科学通报 2013 年 第 58 卷 第 33 期：3471 ~ 3482
论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

一种基于 RNA-Seq 的基因组注解评估方法

王颖*, 刘麟

厦门大学信息科学与技术学院自动化系, 厦门 361000
* 联系人, E-mail: wangying@xmu.edu.cn

2013-03-26 收稿, 2013-07-10 接受
国家自然科学基金(61203282, 61202144)和美国国立卫生研究院基金(NIH/NIMH 5 RC2 MH090047-01)资助

摘要 新一代测序技术下RNA-Seq测序数据为解码真核生物的转录组带来了突破性的变革, 其细致到碱基层面的高分辨率信息, 使得仅采用RNA-Seq作为唯一数据源便可对现有的基因组进行注解. 同样地, 利用 RNA-Seq 信息也能验证现有的剪切位点、外显子乃至转录物的注解信息. 因此本文提出利用 RNA-Seq 数据对现有的基因组注解数据库进行评估, 基于 RNA-Seq 的配准信息提出在基因、转录物、外显子、剪切位点和碱基层面的特异性和敏感性度量指标, 进而评估基因组注解数据库的完整性和精确性. 基于该评估框架, 通过来自人类 16 个组织的 11 亿条 RNA-Seq 读段(read)数据对 5 个代表性的人类基因组注解数据库进行评估, 并基于评价结果构建人体综合准确注解数据库; 此外, 还对现有的恒河猴基因组注解数据库进行了评估, 发现该数据库的完整性有很大欠缺, 同时其注解的精确性与人类数据库的注解水平有较大的差距. 基于该评估体系, 可对各物种的基因组注解信息的完整性和精确性进行全面、快速和高效的评估及验证.

关键词
基因组
转录组
注解数据库
RNA-Seq
敏感性
特异性

RNA-Seq 技术已经成为研究真核生物转录组很重要的工具[1,2], 被用在许多相关的生物研究中, 如量化转录物表达水平[3,4], 高分辨率的分析可变剪切机理[5,6], 识别新的剪切位点和转录物形式[7,8], 探索遗传变异和表达量化特征位点(eQTL)[9,10]. 在上述的计算生物学研究中, 基因组注释的准确性和完整性对于研究结果的正确和可靠性具有重要的影响. 目前, 大量物种的基因组测序和装配完成, 随之产生了不同方式建立的基因组注解数据库; 同时, 对一些模式生物, 存在若干不同的基因组注释数据库. 因此如何评估这些基因组注释数据库成为了一个重要的研究课题. 迄今为止, 有 3 个著名的基因注解竞赛项目, GASP[11], EGASP[12]和 RGASP[13], 它们的目的是对基因的结构及功能的各类注解方法进行深入和客观的评估, 其中仍在进行中的 RGASP 竞赛项目以 RNA-Seq 作为唯一数据源对人类、果蝇和线虫的基因组进行注解竞赛, 评估各类自动注解算法. 这些项目主要通过手工实验对注解进行验证, 由于手工实验成本高, 只能对某段区域的注解进行评估. 但这样的评估方法应用到整个基因组, 将耗费巨大的精力, 且不易于移植到其他资源更少的物种研究中.

受到 RNA-Seq 作为唯一数据源的基因组注解项目的启发, 本文提出利用 RNA-Seq 提供的细致到碱基层面的高分辨率信息, 在全基因组范围内评估基因注解的准确性和完整性. 基于熵、极大似然估计以及 EM 算法等技术提出在基因、转录物、外显子、剪切位点和碱基等不同层面的注解特异性和敏感性评价指标. 利用这些评价指标, 本文对 5 个具有代表性的人类基因组注解数据库 AceView, RefSeq, Ensembl, UCSC 和 Gencode 进行评估. 从整个转录组的注解范围来看, RefSeq 是保守型的数据库, 给出最简明、准确的注解; 而 AceView 是开放型的数据库, 提供更多

引用格式: 王颖, 刘麟. 一种基于 RNA-Seq 的基因组注解评估方法. 科学通报, 2013, 58: 3471–3482
Wang Y, Liu L. RNA-Seq-based assessment for genome annotation databases (in Chinese). Chin Sci Bull (Chin Ver), 2013, 58: 3471–3482, doi: 10.1360/972012-1038

的信息量, 但假阳率也随之增加. 在相同的公共基因的基线上, 结果显示出非常明显的特异性和敏感性分布趋势, 在转录物、外显子和剪切位点层面上, 数据库的特异性随着敏感性的增加而降低. 该结论与 EGASP 项目[14]在局部编码区采用手工验证所获得的相关结论一致, 但更增加了在剪切位点和碱基层面的分析. 同时, 该评估也应用在恒河猴(*Macaca mulatta*)的基因组注解 RefSeq 数据库的评估中, 发现目前的恒河猴基因组注解很不完整, 而且在注解精度上和人类注解相比还有很大的差距, 该结论和 Zhang 等人[15]的文中提到的恒河猴的基因组测序到装配到注解均存在很多不足的结论一致, 但对不足的具体层面给出了具体的定性分析. 相比于手工评估, 本文提出的基于 RNA-Seq 数据的评估方法能高效快速地完成在全基因组范围内各层面上的注解评估, 同时成本低廉, 仅需要足够深度的 RNA-Seq 测序数据和计算资源, 而且很容易应用到各类物种的注解数据库评估上. 实验结果表明, 文中提出的方法可以对真核生物在全基因组范围内实现自动的基因组注释评估, 虽然该方法无法像 EGASP 那样提供非常详细的细节信息, 但是它的结果给出基因组注解的整体和局部的评估信息, 为选择和评价各物种的基因组注释数据提供了量化指标, 为后续的计算生物学研究提供了帮助.

1 基因组注解的评估指标

利用 RNA-Seq 数据对基因组注解数据库进行评估和比较的流程如图 1 所示. 首先利用现有的任意读段(read)匹配工具, 如 PerM[16]和 Bowtie[17], 将 RNA-Seq 数据直接配准到每个数据库的转录物序列上. 基于读段的配准结果, 计算下文定义的不同层面上的特异性和敏感性, 通过本文实现的软件流 Genome Annotation Assessment Pipeline(GA^2P)评估基因组注解, 衡量 RNA-Seq 数据与注解数据库之间的匹配关系.

