

中国野生小鼠来源 1 号染色体替换系 缺失突变的发掘及功能注释

徐福意, 晁天柱, 胡世贤, 李凯, 周宇荀, 周梁良, 肖君华*

(东华大学生物科学与技术研究所, 上海 201620)

【摘要】 目的 基于野生小家鼠来源 1 号染色体替换系群体 (population of specific chromosome 1 substitution strains, PCSSs) 中 18 个品系的全基因组重测序结果, 鉴定 1 号染色体上的缺失突变并对其进行功能注释。**方法** 采用 Illumina 二代测序平台获取 18 个品系的全基因组序列信息, 通过 SpeedSeq 软件鉴定缺失突变, 进一步利用 SnpEff 软件完成功能注释。**结果** 在 18 个品系的 1 号染色体上共检测到 13 803 个缺失突变。缺失长度从 51 bp 到 70 kb 不等, 其中长度 < 500 bp 的缺失突变约占总数的 50%。多数缺失突变位于内含子区 (50.361%) 和基因间隔区 (28.745%)。发现 31 个蛋白编码基因含有功能性缺失, 其中有 3 个基因和人类疾病相关, 7 个基因参与了 11 条 KEGG 通路。**结论** PCSSs 的 1 号染色体上含有丰富的缺失突变, 是在研究复杂性状的重要遗传标记。

【关键词】 缺失突变鉴定; 二代测序; 功能注释; PCSSs

【中图分类号】 Q95-33 **【文献标识码】** A **【文章编号】** 1005-4847(2017) 01-0020-05

Doi:10.3969/j.issn.1005-4847.2017.01.004

Deletion detection and functional annotation of chromosome 1 substitution strains from Chinese wild mice

XU Fu-yi, CHAO Tian-zhu, HU Shi-xian, LI Kai, ZHOU Yu-xun, ZHOU Liang-liang, XIAO Jun-hua*

(Institute of Biological Sciences and Biotechnology, Donghua University, Shanghai 201620, China)

【Abstract】 Objective Deletion detection and annotation of 18 lines from the population of specific chromosome 1 substitution strains (PCSSs) derived from Chinese wild mice based on whole genome re-sequencing data. **Methods** Whole genome re-sequencing of the 18 lines were performed on the Illumina Hiseq platform. SpeedSeq software was used to detect the deletion after read alignment. Further annotation was obtained using SnpEff software. **Results** 13 803 deletions were identified among the 18 lines, the length of deletion was ranged from 51bp to 70 kb, among them nearly 50% were less than 500 bp. Through functional annotation, we found most of the variants were located in intronic (50.361%) and intergenic (28.745%) regions. However, we also identified 31 protein coding genes harboring loss-of-function deletions. Among them, 3 genes were associated with human diseases, 7 genes were participated in 11 KEGG pathways. **Conclusion** The chromosome 1 of PCSSs harbors abundant deletion mutations which can be used as genetic markers in genetic studies.

【Key words】 Deletion detection; Next-generation sequencing; Function annotation; PCSSs

Corresponding author: XIAO Jun-hua, E-mail: xiaojunhua@dhu.edu.cn

【基金项目】 国家自然科学基金面上项目 (编号:31371257); 上海市科技支撑计划 (编号:15140900500)。

【作者简介】 徐福意 (1987-), 男, 博士研究生, 研究方向: 医学分子遗传学。E-mail: xufuyiphd@gmail.com

【通讯作者】 肖君华 (1968-), 男, 教授, 研究方向: 医学分子遗传学。E-mail: xiaojunhua@dhu.edu.cn

现有近交系小鼠品系匮乏的遗传多样性是其在多基因复杂性状研究中进展缓慢的主要原因^[1]。以具有丰富遗传多样性的野生小家鼠为供体建立的 PCSSs, 克服了近交系小鼠在数量性状基因座 (quantitative trait locus, QTL) 研究中的局限性, 是 QTL 精确定位、基因鉴定的理想资源^[2]。缺失突变作为结构变异的一种重要类型, 其大小从 1 bp 到几十 kb 不等, 广泛分布于人和小鼠基因组中^[3,4]。研究表明, 缺失突变与许多人类疾病相关^[3,5]。在小鼠基因组中, 缺失突变已被证明和血清尿素浓度等^[6]性状相关。但迄今, PCSSs 的野生小家鼠来源 1 号染色体中缺失突变的遗传多态性未被解析。

