动 物 学 研 究 2008，Dec. 29(6)：577−584 CN 53-1040/Q ISSN 0254-5853
Zoological Research DOI：10.3724/SP.J.1141.2008.06577

家马核基因组中线粒体核内插入序列分析

姜 枫[1,2]，苗永旺[1,3,*]，刘 斌[2,4]，申 欣[4,5]，任建峰[2,4]，何 帆[2]

（1. 云南农业大学 动物科学技术学院，云南 昆明 650201；2. 中国科学院北京基因组研究所，北京 101300；
3. 云南大学 云南省生物资源保护与利用重点实验室，云南 昆明 650091；
4. 中国科学院海洋研究所, 山东 青岛 266071; 5. 淮海工学院 海洋学院，江苏 连云港 222005）

摘要：在各种真核生物核基因组中，存在一些由线粒体基因组转移进入核基因组中的 DNA 片段，这些被认为是分子化石的片段叫做线粒体核内插入序列（Numt）。由于 Numt 与真实的线粒体序列高度相似，因此它的存在必然会成为 PCR 扩增线粒体 DNA 的不利因素。利用已经公布的家马（*Equus caballus*）基因组序列（2007 年 9 月公布，GenBank 登录号为 NC_009144－NC_009175）对家马 Numt 进行了深入分析，共发现 200 个可能的 Numt，长度范围为 29 到 3727bp，其中有 10 个的长度大于 800bp。分析结果显示由于不存在线粒体控制区域的疑似 Numt，因此对基于此区域的群体遗传学研究不会产生影响。本研究还发现在家马进化过程中，第 1 号和 27 号染色体更倾向于接受线粒体序列的转移。以上结果将为今后马科动物的研究提供重要的参考信息，有助于避免在线粒体 DNA 研究中由于 Numt 污染的存在而得出错误的实验结果。

关键词：线粒体 DNA；核内插入序列；Numt；家马；假基因
中图分类号：Q349.5；Q959.843 **文献标识码**：A **文章编号**：0254-5953-(2008)06-0577-08

Mitochondrial Introgressions into the Nuclear Genome of the Domestic Horse

JIANG Feng[1, 2], MIAO Yong-wang[1, 3, *], LIU Bin[2, 4], SHEN Xin[4, 5], REN Jian-feng[2, 4], HE Fan[2]

（1. *Faculty of Animal Science and Technology, Yunnan Agricultural University, Kunming Yunnan* 650201, *China*;
2. *Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing* 101300, *China*;
3. *Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming Yunnan* 650091, *China; 4. Institute of Oceanology, Chinese Academy of Sciences, Qingdao* 266071, *China; 5. College of Marine Science, Huaihai Institute of Technology, Lianyungang* 222005, *China*）

Abstract: The nuclear insertions of mitochondrial DNA (Numts), which originate from the integration of nuclear DNA by mtDNA, are found as molecular fossils in the nuclear genomes of various eukaryotes. Because integrated Numts tend to have a high sequence similarity to genuine organellar mtDNA sequences, inadvertent amplification of Numts can be a nuisance in studies of mtDNA variation. With the availability of the complete domestic horse genome sequence, we present the first comprehensive analysis of genome-wide distribution and frequency of Numts in the nuclear genome of domestic horse (*Equus caballus*). In the present paper, we detected 200 Numts ranging between 29 and 3 727bp in size, which collectively representing only 0.002154% of the nuclear genome. Furthermore, ten of these segments were found to be longer than 800 bp. The absence of Numts in mitochondrial control region suggested that it would not influence the analysis of horse population genetics studies relating to this region. We also found that during horse evolution, Chromosomes 1 and 27 have been more susceptible to integration by Numts. The results in this study may provide valuable information for future mtDNA studies in Equidae species, including its use as a tool for avoiding Numt contaminations that may result in inauthentic results of experimentation.

Key words: Mitochondrial DNA; Nuclear insertion; Numt; *Equus caballus*; Pesudogene

在真核生物中，存在与线粒体 DNA(mtDNA) 相似的细胞核 DNA 片段，这些片段被称为线粒体

收稿日期：2008-06-30；接受日期：2008-10-28
基金项目：国家自然科学基金项目（30660024）；云南省应用基础研究重点项目（2007C0003Z）；云南省应用基础研究计划面上项目（2006C0034M）
*通讯作者（Corresponding author）：苗永旺，Tel: 13700650615，E-mail: yongwangmiao999@yahoo.com.cn
第一作者简介：姜枫(1981-)，男，硕士研究生，专业方向为动物分子遗传学

