全球微生物组整体结构和功能的搜索

Taxonomic and Functional Search of Global Microbiomes on the Whole-Microbiome Level

赵丰洋 [1]，李坚 [1]，荆功超 [2]，苏晓泉 [1$*]

1 计算机科学技术学院，青岛大学，青岛市，山东省；

2 单细胞中心，中国科学院青岛生物能源与过程研究所，青岛市，山东省；

$现工作单位：计算机科学技术学院，青岛大学，青岛市，山东省

*通讯作者邮箱：suxq@qdu.edu.cn

摘要

来自不同环境的微生物组样本的数量正在迅速增长，随之而来的问题也不断涌现，例如如何快速的找到具有某种群落结构和功能的数据集，以及通过对比发现新的微生物组与现有样本之间的关联等。Microbiome Search Engine 2（以下简称 MSE 2）是第二代的微生物组搜索引擎，为解决以上类型的问题提供良好的解决方案。MSE 2 可以根据某个微生物组整体的物种结构或功能特征，在全球已有的微生物组数据集中搜索与其高度匹配的样本。MSE 2 由以下三部分组成：（i）不断更新的微生物组数据库。该数据库目前包含来自于 819 项研究的 266,000 多个宏基因组和 16S rRNA 扩增子样本，每一个样本的测序数据和元数据（metadata）都进行了统一化处理；（ii）增强的搜索引擎。实时级快速搜索，能够在 0.5 秒内，从整个数据库中搜索到与给定的微生物组在整体物种或功能组成上最相似的样本；（iii）基于 Web 的图形界面。用户可通过 http://mse.ac.cn 免费访问 MSE 2。该网站提供了简单易用的图形界面，方便用户快速上手样本搜索、数据浏览等操作，同时也为自定义的搜索提供了教程。

关键词: 扩增子，宏基因组，微生物组，在线服务，搜索引擎

材料与试剂

本研究不涉及传统试剂耗材。

仪器设备

连接互联网的个人电脑。

软件

MSE 2 (Jing 等， 2021)为 web 端软件，采用浏览器访问使用。建议使用 Firefox、Chrome 或 Microsoft Edge 浏览器。微生物组测序序列数据的预处理，根据测序序列类型和搜索模式，处理软件详见**表 1**。

实验步骤

1. 微生物组测序序列数据预处理

MSE 2 可以从群落物种组成和功能组成两个角度进行搜索，兼容 16S rRNA 基因扩增子（以下简称 16S 扩增子）测序数据和鸟枪法宏基因组测序（Shotgun Metagenomic Whole Genome Sequencing；以下简称 WGS）数据。搜索之前需要将测序序列按照相应的序列类型和搜索类型进行预处理，使之具有与数据库样本相同的分析标准和数据形式。微生物组样本需以 OTU（Operational Taxonomy Unit）、物种（species）或 KO（KEGG Orthology）功能注释作为搜索输入，不同测序类型、不同搜索角度的输入格式及推荐的处理软件如**表 1** 所示：

表 1. 不同测序类型、不同搜索角度的输入格式对应关系

	测序类型	按物种组成搜索	按功能组成搜索
16S 扩增子	搜索输入格式：	OTU （**表 2**）	KO （**表 3**）
	推荐的序列处理软件：	Parallel-Meta 3 (Jing 等， 2017) 详见实验步骤 1.1	Parallel-Meta 3 (Jing 等， 2017) 详见实验步骤 1.2
WGS	搜索输入格式：	物种（**表 4**）	KO （**表 3**）

推荐的序列处理软件：	MetaPhlAn 2 (Tin 等， 2015) 详见实验步骤 1.3	HUMAnN 2 (Franzosa 等， 2018) 详见实验步骤 1.4

1.1 16S 扩增子的 OTU 组成预处理

16S 扩增子序列需要以 97%的相似度与 GreenGenes 参考数据库 (McDonald 等， 2012)比对来得到 OTU 及其丰度信息。推荐使用 Parallel-Meta 3 软件进行序列预处理。以 16S 扩增子测序文件 sample1.fa 为例，采用 Parallel-Meta 3（3.6 版本）的处理命令为：

PM-parallel-meta -r sample1.fa -o sample1.out -f F -v F

输出目录 sample1.out 下，文件 classificatioin.txt 即为符合条件的 OTU 搜索输入文件。该文件首行为标题行，以符号#开头，其余内容主要包含两列，即 GreenGenes 数据库 OTU 编号和其序列数量，如**表 2** 所示。

表 2. OTU 文件格式

#OTU_ID	Count
1082539	412
1023477	322
951711	164

1.2 16S 扩增子的 KO 功能预处理

16S 扩增子序列推荐使用 Parallel-Meta 3 进行功能预测。以 16S 扩增子测序文件 sample1.fa 为例，采用 Parallel-Meta 3（3.6 版本）的处理命令为：

