水生生物学报
ACTA HYDROBIOLOGICA SINICA

**艾美游仆虫大核基因组与转录组测序及结构特征**

何燕 尹家奇 生欣

**MACRONUCLEUS GENOME AND TRANSCRIPTOME SEQUENCING AND STRUCTURE ANALYSIS OF *EUPLOTES AMIETI***

HE Yan, YIN Jia-Qi, SHENG Xin

在线阅读 View online: https://doi.org/10.7541/2021.2020.212

---

doi: 10.7541/2021.2020.212

# 艾美游仆虫大核基因组与转录组测序及结构特征

何 燕 尹家奇 生 欣
(遵义医科大学生物化学与分子生物学教研室, 遵义 563000)

**摘要:** 为获得艾美游仆虫(*Euplotes amieti*)大核基因组结构特征、分析基因功能及其表达调控方式, 研究采用高通量测序技术对艾美游仆虫进行了大核基因组与转录组测序, 结果显示基因组测序最终得到原始reads数据为10.92 Gb, 过滤后得到50287条Contigs。GC含量较低, 为31%; 其中两端同时具有端粒的微染色体数量为27542条, 占54.76%, 只含有一端端粒的基因数量为6118条。Contigs进行基因结构分析, 96.5%的基因能够被预测出功能, 最终得到27650条基因, 平均外显子长度为311.69 bp; 内含子较短, 平均长度为150 bp。转录组测序结果为76219898条, 拼接后获得38588条转录组Unigenes, 其平均长度为1189 bp。将转录组的38588条Unigenes比对发现有2%—3%基因发生了编程性移码, 其中, 绝大多数为+1PRF; 除此之外, 艾美游仆虫的终止密码子还存在重配现象, 其终止密码子为UAA和UAG, 而UGA编码半胱氨酸或硒代半胱氨酸。这与游仆虫属的编程性移码及终止密码子重配的特点一致。将27650条基因组Contigs与38588条转录组Unigenes成功获得注释。基因功能分析显示转录本显著富集于细胞生长与死亡、膜转运、运输与细胞学过程等。将基因组与转录组随机各抽取50个基因进行PCR验证。95%的基因均验证成功。结果表明艾美游仆虫除了具有游仆虫特有的微染色体、密码子重新分配和程序性移码等基因组特征外, 还具有大量“联合微染色体”, 编码大量与感受外界环境变化、细胞周期与蛋白表达调控等相关的特殊蛋白质, 并通过一定量的miRNA与锌指类转录因子对基因进行表达调控。

**关键词:** 艾美游仆虫; 大核基因组; 转录组; 微染色体; 基因注释
**中图分类号:** Q959.116 **文献标识码:** A **文章编号:** 1000-3207(2021)05-1014-10

游仆虫是进化上较为高等的腹毛目纤毛虫, 为一种常见的水生浮游生物, 以水中微生物为食。具有十分复杂的皮层微管骨架系统和背腹分化的纤毛器。常被作为研究微管装配特征、纤毛结构与功能、纤毛形态发生、纤毛基因定位和功能研究的重要模式生物。此外, 游仆虫不仅同时具有负责营养的大核与负责生殖的小核, 而且具有其独特的大核基因结构特征, 如微染色体、特殊的终止密码子和高频率的编程性核糖体移码(Programmed ribosomal frameshifting, PRF)等。因此, 近年来, 游仆虫受到研究者的广泛关注, 目前, 八肋游仆虫(*Euplotes octocarinatus*)、厚游仆虫(*Euplotes crassus*)和扇形游仆虫(*Euplotes vannus*)等的基因组数据库已经建立, 为进一步揭秘这种水生浮游生物的基因、细胞和环境生物学特征提供了不可或缺的基础资料。

艾美游仆虫(*Euplotes amieti*)为一类较大型淡水种游仆虫。活体为(88—125) μm×(55—78) μm, 呈不对称的卵圆形, 背腹扁平, 腹面观右缘较左缘膨出, 体前有一明显的领口区开阔呈三角形约占体长的活体时口侧膜不易观察到横棘毛间有列短而不规则的嵴[1]。本课题组前期对艾美游仆虫微管类细胞骨架、纤毛形态发生过程和γ-微管蛋白在纤毛装配过程中的作用进行了研究, 但由于缺乏该种游仆虫基因组和转录组数据, 限制了对其相关基因功能的研究, 因此, 本研究通过提取艾美游仆虫大核基因组DNA和mRNA, 对其大核基因组DNA和转录组进行了测序和基因注释, 旨在筛选出与微管和纤毛相关蛋白基因, 分析其基因结构, 为进一步探索其功能提供基础资料。

**收稿日期:** 2020-09-17; **修订日期:** 2021-02-06
**基金项目:** 国家自然科学基金(31760616)资助 [Supported by the National Natural Science Foundation of China (31760616)]
**作者简介:** 何燕(1993—), 女, 硕士研究生; 研究方向为纤毛和纤毛疾病的细胞与分子生物学。E-mail: 1458820307@qq.com
**通信作者:** 生欣(1983—), 女, 博士; 研究方向为纤毛和纤毛疾病的细胞与分子生物学。E-mail: xshengbio@163.com

## 1　材料与方法

### 1.1　材料

本实验所用的艾美游仆虫(*Euplotes amieti*)采自上海市青浦区近郊的农田水塘中。本实验所用DNA提取试剂盒, RNA提取试剂盒购买于TaKaRa公司。基因组测序及注释由北京诺禾致源科技有限公司完成。转录组测序由上海派森诺生物科技股份有限公司完成。

### 1.2　方法

**艾美游仆虫的培养及收集**　本研究所用的腹毛类纤毛虫艾美游仆虫使用长梭绿藻(*Chlorogoium elongatum*)喂食。在收集艾美游仆虫细胞前, 先进行饥饿处理5—7d。然后用孔径较小的纱布过滤饥饿后的细胞, 除去较大的杂质。随后用定性滤纸浓缩收集虫体。同时除去较小杂质。最后用纯净水将虫体收集到离心管中。4000 r/min离心5 min收集虫体, 弃掉上清备用。

**艾美游仆虫DNA的提取及基因组测序**　参照TaKaRa公司DNA提取试剂盒说明书, 提取的DNA采用琼脂糖凝胶电泳分析DNA的纯度和完整性。Nanodrop检测DNA的纯度($A_{260}/A_{280}$比值); Qubit对DNA浓度进行精确定量。检验合格的DNA样品通过Covaris破碎机随机打断成长度为350 bp的片段。采用 NEB Next® Ultra DNA Library Prep Kit(NEB, USA)进行建库, 构建好的文库通过Illumina NovaSeq PE150进行测序。然后再采用SPAdes的“careful”模式进行组装。随后, 利用CAP3将SPAdes 的拼接结果进行融合。考虑其一些reads质量较低, 采用blat将长度小于500 bp的序列与长度大于500 bp的序列比对, 去掉一致性百分比≥90%, 覆盖度≥80%的序列、细菌、古菌污染的DNA、无端粒序列及线粒体基因组序列, 11条长度小于100 bp的序列等共去除9508条。最终得到Clean Contigs用于全基因组注释。

