

学号 2017300030039

密级 _____

武汉大学本科毕业论文

基于 GRN 的垂体基因表达差异分析

院（系）名称：弘毅学堂

专业名称：计算机科学与技术

学生姓名：郑晖

指导教师：蔡朝晖 副教授

二〇二一年四月

郑重声明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名：_____ 日期：_____

摘 要

诸如病毒或细菌感染之类的免疫挑战会引起组织炎症，垂体在炎症事件的调节过程中扮演着重要角色。然而，我们对免疫攻击过程中垂体细胞的转录反应知之甚少。使用基于机器学习与统计学习的生信算法，处理单细胞 RNA 测序数据，将为我们提供细胞转录反应层级的生物学见解。现在希望运用 GRN 推断、聚类、基因表达差异分析等算法，对实验科学家收集到的垂体单细胞 RNA 测序数据进行分析，揭示垂体内部各类细胞在中枢神经内分泌炎症调节过程中的角色以及在该过程中起关键作用的调控因子。

在这项研究中，我使用基因组学家标注的调控子-目标基因集合训练了一个 GRN 推断器。依据此模型，我对实验科学家在小鼠炎症模型上收集到的垂体单细胞转录组数据，进行基因调控网络推断。使用聚类等分析手段对得到的基因调控网络进行处理，推断其细胞种类及状态。然后，统计该聚类标签与处理组标签的匹配度，验证刺激的有效性。最后，依据得到的细胞标签，对不同状态的同一种类垂体细胞分别进行基因表达差异分析，并对不同状态的所有细胞进行 GRN 矩阵差异分析，探寻免疫攻击过程中垂体细胞的转录反应变化以及驱动这种变化的关键转录因子。

最终分析结果表明，在垂体各类细胞中，对照组与实验组之间的转录反应都有巨大差异，不同种类垂体细胞都积极参与到炎症调节过程中。但是，不同种类垂体细胞的上调、下调基因集合并不相同，即不同垂体细胞在炎症调节过程中的不同角色。除此之外，我还鉴定出一类在不同种类垂体细胞中共表达的基因调控子，如 Stat、Irf 和 Nfkb 等。

这项研究的结果扩展了我们对炎症激发过程中垂体单细胞转录反应的了解，并提供了将单细胞 RNA 测序用于动态功能研究的新思路。

关键词：基因调控网络；单细胞测序；系统性神经炎症；垂体

ABSTRACT

Immune challenges such as viral or bacterial infections can cause tissue inflammation, and the pituitary gland plays an important role in the regulation of inflammatory events. However, we know very little about the transcriptional response of pituitary cells during immune attack. Using biosynthesis algorithms based on machine learning and statistical learning to process single-cell RNA sequencing data will provide us with biological insights. Now I hope to use GRN inference, clustering, gene expression differential analysis and other algorithms to analyze the pituitary single-cell RNA sequencing data collected by experimental scientists, and reveal the role of various cells within the pituitary in the regulation of central neuroendocrine inflammation and the regulatory factors.

In this study, I trained a GRN inference machine using the regulator-target gene set annotated by genomicists. Based on this model, I performed gene regulatory network inferences on the pituitary single-cell transcriptome data. Then, I use analysis methods to process the obtained gene regulatory network and infer its cell type and state. Finally, according to the obtained cell label, the gene expression difference analysis of the same type of pituitary cells in different states is performed, and the GRN matrix difference analysis is performed on all cells in different states to explore the transcriptional response changes of pituitary cells during immune attack and the key transcription factors.

The final analysis results showed that in various types of pituitary cells, there are huge differences in the transcriptional response between the control group and the experimental group, and different types of pituitary cells are actively involved in the process of inflammation regulation. However, different types of pituitary cells have different sets of up-regulated and down-regulated genes, that is, different pituitary cells have different roles in the regulation of inflammation. In addition, I also identified some gene regulators that are co-expressed in different types of pituitary cells, such as Stat, Irf, and Nfk.

The results of this study expand our understanding of the pituitary single-cell transcriptional response during inflammation stimulation.

Key words: GRN; single cell sequencing; systemic neuroinflammation; pituitary

目 录

1 绪论	1
1.1 研究背景	1
1.2 研究现状与研究内容	2
1.3 研究结果	2
2 相关工作	3
2.1 单细胞 RNA 测序	3
2.2 基因调控网络	5
3 单细胞 RNA 测序数据处理流程	7
3.1 将原始测序数据转化为基因表达矩阵	7
3.2 对基因表达矩阵进行质量控制	7
3.3 依据基因表达矩阵进行聚类	8
3.3.1 特征选取	8
3.3.2 Leiden 聚类算法	8
4 基于 SCENIC 的 GRN 推断	11
4.1 SCENIC 算法原理	11
4.1.1 GRNBoost2	11
4.1.2 RcisTarget	13
4.1.3 AUCell	14
4.2 选择 SCENIC 算法的原因	14
5 实验数据分析	17
5.1 测序数据预处理	17
5.1.1 Scater 质控	17

5.1.2 Seurat 初步分析.....	18
5.2 测序数据 SCENIC 分析.....	19
5.2.1 分析处理条件与垂体细胞状态之间的关系	19
5.2.2 分析不同细胞在炎症状态下的基因表达差异	20
5.2.3 分析导致炎症状态的转录因子	20
5.3 讨论和未来工作	20
6 总结与展望	23
参考文献	25
致谢	31

1 緒論

1.1 研究背景

在病毒或细菌感染期间，免疫因子会在人体中释放，通常会导致组织发炎并导致严重的疾病行为，例如食欲不振，嗜睡，退出正常的社交活动，疲劳，探索力下降等。人们认为疾病行为是由可溶性促炎性细胞因子（IL-1、 $TNF - \alpha$ 、IL-6等）触发的，该因子由感染部位的免疫细胞产生，并会对神经内分泌系统，特别是下丘脑-垂体-肾上腺（HPA）轴 [1, 2] 产生深远的影响。

HPA 轴是体内的压力反应中心，连接中枢神经系统（CNS）和内分泌系统，其在炎症状态下的调节过程如1.1所示。作为 HPA 轴组成部分的垂体，在炎症事件调节过程中起到重要的作用。由炎症事件诱导的细胞因子（IL1, IL6, TNF- α , IFN- γ ）通常循环至垂体前叶，并主要作用于垂体的促肾上腺皮质激素，从而促进释放抗炎激素，例如肾上腺皮质激素（ACTH）。ACTH 被携带到肾上腺并作用于 ACTH 受体，从而上调肾上腺皮质肾上腺皮质细胞中皮质醇的释放。随后，皮质醇在下丘脑和垂体在 HPA 轴上产生负反馈，以抑制促炎性细胞因子的进一步合成和释放。此外，卵泡细胞代表垂体前叶中唯一的非内分泌细胞类型，并释放可能潜在影响垂体局部激素产生的 IL1 和 IL6，从而构成调节炎症反应的复杂系统。

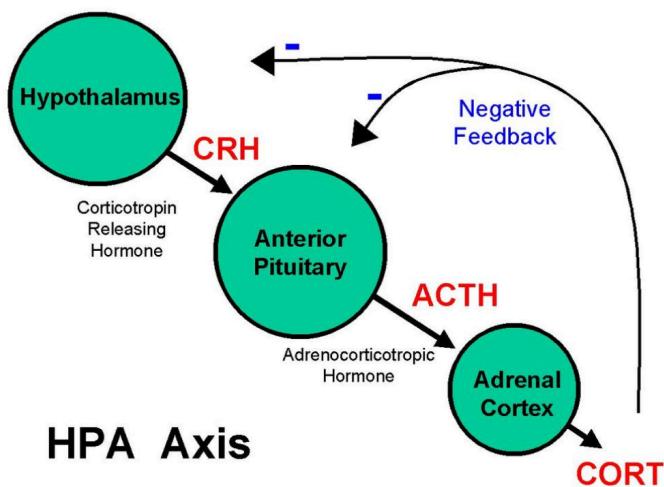


图 1.1 炎症状态下的 HPA 轴

1.2 研究现状与研究内容

以往探究垂体在中枢神经内分泌炎症调节过程中作用的研究 [1, 2] 都没有涉及到单细胞转录层级，没有揭示垂体内部各类细胞在中枢神经内分泌炎症调节过程中的角色以及内在调控因子。

近几年随着单细胞 RNA 测序技术的发展 [3]，研究人员开始从单细胞转录组水平探究垂体相关的一些问题 [4–7]，但这些工作主要关注于某个发育过程中的静态分类问题，很少有研究使用单细胞转录组测序来进行动态功能研究。

在这项研究中，我主要关注不同的垂体细胞如何响应炎症刺激。我使用 Seurat[8, 9] 规定的标准处理流程和基因调控网络推断 SCENIC[10, 11] 算法，对实验科学家收集的垂体单细胞 RNA 测序数据进行分析，以探究垂体内部各类细胞在中枢神经内分泌炎症调节过程中的角色以及内在调控因子，动态炎症研究将成为这项研究中最重要的部分。这项研究可以使我们对人体对病毒或细菌感染的免疫防御具有更清晰的认识。更重要的是，它具有非常重要的临床意义，我们希望获得用于免疫诊断的特定标记。此外，这项研究也可以为我们提供关于单细胞转录组测序技术应用的新思路。

