Genomics data analysis
基因组学数据分析

大作业布置

李程
北京大学生命科学学院、
生物信息中心、统计科学中心
邮件: cheng_li@pku.edu.cn

2023年秋季助教:
胡影绰 <hyc0322@stu.pku.edu.cn>

课件网页

周二下午3-5点，北大二教107 | 联系助教加入课程微信群

大作业布置（page 1 of 3）

作业任务：结合课程所介绍的数据类型、统计分析及做图方法，选择一个或多个组学数据集进行分析（使用R语言或其他语言），使用多种分析方法和做图类型，探索回答科学问题。1A-1C完成一项即可。需要标明选择任务1A/1B/1C。

（1A）从感兴趣的研究方向中选择文献阅读。选择一篇分析方法有代表性、组学数据可下载的论文。可在GEO数据库中搜索，下载数据页面下方的series matrix file、supplementary file或原始测序数据开始分析，例如 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61533。仔细阅读论文的方法部分和补充材料中的分析方法，从尽可能原始的数据开始做数据处理、分析和做图，目标是自己做出论文中的一部分图表并与它们比较。

（1B）对多篇相关文献完成（1A），整合不同文献的组学数据（例如相近样本的转录组、染色质开放性数据），提出和回答新问题。

（1C）使用自己研究课题中的数据集（需征得Lab导师同意、在提交的作业中对敏感信息保密化处理），单独分析或与公共数据做整合分析。选择一篇相似数据类型的文献，参考其分析思路和方法。

大作业布置（page 2 of 3）

提交方式

- 提交（1）数据分析代码，（2）中文分析报告（8-10页PDF文件）
- 报告中包括生物问题、数据描述（来源链接）、分析和作图结果、对结果的生物意义的解读、自己做的图表与原始文献中的图表的比较等内容
- 请将代码(如hw4_YourName_YourStudentID.R)和报告(hw4_YourName_YourStudentID.pdf)打包成文件“大作业-姓名-学号-学校”。教学网用户请在教学网通过教学网=>课程作业提交。非教学网用户发送到邮箱genomics2023@163.com，邮件主题为：《基因组学数据分析》大作业
- 作业提交截止时间为2024年1月7日24:00（周日）
- 鼓励与课程教师、助教、同学交流讨论，严禁照搬他人作业等抄袭行为
- 本作业预期**需要35小时**完成，占课程**总成绩40%**

大作业布置（page 3 of 3）

作业分数分配百分比

- 文献和数据的选择、对文献中问题的理解：10
- 分析代码的完整性、分析思路的科学性：20
- 分析代码的可读性：10
- 分析代码的正确性和算法效率：10
- 分析报告内容翔实：20
- 自己做的图表质量、与原文图表的比较：20
- 分析报告内容有创新发现：10

