大连理工大学本科毕业设计（论文）

真菌界生物合成基因簇的预测、聚类和多样性分析

Prediction, clustering and diversity analysis of biosynthetic gene clusters in Fungi

学 院（系）：生命科学与药学学院

专　　业：生物信息学

学 生 姓 名：郭旭

学　　号：201927176

指 导 教 师：吴琦、韩彦粲

评 阅 教 师：

完 成 日 期：

大连理工大学

Dalian University of Technology

原创性声明

本人郑重声明：本人所呈交的毕业设计（论文），是在指导老师的指导下独立进行研究所取得的成果。毕业设计（论文）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究成果做出重要贡献的个人和集体，均已在文中以明确方式标明。

本声明的法律责任由本人承担。

作者签名：郭旭　　日　期：6月7日

关于使用授权的声明

本人在指导老师指导下所完成的毕业设计（论文）及相关的资料（包括图纸、试验记录、原始数据、实物照片、图片、录音带、设计手稿等），知识产权归属大连理工大学。本人完全了解大连理工大学有关保存、使用毕业设计（论文）的规定，本人授权大连理工大学可以将本毕业设计（论文）的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本毕业设计（论文）。如果发表相关成果，一定征得指导教师同意，且第一署名单位为大连理工大学。本人离校后使用毕业毕业设计（论文）或与该论文直接相关的学术论文或成果时，第一署名单位仍然为大连理工大学。

论文作者签名：郭旭　　日　期：6月7日

指导老师签名：韩秀梅　　日　期：6月7日

摘　要

本设计旨在使用生物信息学手段预测、聚类真菌的生物合成基因簇（biosynthetic gene clusters, BGC），借助基因簇家族（gene cluster family, GCF）揭示整个真菌界的生物合成多样性及其潜力并且探寻一种有效的 BGC 挖掘流程。为此我们收集了目前真菌界全部的基因组 gbk 文件（11，609 个），输入到软件 antiSMASH 预测得到了 293，926 个真菌的 BGC。接着，超过 29 万个 BGC 被大规模 BGC 聚类工具 BiG-SLiCE 聚类为 26，825 个 GCF。根据 GCF 的分布情况我们从门和纲水平出发探索了真菌界的生物合成多样性。而后借助稀疏曲线，我们预测了真菌全部属的生物合成多样性潜力。除具有大量已知天然产物的曲霉属和青霉属外，我们的研究发现镰刀菌属、炭角菌属和炭团菌属虽然已知天然产物数量很少，但却拥有高水平的生物合成多样性及其潜力。仿照单细胞测序数据的处理模式，我们超聚类了真菌界的全部 GCF，并根据已知的 BGC 注释了 GCF 超聚类，仅有不足 1/5 的 GCF 属于未被注释的超聚类。接下来，以曲霉属为例，我们将余下的 GCF 同曲霉属独有的 GCF 取交集，并比对了筛选出的若干 GCF 含有的 BGC 序列。属于同一 GCF 的曲霉属 BGC 拥有很高相似性，而属于不同 GCF 和相同 GCF 超聚类的 BGC 具有相近的核心基因。最终，筛选得到的数十个曲霉属 BGC 将被纳入到后续的外源表达验证实验。

关键词：生物合成基因簇；生物合成多样性；真菌

Prediction, clustering and diversity analysis of biosynthetic gene clusters in Fungi

Abstract

This design aims to use bioinformatics to predict and cluster biosynthetic gene clusters (BGCs) of fungi, reveal the biosynthetic diversity and potential of the whole fungal kingdom by means of gene cluster families (GCFs), and explore an effective BGCs mining process. For this purpose, we collected all the gbk files (11,609) of genomes of fungi at present, and entered the software antiSMASH to predict the BGC of 293,926 fungi. More than 290,000 BGCs were then clustered into 26,825 GCFs by the large-scale BGCs clustering tool BiG-SLiCE. Based on the distribution of GCFs, we explored the biosynthetic diversity of fungi at phylum and class level. Then, using rarefaction curves, we predicted the biosynthetic diversity potential of all fungal genera. In addition to Aspergillus and Penicillium, which have a large number of known natural products, our study found that Fusarium, Carbonocellaria and Xylaria possess a high level of biosynthetic diversity and potential biosynthetic diversity containing despite a small number of known natural products. Following the single-cell sequencing data processing pattern, we super-clustered all the GCFs in the fungal kingdom and annotated the GCFs super-clusters according to the known BGCs, with less than 1/5 of the GCFs belonging to the unannotated super-clusters. Next, taking Aspergillus as an example, we intersected the remaining GCFs with the GCFs unique to Aspergillus, and compared the BGC sequences which contained in several selected GCFS. Aspergillus BGCs belonging to the same GCF have high similarity, while BGCs belonging to different GCFs but the same GCF super-cluster have similar core genes. Finally, dozens of screened Aspergillus BGCs will be included in subsequent exogenous expression verification experiments.

Key Words：Biosynthetic gene cluster; Biosynthetic diversity; Fungi

目　　录

文献综述

活性天然产物是现代新药创制和发现的重要基础，真菌作为活性天然产物的重要来源之一引起了人们极大的关注，主要是因为其丰富的资源、代谢产物结构的多样化及潜在的成药能力，这些代谢产物常常称之为次级代谢产物。对真菌自身而言，这些次级代谢产物有利于真菌在自然环境中竞争营养物质，抵御侵害，实现自我防御，并可作为与环境中生物交流的信号分子[1]。生物合成基因簇（biosynthetic gene clusters，BGC）是指一组在细菌、真菌和植物中发现的基因，这些基因编码了一系列酶和蛋白质，可以合成一种或多种生物活性分子。 这些分子包括抗生素、毒素、色素、激素和其他生物活性化合物。 BGC 的发现和研究对于开发新的药物和化学品具有重要意义。

真菌的生物合成，尤其是次级代谢与人类有着长期而密切的联系，特别是在化学层面上。20 世纪 60 年代发生的土耳其 X 病黄曲霉毒素中毒事件[2]和第一个广谱抗生素——青霉素的发现，使人们认识到真菌既是有害化合物的来源，也是有益化合物的来源。这些生物活性分子被称为次生代谢物(也称为天然产物)，是由特定的真菌类群产生的，主要是归属于子囊菌纲的丝状真菌，以及几个担子菌纲。次级代谢产物来源于中心代谢途径和初级代谢产物库，酰基辅酶 A 是合成聚酮(如黄曲霉毒素)和萜(如胡萝卜素)次级代谢产物的关键初始构件，氨基酸用于合成非核糖体肽次级代谢产物(如青霉素)。与分散在真菌基因组中合成初级代谢物所需的基因相反，编码产生任何次级代谢物的酶活性的基因以连续的方式排列为生物合成基因簇，如黄曲霉毒素的 BGC[3]。次生代谢产物在真菌发育中起着至关重要的作用，并积极地形成与其他生物的相互作用。事实上，BGC 内的基因通常根据其编码的次级代谢物的生态功能共同调节。例如，烟曲霉中编码色素的 BGC 在该真菌的孢子合成过程中被激活[4]，编码小麦赤霉病菌毒力因子毛霉烯的 BGC 在植物定植过程中上调[5]。

天然产物研究通过采用基因组学指导的策略来绘制未开发的生物合成空间而发展[6]。科学家最初专注于挖掘单个基因组，现在同时分析数十、数百甚至数千个基因组。使用所选的 BGC 检测算法识别 BGC 后[7]，来自不同基因组的 BGC 可以根据其整体基因内容和序列同一性的相似性分组到基因簇家族（gene cluster families，GCF）中[8]。BGC 分组的相似性阈值的选择是影响推断 GCF 生物合成产物解释的重要因素。使用严格的相似性阈值形成的 GCF 将是较小的家族，由产生相同代谢物的 BGC 组成，而具有允许阈值产生的 GCF 将更大，并且包含编码结构相关的天然产物家族的 BGC。这些分

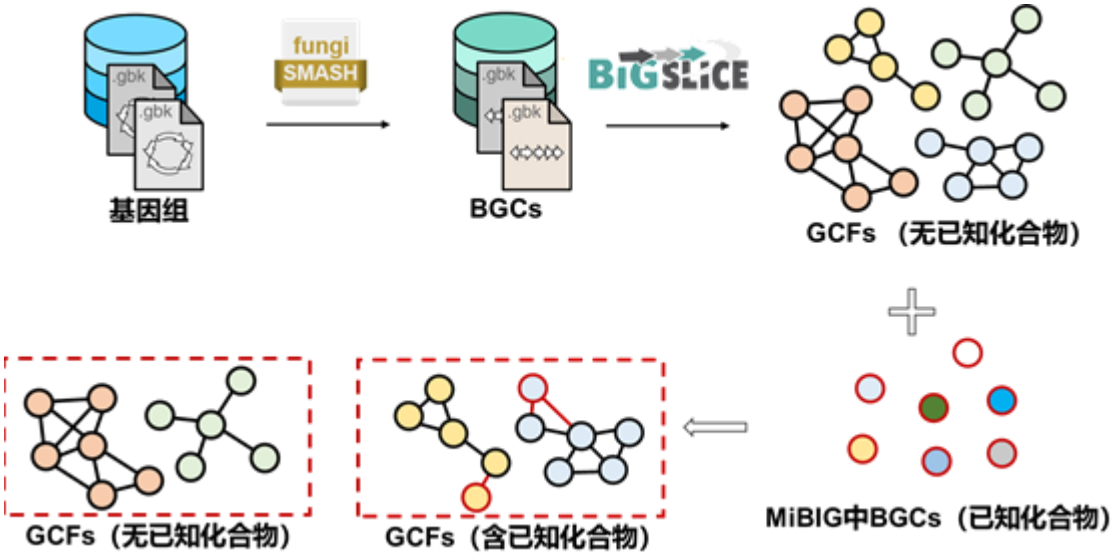
析使研究人员能够发现 BGC 流行和系统发育分布的模式，并将 GCF 网络锚定为实验表征的 BGC。

选择正确的 BGC 的一个挑战是它可能已经被表征了。丰富的测序真菌基因组使先进的生物信息学研究成为可能，这些研究已经确定了 BGC 维持、形成和衰变的进化趋势[9]. 许多 BGC 及其变体在不同的真菌中是保守的，通常是水平转移事件的结果[10]，因此对簇同系物的生物信息学搜索是 BGC 选择的关键因素。鉴于目前所有技术的可用性，我们可以知道一个神秘的簇是否真的会产生代谢物，但仍然有许多失败的案例。在某些情况下，一个 BGC 中的所有基因都是异源表达的，但没有形成产物。当然，转录后机制，包括内含子剪接，肽的错误折叠，缺乏前体和细胞运输，可能会导致执行困难，也许在酿酒酵母中更是如此，它还没有进化到合成复杂的天然产物。但能有其他原因吗？一些 BGC 是否处于进化的死胡同，或者走上了合成的道路？

烟曲霉的 BGCs 为对这一问题的推测提供了一个框架。这种真菌是一种重要的人类病原体，已经完成了 100 多种分离株的基因组序列，并做出了大量努力来鉴定其（潜在有毒的）BGC 产物[11]。目前，在已知 BGC 中编码的次级代谢物中，超过 50%已经被准确分配[12]。对 66 株烟曲霉分离株的 BGC 序列的详细检查揭示了多种类型的 BGC 变异，包括 SNP，基因丢失，基因增益，簇丢失，簇迁移（转移到另一条染色体），簇融合和特异形（多等位）簇[13]. 在 36 个 BGC 中，只有 20 个没有变化或变化很小，主要由特征簇组成，包括两个铁载体簇（铁克罗素和梭沙宁 C）、烟曲霉素、六脱氢阿斯特铬、内巴罗辛、烟粉、DHN 黑色素、胶质毒素、烟喹唑啉、胰氨酸、吡嗪啶 A、新沙托霉素和交织在一起的烟马西林-补骨脂超簇。对于其余 16 个簇，66 株烟曲霉菌株的变异范围从丢失到多重多态性不等，其中只有 2 个 BGC 被表征（富米特雷莫林和海尔沃酸）。剩下的 14 个 BGC 是具有启发性的，例如，一种未表征的 BGC，称为 Fusarielin 样 BGC，具有五个基因，与禾谷镰刀菌中的六基因 Fusarielin 簇同源[14]。该簇在 4 个菌株中的 66 个中不存在，并且在其他 43 个菌株中具有 SNP，其中大多数经检查会导致一个或多个基因产物的功能障碍。也许这个集群代表了进化时期捕获的死亡事件，并且没有通过基因操纵集群的努力会导致至少 43 个菌株的产物形成。Fusarielin 样 BGC 可能反映了从镰刀菌到 B. cinerea 的古老 BGC 水平转移，其中 bikaverin BGC 在 B. cinerea 的许多分离株中反复衰变[15]。然而，烟曲霉只是真菌界内很小的一列分支，要是给整个真菌界的次代产物绘制一张宏伟的蓝图，需要借助更为强大的生物信息学工具。

自 2011 年以来，antiSMASH [16] 已成为鉴定和分析细菌和真菌基因组序列中 BGC 的最广泛使用的工具。antiSMASH 使用基于规则的方法来识别二次代谢生产中涉及的许