随着二代测序技术的不断发展, 低成本、高通量的全基因组测序得以实现。通过全基因组测序, 不仅可以检测单个位点的遗传突变, 如单核苷酸多态性, 也可以实现全基因组范围内复杂的结构变异检测, 包括插入、缺失、倒位、重复等。目前, 针对二代测序数据检测结构变异主要有 5 种策略^[7]: (1) Read pair (也称为 Pair-end Mapping, 简称 PEM); (2) Split read (简称 SR); (3) Read depth (简称 RD); (4) 基于 de novo 组装的方法; (5) 结合上述方法的综合策略 (combinatorial approach, CA)。而 Speed-Seq 软件内置的 lumpy-sv 结合了结构变异检测的前三种策略, 具有较高的灵敏度和准确度, 显著优于其他软件, 广泛应用于结构变异的检测^[8]。

本研究以 18 个“野生小家鼠来源 1 号染色体替换系”品系为样本, 采用 Illumina 二代测序平台获取其全基因组序列信息。利用 SpeedSeq 和 SnpEff 软件完成 1 号染色体上缺失突变的挖掘及其功能注释, 全面了解和评估缺失突变在这一群体中的分布特征, 进一步完善该群体的遗传变异信息, 为该群体用于相关遗传学研究奠定了基础。

1 材料与方法

1.1 实验动物

实验动物来源于本实验室自建的野生小家鼠来源 1 号染色体替换系。动物实验遵守 1988 年动物管理条例, 并得到东华大学实验动物委员会批准。实验在东华大学生物科学与技术研究所屏障动物实验设施进行 [SYXK (沪) 2014 - 0022]。收集 18 个成年 1 号染色体替换系的鼠尾组织, -20℃ 保存

备用。

1.2 DNA 提取和测序

DNA 提取采用 Axygen (爱思进生物技术有限公司) 基因组 DNA 抽提试剂盒。以 0.8% 琼脂糖凝胶电泳、全自动紫外与可见光分析仪 FR-200A (上海复日科技实验技术研究所) 和 NanoDrop 2000c 超微量分光光度计 (Thermo Fisher Scientific, 美国) 确定 DNA 质量和浓度。采用随机打断方式, 选取大小为 400 ~ 500 bp 的 DNA 片段建库。以读长 (reads) 为 150 bp 的 Illumina 高通量双末端 (pair-end) 测序平台 (药明康德新药开发有限公司) 完成全基因组重测序。

1.3 缺失突变检测及功能注释

首先, 采用 NGS QC 软件^[9]对原始数据 (fastq 文件) 进行质控, 保留质量值 $Q > 20$ 碱基数占整条 reads 碱基数 $> 70\%$ 的 reads; 其次利用 SpeedSeq 软件^[10]中的 align 模块将高质量的数据比对到 C57BL/6J 小鼠参考基因组 (GRCm38/mm10), 并生成 bam 文件; 之后通过 SpeedSeq 软件中的 sv 模块鉴定缺失序列, 获得初步的缺失数据集; 最后, 过滤掉质量值低于 100、SU (support reads) < 4 、长度 > 50 bp 以及鉴定为缺失一个拷贝的数据, 最终生成高质量的缺失数据集。

缺失突变的功能注释利用 SnpEff 软件^[11]完成, 基因注释信息版本为 Ensembl 82。此外, 利用 MGI 中的 Human-Mouse: Disease Connection 数据库^[12]和 DAVID: Functional Annotation Tools^[13]分别完成人类疾病基因注释和 KEGG 通路注释。

2 结果

2.1 测序

测序结果表明, 18 个 1 号染色体替换系均获得了大量、有效、高质量的测序数据, 适用于后续生物信息学研究。如表 1 所示, 18 个样本测序数据量均超过 90 Gb; 质控后的高质量数据比对到参考基因组的数据百分比均在 99% 以上; 各品系的平均测序深度均达到或高于 30 倍。18 个样本的原始数据 (fastq 文件) 均已上传到 NCBI (National Center for Biotechnology Information) 的 SRA (Sequence Reads Archive) 数据库 (编号: SRP066591)。

