华中农业大学
HUAZHONG AGRICULTURAL UNIVERSITY

系统遗传学课程论文

孤独症和癫痫星形胶质细胞的差异表达基因的富集、网络构建和 Hub 基因识别

姓名：张逸东

学院：信息学院

学号：2023317110014

1 摘要

本研究收集了孤独症、癫痫和共患个体的单细胞核 RNA Seq 测序数据，进行基因定量和细胞注释后，从 118395 个细胞中筛选出 9216 个取自前额叶皮层的星形胶质细胞进行差异表达分析，找出孤独症、癫痫和共患的差异表达基因，分别进行 GO 富集和 KEGG 富集，找出各自共有的和特有的功能或者通路注释，发现患病个体的星形胶质细胞的能量代谢、蛋白转运和翻译以及细胞的形态存在异常；癫痫影响星形胶质细胞的代谢活动的机制和孤独症不同，癫痫可能和胶质细胞对钙离子的响应和转运存在联系；共患个体的星形胶质细胞对类固醇激素、糖皮质激素的响应存在异常；KEGG 富集结果显示这些差异表达基因和其他的一些精神疾病存在交互，从侧面说明了一些精神疾病的并发机制。最后本研究通过对 2531 个差异表达基因进行 WGCNA 分析，识别到两个共表达网络模块，并确定了这两个共表达模块的 hub 基因。

关键词：孤独症　　癫痫　　单细胞核测序　　星形胶质细胞

2 前言

2013 年发表的一项大型研究观察了近 6000 名孤独症儿童，发现其中 12.5%患有癫痫，在 13 岁以上的儿童中，这一比例上升至 26%。2019 年对近 7000 名孤独症儿童的研究发现，约 10%的儿童患有癫痫。其他研究得出的这一数据从 2% - 46%不等。这些数据都超过了癫痫在普通人群中的患病率。癫痫患者也比其他人更有可能患有孤独症，瑞典对 85000 多名癫痫患者的研究发现，这些人的孤独症发病率是普通人的 10 倍(癫痫与自闭症之间有什么关系？(zhihu.com))。

为研究孤独症、癫痫以及共患患者之间的区别或联系，本研究提出了以下问题：1. 孤独症、癫痫以及共患的差异表达基因是否存在某些联系或者区别；2. 这些差异表达基因都富集到哪些通路和功能；3. 差异表达基因是否存在共表达网络模块；4. 如果存在共表达模块，那这些模块的 hub 基因都有哪些。为研究以上问题，本研究收集了孤独症、癫痫以及共患的大脑皮层的单细胞核 RNA Seq 数据，通过定量获得单细胞基因表达矩阵，进行细胞聚类和注释，然后从 118395 个细胞中筛选出 9216 个星形胶质细胞进行差异表达分析、GO 和 KEGG 富集、WGCNA 分析，构建网络，识别共表达模块，并找出共表达网络模块的 hub 基因。

3 研究方法

3.1 数据获取

本研究收集了 Velmeshev 等研究人员(VELMESHEV, D, et al.,2019)用于研究孤独症患者大脑皮层中特定细胞类型的转录组变化的单细胞核 RNA 测序数据。这一批单细胞核测序数据一共有 118 个样本，分别来自于 43 个死亡个体，2 个大脑区域。43 个个体可以分为癫痫(8)、孤独症(7)、共患(8)以及正常(20)四类。118 个样本采样于前额叶皮层(Prefrontal Cortex, PFC)和前扣带皮层(Anterior Cingulate Cortex, ACC)两个大脑区域。

Velmeshev 等人提供了含有 104559 个细胞的关于孤独症、共患和正常个体的基因表达矩阵以及细胞注释信息的数据集，却没有提供癫痫患者和部分正常个体的基因表达矩阵，需要获取相关样本的测序数据进行单细胞核测序上游分析，对细胞和基因进行定量，获得基因表达矩阵，并对细胞进行注释以进行下游分析。

3.2 单细胞核测序数据处理

本研究一共有 13 个样本需要进行单细胞 RNA 测序的上游分析，包括 8 个癫痫患者的 PFC 样本和 5 个后续添加的正常个体的 PFC 样本。

3.2.1 单细胞测序数据预处理

本研究根据 Velmeshev 等研究人员提供的项目编号从 SRA 数据库(https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA434002)下载对应的 13 个单细胞测序原始数据；使用 Sratoolkit (Version=3.0.7) 中的工具 fasterq-dump (--split-3) 将 SRA 文件转化为 fastq 文件。为节省存储空间，使用多线程压缩工具 pigz (Version=2.8) 对 fastq 文件进行压缩；使用 FastqC (Version=0.11.9) 和 MultiQC (Version=1.18) 对其进行质检，序列质量较好，没有进行后续的序列质控。

3.2.2 细胞和基因定量

CellRanger 是 10X genomics 公司为单细胞 RNA 测序分析量身打造的数据分析软件。本研究使用 CellRanger (Version=7.2.0) 的 count 命令对这 13 个样本的 fastq 文件进行细胞和基因的定量分析。本研究使用了版本号为 GRCh38 的人类

