**从细菌基因组中预测活性前噬菌体工具Prophage Hunter的使用流程和常见问题**

**The Usage Process and Common Problems of Prophage Hunter, a Tool to Capture Active Phage from the Bacterial Genome**

孙海汐1, 2, #, \*，李敏1, 2, 3, #，宋文琛1, 2, #，肖敏凤1, 2, 3, \*

1.深圳华大生命科学研究院,深圳；2.深圳市未知病原体应急检测重点实验室，深圳；3.中国科学院大学华大教育中心，深圳

\*通讯作者邮箱: [xiaominfeng@genomics.cn; sunhaixi@cngb.org](mailto:xiaominfeng@genomics.cn；sunhaixi@cngb.org)

#共同第一作者/同等贡献

**摘要：**识别具有活性的前噬菌体，对于研究噬菌体和细菌的协同进化、噬菌体的生理生化以及工程化设计噬菌体等多种用途至关重要。这里，我们介绍了Prophage Hunter，该工具旨在从细菌的全基因组序列中寻找具有活性的前噬菌体。结合序列相似性的匹配及基于遗传特征的机器学习分类模型，我们开发了一种新颖的评分系统，该系统在验证数据集上预测活性前噬菌体中表现出比当前工具更高的准确性。该工具也可以选择跳过序列相似性匹配，这样有更多可能发现新颖的前噬菌体。Prophage Hunter提供一站式网站服务，包括从细菌基因组中提取前噬菌体基因组、评估前噬菌体的活性、鉴定系统发育相关的噬菌体、注释噬菌体蛋白的功能及可视化前噬菌体基因组位置信息等。Prophage Hunter可在<https://pro-hunter.genomics.cn/>免费使用。