DNA 核内插入序列(nuclear insertion of mitochondrial DNA, Numt)(Lopez et al, 1994)，或者称为线粒体假基因(Wang, 2004)。自从 1967 年首次发现 Numt 以来(du & Riley, 1967)，在多种动植物的细胞核基因组内都发现存在 Numt 现象(Bensasson et al, 2001; Richly & Leister, 2004)。mtDNA 与核基因组发生非同源重组是 Numt 产生与积累的重要原因(Henze & Martin, 2001; Woischnik & Moraes, 2002)，而核基因组倍增是 Numt 数目增多的另外一个原因(Tourmen et al, 2002; Hazkani-Covo et al, 2003) 。因此 Numt 在不同物种中积累的数目差异很大，在人类(*Homo sapiens*)、水稻(*Oryza Sativa*)及拟南芥(*Arabidopsis thaliana*)中存在大量的 Numt，但是在秀丽广杆线虫(*Caenorhabditis elegans*)、原鸡(*Gallus gallus*)和蜜蜂(*Apis mellifera*)中却存在很少的 Numt 或者甚至没有(Pereira & Baker, 2004)。造成这种数目差异的原因目前并未弄清。因此有必要对不同物种的 Numt 系统分析，为阐明 Numt 进化规律奠定基础。

随着基因组序列的测定，对不同物种 Numt 系统分析的工作已经逐步开展。人类基因组的分析显示，最长的 Numt 片段达到 14 654 bp，片段长度大于 800 bp 的数目为 110 个(Mourier et al, 2001)。由于以往未对人类 Numt 进行系统分析，Numt 富集现象造成 PCR 过程中的 Numt 混入，因此得出错误的疾病研究结果(Wallace et al, 1997)，甚至是人类 Numt 被错误认为是恐龙的 DNA 序列(Zischler et al, 1995)。而最近对家猫（*Felis catus*)1.9 倍覆盖的核基因组序列分析，也发现高达 12.5 kb 和 7.9 kb 的疑似 Numt 存在，并建议在猫科动物线粒体群体遗传学研究中应该首先仔细检查 Numt 的混入污染情况(Antunes et al, 2007)。但是并非所有的物种都需要如此谨慎，对家鸡核基因中疑似 Numt 分析发现，只要 PCR 扩增片段大于 1.5 kb 就能获得真实线粒体序列(Pereira & Baker, 2004)。

mtDNA 由于具有严格母系遗传模式、拷贝数目较多和无 DNA 重组现象等优点，因此作为分子标记广泛运用于后生动物的系统发育学和群体遗传学研究。但是在 mtDNA 的分析中，作为线粒体基因的旁系同源物，Numt 的排除不可避免的成为首要解决的问题。Numt 与线粒体基因比较，序列的进化速度相对较慢，因此当使用物种间的通用引物或者基于保守区域设计的引物时，Numt 往往更容易被 PCR 扩增(Mirol et al, 2000; Zhang et al, 2006)，必然会混淆后续的数据分析，甚至得出错误的物种间进化关系和生物系统地理学结论(Thalmann et al, 2004)。受益于越来越多物种的基因组测序项目的完成，使得 Numt 在核基因组中的系统、全面地评估得以实现。最近 Broad 研究所公布的家马(*Equus caballus*)基因组数据为我们分析家马 Numt 提供了有利条件。家马基因组中共有 31 对常染色体和 1 对性染色体。本研究通过分析家马细胞核基因组中 Numt，为马科动物分子进化和群体遗传学研究提供 Numt 的重要参考信息，以期在后续基于 mtDNA 的研究中能够避免由于 Numt 的影响而得出不正确的结论。

1 材料与方法

1.1 材 料

家马基因组序列由美国国立生物技术信息中心(National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/)数据库下载。家马线粒体基因组全序列登录号为 NC_001640(Xu & Arnason, 1994)，其细胞核基因组序列登录号为 NC_009144-NC_009175，核基因组数据中不包括未公布的 Y 染色体序列。核基因组数据是基于 6.8 倍覆盖的鸟枪法测序获得，于 2007 年 9 月公布。

1.2 核基因组中疑似 Numt 鉴定

以核基因组序列建立本地数据库，线粒体全基因序列为查询序列，BLAST(Basic Local Alignment Search Tool)用于搜寻核基因组中的线粒体插入序列(Altschul et al, 1990)。为寻找到具有生物学意义的疑似Numt，按照以往文献中的方法，将BLASTN中的最大期望值设置为$e=10^{-4}$(Pereira & Baker, 2004; Richly & Leister, 2004)。不设定低复杂序列区域过滤设置。核酸序列基对不匹配所罚分数（penalty for a nucleotide mismatch）和核苷酸序列基对匹配所加分数（reward for a nucleotide match）为默认设置，设定值分别为−3 和 1。当两个不同疑似Numt在距离和位置上与其对应的线粒体部分匹配时，这两个疑似Numt就认为是发生于同一次核转移事件并且被连接成为一个疑似Numt(Woischnik & Moraes, 2002)。使用Ensembl在线图形界面基因组浏览器对家马疑似Numt在核基因组中整合位置进行观察，鉴定疑似Numt所在区域的基因结构特征(Hubbard et al, 2002)。

1.3　疑似 Numt 中相应线粒体基因识别

以线粒体单一基因单元为标准序列，使用 Clustal W 软件确定每个疑似 Numt 中相应基因的起始位置(Thompson et al, 1994)，DNA 权重矩阵(Weight Matrix)设置为 IUB，空位开放罚分（Gap Opening Penalty）和空位延伸罚分(Gap Extension Penalty)分别设置为 10.00 和 0.20。使用 DnaSP 软件将所获得的疑似 Numt 序列翻译成蛋白质序列，分别选择哺乳动物线粒体密码子和细胞核通用密码子（Nuclear Universal）作为指导(Rozas & Rozas, 1999)。

