**环境宏病毒组学分析思路与常用工具**

**摘要：**

病毒是地球上丰富度最高的生物个体，几乎能够侵染任何形式的生命体，在全球生态系统和食物链中扮演着不可或缺的角色。其中噬菌体（侵染细菌的病毒）可以影响宿主代谢系统和微生物群落结构，并且能够通过介导基因转移从宿主处获得某些基因。例如，噬菌体编码的辅助代谢基因对地球化学循环的影响以及噬菌体编码的抗生素抗性基因对耐药性传播的影响近年来受到了研究者的高度关注。现今，随着高通量测序技术的发展，病毒组得到了越来越多的研究。与细菌类似，病毒的基因组信息同样能够通过组装技术从宏基因组测序数据集中获取，通过对病毒长序列的识别以及进一步对病毒基因的预测和功能注释，可以评估病毒生物多样性和群落结构，探究病毒-宿主关系以及病毒对微生物群落及生物地球化学循环可能造成的影响。本文主要基于已开展的代表性地球病毒组研究，探讨环境宏病毒组的分析思路，然后介绍目前两种主流的病毒重叠群的识别工具VirSorter和VirFinder的使用方法。

**环境宏病毒组学分析思路：**

**通过以下**环境宏病毒组分析的常规流程（如图1所示），可获取病毒组序列以基因的定性与定量信息。

1）环境**样品采集：**首先，研究者需根据研究目的和环境样品类型进行针对性的宏病毒组采集。针对水样（海水、湖水、河水、污水等）、土壤及、沉积物样本的病毒组采集方法可参考？？ , ？？和？？【待补充文献】。

2）**病毒核酸提取**研究者可以根据具体的研究对象来决定提取的核酸类型，即：提取DNA（针对DNA病毒）或提取RNA（正在活跃侵染宿主的DNA、RNA病毒）。目前针对病毒组的核酸提取有两种策略：

i）首先提取微生物群总DNA或总RNA，然后在后续数据处理过程中再去除大量属于细胞生物的序列及其他污染序列1-3。这种策略没有复杂的核酸提取过程，但在后续数据处理过程中可能会丢失部分低丰度的病毒信号。值得注意的是，该策略产生的海量宏基因组和宏转录组测序数据及其重分析是挖掘宏病毒组信息的合适选择；

ii) 富集病毒颗粒后再进行DNA或（和）RNA的提取，在后续数据处理过程中需要去除的属于细胞生物的序列及其他污染序列将会大大减少4-7。这种策略需首先经过较为复杂的且可能存在一定偏差的病毒富集过程，但得到的测序数据集已经将病毒信号放大，对识别病毒序列更加有利。

**3）扩增建库与高通量测序**：通过病毒核酸样品的建库与高通量测序得到原始序列（raw reads）。通常从环境样品提取的病毒核酸总量低（~ng级别），需要对核酸样本进行全基因组扩增。常见的扩增建库策略包括【待补充内容】

**4）质量控制：**对原始序列进行质量控制（quality control）去除测序接头和低质量序列后得到干净序列（clean reads）。

**5）组装：**将干净序列组装得到长重叠群（contigs），在众多的组装软件中，metaSPAdes， MEGAHIT和IDBA-UD被证明具有最好的病毒组组装效果8。

**6）识别病毒重叠群：**无论使用哪种采样策略，得到的数据集中均包含大量不属于病毒的信号，因此研究者需要识别来自于病毒的信号。目前最常用的识别病毒重叠群（viral contigs）的方法为使用VirSorter软件，同时可结合其他软件的使用（如VirFinder）或直接比对到现有的数据库做辅助验证。上述提及的两款软件使用方法会在后文详细介绍。

