nature protocols

https://doi.org/10.1038/s41596-024-01137-1

Protocol update

Check for updates

CellPhoneDB v5: inferring cell–cell communication from single-cell multiomics data

In the format provided by the authors and unedited

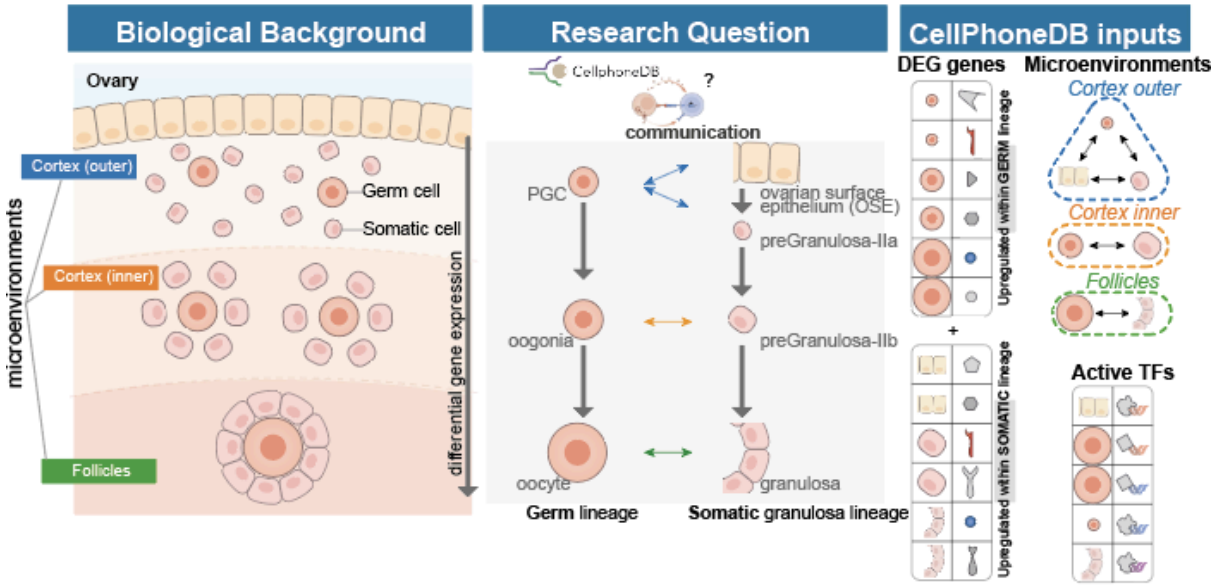
Supplementary Information

Supplementary Notes.

Case Example 1. Germ-somatic communication during ovarian development (with CellSign)

This case study combines **CellPhoneDB Method 3** with **microenvironments** (from Spatial Transcriptomics) and **CellSign** modules (TF activities derived from scATAC-seq) to predict cell-cell interactions specific to each differentiation timepoint of human oocytes. For more details, see Garcia-Alonso et al. (2022), 'The second wave of pre-granulosa cells' section.

Biological background (Figure 3a). Germ cells and somatic cells synchronise their differentiation using local paracrine factors (Supplementary Figure 1a) in the developing ovary. Single-cell transcriptomics analysis revealed several somatic cell states (somatic^{OSE}, somatic$^{preGC\text{-}IIa}$, somatic$^{preGC\text{-}IIb}$, and somatic^{granulosa}) and several germ cell states (germ^{PGCs}, germ^{oogonia}, germ$^{oogonia\text{-}meiotic}$, and germ^{oocytes}). Spatial transcriptomics analysis identified a gradient of differentiation (from undifferentiated cells in the cortex to differentiated cells in the medulla), which we categorised into three microenvironments corresponding to the stages of cell differentiation (see Figure 3a). In other words, germ-somatic cells follow a co-differentiation trajectory that aligns with the microenvironments they inhabit.