1.4　tRNA 结构预测

核插入序列中的一些 tRNA(Numt-tRNA)的序列组成依然完整，因此通过对 Numt-tRNA 及其相应的 Mt-tRNA 的二级结构比较，鉴定 Numt-tRNA 是否发生结构变异，推断其是否仍然存在生物学功能。使用 tRNAscan-SE 软件寻找疑似 Numt 中的 tRNA，参数选择如下：Source 设定为 Mito/Chloroplast; Genetic Code for tRNA Isotype Prediction 设定为 Vertebrate Mito; 其余参数为默认(Lowe & Eddy, 1997)。tRNA 二级结构预测软件使用 tRNAscan-SE 1.21 和 Mfold 软件，折叠的温度设定为 37°，其余参数为默认(Zuker, 2003)。

2　结　果

2.1　家马基因组中疑似 Numt 的数目与长度

当BLASTN的最大期望值设定为$e=10^{-4}$时，家马核基因组中总共搜索到 200 个与线粒体序列相似的疑似Numt片段。覆盖线粒体基因组达到 800 bp以上区域的疑似Numt数目共有 10 个(5%)，而能够达到 3 kb以上的片段仅为 1 个。在片段长度大于 800 bp 的疑似Numt中，仅有 1 个包括控制区域（表 1 中第 14 号疑似Numt），并且所包括的控制区域仅仅只有 49 bp。表 1 为这些疑似NUMT在家马核基因组中的情况统计，Start和End表示疑似Numt在家马核基因组和线粒体基因组中的起始位置。Orientation (Ori)对应整合进入核基因组的方向是 5'>3'(+)或者 3'>5'(−)。E-val和ID分别显示BLAST中返回的期望值和相似性的数值。所有的疑似Numt片段长度之和为 44 373 bp，大约是家马线粒体基因组全序列的 2.66 倍，约为核基因组序列的 0.002154%。图 1 为疑似Numt长度与相似性示意图，其中最长的序列长度为 3 727 bp，相似性为 93.99%；最短的序列长度为 29 bp，相似性为 100%。序列相似性的范围为 79.2%—100%。由表 2 所示，疑似Numt在线粒体基因组中未覆盖的区域长度之和为 789 bp，约线粒体基因组全序列长度的 4.74%。

2.2　家马疑似 Numt 中相应线粒体基因

家马线粒体基因组中共有 13 个编码蛋白、2 个核糖体 RNA 和 22 个 tRNA, 疑似 Numt 中包含结构相对完整的 9 个编码蛋白、20 个 tRNA 和 2 个核糖体 RNA，而 ND2、COX2、ATP8、ND4 和控制区的部分序列在疑似 Numt 中没有发现。尽管疑似 Numt 中存在 9 个相对结构完整的编码蛋白，但是无论是使用哺乳动物线粒体密码子还是使用细胞核通用密码子，在这些编码蛋白中都存在开放阅读框内部的终止密码子和/或移码突变。此外，也未发现结构完整的线粒体控制区域疑似 Numt 存在。

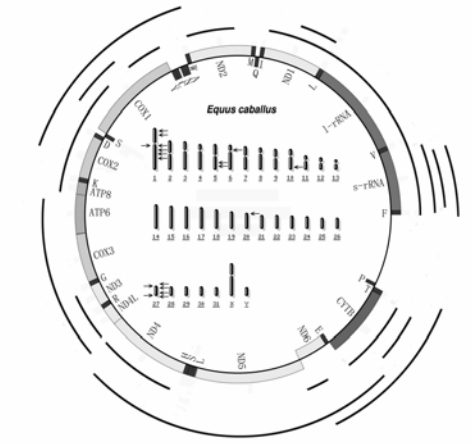

图 1　Numt 长度与相似性示意图

Fig. 1　The distributions of the Numts length and similarities between different length Numts and mtDNA 核型图上的箭头表示疑似 Numt 整合位置；外周线条表示疑似 Numt 在线粒体基因组上覆盖区域。The integration sites in the karyotype map are indicated by solid arrows; the location of each Numts is covered by peripheral line.

tRNA 二级结构预测的结果显示，除了位于 21 号染色体上的 1 个 tRNA-Arg 线粒体插入序列以外，所有的 Numt-tRNA 由于存在非 Waston-crick 碱基对或者被改变茎环结构,因此不能准确的折叠成相应的结构（图 2）。该 21 号染色体上的插入序列长度为 84 bp，含有部分 ND3 和 ND4L 的序列，与线粒体中的对应区域相似性为 100%。

