UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FARMACOLOGIA

Marcelo Luiz Brunatto Falchetti

***SpotComm*: um novo algoritmo para caracterização da comunicação celular em transcriptômica espacial**

Florianópolis
2022

Marcelo Luiz Brunatto Falchetti

SpotComm**:** um novo algoritmo para caracterização da comunicação celular em transcriptômica espacial

Tese submetida ao Programa de Pós-Graduação em Farmacologia da Universidade Federal de Santa Catarina para a obtenção do título de Doutor em Farmacologia
Orientador: Prof. Dr. Alfeu Zanotto-Filho
Coorientador: Prof. Dr. Edroaldo Lummertz da Rocha

Florianópolis
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Falchetti, Marcelo Luiz Brunatto
SpotComm: um novo algoritmo para caracterização da comunicação celular em transcriptômica espacial / Marcelo Luiz Brunatto Falchetti ; orientador, Alfeu Zanotto-Filho, coorientador, Edroaldo Lummertz da Rocha, 2022.
145 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Programa de Pós Graduação em Farmacologia, Florianópolis, 2022.

Inclui referências.

1. Farmacologia. 2. Bioinformática. 3. Comunicação celular. 4. Transcriptômica espacial. I. Zanotto-Filho, Alfeu. II. Lummertz da Rocha, Edroaldo. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Farmacologia. IV. Título.

Marcelo Luiz Brunatto Falchetti

***SpotComm*: um novo algoritmo para caracterização da comunicação celular em transcriptômica espacial**

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Tiago Da Silva Medina, Dr.

Instituição Fundação Antônio Prudente/A.C. Camargo Cancer Center

Prof. Daniel Fernandes, Dr.

Instituição Universidade Federal de Santa Catarina

Gabriela Flavia Rodrigues Luiz, Dra.

Instituição Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em Farmacologia.

Documento assinado digitalmente
Jose Eduardo da Silva Santos
Data: 24/10/2022 08:25:54-0300
CPF: ***.948.289-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. José Eduardo da Silva Santos

Coordenador do Programa de Pós-Graduação

Documento assinado digitalmente
Alfeu Zanotto Filho
Data: 21/10/2022 10:56:57-0300
CPF: ***.553.300-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. Alfeu Zanotto-Filho

Orientador

Documento assinado digitalmente
EDROALDO LUMMERTZ DA ROCHA
Data: 21/10/2022 10:48:26-0300
CPF: ***.282.360-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. Edroaldo Lummertz da Rocha

Coorientador

Florianópolis, 2022.

Este trabalho é dedicado à minha mãe, Sônia Falchetti e ao meu marido, Valdriano Lemos Polla e nosso cachorro, Jon.

AGRADECIMENTOS

Agradeço ao meu orientador Prof. Dr. Alfeu Zanotto-Filho e coorientador Prof. Dr. Edroaldo Lummertz da Rocha pelo ensino, paciência, confiança e orientação durante os meus anos de doutorado.

À banca avaliadora, Prof. Dr. Tiago Da Silva Medina, Prof. Dr. Daniel Fernandes e a Dr.ª Gabriela Flavia Rodrigues Luiz pela disponibilizade de examinar esse trabalho. Agradeço por poder apresentar e defender esse trabalho para profissionais que tanto respeito.

Ao Programa de Pós-graduação em Farmacologia por ter me recebido e oportunizado tantas conjunturas de crescimento social e intelectual e à CAPES e CNPq pela indispensável bolsa de estudos.

A todos os meus amigos do LabCancer pelo carinho, atenção, aprendizado e companherismo nos momentos mais difíceis e mais prazerosos da minha formação no doutorado. Obrigado Barbara dos Santos, Helena Zancanaro, Jonathan Agnes, Karoline Almeida, Marina Delgobo, Rosângela Gonçalves e Raquel Nascimento das Neves do LabCancer.

Obrigado meus amigos e irmãos científicos do doutorado por terem me auxiliado a continuar e chegar nesse ponto, sempre me alegrando e estimulando.

Aos professores, por incitarem minha admiração por essa linda área da ciência. Profissionais que sempre estiveram dispostos a ensinar e crescer juntos. Especialistas que, apaixonados, nos apaixonam diariamente.

A todos os funcionários da Universidade Federal de Santa Catarina que auxiliam e possibilitam a preservação e o exercício dessa instituição. Em especial agradeço os servidores, os profissionais da área da limpeza e do Restaurante Universitário.

Ao meu marido Valdriano Polla pelo amor, calma e apoio em todas as minhas escolhas. À minha família pelo amor incondicional, por acreditarem e me estimularem a acreditar que sou capaz de realizar os meus sonhos. Amo todos vocês!

A todos aqueles que contribuiram e aos que ainda contribuem na minha formação como pessoa, como profissional e na minha saúde física e mental. A todos os presentes na minha história. Muito obrigado!

“DREAM! Dreams shape the world. Dreams create the world anew, every night.”
(Neil Gaiman em *The Sandman* #18, 1990)

RESUMO

Uma vez que a comunicação celular, ao menos de forma autócrina e parácrina é espacialmente limitada, os métodos de transcriptômica espacial podem aumentar a confiabilidade de comunicações preditas. Os autores de análise de comunicação celular vem pensando nisso e desenvolvendo algoritmos que aproveitam a informação espacial em suas predições. Entretanto, os algoritmos de análise de comunicação celular em dados de transcriptômica espacial até agora focavam na análise de populações de *spots*, as análises baseadas em *clusters*, sejam *clusters* transcricionais ou regiões teciduais, e, dito isso, não examinam as particularidades de comunicação do transcriptoma do *spot*, que, dependendo da tecnologia, pode ser uma célula ou um pequeno grupo de células. Com base nisso, nós desenvolvemos um algoritmo de predição de comunicações celulares que utiliza os dados de coordenadas, de *spots*, como unidade de análise, o *SpotComm*. Esse algoritmo possui funções capazes de definir as comunicações intercelulares, intracelulares e sinalizações intracelulares se baseando na presença e na correlação de transcritos em dados de transcriptômica espacial e é capaz de integrar dados de transcriptômica espacial com dados pareados ou não pareados de scRNA-seq. O *SpotComm* é capaz de manipular dados de transcriptômica espacial em análises de uma seção (2D) ou de mais seções (3D), oferecer análises de *clusters* transcricionais inteiros e/ou zonas de contato entre *clusters*, e proporcionar cenários de comunicação e sinalização entre proteínas interatoras monoméricas e complexos proteicos homomultiméricos e heteromultiméricos. O algoritmo é capaz de gerar dados e oferecer metadados ao usuário como a presença, proporção e expressão de interatores e a referência e curagem de comunicações e sinalizações. Ainda, oferece a proporção de células e proporção de co-ocorrência, por tipo celular, com detecção dos elementos de vias de sinalização intracelular. Utilizando o *SpotComm* nós fomos capazes de predizer comunicações que são conhecidas em estruturas linfoides terciárias e em áreas de tumor com o perfil de respostas de interferon tipo I em câncer de mama e predizemos os elementos celulares viáveis às sinalizações. Ainda, detectamos diversas potenciais comunicações inter e intracelulares não detectadas em análise baseada em *clusters* que podem ser importantes no entendimento da homeostase mas também nos prognósticos, diagnósticos e tratamentos de doenças.

Palavras-chave: Bioinformática. Interação celular. Geração de hipóteses. Predição. Análise ômica.

ABSTRACT

Since cell communication, at least in an autocrine and paracrine fashion is spatially limited, spatial transcriptomics methods can increase the reliability of predicted communications. Cell communication analysis authors have been thinking about this and developing algorithms that take advantage of spatial information in their predictions. However, algorithms for analyzing cell communication in spatial transcriptomics data have so far focused on the analysis of spot populations, the cluster-based analyses, whether transcriptional clusters or tissue regions, and, that said, do not examine the communication particularities of the spot transcriptome, which, depending on the technology, can be one cell or a small group of cells. Based on this, we have developed a cell communication prediction algorithm that uses coordinate data, of spots, as the unit of analysis, the *SpotComm*. This algorithm has functions capable of defining intercellular, intracellular communications and intracellular signaling based on the presence and correlation of transcripts in spatial transcriptomics data and is able to integrate spatial transcriptomics data with paired or unpaired scRNA-seq data. *SpotComm* is capable of handling spatial transcriptomics data in single-section (2D) or multi-section (3D) analyses, provide analyses of entire transcriptional clusters and/or contact zones between clusters, and provide scenarios of communication and signaling between monomeric interacting proteins and homomultimeric and heteromultimeric protein complexes. The algorithm is able to generate data and provide metadata to the user such as the presence, proportion and expression of interactors and the reference and curation of communication and signaling. It also provides the proportion of cells and proportion of co-occurrence, by cell type, with detection of the elements of intracellular signaling pathways. Using *SpotComm* we were able to predict communications that are known in tertiary lymphoid structures and in tumor areas profiled by type I interferon responses in breast cancer and predicted the cellular elements viable for signaling. Furthermore, we detected several potential inter- and intracellular communications not detected in cluster-based analysis that may be important in understanding homeostasis but also in disease prognosis, diagnosis and treatment.

Keywords: Bioinformatics. Cellular interaction. Hypothesis generation. Prediction. Omics analysis.

LISTA DE FIGURAS

Figura 1. Principais categorias de métodos de transcriptômica espacial.....22
Figura 2. Elementos de comunicação inter, intracelular e sinalização utilizados por algoritmos.....25
Figura 3. *Hexagon sticker/icon* do *SpotComm.*.....40
Figura 4. Fluxograma de funções, *outputs* e parâmetros do *SpotComm*.....42
Figura 5. Visão geral das funções *get_subset_2D()* e *get_subset_3D()*.....43
Figura 6. Visão geral das funções *get_nearby_2D()* e *get_nearby_3D()*.....45
Figura 7. Visão geral da função define_intercellular_comm*()*.....47
Figura 8. Visão geral da função *buzzer_intercellular_comm()*.....49
Figura 9. Visão geral da função *buzzer_intracellular_comm()*.....53
Figura 10. Visão geral da função *linker_intracellular_comm()*.....57
Figura 11. Visão geral da função *integr_intracellular_comm()*.....60
Figura 12. Visão geral da função *integr_intracellular_comm()*.....62
Figura 13. Região do tecido analisada usando Visium corada com hematoxilina-eosina.....69
Figura 14. Especificidade de assinaturas de tópicos obtidos do *SPOTlight* para cada identidade celular.....71
Figura 15. Áreas de análise transcricional definidas por seu *cluster* transcricional.. .72
Figura 16. Áreas de análise transcricional do *cluster* de “estruturas linfoides terciárias”.....73
Figura 17. Pares de ligantes e receptores com os maiores valores de escore de comunicação.....75
Figura 18. Fatores de transcricão com maior proporção dos alvos detectados.....76
Figura 19**.** Vias de sinalização estimulatórias preditas entre HLA-E e os fatores de transcrição presentes em mais *spots* com base nos transcritos.....77
Figura 20. Pares de ligantes e receptores presentes em mais comunicações entre *spots*.....79
Figura 21. Distribuição de comunicações intercelulares mais autocorrelacionadas. 80
Figura 22. Fatores de transcrição operantes presentes em mais comunicações entre *spots*.....82
Figura 23. Distribuição de comunicações intracelulares mais autocorrelacionadas. 83
Figura 24. Exemplo de diferenças em vias de sinalização preditas de dois *spots*.....84
Figura 25. Elementos celulares viáveis de vias de sinalização preditas.....85

Figura 26. Região do tecido analisada usando Visium corada com hematoxilina-eosina.....88
Figura 27. Especificidade de assinaturas de tópicos obtidos do *SPOTlight* para cada identidade celular.....90
Figura 28. Áreas de análise transcricional definidas por seu *cluster* transcricional.. .91
Figura 29. Áreas de análise transcricional do *cluster* de “respostas de interferon tipo I”.....92
Figura 30. Pares de ligantes e receptores com os maiores valores de escore de comunicação.....94
Figura 31. Fatores de transcricão com maior proporção dos alvos detectados.....96
Figura 32**.** Vias de sinalização estimulatórias preditas entre HLA-E e os fatores de transcrição presentes em mais *spots* com base nos transcritos.....97
Figura 33. Pares de ligantes e receptores presentes em mais comunicações entre *spots*.....99
Figura 34. Distribuição de comunicações intercelulares mais autocorrelacionadas. 100
Figura 35. Fatores de transcrição operantes presentes em mais comunicações entre *spots*.....101
Figura 36. Distribuição de comunicações intracelulares mais autocorrelacionadas. 102
Figura 37. Exemplo de diferenças em vias de sinalização preditas de dois *spots.*. 103
Figura 38. Elementos celulares viáveis de vias de sinalização preditas.....104

LISTA DE ABREVIATURAS E SIGLAS

2D: duas dimensões.

3D: três dimensões.

DNA: Ácido desoxirribonucleico.

cDNA: DNA complementar.

HER2: Receptor do fator de crescimento epidermal tipo 2.

RNA: Ácido ribonucleico.

mRNA: RNA mensageiro.

RNA-seq: Tecnologia de sequenciamento de RNA.

scRNA-seq: Tecnologia de sequenciamento de RNA em células únicas.

Nomes de genes e de algoritmos (quando siglas) não estão contidos nessa lista.

SUMÁRIO

1 INTRODUÇÃO.....17
1.1 TRANSCRIPTÔMICA.....17
1.2 COMUNICAÇÃO CELULAR.....22
1.3 ALGORITMOS DE COMUNICAÇÃO CELULAR EM TRANSCRIPTÔMICA.....24
1.4 ALGORITMOS DE COMUNICAÇÃO CELULAR EM TRANSCRIPTÔMICA E O USO DAS COORDENADAS ESPACIAIS.....34
1.5 JUSTIFICATIVA.....36
2 HIPÓTESE.....38
3 OBJETIVOS.....39
3.1 OBJETIVO GERAL.....39
3.2 OBJETIVOS ESPECÍFICOS.....39
4 MATERIAL E MÉTODOS.....40
4.1 FUNÇÕES E PARÂMETROS DO SPOTCOMM.....40
4.1.1 *get_subset_2D()*.....41
4.1.2 *get_subset_3D()*.....44
4.1.3 *get_nearby_2d()*.....44
4.1.4 *get_nearby_3d()*.....45
4.1.5 *define_intercellular_comm()*.....46
4.1.6 *filter_define_intercellular_comm()*.....47
4.1.7 *buzzer_intercellular_comm()*.....48
4.1.8 *filter_buzzer_intercellular_comm()*.....52
4.1.9 *buzzer_intracellular_comm()*.....52
4.1.10 *filter_buzzer_intracellular_comm()*.....55
4.1.11 *linker_intracellular_sign()*.....56
4.1.12 *filter_linker_intracellular_sign()*.....59
4.1.13 *integr_intracellular_sign()*.....60
4.1.14 *filter_integr_intracellular_sign()*.....62
4.2 ESTUDO DE CASO - PREPARAÇÃO DE DADOS DE ENTRADA.....63
5 RESULTADOS.....66
5.1 ESTUDO DE CASO 1: CÂNCER DE MAMA DO SUBTIPO MOLECULAR TRIPLO-NEGATIVO.....68
5.1.1 Definição de *clusters* transcricionais do paciente “B” de Wu e colaboradores, em 2021.....68

5.1.2 Comunicação celular interna do *cluster* transcricional de "estruturas linfóides terciárias" em análise de forma conjunta (*mode* = *"spot_n"*).....71
5.1.3 Comunicação celular interna do *cluster* transcricional de "estruturas linfóides terciárias" em análise de forma individual (*mode* = *"spot_y"*).....77
5.2 ESTUDO DE CASO 2: CÂNCER DE MAMA DO SUBTIPO MOLECULAR HER2-POSITIVO.....87
5.2.1 Definição de *clusters* transcricionais do paciente "G" de Andersson e colaboradores, em 2021.....87
5.2.2 Comunicação celular interna do *cluster* transcricional de "respostas de interferon tipo I" em análise de forma conjunta (*mode* = *"spot_n"*).....91
5.2.3 Comunicação celular interna do *cluster* transcricional de "respostas de interferon tipo I" em análise de forma individual (*mode* = *"spot_y"*).....96
6 DISCUSSÃO.....106
7 PRINCIPAIS ACHADOS.....127
8 CONCLUSÃO E PERSPECTIVAS.....128
8.1 CONCLUSÃO.....128
8.2 PERSPECTIVAS.....128
REFERÊNCIAS.....130
APÊNDICE A - GENES MARCADORES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS"
APÊNDICE B – TERMOS ENRIQUECIDOS NO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS"
APÊNDICE C - LIGANTES E RECEPTORES UTILIZADOS NA ANÁLISE DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS"
APÊNDICE D - TERMOS ENRIQUECIDOS COM OS LIGANTES E RECEPTORES UTILIZADOS NA ANÁLISE DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS"
APÊNDICE E - COMUNICAÇÕES INTERCELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *CLUSTER*
APÊNDICE F - COMUNICAÇÕES INTRACELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *CLUSTER*

APÊNDICE G - SINALIZAÇÕES INTRACELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *CLUSTER*
APÊNDICE H - COMUNICAÇÕES INTERCELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE I – INDÍCEIS DE MORAN DE COMUNICAÇÕES INTERCELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE J - COMUNICAÇÕES INTRACELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE K – ÍNDICES DE MORAN DE COMUNICAÇÕES INTRACELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE L – SINALIZAÇÕES INTRACELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE M – SINALIZAÇÕES INTRACELULARES DO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" NA ANÁLISE BASEADA EM *SPOTS* PÓS INTEGRAÇÃO COM SCRNA-SEQ
APÊNDICE N - TERMOS ENRIQUECIDOS NO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" COM LIGANTES E RECEPTORES DETECTADOS SOMENTE NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE O – TERMOS ENRIQUECIDOS NO *CLUSTER* DE "ESTRUTURAS LINFOIDES TERCIÁRIAS" COM ALVOS DE FATORES DE TRANSCRIÇÃO ATIVOS DETECTADOS SOMENTE NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE P - GENES MARCADORES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I"
APÊNDICE Q – TERMOS ENRIQUECIDOS NO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I"
APÊNDICE R - LIGANTES E RECEPTORES UTILIZADOS NA ANÁLISE DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I"
APÊNDICE S - TERMOS ENRIQUECIDOS COM OS LIGANTES E RECEPTORES UTILIZADOS NA ANÁLISE DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I"
APÊNDICE T - COMUNICAÇÕES INTERCELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *CLUSTER*

APÊNDICE U - COMUNICAÇÕES INTRACELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *CLUSTER*
APÊNDICE V - SINALIZAÇÕES INTRACELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *CLUSTER*
APÊNDICE W - COMUNICAÇÕES INTERCELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE X – INDÍCEIS DE MORAN DE COMUNICAÇÕES INTERCELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE Y - COMUNICAÇÕES INTRACELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE Z – ÍNDICES DE MORAN DE COMUNICAÇÕES INTRACELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE AA - SINALIZAÇÕES INTRACELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE AB – SINALIZAÇÕES INTRACELULARES DO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" NA ANÁLISE BASEADA EM *SPOTS* PÓS INTEGRAÇÃO COM SCRNA-SEQ
APÊNDICE AC - TERMOS ENRIQUECIDOS NO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" COM LIGANTES E RECEPTORES DETECTADOS SOMENTE NA ANÁLISE BASEADA EM *SPOTS*
APÊNDICE AD - TERMOS ENRIQUECIDOS NO *CLUSTER* DE "RESPOSTAS DE INTERFERON TIPO I" COM ALVOS DE FATORES DE TRANSCRIÇÃO ATIVOS DETECTADOS SOMENTE NA ANÁLISE BASEADA EM *SPOTS*

1 INTRODUÇÃO

1.1 TRANSCRIPTÔMICA

Organismos multicelulares geralmente consistem em tecidos, órgãos e sistemas especializados em um conjunto de funções/processos biológicos realizados por atividades coordenadas de muitas células. Todas as células do organismo multicelular apresentam o mesmo material genético, o mesmo DNA. Entretanto, regulações genéticas e epigenéticas organizam a transcrição de genes e, como resultado, a tradução de proteínas, tornando as células distintas em suas estruturas e funções, ou possibilidades, biológicas (BUCCITELLI; SELBACH, 2020). A transcrição de genes, ou transcrição gênica, se altera durante a vida celular, por exemplo quando a célula se diferencia em um subtipo celular mais especializado como quando a célula responde a estímulos. Muitas das alterações transcricionais são fisiológicas, todavia processos patológicos também podem gerar alterações na expressão de genes.

A transcrição gênica geralmente é definida pela produção de RNA mensageiro (mRNA) e devido a sua relevância no programa da vida os pesquisadores de áreas diversas buscam analisar qualitativamente e quantitativamente as suas moléculas. A detecção de transcritos começou em 1977 quando James Alwine e colaboradores desenvolveram a técnica de *northern blot* (ALWINE; KEMP; STARK, 1977) e desde aquele tempo um grande avanço tecnológico tem sido visto, com a criação da técnica de reação em cadeia da polimerase (PCR do inglês *polymerase chain reaction*) de Kary Mullis (MULLIS; FALOONA, 1987; MULLIS, 1990), da PCR de transcrição reversa e a sintetização de DNA complementar (cDNA) para estabilizar RNAs (HIGUCHI et al., 1992, 1993), da PCR quantitativa de transcrição reversa (LIVAK; SCHMITTGEN, 2001) e derivados que permitiam a detecção e investigação de um ou de poucos RNAs diferentes por vez até o surgimento das técnicas de transcriptômica. Hoje transcriptomas, ou seja, o conjunto completo de transcritos de genes ou espécies de RNA transcritos (como os microRNAs e os RNAs longos não-codificantes, lncRNAs), podem ser

identificados e quantificados em tecidos e em células, sem e com resolução espacial em um experimento.

A primeira tecnologia na análise de transcriptomas bastante divulgada/difundida na pesquisa é o microarranjo de DNA (do inglês *DNA microarray*), ou *DNA chip* (FODOR et al., 1991; PEASE et al., 1994). Essa tecnologia se baseia no isolamento de moléculas de RNA de diferentes tecidos, a transcrição reversa e a marcação das moléculas com um fluoróforo, prata ou quimioluminescência (BUMGARNER, 2013). Essas moléculas se conectam as sondas de DNA dispostas em spots em uma superfície sólida ou membrana, o microarranjo de DNA. Sendo assim, em cada *spot*, em cada coordenada da superfície, há sequências desenhadas e específicas que podem ser seções pequenas de um gene ou de um outro elemento de DNA que será utilizado para hibridização de alvos. A hibridização é detectada e quantificada pela detecção de alvos marcados e dessa forma é possível determinar a abundância relativa de sequências de RNAs . Uma vez que é necessário o conhecimento prévio de sequências de sondas, a pesquisa é limitada a genes conhecidos e a detecção de transcritos formados de splicing alternativo é difícil. (KING; SINHA, 2001). Além do mais, a detecção e quantificação dependente de marcação, ou semi-quantificação, tem baixa precisão a mudanças tênues e a genes com alta e com baixa expressão.

O próximo passo foi a tecnologia do sequenciamento de RNA, ou RNA-seq, que utiliza o sequenciamento de próxima-geração (NGS do inglês *next-generation sequencing*) para caracterizar a presença, a “quantidade” (na profundidade de leituras) e as sequências de RNAs (MARGULIES et al., 2005; MORTAZAVI et al., 2008; NAGALAKSHMI et al., 2008). Há dois passos principais nessa tecnologia, a preparação da biblioteca de sequenciamento e o sequenciamento. Para a preparação da biblioteca de sequenciamento se deve (importante: existem diferentes métodos e requerimentos) isolar o RNA da amostra e “quebrar” as suas moléculas em pequenos fragmentos de 200 a 300 pares de bases (KUKURBA; MONTGOMERY, 2015). Após, deve-se estabilizar as moléculas de RNA as convertendo em cDNA e adicionar adaptadores à sequências. Adaptadores permitem às máquinas reconhecer as sequências/fragmentos e podem permitir o sequenciamento de mais de uma única amostra por vez. Os próximos passos

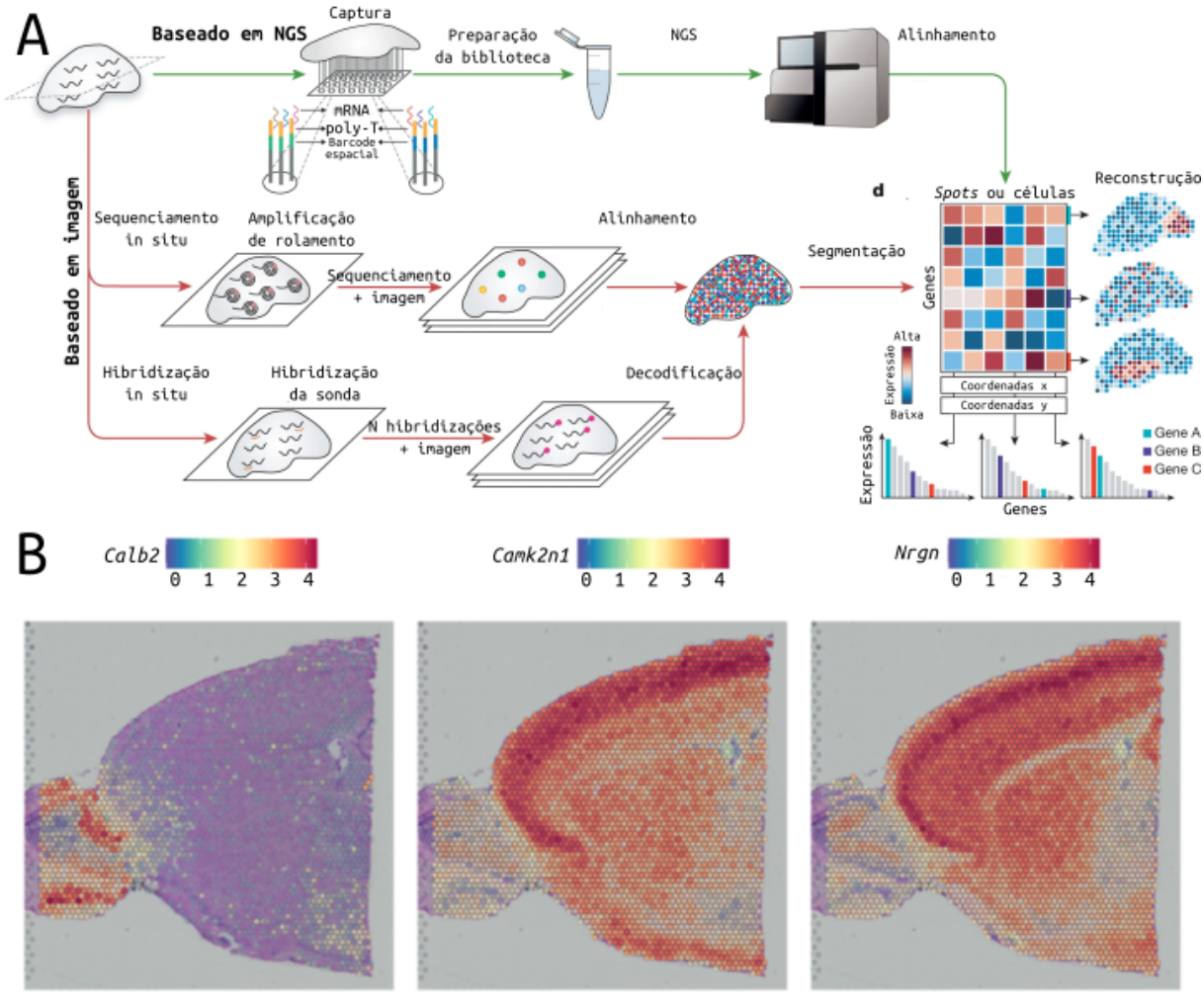
envolvem a amplificação, via PCR, dos fragmentos com adaptadores e o controle de qualidade de sequências. Nesse ponto se pode sequenciar a biblioteca. No sequenciador os fragmentos com adaptadores serão dispostos em uma rede denominada como *flow cell* e sequenciados base-a-base pela interação do nucleotídeo com uma sonda codificada por cor, dependendo do nucleotídeo em que ela se liga. No primeiro *round* do sequenciador as sondas se ligam na primeira base de todas as sequências e o sequenciador "tira uma foto" do *flow cell*. O sequenciador "lava"/retira as marcações de sondas e então novas sondas se ligam na segunda base de todas as sequências e o sequenciador "tira mais uma foto". Esse processo é repetido até que se tenham as sequências resolvidas. Dessa forma, o sequenciador considera a posição e a cor das sequências com sondas. O RNA-seq é uma técnica independente do conhecimento prévio de sequências e a detecção de transcritos formados de *splicing* alternativo é possível, em especial com um bom genoma ou transcriptoma de referência (WANG; GERSTEIN; SNYDER, 2009). Além do mais, com RNA-seq é possível descrever alterações de expressão gênica entre grupos, tempos e tratamentos, modificações pós-transcricionais, mutações, fusão de genes e determinar os limites gênicos de exons, introns, regiões não traduzidas 3' e 5'. Essas análises podem ser efetuadas com transcritos de genes ou espécies de RNA transcritos. Todas essas possibilidades com tal técnica a fazem popular na ciência, com 28.457 conjuntos de dados de RNA-seq e perfil de expressão por sequenciamento de alta taxa de transferência (do inglês *expression profiling by high throughput sequencing*) depositados só no Gene Expression Omnibus, o GEO, do Centro Nacional de Informações sobre Biotecnologia (NCBI do inglês *National Center for Biotechnology Information*) e 43.995 artigos científicos depositados no *PubMed* citando a técnica. Entretanto, métodos como o microarranjo de DNA e o RNA-seq geralmente analisam os RNAs de populações de células, a chamada análise de massa, ou *bulk*. Em populações de células misturadas, avalia-se a a expressão média de todas as células e as diferenças de células individuais não são contabilizadas. Assim sendo, as análises de *bulk* falham em identificar se, por exemplo, a alteração no perfil de expressão é devido a modificação regulatória ou composição celular, onde um tipo celular está mais presente nas amostras.

Com base nisso, ainda no nível de transcriptômica baseada em sequenciamento de próxima-geração, em 2009 Tang e colaboradores traçaram a primeira análise de células individuais, o scRNA-seq (TANG et al., 2009). Após esse trabalho, estudos mostraram que a expressão de genes em células da mesma população/tipo celular é variada, heterogênea, e que podem demonstrar papéis diferentes e formar subpopulações com perfis transcricionais distintos. Estudar os transcriptomas de células individuais nos propicia elucidar as funções celulares em processos biológicos inesperados que podem ser basilares na compreensão de contrapartes saúde/doença (VALLEJO et al., 2021; WU et al., 2021a). Para isso, para essa técnica, um passo chave no processo experimental é a criação da suspensão de células individuais. Os primeiros métodos separavam células individuais em poços separados. Hoje se encapsulam células individuais em “gotas” (do inglês *droplets*) e microesferas em dispositivo microfluídico e nesses *droplets* se realizam a lise celular e ligação dos RNAs aos TAGs das microesferas (HAQUE et al., 2017). Os TAGs contém *barcodes*, "código de barras", de DNA que servem para marcar exclusivamente os RNAs derivados de cada célula. Dentro dos *droplets* são realizadas as reações de síntese do cDNA, os droplets são então “desfeitos” e são realizadas as reações de amplificação. Após a amplificação os cDNAs de células são sequenciados, como no RNA-seq, mapeados em um genoma ou transcriptoma de referência e analisados. Avanços tecnológicos viabilizam a mensuração da expressão gênica de milhares, até mesmo milhões, de células individuais simultaneamente e, juntamente a isso, a criação de atlas de células e estados celulares de diferentes espécies como os 20 representantes dos reinos dos animais, fungos, plantas e protistas de 304 estudos totalizando ~ 8.5 milhões de células no Single Cell Expression Atlas (*www.ebi.ac.uk/gxa/sc/home*). Exclusivamente no *site* do Single Cell Portal estão depositados 430 estudos com scRNA-seq contendo mais de 20 milhões de células sequenciadas e no *PubMed* há 2.575 artigos científicos depositados explorando a variabilidade célula-célula. Entretanto, a criação da suspensão de células individuais usando etapas de dissociação mecânica e enzimática pode potencialmente causar expressão gênica ectópica/anormal sendo capaz de resultar em caracterização imprecisa e/ou incorreta de certas

subpopulações celulares e, ainda, desfazem a disposição e a organização tecidual tão primordial nos organismos multicelulares.

Até recentemente, técnicas de alto rendimento (do inglês *high throughput*) não podiam ser empregadas *in situ*, resultando assim na perda de informação sobre relações espaciais. Os métodos de análise de transcriptoma com informação espacial empregavam séries de fatias/cortes de tecido como os sequenciamentos de microtomia serial (Tomo-seq) (COMBS; EISEN, 2013; JUNKER et al., 2014; LACRAZ et al., 2017), microdissecções (TIROSH et al., 2016), dissociação parcial de tecidos (ProximID) (BOISSET et al., 2018), *cell sorting* de células em interações físicas seguido de scRNA-seq (Pic-seq) (GILADI et al., 2020), os sequenciamentos de amontoados de células (ClumpSeq) (MANCO et al., 2021), o mapeamento de grupo de genes para inferir localização celular em dados de scRNA-seq (PETTIT et al., 2014; ACHIM et al., 2015; SATIJA et al., 2015) entre outros visando reconstituir um eixo espacial. Atualmente vemos técnicas/métodos que reuniram o mantimento da informação espacial e as tecnologias de transcriptômica, as técnicas de transcriptômica espacial . De forma coletiva, os métodos de transcriptômica espacial foram nomeados método do ano de 2020 na revista *Nature Methods* pelo reconhecimento da relevância destes e é esperado que estes métodos transformem rapidamente a análise biológica (MARX, 2021). As tecnologias de transcriptômica espacial variam quanto ao número de genes que podem ser sondados e a dimensão tecidual a ser analisada. Essas tecnologias podem ser categorizadas em 1 - baseadas em sequenciamento de próxima-geração onde há codificação de informação espacial/posicional de transcritos (com *barcodes*) antes da etapa de sequenciamento (Figura 1-A); e 2 - abordagens de imagem; compondo-se de técnicas baseadas em sequenciamento *in situ*, onde os transcritos são amplificados e sequenciados no próprio tecido (Figura 1-A). Independente da técnica, ao final há a elaboração da matriz de expressão gênica com transcriptomas de cada *spot* (ou seja, *pixel*, célula ou grupo de células, dependendo do tecnologia usada) que pode ser utilizada na avaliação de genes (Figura 1-B), clusters transcricionais e outros (RAO et al., 2021). Essas novas possibilidades de analisar transcriptômica espacial, sem sondas, sem dissociação e com sequenciamento de próxima-geração muito atrai a ciência e, apesar da novidade e dos altos valores, estudos estão sendo

Figura 1. Principais categorias de métodos de transcriptômica espacial



A - Ilustração de tecnologias baseadas em sequenciamento de próxima-geração e baseadas em imagem, separadas em sequenciamento *in situ* e hibridização *in situ*. B – Expressão com correlação espacial dos genes *Calb2* (codificador de Calbindina 2), *Camk2n1* (codificador de Inibidor 1 de proteína quinase II dependente de cálcio/calmodulina) e *Nrgn* (codificador de Neurogranina) no cérebro de camundongos avaliados por transcriptômica espacial utilizando a tecnologia *Visium*.

realizados e há só no GEO 99 conjuntos de dados de transcriptômica espacial e perfil de expressão por sequenciamento de alta taxa de transferência e 155 artigos científicos depositados no *PubMed.gov* citando a técnica.

1.2 COMUNICAÇÃO CELULAR

Embora todas as células normais em organismos multicelulares compartilhem o mesmo genoma, os padrões de expressão gênica e morfologia podem ser muito diferentes. Tal variação não é causada/ativada por circuitarias de regulação de expressão interna somente, mas também por sinalização externa, do ambiente do tecido (ROUAULT; HAKIM, 2012). Ainda que tenhamos décadas de estudos voltados para o entendimento de circuitarias de regulação de expressão características a tipo celular, o nosso conhecimento de interações ambientais, externas, segue limitada. A comunicação celular é geralmente coordenada de modo espacial e tipo celular específica e, sendo assim, o estudo das interações célula-célula, das atividades e principalmente da coordenação das atividades das células precisa do contexto da comunidade de cada célula (ZEPP et al., 2017; BACCIN et al., 2020; NIETHAMER et al., 2020; PARK et al., 2020).

As células se comunicam através de moléculas. Algumas moléculas sustentam as interações celulares de modo estrutural, como as moléculas de adesão, enquanto outras mediam as comunicações celulares, como os hormônios, citocinas, quimiocinas e neurotransmissores. Os eventos de comunicação são geralmente mediados por interações entre ligante-receptor, receptor-receptor e matriz extracelular-receptor (ARMINGOL et al., 2021). As células receptoras acionam a sinalização a jusante (do inglês *downstream*) por receptores geralmente culminando em alteração da atividade transcricional. Então estas células, com alteração de expressão, vão interagir com o seu microambiente.

Na farmacologia, o conhecimento de comunicações celulares é de extrema importância, uma vez que fármacos interagem com moléculas de interação, comunicação intra e extracelulares e sinalização como os receptores e podem vir a estimular a alteração na expressão gênica (SANTOS et al., 2016). Ainda, o conhecimento do efeito de comunicações é considerado na elaboração de hipóteses de mecanismos de ação e na criação de novos fármacos. Um exemplo é o prognóstico, diagnóstico e tratamento do câncer de mama. O câncer de mama é uma condição complexa, heterogênea quanto as suas composições celulares, moleculares e desfechos clínicos (HARBECK et al., 2019; LOIBL et al., 2021). Hoje

se sabe que a diversidade fenotípica de prognose e de predição à eficácia da quimioterapia do câncer de mama está relacionada a diversidade de perfis de expressão gênica. Esse conhecimento é a base da taxonomia molecular do câncer de mama. Dependendo da detecção (ou ausência) de receptores, por exemplo, receptores de estrogênio (ER) e receptores do fator de crescimento epidermal tipo 2 (HER2) é definido a taxa de crescimento, o grau de agressividade e o melhor tratamento farmacológico sendo que inibidores competitivos de receptores (como o tamoxifeno para cânceres ER+) e anticorpos monoclonais de receptores e ligantes (como o trastuzumab para cânceres HER2+ e o atezolizumab para cânceres de mama basais ER-/HER2-) estão entre os principais utilizados (HARBECK et al., 2019; LOIBL et al., 2021).

Para entender a atuação celular de cada célula na comunidade, deve-se determinar as comunicações proteicas enviadas/trocadas. As medições de moléculas de mensagem expressas/traduzidas e de suas vias associadas são essenciais no entendimento da direção, magnitude e importância da comunicação celular. Medidas diretas de proteínas envolvidas/implicadas na comunicação celular requerem ensaios de bioquímica especializados e conhecimento prévio extenso do assunto/tópico. Além do mais, geralmente não são ou não podem ser estudadas no microambiente nativo. Técnicas de proteômica e de transcriptômica, principalmente as envolvendo sequenciamento de próxima-geração podem amparar os estudos de interação e comunicação com evidências dos perfis de expressão. Embora as técnicas de proteômica pareçam escolhas lógicas, uma vez que estas técnicas meçam a abundância proteica, os dados de sequenciamento de RNA são mais abundantes e mais facilmente obtidos (HARPER; BENNETT, 2016; ARMINGOL et al., 2021).

1.3 ALGORITMOS DE COMUNICAÇÃO CELULAR EM TRANSCRIPTÔMICA

Os cientistas procuram novos meios de utilizar os dados de transcriptômica, de expressão gênica, para inferir interação e comunicação célula-célula e algumas metodologias já foram criadas. Aqui, visando simplificação, utilizaremos o termo “comunicação celular” para interações e comunicações entre células. Podemos usar

tópicos para entender as pluralidades de algoritmos, como 1 - “Quais elementos (ou interatores) de comunicação são utilizados pelos algoritmos?”, 2 - “Quais escores são utilizados nos elementos de comunicação celular pelos algoritmos?” e 3 - “Em quais conceitos os algoritmos de comunicação celular baseiam as suas inferências?”.

Quanto ao primeiro tópico atualmente há algoritmos que utilizam elementos de comunicação intercelular apenas, como ligantes e receptores enquanto outros utilizamos elementos de comunicação intercelular e intracelular, adicionando informação de reações metabólicas, redes de sinalização intracelular e de regulação gênica (Figura 2).

A maior parte dos estudos com transcriptômica e comunicação celular examina a comunicação intercelular entre ligantes e receptores utilizando/explorando listas de bancos de interação proteína-proteína. Graeber e Eisenberg publicaram, em 2001, um trabalho de inferência de comunicação celular em amostras de microarranjo de câncer utilizando um banco de dados curados com ~ 450 pares de ligantes e receptores de humanos e a chamaram de “Database of Ligand-Receptor Partners” para analisar os *loops* de sinalização autócrina de tumores (GRAEBER; EISENBERG, 2001). Um dos mais utilizados bancos é o de Jordan A. Ramilowski e colaboradores, publicado em 2015, e contendo ~ 2400 pares de ligantes e receptores humanos curados de outros bancos e da literatura (RAMILOWSKI et al., 2015). Esse artigo apresentou o primeiro mapa em grande escala de comunicação celular de 144 células primárias humanas, revelando que a maior parte das células expressam dezenas a centenas de ligantes e receptores criando uma rede muito conectada de sinalização autócrina e parácrina através de diversos caminhos ligante-receptor. Recentemente, Erick Armingol e colaboradores disponibilizaram um compêndio de listas disponíveis de pares ligante-receptor na literatura (https://github.com/LewisLabUCSD/Ligand-Receptor-Pairs) contendo 25 bancos de humanos, camundongos, ratos e *Caenorhabditis elegans* (ARMINGOL et al., 2021). Diversas proteínas requerem a montagem e interação de subunidades para agir e a não transcrição de qualquer subunidade bloqueia a interação ligante-receptor e a comunicação derivada. Em 2018, Roser Vento-Tormo e colaboradores criaram um repositório contendo complexos proteicos com

Figura 2. Elementos de comunicação inter, intracelular e sinalização utilizados por algoritmos.

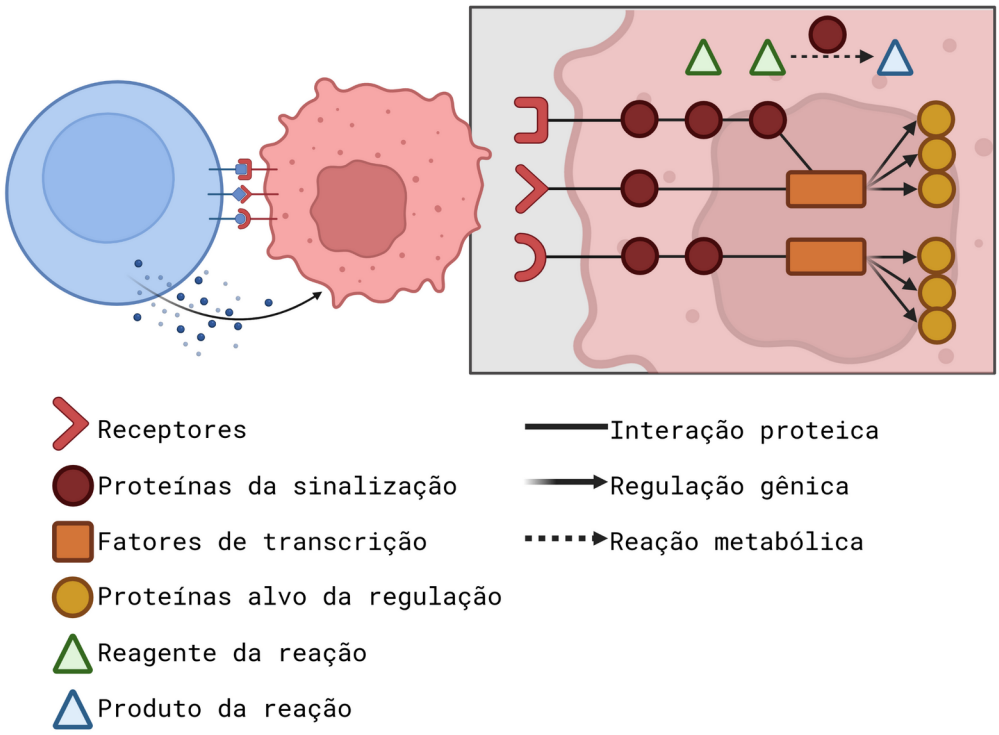


Ilustração da comunicação inter, intracelular e sinalização entre duas células, a azul e a vermelha, e os principais elementos utilizados na compreensão destas em transcriptômica. Nela, a célula vermelha é a célula receptora da comunicação intercelular, representada por interações entre ligantes e receptores e por comunicação na liberação de elementos da célula azul. A célula receptora inicia a sinalização a jusante (interações proteína-proteína na figura simbolizadas com linhas contínuas) por via de receptores cognatos geralmente culminando em alteração da atividade de fatores de transcrição e da expressão gênica, ou comunicação intracelular (na figura simbolizada com setas entre fatores de transcrição e proteínas alvo da regulação). A célula receptora também possui proteínas responsáveis por reações metabólicas (na figura simbolizada com linhas tracejadas).

subunidades de ligante-receptor chamado *CellPhoneDB* e uma ferramenta estatística para prever a especificidade do tipo de célula na comunicação celular por meio das interações moleculares (VENTO-TORMO et al., 2018). Porém esta não é a única ferramenta a abordar complexos proteicos com subunidades de ligantes e receptores, uma vez que considerar a co-expressão de subunidades melhor representa as comunicações funcionais, no momento há o algoritmo *RNA-Magnet* (BACCIN et al., 2020), o *ICELLNET* (NOËL et al., 2021) e o *CellChat* (JIN et al., 2021).

