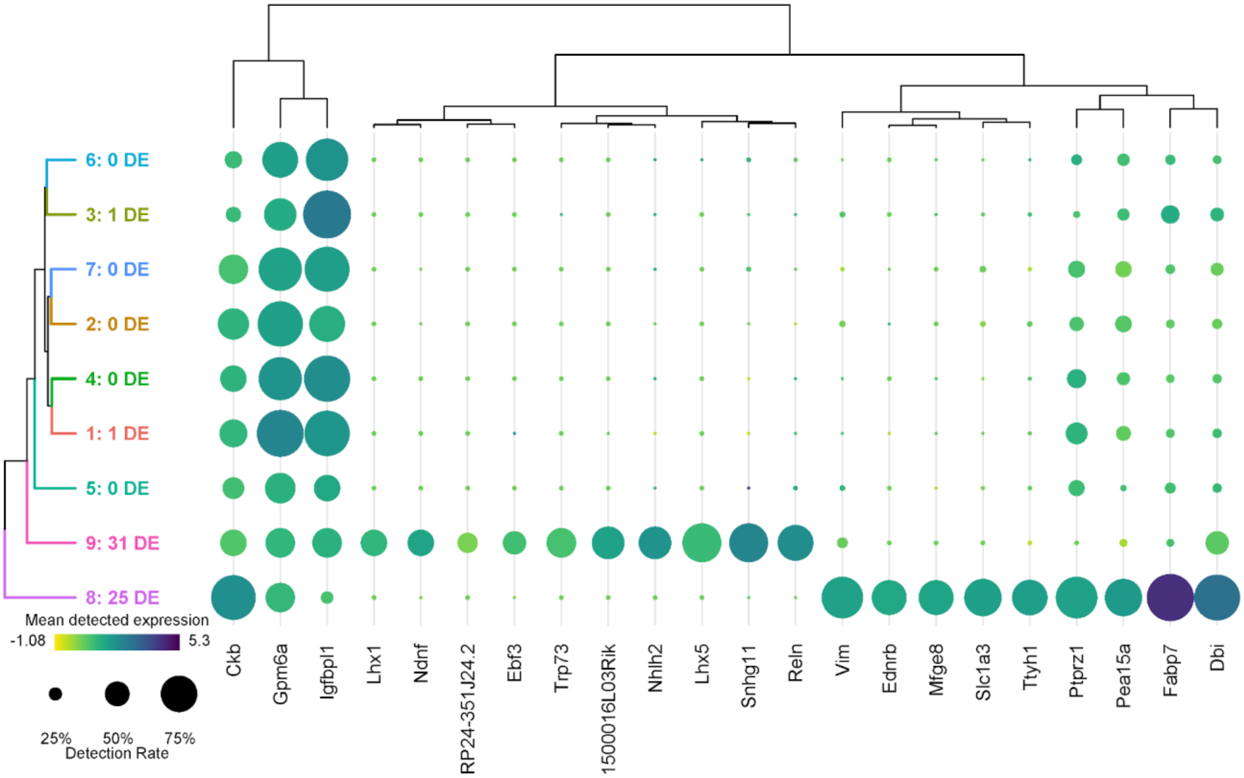


Figure 2-4. Visualizing differential gene expression.
A dot plot showing the relative expression of a subset of marker genes (x-axis) across all clusters (y-axis). A dot plot is a modified heatmap where each dot encodes both detection rate and average gene expression in detected cells for a gene in a cluster. Darker colour indicates higher average gene expression from the cells in which the gene was detected, and larger dot diameter indicates that the gene was detected in greater proportion of cells from the cluster. Cluster colours are indicated for reference on the left side of the plot. Cluster numbers are also indicated on the left side, along with the number of differentially expressed genes in each cluster. The genes included can be changed to reflect those differentially expressed per cluster when compared to the rest of the data set as a whole (i.e. the tissue), the nearest neighbouring cluster, or marker genes unique to that cluster. This figure shows marker genes per cluster. The number of differentially expressed genes contributed per cluster can also be adjusted, here set to 10.

2.3.4 Gene expression distributions per cluster

To more closely inspect the gene expression of an individual cluster, scClustViz presents gene expression data per cluster as a scatter plot with the proportion of cells from that cluster in which a gene is detected (more than zero transcript counts) on the x-axis and mean normalized transcript count from cells in which the gene was detected on the y-axis, as seen in Figure 5a. This visualization method helps separate the contribution of zeros from the mean gene expression value, since like the dot plot it separates magnitude of gene expression from gene detection rate. It also highlights the strong relationship between magnitude of gene expression and likelihood of detection in droplet-based single-cell RNAseq data, since the trend goes from the plot’s bottom left (genes have low expression and are rarely detected) to top right (genes have high expression and are detected often). In this figure, the cortical precursor cluster 8 is shown, but the user can select the cluster shown from a pulldown menu in this panel as well. There are three ways to highlight various genes in this plot. First, the genes passed as known marker genes for expected cell types can be highlighted in colours corresponding to their cell type, if a marker gene list is defined by the user (**Figure 2-5a**). This figure indicates that this cluster was classified as cortical precursors based on the high relative expression of both *Sox2* and *Pax6*, as well as *Nes* and *Cux1* (markers for both cortical precursors and projection neurons). In **Figure 2-5b**, the plot shows differentially expressed genes, specifically the genes contributed by this cluster to the dot plot shown immediately above in the app (**Figure 2-4**). Thus, by changing the differential gene set or number of genes in the heatmap, the user can also adjust the genes highlighted in this scatter plot. Finally, the user can search for genes manually by entering a list of gene symbols or using a regular expression in the search box below the figure. To identify and compare gene expression for any point in this figure, the user can click on the corresponding data point.

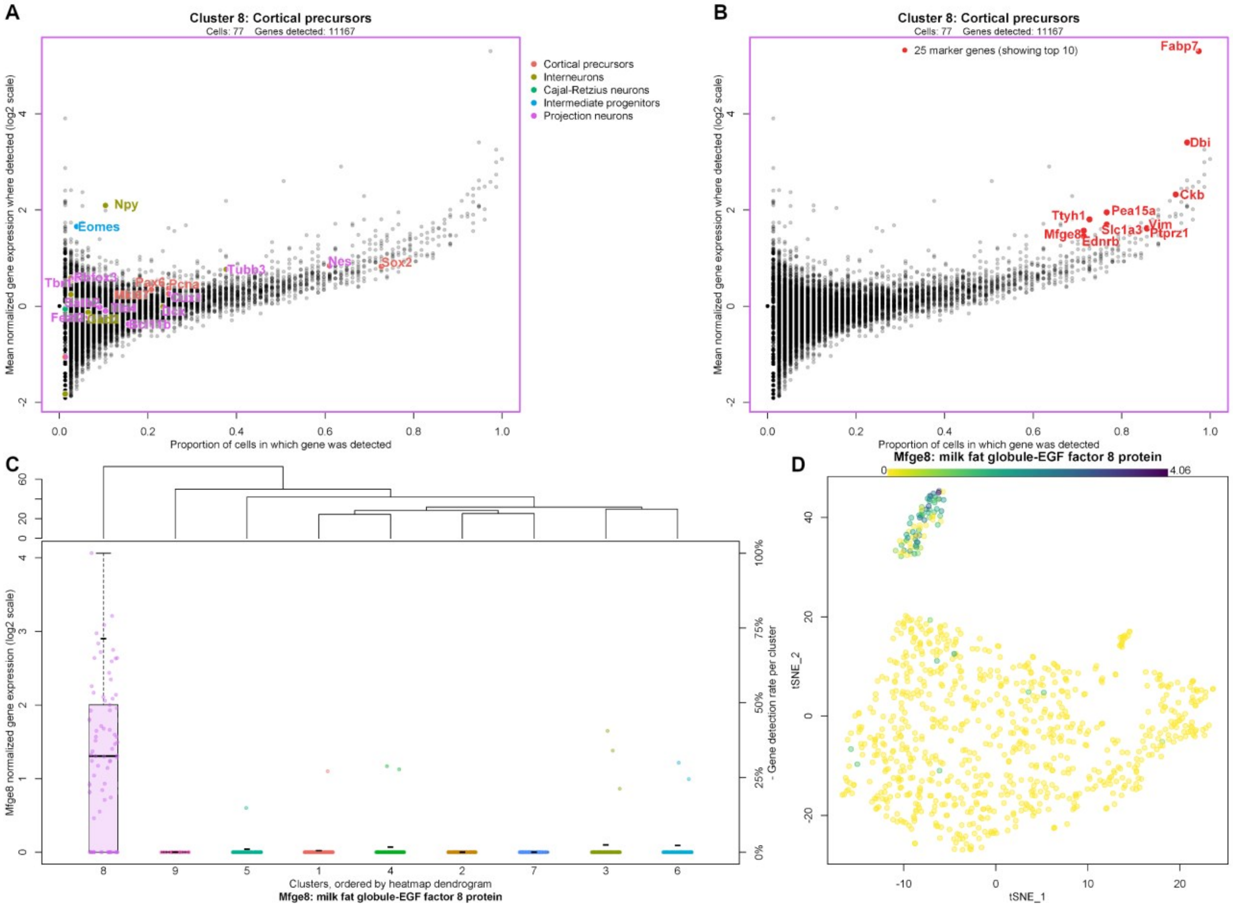


Figure 2-5. Exploring cluster-wise gene expression.

A. A scatter plot representing gene expression in the highlighted cluster, the cortical precursor cluster 8. The x-axis represents the proportion of cells from that cluster in which a gene is detected (more than zero transcript counts), and the y-axis is the mean normalized transcript count from cells in which the gene was detected. The cell type marker genes are highlighted, indicating that this cluster was classified as cortical precursors based on the high relative expression of both Sox2 and Pax6, as well as Nes and Cux1 (markers for both cortical precursors and projection neurons). **B.** The same scatter plot is shown with the top 10 marker genes for cluster 8 highlighted, though the user can choose other differentially expressed gene sets from the heatmap, or search for genes of interest using the interface. The identity of any point can be determined by clicking on it in the interface. **C.** Boxplots comparing the expression of a gene of interest across all clusters. Clusters are arranged on the x-axis based on the cluster dendrogram generated for the dot plot above (**Figure 2-4**), and normalized transcript count for the gene of interest (*Mfge8*, in this case) is represented on the y-axis. The dots on each boxplot represent the individual data points, gene expression per cell. The black dash is an optional indication of the gene detection rate per cluster, as indicated on the y-axis on the right side. This figure shows that *Mfge8* may be a marker of cortical precursors. **D.** Gene expression overlaid on the cell projection. Gene expression is represented by a colour scale on the cells of the two-dimensional projection, where darker indicates higher expression. Clusters can be optionally labelled by number or annotation. This figure shows the distribution of *Mfge8* expression in the data set.

Clicking on a data point in the figure above will generate a series of boxplots comparing gene expression for the selected gene across all clusters (**Figure 2-5c**). Since the above scatter plot can be crowded, all genes near the clicked point are shown in a pulldown menu, so that the user can select their gene of interest. Alternatively, the gene(s) entered in the search box in the previous panel can be used to populate the pulldown list for selecting the gene of interest for this figure. By comparing gene expression across clusters, it is easier to assess the utility of putative marker genes. Here we see that *Mfge8* is expressed nearly exclusively in cluster 8, with rare detection in any other clusters. This suggests that *Mfge8* may be effective for identifying the cells of this cluster *in situ*. In fact, both fluorescence in situ hybridization for *Mfge8* and immunohistochemistry for its protein lactadherin showed specificity for the cortical precursor cells in the embryonic mouse brain, as well as the B1 neural stem cells of the adult ventricular/subventricular zone (Yuzwa et al., 2017).

Finally, the user can directly plot the expression of a gene or genes of interest on the tSNE plot to better visualize the distribution of gene expression in the data set, as shown in **Figure 2-5d**. Genes are selected by entering gene symbols or a using a regular expression and selecting the matching gene symbols from a dropdown list. Gene expression is represented by a colour scale on the cells of the two-dimensional projection. If multiple genes are selected, the maximum gene expression value per cell is shown. This serves as another way of highlighting the specificity of *Mfge8* for the cortical precursor cells in this data set.

2.3.5 Cell set comparisons

The final feature of scClustViz is the ability to generate volcano and MA plots comparing gene statistics for any two clusters, or any two sets of cells specified by the user (**Figure 2-6a**). This is useful for two reasons. First, such detailed investigations of differences between clusters may help identify cell types or classify their relationships. It may also reveal systematic differences in gene expression data between two sets of cells that could indicate a technical or biological confounding factor. Volcano plots show relationships between effect size and statistical significance for sets of differential gene expression comparisons between clusters. MA plots (also known as Tukey's mean-difference plot or Bland-Altman plot) show differences between samples comparing the log-ratio of gene expression between samples to the mean gene expression across those samples. We modify the traditional MA plot by showing the mean on the y-axis and difference on the x-axis to maintain visual consistency to volcano plots. We further expand this plot's utility by giving the user the option of viewing the difference and average of all three gene statistics used in scClustViz: mean gene expression, mean detected gene expression, and detection rate. Furthermore, the user can manually select sets of cells to compare, and scClustViz will calculate differential gene expression statistics between the selected cells and the remaining cells in the data, and between sets of selected cells. Once the calculations are complete, the resulting comparison is represented as a separate "cluster solution" and can be explored in all the figures of scClustViz. These results can be saved to disk by clicking "Save this comparison to disk" when selecting it in the pulldown menu for cluster solution selection. Any saved comparisons will be loaded along with the data any time runShiny is run.

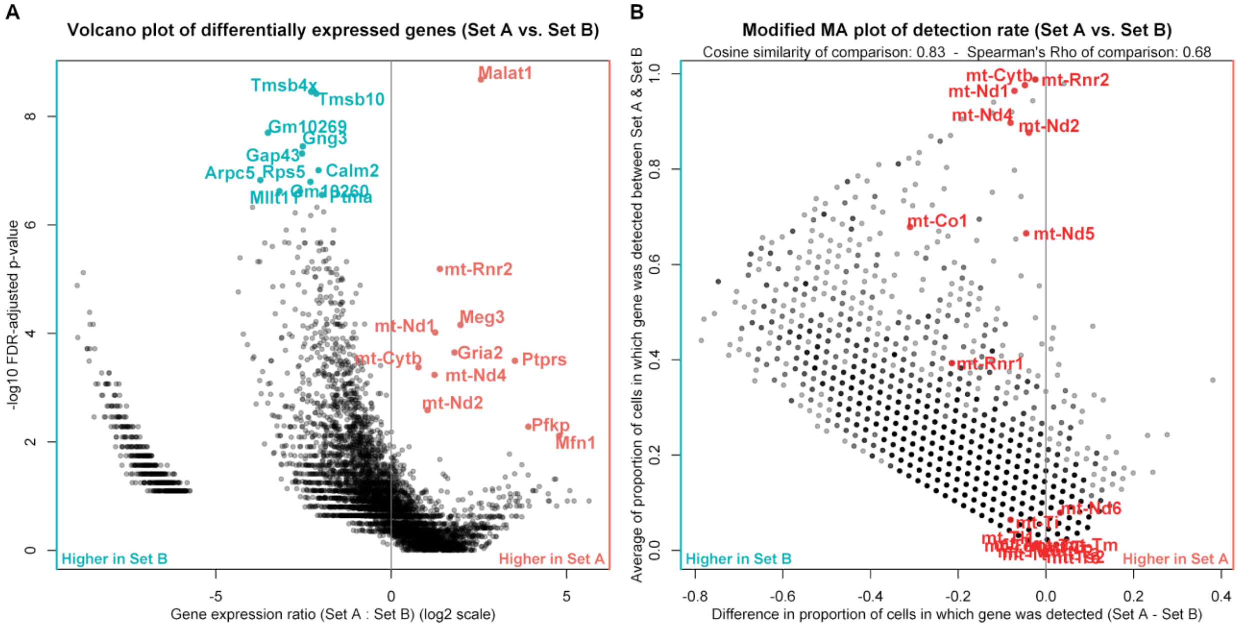


Figure 2-6. Comparing manually defined sets of cells.
A. A volcano plot showing log-ratio of gene expression between cell sets on the x-axis, and differential gene expression significance score (-log10 FDR) on the x-axis. Set A here is a subset of cluster 5 with low library sizes (< 1500 counts per cell), while set B is the subset of cluster 5 with high library sizes (> 1500 counts per cell). Highlighted are the top differentially expressed genes upregulated in set A (red) and set B (blue). **B.** An MA-style plot showing difference in gene detection rate between set A and set B on the x-axis, and average gene detection rate across sets on the y-axis. The vertical line is at zero difference in detection rate. Highlighted in red are

genes from the mitochondrial genome, which are generally used as markers of damaged cells in single-cell RNAseq analyses.

In **Figure 2-6** we're investigating a potential technical artefact in the data, specifically the poor cohesion of cluster 5 as seen in the silhouette plot in **Figure 2-2b**. This poor cohesion could be due to the differences in library size within the cells of the cluster, as seen in **Figure 2-3b**. To investigate this, the cell selection tool in scClustViz was used to select the cells of cluster 5 with low library sizes (Set A, < 1500 UMIs per cell) and those with high library sizes (Set B, > 1500 UMIs per cell). After running the differential gene expression calculations, we can view the differentially expressed genes between the sets in the dot plot or volcano plot (**Figure 2-6a**). Set B seems to have more positively differentially expressed genes, which may be due to improved gene detection rate from higher library sizes. This can be seen in **Figure 2-6b**, where an MA-style plot showing difference in detection rate vs average detection rate across sets is shown. Most genes are more detected in the set with larger library sizes (set B), which might be expected, since more transcripts detected correlates with higher average transcript counts per gene. Clicking on a gene in this figure has the same functionality as the scatter plot in **Figure 2-5**; it will be selected for viewing in the boxplot above (**Figure 2-5c**). Using this, we noticed that genes from the mitochondrial genome were seemingly unaffected by the difference in library sizes, as they tended to fall near zero difference in detection rate. To highlight this, we searched for all genes from the mitochondrial genome using the search tool, which allowed us to highlight them here. If cells are damaged and leaking cytoplasm, they are likely to have smaller library sizes as they lose mRNA. However, since RNA from the mitochondrial genome is sequestered in a separate organelle, they are less likely to lose those transcripts (Ilicic et al., 2016). We can see evidence for this in the cells of cluster 5 with small library sizes, since the detection rate of their mitochondrial genes is unchanged. While this data set was filtered to remove cells with higher than average mitochondrial gene transcript proportions, including that metric in the metadata would allow for tuning of the threshold used. Since these cells have both low library sizes and higher relative detection rate of mitochondrial transcripts, it is safe to assume they are damaged cells and remove them from the analysis.

2.4 Conclusion

We developed scClustViz to aid in the annotation of cell types and identification of marker genes from scRNAseq data. It provides both a metric for cluster assessment based on inter-cluster differential gene expression, as well as a convenient user interface for accomplishing this analysis and interpretation task. Using differential gene expression to assess clustering solutions ensures that the results are suited to addressing the relevant biological task of identifying cell types and their marker genes. The user interface is also focused specifically on this task by generating publication quality figures and providing analyses that help the user determine the appropriate number of clusters, identify cell types, and highlight genes unique to those cell types. There are other user interfaces available for the analysis of scRNAseq data (Zhu et al., 2017; Rue-Albrecht et al., 2018). However, scClustViz fills a niche between existing GUIs, which are either very user-friendly for non-technical users, at the cost of the ability to customize analysis, or very powerful and customizable, at the cost of providing a simple framework for accomplishing a common analysis task. The one-time setup step for scClustViz also simplifies data sharing, as it generates a file that can be shared for viewing by anyone using R. Data sharing can be made more user-friendly by building an R data package with a wrapper function calling scClustViz, as seen in the use case outlined in this paper. Building such a package is a quick process, and a tutorial is available on the scClustViz website. scClustViz is available at https://baderlab.github.io/scClustViz/ as free, open source software under the permissive MIT open source license.

2.5 Software and data availability

scClustViz is available from: https://baderlab.github.io/scClustViz/
Source code is available from GitHub: https://github.com/BaderLab/scClustViz
Archived source code at time of publication: https://doi.org/10.5281/zenodo.2582090
Licence: MIT

The example data set used is available as an R package:
https://github.com/BaderLab/MouseCortex
Archived code at time of publication: https://doi.org/10.5281/zenodo.2582093
Licence: MIT

Chapter 3
Transcriptional signatures of cell-cell interactions are dependent on cellular context

This work is an expanded version of the preprint available on *bioRxiv* (Innes and Bader, 2021). I conceived of, designed, and conducted this project. Gary D. Bader supervised and advised on this project.

3 Transcriptional signatures of cell-cell interactions are dependent on cellular context

Cell-cell interactions are often predicted from single-cell transcriptomics data based on observing receptor and corresponding ligand transcripts in cells. These predictions could theoretically be improved by inspecting the transcriptome of the receptor cell for evidence of gene expression changes in response to the ligand. It is commonly expected that a given receptor, in response to ligand activation, will have a characteristic downstream gene expression signature. However, this assumption has not been well tested. We used ligand perturbation data from both the high-throughput Connectivity Map resource and published transcriptomic assays of cell lines and purified cell populations to determine whether ligand signals have unique and generalizable transcriptional signatures across biological conditions. Most of the receptors we analyzed did not have such characteristic gene expression signatures – instead these signatures were highly dependent on cell type. Cell context is thus important when considering transcriptomic evidence of ligand signaling, which makes it challenging to build generalizable ligand-receptor interaction signatures to improve cell-cell interaction predictions.

3.1 Introduction

The advent of single-cell RNA sequencing has enabled researchers to explore the cell composition of complex tissues and better understand how cells work together in the context of the entire tissue. Cell-cell interaction prediction from scRNAseq data is useful to support this 'cellular ecosystem' study (Shao et al., 2020; Armingol et al., 2021; Dimitrov et al., 2022). The fundamental idea behind cell-cell interaction inference is that if one cell type expresses a ligand, and another cell type expresses its cognate receptor, these cell types may be communicating via paracrine signaling through this putative ligand-receptor interaction. Given an adequate database of ligand-receptor interacting pairs, one can quickly generate lists of candidate interactions between cell types, represented by scRNAseq-derived cell clusters. The disadvantage is that by ignoring myriad factors not directly assayed by scRNAseq that affect ligand-receptor signaling (e.g. post-translational regulation, protein expression level, cellular localization, small molecule cofactors, receptor dimerization), these predictions lack specificity. Most existing cell-cell interaction inference tools attempt to highlight interactions of interest in various ways but do not attempt to improve the accuracy of their predictions beyond simply identifying them. However, scRNAseq data contains information about a cell type's entire transcriptome, not only its expressed ligands and receptors. Next-generation cell-cell interaction inference methods take advantage of this information, using the receptor cell's transcriptome as evidence of predicted ligand-receptor interactions (**Table 3-1**). Evidence of a ligand-receptor interaction may exist as a transcriptional change in the receiving cell's transcriptome mediated by signal transduction and gene regulatory events downstream of the receptor. These methods infer the expected transcriptional signature of a ligand-receptor interaction from multilayer network models constructed from existing signal transduction and gene regulatory network databases. We set out to complement these models by determining ligand-response signatures from experimental data available in pharmacogenomic resources. However, this experimental data suggested that the transcriptional response to perturbation by an individual ligand is not consistent between cell contexts. As most existing multilayer network models do not account for cell type, this work aims to systematically test the assumption that transcriptional response to ligand perturbation is consistent between cell types.

Table 3-1. Cell-cell interaction inference methods using the receptor-expressing cell's transcriptome to provide evidence of ligand-receptor interaction.

SoptSC (Wang et al., 2019a)	Per-cell ligand-interactions scores weighted by product of expression abundance for ligand, receptor, and expected up-/down-regulated receptor pathway targets. Pathway targets manually curated for a handful of signaling pathways.
NicheNet (Browaeys et al., 2020)	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks used to predict expected ligand-responsive genes. Model edge weights trained on a collection of ligand-perturbation gene expression data. Pathways curated from many databases, accessed through Harmonizome (Rouillard et al., 2016)
scMLnet (Cheng et al., 2021)	Multilayer model of ligand-receptor, receptor-TF, and gene regulatory networks. L-R and R-TF graphs pruned of weakly expressed nodes. GRN refined by requiring significant positive correlations between TF and target gene expression.
Domino (Cherry et al., 2021)	Gene regulatory networks inferred by SCENIC (Aibar et al., 2017). Receptor-TF correlations not predicted by GRN are predicted activated receptors. Expression of cognate ligands used to identify sender cells.
CellChat (Hao et al., 2021)	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks. Pathways from OmniPath (Türei et al., 2021). Cell type specific GRNs inferred by coexpression using GRN from DoRothEA (Garcia-Alonso et al., 2019) as a prior.
CytoTalk (Hu et al., 2021)	Multilayer model of ligand-receptor and intracellular gene interaction networks. Prize-collecting Steiner forest algorithm identifies subnetwork with highest L-R correlation across cell types and intracellular cell type specificity.