**RNA 的提取及转录组测序**　RNA 的提取参照TaKaRa公司RNA提取试剂盒说明书, 将提取的RNA采用离子打断的方式打断成300 bp左右的片段。构建好大小为450 bp文库。采用第二代测序技术, 基于 Illumina HiSeq 测序平台, 对这些文库进行双末端 (Paired-end, PE)测序。Cutadapt 去除 3′端带接头的序列和平均质量分数低于Q20 的Reads, 得到的CleanData 用Trinity 软件与基因组比对, 然后根据比对结果, 使用Trinity的Genome-guid模式进行从头拼接、聚类。挑选最长的转录本作为Unigene, 并进行后续基因功能注释和基因结构预测等。

**基因注释**　重复序列注释方法使用Repeatmasker和repeatproteinmas软件对重复序列数据库RepBase库进行同源序列比对, 识别与已知重复序列相似的序列。从头预测使用LTR_FINDER和RepeatScout, RepeatModele等软件首先建立*de novo*重复序列库, 将*de novo*预测出来的重复序列库与同源重复序列数据库Repbase进行整合, 再用RepeatMasker软件对艾美游仆虫基因组进行repeat注释。

基因的结构预测, 主要通过同源预测, *de novo*预测和其他证据支持的预测。同源预测的方法是将已知的同源物种的编码蛋白序列与新物种的基因组序列进行比对(同源物种个数通常8—10个), 通过blast, genewise等比对软件预测基因组中的基因结构。*de novo*预测使用依赖于基因组序列数据统计学特征(如密码子频率和外显子-内含子分布)的软件来预测基因结构, 常用的软件有Augustus、GlimmerHMM和SNAP等。再结合转录组比对数据, 使用EVidenceModeler(EVM)整合软件, 将各种方法预测得到的基因集整合成一个非冗余的, 更加完整的基因集。最后, 使用PASA(http://pasa.sourceforge.net/), 结合转录组组装结果, 对EVM的注释结果进行校正, 加入UTR及可变剪切等信息, 得到最终的基因集。

基因功能注释, 是将基因结构注释得到的基因集, 利用比对软件与已知蛋白数据库(SwissProt、Nr、Pfam、KEGG和InterPro)等比对, 得到基因的功能信息。注释物种包括三伪尖毛虫(*Oxytricha trifallax*, *O. trifallax*)、线虫(*Caenorhabditis elegans*, *C. elegans*)、浮萍棘尾虫(*Stylonychia lemnae*, *S. lemnae*)、扇形游仆虫(*Euplotes vannus*, *E. vannus*)、草履虫(*Paramecium duboscqui*, *P. duboscqui*)、八肋游仆虫(*Euplotes octocarinatus*, *E. octocarinatus*)、四膜虫(*Tetrahymena utriculariae*, *T. utriculariae*)、杜氏利什曼虫(*Leishmania donovani*, *L. donovani*)、有孔虫(*Reticulomyxa filosa*, *R. filosa*)和艾美游仆虫(*Euplotes amieti*, *E. amieti*)等[2—15]。

非编码RNA的注释包括tRNA、rRNA、miRNA和snRNA。根据tRNA的结构特征, 利用tRNAscan-SE(http://lowelab.ucsc.edu/tRNAscan-SE/)软件来寻找基因组中的tRNA序列; 由于rRNA具有高度的保守性, 因此可以选择近缘物种的rRNA序列作为参考序列, 通过blast比对来寻找基因组中的rRNA; 利用Rfam家族的协方差模型, 采用Rfam自带的INFERNAL(http://infernal.janelia.org/)软件可预测基因组上的miRNA和snRNA序列信息。最终得到艾

美游仆虫基因组的ncRNA信息。

## 2 结果

### 2.1 艾美游仆虫基因组和转录组测序数据分析

鉴于游仆虫基因组特点, 利用二代测序技术Illumina测序平台对艾美游仆虫基因组的200 bp小片段数据质控和建库进行测序, 共得到10.92 Gb 原始数据。用 SPAdes拼接软件进行组装, 初始组装产生了64836条Contigs。由于在提取游仆虫基因组DNA的过程中, 无法将其线粒体基因组分离去除, 所以拼接结果中也可能混有线粒体DNA的污染和其他杂菌污染, 初始数据经去除细菌、线粒体基因组DNA污染, 去除<100 bp短序列后, 最终得到50287条艾美游仆虫基因组序列。这些基因序列的GC含量较低, 为 31%, 与其他纤毛虫类似。此外, 两端同时具有端粒的微染色体数量为27542条, 占54.76%, 只含有一端端粒的基因数量为6118条(表 1)。与其他游仆虫相比, 艾美游仆虫大核基因组均大于八肋游仆虫(88.9 Mb)和扇形游仆虫(85.1 Mb)。但微染色体数量小于八肋游仆虫(29413)与扇形游仆虫(37501)。

转录组测序结果去除低质量的Raw Reads, 最终得到68900504条Clean Reads, 经Trinity拼接后得到60691条Transcripts, 其平均长度和$N_{50}$值分别为1326.6和 1759 bp。所有的transcripts进一步聚类后获得38588条Unigenes, Unigenes的平均长度和$N_{50}$值分别为1189.9和1643 bp(表 2), 将38588条转录组序列通过NCBI中进行BLASTX进行比对, 并进一步通过查找开放阅读框, 以$E$值$2\times10^{-5}$作为标准, 发现其中2%—3%基因发生了编程性移码。

**表 1 艾美游仆虫基因组测序数据统计**

Tab. 1 Genomic sequencing data statistics of *Euplotes amieti*

| 拼装结果 Assembler | SPAdes 测序软件 | SPAdes+CAP3拼接软件 | 最终拼装结果 Final assembly |
|---|---|---|---|
| 组装的大小Assembly size (Mb) | 95.50 | 94.56 | 89.46 |
| 片段Contigs (*n*) | 64836 | 59795 | 50287 |
| 具有两端端粒的片段 Number of 2-telomere contigs (*n*) | 27315 | 27544 | 27542 |
| 具有一端端粒的片段 Number of 1-telomere contigs (*n*) | 7719 | 6537 | 6118 |
| 不具有端粒的片段 Number of 0-telomere contigs (*n*) | 29802 | 25714 | 16627 |
| 长度小于500 bp的片段 Number of length<500 bp | 25179 | 22279 | 13472 |
| GC百分比GC (%) | | | 31 |