Conforme dito, os dados de reações metabólicas também são usado por algoritmos, como feito por Kakajan Komurov e colaboradores em 2012 (KOMUROV, 2012) e Anne Richelle e colaboradores em 2019 (RICHELLE; JOSHI; LEWIS, 2019). Ambos utilizaram dados de genes, reações e metabólitos na integração de dados transcriptômicos de microarranjo e RNA-seq em redes de reações bioquímicas visando facilitar a construção de modelos metabólicos específicos ao contexto. Para isso, os autores consideram as expressões de enzimas envolvidas em reações. No artigo de Kakajan Korumov os autores aplicam esses dados (e dados de interação proteína-proteína) para revelar comunicações envolvidas na resistência de inibidores da angiogênese em modelos de xenoenxerto de câncer de pulmão em camundongos (KOMUROV, 2012) enquanto no artigo de Anne Richelle os autores buscam diretrizes, regras, para integração de dados de transcritos e as redes de reações (RICHELLE; JOSHI; LEWIS, 2019). Dessa forma, em ambos os estudos é predita a produção de metabólitos com base nos transcritos, mas não sua concentração.

As redes de sinalização intracelular e de regulação gênica baseadas em banco de dados de interação proteína-proteína são extensamente utilizadas e alguns benefícios destas em comparação com outras abordagens analíticas, como a análise de expressão diferencial, são descritos em contextos não envolvendo e envolvendo a comunicação celular (CHUANG et al., 2007). Estes dados estão sendo utilizados especialmente para esclarecer como as comunicações ativam ou reprimem a transcrição gênica. É importante, entretanto, lembrar que a inferência de comunicação celular baseada em banco de dados de interação proteína-proteína são sensíveis a qualidade e alcance da informação. O algoritmo *CCCExplorer*, publicado em 2015, utiliza os elementos de comunicação intercelular (ligantes e receptores) e de regulação gênica através de fatores de transcrição ativados (alvos *up*-regulados) para caracterizar vias de comunicação entre elementos multicelulares entre o tumor e o estroma em modelo de câncer de pulmão de camundongo utilizando os dados de RNA-seq (CHOI et al., 2015). Em 2019, na era da transcriptômica de células individuais, o algoritmo *SoptSC* buscou contabilizar a heterogeneidade de células do mesmo *cluster* e as comunicações viáveis derivadas da heterogeneidade utilizando os elementos de comunicação intercelular e listas de

alvos *up* ou *down*-regulados derivados da comunicação classificados pelo usuário (WANG et al., 2019a). Os autores utilizaram o *SoptSC* em contextos de embriogênese em camundongos e em humanos, regeneração epidérmica e hematopoiese em camundongos. Ao contrário do *SoptSC* e as listas de genes alvos o algoritmo *NicheNet*, publicado em 2020, optou por ponderar, dar valor, ao conhecimento prévio da literatura (BROWAEYS; SAELENS; SAEYS, 2020). Esse algoritmo utiliza os elementos de comunicação intercelular, sinalização intracelular e regulação gênica com incorporação das redes de interação proteína-proteína e de regulação fator de transcrição-alvos. As redes de comunicações são ponderadas utilizando otimização de parâmetros, onde fontes mais informativas (o quão fortemente o conhecimento prévio ampara a interação/regulação) contribuem mais ao modelo. Os autores utilizaram o *NicheNet* em contexto de câncer de cabeça e pescoço explorando as interações entre células do tumor e o componente imune no microambiente tumoral. O algoritmo CytoTalk, publicado em 2021, abordou a comunicação célula-célula utilizando interações conhecidas entre ligantes e receptores porém as interações intracelulares foram preditas de modo *de novo*, ou seja, sem o uso de referências de interação proteína-proteína (HU et al., 2021). Para isso, o algoritmo define a informação mútua entre interatores intracelulares com os dados de co-expressão gênica e esta é usada/utilizada nos pesos/ponderações das interações. Utilizando o CytoTalk os autores realizaram a comparação de redes de sinalização entre macrófagos e células endoteliais em tecidos humanos fetais e adultos. Recentemente, em abril de 2022, Edroaldo Lummertz da Rocha e colaboradores publicaram o *CellComm*, um algoritmo de predição de interação entre células ou entre *spots* (utilizando dados de transcriptômica espacial) para desvendar os efeitos de comunicações na regulação gênica proporcionando também a construção/reconstrução de vias de sinalizações (LUMMERTZ DA ROCHA et al., 2022). Dessa forma o *CellComm* também utiliza elementos de comunicação intercelular, sinalização intracelular e de regulação gênica através de redes de interação de proteínas e de regulação fatores de transcrição-alvos. Com sua utilização, foi possível detectar sinais do microambiente e redes de sinalizações e regulações envolvidas na formação da hematopoiese e as confirmaram em modelos de peixe-zebra, camundongo e humano.

O tópico número 2 da segregação de algoritmos perguntava “Quais escores são utilizados nos elementos de comunicação celular pelos algoritmos?”. Em estudos de comunicação celular com os dados de transcriptômica os autores definem funções de escores para as moléculas em interação. O escore é geralmente definido para os pares de proteínas interatoras, como ligantes e receptores, utilizando a expressão dos genes codificantes. Essa inferência assume que a abundância proteica está inteiramente relacionada à expressão gênica e que é suficiente para inferir cenário e força de interação proteína-proteína, o que não é verdade. Sabe-se que há fatores muitas vezes basilares ao *binding* de proteínas, como modificações pós-transcricionais, pós-traducionais, especificidades biofísicas, formação de complexos protéicos, etc (ARMINGOL et al., 2021). Atualmente, há escores de comunicação binários ou contínuos.

Os escores binários são encontrados na literatura na forma de duas funções, a de limiar de expressão e a de combinações diferenciais . Na função de limiar de expressão um certo valor é definido como o limiar e se os genes (por exemplo um par de genes em interação) mostrarem expressão maior que o limiar, são considerados “ativos” enquanto na função de combinações diferenciais os interatores ativos são definidos por caracterização de genes diferencialmente expressos em análise de expressão diferencial (ARMINGOL et al., 2021). Limiares diferentes de expressão e de parâmetro diferencial (como o *fold change*) podem vir a ser utilizados em genes diferentes, geralmente empregando valores maiores para ligantes que para receptores, como na publicação de Hemant Suryawanshi e colaboradores em 2018, onde eram considerados somente interatores presentes em mais de 15 % das células, para cada tipo celular, com expressão maior que 5 e 1.5 transcritos por milhão para ligantes e para receptores, nessa ordem (SURYAWANSHI et al., 2018); no algoritmo CellTalker na publicação de Anthony R. Cillo e colaboradores em 2020, onde eram considerados ligantes e receptores presentes em mais de 1 % das células, para cada tipo celular, com expressão maior que 5 *counts* e, de modo adicional, presentes em mais de 5 % das células, para cada tipo celular, em mais de 25 % de amostras por grupo (CILLO et al., 2020) ou no algoritmo *CCCExplorer*, onde eram considerados ligantes *up*-regulados em contrastes de células associadas ao tumor ou do tumor e de suas contrapartes

sadias e receptores expressos/detectados em células do tumor (CHOI et al., 2015). Ambas as técnicas demandam limiares e assumem que valores maiores são requeridos para interação e isso pode resultar em falsos-positivos, por exemplo um gene expresso não condiz absolutamente em proteína traduzida ou funcional, e falsos negativos, por exemplo um gene de baixa expressão capaz de codificar uma proteína com alta força de interação e de propagação de sinal. Desta forma, é necessário o estabelecimento de limiares específicos de genes, uma vez que valores gerais podem não representar a presença/detecção e atividade de proteínas. Estes limiares (idealmente) devem ser tipo celular, estado celular, composição celular (na decomposição de *bulk*) e até mesmo tecnologia-específicos.

Os escores contínuos também são encontrados na literatura na forma de duas funções, a de produto da expressão e a de correlação da expressão. Na função de produto da expressão é calculado o produto (o resultante da multiplicação) de valores de expressão de genes em interação enquanto na função de correlação de expressão é o coeficiente de correlação entre os valores de expressão de genes em interação em amostras (ARMINGOL et al., 2021). Um benefício de valores contínuos é a oportunidade de caracterizar relações e correlações entre variáveis como a correlação entre a comunicação de expansão e a infiltração de células T reguladoras em melanoma metastático descrito por Manu P. Kumar e colaboradores em 2018 (KUMAR et al., 2018) e as validações de interações utilizando o produto da expressão de *Vegfa-Kdr* entre células tipo-alveolar 1 e células endoteliais *Car4* e *Pdgfb-Pdgfrb* entre células endoteliais e pericitos em pulmões de camundongos feitas por Cain e colaboradores em 2020 (CAIN; HERNANDEZ; CHEN, 2020). Ainda, neste mesmo artigo, Cain e colaboradores utilizaram as duas funções (de produto e de correlação) na análise da correlação entre produtos de expressão em duas condições, pulmões de camundongos sadios e infectados pelo vírus Sendai visando identificar o ganho ou a perda de interação entre ligantes-receptores após infecção (CAIN; HERNANDEZ; CHEN, 2020). Logo, os escores contínuos propiciam estudos inviáveis com os escores binários, mas apresentam alguns obstáculos. No caso da função produto da expressão, genes em interação contendo expressões muito diferentes, onde a expressão de um gene domina podem causar falsos-positivos, por exemplo um ligante com alta

expressão pode resultar na detecção de interações de receptores com baixa expressão. Uma mitigação exequível é a normalização dos dados com genes *housekeeping* por tipo/subtipo celular ou considerar a correlação entre RNAs e proteínas. No caso da função correlação de expressão o método requer várias medições, precisando de várias amostras e, ainda, matrizes de valores esparsos (contendo diversos zeros), como os dados de scRNA-seq, podem enviesar os coeficientes de correlação.

Além dos escores com funções simples, alguns autores geraram métodos novos de escores de interação como o *CCCExplorer* (CHOI et al., 2015), *SoptSC* (WANG et al., 2019a), *NicheNet* (BROWAEYS; SAELENS; SAEYS, 2020), *RNA-Magnet* (BACCIN et al., 2020) e outros. O algoritmo *CCCExplorer* considerou apenas ligantes *up*-regulados em contrastes de células associadas ao tumor ou do tumor e de suas contrapartes sadias (semelhante a função de limiar de expressão, porém com parâmetro pós análise de expressão diferencial), os receptores e fatores de transcrição expressos/detectados em células do tumor (sem uma função; apenas detecção) e os alvos *up*-regulados em células epiteliais do tumor em comparação com as células epiteliais sadias (CHOI et al., 2015). O algoritmo *SoptSC* adota um norte bayesiano e considera a detecção e a probabilidade de comunicação entre células com a expressão do ligante e células com a expressão do receptor e expressão diferencial documentada de alvos da via analisada (WANG et al., 2019a). O *NicheNet*, conforme citado, pondera as redes de comunicações com o uso de otimização de parâmetros, onde fontes mais informativas, mais fortemente suportadas no conhecimento prévio, contribuem mais e resultam em escores maiores de interação (BROWAEYS; SAELENS; SAEYS, 2020). O *RNA-Magnet*, por sua vez, foi descrito por Chiara Baccin e colaboradores em 2020 e pode predizer comunicações potenciais entre as células e âncoras (que são células ou populações escolhidas) com base nos padrões de expressão de ligantes e receptores (BACCIN et al., 2020). Os escores de elementos (ligantes e receptores) no *RNA-Magnet* usam as predições de transcritos feitas por MAGIC (*Markov affinity-based graph imputation of cells* ou imputação gráfica de células baseada em afinidade de Markov em inglês), um método de compartilhamento de informações entre células semelhantes, de difusão de dados, para eliminar os ruídos da matriz de contagem de células e

preencher/completar transcritos ausentes (VAN DIJK et al., 2018), e as transformam em uma variável lógica difusa (*fuzzy logic*, do inglês) para codificar se o gene é expresso. Com o *RNA-Magnet* Chiara Baccin pode predizer a organização espacial de células de medula óssea inferida de dados de scRNA-seq.

O último tópico, o número 3 da segregação de algoritmos perguntava: “Em quais conceitos os algoritmos de comunicação celular baseiam as suas inferências?”. Erick Armingol define quatro categorias de conceitos com base nos modelos matemáticos empregados para identificar interações intercelulares (geralmente entre ligantes e receptores): os baseados em combinações diferenciais, em redes, em permutações de expressão e em tensores (ARMINGOL et al., 2021).

Em conceitos baseados em combinações diferenciais os interatores ativos são definidos por caracterização de genes diferencialmente expressos em análise de expressão diferencial. Além do *CCCExplorer* já falado que considera ligantes e alvos *up*-regulados (CHOI et al., 2015), os algoritmos de comunicação celular *iTALK* (WANG et al., 2019c) e *PyMINER* (TYLER et al., 2019) utilizam conceitos baseados em combinações diferenciais. A ferramenta *iTALK* é um pacote em linguagem R que define genes intensamente expressos (quando há apenas 1 grupo) e/ou genes diferencialmente expressos e identifica quais deles estão entre o banco de dados de ligantes e receptores da ferramenta (WANG et al., 2019c). Em casos de análise de expressão diferencial (dois ou mais grupos), o *iTALK* encontra alterações (ganhos ou perdas) significativas de interações, onde há ganhos de interação se um gene interator for *up*-regulado e seu gene de interação for *up*-regulado ou permanecer sem alteração e há perdas de interação se o gene interator for *down*-regulado, não importando o nível de expressão de seu gene de interação. A ferramenta *PyMINER*, por sua vez, é um pacote em liguagem Python (ou *library* de Python) que define marcadores e identifica quais deles estão entre o banco de dados de interações físicas proteína-proteína do STRING v10, as determinando interações de ativação ou inibição (TYLER et al., 2019). As funções e as ferramentas baseadas em combinações diferenciais são excelentes nas análises e na obtenção de interações diferenciais, porém elas são incapazes de obter as interações comuns entre grupos.

Os conceitos baseados em redes, por sua vez, exploram as conexões de dados em estruturas de redes. Como citado, os algoritmos *CCCExplorer* (CHOI et

al., 2015), *SoptSC* (WANG et al., 2019a), *NicheNet* (BROWAEYS; SAELENS; SAEYS, 2020) e *CellComm* (LUMMERTZ DA ROCHA et al., 2022) são exemplos que utilizam redes de interação proteína-proteína e/ou redes de regulação gênica para integrar dados de comunicação intercelular com dados de sinalização intracelular e de controle transcricional relativo a interação. Ainda, o *NicheNet* (BROWAEYS; SAELENS; SAEYS, 2020) emprega a teoria de grafos e medidas de centralidade e conectividade em um algoritmo personalizado chamado PageRank para quantificar a importância na rede de cada nó, de cada elemento, e desta forma, ranquear interações ligante-receptor e o *CellComm* (LUMMERTZ DA ROCHA et al., 2022) implementa um método de otimização de fluxo de redes de modo a resolver "caminhos", vias de sinalização, entre o receptor e os fatores de transcrição em células receptoras (CANG; NIE, 2020). Apesar da integração de dados de comunicação inter e intracelular e de sinalização intracelular utilizando estruturas de redes, essas ferramentas ainda não consideram as regras de regulação gênica e são incapazes de obter as sinalizações de *crosstalk* onde os sinais desencadeados por um receptor podem interferir nos sinais desencadeados por outro.

Conceitos baseados em permutações de expressão definem escores de comunição para cada interação ou interator e avaliam sua significância utilizando diferentes métodos, como a permutação de marcação de *clusters*, onde se gera a distribuição nula de cada interação ou interator empregando as trocas de marcação de *clusters* de todas as amostras *x* vezes; os testes não paramétricos para avaliar diferenças do modelo nulo, onde o modelo nulo é o modelo estatístico em que não há interação ou distinção entre os *clusters* e os métodos empíricos (ARMINGOL et al., 2021). Dados adicionais de interações ou interatores podem colaborar em cálculos de permutação, como os dados de complexos proteicos, retirados do *CellPhoneDB* (VENTO-TORMO et al., 2018), *RNA-Magnet* (BACCIN et al., 2020), *CellChat* (JIN et al., 2021) e *ICELLNET* (NOËL et al., 2021), os dados de modulação positiva/negativa da expressão de receptores em casos de múltiplos receptores co-estimuladores/co-inibidores na interação como no algoritmo *CellChat* (JIN et al., 2021). Os pesquisadores que optam por conceitos baseados em permutações de expressão devem definir limites, *thresholds*, para definir a significância. A permutação de marcação de *clusters*, por exemplo, não discute tal questão, visto

que *clusters* com bastante exemplares resultam em uma não significância. Com base nisso, o algoritmo SingleCellSignalR define um escore regularizado do produto ligante-receptor indicado para variações de profundidade no sequenciamento de scRNA-seq como na presença de *dropouts* e na definição de limites de significância utilizando análises de áreas sob a curva e de estabilidade em taxa de descoberta falsa baseadas em dados de transcriptômica de maior profundidade e de proteômica (CABELLO-AGUILAR et al., 2021).

Por fim, os conceitos baseados em tensores. Tensores são entidades geométricas de *n*-dimensões introduzidas na matemática e na física para generalização de vetores e de matrizes e, atualmente, vem sendo utilizados em estruturas de *machine* e *deep learning* na otimização de seus muitos cálculos. O algoritmo *scTensor* publicado por Koki Tsuyuzaki e colaboradores em 2019, emprega tensores na extração de interações célula-célula representativas (TSUYUZAKI; ISHII; NIKAIDO, 2019). Apesar de ótimos resultados em interações conhecidas e novas, sendo que o *scTensor* distinguiu interações novas em conjuntos de dados simulados e empíricos não detectadas por métodos comparados, a análise, o entendimento e a significação dos resultados não é fácil e pode afastar o usuário e o leitor.

1.4 ALGORITMOS DE COMUNICAÇÃO CELULAR EM TRANSCRIPTÔMICA E O USO DAS COORDENADAS ESPACIAIS

Uma vez que a comunicação celular, ao menos de forma autócrina e parácrina é espacialmente limitada, os métodos de transcriptômica espacial podem aumentar a confiabilidade de comunicações preditas. Os autores de análise de comunicação celular vem pensando nisso e desenvolvendo algoritmos que aproveitam a informação espacial em suas predições. Os principais exemplos à disposição são o *SpaOTsc* (CANG; NIE, 2020), o método utilizado por Falkner-Corbett e colaboradores (FAWKNER-CORBETT et al., 2021a), o *Giotto* (DRIES et al., 2021) e o *CellComm* (LUMMERTZ DA ROCHA et al., 2022).

No *SpaOTsc* (CANG; NIE, 2020), o algoritmo prediz as possíveis coordenadas de células caracterizadas em scRNA-seq utilizando dados de

transcriptômica espacial. Dessa forma, os autores podem obter uma resolução maior do transcriptoma (mais moléculas, mais genes) de experimentos de transcriptômica espacial de baixa resolução. O algoritmo trata a comunicação celular como um problema de transporte, visando o menor custo de transporte de várias origens a vários destinos. Neste, a rede de distância célula-célula é usada como custo de transporte para restringir espacialmente a rede de interação e o transporte “ótimo” representa as probabilidades de comunicação celular.

O método utilizado por Falkner-Corbett e colaboradores (FAWKNER-CORBETT et al., 2021a) analisou a co-localização de pares de ligantes e receptores definindo um modelo linear generalizado para cada par testando se a expressão do receptor é dependente da expressão do ligante em cada *spot* da lâmina. Como alguns pares de ligantes e receptores podem sinalizar em distâncias maiores, os autores definiram, da mesma forma, um modelo linear generalizado para cada spot calculando uma matriz com expressão “suavizada” (do inglês *smoothed*), ou com expressão média, entre o spot e os spots imediatos ao redor.

No *Giotto* (DRIES et al., 2021), o algoritmo calcula a co-expressão espacial de pares de ligantes e receptores para distinguir as comunicações mais plausíveis. Mais especificamente, para cada par de interatores é calculado um escore de comunicação celular entre populações. O escore é calculado da expressão média ponderada de interatores em subconjuntos de células/*spots* próximos de ambas as populações em interação. Então, visando avaliar se o escore é estatisticamente significativo, é calculada a distribuição nula aleatória empregando permutação de posições de células/*spots* de um tipo celular/*cluster* 1000 vezes. No trabalho há também a comparação direta da interação sem e com a informação espacial e, em suma, o poder preditivo é bastante limitado nas análises espacialmente ignorantes.

No *CellComm* (LUMMERTZ DA ROCHA et al., 2022), por sua vez, o algoritmo calcula a co-expressão de pares de ligantes e receptores em todas as combinações de populações. O *CellComm* também usa permutação de marcação de *clusters* para avaliar se a co-expressão é estatisticamente significativa. Em seguida, o algoritmo calcula as distâncias entre os centróides das populações se baseando na distância euclidiana. Define-se um limite de distância de centróides para as

comunicações aceitáveis e o *CellComm* define um escore de comunicação que considera a co-expressão de interatores e a co-localização espacial de populaçoes.

1.5 JUSTIFICATIVA

Conforme mostrado, os algoritmos de análise de comunicação celular em dados de transcriptômica espacial até agora focavam na análise de populações de *spots*, sejam *clusters* transcricionais ou regiões teciduais, e, dito isso, não examinam as particularidades de comunicação do *spot*, que, dependendo da tecnologia, pode ser uma célula ou um pequeno grupo de células ainda que o poder oriundo dessa análise seja fascinante. Uma análise não baseada no *spot* caracterizaria as comunicações celulares, por exemplo, entre uma massa tumoral e o tecido periférico, os entendendo como duas populações resolvidas transcricionalmente. Já uma análise baseada no *spot* poderia investigar 1 - a diversidade de comunicações na massa tumoral (intratumoral): se há heterogeneidade de perfil transcricional dentro do câncer, se há diferentes clones dentro do câncer, alguns com perfis diferentes de prognóstico e de resposta ao tratamento; 2 - a diversidade de comunicações entre regiões distintas da massa tumoral: se há diferenças na comunicação entre a região do córtex e do centro do câncer, ou entre os *spots* próximos e os distantes a vasos, inferindo/indicando a propensão à metástase e 3 – a diversidade de comunicações entre o tumor e o microambiente tumoral e outros.

Nessa tese será mostrado o *SpotComm*, um novo pacote desenvolvido utilizando a linguagem R que tem por objetivo predizer comunicações inter e intracelulares em conjuntos de dados de transcriptômica espacial, com foco no *spot*. O algoritmo analisa dados em duas (uma seção de tecido) ou três dimensões (duas ou mais seções de tecido de corte seriado) de diversas tecnologias. Para comunicação intercelular o algoritmo permite ao usuário estabelecer os interatores intercelulares de interesse, como ligantes e receptores, e utilizando um banco de dados atualizado, com informação de proteínas interatoras monoméricas e complexos proteicos homomultiméricos e heteromultiméricos e, se baseando na detecção e expressão destes prediz as comunicações ativas. Para comunicação intracelular (regulação gênica) o algoritmo define os fatores de transcrição e os alvos

detectados no *spot* e, se baseando na detecção e na expressão desses elementos prediz os regulons ativos. Então, em posse de receptores e regulons ativados o algoritmo prediz *k* caminhos, vias de sinalização possíveis amparadas em interações proteicas descritas na literatura e, também, se baseando na detecção e na expressão destas. O *SpotComm* elabora dados de detecção, proporção, expressão, escores binários e contínuos como produto da expressão de pares de interatores e correlação espacial entre *spots* próximos e outros para concepção de hipóteses e para análises. O algoritmo calcula permutações de expressão, de correlação e de posição espacial para designar significância estatística às predições. Ainda, o *SpotComm* integra os dados de transcriptômica espacial com os dados de scRNA-seq, pareados ou não, para deconvolução de *spots* em probabilidades da presença de tipos celulares e então o cálculo da proporção de células e da proporção de co-ocorrência, por tipo celular, com detecção dos elementos da via de sinalização intracelular (do receptor ao fator de transcrição). Assim, a ferramenta pode predizer as comunicações e as possíveis células envolvidas nelas. Entendemos que assim, com essa nova visão, seja possível localizar comunicações ocultas, heterogeneidades relevantes na criação de hipóteses e de pesquisas que podem revelar novos saberes na saúde e nas doenças, bem como nos tratamentos. Tudo por *spot*.

2. HIPÓTESE

A análise baseada em coordenadas provê novos dados sobre a diversidade de interatores e define a distribuição espacialmente autocorrelacionada de comunicações inter- e intracelulares do câncer de mama.

3. OBJETIVOS

3.1 OBJETIVO GERAL

Desenvolver um algoritmo de predição de comunicações celulares que utiliza os dados de coordenadas, de *spots*, como unidade de análise.

3.2 OBJETIVOS ESPECÍFICOS

- Desenvolver funções capazes de utilizar dados de transcriptômica espacial em análises de uma seção (2D) ou de mais seções (3D) para exploração de comunicações celulares.
- Viabilizar análises de *clusters* transcricionais inteiros e/ou zonas de contato entre *clusters*.
- Garantir cenários de comunicação e sinalização entre proteínas interatoras monoméricas e complexos proteicos homomultiméricos e heteromultiméricos.
- Desenvolver funções capazes de definir as comunicações inter- e intracelulares e sinalizações intracelulares se baseando na presença e na correlação de transcritos.
- Proporcionar ao usuário alcance de metadados como presença, proporção e expressão de interatores e como referência e curagem de comunicações e sinalizações.
- Desenvolver funções capazes de integrar dados de transcriptômica espacial com dados pareados ou não pareados de scRNA-seq.
- Proporcionar ao usuário alcance de dados como proporção de células e proporção de co-ocorrência, por tipo celular, com detecção dos elementos de vias de sinalização intracelular.
- Obter dados de câncer de mama analisados via transcriptômica espacial.
- Utilizar o *SpotComm* em diferentes tecnologias de transcriptômica espacial (Visium 10x Genomics e 1k arrays) e em diferentes casos (câncer de mama subtipo molecular triplo-negativo e HER2-positivo).

4 MATERIAL E MÉTODOS

4.1 FUNÇÕES E PARÂMETROS DO *SpotComm*

Figura 4. *Hexagon sticker/icon* do *SpotComm.*

SpotComm é um novo pacote desenvolvido utilizando a linguagem *R* que tem por objetivo predizer comunicações inter- e intracelulares e sinalizações intracelulares entre *clusters* definidos por seu perfil transcricional em conjuntos de dados de transcriptômica espacial obtidos utilizando os dados de expressão gênica e de coordenação espacial de tecnologias de ST como Visium da 10x Genomics e 1k arrays.

Previamente, o usuário deve ter um objeto (um dado ou conjunto de dados definido) do tipo *Seurat* pré-processado e processado com um conjunto de dados de transcriptômica espacial obtidos. Esse objeto deve conter 1 - uma matriz de expressão em *counts* normalizados, como os *counts* corrigidos derivados de transformação reversa de resíduos de Pearson feita por *sctransform()* de

Hafemeister e Satija em 2019, com genes nas linhas, utilizando seu símbolo gênico oficial como descrito pelo comitê de nomenclaturas gênicas HUGO, e amostras nas colunas, utilizando seu código de cada *spot*, 2 - uma tabela de metadados (dados sobre dados) contendo uma coluna com nomes dos *clusters*, definidos por seu perfil transcricional e 3 - uma tabela com as coordenadas de *spots*.

Nós inicializamos o diretório SpotComm/ com o auxílio do pacote *usethis* e da função *create_package()*. As funções foram criadas em arquivos *.R* e armazenadas no subdiretório R/. O *SpotComm* importa (como dependências) os pacotes de R *dplyr* (>= 1.0.9), *foreach* (>= 1.5.2), *ggplot2* (>= 3.3.6), *igraph* (>= 1.3.4), *Matrix* (>= 1.4-1), *OmnipathR* (>= 3.5.5), *reshape2* (>= 1.4.4), *scatterpie* (>= 0.1.7), *Seurat* (>= 4.1.1), *spdep* (>= 1.2-4), *stringr* (>= 1.4.0), *STutility* (>= 0.1.0), *tibble* (>= 3.1.8) e *yenpathy* (>= 1.0.0), e sugere os pacotes de R *knitr* e *rmarkdown*, para reprodução de vignettes. O pacote conta com a licença MIT que dá aos usuários permissão/anuência expressa para reutilizar o código para qualquer finalidade, às vezes mesmo se o código fizer parte de um software proprietário. Desde que os usuários incluam a cópia original da licença do MIT em sua distribuição, eles podem fazer quaisquer alterações ou modificações no código para atender às suas próprias necessidades. A Figura 4 exibe o fluxo das análises com o *SpotComm* com suas funções base, funções filtro, *outputs* e parâmetros exigidos.

4.1.1 *get_subset_2D()*

A função *get_subset_2D()* produz um *subset*, um subconjunto, do objeto *Seurat* contendo um par de *clusters* de interesse e/ou as zonas de contato entre estes dois *clusters* divergentes quanto à transcrição gênica (Figura 5). A função produz esse *subset* de objetos com somente um *slice*/amostra e é por tal motivo o sufixo *2D* em seu nome. Essa função possui cinco parâmetros: 1 - *sample* (objeto *Seurat*), 2 - *column_with_clusters* (caráter) , 3 - *cluster_01* (carácter), 4 - *cluster_02* (carácter), 5 - *contact* (lógico). Usando tal função nós definimos a “identidade/*cluster*” dos *spots* do objeto do tipo *Seurat* (definido em *sample*), utilizando como base a coluna da tabela de metadados contendo os *clusters* (definido em

Figura 4. Fluxograma de funções, *outputs* e parâmetros do *SpotComm*.

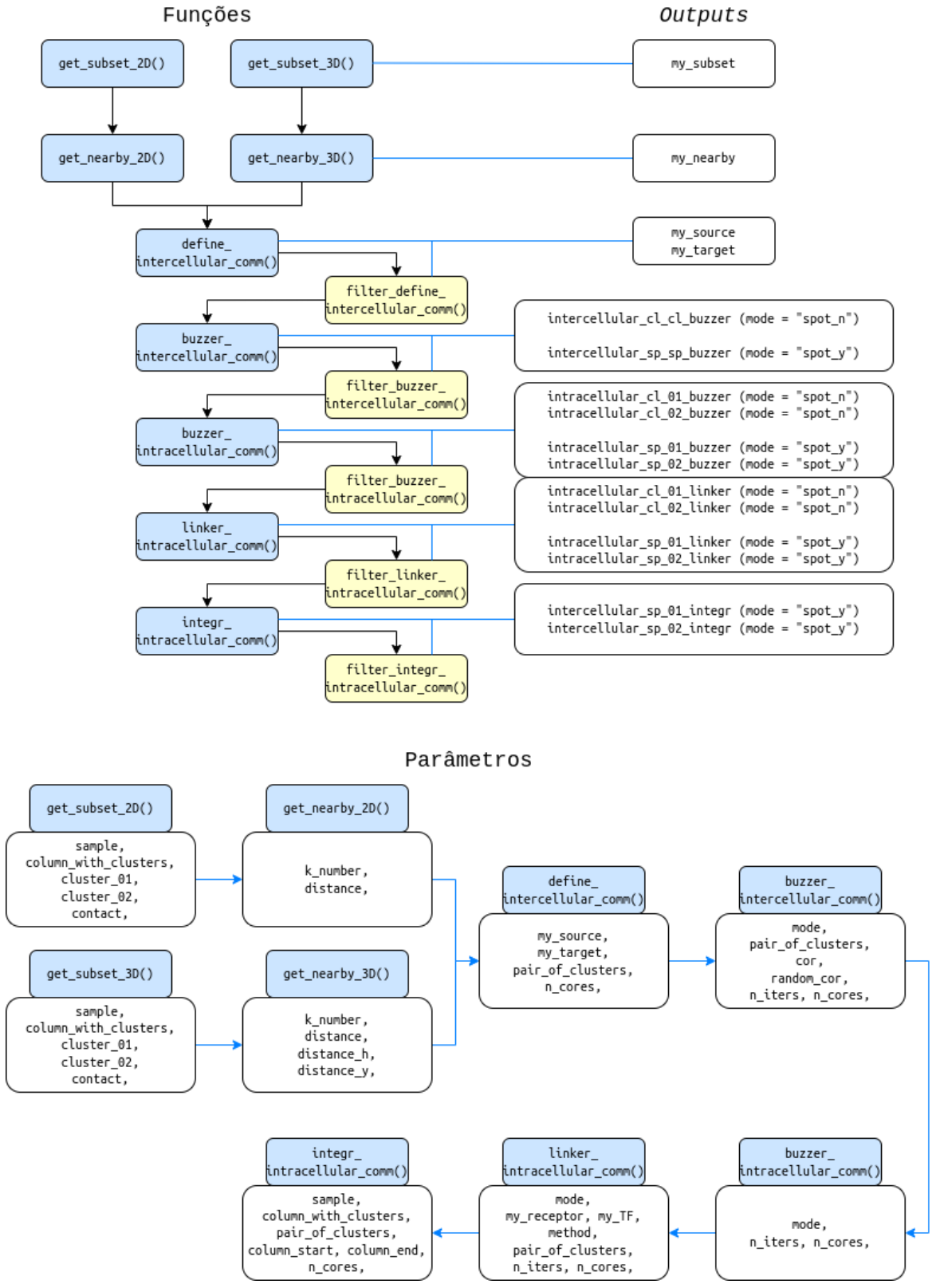
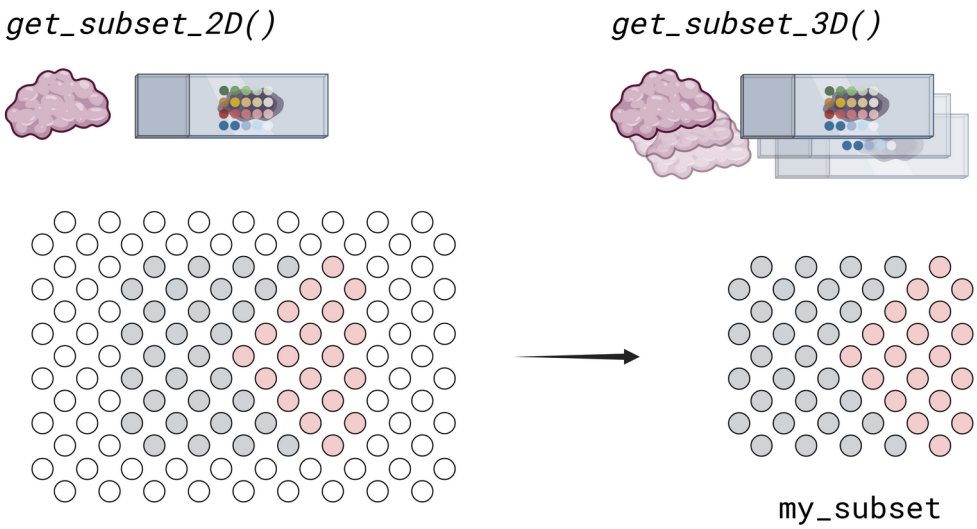


Figura 5. Visão geral das funções *get_subset_2D()* e *get_subset_3D()*.



O objetivo primário da função *get_subset_2D()* e *get_subset_3D()* é filtrar somente os *spots* de interesse do usuário, aqui os *spots* vermelho e cinza. Isso pode ser feito nos estudos com uma (2D) ou mais (3D) seções de tecido.

column_with_clusters). Assim, nós identificamos todos os *spots* vizinhos de dois *clusters* de interesse (definido em *cluster_01* e *cluster_02*) utilizando a função *RegionNeighbors* do pacote *STutility* e definimos um objeto chamado/nomeado *my_subset* se compondo de 1 - todos os *spots* de ambos os *clusters* (caso *contact* = *FALSE*) ou 2 - apenas os *spots* vizinhos de *cluster_01* e *cluster_02* (caso *contact* = *TRUE*). Sendo assim, o parâmetro *contact* oferece análise do *cluster* em totalidade ou somente as zonas de contato. Para utilização da função é necessária a presença de certo objeto chamado *Staffli*, um objeto de classe S4, armazenado dentro do objeto *Seurat*, específico do pacote *STutility.* Este objeto contém os metadados específicos do *STutility*, como coordenadas de pixel, identificadores de amostra, plataforma, etc. Se pode também analisar um único *cluster* de interesse, definindo os parâmetros *cluster_01* e *cluster_02* com o mesmo *cluster*. Assim, é possível investigar a heterogeneidade de comunicações de certo *cluster* isolado.

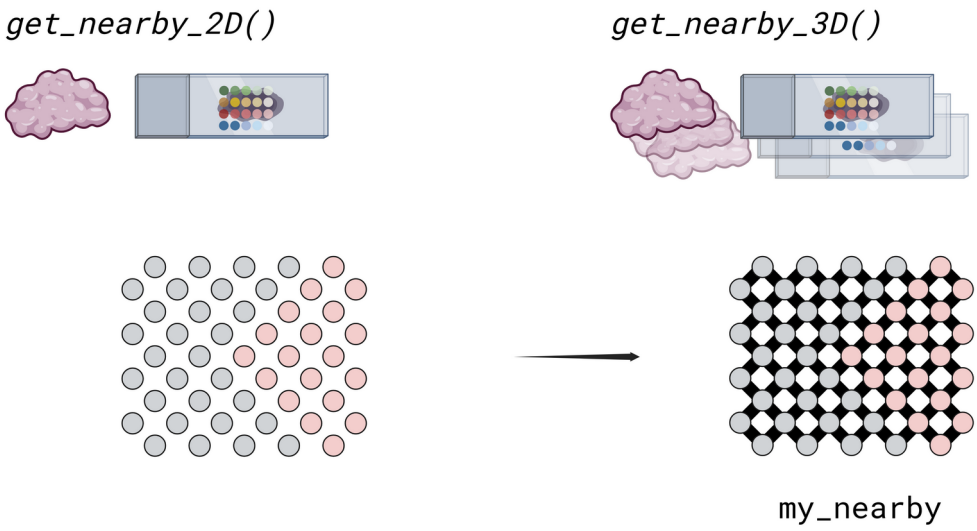
4.1.2 *get_subset_3D()*

A função *get_subset_3D()* produz um *subset*, um subconjunto, do objeto *Seurat* contendo um par de *clusters* de interesse e/ou as zonas de contato entre estes dois *clusters* divergentes quanto à transcrição gênica. A função produz esse *subset* de objetos com dois ou mais *slices*/amostras e é por tal motivo o sufixo 3*D* em seu nome. Essa função possui os mesmos parâmetros da função *get_subset_2D()* e é utilizada da mesma forma.

4.1.3 *get_nearby_2D()*

A função *get_nearby_2D()* produz um objeto com os pares de *spots* próximos (Figura 6). Essa função deve seguir de *get_subset_2D()*. Essa função possui dois parâmetros: 1 - *k_number* (numérico) e 2 - *distance* (numérico). Com ela nós definimos um objeto de três colunas nomeado *my_nearby* com informação sobre os pares de *spots* próximos (dentre os *clusters* em estudo), definidos pela distância um do outro. Nas colunas 1 e 2, nomeadas “Spot_01” e “Spot_02”, nós temos códigos (os nomes de amostras) de cada *spot* e de seu vizinho, na coluna 3 , nomeada “Spot_01_Spot_02”, temos um código do tal par (nome em “Spot_01” e nome em “Spot_02” separados por “_”). O parâmetro *k_number* requer um número de *spots* em geral de vizinhos de cada *spot*. É comum que o valor de *k_number* seja 6 ou 8 (contendo vizinhos a cada 60º como no caso da plataforma Visium da 10x Genomics ou 45º como no caso da plataforma 1k arrays). O parâmetro *distance*, por sua vez, requer um número pouco maior que a maior distância entre um *spot* e os *spots* vizinhos; exemplo, um *spot* da plataforma Visium da 10x Genomics tem geralmente 6 *spots* vizinhos, calcula-se as distâncias entre um *spot* (totalmente cercado) e os *spots* vizinhos, pega-se o maior valor de distância e então se utiliza um valor um pouco maior no parâmetro *distance*, exibindo ao *SpotComm* qual é a

Figura 6. Visão geral das funções *get_nearby_2D()* e *get_nearby_3D()*.



O objetivo primário da função *get_nearby_2D()* e *get_nearby_3D()* é definir os *spots* vizinhos conforme a distância marcada pelo usuário. Isso pode ser feito nos estudos com uma (2D) ou mais (3D) seções de tecido.

distância entre *spots* em que há, ou que se são aceitas, comunicações. Tendo em vista a análise de comunicação do *SpotComm* esse objeto possui/inclui 1 – pares de spots (exemplo, “Spot_01” = “A” e “Spot_02” = “B”), 2 - pares invertidos de *spots* (exemplo, “Spot_01” = “B” e “Spot_02” = “A”), uma vez que a comunicação é bilateral e 3 - pares do mesmo *spot* (exemplo, “Spot_01” = “A” e “Spot_02” = A), uma vez que as células do *spot* podem também se comunicar de modo autócrino e parácrino no espaço do *spot*.

4.1.4 *get_nearby_3D()*

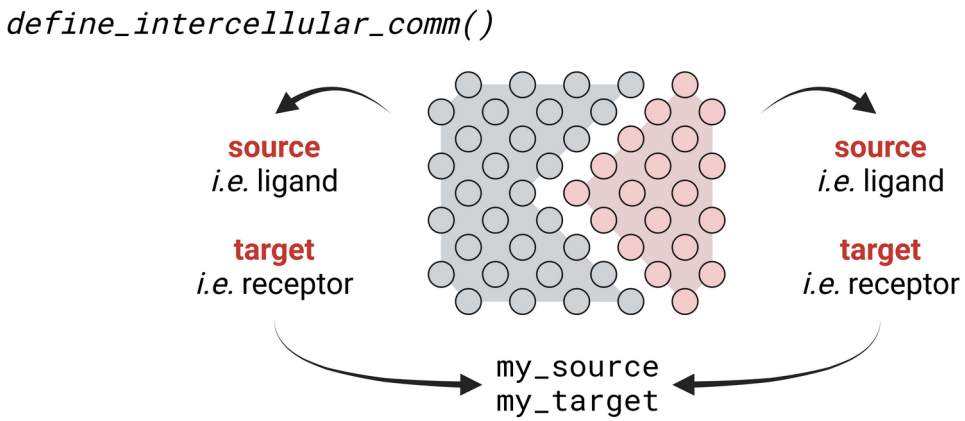
A função *get_nearby_3D()* produz um objeto com os pares de *spots* próximos. Essa função deve seguir de *get_subset_3D()*. Essa função possui quatro parâmetros: 1 - *k_number* (numérico), 2 - *distance* (numérico), 3 - *distance_h* (numérico) e 4 - *distance_z* (numérico). Com ela nós definimos um objeto de três colunas nomeado *my_nearby* com informação sobre os pares de *spots* próximos (dentre os *clusters* em estudo), definidos pela distância um do outro, entretanto esse

objeto contém *spots* próximos de um mesmo corte histólogico, como em *get_nearby_2D()*, e entre cortes histológicos, e é por tal motivo o sufixo 3*D* em seu nome. Os parâmetros *sample*, *k_number* e *distance* servem os mesmos propósitos da função *get_nearby_2D()*, já o parâmetro *distance_h* requer a distância da hipotenusa na triangulação entre os *spots* de diferentes cortes e, portanto, não deve ser maior que o parâmetro *distance* (sugere-se o mesmo valor) e o parâmetro *distance_z* requer a distância entre os cortes histológicos, a distância em terceira dimensão.

