

鸡 Z 染色体基因表达的密码子偏性

饶友生¹, 梁菲菲², 王樟凤¹, 柴学文¹, 张细权²

(1 江西教育学院 生物技术研究, 江西 南昌 330029; 2 华南农业大学 动物科学学院, 广东 广州 510642)

摘要:从 NCBI 数据库(<http://www.ncbi.nlm.nih.gov/projects/mapview/map>)下载鸡 Z 染色体上全部基因完整的 cDNA, 最终共有 626 个基因的 CDS 序列纳入统计分析. 使用 CodonW (1.4.2) 进行密码子偏性分析, 确定了 CGG、AGC、UGC 等 26 个密码子为 Z 染色体基因表达的“最优”密码子. 对应分析表明, 影响鸡 Z 染色体基因表达的密码子偏性的主要因素分别为 GC3s、基因的表达丰度、GC 含量、CDS 长度及氨基酸的疏水性. 鸡 Z 染色体基因表达的密码子用法受到了核苷酸组成偏好的显著影响, 这种核苷酸组成偏好很可能是突变偏倚、固定偏倚及基因转换导致的. 对于鸡这种群体有效规模较大的群体, 密码子的偏性更有可能是核苷酸组成偏好及选择等因素综合作用的结果.

关键词:鸡; Z 染色体; 密码子; 偏性

中图分类号: Q953.3

文献标识码: A

文章编号: 1001-411X(2010)01-0070-05

Study on the Codon Bias of Genes Expression in GGAZ

RAO You-sheng¹, LIANG Fei-fei², WANG Zhang-feng¹, CHAI Xue-wen¹, ZHANG Xi-quan²

(1 Department of Biological Technology, Jiangxi Institute of Education, Nanchang 330029, China;

2 College of Animal Science, South China Agricultural University, Guangzhou 510642, China)

Abstract: The total cDNA sequences were downloaded from NCBI dataset, and 626 CDS sequences were finally included in analyses. By the use of CodonW (1.4.2), 26 codons (eg. CGG, AGC, UGC et al.) were identified as optimal codons. The corresponding analyses indicted that the main factors influencing codon usage of genes in GGAZ were GC3s, expression level, GC content, CDS length as well as Hydrophobicity of amino acids. As the codon usage of genes in GGAZ was mostly influenced by nucleotide composition, the bias of nucleotide composition is more likely due to the mutation bias, fix bias, and gene conversion. For a greater effective population size, the codon usage of chicken was most likely the combination of nucleotide composition and selection and so on.

Key words: chicken; GGAZ; codon; bias

突变和重组是生物进化的根本动力. 突变如果发生在基因的阅读框内, 则必然会导致密码子的突变. 密码子的突变分为 3 大类型: 同义突变 (Synonymous mutation)、非同义突变 (Nonsynonymous mutation)、无义突变 (Nonsense mutation). Li 等^[1]的研究表明, 如果基因表达过程中所有密码子以同一频率出现, 则同义、非同义和无义突变的比例约为 0.25:0.71:0.04. 非同义突变会导致编码的氨基酸发生改变, 从而引起蛋白质的结构和功能发生改变, 使突变个体发生

相对显著的表型变异. 同义突变是一种中性或近中性突变, 虽然不会引起所编码的氨基酸发生变化, 但这些密码子在蛋白质的翻译过程中使用的频率存在显著差异, 此即为密码子的偏性. 影响密码子偏性的因素主要有: 高表达的基因其密码子使用的频率与细胞内同功 tRNA 的丰度有关^[2]; 密码子的偏性同密码子偏倚的突变压有关^[3]; 密码子的偏性同整个基因组的 GC 含量特别是阅读框的 GC 含量有关^[4-6]. 密码子的偏性还与基因的长度、基因的功能、氨基酸

