古菌、生命树和真核细胞的功能演化

肖静¹, 范陆^{2,3*}, 吴顶峰¹, 徐雁冰¹, 赖登训², William F. MARTIN⁴, 朱瑞新^{1,2†}, 张传伦^{2,5}

1. 同济大学生物信息系, 上海 200092;

2. 南方科技大学深圳海洋地球古菌组学重点实验室, 深圳 518055;

3. 南方科技大学前沿与交叉科学研究院, 深圳 518055;

4. Institute for Molecular Evolution, University of Düsseldorf, Düsseldorf 40225, Germany;

5. 青岛海洋科学与技术国家实验室海洋地质过程与环境功能实验室, 青岛 266061

* 通讯作者, E-mail: fanl@sustc.edu.cn

† 通讯作者, E-mail: rxzhu@tongji.edu.cn

收稿日期: 2018-09-11; 收修改稿日期: 2018-11-19; 接受日期: 2019-01-03; 网络版发表日期: 2019-04-02

科技部全球变化国家重点研发计划项目(编号: 2018YFA0605802)、国家自然科学基金项目(批准号: 91851210、41530105、41673073)、广东省自然科学基金项目(编号: 2018B030311016)、深圳海洋地球古菌组学重点实验室和青岛海洋科学与技术国家实验室海洋地质过程与环境功能实验室项目(编号: MGQNLMD201810)资助

摘要 当前对早期生命演化过程的研究仍存在许多尚未解决的问题, 其中以真核生物“基因组起源”和“细胞功能演化”最为重要, 回答该问题的关键在于确定真核生物在生命树中的位置以及20亿年前获得线粒体的宿主与现代古菌之间的关系. 最新研究认为真核生物是从属于古菌的一支, 而新古菌的不断发现将进一步推动“真核细胞功能演化”的研究. 文章先对有关“古菌”、“生命树”和“真核细胞功能演化”的研究进展及挑战进行梳理, 之后单独总结了当前中国的古菌研究现状, 最后从“宏基因组”、“生物信息学方法”和“古菌的分离培养”三方面对通过构建古菌系统发育推动生命演化研究提出展望.

关键词 古菌, 生命树, 真核细胞功能演化, 宏基因组学, 生物信息学

1 引言

早期生命的起源、演化以及真核细胞功能的出现都是当前生命科学领域尚未完全解决的重大问题, 对该类问题已有的研究主要基于基因组序列的系统发生(Phylogeny)分析(Betts等, 2018; Weiss等, 2018). 由基因组序列数据构建的系统发生树可以探寻现存生命间的演化关系, 从而推断早期微生物生命演化的古老历史,

故生命树(Tree of life)结构是理解微生物演化的关键. 尽管基因水平转移(Lateral gene transfer, LGT)在微生物演化中具有十分重要的作用(Ochman等, 2000), 对微生物基因组的组成产生了重要影响, 但现有的系统发生分析工具在构建生命树时还无法准确反映出LGT的影响. 因此, 在承认原核生物是真核生物祖先的前提下, 系统发生研究还是选择通用的单拷贝基因作为早期生命演化和真核细胞功能演化的传递者.

中文引用格式: 肖静, 范陆, 吴顶峰, 徐雁冰, 赖登训, Martin W F, 朱瑞新, 张传伦. 2019. 古菌、生命树和真核细胞的功能演化. 中国科学: 地球科学, 49: 1082–1102, doi: 10.1360/N072018-0187

英文引用格式: Xiao J, Fan L, Wu D, Xu Y, Lai D, Martin W F, Zhu R, Zhang C. 2019. Archaea, the tree of life, and cellular evolution in eukaryotes. Science China Earth Sciences, 62: 489–506, <https://doi.org/10.1007/s11430-018-9326-1>

新古菌类群的发现能够加深我们对原核生物演化的认识, 从而促进我们对生命树深层结构的理解 (Adam等, 2017; Zhou等, 2018a). 近年来, 宏基因组技术和生物信息学的飞速发展有力地推动了与古菌相关的研究 (Bankevich等, 2012; Kang等, 2015; Loman等, 2012; Natarajan等, 2016; Roumpeka等, 2017; Sieber等, 2018), 也加深了我们对微生物多样性、生理学以及进化的理解. 为此, 本文首先梳理了在宏基因组技术推动下发现的新古菌及由其引发的对生命树结构热议, 同时对日益活跃的中国古菌研究进行了专门总结. 对于快速增加的核心数据(宏基因组数据), 本文进一步分析指出了其优势及所带来的挑战. 最后, 本文对古菌系统发育推动的生命演化研究方向进行了展望.

2 生命树研究的发展与争论

2.1 经典的生命树及其争论

2.1.1 三域生命树

由于缺少合适的外群, Woese等人最初提出生命树时尚未确定树根的位置 (Woese和Fox, 1977; Woese和Olsen, 1986). 直到1989年, Iwabe等(1989)和Gogarten等(1989)分别对不同的旁系同源蛋白序列进行系统发生分析, 他们得出的结论均表明系统发生树的根部位于细菌以及真核生物与古菌的共同祖先之间. 此后, Woese于1990年根据这一结论提出了有根的二域树, 并将古菌分为泉古菌门(Crenarchaeota)和广古菌门(Euryarchaeota)这两个门类 (Woese等, 1990). 三域生命树最新的支持证据来自于Da Cunha等(2017), 他们利用RNA聚合酶构建了二域系统发生树(图1a).

2.1.2 二域生命树

1984年, Lake等(1984)通过分析核糖体的结构, 利用简约法(Parsimony)构建的无根二域树显示真核生物与古菌中的一个门类是姐妹关系, 表明真核生物可能起源于古菌. 其后, 基于RNA序列, 利用改进的演化简约法(Evolutionary parsimony)构建了有根二域树, 结果显示在进化树上真核生物与泉古菌关系最近 (Lake, 1988). 不过由于当时无法选择合适的外群, 导致该二域树中古菌的另一个门类与细菌构成了生命树的另外一枝 (Lake, 1988).

其后, Martin和Müller(1998)提出线粒体和氢化酶

体的共同祖先出现在真核生物的祖先中. 该假说认为真核生物的形成源于一种依赖氢的严格自养的厌氧古菌宿主和呼吸产生分子氢的细菌的共生, 指出真核祖先的能量代谢依赖线粒体且是兼性厌氧 (Martin和Müller, 1998). 随着系统发生算法的进步, Embley等人在建树时考虑了分子序列的氨基酸组成异质性, 他们使用了CAT模型 (Lartillot和Philippe, 2004) 并获得了二域树的结果 (Cox等, 2008). 近期, Ettema等人通过对新提出的古菌门类——洛基古菌(Lokiarchaeota) (Spang等, 2015) 的系统发生分析以及细胞功能分析, 表明这一古菌门类与真核的关系较TACK超门 (Guy和Ettema, 2011) 关系更近. 洛基古菌与新发现的另外三个具有真核特征蛋白(Eukaryotic signature protein, ESP)的古菌门并称为Asgard超门, 被认为是原核生物与真核生物之间的过渡, 进一步支持了Lake的二域学说 (Spang等, 2015; Zaremba-Niedzwiedzka等, 2017) (图1a).

2.1.3 生命树争论的演变

生命树是自然分类的基础, 而“单一生命树是否存在”这一的争论却与进化论 (Darwin, 1859) 本身一样古老. 1859年, Darwin在《物种起源》中引入树形结构来解释生物进化的问题, 他认为分类学家发现的树形结构反映了进化上的潜在树形结构, 从而提出“所有现存或已经灭绝的物种都可以在一个树形结构中找到其位置”这一观点 (Darwin, 1859; Doolittle和Brunet, 2016). 其后的系统发生研究便直接利用分类学的表型特征作为构建进化树的依据. 最初利用表型特征建树时Darwin沿用Linnaeus的“二界系统”(Two-kingdom system), 将植物和动物作为生命树的两枝; 其后, 光学显微镜的发明使得许多单细胞的生物被发现, 于是Ernst Haeckel在1866年提出“三界系统”(Three-kingdom system), 将生命树分为植物、动物、原生生物三枝 (Haeckel, 1866; Kutschera, 2016); 显微技术的发展使得更为精细的细胞结构得以被观察到, Chatton(1937)根据细胞核结构的存在与否区分细胞, 后Mayr(1998)将其指定为原核生物、真核生物, 即为“二界系统”(Two-empire system) (Chatton, 1937; Mayr, 1998).

随着蛋白和DNA序列被逐渐鉴定出来, 同时使用这些序列进行进化树的推断成为可能 (Schwartz和Dayhoff, 1978), 生命树的拓扑结构成为了易于探索的科学问题. Woese和Fox(1977)发现了古菌, 其后他们基

图1 生命树发展及其相关理论

(a) “三域学说”与“二域学说”代表性的树。Woese提出“三域学说”，构建三域无根树；Miyata和Gogarten利用旁系同源蛋白分析确定树根位置；Woese构建三域有根树并确定古菌域主要有两个门；Forster在2017年加入了新发现的古菌门类进行系统发生分析得到三域树，由此引发新时代关于二域、三域的争论。Lake根据对核糖体结构分析提出“二域学说”，利用简约法构建二域无根树；其后在1988年利用位点特异模型构建二域有根树；Martin于1998年提出的氢假说，从能量代谢的角度解释真核是由古菌宿主与细菌的内共生形成，支持二域学说；Embley建树时考虑了序列间的氨基酸组成异质性得到二域树；Ettema定义新的古菌超门Asgard(绿色的一支)，构建的系统发生树显示其与真核关系最近，是真核的祖先，建树结果支持二域学说。(b) 进化树中的垂直基因传递与水平基因转移

于rRNA序列提出了三域树(Fox等, 1980; Woese和Olsen, 1986), 且其中古菌形成了一个单系; 在有根树中真核与古菌形成了姐妹关系(Woese等, 1990)。如今我们一般称之为“三域”(3D)树(图1a)。1988年, Lake使用同样的数据, 但选用了不同的方法——进化简约法(Evolutionary parsimony)构建生命树, 提出了不同的真核与古菌的关系, 即真核是古菌的一个分支后代而非其姐妹, Lake称之为泉古菌树(Eocyte tree), 现在通常

叫“二域”(2D)树(图1a)。Lake(1988)和Woese等(1990)的竞争理论之间存在显著差异, 前者在叶绿体和线粒体内共生起源的背景下处理真核生物与古菌的进化关系, 而后者将真核生物视为独立于共生事件的一个谱系。

如今争论仍延续着, 当选择如贝叶斯方法等系统发育方法建树时会获得二域树, 而使用一些的传统的方法(如最大简约法)获得的结果则是三域树(图1a)(Ar-

chibald, 2008; Williams等, 2013), 古菌依旧处于争论的核心(Adam等, 2017). 近期的研究显示新发现的古菌品系是离真核生物尤其近的品系, 支持二域理论(Guy和Ettema, 2011; Spang等, 2015; Zaremba-Niedzwiedzka等, 2017). 这引发了对生命树结构和新古菌品系进化重要性的争论(Da Cunha等, 2018; Spang等, 2018; Zhou等, 2018a).