1.3 研究结果

在这项研究中，主要贡献有：(1) 提供了不同种类垂体细胞参与中枢神经内分泌炎症调节过程的单细胞转录层级证据。(2) 揭示了不同种类垂体细胞在参与中枢神经内分泌炎症调节的过程中的转录水平差异，表明其在炎症调节过程中扮演不同的角色。(3) 发现了一类在不同种类垂体细胞中统一表达的转录因子，表明其在垂体参与中枢神经内分泌炎症调节过程中的重要地位。

文章结构安排如下：第二章介绍单细胞测序以及基因调控网络（GRN）的相关工作；第三章介绍单细胞 RNA 测序数据处理流程；第四章介绍 SCENIC 算法原理及优势；第五章介绍使用 GRN 对小鼠垂体单细胞测序数据进行分析；第六章总结了该项工作的结论并对未来的工作进行了展望。

2 相关工作

2.1 单细胞 RNA 测序

生物体内各种组织之间存在巨大的差异，甚至同一块组织也会有在形态、功能上差异巨大的细胞。Bulk-RNA 测序技术可以很好地用来探究组织异质性，但其无法很好地解决后面一个问题，其原因便在于 Bulk-RNA 测序技术无法提供单细胞层级的转录信息。相较之下，单细胞 RNA 测序（scRNA-seq）技术提供了在单细胞水平观测基因表达的方法，可以更好地研究组织内的细胞异质性 [12–17]。单细胞 RNA 测序技术可解决的常见问题 [18, 19] 包括：

- 探究异质性（Studying heterogeneity）
- 谱系路径分析（Lineage tracing study）
- 随机基因表达研究（Stochastic gene expression study）

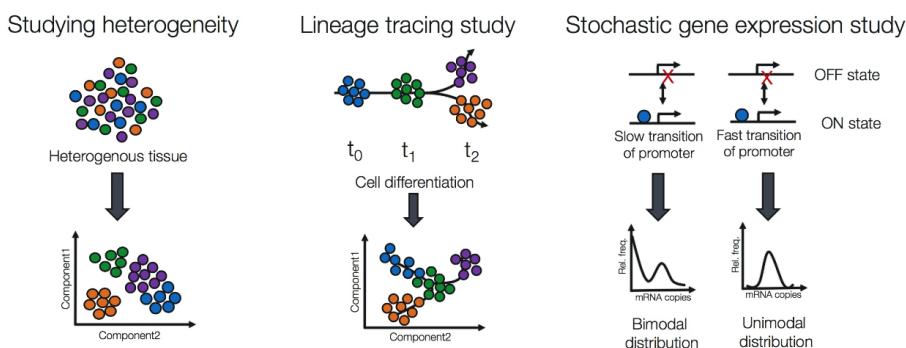


图 2.1 单细胞 RNA 测序可解决的常见问题

单细胞 RNA 测序技术最初起始于 Kurimoto 等人在 2006 年的一项工作 [20]，这项工作对于后来单细胞转录组测序的原理发展有很大的影响。该研究主要特点在于加入了 T7 启动子，这样便把 cDNA 为期 29 个循环的扩增切分为两段，减少 PCR 扩增的偏差。汤富酬等人在 2009 年所做的一项工作 [21]，延用了 Kurimoto 等人在 2006 年工作 [20] 中在末端加 A 的思路。但是在最后读取 cDNA 信息的时候，汤等人使用了 Applied Biosystem 的二代测序 SOLiD system 平台，也就是取代了芯片的读取方式。

目前应用最广泛的是模板转换法，主要代表技术便是 SMART-seq[22, 23]。其实，在 2011 年的 START-seq[24]，就已经用到模板转换法，同时运用 Barcode 标记的思路来达到相对高通量的单细胞转录组测序。稍微改进这种方法，将 Barcode 加

在 3' 端，便可以富集 3' 端测序，同时在一开始就将测序接头设计到引物里去，以后不用再引入，便可以做高通量。如若再不加 Barcode，一个细胞的 cDNA 建立一个库，这样就可以获得全长的 cDNA 信息，也就是有了后来的 SMART-seq1&2。

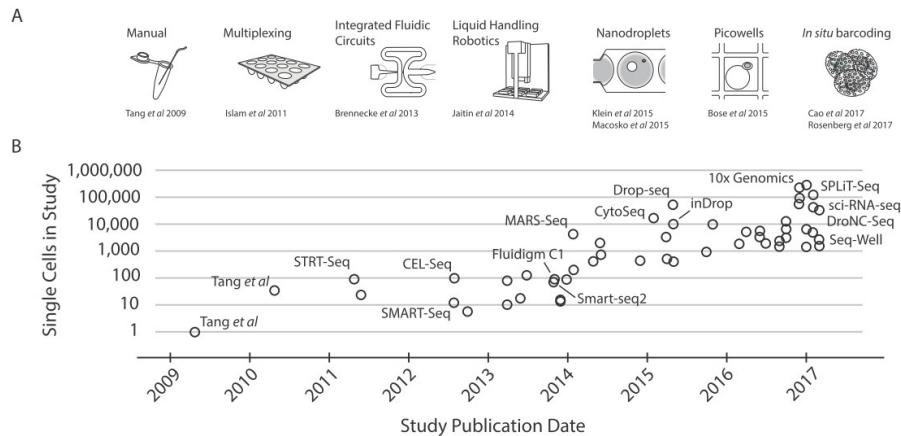


图 2.2 单细胞 RNA 测序技术近年来的发展

对于这项研究而言，重要的是选择一种基于高通量和高度自动化的适当单细胞测序方法，以确保可以获得高灵敏度和准确性的测序数据。我的合作者从 Tang Lab (FuChou Tang 博士) 那里获得了一种新的改进的 Smart-seq2 方法，用来制备垂体单细胞 RNA 测序数据集，其过程如图2.3所示，该方法可最大程度地减少操作并利用单管反应来避免部分材料的损失。反转录时，它将为每个细胞提供特定的细胞条形码，除了为每个 mRNA 提供不同的独特分子标识符 (UMI) 外，还可以将细胞集中在一起以进行测序文库的构建，并使用 UMI 进行质量控制，以便获得高质量的单细胞基因表达数据。

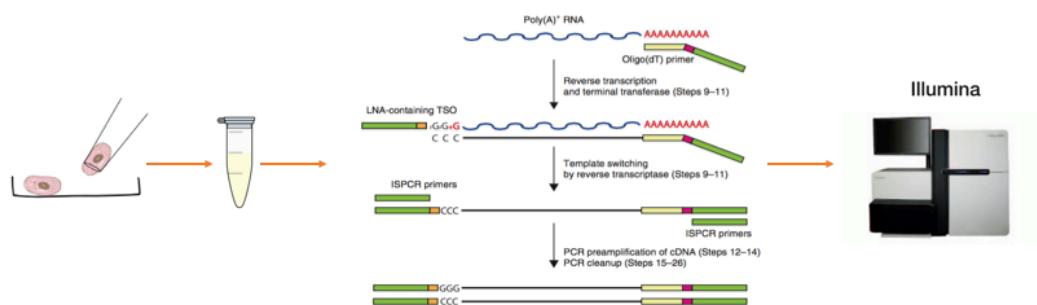


图 2.3 改进的 Smart-seq2 流程

2.2 基因调控网络

基因调控网络 (GRN) 定义并维持特定于细胞类型的转录状态，这反过来又是细胞形态和功能的基础。每种细胞类型或稳定状态均由活性转录因子 (TF) 的特定组合定义，这些转录因子与基因组中的一组顺式调节区域相互作用（与染色质结构相互作用），以产生特定的基因表达谱 [25, 26]。活性 TF 及其靶基因的组合通常表示为 GRN。

揭露 GRN 是基因组研究领域的主要挑战之一。一旦确定了驱动并维持细胞状态行为的关键调节剂，它们最终就可以用来干扰这些调节程序。实例包括通过 Yamanaka 等人 [27] 提出的 TF 组合，将成纤维细胞重编程为诱导性多能干细胞 (iPS)，还有许多其他重编程途径，它们使用 TF 的特定组合将 GRN 从一种状态引导到另一种状态 [28, 29]，以及最近在癌症治疗中，尝试用特定的 TF 组合将癌细胞推入易受特定药物影响的状态 [30, 31]。

基于大规模转录组和表观基因组数据来计算预测 GRN 是一个广泛研究的领域。相关算法包括 GENIE3[32]、GRNBoost2[33] 和 BEELINE[34] 等。在这项研究中我们主要使用基于 GRNBoost2 的 SCENIC[10, 11]，进行基因调控网络推断。