收稿日期: 2009-04-23

作者简介: 饶友生 (1965—), 男, 教授, 博士; 通讯作者: 张细权 (1963—), 男, 教授, 博士, E-mail: xqzhang@scau.edu.cn

基金项目: 江西省教育厅科学技术研究项目 (GJJ8469); 863 计划项目 (2006AA10A120)

的疏水性以及蛋白质的二级结构等有关^[7-9]. 密码子的偏性可能是中性进化的结果(如突变偏畸、基因转换等)也可能是对同义密码子的选择所致,密码子的用法反映的是2种进化力量之间的一种平衡^[10]. 在果蝇和线虫的研究中发现,高表达的基因存在显著的密码子用法偏畸,这种偏畸主要是通过对同义密码子的选择以提高翻译的效率和准确性而导致的^[11-13]. 研究者们还发现,在人的不同组织中,基因表达的密码子用法存在显著差异,组织特异性基因的密码子偏畸是通过对不同组织细胞中 tRNA 丰度的选择导致的^[14-15]. 鸡的核型包括8对大染色体,30对小染色体和1对性染色体($2n = 78$),性别决定方式为ZW型,雄性同配(ZZ). 其他物种的研究表明,在性染色体上存在许多性别偏畸基因(Sex-bias genes),这种偏畸可能对密码子的用法产生影响^[16-17]. 本研究对鸡Z染色体基因的密码子偏性模式进行了探讨,旨在为鸡常染色体基因的密码子用法及线粒体基因组的密码子用法提供参照,同时为鸡的转基因研究中外源基因(如GHR基因,该基因在育种实践中有非常重要的意义)的改造提供理论上的指导.

1 材料与方法

1.1 基因序列

鸡Z染色体上全部基因完整的cDNA序列从NCBI数据库(<http://www.ncbi.nlm.nih.gov/projects/mapview/map>)下载. 编写PERL程序提取相应的CDS序列. 所有下列情形之一的CDS不包括在数据分析中:(1)不以ATG为起始密码子;(2)碱基数全长不为3的倍数;(3)序列内部含有终止密码子;(4)CDS全长小于300个碱基. 包含多个剪接体的基因选择其最长的CDS序列. 最终共有626个基因的CDS序列纳入统计分析.

1.2 统计分析

使用CodonW(1.4.2)进行密码子偏性分析. 密码子偏性的度量指标包括:相对密码子使用度(Relative synonymous codon usage, RSCU)、有效密码子数(Effective number of codons, ENC)、密码子适应指数(Codon adaptation index, CAI)^[18]. RSCU定义为某一同义密码子使用次数的观察值与该密码子出现次数的期望值的比例. 密码子出现次数的期望值为该密码子所编码的氨基酸的所有同义密码子平均使用的次数. 如果密码子的使用无偏好性,则RSCU值为1;如果该密码子相对其他同义密码使用频繁,则RSCU值大于1.

有效密码子数(ENC)反映基因有效使用密码子种类的多少,也即反映了基因密码子使用的偏性程度. 其值一般从20(该基因只有20个密码子,编码所有20种氨基酸)到61(所有的同义密码子以均等频率使用). ENC越大,基因对密码子使用偏性越小,ENC越小,基因对密码子的使用偏性越大.

密码子适应指数(CAI)反映编码区同义密码子与密码子最佳使用相符合的程度,取值范围在0~1之间. 表达量较高的基因具有较高CAI值,表达量较低的基因具有较低的CAI值. 许多研究表明,CAI值最接近于基因表达水平的实际观测值,并已广泛应用于基因表达水平的预测^[19-20]. 本研究采用核糖体核蛋白基因作为计算CAI值的参考数据^[21].

统计全部CDS序列的碱基组成参数:G+C含量(鸟嘌呤和胞嘧啶含量);GC3s(同义密码子第3位的G+C频率);A3s、T3s、G3s、C3s(同义密码子在第3位上腺嘌呤、胸腺嘧啶、鸟嘌呤和胞嘧啶的出现频率). 通过构建这些参数和ENC之间的二维散点图,反映密码子使用偏性与基因碱基组成之间的关系^[22].

