

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

Ohanez Mamigonian

CellComm v2: inferência da dinâmica de redes de comunicação celular durante o
desenvolvimento hematopoiético

Florianópolis, Santa Catarina

2025

Ohanez Mamigonian

CellComm v2: inferência da dinâmica de redes de comunicação celular durante o desenvolvimento hematopoiético

Dissertação submetida ao Programa de Pós-Graduação em Farmacologia da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Farmacologia

Orientador: Prof. Dr. Edroaldo Lummertz da Rocha

Florianópolis, Santa Catarina

2025

Mamigonian, Ohanez

CellComm v2: inferência da dinâmica de redes de comunicação celular durante o desenvolvimento hematopoiético / Ohanez Mamigonian ; orientador, Edroaldo Lummertz da Rocha, 2025.

73 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Farmacologia, Florianópolis, 2025.

Inclui referências.

1. Farmacologia. 2. Interações célula-célula. 3. Bioinformática. 4. Hematopoiese. 5. Microambiente Tumoral. I. Rocha, Edroaldo Lummertz da. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Farmacologia. III. Título.

Ohanez Mamigonian

CellComm v2: inferência da dinâmica de redes de comunicação celular durante o desenvolvimento hematopoietico

O presente trabalho foi julgado adequado para obtenção do título de MESTRE e **APROVADO** em sua forma final pela banca examinadora.

Florianópolis, 29 de agosto de 2025

 Documento assinado digitalmente
Jose Eduardo da Silva Santos
Data: 22/09/2025 15:31:01-0300
CPF: ***.948.289-**
Verifique as assinaturas em <https://v.ufsc.br>

Coordenação do Programa de Pós-Graduação

Banca Examinadora:

 Documento assinado digitalmente
EDROALDO LUMMERTZ DA ROCHA
Data: 22/09/2025 12:56:44-0300
CPF: ***.282.360-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. Edroaldo Lummertz da Rocha, Dr.

Orientador

UFSC/CCB/MIP

 Documento assinado digitalmente
Andre Luiz Barbosa Báfica
Data: 19/09/2025 15:39:43-0300
CPF: ***.356.165-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. André Luiz Barbosa Báfica, Dr.

UFSC/CCB/MIP

 Documento assinado digitalmente
GABRIELA FLAVIA RODRIGUES LUIZ
Data: 19/09/2025 09:29:28-0300
CPF: ***.018.216-**
Verifique as assinaturas em <https://v.ufsc.br>

Gabriela Flavia Rodrigues-Luiz, Dr.^a

Cellertz Bio

Florianópolis, 2025

Dedico este trabalho à minha família, em especial à:
Meus pais, Sheila Vieira e Armen Mamigonian,
Minhas tias, Kátia Lúcia Vieira e Aracy Balabanian,
Minha avó, Walkyria Dias Vieira

Cada momento que passo, ou passei, com vocês me trouxe a onde estou,
e inspirado por vocês, espero ainda conseguir chegar muito longe...
fazendo o melhor possível ao longo do caminho...

... sempre carregando vocês comigo no coração...

AGRADECIMENTOS

Gostaria de agradecer ao meu orientador, Prof. Dr. Edroaldo Lummertz da Rocha, pela atenção e disposição nas nossas inúmeras reuniões. Pela confiança e apoio que me fizeram trilhar um longo caminho ao longo desses dois anos de mestrado. Pela indicação em participar de um evento internacional interessantíssimo que foi o Insight Lab-Research Camp da Roche, com apoio do Serrapilheira. Enfim, sou grato por tudo que vivenciei e sinto que estas experiências todas agregaram não só ao meu trabalho, mas a mim como pessoa. Fico muito feliz por ser parte do laboratório, e orientando do Prof. Edroaldo; anseio pelo que o futuro resguarda.

Nada do que fiz teria sido possível sem o apoio da minha família. Não tenho como agradecer o suficiente pelo apoio oferecido pela minha mãe e minha tia, pessoas muito carinhosas e especiais, que tenho o privilégio de ter na minha vida. Foram imprescindíveis também na minha escolha por trilhar o caminho da ciência, e buscar melhorar a condição humana. Duas mulheres muito dedicadas, trabalhadoras e amáveis, sei que orgulharam e orgulham minha avó, a quem ambas sentem muita saudade.

Quase todas as relações possuem altos e baixos. Se me perguntassem algo que me fez feliz nestes dois anos, eu prontamente responderia – a evolução da relação que tive com meu pai. Sempre fui alguém ansioso e estressado, e meus pais sempre me diziam: “se algo não tem solução, solucionado está”, e “a única coisa que não tem solução é a morte”. Pois bem, a pior coisa que podemos carregar é a amargura de não ter feito as pazes com alguém enquanto podíamos, sobre isso, sou infinitamente grato por ter meu pai hoje como um amigo a quem valorizo cada segundo que temos juntos. E espero que muitos segundos ainda nos restem.

Infelizmente não terei como agradecer aqui a todos que sou grato pelo carinho e afeto que tive nestes anos. Por muita sorte, tenho muitas pessoas queridas fazendo parte de minha vida, a todas elas sou eternamente grato pelo tempo, risadas e sorrisos.

Agradeço aos cidadãos brasileiros pelo investimento nas universidades públicas, e ao contribuinte americano pela minha bolsa de mestrado, oferecida via NIH.

Por fim, agradeço aos integrantes que compõem a presente banca, pelo tempo e conhecimentos. Sei que seus apontamentos contribuirão para o aperfeiçoamento deste e elaboração de futuros trabalhos.

RESUMO

Diversos foram os avanços no cenários de ciências ômicas nos últimos anos. Um *boom* no mundo de *single-cell RNA-sequencing* (scRNA-seq) foi observado tanto no número de trabalhos, quanto no de novos *software* sendo criados. No entanto, velhos hábitos não mudam, e o efeito fundador na área é intransponível. Hoje há dois frameworks para se trabalhar com tecnologias de células únicas, Seurat ou Scanpy, é basicamente uma escolha entre R ou Python, não há alternativas efetivas às suas hegemonias. A validação *in vitro* ou *in vivo* de achados *in silico* é uma peça chave para conquistar a atenção de especialistas, no entanto, por mais consistente e/ou extensa que seja, não garante a real adoção de *software* por parte da comunidade. É neste cenário que nascem os esforços do presente trabalho; visando maior adoção e expansão de suas capacidades, aqui é apresentada uma versão atualizada preliminar do pacote CellComm, um algoritmo para análise de comunicação celular, lançado em 2022, agora funcionando inteiramente dentro do *framework* do Seurat, e com *Non-Negative Matrix Factorization* (NMF) em seu arsenal estatístico. A partir de dados de scRNA-seq e *Spatial Transcriptomics* (ST), seu uso e potencial são demonstrados no contexto de câncer, transição endotélio-hematopoiética e maturação do microambiente da medula óssea.

Palavras-chave: Interações célula-célula; Bioinformática; Hematopoiese; Microambiente Tumoral

ABSTRACT

It's been a busy few years for omics science. The world of single-cell RNA sequencing (scRNA-seq) has seen a boom in both the number of papers and new software being developed. But old habits die hard, and the founder effect in this field is hard to overcome. Today, two frameworks dominate the landscape for working with single-cell technologies: Seurat and Scanpy. It is basically a choice between R and Python, with no real alternatives to their hegemony. In vitro and in vivo validation of in silico findings plays a key role in gaining the attention of experts. However, no matter how consistent and/or extensive such validation may be, it does not guarantee the actual adoption of given software by the community. It is against this backdrop that the efforts of the present work arise. Aiming at greater adoption and expansion of its capabilities, we present here a preliminary updated version of the CellComm package, an algorithm for cellular communication analysis launched in 2022, now running entirely within the Seurat framework and with Non-Negative Matrix Factorization (NMF) in its statistical arsenal. Drawing on scRNA-seq and Spatial Transcriptomics (ST) data, its use and potential are demonstrated in the context of cancer, endothelium-to-hematopoietic transition, and bone marrow microenvironment maturation.

Keywords: Cell–cell interaction; Bioinformatics; Hematopoiesis; Tumor microenvironment

Sumário

AGRADECIMENTOS.....	6
RESUMO.....	7
ABSTRACT.....	8
Lista de Figuras.....	10
Lista de Abreviações.....	12
INTRODUÇÃO.....	13
Non-Negative Matrix Factorization (NMF).....	16
Object Oriented Programming (OOP).....	17
Transição Endotélio-Hematopoiética (EHT) e amadurecimento da medula óssea.....	17
OBJETIVOS.....	18
Objetivo geral.....	18
Objetivos específicos.....	18
MATERIAL E MÉTODOS.....	18
1. Origem de dados.....	18
a. Dataset GSE144240 - TSKs.....	19
b. Dataset GSE160526 - Nascimento Células Tronco Hematopoiéticas - AGM.....	19
c. Dataset GSE178951 - Maturação Medula Óssea - Shannon's.....	19
2. Software usados.....	20
3. Especificações técnicas das máquinas utilizadas.....	21
4. Disponibilidade de Código.....	21
RESULTADOS.....	22
1. ANÁLISE DE TRAJETÓRIA DE DIFERENCIAÇÃO DE MACRÓFAGOS (COLABORAÇÃO).....	22
2. TRABALHANDO SOBRE O CÓDIGO DO CELLCOMM.....	23
FUNÇÕES CRIADAS OU MODIFICADAS.....	25
a. Funções modificadas.....	26
b. Funções criadas.....	30
3. APLICAÇÕES DO CELLCOMM V2 EM DATASETS PÚBLICOS.....	34
Dados (humanos) espaciais de TSKs.....	35
Dados murinos de scRNA do AGM.....	36
Dados murinos de CITE-seq da medula-óssea em maturação.....	43
DISCUSSÃO.....	47
1. DADOS DO AGM.....	47
Aplicando NMF para buscar ligantes-receptores específicos de populações EHT.....	47
2. DADOS DA MEDULA ÓSSEA.....	49
Semelhanças e incongruências com paper original.....	49
Explorando interações iniciais entre HSPCs e outros tipos celulares.....	51
Foco em fatores que melhor diferenciam os time-points.....	55
3. LIMITAÇÕES.....	63
Melhorias a serem feitas no fluxo de trabalho de comunicação espacial.....	63
Reprodutibilidade dos dados de TSKs.....	64
Reprodutibilidade dos dados de AGM.....	64

Reprodutibilidade nos dados da Shannon.....	65
Completude das capacidades de análise do CellComm.....	66
PERSPECTIVAS.....	66
Validação de resultados - EHT.....	67
Validação de resultados - Maturação do Microambiente da Medula Óssea.....	67
Aprofundando análises in silico.....	68
CONCLUSÃO.....	68
REFERÊNCIAS.....	70

Lista de Figuras

Figura 1 - Destrinchando ComputeInteractions().....	24
Figura 2 - Dados AGM - Figura 2C do artigo original.....	38
Figura 3 - Dados AGM - Tentativa de reprodução da Fig.2C original.....	38
Figura 4 - Dados AGM - NMF - Fator X Par-Celular (Rank 2).....	39
Figura 5 - Dados AGM - NMF - Fator X Par-Celular (Rank 8).....	40
Figura 6 - Dados AGM - Dotplot - Fator 1 (Rank 2).....	41
Figura 7 - Dados AGM - Dotplot - Fator 2 (Rank 2).....	42
Figura 8 - Dados AGM - Dotplot - Fator 8 (Rank 8).....	43
Figura 9 - Dados Shannon - Esquema elementos NMF.....	46
Figura 10 - Dados Shannon - Heatmaps - Ligantes-Receptores Originais.....	50
Figura 11 - Dados Shannon - Dotplot - Busca Pelos Ligantes-Receptores Originalmente Destacados.....	51
Figura 13 - Dados Shannon - Pannel Esquemático da Aplicação de NMF com 150 pares.....	56
Figura 12 - Dados Shannon - Dotplot - Interações Autócrinas.....	54
Figura 14 - Dados Shannon - GSEA - Ligantes-Receptores associados ao Fator 6 (Imaturo).....	59
Figura 15 - Dados Shannon - GSEA - Ligantes-Receptores associados ao Fator 1 (Intermediário).....	60
Figura 16 - Dados Shannon - GSEA - Ligantes-Receptores associados ao Fator 3 (Maduro).....	61
Figura 17 - Dados Shannon - Dotplot - Ligantes-Receptores associados aos Fatores 1, 3 e 6.....	63

Lista de Abreviações

ADT	Antibody Derived Tag
AGM	Aorta-Gônoda-Mesonefro
CAR, CARs	CXCL12-Abundant Reticular (CAR) mesenchymal stroma cells
CFU, CFUs	Colony-Forming Unit
CITE-seq	Cellular Indexing of Transcriptomes and Epitopes by Sequencing
E16.5	Dia embrionário 16 e meio
E18.5	Dia embrionário 18 e meio
ECM	matriz extracelular
EHT	Endothelial-to-Hematopoietic Transition
GEO	Gene Expression Omnibus
GSEA	Gene Set Enrichment Analysis (análise de enriquecimento de termos)
HE	Hemogenic Endothelium
hiPSC, hiPSCs	human induced Pluripotent Stem Cells
HSC, HSCs	Hematopoietic Stem Cells
HSPC, HSPCs	Hematopoietic Stem and Progenitor Cells
NMF	Non-Negative Matrix Factorization
OOP	Object Oriented Programming
P0	Dia do nascimento
PC, PCs	Principal Components
scRNA-seq	Sequenciamento do transcriptoma de células
TFs	Fatores de Transcrição
TSKs	Tumor Specific Keratinocyte
WISH	Whole-mount in situ hybridization
LT-HSC, LT-HSCs	Long-Term Hematopoietic Stem Cells
MPP2	Multipotent progenitors (MPPs) 2

INTRODUÇÃO

As tecnologias de sequenciamento de células únicas tiveram um imenso impacto no mundo científico, tornando-se cada vez mais democráticas com o passar dos anos, mas nem sempre foi assim. O título de prováveis “pais” do sequenciamento de RNA de células únicas (scRNA-seq) deve ser atribuído a Fuchou Tang e Catalin Barbacioru, que contribuíram igualmente no artigo de 2009 publicado na revista *Nature Methods*.⁽¹⁾ Neste primeiro trabalho efetuou-se o sequenciamento de 9 células no contexto de embriologia. No entanto, pouco se avançou na área por alguns anos, até a chegada do trabalho que é mais abrangentemente conhecido, e tido como o nascimento da área de scRNA-seq. Embora tecnicamente não seja o primeiro, *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells*⁽²⁾ foi o artigo que serviu de marco para a corrida do desenvolvimento de novas tecnologias capazes de gerar e analisar os dados de scRNA-seq. Publicado na *Nature* em 2013, com primeiras autorias de Alex K. Shalek e Rahul Satija, e coordenado por Aviv Regev, essa publicação sequenciou 18 células dendríticas, trouxe nomes que são expoentes até hoje, e marcou a largada da explosão de desenvolvimentos que ocorreram na última década na área.

Por um bom tempo a comunidade em geral não tinha acesso às tecnologias de sequenciamento de células únicas, mas já fazia uso de algumas das técnicas associadas, como *cell-sorting*, que somado a ensaios de expressão via *Microarray*, poderiam extrair informações úteis de populações celulares enriquecidas. Sendo este, justamente, o caso do CellNet,⁽³⁾ trabalho com coautoria do Prof. Edroaldo Lummertz da Rocha, lançado em 2014, algoritmo desenhado de forma a comparar o perfil de amostras enriquecidas de tipos celulares, de forma a ajudar na melhoria de protocolos de engenharia celular. Para efetivamente comparar amostras de teste com 20 tipos celulares especificamente escolhidos, em 10 contextos de perturbação, aos quais o algoritmo compararia semelhança transcricional, foram necessárias aproximadamente 10.000 amostras, manualmente coletadas, curadas e uniformemente processadas a partir dos dados brutos. Isto viria a ser muito mais simplificado de se obter com sequenciamento de células únicas, o que não era a realidade naquele momento. Mas veio a ser 5 anos depois, o que possibilitou a atualização do *software* na forma do Single Cell Net.⁽⁴⁾ O advento de scRNA-seq como parte do arcabouço de técnicas populares de laboratórios demorou um tanto a chegar; longo ainda seria o percurso da tecnologia de scRNA-seq rumo à sua (debatível) democratização.

Os primeiros trabalhos usando scRNA-seq envolviam a preparação da biblioteca de transcritos via Fluidigm, uma tecnologia de canais microfluídicos. O sequenciamento de

células que passam por esse protocolo era de custo extremamente alto, na ordem de 6-10 dólares por célula. Levando em consideração que era necessário sequenciar uma centena de células para, após controles de qualidade, chegar a 25-30 células devidamente analisáveis, não se apresentava como uma tecnologia acessível. Mas as tecnologias foram evoluindo aos poucos e alguns achados relevantes já surgiam,⁽⁵⁾ demonstrando o potencial oriundo desse novo tipo de dado, scRNA-seq. O próximo passo lógico era, portanto, aumentar o número de células únicas sequenciadas, e embora isso tenha demorado a acontecer em escala comercial, os laboratórios pioneiros da tecnologia usufruíram de uma vantagem considerável e duradoura. O primeiro trabalho do grupo da Aviv Regev, de 2013, gerou o sequenciamento de 18 células,⁽²⁾ 2 vezes mais que o obtido em 2009,⁽¹⁾ mas já em 2014 o grupo chegou ao marco de 1700 células,⁽⁶⁾ quase 100 vezes mais que no ano anterior. Eis que torna-se mais claro qual seria o próximo desafio: como extrair informações úteis desses dados.

O cenário estava lançado, as tecnologias de scRNA-seq avançariam rapidamente no número de células sequenciadas, era inevitável. Tanto protocolos de preparo de bibliotecas tornariam-se mais eficientes e mais baratos (proporcionalmente), quanto o sequenciamento em si haveria de baratear, considerando-se o número de bases sequenciadas por dólar. Mas o que fazer com esses dados? Uma matriz de expressão de scRNA-seq é montada por células (quantas forem preparadas) e por genes (quantos forem identificados como expressos). Logo, tais matrizes podem facilmente chegar a dimensões de 20.000x12.000, caso sejam interrogadas 12 mil células humanas com cobertura próxima ao total de genes que codificam proteínas existentes no genoma humano. O que fazer com essas informações? Extrair informações úteis desse tipo de dados se mostrou outra corrida. Novamente com ênfase no efeito fundador, de dentro do laboratório da Aviv, surge um expoente que domina o cenário até hoje, o *software* Seurat,⁽⁷⁾ de Rahul Satija, um dos primeiros *frameworks* para análises de scRNA-seq, na época (2015) com o provável primeiro trabalho envolvendo transcriptômica em um contexto espacial usando células únicas.

Como em qualquer período de surgimento de inovações, há alta incerteza e volatilidade. Com as expansões tecnológicas de sequenciamento de células únicas não foi diferente. Diversos laboratórios diferentes passaram a elaborar seus próprios conjuntos de códigos formando *frameworks*, na tentativa de extrair descobertas dos dados multidimensionais de scRNA-seq. Um dos mais relevantes foi o Monocle,⁽⁸⁾ que em 2014 lançou um dos primeiros e o mais consolidado algoritmo de predição de trajetória de diferenciação celular. Mesmo após duas atualizações, e se mantendo significativo até hoje no que tange à análise de trajetórias, o *framework* de processamento e análises básicas do

Monocle3⁽⁹⁾ sequer é usado pela maioria dos usuários finais, sendo efetuada a análise de trajetórias do Monocle, somente após a conversão de um objeto Seurat, ainda que o *software* do Satija Lab seja posterior ao Monocle original. Também no tópico de trajetórias, em 2018 foi publicado o primeiro *software* próprio, usado pelo nosso grupo, desenhado para trabalhar com scRNA-seq e usando um *framework* próprio, o CellRouter.⁽¹⁰⁾ Mas, no final, apenas dois *frameworks* dominam o nicho de scRNA-seq, o Seurat e o único outro que perdurou, justamente por não competir diretamente com o primeiro, o Scanpy,⁽¹¹⁾ sendo este um *framework* em Python, onde hoje exerce sua própria hegemonia, enquanto que o Seurat domina o ambiente de R. Mas, claro, embora não haja competição dentro da linguagem em si, ambos *frameworks* competem por usuários, usabilidade e ferramentas disponíveis acabam fazendo com que usuários priorizem o Seurat ou o Scanpy.

Na época do lançamento do CellRouter não era evidente a dominância que o Seurat exerceria na comunidade, então os futuros trabalhos do laboratório continuaram a ser construídos sobre as bases publicadas em 2018. Com o advento do CellRouter poderíamos interrogar quais fatores de transcrição (TFs) ajudariam no percurso de uma célula no caminho a se tornar outro tipo celular, mas essa trajetória ignorava essencialmente o entorno dessas células, sendo que é conhecida a influência do microambiente nos percursos de diferenciação. Eis que, durante o “*morning commute*”, a ideia vem ao Prof. Edroaldo, buscar mensurar efetivamente por quais eixos de interação as células influenciam umas às outras. Não há, até então, trabalhos que explorem esse terreno ainda, os esforços apenas começam. Até que em 2018 é lançado o primeiro artigo sobre comunicação celular, o CellPhone DB.⁽¹²⁾ A ideia deixa de ser inédita, mas a área é pouco explorada ainda, os esforços continuam. 2020 chega, e mais um trabalho passa a popular o ambiente de comunicação celular, o NicheNet.⁽¹³⁾ Até então a maior parte desses algoritmos se preocupou apenas com a validação teórica dos seus achados, e não com a experimentação de novos pares de interação celular, para efetivamente demonstrar a capacidade preditiva do algoritmo, como desejava-se fazer com o CellComm. Esse esmero custaria caro, pois seria mais difícil justificar a publicação em revistas de prestígio sem caráter inédito. Foi apenas em 2022, após extensiva validação *in vitro* e *in vivo* de 12 genes, em três diferentes espécies e de novos achados biológicos – a maior validação experimental de um algoritmo de comunicação celular realizada até então – que a publicação do CellComm se concretizou.⁽¹⁴⁾ As análises de comunicação celular deixam então de mensurar apenas o fenômeno superficial das interações, agora adentrando o citoplasma rumo ao núcleo, elucidando quais TFs seriam os prováveis alvos da cascata de sinalização celular, eis o fator inédito da publicação do CellComm na revista Nature Cell Biology.

Apesar dos méritos do CellComm, como a extensa validação experimental e a inovação na inferência da cascata de sinalização intracelular, pouco se sabe sobre o real uso do *software* pela comunidade. Houve comentários e marcações no GitHub do projeto,⁽¹⁵⁾ o que indica engajamento da comunidade, mas não há métricas concretas a respeito. Ademais, se formos considerar outros pacotes, como é o caso com o Monocle3, há menor adesão quando a comunidade encontra *frameworks* distintos dos estabelecidos (Seurat e Scanpy). Havendo, no caso do Monocle, esforços da própria equipe do Seurat em criar métodos de conversão de objeto para facilitar o uso por usuários finais, há de se ressaltar o efeito fundador do Monocle no quesito trajetórias. Esses diversos pontos levantados, somados ao reconhecimento internacional em encontros de pesquisa, levantaram uma demanda nova ao CellComm: tornar-se compatível aos *frameworks* estabelecidos.

