華中科技大學

本科生毕业设计[论文]

肝脏转分化的多组学整合

院　　系　生命科学与技术学院

专业班级　登峰 2001 班

姓　　名　张展晔

学　　号　U202013309

指导教师　薛宇

2024 年 5 月 11 日

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内容编入有关数据进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密□，在　　年解密后适用本授权书

2、不保密☑　。

（请在以上相应方框内打“√”）

作者签名：　　年　　月　　日

导师签名：　　年　　月

摘要

过表达肝脏转录因子（FOXA3、HNF1A 和 HNF4A，简称为 FHH）可使人成纤维细胞转分化为可增殖的功能肝细胞（hiHep），从而弥补活体肝脏组织短缺。但该生物学过程中的分子调控机制尚未研究清楚。使用多组学的方法对其机制进行探索可以获得全面的景观。在使用 FHH 诱导人成纤维干细胞转分化为肝细胞的 0 到 5 天的过程中，细胞转录组、蛋白质组、磷酸化蛋白质组表达量发生了显著变化。富集分析结果表明，激酶和癌症相关基因在该过程中表达增加，而跨膜运输能力和细胞的信号传导能力在该过程中表达降低。此外，细胞内部骨架结构和物质合成相关基因在转分化早期上调，后逐渐下调，而抗原呈递、病毒感染和自身免疫相关基因在早期下调，随后逐渐上调。分子互作网络图表明，HLA 家族分子在多个重要通路中发挥作用，可能在肝脏转分化过程中具有重要功能。这些发现有助于深入理解肝脏转分化的调控机制，为相关疾病的治疗提供了新的研究思路。

关键词：肝脏转分化，多组学，差异分析，富集分析，分子网络绘制

Abstract

Overexpression of liver transcription factors (FOXA3, HNF1A, and HNF4A, collectively referred to as FHH) can induce human fibroblast cells to transdifferentiate into proliferative functional hepatocytes (hiHep), thereby addressing the shortage of viable liver tissue. However, the molecular regulatory mechanisms underlying this biological process remain poorly understood. Exploring these mechanisms using multi-omics approaches can provide a comprehensive landscape. During the induction of human fibroblast stem cells into liver cells using FHH from day 0 to day 5, significant changes were observed in the cellular transcriptome, proteome, and phosphoproteome. Enrichment analysis revealed an increase in the expression of kinase and cancer-related genes, while the expression of genes related to transmembrane transport and cell signaling decreased during this process. Additionally, genes associated with intracellular cytoskeletal structure and substance synthesis were upregulated in the early stages of transdifferentiation, followed by a gradual decrease, whereas genes related to antigen presentation, viral infection, and autoimmunity were initially downregulated and then gradually upregulated. Molecular interaction network analysis demonstrated that HLA family molecules play a role in multiple crucial pathways, potentially serving important functions in the process of liver transdifferentiation. These findings contribute to a deeper understanding of the regulatory mechanisms underlying liver transdifferentiation and provide new research avenues for the treatment of related diseases.

Key Words： Liver Transdifferentiation； Multi-omics； Differential Analysis； Enrichment analysis； Molecular network mapping

目　录

1 绪论

1.1 研究背景

肝脏是人体最重要的内脏器官之一，承担着多种重要功能，包括解毒、代谢碳水化合物、代谢蛋白质、产生胆汁、吸收和代谢胆红素、辅助产生凝血因子、辅助代谢脂肪、储存维生素和矿物质、过滤血液等，对于维持机体的正常生理功能至关重要。因此一旦肝脏受损，其影响将远远超出肝脏本身，波及到整个机体的健康。

肝脏疾病的发生可能源于多种因素，如慢性病毒性肝炎（乙型肝炎病毒和丙型肝炎病毒感染等）、酗酒、肥胖和代谢综合征、自身免疫性疾病、药物或毒物的滥用、遗传性疾病等。这些因素可能导致肝脏发生不同程度的炎症、纤维化和肿瘤等病变，从而严重影响肝脏的功能和结构。

肝脏疾病不仅对患者的生活质量造成重大影响，也会对医疗系统构成了巨大负担。具体来说，肝脏疾病的治疗通常需要长期的医疗管理和药物治疗，甚至可能需要进行肝移植等高风险的手术。这对医疗资源的需求非常庞大，使得医疗系统负担大幅增加。

因此，寻找一种有效的肝脏修复与再生方法显得尤为迫切。随着生物医学领域的不断发展，已经出现了一些潜在的治疗策略。例如，基因编辑技术的发展为抑制肝脏疾病的发展提供了新的途径，其方法是通过精确地修改细胞的基因组，可以调节肝脏细胞的功能和代谢通路，有望实现肝脏的修复和再生。另外，生物材料工程也在肝脏修复与再生领域展现出巨大潜力，其方法是通过设计和制备具有特定生物学特性的生物材料，可以为受损肝脏提供支架和支持，促进肝细胞的生长和再生。此外，干细胞治疗也被认为是一种具有巨大潜力的方法，干细胞具有自我更新和多向分化的能力，可以分化为肝细胞，用于修复受损的肝脏组织。

产生大量功能性人类肝细胞用于基于细胞的肝病治疗是一个重要的方法，而肝脏转分化就是产生大量功能性人类肝细胞的好方法之一。转分化是指细胞从一种特定类型向另一种类型的细胞转变的过程。这一过程在胚胎发育、组织修复和再生中起着关键作用。惠利健等人前期通过过表达肝脏转录因子（FOXA3、HNF1A 和 HNF4A）将人成纤维细胞转分化为可增殖的功能肝细胞（hiHep）[1]，

进而实现 hiHep 的大规模扩增，构建了基于 hiHep 的新型生物人工肝（hiHep-BAL）[2]。然而，对于该过程中的分子机制尚未完全明确。

在正常生理条件下，转分化可能受到生长因子、细胞外基质、细胞间相互作用等因素的调控，而这些因素可以影响细胞的分化方向和速率。然而，在病理情况下，例如肝脏损伤或肝炎等，转分化过程可能会受到炎症因子、细胞凋亡、纤维化等因素的影响，从而导致肝细胞的功能丧失或异常增生。深入研究肝脏转分化过程的机制，不仅有助于理解肝脏再生与修复的机制，还为开发治疗肝脏疾病的新策略提供了重要的理论基础。但肝脏转分化过程复杂，涉及到多种组学变化。因此，本研究选择使用多组学的方法对其机制进行研究。

多组学研究是一种综合利用多种高通量技术，如基因组学、转录组学、蛋白质组学、代谢组学等组学测序技术，对生物体进行全面分析的方法。这种方法能够提供关于生物体内部复杂生物过程的全面信息，为研究人员深入了解生物体的生理、病理过程提供了重要手段。近年来，多组学研究在生物医学领域取得了显著进展。通过综合分析基因组、转录组、蛋白质组、代谢组等多种组学数据，研究人员能够全面地探索疾病的发病机制、诊断标志物、治疗靶点等方面的信息，为疾病的早期诊断、个性化治疗提供了新的思路和方法。

在多组学研究中，不同组学的数据可以为研究者提供给不同角度的信息。比如，基因组学是多组学研究的重要组成部分之一。通过对基因组的全面测序和分析，可以揭示基因与疾病之间的关联关系，发现致病基因、易感基因等信息。此外，基因组学还可以用于研究群体遗传结构、人类进化历史等问题。与基因组学又所区别的是，转录组学是研究细胞基因表达水平和调控机制的重要手段。通过对转录组的分析，可以发现与疾病相关的差异表达基因、转录调控因子等信息，从而深入理解疾病的发病机制和调控网络。而蛋白质组学是研究细胞蛋白质组成和功能的重要方法。通过对蛋白质组的分析，可以发现蛋白质的表达水平、修饰状态等信息，揭示蛋白质在疾病发生发展中的作用和机制。此外，磷酸化蛋白质组学是研究蛋白质磷酸化修饰的一种方法。磷酸化是一种重要的蛋白质修饰方式，能够调节蛋白质的活性、稳定性和相互作用，影响细胞信号传导、代谢调控等生物学过程。通过磷酸化蛋白质组学的分析，可以发现不同疾病状态下磷酸化水平的变化，揭示磷酸化修饰在疾病发生发展中的作用和机制，为疾病的诊断和治疗

提供新的靶点和策略。

综合利用基因组学、转录组学、蛋白质组学、代谢组学等多种技术手段，可以实现对生物体的全面分析，揭示生物体内部复杂生物过程的全貌。通过多组学研究的方法，不仅有助于深入理解疾病的发病机制和调控网络，还为疾病的早期诊断、个性化治疗提供了新的途径和方法。

目前已有对肝脏转分化过程中的多组学整合分析研究，例如，Yangyang Yuan 等人使用肝脏转分化不同时期的转录组和磷酸化蛋白质组数据，综合考虑不同时期激酶对应 mRNA 数量差异，激酶对应磷酸化位点强度差异，以及激酶对应磷酸化调控网络差异程度三个因素构建了 CKI 算法，之后用此算法预测肝脏转分化过程中的重要激酶，并通过实验验证其机制[3]，但该研究重点关注了肝脏转分化过程中的磷酸化情况，对该生物学过程中其他调控因子情况未能进行全面描述。因此，本研究在 Yangyang Yuan 等人研究的基础上，对肝脏转分化的分子机制进行深入研究。

1.2 研究意义

探寻肝脏转分化过程的机制机理，有助于我们更深入地了解肝脏再生与修复的本质，具有非常重要的研究意义。首先，通过揭示转分化过程中涉及的调控因子及其相互作用，我们可以为肝脏再生的生物学机制提供更为全面和深入的解释，为相关疾病的治疗和预防提供新的思路和策略。另外，人工诱导肝脏细胞合成是目前研究的热点之一，通过深入了解肝脏转分化的调控机制，我们可以有针对性地设计和优化诱导方案，提高肝脏细胞的合成效率和稳定性，为临床治疗提供更加可靠的技术支持。此外，对于肝脏转分化过程的研究还有助于我们理解其他组织或器官再生与修复的机制。肝脏作为一个典型的再生器官，其再生与修复的机制可能具有一定的普适性，因此相关研究的成果也将为其他组织或器官的再生医学研究提供借鉴和启示。