使用对应分析探究样本各变量之间的关系. 对应分析样本中所有基因按密码子的使用频率分布在一个59维(64个密码子去除3个终止密码子以及甲硫氨酸和色氨酸的密码子)的向量空间,通过矩阵数据转换,鉴定出影响密码子使用偏性的主要因素^[23-24]. 对应分析使用CodonW(1.4.2)的correspondence analyses程序(<http://codonw.sourceforge.net/>)完成. 各变量间的相关性分析使用SPSS 11.5完成.

2 结果与分析

2.1 Z染色体基因表达的“最优”密码子

计算所有基因序列的ENC,并对这些值进行排序,取该有序数据集的上下限区域各5%的序列数据,形成2个新的数据子集. 比较2个数据子集中密码子的RSCU值,如果差异大于0.3,且该密码子的RSCU值在高表达基因样本中大于1,在低表达基因样本中小于1,则该密码子定义为“最优”密码子^[25]. 最终确定了CGG、AGC、UGC等26个密码子为Z染色体基因表达的“最优”密码子. 由表1中可以看出“最优”密码子均为以GC结尾的密码子.

2.2 ENC和GC3s的关联分析

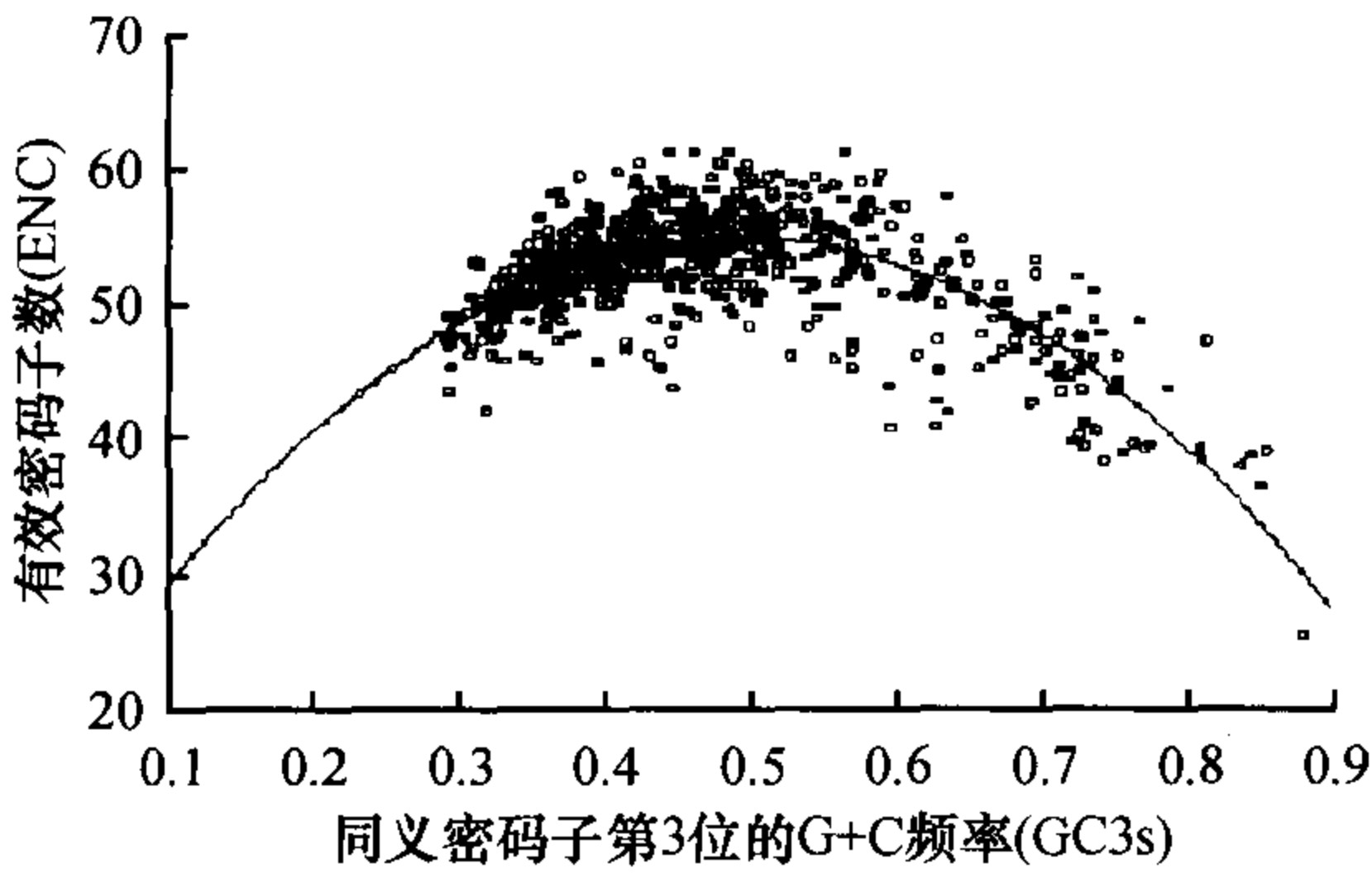
使用ENC与GC3s描绘散点图(Nc-plot). 图1的连续曲线反映了在无选择压力的情况下ENC和GC3s之间的关系. 从图1可知,大多数基因位点的分布偏离了曲线,说明除了核苷酸组成偏好外,自然

