文章编号：1001-4721(2012)02-0001-05

鹿、牛基因组不同区域的微小变异研究

巴恒星，杨福合，李春义

(1.中国农业科学院特产研究所 吉林省特种经济动物分子生物学省部共建实验室，吉林 吉林 132109)

摘要：本研究借助鹿和牛基因组序列，通过同源序列比较的方法研究远缘物种基因组不同区域的微小变异，包括单碱基突变、小片段插入和删除。研究结果验证了基因组功能区的点突变和删除变异相对非功能区是保守的普遍性结论。点突变变异与删除变异在鹿、牛基因组不同区域上表现强的正相关性。比较近缘物种人、黑猩猩基因组点突变变异数据，表明牛、鹿基因组的功能区和非功能区突变速率各自保持大致恒定，符合分子钟理论。

关键词：远源物种 基因组 单碱基突变 插入和删除

中图分类号：Q344+.12　　**文献标识码：**A

Micro-Variation in Different Regions between Deer and Cattle Genome Sequences

BA Heng-xing, YANG Fu-he, LI Chun-yi

(State Key Laboratory of Special Economic Animal Molecular Bioloy Institute of Special Wild Economic Animals and Plants, Chinese Academy of Agricultural Sciences, Jilin, 132109, China)

Abstract: In this study, sika deer genome sequence was used to compare with the cattle reference genome using the methods of homology comparison, which was used to identify the Micro-Variation in different regions between distant species genomes in the past. These Micro-Variations consisted of point mutation variations (SNP) and small nucleotide insertion and deletion variations (Indel). The results verified the general conclusion that the variations of SNP and Indel in the functional regions are more conservative than in the non-functional regions. Also, the results showed a strong positive correlation between the SNP variations and Indel variations among different regions. The variation data between human and chimpanzee are compared with those of this study. The results verified that the mutation rate has kept in consistent in the genome functional regions and non-functional regions respectively. The results validated once again the theory of molecular clock.

Key words: Distant species; Genomes; SNP; Indel

在物种漫长的形成和演变进程中，基因组水平上的同源区域之间积累了大量的微小变异，主要包括单碱基突变(SNP)、小片段的插入和删除(Indel)。这些微小变异提供了物种形成的原始动力，其变异程度与物种亲缘关系远近有直接关系，决定了物种之间在体形外貌和生理结构等方面的差异，是物种多样性的本质体现。基因组水平上的SNP和Indel变异研究在近缘物种之间已经广泛开展，特别是人

收稿日期：2012-04-09

作者简介：巴恒星(1979-)，男，辽宁省抚顺人，在读博士研究生，从事特种动物基因组学研究。

※通讯作者：李春义，E-mail：chunyi.li@agresearch.co.nz.

与黑猩猩基因组之间的变异研究[1,2]。随着新一代高通量测序技术的快速发展，越来越多的物种基因组序列相继完成，因此，提供更多的机会去比较远缘物种（与人、黑猩猩亲缘关系相比）的基因组序列来研究 SNP 和 Indel 变异。

化石信息和分子数据研究表明，反刍类物种中的牛和鹿共同祖先在大约 2 500~2 800 万年前开始分化[3]，早于人与黑猩猩分化时间（500~700 万年）2 000 万年。按照分子钟理论，物种的进化时间与 DNA 等分子进化速度成一定的线性关系，或者说 DNA 分子进化速率保持恒定性[4]。为此，本研究拟利用梅花鹿基因组序列和牛基因组参考序列为研究对象，研究远缘物种基因组不同区域的 SNP 和 Indel 变化规律及其特点，并与近缘物种人、黑猩猩之间的变异数据进行比较。本研究的基因组不同区域包括基因间的间隔区（Intergenic），其中的基因上游 5 000bp 区（Upstream）和基因下游 5 000bp 区（Downstream）以及外显子（Exon）、内含子（Intron）编码区（Coding）和外显子中的上游非翻译区（5′- UTR）和下游非翻译区（3′- UTR）。其中，编码区和调控区属于功能区，基因间隔区和内含子区属于非功能区。