基因组以及注释信息作为参考基因组。由于测序样本为单细胞核测序，相较于单细胞测序，前者包含较多的内含子序列，因此在运行 count 命令的时候，需要指定 -include-introns 参数，避免产生比对到的基因数目较低的情况。

使用 Cell Ranger count 完成对每一个样本的细胞和基因的定量以后，使用 cellranger aggr 命令将 13 个样本的基因表达矩阵合并为一个，方便后续的处理和分析。

3.2.3 细胞过滤和细胞类型注释

为研究孤独症、癫痫以及共患的细胞类型特异性差异表达基因，需要将 Velmeshev 提供的具有细胞类型注释的基因表达矩阵和 3.2.2 生成的 13 个样本的基因表达矩阵合并起来，进行细胞过滤和细胞类型注释。

Seurat 是一个用于质控、分析和探索单细胞 RNA-seq 的 R 包，能够用于鉴定和解释来自单细胞转录本测定的异质性来源，并可以整合成多种类型的单细胞数据。本研究使用 Seurat5 (Version=5.0.1)(HAO, Y, et al.,2023)对基因表达矩阵进行过滤，整合和细胞注释。

Velmeshev 提供的基因表达矩阵具有细胞注释信息，并且通过分析发现这些细胞是经过过滤的，因此没有再对这一批数据进行质控。对于 3.2.2 中获得的基因表达矩阵，本研究统计了每一个细胞的基因种类数目，检测到的基因数目和线粒体基因的比例，并根据基因种类数目≥499，线粒体基因比例<5%的标准对细胞进行过滤，细胞数目由原来的 16880 减少到 13836。

质控过后，使用 Seurat5 的 merge 函数将两个数据集合并起来，然后进行归一化，找前 3000 个高变基因，再对这 3000 个高变基因进行标准化，然后进行 PCA 分析，获取前 50 个主成分。两批数据集存在批次效应，本研究使用 Seurat5 中的 IntegrateLayers 函数，指定 CCAIntegration 对 50 个主成分进行批次处理。使用 CCA 批次处理后的前 30 个主成分进行非线性降维，获取二维的 umap 坐标和 tSNE 坐标。

使用批次处理后的 cca 主成分计算细胞的邻接矩阵(k.param=100)，然后进行聚类。一共得到 34 个 cluster。Velmeshev 提供了细胞的注释信息，104559 个细胞一共注释为 17 个细胞类型，本研究根据这些注释信息，对没有确定细胞类

型的细胞进行有监督的注释，具体操作如下：首先，根据已经有的注释信息确定每一个 cluster 更有可能属于哪一个细胞类型，将细胞类型和 cluster 对应起来；然后对于未注释的细胞，它属于哪一个 cluster，就属于那一个对应的细胞类型，以达到细胞类型注释的目的。

3.3 差异表达分析

两批数据合并后一共有 118395 个细胞，注释得到 15 个细胞类型。本研究只选取了采样位置为 PFC，细胞类型为星形胶质细胞(AST)的细胞进行后续分析。用于后续分析的细胞一共有 9216 个，来自 34 个个体，其中 6 个孤独症患者，7 个癫痫患者，7 个共患个体以及 14 个正常人。使用 Seurat5 的 FindMarkers 函数，默认参数分别获得孤独症患者和正常人、癫痫患者和正常人以及共患患者和正常人的差异表的基因，并依据矫正后的 p 值进行过滤(p_val_ad < 0.05)。

对于以上根据校正后的 p 值进行过滤的差异表达基因，本研究定义 log2FC>1 为上调的差异表达基因(up)，log2FC< -1 的基因为下调的差异表达基因(down)，其他的为无显著差异(ns)。

3.4 富集分析

本研究使用 R 包 clusterpRrofiler(Version=4.6.2)(WU, T, et al.,2021)对差异表达基因进行 GO 富集分析和 KEGG 富集分析。对于 GO 富集分析，enrichGO 函数使用的注释包为 org.Hs.eg.db(Version=3.16.0)；对于 Kegg 富集分析，enrichKEGG 指定的 organism 参数为 hsa(Homo Sapiens)。

3.5 WGCNA 分析

本研究使用 R 包 WGCNA(Version=1.72-1)(LANGFELDER, P and HORVATH, S,2008)进行 WGCNA 分析。使用单细胞基因表达矩阵做 WGCNA 分析，计算量大，并且效果很不理想。本研究通过 Seurat5 的 AggregateExpression 函数，将来自同一个个体的细胞的基因表达量直接相加以聚合起来，将个体的基因表达计数矩阵转化为 log(cpm+1) 基因表达矩阵，用于进行 WGCNA 分析。