FunRes (Jung et al., 2021)	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks. Ligand-receptor interaction inferred by Markov chain from transcription factor expression.
CellCall (Zhang et al., 2021a)	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks. Scoring combines expression of L-R with expression of downstream regulon linked by KEGG and TF databases using GSEA (Subramanian et al., 2005).

To determine the transcriptional signature of a ligand-activated receptor interaction, we used two complementary sources of data measuring transcriptional change in response to ligand perturbation. Connectivity Map is a database of over 1 million gene expression profiles in various cell lines in response to a variety of perturbands, including some secreted protein ligands (Subramanian et al., 2017). To efficiently measure this very large number of gene expression profiles, it uses a ligation-mediated amplification technique to measure the expression of 978 selected genes (referred to as “landmark” genes), from which the expression of a further 11,350 genes is inferred. To complement the breadth of this high-throughput resource, we also used published transcriptomics assays testing individual ligands perturbing a single cell type, originally curated by the authors of NicheNet (Browaeys et al., 2020). These individual microarray datasets were used to corroborate the conclusions drawn from the Connectivity Map data.

Transcriptional responses to individual ligands vary between cell types. Correlation of gene expression change upon treatment with a ligand was significantly worse comparing between cell types than within each cell type. Rarely were the genes whose expression was most responsive to ligand treatment in one cell type common across multiple cell types. This is not due to technical factors, as independent assays (biological replicates) within the same cell type often identify the same responsive genes. Finally, a non-linear machine learning method was not able to extrapolate predictions of ligand-treated versus untreated samples to those from a novel cell line. While different cell types respond differently to the same ligand, transcriptionally similar cell types respond similarly, suggesting that the transcriptome contains information that will enable cell type specific inference of ligand-response signatures. If these ligand-response signatures are to improve the specificity of cell-cell interaction predictions from transcriptomic data, we must consider cell type differences.

3.2 Results

3.2.1 Ligand-driven differential gene expression is not consistent across cell lines

To test if we could identify a characteristic gene expression signature downstream of a ligand-activated receptor, we used the Connectivity Map database of ligand-perturbed cell line gene expression profiles. Connectivity Map used the L1000 assay to measure the transcriptome of cell lines perturbed with individual ligands at given sets of concentrations and durations of treatment, reporting Z-scores for gene abundance comparisons to plate-matched controls (Subramanian et al., 2017). While the Connectivity Map uses the L1000 assay's approximately 1000 (actually 978) measured gene abundances to infer gene expression values for most of the wider transcriptome, we used only the 978 measured Z-scores unless otherwise indicated. Connectivity Map's 2017 release includes assays of 295 peptide ligands across 9 cell lines immortalized from a variety of tissues (**Figure 3-1**). In these cell lines, ligands were often assayed in up to three biological replicates at a single concentration for a single duration of exposure. These data were used to assess the effect of cell context on the transcriptomic signature of ligand perturbation, unless otherwise indicated.

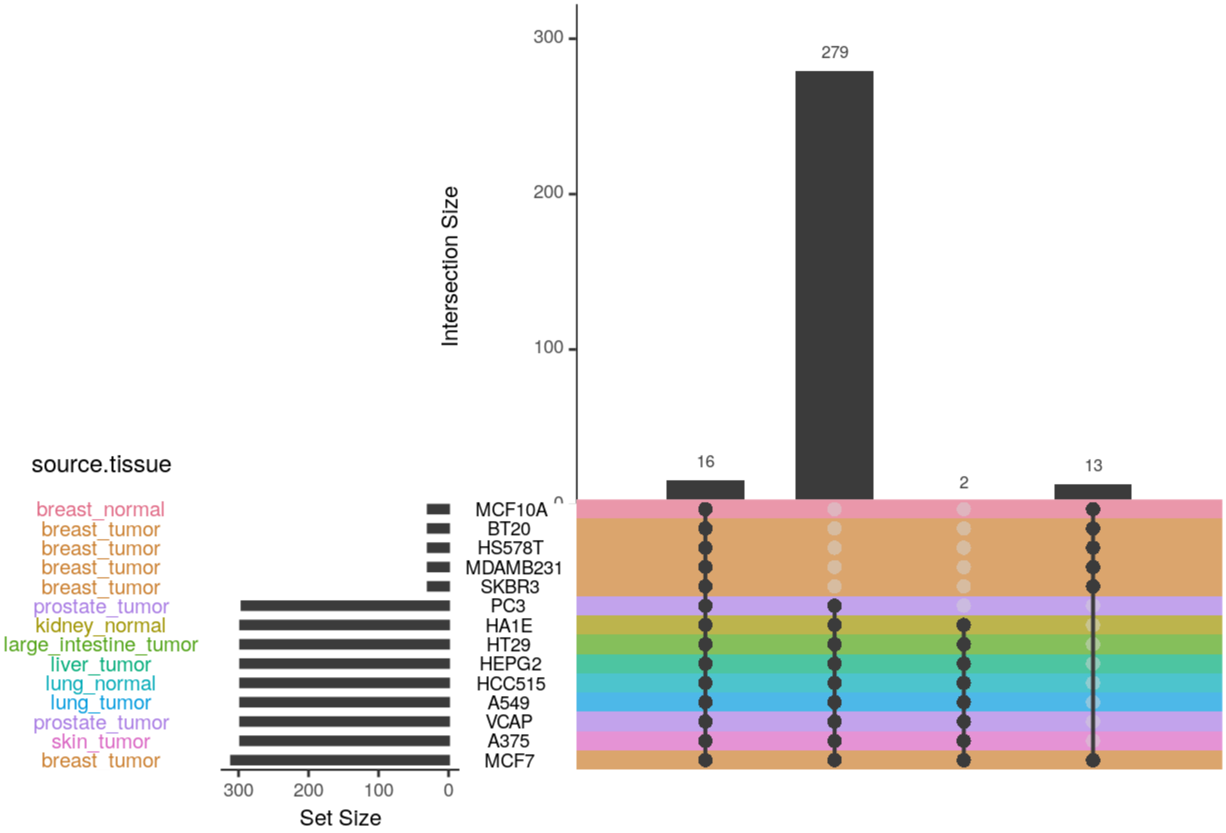


Figure 3-1. Summary of Connectivity Map ligand and cell line coverage.
An UpSet plot (Lex et al., 2014) showing the distribution of ligands assayed in each of the 14 cell lines. 16 ligands were assayed in all 14 cell lines, and 295 ligands were assayed in 9 of the 14 cell lines, covering all seven tissue sites represented by the 14 cell lines.

We first aimed to identify genes whose expression changed in response to ligand treatment consistently across all tested cell lines. To do this, differentially expressed genes were determined for each ligand perturbation by averaging Z-scores across all samples treated with the same ligand (**Figure 3-2**). Statistical significance was determined by converting the mean Z-scores to p-values using the Gaussian distribution and calculating FDR (Benjamini and Hochberg, 1995). This resulted in the unexpected finding that most ligands did not affect gene expression in a sufficiently consistent manner across samples from multiple cell lines to result in statistically significant differential gene expression levels. Only 59 of the 295 assayed ligands caused significant changes in expression of any assayed genes at FDR ≤ 0.05, and only 19 ligands caused differential expression of more genes than expected by chance (p ≤ 0.05 by permutation testing, **Figure 3-3**). If one were to use these differentially expressed genes to build signatures of ligand-mediated transcriptional response, another issue becomes apparent - that relatively few genes are differentially expressed. Of the 978 genes whose expression was measured in these L1000 assays, 58 were significantly differentially expressed (DE) in response to any ligand, and only 32 of these were unique to a single ligand. These unique genes marked only 7 of the 295 assayed ligands, with Interleukin 19 (IL19) being notable for causing significant transcriptional changes in 23 genes not significantly affected by other ligands. Overall, it appears that most ligands in this dataset lack genes that are uniquely and significantly DE upon ligand treatment.

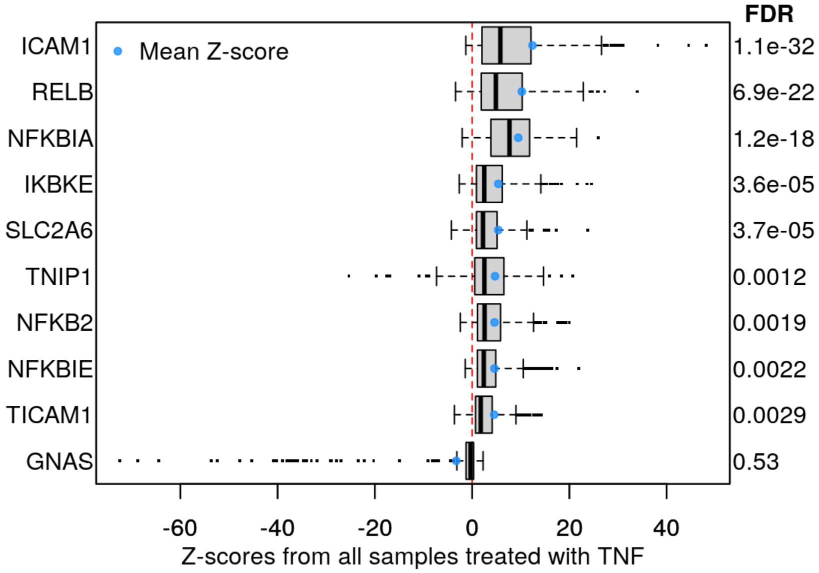


Figure 3-2. Average changes in gene expression upon TNF ligand treatment (Z-scores) across all samples in Connectivity Map.
Shown here are the genes with the highest magnitude mean Z-scores (blue dot) for samples treated with TNF. To determine which gene's expression level was consistently perturbed across samples treated with the same ligand, Z-scores representing normalized change in gene expression for each treated sample were averaged. P-values for each change in gene expression were calculated from the mean Z-score then corrected for multiple testing by the Benjamini-Hochberg procedure to control False Discovery Rate.

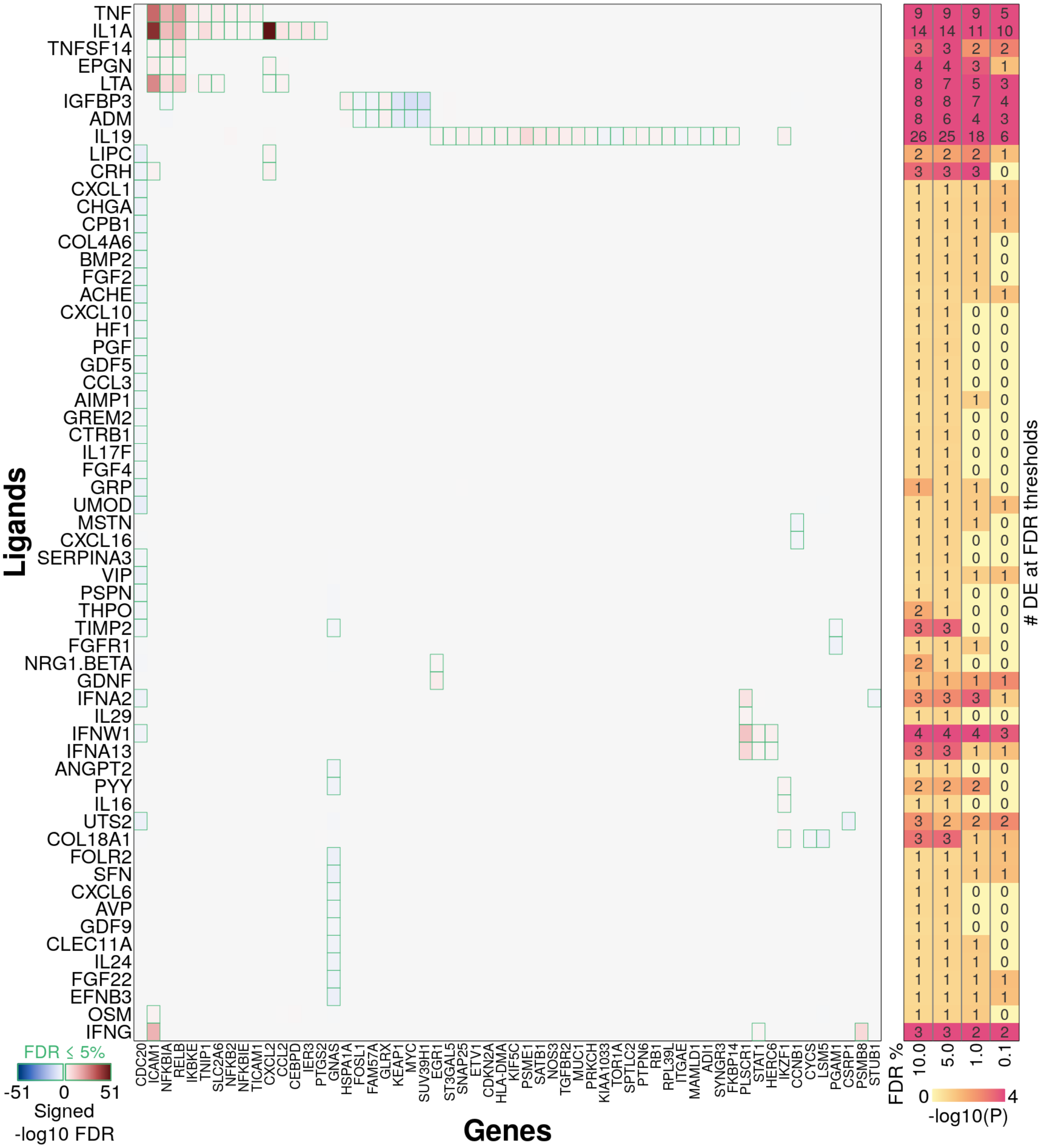


Figure 3-3. Few ligands drive consistent changes in gene expression across cell lines. A heat map showing average change in gene expression caused by ligand treatment averaged across all samples treated with the same ligand. Colours represent -log10 signed, false discovery rate corrected p-values derived from averaging Z-scores of gene expression change across all samples treated with the same ligand. Darker red represents a significant increase in gene expression, while darker blue represents a significant decrease in expression. All 59 (of 295) ligands where at least one gene was significantly different at an FDR threshold of 5% are shown. Green outline on a heat map cell represents a corrected p-value ≤ 0.05. Significantly

differentially expressed genes unique to specific ligands could be used as transcriptional markers of ligand activity. However, the presence of vertical columns representing genes responsive to multiple ligands, and the overall sparsity of this heat map suggest that very few ligands possess such marker genes. On the right is a second heat map showing the probability of detecting the indicated number of differentially expressed genes by chance. The values indicate the number of significantly differentially expressed genes identified per ligand at the given FDR thresholds. Darker pink colour represents decreasing probability of seeing that number of differentially expressed genes by chance as calculated against a background distribution of differentially expressed gene counts determined by permuting sample labels.

3.2.2 The lack of common ligand-responsive genes across cell contexts in the Connectivity Map dataset is not due to unexpected technical variability

To examine whether the Z-scoring procedure may have affected our result, we analyzed quantile-normalized gene abundance measurements (from which ligand-treatment Z-scores were calculated). These gene abundance measurements showed strong pairwise correlations between replicate experiments (mean Spearman correlation coefficient (SCC) of 0.88, **Figure 3-4a**), as did those from samples from the same cell line, whether or not they were treated with the same ligand (mean SCC of 0.85). Pairwise correlations between samples treated by the same ligand were notably less well correlated (mean SCC of 0.72) unless considering only samples from the same cell line (mean SCC of 0.86). Correlation between samples based on Z-scores was far worse (mean SCC of 0.08 between replicates, **Figure 3-4b**), likely due to the relatively small proportion of the transcriptome responding to ligand treatment, as seen in the differential gene expression analysis above. Nevertheless, we observe the same overall pattern in analyses with both normalized gene abundance measurements and Z-scores, that pairwise correlation is improved when limiting the analysis to specific cellular contexts. Specifically, 213 of 295 ligands (72%) had significantly better ($p \leq 0.05$ by Wilcoxon rank-sum test) mean pairwise correlations within at least one cell line versus across all lines, and 281 of 295 (95%) had significantly better correlations in at least one replicate set (same cell line, dosage, and duration of treatment) (**Figure 3-4c**). Furthermore, there was a significant difference in the pairwise correlation of ligand-mediated transcriptional responses, as measured by Z-scores, across all samples treated with the same ligand, and those specifically from the same cell line, or same replicate set (Wilcoxon signed-rank $p < 2.2e-16$, **Figure 3-4d**). Thus, considering cell context improves the correlation between changes in transcriptomes responding to the same ligand and this is not due solely to technical factors.

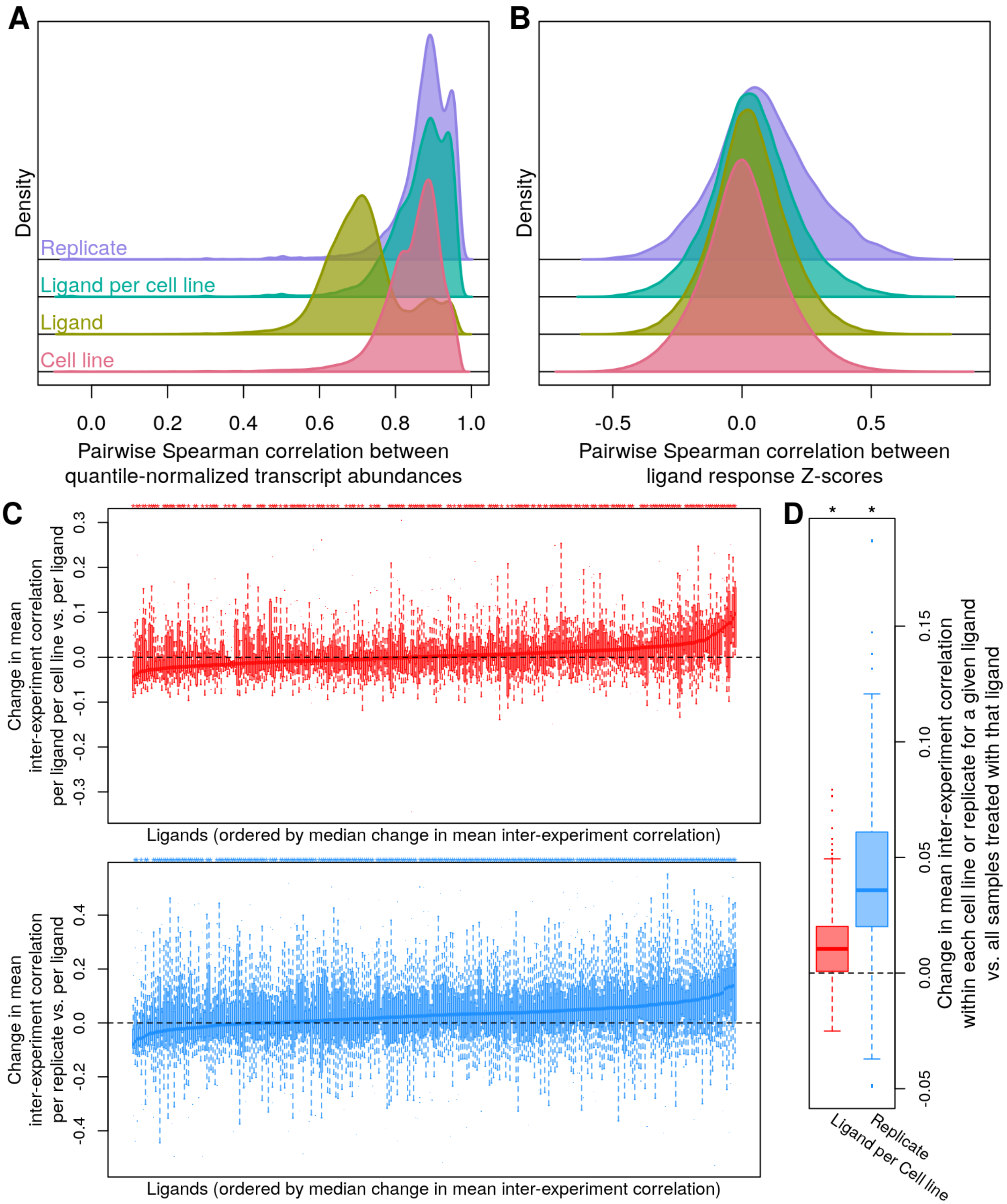


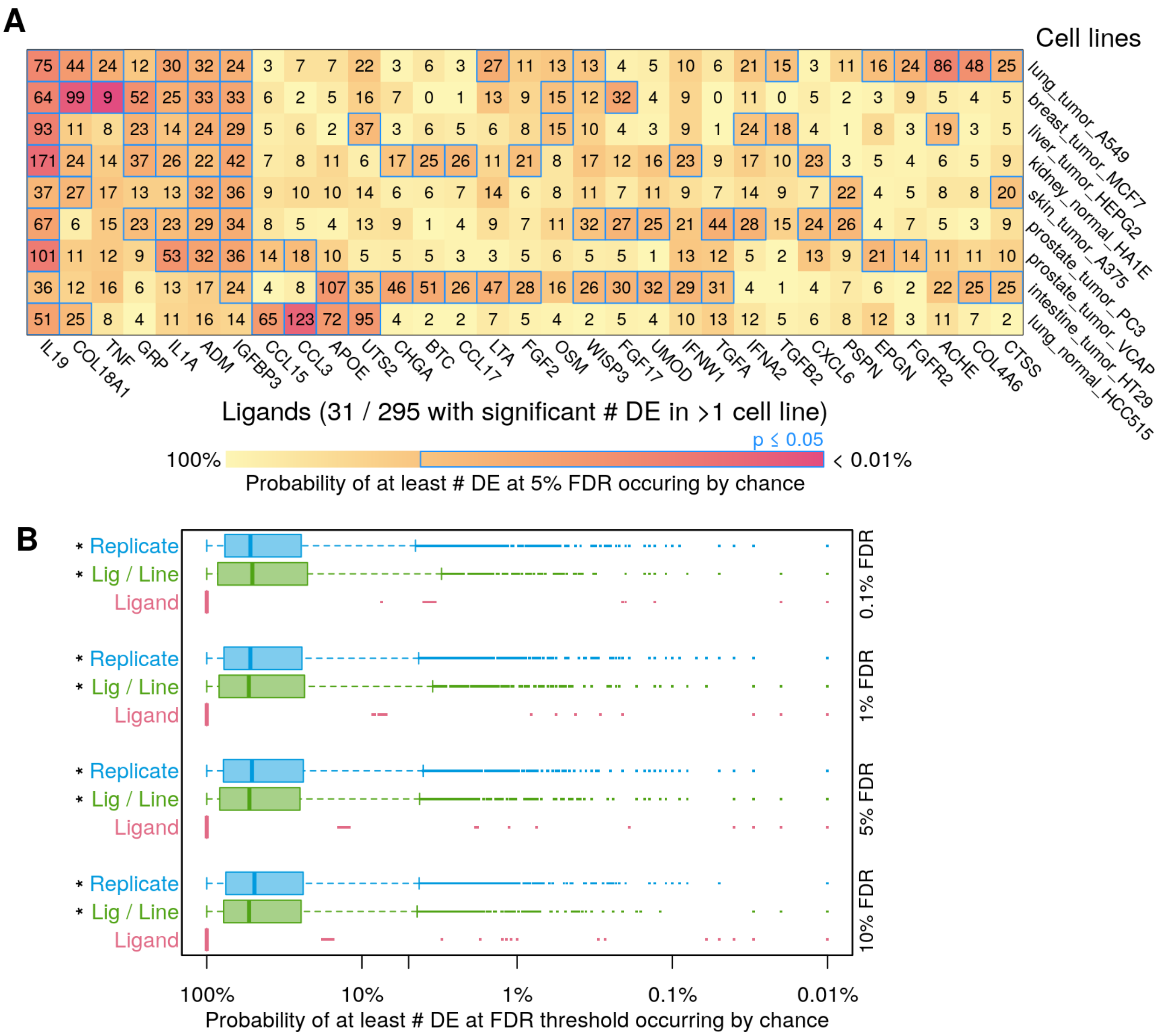
Figure 3-4. Correlation of change in gene expression upon ligand perturbation is decreased when not accounting for cell context, despite good correlation of gene expression between replicates.