**表 2 艾美游仆虫转录组拼接序列**

Tab. 2 Transcriptome splicing sequence of *Euplotes amieti*

| 序列Sequence | 重叠群 Contigs | 转录组序列 Transcript | 转录本代表序列 Unigene |
|---|---|---|---|
| 序列总数Sequence number | 68902 | 60691 | 38588 |
| 序列最大长度Max. length (bp) | 16644 | 16648 | 16648 |
| 序列平均长度Mean length (bp) | 609.45 | 1326.6 | 1189.99 |
| 所有序列长到短排列并相加, 相加的长度达到序列总长度的50%时, 最后一条序列长度 $N_{50}$ (bp) | 1319 | 1759 | 1643 |
| 长度大于$N_{50}$ 的序列总数 $N_{50}$ Sequence No. | 9738 | 15227 | 9000 |
| 所有序列长到短排列并相加, 相加的长度达到序列总长度的90%时, 最后一条序列长度 $N_{90}$ (bp) | 203 | 699 | 581 |
| 长度大于$N_{90}$的序列总数 $N_{90}$ Sequence No. | 39290 | 42291 | 26558 |
| 序列的GC含量GC (%) | 33.47 | 32.99 | 33.19 |

用alveolata_odb10 进行BUSCO评估测序结果的完整度, 其完整性评估基于其他原生动物的数据库建立, 且通过与目前已知的八肋游仆虫和扇形游仆虫基因组比较, 在基因组大小、微染色体数量、基因的GC含量、N50和基因长度分布等方便均与已报道游仆虫类似, 表明基因组测序质量较好。

基因组通过结构预测得到了27650个基因, 其中96.5%的基因能够被预测出功能。分别将基因组27650个基因与转录组38588条Unigenes对比到已知的NR、GO、KEGG、SwissProt、Pfam、InterPro和eggNOG数据库, 其中有26673条大核基因获得注释, 在各数据库中注释成功的基因数与转录本数分别见图 1A和1B。功能分析显示, 在eggNOG数据库中注释最多的功能类为信号转导机制(1790条)和转录后修饰、蛋白折叠和分子伴侣(1157条)。而GO注释主要富集在生殖词条中(1599条); KEGG 分析富集主要在外界环境信息的信号转导(962条)、细胞内物质运输与代谢分解、分泌、细胞增殖与死亡和环境适应等。这些功能与游仆虫捕食浮游生物并进行细胞内消化与营养物质的运输、适应外环境变化、分裂与结合生殖和个体间信息传递等生命活动密切相关。

### 2.2 基因结构分析

将艾美游仆虫基因组测序得到的50287条大核基因组进行重复序列的注释得到含有19.31%的repeat序列。结合转录组测序结果进行基因结构注释; 注释物种包括*O. trifallax*、*C. elegans*、*S. lemnae*、*E. vannus*、*P. duboscqui*、*E. octocarinatus*、*T. utriculariae*、*L. donovani*、*R. filosa*和*E. amieti*等。结

果显示艾美游仆虫的平均transcript长度为680.21 bp, 其中, 平均CDS为560.11 bp, 是几种纤毛虫中最小的(图 2A); 而每个基因的平均外显子数为1.8, 是除八肋游仆虫外最大的(图 2B)。此外, 其平均外显子与内含子大小分别为311.69和150.70 bp。与以上几种纤毛虫相比, 尽管艾美游仆虫的平均trancript、CDS和外显子长度均较小, 而平均内含子长度与平均外显子数量却较大(图 2C)。

通过对转录组38588条Unigenes分析发现艾美游仆虫中同样存在较高频率的编程性核糖体移码现象, 发生率为2%—3%。其中绝大多数为+1PRF基因, 发生移码的位置多见于滑动序列的终止密码子TAR(R为A或G); 也有部分为+2PRF基因, 多发生在滑动序列的终止密码子TAR(R为A或G), 同时跳过T和A, 这些可以证明艾美游仆虫也存在编程性移码突变的现象。此外, 艾美游仆虫的终止密码子既可以做终止密码子也可以编码蛋白质。将艾美游仆虫与其他几种纤毛虫的终止密码子做比较, 艾美游仆虫与八肋游仆虫一样, UAA 和 UAG常作为终止密码子, 而UGA编码半胱氨酸和硒代半胱氨酸。而在四膜虫草履虫和尖毛虫中, 只有 UGA 作为终止密码子, 而UAA和UAG 则编码谷氨酰胺(图 3)。

## 2.3　非编码RNA及转录因子分析

利用tRNAscan-SE和Rfam等软件对基因组中的非编码RNA序列信息注释; 结果显示miRNA的基因有23个, 平均长度为125.78 bp; tRNA有105个, 平均长度为74.32 bp; rRNA有56个, 其中, 18S 26个和28S 2个。平均长度分别为104.68和104.81 bp; snRNA 21个其中包括CD-box 1个, splicing 9个, 平均长度分别为263和145.22 bp(表 3)。转录因子(Transcription factor, TF)是一类能与基因5′端上游特定序列专一性结合, 并与RNA聚合酶Ⅱ形成转录起始复合体, 共同参与转录起始的过程的蛋白质分子。将植物和动物与PlantTFDB和AnimalTFDB数据库比较, 从而预测得到转录因子及转录因子所属的家族信息。在所有预测的转录因子家族中, zf-

