植物学报 *Chinese Bulletin of Botany* 2025, **60** (1): 1–16, www.chinbullbotany.com
doi: 10.11983/CBB24034 cstr: 32102.14.CBB24034

·研究论文·

维管植物质体DNA数据在物种和区域上的空缺研究

邓言[1, 2], 鲁丽敏[2, 3], 张强[4*], 陈之端[2, 3], 胡海花[2, 3*]

[1]广西师范大学生命科学学院, 桂林 541006; [2]中国科学院植物研究所, 植物多样性与特色经济作物全国重点实验室, 系统与进化植物学重点实验室, 北京 100093; [3]国家植物园, 北京 100093; [4]广西壮族自治区中国科学院广西植物研究所, 广西喀斯特植物保育与恢复生态学重点实验室, 桂林 541006

摘要 在植物大数据时代, 测序数据成为众多生物学研究的重要基础, 了解测序数据的现状有利于更好地利用这些数据。质体DNA数据因易获取、单亲遗传及变异速率适中而被广泛应用。基于GenBank公共数据库全面评估和分析了全世界维管植物质体DNA数据取样情况, 结果表明, 仅有33.75%的维管植物种类已测序。已测序物种在不同类群间取样不均衡, 缺失率大致与类群多样性呈显著正相关, 其中缺失最严重的目和科分别是盔被花目(Paracryphiales)、胡椒目(Piperales)和五桠果目(Dilleniales), 以及霉草科(Triuridaceae)、五膜草科(Pentaphragmataceae)和黄眼草科(Xyridaceae)。在地理空间上, 维管植物数据缺失程度从赤道向两极递减, 且生物多样性高的地区缺失更严重, 包括多个生物多样性热点地区。此外, 各地区特有种的数据普遍缺失严重。基于上述结果, 建议针对分子数据缺失程度较高的类群和生物多样性高的地区进行重点采集和测序, 尤其注重对特有种补充取样, 以增加这些类群遗传数据的代表性。

关键词 质体DNA, 维管植物, 数据缺失, 植物大数据, GenBank

邓言, 鲁丽敏, 张强, 陈之端, 胡海花 (2025). 维管植物质体DNA数据在物种和区域上的空缺研究. 植物学报 **60**, 1–16.

随着测序技术的飞速发展, 生物学研究已迈入大数据时代, DNA分子数据成为众多研究的重要基础, 在分子系统学、生物地理学和生态学等研究领域发挥重要作用。现有分子数据库已经积累了大量数据, 如世界上最大的DNA数据库GenBank (https://www.ncbi.nlm.nih.gov/)拥有的分子序列数据高达2亿多条。然而, 目前分子数据的缺失情况较为严重。研究表明, 仅有17%的绿色植物种类同时拥有遗传和分布等在内的基础数据, 31%的绿色植物种类在GenBank中有分子数据(Cornwell et al., 2019)。目前, 对公共数据库中分子数据的取样情况进行全面评估较少, 阻碍了分子数据在相关领域的应用。

数据缺失导致取样代表性不足, 数据取样的偏差可能会导致分析结果的偏差, 产生错误的结论。以系统发生树重建为例, 在建树时某些关键类群未被取样, 可能会得出错误的系统发生关系。例如, 在早期关于无油樟(*Amborella trichopoda*)是否是所有被子植物姐妹种的讨论中, Goremykin等(2003)选用13个类群的质体基因组进行系统发生树重建, 认为无油樟并非所有被子植物的姐妹种, 甚至不是基部类群; 但Soltis和Soltis (2004)增加基部类群的取样后支持无油樟是所有被子植物的姐妹种, 认为Goremykin等(2003)的结果是取样不足导致。除数据取样代表性不足造成结果偏差以外, 数据取样偏差也会影响对结果的判断。在时间估算方面, Linder等(2005)和Schulte (2013)分别选取非洲的帚灯草科(Restionaceae)和美洲鬣蜥科(Iguanidae)测试了物种取样率对分化时间估算的影响, 发现取样率较低会低估物种的分化时间。在评估生物多样性时, Park等(2018)选用菊科(Asteraceae)、菟丝子属(*Cuscuta*)、禾本目(Poales)和水龙骨亚纲(Polypodiidae) 4个类群分析了取样率对系统发生多样性(phylogenetic diversity)评估的影响, 经过对比真实的和模拟的系统发生树, 发现不完全取样会低估系统发生多样性。研究表明, 分子数据的取样可能在类群和区域上存在偏差。例如, Folk等(2018)将蔷薇类(rosids)主要分支已测序物种的比率

收稿日期: 2024-03-06; 接受日期: 2024-05-27
基金项目: 国家自然科学基金(No.32200190, No.32122009)
* 通讯作者。E-mail: qiangzhang04@126.com; huhh@ibcas.ac.cn

映射到系统发生树上，发现蔷薇类的分子数据主要集中在经济价值高和以温带分布为主的分支上。Hu等(2020)对中国维管植物物种生命之树的取样分析表明，横断山、青藏高原西部和新疆西部的被子植物存在较大的数据空缺。因此，了解已有数据中的分子数据现状及空间格局取样情况，识别数据空缺，对于依赖系统发生重建的生物学研究具有重要意义。

质体是植物细胞特有的细胞器，依据所含色素的类型和色素有无，可分为叶绿体、有色体和白色体。大多数陆地植物的质体基因组在结构、基因数目和序列上相对保守(Raman and Park, 2015)。典型的质体基因组结构由2个反向重复序列以及大单拷贝区和小单拷贝区组成，在一个细胞内通常具有高拷贝数(Yurina et al., 2017)，基因组大小为120–190 kb (Wicke et al., 2011; Yu et al., 2014)，包含100–150个基因(Ortelt and Link, 2014)。质体DNA具有进化速率适中、易测序及单亲遗传等特点，在DNA条形码、植物系统发生树重建和生物地理学研究中广泛应用(Nock et al., 2011; 张韵洁和李德铢, 2011; 胡颖等, 2019)。在植物系统发生树重建时，无论是针对单个类群的分子系统发生(Yao et al., 2021; Mo et al., 2022; Lian et al., 2023)，还是群落水平、区域尺度乃至全球植物生命之树重建(Smith and Brown, 2018; Janssens et al., 2020)，质体DNA数据都是重要的基础。较早的被子植物分子系统发生树是基于质体分子标记*rbcL*构建的(Chase et al., 1993)。Li等(2021)则基于质体基因组数据重建了第一棵取样为全部被子植物科的生命之树。由于质体DNA在应用时具有多种优点，因此公共数据库中质体DNA数据的积累越来越多，应用时其在取样代表性上也具有明显优势。目前，在GenBank中超过十分之一的陆地植物的分子数据是质体DNA数据，约有170万条(截至2024年5月)。然而，少有研究对现有公共数据库中质体DNA数据测序物种和基因的取样情况进行评估。

本研究基于目前最大的分子序列公共数据库GenBank，对数据库中维管植物质体DNA数据已测序物种和基因进行全面评估，以期明确以下问题：(1) 现阶段质体所有DNA片段已测序物种比例，明确哪些质体分子标记覆盖的物种最多；(2) 维管植物目和科的质体DNA已测序物种比例，识别数据缺失严重的目和科；(3) 已测序质体DNA的物种在地理空间上的分布格局，识别数据缺失严重的区域。本研究为进一步开发和利用质体DNA数据奠定了理论基础，推动了分类学、分子系统发生学、生物地理学和生态学等学科的发展。

1 研究方法

1.1 数据获取

利用Python第三方库BarcodeFinder 0.9.49版(https://github.com/wpwupingwp/barcodefinder)，从GenBank数据库获取所有维管植物包含注释信息的质体DNA分子序列，Query检索式为“txid58023 [Organism:exp] AND (plants[filter] AND biomol_genomic [PROP] AND ddbj_embl_genbank[filter] AND is_nuccore [filter] AND(chloroplast[filter] OR plastid [filter]))”(截至2023年9月)，共获得原始分子序列1 585 509条。

1.2 数据清洗

按照以下步骤对原始数据进行清洗。(1) 参考相关研究对核苷酸序列长度的限制(Ran et al., 2018; Smith and Brown, 2018)，仅保留长度≥200 bp的分子序列；(2) 仅保留来源物种为现存维管植物的分子序列；(3) 依据每条数据的注释信息，保留注释为质体DNA的分子序列。经过清洗后，得到维管植物质体DNA数据共1 453 217条。

1.3 物种名称标准化

物种名称标准化主要参考WCVP版本12 (World Checklist of Vascular Plants) (https://doi.org/10.34885/jdh2-dr22) (Govaerts et al., 2021)。WCVP是目前收录植物名称最全和最新的数据库之一，且综合了最新的分类学处理结果。物种名称标准化步骤如下：首先根据拉丁名与命名人，利用Python第三方库pykew 0.1.3版(https://github.com/RBGKew/pykew)精准匹配，然后对匹配结果为多个和未匹配上的学名进行复查，重点检查是否存在命名人缩写不一致或拉丁名拼写错误。若某个物种在WCVP中无记录，则根据NCBI保留该学名，暂时作为接受名处理，这部分名称占所有接受名的6.02%。经过上述处理后，将所有名称按照统一标准进行处理，并将异名的数据归并至对应接