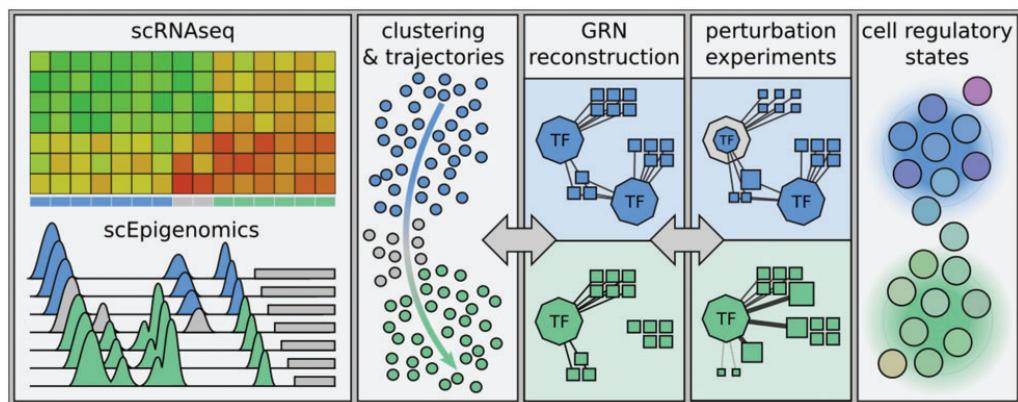


图 2.4 基因调控网络推断过程

以往的单细胞 RNA 测序工作主要使用原始基因表达矩阵进行聚类，进而标记细胞类型以及状态，但结果的真实性经常受到批次效应的挑战。批次效应的产生，是由于涉及单细胞 RNA 测序的实验往往需要在从多只小鼠上收集数据，而这种操作会引入一些与生物信息无关的噪声（如温度、研磨程度等）。

相较之下，SCENIC 在推导基因调控网络的过程中，能够觉察到一整个基因集合的整体趋势，去除批次效应所带来的影响。此外，SCENIC 在推理 GRN 的时候

并不是单纯地依赖 GRNBoost2 得到的相关性，而是对 GRNBoost2 推理得到的相关性利用生物学的先验知识进行修剪，只留下具备因果关系的转录因子及其目标基因，能够获得更加生物合理的结果。因而，使用基因调控网络进行聚类，其结果更贴近生物真实状态。

我们会在第四章进一步阐述选用 SCENIC 算法进行基因调控网络推断的详细原因及相关证据。

3 单细胞 RNA 测序数据处理流程

3.1 将原始测序数据转化为基因表达矩阵

在将生物样品进行测序后，我们会得到内含测序数据的 fastq 文件。但 fastq 文件是由高通量测序产生的输出文件，我们不能直接使用它进行数据挖掘。在这项研究中我们使用 CellRanger 将其转化为基因表达矩阵，并基于此开展下游分析。

CellRanger 是一组分析管道，用于处理 Chromium 单细胞数据以对齐读取、生成 Feature-Barcode 矩阵、执行聚类和其他辅助分析等。CellRanger 有四个处理管线与 3' 单细胞基因表达解决方案及其相关产品有关，我们在这里只关注常用的 cellranger mkfastq 与 cellranger count，其对应流程如图3.1所示。cellranger mkfastq 的主要作用是将 Illumina 测序生成的原始碱基检出（BCL）文件转化为 fastq 文件。cellranger count 对 cellranger mkfastq 获取的 fastq 文件执行对齐、过滤、Barcode 计数以及 UMI 计数，我们使用其产生的计数矩阵进行下游分析。

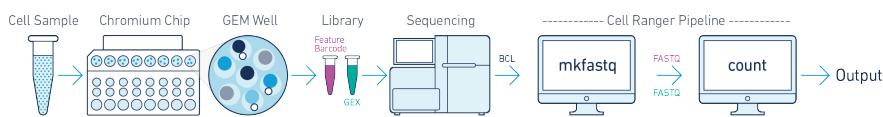


图 3.1 CellRanger 工作流程

3.2 对基因表达矩阵进行质量控制

对基因表达矩阵进行质控的主要目的是去除制备生物样品和单细胞 RNA 测序过程中引入的噪音。例如，在进行单细胞 RNA 测序的过程中，受限于测序芯片的技术条件，会有一定比例将两个细胞吹到同一个测序单位中，又或是使细胞破损等。为了保证下游分析结果的正确性，我们必须将这些离群点检测出来并将其去除。

目前进行质量控制的主流方法还是阈值法，即对基因表达矩阵在细胞水平 (Cell-level QC)、基因水平 (Feature-level QC) 以及变量水平 (Variable-level QC) 依据所处理生物数据特性人为设置生物合理的阈值。正常情况下，所检测到的线粒体 RNA 占所检测到 RNA 总量的 5% 以下，如果高于这个值便可能是在测序的过程中出现细胞破损，导致细胞质 RNA 外流。但是这种情况也不是绝对的，比如在肌肉细胞中代谢旺盛，所检测到的线粒体 RNA 占比便会高于此值，因此我们需要针对组织的特异性设定不同的阈值。

在质量控制之后，我们还可以去除一些已知功能且与研究不相关的 RNA（比如核糖体 RNA、线粒体 RNA），以避免其对下游分析的影响。

3.3 依据基因表达矩阵进行聚类

在对基因表达矩阵进行质量控制之后，为探求单细胞层级的异质性，我们可以依照其基因表达的差异将其聚类为不同的簇，并鉴定不同簇上的基因 marker，为之后的分析提供基础的生物见解。Seurat v3[8, 9] 建议的初步处理流程是：归一化、特征选择、放缩、线性降维、构建最短近邻图、Leiden 聚类和 UMAP 可视化。

3.3.1 特征选取

特征选取是在 Feature-Barcode 矩阵（或者说 Gene-Cell 矩阵）中计算每个基因的方差，并对其进行排序，选取其中具有较高方差的基因集合（即，它们在某些细胞中高表达，而在其他细胞中低表达）用于下游线性降维。这里使用一些方差较大基因的原因在于，从信息学的角度来看，方差越大的变量蕴含的信息越多。同时，Brennecke 等人在 2013 年的一项研究 [35] 表明在下游分析中关注这些基因有助于去除技术噪声，以在单细胞数据集中突出显示生物信息。线性降维也是基于此原理，对数据进行进一步的筛选，去除方差较低成分的噪声影响。

3.3.2 Leiden 聚类算法

Leiden 聚类 [36] 是整个初步处理流程的关键一步，其依据线性降维结果构建的最短近邻图挖掘其中的群落结构，其原型是 Louvain 聚类算法 [37]。我们下面会详细介绍 Louvain 算法的局限以及 Leiden 算法的改进。

Louvain 算法和 Leiden 算法都是基于图数据的群落发现算法，其灵感源于 modularity 的优化。其中，modularity 是一个定义在 $[-1/2, 1]$ 的比例值，用于度量群落内部边缘相对于群落外部边缘的相对密度。对于加权图，modularity 由公式 (3.1) 定义：

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (3.1)$$

其中， A_{ij} 表示节点 i 和 j 之间边的权重， k_i 和 k_j 分别是连接到节点 i 和 j 的边的权重之和， m 是图中所有边缘权重的总和， c_i 和 c_j 是节点的群落， δ 是 Kronecker 函数（如果 $x = y$ ， $\delta(x, y) = 1$ ，否则为 0）。从理论上优化此值会导致给定网络节点的最佳分组，但这在计算上并不可行。因此使用启发式算法，通过迭代来近似 modularity 的最大值，Louvain 算法和 Leiden 算法都是属于这种形式。

在 Traag 等人 2019 年的一项工作 [36] 中，实验证据表明 Louvain 算法得到的分区中会存在连接不良的群落，甚至内部断开的群落。而且他们证明在使用 Louvain 算法时，该问题在实践中经常发生。在 Louvain 算法中存在这样一种情况：一个节点在其旧群落中充当不同部分连接的桥梁，但是却可能在群落更新过程中被移动到另一个群落，从其旧群落中删除这样的节点会断开旧群落的连接。Louvain 算法可能假设旧群落的其他节点也会移动到该新群落，但事实并非如此，尽管旧群落已经断开连接，但其他节点仍可以与其群落保持牢固的联系，如图3.2所示。

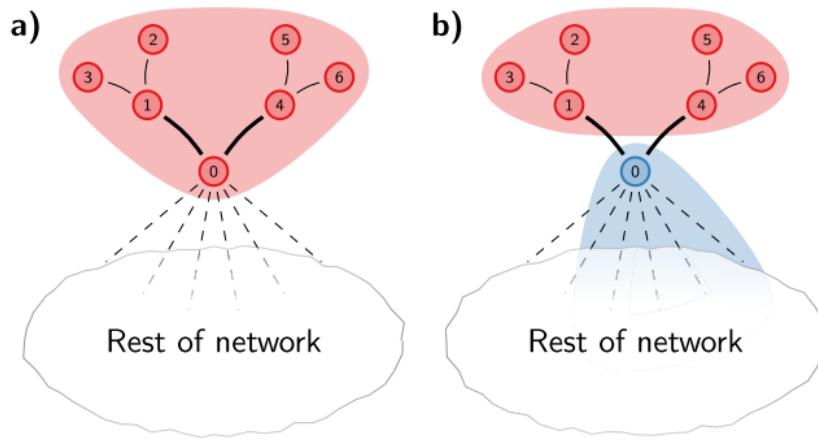


图 3.2 Louvain 算法存在的问题

Leiden 算法的改进主要在于利用了加快节点局部移动 [38, 39] 以及将节点移到随机邻居 [40] 的思想。Leiden 算法由三个阶段组成：(1) 节点的局部移动；(2) 分区的细化；(3) 基于细化的分区进行网络聚合。