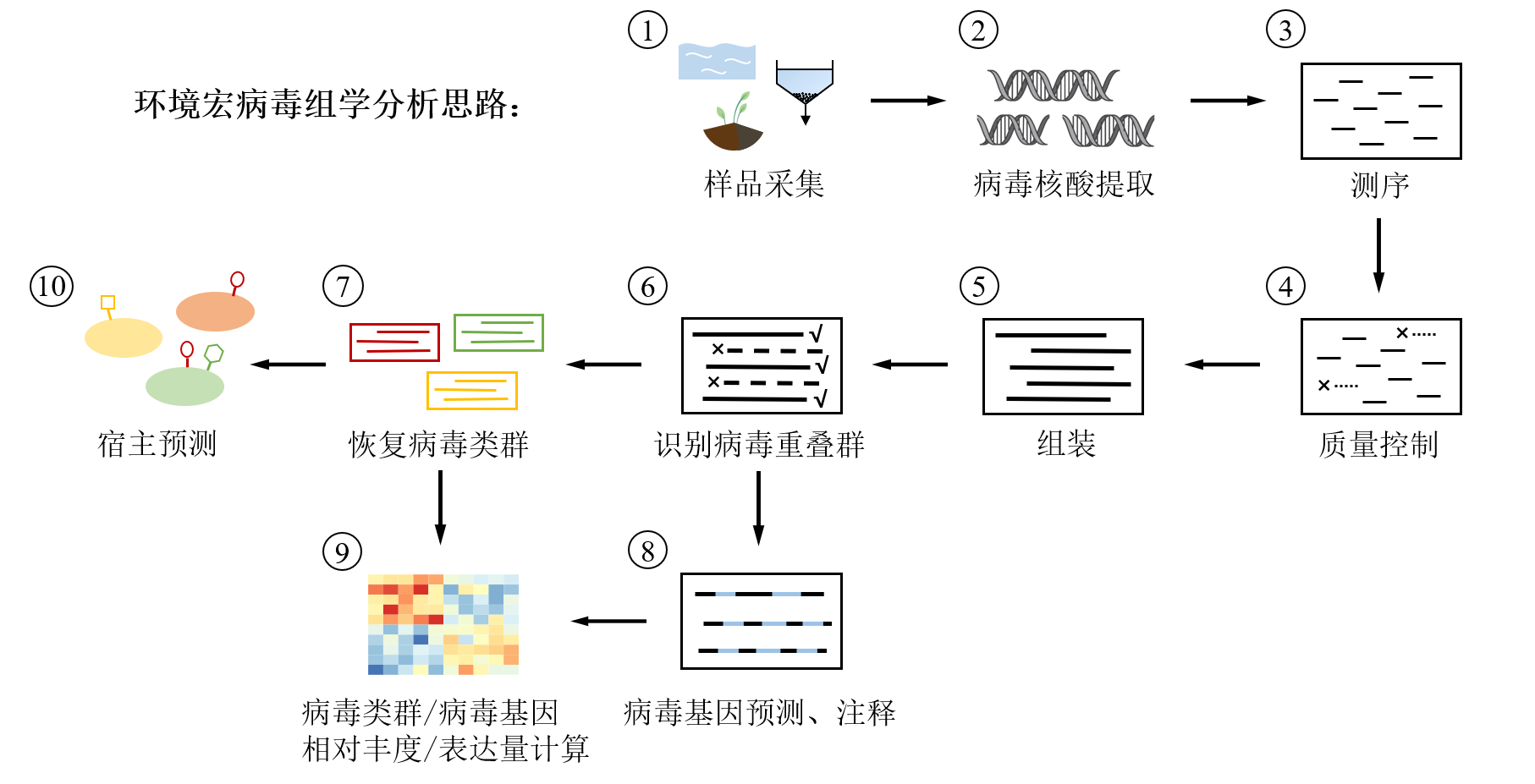


图1 环境宏病毒组分析常规流程示意图

**7）恢复病毒类群：**与细菌不同，病毒没有一定数量普遍存在的单拷贝标记基因（universal single-copy marker genes），这使得属于病毒的基因组（viral genome bins）的恢复更加困难，很难客观评估其完整度和污染度。目前部分病毒组的研究选择跳过本步骤【待插入文献】，仅是基于病毒重叠群的水平进行分析。但也有不少研究通过以下方法更进一步在病毒类群（viral populations）的水平上进行分析【待插入文献】：