**关键词：**前噬菌体，细菌，机器学习，注释，一站式分析

**仪器设备**

1. 个人电脑：安装主流浏览器 (Chrome/Safari) 即可

**实验步骤**

1. 准备输入数据：细菌基因组序列,可包含一条或多条序列 (FASTA格式，图1)。

****

**图1. FASTA格式序列**

1. 在浏览器中输入<https://pro-hunter.bgi.com/>，进入网站主页 (图2)。



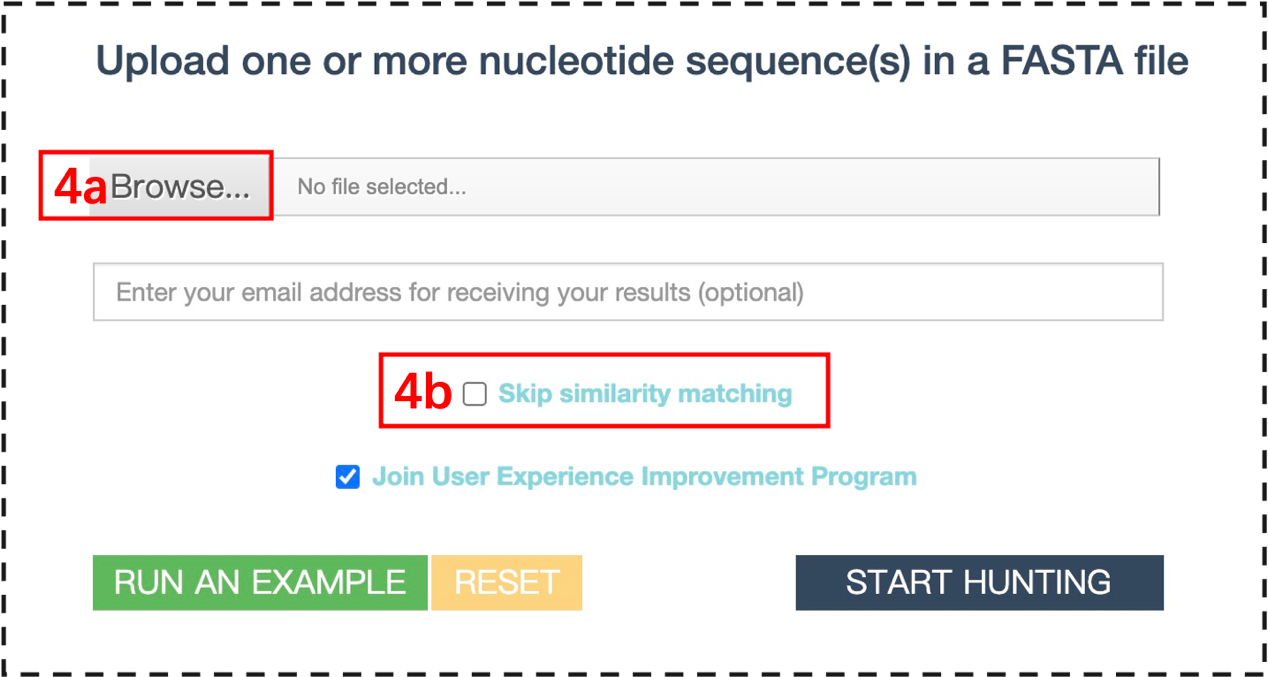
**图2. Prophage Hunter网站主页**

1. 在网页的导航栏上单击"Start Hunting (开始狩猎)"按钮启动Prophage hunter程序 (图3)。

****

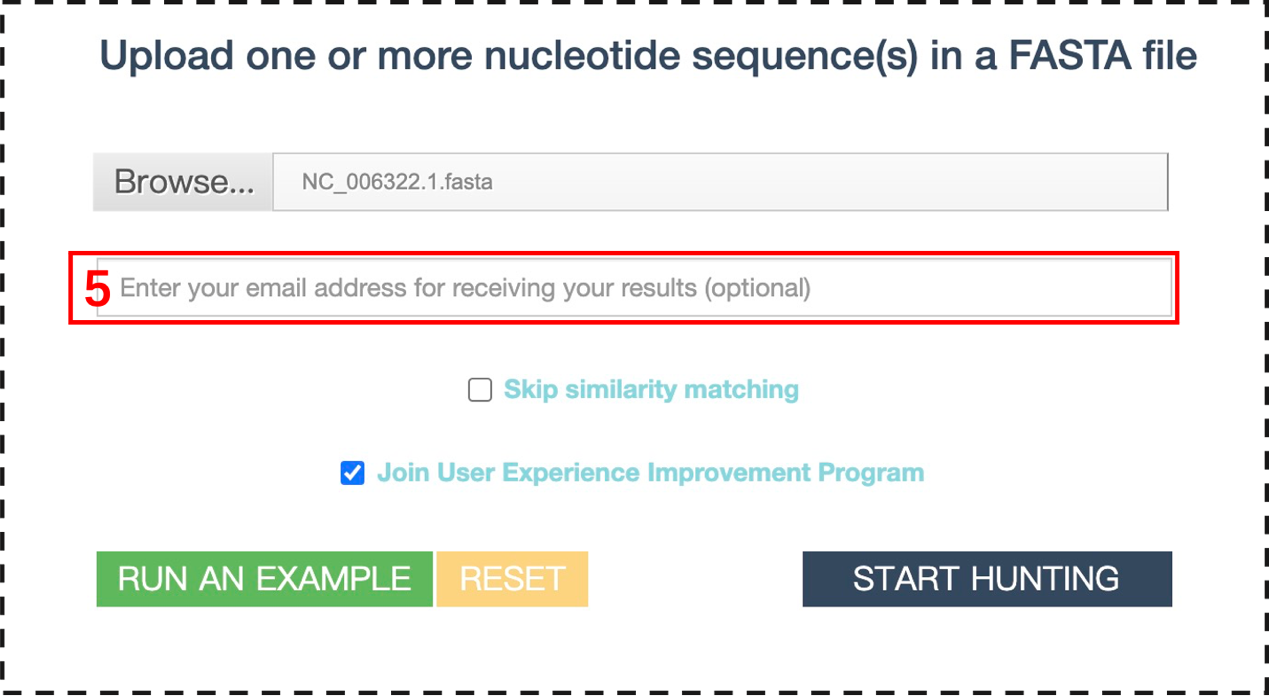
**图3. 单击红色框中的Start Hunting (开始狩猎) 按钮以启动程序**