Research question. *What are the cell-cell interactions that change along the differentiation trajectory of germ and somatic cells?*

Method selection. Because we aim to identify the interactions that change when cells transition from one state to another in their differentiation trajectory (i.e. germ$^{oogonia\text{-}meiotic}$ → germ^{oocytes}), we will use Method 3 to target those specific events. Some ligand-receptor interactions between germ-somatic cells will remain constant throughout the co-differentiation trajectory, but here we are interested in interactions that exhibit dynamic patterns and could drive cell identity changes and differentiation. It's important to note that there are additional cell types in the ovary (e.g., stromal, immune, endothelial); however, we

will limit the analysis to the germ and somatic lineages because: (i) they are in direct contact with one another, and (ii) they undergo synchronised differentiation.

Experimental setup and CellphoneDB INPUTS. We will use **CellphoneDB method 3** to retrieve cell-cell interactions involving genes upregulated by any germ or somatic cell state. Beside the mandatory counts and meta files, we will provide the following inputs.

- DEGs file (mandatory for CellphoneDB Method 3), which contains the genes overexpressed by each cell state in a differentiation trajectory. DEGs were calculated using a one-sided Wilcoxon Rank Sum test implemented in the FindAllMarkers function with Seurat v.3.2.2 in a "one versus rest-of-the-lineage" fashion. Note that, in the case of needing to account for batches or covariates, we could have used any other test to identify differentially expressed genes (for example limma, MAST, edgeR, etc).

Critical step.
To capture subtle expression differences along each lineage trajectory (the somatic and the germinal), we need to compare each cell state against the other cell states of the same lineage-trajectory in a **"one versus rest-of-lineage"** fashion (i.e. hierarchical differential expression analysis) instead of **"one versus rest-of-cells-in-dataset",** which is the test performed by Method 2.

Why "one versus rest-of-the-lineage" and not "one versus all-the-rest"? Because "one versus all-the-rest" does not distinguish cell state-specific from lineage-specific upregulated genes. Note the differences below:

- "one versus rest-of-lineage"
  $germ^{PGCs}$ **vs [**$germ^{oogonia}$ + $germ^{oogonia\text{-}meiotic}$ + $germ^{oocytes}$**]**
- "one versus rest-of-cells-in-analysis"
  $germ^{PGCs}$ **vs [**$germ^{oogonia}$ + $germ^{oogonia\text{-}meiotic}$ + $germ^{oocytes}$ + $somatic^{OSE}$ + $somatic^{preGC\text{-}IIa}$ + $somatic^{preGC\text{-}IIb}$ + $somatic^{granulosa}$**]**
- "one versus rest-of-cells-in-dataset"
  $germ^{PGCs}$ **vs [**immune + endothelial + fibroblats + perivascular + $germ^{oogonia}$ + $germ^{oogonia\text{-}meiotic}$ + $germ^{oocytes}$ + $somatic^{OSE}$ + $somatic^{preGC\text{-}IIa}$ + $somatic^{preGC\text{-}IIb}$ + $somatic^{granulosa}$**]**

Here we want to specifically target ligands/receptors upregulated by a specific cell state along their differentiation trajectory, thus our reference should be the cells in the differentiation trajectory ("rest-of-lineage") and not the cells from other lineages ("rest-of-cells-in-analysis/dataset" including all immune, endothelial, somatic and germ cells together).

- Microenvironments file. Because two cells can only interact paracrinally if they colocalise in space, we will inform CellphoneDB about the cellular composition of each of the three ovarian microenvironments. CellphoneDB will use this information to restrict the analysis to those cell types that co-localise in each microenvironment (i.e. discard interactions between cells that never co-localise). These spatial niches were identified via spatial transcriptomics analysis where Visium Spatial

transcriptomics spots were deconvoluted into single cell type densities using cell2location.

