分类号_____密级_____

UDC _____

本 科 毕 业 论 文

拟菱形藻线粒基因组的组装注释与比较研究

学生姓名 李泽林 学号 16040031011

指导教师 陈刚、陈楠生

院、系、中心 海洋生命学院

专业年级 生物科学 2016 级

论文答辩日期 年 月 日

中 国 海 洋 大 学

拟菱形藻线粒体基因组的组装注释与比较研究

完成日期：_____

指导教师签字：_____

答辩小组成员签字：_____

_____

_____

_____

_____

拟菱形藻线粒体基因组的组装注释与比较研究

摘　要

拟菱形藻属有多个有害藻华物种，其中尖刺拟菱形藻引起的有害藻华危害尤其严重，不仅会造成生态灾害，还会释放毒素威胁人类健康。本研究旨在通过构建多个拟菱形藻物种线粒体基因组，比较分析其遗传多样性，来发掘能够有效区分不同拟菱形藻种间以及种内株系的高分辨率分子标记，尤其是用于区分和跟踪研究尖刺拟菱形藻。本研究组装了 7 个尖刺拟菱形藻、2 个连续拟菱形藻和 1 个疑似的多列拟菱形藻的全基因组；从中筛选出线粒体基因组，在每个物种选一个株系完成注释。结合这 10 个株系和 3 个来自公共数据库的株系进行比较分析，检测了 13 个株株系的单核苷酸差异位点、缺失插入位点，以及插入热点，发现了基因组中的高变异区，并基于这些高变异区探讨了高特异性分子标记的筛选和设计；研究结果为拟菱形藻检测提供了关键信息。

关键词：拟菱形藻；线粒体；基因组组装与注释；基因组多样性

Assembly, Annotation and Comparative research on mitochondrial genome of *Pseudo-nitzschia*

Abstract

Many of *Pseudo-nitzschia* are harmful algae, the harmful algal blooms caused by *Pseudo-nitzschia pungens* are especially damaging, which can not only cause ecological disasters but also release toxins that threaten human health. This study aims to construct mitochondrial genomes of several *Pseudo-nitzschia*. By analyzing their genetic diversity, high-resolution molecular markers that can effectively distinguish between different species and within species strains of them can be found, especially for the identification and tracking of *P. pungens*. The whole genome of 7 *P. pungens*; 2 *P. seriata* and 1 suspected *P. multiseries* have assembled. Their mitochondrion has sieved from the assembly results; one strain has annotated per species. Comparative analysis of 10 strains, together with 3 strains from the public database, has carried out. The indel sites, insertion hot spots, and single nucleotide variant sites of

the 13 strains have detected, and the high variation regions in the genome have found. Based on these high variation regions, the screening and design of specific molecular markers have discussed; the results provide crux information for the detection of *Pseudo-nitzschia*.
Keywords: ***Pseudo-nitzschia*; mitochondrion; genome assembly and annotation; genomic variations**

目　录

1 前言

拟菱形藻(*Pseudo-nitzschia*)是海洋浮游植物的典型组分，在全球各处均有[1]。截止 2020 年有 30 个拟菱形藻属物种被认为是有害藻华(harmful algal blooms, HABs)物种[11]；至少 12 个拟菱形藻属物种被证实可产神经性毒素多莫酸(domoic acid, DA)，而且拟菱形藻 DA 在加拿大东部有经贝类积累致人中毒的记录[8]。其中的尖刺拟菱形藻(*Pseudo-nitzschia pungens*)也产 DA，它是 HAB 物种，造成许多的生态和公共健康危害[1,8,10]。

拟菱形藻与菱形藻(*Nitzschia*)、拟脆杆藻(*Fragilariopsis*)间形态高度相似：拟菱形藻在分类学史上曾于 1899 年被归于菱形藻属，二十世纪初独立为拟菱形藻属，又于 1958 年再次归入菱形藻属，并最终于 1994 年再次恢复为单独的拟菱形藻属[9]。有研究表明在 18 个月、100 公里的时空尺度下使用微卫星标记(microsatellite markers)研究德国北海的尖刺拟菱形藻种群，发现仅有微弱的遗传分化(genetic differentiation)，显著的 FST 值在 0.0018 到 0.0389 之间，被认为是大尺度上无结构的单一类群[2]。另外，HAB 物种尖刺拟菱形藻在我国沿海、世界多处均存在亚种，我国存在未检测到 DA 的尖刺拟菱形藻亚种；尖刺拟菱形藻亚种之间遗传物质上相近，但外观形态上可能存在较突出的差异[5,16]；从分子标记层面(ITS, *rbcL*)可证实这些亚种之间可以发生自然杂交[3]。而其他的拟菱形藻属物种及其可能的亚种也被证实具有较高遗传多样性[6,7]。

可依此认为，拟菱形藻属物种如尖刺拟菱形藻，通常株系关系复杂，通过形态学、微卫星标记或单一分子标记中任意一种方法，难以准确鉴别物种或亚种/种内株系，所以通过生物信息学手段进行比较基因组学研究、开发新的高分辨率分子标记(molecular markers)对它们的分类、进化和生物地理分布研究或将十分重要。

而为了研究拟菱形藻尤其是尖刺拟菱形藻 HAB 的暴发机理，需要利用适当的分子标记跟踪其地理分布规律，及其在 HAB 爆发过程种的动态变化规律。真核生物通用分子标记（例如 18S、28S rDNA, ITS）不能充分区分尖刺拟菱形藻的遗传株系;甚至可能无法有效区分尖刺拟菱形藻和其他拟菱形藻物种如多列拟菱形藻(*P. multiseries*)[17,4]。而常用鉴定硅藻物种的分子标记 *cox1*，也因发现许多硅藻 *cox1* 内部存在内含子而变得难以被广泛使用[12,21]。

本研究对多个不同的拟菱形藻样本进行基因组测序，组装和比较分析，根据线粒体基因组序列的基因组多样性（genomic variations, GVs），开发能够用于区分不同尖刺拟菱形藻遗传株系的高分辨率分子标记，并将其用于不同海域的尖刺拟菱形藻分型鉴定。这一工作将有助于进一步研究 HAB 物种尖刺拟菱形藻的种系分布和进化。

拟菱形藻属是硅藻门(Bacillariophyta)。硅藻的线粒体基因组包含 34 个核心基因，其中包括 32 个蛋白质编码基因[15]：*atp6,8,9*; *cob;cox1,2,3*; *nad1-7,4L,9,11*; *rpl2,5,6,14,16*; *rps3,4,8,10,11,13,14,19*; *tatA,C*; 核糖体 rDNA：*rns, rnl*。部分硅藻线粒体的 *nad11*、*rpl10*、*rps2*、*rps7*、*rps12*、*rrn5* 整体缺失或部分缺失或拷贝数存在差异。这些硅藻线粒体基因组在 tRNAs 的总数量、种类、拷贝数上也存在差异；总数量最多为 26 个，最少为 23 个（表 1）。在这些硅藻基因组内部，部分存在内含子序列，越长的线粒体基因组可能存在的内含子越多。另外，公共数据库的硅藻线粒体基因组中，发现部分物种存在不同类型的重复序列，也有物种不存在；而硅藻与甲藻的三次内共生也可能对硅藻的线粒体基因组有影响[18,19,20]。这些情况说明在硅藻线粒体基因组中存在大量可深

入发掘的变异位点，或将有利于开发高分辨率分子标记。

Class	Species	Accession number	Structure	Size (bp)	A+T (%)	References
Bacillariophyceae (24)	*Nitzschia palea*	AP018512	linear?	>36,830	69.49	Kamikawa et al. (2018)
	Nitzschia sp.	MG182051	circular	36,012	71.18	Guillory et al. (2018)
	Nitzschia sp.	AP018510	circular	35,897	70.87	Kamikawa et al. (2018)
	Nitzschia sp.	AP018509	circular	37,792	69.85	Kamikawa et al. (2018)
	Nitzschia sp.	AP018507	circular	38,056	69.56	Kamikawa et al. (2018)
	Nitzschia sp.	AP018505	linear?	>35,839	69.98	Kamikawa et al. (2018)
	Nitzschia palea (Npa)	MH297491	circular	37,754	69.11	Crowell et al. (2018)
	Nitzschia alba (Nal)	MF997422	circular	36,252	71.57	Pogoda et al. (2018)
	Pseudo-nitzschia multiseries (Pmu)	KR149143	circular	46,283	68.95	Yuan et al. (2015)
	Fistulifera solaris (Fso)	KT363689	circular	39,476	71.86	Tang and Bi (2015)
	Berkeleya fennica (Bfe)	KM886611	circular	35,509	70.28	An et al. (2014)
	Cylindrotheca closterium (Ccl)	MG271845	circular	37,784	67.93	Guillory et al. (2018)
	Halamphora coffeaeformis (Hco)	MF997420	circular	44,653	67.09	Pogoda et al. (2018)
	Halamphora calidilacuna (Hca)	MF997424	circular	103,605	68.79	Pogoda et al. (2018)
	Navicula ramosissima (Nra)	KX343079	circular	48,652	68.89	An et al. (2016)
	Surirella sp. (Ssp)	MF997423	circular	42,867	72.58	Pogoda et al. (2018)
	Entomoneis sp. (Esp)	MF997419	circular	36,078	72.20	Pogoda et al. (2018)
	Eunotia naegelii (Ena)	MG271846	circular	48,049	72.92	Guillory et al. (2018)
	Phaeodactylum tricornutum (Ptr)	HQ840789	circular	77,356	64.99	Oudot-Le Secq and Green (2011)
	Didymosphenia geminata (Dge)	KX889125	circular	37,765	73.07	Aunins et al. (2018)
	Synedra acus (Sac)	GU002153	circular	46,657	68.22	Ravin et al. (2010)
	Asterionella formosa (Afo)	KY021079	circular	61,877	73.38	Villain et al. (2017)
	Proschkinia sp. (Psp)	MH800316	circular	48,863	70.39	Gastineau et al. (2019)
	Haslea nusantara (Hnu)	MH681882	circular	36,288	70.76	Prasetiya et al. (2019)
Mediophyceae (2)	*Psammoneis japonica* (Pja)	MG148339	circular	73,622	69.19	Guillory et al. (2018)
	Toxarium undulatum (Tun)	MG271847	circular	40,429	69.87	Guillory et al. (2018)
Coscinodiscophyceae (3)	*Melosira undulata* (Mun)	MF997421	circular	32,777	78.39	Pogoda et al. (2018)
	Thalassiosira pseudonana (Tps)	DQ186202	circular	43,827	69.89	Armbrust et al. (2004)
	Skeletonema marinoi (Sma)	KT874463	circular	38,515	70.27	An et al. (2015)
Diatom endosymbiont (2)	*Kryptoperidinium foliaceum* (EKfo)	JN378734	circular?	>39,686	67.59	Imanian et al. (2012)
	Durinskia baltica (EDba)	JN378735	circular?	>35,505	68.98	Imanian et al. (2012)

表 1 硅藻线粒体基因组情况统计[13,14]

通常，藻类不同株系的线粒体较叶绿体有更大的遗传差异；有更多的变异位点；而且线粒体基因组较其长度更短，在基因组构建和分析中，工作量一般更小（除非遇到较多短重复序列）。传统的真核生物通用分子标记分辨率针对硅藻可能不足，难以识别种属内株系差异，不利于对株系进行溯源和生物地理学研究；因此针对线粒体上的特定序列或将可以设计出具有更高分辨率、更适合低阶元分析的分子标记或多重分子标记，依据这些分子标记，将更有利于研究者研究拟菱形藻、尖刺拟菱形藻的来源以及系统发育，进而为拟菱形藻藻华、赤潮的预防和控制提供关键信息。

