RNA Seq

2022-06-25

序-生物背景

组学分析包括基因组学 (全外显子组学, 全基因组学), 转录组学 (有参, 无参), 表观组学 (chip-seq, 甲基化等)。

RNA-seq 的重要在于，很多情况下疾病的原因是蛋白的原因，而蛋白是由 mRNA 翻译而来，通过 RNA 我们可以了解到究竟是哪些蛋白发生发生了变化，或者转录出现了问题。对于组学分析而言，你需要知道哪些物种是已经被测序的，对于已测序的动物，你可以在 Ensembl 这个网站找到，植物在 Ensembl Plant 查询。

基因组包括细胞核基因组以及细胞质基因组（主要是线粒体），人基因组大概 3.1Gbp，其中 coding- area 1.1%，RNA genes 以及 regulatory area 3%。线粒体 16.6Kb。

基因在染色上分布不均匀。人类大概有 20000 蛋白基因，7000RNA 基因，6 号染色体基因数目最多。

rRNA 18 / 28 5.8 5 - 40/ 60。可以从 NCBI 下载到 rRNA 序列信息。

tRNA mapping 时的参考序列需要加 CCA, 组氨酸的参考序列 + G。有的 tRNA 含有内含子。可以从 GtRNAdb 下载 tRNA 序列

mRNA 序列信息可以从 Uniprot 下载。

microRNA 可以从 miRbase 中下载。

数量：rRNA(80-90%) tRNA(10%)，mRNA(1-5%)，直接测序大部分都是 tRNA，提取 mRNA(通过其 poly A)，或者去除 tRNA 再测序。

实验篇

RNA 测序文库

mRNA 建库

1. mRNA or total RNA
2. remove contaminant DNA (remove rRNA or select mRNA?)
3. fragment RNA
4. reverse transcrible into cDNA（strand specefic?)
5. ligate sequence adaptors

建库时的几个问题。

- 是去除 tRNA 还是特异性保留 mRNA?

  poly A+ RNA-seq 方法通过与 mRNA 的 poly A 尾巴特异性结合选择 mRNA(无法区分正负链，基因 overlap ，正负链都可能是基因)，但是组蛋白 RNA 没有 poly A 尾巴。

  rRNA - RNA-seq 方法通过酶降解 tRNA。

  两种建库方式的原始数据还是有差异的。

- 如何保留链特异性?

  通过 dUTP method，加入 dUTP 合成，切割该链条。

测序技术

illumina 测序原理可以参考这里。

reads1/reads2（/ - 互补）+ adapater = fragment
reads1/reads)（/ - 互补) = insert distance

- illumina 测不长的原因?

  每测一轮都有可能有错误（同一簇不同步)，后面越来越差，杂色参杂。

- 片段为什么要均一?

  短片段更容易在 flow cell 结合，导致测出来的都是短片段

- 为什么需要保持碱基平衡?

  150 次 flow cell 快照，每簇都同一色不好解方程。

- adapter 序列出现在结果中？insert 短，没有 150bp（只会出现在 3 端），fasta 从左到右边，只有右边有接头；测序是从 5 端到 3 端进行测序。

实验数据质控

total RNA 提取的质控标准：RIN(RNA Integrity Number)，根据 5S, 18, 28S 的峰值进行评估，范围 0-10。(6-6.5，7)

RNASeq 上游分析篇

Quality Control

质控主要需要考虑一下几个方面:

- 去除 adapter
- 去除低质量 reads
- 去除 reads 部分低质量区域
- 为下一步分析做准备；例如研究可变剪切，需要把 reads 修剪成等长。

关于 fasta 文件:

第一行，@reads 名字（测序仪编号:lane:tail:x:y)
第二行，序列
第三行，+（reads 名字/没有）
第四行，phred value + 33 对应的 ASSCI 编码
$Phred("F") = 70 - 33 = 37$(Sanger 标准)
$Phred = -10 * log_{10}(error\ probability)$

Phred40 = 0.0001 error rate
Phred30 = 0.001 error rate
Phred20 = 0.01 error rate
Phred10 = 0.1 error rate

质检报告 html 图表:

per base quality：总 reads 每个位置碱基的犯错概率（下限 q30）。
per tile sequence quality：如果有花的，那么整个 tile 可能都有问题。
GCcontent：不同物种 GC 含量不一样，样品可能被污染。
per base N content: 杂色，未测出碱基是什么。
sequence duplicate level：重复 reads 展示。
adapter content：序列的接头检测

质控参数：

quality -10

1. Phred - 10
2. 数据从右向左累加
3. 在累加的最小值处进行切割，保留前面的。

- 概览

FASTQ(QC, mapping-找 reads 在基因组的位置)- SAM（sequence alignment)- BAM（压缩的 sam 文件）

bam-质控（比对情况），计数，标准化，找差异表达

改进：翻译效率的问题，降解未结合核糖体的 mRNA 区域。RNA 结合蛋白鉴定：fastq-bam-chip-seq 待写。

mapping

- alignment（低通量，两条/多序列）pairwise multiple

  pairwise: 全局，局部（needleman-wunch, smith-waterman）

  blast: one vs many 序列切短，相似再扩展。

- mapping（reads 回溯基因组）many vs one(bwt 算法-bwa bowtie)