在算法的第一阶段，首先将网络中每个节点分配给其自己对应的群落。然后，对每个节点 i ，计算将其从自身的群落移除并将其移至每个邻居节点 j 对应群落所带来的 modularity 变化。这两步的公式比较相似，其中将节点 i 移至其邻居节点 j 对应群落所带来 modularity 的变化由公式 (3.2) 定义：

$$\Delta Q = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] \quad (3.2)$$

其中， Σ_{in} 是节点 i 将要移入群落中所有连接的权重之和， Σ_{tot} 是节点 i 将要移入群落中所有连接到节点的权重之和， k_i 是节点 i 的加权度， k_{in} 是节点 i 与其将要移入群落中所有节点的连接权重之和， m 是网络中所有连接权重之和。利用加快节点局部移动 [38, 39] 的思想，仅计算与节点 i 相连所有群落中刚修改群落的 modularity，其余沿用以前的计算值。

在算法的第二阶段产生的分区 $\mathcal{P}_{refined}$ 是 Louvain 算法对应阶段生成分区 \mathcal{P} 的细分。分区 \mathcal{P} 中的群落可以分为分区 $\mathcal{P}_{refined}$ 中的多个群落。在进行合并的过程中，节点 i 并没有像 Louvain 算法中那样贪婪地与最大化 modularity 的群落合并，而是随机选择任何可以增加 modularity 的群落进行合并。modularity 的增加越大，选择该群落的可能性就越大 [40]。选择群落的随机性允许更广泛的探索分区空间。此外，仅当节点 i 与分区 \mathcal{P} 中的群落充分良好地连接时，才将其与分区 $\mathcal{P}_{refined}$ 中的群落合并，从而避免分区存在连接不良的群落。

在算法的第三阶段，将第二阶段得到的每一个群落折叠为单个节点，并在此基础上建立新的网络。这时，新群落节点上的自环表示同一群落节点之间的所有连接，而群落节点之间的加权边表示同一群落多个节点到不同群落节点的连接。在新网络创建完毕后，其结果可以重新用于算法的第一阶段，进行迭代，最终得到聚类结果。

总之，Leiden 算法可以保证产生的分区中不会存在连接不良的群落。在 Leiden 算法被迭代使用时，它会收敛到一个分区，在该分区中，可以确保所有群落的所有子集都属于局部最优分配。因此，Leiden 算法可以提供准确的聚类见解，供下游分析使用。

4 基于 SCENIC 的 GRN 推断

SCENIC 是一种利用单细胞 RNA 测序数据同时进行基因调节网络重建和细胞状态识别的计算方法。SCENIC 能够帮助我们从基因表达矩阵中鉴定转录因子和细胞状态，提供驱动细胞异质性机制的重要生物学见解。

4.1 SCENIC 算法原理

SCENIC 的工作流程主要有 3 步：GRNBoost2，基于共表达确定潜在的 TF 靶标；RcisTarget，进行 TF 基序富集分析并确定直接的靶标（调节子）；AUCell，用于对单个细胞上调节子（或其他基因集）的活性进行评分。下面我们将对每一步算法原理进行介绍。

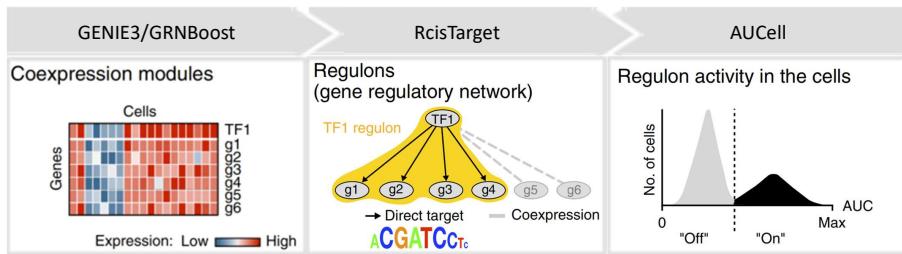


图 4.1 SCENIC 算法流程

4.1.1 GRNBoost2

像 GENIE3[32]一样，GRNBoost2 属于基于回归的 GRN 推理方法 [41]。对于数据集中的每个基因，使用一组候选转录因子（TF）的表达值对基于树的回归模型进行训练，以预测其表达谱。每个模型都会产生部分 GRN，并具有从最佳预测 TF 到目标基因的调控关联。将所有调控关联组合在一起，并按重要性排序，以最终确定 GRN 的输出。

Gradient-Boosting Machines (GBM) [42] 是 GRNBoost2[33] 部分的核心。GBM 是一种用于回归和分类问题的机器学习技术，它以一组弱预测模型（通常为决策树）的形式生成预测模型。当决策树是弱学习者时，结果算法称为梯度提升树，通常胜过随机森林。

GBM 的算法原理与梯度下降类似：梯度下降是在参数空间中寻找一个最优点，而 GBM 则是在函数空间（或者说我们设定的函数集合）寻找一个最优函数。在最优化函数的过程中，我们需要一个优化目标，这通常通过设定损失函数来实现。

在许多有监督的学习问题中，一个输出变量 y 和一个输入向量 x 通过联合概率

分布 $P(x, y)$ 来描述。使用已知 x 和对应 y 所构成的训练集 $\{(x_1, y_1), \dots, (x_n, y_n)\}$ ，目标是找到函数 $F(x)$ 的近似值 $\hat{F}(x)$ 。该近似值应最小化某些指定损失函数 $L(y, F(x))$ 的期望值：

$$\hat{F} = \arg \min_F \mathbb{E}_{x,y}[L(y, F(x))] \quad (4.1)$$

GBM 假设真实值为 y ，并在某些类 \mathcal{H} 中利用函数 $h_i(x)$ 的加权和寻求其的近似值 $\hat{F}(x)$ ，这被称为基学习器：

$$\hat{F}(x) = \sum_{i=1}^M \gamma_i h_i(x) + const \quad (4.2)$$

根据经验风险最小化原理，该方法尝试找到一个近似值 $\hat{F}(x)$ ，该函数将在训练集计算的损失函数平均值最小化，即最小化经验风险。这一过程是从一个由常数函数 $F_0(x)$ 组成的模型开始的，然后以贪心算法的过程逐步扩展它：

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma) \quad (4.3)$$

$$F_m(x) = F_{m-1}(x) + \arg \min_{h_m \in \mathcal{H}} \left[\sum_{i=1}^n L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right] \quad (4.4)$$

这里， $h_m \in \mathcal{H}$ 是一个基学习器函数。

但上面这个过程是在计算上不可行的优化问题，GBM 便将其进一步简化。这个想法便是对这个最小化问题（功能梯度下降）应用最陡峭的下降步骤。如果我们考虑连续的情况，也就是说， \mathcal{H} 在 \mathbb{R} 上是任意微分函数的集合，我们将根据以下方程式更新模型：

$$F_m(x) = F_{m-1}(x) - \gamma \sum_{i=1}^n \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i)) \quad (4.5)$$

$$\gamma_m = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i))) \quad (4.6)$$

这里，微分是对函数 $F_i (i \in \{1, \dots, m\})$ 所计算的来的， γ 是步长。但是，在离散情况下，即当集合 \mathcal{H} 只有有限个元素时，我们选择最接近 L 梯度的候选函数 h ，然后可以根据上述等式通过线搜索来为其计算系数 γ 。但这种方法是一种启发式的方法，无法给出给定问题的精确解决方案，只能提供一个近似的解，其伪代码如算法1所示。

Input: training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, number of iterations M

Output: $F_M(x)$

Initialize model with a constant value:

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma)$$

for $m \leftarrow 1$ **to** M **do**

Compute so-called pseudo-residuals:

$$r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right] \text{ for } i = 1, \dots, n$$

Fit a base learner(e.g. tree) $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^n$.

Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

end

Return $F_M(x)$.

