

DOI: 10.13210/j.cnki.jhmu.20250220.003

网络出版地址: <https://link.cnki.net/urlid/46.1049.R.20250220.1551.006>

TransgeneR:基于测序数据发现基因组整合重组事件的一站式工具

李浩浩¹, 孟国锋²✉, 谭红胜³✉

(1. 香港大学经济及工商管理学院, 中国香港 999077; 2. 上海中医药大学交叉科学研究院, 上海 201203; 3. 上海交通大学医学院, 临床研究中心, 上海 200025)

[摘要] **目的:** 构建一个测序数据分析工具, 发现在转基因动物构建、病毒感染等过程中发生的基因组重组事件。 **方法:** 创建一站式 R 分析工具 TransgeneR, 基于转基因动物或者病毒感染样本的测序数据, 通过两轮序列切割、组装等分析过程, 结合位置、方向、切割位点等信息, 发现外源 DNA 的整合重组事件。在人工模拟的 MAPT 转基因小鼠基因组以及人源 tau (MAPT, 0N4R) 转基因小鼠动物模型上, 评价整合重组事件发现的准确性。 **结果:** 利用模拟动物, TransgeneR 成功发现所有的整合和重组位点, 而且重组位点的数量与切割序列的数量呈线性相关。利用真实转基因动物全基因组测序 (whole genome sequencing, WGS) 数据, TransgeneR 发现 4 号染色体上有一个单一的转基因整合位点 [chr4:129387127 (左侧) 和 chr4:129387416 (右侧)]。 **结论:** 系统性评价表明, TransgeneR 是一个鉴定和研究外源 DNA 整合重组事件的有效工具。

[关键词] 全基因组测序; 基因组重组; 转基因整合位点; DNA 序列

[中图分类号] Q811.4 **[文献标志码]** A **[文章编号]** 2097-5953(2025)18-1408-10

TransgeneR: A one-stop tool for transgene integration and rearrangement discovery using sequencing data

LI Haohao¹, MENG Guofeng²✉, TAN Hongsheng³✉

(1. Faculty of Economics and Business Administration, the University of Hong Kong, Hong Kong 999077, China; 2. Institute of Interdisciplinary Integrative Biomedical Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China; 3. Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China)

[Foundation Project]: This study was supported by Major Project of the "STAR" Program of Shanghai Jiao Tong University (20230101); Shanghai Science and Technology Commission Clinical Research Professional Technical Platform (23DZ2290600)

[Author]: LI Haohao, M.M., E-mail: lihaohao@pku.org.cn.

[Correspondence to]: TAN Hongsheng, M.D., Researcher, E-mail: tanhs@shsmu.edu.cn; MENG Guofeng, M.D., Associate Researcher, E-mail: menggf@gmail.com.

Received: 2024-11-17 Revised: 2025-02-09

JHMu, 2025;31(18):1408-1417

View from specialist: It is creative, and of certain scientific and educational value.

[Abstract] **Objective:** To build a sequencing data analysis tool to discover genome integration and rearrangement events that occur during the construction of transgenic animals, virus infection. **Methods:** A one-stop R analysis tool, TransgeneR was built, which uses sequencing data from transgenic animals or virus-infected samples to discover the integrated rearrangement events of exogenous DNA through two rounds of sequence cleavage, as well as assembly and other analytical processes, incorporating information on position, orientation and cleavage site. The accuracy of integrated rearrangement event discovery was assessed in artificially simulated genomes of MAPT transgenic mice as well as human tau (MAPT, 0N4R) transgenic mice. **Results:** TransgeneR showed that it successfully discovered all the integration sites and rearrangement sites by using simulated animals, and that the number of rearrangement sites appeared to be related linearly with the cutting sequences number. TransgeneR discovered a single transgene integration site on chromosome 4 [chr4:129387127(left side) and chr4:129387416(right side)] by using whole genome sequencing data from real transgenic animals. **Conclusion:** The systematic evaluation shows that TransgeneR is an effective tool to

[基金项目] 上海交通大学“交大之星”(STAR)计划重大项目(20230101);上海市科委临床研究专业技术平台(23DZ2290600)

[作者简介] 李浩浩, 硕士, E-mail: lihaohao@pku.org.cn。

[通讯作者] 谭红胜, 博士, 研究员, E-mail: tanhs@shsmu.edu.cn; 孟国锋, 博士, 副研究员, E-mail: menggf@gmail.com。

[收稿日期] 2024-11-17 **[修回日期]** 2025-02-09 **网络出版时间:** 2025-02-20 17:53:31

identify and study exogenous DNA integration and rearrangements events.

[Key words] Whole genome sequencing; Genetic recombination; Transgene integration sites; DNA sequences

转基因动物,特别是转基因小鼠在药物发现和开发中具有重要的作用^[1,2]。但是,在选择药理学相关实验动物的研究模型解释结果时,通常没有考虑到转基因对动物基因组的影响^[3]。根据国际小鼠品系资源数据库的收集信息,人们已经产生了超过40 000个小鼠品系(截至2022年6月)^[4],但其中绝大多数缺乏对转基因插入事件的描述,尤其是通过前核显微注射产生的转基因动物。转基因的随机整合可能会破坏内源基因,导致部分或完全无法解释的表型(例如,当转基因被培育成纯合子时,胚胎致死)^[5,6]。

同时在病毒感染过程中,病毒将其遗传物质整合到宿主细胞的脱氧核糖核酸(deoxyribo nucleic acid, DNA)中,从而对宿主细胞造成严重后果,包括基因破坏、插入诱变和细胞死亡^[7,8]。如乙型肝炎(hepatitis B virus, HBV)病毒是肝肿瘤发生的主要诱因,超过一半的肝细胞癌患者都存在HBV感染,其原因包括乙肝基因整合、突变引起的基因组不稳定性、促癌信号通路的激活等^[9,10]。在临床样本中检测HBV感染事件、评估HBV感染对宿主基因组潜在影响具有重要的临床应用价值^[11,12]。

