项目分析执行规范

一起实验网 & 中科生信生信部

Specification v1.0.0

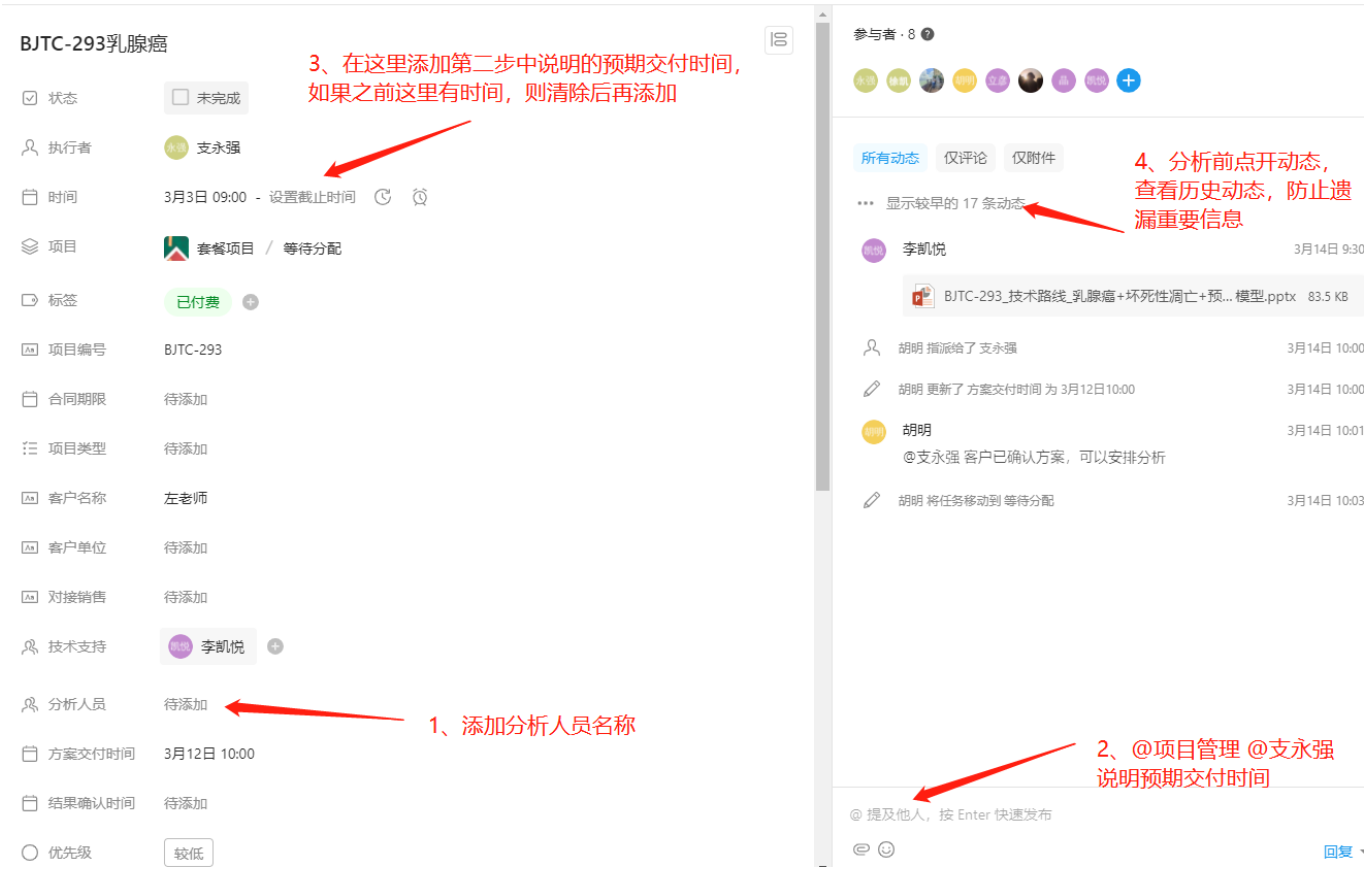
分析人员须知：

1. **项目责任制：项目分配后，该项目由分析人员全权负责，分析人员须确保结果专业准确、项目周期正常、结果可复现等**；
2. **项目责任制由项目进入分析开始，到项目结题为止**；
3. **项目责任制落实过程中遇到的任何问题，须由分析人员自己主动想办法解决，包括但不限于：google、百度、寻求同事帮助、修改分析方案等。**

