张玉波,周正湘,吴小玉,等. 基于转录组的大头金蝇密码子的偏好性分析[J]. 江苏农业科学,2019,47(11):78 - 81.
doi:10.15889/j.issn.1002 - 1302.2019.11.016

基于转录组的大头金蝇密码子的偏好性分析

张玉波[1,2], 周正湘[1,2], 吴小玉[1], 张　斌[3]
(1. 安顺学院农学院,贵州安顺 561000; 2. 贵州省昆虫信息系统与资源开发利用重点实验,贵州安顺 561000;
3. 内蒙古师范大学生命科学与技术学院,内蒙古呼和浩特 010022)

摘要:使用 Codon W 软件分析大头金蝇[*Chrysomya megacephala*(Fabricius,1794)]转录组 10 923 条全长转录序列的密码子偏好性。结果表明,大头金蝇转录组中的 AT 含量(61.81%)远大于 GC 含量(38.19%);PR2(parity rule 2,即密码子偏好性)绘图结果显示,密码子第 3 位碱基 A 的使用频率大于碱基 T,碱基 G 的使用频率大于碱基 C;中性绘图结果显示,该序列密码子的使用更多地受到选择压力的影响;有效密码子数(effective number of codons,简称 ENC)在 25.17 ~ 61.00 个之间,均值为 43.16 个;密码子适应指数(codon adaptation index,简称 CAI)在 0.099 ~ 0.554 之间,均值为 0.215 8。结果共筛选出 29 个同义密码子相对使用度(relative synonymous codon usage,简称 RSCU) >1 的密码子和 28 个最优密码子。

关键词:大头金蝇;转录组;密码子偏好性;同义密码子
中图分类号: Q969.451.9;S186　　**文献标志码**: A　　**文章编号**:1002 - 1302(2019)11 - 0078 - 04

转录组测序(RNA sequencing)是指利用第二代高通量测序技术进行的 cDNA 测序,是一类专注于功能位点的测序策略,能全面快速地获取研究材料的特定组织在某一状态下的全部转录本信息[1]。随着高通量测序技术的应用[2],转录组测序以其较高的性价比而广受各位学者欢迎,被广泛应用于动植物的基因挖掘、功能鉴定等方面的研究,成为当前生物学研究的热点[3]。密码子偏好性指在编码氨基酸合成蛋白时,往往优先使用某一种或几种密码子[4],被优先选用的密码子称为最优密码子,这一现象广泛存在于生物类群中[5]。密码子偏好性具有物种特异性,不同基因组在进化过程中承受不同的选择压力,因此不同物种间密码子的使用偏好性不同[6-7]。分析密码子的偏好性可以深入了解编码序列的碱基含量、二核苷酸偏向性和隐藏的剪接信号等基因序列特征,这些都与密码子使用偏好性相关,都可以影响基因合成的设计与蛋白表达[7]。

大头金蝇[*Chrysomya megacephala* (Fabricius,1794)]为重要的卫生昆虫,隶属于丽蝇科(Calliphoridae)金蝇属(*Chrysomya*)[8]。研究大头金蝇转录组密码子偏好性,可以揭示氨基酸翻译过程中高表达与低表达基因对密码子的偏好选择,有助于解释其特殊生理效应的遗传机制,进一步为相关基因的克隆与表达奠定基础。

1　材料与方法

1.1　序列的获取

本研究数据来源于美国国立生物技术信息中心(National Center for Biotechnology Information,简称 NCBI)网站,序列号为 SRP050024,利用 Codon W 1.4.2 软件分析大头金蝇转录组 10 923 条序列的密码子偏好性。大头金蝇转录组测序数据见表 1。

表 1　大头金蝇转录组的测序数据

NCBI 登录号	样品类型	测序文库	测序数据量(Gb)	测序平台
SRR1663113	卵、幼虫、成虫	双末端(pair - end)	4.3	Illumina HiSeq 2000
SRR1663114	卵、幼虫、成虫	双末端	4.4	Illumina HiSeq 2000
SRR1660427	卵、幼虫、成虫	双末端	4.0	Illumina HiSeq 2000

