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

#共同第一作者/同等贡献

引用格式：孙海汐，李敏，宋文琛，肖敏凤．(2021)．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．Bio-101 e2003952．Doi：10.21769/BioProtoc.2003952．

How to cite: Sun, H. X., Li, M., Song, W. C. and Xiao, M. F. (2021). The Usage Process and Common Problems of Prophage Hunter, a Tool to Capture Active Phage from the Bacterial Genome. Bio-101 e2003952. Doi: 10.21769/BioProtoc. 2003952. (in Chinese)

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

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

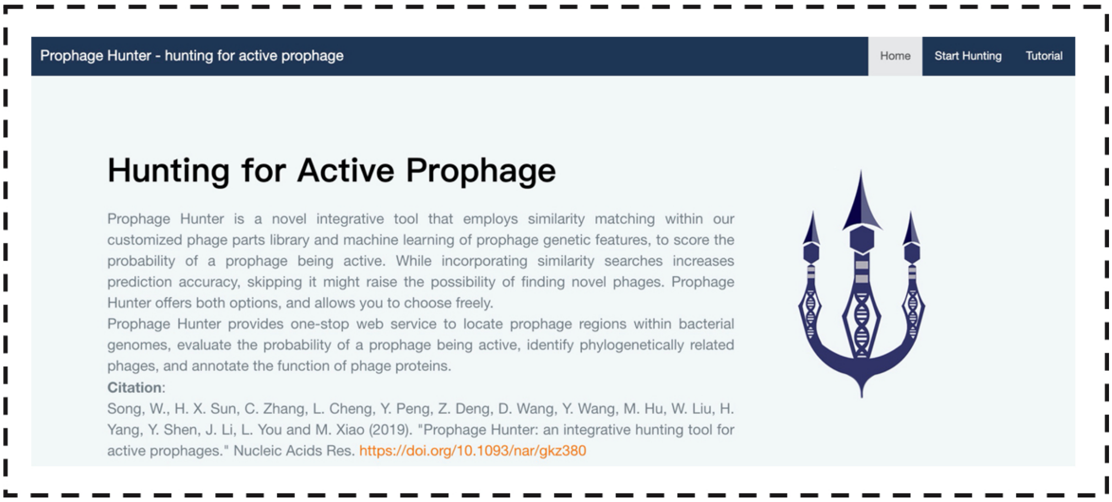
实验步骤

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)。

```
>NC_006322.1 Bacillus licheniformis DSM 13 = ATCC 14580, complete sequence
TGGATAAGTTCCTCGCAACCATTGCAACCACTCGCTTATTCTGATATTATATTTGTGTTTTAACTCTTGA
TAACAAATTGGCTGCCAATCCATTATCCACAAACTGTGGATAAGTTGTGGAGAGTTTTTTCACAGGGTGT
GCAGTATTTTGTCCACATCTTGTGAAAAATGTCGAAAAGACGTTTTTCTACTATATTATATGTTTTCAAC
ATTTCATTAACGAATGGACTCATCCATTTGCTCTTTTTTTGTGTTCTATAACAGGTTGGACAAGCAAATA
TTGCTGGTAAAGGAGGGACGAACCGCCATTATGAAAAACATATCGGATCTTTGGAATCAAGCTTTGGGGC
AGATCGAAAAAAAATTGAGCAAGCCCAGCTTTGAAACATGGATGAAATCGACAAAGGCCCATTCATTGCA
GGGCGATACGCTGATCATCACCGCACCGAACGAGTTCGCCAGAGACTGGCTTGAATCAAGATACCTGCAC
CTGATCGCCGATACGATCTATGATCTGACAGGAGAAGAATTGAGCATTAAATTTGTCATTCCTCAGAATC
AAAATGAAGAAGATTTTATGCCAAAGTCTCCAATCAAAAAAATGTCGAAAGAAGAACCGGCTGATTTTCC
GCAAAACATGCTGAATCCCAAATATACATTTGATACGTTCGTTATCGGTTCAGGAAACCGATTCGCCCAC
GCAGCGTCTTTGGCAGTGGCTGAAGCCCCGGCGAAAGCTTACAATCCGCTGTTTATTTACGGGGGAGTCG
```

图 1. FASTA 格式序列

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



Hunting for Active Prophage

Prophage Hunter is a novel integrative tool that employs similarity matching within our customized phage parts library and machine learning of prophage genetic features, to score the probability of a prophage being active. While incorporating similarity searches increases prediction accuracy, skipping it might raise the possibility of finding novel phages. Prophage Hunter offers both options, and allows you to choose freely.