受名的记录下。对于种下阶元，我们对名称进行标准化处理后，将相应数据合并至对应接受名的物种下。

1.4 质体DNA片段取样情况

数据清洗后，对照所有质体DNA片段，按照注释信息统计每个DNA分子标记的物种数占比，即每个片段覆盖的物种数与有数据的物种数比值。最后，共统计了891种质体DNA基因和基因间区的取样情况。由于数据量大，我们保守地按照GenBank的注释保留了DNA分子标记信息，部分基因和基因间区的名称还有待进一步核实。以模式植物拟南芥(*Arabidopsis thaliana*)质体基因组为参考，统计不属于拟南芥质体基因或基因间区(共558种)的记录条数，发现这部分数据记录仅占所有记录的0.14%，不会对研究结果造成偏差。

1.5 质体DNA数据在目和科水平的缺失情况

在对维管植物质体DNA数据缺失情况进行统计时，分别在目和科水平上统计了维管植物各类群的质体DNA数据缺失率。在本研究中，类群的质体DNA数据缺失率是一个类群在GenBank中未有质体DNA数据的物种数与该类群物种总数的比值。对于科和目的定义，被子植物采用APG IV系统(The Angiosperm Phylogeny Group et al., 2016)，裸子植物采用Christenhusz系统(Christenhusz et al., 2011)，广义蕨类植物采用PPG I系统(PPG I, 2016)。根据质体DNA数据缺失率，按照缺失程度对维管植物各类群进行分类。质体DNA数据缺失率为100%记为完全缺失，≥75%且<100%记为严重缺失，≥50%且<75%记为重度缺失，≥25%且<50%记为部分缺失，>0%且<25%记为轻微缺失，0%记为无缺失。为探究各类群质体DNA数据缺失率与类群多样性的关系，我们对类群包括的物种总数与质体DNA数据缺失率的相关性进行了分析，利用Python第三方库Scipy 1.9.3版(https://scipy.org/)计算两者的Spearman相关系数。由于裸子植物仅有8目12科，样本量较小，我们未将其纳入相关性分析。

1.6 质体DNA数据缺失的空间格局

为识别分子数据采集充分与薄弱的区域，我们对各地区质体DNA数据缺失率进行了统计。在本研究中，一个地区的质体DNA数据缺失率为该地区在GenBank中无质体DNA数据记录的维管植物物种数与该地区维管植物物种总数的比值。本研究使用的分布数据来源于WCVP，该数据库包含全球维管植物的分布数据。我们进一步依据WCVP将全球划分为369个地区(Govaerts et al., 2021)。此外，我们还统计了各地区特有种的质体DNA数据缺失情况。特有种的判断标准依据WCVP的分布记录，仅分布于1个地区的维管植物为该地区的特有种。各地区特有种的质体DNA数据缺失率为该地区在GenBank中无质体DNA数据记录的特有种数与该地区特有种总数的比值。我们根据各地区质体DNA数据缺失率对缺失程度进行分类，划分标准与维管植物质体DNA数据缺失程度的标准相同。最后，每个地区的质体DNA数据缺失率和特有种质体DNA数据缺失率依据数据缺失程度用QGIS 3.32.3版(https://qgis.org/en/site/)在地图上展示。各地区维管植物物种总数和质体DNA数据缺失率以及各地区特有种总数与特有种质体DNA数据缺失率的Spearman相关性均使用Python第三方库Scipy计算完成，计算时去除无特有维管植物分布的地区。

2 结果与分析

2.1 质体DNA数据每个片段的取样情况

本研究对GenBank中维管植物质体DNA数据进行了全面评估和分析，经过清洗和名称标准化后的数据包含139 005种维管植物，占所有维管植物的33.75%，其中被子植物131 220种，来源于64目417科12 332属，裸子植物1 154种，来源于8目12科86属，广义蕨类植物6 631种，来源于14目51科231属(表1)。

表1 维管植物、被子植物、裸子植物和广义蕨类植物的取样物种数和取样率

Table 1 Number of sampled species and sampling ratios in tracheophytes, angiosperms, gymnosperms, and pteridophytes

Group	Number of sampled species	Sampling ratio (%)
Tracheophytes	139005	33.75
Angiosperms	131220	33.16
Gymnosperms	1154	75.92
Pteridophytes	6631	45.35

在维管植物中，覆盖物种最多的10个质体DNA片段是*matK* (*n*=76 596)、*trnL-trnF* (*n*=70 180)、*rbcL* (*n*=65 632)、*trnL* (*n*=54 194)、*trnH-psbA* (*n*=42 881)、*trnK* (*n*=36 715)、*ndhF* (*n*=35 911)、*rps16* (*n*=32 874)、*rpl16* (*n*=27 396)和*atpB* (*n*=22 771)，当质体DNA片段增加到前7个时，所覆盖的物种在有数据的维管植物中占比超过90%，当质体DNA片段增加到前21个时，有数据的物种比例基本不会增加(图1A; 附录1)。在被子植物中，覆盖物种最多的10个质体DNA片段与维管植物相同，前5个质体DNA片段所覆盖的物种在有数据的被子植物中占比超过90%，当质体DNA片段增加到前18个时，有数据的被子植物的比例基本不会增加(图1B; 附录1)。在裸子植物中，覆盖物种最多的10个质体DNA片段是*rbcL* (*n*=989)、*matK* (*n*=939)、*trnL* (*n*=737)、*trnL-trnF* (*n*=674)、*rpoC* (*n*=632)、*trnH-psbA* (*n*=622)、*trnK* (*n*=596)、*rps4* (*n*=592)、*psbB* (*n*=546)和*rpl16* (*n*=539)，前3个质体DNA片段所覆盖的物种在有数据的裸子植物中占比超过90%，当质体DNA片段增加到前7个时，有数据的裸子植物比例基本不会增加(图1C; 附录1)。在广义蕨类植物中，覆盖最多物种的10个质体DNA片段是*rbcL* (*n*=5 708)、*trnL-trnF* (*n*=4 267)、*trnS-rps4* (*n*=3 067)、*rps4* (*n*=2 714)、*atpB* (*n*=2 171)、*trnL* (*n*=1 884)、*trnH-psbA* (*n*=1 782)、*atpA* (*n*=1 518)、*trnG-trnR* (*n*=1 491)和*matK* (*n*=1 462)，前2个质体DNA片段所覆盖的物种在有数据的广义蕨类植物中占比超过90%，当质体DNA片段增加到前13个时，有数据的蕨类植物比例基本不会增加(图1D; 附录1)。

2.2 维管植物质体DNA数据在目和科水平的取样情况

维管植物质体DNA数据在目水平的取样情况普遍表

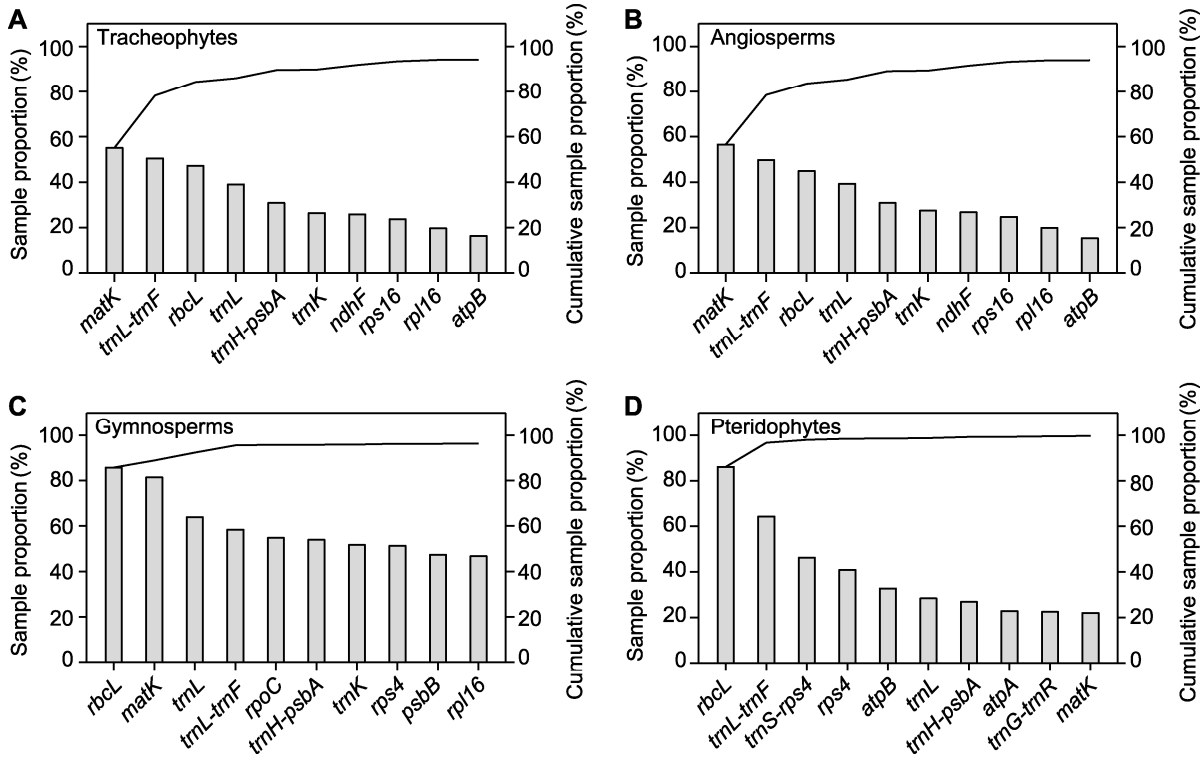


图1 GenBank中维管植物**(A)**、被子植物**(B)**、裸子植物**(C)**和广义蕨类植物**(D)**取样率排名前10的质体DNA分子标记的取样率及累积取样率

柱状图表示每个分子标记的取样率，黑色实线表示累积取样率。

Figure 1 Sample proportion and cumulative sample proportion of the top ten plastid DNA molecular markers in GenBank for tracheophytes **(A)**, angiosperms **(B)**, gymnosperms **(C)**, and pteridophytes **(D)**

Histograms represent the sample proportion of each molecular marker, and black lines represent the cumulative sample proportion.