1 材料与方法

1.1 材料

中国梅花鹿基因组序列数据（利用 SOLiD 测序技术对中国梅花鹿基因组进行了全鸟枪法测序，组装后得到约 400 万个 contigs，包含的碱基总量约 1.62Gbp）和牛基因组参考序列（Ensemble 网站下载，版本 UMD3.1，格式 FASTA）及牛基因组结构注释信息（GTF 格式文件）。

1.2 方法

自定义 Python 脚本，依据 GTF 格式的基因结构注释信息，从牛基因组参考序列中分别提取并产生基因组不同区域的 FASTA 格式序列文件。利用 RepeatMasker 软件（http://www.repeatmasker.org）和牛重复序列数据库分别对鹿基因组序列和牛基因组不同区域序列中的重复序列进行硬掩盖。然后，利用 MUMmer v3 软件包（http://mummer.sourceforge.net）中的 mucmer 软件将鹿基因组序列和牛基因组不同区域序列进行比对。为了降低同源序列匹配的假阳性，要求比对相似性 85%以上，如果同一区域序列发生多次匹配，利用 MUMmer v3 软件包中的 delta- filter 软件筛选最好的序列匹配作为候选。最后，利用 MUMmer v3 软件包中 Show- SNPs 软件进一步对 mucmer 软件输出的结果文件进行格式化，产生包含 SNP 和 Indel 变异的可读文件。

自定义 Python 脚本，对鹿、牛基因组不同区域包含的 SNP 位点和包含 1~6 个碱基的插入和删除（1~6bp Inde）分别计算 SNP 变异百分率（SNP 数量除以总序列比对数量的百分率）和 1~6bp Indel 变异百分率（插入和缺失的碱基数量除以总序列比对数量的百分率）。

2 结果与分析

2.1 鹿、牛基因组不同区域比对

鹿、牛基因组不同区域序列的比对结果见表 1。

表 1 鹿、牛基因组不同区域序列的比对结果

Table 1 The alignment summary in different regions between deer and cattle genome sequences

区域 Regions	匹配对数量（个） Number of alignment	匹配碱基总量（bp） Total of base	平均匹配长度（bp） Average length of alignment
Contigs	2 155 791	792 653 019	368
5′- UTR	1,746	244 896	140
3′- UTR	11 292	3 426 694	303
Coding	66 412	10 411 808	157
Exon	73 833	13 172 447	178
Intron	415 329	148 888 636	358
Upstream	103 561	30 937 307	299
Downstream	102 731	34 067 748	332
Intergenic	1 416 242	517 712 958	366

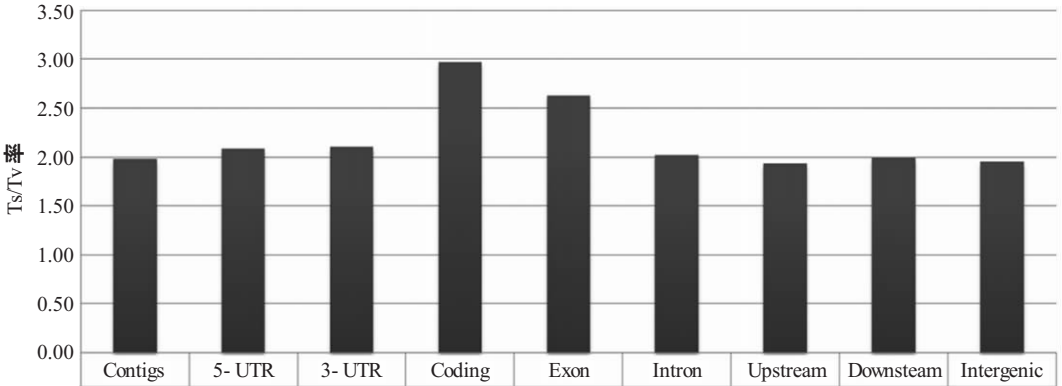
注：Contigs 序列代表鹿基因组序列。 Note:Contigs represents deer genome sequence

其中，Contigs 序列代表鹿基因组序列。Contigs 的匹配碱基总量约为 0.79Gbp，如果排除哺乳动物基因组序列中约 50%重复序列和多拷贝序列，鹿基因组数据中仅一对一匹配的序列碱基量约为 0.81Gbp

