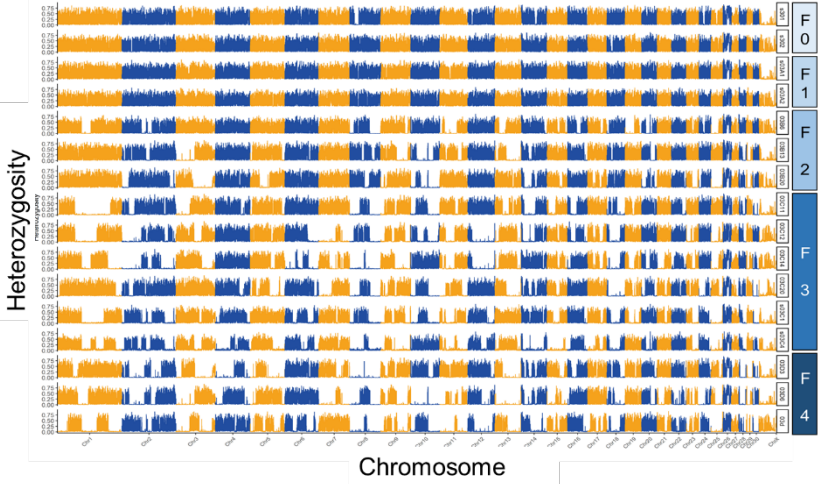


树鼩自发疾病家系样本全基因组测序数据信息

(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	树鼩自发疾病家系样本全基因组测序数据信息
	英文	Whole-genome sequencing data information of spontaneous disease family samples in tree shrews
资源标识		CSTR:13153.11.20211231.OD.TSWGSDI.01.KIZ
数据集内容		a. 描述摘要 本研究针对树鼩自发疾病家系样本开展了全基因组测序分析, 旨在揭示树鼩自发疾病 (尤其是自发性白内障) 的潜在遗传机制。通过高深度测序与系统性变异检测, 共获得 83 个个体的全基因组数据, 并鉴定出 22, 431, 650 个位点的基因组变异。其中包括 14, 730, 085 个单核苷酸多态性位点 (SNPs) 与 7, 701, 565 个小片段插入缺失 (InDels), 后者由 3, 514, 133 个插入和 4, 187, 432 个缺失构成。对所有变异进行功能影响预测后发现, 具有高影响力 (High effect) 的突变 (如移码突变、基因融合、提前终止密码子突变等) 共有 27, 680 个, 占比 0.112%; 中等影响力 (Moderate effect) 的突变 (如非同义突变、转录剪切位点变异等) 共有 119, 726 个, 占比 0.486%; 低影响力 (Low effect) 变异共有 268, 850 个, 占比 1.09%。 这些变异数据在高覆盖度、高准确性测序基础上生成, 结合完善的树鼩参考基因组注释信息, 为群体遗传学及疾病致病研究提供了坚实的数据支撑。在进一步分析中, 我们聚焦于自发性白内障家系。通过检测 32 个患病家系个体及 03 家系患病个体共同的杂合缺失区间 (Runs of Homozygosity, ROH), 共发现 430 个共享 ROH 片段。其中 231 个 ROH 未出现在健康个体中, 被视为潜在致病区域, 总长度约 4, 552.7 kb, 包含 237 个基因。经与白内障致病基因数据库 (OMIM、IMPC、CAT-map) 交叉比对, 发现 5 个基因可能与白内障的发生密切相关, 分布于染色体 17 和染色体 9 的 4 个 ROH 区间内。这些结果提示, 患病个体特异的 ROH 中可能富集了致病突变, 其累积可能导致树鼩自发性白内障的发生, 为进一步功能验证提供了候选靶标。 b. 关键词 树鼩, 白内障, 全基因组测序 c. 数据的时间范围 数据采集于 2016 年 1 月-2021 年 12 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 .gvcf.gz h. 数据量 7.70GB