  1. mapping 回成熟的 mRNA 参考序列，不用处理可变剪切，不能发现新的转录本。
  2. mapping 到参考基因组，可以发现新的转录本，进行 isoform 层次的定量，但是不能使用之前的 DNA mapping 软件。

mapping 的可变剪切问题：

1. exon-first approach(pseudogene-mRNA 逆转录 cDNA 插入基因组，贴到假基因区)
2. seed-extend approch
3. potential limitations of exon-first approches。

参考基因组 (序列）下载：

UCSC genome broswer/download/human/sequence data by chromsome

合并染色体序列信息：
chr2.fa.gz `cat chr1.fa.gz chr2.fa.gz > ref_hg38.fa`

>chr1
NNNNNNNNNN(端粒的占位符)
ATCGGGGGGGG
>chr2
NNNNNNNNNN
ATCGGGGGGGG

Ensembl: ensembl species list/human/gene assemble/download DNA sequence
NCBI: refseq/

参考基因注释文件（GTF,GFF-序列哪些位置是基因等)：
ussc/tools/table browser

注释文件每列含义：待写

Note：注释文件和参考基因组 match

BWT(burrows wheeler transform)

假定：ref- ACAACG
ACAACG& 循环

1. &ACAACG
   G&ACAAC
   CG&ACAA
   ACG&ACA
   AACG&AC
   CAACG&A
2. 得到的字符矩阵根据第一列首字母排序（&ACGT)，得到排序后的字符矩阵。

3. index 就是排序的字符矩阵的最后一列

排序后的字符矩阵有下面两个性质：

- 每一行的第一个字符和最后一个字符一定在原始字符串的相连；且后一个字符在序列中在第一字符前。
- 第一列和最后一列字母（例如同一个 A..) 的相对次序不变；也就是第一列中第一个 A 和最后一列的第一个 A 实际上是序列中的同一个 A。

Index 在 mapping 中的作用如下：

1. 根据 index 还原排序后的字符矩阵矩阵的第一列
2. 根据上面两条性质反推出 reference。
3. 根据排序矩阵第一列和最后一列比对，从序列后面，矩阵第一列开始搜索

后缀树算法- suffix tree

1. 同 bwt 得到未排序的字符矩阵。
2. 按第一列首字符排序得到排序的字符矩阵。
3. 构建 suffix tree（存储树信息以及相对位置信息，index 索引非常大）
4. 从树的根节点出发。

实际上后缀树存储了序列某个字符后的所有可能性，例如某一字符为 A，后缀树就穷举了序列中 A 后面所有字符的可能并保存下来，这种做法就是用空间换时间。

一些 mapping 软件：

tophat/tophat2 构建索引：`bowtie2-build --threds 6 ref_hg38.fa ref_hg38.fa`

star-基于后缀树

1. reads 切成小的 seed，找到 seed 位置
2. 符合的 seed 拼一起

hisat2(tophat 的升级版) - 全基因组 Index，切割为 55000 份建立 index。先定位在哪个小的 index，然后在短的 index 搜索。(两层的 bwt 结构；其优点在于考虑了 SNP 信息（mapping 时可以替换）

构建 hisat2 index：SNP 信息：dbSNP commom（ucsc genome/annotation/sql/common.txt.gz）
可变剪切信息：GTF
参考基因组信息：Genome FASTA

sam flag 信息

`samtool sort` PG bam（哪些操作）

bam 文件建立索引方便查看 mapping 信息。`samtool index`，任意一段 `samtool view -h`

RNA seq 定量

通过参考基因组进行定量

样本内标准方法

除以测序深度（reads 数目），基因长度（bp），进行单位化，分子乘以 10^6（每百万 reads），乘以 10^3(基因长度 1000bp)，1 单位 rpkm 含义就是每百万 reads 中长度为 1kb 的基因的 reads 数目为 1。

r-reads, f-fragments, p-per, k-kilobase, m-million。

RPM or CPM：仅对测序深度进行单位化。

$$\frac{Number\ of\ reads\ mapped\ to\ gene * 10^6}{Total\ number\ of\ mapped\ reads}$$

RPKM, FPKM：同时测序深度以及基因长度进行单位化。对于双端测序，fragment 就是同一对 reads。单端测序 FPKM 等于 RPKM。

$$\frac{Number\ of\ reads\ mapped\ to\ gene * 10^6 * 10^3}{Total\ number\ of\ mapped\ reads * gene\ length\ in\ bp}$$

TPM:
TPM 假设每个样本的基因数值加和相同，其简单的为样本内基因 Fpkm 的百分数 *10^6。

样本间标准化

直接计算比例：

$$C_j = \frac{10^6}{D_j}$$

其中 D_j 为样本 j 测序深度，C_j 为样本 j 校正系数。这种方法容易受到极端值的影响。

quantile：

$$C_j = \frac{exp\left(\frac{1}{N}\sum_{l=1}^{N} log(D_l Q_l^{(p)})\right)}{D_j Q_j^{(p)}}$$