（1.62Gbp× 50%），表明鹿与牛大部分同源单拷贝序列相似性在 85%以上。基因组中 5′- UTR 区域 GC 含量偏高[5]，SOLiD 测序 GC 含量高的序列区域会出现覆盖率偏低的现象[6]，导致鹿基因组序列中低覆盖率的 5′- UTR 序列难以组装，进一步导致表 1 中的 5′- UTR 序列的匹配数量和碱基总量比 3′- UTR 序列少，平均匹配长度也相对短。如果考虑重复序列约占基因组的 50%和 Exon 区域测序偏倚，表 1 中的 Exon 序列和 Contigs 序列碱基量比值将从 1.67%降到 1%左右，这与人外显子测序报道相一致[7]，进一步表明哺乳动物基因组中 Exon 区域约占整个基因组的 1%。另外，Coding 和 Exon 功能区的平均匹配长度偏短，其它非功能区匹配长度偏长，这些结果也与哺乳动物基因组上不同区域结构相一致。

2.2　鹿、牛基因组不同区域的 SNP 转换 / 颠换比率

统计分析表明（图 1），鹿、牛基因组之间的 SNP 转换 / 颠换比率为 1.98:1，符合转换 / 颠换比率为 2:1 的规律。然而，Coding 和 Exon 功能区转换 / 颠换比率都显著高于 2。分析可知，牛、鹿基因组中具有相同功能的同源基因在选择压力下，同义密码子之间的大部分碱基突变是转换而不是颠换。其它非功能区的转换 / 颠换比率也接近 2:1。



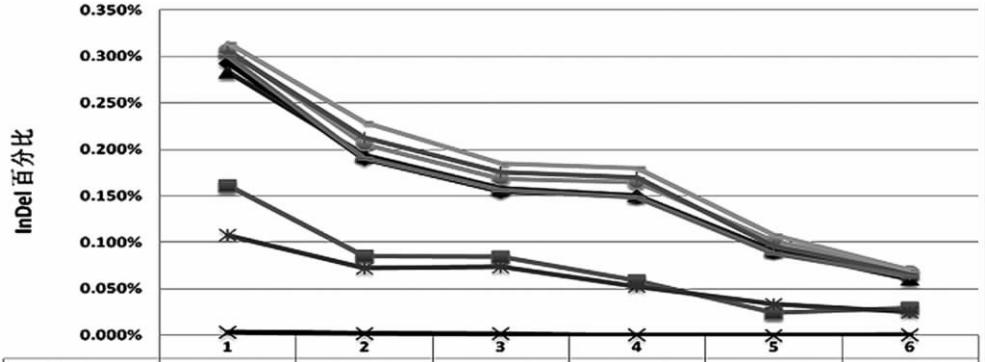
	Contigs	5- UTR	3- UTR	Coding	Exon	Intron	Upstream	Downsteam	Intergenic
Ts/Tv 率	1.98	2.09	2.11	2.97	2.63	2.02	1.94	1.99	1.95

图 1　　鹿、牛基因组不同区域的 SNP 转换 / 颠换比率（Ts/Tv）

Figure 1　The rate of SNP transversion/transition in different regions between deer and cattle genome sequences

2.3　鹿、牛基因组不同区域的 1~6bp Indel

统计分析表明（图 2），基因组中的不同区域中的 Indel 百分率随 Indel 长度的增加而减少。按照 1~6bp Indel 百分率含量，可以将基因组不同区域划



	1	2	3	4	5	6
Contigs	0.293%	0.190%	0.154%	0.150%	0.090%	0.063%
5-UTR	0.161%	0.085%	0.085%	0.059%	0.025%	0.029%
3-UTR	0.283%	0.194%	0.158%	0.150%	0.092%	0.060%
Coding	0.0034%	0.0020%	0.0016%	0.0002%	0.0001%	0.0003%
Exon	0.108%	0.072%	0.074%	0.053%	0.033%	0.025%
Intron	0.307%	0.206%	0.168%	0.165%	0.102%	0.069%
Upstrean	0.305%	0.212%	0.175%	0.170%	0.096%	0.066%
Downstream	0.315%	0.229%	0.185%	0.180%	0.108%	0.070%
Intergenic	0.300%	0.191%	0.156%	0.148%	0.088%	0.063%

注：横坐标表示 1~6bp 的 Indel。　Note: x- axis represents 1~6bp Indel.