Prophage Hunter provides one-stop web service to locate prophage regions within bacterial genomes, evaluate the probability of a prophage being active, identify phylogenetically related phages, and annotate the function of phage proteins.

Citation:

Song, W., H. X. Sun, C. Zhang, L. Cheng, Y. Peng, Z. Deng, D. Wang, Y. Wang, M. Hu, W. Liu, H. Yang, Y. Shen, J. Li, L. You and M. Xiao (2019). "Prophage Hunter: an integrative hunting tool for active prophages." Nucleic Acids Res. https://doi.org/10.1093/nar/gkz380

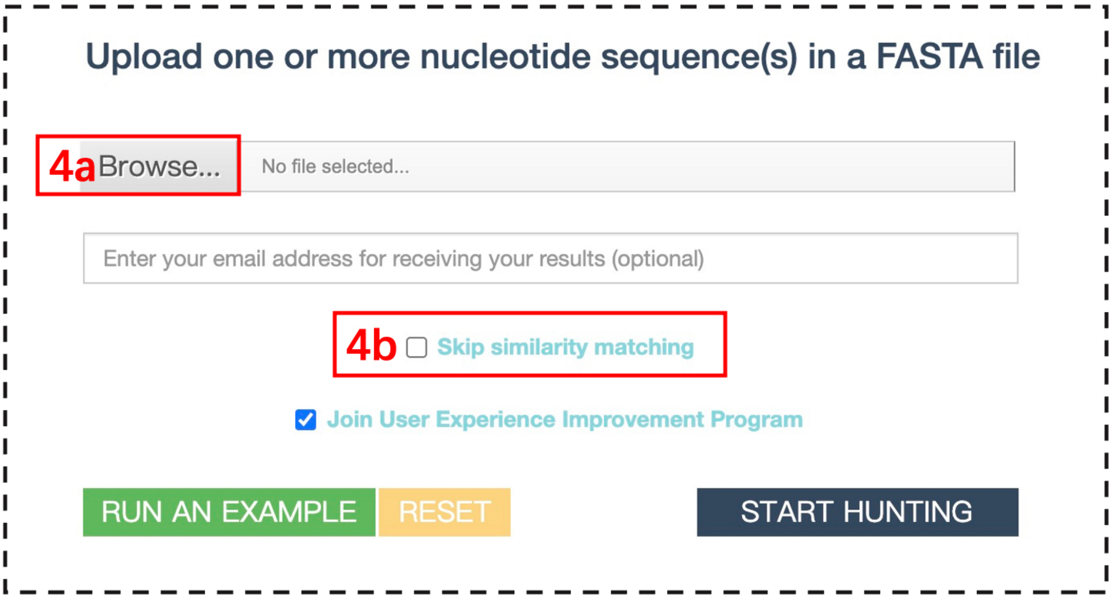
图 2. Prophage Hunter 网站主页

3. 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

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

4. 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。



Upload one or more nucleotide sequence(s) in a FASTA file

No file selected...

Enter your email address for receiving your results (optional)