- TF activity file. Because we have matched scATAC-seq data, we could use chromatin accessibility to estimate TF binding activity and identify active TF. CellphoneDB will use cell state TF activity to prioritise cell-cell interactions where the cell expressing the receptor has also the downstream directly-regulated TF active. In this study, TF activity was estimated from (i) calculating TF motif binding scores with chromVar v.1.12.2 using TF positional weight matrices from JASPAR v2018, HOCOMOCO v10, SwissRegulon, and HOMER, which returns a matrix with TF binding activity per cell, and then (ii) estimating differential TF binding activity between cell states in a “one versus rest-of-lineage” fashion using one-sided Wilcoxon Rank Sum test implemented in the FindAllMarkers function with Seurat v.3.2.2.

Link to input data & notebook:
https://github.com/ventolab/CellphoneDB/tree/master/example_data/NatureProtocols2024_case_studies/CaseExample1_differentiation

CellphoneDB analysis.
Timing 1 min.

1 Define input files

```
cpdb_file_path = 'db/v5/cellphonedb.zip'
meta_file_path = 'input/CaseExample1_differentiation/metadata.tsv'
counts_file_path =
'input/CaseExample1_differentiation/normalised_log_counts.h5ad'
degs_file_path = 'input/CaseExample1_differentiation/DEGs_genes.tsv'
active_tfs_file_path = 'input/CaseExample1_differentiation/active_TFs.tsv'
microenvs_file_path: 'input/CaseExample1_differentiation/microenvironments.tsv'
```

2 Running method 3

```
from cellphonedb.src.core.methods import cpdb_degs_analysis_method

cpdb_results = cpdb_degs_analysis_method.call(
       cpdb_file_path = cpdb_file_path,
       meta_file_path = meta_file_path,
       counts_file_path = counts_file_path,
       degs_file_path = degs_file_path, # method3-specific
       microenvs_file_path = microenvs_file_path, # optional
       active_tfs_file_path = active_tfs_file_path, # optional
       counts_data = 'hgnc_symbol',
       threshold = 0.1,
       result_precision = 3,
       threads = 4,
```

```
        separator = '|',
        debug = False,
        output_path = out_path,
        output_suffix = None)
```

3 Visualise results with ktplotspy

```
kpy.plot_cpdb_heatmap(pvals = cpdb_results['relevant_interactions'],
                      degs_analysis=True,
                      figsize=(5, 5),
                      title="Sum of significant interactions")
kpy.plot_cpdb(
    adata = adata,
    cell_type1 = "Somatic.granulosa",
    cell_type2 = "Germ.oocyte",
    means = cpdb_results['means'],
    pvals = cpdb_results['relevant_interactions'],
    celltype_key = "celltype",
    figsize = (10,15),
    title = "Interactions between oocytes and granulosa in the follicles",
    max_size = 5,
    highlight_size = 0.75,
    degs_analysis = True,
    standard_scale = False,
#     interaction_scores = cpdb_results['interaction_scores'],
    scale_alpha_by_interaction_scores=True,
)

# See interactions from JAG1 across dataset
kpy.plot_cpdb(
    adata = adata,
    cell_type1 = ".",
    cell_type2 = ".",
    means = cpdb_results['means'],
    pvals = cpdb_results['relevant_interactions'],
    celltype_key = "celltype",
    genes = ["JAG1"],
    figsize = (10,5),
    title = "Interactions involving JAG1 ligand",
    max_size = 5,
    highlight_size = 0.75,
    degs_analysis = True,
    standard_scale = False
)
```

Case Example 2. Cell-cell communication in the spatial niches of the endometrium

This case study combines **CellphoneDB method 3** with **microenvironments** (from Spatial Transcriptomics) to predict cell-cell interactions specific to each endometrial niche (i.e. glands, lumen, vessels, basalis layer) along the different phases of the menstrual cycle (Proliferative, Secretory-early and Secretory-mid). Embryos implant in the endometrium in the Secretory-mid timepoint. See Marečková et al. (2024) for more details.

