宏病毒组分析培训导论课

汇　报　人：周之超

日　　期：2025/11/30

目录

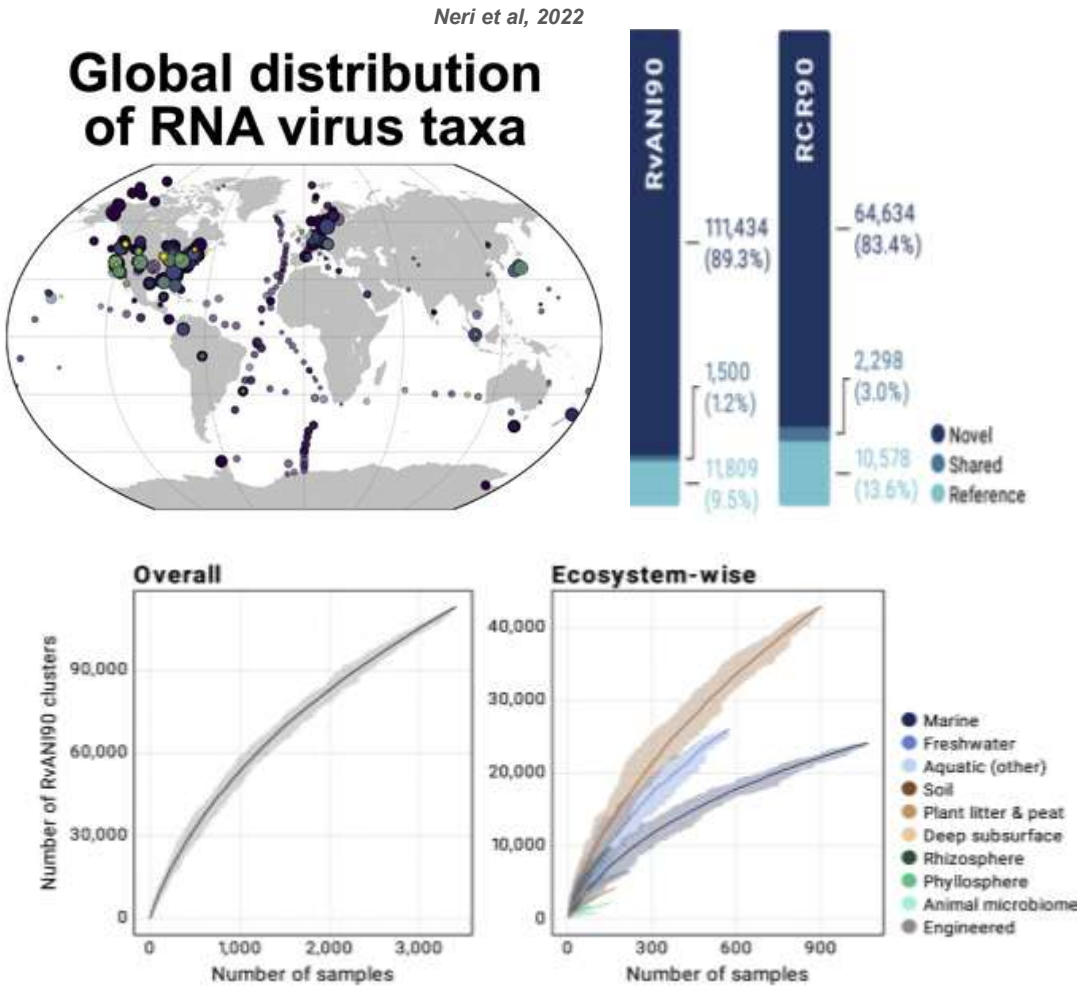
全球病毒多样性的黑箱现状

•病毒在生态系统中数量庞大、种类繁多，但目前我们对其了解仍十分有限。

•全球病毒的多样性就像一个尚未开启的黑箱，隐藏着无数未知等待我们去探索。

•传统研究方法难以全面、深入地揭示病毒的多样性及其生态角色。

Neri et al, 2022



?

需求

传统病毒发现的局限：培养依赖 vs 数据驱动

培养依赖方法：

需要特定的培养条件，操作繁琐且耗时。

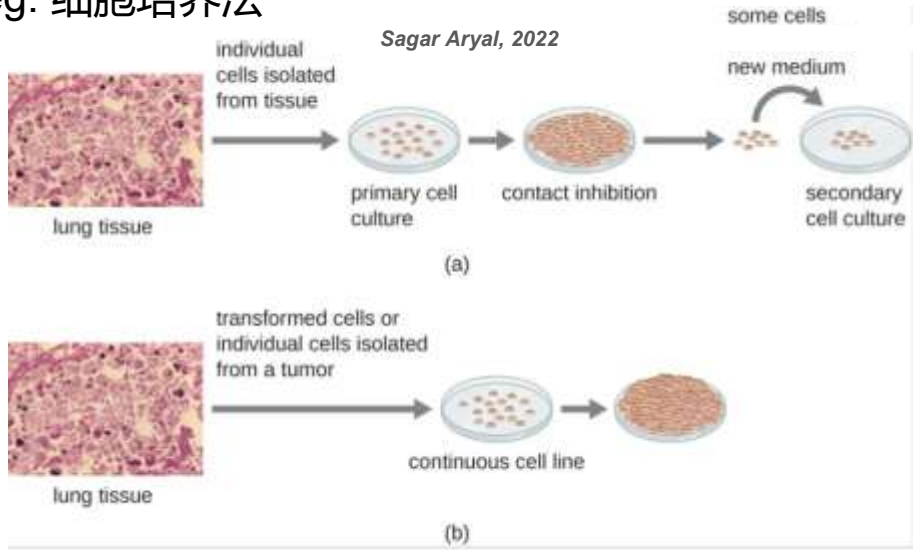
许多病毒对宿主特异性强，难以在实验室中找到合适的培养环境，导致大量病毒无法被培养和研究。

数据驱动方法：

利用高通量测序等技术直接获取病毒的遗传信息，突破了培养条件的限制。

能够更全面、快速地发现病毒，为病毒多样性研究带来新的机遇。

eg. 细胞培养法



需求

快速且准确获得大量病毒序列信息

从病毒暗物质谈起：我们为什么需要宏病毒组

宏转录组的独特价值：聚焦“活跃”的病毒生态系统

•宏转录组技术能够直接从环境样本中获取活跃病毒的遗传信息，无需对病毒进行培养。

•聚焦活跃病毒生态系统，有助于我们了解病毒在自然环境中的真实存在状态、动态变化以及与其他生物的相互作用。

•为研究病毒的生态功能、传播机制和进化规律提供了全新的视角和有力的手段。

例如，在土壤生态系统中，通过宏转录组测序可以揭示土壤病毒活动，进而深入探究不同土壤类型中病毒的生态学效应

病毒研究的核心问题：发现、分类、进化、互作、功能

发现：寻找新的病毒种类，拓展病毒多样性认知的边界。

分类：根据病毒的遗传特征和生物学性质，将其进行系统分类，构建病毒的分类体系。

进化：研究病毒的起源、进化历程以及与其他生物的共同进化关系，揭示病毒的进化规律。

互作：探讨病毒与宿主、病毒与环境之间的相互作用，理解病毒在生态系统中的角色和功能。

功能：解析病毒的生物学功能，包括病毒的致病机制、对宿主的影响以及在生物地球化学循环中的作用等。

研究路径全景：从原始数据到病毒生物学认知

从“拼图”到“故事”：如何从一堆reads中还原病毒生态学意义?

实验流程

样本采集

宏病毒组分析的起点在于从自然环境（如土壤、海洋或宿主组织）中采集代表性样本。这些样本可能包含丰富的微生物群落，包括未知病毒。通过科学的采样策略，确保数据的多样性和代表性，为后续分析奠定基础。

去宿主

在采集的样本中，宿主 DNA/RNA 往往占主导地位，需要通过过滤或生物信息学方法（如比对已知宿主基因组）去除这些序列。这一步骤旨在富集病毒相关序列，提高后续测序和分析的效率和准确性。

建库测序

将富集后的病毒 DNA/RNA 构建成测序文库，并利用高通量测序技术（如 Illumina 或 PacBio）生成海量 reads 数据。

数据分析

质控

对原始测序数据进行质量评估和过滤，去除低质量读长和接头序列

去宿主/去rRNA

利用生物信息学工具去除宿主和核糖体RNA的序列，提高病毒序列的富集度

组装

将高质量的病毒读长进行拼接，生成完整的病毒基因组序列。

注释

对组装后的序列进行功能和分类注释，确定病毒的种类和功能。

挖掘

深入分析病毒的生态学特征、进化关系和与其他生物的相互作用。

病毒组装与发现

为什么选择MetaSpades、MEGAHIT等拼接工具?

MEGAHIT等拼接工具具有高效性和准确性，能够快速处理大规模的测序数据。

它们采用了先进的算法，可有效提高病毒序列的拼接质量和完整性。

如何避免组装假阳性？去宿主、去rRNA的关键细节

在拼接前需对数据进行严格的质控和预处理，去除宿主序列和rRNA序列的干扰

采用多种生物信息学工具和策略相结合，提高病毒序列的特异性识别和拼接准确性。

BLAST/HMM等策略在RNA病毒识别中的适用性

BLAST可用于将拼接后的序列与已知病毒数据库进行比对，快速识别相似的RNA病毒序列。

HMM能够识别序列中的保守结构域和特征模式，提高RNA病毒的识别准确性和灵敏度。

AI如何重塑病毒生物信息学?

- 高效的数据处理能力：AI能够快速处理大规模的病毒测序数据，提高分析效率。
- 精准的序列分析：利用机器学习算法，AI可以更准确地识别病毒序列，发现潜在的新病毒。
- 深度的特征挖掘：AI能够挖掘病毒序列中的隐藏特征，为病毒分类、进化和功能研究提供新的视角。
- 自动化和智能化：减少人工操作，降低误差，实现病毒分析的自动化和智能化。

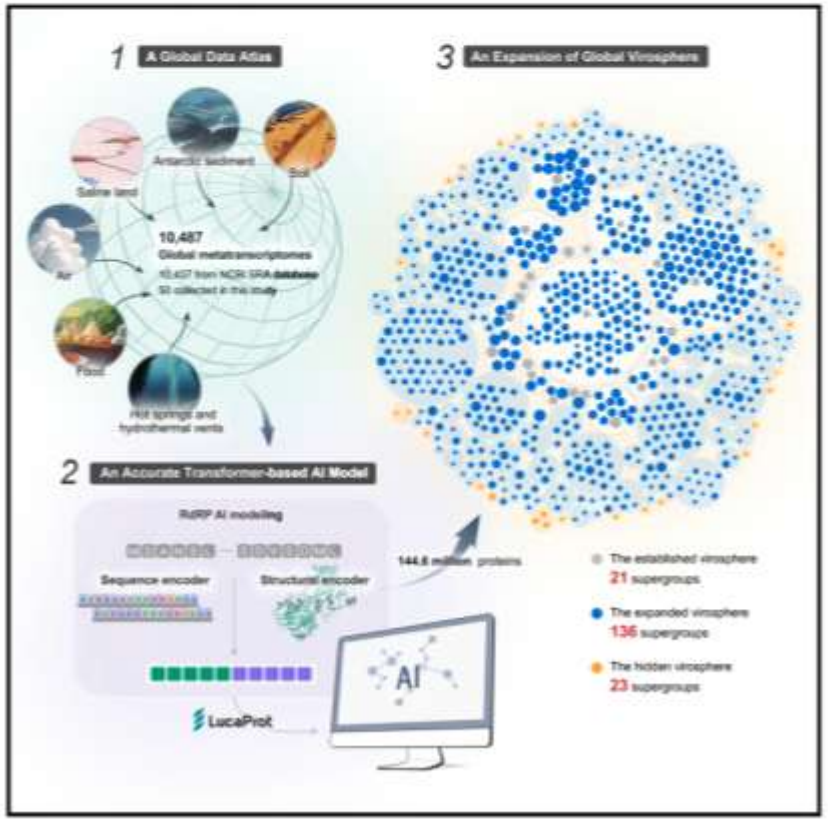
病毒研究进入AI时代：模型助力科学发现

Cell

Article

Using artificial intelligence to document the hidden RNA virosphere

Graphical abstract



Authors

Xin Hou, Yong He, Pan Fang, ..., Edward C. Holmes, Zhao-Rong Li, Mang Shi

Correspondence

edward.holmes@sydney.edu.au (E.C.H.), zhaorong.lzr@alibaba-inc.com (Z.-R.L.), shim23@mail.sysu.edu.cn (M.S.)

In brief

A deep learning algorithm (LucaProt) that integrates both sequence and predicted structural information was employed to identify highly divergent RNA viral "dark matter" in 10,487 metatranscriptomes from diverse global ecosystems. A total of 161,979 potential RNA virus species and 180 RNA virus supergroups were unveiled using this artificial intelligence approach, including many understudied groups.

2024.10

❑ 中山大学 施莽 + 阿里飞天实验室 李兆融

- Transformer大模型 **LucaProt**
- 学习大量病毒和非病毒基因组序列，制定病毒判断标准
- 融合蛋白质序列和隐含结构信息识别功能，用于蛋白质功能鉴定

❑ 核心功能与表现

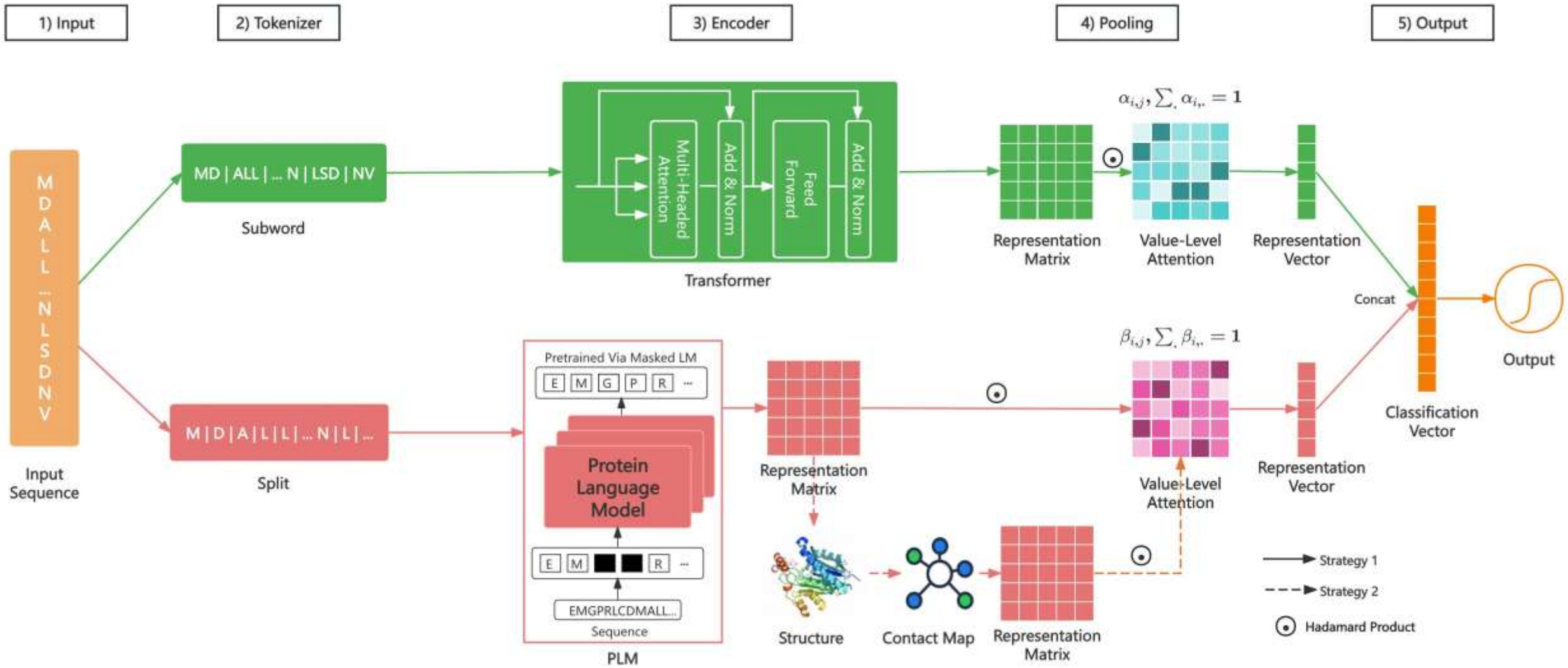
- 假阳性率：0.014%　　RNA病毒挖掘能力
- 假阴性率：1.72%　　（基于RdRp marker）

❑ 关键成果

- 分析全球 **10,487 份 RNA 测序数据**
  - 发现 **超51万条病毒基因组**
  - 涉及 **超16万个潜在病毒种**及 **180 个 RNA 病毒超群**（数量扩增 **9 倍**）
  - 发现 **23**个无法通过序列同源方法识别的超群，被称为**病毒圈的 "暗物质"**
- 识别**最大 RNA 病毒基因组**（**47,250** 核苷酸），展现极高遗传复杂性

什么是LucaProt模型?