4.1.5 *define_intercelullar_comm()*

A função *define_intercellular_comm* define dois objetos, *my_source* e *my_target* incluindo elementos de comunicação entre células (Figura 7). Essa função possui quatro parâmetros: 1 - *my_source* (carácter) e 2 - *my_target* (carácter), 3 - *pair_of_clusters* (lógico) e 4 - *n_cores* (numérico). Para construir esse par de objetos a função aproveita a função *import_omnipath_intercell()* do pacote *OmnipathR* para acessar um banco de dados com elementos de comunicação entre células e filtra da coluna *category* para os *sources* (elementos de origem) e *targets* (elementos de alvo) (definidos em *my_source* e *my_target*) de interesse do usuário. As opções para os parâmetros *my_source* e *my_target* podem ser genéricas (como “*ligand*” e “*receptor*”) ou específicas (como “*ligand_agonist*/*ligand_antagonist*” e “*growth_factor_receptor*”) e as fontes/bancos usados podem ser recuperadas com as funções *get_intercell_categories()* e *get_intercell_resources()* do pacote *OmnipathR*. O parâmetro *pair_of_clusters* informa, caso = *TRUE*, que o usuário analisa a comunicação entre dois *clusters* transcricionais e, caso = *FALSE*, analisa a “autocomunicação” de um *cluster* transcricional. Sendo assim, esse parâmetro está relacionado ao preenchimento em parâmetros *cluster_01* e *cluster_02* das funções *get_subset_2D()* e *get_subset_3D()*. O parâmetro *n_cores* requer o número de núcleos de processamento disponíveis ao usuário, explorando as diferentes arquiteturas para redução do tempo do processo. Os objetos de *output*, *my_source*

Figura 7. Visão geral da função *define_intercellular_comm()*.



O objetivo primário da função *define_intercellular_comm()* é definir as moléculas de comunicação intercelular analisadas/consideradas pelo usuário. As moléculas são definidas por detecção e expressão nos *clusters*, aqui as sombras em vermelho e cinza.

e *my_target,* possuem proteínas monoméricas e complexos proteicos homomultiméricos e heteromultiméricos. Os objetos possuem as colunas “genesymbol”, com nomes das proteínas; “entity_type”; “cl_01_detected_in” e “cl_02_detected_in” contendo a proporção de *spots* dos *clusters* 1 e 2 que contém expressão de cada *source*/*target* em questão; as colunas “cl_01_mean_expression” e “cl_02_mean_expression” contendo a média de expressão (em *counts*) de cada *source*/*target* nos *spots* dos *clusters* 1 e 2 e as colunas “cl_01_mean_expression_in_detected” e “cl_02_mean_expression_in_detected” contendo a média de expressão (em *counts*) de cada *source*/*target* nos *spots* com a expressão detectada dos *clusters* 1 e 2.

Exemplo: Nós podemos começar a análise de comunicação mediada por ligantes e receptores entre um par de *clusters* transcricionalmente distintos com a definição *my_source* = “*ligand*” e *my_target* = “*receptor*”.

4.1.6 *filter_define_intercelullar_comm()*

A função *filter_define_intercellular_comm()* faz parte das funções *filter_*(...) que tem por objetivo filtrar os objetos gerados pelas funções (...) de forma

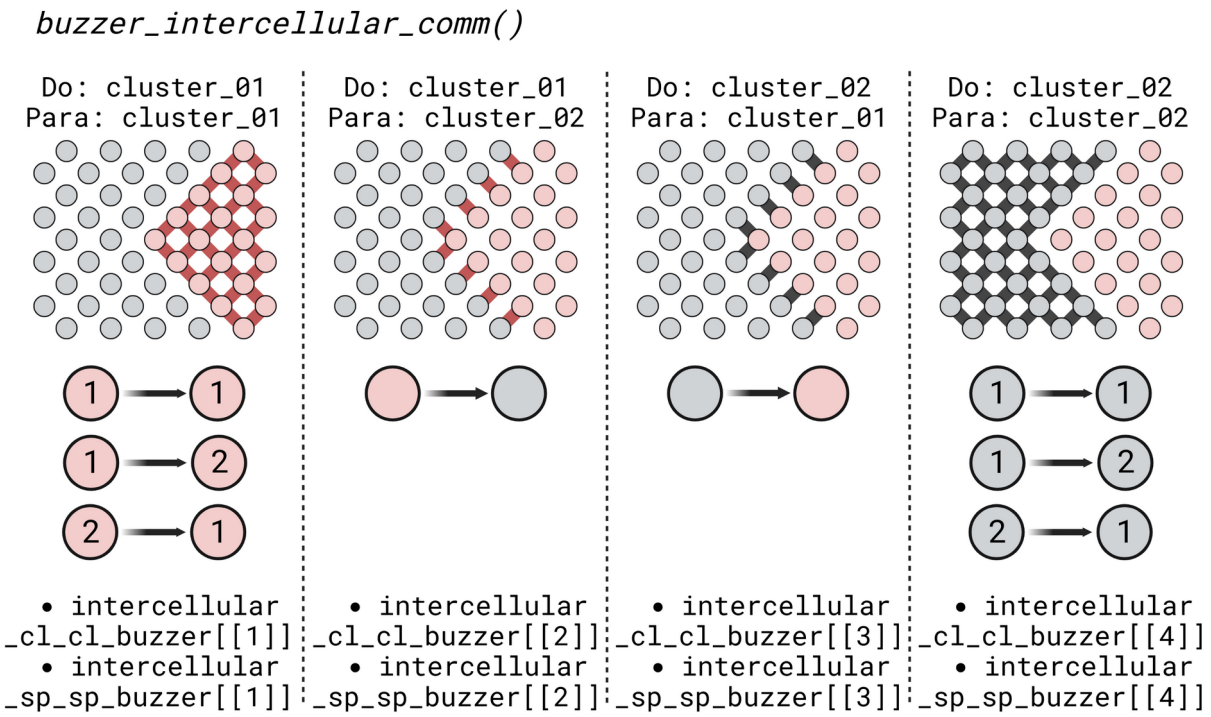
rápida e lógica. Dessa forma, essa função filtra os objetos gerados por *define_intercellular_comm()*. Os parâmetros são referentes ao 1 - objeto (*source*/*target*) que desejamos filtrar, 2 - cluster (*cl_01*/*cl_02*), 3 - coluna (*detected_in* /*mean_expression*/*mean_expression_in_detected*), todos separados por "_". Os parâmetros são definidos por valores numéricos e marcam os cortes de filtragem.

Exemplo: Nós podemos filtrar os *sources* (como ligantes) expressos em mais de 25 % dos *spots* dos *clusters* 1 e 2 utilizando os parâmetros *source_cl_01_detected_in* = 25 e *source_cl_02_detected_in* = 25.

4.1.7 *buzzer_intercelullar_comm()*

A função *buzzer_intercellular_comm()* define as possíveis comunicações intercelulares entre os *clusters* e os *spots* se baseando na presença de transcritos (Figura 8). Essa função possui seis parâmetros: 1 - *mode* (*spot_n*/*spot_y*), 2 - *pair_of_clusters* (lógico), 3 - *cor* (lógico) e 4 - *random_cor* (lógico) e 5 – *n_iters* (numérico) e 6 - *n_cores* (numérico). Essa é a primeira função descrita com o parâmetro *mode* (modo, em inglês) que, por sua vez, define as análises entre os pares de *clusters* com os *spots* analisados de forma conjunta ou individual, a distinção elementar do *SpotComm*. Essa função aproveita a função *import_post_translational_interactions()* do pacote *OmnipathR* para acessar um banco de dados de interações físicas entre proteínas/complexos de proteínas e filtra da coluna *genesymbol* para os *sources* (elementos de origem) e *targets* (elementos de alvo) (definidos em *my_source* e *my_target* na função *define_intercelullar_comm()*) de interesse do usuário. As fontes/bancos usados podem ser recuperadas com as funções *get_intercell_categories()* e *get_intercell_resources()* do pacote *OmnipathR*. Depois, nós definimos um objeto (*intercellular_cl_cl_buzzer* caso *mode* = *spot_n* e *intercellular_sp_sp_buzzer* caso *mode* = *spot_y*) englobando 4 listas, uma para cada possível comunicação entre os *clusters* (1 - *source* e *target* no *cluster* 1, 2 - *source* no *cluster* 1 e *target* no *cluster* 2, 3 - *source* no *cluster* 2 e *target* no *cluster* 1 e, por fim, 4 - *source* e *target* no cluster 2). O parâmetro *pair_of_clusters* informa, caso = *TRUE*, que o usuário analisa a

Figura 8. Visão geral da função *buzzer_intercellular_comm()*.



O objetivo primário da função *buzzer_intercellular_comm()* é definir as comunicações intercelulares entre *clusters* ou *spots* em quatro cenários: 1 - do *cluster* 1 para o *cluster* 1; 2 - do *cluster* 1 para o *cluster* 2; 3 – do *cluster* 2 para o *cluster* 1 e 4 do *cluster* 2 para o *cluster* 2.

comunicação entre dois *clusters* transcricionais e, caso = *FALSE*, analisa a "autocomunicação" de um *cluster* transcricional. Sendo assim, esse parâmetro está relacionado ao preenchimento em parâmetros *cluster_01* e *cluster_02* das funções *get_subset_2D()* e *get_subset_3D()*. Caso pair_of_clusters = FALSE as 4 listas do objeto terá os mesmos conteúdos, ou seja, *source* e *target* no *cluster* 1. O parâmetro *n_iters* requer o número de iterações feitas em análises de *p*-empírico. O parâmetro *n_cores* é o mesmo que o parâmetro na função *define_intercellular_comm().*

No caso de *mode* = *spot_n* (analisados de forma conjunta), as quatro listas de *intercellular_cl_cl_buzzer* possuem quatro colunas de identidade do *source* e do *target* em comunicação (colunas *source*, *target*, *source_genesymbol* e *target_genesymbol*); 11 colunas de dados e referências sobre a comunicação; As colunas *is_directed*, *consensus_direction*, *is_stimulation*, *consensus_stimulation*,

is_inhibition, *consensus_inhibition* (numéricos/binários 0 = não, 1 = sim) indicam se a comunicação é direta, provoca estimulação ou inibição e se tais informações são ou não são consenso na literatura. As colunas *sources* (p.e. = TRIP, do banco de dados "*Transient receptor potential channel-interacting protein*"), *n_resources*, *references* (p.e. = TRIP:11290752), *n_references* e *curation_effort* indicam as fontes, o número de fontes, as referências, o número de referências não curadas e curadas. Ainda, as listas de *intercellular_cl_cl_buzzer* possuem duas colunas com a expressão média do *source* e do *target* nos *clusters* em questão, as colunas *source_mean_expression* e *target_mean_expression* e uma coluna com o produto da expressão média, ou seja, os valores de *source_mean_expression* vezes os valores de *target_mean_expression* armazenado na coluna *mean_expression_product*. Complexos proteicos heteromultiméricos no *source* ou no *target* apresentam a expressão média de todas as proteínas e o produto da expressão média é calculado entre todos os elementos de *source* e *target*. Além do mais, nós comparamos os valores do produto da expressão média com 1000 valores de produto da expressão média de par randômico de genes *source* (da mesma categoria de *source*) e *target* (da mesma categoria de *target*). Dessa forma nós pegamos 1 - a média de 1000 genes *sources* (p.e. 1000 ligantes) e definimos como *sources*, 2 - a média de 1000 genes *targets* (p.e. receptores) e definimos como *targets*, os multiplicamos e obtemos 1000 valores de produto de expressão média de pares randômicos e calculamos a probabilidade (*p* empírico) de que um par randômico tenha expressão média maior que o par *source*/*target* em questão, armazenado na coluna *empirical_pv_01*.

Ainda no caso de *mode* = *spot_n* é possível selecionar os parâmetros *cor* = *TRUE* e *random_cor* = *TRUE*. Caso o parâmetro *cor* seja definido como "TRUE" (ocorra), a função calculará a correlação de Pearson entre os pares de *source*/*target* entre *spots* próximos daqueles *clusters* (os analisados, como p.e. *source* e *target* no *cluster* 1, a lista 1 de *intercellular_cl_cl_buzzer*). Dessa forma utilizamos os pares de *spots* próximos armazenados no objeto *my_nearby* e assim nós temos uma informação de quanto a expressão de um *source* é relacionada a expressão de um *target* em um contexto espacial. Esse parâmetro fornece duas colunas, 1 - *cor_pvalue* e 2- *cor_estimate*, com o *p*-valor e o coeficiente de correlação do teste.

Por sua vez, caso o parâmetro *random_cor* seja definido como “TRUE” (ocorra), a função calculará a correlação de Pearson entre os pares de *source*/*target* entre 1000 pares de *spots* randômicos daqueles *clusters* (os analisados, como p.e. *source* e *target* no *cluster* 1, a lista 1 de *intercellular_cl_cl_buzzer*). Esse parâmetro fornece duas colunas, 1 - *random_cor_pvalue* e 2- *random_cor_estimate*, com o *p*-valor e o coeficiente de correlação do teste. Dessa forma, é possível observar as colunas de correlações que consideram e que desconsideram a informação espacial e checar se há alguma relação entre as proximidades, as posições de *spots*, de *clusters* e a correlação da expressão.

No caso de *mode* = *spot_y* (analisados de forma individual), as quatro listas de *intercellular_sp_sp_buzzer* possuem duas colunas de identidade de *spot*, a coluna *sent* e *received* contendo os *spots* com a expressão do *source* e do *target*, por essa ordem; Quatro colunas de identidade do *source* e do *target* em comunicação (colunas *source*, *target*, *source_genesymbol* e *target_genesymbol*) e 11 colunas de dados e referências sobre a comunicação (*is_directed*, *consensus_direction*, *is_stimulation*, *consensus_stimulation*, *is_inhibition*, *consensus_inhibition*, *sources*, *n_resources*, *references*, *n_references* e *curation_effort*) como para *mode* = *spot_n*. Ainda, as listas de *intercellular_sp_sp_buzzer* possuem duas colunas com a expressão do *source* e do *target* nos *spots* em questão (por vir de um *spot* o *source* o *target* são expressos em *counts* e não média de *counts*), as colunas *source_expression* e *target_expression* e uma coluna com o produto da expressão, ou seja, os valores de *source_expression* vezes os valores de *target_expression* armazenado na coluna *expression_product*. Como descrito, complexos proteicos heteromultiméricos no *source* ou no *target* apresentam a expressão média de todas as proteínas e o produto da expressão média é calculado entre todos os elementos de *source* e *target*. Além do mais, como para *mode* = *spot_n* nós comparamos os valores do produto da expressão com 1000 valores de produto da expressão de par randômico de genes *source* (da mesma categoria de *source*) e *target* (da mesma categoria de *target*) e calculamos a probabilidade (*p* empírico) de que um par randômico tenha expressão maior que o par *source*/*target* em questão, armazenado na coluna *empirical_pv_01*.

4.1.8 *filter_buzzer_intercelullar_comm()*

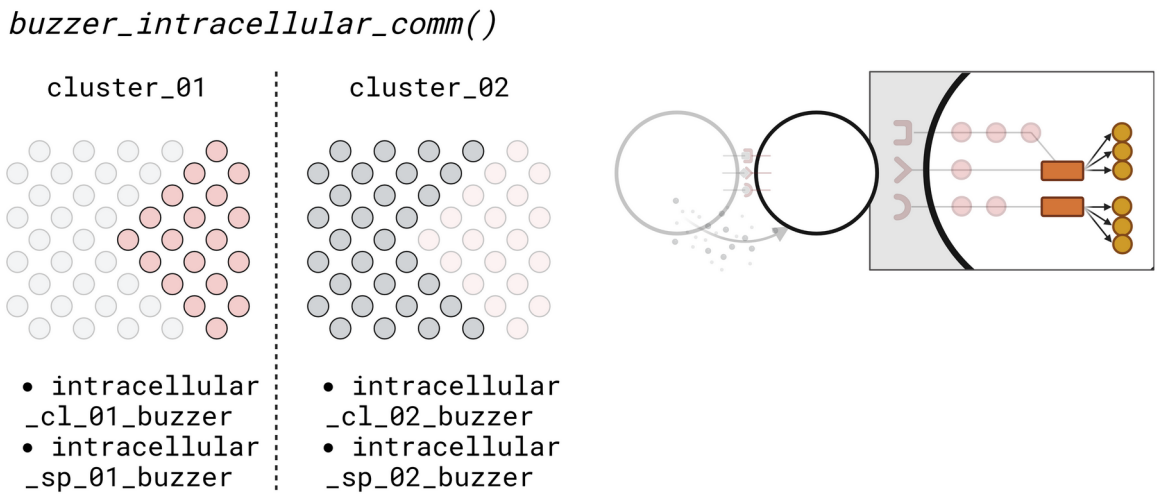
A função *filter_buzzer_intercellular_comm()* pode filtrar os objetos gerados por *buzzer_intercellular_comm()*. Os parâmetros são referentes ao 1 - par de *clusters* (analisados de forma conjunta, *mode* = *spot_n*, ou individual, *mode* = *spot_y*. Em caso de *mode* = *spot_n* se deve utilizar *cl_0X_cl_0X* e em caso de *mode* = *spot_y*, se deve utilizar *sp_0X_sp_0X*, *clusters* em X) que desejamos filtrar, 2 - coluna das quatro listas do objeto *intercellular_cl_cl_buzzer* e *intercellular_sp_sp_buzzer* (*source_mean_expression*, *target_ mean_expression*, *mean_expression_product* e *empirical_pv_01* para *intercellular_cl_cl_ buzzer* e *source_expression*, *target_expression*, *expression_product* e *empirical_pv_01* para *intercellular_sp_sp_buzzer*), todos separados por "_". Os parâmetros são definidos por valores numéricos e marcam os cortes de filtragem.

Exemplo: Nós podemos filtrar os pares *source*/*target*, na comunicação entre o *cluster* 1 com o *cluster* 1 (analisados de forma conjunta, *mode* = *spot_n*), com produtos de expressão média maior que 25 e *p* empírico menor que 0.1 (apenas 10 % dos produtos de expressão média randômicos maior que o obtido) utilizando os parâmetros *cl_01_cl_01_mean_ expression_product* = 25 e *cl_01_cl_01_empirical_pv_01* = 0.1. Todavia, a mesma busca porém com *clusters* analisados de forma individual (*mode* = *spot_y*) se utilizam os parâmetros *sp_01_sp_01_expression_product* = 25 e *sp_01_sp_01_empirical_pv_01* = 0.1.

4.1.9 *buzzer_intracelullar_comm()*

A função *buzzer_intercellular_comm* define as (uma das) possíveis respostas intracelulares, as modificações no perfil de expressão gênica, seguinte a comunicação intercelular entre os *clusters* e os *spots* se baseando na presença de transcritos (Figura 9). Essa função possui três parâmetros: 1 - *mode* (*spot_n*/*spot_y*) que, como dito, define as análises entre os pares de *clusters* com os *spots* analisados de forma conjunta ou individual e 2 – *n_iters* (numérico) e 3 - *n_cores* (numérico), o mesmo que o parâmetro *n_cores* na função

Figura 9. Visão geral da função *buzzer_intracellular_comm()*.



O objetivo primário da função *buzzer_intracellular_comm()* é definir as comunicações intracelulares (regulação entre fatores de transcrição e alvos) em *clusters* ou *spots*. Assim é possível detectar os fatores de transcrição ativos.

define_intercellular_comm(). Essa função aproveita a função *import_transcriptional_interactions* do pacote *OmnipathR* para acessar um banco de dados de interações de *regulons*, ou seja, entre os fatores de transcrição e os alvos (contendo genes codificantes de proteínas). As fontes/bancos usados podem ser recuperadas com as funções *get_intercell_categories()* e *get_intercell_resources()* do pacote *OmnipathR*. Nós consideramos fatores de transcrição em proteínas monoméricas e complexos proteicos homomultiméricos e heteromultiméricos. Além do mais, somente consideramos fatores de transcrição com pelo menos quatro alvos. Depois, nós definimos dois objetos, *intracellular_cl_01_buzzer* e *intracellular_cl_02_buzzer* caso mode = *spot_n*, *intracellular_sp_01_buzzer* e *intracellular_sp_02_buzzer* caso mode = *spot_y*, contendo os *regulons* de *clusters* 1 e 2, nessa ordem. O parâmetro *n_iters* e *n_cores* são os mesmos que os parâmetros nas funções *define_intercellular_comm()* e *buzzer_intercellular_comm().*

No caso de *mode* = *spot_n* (analisados de forma conjunta), os objetos *intracellular_cl_01_buzzer* e *intracellular_cl_02_buzzer* possuem duas colunas de identidade do fator de transcrição e de alvos em comunicação (colunas *TF_name* e *target_name*); duas colunas, *TF_detected_in* e *target_detected_in*, contendo a

proporção de *spots* no *cluster* que contém expressão (*counts* > 0); duas colunas, *TF_mean_expression* e *target_mean_ expression*, contendo a média de expressão (em *counts*) e duas colunas, *TF_mean_ expression_in_detected* e *target_mean_expression_in_detected*, contendo a média de expressão (em *counts*) de *spots* no *cluster* que contém expressão (*counts* > 0). Ainda, os objetos *intracellular_cl_01_buzzer* e *intracellular_cl_02_buzzer* possuem três colunas dirigidas aos alvos, *target_length*, *target_detected* e *target_proportion* contendo o total de alvos do fator de transcrição obtido em *import_transcriptional_interactions*, o total de alvos que contém expressão (*counts* > 0) no *cluster* explorado e a proporção de alvos que contém expressão por o total de alvos do fator de transcrição obtido, nessa ordem. Complexos proteicos heteromultiméricos no fator de transcrição ou em alvos apresentam a detecção, expressão média e etc de todas as proteínas. Por fim, nós comparamos a expressão média dos alvos (cada alvo, elemento) com 1000 valores de expressão média (mediana, porque muitos alvos possuem expressão = 0, e sendo assim, qualquer detecção de expressão, mesmo que 1 *spot* somente, traria a média de expressão para um valor > 0) de genes randômicos do *cluster* e calculamos a probabilidade (*p* empírico) de que os genes randômicos possuam expressão mediana maior que os alvos em questão, armazenado na coluna *empirical_pv_01*.

No caso de *mode* = *spot_y* (analisados de forma individual), os objetos *intracellular_sp_01_buzzer* e *intracellular_sp_02_buzzer* possuem uma coluna de identidade de *spot*, a coluna *spot*, contendo os *spots* com a expressão do fator de transcrição e de alvos, duas colunas de identidade do fator de transcrição e de alvos em comunicação (colunas *TF_name* e *target_name*) e duas colunas, *TF_expression* e *target_expression*, contendo a expressão (em *counts*). Como em *mode* = *spot_n*, os objetos *intracellular_sp_01_buzzer* e *intracellular_sp_02_ buzzer* possuem três colunas dirigidas aos alvos, *target_length*, *target_detected* e *target_proportion* contendo o total de alvos do fator de transcrição obtido em *import_transcriptional_interactions*, o total de alvos que contém expressão (*counts* > 0) no *spot* explorado e a proporção de alvos que contém expressão por o total de alvos do fator de transcrição obtido, nessa ordem. Complexos proteicos heteromultiméricos no fator de transcrição ou em alvos apresentam nome,

expressão e etc de todas as proteínas. Também, nós comparamos a expressão dos alvos (cada alvo, elemento) com 1000 valores de expressão média (mediana, mesmo motivo acima) de genes randômicos do *cluster* 1 - contendo zeros e 2 - não contendo zeros e calculamos a probabilidade (*p* empírico) de que os genes randômicos possuam expressão mediana maior que os alvos em questão, armazenado na coluna *empirical_pv_01* e *empirical_pv_02*.

4.1.10 *filter_buzzer_intracelullar_comm()*

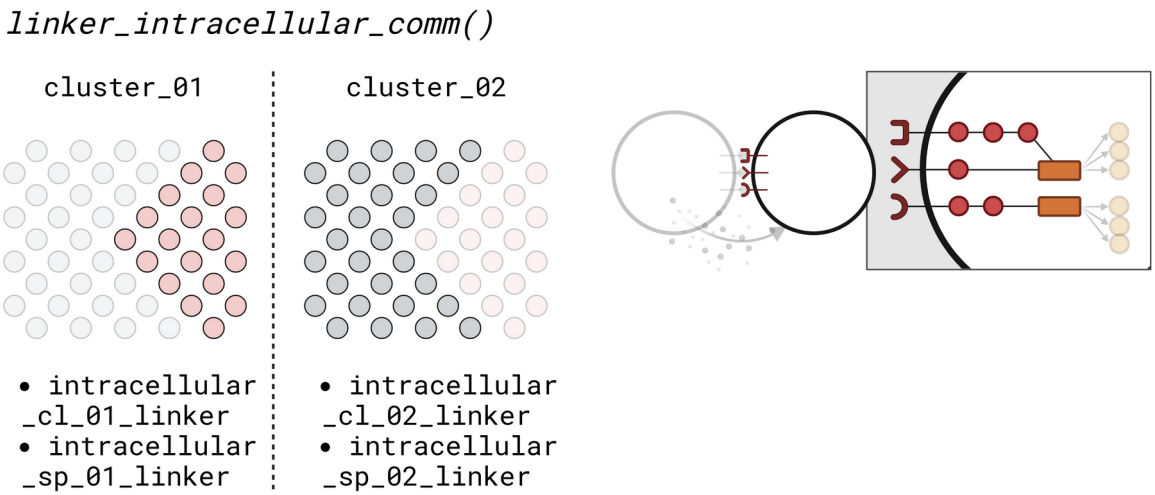
A função *filter_buzzer_intracellular_comm()* pode filtrar os objetos gerados por *buzzer_intracellular_comm*. Os parâmetros são referentes ao 1 - *cluster* (analisado de forma conjunta, *mode* = *spot_n*, ou individual, *mode* = *spot_y*. Em caso de *mode* = *spot_n* se deve utilizar *cl_0X* e em caso de *mode* = *spot_y*, se deve utilizar *sp_0X*, *clusters* em X) que desejamos filtrar, 2 - colunas dos objetos *intracellular_cl_01_buzzer, intracellular_cl_02_ buzzer*, *intracellular_sp_01_buzzer* e *intracellular_sp_02_buzzer* (*TF_detected_in*, *target_ detected_in*, *TF_mean_expression*, *target_mean_expression*, *TF_mean_expression_in_ detected*, *target_mean_expression_in_detected*, *target_proportion* e *empirical_pv_01* para *intracellular_cl_01_buzzer* e *intracellular_cl_02_buzzer* e *TF_detected_in*, *target_ detected_in*, *TF_expression*, *target_expression*, *target_proportion* e *empirical_pv_01* para *intracellular_sp_01_buzzer* e *intracellular_sp_02_buzzer*), todos separados por "_". Os parâmetros são definidos por valores numéricos e marcam os cortes de filtragem.

Exemplo: Nós podemos filtrar os pares de fatores de transcrição e de alvos, no *cluster* 1 (analisado de forma conjunta, *mode* = *spot_n*), com fatores de transcrição com expressão maior que 10 e *p* empírico menor que 0.1 (apenas 10 % dos genes randômicos possuem expressão mediana maior que os alvos em questão) utilizando os parâmetros *cl_01_TF_mean_expression* = 10 e *cl_01_empirical_pv_01* = 0.1. Todavia, a mesma busca porém com *cluster* analisado de forma individual (*mode* = *spot_y*) se utilizam os parâmetros *sp_01_TF_expression_product* = 10 e *sp_01_empirical_pv_01* = 0.1.

4.1.11 *linker_intracelullar_sign()*

A função *linker_intracellular_sign()* define possíveis vias de sinalização intracelular entre receptores e fatores de transcrição em *clusters* e *spots* se baseando na presença e correlação de transcritos (Figura 10). Essa função possui sete parâmetros: 1 - *mode* (*spot_n*/*spot_y*) que, como dito, define as análises entre os pares de *clusters* com os *spots* analisados de forma conjunta ou individual, 2 - *my_receptor* (carácter), 3 - *my_TF* (carácter), 4 - *method* (carácter), 5 - pair_of_clusters (lógico), 6 - *n_iters* (numérico) e 7 – *n_cores* (numérico). Essa função aproveita a função *import_post_translational_interactions()* do pacote *OmnipathR* para acessar um banco de dados de interações de proteínas. As fontes/bancos usados podem ser recuperadas com as funções *get_intercell_categories()* e *get_intercell_resources()* do pacote *OmnipathR*. Nós consideramos proteínas monoméricas e complexos proteicos homomultiméricos e heteromultiméricos. A função define grafos (*G* = *V*, *E*) (redes) com direções contendo as interações proteína-proteína. Os vértices (*V* (*G*)) desses grafos são definidos pelos genes e os *edges* (conectores, em inglês) (*E* (*G*)) são definidos pelos pesos das interações. Os pesos, por sua vez, são definidos e/ou calculados consoante o parâmetro *method*. Especificações sobre a definição/cálculo de pesos dependendo de *mode* e *method* estão abaixo. Para definir as vias, nós usamos 1 - o algoritmo de *shortest path* (caminho mais curto, em inglês) de Dijkstra, descrito em 1959 e que usa pesos das interações para definir o caminho de menor somatório entre dois vértices (como entre um receptor e um fator de transcrição) e 2 - o algoritmo de *k-shortest paths* (*k*-caminhos mais curtos, em inglês) de Yen, descrito em 1971 e que usa o conceito de Dijkstra para designar *k* caminhos de menor somatório entre dois vértices. Sendo assim, para cada combinação de receptores e de fatores de transcrição (definidos em my_receptor e em my_TF, nessa ordem) nós obtemos os *k* caminhos mais curtos, ou seja, os *k* caminhos com o menor somatório (de pesos) entre dois vértices. Depois, nós definimos dois objetos, *intracellular_cl_01_linker* e *intracellular_cl_02_linker* caso mode = *spot_n*, *intracellular_sp_01_linker* e *intracellular_ sp_02 _linker* caso mode = *spot_y*, contendo as vias de sinalização intracelular de *clusters* 1 e 2, nessa ordem. O

Figura 10. Visão geral da função *linker_intracellular_comm()*.



O objetivo primário da função *linker_intracellular_comm()* é definir as sinalizações intracelulares (estimulações entre proteínas que conectem receptores a fatores de transcrição) em *clusters* ou *spots*.

parâmetro *n_iters* e *n_cores* são os mesmos que os parâmetros nas funções *define_intercellular_comm()* e *buzzer_intercellular_comm().*

No caso de *mode* = *spot_n* (analisados de forma conjunta) e *method* = *correlations*, os objetos *intracellular_cl_01_linker* e *intracellular_cl_02_linker* possuem duas colunas de identidade, uma do receptor e uma do fator de transcrição (colunas *receptor* e *TF*), uma coluna, *path*, contendo as vias de sinalização intracelular entre as combinação de receptores e de fatores de transcrição; uma coluna, *weight*, contendo as somas de pesos da via (de cada interação proteína-proteína). Aqui, usando o *mode* = *spot_n* e o método *correlations* para cada *cluster* nós excluímos os genes que não apresentam detecção (*counts* > 0) em todas as amostras da construção do grafo nós calculamos a correlação de Pearson entre todos os pares de genes que apresentam detecção (*counts* > 0) em mais de 25 % dos *spots* (para cada *cluster*) e calculamos os pesos de cada interação proteína-proteína como 1 - coeficiente de correlação de Pearson. Genes com detecção (*counts* > 0) em menos de 25 % são imbuídos com coeficiente de correlação de Pearson = 0 e, portanto, pesos = 1; uma coluna, *n_vertices*, contendo o total de elementos/proteínas da via; uma coluna nós comparamos as somas de pesos

(coluna *weight*) da via de sinalização intracelular com 1000 valores de somas de pesos derivadas de vias com um ponto de partida randômico e “número de passos”, número de interações proteína-proteína, igual ao *n_vertices* (ou seja, utilizamos um recurso de propagação de redes chamado *Random Walk*) e calculamos a probabilidade (*p* empírico) de que vias de sinalização intracelular “randômicas” possuam as somas de pesos menor que as das vias em questão, armazenado na coluna *empirical_pv_01*. Por fim, os objetos contém uma coluna com as expressões médias (em *counts*) de cada gene na via, por seu *cluster*.

No caso de *mode* = *spot_n* e *method* = *counts*, os objetos *intracellular_cl_01_linker* e *intracellular_cl_02_linker* possuem as colunas acima, entretanto nós calculamos os pesos de cada interação proteína-proteína como o máximo log2 do produto da expressão média entre pares - log2 do produto da expressão média das proteínas em interação.

No caso de *mode* = *spot_y* (analisados de forma individual) e *method* = *correlations*, os objetos *intracellular_sp_01_linker* e *intracellular_sp_02_linker* possuem três colunas de identidade, uma do *spot*, do receptor e do fator de transcrição (colunas *spot*, *receptor* e *TF*), uma coluna, *path*, contendo as vias de sinalização intracelular entre as combinação de receptores e de fatores de transcrição; uma coluna, *weight*, contendo as somas de pesos da via (de cada interação proteína-proteína). Aqui, usando o modo = *spot_y* e o método *correlations* para cada *spot* nós excluímos os genes que não apresentam detecção (*counts* > 0) da construção do grafo e nós calculamos a correlação de Pearson entre todos os pares de genes que apresentam detecção em mais de 25 % dos *spots* (para cada *cluster*) e calculamos os pesos de cada interação proteína-proteína como 1 - coeficiente de correlação de Pearson. Genes com detecção em menos de 25 % são imbuídos com coeficiente de correlação de Pearson = 0 e, portanto, pesos = 1; uma coluna, *n_vertices*, contendo o total de elementos/proteínas da via; uma coluna nós comparamos as somas de pesos (coluna *weight*) da via de sinalização intracelular com 1000 valores de somas de pesos derivadas de vias com um ponto de partida randômico e “número de passos”, número de interações proteína-proteína, igual ao *n_vertices* (ou seja, utilizamos um recurso de propagação de redes chamado *Random Walk*) e calculamos a probabilidade (*p* empírico) de que vias de sinalização

intracelular “randômicas” possuam as somas de pesos menor que as das vias em questão, armazenado na coluna *empirical_pv_01*. Por fim, os objetos contém uma coluna com as expressões (em *counts*) de cada gene na via, por seu *spot*.

No caso de *mode* = *spot_y* e *method* = *counts* , os objetos *intracellular_sp_01_linker* e *intracellular_sp_02_linker* possuem as colunas acima, entretanto nós calculamos os pesos de cada interação proteína-proteína como o máximo log2 do produto da expressão (*em counts*) entre pares - log2 do produto da expressão (*em counts*) das proteínas em interação.

4.1.12 *filter_linker_intracelullar_sign()*

A função *filter_linker_intracellular_sign()* pode filtrar os objetos gerados por *linker_intracellular_sign*. Os parâmetros são referentes ao 1 - *cluster* (analisado de forma conjunta, *mode* = *“spot_n”*, ou individual, *mode* = *“spot_y”*. Em caso de *mode* = *“spot_n”* se deve utilizar *cl_0X* e em caso de *mode* = *“spot_y”*, se deve utilizar *sp_0X*, *clusters* em X) que desejamos filtrar, 2 - colunas dos objetos *intracellular_cl_01_linker, intracellular_cl_02_ linker*, *intracellular_sp_01_linker* e *intracellular_sp_02_linker* (*empirical_pv_01* e *counts*), todos separados por “_”. Os parâmetros são definidos por valores numéricos e marcam os cortes de filtragem.

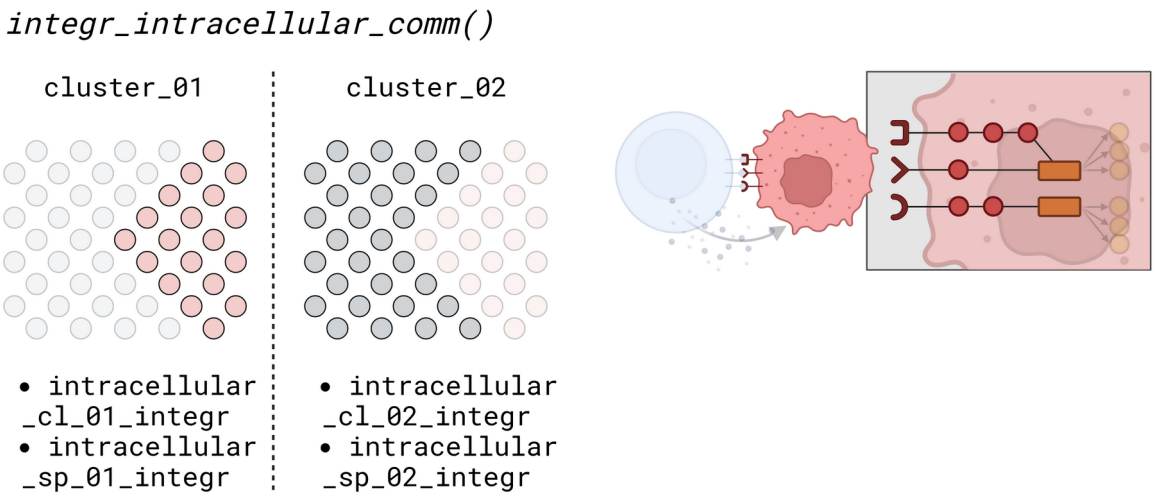
Exemplo: Nós podemos filtrar as vias de sinalização, no *cluster* 1 (analisado de forma conjunta, *mode = “spot_n”*), com elementos (todos) apresentando expressão maior que 3 *counts* e *p* empírico menor que 0.1 (apenas 10 % de vias de sinalização intracelular “randômicas” possuem somas de pesos menor que as das vias em questão) utilizando os parâmetros *cl_01_counts* = 3 e *cl_01_empirical_pv_01* = 0.1. Todavia, a mesma busca porém com *cluster* analisado de forma individual (*mode* = *“spot_y”*) se utilizam os parâmetros *sp_01_counts* = 3 e *sp_01_empirical_pv_01* = 0.1.

4.1.13 *integr_intracelullar_sign()*

O *SpotComm* também possui funções criadas para integrar dados de

single cell RNA-seq ao *spatial* RNA-seq e auxiliar a predição da real comunicação na área estudada (Figura 11). O usuário deve conter um objeto do tipo *Seurat* com um conjunto de dados de transcriptômica espacial após deconvolução para tipos/estados celulares dos *spots* e um objeto do tipo *Seurat* com um conjunto de dados de scRNA-seq pareados ou vindos de tecidos similares ao do transcriptômica espacial (preferentemente o empregue na deconvolução do transcriptômica espacial). Os resultados da deconvolução devem aparecer nos metadados com colunas dedicadas a cada tipo/estado celular contendo a proporção de cada um na composição do *spot*, como realizado pelo algoritmo *SPOTlight*. Os títulos dessas colunas precisam ser iguais aos *clusters* do scRNA-seq.

Figura 11. Visão geral da função *integr_intracellular_comm()*.



O objetivo primário da função *integr_intracellular_comm()* é auxiliar a predição de elementos celulares envolvidos em sinalizações e comunicações intracelulares determinadas em *spots*. Assim, nessa figura substituímos os spots brancos genéricos por imagens de células (comunicação entre o tipo celular “azul” e “vermelho”, sendo que as sinalizaçãos e comunicações intracelulares foram preditas no tipo celular “vermelho”).

A função *integr_intracellular_sign()* define possíveis tipos/estados celulares de ocorrência de vias de sinalização intracelular identificadas em *spots* se baseando na presença de transcritos. Essa função possui 6 parâmetros: 1 - *sample* (objeto *Seurat*), contendo um conjunto de dados de scRNA-seq 2 - *column_with_clusters* (carácter), o título da coluna de metadados contendo os *clusters,* 3 – *pair_of_clusters* (lógico) que informa, caso = *TRUE*, que o usuário analisa a

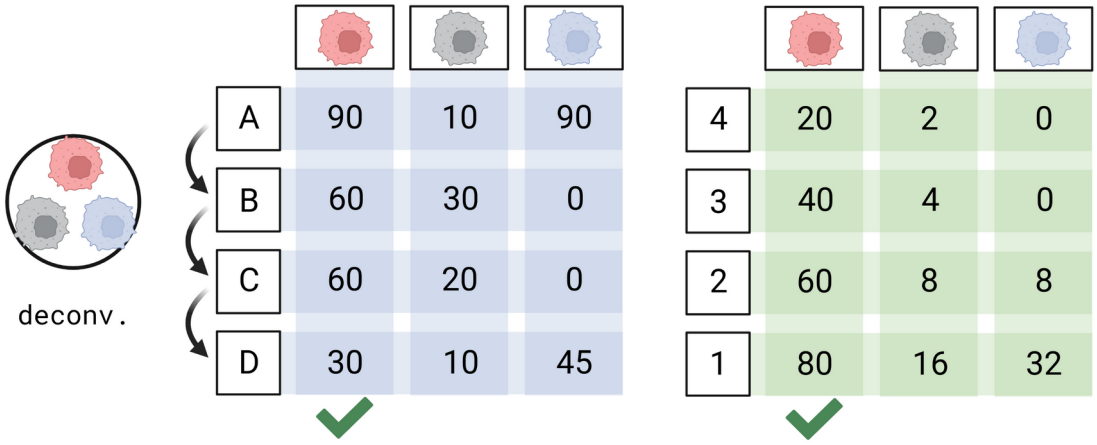
comunicação entre dois *clusters* transcricionais e, caso = *FALSE*, analisa a "autocomunicação" de um *cluster* transcricional, 4 - *column_start* (numérico), 5 - *column_end* (numérico) que informa a posição da primeira e da última coluna contendo no metadado as tipos/estados celulares (metadado do dado de ST, colunas geradas pela deconvolução) e 6 - *n_cores* (numérico), o mesmo que o parâmetro na função *define_intercellular_comm()*. Usando tal função nós definimos dois objetos, *intracellular_sp_01_integr* e *intracellular_sp_02_integr*. Estes, por sua vez, são cópias dos objetos *intracellular_sp_01_linker* e *intracellular_sp_02_linker* mas agora possuem colunas para cada tipo/estado celular presente em *column_with_clusters* contendo a presença, em proporção, das células com detecção de mRNA (*counts* > 0) para cada elemento da via de sinalização intracelular, nomeados iguais ao tipo/estado celular em questão (Figura 12), e colunas contendo a presença, em proporção, das células com detecção de n, n - 1, n - 2… elementos, ou seja, a proporção de co-ocorrência de elementos da via nomeados iguais ao tipo/estado celular em questão e o sufixo "*_co_oc"* (Figura 12). Colunas de tipos/estados celulares ausentes dos *spots* analisados são excluídas. Se algum elemento apresenta detecção em menos de 5 % das células dos tipos/estados celulares, a célula da tabela é definida como NA, para evitar confusão visual.

Exemplo: O *spot* X é composto somente de células do tipo "Epitelial" e apresenta uma via de sinalização intracelular entre o receptor A e o fator de transcrição E, a via A, B, C, D e E. Essa função cria uma coluna chamada "Epitelial" contendo a proporção das células (epiteliais) com detecção de mRNA (*counts* > 0) para cada elemento (ou seja, 5 valores numéricos) e uma coluna chamada "Epitelial_co_oc" contendo a proporção das células com detecção de 5, 4, 3 … elementos da via.

4.1.14 *filter_integr_intracelullar_sign()*

A função *filter_integr_intracellular_sign()* pode filtrar os objetos gerados por *integr_intracellular_sign()*. Os parâmetros são referentes ao 1 - *cluster* (analisado de forma conjunta, *mode* = *"spot_n"*, ou individual, *mode* = *"spot_y"*. Em caso de *mode*

Figura 12. Análise de o*utputs* da função *integr_intracellular_comm()*.