1.2　数据分析

1.2.1　碱基含量及 PR2(parity rule 2,即密码子偏好性)的绘图分析　利用 Codon W(version 1.4,http://bioweb.pasteur.fr/seqanal/ interfaces/codonw.html)软件对大头金蝇的转录组基因序列进行分析,计算每条序列的密码子总 GC 含量、第 3 位密码子的 GC 含量(GC_3)、有效密码子数(effective number of codons,简称 ENC 或 Nc)及密码子适应指数(codon adaptation index,简称 CAI)。分别统计密码子 3 个位置上的 GC 含量,密码子第 1 位的 GC 含量表示为 GC_1,第 2、3 位的 GC 含量依次表示为 GC_2、GC_3。GC_1、GC_2 的平均值记为 GC_{12}。以 GC_{12} 为纵坐标、GC_3 为横坐标进行中性绘图(neutrality plot),分析密码子第 1、2 位与第 3 位碱基组成的相关性,研究密码子偏性的影响因素。选择丝氨酸(TCA、TCC、TCG、TCT)、亮氨酸(CTA、CTC、CTG、CTT)、脯氨酸、精氨酸(CGA、CGC、CGG、CGT)、苏氨酸、缬氨酸、丙氨酸和甘氨

收稿日期:2018 - 05 - 23
基金项目:贵州省教育厅项目(编号:黔教合 KY 字[2014]271、黔教合人才团队字[2015]71);贵州省科技厅项目(编号:黔科合 LH 字[2014]7503);内蒙古高等学校科研重点项目(编号:NJZZ17041)。
作者简介:张玉波(1978—),男,山东济宁人,博士,副教授,主要从事昆虫系统分类研究。E - mail:38615157@qq.com。

酸,计算每个基因的 $A_3/(A_3+T_3)$ 和 $G_3/(G_3+C_3)$,分析各基因密码子中 4 个碱基组分嘌呤(A 和 G)与嘧啶(T 和 C)之间的关系。

1.2.2　ENC 及中性绘图分析　有效密码子数 ENC 用于检测单个基因密码子的使用偏好程度,取值范围在 20 ~ 61 之间,其值越低,表明该基因的密码子使用偏好性越强[9]。以密码子第 3 位上同义密码子 GC 的含量 GC_{3s} 为横坐标、ENC 为纵坐标,作二维散点图,探讨各基因密码子的使用偏性情况,并检测碱基组成对密码子偏性的影响。

1.2.3　同义密码子相对使用度及最优密码子分析　参照 Sharp 等的方法[10],同义密码子相对使用度(relative synonymous codon usage,简称 RSCU)是对同义密码子使用偏好的评估[10],该值等于同义密码子的实际观测值与同义密码子平均使用期望值的比值。如果密码子的使用无偏好性,则 RSCU 值为 1;如果该密码子比其他同义密码子的使用更频繁,则其 RSCU 值大于 1,反之,RSCU 值小于 1。

利用高表达优越密码子分析方法[11],统计所有基因的 ENC 值、有序数据集上下 10% 区间内形成的高 RSCU 集合和低 RSCU 集合,进行最优密码子分析。根据 2 个子集的 ΔRSCU 值及卡方检验结果确定最优密码子。

2　结果与分析

2.1　碱基含量及 PR2、中性绘图分析

对已经得到的大头金蝇转录组数据进行筛选,共获得长度为 300 bp 以上的 10 923 条完全阅读框序列(全长 CDS)。用 Codon W 软件进行密码子偏好性分析,结果表明,大头金蝇转录组序列中平均总 GC 量为 38.19%,分布范围为 24.40% ~ 62.90%,其中第 3 位点 GC_{3s} 的平均值为 25.67%,范围为 10.20% ~ 83.70%;总 A、T、C、G 4 种碱基含量分别为 32.9%、28.9%、18.4%、19.7%,密码子第 3 位点 T_{3s}、C_{3s}、A_{3s} 和 G_{3s} 含量的平均值分别为 50.82%、17.97%、43.10% 和 15.65%;GC_{12} 含量的均值为 44.14%(23.40% ~ 79.40%)(表 2)。可以看出,在大头金蝇转录组序列中的 AT 碱基含量远高于 GC。由图 1 可以看出,经 PR2 分析,大头金蝇转录组序列密码子第 3 位点碱基使用不均衡,密码子第 3 位碱基 A 的使用频率小于碱基 T,碱基 G 的使用频率小于碱基 C,表明大头金蝇转录组序列中基因密码子的使用模式受到突变压力和自然选择等多重因素的影响。

