关于 DIALOGUE 从单细胞或空间转录组学数据中绘制多细胞程序的复现

王迪

摘要

破译组织中细胞的功能相互作用仍然是一个主要的挑战。在这里，我们描述了对话，一种系统地揭示多细胞程序（MCPs）的方法——不同细胞类型的协调细胞程序在组织水平上形成高阶功能单元的组合——从空间数据或没有空间信息获得的单细胞数据。在小鼠下丘脑、小脑、视觉皮层和新皮层的空间数据集上进行测试，对话确定了与动物行为相关的mcp，并在未知数据测试时恢复了空间属性，同时优于其他方法和指标。在来自人类肺癌的空间数据中，对话确定了标记免疫激活和组织重塑的 MCPs。应用于跨个体或区域的单细胞 RNA 测序数据，对话揭示了标记阿尔茨海默病、溃疡性结肠炎和癌症免疫治疗耐药性的 MCPs。这些项目可以预测独立队列中的疾病结局和易感性，并包括来自全基因组关联研究的风险基因。对话使多细胞的分析成为可能。

关键词：多细胞程序，单细胞 RNA 测序，空间数据

1 引言

这篇文章介绍了一种名为对话的方法，用于揭示多细胞程序（MCPs）的功能相互作用。在维持内稳态和疾病发展中，细胞之间的相互作用至关重要。然而，破译组织中细胞的功能相互作用仍然是一个主要的挑战。对话方法通过系统地分析单细胞数据，从空间数据或没有空间信息的数据中识别不同细胞类型的协调细胞程序，形成高阶功能单元的组合。该方法在小鼠和人类的多个组织和疾病中进行了测试，并取得了良好的效果。对话方法的应用为研究细胞间相互作用和组织生物学提供了新的途径，有望在诊断、治疗和预防疾病方面发挥重要作用。

在组织中，不同细胞之间的相互作用对于维持体内平衡至关重要。尽管许多疾病传统上被认为是特定细胞或细胞类型的故障，但越来越多的证据 [1] 和新的治疗策略已经证明了多细胞行动在健康和疾病方面的关键作用，为干预、诊断、疾病监测和预防提供了新的机会。与此同时，单细胞 RNA 测序（scRNA-seq）和空间转录组学的进展现在允许在细胞类型、组织和疾病状态中进行系统地探索。然而，尽管取得了这些进展 [2]，破译多细胞调控仍然是一个挑战，限制了从以细胞为中心到以组织为中心的视角移动的能力。虽然已经开发了许多计算方法来分析单细胞数据，但大多数方法主要通过恢复细胞内的基因-基因共变异结构来绘制和探索单细胞状态。研究细胞相互作用的方法主要集中在重建组织的空间组织上，根据已知的受体-配体对或已知的信号通路推断假定的物理细胞-细胞相互作用，或者使用空间数据突出显示由重复单元类型组成的细胞邻域。然而，揭示协调的多细胞过程的方法仍然缺乏。 [3]

2 相关工作

组织中不同细胞之间的相互作用对于维持内稳态至关重要。虽然许多疾病传统上被认为是一个特定的细胞或细胞类型的故障，越来越多的证据和新的治疗策略证明了多细胞行动的关键作用在健康和疾病，打开新的干预机会，诊断、疾病监测和预防。与此同时，单细胞 RNA 测序（scRNA-seq）和空间转录组学的进展，现在允许在细胞类型、组织、和疾病状态，包括分离细胞和完整组织。然而，尽管取得了这些进展，破译多细胞调控仍然是一个挑战，限制了从以细胞为中心到以组织为中心的视角移动的能力。 [4]

2.1 绝大多数分析单细胞数据方法

虽然已经开发了许多计算方法来分析单细胞数据，但绝大多数方法通过恢复细胞内的基因-基因共变异结构来绘制和探索单细胞状态（例如，PAGDOA21，NMF 实现 25 和扩展，包括 cNMF13 和 LIGER19,26）。方法研究细胞相互作用的发展主要集中在重建组织的空间组织 27-30，推断假知的物理细胞交互基于已知受体配体对或已知信号通路或或使用空间数据突出显示细胞邻域的重复单元类型组成 [?]。虽然这些方法揭示了细胞生物学和组织结构的重要特性，但揭示协调的多细胞过程的方法仍然缺乏。