Biological background. Epithelial cells and stromal cells synchronise their differentiation using local paracrine factors and endocrine hormones generated by the ovaries (estrogen and progesterone). Single cell transcriptomics analysis revealed several epithelial cell states and stromal cell states with mutually exclusive spatiotemporal locations.

Research question. *What epithelial-stromal interactions are niche-specific*?

Method selection. While some interactions will be shared across niches, our focus is on those that are niche-specific—interactions unique to a particular cell state within one niche that are absent in the cell states inhabiting other niches. This follows a similar approach to Case Example 1. Method 3 is the best suited for this question because it allows us to use 'other niches' for the null distribution; Method 2 would compare each cell type against “all cells in the dataset” but here we want our reference to be “all cells in other niches”. Because we do not want to retrieve interaction between cell types that do not coexist, we will define the niches using the microenvironments file.

Experimental setup and CellphoneDB INPUTS. We will use **CellphoneDB method 3** to retrieve niche-exclusive interactions involving genes upregulated by cell states unique to the niche. Beside the mandatory counts and meta files, we will provide the following inputs.

- DEGs file (mandatory for CellphoneDB method 3), which contains the genes overexpressed by any cell state within a lineage (i.e. within epithelials or within mesenchymal/stromal). DEGs were calculated using a one-sided Wilcoxon Rank Sum test implemented in the FindAllMarkers function with Seurat v.3.2.2 in a “one versus rest-of-the-lineage” fashion.

  Critical step.
  To capture subtle expression differences along each lineage, we need to compare each cell state against the other cell states of the same lineage in a **“one versus rest-of-lineage”** fashion (i.e. hierarchical differential expression analysis) instead of **“one versus rest-of-dataset”,** which is the test performed by Method 2.

  Why “one versus rest-of-the-lineage” and not “one versus all-the-rest”? Because “one versus all-the-rest” does not distinguish cell state-specific from lineage-specific upregulated genes.

Here we want to specifically target ligands/receptors upregulated by a specific cell state in a niche, thus our reference should be the lineage-matched cell states in the other niches (”rest-of-lineage”) and not the cells from other lineages (“all-the-rest”).

- Microenvironments file. Because two cells can only interact paracrinally if they colocalise in space and time (aka, menstrual cycle phase), we will inform CellphoneDB about the cellular composition of each endometrial niche. As in previous Case Study 1, the spatial niches were identified via spatial transcriptomics analysis where Visium Spatial transcriptomics spots were deconvoluted into single cell type densities using cell2location. In contrast, temporal information is inferred from the menstrual cycle phase when the endometrial sample was taken from the donor.

Link to input data & notebook.
https://github.com/ventolab/CellphoneDB/tree/master/example_data/NatureProtocols2024_case_studies/CaseExample2_spatialNiches

CellphoneDB analysis.
Timing 1 min.

1 Define input files

```
cpdb_file_path = 'db/v5/cellphonedb.zip'
meta_file_path = 'input/CaseExample2_spatialNiches/dataset_meta.tsv'
counts_file_path = 'input/CaseExample2_spatialNiches/counts_normloqTransformed.h5ad'
degs_file_path = 'input/CaseExample2_spatialNiches/DEGs_upregulated_genes.tsv'
microenvs_file_path: 'input/CaseExample2_spatialNiches/microenvironments.tsv'
```

2 Running method 3

```
from cellphonedb.src.core.methods import cpdb_degs_analysis_method

cpdb_results = cpdb_degs_analysis_method.call(
       cpdb_file_path = cpdb_file_path,
       meta_file_path = meta_file_path,
       counts_file_path = counts_file_path,
       degs_file_path = degs_file_path, # method3-specific
       microenvs_file_path = microenvs_file_path, # optional
       counts_data = 'hgnc_symbol',
       threshold = 0.1,
       result_precision = 3,
       threads = 4,
       separator = '|',
       debug = False,
       output_path = out_path,
       output_suffix = None)
```