	i. 名词解释与量纲 CHROM（染色体）：变异所在的染色体编号或 contig 名称。 POS（位置）：变异在参考基因组上的确切碱基位置。 REF（参考等位基因）：参考基因组在该位置的标准碱基。 ALT（备用等位基因）：样本中检测到的非参考等位基因。若有多个变异，则以逗号分隔。 QUAL（质量值）：表示该位点变异可信度的数值，通常为对数似然得分，值越高，变异越可靠。 FILTER（过滤状态）：标记该变异是否通过质量过滤。 INFO（附加信息）：记录深度（DP）、等位基因频率（AF）、变异类型（INDEL/SNP）等信息。 j. 数据精度 通过严格的过滤标准对假阳性位点进行剔除，仅保留 $QUAL \geq 30$（错误率 $< 0.1\%$）且符合深度阈值的位点进入最终数据集。 k. 数据更新频度 无
缩略图	
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养。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。研究共纳入 83 只树鼯个体样本，包括健康与自发疾病（主要为白内障）家系个体。测序数据由 Illumina 高通量平台生成，平均测序深度约 30×，各样本 Q30 碱基比例均超过 90%，确保测序结果的高准确性与可靠性。数据经严格质控与去噪处理：使用 Trimmomatic 2 去除接头污染、低质量 reads 及含 N 比例高的 reads；保留的有效 reads 比例超过 95%。比对使用 BWA 软件将 reads 精准映射至树鼯参考基因组，重复序列由 Markduplicate 标记并去除。变异检测与过滤遵循 GATK v4.1 标准流程，采用参数 $DP<8 \parallel QD<5.0 \parallel HRun>5 \parallel SB>0.00 \parallel QUAL<50 \parallel FS>60.0 \parallel MQ<40.0 \parallel HaplotypeScore>13.0$ 对低质量变异进行剔除。变异功能注释采用 SnpEff 与 ANNOVAR 完成，确保各类突变影响评估的准确性与一致性。总体而言，数据完整性高、背景噪声低，能够可靠支撑群体变异统计与候选致病区域筛选。
数据产生方式	a. 样本来源与测序 研究共纳入 83 只树鼯个体样本，包括健康与自发疾病（主要为白内障）家系个体。样本来源明确、表型记录完整。基因组 DNA 经质量评估合格后用于文库构建，采用 Illumina 平台进行双端测序（paired-end sequencing）。

	b. 参考基因组与比对 使用树鼯参考基因组作为比对模板。 c. 变异检测与功能注释 使用 GATK v4.1 进行 SNP 和 InDel 检测, 并通过 GATK VariantFiltration 模块进行硬过滤。随后利用 SnpEff 与 ANNOVAR 软件注释变异位点的功能影响等级 (High、Moderate、Low)。 d. ROH 分析与致病区域筛选 通过 PLINK 软件计算患病个体的 ROH 区间, 并筛选在所有患病个体中共享、但不出现在健康个体中的 ROH 区域。进一步将这些区域中的基因与已知白内障致病基因数据库进行交叉比对, 以筛选潜在致病候选基因。
数据采集、加工处理方法	树鼯样本采自多个自然家系与实验室维持种群, 覆盖自发疾病表型个体与对照健康个体。样本经形态及临床表现鉴定后, 提取高纯度基因组 DNA。DNA 浓度与完整性经 Qubit、Nanodrop 及琼脂糖凝胶电泳检测确认。原始 fastq 文件经 Trimmomatic 2 过滤低质量 reads 与接头序列。过滤后数据经 FastQC 评估整体质量, 确保无系统性偏差。使用 BWA-MEM 算法比对至参考基因组, 比对率超过 98%。SAMtools 及 Picard 用于排序、去重复与索引。随后使用 GATK HaplotypeCaller 进行单样本变异检测, GenotypeGVCFs 进行联合分型。利用 GATK VariantFiltration 进行硬过滤, 去除低覆盖、低质量或可能误判的位点。随后使用 ANNOVAR 与 SnpEff 进行功能注释, 评估突变的潜在生物学效应。ROH 分析采用 PLINK (--homozyg 命令) 计算个体连续纯合区间, 筛选出患病个体间共有但健康个体中缺失的 ROH 区域。结合 OMIM、IMPC、CAT-map 等数据库进行候选基因交叉验证, 确定潜在致病基因。
数据使用条件、方法	文件均可用 viscode 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。