值得注意的是, 二域树和三域树的争议只考虑了微生物演化中垂直传递的基因, 但是随着基因组研究的发展, 已发现LGT对生命的三个域的演化历程均产生了重要的影响——包括原核生物间简单的质粒或单基因的转移, 以及真核生物演化过程中内共生事件引发的大量基因的转移(李志江等, 2008; Dagan和Martin, 2007; Martin, 2017; Ochman等, 2000; Zhaxybayeva和Doolittle, 2011). LGT是指不同有机体的遗传物质的转移, 且受体与供体并非亲代——子代的关系(Soucy等, 2015). 在原核生物中, LGT来自正常的DNA重组过程(Forterre, 2015; Gogarten和Townsend, 2005; Popa和Dagan, 2011); 真核生物中最常见的, 也是文献记录最多的LGT形式为线粒体和叶绿体基因向细胞核的转移(Timmis等, 2004). 在研究真核生物与原核生物的关系时, 应该考虑LGT的影响. 然而, 绝大多数真核基因都可追溯到细菌来源而不是古菌来源, 因此在分析时需要考虑所有的基因, 而不仅仅是那些将真核生物与古细菌联系起来的基因(Ku等, 2015).

树形结构的构建是基于基因的垂直传递, 但由于LGT的广泛存在, 有学者认为其对早期生命演化产生了重要影响, 应使用生命网(Net/Web of life)的结构替代生命树结构(Doolittle, 1999, 2000; Hilario和Gogarten, 1993; Martin, 1999)(图1b), 不过该理论仍存在一定的争议, 并未得到广泛认同(Forterre, 2015). 但在构造生命树时需要考虑LGT的影响, 避免在建树过程中产生错误. 理论上, 所有用于建树的基因都应当是垂直传递的, 需要排除LGT, 但是识别LGT仍旧存在一定的困难. 尽管应用了不同的方法, LGT(尤其是一些古老的LGT事件)仍然难以被全部识别出来.

尽管在构建生命树的过程中需要排除所有的LGT, 但LGT在细胞功能演化, 例如线粒体与叶绿体的起源中起到了重要的作用(Zimorski等, 2014), 因此LGT的信息可能有助于生命树的重构. 目前, LGT还无法被整合到树重建中, 因为即使是最新的算法也

不能正确处理它们. 不过, 基因组数据积累和模型开发的进步可能会使得这一操作在将来得以实现.

2.1.4 三域生命树与二域生命树的主要争论

虽然二域理论(图2b)逐渐受到认可(Spang等, 2015; Zaremba-Niedzwiedzka等, 2017), 但许多研究表明那些新的古菌不是与真核关系最近的群体, 相反它们支持三域理论(图2a). 因此, 这两种截然不同的理论引发了热烈的争论(Castelle和Banfield, 2018; Da Cunha等, 2017; Guy和Ettema, 2011; Spang等, 2018). 在这里, 我们总结了这场争论的实质(表1).

Ettema等人通过对其鉴定发现的洛基古菌进行系统发生分析表明其与真核生物关系最近, 同时其还具有部分ESP(Spang等, 2015). 其后有报道指出, 在膜脂合成中, 洛基古菌缺少合成古菌膜脂中醚键的能力却能合成真核脂肪酸的酯键. 因此, 洛基古菌反映了从“古菌型醚键”向“真核型酯键”过渡的表型, 可能导致了当前生物的“膜脂分界”(lipid divide)(Villanueva等, 2017). 其后, Ettema等人根据系统发生分析结果, 将新鉴定出的2个古菌门——海姆达尔古菌门(Heimdallarchaeota)和奥丁古菌门(Odinarchaeota), 与托尔古菌门(Thorarchaeota)和洛基古菌合并为Asgard超门, 并确定其在进化树上是离真核生物最近的支系, 同时在该超门中发现了更多的ESP, 进一步支持了其二域学说的观点(Seitz等, 2016; Zaremba-Niedzwiedzka等, 2017).

然而, Forterre等人对上述研究提出质疑, 认为他们的工作中数据集的选择存在问题, 这些问题包括数据集中存在古菌快速进化物种(Fast-evolving species, FES), 物种分布的不均衡, 宏基因组样本可能存在污染, 以及分子标记的选择不当. 因此他们改进了数据集并选择新的保守蛋白构建了系统发生树, 在他们看来结果支持了三域学说(Da Cunha等, 2017, 2018)(表1).

2.2 基因水平转移与生命树

LGT现象最早在20世纪40年代的微生物研究中被提出(Soucy等, 2015; Tatum和Lederberg, 1947), 其后的研究表明LGT不仅仅存在于原核生物中, 该现象广泛存在于生命的三个域中, 对生物演化的历程产生重要的影响(Hilario和Gogarten, 1993). 最近有研究者主张在建树的时候去除水平转移的基因(Forterre, 2015). 然

图 2 三域树和二域树中Lokiarchaeota以及Asgard超门的系统发生位置比较

星号为Asgard超门古菌成员. (a) 三域树, 根据Da Cunha等(2017)中图8修改; (b)二域树, 根据Zaremba-Niedzwiedzka等(2017)中图1b修改

表 1 近期三域/二域之争论^{a)}

主要工作者(时间)	结论	依据	参考文献
Ettema(2015)	洛基古菌与真核生物具有共同祖先	36个保守蛋白进化树支持洛基古菌与真核生物具有共同祖先	Spang等, 2015
Ettema(2017)	新古菌超门Asgard与真核生物具有共同祖先	48个保守蛋白、55个保守核糖体蛋白和16S/23S核糖体的进化树支持Asgard与真核生物具有共同祖先	Zaremba-Niedzwiedzka等, 2017
Forterre(2017)	支持三域学说, 不存在真核生物的古菌祖先	1. RNA聚合酶进化树支持三域学说 2. 指出(Spang等, 2015)文章中的36个保守蛋白建树结果不一致(其中11个Woese's蛋白支持三域进化树)	Da Cunha等, 2017
Ettema(2018)	Asgard与真核生物具有共同祖先	1. 进化模型(LG+GAMMA)*的过于简单 2. Lokiarchaea的基因组并非受到真核生物基因组的污染	Spang等, 2018
Forterre(2018)	支持三域学说, 不存在真核生物的古菌祖先	1. 系统发育分析中引入古菌快速进化物种会影响进化树的正确结构 2. Lokiarchaea的基因组受到污染	Da Cunha等, 2018

a) *, 该模型用于对11个支持三域树的蛋白的系统分析(Fig. 1B, Da Cunha等, 2017).

而, 另一方面, 有研究者认为不应忽视LGT对生命演化的影响, 因而提出用生命网的结构代替生命树的结构. 他们坚信用生命网结构更能反映真实生命演化的历程(Doolittle, 1999, 2000; Hilario和Gogarten, 1993). 不过生命网结构并未得到广泛认同, 一些研究者坚持树形结构更能反映生命演化历程, 他们指出生命树的构建应当反应各个域间的关系, 据此需要使用一些细胞演化历程中的核心基因, 而非简单的考虑所有的基因组(Forterre, 2015).

尽管用网状结构代替树形结构来描述生命的演化

还存在争议, 但LGT在生命演化过程中的重要性得到了一致的认同(王洽等, 2014; Gogarten和Townsend, 2005; Martin, 2017; Syvanen, 1985; Zhaxybayeva和Doolittle, 2011). 尽管目前对原核生物中发生的LGT已经有了系统的记录(Popa和Dagan, 2011), 但一些所谓的在真核生物中发现的LGT则缺乏可信性, 因为许多大肆宣传的“真核生物中发现的LGT”最后都被证实其实是基因组序列的污染, 这一点已经在其他文章中讨论过了(Martin, 2017).

LGT同样会影响生命的域的关系, 因为在演化的

过程中域可能会产生融合. 非树形演化结构中涉及 LGT 的一个代表性的例子就是细胞器的内共生起源. 线粒体来源于 α -变形杆菌, 叶绿体来源于蓝藻细菌, 不同藻类的次级质体(被3或4层膜包围的质体)来源于细胞的内共生融合(Zimorski等, 2014). 在每个内共生事件中, 一个细胞(内共生体)居住在另外一个远源相关的细胞(宿主)中. 此外, 在每个内共生事件中, 数以百计的基因——某些情况下甚至是数以千计的基因——会从内共生体转移到宿主中(Timmis等, 2004; Ku等, 2015), 因而每次都会形成属于全新类群的全新细胞. 回顾演化的过程中的大事件, 产生的全新的细胞类型分别为: (1) 出现在10亿年前的部分藻类品系中的具有复杂质体的细胞; (2) 出现在至少12亿年以前的具有质体的细胞(植物); (3) 出现在约16~18亿年以前的具有线粒体的细胞(真核生物)(Parfrey等, 2011).

获得线粒体是真核生物产生中的一个至关重要的事件. 在以内共生为基础的理论中, 线粒体的形成是真核起源的关键环节, 因为线粒体不仅作为能量来源, 同时也提供了重要的材料——细菌的基因以及辅助内膜系统形成的线粒体衍生的囊泡(Gould等, 2016; Martin等, 2016, 2017). 尽管如今对内共生的阐释已经与Margulis时期(Sagan, 1967; Margulis, 1970; Whatley等, 1979)有了较大的差异, 但目前可以确定的是线粒体的起源对明确各个域的关系的整体图景是至关重要的.

此外, 还有研究者依据内共生理论, 认为真核生物的形成是古菌与细菌的基因组融合(Fusion)导致的, 从而提出生命环(Ring of life)的结构(Lake和Rivera, 2004), 但是该环状结构没有得到广泛的认同(McInerney等, 2014). 事实上大多数具有原核同源物的真核基因是来自细菌而非古菌(Esser等, 2004; Ku等, 2015; Pisani等, 2007), 可见在生命树的研究中仍然存在许多尚未解决的问题. 因此, 进一步研究所有基因和全基因组的进化, 而不是基因组的小片段或像rRNA这样的单个基因, 将促进我们对早期生命演化过程的理解.

3 系统发育法重构生命树的关键问题

生命树的真实拓扑结构仍旧是未知的, 尽管生命树的实质是探索所有物种(理想情况下全部的基因)的系统发育关系, 但目前这还是一个无法实现的目标. 目前, 基于通用的分子标记的系统发育树的构建方法仍

是主要的探寻生命树物种间关系的方法. 在构建系统发育树时, 尤其是引入宏基因组数据建树的时候, 一些问题尤其需要引起重视.

3.1 纯培养基基因组及宏基因组数据的质量控制

在利用草图基因组序列建树时, 首要的是对基因组序列的质量(尤其是污染度和株系异质性)进行评估. 在宏基因组学普及以前, 微生物基因组主要是通过纯培养获得的(表2). 近年来, 利用宏基因组鸟枪法和单细胞测序技术, 已产生了大量的未培养的微生物基因组数据(Rinke等, 2013; Tully等, 2018). 与纯培养获得的基因组相比, 对这些组装基因组讨论最多的是: 不完整性; 存在嵌合现象; 存在污染; 同时还有基因组的异质性问题.

近些年来已经开发了一些用于评估基因组组装质量的工具(Gurevich等, 2013; Parra等, 2007; Simão等, 2015). 例如, Rinke等(2013)利用139个细菌和162个古菌的保守标记基因用于评估单扩增基因组(Single amplified genomes, SAGs)的质量. 通常而言, 这些评估方法可以分为三类: 一些方法将组装的重叠群(Contig)比对到相近的参考基因组(Mikheenko等, 2016); 一些方法则无需参考基因组, 通过直接计算读长(reads)的组装的指标进行可能性评估(Clark等, 2013; Ghodsi等, 2013); 其他的方法则是评估广泛存在的单拷贝保守基因(Eren等, 2015; Parks等, 2015).