现为高缺失率(图2; 附表1)。对于所有维管植物, 缺失程度表现为重度缺失的目最多(48个), 这些目包含的物种占所有维管植物的96.38%, 质体DNA数据缺失率最高的前5个目是盔被花目(Paracryphiales, 85.71%)、胡椒目(Piperales, 82.81%)、五桠果目(Dilleniales, 82.65%)、黄漆姑目(Vahliales, 80.00%)和水螅花目(Metteniusales, 77.63%), 有6个目无缺失(即所有物种均有分子序列), 分别为无油樟目(Amborellales)、红珊藤目(Berberidopsidales)、银杏目(Ginkgoales)、无叶莲目(Petrosaviales)、昆栏树目(Trochodendrales)和百岁兰目(Welwitschiales)。在被子植物中, 缺失程度表现为重度缺失的目最多(42个), 这些目包含的物种占所有被子植物的97.09%, 除被子植物基部类群外, 木兰类(Magnoliids)、单子叶植物(Monocots)、基部真双子叶植物(Basal eudicots)、超蔷薇类(Superrosids)和超菊类(Superasterids)等分支均存在重度缺失的目, 如木兰类木兰目(Magnoliales, 61.39%)、单子叶植物禾本目(Poales, 58.94%)和超蔷薇类蔷薇目(Rosales, 66.76%)。在被子植物中, 数据缺失率最高的前5个目与维管植物相同, 有4个目无缺失。在裸子植物中, 缺失程度表现为部分缺失的目最多(4个), 这些目包含的物种占所有裸子植物的55.72%, 其中数据缺失率最高的前5个目是买麻藤目(Gnetales, 42.37%)、麻黄目(Ephedrales, 37.63%)、松目(Pinales, 31.94%)、南洋杉目(Araucariales, 26.24%)和柏目(Cupressales, 22.11%), 2个目无缺失。在广义蕨类植物中, 缺失程度表现为部分缺失的目最多(8个), 这些目包含的物种占所有广义蕨类植物的12.46%, 数据缺失率最高的前5个目是里白目(Gleicheniales, 69.50%)、卷柏目(Selaginellales, 61.95%)、水龙骨目(Polypodiales, 56.97%)、膜蕨目(Hymenophyllales, 55.57%)和松叶蕨目(Psilotales, 52.63%), 取样率最高的是槐叶蘋目(Salviniales, 36.05%)。

维管植物质体DNA数据在科水平的取样情况与目水平基本一致, 普遍表现为高缺失率(附表2; 附图1)。对于所有维管植物, 缺失程度表现为重度缺失的科最多, 包括210个科, 这些科包含的物种占所有维管植物的66.21%, 数据缺失率最高的前5个科是霉草科(Triuridaceae, 95.45%)、五膜草科(Pentaphragmataceae, 93.94%)、黄眼草科(Xyridaceae, 93.86%)、花柱草科(Stylidiaceae, 93.41%)和毒鼠子科(Dichapetalaceae, 91.18%), 有74个科无缺失, 包括南茱萸科(Griseliniaceae)、罗伞蕨科(Matoniaceae)和金松科(Sciadopityaceae)等。在被子植物中, 缺失程度表现为重度缺失的科最多(192个), 这些科包含的物种占所有被子植物的66.81%, 被子植物各大分支均存在重度缺失的科。被子植物质体DNA数据缺失率最高的前5个科与维管植物相同, 有65个科无缺失, 包括南茱萸科(Griseliniaceae)、三白草科(Saururaceae)和鱼篓藤科(Ripogonaceae)等。在裸子植物中, 缺失程度表现为轻微缺失的科最多(5个), 这些科包含的物种占所有裸子植物的46.84%, 数据缺失率最高的前5个科是买麻藤科(Gnetaceae, 42.37%)、麻黄科(Ephedraceae, 37.63%)、松科(Pinaceae, 31.94%)、罗汉松科(Podocarpaceae, 27.60%)和柏科(Cupressaceae, 22.92%), 有3个科无缺失, 分别是金松科(Sciadopityaceae)、银杏科(Ginkgoaceae)和百岁兰科(Welwitschiaceae)。广义蕨类植物缺失程度表现为部分缺失的科最多(23个), 这些科包含的物种占所有广义蕨类植物的33.98%, 数据缺失率最高的前5个科是金星蕨科(Thelypteridaceae, 80.31%)、里白科(Gleicheniaceae, 73.51%)、铁角蕨科(Aspleniaceae, 67.74%)、袋囊蕨科(Saccolomataceae, 66.67%)和蹄盖蕨科(Athyriaceae, 64.92%), 有6个科无缺失, 分别是伞序蕨科(Thyrsopteridaceae)、半网蕨科(Hemidictyaceae)、垫囊蕨科(Culcitaceae)、柱囊蕨科(Loxsomataceae)、链脉蕨科(Desmophlebiaceae)和罗伞蕨科(Matoniaceae)。

在目水平, 所有维管植物、被子植物和广义蕨类植物各目的物种总数与质体DNA数据缺失率呈显著正相关(所有维管植物: Spearman'r=0.55, P<0.01, 附图2A; 被子植物: Spearman'r=0.40, P<0.01, 附图2B; 广义蕨类植物: Spearman'r=0.58, P<0.05, 附图2C)。在科水平, 所有维管植物、被子植物和广义蕨类植物各科的物种总数与质体DNA数据缺失率呈显著正相关(所有维管植物: Spearman'r=0.60, P<0.01, 附图2D; 被子植物: Spearman'r=0.59, P<0.01, 附图2E; 广义蕨类植物: Spearman'r=0.60, P<0.01; 附图2F)。

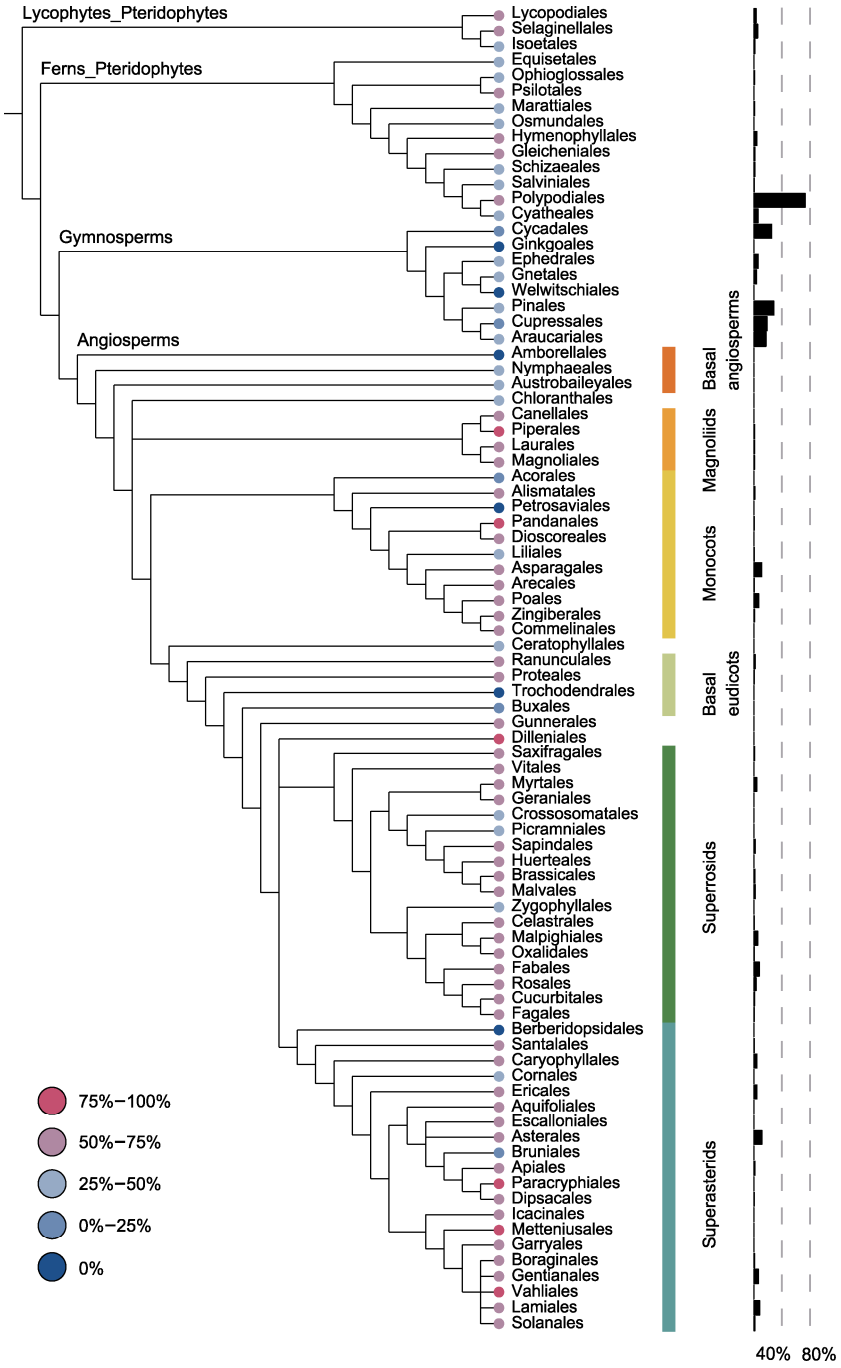


图2 维管植物目水平质体DNA数据缺失情况和每个目的多样性占比

系统发生树末端节点的颜色表示缺失情况，右侧的黑色柱状图表示每个目分别在被子植物、裸子植物和广义蕨类植物中的多样性占比。维管植物系统树改自APG IV (The Angiosperm Phylogeny Group et al., 2016)、Christenhusz (Christenhusz et al., 2011)和PPG I (PPG I, 2016)。被子植物主要分支用不同颜色条带表示。

Figure 2 Missing data in plastid DNA and the proportions of species diversity at the ordinal level of tracheophytes

Colored circles at the terminal of phylogenetic tree represent the proportion of missing data in plastid DNA and black histograms on the right represent the proportion of species diversity of each order in angiosperms, gymnosperms, and pteridophytes, respectively. The tracheophyte phylogenetic tree was modified from APG IV (The Angiosperm Phylogeny Group et al., 2016), Christenhusz (Christenhusz et al., 2011), and PPG I (PPG I, 2016). The major clades of angiosperms are indicated with bars of different colors.

