**Abstract (maximum 300words)**

**Introduction**

**Methodology**

数据来源：

所有项目原始meatdata文件地址来自于网址：<https://raw.githubusercontent.com/LieberInstitute/recount3-docs/master/docs/recount3_metadata_files.csv。>

Project selection

来自SRA[引用], The Cancer Genome Atlas (TCGA)[引用]和Genotype-Tissue Expression (GTEx)[引用]共18830个项目的信息表格被通过recount3 portal（https://jhubiostatistics.shinyapps.io/recount3-study-explorer/）下载。由于本项目关注的是bulk RNA-Seq的转录调控，首先需要将single-cell RNA-Seq项目排除。通过R语言grepl函数匹配project信息表格标题或者摘要中包含single cell, single-cell, scrna-seq和scrna的项目并剔除。

Data collection and processing

在这个项目中，一共有18,418个来自于公共数据库的数据集纳入研究。这些数据集对应的基因水平表达通过recount3提前下载以备后续分析。对于人类样本，选择使用Gencode 26[引用]版本注释的表达谱，而对于老鼠样本，则选择使用Gencode 23[引用]注释的表达谱。来自recount3的raw data在进行差异表达分析前需要通过recount3 package中的transform\_counts函数将raw base-pair coverage data缩放转换成reads raw count以fit in后续差异表达分析所接受的输入[引用]。

另外，由于本课题研究关注的是protein coding gene的转录调控，因此要在差异表达分析之前过滤非编码基因以避免multiple testing problem导致的差异表达基因识别的误差。根据来自Ensembl[引用]的注释信息（人类基因组版本GRCh38.p13，小鼠基因组版本）过滤掉除了protein coding gene以外的其他类型，包括Mt tRNA, lncRNA,polymorphic pseudogene等。

Metadata processing strategy（strategy和assessment，在结果中列出指标）

使用不同的策略分别对来自SRA,TCGA和GTEx的样本信息进行重注释。

SRA（画图，怎么从raw变成annotation）

首先定位到包含有生物学意义的样本信息的列（列名为sample\_attributes），然后通过注释的结构特征提取对应内容。来自SRA的样本注释为以下模式：

attribute A;;content A|attribute B;;content B|attribute C;;content C

其中;;表示对应关系，|作为分隔符。content A,B,C分别对应attribute A,B,C的内容。由于提交至SRA的样本的注释信息有较大的多样性，因此raw metadata中每个样本的注释列含有的attribute不尽相同，且数量也不同。Raw metadata table部分见表1。

表1. raw metadata table部分内容

|  |  |  |
| --- | --- | --- |
| study | external\_id | sample\_attributes |
| SRP107565 | SRR5579328 | agent;;Control|cell line;;BT20|dose;;0 uM|source\_name;;BT20 breast cancer cell line|time;;24 hr |
| SRP107565 | SRR5579329 | agent;;Abemaciclib|cell line;;BT20|dose;;0.3 uM|source\_name;;BT20 breast cancer cell line|time;;24 hr |
| SRP103846 | SRR5445265 | cell cycle stage;;G0 / quiescent|cell line;;NIH3T3|genotype;;Lin37 rescue|source\_name;;NIH3T3 mouse fibroblasts |
| SRP103846 | SRR5445266 | cell cycle stage;;G0 / quiescent|cell line;;NIH3T3|genotype;;Lin37 knockout|source\_name;;NIH3T3 mouse fibroblasts |

考虑到SRA数据库中收录的样本类型的多样性，为了能够更全面地depict每个样本的生物学性质，设计了八个目标annotations。这些目标annotation包括tissue, cell line, cell type, condition, treatment, treatment duration, genotype和subtype。SRA不同样本对于这些属性的描述是不完全相同的，表2中列举本项目捕获不同注释的关键词。

表2.目标annotation名称与相对应用于进行正则表达式捕获对应内容的关键词。

|  |  |
| --- | --- |
| Annotation name | Attribute keywords |
| Tissue | Tissue-type, Tissue type, Source name, Tissue, Organism part |
| Cell line | cell line, cell\_line, cell line/strain, cell line background |
| Cell type | cell type, CellType |
| Condition | disease status, Disease, Diagnosis, Infection agent, Infection, Health state, developmental\_stage, health state, mutational status, sample comment, histological type, Description |
| Treatment | Treatment, Treated with, Agent |
| Treatment duration | Dose, Time |
| Genotype | genotype, genotype/variation |
| Subtype | Subtype, disease stage |