表1 鸡Z染色体基因表达的最优密码子¹⁾
Tab.1 Codon usage of gene's expression in GGAZ

氨基酸	密码子	高表达	低表达	氨基酸	密码子	高表达	低表达
Phe	UUU	0.48	1.37	Ser	UCU	0.51	1.75
	UUC	1.52	0.63		UCC	1.77	0.52
Leu	UUA	0.07	1.19		UCA	0.39	1.59
	UUG	0.38	1.22		UCG	0.75	0.11
	CUU	0.28	1.26	Pro	CCU	0.46	1.57
	CUC	1.83	0.42		CCC	1.75	0.41
	CUA	0.15	0.66		CCA	0.39	1.83
Ile	CUG	3.28	1.24		CCG	1.40	0.20
	AUU	0.51	1.37	Thr	ACU	0.45	1.54
	AUC	2.33	0.60		ACC	1.83	0.45
	AUA	0.16	1.04		ACA	0.66	1.82
Met	AUG	1.00	1.00		ACG	1.06	0.19
Val	GUU	0.26	1.41	Ala	GCU	0.56	1.46
	GUC	1.23	0.44		GCC	1.72	0.51
	GUA	0.22	1.06		GCA	0.45	1.92
	GUG	2.29	1.09		GCG	1.27	0.11
Tyr	UAU	0.41	1.28	Cys	UGU	0.31	1.31
	UAC	1.59	0.72		UGC	1.69	0.69
Ter	UAA	0.56	0.94	Ter	UGA	1.69	1.31
	UAG	0.75	0.75	Trp	UGG	1.00	1.00
His	CAU	0.25	1.43	Arg	CGU	0.48	0.56
	CAC	1.75	0.57		CGC	2.47	0.24
Gln	CAA	0.37	1.01		CGA	0.42	0.56
	CAG	1.63	0.99		CGG	1.86	0.30
Asn	AAU	0.28	1.33	Ser	AGU	0.34	1.34
	AAC	1.72	0.67		AGC	2.25	0.69
Lys	AAA	0.51	1.28	Arg	AGA	0.22	3.07
	AAG	1.49	0.72		AGG	0.55	1.27
Asp	GAU	0.43	1.42	Gly	GGU	0.31	1.15
	GAC	1.57	0.58		GGC	1.87	0.58
Glu	GAA	0.33	1.33		GGA	0.42	1.75
	GAG	1.67	0.67		GGG	1.40	0.51