1. 在主页上，单击"Browse (浏览)"将一个或多个核苷酸序列以FASTA文件格式上传到网站 (图4a)。默认情况下，Prophage Hunter使用相似性搜索策略来标识初始前噬菌体区域。也可以通过勾选"Skip similarity matching (跳过相似性匹配)"框 (图4b)，用户可以跳过此过程以识别新型噬菌体。本示例以地衣芽孢杆菌*Bacillus licheniformis* DSM 13 (Accession Num.:NC\_006322.1) 为输入。

****

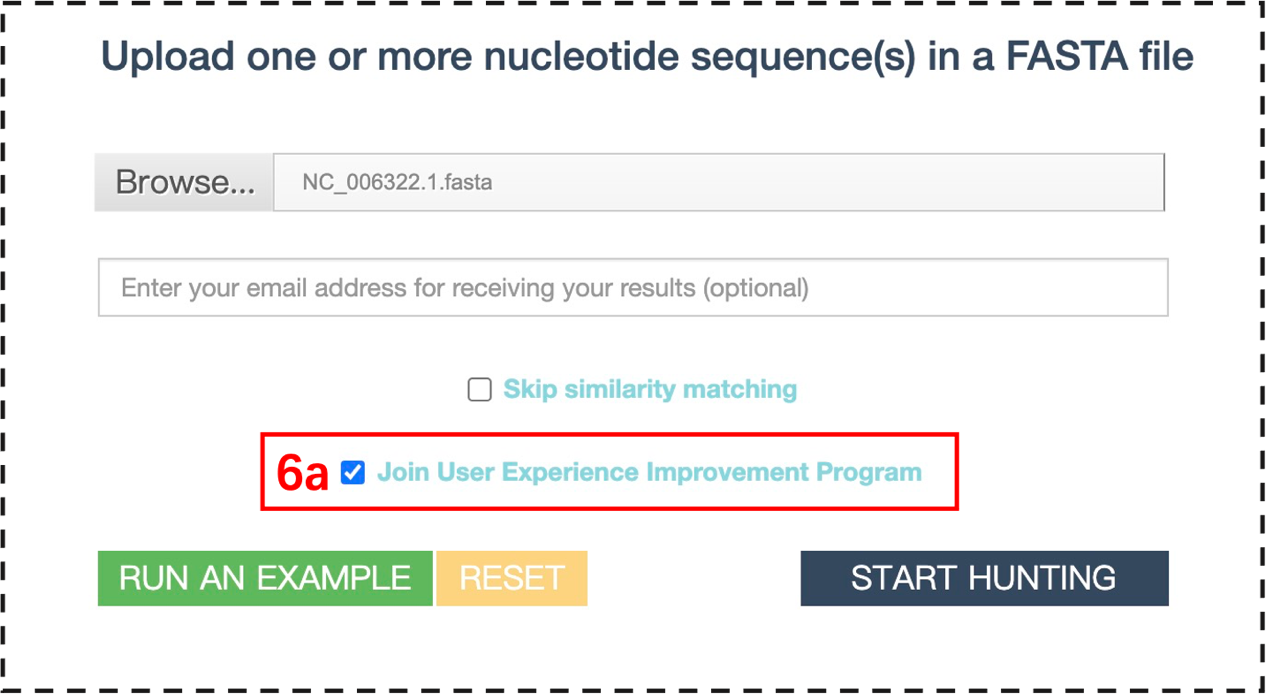
**图4. 单击"Browse (浏览)"上传FASTA文件**

1. 可选择输入电子邮箱地址以接收指向分析报告的超链接 (图5)。 请注意，此超链接将在一周后过期。若不输入，则需在提交序列后跳转的分析页面等待分析完成，或自行将提交序列后跳转的分析页面链接复制保存，以便查看结果（详见步骤7、8）。

