清 华 大 学

综 合 论 文 训 练

题目：通过 RNA-Seq 估计转录本长度和辨识剪切异构体的研究

系　　别 ：自动化系

专　　业 ：自动化

姓　　名 ：李天阳

指导教师 ：张学工 教授

辅导教师 ：江瑞 副教授

2013年 7月1日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签　名：_____ 导师签名：_____ 日　期：_____

中文摘要

RNA-Seq 是最近几年发展起来的通过高通量测序对转录组中的序列进行测序的一种技术。 RNA-Seq 技术的发展使得人们在最近几年中对于生物中的基因表达的规律, 以及基因组上的功能模块有了更为深入的了解。在通过 RNA-Seq 数据确定基因的表达量时, 我们需要知道基因序列的长度。但是在没有基因注释或者没有基因组参考序列时, 我们需要一种得知基因的长度的方法。本文提出了一个通过 RNA-Seq 数据对转录本的长度进行估计的统计方法。通过该方法, 我们可以在基因组参考序列没有基因注释信息, 以及没有基因组参考序列, 的情况下使用 RNA-Seq 数据估计出转录本的长度。同时, 在 RNA-Seq 数据中我们发现读段的分布位置不均匀。此处我们对 RNA-Seq 数据中读段分布的不均匀性做了初步的分析。此外, 真核生物的基因在有多个外显子的情况下会有选择性剪切的现象发生, 同一个基因可能会产生多个剪切异构体。通过 RNA-Seq 数据我们可以辨别一个基因的不同的剪切异构体。本文证明了用最大似然的方法通过真核生物 RNA-Seq 数据辨识基因的剪切异构体是一个 NP 难问题。

关键词：RNA-Seq；转录组；转录本

ABSTRACT

RNA-Seq is a technology developed in the last few years for sequencing the transcriptome using high throughput sequencing. Using RNA-Seq, people have gained much deeper understanding of gene expression patterns, and functional modules in genomes. When estimating a transcript's expression level with RNA-Seq, we need to know the length of the transcript's sequence. However, when no annotations or reference genome sequences are available, we need another method to know the transcript's length. Here, we present a statistical method to estimate transcript length using RNA-Seq. Using this method, we can estimate a transcript's length when no annotations are available for the reference genome sequences, or when the reference genome sequences are not available. We also observed that RNA-Seq reads are non-uniformly distributed. Here, we present a preliminary analysis on the non-uniform distribution of RNA-Seq reads. Furthermore, an eukaryotic gene with multiple exons can correspond to multiple isoforms due to alternative splicing. RNA-Seq enables us to determine a gene's isoroms with RNA-Seq data. Here, we prove that using eukaryotic RNA-Seq data to identify a gene's isoforms by maximum likelihood is NP-hard.

Keywords: RNA-Seq; transcriptome; transcript

目 录

第 1 章 引言 1
1.1 RNA-Seq 1
1.2 RNA-Seq 数据分析方法简介 2
1.2.1 序列比对 4
1.2.2 序列拼装 4
1.2.3 转录本的表达量估计 4
1.2.4 辨识同一个基因的剪切异构体 (真核生物 RNA-Seq 数据) 6
1.3 现有基于 RNA-Seq 数据的转录本表达量估计和剪切异构体辨识方法 (真核生物 RNA-Seq 数据) 6
1.3.1 RNA-Seq 数据定量分析的一般模型 6
1.3.2 转录本表达量估计和剪切异构体辨识 8
1.3.2.1 Cufflinks 8
1.3.2.2 NSMAP, CEM, iReckon, 和 Montebello 9
1.3.2.3 IsoLasso 和 SLIDE 9
1.4 本文内容介绍 10
1.4.1 通过 RNA-Seq 数据估计转录本的长度 10
1.4.2 使用真核生物 RNA-Seq 数据通过最大似然辨识基因的剪切异构体是一个 NP 难问题 11
第 2 章 通过 RNA-Seq 数据估计转录本的长度 12
2.1 介绍 12
2.2 理论分析 12
2.2.1 无参考基因组序列的情况 15
2.3 实际 RNA-Seq 数据分析 19

第 3 章　RNA-Seq 数据分布的不均匀性 23
3.1　介绍 23
3.2　数据分析 23
3.2.1　方法 23
3.2.1.1　检验均匀分布的 p-value 的计算 23
3.2.2　结果 25
3.2.2.1　检验均匀分布的 p-value 25
第 4 章　使用真核生物 RNA-Seq 数据通过最大似然辨识基因的剪切异构体是一个 NP 难问题 26
4.1　问题描述 26
4.1.1　模型 26
4.1.2　M-ISOFORM 判定问题 30
4.2　证明 31
第 5 章　讨论 36
5.1　通过 RNA-Seq 数据估计转录本的长度 36
5.2　通过真核生物 RNA-Seq 数据估计基因的转录组 36
5.3　现有 RNA-Seq 数据处理方法中存在的一些问题 37
5.4　未来 RNA-Seq 数据的定量分析方法的工作 37
5.4.1　RNA-PET 测序技术 37
5.4.2　更长的测序读段 38
5.4.3　原核生物 39
5.5　总结 39
插图索引 41
参考文献 42
致　谢 51
声　明 52

附录 A　文献翻译 53
A.1 文献索引 53
A.2 翻译 53
A.2.1 比对短 RNA-Seq 读段 54
A.2.2 转录组重建 54
A.2.3 估计转录本表达量 56
A.2.4 RNA-Seq 的差异表达分析 56
A.2.5 结论和未来发展的预期 57
附录 B　源代码 58
在学期间参加课题的研究成果 59

主要符号对照表

RNA	ribonucleic acid, 核糖核酸
DNA	deoxyribonucleic acid, 脱氧核糖核酸
cDNA	complementary DNA, 互补 DNA
mRNA	messenger RNA, 信使 RNA
rRNA	ribosomal RNA, 核糖体 RNA
RNA-Seq	RNA sequencing, RNA 测序
RNA-PET	RNA paired end tag, RNA 双末端标签
EST	expressed sequence tag, 表达序列标签
bp	base pair, 碱基对 (用于表示碱基序列长度的单位)
i.i.d.	independent and identically distributed, 独立同分布

第 1 章　引言

1.1　RNA-Seq

RNA-Seq 是对 RNA 序列进行测量的一种技术, 它是近几年来发展起来的通过深度测序用于研究转录组的一种技术。与其他的方法相比起来, RNA-Seq 揭露了生物的转录组中更多的复杂性。同时, RNA-Seq 能够更好地研究生物的转录组中各转录本的表达量。通过了解细胞中在某种特定条件下转录组中的转录本的组成, 以及每一个转录本的表达量, 我们可以了解基因组上的不同的功能模块, 进而了解生物的发育过程, 以及疾病与人体之间的关系。通过使用 RNA-Seq, 我们已经对编码蛋白质的基因以及它们的剪切异构体 (isoform) 有了更为深入的了解。此外, RNA-Seq 也帮助我们对于基因上的非编码区域有了更为深入的认识, 例如 lncRNA (long non-coding RNA)。并且我们对 sRNA (small RNA), microRNA 等也有了更全面的了解。[4–15,15–35]

在 RNA-Seq 技术出现之前, 人们主要通过微阵列 (microarray) 对转录组进行定量分析和研究[36]。但是与 RNA-Seq 相比, 微阵列存在若干问题[1], 例如微阵列只能在序列已知的情况下使用, 微阵列的结果会受交叉杂交 (cross-hybridization) 影响[37,38], 以及微阵列测量的表达量范围有限等。另外, 在 RNA-Seq 技术出现之前, 人们主要通过 Sanger 测序法对互补 DNA (cDNA) 序列或者表达序列标签库 (EST libraries) 进行测序来研究转录组的序列[39,40]。但是 Sanger 测序价格昂贵, 同时测序通量与新的测序技术 (例如 Illumina 的测序技术) 相比偏低, 无法对转录组进行定量分析和研究。高通量测序技术的发展使得我们能够在较短的时间内用较少的成本对大量序列进行测序, 几乎完整地测量转录组的序列, 同时建立测序数量和实际被测序分子的数量间的关系[41]。这些是 RNA-Seq 在当今被广泛应用的一个主要原因。

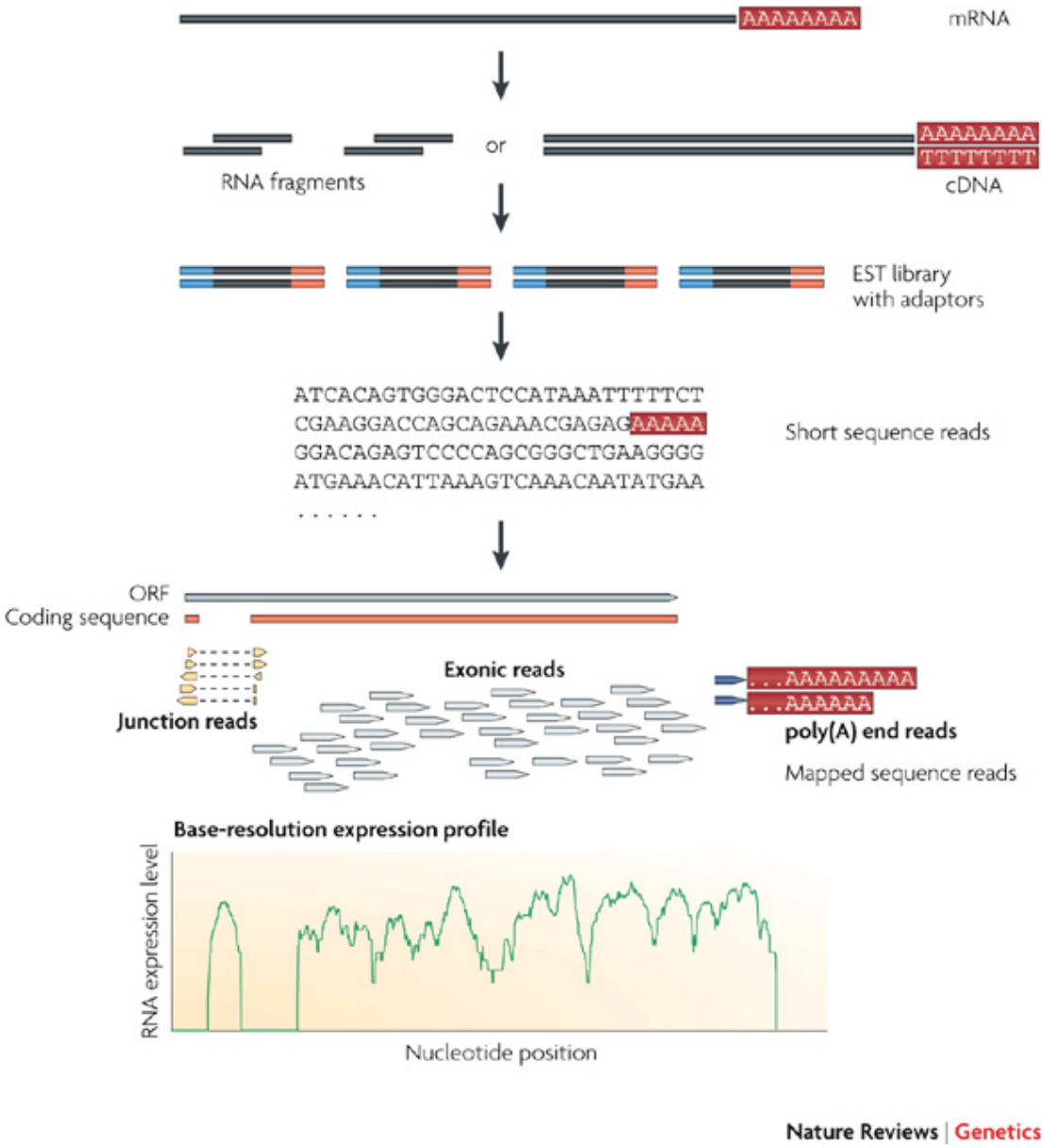


图 1.1　一般的 RNA-Seq 实验流程[1]: 长的 RNA 首先被打断成为 cDNA 库, 然后通过测序得到短的读段。然后这些读段被比对到基因组参考序列上。

1.2　RNA-Seq 数据分析方法简介

RNA-Seq 通常直接用于发现新的剪切异构体[45–48], 研究 RNA 序列的对于生物的调控作用[49] 以及比较不同条件下转录组不同的构成[44] 等生物问题的研究。同时, RNA-Seq 数据还未研究其他生物问题提供了宝贵的原始数据, 例如对于生物系统中的网络的研究[23,50]。

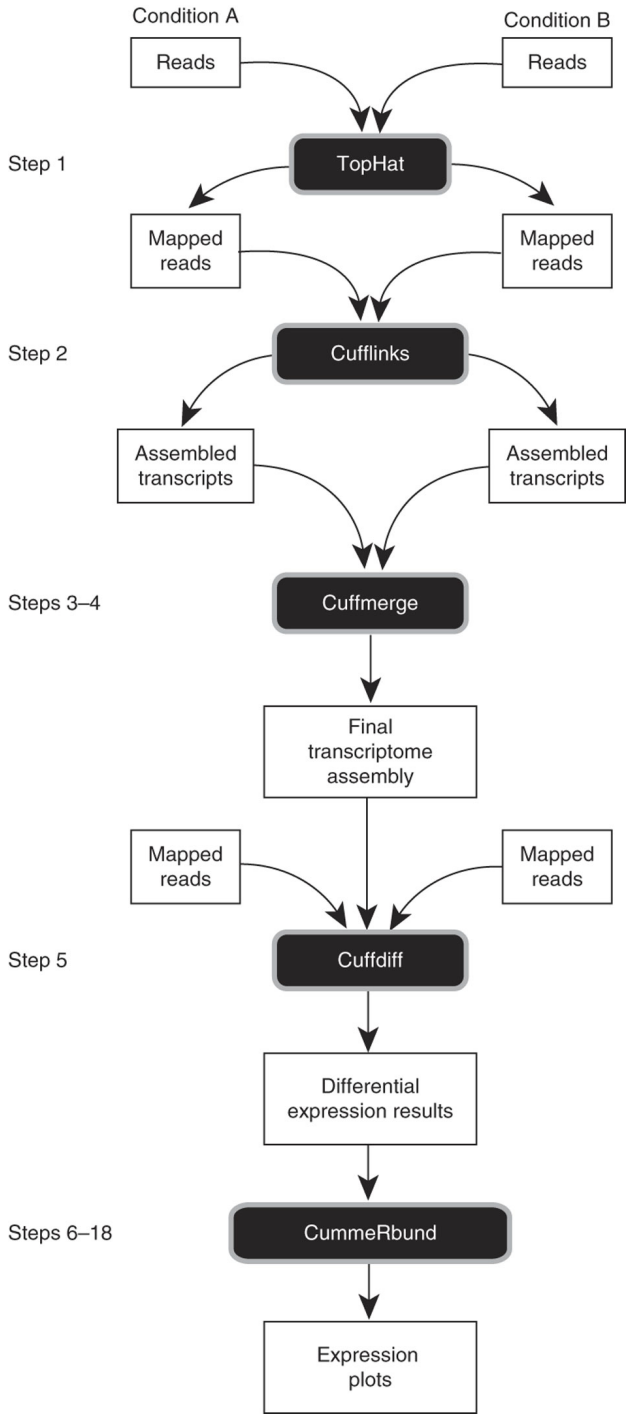


图 1.2　一般的通过 TopHat[42] 和 Cufflinks[43] 分析 RNA-Seq 数据的流程[44]: 在一个有两种不同的条件的实验中, 读段首先被 TopHat 比对到基因组参考序列上, 然后 Cufflinks 根据 TopHat 的比对结果组装转录本, 并且估计每一个转录本的表达量, 最后对两个不同的条件做基因差异表达分析。

为了能够通过 RNA-Seq 数据研究生物问题, 我们通常会对 RNA-Seq 数据用这些方法进行处理:

- 序列比对 (sequence alignment): 将 RNA-Seq 读段找到其在基因组中的位置。
- 序列拼装: 直接将 RNA-Seq 读段拼装成原转录组的各转录本。若拼装时不依赖已知的参考序列, 则成为 *de novo* 拼装 (*de novo* assembly)。
- 转录本的表达量估计: 用一定的单位表示每个被测样本中转录组里每个转录本的含量, 用于在不同的样本之间进行比较。
- 辨识同一个基因的剪切异构体 (真核生物 RNA-Seq 数据): 真核生物中由于选择性剪切同一个基因会产生多个剪切异构体, 通过 RNA-Seq 数据在一定条件下可以辨识出一个基因的不同的剪切异构体。

1.2.1 序列比对

经典的序列比对算法包括 Smith-Waterman 算法[51] 和 Needleman-Wunsch 算法[52]。但是由于 Smith-Waterman 算法和 Needleman-Wunsch 算法复杂度偏高, 它们不适用于大规模的数据。目前人们一般采用一些启发式的方法进行序列比对[53]。对于将 RNA-Seq 读段比对到其在基因组中的位置, 目前比较常用的工具是 BLAT[54] 和 TopHat[42]。

1.2.2 序列拼装

序列拼装是生物信息学和计算生物学中一个经典的问题, 其目的是通过测序的读段恢复出原始的生物序列。对于序列拼装, 目前人们主要采用 OLC (overlap-layout-consensus)[55–59] 和 de Bruijn 图[60] 这两种常用的策略。

目前对于 RNA-Seq 数据的拼装主常用的工具是 Cufflinks[43] 和 Trinity[61]。其中 Cufflinks 拼装序列时使用的是 OLC 策略 (图 1.5), 而 Trinity[61] 拼装序列时使用的是 de Bruijn 图策略 (图 1.3)。