1.3 研究目的

本研究旨在利用 Yangyang Yuan 等人的研究中使用的 bulk RNA-seq、磷酸化蛋白质组和蛋白质组数据，通过多组学整合分析的方法，系统地探究肝脏转分化

过程中的调控因子。具体包括以下几个方面的内容：利用多组学数据，对肝脏转分化过程中的细胞谱系关系进行分析，揭示不同细胞状态下的各组学变化规律；筛选和鉴定参与肝脏转分化调控的关键基因，探究其在转分化过程中的作用机制；建转分化相关的分子互作网络，预测候选调控因子的功能和相互关系。通过上述研究目的的实现，我们将更全面、深入地了解肝脏转分化的调控机制，为肝脏再生与修复的实践应用提供科学依据和技术支持。

1.4 数据来源

本研究所使用的数据来源于 Yangyang Yuan 等人的研究中使用的 bulk RNA-seq、磷酸化蛋白质组和蛋白质组数据。Bulk RNA-seq 数据下载地址为 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169702 ，磷酸化蛋白质组和蛋白质组数据下载地址为 https://www.iprox.cn/page/project.html?id=IPX0001681000 。我们通过直接下载获取了相关的原始数据，并在此基础上展开了后续的分析和研究工作。

2 方法与结果

2.1 样本信息收集整理

之前的研究证明，利用慢病毒编码从而过表达肝重编程调控因子（FOXA3、HNF1A 和 HNF4A，简称为 FHH），可使人成纤维干细胞转分化为肝细胞样细胞（hiHep 细胞）[1]。本研究收集 Yangyang Yuan 等人的研究中慢病毒编码在人成纤维干细胞中过表达 FHH 的数据进行研究。

GFP（Green Fluorescent Protein，绿色荧光蛋白）培养基是一种用于培养表达 GFP 或其他荧光蛋白的细胞的培养基。GFP 是一种来自于水母的蛋白质，在光的作用下能够发出绿色荧光。GFP 培养基通常是含有适当营养成分、缓冲剂和抗生素的培养基，在培养 GFP 表达细胞时，通常会添加适量的 GFP 抗生素以确保细胞的稳定表达。

本研究使用的细胞类型包括使用 GFP 培养 2.25 天的人成纤维干细胞，FOXA3、HNF1A 和 HNF4A 过表达培养 2.25 天的人成纤维干细胞，FOXA3、HNF1A 和 HNF4A 过表达培养 5 天的人成纤维干细胞。其中使用 GFP 培养 2.25 天的人成纤维干细胞数据为对照组，过表达 FHH 培养 2.25 天和过表达 FHH 培养 5 天的人成纤维干细胞数据作为实验组。对每种细胞类型，本研究均收集了 bulk 转录组数据、蛋白质组、磷酸化蛋白质组三种组学数据（如表 1）。具体来说，转录组数据有三种培养条件各一个样本共 3 个样本，而蛋白质组学和磷酸化蛋白质组学各有 6 个和 11 个样本，每个蛋白质组学和磷酸化蛋白质组学样本均使用 TMT 标记同时测量三种培养条件下的组学数据。

2.2 组学数据基本处理

2.2.1 转录组数据处理

对于收集到的三个 bulk RNA-seq 样本数据，分别使用 linux 命令行工具进行处理，并使用 Snakemake 工具包对处理流程进行拼接，从而实现数据处理的自动化。

Snakemake 是一个基于 Python 的工作流管理系统，用于构建和运行数据分析流程。通过定义规则和依赖关系，Snakemake 可以自动化数据处理过程，并在必要时重新运行特定任务，以确保分析的可重复性和可靠性[4]。

对于每个 bulk RNA-seq 样本数据，首先，我们使用 fastq-dump 工具根据 Yangyang Yuan 等人研究中给出的数据集样本编号列表（https://www.ncbi.nlm.nih.gov/Traces/solr-proxy-be/solr-proxy-be.cgi），从 NCBI 数据库的 Sequence Read Archive (SRA) 下载了与我们研究相关的原始测序数据。fastq-dump 是 SRA Toolkit 中的一个工具，它允许我们从 SRA 数据库下载 SRA 格式的原始测序数据，并将其转换成常见的 FASTQ 格式，以便后续的分析。在 Linux 命令行下，我们运行 fastq-dump 命令并指定所需的 SRA 数据集，然后该工具会自动下载并转换数据。

接着，我们使用 fastp 工具[5]对下载的 FASTQ 文件进行质量控制和过滤。fastp 是一个快速且高效的质量控制工具，它能够检测测序数据的质量分数，并根据用户指定的参数进行质量修剪、去除接头序列和过滤低质量读段。我们在命令行中输入 fastp 命令，指定输入和输出文件，并设置相应的参数，以对数据进行质量控制和过滤。

随后，我们使用 STAR 工具[6, 7]将经过质量控制的读段比对到参考基因组上，参考基因组使用 GRCh38 基因组（https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.39/）。STAR 是一种广泛使用的 RNA-seq 比对工具，它采用了基于后缀数组的方法，能够在较短的时间内对大规模的 RNA-seq 数据进行快速而准确的比对。我们在命令行中输入 STAR 命令，并指定参考基因组的索引文件和输入的 FASTQ 文件，然后该工具会自动进行比对，并生成 SAM/BAM 格式的比对结果（如表 1-1）。可以看到，所有样本读段比对率都高于 90%，因此说明比对结果是可靠的。

表 1-1 STAR 工具比对结果

样本编号	总读段数	平均读段长度	唯一映射的读段数	唯一映射读段占比%
SRR14078360	80435542	274	72546872	90.19%
SRR14078361	39566505	275	35749484	90.35%
SRR14078362	42874520	278	38827339	90.56%

最后，我们使用 featureCounts 工具[8]对比对结果进行基因表达量的计算。featureCounts 是一个用于计算基因表达量的工具，它能够从比对到基因组的 BAM 文件中统计读段的数量，并将其映射到基因组上的基因或转录本上。我们在命令行中输入 featureCounts 命令，并指定基因组注释文件和比对结果的 BAM 文件，然后该工具会自动进行基因表达量的计算，并生成相应的结果文件。从表 1-2 中可以发现，使用 featureCounts 得到的计数率均大于 74%，这个结果也说明样本是可靠的。

表 1-2 featureCounts 工具基因表达量计算结果

样本编号	分配成功	映射失败	多次映射	无特征	模糊映射
SRR14078360	127032556	9543242	15075290	13724872	4228051
SRR14078361	62290252	4648920	7261208	7144946	2011543
SRR14078362	66319101	4586587	8538141	8788787	2491144

2.2.2 蛋白质组学和磷酸化蛋白质组学数据处理

在蛋白质组学和磷酸化蛋白质组学数据处理过程中，我们首先关注了不同类型的参考蛋白质数据库。在 Uniprot 数据库中包括了人工审查的和没有人工审查标记的蛋白质数据，这两种参考蛋白质组数据之间存在着显著的差异。人工审查的蛋白质数据经过了严格的审核和验证，具有较高的可信度和准确性，包含丰富的生物学信息。相比之下，没有人工审查标记的蛋白质数据可能是自动预测或计算得到的，可信度和准确性较低，且缺乏详细的生物学注释。我们选择使用了来自 Uniprot 数据库的人工审查的人类参考蛋白质数据作为备用数据。

在蛋白质组学和磷酸化蛋白质组学数据处理过程中，我们采用了 MaxQuant 软件[9]进行数据分析。对于蛋白质组学数据，我们使用了人工审查的参考蛋白质数据，并在 MaxQuant 中设置了相关的参数，包括 reporter 设置为 MS2、6plex TMT、消化酶摄制设置为 Trypsin/P、Variable modifications 设置为 Oxidation(M) 和 Acetyl (Protein N-term)，Fixed modifications 设置为 Carbamidomethyl(C)。对于磷酸化蛋白质组学数据，在 Variable modifications 中额外添加了 Phospho(STV)，

其他参数与蛋白质组处理方式不变。

在得到 MaxQuant 分析得到的 proteinGroups.txt 文件后，我们进行了进一步的处理。具体地，对于文件中 Protein IDs 列中同一行有不同蛋白质 id 的行，我们视作该行 Protein IDs 列中所有蛋白质表达量相等，并将其转换为新的 proteinGroups 数据。这一步处理的目的是为了更准确地反映不同蛋白质在样本中的表达量，从而为后续的数据分析和结果解释提供更可靠的基础。

对于得到的蛋白质组和磷酸化蛋白质组学数据，已知再相同条件下相同蛋白质和相同磷酸化蛋白质表达量差异应该不大。因此我们绘制了三种条件下蛋白质组表达量-磷酸化蛋白质组表达量图（如图 1 所示）。由图 1 我们可以发现，磷酸化蛋白质组表达量并不完全成正比。这可能表明磷酸化水平受到多种因素的调控，包括细胞信号通路的激活和抑制等。这同时也提示了在相同条件下蛋白质的磷酸化状态可能是一个复杂的动态过程，需要进一步的研究来揭示其调控机制。