3.5.1 共表达基因模块识别和网络构建

对于 3.3 中获得的 3 组差异表达基因取并集，从个体基因表达矩阵中选出对应的样本和差异表达基因,确定软阈值以后，使用 blockwiseModules(TOMType = "unsigned",minModuleSize = 10,mergeCutHeight = 0.25)识别网络模块；依次计算基因表达矩阵的邻接矩阵和拓扑重叠矩阵，完成网络的构建。

3.5.2 Hub 基因识别和网络可视化

本研究通过计算 KME 值(module eigengene-based connectivity)来衡量基因和模块的关系，选择|kME| >= 某个阈值(比如 0.99)来筛选出 hub gene，并使用 cytoscape(Version=3.9.1)(SHANNON, P, et al., 2003)对和权重前 500 的边进行网络可视化。

4 结果与结论

4.1 细胞聚类结果

从图 1 可以看出，进行批次处理以后，118395 个细胞聚类为 34 个 cluster，多数 cluster 聚为一团，少部分 cluster，比如 cluster_13 在 tSNE 非线性降维以后可以分成两团相聚相对较远的子 cluster。

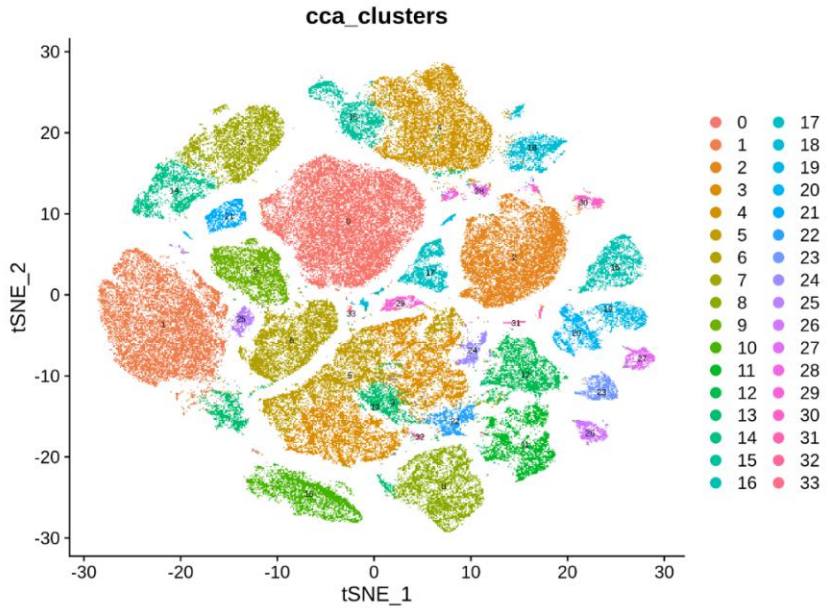

图 1 细胞聚类结果

从图 2 可以看出，有监督地用眼睛对每一个 cluster 属于哪一个细胞类型进行注释效果十分好，相同细胞类型的细胞在 tSNE 坐标上相聚更近。

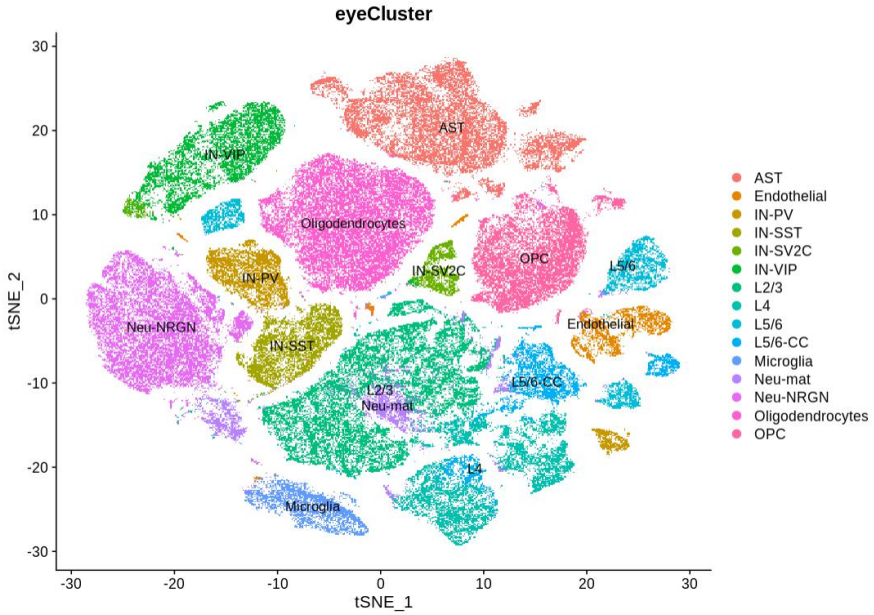

图 2　细胞注释结果

4.2　星形胶状细胞的差异表达分析

如图 3 所示，孤独症一共有 563 个差异表达基因，其中 515 个基因在 ASD 样本中为下调，48 个基因为上调；如图 4 所示癫痫一共有 2254 个差异表达基因，其中 1230 个基因为下调，1024 个基因为上调；如图 5 所示，共患一共有 727 个差异表达基因，其中 694 个为下调，33 个为上调。

