

树鼩群体基因组变异及遗传多样性数据集

(国家非人灵长类实验动物资源库编制, 云南省昆明市盘龙区茨坝街道龙欣路 17 号, 650201, 0871-68424851, nhp@mail.kiz.ac.cn, 2025-12-12)

资源名称	中文	树鼩群体基因组变异及遗传多样性数据集
	英文	Tree shrew population genomic variation and genetic diversity dataset
资源标识		CSTR:13153.11.20220430.OD.TSPGVHDD.01.KIZ
数据集内容		a. 描述摘要 通过对 83 只树鼩个体的全基因组数据进行变异检测与群体遗传分析, 揭示了树鼩群体的遗传多样性特征及近交过程中遗传结构的演化规律。总体来看, PCA 与 Admixture 结果表明野生树鼩各家系间遗传分化较低, 群体结构均一, 说明其具备较高的遗传基础一致性, 有利于后续近交系的构建。在多态性指标方面, 树鼩群体的核苷酸多样性 (π) 约为 0.0010 - 0.0013, 纯合度 (HOM) 在 0.81 - 0.89 之间, 显示出适中的遗传变异水平, 为维持健康种群并开展系统的近交提供了稳定的遗传背景。随着近交代数的推进, 树鼩基因组中连续纯合区段 (ROH) 总长度显著增加, 从 F0 阶段的低水平上升至 F5 代超过 150 万 Kb, 表明基因组纯合度持续增强。同时, 近交系数 (F 值) 从 F0 的约 0 上升至 F5 超过 0.4, 达到大鼠成熟近交系的遗传一致性水平。这一趋势说明树鼩在经过约五代连续自交后, 遗传背景已趋于稳定, 杂合位点显著减少, 遗传漂变作用逐渐固定, 从而标志着树鼩近交系的基本形成。从选育策略上看, F2 至 F3 阶段是近交纯化的关键时期, 纯合度快速上升而群体仍具一定多样性, 此阶段应加强健康筛查和繁殖控制, 以防止近交衰退; 至 F5 阶段后, 个体间遗传一致性高, 可作为标准化的实验用近交系开展疾病模型与基因功能研究。通过结合人类遗传数据库 (如 ClinVar、OMIM、1000Genomes) 对树鼩变异进行注释, 研究还发现了多个与人类疾病相关的功能变异位点, 如经验证的 BRCA1 移码突变, 为疾病模型系的靶向创制提供了明确的遗传依据。总体而言, 本项目不仅系统评估了树鼩群体的遗传多样性与近交系数变化规律, 也为树鼩近交系的创制与遗传评价建立了量化标准。结果表明, 通过连续近交与基因型监测相结合的策略, 可以有效获得遗传背景稳定、纯合度高、可重复利用的树鼩近交家系。这一资源为树鼩在进化生物学、遗传资源保护及人类疾病模型研究中的应用奠定了坚实基础。 b. 关键词 树鼩, 基因组, 遗传变异 c. 数据的时间范围 2022 年 4 月 d. 数据的空间范围 中国科学院昆明动物研究所 e. 学科范围 18031 (遗传学) f. 行业范围 M7310 (自然科学研究和试验发展)