两个分类轨道：**BPE + Transformer** 和 **ESM2-3B**。使用值级注意力池化（Value-Level Attention Pooling）进行向量表示，并通过拼接（Concatenation）完成分类



分类与注释

基因组注释的难点与应对策略

基因组注释面临病毒基因的多样性、功能保守性差等难点。

采用多种蛋白预测和功能分析方法相结合，如同源性比对、保守结构域分析、基因本体论注释等，提高注释的准确性和可靠性。

案例分析

多种数据库比对：KEGG数据库、VOG数据库
EggNOG emapper数据库

HHpred：远源序列搜索（online和本地数据库）

进化与互作分析

病毒重组检测

重组现象的重要性：病毒重组是病毒进化的重要机制之一，可导致新的病毒株的出现，影响病毒的传播和致病性。
常用检测方法：
RDP（Recombination Detection Program）：基于统计学方法检测重组事件，适用于多种类型的病毒序列。
GARD（Genetic Algorithm for Recombination Detection）：利用遗传算法检测重组断点，具有较高的准确性和灵敏度。
Bootscan：通过比较不同区域的序列相似性来检测重组，适用于分析病毒序列的局部重组情况。

系统发育树构建

系统发育树的意义：系统发育树能够展示病毒之间的进化关系，帮助我们理解病毒的起源、进化历程和亲缘关系。
常用构建方法：
邻接法（Neighbor-Joining, NJ）：基于距离矩阵构建系统发育树，计算简单，适用于大规模数据。
最大似然法（Maximum Likelihood, ML）：根据序列数据的最大似然估计构建树，考虑了序列进化的模型，结果较为准确。
贝叶斯法（Bayesian Inference）：利用贝叶斯统计方法构建系统发育树，能够提供分支的置信度，适用于复杂的数据分析。

宿主预测

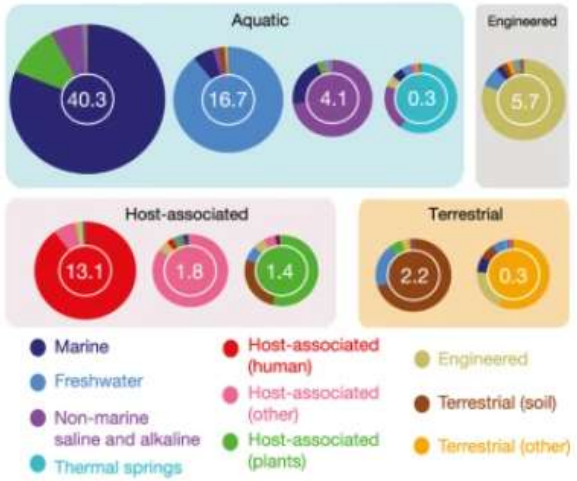
宿主预测的重要性:
宿主预测是理解病毒-宿主相互作用的关键，能够揭示病毒的生态位和传播途径，从而预测潜在的跨种传播风险和流行病暴发可能性。准确识别宿主有助于阐明病毒在自然生态系统中的作用及进化动态。
常用预测方法:
VirHostMatcher: 通过 k-mer 相似性分析病毒与宿主基因组，快速识别可能的宿主，特别适合未注释序列。
CRISPR Spacer Matching: 利用 CRISPR 系统中的 spacer 序列与病毒 DNA 比对，检测病毒与细菌宿主的特异性交互，适用于噬菌体研究。
iPHoP: 基于机器学习和同源性比对，预测病毒与宿主的潜在关系，适用于宏基因组数据，准确性较高

工具介绍 病毒组和从宏基因组中重建病毒

Uncovering Earth's virome

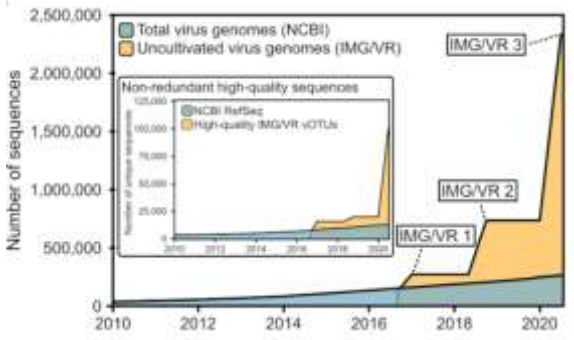
David Paez-Espino & Nikos C. Kyrpides
Nature 2016 (DOE-JGI)

Virome:
metagenomes specifically targeting the viral fraction of environmental samples



Viral genomes reconstructed from metagenome:
viral genomes reconstructed from non-viruses targeted bulk metagenomes

IMG V/R v3, NAR 2021 (DOE-JGI)



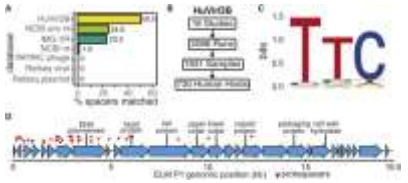
2021
IMG V/R database (v3)
(2023 – IMG V/R v4)

- composed of 18,373 cultivated and 2,314,329 uncultivated viral genomes (UViGs)
- clustered into 935,362 viral Operational Taxonomic Units (vOTUs)
- gathered from 6 primary sources from IMG/M metagenomes, RefSeq/WGS genomes, and other genome/metagenome datasets

2019
Global Oceans Viromes (GOV) v2.0

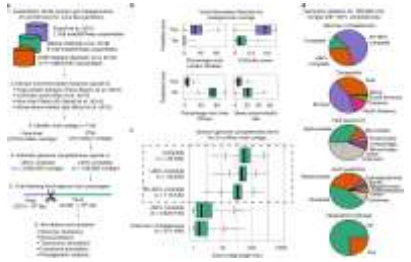
- 200,000 marine viral populations
- separated into five ecological zones, globally: bathypelagic, temperate and tropical epipelagic, mesopelagic, and two Arctic regions

2019
Human virome database (HuVirDB)



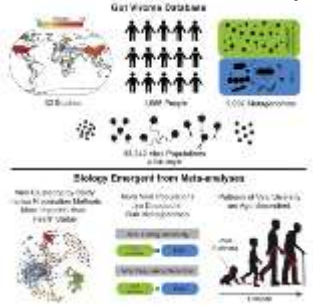
- integrates data from 18 publicly available virome studies representing 1,831 samples from 730 humans from 9 countries
- includes skin, stool, lung and blood samples

2021
Metagenomic Gut Virus (MGV)



- contains 189,680 viral draft genomes from 11,810 publicly available human stool metagenomes
- represents 54,118 candidate viral species.

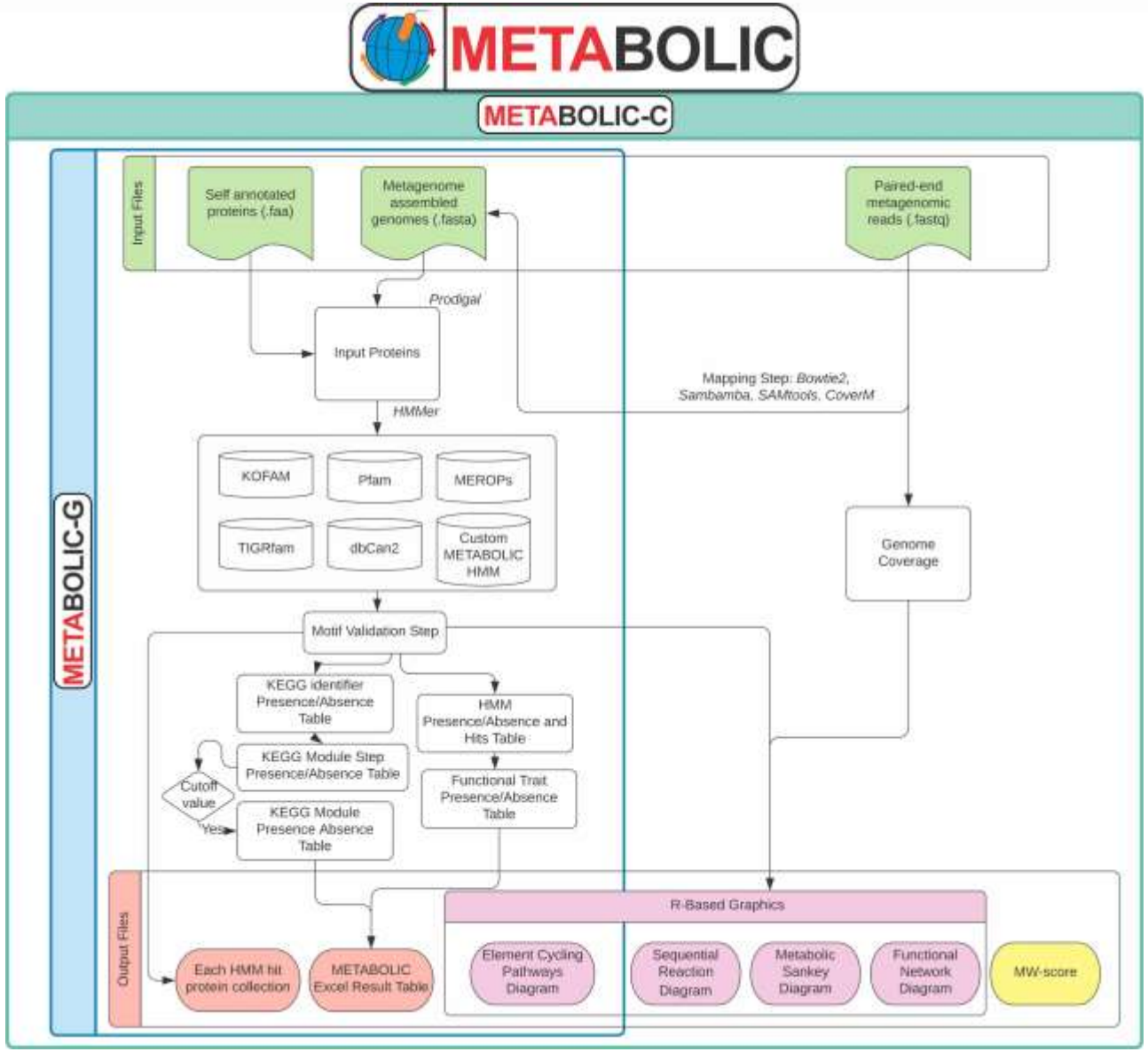
2020
Gut Virome Database (GVD)



(Uses both methods)

- built from 2,697 viral particle or microbial metagenomes from 1,986 individuals representing 16 countries
- contains 33,242 unique viral populations (approximately species-level taxa)

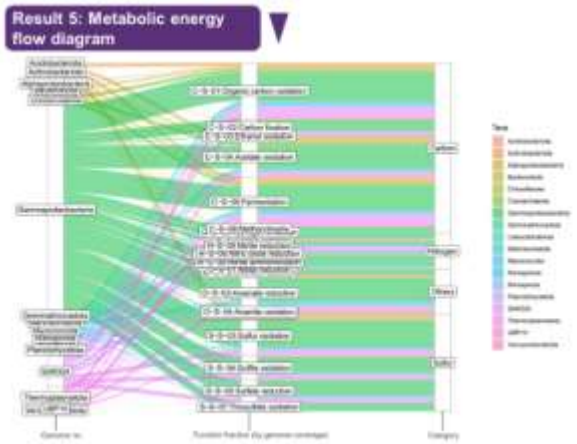
工具1 METABOLIC: 基于微生物基因组的代谢与生物地球化学功能特征分析器



Wide application in many aspects of microbiome studies

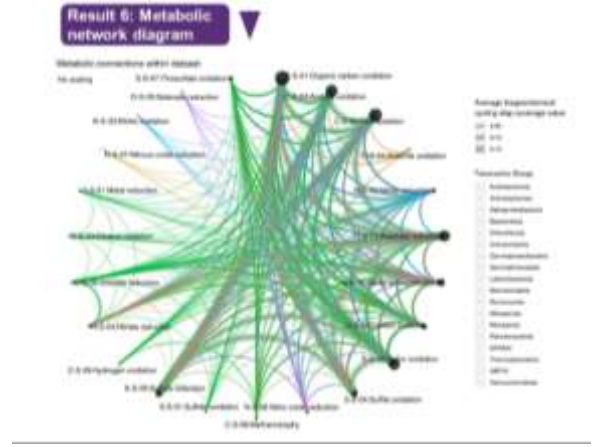
❑ Metabolic Sankey diagram

呈现微生物对各个生物地球化学过程的贡献



❑ Functional network diagram

呈现功能共享网络中的功能连接

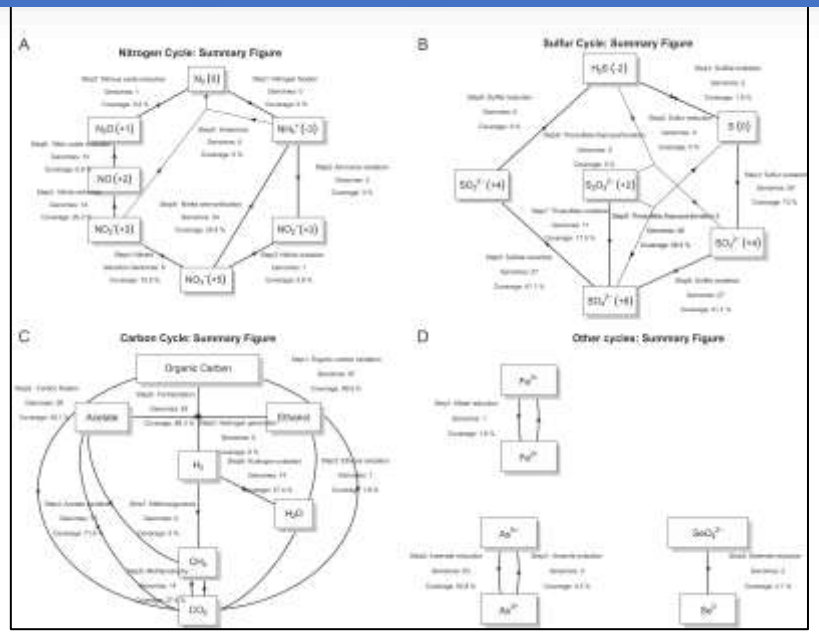


192 WOS citations (截止2024.10)

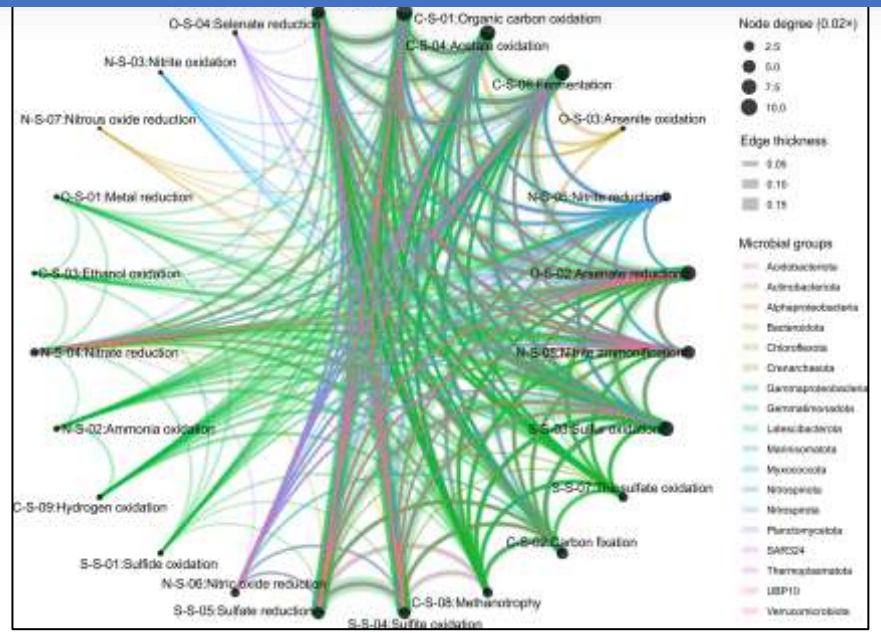
Zhou et al., ***Microbiome*** (2022)