在本文的特异性和敏感性的度量指标中将用到读段完全支持转录物(complete-supported transcripts)的定义, 阐述如下.

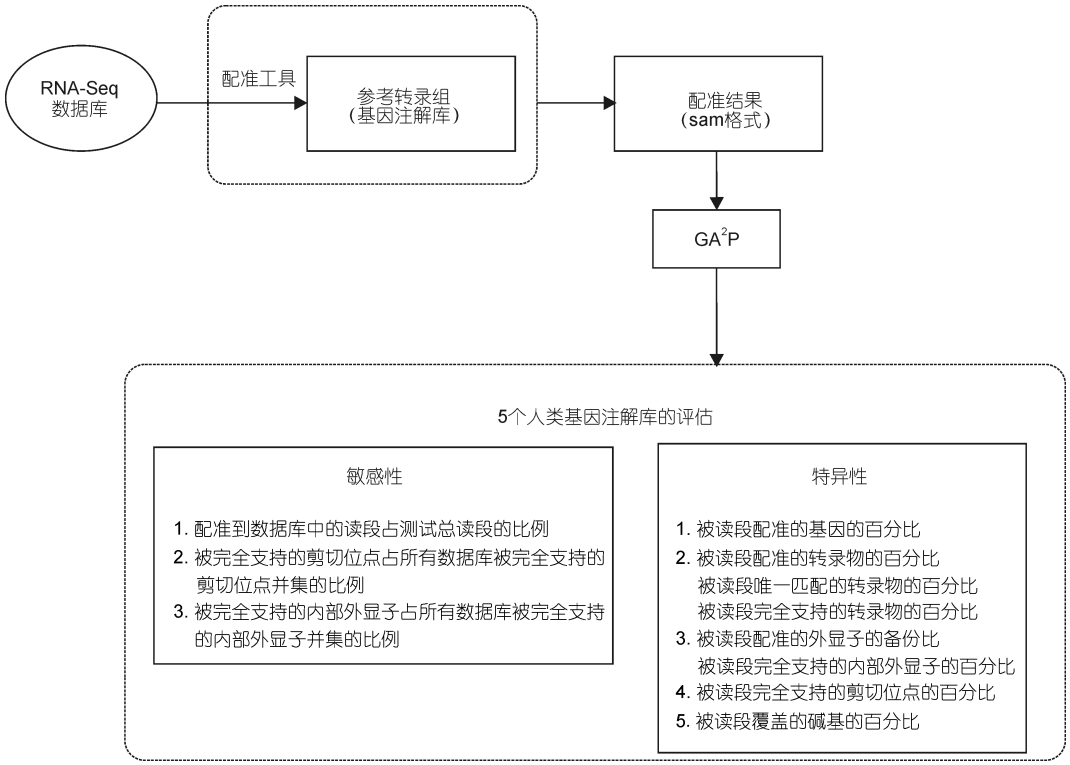

图 1 基因组注解数据库评估流程图

1.1 完全被支持转录物的定义

读段在转录物上的匹配情况是推断该转录物是否被准确注解的重要信息. 因此我们提出了完全被支持转录物的概念, 当转录物被读段完全支持, 则有较强的证据判定其是被准确注解的. 在较高的表达水平和充分的测序深度下, 我们有理由要求转录物从 5′端到 3′端被读段完全、连续覆盖作为完全被读段支持的证据. 但是在转录物表达水平较低或测序深度不充足的情况下, 我们就需要通过统计的方法来推断转录物是否被读段完全支持. 因此, 我们提出在两种情况下的完全被支持转录物的定义:

(ⅰ) 充足的表达丰度和充分的测序深度下. 当某个转录物从 5′端到 3′端被读段完全、连续覆盖, 不存在缺口, 则该转录物被 RNA-Seq 的读段完全支持. 这是对该转录物注解最严苛的评价指标, 也是支持该转录物注解正确性最强有力的证据. 然而, 转录物的起点和终点是基因注解最难准确定位的地方[10], 因此我们将限制条件放松到从第一个外显子末端到最后一个外显子的前端被读段连续覆盖. 对于只有两个或更少外显子的转录物则要求去除 5′端和 3′端 5%的碱基, 剩余碱基位都被完全覆盖, 图示见附件 1 中的图 4.

(ⅱ) 由于表达水平、测序深度、测序偏差等因素导致许多被准确注解的转录物无法获得连续的读段分布. 这种情况下, 本文提出基于当前的基因覆盖深度(GC: Gene Coverage)和观测到的碱基覆盖率(CP: Covered-base Percentage), 建立以下统计模型推断该注解转录物是否真实存在, 将观测到的数据表示为:

$$\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n),\ \boldsymbol{x}_i=(x_{\mathrm{GC}_i},x_{\mathrm{CP}_i}), \tag{1}$$

其中, x_{GC_i} 是转录物 $\boldsymbol{i}$ 所属基因覆盖率的对数值(以 10 为底), 基因覆盖率是指覆盖某一基因的所有读段长度总和与该基因长度的比值; x_{CP_i} 是转录物 i 被读段覆盖的碱基占转录物总长度的比例; n 是转录物的个数. 我们将 http://www.ebi.ac.uk/ena/data/view/ERP000546 中的人类 RNA-Seq 数据配准到 AceView 注解数据库中随机抽取的 5251 个基因、20215 个转录物上, 分析 x_{GC_i} 与 x_{CP_i} 之间的关系, 如图 2 所示.

