

基于逆转座子插入与缺失的系统发育基因组学重建

Phylogenomic Reconstruction with Insertions and Deletions of Retrotransposon

刘高鸣，周旭明*

中国科学院动物生态与保护生物学院重点实验室，中国科学院动物研究所，北京 100101

*通讯作者邮箱: zhouxuming@ioz.ac.cn

引用格式: 刘高鸣，周旭明. (2021). 基于逆转座子插入与缺失的系统发育基因组学重建. **Bio-101 e1010619. Doi: 10.21769/BioProtoc. 1010619.**

How to cite: Liu, G. M. and Zhou, X. M. (2021). Phylogenomic Reconstruction with Insertions and Deletions of Retrotransposon. Bio-101 e1010619. Doi: 10.21769/BioProtoc. 1010619. (in Chinese)

摘要: 逆转座子作为系统发生重建的分子标记，近年来在高级阶元物种的系统发育基因组学研究中应用广泛。目前的基因组序列大部分是基于第二代高通量测序技术测序和相关算法组装完成，短片段序列的组装引起基因组比对精确性下降以及逆转座子假阳性率过高等问题。本研究介绍一种利用生物信息搜索结合传统 Sanger 测序验证从而提高逆转座子鉴定效率的方法。数据分析方面可以概括为四步：(1) 枚举研究物种所有可能的系统发生关系及对应的逆转座子插缺信息；(2) 下载目标基因组的序列和注释信息，并进行全基因组的比对；(3) 根据注释信息和逆转座子的插缺情况，抽取逆转座子的块文件；(4) 依据逆转座子的判断标准，确定潜在的逆转座子。实验方法方面主要流程如下，首先用逆转座子的侧翼序列为参照，设计引物，然后用 PCR 扩增，成功的产物 Sanger 测序，得到存在 (+)/缺失 (-) 矩阵，最后构建最大简约树和网状树，并检测不同拓扑结构之间差异的显著性。综上所述，本方法的数据分析流程和实验操作较为简单，是一种易于掌握的逆转座子鉴定及系统发生关系构建的方法，适用于基因组水平的逆转座子高级阶元物种系统发育基因组学研究。

关键词: 逆转座子，系统发育基因组学，基因组比对，插入和缺失，最大简约树

研究背景

逆转座子 (Retrotransposon), 又称逆转录转座子或者反转座子, 是指通过逆转录的方式插入到寄主基因组并在宿主基因组中稳定遗传的一类基因组元件。由于插入过程不可逆, 在不同物种基因组上同一位置插入相同逆转录转座子的事件在不同时间点独立发生两次的概率几乎为零, 这类元件被认为是高级阶元 (比如, 科目级阶元) 系统发生信号研究的理想分子标记 (Hamdi *et al.*, 1999)。例如, Nikaido *et al.* (1999) 应用逆转座子解析哺乳类的系统发生关系, 发现鲸类和偶蹄目 (Artiodactyla) 的河马科是姐妹群, 支持鲸偶蹄目 (Cetartiodactyla) 单系群的关系。Doronina *et al.* (2015) 通过逆转座子信息厘定熊科 (Ursidae) 在熊超科 (Ursoidae) 基部的系统发生地位。Doronina *et al.* (2017) 基于逆转座子的方法重构劳亚兽类群的系统发生关系, 揭示鲸偶蹄目、奇蹄目 (Perissodactyla)、翼手目 (Chiroptera) 和食肉目 (Carnivora) 是网状树的关系, 猜测这些类群的祖先类群可能发生了杂交事件。

随着高通量测序技术的快速发展, 利用基因组的序列变异进行系统发生重建越来越常见。但目前常用的高通量测序平台, 尤其是第二代高通量测序平台的读长较短, 常出现二代基因组组装断裂的现象, 另外, 由于逆转座子序列的差异性大, 基因组比对的相似性低, 这两者都会导致仅依赖生物信息的方法搜索基因组中潜在的逆转座子分子标记的假阳性率高。比如, Hartig *et al.* (2013) 在探究眼镜猴科 (Tarsiidae) 在灵长目 (Primates) 中的位置时, 生物信息学分析搜索到的转座子有 480 个, 但其中仅有 104 个经过 PCR (Polymerase Chain Reaction) 扩增验证以后确认有系统发生信号。类似的, 相关研究在劳亚兽类动物的基因组中搜索到的逆转座子有 243 个, 但是 PCR 扩增比对之后只有 102 个标记被认为具有有效的插缺信息 (Doronina *et al.*, 2017)。因此, 应用逆转座子信息进行系统发育重建之前, 必须使用 PCR 扩增验证潜在逆转座子插缺信息的有效性, 降低假阳性率。