2 材料与方法

2.1 基因组 DNA 提取与测序

本研究采用全基因组测序、组装思路获得线粒体基因组（图 1）。选择二代测序 Illumina 测序技术[22]，深度 50×。原因是拟菱形藻是硅藻，基因组较小，通常在 100 Mb 以内[8]，故构建其基因组所需的测序深度以及所需数据量不高，进行全基因组测序也可以同时获得除线粒体基因组外更多其他有价值的数据。

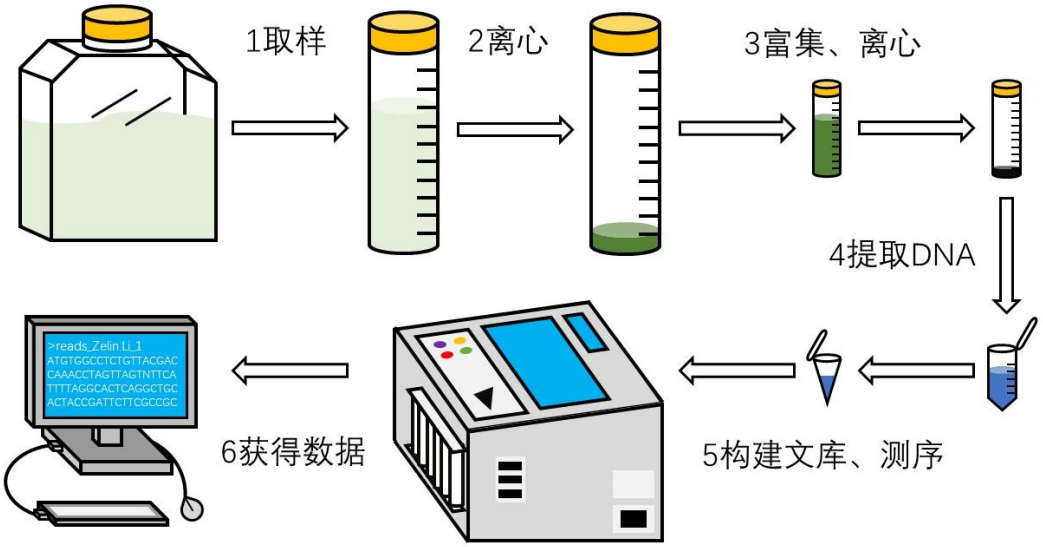

图 1 微型藻类全基因组 DNA 提取、富集、测序流程

2.2 测序数据预处理与组装

本研究具体测序方法是 Illumina 双端测序(paired-end sequencing)，其获得的单一 read 长度为 150 bp(可能有 1-2 bp 的长度差异)[22]。每对 reads 都对应测序前构建测序文库时被打断的一条短序列的两端;分别存储在 2 个文件或压缩包中。。从测序公司返还的每个样本的测序数据有 2 个 FASTQ 文件，均为已经去除接头并经过质量控制的纯净 reads 数据(clean data)。一个存放所有 reads1（双端的一端），另一个存放所有 reads2（双端的另一端)。

因尚未有尖刺拟菱形藻线粒体基因组发表，故无法通过近源物种进行筛选线粒体相关 reads，然后筛选组装。所以只能以全基因组组装(Whole genome de novo assembly)流程处理。组装前，从多个组装软件性能和组装效果中进行挑选了适用于微生物和细胞器基因组组装的[23,24,25]3 个特长各异的软件进行组装：ABySS(2.2.4)[26]、Platanus_allee(2.2.2)[27]、SPAdes(3.14.0)[28]。

2.3 筛选线粒体相关 scaffolds

得到全基因组组装结果后，将每个样本的组装产物单独建立本地 BLAST(Basic Local Alignment Search Tool[30]，2.10.0+)库。已公开线粒体基因组且与尖刺拟菱形藻最近源的物种是多列拟菱形藻(*Pseudo-nitzschia multiseries*)，使用它的线粒体基因组(KR149143.1，即 NC_027265)作为 BLAST query，分别对各个样本组装产物的本地 BLAST 库进行 blastn[29]，找到组装结果当中可能为拟菱形藻线粒体基因组的 scaffolds（下文称为线粒体相关 scaffold(s))。

线粒体相关 scaffold(s)可能唯一，也可能多个，其筛选的依据是将 format 6 的 BLAST 比对结果中的高比值片段对(High Scoring segment Pair, HSP)以 bit score 从高到低排列，从第一条 scaffold 向下依次查看，选择长度比第一条 scaffold 小但又大于 1000 bp 的所有 scaffolds，直到第一条小于 1000 bp 的 scaffold 为止。后把所有线粒体相关 scaffolds 从总的组装结果中摘出即可；依据比对情况所展示的 q. start、q. end(q 代表 query)作 CIRCOS(0.69-8, Perl

v5.16.3)[32]图。这些 CIRCOS 图可以展示代表线粒体基因组的 scaffolds 的 HSP 在参考基因组上的分布状况，便于环化、构建完整线粒体基因组（详细组装情况、更多 CIRCOS 图见附 1）。

以后续研究中完成环化和注释的尖刺拟菱形藻株系 CNS00141 为例：多列拟菱形藻线粒体基因组 KR149143.1 为比对的参考基因组，CNS00141 被三个组装软件分别组装所得的线粒体相关 scaffolds 与此参考基因组的比对情况如图 2 所示。

图 2 CNS00141 使用三个软件组装出线粒体有关 scaffolds 覆盖多列拟菱形藻参考线粒体基因组的情况

这些图最外圈的灰色刻度圈代表多列拟菱形藻线粒体，用于 scaffolds 的定位，内圈的有颜色的圈，表示 BLAST 结果当中每个 HSP 的范围，不同 scaffold 的 HSP 以不同颜色体现，若只有 1 个 scaffold 则颜色唯一。内部的第一段文字是自动比对筛选线粒体有关 scaffolds 的情况，从上向下则是表示脚本 assembly+在执行过程中求出最近源的公开线粒体基因组硅藻 NCBI Accession 号、该线粒体长度(bp)、筛选组装结果所得线粒体有关 scaffolds 数、这些 scaffolds 长度之和。而下半段是由 n50.sh（见附 2）统计的该全基因组组装状况，从上到下分别是组

装产物名、组装产生 scaffolds 序列总数、组装产物大小、最长 scaffold 长度(bp)、N50、N90（N50 表示组装所得 scaffolds/contigs 从大到小排列，依次相加，积累到全长的 50%时，记录该 scaffold/contig 长度为 N50，N90 则为积累到全长的 90%）。

而上述步骤（组装软件组装、BLAST 筛选线粒体基因组、CIRCOS 可视化作图）均被笔者整合为一个生物信息学分析流程脚本集 assembly+（图 3 展示其数据流、信息流、脚本/IO 关系，具体脚本见附 2），可以自动化完成，方便更多类似的分析和研究。

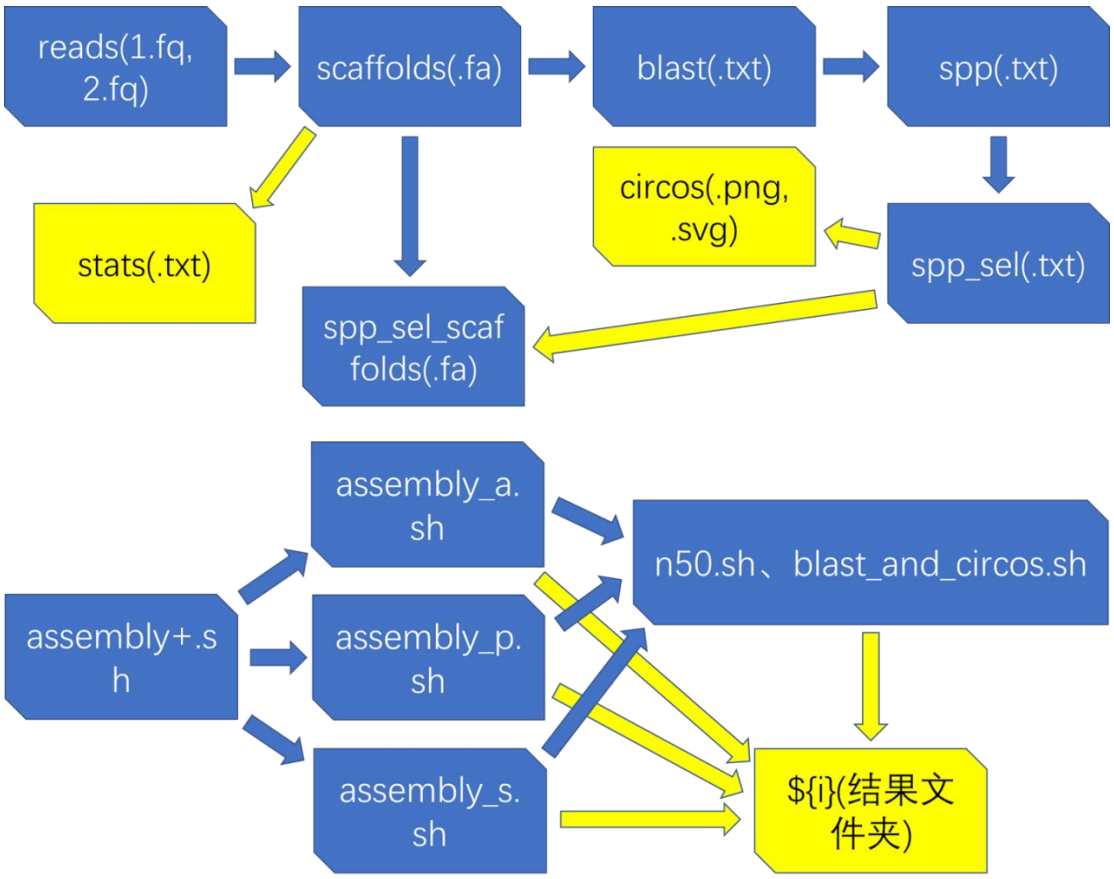

图 3 组装、线粒体筛选、CIRCOS 作图脚本集 assembly+的数据流、信息流、脚本/IO 关系

图的上半部分表示脚本集的数据存储转化迁移，最终得到对后续分析有用的信息如 CIRCOS 图存储于黄色标记框中（黄色框代表此项结果只包含信息，不包含数据，在其基础上不可以继续进行处理），而最终筛选所得相关 scaffolds 结果存储于 spp_sel_scaffolds(.fa)中。图的下半部分表示 assembly+包含的各个脚本之间的运行关系，而黄色框表示脚本运行结果而非脚本本身。