2.3 维管植物质体DNA数据缺失的空间格局

维管植物质体DNA数据缺失在地理空间上表现为不均匀分布，缺失程度由赤道向两极递减(图3A)。质体DNA数据缺失程度为严重缺失的地区有1个，即新几内亚地区。缺失程度为重度缺失的地区有43个，主要为赤道附近的热带地区，少部分在欧洲，大部分地区的质体DNA数据缺失率在50%–60%之间，超过60%的地区仅有11个，主要位于东南亚和北欧；缺失程度为部分缺失的地区最多，共有268个。缺失程度为轻微缺失的地区有55个，主要位于极地附近的群岛以及北美。全球各地区维管植物质体DNA数据缺失率与地区维管植物物种数呈显著正相关(Spearman'r=0.61, P<0.01，图3B)。

全球维管植物特有种数据缺失程度普遍比较严重(图3C)。有50个地区的维管植物质体DNA数据缺失程度为完全缺失，主要分布在北美东北部和北极附近。大多数地区的特有种数据缺失程度为严重缺失，共有204个。缺失程度为重度缺失的地区有76个，大部分在南、北回归线附近。缺失程度为轻微缺失的地区仅有14个，零星分布在北美和非洲。仅有3个地区表现为无缺失，分别为科威特、爱德华王子群岛和密西西比地区。各地区特有种质体DNA数据缺失率与特有种总数呈负相关(Spearman'r=–0.25, P<0.01，图3D)。

3 讨论

本研究基于GenBank数据库对维管植物现有质体DNA数据取样情况按照分子标记片段、类群(目和科)和空间分布进行了深入细致的分析。整体上，GenBank数据库中维管植物质体DNA数据存在较严重的缺失，仅33.75%的维管植物至少有1条质体DNA分

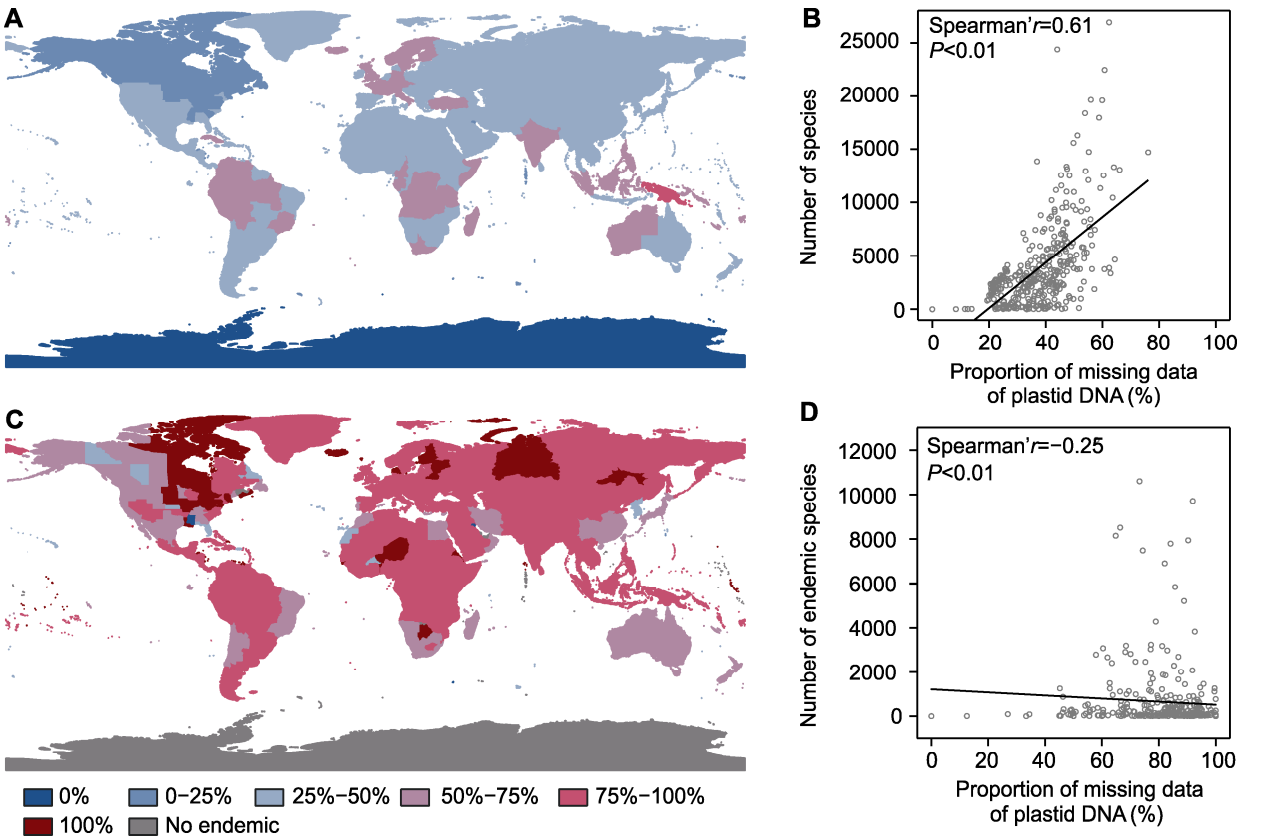


图3 全球维管植物**(A)**和特有维管植物**(C)**的质体DNA数据缺失情况的空间格局以及地区物种多样性**(B)**和特有种多样性**(D)**分别与各自的质体DNA数据缺失率的Spearman相关性

Figure 3 Spatial patterns of missing data in plastid DNA of global tracheophytes **(A)** and endemic tracheophytes **(C)**, and Spearman correlations between regional species diversity and missing data in plastid DNA for tracheophytes **(B)** and endemic tracheophytes **(D)**, respectively.

子序列。Cornwell等(2019)指出，仅31%的绿色植物在GenBank中拥有至少1条DNA分子数据。目前测序技术发展迅速，测序工作多针对已经取样的物种，尤其是经济价值高或模式物种的重新测序或深度测序(Iorizzo et al., 2016; Coe et al., 2023)，而对未测序物种的取样增加较少。尽管对已取样物种开展更多的测序工作有助于深入揭示物种内个体与群体之间的差异，但未测序物种作为生命之树独一无二且必不可少的分支，在资源利用、生物多样性保护和未来环境变化响应等领域具有巨大的潜力，未来需要注重增加未取样物种的采集和测序，以增强物种代表性。

在本研究中，覆盖最多物种的10个质体DNA片段为*matK* (位于*trnK*基因的内含子中)、*trnL-trnF*、*rbcL*、*trnL*、*trnH-psbA*、*trnK*、*ndhF*、*rps16*、*rpl6*和*atpB*，覆盖的物种占有数据的维管植物比例超过90% (图1)。这些基因或基因间区在系统发生树构建、生物多样性评估和DNA条形码鉴定等方面具有广泛应用。*matK*、*trnK*、*rbcL*和*ndhF*常被用于构建某一类群或者区域乃至全球维管植物生命之树(Zanne et al., 2014; Chen et al., 2016; Hu et al., 2020)。越来越多的研究基于区域或全球尺度生命之树度量包含物种演化历史信息的生物多样性维度，即系统发生多样性(Forest et al., 2007; Lu et al., 2018; Hu et al., 2022; Liu et al., 2023)。*matK*、*rbcL*和*trnH-psbA*等也广泛应用于许多类群的DNA条形码研究(刘宇婧等, 2011)，用于区分近缘类群，进行物种鉴定(Liu et al., 2012; Lv et al., 2020; Jiang et al., 2023)以及居群水平的生物地理学研究(Ohi et al., 2003; Huang et al., 2013; Jafari et al., 2015)。这些基因或基因间区由于在居群和物种水平均具有适中的变异速率而被广泛应用。反之，当我们需要选择其它分子标记的数据以提高结果的分辨率时，也可以从这些测序最多物种的片段中选取候选分子标记。此外，本研究发现，DNA片段覆盖的物种并不一定随DNA片段的增多呈现持续显著增长(图1)。对于所有维管植物，当质体DNA片段增加到前7个时，所覆盖的物种在有数据的维管植物中占比超过90%，当质体DNA片段增加到前21个时，继续增加质体DNA片段，有数据的物种占比不会明显增加(图1A; 附录1)。因此，在使用质体DNA分子数据进行研究时，尤其是进行多基因联合建树，通过增加分子标记的数目扩大研究的取样范围是可行的，但从平衡取样代表性与计算效率的角度考虑，选用适量的DNA分子标记即可。需要注意的是，在实际选择哪些分子标记进行联合建树时，还需要考虑分子标记的变异速率(Christin et al., 2014; Zhang et al., 2020; 彭焕文和王伟, 2023)，即是否有足够多的系统发生信息位点解决类群间的系统发生关系，以及分子序列的测序质量等。

