中国科学 C 辑:生命科学 2009 年 第 39 卷 第 10 期: 948 ~ 953
www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

论 文

*Wolbachia pipientis w*Mel 基因组水平上的密码子使用分析

孙铮[①⑥], 马亮[②], MURPHY Robert[③④], 张宪省[⑤], 黄大卫[①②*]

① 山东农业大学植物保护学院, 泰安 271018;
② 中国科学院动物研究所, 北京 100080;
③ Department of Natural History, Royal Ontario Museum, 100 Queen's Park, Toronto, ON Canada M5S 2C6;
④ 中国科学院昆明动物研究所遗传与进化国家重点实验室, 昆明 650223;
⑤ 山东农业大学生命科学学院, 泰安 271018;
⑥ 黄岛出入境检验检疫局, 青岛 266555
* 联系人, E-mail: huangdw@ioz.ac.cn

收稿日期: 2009-07-15; 接受日期: 2009-08-31
国家自然科学基金(批准号: 30570970 和 30770302)、科学技术部科技基础性工作专项重点项目(批准号: 2006FY110500)和国家基础科学人才培养基金(批准号: NSFC-J0630964/J0109)资助

摘要 本实验检测了黑腹果蝇的专性寄生菌 *Wolbachia pipientis w*Mel 基因组的密码子使用模式, 并推测影响其密码子组成的因素. 选择了 478 条蛋白编码基因作为研究对象, 它们的 GC 含量比较低, 约 0.282~0.432. 本研究结果显示, 关联突变(context-dependent mutation)是影响 *W. pipientis w*Mel 基因组密码子组成的主要因素. 同时, Nc-Plot 曲线显示, 基因组的密码子组成还受到了核苷酸组成偏好性(nucleotide composition bias)的影响. 对应分析的结果显示, 基因长度(R=0.123, P<0.01)和基因表达水平(R=–0.312, P<0.01)也对基因组的密码子使用偏好性起到了一定的作用. 因此, 关联突变、核苷酸组成、基因长度和基因表达水平都会影响到基因组的密码子使用偏好性. *Wolbachia* 基因组的几个特征和动物线粒体基因组具有较高的相似性, 而且现在已有报道说明它们有共同的起源, 但仍需进一步验证.

关键词
*Wolbachia pipientis w*Mel
密码子使用
对应分析
关联突变

随着基因组时代的到来, 出现了大量的基因数据, 这使得在分子水平上进行基因组进化和密码子使用的研究成为可能. 密码子使用之所以具有偏好性是为了使出现频率最高的密码子和基因组中丰度最高的tRNA分子上的反密码子相匹配以保证蛋白合成的准确度和效率[1,2]. 密码子使用偏好性在不同的生物体中甚至在同一个生物体基因组的不同基因中都具有显著的差异[3]. 影响密码子使用模式的因素有多种, 如 tRNA 分子的丰度[4,5]、翻译选择[6,7]、突变偏好性[2,8]、基因在染色体上的位置[9]等. 除此之外, 前后相互关联的碱基突变也是影响密码子组成的重要因素[10~12].

Wolbachia 是一种革兰氏阴性α蛋白菌, 该属主要感染节肢动物(*Arthropods*)和丝虫(*Filarial nematodes*)[13]. 对 *Wolbachia* 基因组密码子使用模式的研究已经越来越多. *Wolbachia* 可以影响寄主的生殖, 例

引用格式: 孙铮, 马亮, MURPHY Robert, 等. *Wolbachia pipientis w*Mel 基因组水平上的密码子使用分析. 中国科学 C 辑: 生命科学, 2009, 39: 948—953

如胞质不相容(cytoplasmic incompatibility)、孤雌生殖(thelytokous parthenogenesis)、雄性雌性化(ferminization of genetic males)及杀雄作用(male killing)[14,15]. 人们希望在 *Wolbachia* 的基因和寄主基因之间找到一定的联系[16].

Wolbachia pipientis *w*Mel 是黑腹果蝇(*Drosophila melanogaster*)的专性寄生菌, 目前它的基因组序列已经被测定[14]. 研究表明, 由于受到重复种群瓶颈(repeated population bottlenecks)的影响, 自然选择对于该寄生菌的基因组组成影响不大. 本文将讨论 *W. pipientis* *w*Mel 基因组的密码子使用模式及其受到的主要影响因素.