1) 最优密码子用黑体标记

选择等其他因素对密码子的使用具有显著影响. 相关性分析表明, 基因表达水平(CAI 值) 与 ENC 呈极显著负相关($r = -0.555, P < 0.01$), 与 GC3s 呈极显著正相关($r = 0.978, P < 0.01$). 曲线下方的基因具有较高的 GC3s 含量, 趋向使用较少的密码子(ENC 偏低), 表达水平也相对较高; 曲线上方的基因倾向于随机使用密码子, 表达量相对较低. 从 CAI 和 GC 含量、GC3s 的相关性可以看出 Z 染色体上高表达的基因富集 GC, 尤其偏好以 G 或 C 结尾的密码子. ENC 与 GC 含量、GC3 之间的相关性也验证了上述结论($r = -0.436, P < 0.01$; $r = -0.446, P < 0.01$).



连续曲线表示密码子随机使用时 ENC 和 GC3s 之间的关系

图1 鸡Z染色体基因的散点图分布

Fig.1 Nc-plot of genes in GGAZ

2.3 密码子使用的对应分析

图2的对应分析发现,第1条向量轴能够解释基因间总变异的30.02%,其余3条向量轴只能解释总变异的8.46%、5.87%、4.56%。从对应分析中数据分布的贡献率明显看出,第1条向量轴是第2条向量轴的2倍以上,所以第1条向量轴是解释基因密码子偏好使用的首要参考。第1轴的特征值与偏性参数的相关分析表明,该轴上基因的位置与GC3s、GC含量、CAI值呈极显著正相关($r = 0.984, P < 0.01$; $r = 0.912, P < 0.01$; $r = 0.967, P < 0.01$),与CDS长度及氨基酸的疏水性呈极显著负相关($r = -0.305, P < 0.01$; $r = -0.138, P < 0.01$)。表明影响鸡Z染色体基因表达的密码子偏性的主要因素为GC3s、基因的表达丰度、GC含量、CDS长度及氨基酸的疏水性对密码子的使用偏性也有显著影响。