算法 1 Gradient-Boosting Machine

4.1.2 RcisTarget

RcisTarget 可为基因列表识别富集的 TF 结合基序和候选转录因子。简而言之, RcisTarget 基于两个步骤。首先, 它选择在基因组中的基因转录起始位点 (TSS) 的周围环境中显着过量表达的 DNA 基序。这是通过在数据库中应用基于恢复的方法来实现的, 该数据库包含每个基序的全基因组跨物种排名。保留了注释为相应 TF 并获得归一化富集得分 (NES) > 3.0 的基序。接下来, 对于每个基序和基因组, RcisTarget 预测候选目标基因 (即, 基因组中排在最前沿的基因)。该方法基于 Aerts 等人 [43] 描述的方法, 该方法也可以在 i-cisTarget (网络界面) [44] 和 iRegulon (Cytoscape 插件) [45] 中实现。因此, 当使用相同的参数和数据库时, RcisTarget 可提供与 i-cisTarget 或 iRegulon 相同的结果, 并以 Janky 等人 [45] 中的其他 TFBS 富

集工具为基准。

4.1.3 AUCell

AUCell 可帮助研究人员在单细胞 RNA 测序数据中鉴定具有活跃基因调控网络的细胞。AUCell 的输入是一个基因集，输出是每个细胞中“活跃”的基因集。在 SCENIC 中，这些基因集是调控子，由 TF 及其推定的靶标组成。

AUCell 会计算跨越特定细胞中所有基因排名的恢复曲线下的区域，并将其作为调节子的富集程度，从而根据该表达值对它们进行排名。然后，AUCell 使用 AUC 来计算输入基因集的关键子集是否在每个细胞的排名顶部都得到了富集。通过这种方式，AUC 代表了基因签名中表达基因的比例及其与细胞内其他基因相比的相对表达值。

4.2 选择 SCENIC 算法的原因

由于细胞的状态是受基因调节子调控的，其内部是一个十分复杂的调控网络。我们应当考虑这种潜在的调节网络，以获取更加生物合理的细胞状态定义，因为它对于我们总结结论有着重要的参考作用。我们希望选择的 GRN 推断算法应具备以下优势：

- 聚类结果应与细胞真实的种类有着较好的匹配度，即较高的调整兰德指数 (ARI)。
- 所推断的 GRN 应具备较高鲁棒性，即数据集大小不会对检测出的转录因子有很大影响。
- 具有较低的算法复杂度，对于大数据集仍可以在较短的时间内完成基因调控网络推断。

作为最新提出的 GRN 推断算法 SCENIC，具备以上的所有优势。Aibar 在其 2017 年的工作 [10] 中就以上标准与以往的 GRN 推断算法进行了比较。为了测试 SCENIC 算法的性能，Aibar 使用一个公开的人为标注数据集作为基线，统计所有 GRN 推断方法所得到聚类结果的 ARI。其中，ARI 定义如下：

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]}{\binom{n}{2}}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - \frac{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]}{\binom{n}{2}}} \quad (4.7)$$

其中，设 $X_i (i \in 1, \dots, r)$ 表示真实类别， $Y_j (j \in 1, \dots, s)$ 表示预测类别，则式中 n_{ij} 表示本属于 X_i 类被预测为 Y_j 类数据的数量， a_i 表示数据集中属于 X_i 类数据的

数量， b_j 表示数据集中被预测为 Y_j 类数据的数量。由其统计结果（图4.2a）可见，SCENIC 算法的聚类结果相比其它 GRN 推断算法得到的结果更为精确。

为了测试 SCENIC 算法的鲁棒性，Aibar 在比较性能时使用的数据集中随机抽取 100 个数据组成一个较小的数据集。其将各种 GRN 推断方法在原数据集发掘出的转录因子作为基线，统计小数据集上发掘出转录因子的召回率和准确率。尤其统计结果（图4.2b）可见，SCENIC 算法相比其它 GRN 推断算法具备较高的鲁棒性。

最后，在对大数据集的处理效率方面，SCENIC 算法中使用了 GENIE3 算法 [32] 更为高效的变体——GRNBoost2 算法 [33]。GRNBoost2 基于与 GENIE3 相同的概念，纯粹从基因表达矩阵中推断每个靶基因的调节子。但是，GRNBoost2 使用 Gradient-Boosting Machines (GBM) [42] 代替了随机森林模型。这种实现方式大大减少了推断 GRN 所需的时间（图4.2c），并为在非常大的数据集上进行网络推断铺平道路。

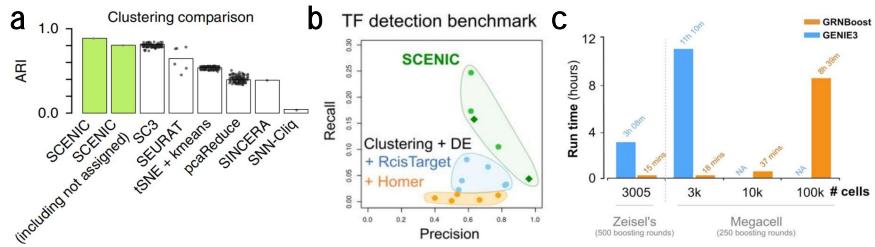


图 4.2 SCENIC 算法与其他 GRN 推断算法的比较

5 实验数据分析

5.1 测序数据预处理

通过图5.1a 所示的流程，我们获取了小鼠垂体细胞的单细胞测序数据。将测序得到的 fastq 文件利用 CellRanger 进行上游分析，序列回帖参考基因组选用 Ensembl (GRCm38)。回帖得到的基因表达矩阵通过 Scater[46] 进行质量控制，筛除低质量细胞，再用 Seurat v3[8, 9] 进行基础下游分析。

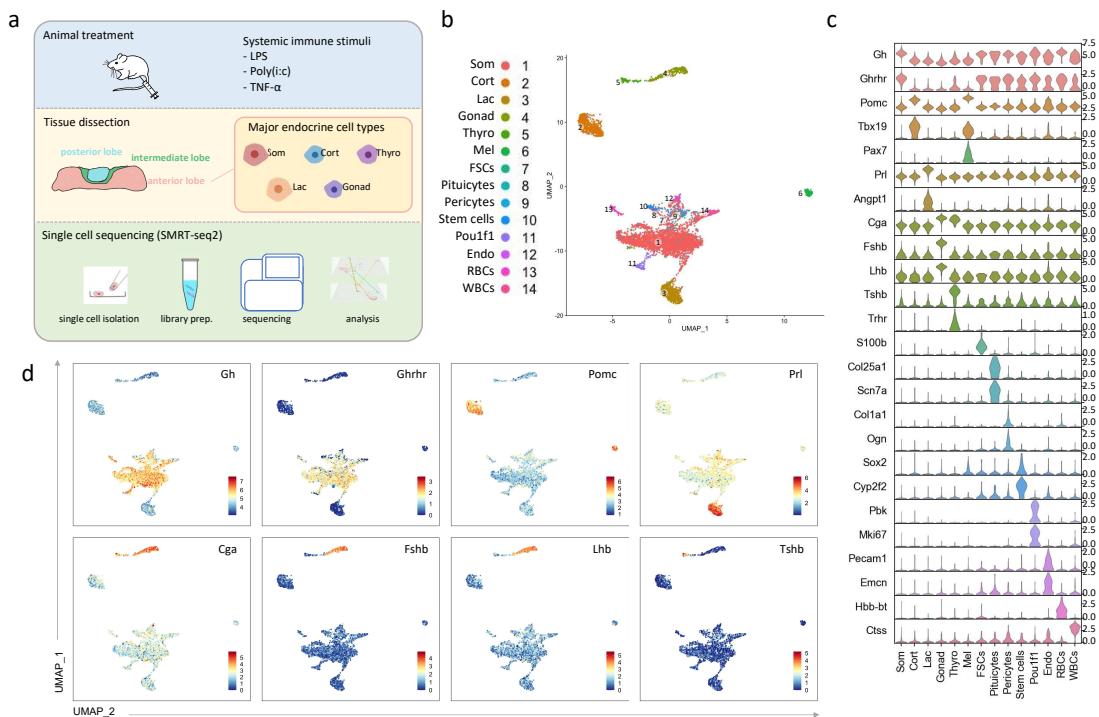


图 5.1 单细胞测序数据预处理

5.1.1 Scater 质控

将测序数据读入 R 工作环境中，将表达矩阵转换为 SingleCellExperiment (SCE) 对象，对该 SCE 对象依次在细胞水平 (Cell-level QC)、基因水平 (Feature-level QC) 以及变量水平 (Variable-level QC) 执行质控。在细胞水平质控中，去除文库含量小 (low library size)、基因含量少 (low features) 和线粒体基因含量高 (high mitochondrial percentage) 的细胞；在基因水平质控中，去除表达量为 0 的基因以及线粒体基因、核糖体基因，以避免在 PCA 部分影响主成分判别；在变量水平中，对每个变量方差解释的贡献率 (variance explained) 做统计，查看是否有异常变量

出现。

通过 STAR-featureCounts 上游管线处理后，我们共收集到 6727 个细胞，平均每个细胞检测到 3546 个基因。在经过 Scater 质控后，保留了 5506 个细胞。

5.1.2 Seurat 初步分析

经过质控处理后的数据转换为 Seurat 对象，执行 Seurat 基础分析流程：归一化、特征选择、放缩、线性降维、构建最短临近图、Leiden 聚类以及 UMAP 可视化。

在此基础上，结合聚类信息以及中枢神经系统各细胞类型已知的基因 marker（表5.1），对数据进行细胞类型注释。我们在垂体腺中鉴定了 6 个主要细胞簇（Somatotropes, Corticotropes, Melanotropes, Lactotropes, Thyrotropes, Gonadotropes），这与先前的知识是一致的。详见图5.1。