2.2 不同于差异基因表达分析

DIALOGUE 不依赖于强有力的基本假设，不同于以往的监督和非监督方法。它不同于差异基因表达分析，因为它更加规范化，不需要强制执行特定结构，保留了识别看似相同表型下不同 MCP 的灵活性。DIALOGUE 也不同于以前的无监督单细胞分析和降维工具，因为它使用细胞内和细胞间的基因-基因相关性。DIALOGUE 不同于从空间转录组中提取空间特征的方法。

3 本文方法

3.1 本文方法概述

在本研究中，我们以一种新的方式来解决这个问题，通过引入 MCPs 的概念，并开发了第一种从单细胞或空间基因组学数据中系统地揭示 MCPs 的方法。我们将 MCPs 定义为不同细胞类型中不同表达程序的组合，它们在组织中协调，从而在组织水平上形成仅在细胞水平而不是高阶的功能单元。为了恢复它们，我们开发了对话，一种计算方法，通过多细胞配置识别来解耦细胞状态的方法，通过使用跨一个组织的生态位或跨来自多个个体的样本的跨细胞类型关联。我们将对话应用于空间转录组或 scRNA-seq 数据，其中它分别使用空间或跨样本变异来识别 mcp。应用于 MERFISH、Slide-seq 和 seq-FISH（序列荧光原位杂交）和小鼠下丘脑、小脑、视觉皮层和新皮层的空间注释 scRNA-seq 数据，对话成功恢复了看不见的测试数据的空间属性，优于其他方法和指标，并确定了标记动物行为的 mcp。应用于来自人类肺癌的空间数据集，对话确定了标记肿瘤边界的免疫激活和组织重塑的 MCPs。最后，应用于患者的 scRNA-seq 数据，对话确定了 (1) 溃疡性结肠炎（UC）MCP，预测治疗的临床反应，包括全基因组关联研究（GWAS）UC 风险基因；(2) 阿尔茨海默病（AD）MCP；(3) 黑

色素瘤的免疫治疗耐药性 MCP。综上所述，我们的方法和方法为研究细胞串扰和连接细胞和组织生物学开辟了一条新的途径 [5]

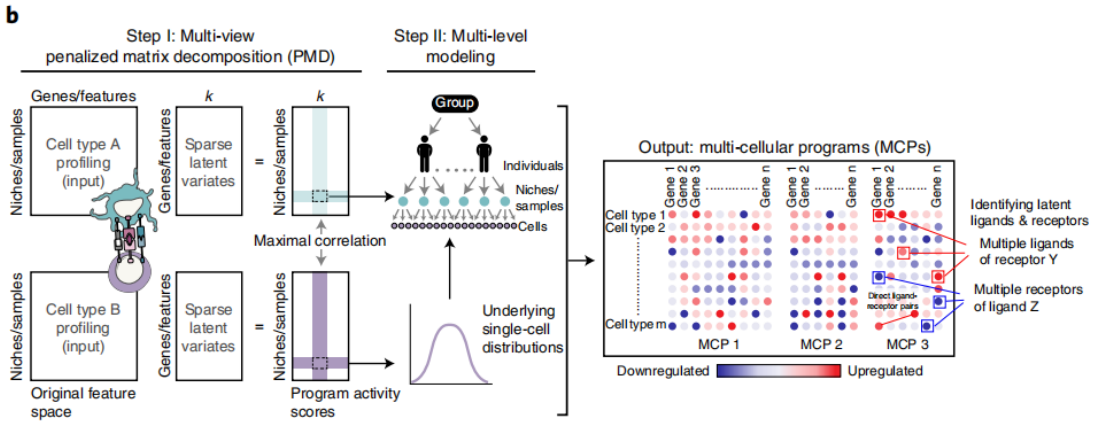


图 1. DIALOGUE：一种 MCP 识别的方法

3.2 对话方法论

概述给定空间或多样本单细胞数据，对话分两步识别潜在的 mcp，每个 mcp 由跨多种细胞类型的共同调控组成的基因集。在 frst 步骤中，它使用 PMD41 来识别稀疏规范变量，将原始特征空间（例如，基因和 pc）转换为一个新的特征空间，其中不同的细胞类型特定表示在不同的样本和环境中是相关的。