表 2　转录组密码子不同位置的碱基含量及 ENC、CAI

类别	T_{3s} 含量(%)	C_{3s} 含量(%)	A_{3s} 含量(%)	G_{3s} 含量(%)
平均值	50.82	17.97	43.10	15.65
范围	8.02 ~ 76.95	3.49 ~ 78.07	5.26 ~ 69.23	1.13 ~ 74.85
类别	**CAI**	**GC 含量(%)**	**GC_{12} 含量(%)**	**ENC(个)**	**GC_{3s} 含量(%)**
平均值	0.215 8	38.19	44.14	43.16	25.67
范围	0.099 0 ~ 0.554 0	24.40 ~ 62.90	23.40 ~ 79.40	25.17 ~ 61.00	10.20 ~ 83.70

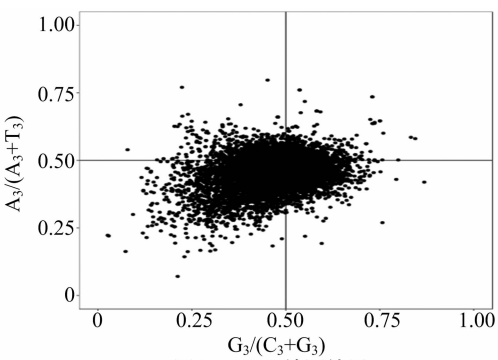

图1　PR2 绘图结果

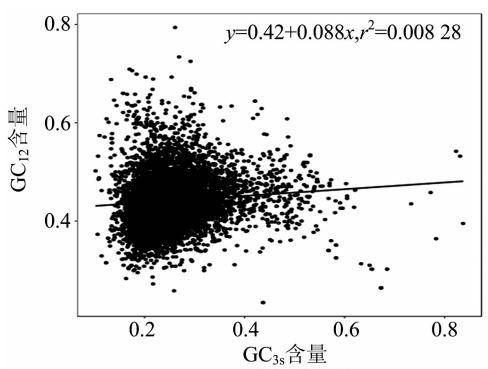

图2　中性绘图结果

由图 2 的中性绘图结果可以看出,GC_{12} 含量与 GC_3 含量呈现出负相关,但相关性不明显($r^2=0.008\ 28$),说明大头金蝇转录组序列的密码子受到的突变压力较小,GC 含量较为保守,其密码子的使用更多地受到选择压力的影响[12-14]。

2.2　ENC、CAI 的分析结果

有效密码子数是衡量基因密码子偏好性的一个重要指标,数值范围为 20 个(每个氨基酸只使用 1 个同义密码子的极端偏好情况) ~ 61 个(每个同义密码子被平均使用的无偏好情况)。研究表明,当 ENC≤35 个时,基因密码子的使用偏好性随 ENC 值的降低而增强[15]。大头金蝇转录组序列的 ENC 在 25.17 ~ 61.00 个之间,均值为 43.16 个(表 2),在 10 923 条序列中仅有 359 条序列的 ENC 小于 35 个[16]。CAI 在 0.099 ~ 0.554 之间,均值为 0.215 8。说明大头金蝇转录组中整体密码子偏好性较低,只有极少部分序列具有较强的密码子偏好性。以 ENC 为纵坐标、GC_3 为横坐标进行 ENC 绘图分析发现,大部分序列沿标准曲线分布,小部分序列位于标准曲线以下较远的位置(图 3),说明大头金蝇转录组的密