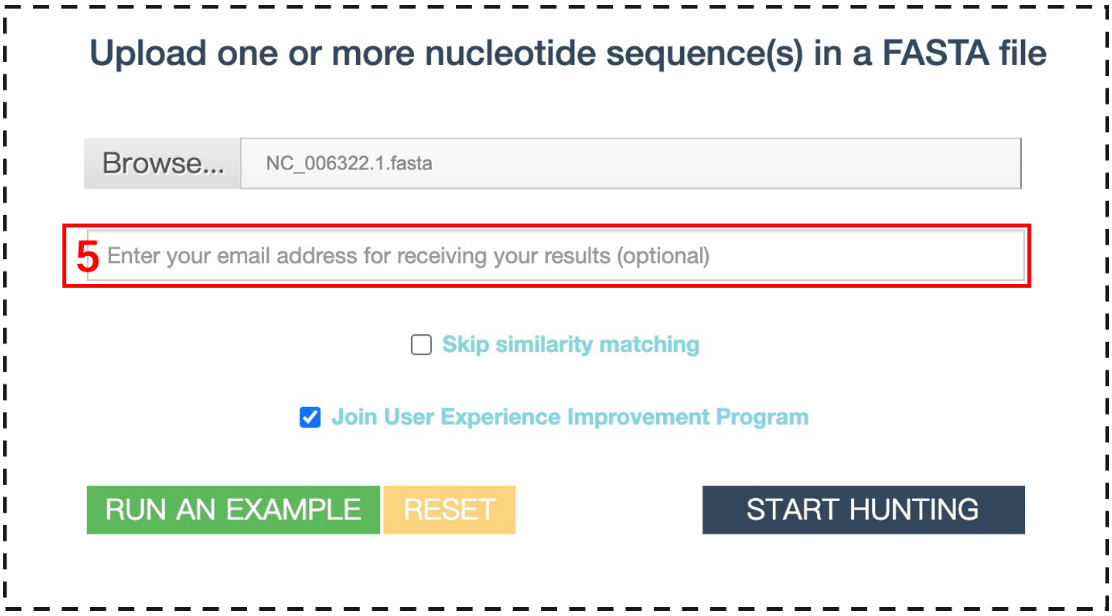
☐ Skip similarity matching

☑ Join User Experience Improvement Program

图 4. 单击"Browse (浏览)"上传 FASTA 文件

5. 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)。

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.