样本的分位数均值与某一个样本的分位数比值。其中 D_j 为样本 j 的测序深度，$Q_j^{(p)}$ 为样本 j 的 p 分位数，C_j 为样本 j 的校正系数。

RLE(relative log expression) - cufdiff；Deseq2 默认方法：

假定行为基因，列为样本。

1. 每行的几何平均数。
2. 每行除以该行的几何平均数，得到新矩阵。
3. 新矩阵每列的中位数就是该列样本的校正因子。

$$C_j = median_g\left(\frac{K_{gj}}{(\prod_{l=1}^{N} K_{gl})^{\frac{1}{N}}}\right)$$

TMM(edge R) Trimmed mean of M-values：

假设前提：total reads 受高表达基因影响。大多数基因的表达量不变。

RNA-seq 的定量在 MA plot 反应为，1）大多数点应该贴近于 M = 0 该直线附近。2）高表达基因的应该贴近于 M = 0 该直线附近———横轴是按基因的几何平均由低（左）到高（右边）排布的。

MA plot：X-A：两组样本的几何平均数

$$A = \frac{1}{2} log_2(RG)$$

M-Y：两组样本的 fold change。

$$M = log_2(R/G)$$

变化基因以及高表达的基因的阈值可选。

RSEM 转录水平定量

reads 直接 mapping 到 mRNA 上，解决可变剪切的问题。reads 来自于哪个 isoform? 从极大似然估计到 EM 算法，可以得出每个 isoform 的 count 数目。

RNASeq 下游分析篇

差异分析

基因的定量是一个抽样的结果 RNA-cDNA-RNA

RNA-Seq 的前提：

- 绝大多数基因的表达量不变
- 高表达基因的表达量不变
- 如果需要绝对定量，使用提前绝对定量的内参（spike-ins），ERCC countrol。

对于 cell line 进行差异分析，需要 2-3 个 repeat，对于生物体进行差异分析需要 3-个 repeat，对于群体而言，需要成百上千个 repeate。

- p-value 校正：

p 排序 → 新 p = p* 检验次数 m → 保持原有顺序（第一次排序的结果）调整顺序

BH 校正：p 排序 → 新 p = p * 检验次数 / 序号 → 保持原有顺序调整 p 值。

- RNA-Seq 定量的分布模型

RNA-Seq 定量的二项分布：考虑一个基因，其它所有基因为一部分。那么基因 A 的出现次数二项分布。

RNA-Seq 定量的泊松分布：因为一个基因出现的概率很低，p 很低，二项分布接近于泊松分布。

RNA-Seq 定量的负二项分布：

实际上 RNA-seq 的数据并不服从泊松分布，泊松分布期望与方差相等，然而 RNA-Seq 图像是 over-dispersion 的。随均值增加，方差也变大。这种现象称为 short noise。

泊松分布：

$$E(K_g) = Var(K_g) = \lambda$$

RNA-seq short noise 现像通过对泊松分布的修正描述。

$$E(K_g) = \lambda,\ Var(K_g) = \lambda + \phi\lambda^2$$

负二项分布具有此特性。所有 RNA-seq 基因的分布修正为负二项分布。

差异检验的零假设：

$$\lambda_g A = \lambda_g B$$

问题变成了估计 λ_g，ϕ_g，这个过程叫做 estimate dispersion，不同软件的估算方法不同。

以前的统计检验思路，列联表卡方检验或 fisher 检验。基因同分布检验，fisher 对大数字敏感，对小数字不敏感。

- RNA-seq 的绝对定量
- ERCC 建库加入定量的已知 ERCC spike-in mRNA
- house kepping(3000-4000）基因定量输入文件是未经过校正的 count 数据，认为不变的基因 list(spike in) 输出就是校正过的数据。

基因注释与富集分析

检验：GO kegg 注释就是列联表的卡方检验。
输入：一个感兴趣基因集，基因注释信息。
输出：基因基是否在某一类注释信息中。
需要认为设定感兴趣基因集的阈值。

GESA：解决人为设定基因集的问题。输入：全部的基因变化信息，一个感兴趣的通路的基因集合（GESA msi)。
输出：是否与整个感兴趣的通路相关。

图片认知（待写）

非模式生物的富集分析：有参，无参。
annotationhub。

多样本数据分析

TCGA 是肿瘤数据库，可以通过 firebrowser 下载其中的数据。GTex 是正常人的数据。

关于 pearson 相关系数以及 spearman 相关系数：spearman 相关系数仅仅是对数据的排序序号进行 pearson 相关分析的结果。

WGCNA：只是简单对基因进行聚类，通过最终聚类效果的评价指标选取距离计算的指数。

多样本差异表达基因：

将负二项模型通过对数连接函数写作广义线性模型，并对其参数进行似然比检验：

$$log\mu_{gi} = x_i^t \beta_g + logN_i$$

其中 $log\mu_{gi}$ 是基因 g 在样本 i 的观测值，x_i^t 设计矩阵，N_i 是样本的测序深度。差异检验就是对 β_g 是否全为 0 的检验。