Comparative analysis of cell–cell communication(CCC) from total and polyadenylated scRNA-seq data

Project raport:

Modeling of Complex Biological Systems

Master's course 1000-719bMSB

Author:

Katsiaryna Dubrouskaya

Abstract

Cell-to-cell communication(CCC) is essential for organizing complex biological processes, especially when the embryo is developing. This report describes how two single-cell RNA sequencing datasets, VASA and Chromium (Atlas), were used to compare intercellular communication networks in mouse embryos at embryonic day 7.5 (7.5E) using the computational tool CellChat. The inclusion of non-polyadenylated RNA species (such as lncRNAs and unspliced RNAs) in the VASA dataset, which significantly increased its transcriptome capture in comparison to the polyA-selected Atlas data, is a significant methodological difference.

Significantly higher predicted interaction counts and strengths in the VASA dataset demonstrated the quantitative dominance of communication, according to my analysis.
Also, CellChat enabled the identification of specific signaling roles for chosen cell types and the ranking of key altered signaling pathways. Any further research would greatly benefit from integrating in-depth knowledge of mouse embryonic development to contextualize these computational insights.

Table of contents

Introduction

Aim of the Project

This project conducts a comprehensive comparative analysis of cell-to-cell communication using the CellChat package, on total scRNA-seq data generated using the VASA-seq method and polyadenylated scRNA-seq data generated using 10x Chromium.

VASA-seq method

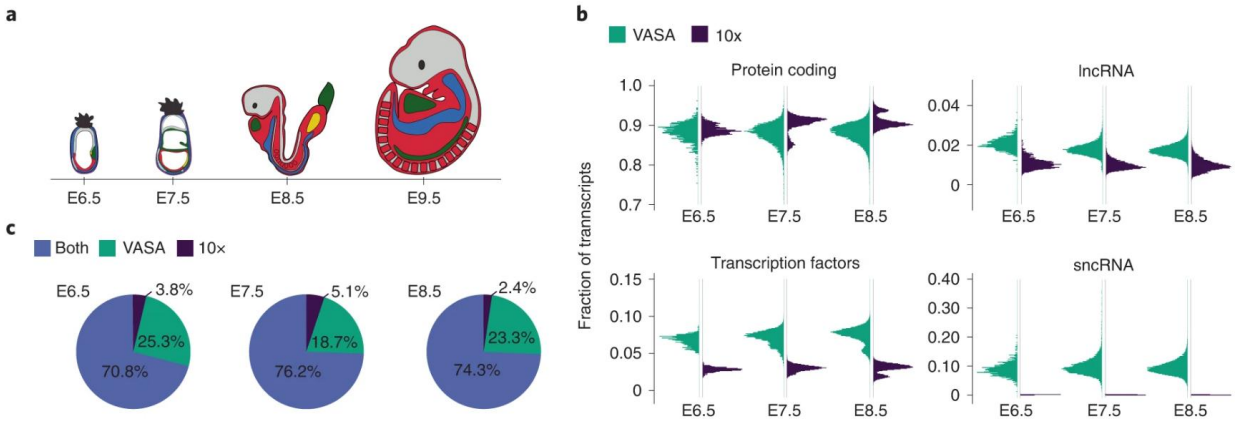
Single-cell RNA sequencing is a powerful technique, which transform the understanding of cellular complexity over the last decade. Most of the method rely on the the hybridization of barcoded oligo-dT primers to the poly(A) end of polyadenylated transcripts for RNA capture and complementary DNA (cDNA) synthesis. This approach leaves out the whole spectrum of non-polyadenylated transcripts.