Para o *spot* X foram determinadas a frequência de células do tipo celular “vermelho”, “cinza” e “azul”. Na análise da via sinalização estimulatória que vai do receptor A ao fator de transcrição D foram detectados mais *counts* normalizados (matriz azul) e proporção de co-ocorrência (matriz verde) da detecção desses *counts* no tipo celular vermelho, sendo que 20 % das células do tipo celular “vermelho” no scRNA-seq mostraram detecção dos 4 genes dessa via.

= *“spot_n”* se deve utilizar cl_0X e em caso de *mode* = *“spot_y”*, se deve utilizar sp_0X, *clusters* em X) que desejamos filtrar, 2 - colunas dos objetos *intracellular_cl_01_integr*, *intracellular_cl_02_ integr*, *intracellular_sp_01_integr* e *intracellular_sp_02_integr* (“counts”, “cell_type_prop” e “cell_type_prop_co_oc”), todos separados por “_”. Os parâmetros são definidos por valores numéricos e marcam os cortes de filtragem.

Exemplo: Nós podemos filtrar as vias de sinalização, no *cluster* 1 (analisado de forma conjunta, *mode* = *“spot_n”*), com elementos (todos) apresentando expressão maior que 3 *counts*, detecção em mais de 10 % das células (por subtipo celular) e co-ocorrência de expressão mínima de 5 % utilizando os parâmetros *cl_01_counts = 3*, *cl_01_cell_type_prop = 10* e *cl_01_cell_type_prop_co_oc = 5*. Todavia, a mesma busca porém com *cluster* analisado de forma individual (*mode* = *“spot_y”*) se utilizam os parâmetros *sp_01_counts = 3*, *sp_01_cell_type_prop = 10* e *sp_01_cell_type_prop_co_oc = 5*.

4.2 ESTUDO DE CASO - PREPARAÇÃO DE DADOS DE ENTRADA

Para validar as funções de análise 2D, nós utilizamos o conjunto de dados disponibilizado na publicação de Wu e colaboradores, em 2021, na qual os autores realizaram transcriptômica espacial em 6 cortes, 2 de pacientes com câncer de mama do subtipo molecular ER-positivo e 4 com câncer de mama do subtipo molecular triplo-negativo. Os cortes tinham 10 µm de espessura e foram analisados em um arranjo de transcriptômica espacial (plataforma Visium da 10x Genomics) contendo um total de 4992 *spots* na área de captura e cada *spot* possui 55 µm de diâmetro com uma distância de centro a centro de 100 µm entre os *spots*.

Para validar as funções de análise 3D, nós utilizamos o conjunto de dados disponibilizado na publicação de Andersson e colaboradores, em 2021, na qual 8 pacientes com câncer de mama do subtipo molecular HER2-positivo foram analisados com ST. Em quatro pacientes (A-D) foram obtidos 6 cortes/seções de tecido e nos outros quatro pacientes (E-H) foram obtidos 3. Os cortes tinham 16 µm de espessura, as seções foram feitas com 32 µm de distância, e foram analisados em um arranjo de transcriptômica espacial (plataforma 1k arrays) contendo um total de 1007 *spots* na área de captura e cada *spot* possui 100 µm de diâmetro com uma distância de centro a centro de 200 µm entre os *spots*.

O pré-processamento, normalização, redução de dimensionalidade, definição de grupos e detecção de marcadores dos dados foi feito da mesma forma que a descrita em Andersson e colaboradores, em 2021, utilizando o texto e o código disponível no *GitHub* (https://github.com/almaan/her2st). Primeiro nós fundimos/juntamos os dados brutos (*counts*) das seções de cada paciente, filtramos os genes codificadores de proteínas ribossomais (RPL- e RPS-), mitocondriais (MT-) e genes MTRNR. Cada gene devia estar presente em mais de 20 *spots* por amostra e cada *spot* deveria possuir mais de 300 genes distintos ou eles seriam filtrados.

Os dados fundidos foram normalizados utilizando o método de regressão binomial negativa regularizada utilizando a função *SCTransform* do pacote *Seurat*. O total de genes variáveis utilizados nesta etapa foram definidos com a aplicação de corte de variância residual de 1,1 (parâmetro *variable.features.rv.th* = 1.1), os parâmetros *return.only.var.genes* e *variable.features.n* foram definidos com *FALSE* e

NULL, nessa ordem e o parâmetro *vars.to.regress* foi definido com os códigos de cada paciente para realizar uma correção de batch em que as diferenças técnicas entre replicatas (seções) foram detectadas e corrigidas. Para realizar a redução de dimensões, nós filtramos os genes variáveis conforme descrito, utilizando um método de ordenação de genes variáveis por autocorrelação espacial, definido por Andersson e colaboradores em 2021, calculando o coeficiente de correlação de Pearson para cada gene entre o vetor de expressão e o vetor de atraso espacial (*spatial lag*), definido da soma da expressão em *spots* vizinhos a todos os *spots*. Dessa forma, genes com o coeficiente de correlação maior que 0.1 foram mantidos entre os genes variáveis. Os cientistas do trabalho original também detectaram 21 genes variáveis que definiam/formavam um padrão de anel em diversas seções. Esse efeito não foi visto em todas as replicatas da biópsia do mesmo paciente e, sendo assim, foi considerada uma variação de caráter técnico. Dessa forma, os 21 genes foram eliminados do conjunto de genes variáveis. Para cada paciente, os genes variáveis foram utilizados como *input* para então calcular a fatoração de matriz não negativa (NMF, do inglês *Non-negative Matrix Factorization*) com 10 fatores utilizando a função *RunNMF* do pacote *STUtility*. Os fatores com padrões consistentes entre cortes/seções do mesmo paciente foram observados e mantidos no trabalho original e aqui. Ainda conforme descrito, nós construímos um grafo de vizinho mais próximo compartilhado (SNN, do inglês *Shared Nearest Neighbor*) a partir da matriz fatorial NMF selecionada utilizando a função *FindNeighbors* do pacote *Seurat*. Com esse grafo nós pudemos identificar grupos/*clusters* de *spots* com o algoritmo de agrupamento baseado em modularidade utilizando a função *FindClusters* do pacote *Seurat* com o parâmetro de resolução definido como 0,4. Para cada paciente nós caracterizamos os marcadores/genes diferencialmente expressos (DEG, do inglês differentially expressed genes) de cada *cluster* utilizando a função *FindAllMarkers* do pacote *Seurat*. Tal função faz um teste (teste de postos sinalizados de Wilcoxon) entre os *spots* de cada *cluster* e os outros (de outros *clusters*). Nós utilizamos os mesmos thresholds/limiares de corte de marcadores do trabalho original, 1 - valor de *p* ajustado menor que 0.01 e 2 - log (logaritmo natural) *fold change* maior que 0.15 e, sendo assim, mantendo apenas marcadores *up*-regulados.

Para as análises 3D nós alinhamos as imagens e obtivemos as coordenadas de *spots* pós-alinhamento destas seções utilizando a função *AlignImages()* do pacote *STUtility*. O alinhamento de imagens significa identificar uma função de transformação rígida que mapeie/remapeie *pixels* entre imagens tal que ambas as imagens estejam alinhadas. A função de transformação é aprendida usando o ICP (*Iterative Closest Point* ou ponto mais próximo iterativo, do inglês) em dois conjuntos de pontos que definem as bordas dos tecidos em imagens de coloração de hematoxilina-eosina das seções.

5. RESULTADOS

Nós utilizamos as funções criadas do *SpotComm* em casos de câncer de mama de subtipos moleculares de pior prognóstico base, o triplo-negativo e o HER2-positivo, focando em estruturas no tecido, definidos por *clusters* transcricionais, de significância no prognóstico e no tratamento destes: o *cluster* de estrutura linfoide terciária no câncer triplo-negativo e o *cluster* de respostas de interferon tipo I no câncer HER2-positivo.

O câncer de mama é uma condição complexa, heterogênea quanto as suas alterações moleculares, composições celulares e desfechos clínicos. Hoje se sabe que a diversidade fenotípica de prognose e de predição à eficácia da quimioterapia do câncer de mama está relacionada a diversidade de perfis de expressão gênica. Esse conhecimento é a base da taxonomia molecular do câncer de mama (BERNARD et al., 2009).

No começo da década de 2000 os cânceres de mama foram classificados em cinco tipos “intrínsecos”, ou subtipos moleculares, consoante a expressão de genes analisados em tecnologia de microarranjo e selecionados por algoritmos de classificação, ou aprendizagem automatizada. Os subtipos moleculares são inicialmente divididos em luminais e basais, por apresentarem perfis de expressão semelhantes as células de tecidos luminal-epitelial e basal-epitelial da mama. A principal distinção transcricional entre esses dois grupos é a expressão em maior nível do receptor de estrogênio (***ER***, do inglês *estrogen receptor*), codificado pelo gene *ESR1*, entre os cânceres do grupo luminal (PEROU et al., 2000).

O subtipo luminal é subdividido entre cânceres que apresentam baixa ou alta expressão do gene *MKI67*, codificador da proteína Ki-67, um marcador de proliferação. Sendo assim, os cânceres luminais pouco proliferativos são chamados de *Luminal A*, ou *tipo-normal* caso tenham um perfil transcricional similar ao de células adiposas, e os muito proliferativos são os *Luminal B* (SØRLIE et al., 2001). Essa diferença de expressão de Ki-67 é também relacionada ao prognóstico, sendo os cânceres de subtipo Luminal A geralmente os luminais de menor taxa de crescimento e grau de agressividade e consecutivamente os de melhor prognóstico. Quanto a terapia farmacológica dos subtipos Luminais A e B (os com mais

transcritos de ER) o tamoxifeno, um modulador seletivo do receptor de estrogênio, segue sendo a alternativa padrão em casos de câncer em estágio inicial ou avançado e em casos de mulheres com alto risco a doença.

Os subtipos de menor nível de expressão de ER, por sua vez, são subdivididos entre os *triplo-negativo* e os *enriquecidos em HER2*. Os triplo-negativo são assim ditos porque apresentam menor nível de expressão de alguns receptores de hormônios, o ER e o receptor de progesterona (*PR*, do inglês *progesterone receptor*) e do receptor do fator de crescimento epidermal conhecido como *HER2* (PEROU et al., 2000). Estes cânceres ER-, PR- e HER2- constituem de 10-20% do total de casos e são conhecidos como os de maior agressividade, com altas taxas de metastização e reaparecimento da doença, sendo assim os de pior prognóstico. Uma vez que os cânceres de subtipo triplo-negativo não têm seu crescimento consequente de hormônios, as terapias voltadas a hormônios, como tamoxifeno retratam baixa eficácia. A terapia farmacológica de cânceres triplo-negativo é composta geralmente de quimioterapia citotóxica (KUMAR; AGGARWAL, 2016). No primeiro semestre de 2019, as agências federais de regulamentação dos EUA e Brasil aprovaram o atezolizumab (nome comercial: Tecentriq) como imunoterapia de cânceres triplo-negativo positivos para expressão do ligante *PD-L1* (do inglês *Programmed Cell Death 1 Ligand 1*, codificado pelo gene *CD274*). Quando essa proteína se liga ao receptor *PD-1*, frequentemente encontrado em células T do sistema imune, ela inibe sua ativação e dessa forma ela inibe a resposta imune. O ligante PD-L1 pode ser expresso em cânceres, os auxiliando na sua não-detecção e decorrente eliminação causada pelo sistema imune. Sendo assim, o atezolizumab em cânceres triplo-negativo com PD-L1 atua inibindo essa proteína e propicia o reconhecimento deste tumor ao controle imune (SCHÜTZ et al., 2017; BARCLAY; CRESWELL; LEÓN, 2018). O segundo subtipo molecular com baixo nível de expressão de ER se chamam enriquecidos em HER2, ou HER2-positivos. São assim ditos por apresentarem geralmente amplificação do locus contendo o gene *HER2*. Essa presença de diversas cópias do gene *HER2* faz com este seja transcrito em abundância (PEROU et al., 2000). A interação entre o HER2 e o fator de crescimento epidermal ativa, dentre outros, vias de proliferação e divisão celular e sendo assim, os cânceres enriquecidos em HER2 exibem proliferação rápida com

altas taxas de metastização e recidiva. Estes cânceres ER-, PR- e HER2+ constituem de 20-30% do total de casos e a prognose é parecida aos dos triplo-negativo, entretanto, para os enriquecidos em HER2 há terapia farmacológica própria como o trastuzumab (nome comercial: Herceptin), anticorpos que se ligam ao HER2 na superfície das células dos tumores e inibe assim o recebimento de sinais de crescimento e divisão (KRISHNAMURTI; SILVERMAN, 2014).

5.1 ESTUDO DE CASO 1: CÂNCER DE MAMA DO SUBTIPO MOLECULAR TRIPLO-NEGATIVO

5.1.1 Definição de *clusters* transcricionais do paciente "B" de Wu e colaboradores, em 2021

Para demonstrar a capacidade do *SpotComm* de lidar com dados 2D na plataforma Visium (10x Genomics) nós utilizamos o conjunto de dados de transcriptômica espacial disponibilizado na publicação de Sunny Wu e colaboradores, em 2021 (WU et al., 2021b), de câncer de mama de subtipo molecular triplo-negativo. Nós analisamos a comunicação interna de um grupo/*cluster*, o *cluster* transcricional de "estruturas linfoides terciárias" definido pela co-localização de células B e T. Nós analisamos o *cluster* na seção do paciente "B" por ser a seção com mais *spots* analisados.

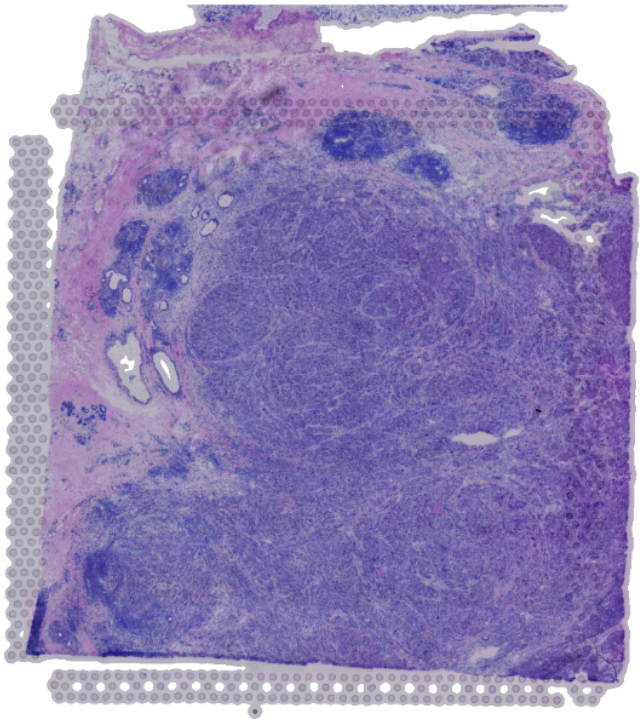
As estruturas linfoides terciárias são regiões com padrões de co-localização de células B e T, órgãos ectópicos linfoides que se desenvolvem em tecidos não linfoides em sítios de inflamação crônica, como tumores. São regiões atribuídas na literatura a respostas antitumorais de cunho imune e as suas relações ao desfecho clínico e as respostas ao tratamento e, sendo assim, sugere-se que as estruturas possuam fator prognóstico e preditivo. Em microambiente tumoral as estruturas linfoides terciárias são locais onde os antígenos do câncer são apresentados às células T, promovendo uma resposta direcionada (KINKER et al., 2021).

Wu e colaboradores realizaram transcriptômica espacial em 1 seção do câncer do paciente "B" e após as filtragens de genes e de *spots*, o conjunto de dados resultou em 16870 genes (19731 genes removidos) e 4890 *spots* (4 *spots*

removidos) e 58552069 identificadores moleculares únicos distribuídos na seção. A média de genes detectados foi de 3505 e a mediana foi de 3586.

Nós carregamos as imagens de coloração de hematoxilina-eosina das seções e “mascaramos” o *background* (plano de fundo) (Figura 13). Dessa forma nós obtivemos os valores de coordenada dos *spots*.

Figura 13. Região do tecido analisada usando Visium corada com hematoxilina-eosina.

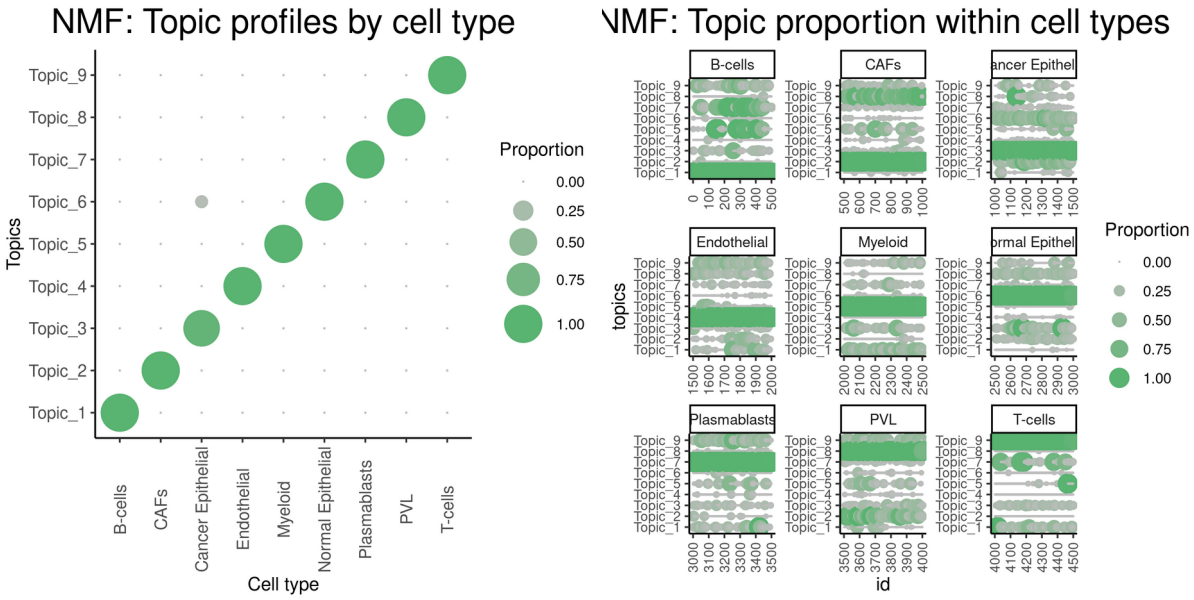


Após a prática de redução de dimensionalidade, nós obtivemos sete *clusters* transcricionais, numerados de zero a seis. O *cluster* número 6 contém o *cluster* transcricional de “estruturas linfoides terciárias”, discutido na publicação original, com 222 *spots*. Nós reproduzimos a análise de identificação de marcadores e identificamos 3255 genes marcadores do *cluster* de “estruturas linfoides terciárias” e 935 restaram após filtragem de log2 de fold change médio > 0.15 e de valor de *p* ajustado < 0.01. Todos os genes marcadores do *cluster* 6 e seus dados analisados estão contidos na Tabela suplementar 1.

Visando obter os termos de ontologia gênica relativos a processos biológicos no *cluster* de “estruturas linfoides terciárias” nós realizamos uma análise de enriquecimento funcional do tipo análise de enriquecimento gênico. Os 10 termos mais significativos no ponto de vista estatístico no *cluster* foram: “resposta imune adaptativa”, “ativação de célula T”, “ativação da regulação de linfócitos”, ativação da resposta imune”, “via de sinalização reguladora da resposta imune”, “regulação positiva da ativação celular”, “via de sinalização mediada por receptor de antígeno”, “proliferação de leucócitos”, “diferenciação de células mononucleares” e “adesão celular leucocitária”. A Tabela suplementar 2 contém os 586 termos enriquecidos no *cluster* de “estruturas linfoides terciárias”.

Nós utilizamos um conjunto de dados de scRNA-seq publicado por Wu e colaboradores em 2021 (WU et al., 2021b) para realizar a deconvolução de *spots* da seção do paciente “B” utilizando o pacote *SPOTlight*. Dessa forma nós realizamos a deconvolução de misturas/combinações de células utilizando uma referência de células únicas. O conjunto de dados de scRNA-seq contém 42512 células de câncer de mama do subtipo molecular triplo-negativo sequenciadas de 10 pacientes. Os autores do artigo original classificaram as células em três níveis de complexidade molecular, ou camadas, chamadas de camadas principais, secundárias e de subconjunto (terminologia de Wu e colaboradores). A camada principal possui nove tipos celulares: câncer epitelial, células B, células T, células semelhantes a perivasculares (PVLs, do inglês perivascular like cells), endoteliais, fibroblastos associados ao câncer (CAFs, do inglês cancer-associated fibroblasts), mielóides, normal epitelial e plasmablastos enquanto na camada de subconjunto há 48 subtipos celulares como subtipos de células B (memória e *naive*) e de células T (CD4:*CCR7; IL7R; MKI67* e outros). Na Figura 14 A é possível observar a alta especificidade do modelo onde cada perfil de tópico contém 100 % de cada tipo celular. Na Figura 14 B vemos os perfis de tópico individuais de cada célula em cada tipo celular e nós observamos que todas as células, de cada tipo celular, mostram distribuições de perfis de tópico similares.

Figura 14. Especificidade de assinaturas de tópicos obtidos do *SPOTlight* para cada identidade celular.



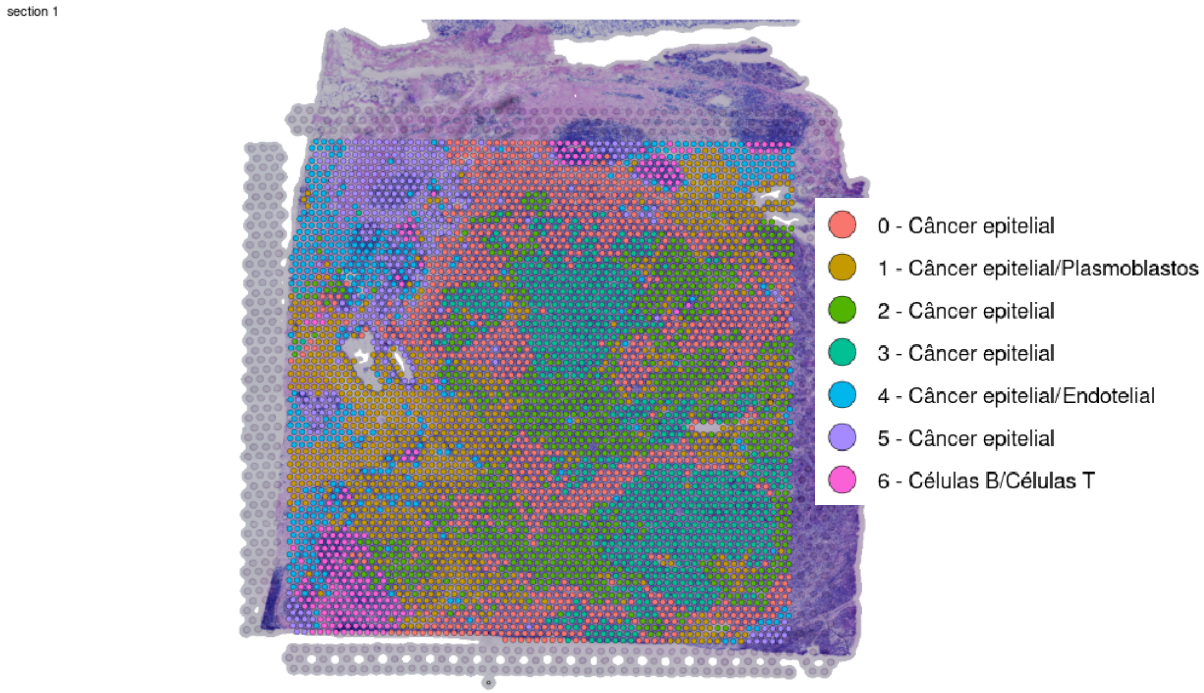
Outputs, objetos de saída, do algoritmo *SPOTlight* exibindo, em A em B, a alta proporção (na legenda, *proportion*) de células de um tipo celular em um dos tópicos. Em A vemos isso em conjunto e em B individualmente, para cada uma das 500 células utilizadas para aferir o modelo.

O próximo passo foi nomear os *clusters* conforme os tipos celulares mais presentes. Primeiro nós nomeamos cada *spot* conforme os tipos celulares presentes em proporção maior que 25 % e nomeamos o *spot* e então nomeamos cada *cluster* conforme os nomes de *spots* do *cluster* presentes em proporção maior que 10 %. Dessa forma nós nomeamos o *cluster* de “estruturas linfoides terciárias” como “Células B/Células T”. Os *clusters* transcricionais 0, 2, 3 e 5 foram nomeados de “Câncer epitelial”, o *cluster* 1 de “Câncer epitelial/Plasmoblastos” e o *cluster* 4 de “Câncer epitelial/Endotelial” (Figura 15).

5.1.2 Comunicação celular interna do *cluster* transcricional de “estruturas linfoides terciárias” em análise de forma conjunta (*mode* = “*spot_n*”)

Nós então analisamos a comunicação interna do *cluster* transcricional de “estruturas linfoides terciárias” utilizando o *SpotComm*. Primeiro nós utilizamos a função *get_subset_2D()* com os parâmetros *column_with_clusters* = “Coluna”

Figura 15. Áreas de análise transcricional definidas por seu *cluster* transcricional.



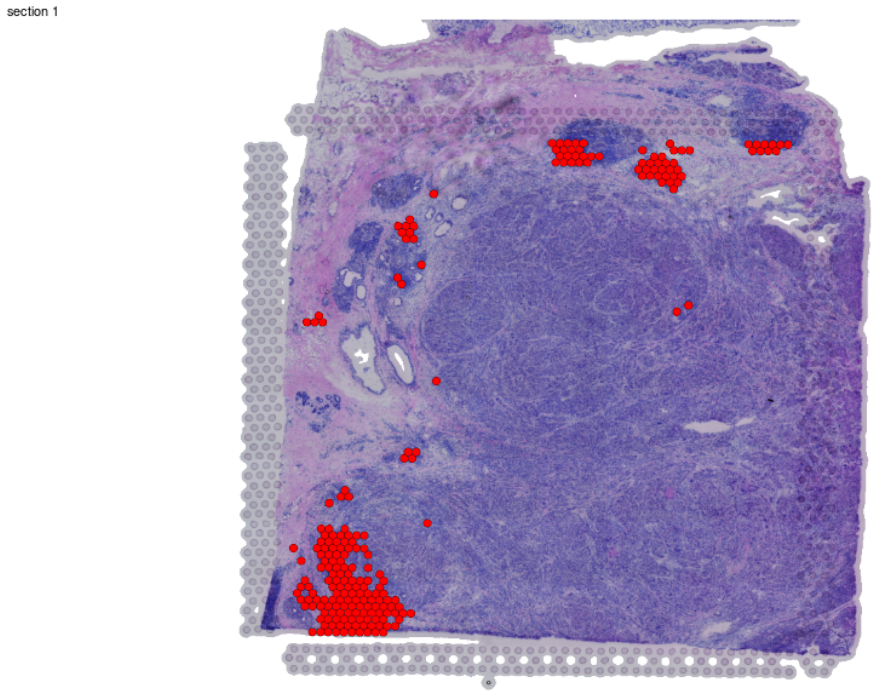
As áreas de análise transcricional, ou seja, os *spots*, coradas por seu *cluster* transcricional. Os *clusters* transcricionais 0, 2, 3 e 5, nomeados “Câncer epitelial”, estão corados de vermelho, verdes e roxo, o *cluster* 1, nomeado “Câncer epitelial/Plasmoblastos”, em amarelo, o *cluster* 4, nomeado “Câncer epitelial/Endotelial”, em azul e o *cluster* 6 aqui examinado, em rosa.

cluster_01 e *cluster_02* = “Células B/Células T” (*cluster* 6) e *contact* = *FALSE*, para obter/selecionar todos os *spots* do *cluster*. O subconjunto (o *subset*) de dados ao final continha 220 *spots* na seção. A Figura 16 mostra no tecido as posições de *spots* do *cluster*, em vermelho.

Na plataforma Visium da 10x Genomics os *spots* possuem 6 vizinhos próximos, ao norte, nordeste, sudeste, sul, sudoeste, e noroeste. Os vizinhos do norte e sul apresentam uma distância (em unidades de imagem) de ~ 22 (~ 160 µm), enquanto os vizinhos do nordeste, sudeste, sudoeste e noroeste (vizinhos da diagonal) de ~ 14 (~ 100 µm). Sendo assim, nós utilizamos a função *get_nearby_2D()* com os parâmetros *k_number* = 6 (número de vizinhos próximos) e *distance* = *25* (~ 180 µm no cálculo 2D). Nós obtivemos 1334 pares de *spots* vizinhos.

Nós buscamos por comunicações intercelulares mediadas por ligantes e

Figura 16. Áreas de análise transcricional do *cluster* de “estruturas linfoides terciárias”.



Os círculos em vermelho indicam as posições de *spots* do *cluster* de “estruturas linfoides terciárias” em seções do paciente “B”.

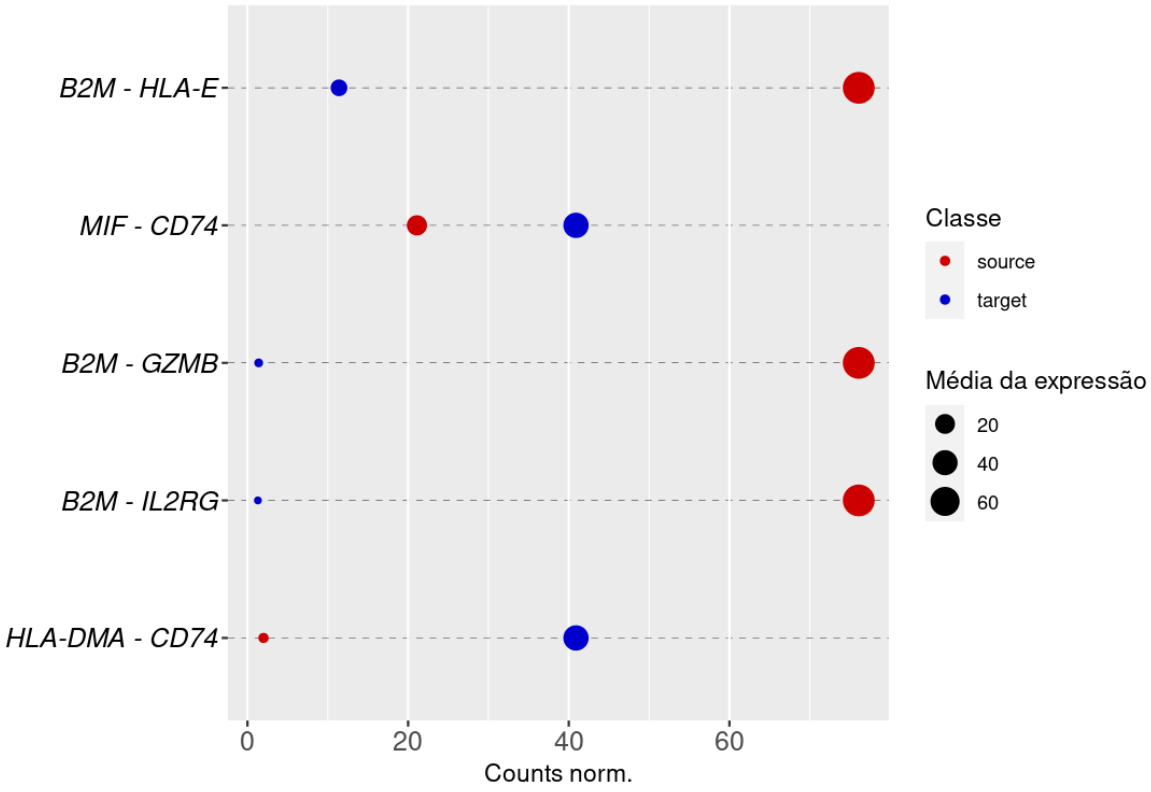
receptores e, para isso, nós utilizamos a função *define_intercellular_comm()* com os parâmetros *my_source = “ligand”, my_target = “receptor”, pair_of_clusters = FALSE*. Com isso, nós identificamos 373 ligantes e 1292 receptores no *cluster* transcricional de “estruturas linfoides terciárias” detectados em mais de 25 % dos *spots*. Os 3 termos mais significativos utilizando os ligantes e receptores detectados como base foram: “resposta imune adaptativa”, “migração de leucócitos” e “imunidade mediada por leucócitos mieloides”. A Tabela suplementar 3 contém os ligantes e receptores e a Tabela suplementar 4 contém os 1381 termos detectados.

Nós utilizamos a função *buzzer_intercellular_comm()* para identificar os pares de ligantes e receptores, utilizando os obtidos em *define_intercellular_comm()*, descritos na literatura. Na análise de forma conjunta, utilizando o parâmetro *mode = “spot_n”*, nós identificamos 168 pares de ligantes e receptores contendo 60 ligantes

e 76 receptores distintos que mostraram uma expressão média do ligante maior que 0.75 e do receptor maior que 0.25 *counts* normalizados (instituindo proporção de 3:1 entre ligante e receptor), um produto da expressão média maior que 1, um valor de *p* empírico menor que 0.05, ou seja, menos de 50 entre 1000 pares randômico de ligantes e receptores apresentam valores de produto da expressão maior que o par de ligante e receptor e a interação entre o ligante e o receptor deve resultar em estimulação (em consenso na literatura). A expressão média dos ligantes foi de 0.9 a 76.1 *counts* normalizados, sendo o ligante mais expresso o *B2M* (presente em todos os *spots* do *cluster*) e dos receptores foi de 0.3 a 40.9 *counts* normalizados, sendo o mais expresso o *CD74* (também presente em todos os *spots* do *cluster*). Ainda, 90 pares dentre esses mostraram valores de *p* do teste de correlação (de Pearson) entre os *spots* próximos menor que 0.1 e do teste de correlação entre os 1000 pares de *spots* randômicos maior que 0.1. Os coeficientes de correlação de testes destas interações foram de -0.05 a 0.29. Os 5 ligantes detectados com mais pares de receptores detectados foram: *FN1* (com interações com 21 receptores), *COL1A1* (10), *COL3A1* (9), *COL1A2* (8) e *CCL5* (7). Por sua vez, os 5 receptores detectados com mais pares de ligantes detectados foram: *ITGB1* (com interações com 10 ligantes), *CD4* (8), *CXCR4* (7), *CD44* (7), *SDC4* (6). Os 5 pares com os maiores valores de escore de comunicação (por produto da expressão média) foram: *B2M:HLA-E*, *MIF:CD74*, *B2M:GZMB*, *B2M:IL2RG* e *HLA-DMA:CD74* (Figura 17). A Tabela suplementar 5 contém o *output* da função com os 168 pares de ligantes e receptores.

Então nós utilizamos a função *buzzer_intracellular_comm*() para identificar os fatores de transcrição e seus genes alvo, ou seja, os *regulons* descritos na literatura e detectados no *cluster* transcricional de “estruturas linfoides terciárias”. Na análise de forma conjunta, utilizando o parâmetro *mode* = *“spot_n”*, nós identificamos 6 *regulons*, contendo 6 fatores de transcrição ou complexos de fatores de transcrição e 63 alvos distintos, que mostraram detecção do fator de transcrição em mais de 25 % dos *spots,* expressão média acima de 0.25, mais de 10 % dos alvos detectados no *cluster* e mostraram um valor de *p* empírico menor que 0.05, ou seja, menos de 50 entre 1000 valores de expressão média (em mediana) de genes randômicos do *cluster* apresentam valores maiores que os alvos em questão. A expressão média de fatores de transcrição foi de 0.5 a 0.9 *counts* normalizados,

Figura 17. Pares de ligantes e receptores com os maiores valores de escore de comunicação.

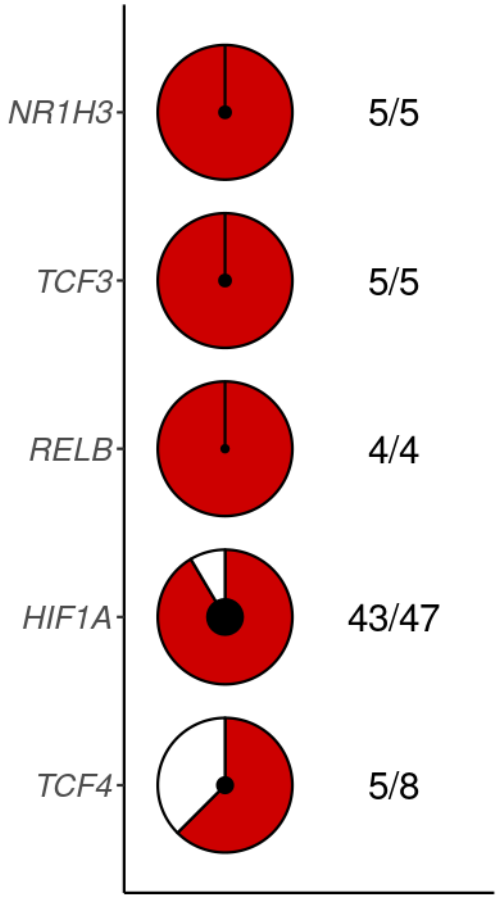


O eixo x define os *counts* normalizados e o eixo y os pares de ligantes e receptores. As cores indicam a classe do interator, em vermelho os *sources* (ligantes) e em azul os *targets* (receptores). Os tamanhos dos círculos indicam a média da expressão de interatores no *cluster*.

sendo o mais expresso o *TCF4* (presente em 50 % de *spots* do *cluster*). Os cinco fatores de transcrição com maior proporção dos alvos detectados foram: *NR1H3* (5/5 detectados), *TCF3* (5/5), *RELB* (4/4) *HIF1A* (43/47), *TCF4* (5/8) (Figura 18). A Tabela suplementar 6 contém o *output* da função com os 6 *regulons*.

O próximo passo foi predizer vias de sinalização intracelular entre os receptores e fatores de transcrição utilizando a informação de *counts* normalizados e de interações proteína-proteína descritas na literatura. Dessa forma, nós utilizamos a função *linker_intracellular_sign()* com os parâmetros *method = "counts"*, *my_receptor = c("HLA-E", "CD74", "GZMB", "IL2RG", "ERBB2")* (receptores de maiores valores de escore de comunicação) e *my_TF = c("NR1H3", "TCF3", "RELB", "HIF1A", "TCF4")*. Na análise de forma conjunta, utilizando o

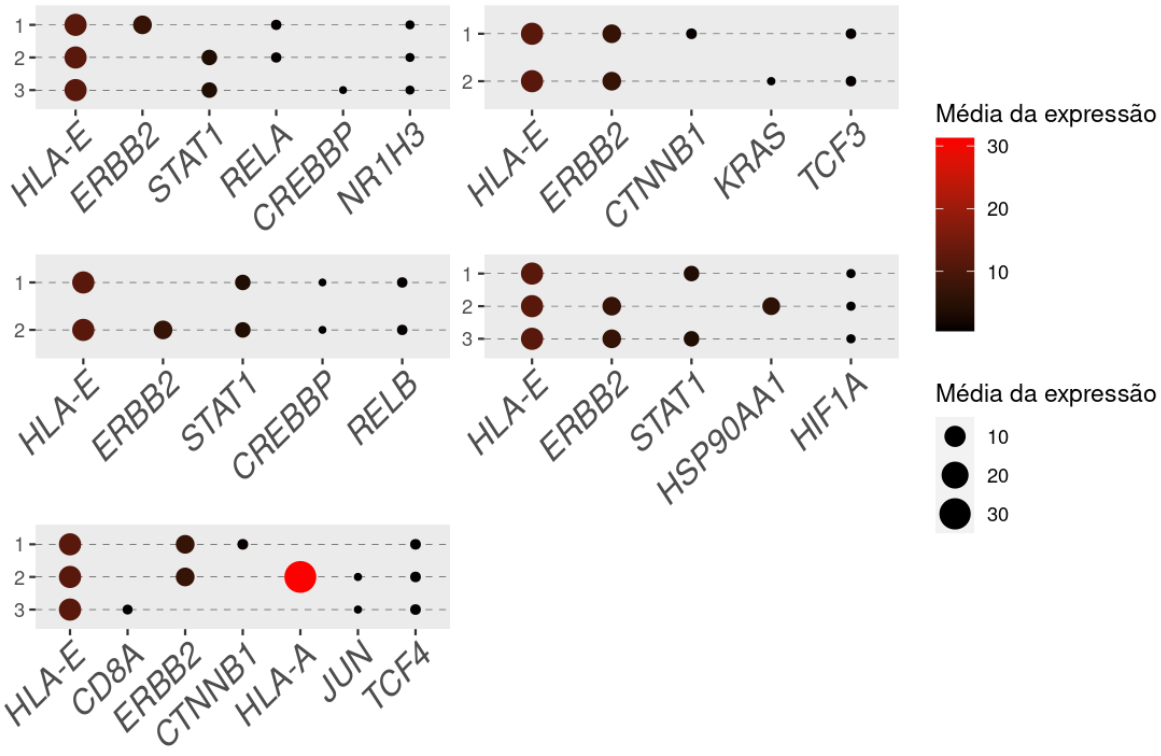
Figura 18. Fatores de transcricão com maior proporção dos alvos detectados.



O eixo y define os fatores de transcrição referidos. Os valores na direita definem o número de alvos detectados/número de alvos descritos. A imagem se refere a um *scatterpie*, é um círculo branco alusivo ao número de alvos descritos que é preenchido por cor vermelha alusiva ao número de alvos detectados. O centro do *scatterpie* (o círculo preto central) também se refere ao número de alvos detectados no *cluster*.

parâmetro *mode* = *"spot_n"*, nós identificamos 75 caminhos, vias de sinalização, contendo 3 caminhos para cada combinação (25) de receptor e fator de transcrição. Entre esses, 69 caminhos apresentaram todos os genes do caminho com uma média maior que 0.25 de *counts* normalizados. Com base nos critérios estabelecidos, nós identificamos caminhos de sinalização proteína-proteína entre os 5 receptores e os 5 fatores de transcrição. A Figura 19 mostra os caminhos identificados entre o receptor HLA-E e os fatores de transcrição . A Tabela suplementar 7 contém o *output* da função com os 69

Figura 19. Vias de sinalização estimulatórias preditas entre HLA-E e os fatores de transcrição presentes em mais *spots* com base nos transcritos.



Vias de sinalização estimulatórias, ou vias de interação proteína-proteína estimulatórias, entre o receptor HLA-E e os fatores de transcrição NR1H3, TCF3, RELB, HIF1A e TCF4. A função *linker_intracellular_sign()* buscou 3 possíveis vias para cada combinação entre receptor e fator de transcrição. As cores (de preto à vermelho) e os tamanhos dos círculos indicam a média da expressão de interatores no *cluster*.

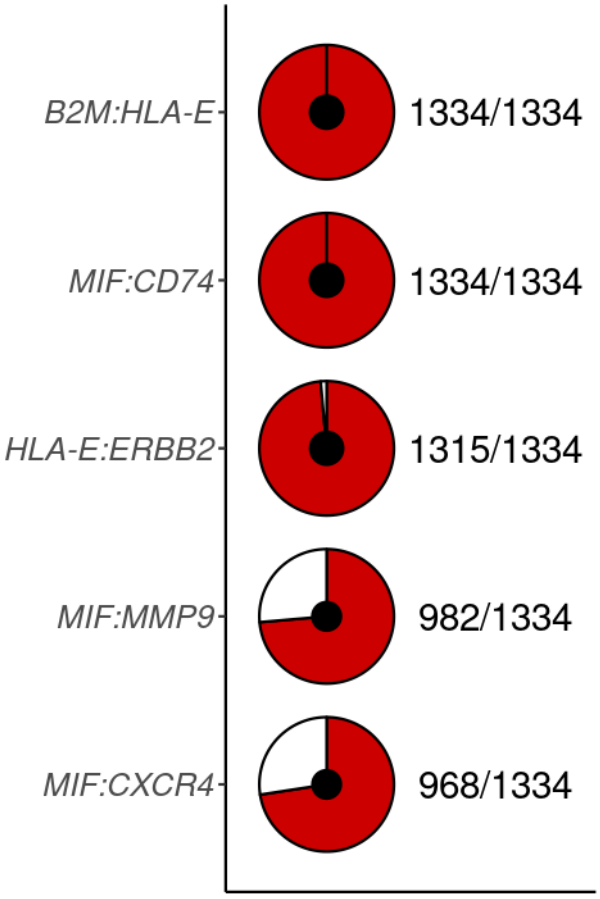
caminhos.