工具1 METABOLIC: 基于微生物基因组的代谢与生物地球化学功能特征分析器

- 群落水平的生物地球化学循环过程（碳、氮、硫等循环）



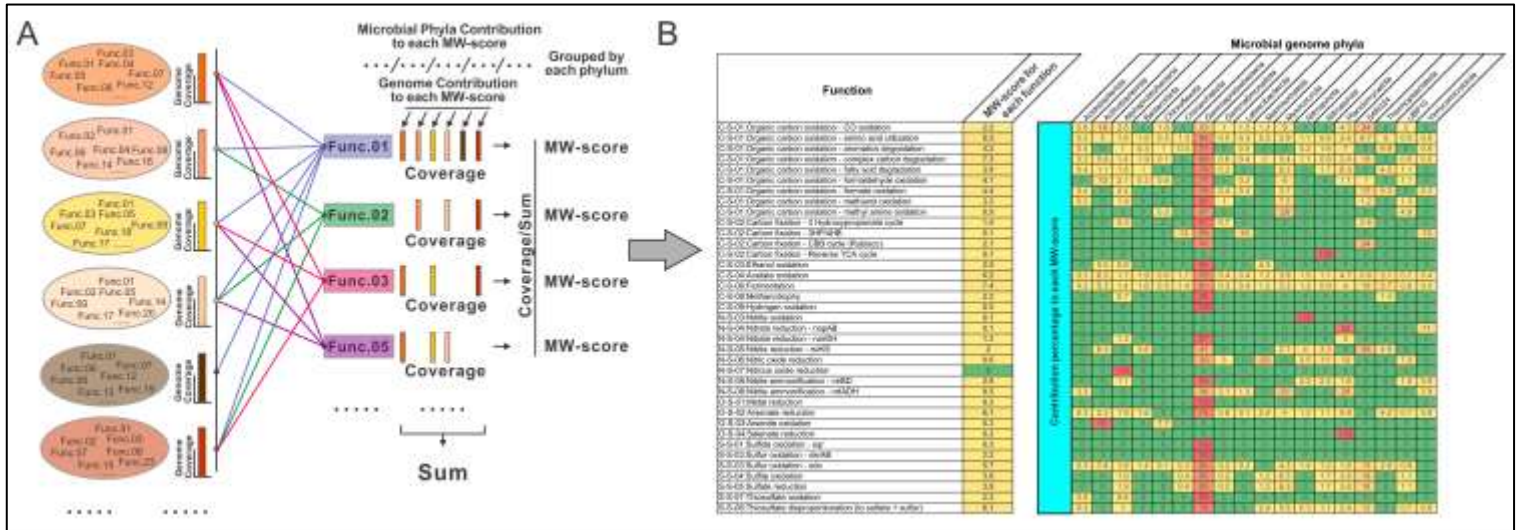
1. Biogeochemical cycling processes at the community level (C, N, S, and other cycles)



2. Functional networks linking different functions in microbial communities

- 连接微生物群落中不同功能的功能网络

3. MW-score (Metabolic-weight score) calculation and table



- 代谢权重评分 (MW-score) 计算及表格呈现

工具1 METABOLIC: 基于微生物基因组的代谢与生物地球化学功能特征分析器

Real life application of METABOLIC

应用于多个生态系统：淡水、海洋、
生物膜、地下水等

Article | Open Access | Published: 09 February 2021

Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika

Patricia Q. Tran, Samantha C. Bachand, Peter B. McIntyre, Benjamin M. Kraemer, Yvonne Vadeboncoeur, Ismael A. Kimirei, Rashid Tamatamah, Katherine D. McMahon & Karthik Anantharaman

The ISME Journal **15**, 1971–1986 (2021) | Cite this article

5831 Accesses | **7** Citations | **72** Altmetric | Metrics

Article | Published: 30 July 2021

Large-scale protein level comparison of Deltaproteobacteria reveals cohesive metabolic groups

Marguerite V. Langwig, Valerie De Anda, Nina Dombrowski, Kiley W. Seitz, Ian M. Rambo, Chris Greening, Andreas P. Teske & Brett J. Baker

The ISME Journal **16**, 307–320 (2022) | Cite this article

2018 Accesses | **4** Citations | **52** Altmetric | Metrics

Article | Published: 14 October 2021

Sulfur cycling and host-virus interactions in *Aquificales*-dominated biofilms from Yellowstone's hottest ecosystems

Luke J. McKay, Olivia D. Nigro, Mensur Dlakić, Karen M. Luttrell, Douglas B. Rusch, Matthew W. Fields & William P. Inskeep

The ISME Journal **16**, 842–855 (2022) | Cite this article

1086 Accesses | **1** Citations | **41** Altmetric | Metrics

Article | Open Access | Published: 25 January 2021

Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems

Christine He, Ray Keren, Michael L. Whittaker, Ibrahim F. Farag, Jennifer A. Doudna, Jamie H. D. Cate & Jillian F. Banfield

Nature Microbiology **6**, 354–365 (2021) | Cite this article

10k Accesses | **19** Citations | **73** Altmetric | Metrics

Article | Open Access | Published: 03 May 2021

Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem

Ya-Jou Chen, Pok Man Leung, Jennifer L. Wood, Sean K. Bay, Philip Hugenholtz, Adam J. Kessler, Guy Shelley, David W. Waite, Ashley E. Franks, Perran L. M. Cook & Chris Greening

The ISME Journal **15**, 2986–3004 (2021) | Cite this article

6403 Accesses | **13** Citations | **62** Altmetric | Metrics

工具2 VIBRANT: 通过迭代注释进行病毒鉴定

Kristopher Kieft, Zhichao Zhou, Karthik Anantharaman ***Microbiome*** (2020)

456 WOS citations (截止2024.10)

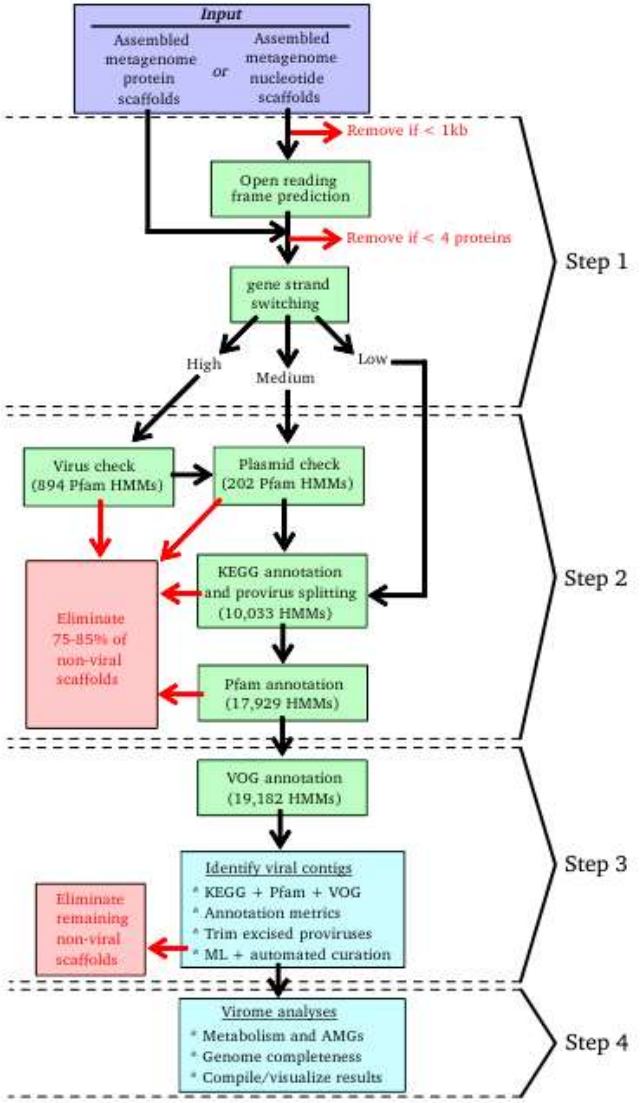
1. 神经网络机器学习 + 后续自动化处理步骤
   - 将病毒序列从混合的宏基因组assembly中分离出来
   - 将原病毒区域从细菌/古菌基因组中剪切出来

2. 代谢功能和基因组质量分析（定性）

3. 全面的病毒基因组功能注释

VIBRANT

Virus Identification By iteRative ANnoTation



KEGG

Pfam

VOGDB

Virus Orthologous Groups

Auxiliary metabolic genes

辅助代谢基因鉴定

(病毒对宿主功能的辅助作用相关基因)

Verified across diverse environments

Freshwater, marine, human, soil etc.

Important biogeochemical cycling functions

工具2 VIBRANT: 通过迭代注释进行病毒鉴定

❑ 27 个 metrics（参数）

total genes
all KEGG
category KEGG
all Pfam
category vPfam
all VOG
category VOG
KEGG int-rep
KEGG zero
Pfam int-rep
Pfam zero
VOG redoxin
VOG rec-tran
VOG int
VOG RnR
VOG DNA
KEGG restriction check
KEGG toxin check
VOG special
annotation check
p_v check
p_k check
k_v check
k check
p check
v check
h check

KEGG v-score sur

Pfam v-score sum

VOG v-score

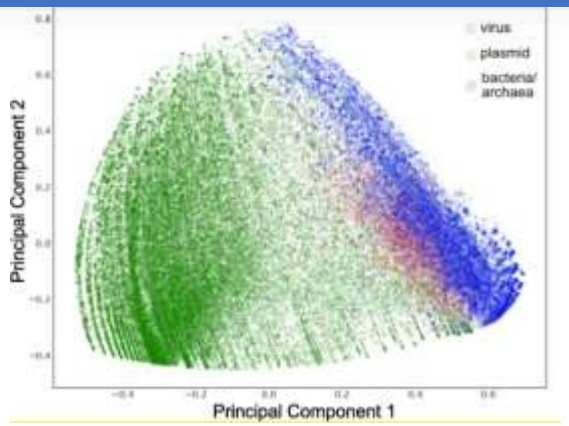
三个v-score是最主要的参数

v-score（0-10）：
衡量HMM注释与病毒蛋白相关性的评分指标

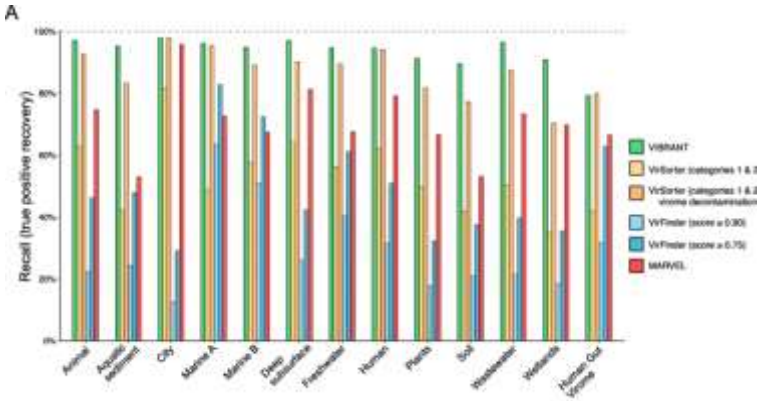
- HMM对病毒蛋白数据库显著命中数进行计算（n / 100）（最大为10）
- 手动调整以增加病毒marker基因的权重（最小为1）

Performance & future development

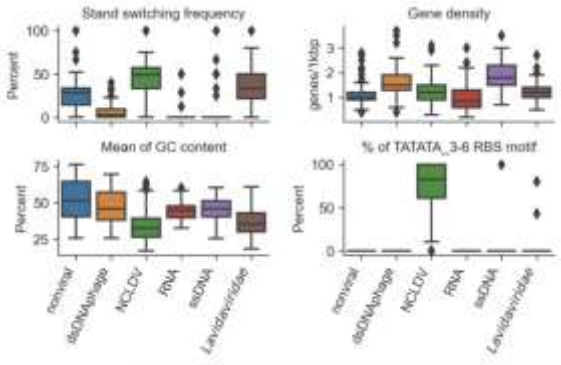
1. Distinguish virus, plasmid, and bacteria/archaea based on Machine Learning algorithm
（**神经网络** + 随机森林）



2. Better or equal performance comparing to other software



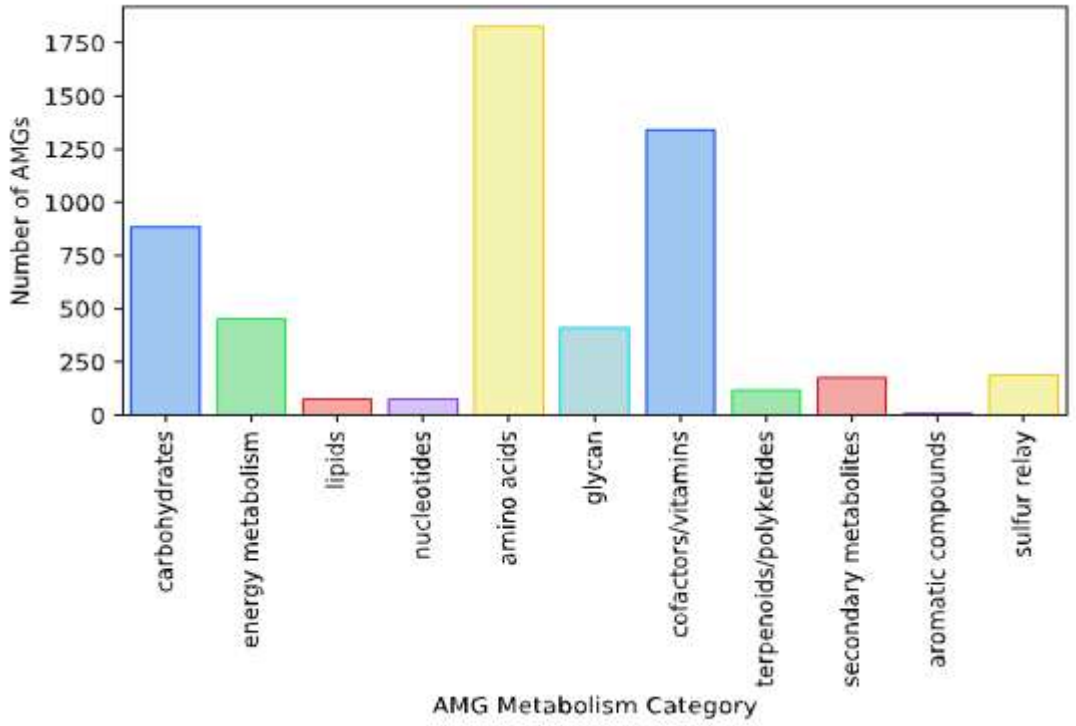
3. Add gene density and TATATA 3-6 RBS motif as additional metrics in the future's version

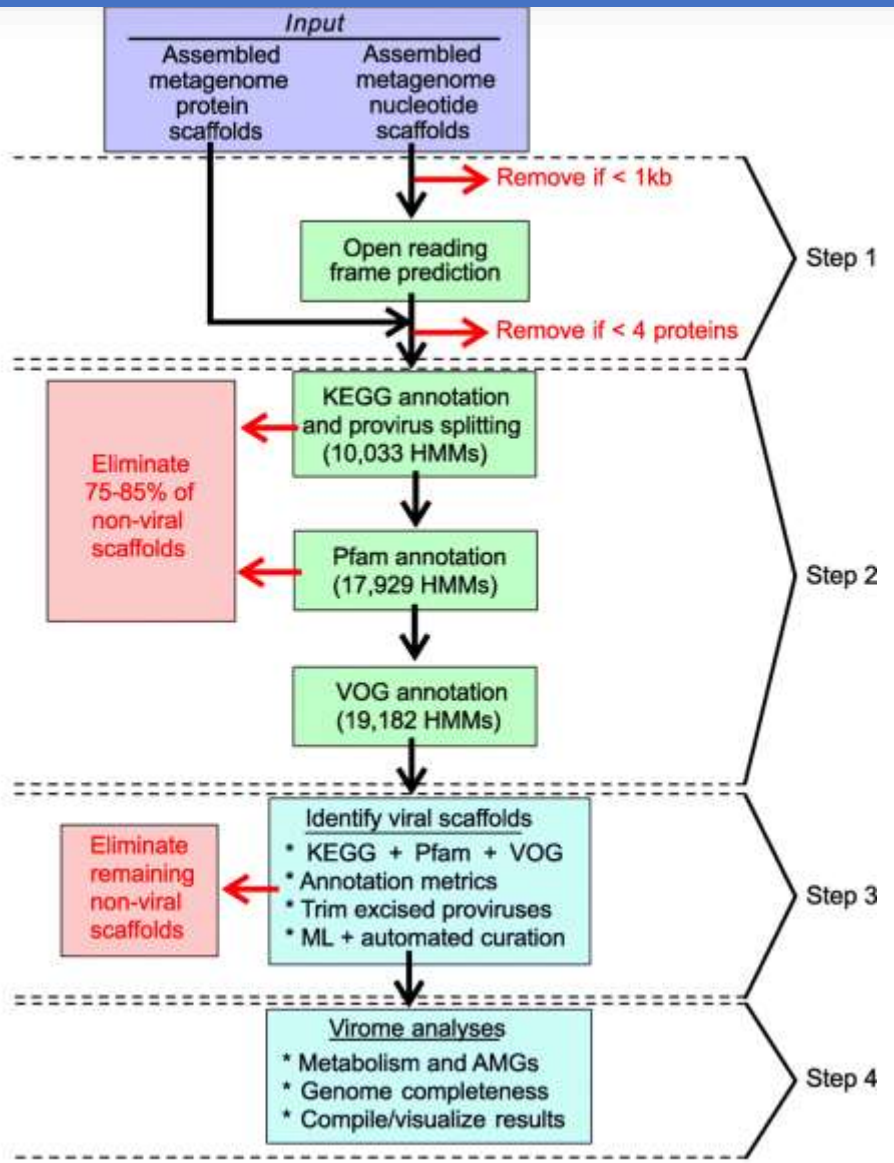


courtesy of VirSorter2

辅助代谢基因AMG注释结果展示

- Class I AMGs and sulfur relay
- AMG individuals, summary counts, and metabolic pathways