Não podendo esperar esforços de terceiros, e reconhecendo o problema intrínseco de usar *frameworks* alternativos, o presente trabalho buscou então tomar o desafio de adaptar os códigos existentes do pacote CellComm, e torná-lo compatível com o objeto Seurat, *framework* hegemônico escrito na linguagem de programação R. Ademais, buscou-se incrementá-lo com novas capacidades associadas para inferir a dinâmica de redes de comunicações intercelulares por meio de *Non-Negative Matrix Factorization* (NMF), uma técnica matemática para identificação de padrões em matrizes. Tal reconstrução de processos dinâmicos a partir dos dados de scRNA-se permanece um desafio para a maioria dos algoritmos atuais de inferência de comunicação celular. Lacuna esta que este trabalho pretende preencher. Através dessas melhorias o presente trabalho buscou: a) validar a nova versão dos códigos; e b) explorar dois *datasets*^(14,16) à procura de novos achados nos cenários de transição endotélio-hematopoética (EHT) e maturação do microambiente da medula óssea. Além disso, durante o percurso, uma nova versão do *software* aqui desenvolvido foi utilizado em uma colaboração científica atualmente publicada no bioRxiv.

Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization (NMF) é uma técnica matemática do tipo *machine learning* que tem como objetivo a identificação de padrões em matrizes de grandes dimensões. A entrada de dados é uma matriz numérica $V_{(m,n)}$, que não pode possuir valores negativos, e a saída (dentre outros subprodutos) são duas matrizes ($W_{(m,k)}$ e $H_{(k,n)}$) que, quando multiplicadas, devem resultar em uma matriz $A_{(m,n)}$, semelhante à matriz $V_{(m,n)}$ original.

Ao aplicar essa metodologia, a ideia é encontrar padrões na matriz $V_{(m,n)}$ que seriam agrupados entre as matrizes W e H , através da dimensão ' k ', agrupando as informações das linhas ' m ' e colunas ' n ' aos fatores da dimensão ' k '. Com isso, podemos selecionar e investigar as informações agrupadas pelos fatores de forma a encontrar padrões relevantes ao contexto biológico aplicado.

Object Oriented Programming (OOP)

Object Oriented Programming (OOP) é um conceito da ciência da computação que diferencia basicamente duas formas de criar funções computacionais: uma forma “focada em ações” (cujos produtos são genericamente chamados de “funções”); e uma forma “focada em estado” (cujos produtos são genericamente chamados de “métodos” ou “classes”, para R ou Python, respectivamente).

O trabalho com ciências ômicas de alto número de componentes celulares sendo medidos, seja scRNA-seq ou ST, acabam compondo matrizes de expressão de genes com dimensões consideráveis. Como esses dados já são grandes por natureza, a forma mais eficiente de trabalhar sobre esses dados é os mantendo em um objeto que será chamado a cada “função” de um *pipeline* de análises. Assim, os *frameworks* possuem objetos que são o primeiro componente de *input* para parametrização das análises. O *framework* Seurat possui objetos do tipo “Seurat”, CellRouter/CellComm usa objetos “CellRouter”, e Scanpy usa “anndata”/“mudata”. Cada framework possui o seu formato, o que os torna estruturalmente incompatíveis entre si, e uma “função” que trabalhe sobre um objeto não é normalmente capaz de lidar com outros formatos.

Transição Endotélio-Hematopoiética (EHT) e amadurecimento da medula óssea

As células tronco hematopoiéticas (HSCs) são responsáveis por formar todas as células do sangue, isso inclui o Sistema Imune, hemácias e plaquetas. As HSCs que compõem a medula óssea amarela, tem origem no endotélio do tecido da aorta-gônada-mesonefro (AGM), onde algumas células que compõem o endotélio hemogênico (células raras ECs/T1preHSCs) passarão pelo processo de transição endotélio-hematopoiética (EHT) (por volta do período E10.5, décimo dia e meio pós-concepção), dando origem às HSC que migrarão rumo à medula, e a colonizarão aproximadamente em E15.5.

As HSCs, junto com suas células levemente mais diferenciadas, compondo o conjunto de células “HSPCs” (*Hematopoietic Stem and Progenitor Cell*), mesmo estando na medula

óssea, não possuirão as características essenciais para efetuar a hematopoiese madura (autorrenovação e diferenciação de todas as linhagens especializadas) até muito próximo do período de nascimento. Esse fenômeno é tido como o período de amadurecimento que as HSPCs precisam para passar de um perfil imaturo para um maduro (hematopoieticamente competente).

OBJETIVOS

Objetivo geral

Expandir o CellComm para reconstruir a dinâmica de redes de comunicações intercelulares.

Objetivos específicos

1. Simplificar o *pipeline* de funções atualmente implementadas no CellComm;
2. Adaptar as funções para funcionamento através do *framework* Seurat;
3. Criar funções para análise de NMF;
4. Demonstrar uso em *dataset* de transcriptômica espacial;
5. Demonstrar uso em *dataset* de scRNA-seq;

MATERIAL E MÉTODOS

Para cumprir com os objetivos do trabalho, foram utilizados dados públicos de sequenciamento de transcriptoma de células únicas (scRNA-seq) e transcriptoma espacial (ST). Foram utilizados alguns *software* para escrever e executar, ou até mesmo facilitar o trabalho com *scripts*. Também estão aqui descritas as especificações técnicas do equipamento computacional utilizado. Detalhes acerca das funções criadas ou alteradas serão descritos no final desta seção, aplicações destas funções encontram-se na seção de resultados.

1. Origem de dados

Foram utilizados dados de scRNA-seq e ST. Os dados passaram por algumas etapas de limpeza e padronização, seguindo padrões cabíveis a cada tipo de dado, seja de células únicas ou espaciais. Os códigos de acesso aos *datasets* são os seguintes: [GSE160526](#),⁽¹⁴⁾ [GSE144240](#)⁽¹⁷⁾ e [GSE178951](#)⁽¹⁶⁾.

Para cada um dos *datasets* usados, os dados já se encontravam minimamente processados, significando que foram baixadas essencialmente tabelas de contagem (genes X células; genes X *spots*), e potencialmente metadados acerca do que os cientistas originalmente já haviam identificado em seus dados (anotação de tipo celular, por exemplo), quando disponíveis. Não foram usados dados brutos de sequenciamento em nenhuma parte deste trabalho (`.FASTA`, `.fastq`, etc).

a. Dataset GSE144240 - TSKs

Este *dataset* traz dados de humanos, os dados são de transcriptômica espacial, e o contexto biológico é de câncer (*Squamous Cell Carcinoma* - Carcinoma de células escamosas da pele). O conjunto de dados depositados consiste em várias lâminas de ST, a amostra usada foi a primeira amostra do paciente 4, cujas informações encontram-se mais detalhadamente descritas na página de acesso do GEO, embora o acesso aos dados em si tenham sido um pouco mais complicados, o que será tratado na seção “Limitações” da Discussão.

Este *dataset* será referenciado como *dataset*/dados de/das “TSKs”, em relação à população celular identificada pelos autores originais como “*Tumor-specific keratinocytes*”.

b. Dataset GSE160526 - Nascimento Células Tronco Hematopoiéticas - AGM

Este *dataset* contém dados de scRNA-seq de camundongo, referentes ao nascimento de células tronco hematopoiéticas (HSCs), que ocorre no tecido Aorta-Gônada-Mesonefro (AGM), próximo do período E10.25.

Alguns detalhes específicos dos dados: a linhagem de camundongos usada foi C57Bl/6; o time-point de coleta foi E10.25; a coleta foi dividida em dois “*batches*”, um com as células totais do AGM, e outra que buscou-se enriquecer (via *sorting*) para as populações raras de células (ECs/T1preHSCs); as bibliotecas de *single-cell* foram preparadas seguindo protocolo da 10x Genomics (10x scRNA-seq v2 protocol) e sequenciadas via Illumina (MiniSeq e NovaSeq).

Este *dataset* será referenciado como *dataset*/dados “AGM”, em relação ao tecido de origem ser a Aorta-Gônada-Mesonefro.

c. Dataset GSE178951 - Maturação Medula Óssea - Shannon’s

Este *dataset* traz dados de camundongos, os dados são de scRNA-seq e *Antibody Derived Tag* (ADT), união conhecida como CITE-Seq (*Cellular Indexing of Transcriptomes*

and Epitopes by Sequencing). O contexto biológico é a maturação do microambiente da medula óssea, passando a melhor aportar as células tronco hematopoiéticas (HSCs) e seus progenitores intimamente relacionados (HSPCs), que ocorre entre o período E16.5 e P0.

Alguns detalhes específicos dos dados: a linhagem de camundongos usada foi C57BL/6J; os *time-points* de coleta foram E16.5, E18.5, P0 e adulto; a coleta foi dividida em dois “*batches*”, um referente ao compartimento hematopoiético da medula óssea, e o outro referente às células estromais do mesmo tecido. Buscou-se enriquecer (via *sorting*) cada compartimento através dos seguintes marcadores, Lineage- e Kit+, para as células hematopoiéticas, e CD45- e Ter119-, para as estromais; as bibliotecas de *single-cell* foram preparadas seguindo protocolo da 10x Genomics (Chromium single cell 3' reagent v3.1 protocol) e sequenciadas via Illumina (NovaSeq).

Este *dataset* será referenciado como *dataset*/dados: “da Shannon”, em relação à última autora e colaboradora do laboratório; “BM” ou “MO”, em relação ao tecido de origem ser a medula óssea (*bone-marrow*).

2. Software usados

Para a realização do trabalho foram gerados *scripts* em linguagem R (v. 4.3.0).⁽¹⁸⁾ Para os trabalhos envolvendo o pacote de análises de comunicação celular desenvolvido pelo laboratório, FUSCA/CellComm,⁽¹⁴⁾ foram utilizados os códigos da última versão disponível até o momento FUSCA 1.3.⁽¹⁵⁾ Para editar e manipular os códigos foi usado o *software* VSCodium,⁽¹⁹⁾ com pacote R instalado. Já para testar e aplicar as funções foi empregado o RStudio Web,⁽²⁰⁾ o qual o laboratório tem acesso controlado a um servidor próprio, carinhosamente apelidado de Ada.

Frequentemente na seção que tratará das funções criadas ou modificadas, se tratará do termo “método” em oposição ao termo “função”. Isso se dá porque dentro do R, esse jargão separa o que são na prática funções “focadas em ação” (funções clássicas) de “focadas em estado” (“métodos”, funções orientadas a objeto). Então cada vez que uma função for dita como “transformada em método”, significa que foi adaptada para funcionar orientada a objeto (OOP).

Para a aplicação das análises e tutoriais gerados, foram utilizados os seguintes pacotes primordiais: Seurat (v. 5.3)⁽⁷⁾ e dplyr (v. 1.1.4),⁽²¹⁾ responsáveis por funções gerais de manipulação de dados, e constituição de um *framework* padrão para análises de bioinformática envolvendo scRNA-seq e ST.

A título de curiosidade, e possível dica para quem venha a ler a presente dissertação, como trabalho em múltiplas máquinas (*desktop* e *notebook*) e em diferentes locais, utilizo alguns *software* que me auxiliam imensuravelmente: Obsidian (v 1.8.10),⁽²²⁾ para anotações e organização pessoal; OneDrive para armazenamento de arquivos não local; e por último, mas talvez mais importante, InSync (v 3.9.6.60027),⁽²³⁾ para sincronização entre servidores de nuvem e múltiplos computadores, viabilizando uma integração de fluxo de trabalho bem fluída entre todos os meus ambientes de trabalho e respectivos equipamentos. Mesmo com o uso desses mecanismos para ter arquivos salvos localmente e na nuvem, indicaria, pessoalmente, que sejam feitos *backups* localmente também, seja manualmente ou usando *software* dedicado, seja apenas uma cópia em outro local do disco ou em um disco dedicado (externo). Há valor em algumas redundâncias.

3. Especificações técnicas das máquinas utilizadas

Nesta seção, para fins de reprodutibilidade computacional, são descritos alguns detalhes técnicos do computador principal que foi usado para a elaboração do presente trabalho. Vale ressaltar que para gerar os tutoriais e resultados propriamente do trabalho, foi usado um dos servidores próprios do laboratório, a Ada. Embora fosse possível rodar os *scripts* em meu computador pessoal, isso foi evitado por questões de inconveniência e desvantagens de segurança. Mesmo assim, as especificações aqui descritas devem ser suficientes para rodar o *software* CellComm v2 *in loco*, para a maioria dos casos.

Caso venha a ser do interesse de um leitor saber quais requisitos computacionais são necessários para rodar análises de comunicação celular, ou outros fluxos de análise, uma aproximação mais ideal dependeria muito de saber com até quantas células deseja-se que o equipamento seja capaz de processar de uma vez. O fator limitador, no geral, acaba sendo a memória RAM, embora para alguns *pipelines* que usam inteligência artificial a VRAM virá a ser limitante.

Especificações de *hardware*: processador 12^a Geração Intel® Core™ i7-12700K, 20 núcleos de processamento; 32 Gb de memória RAM; 12 Gb de memória VRAM; 2 Tb de armazenamento, arquivos usados totalizando em torno de 7,2 Gb.

Especificações de *software*: Linux Ubuntu 22.04.02 LTS como Sistema Operacional.

4. Disponibilidade de Código

O código que compõe o presente trabalho pode ser consultado no repositório temporário do GitHub ([oham9/fusca2](#)), que compila o progresso da adaptação de funções do pacote CellComm, e incremento de novas capacidades, rumo à versão 2.

Tutoriais para visualização, na prática, do novo fluxo de trabalho de análises, podem ser encontrados através dos seguintes *links* externos: Análise [Simples AGM](#) (scRNA-seq), Análise [Espacial TSKs](#) (ST) e Análise [Multicondicional Hematopoiese](#) (scRNA-seq).

RESULTADOS

1. ANÁLISE DE TRAJETÓRIA DE DIFERENCIAÇÃO DE MACRÓFAGOS (COLABORAÇÃO)

Começando em meados de agosto de 2024, passei a trabalhar em um projeto envolvendo análises de trajetória de diferenciação de macrófagos em tecidos adultos e fetais humanos. Os dados partiram de um colaborador internacional do laboratório, Prof. R. Grant Rowe, MD. PhD, Boston Children's Hospital. Meu contato foi com uma pós-doutoranda do Grant Rowe Lab, Frenis Katie, PhD. O trabalho tem como ponto central o percurso diferencial de macrófagos fetais em relação ao caminho de diferenciação convencional de macrófagos. Atualmente o trabalho está em período de revisão, mas uma versão inicial dele pode ser acessada no BioRxiv.⁽²⁵⁾

Como meu trabalho originalmente envolveria redesenhar os algoritmos de inferência de trajetórias de diferenciação do laboratório em um contexto que envolvesse informações de interação ligante-receptor, a participação nesta colaboração encaixava-se muito bem como um primeiro contato com as análises de trajetória em *single-cell*.

Ao trabalhar com os dados da Katie, entrei em contato direto com o fluxo de trabalho que envolvia transformar um objeto Seurat, usado para diversos aspectos do processamento inicial dos dados, em um objeto CellRouter, usado para as análises de trajetórias. Esse contato foi de grande importância para verificar as dificuldades presentes nessa conversão de objetos.

Além da questão da conversão, adquiri experiência com as análises de trajetória, que seriam importantíssimas para o desenvolvimento do meu próprio projeto inicial. Consegui compreender a lógica por trás de diferentes pontos das análises de trajetória: como a construção de grafos onde as células seriam nodos, e a “semelhança” entre células trariam valores às arestas, compondo uma rede que envolve todas as células de um *dataset*. E portanto, uma trajetória seria composta pelo caminho de “menor resistência” entre células indiferenciadas (iniciais) e células diferenciadas alvo, usando algoritmos de *Max Flow*

(Ford-Fulkerson) – Descrevendo aqui superficialmente o funcionamento do *software* CellRouter publicado em 2018.⁽¹⁰⁾ Esse conhecimento me ajudou a entender os desafios à frente no processo de criar os algoritmos que comporiam meu projeto original.

Após construir uma noção mais crítica e palpável sobre os desafios adiante, percebi que, ao invés de construir o algoritmo novo sobre as bases do *framework* do CellRouter original, construí-lo sobre bases mais duradouras e acessíveis à comunidade, trariam muito mais valor ao novo trabalho, e aos *software* prévios do laboratório, como o próprio CellComm e CellRouter. Desta maneira, o norte do trabalho se deslocou na direção de adaptar o que há de funções pivotais à análise de comunicação celular, e, idealmente, implementar novas capacidades.

Tendo essa contextualização, a colaboração aqui citada foi relevantíssima no meu desenvolvimento, onde aprendi a trabalhar sendo parte de uma equipe, na aquisição de novos conhecimentos, e também no realinhamento das prioridades frente ao meu projeto. Os resultados referentes à adaptação do *software* CellComm podem ser observados nas próximas duas sessões, enquanto que a expansão das capacidades do CellComm, através da implementação de NMF, pode ser atestada nas últimas duas seções dos Resultados do presente trabalho, “Dados murinos de scRNA do AGM” e “Dados murinos de CITE-seq da medula-óssea em maturação”.

2. TRABALHANDO SOBRE O CÓDIGO DO CELLCOMM

Nesta segunda seção dos resultados, irei descrever brevemente as estratégias usadas no desenvolvimento das adaptações de código das funções descritas na quinta, e última, porção dos Materiais e Métodos, “Funções criadas ou modificadas”.

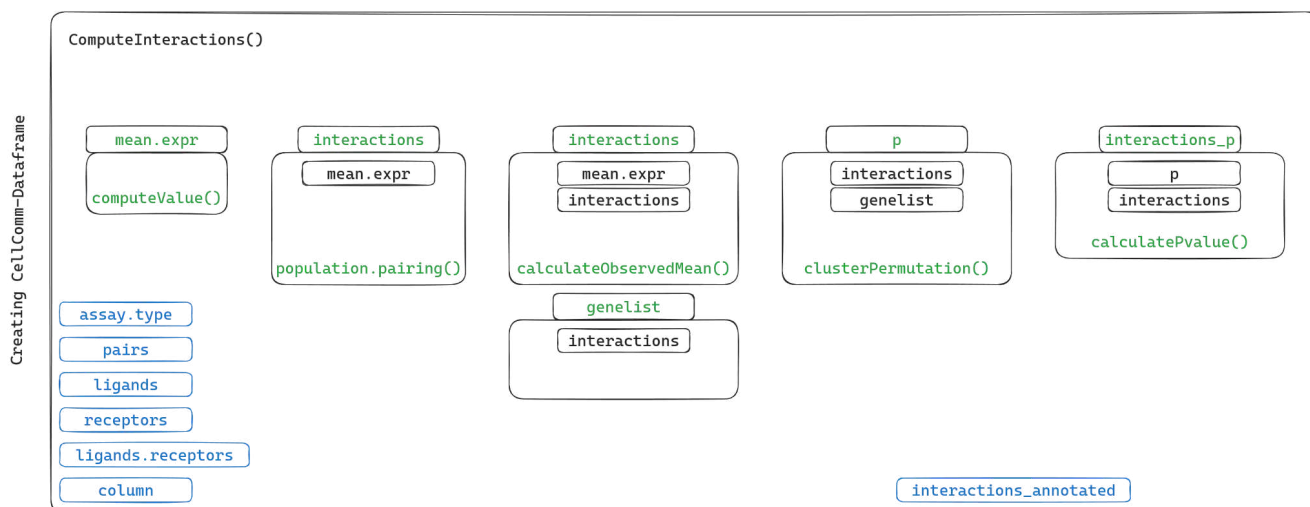
Os pacotes FUSCA, responsável pelo CellComm v1, e Seurat são semelhantes em uma série de aspectos, mas a estrutura de dados possui ligeiras diferenças. Uma questão que une ambos, e é um aspecto comum a algumas áreas da programação, independente da linguagem aplicada, é a natureza de Programação Orientada a Objetos (OOP). O que significa que as funções de ambos os pacotes comumente tem como primeiro *input* o próprio objeto de trabalho, no caso um objeto de classe “CellRouter” ou “Seurat”, seguido de uma série de parâmetros que controlam a execução da referida função. Uma das maravilhas da OOP, é a versatilidade em lidar com diferentes classes de objetos como primeiro *input*, viabilizando diferentes sequências de códigos a serem executadas conforme o tipo do objeto. Isso significa que ao invés de reescrever ou adaptar os códigos antigos do CellComm, que fazem parte do

pacote FUSCA, pude adaptá-los e manter a funcionalidade antiga, caso o objeto a que se aplique a função em questão seja de classe “CellRouter”. Isso permitiu preservar as funcionalidades antigas como uma forma de “legado resquicial”, bem como a possibilidade de aplicar as novas funcionalidades a objetos CellRouter associados a projetos antigos, uma espécie de “retrocompatibilidade”.

Para viabilizar a reconstrução das funções através dessa óptica de preservação de funcionalidades, foi especialmente relevante o conhecimento e mapeamento de como cada bloco de código interage para executar as análises de comunicação celular efetuadas pelo CellComm. Para atingir esse objetivo, foi necessário então mergulhar no código original: conhecer todas as funções que compõem o pacote; inspecionar cada uma das funções associadas à análise de comunicação, anotar quais funções eram chamadas internamente, quais *inputs* e *outputs*; como as funções se relacionam em cascata rumo aos resultados finais de comunicação celular e sinalização. Parte desse processo “cartográfico” é visualizável em um esboço esquemático na Figura 1, ao trabalhar sobre o que viria a ser uma versão inicial da função `ComputeInteractions()`, que compõe o cerne do pacote CellComm, e pode ser visualizado na Figura 1. Este extenso mapeamento foi crucial para a entrega de uma versão renovada funcional, a que este trabalho orgulhosamente foi capaz de atingir até certo ponto, como será descrito e demonstrado ao longo do restante do mesmo.

Figura 1 - Destrinchando `ComputeInteractions()`

Desenho esquemático destrinchando quais as subfunções (em sub-caixas com nomes em verde, seus *inputs* e *outputs* na porção inferior de cada caixa) deveriam compor a função `ComputeInteractions()`, bem como seus parâmetros (em azul, no campo inferior esquerdo) e objetivo (fora da caixa, no canto externo esquerdo, “criação do `cellcomm_dataframe`”).



Fonte: Figura criada pelo autor (2025)

FUNÇÕES CRIADAS OU MODIFICADAS

Nesta seção, para fins de documentação, são citadas e descritas brevemente as funções que sofreram alterações entre a última versão do pacote CellComm,⁽¹⁵⁾ ou que passarão a existir nesta nova versão.

O pacote CellComm, como apresentado em 2022,^(14,15) possui dois módulos principais empregados de forma sequencial, não exatamente declarados: um deles, responsável pela análise inicial de ligantes e receptores dos *clusters* presentes, aqui denominado como “superficial” ou “*ligand-receptor*”; o seguinte, mais complexo, capta o que foi obtido do primeiro como *input*, e computa um caminho nas células receptoras da comunicação celular, conectando os receptores a fatores de transcrição através de cascatas de sinalização, construídas por uma rede de interações proteína-proteína, fator inédito no momento da publicação. As adaptações e trabalhos presentes aqui apenas envolveram-se com o primeiro módulo (“*ligand-receptor*”). A adaptação do módulo mais profundo, abrangendo a sinalização para fatores de transcrição (TFs), será feita num momento futuro, não compondo o presente trabalho.

Funções foram modificadas para que fossem compatíveis de rodar usando objetos Seurat v5, ao invés de serem apenas capazes de rodar sobre objetos CellRouter. A retrocompatibilidade com *workflows* antigos, que possuam objetos CellRouter, foi garantida, viabilizada através de *Object Oriented Programming* (OOP), formas de criar funções que respondem de forma específica à natureza do objeto usado como *input*. Em torno de 12 funções foram modificadas neste trabalho.