i）根据序列长度和核苷酸同一性作为参数进行聚类得到病毒类群，例如在80%序列长度具有95%核苷酸同一性的病毒重叠群将被聚类为同一病毒类群4,6，尽管这仍然会高估病毒的多样性；

ii）基于病毒重叠群在样本中的覆盖度（coverage）和四核苷酸频率（tetranucleotide frequencies）进行分箱获得病毒类群5；iii）对病毒重叠群进行延长、拼接、环化并手动修正组装偏差得到某些在样本中丰度较高的病毒完整基因组1。

**8）****病毒基因预测和注释：**为了得到恢复的病毒序列的物种分类和功能信息，先通过常规方法（如：使用Prodigal）对病毒重叠群中的基因进行预测，然后将预测得到的基因比对到数据库进行注释。常用的数据库均储存了来自病毒的基因信息，如：NCBI RefSeq、GenBank、PFAM、KEGG、UniProt等。由于目前数据库中属于病毒的注释信息还远远不足，这使得研究者即使使用非常宽松的比对参数，仍然会得到大量未知功能的病毒基因。

**9）计算病毒类群/病毒基因的相对丰度/表达量：**通过将干净读序映射（mapping）回病毒重叠群或病毒基因，计算其在样本中的相对丰度（如为转录组数据则为表达量），可以得到病毒的群落结构、病毒基因的表达水平等重要信息。

**10）宿主预测：**病毒组研究的另一重要方面即为病毒的宿主预测。然而目前没有一种全面、准确的方法能够完成宿主预测，一般会选择用几种方法结合使用来进行宿主预测。目前较为常用的病毒宿主预测方法、原理和举例文献列于表1。

表1 目前常用病毒-宿主预测方法

|  |  |  |
| --- | --- | --- |
| 宿主预测方法 | 原理 | 举例 |
| 在可能的宿主基因组中寻找CRISPR spacer，检查恢复的病毒类群中是否含有一致的序列。 | 宿主的CRISPR系统反应了被侵染的历史。 | 1,3,5,9 |
| 如在病毒基因组中找到编码tRNA的序列，可根据tRNA的宿主判断该病毒的宿主。 | 病毒（噬菌体）具有基因转移的能力，可以从宿主处获得某些基因。通过对这些基因的溯源或系统发育分析，可以得到部分病毒的宿主信息。 | 1,3,9 |
| 如在病毒基因组中找到编码辅助代谢基因（AMG）的序列，可根据AMG的系统发育判断该病毒的宿主。 | 1,5 |
| 病毒基因组片段与可能的宿主基因组片段的相似性。 | 5,9 |
| 病毒类群和可能的宿主在样本中的丰度、表达水平的关系。 | 病毒与宿主在样本中的丰度存在共出现的关系，此外，由于来自病毒的转录本一定起源于宿主细胞内，因此病毒和宿主的基因表达存在正相关。 | 1,10 |
| 根据病毒与可能的宿主之间相近的四核苷酸频率预测宿主。 | 经验主义认为宿主和病毒的基因组有着相近的四核苷酸频率。 | 5 |

**VirSorter2使用教程**

原网站：<https://github.com/jiarong/VirSorter2>

**简介：**

VirSorter是一款能够在基因组数据集中识别病毒信号的软件，是目前病毒组研究中使用最广泛的一款病毒长序列(viral contigs)识别软件，已被引用四百余次。

该软件在2020年新推出了VirSorter2新版本，本教程将基于VirSorter2新版本进行介绍。

**安装：**

Option1:

conda create -n vs2 -c bioconda virsorter=2

#conda创建环境vs2并将VirSorter2安装进该环境

conda activate vs2

#激活vs2环境

Option2:

安装最新更新开发版本，官方推荐。

conda create -n vs2 -c bioconda -c conda-forge \

"python>=3.6" scikit-learn=0.22.1 imbalanced-learn \

pandas seaborn hmmer prodigal screed ruamel.yaml \

"snakemake>=5.16,<=5.26" click

#conda创建环境vs2并安装指定版本的相关软件至该环境

conda activate vs2

#激活vs2环境

git clone https://github.com/jiarong/VirSorter2.git

#从github下载源文件

cd VirSorter2

#进入VirSorter2目录

pip install -e .