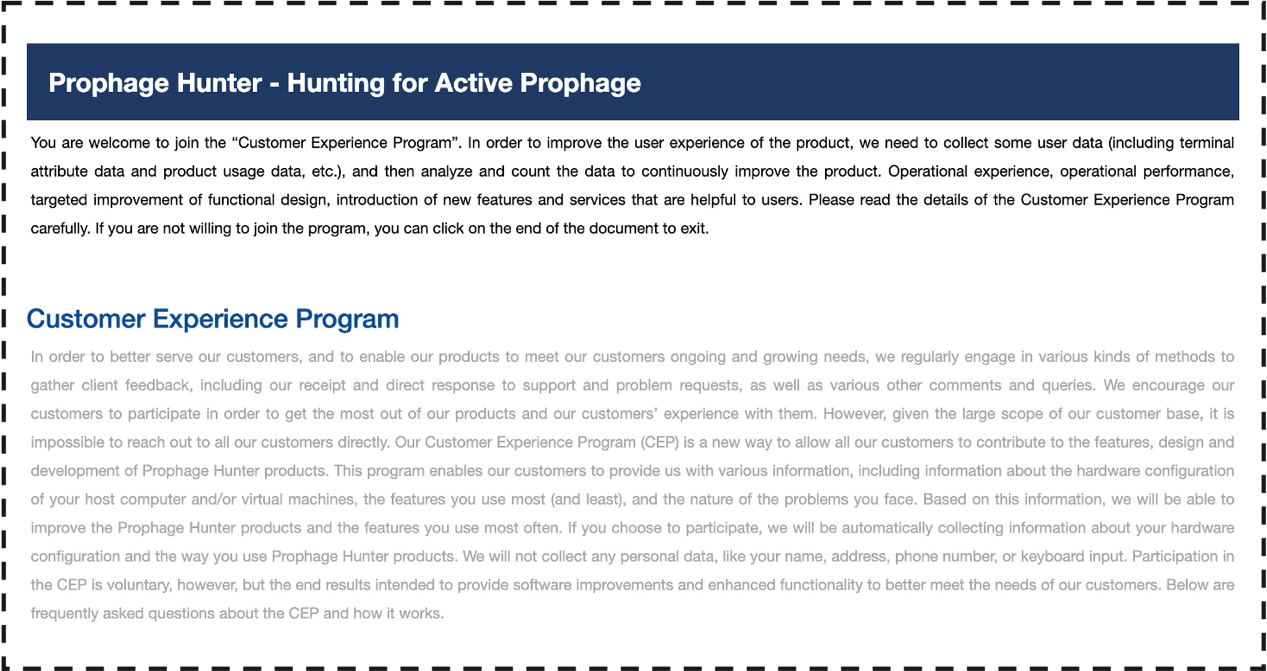
****

**图5. 输入电子邮件地址以接收分析报告**

1. 勾选以加入"Join User Experience Improvement Program (用户体验改善计划)" (图6a)。有关更多详细信息，请单击"Join User Experience Improvement Program (用户体验改善计划)" (图6b)。