通过正则表达式从raw metadata中捕获每个样本的注释信息。正则表达式模式如下：

(?i).\*(Tissue-type|tissue|tissue\_type|organism part|source\_name);;(.\*?)(\\||$)

这是捕获样本的tissue注释信息的正则表达式。其中，(?i)表示case insensitive；.\*表示0或任何次的任何字符；(Tissue-type|tissue|tissue\_type|organism part|source\_name)中以|作为分隔符列出表2中设计的该annotation对应的attribute keywords；然后跟随两个分号以分隔attribute name和后面需要提取的对应该attribute的内容。(.\*?)即为我们所需要提取的，该样本的注释；(\\||$)则是通过转义符号;\\|’匹配两个attributes之间的分隔符vertical bar，或是以dollar sign ( $ )匹配the end of this character.

通过上述正则表达式自动化批量提取raw metadata中的具有生物学意义的annotations，并合并为包含Project id, Sample id, tissue, cell line, cell type, condition, treatment, treatment duration和genotype的processed annotation table，表1对应的部分样本注释处理后如表3所示。

表3.部分样本processed annotation

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Project id | Sample id | Tissue | Cell line | Cell type | Condition | Treatment | Treatment duration | genotype |
| SRP107565 | SRR5579328 |  | BT20 |  |  | Control | 24 hr |  |
| SRP107565 | SRR5579329 |  | BT20 |  |  | Abemaciclib | 24 hr |  |
| SRP103846 | SRR5445265 |  | NIH3T3 |  |  |  |  | Lin37 rescue |
| SRP103846 | SRR5445266 |  | NIH3T3 |  |  |  |  | Lin37 knockout |

最后将具有相同注释的样本合并为同一组。在此之前，留意到部分具有相同生物学条件的样本的注释中由于标注了该样本的biological replicate编号，无法与具有相同条件的样本合并为同一组。通过附表1中的正则表达式识别具有这些特征的样本，并删除其biological replicate编号。

TCGA和GTEx

TCGA和GTEx收录的样本分别具有一致的注释表格。截取部分具有意义的列作为用于网页展示和分析的annotation columns。TCGA的注释表格包括project ID, sample ID, project name, tissue,tumor stage, survival status, tumor status, pathological stage, clinical stage, histological type and submitter ID. GTEx的注释表格包括sample ID, run accession, project name, GTEx Public Donor ID, sex, age.

Differential expression analysis（strategy，分类和validation（sensitivity和PCA））

DESeq2[引用]包用于识别与衡量不同生物学条件下基因表达变化的差异和显著程度。这个研究致力于对注释的每个项目内不同组别进行比较，跨项目的差异表达分析暂时不可接近。面对SRA收录的多样的所有项目提出一个自动化的差异表达分析策略是具有挑战性的。

Contrasts construction strategy

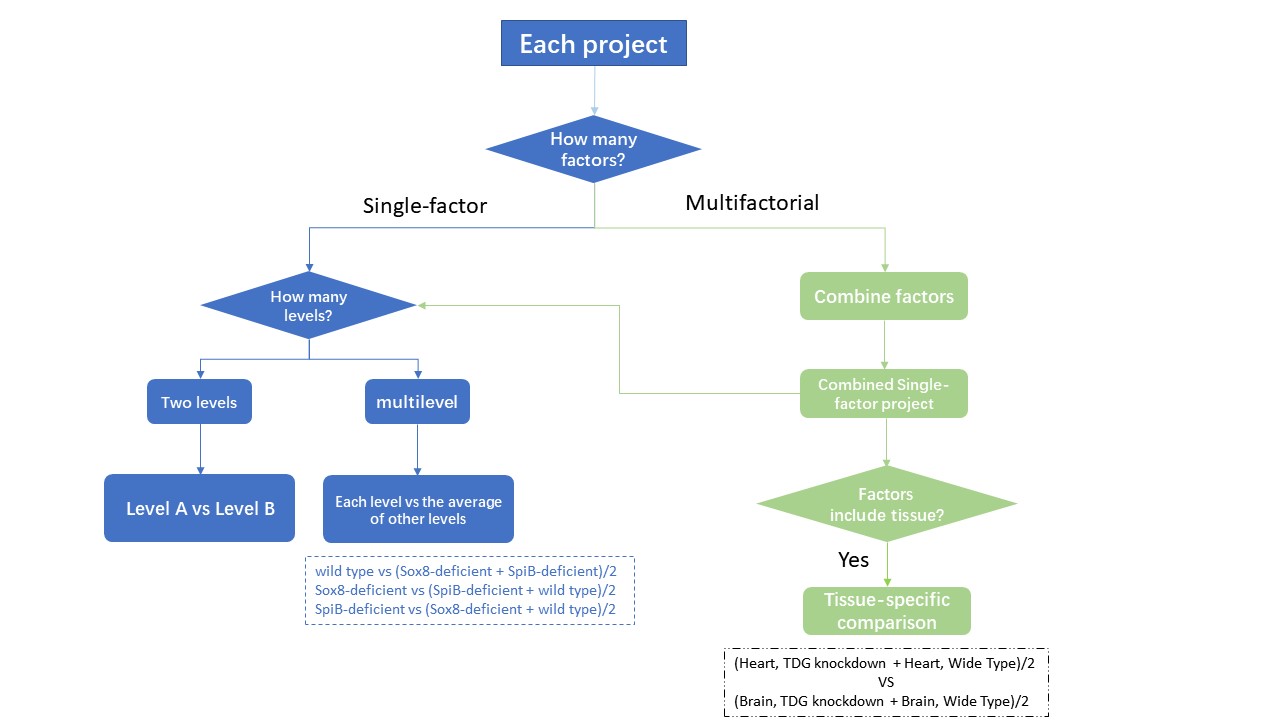
这个研究提出了一个自动批量构建contrasts的策略。该策略依据每个project的variable factor数量和factor level数量区别有不同的构建方法。