To overcome this, Salmen et al. (2022) developed 'vast transcriptome analysis of single cells by dA-tailing' (VASA-seq)[1], which captures both non-polyadenylated and polyadenylated transcripts across their. It was achieved thanks to the protocol in which, after RNA fragmentation, they perform end repair and then add a universal poly(A) tail to all these RNA fragments. Now, this poly(A) tail allows to synthesize cDNA using barcoded oligo-dT probes, regardless of whether the original RNA was polyadenylated or not. In addition, a unique fragment identifier (UFI), a short, random sequence that is ligated or incorporated into each RNA molecule before any amplification steps, allowed them to ensure accurate and strand-specific quantification.

Using VASA-drop, they sequenced 33,662 cells from mouse post-implantation embryos at four different developmental stages: embryonic day (E) 6.5, E7.5, E8.5 and E9.5 (Fig. 1a). The datasets from E6.5, E7.5 and E8.5 were directly compared to a reference dataset generated using the 10x Chromium platform[2].

However, VASA-plate proportionally detected about twice as many long non-coding RNAs (lncRNAs) as 10x Chromium (Fig. 1b). Also VASA-seq detected short non-coding RNAs (sncRNAs), mainly miscellaneous RNA (miscRNA), small nucleolar RNA (snoRNA), ribozymes and small nuclear RNA (snRNA) (Fig. 1b). And 70.8–76.2% of the detected genes were shared between the methods; 18.7–25.3% were detected only in VASA-seq; and 2.4–5.1% were detected only in 10x (Fig. 1c).

Figure 1: VASA-seq enables novel marker gene detection in the developing mouse embryo.
***a**, Schematic figure of mouse embryo morphology at developmental stages E6.5, E7.5, E8.5 and E9.5 (left to right).*
***b**, Fraction of transcripts per biotype in VASA-seq compared to 10x Chromium for mouse embryos at each timepoint using the 20% terminal portion of genes. The comparison includes protein-coding genes (top-left panel), lncRNAs (top-right panel), TFs (bottom-left panel) and sncRNAs (bottom-right panel).*
***c**, Percentage of genes detected in VASA-seq compared to 10x for each timepoint using the 20% terminal portion of genes.*
***Note**: the figure and description comes from: Salmen, F., De Jonghe, J., Kaminski, T.S. et al. High-throughput total RNA sequencing in single cells using VASA-seq. Nat Biotechnol 40, 1780–1793 (2022). https://doi.org/10.1038/s41587-022-01361-8.*



CellChat

Multicellular organism are composed of diverse cell types that need to communicate with each other to work together effectively. Growth, development, differentiation, tissue and organ formation, maintenance, and physiological regulation all require cell–cell communication (CCC). Cells could communicate through direct contact or at a distance using ligand–receptor interactions. So cellular communication includes two essential processes: cells generating and sending signals (signal conduction), and other cells receiving and interpreting those signals (signal transduction). By understanding intercellular communication networks we will be closer to understanding cell differentiation, development, metabolism etc.

One of the tools to do so I CellChat. At its core, CellChat uses a curated database, CellChatDB, which contains 2021 validated ligand-receptor interactions, including 60% of secreting interactions[3]. In addition, 48% of the interactions involve heteromeric molecular complexes.

CellChat either requires user assigned cell labels as input or automatically groups cells based on the low-dimensional data representation supplied as input. On the preprocessed data u can run the core function computeCommunProb(), which is based on the law of mass action (the strength of a reaction is proportional to the product of the concentrations of the reactants).

Then, it computes the probability of communication between cell groups by assessing the expression of ligand-receptor pairs. It also considers cofactors and complex interactions to better model real biological signaling.

Methods and materials

To conduct this comparative analysis, I select two scRNA datasets, both representing mouse embryonic development at E7.5. First data comes from Salmen et al. (2022) paper and was prepared using VASA-seq and are are available at the Gene Expression Omnibus under accession number GSE176588 (https://www.ncbi.nlm.nih.gov/geo/). Another one is generated with 10x Genomics Chromium, and available under accession number E-MTAB6967 (https://www.ebi.ac.uk/biostudies/).

There are 6,505 cells for VASA-seq 7.5E with 151,299 features and 12,876 cells with 29,452 features for 10x Chromium.