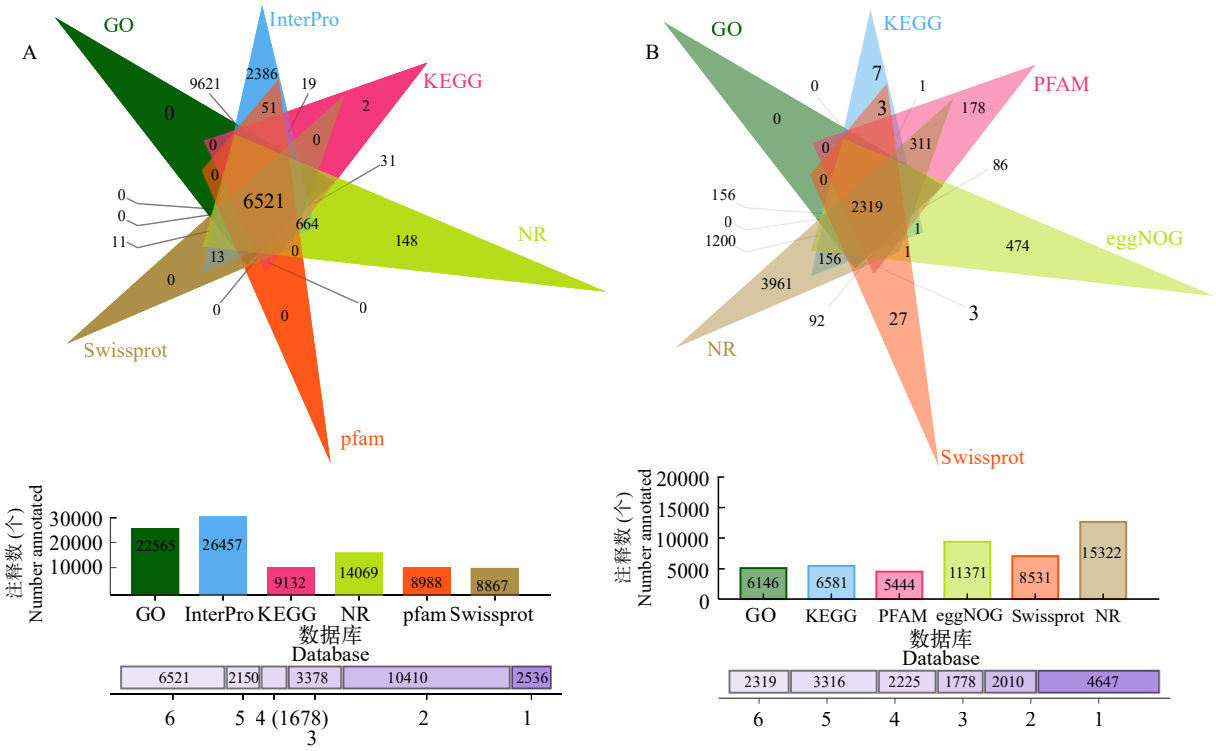


图 1　基因组及转录组测序结果

Fig. 1　Genome and transcriptome sequencing results

A. 基因组注释, B. 转录组注释; GO. GO数据库中被注释到的Conting和Unigene数目; NR. NR数据库中被注释到的Conting和Unigene数目; KEGG. KEGG数据库中被注释到的Conting和Unigene数目; Swissprot. Swissprot数据库中被注释到的Conting和Unigene数目; Pfam. Pfam数据库中注释的Conting数目; Interpro. Interpro 数据库中注释的Conting数目; PFAM. PFAM数据库中注释到的Unigene数目

A. Genome annotation; B. Transcriptome annotation. GO. Number of Contigs and Unigenes annotated in the GO database; NR. Number of Contigs and Unigenes annotated in the NR database; KEGG. Number of Contigs and Unigenes annotated in the KEGG database; Swissprot. Number of Contigs and Unigenes annotated in the Swissprot database; Pfam. Number of Contigs annotated in the Platform database; Interpro. Number of Contigs annotated in the Interpro database; PFAM. Number of Unigenes annotated in the PFAM database

C2H2的Count数最多, 达139条, 其次分别为ZBTB 116条、MYB 89条、bHLH 55条等(图 4)。

### 2.4 基因组及转录组基因验证

将基因组及转录组测序的结果序列中随机取50个基因进行PCR验证。基因组验证结果如图 5A和5B所示, 所用Marker大小为2000 bp, 图中只呈现出1—48个基因的验证条带, 基因组测序结果图 5A和5B可以看出只有1个基因*RABEP*2未P出, 其余各

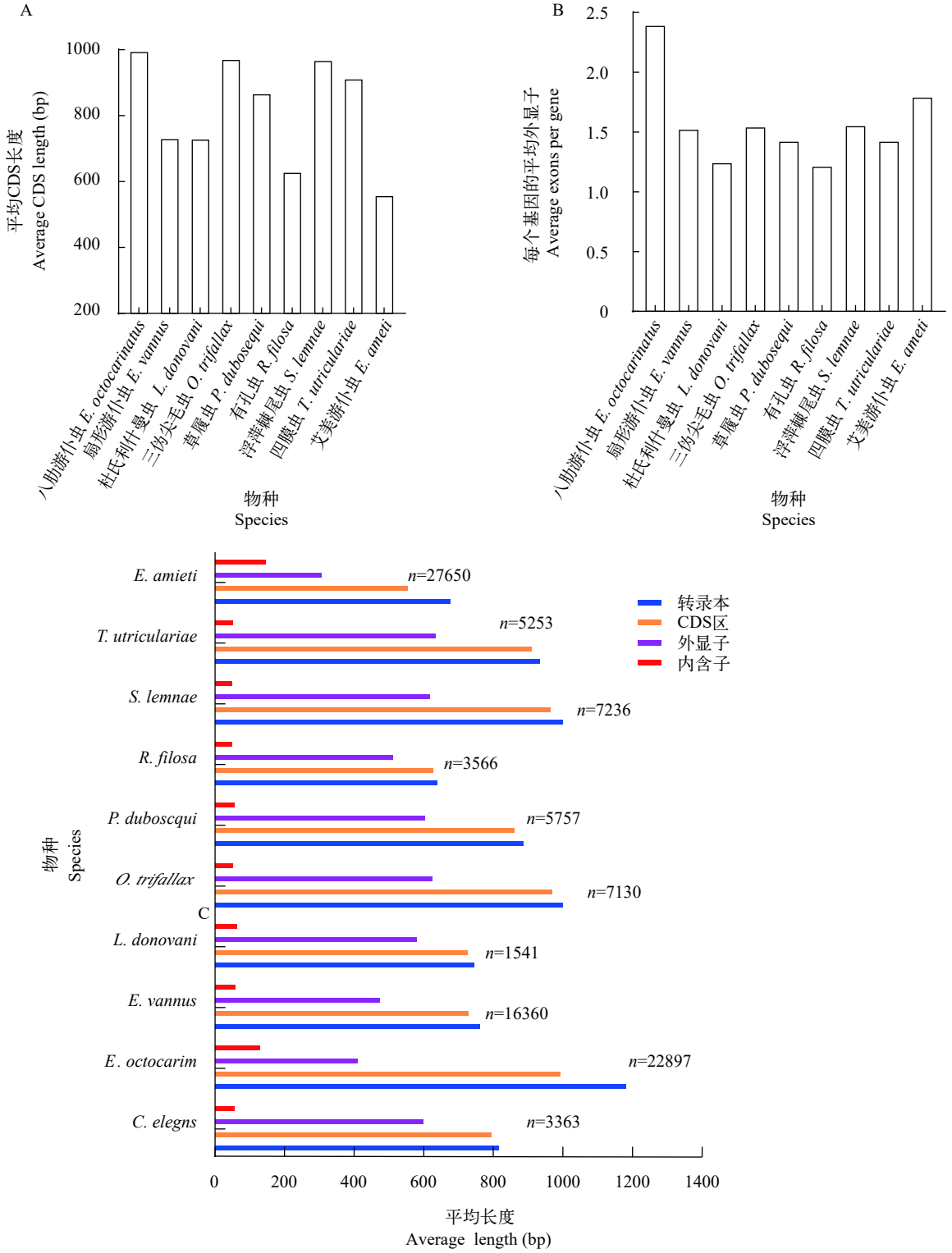


图 2 基因结构特征比较

Fig. 2 Comparison of gene structures

A. 平均CDS长度比较; B. 每个基因平均外显子比较结果; C. 几种腹毛类纤毛虫基因结构比较

A. The average CDS length; B. The average exons of each gene; C. Comparison of the gene structure of some Hypotrichida ciliates