A. Distribution of pairwise Spearman correlation coefficients between quantile-normalized gene expression magnitude values for all sample pairs from the same cell line (red), treated with the same ligand (yellow-green), treated with the same ligand in the same cell line (teal), or from the same replicate set (same ligand, dosage and duration of treatment, in the same cell line; purple). **B.** Distribution of pairwise Spearman correlation coefficients between Z-scores representing change in gene expression upon ligand treatment, for the same sets of samples as panel A. **C.** Boxplots showing the change in mean pairwise Spearman correlation coefficients between Z-scores for each ligand when considering only correlations within the same cell line (top, red) or within the same replicate set (same cell line, dosage and duration of treatment; bottom, blue). Each box represents a ligand, and the y-axis values represent the difference between the mean of all pairwise Spearman correlation coefficients between all samples treated with that ligand; subtracted from the means of all pairwise correlations between samples treated with that ligand in each cell line (red) or in each replicate set (blue). Ligands are ordered on the x-axis by the median value of these differences in means. The small * above the boxplots indicates that at least one cell line (red; 213 / 295 (72%) of ligands) or replicate (blue; 281 / 295 (95%) of ligands) had a significant improvement in mean inter-experiment Z-score correlation compared to the mean of all inter-experiment correlations for that ligand ($p \leq 0.05$ by Wilcoxon rank-sum test). **D.** Boxplots summarizing the results in panel C. Each boxplot contains a data point for each of the 295 ligands. Each point is the difference between the mean of all inter-experiment correlations for a given ligand, subtracted from the mean of all inter-experiment correlations from each cell line (red) or each replicate set (blue). Correlations per ligand within each cell line, and per replicate are both significantly (* = $p < 2.2e\text{-}16$ by Wilcoxon signed-rank test) greater than across all samples treated with the same ligand.

3.2.3 Differential gene expression in response to treatment with an individual ligand is consistent within the same cell type context.

After finding that correlation between ligand-treated transcriptomes improved when considering cell context, we used our differential gene expression analysis framework (**Figure 3-3**) to ask whether ligand mediated DE genes were consistent within the same cell context. We defined cell context by averaging gene expression Z-scores across ligand-treated samples from the same cell line across all experiments. In this analysis, in contrast to the analysis across all cell lines, all ligands cause statistically significant (FDR-corrected p-value ≤ 0.05) changes in the expression of at least one gene in at least one of the nine tested cell lines. This is consistent with the published Connectivity Map signatures (available at https://clue.io/), which are broken down by cell line. Nearly half (123 / 295) of the tested ligands caused differential expression (FDR-corrected p-value ≤ 0.05) of more genes than expected by chance in at least one cell line (p ≤

0.05 by permutation testing), and 31 ligands caused a significant number of DE genes in more than one cell line (**Figure 3-5a**). This is many more than the 19 ligands with more DE genes than expected by chance observed above (**Figure 3-3**). To quantify this context-dependent improvement in consistency of transcriptional response to ligand perturbation, we compared distributions of the probability of detecting as many DE genes by chance for each ligand when averaged across all samples versus the same, but for samples from the same cell line or biological replicate (**Figure 3-5b**). This comparison was made instead of comparing numbers of DE genes because when considering cell type context, we are averaging across fewer samples, and thus would expect higher magnitude average Z-scores by chance. We avoid this bias by permuting sample labels to generate null distributions of Z-scores averaged across the appropriate number of samples and calculating a p-value for the number of DE genes detected. When considering cell line context (i.e. averaging within the same cell line) these p-values improved for 274 (93%) of the 295 ligands (**Figure 3-6**), and overall it was significantly less likely that the number of DE genes per ligand was detected by chance ($p < 2.2e-16$ by Wilcoxon rank-sum test). This finding was robust to the FDR threshold used to determine the number of DE genes. Similar results were obtained when including the duration and dosage of ligand treatment along with cell line context - i.e. averaging Z-scores within replicates ($p < 2.2e-16$ by Wilcoxon rank-sum test compared to per ligand, no change compared to per ligand in the same cell line, **Figure 3-5b**). Thus, for most ligands tested in Connectivity Map, gene expression changes are dependent on cell context.



Cell lines	IL19	COL18A1	TNF	GRP	IL1A	ADM	IGFBP3	CCL15	CCL3	APOE	UTS2	CHGA	BTC	CCL17	LTA	FGF2	OSM	WISP3	FGF17	UMOD	IFNW1	TGFA	IFNA2	TGFB2	CXCL6	PSPN	EPGN	FGFR2	ACHE	COL4A6	CTSS
lung_tumor_A549	75	44	24	12	30	32	24	3	7	7	22	3	6	3	27	11	13	13	4	5	10	6	21	15	3	11	16	24	86	48	25
breast_tumor_MCF7	64	99	9	52	25	33	33	6	2	5	16	7	0	1	13	9	15	12	32	4	9	0	11	0	5	2	3	9	5	4	5
liver_tumor_HEPG2	93	11	8	23	14	24	29	5	6	2	37	3	6	5	6	8	15	10	4	3	9	1	24	18	4	1	8	3	19	3	5
kidney_normal_HA1E	171	24	14	37	26	22	42	7	8	11	6	17	25	26	11	21	8	17	12	16	23	9	17	10	23	3	5	4	6	5	9
skin_tumor_A375	37	27	17	13	13	32	36	9	10	10	14	6	6	7	14	6	8	11	7	11	9	7	14	9	7	22	4	5	8	8	20
prostate_tumor_PC3	67	6	15	23	23	29	34	8	5	4	13	9	1	4	9	7	11	32	27	25	21	44	28	15	24	26	4	7	5	3	9
prostate_tumor_VCAP	101	11	12	9	53	32	36	14	18	10	5	5	5	3	3	5	6	5	5	1	13	12	5	2	13	9	21	14	11	11	10
intestine_tumor_HT29	36	12	16	6	13	17	24	4	8	107	35	46	51	26	47	28	16	26	30	32	29	31	4	1	4	7	6	2	22	25	25
lung_normal_HCC515	51	25	8	4	11	16	14	65	123	72	95	4	2	2	7	5	4	2	5	4	10	13	12	5	6	8	12	3	11	7	2

A
Ligands (31 / 295 with significant # DE in >1 cell line)
p ≤ 0.05
100%
< 0.01%
Probability of at least # DE at 5% FDR occuring by chance
B
* Replicate
* Lig / Line
Ligand
0.1% FDR
1% FDR
5% FDR
10% FDR
100%
10%
1%
0.1%
0.01%
Probability of at least # DE at FDR threshold occurring by chance

Figure 3-5. Many ligands drive consistent changes in gene expression in some individual cell lines.

A. A heat map showing significance scores for the number of differentially expressed genes (at FDR of 5%) caused by a ligand's perturbation of each cell line. The count of differentially expressed genes in response to each ligand (column) in each cell line (row) is shown. Statistical significance (indicated by darker colour) is calculated by determining the probability of seeing that number of differentially expressed genes by chance against a background distribution of differentially expressed gene counts determined by permuting sample labels. Ligands are included in the heat map if they cause differential expression of a significant number of genes in more than one cell line ($p \leq 0.05$ by permutation test). **B.** Boxplots showing the distribution of significance scores (as calculated in panel A) when averaging gene Z-scores across all samples treated with the same ligand (as in **Figure 3-3**), all samples treated with the same ligand in the same cell line ("Lig / Line", from panel A), and all replicate samples (same ligand, cell line, dosage, and duration of treatment). Significance is represented as the probability of seeing as many differentially expressed genes by chance against a background distribution of differentially expressed gene counts determined by permuting sample labels. Numbers of differentially expressed genes were calculated at the FDR thresholds indicated on the right vertical axis, and * indicates that the center of that distribution is significantly different (p < 2.2e-16 by Wilcoxon rank-sum test) than when considering all samples treated with the same ligand.

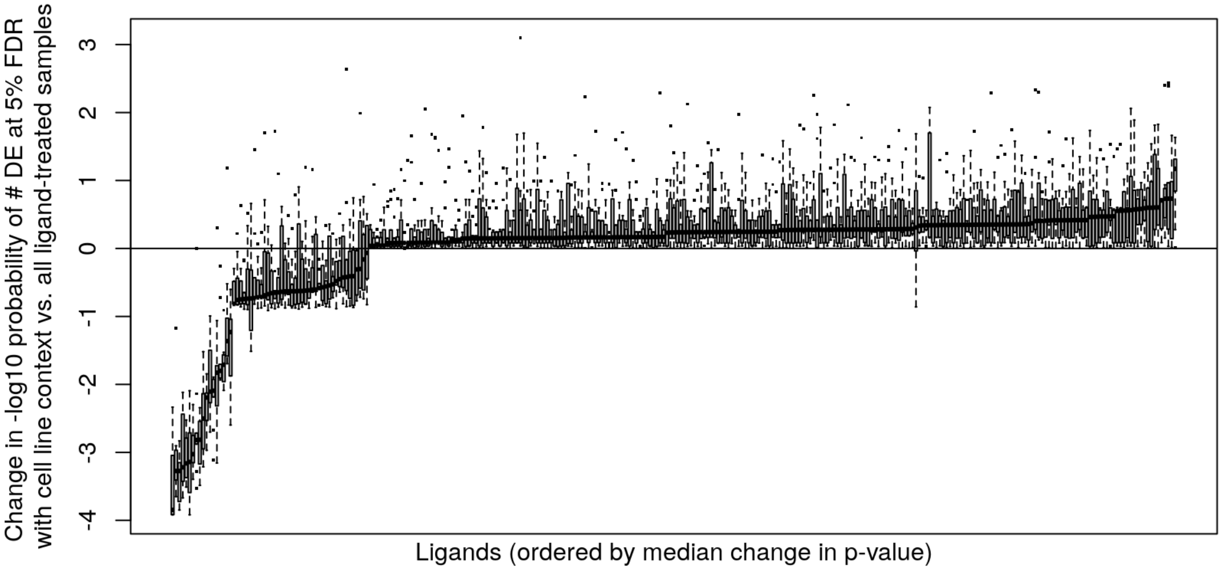
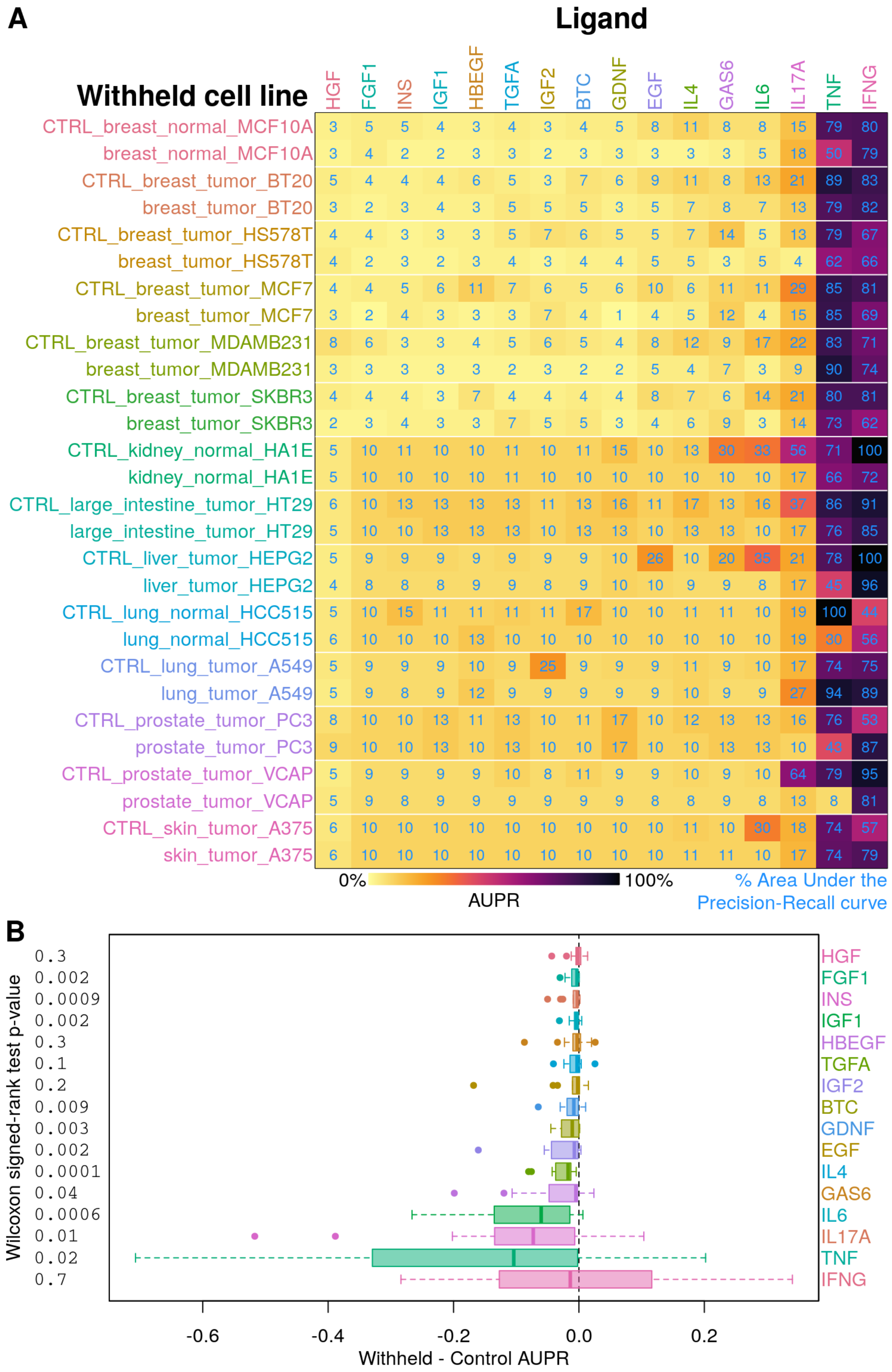


Figure 3-6. Significantly more differentially expressed genes are detected when considering cell line context for most ligands.

Each boxplot represents the change in significance score between averaging across all Z-scores from samples treated with a given ligand, subtracted from the averages of Z-scores from all samples treated with a given ligand in each cell line. Significance score is the negative log10 transformation of the probability of seeing that number of differentially expressed genes by chance, calculated using a background distribution of differentially expressed gene counts determined by permuting sample labels. Values above zero indicate that significantly more differentially expressed genes were caused by the given ligand in that cell line than that ligand causes to be differentially expressed generally across all cell lines.

3.2.4 Modeling non-linear response to ligand stimulus does not improve the transcriptional signature generalizability

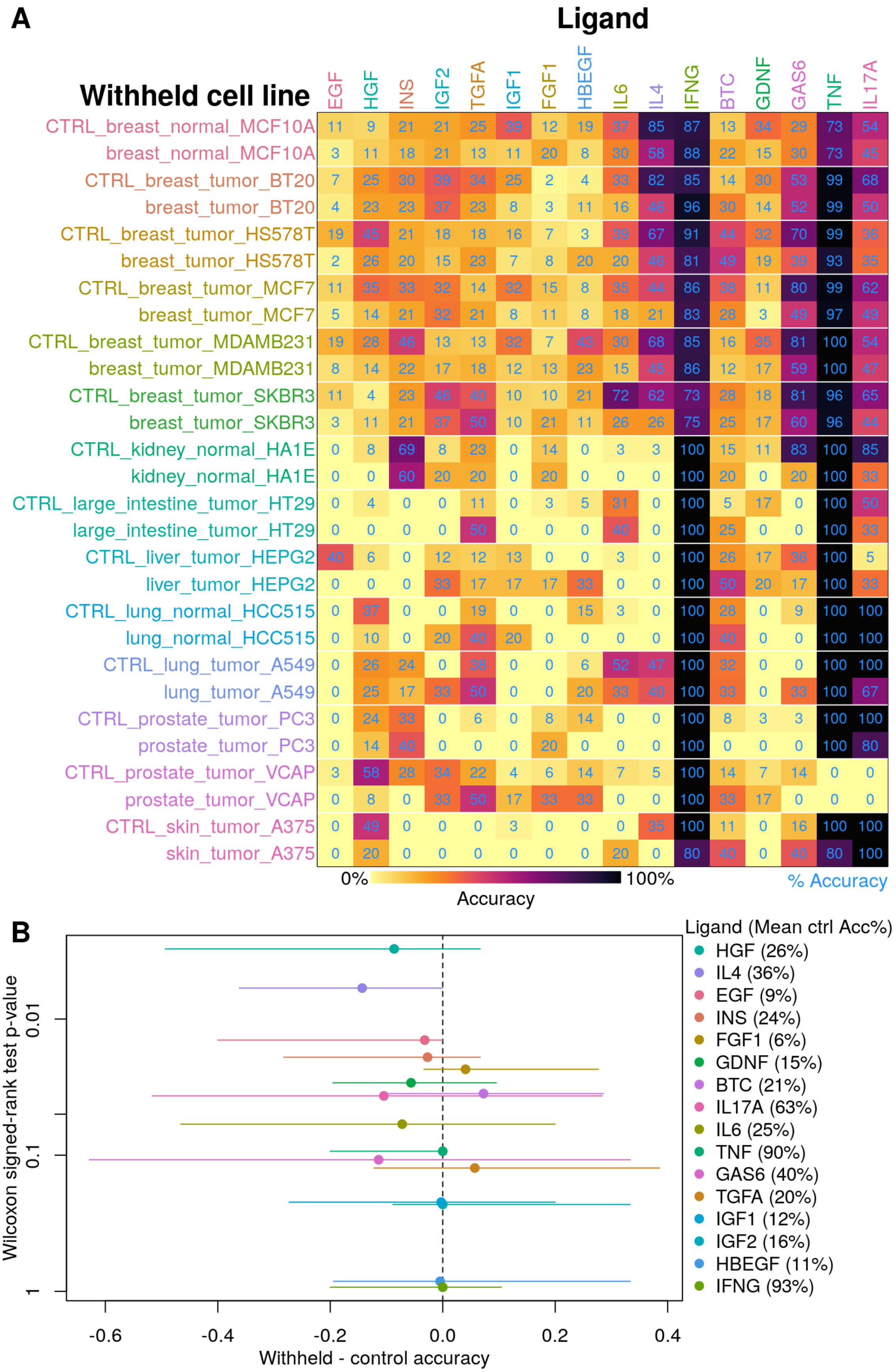
Using significantly differentially expressed genes, it is difficult to identify generalized (i.e. context-independent) transcriptomic signatures for most ligand-receptor interactions. It is possible that there are generalizable non-linear combinations of gene transcription responses to ligand perturbation that could also include genes that vary but are not significantly differentially expressed in the context of whole transcriptome statistical tests. Random forest models can make use of non-linear information, as they use votes from a series of decision trees to build classifiers. They have also proven to be useful as tools for feature selection and classification tasks in transcriptomics and achieve state of the art performance in these tasks (Kursa, 2014; Acharjee et al., 2020; Xu et al., 2021). To assess the ability of a random forest model to identify ligands from their transcriptomic profile, models were trained for each ligand across all assayed cell lines for the binary classification task of determining whether the sample was treated with the given ligand. A subset of 16 ligands (BTC, EGF, FGF1, GAS6, GDNF, HBEGF, HGF, IFNG, IGF1, IGF2, IL17A, IL4, IL6, INS, TGFA, and TNF) with higher sample numbers were used for this task to support good model power. These ligands were tested in the 9 cell lines used in the Connectivity Map data above and assayed at multiple concentrations and exposure durations in an additional 5 breast cancer cell lines. The models' performance on this task was generally poor (**Figure 3-7a** - rows with CTRL prefix), except when predicting IFNG and TNF treatment, which agrees with our differential gene expression results. To determine whether the models' prediction ability is generalizable across cell types, or conversely whether transcriptional responses to ligands are cell type specific, these models were trained again while withholding data from one cell line, and performance was tested on the withheld cell line (**Figure 3-7a**). Performance decreased when extrapolating to a novel cell line, except in the case of IFNG (**Figure 3-7b**, pairwise AUPR decreased significantly ($p < 1e-16$) by Wilcoxon signed-rank test). This is consistent with our previous finding that both correlation in transcriptional change and number of significantly differentially expressed genes increases more than expected by chance when considering cell type specific responses to a ligand stimulus, compared to considering a generalized response across cell types.



Withheld cell line	HGF	FGF1	INS	IGF1	HBEGF	TGFA	IGF2	BTC	GDNF	EGF	IL4	GAS6	IL6	IL17A	TNF	IFNG
CTRL_breast_normal_MCF10A	3	5	5	4	3	4	3	4	5	8	11	8	8	15	79	80
breast_normal_MCF10A	3	4	2	2	3	3	2	3	3	3	3	3	5	18	50	79
CTRL_breast_tumor_BT20	5	4	4	4	6	5	3	7	6	9	11	8	13	21	89	83
breast_tumor_BT20	3	2	3	4	3	5	5	5	3	5	7	8	7	13	79	82
CTRL_breast_tumor_HS578T	4	4	3	3	3	5	7	6	5	5	7	14	5	13	79	67
breast_tumor_HS578T	4	2	3	2	3	4	3	4	4	5	5	3	5	4	62	66
CTRL_breast_tumor_MCF7	4	4	5	6	11	7	6	5	6	10	6	11	11	29	85	81
breast_tumor_MCF7	3	2	4	3	3	3	7	4	1	4	5	12	4	15	85	69
CTRL_breast_tumor_MDAMB231	8	6	3	3	4	5	6	5	4	8	12	9	17	22	83	71
breast_tumor_MDAMB231	3	3	3	3	3	2	3	2	2	5	4	7	3	9	90	74
CTRL_breast_tumor_SKBR3	4	4	4	3	7	4	4	4	4	8	7	6	14	21	80	81
breast_tumor_SKBR3	2	3	4	3	3	7	5	5	3	4	6	9	3	14	73	62
CTRL_kidney_normal_HA1E	5	10	11	10	10	11	10	11	15	10	13	30	33	56	71	100
kidney_normal_HA1E	5	10	10	10	10	11	10	10	10	10	10	10	10	17	66	72
CTRL_large_intestine_tumor_HT29	6	10	13	13	13	13	11	13	16	11	17	13	16	37	86	91
large_intestine_tumor_HT29	5	10	10	13	13	13	10	13	13	10	13	13	10	17	76	85
CTRL_liver_tumor_HEPG2	5	9	9	9	9	9	9	9	10	26	10	20	35	21	78	100
liver_tumor_HEPG2	4	8	8	8	9	9	8	9	10	10	9	9	8	17	45	96
CTRL_lung_normal_HCC515	5	10	15	11	11	11	11	17	10	10	11	11	10	19	100	44
lung_normal_HCC515	6	10	10	10	13	10	10	10	10	10	10	10	10	19	30	56
CTRL_lung_tumor_A549	5	9	9	9	10	9	25	9	9	9	11	9	10	17	74	75
lung_tumor_A549	5	9	8	9	12	9	9	9	9	9	10	9	9	27	94	89
CTRL_prostate_tumor_PC3	8	10	10	13	11	13	10	11	17	10	12	13	13	16	76	53
prostate_tumor_PC3	9	10	10	13	10	13	10	10	17	10	10	13	13	10	43	87
CTRL_prostate_tumor_VCAP	5	9	9	9	9	10	8	11	9	9	10	9	10	64	79	95
prostate_tumor_VCAP	5	9	8	9	9	9	9	9	9	8	8	9	8	13	8	81
CTRL_skin_tumor_A375	6	10	10	10	10	10	10	10	10	10	11	10	30	18	74	57
skin_tumor_A375	6	10	10	10	10	10	10	10	10	10	11	11	10	17	74	79

Figure 3-7. Extrapolating to novel cell lines reduced performance of machine learning models trained to identify ligand-treated transcriptomes.
A. Each column displays the results of a series of random forest models trained to identify transcriptomes treated by the indicated ligand, and each row indicates the performance of the given model in the cell line that was withheld for training and used for testing, or the respective control ("CTRL"), where equal numbers of samples from all cell lines were used for training and testing. Random forest model performance represented by AUPR (darker colour is higher performance, values shown in blue text). Columns are ordered by increasing mean control AUPR. **B.** Boxplots showing change in AUPR between training on all cell lines versus training on all but a withheld cell line and predicting on the withheld cell line. Boxplots are ordered as columns in the above heat map by increasing mean control AUPR to give context to the possible magnitude of change in AUPR. Data to the left of the dashed vertical line indicate that holding out a cell line reduces model performance, while points to the right of the dashed line indicate improvement. Wilcoxon signed-rank test was used to determine significance of pairwise change in accuracy, p-values reported to the left of each boxplot.

A complementary classification task is to identify which ligand treatment drives each set of transcriptional changes. A different set of random forest models were trained to perform this task and this generated similar results to the previous random forest analysis (**Figure 3-8**). Median accuracy when extrapolating to novel cell lines was reduced compared to the already poor accuracy (32%) when training on all cell lines (5% decrease in median accuracy, p = 1.3e-4 by Wilcoxon signed-rank test).