#安装

**下载数据库：**

在使用VirSorter2之前，用户需下载其数据库和相关文件，目前VirSorter2数据库包含dsDNAphage,NCLDV,RNA,ssDNA,lavidaviridae五大类病毒数据。

rm -rf db

#注意，如果用户此前取消下载过，则需先运行该命令移除db目录，如之前未失败下载不需要进行此步骤

virsorter setup -d db -j 4

#正式下载数据库，大约需要10+分钟下载VirSorter2所有的数据库和相关文件，该命令下载内容储存于db目录中。

**使用：**

Usage：

    virsorter run [options] [all|classify]

Options：

all or classify               VirSorter2运行有三个步骤：1)序列预处理，2)提取序列特

                               征，3)分类。如选择all(默认值)则全部三步骤均运行，如选

                               择classify则只运行第三步，适合改变参数重新运行分析的

                               情况。

-w|--working-dir PATH         输出结果路径

-d|--db-dir PATH              数据库路径，安装时路径即为默认路径

-i|--seqfile PATH             输入的序列文件，需为fa或fq格式

-l|--label TEXT               为输出结果文件添加前缀，在使用不同参数重新分析时较有用

--include-groups TEXT         用户需要的病毒类别，可选值有：

                               dsDNAphage,NCLDV,RNA,ssDNA,lavidaviridae，

                               多选需以逗号分隔，无空格。默认值为：

                               dsDNAphage,NCLDV,RNA,ssDNA,lavidaviridae

-j|--jobs INTEGER             最大并行任务数，默认值为256

--min-score FLOAT             被识别为病毒的最小打分值，默认值为0.5

--hallmark-required           在所有序列中均要求有标记基因，默认值为False

--hallmark-required-on-short  在短序列中要求有标记基因，默认值为False。短序列标准在

                               template-config.yaml中确定，用户可自行更改，默认值为

                               3kb。该选项可在丢失可接受的敏感度的同时降低识别假阳性

--viral-gene-required         需要有属于病毒的基因被注释，无基因被注释的可能病毒序列

                               被移除。该选项可在丢失可接受的敏感度的同时降低识别假阳

                               性，默认值为False

--provirus-off                识别完所有序列后不提取溶原性病毒序列，和--max-orf-

                               per-seq结合使用可大大提速运行，默认值为False

--max-orf-per-seq INTEGER     该选项仅在--provirus-off模式下可用，后接计算序列分类

                               特征所需要最大ORF数量，如某序列拥有超过该数量的ORFs，

                               则对其进行subsample至该数量。默认值值为-1

--min-length INTEGER          进行识别的序列的最短长度要求，默认值为0

--prep-for-dramv              为DRAMv生成病毒序列文件和viralaffi-contigs.tab文

                               件，DRAMv是一款能够注释VirSorter2输出的病毒序列的软

                               件，默认值为False

--tmpdir TEXT                 为临时文件命名

--rm-tmpdir                   移除临时文件，默认值为False

--verbose                     显示详细输出，默认值为False

-h|--help                     显示帮助信息

**实例：**

使用VirSorter2官网给出的实例文件进行举例。

wget -O test.fa \ https://raw.githubusercontent.com/jiarong/VirSorter2/master/test/8seq.fa