2.4 线粒体相关 scaffolds 的连接与环化

然后需要将三个软件所组装出的代表线粒体基因组的 scaffolds 合并到同一 FASTA 文件建立本地 BLAST 库，并以自身为 query 对此库进行 blastn，通过所写的脚本 selfblast.sh（附 3）挑选相似度为 100%的 HSP，若该 HSP 的两序列是同一 scaffold 的首尾，则认为该 scaffold 已经通过组装完成环化，去除其重复段完成环化；若该 HSP 的两序列是不同 scaffold 的末端，就可以认为这两条

scaffolds 相连接；而去除其重复段进行连接后，得到的新序列视为两个 scaffolds 的并集（∪），此方法的原理如图 4。

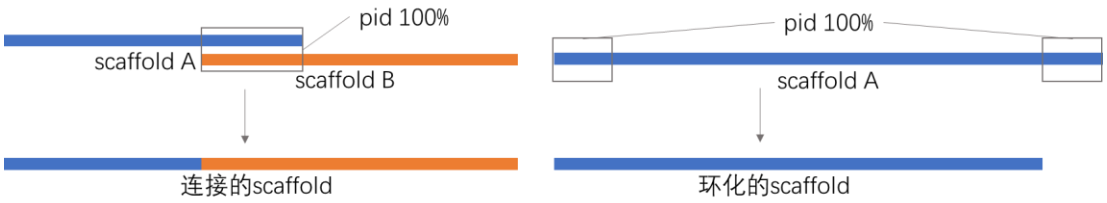

图 4 selfblast.sh 脚本判定 scaffolds 连接/环化原理

BLAST 时，若两条 scaffold 末端局部 pid (percentage identity)为 100%或一条 scaffold 自身首尾 pid 为 100%，则实现连接或环化。

2.5 添加“人工 N 区”

然而并不是所有的组装结果都能产生可以相连甚至环化的 scaffolds。大多数情况下，组装所得线粒体有关 scaffolds 不能通过上述方法完成环化；其原因可能是组装出的序列末端出现短重复片段，导致组装软件无法判定 scaffold 是否组装到序列尽头，故 scaffold 会在此处停止延长。

这样的区域通过 selfblast.sh 无法环化或连接。在这些不能完成连接或环化的 scaffolds 中间，视为存在未填充序列的 gaps。为补上这些 gaps 需要在两个 scaffolds 中间加 N；通过 BWA(Burrows-Wheeler Aligner)[33]比对时尝试通过 reads 覆盖补上 N 区。为了区分人工添加的 N 区和组装软件产生的 N 区，固定“人工 N 区”的 N 数量为 5（组装软件组装所得 scaffold 有 N 区则说明这部分序列被组装软件视为不确定区域，这种不确定通常也是重复序列所致，软件难以确定重复序列的重复次数/长度，但如果有一对 reads 分别在不确定区域两侧，则可以估算出此不确定区域的长度，据此长度组装软件便自发产生 N 区。而“人工 N 区”需要位于 scaffold 断点处，这一区域组装软件算法无法判断不确定区域长度或序列是否终止，所以没有形成 N 区）。

细胞器 DNA 多为环状，这意味着如果细胞器的 scaffolds 连接序列或单一 scaffold 的首尾未能直接环化，则首尾之间存在 gap，需要添加“人工 N 区”，添加这两类“人工 N 区”的方法如图 5。

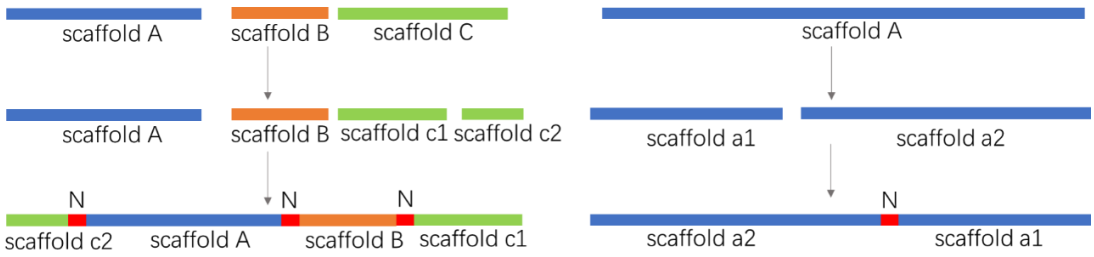

图 5 环化细胞器基因组添加“人工 N 区”方法

左侧意为若某细胞器有 3 条相关 scaffolds（其排列顺序为 A-B-C），那除了在 AB、BC 间添加人工 N 区外，还需要将 A 或 C 从中断开，将一段保留原位另一段补至序列另一侧，在另一段与该侧序列间添加“人工 N 区”。右侧则表示细胞器基因组有 1 条相关 scaffold，需要将其从中断开，两段位置调换，在中间添加“人工 N 区”。

2.6 填补 N 区与质量检验

完成连接或环化及人工 N 区的添加后，需要对先前得到的序列（无论是否有无 N 区）进行质量检验，对于含有 N 区的序列，质量检验的同时还可以同时进行补 N 操作。

质量检验的具体方法是将组装所用的全部纯净 reads 数据以要质量检验的序列为参考（bwa index 序列名.fa）使用 BWA (bwa mem[33])进行比对，然后通过 SAMtools（Sequence Alignment Map tools）[34]提取双端 reads 均比对到质量检验序列上的 reads 并进行排序后输出为 BAM(Binary Alignment Map)格式；将所得 BAM 数据排序、建索引[34];然后以 IGV(Integrative Genomics Viewer)打开此比对排列结果以查看比对情况[36,37]。或用 SAMtools 输出全部位点 reads 覆盖情况；用 VarScan(2.4.4）输出低支持率位点的统计信息[35,38]。

若某些位点支持率极低，大部分 reads 的碱基与质量检验序列的碱基不符或是序列插入/缺失，则应该判断该位点是否满足组装错误的几个特征：

1、质量检验序列此位点碱基支持率低于 25%，即 75%的 reads 此位点碱基与质量检验序列不同；

2、此位点附近区域的 reads 深度与该序列 reads 的平均深度相近（如果此处深度明显高于平均深度，则认为此处发生 reads 的非特异性比对，是核基因组或因样本不纯导致的其他物种的保守 reads 比对到此处）；

3、90%乃至 97%以上的 reads 的碱基一致性地支持另一个非质量检验序列此位点的碱基（如果是多个不同碱基分别占据可观不等的支持率，则此位点可能是多态性位点而非组装错误）。

依次满足上述三个条件则认为该位点是组装错误，需要对此位点进行修正，分三种情况：

1、改为大部分 reads 所支持的碱基；

2、插入 reads 所支持的序列；

3、删除 reads 此处缺失的碱基。

补 N 依赖 reads 在 N 区及其两侧的覆盖状态，以下情形可完成补 N：

1、存在同一条 read 同时匹配到 N 区两侧，则该 read 中部序列相对于质量检验序列是插入，该插入序列即为 N 区实际的序列，将该序列取代 N，即完成该 N 区的补 N 操作；

2、没有 read 同时匹配到 N 区两侧，但 N 区两侧 reads 呈断裂状，比对到的部分均未达 150 bp；它们在 N 区内存在 soft clip，此时可以提取 N 区前后 100 bp 左右比对到的所有 reads，用 SAMtools 将它们筛出（samtools view 序列名.sort.bam 序列名:起始位置-终止位置 | cut -f1 | sort | uniq > IDs.txt），并将它们进行此小范围的局部组装，如果这两侧的 reads 的 soft clip 存在重叠，则组装软件在这少量 reads 的局部组装过程中可以将 N 区两侧 reads 连接(组装）成 contig。如果组装所得 contigs 可以同时匹配 N 区两侧的序列（使用 MUSCLE 进行全局比对[39]），则 N 区实际序列即为该 contig 中部未匹配到两侧的序列，将该序列取代 N，即完成了该 N 区的补 N 操作。通过 IGV 查看补 N 前后的质量检验序列，reads 与质量检验序列的比对形式体现，如图 6。

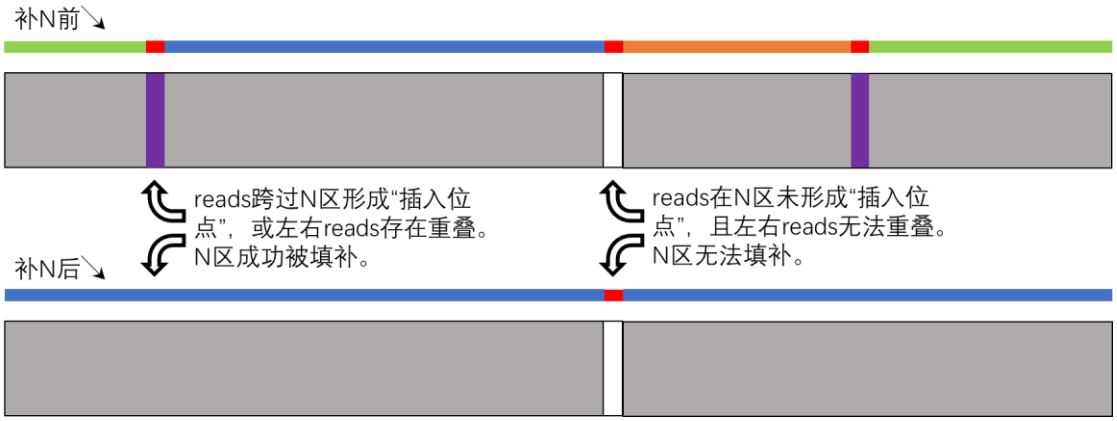

图 6 reads 与质量检验序列比对补 N 前后变化

此图以类 IGV 形式示意 reads 与质量检验序列的比对，细条为质量检验序列，其颜色表示该序列原先所属部分，红色表示 N 区，绿色、蓝色、橙色表示三个不同的 scaffold。下方粗矩形条表示 reads，灰色表示 reads 覆盖且位点支持率良好，紫色表示此区域/位点的 reads 的碱基与质量检验序列碱基不同（可以是碱基差异/插入/缺失任意一种情况），白色表示此处 reads 未覆盖质量检验序列。

序列质量检验、补 N 后，通常情况下仍存在 N 区；为填补这些 N 区需要局部 PCR：两侧引物扩增出的片段可重叠，则视为完成此 gap/N 区已完成验证、成功被补上。但由于本次实验条件所限，未对未能补 N 序列剩余的 N 区进行 PCR 扩增验证。因此本研究直接挑选已经环化或较为完整(N 区较短或覆盖参考基因组 90%以上）的株系的线粒体基因组序列进行后续分析。

2.7 共线性分析

为了便于后续比较分析和注释，需要统一所有线粒体基因组的起始位置与方向，在本次研究当中，所有序列的方向均将 *cox1* 作为正向，起始位置均设置为 *cox1* 的起始密码子 ATG。

调整完成后，可对所有序列进行共线性分析，共线性分析目的在于研究多序列之间是否存在大尺度的结构重排，而不考察其内部细微的缺失或插入变异情况，同时也可检查序列之间是否存在倒位或其他重组现象，这一步分析使用软件为 Mauve(progressiveMauve –output=输出名.xmfa 多序列.fa)[40]。

2.8 构建系统发育树

完成共线性分析后，对全部研究当中用到的线粒体基因组全长进行全局比对（Global Alignment)。鉴于序列较大，比对用 mafft 完成。为保证建树较快速的同时准确，比对结果采用邻接法（Neighbor Joining）设置 bootstrap 拷贝数为 500、Gamma distributed 等于 5；构建系统发育树[42]，从而对 13 个序列的进化关系和遗传距离远近做出直观评估。

2.9 基因组注释

为准确定位变异位点密集部位属于什么基因或基因间区，就要完成基因组注释(Genome annotation)，判断各个蛋白编码基因、rRNA 基因、基因间区在特定株系基因组的位置。注释有两种方式：