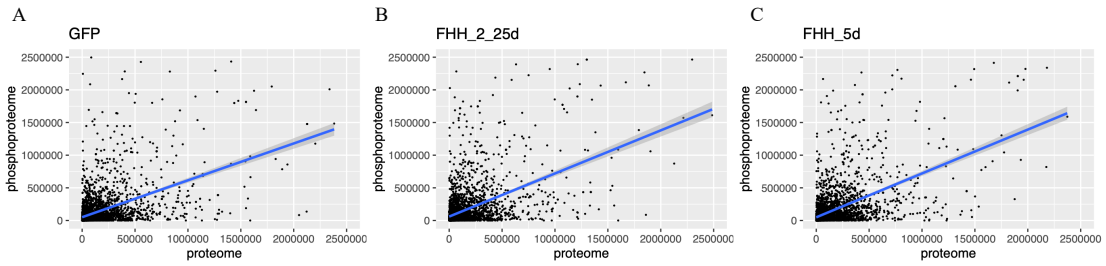

图 1 蛋白质组-磷酸化蛋白质组表达量图，x 轴为蛋白质在蛋白质组中表达量，y 轴为蛋白质在磷酸化蛋白质组中表达量，每一个点为一个蛋白质在蛋白质组和磷酸化蛋白质组中的表达量。A：GFP 培养 2. 25 天样本的蛋白质组-磷酸化蛋白质组表达量图；B：FHH 培养 2. 25 天样本的蛋白质组-磷酸化蛋白质组表达量图；C：FHH 培养 5 天样本的蛋白质组-磷酸化蛋白质组表达量图。

2. 3 差异分析

2. 3. 1 差异分析方法选择

差异分析是基因组学研究中的重要步骤，用于确定在不同条件或处理下基因

表达水平的显著差异。常用的差异分析方法包括 DESeq2、edgeR、limma-voom 和 t 检验等，它们在数据类型、统计原理和适用场景等方面有所不同。

DESeq2 [10]和 edgeR [11]是两种基于负二项分布的差异分析方法，适用于 RNA-seq 数据。它们通过估计基因表达量的离散度和考虑样本间的差异来识别差异表达基因。DESeq2 在设计中更加灵活，适用于多组实验设计和小样本量，而 edgeR 则对样本数量的要求相对较低，适用于单样本或少样本的情况。

limma-voom[12]是一种基于线性模型的差异分析方法，适用于微阵列和 RNA-seq 数据。它使用 voom 转换将 RNA-seq 数据转换为方差稳定的形式，然后利用线性模型和经验贝叶斯方法对差异表达进行统计检验。limma-voom 在样本量较大时表现出色，尤其适用于多组实验设计和复杂数据结构。

另外，t 检验是一种简单直观的差异分析方法，通常用于比较两组间的均值差异。虽然 t 检验不考虑基因表达量的离散度和样本间的方差，但在样本量较小且差异较大的情况下仍然可以提供有用的信息。

在本研究中，转录组数据在 1 个对照组和 2 个实验组共 3 种条件下分别只有一个测序样本，没有相同条件不同样本进行横向比较。而蛋白质组和磷酸化蛋白质组数据虽然在 3 种条件下分别各自有 6 个和 11 个测序样本，但这两个组学数据在实验测量的时候首先将不同测序样本混合在一起，之后再进行测量，因此最终得到的数据在 3 种条件下均各自只有一个总的混合样本，因此也没有相同条件不同样本进行横向比较。

在这种条件下，本研究选择使用 edgeR 包进行差异分析，其主要原因在于其适应性和灵活性。尽管数据中每个条件仅有一个样本，或者只有一个总的混合样本，但是 edgeR 在处理少样本量或单样本情况下表现出色。相比其他方法，edgeR 具有更低的样本数量要求，并且能够处理复杂的数据结构和离散的表达模式。另外，edgeR 提供了基于负二项分布的统计模型，能够考虑到样本间的差异和离散度，产生稳健的统计结果。虽然在这种情况下无法进行横向比较，但是通过 edgeR 可以对每个条件下的表达数据进行内部比较和差异分析，从而识别出显著差异表达的基因。其灵活的参数设置和自定义分析流程也能够满足特定数据条件下的分析需求[13]。

综上所述，尽管在本研究中存在样本数量少且无法进行横向比较的情况，但

选择使用 edgeR 进行差异分析能够充分利用已有数据，获得可靠的差异表达基因，为进一步的生物学解释和实验设计提供重要参考。

2.3.2 差异分析基本结果统计

将 GFP 培养 2.25 天、FHH 培养 2.25 天、FHH 培养 5 天三种条件下的转录组、蛋白质组、磷酸化蛋白质组学样本数据进行两两分组，记为 F225_G225, F5_G225, F5_F225 三组。对三组数据的三种组学数据分别使用 edgeR 进行差异分析。在差异分析结果中，p 值表示差异的显著性，即在两组样本之间观察到的基因表达变化是否统计上显著；而 logFC 值表示基因在不同条件下的倍变差异，即其表达水平的相对变化程度。通常使用 logFC 值和 p 值联用的方法来判定基因在两组之间是上调还是下调。

在本研究中，将 logFC 值大于 1 且 p 值小于 0.05 的基因判定为上调基因，将 logFC 值小于−1 且 p 值小于 0.05 的基因判定为下调基因。三组比对数据在各组学中获得的上调下调基因数量总结（如表 2 所示）。同时，我们使用 EnhancedVolcano 工具包绘制了各组学各对比组差异分析火山图（如图 2 所示）。

表 2 差异基因数量

	基因数量
转录组上调基因数量（个）	2306
转录组下调基因数量（个）	2695
蛋白质组上调基因数量（个）	291
蛋白质组下调基因数量（个）	207
磷酸化蛋白质组上调基因数量（个）	39
磷酸化蛋白质组下调基因数量（个）	82

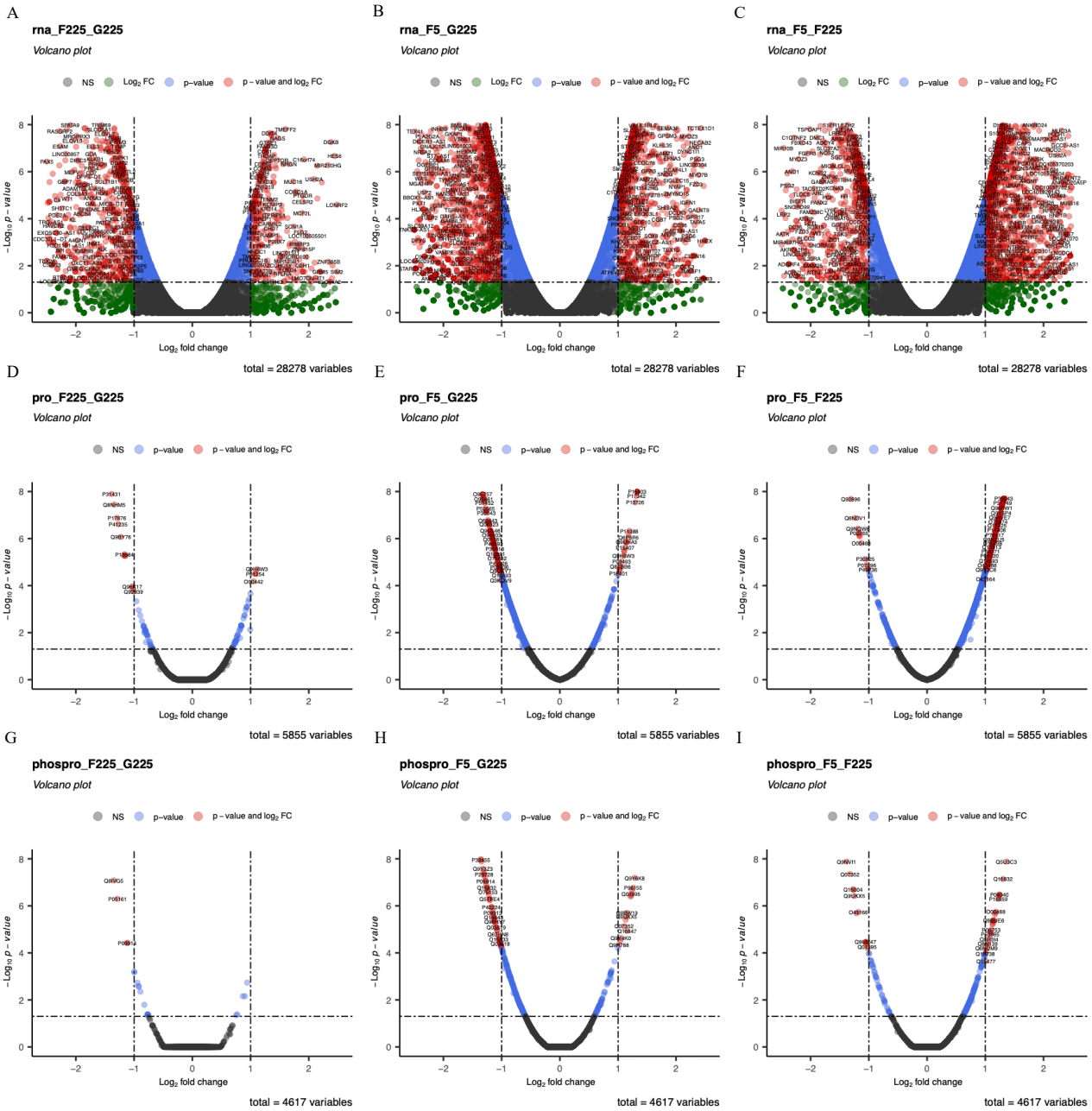

图 2 每幅图对应一个比对条件的一个组学的差异基因情况，图中以 p 值等于 0.05（也即$-\log_{10}(p)$值约等于 1.30103），以及 $\log_2$ Fold Change 值等于 1 或-1 来划分上调或下调的基因。A、B、C：分别对应 F225_G225, F5_G225, F5_F225 三种比对条件下的转录组差异基因分布情况；D、E、F：分别对应 F225_G225, F5_G225, F5_F225 三种比对条件下的蛋白质组差异基因分布情况；G、H、I：分别对应 F225_G225, F5_G225, F5_F225 三种比对条件下的磷酸化蛋白质组差异基因分布情况。