Withheld cell line	EGF	HGF	INS	IGF2	TGFA	IGF1	FGF1	HBEGF	IL6	IL4	IFNG	BTC	GDNF	GAS6	TNF	IL17A
CTRL_breast_normal_MCF10A	11	9	21	21	25	39	12	19	37	85	87	13	34	29	73	54
breast_normal_MCF10A	3	11	18	21	13	11	20	8	30	58	88	22	15	30	73	45
CTRL_breast_tumor_BT20	7	25	30	39	34	25	2	4	33	82	85	14	30	53	99	68
breast_tumor_BT20	4	23	23	37	23	8	3	11	16	46	96	30	14	52	99	50
CTRL_breast_tumor_HS578T	19	45	21	18	18	16	7	3	39	67	91	44	32	70	99	36
breast_tumor_HS578T	2	26	20	15	23	7	8	20	20	46	81	49	19	39	93	35
CTRL_breast_tumor_MCF7	11	35	33	32	14	32	15	8	35	44	86	38	11	80	99	62
breast_tumor_MCF7	5	14	21	32	21	8	11	8	18	21	83	28	3	49	97	49
CTRL_breast_tumor_MDAMB231	19	28	46	13	13	32	7	43	30	68	85	16	35	81	100	54
breast_tumor_MDAMB231	8	14	22	17	18	12	13	23	15	45	86	12	17	59	100	47
CTRL_breast_tumor_SKBR3	11	4	23	46	40	10	10	21	72	62	73	28	18	81	96	65
breast_tumor_SKBR3	3	11	21	37	50	10	21	11	26	26	75	25	17	60	96	44
CTRL_kidney_normal_HA1E	0	8	69	8	23	0	14	0	3	3	100	15	11	83	100	85
kidney_normal_HA1E	0	0	60	20	20	0	20	0	0	0	100	20	0	20	100	33
CTRL_large_intestine_tumor_HT29	0	4	0	0	11	0	3	5	31	0	100	5	17	0	100	50
large_intestine_tumor_HT29	0	0	0	0	50	0	0	0	40	0	100	25	0	0	100	33
CTRL_liver_tumor_HEPG2	40	6	0	12	12	13	0	0	3	0	100	26	17	36	100	5
liver_tumor_HEPG2	0	0	0	33	17	17	17	33	0	0	100	50	20	17	100	33
CTRL_lung_normal_HCC515	0	37	0	0	19	0	0	15	3	0	100	28	0	9	100	100
lung_normal_HCC515	0	10	0	20	40	20	0	0	0	0	100	40	0	0	100	100
CTRL_lung_tumor_A549	0	26	24	0	38	0	0	6	52	47	100	32	0	0	100	100
lung_tumor_A549	0	25	17	33	50	0	0	20	33	40	100	33	0	33	100	67
CTRL_prostate_tumor_PC3	0	24	33	0	6	0	8	14	0	0	100	8	3	3	100	100
prostate_tumor_PC3	0	14	40	0	0	0	20	0	0	0	100	0	0	0	100	80
CTRL_prostate_tumor_VCAP	3	58	28	34	22	4	6	14	7	5	100	14	7	14	0	0
prostate_tumor_VCAP	0	8	0	33	50	17	33	33	0	0	100	33	17	0	0	0
CTRL_skin_tumor_A375	0	49	0	0	0	3	0	0	0	35	100	11	0	16	100	100
skin_tumor_A375	0	20	0	0	0	0	0	0	20	0	80	40	0	40	80	100

Figure 3-8. Accuracy in identifying which ligand treatment caused the changes in the transcriptome of a novel cell line.
A. Random forest models were trained to identify the ligand treatment (columns) causing transcriptomic changes in a withheld cell line (rows), compared to performance when trained on all cell lines (leave-one-out testing, rows labeled 'CTRL'). Darker colour represents higher accuracy, and number of samples per ligand and cell line is shown in blue. **B.** Change in accuracy between predicting on a novel cell line versus the control model trained on all cell lines (x-axis - colours match ligand column names above). Lines represent range across cell lines with median indicated by a point on each line. Lines extending to the left of the dashed vertical line indicate that holding out a cell line reduces model performance, and vice versa for lines extending to the right of the dashed line. Wilcoxon signed-rank test was used to determine significance of pairwise change in accuracy (y-axis). Mean control accuracy is indicated in brackets in the legend to give context to the possible magnitude of accuracy change.

One possible explanation for the general difficulty in establishing transcriptional signatures of ligand response (either linear or nonlinear) is poor signal in this data. This may indicate a lack of response to the ligand perturbation, possibly because the ligand's cognate receptor is not expressed in that cell line. To assess this possibility, receptor expression was correlated with ligand prediction accuracy in our random forest model tasks (**Figure 3-9**). Cognate receptors for each ligand were determined from a database of known ligand-receptor interactions (Ximerakis et al., 2019), and their quantile-normalized gene abundance per cell line were calculated from the Connectivity Map control data. While epidermal growth factor (EGF) receptor family members showed reasonable correlation with prediction accuracy of transcriptional response to the EGF ligand family member TGFA, for most ligands the correlation between cognate receptor expression and ligand prediction accuracy was unclear. Thus, we do not find that receptor expression level explains ligand perturbation signal in general.

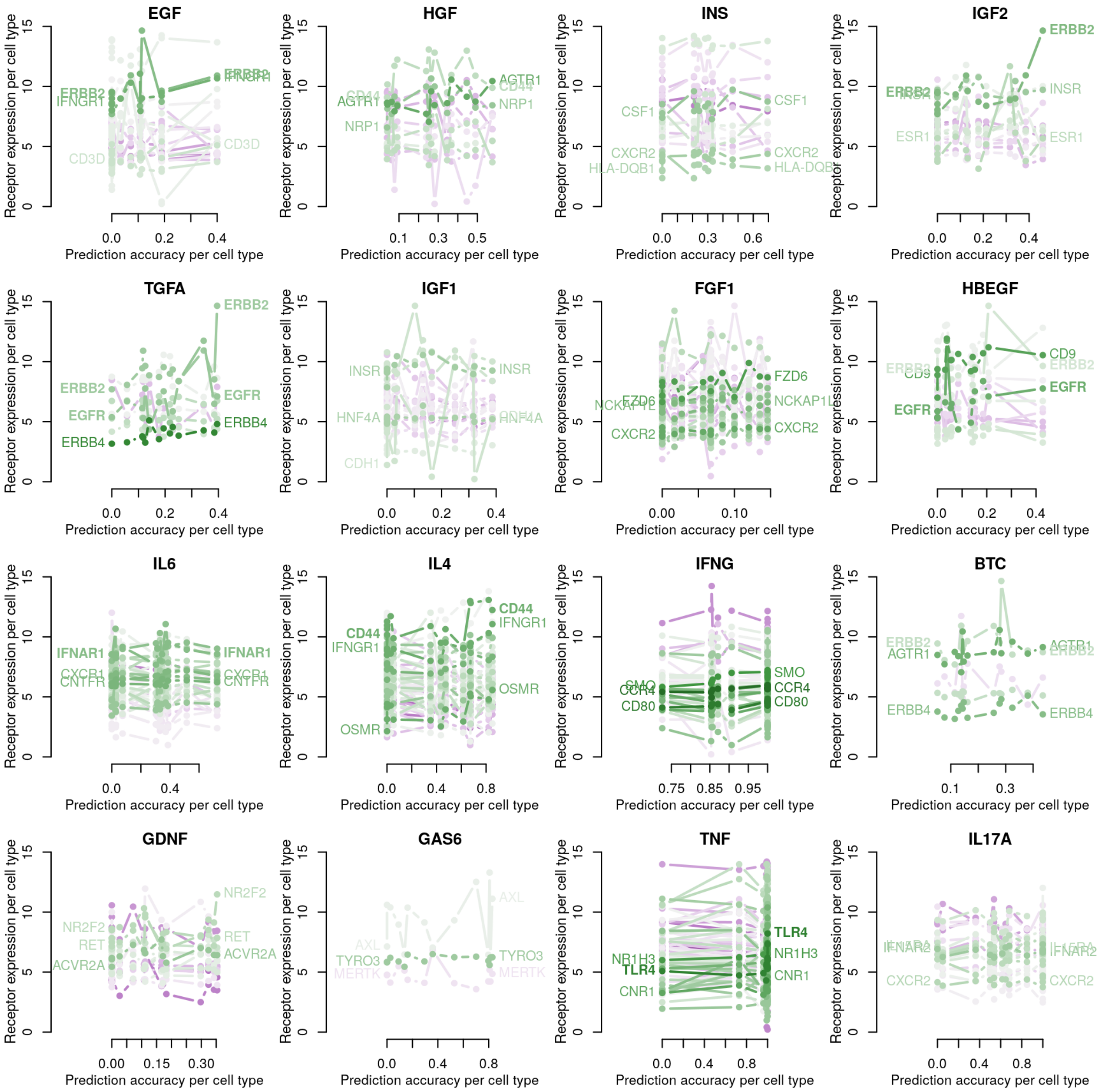


Figure 3-9. Cognate receptor gene expression does not sufficiently explain differences in transcriptional response to ligand perturbation.

Cognate receptor expression for each ligand was compared to ligand classification accuracy per cell line. Lines for each receptor are coloured by the Spearman correlation coefficient between expression and accuracy per cell type (purple is negative, green is positive), and the three most correlated receptor genes are labeled.

3.2.5 Cell type specific transcriptional response to ligand perturbation is corroborated by independent microarray assays

To corroborate our main finding from the Connectivity Map database that transcriptional response to ligand perturbation is cell type dependent, we repeated our analysis on a complementary transcriptomic dataset. The authors of the cell-cell interaction prediction method NicheNet (Browaeys et al., 2020) used the NCBI Gene Expression Omnibus (Barrett et al., 2013) to curate a collection of microarray experiments in which transcriptional response to ligand stimulation was assayed. Twenty ligands appeared more than once in the collected data, of which eleven ligands were assayed in more than one independent experiment from different cell types and thirteen were assayed more than once in the same cell type by independent labs in separate experiments (**Figure 3-10**). As was done with the Connectivity Map data, we calculated pairwise Spearman correlation coefficients between all fold-change values from experiments where the same ligand was used to perturb cells (**Figure 3-11a**). As with the Connectivity Map data, change in gene expression upon ligand treatment in different cell lines was poorly correlated (mean SCC of 0.1). Ligand-induced gene expression changes were significantly better correlated when comparing independent experiments in the same cell type (mean SCC of 0.35, p = 3.4e-7 by Wilcoxon rank-sum test). This mirrors our findings analyzing the Connectivity Map data, albeit with a much stronger effect and better coverage of the whole transcriptome. By comparing overall ligand-induced changes in gene expression within and between cell types, the published transcriptomic data clearly support the finding that transcriptional response to ligand treatment is highly dependent on cell type.

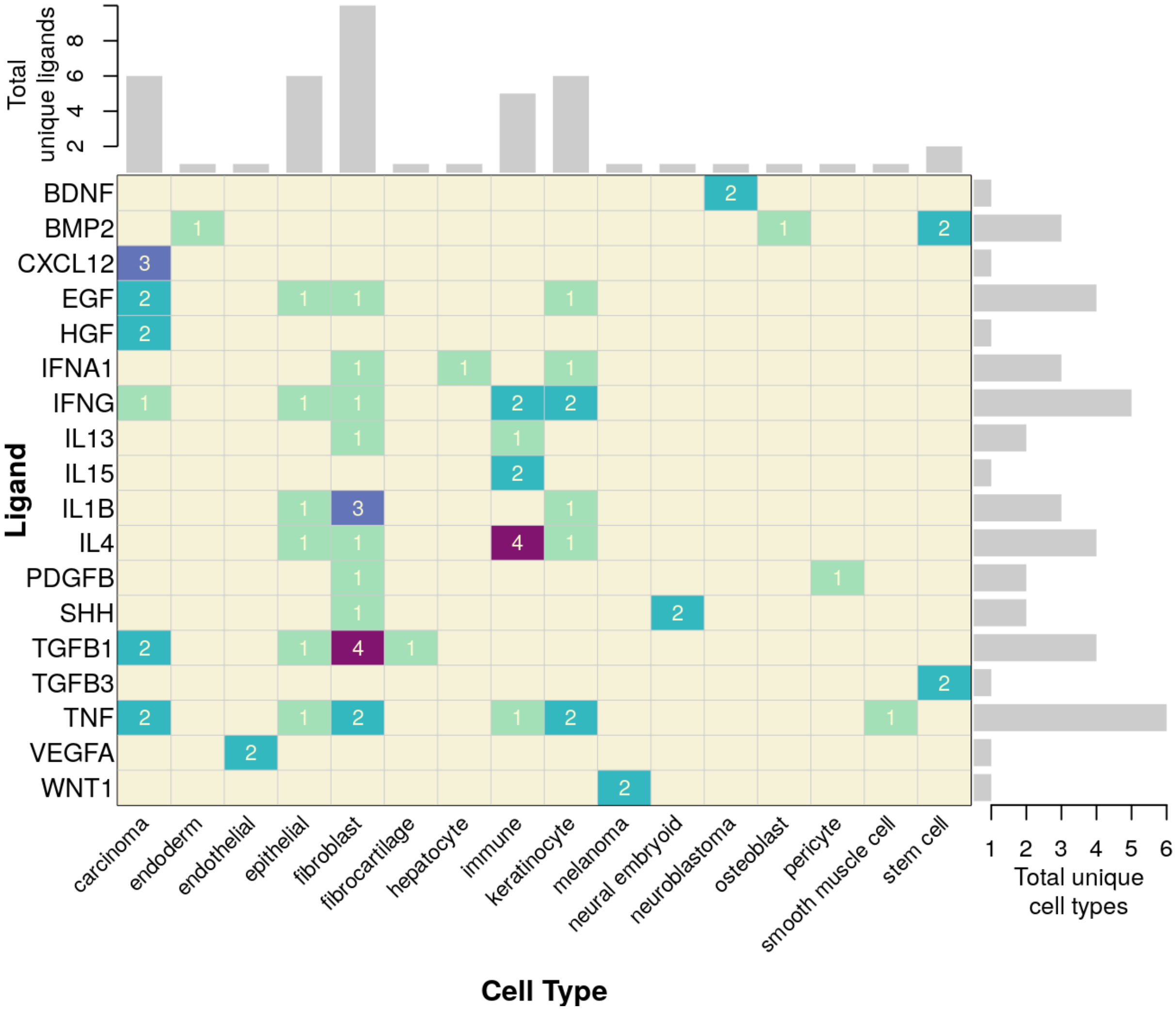


Figure 3-10. Distribution of experimental conditions from NicheNet-curated transcriptomics assays assessing change in gene expression caused by ligand treatment.

The matrix indicates which ligands (rows) were assayed in which cell types (columns), with the number at each matrix element indicating how many transcriptional experiments from independent labs contained the same combination of ligand and cell type (darker colour reflecting higher count). The bar plots indicate row and column sums for the unique cell types in which each ligand was assayed, or unique ligands assayed in each cell type, respectively.

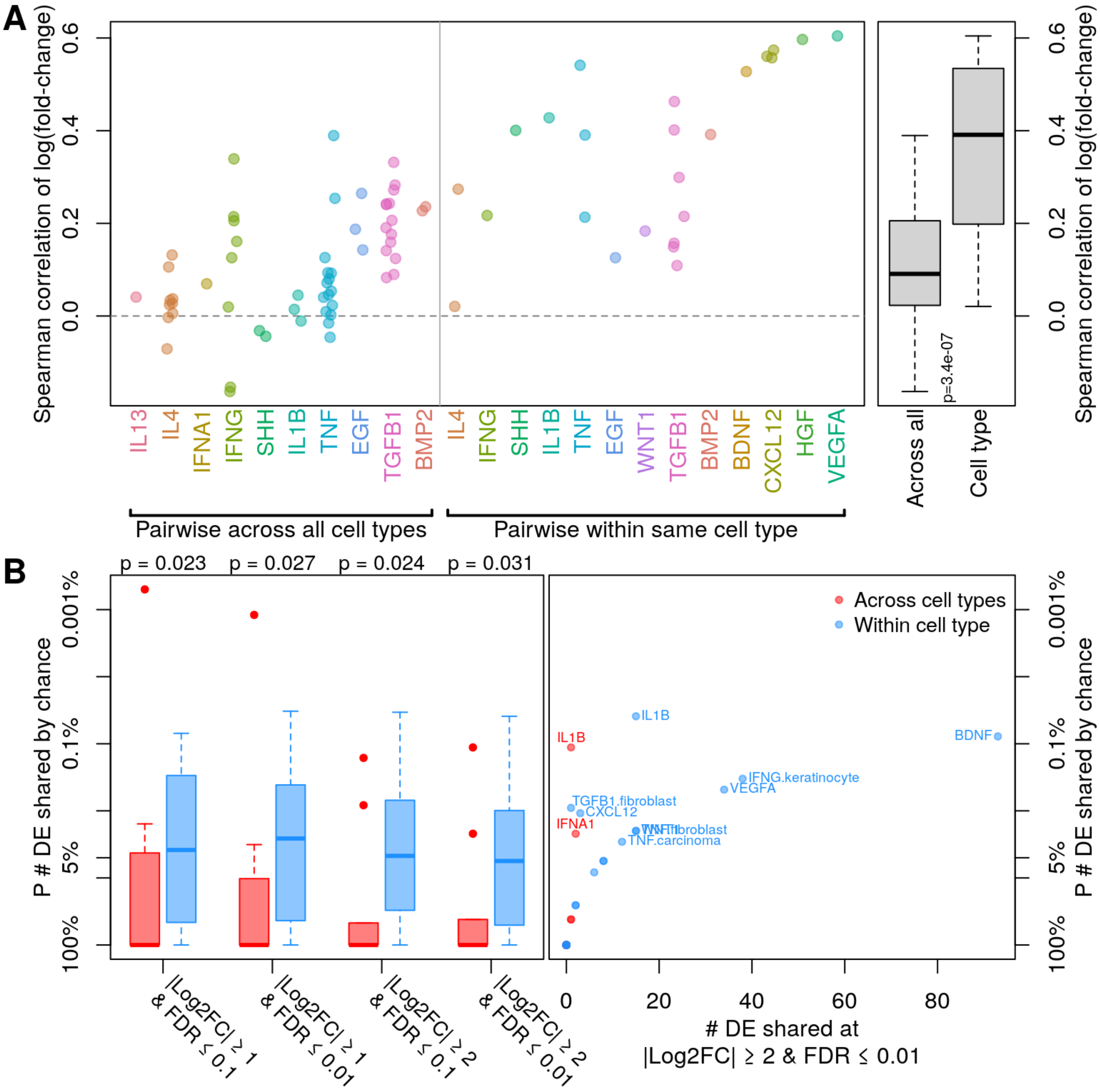


Figure 3-11. Context-specific response to ligand stimulus is corroborated by independent transcriptomics assays.

A. A strip plot showing pairwise Spearman correlations between log fold-change values from NicheNet-curated transcriptomics assays assessing change in gene expression caused by ligand treatment, for all pairs of assays for each ligand (left) or pairs of assays using the same cell type (right). Correlations in gene expression change between samples in different cell types treated with the same ligand are shown on the left, with correlations between samples from the same cell type and ligand treatment on the right. The data is summarized in the boxplot on the far right, which indicates that ligand-dependent transcriptional changes are significantly more correlated in the same cell type (p = 3.4e-7 by Wilcoxon rank-sum test). **B.** Ligand-dependent gene expression changes from published transcriptomics assays common across all assays. On the left are box plots showing the statistical significance of overlapping differentially expressed genes (y-axis), using various cutoffs to define differential expression (x-axis). Statistical significance is represented as the probability of seeing such overlap occurring by chance, as calculated by permutation testing. Indicated p-values (top) are from Wilcoxon rank-sum tests comparing probabilities of overlap by chance between samples from different cell types (red) versus within the same cell type (blue). On the right is a representative scatter plot showing the data from one pair of boxplots, with the x-axis representing the number of differentially expressed (DE) genes shared between all datasets treated with the same ligand (in red) or all datasets from the same cell type and treated with the same ligand (in blue). Ligands or ligand and cell type were labeled if the overlap p-value was ≤ 0.05.

We also used the NicheNet-curated gene expression data to assess the finding that some genes display significant and consistent changes in gene expression upon ligand treatment, but these are context-specific. Significantly DE genes in response to ligand treatment were defined from each ligand-perturbed microarray dataset at multiple effect size and statistical significance thresholds to ensure robustness of findings. The number of significantly DE genes shared across all datasets treated with the same ligand or treated with the same ligand in the same cell type, was then counted, and the probability of those numbers of shared DE genes occurring by chance was calculated by permutation testing (**Figure 3-11b**). The chance of differential gene overlap between datasets was significantly less likely when considering ligand-induced changes in gene expression within each cell type, irrespective of the fold change and false discovery rate thresholds used to define significant differential expression (**Figure 3-11b** right panel, $p < 0.05$ by Wilcoxon rank-sum tests). This is consistent with the finding from Connectivity Map data analysis that most ligands show cell line specificity in transcriptional response (**Figure 3-5b**).

From this analysis, we find that only two ligands, IL1B and IFNA1, share a significant number of ligand-responsive DE genes across all cell types in the NicheNet-curated data. None of these genes were DE across cell types in the Connectivity Map data (**Figure 3-12**). Of the ligands assayed in NicheNet-curated transcriptomics data, TNF and IFNG caused the most generalizable transcriptional changes in the Connectivity Map analysis (respectively 9 and 3 significantly DE genes at 5% FDR, **Figure 3-3**). The most consistently DE gene in response to both TNF and IFNG is *ICAM1*, but it is not upregulated in all the TNF-treated cell lines, lacking significant change in expression in both smooth muscle and immune cell contexts from the NicheNet-curated data, and in the prostate tumour line VCAP from Connectivity Map (**Figure 3-13**). Two other consistently upregulated genes in Connectivity Map in response to TNF, *NFKBIA* and *RELB*, were similarly inconsistent, with no significant differential expression in the breast cell line MCF10A or prostate tumour line VCAP in Connectivity Map. In response to IFNG, *ICAM1* was upregulated in many, but not all, cell contexts. It showed no significant response in immune or fibroblast cells from the NicheNet curated data, nor in the breast tumour line HS578T from Connectivity Map (**Figure 3-14**). Thus, even the most consistent transcriptional response signatures across many independent data are not always observed, emphasizing that cellular context is still important even in these special cases.