1、 从头注释。首先需将得到的线粒体基因组序列通过 tRNAscan-SE[43]寻找 tRNA，并通过相近物种的 rDNA（*rns*、*rnl*)序列使用 blast 定位寻找并注释该株系 rRNA 基因；然后在 tRNA、rRNA 基因的间区通过 ORF(Open Reading Frame) finder 寻找潜在的蛋白编码基因；使用 smartblast（一种 blastp）[44,29]注释它们，最后将这些基因的注释信息转换为 genbank 格式，最后将得到的 genbank 文件通过 OGDRAW 绘制线粒体基因组注释圈图[45]。

2、 同源注释。如果在公共数据库如 NCBI 上已有完成注释的相同或极近源物种基因组参考序列；可使用参考序列在 GeSeq[46]下对线粒体基因组进行同源注释，输入同源物种线粒体基因组 genbank 格式文件及待注释 FASTA 格式序列，以 genbank 格式输出该序列注释情况后，作图方式同 1。

2.10 基因组多样性分析

完成注释后，为了发掘高分辨率分子标记或研究不同基因的变异速率、保守性差异。在一般的重测序流程中，是使用已完成注释的参考基因组作为 BWA 比对的参考序列，然后将重测序纯净 reads 数据与其比对，然后使用 SAMtools 等软件对比对结果进行过滤筛选得到不同重测序株系的单核苷酸差异位点和短的缺失、插入位点。这一方法较为成熟并且使用广泛，但这样做，对于已经完成组装与注释的完整序列来说，并不是最佳比较分析方式，因为其会产生非特异性比对，并且在基因组差异较大的情况下（比如不同物种），由于单一 read 本身过短，会出现部分片段相似性过低因而无法比对到 reads 的情况，不利于基因组的精确比较分析。

为了改善这种精确性不足的分析方式，笔者开发了一套适用于使用细胞器基因组全局比对结果进行基因组差异分析的系列程序，其以速度较快且程序健壮性强的 C 语言[47]编写，主要包括 4 个模块化程序（详见附 4）：

1、 提取 gaps 位置的程序（设计原理如图 7 所示）；
2、 合并全部株系 gaps 位置后识别这些株系插入热点位置的程序；
3、 删除全局比对结果当中特定株系的全部 gaps 位置的程序；
4、 提取删除 gaps 位置后剩余比对中其他株系与该特定株系碱基不同的位点（单核苷酸差异位点）位置的程序。

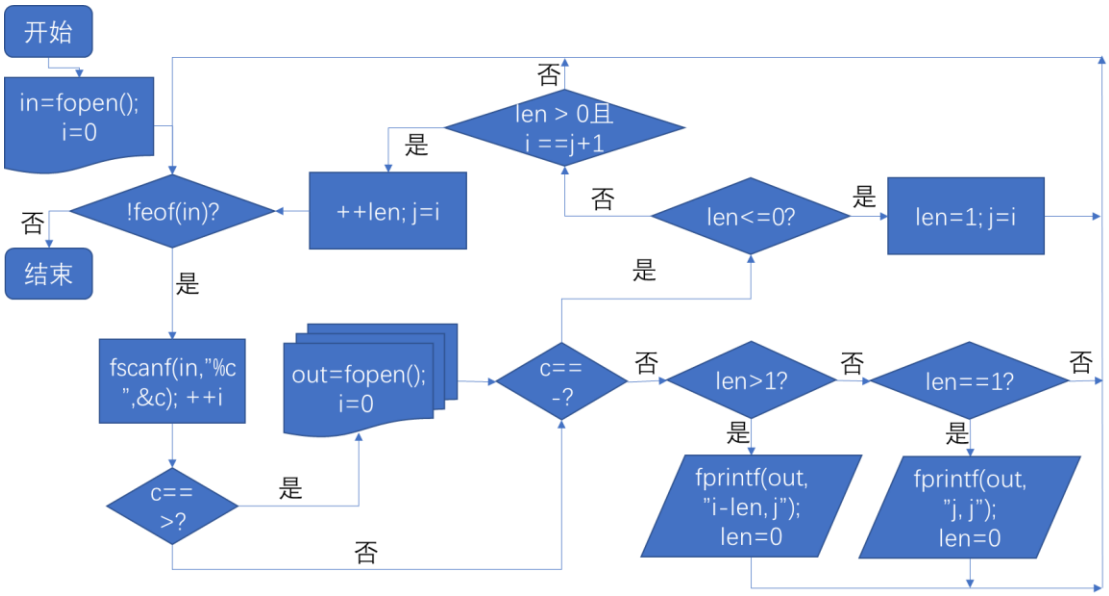

图 7 从全局比对中提取 gaps 位置程序原理

除此之外，另外 3 个程序采用类似的数据逻辑对全局比对结果进行处理，它们的共同特征是识别多序列比对当中的特定字符并记录它们的延伸范围，输出为位置信息，然后选择性地处理多序列比对数据本身。

经程序 1 处理后，得到的是每个株系的 gaps 在全局比对结果的位置，并非 gaps 在每个株系的绝对位置。这些位置的参考可视为属于一条“全部比对株系序列的并集序列”，它并不真实存在，但它的长度即为全局比对结果的长度。

程序 2 可以得到全部株系长度大于 10 bp 的 gaps 的位置的并集并将相邻距离小于 100 bp 的 gaps 合并，并将这一区域识别为插入热点。

程序 3 是程序 4 的前体，目的在于得到其他株系仅比对到选定的某一参考基因组的序列，这一过程避免了参考基因组 gaps 的干扰，利于后续程序 4 识别出各个株系相对参考基因组单核苷酸差异位点的位置。最终程序 4 提取的各个株系单核苷酸差异位点的位置，是所选定的参考基因组的绝对位置，参考基因组本身在单核苷酸差异位点检测时将不会被检测出任何差异位点。

得到的所有株系线粒体基因组 gaps 位置数据，全部录入一个 CIRCOS 图（全部株系的 gaps 位置以线为标识记录在该图内侧）而插入热点在全局比对上的位置，也被录入该图的最外圈以白色区域标识。

基因组单核苷酸差异分析选定的参考基因组为 CNS00141 这一尖刺拟菱形藻线粒体基因组，以其为坐标录入所有株系的单核苷酸差异位点位置形成另一 CIRCOS 图：全部株系相对 CNS00141 的单核苷酸差异位点位置以线为标识记录在该图内侧，另外，为了分析单核苷酸差异密集区和其在基因组上的密度分布，笔者合并 13 个株系的单核苷酸差异位点位置数据，另行写了一个针对该数据进行滑动窗口(sliding window)分析的 C 程序，设定滑动窗口大小为 500 bp，以窗口中心碱基为坐标点，窗口间距 1 bp（最高密度分辨率），根据此结果得到综合 13 个株系线粒体基因组单核苷酸差异位点在 CNS00141 参考坐标上的总密度分布，将每个滑动窗口的单核苷酸差异位点值以连续柱状图绘制在该 CIRCOS 图的最外圈，以其值对数变换（以 0.5 为底）后的值作热力图置于该图次外圈；这两个图配置见附 5。

3 结果

3.1 组装情况

本次研究所用数据是 11 个单细胞拟菱形藻株系的全基因组测序数据；通过全长 18S rDNA 分子标记的相似性注释物种，表明这 11 个单细胞硅藻株系中 9 个为尖刺拟菱形藻(*P. pungens*)，2 个为连续拟菱形藻(*P. seriata*)。而在三个软件的全基因组组装结果筛选线粒体有关 scaffolds 后，获取各株系线粒体相关 scaffolds，统计其中可用于后续连接、环化、补 N 的 scaffolds 数量，记录了它们的连接方式（以便后续追溯问题），统计完成质量检验和补 N 的序列的长度、N 区长度（表 2）。

海域	株系	18S 分型	相关 scaffolds 数	scaffolds 连接方式	N 区长度	序列长度(bp)
胶州湾	CNS00055	*P. pungens*	1	P	5	38535

东海	CNS00089	*P. pungens*	1	S	100,5	38281
渤黄海	CNS00110	*P. pungens*	2	A∪S	50,5	39368
渤海	CNS00141	*P. pungens*	2	P∪S	1	40117
渤海	CNS00153	*P. pungens*	2	S∪S	5	37276
渤海	CNS00154	*P. pungens*	1	S	100,5	39475
渤海	CNS00155	*P. pungens*	5	A+A+A∪P∪A	3*5,4*未计	32855
渤海	CNS00156	*P. pungens*	1	S	5	39588
渤海	CNS00158	*P. pungens*	1	P	0	45219
西太平洋	CNS00090	*P. seriata*	1	S	0	37526
西太平洋	CNS00097	*P. seriata*	1	S	5	37084

表 2 组装、质量检验、环化后序列情况

其中相关 scaffolds 数表示每个株系全基因组组装筛选所得的最终用于连接环化/补 N 的 scaffolds 数量，这些 scaffolds 可能源于不同软件；而 scaffolds 连接方式表示“scaffolds 数量”中的 scaffold 各源于什么软件，它们又是以何种形式连接，其中的加号“+”表示两个 scaffolds 在参考基因组上无重叠，而且互相比对时它们之间也没有互相重叠，视其间存在 gaps，以补 N 形式连接，而并集“∪”表示两个 scaffolds 在互相比对时发现首尾 100%重叠部分，根据序列方向和在参考上的排布位置将它们的重复部分删除一段并将二者连接。N 区长度表示一个连接完成的株系序列中仍存在的 N 区大小和数量（N 区长度间以逗号“,”分开），其中长度为 5 的 N 区均为“人工 N 区”。序列长度指完成连接/环化、序列添加“人工 N 区”和质量检验、补 N 后得到的株系序列长度。

在 11 个序列中，CNS00158 虽然 18S 分型为尖刺拟菱形藻，但实际得到的线粒体序列长度达到 45 kb，远超其他尖刺拟菱形藻株系线粒体基因组长度。这或许表明 CNS00158 在进化地位上与其他尖刺拟菱形藻株系不同。

上述 11 个完成质量检验和计算方法补 N 的线粒体基因组序列，有 10 条完成环化或较为完整，其中 CNS00155 得到最终序列长度与预期实际线粒体基因组长度相差较大，比较不完整，故不将其纳入后续比较分析。本研究另外加入了 3 个公共数据库序列线粒体基因组；它们是 2 个菱形藻属物种：*Nitzschia alba* (NC_037729, 36252 bp)[12]、*Nitzschia palea* (MH297491, 37754 bp)[48]以及 1 个多列拟菱形藻（*Pseudo-nitzschia multiseries*, NC_027265, 46283 bp）[31]。

3.2 共线性

根据 13 个序列共线性的结果（图 8, 详细 xmfa 结果见附 8），可以认为大尺度上，虽然这些株系长度各异，相似度有的高，有的不够高，但具有良好的共线性：13 个线粒体基因组内也没有任何一段序列被识别为与其他株系方向相反或位置发生调换。