1.2.3 转录本的表达量估计

通过量化转录组中各转录本, 从实验数据中估计其表达量, 我们可以对生物系统中的结构和过程有更好的量化的认识。 RNA-Seq 技术的发展对于转录本表达量的估计有很大的帮助。在 RNA-Seq 实验中, 通常使用 RPKM (reads per

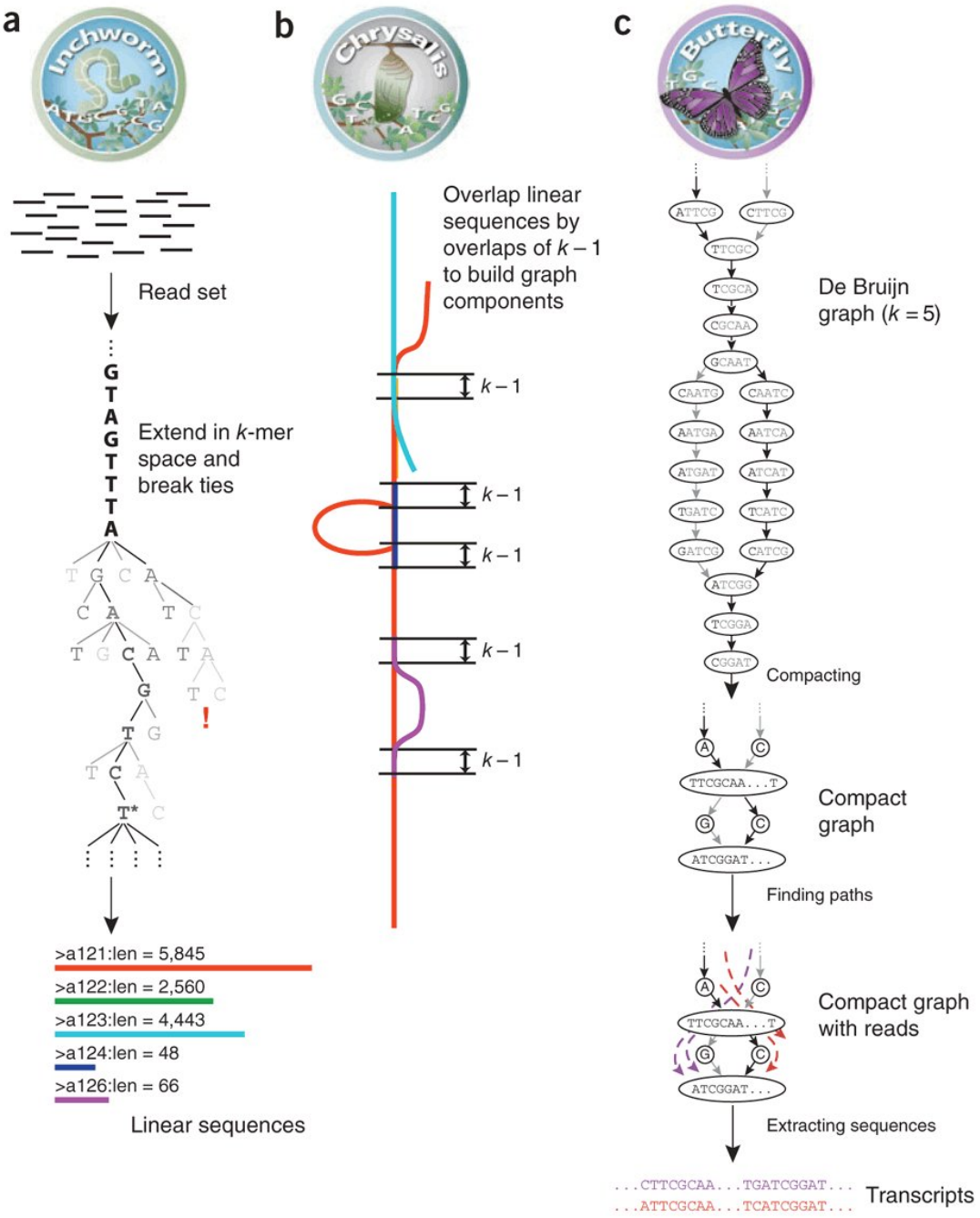


图 1.3　Trinity 拼装策略[61]: Trinity 拼装序列时使用的是 de Bruijn 图策略, 通过链接有 $k-1$ bp 重叠的 k-mer 将短序列拼接成更长的序列。

kilobase per million reads sequenced)[9] 作为一个转录本表达量的单位。目前在 RNA-Seq 实验中通常使用的转录本表达量估计的工具是 Cufflinks[43] (图 1.4) 和 RSEM[62]。

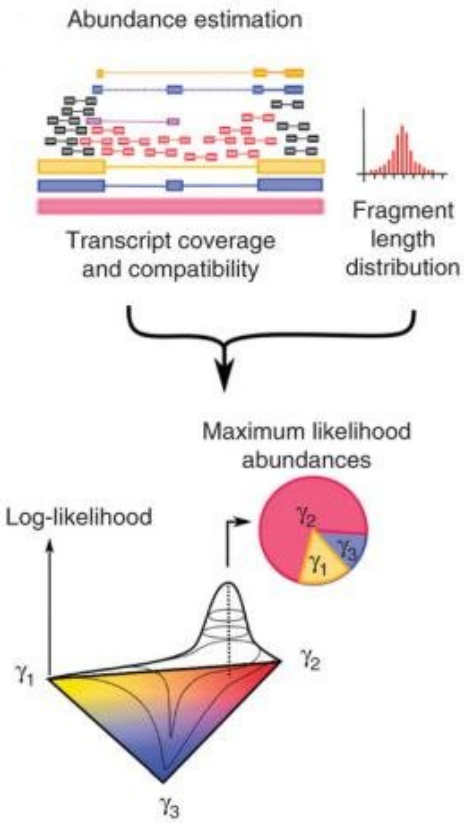


图 1.4　Cufflinks 对转录本的表达量的估计[43]: Cufflinks 使用一个多项式模型通过最大似然对转录本的表达量进行估计。

1.2.4　辨识同一个基因的剪切异构体 (真核生物 RNA-Seq 数据)

经过多年的研究, 人们认识到真核生物的同一个基因会有多个剪切异构体[63–67]。 RNA-Seq 技术的发展大大促进了人们对于真核生物中选择性剪切的认识, 因为 RNA-Seq 数据中的经过剪切位点的读段可以直接反应真核生物中的选择性剪切的现象。目前在真核生物相关的 RNA-Seq 实验中通常使用的用于辨识同一个基因的剪切异构体的工具是 Cufflinks[43] (图 1.5)。

1.3　现有基于 RNA-Seq 数据的转录本表达量估计和剪切异构体辨识方法 (真核生物 RNA-Seq 数据)

1.3.1　RNA-Seq 数据定量分析的一般模型

Pachter 在 [68] 中对 RNA-Seq 数据定量分析建立了一个一般的模型, 用统计的方法对 RNA-Seq 数据进行定量分析。在这里我们对这个模型进行介绍。

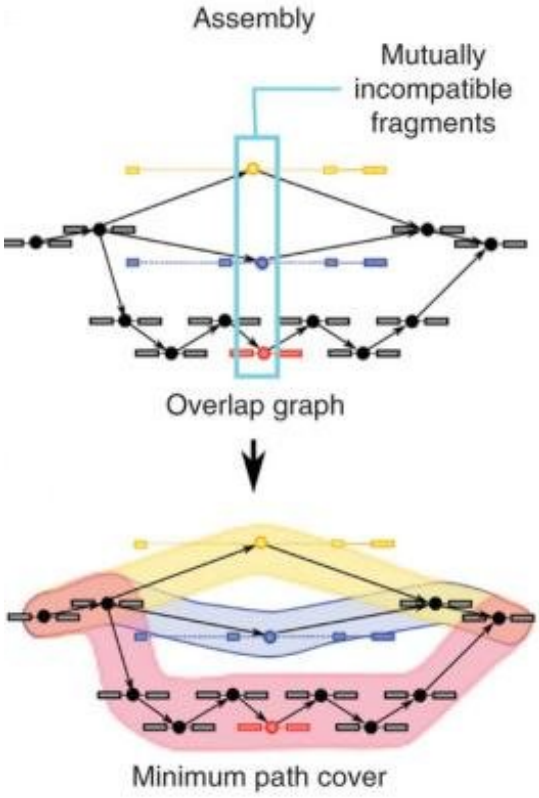


图 1.5　Cufflinks 序列拼装策略和对同一个基因的剪切异构体的辨识[43]: Cufflinks 拼装序列时使用的是 OLC 策略。另外，Cufflinks 通过在读段位置关系构成的图中寻找一个最小路径覆盖辨识一个基因的剪切异构体。

下面是为了描述该模型所使用的一些符号, 以及一些假设:

- G 表示转录组中所有包含的基因所组成的集合。
- 所有读段构成的集合表示为 R, 同时所有的读段都是单端测序得到的等长读段。 (对于双端测序数据以及不等长读段的数据的处理与这里的模型相似, 此处为了叙述简便不再赘述, 有兴趣的读者可参考 [43] 和 [68]。)
- I_g, $g \in G$ 表示基因 g 在转录组中所有的剪切异构体组成的集合, $I = \bigcup_{g \in G} I_g$ 是所有的剪切异构体所构成的集合。此处为了方便描述, 我们也将无选择性剪切的基因所对应的转录本称作是它的剪切异构体。
- $\theta_i \geq 0$, $i \in I$ 是剪切异构体 i 的表达量, 它们满足 $\sum_{i \in I} \theta_i = 1$。记 $\theta_I = \{\theta_i | i \in I\}$。在该模型下, 此处定义的表达量和实际中 RPKM 表达的表达量成正比[43]。
- $q(p, i)$ 表示对于一个剪切异构体 $i \in I$, 来自 i 的一个读段的起始位置 (5' 段)

是 p 的概率。对于给定的一个剪切异构体 i, $\sum_{p\in\{\text{所有读段起始位置}\}} q(p,i) = 1$。对于处理实际 RNA-Seq 数据中遇到的读段分布不均匀的问题我们可以通过调整 $q(p,i)$ 以改进该方法的准确性[69]。

- 当给定一个剪切异构体 i 和它上面一个读段的起始位置 p 时, $q(p,i)$ 是已知的。
- 一个读段 $r \in R$ 可以是来自多个 I 中的剪切异构体, 但是给定一个剪切异构体 i, 用序列比对的方法将 r 比对到 i 上的所有位置是唯一的。 (对于基因序列的分析表明这样一个假设是合理的[70]。)
- 对于一个读段 r 和一个剪切异构体 i, 定义 a_{ri}:

$$a_{ri} = \begin{cases} 0 & \text{当 } r \text{ 不能比对到 } i \text{ 上时} \\ q(r \text{ 比对到 } i \text{ 上的起始位置}, i) & \text{当 } r \text{ 能比对到 } i \text{ 上时} \end{cases}$$

- 对于每一个读段 $r \in R$, 都存在一个剪切异构体 $i \in I$ 使得 r 是可以由 i 产生的一个读段。
- 所有的读段都是独立的。

由以上我们给出到观察到所有读段 R 的似然函数[68]:

$$L(R) = \prod_{r\in R}(\sum_{i\in I} a_{ri}\theta_i) \tag{1-1}$$

1.3.2 转录本表达量估计和剪切异构体辨识

在已知所有剪切异构体的集合 I 时, 通过式 (1-1) 中的似然函数, 我们可以对各剪切异构体的表达量进行估计。

当我们不知道剪切异构体的集合 I 时, 我们可以通过过优化通过式 (1-1) 中的似然函数, 或者其他的目标函数, 在进行剪切异构体集合 I 的辨识的同时对各转录本的表达量进行估计[71]。已发表的方法包括 NSMAP[72], IsoLasso[73], CEM[74], iReckon[75], SLIDE[76], 和 Montebello[77]。

下面我们对若干已发表的通过 RNA-Seq 数据进行转录本表达量估计和剪切异构体辨识的工具进行介绍。

1.3.2.1 Cufflinks

Cufflinks 对于剪切异构体的辨识和转录本表达量的估计如图 1.5 和图 1.4 所示。 Cufflinks 是分开处理剪切异构体的辨识和转录本表达量的估计这两个流程

的。对于剪切异构体的辨识, Cufflinks 在读段所构成的剪切图 (splicing graph)[78] 寻找一个最小的路径覆盖, 但是 Cufflinks 并不是直接在剪切图上进行操作的。Cufflinks 通过在读段之间定义一个序, 并且在这个序构成的二分图上计算最小花费最大匹配来构成所求的最小路径覆盖。在定义该二分图各边的权重时, Cufflinks 采用了 Wang et al. 在 [10] 中的方法启发式地定义权重。

1.3.2.2 NSMAP, CEM, iReckon, 和 Montebello

NSMAP, CEM, iReckon, 和 Montebello 这 4 个工具是同时进行剪切异构体的辨识和转录本表达量的估计的。由于一个基因可能的剪切异构体的个数是随它的外显子的个数的增长按指数规律增长的, 一个基因有大量可能的剪切异构体, NSMAP, CEM, iReckon, 和 Montebello 采用了带惩罚项的似然函数进行剪切异构体的辨识和转录本表达量的估计。其具体做法是对式 (1-1) 中的似然函数增加一个惩罚项, 变为求剪切异构体的集合 I 和它们的表达量 θ_I 使得

$$e^{-\Phi}L(R) = e^{-\Phi}\prod_{r\in R}(\sum_{i\in I} a_{ri}\theta_i) \tag{1-2}$$

最大化, 其中 $\Phi \geq 0$ 是一个惩罚项, 可以使用 BIC (Bayesian information criterio)[79] ($\Phi = \frac{\sum_{i\in I,\theta_i\neq 0} 1}{2}\log|R|$) 等方法。

NSMAP 和 iReckon 均使用的是 Lasso, 而 CEM 则使用了一个 Dirichlet 分布作为先验概率。 NSMAP, CEM, 和 iReckon 在计算时采用的策略均是枚举所有剪切异构体集合 I 中的可能的剪切异构体。由于一个基因可能的剪切异构体的个数是随它的外显子的个数的增长按指数规律增长的, 所以 NSMAP, CEM, 和 iReckon 在计算时无法处理外显子较多的基因。 Montebello 则是采用一个 MCMC (Markov chian Monte Carlo)[80] 的策略, 在计算的过程中避免了枚举一个基因所有可能的剪切异构体。

1.3.2.3 IsoLasso 和 SLIDE

IsoLasso 和 SLIDE 也是同时进行剪切异构体的辨识和转录本表达量的估计的。 IsoLasso 和 SLIDE 采用的策略与 1.3.2.2 中提到的若干工具相似, 但是 IsoLasso 和 SLIDE 并不是对式 (1-1) 中的似然函数进行最大化。这两个工具均通过将一个基因按该基因的每个外显子分为若干个区域, 对每个区域计算读段的覆盖 (coverage), 并根据此通过线性回归同时进行剪切异构体的辨识和转录本表达

量的估计。这里对 IsoLasso 和 SLIDE 中所用的数学模型进行简单的介绍。通过 RNA-Seq 数据我们可以计算得每一个基因中的每一个外显子的覆盖 c, IsoLasso 和 SLIDE 均假设该外显子的覆盖 c 是所有包含该外显子的剪切异构体的表达量之和, 即在理想情况下应有

$$c = \epsilon + \sum_{i \in \text{所有包含该外显子的剪切异构体}} e_i \tag{1-3}$$

其中 e_i 是剪切异构体 i 的表达量 (这里的 e_i 与之前所描述的 θ_i 不同), ϵ 是一个期望 $[E] = 0$ 的噪声。 IsoLasso 和 SLIDE 对目前转录组中所有的基因所包含的外显子进行类似式 (1-3) 中的处理, 通过线性回归估计每一个剪切异构体的表达量。注意到通过该方法计算得的表达量值和用 RPKM 单位表达的表达量是成正比的。由于一个基因可能的剪切异构体的个数是随它的外显子的个数的增长按指数规律增长的, 一个基因有大量可能的剪切异构体, 在线性回归时 IsoLasso 和 SLIDE 均采用了 Lasso[81] 用于正则化。与 1.3.2.2 中介绍的 NSMAP, CEM, 和 iReckon 类似, IsoLasso 和 SLIDE 在计算时采用的策略均是枚举所有剪切异构体集合 I 中的可能的剪切异构体。由于一个基因可能的剪切异构体的个数是随它的外显子的个数的增长按指数规律增长的, 所以 IsoLasso 和 SLIDE 在计算时无法处理外显子较多的基因。

1.4 本文内容介绍

本文提出了一个通过 RNA-Seq 数据估计转录本长度的统计方法 (第 2 章)。同时, 本文对于 RNA-Seq 数据中读段分布的不均匀性进行了初步的分析 (第 3 章)。此外, 本文证明了使用真核生物 RNA-Seq 数据通过最大似然辨识基因的剪切异构体是一个 NP 难问题 (第 4 章)。

1.4.1 通过 RNA-Seq 数据估计转录本的长度

在通过 RNA-Seq 数据估计估计转录本的表达量时, 我们需要通过每个转录本的长度对该转录本上的读段个数进行归一化[9,43,82]。在没有基因组参考序列的情况下, 我们需要一种方法从 RNA-Seq 数据中得知转录本的长度。此处我们提出一个通过 RNA-Seq 数据估计转录本长度的统计方法。当我们假设 RNA-Seq