参考阅读：组学课题、作图与文稿设计
https://share.weiyun.com/HTe8f2ey

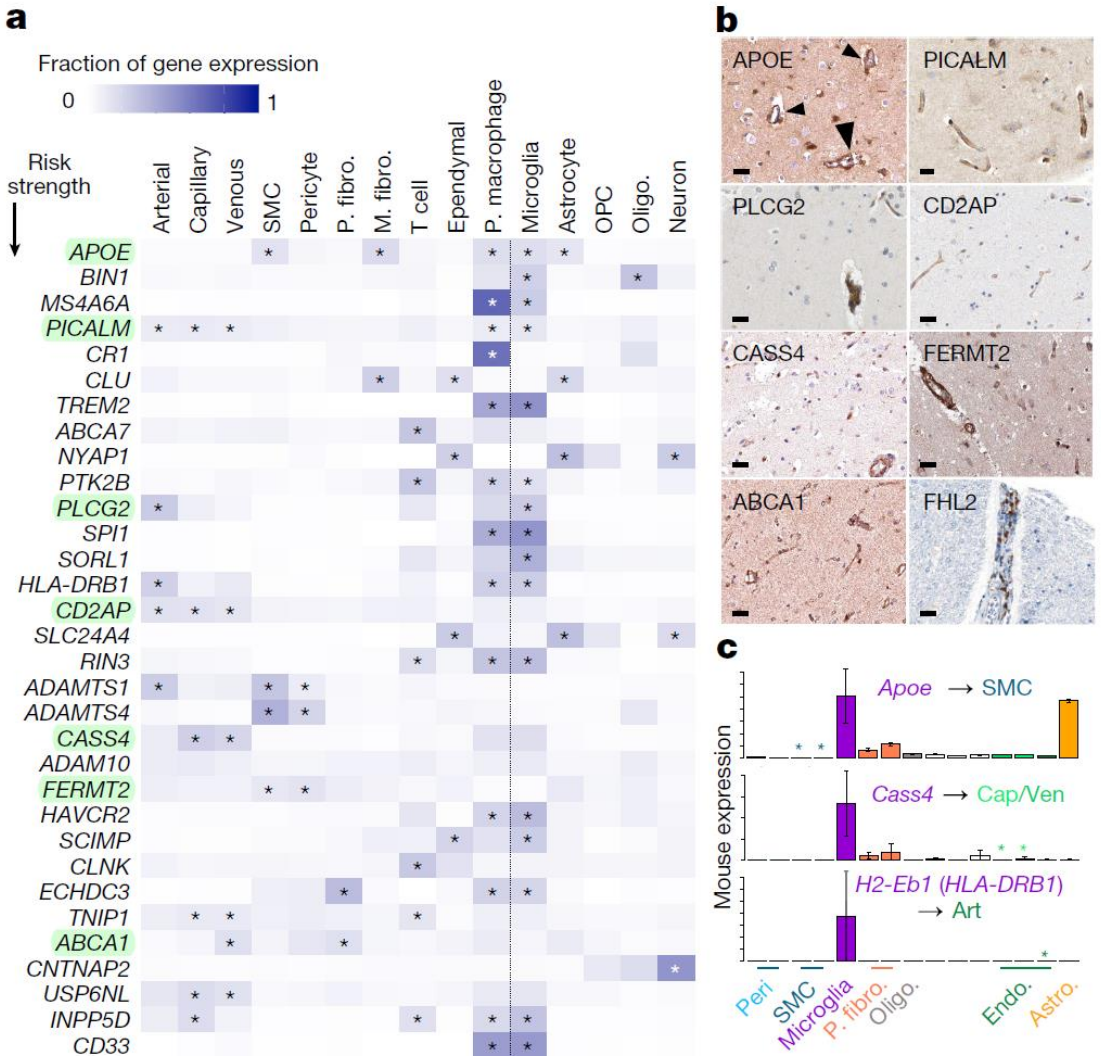
数据来源举例：人类脑疾病的多组学数据资源

Brain-related consortia

Consortium	Access	Diseases [α]	Subject size [β]	Omics assay type [γ]
psychENCODE	Controlled	Healthy, SCZ, BP, ASD	2793	(sc)RNA-seq, ChIP-seq, Hi-C, ATAC-seq, methylation, genotyping
CommonMind	Controlled	Healthy, SCZ, BP	1143	WGS, RNA-seq, ATAC-seq
AMP-PD	Controlled	Healthy, PD	10 247	WGS, total RNA-seq
AMP-AD	Controlled	Healthy, AD, PSP	4470	WGS, genotyping, RNA-seq
BRAINcode	Controlled	Healthy, PD, AD	200	Total RNA-seq, genotyping
BrainSeq	Open	Healthy, SCZ, MDD, BP	861	RNA-seq, genotyping, methylation
Aging, Dementia and TBI Study	Controlled	Healthy, TBI, AD	107	Total RNA-seq
NIAGADS - ADSP umbrella	Controlled	Healthy, AD, CBD, PSP	23 868	WGS, WES
NIAGADS - NG00057	Open	Healthy, PD, AD	18	Total RNA-seq
NIAGADS - NG00038	Open	Healthy, AD	492	Expression array, genotyping
BrainSpan	Open	Healthy	42	polyA+ RNA-seq, expression array, methylation
IMAGE-CpG	Open	Healthy, EP	17	Methylation
BOCA	Open	Healthy	5	ATAC-seq
Braineac	Dead link	Healthy	134	Expression array, genotyping
HBT	Open	Healthy	57	Expression array, genotyping
BrainCloud	Open	Healthy	270	Expression array, genotyping, methylation
NeMO	Open	Healthy	418	scRNA-seq, scATAC-seq