目前已经开发了多种实验依赖的方法来研究基因组整合重组事件,包括利用荧光原位杂交定位,聚合酶链式反应(polymerase chain reaction, PCR)等^[13-15]。然而,这些方法都有其自身的局限性和缺点。下一代测序(next-generation sequencing, NGS)分析技术已经开始应用于基因组整合重组事件发现,如目标位点扩增^[16]、VISPA2^[17,18]、CON-TRAILS^[19]、DeepVISP^[20]、Isling^[21]、VIRUSBreak-end^[22]、Exogene^[23]。其中,目标位点扩增是结合实验靶向位点扩增和测序技术研究转基因事件的综合平台,其他计算工具通常利用断裂的测序序列来寻找宿主基因组断点。大多数工具没有考虑转基因序列的多拷贝、转基因缺失/重组和重复区域的整合等复杂性。对宿主基因组中内源性同源序列考虑不足也会干扰结果的解读,甚至有时会识别出虚假的插入位点^[24,25]。另外这些工具普遍没有发布代码或者仅提供在线工具,不适合开展NGS数据依赖的深度分析,限制了应用。

本研究开发出一个新的R软件包-TransgeneR,被设计为一个通用一站式计算工具,基于测序数据发现转基因动物或者病毒插入宿主基因组时所发生的整合和重组事件。该工具采用两轮序列比对,

针对测序reads切割并进行组装,以预测基因整合和重组事件。当全基因组测序(whole genome sequencing, WGS)数据作为输入时,TransgeneR可以预测外源DNA序列在宿主基因组上因为重组而产生的转基因片段使用情况。在应用于模拟转基因小鼠时,TransgeneR成功预测了所有整合位点插入和重组数目,同时过滤了噪声信号。利用真实转基因动物的WGS数据,TransgeneR准确预测了转基因片段,尤其是转基因序列的全长插入及其拷贝。

1 材料与方法

1.1 转基因动物

动物研究方案经动物护理和使用委员会(IA-CUC)审查通过,AUP编码为0084。所有动物均在葛兰素史克公司根据机构动物护理和使用委员会的指导方针进行饲养和处理。动物被饲养在温度和光照可控的隔离设施中,12 h光照/12 h黑暗循环(7:00/19:00),温度和光照都受到控制。在小鼠钙/钙调素依赖性蛋白激酶II α (CaMk II α)启动子驱动下,将具有单一编码变异体(P301L)的0N4R tau异构体的人类cDNA显微注射到C57BL/6J胚胎中,并植入代孕动物体内。筛选出携带hTau cDNA并通过种系传播将转基因稳定传递给后代的创始者,通过蛋白免疫印迹和/或免疫组织化学证实转基因的表达,并将筛选出的品系(TauD35)作为一个群体维持。

1.2 全基因组测序

从尾部提取并纯化转基因小鼠的基因组DNA,根据制造商的说明书(Illumina Hiseq 2500和X10)生成文库进行测序。

合成了覆盖Tau启动子区域的正向引物“GGCCTCCCTGTCCATAGA”和覆盖Tau启动子区域的反向引物“AAGTTCCTCGCCGT-CATC”来扩增整合的cDNA。采用Qiagen试剂盒(Cat#51304)提取小鼠尾部基因组DNA。在95℃变性30 s, 58℃退火30 s, 72℃延伸30 s的条件下进行30个循环的PCR扩增。扩增的DNA用1%琼脂糖凝胶和1 kb DNA片段(Invitrogen Cat#10787018)分离。纯化DNA采用Illumina X10平台进行测序。

1.3 测序数据分析

配对测序(paired-end)数据使用Fastqc进行质量控制分析,并修剪低质量read。使用Short-Reads^[26]中实现的方法过滤重复read。在局部比对

模式下,使用 bowtie2^[27] 将这些 read 与小鼠基因组 (mm10) 和外源 DNA 序列分别进行比对。局部比对模式的特点是允许把测序 read 部分比对到基因组和外源 DNA 序列,即软剪切 (soft clips)。其中这些软剪切位点被认为是整合或重组的潜在位点。相对应的那些能够完全比对的 read,由于不提供额外信息,不再进一步考虑。

为了发现整合重组事件,将不能够比对成功的软剪切部分重新与基因组和外源 DNA 序列比对。在第二轮 read 比对中,确保每个切割 read 都有唯一的映射位置。每个映射片段将被视为从原始映射位置到片段映射位置的一个切割位点证据。

完成两轮序列比对之后,TransgeneR 对对比结果组合分析。如果 read 被分割并分别比对到基因组和外源 DNA 序列,则 read 将被视为整合的证据,这类整合证据的数量,用于表示整合重组事件的可靠性计数。但是如果一个 read 的映射部分重叠,即该 read 的前半部分同时映射到基因组和外源 DNA 序列中,那么根据映射得分、映射一致性和同源物注释,该 read 将被唯一分配到基因组或外源 DNA 序列中。

1.4 同源序列分析

在“1.3”部分的序列比对过程中,经常会出现序列同时被比对到基因组和外源 DNA 的情况,尤其是外源 DNA 和基因组有同源性的时候。为了解决问题,本研究为外源 DNA 建立了与基因组同源性注释信息。外源 DNA 序列被人为分割成 200 bp,允许部分重叠约 50 bp。使用 bowtie2 将所有片段与宿主基因组进行比对,并将外源 DNA 序列的相似度高的部分,组装注释为外源 DNA 序列的同源部分。同源信息,包括在宿主基因组位置及方向都被记录下来,以供后续分析。