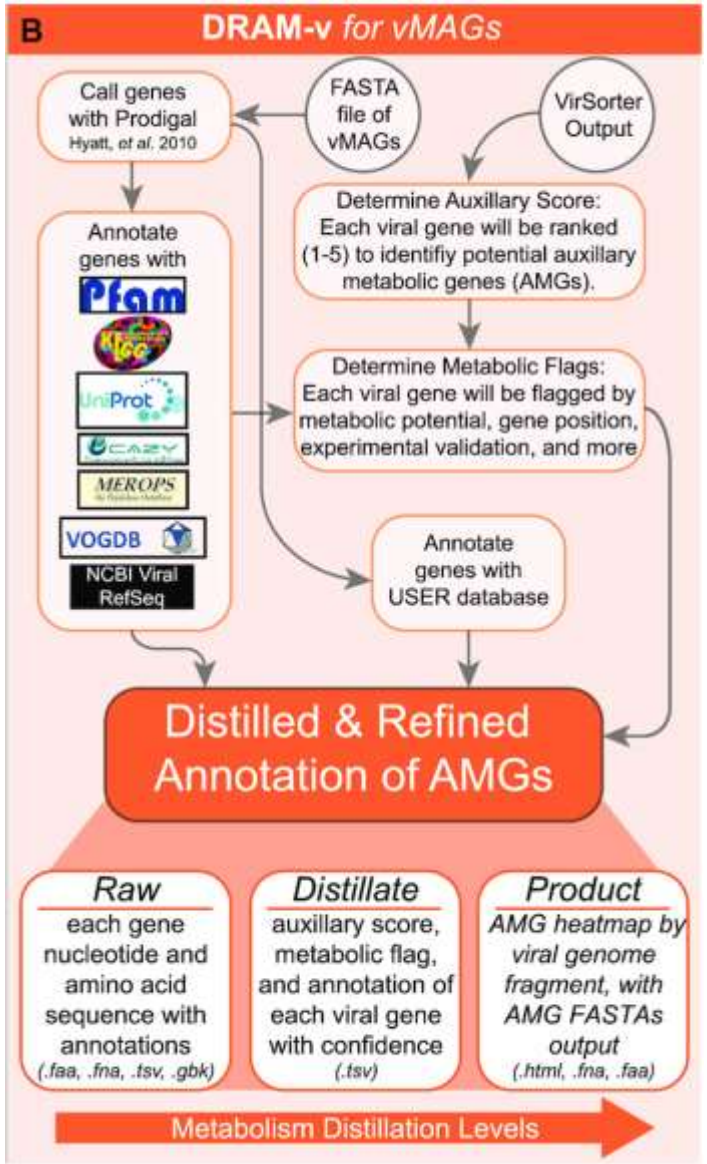




Workflow of using protein annotation VIBRANT, Kieft et al. 2020

将AMG和KEGG注释连接起来 – 两个软件的对比

- VIBRANT AMGs
  - 2,835 KEGG
    - “metabolic pathways” or “sulfur relay system” category (legitimate metabolic gene, exclude glycosyl transferases, glycoside hydrolase, *dcm*)
    - Manual inspection; Remove non-AMG annotations, e.g., *nrdAB* (核糖核苷酸还原酶) and *thyAX* (胸苷酸合成酶)
    - Remove annotation associated with direct nucleotide to nucleotide conversions
  - 104 with a v-score > 0.02
- How to relate KEGG annotations to Pfam and other databases?
  - DRAM-v has done this with curated list
    - 279 AMGs (34 KEGG)
    - 19 KEGG modules



全面注释的必要性

- See IMG/VR marine annotations
- K21140
  - mec; [CysO sulfur-carrier protein]-S-L-cysteine hydrolase

protein	scaffold	KO	AMG	KO name	KO evalue	KO score	KO v-score	Pfam	Pfam name	Pfam evalue	Pfam score	Pfam v-score	VOG	VOG name	VOG evalue	VOG score	VOG v-score
3300006749_	3300006749_	K21140	AMG	mec; [CysO s	2.80E-15	57.9	0.89	PF14464.6	Prokaryotic h	5.80E-14	53.4	0.27	VOG01039	sp|O64333|	1.80E-51	177.2	4.6
3300010148_	3300010148_	K21140	AMG	mec; [CysO s	2.60E-14	54.7	0.89	PF00877.19	NlpC/P60 far	2.20E-11	45.2	1.39	VOG01039	sp|O64333|	6.30E-51	175.4	4.6
3300002482_	3300002482_	K21140	AMG	mec; [CysO s	5.30E-16	60.2	0.89	PF14464.6	Prokaryotic h	3.80E-14	53.9	0.27	VOG01039	sp|O64333|	7.20E-52	178.5	4.6
3300006730_	3300006730_	K21140	AMG	mec; [CysO s	1.80E-19	71.5	0.89	PF00877.19	NlpC/P60 far	6.40E-20	72.6	1.39	VOG01039	sp|O64333|	9.50E-68	230.9	4.6

scaffold	KO	AMG	KO name	KO evalue	KO score
3300006749_	K21140	AMG	mec; [CysO sulfur-carrier protein]-S-L-cysteine hydrolase [E	2.80E-15	57.9
3300010148_	K21140	AMG	mec; [CysO sulfur-carrier protein]-S-L-cysteine hydrolase [E	2.60E-14	54.7
3300002482_	K21140	AMG	mec; [CysO sulfur-carrier protein]-S-L-cysteine hydrolase [E	5.30E-16	60.2
3300006730_	K21140	AMG	mec; [CysO sulfur-carrier protein]-S-L-cysteine hydrolase [E	1.80E-19	71.5

KEGG

Pfam	Pfam name	Pfam evalue	Pfam score	Pfam v-score
PF14464.6	Prokaryotic homologs of the JAB domain	5.80E-14	53.4	0.27
PF00877.19	NlpC/P60 family	2.20E-11	45.2	1.39
PF14464.6	Prokaryotic homologs of the JAB domain	3.80E-14	53.9	0.27
PF00877.19	NlpC/P60 family	6.40E-20	72.6	1.39

Pfam

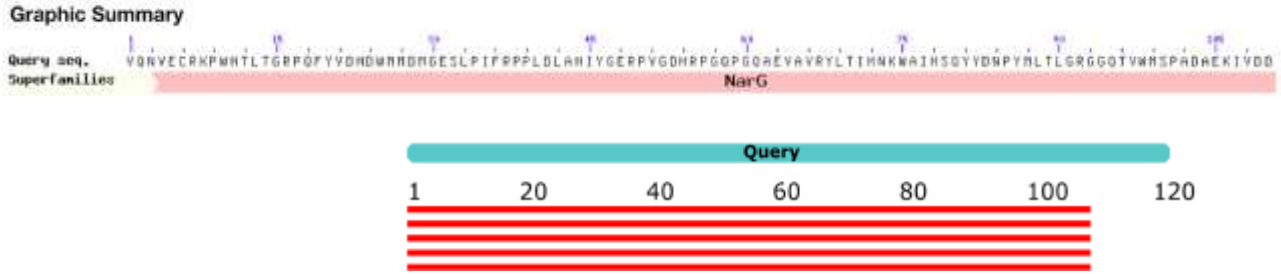
VOG	VOG name	VOG evalue	VOG score	VOG v-score
VOG01039	sp|O64333|TIPK_BPN15 Tail tip assembly protein K	1.80E-51	177.2	4.6
VOG01039	sp|O64333|TIPK_BPN15 Tail tip assembly protein K	6.30E-51	175.4	4.6
VOG01039	sp|O64333|TIPK_BPN15 Tail tip assembly protein K	7.20E-52	178.5	4.6
VOG01039	sp|O64333|TIPK_BPN15 Tail tip assembly protein K	9.50E-68	230.9	4.6

VOG

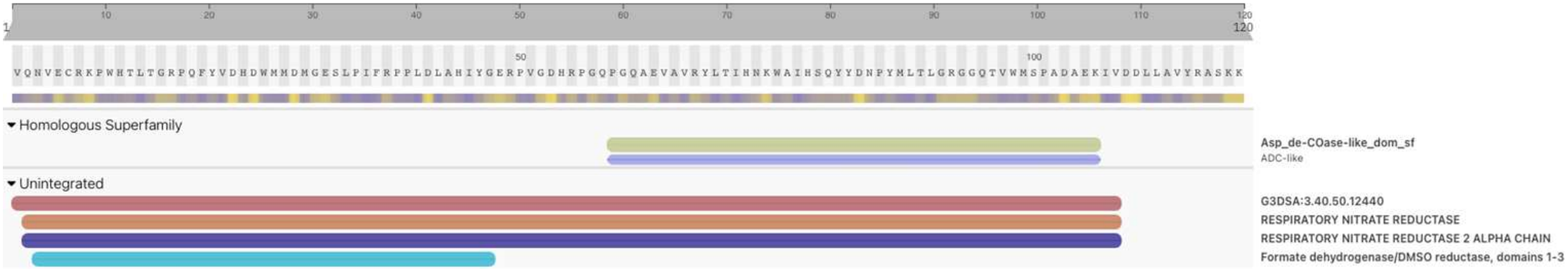
注释的一些简单/快捷方法

Conserved sites

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
nitrate reductase subunit alpha [Isoptericola variabilis]	Isoptericola variabilis	232	232	89%	2e-73	98.13%	362	MBF1232883.1
nitrate reductase subunit alpha [Actinomyces sp.]	Actinomyces sp.	232	232	89%	6e-73	98.13%	376	MBS4875505.1
nitrate reductase subunit alpha [Schaalia odontolytica]	Schaalia odontolytica	234	234	89%	1e-68	100.00%	1256	WP_003791817.1
respiratory nitrate reductase subunit alpha apoprotein [Mycobacteroides abscessus subsp. abscessus]	Mycobacteroides abscessus subsp. abscessus	233	233	89%	1e-68	99.07%	956	SIA93296.1
nitrate reductase subunit alpha [Actinomyces sp.]	Actinomyces sp.	234	234	89%	1e-68	100.00%	1151	MBF0936762.1