ScRNA-seq data analysis and integration

The initial preprocessing for the VASA-seq 7.5E dataset involved making multiple individual Seurat objects from each sample, and merging them after into a single Seurat object (VASA_seq_analysis.R). And no pre-process for 10x Chromium.

Also, no additional filtering based on quality control metrics was performed on either the VASA-seq or 10x Chromium datasets. Both datasets were already selected due to the standards for the embryonic cells, as for both, the overall mitochondrial percentage was lower than 3%, and the number of features per cell was lower than 6,000.

For consistency in metadata, unnecessary columns from the 10x Chromium (atlas) Seurat object were removed, retaining only “orig.ident”, “nCount_RNA”, “nFeature_RNA”, “sample”, and “celltype”.

Both datasets were first normalized using NormalizeData, next the highly variable features were identified with FindVariableFeatures (selection.method = "vst", nfeatures = 2000). A shared set of 2000 integration features was selected using SelectIntegrationFeatures. Data were then scaled independently for these features using ScaleData.

To reduce memory consumption and equel a larger Atlas dataset, it was subsampled to 7,000 cells. PCA was performed on both datasets using the shared integration features. Integration anchors were identified using FindIntegrationAnchors with reduction = "rpca". Next, the datasets were integrated using IntegrateData.

ScaleData, RunPCA, RunUMAP (dims = 1:15), FindNeighbors (dims = 1:15), and FindClusters (resolution = 0.5) were used to analyze the integrated dataset as a standard Seurat downstream method. Clustering was performed on the integrated data to enable the comparative analysis of CCC. Cell type annotation was done using the SingleR package.

CCC with CellChat

To start with CCC analysis, the integrated Seurat object was first split into two separate Seurat objects corresponding to the VASA-seq and 10x Chromium datasets. From the VASA-seq subset, cells belonging to seurat_clusters "13" were removed, cause there it is not represented in Chromium dataset, which implement the errors into the forward analysis, due to unequal amount of layers.

Individual CellChat objects were then created for each subset.

Following tutorial from https://github.com/sqjin/CellChat/tree/master, each CellChat object underwent standard preprocessing steps:

- subsetData(): filters the expression data to leave only the genes present in the CellChatDB
- identifyOverExpressedGenes(): identifies genes that are significantly over-expressed in each cell group, because CellChat often focuses on interactions where at least one of the components is over-expressed in its respective cell group.

- identifyOverExpressedInteractions(): identifies ligand-receptor pairs from previously selected over-expressed data.
- projectData(): overlay gene expression from data onto protein-protein interaction (PPI) networks using PPI.mouse to infer protein activity.
- computeCommunProb(): computes communication probabilities between cell groups using on the law of mass action to determine the strength of the connection.
- filterCommunication(): removes interactions, which have less than 3 cells expressing a ligand/receptor.
- computeCommunProbPathway(): infer the overall strength and significance of the calculated ligand-receptor interactions at the pathway level.
- aggregateNet(): create a more summarized view of the communication network.
- netAnalysis_computeCentrality(): helps to identify key signaling roles of cell types, by computing the network centrality.

Finally, the individual cellchat objects were merged into a single cellchat_merge object using mergeCellChat().

Results

As shown in "Figure 2" the VASA dataset significantly outperforms the Chromium (Atlas) one in terms of number of total interaction (within and between all cell types) and the total aggregated communication strength. This observation is consistent, given that the number of genes detected in the VASA dataset is at least five times larger than in the Chromium dataset.