1 材料与方法

1.1 数据获取

W. pipientis *w*Mel 的全基因组序列从 TIGR (http://cmr.tigr.org/tigr-scripts/CMR/GenomePage.cgi?database=dmg)下载. 基于数据库中的注释, 基因组中的蛋白编码基因通过批下载(batch download)获得. 为了计算更加精确, 删除了少于 300 bp 的基因以及核苷酸序列不能翻译成蛋白序列的基因, 共有 478 条蛋白基因被用来分析.

1.2 同义密码子偏好性检测

相对同义密码子使用(relative synonymous codon usage, RSCU)用来检测基因中全部密码子使用的变化, 它的值等于同义密码子的实际观测值与同义密码子平均使用时期望值的比值[17,18]. 对于同义密码子家族中的密码子来说, 如果这个同义密码子的 RSCU 值大于 1, 则表明该密码子的使用频率高于期望值, 反之亦然[19].

密码子适应指数(codon adaption index, CAI)可以根据已知高表达基因的序列来估计未知基因密码子使用的偏好性程度[20]. CAI 的值在 0~1 之间, 如果越高则表明该基因的密码子使用偏好性越强, 且表达水平也越高[18].

有效密码子数目(effective number of codons, ENC or Nc)用于检测单个基因密码子使用偏好性的高低, 它的变化范围在 20~61 之间. 假如一个基因的密码子使用偏好性程度达到最强的话, 也就是说每种氨基酸只使用一种密码子, 那么此时 ENC 的值为 20; 若基因的密码子使用偏好性最弱, 即每个同义密码子使用的频率几乎是随机的, 则 ENC 的值为 61[18,21].

W. pipientis *w*Mel 基因组的碱基组成偏好性主要通过 GC3s 来进行检测. GC3s 是指除 Met, Trp 和终止密码子之外的同义密码子第 3 位上的 GC 含量[18]. 为了检测密码子使用偏好性是否受到基因组碱基组成的影响, 本实验绘制了 *Nc*-plot 曲线. 若一个基因的密码子使用模式受到 GC 碱基组成影响的话, 那么这个基因将落在 *Nc*-plot 期望曲线的上面或较近的位置, 否则该基因将落在离期望曲线比较远的位置[21]. ENC 期望值可以通过以下公式计算: ENC 期望值 $= 2 + s + \{29 \div [s^2 + (1-s)^2]\}$, 其中 s 代表 GC3s[21,22].

为了研究基因长度、基因表达水平和密码子偏好性之间的相互关系, 本实验把这些基因根据基因长度分成 5 组, 将进一步检测 *W. pipientis* *w*Mel 基因组中基因长度对每组基因的表达水平、*ENC* 和 GC3s 的影响.

1.3 对应分析

对应分析(correspondence analysis, COA)主要是检测基因 RSCU 值的变化并推测影响基因密码子组成的主要因素[7,23]. 所有的基因根据它们所包含的 59 个有义密码子的使用情况被分布在一个 59 维空间中. 通过检测相对向量和基因在主向量轴上的位置可以确定一组数据的主要变化趋势[18].

1.4 关联突变的统计检验

本实验主要集中在 8 个四重密码子家族(4-fold degenerate families, FFD)(表 1). 可以进行一个简单的假设, 即单个位点之间的突变是相互独立的并且在密码子第 3 位上没有选择压力.

让 n(XYZ)作为密码子 XYZ 在基因中出现的次数, n(YZ)作为在密码子的 2 和 3 位置上出现的 YZ 碱基对的总次数, 这里仅仅计算四重密码子家族的第三位. 因此, n(UZ)=n(CUZ)+n(GUZ); n(CZ)=n(UCZ)+n(CCZ)+n(ACZ)+n(GCZ); n(GZ)=n(CGZ)+n(GGZ) n(YZ)可以形成一个 3×4 的联表(contingency table). 可以定义这样一个无效假说(null hypothesis): Y 和 Z 是独立的, 则可以通过卡方检验来验证这个假说是

否成立, 它的自由度为 6[12].

使用 DAMBE 5.0.52 软件计算各个基因中密码子的 RSCU 值[24,25], Java Codon Adaptation Tool(http://www.jcat.de/)根据 *W. pipientis* *w*Mel 基因组中的参照基因计算各个基因的 CAI 值[26]. CodonW 1.4(http://codonw.sourceforge.net/)被用来对基因中密码子的 RSCU 值执行对应分析. Spearman 相关度分析和标准卡方检验使用多元统计软件 SPSS 16.0 完成[18,27].