读段在转录本上的位置的分布是均匀时, 我们的方法给出的转录本长度的估计是无偏的。

1.4.2 使用真核生物 RNA-Seq 数据通过最大似然辨识基因的剪切异构体是一个 NP 难问题

在真核生物中一个基因会有多个外显子, 这些外显子不同的组合会产生多个不同的剪切异构体。通过 RNA-Seq 数据可以对每个基因的不同的剪切异构体进行辨识。现有的方法 (IsoLasso, CEM, 和 SLIDE 等) 都注意到辨识剪切异构体时会需要大量的计算资源。此处我们证明使用真核生物 RNA-Seq 数据通过最大似然辨识基因的剪切异构体是一个 NP 难问题。

第 2 章　通过 RNA-Seq 数据估计转录本的长度

2.1　介绍

在这里我们给出一个能够通过 RNA-Seq 数据在没有已知基因注释的情况下估计一个转录本的长度的统计方法。

对于目前大部分现有的 RNA-Seq 数据, 我们都无法直接从数据中直接获得其所包含的转录本在基因组上的 5' 段和 3' 段的位置。但是在 RNA-Seq 实验中为了能够测量每一个转录本的表达量, 我们在计算其表达量时均需要使用到该转录本的长度[9,43,82]。在使用基因组及其所对应的基因注释的情况下 (例如 RefSeq[83]), 在分析时可以直接采用注释中的基因位置得出各转录本的长度。在没有基因注释, 或者在没有基因组参考序列的情况下, 如果不采用更加复杂的技术, 如 RNA-PET[84] 技术, 则需要通过 RNA-Seq 数据估计出转录本的长度。对于现有的大部分 RNA-Seq 数据, 仍然没有对应的 RNA-PET 数据用语确定各转录本的 5' 和 3' 段。我们此处给出的方法可以通过现有的 RNA-Seq 数据估计转录本的长度。

2.2　理论分析

在通过 RNA-Seq 数据估计转录本的时候我们对 RNA-Seq 数据做如下的假设以简化模型:

- 所有的读段在原转录本上的位置的分布都是独立的。
- 读段的 5' 段位置 (以下称之为起始位置) 在原转录本上的分布是均匀的。同样的, 读段的 3' 段位置 (以下称之为终止位置) 在原转录本上的分布也是均匀的。

此处为了方便我们的讨论, 我们还假设:

- 所有的读段的长度都是相同的, 为 R bp 长。

为了在基因组参考序列没有基因注释, 甚至于没有基因组参考序列, 的情况下能够通过 RNA-Seq 数据估计出转录本的长度, 我们在这里给出定理 1 和定理

2 及其证明。通过定理 1 和定理 2 我们可以通过 RNA-Seq 数据估计出转录本的长度。

定理 1： 假设 $X_1, X_2, \ldots, X_N$ 是来自定义在 $1, 2, \ldots, L, L \in \mathbb{Z}^+$ 的均匀分布, 即

$$P(X = i) = \begin{cases} \frac{1}{L} & \text{当 } 1 \le i \le L \\ 0 & \text{其余情况} \end{cases}$$

则对于 L 的最小方差无偏估计 $\hat{L}$ 为

$$\hat{L} = \frac{{X_{(N)}}^{N+1} - (X_{(N)} - 1)^{N+1}}{{X_{(N)}}^{N} - (X_{(N)} - 1)^{N}} \tag{2-1}$$

其中 $X_{(N)} = \max_{1 \le i \le N} X_i$。

通过定理 1 我们可以在已知转录本上一个固定的位点 (例如通过读段得到的剪切位点), 根据读段的起始位置 (或者终止位置), 估计出着一个位于转录本上的固定的位点到转录本的一端的序列的长度, 从而能够估计出转录本的长度。

下面我们给出定理 1 的证明。

证明 我们首先说明

$$X_{(N)} = \max_{1 \le i \le N} X_i$$

在定理中的假设下是一个充分统计量。

由于在这里考虑的是一个离散的均匀分布, 我们容易注意到

$$P(X_1, X_2, \ldots, X_N | X_{(N)}) = \frac{1}{{X_{(N)}}^{N} - (X_{(N)} - 1)^{N}}$$

与该分布的参数 L无关, 所以 $X_{(N)}$ 是一个充分统计量。

另外, 根据离散的均匀分布的性质我们可以得到

$$P(X_{(N)} = i) = \frac{i^N - (i-1)^N}{L^N}$$

对于定理 1 中给出的估计 $\hat{L}$ (式 (2-1)), 我们可以进一步计算 $\mathrm{E}[\hat{L}]$ 得

$$\mathrm{E}[\hat{L}] = \sum_{i=1}^{L} P(X_{(N)} = i) \frac{i^{N+1} - (i-1)^{N+1}}{i^N - (i-1)^{N-1}}$$

$$= L$$

从而我们可以知道定理 1 中给出的估计 $\hat{L}$ (式 (2-1)) 是无偏的。

下面我们说明 $X_{(N)}$ 是一个完全统计量。即对于所有的 $L \in \mathbb{Z}^+$, 若有一个函数 $g(x)$, $x \in \mathbb{Z}^+$ 使得

$$\mathrm{E}[g(X_{(N)})] = 0$$

则 $g(x) = 0$ 对所有 $x \in \mathbb{Z}^+$ 成立。

注意到

$$\mathrm{E}[g(X_{(N)})] = 0$$

对所有的 $L \in \mathbb{Z}^+$ 成立意味着

$$\begin{aligned}
\frac{1^N - 0^N}{1^N} g(1) &= 0 \\
\frac{1^N - 0^N}{2^N} g(1) + \frac{2^N - 1^N}{2^N} g(2) &= 0 \\
\frac{1^N - 0^N}{3^N} g(1) + \frac{2^N - 1^N}{3^N} g(2) + \frac{3^N - 2^N}{3^N} g(3) &= 0 \\
&\cdots
\end{aligned}$$

所以根据数学归纳法, 我们容易得出 $g(x) = 0$ 对所有 $x \in \mathbb{Z}^+$ 成立。进而我们可以知道 $X_{(N)}$ 是一个完全统计量。

由以上的分析我们知道:

- $X_{(N)}$ 是一个充分统计量。
- $X_{(N)}$ 是一个完全统计量。
- 定理 1 中给出的估计 $\hat{L}$ (式 (2-1)) 对于此处的分布的参数 L 的估计是无偏的。

从而根据 Lehmann-Scheffé 定理[86,87] 我们可以得出定理 1 中给出的估计 $\hat{L}$ (式 (2-1)) 是对该分布的参数 L 的最小方差无偏估计。

□

定理 2：假设X_1, X_2, ..., X_N 是来自定义在 $p+1, p+2, \ldots, p+L$, $p \in \mathbb{Z}$, $L \in \mathbb{Z}^+$ 的均匀分布, 即

$$P(X=i)=\begin{cases}\frac{1}{L} & \text{当 } p+1 \le i \le p+L \\ 0 & \text{其余情况}\end{cases}$$

则对于 L 的一个无偏估计 $\hat{L}$ 为

$$\hat{L}=\frac{D^{N+1}-2(D-1)^{N+1}+(\max(D-2,0))^{N+1}}{D^N-2(D-1)^N+(\max(D-2,0))^N} \tag{2-2}$$

其中 $D=\max_{1\le i\le N} X_i - \min_{1\le i\le N} X_i + 1$。

通过定理 2 我们可以在转录本上无法找到一个固定的位点时, 通过 RNA-Seq 数据对转录本的长度进行估计。这种情况适用于在一个基因没有出现选择性剪切的现象时估计该转录本的长度, 例如真核生物中只有一个外显子的基因或者是原核生物中的单个操纵子。

下面我们给出定理 2 的证明。

证明 根据均匀分布的性质, 我们容易得到

$$P(D=i)=(L-D+1)\frac{D^N-2(D-1)^N+(\max(D-2,0))^N}{L^N}$$

从而我们可以知道 $\hat{L}$ 是无偏的

$$\begin{aligned}\mathrm{E}[\hat{L}] &= \sum_{i=1}^{L} P(D=i)\frac{i^{N+1}-2(i-1)^{N+1}+(\max(i-2,0))^{N+1}}{i^N-2(i-1)^N+(\max(i-2,0))^N} \\ &= \frac{1}{L^N}\sum_{i=1}^{L}(L-i+1)(i^{N+1}-2(i-1)^{N+1}+(\max(i-2,0))^{N+1}) \\ &= L\end{aligned}$$

□

2.2.1 无参考基因组序列的情况

在没有参考基因组序列的情况下, 我们希望从拼装的所有连续段 (contig) 中挑选出覆盖原先转录本部分较多 (例如该连续段是该转录本唯一的一个连续段) 的连续段。因为在没有参考基因组序列的情况下, 若要认为一个连续段代表原基

因, 则必须要求该连续段覆盖了原转录本的大部分的区域。如果一个转录本的测序深度偏低、覆盖低, 则通过测序的结果我们可能看到这个转录本会拥有多个连续段。在没有参考序列时的情况下由于无法知道来自同一个转录本的多个连续段之间的关系, 这样的连续段几乎是无法分析的。

下面我们给出我们对于选择 “覆盖原转录本较多” 的连续段的一种思路。

首先, 我们在这里介绍一个用来描述测序过程的 Poisson 模型[89]。

我们定义测序的覆盖 (coverage) c

$$c = \frac{NR}{L} \tag{2-3}$$

其中 R 是测序读段的长度, N 是转录本上测序时得到的读段的个数, L 是转录本的有效长度 (转录本实际长度 $- R + 1$)。

这里假设我们有一个无限长的序列, 每一个位置上起始的读段个数都是一个参数为 λ 的 Poisson 分布。并且所有位置的起始的读段个数均是独立的。在这里测序的覆盖和 λ 的在理想的情况下满足如下的关系

$$\lambda = cR$$

当两个读段至少有一个碱基对重合时, 我们认为这两个读段能够连接到一起组成一个更长的连续段。通过这样的过程, 我们将若干读段连接在一起构成连续段。

为了挑选覆 “覆盖原转录本较多” 的连续段, 我们在上面所述的测序模型下对每一个连续段考虑如下的一个概率

$$P(\text{观察到一个连续段其长度} \leq \text{目前的连续段} \mid \text{目前的连续段的覆盖}) \tag{2-4}$$

这里的想法是, 如果对于一个给定的连续段, 式 (2-4) 中的概率较大, 则在这个连续段 “变长” 的过程中就已经 “碰到了转录本的一端” 无法 “变得更长”。

为了更深入地了解式 (2-4) 中的概率, 我们在这里分析

$$P(\text{连续段长度} = l \mid \text{目前的连续段的覆盖}\lambda) \tag{2-5}$$

即给定覆盖, 连续段的长度的分布。下面为方便讨论我们记

$$p_n = \frac{P(\text{连续段长度} = n + R - 1 \mid \text{目前的连续段的覆盖}\lambda)}{e^{-(R-1)\lambda}} \tag{2-6}$$

则

$$p_n = 0 \qquad n \le 0 \tag{2-7}$$

$$p_1 = 1 \tag{2-8}$$

$$p_{n+R} = \sum_{i=1}^{R-1}(1-e^{-\lambda})e^{-(R-1-i)\lambda}p_{n+i} \qquad n \ge 2-R \tag{2-9}$$

为了估计式 (2-4) 中的概率, 在这里我们先介绍一个改进的 Lander-Waterman 公式 (定理 3), 原先的 Lander-Waterman 公式[90] 在推导中使用了一些近似, 在这里我们不使用 Lander 和 Waterman 在 [90] 中所使用的近似。

定理 3 (改进的 Lander-Waterman 公式)： 在上述的 Poisson 模型下, 一个连续段的期望的长度

$$\mathrm{E}[\text{一个连续段的长度}] = \frac{e^{R\lambda}-1}{e^{\lambda}-1} \tag{2-10}$$

下面我们给出定理 3 的证明。

证明 这里为了描述的1简便, 我们记

$$q = 1 - e^{-\lambda}$$
$$1 - q = e^{-\lambda}$$

为了证明定理 3 中的式 (2-10), 我们考虑相邻的读段起始位置的个数 N_D (即两个位置是相邻的两个读段起始位置, 同时这两个位置上的读段连接在了一起), 以及两个相邻的有读段的位置间的差值 D_i, $i = 1, 2, \ldots, N_D$ (即用 $X_1 = 1, X_2, \ldots, X_{N_D}, X_{N_D+1}$ 表示若干个相邻的读段起始位置时, $D_i = X_{i+1} - X_i$, $i = 1, 2, \ldots, N_D$)。根据这里所使用的模型, 我们容易知道 D_i, $i = 1, 2, \ldots, N_D$ 是 i.i.d.。

所以我们容易看出来

$$\begin{aligned}\mathrm{E}[\text{一个连续段的长度}] &= \mathrm{E}[X_{N_D+1}] + R - 1 \\ &= \mathrm{E}[\sum_{i=1}^{N_D} D_i] + R\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}[\mathrm{E}[\sum_{i=1}^{N_D} D_i | N_D]] + R \\
&= \mathrm{E}[D]\,\mathrm{E}[N_D] + R
\end{aligned}
$$

下面我们分别计算

$$
\mathrm{E}[D]
$$

$$
\mathrm{E}[N_D]
$$

根据这里所使用的模型, 我们可以知道

$$
P(D = i) \propto \begin{cases} (1-q)^{i-1} q & i = 1, 2, \ldots, R-1 \\ 0 & \text{其余情况} \end{cases}
$$

所以可以计算得

$$
\begin{aligned}
\mathrm{E}[D] &= \frac{1}{1-(1-q)^{R-1}} \sum_{i=1}^{R-1} i(1-q)^{i-1} q \\
&= \frac{(q(R-1)+1)(1-q)^R - (1-q)}{-(1-q)q(1-(1-q)^{R-1})}
\end{aligned}
$$

同时, 我们根据模型可知

$$
P(N_D = j) = (1-q)^{R-1}(1-(1-q)^{R-1})^j,\ j = 0, 1, \ldots
$$

根据几何分布的性质我们可知

$$
\mathrm{E}[N_D] = \frac{1-(1-q)^{R-1}}{(1-q)^{R-1}}
$$

综上所述, 进一步化简我们可得

$$
\mathrm{E}[D]\,\mathrm{E}[N_D] = \frac{e^{R\lambda} - 1}{e^{\lambda} - 1} - R
$$

故定理得证。

□

为了估计式 (2-4) 中在给定覆盖 λ 的情况下, 在无穷长的序列上观察到长度为 l 的概率, 我们将该分布近用一个期望是 E[一个连续段的长度] $- R + 1$ 的几何

分布, 从而估计出式 2-4 中在无穷长的序列上给定覆盖 λ 观察到一个连续段其长度小于等于目前的连续段的概率

$$
\begin{aligned}
&P(\text{观察到一个连续段其长度} \leq \text{目前的连续段} l | \text{目前的连续段的覆盖} \lambda) \\
\approx &\frac{l - R + 1}{\frac{e^{R\lambda}-1}{e^{\lambda}-1} - R + 1}
\end{aligned}
\tag{2-11}
$$

在有了式 (2-11) 中的近似估计之后, 我们可以在没有参考基因组序列的情况下挑选出来拼装结果中那些可能是 “大部分覆盖原转录本” 的那些连续段, 对这些连续段的长度通过前面定理 1 和定理 2 中所述的方法进行估计, 从而进一步估计得这些连续段所对应的转录本的表达量。

2.3 实际 RNA-Seq 数据分析

我们对来自 ENCODE[5] 的 RNA-Seq 数据 ① 进行了分析, 用来在真实 RNA-Seq 数据上测试前面所叙述的估计转录本长度的方法。在分析中, 我们只对只有一个外显子的基因进行了分析。

图 2.1 给出了我们估计的转录本的长度和真实的转录本的长度的比较。转录本实际长度和真实长度的 Pearson 相关系数 95% 置信区间为 [0.8749423, 0.8948220]。

图 2.2 给出了估计的转录本长度和真实的转录本长度之比 ($\frac{\text{估计的转录本长度}}{\text{真实的转录本长度}}$) 和测序 λ 之间的关系。通过图 2.2 中无法看出估计的转录本长度和真实的转录本长度之比和测序 λ 之间有明显的关系。

图 2.3 给出了估计的转录本长度和真实的转录本长度之比 ($\frac{\text{估计的转录本长度}}{\text{真实的转录本长度}}$) 和第 3 章中计算得的检验均匀分布经过 FDR 调整的 p-value 之间的关系。通过图 2.3 中无法看出估计的转录本长度和真实的转录本长度之比和和第 3 章中计算得的检验均匀分布经过 FDR 调整的 p-value 之间有明显的关系。

① http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/wgEncodeCaltechRnaSeqGm12878R1x75dAlignsRep2V2.bam