在第二步中，给定新的表示，对话使用多层次的层次模型来识别包含潜在特征的基因，同时考虑单细胞分布和控制潜在的混杂因素。对话以已知空间坐标或样本成员的单单元轮廓作为输入。虽然我们展示了 RNA 分析的应用，但其他测量类型（例如，蛋白质水平和基于细胞形状图像的特征）可以替代或另外使用。在非空间数据中，建议使用 frst20-30pc 或对每个单元类型分别使用其他降维方法导出的类似元特征来提供数据的紧凑表示。根据方差分析，对话删除了输入，只包含了在样本间显示出更大变化（BH FDR<0.05）的特征。对数据驱动的参与每个 MCP 的细胞类型的识别。

在 PMD 步骤之后，对话使用相关系数和排列测试来确定哪些细胞类型参与了每个 MCP，并仅对这些细胞类型应用多级测试。其次，给定一组单元格类型，对话以多路方式和成对的方式执行 frst PMD 步骤。如果它识别出了成对版本中唯一的程序，那么它就会将它们包含在多层建模步骤和 fnal 输出中。利用巨噬细胞、B 细胞、CD4 和 CD8 T 细胞的黑色素瘤 scRNA-seq 数据 55，评估了对话细胞识别参与每个 MCP 的相关细胞类型的能力。每次，其中一种细胞类型的表达矩阵被随机打乱，并与其他三种细胞类型的真实（未打乱）矩阵一起进行对话。所有识别的 mcp 只涉及真实的细胞类型，并且与仅提供三种细胞类型的未打乱数据时识别的 mcp 相同。

3.3 多层次建模步骤

由于对话开辟了一个新的特征空间，它通过询问单细胞分布，识别每个稀疏典型变量的基因签名，同时在不同的水平上考虑潜在的混杂因素（例如，患者年龄、性别、样本类型和细胞测序质量）。对于 K 个潜在特征集中的每一个，对话定义了一组 N 个签名（每个单元格类

型一个），记为（s1，k，...，sN，k）K k=1。如果一个基因 g 在细胞类型 r 中的表达与细胞类型 r 的 Xrwr，= 相关，而与 Xzwz，=，= 与其他细胞类型的相关潜在特征相关，则该基因 g 在 sr，= 中。在控制细胞质量时，使用部分 Spearman 相关性（BHFDR<0.05,0.05，每个细胞类型最多 250 个上调或下调的基因），使用每个细胞中检测到的 reads 的对数转换数。

为了通用化和减轻过拟合，对话方法采取了以下措施：首先，通过比较空模型来自洗牌数据的经验 P 值，重新运行 MCP 检测程序，并量化 MCP 的统计意义与真实数据进行比较，来计算每个 MCP 的经验 P 值。其次，对原始数据进行训练集和测试集的分割，并使用看不见的外部数据集来检查 MCP 的通用性。如果某个 MCP 不能显示出足够的统计意义或不能泛化，建议使用更少的特征作为输入或调整正则化参数，以增加 PMD 解决方案的稀疏性。

4 复现细节

4.1 与已有开源代码对比

引用的代码：

```
> param <- DLG.get.param(k = 3,
                         results.dir = "DLG.results/",
                         conf = c("gender","sample.quality","cellQ")
                         # Confounding factors
                         pheno = "pathology") # Phenotype (optional)

> R<-DIALOGUE.run(rA = rA, # list of cell.type objects
                  main = "RunName",
                  param = param)
```

我们注意到无论使用哪个 k，DIALOGUE 总是会找到相同的 MCPs 或其子集。

在作者的方法框架基础上，进行了额外的实验数据测试，并对数据进行了预处理以确保其符合分析需求。预处理包括数据清理、格式转换和特征工程等步骤，以使数据能够有效地被处理。这个个性化的数据处理方法使得我们能够更好地适应不同结构和形式的实验数据，并能够提供准确和可靠的分析结果。

4.2 实验环境搭建

实验要求: R（在 R 版本 3.4.0 中经过测试）。

R 库：lme4、lmerTest、PMA、plyr、matrixStats、psych、stringi、RColorBrewer、unikn、reshape2、ggplot2、grid、beanplot、parallel.