2 结果

2.1 基因组和基因组成分析

W.pipientis *w*Mel 的基因组具有比较高的碱基偏好性. 它所包含基因的 GC 含量变化在 0.282~0.432 之间, GC3s 的变化范围在 0.157~0.361 之间. 结果表明基因组的密码子组成可能受到组成限制性的影响并且与核苷酸碱基组成直接相关. 仅有一部分基因落在期望曲线的上面或附近(图 1), 大部分基因还是落在了离期望曲线比较远的位置表明还有其他因素影响密码子组成.

表 1 *Wolbachia pipientis* *w*Mel 基因组中 8 个四重密码子的密码子使用情况

氨基酸	密码子	数目	氨基酸	密码子	数目
A	GCU	4031	L	CUA	2426
A	GCG	1077	L	CUC	1035
A	GCC	818	L	CUG	1454
A	GCA	5452	L	CUU	3438
G	GGU	3920	P	CCA	2556
G	GGG	1271	P	CCC	368
G	GGC	1887	P	CCU	2042
G	GGA	3573	P	CCG	570
R	CGA	399	S	UCA	3038
R	CGC	637	S	UCC	840
R	CGG	193	S	UCG	724
R	CGU	1218	S	UCU	3271
T	ACC	957	V	GUU	4621
T	ACA	3044	V	GUG	2390
T	ACG	797	V	GUC	807
T	ACU	3339	V	GUA	3827

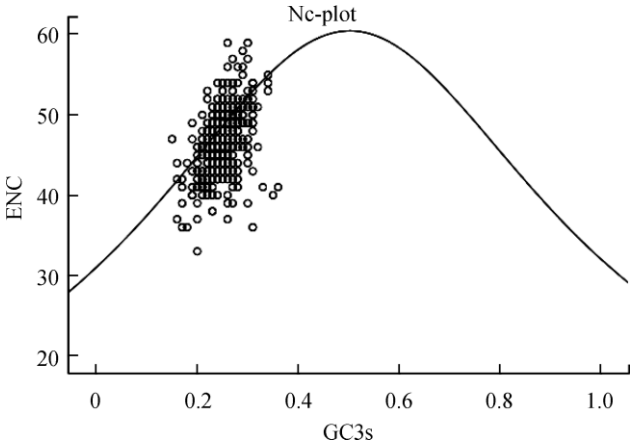

图 1 *Wolbachia pipientis* *w*Mel 的 Nc-plot 曲线
连续的曲线表示当密码子随机使用时的期望曲线

2.2 基因长度、基因表达水平和密码子偏好性之间的相互关系

基因长度的分析结果表明越短的基因具有越低的 Nc 值, 同时具有越高的密码子使用偏好性(表 2). 并且, 越短的基因具有越低的 GC3s 含量和越高的 CAI 值, 即具有较高的基因表达水平. 由于对应分析的第一个轴和基因长度成正相关(R=0.123, P<0.01), 可以推测基因长度也是影响其密码子组成的一个因素.

2.3 对应分析的结果

对每个基因密码子的 RSCU 值进行对应分析以

表 2 不同长度基因的 Nc, GC3s 和 CAI 的比较

组号	基因长度/bp	基因数目	$\bar{x}$ ± SD*		
			Nc	GC3s	CAI
1	≥3000	9	48.800±1.861	0.274±0.029	0.476±0.034
2	2000 ≤ and < 2999	32	47.371±3.278	0.260±0.036	0.502±0.049
3	1000 ≤ and < 1999	190	47.203±3.266	0.256±0.031	0.508±0.040
4	500 ≤ and < 999	190	46.873±3.967	0.251±0.033	0.514±0.044
5	< 500	57	45.703±5.377	0.242±0.039	0.536±0.060

最小化氨基酸组成的影响. 由对应分析产生的第一个轴代表总向量的 6.59%, 第二个轴代表总向量的 5.92%. 由于第一个轴只能解释密码子使用的部分变化, 意味着影响该细菌基因组密码子组成的因素不止一个. 对应分析的第一个轴与 ENC(R=0.270, P<0.01), GC(R=0.260, P<0.01)和基因长度成正相关(R=0.123, P<0.01), 但和 CAI 值成微弱的负相关(R=–0.312, P<0.01). 其中, CAI 与 ENC(R=–0.584, P<0.01), GC3s(R=–0.779, P<0.01)成负相关关系, 表明密码子使用也可能会受到基因表达水平的影响.