1.5 预测评分

切割 read 位点两侧对比打分可以用来表示预测整合和重组的置信度。对于一个切割 read i , 如果其切割位点左右两侧部分的比对得分,分别为 S'_i 和 S''_i , 则该切割 read 在该位点的得分可以描述为:

$$S_i = \sqrt{S'_i \times S''_i}$$

外源 DNA 整合或重组位点的得分是根据通过该切割位点的所有 reads 的平均值计算得出:

$$S = \sum_{i=1}^n S_i / n$$

1.6 外源 DNA 片段使用情况分析

外源 DNA 序列在插入到基因组前,可能发生重组,导致复杂的插入序列,比如多份拷贝,不完整

插入序列等。为了检测外源 DNA 片段潜在构成,TransgeneR 根据测序深度对片段使用情况进行建模。在此过程中,TransgeneR 利用比对到整个染色体上的 read 估算整体测序深度,然后将观察到的外源 DNA 序列的测序深度归一化为拷贝数 (C)。假设 C 值是所有外源 DNA 片段的总和。

为了寻找外源 DNA 片段,TransgeneR 首先通过构建一组外源 DNA 片段 $X = \{x_1, x_2, \dots, x_i\}$, 连接外源 DNA 整合和重组事件的起点 (S) 和终点 (E)。TransgeneR 采用 nnls 软件包中的非负线性模型来拟合 C 值:

$$f(x) = lm(C \sim x_1 + x_2 + \dots + x_i)$$

进一步寻找外源 DNA 片段的整数拷贝,TransgeneR 会选择系数在 1 左右或其他整数值值的片段,并将其作为外源 DNA 片段的初始集,然后添加或删除其他片段,以检查是否可以提高片段拷贝 (Y) 与测序深度 (C) 的相关值 r , 其中 m 为所选片段的数量。

$$Y = \sum_{i=1}^m x_i$$

2 结果

2.1 TransgeneR 概述

TransgeneR 主要通过捕获跨切割位点的测序 read 来识别外源 DNA 整合和重组事件。鉴于外源 DNA 模型可能在宿主基因组中携带同源序列,在宿主基因组中,TransgeneR 会基于序列相似性对同源区域进行注释,当它们被映射到这些区域时,同源信息可用于确定 read 的来源。

图 1(A) 描述了外源 DNA 工具的流程图。使用参数设置相同的 bowtie2 将对末端 read 映射到宿主基因组和外源 DNA 序列上。为了识别切割 read,可通过局部比对算法 (local alignment) 发现软剪切 (soft clips)。当提供全基因组测序数据作为输入时,大部分 read 都来自基因组序列,因此,与基因组精确映射的 read 将被过滤,不再进一步考虑。

进一步 transgeneR 对映射到基因组和外源 DNA 序列的 read 进行第一轮 read 对比。如果一个 read 被分割映射到基因组和外源 DNA 序列中,例如前半部分映射到基因组,后半部分映射到外源 DNA 序列,则该 read 将被视为整合的证据,因此会在该位置加上 1 的计数。如果一个 read 的映射部分重叠,即该 read 的前半部分同时映射到基因组和外源 DNA 序列中,那么根据映射得分、映射一致性和同源物注释,该 read 将被唯一分配到基因组或外源 DNA 序列中。由于外源 DNA 随机整合到宿主基因

组中同源基因位点的概率极其罕见,甚至可以忽略不计,因此如果观察到相同的映射得分,映射到同源区域的 read 将优先分配到整合位点。

然后,将软剪切部分映射 read 作为片段,并重新与基因组和外源 DNA 序列比对。进行第二轮 read 分配,以确保每个片段有唯一的映射位置。每个映射片段将被视为从原始映射位置到片段映射位置的一个切割位点证据。在这一步中,新的映射位于不同的序列上,例如,原始映射位置在宿主基因组上,而片段映射转到基因序列上。因此,可以报道的情况如下:(1)外源 DNA 整合,其中部分映射 read 位于染色体上,其余位于外源 DNA 序列上;(2)基因组重组,其中两部分 read 都被映射到染色体上;(3)外源 DNA 重组,两个部分 read 都映射到外源 DNA 序列上。最后,对具有相同切割位置的 read 进行汇总,以预测外源 DNA 整合和重组位点。在这个过程中,切割 read 的数量是最为关键的标准。只有拆分 read 足够多的位点才会被考虑用于最终的预测。TransgeneR 还会考虑其他因素,如预测位点周围映射 read 的平衡分布、read 的切割模式、预测基因组位置的复杂性和映射质量。在最终输出中,TransgeneR 会报告每个整合或重组位点的多个信息,包括(1)切割、非切割和交叉 read 的数量;(2)组织方向,如正向-正向(ff)、正向-反向(fr)、反向正向(rf)和反向反向(rr);(3)片段间隙;(d)配对的宿主基因组整合位点位置。TransgeneR 还能生成整合位点的可视化 read 比对展示图(图 3A)。

在使用无偏差测序数据时(比如全基因组测序),TransgeneR 可以预测插入的外源 DNA 片段及其拷贝(图 1B)。整合和重组位点首先根据其在宿主基因组中的组织方向,将其分配为起点(S)或终点(E)。所有可能的外源 DNA 片段都是通过连接 S 点和 E 点生成的。TransgeneR 采用一种寻找外源 DNA 片段子集组合的方法,使片段覆盖率与测序深度具有最佳相关性。在这一步骤中,切割 read 和非切割 read 的比例可以用来估算 S 点和 E 点的使用频率,从而确定由相同 S 点或 E 点产生的片段拷贝数。

2.2 利用模拟转基因基因组进行评估

为了评估 TransgeneR 的性能,以 Tg(Camk2a-MAPT*P301L)D35 为转基因外源 DNA 序列,构建了人工转基因小鼠基因组。任意设计转基因组,使其有 3 个转基因整合事件(chr1、chr6 和 chr17),在小鼠基因组中共有 12 个转基因重组片段,6 个整合位点(1-6)和 13 个后排位点(a-k)(图 2A),其中,重组

位点(a)、(c)和(e)出现了两次。模拟的 WGS 数据由 ART^[28]生成,设置为:(1)50 倍读取深度,(2)HiSeq 2500 测序平台,(3)150 bp 成对末端 read。