最常见的包含two levels的single factor项目，其对应的contrast被设计为comparison of level A and level B samples of this factor。对于含有多个level的单因素项目，其contrasts被设计为每个level与其余level的均值进行比较。则multilevel项目的contrasts数量与levels的数量相同。更复杂的，对于multifactorial的project，需要进行简化。通过合并variable annotation column，将multifactorial转化为single factor project，然后按照single factor的分类策略自动构建contrasts。如果multifactorial project中其中一个variable factor是tissue的话，则额外增加tissue-specific comparison。

值得注意的是，在multilevel comparison中，null hypothesis是the full change of this comparison zero。比如在comparison level A versus the average of level B and C中，reject null hypothesis并不意味着level A对应样本该基因的表达与level B和level C的表达都显著不同，而是与它们的平均有显著的差异。

使用这个策略，可以在不需要理解每一个project所有样本的生物学意义的情况下批量自动化构建contrasts。对比其他可选的自动构建contrasts的策略，即对于multilevel和multifactorial的项目递归地进行two-group pairwise comparison，这个策略避免了在motif prediction中引起的multiple testing problem。尤其是对于超过三个levels的projects，这些项目如果采用典型的two-group pairwise comparison，则每个project的comparison数量则是一个nchoose 2 combination problem，其中n是level的数量。当levels数量较大时，采用two-group pairwise comparison方法的数量会远大于本研究所采用的策略。这个策略虽然在一些情况中并没有给出最理想的contrasts design，但却能够eliminate multiple testing problem。

图1.Contrasts construction strategy.虚线框中的comparison为部分实例。



对于来源于TCGA的样本，在每个project内进行三个维度的差异表达分析。这包括tissue type、pathologic stage和histological type。Tissue type属性主要有solid tissue normal、primary tumor和metastatic三个levels。与SRA策略不同的是，在这一维度的comparison中，递归地进行two-group comparison。特别的，solid tissue normal samples是同一个患者的primary tumor或metastatic样本的配对样本，因此tissue type comparison使用paired sample test。

GTEx项目中，对于每种组织进行一次差异表达分析，使用该组织的全部样本与同样数目的来自其他组织的样本进行比较。来自其他组织的样本通过随机函数进行挑选。

Differential expression analysis implementation

在本研究中，无论是single factor还是multifactor的项目在进行差异表达分析时都转化为了single factor comparisons。因此，在构建DESeqDataSet object时，输入的coldata是有且仅有一列代表variable factor的data frame。