R 库
lme4
lmerTest
PMA
plyr
matrixStats
psych
stringi
RColorBrewer
unikn
reshape2
ggplot2
grid
beanplot
parallel

要安装 DIALOGUE，可以使用

```
devtools::installgithub(repo = "https://github.com/livnatje/DIALOGUE")
```

或者直接下载其 R 包并使用

```
devtools::install("DIALOGUE").
```

生成对象：输入以 cell.type 对象的形式提供，每个对象表示特定的细胞类型或亚型。可以使用以下方法生成这些 cell.type 对象：

```
make.cell.type(name = "Cell.type.name",tpm,samples,X,metadata)
```

1.name 是细胞类型的名称（例如，”macrophage”，”DC” 等）；

2.tpm（m x n）是单细胞基因表达；

3.samples（n x 1）是每个细胞的样本标识；如果数据具有空间坐标，则组织中的每个小环境被视为不同的” 样本”，从而提供更多统计学能力来识别多细胞程序（MCPs）；

4.X（n x k1）是” 原始特征空间”；这些可以是主成分（PCs），NMF 成分，基因表达矩阵或任何其他表示。建议使用具有 k1 lt;lt; n 的表示，并且在仅使用特定类型或亚型的细胞时执行初始降维，以充分捕捉该特定子集内的变化；

5.metadata（n x k2）包括可能希望包含为潜在混杂因子或具有生物学意义的细胞的任何其他特征。

4.3 使用说明

输入数据：包括不同细胞类型的单细胞转录组，通常还包括更紧凑的表示（例如主成分）。

输出结果：将是跨不同细胞类型共同调控基因的多细胞程序（MCPs），它们在各个细胞中的表达，以及与特定感兴趣的表型的关联。每个 MCP 包括多个细胞类型特定的基因子集。

为了运行 DIALOGUE，首先生成一个包含你想在分析中包括的细胞类型对象的 list。DIALOGUE 将识别在不同细胞类型中共调控的基因集，我们称之为多细胞程序（MCPs）。

输出 R 包括：

MCPs - 以基因集列表形式给出的 MCPs；

scores - 每个细胞中 MCPs 的得分；

gene.pval - 每个 MCP 中每个基因的跨细胞类型 p 值；

pref - 每个 MCP 的细胞类型特异成分之间的相关性（R）和关联性（混合效应 p 值）

pheno - 每个 MCP 与感兴趣表型的关联，以方向乘以-log10(p 值) 的形式给出

MCPs	以基因集列表形式给出的 MCPs
scores	每个细胞中 MCPs 的得分
gene.pval	每个 MCP 中每个基因的跨细胞类型 p 值
pref	每个 MCP 的细胞类型特异成分之间的相关性（R）和关联性（混合效应 p 值）
pheno	每个 MCP 与感兴趣表型的关联，以方向乘以-log10(p 值) 的形式给出

4.4 创新点

在作者的方法框架基础上，进行了额外的实验数据测试。由于测试数据的结构与原始数据不同，需要在进行分析之前对数据进行预处理。在这个过程中，根据数据的形式做出了一些相应的改进。具体而言，对数据进行了适当的调整，以确保其符合的分析需求。这涉及到数据清理、格式转换以及可能的特征工程，以使数据能够有效地被处理。这个预处理的阶段是为了确保我的测试数据能够与原始方法相兼容，并能够提供准确和可靠的分析结果。这种个性化的数据处理方法使其能够更好地适应不同结构和形式的实验数据，从而更全面地评估此方法的有效性。