表 5.1 中枢神经系统各细胞类型已知的基因 marker

细胞类型	基因 marker
Somatotropes	Gh, Ghrhr, Pappa2, Gnm
Lactotropes	Prl, Angpt1
Corticotropes	Pomc, Crhr1, Tbx19
Melanotropes	Pomc, Tbx19, Pax7, Pcsk2, Rbfox3
Gonadotropes	Fshb, Lhb, Gnrhr, Cga, Nr5a1
Thyrotropes	Tshb, Trhr, Cga
Pou1f1 progenitors	Pbk, Top2a, Mki67
RBCs	Hbb-bt, Hbb-bs
WBCs	C1qa, Ctss, Ptprc
Folliculostellate cells	S100b, Fxyd1
Endothelial cells	Pecam1, Emen, Plvap
Pituicytes	Gja1, Scn7a, Col25a1
Pericytes	Colia1, Dcn, Ogn, Lum, Pdgfrb
Stem cells	Sox2, Aldh3a1, Aldh1a2, Cgp2f2

在这项研究中，我们主要关注的是系统性神经炎症对于垂体细胞的单细胞转录水平影响。因而，我们依据上面得到的细胞注释对数据进行筛选，只留下 Somatotropes、Corticotropes、Lactotropes、Thyrotropes 以及 Gonadotropes 五类细胞，共 3788 个细胞。我们之后的分析便都以此数据上开展。

5.2 测序数据 SCENIC 分析

我们对筛选出来的测序数据进行 SCENIC 分析，以推断其潜在转录因子及对应目的基因。这一步的输出是一个 AUC-score 矩阵，矩阵的每一个单元表示每一个细胞中对每个基因集合的整体趋势评分，当其符号为正时上调，反之下调。我们对该矩阵进行降维聚类，并用其聚类结果作为细胞是否处于炎症状态的判别标准，其原因已在相关工作的基因调控网络部分说明。使用该聚类结果标注的基因表达矩阵降维结果展示在图5.2a,b。对于每一类细胞，该分类结果可以很好地匹配 UMAP 可视化后的数据分布。

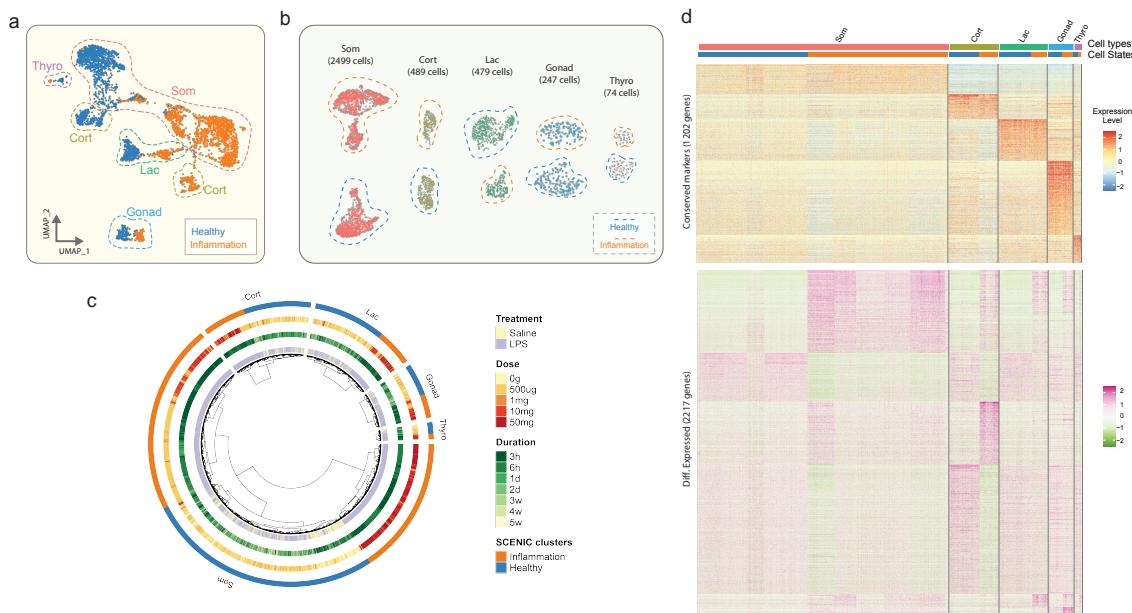


图 5.2 单细胞测序数据 SCENIC 分析

5.2.1 分析处理条件与垂体细胞状态之间的关系

在我们所构建小鼠炎症模型中，我们依据免疫刺激剂、给药剂量与恢复时间尺度等因素建立了一系列实验。为了揭示这一系列因素与垂体细胞状态之间的关系，我们将这些处理条件与 SCENIC 聚类结果进行整合，如图5.2c 所示。

我们可以看到所有注射 saline 的处理组，基本都处于健康 (healthy) 状态。除此之外，还有注射 LPS 的处理组也对应到了健康 (healthy) 状态。通过观察给药剂量与恢复时间尺度因素，我们可以发现这些细胞大多是注射低剂量 LPS 或者经历了长时程的恢复，在这种情况下神经系统仅有少量垂体细胞处于免疫应激状态。相比之下，注射高剂量 LPS 且仅经历短时程恢复的处理组则大多处于炎症 (inflammation) 状态。

5.2.2 分析不同细胞在炎症状态下的基因表达差异

我们进一步比较了垂体中不同细胞在炎症状态下的基因表达差异，见图5.2d。该图的上半部分是垂体各类细胞在其对应基因 marker 上的基因表达热图。同类细胞对应的基因 marker 无论中枢神经系统是否处于炎症状态，都会在该类细胞中稳定表达。然后，我们对每一类细胞分析其炎症状态与健康状态下的差异表达基因集合，并将其整合，便得到了该图的下半部分。

我们发现同处于炎症状态，垂体中不同细胞应对炎症所做出的基因表达调整并不一致。例如，在炎症状态 Somatotropes 中上调最显著的基因集合并不是在炎症状态 Corticotropes 中上调最显著的基因集合。这便说明，在中枢神经内分泌系统处于炎症状态时，垂体内各类细胞会采取不同的应激方式，组成一个调节炎症反应的复杂系统。

5.2.3 分析导致炎症状态的转录因子

我们依据 SCENIC 过程得到的 AUC-score 矩阵，统计出每个转录因子的 AUC-score 密度分布，我们希望找到具备双峰分布或者重尾分布的转录因子。对这些转录因子的分布进行自适应二值化，我们发现其标签可以和之前使用 SCENIC 聚类结果判定的细胞状态很好地匹配起来（见图5.3）。

我们通过上面过程找到的转录因子涉及 Stat、Irf 以及 NfkB 等转录因子家族，这些转录因子大多是与免疫过程相关的。例如，Irf7 编码干扰素调节因子 7，以往实验数据表明其在病毒诱导的细胞基因（包括 I 型干扰素基因）的转录中起作用。这些转录因子并没有像之前分析的差异表达基因那样展现出细胞种类特异性，而是在整个炎症状态的垂体细胞中广泛表达。

这就表明 Stat、Irf 以及 NfkB 等转录因子家族在垂体参与中枢神经内分泌炎症调节过程中扮演着重要的角色，影响着各类垂体细胞的调控路径。换句话说，Stat、Irf 以及 NfkB 等转录因子家族是垂体参与中枢神经内分泌炎症调节过程中的 Master Regulator Genes(MRs)[47]。

5.3 讨论和未来工作

我们在实验中发现，在给以小鼠 $TNF - \alpha$ 刺激之后，其部分垂体细胞在经历 UMAP 可视化降维之后，呈现出与其他炎症状态细胞相分离的现象。这似乎意味着垂体细胞在参与中枢神经内分泌炎症调节过程中，会因免疫刺激不同而进入不同的调节状态。

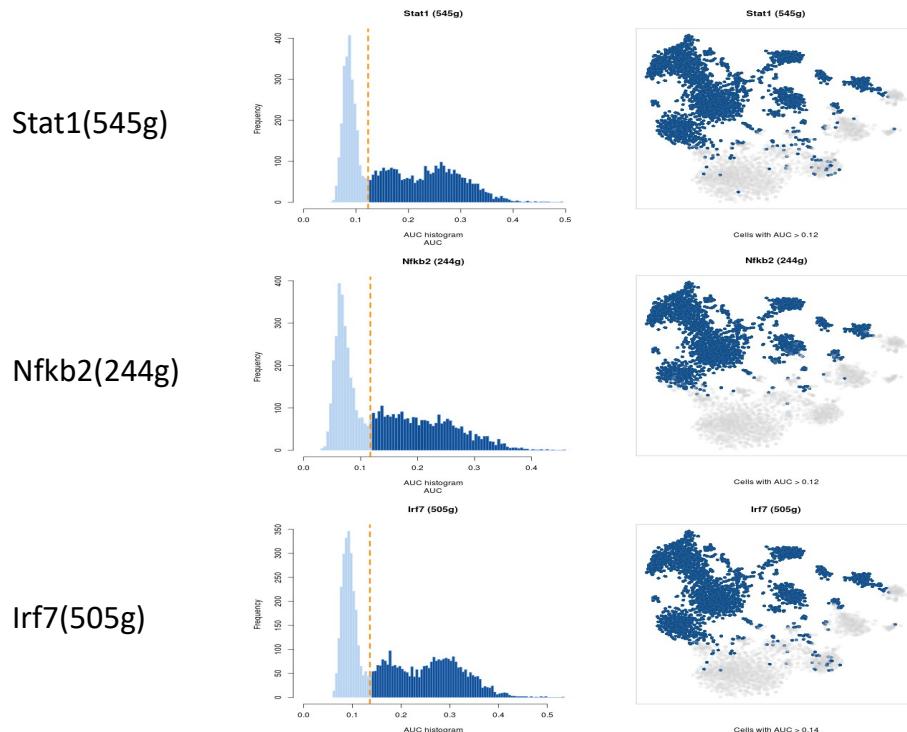


图 5.3 具备双峰分布或者重尾分布的转录因子

我们初步推断是 $TNF - \alpha$ 作为一种较强的免疫刺激剂，使垂体细胞进入一种不可逆转的免疫状态。在实验设计上，针对 LPS、saline 与 Ploy I:C，我们都设计了长短恢复时程的对比试验，但 $TNF - \alpha$ 只有短时程恢复处理。其原因在于 $TNF - \alpha$ 相比 LPS、Ploy I:C 产生的免疫反应过于剧烈，小鼠在被腹腔注射 500ug $TNF - \alpha$ 之后无法恢复，6h 内便会死亡。