图 2　　鹿、牛基因组不同区域的 1~6bp Indel 百分率及其分布

Figure 2　The percentage of 1~6bp Indel in different regions between deer and cattle genome sequences

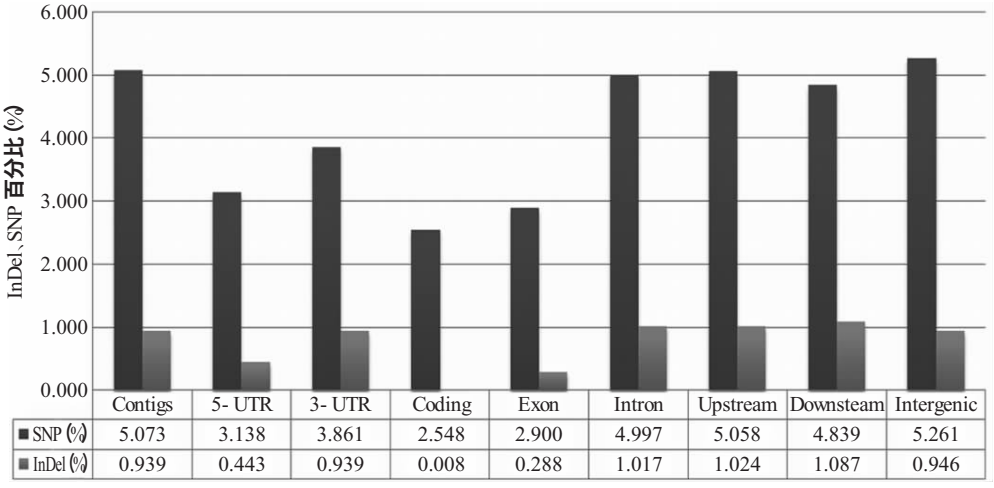
分为 3 组，第 1 组仅包括 Coding 区，该区域中出现的 Indel 百分率极低，原因是 Coding 区编码氨基酸，Indel 突变很可能导致移码突变，进而影响蛋白质生物活性，对基因功能破坏极大且大多数是致命的。更值得注意的是，6bp Indel 百分率比 4bp 和 5bp Indel 偏多，这可能是由于 6bp Indel （3 个碱基编码 1 个密码子） 不容易导致移码突变而使整个蛋白质生物活性完全消失 第 2 组包括 5′- UTR 区和 Exon 区，5′- UTR 区在蛋白质翻译过程中起到重要调控作用，Indel 变异对其也有重要影响，1~6bp Indel 百分率在 3 组中处于中间水平；第 3 组包括所有其它无功能基因组区域，1~6bp Indel 百分率接近相同且在 3 组中最高。

2.4 鹿、牛基因组不同区域的 SNP 和总 1~6bp Indel

统计分析表明 （图 3），Coding 区的 SNP 和总 1~6bp Indel 百分率都低于其它不同区域，特别是总 1~6bp Indel 百分率更低，Exon 区和 5′- UTR 区的 SNP 和总 1~6bp Indel 百分率也相对较低，这些都是基因组中重要的功能区或调控区。有趣的是，3′- UTR 调控区与其它非功能区相比，SNP 百分率要低，而总 1~6bp Indel 百分率却相当，这可能与蛋白质翻译终止的特殊结构序列的进化相关。