Funções foram criadas para preencher lacunas no *workflow* original, facilitar e simplificar o uso para novos usuários (ou inexperientes no nosso pacote em específico), ou ainda, para apresentar capacidades totalmente novas e potencialmente inéditas ao ambiente Seurat (como a análise de NMF sobre interações celulares, apresentados inicialmente na porção “Dados murinos de scRNA do AGM”, Figura 5, dos Resultados). Em torno de 15 funções foram criadas neste trabalho.

Os resultados apresentados ao longo deste trabalho usaram apenas o *framework* do Seurat para fazer as análises, tudo foi obtido usando objetos de tipo “Seurat Object”. Não foram executadas análises comparativas para saber exatamente quais os ganhos em tempo e capacidade computacional de usar apenas o CellRouter para usar o Seurat. Porém o dentro das experiências descritas pelo Edroaldo Lummertz da Rocha, o CellRouter já foi usado por ele com algo próximo de 200 mil células, enquanto que o Seurat já foi usado *in-house* (no nosso

laboratório) em análises de >400 mil células (algumas demonstrações *online* trazem o Seurat para o patamar de >1 milhão de células, mas usando computação em disco). O maior ganho ao se mudar para o framework Seurat não é a questão da escalabilidade, embora ela seja útil, mas principalmente a experiência e usabilidade para o usuário final (integrando com seus próprios *pipelines* de análises com menos atrito).

a. Funções modificadas

Aqui encontram-se listadas as principais funções que foram modificadas em comparação com a versão anterior do pacote CellComm; funções que não estão listadas provavelmente foram omitidas por não estarem prontas para o uso no novo fluxo de trabalho, referente ao uso prioritário de objetos Seurat. Entretanto, isto não significa que todas as funções omitidas aqui não funcionam no novo fluxo de trabalho, algumas foram omitidas por não haver modificações relevantes a se relatar, podendo essas mesmas funções serem meramente auxiliares a outras funções aqui descritas. Para uma consulta mais próxima da realidade quanto às funções que compõem a atualização do pacote, é indicado que seja acessada a página referente ao trabalho no GitHub, cujo link pode ser encontrado na seção “Disponibilidade de Código”.

Como apenas um resumo de cada função será dado, para uma melhor compreensão de como o trabalho foi realizado, é necessário esclarecer sobre o processo destas modificações. Estas consistem primariamente em duplicar o código de uma função, e adaptar cada *input* oriundo do objeto para que a mesma informação seja extraída (seja expressão de genes, nomes de *clusters*, etc), de forma a possibilitar ao restante do código funcionar o mais coerentemente possível, sem mais alterações. Em muitos casos, além dessas duplicações, foi incrementado o código para que este ficasse mais bem documentado, limpo, ou até mesmo funcionando melhor através de formas que automatizam a inferência de parâmetros omitidos pelos usuários (tornando o uso do pacote mais “*user-friendly*”).

ComputeValue()

Função responsável pelo cálculo de expressão de todos os genes (incluindo ligantes e receptores) dentro de todos os clusters.

Foram feitas modificações transformando a função em um “método”, fazendo com que esta funcione de forma “orientada a objeto” (OOP), seja recebendo um objeto CellRouter ou

Seurat como *input* principal. Parâmetros podem ter sido modificados de forma a acrescentar funcionalidades novas ou permitir automação de processos na omissão de *input* do usuário.

Population.pairing()

Uma das funções diretamente responsáveis pelo cálculo de coexpressão entre ligante e receptor dentro dos clusters que enviam (*sender*) e os que recebem (*receiver*) o sinal de comunicação celular.

Alterações feitas foram apenas para padronizar parâmetros e/ou organizar o código.

CalculateObservedMean()

Uma das funções diretamente responsáveis pelo cálculo de coexpressão entre ligante e receptor dentro dos clusters que enviam (*sender*) e os que recebem (*receiver*) o sinal de comunicação celular.

Alterações feitas foram apenas para padronizar parâmetros e/ou organizar o código.

ClusterPermutation()

Função responsável pelo cálculo de permutação estatística de todos os genes (incluindo ligantes e receptores) dentro de todos os clusters.

Foram feitas modificações transformando a função em um “método”, fazendo com que esta funcione de forma “orientada a objeto” (OOP), seja recebendo um objeto CellRouter ou Seurat como *input* principal. Parâmetros foram modificados de forma a acrescentar funcionalidades novas ou permitir automação de processos na omissão de *input* do usuário. Também foram corrigidos problemas associados à paralelização dos cálculos.

calculateCentroids()

Função responsável pelo cálculo da posição espacial de *clusters* (ou *subclusters*) baseado nas coordenadas espaciais do conjunto de *spots* que compõem um *cluster/subcluster*.

Foram feitas modificações transformando a função propriamente em um “método”, fazendo com que esta funcione tanto com objetos CellRouter quanto Seurat. Parâmetros foram modificados de forma a acrescentar funcionalidades novas ou permitir automação de processos na omissão de *input* do usuário.

calculateDistanceMatrix()

Função responsável pelo cálculo da distância entre *clusters* (ou *subclusters*) baseado na posição de seus centroides (calculado previamente pela função `calculateCentroids()`), servindo como métrica posterior para selecionar *clusters/subclusters* mais prováveis de estarem se comunicando, considerando um cenário espacial.

Foram feitas modificações transformando a função propriamente em um “método”, fazendo com que esta funcione tanto com objetos CellRouter quanto Seurat. Poucas alterações foram feitas, essencialmente no âmbito de tornar o código mais legível.

calculateLigandReceptor()

Não foram feitas alterações, mas a função já foi verificada para funcionar corretamente na transição entre CellComm v1.3 para v2. Funções não adaptadas ou verificadas ficarão separadas aguardando inspeção para integrar a nova versão.

predictCellInteractions() → PlotPredictedCellInteractions

Função responsável pela visualização do escore de comunicação potencial entre *clusters* (ou *subclusters*). Basicamente é calculado um *heatmap* que ajuda a identificar quais pares de *clusters* têm maior número de interações potenciais (número relativo de combinações entre genes correspondentes a ligantes em um *cluster* e receptores em outro). É aplicável tanto para dados de scRNA-seq quanto ST; no último caso é considerada a proximidade relativa entre *clusters/subclusters* para calcular quais pares possuem proximidade relevante. No caso espacial também é computada a multiplicação entre o potencial de pares ligantes-receptores com o potencial de proximidade espacial, resultando em uma matriz de escore misto.

Foram feitas modificações transformando a função propriamente em um “método”, fazendo com que esta funcione tanto com objetos CellRouter quanto Seurat. O código foi modificado de forma a tornar erros mais compreensíveis ao usuário final, além de permitir automação de processos na omissão de input do usuário em alguns parâmetros.

PlotPairDotplot()

Função responsável por produzir o principal *plot* de interações de comunicação a nível superficial (ligantes-receptores).

Foram feitas modificações transformando a função propriamente em um “método”, fazendo com que esta funcione tanto com objetos CellRouter quanto Seurat, ao invés de com o output de outra função (um grafo), o que tornava o fluxo de trabalho mais complexo. A função também foi pareada com a recém criada `FilterCellComm_Dataframe()`, de forma a expandir o controle do usuário sobre o que deve ser selecionado para visualização das comunicações. Estas melhorias tornam a função `PlotPairDotplot()` mais dinâmica e *user-friendly*. Outras melhorias visuais estão sendo testadas, mas não estão prontas para lançamento ainda.

FindSubclusters()

Função responsável por delimitar subdivisões espaciais dentro de *clusters* transcricionais (num determinado corte histológico sequenciado).

Foram feitas modificações transformando a função propriamente em um “método”, fazendo com que esta funcione tanto com objetos CellRouter quanto Seurat. Parâmetros foram modificados de forma a acrescentar funcionalidades novas ou permitir automação de processos na omissão de *input* do usuário. O código também foi tornado ligeiramente mais limpo.

ComputeValueSubclusters()

Função essencialmente igual à `ComputeValue()`, mas dedicada apenas à computação de *subclusters* espaciais.

Foram feitas modificações transformando a função propriamente em um “método”, fazendo com que esta funcione tanto com objetos CellRouter quanto Seurat. Parâmetros foram modificados de forma a acrescentar funcionalidades novas ou permitir automação de processos na omissão de *input* do usuário. O código também foi tornado ligeiramente mais limpo.

ClusterPermutationSubcluster()

Função essencialmente igual à `ClusterPermutation()`, mas dedicada apenas à computação de *subclusters* espaciais.

Foram feitas modificações transformando a função em um “método”, fazendo com que esta funcione de forma “orientada a objeto” (OOP), seja recebendo um objeto CellRouter ou

Seurat como *input* principal. Parâmetros foram modificados de forma a acrescentar funcionalidades novas ou permitir automação de processos na omissão de *input* do usuário. Também foram corrigidos problemas associados a paralelização dos cálculos. A forma de calcular a permutação teve de ser mudada ligeiramente pois o código anterior estava apresentando erros.

b. Funções criadas

Aqui encontram-se listadas as principais funções que não existiam na versão anterior do pacote CellComm, funções menores de função auxiliar podem ter sido omitidas para dar relevância às que realmente impactarão o fluxo de trabalho de forma positiva. Apenas um resumo de cada função é dado, mas a título de reconhecimento de trabalho: há ~15 funções nesta seção, a mais simples delas é (por acaso) a última, contendo de 30-60 linhas de código. A mais complexa é a primeira, `ComputeInteractions()`, com ~630 linhas de código.

ComputeInteractions()

Esta é uma função do tipo “agregadora”, que junta várias funções antigas de forma a diminuir o número de funções que o usuário final tem de rodar para efetuar a análise, simplificando o fluxo de trabalho, o que é relevante para atrair mais usuários ao pacote.

Uma versão simplificada desta função já existia, mas ainda muito rudimentar e não verdadeiramente funcional, mas pode-se dizer que foi o ponto inicial para o trabalho como um todo. Agora esta é a função mais importante do pacote, efetuando as análises de comunicação ao nível superficial (ligante-receptor) de forma eficiente, sendo usada tanto para dados de scRNA-seq, quanto para ST (o qual a própria função faz a identificação e ajustes, processo que antes era manual). A função também é usada no âmbito de análise completa de *clusters*, *subclusters* ou *subsetting* condicional, capacidades antes não existentes. A função conta com 15 parâmetros diferentes, contando o objeto em si. Para facilitar seu uso, muitos desses parâmetros possuem valores padrão ou a própria função tenta atribuir valores a essas variáveis sozinha, se possível. Esta função, assim como todas as que recebem um objeto como *input*, funciona agora com Seurat, mas preservando uma retrocompatibilidade com CellRouter.

Uma função-irmã desta foi criada especificamente para aplicação em *loop* associada a valores dentro de uma coluna no metadado, `ComputeMultipleInteractions()`, como foi o caso com os trabalhos no dados da Shannon, usando como coluna condicional a associada aos *time-points*. Esta pequena melhoria facilitou extremamente o fluxo de trabalho,

evitando uma manipulação excessiva dos dados para executar uma tarefa “simples”: calcular as comunicações em um *dataset* de forma independente para cada tempo amostral.

LoadLigandReceptorPairs()

Esta função é responsável por carregar listas de genes ligantes e receptores que interagem. Por hora a única fonte disponível é a lista curada de pares presente no trabalho NicheNet,⁽¹³⁾ que também é a originalmente usada no lançamento do pacote CellComm.⁽¹⁴⁾ No futuro a função também receberá a possibilidade de incorporar pares do banco de dados Omnipath.⁽²⁴⁾ Antes a incorporação de pares era uma tarefa manual, agora é muito mais simplificada dentro desta função, que está imbuída dentro da função principal **ComputeInteractions()**.

PlotPairSankey()

Esta é uma função que adiciona mais uma visualização para os pares de interação celular, de forma a identificar fluxos possíveis de comunicação entre ligantes dentro do *cluster* “expedidor” (*sender*) e receptores dos *clusters* aceptores (*receivers*).

Uma versão simples dessa função já havia sido criada, mas não implementada ao arcabouço de funções do pacote CellComm até então.

LoadSpatialPositions()

Esta função é responsável por carregar informações espaciais dos *spots* em dados de ST. Foi necessário criá-la pois o *pipeline* do Seurat não inclui as informações de coordenadas dos *spots* na criação do objeto. Com a função esta informação é extraída dos arquivos oriundos do sequenciamento e incorporada no objeto Seurat, sendo desnecessária no fluxo de trabalho usando CellRouter.

CalculateSpatialMetrics()

Esta é uma função do tipo “agregadora” que compila as funções `calculateCentroids()`, `calculateDistanceMatrix()` e `LoadSpatialPositions()`. É empregada diretamente após a função `ComputeInteractions()`, e é responsável por calcular as posições espaciais dos *clusters* ou *subclusters* em análises de dados ST (Visium 10x).

AuxiliaryCellComm()

Esta função é também do tipo “agregadora” e une tanto as funções existentes `cellnetwork3()` e `interactionmatrix()`, quanto a nova, previamente descrita, `LoadSpatialPositions()`.

InternalizeCellCommDF()

Esta função foi desenvolvida para facilitar o fluxo de trabalho de usuários mais experientes, que desejem manipular o *output* da função `ComputeInteraction()` de forma isolada, fora do objeto, facilitando a incorporação desse *dataframe* processado. Também tem como objetivo retrocompatibilidade com o caso de projetos antigos, os quais possuíam o equivalente ao output de `ComputeInteraction()` como um *dataframe*, e este era usado como input de algumas funções *downstream*. De forma geral isso complexificava o *pipeline* de análises, o que se deseja evitar, por isso, agora a maioria das funções é orientada ao objeto principal (Seurat ou CellRouter), evitando-se complicações desnecessárias.

load_method_library()

Está é uma função simples mas que ajudou bastante para o trabalho de desenvolvimento de funções, e poderá ser usada em novas funções futuras. Ela é responsável por verificar se pacotes de R estão carregados, se não estiverem serão carregados, se não estiverem instalados, um aviso é emitido e a função principal é interrompida. Diversas funções novas e modificadas carregam esta nova função no início de seus códigos para evitar erros e facilitar alguns episódios de *debugging*.

GetHumanMouseHomologs()

Esta função foi criada para facilitar o processo de transformação de uma lista de genes humanos em seus equivalentes de camundongo. Sua criação se demonstrou necessária para facilitar o trabalho com dados murinos e humanos sem precisar de controle manual sobre a lista de pares de interação oferecidos pela função `LoadLigandReceptorPairs()`, embora ela possa ser usada avulsamente (fora da função que a incorporou).

check_gene_species()

Esta é uma função que oferece “qualidade-de-vida”, como a anterior, facilitando a detecção de padrões de nomeação de genes, e lidando com a transformação de lista de pares de interação de forma automática, facilitando o trabalho de usuários finais. Não é uma função que espera-se que os usuários usem, mas está devidamente embutida à `ComputeInteractions()`, de forma a rodar a `GetHumanMouseHomologs()`, caso seja verificado que os genes presentes no objeto não são humanos.

SubsettingObject()

Esta função é muito útil para “*subsetar*” objetos. É usada em vários pontos internos a funções como `ComputeInteractions()`, mas pode ser usada avulsamente também. A criação dela foi tremendamente importante pois originalmente era necessário efetuar o *subset* do objeto `CellRouter` manualmente, o que é uma tarefa bem mais complicada do que o equivalente a se fazer num objeto `Seurat`. Essa função simplifica tudo e homogeneiza o processo para ser feito de forma orientada a objeto, e executada de forma automática quando necessário dentro das diferentes etapas de análise sobre objetos `Seurat` ou `CellRouter`.

FilterCellComm_Dataframme()

Esta função foi criada para complementar e simplificar o código da função `PlotPairDotplot()`, tornando mais fácil ao usuário ter controle sobre quais interações devem ser exibidas no *plot*, conforme interesse e foco da análise. Isso inclui controle sobre os tipos celulares que enviam sinal, quais recebem, quais genes estão envolvidos como ligantes ou receptores, etc. As opções originais eram muito mais limitadas e menos intuitivas a usuários externos. A forma antiga de filtrar não foi eliminada, apenas foram acrescentadas opções novas e mais intuitivas, dando mais controle e poder aos usuários finais.

ComputeNMFInteractions()

Esta função é responsável por performar a fatorização sobre os dados de interações obtidos ao final da análise de comunicação superficial (ligantes-receptores). Ela encaixa-se no fluxo de análise diretamente após a função `PlotPairDotplot()`, e o algoritmo associado a ela é o algoritmo de fatoração de matrizes não-negativas (*Non-Negative Matrix Factorization* (NMF)), que acaba servindo como uma forma de *clusterizar* interações em

âmbito de pares ligante-receptor e *sender-receiver*. Seu uso pode ser interessante para identificar conjuntos de clusters celulares (*sender-receiver*) ou ligantes-receptores importantes para diferentes condições, sejam elas temporalmente definidas ou condicionais (fisiopatológicas, perturbadas, espacialmente delimitadas, etc). Esta foi uma inovação ao pacote que viabilizou (e viabilizará mais) análises como as presentes neste trabalho.

plot_heatmap_NMF()

Está é uma função simples de visualização dos resultados obtidos pelos cálculos de NMF. A função encontra-se internalizada à função `ComputeNMFInteractions()`, e pode ser modularmente modificada no futuro para criar plots mais interessantes aos usuários finais.

ComputeMultipleInteractions()

Esta função executa essencialmente o mesmo código que a função `ComputeInteractions()`, mas com o detalhe importante de fazer isso em um contexto de *subset* automático do objeto conforme uma coluna específica do metadado. Isso viabiliza que seja computada a análise de comunicação celular ao longo de todo o *dataset*, mas criando valores de análise individualmente para cada condição, possibilitando analisar as dinâmicas de comunicação celular ao longo de vários *time-points*, ou diferentes condições experimentais, com apenas uma função.

3. APLICAÇÕES DO CELLCOMM V2 EM DATASETS PÚBLICOS

De forma a executar os trabalhos de adaptação, como demonstrado na seção anterior, foram usados dois *datasets* diferentes. O primeiro com dados de ST, denominado como “TSK”, por causa do grupo de células *tumor-specific keratinocyte* (TSK), como feito pela publicação original.⁽¹⁷⁾ O segundo com dados de scRNA-seq, denominado como “AGM”, por causa do tecido em foco no estudo, *aorta-gonad-mesonephros* (AGM), local onde algumas células endoteliais se tornarão hemogênicas.⁽¹⁴⁾ Além da diferença na natureza dos dados entre o *dataset* TSK e o AGM, ST e scRNA-seq, respectivamente, há também a importante diferença no organismo sequenciado, sendo o primeiro de origem humana (carcinoma escamoso), e o segundo murina (AGM).

Dados (humanos) espaciais de TSKs

O objetivo desta etapa foi usar os dados de TSK para dar início às adaptação das funções do CellComm ao *framework* Seurat, especialmente as que compõem as análises do módulo “superficial” (referentes a ligantes-receptores). A ideia de começar com esses dados originou-se no fato de o trabalho original do CellComm em 2022 ter explorado esses dados, então parte das interações já havia sido descrita, sendo, portanto, um “terreno conhecido”. Além de que, fazendo funcionar para os dados de ST, tornaria-se mais fácil adaptar o que faltasse à análise de scRNA-seq, considerando que a primeira possui mais informações (e etapas) que a segunda.

Os trabalhos foram feitos, a adaptação das funções iniciais foi feita, viabilizando inclusive as funcionalidades de análise de comunicação em contexto espacial, aproveitando a natureza espacial dos dados de ST. Mas não foi sem contratempos; acontece que a criação do objeto Seurat com dados de ST não extrai dos dados a mesma informação que o pacote CellComm fazia, na criação do objeto CellRouter. A informação da posição dos spots não era importada no Seurat. Foi necessário criar uma função que fizesse a extração dessa informação dos dados de sequenciamento e a salvasse no objeto Seurat, de forma a ser usada posteriormente nas análises espaciais. Essa função veio a ser a `LoadSpatialPositions()`. Quem sabe, em uma iteração futura, possamos eliminar a necessidade de rodar tal função. Entretanto, atualmente ela se faz necessária se o usuário deseja fazer inferências espaciais no corte histológico em questão. O que é um problema, considerando que nem sempre os autores disponibilizam os dados de sequenciamento contendo a informação que precisamos extrair. O compartilhamento do objeto Seurat é também mais comum em casos de colaborações, que podem nem possuir os dados de sequenciamento de cada amostra analisada. Esta é, portanto, uma limitação do trabalho na atual versão.

Apesar do que foi previamente explanado, a adaptação do *pipeline* nesse conjunto de dados foi, em suma, um sucesso. Uma amostra do funcionamento da nova versão do pacote pode ser acessada pelo presente [link](#), funcionando como um tutorial de análises para dados espaciais usando o novo fluxo de trabalho compatível com Seurat.

Dados murinos de scRNA do AGM

O trabalho sobre os dados de AGM teve como função verificar se as adaptações feitas para dados de ST conseguiriam ser generalizadas para os dados de scRNA-seq, além de identificar quais mudanças deveriam ser feitas para facilitar o trabalho com dados de camundongo, com o menor atrito possível. A escolha deste *dataset*, além das claras diferenças que promoveriam uma testagem mais robusta sobre os “alicerces” do código desenvolvido, possuía mais dois componentes que o favoreciam: esses dados já estavam devidamente anotados, o que facilita muito o fluxo de trabalho; e o *dataset* em questão foi gerado especialmente para o trabalho que resultou no desenvolvimento da primeira versão do pacote CellComm, publicado em 2022.⁽¹⁴⁾

Com a aplicação do novo fluxo de trabalho, via Seurat, nos dados de AGM, foi verificado que grande parte das alterações, originalmente desenhadas para os dados de TSKs, podiam ser generalizadas para os dados de scRNA-seq com sucesso. Funcionalidades recém construídas lidavam com a natureza dos dados e outros parâmetros, de forma implícita, com êxito, sem a necessidade de o usuário manualmente lidar com essas questões. Assegurando a proposta inicial de tornar o fluxo de trabalho mais acessível para usuários finais, facilitando o uso do pacote em geral.