TransgeneR 分析显示,成功发现所有的整合位点(a-k)和重组位点(1-6),见图 2C。切割 read 的数量被用来衡量对整合或重组位点的置信度。所有真实整合位点的中位数为 31 分,最低分为 17 分,这与 17 号染色体左侧的转基因整合位点(2)有关。在这个人工基因组中,(a)、(c)和(e)有两个重组位点,观察到的切割 read 中位数约为两倍。除了真正的重组事件或整合位点外,TransgeneR 还报告了其他假阳性位点。不过,支持其切割 read 通常少很多,例如只有 3 个切割 read,远远少于真实位点的最低数量。总之,TransgeneR 可以检测到真正的转基因整合位点和重组位点,而且重组位点的数量与切割 read 的数量呈线性相关。此外,TransgeneR 还能恢复插入序列中所有转基因片段的方向,见图 2B、C。

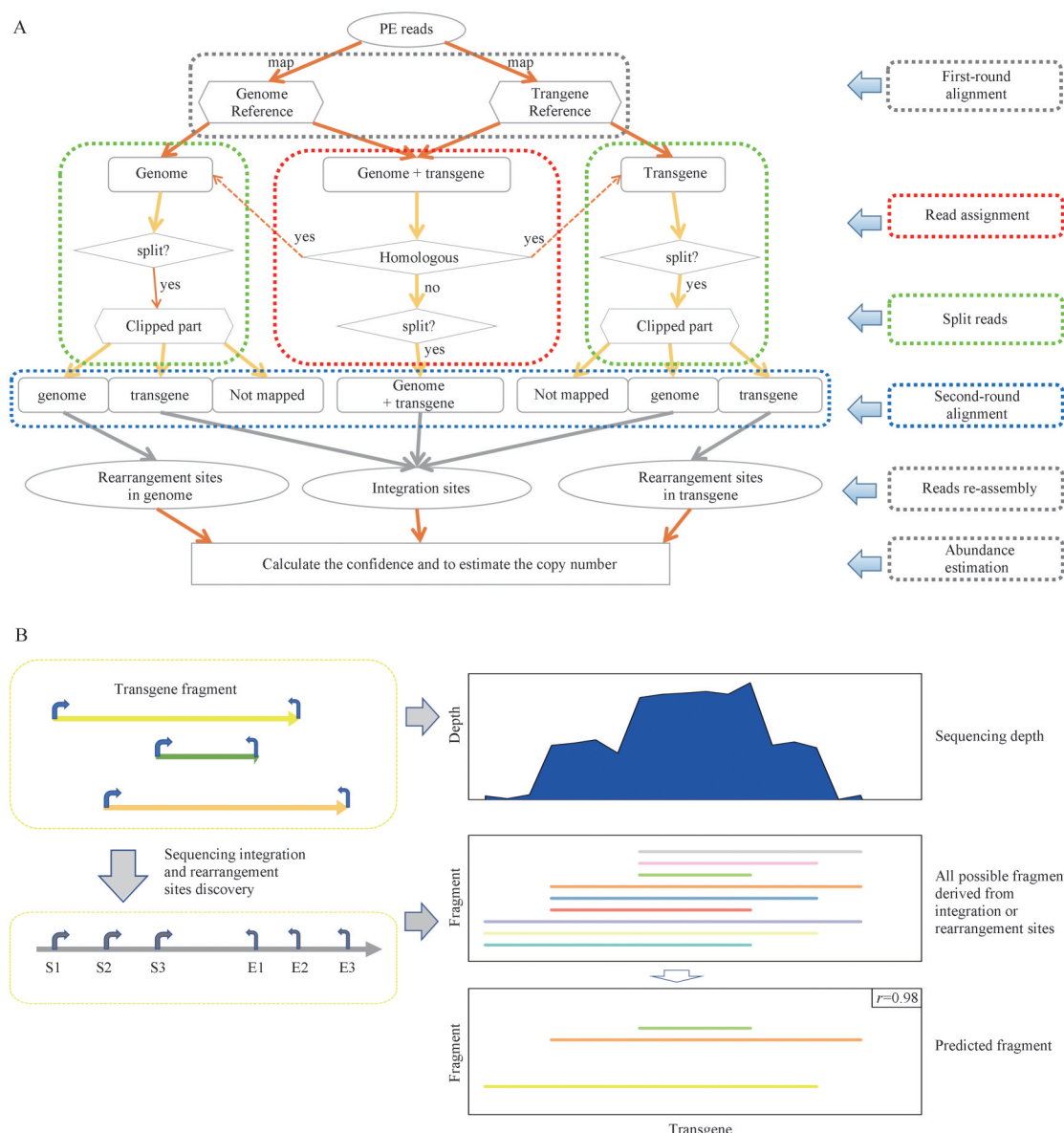
TransgeneR 利用 WGS 数据估算转基因片段长度和转基因组中的拷贝数。在该模型中,根据 WGS 数据计算的覆盖率被建模为转基因片段的总和,而转基因片段是由整合或重组位点的连续组合初始化的。使用模拟基因组,将预测片段及其拷贝相加得到的预测覆盖率与真实覆盖率值的皮尔逊相关性 $r=0.995$ 。图 2D 显示了转基因片段的预测结果。与模拟转基因基因组的真实片段相比,TransgeneR 恢复了 90% 的转基因片段和 100% 的完整转基因序列。

为了评估 transgeneR 的稳健性,将测序深度从 50 倍降低到 5 倍,进行了另一轮评估。图 2E 显示了整合和重组位点发现的准确性。总体而言,TransgeneR 仍能发现真正的整合和重组位点。然而,随着测序深度的降低,TransgeneR 识别错误位点的能力也相应降低。本研究还预测了片段的使用。图 2E 显示了使用预测的转基因片段估计的测序深度与真实测序深度之间的相关值,这表明转基因片段使用的预测准确性对测序深度更为敏感。