高表达基因和低表达基因具有相似的密码子使用模式. 对位于 axis1 上两端基因(5%)的 RSCU 值进行标准卡方检验(表 3), 表明有 11 个密码子的使用频率在高表达基因中明显增多, 它们编码 11 种氨基酸.

表 3 ***Wolbachia pipientis w*Mel 基因组中的最优密码子**

氨基酸	密码子 a)	高表达基因		低表达基因		氨基酸	密码子 a)	高表达基因		低表达基因	
		数目	RSCU	数目	RSCU			数目	RSCU	数目	RSCU
Phe	UUU	211	1.60	221	1.66	Ser	UCU*	105	1.79	101	1.42
	UUC	52	0.40	45	0.34		UCC	20	0.34	35	0.49
Leu	UUA*	183	2.38	169	1.95		UCA	75	1.28	104	1.46
	UUG	88	1.15	85	0.98		UCG*	29	0.50	14	0.20
	CUU	71	0.92	104	1.20	Pro	CCU*	62	2.05	46	1.13
	CUC	17	0.92	41	0.47		CCC	9	0.30	9	0.22
	CUA	80	1.04	73	0.84		CCA	43	1.42	87	2.13
	CUG	53	0.52	81	0.44		CCG	7	0.23	21	0.52
Ile	AUU	171	1.13	205	1.24	Thr	ACU	81	1.59	96	1.54
	AUC	58	0.38	78	0.47		ACC	24	0.47	24	0.39
	AUA*	225	1.49	213	1.29		ACA	82	1.61	106	1.70
Met	AUG	110	1.00	160	1.00		ACG	17	0.33	23	0.37
Val	GUU	142	1.73	144	1.56	Ala	GCU	106	1.68	127	1.52
	GUC	17	0.21	28	0.30		GCC	10	0.16	28	0.33
	GUA*	131	1.59	113	1.22		GCA	124	1.97	147	1.76
	GUG	39	0.47	84	0.91		GCG	12	0.19	33	0.39
Tyr	UAU	121	1.48	130	1.37	Cys	UGU*	34	1.19	31	0.78
	UAC	43	0.52	60	0.63		UGC	23	0.81	49	1.23
TER	UAA	11	1.43	13	1.70	TER	UGA	8	1.04	3	0.39
	UAG	4	0.52	7	0.91	Trp	UGG	43	1.00	57	1.00
His	CAU	43	1.23	74	1.38	Arg	CGU	8	0.37	57	1.98
	CAC	27	0.77	33	0.62		CGC	5	0.23	32	1.11
Gln	CAA	76	1.42	112	1.48		CGA	1	0.05	14	0.49
	CAG	31	0.58	39	0.52		CGG	0	0.00	10	0.35
Asn	AAU*	182	1.56	219	1.38	Ser	AGU	77	1.32	103	1.45
	AAC	51	0.44	99	0.62		AGC	45	0.77	70	0.98
Lys	AAA	194	1.30	326	1.47	Arg	AGA	53	2.45	53	1.84
	AAG*	105	0.70	118	0.53		AGG*	63	2.91	7	0.24
Asp	GAU	134	1.49	211	1.60	Gly	GGU	73	1.40	139	1.55
	GAC	46	0.51	52	0.40		GGC	31	0.60	68	0.76
Glu	GAA	145	1.39	232	1.36		GGA	75	1.44	126	1.40
	GAG	63	0.61	110	0.64		GGG*	29	0.56	26	0.29

a) * 示最优密码子

本实验推测这些密码子是 *W. pipientis* *w*Mel 基因组中的最优密码子.

2.4 关联突变

密码子第 3 位上的碱基频率并不是独立于第 2 位的, 由表 1 得到的卡方检验结果为 6688.00(P<0.01). 因此, 前后相互关联的碱基突变也可以影响到 *W.pipientis* *w*Mel 基因组的密码子组成.

3 讨论

本实验结果表明, 关联突变可能是影响 *W. pipientis* *w*Mel 基因组密码子组成的主要因素. 然而, 尽管统计结果是显著的, 但各个参数之间的相关系数却较低. 也许这些因素可以影响到基因组的密码子组成, 但却不是主要影响因素.

基因组具有较高的 AT 碱基含量. 不论是高表达基因还是低表达基因的密码子第 3 位都倾向于选择 A 或 T. Nc-plot 曲线表明, 核苷酸组成会影响到基因组的密码子组成. 对应分析的结果表明, 基因的密码子使用还可能受到基因长度和基因表达水平的影响.