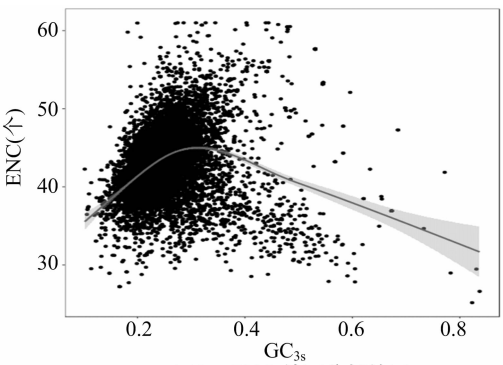

图3　ENC 绘图分析结果

码子偏好性主要是受突变的影响,而选择压力仅在小部分序列中表现得比较明显。

2.3　同义密码子及最优密码子分析

经过计算可知,13 条编码蛋白基因密码子中 RSCU 大于 1 的共有 29 个,除色氨酸(Trp)外,其他 19 种氨基酸及终止子均有 RSCU 值大于 1 的密码子。在这 29 个使用度较高的密码子中,第 3 位点嘌呤为 U 的有 15 个,为 A 的有 12 个,为 G 的有 1 个,为 C 的有 1 个,可以看出,在使用度较高的密码子中,绝大部分以 A 或 U 结尾(表 3)。

表 3　同义密码子的相对使用度分析结果

氨基酸	密码子	数目(个)	RSCU 值	氨基酸	密码子	数目(个)	RSCU 值
苯丙氨酸(Phe)	UUU	156 025	1.34	天门冬氨酸(Asp)	GAU	297 609	1.72
	UUC	76 374	0.66		GAC	48 709	0.28
酪氨酸(Tyr)	UAU	158 352	1.61	谷氨酸(Glu)	GAA	347 854	1.61
	UAC	38 600	0.39		GAG	83 383	0.39
半胱氨酸(Cys)	UGU	87 665	1.54	亮氨酸(Leu)	UUA	208 862	2.22
	UGC	26 150	0.46		UUG	174 183	1.85
Ter(终止子)	UAA	7 217	1.98		CUU	65 110	0.69
	UAG	1 733	0.48		CUC	25 745	0.27
	UGA	1 793	1.54		CUA	57 114	0.61
色氨酸(Trp)	UGG	60 051	1.00		CUG	33 559	0.36
脯氨酸(Pro)	CCU	105 500	1.33	组氨酸(His)	CAU	129 766	1.61
	CCC	78 771	1.00		CAC	31 123	0.39
	CCA	105 410	1.33	谷氨酰胺(Gln)	CAA	240 805	1.59
	CCG	26 879	0.34		CAG	62 525	0.41
精氨酸(Arg)	AGA	76 914	1.49	异亮氨酸(Ile)	AUU	170 754	1.41
	AGG	14 133	0.27		AUC	43 262	0.36
	CGU	149 393	2.88	甲硫氨酸(Met)	AUA	148 840	1.23
	CGC	35 986	0.69		AUG	151 370	1.00
	CGA	26 588	0.51	苏氨酸(Thr)	ACU	132 600	1.30
	CGG	7 704	0.15		ACC	83 938	0.83
天冬酰胺(Asn)	AAU	322 457	1.64		ACA	155 224	1.53
	AAC	69 771	0.36		ACG	34 798	0.34
丝氨酸(Ser)	AGU	125 139	1.42	赖氨酸(Lys)	AAA	321 131	1.51
	AGC	40 866	0.46		AAG	105 238	0.49
	UCU	116 548	1.32	丙氨酸(Ala)	GCU	192 841	1.87
	UCC	64 869	0.74		GCC	109 409	1.06
	UCA	127 245	1.44		GCA	89 008	0.86
	UCG	54 510	0.62		GCG	21 460	0.21
缬氨酸(Val)	GUU	147 530	1.64	甘氨酸(Gly)	GGU	200 763	2.32
	GUC	39 049	0.43		GGC	67 751	0.78
	GUA	109 841	1.22		GGA	69 502	0.80
	GUG	62 740	0.70		GGG	8 839	0.10