2.3　核基因中疑似 Numt 的所在区域

根据疑似 Numt 在染色体上的位置，通过登陆 Ensembl 在线图形界面基因组浏览器对疑似 Numt

表 1 疑似 Numt(>500 bp)在家马核基因组中的分布

Tab. 1 Numts Distribution (>500 bp) in the domestic horse nuclear genome

编号 NO.	所包括基因 Genes included	mtDNA 起点 Start mtDNA	mtDNA 终点 End mtDNA	长度 Length	染色体 Chrom	染色体起点 Start chrom	染色体终点 End chrom
1	tRNA-Phe s-rRNA tRNA-Val l-rRNA tRNA-Leu ND1	1	3727	3727	1	136277800	136281515
2	tRNA-Phe s-rRNA	41	959	919	1	21135884	21136799
3	s-rRNA	159	805	648	1	108586953	108586306
4	COX1	5364	5880	517	1	66913138	66913654
	COX1	5966	6736	771	1	66913721	66914490
5	CYTB	14194	15093	900	1	105391947	105391053
6	CYTB	14563	15091	531	1	136275828	136276358
7	COX1 tRNA-Ser tRNA-Asp COX2 tRNA-Lys ATP8	6267	8006	1743	5	64470402	64468670
8	ND5 ND6	13358	13887	530	5	60735755	60736284
9	ND3 tRNA-Arg ND4L ND4	9684	10224	541	6	29859449	29858912
10	tRNA-Leu ND1	2698	3234	537	10	70230223	70229687
11	tRNA-Phe s-rRNA	1	804	806	20	3169875	3170675
12	ND2	4250	4934	686	27	4654943	4654258
	ND2 tRNA-Trp tRNA-Ala tRNA-Asn origin_of_L_strand_replication tRNA-Cys tRNA-Tyr COX1 tRNA-Ser tRNA-Asp COX2	4948	7457	2510	27	4654258	4651753
13	ATP6 COX3 tRNA-Gly ND3 tRNA-Arg ND4L ND4	8065	10235	2177	27	36711445	36709276
	ND4	10320	11438	1123	27	36709186	36708064
	ND4 tRNA-His tRNA-Ser tRNA-Leu ND5 ND6 tRNA-Glu CYTB	11499	14435	2938	27	36707963	36705026
14	ND6 tRNA-Glu CYTB tRNA-Thr tRNA-Pro control_region	14036	15518	1483	27	4659109	4657629

编号 NO.	所包括基因 Genes included	不匹配数目 Mismatch	数目 Gap	分值 Blast score	E 值 E-val	相似性 ID (%)	方向 Ori
1	tRNA-Phe s-rRNA tRNA-Val l-rRNA tRNA-Leu ND1	213	6	5574	0.00E+00	93.99	+
2	tRNA-Phe s-rRNA	32	3	1520	0.00E+00	96.19	+
3	s-rRNA	7	1	1213	0.00E+00	98.77	−
4	COX1	101	0	224	8.00E-56	80.46	+
	COX1	129	1	490	9.00E-136	83.14	+
5	CYTB	154	4	494	5.00E-137	82.33	−
6	CYTB	46	2	656	0.00E+00	90.96	+
7	COX1 tRNA-Ser tRNA-Asp COX2 tRNA-Lys ATP8	261	8	1229	0.00E+00	84.28	−
8	ND5 ND6	32	0	797	0.00E+00	93.96	+
9	ND3 tRNA-Arg ND4L ND4	81	3	383	6.00E-104	84.47	−
10	tRNA-Leu ND1	5	0	1025	0.00E+00	99.07	−
11	tRNA-Phe s-rRNA	84	7	821	0.00E+00	88.71	+
12	ND2	77	1	733	0.00E+00	88.63	−
	ND2 tRNA-Trp tRNA-Ala tRNA-Asn origin_of_L_strand_replication tRNA-Cys tRNA-Tyr COX1 tRNA-Ser tRNA-Asp COX2	239	3	3027	0.00E+00	90.32	−
13	ATP6 COX3 tRNA-Gly ND3 tRNA-Arg ND4L ND4	31	11	3883	0.00E+00	97.98	−
	ND4	16	4	2036	0.00E+00	98.22	−
	ND4 tRNA-His tRNA-Ser tRNA-Leu ND5 ND6 tRNA-Glu CYTB	33	1	5547	0.00E+00	98.84	−
14	ND6 tRNA-Glu CYTB tRNA-Thr tRNA-Pro control_region	144	1	1776	0.00E+00	90.16	−

整合位置的邻近区域进行分析。发现整合位置一般都不在已知或者预测的基因内部，邻近区域也没有 cDNA，表达序列标签或者同源蛋白质质存在。整合位置的邻近区域一般富含简单重复，串联重复或者长散布重复序列，但是未发现这些重复序列与疑似 Numt 存在任何明显的关联。如图 3，在相对较长的 14 个疑似 Numt（>500 bp）中，位于第 1、27 和 5 号染色体的疑似 Numt 数目分别为 6、3 和 2，第 6、10 和 20 号染色体上的疑似 Numt 数目皆为 1 个。所有疑似 Numt 中长度最长的片段位于第 1 号染色体，该疑似 Numt 的覆盖区域包含有 tRNA-Phe、12S rRNA、tRNA-Val、16S rRNA、tRNA-Leu 和 ND1 基因部分同源序列。对位于 21 号染色体上、似性为 100%、长度为 84 bp 的这一个疑似 Numt 整