我们未来的工作便打算在该研究的基础上，进一步探讨面对不可恢复炎症刺激与可恢复炎症刺激时垂体细胞在转录水平上的差异，揭示由健康 (healthy) 状态向这两种炎症 (inflammation) 状态转变的关键转录因子。

6 总结与展望

诸如病毒或细菌感染之类的免疫挑战会引起组织炎症，并对下丘脑-垂体-肾上腺（HPA）轴产生深远影响。其中，垂体是大脑的内分泌中心，并积极参与炎症事件的调节。在这项研究中，我们对免疫攻击过程中垂体细胞的转录反应进行了探究。

首先，这项工作提供了不同种类垂体细胞都参与中枢神经内分泌炎症调节过程的单细胞转录层级证据。LPS 和 Poly I:C 在炎症过程中是有效的免疫刺激，将此类免疫激活剂应用于小鼠科引起严重的免疫反应。虽然以往的研究也使用这些刺激建立了小鼠炎症模型，但是其所收集到的数据集局限于垂体组织水平，无法证明不同垂体细胞均参与到炎症调节过程中。我的合作者则建立了其单细胞转录水平的数据集，弥补了垂体单细胞测序数据在炎症状态下的空缺。我在对该数据集分析时发现，垂体中不同种类细胞在炎症状态和非炎症状态下展现出较大的转录差异，表明不同种类垂体细胞都积极参与到中枢神经内分泌炎症的调节过程中。

其次，这项工作揭示了不同种类垂体细胞在参与中枢神经内分泌炎症调节的过程中的转录水平差异，表明其在炎症调节过程中扮演不同的角色。我对收集到的测序数据进行基因调控网络推断，得到每一个细胞对于各基因调控通路的 AUC-score 矩阵。并依据此进行了 Leiden 聚类，将聚类结果作为判别细胞处于健康（healthy）状态还是炎症（inflammation）状态的标准。我对不同种类垂体细胞在两种细胞状态下的基因表达差异进行了分析，发现不同种类垂体细胞在炎症状态下表达量主要调整的基因集合并不相近。这项工作从转录水平上证明了不同细胞在参与中枢神经内分泌炎症调节的过程中扮演着不同的角色。

此外，这项工作发现了一类在不同种类垂体细胞中统一表达的转录因子，表明其在垂体参与中枢神经内分泌炎症调节过程中的重要地位。在转录因子的 AUC-score 密度分布中，我找出具备双峰分布或者重尾分布的转录因子，比如 Stat、Irf 和 NfkB 等转录因子家族，这些转录因子大多是与免疫过程相关的。我发现这些转录因子在不同种类垂体细胞中有统一的表达，这表明其在垂体参与中枢神经内分泌炎症调节的过程中在多条调节路径上扮演着重要角色。

最后，我在对转录因子 AUC-score 矩阵进行统计的时候，发现 $TNF - \alpha$ 刺激组在聚类结果中呈现出与其他免疫刺激组相互分离的情况。这暗示由 $TNF - \alpha$ 介

导的免疫状态转变可能涉及不同的关键转录因子，而这些转录因子将使细胞进入一种不可逆转的状态。找到这些关键转录因子将为我们设计相应的调节剂提供重要见解，有助于研发相应的抗炎症药物。

总之，在论文中从单细胞转录水平分析了垂体细胞在参与中枢神经内分泌炎症调节过程中的共性与差异，并为未来的中枢神经内分泌炎症调节过程研究分析了方向，作为我们的未来工作。

参考文献

- [1] CHROUSOS G P. The hypothalamic–pituitary–adrenal axis and immune-mediated inflammation[J]. New England Journal of Medicine, 1995, 332(20) : 1351 – 1363.
- [2] SHANKS N, WINDLE R J, PERKS P A, et al. Early-life exposure to endotoxin alters hypothalamic–pituitary–adrenal function and predisposition to inflammation[J]. Proceedings of the National Academy of Sciences, 2000, 97(10) : 5645 – 5650.
- [3] SVENSSON V, VENTO-TORMO R, TEICHMANN S A. Exponential scaling of single-cell RNA-seq in the past decade[J]. Nature protocols, 2018, 13(4) : 599 – 604.
- [4] CHEN Q, LESHKOWITZ D, BLECHMAN J, et al. Single-cell molecular and cellular architecture of the mouse neurohypophysis[J]. Eneuro, 2020, 7(1).
- [5] CHEUNG L Y, GEORGE A S, MCGEE S R, et al. Single-cell RNA sequencing reveals novel markers of male pituitary stem cells and hormone-producing cell types[J]. Endocrinology, 2018, 159(12) : 3910 – 3924.
- [6] HO Y, HU P, PEEL M T, et al. Single-cell transcriptomic analysis of adult mouse pituitary reveals sexual dimorphism and physiologic demand-induced cellular plasticity[J]. Protein & Cell, 2020 : 1 – 19.
- [7] FLETCHER P A, SMILJANIC K, MASO PRÉVIDE R, et al. Cell type-and sex-dependent transcriptome profiles of rat anterior pituitary cells[J]. Frontiers in Endocrinology, 2019, 10 : 623.
- [8] BUTLER A, HOFFMAN P, SMIBERT P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species[J]. Nature biotechnology, 2018, 36(5) : 411 – 420.
- [9] STUART T, BUTLER A, HOFFMAN P, et al. Comprehensive integration of single-cell data[J]. Cell, 2019, 177(7) : 1888 – 1902.
- [10] AIBAR S, GONZÁLEZ-BLAS C B, MOERMAN T, et al. SCENIC: single-cell regulatory network inference and clustering[J]. Nature methods, 2017, 14(11) : 1083 – 1086.
- [11] Van de SANDE B, FLERIN C, DAVIE K, et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis[J]. Nature Protocols, 2020, 15(7) :

- [12] HAMMOND T R, DUFORT C, DISSING-OLESEN L, et al. Single-cell RNA sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes[J]. *Immunity*, 2019, 50(1): 253–271.
- [13] KEREN-SHAUL H, SPINRAD A, WEINER A, et al. A unique microglia type associated with restricting development of Alzheimer’s disease[J]. *Cell*, 2017, 169(7): 1276–1290.
- [14] LI Q, CHENG Z, ZHOU L, et al. Developmental heterogeneity of microglia and brain myeloid cells revealed by deep single-cell RNA sequencing[J]. *Neuron*, 2019, 101(2): 207–223.
- [15] MASUDA T, SANKOWSKI R, STASZEWSKI O, et al. Spatial and temporal heterogeneity of mouse and human microglia at single-cell resolution[J]. *Nature*, 2019, 566(7744): 388–392.
- [16] MASUDA T, AMANN L, SANKOWSKI R, et al. Novel Hexb-based tools for studying microglia in the CNS[J]. *Nature Immunology*, 2020, 21(7): 802–815.
- [17] MATCOVITCH-NATAN O, WINTER D R, GILADI A, et al. Microglia development follows a stepwise program to regulate brain homeostasis[J]. *Science*, 2016, 353(6301).
- [18] LIU S, TRAPNELL C. Single-cell transcriptome sequencing: recent advances and remaining challenges[J]. *F1000Research*, 2016, 5.
- [19] JUNKER J P, van OUDENAARDEN A. Every cell is special: genome-wide studies add a new dimension to single-cell biology[J]. *Cell*, 2014, 157(1): 8–11.
- [20] KURIMOTO K, YABUTA Y, OHINATA Y, et al. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis[J]. *Nucleic acids research*, 2006, 34(5): e42–e42.
- [21] TANG F, BARBACIORU C, WANG Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell[J]. *Nature methods*, 2009, 6(5): 377–382.
- [22] RAMSKÖLD D, LUO S, WANG Y-C, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells[J]. *Nature biotechnology*, 2012, 30(8): 777–782.