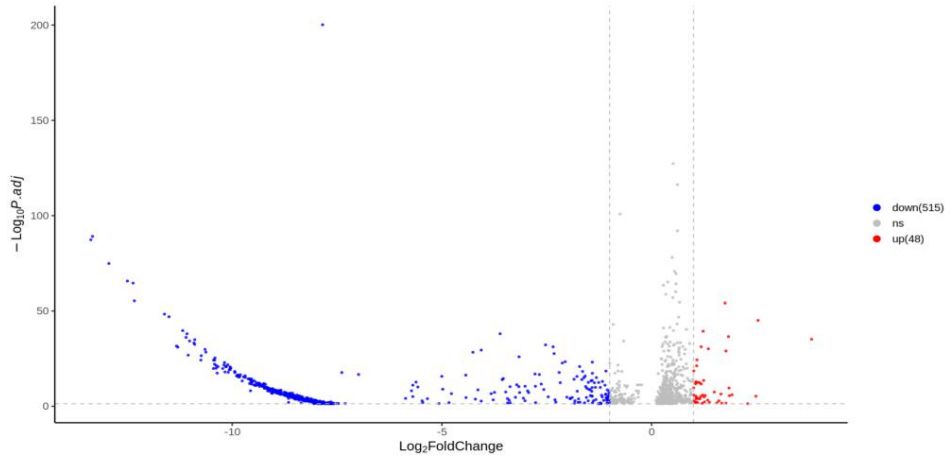

图 3　ASD 和 Control 的火山图

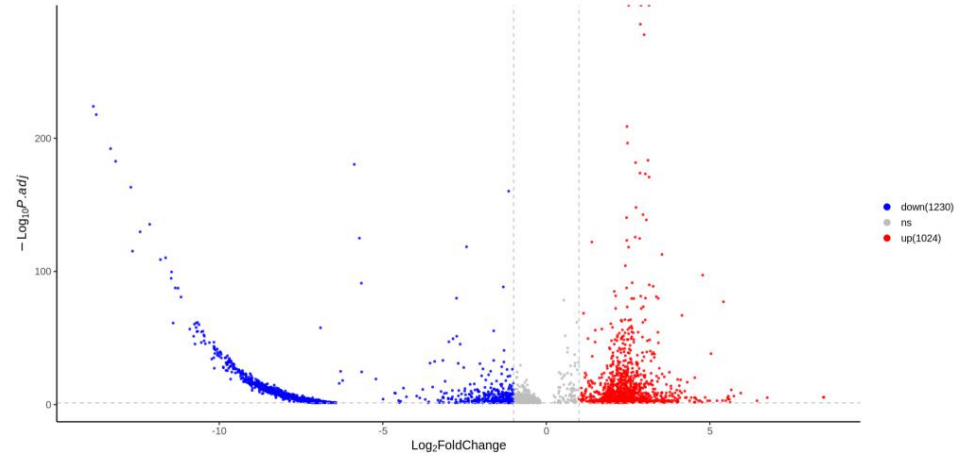

图 4　Seizure 和 Control 的火山图

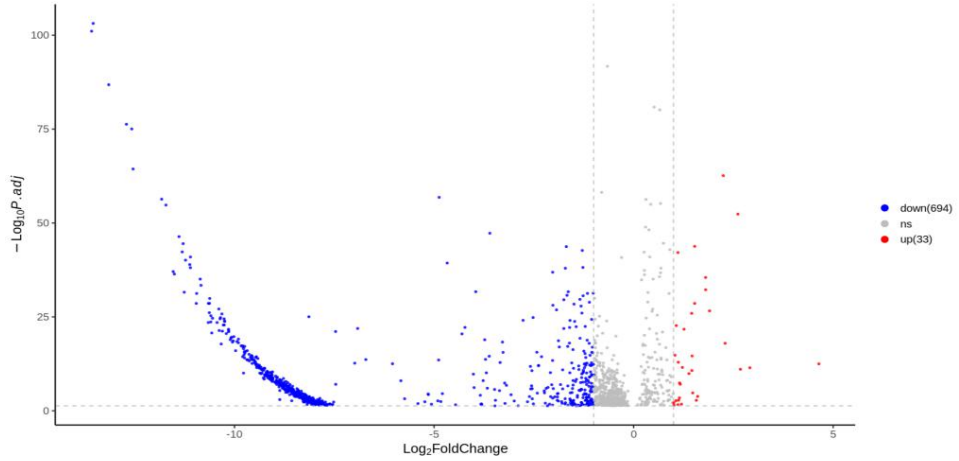

图 5　共患和 Control 的火山图

从火山图可以看出，对于共患和孤独症样本的单细胞差异表达基因的数目相对较少，而癫痫个体的的差异表达基因的数目要远远大于共患和孤独症样本的。并且对于孤独症和共患个体，绝大多数差异表达基因为负调控，而癫痫个体的差异表达基因中正调控和负调控的比例接近 1:1。从火山图中还可以看出，Log2FC 值的变化存在断层，合并后的单细胞 RNA Seq 的差异表达基因的倍数变化不连续。