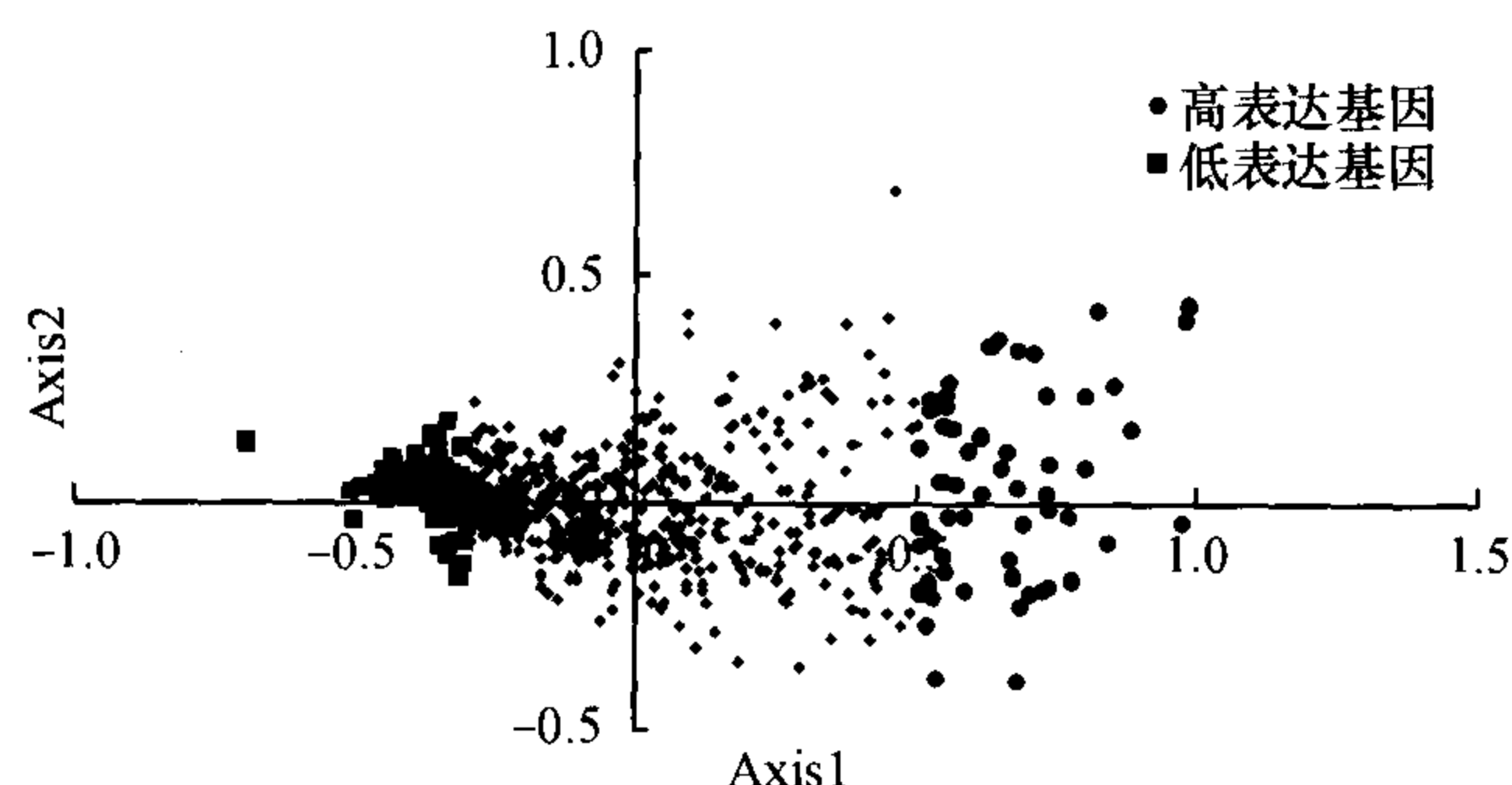


图2 Z染色体上的基因在2个主向量轴上的分布

Fig. 2 Distribution of genes in GGZ on the plane defined by the first two main axes of the correspondence analysis

3 讨论

使用NCBI数据库,笔者下载了鸡Z染色体上全部基因的CDS序列,探讨了鸡Z染色体基因表达的密码子使用模式,确定了26个“最优”密码子。研究发现,影响鸡Z染色体基因表达密码子偏性的主要因素是GC3s(同义密码子第3位的G+C频率)、基因的表达丰度、GC含量、CDS长度及氨基酸的疏水性。基因的碱基组成是影响密码子使用的一个主要因素,鸡Z染色体上基因的碱基组成存在很大的异质性,GC3s值的范围为0.26~0.91,GC含量和ENC的变化范围也分别为0.376~0.814和25~61。结合对应分析和相关分析的结果,第1向量轴基因的位置和GC含量呈极显著正相关($r = 0.912, P < 0.01$),可以推测Z染色体基因的密码子用法受到了核苷酸组成偏好的显著影响。笔者以为,这种核苷酸组成偏好很可能是突变偏倚、固定偏倚及基因转换导致的。根据群体遗传学的观点,对于有效规模较小的群体,密码子的偏性最有可能是核苷酸组成偏好

及遗传漂变的结果;但对于鸡这种群体规模较大的群体,密码子的偏性更有可能是核苷酸组成偏好及选择等因素综合作用的结果。基因表达水平(CAI值)与ENC的相关性($r = -0.555, P < 0.01$)以及GC含量与ENC的相关性验证了上述结论($r = -0.436, P < 0.01$; $r = -0.446, P < 0.01$)。