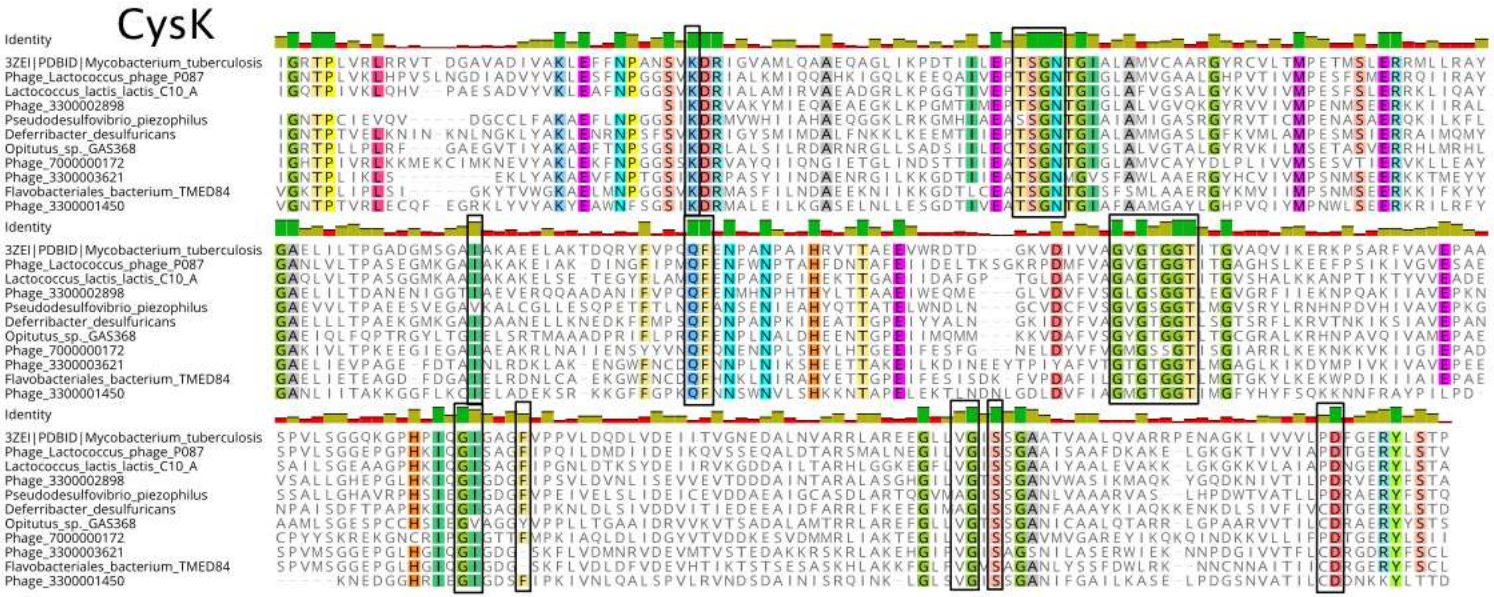
InterProScan

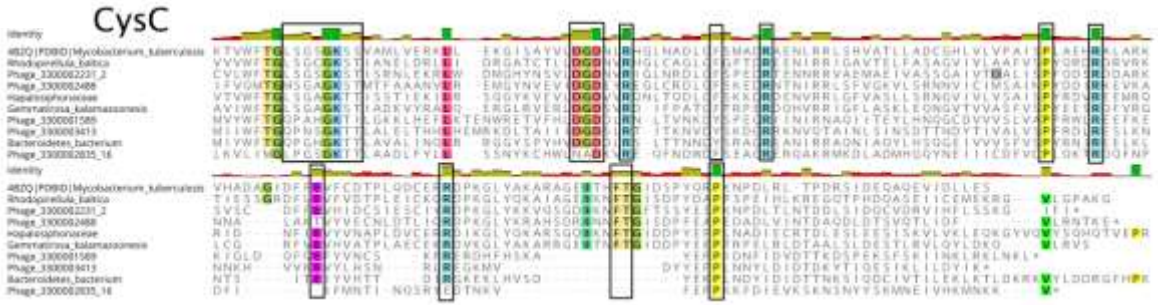


Blastp

功能保守的氨基酸残基

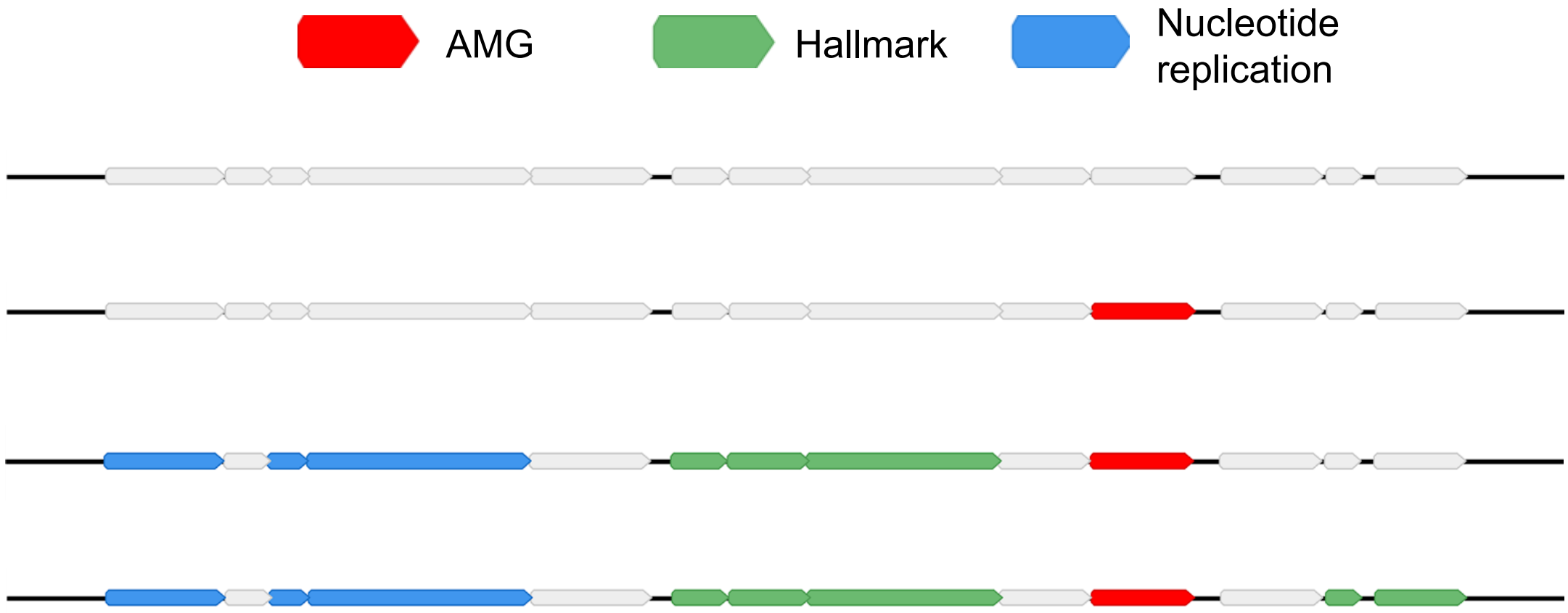
Virus AMGs aligned to PDB references





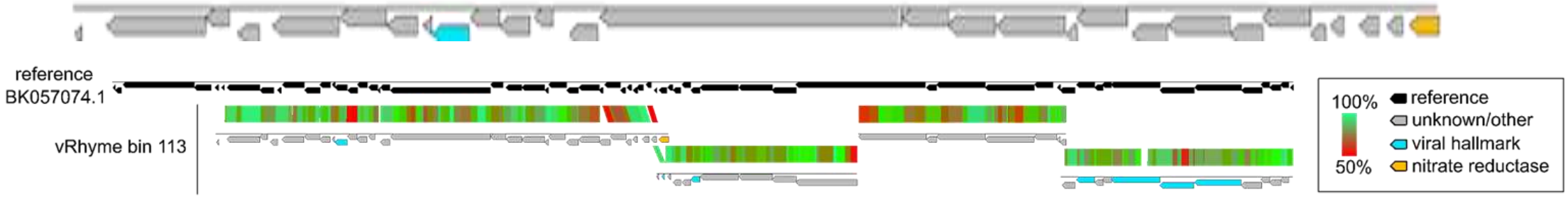
Contig尾部AMG的鉴定

Validating AMGs on contig ends

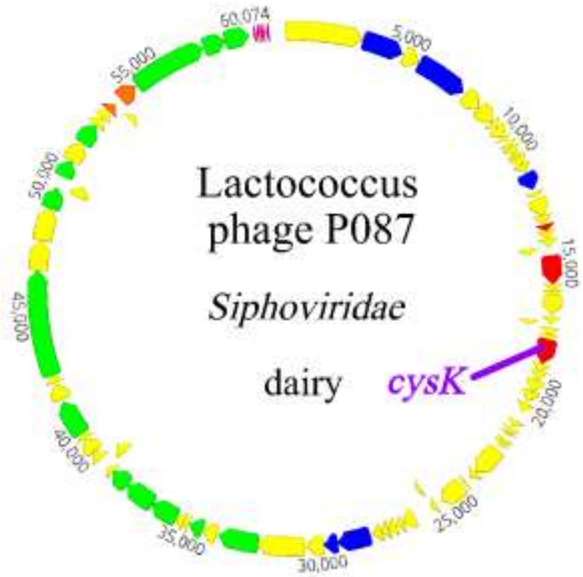


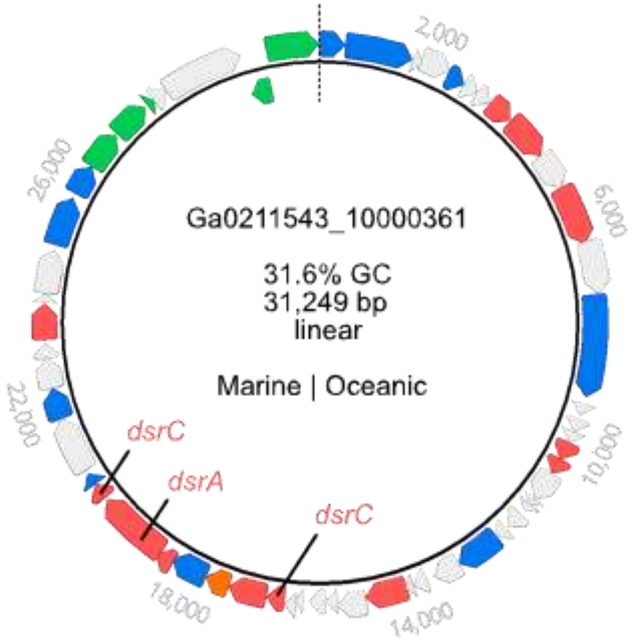
Contig尾部AMG的鉴定

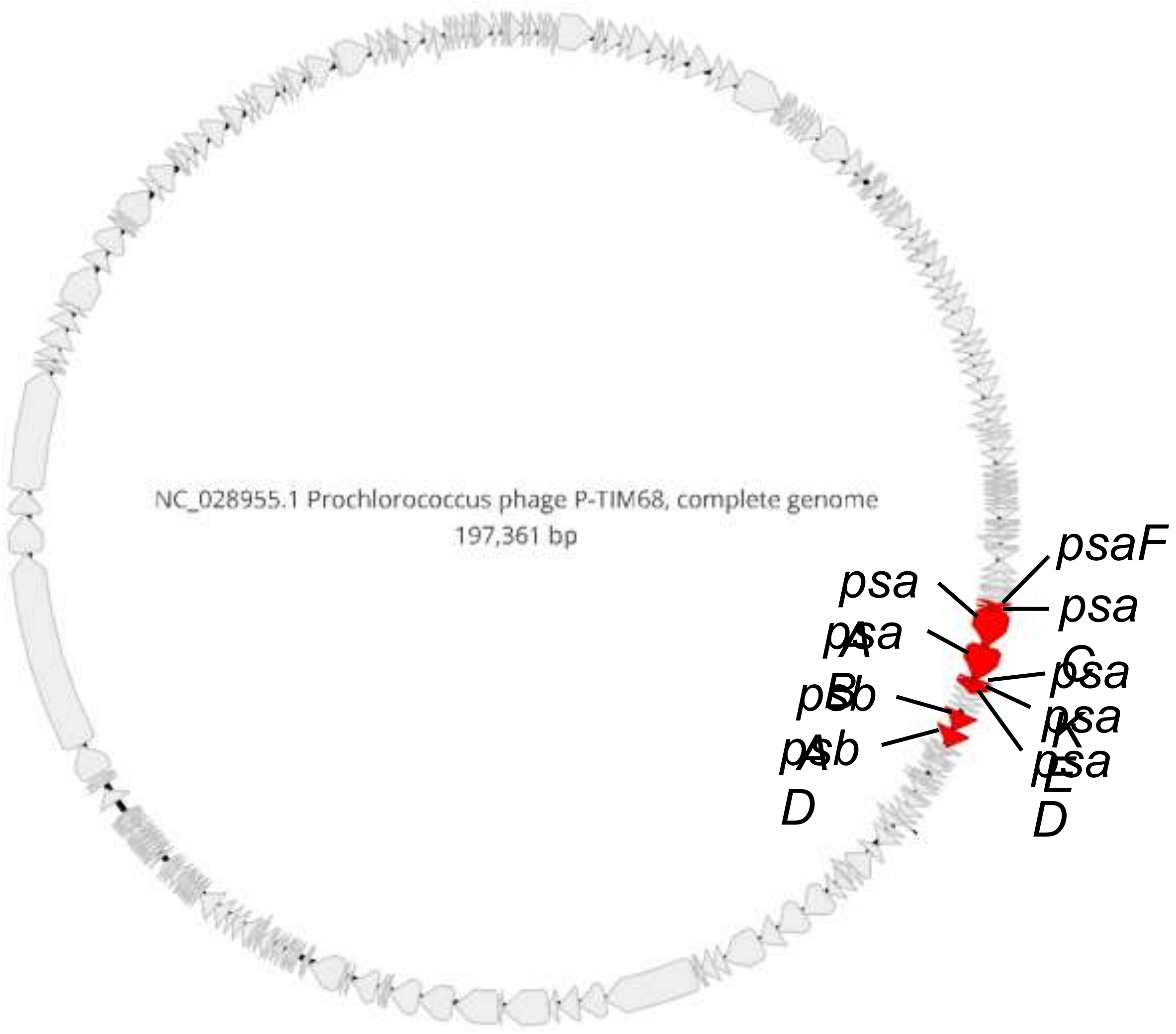
Validating AMGs on contig ends with binning



Singles, multiples, duplicates, and cassettes







In silico 去除/过滤 AMG的方法

- **(1) Filter AMGs at scaffold ends**

  Any AMG placed at either ends of a scaffold or any number of AMGs placed at either ends of a scaffold in tandem should be filtered.

Reason: ambiguous on whether cellular or viral

Scaffold边缘的基因不好判定是病毒的还是宿主细胞来源的

In silico 去除/过滤 AMG的方法

- **(2) Filter AMGs that have any v-scores (KEGG and Pfam v-scores) >= 1**

Parse KEGG v-scores and Pfam v-scores for individual AMGs from the VIBRANT result. If any AMG has any v-scores (KEGG or Pfam v-scores) ≥ 1 (representing a viral-like nature), this AMG should be filtered.

Reason: much like a viral-like/hallmark gene instead of a metabolic gene

V-score >= 1 病毒特性较高，可能属于病毒特征基因而不是AMG

In silico 去除/过滤 AMG的方法

- **(3) Filter AMGs with flanking genes of v-scores (only KEGG v-scores) < 0.25**

For any AMG (or multiple AMGs placed in tandem) surrounded by all their flanking genes with v-scores < 0.25 (only KEGG v-scores considered here), these AMGs are considered to be surrounded by non-viral (cellular) genes. These AMGs should be filtered since they were likely to be non-viral (cellular) in origin.

Reason: much like from cellular fragments

AMG两边的flanking基因如果病毒特性不明显的话，它们可能是来源于宿主细胞

In silico 去除/过滤 AMG的方法

- **(4) Filter AMGs that have COG category as T or B**

Use eggNOG-mapper v2 (in March 2023) to annotate all AMG proteins to get COG category assignment. If any AMG has COG category assigned as “T” (Signal Transduction) or “B” (Chromatin Structure and dynamics), then this AMG should be filtered.

Reason: not likely to be microbial-related metabolisms contained in bacteriophages

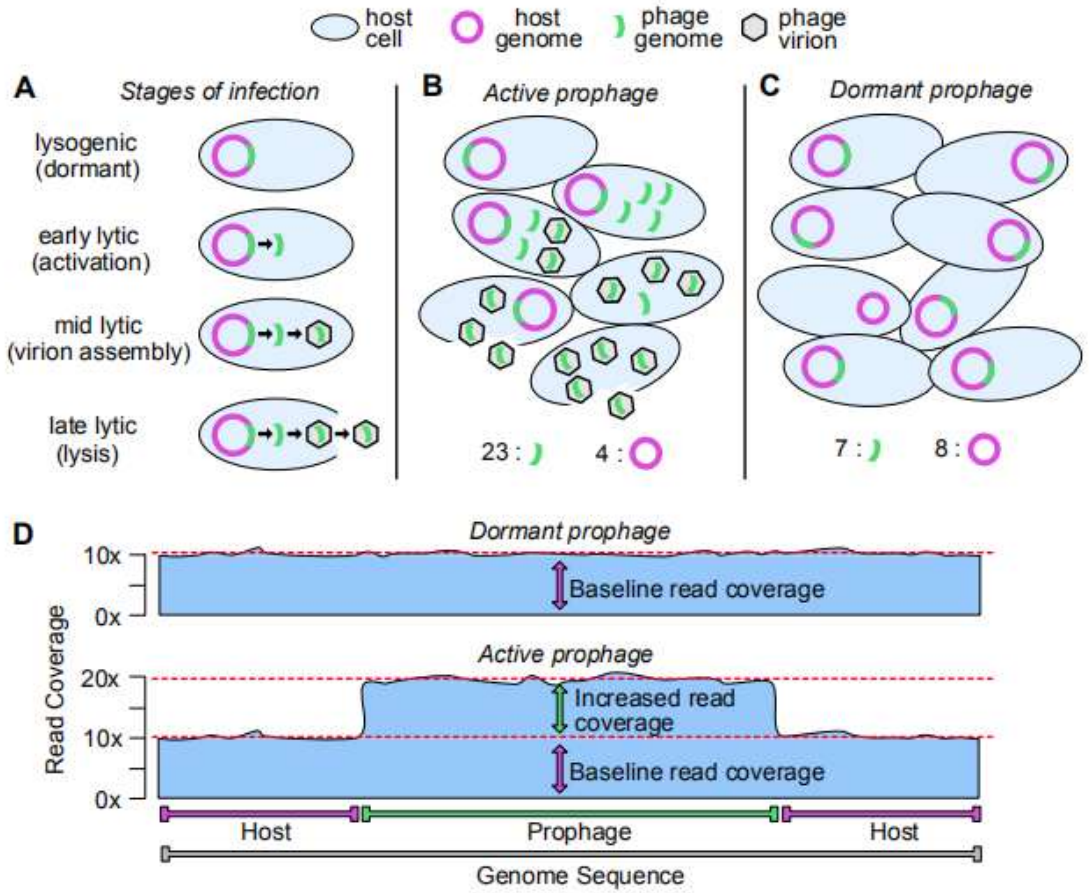
AMG不太可能的一些COG功能组别

工具3 PropagAtE：原噬菌体活性估算

PropagAtE
(Prophage Activity Estimator)

PropagAtE
Prophage Activity Estimator

Kieft K, Anantharaman K. ***mSystems*** 2022



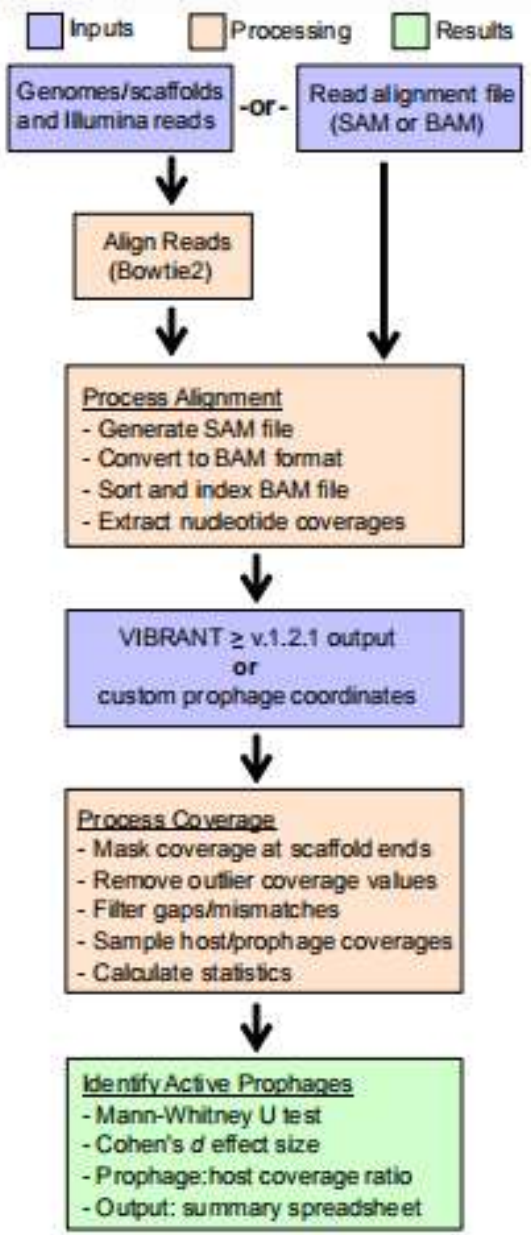
Determine if prophage was actively replicating its genome (active state) or stay dormant, integrated in the genome