表 2　家马疑似 Numt 在线粒体中未覆盖的对应区域

Tab. 2　Uncovered region in the horse mitochondrial counterparts of Numt

未覆盖区域 Uncovered Region		长度 Length (bp)	所包括基因 Gene Included	基因位置 Gene Location	
起点 Start	终点 End			起点 Start	终点 End
3954	3980	27	ND2	3937	4977
4167	4249	83	ND2	3937	4977
4935	4936	2	ND2	3937	4977
7566	7649	84	COX2	7048	7731
7746	7748	3	tRNA-Lys	7735	7802
7813	7948	136	ATP8	7804	8007
10236	10264	29	ND4	10205	11582
10317	10319	3	ND4	10205	11582
11439	11498	60	ND4	10205	11582
15590	15684	95	tRNA-Pro	15468	15403
16121	16387	267	Control region	15469	16660

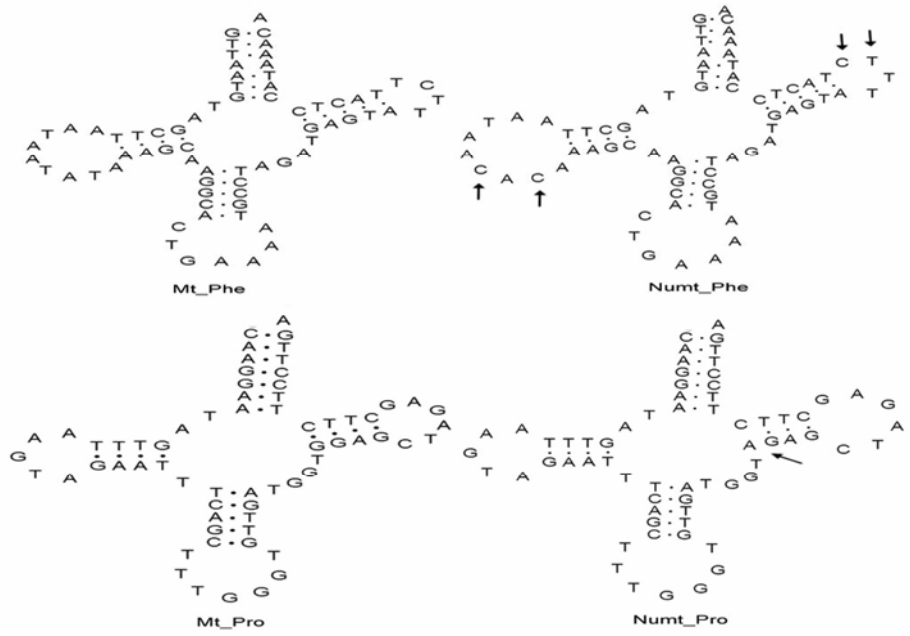

图 2　tRNA 二级结构预测

Fig. 2　Prediction of secondary structure for tRNAs

箭头表示疑似 Numt 中发生变异的位点(Mismatches in the alignment are indicated by arrows)。

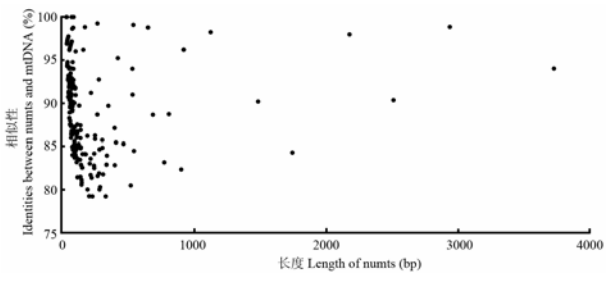

图 3　疑似 Numt 在家马线粒体中的分布

Fig. 3　Representation of the BLASTN Numts detected in the domestic horse genomes surveyed and their homology with the mtDNA genome

合位点附近区域进行检查，未发现存在由于插入、缺失或者重排等原因造成疑似 Numt 断裂的现象。

3　讨　论

在真核生物的核基因组人工细菌染色体（Bacterial artificial chromosome, BAC）文库构建、序列组装和基因组图谱测定的过程中，mtDNA 污染是一个不可忽视的问题。由于高等植物中的线粒体基因组长度较长（NCBI Organelle Genome Resources, http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html），容易在 BAC 文库构建过程中被克隆进入载体，因此对于植物基因组

文库中 mtDNA 污染的评估成为一项常规性操作(Yoo et al, 2004; Wang et al, 2005)。而类似家马这种高度脊椎动物的线粒体基因组一般约 16 kb 左右，对于平均插入长度至少大于 100 kb 的 BAC 文库，mtDNA 的污染很容易在文库构建过程中被排除，因此存在 mtDNA 污染的可能性微乎其微(Leeb T et al, 2006)。故本研究中所使用的家马核基因组序列中，所鉴定的疑似 Numt 肯定并非源自基因组序列组装过程中的 mtDNA 污染。