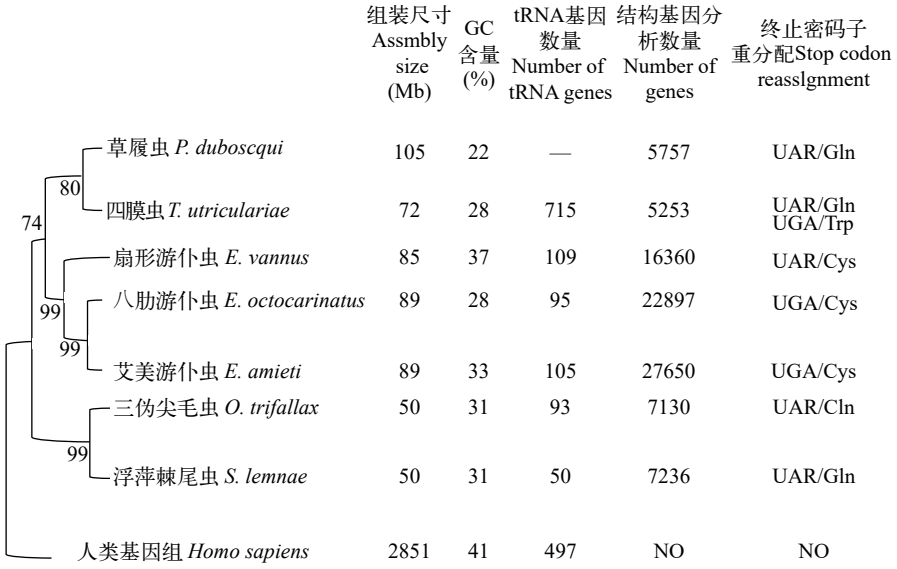


图 3　艾美游仆虫与多种真核生物基因组比较(进化树基于18S rRNA绘制)

Fig. 3　CComparison of representative eukaryotic genome. The tree was constructed based on the sequences of 18S rRNA genes

**表 3　非编码RNA序列注释**

Tab. 3　Noncoding RNA sequence annotation

| 类型 Type | | 拷贝数 Copy (w*) | 平均长度 Average length (bp) | 总长度 Total length (bp) | 占基因组比例Of genome (%) |
|---|---|---|---|---|---|
| miRNA | | 23 | 125.78 | 2893 | 0.003234 |
| tRNA | | 105 | 74.32 | 7804 | 0.008724 |
| rRNA | | 28 | 104.68 | 2931 | 0.003276 |
| rRNA | 18S | 26 | 104.81 | 2725 | 0.003046 |
| | 28S | 2 | 103 | 206 | 0.000230 |
| | 5.8S | 0 | 0 | 0 | 0 |
| | 5S | 0 | 0 | 0 | 0 |
| snRNA | | 11 | 164.73 | 1812 | 0.002026 |
| snRNA | CD-box | 1 | 263 | 263 | 0.000294 |
| | HACA-box | 0 | 0 | 0 | 0 |
| | splicing | 9 | 145.22 | 1307 | 0.001461 |

注: *全基因组注释, 综合数据计算平均长度和总长度

Note: *Genome annotation, synthesis data calculate average length and total length

个基因条带大小与测序的基因大小一致。转录组验证结果如图 5C和5D所示, 图中除了*POC1A*外, 其余各基因条带均与测序结果一致。

## 3　讨论

### 3.1　艾美游仆虫具有游仆虫属共有的基因组特征

原生动物纤毛虫种类繁多, 不同类型的纤毛虫在基因表达调控、大分子相互作用和应激反应等方面呈现极大的多样性特征。如游仆虫中存在的编程性移码现象[16], 尖毛虫中出现的由长非编码RNA指导的大规模基因乱序和重排等[17]。因此, 揭示这些存在于不同类群纤毛虫中的分子与细胞生物学特征对于研究纤毛虫的系统发育和进化关系具有重要的意义。然而, 基因组信息的缺失也极大地限制了纤毛虫分子和细胞生物学水平的多样性研究。目前为止, 已经获得基因组测序资料的纤毛虫有20多种, 而游仆虫只有3种, 分别为厚游仆虫、八肋游仆虫和扇形游仆虫, 这些研究揭示了游仆虫的共有特征: 含有微染色体, 即一个基因一个染色体; 终止密码密码子的重新分配; 程序性核糖体框架转移和对环境应激源的强烈抵抗力等[18]。

本研究对艾美游仆虫基因组和转录组测序组装后, 最终成功注释了50287条大核基因组序列, 其N50为2774 bp, GC含量为31%, 54.8%的基因含有两端端粒, 与以上3种游仆虫大核基因组的组装结果相似[19, 20]。这表明艾美游仆虫测序与拼装结果较好, 且符合游仆虫较低GC含量和微染色体的基因组基本结构特点[19]。同时也存在一定的差异, 其GC含量和微染色体百分比均介于厚游仆虫与八肋游仆虫之间, 且艾美游仆虫中长度小于500 bp的基因与含有一个端粒或无端粒的基因数量均较八肋游仆虫和扇形游仆虫多。排除细菌与小核基因组污染, 推测可能是艾美游仆虫中还含有大量“联合微染色体”, 这些染色体中含有2个以上基因[20]。尽管目前有研究“联合微染色体”上不同基因的表达方向的可能相同或相反, 但微染色体所占的比例是否与生物进化和物种亲缘性相关还未见报道, 但可以确定的是, 不同游仆虫之间存在多样化的基因组特征将为确定种属之间的进化地位提供新的依据。另一方面, 基因结构分析显示基因组序列中含