据此可以认为菱形藻、拟菱形藻的线粒体基因组在整体结构上较相似，内部可能存在不同程度的变异、插入、缺失事件，但几乎没有发生重组，因而认为它们的遗传关系可能较近。

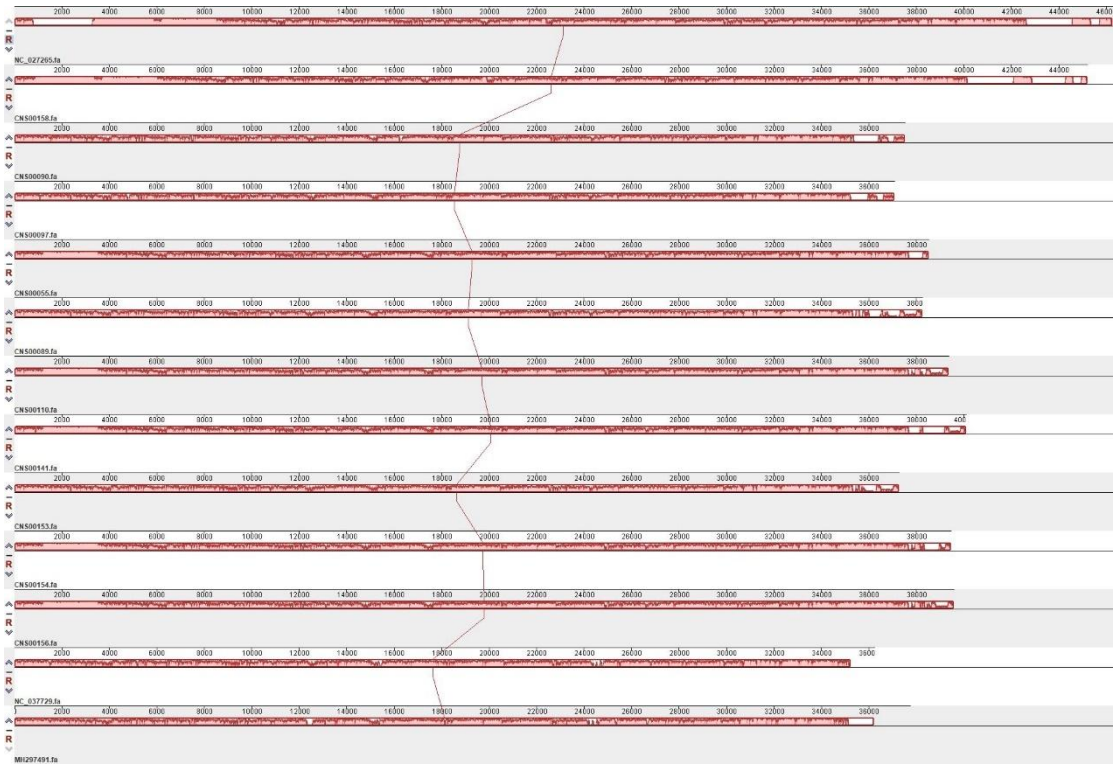

图 8 13 个线粒体基因组序列共线性分析

序列共线性则以相同颜色的色块表示；如果序列之间存在结构变异，则会产生多个分布位置不同的色块，图中的 Mauve 比对结果 13 个株系均为相同单一红色色块，它们的内部虽然各含大量其他类型的变异，但彼此之间仍呈现极佳的共线性。

3.3 演化关系

在确定共线性极佳的基础之上，对 13 个线粒体基因组序列进行全局比对后进行系统发育分析，以评估株系之间的相似性和遗传距离，最终得到基于全线粒体基因组序列构建的系统发育树（图 9）。根据该树，可大致分为 5 个类群，第一大类群包含了大部分本次研究中用到的尖刺拟菱形藻株系（CNS00110、CNS00153、CNS00089、CNS00156、CNS00154、CNS00055、CNS00141），它们遗传距离较近，但也分为了两个亚群，可以认为是尖刺拟菱形藻的两个类群，二者是否可以被认为是 2 个亚种，有待进一步验证。

而 18S rDNA 分型也为尖刺拟菱形藻的株系 CNS00158 在此树中与 NCBI 的多列拟菱形藻参考基因组（NC_027265）遗传距离极近，加上其长度亦接近多列拟菱形藻，故认为该株系可能是一种特殊的杂交——核基因组存在类似尖刺拟菱形藻的序列，而线粒体基因组类似多列拟菱形藻。

另外，2 个连续拟菱形藻株系（CNS00090、CNS00097）虽然可以认为处于同一类群，但二者也呈现出较大差异，遗传距离较远，这或许表明连续拟菱形藻具有更大的变异性，也可能表明这 2 个株系起源上存在较大差异，并也有可能是不同亚种（或不同物种）。

而作为此树外类群的 2 个菱形藻属的物种（MH297491，NC_037729）符合预期的位于遗传距离更远的进化枝上。

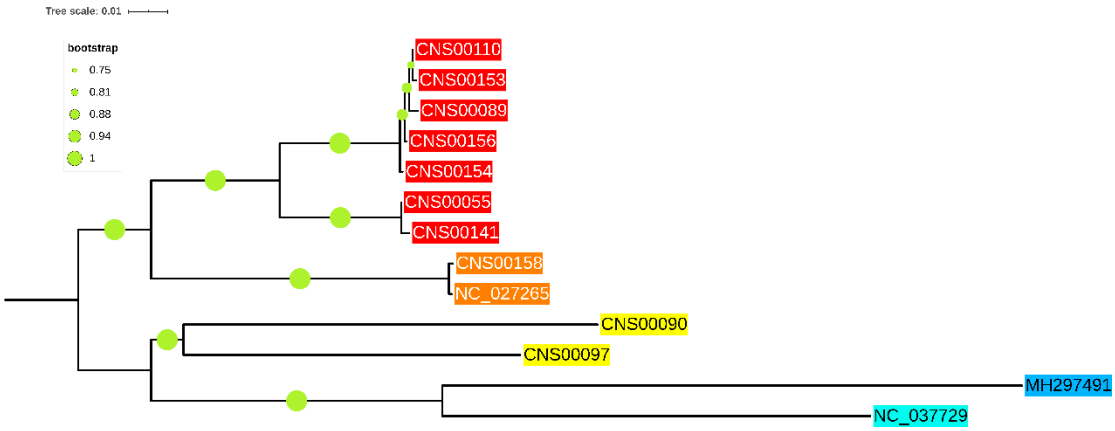

图 9 13 个线粒体基因组系统发育树

红色标签株系为尖刺拟菱形藻，橙色标签株系为多列拟菱形藻，黄色标签株系为连续拟菱形藻，蓝色标签株系为 *palea* 菱形藻，青色标签株系为 *alba* 菱形藻；图中 bootstrap 检验自展值（只显示 75%以上的）以淡绿色圆圈大小显示在进化枝中部。

3.4 线粒体基因组注释

从这 13 个线粒体基因组中选取先前已经环化或较为完整的 3 个（CNS00158-可能为多列拟菱形藻、CNS00090-连续拟菱形藻、CNS00141-尖刺拟菱形藻）兼用从头注释和同源注释两种方法进行注释得到 genbank 文件并制作圈图。同时将调整方向和序列起点后的 3 个公共数据库序列的 genbank 注释转化为圈图(图 10)，综合评估他们的基因组成情况，统计它们的 rRNA 基因、蛋白编码基因以及 tRNA 基因组成汇总至表 3、4。

从注释情况易发现，多列拟菱形藻 NC_027265 和 CNS00158 的 *cox1* 含有内含子，内含子在线粒体当中较为罕见，另一个现象是：这二者的内含子长度不同，在 NC_027265 的该内含子中包含 2 个蛋白编码基因（标注为 *orf742* 和 *orf790*），然而在 CNS00158 的该内含子中仅有 1 个蛋白编码基因（*orf742*），二者的基因组成除此之外便没有其他差异，这可能是一次插入事件；考虑到 CNS00158 株系的 18S rDNA 与线粒体基因组各自相近的物种不同（前者与尖刺拟菱形藻相近，后者与多列拟菱形藻相近），这一事件很有可能与种内杂交或种间杂交有关，而且 18S rDNA 可能本也无法区分这两个物种。

这 6 个完成注释的线粒体基因组的基因的排布顺序符合先前共线性分析得出的结论，彼此之间大尺度上没有重排，并且在序列尾部（*cox1* 之前）呈现出相同特征：存在一个较大的基因间区，该区域在部分株系当中出现了 23S rDNA 的拷贝或片段，具有这一现象的序列可能在该区域发生过插入、缺失或重组事件，并且不同株系线粒体基因组总长度具有差别，也与该区域的长度不同相关。

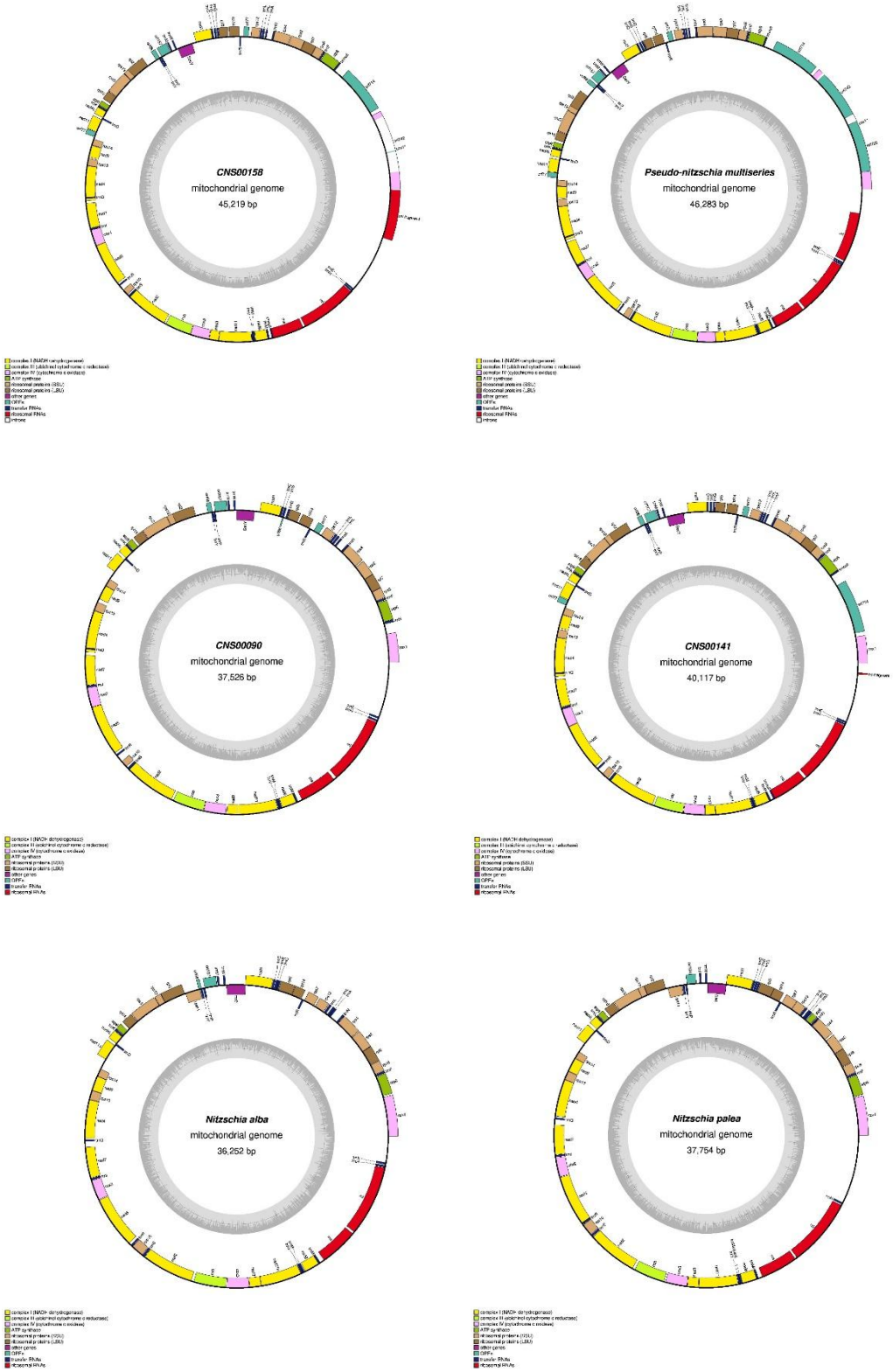

图 10 6 个线粒体基因组注释

6 个注释圈图从上到下分别是疑似多列拟菱形藻株系 CNS00158、多列拟菱形藻株系 NC_027265、连续拟菱形藻株系 CNS00090、尖刺拟菱形藻株系 CNS00141、*alba* 菱形藻株系 NC_037729、*palea* 菱形藻株系 MH297491。其中标有物种拉丁名的株系即为公开于公共数据库的株系，而 CNS 开头的株系为本研究自有株系，它们的