家马 Numt 存在下列 3 种现象，因此推测这些转入细胞核中的序列已经丧失功能：(1)由于线粒体和细胞核内遗传密码的差异使得在开放阅读框中出现非正常终止密码子或者移码突变；(2)控制区域的缺乏或者不完全造成无法进行正常的转录；(3)某些功能区域出现序列缺失，使得蛋白质或者 RNA 高级结构受到破坏。上述 3 种现象在原鸡和家猫基因组中也被发现(Pereira & Baker, 2004; Kim et al, 2006)。

一般而言，所有的动物线粒体 tRNA 都属于一型 tRNA，具有典型的三叶草结构特征。家马的疑似 Numt_tRNA 与牛(*Bos taurus*)、原鸡一样，整合进入核基因组中以后，由于受到的选择压力减轻而逐渐退化产生许多变异位点。存在这些变异位点造成 tRNA 二级结构中形成非 Waston-crick 碱基对或者改变茎环结构，必然影响其结构的稳定性(Pereira & Baker, 2004; Liu & Zhao, 2007)。因此，二级结构预测结果显示绝大部分的 Numt-tRNA 都不能准确的折叠成相应结构，推测这些 Numt-tRNA 可能已经丧失功能，被称作“到达即死”（dead on arrival）(Antunes & Ramos, 2005)。但是，在本研究中发现一个疑似 Numt 包括序列相似性为 100% 的 tRNA-Arg，高度的序列相似性推测该插入序列可能是最近转移进入线粒体基因组的。由于该 Numt 整合区域附近未发现存在由于插入、缺失或者重排等原因造成的 Numt 断裂现象，说明此 84 bp 的 Numt 插入是经过一次独立的核转移过程，暗示片段较短的 Numt 也有可能转入核基因中。

对家马基因组中 Numt 特征和分布的系统地分析，为马科群体遗传学和系统发生学研究中的 Numt 识别提供参考资料。在 PCR 过程中，Numt 与线粒体的扩增产物往往长度相似，并且偏好扩增或者共增(Zhang & Hewitt, 1996)。因此，在线粒体 DNA 研究中当其不慎被扩增容易混淆分析结论，特别是在使用进化速率相对较慢的基因序列作为分子标记的时候(van et al, 1995)。在群体遗传学和生物地理学研究中，线粒体控制区域的序列常常被用作分子标记物，通过 PCR 产物测序的策略获得该区域的多态信息(Aberle et al, 2007; Kakoi et al, 2007)。比较幸运的是在马群体遗传学研究中, 通常覆盖控制区域的片段长度远大于 49 bp，PCR 扩增的片段往往大于 800 bp，因此 Numt 对于基于线粒体控制区域序列的群体遗传学研究的结果不会产生影响(Harrison & Turrion-Gomez, 2006; Aberle et al, 2007; Kakoi et al, 2007)。如果需要使用线粒体控制区域以外的其他序列作为分子标记物时，应该采取一些措施避免 Numt 的扩增(Greenwood & Paabo, 1999; Wang, 2004)：(1)选择线粒体含量相对较高的组织提取 DNA；(2)根据所研究的特定物种分类设计扩增引物，特别是在引物的 3'端尤为重要；(3)以长片段 PCR 产物为模板进行二次 PCR。此外，如果获得的序列结果中编码基因区域出现移码突变、终止密码子、碱基组成或者转换/颠换比率的较大差异等现象，也说明可能存在 Numt 的污染(Mirol et al, 2000)。

随着日益增多的全基因组序列公布，比较和分析 Numt 在物种间的进化过程得以实现。对 Numt 进化史、染色体上分布位置及数量的特征分析，尤其是对亲缘关系非常接近物种之间的分析，能够更好地阐明 Numt 的进化规律(Bensasson et al, 2001)。利用近缘物种研究生物进化的手段目前已经基本具备，而且由于人类及其相近物种核基因组序列的测定，因此受到更多研究者的青睐(Chen et al, 2001; Ricchetti et al, 2004)。但是，目前正在进行或者已经完成的基因组计划更多的关注于生物医学领域或者分类地位上重要的一些物种(O'Brien et al, 2001)，除了人类以外的近缘物种在整个基因组水平比较的 Numt 研究较少，更多的只是关注于部分区域的研究(Mirol et al, 2000; Kim et al, 2006; Martins et al, 2007)。相信在不久的将来，受益于新一代测序技术的发展和测序成本的降低，越来越多物种的全基因组序列将会被解析。对更大范围近缘物种间的比较，有助于研究 Numt 的起源、分化和进化历史，并且为更准确的基因组功能注释提供参考。

参考文献：

Aberle KS, Hamann H, Drogemuller C, Distl O. 2007. Phylogenetic relationships of German heavy draught horse breeds inferred from mitochondrial DNA D-loop variation [J]. *J Anim Breed Genet*, **124**(2): 94-100.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool [J]. *J Mol Biol*, **215**(3): 403-410.

Antunes A, Pontius J, Ramos MJ, O'Brien SJ, Johnson WE. 2007. Mitochondrial introgressions into the nuclear genome of the domestic cat [J]. *J Heredy*, **98**(5): 414-420.

Antunes A, Ramos MJ. 2005. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes [J]. *Genomics*, **86**(6): 708-717.

Bensasson D, Zhang D, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses [J]. *Trends Ecol Evol* (*Personal edition*), **16**(6): 314-321.