本论文提供一套完整的数据分析流程和实验验证过程, 操作简单, 不仅适用于哺乳动物的系统发生关系研究, 还可以应用于鸟类等脊椎生物 (Suh *et al.*, 2011)。该方案的有效性已经在灵长目和食肉目科级阶元以及劳亚兽类目级阶元的分子生物学研究中得以验证, 详细-操作请参阅 Hartig *et al.* (2013)、Doronina *et al.* (2015) 和 Doronina *et al.* (2017)。

材料与试剂

1. 200 μ l 和 2 ml 离心管 (BBI Life Sciences)
2. 各种型号移液枪枪头 (10 μ l, 100 μ l, 200 μ l, 1 ml, BBI Life Sciences)
3. PCR 用八连管 (BBI Life Sciences)
4. DNA 提取试剂盒 (DNeasy® Blood & Tissue Kit, QIAGEN, catalog number: 69504)
5. PCR 扩增的 DNA 聚合酶 Mix (2x Es Taq MasterMix (Dye), CWBIO, catalog number: CW0690L)
6. 双蒸水 (喀斯玛中国科学院生物试剂仓储派送)
7. 琼脂粉 (AGAROSE-G10, BIOWEST, catalog number: 111935)
8. DNA 分子 Marker (两种, D2000, TIANGEN, catalog number: MD114-02; D15000, TIANGEN, catalog number: MD110-4)
9. 核酸燃料 (GelRed, BIORIGIN, catalog number: BN20292)
10. 乙二胺四乙酸二钠 (Na_2EDTA) (国药集团化学试剂有限公司)
11. 冰醋酸 (北京伊诺凯科技有限公司)
12. Tris (北京中生达生物技术有限公司)

仪器设备

1. 涡旋仪 (上海沪析实业有限公司)
2. 离心机 (赛默飞世尔科技 (中国) 有限公司)
3. 小离心机 (大龙兴创实验仪器 (北京) 股份公司)
4. PCR 仪 (美国应用生物系统公司)
5. 灭菌锅 (上海博讯医疗生物仪器股份有限公司)
6. 电泳仪 (北京六一生物科技有限公司)
7. 制胶槽和梳子 (北京六一生物科技有限公司)
8. 凝胶成像系统 (Tanon 1600 天能全自动数码凝胶图像分析系统)
9. 恒温水浴锅 (上海禾汽玻璃仪器有限公司)
10. 电子天平 (上海禾汽玻璃仪器有限公司)
11. 微波炉 (佛山市顺德区美的微波电器制造有限公司)
12. 移液枪一套 (赛默飞世尔科技 (中国) 有限公司)

13. 大型服务器 (Linux 系统, 196 G 内存, 4 T 存储, 戴尔 (中国) 有限公司)

软件

1. RepeatMasker v4.0.5 (<http://www.repeatmasker.org>)
2. LASTZ v1.02.00 (<https://github.com/lastz/lastz.git>)
3. MULTIZ (<https://github.com/multiz/multiz.git>)
4. MafFilter v1.3.1 (<https://github.com/jydu/maffilter.git>)
5. Perl v5.32.1 (<https://www.cpan.org/src/5.0/perl-5.32.1.tar.gz>)
6. UCSC2FASTA (<http://retrogenomics.uni-muenster.de/tools/UCSC2FASTA/index.html?>)
7. MAFFT v7.471 (<https://github.com/GSLBiotech/mafft.git>)
8. PHYLIP v3.697 (<https://evolution.gs.washington.edu/phylip/download/phylip-3.697.tar.gz>)
9. SplitsTree v5.2.25 (https://software-ab.informatik.uni-tuebingen.de/download/splits/tree5/SplitsTree5_unix_5_2_25-beta.sh)
10. R 语言 (<https://mirrors.tuna.tsinghua.edu.cn/CRAN/src/base/R-4/R-4.0.4.tar.gz>)
11. KKSC 检验 (http://retrogenomics.uni-muenster.de:3838/KKSC_significance_test/)
12. PriFi (<http://cgi-www.daimi.au.dk/cgi-chili/PriFi/main>)