线粒体基因组序列长度写于图中央，注释基因方向为黑色实环外为正向，实环内为反向；按正向读的顺序为逆时针，每个序列的起点位于每张图右侧的 *cox1* 基因底部。内部的灰色峰图圆环，代表序列各段的 GC%。这 6 个图的细节和它们对应的 genbank 文件详见附 7。

分析这 6 个线粒体基因组的基因组成（表 3），发现三种拟菱形藻和 *alba* 菱形藻均缺失 *atp8*，三种拟菱形藻相对菱形藻缺失 *rpl16*、*rps7*，*rps11*、*tatC*，而两种菱形藻相对拟菱形藻缺失 *rpl17*、*SecY*。值得注意的是，无论是菱形藻还是拟菱形藻，都存在核心基因缺失，例如本研究中的 6 个线粒体中仅有 *palea* 拟菱形藻含 *atp8* 基因，全部拟菱形藻株系也均缺失 *rpl16*、*rps11*。另外，四个拟菱形藻株系也包含硅藻线粒体基因组中较少出现的基因如 *rpl17*。

另外，也可能本次注释过程发现的部分 ORF 由于没有被在基因组上被注释过，而其与其他更远源物种的直系同源基因相似性过低，因而无法通过 smartblast 找到，未能完成注释。

基因\株系	CNS00090	CNS00141	CNS00158	NC_027265	NC_037729	MH297491
atp6	1	1	1	1	1	1
atp8						1
atp9	1	1	1	1	1	1
cob	1	1	1	1	1	1
cox1	1	1	1	1	1	1
cox2	1	1	1	1	1	1
cox3	1	1	1	1	1	1
nad1	1	1	1	1	1	1
nad11(a, b)	2	2	2	2	2	2
nad2	1	1	1	1	1	1
nad3	1	1	1	1	1	1
nad4	1	1	1	1	1	1
nad4L	1	1	1	1	1	1
nad5	1	1	1	1	1	1
nad6	1	1	1	1	1	1
nad7	1	1	1	1	1	1
nad9	1	1	1	1	1	1
orf124						1
orf157	1	1	1	1	1	
orf66	2	1	1	1	1	
orf714		1	1	1		
orf742			1	1		
orf77	1	2	2	2		
orf790				1		
rnl*	1	1.1**	2	2	1	1
rns*	1	1	1	1	1	1
rpl14	1	1	1	1	1	1
rpl16	1	1	1	1	1	1
rpl2	1	1	1	1	1	1
rpl5	1	1	1	1	1	1
rpl6					1	1
rpl7	1	1	1	1		
rps10	1	1	1	1	1	1
rps11					1	1
rps12	1	1	1	1	1	1
rps13	1	1	1	1	1	1
rps14	1	1	1	1	1	1
rps19	1	1	1	1	1	1

rps2	1	1	1	1	1	1
rps3	1	1	1	1	1	1
rps4	1	1	1	1	1	1
rps7					1	1
rps8	1	1	1	1	1	1
SecY	1	1	1	1		
tatC					1	1

*rRNA 基因与蛋白编码基因与其他蛋白编码基因一同显示于此表

**CNS00141 株系的 *rnl* 存在片段(rnl-fragment)，此片段记 0.1

表 3 6 个线粒体基因组基因组成

在 tRNA 方面（表 4），*palea* 菱形藻缺失 *trnE*；但是所有株系则均有 23-24 个 tRNA 基因（CNS00090 有 24 个，其余株系均为 23 个），与大部分其他硅藻相比，该数量属于较少的。但从中也可以发现，种属关系越接近的株系的 tRNA 基因数量、类型越接近，说明在遗传演化过程中，tRNA 基因由于较小，更容易伴随线粒体基因组的插入或缺失事件而发生变化，据此可以认为其数量和种类一定程度上可以作为种或属的鉴别标志。

tRNA 基因\株系	CNS00090	CNS00141	CNS00158	NC_027265	NC_037729	MH297491
trnA	1	1	1	1	1	1
trnC	1	1	1	1	1	1
trnD	1	1	1	1	1	1
trnE	1	1	1	1	1	
trnF	1	1	1	1	1	1
trnG	1	1	1	1	1	1
trnH	2	1	1	1	1	1
trnI	1	1	1	1	1	2
trnK	1	1	1	1	1	1
trnL	1	1	1	1	1	2
trnM	3	3	3	3	3	2
trnN	1	1	1	1	1	1
trnP	1	1	1	1	1	1
trnQ	1	1	1	1	1	1
trnR	2	2	2	2	2	2
trnS	2	2	2	2	2	2
trnV	1	1	1	1	1	1
trnW	1	1	1	1	1	1
trnY	1	1	1	1	1	1

表 4 6 个线粒体基因组含有 tRNA 基因情况

3.5 基因组多样性

3.5.1 插入缺失位点

对 13 个线粒体基因组全局比对得到的 gaps 进行比较分析，定位了 13 个株系全局比对中每个株系上的 gaps 相对全局比对的位置，并计算出联合这 13 个株系的 gaps 得到的插入热点位置（图 11）。

据这一分析结果可认为拟菱形藻、菱形藻之间的插入/缺失变异集中于序列末端 2-5 kb，此位置对应先前注释结果中的序列末端基因间区。

另外，在 *cox1* 内，多列拟菱形藻 NC_027265、疑似多列拟菱形藻 CNS00158 相对其他所有株系有明显的大片段插入，体现为其他株系在比对中形成长逾

2000 bp 的 gaps。其中 NC_027265 插入了两大段序列（对应 *orf742* 和 *orf790*），而 CNS00158 仅插入一段（对应 *orf790*），这些插入事件形成了多列拟菱形藻线粒体基因组 *cox1* 的内含子。

而在全局比对 6-8.5 kb 附近，可以对应注释结果发现：多列拟菱形藻 NC_027265、疑似多列拟菱形藻 CNS00158、7 个尖刺拟菱形藻中的 5 个（CNS00141、CNS00055、CNS00154、CNS00156、CNS00110）相对其他株系存在 *orf714* 的插入（或可理解为其他株系丢失 *orf714*）。

除了这三个特性，基因组的其他区域也存在大小长度不等的插入缺失事件，但它们均长度较短且相对不密集。

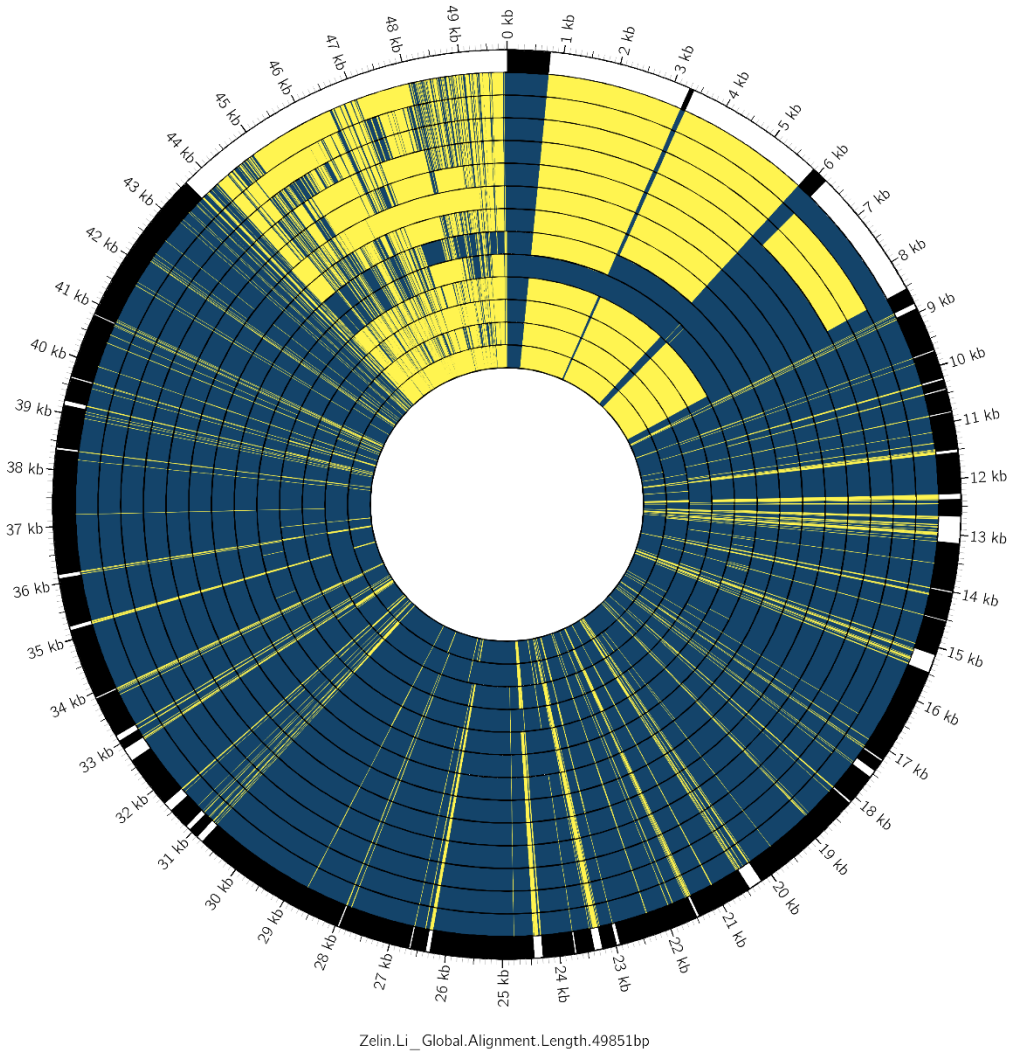

图 11 13 个株系 gaps 分布与插入热点分布

以深蓝色区域表示非 gaps 位点，黄色区域表示 gaps 位点，株系由内向外分别是 2 个菱形藻（NC_037729、MH297491）、2 个连续拟菱形藻（CNS00097、CNS00090）、多列拟菱形藻 NC_027265、疑似多列拟菱形藻 CNS00158、7 个尖刺拟菱形藻（CNS00141、CNS00055、CNS00154、CNS00156、CNS00089、CNS00153、CNS00110），此排序与图 9 的 13 个线粒体基因组系统发育树中从下到上顺序、图 12 的序列比对的株系从下到上排序、图 13 的单核苷酸差异位点分布与密度状况图的由内向外株系排序一致。圆环参考坐标总长度与全局比对长度相同，为 49851 bp。