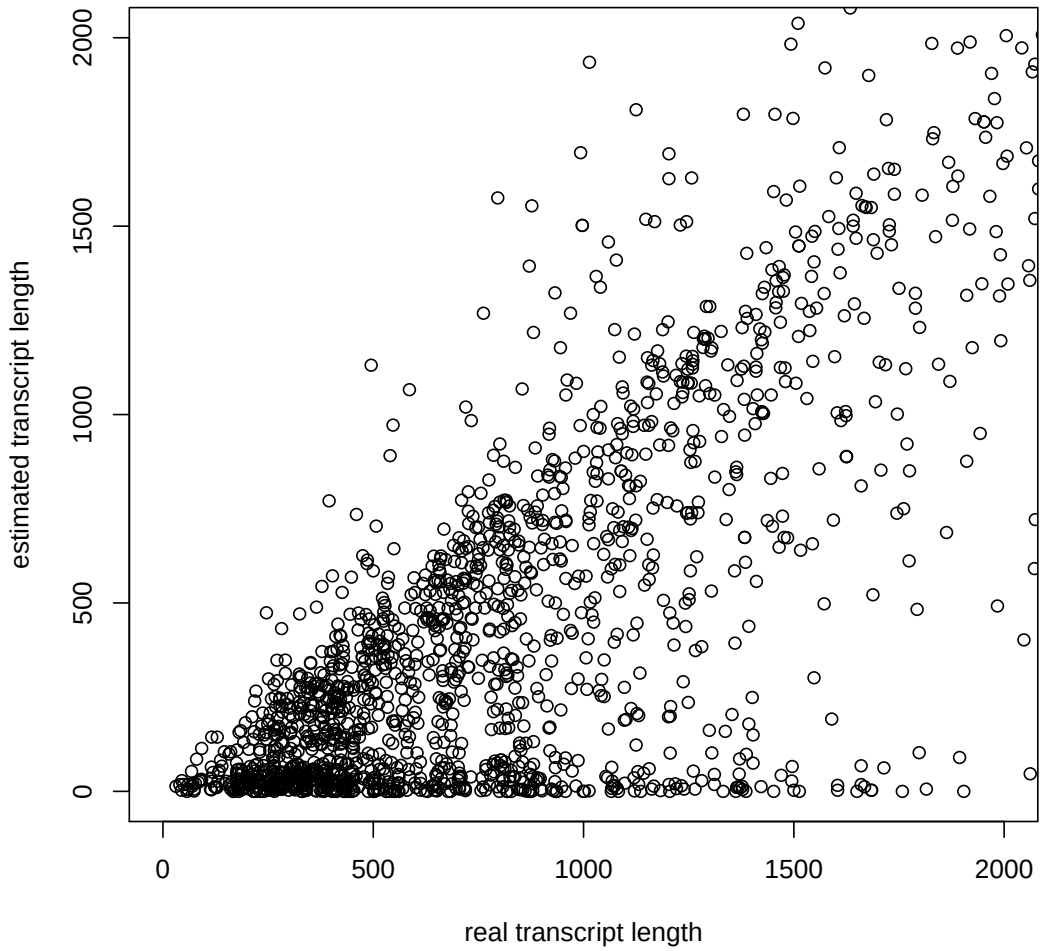
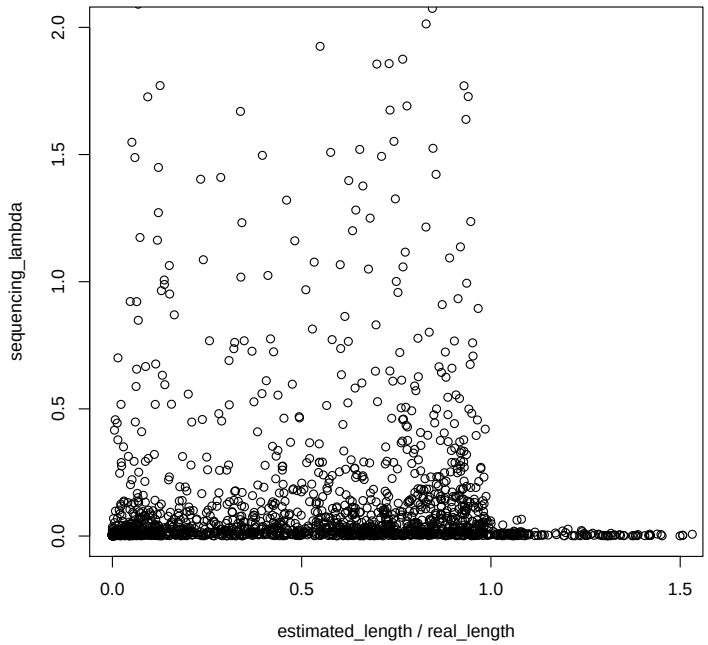


图 2.1　实际的转录本长度和估计的转录本长度的比较: 每一个点代表一个转录本。



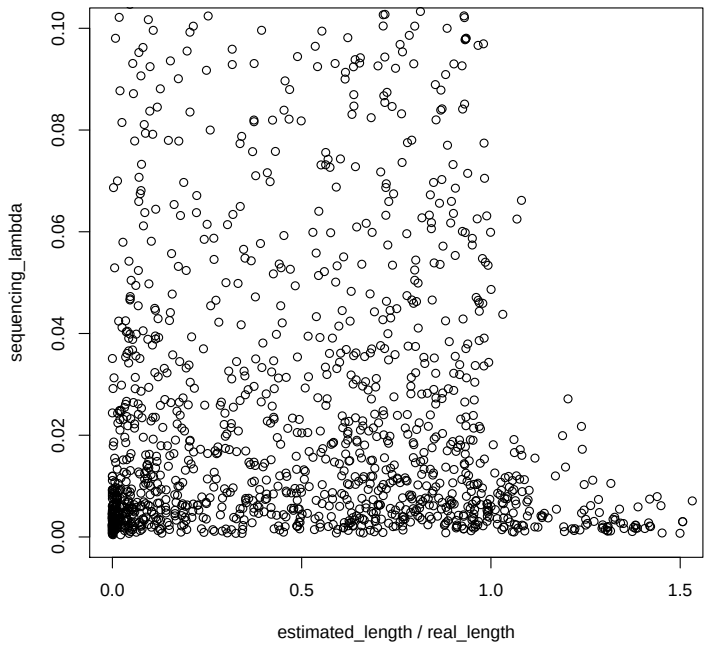


图 2.2　估计的转录本长度和真实的转录本长度之比 ($\frac{\text{估计的转录本长度}}{\text{真实的转录本长度}}$) 和测序 λ 之间的关系: 每一个点代表一个转录本。从该图中无法看出二者之间有明显的关系。

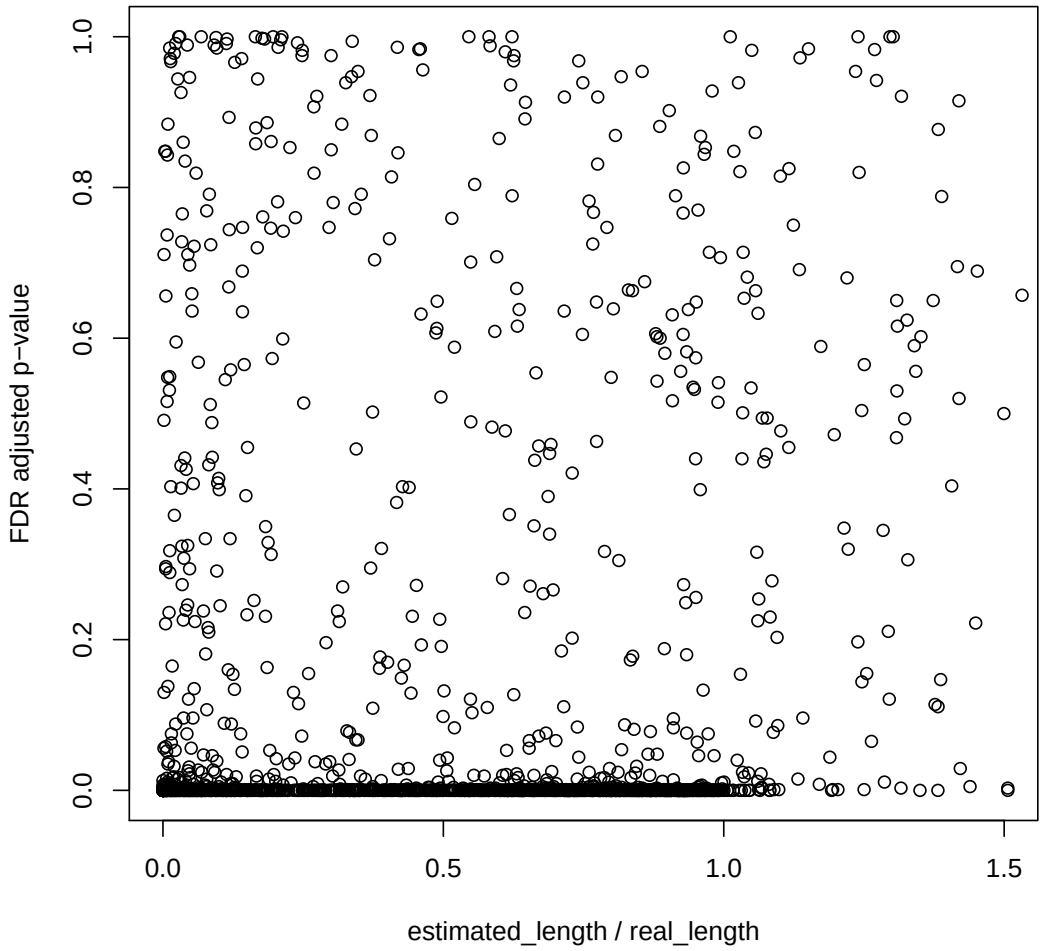


图 2.3　估计的转录本长度和真实的转录本长度之比 ($\frac{\text{估计的转录本长度}}{\text{真实的转录本长度}}$) 和和第 3 章中计算得的检验均匀分布经过 FDR 调整的 p-value 之间的关系: 每一个点代表一个转录本。从该图中无法看出二者之间有明显的关系。

第 3 章　RNA-Seq 数据分布的不均匀性

3.1　介绍

根据 2.3 中对于实际 RNA-Seq 数据的分析, 我们发现 RNA-Seq 数据并不符合理想的均匀分布。已有的工作[69,91–95] 也对 RNA-Seq 实验数据分布的不均匀性进行了研究和讨论 (图 3.1)。

3.2　数据分析

我们对来自 ENCODE[5] 的 RNA-Seq 数据 ① 进行了分析, 试图发现 RNA-Seq 数据读段分布的一些规律。在分析中, 我们只对只有一个外显子的基因进行了分析。

3.2.1　方法

3.2.1.1　检验均匀分布的 p-value 的计算

给定一个转录本, 在计算检验均匀分布的 p-value 时, 我们将转录本划分成若干段, 在统计出每一段中的读段起始书目后, 使用 χ^2 统计量, 通过 parametric bootrstrap[96] 计算 p-value。

对于一个转录本, 其长度为 l bp, 读段的长度为 r bp, 每个位置 $p = 1, 2, \ldots, l-r+1$ 上的读段个数为 n_p, 该转录本上所有的读段个数 $n = \sum_{p=1}^{l-r+1} n_p$, 我们所用的统计量为

$$\sum_{p=1}^{l-r+1} \frac{(n_p - \frac{n}{l-r+1})^2}{\frac{n}{l-r+1}} \tag{3-1}$$

① http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/wgEncodeCaltechRnaSeqGm12878R1x75dAlignsRep2V2.bam

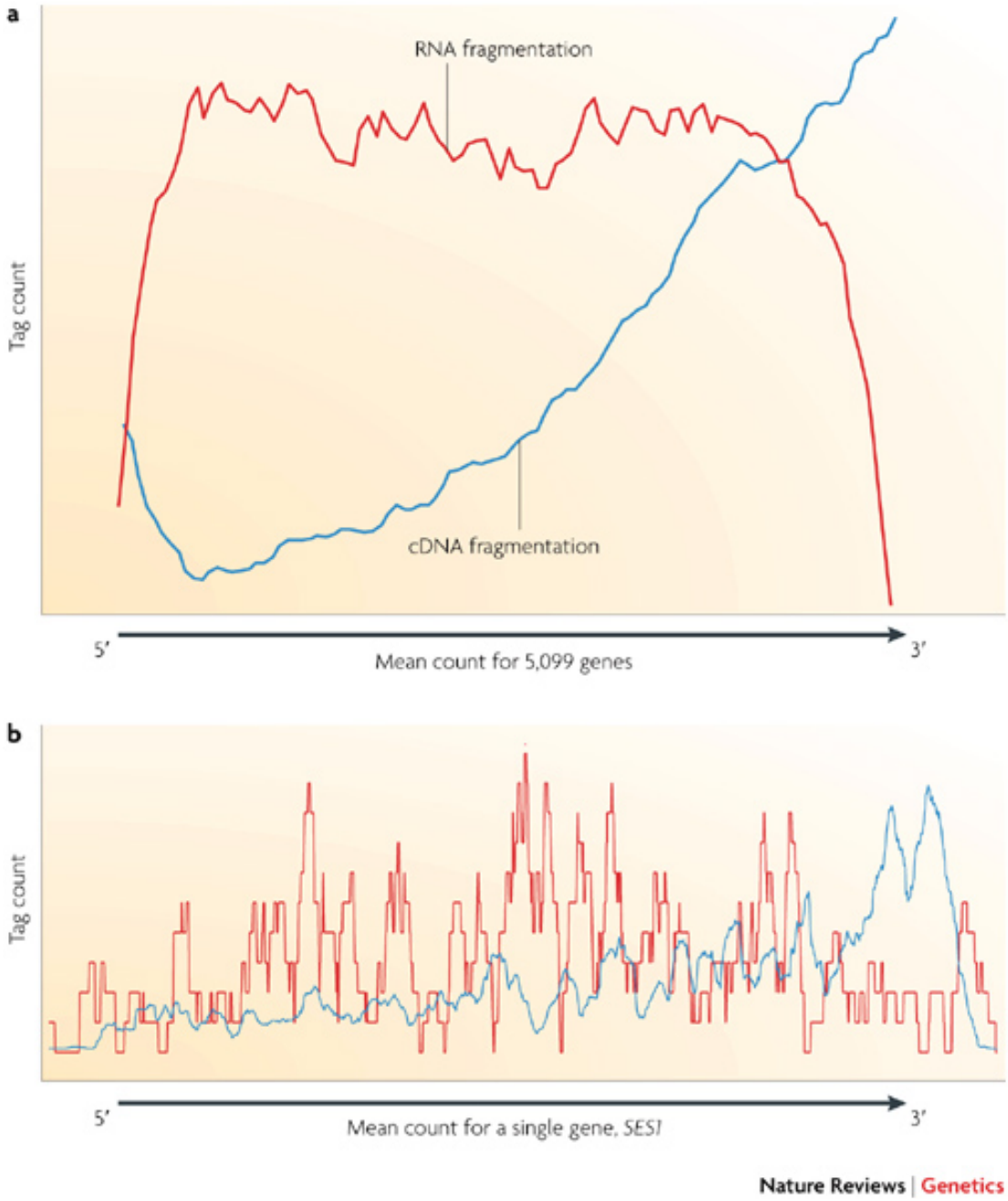


图 3.1　实际 RNA-Seq 数据中读段的位置分布[1]: 在 RNA-Seq 实验中, 使用 cDNA 打断方式时读段的位置更靠近 3’ 端, 使用 RNA 打断方式时读段的位置更多地集中在转录本的中间。

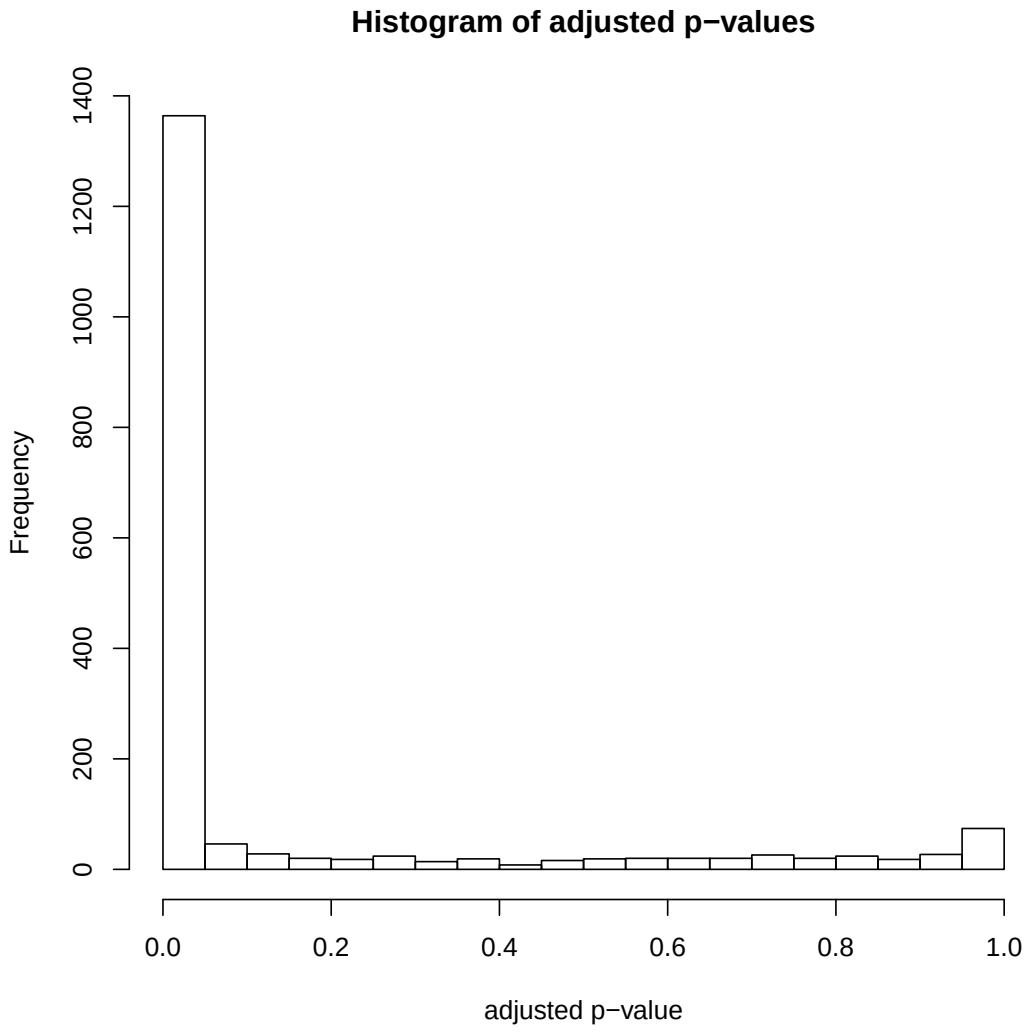


图 3.2　FDR 调整[97] 后的检验均匀分布的 p-value: 对于大部分的转录本, 读段的分布都是较不均匀的。

3.2.2　结果

3.2.2.1　检验均匀分布的 p-value

我们对于只有一个外显子的基因中读段的分布计算了检验均匀分布的 p-value, 做过 FDR 调整[97] 的 p-value 的柱状图如图 3.2 所示。从图 3.2 中我们可以看出, 对于我们分析中考虑的大部分基因, 读段分布都是较不均匀的