生物信息分析方法

逆转座子根据其嵌入寄主基因组的形式, 可以分为长末端序列 (Long Terminal Repeat, LTR)、内源性逆转录病毒和非长末端序列 (Non-Long Terminal Repeat, non-LTR)。其中的非长末端序列逆转录转座子, 根据序列特征, 又可分为长散核元件 (Long Interspersed Nuclear Element, LINE) 和短散核元件 (Short Interspersed Nuclear Element, SINE)。我们这里的系统发生分子标记, 主要是 LTR、LINE 和 SINE。

目前, 逆转座子的系统发育基因组学的方法主要应用于三个或者四个物种的研究, 这里以四个物种的逆转座子存在和缺失信息为例, 概述逆转座子的系统发育基因组学分析的流程。

1. 假定四种物种的系统发生关系, 物种分别用 A、B、C、D 表示, 用逆转录转座子的插入 (+)/缺失 (-) 信息描述其关系, 共有 10 种插缺状态如表 1 所示。

表 1. 四物种的转座子插入/缺失状态

	1	2	3	4	5	6	7	8	9	10
A	+	+	+	-	-	-	+	+	+	-
B	+	-	-	+	+	-	+	+	-	+
C	-	+	-	+	-	+	+	-	+	+
D	-	-	+	-	+	+	-	+	+	+

2. 下载 UCSC 四物种的基因组序列和 RepeatMasker 注释的重复信息 (包含 LTR、LINE 和 SINE 的位置信息)。

RepeatMasker -parallel 30 -specie A -html -gff -dir repeat A.fa

3. 选取物种 A 为参考基因组, 用 LASTZ 进行全基因组的两两比对, 然后基于物种 A 的位置信息, 用 MULTIZ 合并两两物种的比对文件为多物种的 maf 文件 (图 1A-C)。

lastz A.fa B.fa --notransition --step=20 --format=maf > A_B.maf

multiz A_B.maf A_C.maf v1 A_B_C.out1_1 A_B_C.out2_1 > A_B_C.maf

4. 根据物种的 RepeatMasker 报告的 LTR、LINE 和 SINE 的位置信息, mafFilter 提取 maf 文件中有逆转座子序列及质量信息的块文件。
5. 判断序列的存在/缺失信息, 依据的标准主要是: 1) 逆转座子元件的最短长度是 50 nucleotide, nt; 2) 逆转座子元件的最长长度是 6000 nt; 3) 序列两两比较之间的分化程度不能高于 30% (即相似度高于 70%); 4) 逆转座子缺失的最短长度是 30 nt 并且一个逆转座子的序列长度至少 70% 是缺失的。

6. 用 JAVA 语言撰写的 UCSC2FASTA 把 maf 文件转换成 fasta 格式 (图 1D)。

在 UCSC2FASTA.ini 中添加 species=A,B,C,D

java -Xmx=1024m -jar ucsc2fasta.jar

7. 提取获得的潜在逆转座子序列用 MAFFT 重新比对, 并且做人工矫正。选取具有存在/缺失信息的逆转座子且侧翼序列约 500 nt 的序列作为备选的后续系统发生信号的分子标记。

mafft A_B_C_D.fa >A_B_C_D_mafft.fa

8. 根据 PCR 验证之后的比对序列, 构建逆转座子的存在/缺失 (1/0) 数据集 (如表 2), 用于 PHYLIP 软件构建最大简约树, 同时用 SplitsTree 构建邻近-网状树 (neighbor-net), bootstrap 值设置为 1000 (如图 2)。

9. 三个物种 (或者枝系) 的系统发生拓扑结构差异的检验, 用 R 语言撰写的 KKSC 显

著性检验。而四个物种 (或者枝系) 拓扑结构差异性检验, 是在 KKSC 检验的基础上用扩散过程去建模描述拓扑结构, 而后用似然比的卡方检验研究拓扑结构间的差异显著性 (如图 2)。