NC_006322.1.fasta

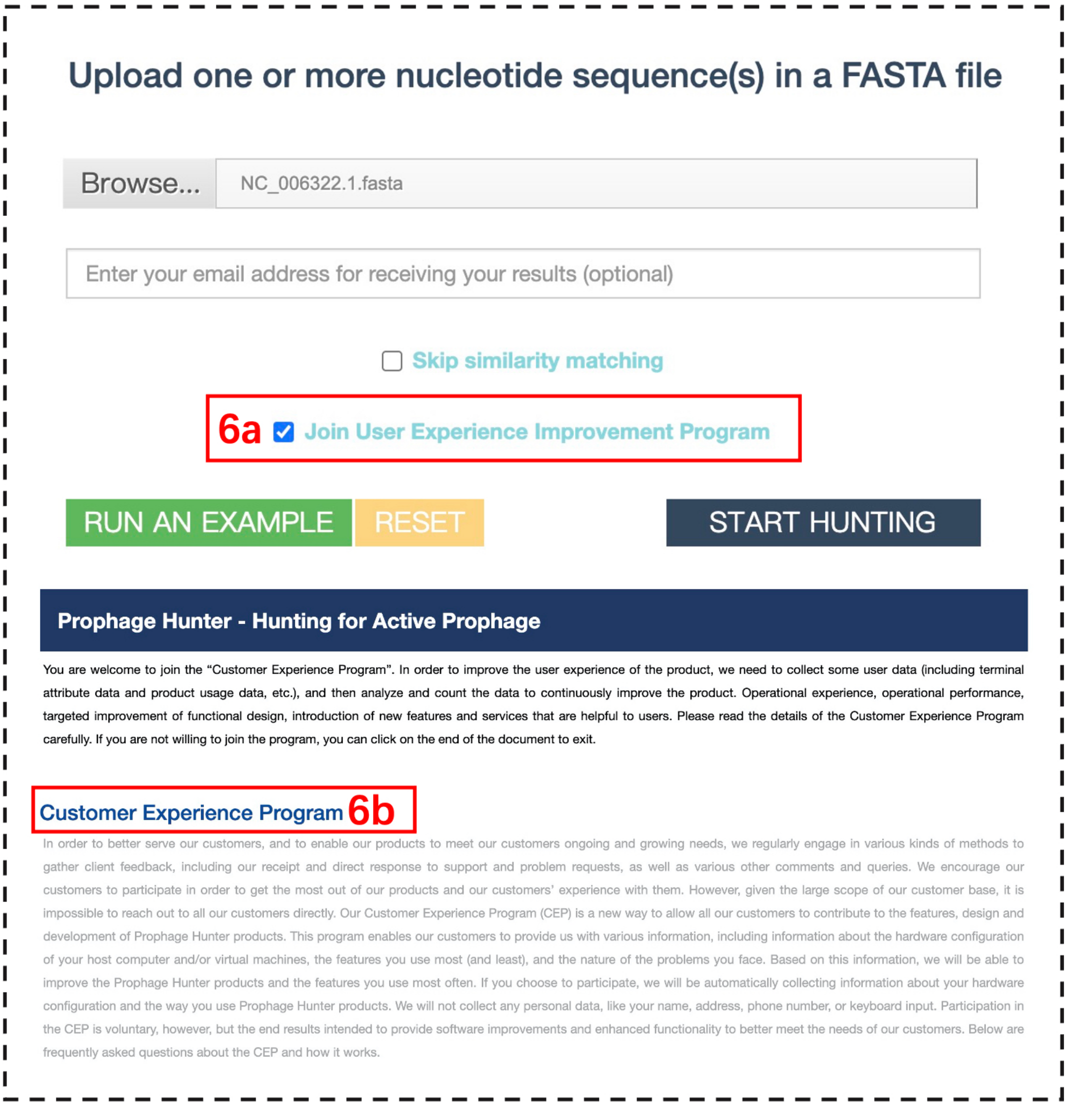
☐ Skip similarity matching

☑ Join User Experience Improvement Program

图 5. 输入电子邮件地址以接收分析报告

6. 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)。

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.



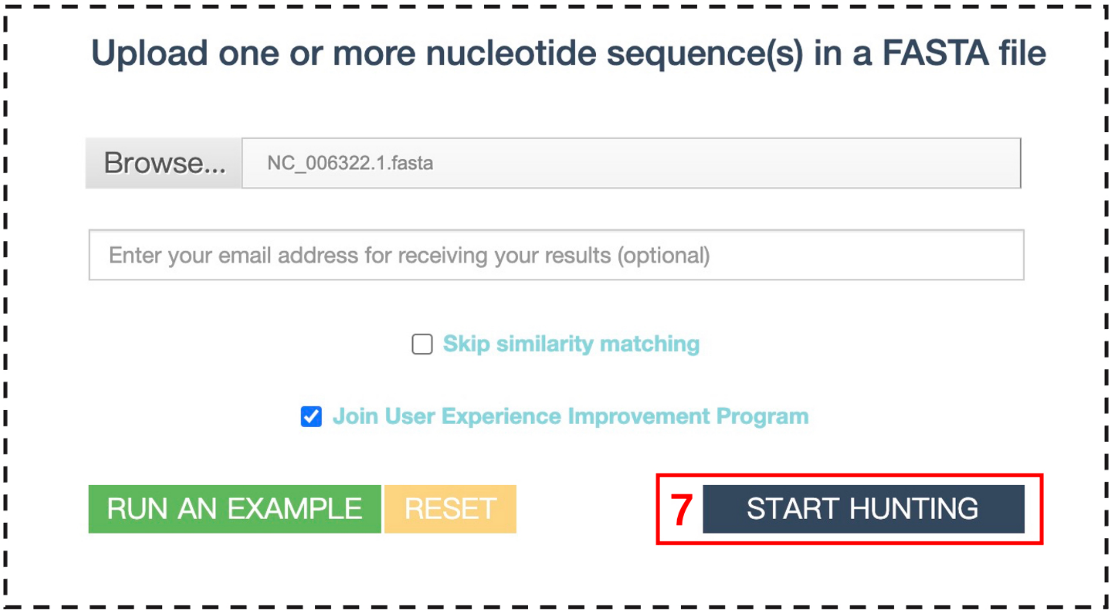
You are welcome to join the "Customer Experience Program". In order to improve the user experience of the product, we need to collect some user data (including terminal attribute data and product usage data, etc.), and then analyze and count the data to continuously improve the product. Operational experience, operational performance, targeted improvement of functional design, introduction of new features and services that are helpful to users. Please read the details of the Customer Experience Program carefully. If you are not willing to join the program, you can click on the end of the document to exit.

In order to better serve our customers, and to enable our products to meet our customers ongoing and growing needs, we regularly engage in various kinds of methods to gather client feedback, including our receipt and direct response to support and problem requests, as well as various other comments and queries. We encourage our customers to participate in order to get the most out of our products and our customers' experience with them. However, given the large scope of our customer base, it is impossible to reach out to all our customers directly. Our Customer Experience Program (CEP) is a new way to allow all our customers to contribute to the features, design and development of Prophage Hunter products. This program enables our customers to provide us with various information, including information about the hardware configuration of your host computer and/or virtual machines, the features you use most (and least), and the nature of the problems you face. Based on this information, we will be able to improve the Prophage Hunter products and the features you use most often. If you choose to participate, we will be automatically collecting information about your hardware configuration and the way you use Prophage Hunter products. We will not collect any personal data, like your name, address, phone number, or keyboard input. Participation in the CEP is voluntary, however, but the end results intended to provide software improvements and enhanced functionality to better meet the needs of our customers. Below are frequently asked questions about the CEP and how it works.