3 Visualise results with ktplotspy

```
kpy.plot_cpdb_heatmap(pvals = cpdb_results['relevant_interactions'],
                      degs_analysis=True,
                      figsize=(8, 8),
                      title="Sum of significant interactions")

kpy.plot_cpdb(
    adata = adata,
    cell_type1 = "epi.SOX9_functionalis_II|epi.preGlandular|epi.Glandular",
    cell_type2 = "mes.eStromal|mes.dStromal_early|mes.dStromal_mid",
    means = cpdb_results['means'],
    pvals = cpdb_results['relevant_interactions'],
    celltype_key = "celltype",
#     genes = ["JAG1"],
    figsize = (8,20),
    title = "Interactions between epithelial \nand stroma in the secretory glands",
    max_size = 5,
    highlight_size = 0.75,
    degs_analysis = True,
    standard_scale = False,
#     interaction_scores = cpdb_results['interaction_scores'],
    scale_alpha_by_interaction_scores=True,
)

# Group results by pathway
from plotnine import facet_wrap

p = kpy.plot_cpdb(
    adata = adata,
    cell_type1 = "epi.SOX9_functionalis_II|epi.preGlandular|epi.Glandular",
    cell_type2 = "mes.eStromal|mes.dStromal_early|mes.dStromal_mid",
    means = cpdb_results['means'],
    pvals = cpdb_results['relevant_interactions'],
    celltype_key = "celltype",
#     genes = ["JAG1"],
    figsize = (15,25),
    title = "Interactions between epithelial \nand stroma in the secretory glands",
    max_size = 5,
    highlight_size = 0.75,
    degs_analysis = True,
    standard_scale = False,
#     interaction_scores = cpdb_results['interaction_scores'],
```

```
        scale_alpha_by_interaction_scores=True,
    )

    p + facet_wrap("~ classification", ncol = 2, scales='free_y')
```

Case Example 3. Overexpressed interactions between trophoblasts and maternal vessels.

This case study combines **CellphoneDB method 2** with **microenvironments** (from Spatial Transcriptomics) and **specificity scoring** to explore all cell-cell interactions between the fetal trophoblast and the maternal vessels. See Arutyunyan et al. (2023, Nature) “eEVT interactions with spiral arteries” section for more details.

Biological background. Fetal trophoblast cells invade the maternal decidua to remodel maternal vessels. Arutyunyan et al. (2023) used single cell transcriptomics analysis to explore the usage of L/R between the fetal extravillous trophoblast (EVTs) and the maternal perivascular cells (PV). The microenvironments file was used to restrict the output to interactions between EVTs and PVs, and do not report interaction between other cells.

Research question. *What are the cell-cell interactions that are OVEREXPRESSED between the fetal extravillius trophoblast (EVTs) and any maternal perivascular cell type (PV) in the maternal-fetal interface?*

Method selection. While there will be interactions expressed in all cells, here the aim is to extract the interactions that are overexpressed. This is an exploratory analysis to identify overexpressed interactions in a broad sense, therefore, we chose to use Method 2 to detect interactions upregulated by a cell type pair across the entire dataset. Because we anticipate a large number of overexpressed interactions, we will also use the scoring approach on top of Method 2 to further prioritise the topmost *EVT-PV* specific interactions (i.e. interactions rarely used by other cell type pairs).

Experimental setup and CellphoneDB INPUTS. We will use **CellphoneDB method 2** to retrieve interactions that are upregulated between the fetal extravillous trophoblast (EVTs) and the maternal perivascular cells (PV). Beside the mandatory counts and meta files, we will provide the following inputs.

- Microenvironments file. We will use this file to tell CellphoneDB that we only want to output interactions between the invading trophoblast cells (the EVTs) and the cells around the vessels (the perivascular cells - PVs). Thus, this file will have only one microenvironment that will include all PVs and all EVT cell types.