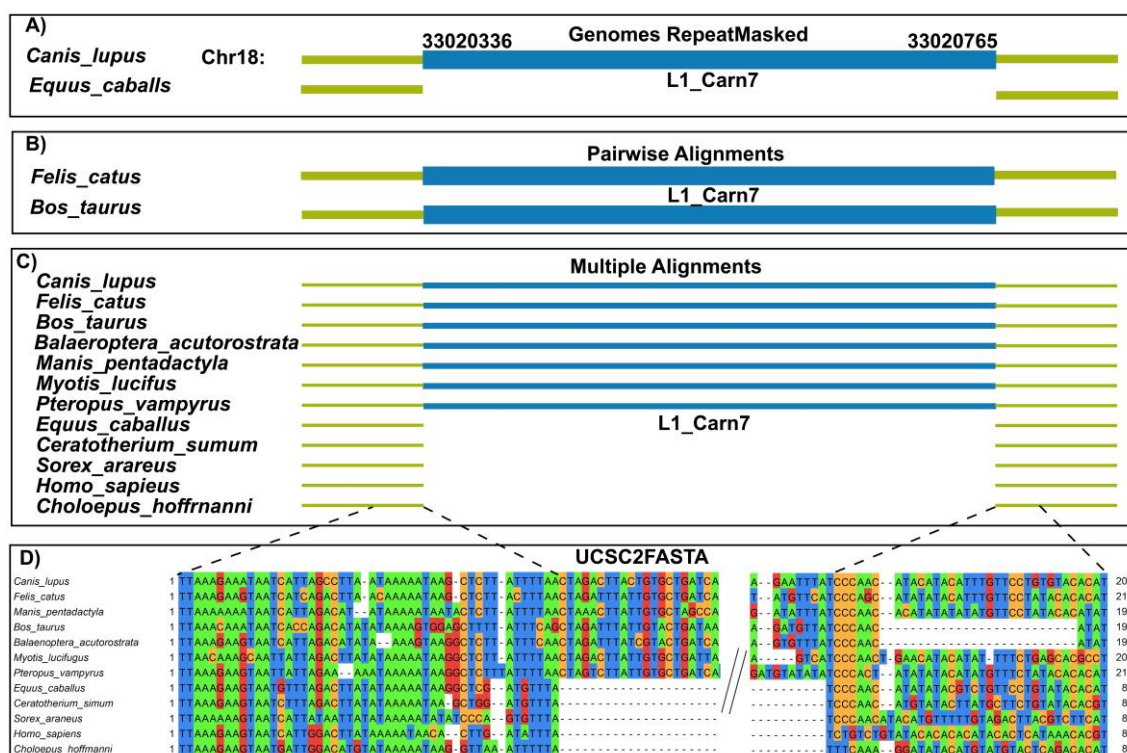


图 1. 在目标基因组中搜索逆转座子 L1_Carn7 的简单流程 A) L1_Carn7 在重复注释的狗 (*Canis_lupus*) 基因组中是插入状态, 马 (*Equus_cabalis*) 是缺失状态; B) 猫 (*Felis_catus*) 和牛 (*Bos_taurus*) 的同源基因组比对; C) 多物种序列比对中 L1_Carn7 的插缺状态; D) 具体的多物种序列的部分展示。双斜线表示省略。

表 2. 逆转座子的插入/缺失信息

物种及代表类群	LTR1	LTR2	SINE1	SINE2	SINE3	LINE1
蝙蝠 (Chiroptera)	0	0	0	1	0	1
马 (Perissodactyla)	0	1	1	1	1	1
牛 (Cetartiodactyla)	1	0	1	0	1	1
狗 (Carnivora)	1	1	0	0	0	0