表 1 18 个品系 DNA 测序及 1 号染色体缺失检测统计

Tab. 1 An overview of the DNA sequencing statistics and chromosome 1 deletion called from 18 lines

品系 Strain	数据量 Yield /Gbases	Q30 碱基比例 % of > = Q30 Bases	映射比例 Mapping rate	测序深度 Average depth/x	缺失数量 Number of deletions	最大长度 Maximum length/bp	中值 Median /bp	平均值 Mean /bp	占 chr1 比例 % Accounted for chr1
B6-Chr1 ^{BLD}	105.78	84.09	99.92	33.38	4573	40 112	1768	491	4.14
B6-Chr1 ^{CM}	120.70	93.60	99.91	40.01	4958	68 164	1732	448	4.39
B6-Chr1 ^{DX}	105.85	85.02	99.94	34.16	4521	40 111	1765	473	4.08
B6-Chr1 ^{FX}	93.10	81.50	99.90	30.12	4550	60 390	1813	504	4.22
B6-Chr1 ^{HZ}	106.30	83.91	99.94	33.58	4370	60 390	1757	482	3.93
B6-Chr1 ^{JD}	98.43	83.74	99.92	31.04	4542	66 693	1784	484	4.15
B6-Chr1 ^{KM}	126.34	80.14	99.92	36.29	1510	24 387	2110	613	1.63
B6-Chr1 ^{PD}	135.30	93.90	99.91	47.36	4642	60 389	1756	452	4.17
B6-Chr1 ^{QP}	91.60	80.80	99.88	29.39	4591	69 356	1777	491	4.17
B6-Chr1 ^{SMX}	92.65	85.63	99.93	30.2	4708	40 111	1766	483	4.25
B6-Chr1 ^{SY}	96.80	80.80	99.92	30.41	4668	40 110	1757	488	4.20
B6-Chr1 ^{TW}	95.10	80.90	99.91	32.01	4716	40 108	1749	473	4.22
B6-Chr1 ^{TZ}	101.88	86.37	99.93	32.84	4528	37 724	1785	492	4.13
B6-Chr1 ^{YP}	107.74	87.09	99.93	35.19	4811	61 957	1754	473	4.32
B6-Chr1 ^{YX}	127.90	93.50	99.91	44.54	4926	40 112	1714	452	4.32
B6-Chr1 ^{ZC}	111.80	93.30	99.90	38.90	4926	29 147	1683	445	4.04
B6-Chr1 ^{ZZ1}	118.38	84.41	99.92	37.49	4820	40 111	1758	471	4.33
B6-Chr1 ^{ZZ2}	128.90	93.70	99.89	45.01	4897	40 112	1712	458	4.29

2.2 缺失检测

缺失突变检测结果如表 1 所示,在 18 个品系的 1 号染色体上,共鉴定 13 803 个缺失。其中,17 个品系的缺失数量在 4370 ~ 4958 个之间 ($\bar{x} = 4691$, $s = 169$)。而 B6-Chr1^{KM} 品系只检测到 1510 个,但其缺失长度的中值和平均值在所有品系中最大。此外, B6-Chr1^{KM} 品系中缺失长度占 1 号染色体 1.63%,而另 17 个品系的比例均大于 4% ($\bar{x} = 4.2\%$, $s = 0.1\%$)。

从长度分布来看,有超过半数的缺失小于 500 bp (51.9%, 7164 个);长度在 0.5 ~ 1 kb 和 1 ~ 2 kb 的缺失分别占 15.5% 和 11.6%;而大于 10 kb 的只有 122 个,说明缺失突变数量与缺失长度呈负相关。但长度在 6 ~ 7 kb 的有 972 个,占整体数量的 7%,高于其他相近长度(图 1)。此外,对缺失突变在 18 个小鼠品系中的频率分布统计发现(图 2),约 40% 的缺失频率小于 0.1,即这些缺失仅存在于单个品系中;另近 1300 个缺失频率在 0.9 ~ 1.0 之间,即同时存在于 17 或 18 个品系中;频率在 0.5 ~ 0.6 之间的最少,仅占总数的 2% (278 个)。