这些结果表明,TransgeneR 不仅能准确预测转基因整合和后排位点,还能准确预测转基因片段的使用情况,尤其是完整插入片段。

2.3 利用转基因动物全基因组测序数据评估

克隆了含有 *CaMk II α* 启动子驱动的 P301L 变体的人类 MAPT(0N4R)cDNA,并将其微注射到胚胎中,以产生人类 tau 转基因小鼠。利用转基因特异性聚合酶链反应 PCR 评估创始品系的基因型,以确定种系传递是否成功,并通过 qPCR 确定转基因



A:转基因整合和重组位点发现的流程;B:预测转基因片段使用的方法

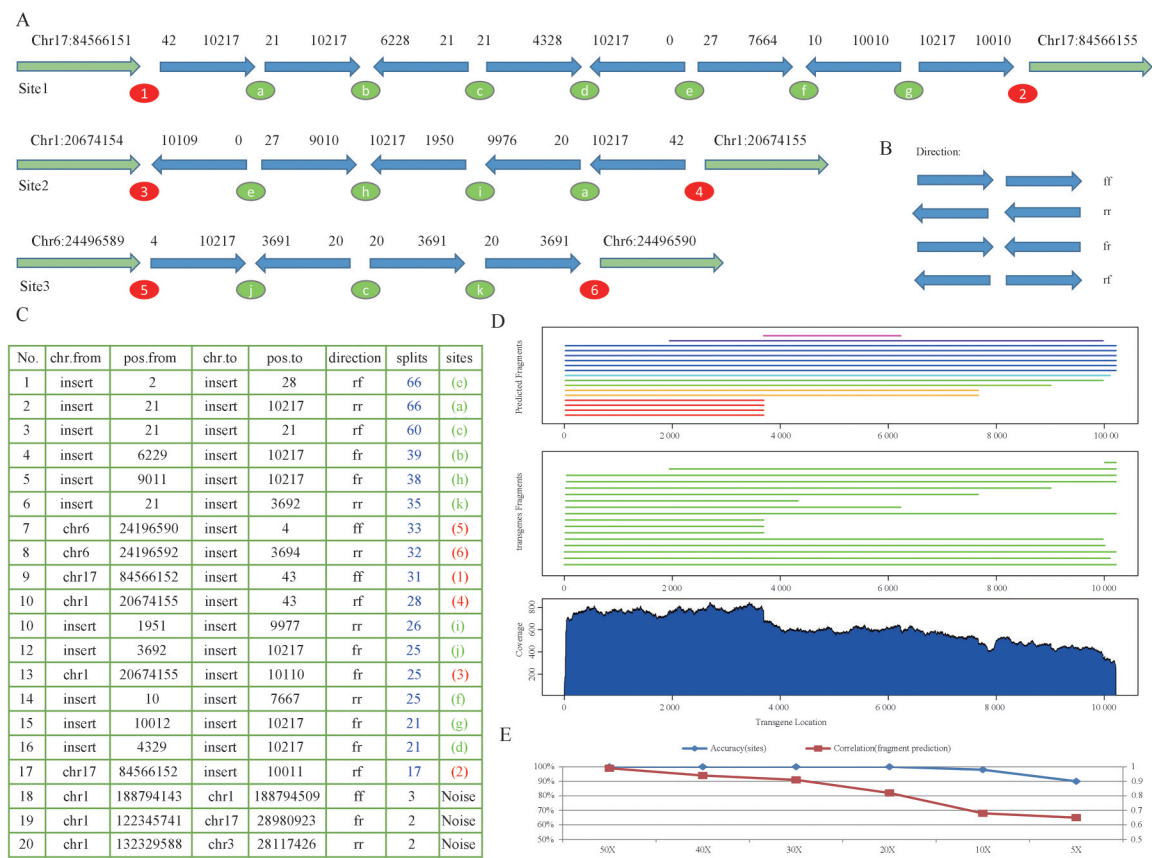
图1 TransgeneR的工作流程

Fig 1 Workflow of TransgeneR

的相对表达水平。其中一个以C57B16/J为背景的品系(在此被命名为TauD35,作为杂合子保留),通过HiSeq X10进行了全基因组测序分析。

在对重复 read 进行质量控制和过滤后,应用TransgeneR对转基因整合和重组进行分析。实际测序深度是通过分析1号染色体估算的,平均深度为49.5倍。考虑到转基因小鼠是杂合子,基于此覆盖率的预期切割 read 不超过25个。与人工模拟的WGS数据不同,实验得出的WGS数据的分析将面临一定的挑战,尤其是在重复区域。因此,不可能只以切割 read 的数量作为选择转基因整合位点的

唯一标准。因此,TransgeneR还考虑了切割 read 在预测整合位点上的分布是否均衡,即只在一侧映射有剪切片段的整合位点不会被选为真正的整合位点。此评估工作还发现,考虑到映射基因组区域的DNA序列复杂性,这是一种过滤错误预测的有效方法。TransgeneR通过检查映射 read 的唯一性来评估基因组的复杂性。映射到同一类型复杂区域,如LINE的切割 read 会被过滤为整合或重组位点发现。所有这些策略都有助于识别真正的整合位点。对测序数据的分析表明,4号染色体上有一个单一的转基因整合位点[chr4:129387127(左侧)和chr4:



A:模拟小鼠转基因基因组,包括chr17、chr1和chr6上的3个转基因整合;B:使用4个双字母短语定义转基因序列组织方向;C:预测转基因整合和重组位点,包括基因组/转基因位置(chr.from, pos.from, chr.to和 pos.to)、方向(direction)和支持读段数(splits);D:预测的转基因片段使用情况

图2 使用模拟WGS数据的计算评估
Fig 2 Computational evaluation using simulated WGS data

129387416(右侧)],见图3A。左侧和右侧分别有19个切割read和18个切割read支持。此外,还有8个read穿过该整合位点。

此外,transgeneR还在基因组中发现了大约20个转基因重组位点,其中大部分位于完整转基因序列的头尾方向。通过对映射到转基因序列上的read进行计算,相当于约16份转基因序列。TransgeneR预测了转基因重组产生的片段,见图3B。其中,完整拷贝8个,其余为部分拷贝。在测序深度图上,有两个区域的read覆盖率较高,分别在1200和9500bp左右。其中的技术原因目前尚不清楚。它们似乎是转基因重组的热点区域。TransgeneR预测这两个区域有许多短片段。