PM-parallel-meta -r sample1.fa -o sample1.out -v F

输出目录 sample1.out 下，文件 functions.txt 即为符合条件的 KO 搜索输入文件。如果该 16S 扩增子序列已经经过了 1.1 中的处理并得到了表 2 的 OTU 结果，也可以通过 Parallel-Meta 3 的以下命令生成 KO 搜索输入文件：

PM-predict-func -i samples1.out/classification.txt -o sample1.func.out

输出目录 sample1.func.out 下，文件 functions.txt 即为符合条件的 KO 搜索输入文件。该文件首行为标题行，以符号#开头，其余内容主要包含两列，即 KO 编号和其丰度，如**表 3** 所示。

表 3. KO 文件格式

#KO	Count
K01992	151
K01990	146
K06147	107

1.3 WGS 的物种组成预处理

WGS 序列推荐使用 MetaPhlAn 2 进行物种解析。以 WGS 测序文件 sample1.fa 为例，采用 MetaPhlAn 2 的处理命令为：

metaphla2.py sample_1.fa --input_type fasta --tax_lev s --ignore_viruses --ignore_eukaryotes --ignore_archaea > profiled_sample_1.sp.txt

输出文件 profiled_sample_1.sp.txt 即为符合条件的物种信息。该文件首行为标题行，以符号#开头，其余内容主要包含两列，即物种名称（以物种名称标识 “s__” 为开头）和其相对丰度，如**表 4** 所示。

表 4. 物种文件格式

#Species	Abundance
s__Rothia_aeria	10
s__Actinomyces_naeslundii	12.49
s__Corynebacterium_matruchotii	11.27

1.4 WGS 的 KO 功能预处理

WGS 序列推荐使用 HUMAnN 2 进行 KO 功能解析。以测序文件 sample1.fa 为例，采用 HUMAnN 2 的命令为：

humann2 --input sample1.fa --output sample1.out

输出目录 sample1.out 中即包含符合条件的 KO 搜索输入文件，格式与**表 3** 一致。

2. 数据库搜索

2.1 访问 MSE 2

通过浏览器访问 MSE 2 在线平台 http://mse.ac.cn，点击首页或导航栏的“Search”按钮。

2.2 选择搜索类型

根据序列预处理后的输入格式，选择相应的搜索类型。具体为，16S 扩增子的 OTU（**表 2**）选择“Search by OTU”；WGS 的物种信息（**表 4**）选择“Search by species”；16S 扩增子及 WGS 的 KO 功能（**表 3**）选择“Search by function”。

2.3 输入与搜索

待搜索数据能够以两种方式上传（只需任选其一）：

a. 以文件的形式上传。如**图 1** 中在“upload your query OTUs”栏目下点击“Select”按钮选择文件；

b. 以纯文本的形式直接粘贴。如**图 1** 中点击“or Paste the input OTUs here”按钮出现文本框即可粘贴。

输入完毕后，点击最下方“Launch search”即可启动搜索。需要注意，由于网络延迟，浏览器可能需要 2-3 秒的反应时间，在此会出现“This search might take several seconds, please do not close this window.”提示，请勿关闭浏览器窗口，否则将无法显示结果，需要重新上传并搜索。

2.4 自定义搜索参数

在搜索页面（如**图 1**），点击“Other parameters”可以指定搜索参数，其中：

a. “最多匹配数”表示从数据库中返回匹配样本的上限数量，默认为 10；

b. “最低相似度”表示搜索结果将丢弃低于该相似度的匹配样本，默认为 0.6。相似度的定义见**结果与分析**。

2.5 范例数据

MSE 2 在线平台提供各种搜索类型的范例输入。例如，在**图 1** 的“Search by OTU”模式下：

a. 文件形式的范例。点击右下方“Demo query”按钮展开栏目会出现 OTU 格式的输入文件范例。继续点击“View”按钮可以查看该群落的结构，点击“Download”即可下载该范例文件作为输入，或者点击“Demo run”即可直接启动以该文件为输入的搜索。

b. 纯文本形式的范例。点击“or Paste the input OTUs here”按钮出现文本框，勾选文本框上方的“paste the demo query”可以直接将范例样本的内容粘贴到文本框中，该范例样本内容与文件形式的范例内容一致。点击旁边的“view”链接同样可以查看该群落的结构。