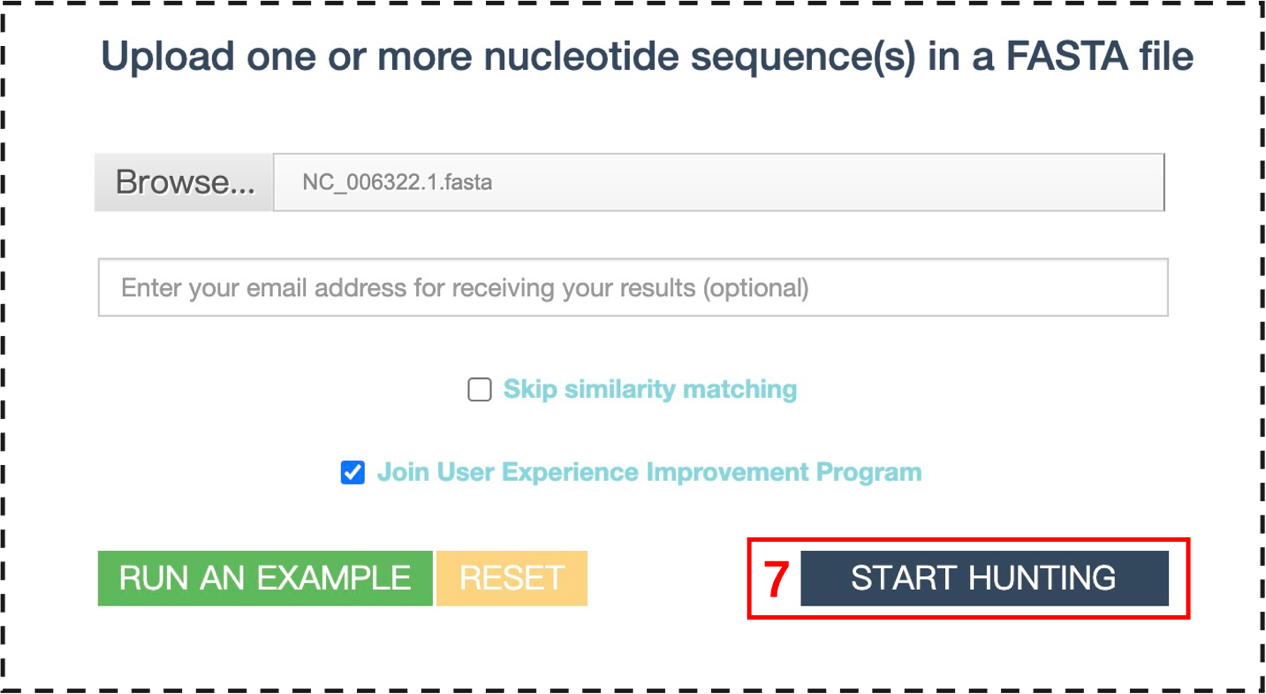
****

**图6a. 勾选加入用户体验改善计划**

****

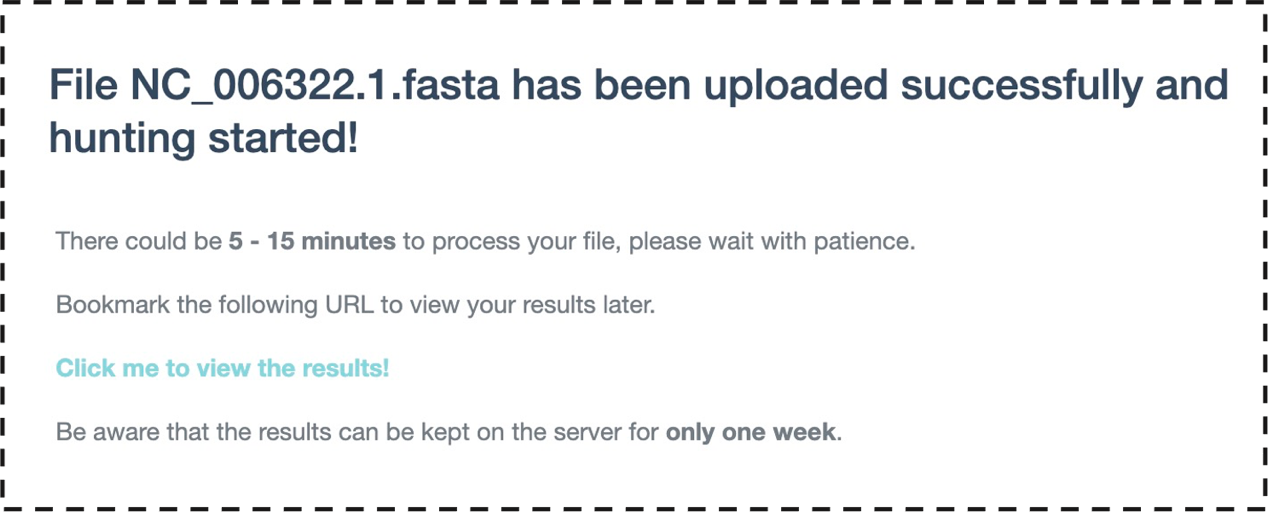
**图6b. 用户体验改善计划具体信息**

1. 点击"START HUNTING (开始狩猎)"按钮开始分析。

****

**图7. 开始分析**

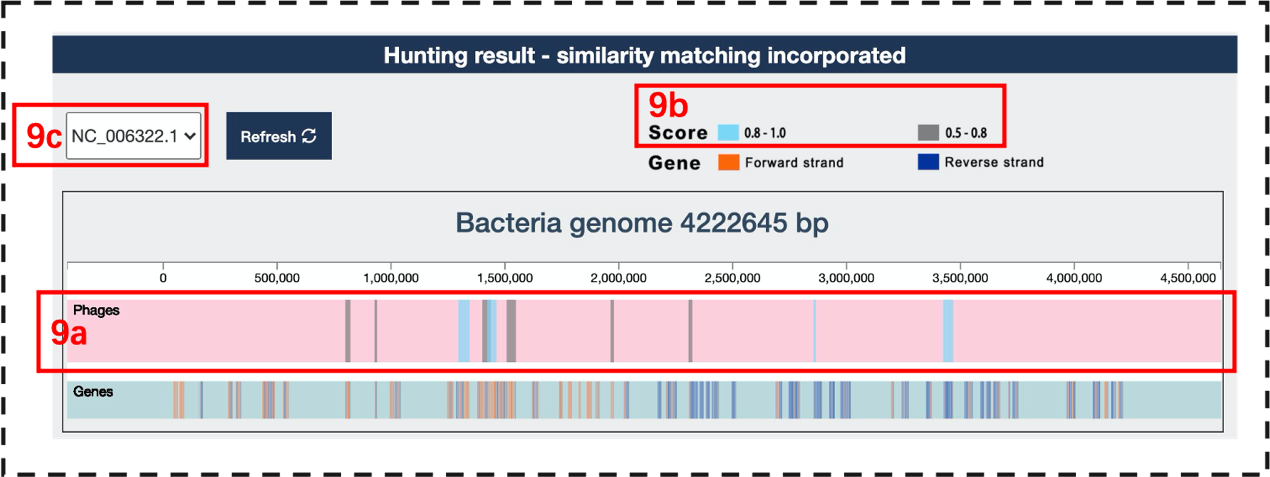
1. 分析完成需要等待5-15分钟，分析完成后点击蓝色区域链接即可到达结果页面 (图8)。

****

**图8. 等待分析完成**

**结果**

1. 在结果页面的顶部是一个基因组浏览器，显示了每个预测的前噬菌体区域 (图9a)。 活性前噬菌体区域（Category 为Active）和模糊区域（Category 为Ambiguous，即难以判断活性的前噬菌体区域）分别以天蓝色和灰色着色 (图9b)。若提交文件中含有多个序列，用户可以通过单击左上方的下拉菜单切换到基因组的其他染色体或Scaffold序列 (图9c)。