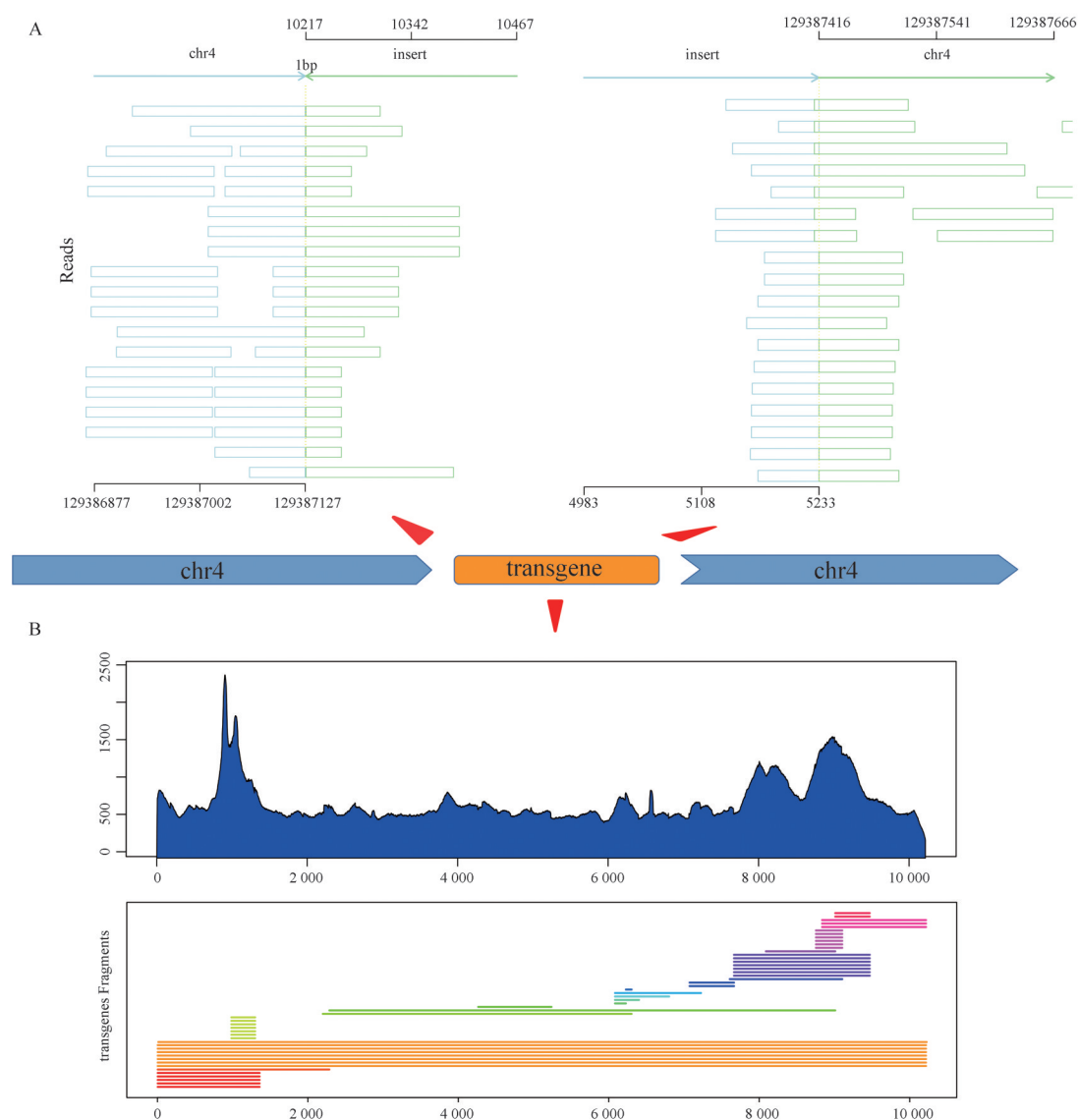
2.4 TransgeneR在基于PCR测序数据的实用性

基于PCR的测序方法,利用转基因特异引物与宿主基因组退化引物绘制转基因插入位点图。该方法还应用于TauD35小鼠分离的尾部DNA,使用

MiSeq进行测序,并用targetR进行分析,以评估该软件包与使用不同方法获得的序列的兼容性。与WGS数据不同的是,大部分的切割read(>95%)均映射到转基因重组位点,尤其是完整转基因序列的首尾重组位点。这在意料之中,并且反映了转基因的靶向富集,因为序列特异性引物与退化引物结合使用。表1列出了TransgeneR预测的重组位点,其中大部分与WGS分析结果一致。有趣的是,除了重组位点的识别外,4号染色体上的整合位点也被观察到了,尽管只有一个切割read,尽管其原因目前尚不清楚。

2.5 性能评估

首先,横向对比了TransgeneR和已发表软件的性能对比。为了更好反映不同场景下的重组整合事件,利用模拟全基因组测序数据为测试数据。测试软件包括Exogene, VISPA2, CONTRAILS, DeepVISP, Isling, VIRUSBREKEND等。通过公共



A: 预测的转基因整合位点(包括左侧和右侧位点)和可视化的切割 read; B: 预测的转基因片段使用情况

图3 将 TransgeneR 应用于 WGS 数据

Fig 3 Application of TransgeneR to the WGS data analysis

来源方式下载并安装这些软件。其中,CON-TRAILS 没有公共来源的代码可以下载,没有分析结果用于评估。另外,DeepVISP 是在线版,并且不支持测序数据的输入,设定其输出为 0。表 2 展示了分析结果。TransgeneR 在本测试数据中,表现最佳,有较高的灵敏度和准确性。通过深入分析发现,其他测试软件由于缺乏对同源序列的分析策略,对转基因插入序列中的 read 没法有效判别来源,导致经常随机预测为插入 DNA 来源或基因组来源,降低了预测的准确性。