4.3　差异表达基因富集分析

4.3.1　GO 富集

本研究分别对孤独症，癫痫和共患的差异表达基因做 GO 富集分析，从图 6 可以看出，一共有 41 个 GO Term 是孤独症、癫痫和共患所共有的，孤独症、癫

痫和共患各自特有的 GO Term 的数目分别是 2，105 和 9。

通过对 41 个共有的 GO Term 进行分析发现，这些 Term 涉及到多种蛋白质复合物和酶的活动，与 ATP 的合成和代谢过程，细胞内的呼吸过程、质子跨越细胞膜、蛋白质翻译以及细胞形态的维持相关，由此可以看出，无论是孤独症、癫痫还是共患，处于疾病状态下的星形胶质细胞的能量代谢、蛋白质翻译甚至细胞的形态均存在异常

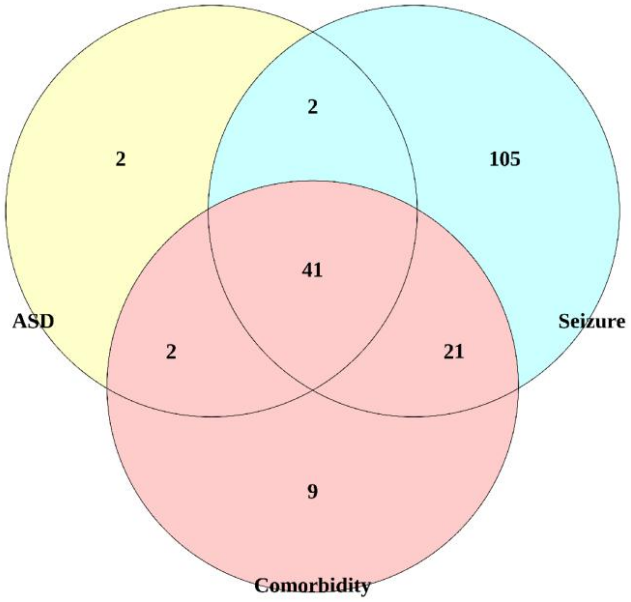

图 6　GO 富集韦恩图

孤独症所特有的两个 GO Term 分别时 cell division site 和 ligase activity。cell division site 的富集表明孤独症个体的星形胶质细胞的细胞周期调控过程过程受到了影响；ligase activity 的富集表型星形胶质细胞的连接酶活性发生了显著变化，这可能涉及到 DNA 修复、蛋白质修饰等分子过程的调节。

癫痫所特有的 GO Term 数目较多，通过分析发现这些 Term 和有氧电子传递、呼吸电子传递、线粒体呼吸链复合物的组装和线粒体电子传递等能量代谢过程相关；与神经元突触的组织和结构相关；与对钙离子的相应和钙离子跨膜转运蛋白的活性相关。由此可以看出，癫痫影响星形胶质细胞的代谢活动的机制和孤独症不同；癫痫可能和胶质细胞对钙离子的反应和转运存在联系。

对于共患的差异表达基因，通过分析发现这些 GO Term 和蛋白质折叠相关，

和 MHC I 类蛋白质结合相关；并且共患个体的星形胶质细胞对类固醇激素、糖皮质激素的响应存在异常。

4.3.2 KEGG 富集

通过图 7 可以看出，有 9 个通路是孤独症、癫痫和共患所共富集到的；只有癫痫有 3 个特有的通路富集。通过分析发现，这 9 个共有的通路分别是氧化磷酸化、热能产生、化学致癌 - 活性氧物质、帕金森病、亨廷顿病、朊病、肌萎缩性侧索硬化症、糖尿病心肌病，阿尔茨海默病，这可能说明孤独症和癫痫导致的差异表达和一些其他的精神疾病存在交互，能够从侧面说明一些精神疾病并发的机制。癫痫所特有的三个富集通路分别是逆行内源性大麻素信号传导、安非他命成瘾和心肌收缩。有研究表明，大麻二酚(Cannabidiol, CBD)，对于治疗耐药性癫痫是有效的；安非他命类药物可以让人注意力更专注，更容易注意细节，并且长时间保持很好的精神，识别到的这两个通路的富集可能和癫痫患者的服药习惯有较大关系。