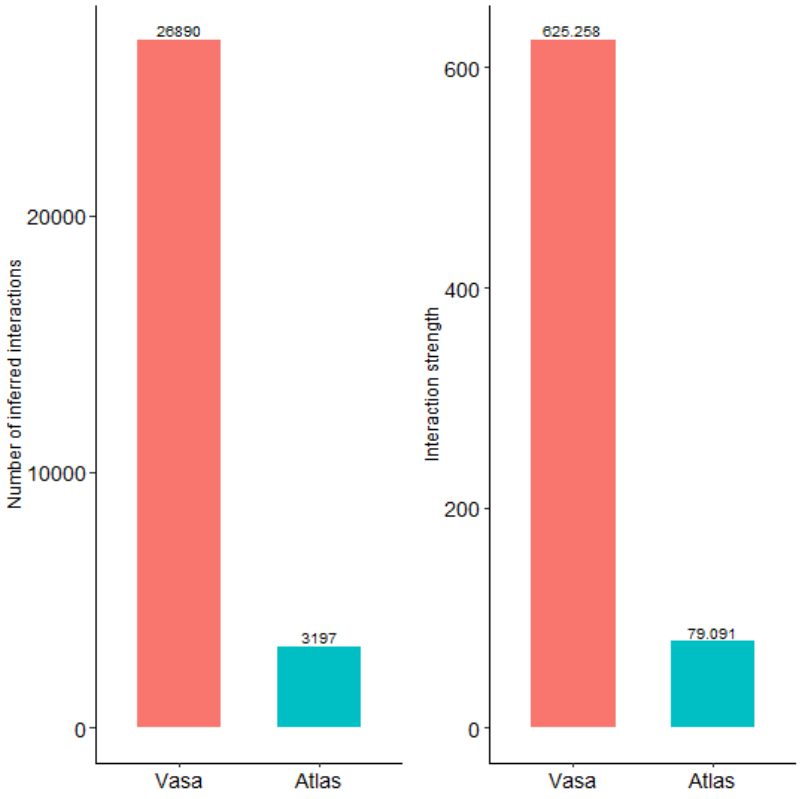


Figure 2**: **The high-level summary plots of the cell-cell communication.
*The **left panel**, representing the total number of cell-cell interactions, shows that VASA dataset exhibits a significantly higher number of communication links compared to Atlas dataset. Similarly, the **right panel**, which depicts the total aggregated communication strength, reinforces this observation*

Initial attempts to visualize the differential number of interactions or interaction strength using heatmaps (as shown in Fig. 3) were largely uninformative. The high overall activity in the VASA dataset dominated the visualization scale, rendering these plots unreadable and making it challenging to discern specific increases or decreases between the datasets.

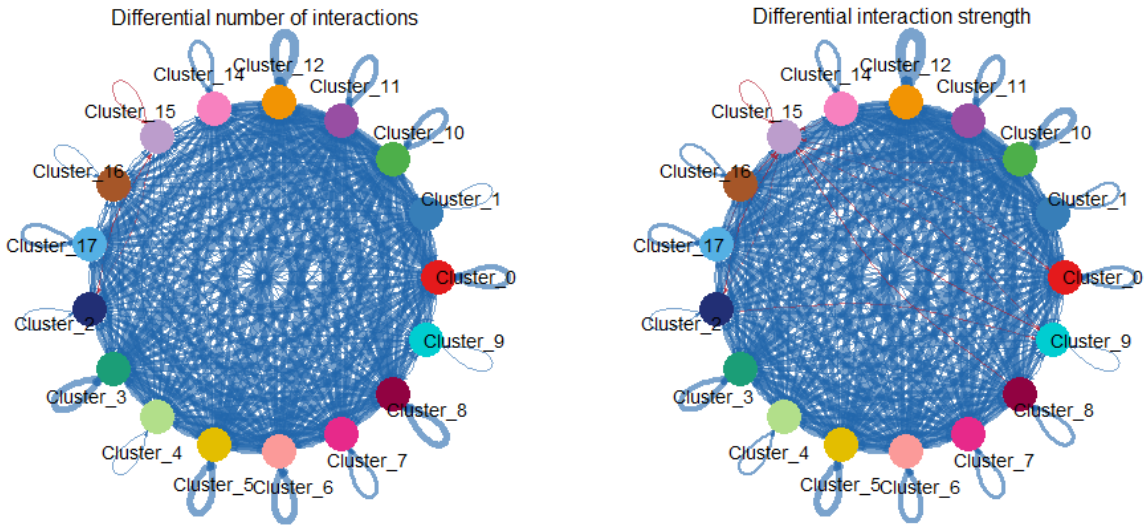


Figure 3:Differential number of interactions or interaction strength among different cell datasets
Where, blue for increase in VASA data and red for increase Atlas data.