significantly more prophage genome copies than host copies

工具3 PropagAtE：原噬菌体活性估算

Workflow

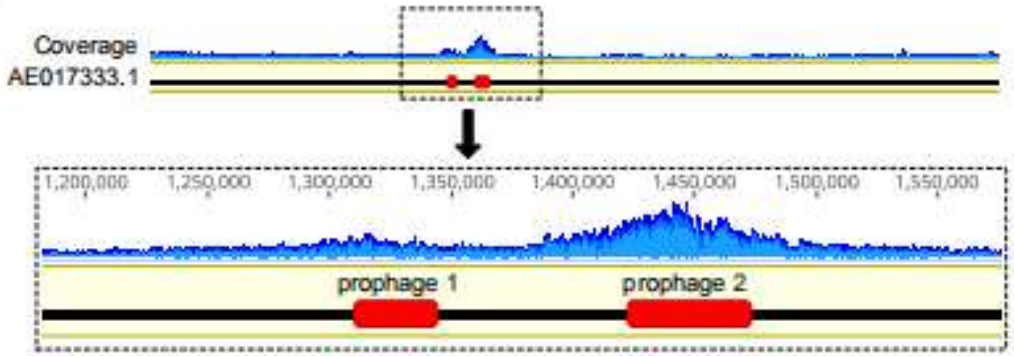
A



B

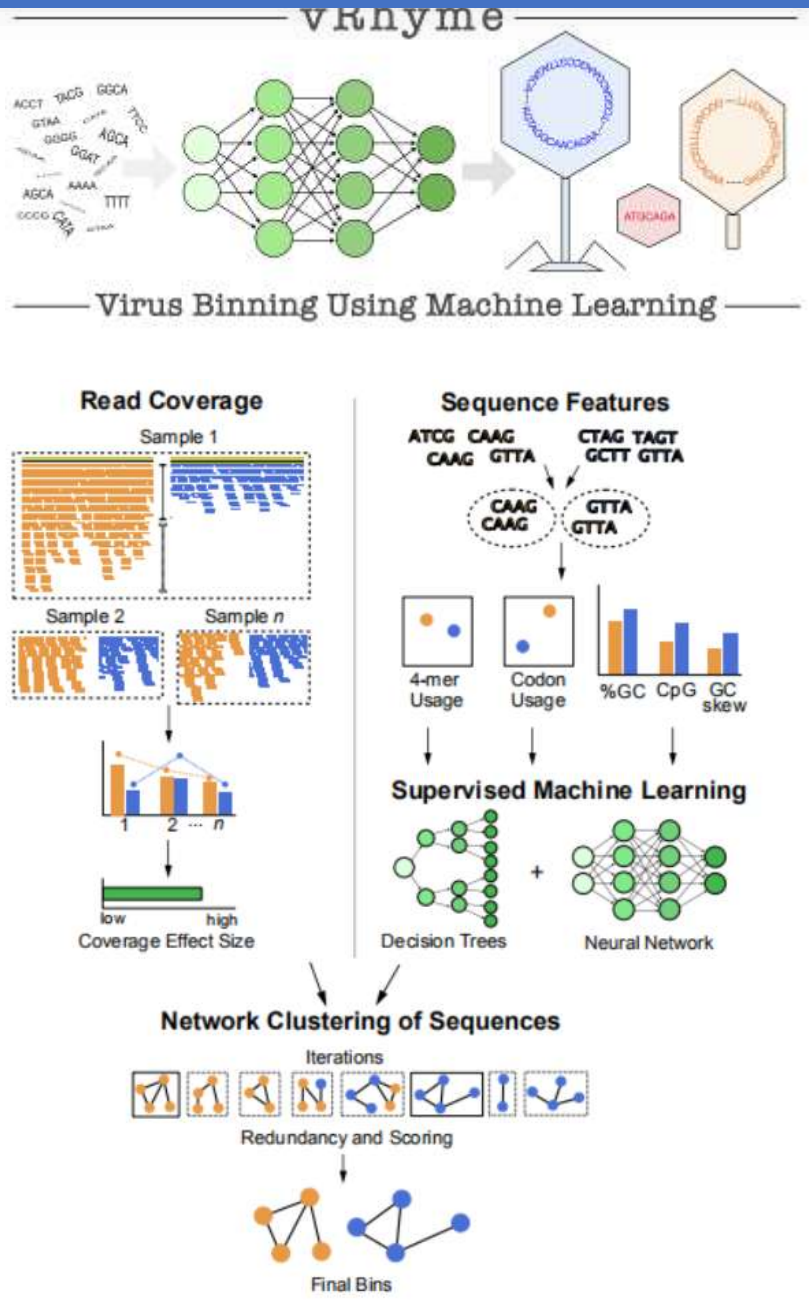
Metric	Description	Default	Options	Activity Metric
Mann-Whitney U Test	Compare average prophage and host coverages; *n* subsamples of 100 base-pairs each; yield p-value (p)	p <= 0.05 n = 5	p = any n >= 5	yes
Cohen's *d*	Effect size (*d*) of coverage difference between prophage and host (<0.6: low, 0.75: medium, >1: high)	d >= 0.75	d > 0.6	yes
Prophage:host coverage ratio	Ratio (*r*) of prophage to host average coverage; approximate prophage to host genome copy ratio	r >= 1.75	r > 1.5	yes
Alignment gaps	Remove reads with greater than *g* gap extensions	g <= 1	g = 0-3, off	no
Alignment mismatches	Remove reads with greater than *m* base mismatches	m <= 3	m = 0-10, off	no
Coverage outliers	Remove outlier coverages greater than *s* standard deviations greater than the average coverage	s = 4	s = 0-4	no
Mask scaffold ends	Mask coverage values of *x* nucleotides at both ends of the input scaffold, where *x* is the input read length	x = read length	none	no

C



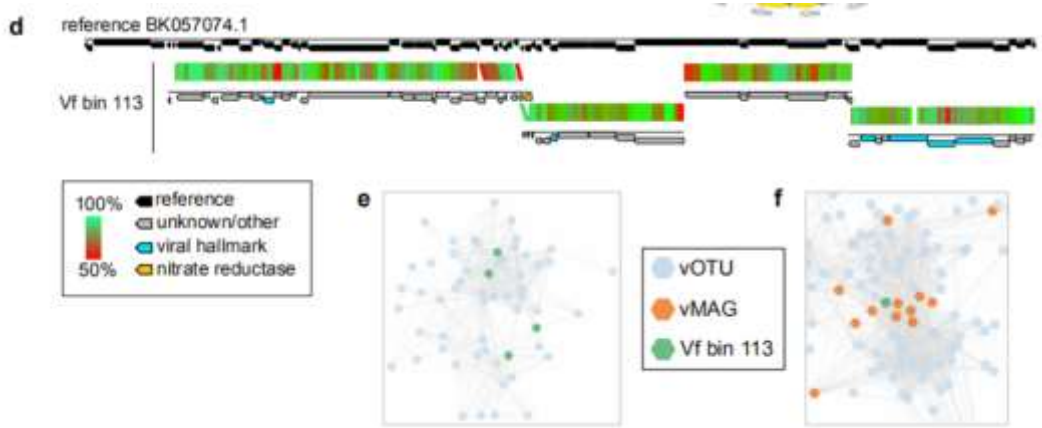
Prophage 1		Prophage 2	
Status:	active	Status:	active
Avg. coverage:	60x	Avg. coverage:	169x
Cohen's *d*:	1.65	Cohen's *d*:	3.70
Coverage ratio:	2.17	Coverage ratio:	6.07

工具4 vRhyme：病毒基因组分箱工具



Advantages of vRhyme compared to other software for prokaryote binning

- **vRhyme** yielded the highest, or equal to highest, F_1 score, average precision, accuracy, and specificity, comparing to MetaBat2, VAMB, CONCOCT, and BinSanity
- CONCOCT and BinSanity yielded the greatest average recall values but at the expense of precision.
- VAMB requires many sequences as input for optimal performance
- CONCOCT and BinSanity have biased to over-bin distinct genomes into a single bin.
- **vRhyme** can better distinguish the source scaffolds as complete genomes, overcomplete MetaBat2



Kieft K et al. ***Nucleic Acids Research*** 2022

工具5 ViWrap：一步式病毒基因组分析工具软件包

VIBRANT

通过混合机器学习和蛋白质相似性方法进行病毒鉴定

Anantharaman实验室, 威斯康星大学麦迪逊分校, 2020年

jiarong/**VirSorter2**

customizable pipeline to identify viral sequences from (meta)genomic data

一个多分类器、由专业指导的识别器

Sullivan & Roux 实验室，俄勒冈州立大学和美国能源部基因组科学研究所，2021年

jessieren/**DeepVirFinder**

Identifying viruses from metagenomic data by deep learning

基于kmer的卷积神经网络算法识别器

Sun实验室，南加州大学，2020年

用于病毒鉴定的工具

NCBI RefSeq Virus genomes

(NCBI)

VOG HMMs

(http://vogdb.org)

IMG/VR V4 具有分类分配的高质量虚拟操作性单元（vOTUs）

(美国能源部基因组科学研究所)

病毒分类学分类的数据库

vRhyme

Virus Binning Using Machine Learning

基于scaffolds覆盖效果大小和核苷酸特征的病毒分箱

Anantharaman实验室, 威斯康星大学麦迪逊分校, 2022年

病毒分箱

vConTACT2

构建基于整个基因组的基因共享网络，用于基于距离的分层聚类和分类预测

Sullivan实验室，俄勒冈州立大学，2019年

MrOlm/**drep**

Rapid comparison and dereplication of genomes

基于序列相似性对微生物/病毒基因组进行聚类和去冗余处理

Banfield实验室，加州大学伯克利分校，2017年

病毒分组

CheckV

检查病毒基因组的质量和完整性

Banfield实验室, 加州大学伯克利分校, 2017年

质量检查

iPHoP

一种集成的机器学习框架用于宿主预测

Roux实验室，美国能源部基因组科学研究所，2022年

宿主预测

缺乏一个集成的流程/封装器

ViWrap

软件流程

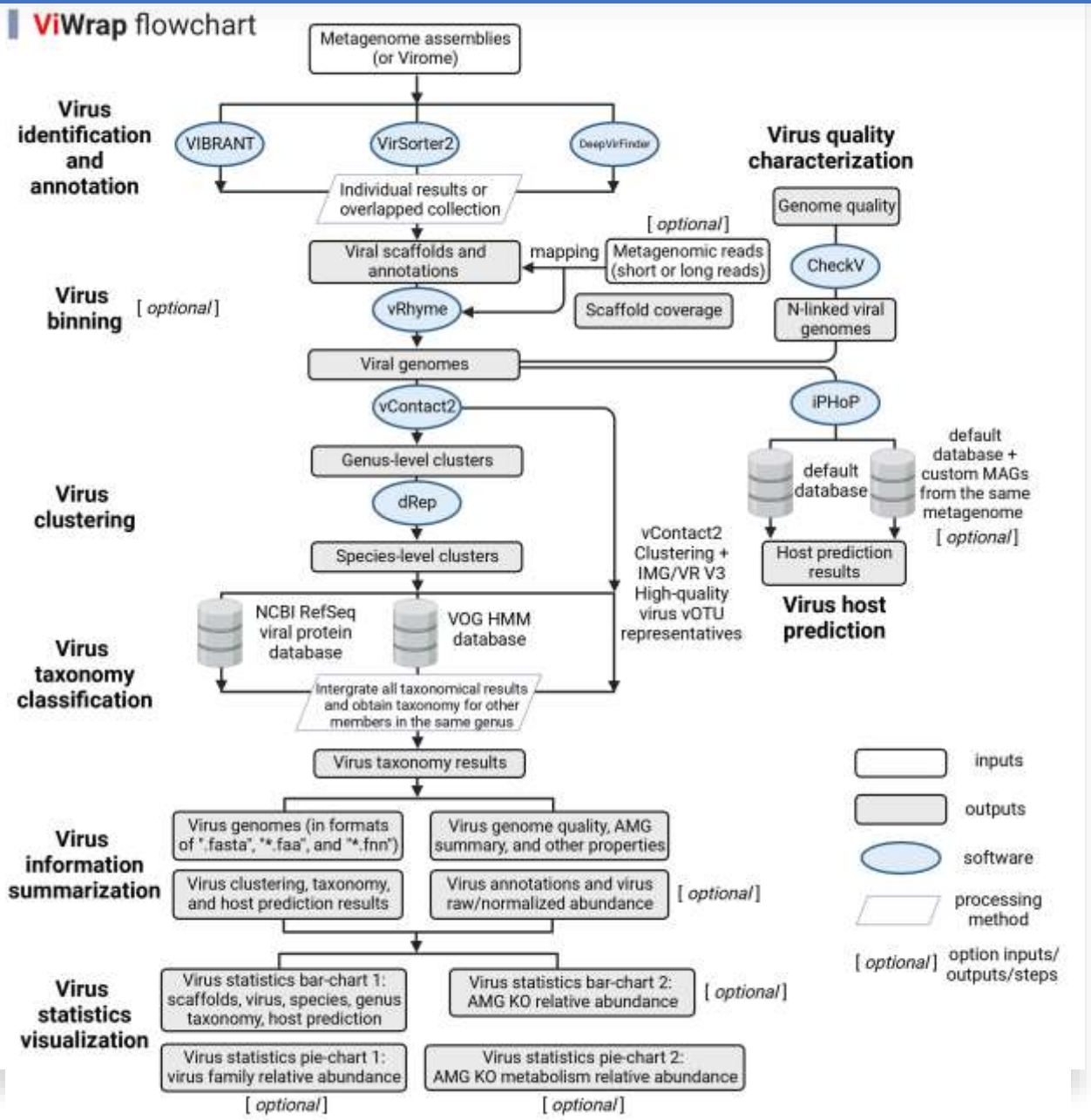
第一步： 三个病毒识别器

第二步： 宏基因组reads mapping、病毒分箱和质量检查

第三步： 将病毒聚类成属和种，并分配分类学信息

第四步： 使用iPHoP预测病毒的宿主

第五步： 总结结果并可视化病毒统计信息



工具5 ViWrap：一步式病毒基因组分析工具软件包

All result folders

组织好的中间文件夹

- `00_VIBRANT_input_metageome_stem_name` : the virus identification result (would be "00_VirSorter_input_metageome_stem_name", "00_DeepVirFinder_input_metageome_stem_name", "00_VIBRANT_VirSorter_input_metageome_stem_name", or "00_VIBRANT_VirSorter_DeepVirFinder_input_metageome_stem_name")
- `01_Mapping_result_outdir` : the reads mapping result
- `02_vRhyme_outdir` : vRhyme binning result
- `03_vConTACT2_outdir` : vConTACT2 classifying result
- `04_Nlinked_viral_gn_dir` : N-linked viral genome as CheckV inputs
- `05_CheckV_outdir` : CheckV result
- `06_dRep_outdir` : dRep clustering result
- `07_iPHoP_outdir` : iPHoP result for host prediction
- `08_ViWrap_summary_outdir` : Summarized results
- `09_Virus_statistics_visualization` : Visualized statistics of viruses
- `ViWrap_run.log` : running log file containing the issued command and time log

ViWrap总结文件夹里面的结果

```
> 08_ViWrap_summary_outdir
    - Genus_cluster_info.txt # Virus genus clusters
    - Species_cluster_info.txt # Virus species clusters
    - Host_prediction_to_genome_m90.csv # Host prediction result at genome level
    - Host_prediction_to_genus_m90.csv # Host prediction result at genus level
    - Sample2read_info.txt # Reads counts and bases
    - Tax_classification_result.txt # Virus taxonomy result
    - Virus_annotation_results.txt # Virus annotation result
    > Virus_genomes_files # Contains all fasta, faa, ffn files for virus genomes
        - vRhyme*.fasta
        - vRhyme*.faa
        - vRhyme*.ffn
    > Virus_normalized_abundance.txt # Normalized virus genome abundance (normalized by 100M reads/sample)
    > Virus_raw_abundance.txt # Raw virus genome abundance
    > Virus_summary_info.txt # Summarized property for all virus genomes
```

结果呈现

可视化文件夹中的病毒统计可视化结果

```
> 09_Virus_statistics_visualization
    > Result_visualization_inputs
        - virus_statistics.txt
        - virus_family_relative_abundance.txt
        - KO_ID_relative_abundance.txt
        - KO_metabolism_relative_abundance.txt
    > Result_visualization_outputs
        - virus_statistics.png # the 1st bar-chart
        - virus_family_relative_abundance.png # the 1st pie-chart
        - KO_ID_relative_abundance.png # the 2nd bar-chart
        - KO_metabolism_relative_abundance.png # the 2nd pie-chart
        - virus_statistics.pdf
        - virus_family_relative_abundance.pdf
        - KO_ID_relative_abundance.pdf
        - KO_metabolism_relative_abundance.pdf
```

总结要点

- ViWrap整合了目前可用的工具和数据库，用于全面且严格的病毒筛查
- 它对于鉴定方法、宏基因组reads和自定义微生物基因组的选择具有灵活性，适用于各种应用场景
- 它具有一站式、用户友好的工作流程，并生成易于阅读和解析的结果

ViWrap使用

- ViWrap可以用于各种环境设置，包括自然环境、人为环境和与人类微生物组相关的环境。
- ViWrap可以通过GitHub（https://github.com/AnantharamanLab/ViWrap）公开获取。软件的使用方法和结果解释的详细描述可以在该网站上找到。

nature microbiology

Perspective

https://doi.org/10.1038/s41564-025-02095-4

A call for caution in the biological interpretation of viral auxiliary metabolic genes