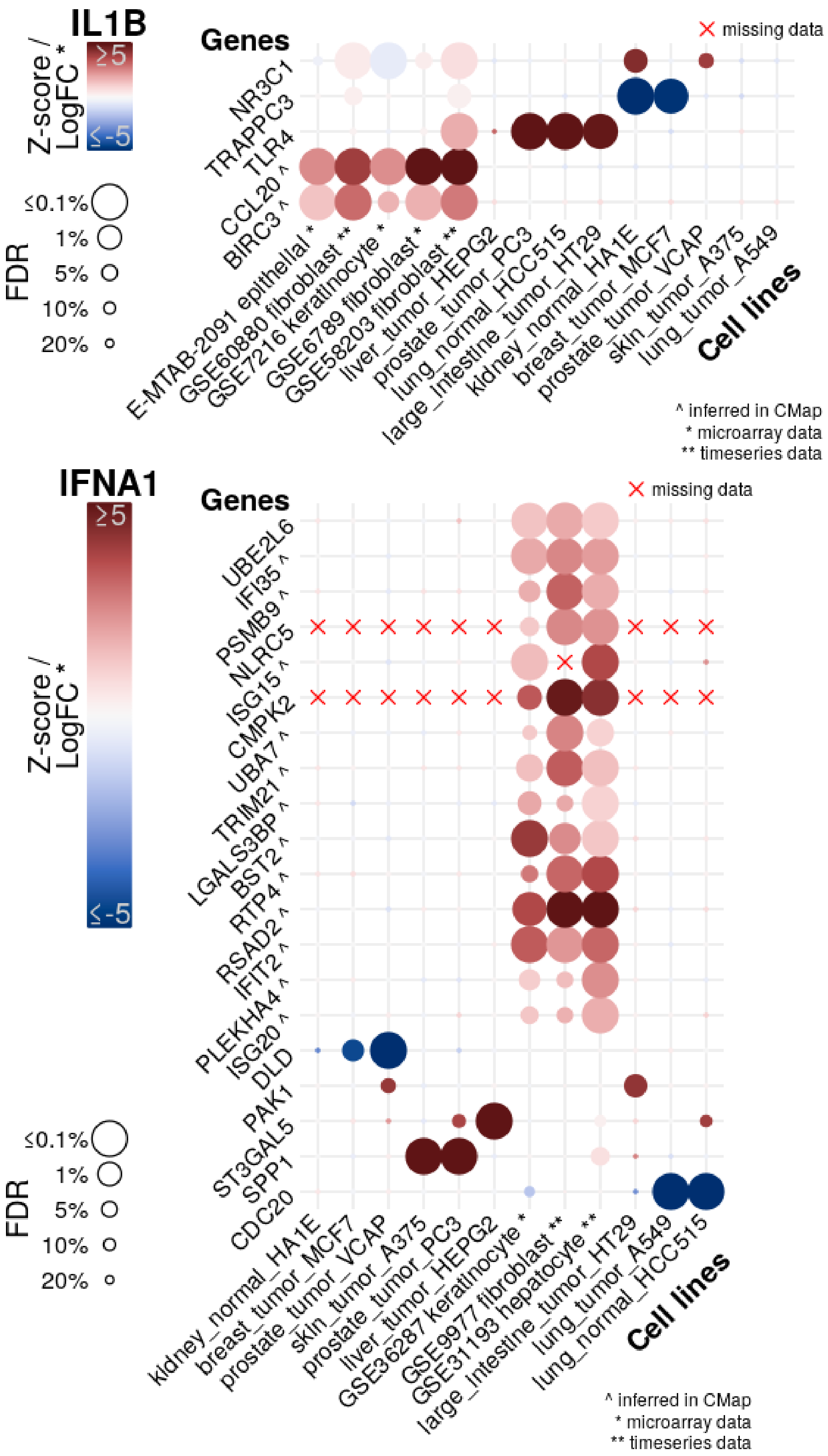


Figure 3-12. Gene expression changes in response to IL1B or IFNA1 differ between cell types. Matrix plots showing mean change in gene expression per cell line upon IL1B or IFNA1 treatment for selected genes from both Connectivity Map and NicheNet-curated transcriptomics data. Cell lines are on each column, with NicheNet-curated transcriptomes indicated with *, and

those from time series experiments where only magnitude of gene expression change was recorded with a **. Genes (in rows) were included if they were significantly differentially expressed (at 10% FDR) in at least two Connectivity Map cell lines or the maximum possible number of overlapping NicheNet-curated cell lines. Genes whose expression was inferred in Connectivity Map are indicated with ^. Genes whose expression was not reported in an experiment are indicated with a red x. Dot colour indicates difference in gene expression upon ligand treatment, measured as Connectivity Map Z-score and log2 fold-change from NicheNet-curated transcriptomes. Dot size indicates false discovery rate-corrected significance.

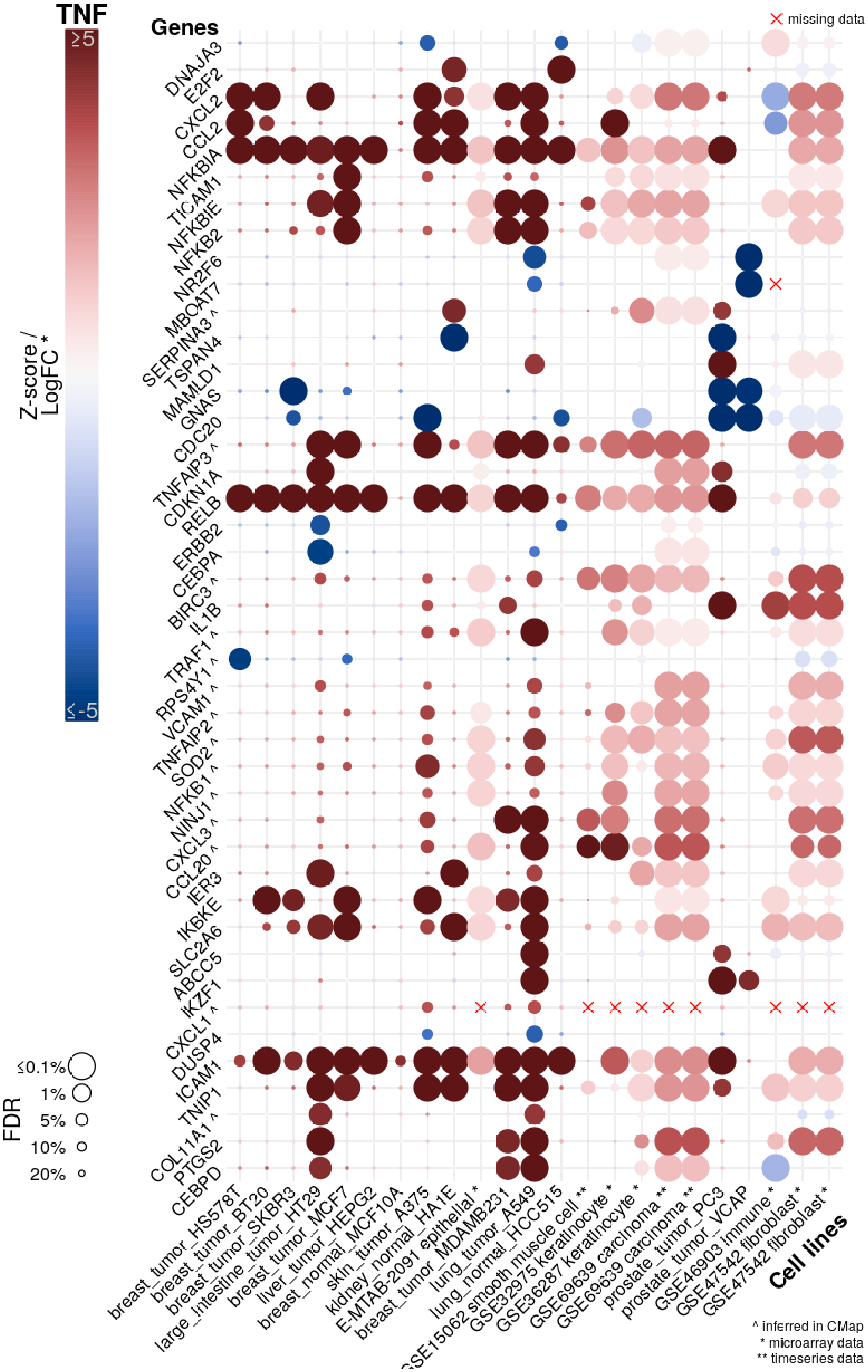

TNF
≥5
≤-5
Z-score /
LogFC *
FDR
≤0.1%
1%
5%
10%
20%
Genes
× missing data
DNAJA3
E2F2
CXCL2
CCL2
NFKBIA
TICAM1
NFKBIE
NFKB2
NR2F6
MBOAT7
SERPINA3 ^
TSPAN4
MAMLD1
GNAS
CDC20
TNFAIP3 ^
CDKN1A
RELB
ERBB2
CEBPA
BIRC3 ^
IL1B
TRAF1 ^
RPS4Y1 ^
VCAM1 ^
TNFAIP2 ^
SOD2 ^
NFKB1 ^
NINJ1 ^
CXCL3 ^
CCL20 ^
IER3
IKBKE
SLC2A6
ABCC5
IKZF1
CXCL1 ^
DUSP4
ICAM1
TNIP1
COL11A1 ^
PTGS2
CEBPD
breast_tumor_HS578T
breast_tumor_BT20
breast_tumor_SKBR3
large_intestine_tumor_HT29
breast_tumor_MCF7
liver_tumor_HEPG2
breast_normal_MCF10A
skin_tumor_A375
kidney_normal_HA1E
E-MTAB-2091 epithelial *
breast_tumor_MDAMB231
lung_tumor_A549
lung_normal_HCC515
GSE15062 smooth muscle cell **
GSE32975 keratinocyte *
GSE36287 keratinocyte *
GSE69639 carcinoma **
GSE69639 carcinoma **
prostate_tumor_PC3
prostate_tumor_VCAP
GSE46903 immune *
GSE47542 fibroblast *
GSE47542 fibroblast *
Cell lines
^ inferred in CMap
* microarray data
** timeseries data

Figure 3-13. Gene expression changes in response to TNF differ between cell types.
A matrix plot showing mean change in gene expression per cell line upon TNF treatment for selected genes from both Connectivity Map and NicheNet-curated transcriptomics data. Cell lines are on each column, with NicheNet-curated transcriptomes indicated with *, and those from time series experiments where only magnitude of gene expression change was recorded with a **. Genes (in rows) were included if they were significantly differentially expressed (at 10% FDR) in at least two Connectivity Map cell line assays or the maximum possible number of overlapping NicheNet-curated cell line assays. Genes whose expression was inferred in Connectivity Map are indicated with ^. Genes whose expression was not reported in a given experiment are indicated with a red x instead of a dot. Dot colour indicates difference in gene expression upon ligand treatment, measured as Connectivity Map Z-score and log2 fold-change from NicheNet-curated transcriptomes. Dot size indicates false discovery rate-corrected significance. If a gene were consistently differentially expressed in response to ligand treatment, an unbroken horizontal line of circles of the same colour would appear on this plot.

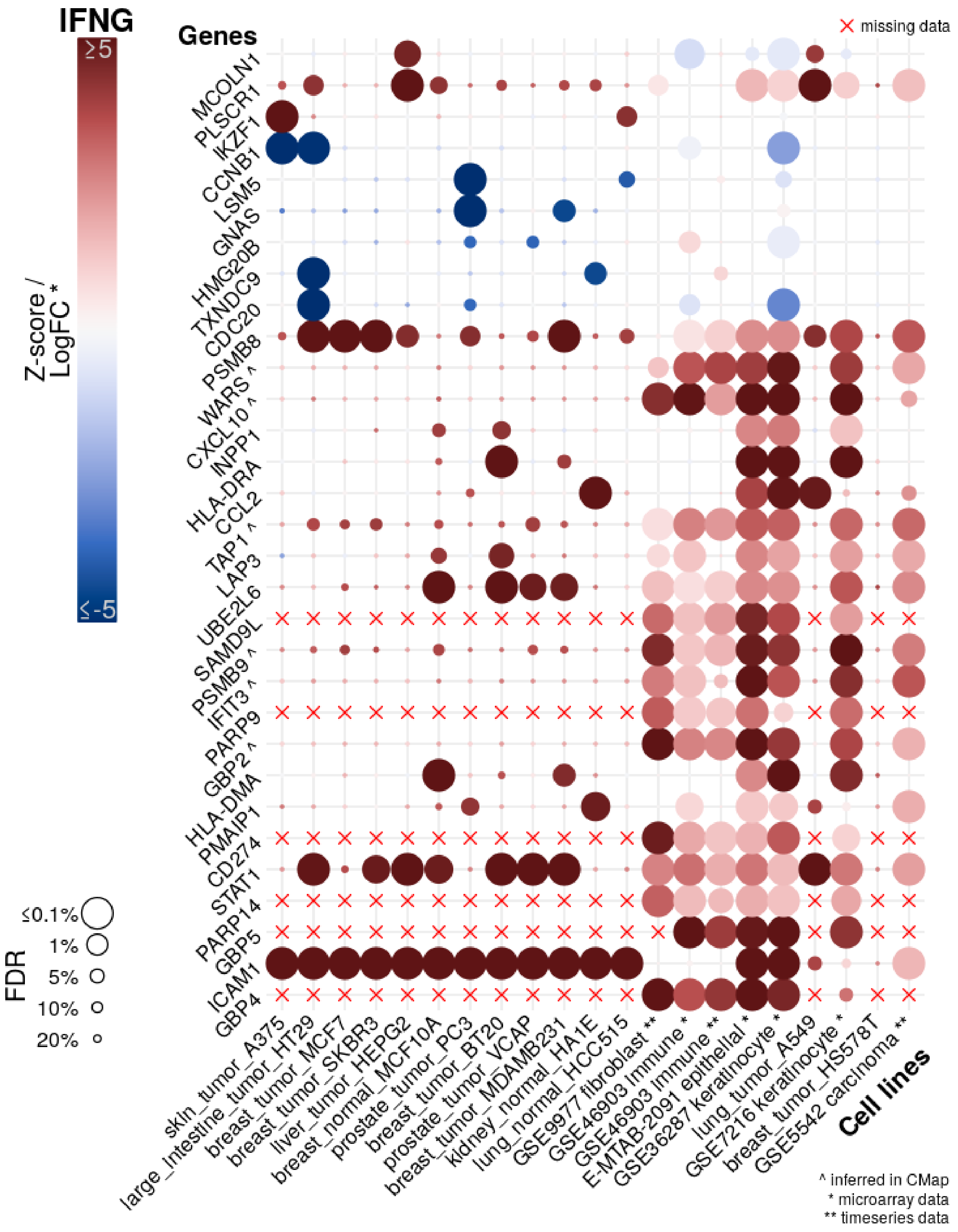


Figure 3-14. Gene expression changes in response to IFNG differ between cell types. A matrix plot showing mean change in gene expression per cell line upon IFNG treatment for selected genes from both Connectivity Map and NicheNet-curated transcriptomics data. Cell lines are on each column, with NicheNet-curated transcriptomes indicated with *, and those from time series experiments where only magnitude of gene expression change was recorded with a

**. Genes (in rows) were included if they were significantly differentially expressed (at 10% FDR) in at least two Connectivity Map cell lines or the maximum possible number of overlapping NicheNet-curated cell lines. Genes whose expression was inferred in Connectivity Map are indicated with ^. Genes whose expression was not reported in an experiment are indicated with a red x. Dot colour indicates difference in gene expression upon ligand treatment, measured as Connectivity Map Z-score and log2 fold-change from NicheNet-curated transcriptomes. Dot size indicates false discovery rate-corrected significance.

3.2.6 Difference in ligand response is predicted by dissimilarity of cell type transcriptomes

There are many potential mechanisms facilitating heterogenous responses between cell types to the same ligand treatment. Agonist and antagonist cofactors, both extracellular and present at the cell membrane, may be modulating receptor activation. Availability of components of the downstream intracellular signaling cascades, especially transcription factors, may allow cells to tune response to ligand treatment. Finally, chromatin state may dictate what genes are available to be expressed. The status of at least some of these factors affecting transcriptional response to a ligand may be encoded in the receiving cell's transcriptome. To test this hypothesis, the transcriptomes of cell types being treated with the same ligand were compared. If the transcriptome contains information guiding the cell's response to ligand stimulus, cells with similar transcriptomes should respond more similarly to treatment with the same ligand. By comparing pairwise correlation of untreated cell transcriptomes with pairwise correlation of those cells' change in gene expression in response to the same ligand, we can see evidence that this is true (**Figure 3-15**). There was a strong correlation between transcriptome similarity and similarity of ligand response (Pearson correlation coefficient of 0.79) with good distribution of tested ligands across the range of comparisons. Thus, the transcriptome of the receptor-expressing cell contains information that can help predict its transcriptional response to a ligand.

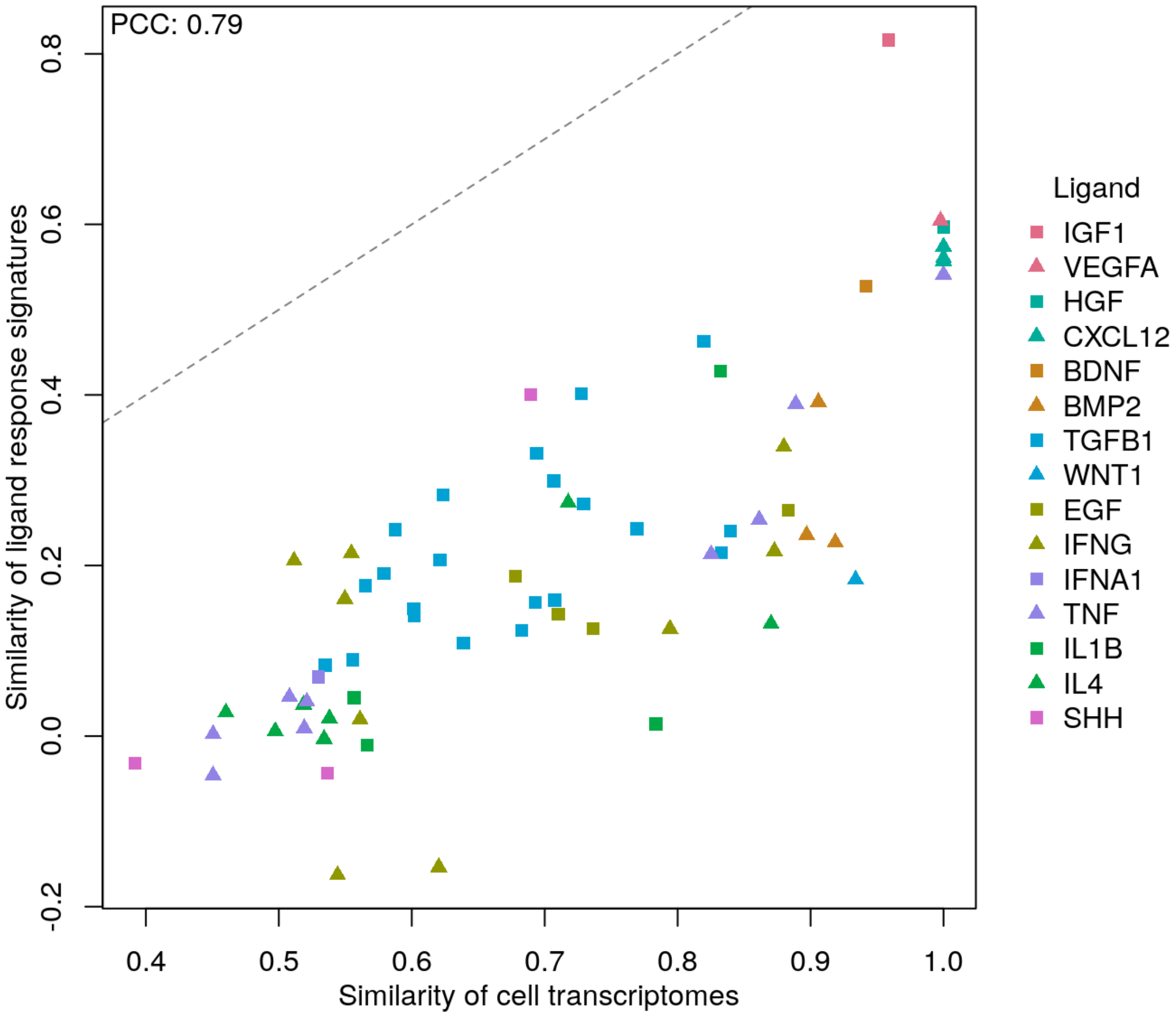


Figure 3-15. Correlation between cell transcriptome similarity and response to ligand stimulus. Spearman rank correlations of untreated transcriptomes from pairs of cell lines (x-axis) were compared to Spearman rank correlations of changes in gene expression upon treatment of those cell line pairs with the same ligand (y-axis). The identity line is shown dotted for reference. There is a Pearson correlation coefficient (PCC) of 0.79 between untreated transcriptome correlation and correlation of log fold-change upon treatment.

3.3 Discussion

If the various cell types of a tissue analyzed by single-cell RNAseq are communicating through ligand-receptor interactions, one may expect to see evidence of these ligand-mediated signaling events in the receiving cell's transcriptome. In this study, two independent and complementary transcriptome assay databases measuring gene expression change in response to ligand perturbation were used to investigate this hypothesis. Within each cell type tested, change in gene expression in response to each ligand was often consistent and predictable, as evidenced by nearly half of the ligands in the Connectivity Map data causing significant differential expression of more genes than expected by chance. However, ligand response across cell types was generally inconsistent and unpredictable. These results have important implications for cell-cell interaction prediction from scRNAseq data, which cannot rely on generalizable, characteristic, and easily predictable ligand response signatures to identify receptors that are actively signaling to effect downstream transcriptional changes. Instead, only cell type-specific ligand response signatures will be useful. Unfortunately, this cellular context dependent ligand response information is not available for most cell types.

We also found that certain genes are activated by many tested ligands, as evidenced by the vertical lines of significant differential expression in **Figure 3-3**. That independent experimental perturbations can have similar transcriptional effects has been noted before, most notably by Crow et al., who built a prior probability of differential expression for all genes with 80% accuracy across a wide variety of transcriptomic datasets (Crow et al., 2019). While one might attempt to mitigate the apparent redundancy in transcriptional response to ligand perturbation by making more coarse-grained predictions by ligand family, Crow et al. suggests that by virtue of either the way transcriptomic data is collected, or biological redundancy in transcriptional response to stimuli, that may not be sufficient to ensure uniqueness of ligand signaling signatures.

Cell-cell communication inference is a popular method for understanding tissue function from scRNAseq data. The modern methods outlined in **Table 3-1** use the receiving cell's transcriptome to support their ligand-receptor predictions. To define the expected transcriptional response to receptor stimulus, they use canonical pathways to link receptors to transcription factors and then gene regulatory networks to link those transcriptional factors to their target genes. These GRNs are either sourced from databases or inferred from the input scRNAseq data. Unfortunately, current pathway databases do not consider cell type context in their networks (Larsen et al., 2019), meaning that methods using database derived GRNs implicitly assume that ligand response signatures are consistent between cell types. While inferring the GRN from the input scRNAseq data presumably ensures cell type specificity, even state-of-the-art methods struggle to perform this task accurately (Holland et al., 2020; Pratapa et al., 2020; Kang et al., 2021). Thus, the prediction of cell type specific ligand-response signatures is an unsolved problem that must be addressed before these cell-cell interaction inference methods can deliver their promised improvement in prediction specificity.

Though this work highlights concerns with current methods, it also suggests paths forward. As seen in **Figure 3-15**, a cell's transcriptome correlates with its response to ligand stimulus. This indicates that cell type specific ligand-response signatures may be inferred from a multilayer network model that includes cell type specificity. Simply weighting nodes in the signal transduction network by their expression in the cells of interest may be sufficient to improve the specificity of that layer of the network. The rise of single cell multiome assays that combine transcriptome with gene accessibility (scATACseq) information may improve cell type specific GRN inference (Bravo González-Blas et al., 2022; Kamimoto et al., 2023). Finally, the strong correlation between transcriptome similarity and ligand response similarity suggests that if experimentally derived ligand response signatures are available for a cell type similar to the cells being studied, applying that those signatures to support cell-cell interaction predictions is reasonable.

Ultimately, current cell-cell interaction predictions can provide a non-specific list of potential ligand-receptor pairs as a hypothesis generation exercise, but study authors must test these predictions to both identify true sources of cell-cell signaling and provide the sort of ground-truth data necessary for further improvement of these predictions.

3.4 Materials & Methods

3.4.1 Data & code availability

All analysis was performed in the R statistical programming language (R Core Team, 2018), with all code available at https://github.com/BaderLab/Brendan_CCInxPred. Connectivity Map data is deposited in NCBI GEO with accession GSE92742 (Subramanian et al., 2017). NicheNet-curated data is from the 'expression_settings.rds' file deposited in Zenodo as record 3260758 (Browaeys et al., 2019).

3.4.2 Connectivity Map differentially expressed gene expression analysis

Connectivity Map calculates robust Z-scores for gene *i* of the n^{th} sample on the plate as:

$$z_{i,n} = \frac{x_{i,n} - \text{median}(x_i)}{1.4826 \cdot \text{MAD}(x_i)}$$

where x_i is the normalized gene expression of gene i across all samples on the plate, and MAD is the median absolute deviation (Subramanian et al., 2017). These scores, representing normalized changes in gene expression upon ligand treatment, were averaged across all samples treated with the same ligand, same ligand in the same cell line, or from the same technical replicate set. False Discovery Rate (FDR)-corrected p-values were calculated from averaged Z-scores using the Gaussian distribution. A background distribution of FDR-corrected p-values was generated by averaging each gene's Z-scores across a random set of samples of equivalent size. This background distribution was used to calculate the probability of seeing as many differentially expressed genes by chance for each set of samples. The R scripts to perform these calculations and generate **Figure 3-3** and **Figure 3-5** are 'lig295_DE_FDR.R' and 'DEoverlap_FDR_forPaper.Rmd' in the GitHub repository cited above.

3.4.3 Connectivity Map sample correlations

Spearman correlation coefficients were calculated for all pairs of samples from the same cell line, treated with the same ligand, same ligand in the same cell line, or from the same technical replicate set. Differences in the distributions of pairwise correlation coefficients were tested using the Wilcoxon rank-sum test (Wilcoxon, 1945). The R scripts to perform these calculations and generate **Figure 3-4** are 'Zcorr.R' and 'Zcorr_forPaper.Rmd' in the GitHub repository cited above.