2.5 鹿、牛基因组不同区域 SNP 与总 1~6bp Indel 的相关关系

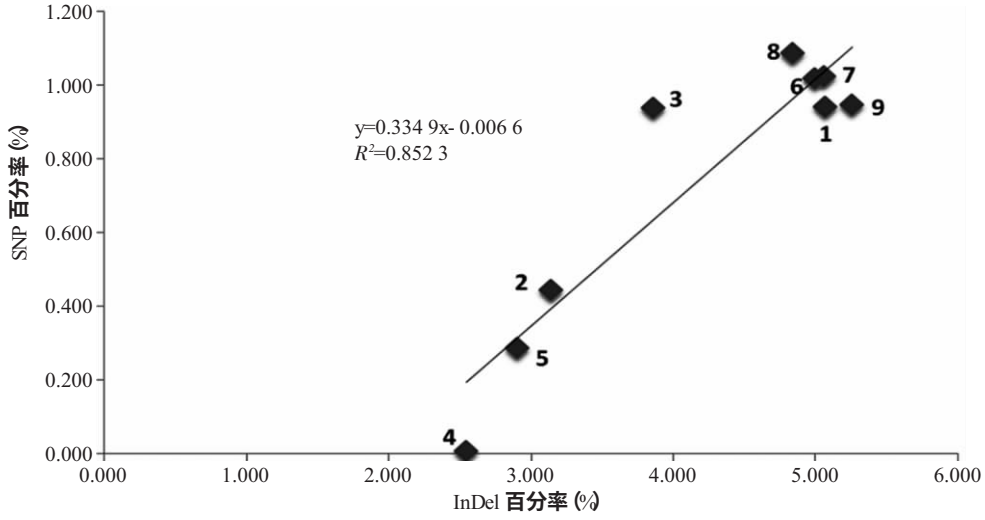
统计分析表明 (图 4)，SNP 与总 1~6bp Indel 百



	Contigs	5- UTR	3- UTR	Coding	Exon	Intron	Upstream	Downsteam	Intergenic
SNP (%)	5.073	3.138	3.861	2.548	2.900	4.997	5.058	4.839	5.261
InDel (%)	0.939	0.443	0.939	0.008	0.288	1.017	1.024	1.087	0.946

图 3 鹿、牛基因组不同区域的 SNP 和总 1~6bp Indel 百分率及其分布

Figure 3 The percentage of SNP and 1bp~6bp Indel in different regions between deer and cattle genome sequences



1. Contigs ；2. 5′- UTR ；3. 3′- UTR ；4. Coding ；5. Eoxn ；6. Intron ；7. Upstream ；8. Downsteam ；9. Intergenic

图 4 鹿、牛基因组不同区域 SNP 百分率与总 1~6bp Indel 百分率的相关关系

Figure 4 Plots of percentage of SNP in different regions between deer and cattle genome sequences against 1~6bp Indel

（下转第 8 页）

奠定了基础。

参 考 文 献

[1] Reichel MP, Hill FI, Voges H. Does control of bovine viral diarrhoea infection make economic sense? [J]. New Zealand Veterinary Journal, 2008, 56(2):60- 66.

[2] Olafson P, MacCallum AD, Fox FH. An apparently new transmissible disease of cattle[J]. Cornell Vet, 1946, 36:205- 213.

[3] 李佑民,刘振润,武银莲,等. 牛病毒性腹泻 - 黏膜病病毒株长春 184》的分离和鉴定[J]. 中国兽医学报,1983, 3(2):113- 121.

[4] 白文彬,于康震.动物传染病诊断学[M].北京:中国农业出版社, 2002:144.

[5] Luzzago C, Bandi C, Bronzo V, et al. Distribution pattern of bovine viral diarrhea virus strains in intensive cattle herds in Italy[J]. Vet Microbiolgy, 2001, 83(3):265~274.

[6] Ridpath JF. Practical significance of heterogeneity among BVDV strains: impact of biotype and genotype on U.S. control programs[J]. Prev Vet Med, 2005, 72(1- 2):17- 30.

[7] 李海涛,苗利光,刘艳环,等. 牛病毒性腹泻病活疫苗 BVDV2/JZ05- 1 株毒力返强试验[J].特产研究, 2012, 34(1): 6- 7.

[8] 李庆超. 牛病毒性腹泻病毒 JZ05- 1 株基因分型、全基因序列测定及分析[D]. 北京:中国农业科学院特产研究所, 2010.