图 2 中所包含为被读段完全支持和不完全支持两类转录物, 因此可设观测到的数据 $(x_{\mathrm{GC}_i},x_{\mathrm{CP}_i})$ 服从如下混合 Gaussian 分布:

$$p(\boldsymbol{X}\mid\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\sigma}^2)=\sum_{j=1}^{2}\omega_j\frac{1}{\sqrt{2\pi\sigma_j^2}}\mathrm{e}^{-\frac{(\boldsymbol{X}_{\mathrm{CP}}-\boldsymbol{\mu}_j)^2}{2\sigma_j^2}}, \tag{2}$$

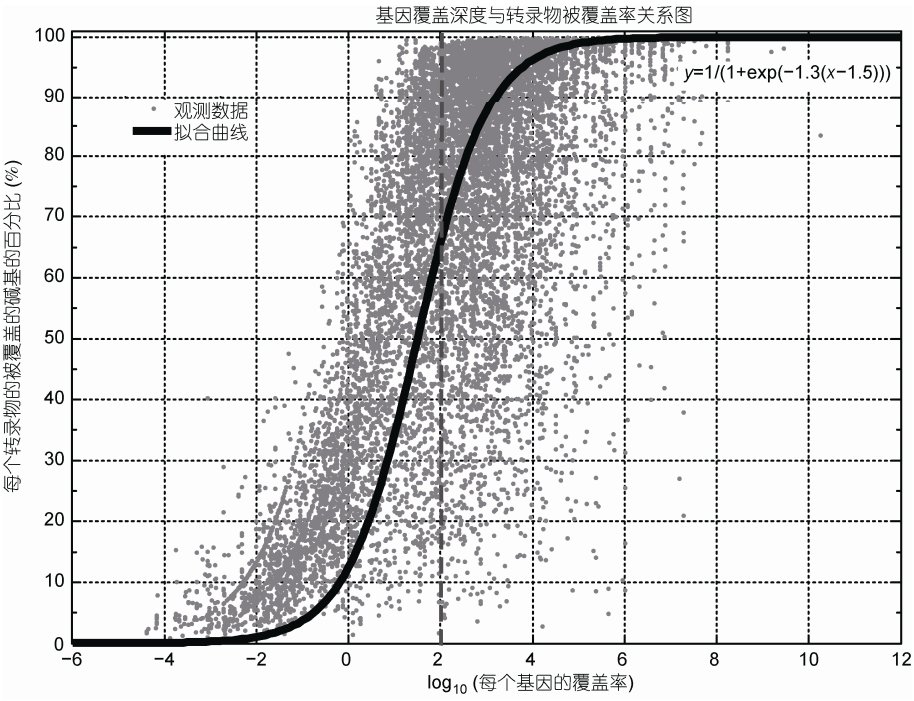

图 2 基因覆盖深度与转录物被覆盖率关系图

其中，$\boldsymbol{X}_{\mathrm{CP}}=(x_{\mathrm{CP}_1},\cdots,x_{\mathrm{CP}_i},\cdots,x_{\mathrm{CP}_n})$ 是观测到的所有转录物其各自所属基因覆盖率的对数值，$\boldsymbol{\sigma}^2=(\sigma_1^2,\sigma_2^2)^{\mathrm{T}}$ 是两个 Gaussian 分布的方差，$\boldsymbol{\omega}=(\omega_1,\omega_2)^{\mathrm{T}}$ 是两个分布的权重，且满足 $\omega_1+\omega_2=1$．而 $\boldsymbol{\mu}=(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)^{\mathrm{T}}=\begin{pmatrix}\mu_{11},\cdots,\mu_{1i},\cdots,\mu_{1n}\\ \mu_{21},\cdots,\mu_{2i},\cdots,\mu_{2n}\end{pmatrix}$ 为两个 Gaussian 分布的均值，为两种情况下的碱基覆盖率．当注解转录物 i 被完全支持，则其服从均值为 μ_{1i}，方差为 σ_1 的正态分布，否则服从均值为 μ_{2i}，方差为 σ_2 的正态分布．转录物 i 的均值 $\mu_{ji},j=1,2$，分别来自两个 Sigmoid 函数:

$$\mu_{ji}=\frac{1}{1+\mathrm{e}^{a_j+b_jx_{\mathrm{GC}_i}}},j=1,2;\ i=1\cdots n, \tag{3}$$

其中，a_j 和 b_j 为 Sigmoid 函数的参数，因此混合模型中的未知参数可记为

$$\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\omega_1,a_1,b_1,\sigma_1,\omega_2,a_2,b_2,\sigma_2). \tag{4}$$

由于被注解的转录物是否真实存在是未知的，因此引入缺失数据 $\boldsymbol{Y}=(y_1,\cdots,y_i,\cdots,y_n)$，$y_i$ 表明观测数据 x_i 来自哪个子分布．当 $y_i=1$，表示第 i 个转录物是被读段完全支持的；当 $y_i=2$，表示第 i 个转录无法被读段完全支持．因此该问题可以通过 EM 算法优化极大似然模型，获得模型未知量的最优值 $\boldsymbol{\theta}^*$，详细的求解过程见附件 1 的第 1 节．基于以上步骤获得的模型，即可推断出在当前的覆盖深度和碱基覆盖率下转录物是否被读段完全支持，也就是该转录物是否被准确注解的强有力证据．

综合以上两种情况的结果可以获得当前数据库中准确注解的转录物集合．因此下文中提到的完全被支持转录物，定义参照本节．

1.2 特异性的度量指标

特异性指的是数据库中真实的信息占全部注解信息的比例，是数据库准确性的衡量．下面从基因、转录物、外显子、剪切位点和碱基层面给出特异性的度量指标．

(ⅰ) 基因层面的特异性

$$\text{被读段配准的基因的比例}=\frac{\text{被读段配准的基因数}}{\text{当前数据库中已注解的基因数}}.$$

“被读段配准(read-hit)”要求至少有一个读段配准到该区域．该指标认为至少被一个读段匹配是一个基因存在的最基本的要求，该指标从基因层面反映基因组注解的特异性，但该指标过于宽松，仅被读段匹配不足以充分证明一个基因的存在，因此我们提出下面更加严格的基因层面的度量指标．

$$\text{包含完全被支持转录物的基因的比例}=\frac{\text{包含完全被支持转录物的基因数}}{\text{当前数据库中已注解的基因数}}.$$

(ⅱ) 转录物层面的特异性

$$\text{被读段配准的转录物的比例}=\frac{\text{被读段配准的转录物的数量}}{\text{当前数据库中已注解的转录物数量}}.$$

同样地，该指标从转录物层面宽松地衡量了基因组注解的特异性．如果转录物被读段配准，表明该转录物的部分注解是正确的，但单纯被读段匹配的转录物有可能是伪基因(pseudo gene)，更进一步，我们给出以下两个衡量指标为转录物注解提供更充分的验证支持:

$$\text{被读段唯一配准的转录物的比例}=\frac{\text{被读段唯一配准的转录物的数量}}{\text{当前数据库中已注解的转录物数量}}.$$

“被读段唯一配准(read-unique-hit)”是指读段唯一匹配到该转录物而不再匹配到其他的区域，该信息更有力地支持了该转录物存在和表达．

$$\text{完全被支持转录物的比例}=\frac{\text{完全被支持转录物数量}}{\text{当前数据库中已注解的转录物数量}}.$$

当转录物被读段完全支持，是被注解的转录物存在最严格的证明，最充分有力地支持了转录物的精确注解．对于“完全被支持转录”的定义充分考虑了在各种表达水平和测序深度前提下的读段分布，因此在转录物层面最准确地衡量了注解的特异性．