Chen FC, Vallender EJ, Wang H, Tzeng CS, Li WH. 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences [J]. *J Hered*, **92**(6): 481-489.

du Buy HG, Riley FL. 1967. Hybridization between the nuclear and kinetoplast DNA's of *Leishmania enriettii* and between nuclear and mitochondrial DNA's of mouse liver [J]. *Proc Natl Acad Sci*, **57**(3): 790-797.

Greenwood AD, Paabo S. 1999. Nuclear insertion sequences of mitochondrial DNA predominate in hair but not in blood of elephants [J]. *Mol Ecol*, **8**(1): 133-137.

Harrison SP, Turrion-Gomez JL. 2006. Mitochondrial DNA: an important female contribution to thoroughbred racehorse performance [J]. *Mitochondrion*, **6**(2): 53-63.

Hazkani-Covo E, Sorek R,Graur D. 2003. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications [J]. *J Mol Evol*, **56**(2): 169-174.

Henze K, Martin W. 2001. How do mitochondrial genes get into the nucleus? [J]. *Trends Genet*, **17**(7): 383-387.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M. 2002. The Ensembl genome database project [J]. *Nucleic Acids Res*, **30**(1): 38-41.

Kakoi H, Tozaki T,Gawahara H. 2007. Molecular analysis using mitochondrial DNA and microsatellites to infer the formation process of Japanese native horse populations [J]. *Biochem Genet*, **45**(3-4): 375-395.

Kim JH, Antunes A, Luo SJ, Menninger J, Nash WG, O'Brien SJ, Johnson WE. 2006. Evolutionary analysis of a large mtDNA translocation (numt) into the nuclear genome of the Panthera genus species [J]. *Gene*, **366**(2): 292-302.

Leeb T, Vogl C, Zhu B, de Jong PJ, Binns MM, Chowdhary BP, Scharfe M, Jarek M, Nordsiek G, Schrader F, Blocker H. 2006. A human-horse comparative map based on equine BAC end sequences [J]. *Genomics*, **87**: 772-776

Liu Y, Zhao X. 2007. Distribution of nuclear mitochondrial DNA in cattle nuclear genome [J]. *J Anim Breed Genet*, **124**(5): 264-268.

Lopez JV, Yuhki N, Masuda R, Modi W,O'Brien SJ. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat [J]. *J Mol Evol*, **39**(2): 174-190.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence [J]. *Nucleic Acids Res*, **25**(5): 955-964.

Martins JJ, Solomon SE, Mikheyev AS, Mueller UG, Ortiz A, Bacci M Jr. 2007. Nuclear mitochondrial-like sequences in ants: evidence from *Atta cephalotes* (Formicidae: Attini) [J]. *Insect Mol Biol*, **16**(6): 777-784.

Mirol PM, Mascheretti S, Searle JB. 2000. Multiple nuclear pseudogenes of mitochondrial cytochrome b in Ctenomys (Caviomorpha, rodentia) with either great similarity to or high divergence from the true mitochondrial sequence [J]. *Heredity*, **84** (Pt 5): 538-547.

Mourier T, Hansen AJ, Willerslev E, Arctander P. 2001. The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus [J]. *Mol Biol Evol*, **18**(9): 1833-1837.

O'Brien SJ, Eizirik E, Murphy WJ. 2001. Genomics: On choosing mammalian genomes for sequencing [J]. *Science*, **292**(5525): 2264-2266.

Pereira SL, Baker AJ. 2004. Low number of mitochondrial pseudogenes in the chicken (*Gallus gallus*) nuclear genome: implications for molecular inference of population history and phylogenetics [J]. *BMC Evol Biol*, **4**: 17.

Ricchetti M, Tekaia F, Dujon B. 2004. Continued colonization of the human genome by mitochondrial DNA [J]. *PLoS Biol*, **2**(9): E273.

Richly E, Leister D. 2004. NUMTs in sequenced eukaryotic genomes [J]. *Mol Biol Evol*, **21**(6): 1081-1084.

Rozas J, Rozas R. 1999. DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis [J]. *Bioinformatics*, **15**(2): 174-175.

Thalmann O, Hebler J, Poinar HN, Paabo S, Vigilant L. 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes [J]. *Mol Ecol*, **13**(2): 321-335.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice [J]. *Nucleic Acids Res*, **22**(22): 4673-4680.

Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P. 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes [J]. *Genomics*, **80**(1): 71-77.

van der Kuyl AC, Kuiken CL, Dekker JT, Perizonius WR, Goudsmit J. 1995. Nuclear counterparts of the cytoplasmic mitochondrial 12S rRNA gene: a problem of ancient DNA and molecular phylogenies [J]. *J Mol Evol*, **40**(6): 652-657.

Wallace DC, Stugard C, Murdock D, Schurr T, Brown MD. 1997. Ancient mtDNA sequences in the human nuclear genome: a potential source of errors in identifying pathogenic mutations [J]. *Proc Natl Acad Sci*, **94**(26): 14900-14905.