多不同类型的生物合成途径。其对编码非核糖体肽合成酶 （NRPS）、I 型和 II 型聚酮合成酶（PKS）、镧肽、套索肽、镰肽和硫代肽的 BGC 进行了更深入的分析，并基于集群特异性分析提供有关所执行的生物合成步骤的更多信息，从而还可以对产生的化合物提供更详细的预测。BiG-SLiCE[17]，是一个旨在聚类大量 BGC 的工具。BiG-SLiCE 在它的前身 BiG-SPACE[18]的基础上重点地提升了计算的效率。BiG-SLiCE 基于 Britch 算法原理，将原来复杂度为 n^2 的聚类算法简化成了线性，这使得程序的运行速度有了本质上的飞跃，因此具备了 BiG-SPACE 所欠缺的聚类数以万计的 BGC 的能力。它能通过在欧几里得空间中表示出数以万计的 BGC，并可以以非成对、近线性的方式将 BGC 分组为 GCF。通过调整聚类参数可以使得 GCF 的聚类模式与二次代谢产物的基于化学指纹[19]的聚类模式相仿。因此，分析 GCF 的数量和分布就能模拟真实的二次代谢产物多样性。2022 年 5 月报道了一篇关于细菌基因组中编码的特化代谢物生物合成多样性的文章，Nadine Ziemert 等人使用 antiSMASH 预测了 NCBI RefSeq 数据库中全部基因组的 BGC[20]，并使用 BiG-SLiCE 聚类了得到的 BGC 产生了超过六万个 GCF。该研究又进一步分析了属水平的次代产物多样性和地理分布，为细菌次代产物的挖掘建立了有力的基础[20]。



设计流程图

生物信息学工具在细菌次级代谢多样性挖掘的过程中起到了决定性的作用，因此，我们的研究旨在通过相似的研究思路获取真菌次级代谢多样性的宏图，并以此为真菌的次级代谢研究铺设好崭新且坚固的道路。我们的研究主要分为以下四个阶段：第一阶段，收集目前 NCBI 数据库中全部的真菌基因组，并使用 antiSMASH 和 BiG-SLiCE 两种生

物信息学工具分别预测和聚类 BGC，得到真菌已知 GCF。在这个过程中需要确定真菌 BiG-SLiCE 的聚类参数，使聚类结果尽可能地逼近真实的化学指纹聚类结果。第二阶段，建立系统的真菌界次级代谢多样性宏图，并且通过稀疏曲线预测不同类别真菌的潜在 GCF（pGCF）数量，以从整体上揭示真菌的次代产物多样性和其潜力。第三阶段，选择特殊真菌类群，完成更为深入和细致的信息学分析，借助 GCF 的分布情况、GCF 和 pGCF 的数量、BGC 的化学分类和 MiBiG 数据库[21]中已知的一部分 BGC 的化学注释信息等生物学信息来进一步挖掘有价值、易验证的 BGC。最后，在第四阶段中，我们将根据前三个阶段中通过一系列生物信息学分析手段得到的若干 BGC 进行实验验证（BGC 的内源或外源表达实验）。由于 BGC 的人工表达难度极大，筛选出合适的 BGC 尤为重要，第四阶段验证过程的顺利与否很大程度上依赖第三阶段的挖掘过程。因此，我们的工作也将为其他有关真菌次代产物的研究提供新思路。

1　服务器软件安装与环境配置

1.1 生物信息学软件介绍

1.1.1 antiSMASH

antiSMASH 是目前寻找代谢基因簇最好的软件，一般情况下，参与代谢途径中生物合成酶的基因在染色体上成簇排列，基于指定类型的模型，可以准确鉴定所有已知的次级代谢基因簇。在 antiSMASH 中，将次级代谢基因簇分为了 24 类。antiSMASH 依赖的软件有 blastx、hmmer、glimmer3、GlimmerHMM 和 muscle。

使用 antiSMASH 寻找基因簇，这依赖基因组的结构信息，进而开始寻找次级代谢相关的基因在基因组上成簇排列的情况。输入文件仅有一个，就是 genebank 格式的文件（gbk 文件），如缺少 gbk 文件，也可以用 fasta 格式的文件代替，这时 antiSMASH 会自动调用 GlimmerHMM 对基因进行预测后再进行次级代谢基因簇的鉴定，然而这可能导致预测结果不准确，因此我们的输入文件全部为基因组的 gbk 文件。

本设计采用的是 antiSMASH 版本 6.0[16]，其相对于 5.0 版本提供了进一步的优化。对于镧肽、套索肽、镰肽和硫肽的 BGC，antiSMASH 5.0[22]已经通过检测簇的前肽和常用的定制酶提供了更详细的产物预测。由于一些定制酶可以匹配相对通用的功能特征，因此确定给定的酶是否确实作为剪切酶与核糖体合成和翻译后修饰肽（Ribosomally synthesized and post-translationally modified peptides，RiPP）的前体肽相互作用或者它是否只是碰巧在附近编码的无关酶是十分必要的。通常，RiPP 定制酶将包含 RiPP 识别元件（RRE）结构域，该结构域可以识别和结合 RiPP 前体肽。RRE-Finder[23]能够注释这些 RRE 结构域，现已纳入到 antiSMASH 6.0 中，从而有助于更可靠地识别 antiSMASH 检测到的 RiPP 簇中的剪切酶。

1.1.2 BiG-SLiCE

本工作采用的 BiG-SLiCE[17]软件是一个旨在聚类大量 BGC 的生物信息学工具。通过在欧几里得空间中表示 BGC，BiG-SLiCE 可以以非成对、近线性的方式将 BGC 分组为 GCF。相较于其前身 BiG-SPACE，BiG-SLiCE 将二阶非线性的算法时间复杂度降低为一阶线性，这赋予其聚类数以万计的 BGC 的强大能力。在软件开发者 Kautsar SA 的研究中，他们使用了具有 36 个运算核心的服务器仅在 10 天内就完成了对超过 122 万个细菌 BGC 的聚类工作。而事实上据我们估算，由于目前真菌的已测基因组数量远少于细菌，通过 antiSMASH 预测出的 BGC 数量也会少于细菌。因此，基于 Big-SLiCE 这一强大的

生物信息学软件我们完全可以在有限时间内完成真菌 BGC 的聚类。此外，Big-SLiCE 可以结合 MIBiG 数据库中的“已经通过实验手段表达出对应化合物的 BGC”（已注释 BGC）的序列信息，来判断预测得到的新 BGC 的对应化合物类型。

1.2 服务器软件下载、安装及环境配置

1.2.1 服务器环境配置

我们的计算设备包括一台具有 24 个 CPU 和 64G 内存的服务器以及一台具有 40 个 CPU 和 256G 内存的服务器。我们分别在两台服务器上下载了 Anaconda，该软件为服务器提供了 Python 环境 并简化了下载各种生物信息学软件的步骤。而后，我们使用 conda create -n 命令分别在两台服务器上建立了基于 Python3.6 版本的 bigslice 环境和 antismash 环境。

1.2.2 软件下载与安装

antiSMASH 的安装：依照官方网站（https://docs.antismash.secondarymetabolites.org/install/）的描述，使用 anaconda 的内部命令 conda create -n antismash antismash，在之前建立的 antismash 环境内下载并安装该软件。

BiG-SLiCE 的安装：安装过程参照软件开发者的 github 主页的安装指南（https://github.com/medema-group/bigslice）。开启 bigslice 环境后，使用命令 pip install bigslice 下载安装。此外需要通过命令 download_bigslice_hmmdb 下载所需的 hmm 数据库。使用 bigslice –version 命令检查安装情况。

2. 数据下载与来源

2.1 数据来源

2.1.1 NCBI 数据库

NCBI (National Center for Biotechnology Information)，美国国家生物技术信息中心。NCBI 开发有 Genbank 等公共数据库，提供 Pubmed、BLAST、Entres、OMIM、Taxonomy、Structure 等工具，可对国际分子数据库和生物医学文献进行检索和分析，并开发用于分析基因组数据和传播生物医学信息的软件工具。本研究访问了其基因组数据库（https://www.ncbi.nlm.nih.gov/genome/）和分类数据库（https://www.ncbi.nlm.nih.gov/taxonomy/）。

2.1.2 NP Atlas 数据库

Natural Products Atlas (NP Atlas)数据库旨在涵盖所有发表的主要科学文献中的微生物来源的天然产物。这包括细菌、真菌和蓝藻化合物，但不包括来自植物、无脊椎动物或其他高等生物的化合物，除非这些化合物也已从微生物来源明确识别。其中包括来自地衣、蘑菇和其他高等真菌的化合物，但海洋大型藻类和硅藻的化合物不包括在内。模板中已经自动设置为缺省值。

2.1.3 MIBiG 数据库

MIBiG（Minimum Information about a Biosynthetic Gene Cluster）数据标准和存储库，其中包含 1170 个 BGC 条目，它们是由整个领域的研究者们手动注释而来，其结果可以通过一个相当简单的 Web 应用程序访问[24]。现在，MIBiG 存储库已成为已知功能 BGC 的中央参考数据库，并通过 antiSMASH 为比较分析提供了基础。它使 BGC 功能和新颖性的许多计算分析成为可能，这对于微生物和微生物群落的小规模和大规模研究至关重要。例如，Crits-Cristoph 等人[25]于几年前使用 MIBiG 来评估和强调 BGCs 在 376 个未被研究的门的未栽培土壤细菌宏基因组组装基因组中的非凡新颖性，表明这些 BGC 中的大多数与 MIBiG 的基因簇没有任何同源性。

2.2 数据下载

2.2.1 基因组数据下载

于 2022 年 10 月 1 日从 NCBI 下载截止于当时的全部真菌界的基因组 gbk 文件并分批保存在两台服务器中，共 11609 个 gbk 文件。此外，从 NCBI 分类学数据库下载全部生物的分类信息（数据库格式），并使用 python 编程处理数据库信息，整理出 11609 个基因组对应的上级分类情况（包括所属的界、门、纲、目、科、属、种和菌株）数据库的全部分类信息，形成一个具有 11609 行的分类信息列表。

2.2.2 化合物信息下载

化合物信息同样于 2022 年 10 月 1 日于 NP Altas 官网下载得到。包括 20304 个已经发现的真菌天然产物，以及它们的来源种属信息、化合物名称和所属化合物聚类等信息。NP Altas 化合物的聚类信息通过基于分子指纹的大规模分子聚类获得，是从天然产物的化学构成出发的聚类过程。

2.2.3 参考 BGC 的下载

于 MIBiG 数据库官网下载截止于 2022 年 10 月 1 日的全部已知 BGC 的信息。包括 BGC 的 gbk 文件、对应天然产物的信息和天然产物的来源种属信息。而后我们剔除了部分来源种属信息不明的 BGC，最终筛选得到 143 个 BGC。与 NP Altas 的数据整合后，获取了 143 个 BGC 对应化合物的化学指纹聚类信息，用于后续分析。

3．BGC 的预测与聚类

3.1 antiSMASH 预测 BGC

3.1.1 基因组文件处理

并非所有基因组的 gbk 文件都能够成功运行 antiSMASH 软件。这些基因组往往具有一下几类问题：

（1） 基因组组装质量不佳

（2） 基因组过于庞大

（3） 蛋白编码序列（CDS）编号重名或重复出现

其中（3）问题较为常见，使用 Python 编写脚本批量处理，删除重复名称即可。（2）问题可以将基因组于 CDS 的间隙切割，再将切割后的片段输入到 antiSMASH 中预测 BGC。对于（1）问题，组装不佳的基因组的 gbk 文件往往数据量巨大，通过人工筛选剔除了 1 个基因组,其 gbk 文件达到 4G。

3.1.2 antiSMASH 的运行

全部 11608 个（1 个基因组被剔除）真菌基因组 gbk 文件被分为两组分别存入两台服务器中。两台服务器全部多线程运行，并使用全部运算核心。最终，24 核心的服务器于 13 天内预测了 3608 个基因组 gbk 文件（部分基因组 gbk 文件被切割为多个更小的 gbk 文件）的 BGC，40 核心的服务器于 10 天内预测了余下的 8000 个基因组 gbk 文件的 BGC。最终，我们整合了两个服务器得到的全部 BGC 的 gbk 文件，共预测得到 293，926 个 BGC。

3.2 最佳聚类参数 T 值搜索

BiG-SLiCE 在聚类时有一个超参数需要提前确定及“threshold”，简称“T 值”。该参数表示了聚类的搜索半径，形象地讲就是 T 值越大，一个聚类的搜索范围就越大，其内的个体就越多，反过来讲，T 值越小，搜索半径越小，则包含的个体只可能更少。我们使用 BiG-SLiCE 聚类的目的就是想通过 BGC 的聚类来模拟真菌天然产物的聚类，每个 BGC 对应一种真菌的天然产物，而每个 BGC 家族（GCF）对应一个天然产物簇及一类天然产物。因此，T 值的大小应当满足上述的要求，使得 BiG-SLiCE 的聚类结果能够尽可能地模拟真实的化合物聚类。那么确认合适的 T 值就是一项必须优先完成的工作。

前面的工作中，我们从 MIBiG 数据库下载了 143 个有明确种属来源信息的已知 BGC，

这些 BGC 都有对应的已鉴定的天然产物。接下来，我们将这 143 个 BGC 作为测试集输入到 BiG-SLiCE 中，分别以在 100~1500 区间，间隔为 50 的 29 种 T 值完成聚类。得到 29 个聚类结果后为判断 T 值的优劣，我们使用了两个指标一个是 Δ GCF，另一个是同质性度量和完整性度量的调和平均（V-Score）[26]。前者表示对于同一组 BGC，BiG-SLiCE 聚类数与这些 BGC 对应的真实化合物聚类数（基于化学指纹）的差值。而 V-Score 是同质性 homogeneity 和完整性 completeness 的调和平均数，其取值范围为[0,1]，越接近 1 表示 BiG-SLiCE 聚类与真实化合物聚类越接近。V-Score 和 Δ GCF 是两种不同的外部评估指标，两者被绘制到下图 3.1 中以找出最优的 T 值。