(ⅲ) 外显子层面的特异性

$$\text{被读段配准的外显子的比例}=\frac{\text{被读段配准的外显子的比例}}{\text{当前数据库中已注解的外显子数量}}.$$

类似地，该指标从外显子层面衡量基因组注解的特异性．计算中，多个同源异构体中的相同外显子被视为同一外显子．进一步，我们提出下面更加严格的验证指标:

$$\text{完全被支持的内部外显子的比例}=\frac{\text{完全被支持的内部外显子数量}}{\text{当前数据库中已注解的内部外显子数量}}.$$

如果外显子 5′端和 3′端的剪切位点以及 90%的碱基都有读段覆盖，或该短序列来自完全被支持转录物，则称该外显子被读段完全支持(complete-supported

exon). 读段支持两端剪切位点, 则验证了外显子两端注解的准确性, 如附件 1 中图 5 所示. 该指标是外显子准确注解更确凿的证据.

(ⅳ) 剪切位点层面的特异性

$$\text{完全被支持的剪切位点的比例}=\frac{\text{完全被支持的剪切位点数量}}{\text{当前数据库中已注解的剪切位点数量}}.$$

该指标从剪切位点(基因结构)层面衡量注解的特异性. 一个注解的剪切位点满足以下条件之一, 则认为该剪切位点被读段完全支持(complete-supported splice junction):

条件 1: 该剪切位点来自被完全支持的转录物.

条件 2: 基于读段匹配在当前剪切位点上的位置, 通过计算匹配熵 E 衡量读段对该剪切位点提供的支持信息的多样性:

$$E=\sum_{i=1}^{n} p_i \ln p_i, \tag{5}$$

其中, n 是所有匹配到该剪切位点读段的起始位置的数目; 当一个剪切位点被 10 个读段覆盖, 但这 10 个读段均起始于同一碱基位置, 则 $n=1$; p_i 是起始于某一个碱基位置的读段占该剪切位点所有匹配读段数目的比例. 因此 E 指示了当前剪切位受读段支持的多样性. E 值越高, 该剪切位点获得的支持越强. 当熵值高于设定的阈值, 则认为该剪切位点被读段完全支持.

条件 3: E=0 表示极其类似的读段匹配到该剪切位点上相同的碱基位置. 在这种情况下, 我们要求只有这些读段都唯一匹配到该剪切位点且在剪切点两端至少有 10 bp(base pair)以上的覆盖, 则认为该剪切位点被读段完全支持.

(ⅴ) 碱基层面的特异性

$$\text{被读段覆盖的碱基的比例}=\frac{\text{被读段覆盖的碱基数}}{\text{当前数据库中非重叠碱基数}}.$$

非重叠碱基数(non-overlapping bases)是各转录物序列并集的长度. 该指标衡量了被读段覆盖的碱基占数据库中注解的所有碱基的比例, 该指标从碱基层面衡量了基因组注解的特异性.

1.3 敏感性的度量指标

敏感性是指数据库中所含真实信息占全部真实注解信息量的比例, 是数据库完整性的衡量. 但由于目前无法获取完全准确全面的注解集合(ground truth), 因此我们将测序结果 RNA-Seq 提供的信息或者若干数据库提供的真实信息的并集作为真实集合.

(ⅰ)

$$\text{配准读段的比例}=\frac{\text{配准到数据库中的读段数}}{\text{RNA-Seq数据库中的总读段数}}.$$

该指标计算的是能配准到注解数据库中的读段占测试总读段的比例, 衡量了当前注解数据库中包含真实表达的转录物序列信息的比例.

(ⅱ)

$$\text{在数据库并集中被完全支持剪切位点的比例}=\frac{\text{当前数据库中被完全支持剪切位点的数量}}{\text{所有待比较数据库中被完全支持的剪切位点求并集的元素个数}}.$$

该指标反映了注解数据库在剪切位点层面的敏感度. 与特异性中的"完全被支持的剪切位点的比例"的不同之处在于分母. 特异性中该指标的分母为当前数据库中注解的剪切位点的数目; 而敏感性中该指标的分母是将待比较的所有数据库中被完全支持的剪切位点求并集, 将其作为数据真集的元素个数. 该指标从剪切位点层面严格地评估了注解的敏感性.

(ⅲ)

$$\text{在数据库并集中被完全支持外显子的比例}=\frac{\text{当前数据库中被完全支持内部外显子的数量}}{\text{所有待比较数据库中被完全支持的外显子求并集的元素个数}}.$$

该指标反映了数据库外显子层面注解的敏感度. 将待比较的所有数据库中被完全支持的内部外显子的并集作为真实集, 从外显子层面严格地评估了注解的敏感性.

1.4 对人类基因组注解数据库的评估和比较

采用以上各指标对 5 个具有代表性的人类基因组注解数据库进行评估. 基于 RNA-Seq 数据集到注解转录物序列的配准结果, 从基因、转录物、外显子、剪切位点和碱基层面对数据库注解的完整性和准确性进行评估. 基于整个转录组(蛋白质编码基因集合)和 9502 个公共基因子集, 分别获得对基因组注解全局、综合的评价以及统一的衡量基线下的局部更精细的比较.

评估结果是通过本文的 RNA-Seq 分析软件获得的. 该软件将利用 RNA-Seq 数据分析注解数据库的过程实现为软件流平台, 以供不同的物种下不同的数据库的评估使用, 在附件 5 中提供该软件的下载.

将 RNA-Seq 到参考转录组序列的配准结果及注解文件输入软件流平台, 运行后将输出从基因、转录物、外显子和剪切位点层面的评估结果, 该结果不仅反映了注解数据库的特异性和敏感性, 同时也给出了 RNA-Seq 数据与该注解数据库的关系. 该软件流平台只将 RNA-Seq 数据配准到转录组, 大大减少了运行时间和存储空间. 当运行在单 CPU, 10 GB 内存下, 针对 Ensembl 人类基因组注解数据库, 整个运行程序需花费 6.5 h, 47 GB 的存储空间.