5.1.3 Comunicação celular interna do *cluster* transcricional de "estruturas linfoides terciárias" em análise de forma individual (*mode* = *"spot_y"*)

Então nós comparamos as análises de forma conjunta e de forma individual. Para isso, utilizamos os mesmos 168 ligantes e 373 receptores do *cluster* transcricional de "estruturas linfoides terciárias" detectados em mais de 25 % de *spots* obtidos em *define_intercellular_comm()*. Na análise de forma conjunta, utilizando a função *buzzer_intercellular_comm()* com o parâmetro *mode* = *"spot_y"*,

nós identificamos 91175 interações distribuídas entre 214 *spots*, contendo 540 pares distintos (525 em mais de 5 comunicações) entre 123 ligantes e 183 receptores que mostraram uma expressão do ligante maior que 3 e do receptor maior que 1 *counts* normalizados (ainda com a proporção de 3:1 entre ligante e receptor), um produto da expressão média maior que 3, um valor de *p* empírico menor que 0.05, ou seja, menos de 50 entre 1000 pares randômico de ligantes e receptores do *spot* em questão apresentam valores de produto da expressão maior que o par de ligante e receptor do *spot* e a interação entre o ligante e o receptor deve resultar em estimulação (em consenso na literatura). A expressão dos ligantes foi de 10 a 99 *counts* normalizados, sendo o ligante mais expresso o *B2M* (presente em todos os *spots* do *cluster* e com expressão de 99 em 40 comunicações) e dos receptores foi de 1 a 9 *counts* normalizados, sendo o mais expresso o *HLA-E* (também presente em todos os *spots* do *cluster* e com expressão de 9 em 103 comunicações). A Tabela suplementar 8 contém o *output* da função com os 91175 pares de ligantes e receptores. Os cinco pares de ligantes e receptores presentes em mais comunicações entre *spots* foram *B2M*:*HLA-E* (presente em 100 % das comunicações), *MIF:CD74* (100 %), *HLA-E:ERBB2* (98.5 %), *MIF:MMP9* (73.5 %), *MIF:CXCR4* (72.5 %) (Figura 20). Os 5 ligantes detectados com mais pares de receptores detectados foram: *FYN* (com interações com 25 receptores), *FN1* (21), *SYK* (19), *TGFB1* (18) e *TP53* (17). Por sua vez, os 5 receptores detectados com mais pares de ligantes detectados foram: *ITGB1* (com interações com 18 ligantes), *ITGB7* (17), *ITGA5* (16), *MAPK1* (15), *ITGAV* (15). Para cada interação *entre spots* nós comparamos o perfil de ligantes e receptores com as outras interações e, em média, as interações se asselham em 61.2 % (com mediana = 62.1 % e intervalo interquartil = 22.1 %). Ainda, nós analisamos a autocorrelação, ou correlação cruzada, espacial de pares de ligantes e receptores em comunicações de *spots* visando encontrar padrões espaciais na detecção da comunicação. Para isso, nós utilizamos o índice *I* de autocorrelação de Moran e detectamos 212 interações ligante-receptor espacialmente organizadas (valor de $p < 0.05$) com os índices de +0.02 a +0.24, ou seja, todos as interações significativas tendem ao *clustering* perfeito e não a dispersão perfeita. As 9 interações mais organizadas em coordenadas foram: *B2M:HLA-E* (Moran's I = 0.24), *CXCL13:CXCR4* (0.23),

Figura 20. Pares de ligantes e receptores presentes em mais comunicações entre *spots*.

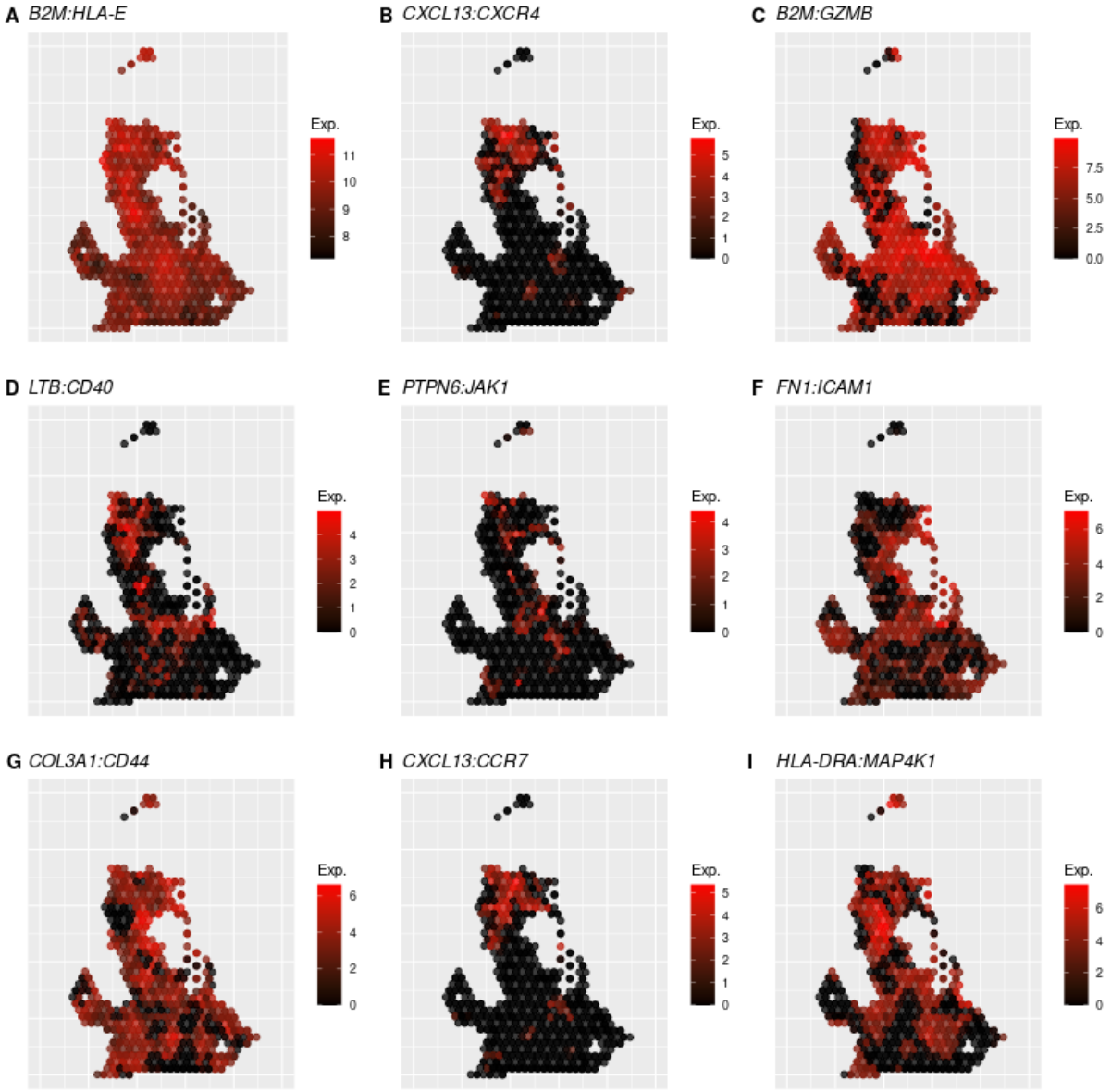


O eixo y define os pares de ligantes e receptores referidos. Os valores na direita definem o número de comunicações detectadas/número de comunicações viáveis no *cluster*. A imagem se refere a um *scatterpie*, é um círculo branco alusivo ao número de comunicações viáveis que é preenchido por cor vermelha alusiva ao número de comunicações detectadas. O centro do *scatterpie* (o círculo preto central) também se refere ao número de comunicações detectadas no *cluster*.

B2M:GZMB (0.22), *LTB:CD40* (0.19), *PTPN6:JAK1* (0.18), *FN1:ICAM* (0.17), *COL3A1:CD44* (0.17), *CXCL13:CCR7* (0.17) e *HLA-DRA:MAP4K1* (0.16) e estão mostradas na Figura 21. A Tabela suplementar 9 contém os valores de índices de Moran para as interações analisadas.

Então nós utilizamos a função *buzzer_intracellular_comm*() para identificar os *regulons* descritos na literatura e detectados no *cluster* transcricional de “estruturas linfoides terciárias”, porém com o parâmetro *mode* = *“spot_y”*, ou seja, na

Figura 21. Distribuição de comunicações intercelulares mais autocorrelacionadas.



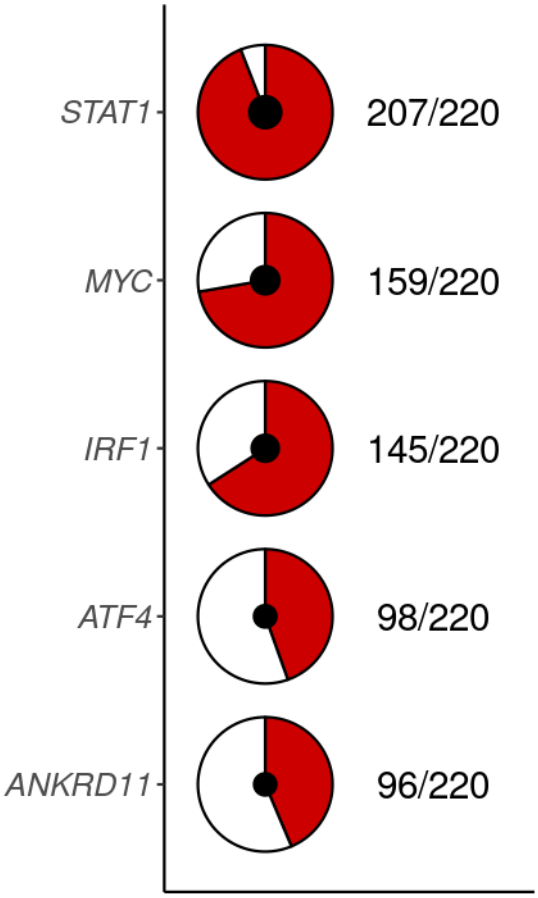
Distribuição e expressão de comunicações intercelulares expondo *spots* e espaços entre *spots.* As cores (de preto à vermelho) de círculos indicam o log2 produto da expressão de interatores na comunicação.

análise de forma individual. Nós identificamos 2770 *regulons* distribuídos entre 220 *spots*, contendo 80 fatores de transcrição ou complexos de fatores de transcrição (63 em mais de 5 *spots*) e 726 alvos distintos, que mostraram mais de 10 % dos alvos detectados no *spot* e mostraram um valor de *p* empírico menor que 0.1, ou seja, menos de 100 entre 1000 valores de expressão de genes randômicos do *spot*

apresentam valores maiores que os alvos em questão. A expressão de fatores de transcrição foi de 1 a 10 *counts* normalizados, sendo o mais expresso o *IRF1* e *STAT1* (presentes em 65.9 e 94.1 % de *spots* do *cluster* e com expressão de 10 em 1 *spot* cada). A Tabela suplementar 10 contém o *output* da função com os 2770 *regulons*. Os cinco fatores de transcrição presentes em mais *spots* foram *STAT1* (presente em 94.1 % dos *spots*), *MYC* (72.2 %), *IRF1* (65.9 %), *ATF4* (44.5 %), *ANKRD11* (43.6 %) (Figura 22). Para cada *spot* nós comparamos o perfil dos fatores de transcrição com os outros *spots* e, em média, as *spots* se asselham em 59.9 % (com mediana = 61.1 % e intervalo interquartil = 27.4 %). Utilizando o índice *I* de autocorrelação de Moran e detectamos 7 expressões de fatores de transcrição espacialmente organizadas (valor de $p < 0.05$) com os índices de +0.02 a +0.05, todas as expressões significativas tendem ao *clustering*. As 7 expressões mais organizadas em coordenadas foram: *RELA* (Moran's I = 0.05), *ANKRD11* (0.03), *IRF1* (0.03), *YY1* (0.02), *TCF3* (0.02), *NFKB1* (0.02), *SREBF1* (0.02) e estão mostradas na Figura 23. A Tabela suplementar 11 contém os valores de índices de Moran para as expressões analisadas.

O próximo passo foi predizer vias de sinalização intracelular entre os receptores e fatores de transcrição presentes em mais *spots* utilizando a informação de *counts* normalizados e de interações proteína-proteína descritas na literatura, porém com o parâmetro *mode = "spot_y"*, ou seja, na análise de forma individual. Dessa forma, nós utilizamos a função *linker_intracellular_sign()* com os parâmetros *method = "counts"*, *my_receptor = c("HLA-E", "CD74", "ERBB2", "MMP9", "CXCR4")* e *my_TF = c("STAT1", "MYC", "IRF1", "ATF4", "ANKRD11")*. Nós identificamos 9091 caminhos, vias de sinalização em potencial, contendo caminhos entre todos os fatores de transcrição e os receptores distribuídos em 220 *spots* com todos os genes do caminho com uma expressão maior que 1 *count* normalizado. A Tabela suplementar 12 contém o *output* da função com os 9091 caminhos. Os cinco pares entre os fatores de transcrição e os receptores presentes em mais *spots* (onde cada *spot* tem ao menos um caminho aceito) foram *HLA-E:STAT1* (presente em 94.1 % dos *spots*), *ERBB2:STAT1* (92.7 %), *ERBB2:ATF4* (90.1 %), *HLA-E:ATF4* (90.0 %) e *HLA-E:MYC* (72.2 %). A Figura 24 mostra como as vias de sinalização intracelular podem ser diferentes entre *spots* de um mesmo *cluster* transcricional.

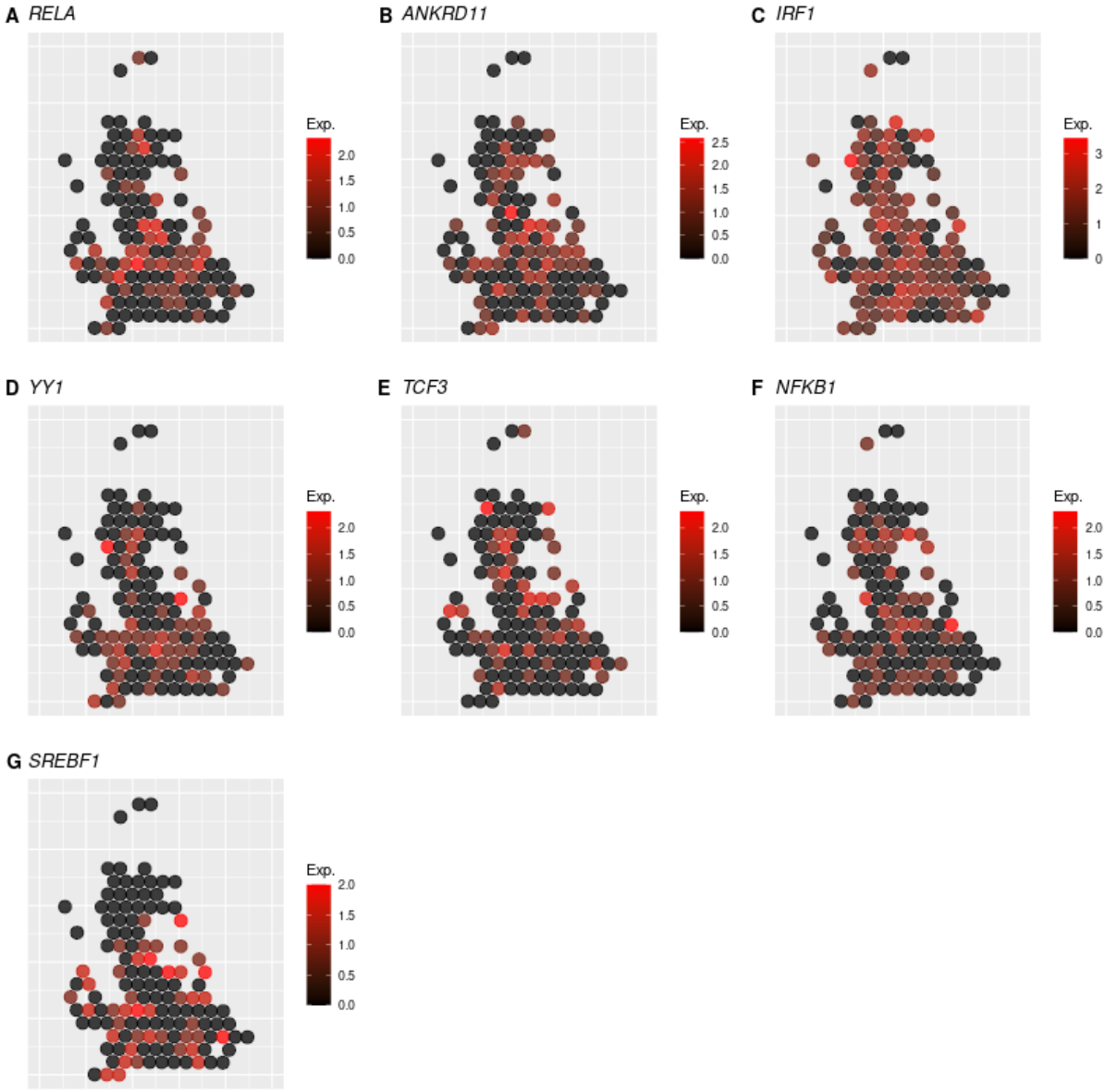
Figura 22. Fatores de transcrição operantes presentes em mais comunicações entre *spots*.



O eixo y define os fatores de transcrição referidos. Os valores na direita definem o número de ativações detectadas/número de *spots* no *cluster*. A imagem se refere a um *scatterpie*, é um círculo branco alusivo ao número de *spots* que é preenchido por cor vermelha alusiva ao número de ativações detectadas. O centro do *scatterpie* (o círculo preto central) também se refere ao número de ativações detectadas no *cluster*.

Mais uma vez nós utilizamos o conjunto de dados de scRNA-seq contendo as 42512 células de câncer de mama do subtipo molecular triplo-negativo para integrar com os dados de vias de sinalização obtidas utilizando a função *integr_intracellular_sign()*. Dos 9091 caminhos, 5050 retratam detecção de todos os genes em mais de 5 % de um tipo celular contido no *spot*. A Tabela suplementar 13 contém o *output* da função com os 5050 caminhos. Foram detectados possíveis caminhos em células B (2228 caminhos, 44.1 %), células T (3147, 62.3 %) e células

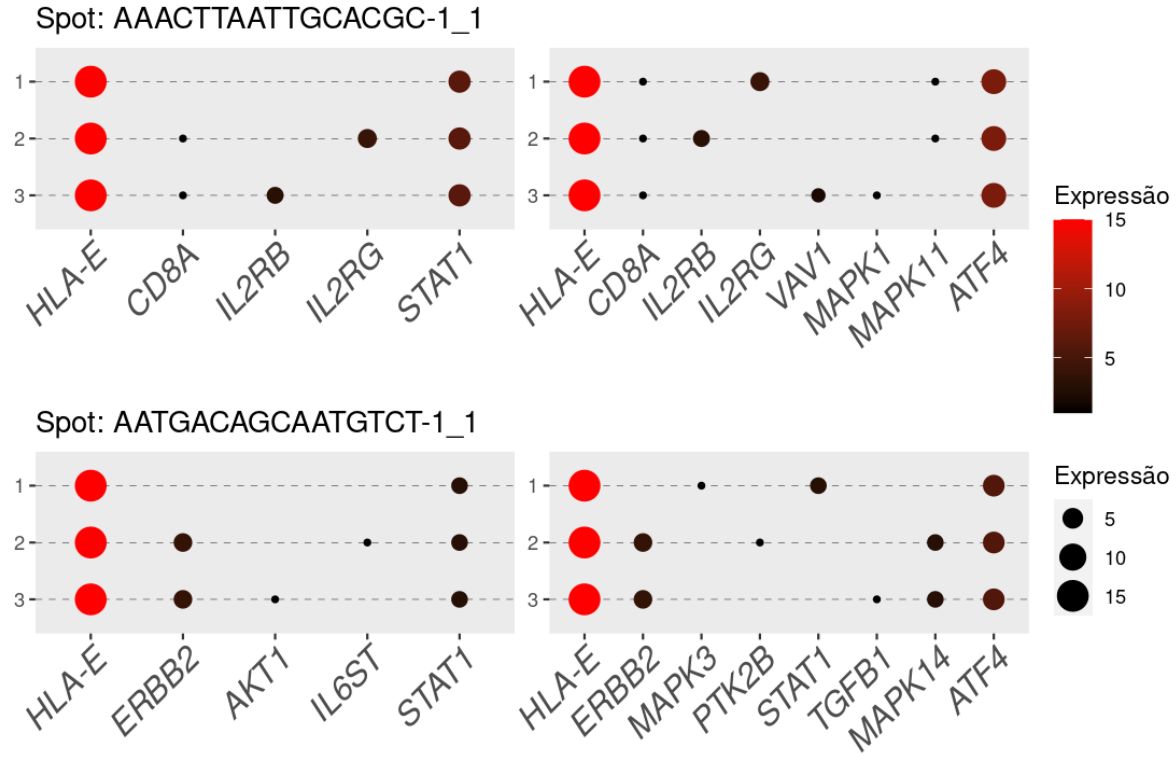
Figura 23. Distribuição de comunicações intracelulares mais autocorrelacionadas.



Distribuição e expressão de comunicações intracelulares expondo *spots.* As cores (de preto à vermelho) de círculos indicam o log2 da expressão de fatores de transcrição.

epiteliais (câncer) (2927, 57.9 %). Os caminhos supracitados envolvem os receptores *CD74*, *CXCR4* e *HLA-E* e os fatores de transcrição a *ATF4*, *IRF1*, *MYC* e *STAT1* em células B, T e epiteliais de câncer e, ainda, o receptor *ERBB2* em epiteliais de câncer. Os 3 caminhos mais detectados em células B e T foram: *HLA*

Figura 24. Exemplo de diferenças em vias de sinalização preditas de dois *spots.*

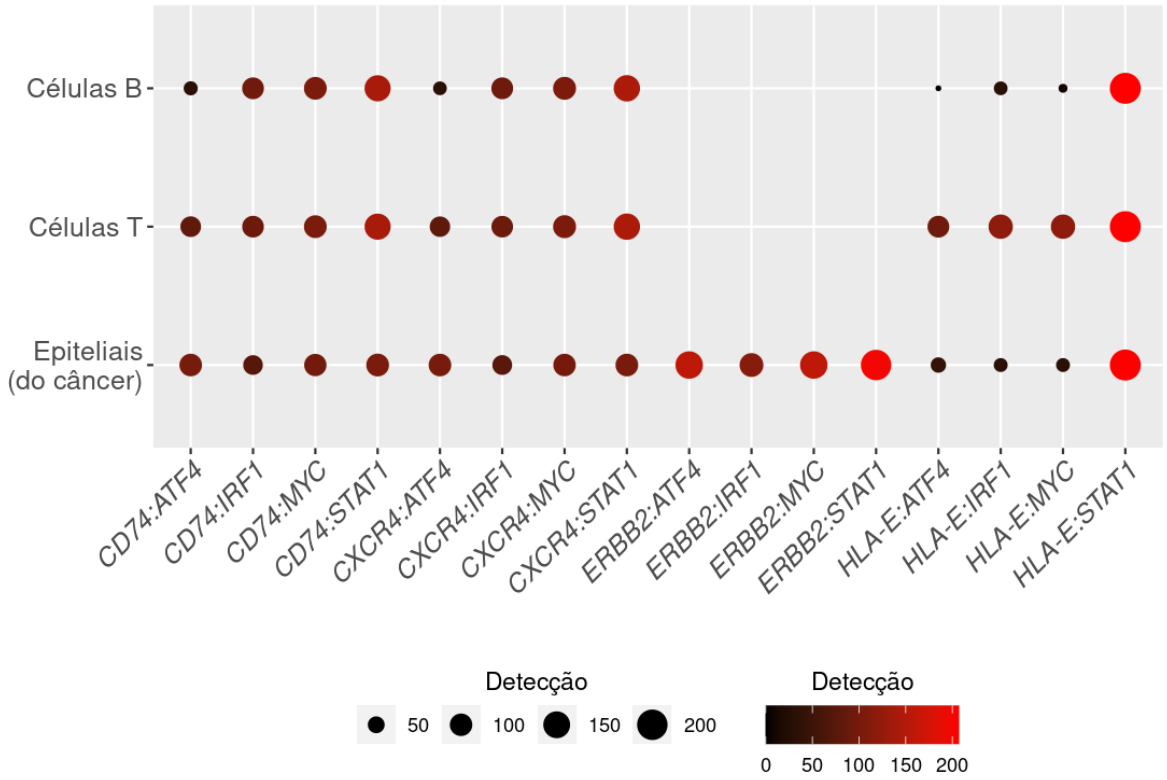


Vias de sinalização estimulatórias, ou vias de interação proteína-proteína estimulatórias, entre o receptor HLA-E e os fatores de transcrição STAT1 e ATF4 em dois *spots*. A função *linker_intracellular_sign()* buscou 3 possíveis vias para cada combinação entre receptor e fator de transcrição. As cores (de preto à vermelho) e os tamanhos dos círculos indicam a expressão de interatores no *spot*.

-E:STAT1 (presente em 207 *spots*), *CXCR4:STAT1* (144) e *CD74:STAT1* (141); em células epiteliais (câncer) foram: *HLA-E:STAT1* (presente em 207 *spots*), *ERBB2:STAT1* (199) e *ERBB2:MYC* (159). A Figura 25 mostra a proporção dos *spots* com a predição de cada uma das vias em cada tipo celular contido no *spots.*

Assim, a análise baseada em *spots*, utilizando o *mode* = *"spot_y"*, detectou, após filtragens, comunicações intercelulares envolvendo 123 ligantes e 183 receptores (representando 32.9 e 14.1 % dos ligantes e receptores expressos em mais de 25 % de *spots* do *cluster*) contra 61 ligantes e 76 receptores detectados na análise baseada em *clusters,* utilizando o *mode* = *"spot_n"* (representando 16.3 e 5.8 % dos ligantes e receptores expressos em mais de 25 % de *spots* do *cluster*). Um total de 172 genes de ligantes e receptores foram identificados em *mode* = *"spot_y"*

Figura 25. Elementos celulares viáveis de vias de sinalização preditas.



Proporção dos *spots* com a predição de cada uma das vias em linfócitos B, linfócito T e células epiteliais de câncer. As cores (de preto à vermelho) e os tamanhos dos círculos indicam o total de *spots* com a predição indicada.

e não em *mode* = *"spot_n"* e os 5 termos de ontologias contendo mais genes dessa lista foram: adesão célula-célula leucocitária, migração leucocitária, regulação positiva da adesão celular, proliferação leucocitária, regulação da adesão célula-célula. Ainda, há 33 termos envolvendo "células T", 8 envolvendo "células B", 15 envolvendo "linfócitos", 24 envolvendo "leucócitos" e outros. No total 703 ontologias tiveram enriquecimento e estão presentes na Tabela suplementar 14. A análise baseada em *spots* detectou 541 pares ligantes/receptores distintos (525 presentes em mais de 5 comunicações entre *spots*) contra 168 da análise baseada em *clusters*. Ainda, a análise baseada em *spots* detectou, após filtragens, comunicações intracelulares envolvendo 81 fatores de transcrição ou complexos de fatores de transcrição contra 6 detectados na análise baseada em *clusters*. Um total de 78 genes de fatores de transcrição foram identificados em *mode* = *"spot_y"* e não em

mode = *“spot_n”* e os alvos detectados em *spots* foram capazes de enriquecer 1685 termos de ontologias, presentes na Tabela suplementar 15, em que os 5 termos contendo mais genes dessa lista foram: resposta à drogas, resposta à molécula de origem bacteriana, resposta à substância inorgânica, proliferação de células epiteliais e transição de fase G1/S do ciclo celular. Ainda, há 51 termos envolvendo “células T”, 16 envolvendo “células B”, 16 envolvendo “linfócitos”, 34 envolvendo “leucócitos” e outros. Obtivemos a distribuição de comunicações e sinalizações, tornando possível as análises de autocorrelação espacial, onde diversos interatores inter e intracelulares estavam organizados em áreas. Com as vias de interação proteína-proteína resolvidas em *spots* foi possível predizer com maior segurança os elementos celulares responsáveis de sinalização, sendo que nesse exemplo, os conjuntos receptor/possíveis proteínas estimuladas/fator de transcrição (com os cinco receptores e fatores de transcrição detectados em mais *spots*) foram preditos em células B, T e epiteliais do câncer.

5.2 ESTUDO DE CASO 2: CÂNCER DE MAMA DO SUBTIPO MOLECULAR HER2- POSITIVO

5.2.1 Definição de *clusters* transcricionais do paciente “G” de Andersson e colaboradores, em 2021

Para demonstrar a capacidade do *SpotComm* de lidar com dados 3D, ou seja, fatias/*slices* de transcriptômica espacial obtidos de seções da mesma amostra, tratadas como replicatas, da plataforma “1k arrays”, nós utilizamos o conjunto de dados de transcriptômica espacial disponibilizado na publicação de Alma Andersson e colaboradores, em 2021 (ANDERSSON et al., 2021), de câncer de mama do subtipo molecular HER2-positivo. Nós analisamos a comunicação interna de um grupo/*cluster* discutidos na publicação original, o *cluster* transcricional de “respostas de interferon tipo I” definido pela co-localização de “macrófagos 2: *CXCL10*” e “células T: *IFIT1*”, ambos subtipos celulares bastante específicos, definidos por marcadores transcricionais. Nós analisamos o *cluster* em seções do paciente “G”, o mesmo utilizado pelos autores em figuras da publicação original.

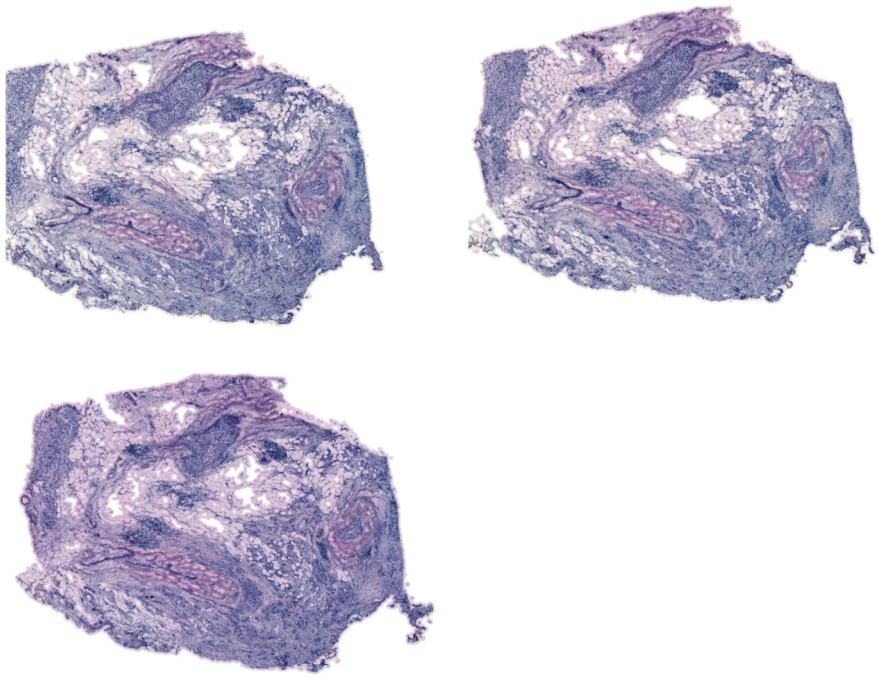
No trabalho de Andersson e colaboradores é descrito, em análises de scRNA-seq, que os “macrófagos 2: *CXCL10*” expressam um maior nível de quimioatraentes *CXCL9* e *CXCL10* e células mielóides associadas ao tumor, como macrófagos, que expressam essas moléculas foram atribuídas na literatura com funções antitumorais induzidas por imunoterapia. Ambas as quimiocinas *CXCL9* e *CXCL10* se ligam no receptor *CXCR3*, geralmente encontrado em células T e NK. Além do mais, a expressão de *CXCL9* e *CXCL10* pode ser induzida por estímulo de interferon tipo I e, da mesma forma, diversos marcadores de “células T: *IFIT1*” podem ser associados a respostas de interferon tipo I. A ativação de interferon tipo I em microambiente tumoral pode agir diretamente em células do tumor inibindo a proliferação ou estimulando a apoptose ou indiretamente ativando o sistema imune com propósitos antitumorais. Também é sabido que certas terapias anticâncer induzem e dependem de ativação de interferon tipo I. Devido a tal relevância das respostas de interferon tipo I no tratamento do câncer, nós avaliamos a comunicação celular da região definida pela co-localização de “macrófagos 2: *CXCL10*” e “células

T: *IFIT1*” e com genes marcadores associados às respostas de interferon tipo I.

Andersson e colaboradores realizaram transcriptômica espacial em 3 seções do câncer do paciente “G” e após as filtragens de genes e de *spots*, o conjunto de dados resultou em 11.213 genes (5.135 genes removidos) e 1.352 *spots* (19 *spots* removidos; 437, 466 e 449 *spots* em cada seção) e 7.770.417 identificadores moleculares únicos distribuídos entre as três seções. Para cada uma das seções a média de genes detectados por *spot* foi de 1.271, 1.354 e 1.190 e a mediana foi de 1.045, 1.158 e 1.006, respectivamente.

Nós carregamos as imagens de coloração de hematoxilina-eosina das seções, “mascaramos” o *background* (plano de fundo) e alinhamos as seções (Figura 26). Dessa forma nós obtivemos os valores de coordenada pós-alinhamento dos *spots*.

Figura 26. Região do tecido analisada usando Visium corada com hematoxilina-eosina.



Após a prática de redução de dimensionalidade, nós obtivemos sete *clusters* transcricionais, numerados de zero a seis. O *cluster* com a menor diferença de proporção de *spots* de diferentes seções foi o *cluster* 3 com 33.5, 33.5 e 33.0 % e o com a maior diferença foi o *cluster* 4 com 40.5, 31.1 e 28.4 % dos *spots* das seções 1, 2 e 3, respectivamente. Sendo assim, é possível observar que os *clusters* transcricionais estão presentes em todas as seções e em proporções similares.

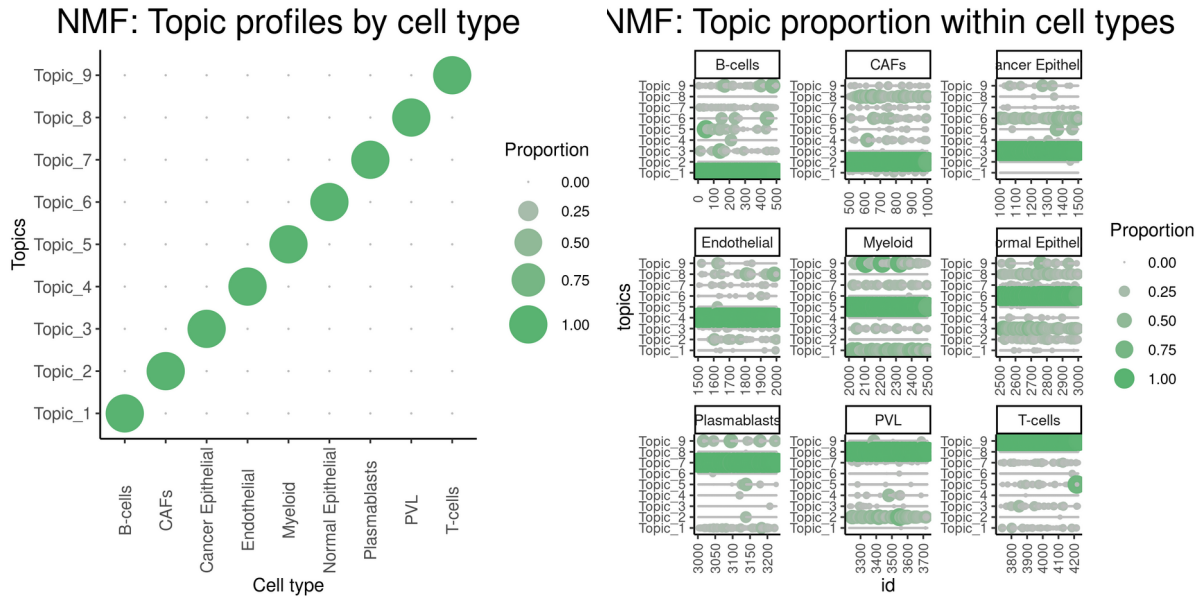
O *cluster* número 4 contém o *cluster* transcricional de “respostas de interferon tipo I” caracterizado e discutido na publicação original. Nós reproduzimos a análise de identificação de marcadores e identificamos 1130 genes marcadores do *cluster* de “respostas de interferon tipo I” e 416 restaram após filtragem de log2 de *fold change* médio > 0.15 e de valor de *p* ajustado < 0.01. Todos os genes marcadores do *cluster* 4 e seus referentes dados analisados estão contidos na Tabela suplementar 16.

Visando obter os termos de ontologia gênica relativos a processos biológicos no *cluster* de “respostas de interferon tipo I” nós realizamos uma análise de enriquecimento funcional do tipo análise de enriquecimento gênico. Os 10 termos mais significativos no ponto de vista estatístico no *cluster* de respostas de interferon tipo I foram: “resposta a interferon do tipo I”, “regulação negativa da replicação do genoma viral”, “cornificação”, “queratinização” e “regulação negativa do processo viral”, “regulação da ciclagem viral”, “replicação do genoma
viral”, “resposta a vírus” e resposta de defesa a vírus” consistente com os processos biológicos atribuidos a este *cluster* na publicacao inicial. A Tabela suplementar 17 contém os 60 termos enriquecidos no *cluster* de “respostas de interferon tipo I”.

Nós utilizamos um conjunto de dados de scRNA-seq publicado por Wu e colaboradores em 2021 (WU et al., 2021b) para realizar a deconvolução de *spots* de seções do paciente G utilizando o pacote *SPOTlight*. Dessa forma nós realizamos a deconvolução de misturas ou combinações de células utilizando uma referência de células únicas. O conjunto de dados de scRNA-seq contém 19311 células de câncer de mama do subtipo molecular HER2-positivo sequenciadas de 5 pacientes. Os autores do artigo original classificaram as células em três níveis de complexidade molecular, ou camadas, chamadas de camadas principais, secundárias e de subconjunto (terminologia de Wu e colaboradores). A camada principal possui nove

tipos celulares: câncer epitelial, células B, células T, células semelhantes a perivasculares (PVLs, do inglês *perivascular like cells*), endoteliais, fibroblastos associados ao câncer (CAFs, do inglês *cancer-associated fibroblasts*), mielóides, normal epitelial e plasmablastos enquanto na camada de subconjunto há 48 subtipos celulares como macrófagos 2: *CXCL10* e células T: *IFIT1*. Na Figura 27 A é possível observar a alta especificidade do modelo onde cada perfil de tópico contém 100 % de cada tipo celular. Na Figura 27 B vemos os perfis de tópico individuais de cada célula em cada tipo celular e nós observamos que todas as células, de cada tipo celular, mostram distribuições de perfis de tópico similares.

Figura 27. Especificidade de assinaturas de tópicos obtidos do *SPOTlight* para cada identidade celular.

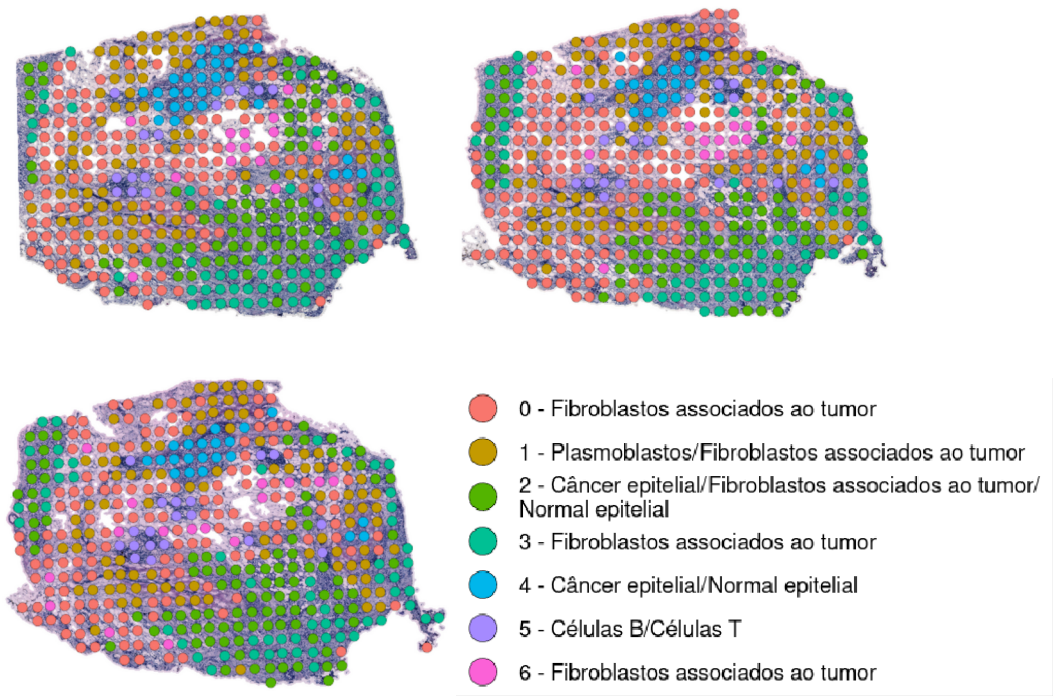


Outputs, objetos de saída, do algoritmo *SPOTlight* exibindo, em A em B, a alta proporção (na legenda, *proportion*) de células de um tipo celular em um dos tópicos. Em A vemos isso em conjunto e em B individualmente, para cada uma das 500 células utilizadas para aferir o modelo.

O próximo passo foi nomear os *clusters* conforme os tipos celulares mais presentes. Primeiro nós nomeamos cada *spot* conforme os tipos celulares presentes em proporção maior que 25 % e nomeamos o *spot* e então nomeamos cada *cluster* conforme os nomes de *spots* do *cluster* presentes em proporção maior que 10 %. Dessa forma nós nomeamos o de respostas de interferon tipo I como “Câncer

epitelial/Normal epitelial". Os *clusters* transcricionais 0, 3 e 6 foram nomeados de "Fibroblastos associados ao tumor", o *cluster* 1 de "Plasmoblastos/Fibroblastos associados ao tumor", o *cluster* 2 de "Câncer epitelial/Fibroblastos associados ao tumor/Normal epitelial", o *cluster* 4 de "Câncer epitelial/Normal epitelial", e o *cluster* 5 de "Células B/Células T" (Figura 28).

Figura 28. Áreas de análise transcricional definidas por seu *cluster* transcricional.



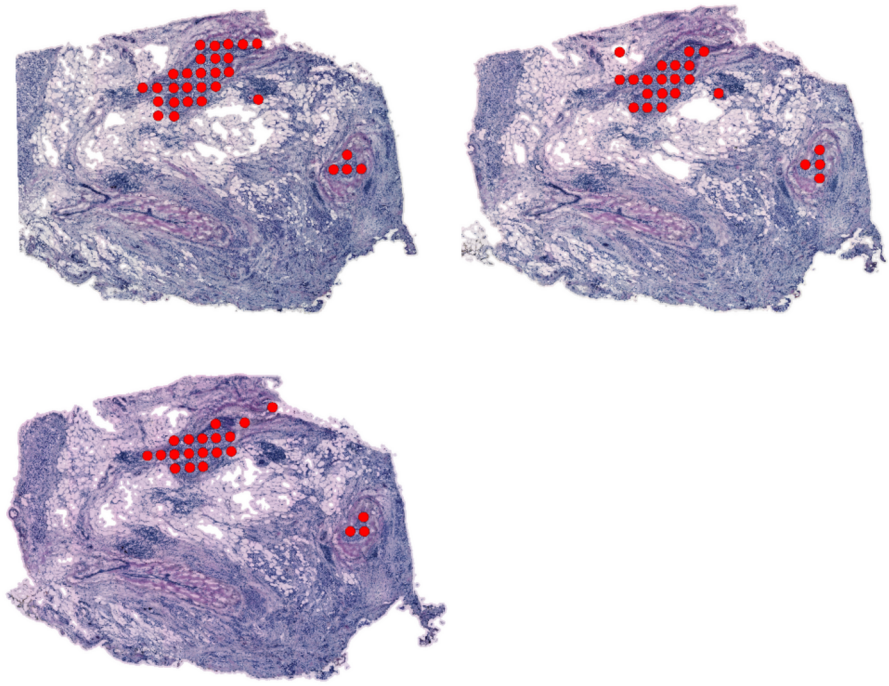
As áreas de análise transcricional, ou seja, os *spots*, coradas por seu *cluster* transcricional. Os *clusters* transcricionais 0, 3 e 6, nomeados "Fibroblastos associados ao tumor", estão corados de vermelho, verde-claro e rosa, o *cluster* 1, nomeado "Plasmoblastos/Fibroblastos associados ao tumor", em amarelo, o *cluster* 2, nomeado "Câncer epitelial/Fibroblastos associados ao tumor/Normal epitelial", em verde, o *cluster* 4 (aqui analisado), nomeado "Câncer epitelial/Normal epitelial", em azul e o *cluster* 5, nomeado "Células B/Células T", em rosa.