CheckM是目前使用最为广泛的评估软件之一(Parks等, 2015), 其提出评估时完整度 $\geq 70\%$ 的基因组可以被视为基本完整的基因组(CheckM根据特定标记基因集中单拷贝标记基因出现的比例评估完整度), 而污染度 $\leq 10\%$ 则可视为是轻微污染(CheckM依据每个标记基因集里鉴定出的多拷贝标记基因数量来评估污染度; 此外, 菌株异质性指数是指“高于特定AAI阈值的多拷贝基因对的比例”, 默认阈值为0.9). 不过, CheckM并非评估宏基因组组装基因组(Metagenome-assembled genomes, MAGs)的唯一标准, 上文中提到的其他方法都可以用于评估基因组质量(Cornet等, 2018; Meyer等, 2018).

根据CheckM的标准, 组装和分箱的过程会产生大量基因组的污染(Lasken和Stockwell, 2007; Nurk等, 2013). 例如, Lokiarchaeota是解决古菌和真核系统发生关系的关键物种(Spang等, 2018), 但Loki sp.

表 2 系统发育分析中的重要问题

关键问题	相关步骤	参考文献
质量控制	1. 数据库下载的纯培养的基因组数据. 评估纯培养物种的基因组的完整性	Cox等, 2008; Foster等, 2009; Guy和Ettema, 2011
	2. 单细胞组装的基因组, 通过两个保守的单拷贝基因集估计基因组大小以及完整性	Rinke等, 2013
	3. 基因组中序列受到污染会影响建树的结构	Da Cunha等, 2017; Williams和Embley, 2014
	4. 需要对宏基因组组装基因组的完整度和污染度进行仔细的评估(软件: ALE; CheckM; MetaQUAST; BUSCO)	Clark等, 2013; Mikheenko等, 2016; Parks等, 2015; Simão等, 2015; Spang等, 2015; Zaremba-Niedzwiedzka等, 2017
标记蛋白	1. 建树时使用串联保守蛋白	Cox等, 2008; Foster等, 2009; Williams和Embley, 2014; Williams等, 2012
	2. 建树时使用原始蛋白&重聚类蛋白	Cox等, 2008; Foster等, 2009; Spang等, 2018; Williams等, 2017; Zaremba-Niedzwiedzka等, 2017
	3. 蛋白重聚类或是在蛋白比对的超级矩阵中直接删除一些位点都会使部分导致系统发生信号丢失	Lasek-Nesselquist和Gogarten, 2013
	4. 仅使用少量可以反映进化真实情况的蛋白构树, 而不是仅仅串联一些保守蛋白: 如使用2个最大的RNA聚合酶亚基联合构树	Da Cunha等, 2017
模型使用	1. 重构祖先的关系时组成异质性是一个广泛存在的问题, 应该使用可以处理这一问题的模型, 如NCDH和CAT模型	Cox等, 2008; Foster等, 2009
	2. 使用LG+Gamma模型和CAT模型	Guy和Ettema, 2011
	3. 进化模型的选择会影响系统发育分析的不一致水平	Williams等, 2012
	4. 比较了4个模型(LG4M、LG4X、UL3、CAT), 其中CAT模型更优	Lasek-Nesselquist和Gogarten, 2013
	5. 使用CAT-GTR 混合模型	Williams和Embley, 2014; Zaremba-Niedzwiedzka等, 2017
	6. 后验预测模拟的结果表明, 这些模型(LG, CAT+Pisson, CAT+GTR)中没有一个能够充分考虑位点间或分支间组成的异质性	Williams等, 2017

GC14_7的基因组的污染度(45.15%)和种系异质性(78.12%)都非常高(Da Cunha等, 2017, 2018). Spang等(2018)回应菌株特异性不应被解释为污染, 他们强调了Lokiarchaeum的基因组包含关系很近的菌株的序列, 但并未对该基因组的高污染度(45.15%)做出回应. 最近, 在生命树的构建中标记蛋白的质量受到了格外的关注. Da Cunha等(2017)指出Lokiarchaeota和Heimdal-larchaeota基因组中的蛋白——延伸因子EF2受到了真核序列的污染, 而在系统发生树构建时该蛋白是否存在会使生命树的拓扑结构发生彻底的改变. 近期的一项研究指出, EF2在一些古菌基因组中具有旁系同源物, 其具有复杂的进化历史(Narrowe等, 2018). 因此, 应谨慎选择用于构建系统发育树的标记蛋白. 总之, 在系统发育分析中, 对纯培养以及宏基因组获取的基因组数据进行质量控制, 是排除随机误差的关键步骤.

3.2 标记蛋白的质量控制与重聚类/非重聚类处理

系统发育树的构建过程中通常会使用保守的标记蛋白. 这些保守的标记蛋白通常为单拷贝的, 且理想状况下, 在研究中使用的数据集涉及的物种中不会发生LGT(Guy等, 2014). 以往的研究在构建生命树时, 通常会把多个标记蛋白串联起来构建一个超级序列比对矩阵(表2). 此外, 一些研究中使用的是原始蛋白(非重聚类蛋白)而有些研究则使用的是重聚类蛋白(表2).

对标记蛋白的重聚类/非重聚类处理同样也会影响生命树的拓扑结构. 在构建系统发育树时, 选择原始蛋白还是重聚类蛋白仍缺乏定论. 蛋白重编码是一种数据简化方式, 具体做法是将20种氨基酸字母表转换为一个更小的字母表, 通常是转换成4种字母表(SR-4重编码)(Susko和Roger, 2007)或者6种字母表(6阶Dayhoff重编码)(Dayhoff等, 1978). 重编码的本质是一

种粗粒化(Coarse-Grain model, CG)模型(Kmiecik等, 2016), 在计算资源不足的年代, 该方法牺牲了一定的精确度以保证任务得以顺利完成(Kmiecik等, 2016; Lettieri和Zuckerman, 2012). 粗粒化模型在分子动力学模拟领域被广泛使用, 然而随着近年来高通量计算和云计算的发展, 当原子级别的计算已经不再是问题的时候, 这些模型就逐渐被淘汰了(Zhang等, 2009). 此外, 重聚类蛋白会导致进化信息的缺失并且可能会造成系统发育的假阳性现象, 关于这一点, Lasek-Nesselquist和Gogarten(2013)已经在他们的研究中进行过评估了. 在利用重聚类蛋白构建的系统发育树中, 无关的物种更倾向于聚在一起且具有一个共同的祖先. 一个代表性的例子就是氢化酶相关的NADH脱氢酶模块: 不进行蛋白重聚类时, 滴虫属(*Trichomonas*)的氢化酶体的分支是独立于线粒体的(Dyall等, 2004); 但对蛋白进行重聚类后, 其会与线粒体聚在一起形成一个单系组(Hrdy等, 2004). 此外, 在Ajith Harish的近期研究中, 数据展示网络(data-display network, DDN)分析表明, 相比于原始蛋白(非重编码蛋白), 重编码蛋白增加了系统发育的噪音(Harish, 2018). 近期对古菌和真核关系的研究中, 有的使用原始蛋白(Spang等, 2015), 而有的则使用重聚类蛋白(Zaremba-Niedzwiedzka等, 2017).

除了原始蛋白和重编码蛋白序列, 近期研究提出还可将保守蛋白域的存在/缺失信息作为生命树重构的标记分子(Harish, 2018). 与蛋白序列相比, 这些蛋白域标记分子被认为携带更多的位点信息, 不过它们缺乏物种演化过程中积累的丰富的位点替换信息.

3.3 长枝吸引效应(LBA)和GC组成偏差导致的趋同性

趋同性(Homoplasy)是指相似但非同源的生物性状(Torres-Montúfar等, 2018), 是分子进化中公认的系统发育噪音. 在系统发育分析中, 长枝吸引效应(Long branch attraction, LBA)和GC组成偏差是引起趋同性的主要因素. 长枝吸引效应是指, 当两个(及以上)的世系比其他支系的枝长都长时, 它们会无视它们真实的关系却更倾向于聚在一起(Quince等, 2017). 对FES稀疏的取样也会在构建生命树时造成长枝吸引效应(Bodilis等, 2011; Brinkmann等, 2005). 加入充足的与长枝(引起长枝吸引效应的品系)关系较近的类群样本有助于解决这一问题(Quince等, 2017). 除了快速进化物种,

外群(Outgroup)也可能是造成长枝吸引效应的原因之一, 当长枝类群被外群吸引时, 它们会出现在树的靠近根部的位置(Philippe和Laurent, 1998). GC组成偏差是另一个影响系统发育树拓扑结构的重要因素. 在系统发育分析中, 具有相似核苷酸组成的物种(AT富集或者GC富集物种)倾向于聚集在一起, 尽管他们不是由同一祖先进化形成的. 总之, 长枝吸引效应和GC组成偏差仍然是生命树重构中亟待解决的关键难题.

3.4 系统发育重构中方法与模型的选择

系统发育树重构的核心算法主要包括“位点替换模型”和“位点进化速率模型”两块. 这些模型会影响用于评估不同系统发育树可靠性的指标(例如似然性). 因此, 选择合适的模型非常重要. 可以使用ModelFinder(Kalyanamoorthy等, 2017)这样的方法选择模型. 但是, 应该注意的是, ModelFinder不支持类似CAT(Lartillot和Philippe, 2004)和C60(Quang等, 2008)这样的更高级的位点进化速率模型.

在位点替换模型中, 广义时间可逆(General Time Reversible, GTR)模型(Waddell和Steel, 1997)因为考虑了每两种氨基酸或核苷酸之间的替换频率, 被视为目前最精准的模型. 在位点进化速率模型中, CAT模型(Lartillot和Philippe, 2004)因为考虑了每个位点的进化速率, 因此比其他位点进化速率模型更为高级. 蛋白质位点的异质性和多种蛋白质的进化速率差异是物种进化的本质特征(Williams等, 2013).

一个强大的模型应该考虑到位点的异质性. 目前仅有贝叶斯推断能同时集成GTR和CAT模型, 这使其成为许多研究者的首选. 此外, 在极大似然法中存在一个替代模型C60(Quang等, 2008), 该模型是CAT的一个简化版本, 因为极大似然法没有用于集成CAT模型的计算框架. 考虑到模型的优越性, 贝叶斯推断通常优于极大似然法(表2). 不过, 哪种方法能产生具有最优可能性的树与哪种方法能产生最能反映生物历史的树是不同的问题. 此外, 随着模型复杂性的增加会引入额外的自由度, 这也会影响模型辨别选择树的能力.

总之, 获取高质量的基因组和标记蛋白是构建稳健的生命树的关键. 当大量基因组是从宏基因组数据中分离出来时, 这一点尤为重要. 在构建生命树时, 我们认为最优的策略是使用集成CAT+GTR模型的贝叶斯方法, 同时选择高质量的基因组中获取的原始蛋

白序列. 与其他方法相比, 集成CAT+GTR模型的贝叶斯方法是唯一的使用非参数方法描述特异位点的氨基酸异质性的方法. 尽管极大似然法中的C60模型试图使用参数法来模拟位点异质性(Quang等, 2008), 但其不能反应出位点特异的变化. 因此, 贝叶斯方法是更好的选择. 然而, 同样需要注意贝叶斯方法也存在一定的局限性. 因为同时集成了CAT和GTR模型, 贝叶斯在运行时可能会耗费更多的时间, 且当数据集非常大时会消耗更多的运算资源. 当数据集中的物种分布更加均匀时贝叶斯方法会是更好的选择, 因为如果数据过于稀疏会导致系统发生树在构建时难以收敛. 此外, 物种数量, 标记蛋白质和LGT等其他因素也可能影响系统发育树的结构.