除此内侧的13圈外，计算所得插入热点的分布显示在了图11的最外侧一圈。此处定义插入热点为：全局比对当中的特定 gaps 区域，其在不同株系中存在不同数目的 gaps，即此区域存在比其他位置更为频繁的插入或缺失事件，图 12 展示了一个大约位于全局比对结果中 24300-24440 bp 的插入热点的结构（在图 11 中此插入热点显示为白色，位于图最外圈正下方 24.30-24.44 kb 处）。

而考虑到真正的插入或缺失事件一般片段达到一定程度，所以寻找插入热点时过滤掉了每个株系中长度小于 10 bp 的 gaps，之后仍需定位插入热点的范围，设置相邻小于 100 bp 的两个 gaps 结合，算作同一个插入热点，依此算法得出这 13 个株系的插入热点共有 39 个，它们的位置即为图 11 最外圈的白色区域。这一寻找插入热点模式可能存在误差，但可以找到大部分插入热点，它们大多在短蛋白编码基因、tRNA 基因处。

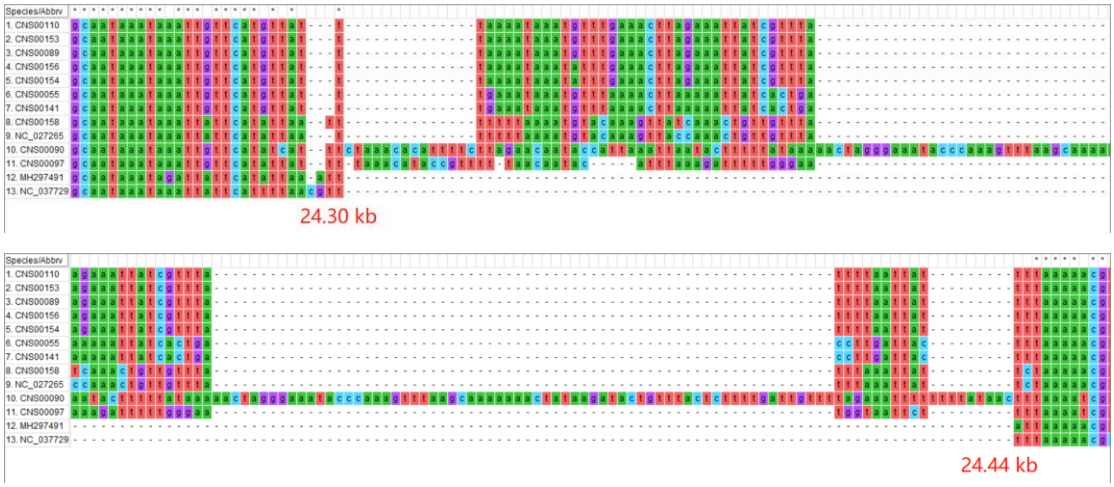

图 12 MEGA[49]中显示多序列比对 24.30-24.44 kb

不同株系在此处有不同程度的 gaps，从中可以发现，此区域（插入热点）只有连续拟菱形藻 CNS00090 基本没有 gaps，其他株系相对它有长度、位置不同的 gaps，例如前 9 个拟菱形藻株系（CNS00110、CNS00153、CNS00089、CNS00156、CNS00154、CNS00055、CNS00141 这些尖刺拟菱形藻；CNS00158、NC_027265 两个多列拟菱形藻）主要 gaps 分为 3 段，而连续拟菱形藻 CNS00097 则 gaps 分为 2 段，后 2 个菱形藻（MH297491、NC_037729）则仅有 1 段长 gap。据此认为不同株系在此区域发生过多次插入（或缺失）事件，其两侧均为保守序列，故称此区域是插入热点。

3.5.2 单核苷酸差异位点

以 CNS00141 为参考，检测其他 12 个株系相对它的单核苷酸差异位点位置，得到单核苷酸差异位点分布与密度图（图 13），总体上看，不同属的物种之间单核苷酸差异位点数量多于同一个属的不同物种之间，后者亦多于同一个属的相同物种的不同株系之间。7 个尖刺拟菱形藻明显分为 2 个类群，它们二者的主要差异比较均匀地分布在线粒体基因组各处。而本分析更进一步证实了 CNS00158 与多列拟菱形藻更相似，它们二者相对 CNS00141 的单核苷酸差异位点高度一致。同时也可以发现 CNS00141 与 CNS00055 高度相似，仅在序列末端的基因间区存在单核苷酸差异位点。

计算得到单核苷酸差异位点密度分布，得出变异率极高的区域主要有 5 个—

—6.2 kb 附近、14.7-15 kb、26.7-27 kb、38 kb 附近、38.7-40 kb，结合 CNS00141 的注释情况（图 10），发现这些位置对应的分别是 *rps2*、*rps3*、*nad2*、基因间区、基因间区。除此之外，密度略低于这些位置，但仍具有较高变异率的位点还有 3 个，是 11.2-12.3 kb、17.3-17.8 kb、25-25.4 kb，这些位置对应的分别是 *SecY*、*orf77* 和 *rps14*、*rps10*。

另外值得注意的是，这 13 个株系存在一段明显比其他部位更保守的片段，即 1.2-3.5 kb，这一区域对应 orf714。另一个较为保守的区域是 33-37 kb，这一位置有两个 rRNA 基因，即 *rnl* 和 *rns*。

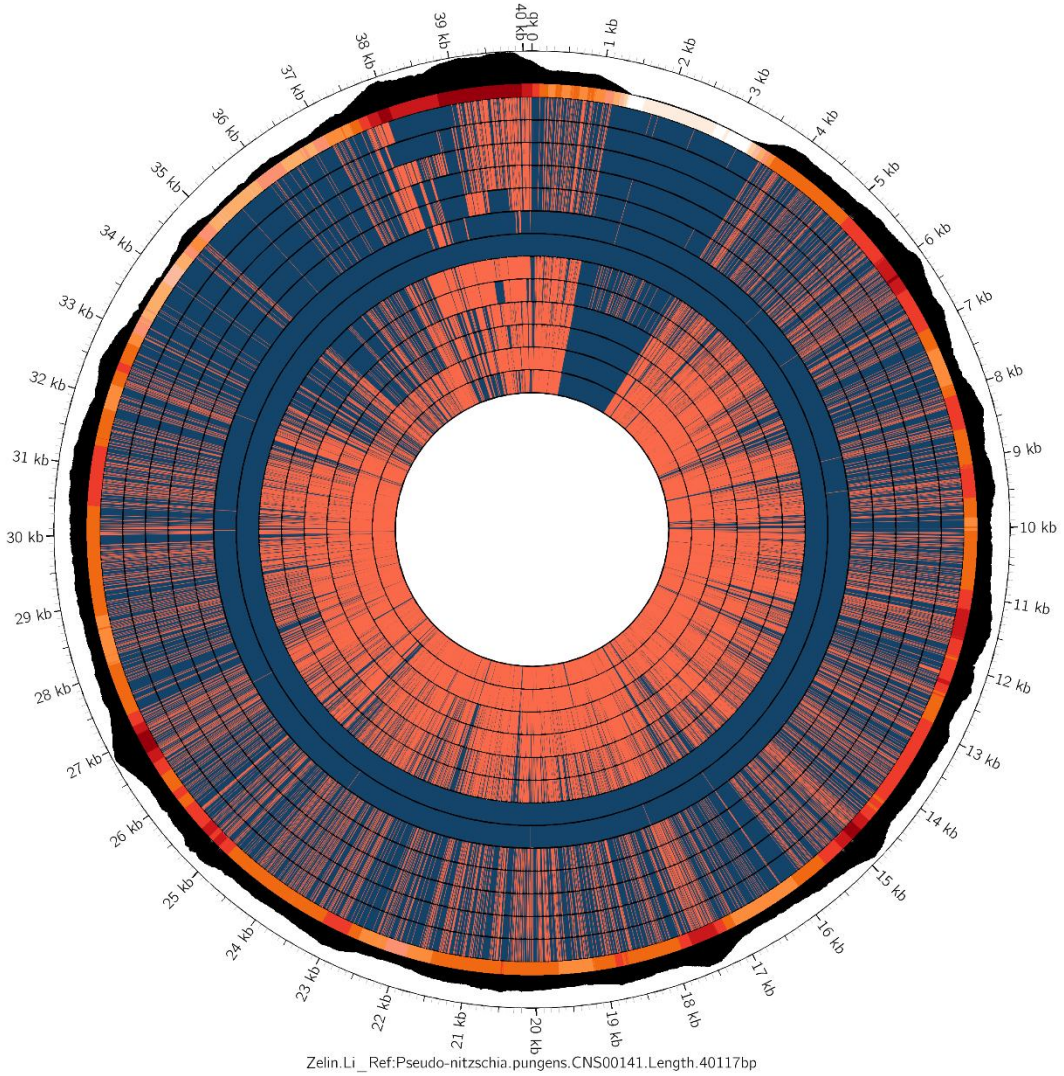

图 13 13 个株系单核苷酸差异位点分布与密度状况

以深蓝色区域表示与 CNS00141（由内向外第 7 个环）匹配一致的位点，以红色区域表示与之不一致的位点。最外圈为连续柱状图，以黑色填充，其高度表示以该位点为中心左右各 250 bp 的滑动窗口中包含全部株系取并集的单核苷酸差异位点数目（单窗口大小 501 bp），此数目越大则柱状图越高，最高为 408，最低为 4，在其内侧的圈的为热力图，是将此数值以 0.5 为底进行对数变换，设置 11 个颜色按梯度展示变换后数值大小，最高区间颜色为深红，最低区间为白色。该图株系从内向外顺序与图 11 相同，圆环参考坐标长度与 CNS00141 线粒体基因组长度一致，为 40117bp。

4 讨论

综合本研究过程和结果，认为这一新开发的线粒体基因组比较分析方法具有较好的准确性和自动化潜力，可制成生物信息分析流程。

本研究发现 CNS00158 株系的 18S rDNA 无法正确鉴定该株系物种所属，线粒体基因组与 18S rDNA“相悖”，导致这一现象有两种可能：

1、尖刺拟菱形藻与多列拟菱形藻 18S 部分株系极其相似（全长相似度＞99%），利用 18S 作为分子标记无法区分二者；

2、尖刺拟菱形藻部分株系与多列拟菱形藻部分株系间可能发生过自然杂交，其线粒体、核基因发生过重组或其他遗传事件。考虑到已有记录证实尖刺拟菱形藻亚种间可以发生杂交，其他某些海洋硅藻也可以发生种间杂交，这一可能性较大。

为了进一步检验 CNS00158 株系的种系，需要对其全基因组测序结果进行更深入挖掘。

在针对插入缺失位点和插入热点的分析当中，认为拟菱形藻和菱形藻均在序列末端的基因间区（*cox1* 和 *rnl* 之间）有最多最频繁的插入缺失事件，这一现象与多列拟菱形藻 *cox1* 存在不等长度的内含子或许也有联系。关于插入热点现象的成因，更值得深入挖掘。

据本研究结果可认为菱形藻与拟菱形藻线粒体基因组的基因组成存在明显差异，可以根据基因组成区分二者，但为了更标准化更简易地进行物种鉴定，可以通过 *rns* 或 *rnl* 作为分子标记进行扩增子测序即可实现属间鉴定。

而如果目的是区分拟菱形藻属某些物种间、物种内株系的差异，较为传统的分子标记 *cox1* 在大部分情况可以使用，例如，*cox1* 在尖刺拟菱形藻株系中均没有内含子，而且 *cox1* 本身也具有较高变异率，足够区分出不同拟菱形藻物种。但是在多列拟菱形藻中，其含有 1-2 个较长内含子；其中存在内含子，使得这一分子标记通过测序区分株系变得困难，但或许可以通过对 *cox1* 扩增产物通过其他途径如电泳，区分出“杂交多列拟菱形藻株系”（如 CNS00158）和一般的多列拟菱形藻株系。

如果要更进一步以更高分辨率区分各拟菱形藻属及菱形藻属的物种，可以考虑使用 *rps2*、*rps3*、*nad2* 作为分子标记，它们在拟菱形藻和菱形藻物种中均存在，且具有相对其他基因更高的变异率。

如果要独立设计分子标记，排除序列末尾的基因间区，根据滑动窗口分析结果（详见附 6），在 CNS00141 株系以 500 bp 窗口大小滑动检索到单核苷酸差异位点最多的区域是 26594-27094 bp，达到 382 个，这一区域位于 *nad2* 基因内，但这一区域上下游变异率均较高，但根据比对结果，在 26.7 kb 和 27.4 kb 附近，具有保守性较高的序列，可以考虑在这两处设计左右引物、分子标记，预期此分子标记将能以较高分辨率程度区分拟菱形藻属和菱形藻属物种。

参考文献

〔1〕Casteleyn G, Chepurnov V A, Leliaert F, et al. Pseudo-nitzschia pungens (Bacillariophyceae): a cosmopolitan diatom species?[J]. Harmful Algae, 2008, 7(2): 241-257.