****

**图9. 基因组浏览器显示预测的前噬菌体区域的基因座和活性类别**

1. 在结果页面底部的表格显示了每个预测的前噬菌体基因组区域的详细信息 (图10a)：

- Candidate ID (候选ID)：预测的前噬菌体基因组区域的ID；

- Sequence ID (序列ID)：输入文件中显示的细菌基因组序列ID；

- Start (开始)：预测的前噬菌体基因组区域的开始位置；

- End (结束)：预测的前噬菌体基因组区域的结束位置；

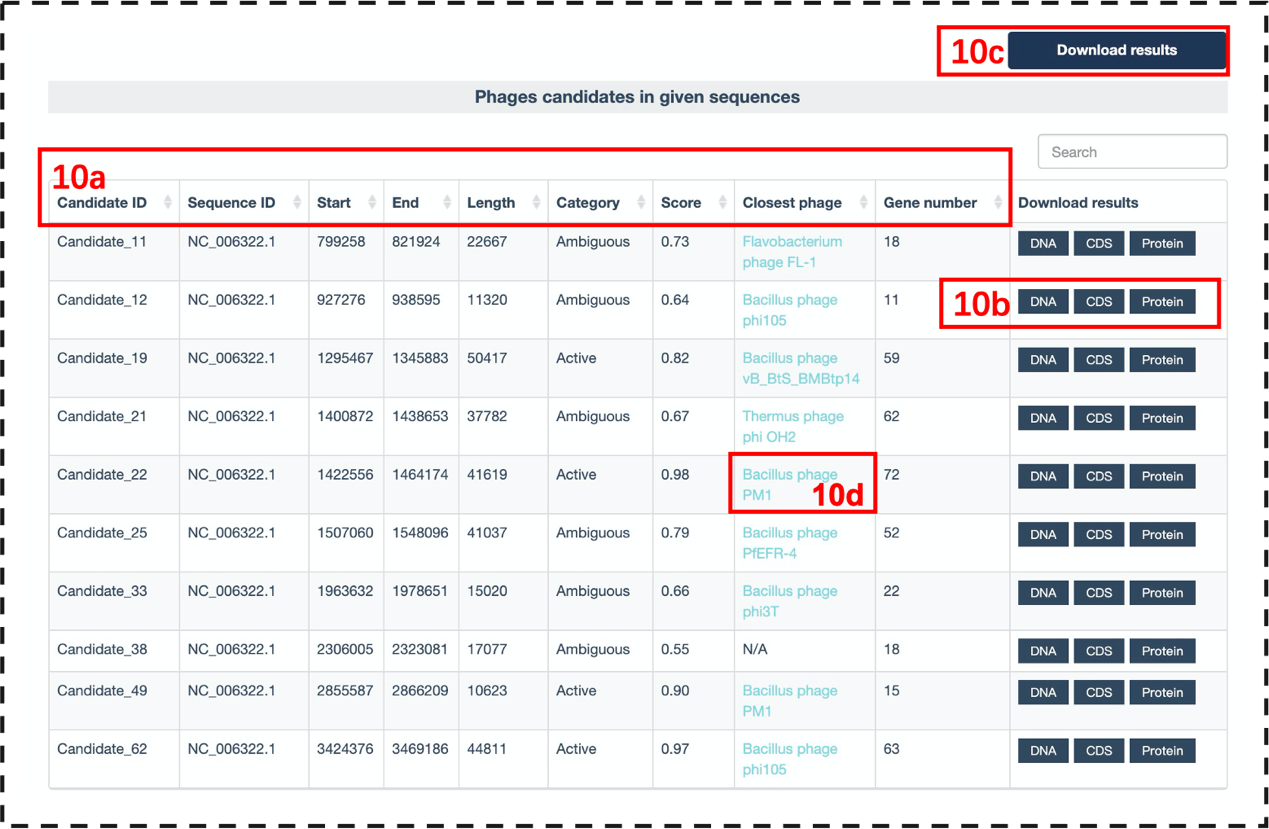
- Length (长度)：预测的前噬菌体基因组区域的长度；

- Category (类别)：预测的前噬菌体基因组活性类别，Active表示预测为具有活性(Score > 0.8)，Ambiguous表示该预测区域活性难以判断 (0.5 < Score < 0.8)；

