Raw

每个检测的样本经过基因测序后会获得一个原始测序数据**Raw**。

**质控分析**

对每个**Raw**进行质控分析，获得**clean**。**clean**代表每个样本剩下的合格的数据。**clean**除参与拼接分析外，还将参与后面的基因表达定量分析。

Clean

将所有样本的clean混合成一个数据集，然后进行拼接分析。

拼接生成的序列集合

**拼接**

拼接分析完成后，获得一套**拼接序列集合**（**assembly**）。

Database4

Database3

Database2

Database1

assembly分别于已知序列信息的数据进行比对分析，获得**比对结果**，其中**比对结果3**和**比对结果4**的信息还将用于富集分析。

比对结果4

比对结果3

比对结果2

比对结果1

基因表达数据

结合clean和assembly，分析产生**基因表达数据**。

已知比较分组信息

结合基因表达数据和**已知的比较分析信息**，分析产生**基因表达比较结果**。已知的比较分组数目同产生的基因比较结果的数目一致。

基因比较结果3

基因比较结果2

基因比较结果1

结合**基因比较结果**、**比对结果3**和**比对结果4**，分析产生**富集结果**。

富集分析结果3

富集分析结果2

富集分析结果1

最后汇总以上分析结果，形成网页版的报告。

以上主要是从数据信息流向来概括整个转录组数据分析，每个环节还有诸多细节，但整体逻辑服从以上框架。目前我们已完成的工作：

1. 以上每步都能独立运行

期望完成的工作：

1. 将所有分析步骤整合成为1个pipeline，涉及多个步骤之间的接口问题
2. Pipeline需要具备监控功能，能够对每次运行状况参数进行记录，并将记录结果导入到数据库进行存储，便于对项目运行进行统计分析。
3. 可以考虑设计一个简单易上手的UI界面，UI界面整合整个项目所有的input信息。
4. 分析运行完后，可以直接利用产生的结果动态地生成网页版报告（存在某些由于样本量不足2个导致基因比较结果和富集结果为空）。