DESeq函数在这个项目中被用于执行差异表达分析的核心部分。它是基于negative binomial distribution对基因表达量进行模型拟合以识别组间表达量变化的函数。由于该函数会自动根据library size进行缩放，因此无需对raw count进行normalization。然而，对于低表达丰度的基因的筛选是有必要的。在这个研究中，raw count在所有的样本中的皆为0的基因（即rowSums(counts(dds)) < 1）被认为是低丰度，会被过滤掉。另外，在至少最小的一组样本数的一半样本中，该基因的CPM大于lower empirical quartile（即rowSums(cpm(dds) >= quantile(cpm(dds),obs=0.25)[2]) >= min(sample\_size)/2），否则也要exclude这个基因。通过设置参数parallel=TRUE开启并行计算以handle大量的comparisons。

最后通过results函数提取DESeq函数返回的分析结果。对于multilevel comparison，通过给contrast参数赋值相应的vector to specify the test。具体来说，如果比较level A versus the average of level B and level C，则assign level A with 1 as coefficient, while level B和level C的coefficient都是-0.5以保持平衡。那么传递给contrast的vector则为(1,0.5,0.5)。

TCGA项目的pathologic stage和histological type差异表达分析与上述的实施方法相同。而tissue type涉及到paired sample则需要借助multi-factor design来将配对的样本信息纳入design formula，并将tissue type放置于design formula的结尾。通过这种方式，deseq在估计tissue type引起的差异时会考虑paired variation。

在本研究中，GTEx样本的年龄信息被认为hidden grouping factor that需要被除去，减少噪音和更精准的探测到真正由于组织引起的差异信号。Sva（surrogate variable analysis） package被用于去除年龄这一unwanted variation。具体来说，构建两个矩阵，包括使用biological condition的矩阵和空模型矩阵作为截距项。将空模型、用于拟合模型的model矩阵以及DESeq2归一化后的count data传入svaseq函数中，该函数会estimate并返回surrogate variables。svaseq函数返回的surrogate variables

通过DESeq2处理raw count后获得的normalized count以及

sva

Case study

Validataion

PCA(DEG=0的样本)