质体DNA数据缺失率在维管植物中存在类群异质性，整体表现为被子植物的缺失率高于广义蕨类植物和裸子植物。在目和科水平，对于所有维管植物和被子植物，大多数类群表现为重度缺失，而且除被子植物基部类群外，被子植物各大分支均存在重度缺失的目和科(图2，附图1；附表1，附表2)。对于裸子植物和广义蕨类植物，大多数类群分别表现为轻微缺失和部分缺失质体DNA数据。缺失率最高的目和科分别达到85.71% (盔被花目(Paracryphiales))和95.45% (霉草科(Triuridaceae))，均属于被子植物。类群所包含的物种数与相应质体DNA数据缺失率的相关性分析结果显示，二者呈显著正相关，表明维管植物各类群在目和科水平的缺失程度可能与类群的多样性有关(附图2)。在维管植物中，物种数超过10 000种的目或科，平均缺失率均超过65%，多样性最高的菊目(Asterales)和菊科(Asteraceae)质体DNA数据缺失率分别为74.36%和76.11%。而质体DNA数据无缺失的类群，物种数均未超过10个。以上结果表明，质体DNA数据缺失率可能受类群多样性影响，未测序的物种随着类群多样性的增加而增多，即多样性越高的类群可能表现出更高的数据缺失率。Rudbeck等(2022)的研究结果也证实已测序物种数曲线随类群多样性的增加趋于平缓。但类群多样性与质体DNA数据缺失率的关系也存在例外情况，多样性较高的类群质体DNA数据缺失率可能比多样性较低的类群更低。在目和科水平，质体DNA数据缺失程度为严重缺失(数据缺失率在75%–100%之间)的类群大部分多样性较低(物种数小于1 000)，而物种数超过10 000的类群多表现为重度缺失(缺失率为50%–75%)。例如，唇形目(Lamiales)、豆目(Fabales)和禾本目(Poales)的多样性均高于杜鹃花目(Ericales)、金虎尾目(Malpighiales)和龙胆目(Gentianales)，但前三者的质体DNA数据缺失率均低于后三者。这可能与类群的经济

价值有关, 唇形目含有许多药用和香料植物, 如黄芩(*Scutellaria baicalensis*)和迷迭香(*Rosmarinus officinalis*); 豆目和禾本目包含重要的农作物, 如大豆(*Glycine max*)、水稻(*Oryza sativa*)和小麦(*Triticum aestivum*), 这些类群更容易受到科研人员的关注和研究。Folk等(2018)统计的蔷薇类(Rosid) DNA数据在系统发生树上的分布也表现出温带分布和经济价值高的分支比热带分布和经济价值低的分支有更高的取样率。

维管植物质体DNA数据缺失率的空间格局大致表现为纬度梯度趋势, 缺失的严重程度从热带向两极递减(图3A)。各地区质体DNA数据缺失率与维管植物多样性呈显著正相关, 即维管植物种类越多的地区质体DNA数据缺失程度越严重(Spearman'*r*=0.61, *P*< 0.01) (图3B)。例如, 拥有全球物种丰富度最高的新几内亚地区(Cámara-Leret et al., 2020)是唯一质体DNA数据为严重缺失的地区, 这可能与采样不充分有关。维管植物质体DNA数据缺失率的空间格局与全球种子植物的系统发生关系认知(knowledge of phylogenetic relationships)格局(Rudbeck et al., 2022)基本相反, 后者通过统计种子植物系统发生研究中常用DNA片段(Hinchliff and Smith, 2014)的取样比例, 发现系统发生关系认知水平从赤道向两极递增, 即有数据的物种比例随纬度升高而增加。多数生物多样性热点地区(Myers et al., 2000)质体DNA数据的缺失程度为重度缺失, 如马达加斯加地区的质体DNA数据缺失率为66.10%; 少数热点地区虽表现为部分缺失, 但质体DNA数据缺失率基本偏高, 如加利福尼亚州为44.36%。斯堪的纳维亚半岛位于欧洲西北角, 处在高纬度地区, 该地区维管植物多样性并不突出, 但质体DNA数据缺失率较高, 这可能与近期对该区域部分类群进行了较大规模的分类修订有关, 基于系统发生关系认知的空间格局研究也得到相似的结果(Rudbeck et al., 2022)。Rudbeck等(2022)指出, 近年开展的分类修订(Tyler, 2017)中描述了瑞典植物区系的主要成分之一山柳菊属(*Hieracium*)的多个新种, 使该属物种大量增加, 造成瑞典植物区系中已有DNA数据的物种比例降低。特有维管植物质体DNA数据缺失率的空间格局未表现出明显的纬度梯度趋势, 大多数地区的数据缺失程度为严重缺失(图3C)。总体上, 各地区特有种质体DNA数据缺失率普遍较高, 说明在全球范围特有种的数据仍有较大空缺, 这可能与特有种的分布范围较小及野外采集难度大有关。全球各地区特有种总数与质体DNA数据缺失率呈负相关(Spearman'*r*=0.25, *P*<0.01) (图3D), 这可能是由于特有性较高的地区更加重视特有种的采集和测序工作; 而特有性较低的地区特有种受重视程度较低, 因而特有种质体DNA数据缺失率更高。特有种是区系独有的组分, 反映了区系独特的演化历史(He et al., 2018; Ye et al., 2023), 通常比广布种具有更高的灭绝风险(Vischi et al., 2004; Krause et al., 2015; Coelho et al., 2020), 加强对各地区特有维管植物的测序和研究十分必要。Rudbeck等(2022)对全球种子植物系统发生空间格局的研究表明, 增加对狭域分布的物种尤其是特有种的测序和研究, 有利于加深对地区系统发生关系的认识。

现今的生物学研究已迈入基因组学时代, 基因组相对于多基因片段提供的信息位点数量存在指数级差距。随着测序技术的发展, 测序成本降低且测序效率显著提高, 对细胞器基因组、转录组甚至全基因组进行测序研究增多(Kuang et al., 2011; Novikova et al., 2016; Li et al., 2019)。越来越多的研究基于质体基因组进行类群系统发生关系、演化历史、作物驯化及生物群落演化研究(Mo et al., 2022; Chen et al., 2023; Li et al., 2023)。然而, 受限于材料采集费用、测序成本和计算效率, 基因组研究多局限于单一类群, 基因组数据在应用上仍存在诸多挑战。在物种取样方面, 目前GenBank数据库中仅有13 000个质体基因组数据(Wang et al., 2024)。经过多年的积累, 公共数据库已积累了较为丰富的质体DNA数据, 在取样上具有显著优势。此外, 植物标本馆作为重要资源宝库(Kistler et al., 2020; 王露露等, 2023; Burbano and Gutaker, 2023), 可提供丰富的分子数据。植物标本中的DNA受制作及保存方法和年代久远的影响发生了不同程度的降解(Staats et al., 2011), 但是现今的测序技术能从中获取可用的DNA数据(Gutaker and Burbano, 2017; Lang et al., 2020), 为增加质体DNA分子标记, 尤其是常用分子标记提供了基础数据。目前, 在从植物标本中获取DNA数据方面取得了重要进展, 未来可进一步优化DNA提取和测序方法,

充分利用植物标本资源。综上，尽管基因组数据具有非常广阔的应用前景，但基于当前的数据积累现状，细致深入地了解质体DNA数据的空缺状况，从而充分利用和开发这些数据，并通过重点采集和从标本中获取分子数据补充取样，对生物学研究意义重大。需要注意的是，本研究仅基于GenBank公共数据库对全球维管植物质体DNA数据取样情况进行评估和分析，未将其它保存DNA序列的数据库或自建网站纳入分析，如中国国家生物信息中心(China National Center for Bioinformation, https://www.cncb.ac.cn/)和Dryad (https://datadryad.org/)等。因此，本研究可能在一定程度上高估了质体DNA数据缺失情况，未来有必要充分收集这些数据，以期真实反映质体DNA数据缺失情况。

4 结论

本研究基于GenBank数据库，全面深入地评估了维管植物质体DNA数据在物种和空间分布上的取样现状。结果表明，GenBank数据库中的维管植物质体DNA数据量非常可观，拥有质体DNA数据的维管植物物种总数超过10万种，但整体仍有较大的数据缺失。数据缺口主要体现在3个方面。(1) 有数据的维管植物在维管植物中占比较低(仅为1/3)；(2) 类群间取样不均衡，被子植物数据缺失最严重，其次是广义蕨类植物和裸子植物；(3) 地区间取样不均衡，数据缺失率从赤道向两极递减，生物多样性热点地区和特有种存在比较严重的数据缺失。分子数据作为生物学研究的重要基础，补充数据缺口对正确认识并保护生物多样性具有重要意义。基于本研究结果，我们建议对高缺失率类群以及生物多样性丰富的地区加强野外采集与物种测序工作，同时要注重对特有种的取样。

致谢 中国科学院植物研究所刘冰副研究员和杨宇昌博士(已毕业)指导数据分析，在此表示感谢！

作者贡献声明

邓言：设计研究，分析数据，撰写并修改论文；鲁丽敏：修改论文；张强：总体构思及设计研究，修改论文；陈之端：修改论文；胡海花：总体构思及设计研究，指导数据分析，撰写并修改论文。

参考文献

Burbano HA, Gutaker RM (2023). Ancient DNA genomics and the renaissance of herbaria. *Science* **382**, 59–63.