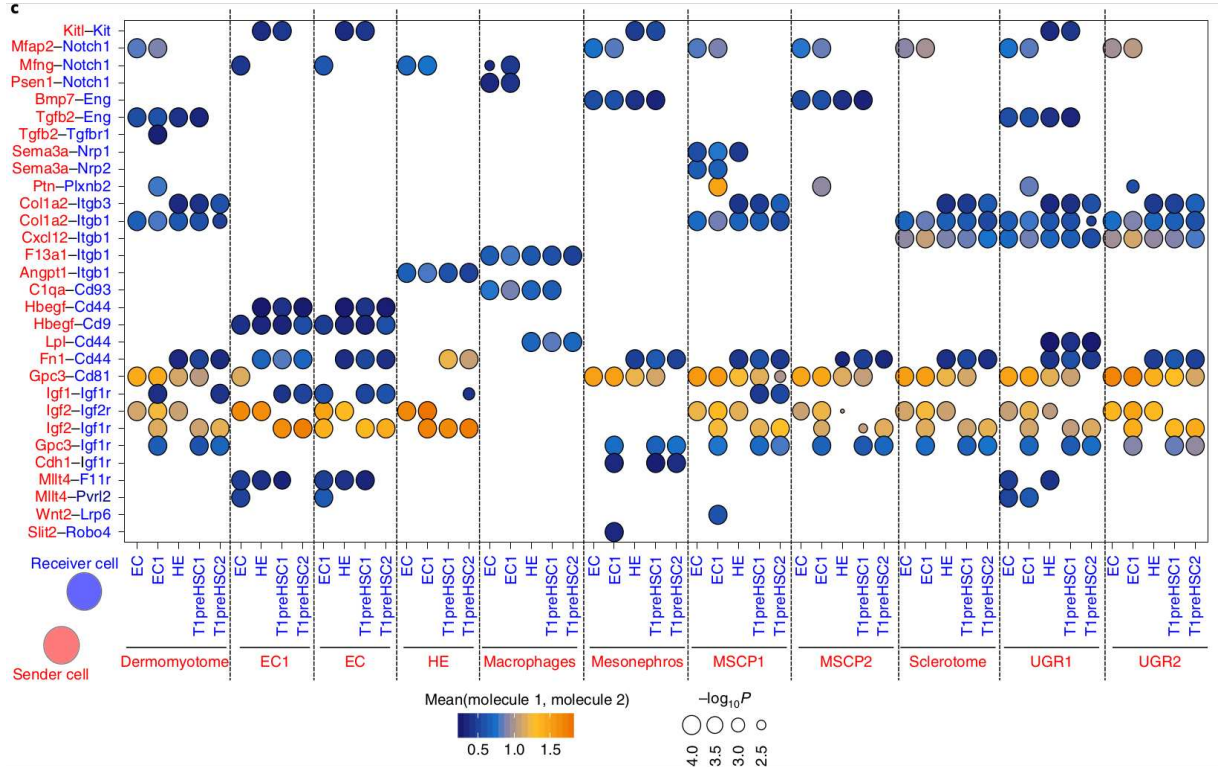
As questões envolvendo a natureza dos dados ser murina, tiveram de ser tratadas pontualmente. O único verdadeiro entrave encontrado, ao tratar com dados dessa natureza no novo fluxo de trabalho, ocorreu no momento em que teve de se aplicar a principal função de comunicação celular desenvolvida, `ComputeInteractions()`, pois o banco de ligantes-receptores usados, o qual foi gerado pelo trabalho NicheNet,⁽¹³⁾ é de natureza humana. Para resolver esta questão, foram desenvolvidas funções responsáveis por lidar com diferentes pontos desse entrave: primeiramente identificar qual a natureza dos dados; em segundo lugar, executar a identificação de genes homólogos humano-murinos e convertê-los apropriadamente. Essas funcionalidades deveriam entrar em ação em conjunto, de forma a lidar com o problema, idealmente, sem necessidade de *input* do usuário. As novas funções desenvolvidas e empregadas na resolução de cada uma das questões aqui expostas, foram denominadas `check_gene_species()` e `GetHumanMouseHomologs()`, respectivamente. E assim, passaram a compor parte do código executável que roda implicitamente sob a função `ComputeInteractions()`.

O trabalho sobre o dataset de AGM serviu, primeiramente, como forma de comparar os resultados obtidos pelo novo fluxo de trabalho, com os resultados antigos publicados

conjuntamente à versão original do CellComm.⁽¹⁴⁾ Não houve um esforço grande no sentido de efetuar o processamento dos dados de forma exatamente igual ao executado no passado, até para tentar verificar se, mesmo com configurações ligeiramente diferentes, padrões em comum seriam observados. Para efetuar essa comparação de forma pontual foi gerado o gráfico oriundo da função `PlotPairDotplot()`, usado na Figura 2C da publicação original,⁽¹⁴⁾ buscando filtrar pelos mesmos pares de ligantes-receptores e tipos celulares interagindo. Houve uma discrepância entre os ligantes-receptores originais (Figura 2) e os obtidos na tentativa de reprodução (Figura 3). Provavelmente as diferenças entre os pares de interação que surgiram, deixaram de aparecer ou mudaram de valor, na tentativa de reprodução da imagem, são oriundas das diferenças entre as configurações originais de processamento dos dados usando o pacote FUSCA do CellComm v1, e as configurações usadas aqui via *pipeline* do Seurat, além de outras possibilidades discutidas posteriormente na seção Limitações, “Reprodutibilidade dos dados de AGM”. Mas, de forma geral, pode-se dizer que os resultados são compatíveis, visto que comunicações associadas a receptores como *Igf1r*, *Igf2r*, e *Kit* se mantiveram. Porém, uma busca maior visando elucidação da raiz dessas incongruências deve ser conduzida, para garantir a total reprodutibilidade e eficácia na adaptação do código ao *framework* do Seurat.

Figura 2 - Dados AGM - Figura 2C do artigo original

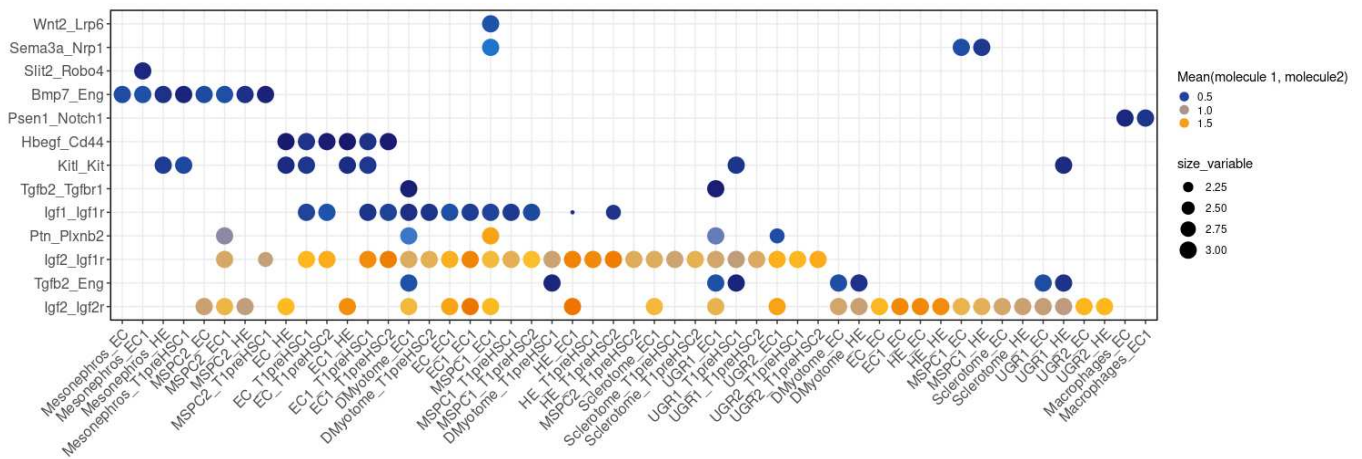
Gráfico de dispersão dos pares de ligantes-receptores (eixo y) para tipos celulares selecionados associados à EHT (eixo x). Figura extraída do artigo original do CellComm.



Fonte: Figura adaptada de LUMMERTZ DA ROCHA, E. et al., 2022 ⁽¹⁴⁾

Figura 3 - Dados AGM - Tentativa de reprodução da Fig.2C original

Gráfico de dispersão dos pares de ligantes-receptores (eixo y) para tipos celulares selecionados associados à EHT (eixo x). Esta figura tenta reproduzir os resultados da Figura 2.



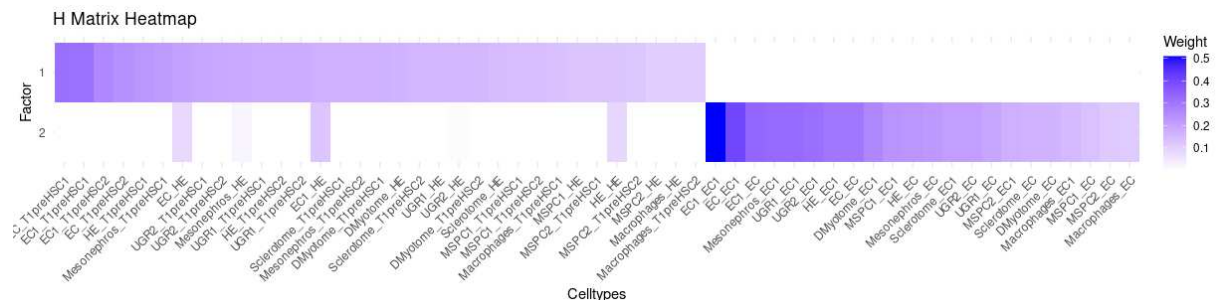
Fonte: Figura gerada pelo autor (2025)

Além da comparação de reprodutibilidade entre as versões do código, buscou-se explorar como as novas funcionalidades do código, em especial as análises de *Non-Negative Matrix Factorization* (NMF), poderiam agregar na busca por ligantes-receptores no contexto de nascimento das células tronco hematopoiéticas. Para cumprir com esse papel foi feito o que

não seria viável anteriormente de forma manual: foi gerado um gráfico de comunicação *Dotplot*, através da função `PlotPairDotplot()`, semelhantemente ao mostrado a pouco, mas, ao invés de selecionar os 10 pares de ligantes-receptores mais relevantes de cada população (par de *clusters* celulares comunicando), foram usados 200 pares. O resultado desse plot não é algo viável de se analisar pelos padrões anteriores, através de curadoria individual do pesquisador, mas é ideal para se usar como *input* na função que calcula o NMF, `ComputeNMFInteractions()`. Através da análise de NMF, com os *clusters* chave no processo de Transição Endotélio-Hematopoiética (EHT) como receptores da comunicação, e com 200 pares de ligantes-receptores, obteve-se que a matriz de comunicação em AGM possuía alguns *fatores* interessantes que a decompunham. Foram feitas duas análises de NMF, uma usando o número de fatores (*rank*) para a fatorização como 2, e outra como 8. Esses valores foram escolhidos baseando-se no cálculo de estimativa de *optimal-rank*, do próprio pacote “NMF” do R. Os resultados do experimento com “rank=2” interessantemente selecionaram comunicações que eram predominantes, ou no conjunto de células T1PreHSC e HE, ou EC, o que pode ser visualizado na Figura 4. Já no caso da fatorização com “rank=8”, o oitavo fator estava enriquecido para populações envolvendo um dos *clusters* T1PreHSC, que pode ser observado na Figura 5. Na posse dessas novas informações, pudemos retornar à função `PlotPairDotplot()` com algo mais focado em mente. Através da análise de NMF como intermediária para seleção de pares de ligantes-receptores, permitiu-se chegar às Figuras Figura 6, Figura 7 e Figura 8, ressaltando-se os ligantes-receptores dos fatores 2-1 e 8-8, que possuíam uma expressiva associação às populações mais relevantes à emersão das células tronco hematopoiéticas do endotélio.

Figura 4 - Dados AGM - NMF - Fator X Par-Cellular (*Rank 2*)

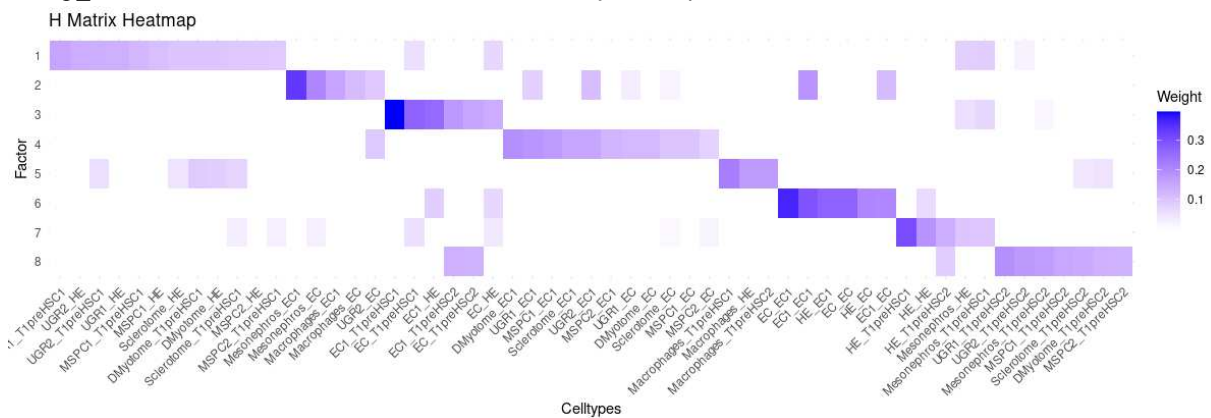
Heatmap associando os dois fatores encontrados por melhor dividir os grupos de células interagindo associados à EHT. Fatores compõem as linhas (eixo y) e cada par de comunicação celular compõe as colunas (eixo x). Os valores de preenchimento são calculados pela relação do fator em questão com o cálculo: $\sqrt{(-\log_{10} p_value) \times mean}$, onde *p_value* e *mean* vêm dos cálculos oriundos da função `ComputeInteractions()`.



Fonte: Figura gerada pelo autor (2025)

Figura 5 - Dados AGM - NMF - Fator X Par-Cellular (*Rank 8*)

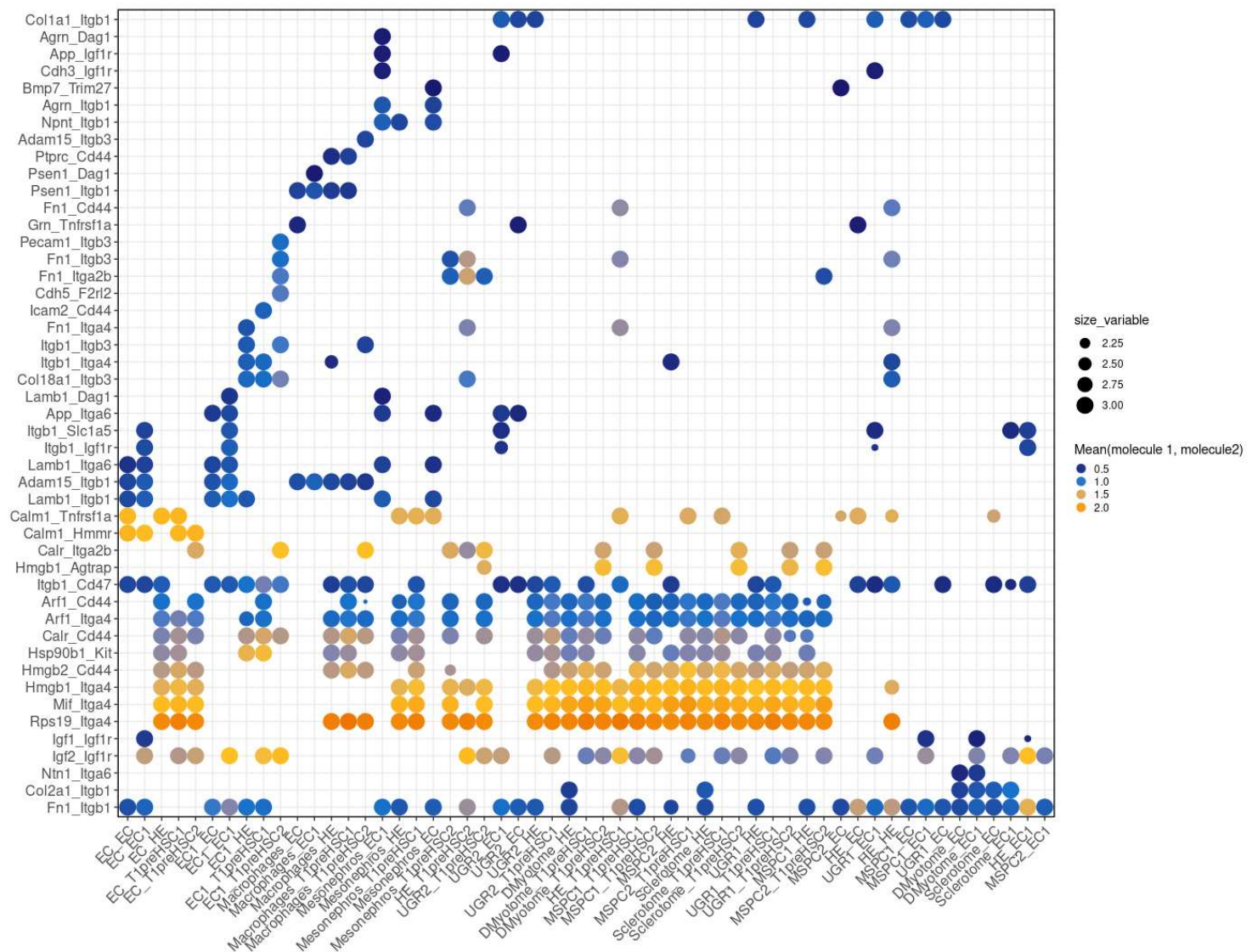
Heatmap associando os oito fatores encontrados por melhor dividir os grupos de células interagindo associados à EHT. Fatores compõem as linhas (eixo y) e cada par de comunicação celular compõe as colunas (eixo x). Os valores de preenchimento são calculados pela relação do fator em questão com o cálculo: $\sqrt{(-\log_{10} p_value) \times mean}$, onde p_value e mean vêm dos cálculos oriundos da função `ComputeInteractions()`.



Fonte: Figura gerada pelo autor (2025)

Figura 6 - Dados AGM - Dotplot - Fator 1 (Rank 2)

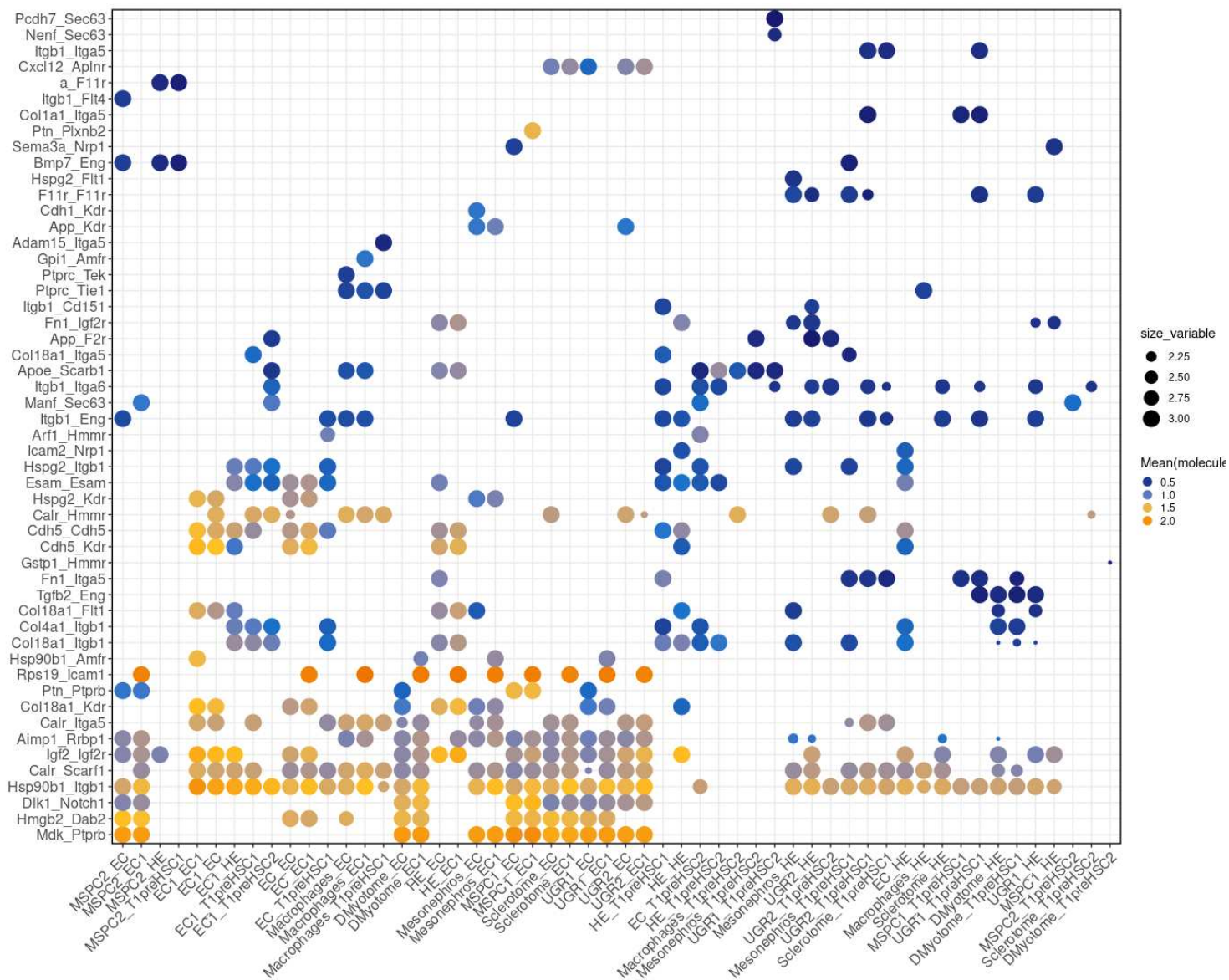
Gráfico de dispersão dos pares de ligantes-receptores para tipos celulares selecionados associados à EHT. Foram filtrados para aparecer apenas os pares de ligantes-receptores associados ao fator 1 do cálculo de `ComputeNMFInteractions()`, usando `rank=2` para a construção da matriz.



Fonte: Figura gerada pelo autor (2025)

Figura 7 - Dados AGM - Dotplot - Fator 2 (Rank 2)

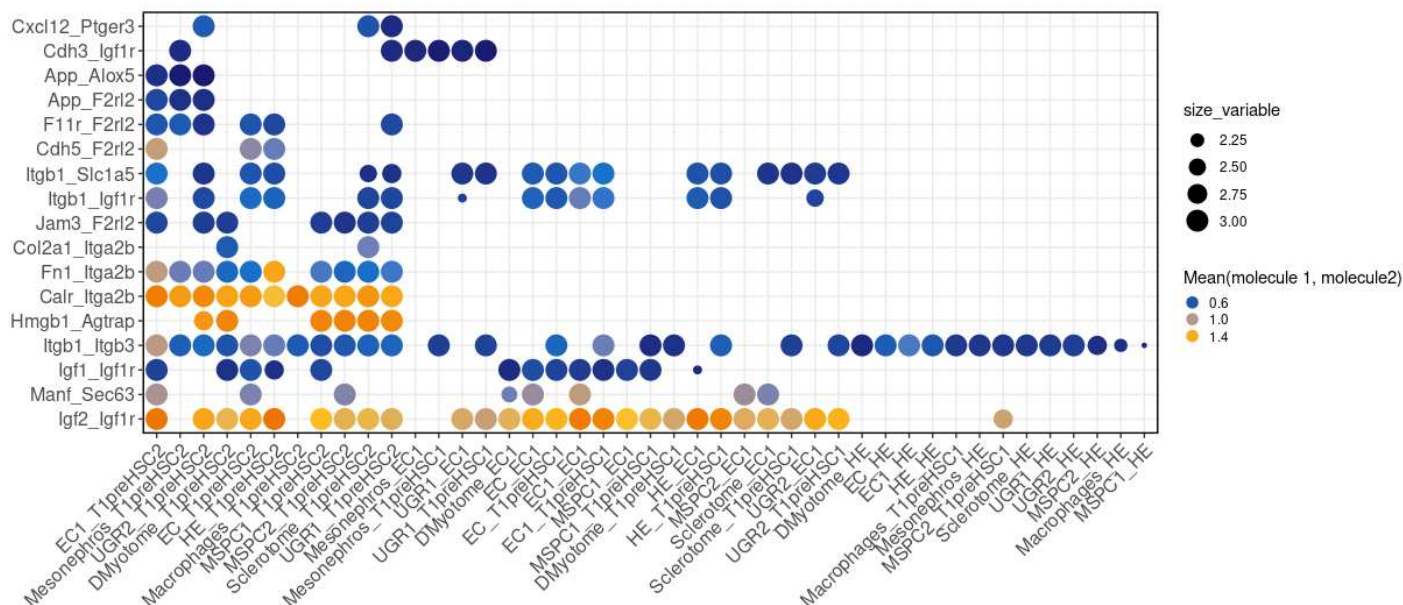
Gráfico de dispersão dos pares de ligantes-receptores (eixo y) para tipos celulares selecionados associados à EHT (eixo x). Foram filtrados para aparecer apenas os pares de ligantes-receptores associados ao fator 2 do cálculo de `ComputeNMFInteractions()`, usando `rank=2` para a construção da matriz.