5 实验结果分析

多细胞程序（MCPs）与特定感兴趣表型的关联，由 DIALOGUE.run 中的 phenoZ 参数给出。该值的大小为 -log10(p-value)，其符号表示关联是正向还是负向的。

```
> print(round(R$phenoZ,2))
```

	MCP1	MCP2
A	-3.04	1.39
B	-3.30	1.07
C	-4.43	-1.05
All	-3.86	0.86

这个描述表明在这个例子中，MCP2 在细胞类型 A 中表现出与表型的强烈负向关联，而 MCP1 在所有细胞类型中，特别是在细胞类型 C 中，表现出与表型的强烈正向关联。这些观察可以帮助解释基因集与表型之间的关系，以及在不同条件下这种关系的差异。

在这个示例中，MCP2 显示出与表型的强烈负相关，特别是在 A 型细胞中，而 MCP1 在所有细胞类型中，特别是在 C 型细胞中，显示出与表型的正相关。把解压后的 txt 文件用 seutat 转换为.rds 类型的文件。

```
    install.packages("Seurat")
```

```
    library(Seurat)
# 例如，假设你的txt文件包含细胞名称在第一列，
基因表达值从第二列开始
data <- read.table("D:\\...Infect_HP_P_01.txt",
header = TRUE, row.names = 1)
 # 创建一个空的Seurat对象
 seurat_obj <- CreateSeuratObject(counts = data)
 # 进行一些预处理，例如，归一化和标准化数据
 seurat_obj <- NormalizeData(seurat_obj)
 seurat_obj <- ScaleData(seurat_obj)
# 使用Seurat的聚类和降维函数对数据进行分析。
例如，运行聚类和UMAP
 seurat_obj <- FindNeighbors(seurat_obj)
 seurat_obj <- FindClusters(seurat_obj)
 seurat_obj <- RunUMAP(seurat_obj) #可能运行不出来但是不影响
 # 使用Seurat的可视化函数和分析工具对结果进行可视化和进一步分析
 例如，绘制UMAP图和基因表达热图
 DimPlot(seurat_obj)
 FeaturePlot(seurat_obj, features = c("gene1", "gene2"))
 # 将Seurat对象保存到文件中
 saveRDS(seurat_obj, file = "D:\\05.HP\\Infect_HP_P_01.rds")
```