4 近期古菌分类研究促进生命树研究

古菌是二域和三域的争论中的关键物种. 最近的研究表明, 新发现的古菌谱系(即Asgard超门)是真核生物最近的支系, 支持二域理论(Spang等, 2015; Zaremba-Niedzwiedzka等, 2017). 这一发现引起了人们对生命树结构的热烈争论(Da Cunha等, 2018; Spang等, 2018). 同时, 在这些古菌中发现的ESP也反映出新古菌品系在进化研究中的重要性(Zhou等, 2018a), 这表明对古菌分类学的研究可以极大地促进对生命树的研究.

4.1 古菌门类的发展

古菌由最初的2个门(泉古菌门和广古菌门)(Woese和Fox, 1977; Woese和Olsen, 1986; Woese等, 1990)发展到目前的20多个(部分古菌分类地位存在争议). 研究者们根据对这些新发现古菌门类的研究提出了三个古菌超门: TACK超门(Guy和Ettema, 2011)、DPANN超门(Rinke等, 2013), 与Asgard超门(Zaremba-Niedzwiedzka等, 2017).

随着古菌基因组数据的积累以及分析技术的进步, 在门(Phylum)、纲(Class)、目(Order)、科(Family)、属(Genus)等各个分类水平上, 新的古菌类群不断被鉴定出来, 极大的推动了古菌的相关研究. Adam等(2017)的此前的研究中, 总结了到2017年为止从超门到科水平的一些新命名的古菌品系, 本文在网络版附表1(<http://earthen.scichina.com>)中整理了2018年新提

出的古菌门的信息, 包括最近提出的TACK超门中的Marsarchaeota和DPANN超门中的Huberarchaeota; 此外一些传统的门及其代表的品系也在表格中列出了(附表1). 这些新发现的古菌门推动了对早期生命演化的研究(Spang等, 2015; Zaremba-Niedzwiedzka等, 2017; Zhou等, 2018a).

通过总结目前大部分古菌门类的研究发展(附表1), 可以看出, 近5年来大量新的古菌门被确定. 随着技术的发展, 一些原本被划分为Crenarchaeota的古菌得以被识别并独立为一个门, 如奇古菌(Thaumarchaeota, 原为Marine group I)、深古菌(Bathyarchaeota, 原为Miscellaneous Crenarchaeotic Group)(DeLong, 1992; Meng等, 2014). 随着数据的积累和技术的进步, 这些古菌的分类地位得到了重新确立(Brochier-Armanet等, 2008; Meng等, 2014). 洛基古菌也是如此, Christa Schleper等人首次发现并将其归为深海古菌群(Deep-Sea Archaeal Group)(Jørgensen等, 2012, 2013), 该类群如今已经基于组装的宏基因组信息被重新分类定名(Spang等, 2015; Zaremba-Niedzwiedzka等, 2017). 同时, 已有在这些古菌中发现了ESP的报道, 表明它们可能是真核生物进化过程中重要的过渡类群(Zaremba-Niedzwiedzka等, 2017). 然而, 仍然没有人见过这些新的古菌类群的真实形象, 因此还无法支持它们真的是过渡形式这一观念(Martin等, 2017). 此外, 新的古菌类群数据的积累有助于对古菌生理生化特点、代谢特征、生态功能等进行进一步研究, 加深研究者对古菌的了解. 对产甲烷古菌、极端嗜盐古菌, 热泉古菌等特定古菌类群的研究, 可以解释古菌在特异环境中的遗传、代谢、进化等机制, 使得研究者们对古菌多样性的认识大大提高(Cao等, 2018; Chu等, 2016; Han等, 2017; Ma等, 2013).

4.2 古菌分类标准的发展

尽管新古菌类群的发现丰富了古菌多样性, 同时极大地推动了生命早期演化研究, 但古菌的类群研究中一个问题亟待解决, 即古菌的门及以上水平(超门)的类群划分缺乏明确的标准, 而这会导致分类的混乱(Gribaldo和Brochier-Armanet, 2012). 虽然在低一些的分类水平(如种、属)上, 古菌分类是具有一定标准的, 如基因组DNA-DNA杂交的百分比大于70%则认为是一个种, 或16S rRNA基因的相似性不高于98.7%则视

为两个种(Kim等, 2014; Schleifer, 2009), 但在“门”及以上的水平上则缺乏这种一般性的标准(Gribaldo和Brochier-Armanet, 2012). Sliva数据库对古菌的类别的划分主要是基于rRNA基因的系统发生树结果(Yilmaz等, 2014), 也有研究是根据16S rRNA基因的相似性或全基因组相似性(例如: Average nucleotide identity, ANI; Average amino acid identity, AAI)作为划分依据. 不同研究中标准选择的差异会导致结果的不一致, 目前古菌的类别划分没有统一结果.

此前有研究根据16S rRNA基因的相似性为古菌和细菌分类提出了量化的标准, 其针对每个分类单元都给出一个阈值, 门为75.0%, 纲为78.5%, 目为82.0%, 科为86.5%, 属为94.5%(Yarza等, 2014). 该研究结果对古菌门这一分类水平的划分提供了一个参考. 同时, 随着测序技术的发展, 产生了大量的古菌全基因组数据, 以全基因组数据为分类标准的方法得到了广泛应用, 尽管全基因组可以提供物种代谢、功能、结构等信息, 但该方法在较高分类水平的区分上仍存在一定缺点(Garrity, 2016). 综合目前研究中常用的古菌门类划分标准, 未来古菌门水平的分类标准可以综合考虑16SrRNA基因的相似性与标记基因的系统发生树的结果(Gribaldo和Brochier-Armanet, 2012). 不过16SrRNA基因相似性的阈值的划分, 以及构建系统发生树需要的标记基因等都需要谨慎的选择. 这些标准的确定最终会构建一个完善的古菌分类体系, 使得古菌的分类更加严谨, 有助于后续古菌研究的系统发展.

4.3 古菌在真核细胞的功能演化的作用

早期生命演化过程的相关研究中, 对真核细胞的功能演化的研究主要是对ESP的演化及其功能的研究. ESP是指那些在真核生物主要支系都具有同源蛋白但在古菌与细菌中都不具有同源蛋白的一类蛋白质(Hartman和Fedorov, 2002). 关于ESP的早期研究中获得了347个ESP, 对这些蛋白的功能分析表明其主要是与细胞核、细胞质和膜系统、以及信号系统等相关的蛋白(Hartman和Fedorov, 2002). 最近的研究涉及了对数十个测序的真核基因组和数千个原核基因组的比较, 确定了数千种ESP(Ku等, 2015). 已发现在一些新鉴别出的古菌中存在部分ESP, 其被作为真核细胞功能起源于古菌的证据, 支持二域学说(Spang等, 2015; Zaremba-Niedzwiedzka等, 2017).

需要强调的是, 尽管人们对新发现古菌类群的生物学重要性以及新古菌类群中预测出的ESP的作用进行了广泛的讨论, 但这些研究均是基于吞噬作用(Spang等, 2015; Zaremba-Niedzwiedzka等, 2017)和一个50年前提出的假设(细胞获得线粒体的必要机制是吞噬作用)(Martin等, 2017). 可是在考虑吞噬作用的生理学功能时, 该理论只适用于那些已经具有线粒体的细胞(Martin等, 2017). 这一认知提示新的研究需要重新关注这些新发现的古菌类群的生理学特点(He等, 2016; Sousa等, 2016), 而不仅仅只对真核相关的蛋白进行研究. 此外, 对动态内膜系统、囊泡转运、核膜、吞噬作用、减数分裂和细胞周期等在线粒体起源后演化出的这些“真正的”真核特性的蛋白进行研究——探寻其由古菌ESP的“原始材料”到成为具有功能的ESP的转化过程, 是完全可以实现的(Garg和Martin, 2016; Gould等, 2016).

真核生物的内膜系统的起源是真核组织复杂性、细菌的膜外囊泡(Gould等, 2016)、以及线粒体同源物的关键, 而线粒体衍生的囊泡(Sugiura等, 2017)在该问题中占有突出的地位. 在新的古菌品系中发现的Rab家族的同源的小GTP酶并不是真核特有的RabGTP酶(Spang等, 2015; Zaremba-Niedzwiedzka等, 2017), 因为其缺少在膜转运的膜识别过程中最关键的膜锚定位点(Surkont和Pereira-Leal, 2016). 此外, 有报道指出Asgard古菌品系中真核外壳蛋白的 α -螺旋和 β -螺旋结构域是融合的(Spang等, 2015; Zaremba-Niedzwiedzka等, 2017), 但在目前已研究过的原核生物(包括Asgard品系)中, 根本没有发现这种融合; 相反, 这种融合只在真核生物中出现(Martin等, 2017).

研究显示ESP数量有限, 目前的研究对于所有ESP都进行了分类(Hartman和Fedorov, 2002; Jian和Lesley, 2012), 所以目前对于在古菌中找到的ESP的分析一般基于这些已有的分类结果. 对古菌中发现的ESP的整合分析表明, 这些蛋白的功能主要集中在转录、翻译、细胞骨架的构成, 以及蛋白降解相关的过程, 提示了未来对真核细胞的功能演化的研究方向(Guy和Ettema, 2011; Spang等, 2015; Zaremba-Niedzwiedzka等, 2017). 附表2中总结了部分ESP并简要介绍了其涉及生物学过程. 尽管古菌中的ESP表明了真核特异的细胞功能起源, 但它们并不适合作为系统发育树构建的标记蛋白. 同时, 古菌中识别出的ESP的数量仍旧十

分稀少, 这也阻碍了原核和真核生物中ESP的系统发育的重构。

4.4 中国古菌研究的进展

4.4.1 古菌的生物学研究

当前古菌的研究方兴未艾, 相关研究发展迅速, 中国在古菌这一领域的研究拓展了对古菌的生态学、基因组学、生理学, 以及代谢途径等的认识. 一些研究对特定环境中的特殊古菌类群的生理生化、代谢遗传途径进行了研究. 例如, 氨氧化古菌(Ammonia-oxidizing archaea, AOA)是一种在全球范围内分布的古菌, 其在硝化作用中具有重要的作用. 据报道, AOA例如奇古菌(Thaumarchaea)在土壤中的自养生活(Zhang等, 2010); 还有研究分析了影响AOA的全球分布的演化机制以及环境因素(Cao等, 2013). 嗜热古菌生活在高温环境中, 具有很高的多样性. 有研究报道了深海嗜热古菌的适应性机制(Xie等, 2011); 此外, 还有研究对嗜热古菌特异的翻译后修饰进行了阐释(Cao等, 2018; Chu等, 2016).

碳代谢和脂质生物标记物的研究是古菌研究中的重要方向. 产甲烷菌——能够进行甲烷合成的古菌, 对其的研究已经有很多年了: 已有报道阐释了产甲烷古菌在稻田和湿地中的生物地理分布模式(Ma等, 2013; Zhang J等, 2018); 有研究发现产甲烷菌种存在基于AHL的群体感应现象(Zhang等, 2012). 深古菌(Bathyarchaeota)作为一种新命名的未培养古菌(Zhou等, 2018b), 对其的研究启发了对光合基因的起源和演化的研究, 同时推动了对深海古菌的生态学和地球化学功能的研究(陈玉连等, 2017; Meng等, 2014; Yu等, 2018a); 此外有报道显示深古菌具有分解木质素的能力(Yu等, 2018b). 最近有研究对Marine Group II (MGII)古菌在河口区有机质降解的组学特征进行了报道(Xie等, 2018; Zhang等, 2015). 除了上述碳代谢相关的古菌的研究, 关于古菌的脂质生物标记物的研究也取得了一定的进展. 对墨西哥湾冷泉和甲烷水合物中的古菌脂类(Zhang等, 2003, 2002)、热泉古菌脂类(Pearson等, 2004; Zhang等, 2006)和中国领海的古菌脂类(Ge和Zhang, 2016; Ge等, 2016; Jia等, 2012; Wang等, 2015, 2017; Wei等, 2011)等研究揭示了生物标志物代表的关于气候变化, 生物地球化学过程和氨氧化古菌分布的情况. 同时, 已有工作总结了中国古

菌脂质的生物地球化学的相关研究(Zhang C L等, 2018).