有19.31%的重复序列, 这种基因的高度串联重复性最早在1983年在鬃棘尾虫小(*Stylonychia pustulata*)核中被报道[21—24], 此后这种现象在纤毛虫大核基因组中被广泛报道, 可见高度重复性是纤毛虫的基因组特征之一。另外, 与几种已知的腹毛类纤毛虫基因结构相比, 艾美游仆虫也具有内含子较短的特征, 平均长度仅为150 bp, 与八肋游仆虫的189 bp类似, 而艾美游仆虫平均CDS区和外显子大小均小于其他物种, 表明艾美游仆虫基因同样具有高度片段化特征。

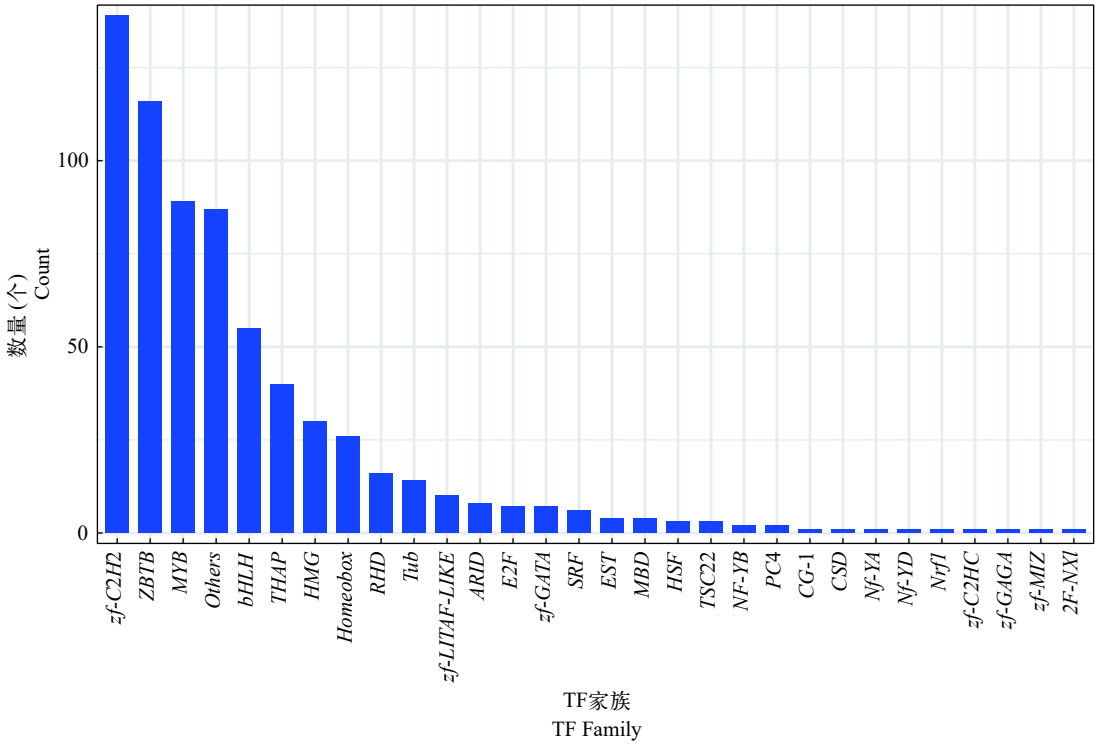


图 4 转录因子家族统计图

Fig. 4 Transcription factor family statistic

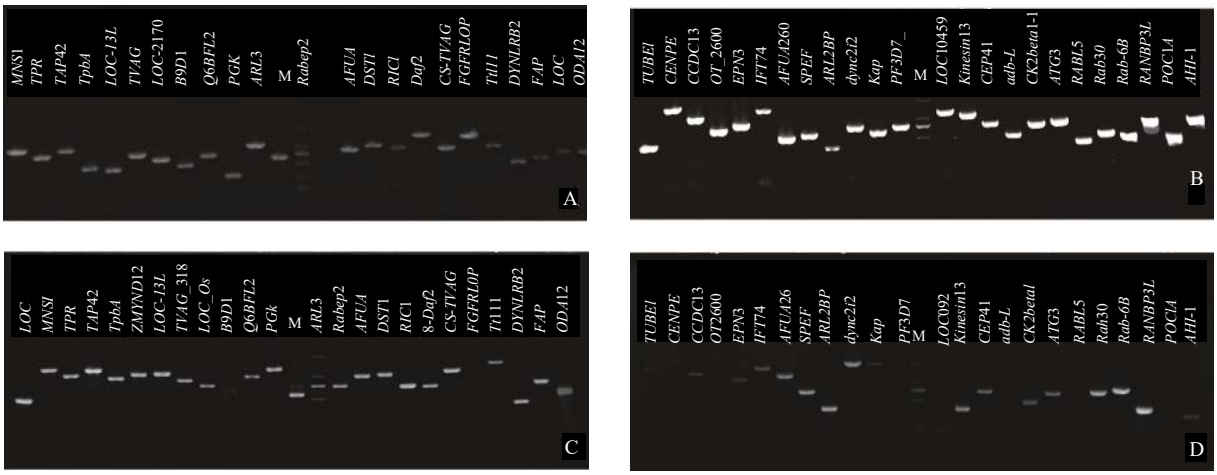


图 5 基因组及转录组结果验证

Fig. 5 Genome and transcriptome validation

A. 基因组测序1—24号基因验证结果; B. 基因组测序26—50号基因验证结果; C. 转录组测序1—24号基因验证结果; D. 转录组测序26—50号基因验证结果

Marker从上到下依次为2000、1000、750、500、250和100 bp

A. Genome sequencing 1—24 gene validation results; B. Genome sequencing 26—50 gene validation results; C. Transcriptome sequencing 1—24 gene validation results; D. Transcriptome sequencing 26—50 gene validation results