https://mp.weixin.qq.com/s/g9olqJpaSxw2cBR32m_Gtg

公共数据使用举例



- 这篇文章的图6利用公共蛋白染色图片（human protein atlas数据库）验证细胞类型特异表达的marker基因

- **b**, Immunohistochemical confirmation of the vascular localization of proteins encoded by top AD GWAS genes from a. Scale bars, 25 μm. Arrowheads in APOE point to signal around larger-diameter vessels, consistent with SMC expression. Image credit: Human Protein Atlas (http://www.proteinatlas.org) .

Nature 22 A human brain vascular atlas reveals diverse mediators of Alzheimer's risk.pdf

如何提出科学问题？(待补充)

- 通读一本研究方向的**经典教材**（如骆利群的《神经生物学原理》）
- 多组学数据**比较与关联**
  - **生物问题**：正常vs.疾病，年轻vs.衰老，不同发育时间点，跨物种
  - **变化层面**：表达差异，细胞亚群比例变化，细胞通信
  - **多组学、多模态**：TF motif的开放性变化作为转录变化的上游机制
  - **关联人类疾病**：差异基因/位点是否富集GWAS位点?

骆利群：“我在自己写的大学教科书《神经生物学原理》的每一章结尾设置了一些「开放问题」，意在提出那些可能在未来 5~10 年有望被解决的难题，我希望这些提出的「开放问题」能够引起大家关注，成功被大家解决，从而推动神经生物学不断前进与发展。”

课题思路举例

张国捷教授回答：

- 虽然做基因组学可以挖掘很多信息，但是从课题设计之初，就必须带着问题去设计实验，大概需要回答哪些问题，采集数据之后，在这些问题的引导下，做数据挖掘会更高效。否则不带着问题去思考工作效率会很低。

https://mp.weixin.qq.com/s/p0Nx2IHxsoqS2bawJ49-sA

“对生物学抱有真挚的爱”

- Michael Q. Zhang教授在CSHL从物理学转向生物学研究时，他认真学习了生物学及基因组学相关课程及实验操作方法。
- 这段学习经历使他意识到：当一个生物信息学专业的学生收到原始数据时，**首要任务不是深入分析数据**，而是要知道：
  - 数据是如何产生的，以及可能发生错误的主要根源;
  - 检查数据的质量和数量是否足够使用，并立即反馈此类QC信息。
- **任何计算生物学家都必须对生物学抱有真挚的爱**，并且对实验工作要有真挚的欣赏，这样才能赢得合作者的信任。

QB期刊:纪念人类基因组草图发表20周年系列文章
Michael Q. Zhang教授分享个人研究历程及学科发展思考
https://mp.weixin.qq.com/s/xWO-1DfXMZQ0AZgEE0L3CA

大作业备选问题

梁瀛（北医三院呼吸科）

- 李老师，我大概读了一下您这篇文章的结果，我看文章里的实验是基于小鼠的组织，请问有无可能用您这个算法去评价外周血免疫细胞的衰老程度或者表型？
- 目前初步思路如下：慢性阻塞性肺疾病（慢阻肺）是一种重大慢性呼吸道疾病，社会经济负担重，且发病机制尚不清楚，衰老参与慢阻肺发病机制，也是研究的热点之一；可能存在呼吸道上皮细胞衰老，也存在免疫细胞的衰老，而后者可能会参与肺部慢性炎症并引起肺结构破坏，现在希望基于李老师开发的新算法，从已有数据库中先进行一些挖掘，看是否能找到一些新的研究切入点。

Method

A transcriptome-based single-cell biological age model and resource for tissue-specific aging measures

Shulin Mao,[1,2,7] Jiayu Su,[2,3,7] Longteng Wang,[2,4] Xiaochen Bo,[5] Cheng Li,[2,6] and Hebing Chen[5]

https://mp.weixin.qq.com/s/J6hjcws87_pY4QZLb_DCCw