- [23] PICELLI S, BJÖRKLUND Å K, FARIDANI O R, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells[J]. *Nature methods*, 2013, 10(11) : 1096–1098.
- [24] ISLAM S, KJÄLLQUIST U, MOLINER A, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq[J]. *Genome research*, 2011, 21(7) : 1160–1167.
- [25] FIERS M W, MINNOYE L, AIBAR S, et al. Mapping gene regulatory networks from single-cell omics data[J]. *Briefings in functional genomics*, 2018, 17(4) : 246–254.
- [26] ARENDT D, MUSSER J M, BAKER C V, et al. The origin and evolution of cell types[J]. *Nature Reviews Genetics*, 2016, 17(12) : 744–757.
- [27] TAKAHASHI K, YAMANAKA S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors[J]. *cell*, 2006, 126(4) : 663–676.
- [28] MARRO S, PANG Z P, YANG N, et al. Direct lineage conversion of terminally differentiated hepatocytes to functional neurons[J]. *Cell stem cell*, 2011, 9(4) : 374–382.
- [29] IEDA M, FU J-D, DELGADO-OLGUIN P, et al. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors[J]. *Cell*, 2010, 142(3) : 375–386.
- [30] CREIXELL P, SCHOOF E M, ERLER J T, et al. Navigating cancer network attractors for tumor-specific therapy[J]. *Nature biotechnology*, 2012, 30(9) : 842–848.
- [31] WOUTERS J, ATAK Z K, AERTS S. Decoding transcriptional states in cancer[J]. *Current opinion in genetics & development*, 2017, 43 : 82–92.
- [32] HUYNH-THU V A, IRRTHUM A, WEHENKEL L, et al. Inferring regulatory networks from expression data using tree-based methods[J]. *PloS one*, 2010, 5(9) : 1–10.
- [33] MOERMAN T, AIBAR SANTOS S, BRAVO GONZÁLEZ-BLAS C, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks[J]. *Bioinformatics*, 2019, 35(12) : 2159–2161.

- [34] PRATAPA A, JALIHAL A P, LAW J N, et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data[J]. *Nature methods*, 2020, 17(2) : 147–154.
- [35] BRENNCKE P, ANDERS S, KIM J K, et al. Accounting for technical noise in single-cell RNA-seq experiments[J]. *Nature methods*, 2013, 10(11) : 1093.
- [36] TRAAG V A, WALTMAN L, VAN ECK N J. From Louvain to Leiden: guaranteeing well-connected communities[J]. *Scientific reports*, 2019, 9(1) : 1–12.
- [37] BLONDEL V D, GUILLAUME J-L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. *Journal of statistical mechanics: theory and experiment*, 2008, 2008(10) : P10008.
- [38] OZAKI N, TEZUKA H, INABA M. A simple acceleration method for the Louvain algorithm[J]. *International Journal of Computer and Electrical Engineering*, 2016, 8(3) : 207.
- [39] BAE S-H, HALPERIN D, WEST J D, et al. Scalable and efficient flow-based community detection for large-scale graph analysis[J]. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2017, 11(3) : 1–30.
- [40] TRAAG V A. Faster unfolding of communities: Speeding up the Louvain algorithm[J]. *Physical Review E*, 2015, 92(3) : 032801.
- [41] SANGUINETTI G, OTHERS. Gene regulatory network inference: an introductory survey[G] // *Gene Regulatory Networks*. [S.l.] : Springer, 2019 : 1–23.
- [42] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. *Annals of statistics*, 2001 : 1189–1232.
- [43] AERTS S, QUAN X-J, CLAEYS A, et al. Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in *Drosophila* uncovers a regulatory basis for sensory specification[J]. *PLoS Biol*, 2010, 8(7) : e1000435.
- [44] HERRMANN C, Van de SANDE B, POTIER D, et al. i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules[J]. *Nucleic acids research*, 2012, 40(15) : e114–e114.
- [45] VERFAILLIE A, IMRICOVÁ H, Van de SANDE B, et al. iRegulon: from a gene

- list to a gene regulatory network using large motif and track collections[J]. PLoS Comput Biol, 2014, 10(7) : e1003731.
- [46] MCCARTHY D J, CAMPBELL K R, LUN A T, et al. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R[J]. Bioinformatics, 2017, 33(8) : 1179–1186.
- [47] MATTICK J S, TAFT R J, FAULKNER G J. A global view of genomic information—moving beyond the gene and the master regulator[J]. Trends in genetics, 2010, 26(1) : 21–28.

致谢

感谢蔡朝晖教授。她是大学期间为我们代课最多的老师，在我大学的学习成长过程和毕业论文撰写过程中提供了很大的指导和帮助。大一入学常听陈子轩和原昊博提起蔡老师，经过一段时间的接触，果然对学生十分负责！感谢蔡老师大学四年对我的信任，让我在计算机本科学习中更加自信。

感谢北京脑科学与类脑研究中心的罗敏敏教授。他在毕业论文撰写过程中给我提供了很大指导和帮助。我很敬佩罗老师对于科学的近乎狂热般的痴迷与对学生的关心。出于对神经科学的痴迷，我在保研的时候选择放弃我所擅长的计算机系统领域，转而投入神经科学领域。在刚保研完的时候，我深知我在神经科学领域的基础十分薄弱，便向罗老师申请到其实验室做毕设。罗老师并没有排斥一个本科其他专业的学生，而是在实验室本已满员的情况下十分认真地为我安排了一个相关方向的学长来指导我，让我更多地参与到实验室的课题中来，这为我以后从事神经科学领域研究打下了坚实的基础。

与北京大学王睿宇学长的讨论让这篇论文更加完善。我们之后在神经科学领域会持续合作。

感谢武汉大学的陈丹教授。大三上学期，选了他的一门专业选修课——计算机体系结构。陈老师在讲课的时候不拘泥于课本内容，而是十分注重培养我们的科研阅读素养。在与陈老师的交谈中，我了解到了很多体系结构前沿的研究，比如存内计算、类脑计算等，极大地开阔了我对于计算机系统的认识。也是通过陈老师的课，我对类脑计算、脑机接口与计算神经学等领域之间的联系有了初步的认知。最终走上科研道路、决定读博并且申请到北京大学前沿交叉学科研究院的PhD，陈老师起了巨大的引导作用。

感谢武汉大学的艾浩军教授。大三下学期，由艾老师指导，和李蕴哲、朱赫合作完成的空中手写字符迁移学习项目最终成功发表，成为我人生中的第一篇论文！当时由于疫情，我们所有的交谈都被局限在线上。但艾老师每周都会与我们进行两小时以上的进展沟通，并对我们的工作提出建设性的指导意见。投稿前艾老师帮我们反复修改，在进行线上会议之前，他多次帮助我们修改海报以及展示视频，最终成功的展示离不开他的热心帮助。我们一直保持着联系。

感谢浙江大学的潘纲教授。虽然接触的时间不长，但他的研究态度给我留下

了深刻的印象——我们的电话交流永远发生在凌晨。实验室的博士生谭显瀚和祝歆韵日后都会是优秀的类脑计算研究者！这一段实习经历也让我更加坚定了从事神经科学的研究的决心。

感谢北京脑科学与类脑研究中心的周景峰教授，也是我未来的博士导师。在罗老师实验室便一直听说周师兄读博期间的各种经历，被实验室的师兄师姐一致称赞。之前在与他的交流中，我能深刻感受到他对于神经科学独到的理解与深厚的跨学科背景。虽然接触的时间还不长，但我能感受到他完美的性格！相信我们会有愉快、高产的合作。

感谢北京生命科学联合中心的吴思教授和北京脑科学与类脑研究中心的柳昀哲教授。在与他们的交流中，我更加坚定了从神经元层级研究 schema 表示与修正过程的决心。我会在博士一年级到他们的实验室进行轮转。

感谢北京生命科学研究所的王睿宇、卢立辉、袁正巍、刘志祥、曾佳为、黎亨、左鹏、于涛、全竞、严婷等同学。我十分享受与他们的每一次交流，他们对科学的严谨态度对我产生了深远的影响。

感谢彭鹏、朱赫、李蕴哲、范文騫、章博文、周稚璇、陈子轩、原昊博等同学。感谢院学生会的同事们、WHU-MSC 的朋友们、WHU-ICRobo 的队友们，感谢所有的朋友。你们给我留下了永远的美好回忆！

感谢武汉大学和弘毅学堂。学院“宽口径、厚基础、强能力”的教育方针，为我从事交叉学科的研究打下了坚实的基础。感谢弘毅学堂石兢、方萍、李瑶、董甲庆老师和辅导员王璐。

感谢父母和家人长期的支持和鼓励。没有你们，我不会取得今天的成绩！

四年的时光弹指一挥间，从青涩地踏进校园，到即将本科毕业，步入博士生涯。感谢过得飞快的时间，告诉我要不断努力，永不止步！

最后用我很喜欢的一句话结束。感谢陈立杰的这句话。“能够生在这样一个黄金时代里，我感到无比的荣幸。我梦想能够成为黄金时代浪潮中的一朵浪花，为人类的智慧添砖加瓦！”