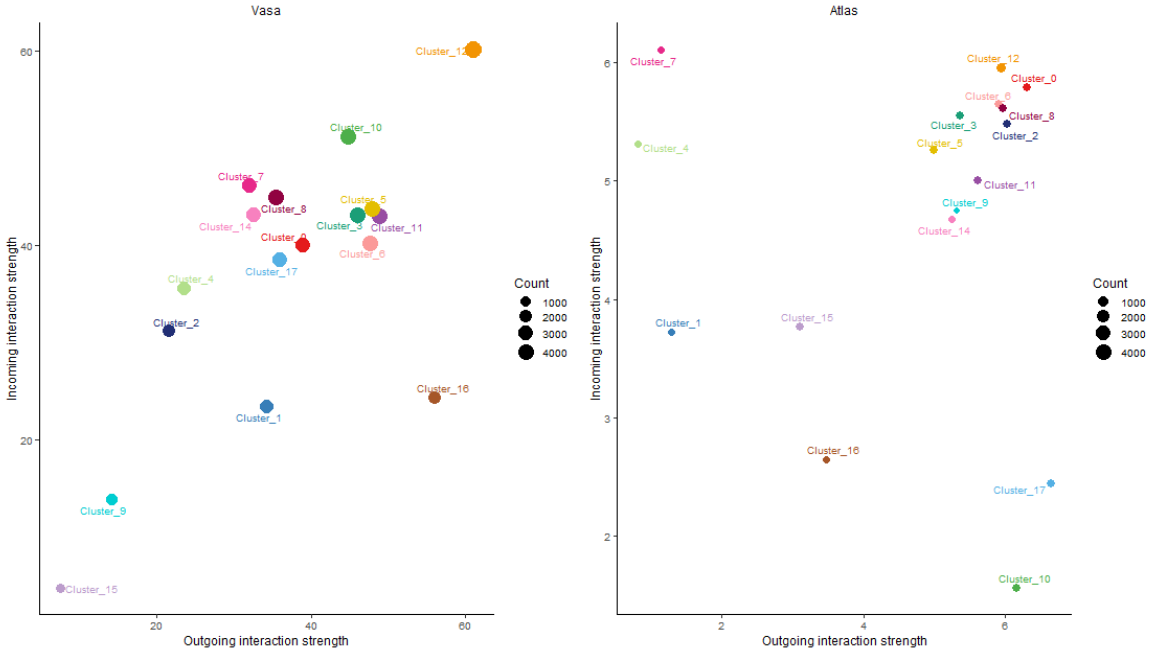


Figure 4: Signaling Role Scatter Plots
*The **left** panel corresponds to the VASA dataset, and the **right** panel to the Atlas dataset. Each point represents a cell cluster, positioned by its total outgoing and incoming interaction strength. Point size indicates the total number of inferred communication links for that cluster.*

The next stage of CCC analysis involved analyzing the signaling roles of individual cell types within each dataset. Fig. 4 presents scatter plots showing each cluster's total outgoing versus incoming interaction strength.

The VASA dataset shows significantly higher values for both outgoing and incoming interaction strengths, with scales extending to 60, whereas the Atlas dataset shows scales limited to 6-7.

To go deeper into the data you can investigate the signaling changes within the cluster, comparing two datasets. As an example Figure 5 displays scatter plots detailing pathway-specific changes for Cluster_12 and Cluster_15.