1 项目分配与交付

1.1 项目分配

上一个项目快做完时，需要提前领取新项目，并对新项目中的数据集进行核对，以防数据集无法使用。



接到项目之后，要做**五件事**，其中四件如上图所示：

1. 填写分析人员为自己；

2. 填写预期交付结果时间，并@项目管理、@支永强；
3. **更新执行者下面的截止时间，和第二条中的预期交付时间保持一致，如果此处有时间，则清除掉再添加**；
4. 注意查看历史动态，防止遗漏重要信息；
5. 将项目从等待分析列表拖到分析中列表。

注：小项目领取和套餐项目基本一致。

1.2 项目交付

1. 将报告（word格式）上传到Teambition，进行审核；
2. 项目审核共有三次，分别为：分析部审核->方案部审核->框架部审核。每次审核不通过都要按要求修改后重新提交审核，最后框架审核通过后才可以进行结果交付；
3. 终审通过后，需要将结果打包上传到服务器 /data/nas2/project_results 目录下，此处需要做到以下几点：
    - 结题报告转换为pdf格式，将word格式和pdf格式的报告一起交付；
    - 将方案文档转为pdf格式，将word格式和pdf格式和结题报告一起交付；
    - 删除结果中不适合交付的文件，例如RData、不必要的中间大文件、脚本等；
    - 各个分析点结果需存放在一个单独的目录，每个目录命名格式必须以数字开头，加下划线和分析点简写英文，例如：00_rawdata、01_DEG、02_WGCNA等；
    - 将结果文件压缩，压缩文件名称格式为：项目编号_results.zip；
    - 将原始数据压缩，压缩文件名称格式为：项目编号_rawdata.zip；
        - **原始数据指整个项目最开始用的原始数据，不是各个分析点的输入数据；各个分析点的输入数据应和其结果文件保存在一个文件夹下；**
        - **此条仅适用于原始数据很大的项目，普通项目可以不单独设“项目编号_rawdata.zip”；**
    - **注意压缩格式均为zip格式，不允许使用其他压缩格式**；
    - 将 项目编号_results.zip 和 项目编号_rawdata.zip 复制到 /data/nas2/project_results 目录下；
    - 到Teambition @支永强和@项目管理，告知文件已上传，并写明具体路径和名称，例如：/data/nas2/project_results/BJTC-111_results.zip、/data/nas2/project_results/BJTC-111_rawdata.zip。

1.3 项目返修

返修指项目审核通过之后由于客户或者框架老师提出问题而导致的返工。 **返修**项目会有两种方式分配：

1. 未收费的项目返修，由提出返修的人建立**返修项目追踪记录**，并关联主项目，交给@支永强分配，返修完成后由提出人确认该任务状态；
2. 收费的项目返修，由项目管理建立小项目，并关联主项目，交给@支永强分配，返修完成后由项目管理确认该任务状态。

返修项目类型分为以下几种情况：

1. 原结果更新：
    - 将需要更新的结果文件命名为：原文件名称_修改日期_修改1.文件名称后缀 的格式；
    - 将新文件放到整体结果中，路径和旧结果保持一致；
    - 不删除旧结果；
    - 将更新后的整体结果重新压缩，压缩文件格式为zip格式，更新服务器/data/nas2/project_results目录下的压缩包；
    - 在teambition上@支永强和项目管理，提示更新完成。

2. 新增结果：
    - 和第一条类似，只不过命名时没有源文件名称，最终文件名称格式为：新文件名称_修改日期_修改1.文件名称后缀。
    - 其余和第一条相同；
3. 没有结果变动，只是回答问题：
    - 将回复写在word中，上传到teambition；
    - 在teambition上@支永强和项目管理，提示回复完成。