2.3.3 差异分析结果整合

之后，我们对三组比对条件下三个组学数据所产生的差异基因进行整合，试图发现一些内在规律。为了方便后续的整合分析，我们使用了 clusterprofiler 工具包[14, 15]，将蛋白质组学和磷酸化蛋白质组学的差异分析结果中得到的 UNIPROT protein ID 转换为 SYMBOL 基因名。

首先，我们查看了在相同组学中不同比较条件的基因交集情况（如图 3 所示）。从中可以发现，在转录组学数据中，15 个基因是共同上调的，73 个基因是共同下调的；在蛋白质组学数据中，3 个基因是共同上调的，10 个基因是共同下调的；在磷酸化蛋白质组学数据中，3 个基因是共同上调的，9 个基因是共同下调的。

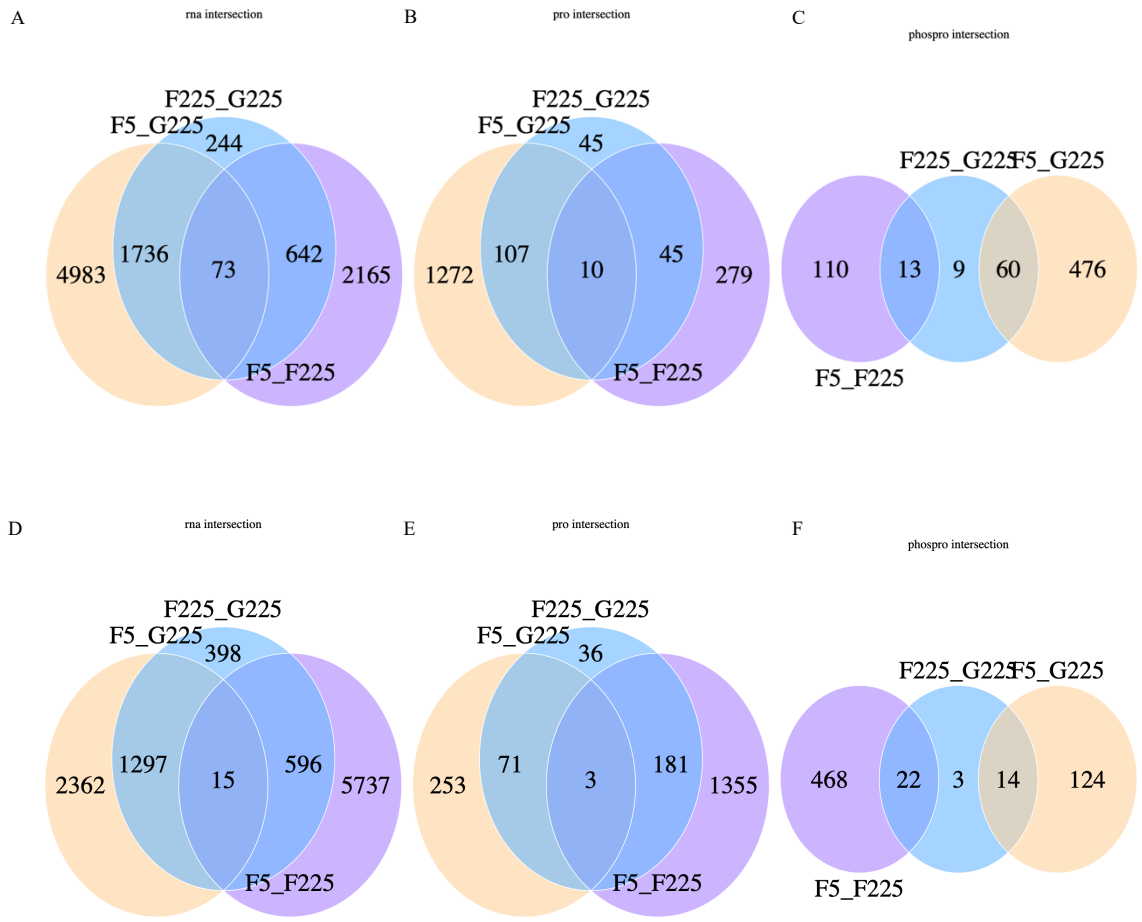

图 3 每幅图对应一个组学不同比对条件中的差异基因交集情况。A、B、C：分别对应转录组、蛋白质组、磷酸化蛋白质组的上调基因在 F225_G225, F5_G225, F5_F225 三种比对条件中的交集情况；D、E、F：分别对应转录组、蛋白质组、磷酸化蛋白质组的下调基因在 F225_G225, F5_G225, F5_F225 三种比对条件中的交集情况。

之后，我们查看了在相同对比条件中不同组学的基因交集情况（如图 4 所示）。从中可以发现，在 FHH2. 25d-GFP 比对组中，1 个基因是共同上调的，6 个基因是共同下调的；在 FHH5d-GFP 比对组中，18 个基因是共同上调的，94 个基因是共同下调的；在 FHH5d-FHH2. 25d 比对组中 74 个基因是共同上调的，10 个基因是共同下调的。

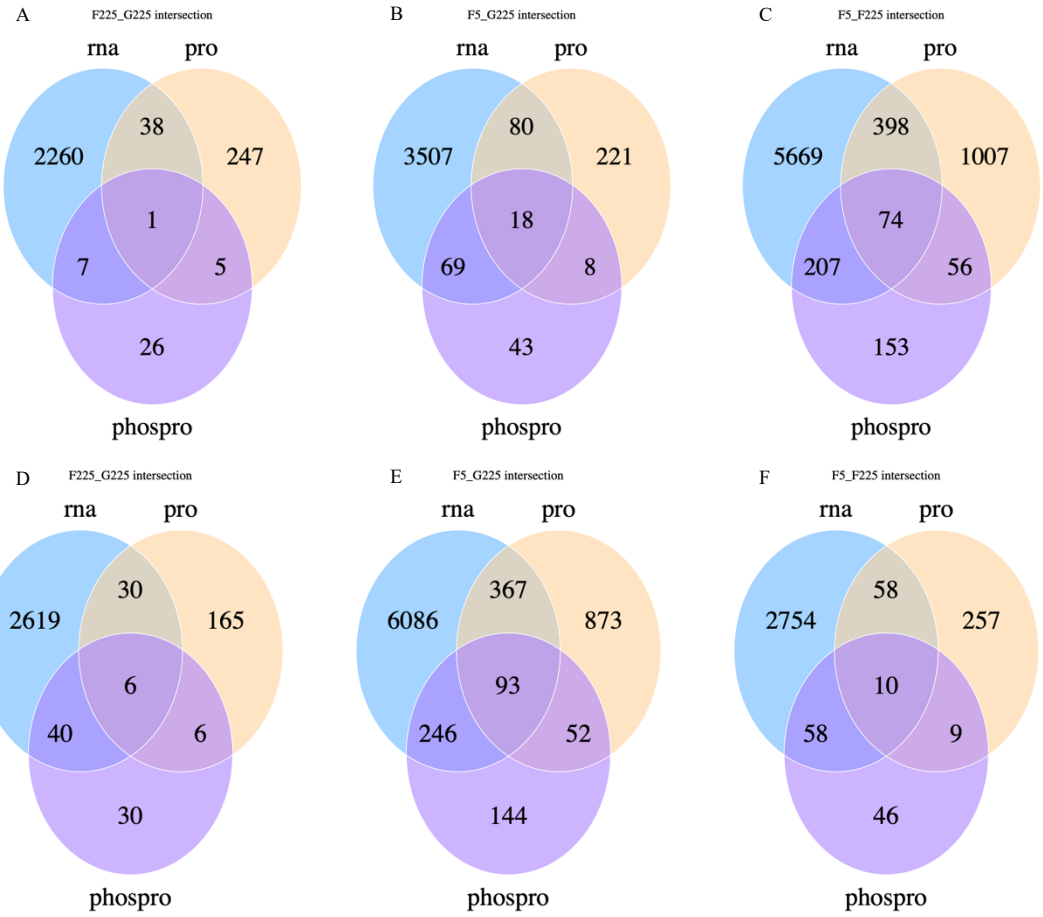

图 4 每幅图对应一个比对条件在不同组学中的差异基因交集情况。A、B、C：分别对应 F225_G225, F5_G225, F5_F225 三种比对条件的上调基因在转录组、蛋白质组、磷酸化蛋白质组三个组学中的交集情况。D、E、F：分别对应 F225_G225, F5_G225, F5_F225 三种比对条件的下调基因在转录组、蛋白质组、磷酸化蛋白质组三个组学中的交集情况。

在 FHH 诱导成纤维干细胞转分化为肝细胞样细胞 （hiHep 细胞）的过程中，我们假设存在一些基因只在转分化前期过程（0 天到 2. 25 天）起重要作用，而在后期过程（2. 25 天到 5 天）中作用消失或减弱。这部分基因我们假设他们表达量变化程度在前期和后期是相反的。因此，我们收集了在 0 天到 2. 25 天上调，