3.4.4 Connectivity Map random forest modelling

Random forest models were trained for each ligand using *ranger* to classify samples as either treated with the ligand or not (Wright and Ziegler, 2017). Training sets were balanced by randomly selecting an equivalent number of samples treated with other ligands to represent the negative case for each ligand. "Control" models were trained on samples from all cell lines, while "test" models for each cell line were trained on all lines except the cell line in question. The models were tested on all withheld data. The R scripts to perform these calculations are generate **Figure 3-7** are `200813_newVall_probs_lvl4.R` and 'RFresults_forPaper.Rmd' in the GitHub repository cited above.

Random forest models were also trained on the same datasets as above to identify which of the 16 ligands each sample was treated with. The R scripts to perform these calculations and generate **Figure 3-8** are '200524_lvl5_mixall_leave1out.R' and 'LigPred_extrapolate.Rmd' in the GitHub repository cited above.

3.4.5 Connectivity Map receptor availability analysis

Average of quantile-normalized gene expression values of cognate receptors for each ligand in each cell line were correlated with accuracy of binary classification by the relevant random forest model. Cognate receptors for each ligand were identified from the ligand-receptor database used by *CCInx* (Ximerakis et al., 2019). The R script to perform these calculations and generate **Figure 3-9** is `RecExpr.Rmd' in the GitHub repository cited above.

3.4.6 NicheNet database correlation and differential expression analysis

Spearman correlation coefficients between log-scaled fold-change values of gene expression difference between treated and untreated samples were calculated for all pairs of samples treated with the same ligand or treated with the same ligand in the same cell type. Differences in the distributions of pairwise correlation coefficients were tested using the Wilcoxon rank-sum test.

Differentially expressed genes in each sample were determined using multiple fold-change and FDR thresholds. To determine the statistical significance of overlap between differentially expressed gene lists from samples treated with the same ligand or treated with the same ligand in the same cell type, the null probability of overlap was calculated by permuting sample labels. Differences in the distributions of overlap probabilities were tested using the Wilcoxon rank-sum test.

The R scripts to perform these calculations and generate **Figure 3-11** are 'NicheNet_calc.R' and `NicheNet_forPaper.Rmd' in the GitHub repository cited above. The R script to generate **Figure 3-12**, **Figure 3-13**, and **Figure 3-14** is 'NicheNet.Rmd' in the GitHub repository cited above. **Figure 3-15** code can be found in '220607_NicheNet_predictions.Rmd'.

Chapter 4
Conclusion

4 Conclusion

This thesis provides a framework for building models of intercellular signaling in tissue from high-throughput single-cell transcriptomic data using current best practices. I outline the analysis workflow required to cluster single-cell sequencing libraries into cell type transcriptomes, describe scClustViz, the software tool I built to facilitate the annotation of these clusters, and summarize the many tools available, including my CCInx software, for predicting ligand-receptor interaction networks from cell type transcriptomes. Finally, I highlight a major barrier to improving ligand-receptor inference methods and suggest a path forward.

4.1 Impact, caveats, and future of scClustViz

The software tool scClustViz was one of the first tools available to generate interactive reports of scRNAseq results, facilitating interpretation of clustering results for study authors. It has empowered my collaborations studying the neural stem cell niche in the developing mouse brain (Yuzwa et al., 2017; Borrett et al., 2020, 2022), generating the human liver cell atlas and assessing liver cells grown *in vitro* (MacParland et al., 2018; Gage et al., 2020), characterizing retinal stem cells (Coles et al., 2021), and identifying senescence as a pathology in traumatic brain injury (Schwab et al., 2022). Perhaps more importantly, it allows the data from published studies to be accessed and interpreted without the need for R coding skills through its functionality as a web-based portal (e.g., http://shiny.baderlab.org/).

The scClustViz paper was also the first to highlight the bias introduced in differential expression magnitude calculations due to the use of pseudocounts during normalization. A recent paper has highlighted the impact of this bias on data interpretation in the context of SARS-Cov-2 (Booeshaghi and Pachter, 2021).

Finally, scClustViz uses a novel method for determining the appropriate number of clusters, motivated by the expectation that distinct cell types should have statistically significant differences in gene expression when compared to their most similar clusters. Unfortunately clusters are defined by differences in gene expression, so testing differences in gene expression on the same data results in uncontrolled false positives. This “double-dipping” error is widespread in scRNAseq analysis, and a few methods have been proposed to allow for statistical testing of differential gene expression between clusters while appropriately controlling the false positive rate. They rely on the concept of selective inference, the notion that to obtain a valid test of the null hypothesis between groups defined by the data being tested, one must condition on the aspect of the data that defines the groups (Lee et al., 2016). The first is the truncated-normal or TN test, where a subset of the data is clustered and a hyperplane defining cluster separation is learned, and then the differential expression test is performed on the remainder of the data labeled using the hyperplane with the null model being a pair of normal distributions truncated by the hyperplane (Zhang et al., 2019a). This requires that the clusters be approximately linearly separable in multidimensional space. Alternative models have been developed specifically for hierarchical clustering (Gao et al., 2022) and K-means clustering (Chen and Witten, 2022).

The scClustViz interface allows the user to define their own groups for differential expression testing, a feature commonly requested to test differences in expression in an individual cell type between case and control. This feature then performs a Wilcoxon rank-sum test between the user-defined groups, which treats each cell as a sample. This results in "pseudoreplication bias", another statistical error widespread in the scRNAseq community (Zimmerman et al., 2021). The solution to this seems relatively simple: properly account for sample number by either aggregating cells from each tested specimen into "pseudobulk" transcriptomes and treat each of these as a sample (Squair et al., 2021; Murphy and Skene, 2022), or account for the source specimen of each cell explicitly in the statistical model, whether a two-part hurdle model (Zimmerman et al., 2021, 2022) or Poisson multilevel model (Svensson, 2019b). Unfortunately, these solutions all require knowledge of which cells came from which specimen. Unless single-cell cDNA libraries are generated from each specimen separately (confounding technical variability with biological variability), or the specimens are sufficiently genetically distinct to deconvolute their origin *in silico*, it is impossible to know which cell came from which specimen. An alternative experimental approach is to use cell hashing, a technique to label cells prior to pooling for library generation using barcoded oligos unique to each sample, but this technique has proven difficult to implement in many sequencing facilities. A recent preprint may have the solution (Lee and Han, 2023). They argue that the loss of type I error control (false positive rate) is not due to pseudoreplication but rather distributional misspecification, specifically that the true distribution may be multimodal due to cell type heterogeneity. They show that using a robust standard error (also known as the sandwich estimator) in a generalized linear model to address this distributional misspecification appropriately controls type I error, and more importantly for the *post hoc* analysis of existing datasets, does not require explicit modeling of donor specimen.

While scClustViz was one of the first graphical user interfaces designed for the interpretation of single-cell RNAseq data, it is no longer unique. Well supported tools such as the open-source CELLxGENE software from the Chan Zuckerberg Initiative have been developed to fill this role. Thus I have no future plans for the development of scClustViz, though I will continue to support its users through GitHub.

4.2 Specificity of transcriptional signatures affects the future of intercellular network inference

Cell-cell interaction prediction from transcriptomics traditionally assumes that if a ligand is expressed in one cell and its cognate receptor is expressed in another, they will interact. This assumption leads to very poor specificity in predictions, as many factors affecting ligand-receptor interaction are overlooked. To improve the specificity of their predictions, the next generation of ligand-receptor interaction inference methods score their predictions using evidence from the receptor-expressing cell's transcriptome. As transcriptional signatures of ligand response are generally not known, they must be inferred from other available data. These methods link published intracellular signaling networks and downstream gene regulatory networks (GRNs) with their predicted intercellular signaling network to infer the expected transcriptional response to predicted ligand stimulation. This assumes that these published networks are consistent between cell types. In Chapter 3 I show that this fundamental assumption of the next generation of ligand-receptor inference methods is invalid. Transcriptional response to ligand perturbation is not consistent across cell types. However, I also show that the transcriptome does at least partially explain the differences in transcriptional response to ligand stimulus between cell types. Thus, if these next-gen methods can refine their multilayer networks to be cell type specific, they should better infer transcriptional responses to their predicted ligands.

Most of these next-gen methods have been benchmarked in a recent study (**Figure 1-12b**, (Liu et al., 2022)). Benchmarking was based on the expected spatial distribution of cell types and their interactions. The evaluation metric assigned ligand-receptor interactions as short- or long-range based on distribution of expressed genes in the spatial transcriptomics datasets used for testing. They then evaluated ligand-receptor interaction predictions for their ability to predict short-range interactions between proximal cell types and long-range interactions between distant cell types. If some of these next-gen methods perform better in benchmarking, it may be thanks to cell type specific refinements of their multilayer networks. The top performing method in this study with an average rank of 2.4 across the various test sets was CellChat v1 (Jin et al., 2021). This method scores predicted interactions by weighting expression of ligand, receptor, and functional cofactors, but does not use the receiving cell's transcriptome to improve specificity. NicheNet (Browaeys et al., 2020) is the top-scoring next-gen method with an average rank of 4.4. To refine its multilayer network, the network was trained on a curated set of published ligand perturbation experiments. This may have improved the accuracy of their ligand response inference in the cell types present in their training data, but does not make their network specific to the cell types a user's data may contain. scMLnet (Cheng et al., 2021) had an average rank of 6.4, and prunes their GRN by requiring significant positive correlation between transcription factors and predicted target genes. This may result in cell type specific GRNs, as it does use expression data specific to the cells of interest. Domino (Cherry et al., 2021) had an average score of 11.2, and uses SCENIC (Aibar et al., 2017) to learn cell type specific GRNs. Finally CellCall (Zhang et al., 2021a) had an average rank of 13.8, and uses curated TF-target gene networks with no refinement. From this benchmark, it doesn't seem like next-gen methods perform particularly well, with or without cell type specific networks. Unfortunately, none of these next-gen methods were included in a complementary benchmarking study (Dimitrov et al., 2022), but the two studies have contradictory findings when using spatial association as the evaluation metric (**Figure 1-12b**). This casts doubt on this metric's value in benchmarking these methods.

Evaluating ligand-receptor interaction predictions is a challenge, as there is very little transcriptomic data available where the interactions are known. But next-gen methods rely on inferring transcriptomic signatures of ligand perturbation to improve their ligand-receptor interaction predictions, so they can be evaluated based on their perturbation response predictions. This is an easier task, as the resources used in Chapter 3 can serve as test datasets. Predicting gene expression is also a more mature field of research, and knowledge from both the machine learning communities (Li et al., 2019; Avsec et al., 2021) and genetic regulation communities

(Huynh-Thu et al., 2010; Ding et al., 2018). I suspect that methods such as CellChat v2 (Hao et al., 2021) and Domino (Cherry et al., 2021) that make an effort to predict cell type specific gene regulatory networks may perform best at predicting ligand perturbation signatures. However, predicting GRNs from transcriptomic data is challenging, with benchmarking studies generally noting that all methods struggle (Holland et al., 2020; Pratapa et al., 2020; Kang et al., 2021). Single-cell chromatin accessibility assays (scATACseq) improve GRN inference by including epigenetic information (Kamimoto et al., 2023). While most users of intercellular interaction inference methods don't have access to multimodal data, it may be possible to use published scATACseq data from their assayed cell types in conjunction with their scRNAseq data.

Predicting transcriptional response to intercellular signaling is not without its own complications, most notably the combinatorial nature of signal processing. Unlike in traditional transcriptomic assays of ligand perturbation, *in vivo* a cell is exposed to multiple ligands at the same time. Together these signals may cause a different response than if presented to a cell individually. The multilayer networks built by next-gen ligand-receptor interaction inference methods actually have an advantage in this context, as they already model all potential interactions. It remains to be seen whether they can appropriately handle combinatorial responses. There is a dataset available to test this, thanks to a recent study of combinatorial signaling amongst the bone morphogenic protein (BMP) ligand family (Klumpe et al., 2022). With combinatorial logic modeled in their multilayer networks and the addition of cell type specific gene regulatory networks, next-gen ligand-receptor inference methods could be a major improvement in the field.

4.3 Single-cell transcriptomics beyond cell types

High-throughput single-cell RNA sequencing was first used to systematically identify the cell types present in a tissue, which raised the question of how to define a cell type. If cell types are defined by the gene products they express, is scRNAseq sufficiently sensitive to distinguish them? My work includes many examples where scRNAseq data identified rare cells and distinguished similar cell types. In our first scRNAseq study of the developing forebrain, we identified the rare and transient Cajal-Retzius cells by their expression of the patterning cue Reelin, and were able to categorize nascent neurons into their cortical layers by their transcriptomes (Yuzwa et al., 2017).

The collection of many transcriptomes in a single assay has renewed interest in the concept of cells as points in transcriptional space, as exemplified in the common tSNE and UMAP projections to visualize scRNAseq data in two dimensions. Rather than defining cell types as discrete categories, they may be regions of transcriptomic space, with individual cells able to transition through the region depending on their state. I explore this concept in our study relating embryonic and adult neural stem cells (**Figure 1-13**), where we identified cell fate and activation axes in transcriptional space and showed that adult NSCs derived from glutamatergic embryonic radial precursors adopt an GABAergic embryonic-like state when activated (Borrett et al., 2020).

What drives cells to make these purposeful moves through transcriptional space to acquire various cell states? Does the comprehensive definition of a cell type include its anatomical position and functional role? These related questions will define the future of single-cell research as we move beyond annotating cell types to placing them in their functional context in a tissue, a goal shared by major research consortia including the Human Cell Atlas and the Multiscale Human (Regev et al., 2018; CIFAR, 2023). To realize this goal, we need to add two major pieces of information beyond cell type. This thesis has already argued the need to identify communication between cell types and their environment and outlined the state of this field. The other addition would be to place cells in their spatial context. Adding spatial information not only has the potential to improve cell-cell interaction prediction but may also help understand the function of both individual cells and their role in a tissue. Spatial transcriptomic technologies like the 10x Genomics Visium (Ståhl et al., 2016), which captures whole transcriptomes at nearly single-cell resolution, and seqFISH (Eng et al., 2019), which can probe for increasingly large subsets of the transcriptome at subcellular resolution, are facilitating this next step towards molecular models of tissue function.

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* 20, 194. doi:10.1186/s13059-019-1795-z.

Acharjee, A., Larkman, J., Xu, Y., Cardoso, V. R., and Gkoutos, G. V. (2020). A random forest based biomarker discovery and power analysis framework for diagnostics research. *BMC Med. Genomics* 13, 178. doi:10.1186/s12920-020-00826-6.

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi:10.1038/nmeth.4463.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). *Molecular biology of the cell*. 5th ed. W.W. Norton & Company doi:10.1201/9780203833445.

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* 17, 137–145. doi:10.1038/s41592-019-0654-x.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106. doi:10.1186/gb-2010-11-10-r106.

Andrews, T. S., and Hemberg, M. (2019). M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* 35, 2865–2867. doi:10.1093/bioinformatics/bty1044.

Anscombe, F. J. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika* 35, 246–254. doi:10.1093/biomet/35.3-4.246.

Armingol, E., Baghdassarian, H. M., Martino, C., Perez-Lopez, A., Aamodt, C., Knight, R., and Lewis, N. E. (2022a). Context-aware deconvolution of cell-cell communication with Tensor-cell2cell. *Nat. Commun.* 13, 3665. doi:10.1038/s41467-022-31369-2.

Armingol, E., Ghaddar, A., Joshi, C. J., Baghdassarian, H., Shamie, I., Chan, J., Her, H.-L., Berhanu, S., Dar, A., Rodriguez-Armstrong, F., et al. (2022b). Inferring a spatial code of cell-cell interactions across a whole animal body. *PLoS Comput. Biol.* 18, e1010715. doi:10.1371/journal.pcbi.1010715.

Armingol, E., Officer, A., Harismendy, O., and Lewis, N. E. (2021). Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* 22, 71–88. doi:10.1038/s41576-020-00292-x.

Arneson, D., Zhang, G., Ying, Z., Zhuang, Y., Byun, H. R., Ahn, I. S., Gomez-Pinilla, F., and Yang, X. (2018). Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nat. Commun.* 9, 3894. doi:10.1038/s41467-018-06222-0.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18, 1196–1203. doi:10.1038/s41592-021-01252-x.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* 41, D991–5. doi:10.1093/nar/gks1193.

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., and Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. *Science* 347, 664–667. doi:10.1126/science.1260793.

Baxevanis, A. D., Bader, G. D., and Wishart, D. S. (2020). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. 4th ed. Hoboken, NJ: Wiley.

Ben-Shlomo, I., Yu Hsu, S., Rauch, R., Kowalski, H. W., and Hsueh, A. J. W. (2003). Signaling receptome: a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Sci STKE* 2003, RE9. doi:10.1126/stke.2003.187.re9.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., and Jensen, L. J. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)* 2014, bau012. doi:10.1093/database/bau012.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008, P10008. doi:10.1088/1742-5468/2008/10/P10008.

Booeshaghi, A. S., and Pachter, L. (2021). Normalization of single-cell RNA-seq counts by log(x + 1)† or log(1 + x)†. *Bioinformatics* 37, 2223–2224. doi:10.1093/bioinformatics/btab085.

Borrett, M. J., Innes, B. T., Jeong, D., Tahmasian, N., Storer, M. A., Bader, G. D., Kaplan, D. R., and Miller, F. D. (2020). Single-Cell Profiling Shows Murine Forebrain Neural Stem Cells Reacquire a Developmental State when Activated for Adult Neurogenesis. *Cell Rep.* 32, 108022. doi:10.1016/j.celrep.2020.108022.

Borrett, M. J., Innes, B. T., Tahmasian, N., Bader, G. D., Kaplan, D. R., and Miller, F. D. (2022). A Shared Transcriptional Identity for Forebrain and Dentate Gyrus Neural Stem Cells from Embryogenesis to Adulthood. *eNeuro* 9. doi:10.1523/ENEURO.0271-21.2021.

Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., and Aerts, S. (2022). SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *BioRxiv*. doi:10.1101/2022.08.19.504505.

Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., Winsor, G. L., Hancock, R. E. W., Brinkman, F. S. L., and Lynn, D. J. (2013). InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Res.* 41, D1228–33. doi:10.1093/nar/gks1147.

Browaeys, R., Saelens, W., and Saeys, Y. (2019). Development, evaluation and application of NicheNet: datasets. *Zenodo*. doi:10.5281/zenodo.3260758.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* 17, 159–162. doi:10.1038/s41592-019-0667-5.

Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *Int. J. Mol. Sci.* 10, 2763–2788. doi:10.3390/ijms10062763.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi:10.1038/nbt.4096.

Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., and Colinge, J. (2020). SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* 48, e55. doi:10.1093/nar/gkaa183.

Camp, J. G., Sekine, K., Gerber, T., Loeffler-Wirth, H., Binder, H., Gac, M., Kanton, S., Kageyama, J., Damm, G., Seehofer, D., et al. (2017). Multilineage communication regulates human liver bud development from pluripotency. *Nature* 546, 533–538. doi:10.1038/nature22796.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* 39, D685–90. doi:10.1093/nar/gkq1039.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. RStudio.

Chasman, D., and Roy, S. (2017). Inference of cell type specific regulatory networks on mammalian lineages. *Current Opinion in Systems Biology* 2, 130–139. doi:10.1016/j.coisb.2017.04.001.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 45, D369–D379. doi:10.1093/nar/gkw1102.

Chen, X., Chen, L., Kürten, C. H. L., Jabbari, F., Vujanovic, L., Ding, Y., Lu, B., Lu, K., Kulkarni, A., Tabib, T., et al. (2022). An individualized causal framework for learning intercellular communication networks that define microenvironments of individual tumors. *PLoS Comput. Biol.* 18, e1010761. doi:10.1371/journal.pcbi.1010761.

Chen, Y. T., and Witten, D. M. (2022). Selective inference for k-means clustering. *arXiv*. doi:10.48550/arxiv.2203.15267.

Chen, Y., Zhang, Y., Yin, Y., Gao, G., Li, S., Jiang, Y., Gu, X., and Luo, J. (2005). SPD--a web-based secreted protein database. *Nucleic Acids Res.* 33, D169–73. doi:10.1093/nar/gki093.

Cheng, J., Zhang, J., Wu, Z., and Sun, X. (2021). Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19. *Brief. Bioinformatics* 22, 988–1005. doi:10.1093/bib/bbaa327.

Cherry, C., Maestas, D. R., Han, J., Andorko, J. I., Cahan, P., Fertig, E. J., Garmire, L. X., and Elisseeff, J. H. (2021). Computational reconstruction of the signalling networks surrounding implanted biomaterials from single-cell transcriptomics. *Nat. Biomed. Eng.* 5, 1228–1238. doi:10.1038/s41551-021-00770-5.

Choi, H., Sheng, J., Gao, D., Li, F., Durrans, A., Ryu, S., Lee, S. B., Narula, N., Rafii, S., Elemento, O., et al. (2015). Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. *Cell Rep.* 10, 1187–1201. doi:10.1016/j.celrep.2015.01.040.

Chong, Z.-S., Ohnishi, S., Yusa, K., and Wright, G. J. (2018). Pooled extracellular receptor-ligand interaction screening using CRISPR activation. *Genome Biol.* 19, 205. doi:10.1186/s13059-018-1581-3.

CIFAR (2023). The Multiscale Human. Available at: https://cifar.ca/research-programs/the-multiscale-human/ [Accessed September 14, 2023].

Clarke, Z. A., Andrews, T. S., Atif, J., Pouyabahar, D., Innes, B. T., MacParland, S. A., and Bader, G. D. (2021). Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat. Protoc.* 16, 2749–2764. doi:10.1038/s41596-021-00534-0.

Clevers, H., Rafelski, S., Elowitz, M., Klein, A., Shendure, J., Trapnell, C., Lein, E., Lundberg, E., Uhlen, M., Martinez-Arias, A., et al. (2017). What is your conceptual definition of "cell type" in the context of a mature organism? *Cell Syst.* 4, 255–259. doi:10.1016/j.cels.2017.03.006.

Cohen, M., Giladi, A., Gorki, A.-D., Solodkin, D. G., Zada, M., Hladik, A., Miklosi, A., Salame, T.-M., Halpern, K. B., David, E., et al. (2018). Lung Single-Cell Signaling Interaction Map Reveals Basophil Role in Macrophage Imprinting. *Cell* 175, 1031–1044.e18. doi:10.1016/j.cell.2018.09.009.

Coles, B. L. K., Labib, M., Poudineh, M., Innes, B. T., Belair-Hickey, J., Gomis, S., Wang, Z., Bader, G. D., Sargent, E. H., Kelley, S. O., et al. (2021). A microfluidic platform enables comprehensive gene expression profiling of mouse retinal stem cells. *Lab Chip* 21, 4464–4476. doi:10.1039/d1lc00790d.

Combes, A. N., Phipson, B., Lawlor, K. T., Dorison, A., Patrick, R., Zappia, L., Harvey, R. P., Oshlack, A., and Little, M. H. (2019). Single cell analysis of the developing mouse kidney provides deeper insight into marker gene expression and ligand-receptor crosstalk. *Development* 146. doi:10.1242/dev.178673.

Cosacak, M. I., Bhattarai, P., Reinhardt, S., Petzold, A., Dahl, A., Zhang, Y., and Kizil, C. (2019). Single-Cell Transcriptomics Analyses of Neural Stem Cell Heterogeneity and Contextual Plasticity in a Zebrafish Brain Model of Amyloid Toxicity. *Cell Rep.* 27, 1307–1318.e3. doi:10.1016/j.celrep.2019.03.090.

Crow, M., Lim, N., Ballouz, S., Pavlidis, P., and Gillis, J. (2019). Predictability of human differential gene expression. *Proc. Natl. Acad. Sci. USA* 116, 6491–6500. doi:10.1073/pnas.1802973116.

Crow, M., Paul, A., Ballouz, S., Huang, Z. J., and Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* 9, 884. doi:10.1038/s41467-018-03282-0.

Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P. L., Nagai, J. S., Boys, C., Ramirez Flores, R. O., Kim, H., Szalai, B., Costa, I. G., et al. (2022). Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat. Commun.* 13, 3224. doi:10.1038/s41467-022-30755-0.

Ding, H., Douglass, E. F., Sonabend, A. M., Mela, A., Bose, S., Gonzalez, C., Canoll, P. D., Sims, P. A., Alvarez, M. J., and Califano, A. (2018). Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat. Commun.* 9, 1471. doi:10.1038/s41467-018-03843-3.

Dumitrascu, B., Villar, S., Mixon, D. G., and Engelhardt, B. E. (2021). Optimal marker gene selection for cell type discrimination in single cell analyses. *Nat. Commun.* 12, 1186. doi:10.1038/s41467-021-21453-4.