此外, 还有很多研究揭示了古菌的分子生物学特征, 包括DNA复制(Lang和Huang, 2015; Xing等, 2014), DNA修复机制(Han等, 2014; Sun等, 2018), 古菌的CRISPR系统(Liu等, 2016), 以及Exoribonucleolysis机制(Zheng等, 2017); 同时还有一些研究探索了嗜盐古菌的代谢机制并将代谢产物其应用于生物工程(Han等, 2017).

4.4.2 古菌的系统发生学

另外, 一些研究集中在古菌的系统发生上, 扩大了古菌的类群. 一方面是对特定古菌类群多样性的报道. 例如, 属于Micrarchaeota和Parvarchaeota古菌门的小的嗜酸古菌, 由于缺乏纯培养, 对该类古菌的了解还十分有限, 而近期有研究报道了该类古菌的多样性和可能的代谢途径(Chen等, 2018). 另一方面, 特定环境中(如热泉环境)的古菌多样性也在不断被阐明(Jiang等, 2010; Song等, 2013). 如今越来越多的古菌品系被提出, 极大的丰富了古菌的类群, 包括新的嗜盐古菌(Cui等, 2017; Ding等, 2017; Liu等, 2014; Lu等, 2017)、深海热泉沉积物中的新的嗜热古菌(Zeng等, 2013)、以及从马里亚纳海沟取样获得的新MGIIC的组装基因组(数据尚未发表). 值得注意的是, Bathyarchaeota(原命名为Miscellaneous Crenarchaeotal Group, MCG)是首个由中国学者提出的古菌门(Meng等, 2014). 此外, 虽然新的古菌门不断被提出, 但是对这些门的了解仍旧非常有限. 最近的研究对这些新的古菌门的生态学和演化特征有了整体的研究, 包括Woese archaeota(Liu X等, 2018)、Thorarchaeota(Liu Y等, 2018)、Aigarchaeota(Hua等, 2018)和Asgard超门(Cai等, 2018).

中国的这些古菌研究的进步丰富了我们不同环境中古菌功能的认识. 此外, 微生物碳泵(Microbial Carbon Pump, MCP)理论(Jiao等, 2010)以及CVTree理论(左光宏和郝柏林, 2017)分别在全球碳循环和微生物类群研究这两方面进一步扩大了古菌研究的重要性.

5 古菌的系统发生研究展望: 需要更高级的基因组技术

今后技术的发展将进一步推动古菌的研究发展.

过去对古菌的研究主要基于SSU rRNA基因, 而有研究指出, 传统研究中由于引物的限制会丢失环境中大量古菌的信息. 使用某些特定引物时, 研究环境中约50%的古菌都不能被识别出来, 但同时细菌则只有约2%不能被识别出来(Eloe-Fadrosh等, 2016). 显然, 相比于对细菌的研究, 对古菌的研究还远不够彻底(Eloe-Fadrosh等, 2016). 宏基因组数据的急速增长见证了宏基因组技术的发展(图3a和3b), 其对古菌研究产生了深远的影响(Adam等, 2017).

这里, 我们对两方面的技术发展进行了展望: 宏基因组技术和生物信息学方法(图3c), 它们的发展将进一步推动古菌研究. 采样技术和测序技术是宏基因组技术中最基础的两块, 它们的发展可以促进各种环境中新古菌品系的发现. 随着宏基因组数据的积累, 宏基因组相关的算法和系统发生算法不断得到改进, 以确定新古菌的演化位置和细胞功能.

5.1 宏基因组技术推动古菌研究

5.1.1 采样技术扩大研究范围

采样技术的发展是宏基因组研究的基础. 古菌的分布广泛, 在一些极端环境中(如深海沉积物、火山热泉、极地海洋)也有分布. 因此相应极端环境采样技术的发展对古菌的研究发展就显得尤为重要. 目前, 深海钻探、远洋采样等技术的发展, 推动了古菌的研究发现(孙松和孙晓霞, 2017). 可见采样技术的进一步发展将使得对各类环境中的古菌探索更加充分, 推动古菌相关研究.

5.1.2 测序技术推动研究进展

此外, 宏基因组相关实验技术的发展对于古菌研究也至关重要(Thomas等, 2012; Woyke和Rubin, 2014)(图3c). 目前实验中仍存在一些制约古菌研究发展的问题, 如难以从环境样本中提取全部古菌的DNA. 古菌的细胞壁与细菌的结构不同, 更加难以破坏, 在实验过程中难以获得环境样本中全部古菌的DNA, 会对后续的研究产生影响. 此外, 测序深度的限制可能导致一些样本中丰度较低的古菌难以被发现, 或是导致组装(Assembly)时产生不完整的基因组, 从而影响后续分析结果. 这些实验技术的发展也是未来古菌研究的重要发展方向, 而这些问题的解决也将促进古菌研究的发展.

5.2 生物信息学方法加速研究进程

5.2.1 宏基因组特异算法提高

随着取样技术和宏基因组学技术的进步, 已经积累了大量与古菌相关的研究数据. 因此, 除了特定的平台架构之外, 还需要更准确有效的生物信息学工具. 处理宏基因组数据时, 计算时间和组装所需的计算内存会急速增加. 对于一些关键技术, 例如当使用MCL (Markov cluster)方法对数千个基因组进行聚类时(Ku等, 2015), 只有在有一台有数TB的共享内存机器上才能进行计算. 因此, 在未来的研究中进一步优化算法尤为重要(图3c).

根据组成相似性和丰度分布模式, 将读长(Reads)或重叠群(Contigs)聚类到同一物种中, 这一过程称为分箱(Binning). 目前常用的分箱的方法分为监督方法和无监督方法. 前者依赖于已有的参考基因组数据, 但这类数据仍然非常有限, 而后的准确性需要进一步提高(Quince等, 2017). 分箱是一项关键技术, 因为它是介于原始宏基因组数据和未培养物种基因预测的中间步骤. 因此, 需要仔细估计宏基因组分箱获得的bin的质量, 并在下游分析中(特别是在系统发育分析中)去除低质量的bin. 质量评估方法(评估基因组质量或标记蛋白质量)的发展有助于重构稳定的生命树. 总之, 高质量基因组是系统发育分析的重要先决条件, 因此需要严格控制用于下游分析的宏基因组bin的质量.

5.2.2 系统发育重构算法的优化

在进行系统发育分析算法开发时, 需要考虑到一些会影响树形拓扑结构的因素, 包括趋同性(Homoplasy), 以及对进化模型中的位点异质性处理.

(1) 物种进化中的趋同性. 物种进化中的趋同性主要由长枝吸引效应和GC组成偏差造成, 而引起长枝吸引效应的物种(例如快速进化物种)具有一些快速进化的位点, 其构成了蛋白位点的异质性, 是物种演化的本质特征(Williams等, 2013). 所以, 理论上当使用的模型足够稳健时, 快速进化物种也可以被包含在系统发生分析中. GC组成偏差是另一个引起趋同性的重要因素. 从系统发育方法的角度, 尽管利用了LogDet/Paralinear转换的方法或者使用去除了偏差位点的方法, 仍旧无法完全解决GC组成偏差造成的问题(Foster和Hickey, 1999). 不过, 可以从物种选择的角度进行检验,

图3 古菌研究的发展

(a) 过去十年中新古菌门的快速发现。数据总结见附表1。(b) 古菌基因数据的累积。数据来源于NCBI(Nucleotide, <https://www.ncbi.nlm.nih.gov/nucleotide>), 搜索关键词为“metagenomic [All Fields] AND archaea [filter]”, 截止于2018/05/08。(c) 宏基因组学技术和生物信息学方法推动古菌的研究。宏基因组学技术包含取样技术和测序技术, 它们的发展对于古菌实验研究非常重要。生物信息学方法包括宏基因组特定算法和广义算法, 它们的进展将极大地影响古菌分析研究。火山热泉和深海图片来源于<https://pixabay.com/>

即用GC平衡的物种集合替换掉GC偏差的物种集合, 这样便可判断GC聚集的关系是反映的是同源性(真实进化历史)还是趋同性(系统发育分析假阳性)。

(2) 进化模型中的位点异质性处理。一个好的模型可以顾及位点的异质性, 并准确地表征物种进化过程中特定位点的变化。目前, 贝叶斯框架下的GTR+CAT模型是表征特定位点变异的最准确模型, 其被证明能够在一定程度上解决LBA的影响(Lartillot等, 2007)。因此, 当涉及引起LBA的谱系间的演化关系的判断, 尤其是在模拟位点的特异性突变时, 系统发育算法的优化将有助于解决这一问题。

CAT+GTR作为一种高级的进化模型, 目前仅被应用于PhyloBayes(Lartillot等, 2013)中, 这是一种常用的

贝叶斯推断软件。但PhyloBayes的运行时容易陷入局部最大值, 因此很难达到收敛。PhyloBayes收敛的难度与其估计参数后验分布时使用的抽样方式密切相关。PhyloBayes采用蒙特卡罗马尔可夫链(Markov chain Monte Carlo, MCMC)采样器来推断分子序列中一组进化参数的后验分布。通常MCMC采样器需要数千次采样才能达到“收敛”状态, 这是主要的限速步骤。值得注意的是, MCMC采样器的温度(即参数估计中的随机变量波动大小)不可加热, 这是MCMC采样器默认的。如果链条是可加热的, 则可以调整随机变量的波动以避免陷入局部最大值, 因此多次运行会更容易达到收敛。

不同于MCMC采样器, 一些其他的采样中链是可

加热的, 例如Metropolis耦合的蒙特卡罗马尔可夫链[(MC)3]. 相应的, (MC)3已被应用于其他一些贝叶斯推断软件中来解决局部最大问题并缩短收敛时间, 如BAMM(Moore等, 2016)、MrBayes(Ronquist和Huel-senbeck, 2003)和BEAST(Drummond和Rambaut, 2007). (MC)3可以在BAMM(Drummond和Rambaut, 2007)中使用名为“多核Metropolis耦合MCMC[(MC)4]”进行多个CPU并行, 或是在MrBayes v3.0中使用“parallel(MC)3[p(MC)3]”(Ronquist和Huel-senbeck, 2003). 此外, 随着计算技术的创新, 系统发育的MCMC分析已逐步实现在图形处理器(一个或多个图形处理单元(GPU))上运行, 这显著提高了MCMC的计算效率(Pang等, 2015; Ronquist等, 2012).

然而, 将可加热的采样器和CAT+GTR模型集成到PhyloBayes软件中是一项紧迫而艰巨的任务. 因此, 在贝叶斯推断的系统发生算法中, 可加热的采样器和CAT+GTR模型的集成将提高模型的计算效率和收敛性, 这也是贝叶斯推断的发展趋势. 此外, 近年来一些其他方法也被用于计算后验分布, 例如变分贝叶斯推断(Variational Bayesian inference)(Hao, 2016), 这将有助于改进系统发育推断的算法.