2.3 功能注释

功能注释结果如表 2 所示,多数缺失位于内含子区 (49.655%) 和基因间隔区 (27.955%),其次为基因上下游,其余区域所占比例均接近或小于 1%。共发现 605 个缺失对基因功能可能产生功能性的影

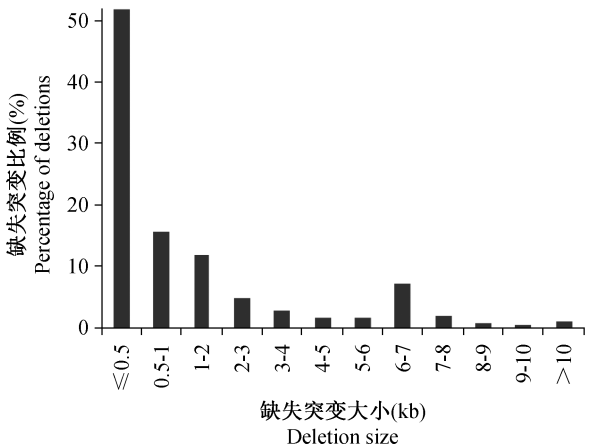


图 1 缺失长度频率分布

Fig. 1 Size distribution of the identified deletions

响(loss of function deletion),其中包括 99 个外显子缺失、15 个移码突变、188 个剪切受体或供体突变、4 个起始或终止密码子突变以及 199 个转录本 ablation。这 605 个缺失突变共计影响 205 个基因,其中包括 31 个蛋白编码基因(其中 21 个基因含有外显子缺失突变)、9 个 LincRNA、10 个 miRNA、3 个 snRNA 和 7 个 snRNA。此外,还有部分缺失位于 3' UTR 和 5' UTR 区 (0.36%),这些缺失突变可能通过影响基因表达调控序列而发挥作用。

2.4 蛋白编码基因疾病与通路注释

31 个含缺失突变的蛋白编码基因的注释结果

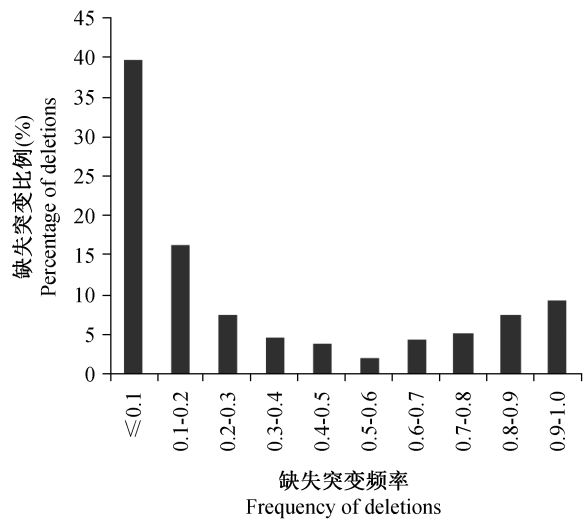


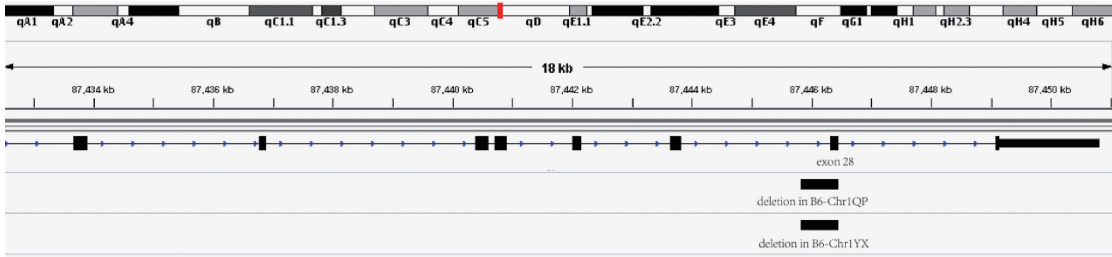
图2 缺失在 18 个品系中的频率分布
Fig.2 Frequency distribution of identified deletions among 18 lines