Fonte: Figura gerada pelo autor (2025)

Figura 8 - Dados AGM - Dotplot - Fator 8 (Rank 8)

Gráfico de dispersão dos pares de ligantes-receptores (eixo y) para tipos celulares selecionados associados à EHT (eixo x). Foram filtrados para aparecer apenas os pares de ligantes-receptores associados ao fator 8 do cálculo de `ComputeNMFInteractions()`, usando `rank=8` para a construção da matriz.



Fonte: Figura gerada pelo autor (2025)

Novamente, a aplicação do novo fluxo de trabalho neste “novo mas explorado” conjunto de dados, demonstrou a eficácia dos novos alicerces em identificar resultados previamente validados, bem como possíveis novas perspectivas a se explorar com a aplicação de NMF. Ademais, o presente trabalho também deixa disponível, através do [link](#), o código usado para a obtenção dos resultados sobre os dados de AGM usando o novo fluxo de trabalho, servindo este como um tutorial de uso, considerando dados de scRNA-seq.

Dados murinos de CITE-seq da medula-óssea em maturação

Após a aplicação do novo conjunto de códigos (NMF) aos dados de AGM, chegou o momento de aplicá-lo ao conjunto de dados que realmente era o alvo inicial de seu desenvolvimento, os dados multi-temporais de medula óssea, gerados por uma antiga colaboradora do laboratório, Shannon McKinney-Freeman. Estes dados são de medula óssea de camundongo coletados em 4 tempos diferentes, E16.5, E18.5, P0 e Adulto, como já fora previamente descrito na primeira seção dos Materiais e Métodos. A ideia com estes dados era buscar por pares de ligantes-receptores que ajudassem a explicar o processo de maturação do microambiente da medula óssea, entre o final do período embrionário e o nascimento, de modo a explicar por que a medula óssea adulta suporta hematopoiese enquanto a fetal, não.

Durante o decorrer dos trabalhos sobre esses dados, aos quais me referirei como “dados da Shannon”, foi nitidamente perceptível como as adaptações e novas funcionalidades ao CellComm facilitaram o fluxo de análises. Foram mensurados RNA e proteínas de superfície das células, usando CITE-seq, e o CellComm funcionou sem a necessidade de explicitar qual o tipo de dado a se usar, dando preferência ao “*active-assay*”, setado pelo próprio Seurat. Ademais, o cálculo de comunicações não só passou pela conversão automática de genes (ligantes-receptores) de humanos para murinos, como ao submeter o metadado (coluna da matriz) que continha a informação temporal do *dataset* (com os 4 *time-points*) à função `ComputeMultipleInteractions()`, esta foi capaz de executar os cálculos subdividindo o objeto de acordo com cada condição presente na coluna, automaticamente. Ou seja, ao usar a nova função, desenhada para análise de múltiplas condições, foram feitas análises de comunicação independentes para cada valor dentro da variável fornecida (tempo, no caso), com *subset* automático. Resultando, por fim, num quinto *data-frame* que concatena a informação de todos os tempos numa só tabela, preservando a informação temporal, viabilizando plots mais elaborados e facilitando muito os esforços do usuário final.

Para os trabalhos sobre os dados da Shannon, algumas funções antigas foram melhoradas, incluindo uma categoria nova de *plots* experimentais dentro de `PlotPairDotplot()` e `ComputeNMFInteractions()`. Algumas funções novas foram implementadas, como a `ComputeMultipleInteractions()`, que faz essencialmente o mesmo que a `ComputeInteractions()`, mas com o foco em análise de múltiplas condições simultaneamente, com o menor esforço adicional do usuário possível.

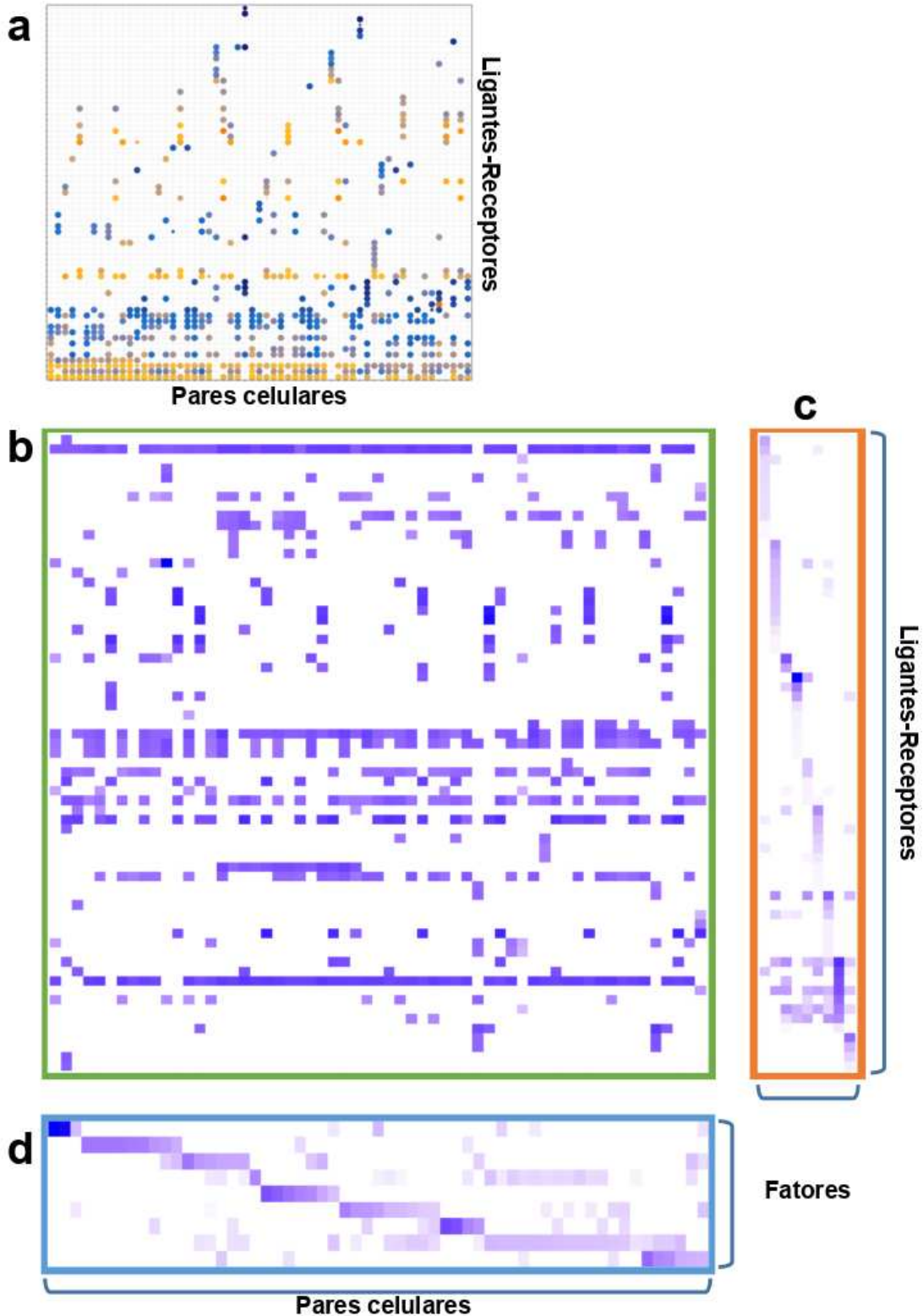
Os passos que compõem o processo de execução das análises sobre este *dataset* serão descritos brevemente a seguir. Primeiro foram carregados os dados via GEO para dentro do servidor de trabalho, sendo identificadas quais colunas do metadado traziam informações relevantes, como as amostras, tempo, componente tecidual e anotação de tipo celular. Logo após o cálculo de redução de dimensionalidade sem integração, notou-se um efeito de *batch* claro para algumas amostras. Foi executada a integração via BBkNN,⁽²⁶⁾ de forma a valorizar os tipos celulares anotados pelos autores originais. Concluída a integração dos dados, passou-se então para a análise de comunicação celular usando CellComm, tendo sido empregada a função `ComputeMultipleInteractions()`, que tornou o cálculo das análises muito mais rápido e simples. Foram usados os 10 pares de ligantes-receptores mais relevantes para cada par celular, cujo eixo de comunicação deveria ter as células HSPC como receptoras, para criar um *dotplot*, que podemos observar na Figura 9-a. A partir deste *plot*, que

estruturou-se como uma matriz, foram efetuados os cálculos de NMF, através da função `ComputeNMFInteractions()`, seguindo os componentes b-d, também da Figura 9. Com as matrizes oriundas da análise de NMF, foi observado que alguns fatores apresentavam-se em padrões interessantes, seja associados aos tipos celulares que enviavam o sinal, ou agrupando-se pelos próprios *time-points*. Ambos os casos foram investigados, mas para o primeiro caso, poucas comunicações se mostraram diferenciais ao longo dos intervalos de tempo. Ao obter-se esses fatores de interesse, o passo seguinte foi retornar aos *dotplots* para procurar por padrões associados à passagem de tempo, e, portanto, maturação do microambiente da medula. Alguns ligantes-receptores foram identificados concomitantemente pela análise dos autores originais, outros foram exclusivos de cada análise. Houve pares que só foram identificados na análise original, e também pares únicos da análise do CellComm; detalhes serão tratados posteriormente. Em seguida, houve o esforço de procurar por ligantes-receptores identificados pelo artigo da Shannon,⁽¹⁶⁾ verificando-se que nem todos os achados puderam ser replicados, como será tratado na seção “Limitações” da Discussão. Por fim foi executada uma análise de NMF com os 150 pares mais relevantes de cada eixo de comunicação, ao invés de apenas 10, como previamente feito. Esse último esforço foi focado em encontrar ligantes-receptores associados aos fatores mais marcantes de cada ponto temporal, buscando captar ao máximo o potencial intrínseco da análise de NMF.

Figura 9 - Dados Shannon - Esquema elementos NMF

Painel de gráficos em um arranjo esquemático representando os componentes associados à análise de NMF. Componentes:

- a.** *Dotplot* de ligantes-receptores (linhas) por pares de tipos celulares interagindo (colunas);
- b.** Matriz expressa em um *Heatmap* com ligantes-receptores (linhas) por pares de tipos celulares interagindo (colunas), cujos valores de preenchimento são oriundos da raiz quadrada da multiplicação entre a expressão média dos ligantes-receptores e o seu respectivo $-\log_{10}$ p-valor, em cada par de tipos celulares interagindo;
- c.** Matriz expressa em um *Heatmap* com ligantes-receptores (linhas) pelos fatores (colunas), identificados pelo cálculo de NMF, cujos valores de preenchimento expressam a contribuição de cada fator para cada par de ligantes-receptores;
- d.** Matriz expressa em um *Heatmap* com fatores (linhas), identificados pelo cálculo de NMF, pelos pares celulares interagindo (colunas), cujos valores de preenchimento expressam a contribuição de cada fator para cada par de tipos celulares interagindo.



Fonte: Figura esquemática criada pelo autor (2025)

Alguns dos novos ligantes-receptores relevantes encontrados nas análises, focando em comunicação de populações restritas (olhando especialmente para um conjunto específico de *clusters* celulares), foram: comunicação autócrina, HSPC-HSPC (Itgb1_Rack1, Hmgb1_Agtrap, Calr_Cd44, Calr_Hmmr); Fibroblastos-HSPC ou progenitores-HSPC (Comp_Cd47, Col2a1_Itgb1, Comp_Itgb1, Fn1_Itga4, Fn1_Cd44); endotélio-HSPC, Osteo-HSPC ou CARs-HSPC (Hsp90b1_Kit, Hspg2_Itgb1, Clst1_Cd47, Itgb1_Eng, Itgb1_Igflr); estroma total (App_Igflr, Spp1_Itgb1, Spp1_Cd44, Spp1_Itga4).

A aplicação deste novo fluxo de trabalho no conjunto de dados de medula óssea demonstrou a capacidade do CellComm em identificar resultados previamente validados, no que tange aos poucos ligantes-receptores testados originalmente, bem como possíveis novas perspectivas a se explorar oriundas da aplicação de NMF. Ademais, o presente trabalho novamente deixa disponível, através do [link](#), o código usado para a obtenção dos resultados sobre os dados da Shannon usando o novo fluxo de trabalho, servindo este como um tutorial de uso, considerando dados de scRNA-seq com múltiplas condições temporais.

DISCUSSÃO

A discussão do presente trabalho foi estruturada em tópicos, dividindo-se por dataset ou questões gerais do trabalho, tendo-se como exemplo as limitações atuais, com as quais esforços futuros podem tentar resolvê-las, caso haja solução aplicável. Serão discutidos brevemente alguns dos achados referentes aos dados de AGM e medula óssea.

1. DADOS DO AGM

Aplicando NMF para buscar ligantes-receptores específicos de populações EHT

Uma série de achados, na forma de ligantes-receptores, foram expostos nas Figuras 6-8, que tratam da aplicação de NMF para encontro de padrões de comunicação nas células responsáveis pela transição endotélio-hematopoiética. Nas Figuras 4 e 5 vemos que são os fatores 1 (do NMF com rank=2) e 8 (do NMF com rank=8) que trazem os achados potencialmente mais relevantes para explicar o processo de EHT. A visualização de quais seriam as interações associadas com esses fatores fica visível nas Figuras 6 e 8.

Com as informações computacionais obtidas para a publicação original, haviam sido identificados ligantes-receptores provavelmente associados com o processo de EHT, mas

apenas alguns foram efetivamente testados *in vitro* ou *in vivo*, devido ao imenso trabalho de validação necessário para cada alvo a se testar. A partir das novas informações oriundas da análise de NMF, alguns eixos de comunicação antigos se perpetuaram como relevantes, assim como houve o aparecimento de novos possíveis alvos. Itgb1 e Itgb3 foram dois dos ligantes identificados e testados no trabalho original, escolhidos especialmente após a análise de sinalização intracelular, que não foi completa aqui, como será descrito na última seção de “Limitações”. No entanto, através das análises de comunicação superficial somadas à estratégia de NMF, já conseguimos chegar ao alvo Itgb1, como podemos ver nas Figuras 6 e 8. Além de que foram encontrados possíveis novos alvos de interesse, como: os ligantes Fn1, Igf1, Igf2, Rps19, Mif, Hgmb1, Hgmb2, Hsp90b1, Calr, Arf1, Calm1, App; e dos receptores Igf1r, Agtrap, Itga2b, F2lr2, Itga4, Kit, Itgb1, Itga6, Cd44, Hmnr, Tnfrsf1a.

Dos pares observados nas Figuras 6 e 8, os genes Kit e Cd44 já são conhecidamente relevantes no desenvolvimento hematopoiético, sendo o último já reconhecido como atuante na própria transição endotélio-hematopoiética.⁽²⁷⁾ HADLAND B. e colaboradores encontraram que os genes Icam1, Vcam1, Fn1 (Fibronectin), Tgm2 e Cxcl12 estavam mais expressos nas células do AGM que davam suporte ao emergir das células tronco hematopoiéticas. Por fim, validaram que Itga4/Itgb1 e Fibronectin (Fn1) eram necessários para dar suporte à geração de HSCs a partir de precursores hemogênicos da AGM. Também no contexto de nichos hematopoiéticos, conexões com APP foram identificadas em nichos como a AGM, fígado fetal (FL) e medula óssea (BM) por FENG C. e colaboradores.⁽²⁸⁾ E, embora não exploradas, comunicações através de FN1 também foram observadas. Desta forma, dos achados do presente trabalho, pelo menos Fn1, App, Kit e Cd44 já foram descritos como tendo relações com o processo de EHT, ou pelo menos com os sítios embrionários das HSCs.

As menções acerca dos demais alvos encontrados, no escopo de HSCs ou EHT, passam a se tornar mais rarefeitas. High mobility group box 1 (HMGB1) parece ter um papel crucial em cânceres líquidos,⁽²⁹⁾ e Calreticulin (Calr) parece enviesar a diferenciação de HSCs rumo à perfis eritróides e megacarióides, além de neoplasmas mieloproliferativos.^(30,31) Igf1, Igf2 e Igf1r são pares de ligantes-receptor associados à sinalização de crescimento pela via “*insulin-like*“, e já foram previamente relatados como relevantes no cenário hematopoiético.^(32,33,34) Tnfrsf1a é um receptor de TNF-alpha, e seu papel no controle de sinais inflamatórios já foi associado a nascimento das HSCs.⁽³⁵⁾ Os demais genes não possuem um papel tão conhecido no desenvolvimento hematopoiético.

Em suma, diversos dos achados sustentam-se quando comparados com conhecimentos previamente descritos na literatura. Embora os demais alvos não possuam uma relação tão clara até então, isto os torna candidatos ideais para uma investigação mais profunda e formulação de futuros experimentos.

2. DADOS DA MEDULA ÓSSEA

Semelhanças e incongruências com paper original

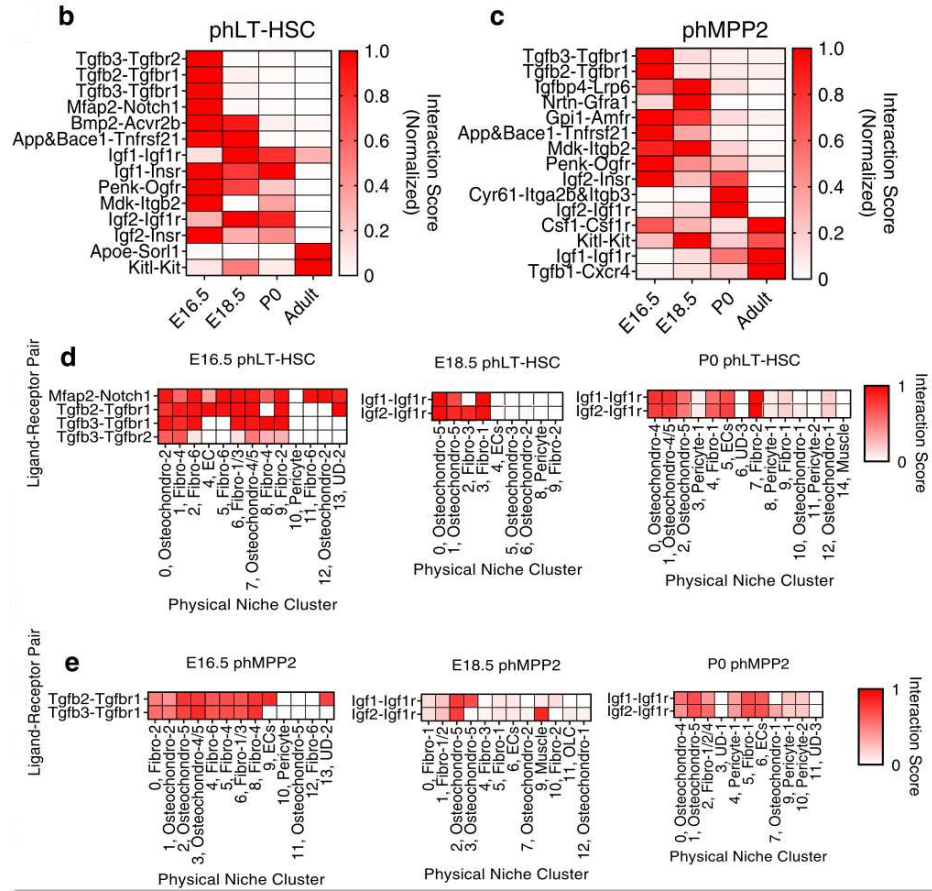
Os autores do artigo original⁽¹⁶⁾ relataram alguns ligantes-receptores na busca por associações com a maturação do microambiente da medula óssea, como podemos ver no recorte trazido na Figura 10. Tendo em vista o recorte mencionado e os resultados descritos no próprio artigo, um dos objetivos lógicos era tentar reproduzir estes achados.

A partir dos conjuntos de pares de ligantes-receptores presentes nos gráficos b e c da Figura 10, foi elaborada a Figura 11-a, e como podemos ver ao compará-la com os achados originais, ocorrem algumas dissimilaridades. Porém, vale ressaltar que os autores validaram apenas o impacto de Igf1 e Igf2, genes estes cuja relevância também foi captada pelo nosso algoritmo, como pode-se observar tanto na Figura 11-a, quanto em 11-b – demonstrando a consistência de nossa análise. Detalhes explorando o porquê das diferenças entre resultados gerais serão tratados dentro da seção Limitações, subtópico associado aos “dados da Shannon”. Mas, tratando brevemente aqui, a diferença ou “variedade” entre os achados aponta para dois fatores mais prováveis: diferenças de *software* de análise de comunicação, ou pré-processamento dos dados de forma diferente. Mesmo parte dos ligantes-receptores não aparecendo conjuntamente nas duas análises, e alguns dos que apareceram não demonstrarem os mesmos padrões, houve pares que se mantiveram comparáveis, como o padrão de comunicação Kitl_Kit.

Além da tentativa de reproduzir a figura original com os mesmos pares de ligantes-receptores, também buscou-se pelos genes melhor explorados na argumentação dos autores originais em defender seus achados, os genes associados à sinalização de TGF- β e *Insulin-like Growth Factor*. O conjunto de ligantes e receptores destas vias foram então usados na criação da Figura 11-b, onde os nossos resultados se alinham de forma geral com a narrativa proposta pelos autores originais. Principalmente no que tange à sinalização de Igf1 e Igf2, que não é encontrada em E16.5, mas passam a aparecer nos tempos seguintes.

Figura 10 - Dados Shannon - *Heatmaps* - Ligantes-Receptores Originais

Conjunto de *Heatmaps* de pares de ligantes-receptores associados a cada um dos *time-points* dos dados. Primeiro são observados padrões de cada *time-point*, em seguida são destrinchadas quais populações estão associadas com as comunicações nos painéis inferiores.