结合系统发育分析方法的发展现状, 目前推荐的系统发育分析流程是: 首先, 检查用于系统发育树重构的基因组和标记蛋白的质量; 其次, 在原始蛋白序列基础上, 利用贝叶斯CAT+GTR进化模型进行系统发育重构. 同时, 对引起长枝吸引效应或者GC偏差的物种进行选择性的删除, 以消除趋同性造成的系统发育的假阳性.

5.2.3 其他算法

理论上, 深度学习技术可以改进相关算法, 这将成为未来的重要研究领域. 新的统计理论方法以及深度学习等技术在数据处理流程中的应用, 都将进一步提高数据处理的效率和准确性, 为后续的古菌比较基因组学和功能基因组学分析奠定基础(Arango-Argoty等, 2018; Quince等, 2017; Roumpeka等, 2017). 对古菌宏基因组数据的后续研究的主要目标是获得关于古菌的类群、基因功能、以及在进化上所处的位置等信息, 从而更好地分析其在真核细胞功能演化过程中起到的作用(Thomas等, 2012). 生物信息技术的不断发展, 深度学习技术的应用, 统计理论与方法的优化, 都将促进古

菌的研究, 从而加快生命演化历程的研究, 推动真核细胞功能演化的研究.

5.3 培养和分离技术促进古菌研究发展

除了宏基因组技术和方法, 富集、分离和纯培养技术在可培养微生物的生理学研究中也发挥着关键作用. Asgard品系反应了一个共同的问题, 即在验证古菌的特有演化或功能特性时, 这些古菌无法被培养. Eloë-Fadros等(2016)的研究显示, 大约50%的古细菌尚未被鉴定, 并且大多数目前已知的物种尚未被分离. 研究人员一致认同分离那些已经作为环境基因组存在的古菌已经是迫在眉睫了(Zhang等, 2015), 而这会帮助我们更好地理解古菌生物学, 并建立更好的古菌生态学(生理学)的预测模型.

古菌研究已经进入微生物演化、代谢、生态功能和古生物地球化学研究领域的黄金时代. 最近在德国莱比锡举行的ISME 17会议(2018年8月12~17日)上举行的“古菌——真核生物演化途径”(张传伦和朱瑞新, 会议摘要)圆桌会议证明了这一点, 该会议还提供了未来的研究方向: 强调了需要构建更准确的系统发育树的拓扑结构; 同时, 要更关注“古菌宿主”或“古菌/真核共同的祖先”; 阐明真核形成的途径; 并发展分离纯培养的这些目前未知古菌的新方法.

致谢 范陆(共同通讯作者)同时作为共同第一作者贡献了本文的构思和写作.

参考文献

- 陈玉连, 潘杰, 周之超, 王风平, 李猛. 2017. 滨海深古菌的研究进展. 微生物学通报, 44: 1690–1698
- 李志江, 李海权, 刁现民. 2008. 基因水平转移的评判方法和转移方式研究进展. 遗传, 30: 1108–1114
- 孙松, 孙晓霞. 2017. 全面提升海洋综合探测与研究能力——中国科学院海洋先导专项进展. 海洋与湖沼, 48: 1132–1144
- 王治, 乐霖培, 张体操, 黄锦岭, 孙航. 2014. 水平基因转移在生物进化中的作用. 科学通报, 59: 2055–2064
- 左光宏, 郝柏林. 2017. 基因组时代的计算微生物学. 中国科学: 生命科学, 47: 159–170
- Adam P S, Borrel G, Brochier-Armanet C, Gribaldo S. 2017. The growing tree of Archaea: New perspectives on their diversity, evolution and ecology. ISME J, 11: 2407–2425
- Arango-Argoty G, Garner E, Pruden A, Heath L S, Vikesland P, Zhang

- L. 2018. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6: 23
- Archibald J M. 2008. The eocyte hypothesis and the origin of eukaryotic cells. *Proc Natl Acad Sci USA*, 105: 20049–20050
- Bankevich A, Nurk S, Antipov D, Gurevich A A, Dvorkin M, Kulikov A S, Lesin V M, Nikolenko S I, Pham S, Pribelski A D, Pyshkin A V, Sirotkin A V, Vyahhi N, Tesler G, Alekseyev M A, Pevzner P A. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 19: 455–477
- Betts H C, Puttick M N, Clark J W, Williams T A, Donoghue P C J, Pisani D. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol*, 2: 1556–1562
- Bodilis J, Nsique Meilo S, Cornelis P, De Vos P, Barray S. 2011. A long-branch attraction artifact reveals an adaptive radiation in pseudomonas. *Mol Biol Evol*, 28: 2723–2726
- Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol*, 54: 743–757
- Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. 2008. Mesophilic Crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol*, 6: 245–252
- Cai M, Liu Y, Zhou Z, Yang Y, Pan J, Gu J D, Li M. 2018. Asgard archaea are diverse, ubiquitous, and transcriptionally active microbes. bioRxiv, doi: [10.1101/374165](https://doi.org/10.1101/374165)
- Cao H, Auguet J C, Gu J D. 2013. Global ecological pattern of ammonia-oxidizing archaea. *PLoS ONE*, 8: e52853
- Cao J, Wang Q, Liu T, Peng N, Huang L. 2018. Insights into the post-translational modifications of archaeal Sis10b (Alba): Lysine-16 is methylated, not acetylated, and this does not regulate transcription or growth. *Mol Microbiol*, 109: 192–208
- Castelle C J, Banfield J F. 2018. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell*, 172: 1181–1197
- Chatton É. 1937. Titres et Travaux Scientifiques. Italy: Sette: Sottano
- Chen L X, Méndez-García C, Dombrowski N, Servín-Garcidueñas L E, Eloë-Fadrosch E A, Fang B Z, Luo Z H, Tan S, Zhi X Y, Hua Z S, Martínez-Romero E, Woyke T, Huang L N, Sánchez J, Peláez A I, Ferrer M, Baker B J, Shu W S. 2018. Metabolic versatility of small archaea Micrarchaeota and Parvarchaeota. *ISME J*, 12: 756–775
- Chu Y, Zhu Y, Chen Y, Li W, Zhang Z, Liu D, Wang T, Ma J, Deng H, Liu Z J, Ouyang S, Huang L. 2016. aKMT Catalyzes Extensive Protein Lysine Methylation in the Hyperthermophilic Archaeon *Sulfolobus islandicus* but is Dispensable for the Growth of the Organism. *Mol Cell Proteomics*, 15: 2908–2923
- Clark S C, Egan R, Frazier P I, Wang Z. 2013. ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, 29: 435–443
- Cornet L, Meunier L, Van Vlierberghe M, Léonard R R, Durieu B, Lara Y, Misztak A, Sirjacobs D, Javaux E J, Philippe H, Wilmette A, Baurain D. 2018. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLoS ONE*, 13: e0200323
- Cox C J, Foster P G, Hirt R P, Harris S R, Embley T M. 2008. The archaeobacterial origin of eukaryotes. *Proc Natl Acad Sci USA*, 105: 20356–20361
- Cui H L, Lu Z Z, Li Y, Zhou Y. 2017. *Salinirussus salinus* gen. nov., sp. nov., isolated from a marine solar saltern. *Int J Syst Evol Microbiol*, 67: 3622–3626
- Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet*, 13: e1006810
- Da Cunha V, Gaia M, Nasir A, Forterre P. 2018. Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet*, 14: e1007215
- Dagan T, Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA*, 104: 870–875
- Darwin C. 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray
- Dayhoff M O, Schwartz R M, Orcutt B C. 1978. A model of evolutionary change in proteins. In: Dayhoff M O, ed. Atlas of Protein Sequence and Structure. National Biomed Res Foundation, 5: 345–352
- DeLong E F. 1992. Archaea in coastal marine environments. *Proc Natl Acad Sci USA*, 89: 5685–5689
- Ding J Y, Chen S C, Lai M C, Liao T L. 2017. Haloterrigena mahii sp. nov., an extremely halophilic archaeon from a solar saltern. *Int J Syst Evol Microbiol*, 67: 1333–1338
- Doolittle W F. 1999. Phylogenetic classification and the universal tree. *Science*, 284: 2124–2128
- Doolittle W F. 2000. Uprooting the tree of life. *Sci Am*, 282: 90–95
- Doolittle W F, Brunet T D P. 2016. What is the tree of life? *PLoS Genet*, 12: e1005912
- Drummond A J, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7: 214
- Dyall S D, Yan W, Delgadillo-Correa M G, Lunceford A, Loo J A, Clarke C F, Johnson P J. 2004. Non-mitochondrial complex I

- proteins in a hydrogenosomal oxidoreductase complex. *Nature*, 431: 1103–1107
- Eloe-Fadrosch E A, Ivanova N N, Woyke T, Kyrpides N C. 2016. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol*, 1: 15032
- Eren A M, Esen Ö C, Quince C, Vineis J H, Morrison H G, Sogin M L, Delmont T O. 2015. Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ*, 3: e1319
- Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, Bryant D, Steel M A, Lockhart P J, Penny D, Martin W. 2004. A genome phylogeny for mitochondria among-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol*, 21: 1643–1660
- Forster P. 2015. The universal tree of life: An update. *Front Microbiol*, 6: 717
- Foster P G, Cox C J, Embley T M. 2009. The primary divisions of life: A phylogenomic approach employing composition-heterogeneous methods. *Phil Trans R Soc B*, 364: 2197–2207
- Foster P G, Hickey D A. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol*, 48: 284–290
- Fox G E, Stackebrandt E, Hespell R B, Gibson J, Maniloff J, Dyer T A, Wolfe R S, Balch W E, Tanner R S, Magrum L J, Zablen L B, Blakemore R, Gupta R, Bonen L, Lewis B J, Stahl D A, Luehrsens K R, Chen K N, Woese C R. 1980. The phylogeny of prokaryotes. *Science*, 209: 457–463
- Garg S G, Martin W F. 2016. Mitochondria, the cell cycle, and the origin of sex via a syncytial eukaryote common ancestor. *Genome Biol Evol*, 8: 1950–1970
- Garrity G M. 2016. A new genomics-driven taxonomy of Bacteria and Archaea: Are we there yet? *J Clin Microbiol*, 54: 1956–1963
- Ge H M, Zhang C L. 2016. Advances in GDGT research in Chinese marginal seas: A review. *Sci China Earth Sci*, 59: 1173–1186
- Ge H M, Zhang C L, Versteegh G J M, Chen L L, Fan D D, Dong L, Liu J J. 2016. Evolution of the East China Sea sedimentary environment in the past 14 kyr: Insights from tetraethers-based proxies. *Sci China Earth Sci*, 59: 927–938
- Ghods M, Hill C M, Astrovskaya I, Lin H, Sommer D D, Koren S, Pop M. 2013. De novo likelihood-based measures for comparing genome assemblies. *BMC Res Notes*, 6: 334
- Gogarten J P, Kibak H, Dittrich P, Taiz L, Bowman E J, Bowman B J, Manolson M F, Poole R J, Date T, Oshima T, Konishi J, Denda K, Yoshida M. 1989. Evolution of the vacuolar H⁺-ATPase: Implications for the origin of eukaryotes. *Proc Natl Acad Sci USA*, 86: 6661–6665
- Gogarten J P, Townsend J P. 2005. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol*, 3: 679–687
- Gould S B, Garg S G, Martin W F. 2016. Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system. *Trends Microbiol*, 24: 525–534
- Gribaldo S, Brochier-Armanet C. 2012. Time for order in microbial systematics. *Trends Microbiol*, 20: 209–210
- Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29: 1072–1075
- Guy L, Ettema T J G. 2011. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol*, 19: 580–587
- Guy L, Saw J H, Ettema T J G. 2014. The archaeal legacy of eukaryotes: A phylogenomic perspective. *Cold Spring Harb Perspect Biol*, 6: a016022
- Haeckel E. 1866. *Generelle Morphologie der Organismen*. Berlin: G. Reimer
- Han J, Wu L P, Liu X B, Hou J, Zhao L L, Chen J Y, Zhao D H, Xiang H. 2017. Biodegradation and biocompatibility of haloarchaea-produced poly(3-hydroxybutyrate-co-3-hydroxyvalerate) copolymers. *Biomaterials*, 139: 172–186
- Han W, Shen Y, She Q. 2014. Nanobiomotors of archaeal DNA repair machineries: Current research status and application potential. *Cell Biosci*, 4: 32
- Hao X. 2016. Variational Bayesian inference on phylogenetic trees, with applications to metagenomics. Doctoral Dissertation. Nebraska: The University of Nebraska-Lincoln
- Harish A. 2018. What is an archaeon and are the Archaea really unique? *PeerJ*, 6: e5770
- Hartman H, Fedorov A. 2002. The origin of the eukaryotic cell: A genomic investigation. *Proc Natl Acad Sci USA*, 99: 1420–1425
- He Y, Li M, Perumal V, Feng X, Fang J, Xie J, Sievert S M, Wang F. 2016. Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nat Microbiol*, 1: 16035
- Hilario E, Gogarten J P. 1993. Horizontal transfer of ATPase genes—The tree of life becomes a net of life. *Biosystems*, 31: 111–119
- Hrdy I, Hirt R P, Dolezal P, Bardonová L, Foster P G, Tachezy J, Martin Embley T. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature*, 432: 618–622
- Hua Z S, Qu Y N, Zhu Q, Zhou E M, Qi Y L, Yin Y R, Rao Y Z, Tian Y, Li Y X, Liu L, Castelle C J, Hedlund B P, Shu W S, Knight R, Li W J. 2018. Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. *Nat Commun*, 9: 2832
- Iwabe N, Kuma K I, Hasegawa M, Osawa S, Miyata T. 1989.

- Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA*, 86: 9355–9359
- Jia G D, Zhang J, Chen J F, Peng P A, Zhang C L L. 2012. Archaeal tetraether lipids record subsurface water temperature in the South China Sea. *Org Geochem*, 50: 68–77
- Jian H, Lesley J C. 2012. Eukaryotic signature proteins. *Proteomics Genomics Res*, 1: 2–8
- Jiang H, Huang Q, Dong H, Wang P, Wang F, Li W, Zhang C. 2010. RNA-based investigation of ammonia-oxidizing archaea in hot springs of Yunnan Province, China. *Appl Environ Microbiol*, 76: 4538–4541
- Jiao N, Herndl G J, Hansell D A, Benner R, Kattner G, Wilhelm S W, Kirchman D L, Weinbauer M G, Luo T, Chen F, Azam F. 2010. Microbial production of recalcitrant dissolved organic matter: Long-term carbon storage in the global ocean. *Nat Rev Microbiol*, 8: 593–599
- Jørgensen S L, Hannisdal B, Lanzén A, Baumberger T, Flesland K, Fonseca R, Ovreås L, Steen I H, Thorseth I H, Pedersen R B, Schleper C. 2012. Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proc Natl Acad Sci USA*, 109: 2846–2855
- Jørgensen S L, Thorseth I H, Pedersen R B, Baumberger T, Schleper C. 2013. Quantitative and phylogenetic study of the Deep Sea Archaeal Group in sediments of the Arctic mid-ocean spreading ridge. *Front Microbiol*, 4: 299
- Kalyaanamoorthy S, Minh B Q, Wong T K F, von Haeseler A, Jermini L S. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Meth*, 14: 587–589
- Kang D D, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3: e1165
- Kim M, Oh H S, Park S C, Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol*, 64: 346–351
- Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid A E, Kolinski A. 2016. Coarse-grained protein models and their applications. *Chem Rev*, 116: 7898–7936
- Ku C, Nelson-Sathi S, Roettger M, Sousa F L, Lockhart P J, Bryant D, Hazkani-Covo E, McInerney J O, Landan G, Martin W F. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*, 524: 427–432
- Kutschera U. 2016. Haeckel's 1866 tree of life and the origin of eukaryotes. *Nat Microbiol*, 1: 16114
- Lake J A. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature*, 331: 184–186
- Lake J A, Henderson E, Oakes M, Clark M W. 1984. Eocytes: A new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci USA*, 81: 3786–3790
- Lake J A, Rivera M C. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. *Mol Biol Evol*, 21: 681–690
- Lang S, Huang L. 2015. The *Sulfolobus solfataricus* GINS complex stimulates DNA binding and processive DNA unwinding by minichromosome maintenance helicase. *J Bacteriol*, 197: 3409–3420
- Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*, 7: S4
- Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, 21: 1095–1109
- Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol*, 62: 611–615
- Lasek-Nesselquist E, Gogarten J P. 2013. The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol Phylogenets Evol*, 69: 17–38
- Lasken R S, Stockwell T B. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol*, 7: 19
- Lettieri S, Zuckerman D M. 2012. Accelerating molecular monte carlo simulations using distance and orientation-dependent energy tables: Tuning from atomistic accuracy to smoothed “coarse-grained” models. *J Comput Chem*, 33: 268–275
- Liu B B, Zhao W Y, Chu X, Hozzein W N, Prabhu D M, Wadaan M A, Tang S K, Zhang L L, Li W J. 2014. Haladaptatus palliduribidus sp. nov., a halophilic archaeon isolated from saline soil samples in Yunnan and Xinjiang, China. *Antonie Van Leeuwenhoek*, 106: 901–910
- Liu G, She Q, Garrett R A. 2016. Diverse CRISPR-Cas responses and dramatic cellular DNA changes and cell death in pKEF9-conjugated *Sulfolobus* species. *Nucleic Acids Res*, 44: 4233–4242
- Liu X, Li M, Castelle C J, Probst A J, Zhou Z, Pan J, Liu Y, Banfield J F, Gu J D. 2018. Insights into the ecology, evolution, and metabolism of the widespread Woese archaeal lineages. *Microbiome*, 6: 102
- Liu Y, Zhou Z, Pan J, Baker B J, Gu J D, Li M. 2018. Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota. *ISME J*, 12: 1021–1031
- Loman N J, Constantinidou C, Chan J Z M, Halachev M, Sergeant M,

- Penn C W, Robinson E R, Pallen M J. 2012. High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. *Nat Rev Microbiol*, 10: 599–606
- Lu Z Z, Li Y, Zhou Y, Cui H L, Li Z R. 2017. *Halobacterium litoreum* sp. nov., isolated from a marine solar saltern. *Int J Syst Evol Microbiol*, 67: 4095–4099
- Ma K, Conrad R, Lu Y. 2013. Dry/Wet cycles change the activity and population dynamics of methanotrophs in rice field soil. *Appl Environ Microbiol*, 79: 4932–4939
- Margulis L. 1970. *Origin of Eukaryotic Cells*. New Haven: Yale University Press
- Martin W. 1999. Mosaic bacterial chromosomes: A challenge en route to a tree of genomes. *Bioessays*, 21: 99–104
- Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature*, 392: 37–41
- Martin W F. 2017. Too much eukaryote LGT. *Bioessays*, 39: 1700115
- Martin W F, Neukirchen S, Zimorski V, Gould S B, Sousa F L. 2016. Energy for two: New archaeal lineages and the origin of mitochondria. *Bioessays*, 38: 850–856
- Martin W F, Tielens A G M, Mentel M, Garg S G, Gould S B. 2017. The physiology of phagocytosis in the context of mitochondrial origin. *Microbiol Mol Biol Rev*, 81: doi: 10.1128/MMBR.00008-17
- Mayr E. 1998. Two empires or three? *Proc Natl Acad Sci USA*, 95: 9720–9723
- McInerney J O, O'Connell M J, Pisani D. 2014. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat Rev Microbiol*, 12: 449–455
- Meng J, Xu J, Qin D, He Y, Xiao X, Wang F. 2014. Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses. *ISME J*, 8: 650–659
- Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A, Sczyrba A, McHardy A C. 2018. AMBER: Assessment of Metagenome BinnERs. *Gigascience*, 7
- Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics*, 32: 1088–1090
- Moore B R, Höhna S, May M R, Rannala B, Huelsenbeck J P. 2016. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *Proc Natl Acad Sci USA*, 113: 9569–9574
- Narrowe A B, Spang A, Stairs C W, Caceres E F, Baker B J, Miller C S, Ettema T J G. 2018. Complex evolutionary history of translation elongation factor 2 and diphthamide biosynthesis in Archaea and Parabasalids. *Genome Biol Evol*, 10: 2380–2393
- Natarajan V P, Zhang X, Morono Y, Inagaki F, Wang F. 2016. A modified SDS-Based DNA extraction method for high quality environmental DNA from seafloor environments. *Front Microbiol*, 7: 986
- Nurk S, Bankevich A, Antipov D, Gurevich A A, Korobeynikov A, Lapidus A, Prjibelski A D, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel S R, Woyke T, McLean J S, Lasken R, Tesler G, Alekseyev M A, Pevzner P A. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol*, 20: 714–737
- Ochman H, Lawrence J G, Groisman E A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405: 299–304
- Pang S, Stones R J, Ren M M, Liu X G, Wang G, Xia H, Wu H Y, Liu Y, Xie Q. 2015. GPU MrBayes V3.1: MrBayes on graphics processing units for protein sequence data: Table 1. *Mol Biol Evol*, 32: 2496–2497
- Parfrey L W, Lahr D J G, Knoll A H, Katz L A. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci USA*, 108: 13624–13629
- Parks D H, Imelfort M, Skennerton C T, Hugenholtz P, Tyson G W. 2015. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*, 25: 1043–1055
- Parra G, Bradnam K, Korf I. 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23: 1061–1067
- Pearson A, Huang Z, Ingalls A E, Romanek C S, Wiegell J, Freeman K H, Smittenberg R H, Zhang C L. 2004. Nonmarine crenarchaeol in Nevada hot springs. *Appl Environ Microbiol*, 70: 5229–5237
- Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? *Curr Opin Genets Dev*, 8: 616–623
- Pisani D, Cotton J A, McInerney J O. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol*, 24: 1752–1760
- Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol*, 14: 615–623
- Quang le S, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24: 2317–2323
- Quince C, Walker A W, Simpson J T, Loman N J, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol*, 35: 833–844
- Rinke C, Schwientek P, Sczyrba A, Ivanova N N, Anderson I J, Cheng J F, Darling A, Malfatti S, Swan B K, Gies E A, Dodsworth J A, Hedlund B P, Tsiamis G, Sievert S M, Liu W T, Eisen J A, Hallam S J, Kyrpides N C, Stepanauskas R, Rubin E M, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499: 431–437
- Ronquist F, Huelsenbeck J P. 2003. MrBayes 3: Bayesian phylogenetic