Cámara-Leret R, Frodin DG, Adema F, Anderson C, Appelhans MS, Argent G, Arias Guerrero S, Ashton P, Baker WJ, Barfod AS, Barrington D, Borosova R, Bramley GLC, Briggs M, Buerki S, Cahen D, Callmander MW, Cheek M, Chen CW, Conn BJ, Coode MJE, Darbyshire I, Dawson S, Dransfield J, Drinkell C, Duyfjes B, Ebihara A, Ezedin Z, Fu LF, Gideon O, Girmansyah D, Govaerts R, Fortune-Hopkins H, Hassemer G, Hay A, Heatubun CD, Hind DJN, Hoch P, Homot P, Hovenkamp P, Hughes M, Jebb M, Jennings L, Jimbo T, Kessler M, Kiew R, Knapp S, Lamei P, Lehnert M, Lewis GP, Linder HP, Lindsay S, Low YW, Lucas E, Mancera JP, Monro AK, Moore A, Middleton DJ, Nagamasu H, Newman MF, Nic Lughadha E, Melo PHA, Ohlsen DJ, Pannell CM, Parris B, Pearce L, Penneys DS, Perrie LR, Petoe P, Poulsen AD, Prance GT, Quakenbush JP, Raes N, Rodda M, Rogers ZS, Schuiteman A, Schwartsburd P, Scotland RW, Simmons MP, Simpson DA, Stevens P, Sundue M, Testo W, Trias-Blasi A, Turner I, Utteridge T, Walsingham L, Webber BL, Wei R, Weiblen GD, Weigend M, Weston P, de Wilde W, Wilkie P, Wilmot-Dear CM, Wilson HP, Wood JRI, Zhang LB, van Welzen PC (2020). New Guinea has the world's richest island flora. *Nature* **584**, 579–583.

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu YL, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim KJ, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang QY, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA (1993). Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Ann Mo Bot Gard* **80**, 528–548, 550–580.

Chen QH, Chen L, Teixeira da Silva JA, Yu XN (2023). The plastome reveals new insights into the evolutionary and domestication history of peonies in East Asia. *BMC Plant Biol* **23**, 243.

Chen ZD, Yang T, Lin L, Lu LM, Li HL, Sun M, Liu B, Chen

M, Niu YT, Ye JF, Cao ZY, Liu HM, Wang XM, Wang W, Zhang JB, Meng Z, Cao W, Li JH, Wu SD, Zhao HL, Liu ZJ, Du ZY, Wang QF, Guo J, Tan XX, Su JX, Zhang LJ, Yang LL, Liao YY, Li MH, Zhang GQ, Chung SW, Zhang J, Xiang KL, Li RQ, Soltis DE, Soltis PS, Zhou SL, Ran JH, Wang XQ, Jin XH, Chen YS, Gao TG, Li JH, Zhang SZ, Lu AM, China Phylogeny Consortium (2016). Tree of life for the genera of Chinese vascular plants. *J Syst Evol* **54**, 277–306.

Christenhusz MJM, Reveal JL, Farjon A, Gardner MF, Mill RR, Chase MW (2011). A new classification and linear sequence of extant gymnosperms. *Phytotaxa* **19**, 55–70.

Christin PA, Spriggs E, Osborne CP, Strömberg CAE, Salamin N, Edwards EJ (2014). Molecular dating, evolutionary rates, and the age of the grasses. *Syst Biol* **63**, 153–165.

Coe K, Bostan H, Rolling W, Turner-Hissong S, Macko-Podgórni A, Senalik D, Liu S, Seth R, Curaba J, Mengist MF, Grzebelus D, Van Deynze A, Dawson J, Ellison S, Simon P, Iorizzo M (2023). Population genomics identifies genetic signatures of carrot domestication and improvement and uncovers the origin of high-carotenoid orange carrots. *Nat Plants* **9**, 1643–1658.

Coelho N, Gonçalves S, Romano A (2020). Endemic plant species conservation: biotechnological approaches. *Plants* (*Basel*) **9**, 345.

Cornwell WK, Pearse WD, Dalrymple RL, Zanne AE (2019). What we (don't) know about global plant diversity. *Ecography* **42**, 1819–1831.

Folk RA, Sun M, Soltis PS, Smith SA, Soltis DE, Guralnick RP (2018). Challenges of comprehensive taxon sampling in comparative biology: wrestling with rosids. *Am J Bot* **105**, 433–445.

Forest F, Grenyer R, Rouget M, Davies TJ, Cowling RM, Faith DP, Balmford A, Manning JC, Procheş Ş, van der Bank M, Reeves G, Hedderson TAJ, Savolainen V (2007). Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* **445**, 757–760.

Goremykin VV, Hirsch-Ernst KI, Wölfl S, Hellwig FH (2003). Analysis of the *Amborella* trichopoda chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Mol Biol Evol* **20**, 1499–1505.

Govaerts R, Nic Lughadha E, Black N, Turner R, Paton A (2021). The world checklist of vascular plants, a continuously updated resource for exploring global plant diversity. *Sci Data* **8**, 215.

Gutaker RM, Burbano HA (2017). Reinforcing plant evolutionary genomics using ancient DNA. *Curr Opin Plant Biol* **36**, 38–45.

He JK, Gao ZF, Su YY, Lin SL, Jiang HS (2018). Geographical and temporal origins of terrestrial vertebrates endemic to Taiwan. *J Biogeogr* **45**, 2458–2470.

Hinchliff CE, Smith SA (2014). Some limitations of public sequence data for phylogenetic inference (in plants). *PLoS One* **9**, e98986.

Hu HH, Liu B, Liang YS, Ye JF, Saqib S, Meng Z, Lu LM, Chen ZD (2020). An updated Chinese vascular plant tree of life: phylogenetic diversity hotspots revisited. *J Syst Evol* **58**, 663–672.

Hu HH, Ye JF, Liu B, Mao LF, Smith SA, Barrett RL, Soltis PS, Soltis DE, Chen ZD, Lu LM (2022). Temporal and spatial comparisons of angiosperm diversity between eastern Asia and North America. *Natl Sci Rev* **9**, nwab 199.

Hu Y, Wang X, Zhang XX, Zhou W, Chen XY, Hu XS (2019). Advancing phylogeography with chloroplast DNA markers. *Biodiv Sci* **27**, 219–234. (in Chinese)
胡颖，王茜，张新新，周玮，陈晓阳，胡新生 (2019). 叶绿体DNA标记在谱系地理学中的应用研究进展. 生物多样性 **27**, 219–234.

Huang WD, Zhao XY, Zhao X, Li YQ, Zuo XA, Lian J, Luo YQ (2013). Genetic diversity in *Artemisia halodendron* (Asteraceae) based on chloroplast DNA *psbA-trnH* region from different hydrothermal conditions in Horqin sandy land, northern China. *Plant Syst Evol* **299**, 107–113.

Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang JY, Bowman M, Iovene M, Sanseverino W, Cavagnaro P, Yildiz M, Macko-Podgórni A, Moranska E, Grzebelus E, Grzebelus D, Ashrafi H, Zheng ZJ, Cheng SF, Spooner D, Van Deynze A, Simon P (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat Genet* **48**, 657–666.

Jafari F, Osaloo SK, Mozffarian V (2015). Molecular phylogeny of the tribe Astereae (Asteraceae) in SW Asia based on nrDNA ITS and cpDNA *psbA-trnH* sequences. *Willdenowia* **45**, 77–92.

Janssens SB, Couvreur TLP, Mertens A, Dauby G, Dagallier LPMJ, Vanden Abeele S, Vandelook F, Mascarello M, Beeckman H, Sosef M, Droissart V, van der Bank M, Maurin O, Hawthorne W, Marshall C, Réjou-Méchain M, Beina D, Baya F, Merckx V, Verstraete B, Hardy O (2020). A large-scale species level dated an-

giosperm phylogeny for evolutionary and ecological analyses. *Biodivers Data J* **8**, e39677.

Jiang ZH, Zhang MQ, Kong LY, Bao YH, Ren WC, Li HY, Liu XB, Wang Z, Ma W (2023). Identification of Apiaceae using ITS, ITS2 and *psbA-trnH* barcodes. *Mol Biol Rep* **50**, 245–253.

Kistler L, Bieker VC, Martin MD, Pedersen MW, Ramos Madrigal J, Wales N (2020). Ancient plant genomics in archaeology, herbaria, and the environment. *Annu Rev Plant Biol* **71**, 605–629.

Krause CM, Cobb NS, Pennington DD (2015). Range shifts under future scenarios of climate change: dispersal ability matters for Colorado plateau endemic plants. *Nat Areas J* **35**, 428–438.

Kuang DY, Wu H, Wang YL, Gao LM, Zhang SZ, Lu L (2011). Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome* **54**, 663–673.

Lang PLM, Weiß CL, Kersten S, Latorre SM, Nagel S, Nickel B, Meyer M, Burbano HA (2020). Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. *Mol Ecol Resour* **20**, 1228–1247.

Li FD, Tong W, Xia EH, Wei CL (2019). Optimized sequencing depth and *de novo* assembler for deeply reconstructing the transcriptome of the tea plant, an economically important plant species. *BMC Bioinformatics* **20**, 553.