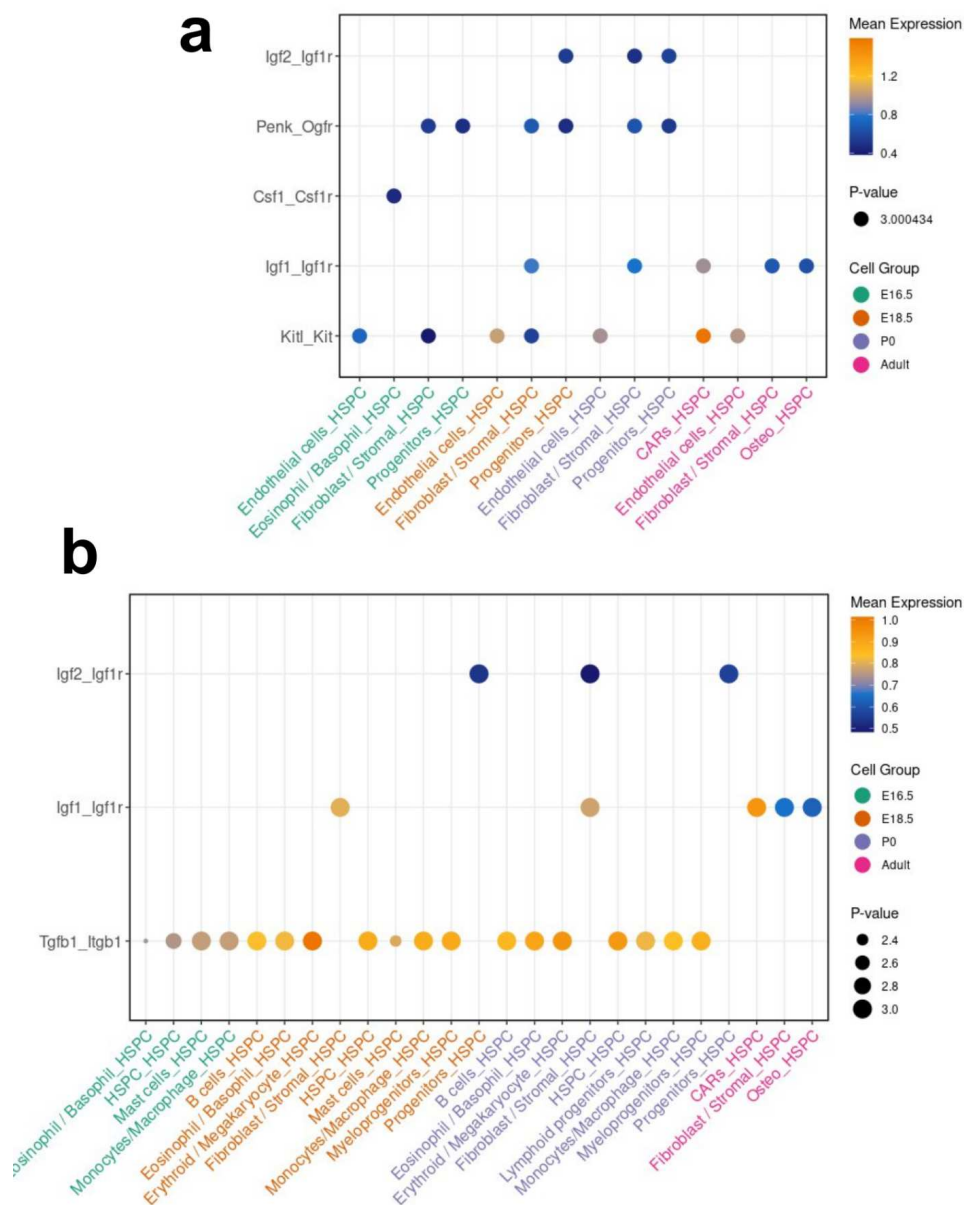


Fonte: Figura adaptada de HALL T. D. et al, 2022 ⁽¹⁶⁾

Figura 11 - Dados Shannon - *Dotplot* - Busca Pelos Ligantes-Receptores Originalmente Destacados
Gráfico de dispersão dos pares de ligantes-receptores e tipos celulares interatores.

Componentes:

- a.** *Dotplot* de interações encontradas no artigo original, usando mesmos ligantes-receptores que em Fig. 10-b e 10-c;
- b.** *Dotplot* de interações descritas no artigo original, filtrando por qualquer interação usando os ligantes e receptores canônicos dos genes considerados associados à maturação da medula óssea por HALL T. D. et al, 2022.



Fonte: Figura gerada pelo autor (2025)

Explorando interações iniciais entre HSPCs e outros tipos celulares

Após passar pelas etapas de comparação com os resultados originalmente publicados, o próximo passo natural era efetuar a exploração dos ligantes-receptores obtidos da análise de comunicação usando CellComm via `ComputeMultipleInteractions()`. Dos gráficos

com maior valor científico, foi elaborado o painel presente na Figura 12, onde podemos visualizar comunicações entre diversos tipos celulares diferentes, como fibroblastos e células progenitoras hematopoiéticas (12-b), endotélio (12-c) e um conjunto expandido de células do estroma (12-d), até mesmo comunicação autócrina entre HSPCs (12-a).

Dos componentes da Figura 12, alguns pares de comunicação se destacam como aparentemente associados ao processo de maturação que podemos ver sequencialmente na passagem do eixo-x (com diferentes cores para cada *time-point*). São estes quanto à: comunicação autócrina, 12-a (Itgb1_Rack1, Hmgb1_Agtrap, Calr_Cd44, Calr_Hmmr); fibroblastos e progenitores hematopoiéticos, 12-b (Comp_Cd47, Col2a1_Itgb1, Comp_Itgb1, Fn1_Itga4, Fn1_Cd44); células do endotélio, 12-c (Hsp90b1_Kit, Hspg2_Itgb1, Clst1_Cd47, Itgb1_Eng, Itgb1_Igflr); conjunto de células estromais, 12-d (App_Igflr, Spp1_Itgb1, Spp1_Cd44, Spp1_Itga4).

Ao buscar por esses conjuntos de ligantes-receptores na literatura, alguns paralelos puderam ser traçados. Embora esparsas, e por ora rasas, algumas conexões com a literatura trouxeram evidências de que genes encontrados pela nossa análise são, de fato, impactantes no processo de maturação do microambiente da medula óssea como um nicho hematopoiético, ou pelo menos seriam associados à homeostase de células tronco hematopoiéticas (HSCs).

Do conjunto total de 18 pares de ligantes receptores elencados acima, apenas 5 genes que compõem tais comunicações (Hmgb1, Cd44, Cd47, Endoglin/Cd105 e Spp1), possuíam descrição mais evidente e validada na literatura. Seguindo a ordem em que os mesmos foram expostos temos:

Hmgb1, aparecendo em par Hmgb1-Agtrap na Figura 12-a (comunicação via autócrina). High mobility group box 1 (HMGB1) é um gene associado a cromatina, servindo como um *damage-associated molecular pattern* (DAMP). Como ligante extracelular ele pode regular diversos processos como proliferação, diferenciação, mobilização e senescência de HSCs.⁽²⁹⁾

Cd44, aparecendo múltiplas vezes (12-a-b-d, comunicação via autócrina, fibroblastos, progenitores e CARs e osteo), em pares como Calr-Cd44, Fn1-Cd44 e Spp1-Cd44 (Opn-Cd44). Previamente descrito também como associado à EHT,⁽²⁷⁾ o gene tem igualmente diversos papéis na regulação de HSCs, como migração, adesão e manutenção.^(36,37) Inclusive seu papel de adesão parece se tornar ainda mais significativo com múltiplas demonstrações associando HSCs interagindo com ácido hialurônico (componente da ECM), interação essa mediada via CD44.^(38,36,37) De fato, ainda outros paralelos serão traçados com relação ao ácido

hialurônico e a maturação do microambiente da medula óssea, como veremos na próxima sessão, através das Figuras 15-a-c-d e 16-d.

Cd47, aparecendo na Figura 12-b-c (comunicação via progenitores, fibroblastos e endotélio), sob interações como Comp-Cd47 e Clstl-Cd47. Um gene com importância transiente, ao que se conhece pelo menos, sendo descrito como superexpresso em momentos chave de inflamação, servindo como proteção das HSCs frente ao risco de serem fagocitadas durante períodos de migração.⁽³⁹⁾

Endoglin (Eng), ou CD105, aparecendo aqui através da interação Itgb1-Eng (12-c, comunicação via endotélio). Na verdade, os receptores não aparecem diretamente sendo expressos em HSCs na literatura, mas em células do estroma responsáveis por dar melhor apoio à hematopoiese por formar nichos mais apropriados.⁽⁴⁰⁾ Não é claro se o papel de sinalização responsável por fomentar este microambiente poderia ser parcialmente executado pelas próprias HSCs em um momento embrionário, mesmo assim, alguma conexão parece ser traçável.

Spp1, a última conexão encontrada na literatura, mas possivelmente a mais rica. Spp1 ou *Osteopontin* (OPN) apareceu três vezes na Figura 12-d (via estroma expandido), comunicando através de Spp1-Itgb1, Spp1-Cd44 e Spp1-Itga4. *Osteopontin* é uma glicoproteína de ação variada (adesão, inflamação, angiogênese, metástase),^(41,37) inclusive sendo responsável pela localização (*homing*) e proliferação de HSCs.⁽⁴¹⁾ Há evidências de que a sinalização por esta proteína seja responsável pelo controle da localização das HSCs, no evento da migração trans-medula até o nicho endosteal, além de servir como controle negativo da proliferação desta população celular.⁽⁴¹⁾ Também é reportada a diminuição de sua expressão em medulas ósseas de animais mais velhos, comprometendo a sua renovação e diferenciação rumo a um viés mielóide. Estes comprometimentos são parcialmente revertidos no evento de reposição extrínseca de OPN.⁽⁴²⁾

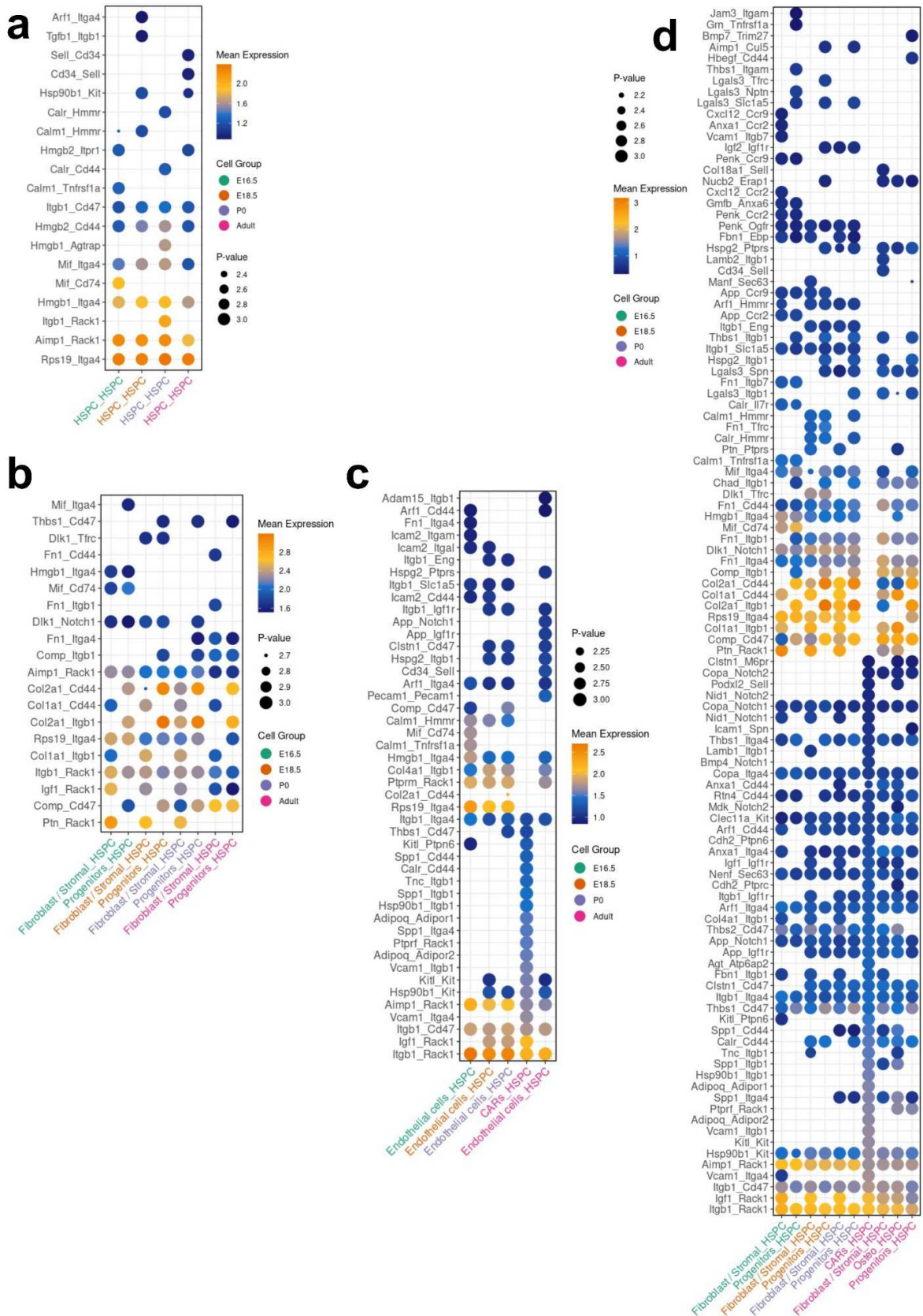
De forma conjunta, o achado destes genes acaba endossando uma pré-confirmação de que nossos resultados *in silico* de comunicação celular trazem informações consistentes com a literatura atual, no contexto de maturação do microambiente da medula. Os passos seguintes incluem uma passagem intermediária na análise de comunicação, onde é empregada a técnica de NMF para encontrar padrões em um número elevado de ligantes-receptores. Com a aplicação destes cálculos, cuja lógica fora sutilmente descrita na Figura 9, os resultados (Fig. 13) nos guiarão para a aplicação de novas análises derivadas (Fig. 14-17).

Figura 12 - Dados Shannon - *Dotplot* - Interações Autócrinas

Gráfico de dispersão dos pares de ligantes-receptores para tipos celulares selecionados associados à maturação do microambiente da medula óssea.

Componentes:

- Dotplot* de interações autócrinas;
- Dotplot* de provindas de fibroblastos e progenitores hematopoieticos;
- Dotplot* de interações do endotélio;
- Dotplot* de interações do componente estromal expandido.



Fonte: Figura gerada pelo autor (2025)

Foco em fatores que melhor diferenciam os *time-points*

Nesta última porção tratando propriamente dos resultados do trabalho, serão discutidos os achados ao aplicar a técnica de NMF usando os 150 pares de ligantes-receptores mais relevantes ao *dataset* de maturação da medula óssea murina. Primeiramente será descrita a lógica na escolha de ligantes-receptores (Figura 13), seguida da análise de enriquecimento de termos associados a estes genes (Figuras 14, 15 e 16), e finalizando com a visualização de alguns desses pares de comunicação expostos em eixos envolvendo fibroblastos, progenitores hematopoiéticos e células do endotélio (Figura 17).

A partir do *Dotplot* com os 150 pares de ligantes-receptores, análogo ao observado na Figura 9-a, mas substancialmente maior, tornando-o inapropriado de expor aqui e performar curadoria humana manual, foram gerados os cálculos de NMF, na busca por padrões que ajudassem a compreender as diferenças de comunicações presentes entre os estágios de maturação do microambiente da medula. Os resultados que nos ajudam a compreender estes padrões são expostos primeiramente na Figura 13-a, onde podemos observar que os fatores 1, 3 e 6 estão “enriquecidos” entre os *time-points* observados pelo estudo. O fator 1 traz padrões “intermediários”, associados a tipos celulares dos tempos E18.5 e P0. Já o fator 3 traz padrões “maduros”, associados a tipos celulares do tempo adulto, enquanto o fator 6 apresentaria o padrão mais “imaturo”, enriquecido por pares celulares do primeiro tempo observado, E16.5.

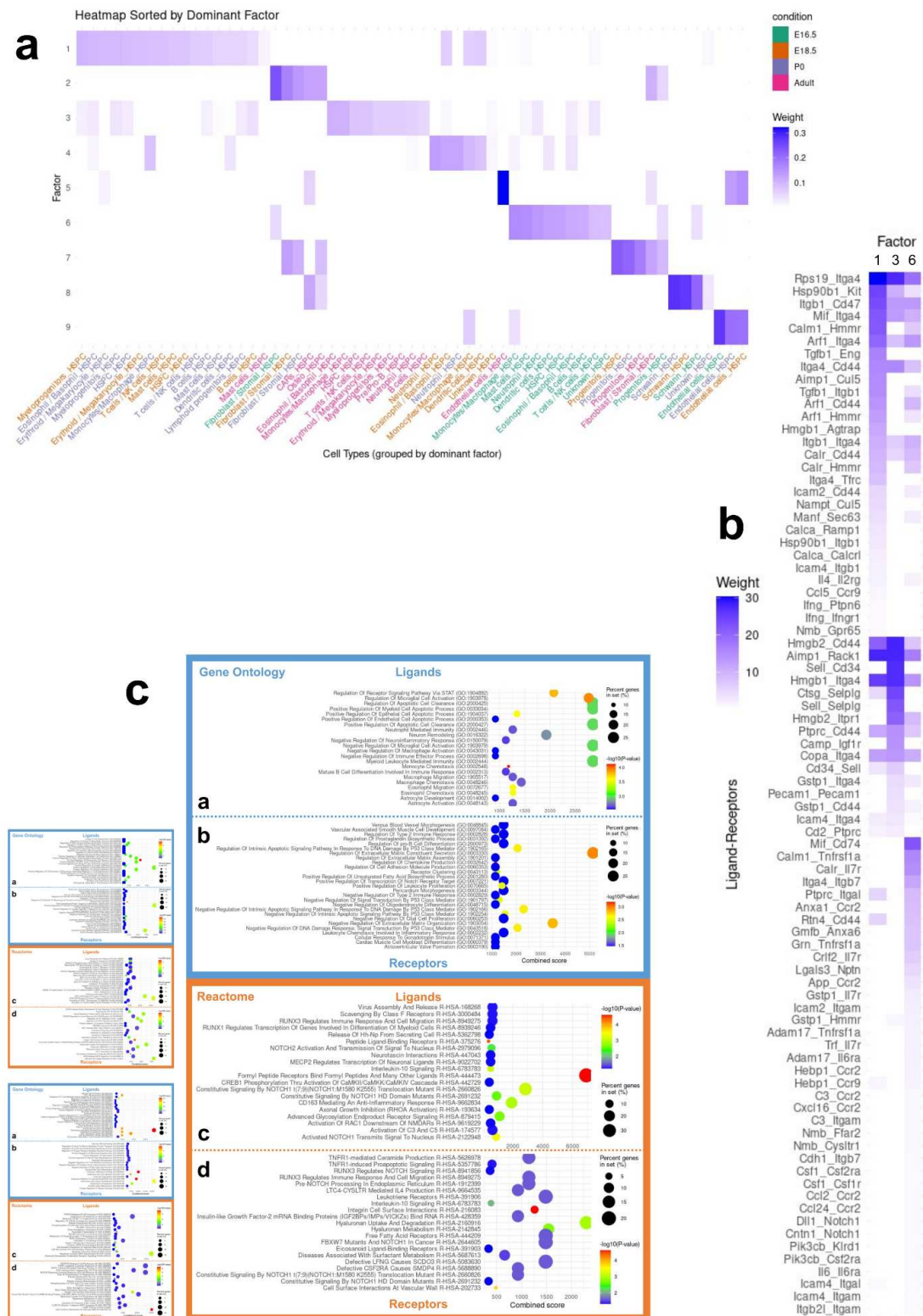
Com o destaque destes fatores, o próximo passo lógico seria a identificação dos ligantes-receptores associados a tais fatores, como podemos ver na Figura 13-b. Como podemos observar, os pares de comunicação selecionados possuem uma associação maior com um dos fatores, mas estão também associados, em algum nível, aos outros 2. Desta forma, não é incomum ver que as comunicações associadas ao fator 3, pares “maduros”, vinham crescendo dos estágios anteriores, normalmente menos fortes para o fator 6 (imaturo) que para o 1 (intermediário), por exemplo. E a mesma narrativa pode ser feita para os ligantes-receptores dos demais fatores.

Com o conjunto de genes de ligantes e receptores em mãos, a próxima etapa foi investigá-los; para tal, duas abordagens foram seguidas: gerar o enriquecimento de termos desses genes, como (meramente) ilustrado no componente c da Figura 13 (adequadamente exposto nas Figuras 14-16 adiante), e explorar graficamente o comportamento destes pares em gráficos de comunicação celular. A Primeira abordagem resultou nos próximos três conjuntos de figuras, 14-16, enquanto os últimos gráficos de comunicação celular podem ser observados na Figura 17.

Figura 13 - Dados Shannon - Paineis Esquemáticos da Aplicação de NMF com 150 pares
Conjunto de gráficos em um arranjo esquemático representando os componentes associados à análise de NMF usando os top 150 pares de interação em cada par de tipos celulares em comunicação.

Componentes:

- Matriz expressa em um *Heatmap* com fatores (linhas) por pares de tipos celulares interagindo (colunas). Percebe-se uma correlação maior dos fatores 1, 3 e 6 com *time-points* diferentes, E18.5 e P0, Adult e E16.5, sugerindo que podem ser úteis na identificação de padrões na maturação do microambiente;
- Matriz expressa em um *Heatmap* com ligantes-receptores (linhas) por fatores (colunas). Recorte dos fatores focando apenas nos três fatores associados “estágios de maturação”, 1, 3 e 6. Valores de preenchimento são oriundos da raiz quadrada da multiplicação entre a expressão média dos ligantes-receptores e o seu respectivo $-\log_{10}$ p-valor, em cada par de tipos celulares interagindo;
- Overview* das análises de enriquecimento de termos (GSEA) oriundos dos genes demonstrados no componente b. Gráficos encontram-se aqui por contextualização apenas, serão propriamente expostos e explorados nas Figs 14-16.



Fonte: Figura gerada pelo autor (2025)

Nas próximas figuras veremos os resultados da análise de enriquecimento de termos dos genes de ligantes e receptores associados a cada fator previamente identificado. As figuras estão em ordem de estágio de maturação, ou seja: a Figura 14 é associada ao fator 6, enriquecendo termos associados aos pares de comunicação do *time-point* E16.5; a Figura 15 é associada ao fator 1, enriquecendo termos do estágio “intermediário”, com comunicações dos pontos E18.5 e P0; e por último, a Figura 16 provém os termos associados às interações da medula adulta, “madura”.

Como podemos ver na Figura 14, trazendo o enriquecimento dos genes de ligantes-receptores do fator 6, há uma prevalência geral de termos associados às vias de RUNX e NOTCH, visível nos componentes b-c-d da figura. Termos associados a apoptose também são numerosos, como podemos ver expressivamente no componente 14-a. Há também alguns termos associados à matriz-extracelular, inclusive regulação negativa desta, como podemos ver na Figura 14-b. Alguns termos com destaques imunes são a sinalização de IL10, na Figura 14-c-d, bem como a regulação negativa de resposta imune de tipo 2 na Figura 14-b e resposta antiinflamatória via CD163 na Figura 14-c. Também há diversos termos associados à quimiotaxia de células imunes, especialmente de tipos celulares da imunidade inata, monócitos, macrófagos e eosinófilos, visível na Figura 14-a.

Já na Figura 15, vemos os termos associados aos genes do perfil “intermediário”, do fator 1. Os sinais de migração permanecem, como podemos ver na Figura 15-a; o número de termos é mais limitado, mas ainda presente, com sinal consideravelmente elevado para migração de fibroblastos. O mesmo pode ser dito da sinalização via RUNX, embora também mais restritos, termos são observados na Figura 15-c-d. Outros termos associados ao contexto imune, como sinalização via TGF-beta e IL-18 podem ser vistos na Figura 15-c. Há mais conexões aparecendo com a matriz celular, como podemos ver na Figura 15-a-c-d, desta vez tanto para ácido hialurônico, quanto para fibronectina. Bem como uma expansão no número de termos referentes a proliferação endotelial, células da musculatura de vasos e produção de óxido-nítrico (15-a-b). Por fim, vemos também a sinalização da via de receptores de calcitonina.