在果蝇和线虫的研究中发现,基因的偏性和CDS的长度及基因的长度呈极显著负相关,该现象被解释为在自然选择压力下缩短表达量高的基因对生物体本身有利,可以减少物质、能量和时间的消耗^[2]。相反地,对于大的基因,可能存在更多的更严格的调控因子,因而在自然选择过程中承受着更大的选择压^[4]。本研究发现,鸡Z染色体上的基因虽然表现出相同的趋势,但基因的偏性和基因的长度之间的相关性并未达到显著水平。有趣的是,研究样本中65个没有内含子的基因,其长度和基因的偏性(ENC)呈极显著正相关($r = 0.292, P = 0.0204$),印证了上述研究结果。这是否意味着基因中内含子的存在与否(Polymorphism of presence/absent of intron)以及内含子的数目和大小与基因的表达偏性相关联,抑或是性染色体上基因的密码子用法和常染色体有别?相关的研究我们还在进行中。

参考文献:

- [1] LI Wen-hsiung. Molecular evolution [M]. Sunderland MA: Sinauer Associates, 1997: 28-46.
- [2] MORIYAMA E N, POWELL J R. Codon usage bias and tRNA abundance in *Drosophila* [J]. J Mol Evol, 1996, 13: 261-277.
- [3] 刘庆坡, 薛庆中. 遗传密码子及其应用[J]. 中国生物化学与分子生物学报, 2006, 22(11): 851-855.
- [4] HOLMQUIST G P, FILIPSKE J. Organization of mutations along the genome: A prime determinant of genome evolution [J]. Trends Ecol Evol, 1994, 9: 65-69.
- [5] BERNARDI G. The human genome: Organization and evolutionary history [J]. Annu Rev Genet, 1995, 29: 445-476.
- [6] 石秀凡, 黄京飞, 柳树群, 等. 人类基因同义密码子偏好的特征以及与基因GC含量的关系[J]. 生物化学与生物物理进展, 2002, 29(3): 411-414.
- [7] KNIGHT R D, FREELAND S J, LANDWEBER L F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes [J]. Genome Biol, 2001, 2: RESEARCH0010.
- [8] MARIN A, GONZALEZ F, GUTIERREZ G, et al. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli* [J]. Nucleic Acids Res, 1998, 26(19): 4540.