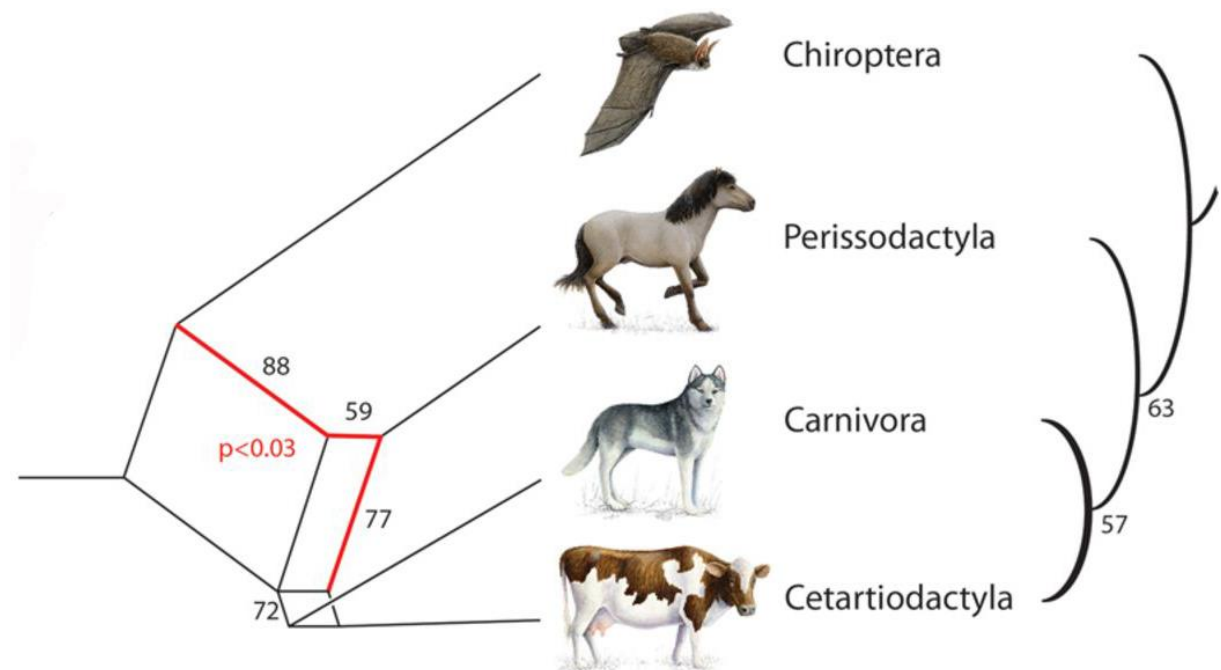


图 2. 基于转座子重建的劳亚兽类群的邻近网状树 (左侧) 和最大简约树 (右侧) (来自于 Doronia *et al.*, 2017) 数字是 bootstrap 值, p 值是四个物种拓扑结构差异性检验。

实验步骤

1. 用手术剪剪取目标物种的组织 (主要是肺和肾组织) 样品到 1.5 ml 离心管中, 剪碎组织, 加入裂解液和严格按照 DNeasy® Blood and Tissue Kit (QIAGEN) 的操作步骤提取目标物种的基因组 DNA。
2. 根据生物信息方法得到的潜在逆转录转座子侧翼序列, 应用 PriFi 设计简并引物。为保证扩增成功率, 一个逆转录转座子选取最优的三对引物发送到生物技术公司, 进行合成。
3. PCR 反应的体系和条件严格按照生物公司的 DNA 聚合酶说明书要求。可参考表 3 和表 4。

表 3. 反应体系 (30 μ l)

试剂	体积 (μ l)
DNA 模板 (~25 ng/ μ l)	1-1.5
Primer Forward (10 mM)	2
Primer Reverse (10 mM)	2
2x PCR Buffer (含 DNA 聚合酶)	15
ddH ₂ O	4.5-5
Total	25

PCR 体系配好后轻微涡旋混匀，分装到八连管中，微离心。

表 4. PCR 反应条件

流程	温度	时间
步骤 1	95 °C	5 min
步骤 2	95 °C	30 s
步骤 3	55-46 °C	40 s
步骤 4	72 °C	2 min
步骤 5	重复步骤 2-4 至 10 个循环，每个循环步骤 3 降低 1 °C	
步骤 6	95 °C	30 s
步骤 7	45 °C	45 s
步骤 8	72 °C	2 min
步骤 9	重复步骤 6-8 至 40 个循环	
步骤 10	72 °C	7 min
步骤 11	4 °C	保存