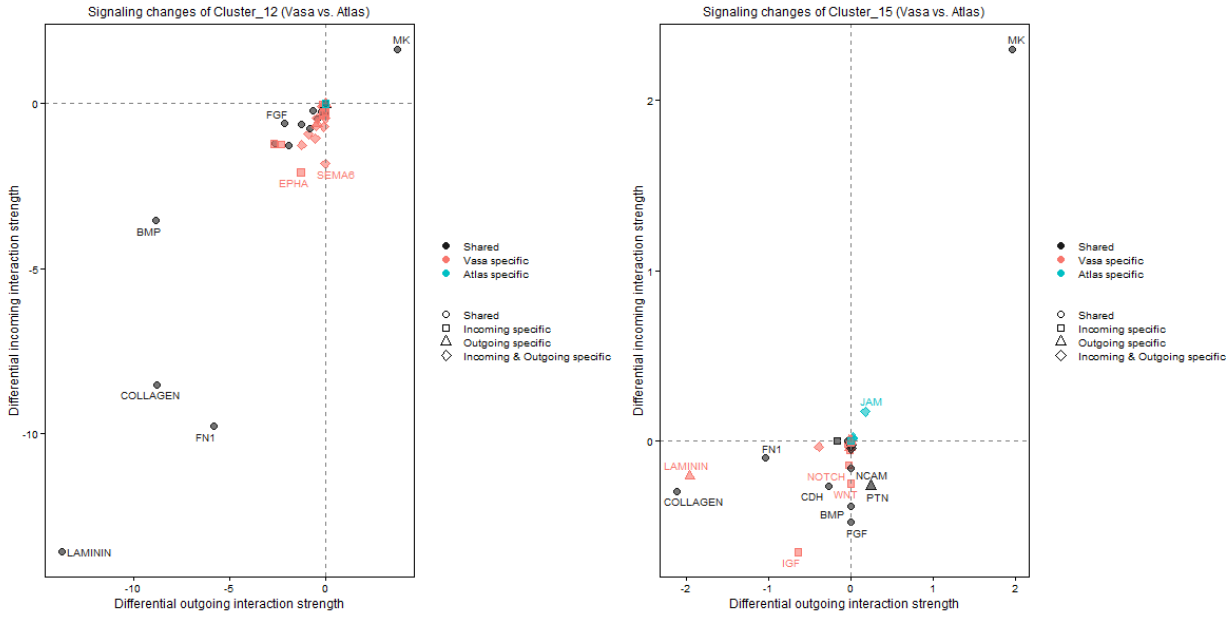


Figure 5: Pathway-specific changes in chosen cell types
*The **left** panel illustrates signaling changes for Cluster_12, and the **right** panel for Cluster_15.*

To identify the most significantly altered signaling pathways between the two conditions, pathways were ranked based on their differential information flow (the overall activity of a signaling pathway) (Fig. 6).

The VASA dataset shows a significant increase in information flow across the vast majority of pathways, as previously observed. Pathways such as SEMA3/5/6, VISFATIN, DESMOSOME, NRG, and COLLAGEN are notably more active in Vasa with minimal corresponding activity in Atlas. On the other hand, only a few pathways show relatively higher activity in the Atlas dataset, including MK, CLDN, GALECTIN and TWEAK.