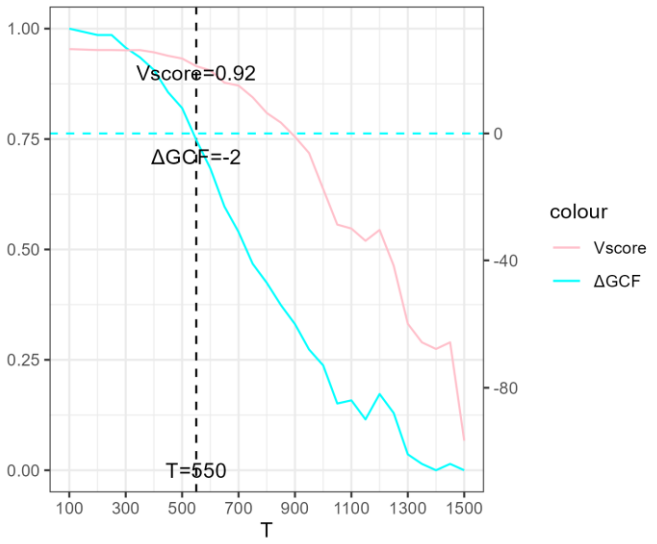


图 3.1 最优 T 值的搜索

从图 3.1 可知，当 T 值取 550 时，V-Score=0.92，Δ GCF=-2，此时 Δ GCF 接近于 0 且 V-Score 的值仍保持在较高水平，因此 T=550 是聚类的最优参数。

3.3 BiG-SLiCE 聚类全体 BGC

获取最佳的 T 值后，29 万个 BGC 的 gbk 文件被整合到具有 40 个核心的服务器中。而后基于 11608 个真菌基因组的分类信息，我们绘制了 29 万了 BGC 的分类信息表，作为 BiG-SLiCE 的额外输入信息。按照其官方说明，我们将输入文件整理到一个文件夹中命名为 input，开启 bigslice 环境并在相应路径运行命令 bigslice --threshold 550 -i input output。服务器全核心运算约 30 个小时后，得到输出文件，包括可视化的 http 文件和存储结果信息的 sql 数据库文件。基于 Python，我们提取出大量信息，其中最为关键的就是 BGC 的聚类信息和 BGC 的化合物类型预测信息。这些信息被汇总到一张表格中，命名为 all_information。

表 3.1 all_information 信息汇总表（前 10 行）

基因组编号	BGC 编号	化合物类型	GCF 编号	界	门	纲	目	科	属	种
GCA_000002495.2	1	172	54	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	2	144	2801	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	3	172	965	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	4	172	14	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	5	172	634	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	6	145	10	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	7	169	14937	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	8	169	23650	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	9	172	1	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae
GCA_000002495.2	10	others	1968	Fungi	Ascomycota	Sordariomycetes	Magnaporthales	Pyriculariaceae	Pyricularia	Pyricularia oryzae

基于表 3.1 的信息，使用 R 语言和 Excel 等分析工具，我们先简单地查看了各个分类等级中不同物种包含的 GCF 数量，并绘制了图片 3.2（A）。从图 3.2（A）中我们可以得到以下判断，相对于其他分类层级，不同属包括的 GCF 数目更为接近，方差更小。与属级紧邻的种水平和科水平 GCF 分布的方差都要高于属。而余下的三种分类层级——门、纲、目的方差都明显高于科水平。进一步，我们绘制饼图 3.2（B）展示了真菌界的 BGC 类型分布情况，其中非核糖体多肽的 BGC 数量占比最大，达到 36.5%。

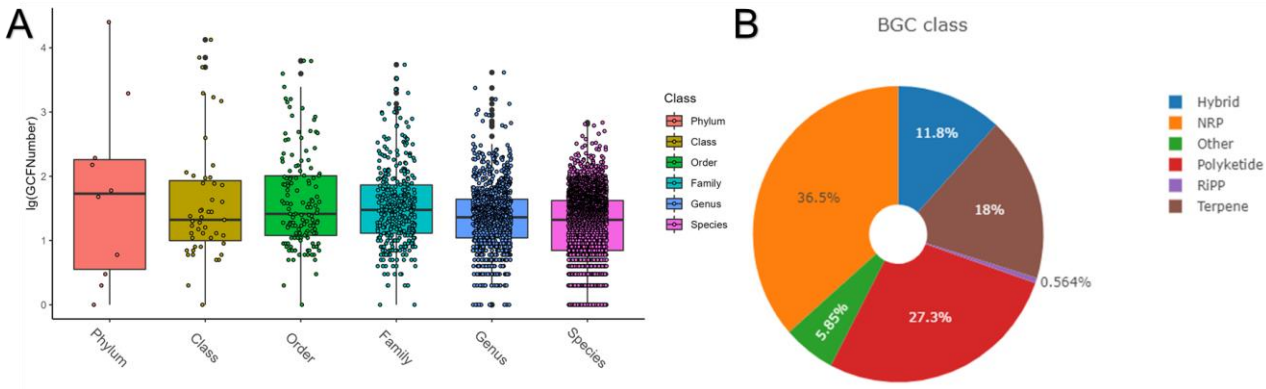


图 3.2 GCF 分布和 BGC 类型分布总览

除了使用 T=550 聚类外，我们还以 T=350,T=400,T=450,T=650,T=750 分别完成了 BiG-SLiCE 的运行，用时 7 天。同样从聚类结果的数据库文件中提取相关信息，用于后续分析。

3.4 聚类结果校验

不同的 T 值会产生不同数量的 GCF，理论上讲搜索半径 T 值越大所得的聚类数目越小，那么我们得到的真菌 GCF 数量是否满足上述规律就需要进行检验。为此，我们找到了在 T=550 的情况下包含 GCF 数目最多的 10 个属，观察它们在不同 T 值下（T=350,T=450,T=550,T=650,T=750）的 GCF 数目变化，绘制了柱状图 3.3（A 图为科，B 图为属）。其中包含 GCF 数量最多的十个科和属随 T 值增加完全按照递减的规律，只是后五种科或属的数量排名有一定浮动。我们可以得出结论，GCF 的绝对数量因阈值而异，但一般趋势（随 T 值增加，从最高到最低的 GCF 数量）在它们之间是一致的。

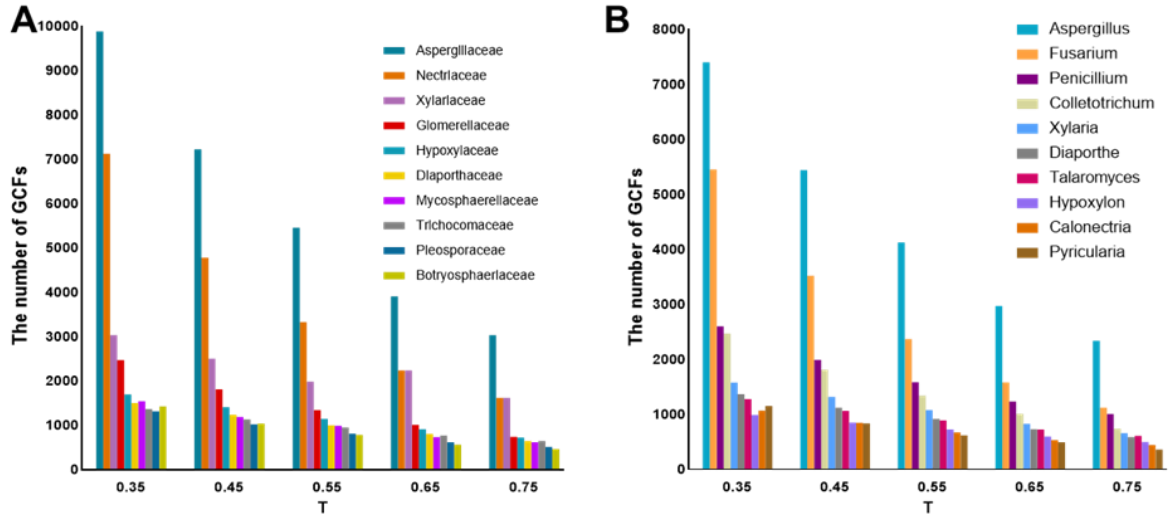


图 3.3 十个科和属在不同 T 值下包括的 GCF 数量柱状图

4. 真菌生物合成多样性的探索

4.1 真菌界生物合成多样性的整体描述

4.1.1 稀疏曲线

稀疏曲线常用于评估样本量是否足够，是从样本中随机抽取一定数量的个体，统计出这些个体所代表物种数目，并以个体数与物种数来构建曲线。它可以用来比较测序不同数量样本的物种丰富度，也可以用来说明样本量大小是否合理。在 Athina Gavriilidou 等人的研究中[20]，稀疏曲线被用于判断生物合成多样性的潜力水平。基于 R 包 iNEXT[27]，在本研究中也使用稀释曲线揭示生物合成多样性潜力，以及真菌基因组的在生物合成问题上的测序深度如何。

4.1.2 真菌界的 GCF 稀释曲线

根据 2 万 6 千个 GCF 每个 GCF 包含基因组的数量，统计得到一个长度为 26825 的向量，输入到函数 iNEXT()中绘制出真菌全部 GCF 的稀疏曲线，其中 GCF 数量代表物种丰富度，而基因组数目反应了样本量大小。从图 4.1 中可知目前真菌界的生物合成多样性随着完成测序的真菌基因组的数目增多仍在高速增长。目前已测序的基因组数目略多余 1 万，稀释曲线显示测序数目达到 5 万时，真菌的 GCF 增速才会缓慢，达到 10 万时，才能揭示整个真菌界的生物合成多样性。因此，探明整个真菌界生物合成多样性的任务我们目前可能才完成 1/10 的工作。目前通过 1 万余个基因组预测得到的 BGC 聚为 2 万 6 千多个 GCF（拐点以后为外推曲线），而稀释曲线预测值的极限趋于 5 万个 GCF。根据 Athina Gavriilidou 等人的研究，在列向量的数据方差很大时，稀疏曲线的极限值会随着 GCF 基数的增加而增加，由此可知，整个真菌界实际的 GCF 数量可能远低于 5 万，目前已经预测得到的生物合成多样性可能不足整体的一半。总之，真菌界的生物合成多样性潜力还有相当大的空间待发掘。

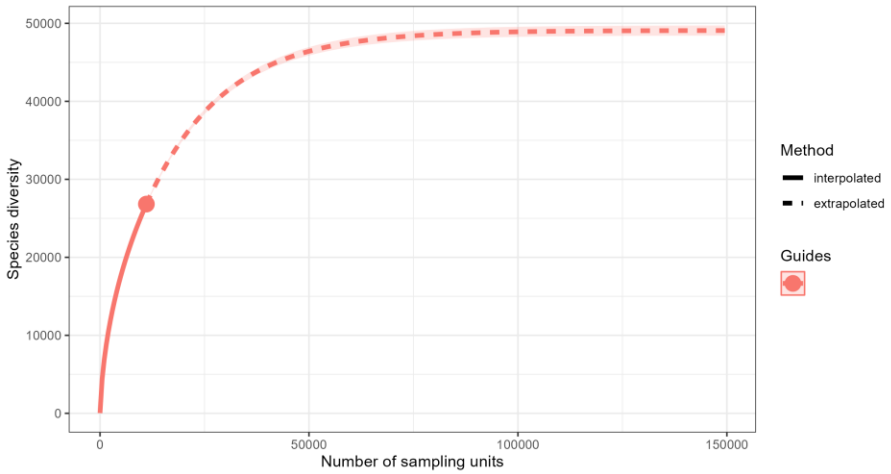


图 4.1 真菌 GCF 的稀释曲线

4.1.3 真菌与细菌的多样性潜力比较

Athina Gavriilidou 等人关于细菌界的生物合成多样性的研究显示，通过预测目前细菌界的全部基因组（超过 17 万个基因组），预测得到了 118 万余个 BGC，能够聚类到 6 万 2 千多个 GCF 中（BiG-SLiCE，T=400）。进而，通过细菌的稀疏曲线，他们发现若要描述细菌界的全部生物合成多样性，需测序超过 150 万个基因组才有可能实现。

在我们的研究中，我们共预测了 11608 个真菌基因组，得到了超过 29 万个 BGC，并通过聚类得到了 2 万 6 千余个基因簇。通过稀疏曲线，我们判断至少需要测序 10 万余个真菌基因组才能揭示整个真菌界的生物合成多样性。

此外，根据 NP Atlas 数据库的信息，目前已发现的细菌天然产物数超过 1 万 2 千种，而真菌的天然产物数超过两万种，两者的聚类数分别是四千余种和六千余种。上述信息汇总为下表：

表 4.1 真菌细菌的比较

数量信息	细菌	真菌
聚类 T 值	400	550
基因组数	>17 万	>1 万 1 千
BGC 数	>118 万	>29 万
GCF 数	>6 万 2 千	>2 万 6 千
待测序基因组数	>150 万	>10 万
已发现天然产物数	>1 万 2 千	>2 万
已发现天然产物聚类数	>4 千	>6 千

虽然本研究和 Athina Gavriilidou 等人的研究方法基本一致，但是直接比较上表中的信息来得出“细菌生物合成多样性高”的结论**并不可靠**。原因如下：

（1） 若以比较真菌细菌间生物合成多样性为目的来设计实验，那么最合理的方法是将真菌和细菌的 BGC 合并，一起聚类。但这与本设计的研究内容关系不大，固没有进一步采用上述方法。