Terminando com a Figura 16, com os resultados do enriquecimento dos ligantes-receptores do estágio mais maduro, observamos alguns termos inéditos aos previamente trazidos, bem como a permanência de outros. Ativação de RUNX3 aparece novamente, mas substancialmente menos destacada que anteriormente (Fig. 16-d), o mesmo pode ser observado referente ao ácido hialurônico (Fig. 16-d). As interações com a

vasculatura (16-c-d) e desenvolvimento de endotélio (Fig. 16-a) tomaram o lugar dos termos de proliferação endotelial. Em sinais antes mais sutis (Fig. 15-c), agora termos associados à motilidade de leucócitos se destacam dos demais (Fig. 16-b). Termos associados a apoptose (Fig. 16-c) e TNFR1 (Fig. 16-d) aparecem de forma mais expressiva. Por fim, regulação de proliferação de células tronco, bem como alguns termos sinalizadores de processos mais maduros aparecem, como recombinação V(D)J, proliferação de leucócitos, ativação de CD4 e regulação de sinalização de antígenos (Fig. 16-a).

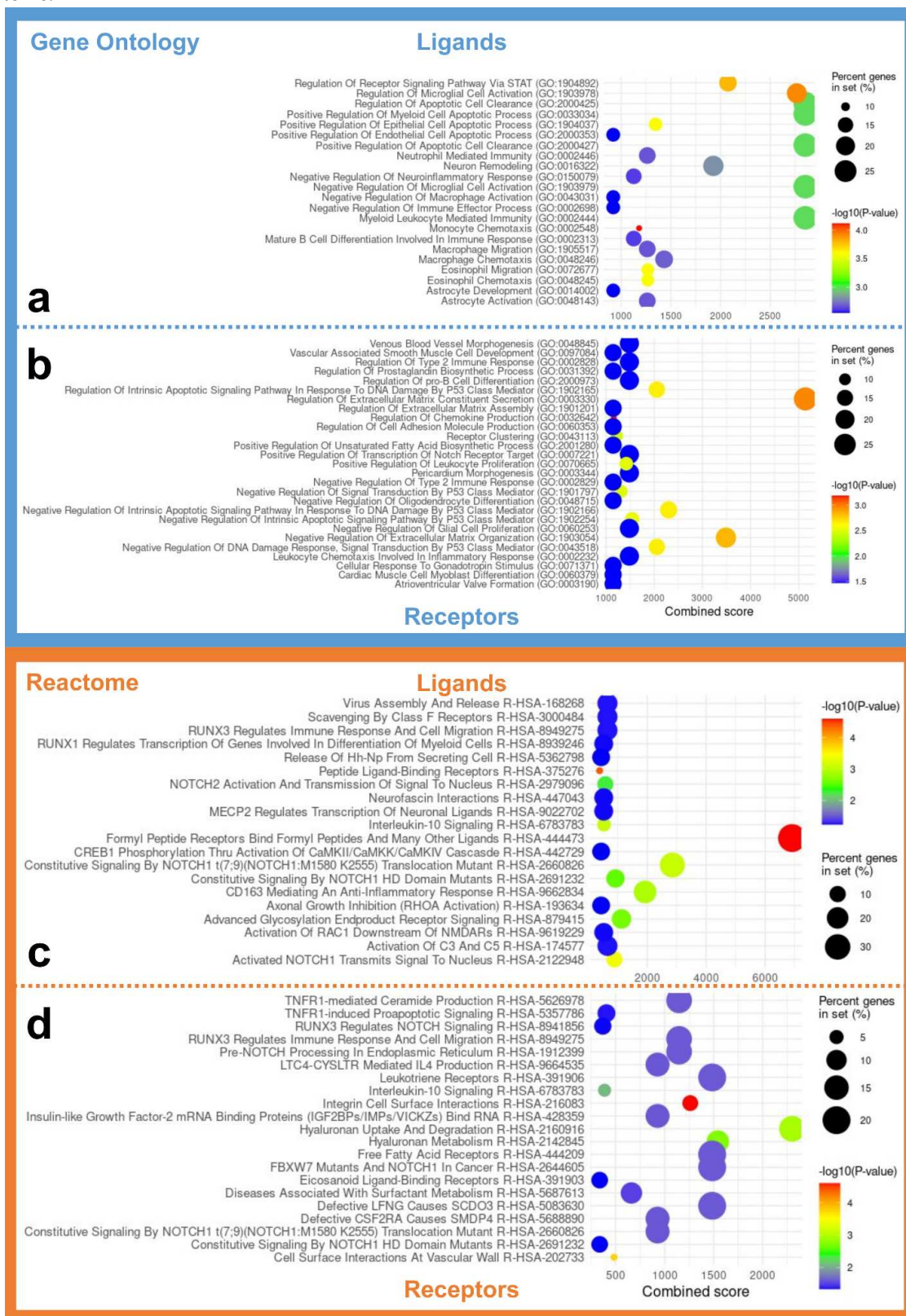
De forma geral o panorama aqui exposto pode ser interpretado da seguinte forma: os sinais de quimiotaxia/migração celular prevalecem no primeiro estágio, mas se mantêm sutilmente ainda no período intermediário, especialmente associado a fibroblastos. Sinais como NOTCH e RUNX3 seguem o mesmo padrão, começando com maior ênfase e terminando como um sinal mais brando. Os agentes aparentemente mais expressivos da maturação parecem ser os componentes da matriz extracelular (ECM), que foram inicialmente regulados de forma negativa, e expressamente expandidos no fator intermediário, com destaque para a atuação do ácido hialurônico. Somada à influência da matriz temos a ação das células endoteliais e toda a vasculatura associada, também com expansão intermediária e maturação no fator final. Sendo assim, há indicações de que as comunicações com células endoteliais e fibroblastos parecem ter um papel relevante na narrativa de maturação do microambiente da medula óssea, bem como nas comunicações entre as células HSPCs e a ECM. Com isso em mente, foram gerados gráficos de comunicação celular filtrando por essas populações e usando os ligantes-receptores responsáveis pelos fatores 1, 3 e 6.

Figura 14 - Dados Shannon - *GSEA* - Ligantes-Receptores associados ao Fator 6 (*Imaturo*)

Conjunto de gráficos de dispersão (enriquecimento de termos). Termos enriquecidos no eixo-y, tamanho traz a % de genes associados ao termo, cor traz relevância estatística, posição do eixo-x traz escore misto destes valores;

Componentes:

- Enriquecimento de termos a partir dos genes de **ligantes** considerados “imatuross” (fator 6), usando banco do **Gene Ontology**;
- Enriquecimento de termos a partir dos genes de **receptores** considerados “imatuross” (fator 6), usando banco do **Gene Ontology**;
- Enriquecimento de termos a partir dos genes de **ligantes** considerados “imatuross” (fator 6), usando banco do **Reactome**;
- Enriquecimento de termos a partir dos genes de **receptores** considerados “imatuross” (fator 6), usando banco do **Reactome**.



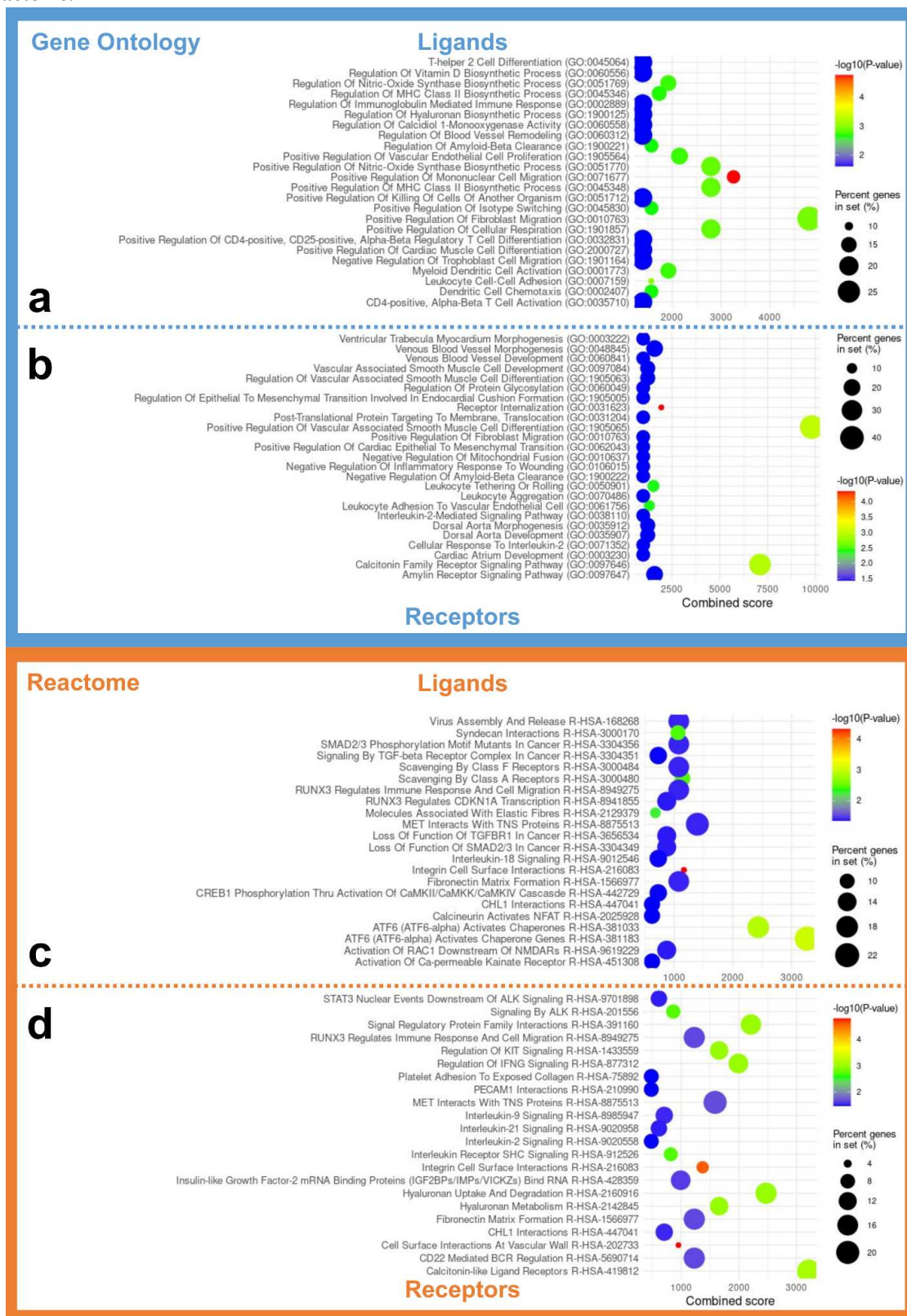
Fonte: Figura gerada pelo autor (2025)

Figura 15 - Dados Shannon - *GSEA* - Ligantes-Receptores associados ao Fator 1 (*Intermediário*)

Conjunto de gráficos de dispersão (enriquecimento de termos). Termos enriquecidos no eixo-y, tamanho traz a % de genes associados ao termo, cor traz relevância estatística, posição do eixo-x traz escore misto destes valores;

Componentes:

- Enriquecimento de termos a partir dos genes de **ligantes** considerados “intermediários” (fator 1), usando banco do **Gene Ontology**;
- Enriquecimento de termos a partir dos genes de **receptores** considerados “intermediários” (fator 1), usando banco do **Gene Ontology**;
- Enriquecimento de termos a partir dos genes de **ligantes** considerados “intermediários” (fator 1), usando banco do **Reactome**;
- Enriquecimento de termos a partir dos genes de **receptores** considerados “intermediários” (fator 1), usando banco do **Reactome**.



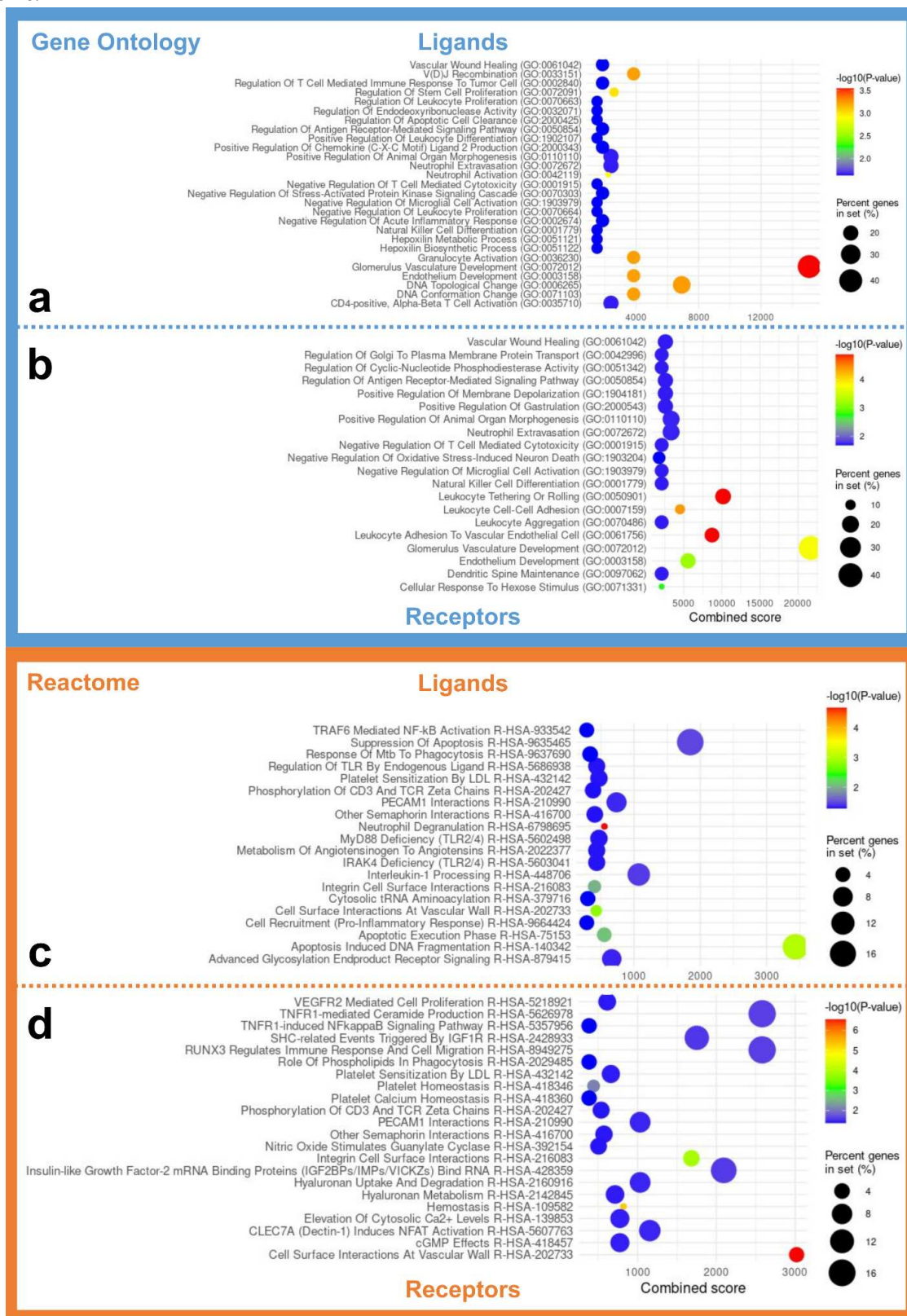
Fonte: Figura gerada pelo autor (2025)

Figura 16 - Dados Shannon - *GSEA* - Ligantes-Receptores associados ao Fator 3 (*Maduro*)

Conjunto de gráficos de dispersão (enriquecimento de termos). Termos enriquecidos no eixo-y, tamanho traz a % de genes associados ao termo, cor traz relevância estatística, posição do eixo-x traz escore misto destes valores;

Componentes:

- Enriquecimento de termos a partir dos genes de **ligantes** considerados “maduro” (fator 3), usando banco do **Gene Ontology**;
- Enriquecimento de termos a partir dos genes de **receptores** considerados “maduro” (fator 3), usando banco do **Gene Ontology**;
- Enriquecimento de termos a partir dos genes de **ligantes** considerados “maduro” (fator 3), usando banco do **Reactome**;
- Enriquecimento de termos a partir dos genes de **receptores** considerados “maduro” (fator 3), usando banco do **Reactome**.



Fonte: Figura gerada pelo autor (2025)

Chegando, enfim, aos últimos resultados a se discutir, temos a Figura 17, onde observamos como os padrões de comunicação dos ligantes-receptores associados aos fatores 1, 3 e 6, comportam-se ao longo dos diferentes tempos ao focarmos em dois grupos se comunicando com as células HSPCs.

Como foram incluídos apenas os pares associados aos fatores previamente citados, o conjunto de ligantes-receptores é consideravelmente mais limitado, porém tendem a ser mais informativos das diferenças entre estágios, como podemos ver na Figura 17. Dentre ambos componentes, alguns pares de interação foram selecionados para maior investigação, sendo eles: na comunicação com fibroblastos e progenitores (17-a), Aimp1-Cul5, Cd34-Sell, Calm1-Hmmr, Calr-Hmmr, Calr-Cd44; e Tgfb1-Eng, Arf1-Hmmr, Cd34-Sell, Pecam1-Pecam1, Calr-Cd44, Hsp90b1-Kit, na Figura 17-b, tratando das comunicações com os tipos celulares endoteliais e do tipo “osteo”.

Dos 11 pares elencados, alguns já haviam surgido na seção anterior, “Explorando interações iniciais entre HSPCs e outros tipos celulares”, que tratou das análises de comunicação pré-NMF. Essa intersecção foi boa, pois trouxe alguns genes (Cd44 e Eng) que inclusive foram previamente validados na literatura, denotando a consistência da presente análise.

Infelizmente a maior parte dos genes encontrados não possui relatos claros de sua relação com o desenvolvimento de HSCs ou do microambiente da medula óssea. Além dos dois previamente descritos, apenas Pecam1 e Sell tiveram alguma correlação descrita, mas de forma sutil.

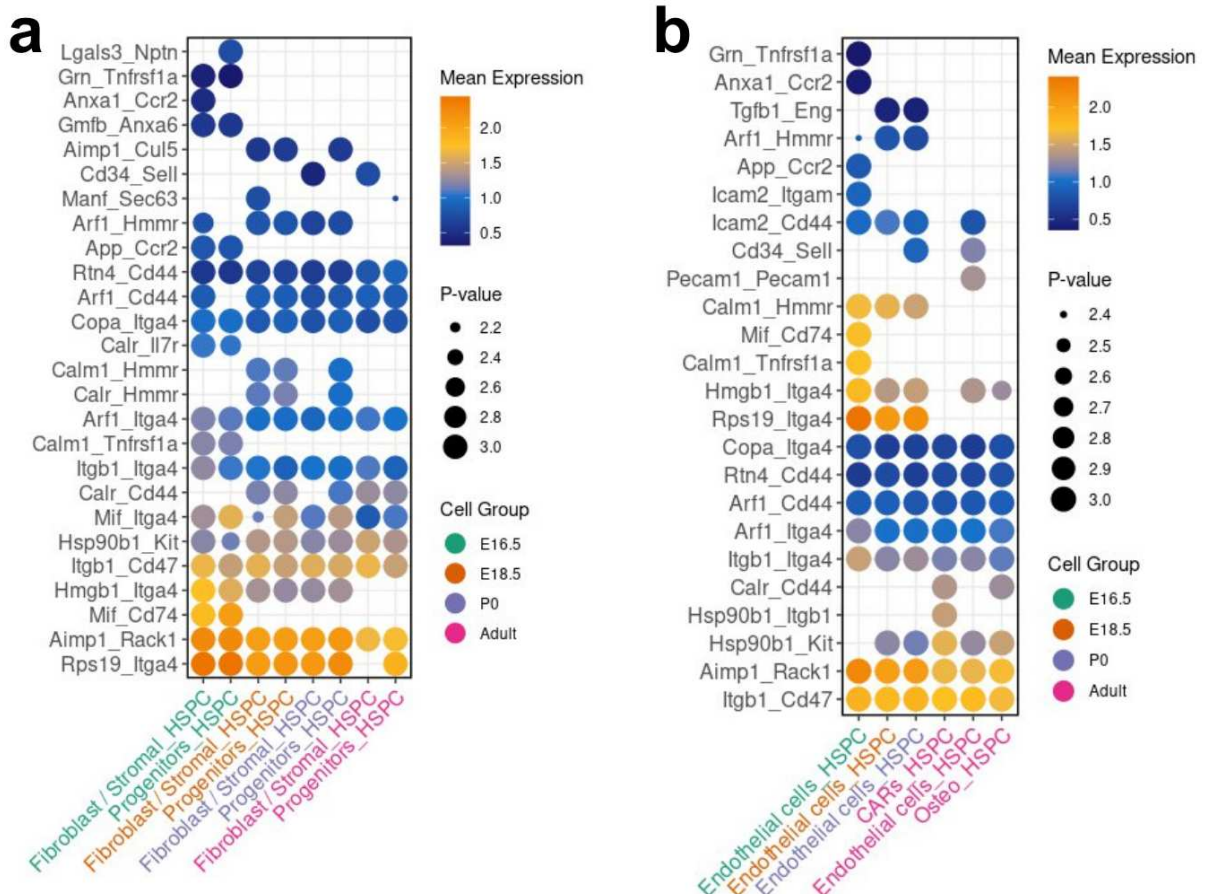
Primeiramente Platelet endothelial cell adhesion molecule-1 (PECAM-1), ou CD31, parece ter um papel positivo na diferenciação de HSCs rumo a células da linhagem eritróide, e um negativo em ramos linfóides, em especial células B.^(43,44)

Por fim, Sell, ou L-selectin, não possui sua relação bem descrita, inclusive há relatos conflitantes acerca de sua natureza ser pró-manutenção ou pró-proliferação de HSCs. Mas parece certo que a proteína tem algum papel na homeostasia das células tronco hematopoiéticas.^(45,46)

Figura 17 - Dados Shannon - *Dotplot* - Ligantes-Receptores associados aos Fatores 1, 3 e 6
Paineis de gráficos de dispersão. Ligantes-receptores (eixo-y) por pares celulares comunicando (eixo-x), tamanho traz a relevância estatística, cor traz a expressão média de ligante e receptor entre o par celular;

Componentes:

- Dotplot* de fibroblastos e progenitores interagindo com HSPCs, sobre os eixos dos ligantes-receptores referentes aos fatores 1, 3 e 6;
- Dotplot* de populações endoteliais e linhagem óssea interagindo com HSPCs, sobre os eixos dos ligantes-receptores referentes aos fatores 1, 3 e 6.



Fonte: Figura gerada pelo autor (2025)

3. LIMITAÇÕES

Melhorias a serem feitas no fluxo de trabalho de comunicação espacial

Não há muito o que se discutir dos trabalhos sobre o conjunto de dados de TSKs, apenas ressaltar o fato de que: como o Seurat não faz a importação da posição espacial dos *spots*, como o CellComm faz em seu lançamento original, melhorias devem ser pensadas sobre como tentar extrair essa informação a partir dos dados naturalmente extraídos pelo Seurat. Isto porque as funcionalidades da presente atualização dependem do acesso aos dados de processamento do *software* Cell-Ranger, oriundos do sequenciamento, o que pode nem sempre estar disponível publicamente, ou não ser de tão fácil acesso, no mínimo.