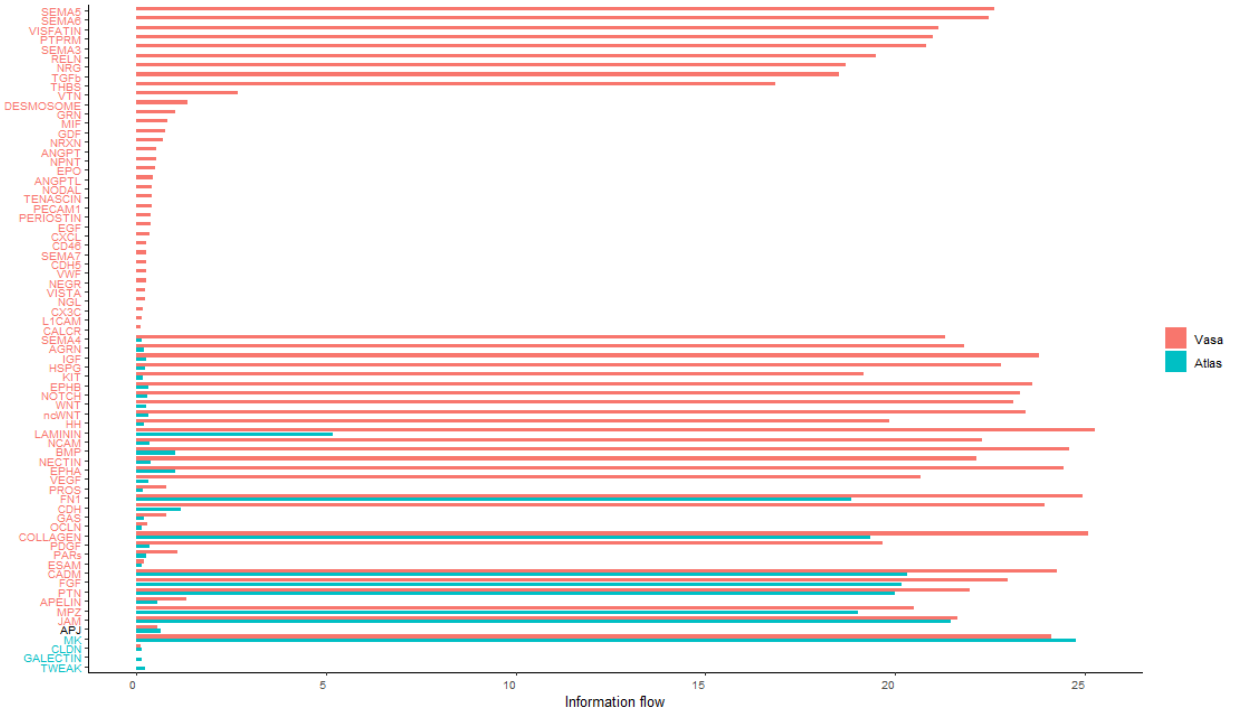


Figure 6: The horizontal bar plots ranking of differential signaling pathways
Red bars for VASA data, blue for Atlas

Discussion

CellChat is a very powerful tool which analyze cell-cell communication networks from scRNA-seq data, in addition CellChat offers varied visualization outputs for different analytical tasks. This report is an investigation of CellChat's methodology, demonstrating its capabilities in comparing complex communication network between different conditions.

In this analysis of mouse embryo 7.5E data, it was found that VASA and Chromium (Atlas) datasets had significant quantitative differences in CCC. A key factor in this observation might be the inclusion of non-polyadenylated RNAs (such as lncRNAs, sncRNAs, and unspliced RNAs) in the VASA dataset, which are typically excluded from standard polyA-selected sequencing, as in the Atlas data. This broader capture of the transcriptome in VASA led to a higher number of detected genes and, consequently, a denser and more active communication network.

One of the limitation in this analysis is my lack of fundamental biological knowledge about the mouse embryo about development, but if you know what to look out for, it can be very helpful. One way or another, further biological validation with in-depth knowledge of mouse embryonic development, is essential to contextualize these computational findings.

Also, as this raport focused on CellChat, there are other packages like NicheNet and Scriabin, which offer alternative approaches, instead of clusters they based on the single-cell or individual ligand-receptor level, providing even more complex ground for research and analyses.

References

1. Salmen, F. *et al*. High-throughput total RNA sequencing in single cells using VASA-seq. *Nat Biotechnol* **40**, 1780–1793 (2022).
2. Pijuan-Sala, B. *et al*. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
3. Jin, S. *et al*. Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12**, 1088 (2021).