（2） 两次实验真菌和细菌的聚类 T 值不同，在一定程度上影响最终的 GCF 数量。图 4.2 中，在 T 值为 400 聚类真菌的 29 万个 BGC 可以得到近 5 万个 GCF。同时，我们从 118 万个细菌 BGC 中随机抽取 29 万个 BGC，如此抽样 30 次，形成 30 个 BGC 集，计算这些 BGC 集的 GCF 数量的平均值，得到图中稀疏曲线的拐点值约为 3 万个 BGC。进而通过 R 包 iNEXT 计算每个 BGC 集，得到 30 条细菌以 29 万 BGC 为起点的稀疏曲线，将这些稀疏曲线取平均值得到图中红色曲线。而蓝色曲线则是真菌 29 万 BGC 在 T 值为 400 情况下计算得到的稀疏曲线。比较两条曲线我们可以得知，在相同起点（等量的 BGC）相同聚类标准的情况下，真菌预测得到的生物合成多样性要高于真菌。

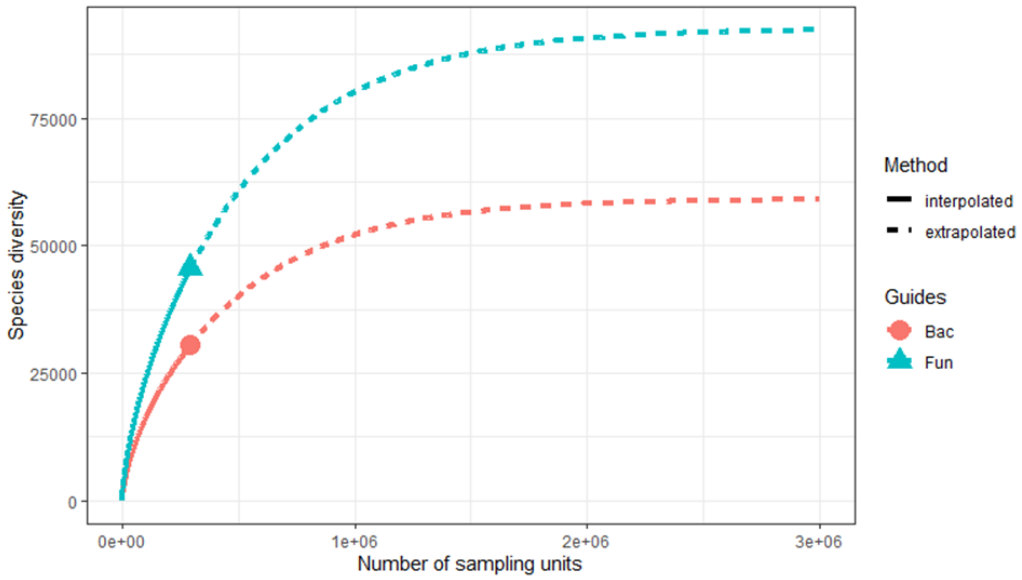


图 4.2 T=400，BGC 为 29 万时真菌和细菌的稀释曲线

（3） 真菌和细菌的基因组和生物合成基因簇存在差别。真菌基因组远比细菌庞大，且生物合成基因簇中的各个基因是非连续的，中间有大量非编码片段相隔开。这使得两者无论在 BGC 的预测上还是在 BGC 的聚类上都有一定差别。真核生物和原核生物的基因组和基因簇本身存在很大区别，加之真菌和细菌在进化上的速度和时间尺度都有不同，因此直接比较两者的 BGC 数目或是 GCF 数目可能存在诸多难以探寻和阐释的问题。

4.2 更低分类水平的生物合成多样性描述

4.2.1 门水平的生物合成多样性

真菌界下一级分类水平就是门，因此基于门水平的 GCF 分布分析能够概览整个真菌界的情况。我们使用 R 语言的 ggplot2 包和 circle 包完成了数据分析和可视化，结果由图 4.3 描述。

由图 4.3 的 A-D 四张图所示结果我们可以得知，在已探明的生物合成多样性（GCF）中，子囊菌门占据了绝大多数（93.6%），而子囊菌门独有的 GCF 的占比也能达到 91.6%，而且在子囊菌门内部其独有的 GCF 更是独占江山。这意味着子囊菌门将是未来挖掘新天然产物的重点目标。占比第二位的担子菌门（7.2%），也同样拥有很高的生物合成多样性潜力，其内部独有的 GCF 占到其全部 GCF 的近八成。另一方面，已知天然产物的聚类数似乎印证了上述结论，子囊菌门以超过 4 千个聚类数占据第一，担子菌门以 819 个聚类数紧跟其后，其他的门类无论 GCF 数量还是化合物聚类数量都很稀少。当然，子囊菌门能够独占鳌头与其已测序基因组的数量远超其他门类有关，达到 9 千多个，因此仅靠这些信息并不能比较各个门潜在的多样性，其与目前真菌研究的偏好和进度有相当大的关系。不过值得注意的是，子囊菌门含有 GCF 数量与其余门类含有 GCF 数量之比要远高于其已测基因组数与其余门已测基因组数之比或是化合物聚类数之比。由此可知，基于目前已探明的 GCF，子囊菌门具有很高的生物合成多样性，最适合深入挖掘。

不同门之间的GCF差别非常显著，其中一部分原因在于不同门内囊括的物种数不同。一般而言物种数量越少具有的生物合成多样性也越少。因此，即使测序的基因组数量足够多，门之间的差异也会显著存在，我们很难从门水平来发掘更多有价值的信息。

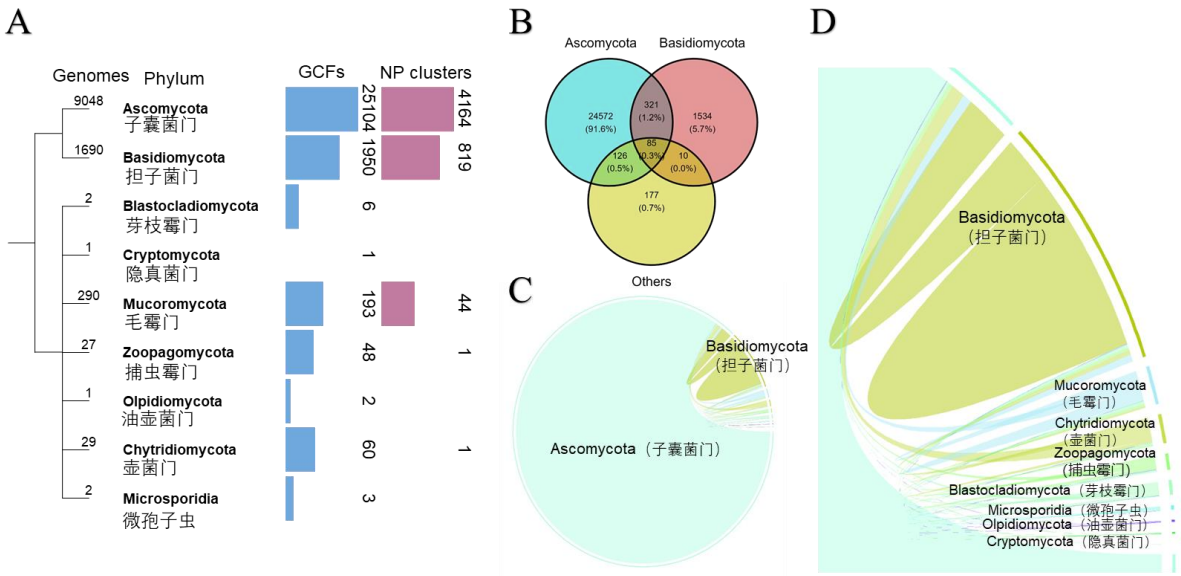


图 4.3 门水平的生物合成多样性刨析

4.2.2 纲水平的生物合成多样性

纲水平相对于门能够揭示更为细致的生物合成信息。通过在线绘图软件和 R 语言的 circle 包我们制作了图 4.4 的 A-D。结合每个基因组含有 BGC 数量以及这些 BGC 的化合物类型信息，我们分别计算了每个纲内的平均每基因组包含的 BGC 数，以及每个基因组包含不同类型 BGC 的平均个数，将这些信息整合并以热图和箱线图的形式可视化得到图 4.4 E，该过程通过 R 语言的 pheatmap 包和 ggplot2 包编程实现。

由图 4.4 的 A-D 可知，纲水平的 GCF 数量分布相对门水平更为均匀，排名前五的纲都来自于子囊菌门。图 4.4 A 中的蓝色柱表示对应纲中包括的 GCF 数量，红色柱表示对应纲中包括的已知天然产物簇数量。可以看出，GCF 数量排名前五的粪壳菌纲、散囊菌纲、座囊菌纲、锤舌菌纲和茶渍菌纲也有大量天然产物簇被发现。此外，壶菌门和捕虫菌门的全部纲都预测有一定数量的 GCF，然而目前几乎没有天然产物被发现，有很高的挖掘潜力。

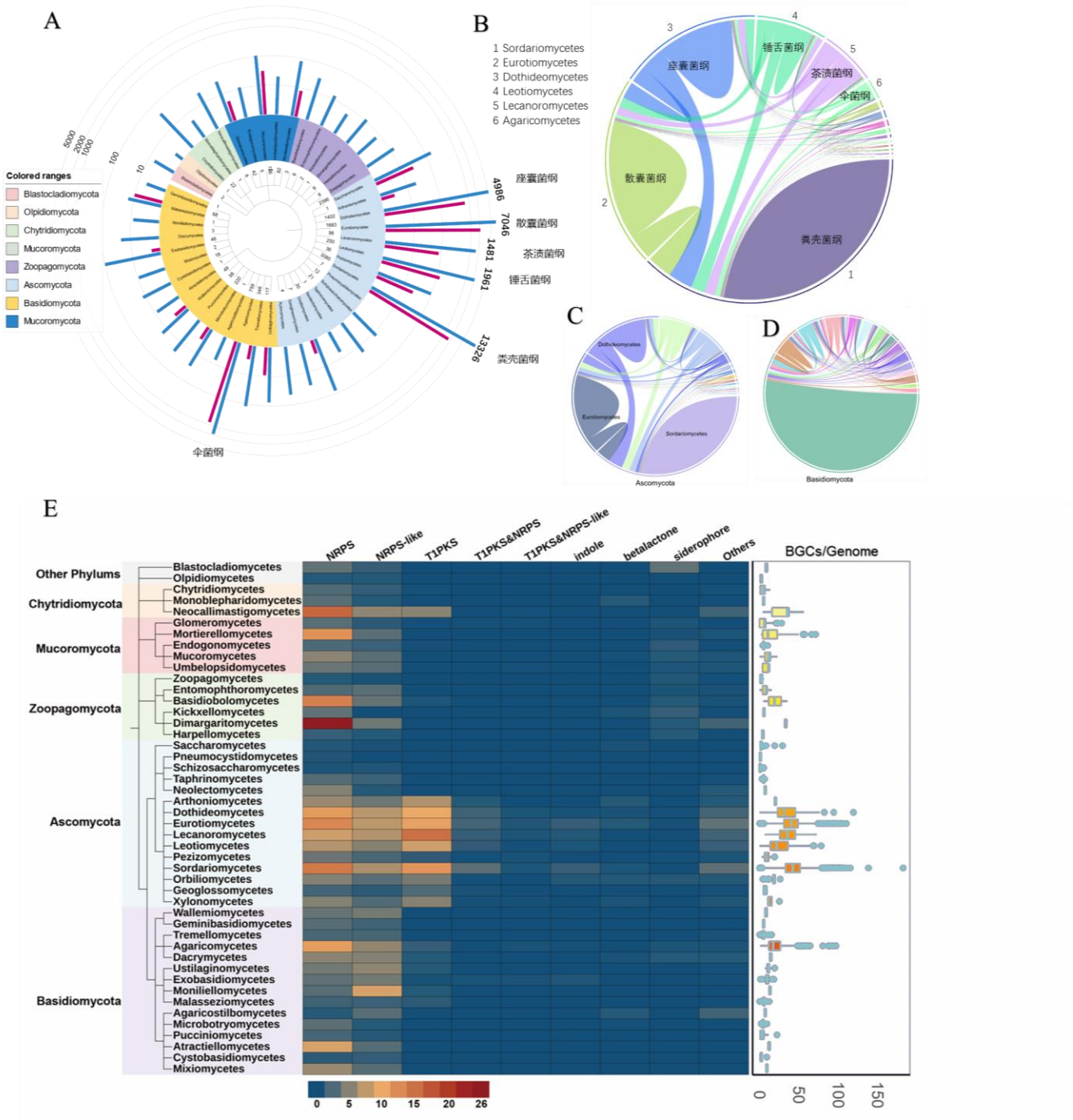


图 4.4 纲水平的生物合成多样性刨析

和弦图 4.4 B 描述了所有纲之间 GCF 的共有和独有关系。图 4.4 C 和 D 分别描述了子囊菌门和担子菌门内部纲之间 GCF 的关系。相对来说，子囊菌门内部的 GCF 分布相对均匀，共有的 GCF 数量也较多，而担子菌门的 GCF 主要分布于伞菌纲内，而各个纲

之间共有的 GCF 数量不多，独有 GCF 的比例较高。另一个规律是，在所有纲中，GCF 含量高的纲往往具有更大比例的独有 GCF，而纲的 GCF 数量越低则独有 GCF 的比例越小。于门水平一样，由于研究偏好和物种数分布不均，纲水平对于生物合成多样性的描述也存在一定偏差。