- inference under mixed models. *Bioinformatics*, 19: 1572–1574
- Ronquist F, Teslenko M, van der Mark P, Ayres D L, Darling A, Höhna S, Larget B, Liu L, Suchard M A, Huelsenbeck J P. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*, 61: 539–542
- Roumpika D D, Wallace R J, Escalettes F, Fotheringham I, Watson M. 2017. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Front Genet*, 8: 23
- Sagan L. 1967. On the origin of mitosing cells. *J Theor Biol*, 14: 255–274
- Schleifer K H. 2009. Classification of Bacteria and Archaea: Past, present and future. *Syst Appl Microbiol*, 32: 533–542
- Schwartz R M, Dayhoff M O. 1978. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science*, 199: 395–403
- Seitz K W, Lazar C S, Hinrichs K U, Teske A P, Baker B J. 2016. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J*, 10: 1696–1705
- Sieber C M K, Probst A J, Sharrar A, Thomas B C, Hess M, Tringe S G, Banfield J F. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*, 3: 836–843
- Simão F A, Waterhouse R M, Ioannidis P, Kriventseva E V, Zdobnov E M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31: 3210–3212
- Song Z Q, Wang F P, Zhi X Y, Chen J Q, Zhou E M, Liang F, Xiao X, Tang S K, Jiang H C, Zhang C L, Dong H, Li W J. 2013. Bacterial and archaeal diversities in Yunnan and Tibetan hot springs, China. *Environ Microbiol*, 15: 1160–1175
- Soucy S M, Huang J, Gogarten J P. 2015. Horizontal gene transfer: Building the web of life. *Nat Rev Genet*, 16: 472–482
- Sousa F L, Neukirchen S, Allen J F, Lane N, Martin W F. 2016. Lokiarchaeon is hydrogen dependent. *Nat Microbiol*, 1: 16034
- Spang A, Eme L, Saw J H, Caceres E F, Zaremba-Niedzwiedzka K, Lombard J, Guy L, Ettema T J G. 2018. Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet*, 14: e1007080
- Spang A, Saw J H, Jørgensen S L, Zaremba-Niedzwiedzka K, Martijn J, Lind A E, van Eijk R, Schleper C, Guy L, Ettema T J G. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521: 173–179
- Sugiura A, Mattie S, Prudent J, McBride H M. 2017. Newly born peroxisomes are a hybrid of mitochondrial and ER-derived peroxisomes. *Nature*, 542: 251–254
- Sun M, Feng X, Liu Z, Han W, Liang Y X, She Q. 2018. An Orc1/Cdc6 ortholog functions as a key regulator in the DNA damage response in Archaea. *Nucleic Acids Res*, 46: 6697–6711
- Surkont J, Pereira-Leal J B. 2016. Are there Rab GTPases in archaea? *Mol Biol Evol*, 33: 1833–1842
- Susko E, Roger A J. 2007. On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol*, 24: 2139–2150
- Syvanen M. 1985. Cross-species gene transfer: Implications for a new theory of evolution. *J Theor Biol*, 112: 333–343
- Tatum E L, Lederberg J. 1947. Gene recombination in the bacterium *Escherichia coli*. *J Bacteriol*, 53: 673–684
- Thomas T, Gilbert J, Meyer F. 2012. Metagenomics—A guide from sampling to data analysis. *Microb Inform Exp*, 2: 3
- Timmis J N, Ayliffe M A, Huang C Y, Martin W. 2004. Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet*, 5: 123–135
- Torres-Montúfar A, Borsch T, Ochoterena H. 2018. When homoplasy is not homoplasy: Dissecting trait evolution by contrasting composite and reductive coding. *Syst Biol*, 67: 543–551
- Tully B J, Graham E D, Heidelberg J F. 2018. The reconstruction of 2631 draft metagenome-assembled genomes from the global oceans. *Sci Data*, 5: 170203
- Villanueva L, Schouten S, Damsté J S S. 2017. Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the ‘lipid divide’. *Environ Microbiol*, 19: 54–69
- Waddell P J, Steel M A. 1997. General time-reversible distances with unequal rates across sites: Mixing Γ and inverse gaussian distributions with invariant sites. *Mol Phylogenets Evol*, 8: 398–414
- Wang J X, Wei Y, Wang P, Hong Y, Zhang C L. 2015. Unusually low TEX 86 values in the transitional zone between Pearl River estuary and coastal South China Sea: Impact of changing archaeal community composition. *Chem Geol*, 402: 18–29
- Wang J X, Xie W, Zhang Y G, Meador T B, Zhang C L. 2017. Evaluating production of cyclopentyl tetraethers by Marine Group II Euryarchaeota in the Pearl River estuary and coastal South China Sea: Potential impact on the TEX86 paleothermometer. *Front Microbiol*, 8: 2077
- Wei Y, Wang J, Liu J, Dong L, Li L, Wang H, Wang P, Zhao M, Zhang C L. 2011. Spatial variations in archaeal lipids of surface water and core-top sediments in the South china sea and their implications for paleoclimate studies. *Appl Environ Microbiol*, 77: 7479–7489
- Weiss M C, Preiner M, Xavier J C, Zimorski V, Martin W F. 2018. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet*, 14: e1007518
- Whatley J M, John P, Whatley F R. 1979. From extracellular to intracellular: The establishment of mitochondria and chloroplasts. *Proc R Soc Lond B*, 204: 165–187

- Williams T A, Embley T M. 2014. Archaeal “dark matter” and the origin of eukaryotes. *Genome Biol Evol*, 6: 474–481
- Williams T A, Foster P G, Cox C J, Embley T M. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504: 231–236
- Williams T A, Foster P G, Nye T M W, Cox C J, Embley T M. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc R Soc B-Biol Sci*, 279: 4870–4879
- Williams T A, Szöllösi G J, Spang A, Foster P G, Heaps S E, Boussau B, Ettema T J G, Embley T M. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci USA*, 114: E4602–E4611
- Woese C R, Fox G E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA*, 74: 5088–5090
- Woese C R, Kandler O, Wheelis M L. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*, 87: 4576–4579
- Woese C R, Olsen G J. 1986. Archaeobacterial phylogeny: Perspectives on the urkingdoms. *Syst Appl Microbiol*, 7: 161–177
- Woyke T, Rubin E M. 2014. Evolution. Searching for new branches on the tree of life. *Science*, 346: 698–699
- Xie W, Luo H, Murugapiran S K, Dodsworth J A, Chen S, Sun Y, Hedlund B P, Wang P, Fang H, Deng M, Zhang C L. 2018. Localized high abundance of Marine Group II archaea in the subtropical Pearl River Estuary: Implications for their niche adaptation. *Environ Microbiol*, 20: 734–754
- Xie W, Wang F, Guo L, Chen Z, Sievert S M, Meng J, Huang G, Li Y, Yan Q, Wu S, Wang X, Chen S, He G, Xiao X, Xu A. 2011. Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME J*, 5: 414–426
- Xing X, Zhang L, Guo L, She Q, Huang L. 2014. Sulfolobus replication factor C stimulates the activity of DNA polymerase B1. *J Bacteriol*, 196: 2367–2375
- Yarza P, Yilmaz P, Pruesse E, Glöckner F O, Ludwig W, Schleifer K H, Whitman W B, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol*, 12: 635–645
- Yilmaz P, Parfrey L W, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner F O. 2014. The SILVA and “all-species living tree project (LTP)” taxonomic frameworks. *Nucl Acids Res*, 42: D643–D648
- Yu T, Li M, Niu M, Fan X, Liang W, Wang F. 2018a. Difference of nitrogen-cycling microbes between shallow bay and deep-sea sediments in the South China Sea. *Appl Microbiol Biotechnol*, 102: 447–459
- Yu T, Wu W, Liang W, Lever M A, Hinrichs K U, Wang F. 2018b. Growth of sedimentary *Bathymarchaeota* on lignin as an energy source. *Proc Natl Acad Sci USA*, 115: 6022–6027
- Zaremba-Niedzwiedzka K, Caceres E F, Saw J H, Bäckström D, Juzokaite L, Vancaester E, Seitz K W, Anantharaman K, Starnawski P, Kjeldsen K U, Stott M B, Nunoura T, Banfield J F, Schramm A, Baker B J, Spang A, Ettema T J G. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541: 353–358
- Zeng X, Zhang X, Jiang L, Alain K, Jebbar M, Shao Z. 2013. *Palaeococcus pacificus* sp. nov., an archaeon from deep-sea hydrothermal sediment. *Int J Syst Evol Microbiol*, 63: 2155–2159
- Zhang C L, Pancost R D, Sassen R, Qian Y, Macko S A. 2003. Archaeal lipid biomarkers and isotopic evidence of anaerobic methane oxidation associated with gas hydrates in the Gulf of Mexico. *Org Geochem*, 34: 827–836
- Zhang C L, Pearson A, Li Y L, Mills G, Wiegel J. 2006. Thermophilic temperature optimum for crenarchaeol synthesis and its implication for archaeal evolution. *Appl Environ Microbiol*, 72: 4419–4422
- Zhang C L, Peng P A, Zhao M X, Xie S C, Chen J F, Liu W G. 2018. Bio-organic geochemistry research in China: Advances, opportunities and challenges. *Sci China Earth Sci*, 61: 1775–1780
- Zhang C L, Xie W, Martin-Cuadrado A B, Rodriguez-Valera F. 2015. Marine Group II Archaea, potentially important players in the global ocean carbon cycle. *Front Microbiol*, 6: 1108
- Zhang C L, Li Y, Wall J D, Larsen L, Sassen R, Huang Y, Wang Y, Peacock A, White D C, Horita J, Cole D R. 2002. Lipid and carbon isotopic evidence of methane-oxidizing and sulfate-reducing bacteria in association with gas hydrates from the Gulf of Mexico. *Geology*, 30: 239–242
- Zhang G, Zhang F, Ding G, Li J, Guo X, Zhu J, Zhou L, Cai S, Liu X, Luo Y, Zhang G, Shi W, Dong X. 2012. Acyl homoserine lactone-based quorum sensing in a methanogenic archaeon. *ISME J*, 6: 1336–1344
- Zhang J, Jiao S, Lu Y. 2018. Biogeographic distribution of bacterial, archaeal and methanogenic communities and their associations with methanogenic capacity in Chinese wetlands. *Sci Total Environ*, 622–623: 664–675
- Zhang J, Li W, Wang J, Qin M, Wu L, Yan Z, Xu W, Zuo G, Wang W. 2009. Protein folding simulations: From coarse-grained model to all-atom model. *IUBMB Life*, 61: 627–643
- Zhang L M, Offre P R, He J Z, Verhamme D T, Nicol G W, Prosser J I. 2010. Autotrophic ammonia oxidation by soil thaumarchaea. *Proc Natl Acad Sci USA*, 107: 17240–17245

- Zhaxybayeva O, Doolittle W F. 2011. Lateral gene transfer. *Curr Biol*, 21: R242–R246
- Zheng X, Feng N, Li D, Dong X, Li J. 2017. New molecular insights into an archaeal RNase J reveal a conserved processive exoribonucleolysis mechanism of the RNase J family. *Mol Microbiol*, 106: 351–366
- Zhou Z, Liu Y, Li M, Gu J D. 2018a. Two or three domains: A new view of tree of life in the genomics era. *Appl Microbiol Biotechnol*, 102: 3049–3058
- Zhou Z, Pan J, Wang F, Gu J D, Li M. 2018b. Bathyarchaeota: Globally distributed metabolic generalists in anoxic environments. *FEMS Microbiol Rev*, 42: 639–655
- Zimorski V, Ku C, Martin W F, Gould S B. 2014. Endosymbiotic theory for organelle origins. *Curr Opin Microbiol*, 22: 38–48
- (责任编辑: 李超)