另外,TransgeneR 工具提供了灵活使用计算能力的选项。几乎整个分析过程都支持多线程计算,

尤其是计算量大的部分。用户可以设置“core”参数来定义使用多少线程,以及“buffer size”选项来调整内存使用量。“buffer size”越小,需要的内存就越少。本研究测试了 TransgeneR 在 4 核(8 线程)AMD Ryzen 1400 CPU 和 16 G ROM 计算机上的性能。当使用 6 个线程时,分析小鼠基因组的 50 倍测序深度的 WGS 数据,平均运行时间约为 11.5 h(不包括 bowtie2 读取比对的时间),在“buffer size”设置为 500 000 时,最大内存使用量约为 12 GB。本研究还使用基于 PCR 扩增的测序数据对该工具进行了评估,这些数据大约有 50~200 万个 read。平均分析时间小于 6 min。在分析过程中,会生成许多临时文

表 1 利用 WGS 和基于 PCR 的测序数据预测的整合和重组位点
Tab 1 Predicted integration and rearrangement sites using WGS and PCR-based sequencing data

PCR-amplification						WGS					
chr.from	pos.from	chr.to	pos.to	dir	reads	chr.from	pos.from	chr.to	pos.to	dir	reads
transgene	2	transgene	10 217	rr	1 739	transgene	3	transgene	10 217	rr	1 556
transgene	4	transgene	4 328	rf	1 726	transgene	4	transgene	4 325	rf	153
transgene	2	transgene	7 664	rr	1 483	transgene	3	transgene	7 666	rr	147
transgene	2	transgene	9 010	rr	1 403	transgene	3	transgene	9 010	rr	149
transgene	4	transgene	6 228	rf	1 124	transgene	4	transgene	6 228	rf	159
transgene	4	transgene	2 223	rf	1 070						
transgene	4	transgene	1 950	rf	633	transgene	4	transgene	1 950	rf	154
transgene	4	transgene	93	rf	304	transgene	4	transgene	74	rf	149
transgene	52	transgene	10 106	rf	170						
transgene	6 402	transgene	7 075	ff	118	transgene	6 402	transgene	7 075	ff	149
transgene	4	transgene	70	rf	97	transgene	4	transgene	74	rf	149
transgene	152	transgene	8 896	rf	87						
transgene	4	transgene	128	rf	86						
transgene	152	transgene	8 912	rf	81						
transgene	4	transgene	10	rf	77						
chr4	129 387 127	transgene	10 217	fr	19						

表 2 不同软件与 TransgeneR 的性能对比
Tab 2 Comparison of different software performance with TransgeneR

	真实位点	错误位点	结果分析	使用方便程度
TransgeneR	17	2	发现了全部整合位点;错误位点打分偏低	方便,github R 源码安装
Exogene	9	18	由于缺乏同源分析的策略,多个整合位点被预测为鼠源基因组重组事件	通过 docker 使用;国内使用有一定难度
VISPA2	11	19	由于缺乏同源分析的策略,多个整合位点被预测为鼠源基因组重组事件	方便,github python 源码安装
CONTRAILS	/	/	没有通过公共来源发现源代,未评估	没有通过公共来源发现源代
DeepVISP	0	0	通过病毒整合事件训练深度学习模型,不适用于更多病毒或复杂转基因整合事件	在线工具
Isling	15	7	主要针对病毒整合事件开发,考虑了不明确重组事件	方便,github python 源码安装
VIRUSBreak-end	16	10	主要针对病毒整合事件开发,考虑了不明确重组事件	整合在 GRIDSS 软件包中,安装方便

件作为输出,这可能需要很大的磁盘空间。用户也可以设置压缩输入的 .fastq 文件和临时文件,这样可以节省约 70% 左右的存储空间。

3 讨论

TransgeneR 是一款用于发现外源 DNA 整合位点和重组事件的通用性集成工具(参见软件包文档 <https://github.com/menggf/transgeneR/>)。适用于不同方法生成的测序数据,即全基因组测序和基于扩增的测序,而且只有一个统一 R 函数作为接口。如果用户遇到错误或需要调整参数,TransgeneR 支持对整个流程部分进行分析,将每一步的分析结果输出为可读文件,还可以自动判断每个步骤的输出结果,并决定哪些步骤已经执行。

评估结果表明,TransgeneR 在恢复模拟基因组序列和实验得出的基因组序列中的所有外源 DNA 整合和后伸位点方面都有很好的表现。TransgeneR 已经过优化,可根据一系列标准输出最终结果,这些标准包括切割 read 的数量、预测位点周围的 read 分布以及预测区域的 DNA 复杂性。然而,外源 DNA 事件的复杂情况可能会导致错误的预测。例如,低复杂区域的 read 分布存在偏差,导致不存在唯一的阈值来确定真正的外源 DNA 整合位点。在许多情况下,需要人工评估才能得到可靠的结论。作为支持,TransgeneR 输出提供了一些必要的信息,包括该工具可用于测序深度和分割 read 的基因组学可视化。

这个软件包最初是为鉴定外源DNA动物的基因组修饰而设计的。然而,这一策略也适用于其他与DNA整合相关的研究。其中一个例子是预测宿主基因组中的病毒整合位点。这可以通过修改一些参数来实现,以调整bowtie2对病毒基因组中错配的耐受性来实现。在其他研究中,充分证明TransgeneR可以用于HBV整合和重组事件的发现^[29]。

虽然TransgeneR工具易于使用,并且其在新产生的TauD35小鼠中的应用通过WGS和基于PCR的方法确定了Chr4中的hMAPT插入位点,但也存在一定的局限性:(1)预测的位点没有通过其他方法进行实验验证,如使用序列特异性引物扩增整个基因组插入位点,跨越切割read,然后使用NGS来确认和确定插入位点;(2)未对整合外源DNA的宿主基因的表达水平及其对全基因组表达的影响进行实验验证;(3)TransgeneR对于发生的基因组重复区域的重组整合事件没有开展足够评估,没法确保其输出结果的准确性。未来的研究将会解决这个问题。