- [9] GUPTA S K, MAJUMDAR S K, BHATTACHARYA T, et al. Studies on the relationships between the synonymous codon usage and protein secondary structural units [J]. Biochem Biophys Res Commun, 2000, 269(3): 692-696.
- [10] MARAIS G D, MOUCHIROUD D, DURET L. Neutral effect of recombination on base composition in *Drosophila* [J]. Genet Res, 2003, 81: 79-87.
- [11] COMERON J M, KREITMAN M. The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: Mutation, selection or relaxed constraints? [J]. Genetics, 1998, 150: 767-775.
- [12] DURET L. Evolution of synonymous codon usage in metazoans [J]. Curr Opin Genet Dev, 2002, 12: 640-669.
- [13] AKASHI H. Translational selection and yeast proteome evolution [J]. Genetics, 2003, 164: 1291-1303.
- [14] PLOTKIN J B, ROBINS H, LEVINE A J. Tissue-specific codon usage and the expression of human genes [J]. Proc Natl Acad Sci, 2004, 101: 12588-12591.
- [15] SE' MON M, LOBRY J R, DURET L. No evidence for tissue-specific adaptation of synonymous codon usage in humans [J]. Mol Biol Evol, 2006, 23(3): 523-529.
- [16] PARISI M, NUTTALL R, EDWARDS P, et al. A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults [J]. Genome Biol, 2004, 5: R40.
- [17] RANZ J M, CASTILLO-DAVIS C I, MEIKLEJOHN C D, et al. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome [J]. Science, 2003, 300: 1742-1745.
- [18] 吴宪明, 吴松锋, 任大明, 等. 密码子偏性的分析方法及相关研究进展[J]. 遗传, 2007, 29(4): 420-426.
- [19] GUPTA S K, BHATTACHARYA T K, GHOSH T C. Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection [J]. J Biomol Struct Dyn, 2004, 21: 1-9.
- [20] PEIXOTO L, ZAVALA A, ROMERO H, et al. The strength of translational selection for codon usage varies in three replicons of *Sinorhizobium meliloti* [J]. Gene, 2003, 320: 109-116.
- [21] HOU Z C, YANG N. Factors affecting codon usage in *Yersinia pestis* [J]. Acta Biochimica et Biophysica Sinica, 2003, 25: 580-586.
- [22] WRIGHT F. The "effective number of codon" used in a gene [J]. Gene, 1990, 87: 23-29.
- [23] SINGER G A C, HICKEY D A. Thermophilic prokaryotes have characteristic pattern of codon usage, amino acid composition and nucleotide content [J]. Gene, 2003, 317: 39-47.
- [24] ROMERO H, ZAVALA A, MUSTO H, et al. The influence of translational selection on codon usage in fishes from the family Cyprinidae [J]. Gene, 2003, 317: 141-147.
- [25] DURET L, MOUCHIROUD D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis* [J]. Proc Natl Acad Sci, 1999, 96: 4482-4487.

【责任编辑 柴 焰】

(上接第69页)

- [12] 张清峰, 许尚忠, 李俊雅, 等. 鲁西黄牛肉用品系育种目标性状和选择性状研究[J]. 西北农林科技大学学报: 自然科学版, 2007, 35(2): 33-37.
- [13] KOOTS K R, GIBSON J P, SMITH C, et al. Analyses of published genetic parameter estimates for beef production traits: 1: Heritability [J]. Animal Breeding Abstract, 1994, 62: 309-338.
- [14] KOOTS K R, GIBSON J P, SMITH C, et al. Analyses of published genetic parameter estimates for beef production traits: 2: Phenotypic and genetic correlations [J]. Animal Breeding Abstract, 1994, 62: 825-853.
- [15] KEALEY C G, MACNEIL M D, TESS M W, et al. Estimation of genetic parameters of yearling scrotal circumference and semen characteristics in line 1 hereford bulls [J]. Animal Science, 2006, 84: 283-290.
- [16] GREGORY K E, CUNDIFF L V, KOCH R M. Genetic and phenotypic (co) variances for growth and carcass traits of purebred and composite populations of beef cattle [J]. Journal of Animal Science, 1995, 73: 1920-1926.
- [17] RILEY A C, CHASE C C, Jr, HAMMOND A C, et al. Estimated genetic parameters for carcass traits of Brahman cattle [J]. Journal of Animal Science, 2002, 80: 955-962.
- [18] LÔBO R N B, MADALENA F E, PENNA V M. Evaluation of alternative breeding programs for dual purpose zebu cattle [J]. Rev Bras Zootec, 2000, 29(5): 1361-1370.
- [19] 李俊雅. 中国西门塔尔牛核心群优化育种规划的研究[D]. 北京: 中国农业大学动物科学院, 2002.
- [20] 张沅. 家畜育种规划[M]. 北京: 中国农业大学出版社, 2000: 96-179.
- [21] DUBOIS C, MANFREDI E, RICARD A. Optimization of breeding schemes for sport horses [J]. Livestock Science, 2008, 18: 99-112.
- [22] GICHEHA M G, KOSGEY I S, BEBE B O, et al. Evaluation of the efficiency of alternative two-tier nucleus breeding systems designed to improve meat sheep in Kenya [J]. Journal of Animal Breeding and Genetics, 2006, 123(4): 247-257.

【责任编辑 柴 焰】