第 4 章　使用真核生物 RNA-Seq 数据通过最大似然辨识基因的剪切异构体是一个 NP 难问题

4.1　问题描述

在真核生物 RNA-Seq 数据处理过程当中, 重要的一步是通过 RNA-Seq 数据识基因的不同的剪切异构体。由于在真核生物的转录过程当中会有选择性剪切的现象发生, 一个基因可能转录出多个剪切异构体。在这种情况下, 我们需要通过 RNA-Seq 数据辨识出基因的不同的剪切异构体。

在 RNA-Seq 实验中得到的读段数据在比对到参考基因组序列上如图 4.1 所示. 对于 RNA-Seq 数据, 我们可以通过序列拼装工具 (例如 Trinity[61] 等) 进行拼装。得到每一个基因的外显子的序列以及这些外显子之间是如何连接在一起的关系, 从而得到每一个基因的剪切图[78]。在有基因的参考序列以及基因注释的情况下, 也可以直接通过基因注释获得每一个基因的剪切图。对于图 4.1 中所示的基因和 RNA-Seq 数据, 通过 Trinity 拼装获得的剪切图如图 4.2 所示。

4.1.1　模型

为了能够通过 RNA-Seq 数据辨识每个基因的剪切异构体, 我们在这里介绍一个 Poisson 模型用来描述 RNA-Seq 数据[82]。该模型与 1.3.1 中所介绍的多项式模型在数学上是等价的[68]。在这里采用 RNA-Seq 数据的 Poisson 模型是为了简化问题的描述。

另外, 为了叙述的方便, 我们在这里纸考虑辨识一个基因的剪切异构体。

对于真核生物中的一个基因, 根据剪切图的定义, 我们将其表示为一个有向无环图 $G = (V, E)$, 其中 G 中的一个节点代表一个外显子, 节点 v_i 到 v_j 有一条边当且仅当在这个基因当中有一个剪切异构体同时包含 v_i 和 v_j 所代表的外显子, 同时在这个剪切异构体中 v_i 和 v_j 所代表的外显子是相邻的, 并且 v_i 代表的外显子在 v_j 代表的外显子的 5' 上端。

该基因的一个剪切异构体是图 G 中的一个路径 $v_1 \to v_2 \to \ldots \to v_k$, 其中 v_i, $i = 1, 2, \ldots, k$ 是构成该剪切异构体的外显子。我们用 I 表示这个基因的剪切异

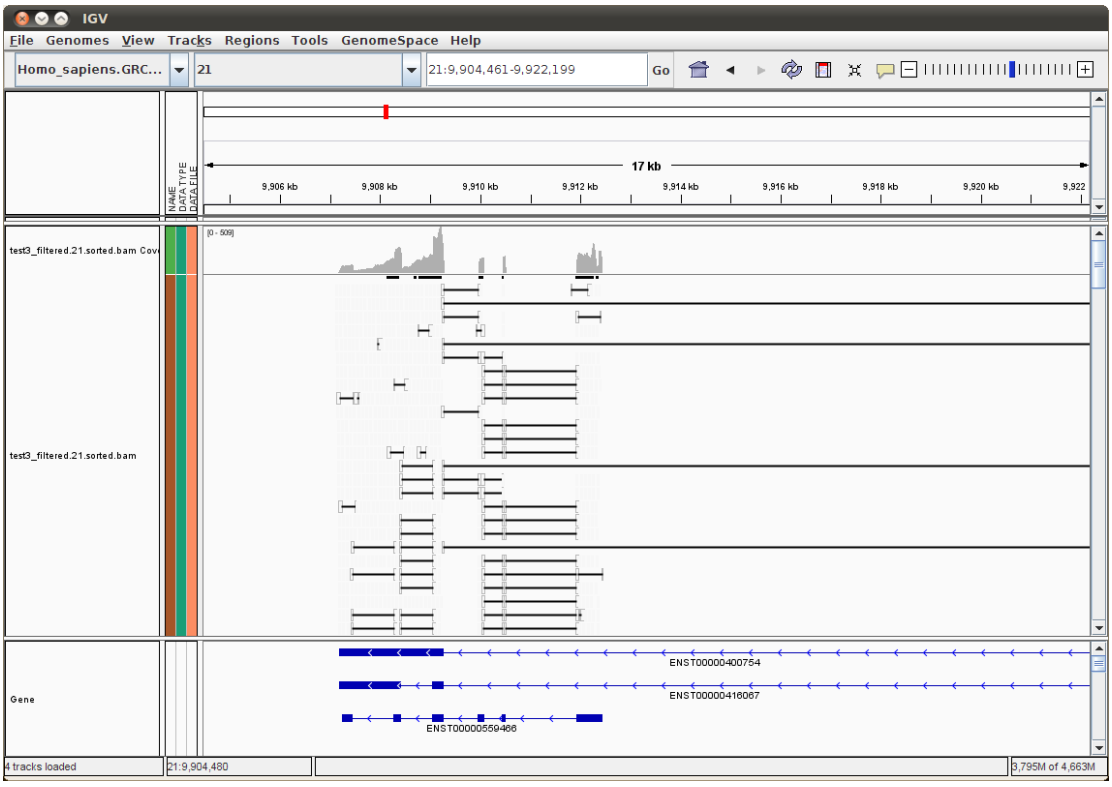


图 4.1　一个基因的比对到基因组参考序列上的 RNA-Seq 数据: 根据每一个读段比对到基因组参考序列上的位置可以确定其所对应的剪切异构体。

构体构成的集合, $\lambda_i \geq 0, i \in I$ 是剪切异构体 i 的表达量。

在这里, 我们假设所有的 RNA-Seq 数据均是单端读段, 且长度均为 k。 (对于长度不均的读段或者双端测序的读段我们可以采用与这里类似的方法进行处理, 此处不再赘述, 感兴趣的读者可以参看 [43]。) 一个读段对应到剪切图中时, 该读段对应到原转录本上的每一个外显子在剪切图中所对应的节点, 同时在读段经过多个外显子时该读段也会经过剪切图中所对应的边。为了描述的方便, 我们把每一个读段对应在原转录本上的 5' 端的位置称之为该读段的起始位置, 对应在原转录本上的 3' 端的位置称之为该读段的终止位置。

我们假设每一个读段可以在该基因中的位置是已知的并且是唯一的 (对于基因序列的分析表明这样一个假设是合理的[70])。

同时, 我们把该基因在 RNA-Seq 实验中所有对应的读段构成的集合记为 R。

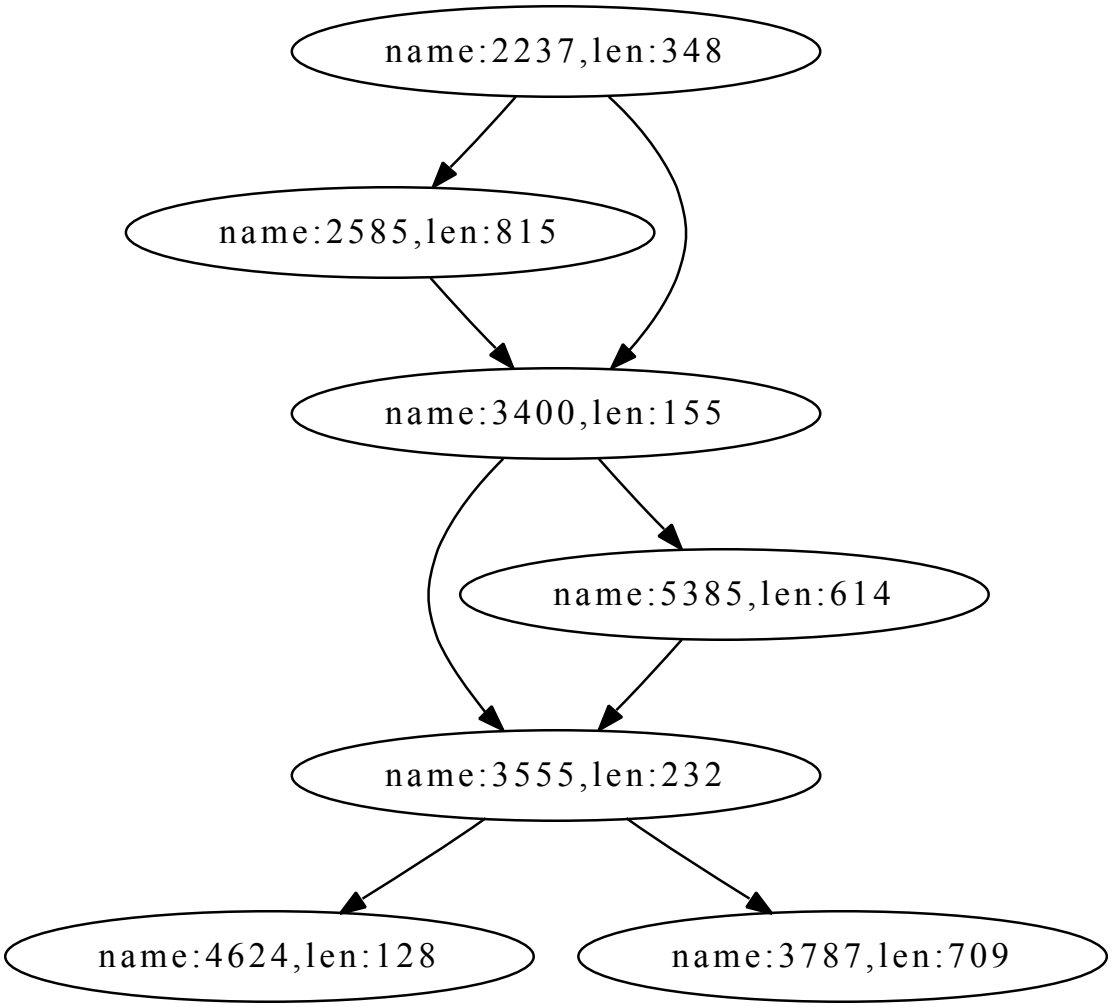


图 4.2　图 4.1 中所示的基因和 RNA-Seq 数据通过 Trinity[61] 拼装得到的剪切图: 该基因的一个剪切异构体是这个图中的一条路径。

在用来描述 RNA-Seq 数据的 Poisson 模型中, 我们假设对于基因中的每一个位置 p, 在这个位置起始的读段的个数满足参数为

$$\lambda_p = \sum_{\text{剪切异构体 } i \in I \text{ 经过位置 } p} a_{pi}\lambda_i \tag{4-1}$$

的 Poisson 分布。 (此处可以参考图 4.1 进行理解。) 这里 $a_{pi} \geq 0$ 是与目前该位置 p 有关的一个采样率[98]。我们假设在给定基因中的一个位置 p 之后, 以及一个剪切异构体 $i \in I$, a_{pi} 是已知的。对于理想的情况我们认为对于所有的基因中的位置 p 和剪切异构体 $i \in I$, 我们均有 $a_{pi} = 1$。通过设定不同的位置和不同的剪切异构体 $i \in I$ 的 a_{pi}, 我们可以对 RNA-Seq 实验中存在的不均匀性进行建模。为了方便下面的陈述, 我们记基因中所有可以有读段起始的位置的集合记作 P。

(例如, 对于一个只有一个外显子的基因, 其长度为 L, P 由该基因上的前 $L-k+1$ 个位置组成。)

同时, 我们假设对于该基因中不同的位置, 在这些位置起始的读段的个数的分布是独立的。

根据上面的描述, 我们得出观察到 RNA-Seq 读段的似然函数

$$L(R) = \prod_{p \in P} \frac{{\lambda_p}^{N(p)} e^{-\lambda_p}}{N(p)!} \tag{4-2}$$

其中 $N(p)$ 是在位置 $p \in P$ 所管查到的在该位置起始的读段的数目, λ_p 如式 (4-1) 中所定义。注意到 $|R| = \sum_{p \in P} N_p$。

根据式 (4-2) 中所给出来的似然函数, 我们可以通过最大化该似然函数确定该基因的剪切异构体的集合 I, 辨识该基因的剪切异构体[71]。但是由于一个基因可能的剪切异构体的个数是随着该基因中外显子的个数增长按指数规律增长的, 所以我们可以使用一个带惩罚项的似然函数进行最大化

$$\begin{aligned} \tilde{L}(R) &= e^{-\Phi} L(R) \\ &= e^{-\Phi} \prod_{p \in P} \frac{{\lambda_p}^{N(p)} e^{-\lambda_p}}{N(p)!} \end{aligned} \tag{4-3}$$

其中 $\Phi \geq 0$ 是一个惩罚函数。

对于这里的乘法函数 Φ 我们可以选取:

- L_0 正则化:

  $$\Phi = \Lambda \sum_{i \in I, \lambda_i \neq 0} 1$$

  其中 $\Lambda \geq 0$ 是一个常数。

  当我们取

  $$\Lambda = \frac{\log |R|}{2}$$

  此时相当于使用 BIC (Bayesian information criterion)[79]。

- Lasso[81]:

  $$\Phi = \Lambda \sum_{i \in I} |\lambda_i|$$

  其中 $\Lambda \geq 0$ 是一个常数。

该方法是 NSMAP[72] 所使用的方法。

- L_q $(0 < q < 1)$ 正则化:

$$\Phi = \Lambda \sum_{i \in I} |\lambda_i|^q$$

其中 $\Lambda \geq 0, 0 < q < 1$ 为两个常数。

该方法是 Montebello[77] 所使用的方法。

4.1.2 M-ISOFORM 判定问题

为了说明通过真核生物 RNA-Seq 数据估计基因的转录组的问题, 也即辨识基因的剪切异构体的问题, 我们在这里定义一个判定问题 M-ISOFORM。

根据 [99] 我们可以知道, 当我们使用 L_0 正则化作为式 (4-3) 中的惩罚函数最大化带惩罚项的似然函数时, 此时等价于

$$\begin{aligned} \max L(R) &= \prod_{p \in P} \frac{{\lambda_p}^{N(p)} e^{-\lambda_p}}{N(p)!} \\ \text{s.t.} \sum_{i \in I, \lambda_i > 0} 1 &\neq t \end{aligned} \tag{4-4}$$

其中 $t \geq 0$ 是一个常数。

同时我们对剪切异构体的集合 I 也有要求: 对于每一个属于 R 的读段 r, 我们均要求至少存在一个剪切异构体 $i \in I$ 使得读段 r 是可以来自剪切异构体 i 的。这对于式 4-4 中的优化问题又增加了一项组合约束。

给出了这些观察之后, 我们可以将此处通过 RNA-Seq 数据估计转录组中转录本的问题, 也即基因的剪切异构体的辨识的问题, 转换为一个判定问题:

给定一个阈值, 能不能找到一个剪切异构体的集合以及这些转录本的表达量使得似然函数的值大于该阈值?

下面我们采用前面 4.1.1 中所使用的符号来描述判定问题 M-ISOFORM。

M-ISOFORM：
实例: 一个基因的剪切图 G, 以及对应到该剪切图中的读段集合 R (这里我们保证剪切图 G 中的所有定点所对应的外显子和边所表示的外显子的连接关系均能由 R 中的读段解释)。常数 $q \in \mathbb{R}$, $m \in \mathbb{Z}^+$。
问题: 是否存在该基因的一个剪切异构体的集合 I, $|I| \leq m$, 以及这些转录本的表达量 $\theta_i > 0$ 使得式 (4-2) 中的似然函数 $L(R) \geq q$, 并且对每一个读段 $r \in R$ 均存在至少一个该基因的剪切异构体 $i \in I$ 让读段 r 来自剪切异构体 i?

在实数 RAM 模型[100] 下, 我们下面的定理说明了 M-ISOFORM 是一个 NP 完全问题。

定理 4： M-ISOFORM 是一个 NP 完全问题。

4.2 证明

下面我们给出定理 4, 说明 M-ISOFORM 是一个 NP 完全问题。定理 4 的证明主要受到了 Tomescu et al. 在 [101] 的工作的启发, 将通过 RNA-Seq 数据辨识基因的剪切异构体的问题看作是一个网络流问题。证明中我们采用了一个类似 Vatinlen et al. 在 [102] 中所使用的方法, 将 3-PARTITION[103,104] 规约到 M-ISOFORM。

证明 首先注意到在实数 RAM 模型[100] 下, 在给定该基因的一个剪切异构体的集合 I 的情况下, 及个转录本的表达量 θ_i, $i \in I$, 我们可以在多项式时间内验证是否式 (4-2) 中的似然函数 $L(R) \geq q$, 并且对所有的读段 $r \in R$, 是否存在一个剪切异构体 $i \in I$ 让读段 r 可以来自剪切异构体 i。所以我们可以看出来 M-ISOFORM $\in$ NP。

下面处我们介绍判定问题 3-PARTITION[103,104], 并给出一个方法在多项式时间内规将其约到 M-ISOFORM。

3-PARTITION：
实例: 一个包含 $3w$ 个元素的集合 X, 一个界 $Y \in \mathbb{Z}^+$, 和一个量 $u(x) \in \mathbb{Z}^+$ 使得对每个 $x \in X$ 都有 $\frac{Y}{4} < u(x) < \frac{Y}{2}$, 并且 $\sum_{x \in X} u(x) = wY$。
问题: X 能否被划分为 w 个不相交的集合 $X_1, X_2, \ldots, X_w$ 使得, 对所有的 $1 \le i \le w$, $\sum_{x \in X_i} u(x) = Y$ (注意到每个 X_i 必须包含恰好 3 个来自 X 的元素)?