采用 ΔRSCU 值法对大头金蝇转录组序列进行最优密码子的确定,共筛选出 UUC、UUG、CUC、AUU、AUC、GUU、GUC、UAC、CAC、CAA、AAC、AAG、GAC、GAA、UCU、UCC、AGC、CCU、CCC、ACU、ACC、GCU、GCC、UGC、CGU、CGC、GGU、GGC 共 28 个最优密码子,分别编码 Phe、Leu、Ile、Val、Tyr、His、Gln、Asn、Lys、Asp、Glu、Ser、Pro、Thr、Ala、Cys、Arg、Gly 共 18 种氨基酸(表 4)。这 28 个最优密码子中以 C 结尾的有 16 个,以 U 结尾的有 8 个,以 A、G 结尾的均为 2 个,这与高频密码子的统计结果相似,说明大头金蝇最优密码子偏向于以 C、U 结尾。

3　讨论

目前已完成的双翅目类群转录组的测序工作不多,基于昆虫转录组的密码子偏好性分析结果更少,本研究结果与其他昆虫类群转录组密码子的使用模式是否一致,还需进一步分析确定。因此,若需要明确昆虫基因组密码子的使用模式及其与基因表达等之间的深入关系,则需要对昆虫线粒体基因组数据进行大量统计分析,而目前各数据库中有关昆虫线粒体基因组的数据相对较少,是否可以借鉴真菌、植物等真核生物线粒体成功测序的经验完成大量昆虫基因组的测序,进而为其密码子的真正“解密”提供原始材料,有待进一步研究。

参考文献:

[1]贾新平,孙晓波,邓衍明,等. 鸟巢蕨转录组高通量测序及分析[J]. 园艺学报,2014,41(11):2329-2341.

[2]Margulies M,Egholm M,Altman W E,et al. Genome sequencing in microfabricated high-density picolitre reactors[J]. Nature,2005,437(757):376-380.

[3]张棋麟,袁明龙. 基于新一代测序技术的昆虫转录组学研究进展[J]. 昆虫学报,2013,56(12):1489-1508.

表 4　大头金蝇中高/低表达样本组密码子使用频率的比较

氨基酸	密码子	高 RSCU 组与其数目(个)	低 RSCU 组与其数目(个)	氨基酸	密码子	高 RSCU 组与其数目(个)	低 RSCU 组与其数目(个)
Phe	UUU	0.54(2 397)	1.55(5 888)	Ser	UCU*	1.91(3 396)	1.26(9 350)
	UUC*	1.46(6 461)	0.45(1 695)		UCC*	1.53(2 714)	0.60(4 427)
Tyr	UAU	1.02(3 405)	1.58(5 649)		UCA	0.67(1 192)	1.76(13 020)
	UAC*	0.98(3 248)	0.42(1 511)		UCG	0.37(663)	0.58(4 322)
Cys	UGU	1.17(1 910)	1.50(2 986)		AGU	0.80(1 420)	1.32(9 742)
	UGC*	0.83(1 368)	0.50(1 004)		AGC*	0.72(1 720)	0.48(3 539)
Ter	UAA	2.73(497)	1.88(342)	Leu	UUA	0.72(1 992)	2.61(9 906)
	UAG	0.19(34)	0.40(73)		UUG*	3.89(10 822)	1.36(5 153)
	UGA	0.08(15)	0.72(131)		CUU	0.74(2 050)	0.74(2 798)
Trp	UGG	1.00(2 016)	1.00(1 422)		CUC*	0.44(1 226)	0.23(876)
Pro	CCU*	1.47(3 259)	1.14(6 117)		CUA	0.14(396)	0.72(2 731)
	CCC*	1.78(3 943)	0.69(3 691)		CUG	0.08(209)	0.35(1 341)
	CCA	0.71(1 576)	1.75(9 396)	His	CAU	1.13(2 470)	1.59(8 084)
	CCG	0.03(72)	0.41(2 220)		CAC*	0.87(1 910)	0.41(2 056)
Arg	CGU*	4.01(6 224)	2.25(5 442)	Gln	CAA*	1.81(6 335)	1.53(16 935)
	CGC*	0.97(1 513)	0.65(1 567)		CAG	0.19(653)	0.47(5 213)
	CGA	0.04(60)	0.84(2 029)	Ile	AUU*	1.97(7 510)	1.25(6 449)
	CGG	0.02(33)	0.21(507)		AUC*	0.87(3 336)	0.28(1 429)
	AGA	0.81(1 252)	1.66(4 017)		AUA	0.16(611)	1.47(7 592)
	AGG	0.15(238)	0.41(982)	Thr	ACU*	1.70(4 291)	1.09(8 669)
Asn	AAU	1.02(4 379)	1.57(2 261)		ACC*	1.74(4 395)	0.52(4 121)
	AAC*	0.98(4 241)	0.43(6 168)		ACA	0.49(1 223)	2.00(15 897)
Val	GUU*	1.97(7 196)	1.63(5 774)		ACG	0.07(166)	0.40(3 177)
	GUC*	0.81(2 953)	0.39(1 373)	Ala	GCU*	2.35(10 069)	1.49(8 081)
	GUA	0.95(3 461)	1.30(4 602)		GCC*	1.46(6 282)	0.76(4 107)
	GUG	0.27(1 003)	0.68(2 423)		GCA	0.14(608)	1.45(7 860)
Asp	GAU	1.61(8 285)	1.66(12 120)		GCG	0.05(196)	0.31(1 677)
	GAC*	0.39(2 002)	0.34(2 520)	Gly	GGU*	2.94(11 230)	1.95(7 343)
Glu	GAA*	1.74(11 564)	1.60(14 015)		GGC*	0.78(2 982)	0.72(2 736)
	GAG	0.26(1 754)	0.40(3 467)		GGA	0.25(951)	1.15(4 359)
Lys	AAA	1.04(7 306)	1.58(15 152)		GGG	0.02(94)	0.18(662)
	AAG*	0.96(6 685)	0.42(4 025)				