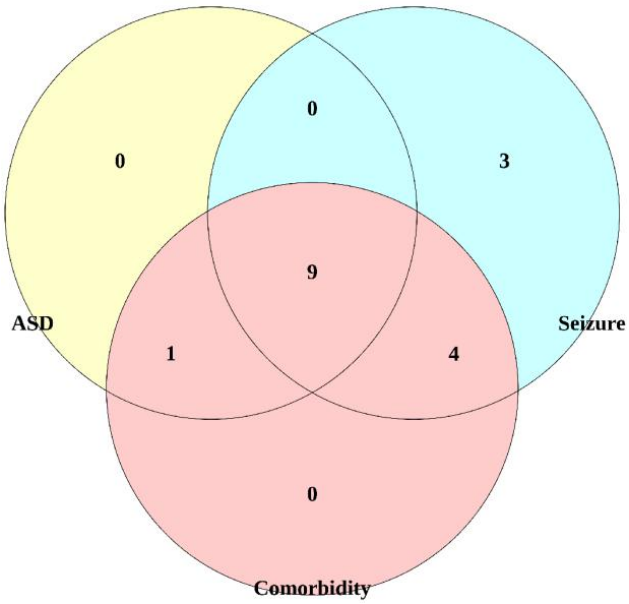

图 7 KEGG 富集韦恩图

4.4 WGCNA 分析

通过对 PFC 区域的星形胶质细胞的单细胞 RNA seq 进行 bulk 处理后，一共

有 34 个样本，2531 个基因进行 WGCNA 分析。

通过将基因的连通性分为多个区段(bins)，对每一个区段内基因连通性的均值取 log10，该区段的基因的频率取 log10，如果两者存在线性关系，那么这些基因的构成的网络符合幂律分布。如图 8 所示，通过计算不同软阈值下的线性相关系数 R^2，当软阈值为 40 的时候，R^2 接近 0.9。然而识别网络模块最高可以指定的软阈值为 30，因此本研究在软阈值为 30 的条件下构建构建网络，识别共表达基于模块。如图 9 所示，当软阈值为 30 的时候，基因的平均连通度为 99。