根据 Garey 和 Johnson 在 [103] 中以及 Garey 和 Johnson 在 [104] 的工作, 我们可以知道 3-PARTITION 是强 NP 完全的, 即 3-PARTITION 的输入用 unary 表示时 3-PARTITON 仍然是 NP 完全的。

我们下面将用 unary 输入的一个 3-PARTITION 实例在多项式时间内规约到一个 M-ISOFORM 实例。

给定一个 unary 输入的 3-PARTITION 的实例, 我们构造一个基因如图 4.3 所示, 这里每一个外显子的长度均为 $k-1$ bp ($k \ge 2$ 是一个固定的常数)。同时我们构造一个读段的集合 R, 其中共包含 $4wY(k-1)$ 个长度为 k bp 的读段。在这里为了方便描述, 我们不妨设 $X = \{1, 2, \ldots, 3w\}$。

对于外显子 A, 在其中的 $k-1$ 个位置中的每一个位置共有 wY 个读段将该位置作为起始位置, 其中对于 $1 \le i \le 3w$, 这 wY 个读段中有 $u(i)$ 个读段是从外显子 A 连接到到外显子 B_i 的。

对于 $1 \le i \le 3w$, 外显子 B_i 的 $k-1$ 个位置的每个位置均有 $u(i)$ 个读段将该位置作为起始位置, 连接到外显子 C 中。

对于外显子 C, 其中的 $k-1$ 个位置的每个位置均有 wY 个读段将该位置作为起始位置, 其中对于 $1 \le i \le w$, 这 wY 个读段中有 Y 个读段是从外显子 C 连接到外显子 D_i 的。

对于 $1 \le i \le w$, 外显子 D_i 的 $k-1$ 个位置的每个位置均有 Y 个读段将该位置作为起始位置, 连接到外显子 E 中。

注意到, 由于所有的外显子的长度均为 $k-1$ bp, 而读段的长度为 k bp, 所以 D 的所有位置中都没有是任何读段的起始位置。

同时, 我们对于该 3-PARTITION 实例, 对于图 4.3 中的所有的读段起始位置, 采样率均为 1。

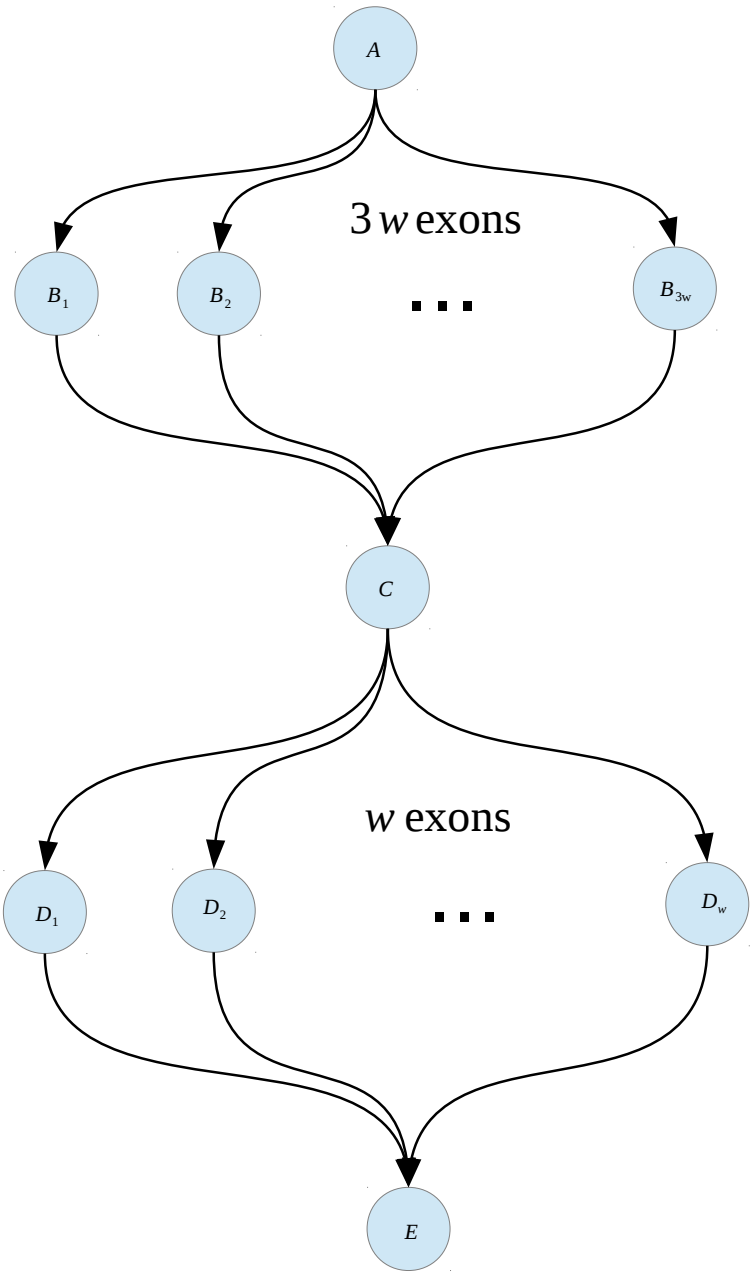


图 4.3　一个 3-PARTITION 实例对应的 M-ISOFORM 实例的构造的基因: 这里每一个外显子的长度均为 $k-1$ bp ($k \geq 2$ 是一个固定的常数)。

在此我们有了对应该 3-PARTITION 的实例的一个 M-ISOFORM 实例, 并且我们有一个如下的问题:

是否存在一个剪切异构体的集合 I, $|I| \leq 3w$, 以及这些转录本的表达量 $\lambda_i \geq 0$, $i \in I$, 使得式 (4-2) 中的似然函数

$$\begin{aligned} L(R) &= \prod_{p \in P} \frac{{\lambda_p}^{N(p)} e^{-\lambda_p}}{N(p)!} \\ &\geq \prod_{p \in P} \frac{N(p)^{N(p)} e^{-N(p)}}{N(p)!} \end{aligned} \tag{4-5}$$

(这里所使用的阈值是 $\prod_{p \in P} \frac{N(p)^{N(p)} e^{-N(p)}}{N(p)!}$。)

注意到, 前面对于给定的 3-PARTITION 实例构造这一个 M-ISOFORM 实例是在多项式时间之内完成的。下面我们说明给出的 3-PARTITION 有解当且仅当我们构造的着一个 M-ISOFORM 实例是有解的。

首先我们容易看出来, 对于下面的函数

$$f(x) = \begin{cases} 0 & x < 0 \\ x^N e^{-x} & x \geq 0 \end{cases}$$

其中 $N \in \mathbb{Z}^+$。该函数取得最大值当且仅当 $x = N$。

所以, 式 (4-5) 成立当且仅当

$$\lambda_p = N(p) \tag{4-6}$$

对所有的 $p \in P$ 成立。

我们首先说明在这里的 3-PARTITION 的实例有解时, 我们所构造的 M-ISOFORM 的实例也有解。当该 3-PARTITION 的实例有解时, 假设 $i \in X$ 所属于的集合是 X_j, 则我们给剪切异构体的集合 I 中添加剪切异构体

$$A \to B_i \to C \to D_j \to E$$

同时将该剪切异构体的表达量定为

$$u(i)$$

此时 I 中恰好有 $3w$ 个剪切异构体, 并且根据式 (4-1) 我们知道式 (4-6) 此时也是满足的。另外这里容易看出来对于每个读段 $r \in R$, 均存在一个剪切异构体 $i \in I$ 让读段 r 是可以来自剪切异构体 i 的。所以我们知道对应的我们构造的 M-ISOFORM 实例也有解。

下面我们说明当我们所构造的 M-ISOFORM 的实例有解时, 这里的 3-PARTITION 问题也是有解的。我们容易注意到, 若存在满足要求的剪切异构体的集合 I 时, 其中一定包含恰好 $3w$ 个剪切异构体, 因为图 4.3 所示的基因结构意味着至少有 $3w$ 个剪切异构体, 这样才能保证对于每个读段 $r \in R$ 至少存在一个剪切异构体 $i \in I$ 让读段 r 是可以来自剪切异构体 i 的。同时根据图 4.3 所示的基因结构, 我们可以知道每个剪切异构体必须是

$$A \to B_i \to C \to D_j \to E$$

对于某 $1 \le i \le 3w$ 和 $1 \le j \le w$, 否则为了保证保证对于每个读段 $r \in R$ 至少存在一个剪切异构体 $i \in I$ 让读段 r 是可以来自剪切异构体 i 的, 我们需要 I 中有多于 $3w$ 个剪切异构体。再注意到, 此时由于还有式 (4-1) 和式 (4-6) 的约束, 对于前面的剪切异构体 $A \to B_i \to C \to D_j \to E$, 它的表达量应该是

$$u(i)$$

同时, 根据式 (4-1) 和式 (4-6) 的约束, 以及 $\frac{Y}{4} < u(x) < \frac{Y}{2}$, 我们可以知道对于每个 $q \le j \le w$, 外显子 D_j 中恰好有 3 个剪切异构体通过, 并且这 3 个剪切异构体的表达量之和是 Y。所以在这个时候我们根据每个外显子 D_j, $1 \le j \le w$ 中所经过的剪切异构体的情况, 可以把 X 分为不相交的 w 个集合 $X_1, X_2, \ldots, X_w$ 使得对所有的 $1 \le i \le w$, $\sum_{x \in X_i} u(x) = Y$, 并且每个 X_j, $1 \le j \le w$ 中恰好有 X 中的 3 个元素。

所以该定理得证, 我们知道 M-ISOFORM 是一个 NP 完全问题。

□

第 5 章　讨论

前面我们分别描述了现有的 RNA-Seq 数据分析方法 (第 1 章)。此外, 我们提出了一个通过 RNA-Seq 数据估计转录本的长度的方法 (第 2 章)。同时, 我们初步分析了 RNA-Seq 数据分布的不均匀性 (第 3 章)。另外, 我们证明了通过真核生物 RNA-Seq 数据辨识基因的转录组问题是一个 NP 难问题 (第 4 章)。在这一章中我们对前面的内容和 RNA-Seq 数据分析方法的未来的发展方向进行讨论, 并且对本文进行总结。

5.1　通过 RNA-Seq 数据估计转录本的长度

在第 2 章的分析中我们只对单端等长读段测序数据进行了考虑, 对于双端数据或者长度不均的读段均未作具体分析。在实际应用中可以只对双端数据考虑等长的在原转录本的 5' (或者 3') 端, 或者长度不均的读段考虑等长的在原转录本的 5' (或者 3') 端, 从而可以用第 2 章中的方法进行分析。

此外, 第 2 章中所介绍的方法并没有考虑 RNA-Seq 数据中读段分布受转录本序列以及转录本长度等因素造成的分布不均匀性。另外, 第 2 章中所介绍的方法也未考虑读段分布的非独立性。这些问题需要进一步的研究来解决。

5.2　通过真核生物 RNA-Seq 数据估计基因的转录组

根据第 4 章中的讨论, 我们知道通过目前的 RNA-Seq 技术从 RNA-Seq 数据中直接估计转录组中转录本的组成在计算上是比较困难的。同样的问题在通过测序数据估计单体型 (haplotype)[105,106] 时也是同样存在的[107]。

随着测序技术的发展, RNA-PET[84] 技术可以对直接用来确定转录本的 5' 端和 3' 端, 另外更长的测序读段 (例如 PacBio 研发的单分子测序技术[108]) 可以帮助我们在测序数据得到后不需要使用类似第 4 章给出的复杂的方法就能简单地直接从数据中辨识基因的剪切异构体。

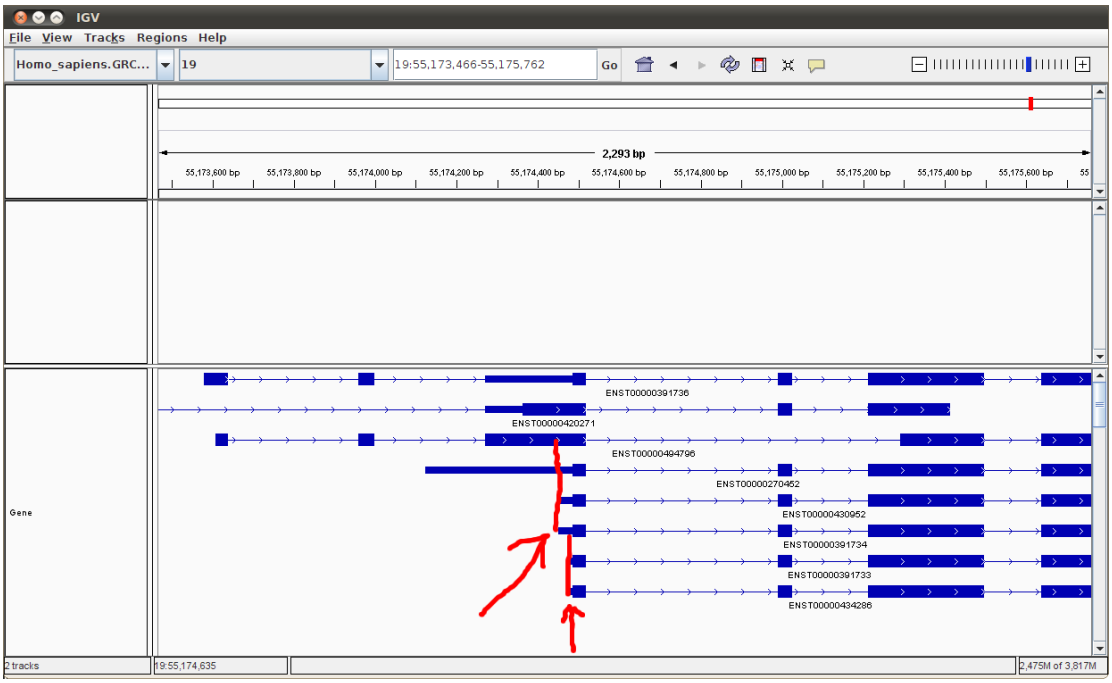


图 5.1　人的基因组中出现的一个转录本的一端位于另一个转录本的中间的现象: 图中可见若干个基因的 5' 端位于另外的基因的中间。

5.3　现有 RNA-Seq 数据处理方法中存在的一些问题

目前已有的用于 RNA-Seq 数据辨识基因的剪切异构体和估计表达量的方法 (1.3) 仍然存在不足。对于真核生物 RNA-Seq 数据, 辨识基因的剪切异构体复杂度高。此外, 当有两个转录本其中一个转录本的一端位于另一个转录本的中间时 (图 5.1), 目前没有有效的方法通过 RNA-Seq 数据对这两个转录本进行区分。另外, 对于原核生物, 尤其是原核生物中的操纵子, 的分析方法仍然没有较规范的分析方法[109] (图 5.2)。同时, 估计转录本表达量时仍无系统的方法处理 RNA-Seq 数据中读段分布的不均性, 以及转录本序列的组成带来的误差[91,110]。

5.4　未来 RNA-Seq 数据的定量分析方法的工作

5.4.1　RNA-PET 测序技术

Fullwood et al. 在 [84] 中介绍了 RNA-PET 技术。通过 RNA-PET 技术我们可以对转录本的 5' 端和 3' 端的序列进行测序。这样我们就可以较为准确地确定转录本的 5' 端和 3' 端。为了处理 RNA-PET 数据, 我们需要实开发新的生物