Link to input data & notebook
https://github.com/ventolab/CellphoneDB/tree/master/example_data/NatureProtocols2024_case_studies/CaseExample3_specificityScoring

CellphoneDB analysis.
Timing 1 min.

1 Define input files

```
cpdb_file_path = 'db/v5/cellphonedb.zip'
meta_file_path = 'input/CaseExample3_specificityScoring/dataset_meta.tsv'
counts_file_path = 'input/CaseExample3_specificityScoring/counts_normloqTransformed.h5ad'
microenvs_file_path: 'input/CaseExample3_specificityScoring/microenvironments.tsv'
```

2 Running method 2

```
from cellphonedb.src.core.methods import cpdb_statistical_analysis_method

cpdb_results = cpdb_statistical_analysis_method.call(
       cpdb_file_path = cpdb_file_path,
       meta_file_path = meta_file_path,
       counts_file_path = counts_file_path,
       counts_data = 'hgnc_symbol',
       microenvs_file_path = microenvs_file_path,
       score_interactions = True,
       threshold = 0.1,
       result_precision = 3,
       separator = '|',
       debug = False,
       output_path = out_path,
       output_suffix = None,
       threads = 25
    )
```

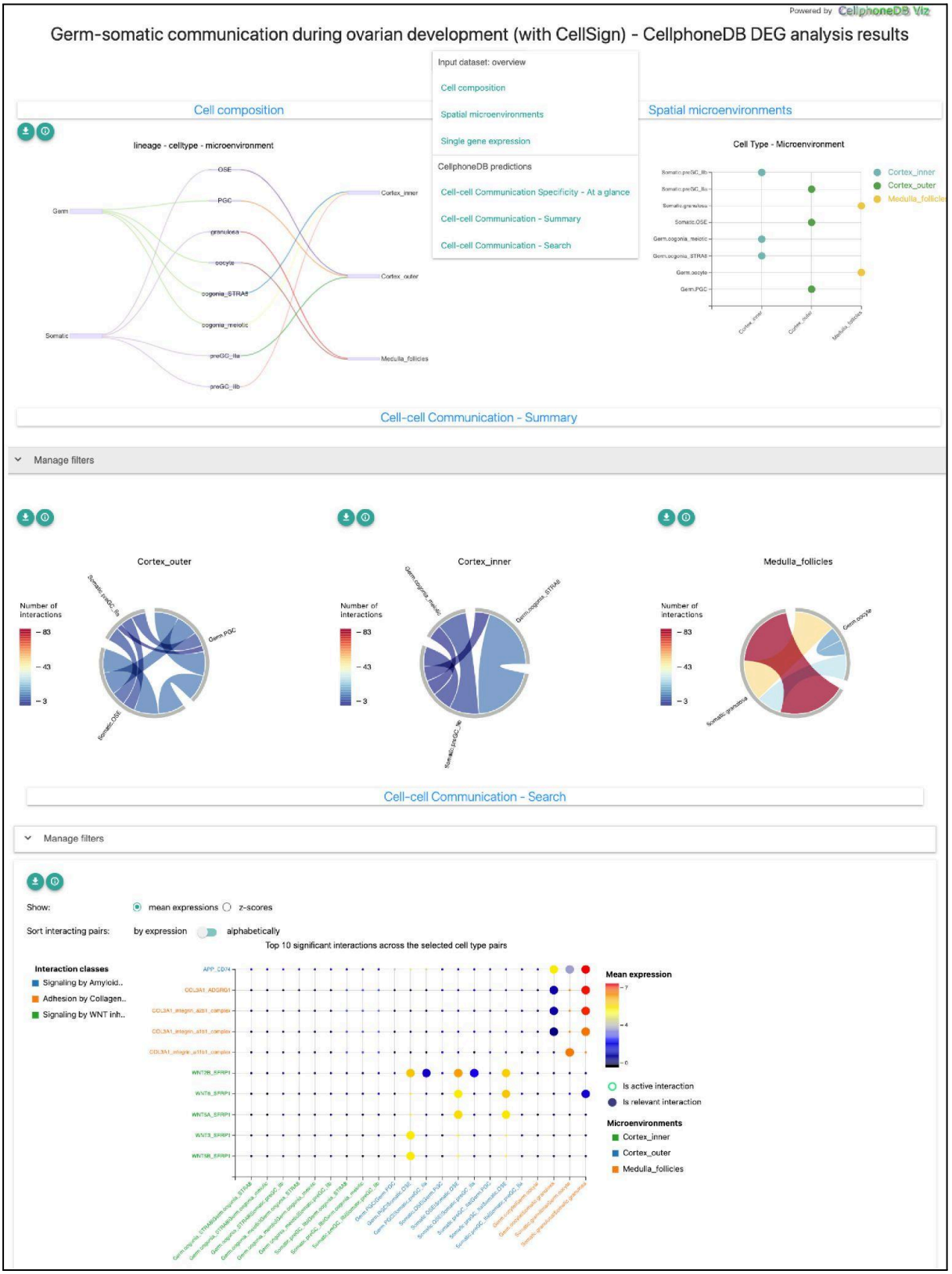
3 Visualise results with ktplotspy

```
kpy.plot_cpdb_heatmap(pvals = cpdb_results['significant_means'],
                      degs_analysis=True,
                      figsize=(4, 4),
                      title="Sum of significant interactions")
from plotnine import facet_wrap

p = kpy.plot_cpdb(
    adata = adata,
    cell_type1 = "PV MMP11|PV AOC3|PV STEAP4",
    cell_type2 = "VCT_CCC|EVT_1|EVT_2|GC|eEVT|iEVT",
    means = cpdb_results['means'],
    pvals = cpdb_results['pvalues'],
    celltype_key = "cell_labels",
#     genes = ["JAG1"],
    figsize = (17,50),
    title = "Interactions between PV \nand trophoblast in the maternal decidua",
    max_size = 5,
    highlight_size = 0.75,
    degs_analysis = False,
    standard_scale = False,
    interaction_scores = cpdb_results['interaction_scores'],
    scale_alpha_by_interaction_scores=True,
)
```

```
p + facet_wrap("~ classification", ncol = 2, scales='free_y')
```