Li HT, Luo Y, Gan L, Ma PF, Gao LM, Yang JB, Cai J, Gitzendanner MA, Fritsch PW, Zhang T, Jin JJ, Zeng CX, Wang H, Yu WB, Zhang R, van der Bank M, Olmstead RG, Hollingsworth PM, Chase MW, Soltis DE, Soltis PS, Yi TS, Li DZ (2021). Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biol* **19**, 232.

Li L, Wang WY, Zhang GQ, Wu KL, Fang L, Li MZ, Liu ZJ, Zeng SJ (2023). Comparative analyses and phylogenetic relationships of thirteen *Pholidota* species (Orchidaceae) inferred from complete chloroplast genomes. *BMC Plant Biol* **23**, 269.

Lian L, Peng HW, Ortiz RDC, Jabbour F, Gao TG, Erst AS, Chen ZD, Wang W (2023). Phylogeny and biogeography of Tiliacoreae (Menispermaceae), a tribe restricted to tropical rainforests. *Ann Bot* **131**, 685–695.

Linder HP, Hardy CR, Rutschmann F (2005). Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. *Mol Phylogenet Evol* **35**, 569–582.

Liu YJ, Liu Y, Huang YJ, Long CL (2011). Progress and application of DNA barcoding technique in plants. *J Plant Resour Environ* **20**, 74–82, 93. (in Chinese)
刘宇婧，刘越，黄耀江，龙春林 (2011). 植物DNA条形码技术的发展及应用. 植物资源与环境学报 **20**, 74–82, 93.

Liu YM, Zhang LH, Liu Z, Luo K, Chen SL, Chen KL (2012). Species identification of *Rhododendron* (Ericaceae) using the chloroplast deoxyribonucleic acid *psbA-trnH* genetic marker. *Pharmacogn Mag* **8**, 29–36.

Liu YP, Xu XT, Dimitrov D, Pellissier L, Borregaard MK, Shrestha N, Su XY, Luo A, Zimmermann NE, Rahbek C, Wang ZH (2023). An updated floristic map of the world. *Nat Commun* **14**, 2990.

Lu LM, Mao LF, Yang T, Ye JF, Liu B, Li HL, Sun M, Miller JT, Mathews S, Hu HH, Niu YT, Peng DX, Chen YH, Smith SA, Chen M, Xiang KL, Le CT, Dang VC, Lu AM, Soltis PS, Soltis DE, Li JH, Chen ZD (2018). Evolutionary history of the angiosperm flora of China. *Nature* **554**, 234–238.

Lv YN, Yang CY, Shi LC, Zhang ZL, Xu AS, Zhang LX, Li XL, Li HT (2020). Identification of medicinal plants within the Apocynaceae family using ITS2 and *psbA-trnH* barcodes. *Chin J Nat Med* **18**, 594–605.

Mo ZQ, Fu CN, Zhu MS, Milne RI, Yang JB, Cai J, Qin HT, Zheng W, Hollingsworth PM, Li DZ, Gao LM (2022). Resolution, conflict and rate shifts: insights from a densely sampled plastome phylogeny for *Rhododendron* (Ericaceae). *Ann Bot* **130**, 687–701.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000). Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858.

Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ (2011). Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol J* **9**, 328–333.

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM, Holm S, Säll T, Schlötterer C, Marhold K, Widmer A, Sese J, Shimizu KK, Weigel D, Krämer U, Koch MA, Nordborg M (2016). Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat Genet* **48**, 1077–1082.

Ohi T, Kajita T, Murata J (2003). Distinct geographic structure as evidenced by chloroplast DNA haplotypes and ploidy level in Japanese *Aucuba* (Aucubaceae). *Am J*

Bot **90**, 1645–1652.

Ortelt J, Link G (2014). Plastid gene transcription: promoters and RNA polymerases. *Methods Mol Biol* **1132**, 47–72.

Park DS, Worthington S, Xi ZX (2018). Taxon sampling effects on the quantification and comparison of community phylogenetic diversity. *Mol Ecol* **27**, 1296–1308.

Peng HW, Wang W (2023). Phylogenetic tree reconstruction based on molecular data. *Chin Bull Bot* **58**, 261–273. (in Chinese)

彭焕文，王伟 (2023). 基于分子数据的系统发生树构建. 植物学报 **58**, 261–273.

PPG I (2016). A community-derived classification for extant lycophytes and ferns. *J Syst Evol* **54**, 563–603.

Raman G, Park S (2015). Analysis of the complete chloroplast genome of a medicinal plant, *Dianthus superbus* var. *longicalyncinus*, from a comparative genomics perspective. *PLoS One* **10**, e0141329.

Ran JH, Shen TT, Wang MM, Wang XQ (2018). Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proc Biol Sci* **285**, 20181012.

Rudbeck AV, Sun M, Tietje M, Gallagher RV, Govaerts R, Smith SA, Svenning JC, Eiserhardt WL (2022). The Darwinian shortfall in plants: phylogenetic knowledge is driven by range size. *Ecography* **2022**, e06142.

Schulte JA (2013). Undersampling taxa will underestimate molecular divergence dates: an example from the South American Lizard clade Liolaemini. *Int J Evol Biol* **2013**, 628467.

Smith SA, Brown JW (2018). Constructing a broadly inclusive seed plant phylogeny. *Am J Bot* **105**, 302–314.

Soltis DE, Soltis PS (2004). *Amborella* not a “basal angiosperm”? Not so fast. *Am J Bot* **91**, 997–1001.

Staats M, Cuenca A, Richardson JE, Ginkel RVV, Petersen G, Seberg O, Bakker FT (2011). DNA damage in plant herbarium tissue. *PLoS One* **6**, e28448.

The Angiosperm Phylogeny Group, Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J Linn Soc* **181**, 1–20.

Tyler T (2017). The last step towards a full revision of *Hieracium* sect. *Vulgata* in Sweden. *Nord J Bot* **35**, 305– 321.

Vischi N, Natale E, Villamil C (2004). Six endemic plant species from central Argentina: an evaluation of their conservation status. *Biodivers Conserv* **13**, 997–1008.

Wang J, Kan SL, Liao XZ, Zhou JW, Tembrock LR, Daniell H, Jin SX, Wu ZQ (2024). Plant organellar genomes: much done, much more to do. *Trends Plant Sci* **29**, 754–769.

Wang LL, Yang Z, Yang Y (2023). Plant ultra-barcoding using herbariomics. *Chin Bull Bot* **58**, 831–842. (in Chinese)

王露露，杨智，杨永 (2023). 利用标本组学推进植物超级DNA条形码研究. 植物学报 **58**, 831–842.

Wicke S, Schneeweiss GM, de Pamphilis CW, Müller KF, Quandt D (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* **76**, 273–297.

Yao X, Song Y, Yang JB, Tan YH, Corlett RT (2021). Phylogeny and biogeography of the hollies (*Ilex* L., Aquifoliaceae). *J Syst Evol* **59**, 73–82.

Ye JW, Yang ZZ, Tian B (2023). Tempo-spatial evolution of seed plant endemism in Taiwan island. *J Biogeogr* **50**, 1981–1991.

Yu QB, Huang C, Yang ZN (2014). Nuclear-encoded factors associated with the chloroplast transcription machinery of higher plants. *Front Plant Sci* **5**, 316.

Yurina NP, Sharapova LS, Odintsova MS (2017). Structure of plastid genomes of photosynthetic eukaryotes. *Biochemistry* (*Mosc*) **82**, 678–691.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O’Meara BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ, Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J, Soltis PS, Swenson NG, Warman L, Beaulieu JM (2014). Three keys to the radiation of angiosperms into freezing environments. *Nature* **506**, 89–92.

Zhang X, Sun YX, Landis JB, Lv ZY, Shen J, Zhang HJ, Lin N, Li LJ, Sun J, Deng T, Sun H, Wang HC (2020). Plastome phylogenomic study of Gentianeae (Gentianaceae): widespread gene tree discordance and its association with evolutionary rate heterogeneity of plastid genes. *BMC Plant Biol* **20**, 340.

Zhang YJ, Li DZ (2011). Advances in phylogenomics based on complete chloroplast genomes. *Plant Diversity Resour* **33**, 365–375. (in Chinese)

张韵洁，李德铢 (2011). 叶绿体系统发育基因组学的研究进展. 植物分类与资源学报 **33**, 365–375.