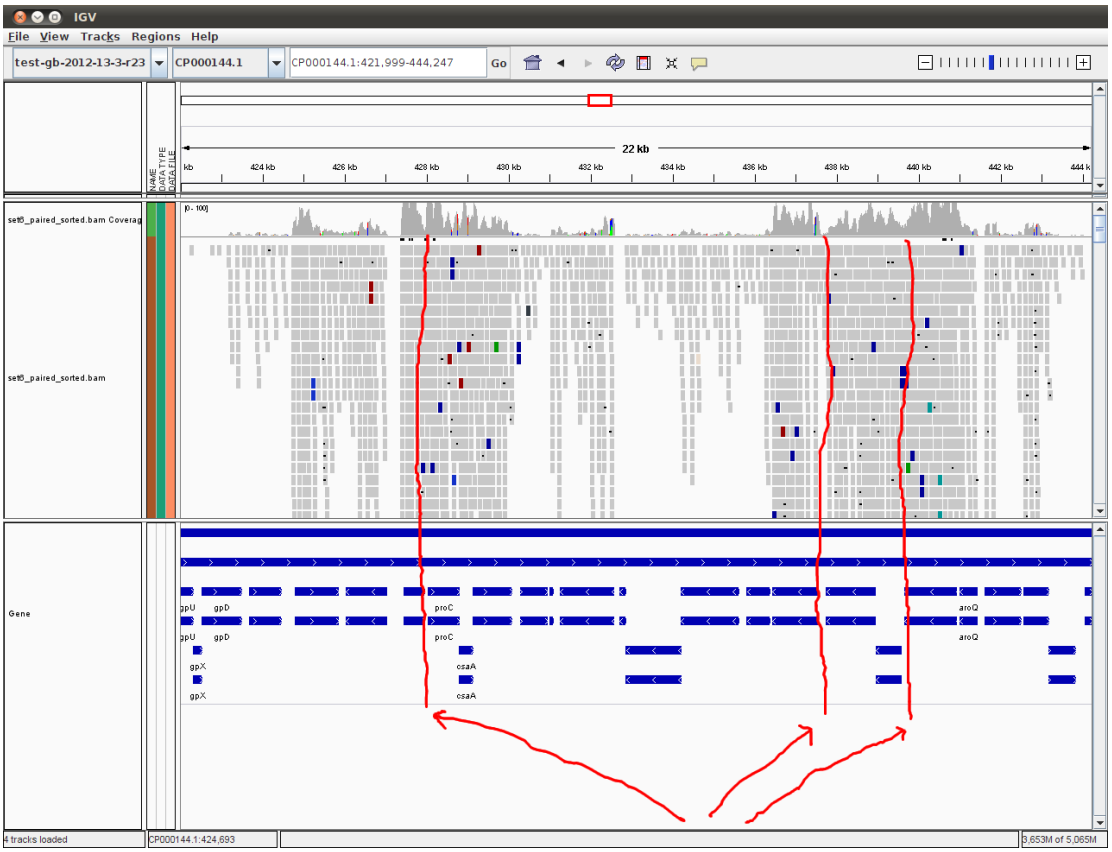


图 5.2　Giannoukos et al. 在 [111] 中提供的细菌的 RNA-Seq 数据: 其中可以看见多个基因构成的操纵子

信息工具能够有效地将 RNA-PET 数据中的读段比对到参考基因序列上。此外, 我们还需要开发新的生物信息工具能够有效地将 RNA-PET 数据中的读段融合到序列拼装中, 提高转录组序列拼装的质量。

5.4.2　更长的测序读段

目前大部分 RNA-Seq 都是使用 Illumina 的平台进行测序的, 这使得大部分的 RNA-Seq 数据都是长度小于 100 bp 的读段。虽然 454 的平台能够测量长度为数百 bp 的读段, 但是 454 的平台的测序通量与 Illumina 的平台的测序通量相比较低, 不适合通过测序的放啊对转录组进行研究。

新的测序技术, 例如 PacBio 开发的单分子测序技术, 能够提供更长的测序读段。但是由于其中测序错误较多, 并且测序通量和 Illumina 的平台的测序通量相比较低, 所以目前的做法是将 PacBio 的单分子测序技术得到的少量的长读段和 Illumina 的平台上通过高通量测序得到的大量的短读段结合在一起进行分

析[108]。

我们需要开发出新的生物信息工具能够针对新的测序技术提供的更长的读段进行处理, 其中主要包括针对新的测序技术的序列比对工具和针对新的测序技术的序列拼装工具

5.4.3 原核生物

原核生物的 RNA-Seq 数据分析对于对于我们研究微生物中的基因表达是十分有意义的。此外, 近几年来兴起的宏转录组 (metatranscriptome) 的研究也帮助了我们对于微生物群落有了更为深入的了解[112–118]。

与真核生物不同, 原核生物中多个基因会被同时翻译到一个转录本中形成一个操纵子 (图 5.3)。最近几年的研究表明[120,121], 原核生物在不同的环境条件下在基因组的同一段区域可能产生不同的操纵子。通过 RNA-Seq 数据辨识原核生物操纵子也将可能成为原核生物 RNA-Seq 数据分析中重要的一步。

另外, 在原核生物的转录组中有大量的 rRNA (ribosomal RNA)。在 RNA-Seq 实验中, rRNA 和其他种类的 RNA (例如 mRNA) 相比起来 rRNA 含量可能超过 90%[111]。此时, rRNA 的含量会对于研究其他种类的 RNA 产生较大的负面影响。目前 Giannoukos et al. 在 [111] 中使用了一种新的实验技术, 可以在测序前减少样本中 rRNA 的含量。

5.5 总结

本文的主要贡献在于本文提出了一个能够在没有基因注释信息的情况下通过 RNA-Seq 数据估计转录本长度的方法, 同时证明了使用真核生物 RNA-Seq 数据通过最大似然辨识基因的剪切异构体是一个 NP 难问题。我们期待测序技术的发展能够缓解 RNA-Seq 数据中读段分布不均匀的现象。此外, 我们期待新的测序技术的的发展能够降低测序数据计算处理的要求, 同时方便我们通过 RNA-Seq 技术对生物问题进行研究。

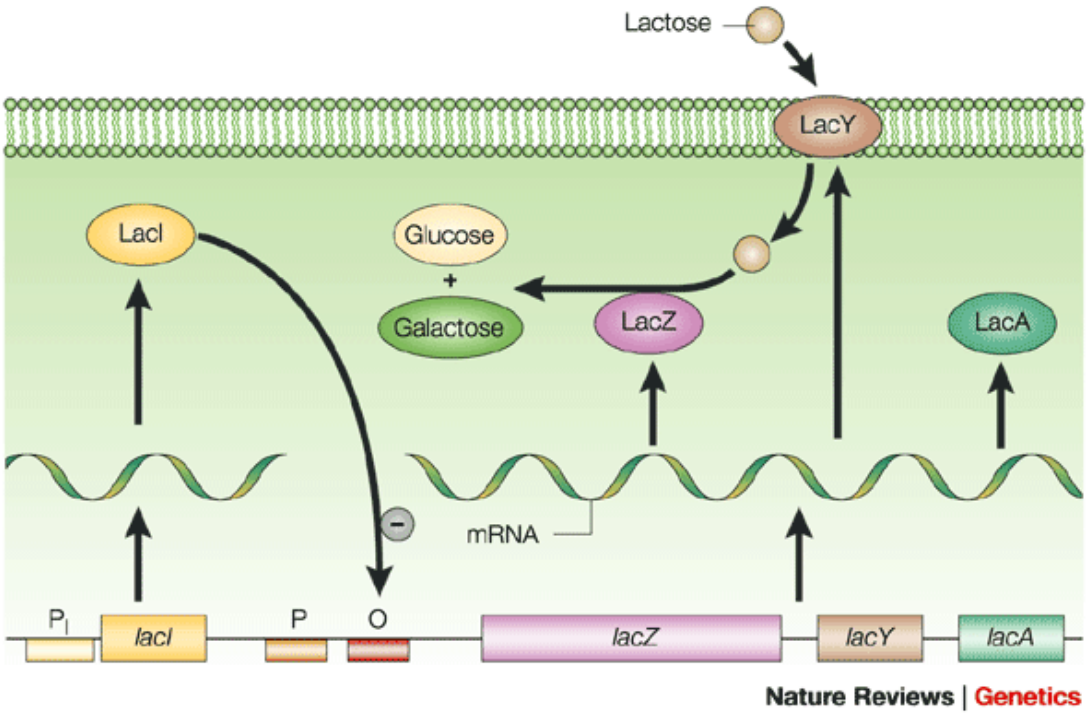


图 5.3　*E. coli* 中的乳糖操纵子[122]: 基因 *lacZ*, *lacY*, 和 *lacA* 被转录到同一个 mRNA 中

插图索引

图 1.1 一般的 RNA-Seq 实验流程[1] 2
图 1.2 一般的通过 TopHat[42] 和 Cufflinks[43] 分析 RNA-Seq 数据的流程[44] .. 3
图 1.3 Trinity 拼装策略[61] 5
图 1.4 Cufflinks 对转录本的表达量的估计[43] 6
图 1.5 Cufflinks 序列拼装策略和对同一个基因的剪切异构体的辨识[43] 7
图 2.1 实际的转录本长度和估计的转录本长度的比较 20
图 2.2 估计的转录本长度和真实的转录本长度之比和测序 λ 之间的关系 ... 21
图 2.3 估计的转录本长度和真实的转录本长度之比和和第 3 章中计算得的检验均匀分布经过 FDR 调整的 p-value 之间的关系 22
图 3.1 实际 RNA-Seq 数据中读段的位置分布[1] 24
图 3.2 FDR 调整[97] 后的检验均匀分布的 p-value 25
图 4.1 一个基因的比对到基因组参考序列上的 RNA-Seq 数据 27
图 4.2 图 4.1 中所示的基因和 RNA-Seq 数据通过 Trinity[61] 拼装得到的剪切图 28
图 4.3 一个 3-PARTITION 实例对应的 M-ISOFORM 实例的构造的基因 33
图 5.1 人的基因组中出现的一个转录本的一端位于另一个转录本的中间的现象 37
图 5.2 Giannoukos et al. 在 [111] 中提供的细菌的 RNA-Seq 数据 38
图 5.3 *E. coli* 中的乳糖操纵子[122] 40

参考文献

[1] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 2009, 10(1):57–63

[2] Ozsolak F, Milos P M. RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics, 2010, 12(2):87–98

[3] Garber M, Grabherr M G, Guttman M, et al. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods, 2011, 8(6):469–477

[4] Pickrell J K, Marioni J C, Pai A A, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 2010, 464(7289):768–772

[5] Encyclopedia of DNA Elements (ENCODE). http://encodeproject.org

[6] Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008, 320(5881):1344–1349

[7] Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods, 2009, 6(5):377–382

[8] Bánfai B, Jia H, Khatun J, et al. Long noncoding RNAs are rarely translated in two human cell lines. Genome Research, 2012, 22(9):1646–1657

[9] Mortazavi A, Williams B A, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 2008, 5(7):621–628

[10] Wang E T, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature, 2008, 456(7221):470–476

[11] Katz Y, Wang E T, Airoldi E M, et al. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods, 2010, 7(12):1009–1015

[12] Deng N, Puetter A, Zhang K, et al. Isoform-level microRNA-155 target prediction using RNA-seq. Nucleic Acids Research, 2011, 39(9):e61–e61

[13] Lu T, Lu G, Fan D, et al. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Research, 2010, 20(9):1238–1249

[14] Mercer T R, Gerhardt D J, Dinger M E, et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nature Biotechnology, 2011, 30(1):99–104

[15] Howald C, Tanzer A, Chrast J, et al. Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. Genome Research, 2012, 22(9):1698–1710

[16] Lalonde E, Ha K C, Wang Z, et al. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. Genome Research, 2011, 21(4):545–554

[17] Djebali S, Davis C A, Merkel A, et al. Landscape of transcription in human cells. Nature, 2012, 489(7414):101–108

[18] Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research, 2012, 22(9):1775–1789

[19] Gerstein M B, Kundaje A, Hariharan M, et al. Architecture of the human regulatory network derived from ENCODE data. Nature, 2012, 489(7414):91–100

[20] Fairfax B P, Makino S, Radhakrishnan J, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nature Genetics, 2012, 44(5):502–510

[21] Morrissy A S, Griffith M, Marra M A. Extensive relationship between antisense transcription and alternative splicing in the human genome. Genome Research, 2011, 21(8):1203–1212

[22] Park E, Williams B, Wold B J, et al. RNA editing in the human ENCODE RNA-seq data. Genome Research, 2012, 22(9):1626–1633

[23] Shalek A K, Satija R, Adiconis X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature, 2013.

[24] Tilgner H, Knowles D G, Johnson R, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Research, 2012, 22(9):1616–1625

[25] Ørom U A, Derrien T, Beringer M, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell, 2010, 143(1):46–58

[26] Mercer T R, Neph S, Dinger M E, et al. The human mitochondrial transcriptome. Cell, 2011, 146(4):645–658

[27] Chung W J, Agius P, Westholm J O, et al. Computational and experimental identification of mirtrons in Drosophila melanogaster and Caenorhabditis elegans. Genome Research, 2011, 21(2):286–300

[28] Gingeras T R. Implications of chimaeric non-co-linear transcripts. Nature, 2009, 461(7261):206–211

[29] Roy S, Ernst J, Kharchenko P V, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science, 2010, 330(6012):1787–1797

[30] Axtell M J, Westholm J O, Lai E C, et al. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. Genome Biology, 2011, 12(4):221

[31] Berezikov E, Liu N, Flynt A S, et al. Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. Nature Genetics, 2010, 42(1):6–9

[32] Cherbas L, Willingham A, Zhang D, et al. The transcriptional diversity of 25 Drosophila cell lines. Genome Research, 2011, 21(2):301–314

[33] Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Research, 2012, 22(10):2008–2017

[34] Stoeckius M, Maaskola J, Colombo T, et al. Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. Nature Methods, 2009, 6(10):745–751

[35] Lau N C, Robine N, Martin R, et al. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome Research, 2009, 19(10):1776–1785

[36] Schena M, Shalon D, Davis R W, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995, 270(5235):467–470

[37] Okoniewski M J, Miller C J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics, 2006, 7(1):276

[38] Royce T E, Rozowsky J S, Gerstein M B. Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. Nucleic Acids Research, 2007, 35(15):e99

[39] Boguski M S, Tolstoshev C M, Bassett Jr D E, et al. Gene discovery in dbEST. Science, 1994, 265(5181):1993–1994

[40] Gerhard D S, Wagner L, Feingold E, et al. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Research, 2004, 14(10b):2121

[41] Marioni J C, Mason C E, Mane S M, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research, 2008, 18(9):1509–1517

[42] Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25(9):1105–1111

[43] Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 2010, 28(5):511–515

[44] Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. nature protocols, 2012, 7(3):562–578

[45] Merkin J, Russell C, Chen P, et al. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. Science, 2012, 338(6114):1593–1599

[46] Wang Z, Fang B, Chen J, et al. De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (Ipomoea batatas). BMC Genomics, 2010, 11(1):726

[47] Roberts A, Pimentel H, Trapnell C, et al. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics, 2011, 27(17):2325–2329

[48] Wang K, Singh D, Zeng Z, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research, 2010, 38(18):e178–e178

[49] Van Dijk E, Chen C, Carafa Y, et al. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature, 2011, 475(7354):114–117

[50] Sinicropi D, Qu K, Collin F, et al. Whole Transcriptome RNA-Seq Analysis of Breast Cancer Recurrence Risk Using Formalin-Fixed Paraffin-Embedded Tumor Tissue. PLoS One, 2012, 7(7):e40092

[51] Smith T F, Waterman M S. Identification of common molecular subsequences. Journal of Molecular Biology, 1981, 147(1):195–197

[52] Needleman S B, Wunsch C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 1970, 48(3):443–453

[53] Isaev A. Introduction to Mathematical Methods in Bioinformatics. Universitext, Springer, 2004

[54] Kent W J. BLAT - the BLAST-like alignment tool. Genome Research, 2002, 12(4):656–664

[55] Green P. Phrap documentation. http://www.phrap.org/phredphrap/phrap.html

[56] Bonfield J K, Smith K F, Staden R. A new DNA sequence assembly program. Nucleic Acids Research, 1995, 23(24):4992–4999

[57] Kececioglu J D, Myers E W. Combinatorial algorithms for DNA sequence assembly. Algorithmica, 1995, 13(1-2):7–51

[58] Myers E W. Toward simplifying and accurately formulating fragment assembly. Journal of Computational Biology, 1995, 2(2):275–290

[59] Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome research, 1999, 9(9):868–877

[60] Pevzner P A, Tang H, Waterman M S. An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences, 2001, 98(17):9748–9753

[61] Grabherr M G, Haas B J, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, 2011, 29(7):644–652

[62] Li B, Dewey C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 2011, 12(1):323

[63] Gilbert W. Why genes in pieces? Nature, 1978, 271(5645):501

[64] Rosenfeld M G, Lin C R, Amara S G, et al. Calcitonin mRNA polymorphism: peptide switching associated with alternative RNA splicing events. Proceedings of the National Academy of Sciences, 1982, 79(6):1717–1721

[65] Early P, Rogers J, Davis M, et al. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. Cell, 1980, 20(2):313–319

[66] Sharp P A. Split genes and RNA splicing. Cell., 1994, 77(6):805–815

[67] Modrek B, Lee C. A genomic view of alternative splicing. Nature Genetics, 2002, 30(1):13–19

[68] Pachter L. Models for transcript quantification from RNA-Seq. ArXiv e-prints, 2011.

[69] Roberts A, Trapnell C, Donaghey J, et al. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology, 2011, 12(3):R22

[70] Peng Y, Leung H C, Yiu S, et al. T-IDBA: A de novo iterative de Bruijn graph assembler for transcriptome. Proceedings of Research in Computational Molecular Biology. Springer, 2011. 337–338

[71] Xing Y, Yu T, Wu Y N, et al. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. Nucleic Acids Research, 2006, 34(10):3150–3160

[72] Xia Z, Wen J, Chang C C, et al. NSMAP: A method for spliced isoforms identification and quantification from RNA-Seq. BMC Bioinformatics, 2011, 12(1):162

[73] Li W, Feng J, Jiang T. IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly. In: Bafna V, Sahinalp S, (eds.). Proceedings of Research in Computational Molecular Biology. Springer Berlin Heidelberg, 2011: 168–188

[74] Li W, Jiang T. Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. Bioinformatics, 2012, 28(22):2914–2921

[75] Mezlini A M, Smith E J, Fiume M, et al. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Research, 2012.