5.2.2 Comunicação celular interna do *cluster* transcricional de "respostas de interferon tipo I" em análise de forma conjunta (*mode* = *"spot_n"*)

Nós então analisamos a comunicação interna do *cluster* transcricional de

“respostas de interferon tipo I” utilizando o *SpotComm*. Primeiro nós utilizamos a função *get_subset_3D()* com os parâmetros *cluster_01* e *cluster_02* = “Câncer epitelial/Normal epitelial” (*cluster* 4), e *contact* = *FALSE*, para obter/selecionar todos os *spots* do *cluster*. O subconjunto (o *subset*) de dados ao final continha 74 *spots* 30, 23 e 21 na primeira, segunda e terceira seção, nessa ordem. A Figura 29 mostra no tecido as posições de *spots* do *cluster*, em vermelho.

Figura 29. Áreas de análise transcricional do *cluster* de “respostas de interferon tipo I”.



Legenda: Os círculos em vermelho indicam as posições de *spots* do *cluster* de “respostas de interferon tipo I” em seções do paciente G.

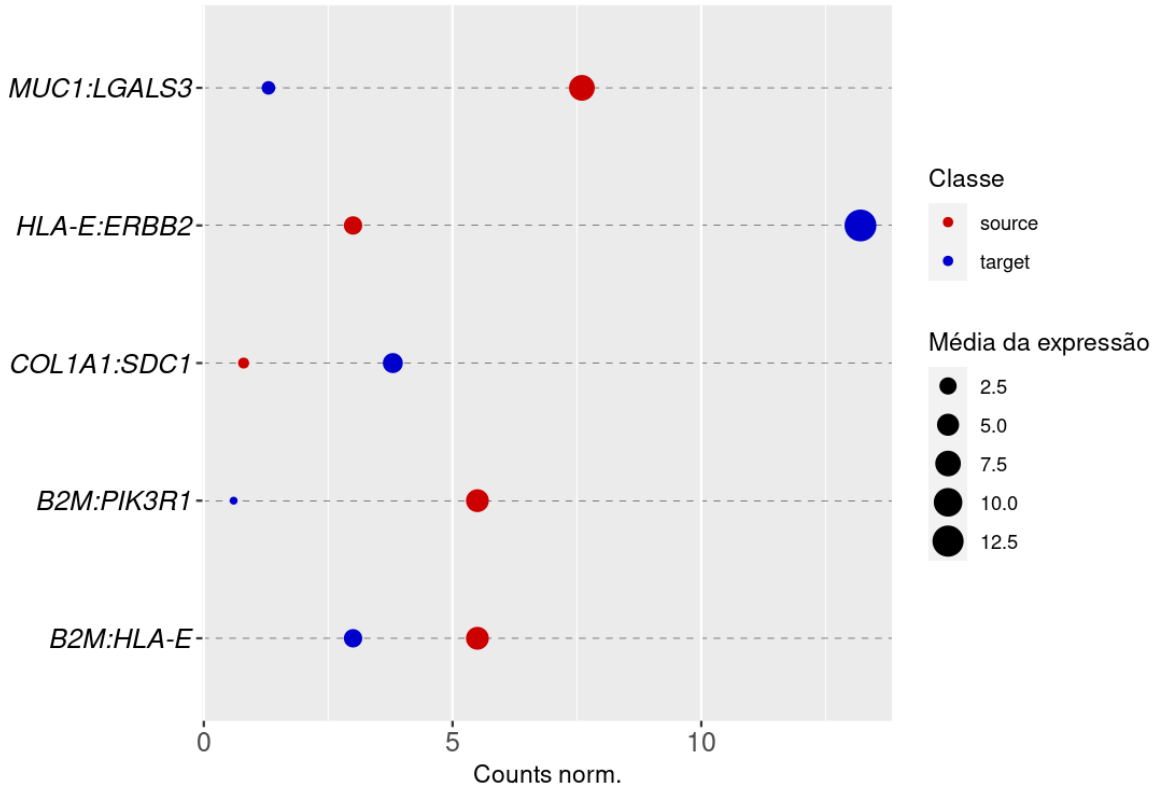
Na plataforma “1k arrays” cada *spot* possui 8 vizinhos próximos, ao norte, nordeste, leste, sudeste, sul, sudoeste, oeste e noroeste. Os vizinhos do norte, leste, sul e oeste apresentam uma distância (em unidades de imagem) de ~ 300 (~ 200 µm), enquanto os vizinhos do nordeste, sudeste, sudoeste e noroeste (vizinhos da diagonal) de ~ 400 (~ 265 µm). Sendo assim, nós utilizamos a função *get_nearby_3D()* com os parâmetros *k_number = 8*, *distance* e *distance_h = 450*

(para seção, cálculo 2D, e para seções, cálculo 3D, ou distância da hipotenusa) e *distance_z = 32* (distância entre as seções/cortes). Nós obtivemos 762 pares de *spots* vizinhos, 378 na mesma seção e 384 entre uma seção e outra.

Nós buscamos por comunicações intercelulares mediadas por ligantes e receptores de um *cluster* e, para isso, nós utilizamos a função *define_intercellular_comm()* com os parâmetros *my_source = "ligand"*, *my_target = "receptor"* e *pair_of_clusters = FALSE*. Com isso, nós identificamos 168 ligantes e 370 receptores no *cluster* transcricional de respostas de interferon tipo I detectados em mais de 25 % dos *spots*. Os 3 termos mais significativos utilizando os ligantes e receptores detectados como base foram: "migração de leucócitos", "ativação de leucócitos mielóides e "imunidade mediada por leucócitos mielóides". A Tabela suplementar 18 contém os ligantes e receptores e a Tabela suplementar 19 contém os 808 termos detectados. Nós utilizamos a função *buzzer_intercellular_comm()* para identificar os pares de ligantes e receptores, utilizando os obtidos em *define_intercellular_comm()*, descritos na literatura. Na análise de forma conjunta, utilizando o parâmetro *mode = "spot_n"*, nós identificamos 17 pares de ligantes e receptores contendo 12 ligantes e 13 receptores distintos que mostraram uma expressão média do ligante maior que 0.75 e do receptor maior que 0.25 *counts* normalizados (instituindo proporção de 3:1 entre ligante e receptor, 3 e 1 *counts* em 25 % de *spots* é igual a 0.75 e 0.25 *counts* em expressão média), um produto da expressão média maior que 1, um valor de *p* empírico menor que 0.05, ou seja, menos de 50 entre 1000 pares randômico de ligantes e receptores apresentam valores de produto da expressão maior que o par de ligante e receptor e a interação entre o ligante e o receptor deve resultar em estimulação (em consenso na literatura). A expressão média dos ligantes foi de 0.8 a 7.6 *counts* normalizados, sendo o ligante mais expresso o *MUC1* (presente em todos os *spots* do *cluster*) e dos receptores foi de 0.4 a 13.2 *counts* normalizados, sendo o mais expresso o *ERBB2* (também presente em todos os *spots* do *cluster*). Ainda, 7 pares ligantes e ereceptores dentre esses mostraram valores de *p* do teste de correlação entre os *spots* próximos menor que 0.1 e do teste de correlação entre os 1000 pares de *spots* randômicos maior que 0.1. Os coeficientes de correlação de testes destas interações foram de -0.14 a 0.12. Quatro ligantes detectados exibiram ligações com dois ou

mais receptores detectados: *COL1A1* (com interações com 3 receptores), *TNFSF10* (2), *COL1A2* (2) e *B2M* (2). Por sua vez, três receptores detectados exibiram ligações com dois ou mais ligantes detectados: *PIK3R1* (com interações com 3 ligantes), *SDC4* (2) e *SDC1* (2). Os 5 pares com os maiores valores de escore de comunicação (por produto da expressão média) foram: *HLA-E:ERBB2*, *B2M:HLA-E*, *MUC1:LGALS3*, *B2M:PIK3R1* e *COL1A1:SDC1* (Figura 30). A Tabela suplementar 20 contém os o *output* da função com os 17 pares de ligantes e receptores.

Figura 30. Pares de ligantes e receptores com os maiores valores de escore de comunicação.



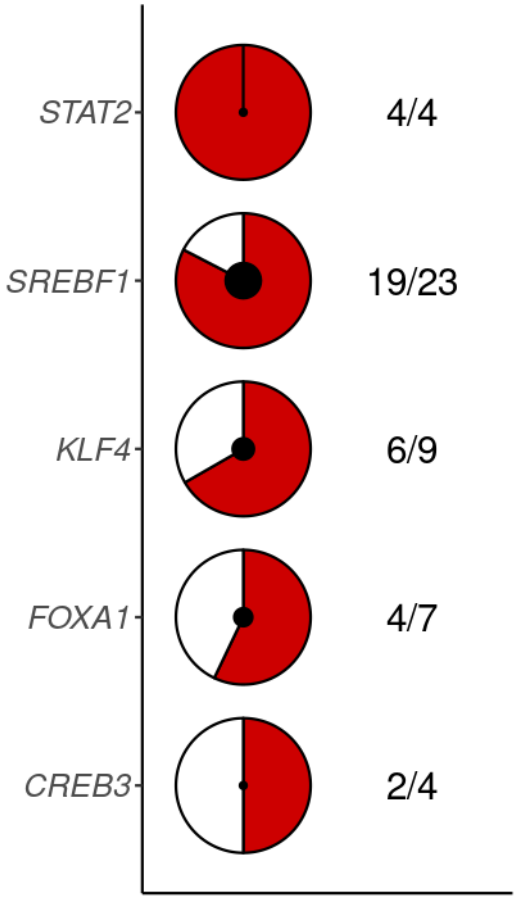
O eixo x define os *counts* normalizados e o eixo y os pares de ligantes e receptores. As cores indicam a classe do interator, em vermelho os *sources* (ligantes) e em azul os *targets* (receptores). Os tamanhos dos círculos indicam a média da expressão de interatores no *cluster*.

Então nós utilizamos a função *buzzer_intracellular_comm()* para identificar os fatores de transcrição e seus genes alvo, ou seja, os *regulons* descritos na literatura e detectados no *cluster* transcricional de “respostas de interferon tipo I”. Na

análise de forma conjunta, utilizando o parâmetro *mode* = *“spot_n”*, nós identificamos 7 *regulons*, contendo 7 fatores de transcrição ou complexos de fatores de transcrição e 38 alvos distintos, que mostraram detecção do fator de transcrição em mais de 25 % dos *spots* e expressão média acima de 0.25 e mais de 10 % dos alvos detectados no *cluster* e mostraram um valor de *p* empírico menor que 0.05, ou seja, menos de 50 entre 1000 valores de expressão média (em mediana) de genes randômicos do *cluster* apresentam valores maiores que os alvos em questão. A expressão média de fatores de transcrição foi de 0.3 a 0.7 *counts* normalizados, sendo o mais expresso o *STAT2* (presente em 52 % de *spots* do *cluster*). Os cinco fatores de transcricão com maior proporção dos alvos detectados foram: *STAT2* (4/4 detectados), *SREBF1* (19/23), *KLF4* (6/9) *FOXA1* (4/7), *CREB3* (2/4) (Figura 31). A Tabela suplementar 21 contém os o *output* da função com os 7 *regulons*.

O próximo passo foi predizer vias de sinalização intracelular entre os receptores e fatores de transcrição presentes em mais *spots* utilizando a informação de *counts* normalizados e de interações proteína-proteína descritas na literatura. Dessa forma, nós utilizamos a função *linker_intracellular_sign()* com os parâmetros *method = “counts”*, *my_receptor = c("ERBB2", "HLA-E", "LGALS3", "PIK3R1", "SDC1")* e *my_TF = c("STAT2", "SREBF1", "KLF4", "FOXA1", "CREB3")*. Na análise de forma conjunta, utilizando o parâmetro *mode = “spot_n”*, nós identificamos 75 caminhos, vias de sinalização, contendo 3 caminhos para cada combinação (25) de receptor e fator de transcrição. Entre esses, 15 caminhos apresentaram todos os genes do caminho com uma média maior que 0.25 de *counts* normalizados. Com base nos critérios estabelecidos, nós identificamos caminhos de sinalização proteína-proteína de 3 receptores (*ERBB2*, *HLA-E* e *PIK3R1*) e 4 fatores de transcrição (*CREB3*, *FOXA1*, *KLF4* e *SREBF1*). A Figura 32 mostra os caminhos identificados no entre o receptor HLA-E e os fatores de transcrição presentes em mais *spots*. A Tabela suplementar 22 contém o *output* da função com os 15 caminhos.

Figura 31. Fatores de transcricão com maior proporção dos alvos detectados.

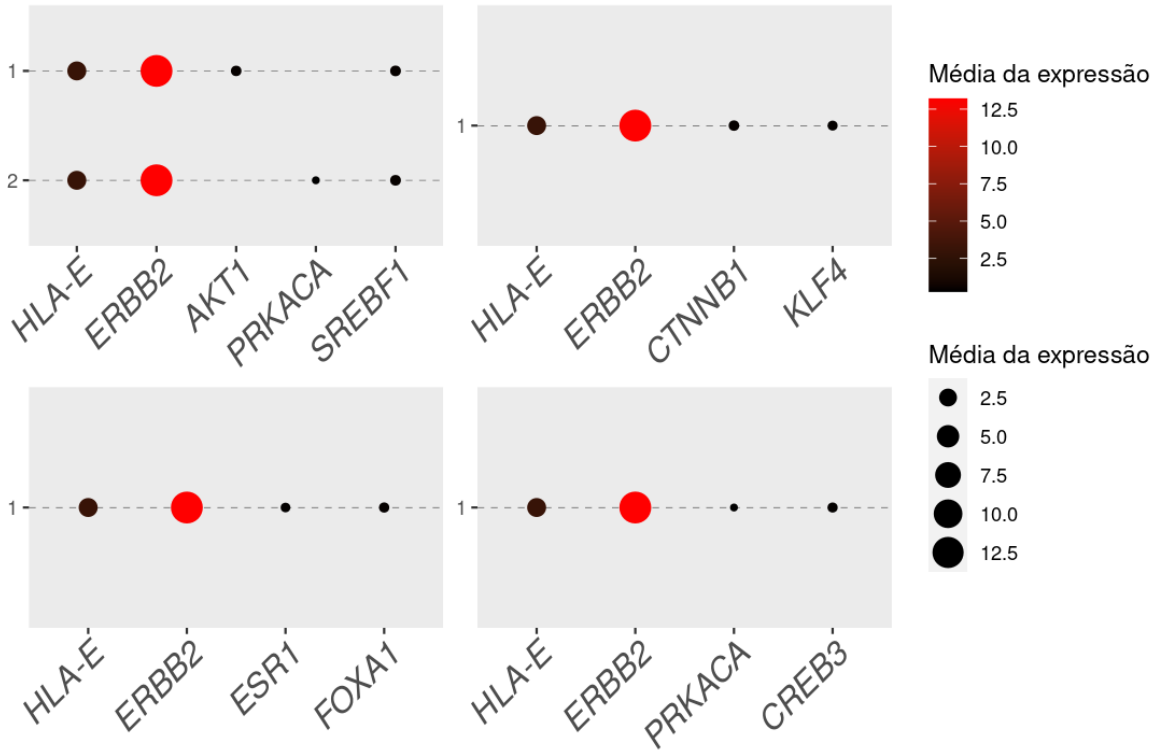


O eixo y define os fatores de transcrição referidos. Os valores na direita definem o número de alvos detectados/número de alvos descritos. A imagem se refere a um *scatterpie*, é um círculo branco alusivo ao número de alvos descritos que é preenchido por cor vermelha alusiva ao número de alvos detectados. O centro do *scatterpie* (o círculo preto central) também se refere ao número de alvos detectados no *cluster*.

5.2.3 Comunicação celular interna do *cluster* transcricional de "respostas de interferon tipo I" em análise de forma individual (*mode* = "*spot_y*")

O próximo passo foi comparar os resultados de análises de forma conjunta e de forma individual, a análise por *spots*. Sendo assim, utilizamos os 168 ligantes e 370 receptores do *cluster* transcricional de respostas de interferon tipo I detectados em mais de 25 % de *spots* obtidos em *define_intercellular_comm()*. Na análise de forma conjunta, utilizando a função *buzzer_intercellular_comm()* com o parâmetro

Figura 32. Vias de sinalização estimulatórias preditas entre HLA-E e os fatores de transcrição presentes em mais *spots* com base nos transcritos.



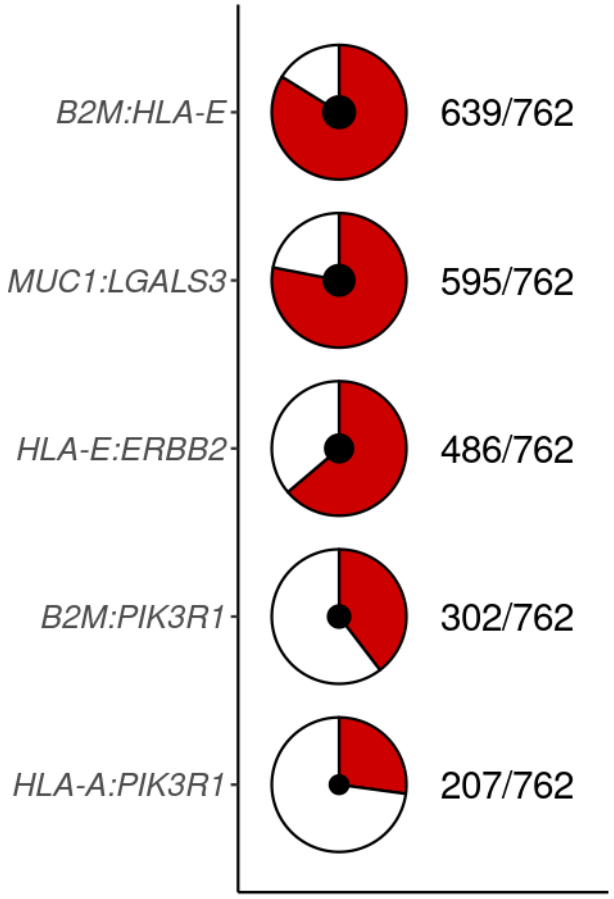
Vias de sinalização estimulatórias, ou vias de interação proteína-proteína estimulatórias, entre o receptor HLA-E e os fatores de transcrição NR1H3, TCF3, RELB, HIF1A e TCF4. A função *linker_intracellular_sign()* buscou 3 possíveis vias para cada combinação entre receptor e fator de transcrição. As cores (de preto à vermelho) e os tamanhos dos círculos indicam a média da expressão de interatores no *cluster*.

mode = *"spot_y"*, nós identificamos 4303 interações distribuídas entre 72 *spots*, contendo 84 pares distintos (60 em mais de 5 comunicações) entre 30 ligantes e 31 receptores que mostraram uma expressão do ligante maior que 3 e do receptor maior que 1 *counts* normalizados (ainda com a proporção de 3:1 entre ligante e receptor), um produto da expressão média maior que 3, um valor de *p* empírico menor que 0.05, ou seja, menos de 50 entre 1000 pares randômico de ligantes e receptores do spot em questão apresentam valores de produto da expressão maior que o par de ligante e receptor do *spot* e a interação entre o ligante e o receptor deve resultar em estimulação (em consenso na literatura). A expressão dos ligantes foi de 3 a 17 *counts* normalizados, sendo o ligante mais expresso o *MUC1* (presente

em todos os *spots* do *cluster* e com expressão de 17 em 6 *spots*) e dos receptores foi de 1 a 34 *counts* normalizados, sendo o mais expresso o *ERBB2* (também presente em todos os *spots* do *cluster* e com expressão de 34 *counts* normalizados em 2 *spots*). A Tabela suplementar 23 contém os o *output* da função com os 4303 pares de ligantes e receptores. Os cinco pares de ligantes e receptores presentes em mais interações entre *spots* foram *B2M*:*HLA-E* (presente em 83.8 % das interações), *MUC1*:*LGALS3* (78.0 %), *HLA-E*:*ERBB2* (63.8 %), *B2M*:*PIK3R1* (39.6 %), *HLA-A*:*PIK3R1* (27.1 %) (Figura 33). Os 6 ligantes detectados com mais pares de receptores detectados foram: *FN1* (com interações com 11 receptores), *COL1A1* (10), *COL3A1* (8), *COL1A2, THBS1* e *TNC* (7). Por sua vez, os 6 receptores detectados com mais pares de ligantes detectados foram: *PIK3R1* (com interações com 9 ligantes), *ITGA6, ITGAV, ITGB6, SDC1* e *SDC4* (6). Para cada interação entre *spots* nós comparamos o perfil de ligantes e receptores com as outras interações e, em média, as interações se asselham em 59.9 % (com mediana = 62.5 % e intervalo interquartil = 33.3 %). Ainda, nós analisamos a autocorrelação, ou correlação cruzada, espacial de pares de ligantes e receptores em comunicações de *spots* visando encontrar padrões espaciais na detecção da comunicação. Para isso, nós utilizamos o índice *I* de autocorrelação de Moran e detectamos 17 interações ligante-receptor espacialmente organizadas (valor de $p < 0.05$) com os índices de +0.02 a +0.15, ou seja, todos as interações significativas tendem ao *clustering* perfeito e não a dispersão perfeita. As 9 interações mais organizadas em coordenadas foram: *B2M:HLA-E* (Moran's I = 0.15), *HLA-E:ERBB2* (0.12), *AREG:ERBB2* (0.07), *B2M:PIK3R1* (0.07), *COL1A2:SDC1* (0.06), *PSAP:LRP1* (0.06), *TNFSF10:PIK3R1* (0.05), *HLA:A:PIK3R1* (0.05) e *CDH1:PIK3R1* (0.04) e estão mostradas na Figura 34. A Tabela suplementar 24 contém os valores de índices de Moran para as interações analisadas.

Então nós utilizamos a função *buzzer_intracellular_comm()* para identificar os *regulons* descritos na literatura e detectados no *cluster* transcricional de "respostas de interferon tipo I", porém com o parâmetro *mode* = *"spot_y"*, ou seja, na análise de forma individual. Nós identificamos 682 *regulons* distribuídos entre 74 *spots*, contendo 92 fatores de transcrição ou complexos de fatores de transcrição (34 em mais de 5 *spots*) e 902 alvos distintos, que mostraram mais de 10 % dos

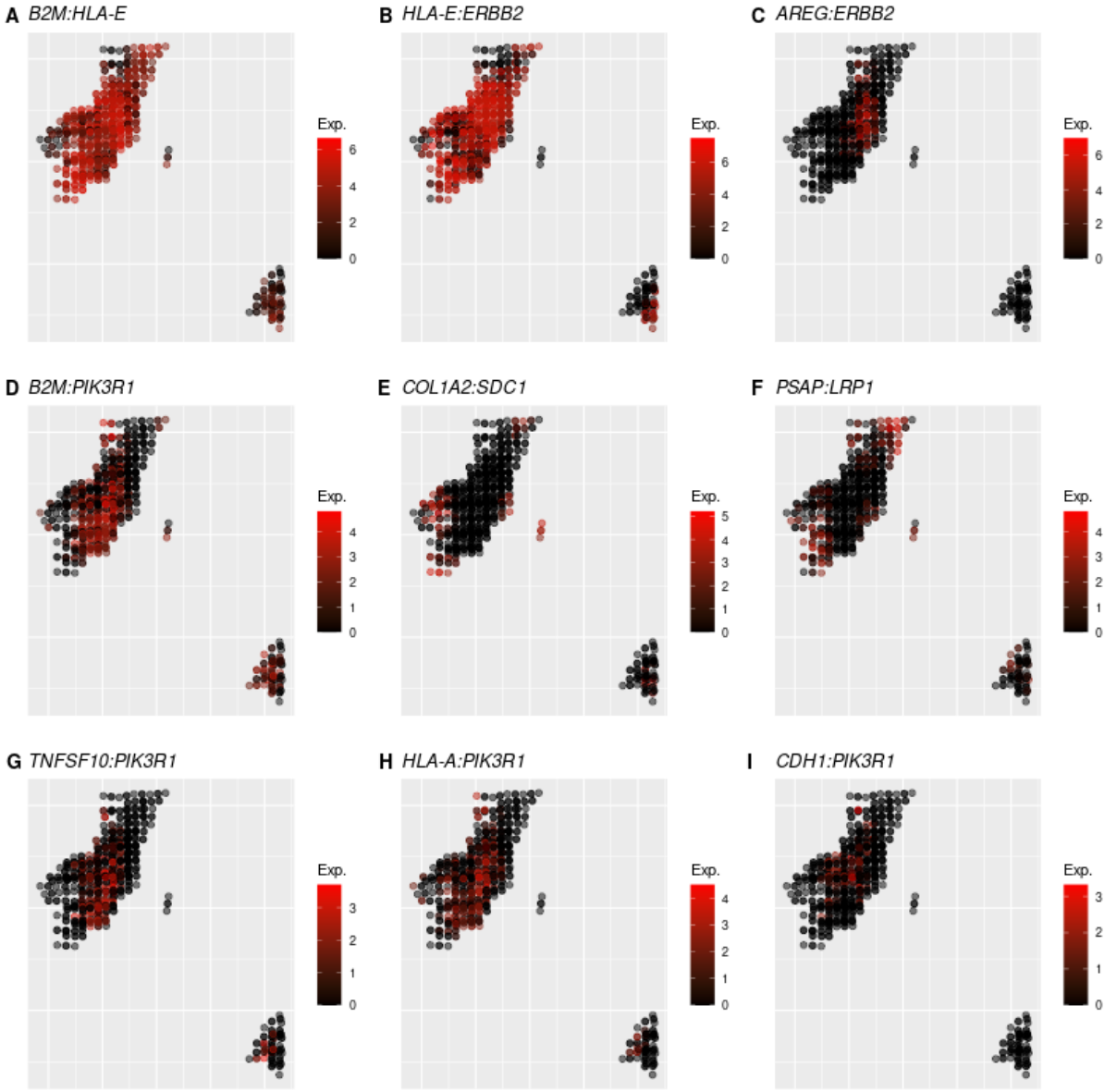
Figura 33. Pares de ligantes e receptores presentes em mais comunicações entre *spots*.



O eixo y define os pares de ligantes e receptores referidos. Os valores na direita definem o número de comunicações detectadas/número de comunicações viáveis no *cluster*. A imagem se refere a um *scatterpie*, é um círculo branco alusivo ao número de comunicações viáveis que é preenchido por cor vermelha alusiva ao número de comunicações detectadas. O centro do *scatterpie* (o círculo preto central) também se refere ao número de comunicações detectadas no *cluster*.

alvos detectados no *spot* e mostraram um valor de *p* empírico menor que 0.1, ou seja, menos de 100 entre 1000 valores de expressão de genes randômicos do *spot* apresentam valores maiores que os alvos em questão. A expressão de fatores de transcrição foi de 1 a 6 *counts* normalizados, sendo o mais expresso o *FOS* (presente em 52 % de *spots* do *cluster* e com expressão de 6 em 2 *spots*). A Tabela suplementar 25 contém os o *output* da função com os 682 *regulons*. Os cinco fatores de transcrição presentes em mais *spots* foram *STAT1* (presente em 58.1 % dos

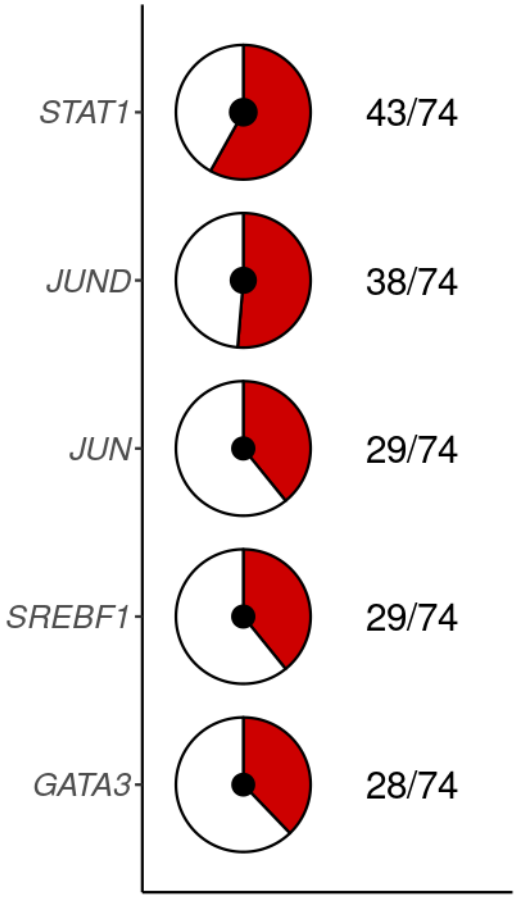
Figura 34. Distribuição de comunicações intercelulares mais autocorrelacionadas.



Distribuição e expressão de comunicações intercelulares expondo *spots* e espaços entre *spots.* As cores (de preto à vermelho) de círculos indicam o log2 produto da expressão de interatores na comunicação.

spots), *JUND* (51.3 %), *JUN* (39.1 %), *SREBF1* (39.1 %) e *GATA3* (37.8 %) (Figura 35). Para cada *spot* nós comparamos o perfil dos fatores de transcrição com os outros *spots* e, em média, as *spots* se asselham em 55.7 % (com mediana = 67.1 % e intervalo interquartil = 23.7 %). Utilizando o índice *I* de autocorrelação de Moran e detectamos 2 expressões de fatores de transcrição espacialmente organizadas

Figura 35. Fatores de transcrição operantes presentes em mais comunicações entre *spots*.

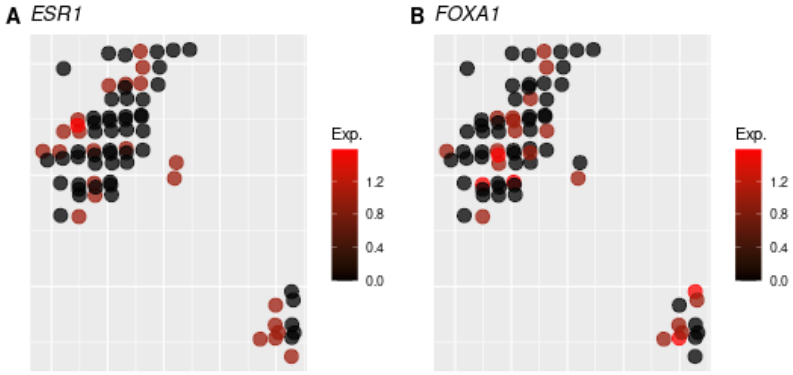


O eixo y define os fatores de transcrição referidos. Os valores na direita definem o número de ativações detectadas/número de *spots* no *cluster*. A imagem se refere a um *scatterpie*, é um círculo branco alusivo ao número de *spots* que é preenchido por cor vermelha alusiva ao Continuação da Figura 35: número de ativações detectadas. O centro do *scatterpie* (o círculo preto central) também se refere ao número de ativações detectadas no *cluster*.

(valor de $p < 0.05$), o ESR1 e o FOXA1 com os índices de +0.07 e +0.05, tendendo ao *clustering* e estão mostradas na Figura 36. A Tabela suplementar 26 contém os valores de índices de Moran para as expressões analisadas.

O próximo passo foi predizer vias de sinalização intracelular entre os receptores e fatores de transcrição presentes em mais *spots* utilizando a informação de *counts* normalizados e de interações proteína-proteína descritas na literatura, porém com o parâmetro *mode = "spot_y"*, ou seja, na análise de forma individual.

Figura 36. Distribuição de comunicações intracelulares mais autocorrelacionadas.

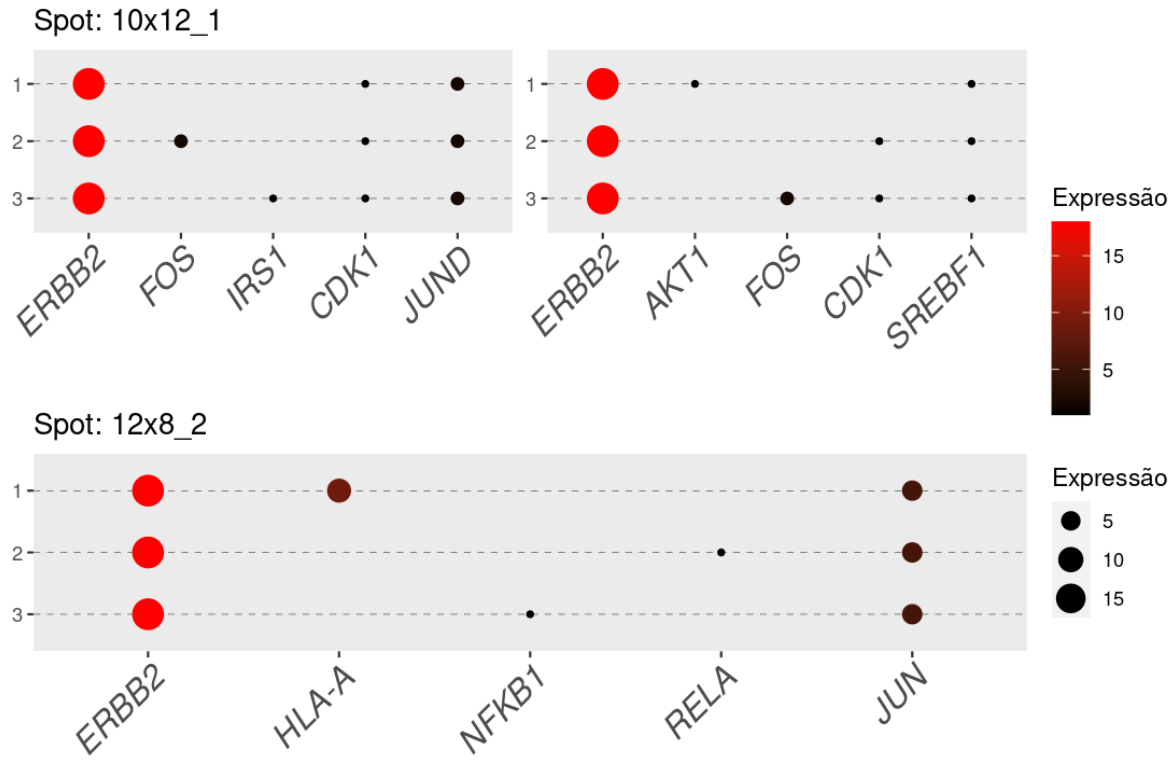


Distribuição e expressão de comunicações intracelulares expondo *spots.* As cores (de preto à vermelho) de círculos indicam o log2 da expressão de fatores de transcrição.

Dessa forma, nós utilizamos a função *linker_intracellular_sign()* com os parâmetros *method = "counts"*, *my_receptor = c("ERBB2", "HLA-E", "LGALS3", "PIK3R1", "SDC1")* e *my_TF = c(STAT1*, *JUND*, *JUN*, *SREBF1* e *GATA3)*. Nós identificamos 1348 caminhos, vias de sinalização em potencial, contendo caminhos entre todos os fatores de transcrição e os receptores distribuídos em 69 *spots* com todos os genes do caminho com uma expressão maior que 1 *count* normalizado. A Tabela suplementar 27 contém os o *output* da função com os 1348 caminhos. Os cinco pares entre os fatores de transcrição e os receptores presentes em mais *spots* (onde cada *spot* tem ao menos um caminho aceito) foram *ERBB2*:*STAT1* (presente em 58.1 % dos spots), *HLA-E*:*STAT1* (52.6 %), *ERBB2*:*JUN* (50.0 %), *ERBB2*:*GATA3* (43.2 %) e *ERBB2*:*JUND* (40.4 %). A Figura 37 mostra como as vias de sinalização intracelular podem ser diferentes entre *spots* de um mesmo *cluster* transcricional.

Mais uma vez nós utilizamos o conjunto de dados de scRNA-seq contendo as 19311 células de câncer de mama do subtipo molecular HER2-positivo para integrar com os dados de vias de sinalização obtidas. Dos 1348 caminhos, 470 retratam detecção de todos os genes em mais de 5 % de um tipo celular do *spot*. A Tabela suplementar 28 contém o *output* da função com os 470 caminhos. Foram detectados possíveis caminhos em células epiteliais (sadias) (422 caminhos, 89.5 %), células epiteliais (câncer) (355, 75.3 %), células endoteliais (65, 13,8 %) e plasmablastos (96, 8.2 %). Os caminhos supracitados em células epiteliais sadias e de câncer envolvem os receptores *ERBB2*, *HLA-E*, *LGALS3* e *PIK3R1* e os fatores

Figura 37. Exemplo de diferenças em vias de sinalização preditas de dois *spots.*

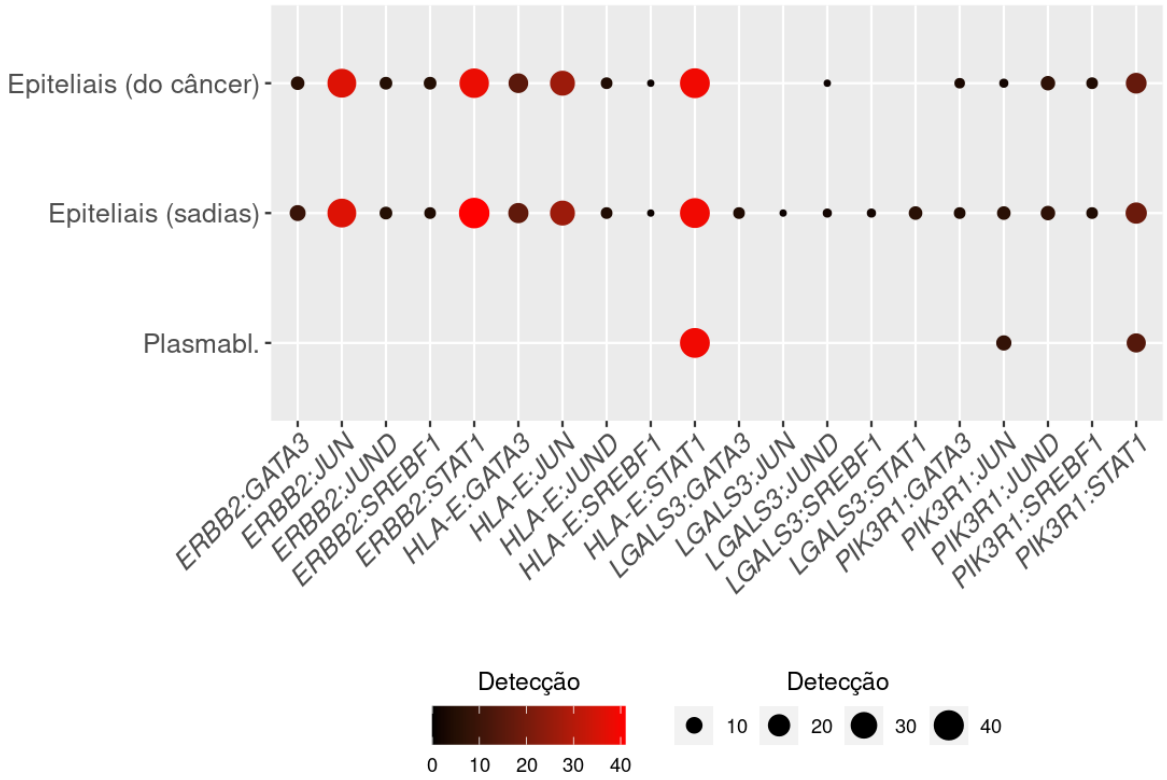


Vias de sinalização estimulatórias, ou vias de interação proteína-proteína estimulatórias, entre o receptor HLA-E e os fatores de transcrição STAT1 e ATF4 em dois *spots*. A função *linker_intracellular_sign()* buscou 3 possíveis vias para cada combinação entre receptor e Continuação da Figura 37: fator de transcrição. As cores (de preto à vermelho) e os tamanhos dos círculos indicam a expressão de interatores no *spot*.

de transcrição *GATA3*, *JUN*, *JUND*, *SREBF1* e *STAT1*, enquanto em plasmablastos envolvem os receptores *HLA-E* e *PIK3R1* e os fatores de transcrição *JUN* e *STAT1*. Os 3 caminhos mais detectados em células epiteliais sadias e de câncer: *ERBB2:STAT1* (presente em 39 spots em células epitelias de câncer e 41 em células epiteliais sadias), *HLA-E:STAT1* (39, 39) e *ERBB2:JUN* (36, 36) e em plasmablastos foram: *HLA-E:STAT1* (presente em 39 *spots*), *PIK3R1:STAT1* (14) e *PIK3R1:JUN* (8). A Figura 38 mostra a proporção dos *spots* com a predição de cada uma das vias em cada tipo celular contido no *spots.*

Assim, a análise baseada em *spots*, utilizando o *mode* = *"spot_y"*, detectou, após filtragens, comunicações intercelulares envolvendo 30 ligantes e 31 receptores (representando 17.8 e 8.4 % dos ligantes e receptores expressos em mais de 25 %

Figura 38. Elementos celulares viáveis de vias de sinalização preditas.



Proporção dos *spots* com a predição de cada uma das vias em linfócitos B, linfócito T e células epiteliais de câncer. As cores (de preto à vermelho) e os tamanhos dos círculos indicam o total de *spots* com a predição indicada.

de *spots* do *cluster*) contra 12 ligantes e 13 receptores detectados na análise baseada em *clusters,* utilizando o *mode* = *"spot_n"* (representando 7.2 e 3.5 % dos ligantes e receptores expressos em mais de 25 % de *spots* do *cluster*). Um total de 36 genes de ligantes e receptores foram identificados em *mode* = *"spot_y"* e não em *mode* = *"spot_n"* e os 5 termos de ontologias contendo mais genes dessa lista foram: organização da estrutura de encapsulamento externo, adesão da matriz celular, adesão do substrato celular, via de sinalização mediada por integrina e migração leucocitária. Ainda, há 24 termos envolvendo "migrar", 21 envolvendo "desenvolver" (epitélio da glândula mamária, por exemplo), 20 envolvendo "aderir", 16 envolvendo "crescer/ampliar" e outros. No total 323 ontologias tiveram enriquecimento e estão presentes na Tabela suplementar 29. A análise baseada em *spots* detectou 84 pares ligantes/receptores distintos (60 presentes em mais de 5 comunicações entre *spots*)

contra 17 da análise baseada em *clusters*. Ainda, a análise baseada em *spots* detectou, após filtragens, comunicações intracelulares envolvendo 92 fatores de transcrição ou complexos de fatores de transcrição contra 7 detectados na análise baseada em *clusters*. Um total de 85 genes de fatores de transcrição foram identificados em *mode* = *"spot_y"* e não em *mode* = *"spot_n"* e os alvos detectados em *spots* foram capazes de enriquecer 2012 termos de ontologias, presentes na Tabela suplementar 30, em que os 5 termos contendo mais genes dessa lista foram: resposta inflamatória, resposta ao oxigênio, regulação da morte celular, via de sinalização apoptótica e via de sinalização apoptótica intrínseca. Ainda, há 113 termos envolvendo "desenvolver", 107 envolvendo "diferenciar", 55 envolvendo "morfogênese", 55 envolvendo "proliferar" e outros. Obtivemos a distribuição de comunicações e sinalizações, tornando possível as análises de autocorrelação espacial, onde diversos interatores inter e intracelulares estavam organizados em áreas. Com as vias de interação proteína-proteína resolvidas em *spots* foi possível predizer com maior segurança os elementos celulares responsáveis de sinalização, sendo que nesse exemplo, os conjuntos receptor/possíveis proteínas estimuladas/fator de transcrição (com os cinco receptores e fatores de transcrição detectados em mais *spots*) foram preditos em células epiteliais do câncer, epiteliais sadias e plasmoblastos.