Published online: 27 August 2025

1.1 Author Introduction

Karthik Anantharaman

/ˈkɑːrθɪk/ /əˈnʌntərəˌmən/

Web of Science Core Collection metrics

40 H-Index	91 Publications
8,488 Sum of Times Cited	6,526 Citing Articles
8,275 Sum of Times Cited without self-citations	6,460 Citing Articles without self-citations
169 Sum of Times Cited by Patents	156 Citing Patents

- 威斯康星大学麦迪逊分校细菌学系副教授
- 研究方向：微生物与病毒生态学，以及其在生物地球化学循环中的作用
- 多项重要荣誉：NSF CAREER 奖、NIH MIRA 奖、ASM 早期职业环境微生物学研究奖，2023年入选科睿唯安**全球高被引科学家**名单
- 开发了许多著名软件用以表征病毒特征、研究微生物代谢和微生物组的相互作用：VIBRANT、vRhyme、PropagAtE、**ViWrap**、vClassifier、Protein Set Transformer、METABOLIC

★ 什么是病毒辅助代谢基因（Virus-encoded auxiliary metabolic genes，AMGs）？

viral genes, derived from the host genome

病毒编码的非必需基因

augment or redirect host metabolism for the ultimate benefit of the virus during infection

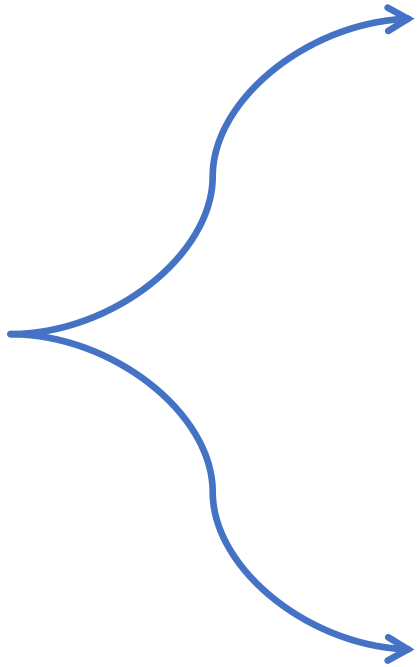
操纵宿主代谢（增强或重定向）以利于病毒感染

a form of virus–host co-evolution

病毒-宿主共进化的产物

泛指病毒中任何具有宿主类似功能且可能提供优势的基因

★ AMGs 的两条进化路径及其研究意义



Direct transfer of host genes （直接传递 – 进化和功能上）

- the virus uses the acquired gene in the same way as the host during infection
- require relatively less evidence（因为功能明确且与宿主同源）

e.g. the first AMGs reported in cyanophages were homologues of the host proteins that formed photosynthetic reaction centres

Exaptation（扩展适应）

- Viral homolog function in different biological and/or biochemical capacities
- may retain the same biochemical function of the host protein, but could be tailored to the unique needs of viral development during infection
- the practical limitations of protein functional annotation
- experimental data are needed to resolve the competing hypotheses

★ 当前 AMGs 研究中的主要问题

过度依赖自动化功能预测流程

↓

大量常见病毒基因被**误判**和**误报**为 AMGs

↓

夸大病毒对生态过程的直接影响力

the caution needed

★ 宿主基因组污染始终存在

☒ 宿主基因天生就和代谢功能相关，更容易被注释为AMG，造成假阳性。

✓ 解决方案：

- 只使用高质量、高完整度的病毒基因组。
- 真正的病毒AMG应该被典型的病毒基因（如衣壳蛋白、尾蛋白）所包围，而不是孤零零地出现在序列片段末端。

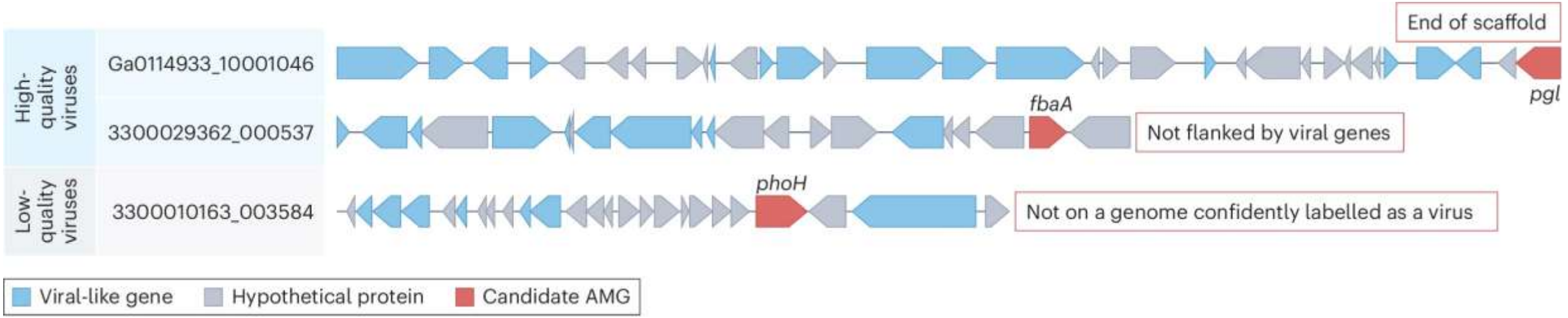


Fig. 1: Common limitations when studying viral AMGs.

★ AMG 识别的半自动化工具

community guidelines

high-confidence viral identification

AMG annotation and validation

downstream ecological, experimental and phylogenetic investigations

但没有“一键式”工具能完美完成

VIBRANT：

功能全面，但缺乏基因组上下文过滤，可能产生较多假阳性。

DRAM-v：

更先进，会检查基因两侧环境（上下文）并过滤掉常见的、可能与代谢无关的病毒基因（如肽酶）。

users manually validate the results

一项研究表明，通过手动排除那些更可能用于病毒必需过程（如核苷酸代谢）或功能过于通用的基因，可以将 VIBRANT 和 DRAM-v 报告的 AMGs 数量分别减少 33% 和 44%。

★ 注释病毒蛋白功能十分困难

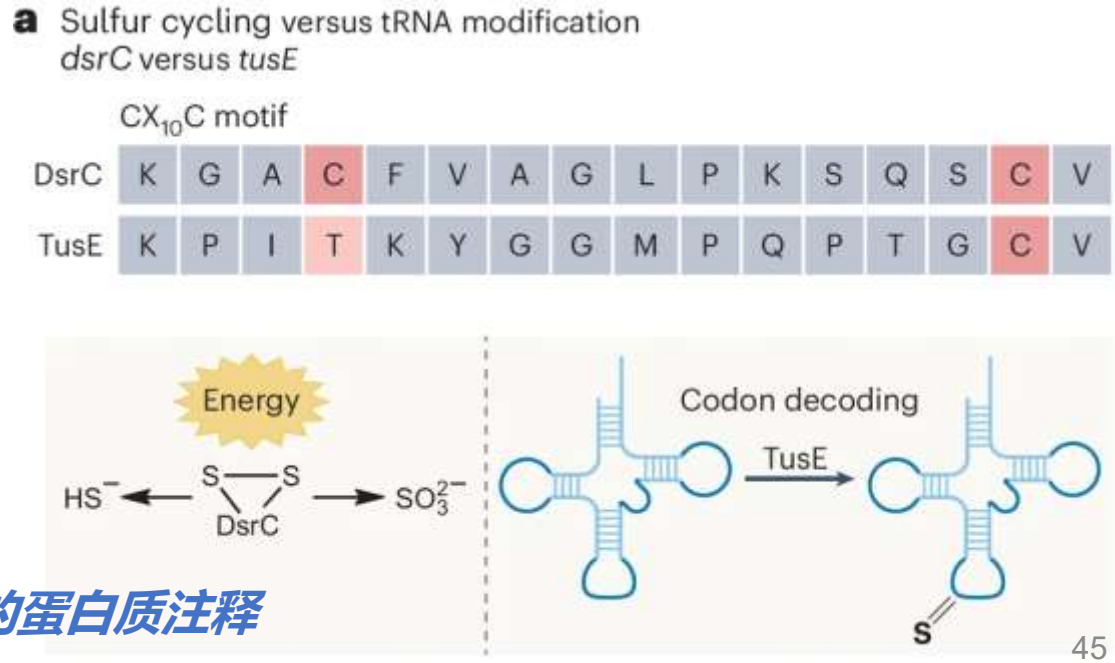
- 标准方法：对比数据库中已知代表序列的相似性，来注释蛋白质功能。
- 三大难题：
  - 病毒蛋白质遗传多样性极高，导致新蛋白的生化功能难预测、易误判。病毒基因的扩展适应会给注释带来难度。
  - 病毒蛋白质中蛋白质结构域的大量镶嵌重排。仅凭一个结构域（如果胶裂解酶域）就判断整个蛋白功能（降解植物有机物）会出错，因为该域也可能用于病毒识别和进入宿主。
  - 通过 HMMs（概率隐马尔可夫模型）检测到的远缘同源性很常见，但进化时间长会导致功能分化，序列相似的蛋白可能功能完全不同。

e.g. DsrC 在参与异化硫氧化或亚硫酸盐还原的不同蛋白质之间传递硫原子

TusE 则参与用于 tRNA 修饰和氧化还原平衡的硫传递系统。

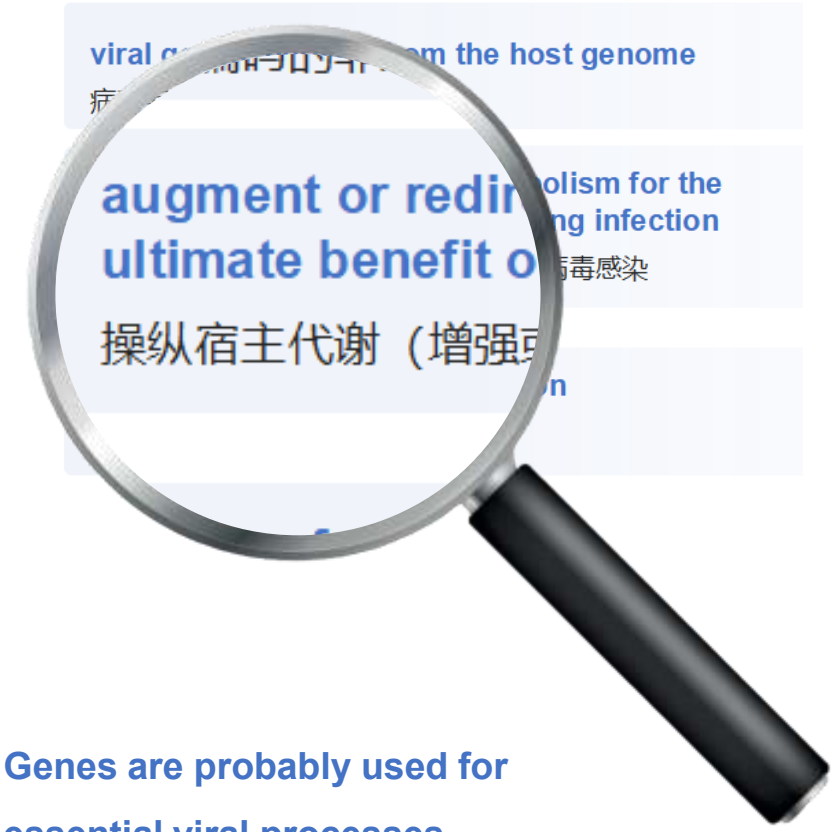
与 TusE 不同，DsrC 的 C 端有两个保守半胱氨酸残基，这对其在硫转化过程中与硫的相互作用至关重要。

然而，目前的 HMMs 无法区分。



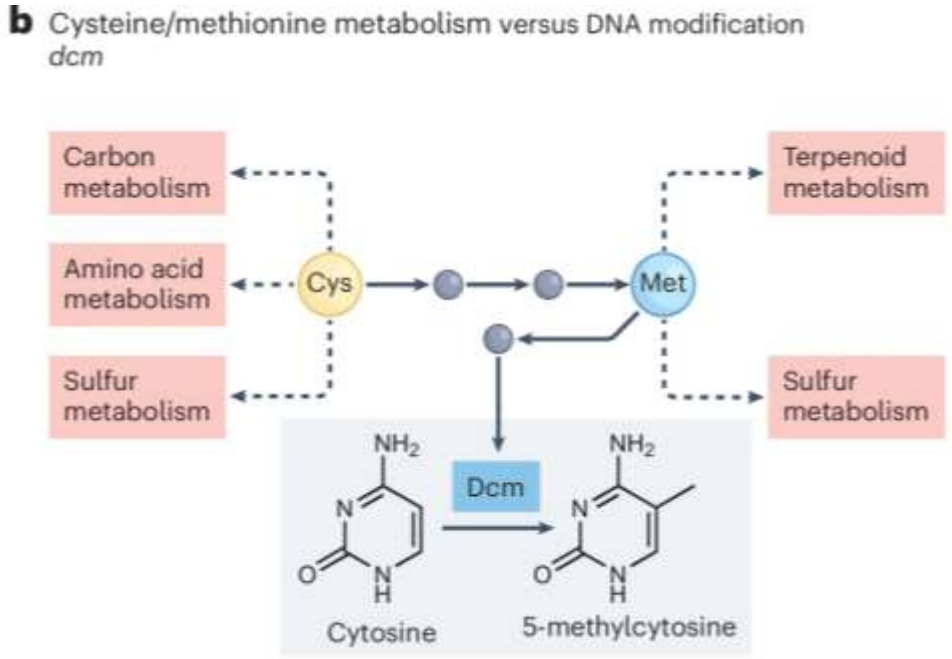
需要进行全面的蛋白质注释

★ 病毒中的常见代谢基因可能并非AMGs



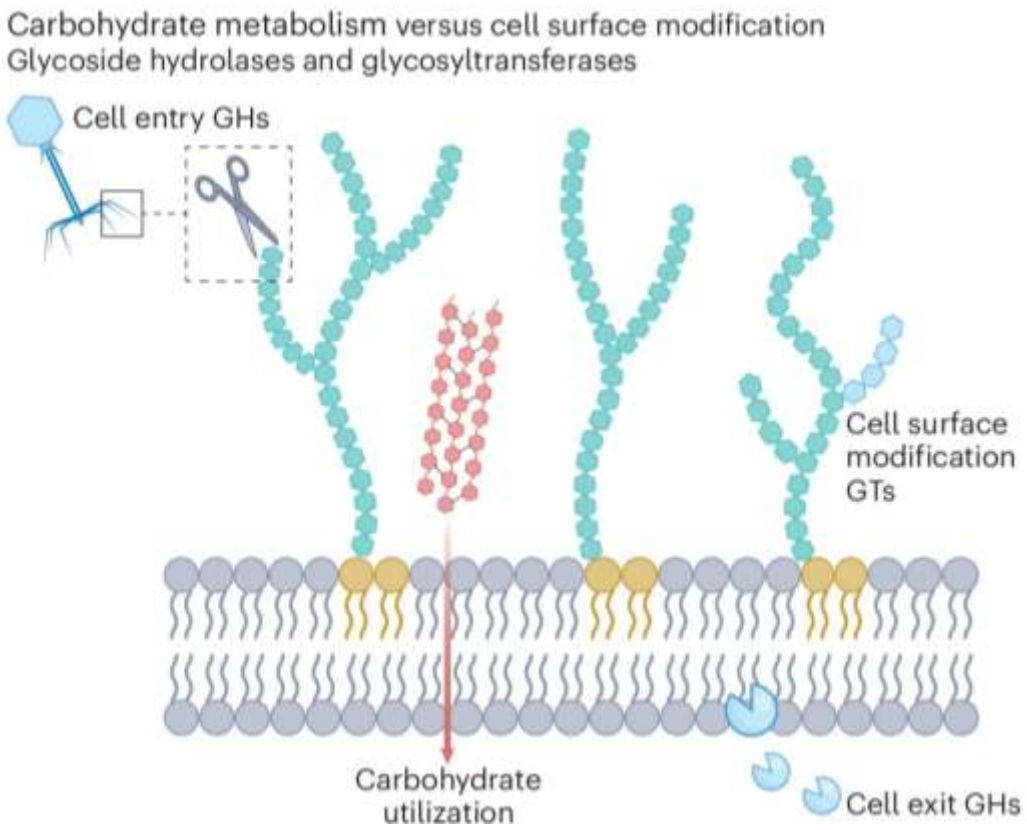
Example 1：*dcm* and *queC*

- 这两个基因被 VIBRANT、DRAM-v 等工具误判为 AMG。
- *dcm* 是 DNA 胞嘧啶甲基转移酶，用于修饰病毒 DNA 以逃避宿主防御，而非参与宿主代谢；
- *queC* 参与 tRNA 修饰或病毒染色体修饰，均为病毒自身必需功能（也包括其他的*queDEF*）。