（上接第 4 页） 分率呈现强的正相关关系，皮尔逊相关系数为 0.92。以人和黑猩猩基因组之间 1Mbp 非重叠窗口统计所有的 SNP 和 Indel 百分率，同样得出 2 种变异呈现正相关关系，皮尔逊相关系数是 0.64 [7]。我们的结果也表明，SNP 与 1~6bp Indel 在鹿、牛基因组不同区域之间有 85%的共同变异趋势。

2.6 鹿、牛与人、黑猩猩的基因组 SNP 变异数据比较

在系统进化上，鹿、牛分属于鹿科和牛科，二者遗传距离比人亚科中的人和黑猩猩遗传距离要大，鹿和牛的分化时间与人和黑猩猩的分化时间大约相差 2 000 万年。在本研究中，鹿、牛基因组功能区 SNP 百分率为 2.55%（图 3Coding 区），非功能区 SNP 百分率为 5.13%（图 3Intron 区和 Intergenic 区平均值），人、黑猩猩基因组的功能区 SNP 百分率为 0.86%，非功能区 SNP 百分率为 1.57%[8]。鹿、牛功能区 SNP 百分率约是人、黑猩猩功能区 SNP 百分率的 3.0 倍，同样非功能区 SNP 百分率也约是 3.0 倍。另外，在鹿和牛之间，非功能区 SNP 百分率约是功能区的 2.0 倍，同样在人和黑猩猩之间，非功能区 SNP 百分率也约是功能区的 2.0 倍。

这些结果显示，远缘物种功能区与非功能区序列之间进化速率保持恒定，同样，近缘物种功能区与非功能区序列之间也保持恒定；远缘与近缘物种之间的同源功能区进化速率保持恒定，同样，远缘与近缘物种之间的同源非功能区进化速率也保持恒定，表明基因组不同区域序列的分子进化速率在不同种系中恒定，符合分子钟理论。

本研究分析的是远缘物种鹿、牛全基因组的 1~6bp Indel 数据，没有找到合适的人、黑猩猩的 1~6bp Indel 参考数据，因此没有对 1~6bp Indel 数据做进一步地比较分析。

3 结语

本研究首先分析了组装的鹿基因组序列中的不同区域的序列特点。然后，通过远缘物种鹿、牛基因组不同区域的微小变异研究，得到如下结论：①相对于非功能区变异，鹿、牛基因组功能区的变异更趋保守；②基因组水平上的 SNP 变异与 Indel 变异有强的正相关性；③鹿、牛与人、黑猩猩的基因组 SNP 变异数据比较结果支持分子钟理论。

参 考 文 献

[1] Mikkelsen TS, Hillier LW, Eichler EE,et al. Initial sequence of the chimpanzee genome and comparison with the human genome[J]. Nature, 2005, 437: 69- 87.

[2] Watanabe H, Fujiyama A, Hattori M,et al. DNA sequence and comparative analysis of chimpanzee chromosome 22[J]. Nature, 2004, 429: 382- 388.

[3] Gilbert C, Ropiquet A, Hassanin A. Mitochondrial and nuclear phylogenies of Cervidae (Mammalia, Ruminantia): Systematics, morphology, and biogeography[J]. Mol Phylogenet Evol, 2006, 40(1): 101- 117.

[4] Kimura, Motoo .Evolutionary rate at the molecular level[J]. Nature, 1968, 217: 624- 626.

[5] 陈祥贵,胡军,杨潇. 人类蛋白编码基因局部 GC 水平相关性分析[J]. 遗传, 2009, 30(9): 1169- 1174.

[6] Dohm, J.C., Lottaz, et al. Substantial biases in ultra- short read data sets from high- throughput DNA sequencing [J]. Nucleic Acids Res, 2008, 36(16): 105.

[7] Wetterbom A, Sevov M, Cavelier L, Bergstrom TF. Comparative genomic analysis of human and chimpanzee indicates a key role for Indels in primate evolution [J]. J Mol Evol, 2006, 63 (15): 682- 690.

[8] Li Y, Vinckenbosch N, Tian G,et al. Resequencing of 200 human exomes identifies an excess of low- frequency non- synonymous coding variants[J]. Nat Genet, 2010, 42: 969- 972.