6. DISCUSSÃO

No presente estudo nós desenvolvemos um novo algoritmo de predição de comunicações celulares que utiliza os dados de coordenadas, os *spots*, como unidade de análise, o *SpotComm*. Esse algoritmo oferece funções de predição de possíveis comunicações intercelulares (geralmente entre ligantes e receptores), comunicações intracelulares (entre fatores de transcrição e alvos) e sinalizações intracelulares, ou seja, possíveis interações proteína-proteína que conectam o receptor e o fator de transcrição ativados. Isso com base nos transcritos. Ainda, o *SpotComm* é capaz de integrar dados de transcriptômica espacial e scRNA-seq gerando dados de proporção de células e proporção de co-ocorrência, por tipo celular, com detecção dos elementos de vias preditas de sinalização intracelular. Ainda, nós testamos o *SpotComm* com conjuntos de dados de transcriptômica espacial e scRNA-seq de câncer de mama em *clusters* transcricionais específicos.

O *SpotComm* surgiu como um algoritmo irmão à ferramenta *CellComm* (LUMMERTZ DA ROCHA et al., 2022), porém explorando os dados de coordenadas oferecidos pela transcriptômica espacial de modo distinto, possibilitando análises orientadas pelos *clusters*, semelhante ao que já foi feito, mas também orientadas pelos *spots,* de um modo que ao nosso conhecimento não foi ainda feito e que entendemos que pode gerar hipóteses e, à vista disso, conhecimento.

Para o *SpotComm* detectar comunicações e sinalizações há dois pontos cruciais, 1 - o banco de dados, tendo em vista que o algoritmo não é focado em predizer interações proteicas e regulações de expressão de modo *de novo* e 2 – a “profundidade” do sequenciamento. Assim, integrar ao *SpotComm* um banco de dados atualizado, bem curado e com acesso a metadados de interações e regulações é essencial. Nosso algoritmo integra em suas funções o Omnipath (utilizando o pacote R/Bioconductor *OmnipathR*), um banco de dados de conhecimento prévio de biologia de sistemas (TÜREI; KORCSMÁROS; SAEZ-RODRIGUEZ, 2016). Ele contém os dados de mais de 100 fontes de interações ligante-receptor, regulação transcricional, de RNAs, anotação gênica quanto a via/função/doença, estruturas de proteínas, complexos proteicos, interações proteicas, intervenções, etc e por sua utilidade ele foi capa da revista Nature

Methods na edição de dezembro de 2016. Dessa forma, o Omnipath se mostra um banco de dados magnífico com nossos objetivos. Recentemente, em março de 2021 o Omnipath foi alterado e renovado, contendo mais de 4000000 de dados e referências para mais de 20000 proteínas, uma estrutura moderna capaz de generalizar conceitos de ligante (*ligand*) e de receptor (*receptor*) para papéis de transmissor/emissor (*transmitter* ou *sources*) e receptor (*receiver* ou *targets*) e outros completando a nossa necessidade de dados recentes (TÜREI et al., 2021). Quanto a profundidade de sequenciamento, envolve a tecnologia de ST, a plataforma optada e gastos. Um sequenciamento mais profundo requer mais investimento, mas fornece/confere a identificação de transcritos raros, resultantes de baixa expressão ou de expressão em população rara no exemplar, aqui, no *spot*.

Em ambas as plataformas estudadas nesse trabalho os *spots* se formavam, se constituiam, de grupos ou populações de células. Os números de células capturadas em *spots* é baseado no tipo de tecido, tamanho das células e espessura da seção. A plataforma Visium da 10x Genomics tem *spots* com 55 µm de diâmetro e, segundo a 10x Genomics, geralmente os números de células variam entre 1-10 células (em seções de 10 µm de espessura, assim como nos dados utilizados) (https://kb.10xgenomics.com), enquanto a plataforma 1k arrays tem *spots* com 100 µm de diâmetro e as seções tinham 16 µm de espessura (ANDERSSON et al., 2021). Com base nisso, a predição de interações ligante e receptor de modo *de novo*, baseado p.e. em correlação, ou qualquer exploração de correlação, na verdade, é um desafio. Os *spots* possuem, além da variação transcricional dependente de nicho, de posição, a variação dependente de composição celular e tal foi visto nos limitados índices de correlação de Pearson de interações e autocorrelação espacial de Moran. O ideal agora é utilizar o *SpotComm* em plataformas de transcriptômica espacial de resolução celular como MERFISH (*Multiplexed Error-Robust Fluorescence in situ Hybridization* ou hibridização *in situ* de fluorescência robusta com erros multiplexados, em inglês) (WANG; MOFFITT; ZHUANG, 2018), uma tecnologia capaz de medir simultaneamente o número de cópias e a distribuição espacial de centenas a dezenas de milhares de espécies de RNA em células individuais. Os valores de distância entre *spots* (agora células) que o *SpotComm* utilizaria seria baixo, preferentemente alcançando os vizinhos mais

próximos. Assim, creio que seria possível obter índices de correlações e autocorrelações maiores e predizer interações novas.

Os algoritmos de análise de comunicação celular em transcriptômica espacial de Fawkner-Corbett (FAWKNER-CORBETT et al., 2021b) e colaboradores e o Giotto (DRIES et al., 2021) definem comunicações segundo a co-localização e co-expressão e, assim, encaram desafios parecidos com os de correlação/autocorrelação por motivos citados. Para contornar esse problema, Fawkner-Corbett e colaboradores usaram uma matriz com expressão “suavizada” (do inglês *smoothed*), ou com expressão média, entre o *spot* e os *spots* imediatos ao redor (FAWKNER-CORBETT et al., 2021b) e o Giotto usa um escore de comunicação calculado da expressão média ponderada de interatores em subconjuntos de células/*spots* próximos de ambas as populações em interação (DRIES et al., 2021). Desse jeito, aumentando o tamanho do *bulk*, do volume, de células estudadas no *spot* se confia que as distinções na localização e na composição tecidual sejam suavizadas/mitigadas. O *SpotComm* caminha na antemão desse pensamento, não valorizando um sobre o outro, mas presumindo que ambos possam funcionar juntos. O *SpotComm* é para tais algoritmos como o scRNA-seq é para o RNA-seq (*bulk*), ou seja, o *SpotComm* se baseia na heterogeneidade de *spots* e em possibilidades que *spots* com perfis transcricionais distintos possam ter quando em contato/interação para achar comunicações novas e quiçá importantes. O scRNA-seq mostra a heterogeneidade no conteúdo de RNAs entre células de um tipo celular da mesma forma que o *SpotComm* mostra a heterogeneidade no perfil de comunicação entre *spots* de um *cluster* transcricional e ambos apostam nessa riqueza.

Certos controles são basilares para isso, como um total de *reads* adequado em *spots* e um controle de *spot swapping*. Idealmente, identificadores moleculares únicos em um *spot* aferem a expressão específica do *spot*, contudo pode ocorrer o “sangramento” de *spots* próximos, um artefato que chamamos de *spot swapping*. Para isso, pesquisadores vem criando métodos visando refinar o poder e a precisão das análises *downstream* de transcriptômica espacial como o *SpotClean*, um modelo probabilístico na tirada do *spot swapping* (NI et al., 2022). Em nossas análises de comunicação externa aos *clusters* nós detectamos que mais de 97 % de interações

obtidas eram compartilhadas entre os *clusters* em contato e nós entendemos que tal similaridade em zonas de contato pode vir de *spot swapping*. Pensando nisso, nós queremos incluir em nosso *pipeline* de preparação de objetos para o *SpotComm* um algoritmo como o *SpotClean*.

É possível usar o *SpotComm* em análises com muitas plataformas de transcriptômica espacial como Visium, 1k arrays (ANDERSSON et al., 2021), Slide-seq (RODRIQUES et al., 2019), Slide-seqV2 (STICKELS et al., 2021), MERFISH (WANG; MOFFITT; ZHUANG, 2018) e outros. Isso porque o *Spotcomm* requer a matriz de expressão e as tabelas de metadados e coordenadas. Então o usuário tem pleno domínio da distância entre *spots* a ser analisada, focando na resolução da plataforma/tecnologia. Esses fatores fazem do *SpotComm* um algoritmo útil em muitos cenários, para muitos estudos e para vários objetivos e assim tardando o seu obsoletismo. Contudo, ainda que os dados de entrada sejam simples e genéricos, hoje eles tem que se encontrar em um tipo de objeto específico, o objeto *Seurat* (com a matrix de expressão em objeto_Seurat@assays$SCT@counts e tabela de metadados em objeto_Seurat@meta.data) com o objeto *Staffli* definido (com a tabela de coordenadas em objeto_Seurat@tools$Staffli). O *Seurat* é um pacote popular de R que faz pré-processamento e processamento de dados de scRNA-seq e transcriptômica espacial e o Staffli é um objeto que contém os metadados específicos do pacote *STutility*, como coordenadas de pixel, identificadores de amostra, plataforma, etc. Entendemos que seria ideal retirar a dependência do pacote *Seurat*, uma vez que há diversas outras fontes para preparação de dados de scRNA-seq e ST, algumas com ferramentas distintas p.e. as análises de tragetória obtidas pelo *CellRouter* (LUMMERTZ DA ROCHA et al., 2018), vindos de outras linguagens como o *Spacemake* em linguagem Python (SZTANKA-TOTH et al., 2022), que processa dados desde arquivos fastq, passando por controle de qualidade de sequências, trimagem, etc. Sendo assim, visamos adaptar o *SpotComm* para receber uma matriz de expressão e duas tabelas de metadados e coordenadas como *input* básico e outros como caminhos para imagens em funções específicas.

Os algoritmos de comunicação celular definem escores e funções de escores para as moléculas em interação. O *SpotComm* produz dados que podem

ser utilizados na predição de comunicação e sinalização, como escores binários e contínuos e ainda metadados. Nós utilizamos o escore de limiar de expressão (binário) em muitos momentos em nossos estudos de caso, como 1 - ao filtrar ligantes e receptores por detecção (detectados em > 25 % de *spots* do *cluster*) e por expressão (ligantes e receptores expressos em média > 0.75 e 0.25 *counts* normalizados no *mode* = *"spot_n"* e expressos > 3 e 1 *counts* normalizados no *mode* = *"spot_y"*), 2 - filtrar *regulons* por detecção (fatores de transcrição detectados em > 25% dos *spots*) e por expressão (fatores de transcrição expressos em média > 0.25 *counts* normalizados no *mode = "spot_n"* e expressos > 1 *count* normalizados no *mode* = *"spot_y",* > 10 % dos alvos detectados no *spot*, ou seja, *count* > 1), 3 - filtrar caminhos por expressão (todos os genes do caminho expressos em média > 0.25 *counts* normalizados no *mode = "spot_n"* e expressos > 1 *count* normalizados no *mode = "spot_y*) e outros. Os "*counts* normalizados" usados se referem os *counts* corrigidos derivados de transformação reversa de resíduos de Pearson feita por *sctransform()* (HAFEMEISTER; SATIJA, 2019). O *count* então é o produto da normalização e estabilização de variância de dados de scRNA-seq usando regressão binomial negativa regularizada onde a profundidade do sequenciamento é usada como uma covariável em um modelo linear generalizado, removendo a influência das características técnicas mas preservando a heterogeneidade biológica (HAFEMEISTER; SATIJA, 2019). Ainda que a função tenha sido criada para scRNA-seq, a dispersão da relação média-variância e da taxa de detecção média de genes em transcriptômica espacial mostram superdispersão em comparação com o que seria esperado em um modelo de Poisson. Por essas propriedades serem compartilhadas entre os dados de scRNA-seq e transcriptômica espacial e scRNAseq o método é adequado para os dados de transcriptômica espacial (mostrado em ludvigla.github.io/STUtility_web_site/ Normalization.html). De fato, vários estudos utilizam a função, como os artigos originais de estudos de caso 1 e 2 (ANDERSSON et al., 2021; WU et al., 2021b). Usar valores diferentes de limiar de ligantes e receptores, como optado aqui na relação 3 ligantes: 1 receptor, tem base nas referências como Hemant Suryawanshi e colaboradores em 2018 (SURYAWANSHI et al., 2018), que lidaram com a relação 5.0:1.5 transcritos por milhão para ligantes e para receptores ou no algoritmo *CCCExplorer*, onde eram

considerados ligantes *up*-regulados em contrastes estudados e receptores somente expressos/detectados (CHOI et al., 2015). Os escores contínuos de produto de expressão e correlação de expressão também são usados no *SpotComm*. Ambos estão contifdos no *output* da função *buzzer_intercellular_comm()* para cada par de interatores. O produto de expressão está sempre presente, enquanto a correlação de expressão está presente quando o parâmetro *cor* = *TRUE* (porque a função *cor.test()* com o parâmetro *method* = *"pearson"* é utilizada). Nós entendemos que os escores possuem alguns viéses, no caso do escore de limiar de expressão se pode falsos positivos, onde um gene expresso não condiz com a proteína traduzida ou funcional, e falsos negativos, onde um gene de baixa expressão é capaz de codificar uma proteína com alta força de interação e de propagação de sinal. Então, é importante 1 - incorporar a chance de haver dados pareados de proteômica ao *SpotComm*, que demonstrem a detecção proteica dos interatores como em estudos estão criando (BACCIN et al., 2020; JI et al., 2020) e confirmar a presença da proteína e reforçar a comunicação e 2 – incorporar escores para termos de ontologias em *spots* que demonstrem um processo/função decorrente da ativação de um receptor/via e isso reforce a comunicação como no enriquecimento *spot-wise* da via de sinalização de interferon tipo I (ANDERSSON et al., 2021). O escore de correlação de expressão contém problemas já citados e no produto de expressão genes em interação com expressões muito diferentes pode resultar na detecção de interações dominadas por um dos interatores. Isso é possível de se observar em nossos resultados de produto de expressão, onde a alta expressão de genes como *HLA* (*HLA-A*, *HLA-E*), *B2M* e *ERBB2* por serem muito expressos dominavam as nossas conclusões gerais. Para isso, pensamos em incorporar um cálculo de produto que considere os dados de expressão escalonados, como um *z*-score. Assim, um gene expresso de modo homogêneo pelo tecido, ainda que muito expresso em counts normalizados, não teria tanto peso (tendendo a 0). Isso, somado ao produto de expressão básico nos dariam noções significativas de cada interator/interação.

O nosso algoritmo faz uso de redes de comunicação intercelular, de interação proteína-proteína para sinalização intracelular e de regulatórias na comunicação intracelular. Conforme citado, diversos algoritmos utilizam as

estruturas de redes em seus conceitos na predição de comunicação celular (CHOI et al., 2015; WANG et al., 2019b; BROWAEYS; SAELENS; SAEYS, 2020; LUMMERTZ DA ROCHA et al., 2022). Tanto nas redes de comunicação intercelular quanto nas redes de sinalização intracelular nós utilizamos somente interações classificadas como *consensus stimulation* (consenso de estimulação), ou seja, somente interações estimulatórias, de ativação e de propagação de sinal. Ainda não testado, nós poderíamos ter adotado somente interações de *consensus inhibition* (consenso de inibição) e analisado os contextos em nova perspectiva, com uma alteração simples. Analisar redes com o metadado de estimulação/inibição nos faz pensar que não há ainda modelos, entre esses algoritmos, capazes de obter as sinalizações de *crosstalk*, onde os sinais desencadeados por um receptor podem interferir nos sinais desencadeados por outro. Pensando nisso, nós entendemos que poderíamos utilizar o *SpotComm* em um conjunto de transcriptômica espacial com maior resolução, como o Slide-seq (RODRIQUES et al., 2019), Slide-seqV2 (STICKELS et al., 2021) ou MERFISH (WANG; MOFFITT; ZHUANG, 2018), e predizer a expressão gênica ou ativação de *regulons*, com base nos perfis de ligantes em comunicações próximas e de receptores de cada célula utilizando redes neurais profundas. Nós utilizamos alguns algoritmos nas análises de redes, o *shortest path* de Dijkstra (DIJKSTRA, 1959), que utiliza os pesos das interações nas definições de caminho de menor somatório entre dois vértices, o *k-shortest paths* que usa o conceito de Dijkstra (YEN, 1971) para designar *k* caminhos de menor somatório entre dois vértices em que os pesos são os coeficientes de correlação de Pearson ou os produtos de expressão entre os pares em interação e o algoritmo de *Random Walk* de propagação de redes chamado no cálculo da probabilidade de que vias de sinalização intracelular “randômicas” possuam as somas de pesos menor que as das vias em questão. Com esses algoritmos, na função *linker_intracellular_comm()* e com a integração com os dados de scRNA-seq na função *integr_intracellular_comm()* nosso objetivo era detectar a *possibilidade* de, com base nos transcritos de *spots* ou de células, ter possíveis proteínas que devido a interações de estimulações estimulassem os fatores de transcrição definidos como ativos. Entendemos, outra vez, que RNA transcrito não reflete proteína traduzida em todos os casos, que as vias de sinalização não são sempre os caminhos mais curtos

e que controles de *crosstalk* existem, mas ainda assim os algoritmos e as funções nos dão um panorama da probabilidade útil. Entre os algoritmos presentes na literatura o *CellComm* (LUMMERTZ DA ROCHA et al., 2022) é o mais próximo em suas análises de redes substituindo o *shortest* e o *k-shortest paths* por um algoritmo de otimização de fluxo de redes (*flow networks*) generalizando o *framework one-source/one-sink* para resolver um problema de *multiple-source/multiple-sink* para detectar caminhos conectando um conjunto de nós a outro conjunto de nós em uma rede proteína-proteína ponderada, como descrito (DA ROCHA et al., 2016; LUMMERTZ DA ROCHA et al., 2018). Agora, nós queremos utilizar algoritmos de teoria de grafos como as medidas de centralidade e conectividade como executado no *NicheNet* (BROWAEYS; SAELENS; SAEYS, 2020) para quantificar a importância na rede de cada nó, de cada elemento, e desta forma, caracterizar a relevância dos integrantes das vias de sinalização e, ainda, como no *NicheNet* e com o auxílio do *Omnipath* (TÜREI; KORCSMÁROS; SAEZ-RODRIGUEZ, 2016; TÜREI et al., 2021), podemos ponderar as redes de comunicações e sinalizações utilizando otimização de parâmetros, onde fontes mais informativas (o quão fortemente o conhecimento prévio ampara a interação/regulação/sinalização) contribuem mais ao modelo (BROWAEYS; SAELENS; SAEYS, 2020).

Ainda sobre o *SpotComm*, o nosso algoritmo utiliza em diversos momentos as concepções baseadas em permutações para auxiliar a predizer comunicações. No *SpotComm* chamamos de *p*-empírico, ou distribuição nula, e é presente 1 - na função *buzzer_intercellular_comm()*, onde nós comparamos os valores do produto da expressão média (ou produto da expressão) com 1000 valores de produto da expressão média (ou produto da expressão) de par randômico de genes *source* (da mesma categoria de *source*) e *target* (da mesma categoria de *target*) e caso o parâmetro *random_cor* seja definido como “TRUE” (ocorra), a função calculará a correlação de Pearson entre os pares de *source*/*target* entre 1000 pares de *spots* randômicos de *clusters*.; 2 - na função *buzzer_intracellular_ comm()*, onde nós comparamos a expressão média dos alvos (cada alvo, elemento) com 1000 valores de expressão média de genes randômicos do *cluster* do *cluster* contendo zeros e não contendo zeros; 3 - na função *linker_intracellular_comm()*, onde nós comparamos as somas de pesos (coluna *weight*) da via de sinalização intracelular

com 1000 valores de somas de pesos derivadas de vias com um ponto de partida randômico e “número de passos”, número de interações proteína-proteína, igual ao *n_vertices* (ou seja, utilizamos um recurso de propagação de redes chamado *Random Walk*). Assim, ainda que com a distribuição nula, nós temos cautela na criação de distribuições de unidades análogas ao comparado, p.e. utilizando genes da mesma categoria de *source/target* na comparação em *buzzer_intercellular_comm()*, visto que estaríamos comparando genes com traços (e, talvez, perfil de expressão de ligantes/receptores. Contrapor a expressão de p.e. genes codificantes de receptores, que costumam ter baixa expressão, com outros não receptores, poderia interferir nos resultados. Na função *buzzer_intracellular_comm()* o cuidado incluiu o uso das comparações com medianas e não médias, porque muitos alvos possuem expressão = 0, e sendo assim, qualquer detecção de expressão, mesmo que 1 *spot* somente, traria a média de expressão para um valor > 0 e a comparamos a expressão dos alvos com valores de expressão mediana de genes randômicos do *cluster* contendo zeros e não contendo zeros, em razão da detecção de zeros. Assim, comparamos a detecção/expressão dos alvos (usual, contando zeros), mas também a expressão dos alvos detectados com genes detectados e podemos contrapor as respostas. Na função *linker_intracellular_comm()* o cuidado incluiu o uso dos caminhos de similar tamanhos na rede de interação proteina-proteina do cluster ou do spot e a comparação da improbilidade de ocorrência/sucesso. Nós entendemos que algo deve ser executado para estipular limiares específicos de *p*-valor de genes, uma vez que genes podem não ser muito detectados/expressos para determinar atividade de proteínas e, portanto, determinar nossa predição. Um controle positivo do *SpotComm* na análise de permutações é a capacidade de manipular complexos proteicos homomultiméricos e heteromultiméricos em suas funções e, assim, em casos de complexos, executar permutações de todos os elementos, ou em casos de interações de complexos, de todas as combinações de elementos. Dados de complexos proteicos são utilizados por alguns algoritmos e de fato, auxiliam a predição (VENTO-TORMO et al., 2018; BACCIN et al., 2020; JIN et al., 2021; NOËL et al., 2021).

Nosso objetivo utilizando os estudos de caso foi mostrar que o *SpotComm* distingue moléculas de interação inter e intracelular e de regulação gênica conhecidas em câncer de mama triplo-negativo e HER2-positivo que são modelos bem estudados para termos base, mas ainda assim, modelos que são desafios na prática clínica (HARBECK et al., 2019) e que podem se beneficiar de achados novos que o *SpotComm* pode trazer.

As análises de comunicação utilizaram, em ambos os *clusters* explorados, só os ligantes e os receptores detectados em mais de 25 % dos *spots*. Assim, as análises de forma conjunta (*mode = "spot_n"*) ou individual (*mode = "spot_y"*) detectaram comunicações entre moléculas bem expressas. Nós optamos por tal porque queríamos passar ao leitor os resultados que mostrassem que a visão geral do *cluster* condizia com o esperado. Caso contrário, sem tal filtragem, o usuário poderá obter mais comunicações, menos "globais" e mais específicas a alguma região e, dessa forma, criar novas hipóteses quiçá importantes para prognóstico, diagnóstico e tratamento. Ainda, com a deconvolução em subtipos mais específicos o usuário poderá ligar essas comunicações raras a tipos celulares raros ou a interações raras no *cluster* tal como entre as células T da zona externa da estrutura linfoide terciária e um vaso.

As estruturas linfoides terciárias são agregações ectópicas de células do sistema imunológico que se desenvolvem em tecidos periféricos em resposta a diversos cenários inflamatórios crônicos, como infecções e câncer. No microambiente tumoral, as estruturas das estruturas linfoides terciárias se consistem em áreas prevalentemente populadas com células B e T e indicam que essas estruturas podem ser o local do início e da manutenção das respostas imunes humorais e celulares contra o câncer (ANDERSSON et al., 2021). Numerosos estudos avaliaram a expressão de estruturas linfoides terciárias em câncer e sua associação com prognósticos de pacientes com câncer (DE CHAISEMARTIN et al., 2011; GOC et al., 2014; HENNEQUIN et al., 2016).

Na região do tecido analisada do paciente "B" um *cluster*, o *cluster* 6 foi classificado como o *cluster* transcricional de "estruturas linfoides terciárias". Esse *cluster* mostrou a maior concentração de células B e T, com *spots* contendo de 0 a 62 % de células B e 0 a 56 % de células T. Ainda, o *cluster* possuia *spots* com 49 %

a mais de células B que T e 51 % a mais de células T, indicando regiões ricas em um ou outro tipo celular. Essa distribuição é bem comum nas estruturas linfoides terciárias, uma vez que estruturas maduras possuem zonas de células T externas e zonas de células B internas, o centro germinal (KINKER et al., 2021). Os genes marcadores do *cluster* estavam associados a termos de ativação da resposta imune, especialmente adaptativa, com a ativação e regulação de linfócitos como células T, sinalização mediado por receptor de antígeno e outros. Então, as distribuições celulares, os genes marcadores e os processos em que os genes estão associados nos garantiram a confiança de estar de olhos em um *cluster* de estruturas linfoides ectópicas, possivelmente terciárias. Além de células B e T, os *spots* do *cluster* 6 também mostraram células mielóides, epiteliais de câncer, fibroblásticas associadas ao câncer, na maioria, fenômeno já argumentado (HELMINK et al., 2020). Em média as presenças de células B em outros clusters são de 2 %, tendo *spots* de 0 a 38 % dessas enquanto de células T em média as presenças são de 3 %, tendo *spots* de 0 a 40 % dessas. Assim, há presença de tais linfócitos em mais *spots*, mas sem estrutura. Uma próxima análise irá se focar em classificações mais exploradas, mais especificadas de tipos celulares, utilizando p.e. a camada de subconjunto de Wu e colaboradores (WU et al., 2021b), para podermos caracterizar a localização de células T CD4+, CD8+, e B para detectar centro germinal e outros que determinam estruturas linfoides terciárias maduras (KINKER et al., 2021).

O ligante mais expresso em ambas as análises (não focada e focada nos *spots*) foi o *B2M*, enquanto o receptor mais expresso na análise não focada nos *spots* foi o *CD74* e na análise focada nos *spots* foi o *HLA-E*. O gene *B2M* (beta-2-microglobulina) codifica uma proteína sérica encontrada em associação com a cadeia pesada de classe I do complexo principal de histocompatibilidade (MHC) na superfície de quase todas as células nucleadas e, dessa forma, é vinculada a exposição de antígenos ao sistema imunológico (O'LEARY et al., 2016). É relativo a ontologias como sistema imune e citotoxicidade mediada a células T, é um gene marcador do *cluster* e é expresso 1,61 vezes mais que em outros *clusters* (LIBERZON et al., 2015). A proteína codificada pelo gene *CD74* (molécula CD74), por sua vez, associa-se ao MHC de classe II e regula a apresentação de antígenos para a resposta imune. Também serve como receptor de superfície celular para o

fator inibidor de migração de macrófagos de citocinas (*MIF*) que, quando ligado à proteína codificada, inicia vias de sobrevivência e de proliferação celular (O'LEARY et al., 2016). É relativo a ontologias como regulação da via de sinalização mediada por citocinas e ativação de células T envolvidas na resposta imunológica, é um gene marcador do *cluster* e é expresso 2,44 vezes mais que em outros *clusters* (LIBERZON et al., 2015). O *HLA-E* (complexo principal de histocompatibilidade, classe I, E) é um parálogo de cadeia pesada HLA de classe I. Esta molécula de classe I é um heterodímero constituído por uma cadeia pesada e uma cadeia leve (beta-2-microglobulina). A cadeia pesada está ancorada na membrana. A proteína HLA-E liga-se a um subconjunto restrito de peptídeos derivados dos peptídeos líderes de outras moléculas de classe I (O'LEARY et al., 2016). É relativo a ontologias como regulação da citotoxicidade celular dependente de anticorpos e regulação da citotoxicidade mediada por células T, é um gene marcador do *cluster* e é expresso 1,74 vezes mais que em outros *clusters*. De nota, no *SpotComm* comunicações envolvendo *B2M*, *CD74* e *HLA-E* estão entre as comunicações cujos pares possuem os maiores valores de escores de comunicação na análise de forma conjunta (*B2M:HLA-E*, *MIF:CD74* e outras); comunicações cujos pares estavam presentes em mais comunicações entre *spots* na análise de forma individual (*B2M:HLA-E*, *MIF:CD74*, ambos em 100 % das comunicações e outras); comunicações cujos pares estão bem organizados em coordenadas (*B2M:HLA-E*, Moran's *I* de 0.24) e outros. A interação *B2M:HLA-E* se liga a autopeptídeos nonâmeros derivados da sequência sinal de moléculas clássicas de MHC classe Ia (peptídeos VL9) (LLANO et al., 1998; SULLIVAN et al., 2007; HOARE et al., 2008) . O complexo heterotrimérico *B2M:HLA-E* ligado a peptídeo funciona principalmente como um ligante para o receptor inibitório de células natural killer (NK) *KLRD1-KLRC1*, permitindo que as células NK monitorem a expressão de outras moléculas de MHC classe I em células saudáveis e se tolerem (BRAUD et al., 1998; LLANO et al., 1998; COUPEL et al., 2007). Por sua vez, a interação *MIF:CD74* como dito inicia vias de sobrevivência e proliferação celular e, ainda, a CD74 interage com CD44 e este é um complexo essencial para a cascata de sinalização induzida por MIF que resulta na sobrevivência de células B (GORE et al., 2008), sendo que o *CD74* é expresso em 93.3 % de *spots* do *cluster*. Ainda, a interação MIF:CD74 participa na

imunopatogênese do lúpus eritematoso sistêmico sendo que a expressão elevada de *MIF* e *CD74* se correlaciona com piora das inflamações (FARR; GHOSH; MOONAH, 2020). A detecção de tais interatores, as ontologias associadas e os dados da literatura associados nos garantiram padrões e dinâmicas imunes de estruturas linfoides ectópicas e o *SpotComm* nos permitiu os detectar em regiões de ocorrência.

Entre os genes de fatores de transcrição, o *STAT1* (transdutor de sinal e ativador de transcrição 1) se destacou em análises de forma individual (*mode* = *"spot_y"*) por estar entre os fatores mais expressos, mais presentes e com mais *links*, caminhos, em *spots*, com mais *links* em células B, T e epiteliais. A proteína codificada por *STAT1* é um membro da família de proteínas STAT. Em resposta a citocinas e fatores de crescimento, os membros da família STAT são fosforilados pelas quinases associadas ao receptor, e formam homo ou heterodímeros que se translocam ao núcleo da célula onde atuam como ativadores de transcrição (O'LEARY et al., 2016). A STAT1 pode ser ativada por diversos ligantes como interferon-alfa, interferon-gama, EGF, PDGF e IL6. Esta proteína medeia a expressão de genes em que se acredita serem importantes para a viabilidade celular em resposta a diferentes estímulos celulares e patógenos. Como parte da cascata de sinalização chave para citocinas e fatores de crescimento, a JAK-STAT, desempenha um papel central no sistema imunológico inato e adaptativo. As citocinas controlam a estabilidade, o comprometimento e a maturação de células T citotóxicas e auxiliares, partes do sistema imunológico adaptativo que medeiam a imunidade a patógenos e estão ligadas a doenças inflamatórias (GOSWAMI; KAPLAN, 2017). Em células B e T nós predizemos caminhos, vias de interação proteína-proteína (entre os receptores e fatores de transcrição presentes em mais *spots*), entre *HLA-E:STAT1* (presente em 207 *spots* de 222), *CXCR4:STAT1* (144) e *CD74:STAT1* (141). Entre esses, a indução de HLA de classe I induzida por STAT1 aumenta a imunogenicidade e a resposta clínica à terapia com anticorpos monoclonais anti-EGFR com cetuximab em pacientes com câncer de cabeça e pescoço (SRIVASTAVA et al., 2015), o *STAT1* é contido na via de "eventos de sinalização mediados por *CXCR4*" do *PubChem* e a superexpressão de *STAT1* e *CD74* é co-dependente e ligada ao aumento da invasão e metástase linfonodal em

câncer de mama triplo negativo (GREENWOOD et al., 2012). Em células epiteliais nós predizemos caminhos entre *HLA-E:STAT1* (207) e *ERBB2:STAT1* (199). A interação entre HLA-E e STAT1 direta foi previamente descrita por ensaios de ligação por proximidade (CHEN et al., 2014) e arrays de proteínas (JONES et al., 2006). Bem como nas interações intercelulares, as comunicações e sinalizações intracelulares indicam os resultados expectados no cenário destacado.

A capacidade de expor as interações em estruturas espaciais, por meio das coordenadas, é uma das ricas evidências do *SpotComm*. Utilizando os espaços de comunicações intermédios entre *spots* e os produtos de comunicação como seus escores nós definimos interações entre ligantes e receptores autocorrelacionados espacialmente. Neste detectamos autocorrelações muito distintas, algumas dispersas na estrutura, algumas focadas em pontos da estrutura. Nós percebemos que a comunicaçãe *B2M:HLA-E* é mais intensa em pontos onde a comunicação *B2M:GZMB* é menos intensa, predizendo regiões onde o ligante pode, ao acaso, interagir com diferentes receptores. O gene *GZMB* (granzima B) codifica um membro da subfamília de proteínas granzima, parte da família peptidase S1 de serina proteases. A pré-proproteína codificada é secretada por células *natural killer* e linfócitos T citotóxicos e processada proteoliticamente para gerar a protease ativa, que induz apoptose da célula alvo. Essa proteína também processa citocinas e degrada proteínas da matriz extracelular, e esses papéis estão implicados na inflamação crônica e na cicatrização de feridas (O'LEARY et al., 2016). Em nosso trabalho os escores da interação *B2M:GZMB* foram maiores em zonas de maior proporção células T/células B.

As autocorrelações de interações envolvendo a quimiocina *CXCL13* (motivo C-X-C quimiocina ligante 13) se distinguiram por sua concentração em sítios específicos do tecido. Nós detectamos interações *CXCL13:CXCR3*, *CXCL13:CXCR4* e *CXCL13:CCR7* com Moran's *I* significativo (como também *CXCL13:CXCR5*, mas nós detectamos o *CXCR5* em apenas 21.8 % de *spots* do *cluster* e, assim, foi filtrado por *filter_define_intercellular_com()*). A CXCL13 é descrita como molécula quimiotática para linfócitos B e linfócitos T auxiliares foliculares, mas não para outros linfócitos T, monócitos ou neutrófilos. Autores a retrataram na formação do centro germinativo, no desenvolvimento de linfonodos, na regulação da imunidade humoral

e na regulação da proliferação celular (COPPOLA et al., 2011; DE CHAISEMARTIN et al., 2011; BINDEA et al., 2013; GU-TRANTIEN et al., 2013; HENNEQUIN et al., 2016). Por sua vez os genes *CXCR3*, *CXCR4*, *CXCR5* e *CCR7* (motivo C-X-C receptor de quimiocinas 3, 4, 5 e motivo C-C receptor de quimiocinas 7) podem estar envolvidos na estruturação e função da estrutura linfoide terciária. A ligação de quimiocinas a proteína CXCR3 induz respostas celulares que estão envolvidas no tráfego de leucócitos, principalmente ativação de integrinas, alterações citoesqueléticas e migração quimiotática (GOC et al., 2014). A proteína CXCR5 é encontrada em células B maduras, se liga ao quimioatraente de linfócitos B e é envolvido na migração dessas. Ainda, o *CXCR5* é fundamental para a neogênese do tecido linfoide associado à mucosa ectópica na inflamação crônica induzida por *Helicobacter pylori* (WINTER et al., 2010). O par *CXCL13*/*CXCR5* é induzido após a sinalização de LT-βR durante a gênese linfóide (DEJARDIN et al., 2002). Ambos são superexpressos em estruturas linfoides terciarias de pacientes com câncer de pele (MESSINA et al., 2012), colorretal (COPPOLA et al., 2011) e de pulmão (DE CHAISEMARTIN et al., 2011).

Nossa segunda análise focou no cluster transcricional de “resposta a interferon tipo I”. Os interferons são citocinas com longa história de envolvimento no desenvolvimento e tratamento do câncer. Existem três tipos principais de interferons, distinguidos por sua identidade de sequências, natureza, distribuição de receptores, estímulo indutor e célula de origem (SCHRODER et al., 2004; DONNELLY; KOTENKO, 2010; HERTZOG; WILLIAMS, 2013). Os genes humanos de interferons tipo I estão agrupados no braço curto do cromossomo 9 e codificam 17 proteínas distintas que se ligam ao seu receptor composto por subunidades IFNAR1 e IFNAR2 (receptor IFNα/β 1 e 2). A expressão de interferons tipo I é induzida por meio de vias de receptores de reconhecimento de padrões moleculares associados ao dano. Os interferons regulam as expressões de diversos genes os quais impactam diretamente o crescimento, proliferação, diferenciação, sobrevivência, migração e outras funções especializadas de cânceres (PARKER; RAUTELA; HERTZOG, 2016). Em linhagens celulares de câncer de mama humano, o tratamento utilizando interferons tipo I teve um efeito antiproliferativo direto que foi atribuído ao expansão de todas as fases do ciclo celular induzida por interferons (BALKWILL; WATLING;

TAYLOR-PAPADIMITRIOU, 1978). Ainda, o tratamento com interferon α em linhagens celulares de câncer de próstata regula positivamente os inibidores endógenos de quinases dependentes de ciclina, como p21, e paralisa a progressão do ciclo celular (HOBEIKA; SUBRAMANIAM; JOHNSON, 1997). Ainda, os interferons podem induzir apoptose em linhagens celulares de vários cânceres via regulação das duas respostas apoptóticas básicas: a via extrínseca, ou mediada pelo receptor de morte, e a via intrínseca, ou mitocondrial (THYRELL et al., 2002; CHOI et al., 2003; FULDA; DEBATIN, 2006). Os interferons também exibem impactos extrínsecos em tumores através da regulação de processos como angiogênese, osteoclastogênese e imunidade (CHEON; BORDEN; STARK, 2014).

No estudo de caso 2 o *SpotComm* foi empregue em um *cluster* transcricional de “respostas de interferon tipo I” definido pela co-localização de macrófagos 2: *CXCL10* e células T: *IFIT1* e com genes marcadores associados às respostas de interferon tipo I pelos autores da pesquisa e do conjunto de dados. Nós avaliamos a comunicação celular da região devido a tal relevância das respostas de interferon do tipo I na biologia e tratamento do câncer uma vez que a ativação de interferon tipo I em microambiente tumoral pode agir diretamente em células do tumor inibindo a proliferação ou estimulando a apoptose ou indiretamente ativando o sistema imune com propósitos antitumorais (ANDERSSON et al., 2021).

Na região do tecido analisada do paciente “G” um *cluster*, o *cluster* 4 foi classificado como o *cluster* transcricional de “respostas de interferon tipo I”. Esse *cluster* mostrou a maior concentração de células epiteliais de câncer e sadias, com *spots* contendo de 0 a 50 % (média de 25.9 %) de epiteliais de câncer e 0 a 47 % (média de 25.5 %) de células epiteliais sadias. Ainda, o *cluster* possuia 3 *spots* (4 % dos *spots*) com mais de 25 % a mais de células epiteliais de câncer que sadias e 2 *spots* (3 % dos *spots*) com mais de 25 % a mais de células epiteliais sadias que de câncer, indicando um balança entre as populações no *cluster*. Os genes marcadores do *cluster* estavam associados a termos de resposta a interferon tipo I como as respostas de defesa a vírus e a termos de células epiteliais como cornificação e queratinização. Então, as distribuições celulares, os genes marcadores e os processos em que os genes estão associados nos garantiram a confiança de estar de olhos em um *cluster* de células epiteliais de câncer e sadias exibindo perfil

transcricional de resposta a interferon tipo I. Além de células epiteliais, os *spots* do *cluster* 4 também mostraram células endoteliais, plasmablásticas e fibroblásticas associadas ao câncer, na maioria. Somente um outro *cluster* mostrou médias de presença de células epiteliais de câncer ou sadias maior que 10 %, o *cluster* 2, que nós classificamos como “Câncer epitelial/Fibroblastos associados ao tumor/Normal epitelial” que mostrou uma média de 21 % de células epitelias de câncer e 16 % de células epiteliais sadias em seus *spots*. Contudo, os marcadores do *cluster* 2 não estão associados a respostas de interferon tipo I (ANDERSSON et al., 2021). Assim, há presença de tais populações em mais *spots*, mas sem detecção do processo biológico particular. Uma próxima análise, assim como no caso 1, irá se focar em classificações mais exploradas, mais especificadas de tipos celulares, utilizando p.e. a camada de subconjunto de Wu e colaboradores (WU et al., 2021b) e pesquisar se há populações específicas distintas entre os *clusters* 4 e 2.

O ligante e o receptor mais expresso em ambas as análises (não focada e focada nos *spots*) foi o ligante *MUC1* e o receptor *ERBB2*. O gene *MUC1* (mucina 1, associada à superfície celular) codifica uma proteína ligada à membrana, membro da família das mucinas. As mucinas são proteínas que desempenham um papel importante na formação de obstruções mucosas protetoras nas superfícies epiteliais além de participar da sinalização intracelular. A proteína codificada é expressa na superfície de células epiteliais que revestem as superfícies mucosas de diferentes tecidos, incluindo a mama. A superexpressão, a localização intracelular anômala e as alterações na glicosilação dessa proteína vem sendo associadas a carcinomas (O’LEARY et al., 2016). O gene é relativo a ontologias como resposta a danos no DNA: transdução de sinal pelo mediador da classe p53 resultando na parada do ciclo celular e na transcrição do mediador da classe p21, é um gene marcador do *cluster* e é expresso 2,54 mais que em outros *clusters* (LIBERZON et al., 2015). A proteína codificada pelo gene *ERBB2* (receptor Erb-B2 tirosina quinase 2, ou HER2), por sua vez, é um membro da família de receptores do fator de crescimento epidérmico de receptores de tirosina quinases. Tal proteína não possui domínio próprio de ligação ao ligante, contudo ela se liga em outros membros da família de receptores do fator de crescimento epidérmico ligados ao ligante para formar um heterodímero, estabilizando a ligação do ligante e aumentando a ativação mediada

por quinase de vias de sinalização a jusante, como as envolvendo proteína quinase ativada por mitógeno e fosfatidilinositol-3 quinase. A amplificação e/ou superexpressão do *ERBB2* vem sendo relatada em vários cânceres, como o câncer de mama HER2-positivo (O'LEARY et al., 2016). É relativo a ontologias como regulação positiva da fosforilação de proteínas e transdução de sinal, é um gene marcador do *cluster* e é expresso 2,62 vezes mais que em outros *clusters* (LIBERZON et al., 2015). De nota, comunicações envolvendo *MUC1* e *ERBB2* estão entre as comunicações cujos pares possuem os maiores valores de escores de comunicação na análise de forma conjunta (*HLA-E:ERBB2*, *MUC1:LGALS3* e outras); comunicações cujos pares estavam presentes em mais comunicações entre *spots* na análise de forma individual (*MUC1:LGALS3*, *HLA-E:ERBB*, presentes em 78 e 64 % das comunicações, e outras); comunicações cujos pares estão bem organizados em coordenadas (*HLA-E:ERBB2* e *AREG:ERBB2* Moran's *I* de 0.12 e 0.07) e outros.

Sabe-se que a interação de MUC1 e LGALS3 (galectina 3) na superfície celular promove dimerização e ativação do receptor do fator de crescimento epidérmico (EGFR) induzido por EGF em células de câncer epitelial humano de mama e de cólon (PIYUSH et al., 2017). O EGFR é um importante regulador do crescimento e sobrevivência das células epiteliais em tecidos normais e cancerosos e é um importante alvo terapêutico para o tratamento do câncer. A ligação entre o domínio extracelular da MUC1 e a LGALS3 induz a polarização MUC1 da superfície celular e aumenta a associação MUC1-EGFR. Isto leva a um rápido aumento da homo-/hetero-dimerização de EGFR e subsequentemente aumentada, e também prolongada, ativação e sinalização de EGFR. Uma vez que ambas MUC1 e LGALS3 são geralmente superexpressos em cânceres epiteliais, sua interação e impacto na ativação do EGFR deve supostamente contribuir para a tumorigênese e a progressão do câncer e influenciar a eficácia de terapias direcionada ao EGFR (PIYUSH et al., 2017). Por sua vez, não obtivemos dados sobre a interação *HLA-E:ERBB2* e ela é contida no banco de dados Omnipath vinda de tão somente uma referência (AWAN et al., 2007) e talvez esse seja um ponto a ser considerado no futuro. Felizmente, os *outputs* gerados pelo *SpotComm* mantém os metadados de cada comunicação no arquivos intercellular_cl_cl_buzzer (para *mode = "spot_n"*) e

intercellular_sp_sp_buzzer (para *mode* = *“spot_y”*) e, assim, é possível executar a filtragem a qualquer instante. A detecção de tais interatores, as ontologias associadas e os dados da literatura associados nos garantiram padrões e dinâmicas típicas de células epiteliais sadias e de câncer de mama HER2-positivo e o *SpotComm* nos permitiu os detectar os padrões em regiões de ocorrência.