作者贡献度说明:

李浩浩:清洗整理测序数据、撰写论文;孟国锋:数据分析及建模;谭红胜:修改论文,支持研究经费。

所有作者声明不存在利益冲突关系。

参考文献

- Dunn DA, Pinkert CA, Kooyman DL. Foundation review: Transgenic animals and their impact on the drug discovery industry [J]. *Drug Discov Today*, 2005, 10: 757-767.
- Shakweer WME, Krivoruchko AY, Dessouki SM, et al. A review of transgenic animal techniques and their applications [J]. *J Genet Eng Biotechnol*, 2023, 21: 55.
- Yoshiki A, Ballard G, Perez AV. Genetic quality: A complex issue for experimental study reproducibility [J]. *Transgenic Res*, 2022, 31: 413-430.
- Eppig JT, Motenko H, Richardson JE, et al. The international mouse strain resource (IMSR): Cataloging worldwide mouse and ES cell line resources [J]. *Mamm Genome*, 2015, 26: 448-455.
- Yoshihara M, Sato T, Saito D, et al. A deletion in the intergenic region upstream of Ednrb causes head spot in the rat strain KFRS4/Kyo [J]. *BMC Genet*, 2017, 18:29.
- Goodwin LO, Splinter E, Davis TL, et al. Large-scale discovery of mouse transgenic integration sites reveals frequent structural variation and insertional mutagenesis [J]. *Genome Res*, 2019, 29:494-505.
- Desfarges S, Ciuffi A. Viral Integration and consequences on host gene expression [M]. *Viruses: Essential Agents of Life*, 2012, 25:147-175.
- Wu P, Ma D, Wu P. Oncogenic virus integration: Moving toward clinical applications [J]. *Med*, 2023, 4: 347-352.
- Pollicino T, Caminiti G. HBV-integration studies in the clinic: Role in the natural history of infection [J]. *Viruses*, 2021,13: 368.
- Levrero M, Zucman-Rossi J. Mechanisms of HBV-induced hepatocellular carcinoma [J]. *J Hepatol*, 2016, 64: S84-S101.
- Jiang Y, Han QJ, Zhao HJ, et al. The mechanisms of HBV-induced hepatocellular carcinoma [J]. *J Hepatocell Carcinoma*, 2021, 8: 435-450.
- Abdelhamed W, El-Kassas M. Hepatitis B virus as a risk factor for hepatocellular carcinoma: There is still much work to do [J]. *Liver Res*, 2024, 8: 83-90.
- Schwarzacher T. Fluorescent in situ hybridization to detect transgene integration into plant genomes [J]. *Methods Mol Biol*, 2009, 478: 227-246.
- Chen S, Evans GA. Use of polymerase chain reaction for screening transgenic mice [J]. *Methods Mol Biol*, 1993, 15: 75-80.
- Hernández M, Pla M, Esteve T, et al. A specific real-time quantitative PCR detection system for event MON810 in maize YieldGard based on the 3'-transgene integration sequence [J]. *Transgenic Res*, 2003, 12: 179-189.
- Cain-Hom C, Splinter E, Min MV, et al. Efficient mapping of transgene integration sites and local structural changes in Cre transgenic mice using targeted locus amplification [J]. *Nucleic Acids Res*, 2017, 45: e62.
- Calabria A, Leo S, Benedicenti F, et al. VISPA: A computational pipeline for the identification and analysis of genomic vector integration sites [J]. *Genome Med*, 2014, 6: 67.
- Spinozzi G, Calabria A, Brasca S, et al. VISPA2: A scalable pipeline for high-throughput identification and annotation of vector integration sites [J]. *BMC Bioinformatics*, 2017,18: 520.
- Lambirth KC, Whaley AM, Schlueter JA, et al. Con-trails: A tool for rapid identification of transgene integra-

- tion sites in complex, repetitive genomes using low-coverage paired-end sequencing [J]. *Genom Data*, 2015, 6: 175-181.
- 20 Xu HD, Jia PL, Zhao ZM. DeepVISP: Deep learning for virus site integration prediction and motif discovery [J]. *Adv Sci (Weinh)*, 2021, 8: 2004958.
- 21 Scott S, Hallwirth CV, Hartkopf F, et al. Isling: A tool for detecting integration of wild-type viruses and clinical vectors[J]. *J Mol Biol*, 2022, 434:167408.
- 22 Cameron DL, Jacobs N, Roepman P, et al. VIRUS-Breakend: Viral integration recognition using single breakends[J]. *Bioinformatics*, 2021, 37: 3115-3119.
- 23 Stephens Z, O'Brien D, Dehankar M, et al. Exogene: A performant workflow for detecting viral integrations from paired-end next-generation sequencing data [J]. *PLoS One*, 2021, 16: e0250915.
- 24 Meng GF, Xu H, Lu D, et al. Three-dimensional chromatin architecture datasets for aging and Alzheimer's disease[J]. *Sci Data* 2023, 10: 51.
- 25 Hayut SF, Bessudo CM, Levy AA. Targeted recombination between homologous chromosomes for precise breeding in tomato[J]. *Nat Commun*, 2017, 8: 15605.
- 26 Morgan M, Anders S, Lawrence M, et al. ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data [J]. *Bioinformatics*, 2009, 25: 2607-2608.
- 27 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2[J]. *Nat Methods*, 2012; 9: 357-359.
- 28 Huang WC, Li LP, Myers JR, et al. ART: A next-generation sequencing read simulator [J]. *Bioinformatics*, 2012, 28: 593-594.
- 29 Meng GF, Tan Y, Fan Y, et al. TSD: A computational tool to study the complex structural variants using pacbio targeted sequencing data [J]. *G3 (Bethesda)*, 2019, 9: 1371-1376.

[本文编辑] 雷 燕 宋睿璞 朱金月