注:“*”代表最优密码子。

[4] Olejniczak M, Uhlenbeck O C. tRNA residues that have coevolved with their anticodon to ensure uniform and accurate codon recognition [J]. Biochimie, 2006, 88(8): 943 – 950.

[5] Campos J L, Zeng K, Parker D J, et al. Codon usage bias and effective population sizes on the X chromosome versus the autosomes in *Drosophila melanogaster* [J]. Molecular Biology and Evolution, 2013, 30(4): 811 – 823.

[6] 时　慧, 王　玉, 杨路成, 等. 茶树抗寒调控转录因子 ICE1 密码子偏性分析[J]. 园艺学报, 2012, 39(7): 1341 – 1352.

[7] Quax T E F, Claassens N J, Söll D, et al. Codon bias as a means to fine – tune gene expression[J]. Molecular Cell, 2015, 59(2): 149 – 161.

[8] 薛万琦, 赵建铭. 中国蝇类(下册)[M]. 沈阳: 辽宁科学技术出版社, 1998: 1438 – 1452.

[9] Wright F. The ‘effective number of codons’ used in a gene[J]. Gene, 1990, 87(1): 23 – 29.

[10] Sharp P M, Li W H. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications[J]. Nucleic Acids Research, 1987, 15(3): 1281 – 1295.

[11] Bellgard M, Schibeci D, Trifonov E, et al. Early detection of G + C differences in bacterial species inferred from the comparative analysis of the two completely sequenced *Helicobacter pylori* strains [J]. Journal of Molecular Evolution, 2001, 53(4/5): 465 – 468.

[12] Sueoka N. Directional mutation pressure and neutral molecular evolution[J]. Proceedings of the National Academy of Sciences of the United States of America, 1988, 85(8): 2653 – 2657.

[13] Sueoka N. Two aspects of DNA base composition: G + C content and translation – coupled deviation from intra – strand rule of A = T and G = C[J]. Journal of Molecular Evolution, 1999, 49(1): 49 – 62.

[14] Nie X J, Deng P C, Feng K W, et al. Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family[J]. Plant Molecular Biology Reporter, 2014, 32(4): 828 – 840.

[15] Comeron J M, Aguadé M. An evaluation of measures of synonymous codon usage bias[J]. Journal of Molecular Evolution, 1998, 47(3): 268 – 274.

[16] Rai A, Yamazaki M, Takahashi H, et al. RNA – seq transcriptome analysis of *Panax japonicus*, and its comparison with other panax species to identify potential genes involved in the saponins biosynthesis[J]. Frontiers in Plant Science, 2016, 7: 481.