Entre os genes de fatores de transcrição, o *STAT1* (transdutor de sinal e ativador de transcrição 1) se destacou em análises de forma individual (*mode* = *“spot_y”*) por ser o fator mais presente e com mais *links*, caminhos, em *spots* especialmente em epiteliais com o receptor *ERBB2*. A proteína codificada por *STAT1*, aludida antes, pode ser ativada por diversos ligantes como interferons tipo I, II e EGF, por exemplo. Sabe-se que a fosforilação de STAT1 na tirosina 701, essencial para a sinalização de interferon, é aumentada em células humanas de câncer de mama com níveis elevados de ERBB2 e em células transfectadas com ERBB2. No entanto, a inibição farmacológica de ERBB2 resulta na inibição da fosforilação de STAT1 na tirosina 701 (RAVEN et al., 2011). Assim, é possível que caminhos alcançados com o *SpotComm* entre o receptor ERBB2 e o fator de transcrição STAT1 em células epiteliais sejam essenciais na sinalização de interferon, particularidade desse *cluster* transcricional. Ainda, em células epiteliais nós predizemos caminhos, vias de interação proteína-proteína (entre os receptores e fatores de transcrição presentes em mais *spots*) entre *ERBB2:JUN* (presente em 37 *spots* de 74), *ERBB2:GATA3* (32) e *ERBB2:JUND* (30). Entre esses, sabe-se que JUN endógeno desempenha um papel importante na migração induzida por ERBB2 e na invasão de células epiteliais mamárias utilizando estas células e tumores mamários derivados do cruzamento de camundongos transgênicos expressando vírus do tumor mamário murino-*ERBB2* com alelos “floxed” *JUN* (JIAO et al., 2010) e embora o imunoconteúdo de GATA3 tenha correlação positiva com o carcinomas de mama ER-positivo, ele foi detectado em 68.5 % dos carcinomas HER2-positivo (SHAOXIAN et al., 2017). Bem como nas interações intercelulares, as comunicações e sinalizações intracelulares indicam os resultados expectados no cenário destacado.

Ainda no *cluster* de “resposta a interferon tipo I” utilizamos a capacidade de expor as interações em estruturas espaciais do *SpotComm* utilizando os espaços de

comunicações intermédios entre *spots* e os produtos de comunicação como seus escores. Nós percebemos que a comunicaçãe *B2M:HLA-E e HLA-E:ERBB2* parecem correlatas, ostentando expressões no *cluster* inteiro. Esses três genes (*B2M*, *HLA-E* e *ERBB2*) são super expressos no câncer de mama HER2-positivo, conforme debatido. Curiosamente vemos o *HLA-E* como “ligante” e como “receptor” nos exemplos e essa é um dos benefícios de usarmos um banco de dados que utilize conceitos de *sources* e *targets* no lugar de ligantes e receptores. Entre as nove principais interações espacialmente organizadas quatro envolvem o receptor *PIK3R1*, interagindo com os ligantes *B2M*, *TNFSF10*, *HLAA* e *CDH1*. A via fosfoinositídeo 3-quinase (PI3K)/proteína quinase B (PKB ou AKT)/alvo mamífero da via de sinalização da rapamicina (mTOR) desempenha um papel importante na progressão do câncer de mama e é intimamente associada a resistência à terapia endócrina, terapia direcionada ao receptor ERBB2 e terapia citotóxica no câncer de mama (NAHTA, 2012; O’REGAN; PAPLOMATA, 2013). Novos estudos comprovam que há ativação residual de ERBB2 da sinalização da via PI3K/AKT que é impulsionada por ERBB2 através de mecanismos diretos e indiretos. Mecanismos indiretos envolvem vias de segundos mensageiros, incluindo RAS ou GRB2. O mecanismo direto envolve a ligação, geralmente fraca, de ERBB2 a PI3K, que se torna significativa em alta expressão e/ou fosforilação (da tirosina 1139 do ERBB2) (RUIZ-SAENZ et al., 2018). O gene autocorrelato *PIK3R1* é parte da via canonica de ativação de PI3K-AKT, ativação de PI3K-AKT em câncer e eventos de PI3K na sinalização ERBB2 e outros do Reactome (LIBERZON et al., 2015). Ainda, digno de nota, nós percebemos que as comunicações envolvendo PIK3R1 é mais intensa onde a ativação de ESR1 é menos intensa. ESR1, ou receptor de estrogênio 1, é um fator de transcrição ativo por ligantes cujo imunoconteúdo é um marcador de câncer de mama luminais, não sendo ou sendo pouco detectado em câncer de mama triplo-negativo e HER2-positivo. Analisando os dados de scRNA-seq vimos que o *PIK3R1* e o *ESR1* são mais expressos células epiteliais sadias que de câncer (*PIK3R1*: 27.9 contra 22.3 %; *ESR1*: 3.0 contra 0.4 %), entretanto o *PIK3R1* é expresso apenas 1.2 vezes a mais, o *ESR1* é expresso 7.5 vezes a mais. Isso nos indica que o centro do *cluster* transcricional é composto prevalentemente de células epiteliais de câncer

com comunicação envolvendo *PIK3R1* e as bordas do *cluster* é composto prevalentemente de células epiteliais sadias com ativação de *ESR1*.

A análise baseada em *spots* em ambos os estudos nos forneceram uma série de ligantes, receptores, pares de interatores na comunicação intercelular e fatores de transcrição ativos não detectados na análise baseada em *clusters*. No estudo de caso 1 e 2 detectamos comunicações intercelulares envolvendo 101 e 150 % a mais de ligantes e 140 e 138 % a mais de receptores na análise baseada em *spots*. Isso se refletiu também no número de pares ligantes/receptores distintos detectados, sendo que observamos 222 e 373 % a mais nos estudos 1 e 2, nessa ordem. No *cluster* de “estruturas linfoides terciárias” os ligantes e receptores exclusivamente detectados em *mode = “spot_y”* estão envolvidos primeiro com ontologias com leucócitos e adesão celular, mas ontologias com os linfócitos também são vistos. Por sua vez, no *cluster* de “resposta a interferon tipo I” estão envolvidos com ontologias de montagem e organização de estruturas externas, adesão celular e migração leucocitária, mas ontologias envolvendo hallmarks de câncer também são vistos como termos “migrar”, “desenvolver” (epitélio da glândula mamária, por exemplo), “aderir” e “crescer/ampliar”. Ainda, nós detectamos 1250 e 1215 % a mais de fatores de transcrição ativos distintos na baseada em *spots*. No *cluster* de “estruturas linfoides terciárias” os alvos de fatores de transcrição ativos exclusivamente detectados em *mode = “spot_y”* estão envolvidos primeiro com ontologias com respostas a estímulos (como drogas, moléculas de origem bacteriana, substância inorgânica) e proliferação celular. Por sua vez, no *cluster* de “resposta a interferon tipo I” estão envolvidos com ontologias de respostas (inflamatória e ao oxigênio) e morte celular por apoptose (via intrínseca). Nós entendemos que com essa nova visão nós localizamos comunicações escondidas na análise baseada em *cluster*, heterogeneidades que serão importantes na formação do conhecimento em diversos cenários na saúde e nas doenças, bem como nos tratamentos.

7. PRINCIPAIS ACHADOS

- Criou-se um algoritmo, o *SpotComm*, com funções capazes de definir as comunicações intercelulares, intracelulares e sinalizações intracelulares se baseando na presença e na correlação de transcritos em dados de transcriptômica espacial e capaz de integrar dados de transcriptômica espacial com dados pareados ou não pareados de scRNA-seq.
- O algoritmo é capaz de manipular dados de transcriptômica espacial em análises de uma seção (2D) ou de mais seções (3D), oferecer análises de *clusters* transcricionais inteiros e/ou zonas de contato entre *clusters*, e proporcionar cenários de comunicação e sinalização entre proteínas interatoras monoméricas e complexos proteicos homomultiméricos e heteromultiméricos.
- O *SpotComm* gerou dados de ofereceu metadados ao usuário como a presença, proporção e expressão de interatores e a referência e curagem de comunicações e sinalizações. Ainda, ofereceu a proporção de células e proporção de co-ocorrência, por tipo celular, com detecção dos elementos de vias de sinalização intracelular.
- Utilizando o algoritmo nós fomos capazes de predizer comunicações inter, intracelulares e sinalizações celulares que são conhecidas em estruturas linfoides terciárias e em áreas de tumor com o perfil de respostas de interferon tipo I em câncer de mama. Ainda, nós predizemos os elementos celulares viáveis às sinalizações.
- Detectamos diversas potenciais comunicações inter e intracelulares não detectadas em análise baseada em clusters que podem ser importantes no entendimento da homeostase mas também nos prognósticos, diagnósticos e tratamentos de doenças.

8. CONCLUSÃO E PERSPECTIVAS

8.1 CONCLUSÃO

Com a criação e a utilização de um algoritmo de análise de comunicação celular que utiliza os dados de coordenadas como unidade de análise, esse trabalho pôde caracterizar a distribuição de comunicações intercelulares e intracelulares em câncer de mama, podendo ser espacialmente organizadas e relevantes e quiçá essenciais para descrição da interação do tecido.

8.2 PERSPECTIVAS

- Retirar a dependência de pacotes de preparação de *input* do *SpotComm*, como o *Seurat* e o *STutility*.
- Utilizar o *SpotComm* em plataformas de transcriptômica espacial de resolução celular como MERFISH empregando valores moderados de distância entre as células, preferentemente alcançando os vizinhos mais próximos, visando obter índices de correlações e autocorrelações maiores e predizer interações novas utilizando redes neurais profundas.
- Incluir no nosso *pipeline* de preparação de *input* do *SpotComm* um algoritmo de retirada do *spot swapping*, como o *SpotClean*.
- Incorporar escores de termos de ontologias em *spots* que demonstrem um processo/função decorrente da ativação de um receptor/via e isso reforce as regulações/comunicações preditas.
- Incorporar a possibilidade de haver dados pareados de epigenômica e proteômica para reforçar as regulações/comunicações preditas.
- Adicionar um cálculo de produto da expressão que considere os dados de expressão escalonados.
- Inserir algoritmos de teoria de grafos como as medidas de centralidade e conectividade para quantificar a importância na rede de cada nó.
- Em estudos de caso, utilizar deconvolução em subtipos mais específicos onde poderemos ligar as comunicações raras a tipos celulares raros ou a

interações raras no *cluster*.

- Testar o *SpotComm* em tecidos de diferentes complexidades em suas populações celulares em geral e em *spots*.
- Validar as áreas de análise do SpotComm, como as estruturas linfóides terciárias com patologistas.
- Definir filtros e métodos de classificação de achados que auxiliem e promovam a descoberta do conhecimento.

REFERÊNCIAS

ACHIM, K. et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*, v. 33, n. 5, p. 503–509, 12 maio 2015. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/25867922/>. Acesso em: 3 jun. 2022.

ALWINE, J. C.; KEMP, D. J.; STARK, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, v. 74, n. 12, p. 5350–5354, 1977. Disponível em: <https://www.pnas.org>. Acesso em: 2 jun. 2022.

ANDERSSON, A. et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, v. 12, n. 1, 1 dez. 2021. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/34650042/>. Acesso em: 22 set. 2022.

ARMINGOL, E. et al. *Deciphering cell–cell interactions and communication from gene expressionNature Reviews Genetics*Nature Publishing Group, , 9 nov. 2021. . Disponível em: <https://www.nature.com/articles/s41576-020-00292-x>. Acesso em: 30 maio. 2022.

AWAN, A. et al. Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. In: IET Systems Biology, 5, *Anais*...set. 2007.

BACCIN, C. et al. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nature Cell Biology*, v. 22, n. 1, p. 38–48, 23 dez. 2020. Disponível em: <https://www.nature.com/articles/s41556-019-0439-6>. Acesso em: 12 maio. 2022.

BALKWILL, F.; WATLING, D.; TAYLOR-PAPADIMITRIOU, J. Inhibition by lymphoblastoid interferon of growth of cells derived from the human breast. *International Journal of Cancer*, v. 22, n. 3, p. 258–265, 1978. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/700890/>. Acesso em: 25 set. 2022.

BARCLAY, J.; CRESWELL, J.; LEÓN, J. *Cancer immunotherapy and the PD-1/PD-L1 checkpoint pathwayArchivos Espanoles de Urologia*Iniestares, S.A., , 1 maio 2018. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/29745928/>. Acesso em: 11 set. 2020.

BERNARD, P. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, v. 27, n. 8, p. 1160–1167, 2009.

BINDEA, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, v. 39, n. 4, p. 782–795, 17 out. 2013. Disponível em: <http://www.cell.com/article/S1074761313004378/fulltext>. Acesso em: 25 set. 2022.

BOISSET, J. C. et al. Mapping the physical network of cellular interactions. *Nature Methods*, v. 15, n. 7, p. 547–553, 1 jul. 2018. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/29786092/>. Acesso em: 3 jun. 2022.

BRAUD, V. M. et al. HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature*, v. 391, n. 6669, p. 795–799, 19 fev. 1998. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/9486650/>. Acesso em: 22 set. 2022.

BROWAEYS, R.; SAELENS, W.; SAEYS, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, v. 17, n. 2, p. 159–162, 9 dez. 2020. Disponível em: <https://www.nature.com/articles/s41592-019-0667-5>. Acesso em: 12 maio. 2022.

BUCCITELLI, C.; SELBACH, M. *mRNAs, proteins and the emerging principles of gene expression control*Nature Reviews Genetics Nature Publishing Group, , 24 jul. 2020. . Disponível em: <https://www.nature.com/articles/s41576-020-0258-4>. Acesso em: 16 jun. 2022.

BUMGARNER, R. Overview of dna microarrays: Types, applications, and their future. *Current Protocols in Molecular Biology*, v. 0 22, n. SUPPL.101, p. Unit, 2013. Disponível em: </pmc/articles/PMC4011503/>. Acesso em: 16 jun. 2022.

CABELLO-AGUILAR, S. et al. SingleCellSignalR: Inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Research*, v. 48, n. 10, 2021. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/32196115/>. Acesso em: 30 maio. 2022.

CAIN, M. P.; HERNANDEZ, B. J.; CHEN, J. Quantitative single-cell interactomes in normal and virus-infected mouse lungs. *DMM Disease Models and Mechanisms*, v. 13, n. 6, 1 jun. 2020. Disponível em: <https://journals.biologists.com/dmm/article/13/6/dmm044404/225260/Quantitative-single-cell-interactomes-in-normal>. Acesso em: 12 maio. 2022.

CANG, Z.; NIE, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, v. 11, n. 1, 1 dez. 2020. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/32350282/>. Acesso em: 30 maio. 2022.

CHEN, T. C. et al. Using an in situ proximity ligation assay to systematically profile endogenous protein-protein interactions in a pathway network. *Journal of Proteome Research*, v. 13, n. 12, p. 5339–5346, 5 dez. 2014. Disponível em: <https://pubs.acs.org/doi/abs/10.1021/pr5002737>. Acesso em: 25 set. 2022.

CHEON, H.; BORDEN, E. C.; STARK, G. R. Interferons and their stimulated genes in the tumor microenvironment. *Seminars in Oncology*, v. 41, n. 2, p. 156–173, 2014. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/24787290/>. Acesso em: 25 set. 2022.

CHOI, E. A. et al. Stat1-dependent induction of tumor necrosis factor-related apoptosis-inducing ligand and the cell-surface death signaling pathway by interferon β in human cancer cells. *Cancer Research*, v. 63, n. 17, p. 5299–5307, 2003. Disponível em: <https://aacrjournals.org/cancerres/article/63/17/5299/510466/Stat1-dependent-Induction-of-Tumor-Necrosis-Factor>. Acesso em: 25 set. 2022.

CHOI, H. et al. Transcriptome Analysis of Individual Stromal Cell Populations Identifies Stroma-Tumor Crosstalk in Mouse Lung Cancer Model. *Cell Reports*, v. 10, n. 7, p. 1187–1201, 24 fev. 2015. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/25704820/>. Acesso em: 10 maio. 2022.

CHUANG, H. Y. et al. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, v. 3, 2007.

CILLO, A. R. et al. Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity*, v. 52, n. 1, p. 183- 199.e9, 14 jan. 2020. Disponível em: <http://www.cell.com/article/S1074761319304959/fulltext>. Acesso em: 12 maio. 2022.

COMBS, P. A.; EISEN, M. B. Sequencing mRNA from Cryo-Sliced Drosophila Embryos to Determine Genome-Wide Spatial Patterns of Gene Expression. *PLoS ONE*, v. 8, n. 8, p. 71820, 12 ago. 2013. Disponível em: </pmc/articles/PMC3741199/>. Acesso em: 3 jun. 2022.

COPPOLA, D. et al. Unique ectopic lymph node-like structures present in human primary colorectal carcinoma are identified by immune gene array profiling. *American Journal of Pathology*, v. 179, n. 1, p. 37–45, 1 jul. 2011. Disponível em: <http://ajp.amjpathol.org/article/S0002944011003208/fulltext>. Acesso em: 25 set. 2022.

COUPEL, S. et al. Expression and release of soluble HLA-E is an immunoregulatory feature of endothelial cell activation. *Blood*, v. 109, n. 7, p. 2806–2814, 1 abr. 2007. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/17179229/>. Acesso em: 22 set. 2022.

DA ROCHA, E. L. et al. NetDecoder: A network biology platform that decodes context-specific biological networks and gene activities. *Nucleic Acids Research*, v. 44, n. 10, p. e100–e100, 2 jun. 2016. Disponível em: <https://academic.oup.com/nar/article/44/10/e100/2516396>. Acesso em: 25 set. 2022.

DE CHAISEMARTIN, L. et al. Characterization of chemokines and adhesion molecules associated with T cell presence in tertiary lymphoid structures in human lung cancer. *Cancer Research*, v. 71, n. 20, p. 6391–6399, 15 out. 2011. Disponível em: <https://aacrjournals.org/cancerres/article/71/20/6391/568120/Characterization-of-Chemokines-and-Adhesion>. Acesso em: 25 set. 2022.

DEJARDIN, E. et al. The lymphotoxin-β receptor induces different patterns of gene expression via two NF-κB pathways. *Immunity*, v. 17, n. 4, p. 525–535, 1 out. 2002. Disponível em: <http://www.cell.com/article/S1074761302004235/fulltext>. Acesso em: 25 set. 2022.

DIJKSTRA, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik*, v. 1, n. 1, p. 269–271, dez. 1959. Disponível em: <https://link.springer.com/article/10.1007/BF01386390>. Acesso em: 24 set. 2022.

DONNELLY, R. P.; KOTENKO, S. V. *Interferon-lambda: A new addition to an old familyJournal of Interferon and Cytokine Research*J Interferon Cytokine Res, , 1 ago. 2010. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/20712453/>. Acesso em: 25 set. 2022.

DRIES, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, v. 22, n. 1, p. 1–31, 1 dez. 2021. Disponível em: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02286-2>. Acesso em: 30 maio. 2022.

FARR, L.; GHOSH, S.; MOONAH, S. *Role of MIF Cytokine/CD74 Receptor Pathway in Protecting Against Injury and Promoting RepairFrontiers in Immunology*Frontiers Media S.A., , 23 jun. 2020. .

FAWKNER-CORBETT, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell*, v. 184, n. 3, p. 810- 826.e23, 4 fev. 2021a. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/33406409/>. Acesso em: 15 jun. 2022.

FAWKNER-CORBETT, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell*, v. 184, n. 3, p. 810- 826.e23, 4 fev. 2021b.

FODOR, S. P. A. et al. Light-directed, spatially addressable parallel chemical synthesis. *Science*, v. 251, n. 4995, p. 767–773, 1991. Disponível em: <https://www.science.org/doi/abs/10.1126/science.1990438>. Acesso em: 2 jun. 2022.

FULDA, S.; DEBATIN, K. M. *Extrinsic versus intrinsic apoptosis pathways in anticancer chemotherapyOncogene*Oncogene, , 7 ago. 2006. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/16892092/>. Acesso em: 25 set. 2022.

GILADI, A. et al. Dissecting cellular crosstalk by sequencing physically interacting cells. *Nature Biotechnology*, v. 38, n. 5, p. 629–637, 1 maio 2020. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/32152598/>. Acesso em: 3 jun. 2022.

GOC, J. et al. Dendritic cells in tumor-associated tertiary lymphoid structures signal a th1 cytotoxic immune contexture and license the positive prognostic value of infiltrating CD8+ t cells. *Cancer Research*, v. 74, n. 3, p. 705–715, 1 fev. 2014. Disponível em: <https://aacrjournals.org/cancerres/article/74/3/705/599302/Dendritic-Cells-in-Tumor-Associated-Tertiary>. Acesso em: 25 set. 2022.

GORE, Y. et al. Macrophage migration inhibitory factor induces B cell survival by activation of a CD74-CD44 receptor complex. *Journal of Biological Chemistry*, v. 283, n. 5, p. 2784–2792, 1 fev. 2008. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/18056708/>. Acesso em: 22 set. 2022.

GOSWAMI, R.; KAPLAN, M. H. STAT Transcription Factors in T Cell Control of Health and Disease. In: *International Review of Cell and Molecular Biology*. [s.l.] Academic Press, 2017. 331p. 123–180.

GRAEBER, T. G.; EISENBERG, D. Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nature Genetics*, v. 29, n. 3, p. 295–300, 29 out. 2001. Disponível em: <https://www.nature.com/articles/ng755>. Acesso em: 10 maio. 2022.

GREENWOOD, C. et al. Stat1 and CD74 overexpression is co-dependent and linked to increased invasion and lymph node metastasis in triple-negative breast cancer. *Journal of Proteomics*, v. 75, n. 10, p. 3031–3040, 6 jun. 2012.

GU-TRANTIEN, C. et al. CD4+ follicular helper T cell infiltration predicts breast cancer survival. *Journal of Clinical Investigation*, v. 123, n. 7, p. 2873–2892, 1 jul. 2013. Disponível em: <http://www.jci.org>. Acesso em: 25 set. 2022.

HAFEMEISTER, C.; SATIJA, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, v. 20, n. 1, p. 1–15, 23 dez. 2019. Disponível em:

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1>. Acesso em: 23 set. 2022.

HAQUE, A. et al. *A practical guide to single-cell RNA-sequencing for biomedical research and clinical applicationsGenome Medicine*BioMed Central Ltd., , 18 ago. 2017. . Disponível em: <https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4>. Acesso em: 16 jun. 2022.

HARBECK, N. et al. Breast cancer. *Nature Reviews Disease Primers*, v. 5, n. 1, p. 1–31, 23 set. 2019. Disponível em: <https://www.nature.com/articles/s41572-019-0111-2>. Acesso em: 3 jun. 2022.

HARPER, J. W.; BENNETT, E. J. *Proteome complexity and the forces that drive proteome imbalanceNature*Nature Publishing Group, , 14 set. 2016. . Disponível em: <https://www.nature.com/articles/nature19947>. Acesso em: 3 jun. 2022.

HELMINK, B. A. et al. B cells and tertiary lymphoid structures promote immunotherapy response. *Nature*, v. 577, n. 7791, p. 549–555, 15 jan. 2020. Disponível em: <https://www.nature.com/articles/s41586-019-1922-8>. Acesso em: 22 set. 2022.

HENNEQUIN, A. et al. Tumor infiltration by Tbet+ effector T cells and CD20+ B cells is associated with survival in gastric cancer patients. *OncoImmunology*, v. 5, n. 2, 1 fev. 2016. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/2162402X.2015.1054598>. Acesso em: 25 set. 2022.

HERTZOG, P. J.; WILLIAMS, B. R. G. *Fine tuning type I interferon responsesCytokine and Growth Factor Reviews*Cytokine Growth Factor Rev, , jun. 2013. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/23711406/>. Acesso em: 25 set. 2022.

HIGUCHI, R. et al. Simultaneous amplification and detection of specific DNA sequences. *Bio/Technology*, v. 10, n. 4, p. 413–417, 1992. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/1368485/>. Acesso em: 2 jun. 2022.

HIGUCHI, R. et al. Kinetic PCR analysis: Real-time monitoring of DNA amplification reactions. *Bio/Technology*, v. 11, n. 9, p. 1026–1030, 1993. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/7764001/>. Acesso em: 2 jun. 2022.

HOARE, H. L. et al. Subtle Changes in Peptide Conformation Profoundly Affect Recognition of the Non-Classical MHC Class I Molecule HLA-E by the CD94-NKG2 Natural Killer Cell Receptors. *Journal of Molecular Biology*, v. 377, n. 5, p.

1297–1303, 11 abr. 2008. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/18339401/>. Acesso em: 22 set. 2022.

HOBEIKA, A. C.; SUBRAMANIAM, P. S.; JOHNSON, H. M. IFNα induces the expression of the cyclin-dependent kinase inhibitor p21 in human prostate cancer cells. *Oncogene*, v. 14, n. 10, p. 1165–1170, 1997. Disponível em: <https://www.nature.com/articles/1200939>. Acesso em: 25 set. 2022.

HU, Y. et al. CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Science Advances*, v. 7, n. 16, 14 abr. 2021. Disponível em: <https://www.science.org/doi/full/10.1126/sciadv.abf1356>. Acesso em: 30 maio. 2022.

JI, A. L. et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Journal of Cleaner Production*, v. 182, n. 2, 2020. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/32579974/>. Acesso em: 23 set. 2022.

JIAO, X. et al. c-Jun induces mammary epithelial cellular invasion and breast cancer stem cell expansion. *Journal of Biological Chemistry*, v. 285, n. 11, p. 8218–8226, 12 mar. 2010. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/20053993/>. Acesso em: 25 set. 2022.

JIN, S. et al. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, v. 12, n. 1, p. 1–20, 17 fev. 2021. Disponível em: <https://www.nature.com/articles/s41467-021-21246-9>. Acesso em: 30 maio. 2022.

JONES, R. B. et al. *A quantitative protein interaction network for the ErbB receptors using protein microarraysNature*Nature Publishing Group, , 6 nov. 2006. . Disponível em: <https://www.nature.com/articles/nature04177>. Acesso em: 25 set. 2022.

JUNKER, J. P. et al. Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell*, v. 159, n. 3, p. 662–675, 23 out. 2014.

KING, H. C.; SINHA, A. A. *Gene expression profile analysis by DNA microarrays: Promise and pitfallsJournal of the American Medical Association*American Medical Association, , 14 nov. 2001. . Disponível em: <https://jamanetwork.com/journals/jama/fullarticle/194368>. Acesso em: 2 jun. 2022.

KINKER, G. S. et al. *B Cell Orchestration of Anti-tumor Immune Responses: A Matter of Cell Localization and CommunicationFrontiers in Cell and Developmental Biology*Front Cell Dev Biol, , 7 jun. 2021. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/34164398/>. Acesso em: 22 set. 2022.

KOMUROV, K. Modeling community-wide molecular networks of multicellular systems. *Bioinformatics*, v. 28, n. 5, p. 694–700, mar. 2012. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/22210865/>. Acesso em: 10 maio. 2022.

KRISHNAMURTI, U.; SILVERMAN, J. F. *HER2 in breast cancer: A review and updateAdvances in Anatomic Pathology*Lippincott Williams and Wilkins, , 2014. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/24508693/>. Acesso em: 11 set. 2020.

KUKURBA, K. R.; MONTGOMERY, S. B. RNA sequencing and analysis. *Cold Spring Harbor Protocols*, v. 2015, n. 11, p. 951–969, 1 nov. 2015. Disponível em: </pmc/articles/PMC4863231/>. Acesso em: 16 jun. 2022.

KUMAR, M. P. et al. Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Reports*, v. 25, n. 6, p. 1458- 1468.e4, 6 nov. 2018. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/30404002/>. Acesso em: 12 maio. 2022.

KUMAR, P.; AGGARWAL, R. *An overview of triple-negative breast cancerArchives of Gynecology and Obstetrics*Springer Verlag, , 1 fev. 2016. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/26341644/>. Acesso em: 11 set. 2020.

LACRAZ, G. P. A. et al. Tomo-Seq Identifies SOX9 as a Key Regulator of Cardiac Fibrosis during Ischemic Injury. *Circulation*, v. 136, n. 15, p. 1396–1409, 10 out. 2017. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/28724751/>. Acesso em: 3 jun. 2022.

LIBERZON, A. et al. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, v. 1, n. 6, p. 417–425, 12 dez. 2015. Disponível em: </pmc/articles/PMC4707969/>. Acesso em: 22 set. 2022.

LIVAK, K. J.; SCHMITTGEN, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods*, v. 25, n. 4, p. 402–408, 1 dez. 2001.

LLANO, M. et al. HLA-E-bound peptides influence recognition by inhibitory and triggering CD94/NKG2 receptors: Preferential response to an HLA-G-derived nonamer. *European Journal of Immunology*, v. 28, n. 9, p. 2854–2863, 1998.

LOIBL, S. et al. *Breast cancerThe Lancet*Elsevier, , 8 maio 2021. . Disponível em: <http://www.thelancet.com/article/S0140673620323813/fulltext>. Acesso em: 3 jun. 2022.

LUMMERTZ DA ROCHA, E. et al. Reconstruction of complex single-cell trajectories using CellRouter. *Nature Communications*, v. 9, n. 1, p. 1–13, 1 mar.

2018. Disponível em: <https://www.nature.com/articles/s41467-018-03214-y>. Acesso em: 23 set. 2022.

LUMMERTZ DA ROCHA, E. et al. CellComm infers cellular crosstalk that drives haematopoietic stem and progenitor cell development. *Nature Cell Biology*, v. 24, n. 4, p. 579–589, 12 abr. 2022. Disponível em: <https://www.nature.com/articles/s41556-022-00884-1>. Acesso em: 12 maio. 2022.

MANCO, R. et al. Clump sequencing exposes the spatial expression programs of intestinal secretory cells. *Nature Communications*, v. 12, n. 1, 1 dez. 2021. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/34031373/>. Acesso em: 3 jun. 2022.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, v. 437, n. 7057, p. 376–380, 15 set. 2005. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/16056220/>. Acesso em: 2 jun. 2022.

MARX, V. Method of the Year: spatially resolved transcriptomics. *Nature Methods*, v. 18, n. 1, p. 9–14, 6 jan. 2021. Disponível em: <https://www.nature.com/articles/s41592-020-01033-y>. Acesso em: 3 jun. 2022.

MESSINA, J. L. et al. 12-chemokine gene signature identifies lymph node-like structures in melanoma: Potential for patient selection for immunotherapy? *Scientific Reports*, v. 2, n. 1, p. 1–6, 24 out. 2012. Disponível em: <https://www.nature.com/articles/srep00765>. Acesso em: 25 set. 2022.

MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, v. 5, n. 7, p. 621–628, 30 maio 2008. Disponível em: <https://www.nature.com/articles/nmeth.1226>. Acesso em: 2 jun. 2022.

MULLIS, K. B. The unusual origin of the polymerase chain reaction. *Scientific American*, v. 262, n. 4, p. 56–65, 1990. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/2315679/>. Acesso em: 2 jun. 2022.

MULLIS, K. B.; FALOONA, F. A. Specific Synthesis of DNA in Vitro via a Polymerase-Catalyzed Chain Reaction. *Methods in Enzymology*, v. 155, n. C, p. 335–350, 1987. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/3431465/>. Acesso em: 2 jun. 2022.

NAGALAKSHMI, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, v. 320, n. 5881, p. 1344–1349, 6 jun. 2008. Disponível em: <https://www.science.org/doi/abs/10.1126/science.1158441>. Acesso em: 2 jun. 2022.

NAHTA, R. Pharmacological Strategies to Overcome HER2 Cross-Talk and Trastuzumab Resistance. *Current Medicinal Chemistry*, v. 19, n. 7, p. 1065–1075, 16

out. 2012. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/22229414/>. Acesso em: 25 set. 2022.

NI, Z. et al. SpotClean adjusts for spot swapping in spatial transcriptomics data. *Nature Communications*, v. 13, n. 1, p. 1–11, 27 maio 2022. Disponível em: <https://www.nature.com/articles/s41467-022-30587-y>. Acesso em: 23 set. 2022.

NIETHAMER, T. K. et al. Defining the role of pulmonary endothelial cell heterogeneity in the response to acute lung injury. *eLife*, v. 9, 1 fev. 2020. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/32091393/>. Acesso em: 3 jun. 2022.

NOËL, F. et al. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nature Communications*, v. 12, n. 1, p. 1–16, 17 fev. 2021. Disponível em: <https://www.nature.com/articles/s41467-021-21244-x>. Acesso em: 13 maio. 2022.

O'LEARY, N. A. et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, v. 44, n. D1, p. D733–D745, 2016. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/26553804/>. Acesso em: 22 set. 2022.

O'REGAN, R.; PAPLOMATA. New and emerging treatments for estrogen receptor-positive breast cancer: focus on everolimus. *Therapeutics and Clinical Risk Management*, v. 9, p. 27, jan. 2013. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/23345981/>. Acesso em: 25 set. 2022.

PARK, J. E. et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science*, v. 367, n. 6480, 21 fev. 2020. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/32079746/>. Acesso em: 3 jun. 2022.

PARKER, B. S.; RAUTELA, J.; HERTZOG, P. J. *Antitumour actions of interferons: Implications for cancer therapyNature Reviews Cancer*Nat Rev Cancer, , 1 mar. 2016. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/26911188/>. Acesso em: 25 set. 2022.

PEASE, A. C. et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America*, v. 91, n. 11, p. 5022–5026, 24 maio 1994. Disponível em: <https://www.pnas.org>. Acesso em: 2 jun. 2022.

PEROU, C. M. et al. Molecular portraits of human breast tumours. *Nature*, v. 406, n. 6797, p. 747–752, 2000.

PETTIT, J. B. et al. Identifying Cell Types from Spatially Referenced Single-Cell Expression Datasets. *PLoS Computational Biology*, v. 10, n. 9, 1 set. 2014.

Disponível em: <https://pubmed.ncbi.nlm.nih.gov/25254363/>. Acesso em: 3 jun. 2022.

PIYUSH, T. et al. Interaction of galectin-3 with MUC1 on cell surface promotes EGFR dimerization and activation in human epithelial cancer cells. *Cell Death and Differentiation*, v. 24, n. 11, p. 1931–1947, 21 jul. 2017. Disponível em: <https://www.nature.com/articles/cdd2017119>. Acesso em: 25 set. 2022.

RAMILOWSKI, J. A. et al. A draft network of ligand-receptor-mediated multicellular signalling in human. *Nature Communications*, v. 6, n. 1, p. 1–12, 22 jul. 2015. Disponível em: <https://www.nature.com/articles/ncomms8866>. Acesso em: 10 maio. 2022.

RAO, A. et al. *Exploring tissue architecture using spatial transcriptomicsNature*Nature Publishing Group, , 11 ago. 2021. . Disponível em: <https://www.nature.com/articles/s41586-021-03634-9>. Acesso em: 3 jun. 2022.

RAVEN, J. F. et al. Stat1 is a suppressor of ErbB2/Neu-mediated cellular transformation and mouse mammary gland tumor formation. *Cell Cycle*, v. 10, n. 5, p. 794–804, 1 mar. 2011. Disponível em: <https://www.tandfonline.com/doi/abs/10.4161/cc.10.5.14956>. Acesso em: 25 set. 2022.

RICHELLE, A.; JOSHI, C.; LEWIS, N. E. Assessing key decisions for transcriptomic data integration in biochemical networks. *PLoS Computational Biology*, v. 15, n. 7, 1 jul. 2019. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/31323017/>. Acesso em: 30 maio. 2022.

RODRIQUES, S. G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, v. 363, n. 6434, p. 1463–1467, 29 mar. 2019. Disponível em: <https://www.science.org/doi/10.1126/science.aaw1219>. Acesso em: 23 set. 2022.

ROUAULT, H.; HAKIM, V. Different cell fates from cell-cell interactions: Core architectures of two-cell bistable networks. *Biophysical Journal*, v. 102, n. 3, p. 417–426, 8 fev. 2012. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/22325263/>. Acesso em: 3 jun. 2022.

RUIZ-SAENZ, A. et al. HER2 amplification in tumors activates PI3K/Akt signaling independent of HER3. *Cancer Research*, v. 78, n. 13, p. 3645–3658, 1 jul. 2018. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/29760043/>. Acesso em: 26 set. 2022.

SANTOS, R. et al. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, v. 16, n. 1, p. 19–34, 29 dez. 2016. Disponível em: </pmc/articles/PMC6314433/>. Acesso em: 3 jun. 2022.

SATIJA, R. et al. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, v. 33, n. 5, p. 495–502, 12 maio 2015. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/25867923/>. Acesso em: 3 jun. 2022.

SCHRODER, K. et al. Interferon-γ: an overview of signals, mechanisms and functions. *Journal of Leukocyte Biology*, v. 75, n. 2, p. 163–189, fev. 2004. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/14525967/>. Acesso em: 25 set. 2022.

SCHÜTZ, F. et al. *PD-1/PD-L1 Pathway in Breast CancerOncology Research and Treatment*S. Karger AG, , 1 abr. 2017. . Disponível em: <https://pubmed.ncbi.nlm.nih.gov/28346916/>. Acesso em: 11 set. 2020.

SHAOXIAN, T. et al. Characterisation of GATA3 expression in invasive breast cancer: Differences in histological subtypes and immunohistochemically defined molecular subtypes. *Journal of Clinical Pathology*, v. 70, n. 11, p. 926–934, 1 nov. 2017. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/28428285/>. Acesso em: 25 set. 2022.

SØRLIE, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, v. 98, n. 19, p. 10869–10874, 11 set. 2001. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/11553815/>. Acesso em: 11 set. 2020.

SRIVASTAVA, R. M. et al. Stat1-induced HLA class i upregulation enhances immunogenicity and clinical response to anti-EGFR mab cetuximab therapy in HNC patients. *Cancer Immunology Research*, v. 3, n. 8, p. 936–945, 1 ago. 2015. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/25972070/>. Acesso em: 22 set. 2022.

STICKELS, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology*, v. 39, n. 3, p. 313–319, 1 mar. 2021. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/33288904/>. Acesso em: 23 set. 2022.

SULLIVAN, L. C. et al. The Heterodimeric Assembly of the CD94-NKG2 Receptor Family and Implications for Human Leukocyte Antigen-E Recognition. *Immunity*, v. 27, n. 6, p. 900–911, 21 dez. 2007. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/18083576/>. Acesso em: 22 set. 2022.

SURYAWANSHI, H. et al. A single-cell survey of the human first-trimester placenta and decidua. *Science Advances*, v. 4, n. 10, 31 out. 2018. Disponível em: <https://www.science.org/doi/full/10.1126/sciadv.aau4788>. Acesso em: 12 maio. 2022.

SZTANKA-TOTH, T. R. et al. Spacemake: processing and analysis of large-scale spatial transcriptomics data. *GigaScience*, v. 11, p. 1–14, 20 set. 2022. Disponível em: <https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac064/6646447>. Acesso em: 23 set. 2022.

TANG, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, v. 6, n. 5, p. 377–382, 2009. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/19349980/>. Acesso em: 2 jun. 2022.

THYRELL, L. et al. Mechanisms of interferon-alpha induced apoptosis in malignant cells. *Oncogene*, v. 21, n. 8, p. 1251–1262, 20 fev. 2002. Disponível em: <https://www.nature.com/articles/1205179>. Acesso em: 25 set. 2022.

TIROSH, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, v. 352, n. 6282, p. 189–196, 8 abr. 2016. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/27124452/>. Acesso em: 3 jun. 2022.

TSUYUZAKI, K.; ISHII, M.; NIKAIDO, I. Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data. *bioRxiv*, n. Cci, p. 566182, 4 mar. 2019. Disponível em: <https://www.biorxiv.org/content/10.1101/566182v1>. Acesso em: 30 maio. 2022.

TÜREI, D. et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, v. 17, n. 3, p. e9923, 1 mar. 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.15252/msb.20209923>. Acesso em: 23 set. 2022.

TÜREI, D.; KORCSMÁROS, T.; SAEZ-RODRIGUEZ, J. *OmniPath: Guidelines and gateway for literature-curated signaling pathway resourcesNature Methods*Nature Publishing Group, , 29 nov. 2016. . Disponível em: <https://www.nature.com/articles/nmeth.4077>. Acesso em: 23 set. 2022.

TYLER, S. R. et al. PyMINEr Finds Gene and Autocrine-Paracrine Networks from Human Islet scRNA-Seq. *Cell Reports*, v. 26, n. 7, p. 1951- 1964.e8, 12 fev. 2019. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/30759402/>. Acesso em: 17 maio. 2022.

VALLEJO, J. et al. *Heterogeneity of immune cells in human atherosclerosis revealed by scRNA-SeqCardiovascular Research*Oxford University Press, , 11 nov. 2021. . Disponível em: </pmc/articles/PMC8921647/>. Acesso em: 2 jun. 2022.

VAN DIJK, D. et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, v. 174, n. 3, p. 716- 729.e27, 26 jul. 2018. Disponível em: <http://www.cell.com/article/S0092867418307244/fulltext>. Acesso em: 30 maio. 2022.

VENTO-TORMO, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, v. 563, n. 7731, p. 347–353, 14 nov. 2018. Disponível em: <https://www.nature.com/articles/s41586-018-0698-6>. Acesso em: 10 maio. 2022.

WANG, G.; MOFFITT, J. R.; ZHUANG, X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Scientific Reports*, v. 8, n. 1, p. 1–13, 19 mar. 2018. Disponível em: <https://www.nature.com/articles/s41598-018-22297-7>. Acesso em: 23 set. 2022.

WANG, S. et al. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Research*, v. 47, n. 11, 20 jun. 2019a. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/30923815/>. Acesso em: 30 maio. 2022.

WANG, S. et al. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Research*, v. 47, n. 11, p. e66–e66, 20 jun. 2019b. Disponível em: <https://academic.oup.com/nar/article/47/11/e66/5421812>. Acesso em: 10 maio. 2022.

WANG, Y. et al. ITALK: An R Package to Characterize and Illustrate Intercellular Communication. *bioRxiv*, p. 507871, 4 jan. 2019c. Disponível em: <https://www.biorxiv.org/content/10.1101/507871v1>. Acesso em: 17 maio. 2022.

WANG, Z.; GERSTEIN, M.; SNYDER, M. *RNA-Seq: A revolutionary tool for transcriptomicsNature Reviews Genetics*Nature Publishing Group, , jan. 2009. . Disponível em: <https://www.nature.com/articles/nrg2484>. Acesso em: 2 jun. 2022.

WINTER, S. et al. The chemokine receptor CXCR5 is pivotal for ectopic mucosa-associated lymphoid tissue neogenesis in chronic Helicobacter pylori-induced inflammation. *Journal of Molecular Medicine*, v. 88, n. 11, p. 1169–1180, 27 nov. 2010. Disponível em: <https://link.springer.com/article/10.1007/s00109-010-0658-6>. Acesso em: 25 set. 2022.

WU, F. et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nature Communications*, v. 12, n. 1, p. 1–11, 5 maio 2021a. Disponível em: <https://www.nature.com/articles/s41467-021-22801-0>. Acesso em: 2 jun. 2022.

WU, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, v. 53, n. 9, p. 1334–1347, 1 set. 2021b. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/34493872/>. Acesso em: 23 set. 2022.

YEN, J. Y. Finding the K Shortest Loopless Paths in a Network. *Management Science*, v. 17, n. 11, p. 712–716, 1 jul. 1971. Disponível em: <https://pubsonline.informs.org/doi/abs/10.1287/mnsc.17.11.712>. Acesso em: 24 set. 2022.

ZEPP, J. A. et al. Distinct Mesenchymal Lineages and Niches Promote Epithelial Self-Renewal and Myofibrogenesis in the Lung. *Cell*, v. 170, n. 6, p. 1134-1148.e10, 7 set. 2017. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/28886382/>. Acesso em: 3 jun. 2022.

APÊNDICES

Os apêndices se encontram na página: https://bit.ly/3CPH5R0. Caso haja necessidade, contatar via endereço eletrônico: mlbfalchetti@gmail.com