其他搜索模式下也同样提供以上范例。

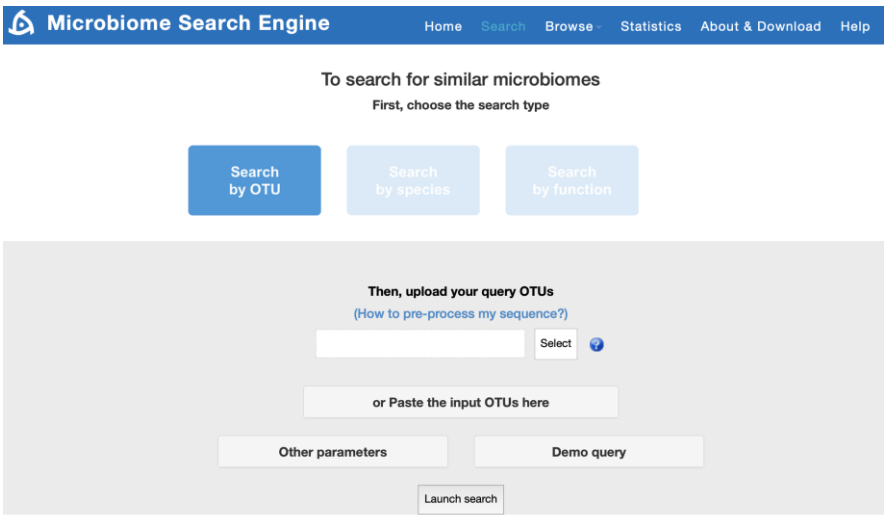

图 1. MSE 2 的“Search by OTU”搜索页面

3. 数据库浏览与下载

MSE 2 网站提供了两种样本浏览方式：

按项目/研究浏览。在项目列表页面，样本按照项目进行排列，所有的项目会按照项目 ID 进行排序。单击项目 ID 可以进入项目页面，该页面包含每个项目的统一化元数据（例如，研究标题、样本数量、发表情况等），该项目的完整原始元数据，以及访问该数据原始发布页的链接。

按样本浏览。在样本列表页面中，所有的样本会以列表的形式展示并按照样本号进行排序。用户可以对样本列表进行筛选，目前支持的筛选条件有元数据过滤器，环境，序列类型，采样年份等等。单击样本 ID 可以进入样本页面，可以查看其详细的统一化元数据（例如，来源研究、采样地点、序列类型等）和由 Krona 绘制的动态物种组成图。

MSE 2 数据库会保持动态更新。在 MSE 2 网站的“About & Download”页面中，所有样本统一的元数据可在“Database Information”栏目下载。与此同时，单机版的 MSE 2 搜索引擎内核软件也可以在“Download”栏目下载，从而实现本地化的数据库构建与搜索。该软件能够以独立软件的形式安装使用，也能够以 QIIME2 插件的形式使用。

结果与分析

以 OTU 搜索为例，将预处理中得到的 classificatioin.txt 文件（格式见**表 2**）作为输入。范例该文件可以从实验步骤 2.5a 中下载。得到输出结果如**图 2** 所示：