Markers from top to bottom are 2000, 1000, 750, 500, 250 and 100 bp

### 3.2　艾美游仆虫基因组同样具有编程性移码

编程性核糖体移码是一种重编码事件, 它是指翻译中的核糖体能够在mRNA 上的特定位置, 从起始的0读框转换到+1或者−1读框, 然后继续进行翻译的现象。这种现象的发生是可调控的, 发生频率高达80%。目前已经报道的核糖体编程性移码信号常见于3个主要元件, 即七核苷酸滑动序列(5′—AAA-UAR-V-3′ R为A 或G; V≠U), 及其上游都有SD相似序列CAAGAA, 和5—12个核苷酸组成的间隔序列以及假结(Pseudonotes)或茎环; 其可产生牵引效应将XXX(AAA、GGG或UUU)和ZZZ(AAA或UUU)引入P和A位点然后引起核糖体结构重排[16]。本研究中2%—3%的编程性移码基因, 与目前报道的八肋游仆虫和扇形游仆虫中3.5%和2.8%的编程性移码基因发生率相当[10,25], 且其中绝大多数为+1PRF。此外, 与其他纤毛虫相比, 艾美游仆虫也发生了终止密码子的重新分配, 与八肋游仆虫和扇形游仆虫一样, UAA 和 UAG作为终止密码子, 而UGA编码半胱氨酸和硒代半胱氨酸。相比之下, 在四膜虫、草履虫和尖毛虫中, 只有 UGA 作为终止密码子, 而UAA和 UAG 则编码谷氨酰胺。以上结果进一步证实游仆虫中存在着高频率的编程性移码现象, 这种现象是在翻译水平上进行基因表达调控的独特方式, 对于游仆虫有限的大核基因信息来说, 通过编程性移框与终止密码子的重新分配, 能够产生多样化的蛋白表型, 是其适应外界环境的变化的分子基础。目前, 部分研究显示游仆虫中的编程性移码在不同类型之间还存在转变现象, 这种现象存在简约和精致的调控机制[16], 这种调控机制的揭示将有利于阐述游仆虫高频率编程性移码与基因表达调控和环境适应性之间的关系[22—24]。

### 3.3　艾美游仆虫的基因表达与适应环境密切相关

另外, 转录组测序获得了60691个转录本, 分别是扇形游仆虫与八肋游仆虫的1.5倍和的3倍。其中, 38588个基因被成功注释, mRNA序列平均长度为1189.99 bp, 96.5%的基因被成功预测功能。通过基因功能分析显示, 基因组绝大多数基因富集到信号转导、转录后修饰和蛋白质折叠, 而GO与KEGG功能分析显示其转录本显著富集于生殖、单细胞过程、外界环境信息的信号转导、细胞内物质运输与代谢分解、分泌、细胞增殖与死亡和环境适应等。与八肋游仆虫和扇形游仆虫的功能富集上具有较大差别。这提示艾美游仆虫在营养生长期, 除了编码游仆虫属特有的功能蛋白外, 还编码大量蛋白质以适应外界环境变化。尽管本研究并未对不同温度、离子浓度及pH等应激状态下的游仆虫进行转录组测序, 从扇形游仆虫在营养条件、极端温度和盐浓度等条件下的基因表达情况可见, 游仆虫中存在具有应对不同理化刺激的基因簇, 这些基因的表达对于其适应外界环境具有重要意义。而不同游仆虫中应对不同刺激基因簇是否一致还有待进一步探索。此外, 非编码RNA与转录因子分析显示, 在艾美游仆虫中存在一定量的非编码RNA, 其中miRNA的数量为23个, 未见长非编码RNA, 这在游仆虫中还未见报道。有研究显示长非编码RNA在尖毛虫中具有指导大规模基因乱序与重排等作用[25], 因此, 这些miRNA是否也参与了游仆虫基因高度片段化和移码突变等过程还有待进一步研究。而在预测的转录因子家族中, 锌指蛋白基因*ZF-C2H2*与*ZBTB*的数量最多, 尽管有研究显示转录因子E2Fl1在嗜热四膜虫中参与了减数分裂[26], 但这2种转录因子在纤毛虫中的功能研究还未见报道, 在植物与细菌中的研究显示, ZF-C2H2与ZBTB参与了真核生物生长发育及逆境胁迫的耐受等[27], 表明艾美游仆虫可能通过这两类转录因子对基因表达调控, 从而适应外界环境的变化。

### 参考文献:


[1] Sheng X, Wang Z J, Yin F, *et al*. Morphology and morphogenesis of the cell microtubular organelles in the cortex of *Euplotes eurystomus* [J]. *Journal of East China Normal University*, 2010(6): 83-90. [生欣, 王正君, 尹飞, 等. 纤毛虫阔口游仆虫皮层微管胞器的形态及形态发生 [J]. 华东师范大学学报, 2010(6): 83-90.]

[2] Eisen J A, Coyne R S, Wu M, *et al*. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila* [J]. *Model Eukaryote*, 2006, **4**(9): e286.

[3] Aury J M, Jaillon O, Duret L, *et al*. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia* [J]. *Nature*, 2006, **444**(7116): 171-178.

[4] Coyne R S, Hannick L, Shanmugam D, *et al*. Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control [J]. *Genome Biology*, 2011, **12**(10): R100.

[5] Swart E C, Bracht J R, Magrini V, *et al*. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16, 000 tiny chromosomes [J]. *PLoS Biology*, 2013, **11**(1): e1001473.

[6] Aeschlimann S H, Jönsson F, Postberg J, *et al*. The draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome [J]. *Genome Biology and Evolution*, 2014, **6**(7): 1707-1723.

[7] Xiong J, Wang G, Cheng J, *et al*. Genome of the facultative scuticociliatosis pathogen *Pseudocohnilembus persalinus* provides insight into its virulence through horizontal gene transfer [J]. *Scientific Reports*, 2015, **5**(1): 15470.

[8] Wang R, Miao W, Wang W, *et al*. EOGD: the *Euplotes octocarinatus* genome database [J]. *BMC Genomics*, 2018, **19**(1): 1-6.

[9] Zhang W, Gao Y, Long M, *et al*. Origination and evolution of orphan genes and de novo genes in the genome of *Caenorhabditis elegans* [J]. *Science China*, 2019, **62**(4): 579-593.

[10] Chen X, Jiang Y, Gao F, *et al*. Genome analyses of the new model protist *Euplotes vannus* focusing on genome rearrangement and resistance to environmental stressors [J]. *Molecular Ecology Resources*, 2019, **19**(5): 1292-1308.

[11] Downing T, Imamura H, Decuypere S, *et al*. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance [J]. *Genome Research*, 2011, **21**(12): 2143-2156.

[12] Chen B, Xu J, Wang W. Anti-silencing factor Asf1 regulates the nuclear stability in *Tetrahymena thermophila* [J]. *Chinese Journal of Biochemistry and Molecular Biology*, 2015, **31**(1): 55-63. [陈波, 许静, 王伟. 抗沉默因子Asf1调控嗜热四膜虫细胞核的稳定性 [J]. 中国生物化学与分子生物学报, 2015, **31**(1): 55-63.]

[13] Wang X Y. Species diversity and molecular phylogeny of periphyton ciliates in the littoral zone of Yantai[D]. Yantai: Yantai University, 2019: 270-295. [王雪艳. 烟台近海周丛类纤毛虫物种多样性及系统发育研究 [D]. 烟台: 烟台大学, 2019: 270-295.]