Duò, A., Robinson, M. D., and Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* 7, 1141. doi:10.12688/f1000research.15666.3.

Ecker, J. R., Geschwind, D. H., Kriegstein, A. R., Ngai, J., Osten, P., Polioudakis, D., Regev, A., Sestan, N., Wickersham, I. R., and Zeng, H. (2017). The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. *Neuron* 96, 542–557. doi:10.1016/j.neuron.2017.10.007.

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* 15, 1484–1506. doi:10.1038/s41596-020-0292-x.

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 568, 235–239. doi:10.1038/s41586-019-1049-y.

Fagerberg, L., Jonasson, K., von Heijne, G., Uhlén, M., and Berglund, L. (2010). Prediction of the human membrane proteome. *Proteomics* 10, 1141–1149. doi:10.1002/pmic.200900258.

Farbehi, N., Patrick, R., Dorison, A., Xaymardan, M., Janbandhu, V., Wystub-Lis, K., Ho, J. W., Nordon, R. E., and Harvey, R. P. (2019). Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. *Elife* 8. doi:10.7554/eLife.43882.

Fernandez, D. M., Rahman, A. H., Fernandez, N. F., Chudnovskiy, A., Amir, E.-A. D., Amadori, L., Khan, N. S., Wong, C. K., Shamailova, R., Hill, C. A., et al. (2019). Single-cell immune landscape of human atherosclerotic plaques. *Nat. Med.* 25, 1576–1588. doi:10.1038/s41591-019-0590-4.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278. doi:10.1186/s13059-015-0844-5.

Frankenstein, Z., Alon, U., and Cohen, I. R. (2006). The immune-body cytokine network defines a social architecture of cell interactions. *Biol. Direct* 1, 32. doi:10.1186/1745-6150-1-32.

Friedlander, M. P., and Hatz, K. (2008). Computing non-negative tensor factorizations. *Optimization Methods and Software* 23, 631–647. doi:10.1080/10556780801996244.

Fuentealba, L. C., Rompani, S. B., Parraguez, J. I., Obernier, K., Romero, R., Cepko, C. L., and Alvarez-Buylla, A. (2015). Embryonic origin of postnatal neural stem cells. *Cell* 161, 1644–1655. doi:10.1016/j.cell.2015.05.041.

Gage, B. K., Liu, J. C., Innes, B. T., MacParland, S. A., McGilvray, I. D., Bader, G. D., and Keller, G. M. (2020). Generation of Functional Liver Sinusoidal Endothelial Cells from Human Pluripotent Stem-Cell-Derived Venous Angioblasts. *Cell Stem Cell* 27, 254–269.e9. doi:10.1016/j.stem.2020.06.007.

Gao, L. L., Bien, J., and Witten, D. (2022). Selective inference for hierarchical clustering. *J. Am. Stat. Assoc.*, 1–27. doi:10.1080/01621459.2022.2116331.

Garcia, A. D. R., Han, Y.-G., Triplett, J. W., Farmer, W. T., Harwell, C. C., and Ihrie, R. A. (2018). The elegance of sonic hedgehog: emerging novel functions for a classic morphogen. *J. Neurosci.* 38, 9338–9345. doi:10.1523/JNEUROSCI.1662-18.2018.

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* 29, 1363–1375. doi:10.1101/gr.240663.118.

Garnier, S. (2018). *viridis: Default Color Maps from' ' 'matplotlib*. R package.

Gauthier-Fisher, A., and Miller, F. D. (2013). "Environmental Cues and Signaling Pathways that Regulate Neural Precursor Development," in *Patterning and Cell Type Specification in the Developing CNS and PNS* (Elsevier), 355–383. doi:10.1016/B978-0-12-397265-1.00066-6.

Ghoshdastider, U., Rohatgi, N., Mojtabavi Naeini, M., Baruah, P., Revkov, E., Guo, Y. A., Rizzetto, S., Wong, A. M. L., Solai, S., Nguyen, T. T., et al. (2021). Pan-Cancer Analysis of Ligand-Receptor Cross-talk in the Tumor Microenvironment. *Cancer Res.* 81, 1802–1812. doi:10.1158/0008-5472.CAN-20-2352.

Graeber, T. G., and Eisenberg, D. (2001). Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet.* 29, 295–300. doi:10.1038/ng755.

Griffiths, J. I., Cosgrove, P. A., Castaneda, E. M., Nath, A., Chen, J., Adler, F. R., Chang, J. T., Khan, Q. J., and Bild, A. H. (2022). Cancer cells communicate with macrophages to prevent T cell activation during development of cell cycle therapy resistance. *BioRxiv*. doi:10.1101/2022.09.14.507931.

Gu, W., Ni, Z., Tan, Y.-Q., Deng, J., Zhang, S.-J., Lv, Z.-C., Wang, X.-J., Chen, T., Zhang, Z., Hu, Y., et al. (2019). Adventitial Cell Atlas of wt (Wild Type) and ApoE (Apolipoprotein E)-Deficient Mice Defined by Single-Cell RNA Sequencing. *Arterioscler. Thromb. Vasc. Biol.* 39, 1055–1071. doi:10.1161/ATVBAHA.119.312399.

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., et al. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* 5, 11432. doi:10.1038/srep11432.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 172, 1091–1107.e17. doi:10.1016/j.cell.2018.02.001.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14, 7. doi:10.1186/1471-2105-14-7.

Hao, M., Zou, X., and Jin, S. (2021). Identification of Intercellular Signaling Changes Across Conditions and Their Influence on Intracellular Signaling Response From Multiple Single-Cell Datasets. *Front. Genet.* 12, 751158. doi:10.3389/fgene.2021.751158.

Harding, S. D., Armstrong, J. F., Faccenda, E., Southan, C., Alexander, S. P. H., Davenport, A. P., Pawson, A. J., Spedding, M., Davies, J. A., and NC-IUPHAR (2022). The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. *Nucleic Acids Res.* 50, D1282–D1294. doi:10.1093/nar/gkab1010.

Harmar, A. J., Hills, R. A., Rosser, E. M., Jones, M., Buneman, O. P., Dunbar, D. R., Greenhill, S. D., Hale, V. A., Sharman, J. L., Bonner, T. I., et al. (2009). IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.* 37, D680–5. doi:10.1093/nar/gkn728.

Holland, C. H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M. P., Mereu, E., Joughin, B. A., Stegle, O., Lauffenburger, D. A., Heyn, H., et al. (2020). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* 21, 36. doi:10.1186/s13059-020-1949-z.

Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A., and Forrest, A. R. R. (2020). Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* 11, 5011. doi:10.1038/s41467-020-18873-z.

Hu, Y., Peng, T., Gao, L., and Tan, K. (2021). CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Sci. Adv.* 7. doi:10.1126/sciadv.abf1356.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5. doi:10.1371/journal.pone.0012776.

Ibelgaufts, H. (1997). Cell Communication Encyclopedia (Horst Ibelgaufts' COPE). Available at: http://www.copewithcytokines.org/cope.cgi [Accessed August 6, 2021].

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17, 29. doi:10.1186/s13059-016-0888-1.

Innes, B. T., and Bader, G. D. (2018). scClustViz - Single-cell RNAseq cluster assessment and visualization. [version 2; peer review: 2 approved]. *F1000Res.* 7. doi:10.12688/f1000research.16198.2.

Innes, B. T., and Bader, G. D. (2021). Transcriptional signatures of cell-cell interactions are dependent on cellular context. *BioRxiv*. doi:10.1101/2021.09.06.459134.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11, 163–166. doi:10.1038/nmeth.2772.

Isserlin, R., Voisin, V., Ailles, L., and Bader, G. D. (2020). Cell-Cell Interaction Database. *Zenodo*. doi:10.5281/zenodo.7589953.

Jakobsson, J. E. T., Spjuth, O., and Lagerström, M. C. (2021). scConnect: a method for exploratory analysis of cell-cell communication based on single-cell RNA-sequencing data. *Bioinformatics* 37, 3501–3508. doi:10.1093/bioinformatics/btab245.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi:10.1093/nar/gkz1031.

Jiang, P., Zhang, Y., Ru, B., Yang, Y., Vu, T., Paul, R., Mirza, A., Altan-Bonnet, G., Liu, L., Ruppin, E., et al. (2021). Systematic investigation of cytokine signaling activity at the tissue and single-cell levels. *Nat. Methods* 18, 1181–1191. doi:10.1038/s41592-021-01274-5.

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M. V., and Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* 12, 1088. doi:10.1038/s41467-021-21246-9.

Jin, Z., Zhang, X., Dai, X., Huang, J., Hu, X., Zhang, J., and Shi, L. (2022). InterCellDB: A User-Defined Database for Inferring Intercellular Networks. *Adv Sci (Weinh)* 9, e2200045. doi:10.1002/advs.202200045.

Jung, S., Singh, K., and del Sol, A. (2021). FunRes: resolving tissue-specific functional cell states based on a cell–cell communication network model. *Brief. Bioinformatics* 22. doi:10.1093/bib/bbaa283.

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature* 614, 742–751. doi:10.1038/s41586-022-05688-9.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–62. doi:10.1093/nar/gkv1070.

Kaneko, N., Kurata, M., Yamamoto, T., Morikawa, S., and Masumoto, J. (2019). The role of interleukin-1 in general pathology. *Inflamm. Regen.* 39, 12. doi:10.1186/s41232-019-0101-5.

Kang, Y., Thieffry, D., and Cantini, L. (2021). Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms. *Front. Genet.* 12, 617282. doi:10.3389/fgene.2021.617282.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Res.* 37, D767–72. doi:10.1093/nar/gkn892.

Kirouac, D. C., Ito, C., Csaszar, E., Roch, A., Yu, M., Sykes, E. A., Bader, G. D., and Zandstra, P. W. (2010). Dynamic interaction networks in a hierarchically organized tissue. *Mol. Syst. Biol.* 6, 417. doi:10.1038/msb.2010.71.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486. doi:10.1038/nmeth.4236.

Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* 15, 359–362. doi:10.1038/nmeth.4644.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74. doi:10.1038/nmeth.1778.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi:10.1016/j.cell.2015.04.044.

Klumpe, H. E., Langley, M. A., Linton, J. M., Su, C. J., Antebi, Y. E., and Elowitz, M. B. (2022). The context-dependent, combinatorial logic of BMP signaling. *Cell Syst.* 13, 388–407.e10. doi:10.1016/j.cels.2022.03.002.

Kumar, M. P., Du, J., Lagoudas, G., Jiao, Y., Sawyer, A., Drummond, D. C., Lauffenburger, D. A., and Raue, A. (2018). Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Rep.* 25, 1458–1468.e4. doi:10.1016/j.celrep.2018.10.047.

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics* 15, 8. doi:10.1186/1471-2105-15-8.

Lagger, C., Ursu, E., Equey, A., Avelar, R. A., Pisco, A. O., Tacutu, R., and de Magalhaes, J. P. (2021). scAgeCom: a murine atlas of age-related changes in intercellular communication inferred with the package scDiffCom. *BioRxiv*. doi:10.1101/2021.08.13.456238.

Lambiotte, R., Delvenne, J. C., and Barahona, M. (2008). Laplacian Dynamics and Multiscale Modular Structure in Networks. *arXiv*. doi:10.48550/arxiv.0812.1770.

Larsen, S. J., Röttger, R., Schmidt, H. H. H. W., and Baumbach, J. (2019). E. coli gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Res.* 47, 85–92. doi:10.1093/nar/gky1176.

Lee, H., and Han, B. (2023). Why do cell-level test methods for detecting differentially expressed genes fail in single-cell RNA-seq data? *BioRxiv*. doi:10.1101/2023.01.09.523212.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* 44, 907–927. doi:10.1214/15-AOS1371.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., et al. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 162, 184–197. doi:10.1016/j.cell.2015.05.047.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). Upset: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* 20, 1983–1992. doi:10.1109/TVCG.2014.2346248.

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., Kong, S. L., Chua, C., Hon, L. K., Tan, W. S., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* 49, 708–718. doi:10.1038/ng.3818.

Li, J. J., Bickel, P. J., and Biggin, M. D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2, e270. doi:10.7717/peerj.270.

Li, W., Yin, Y., Quan, X., and Zhang, H. (2019). Gene expression value prediction based on xgboost algorithm. *Front. Genet.* 10, 1077. doi:10.3389/fgene.2019.01077.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi:10.1016/j.cels.2015.12.004.

Liu, Z., Sun, D., and Wang, C. (2022). Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. *Genome Biol.* 23, 218. doi:10.1186/s13059-022-02783-y.

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016a). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17, 75. doi:10.1186/s13059-016-0947-7.

Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. [version 2; peer review: 3 approved, 2 approved with reservations]. *F1000Res.* 5, 2122. doi:10.12688/f1000research.9501.2.

Lun, A. T., and Risso, D. (2017). *SingleCellExperiment: S4 Classes for Single Cell Data*. R package.

Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202–1214. doi:10.1016/j.cell.2015.05.002.

MacParland, S. A., Liu, J. C., Ma, X.-Z., Innes, B. T., Bartczak, A. M., Gage, B. K., Manuel, J., Khuu, N., Echeverri, J., Linares, I., et al. (2018). Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* 9, 4383. doi:10.1038/s41467-018-06318-7.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2018). *cluster: Cluster Analysis Basics and Extensions*. R package.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583, 3966–3973. doi:10.1016/j.febslet.2009.10.036.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. *Nature* 560, 494–498. doi:10.1038/s41586-018-0414-6.

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., and Califano, A. (2006). Reverse engineering cellular networks. *Nat. Protoc.* 1, 662–671. doi:10.1038/nprot.2006.106.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C., Chou, A., Ienasescu, H., et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142–7. doi:10.1093/nar/gkt997.

McGinnis, C. S., Murrow, L. M., and Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* 8, 329–337.e4. doi:10.1016/j.cels.2019.03.003.

McInnes, L., and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*.

Murphy, A. E., and Skene, N. G. (2022). A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat. Commun.* 13, 7851. doi:10.1038/s41467-022-35519-4.

Muskovic, W., and Powell, J. E. (2021). DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. *Genome Biol.* 22, 329. doi:10.1186/s13059-021-02547-0.

Nagai, J. S., Leimkühler, N. B., Schaub, M. T., Schneider, R. K., and Costa, I. G. (2021). CrossTalkeR: analysis and visualization of ligand-receptorne tworks. *Bioinformatics* 37, 4263–4265. doi:10.1093/bioinformatics/btab370.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package.

Noël, F., Massenet-Regad, L., Carmi-Levy, I., Cappuccio, A., Grandclaudon, M., Trichot, C., Kieffer, Y., Mechta-Grigoriou, F., and Soumelis, V. (2021). Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat. Commun.* 12, 1089. doi:10.1038/s41467-021-21244-x.

Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L., and Tse, D. N. (2016). Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.* 17, 112. doi:10.1186/s13059-016-0970-8.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F. S. L., Brinkman, F., et al. (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* 9, 345–350. doi:10.1038/nmeth.1931.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* 30, 187–200. doi:10.1002/pro.3978.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2, e190. doi:10.1371/journal.pgen.0020190.

Pavličev, M., Wagner, G. P., Chavan, A. R., Owens, K., Maziarz, J., Dunn-Fletcher, C., Kallapur, S. G., Muglia, L., and Jones, H. (2017). Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface. *Genome Res.* 27, 349–361. doi:10.1101/gr.207597.116.

Pesaresi, M., Sebastian-Perez, R., and Cosma, M. P. (2019). Dedifferentiation, transdifferentiation and cell fusion: in vivo reprogramming strategies for regenerative medicine. *FEBS J.* 286, 1074–1093. doi:10.1111/febs.14633.

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17, 147–154. doi:10.1038/s41592-019-0690-6.

Qiao, W., Wang, W., Laurenti, E., Turinsky, A. L., Wodak, S. J., Bader, G. D., Dick, J. E., and Zandstra, P. W. (2014). Intercellular network structure and regulatory motifs in the human hematopoietic system. *Mol. Syst. Biol.* 10, 741. doi:10.15252/msb.20145141.

Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., and Crowley, T. M. (2019). A field guide for the compositional analysis of any-omics data. *Gigascience* 8. doi:10.1093/gigascience/giz107.

R Core Team (2018). *R: A Language and Environment for Statistical' ' Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramilowski, J. A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V. P., Itoh, M., Kawaji, H., Carninci, P., Rost, B., et al. (2015). A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat. Commun.* 6, 7866. doi:10.1038/ncomms8866.

Ranjan, B., Sun, W., Park, J., Mishra, K., Schmidt, F., Xie, R., Alipour, F., Singhal, V., Joanito, I., Honardoost, M. A., et al. (2021). DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nat. Commun.* 12, 5849. doi:10.1038/s41467-021-26085-2.

Raredon, M. S. B., Yang, J., Garritano, J., Wang, M., Kushnir, D., Schupp, J. C., Adams, T. S., Greaney, A. M., Leiby, K. L., Kaminski, N., et al. (2022). Computation and visualization of cell-cell signaling topologies in single-cell systems data using Connectome. *Sci. Rep.* 12, 4187. doi:10.1038/s41598-022-07959-x.

Raredon, M. S. B., Yang, J., Kothapalli, N., Lewis, W., Kaminski, N., Niklason, L. E., and Kluger, Y. (2023). Comprehensive visualization of cell-cell interactions in single-cell and spatial transcriptomics with NICHES. *Bioinformatics* 39. doi:10.1093/bioinformatics/btac775.

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* 9, 405. doi:10.1186/1471-2105-9-405.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell atlas. *Elife* 6. doi:10.7554/eLife.27041.

Regev, A., Teichmann, S., Rozenblatt-Rosen, O., Stubbington, M., Ardlie, K., Amit, I., Arlotta, P., Bader, G., Benoist, C., Biton, M., et al. (2018). The Human Cell Atlas White Paper. *arXiv*. doi:10.48550/arxiv.1810.05192.

Ren, X., Zhong, G., Zhang, Q., Zhang, L., Sun, Y., and Zhang, Z. (2020). Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. *Cell Res.* 30, 763–778. doi:10.1038/s41422-020-0353-2.

Rieckmann, J. C., Geiger, R., Hornburg, D., Wolf, T., Kveler, K., Jarrossay, D., Sallusto, F., Shen-Orr, S. S., Lanzavecchia, A., Mann, M., et al. (2017). Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat. Immunol.* 18, 583–593. doi:10.1038/ni.3693.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018a). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9, 284. doi:10.1038/s41467-017-02554-5.

Risso, D., Purvis, L., Fletcher, R. B., Das, D., Ngai, J., Dudoit, S., and Purdom, E. (2018b). clusterExperiment and RSEC: A Bioconductor package and framework for clustering of single-cell and other large gene expression datasets. *PLoS Comput. Biol.* 14, e1006378. doi:10.1371/journal.pcbi.1006378.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25. doi:10.1186/gb-2010-11-3-r25.

Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., Franz, M., Siper, M. C., Cheung, M., Wrana, M., et al. (2020). Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.* 48, D489–D497. doi:10.1093/nar/gkz946.

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182. doi:10.1126/science.aam8999.

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)* 2016. doi:10.1093/database/baw100.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65. doi:10.1016/0377-0427(87)90125-7.

Rue-Albrecht, K., Marini, F., Soneson, C., and Lun, A. T. L. (2018). iSEE: Interactive SummarizedExperiment Explorer. [version 1; peer review: 3 approved]. *F1000Res.* 7, 741. doi:10.12688/f1000research.14966.1.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502. doi:10.1038/nbt.3192.

Saunders, A., Macosko, E., Wysoker, A., Goldman, M., Krienen, F., Bien, E., Baum, M., Wang, S., Goeva, A., Nemesh, J., et al. (2018). A Single-Cell Atlas of Cell Types, States, and Other Transcriptional Patterns from Nine Regions of the Adult Mouse Brain. *BioRxiv*. doi:10.1101/299081.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* 176, 928–943.e22. doi:10.1016/j.cell.2019.01.006.

Schwab, N., Taskina, D., Leung, E., Innes, B. T., Bader, G. D., and Hazrati, L.-N. (2022). Neurons and glial cells acquire a senescent signature after repeated mild traumatic brain injury in a sex-dependent manner. *Front. Neurosci.* 16, 1027116. doi:10.3389/fnins.2022.1027116.

Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., Marioni, J. C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* 85, 54–61. doi:10.1016/j.ymeth.2015.06.021.

Shao, C., and Höfer, T. (2017). Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* 33, 235–242. doi:10.1093/bioinformatics/btw607.

Shao, X., Liao, J., Li, C., Lu, X., Cheng, J., and Fan, X. (2021). CellTalkDB: a manually curated database of ligand-receptor interactions in humans and mice. *Brief. Bioinformatics* 22. doi:10.1093/bib/bbaa269.

Shao, X., Lu, X., Liao, J., Chen, H., and Fan, X. (2020). New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. *Protein Cell* 11, 866–880. doi:10.1007/s13238-020-00727-5.

Skelly, D. A., Squiers, G. T., McLellan, M. A., Bolisetty, M. T., Robson, P., Rosenthal, N. A., and Pinto, A. R. (2018). Single-Cell Transcriptional Profiling Reveals Cellular Diversity and Intercommunication in the Mouse Heart. *Cell Rep.* 22, 600–610. doi:10.1016/j.celrep.2017.12.072.

Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., Herbst, R. H., Rogel, N., Slyper, M., Waldman, J., et al. (2019). Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* 178, 714–730.e22. doi:10.1016/j.cell.2019.06.029.

Snow, G. (2016). *TeachingDemos: Demonstrations for Teaching and Learning*. R package.

Soneson, C., and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15, 255–261. doi:10.1038/nmeth.4612.

Song, Q., Hawkins, G. A., Wudel, L., Chou, P.-C., Forbes, E., Pullikuth, A. K., Liu, L., Jin, G., Craddock, L., Topaloglu, U., et al. (2019). Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. *Cancer Med* 8, 3072–3085. doi:10.1002/cam4.2113.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. *Nat. Commun.* 12, 5692. doi:10.1038/s41467-021-25960-2.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82. doi:10.1126/science.aaf2403.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888–1902.e21. doi:10.1016/j.cell.2019.05.031.

Su, K., Yu, T., and Wu, H. (2021). Accurate feature selection improves single-cell RNA-seq cell clustering. *Brief. Bioinformatics* 22. doi:10.1093/bib/bbab034.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452.e17. doi:10.1016/j.cell.2017.10.049.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102, 15545–15550. doi:10.1073/pnas.0506580102.

Sun, S., Zhu, J., Ma, Y., and Zhou, X. (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol.* 20, 269. doi:10.1186/s13059-019-1898-6.

Svensson, V. (2019a). Droplet scRNA-seq is not zero-inflated. *BioRxiv*. doi:10.1101/582064.

Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* 38, 147–150. doi:10.1038/s41587-019-0379-5.

Svensson, V. (2019b). Handling confounded samples for differential expression in scRNA-seq experiments. *What do you mean "heterogeneity"?* Available at: https://www.nxn.se/valent/2019/2/15/handling-confounded-samples-for-differential-expression-in-scrna-seq-experiments [Accessed February 11, 2023].

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., and Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387. doi:10.1038/nmeth.4220.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–52. doi:10.1093/nar/gku1003.

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators

(2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562, 367–372. doi:10.1038/s41586-018-0590-4.

Tabula Sapiens Consortium, Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., et al. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* 376, eabl4896. doi:10.1126/science.abl4896.

Tang, M., Kaymaz, Y., Logeman, B. L., Eichhorn, S., Liang, Z. S., Dulac, C., and Sackton, T. B. (2021). Evaluating single-cell cluster stability using the Jaccard similarity index. *Bioinformatics* 37, 2212–2214. doi:10.1093/bioinformatics/btaa956.

Taylor, S. R., Santpere, G., Weinreb, A., Barrett, A., Reilly, M. B., Xu, C., Varol, E., Oikonomou, P., Glenwinkel, L., McWhirter, R., et al. (2021). Molecular topography of an entire nervous system. *Cell* 184, 4329–4347.e23. doi:10.1016/j.cell.2021.06.023.

Tjalsma, H., Antelmann, H., Jongbloed, J. D. H., Braun, P. G., Darmon, E., Dorenbos, R., Dubois, J.-Y. F., Westers, H., Zanen, G., Quax, W. J., et al. (2004). Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome. *Microbiol. Mol. Biol. Rev.* 68, 207–233. doi:10.1128/MMBR.68.2.207-233.2004.

Tracy, C. A., and Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Commun.Math. Phys.* 159, 151–174. doi:10.1007/BF02100489.

Tsuyuzaki, K., Ishii, M., and Nikaido, I. (2019). Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data. *BioRxiv*. doi:10.1101/566182.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13, 966–967. doi:10.1038/nmeth.4077.

Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., et al. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* 17. doi:10.15252/msb.20209923.