A Comprehensive Evaluation of the Plastid DNA Data Gaps of Vascular Plants in Species and Geographic Area

Yan Deng[1, 2], Limin Lu[2, 3], Qiang Zhang[4*], Zhiduan Chen[2, 3], Haihua Hu[2, 3*]

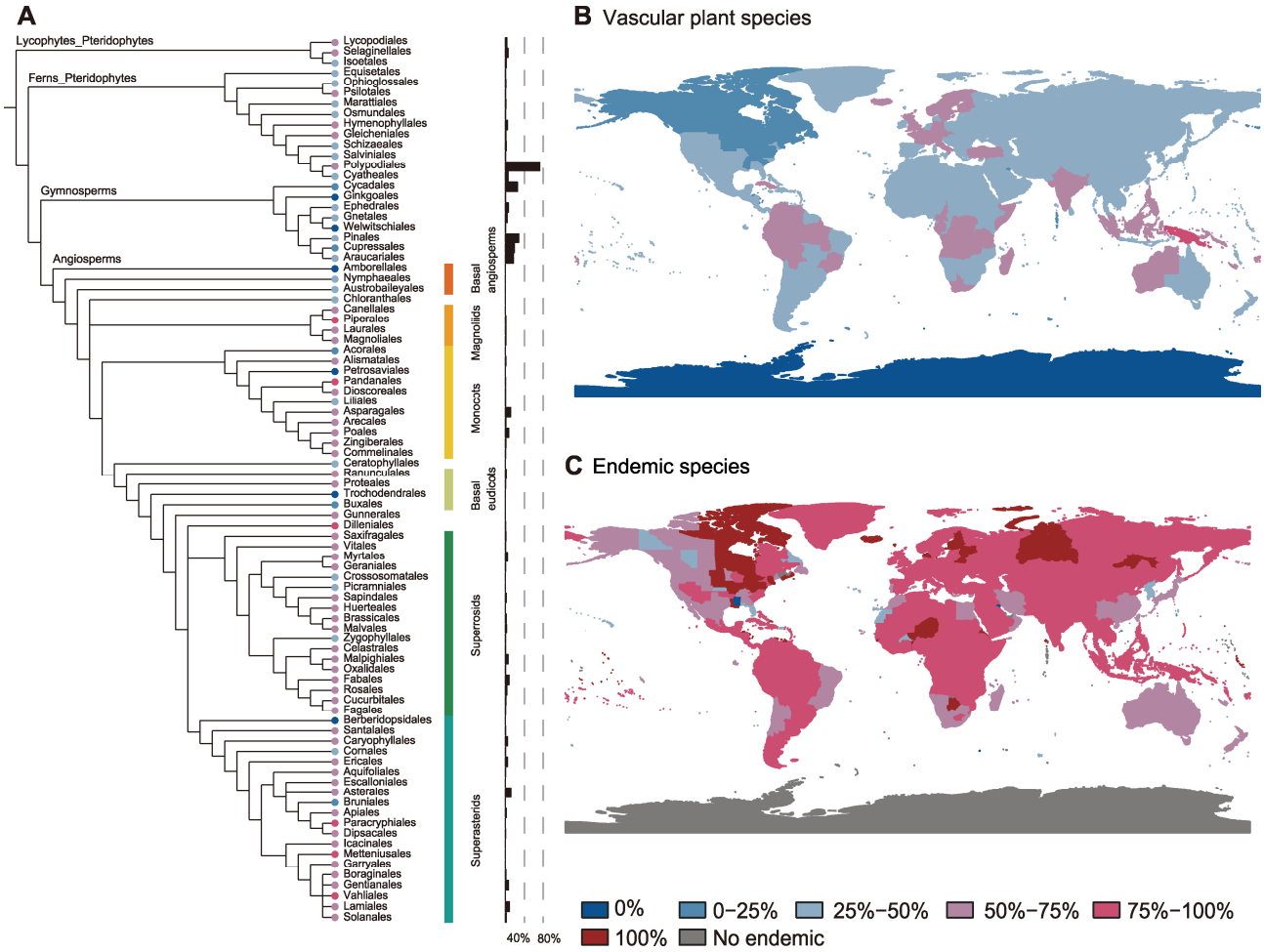
[1]College of Life Sciences, Guangxi Normal University, Guilin 541006, China; [2]Key Laboratory of Systematic and Evolutionary Botany, State Key Laboratory of Plant Diversity and Specialty Crops, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China; [3]China National Botanical Garden, Beijing 100093, China; [4]Guangxi Key Laboratory of Plant Conservation and Restoration Ecology in Karst Terrain, Guangxi Institute of Botany, Chinese Academy of Sciences, Guilin 541006, China

INTRODUCTION: Molecular data is one of the most important bases for many biological studies, including phylogeny, ecology, and biogeography etc. Incomplete sampling may lead to biased results and inadequate conclusions. However, few studies have evaluated current state of sampling density for sequencing DNA data comprehensively. Plastid DNA sequences have been applied in scientific studies of plants extensively due to their easy accessibility, uniparental inheritance, and moderate rate of mutation. Therefore, it is essential to investigate the current state of sampling density for sequencing plastid DNA data in species and geographic area for researchers to better utilize it.

RATIONALE: The GenBank is the biggest and most commonly used database of sequencing DNA data. The data gap of plastid DNA in species and geographic area for vascular plants was investigated based on the GenBank database in this study. Firstly, the plastid DNA data of vascular plant species were downloaded from the GenBank database and cleaned. Secondly, species names were standardized according to the World Checklist of Vascular Plants (WCVP) database. Thirdly, to evaluate the current state of sampling density for plastid DNA data of vascular plants, we counted the number of species with plastid DNA sequenced and the proportion of missing data of lineages representing orders and families. We also mapped the proportion of missing data in each region to evaluate the current state of sampling density of plastid DNA data geographically. To further investigate the potential influencing factors of the plastid DNA data gap, Spearman's correlations between the proportion of missing data and species diversity among major groups of vascular plants or regions were calculated.

RESULTS: Only 33.75% vascular plant species have at least one record of DNA in GenBank, covering 139 005 vascular plant species (angiosperms: 131 220 species, gymnosperms: 1 154 species, and pteridophytes: 6 631 species). For data gap in species, sequenced species were unevenly sampled among lineages, with the proportion of missing data generally correlated with species richness within the lineages. The top three orders of the highest proportion of missing data were Paracryphiales, Piperales, and Dilleniales, and the top three families were Triuridaceae, Pentaphragmataceae, and Xyridaceae. For data gap in geographic area, the proportion of missing data of plastid DNA of vascular plant species showed a trend of latitudinal gradient, with the degree of missing data decreasing from the equator to the poles. Regions with high proportion of missing data usually possess high biodiversity, including many biodiversity hotspots. In addition, endemic species were generally with the high proportion of missing data in the majority of regions.

CONCLUSION: Our research evaluated the current state of sampling density for plastid DNA data in species and geographic area comprehensively. Our results suggested that about 140 000 vascular plant species have been sequenced for the plastid DNAs. However, there are still large data gaps for the plastid DNA of vascular plants in the following three aspects: (1) Only 1/3 vascular plant species have been sequenced; (2) Ratios of species with plastid DNA sequenced are uneven among lineages; (3) The proportion of missing data decreases from the equator to the poles, with more deficiencies in biodiversity hotspots and endemic species. Based on the results of this study, we propose to give priority to collection and sequencing of vascular plants for groups with high proportion of missing data and regions with high biodiversity, particularly for the endemic species. Our research points out the direction of filling plastid DNA data gap and will be beneficial to biodiversity protection.



The plastid DNA data gaps at the ordinal level of vascular plants and geographic area

(A) There is uneven distribution of sampled species among lineages representing orders of vascular plants; **(B)** The degree of missing data decreasing from the equator to the poles for vascular plant species; **(C)** The proportion of missing data of endemic species

Key words plastid DNA, vascular plants, missing data, big data of plant, GenBank

Deng Y, Lu LM, Zhang Q, Chen ZD, Hu HH (2025). A comprehensive evaluation of the plastid DNA data gaps of vascular plants in species and geographic area. *Chin Bull Bot* **60**, 1–16.

---

* Authors for correspondence. E-mail: qiangzhang04@126.com; huhh@ibcas.ac.cn

(责任编辑: 白羽红)

附录 1 GenBank 中质体 DNA 分子标记在维管植物、被子植物、裸子植物和广义蕨类植物中的取样率(doi:10.57760/sciencedb.j00154.00007)
Appendix 1 Sample proportion of molecular markers of plastid DNA in tracheophytes, angiosperms, gymnosperms, and pteridophytes (doi:10.57760/sciencedb.j00154.00007)

附表 1 维管植物在目水平的质体 DNA 数据缺失率
Appendix table 1 The proportion of missing data of plastid DNA at the ordinal level of tracheophytes

附表 2 维管植物在科水平的质体 DNA 数据缺失率
Appendix table 2 The proportion of missing data of plastid DNA at the family level of tracheophytes

附图 1 被子植物、裸子植物和广义蕨类植物在科水平的质体 DNA 数据缺失率
Appendix figure 1 The proportion of missing data of plastid DNA at the family level of angiosperms, gymnosperms, and pteridophytes

附图 2 所有维管植物、被子植物和广义蕨类植物目和科水平的物种数与质体 DNA 数据缺失率的 Spearman 相关性
Appendix figure 2 Spearman correlations between the proportion of missing data of plastid DNA and species number at the ordinal and family levels of tracheophytes, angiosperms, and pteridophytes

https://www.chinbullbotany.com/fileup/1674-3466/PDF/24-034-1.pdf

通讯作者/团队简介

胡海花, 中国科学院植物研究所副研究员。长期从事维管植物生命之树重建及生物多样性格局与保护研究。已发表研究论文 20 余篇, 其中以第一作者(含共同第一作者)身份在 *National Science Review*、*Fundamental Research*、*Journal of Systematics and Evolution* 等期刊发表研究论文 4 篇。所在的植物大数据与生物多样性保护研究团队利用多学科手段, 在较高分类阶元上探讨植物的系统发育关系和演化, 并将形态学、古植物学和分子系统学的研究结果相结合, 研究植物类群的起源、分化和现代地理分布格局及其成因。近年来, 以生命之树为依托, 结合海量物种分布数据, 从时间和空间维度探究植物区系的演化历史、多样性格局及其成因及生物多样性保护策略。

张强, 广西壮族自治区中国科学院广西植物研究所研究员。主要从事被子植物(金粟兰科等类群)系统发育与生物地理、喀斯特植物(毛茛科天葵属和苦苣苔科等)适应性演化以及分子进化和生物信息学方法研究。已发表研究论文30余篇。开发出同源序列比对矩阵质量过滤软件alignmentFilter。