2 项目分析

1. 每个项目分析路径应为：/data/nas1/自己姓名/project下，对应的项目编号目录名称中，并加上序号加以区分，例如：01_project_BJTC-111;**不能在家目录（/home/自己名字）下进行数据分析**；
2. 建议一个项目对应一个R工作空间，如果项目比较复杂，也可以一个项目分为多个R工作空间；
3. 分析时注意代码的编写规范性以及确保代码可重复性；
4. 分析时将方案内容梳理清楚，为每个分析点建立一个目录，将对应的结果放在目录下，各个分析点的目录名称须有序号区分，例如：00_rawdata、01_DEG、02_WGCNA等；
    - 建议使用如下代码管理目录：

```
setwd("/data/nas1/zhiyq/project/09_project_BJTC-182_analysis/")
if (!dir.exists("00_rawdata")){
  dir.create("00_rawdata")
}
setwd("./00_rawdata")
```

5. 建议使用git进行版本控制，注意使用.gitignore规避大文件（github仓库容量限制在1G，单个文件不能超过100M，有50M的文件，就会警告了）；
6. **分析时，有任何问题请在teambition上说明，包括项目遇到问题、项目方案更改等等**；
    - **所有备注须@支永强，若有其他需求可再@其他人，例如该项目对应的方案同事；**
    - **备注描述需要逻辑清楚、前因后果明确；**
    - **如果备注为项目遇到的问题，那么备注除了描述问题本身之外，还需要写清楚想要得到什么样的帮助，并@对应人员；**
    - **想更改方案需@方案同事，任何方案修改均需方案确认，否则视为无效，审核不通过。**
7. 差异分析
    - 使用箱线图查看数据分布；
    - Count数据：DESeq2
    - 芯片数据：limma
    - DESeq2的result函数返回的结果中第一行可以查看差异分析的分组顺序；
    - limma包的make.contrast函数可以查看差异分析的分组顺序，其结果中的+1为Case，-1为Control，则差异分析为Case vs Control；
8. 预后模型类项目的一些问题：
    - 单因素+lasso，使用lasso系数构建风险分数模型；
    - 单因素+lasso+多因素，使用多因素系数构建风险分数模型（不推荐）；
    - 单因素+多因素，使用多因素系数构建风险分数模型；
    - 将riskScore和临床信息整合，分别做单因素+多因素，筛选独立预后因子，构建独立预后模型，绘制独立预后模型的列线图；如果最后临床因素都过滤掉了，那就只用riskScore建立独立预后模型；如果riskScore也不显著，则说明分析有问题；

    - 将riskScore和临床信息整合，将全部特征全部纳入构建模型，即使不显著，也可以放在模型中，列线图中也可以放全部的；（不推荐）
9. 诊断模型ROC最低标准0.7；预后模型ROC最低标准0.65；
10. 单因素Cox分析一般阈值是0.05或者更严格；如果遇到问题项目，做大可以放宽到0.2，宽进严出；
11. 各种相关性系数最低阈值0.3；
12. 待补充。

3 结题报告

3.1 报告排版

报告排版问题参考Teambition上的“项目交付报告模板.docx”，具体下载方式如下：

“项目交付报告模板.docx”可能会不定期更新，请大家注意一下，有更新时及时替换更新后的模板。

项目报告中图片和表格的序号，请使用Word或者WPS的题注功能，禁止手动编写；项目完成后需更新全域。

3.2 原始数据

1. 写清数据来源、样本数目、基因数目、数据类型；
2. 是否对数据进行过清洗过滤，如果有，写清楚过滤标准及过滤后的样本和基因数目；
3. 如果有分组，写清分组标准及各组样本数目；
4. 不仅在写作中，分析之前也要注意，对于方案中提到的数据集：
    - 要注意核对各样本的取样来源是否一致，如果不一致，判断是否对分析有影响，同时也要体现在报告中；