1.5 5 个人类基因组注解数据库

我们从 UCSC Genome Bioinformatics (http://genome.ucsc.edu/)下载基于 hg18 的人类基因组注解数据库(截止到 2010 年 4 月 14 日): AceView assembly genes, RefSeq refGene, Ensembl ensGene, UCSC knownGene 和 Gencode, 其中包括 5′端、3′端和编码区的外显子的注解以及各基因转录物的参考序列. 进一步的信息见附件 1 中的表 1. 同时, 我们通过不同数据库之间基因命名的对应关系选择 5 个数据库中都包含的 9502 个公共基因作为基线进行更细致的比对.

1.6 RNA-Seq 评估数据集

我们采用来自 Human BodyMap 2.0 中人类 16 个组织、长度为 75 bp 的 RNA-Seq 数据, 可通过 http://www.ebi.ac.uk/ena/data/view/ERP000546 下载, 这些数据由 Illumina 公司的 HiSeq 2000 测序获得, 共包含 1113014448 个单向 75 bp 读段. 我们采用读段匹配软件 PerM[16]作为读段的匹配工具.

首先测试该 RNA-Seq 数据集的数据质量, 当最多允许 8 个碱基位错配的设置下, 共有 92.95%的读段可以匹配到 NCBI36/hg18 基因组序列和转录组序列(AceView)中, 充分证明该数据集的高质量. 另外, 平均每个 RefSeq 基因有超过 25000 个读段的匹配量, 因此说明在如此充分的测序深度下, 集合人类 16 个组织的测序结果, 该数据集能覆盖绝大部分被表达的基因转录物.

在采用 RNA-Seq 数据集对人类注解数据库进行评估时, 我们采用的是 PerM version 0.3.3 在 F2 种子下最多 5 个碱基位容错配设置, 该设置对 5 以内的所有的错配完全敏感.

1.7 在全转录组范围内的评估

经分析, 5 个数据库具有较明显的注解特征, 整体趋势分析如下, 具体数据见表 1~3, 更详细的结果分析见附件 1 中的第 2 节.

从敏感度来看, AceView 数据库无论从被匹配的读段数目还是从剪切位点层面都具有最高的敏感度; 而从特异性来看, 在基因层面上, RefSeq, AceView 和 UCSC 都具有很高的特异性; 在转录物和同源异构体层面上, RefSeq 同样具有很高的特异性; 在外显子和剪切位点层面, RefSeq, UCSC 和 Ensemble 都具有很高的特异性, 在碱基层面, UCSC 有非常出色的特异性. 同时, Gencode, Ensemble, UCSC 在转录物、剪切位点层面在敏感性和特异性之间具有较好的权衡. 而 AceView 在碱基层面较好地平衡了敏感性和特异性.

以上是在全转录组范围内对每个基因组注解数据库内所有的注解数据进行了全局的评估, 因此得到的结论在某种程度上会依赖于该数据库注解的基因个数等信息丰度.

1.8 在公共基因范围内的评估

本节将提取 5 个数据库内注解的 9502 个公共基因, 在统一基线上, 去除注释基因的数量对评估结果的影响, 对每个基因的注解进行更加细致和深入地

表 1 转录组范围内基因层面及转录物层面的比较

数据库	匹配读段的百分比(%)	基因			转录物		
		基因数量	读段配准的基因数量	被读段配准的基因(%)	转录物数量	被读段配准的转录物的比例(%)	被读段唯一配准的转录物的比例(%)
AceView	84.15	85684	80606	94.07	258618	97.97	75.11
Gencode	74.77	43422	35655	82.11	132101	91.53	65.36
UCSC	75.40	29367	27699	94.32	65260	96.76	74.17
Ensembl	72.58	36655	27852	75.98	62031	86.90	73.54
RefSeq	66.70	21498	20423	95.00	33675	94.21	81.81

表 2 转录组范围内同源异构体层面、范围内外显子和剪切位点层面的比较

数据库	同源异构体			外显子		剪切位点(共 95460)		
	同源异构体数量	被读段配准的同源异构体的比例(%)	被读段唯一配准的同源异构体的比例(%)	外显子数量	被读段配准的外显子的比例(%)	剪切位点数量	被读段配准的剪切位点的比例(%)	剪切位点敏感度(%)
AceView	258618	97.97	75.11	667875	88.94	377263	73.49	93.83
Gencode	132101	91.53	65.36	421918	87.55	288246	79.81	77.87
UCSC	65260	96.76	74.17	295438	94.42	266979	90.47	81.75
Ensembl	62031	86.90	73.54	272307	90.28	222389	88.51	66.62
RefSeq	33675	94.21	81.81	220217	95.10	200035	93.96	63.61

表 3 转录组范围内碱基层面的比较

数据库	碱基数(bp)	非重叠的碱基的比例(%)	被读段覆盖的非重叠碱基的比例(%)
AceView	344923392	53.36	97.13
Gencode	224738664	42.64	96.85
UCSC	171099779	51.75	98.62
Ensembl	133201223	52.82	93.81
RefSeq	101656644	65.58	96.02

比较和评估. 更详细的结果分析描述和比较图像见附件 1 中的第 3 节和图 2.

这 9502 个基因的覆盖深度中位数从 100 到 1000, 最高可达 10^5, 最低为 10^{-1}. 基因覆盖深度的盒图如图 3 所示.

注解数据库的整体分布趋势分析如下, 具体数值见表 4 和 5, 更详细的描述见附件 1 中的第 3 节.

(ⅰ) 在基因层面上, 5 个数据库的特异性非常接近. 进一步考虑每个数据库中至少包含一个准确注解转录物的基因比例时, AceView 具有最高的特异性, 这与 AceView 的每个基因包含更多的注解转录物有关. 在基因层面上, AceView 与 Gencode 同时具有最大的敏感性和特异性.

(ⅱ) 在转录物层面, 9502 个基因中, AceView 注解了 89092 个转录物, 而 RefSeq 仅注解了 15760 个. 其中, AceView 准确注解了 42413 个完全被支持的转录物, 而 RefSeq 和 UCSC 具有最高的特异性. 除了 Ensembl, 在转录物层面, 特异性随着敏感性的增加而降低.