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)* 2010, baq023. doi:10.1093/database/baq023.

Tyler, S. R., Rotti, P. G., Sun, X., Yi, Y., Xie, W., Winter, M. C., Flamme-Wiese, M. J., Tucker, B. A., Mullins, R. F., Norris, A. W., et al. (2019). PyMINEr Finds Gene and Autocrine-Paracrine Networks from Human Islet scRNA-Seq. *Cell Rep.* 26, 1951–1964.e8. doi:10.1016/j.celrep.2019.01.063.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347, 1260419. doi:10.1126/science.1260419.

UniProt Consortium (2023). Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* 51, D523–D531. doi:10.1093/nar/gkac1052.

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J.-E., Stephenson, E., Polański, K., Goncalves, A., et al. (2018). Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353. doi:10.1038/s41586-018-0698-6.

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* 10, 4667. doi:10.1038/s41467-019-12266-7.

Voet, D., and Voet, J. G. (2010). *Biochemistry, 4th Edition*. 4th ed. Hoboken, NJ: Wiley.

Vogel, C., and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232. doi:10.1038/nrg3185.

Waddington, C. H. (1957). *The strategy of the genes*. Routledge doi:10.4324/9781315765471.

Wang, S., Karikomi, M., MacLean, A. L., and Nie, Q. (2019a). Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res.* 47, e66. doi:10.1093/nar/gkz204.

Wang, X., Li, F., Xu, J., Rong, J., Webb, G. I., Ge, Z., Li, J., and Song, J. (2022). ASPIRER: a new computational approach for identifying non-classical secreted proteins based on deep learning. *Brief. Bioinformatics* 23. doi:10.1093/bib/bbac031.

Wang, Y., Wang, R., Zhang, S., Song, S., Jiang, C., Han, G., Wang, M., Ajani, J., Futreal, A., and Wang, L. (2019b). iTALK: an R Package to Characterize and Illustrate Intercellular Communication. *BioRxiv*. doi:10.1101/507871.

Wickham, H. (2017). *scales: Scale Functions for Visualization*. R package.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* 1, 80. doi:10.2307/3001968.

Wright, M. N., and Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in*C*++ and*R*. *J. Stat. Softw.* 77, 1–17. doi:10.18637/jss.v077.i01.

Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J. R., Bartonicek, N., et al. (2021). A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* 53, 1334–1347. doi:10.1038/s41588-021-00911-1.

Ximerakis, M., Lipnick, S. L., Innes, B. T., Simmons, S. K., Adiconis, X., Dionne, D., Mayweather, B. A., Nguyen, L., Niziolek, Z., Ozek, C., et al. (2019). Single-cell transcriptomic profiling of the aging mouse brain. *Nat. Neurosci.* 22, 1696–1708. doi:10.1038/s41593-019-0491-3.

Xiong, X., Kuang, H., Ansari, S., Liu, T., Gong, J., Wang, S., Zhao, X.-Y., Ji, Y., Li, C., Guo, L., et al. (2019). Landscape of Intercellular Crosstalk in Healthy and NASH Liver Revealed by Single-Cell Secretome Gene Analysis. *Mol. Cell* 75, 644–660.e5. doi:10.1016/j.molcel.2019.07.028.

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi:10.1093/bioinformatics/btv088.

Xu, F., Wang, S., Dai, X., Mundra, P. A., and Zheng, J. (2021). Ensemble learning models that predict surface protein abundance from single-cell multimodal omics data. *Methods* 189, 65–73. doi:10.1016/j.ymeth.2020.10.001.

Xu, J., Falconer, C., Nguyen, Q., Crawford, J., McKinnon, B. D., Mortlock, S., Senabouth, A., Andersen, S., Chiu, H. S., Jiang, L., et al. (2019). Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol.* 20, 290. doi:10.1186/s13059-019-1852-7.

Young, M. D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* 9, giaa151. doi:10.1093/gigascience/giaa151.

Yu, A., Li, Y., Li, I., Ozawa, M. G., Yeh, C., Chiou, A. E., Trope, W. L., Taylor, J., Shrager, J., and Plevritis, S. K. (2022). Reconstructing codependent cellular cross-talk in lung adenocarcinoma using REMI. *Sci. Adv.* 8, eabi4757. doi:10.1126/sciadv.abi4757.

Yuan, D., Tao, Y., Chen, G., and Shi, T. (2019). Systematic expression analysis of ligand-receptor pairs reveals important cell-to-cell interactions inside glioma. *Cell Commun. Signal.* 17, 48. doi:10.1186/s12964-019-0363-1.

Yuan, Y., Cosme, C., Adams, T. S., Schupp, J., Sakamoto, K., Xylourgidis, N., Ruffalo, M., Li, J., Kaminski, N., and Bar-Joseph, Z. (2022). CINS: Cell Interaction Network inference from Single cell expression data. *PLoS Comput. Biol.* 18, e1010468. doi:10.1371/journal.pcbi.1010468.

Yuzwa, S. A., Borrett, M. J., Innes, B. T., Voronova, A., Ketela, T., Kaplan, D. R., Bader, G. D., and Miller, F. D. (2017). Developmental Emergence of Adult Neural Stem Cells as Revealed by Single-Cell Transcriptional Profiling. *Cell Rep.* 21, 3970–3986. doi:10.1016/j.celrep.2017.12.017.

Yuzwa, S. A., Yang, G., Borrett, M. J., Clarke, G., Cancino, G. I., Zahr, S. K., Zandstra, P. W., Kaplan, D. R., and Miller, F. D. (2016). Proneurogenic ligands defined by modeling developing cortex growth factor communication networks. *Neuron* 91, 988–1004. doi:10.1016/j.neuron.2016.07.037.

Zahr, S. K., Yang, G., Kazan, H., Borrett, M. J., Yuzwa, S. A., Voronova, A., Kaplan, D. R., and Miller, F. D. (2018). A Translational Repression Complex in Developing Mammalian Neural Stem Cells that Regulates Neuronal Specification. *Neuron* 97, 520–537.e6. doi:10.1016/j.neuron.2017.12.045.

Zappia, L., and Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* 7, giy083. doi:10.1093/gigascience/giy083.

Zhang, J., Guan, M., Wang, Q., Zhang, J., Zhou, T., and Sun, X. (2020). Single-cell transcriptome-based multilayer network biomarker for predicting prognosis and therapeutic response of gliomas. *Brief. Bioinformatics* 21, 1080–1097. doi:10.1093/bib/bbz040.

Zhang, J. M., Kamath, G. M., and Tse, D. N. (2019a). Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq. *Cell Syst.* 9, 383–392.e6. doi:10.1016/j.cels.2019.07.012.

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., and Wang, J. (2019b). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol. Cell* 73, 130–142.e5. doi:10.1016/j.molcel.2018.10.020.

Zhang, Y., Liu, T., Hu, X., Wang, M., Wang, J., Zou, B., Tan, P., Cui, T., Dou, Y., Ning, L., et al. (2021a). CellCall: integrating paired ligand-receptor and transcription factor activities for cell-cell communication. *Nucleic Acids Res.* 49, 8520–8534. doi:10.1093/nar/gkab638.

Zhang, Y., Liu, T., Wang, J., Zou, B., Li, L., Yao, L., Chen, K., Ning, L., Wu, B., Zhao, X., et al. (2021b). Cellinker: a platform of ligand-receptor interactions for intercellular communication analysis. *Bioinformatics*. doi:10.1093/bioinformatics/btab036.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049. doi:10.1038/ncomms14049.

Zhou, J. X., Taramelli, R., Pedrini, E., Knijnenburg, T., and Huang, S. (2017). Extracting Intercellular Signaling Network of Cancer Tissues using Ligand-Receptor Expression Patterns from Whole-tumor and Single-cell Transcriptomes. *Sci. Rep.* 7, 8815. doi:10.1038/s41598-017-09307-w.

Zhu, X., Wolfgruber, T. K., Tasato, A., Arisdakessian, C., Garmire, D. G., and Garmire, L. X. (2017). Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.* 9, 108. doi:10.1186/s13073-017-0492-3.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* 65, 631–643.e4. doi:10.1016/j.molcel.2017.01.023.

Zimmerman, K. D., Espeland, M. A., and Langefeld, C. D. (2021). A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* 12, 738. doi:10.1038/s41467-021-21038-1.

Zimmerman, K. D., Evans, C., and Langefeld, C. D. (2022). Reply to: A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat. Commun.* 13, 7852. doi:10.1038/s41467-022-35520-x.

Žurauskienė, J., and Yau, C. (2016). pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 17, 140. doi:10.1186/s12859-016-0984-y.

Appendices

Appendix A. A comprehensive listing of novel methods and resources for cell-cell interaction prediction from transcriptomic data. Methods that rely on additional spatial context are omitted. Relevant historical non-transcriptomic efforts have been included.

	Paper	Ligand-Receptor Source	Inference Method
1.	Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles (Graeber and Eisenberg, 2001)	Curated from literature to create DLRP.	Correlation of ligand & cognate receptor expression from microarray data to infer autocrine signaling.
2.	Dynamic interaction networks in a hierarchically organized tissue (Kirouac et al., 2010)	Curated mostly from COPE (Ibelgaufts, 1997)	Ligand & receptor expression above threshold in microarray experiments of purified cell types.
3.	Intercellular network structure and regulatory motifs in the human hematopoietic system (Qiao et al., 2014)	GO terms & manual curation for ligands, receptors. iRefWeb (Turner et al., 2010) for interactions.	Ligand & receptor expression above threshold in microarray experiments of purified cell types and deconvolved bulk RNAseq.
4.	Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model (Choi et al., 2015)	DLRP (Graeber and Eisenberg, 2001); HPRD (Keshava Prasad et al., 2009); KEGG (Kanehisa et al., 2016)	Ligand & receptor expression above threshold in bulk RNAseq experiments of purified cell types.
5.	A draft network of ligand–receptor-mediated multicellular signalling in human (Ramilowski et al., 2015) *Ligand-Receptor pairs from this paper are extensively referenced, referred to as the FANTOM5 database here.*	IUPHAR (Harding et al., 2022); HPMR (Ben-Shlomo et al., 2003); DLRP; curation for ligands, receptors. STRING (Szklarczyk et al., 2015); HPRD; curation for interactions.	Ligand & receptor expression above threshold in CAGE analysis of purified cell types (FANTOM5 dataset).
6.	Proneurogenic ligands defined by modeling developing cortex growth factor communication networks (Yuzwa et al., 2016)	GO terms & manual curation for ligands, receptors. iRefWeb for interactions.	Ligand & receptor expression above threshold in microarray experiments of purified cell types and deconvolved bulk. Cell surface proteomics for validation of receptor expression.
7.	Social network architecture of human immune cells unveiled by quantitative proteomics (Rieckmann et al., 2017)	Secretome from experiments + STRING.	Proteomics from purified cell types.

8. Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface (Pavličev et al., 2017)	DLRP; IUPHAR	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
9. Extracting Intercellular Signaling Network of Cancer Tissues using Ligand-Receptor Expression Patterns from Whole-tumor and Single-cell Transcriptomes (Zhou et al., 2017)	FANTOM5	Coexpression of ligand & receptor in TCGA samples. Ligand & receptor expression above threshold in single-cell RNAseq experiments.
10. Multilineage communication regulates human liver bud development from pluripotency (Camp et al., 2017)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
11. Single-Cell Transcriptional Profiling Reveals Cellular Diversity and Intercommunication in the Mouse Heart (Skelly et al., 2018)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
12. Lung Single-Cell Signaling Interaction Map Reveals Basophil Role in Macrophage Imprinting (Cohen et al., 2018)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
13. Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics (Kumar et al., 2018)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Scored by expression.
14. Single cell molecular alterations reveal target cells and pathways of concussive brain injury (Arneson et al., 2018)	UniProt "secreted" peptides as ligands.	Interaction score calculated by summing correlations between ligand and all genes (across cell types). Significance determined by permutation testing.
15. Single-cell transcriptomic profiling of the aging mouse brain (Ximerakis et al., 2019)	GO, Uniprot, HPA (Uhlén et al., 2015) for ligands, receptors. iRefIndex (Razick et al., 2008), Pathway Commons (Cerami et al., 2011), BioGRID(Chatr-Aryamontri et al., 2017) for interactions. (Isserlin et al., 2020)	Ligand & receptor nodes scored by differential expression magnitude between conditions. Edges ranked by summed node scores scaled per cell type. Avoids arbitrary thresholding.
16. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming (Schiebinger et al., 2019)	GO terms for ligands, receptors. Protein-protein interaction databases for links. Curated.	Ligand & receptor interaction scored by positive coexpression in single-cell RNAseq experiments.
17. Landscape of Intercellular Crosstalk in Healthy and NASH Liver Revealed by Single-Cell Secretome Gene Analysis (Xiong et al., 2019)	SPD (Chen et al., 2005)	Ligand & receptor expression above threshold in single-cell RNAseq experiments.

18. iTALK: an R Package to Characterize and Illustrate Intercellular Communication (Wang et al., 2019b)	DLRP ; IUPHAR; FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
19. Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq (Song et al., 2019)	IUPHAR; FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
20. Single-Cell Transcriptomics Analyses of Neural Stem Cell Heterogeneity and Contextual Plasticity in a Zebrafish Brain Model of Amyloid Toxicity (Cosacak et al., 2019)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
21. Adventitial Cell Atlas of wt (Wild Type) and ApoE (Apolipoprotein E)-Deficient Mice Defined by Single-Cell RNA Sequencing (Gu et al., 2019)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
22. Systematic expression analysis of ligand-receptor pairs reveals important cell-to-cell interactions inside glioma (Yuan et al., 2019)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Filtered for ligand-receptor pairs coexpressed in corresponding TCGA samples.
23. Single cell analysis of the developing mouse kidney provides deeper insight into marker gene expression and ligand-receptor crosstalk (Combes et al., 2019)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Networked weighted by gene expression, statistical model to find highly expressed interactions.
24. Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury (Farbehi et al., 2019)	FANTOM5	Ligand & receptor expression weighted by DE in cluster vs all cells. Cell-cell interactions are summed ligand-receptor edge weights, significance testing by permutation of ligand-receptor edges.
25. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis (Smillie et al., 2019)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Statistical model of cell-cell communication from number of ligand-receptor interactions per cell type pair.
26. Single-cell reconstruction of the early maternal-fetal interface in humans (Vento-Tormo et al., 2018); CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes (Efremova et al., 2020)	IUPHAR; IMEx (Orchard et al., 2012); InnateDB (Breuer et al., 2013).	Statistical model prioritizes ligand-receptor enrichment in pair of cell types relative to all cell pairs. Models complexes explicitly.
27. Cell lineage and communication network inference via optimization for single-cell transcriptomics (Wang et al., 2019a)	Curated selected pathways	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks. Probability of interaction is a function of ligand, receptor, and pathway gene expression.

28. Single-cell immune landscape of human atherosclerotic plaques (Fernandez et al., 2019)	FANTOM5	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Scored by expression.
29. Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data (Tsuyuzaki et al., 2019)	Organism-specific DBs using UniProt, HPRD; STRING	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Modeled as a hypergraph, with cell types as nodes and L-R interactions are hyperedges, which allows for one/many-to-many interactions.
30. PyMINEr finds gene and autocrine-paracrine networks from human islet scRNA-seq (Tyler et al., 2019)	Extracellular domain or secreted GO terms; STRING	Ligand & receptor expression above threshold in single-cell RNAseq experiments.
31. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics (Cabello-Aguilar et al., 2020)	FANTOM5; Reactome (Jassal et al., 2020); manual curation	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Scored by normalized expression.
32. Single-cell transcriptome-based multilayer network biomarker for predicting prognosis and therapeutic response of gliomas (Zhang et al., 2020)	DLRP ; HPRD; KEGG	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks. Supported by prognostic power of EGFR signal in glioma. No code published.
33. NicheNet: modeling intercellular communication by linking ligands to target genes (Browaeys et al., 2020)	Many databases, mostly accessed through Harmonizome (Rouillard et al., 2016)	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks. Weights trained on collection of ligand-perturbation gene expression data.
34. Predicting cell-to-cell communication networks using NATMI (Hou et al., 2020)	Updated FANTOM5 to connectomeDB2020	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Weighted by expression or specificity.
35. FunRes: resolving tissue-specific functional cell states based on a cell–cell communication network model (Jung et al., 2021)	FANTOM5; OmniPath (Türei et al., 2016, 2021); Reactome; MetaCore (Clarivate)	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks. Ligand-receptor interaction inferred by Markov chain from transcription factor expression.
36. Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly (Ren et al., 2020)	FANTOM5 & curated list of immune cytokines / chemokines.	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Scored by expression.
37. Pan-Cancer Analysis of Ligand–Receptor Cross-talk in the Tumor Microenvironment (Ghoshdastider et al., 2021)	FANTOM5 & curated list of immune checkpoint L-R pairs.	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Direction of cell-cell signaling scored by relative expression of ligand/receptor between cell types.
38. Molecular topography of an entire nervous system (Taylor et al., 2021)	Curated list of *C. elegans* cell adhesion molecules	Proposes network differential gene expression (nDGE), calculating bivariate genetic effects between groups of interacting cell pairs.

39. CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data (Hu et al., 2021)	FANTOM5	Multilayer model of ligand-receptor and intracellular gene interaction networks. Prize-collecting Steiner forest algorithm identifies subnetwork with highest L-R correlation across cell types and intracellular cell type specificity.
40. Dissection of intercellular communication using the transcriptome-based framework ICELLNET (Noël et al., 2021)	Manual curation of immune signaling L-R pairs.	Scored by product of scaled ligand & receptor expression.
41. Inference and analysis of cell-cell communication using CellChat (Jin et al., 2021)	KEGG, STRING & literature curation	Ligand & receptor expression above threshold in single-cell RNAseq experiments after imputation by PPI network propagation. Scoring model based on law of mass action, includes agonist + antagonist cofactor expression. Cell-cell edges scored by number of ligand-receptor edges over expected by permutation.
42. scConnect: a method for exploratory analysis of cell-cell communication based on single cell RNA sequencing data (Jakobsson et al., 2021)	IUPHAR & manual association of non-peptide ligands with gene products required for synthesis	Ligand & receptor expression above threshold in single-cell RNAseq experiments. Scored by expression. For non-peptide ligands, score is the geometric mean of synthesis genes.
43. CrossTalkeR: Analysis and Visualisation of Ligand Receptor Networks (Nagai et al., 2021)	CellphoneDB (Efremova et al., 2020)	Summarizes CellPhoneDB output as a directed graph for the purpose of calculating network topology statistics.
44. CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice (Shao et al., 2021)	STRING; text mining of GO terms, NCBI gene annotation, PubMed abstracts	A ligand-receptor database meant to be used with existing cell-cell interaction algorithms.
45. Cellinker: a platform of ligand–receptor interactions for intercellular communication analysis (Zhang et al., 2021b)	Literature curation; DLRP; HPRD; IUPHAR; CellphoneDB	Interaction database includes endogenous small molecules annotated as inorganic, metabolite, natural product, peptide and synthetic organic. Interaction database categorizes cell-cell interaction types. Statistical model prioritizes L-R interactions specific to cell type pairs.
46. CellCall: integrating paired ligand–receptor and transcription factor activities for cell–cell communication (Zhang et al., 2021a)	connectomeDB2020 (Hou et al., 2020); CellTalkDB (Shao et al., 2021); Cellinker (Zhang et al., 2021b); CellChat (Jin et al., 2021)	Multilayer model of ligand-receptor, receptor-pathway, and gene regulatory networks. Scoring combines expression of L-R with expression of downstream regulon linked by KEGG and TF databases using enrichment statistics from GSEA (Subramanian et al., 2005).

47. scAgeCom: a murine atlas of age-related changes in intercellular communication inferred with the package scDiffCom (Lagger et al., 2021)	Curated from CellChat; CellPhoneDB; CellTalkDB; connectomeDB2020; ICELLNET; NicheNet; SingleCellSignalR	Geometric mean of ligand & receptor expression above threshold, specific (using CellPhoneDB's permutation test), and differential between conditions by permutation testing. Computationally expensive.
48. Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19 (Cheng et al., 2021)	Collected ligand-receptor and TF-target interactions from multiple databases	Multlayer model of ligand-receptor, receptor-TF, and gene regulatory networks. L-R and R-TF graphs pruned of weakly expressed nodes. GRN refined by requiring significant positive correlations between TF and target gene expression.
49. Identification of Intercellular Signaling Changes Across Conditions and Their Influence on Intracellular Signaling Response From Multiple Single-Cell Datasets (Hao et al., 2021)	CellChat	Multlayer model of ligand-receptor (CellChat), receptor-TF (OmniPath), and gene regulatory networks. Cell type specific GRNs inferred by coexpression analysis using TF networks from DoRothEA (Garcia-Alonso et al., 2019) as a prior.
50. Computational reconstruction of the signalling networks surrounding implanted biomaterials from single-cell transcriptomics (Cherry et al., 2021)	CellphoneDB for L-R interactions, SCENIC (Aibar et al., 2017) for GRN.	Receptors positively correlated with TFs (excluding receptors that were TF targets by SCENIC GRN) were considered to be activating those TFs. Expression of cognate ligands used to identify sender cells.
51. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis (Türei et al., 2021)	Curation from 26 external databases.	OmniPath now includes annotations for protein roles in intercellular signaling: surface/secreted ligand; receptor; ECM; adhesion; surface/secreted enzymes; transporter.
52. Deciphering cell–cell interactions and communication from gene expression (Armingol et al., 2021)	Collected from many published ligand-receptor lists	Indexed available ligand-receptor lists: https://github.com/LewisLabUCSD/Ligand-Receptor-Pairs
53. Inferring a spatial code of cell-cell interactions across a whole animal body (Armingol et al., 2022b)	*C. elegans* orthologs mapped from FANTOM5 + manual curation	Ligand & receptor expression above threshold. Cell-cell interaction scored by modified Bray-Curtis: number of cell pair-specific interacting ligands + receptors relative to total ligands + receptors expressed in pair.
54. Context-aware deconvolution of cell–cell communication with Tensor-cell2cell (Armingol et al., 2022a)	CellChat	Ligand & receptor mean expression per cell pair per sample into 4D matrix. Non-negative TCA (Friedlander and Hatz, 2008) to associate contributions of ligand-receptor pairs with samples.

55. Reconstructing co-dependent cellular crosstalk in lung adenocarcinoma using REMI (Yu et al., 2022)	FANTOM5	Calculates partial correlations in graph of ligand-receptor interactions to improve specificity of conditionally-dependent interaction prediction.
56. Computation and visualization of cell–cell signaling topologies in single-cell systems data using Connectome (Raredon et al., 2022)	FANTOM5	Ligand-receptor network edges weighted by expression or relative expression. Edges pruned by specificity or differential expression of ligands/receptors. Will also calculate network topology statistics.
57. InterCellDB: A User-Defined Database for Inferring Intercellular Networks (Jin et al., 2022)	All human & mouse proteins in Ensembl & NCBI annotated by STRING for interactions, COMPARTMENTS (Binder et al., 2014) for locations, Uniprot and GO for functions.	All protein-protein interactions considered, with optional filtering by function/location. Edge weight is product of gene expression, significance determined by permutation testing. Edges annotated with expected interaction effect from STRING. Plotting includes location.
58. CINS: Cell Interaction Network inference from Single cell expression data (Yuan et al., 2022)	NicheNet's multiscale network to associate ligands with transcriptional response	Learns a Bayesian network of cell type dependencies from changes in cell type proportions in case-control studies. Causal ligands from cell-cell edges are predicted by LASSO regression of ligands and response genes.
59. An individualized causal framework for learning intercellular communication networks that define microenvironments of individual tumors (Chen et al., 2022)	Not reported	Nested hierarchical Dirichlet process to identify context-specific gene expression modules. Learn causal Bayesian network from cell type – gene expression module associations. No attempt to identify ligand-receptor pairs involved in interactions predicted by cell-cell network.
60. Cancer cells communicate with macrophages to prevent T cell activation during development of cell cycle therapy resistance (Griffiths et al., 2022)	FANTOM5	Proposes TWISTER to consider cell type-niche interactions. Sums ligand expression proportional to relative prevalence of contributing cell types for each receiving cell type.
61. Comprehensive visualization of cell–cell interactions in single-cell and spatial transcriptomics with NICHES (Raredon et al., 2023)	OmniPath, FANTOM5	Generates "cell-cell" and "niche" matrices interpretable using scRNAseq analysis workflows (ie. Seurat/Scater). Cell-cell matrix rows are ligand-receptor pairs, columns are individual cell pairs, value is product of ligand & receptor gene expression. Niche matrix contains mean ligand expression in system for each receptor in each cell.