对数据进行了简单的分类可视化

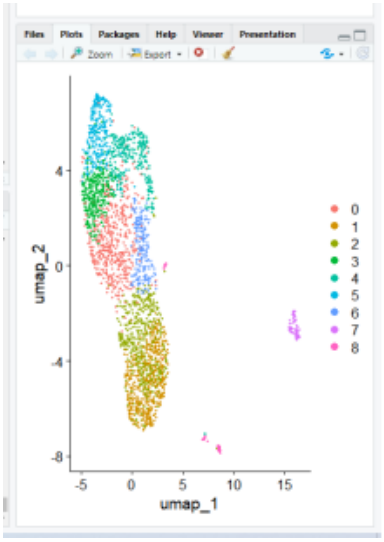


图 2. 实验结果示意

最后用

```
subset = adata[adata.obs['type'] == 'HBV']
```

读到了类型为 HBV 的数据这里有显示数据信息。

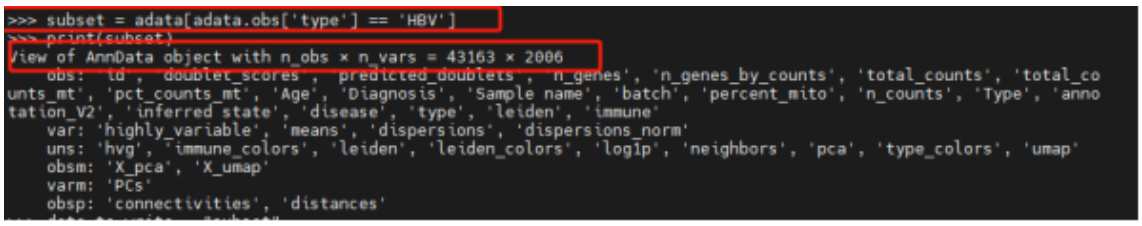
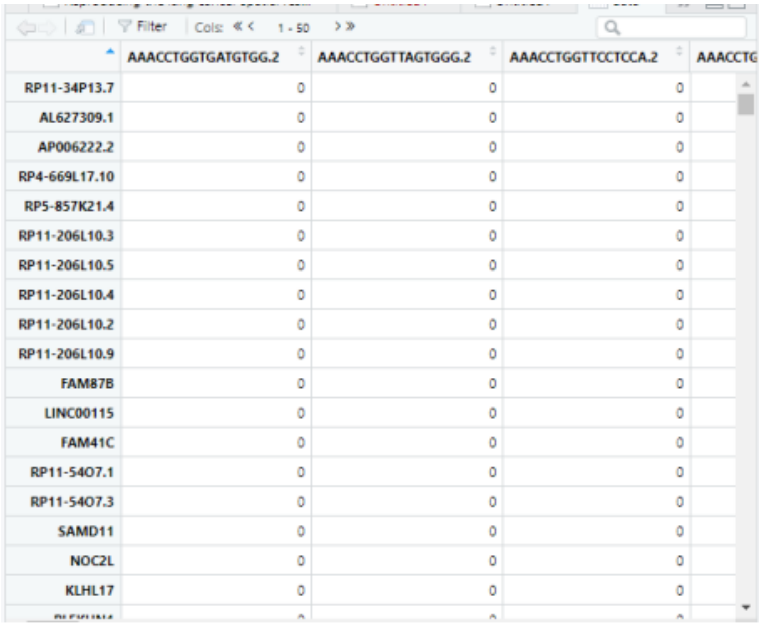


图 3. 数据结构显示



Filter Cols: 1 - 50

	AAACCTGGTGATGTGG.2	AAACCTGGTTAGTGGG.2	AAACCTGGTTCCTCCA.2	AAACCTG
RP11-34P13.7	0	0	0	
AL627309.1	0	0	0	
AP006222.2	0	0	0	
RP4-669L17.10	0	0	0	
RP5-857K21.4	0	0	0	
RP11-206L10.3	0	0	0	
RP11-206L10.5	0	0	0	
RP11-206L10.4	0	0	0	
RP11-206L10.2	0	0	0	
RP11-206L10.9	0	0	0	
FAM87B	0	0	0	
LINC00115	0	0	0	
FAM41C	0	0	0	
RP11-54O7.1	0	0	0	
RP11-54O7.3	0	0	0	
SAMD11	0	0	0	
NOC2L	0	0	0	
KLHL17	0	0	0	

图 4. 数据结构显示

在 results.dir 目录中，可以找到显示我们识别的 MCPs 的表达、组成和特定表型分布的图表。

下面的图表中，每个点对应一个样本，显示了一个 MCP 组件的表达水平与另一个 MCP 组件的表达水平之间的关系，同时还显示了皮尔逊相关系数（Pearson correlation coefficient）：