- 对于某些肿瘤，例如肺癌，有多种亚型，需要核对数据集中是否是同一种亚型；如果是多种亚型，判断多种亚型是否可以混合分析；如果不能混合分析需要剔除非目标亚型，以上也要体现在报告中；
- 其他。

3.3 分析方法

1. 语言描述要求逻辑清晰，没有前后矛盾、模棱两可；
2. 语言描述后添加技术路线图；

3.4 结果展示

3.4.1 表格

1. 要求三线表或者使用Word/WPS的表格样式；
2. 不建议展示超过一页的大表，此类表格可以仅展示前10行；如确实需要展示，跨页后的表格需要添加标题行和附表提示；
3. 表格字体和正文一致，标题行加粗，全部居中；
4. 避免出现表格和表注分隔两页的情况；

示例1：

表 7. 显著性差异翻译基因

GeneName	BaseMean	log2FoldChange	lfcSE	stat	pvalue	padj
4833412C05Rik	193.5880	4.0944	0.2865	14.2915	2.47196E-46	4.44112E-42
Adamts20	116.5699	3.3494	0.2523	13.2758	3.19906E-40	2.87371E-36
Lman1l	162.2596	2.3682	0.1819	13.0182	9.63875E-39	5.77233E-35
Serpinb1c	293.6397	2.9253	0.2252	12.9906	1.38357E-38	6.21432E-35
Nox4	127.6317	2.4624	0.1962	12.5524	3.8542E-36	1.15408E-32
Cda	169.9899	2.4791	0.1974	12.5583	3.58061E-36	1.15408E-32
Gstk1	1459.9085	-1.0225	0.0820	-12.4678	1.11837E-35	2.87038E-32
Ces1d	1444.3142	-2.7984	0.2367	-11.8209	3.04281E-32	6.83338E-29
Nppa	169668.7252	4.4695	0.3949	11.3175	1.07532E-29	2.14658E-26
Ech1	19546.8436	-1.1751	0.1132	-10.3823	2.98623E-25	5.36506E-22

注：各列含义分别为：GeneName：基因名称；BaseMean：所有样本经过校正的平均 reads 数；log2FoldChange：取 log2 后的表达量差异；lfcSE：log2FoldChange 标准误差值；stat：log2FoldChange 除以 lfcSE，用于计算 pvalue；pvalue：统计学差异显著性检验指标；padj：校正后的 pvalue，padj 越小，表示基因表达差异越显著。BaseMean 值较低，padj 值将设置为 NA。

示例2：

11 个基因的系数如下表所示：

表 1. 基因系数

Gene	Coefficient
TPD52	-0.22136192574161
CDK6	0.0125944143533124
TMEM109	0.307841612891996
SATB2	0.153591821653682
SGPP1	0.0152213079726379
PDGFRA	0.113837572212426
MYC	0.0695108900727681
AXL	-0.0611874388829269
MAP2K1	0.18532415566336
FOSL1	0.027626620531137
GRM7	0.306660674477604

3.4.2 图片

1. **结果文件中每张图片包含两种格式：png和pdf**；
2. 图片颜色建议选择亮色、辨识度高的颜色；
3. 尽量避免图片占一整页的情况，在不影响可读性的情况下，可以将图片适当调小；
4. 避免图片和图注分隔两页的情况。
5. 多个图形拼接：首选R包patchwork等；如果patchwork拼接后影响图片质量，可以选择线下手动方式；
    - 图形拼接仅限于同一类图形拼接，例如10个基因的箱线图等；
    - 除第一种情形外，不建议将一个分析点的所有图拼接在一起展示在报告中，应该单独展示。
6. 图片保存：首选R包lpaper或ggsave等；如果命令保存存在问题，可以选择手动保存。