	g. 数据格式 .fastq.gz h. 数据量 260GB i. 名词解释与量纲 j. 数据精度 基因组层面的 PacBio SMRT 长读长测序平均读长超过 10 kb，单碱基错误率经多轮 Quiver/Pilon 校正后显著降低，最终基因组装配的整体质量值达到约 28–30（对应错误率 <0.002%），并在 BUSCO 分析中获得 >91% 完整度，显示较高的组装精度。Illumina 短读长测序用于错误校正与变异检测，其原始数据中 Q30 比例普遍高于 80%，确保基因组校正、SNP/indel 识别的高可信度。 k. 数据更新频度 无																								
缩略图	Varitations of <i>PSEN1</i> Show all variations Liftover variations (from treeshrew genome position to human genome position) Positive selection (only calculated for one-one orthologue among 8 mammal species) <table><tr><th>Model</th><th>2 delta lnL</th><th>P-value</th><th>Positively selected sites (BEB)</th></tr><tr><td>Branch site</td><td>0</td><td>1.000000</td><td>-</td></tr><tr><td>Branch</td><td>1.57430599999952</td><td>0.2096</td><td>-</td></tr></table> Population genetics <table><tr><th>Theta pi</th><th>Watterson's Theta</th><th>Tajima's D</th><th>Fu and Li's D test</th><th>Fu and Li's F test</th><th>Fay and Wu's H test</th></tr><tr><td>49.788</td><td>53.493</td><td>-0.317</td><td>0.923</td><td>0.679</td><td>-0.744</td></tr></table> Population Genetics Theta piWatterson ThetaTajima DF testF testH test	Model	2 delta lnL	P-value	Positively selected sites (BEB)	Branch site	0	1.000000	-	Branch	1.57430599999952	0.2096	-	Theta pi	Watterson's Theta	Tajima's D	Fu and Li's D test	Fu and Li's F test	Fay and Wu's H test	49.788	53.493	-0.317	0.923	0.679	-0.744
Model	2 delta lnL	P-value	Positively selected sites (BEB)																						
Branch site	0	1.000000	-																						
Branch	1.57430599999952	0.2096	-																						
Theta pi	Watterson's Theta	Tajima's D	Fu and Li's D test	Fu and Li's F test	Fay and Wu's H test																				
49.788	53.493	-0.317	0.923	0.679	-0.744																				
数据质量描述	实验用成年中华树鼯由中国科学院昆明动物研究所实验动物中心饲育。所有动物实验均经中国科学院昆明动物研究所动物伦理委员会批准。动物经戊巴比妥钠麻醉后，采用磷酸盐缓冲盐水（PBS）进行心脏灌注，采集多组织进行提取，包括大脑、肌肉、肝脏、肠道、睾丸等。RNA 提取、文库构建和测序由安诺优达基因科技（中国）完成。测序采用 Illumina 高通量平台，平均测序深度约为 30×，各样本的 Q30 碱基比例均超过 90%，确保了数据的高准确性。原始数据经 Trimmomatic 2 软件严格过滤以去除接头序列、低质量 reads 及含 N 比例高的 reads。过滤后数据的平均保留率超过 95%。变异检测采用 GATK v4.1 标准流程，结合多重过滤标准（DP<8、QD<5.0、HRun>5、SB>0.00、QUAL<50、FS>60.0、MQ<40.0、HaplotypeScore>13.0），剔除了潜在的低置信度变异。最终生成的变异集经过多轮交叉验证与一致性评估，确保变异识别的准确性和群体层面的代表性。																								
数据产生方式	a. 样本收集与测序： 共采集 83 只树鼯个体的组织样本，提取高质量基因组 DNA。使用 Illumina 测序平台构建短片段文库（平均插入片段长度约 350 bp），并进行双端测序。																								

	b. 参考基因组： 所有 reads 均比对至树鼩参考基因组。 c. 变异检测流程： 质量控制（QC）： 使用 Trimmomatic 2 过滤低质量 reads。 d. 比对（Alignment）： 使用 BWA 软件将清洗后的 reads 映射至参考基因组。 e. 重复标记（Duplicate Marking）： 使用 Markduplicate 识别并标记 PCR 重复。 f. 变异识别（Variant Calling）： 采用 Genome Analysis Toolkit (GATK) v4.1 鉴定 SNVs 和 InDels。 g. 变异过滤（Variant Filtering）： 利用 GATK VariantFiltration 模块进行高置信度变异筛选，参数设置为： DP<8 QD<5.0 HRun>5 SB>0.00 QUAL<50 FS>60.0 MQ<40.0 HaplotypeScore>13.0。
数据采集、加工处理方法	原始 fastq 数据经过严格的质量控制，剔除测序接头、低质量和含 N 比例>10%的 reads。经清洗后的数据用于后续分析。高质量 reads 比对至 TS_3.0 参考基因组后，使用 SAMtools 及 Picard 对比对结果进行排序、重复标记和索引。变异检测采用 GATK HaplotypeCaller 进行单样本初步检测，之后使用 GenotypeGVCFs 进行联合分型。对获得的原始变异结果进行硬过滤（Hard Filtering）与质量控制（QC），以去除假阳性位点。过滤后的高置信度变异使用 ANNOVAR、SnEff 等软件进行功能注释，涵盖基因区、外显子区、同义/非同义突变、剪接位点变异等信息。在获得高质量变异数据的基础上，计算各样本及群体的多态性参数、遗传距离、连锁不平衡（LD）、主成分分析（PCA）及群体结构分析（STRUCTURE/ADMIXTURE）。结果为树鼩种群的遗传多样性研究提供基础数据支撑。
数据使用条件、方法	文件均可用 viscode 等软件打开。
知识产权	a. 标注知识产权说明 使用本数据集时，请在文章中引用以下文献：无 b. 数据标注参考以下规范： 数据来源引用参考以下规范： 中文表达方式：数据来源于国家科技基础条件平台—国家非人灵长类实验动物资源库（https://nhp.kiz.ac.cn）。 英文表达方式：National Resource Center for Non-Human Primates, National Science & Technology Infrastructure of China（https://nhp.kiz.ac.cn）。 致谢方式参考以下规范： 中文致谢方式：感谢国家科技基础条件平台-国家非人灵长类实验动物资源库（https://nhp.kiz.ac.cn）提供数据支撑。 英文致谢方式：Acknowledgement for the data support from National Resource Center for Non-Human Primates, National Science & Technology Infrastructure of China.（https://nhp.kiz.ac.cn）。

	c. 数据贡献者信息 姓名：叶茂森 单位：中国科学院昆明动物研究所（姚永刚课题组） 电话：15682010335 邮箱：yemaosen@mail.kiz.ac.cn
其它说明内容	若使用方希望利用该资源的任何材料开展宣传等活动，须事先得到资源管理方的书面授权。