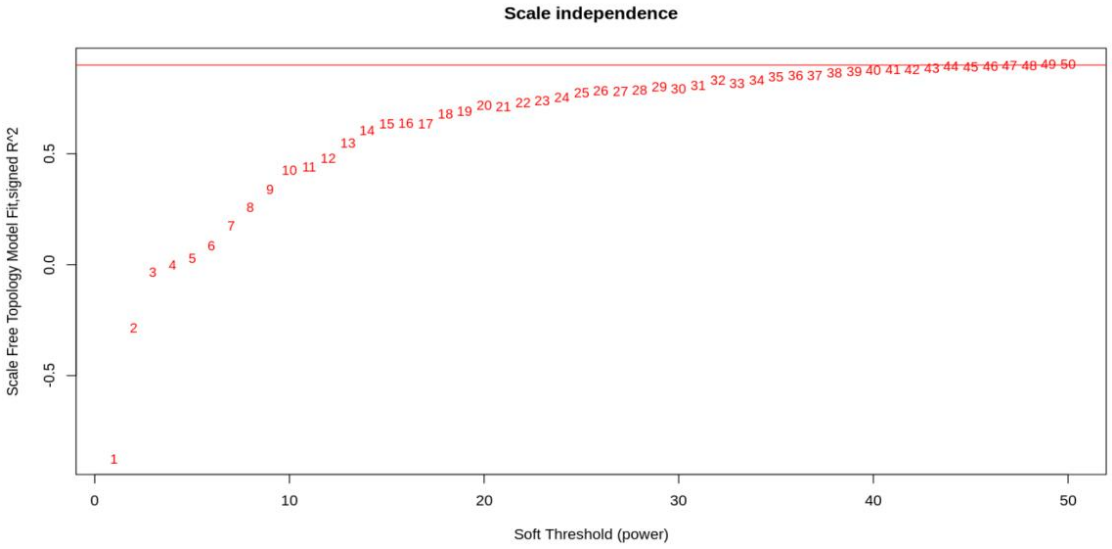

图 8　不同软阈值下幂律分布的拟合指数

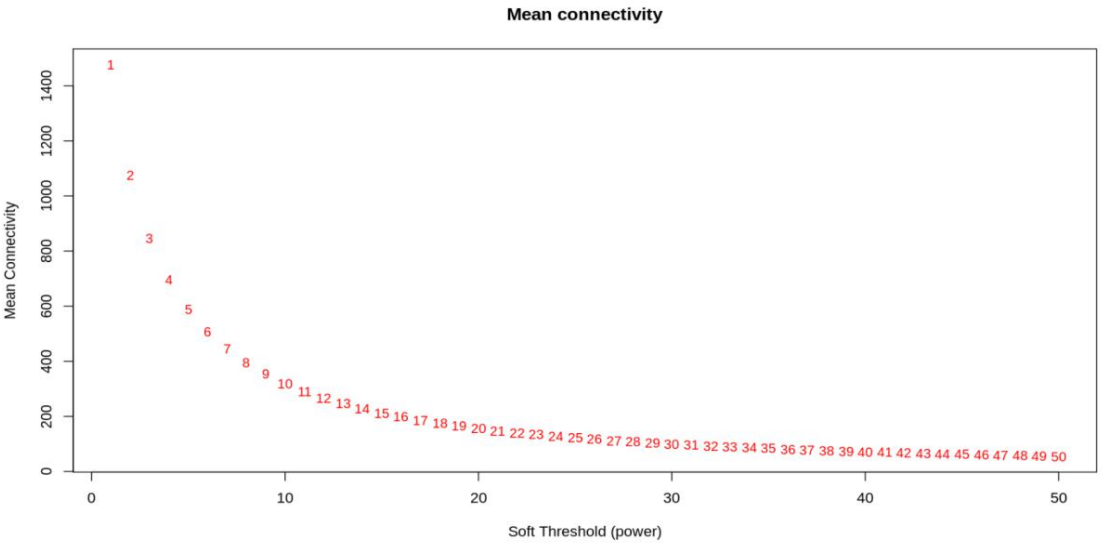

图 9　不同软阈值下的平均连通度

如图 10 所示，2531 个基因可以分为 3 个模块，蓝色模块有 952 个基因，青绿色模块有 981 个基因，有 598 个基因不分配到其他任何共表达模块(灰色模块)。图 11 展示了该网络权重前 500 的边构成的网络，节点代表基因，节点的颜色代表基因所述模块。

KME (eigengene connectivity)为基因和模块相关系数，KME 值接近 0,说明这个基因不是该模块的成员；KME 接近 1 或者-1,说明这个基因与该模块密切相关(正相关或者负相关)。本研究通过 KME 值筛选每一个共表达模块中的 hub 基因。