示例1：

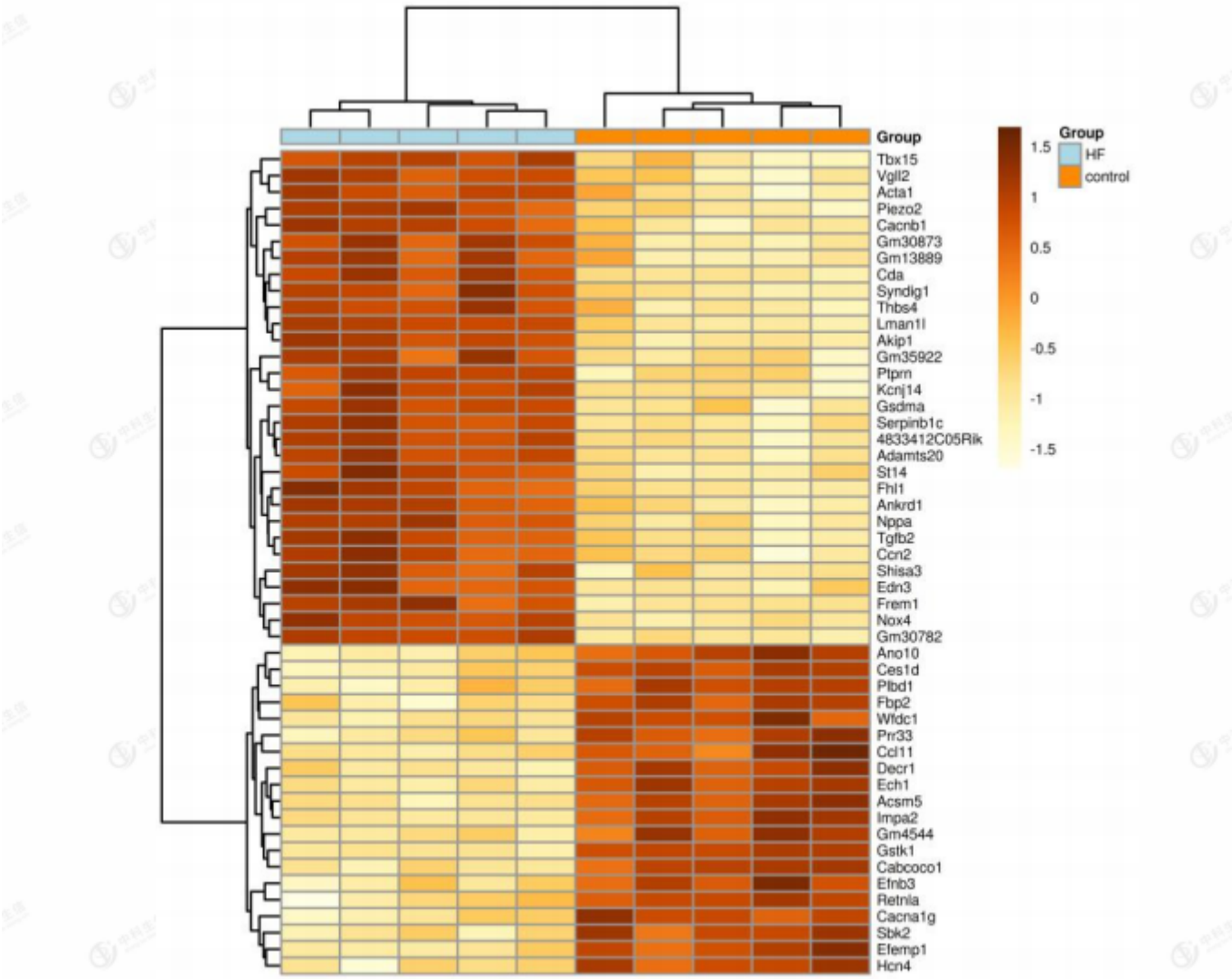


图 6. 差异翻译基因热图

3.4.3 内容

1. 各个分析点的分析方式、阈值选择需要描述清楚；
2. 各个分析点的结果，获取方式如果是通过筛选、交集、并集等额外的方式获得，需要着重说明；交集方式需要提供venn图；
3. 各个分析点之间的衔接需要逻辑自洽，这种逻辑上的承接关系，在正文描述中体现出来；
4. 正文语句需要逻辑通顺，没有语病和标点符号错误等写作问题；
5. 每个分析点的结果需要简要用正文描述一下；例如富集分析可以将富集到的通路名称写到报告中，尽量挑选和分析方向相关的通路；如果没有，则取前10个进行描述；
6. 全文中文字体应为宋体四号、英文字体为Times New Roman，其他格式和模板保持一致。