[14] Gu F K. The morphology and development of ciliates' large and small nuclei [J]. *Chinese Journal of Zoology*, 1993, **28**(5): 54-59. [顾福康. 纤毛虫大核和小核的形态及其发育过程 [J]. 动物学杂志, 1993, **28**(5): 54-59.]

[15] Yi Z Z, Miao M, Gao S, *et al*. On molecular biology of ciliated protozoa: Frontier areas and progresses [J]. *Chinese Science Bulletin*, 2016, **61**(20): 2227-2238. [伊珍珍, 苗苗, 高珊, 等. 纤毛虫原生动物的分子生物学研究: 若干热点领域及新进展 [J]. 科学通报, 2016, **61**(20): 2227-2238.]

[16] Xiao Y, Wang R L, Du J, *et al*. Slippery sequence is important for distinguishing between +1 and +2 programmed ribosomal frameshifting in Euplotes [J]. *Chinese Journal of Biochemistry and Molecular Biology*, 2020, **36**(3): 289-299. [肖羽, 王软林, 杜军, 等. 滑动序列对游仆虫中识别+1位和+2位编程性核糖体移码具有关键作用 [J]. 中国生物化学与分子生物学报, 2020, **36**(3): 289-299.]

[17] Yang R, Chen T B, Huang J, *et al*. Evolution of the scrambled pattern of the actin Ⅰ, α-TBP and DNA pol α gene within the genus *Oxytricha* (Protozoa, Ciliates) [J]. *Acta Hydrobiologica Sinica*, 2017, **41**(2): 285-295. [杨然, 陈天兵, 黄俊, 等. 尖毛虫属Actin Ⅰ、α-TBP和DNA pol α乱序基因的模式研究 [J]. 水生生物学报, 2017, **41**(2): 285-295.]

[18] Sheng X. Study on the microtubular cytoskeleton and γ-tubulin in the hypotrich ciliate *Euplotes eurystomus* [D]. Shanghai: East China Normal University, 2011: 275-286. [生欣. 腹毛类纤毛虫阔口游仆虫微管类细胞骨架及γ-微管蛋白的研究 [D]. 上海: 华东师范大学, 2011: 275-286.]

[19] Wang R L. Identification of programmed ribosomal frameshifting genes and exploration of the frameshift mechanism in *Euplotes octocarinatus* [D]. Taiyuan: Shanxi University, 2017: 235-237. [王软林. 八肋游仆虫中编程性核糖体移码基因的鉴定及其分子机制研究 [D]. 太原: 山西大学, 2017: 235-237.]

[20] Chen X, Jiang Y, Gao F, *et al*. Genome analysis of the unicellular eukaryote *Euplotes vannus* reveals molecular basis for sex determination and tolerance to environmental stresses [J]. *bioRxiv*, 2018: 357715.

[21] Dawson D, Stetler D J, Swanton M T, *et al*. Tandemly repeated sequence families in micronuclear DNA of the ciliate *Stylonychia pustulata* [J]. *Eukaryot Cell*, 1983, **30**(4): 7.

[22] Tan M, Liang A H, Brünen-Nieweler C, *et al*. Programmed translational frameshifting is likely required for expressions of genes encoding putative nuclear protein kinases of the ciliate *Euplotes octocarinatus* [J]. *Journal of Eukaryotic Microbiology*, 2001, **48**(5): 575-582.

[23] Tan M, Heckmann K, Brünen N C. Analysis of micronuclear, macronuclear and cDNA sequences encoding the regulatory subunit of cAMP-dependent protein kinase of *Euplotes octocarinatus*: Evidence for a ribosomal frameshift [J]. *Journal of Eukaryotic Microbiology*, 2001, **48**(1): 80-87.

[24] Wang Z. Study of the Programmed translational frameshifting genes in *Euplotes octocarinatus* [D]. Taiyuan: Shanxi University, 2016: 288-290. [王昭. 八肋游仆虫中编程性翻译移码基因的研究 [D]. 太原: 山西大学, 2016: 288-290.]

[25] Chen X, Bracht J R, Goldman A D, *et al*. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development [J]. *Cell*, 2014, **158**(5): 1187-1198.

[26] Zhang J, Tian M, Yan G X, *et al*. E2fl1 is a meiosis-specific transcription factor in the protist *Tetrahymena thermophila* [J]. *Cell Cycle*, 2017, **16**(1): 123-135.

[27] Zhang K, Qiu S X, Wang T M, *et al*. Research progress of ZBTB32 biological function [J]. *Journal of Chengdu Medical College*, 2016, **11**(3): 381-386. [张科, 邱世香, 王茂田, 等. ZBTB32生物学功能的研究进展 [J]. 成都医学院学报, 2016, **11**(3): 381-386.]

# MACRONUCLEUS GENOME AND TRANSCRIPTOME SEQUENCING AND STRUCTURE ANALYSIS OF *EUPLOTES AMIETI*

HE Yan, YIN Jia-Qi and SHENG Xin

(*Department of Biochemistry and Molecular Biology*, *Zunyi Medical University*, *Zunyi* 563000, *China*)

**Abstract**: To explore genome structure characteristics, gene function and expression regulation of *Euplotes amieti*, high-throughput sequencing technique was used to sequence the macronucleus genome and transcriptome of *Euplotes amieti*. A total of 10.92 Gb data and 50287 Contigs were obtained after filtration. The GC content was 31% and the average length of exons and introns were 311.69 bp and 150 bp, respectively. The number of microchromosomes with telomeres at both ends was 27542, accounting for 54.76%, and the number of genes containing only one telomere was 6118. 38588 Unigenes were split, and the average length was 1189 bp. There were 2%—3% PRF in the total Unigenes, and most of which were +1PRF. Besides, the stop codon of *Euplotes amieti* also has a reconfiguration phenomenon. The stop codons are UAA and UAG, while UGA codes cysteine and selenocysteine. This was consistent with the characteristics of programmed ribosomal frameshifting and stop codon reassignments in Euplotes. 27650 genomic contigs and 38588 transcriptome Unigenes were successfully annotated. Gene function analysis showed that transcripts were significantly enriched in multiple biological processes, mainly involved cell growth and death, membrane transport, transport, and cytology. Moreover, 50 genes were randomly selected from the genome and transcriptome for PCR verification, and 95% of them were successfully verified. The results indicated that in addition to the characteristic microchromosomes and programmed frameshift, the *Euplotes amieti* had a large number of "combined microchromosomes" and encoded a large number of special proteins related to the perception of changes in the external environment, cell cycle, and protein expression regulation. In addition,, the gene expression may be regulated by some miRNA and zinc finger transcription factors.

**Key words**: *Euplotes amieti*; Hronucleus genome; Transcriptome; Microchromosome; Gene annotation