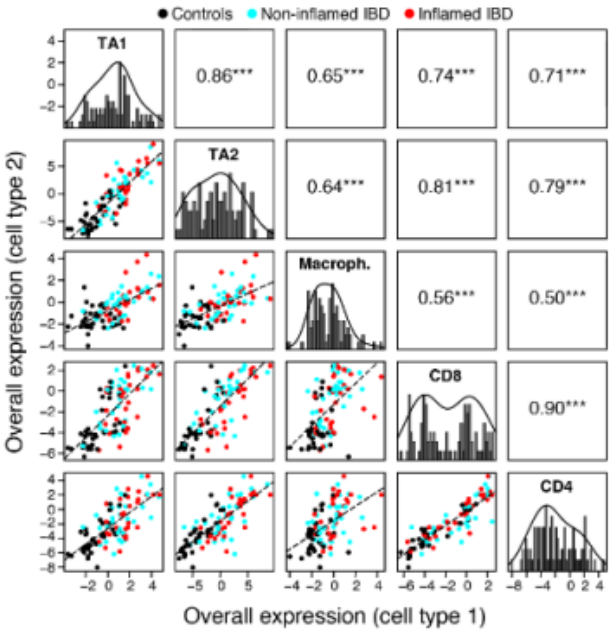


图 5. 细胞类型 1 的整体表达

下面的图表显示了每个多细胞程序（MCP）的细胞类型特异性和非特异性组分中基因的数量：

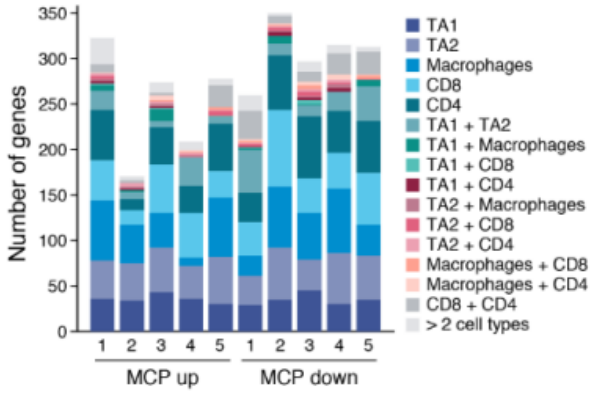


图 6. 基因数量

如果存在感兴趣的特定特征或表型，DIALOGUE 将绘制根据该分类对细胞进行分层的 MCP 表达图：

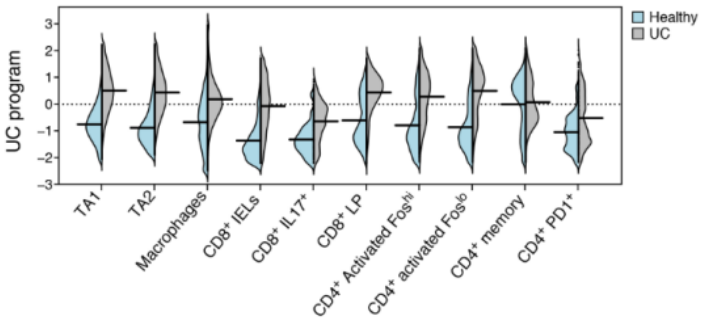


图 7. 疾病状态

我们可以看到在结肠样本中识别的 MCP 在疾病状态（UC）中更高。

6 总结与展望

开发了 DIALOGUE, 这是一种系统地揭示多细胞程序 (Multicellular program, MCP) 的方法—不同细胞类型的协调细胞程序的组合，从空间数据或在没有空间信息的情况下获得的单细胞数据在组织水平上形成高阶功能单元。

将 DIALOGUE 应用于跨个体或区域的单细胞 RNA 测序数据，作者发现了标记阿尔茨海默病、溃疡性结肠炎和癌症免疫治疗抵抗的 MCP。这些程序可以预测独立队列中的疾病结果和易感性, 并包括来自全基因组关联研究的风险基因。DIALOGUE 还可以分析健康和疾病中的多细胞调节。

DIALOG 目前的局限性之一是，它不仅取决于分析的细胞的数量，而且还取决于样本的数量或空间位置。随着单细胞研究的数量、规模和多样性迅速增长，以及空间转录组学技术得到更广泛的应用, DIALOGUE 应该有助于分析未来的单细胞和空间数据集, 并可能将它们结合使用。本文提出的概念、方法和问题为全面绘制健康和疾病中 MCP 潜在组织功能提供了基础。

参考文献

[1] Soyon Hong and Beth Stevens. Microglia: phagocytosing to clear, sculpt, and eliminate, 2016.

[2] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, 2015.

[3] Darren J Burgess. Spatial transcriptomics coming of age, 2019.

[4] Bogdan A Luca, Chloé B Steen, Magdalena Matusiak, Armon Azizi, Sushama Varma, Chunfang Zhu, Joanna Przybyl, Almudena Espín-Pérez, Maximilian Diehn, Ash A Alizadeh, et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors, 2021.

[5] Livnat Jerby-Arnon and Aviv Regev. Dialogue maps multicellular programs in tissue from single-cell or spatial transcriptomics data, 2022.