〔2〕Evans K M, Kühn S F, Hayes P K. High levels of genetic diversity and low levels of genetic differentiation in North Sea Pseudo-nitzschia pungens (Bacillariophyceae)

populations 1[J]. Journal of Phycology, 2005, 41(3): 506-514.
〔3〕Casteleyn G, Adams N G, Vanormelingen P, et al. Natural hybrids in the marine diatom Pseudo-nitzschia pungens (Bacillariophyceae): genetic and morphological evidence[J]. Protist, 2009, 160(2): 343-354.
〔4〕Manhart J R, Fryxell G A, Villac M C, Segura L Y, Pseudo‐nitzschia pungens and P. multisenes (Bacillariophyceae): Nuclear ribosomal DNAs and species difference 1[J]. Journal of Phycology, 1995, 31: 421-427.
〔5〕Villac M C, Fryxell G A. Pseudo-nitzschia pungens var. cingulata var. nov. (Bacillariophyceae) based on field and culture observations[J]. Phycologia, 1998, 37(4): 269‐274.
〔6〕Lim H C, Lim P T, Su S N P, et al. Genetic diversity of Pseudo-nitzschia brasiliana (Bacillariophyceae) from Malaysia[J]. Journal of applied phycology, 2012, 24(6): 1465-1475.
〔7〕Lim H C, Teng S T, Leaw C P, et al. Three novel species in the Pseudo‐nitzschia Pseudodelicatissima complex: P. batesiana sp. nov., P. lundholmiae sp. nov., and P. fukuyoi sp. nov. (Bacillariophyceae) from the strait of Malacca, Malaysia[J]. Journal of phycology, 2013, 49(5): 902-916.
〔8〕Bates S S, Trainer V L. The ecology of harmful diatoms. In: Grane´li, E., Turner, J. (Eds.), Ecology of Harmful Algae[M]. Berlin: Springer-Verlag Heidelberg, 2006, 81‐93.
〔9〕林昕. 中国东南沿海五种拟菱形藻形态、遗传多样性与系统进化研究[D]. 厦门大学,2008.
〔10〕Trainer V L, Wekell J C, Horner R A, et al. Domoic acid production by Pseudo-nitzschia pungens[C]//Reguera B, Blanco J, Fernandez M L, et al. Harmful Algae. Paris: Xunta de Galicia and the IOC of UNESCO, 1998, 337‐340.
〔11〕Lundholm N. Bacillariophyceae, in IOC-UNESCO taxonomic reference list of harmful micro algae [EB/OL]. 2017, http://www.marinespecies.org/hab.
〔12〕Pogoda C S, Keepers K G, Hamsher S E, et al. Comparative analysis of the mitochondrial genomes of six newly sequenced diatoms reveals group II introns in the barcoding region of cox1[J]. Mitochondrial DNA Part A, 2019, 30(1): 43-51.
〔13〕Liu F. Comparative mitogenomics in diatoms[C]. Institute of Oceanology, conference about introns in diatoms, 20 Oct. 2019.
〔14〕Liu F, Li X, Che Z. Mitochondrial genome sequences uncover evolutionary relationships of two Sargassum subgenera, Bactrophycus and Sargassum[J]. Journal of Applied Phycology, 2017, 29(6): 3261-3270.
〔15〕Ševčíková T, Horák A, Klimeš V, et al. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte?[J]. Scientific reports, 2015, 5: 10134.
〔16〕董焕嫦, 黄春秀, 徐国双,等. 中国沿海尖刺拟菱形藻的种下分类学研究[J]. 热带海洋学报, 2018, 037(001):12-19.
〔17〕Lim H C, Teng S T, Lim P T, et al. 18S rDNA phylogeny of Pseudo-nitzschia (Bacillariophyceae) inferred from sequence-structure information[J]. Phycologia, 2016, 55(2): 134-146.
〔18〕Oudot-Le Secq M P, Green B R. Complex repeat structures and novel features

in the mitochondrial genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana[J]. Gene, 2011, 476(1-2): 20-26.
〔19〕An S M, Noh J H, Choi D H, et al. Repeat region absent in mitochondrial genome of tube-dwelling diatom Berkeleya fennica (Naviculales, Bacillariophyceae)[J]. Mitochondrial DNA Part A, 2016, 27(3): 2137-2138.
〔20〕Imanian B, Pombert J F, Dorrell R G, et al. Tertiary endosymbiosis in two dinotoms has generated little change in the mitochondrial genomes of their dinoflagellate hosts and diatom endosymbionts[J]. PLoS One, 2012, 7(8).
〔21〕Wilson X. Guillory, Anastasiia Onyshchenko, Elizabeth C. Ruck, Matthew Parks, Teofil Nakov, Norman J. Wickett, and Andrew J. Alverson, Recurrent Loss, Horizontal Transfer, and the Obscure Origins of Mitochondrial Introns in Diatoms (Bacillariophyta)[J]. Genome Biol. Evol., 2018, 10(6):1504–1515.
〔22〕Illumina Inc., Power and efficiency for large-scale genomics[EB/OL]. 2020, https://www.illumina.com/systems/sequencing-platforms/hiseq-2500.html.
〔23〕Sajeet Haridas, Colette Breuill, Joerg Bohlmann, Tom Hsiang, A biologist's guide to de novo genome assembly using next-generation sequence data: A test with fungal genomes[J]. Journal of Microbiological Methods, 2011, 86:368–375.
〔24〕Esmaeil Forouzan, Masoumeh Sadat Mousavi Maleki, Ali Asghar Karkhane, Bagher Yakhchali. Evaluation of nine popular de novo assemblers in microbial genome assembly[J]. Journal of Microbiological Methods, 2017, 143:32–37.
〔25〕Nurk S. et al., Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In: Deng M., Jiang R., Sun F., Zhang X. (Eds). Research in Computational Molecular Biology, RECOMB 2013. Lecture Notes in Computer Science[M], vol 7821. Springer, Berlin, Heidelberg.
〔26〕Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, and Inancx Birol. ABySS: A parallel assembler for short read sequence data[J]. Genome Research, 2009, 19:1117–1123.
〔27〕Rei Kajitani, Dai Yoshimura, Miki Okuno, Yohei Minakuchi, Hiroshi Kagoshima, Asao Fujiyama, Kaoru Kubokawa, Yuji Kohara, Atsushi Toyoda & Takehiko Itoh. Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions[J]. Nature Communications, 2019, 10:1702.
〔28〕Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A S, Lesin V, Nikolenko S, Pham S, Prjibelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev M A, Pevzner P A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing[J]. Journal of Computational Biology, 2012, 0021.
〔29〕Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications[J]. BMC Bioinformatics. 2009;10:421.
〔30〕Altschul S, Gish W, Miller W, et al. Basic local alignment search tool[J]. J. Mol. Biol.. 1990, 215(3):403–410.
〔31〕XiaoLong Yuan, Min Cao, GuiQi Bi. The complete mitochondrial genome of Pseudo-nitzschia multiseries (Baciuariophyta)[J]. Mitochondrial DNA A: DNA mapped sequence analysis, 2016, 27(4):2777-2778.
〔32〕Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics.

Genome Res., 2009, 19:1639-1645.
〔33〕Li H. and Durbin R.. Fast and accurate short read alignment with Burrows-Wheeler Transform[J]. Bioinformatics, 2009, 25:1754-60.
〔34〕Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools[J]. Bioinformatics, 2009, 25(16):2078-9.
〔35〕Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data[J]. Bioinformatics, 2011, 27(21):2987-93.
〔36〕James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer[J]. Nature Biotechnology, 2011, 29:24–26.
〔37〕Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration[J]. Briefings in Bioinformatics, 2013, 14:178-192.
〔38〕Koboldt D, Zhang Q, Larson D, Shen D, McLellan M, Lin L, Miller C, Mardis E, Ding L, & Wilson R. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing[J]. Genome Research, 2012, DOI:10.1101/gr.129684.111.
〔39〕Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. The EMBL-EBI search and sequence analysis tools APIs in 2019[J]. Nucleic Acids Research, 01 Jul 2019, 47(W1):W636-W641.
〔40〕 Aaron C E Darling, Bob Mau, Frederick R Blattner, Nicole T Perna. Mauve:Multiple Alignment of Conserved Genomic Sequence With Rearragements.[J]. Genome Research, 2004, 14(7):1394-403.
〔41〕Katoh, Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability[J]. Molecular Biology and Evolution, 2013, 30:772-780.
〔42〕Hall B G . Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences[J]. Molecular Biology & Evolution, 2005, 22(3):792.
〔43〕Chan, P.P. and Lowe, T. M.. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol. Biol., 2019, 1962:1-14.
〔44〕Stothard P.. The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences[J]. Biotechniques, 2000, 28:1102-1104.
〔45〕Greiner S, Lehwark P and Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes[J]. Nucleic Acids Research, 2019, 47:W59-W64.
〔46〕Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R and Greiner S. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Research, 2017, 45: W6-W11.
〔47〕Kernighan B W , Ritchie D M . The C Programming Language[M]. The Bell System Technical Journal, 1988, 57.
〔48〕Crowell,R.M., Nienow,J.A. and Bruce Cahoon,A.,The complete chloroplast and mitochondrial genomes of the diatom Nitzschia palea (Bacillariophyceae) demonstrate

high sequence similarity to the endosymbiont organelles of the dinotom Durinskia baltica, Journal of Phycology, 2019.
〔49〕Kumar S, Stecher G, and Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets[J]. Molecular Biology and Evolution, 2016, 33:1870-1874.

致谢

能够完成这篇毕业论文，我要感谢我在中科院海洋所的导师陈楠生老师，他给我在科研方法、思考方式上提供了大量指导和帮助；我还要感谢我在海大的导师陈刚老师，他悉心监督我的毕业论文进展并予以许多帮助。同时，我要感谢为拟菱形藻相关研究做出先期工作的陈阳学长和王毅超学长，以及教导我生物信息学、计算机技能的徐青师姐。

这篇毕业论文的写就，于我个人也是一种成长，我从中学习到大量计算机编程知识和科研结果可视化、论文写作方式；是我在海大四年学习的见证。