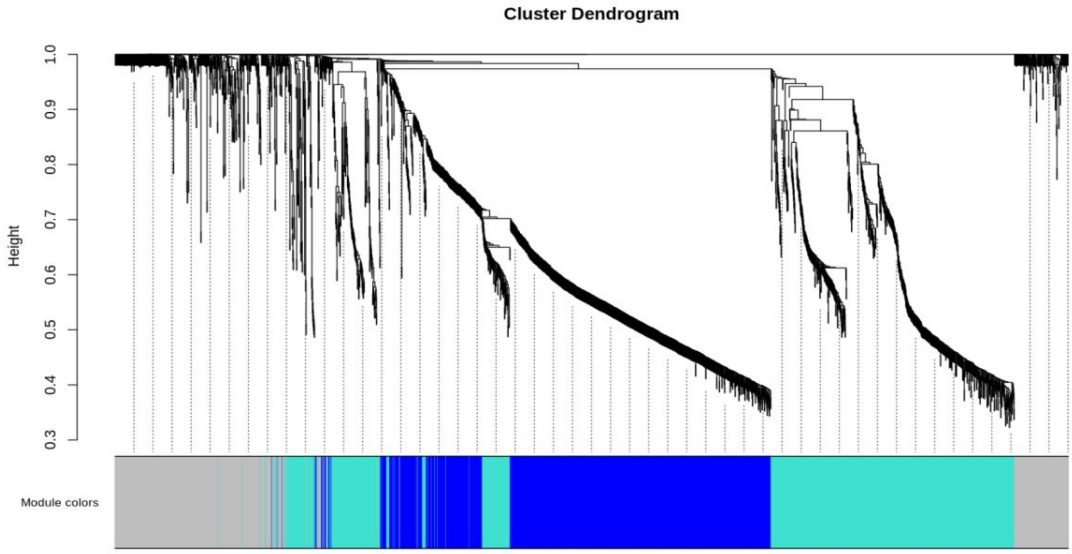

图 10　基因聚类树状图

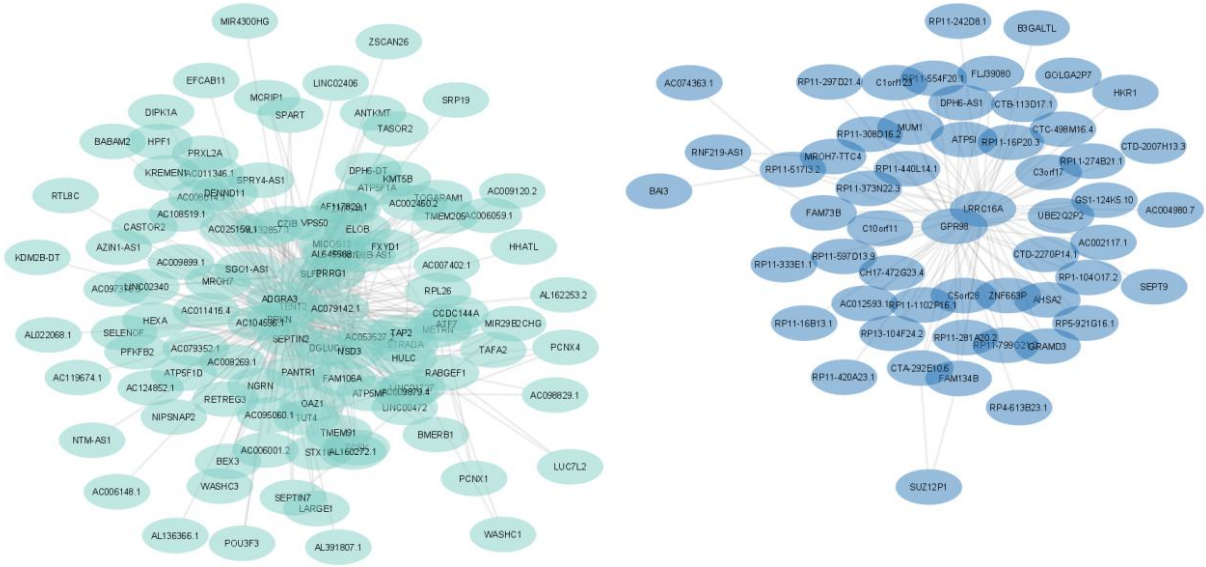

图 11　权重前 500 的边的网络图

当定义 |KME| 阈值为 0.99 时，蓝色模块有 158 个 hub 基因，青绿色模块有 24 个 hub 基因。蓝色模块中排名前五的 hub 基因分别时 AC074391.1、SUZ12P1、GUSBP1、SEPT2 和 BRE；青绿色模块中排名前五的 ALDOC、ERBIN、AFDN、WAPL、LINC01896。

5 讨论

在这项研究中，本研究收集了孤独症、癫痫和共患个体的单细胞核 RNA Seq 测序数据。经过基因定量和细胞注释后，从 118,395 个细胞中筛选出了 9,216 个位于前额叶皮层的星形胶质细胞，以进行差异表达分析。本研究的目标是找出这些不同疾病条件下的星形胶质细胞中的差异表达基因，并进一步进行 GO 富集和 KEGG 富集分析，以找出各自疾病的共有和特有的功能及通路注释。

本研究发现，患有孤独症、癫痫以及两者共患的个体的星形胶质细胞存在着能量代谢、蛋白转运、翻译以及细胞形态方面的异常。癫痫影响星形胶质细胞的代谢活动的机制和孤独症不同，癫痫可能和胶质细胞对钙离子的响应和转运存在联系。共患个体的星形胶质细胞则在对类固醇激素和糖皮质激素的响应上存在异常。此外，KEGG 富集分析结果还显示，这些差异表达基因与其他一些精神疾病存在交互作用，从侧面说明了一些精神疾病可能具有共同的并发机制。

最后，本研究对 2531 个差异表达基因进行了 bulk WGCNA 分析，成功识别出了两个共表达网络模块，并确定了这两个模块的 hub 基因。

本研究存在一些缺陷。本研究在细胞注释过程中一共注释到 15 个细胞类型，然而有的细胞类型在孤独症、癫痫以及共患中细胞数目相差较大，考虑到不同分组中的数目对统计分析结果的影响以及时间的限制，本研究只对来自前额页皮层的星形胶质细胞进行相关分析，而没有挖掘其他细胞类型的差异表达基因的信息；为节约时间和计算资源，本研究只对癫痫和部分正常样本的单细胞核测序进行基因和细胞定量，然后将产生的基因表达矩阵和 Velmeshev 提供的孤独症和共患的基因表达矩阵合并起来分析，这意味着需要批次效应进行处理，本研究对基因表达矩阵的批次效应的处理并不好，在找孤独症、癫痫和共患细胞的差异表达基因时，孤独症、癫痫和共患识别到的差异表达基因的 log2FC 值存在很明显的断层，孤独症和共患的表达矩阵由 Velmeshev 提供，识别到的显著的差异表达基因多为

负调控，而癫痫的表达矩阵由笔者定量获得，能够识别到更多的正调控的差异表达基因。

本研究下一步应该使用一个一致的单细胞处理流程，一致的参考基因组对所有的单细胞核测序样本进行比对、定量和注释，从源头上获得一个批次效应相对更小的基因表达矩阵进行下游分析。本研究还能根据这一个分析流程，进一步对其他细胞类型进行分析。对于识别到的共表达模块的 hub 基因，可以对其进行富集分析，以揭示 hub 基因的通路和功能。

6 参考文献

1. Hao Y, Stuart T, Kowalski M H, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, Satija R. Dictionary Learning for Integrative, Multimodal and Scalable Single-Cell Analysis. *Nat Biotechnol*, 2023.

2. Langfelder P, Horvath S. WGCNA: An R Package for Weighted Correlation Network Analysis. *Bmc Bioinformatics*, 2008, 9: 559.

3. Shannon P, Markiel A, Ozier O, Baliga N S, Wang J T, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*, 2003, 13 (11): 2498~2504.

4. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, Bhaduri A, Goyal N, Rowitch D H, Kriegstein A R. Single-Cell Genomics Identifies Cell Type-Specific Molecular Changes in Autism. *Science*, 2019, 364 (6441): 685~689.

5. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G. ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation (Camb)*, 2021, 2 (3): 100141.