同时在 2. 25 天到 5 天下调的基因；以及在 0 天到 2. 25 天下调，同时在 2. 25 天到 5 天上调的基因，并绘制韦恩图（如图 5 所示）。

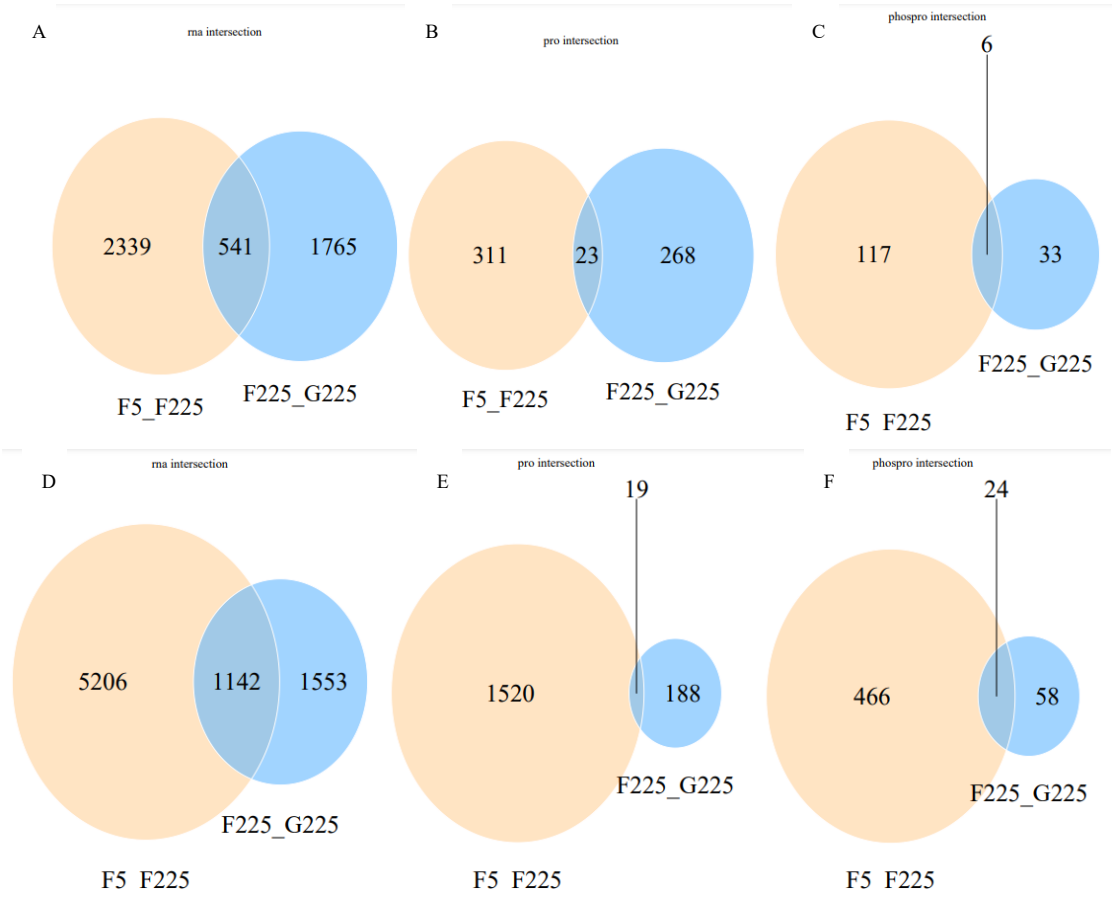

图 5 每幅图表示一个组学在不同比对条件下先上调后下调，或先下调后上调的基因交集情况。A、B、C：分别对应转录组、蛋白质组、磷酸化蛋白质组在 F225_G225 条件下的上调基因 ，以及在 F5_F225 条件下的下调基因的交集情况。D、E、F：分别对应转录组、蛋白质组、磷酸化蛋白质组在 F225_G225 条件下的下调基因 ，以及在 F5_F225 条件下的上调基因的交集情况

总的来说，我们收集了在相同组学中不同比较条件共同上调或共同下调的基因集，在相同对比条件中不同组学共同上调或共同下调的基因集，以及在转分化过程中先上调再下调，先下调再上调的基因集。下面，我们使用富集分析的方法查看这部分基因的功能。

2.4 富集分析

2.4.1 富集分析方法介绍

富集分析是一种常用的生物信息学方法，用于解释基因或蛋白质集合的功能特征和生物学意义。它通过比较实验观察到的基因或蛋白质集合与已知的功能数据库或通路信息，来确定哪些功能或通路在观察基因或蛋白质集合中富集，从而使研究人员可以深入理解这些基因或蛋白质在特定生物学过程中的作用和相互关系，并揭示细胞信号传导、代谢通路、疾病发生机制等重要生物学问题。

富集分析的计算方法主要包括超几何分布检验、Fisher’s 精确检验、基于积分的方法以及基于排序的方法（如基因集合富集分析，GSEA）。这些方法在统计学原理和计算方法上有所不同，但都旨在确定观察到的基因或蛋白质集合中哪些功能或通路富集程度显著。

在具体应用中，富集分析的对象包括各种功能数据库，如 Gene Ontology (GO)[16]、Kyoto Encyclopedia of Genes and Genomes (KEGG)[17]、Reactome[18]等。这些数据库提供了大量的功能注释、通路信息和生物学过程描述，为富集分析提供了重要的参考资源。

其中，Gene Ontology（GO）数据库是一个用于描述基因和蛋白质功能的重要资源。它将生物学过程、分子功能和细胞组分分为三个主要的功能分类，并提供了一个层次化的结构来组织这些功能术语。GO 数据库中的每个功能术语都与一组基因或蛋白质相关联，使得研究人员能够了解特定基因或蛋白质在细胞中的功能角色。通过富集到 GO 数据库中，可以对基因集合进行功能注释和分类，进一步了解实验数据中基因的功能特征，以及这些基因在生物过程中的参与情况，从而揭示其生物学意义。

而 Kyoto Encyclopedia of Genes and Genomes（KEGG）数据库是一个用于研究生物系统功能和代谢通路的重要资源。它包含了大量的生物学通路信息，包括代谢通路、细胞信号传导通路、遗传信息处理通路等。KEGG 数据库通过图形化地展示这些通路，并提供了详细的生物学注释和相关的基因信息。通过富集到 KEGG 数据库中，可以深入了解基因集合在细胞内各种通路中的参与情况，发现哪些通路在基因集合中显著富集，从而进一步探索基因在生物系统中的功能和调

控关系。

同时考虑 Gene Ontology（GO）和 Kyoto Encyclopedia of Genes and Genomes（KEGG）数据库能够提供对基因集合功能和调控机制的全面了解。GO 数据库关注基因的功能特征，如生物过程和分子功能，而 KEGG 数据库则提供了基因在生物通路中的参与情况。两者结合使用可以从不同层面对基因集合进行综合分析，验证分析结果，增强结果的可信度和解释性。这种综合分析能够帮助研究人员更好地理解基因在生物系统中的功能和作用机制。

在本研究中，对获取的差异表达基因，我们使用了超几何分布检验和 GSEA[19] 的方法，同时在 GO 和 KEGG 数据库中进行富集分析。这样做我们不但可以探究两种不同富集分析算法的优劣性和差异性，同时由于 GO 和 KEGG 数据库的互补性，我们可以从不同角度揭示基因或蛋白质集合的生物学意义，从而帮助我们更加深入的理解肝脏转分化过程中的分子机制。

2.4.2 ORA 富集分析

2.4.2.1 ORA 富集分析方法介绍

ORA（Over−Representation Analysis，代表比例分析）是一种常用的富集分析方法，用于确定在给定的基因列表中是否存在与特定功能、通路或分类的显着关联。该方法主要用于解释高通量基因表达或蛋白质组学数据中的生物学意义，帮助研究人员理解不同条件下基因表达或蛋白质表达的差异，并为实验结果提供生物学解释。

ORA 方法的基本思想是将感兴趣的基因集与已知的功能、通路或分类进行比较，然后统计在感兴趣的基因集中出现某一功能、通路或分类的数量，再将这个数量与整个基因组的期望数量进行比较，从而判断这个功能、通路或分类是否在感兴趣的基因集中显著富集。

在进行 ORA 分析时，通常需要先将感兴趣的基因集与已知的功能、通路或分类进行映射或注释，然后使用统计方法对富集程度进行评估。常用的统计方法包括超几何分布检验、Fisher 精确检验等。通过计算统计学显著性指标（如 p 值、FDR 等），可以确定哪些功能、通路或分类在感兴趣的基因集中富集得到显著性。

ORA 方法的优点之一是简单易行，不需要复杂的数学模型和大量的计算。它

可以快速识别出在感兴趣的基因集中显著富集的功能、通路或分类，为后续的生物学解释和实验设计提供重要线索。另外，由于 ORA 方法只考虑了单个功能、通路或分类的富集情况，因此可以避免多重比较问题，结果更易解释和理解。

然而，ORA 方法也存在一些局限性。首先，它假设基因之间是独立的，但在生物学上，基因往往是相互关联的，因此可能会忽略基因间的相互作用信息。其次，ORA 方法只能发现显著富集的功能、通路或分类，对于那些在样本中低表达或未富集的功能、通路或分类无法提供有效的信息。因此，在使用 ORA 方法时，需要结合其他富集分析方法进行综合分析，以获取更全面的生物学信息。

2. 4. 2. 1 ORA 富集分析结果分析

对于 ORA 富集分析，我们使用了 clusterprofiler 的 R 包来进行富集分析。clusterProfiler 是一个功能强大的 R 包，用于生物信息学和生物统计学中的基因和蛋白质注释以及富集分析。它提供了丰富的工具和函数，用于基因和蛋白质的功能注释、通路富集和网络分析。该包支持多种注释数据库，包括 GO、KEGG、Reactome 等，以及多种富集分析方法，如超几何分布检验、Fisher 精确检验等。clusterProfiler 还提供了多种可视化工具，包括饼图、柱状图、热图等，帮助用户直观地展示富集分析结果和功能注释信息。此外，该包还支持基因网络分析，帮助用户理解基因和蛋白质之间的相互作用关系。我们使用 clusterprofiler 包中超几何分布检验的方法将之前得到的差异表达基因集分别富集到了 GO 和 KEGG 数据库中。

首先，我们查看了在不同比对条件下上调基因集或下调基因集在两个通路数据库中的富集情况。具体来讲，对于每个比对条件，我们将所有三种组学的上调或下调基因的并集富集到了 GO 和 KEGG 数据库中（如图 6 所示）。