图 **6 a.** 勾选加入用户体验改善计划 **b.** 用户体验改善计划具体信息

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

7. 点击"START HUNTING (开始狩猎)"按钮开始分析。



Upload one or more nucleotide sequence(s) in a FASTA file

Browse... NC_006322.1.fasta

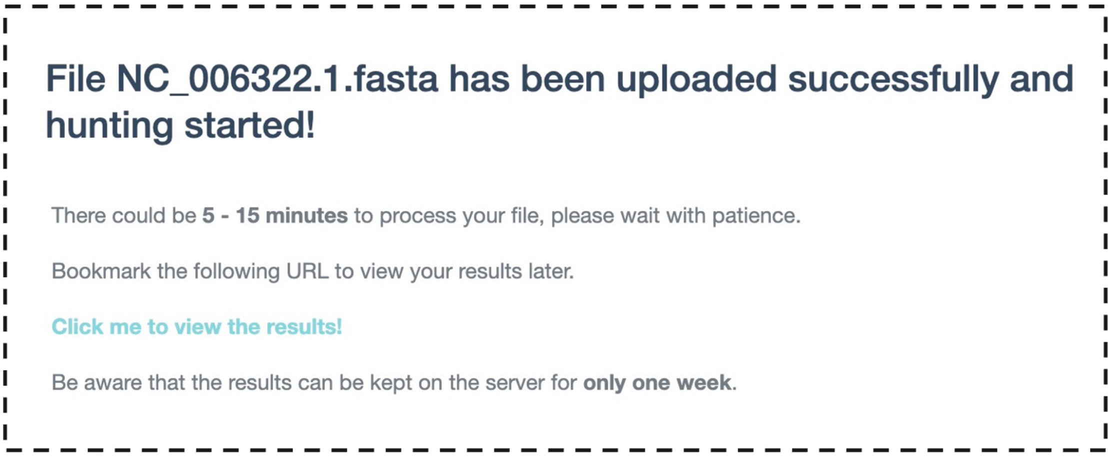
Enter your email address for receiving your results (optional)

☐ Skip similarity matching

☑ Join User Experience Improvement Program

图 7. 开始分析

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



File NC_006322.1.fasta has been uploaded successfully and hunting started!

There could be **5 - 15 minutes** to process your file, please wait with patience.

Bookmark the following URL to view your results later.

Click me to view the results!

Be aware that the results can be kept on the server for **only one week**.