(ⅲ) 在外显子层面上, 特异性的分布趋势与转录物层面相似. AceView 准确注解了最多的外显子, 但 RefSeq 具有最高的特异性. 在该层面上 UCSC 和 Ensembl 在特异性和敏感性之间达到较好的平衡.

(ⅳ) 在剪切位点层面上, AceView 具有最高的敏感度和最低的特异性, 而 RefSeq 具备最高的特异性和最低的敏感度, 而 Gencode, Ensembl 和 UCSC 在这两个指标之间达到较好的平衡. 该结果与外显子体现的趋势一致.

总体看来, 在转录物、外显子和剪切位点层面, 数据库的特异性随着其敏感性的增加而降低.

从转录组范围内以及公共基因范围内对以上 5 个人类基因组注解数据库的比较中可以看出, 所有比较的结果趋势都较明显, Aceview 的注解最完整, 而 RefSeq 的注解最准确, 而其他 3 个数据库在完整性和准确性之间达到平衡.

1.9 综合准确注解数据库

基于利用人类 16 个组织 RNA-Seq 数据对 5 个人类基因组注解数据库的评估结果, 我们抽取所有数据库中被读段完全支持的转录物建立针对每个组织及整个人体的准确注解基因集合(附件 2). 被读段完全支持是转录物准确注解的强烈证据, 不同数据库中仅起始和结束端不同的转录物被视为同一转录物加以合并. 针对每个组织, 抽取准确注解的基因转录物, 分别建立针对每个组织的准确注解转录物集合. 并基于 RefSeq 中的基因目录, 综合 5 个数据库中的准确注解, 建立人体准确注解集合和转录物参考序列. 该数据库适用于对注解准确度有高要求的生物

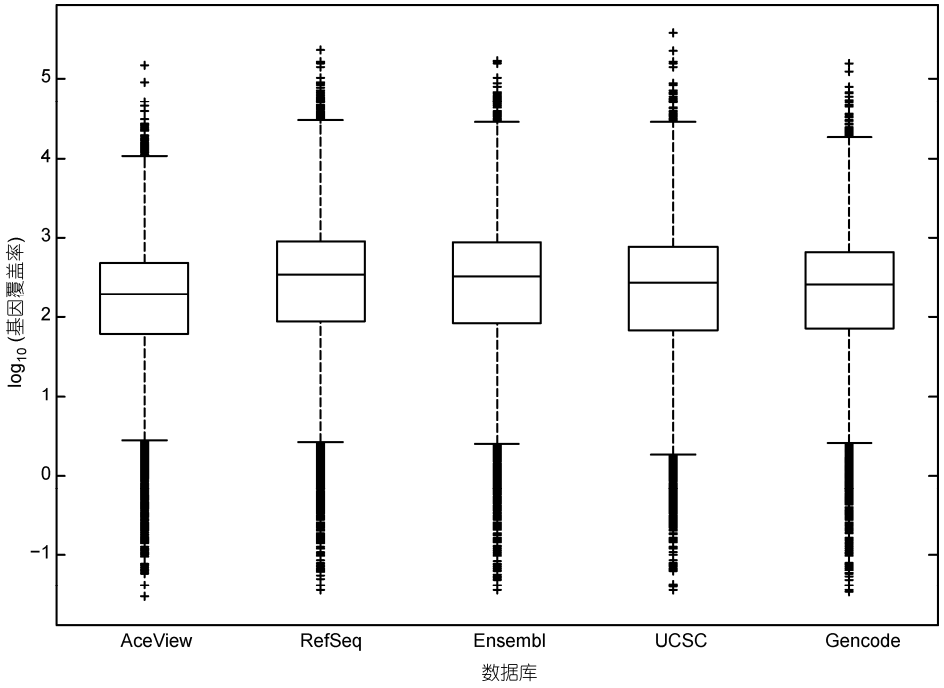

图 3 基因覆盖深度盒图

表 4 9502 个公共基因内基因层面及转录物层面的比较

数据库	匹配读段的百分比(%)	基因			转录物				完全被支持的转录物	
		基因数量	被读段配准的基因数量	被读段配准的基因比例(%)	转录物数量	转录物数量 (1 个外显子)	转录物数量 (2 个外显子)	转录物数量 (3 个以上外显子)	总数	百分比(%)
AceView	39.65	98.64	87.03	39.65	89092	8147	10503	70442	42413	47.61
Gencode	38.05	98.63	86.14	38.05	59942	648	6749	52545	33137	55.28
UCSC	36.71	98.58	84.60	36.71	27109	986	1326	24797	19026	70.18
Ensembl	36.59	98.12	81.90	36.59	25894	647	1155	24092	15536	60.00
RefSeq	36.09	97.71	84.41	36.09	15760	561	770	14429	10956	69.52

表 5 9502 个公共基因内外显子层面的比较

数据库	外显子(共 109153)				剪切位点(共 141178)		
	外显子数量	内部外显子数量	完全被支持的外显子比例(%)	外显子敏感度(%)	剪切位点数量	完全被支持的剪切位点比例(%)	剪切位点敏感度(%)
AceView	271839	117650	86.48	93.21	167045	79.57	94.15
Gencode	132786	95543	91.41	80.01	140173	87.11	83.92
UCSC	123307	95256	92.93	81.10	114156	91.03	73.61
Ensembl	110941	89077	94.37	77.01	115092	92.60	75.49
RefSeq	27839	117650	86.48	93.21	102654	95.15	69.18

研究.

该综合准确注解数据库中包含 16336 个基因、95759 个转录物、324302 个外显子和 209349 个剪切位点. 进一步地, 我们利用来自 MAQC-2 项目的 RNA-Seq 数据集对该数据库的敏感性和特异性进行评估.