- Score (分数)：预测的前噬菌体基因组活性分数，分数越高，支持预测区域的活性证据越强；

- Closest phage (最近源的噬菌体)：与现有噬菌体库相比，预测的前噬菌体区域最近源的噬菌体，" N/A"表示未在噬菌体库中找到近源噬菌体；

- Gene number (基因数目)：预测的前噬菌体区域内基因的数目。

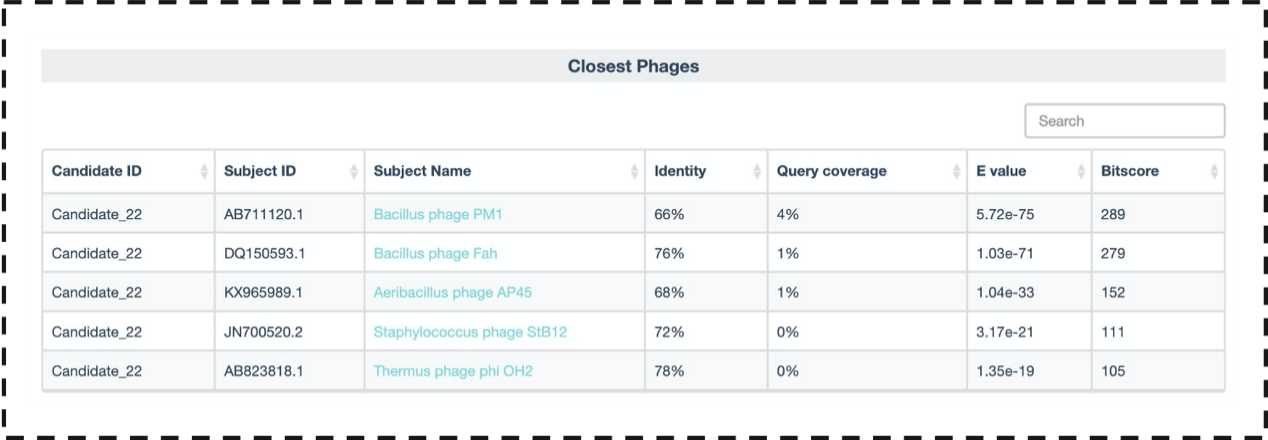
****

**图10. 每个预测的前噬菌体区域详细信息。**

用户可以单击相应的按钮以下载对应预测的前噬菌体区域的基因组DNA序列、CDS序列或蛋白质序列 (图10b) 或所有分析结果 (图10c)。 用户还可以单击最接近的噬菌体的分类名称以查看详细信息 (图10d)。

1. 在结果页面顶部的基因组浏览器或下方的表格中单击预测的前噬菌体区域，可以查看预测的前噬菌体区域每个基因的注释情况 (图11a) 和与该区域同源的前5个最接近的噬菌体 (图11b)。本示例为点击表格中 Candidate 22 。

** 图11a. 预测的前噬菌体中基因在数据库中 (NCBI NR、Pfam和InterPro) 的注释情况**

****

**图11b. 与预测区域同源的前5个最接近的噬菌体**

**致谢**

本项目由国家重点研发计划项目 (2020YFA0908700)、深圳市孔雀团队项目(KQTD2015033117210153)支持。

**参考文献**

1. Song, W., Sun, H. X., Zhang, C., Cheng, L., Peng, Y., Deng, Z., Wang, D., Wang, Y., Hu, M., Liu, W., Yang, H., Shen, Y., Li, J., You, L. and Xiao, M. (2019). [Prophage Hunter: an integrative hunting tool for active prophages.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6602508/) *Nucleic Acids Res* 47(W1):W74-W80.