[76] Li J J, Jiang C R, Brown J B, et al. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proceedings of the National Academy of Sciences, 2011, 108(50):19867–19872

[77] Hiller D, Wong W. Simultaneous Isoform Discovery and Quantification from RNA-Seq. Statistics in Biosciences, 2012. 1–19

[78] Heber S, Alekseyev M, Sze S H, et al. Splicing graphs and EST assembly problem. Bioinformatics, 2002, 18(suppl 1):S181–S188

[79] Schwarz G. Estimating the dimension of a model. Annals of Statistics, 1978, 6(2):461–464

[80] Robert C, Casella G. Monte Carlo Statistical Methods. Springer Texts in Statistics, Springer, 2004

[81] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 1996. 267–288

[82] Jiang H, Wong W H. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics, 2009, 25(8):1026–1032

[83] RefSeq: NCBI Reference Sequence Database. http://www.ncbi.nlm.nih.gov/refseq/

[84] Fullwood M J, Wei C L, Liu E T, et al. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Research, 2009, 19(4):521–532

[85] Casella G, Berger R. Statistical inference. Duxbury advanced series in statistics and decision sciences, Thomson Learning, 2002

[86] Lehmann E, Scheffé H. Completeness, Similar Regions, and Unbiased Estimation - Part I. Proceedings of Selected Works of EL Lehmann. Springer, 2012: 233–268

[87] Lehmann L, Scheffé H. Completeness, Similar Regions, and Unbiased Estimation - Part II. Selected Works of EL Lehmann, 2012. 269–286

[88] Ewens W, Grant G. Statistical Methods in Bioinformatics: An Introduction. Number v. 10, Springer, 2005

[89] Waterman M S, Vingron M. Sequence comparison significance and Poisson approximation. Statistical Science, 1994. 367–381

[90] Lander E S, Waterman M S. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 1988, 2(3):231–239

[91] Oshlack A, Wakefield M J, et al. Transcript length bias in RNA-seq data confounds systems biology. Biology Direct, 2009, 4(1):14

[92] Young M, Wakefield M, Smyth G, et al. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology, 2010, 11(2):R14

[93] Oshlack A, Robinson M, Young M. From RNA-seq reads to differential expression results. Genome Biology, 2010, 11(12):220

[94] Bullard J, Purdom E, Hansen K, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics, 2010, 11(1):94

[95] Li B, Ruotti V, Stewart R M, et al. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics, 2010, 26(4):493–500

[96] Efron B, Tibshirani R. An Introduction to the Bootstrap. Monographs on statistics and applied probabilities, Chapman & Hall/CRC, 1993

[97] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 1995. 289–300

[98] Salzman J, Jiang H, Wong W H. Statistical Modeling of RNA-Seq Data. ArXiv e-prints, 2011.

[99] Luenberger D, Ye Y. Linear and Nonlinear Programming. International Series in Operations Research & Management Science, Springer, 2008

[100] Preparata F P, Shamos M I. Computational Geometry: An Introduction. New York, NY, USA: Springer-Verlag New York, Inc., 1985

[101] Tomescu A, Kuosmanen A, Rizzi R, et al. A novel min-cost flow method for estimating transcript expression with RNA-Seq. BMC Bioinformatics, 2013, 14(Suppl 5):S15

[102] Vatinlen B, Chauvet F, Chrétienne P, et al. Simple bounds and greedy algorithms for decomposing a flow into a minimal set of paths. European Journal of Operational Research, 2008, 185(3):1390 – 1401

[103] Garey M, Johnson D. Complexity Results for Multiprocessor Scheduling under Resource Constraints. SIAM Journal on Computing, 1975, 4(4):397–411

[104] Garey M R, Johnson D S. Computers and Intractability; A Guide to the Theory of NP-Completeness. New York, NY, USA: W. H. Freeman & Co., 1990

[105] Li L M, Kim J H, Waterman M S. Haplotype reconstruction from SNP alignment. Journal of Computational Biology, 2004, 11(2-3):505–516

[106] Xing E P, Jordan M I, Sharan R. Bayesian haplotype inference via the Dirichlet process. Journal of Computational Biology, 2007, 14(3):267–284

[107] Sharan R, Halldorsson B, Istrail S. Islands of Tractability for Parsimony Haplotyping. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2006, 3(3):303–311

[108] Koren S, Schatz M C, Walenz B P, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology, 2012, 30(7):693–700

[109] McClure R, Balasubramanian D, Sun Y, et al. Computational analysis of bacterial RNA-Seq data. Nucleic Acids Research, 2013.

[110] Jones D C, Ruzzo W L, Peng X, et al. A new approach to bias correction in RNA-Seq. Bioinformatics, 2012, 28(7):921–928

[111] Giannoukos G, Ciulla D M, Huang K, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biology, 2012, 13(3):R23

[112] Gilbert J A, Field D, Huang Y, et al. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One, 2008, 3(8):e3042

[113] Urich T, Lanzén A, Qi J, et al. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One, 2008, 3(6):e2527

[114] Gifford S M, Sharma S, Rinta-Kanto J M, et al. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. The ISME journal, 2010, 5(3):461–472

[115] Helbling D E, Ackermann M, Fenner K, et al. The activity level of a microbial community function can be predicted from its metatranscriptome. The ISME Journal, 2011, 6(4):902–904

[116] Mason O U, Hazen T C, Borglin S, et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. The ISME Journal, 2012.

[117] Huson D H, Mitra S, Ruscheweyh H J, et al. Integrative analysis of environmental sequences using MEGAN4. Genome Research, 2011, 21(9):1552–1560

[118] Lesniewski R A, Jain S, Anantharaman K, et al. The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. The ISME Journal, 2012.

[119] Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nature Reviews Genetics, 2009, 11(1):9–16

[120] Guell M, Noort V, Yus E, et al. Transcriptome Complexity in a Genome-Reduced Bacterium. Science, 2009, 326(5957):1268–1271

[121] Koide T, Reiss D J, Bare J C, et al. Prevalence of transcription promoters within archaeal operons and coding sequences. Molecular Systems Biology, 2009, 5(1)

[122] Shuman H A, Silhavy T J. The art and design of genetic screens: Escherichia coli. Nature Reviews Genetics, 2003, 4(6):419–431

致　谢

衷心感谢张学工和江瑞两位老师对本人的精心指导。
同时也感谢 ThuThesis, 以及其他各种开源项目给予我的帮助和支持。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

附录 A　文献翻译

A.1　文献索引

Garber, Manuel, et al. “Computational methods for transcriptome annotation and quantification using RNA-seq.” Nature Methods 8.6 (2011): 469-477.

A.2　翻译

高通量 RNA 测序（RNA-Seq）帮助我们对转录有一个全面的了解，从而完整的注释和定量的样本之间的基因。实现这一目标，需要越来越复杂的计算方法。这些计算的挑战可分为三大类：（一）读段比对，（二）转录重建及（三）表达量化。在这里，我们解释的主要概念和实践的挑战，和为每个类别一般类的解决方案。最后，我们强调这些类之间的相互依存和讨论不同生物应用的好处。

定义的精确地研究所有基因在不同的细胞类型和表达是理解生物学的关键。直到最近，这样的数据是非常昂贵的生产和实验费力。注释转录的主要方法，需要缓慢而昂贵的过程中克隆的 cDNA 或表达序列标签（EST）库，然后由毛细管测序。

DNA 测序技术的最新进展使人们有可能来自细胞的 RNA 序列的 cDNA 的大规模并行测序技术，这一过程被称为 RNA-Seq。

在这里，我们重点介绍 RNA-Seq 分析面临的核心挑战和需要解决的计算方法。首先，我们描述对齐的方法直接读段到一个参考转录或基因组（’读段比对’）。第二，我们讨论的方法来确定表达的基因和亚型（’转录重建’）。第三，我们提出了基因及变异体丰度估计的方法，以及用于分析样品（’表达定量’）之间的差异表达的方法。

由于 RNA-Seq 的数据生成的持续改善，有不同的成熟可用的计算工具。在某些领域，如读段比对，丰富的算法存在，但在其他方面，如差异表达分析，解决方案才刚刚开始出现。我们并不较全面地描述每一种方法，而是突出重点的共同原则以及相关 RNA-Seq 分析中每种方法及其应用的关键差异。我们还

讨论了这些不同的方法如何可能会影响数据的结果和解释。虽然我们讨论每三个类别作为独立的单位，RNA-Seq 的数据分析往往需要从所有三个类别使用方法。这里描述的方法在很大程度上是独立选择库建设协议，同时我们介绍双端测序（读从两端的一个片段），它提供了有价值的信息，在 RNA-Seq 的分析各个阶段有显着的应用。

A.2.1 比对短 RNA-Seq 读段

RNA-Seq 分析中最基本的任务之一是将读段比对到基因组参考序列上。序列是生物信息学中的一个经典问题, 并且有多种解决方案为 EST 进行比对。

比对没有剪切的读段到基因组参考序列对于定量分析有十分重要的作用。

没有剪切的读段仅限于识别已知的外显子和剪切位点，不允许涉及新外显子的剪接事件的识别。另外，读段可以对齐到整个基因组，其中包括内含跨越读段，可能会有很大的差距，需要妥善安置。有几种方法，统称为“拼接对准’，落入两大类：’第一外显子’和’种子和延长。第一外显子也支持配对末端读比对，从而增加了对准特异性。

第一外显子的方法更快，需要更少的计算资源相比，种子扩展方法。例如，种子扩展方法 (GSNAP) 需要更长的时间 8x (340个CPU小时) 比第一个外显子的方法（TopHat） 1.5x 拼接读段。然而，这些额外的剪接点的生物学意义还没有被证实。

第一外显子的方法可能发生基因有重组的错误从而错过拼接路线。与此相反，种子扩展方法评价剪接和未拼接的路线在同一步骤中，这降低偏向未拼接的路线，得到每次读操作的最佳位置。种子扩展方法比第一外显子的方法比对读段时多态性物种时有更好的表现。

A.2.2 转录组重建

精确地比对并且通过一个特定样本组装读段成转录单位。总的来说，我们这个过程称为转录重建。转录组重建是一项困难的计算任务，主要有三个原因。首先，基因表达跨越几个数量级，表示仅由少数读段某些基因。其次，读段源于成熟的 mRNA（外显子），以及来自不完全剪接的前体 RNA（包含内含子序列），使得难以识别的成熟转录本。第三，读段很短，基因可以有许多亚型，它具有挑战性，以确定哪些亚型产生每个读。

有几种方法来重建转录，它们分为两大类：“基因组引导’和’基因组独立’。基因引导的方法依赖于一个参考基因组首先比对所有读段的基因组，然后组装重叠读入笔录。相比之下，独立于基因组的方法组装成的转录本，而无需使用参考基因组直接读段。

基因引导重建。现有基因制导方法可分为两大类：’外显子鉴定“和”基因制导组装“的方法。

外显子识别早期开发时，读段短（36个碱基）和几个外显子 - 外显子路口对齐。他们首先定义假定外显子覆盖的岛屿，然后使用拼接读段这些覆盖岛屿定义跨越外显子边界和外显子之间建立连接。外显子识别方法提供转录本重建问题最合适的短读段第一个办法来解决，但他们是动力不足，以确定低表达，长和选择性剪接基因的全长结构。

要利用更长的读长，基因组制导组件方法，如 Cufflinks。

Scripture 和 Cufflinks 也有类似的计算要求，可以在个人计算机上运行。组装类似的转录本，在高表达水平，但显着差异表达的转录袖扣报告 3x 更多的位点（70,000 vs 25,000），其中大部分不及格Scripture所使用的统计学意义阈值较低。在最极端的情况下，超过300个亚型为一个单一的轨迹，而袖扣的Scripture报告报告相同基因的11种亚型。

基因无关重建。而非比对读段参考序列，基因组独立的转录重建算法如 transABySS 。重叠的 k - 1 之间的碱基通过k个聚体构成的曲线图中，可以构造所有可能的序列。接下来，在图中遍历路径，指导阅读和配对末端覆盖水平，消除虚假的分支点引入 k-mer 所共享的不同转录，由读段和配对两端的，但不支持。所有剩余的路径，然后通过图形报告作为一个单独的转录本。

虽然基因组独立重建的概念很简单，有两个主要的并发症：区别序列变异的错误，并找到灵敏度和图形的复杂性之间的最佳平衡。比对的策略不同，测序错误介绍分支点在图中，增加其复杂性。要消除这些影响，基因组独立的方法在不同的路径图中的覆盖面和覆盖截止申请来决定何时遵循的路径，或将其删除时。

为了配合转录丰度的变化，内在的表达谱数据的几种方法，如 transABySS，使用变 k-mer 的战略来获得权力，表达水平之间组装的转录本，尽管在耗费 CPU 和需要并行执行。

相比重建战略。基因组引导和基因组独立的算法已上报准确地重建数千转

录本和许多替代剪接形式，对上述问题的答案变得不那么清晰，取决于分析目标上。在许多情况下，一种混合的方法结合了基因组独立和基因组制导的战略可能最适合捕捉已知信息，以及把握新的变化。在实践中，基因组独立的方法需要相当大的计算资源（650的CPU小时和16千兆字节的随机存取存储器（RAM））相比，基因组的引导方法（4的CPU时间和<4千兆字节RAM）。

A.2.3 估计转录本表达量

表达定量一直是一个重要的应用。在过去的十年中，DNA 微阵列技术高通量转录组分析的首选技术。当使用 RNA-Seq 的估计基因表达，阅读计数需要得到适当的标准化提取有意义的表达估计。

考虑到这些问题，读段每碱基每百万比对读段（RPKM）度量的转录本，转录本的读段计数正常化它的长度和总数的比对读段样品中。

由于许多基因有多种亚型，其中许多股票外显子，和许多基因家族有密切的旁系，一些读段不能被明确分配的转录本。

我们注意到，一些潜在的亚型极大地影响结果，引入不确定性与不正确或拼接错误的异构体。因此，工作时，产生最大的异构体集的方法，它是必要的预过滤转录之前，某些基因的表达估计。这适用于两个基因组的引导下，以及基因组独立的算法。

通常情况下，目标是要估计每个基因的表达，而不是为每个异构体或转录本，并且，如下面所讨论的交点，可以降低功耗的差异表达分析。

A.2.4 RNA-Seq 的差异表达分析

量化和规范化的表达式的值，一个重要的问题是要了解如何将这些表达水平的不同而不同的条件。过去的十年中看到了发展广泛的方法表达差异进行统计分析，使用微阵列。

为了适应基于计数性质的 RNA-Seq 的数据，最初的方法为蓝本所观察到的读段使用计数型分布，如泊松分布。

重要的是要注意，虽然这些方法可以分配差异表达的意义，生物的结论，必须小心诠释。例如，虽然在测序过程的可变性比较低的微阵列杂交和最重要的，因为在生物样品中的内在变异。至于与任何生物测量，生物复制提供内在

的，非技术性的转录表达变异的唯一衡量和因而是有史以来差异表达分析的关键。

对基因的差异表达分析的量化策略。当执行差异表达分析，大多数方法采取输入归一读的计数每个基因在每个条件。使用简化基因定量模型，如外显子交会法或外显子工会法，可能会导致意想不到的结论。当一个基因有多个异构体，基因的表达的变化可能不会导致相应的变化，原始的基因水平计数。

A.2.5 结论和未来发展的预期

随着测序技术的成熟，现有的计算工具将需要发展以满足新的要求，新的工具，会出现使新的应用。例如，读段长度的不断增加，新的比对方法将需要有效地使数以百万计的长读段一个艰巨的任务。更长的读段经常跨越多个外显子 - 外显子剪切位点，转录本重建和定量分析方法，将有利于纳入更完整的异构信息编码较长读段。标准 RNA-Seq 的方法不适合注释5’起始位点和3’端的转录本通过使用专门的 RNA-Seq 的库，确定两端转录重建方法，将改善转录本注解。从 RNA-Seq 的数据估算表达方法需要改进，以更好地处理日益普及生物复制实验和理想模式（自动减去）系统的误差来源所引入实验室方法（如3’-端偏见）。提供的序列表达的转录，RNA-Seq 的编码信息等位基因变异和 RNA 加工，所以重建方法应适应这种变异并且报告这些变异。计算技术的改进，既可以在实验室以及不断循环，将继续扩大的 RNA-Seq 的可能性，使得这项技术适用于更多种类的生物学问题。

附录 B　源代码

- https://github.com/tianyang-li/de-novo-rna-seq-quant-1
- https://github.com/tianyang-li/thu-undegrad-thesis-code
- https://github.com/tianyang-li/aarsa
- https://github.com/tianyang-li/rna-seq-len-est-0
- https://github.com/tianyang-li/misc-bioinfo-0
- https://github.com/tianyang-li/de-novo-metatranscriptome-analysis--the-uniform-model
- https://github.com/tianyang-li/human-rna-seq-analysis-0
- https://github.com/tianyang-li/de-novo-rna-seq-quant-with-contigs-py-0
- https://github.com/tianyang-li/bi-misc
- https://code.google.com/p/meta-transcriptome/

在学期间参加课题的研究成果

个人简历

2009 年 8 月考入清华大学自动化系自动化专业, 2013 年 7 月本科毕业并获得工学学士学位。

发表的学术论文

[1] T. Li, R. Jiang, and X. Zhang. Isoform reconstruction using short RNA-Seq reads by maximum likelihood is NP-hard. ArXiv e-prints, May 2013. http://arxiv.org/abs/1305.0916.

[2] Tianyang Li, Rui Jiang and Xuegong Zhang. *De novo* transcript reconstruction and abundance estimation in eukaryotic RNA-Seq data analysis. RECOMB 2013. (Poster)