显示,其中 3 个基因与人类疾病相关,分别为 *Gig**yf*2, *Ptpn*14 和 *Cfh*。如图 3 中所示,在 *B6-Chr1*^{QP} 和 *B6-Chr1*^{YX} 小鼠中,与人类帕金森病相关 (OMIM: 607688) 的 *Gig**yf*2 基因,其第 28 个外显子含有一 634 bp 的缺失。与后鼻孔闭锁和淋巴水肿相关 (OMIM:613611),属于蛋白酪氨酸磷酸酶家族的 *Ptpn*14 基因,其第 3 个外显子完全丢失 (*B6-Chr1*^{BLD} 小鼠)。而与补体因子 H 缺乏 (OMIM:126700)、溶血性尿毒综合征 (OMIM:235400) 以及年龄相关性

黄斑变性 (OMIM:610698) 有关的 *Cfh* 基因的剪切受体区域存在序列缺失。31 个基因的 KEGG 通路注释结果如表 3 所示,共发现 7 个基因参与了 11 条代谢通路,包括补体和凝血级联、药物代谢、嗅觉转导以及细胞因子-细胞因子受体相互作用等。

表 2 缺失突变功能注释结果

Tab. 2 Predicted functional consequence of deletions				
类型 Type	数量 Count	比例/% Percent	影响 Influence	
3_prime_UTR_truncation	4	0.013	低	
3_prime_UTR_variant	53	0.168	低	
5_prime_UTR_truncation	42	0.133	低	
5_prime_UTR_variant	15	0.048	低	
disruptive_inframe_deletion	14	0.044	低	
downstream_gene_variant	2953	9.371	低	
exon_loss_variant	99	0.314	高	
frameshift_variant	15	0.048	高	
inframe_deletion	1	0.003	低	
intergenic_region	8809	27.955	低	
intragenic_variant	1	0.003	低	
intron_variant	15 650	49.665	低	
non_coding_exon_variant	337	1.069	低	
non_coding_transcript_variant	31	0.098	低	
splice_acceptor_variant	145	0.460	高	
splice_donor_variant	143	0.454	高	
splice_region_variant	345	1.095	高	
start_lost	1	0.003	高	
stop_lost	3	0.010	高	
transcript_ablation	199	0.632	高	
upstream_gene_variant	2651	8.413	低	



注:图中上部分为 *Gig**yf*2 基因示意图,下部分为 *B6-Chr1*^{QP} 和 *B6-Chr1*^{YX} 品系在 *Gig**yf*2 基因上的缺失突变。

图3 *Gig**yf*2 基因在 *B6-Chr1*^{QP} 和 *B6-Chr1*^{YX} 小鼠中的缺失突变

Note. Show in upper panel is part of *Gig**yf*2 gene model based on Ensembl, lower panel are deletions in *B6-Chr1*^{QP} and *B6-Chr1*^{YX} lines.

Fig.3 An exon loss of *Gig**yf*2 gene in *B6-Chr1*^{QP} and *B6-Chr1*^{YX} lines

3 讨论

18 个 1 号染色体替换系小鼠的基因组中,除 1 号染色体来源于不同供体小鼠外,其他基因组背景均来源于近交系品系 C57BL/6J。因此,本研究只针对 1 号染色体进行缺失的鉴定与分析。此外,因 18 个小鼠品系已近交化,因此,不可能出现单拷贝缺

失,利用这一特征,我们对鉴定的缺失做了进一步的过滤,最终获得了高质量的缺失数据集。作为遗传变异的一种类型,缺失可通过诸如外显子缺失、移码框改变等方式改变蛋白结构,使其丧失原有功能,也可通过改变基因拷贝数或基因转录调控序列而影响其表达量。本次研究,在 18 个品系的 1 号染色体上注释到 605 个功能性缺失,共影响

205 个基因,其中大部分基因为假基因,只有 31 个为蛋白编码基因。通过对 31 个基因做进一步的疾病与通路注释,发现 3 个基因与人类疾病相关,7 个基因参与了 11 条已知通路,提示这些基因的相关表

型可能在该群体中存在差异。此外,我们还发现了一些 lincRNA、miRNA 等存在功能性的缺失。作为非编码 RNA,二者在基因转录调控方面发挥着重要作用^[14,15]。