图 4.4E 的左侧为截止于纲的物种树，中间是每基因组包含不同种类 BGC 的平均个数。右侧的箱线图描述的是纲中每个基因组包含 BGC 数的数据分布情况。综合热图和箱线图我们可以大致判断出不同纲的生物合成多样性水平。热图整行的颜色越深，暖色块囊括的化合物种类越丰富，则说明该纲的生物合成多样性越高。此外，箱线图的水平越高则基因组内含有的 BGC 数量越多说明其生物合成能力越强。一些亲缘关系较近的纲也具有相似的生物合成能力以及合成多样性。例如酵母纲和裂殖酵母纲具有较近的亲缘关系，其生物合成能力和生物合成多样性都极低，而散囊菌纲和茶渍菌纲则共同具有很高的生物合成能力和生物合成多样性。除此以外，将图 4.4 E 同图 4.4 A 的数据相结合，可以发现具有较高 GCF 数量的排名前五的属同样具有很高的每基因组 BGC 数量和更丰富的天然产物类型，可见一个基因组内含有的 BGC 数量越多，其生物合成多样性也趋于越高水平。

4.3 潜在 GCF（pGCF）的预测

4.3.1 寻找适合预测的分类水平

稀疏曲线具备预测潜在多样性的能力，但是需要在列向量方差较小时，才具备一定准确性。为准确预测 pGCF 的数量，我们需要优先选出最适合的分类层级，该水平的列向量方差最小。为此，我们统计了不同水平的低阶 GCF 数量方差分布情况。以属水平为例，低阶就是种水平，曲霉属的低阶 GCF 数量方差就是曲霉属包括的所有物种，它们各自包含 GCF 数量不同而产生的方差，那么属水平的低阶 GCF 数量方差分布就指的是真菌界所有属的低阶 GCF 数量方差的大小分布情况。通过 R 语言的统计学分析和 ggplot2 包的可视化得到图 4.5(A)，该雨云图反映了不同分类水平低阶 GCF 数量方差的差别，其纵坐标取以 10 为底的对数以缩小视觉差异方便绘图。图 4.5(B)展开了含有所有目种含有 GCF 最多的散囊菌目，又进一步展开了散囊菌目中含 GCF 最多的曲霉科，进而是展开曲霉属，柱状图的纵坐标表示包含 GCF 的数量。

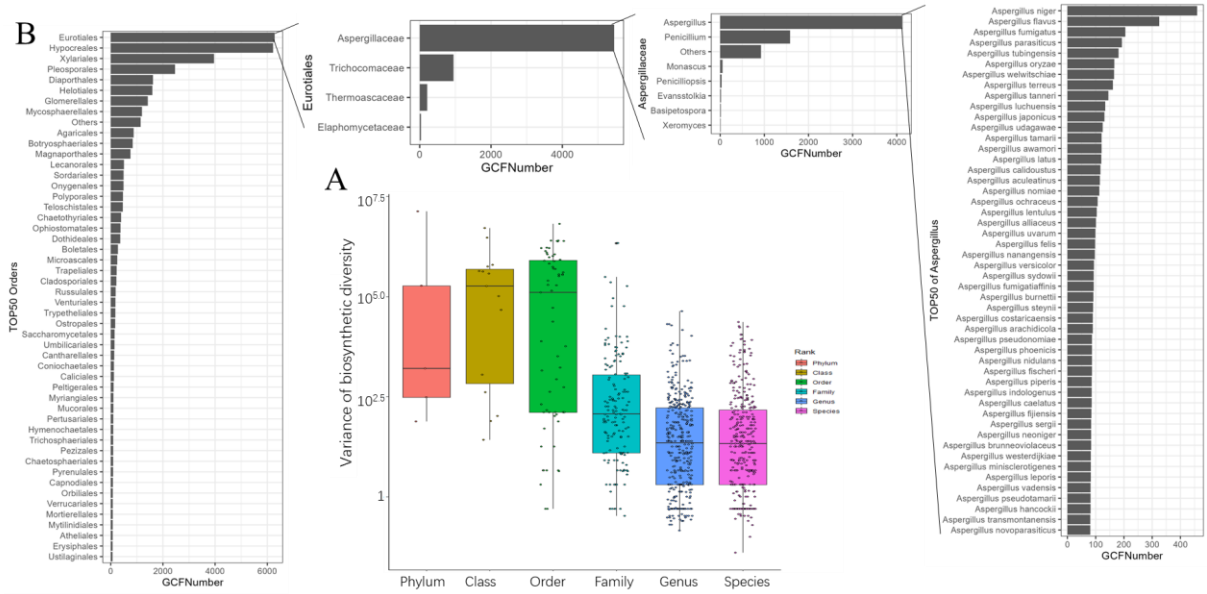


图 4.5 探索合适的分类水平

如图 4.5 A 所示，门、纲、目的方差很大，无法使用稀疏曲线预测得到可靠的 pGCF 数量。科相对于属和种水平其方差较大，属、种是更好的选择。而相较于种，属包含的 GCF 数量更多，分类单位更大，包括物种数更多，深入研究的实际价值高于种水平。系列柱状图 4.5 B 直观地证实了属内部 GCF 的要远比目、科和属的内部分布更均匀，进而方差也更小。因此，属水平是进行多样性比较的最优选。

4.3.2 基于属水平的生物合成多样性挖掘

根据 NCBI 的分类数据，目前已有 1007 个属具有一个及以上 GCF。通过 R 语言的 iNEXT 包，从 1 千多个属中我们筛选出了“含有三个及以上具有 GCF 的物种”的属，绘制了稀疏曲线，并取得了外推值，其与已预测的 GCF 数量的差值即为该属的 pGCF 数。全部稀疏绘制成图 4.6 B。综合 pGCF 值、GCF 数量、到属级的物种树我们绘制了圈图 4.6 A、C、D。进而，我们找到了总 GCF 数量（pGCF 数加上已知 GCF 数）最高的十个属，基于它们的 pGCF 数、GCF 数和天然产物聚类数绘制了柱状图 4.6 E。最后，我们找出了在属水平分布最广泛的前 500 个 GCF，并结合到属级的物种树绘制了 GCF 的分布热图见图 4.6 F。

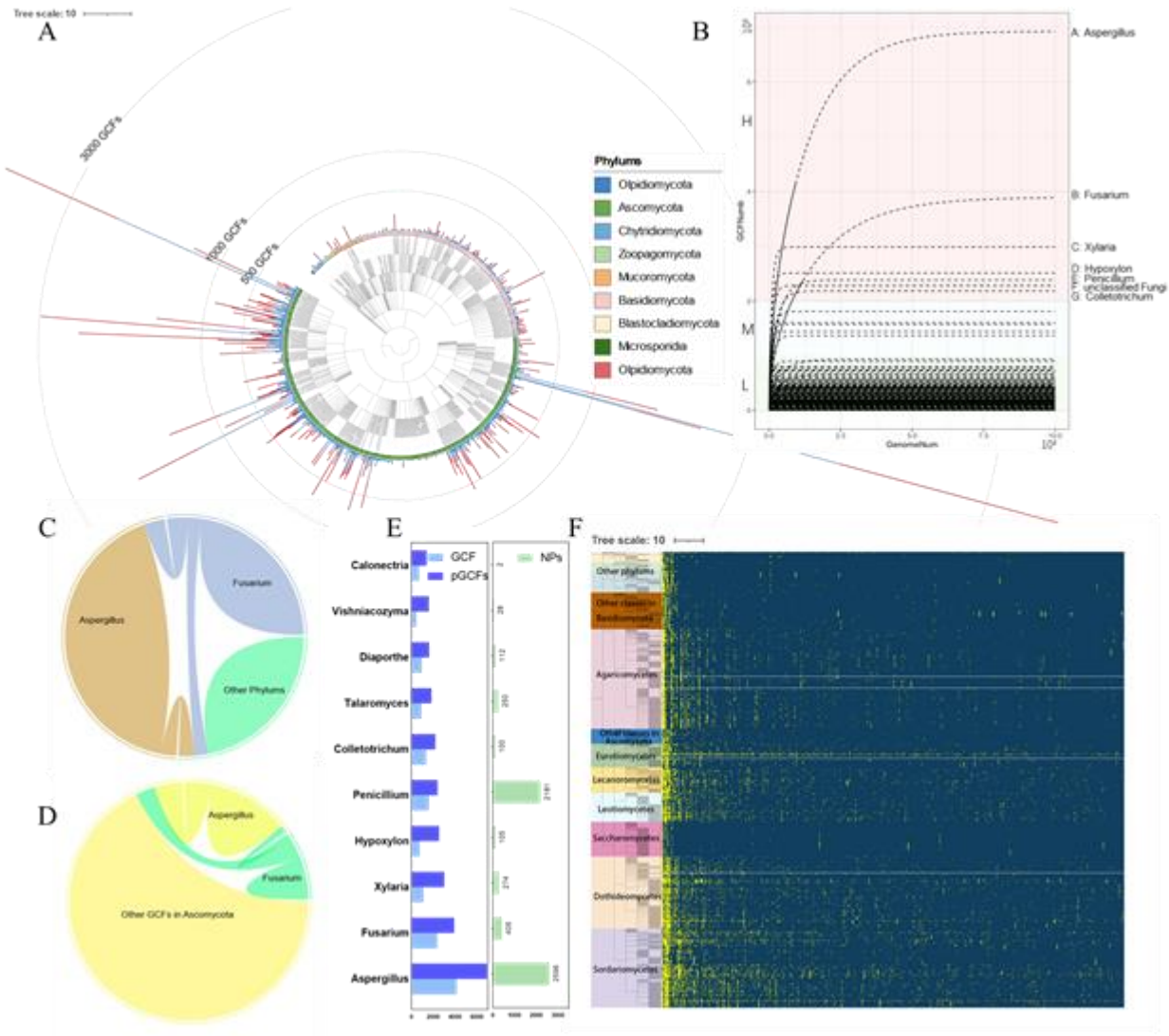


图 4.6 属级生物合成多样性的分析与预测

图 4.6 A 中的柱子高度坐标是实际值，并未处理，对生物合成多样性的展示更加直观。能够明显看出，子囊菌门无论是已知的还是潜在的生物多样性，都远远高于其他门。此外，分类关系密切的属之间的生物合成多样性也极为相似，柱状图的高低起伏有明显的区域性。结合图 4.6 B 可以发现，真菌界的生物合成多样性集中于某几个属内，换句话说，就是处于 H 和 M 区间的十几个属贡献了整个真菌界的绝大多数生物合成多样性。此外，根据稀疏曲线的外推，这些已探明具有很高的生物合成多样性的属仍具有相当大的潜力。曲霉属和镰刀菌属的外推 GCF 数都超过了 3000，两个属的生物合成多样性及其潜力都远远超过其他属，曲霉属更是遥遥领先。由于两个属同属于子囊菌门，和弦图 C 和 D 分别展示了两个属与子囊菌门外的其他门间的分布关系以及与子囊菌门内其他属

之间的分布关系。两个属跟其他门的共有 GCF 数量很少，而在子囊菌门内部两者都有很高的共有比例。不过无论是跟子囊菌门外的属还是子囊菌门内的其他属相比较，两者独有的 GCF 都占据大多数，其比例都要高于共有的 GCF。

图 4.6 E 揭示了很多有趣的信息。十个总 GCF 数最高的属，它们的已探明 GCF 数都少于潜在的 pGCF 数，都有大量的生物合成潜力有待挖掘。比较特殊的属，比如青霉属，因为是研究的热点方向，其已发现的化合物数远比总 GCF 数多于它的几个属要多，比如镰刀菌属无论时已探明 GCF 还是 pGCF 的数量都要大于青霉属，然而其发现的天然产物数量却不足前者的 1/5。由此可见，仅从天然产物的发现数目来判断一类真菌的生物合成多样性是不可靠的，诸如镰刀菌属、炭角菌属和炭团菌属，由于研究冷门，它们已发现的化合物数目都不多，然而它们却具有很强的生物合成多样性及潜力，我们的研究将为天然产物的挖掘提供崭新的方向。

图 4.6 F 所示的热图显的横坐标示了分布最广的前 500 个 GCF 的分布矩阵，右侧的物种树截止到属，其中黄色色块表示该属中含有对应横坐标的 GCF，蓝色表示不包含。GCF 的分布在不同纲之间呈现出显著的差异性，例如子囊菌纲和酵母纲，纲之间的似乎都有一条明显的分界线存在。此外，我们也可以得到以下结论，分布较广的 GCF 会出现在真菌界大多数的属内，相反，真菌界的大多数属内也都包含数量很多的广泛分布 GCF。由此不难推测，这些广布的 GCF 很可能参与真菌必不可少的代谢，产生的天然产物也是真菌赖以生存的化学物质，因此，在这些 GCF 中挖掘出次级代谢产物的可能性不大。

5. 挖掘潜在的真菌 BGC

在前面的工作中，我们从真菌界、门、纲和属水平对真菌生物合成多样性进行了详细的讨论，并了解了生物合成多样性及潜力的分布情况，基本可以确定蕴藏丰富待发掘的天然产物的“矿脉”所在。那么接下来的任务就是如何挖掘出全新的真菌次级代谢产物及其 BGC，我们为此提供了一个具体的挖掘流程。

5.1 GCF 的筛选

通过之前的结论我们可以得知，在真菌界分布广泛的 GCF 往往不具有稀有性，很难发现全新的有价值的天然产物，所以在深入挖掘某一类别的真菌时，应该优先选择独有的 GCF，它们很可能参与合成一些与真菌正常代谢无关的次级代谢产物，发现新成果的概率更高。

除了 GCF 的物种间的分布规律外，GCF 的化学类型在一定程度上也能提供一些信息。每个 GCF 内都含有一个以上的 BGC，这些 BGC 可能属于不同类型的天然产物，根据 BGC 化学类型的比例，我们用比例最大的种类代表整个 GCF 的化合物类别，如此 2 万 6 千多个 GCF 都得到了化合物信息的注释。不同化合物类型之间包括的 GCF 数量不同，其 BGC 序列的复杂度和特异程度也不同，比如非核糖体肽类的基因簇普遍简单，而杂合类（hybrid）的基因组一般具有一些特殊的序列结构，其合成全新化合物的可能性更高。