从 MAQC-2[18]项目中下载人类 RNA-Seq 数据, 该 RNA-Seq 数据集有 90593559 个 35 bp 的读段, 分别配准到 5 个数据库和综合准确注解数据库. 尽管综

合注解数据库包含的基因个数大大少于其他 5 个数据库(比 RefSeq 少 5000 个), 但仍具有很令人满意的敏感度. 同时, 该数据库无论在基因、转录物、外显子和剪切位点层面都具有非常出色、远高于其他数据库的特异性, 具体数值见表 6 和附件 1 中的图 3. 尽管比 RefSeq 数据库少了 5000 多个基因, 综合准确注解数据库的敏感度仍优于 RefSeq, 而且在特异性方面, 与 5 个数据库相比有非常突出的优势. 因此该综合准确注解数据库很适用于对注解精度具有很高要求的相关生物研究. 该数据库的注解和参考序列文件在附件 3 和 4 下载.

2 对恒河猴基因注解数据库的评估

为了进一步说明基于以上指标对基因组注解的评估作用, 我们对恒河猴的注解数据库 RefSeq Genes 进行评估. 该注解构建于 2006 年, 基于恒河猴基因组(MGSC Merged 1.0/rheMac2), 此后一直没有更新, 所有对恒河猴的研究都基于该注解数据库. 因此利用 RNA-Seq 数据对该注解数据库进行评估是极为必要和重要的. 我们利用长度为 102 bp 的 101326924 条 RNA-Seq 测序数据进行评估, 如表 7 所示.

从注解的数据量来看, 恒河猴的注解数据库包含 24189 个基因、28943 个转录物、205581 个外显子、181256 个剪切位点以及 67461250 个核苷酸碱基. 与人类的 Refseq 注解相比, 被注解的基因数量相当, 其他注解数量略有减少, 其中被注解的碱基数量减少了 30%左右. 由于恒河猴与人类的基因组非常近似, 规模相当, 因此其注解数量具有可比性.

从敏感度分析中, 我们发现相比于人类注解数据库 60%~85%的敏感度, 恒河猴注解数据库的敏感度只有 26.65%, 远远低于人类注解数据库的敏感度水平. 而当把 RNA-Seq 数据配准到恒河猴的基因组时, 敏感度能达到 75%. 由此说明目前恒河猴注解数据库的完整性很低, 有待进一步补充.

从特异性分析中, 恒河猴的特异性在基因、外显子和转录物层面, 其特异性比人类 RefSeq 注解低了 10%, 而在剪切位点层面, 其特异性比人类注解降低 30%, 说明在恒河猴的注解中, 剪切位点的注解精确性尤为薄弱. 而完全被支持转录物和外显子的数目则远远低于人类的相关指标, 除了由于测试的 RNA-Seq 数据量不够充分外, 注解的精确性差是另外一个很主要的原因.

综合以上分析, 恒河猴目前的注解水平的敏感度过低, 特异性相比于人类注解数据库来说, 也相差很大. 说明当前的注解数据库很不完整, 同时精确性也有待进一步完善. 文献[15]通过大量生物实验检查若干区域发现当前恒河猴从基因组测序序列、基因组装配到基因组注解都存在错误的结论一致, 从实验角度支持了本方法得到的结论, 也进一步证明了本文提出的评估框架的有效性.

3 结论与展望

虽然 RNA-Seq 数据本身存在许多的不确定性, 例如单核苷酸多态性(SNPs)、测序制备过程中的质量

表 6 利用 MAQC-2 RNA-Seq 数据的评估

数据库	AceView	RefSeq	Ensemble	UCSC	Gencode	Combinate
匹配的读段(%)	72.23	62.76	65.33	66.59	67.07	62.20
基因	85684	21498	36655	29367	43422	16336
被读段配准的基因	74821	19895	29140	26930	35511	16015
百分比(%)	87.32	92.54	79.50	91.70	81.78	98.04
外显子	667875	220217	272307	295438	421918	324302
被读段配准的外显子	551073	189793	227712	252465	343200	298568
百分比(%)	82.51	86.18	83.62	85.45	81.34	92.06
转录物	258618	33675	62031	65260	132101	95759
被读段配准的转录物	238949	31019	53883	61750	118548	94513
百分比(%)	92.39	92.11	86.86	94.62	89.74	98.70
剪切位点	377263	200035	222389	266979	288246	209349
完全被支持的剪切位点	195682	148778	156021	187002	173857	157613
百分比(%)	51.87	74.38	70.16	70.04	60.32	75.29

表 7 恒河猴基因组数据评估

匹配读段的百分比(%)	基因		外显子		
	基因数量	被读段配准的基因数量	外显子数量	被读段配准的外显子数量	完全被支持的外显子数量
25.63	24189	21744	205581	150767	91841

剪切位点		转录物			
剪切位点数量	完全被支持的剪切位点数量	转录物数量	被读段配准的转录物数量	被读段唯一配准的转录物数量	完全被支持的转录物数量
181256	113779	28943	24189	17731	5956

碱基数(bp)	非重叠的碱基数(bq)	被读段配准的非重叠碱基数(bp)
67461250	49717905	46049049

控制是否到位、样本的组织来源、转录物的表达水平以及读段配准过程中的参数设置等, 都会影响 RNA-Seq 数据对注解数据库的评估结果. 但是这些因素都可以通过评估数据的筛选和评估过程的参数设计加以克服. 通过合理的设置最大容错碱基数目, 可以在保证配准精度的前提下克服单核苷酸多态性和测序偏差, 而有一些配准工具, 例如 bowtie2[19]甚至可以解决具有删除插入错误下的读段配准问题; 目前 RNA-Seq 数据的大量出现, 通过整合同物种多组织样本的测序结果, 可以克服组织特异性对评估带来的偏差; 随着新一代测序技术的不断完善, RNA-Seq 数据的测序结果更加纯化、精确和完整; 对于转录物自身的低表达带来的配准读段过少等问题, 可以通过建模等方式进行弥补和克服, 例如本文中针对完全被支持转录物的定义就引入统计模型解决基因覆盖深度过低情况下的读段分布不连续、不均匀的问题. 通过数据控制、参数控制和方法控制, 使得 RNA-Seq 数据在全基因组范围对基因注释进行自动化的评估成为可能. 同时, RGASP 项目采用 RNA-Seq 作为唯一数据源对基因组进行注解, 也从另一个角度证明了 RNA-Seq 在评估基因注解方面的可靠性.