将分装好 PCR 体系的八连管放入 PCR 仪中，设置运行程序。

4. 琼脂糖凝胶电泳检测

用 1x TAE 配置的 1.2%浓度琼脂糖凝胶 (100 ml TAE 加入 1.2 g 琼脂糖)，微波炉加热煮沸至琼脂糖完全溶解，冷却至约 60-70 °C，摇匀后缓慢倒入制胶槽中，插入梳子，冷却至完全凝固，大约 25 min。适量的 PCR 产物加入到琼脂糖胶孔中，加入

适当大小的 Marker，120 V 电泳 25-30 min。

5. 取出琼脂糖胶放入凝胶成像系统中拍照，检查目的条带是否扩增成功 (如图 3)。
6. 物种全部扩增成功的 PCR 产物，送至生物技术公司 Sanger 测序。
7. 测序的结果，经过人工校正，与潜在的逆转录转座子序列比较，排除非特异性扩增。

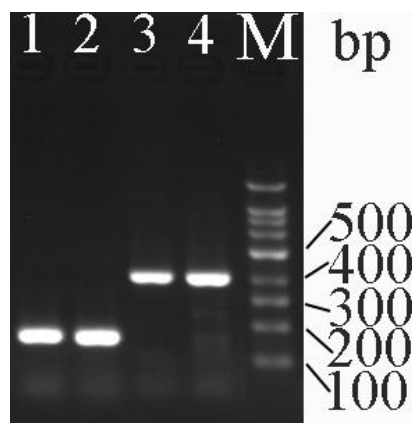


图 3. 逆转座子的 PCR 扩增成功的胶图 M 表示分子 Marker，bp 表示 base pair，1-4 表示四个物种，其中物种 1，2 的逆转座子是缺失状态；物种 3，4 的逆转座子是插入状态。

注意事项

1. PCR 试剂应放置在冰上融化，用完后及时放回-20 °C 冰箱。
2. PCR 成功扩增是建立在模板与引物特异性结合，因此，引物的设计尤为重要。
3. PCR 测序序列重新比对，由于大量的缺失，导致逆转录转座子比对的效果堪忧，需要人工肉眼的校对。

溶液配方

1. 50x TAE 缓冲液

称取 242 g Tris 碱、37.2 g Na₂EDTA·2H₂O 倒入 1 L 容量瓶中，然后加入约 800 ml 的双蒸水，充分搅拌溶解，加入 57.1 ml 的冰醋酸，充分混匀。加双蒸水定容至 1 L，室温保存。需要使用的時候，稀释为 1x TAE 即可使用。

竞争性利益声明

作者声明没有利益冲突

参考文献

1. Doronina, L., Churakov, G., Shi, J., Brosius, J., Baertsch, R., Clawson, H. and Schmitz, J. (2015). [Exploring massive incomplete lineage sorting in Arctoids \(Laurasteria, Carnovira\)](#). *Mol Biol Evol* 32(12): 3194-3204.
2. Doronina, L., Clawson, H., Kuritzin, A., Shi, J., Baertsch, R., Clawson, H. and Schmitz, J. (2017). [Speciation network in Laurasiatheria: retrophylogenomic signals](#). *Mol Biol Evol* 27: 997-1003.
3. Hamdi, H., Nishio, H., Zielinski, R. and Dugaiczyk, A. (1999). [Origin and phylogenetic distribution of A/u DNA repeats: irreversible events in the evolution of primates](#). *J Mol Biol* 289(4): 861-871.
4. Hartig, G., Churakov, G., Warren, W. C., Brosius, J., Makalowski, W. and Schmitz, J. (2013). [Retrophylogenomics place tarsiers on the evolutionary branch of anthropoids](#). *Sci Rep* 3: 1756.
5. Nikaido, M., Rooney, A. P. and Okada, N. (1999). [Phylogenetic relationships among Cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are closest extant relatives of whales](#). *Proc Natl Acad Sci USA* 96(18): 10261-10266.
6. Suh, A., Paus, M., Kiefmann, M., Churakov, G., Franke, F. A., Brosius, J., Kriegs, J. O., Schmitz, J. (2011). [Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds](#). *Nat Comm* 2: 443.