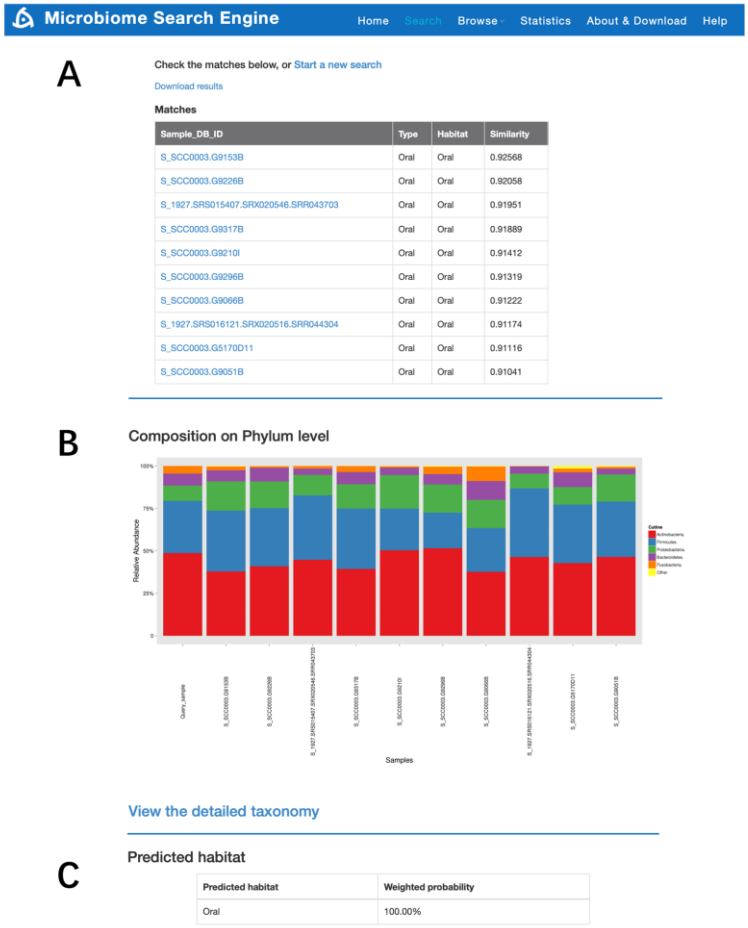

图 2. MSE 2 的“Search by OTU”搜索结果。A 搜索匹配结果列表。B 搜索样本与匹配结果结构对比图。C 搜索样本来源环境预测结果

1. 搜索匹配结果列表

如**图 2A** 所示，其中：第一列为匹配样本的数据库样本 ID，点击后可查看详细的元数据；第二列为匹配样本的来源类型；第三列为匹配样本的来源环境；第四列为搜索输入样本与匹配样本的相似度，OTU 搜索采用 Meta-Storms 相似度(Su 等， 2012)，物种搜索采用 Dynamic Meta-Storms 相似度(Jing 等， 2020)，KO 功能搜索采用 Bray-Curtis 相似度。

2. 搜索样本与匹配结果结构对比图

如**图 2B** 所示，该柱状图展示了门（Phylum）层次上输入样本与匹配样本的相对丰度的差异。点击下方“View the detailed taxonomy”可以查看更多层次上的对比。

3. 来源环境预测结果

图 2C 为根据匹配结果的来源环境及相似度（即**图 2A** 中内容）对搜索样本的环境预测结果以及概率。

4. 搜索结果下载

点击**图 2A** 中“Download results”链接可下载以上搜索结果的压缩包，其中：

a. query.out，文本文件，包含**图 2A** 中匹配样本的 ID 及相似度；

b. Query_sample.png，png 格式的图，为**图 2B** 中柱状图；

c. Query_sample.phylum.Abd，文本文件，为**图 2B** 的门层次的丰度信息；

d. Query_sample.OTU.Abd，文本文件，为**图 2B** 所对应的 OTU 层次的丰度信息；

e. multi-view，文件夹，其中的“taxonomy.html”网页文件为搜索样本和其匹配结果在所有 taxonomy 分类层次的展示，其他文件为显示辅助文件。

其他类型的搜索结果与 OTU 搜索基本一致。需要注意的是，KO 功能搜索（Search by function）的搜索结果中，**图 2B** 展示的为 KO BRITE Level 2 层次上的代谢通路的差异，其结果的下载包中也不包含 multi-view 文件夹。

5. 搜索匹配结果的详细信息

搜索匹配结果**图 2A** 中，每个样品的 ID 均链接到其样本页面，可以查看其详细的元数据（例如，来源研究、采样地点、序列类型等）。此外，在该样本页面点击项目 ID 来进入项目页面，也可以通过“Download raw metadata”下载该项目的完整原始元数据。

失败经验

常见问题：No-Hit。

问题原因：输入数据格式错误，或者最低相似度阈值太高。

解决方法：

a. 根据待搜索样本类型和搜索类型，按照**表 1** 检查预处理方法和输入格式；

b. 根据**实验步骤**中 2.4b，降低“最低相似度”。

致谢

感谢中国科学院青岛生物能源与过程研究所乔英合工程师对服务器的管理和硬件维护。该工作得到了国家自然科学基金 **31771463**、**32070086** 的资助。

参考文献

1. Franzosa, E. A., Mciver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G. and Segata, N. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* 15:962-968.
2. Jing, G., Liu, L., Wang, Z., Zhang, Y., Qian, L., Gao, C., Zhang, M., Li, M., Zhang, Z., Liu, X., Xu, J. and Su, X. (2021). Microbiome Search Engine 2: a Platform for Taxonomic and Functional Search of Global Microbiomes on the Whole-Microbiome Level. *mSystems* 6(1): e00943-00920.
3. Jing, G., Sun, Z., Wang, H., Gong, Y., Huang, S., Ning, K., Xu, J. and Su, X. (2017). Parallel-META 3: Comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. *Scientific Reports* 7: 40371
4. Jing, G., Zhang, Y., Ming, Y., Liu, L., Xu, J. and Su, X. (2020). Dynamic Meta-Storms enables comprehensive taxonomic and phylogenetic comparison of shotgun metagenomes at the species level. *Bioinformatics* 36:2308–2310
5. McDonald, D., Price, M., N., Goodrich, J., Nawrocki, E. and P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *Isme Journal* 6:610–618
6. Su, X., Xu, J. and Ning, K. (2012). Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. *Bioinformatics* 28:2493-2501.
7. Tin, D., Truong, Eric, A., Franzosa, Timothy, L., Tickle, Matthias, Scholz, George and Weingart (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods* 12:902-903.