Supplementary Figures.



Supplementary Figure 1. CellPhoneDBViz web overview. Screenshots of CellPhoneDBViz reporting CellPhoneDB results for the Case Study 1 (Supplementary Note 1). **a,** Stankey plot and dotplot depicting cell types (second row or y-axis) per microenvironment (thirds row or x-axis). **b,** Chord plots showing the total number of interactions identified for each cell type pair in each microenvironment (same data coils be represented as heatmaps, feature that is customisable in

CellPhoneDBViz). **c,** Mean expression for a subset of selected interactions (y-axis) per cell type pair (x-axis). If the interaction involves a receptor which downstream TF is active (CellSign) then an outer green ring is shown. Cell type pairs are coloured according to their microenvironment definition.

Supplementary Tables.

Output file	Content	Method specificity
means	Contains mean expression values for each interaction in each pair of cell types, where each row represents a molecular interaction, and each column represents a cell type pair.	`Methods 1-3`
pvalues	Follows the same format as "means" and contains the mean's associated p-values derived from Method 2.	`Method 2`
significant_means	Follows the same format as "means" and contains mean expression values for significant interactions. An interaction is considered significant if it gets a p-value < 0.05 in the "pvalues" file.	`Method 2`
relevant_interactions	Follows the same format as "significant_means" but in this case the values indicate whether an interaction is relevant (1) or not (0), based on Method 3. An interaction is relevant and will get "1" if at least one of the partners is differentially expressed and the other is expressed in more than k% of the cells.	`Method 3`
deconvoluted	Provides additional information about the expression of each individual gene (not just interactions) in each cell type pair.	`Methods 1-3`
percentages	Offers further information about the percentage of cells expressing the gene in each type, useful for tracking the expression of genes in heteromers.	`Methods 1-3`
CellSign (CellSign_active_interactions and CellSign_active_interactions_deconvoluted)	Follow the same format and information as "significant_means" or "relevant_interactions". Here interactions not supported by downstream TF activation will get a value=0. These outputs are generated only if TF data is provided via CellSign. CellSign_active_interactions.txt: this file denotes sets of relevant interactions that have been found to have an active TF downstream of the receptor. CellSign_active_interactions_deconvoluted.txt: deconvoluted means expression value for the transcript encoding each protein for those partners for which an active TF is found.	`CellSign module`
Interaction_scores	Contains the results from the cell type pair specificity scoring methodology. Reports the interaction specificity score for each interaction in each cell type pair.	`Specificity scoring (score_interactions = True)`

Supplementary Table 1a. Description of the output files. CRITICAL: Interactions are not symmetric. This means that when evaluating a ligand→receptor pair (L→R) between clusters "x" and "y", the expression of partner L is measured in the first cluster (x) and partner R in the second cluster (y). As a result, the comparisons xL→yR and xR→yL are distinct and will yield different means, p-values and scores.

Identifier	Definition	Output file	Example
id_cp_interaction	Unique CellPhoneDB identifier for each interaction stored in the database.	CellSign_active_interactions.txt; CellSign_active_interactions_deconvoluted.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	CPI-SS0A28DCA72
interacting_pair	Name of the interacting pairs separated by '_'	CellSign_active_interactions.txt; CellSign_active_interactions_deconvoluted.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	CXCL12_CXCR4
partner a or b	Identifier for the first interacting partner (A) or the second (B). It could be: UniProt (prefix 'simple:') or complex (prefix 'complex:').	CellSign_active_interactions.txt; CellSign_active_interactions_deconvoluted.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	simple:P48061
gene a or b	Gene identifier for the first interacting partner (A) or the second (B). The identifier will depend on the input user list.	CellSign_active_interactions.txt; CellSign_active_interactions_deconvoluted.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	CXCL12
secreted	True if one of the partners is secreted	CellSign_active_interactions.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	TRUE
receptor a or b	True if the first interacting partner (A) or the second (B) is annotated as a receptor in our database.	CellSign_active_interactions.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	FALSE
annotation_strategy	Curated if the interaction was annotated by the CellPhoneDB developers. Otherwise, the name of the database where the interaction has been downloaded from.	CellSign_active_interactions.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	curated
is_integrin	True if one of the partners is an integrin.	CellSign_active_interactions.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	FALSE
directionality	Directionality annotation as denoted in interaction_input file.	CellSign_active_interactions.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	Ligand-Receptor
classification	Pathway classification as denoted in interaction_input file.	CellSign_active_interactions.txt; interaction_scores.txt;relevant_interactions.txt;significant_interactions.txt	Signaling by Chemokines
active_TF	Denotes the active TF downstream the receptor.	CellSign_active_interactions_deconvoluted.txt	STAT1
celltype_pairs	Denotes the pair of interacting cells for which the receptor-TF interaction is found active.	CellSign_active_interactions_deconvoluted.txt	PV MMP11|EVT_1
active_celltype	Cell type expressing the active transcription factor.	CellSign_active_interactions_deconvoluted.txt	EVT_1
active	Denotes whether downstream the receptor the transcription factor is active (1) or not (.)	CellSign_active_interactions.txt	1
score	Score generated by the scoring methodology.	interaction_scores.txt	78.23
relevancy	For method 3, denotes whether the interaction has been found relevant (1) or not (0) based on the differential expression of the interacting partners.	relevant_interactions.txt	1

P-value	For method 2, denotes the P-value obtained after random shuffling procedure.	significant_interactions.txt	0.05

Supplementary Table 1b. Description of the meaning of the columns in the output file.