Database

Website backend

Server端分页，连接数据库，submit转跳，GET和POST

**Results**

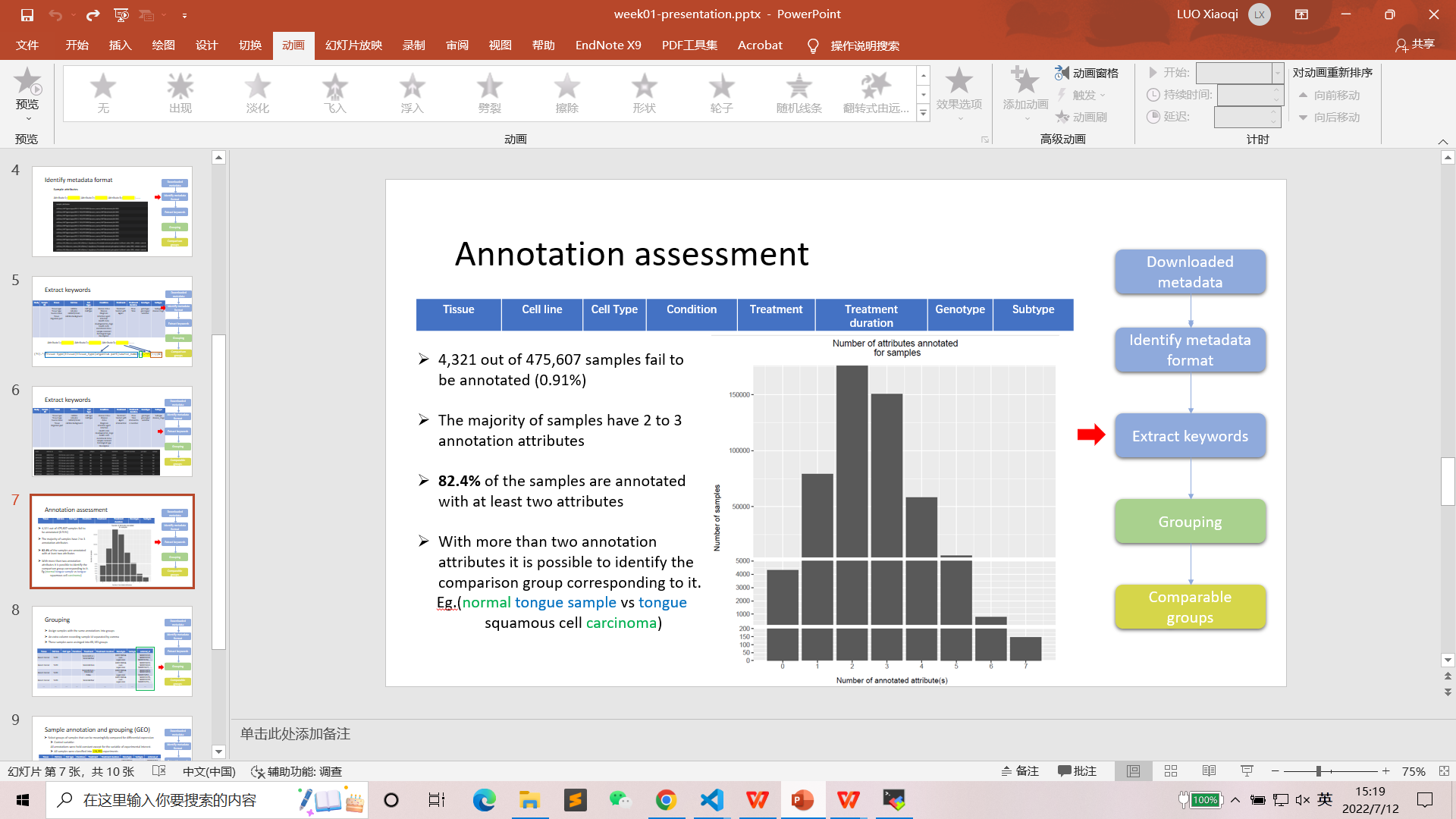
Project selection

412个单细胞测序项目。（画图：SRA,TCGA,GTEx,单细胞project数量柱状图）

Data collection and processing

过滤后又多少protein coding gene？

Metadata



有多少个样本被注释，多少project被注释。

注释数量图。

有多少group。

加一个表，对应表1。（？）

除去每组只有一个样本、每个project只有一个组后共有多少个group？

DEA

summary表格包括什么内容。画图，每种有多少个comparison。

**Discussion**

**Conclusion**

**Bibliography and referencd**

**Appendix**

附录表1

|  |  |
| --- | --- |
| Biological replicate id | Regular expression pattern |
| ①biological replicate/donor number;;1  ②donor id;;DC4  ③fresh kidney tissue endothelial cells donor 80 (M, 135 d)  ④Donor 1  ⑤PBMCs, DMSO, donor 490 | (?i)donor(#)\*( )\*(\_)\*\\d+(;)\* |
|  | (?i)(\\.|\\\_|\\#|, repeat |rep|replicate \\#|Control-rep|control replicate|control |control)\\d{1,5}(\\.){0,1}$ |
|  | (?i)[Patient|Donor|rep]( ){0,2}\\d{1,5}( ){0,2} |
|  | (?i)Repeat [A-Z]\\+[A-Z] |
|  | (?i) [A-Ea-e]$ |
|  | (?i)(\\(){0,1}Replicate(\\\_|-#| #|#)[0-9]{1,3}(\\)){0,1}$ |
|  | (?i) \\({0,1}Replicate [0-9]{1,3}\\){0,1}$ |
|  | (?i)KiaZcKO-\\d$ |
|  | (?i)control\\\_\\d |
|  |  |
|  | (?i)KiaZcKO\\-\\d$ |
|  | (?i)Biological replicate \\d{1,5} |
|  |  |
|  | (?i)replicate \\d$ |
|  |  |
|  | donation(#)\*( )\*(\_)\*\\d+(;)\* |
|  | (?i).Biological Replicate \\d$ |
|  | (?i).Technical Replicate \\d$ |
|  | (?i).Replicate( )\*\\d$ |
|  | (?i)sequence replicate( )\*\\d. |
|  | (?i)of participant( )\*\\d{1,5} |
|  | (?i), ([a-zA-Z])+ mouse |
|  | (?i)[patient |person ]\\d{1,3} |
|  | (?i)\\({0,1}(Biological Replicate |replicate |replicate)(\\d|[A-Z])\\){0,1}$ |
|  | repeat\\-\\d |
|  | replicate\\\_\\d |
|  | replicate\\\_\\d |
|  | (?i)(control)\\-\\d$ |
|  | (?i)(infection)\\-\\d$ |
|  | (?i)control \\d$ |
|  | (?i)Silence \\d$ |
|  | (?i)knock-down( ){0,1}\\d$ |
|  | (?i)knock-out( ){0,1}\\d$ |
|  | replicate \\d |
|  | \\(replicate\\d\\) |
|  | control \\d$ |
|  | (?i)control#\\d$ |
|  | (?i)replicate(\_| )[0-9]{1,3} |