3.4.4 总结

1. 对整个分析的结果进行简单描述、总结，要求前后一致、逻辑自洽；并且对关键结果进行体现；
2. 字数上不能太少，原则上需要500字以上。

3.4.5 软件列表

1. 格式要求同上述表格要求；
2. 要求内容准确：包括软件名称、软件版本、软件用途、R包官方链接。

3.4.6 方案修改

如果方案有过修改，需要添加新的章节“方案调整说明”，在该章节说明：方案做了什么修改、为什么修改、修改后是什么样等。例如：

7. 方案调整说明

1. 原方案中根据29种免疫细胞评分进行聚类，分为高低免疫浸润组，筛选差异表达的基因与铁死亡相关基因取交集。由于聚类效果极差，考虑到本研究为铁死亡-免疫相关的方案，所以调整为：根据ssGSEA得到的铁死亡评分的最佳阈值，将AML患者分为高低铁死亡评分组筛选差异表达的基因与免疫相关基因取交集。

2. 后续部分根据原方案进行。

3.4.7 结果文件列表

在报告中的每个分析点后面列出该分析点的结果文件列表，要求使用linux tree命令得到的目录树的形式。

示例如下：

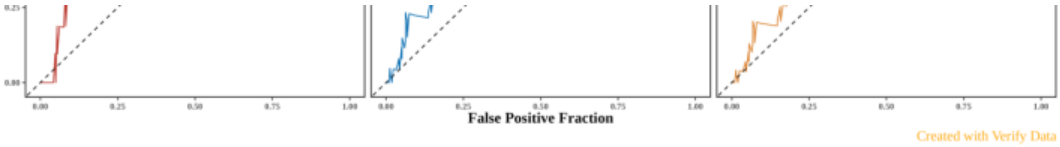


图 25. ROC 曲线

该部分结果见 07_riskModel_verify，其内容如下所示：

```
07_riskModel_verify
|-- 00_rawdata      #存放验证集原始数据的文件夹
|    |-- BLCA_GSE13507_expr.xls      #验证集表达矩阵文件
|    `-- BLCA_GSE13507_phenotype.xls      #验证集临床信息原始文件
|-- 01.verify_logdat_cli.xls      #验证集基因表达数据和生存信息整合文件
|-- 02.verify_riskScore.txt      #验证集风险分数
|-- 03.verify_km.pdf      #图 24 pdf 文件
|-- 03.verify_km.png      #图 24 png 文件
|-- 04.verify_roc.pdf      #图 25 pdf 文件
|-- 04.verify_roc.png      #图 25 png 文件
|-- 05.verify_pvalue.txt      #验证集 KM 曲线 p 值
|-- 06.verify_roc_value.txt      #验证集 ROC 曲线 AUC
|-- 07.riskScore_dis.pdf      #图 22 pdf 文件
|-- 07.riskScore_dis.png      #图 22 png 文件
|-- 08.OS_dis.pdf      #图 23 pdf 文件
`-- 08.OS_dis.png      #图 23 png 文件
```

4 报告解读

1. 分析人员需要对客户解读报告；
2. **解读完毕后，到Teambition @支永强和@项目管理，告知结题报告解读已完成；如果解读过程中客户提出了返修，也需要备注清楚。**
3. 解读过程中：
    - 每个小结讲完需询问老师有没有疑惑；
    - 背景可以略过；
    - 方法讲解以技术路线图为主；
    - 需讲解每个结果的用途；
    - 每个图片、表格都要理解到位、讲解清楚；
    - 小结不能省略，需要将整个分析串一遍。
    - 对报告中的优势部分着重强调，体现我们的专业性和优势，例如哪里的阈值选的好、哪里的结果做的好、哪里的图片质量很高等等。