表 3 KEGG 通路注释结果

Tab.3 The annotation results of KEGG pathway			
基因名 Gene name	Ensembl 编号 Ensembl ID	KEGG 通路 KEGG pathway	描述 Description
B3gat2	ENSMUSG00000026156	mmu00532	硫酸软骨素合成
		mmu00534	硫酸乙酰肝素生物合成
Cfh	ENSMUSG00000026365	mmu04610	补体和凝血级联
Fmo1	ENSMUSG00000040181	mmu00982	药物代谢
		mmu04060	细胞因子-细胞因子受体相互作用
		mmu04062	趋化因子信号传导途径
Cxcr2	ENSMUSG00000026180	mmu04144	胞吞作用
		mmu00230	嘌呤代谢
Nme7	ENSMUSG00000026575	mmu00240	嘧啶代谢
		mmu04740	嗅觉转导
Olfir248	ENSMUSG00000059503	mmu04740	嗅觉转导
G2000	ENSMUSG00000078193	mmu03010	核糖体

中国野生小家鼠包含两个亚种 *M. m. musculus* 和 *M. m. castaneus*^[16],而 KM 鼠是上世纪 50 年代引入的 Swiss 小鼠,与现有近交系小鼠来源于同一亚种 *M. m. domesticus*^[17],本研究结果揭示,*B6-Chr1^{KM}* 品系 1 号染色体上缺失突变的数量(1510 个)与其他 17 个品系相差较远,说明野生小家鼠缺失突变的多样性高于实验室小鼠。

总之,本次研究,利用二代测序技术,结合生物信息学分析手段准确解析了 18 个 1 号染色体替换系缺失突变的遗传多样性,进一步完善了该群体的遗传变异信息,为该群体用于相关遗传学研究奠定了基础。

参 考 文 献

[1] Flint J, Valdar W, Shifman S, et al. Strategies for mapping and cloning quantitative trait genes in rodents[J]. Nat Rev Genet, 2005, 6(4):271-286.

[2] Xiao J, Liang Y, Li K, et al. A novel strategy for genetic dissection of complex traits: the population of specific chromosome substitution strains from laboratory and wild mice[J]. Mammalian Genome, 2010, 21(7-8):370-376.

[3] Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease[J]. Ann Rev Med,2010, 61(1):437-455.

[4] Yalcin B, Wong K, Bhomra A, et al. The fine-scale architecture of structural variants in 17 mouse genomes[J], Genome Biol, 2012, 13(3):405-413.

[5] Weischenfeldt J, Symmons O, Spitz F, et al. Phenotypic impact of genomic structural variation: insights from and for human disease[J]. Nat Rev Genet, 2013, 14(2):125-138.

[6] Keane TM, Wong K, Adams DJ, et al. Identification of structural variation in mouse genomes. [J]. Front Genet,2014, 5:192.

[7] Zhao M, Wang Q, Wang Q, et al. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives[J]. BMC Bioinformatics, 2013, 14(S11):1-16.

[8] Layer RM, Chiang C, Quinlan AR, et al. LUMPY: a probabilistic framework for structural variant discovery[J]. Genome Biol, 2012, 15(6):1-19.

[9] Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data[J]. PLoS One, 2012, 7(2):e30619

[10] Chiang C, Layer RM, Faust GG, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation[J]. Nat Methods, 2015, 12(10):966-968.

[11] Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3[J]. Fly, 2012, 6(2):80-92.

[12] Eppig JT, Blake JA, Bult CJ, et al. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease[J]. Nucleic Acids Res,2014, 43(D1):726-736.

[13] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nat Protoc, 2009, 4(1):44-57.

[14] Ulitsky I, Bartel DP. lincRNAs: Genomics, evolution, and mechanisms[J]. Cell, 2013, 154(1):26-46.

[15] Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing[J]. Nat Rev Genet,2015, 16(7):421-433.

[16] Boursot P, Auffray JC, Brittondavidian J, et al. The evolution of house mice[J]. Ann Rev Ecol Syst,1993, 24(4):119-152.

[17] 岳乘飞, 刘双环, 刘殿峰,等. 五个昆明小鼠封闭群遗传生化位点比较研究[J]. 实验动物科学与管理, 2003, 20(z1):58-62.

[收稿日期] 2016-06-28