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

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

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）。

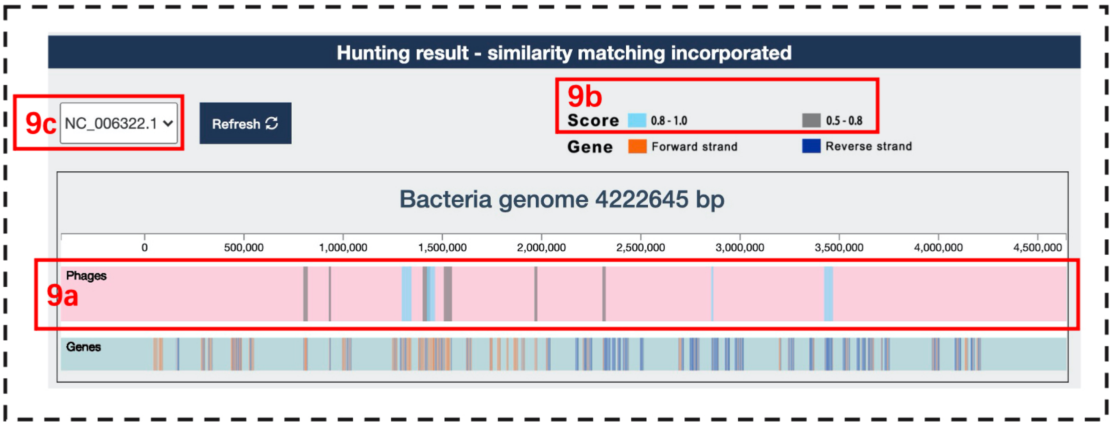


图 9. 基因组浏览器显示预测的前噬菌体区域的基因座和活性类别

2. 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。

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

Phages candidates in given sequences

Candidate ID	Sequence ID	Start	End	Length	Category	Score	Closest phage	Gene number	Download results
Candidate_11	NC_006322.1	799258	821924	22667	Ambiguous	0.73	Flavobacterium phage FL-1	18	DNA CDS Protein
Candidate_12	NC_006322.1	927276	938595	11320	Ambiguous	0.64	Bacillus phage phi105	11	DNA CDS Protein
Candidate_19	NC_006322.1	1295467	1345883	50417	Active	0.82	Bacillus phage vB_BtS_BMBtp14	59	DNA CDS Protein
Candidate_21	NC_006322.1	1400872	1438653	37782	Ambiguous	0.67	Thermus phage phi OH2	62	DNA CDS Protein
Candidate_22	NC_006322.1	1422556	1464174	41619	Active	0.98	Bacillus phage PM1	72	DNA CDS Protein
Candidate_25	NC_006322.1	1507060	1548096	41037	Ambiguous	0.79	Bacillus phage PfEFR-4	52	DNA CDS Protein
Candidate_33	NC_006322.1	1963632	1978651	15020	Ambiguous	0.66	Bacillus phage phi3T	22	DNA CDS Protein
Candidate_38	NC_006322.1	2306005	2323081	17077	Ambiguous	0.55	N/A	18	DNA CDS Protein
Candidate_49	NC_006322.1	2855587	2866209	10623	Active	0.90	Bacillus phage PM1	15	DNA CDS Protein
Candidate_62	NC_006322.1	3424376	3469186	44811	Active	0.97	Bacillus phage phi105	63	DNA CDS Protein

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）。

3. 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 。

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

Candidate 22 Annotation

Gene ID	Protein length (aa)	NCBI_nr	Pfam	InterPro
gene_1437	379	integrase	Arm DNA-binding domain | Phage integrase, N-terminal SAM-like domain | Phage integrase family	AP2-like integrase, N-terminal domain | Integrase, SAM-like, N-terminal | Integrase, catalytic domain
gene_1438	70	hypothetical protein AV945_gp40	IrrE N-terminal-like domain	IrrE N-terminal-like domain
gene_1439	115	hypothetical protein Waukesha92_40	Protein of unknown function (DUF4064)	Domain of unknown function DUF4064
gene_1440	103	hypothetical protein LP101_031	#N/A	#N/A
gene_1441	35	#N/A	#N/A	#N/A
gene_1442	82	XRE family transcriptional regulator (endogenous virus)	Helix-turn-helix	Cro/C1-type helix-turn-helix domain
gene_1443	122	Cro/CI family XRE family transcriptional regulator	Helix-turn-helix	Cro/C1-type helix-turn-helix domain
gene_1444	68	Cro-like repressor	Helix-turn-helix	Cro/C1-type helix-turn-helix domain
gene_1445	58	hypothetical protein	#N/A	#N/A
gene_1446	79	hypothetical protein	#N/A	#N/A
gene_1447	143	RecT	#N/A	#N/A
gene_1448	244	Rha type regulatory protein	Phage regulatory protein Rha (Phage_pRha)	Bacteriophage regulatory protein, Rha family
gene_1449	172	gp43	#N/A	#N/A
gene_1450	92	unnamed protein product	Uncharacterized YqaH-like	Uncharacterized protein YqaH-like
gene_1451	41	hypothetical protein mEp213_026	#N/A	#N/A
gene_1452	62	hypothetical protein phiNJ2_0027	#N/A	#N/A
gene_1453	34	hypothetical protein O4_27	#N/A	#N/A
gene_1454	312	phage-type endonuclease	YqaJ-like viral recombinase domain	YqaJ viral recombinase
gene_1455	279	DNA-binding phage-related protein	RecT family	RecT family
gene_1456	233	putative DNA binding protein	Replication initiation and membrane attachment	DnaD domain

Showing 1 to 20 of 72 rows 20 rows per page

Closest Phages

Candidate ID	Subject ID	Subject Name	Identity	Query coverage	E value	Bitscore
Candidate_22	AB711120.1	Bacillus phage PM1	66%	4%	5.72e-75	289
Candidate_22	DQ150593.1	Bacillus phage Fah	76%	1%	1.03e-71	279
Candidate_22	KX965989.1	Aeribacillus phage AP45	68%	1%	1.04e-33	152
Candidate_22	JN700520.2	Staphylococcus phage StB12	72%	0%	3.17e-21	111
Candidate_22	AB823818.1	Thermus phage phi OH2	78%	0%	1.35e-19	105

图 11 a. 预测的前噬菌体区域中每个基因在数据库中（NCBI NR、Pfam 和 InterPro）的注释情况 b. 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。

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

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. *Nucleic Acids Res* 47(W1): W74-W80.

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.