★ 病毒中的常见代谢基因可能并非AMGs

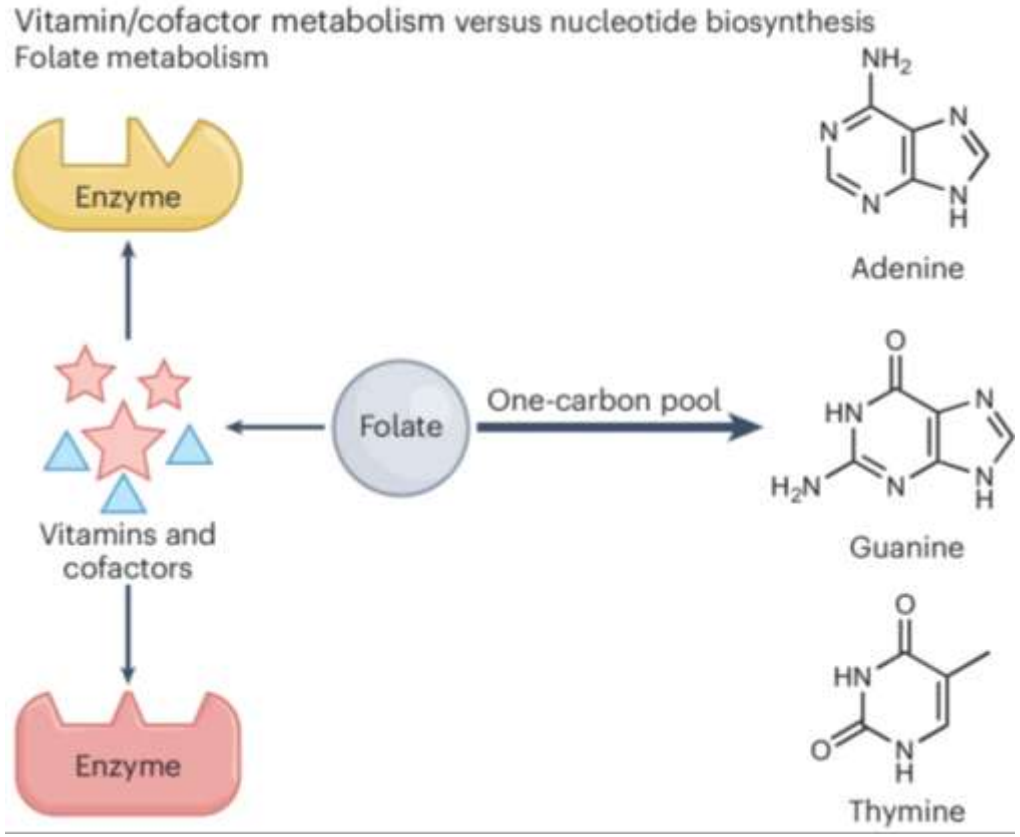
Example 2：glycoside hydrolases (GHs)



- 糖苷水解酶（GHs）虽能断裂糖苷键，但病毒中的 GHs 多用于自身侵染相关过程，而非为宿主获取营养。如：
  - 降解宿主细胞表面糖苷键实现入侵 cell entry
  - 破坏生物膜接触宿主 carbohydrate utilization
  - 维持自身结构稳定 cell surface modification
  - 辅助病毒释放 cell exit
- 其功能需结合宿主细胞表面生化特征、生态系统分析验证，不能仅凭 CAZyme 数据库注释判定为 AMG。

★ 病毒中的常见代谢基因可能并非AMGs

Example 3：folate biosynthesis genes 叶酸代谢基因



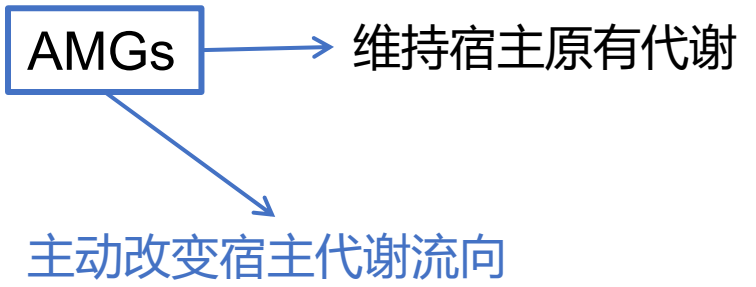
虽能辅助维生素合成，但更可能用于提供核苷酸合成所需的一碳单位，这是病毒基因组复制的核心需求。

观点：

在描述推定的 AMGs 时应采取保守态度，以避免夸大基于 AMGs 的结论。

鼓励对推定 AMGs 可能发挥的更广泛作用进行推测，但需明确指出哪些证据支持这些推测。

★ 病毒也可以重编程宿主细胞代谢



- 海洋病毒的普遍特征：碳代谢通量在关键分支点被转移（而非增强线性途径），实现中心碳代谢重编程。
  - 蓝藻噬菌体的 *cp12* 基因：抑制宿主卡尔文循环，将碳代谢流转向磷酸戊糖途径（PPP），为病毒自身的核糖和核苷酸合成供能。
- 病毒对宿主代谢的彻底重塑
  - 如病毒谷氧还蛋白高表达时，宿主的硫还原和乙醛酸循环基因上调，以支持病毒增殖。
- 病毒和 AMGs 的影响还涉及其他细胞过程（如细胞生理、基因调控）
  - 如感染产芽孢细菌的病毒携带孢子形成转录基因，可能影响宿主的孢子形成过程。

3. AMG predictions must be supported by robust evidence

★ 计算预测证据

silico-only evidence is appropriate only for AMGs whose biological roles are unambiguous

以 *rdsr* 和 *amoC* 为例：

- 基因组上下文：**gene synteny**，基因两侧是典型的病毒基因，确保其是病毒基因组的一部分，而非污染。
- 关键活性位点：**multiple sequence alignments**，证明病毒编码的蛋白含有执行该功能所必需的关键氨基酸残基。
- 系统发育分析：**clustering**
  - *rdsr* 案例：病毒基因与已知的硫氧化细菌（如SUP05）的基因聚在一起，佐证了其宿主来源和功能一致性。
  - *amoC* 案例：病毒基因与宿主基因在进化树上明显分开，这反而有利于在环境中区分病毒和宿主对代谢的贡献。

3. AMG predictions must be supported by robust evidence

★ 环境关联证据

in situ field and mesocosm experiments can be used as complementary evidence

- 核心逻辑：通过原位 / 中型生态系统实验，观察环境变化与 AMG 丰度的关联，间接支撑功能。

海带降解实验

在含有海带（富含laminarin多糖；海带多糖 ）的生态缸中，携带褐藻酸酶（laminarinase GHs）的病毒显著富集。这强烈暗示这些病毒AMG参与了海带多糖的降解利用。

尿素添加实验

在土壤中添加尿素（氨氧化古菌（AOA）的底物）后，携带多铜氧化酶基因（multicopper oxidase type 1 ， MCO1）的病毒随之富集，暗示该AMG在氨氧化过程中起作用。

3. AMG predictions must be supported by robust evidence

★ 蛋白水平实验验证

Experimentally dissecting both the biochemical activity and biological role of putative AMGs is required to confirm the auxiliary role of any gene

01

验证标准

需满足 “提升病毒适应性，但不直接满足病毒基本需求”即直接证明该基因的产物具有预期的酶活性能，并且其表达能给病毒复制带来可衡量的好处。

02

典型案例

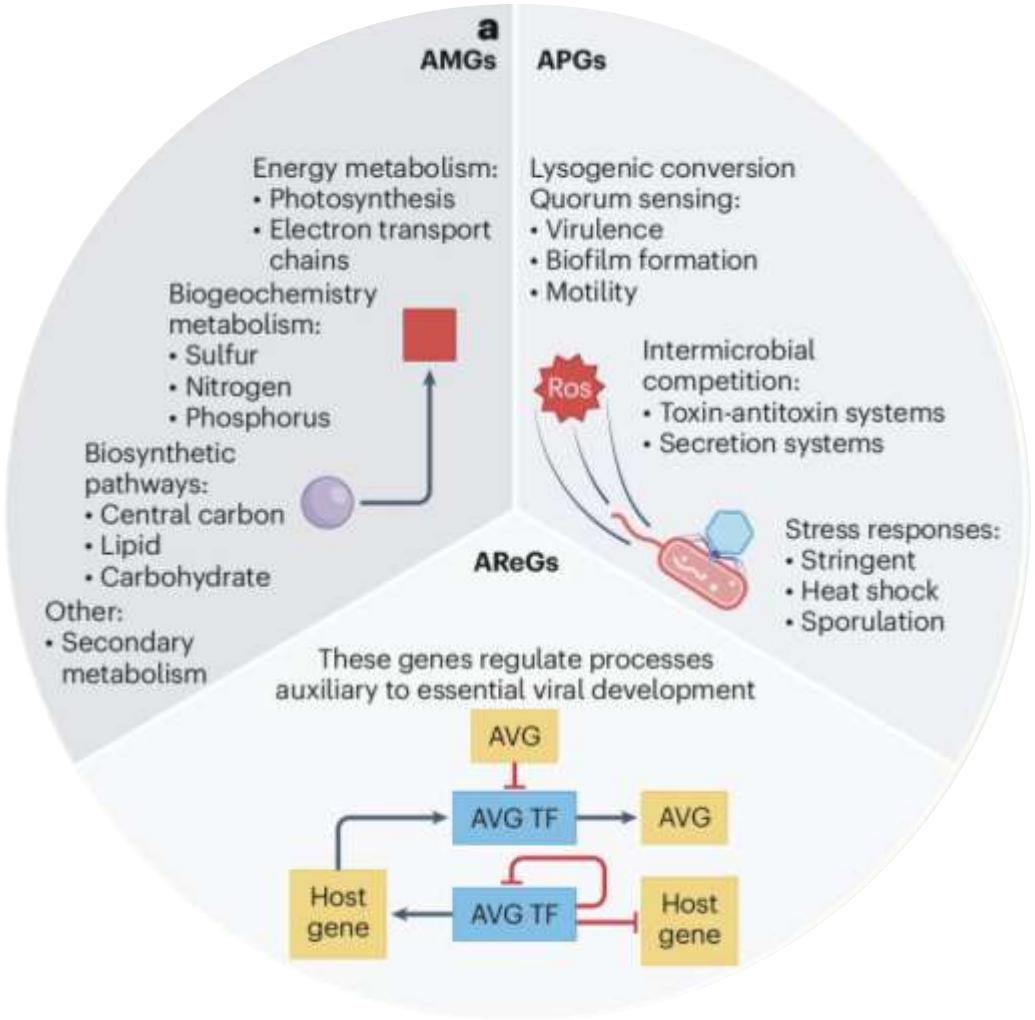
蓝藻噬菌体 *psbA*/*psbD*：高光下促进病毒复制，补偿宿主光合基因表达下降，明该 AMG 直接增强了病毒在特定环境下的适应度。

03

现存不足

多数 AMG（如视紫红质、氮 / 磷获取基因）仅证实生化功能，但其对 “病毒适应性的直接提升作用” 仍未明确验证。

★ 提出“辅助病毒基因（auxiliary viral genes, AVGs）”



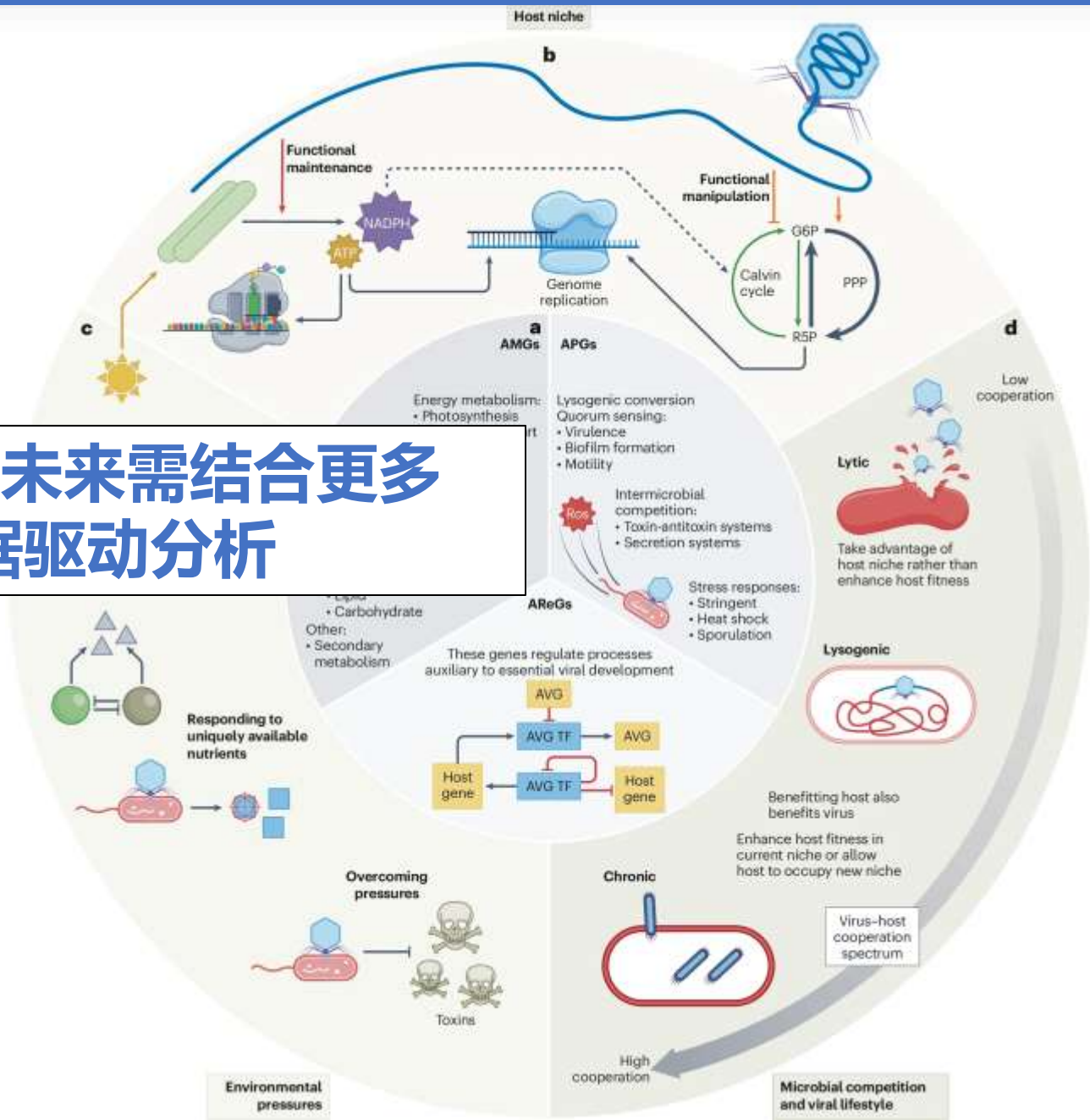
- 以包容性视角涵盖所有辅助核心病毒功能的基因
- 鼓励整体性的 AVG 研究
- 不仅关注 AVGs 在群落层面的影响，也关注特定的病毒 - 宿主互作

AMGs are one type of AVG that focus on **metabolism**, but other functional categories such as **physiology** (auxiliary physiology genes, APGs) and **gene regulation** (auxiliary regulatory genes, AReGs) are potential candidates for AVGs

★ AVG 进化的三大驱动力量

- **宿主生态位**：病毒通过 AVG 适应宿主生态位，分为 “功能维持型”（如维持宿主光合）和 “功能操纵型”（如重编程宿主碳代谢）。
- **环境压力**：筛选出应对压力（如降解农药）利用特定营养的 AVGs 是辅助功能。
- **微生物竞争与病毒生活方式**：
  - 烈性病毒：AVGs 多利于病毒工厂，可能间接惠及群落；
  - 温和病毒：可能携带利于宿主的基因（如毒力基因），助力宿主竞争，间接利于自身；
  - 慢性病毒：更可能携带利于宿主的 AVGs。

并非独立作用，未来需结合更多 AVG 案例，数据驱动分析



substantial AVG analyses should not be required or expected in viral ecological studies

viral fitness analyses alone do not confirm the auxiliary role of the gene

the caution needed

AVGs must not be used for essential viral functions

acknowledge the technical difficulties associated with experimentally verifying claims made about AVGs

5. Future perspectives

01

building a rigorous and biologically plausible story that clearly indicates when claims require further evidence is still important

构建严谨且生物学合理的研究脉络（明确指出何时结论需要进一步证据支持）仍至关重要

02

encourage holistic investigations of viral datasets that may or may not include AVG analyses

鼓励对病毒数据集进行全面研究（无论是否包含 AVG 分析）

谢谢大家！