基于 MiBiG 数据库，我们以及获取了一些已注释 BGC（这些 BGC 被实验验证能够表达出某种天然产物）的信息，这些包含有“已经被研究的 BGC”的 GCF 就不具备深入探索的价值了，我们称这些 GCF 为**已注释 GCF** 因此，在筛选的过程中我们需要优先排除这些 GCF。

根据前面的研究，我们发现在担子菌门的所有属中曲霉属具有最高的生物合成多样性，而所有纲中酵母纲具有最低的生物合成多样性，并且因为曲霉属和酵母是真菌学研究的重点内容，所以本研究中间分析了这两类真菌的内部构成。

通过 R 语言的 pheatmap 包，我们绘制了图 5.1 曲霉属内部 GCF 的分布情况，热图的纵坐标是内部含有的基因组，横坐标是内部含有的 GCF。

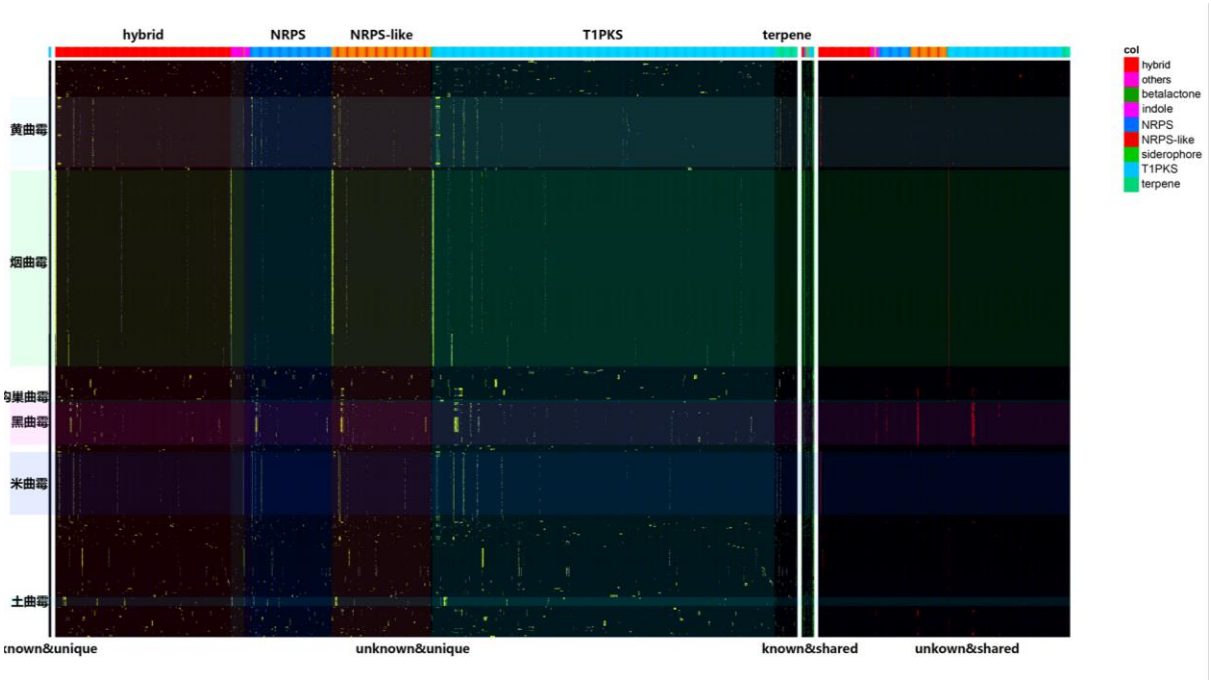


图 5.1 曲霉属的 GCF 分布信息图

曲霉属的 GCF 分布热图被分成四个板块，第一板块是最左侧很细的版块，表示的是曲霉属独有的 GCF 中含有已测定 BGC 的 GCF，第二板块是中间面积最大的板块表示的曲霉属中独有的 GCF 中未含有已测定 BGC 的 GCF，第三板块则表示曲霉属和其他属共有的 GCF 中含有已测定 BGC 的 GCF，而第四板块共有 GCF 中未含有已测定 BGC 的 GCF，在图片最左侧标注了基因组来源的曲霉物种，上方的标签展示了 GCF 的化学类型归属。每个板块又被化合物类型标签分块。

通过图 5.1 我们可以获得以下信息：

（1） 在标注的物种中似乎都具有其种类独有且在种内分布广泛的 GCF。

（2） 曲霉属独有的 GCF 数量要高于共有的 GCF，并且在属内分布也比后者更广泛。

（3） 热图中有很多黄色长线条，我们初步认为这些长线条很可能是对应物种中独有的特异性天然产物的 GCF。然而经过观察这些 GCF 内部含有的 BGC 的序列情况，我们发现这些序列包括的酶类似乎并不具有很高的特异性。我们推测可能是在进化过程中发生过水平基因转移造成了上述现象。

总而言之，从 GCF 分布情况去判断其含有 BGC 是否参与一种全新化合物的合成非常不可靠。而且曲霉属内部的未注释且独有的 GCF 非常多，仅凭这一筛选条件无法有效地筛选出高价值的潜在 BGC。因此，我们需要找到一个更为有效的筛选手段。

5.2 GCF 超聚类

5.2.1 挑选 GCF 的新角度

从大量未知的 BGC 中选出与目前已验证的 BGC 都不相似的 BGC 是一项艰巨的工作。BGC 通过聚类成为 GCF 后，筛选的难度降低，29 万个样本减为 2 万 6 千个。用已验证的 BGC 注释 GCF 后，最终得到 100 余个已注释 GCF，相较于全部的 GCF 这些已注释的部分只是冰山一角。即使我们只考虑曲霉属，并且只保留未注释且独有的 GCF，可供选择的 GCF 依然达到上千个。那么如何缩小搜索范围，准确地找到全新的 BGC 是本研究必须攻克的壁垒。

在 5.1 节中，我们探讨了一些剔除或筛选 GCF 的思路，但是似乎都不是有效且精准的途径。BGC 的本质还是脱氧核糖核苷酸序列，比较 BGC 的相似性或是 GCF 的相似性还是应该回归到序列本身上。幸而，BiG-SLiCE 是一个基于序列比对的聚类工具，它的结果文件提供了包括全体 GCF 的特征矩阵，即为一个行名是 GCF 编号 1~26，825、列名是蛋白域（Domain）编号 1~4，533、值处于 1~255 区间内的特征矩阵。上述的蛋白域来自于各种蛋白质数据库，这些区域往往参与某种代谢途径，如果某个 BGC 的序列恰好含有其对应的 DNA 序列（或是具有很高的相似性），那么这个 BGC 隶属于的 GCF 就视为拥有此特征。矩阵中的值是一种专门用来记录特征量的特殊数据格式，本文不再赘述，详见参考文献[17]。基于该矩阵，我们可以进一步得到 2 万 6 千个 GCF 之间的关系。

5.2.2 GCF 再聚类

在已知 100 余个已注释 GCF 的情况下，如果我们可以把 GCF 再进行聚类，得到 **GCF 超聚类**（GCF Group），那么只要一个超聚类里含一个已注释 GCF，我们就可以把整个超聚类判断为一个**已注释的超聚类**。如此，我们可以剔除大量的已注释超聚类，从而快速缩小搜索空间，该流程如图 5.2 A 所示。

GCF 的再聚类基于 5.2.1 中 GCF 的特征矩阵展开。由于矩阵过于庞大，用一般的聚类手段效率低且精度差，而且聚类半径的超参调节是一个巨大难题。于是，我们打开思维，借鉴最近非常火热的单细胞测序结果分析流程，解决了这一难题。

就像前人把原本是来判断区域内生物多样性的稀疏曲线用于揭示生物合成多样性的潜力一样，我们也把单细胞测序中处理“细胞-基因特征矩阵”的方法挪用到了“GCF-蛋白域特征矩阵”上。由于计数方式的特殊，GCF 特征矩阵无需进行数据的归一化，又因为研究目的不同，我们也不需要对数据进行过滤。因此，在找到方差最大的前 1500

个特征后，我们直接对其进行 PCA 降维处理，而后使用 TSNE[28]方法实现聚类和二维可视化，此过程使用 R 包 Seurat 4.3.0[29]完成，结果由图 5.2 B 展示。

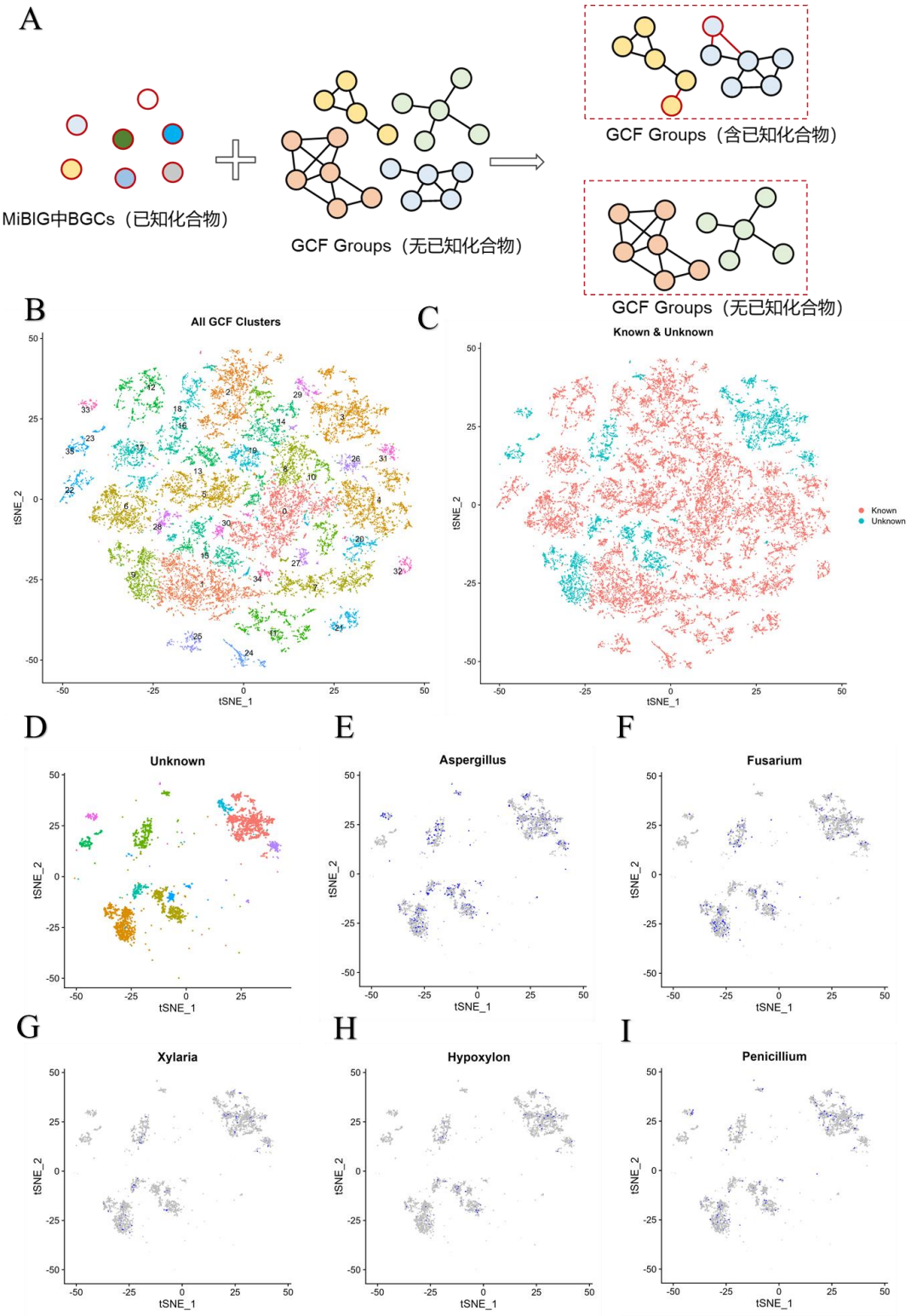


图 5.2 GCF 的超聚类、注释和筛选

而后，我们注释了含有已注释 GCF 的所有超聚类，用红色表面，余下未注释的超聚类用蓝色标明，聚类图变化为图 5.2 B。图 5.2 C 只显示了未注释的超聚类，而图 5.2 E-J 分别为 GCF 数量最高的五个属在图 5.2 C 上的映射图，灰色的点表示对应属不包括的 GCF，蓝色的点表示包含在对应属内的 GCF。

在图 5.2 B 中 2 万 6 千个 GCF 被聚类为 36 个超聚类，经过 GCF 的注释后，26 个超聚类被注释，而仍有 10 个超聚类内不具有任何已注释的 GCF。经过注释后如图 5.2 C，我们下一步的搜索空间缩小为 10 个超聚类的 GCF，超过五分之四的 GCF 被移出。

图 5.2 D 显示了所有未注释的超聚类，进而我们又查看了 GCF 数量最高的五个属在十个未注释的超聚类内的分布情况，其结果显示这些属的 GCF 并没有聚集在某个超聚类内而是广泛分布。

为更加深入的探究 10 个超聚类的内部情况，我们分析了其中 GCF 的数量和物种组成情况。结果显示这些聚类各自都有明显的特异性。比如某种化合物类型的 GCF 比例超过 80%，有的甚至接近于 100%，或是含有两种不同化合物类型的 GCF，它们的比例都很高，占据了绝大比例。上述结果印证了基于单细胞测序原理的聚类的可靠性，也为进一步的筛选和探索提供了重要信息。