通过查看在 FHH 培养 0 天到 5 天的过程中上调的通路，我们可以发现，DNA 解旋酶活性、DNA 结合、转录共调节因子结合、ATP 活性等转录相关通路有所上调，说明在肝脏转分化过程中会转录产生新的蛋白质或其他细胞内物质，从而改变细胞内容物组成结构；腺苷酸环化酶活性、JUN 激酶（Jun kinase，JNK）活性、蛋白质激酶 A 结合、有丝分裂原激活蛋白激酶结合等通路表达增加，说明肝脏转分化过程中激酶或许起到重要调节作用，这与 Yangyang Yuan 等人的研究结果符合。值得注意的是，甲状腺癌、非小细胞肺癌、胃癌、细胞紧密连接、

细胞衰老、细胞骨架运动活性等癌症相关通路表达增加(图中未展示，其通路富集分析的 p. adjust 值不在前 20 个内，但均小于 0. 05)，说明 FHH 有可能对成纤维干细胞有一定的致癌性。

通过查看在 FHH 培养 0 天到 5 天的过程中下调的通路，我们可以发现，糖跨膜转运蛋白活性、长链脂肪酸跨膜转运蛋白活性、胆固醇转运活性等物质跨膜运输通路表达下调，说明肝脏转分化过程中跨膜运输能力可能下降；CXCR 趋化因子受体结合、趋化因子活性、细胞因子受体结合、细胞因子活性、细胞因子–细胞因子受体相互作用通路表达下调，这可能表明细胞的趋化和信号传导能力受到了抑制，可能导致细胞对外部刺激的应答能力下降，从而影响了细胞间的相互作用和信号传递过程。

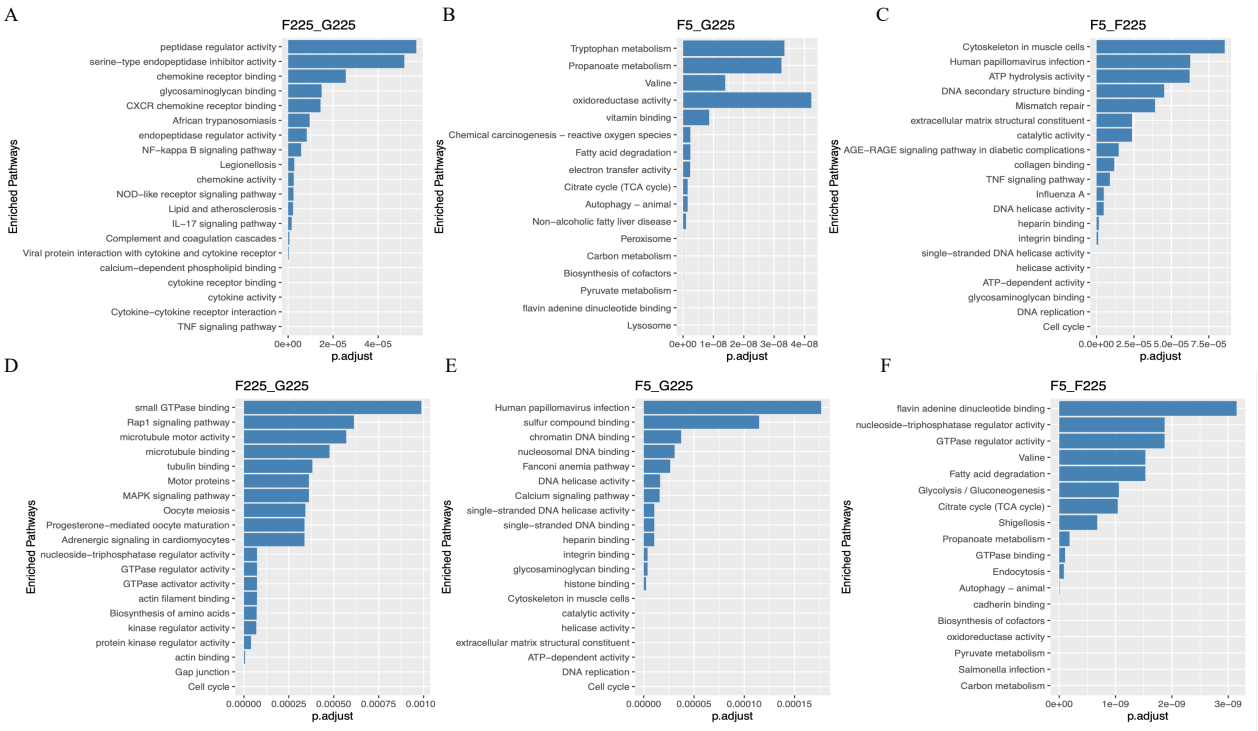

图 6 将不同比对条件下的上调或下调基因富集到 GO 和 KEGG 数据库中，整合三个组学两个数据库的富集结果后，取每个比对条件下的根据富集分析 p. adjust 值由小到大排列的前 20 个富集通路进行作图。A、B、C：分别对应 F225_G225, F5_G225, F5_F225 三种比对条件下所有组学上调基因的前 20 个富集通路。D、E、F：分别对应 F225_G225, F5_G225, F5_F225 三种比对条件下所有组学下调基因的前 20 个富集通路。

之后，我们查看了先上调再下调和先下调再上调基因在通路中的富集情况（如图 7 所示）。

通过查看在 FHH 培养 0 天到 5 天的过程中先上调后下调的基因，我们可以发现，基因在微管结合、微管蛋白结合、细胞骨架的结构成分等通路富集，这可能反映了在肝脏转分化过程中细胞内部骨架结构的动态变化；基因在 DNA 聚合酶结合、DNA 导向的 5’-3’ RNA 聚合酶活性、5’-3’ RNA 聚合酶活性、RNA 聚合酶活性等通路富集, 说明在转分化前期过程中，细胞可能增强了对 DNA 和 RNA 的合成能力，从而为细胞功能和结构的改变提供了必要的生物学基础，而在后期过程中细胞功能结构逐渐稳定，则减少了对 DNA 和 RNA 的合成能力。

通过查看在 FHH 培养 0 天到 5 天的过程中先下调后上调的基因，我们可以发现，基因在肽抗原结合、抗原加工提呈等抗原呈递相关通路富集，这暗示着在转分化过程中免疫系统可能经历了变化，可能涉及到对抗原的识别和呈递的调节；同时在人类免疫缺陷病毒 1 型感染、人类巨细胞病毒感染、病毒性心肌炎等病毒感染相关通路中富集，这可能意味着在肝脏转分化的过程中，细胞对病毒感染的免疫应答发生了调整；也在移植物抗宿主病、1 型糖尿病等自身免疫相关通路中富集，这可能暗示着免疫调节在肝脏转分化中也扮演着重要角色。这些富集结果为我们理解肝脏转分化过程中免疫相关机制的调节提供了线索。

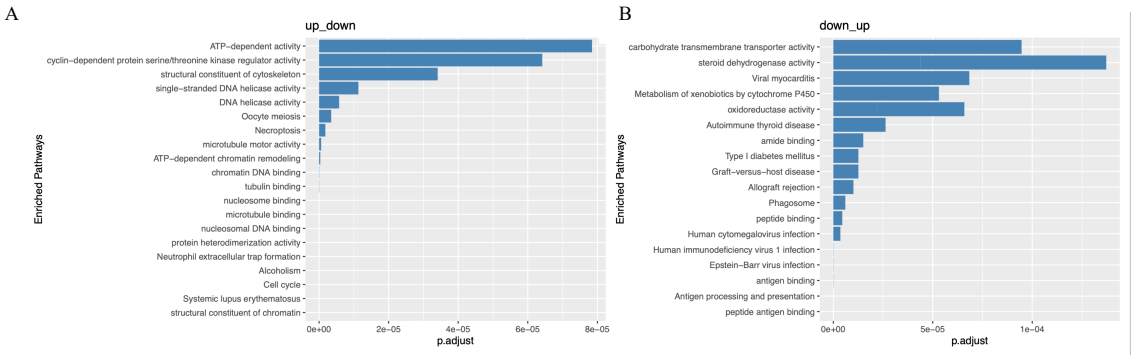

图 7 将培养 0 到 5 天过程中先上调再下调的基因，以及先下调再上调的基因富集到 GO 和 KEGG 数据库中，整合两个数据库的富集结果后，取每个比对条件下根据富集分析 p. adjust 值由小到大排列的前 20 个富集通路进行作图。A：在 F225 天组比 G225 天组上调，同时 F5 组比 F225 组下调的基因的富集结果。B：在 F225 天组比 G225 天组下调，同时 F5 组比 F225 组上调的基因的富集结果。

2. 4. 3 GSEA 富集分析

2. 4. 3. 1 GSEA 富集分析方法介绍

基因集富集分析（Gene Set Enrichment Analysis，GSEA）[19]是一种用于解释基因表达数据中生物学意义的计算方法。该方法的主要目的是确定已知生物学通路、功能或其他基因集在给定基因表达数据中的富集程度，以帮助研究人员理解基因表达的调控机制和生物学过程。

GSEA 方法的基本思想是将基因按其表达水平进行排序，然后检查已知的基因集中的基因在排序列表中是否倾向于聚集在一起。如果某个基因集中的基因在排序列表的某一端集中分布，那么就说明这个基因集在表达数据中是富集的。GSEA 不像传统的富集分析方法那样依赖于一个预先定义的基因列表，而是基于整个基因集的表达模式来进行分析。