Reprodutibilidade dos dados de TSKs

Houve dois pontos diferentes que dificultaram a reprodutibilidade dos resultados no dataset de TSKs. O primeiro, e mais trivial, foi o fato de que os autores tiveram problemas no depósito dos arquivos no GEO; os arquivos estão lá, mas há algo corrompido ou incorreto nos mesmos, o que impede o objeto Seurat de ser criado. Para solucionar essa questão, foi necessário ter acesso a outros arquivos tecnicamente não públicos, fornecidos pelos autores após contato individual. O segundo, e mais complicado, foi relacionado à anotação dos *clusters* espaciais. Não foi fornecido pelos autores um arquivo referente aos metadados anotados do trabalho deles, o que fez com que não fosse possível recriar as comunicações tais quais as originais. Foi usada uma assinatura para classificar os *spots*, mas no final, a análise sobre este *dataset* não era de grande relevância, acabando por servir apenas como forma de verificar se os códigos associados à análise espacial estavam funcionando corretamente.

Reprodutibilidade dos dados de AGM

Referente aos dados de AGM houve um ponto, apenas, de incongruência entre as análises, mas de maior seriedade que a anteriormente trazida sobre TSKs. Como apresentado anteriormente através das Figuras 2 e 3, houve incongruências na tentativa de replicar os mesmos achados feitos pela publicação original do CellComm.⁽¹⁴⁾ Embora os valores das comunicações em comum se mantenham comparáveis, houve o desaparecimento de vários ligantes-receptores. A diferença sutil nos valores é provável que tenha ocorrido pela mudança de configurações no *pipeline* de pré-processamento, ou seja, na forma de normalização, e escala dos dados de expressão. No entanto, a ausência de pares de interação é um problema mais grave, e afeta a reprodutibilidade dos resultados.

Ao conduzir uma busca rumo à raiz do problema dos ligantes-receptores desaparecidos, foi verificado que alguns dos pares nem sequer constavam na tabela de importação do NicheNet pós conversão murina, e foi então que tornou-se mais claro qual poderia ser o problema. A listagem de pares de ligantes-receptores usada pelo CellComm, até então sempre veio de uma lista depurada pelos autores do NicheNet.⁽¹³⁾ Acontece que nessa tabela, os pares de genes são humanos, assim sendo, para trabalhar sobre um *dataset* de camundongo, ou tecnicamente qualquer outra espécie, é necessário converter os genes de humanos para os homólogos da espécie em questão. O problema provavelmente ocorreu, pois a forma com que os genes foram convertidos na análise original do CellComm, em 2022, e a iteração referente ao presente trabalho, não captaram o mesmo conjunto de

ligantes-receptores. Em conversa com o Prof. Edroaldo, responsável por essa parte no artigo original, foi constatado o fato de que a lista de conversão usada por ele era diferente da empregada por mim através da função `GetHumanMouseHomoLogs()`, onde a conversão foi feita via Ensembl.⁽⁴⁷⁾ Novas formas de efetuar a conversão serão testadas e possivelmente incorporadas no futuro para garantir essa retrocompatibilidade.

Reprodutibilidade nos dados da Shannon

Houve uma série de problemas associados à reprodutibilidade dos resultados do artigo original de amadurecimento da medula óssea.⁽¹⁶⁾ Antes de falar dos problemas, deve-se reconhecer o esforço dos autores em deixar boa parte do código de processamento disponível em um GitHub,⁽⁴⁸⁾ embora os códigos fossem extremamente complexos e os *scripts* de processamento extensamente longos, o que é de se esperar de uma análise tão grande. Mas, retomando os problemas encontrados, podemos dividi-los em quatro principais, de mais simples a mais complexo.

Primeiro, algumas escolhas heterodoxas no processamento dos dados, como por exemplo, processar os dados individualmente sobre cada *time-point*, usando números de *Principal Components* (PCs) diferentes para cada análise.

Segundo, considerar que as amostras que compõem o *dataset* não possuíam efeito de *batch*, sem argumentação ou evidência alguma dessa afirmação. Isto demonstrou-se, aparentemente, um erro, através das análises feitas pelo presente trabalho, ao se identificar *batch effect* entre algumas amostras. E, naturalmente, a correção desse efeito pode ter trazido mudanças nos resultados finais.

Terceiro, não fica bem claro como funciona a relação do *software* empregado pelos autores, o RNA-Magnet,⁽⁴⁹⁾ e a anotação celular encontrada no arquivo de metadados (encontrado no GEO), que não segue os mesmos nomes usados na publicação do trabalho em questão. Como as anotações fornecidas e as aparentemente usadas na geração dos principais *plots* não se conversam, fica difícil ter confiança na solidez da anotação fornecida. No futuro, será necessário aplicar-se análises de verificação de consistência da anotação dos *clusters* fornecidos, provavelmente usando a ferramenta Single Cell Net⁽⁴⁾ e algum dataset de células da medula, devidamente anotado.

O quarto e último problema associado à reprodutibilidade, está no fato de que o *software* RNA-Magnet performa a análise de comunicação celular muito diferentemente do CellComm; pelo menos, pela forma empregada pelos autores, fica difícil a reprodução dos

mesmos resultados. Sem entrar em muitos pormenores, a análise originalmente empregada computa a comunicação primeiramente num *pool* de células “*pseudo-bulk*” do *time-point* como um todo, sem discernir tipos celulares enviando o sinal; depois divide-se em quais tipos celulares que aparentemente são responsáveis pelas comunicações em questão. Essa abordagem é consideravelmente diferente da performada pelo CellComm, então, é possível que o sumiço de parte dos ligantes-receptores, entre a publicação original e a nossa reanálise, que pode ser observado entre as Figuras 10 e 11, possa ser explicado por essa forma diferenciada de computar as comunicações.

Compleitude das capacidades de análise do CellComm

Como conhecedores do artigo original do CellComm podem perceber, não houve resultado algum demonstrado no presente trabalho, que explice o fator inicialmente inédito da publicação original na revista Nature Cell Biology, as análises de comunicação *downstream*. A funcionalidade de identificar quais Fatores de Transcrição (TFs) estão sendo acionados, através da comunicação superficial ligante-receptor somada à possível cascata de sinalização de sinal, não foi agraciada por devida adaptação de suas funções, no que tangem os esforços do presente trabalho. Tal funcionalidade foi devidamente mapeada para que se conheçam quais funções devem ser adaptadas e transformadas em sub-funções dentro de agregadoras, como feito no caso da `ComputeInteractions()`. Porém não houve tempo hábil para se trabalhar sobre as adaptações necessárias, além do fato de que é nessa porção de código que residem as funções que efetuam cálculos usando Java, e trabalhar sobre essas funções requer conhecimentos fora das habilidades atuais do presente autor. Contudo, deve-se anunciar que a completude dos trabalhos rumo ao CellComm v2, por mais que não atingida com o agora, tornou-se uma realidade mais tangível com a conclusão dos trabalhos aqui obtidos, e espera-se ser alcançada em um horizonte próximo.

PERSPECTIVAS

Considerando os pontos previamente relatados, tanto no que tange achados do presente, quanto limitações descritas, novos direcionamentos hão ser traçados. A respeito de possíveis genes de ligantes-receptores encontrados aqui, experimentos que os validem em modelos *in vitro* e *in vivo* devem ser propriamente desenhados. Enquanto que as atuais limitações do *software* devem ser trabalhadas em futuros esforços, desde identificar possíveis

problemas na conversão de genes, à completude na adaptação do módulo de análises intracelulares do CellComm.

Validação de resultados - EHT

Referente à validação dos achados, sugere-se o desenho de experimentos usando estratégias de análise de expansão, competência de diferenciação e perfil transcricional das células alvo de cada contexto biológico trabalhado (EHT e maturação da medula). Primeiramente para o escopo de EHT: 1) a avaliação da influência dos ligantes (Igf1, Igf2, Hgmb1, Hgmb2, Calr, Calm1, Hsp90b1 e App) e receptores (Igf1r, Agtrap, Itga2b, F2lr2, Itga6 e Hmnr), pode ser testada através do impacto de *Morpholino-based Knockdown* em modelo *Zebrafish*, seguidos por análise *Whole-mount in situ hybridization* (WISH) da expressão de *runx1* e *cmyb* em cortes histológicos dos animais; 2) Conforme os resultados, poderia-se partir para ensaios de cultura e diferenciação de *Hemogenic Endothelium* (HE) a partir de *human induced Pluripotent Stem Cells* (hiPSC), seguindo de diferenciação guiada para perfis pro-T, usando estratégias similares às descritas originalmente.⁽¹⁴⁾ Nesta etapa seriam aplicadas intervenções químicas para inibir/estimular os ligantes-receptores, sendo que, dos citados, apenas Calr, Agtrap, Itga6 e Hmnr não são molecularmente acionáveis segundo o DrugBank,⁽⁵⁰⁾ porém, estratégias de siRNA ainda seriam cabíveis. Culminando, por fim, em ensaios de *colony-forming unit* (CFU) para identificar o potencial de expansão e diferenciação dos diferentes grupos avaliados.

Validação de resultados - Maturação do Microambiente da Medula Óssea

Buscando a validação dos genes no contexto de maturação de medula óssea, há, naturalmente, diferentes formas de validar os achados, duas delas são: avaliar a capacidade de suporte do nicho, e contagem de CFUs a partir de células individuais MPP2 e LT-HSC. A segunda abordagem é mais simples, mas potencialmente apenas viável para testar o potencial de ligantes na capacidade das células de se diferenciar. A primeira é parecida, mas mais trabalhosa, necessitando de uma etapa anterior de co-cultura antes do isolamento para verificar o potencial de diferenciação através de CFUs. Conforme tempo e recursos, a escolha da abordagem simples seria mais cabível, mas a primeira parece que seria mais robusta para os testes de validação. A título de simplicidade, tratarei apenas da primeira abordagem a seguir.

Para melhor avaliar o efeito dos genes encontrados, um experimento em camundongos plausível de se aplicar consistiria em: 1) isolar LT-HSCs do fígado fetal nos tempos E16.5 e P0; 2) aplicar ambos grupos celulares a dois tipos diferentes de meio, co-cultura com células do compartimento estromal providas dos tempos E16.5 e P0, totalizando 4 grupos até então; 3) submeter parte das amostras de cada grupo a intervenções moleculares/farmacológicas; 4) avaliar potencial de diferenciação através de ensaios de CFUs de cada grupo experimental.

Para a escolha de métodos para intervenção molecular, infelizmente, não foram encontradas *small molecules* disponíveis que tenham os genes previamente encontrados (CD44, Calr, Spp1, Fn1, CD47, Comp, Clst1, Endoglin, Hmgb1, Agtrap, Hmnr, Pecam1) como seus alvos. Desta forma, abordagens biológicas teriam de entrar em cena. Há alguns anticorpos que possuem algumas dessas proteínas de superfície como alvo, mas a forma mais segura, e potencialmente mais barata, seria usar siRNAs para avaliar o impacto do *knockdown* destes genes. Assim, após a administração em cada grupo testável, ao comparar os resultados da expansão de CFUs, seria possível verificar o real impacto destes genes ligantes-receptores no processo de maturação da medula óssea rumo a fomentar uma hematopoiese madura.

Aprofundando análises *in silico*

De posse das melhorias vindouras ao *software* CellComm, mais precisamente com integridade da adaptação do módulo de sinalização intracelular, enriquecimento de TFs em vias de diferenciação, somados à integração conjunta destas capacidade, novos experimentos *in silico* poderiam ser aplicados. Tais experimentos consistiriam na busca por comunicações celulares que tivessem como alvo de sinalização TFs enriquecidos nas trajetórias de cada contexto biológico. Ou seja, poderíamos identificar quais ligantes-receptores seriam mais prováveis de afetar as células alvo diretamente no escopo da diferenciação celular sofrida pela EHT, e quais afetariam mais o perfil de HSPCs rumo ao estado maduro, no contexto de medula óssea.

CONCLUSÃO

Ao que tudo indica, o presente trabalho conseguiu completar o que se propôs inicialmente cumprir: a expansão do CellComm para inferir a dinâmica ou diferenças de comunicação celular ao longo do tempo ou condições experimentais. Havendo dois objetivos específicos considerando a “renovação”, seriam estes a simplificação do *pipeline* de análises e

adaptação ao *framework* do Seurat. E também os objetivos de incremento através das funcionalidades de NMF, e demonstrações de uso em *datasets* de diferentes naturezas.

Acerca do aspecto “renovação”, podemos explicar os avanços de forma prática: antes o *pipeline* para chegar num resultado visual da comunicação celular (com ligantes-receptores e pares de tipos celulares interagindo), levava entre 11-14 linhas/comandos diferentes de códigos a se executar, com diferentes funções e parâmetros a se organizar. Agora, esse mesmo processo é feito em apenas duas linhas de código, três se quisermos avaliar quais populações possuem maior índice de comunicação, incluindo a própria criação do resultado visual em questão. Além da facilidade e simplicidade, o código está mais robusto e completo para lidar com imprevistos, como ausência de parâmetros que tratam da natureza dos dados serem de *single-cells* ou *spots*, a espécie ser *Homo sapiens* ou *Mus musculus*. A implementação também deixou lacunas propositais, de forma que o código funcione modularmente, viabilizando a implementação de novas funcionalidades no futuro mais facilmente.

Tratando da tarefa de adaptação, vale ressaltar que toda a porção associada à comunicação superficial de ligantes-receptores está agora compatível com o *framework* do Seurat, além de ter sido mantida retrocompatibilidade com CellRouter. Sendo testado e assegurado o funcionamento em cenários com dados de diferentes naturezas, *seamlessly*, do clássico scRNA-seq à ST, os quais foram agraciados também com 3 tutoriais atualizados para a nova implementação do pacote. No que tange o desenvolvimento de novas funcionalidades, foi incorporada a análise de NMF ao *pipeline* do CellComm de forma integrada, com demonstração em dois dos tutoriais disponíveis.

No presente trabalho foram apontadas lacunas e limitações que demonstram, no estado atual, serem incompatíveis com o título de “CellComm v2”. Contudo, caminhos para tal foram traçados, inclusive na forma de experimentos para validação dos achados e próximos passos no desenvolvimento do pacote. Assim, em virtude das melhorias, novas capacidades e futuras perspectivas, expõe-se o CellComm v2 para análises sistemáticas de processos de comunicação celular a partir de dados de scRNA-seq e transcriptômica espacial.

“A dream without a plan is just a wish”

~ Katherine Paterson

REFERÊNCIAS

- [1] TANG, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. **Nature Methods**, v. 6, n. 5, p. 377–382, maio 2009.
- [2] SHALEK, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. **Nature**, v. 498, n. 7453, p. 236–240, jun. 2013.
- [3] CAHAN, P. et al. CellNet: Network Biology Applied to Stem Cell Engineering. **Cell**, v. 158, n. 4, p. 903–915, 14 ago. 2014.
- [4] TAN, Y.; CAHAN, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. **Cell Systems**, v. 9, n. 2, p. 207–213.e2, 28 ago. 2019.
- [5] GAO, X. et al. TGF- β inhibitors stimulate red blood cell production by enhancing self-renewal of BFU-E erythroid progenitors. **Blood**, v. 128, n. 23, p. 2637–2641, 8 dez. 2016.
- [6] SHALEK, A. K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. **Nature**, v. 510, n. 7505, p. 363–369, jun. 2014.
- [7] SATIJA, R. et al. Spatial reconstruction of single-cell gene expression data. **Nature Biotechnology**, v. 33, n. 5, p. 495–502, maio 2015.
- [8] TRAPNELL, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. **Nature Biotechnology**, v. 32, n. 4, p. 381–386, abr. 2014.
- [9] CAO, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. **Nature**, v. 566, n. 7745, p. 496–502, fev. 2019.
- [10] LUMMERTZ DA ROCHA, E. et al. Reconstruction of complex single-cell trajectories using CellRouter. **Nature Communications**, v. 9, n. 1, p. 892, 1 mar. 2018.
- [11] WOLF, F. A.; ANGERER, P.; THEIS, F. J. SCANPY: large-scale single-cell gene expression data analysis. **Genome Biology**, v. 19, n. 1, p. 15, 6 fev. 2018.
- [12] VENTO-TORMO, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. **Nature**, v. 563, n. 7731, p. 347–353, nov. 2018.
- [13] 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, fev. 2020.
- [14] 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, abr. 2022.
- [15] EDROALDO. FUSCA. , 23 maio 2022. Disponível em: <<https://github.com/edroaldo/fusca>>. Acesso em: 28 maio. 2022

- [16] HALL, T. D. et al. Murine fetal bone marrow does not support functional hematopoietic stem and progenitor cells until birth. **Nature Communications**, v. 13, n. 1, p. 5403, 15 set. 2022.
- [17] JI, A. L. et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. **Cell**, v. 182, n. 2, p. 497- 514.e22, 23 jul. 2020.
- [18] **R: The R Project for Statistical Computing**. Disponível em: <<https://www.r-project.org/>>. Acesso em: 29 jul. 2025.
- [19] **VSCodium - Open Source Binaries of VSCode**. Disponível em: <<https://vscodium.com/>>. Acesso em: 29 jul. 2025.
- [20] **Posit. Posit**, [s.d.]. Disponível em: <<https://www.posit.co/>>. Acesso em: 29 jul. 2025
- [21] **A Grammar of Data Manipulation**. Disponível em: <<https://dplyr.tidyverse.org/>>. Acesso em: 29 jul. 2025.
- [22] **Obsidian - Sharpen your thinking**. Disponível em: <<https://obsidian.md/>>. Acesso em: 29 jul. 2025.
- [23] **Google Drive, OneDrive and Dropbox Syncing - Insync**. Disponível em: <<https://www.insynchq.com/>>. Acesso em: 29 jul. 2025.
- [24] LAB, S.-R. **OmniPath :: Intra- & intercellular signaling knowledge**. Disponível em: <<https://omnipathdb.org/>>. Acesso em: 29 jul. 2025.
- [25] FRENIS, K. et al. **Path of differentiation defines human macrophage identity**. bioRxiv, , 24 jan. 2025. Disponível em: <<https://www.biorxiv.org/content/10.1101/2025.01.24.634694v1>>. Acesso em: 29 jul. 2025
- [26] POLAŃSKI, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. **Bioinformatics**, v. 36, n. 3, p. 964–965, 1 fev. 2020.
- [27] OATLEY, M. et al. Single-cell transcriptomics identifies CD44 as a marker and regulator of endothelial to haematopoietic transition. **Nature Communications**, v. 11, n. 1, p. 586, 29 jan. 2020.
- [28] FENG, C. et al. Deciphering the evolving niche interactome of human hematopoietic stem cells from ontogeny to aging. **Frontiers in Molecular Biosciences**, v. 11, 4 dez. 2024.
- [29] YUAN, S. et al. High mobility group box 1 (HMGB1): a pivotal regulator of hematopoietic malignancies. **Journal of Hematology & Oncology**, v. 13, n. 1, p. 91, 13 jul. 2020.
- [30] SALATI, S. et al. Calreticulin Affects Hematopoietic Stem/Progenitor Cell Fate by Impacting Erythroid and Megakaryocytic Differentiation. **Stem Cells and Development**, v. 27, n. 4, p. 225–236, 15 fev. 2018.
- [31] SHIDE, K. et al. Calreticulin haploinsufficiency augments stem cell activity and is

- required for onset of myeloproliferative neoplasms in mice. **Blood**, v. 136, n. 1, p. 106–118, 2 jul. 2020.
- [32] YOUNG, K. et al. Decline in IGF1 in the bone marrow microenvironment initiates hematopoietic stem cell aging. **Cell Stem Cell**, v. 28, n. 8, p. 1473–1482.e7, 5 ago. 2021.
- [33] VALENTINIS, B.; BASERGA, R. IGF-I receptor signalling in transformation and differentiation. **Molecular Pathology**, v. 54, n. 3, p. 133–137, 1 jun. 2001.
- [34] LOU, Q. et al. IGF1R signaling in perinatal mesenchymal stem cells determines definitive hematopoiesis in bone marrow. **Blood**, v. 144, n. 26, p. 2773–2787, 26 dez. 2024.
- [35] ESPÍN-PALAZÓN, R. et al. Proinflammatory Signaling Regulates Hematopoietic Stem Cell Emergence. **Cell**, v. 159, n. 5, p. 1070–1085, 20 nov. 2014.
- [36] CAO, H. et al. The role of CD44 in fetal and adult hematopoietic stem cell regulation. **Haematologica**, v. 101, n. 1, p. 26–37, 1 jan. 2016.
- [37] ZANETTI, C.; KRAUSE, D. S. “Caught in the net”: the extracellular matrix of the bone marrow in normal hematopoiesis and leukemia. **Experimental Hematology**, v. 89, p. 13–25, 1 set. 2020.
- [38] NILSSON, S. K. et al. Hyaluronan is synthesized by primitive hemopoietic cells, participates in their lodgment at the endosteum following transplantation, and is involved in the regulation of their proliferation and differentiation in vitro. **Blood**, v. 101, n. 3, p. 856–862, 1 fev. 2003.
- [39] JAISWAL, S. et al. CD47 Is Upregulated on Circulating Hematopoietic Stem Cells and Leukemia Cells to Avoid Phagocytosis. **Cell**, v. 138, n. 2, p. 271–285, 23 jul. 2009.
- [40] CHAN, C. K. F. et al. Clonal precursor of bone, cartilage, and hematopoietic niche stromal cells. **Proceedings of the National Academy of Sciences**, v. 110, n. 31, p. 12643–12648, 30 jul. 2013.
- [41] NILSSON, S. K. et al. Osteopontin, a key component of the hematopoietic stem cell niche and regulator of primitive hematopoietic progenitor cells. **Blood**, v. 106, n. 4, p. 1232–1239, 15 ago. 2005.
- [42] GUIDI, N. et al. Osteopontin attenuates aging-associated phenotypes of hematopoietic stem cells. **The EMBO Journal**, v. 36, n. 7, p. 840–853, 3 abr. 2017.
- [43] BAUMANN, C. I. et al. PECAM-1 is expressed on hematopoietic stem cells throughout ontogeny and identifies a population of erythroid progenitors. **Blood**, v. 104, n. 4, p. 1010–1016, 15 ago. 2004.
- [44] JACKSON, D. E. et al. Platelet endothelial cell adhesion molecule-1 (PECAM-1/CD31) is associated with a naïve B-cell phenotype in human tonsils. **Tissue Antigens**, v. 56, n. 2, p. 105–116, 2000.
- [45] SESHADRI, M.; QU, C.-K. Microenvironmental regulation of hematopoietic stem cells and its implications in leukemogenesis. **Current Opinion in Hematology**, v. 23, n.

- 4, p. 339, jul. 2016.
- [46] WINKLER, I. G. et al. Vascular niche E-selectin regulates hematopoietic stem cell dormancy, self renewal and chemoresistance. **Nature Medicine**, v. 18, n. 11, p. 1651–1657, nov. 2012.
- [47] **Ensembl genome browser 114**. Disponível em:
<<https://www.ensembl.org/index.html?redirect=no>>. Acesso em: 29 jul. 2025.
- [48] **Hall-et-al/codes at main · trenthall612/Hall-et-al · GitHub**. Disponível em:
<<https://github.com/trenthall612/Hall-et-al/tree/main/codes>>. Acesso em: 29 jul. 2025.
- [49] 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, jan. 2020.
- [50] KNOX, C. et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. **Nucleic Acids Research**, v. 52, n. D1, p. D1265–D1275, 5 jan. 2024.