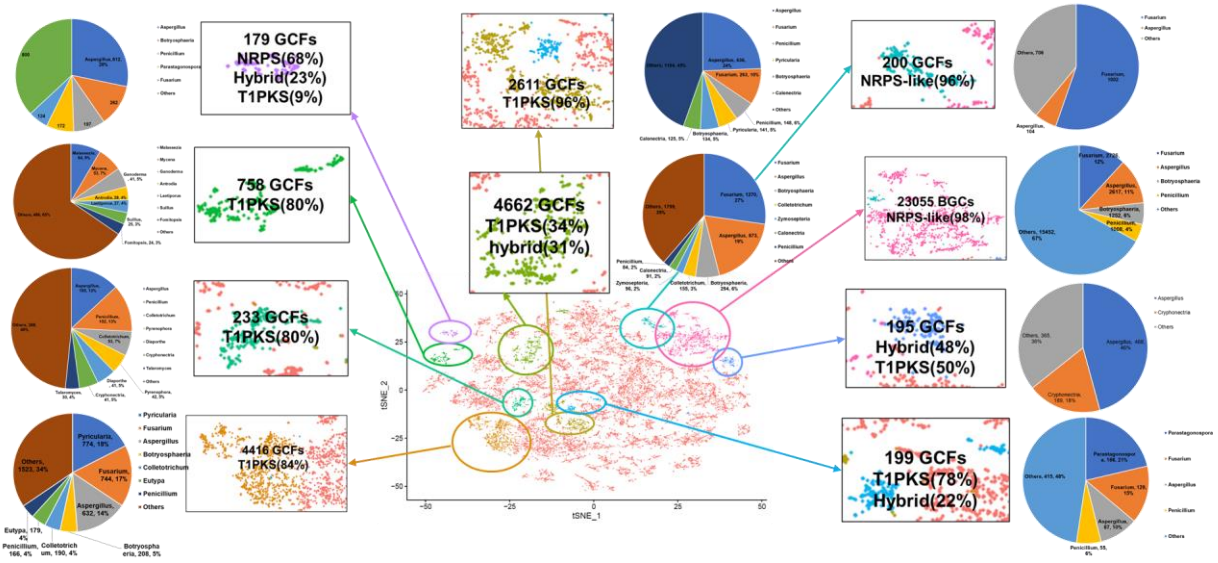


图 5.3 各未知超聚类的内部分析

5.3 目标 BGC 的筛选与验证

根据 5.2 节的研究，我们把筛选范围缩小到了十个超聚类的 GCF 内，并且详细探究了 10 个超聚类 GCF 的组成情况。那么接下来的工作我们以曲霉属为例，进一步挖掘全新的 BGC 和对应的天然产物，并验证挖掘的结果。

在继续挖掘之前，为了验证已注释的超聚类内的其他 GCF 是否与“用于注释它们的已知 BGC”具有相似性，我们选择了超聚类 2 中的几个 GCF 的 BGC，其中包括已知 BGC，进行了序列比对。这些 GCF 包括五个不含有已知 BGC 的 GCF，分别为 GCF19235，GCF18893，GCF16139，GCF14136，GCF14317 和两个包含已知 BGC 的 GCF，为 GCF23558 和 GCF24078。GCF23558 和 GCF24078 内部都含有天然产物洛伐他汀的对应基 BGC，除了这两个 BGC 外，我们又额外比对了两个 GCF 内各一个其他的 BGC。在另五个 BGC 中，我们从中分别随机选择了一个 BGC 加入比对，因此我们共比对了 9 条 BGC 序列，比对结果如图 5.4 所示。

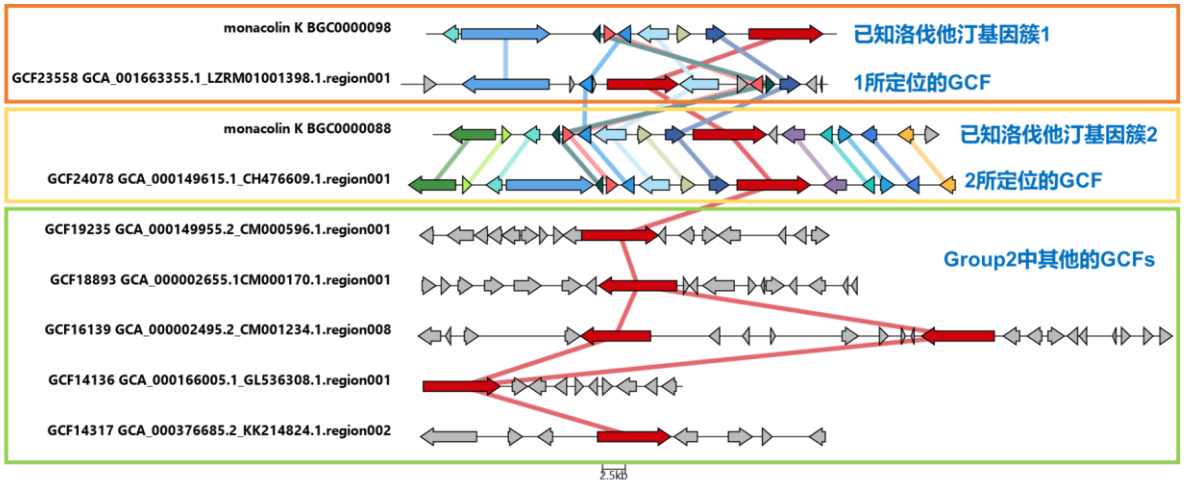


图 5.4 已注释 GCF 超聚类的内部相似性验证

从图 5.4 中我们可以得知，9 个来自同一个超聚类和 7 个不同 GCF 的 BGC 具有相同的核心基因，并且隶属于同一个 GCF 的 BGC 具有很高的序列相似性。通过上述比对，我们可以推断，一个超聚类内的 BGC 都有某个共同的核心基因（或核心蛋白域），这也是它们能被聚类于一个超聚类内的重要原因。这一结论与单细胞测序的注释方式相符合。在单细胞测序中，每个细胞簇都具有几个“标志基因”，也就是通过这些基因研究者可以注释每个细胞簇的细胞类型。那么，在我们的研究中，“标志基因”仍然存在，只不过变成了整个簇内 BGC 的核心蛋白域或是核心基因，反过来，这些特征也是聚类

分群的重要参考。综上，借以单细胞测序技术来完成 GCF 的超聚类具有充分的生物学意义和准确性。

在证实了超聚类的可靠性后，接下来我们将把目光移向我们更关心的属于未注释超聚类的 GCF 中。在本研究中，我们选择以曲霉属为例探究其潜在的高价值 GCF。选择曲霉属作为样例的原因如下：第一，曲霉属的 GCF 和 pGCF 数量都是所有属中最高的，无论时已知的生物合成多样性还是潜在的多样性都独占鳌头，理应优先研究。其次，曲霉属一直是真菌研究的重点物种，能够从中挖掘出全新的天然产物将为真菌次级代谢的研究做出巨大贡献；第三，在十个超聚类中，曲霉属在其中九个中都有出现，有更好的代表性。最后，曲霉属的菌株较易获取和培养，有利于开展后续的实验验证。

我们先是筛选出了曲霉属独有的全部 GCF 并与来自 9 个未知超聚类的 GCF 取交集，再从交集的 GCF 中手动挑选出一些结构较为特殊的 BGC。通过可视化软件，我们展示了这些 BGC 的核心基因组成和序列结构，在综合了全部 9 个超聚类的 BGC 序列后，绘制了图 5.5。

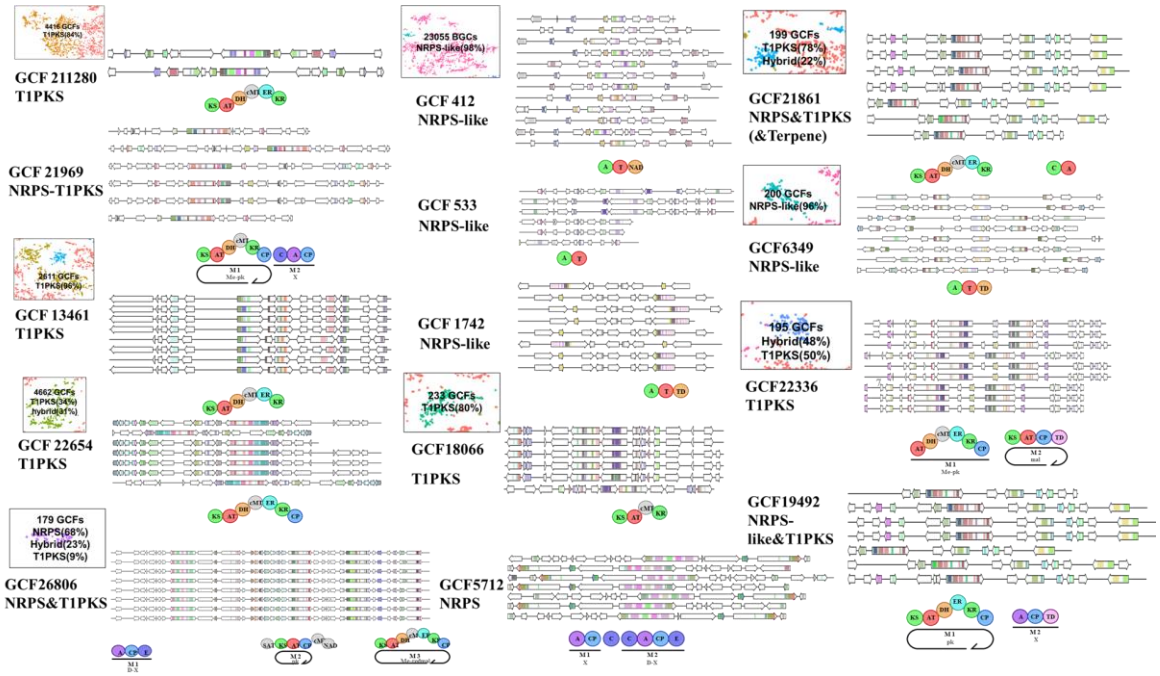


图 5.5 各未知 GCF 超聚类的 BGC 序列信息总览

从图 5.5 中我们不难看出，来自于同一 GCF 的曲霉属独有 BGC 之间有着极高的序列相似性，并且都具有相同的核心基因。而来自不同 GCF 却是同一超聚类的曲霉属独有 BGC，它们也具有一定的序列相似性，并且它们的核心基因也十分相似。

总而言之，通过整个低 5 节的工作，我们筛选出了如图 5.5 所展示的数十个潜在的 BGC 以供参考，它们很可能具有表达出全新天然产物的能力。那么下一步的工作就是通过实际的微生物学实验，外源表达出上述 BGC 的产物，并鉴定这些产物的化学组成，以证明我们挖掘的成果。当然，这一部分工作并不包括在本设计中，不过仍与本设计归属于同一研究课题。

6. 讨论与展望

从真菌界的角度出发，我们能够得到一些令人兴奋的结论：对于真菌，我们目前能够预测到的生物合成多样性很可能不足实际的一半，而我们已经实验测定的已知的生物合成多样性水平又不足能够预测到的 1/4。这意味着，对于真菌，我们已探明的天然产物种类只有自然界全部天然产物种类数的 1/10 左右，宛如冰山一角。而假若我们想探明水面下的冰山（预测出全部的 GCF），我们要做的全部工作大概是目前的十倍，即测序超过 10 万个真菌基因组。若进一步要挖掘冰山中的宝藏（实验发现全部的天然产物簇），那么需要完成的工作将难以估量。因此，真菌的生物合成多样性就像一个待开采的大矿脉，人们目前只是敲开了其一角，就算这样人类也已经从真菌中获得了大量宝贵的化学产物，诸如各种抗生素、发酵物，那么可想而知，如果有一天人类能够找到一种能快速探明和挖掘天然产物的高效手段并借此把整个真菌界的天然产物为己所用，则带来的价值将无法想象。千里之行，始于足下，本研究为探索真菌天然产物冰山迈出坚实一步。

从纲和门水平我们也能窥探到大量生物合成多样性信息。已知生物合成多样性在门和纲水平的分布相当不均匀，不同真菌的研究热度不同可能是影响因素之一，但是可以明确的结论是子囊菌门内蕴藏的天然产物仍然浩如烟海。此外，另一个在意料之中的结论是，亲缘关系较近的纲其生物合成多样性也相似，甚至是包含的 BGC 类型也趋于一致。从稀疏曲线预测的 pGCF 数据来看，几乎所有纲的 GCF 数目都少于其 pGCF 数目。可以得知，对于真菌界的全部纲，目前已探明的生物合成多样性都未达到它们全部生物合成多样性的一半。在真菌界的很多纲中，我们能够利用的天然产物只占已发现的天然产物极小部分，而已发现的天然产物又是已探明的生物合成多样性的极小部分，现在我们知道，已探明的生物合成多样性又是整个生物合成多样性的极小部分。由此可见，人类在真菌生物多样性的挖掘道路上还有多少工作需要完成。

在属水平上，我们可以更加清晰地观察到亲缘关系的远近与生物合成多样性水平高低的相关性。不过，令人欣慰的是，在属水平上，我们能够看到科学家们为真菌生物合成研究做出的努力。我们可以看到，青霉属的 pGCF 数量要小于 GCF 的数量，这意味着青霉属的大多数生物合成多样性已经能够探明。相较于纲水平来说，在更细化的某个小领域中，人们夜以继日地探索确实能够取得一定成果（当然，若要探明全部 pGCF，付出的努力肯定远比之前的多，具体原因参见 4.1.2 节内容）。除此之外，青霉属内已发现的天然产物类别数目也是惊人的，虽然这个数目相较于其全部的天然产物可能只是沧海一粟，但是仍然证明了人类已有的天然产物探索工作确实能够实现一定突破。而另