本文提出了一种利用 RNA-Seq 数据便捷、高效地在全转录组范围内评估蛋白质编码基因的方法, 有效地替代了传统的手工实验验证, 从基因、转录物、外显子、剪切位点和碱基层面对数据库的敏感性和特异性进行分析. 该方法不仅可以高效、快速、低成本地对基因组注解进行评估, 而且很容易应用到其他物种的注解数据库评估上. 虽然该方法无法像 EGASP 那样提供非常详细的细节信息, 但是它的结果给出基因组注解的整体和局部的评估信息, 为后续的研究等提供有效的参考.

对 5 个代表性的人类基因组注解数据库的评估表明, 从整个转录组的注解范围来看, RefSeq 是保守型的数据库, 给出最简明、准确的注解; 而 AceView 是开放型的数据库, 提供更多的信息量, 但假阳率也随之增加. Gencode, UCSC 和 Ensembl 都在敏感性和特异性, 也就是信息的准确性和完整性之间达到均衡. 在相同的公共基因的基线上, 结果显示出非常明显的特异性和敏感性分布趋势, 在转录物、外显子和剪切位点层面上, 数据库的特异性随着敏感性的增加而降低. 敏感性最高的前两位数据库是 AceView 和 Gencode, 特异性最高的前两位数据库是 RefSeq 和 UCSC. 每个数据库都有其各自的针对性和侧重点, 因此可以用于不同的研究目的. 从 5 个数据库中收集准确注解的基因转录物, 组成综合准确注解数据库, 该数据库具有较好的敏感度和非常突出的特异性. 如果需要完整、丰富的信息, AceView 是一个很好的选择, 而当需要准确的注解信息时, RefSeq 和综合准确注解数据库可以提供满意的注解. 如果需要在敏感性和特异性的平衡折衷时, 可以选择 Gencode, Ensembl 和 UCSC. 通过对恒河猴注解数据库的评估发现, 目前该注解数据库很不完整, 同时精确性特别是剪切位点的精确性上与人类基因组注解相比有较大的欠缺, 其精确性有待进一步完善.

致谢 感谢 Illumina 公司的 Gary Shroth 博士为我们提供了高质量的 RNA-Seq 测序数据.

参考文献

1 Carninci P. Is sequencing enlightenment ending the dark age of the transcriptome? Nat Methods, 2009, 6: 711–713
2 Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
3 Mortazavi A, Williams B A, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008, 5: 621–628
4 Jiang H, Wong W H. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics, 2009, 25: 1026–1032
5 Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008, 320: 1344–1349
6 Pan Q, Lee L J, Frey B J, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet, 2008, 40: 1413–1415
7 Sultan M, Schulz M H, Richard H, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008, 321: 956–960
8 Trapnell C, Williams B A, Pertea G. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010, 28: 511–515
9 Montgomery S B, Sammeth M, Guigo R, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature, 2010, 464: 773–777
10 Pickrell J K, Marioni J C, Pai A A, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 2010, 464: 768–772
11 Reese M, Hartzell G, Harris N, et al. Genome annotation assessment in Drosophila melanogaster. Genome Res, 2000, 10: 483–501
12 Reese M G, Guigó R. EGASP: Introduction. Genome Biol, 2006, 7(Suppl 1): S1, doi: 10.1186/gb-2006-7-s1-s1
13 Kokocinski F. RGASP: The RNA-Seq genome annotation assessment project. Technical Report, Wellcome Trust Sanger Institute, United Kingdom. 2010
14 Guigó R, Flicek P, Abril J, et al. EGASP: The human ENCODE genome annotation assessment project. Genome Biol, 2006, 7(Suppl 1): S2, doi: 10.1186/gb-2006-7-s1-s2
15 Zhang X, Goodsell J, Norgren R J. Limitations of the rhesus macaque draft genome assembly and annotation. BMC Genomics, 2012, 13: 206, doi: 10.1186/1471-2164-13-206
16 Chen Y, Souaiaia T, Chen T. PerM: Efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. Bioinformatics, 2009, 25: 2514–2521
17 Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, 10: R25, doi: 10.1186/gb-2009-10-3-r25
18 Bullard J H, Purdom E, Hansen K D, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics, 2010, 11: 94–108
19 Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2. Nat Methods, 2012, 9: 357–359

RNA-Seq-based assessment for genome annotation databases

WANG Ying & LIU Lin

Department of Automation, School of Information Science and Technology, Xiamen University, Xiamen 361000, China

RNA-Seq brings a breakthrough to decode eukaryotic transritptomes. With the high resolution to nucleotide level, RNA-Seq can be adopted as an only data resources to annotate a whole genome. Similarily, RNA-Seq should be able to validate the annotated splicing junction, exon and transcript sets. Therefore, this study proposed an evaluation scheme for the accuracy (specificity) and completeness (sensitivity) of genome annotation databases at gene/transcript/exon/splice-junction/nucleotide base levels with RNA-Seq datasets as only resources. The scheme was applied to assess 5 widely-used human genome annotation databases using 1.1 billion high-quality RNA-Seq reads from 16 human tissues. Accurate-annotated transcripts were collected from the 5 databases to build combined accurate-annotated transcripts databases for the 16 tissues and the whole human body. Furthermore, the assessment for current rhesus annotation database showed that it is far from complete, and not so accurate as Human's annotations. The RNA-Seq analysis pipeline was constructed to implement an express and efficient assessment of various organisms' genome annotations over the whole transcriptome. The implementing pipeline can be downloaded from http://code.google.com/p/genome-annotation-assessment-pipeline/downloads/.

genome, transcriptome, annotation database, RNA-Seq, sensitivity, specificity

doi: 10.1360/972012-1038

补充材料

附件 1: 对 5 个人类基因组注解数据库的详细比较和分布图以及补充的说明和补充的详细数据

http://genome-annotation-assessment-pipeline.googlecode.com/files/supplement.doc

附件 2: 每个人类组织的准确注解转录物集合

http://genome-annotation-assessment-pipeline.googlecode.com/files/supplement3-ValidataedTranscriptsfor16Tissues.rar

附件 3: 综合准确注解数据库

http://genome-annotation-assessment-pipeline.googlecode.com/files/supplement4-combine_annotation_final.txt

附件 4: 综合准确注解转录物参考序列数据库

http://genome-annotation-assessment-pipeline.googlecode.com/files/supplement5-combine_final.zip

附件 5: RNA-Seq 分析软件流

http://genome-annotation-assessment-pipeline.googlecode.com/files/GAAP_ver1.00_unix.zip

本文以上补充材料见网络版 csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.