GSEA 方法的具体步骤如下：数据准备：首先，需要准备基因表达数据和与之相关的基因注释信息，通常是一个基因表达矩阵和一个基因集数据库；基因集的选择：选择一个或多个感兴趣的基因集作为分析对象，这些基因集通常来自于已知的生物学通路、功能、疾病相关基因等；基因排序：将样本中的基因按照其在基因表达数据中的表达水平进行排序，通常是根据差异表达分析的结果或其他统计指标来排序；富集分析：对于每个基因集，计算该基因集中的基因在排序列表中的富集程度。常用的统计指标包括富集得分（enrichment score）、标准化富集得分（normalized enrichment score，NES）、p 值等；统计显著性检验：使用适当的统计方法对富集得分进行显著性检验，通常是通过对基因集的排列来估计 p 值或者计算 FDR 等多重检验校正；结果解释：根据富集分析的结果，解释哪些生物学通路、功能或基因集在表达数据中是显著富集的，以及它们可能与研究感兴趣的生物学过程相关联。

GSEA 方法的优点之一是它不受预定义基因集的限制，能够全面、无偏地分析整个基因集的表达模式。此外，GSEA 还能够发现基因表达数据中的微小但一致的变化，因此在分析复杂生物学过程和疾病机制时具有一定的优势。

然而，GSEA 方法也存在一些局限性。首先，它对基因表达数据的排序十分敏感，因此对数据质量和准确性要求较高。其次，GSEA 需要大量的计算资源和时间，尤其是在分析大规模基因表达数据时，需要进行大量的排列来估计统计显

著性。因此，在实际应用中，需要权衡计算成本和结果的可解释性。

2.4.3.2 GSEA 富集分析结果分析

对于 GSEA 富集分析，由于需要进行富集分析的数据较少，因此我们使用了 WebGestalt 在线分析工具[20]。WebGestalt 是一个在线生物信息学工具，提供 ORA（Over-Representation Analysis）、GSEA（Gene Set Enrichment Analysis）和 NTA（Network Topology-based Analysis）三种富集分析方法。WebGestalt 还提供了多种可视化工具，如饼图、柱状图、网络图等，帮助用户直观地展示分析结果。此外，WebGestalt 还支持基因集集成和比较分析，帮助用户发现不同实验条件下的共同和特异性富集通路或功能。

我们使用 WebGestalt 在线分析工具查看了在不同比对条件下上调基因集或下调基因集在两个通路数据库中的富集情况。具体来讲，对于每个比对条件，我们将所有三种组学的上调下调基因的并集富集到了 KEGG 数据库中（如图 8 所示）。

通过查看在 FHH 培养 0 天到 5 天的过程中上调的通路，我们可以发现，类固醇激素生物合成、粘多糖生物合成、其他类型的氧-聚糖生物合成、氮-聚糖生物合成等物质合成通路上调，这与 ORA 富集分析结果相符。

通过查看在 FHH 培养 0 天到 5 天的过程中下调的通路，我们可以发现，NOD 样受体信号通路、T 细胞受体信号通路、JAK-STAT 信号通路表达下调，说明肝脏转分化过程中免疫应答、炎症反应和细胞功能调节等生物学过程可能有所抑制。

同时，我们发现戊糖和葡萄糖酸盐的相互转化通路、抗坏血酸和醛酸盐代谢通路、类固醇激素生物合成通路先上调后下调，这说明部分代谢过程可能在肝脏转分化早期过程中起重要作用，且类固醇激素可能是肝脏转分化早期过程中起重要调节作用的激素。

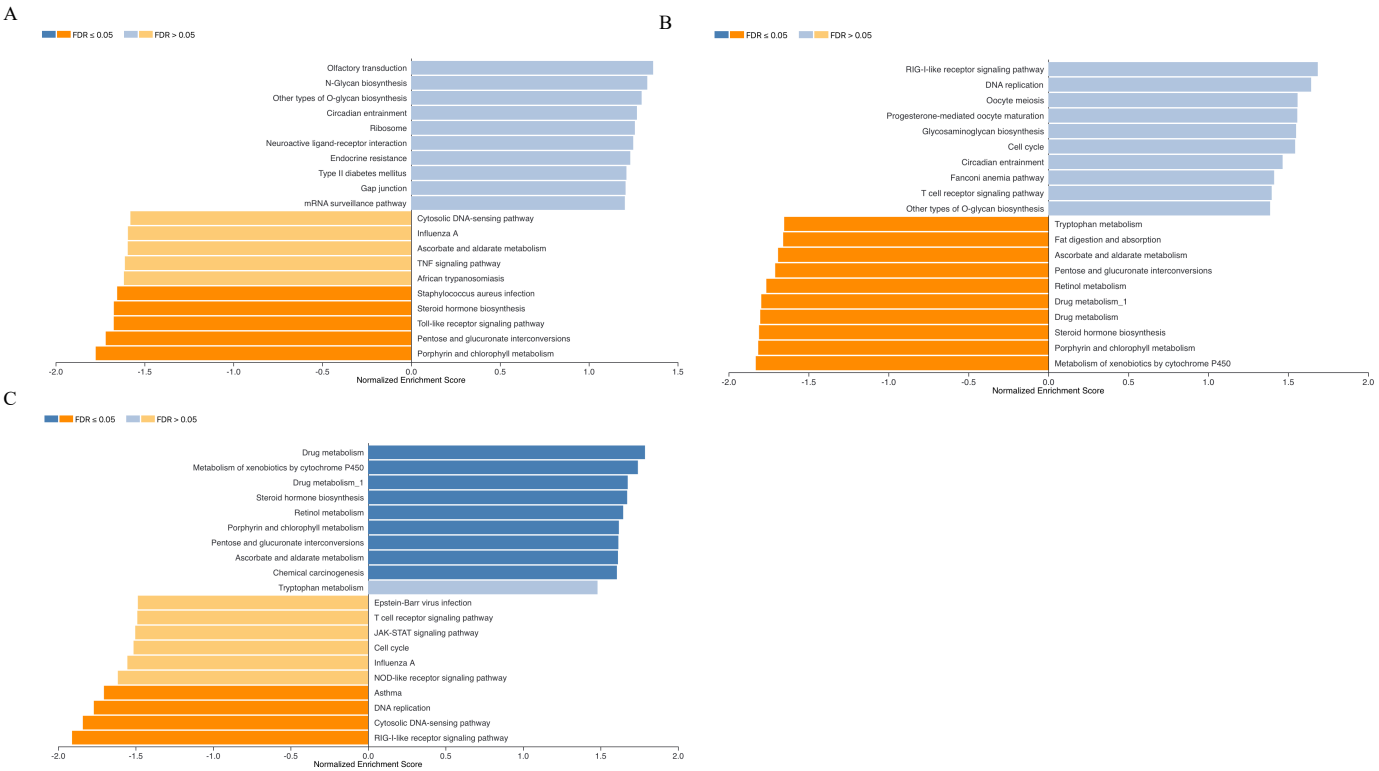

图 8 使用 WebGestalt 在线分析工具对三种比对条件下的差异基因进行 GSEA 富集分析，使用的差异基因为三个组学得到的差异基因的并集，条形图结果中只展示上调或下调通路各自的前 10 个结果，结果由富集分析的 FDR 值由小到大排列。A、B、C：分别对应 F225_G225, F5_G225, F5_F225 三种比对条件下所有组学 GSEA 的富集通路情况。

2. 4. 4 富集分析结果总结

在使用 FHH 诱导人成纤维干细胞转化为肝脏细胞的 0 到 5 天过程中，我们观察到以下几个主要变化：在上调的通路中，我们发现部分激酶和部分癌症相关基因表达水平增加，这可能与细胞信号传导、调节和肿瘤相关途径的激活有关。在下调的通路中，我们发现跨膜运输能力、细胞的趋化、信号传导能力相关基因表达水平降低。这可能意味着在转分化过程中细胞的运输能力和细胞间相互作用能力下降。在上调后下调的通路中，我们发现细胞内部骨架结构和物质合成相关基因表达在转分化早期上调，随后逐渐下调。这可能反映了细胞结构的调整和代

谢活动的变化。在下调后上调的通路中，我们发现抗原呈递、病毒感染和自身免疫相关基因表达在转分化早期下调，随后逐渐上调。这可能表明在转分化过程中免疫反应和病毒感染等生物学过程的暂时抑制，随后逐渐恢复到正常水平。

2.4 分子网络图绘制

之后，我们对富集分析得到的通路基因进行分子互作网络图的绘制，试图探明通路基因之间的相互作用关系。我们使用 String 在线数据库寻找蛋白质间互作关系网（protein–protein interaction, PPI）和 Cytoscape 工具软件[21]进行分子网络图绘制。

String 是一个广泛使用的在线数据库，专注于收集和整合蛋白质间相互作用的信息。它提供了全面的蛋白质互作数据，包括已知的直接和间接相互作用，以及不同来源的实验数据和预测数据。通过 String，研究人员可以获取有关蛋白质之间相互作用的详细信息，如结合强度、功能注释等。

Cytoscape 是一款用于分子网络分析和可视化的开源软件。它提供了丰富的功能和工具，用于构建、分析和可视化生物网络，包括蛋白质互作网络。使用 Cytoscape，用户可以将从 String 等数据库中获取的蛋白质互作数据导入软件中，并利用其强大的可视化功能，生成美观直观的网络图。此外，Cytoscape 还提供了各种插件和工具，用于网络分析、布局、注释和数据整合，帮助用户深入挖掘生物网络的结构和功能。

我们查看了在富集分析中找到的，在三个比较组中重要的通路基因之间的相互作用关系网络。具体来讲，我们绘制了细胞骨架结构成分、蛋白激酶 A 结合、趋化因子活性、胆固醇转移活性、抗原加工与递呈、病毒性心肌炎、同种异体移植排斥共 7 个重要通路基因的分子作用网络预测图。从图中我们发现，HLA 家族分子在几个重要通路中存在交集，可能为重要基因。HLA（人类白细胞抗原）家族是一组高度多态性的分子，主要存在于人体的细胞表面，它们的主要功能是向免疫系统呈递外源性抗原和内源性抗原，从而激活免疫应答。肝脏转分化过程中 HLA 家族分子的重要作用提示免疫系统参与调控，它们影响抗原呈递、免疫调节和免疫耐受，对肝脏细胞转分化和功能可能具有重要影响。