Wang JW. 2004. Numts Identification and utility in evolutionary biology [J]. *Chinese Journal of Zoology*, **39**(3): 103-108.[王继文. 2004. 动物线粒体假基因的识别及其在进化生物学中的应用. 动物学杂志, **39**(3): 103-108.]

Wang W, Tanurdzic M, Luo M, Sisneros N, Kim HR, Weng JK, Kudrna D, Mueller C, Arumuganathan K, Carlson J, Chapple C, de Pamphilis C, Mandoli D, Tomkins J, Wing RA, Banks JA. 2005. Construction of a bacterial artificial chromosome library from the spikemoss *Selaginella moellendorffii*: a new resource for plant comparative genomics [J]. *BMC plant biology*, **5**: 10.

Woischnik M, Moraes CT. 2002. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome [J]. *Genome Res*,

12(6): 885-893.

Xu X, Arnason U. 1994. The complete mitochondrial DNA sequence of the horse, *Equus caballus*: Extensive heteroplasmy of the control region [J]. *Gene*, **148**(2): 357-362.

Yoo EY, Kim S, Kim YH, Lee CJ, Kim BD. 2003. Construction of a deep coverage BAC library from *Capsicum annuum*, 'CM334' [J]. *TAG Theoretical and applied genetics*, **107**: 540-543.

Zhang DX, Hewitt GM. 1996. Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: some implications for population studies [J]. *Mol Ecol*, **5**(2): 295-300.

Zhang FY, Ma LB, Qiao ZG, Ma CY. 2006. Separation and characterization of mitochondrial COI pseudogenes in *Scylla paramamosain* [J]. *Heredity*, **28**(1): 43-49[张凤英, 马凌波, 乔振国, 马春艳. 2006. 青蟹线粒体 COI 假基因的分离和特征分析. 遗传, **28**(1): 43-49.]

Zischler H, Hoss M, Handt O, von Haeseler A, van der Kuyl AC, Goudsmit J. 1995. Detecting dinosaur DNA [J]. *Science*, **268**(5214): 1192-1193; author reply 1194.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction [J]. *Nucleic Acids Res*, **31**(13): 3406-3415.

中国鸟类新记录——斑[姬]鹟

The European Pied Flycatcher, *Ficedula hypoleuca*, a New Record of Bird in China

马 鸣，梅 宇，胡宝文

MA Ming, MEI Yu, HU Bao-wen

（中国科学院新疆生态与地理研究所，新疆 乌鲁木齐 830011）

（*Xinjiang Institute of Ecology and Geography, the Chinese Academy of Sciences, Urumqi* 830011, *China*）

关键词：斑[姬]鹟；新记录；克里雅河；昆仑山；新疆

Key words：*Ficedula hypoleuca*；New record in China；Keriya River；Kunlun Mountains；Xinjiang

中图分类号：Q959.7 **文献标识码**：A **文章编号**：0254-5853-(2008)06-0584-01

2008 年 10—11 月在昆仑山考察期间，在野外拍摄到 20 余张斑[姬]鹟[*Ficedula hypoleuca*（Pallas, 1764）]的生态照片（其一请见本期封面照片），经过专家比对和文献查证，确定为中国鸟类一新记录种。该鸟的发现地点在新疆南部和田地区于田县阿羌乡普鲁村（36º11'15.6" N，81º28'56.9" E），海拔 2 601 m。属于昆仑山南麓与塔克拉玛干沙漠之间的一个峡谷绿洲，位于克里雅河的上游。生境有园林、农田和荒漠草原。记录时间为 2008 年 10 月 23 日下午，当时与其混群的鸟种有红背红尾鸲（*Phoenicurus erythronotus*）、红胁蓝尾鸲（*Tarsiger cyanurus*）、戴菊（*Regulus regulus*）、淡眉柳莺（*Phylloscopus humei*）、叽咋柳莺（*Phylloscopus collybita*）、赤颈鸫（*Turdus atrogularis*）等。附近山坡上有雀鹰（*Accipiter nisus*）、红隼（*Falco tinnunculus*）、石鸡（*Alectoris chukar*）、岩鸽（*Columba rupestris*）、小鸮（*Athene noctua*）、角百灵（*Eremophila alpestris*）、地山雀（*Pseudopodoces humilis*）、褐岩鹨（*Prunelle fulvescens*）、树麻雀（*Passer montanus*）等几十种鸟类。

斑[姬]鹟体长 120—140 mm，翅长 75—84 mm，体重 10—15 g，迁徙季节可达 20.5 g。斑[姬]鹟个体略小于斑鹟（*Muscicapa striata*），而大于小斑[姬]鹟（*Ficedula westermanni*）。通体为黑白二色（♂）或者褐色及白色（♀）。雄鸟上体黑色，额前有白色斑，尾和翅黑色，翼斑及尾基部羽缘白色，下体包括尾下覆羽均为白色。雌鸟上体灰褐色，翼斑白色，下体近白。亚成鸟与雌鸟接近，通体为褐色及白色。斑[姬]鹟与小斑[姬]鹟比较相似，主要区别在于斑[姬]鹟缺白色眉纹。

(下转第 602 页)

收稿日期: 2008-11-20; 接受日期: 2008-12-02