作为革兰氏阴性 alpha-蛋白菌, *Wolbachia* 具有较高的 AT 碱基含量, 并且是母系遗传. 动物线粒体基因的祖先同样被认为是一种 alpha-蛋白菌[28], 并且动物线粒体基因也是母系遗传. 对于动物线粒体基因来说, 尤其是昆虫的线粒体基因, 同样具有较高的 AT 碱基含量. 众所周知, *Wolbachia* 在昆虫体内的寄生是最广泛的. 对于两种基因组来说, 尽管有其他因素影响其密码子组成, 但核苷酸组成都是影响因素之一. 这些相似的特征可能是由于 *Wolbachia* 和动物线粒体基因具有密切的进化关系所导致的.

致谢 感谢山东农业大学动物科技学院王宁新老师提出的修改意见.

参考文献

1 Moriyama E N, Powell J R. Codon usage bias and tRNA abundance in *Drosophila*. J Mol Evol, 1997, 45: 514—523
2 Xia X. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene, 2005, 345: 13—20
3 Lavner Y, Kotlar D. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene, 2005, 345: 127—138
4 Bulmer M. Coevolution of codon usage and transfer RNA abundance. Nature, 1987, 325: 728—730
5 Olejniczak M, Uhlenbeck O C. tRNA residues that have coevolved with their anticodon to ensure uniform and accurate codon recognition. Biochinmie, 2006, 88: 943—950
6 Musto H, Cruveiller S, D'Onofrio G, et al. Translational selection on codon usage in *Xenopus laevis*. Mol Biol Evol, 2001, 18: 1703—1707
7 Naya H, Romero H, Carels N, et al. Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*. FEBS Lett, 2001, 501: 127—130
8 Chen S L, Lee W, Hottes A K, et al. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA, 2004, 101: 3480—3485
9 Romero H, Zavala A, Musto H, et al. The influence of translational selection on codon usage in fishes from the family Cyprinidae. Gene, 2003, 317: 141—147
10 Fedorov A, Saxonov S, Gilbert W. Regularities of context-dependent codon bias in eukaryotic genes. Nucl Acids Res, 2002, 30: 1192—1197
11 Morton B R. The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. J Mol Evol, 2003, 56: 616—629
12 Jia W, Higgs P G. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. Mol Biol Evol, 2008, 25: 339—351
13 Pfarr K M, Hoerauf A. A niche for *Wolbachia*. Trends Parasitol, 2007, 23: 5—7
14 Wu M, Sun L V, Vamathevan J, et al. Phylogenomics of the reproductive parasite *Wolbachia pipientis* *w*Mel: a streamlined genome

overrun by mobile genetic elements. PLoS Biol, 2004, 2: e69
15 Braquart-Varnier C, Greve P, Felix C, et al. Bacteriophage WO in *Wolbachia* infecting terrestrial isopods. BBRC, 2005, 337: 580—585
16 Rand D M, Haney R A, Fry A J. Cytonuclear coevolution: the genomics of cooperation. Trends Ecol Evol, 2004, 19: 645—653
17 Sharp P M, Tuohy T M F, Mosurski K R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res, 1986, 14: 5125—5143
18 Liu Q, Feng Y, Xue Q. Analysis of factors shaping codon usage in the mitochondrion genome of *Oryza sativa*. Mitochondrion, 2004, 4: 313—320
19 Sau K, Gupta S K, Sau S, et al. Factors influencing synonymous codon and amino acid usage biases in Mimivirus. Biosystems, 2006, 85: 107—113
20 Sharp P M, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res, 1987, 15: 1281—1295
21 Wright F. The `effective number of codons' used in a gene. Gene, 1990, 87: 23—29
22 Novembre J A. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol, 2002, 19: 1390—1394
23 Romero H, Zavala A, Musto H. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res, 2000, 28: 2084—2090
24 Xia X, Xie Z. DAMBE: software package for data analysis in molecular biology and evolution. J Heredity, 2001, 92: 371—373
25 Xia X. An improved implementation of codon adaptation index. Evolutionary Bioinformatics, 2007, 3: 53—58
26 Grote A, Hiller K, Scheer M, et al. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res, 2005, 33: W526—531
27 Liu Q, Dou S, Ji Z, et al. Synonymous codon usage and gene function are strongly related in *Oryza sativa*. Biosystems, 2005, 80: 123—131
28 Burger G, Gray M W, Franz Lang B. Mitochondrial genomes: anything goes. Trends Genet, 2003, 19: 709—716