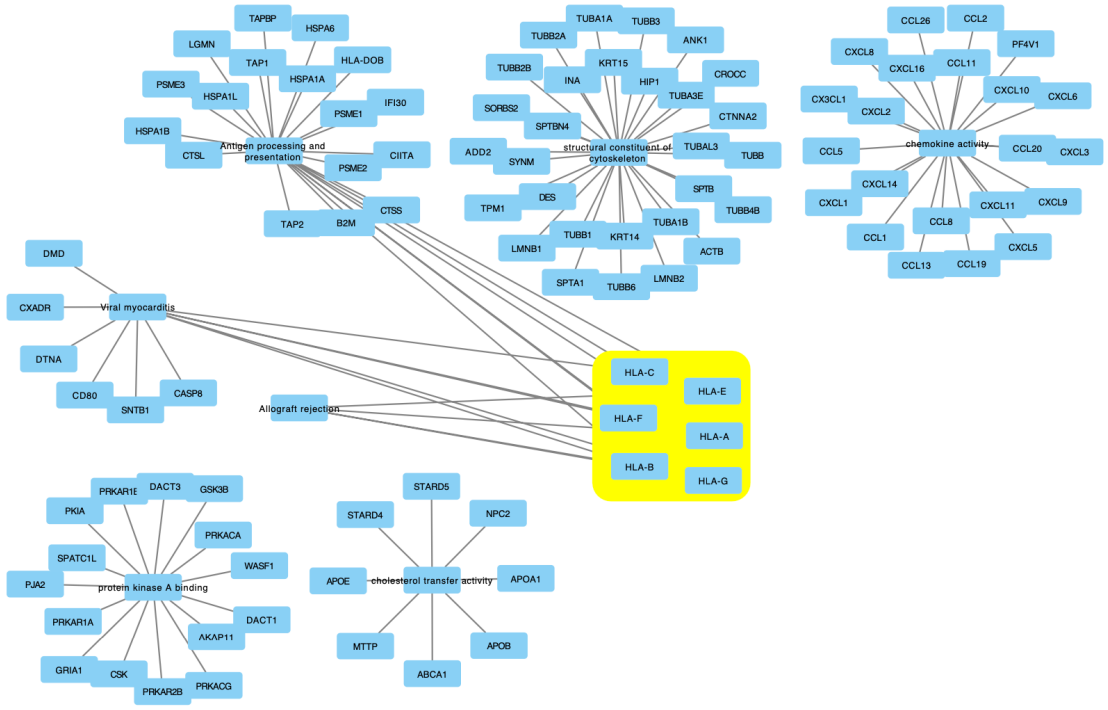

图 9 富集分析结果分子网络图，使用在细胞骨架结构成分、蛋白激酶 A 结合、趋化因子活性、胆固醇转移活性、抗原加工与递呈、病毒性心肌炎、同种异体移植排斥共 7 个重要通路基因中富集的差异基因绘制，差异基因来源为 ORA 富集分析上调、下调、先上调后下调和先下调后上调的差异基因结果。

3 讨论

肝脏转分化是一种重要的生物学过程，对于肝脏再生和疾病治疗具有重要意义。本研究旨在通过多组学分析，深入探究在 FHH 培养 0 天到 5 天的转分化过程中的基因表达、蛋白质组学以及磷酸化蛋白质组学变化，以揭示其调控机制及相关意义。

我们使用了转录组学、蛋白质组学和磷酸化蛋白质组学的数据，涵盖了 FHH 培养 2. 25 天和 5 天两个时间点的样本作为实验组，GFP 培养 2. 25 天的样本作为对照组, 数据包含 bulk RNA 测序数据，蛋白质组学数据，以及磷酸化蛋白质组学数据，数据来源为 Yangyang Yuan 等人的研究。

肝脏转分化过程涉及复杂的信号通路和生物学过程，其调控机制至今尚未完全阐明。通过本研究的分析，我们尝试深入理解肝脏转分化的调控机制，为肝脏再生与修复提供新的思路和策略。此外，对于肝脏疾病的治疗和预防也具有重要意义。

我们的研究发现，在肝脏转分化的过程中，涉及到许多信号通路和生物学过程的基因和蛋白质发生了显著变化。具体来说，在上调的通路中，激酶和癌症相关基因的表达水平增加，而在下调的通路中，跨膜运输能力、细胞的趋化、信号传导能力相关基因的表达水平降低。此外，细胞内部骨架结构和物质合成相关基因在转分化早期上调，随后逐渐下调，而抗原呈递、病毒感染和自身免疫相关基因的表达在早期下调，随后逐渐上调。最后，通过绘制分子互作网络图，我们发现 HLA 家族分子在几个重要通路中存在交集，可能为肝脏转分化过程中的重要基因。

本研究虽然取得了一些重要的发现，但也存在一些局限性。首先，由于实验条件和样本数量的限制，我们的研究结果可能存在一定的偏差和不确定性。其次，我们的研究仅仅是在细胞水平上进行的，并且只通过组学数据进行生物信息学研究，对于在体动物模型的验证和应用尚需进一步研究。未来，我们将继续深入探究肝脏转分化的调控机制，结合更多的实验手段和技术手段，为肝脏再生与修复提供更加全面和深入的理解，为相关疾病的治疗和预防提供更为可靠的技术支持。

致谢

四年时间不长也不短，马上也该轮到我们本科毕业了。回望我的本科时光，酸甜苦辣历历在目。

值此之际，我想首先感谢我的论文指导老师薛宇老师。在我本科毕业论文的写作过程中，您给予了我悉心的指导和支持，帮助我克服了许多困难，让我更加深入地理解了研究课题。您的耐心指导和宝贵建议不仅提升了我的学术能力，也培养了我解决问题的能力。再次衷心感谢您在我学习道路上的指引和帮助。

此外，我也要感谢薛老师实验室的张玮之师兄，在我的论文写作过程中不厌其烦的回答我提出的各种问题，为我论文写作提出了宝贵的建议，帮助我对毕设解决的问题有了更深层次的见解，同时也提高了我论文撰写的水平。

另外，我想感谢我的本科科研导师郭安源老师。在我科研探索的道路上，郭老师给予了我无私的指导和鼓励，让我深入了解科研的精髓，并在课题设计和数据分析方面获得了宝贵的经验。郭老师严谨的治学态度和对科研的热爱深深感染着我，激励我不断前行。再次感谢您对我的关心和支持。

同时，我也要感谢郭老师实验室的陈思义师兄、岳涛师兄、罗涛师兄、谢贵燕师姐，以及其他师兄师姐，在我的科研起始阶段对我进行各方面的技术思路指导，帮助我为日后更深入的科学研究打下夯实的基础。

最后，我想感谢我的家人和同学朋友。感谢我的父母，是你们无私的支持和关爱，让我得以顺利完成学业，迈向人生新的阶段。你们的辛勤付出和包容支持是我不断成长的力量源泉。感谢我的同学以及朋友们，虽然我们不一定是同一个专业，甚至不一定在同一个学院同一个学校，但我们在大学生活中一起玩耍，共同进步的日子我永远不会忘记。希望我们都有美好的明天。愿友谊长存。

最后的最后，我想特别感谢我的女朋友帅扬。是你在我目前人生最艰难的时候在我身边陪伴我，温暖我。希望我们可以幸福地携手共进，在科研领域共创辉煌。

蒹葭苍苍，白露为霜。所谓伊人，在水一方。道阻且长，溯洄从之。

与君共勉。

参考文献

1. Huang P, Zhang L, Gao Y, He Z, Yao D, Wu Z, et al. Direct reprogramming of human fibroblasts to functional and expandable hepatocytes [J]. Cell Stem Cell. 2014;14(3):370-84.
2. Shi XL, Gao Y, Yan Y, Ma H, Sun L, Huang P, et al. Improved survival of porcine acute liver failure by a bioartificial liver device implanted with induced human functional hepatocytes [J]. Cell Res. 2016;26(2):206-16.
3. Yuan Y, Wang C, Zhuang X, Lin S, Luo M, Deng W, et al. PIM1 promotes hepatic conversion by suppressing reprogramming-induced ferroptosis and cell cycle arrest [J]. Nat Commun. 2022;13(1):5237.
4. Menzel P. Snakemake workflows for long-read bacterial genome assembly and evaluation [J]. GigaByte. 2024;2024:gigabyte116.
5. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor [J]. Bioinformatics. 2018;34(17):i884-i90.
6. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR [J]. Curr Protoc Bioinformatics. 2015;51:11 4 1- 4 9.
7. Dobin A, Gingeras TR. Optimizing RNA-Seq Mapping with STAR [J].. Methods Mol Biol 2016;1415:245-62.
8. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features [J]. Bioinformatics. 2014;30(7):923-30.
9. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification [J]. Nat Biotechnol. 2008;26(12):1367-72.
10. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 [J].. Genome Biol. 2014;15(12):550.
11. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data [J]. Bioinformatics. 2010;26(1):139-40.
12. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies [J]. Nucleic Acids Res. 2015;43(7):e47.
13. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2 [J]. J Vis Exp. 2021(175).
14. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters [J]. OMICS. 2012;16(5):284-7.
15. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data [J]. Innovation (Camb). 2021;2(3):100141.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium [J].

Nat Genet. 2000;25(1):25-9.
17. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes [J]. Nucleic Acids Res. 2000;28(1):27-30.
18. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, et al. Reactome: a knowledge base of biologic pathways and processes [J]. Genome Biol. 2007;8(3):R39.
19. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles [J]. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.
20. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs [J]. Nucleic Acids Res. 2019;47(W1):W199-W205.
21. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks [J]. Genome Res. 2003;13(11):2498-504.