一个具有更大量已知天然产物的曲霉属，它也是人类生物合成多样性探索的主要目标之一。曲霉属已测序的真菌基因组数目是所有属中最多的，目前已发现的天然产物也达 2 千余类。然而我们的研究发现，目前针对曲霉属的研究仍然只是一个开始，其稀疏曲线显示，如要探明曲霉属的全部 GCF，已有的测序深度不足需求的十分之一。由于曲霉属稀疏曲线真实值的截止点相对靠前，因此在未来的很长一段时间里，每测序一条曲霉属的基因组都可能带来全新的 GCF，也就是一类全新的天然产物。而对于曲霉属和青霉属以外的其他属，其 GCF 的探索工作仍然是一项有待完成的任务，更不用说挖掘出多少天然产物，其蕴藏的生物合成宝藏是难以估量的。

在我们工作的后半部分旨在发现一种行之有效的 BGC 挖掘流程。Nickles 的研究提出基于核心基因序列比对的 GCF 超聚类手段[30]，而 BiG-SliCE 的开发者 Kautsar 等人也实现过 GCF 超聚类[17]。在我们的研究中，一种全新的思路被采用。我们借鉴单细胞测序数据分析过程中细胞聚类的原理，将 26，825 个 GCF 聚类为 37 个超聚类。完成注释后，只剩下 10 个未注释的超聚类，约占 4/5 的 GCF 被去除。从未知的 GCF 中挑出一些更可能包含全新类型天然产物的 GCF 是一个在未知中筛选未知的困难过程。因此在这个过程中，我们需要排除模棱两可的目标，而保留那些与目前已知的 BGC 序列相似度很低的 GCF。整个 GCF 超聚类里包含成千上万个 BGC，只要其中一个 BGC 是已知的，我们就认定整个超聚类为“已注释”。虽然超聚类里极有可能包含很多全新的 BGC，但是因为我们还有相当数量“更优”的超聚类（它们内部连一个已注释的 BGC 都没有）以供选择，所以我们可以大胆地舍去这些已注释的超聚类。经过上述工作，在余下的 GCF 内，我们选取任何一个 BGC 作为实验对象都有发现全新天然产物的可能性，因此后续过程中我们以实验为导向，挑选了一些 GCF 和 BGC 以供参考。

我们的研究为真菌生物合成的研究提供了一张较为详细的蓝图，当然这张图纸还有很多空白等待填充。目前，人们探索生物合成多样性的进度还未进行到一半，而挖掘真菌天然产物的进度更是刚刚起始。对于生物合成多样性的探索我们可以保持一个乐观的预期，随着真菌基因组的测序深度增加，凭借目前已有的生物信息学工具探明绝大多的数真菌界生物合成多样性指日可待。但是我们目前缺少高效且低成本的天然产物挖掘，并且挖掘的范围也只是集中在某几个热门的属内，对于绝大多数其他属天然产物的认知仍停留在很低的水平。真菌的生物合成资源是一个浩瀚如烟的巨大矿藏，通过生物信息学的照明器我们能够看清矿脉的潜在位置，也能指导开采的方向，可是决定开采速度的既有探测效率还有挖掘效率。开发一种全新的挖掘手段，可能是真菌生物合成研究的下一关键突破口。

结　论

近年来基于几种强大的生物信息软件，从真菌 BGC 到真菌次级代谢产物这条全新的挖掘路线愈发清晰。真菌的 BGC 除了能够用于表达对应的天然产物外，凭借其序列信息还能揭示进化轨迹。在我们的研究中，经过聚类，BGC 被归纳到上万个 GCF 中，根据 GCF 的分布情况和外推值我们成功地探明了真菌界的生物合成多样性及潜力。在我们的研究进行的同时，全世界的科研工作者也在致力于真菌次级代谢产物的开采。比如 Lindsay K. Caesar 等人基于化学的质谱数据，尝试从代谢组学的角度挖掘全新的真菌天然产物，并取得了重大突破[31]。再比如 Grant R. Nickles 等人通过 BGC 的进化关系和核心基因将 GCF 再聚类，最终成功指导了非规范异腈合酶类 BGC 的挖掘工作。可见，如何快速有效地挖掘出新颖的 BGC 及其天然产物是真菌代谢领域正在面临的挑战。同样，我们的工作参照单细胞测序数据的聚类流程解决了 GCF 的再聚类问题，进一步地缩小了搜索空间，也为 BGC 的挖掘难题提供了一种特别的解决思路。

我们的研究从整个真菌界入手，收集了 11，608 个真菌基因组，并使用 BGC 预测工具 antiSMASH 预测了全部基因组的 BGC。而后，使用大规模聚类工具 BiG-SLiCE 聚类了 293，926 个真菌 BGC，并得到了 26，825 个 GCF。GCF 模拟了真菌天然产物的聚类，因此，基于 GCF 的分布和数量我们真菌界门、纲和属水平的生物合成多样性。进一步，借助稀疏曲线，我们在属水平上预测潜在的 GCF 数量，并结合已知的化合物数据库揭示了具有挖掘潜力的属。借助单细胞测序数据的处理流程，我们依据 GCF-蛋白域特征矩阵聚类超聚类了 GCF，并通过已知的 BGC 注释了这些超聚类。去除被注释的超聚类内的 GCF 后，超过 4/5 的 GCF 不再纳入搜索范围。在余下不足 1/5 的 GCF 内我们以曲霉属为例找出了数个曲霉属独有的 GCF。这些 GCF 内所包括的 BGC 将作为我们挖掘的成果，接受进一步的实验验证。

总而言之，我们的研究窥探了真菌界的生物合成多样性及其潜力，探寻了真菌天然产物的全新挖掘方向，并且提供了一种基于生物信息学的挖掘流程，通过该方法我们最终筛选出了数十个很可能表达全新代谢产物的 BGC 以供实验鉴定。

参 考 文 献

[1] 马紫卉, 李伟, 尹文兵. 真菌天然产物异源生产研究进展 %J 微生物学报 [J]. 2016, 56(03): 429-40.

[2] Nesbitt B F, O'kelly J, Sargeant K, et al. Aspergillus flavus and turkey X disease. Toxic metabolites of Aspergillus flavus [J]. Nature, 1962, 195: 1062-3.

[3] Trail F, Mahanti N, Rarick M, et al. Physical and transcriptional map of an aflatoxin gene cluster in Aspergillus parasiticus and functional disruption of a gene involved early in the aflatoxin pathway [J]. Applied and environmental microbiology, 1995, 61(7): 2665-73.

[4] Lind A L, Lim F Y, Soukup A A, et al. An LaeA- and BrlA-Dependent Cellular Network Governs Tissue-Specific Secondary Metabolism in the Human Pathogen Aspergillus fumigatus [J]. mSphere, 2018, 3(2).

[5] Lysøe E, Seong K Y, Kistler H C. The transcriptome of Fusarium graminearum during the infection of wheat [J]. Molecular plant-microbe interactions : MPMI, 2011, 24(9): 995-1000.

[6] Caesar L K, Montaser R, Keller N P, et al. Metabolomics and genomics in natural products research: complementary tools for targeting new chemical entities [J]. Natural product reports, 2021, 38(11): 2041-65.

[7] Chavali A K, Rhee S Y. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites [J]. Briefings in bioinformatics, 2018, 19(5): 1022-34.

[8] Robey M T, Caesar L K, Drott M T, et al. An interpreted atlas of biosynthetic gene clusters from 1,000 fungal genomes [J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(19).

[9] Stępień L. The use of Fusarium secondary metabolite biosynthetic genes in chemotypic and phylogenetic studies [J]. Critical reviews in microbiology, 2014, 40(2): 176-85.

[10] Campbell M A, Rokas A, Slot J C. Horizontal transfer and death of a fungal secondary metabolic gene cluster [J]. Genome biology and evolution, 2012, 4(3): 289-93.

[11] Keller N P. Fungal secondary metabolism: regulation, function and drug discovery [J]. Nature reviews Microbiology, 2019, 17(3): 167-80.

[12] Bignell E, Cairns T C, Throckmorton K, et al. Secondary metabolite arsenal of an opportunistic pathogenic fungus [J]. Philosophical transactions of the Royal Society of London Series B, Biological sciences, 2016, 371(1709).

[13] Lind A L, Wisecaver J H, Lameiras C, et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species [J]. PLoS biology, 2017, 15(11): e2003583.

[14] Droce A, Saei W, Jørgensen S H, et al. Functional Analysis of the Fusarielin Biosynthetic Gene Cluster [J]. Molecules (Basel, Switzerland), 2016, 21(12).

[15] Campbell M A, Staats M, Van Kan J A, et al. Repeated loss of an anciently horizontally transferred gene cluster in Botrytis [J]. Mycologia, 2013, 105(5): 1126-34.

[16] Blin K, Shaw S, Kloosterman A M, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities [J]. Nucleic acids research, 2021, 49(W1): W29-w35.

[17] Kautsar S A, Van Der Hooft J J J, De Ridder D, et al. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters [J]. GigaScience, 2021, 10(1).

[18] Navarro-Muñoz J C, Selem-Mojica N, Mullowney M W, et al. A computational framework to explore large-scale biosynthetic diversity [J]. Nature chemical biology, 2020, 16(1): 60-8.

[19] Yang J, Cai Y, Zhao K, et al. Concepts and applications of chemical fingerprint for hit and lead screening [J]. Drug Discovery Today, 2022, 27(11): 103356.

[20] Gavriilidou A, Kautsar S A, Zaburannyi N, et al. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes [J]. Nature Microbiology, 2022, 7(5): 726-35.

[21] Terlouw B R, Blin K, Navarro-Muñoz J C, et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters [J]. Nucleic acids research, 2023, 51(D1): D603-d10.

[22] Blin K, Shaw S, Steinke K, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline [J]. Nucleic acids research, 2019, 47(W1): W81-w7.

[23] Kloosterman A M, Shelton K E, Wezel G P V, et al. RRE-Finder: A Genome-Mining Tool for Class-Independent RiPP Discovery [J]. 2020: 2020.03.14.992123.

[24] Medema M H, Kottmann R, Yilmaz P, et al. Minimum Information about a Biosynthetic Gene cluster [J]. Nature chemical biology, 2015, 11(9): 625-31.

[25] Crits-Christoph A, Diamond S, Butterfield C N, et al. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis [J]. Nature, 2018, 558(7710): 440-4.

[26] Rosenberg A, Hirschberg J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure; proceedings of the Conference on Empirical Methods in Natural Language Processing, F, 2007 [C].

[27] Hsieh T C, Ma K H, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers) [J]. 2016, 7(12): 1451-6.

[28] Zhou H, Wang F, Tao P. t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations [J]. Journal of Chemical Theory and Computation, 2018, 14(11): 5499-510.

[29] Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data [J]. Cell, 2019, 177(7): 1888-902.e21.

[30] Nickles G R, Oestereicher B, Keller N P, et al. Mining for a New Class of Fungal Natural Products: The Evolution, Diversity, and Distribution of Isocyanide Synthase Biosynthetic Gene Clusters [J]. bioRxiv : the preprint server for biology, 2023.

[31] Caesar L K, Butun F A, Robey M T, et al. Correlative metabologenomics of 110 fungi reveals metabolite–gene cluster pairs [J]. Nature chemical biology, 2023.

修改记录

一、 毕业设计（论文）题目修改

原题目：真菌次级代谢产物基因簇的预测和多样性分析

新题目：真菌界生物合成基因簇的预测、聚类和多样性分析

二、校外毕业设计（论文）时间节点记录

本人于 2023 年 4 月申请到中国科学院微生物所做毕业设计（论文）

指导教师为：吴琦

校内指导教师为：韩彦粲

2023 年 5 月 24 日回到学校。

三、 毕业论文（设计）内容重要修改记录

1.结论内容修改：总而言之，我们的研究窥探了真菌界的生物合成多样性及其潜力，探寻了真菌天然产物的全新挖掘方向，并且提供了一种基于生物信息学的挖掘流程，通过该方法我们最终筛选出了数十个很可能表达全新代谢产物的 BGC 以供实验鉴定。

四、毕业论文（设计）内容重要修改记录。

2.图 5.2 图片清晰度和字母标号修正。

3.所有表的表头位置修改

4.首次出现的英文缩写添加英文全文

5.图 4.1 清晰度提高

五、毕业论文（设计）外文翻译修改记录

1.修改字体字号

2.修改各小标题格式

3.提高图 1 的清晰度

六、毕业论文（设计）正式检测重复比

重复比为 2%。

记录人（签字）：

指导教师（签字）：

致　　谢

感谢微生物所吴琦研究员指导并监督本次毕业设计，为实验提供新思路、解答技术难题、完成技术路线设计。感谢微生物所尹文兵课题组的尹文兵研究员、研究生张姝和研究生王子陌，课题组为实验提供了强有力的理论支持、设备支持和数据支持；尹文兵老师参与了实验设计、引导实验大方向、审核实验进度；张姝同学参与了部分工作内容，付出了大量时间和经历制作图片、处理数据、准备汇报内容;王子陌同学在工作进行的前期积极地参加问题讨论和思路设计，为前期实验的推进做出了巨大贡献。同时感谢中科院微生物所为本人提供了良好的工作环境和优质的科研平台。

此外，感谢学院的韩彦粲副教授，与校外指导老师共同监督和审查本次设计；感谢学院贺雷雨老师，解决了为本研究使用服务器的系列问题；感谢学院、校区和学校能够同中科院微生物所进行联合培养计划，为本人在北京的生活提供了充足的政策支持和经济支持。在此我也希望学院能够继续同中科院微生物所深入合作，为学校为国家培养更多的科研人才。