#下载示例序列文件

virsorter run -w test.out -i test.fa -j 4 all

#例1:以test.fa作为输入，使用默认下载数据库，使用4个线程，运行VirSorter2全部三步骤进行病毒序列的识别，结果输出至test.out

virsorter run all -w test.out -i test.fa -j 20 \

-l DNA\_mins0.7\_minl1.5 \

--include-groups dsDNAphage,ssDNA \

--min-score 0.7 \

--hallmark-required-on-short 、

--min-length 1500 \

--prep-for-dramv

#例2:以test.fa作为输入，使用默认下载数据库，使用20个线程，运行VirSorter2全部三步骤进行病毒序列的识别，仅对dsDNAphage和ssDNA病毒进行识别，长度>1.5kb且打分>0.7的结果保留，短于3kb的识别序列必须有hallmark基因，为DRAMv输出相关文件，结果输出至test.out并加上DNA\_mins0.7\_minl1.5前缀区分

**核心输出结果：**

VirSorter2运行后会生成三个核心输出文件，分别为：

final-viral-combined.fa #储存识别得到的病毒序列

final-viral-score.tsv #每条序列的各分类类别得分表格

final-viral-boundary.tsv #每条序列的信息表格

核心输出结果的详细描述：

1. final-viral-combined.fa

该fa文件储存VirSorter2识别得到的病毒序列，每条序列的名称为原始序列名称加后缀，后缀有以下三种可能：

i) ||full 代表整条序列都被识别为病毒序列

ii) ||{i}\_partial 代表该序列仅有部分区域被识别为病毒序列，其中i可以为0至该序列中找到的病毒片段数量的最大值

iii) ||lt2gene 拥有病毒标记基因的短序列（少于2个基因的序列）

2. final-viral-score.tsv

该表格主要用于进一步筛选识别的病毒序列，其中主要内容包括：

seqname                                  序列名称

dsDNAphage,NCLDV,RNA,ssDNA,lavidaviridae  各类别病毒下的打分

max\_score                                  最高分

max\_score\_group                           最高分所述类别

length                                    序列长度

hallmark                                  标记基因个数

viral                                     病毒基因比例

cellular                                  非病毒基因比例

#需要注意的是，VirSorter在此给出的分类学分类并不可靠，VirSorter2的目的仅限于病毒识别。

3.final-viral-boundary.tsv

该表格主要内容包括：

seqname                                    序列名称

trim\_orf\_index\_start,trim\_orf\_index\_end   该序列被识别为病毒的部分的ORF的起止位置

trim\_bp\_start,trim\_bp\_end                 该序列被识别为病毒的部分的碱基的起止位置

trim\_pr                                    该序列被识别为病毒的部分的最终得分

partial                                   该序列是否全部被识别为病毒，如果该序列的整

                                          体得分>cutoff，则该序列整体都被视为病毒序

                                          列(0)，反之则仅有部分被视为病毒序列(1)

pr\_full                                  该序列的整体得分

hallmark\_cnt                              标记基因个数

group                                     最高分所述病毒类别

**VirFinder使用教程**

原网站：<https://github.com/jessieren/VirFinder>

**简介：**

VirFinder是另一款常用的病毒长序列(viral contigs)识别软件，该软件不基于现有数据库比对，而是基于病毒和宿主具有不同的k-mer频率特征，使用机器学习的方法来识别病毒序列，常常与VirSorter结合使用。

**安装：**

Linux系统推荐使用conda安装

conda create -n VF -c bioconda r-virfinder

#conda创建环境VF并将VirFinder安装进该环境

conda activate VF

#激活VF环境

Windows系统需先进入<https://github.com/jessieren/VirFinder/blob/master/windows/VirFinder_1.1.zip>，下载[VirFinder\_1.1.zip](https://github.com/jessieren/VirFinder/blob/master/windows/VirFinder_1.1.zip)至本地，然后进入R安装。

install.packages("glmnet", dependencies=TRUE)

install.packages("Rcpp", dependencies=TRUE)

source("https://bioconductor.org/biocLite.R")

biocLite("qvalue")

#安装相关依赖包

install.packages("<path\_to\_the\_file>/VirFinder\_1.1.zip", repos = NULL, type="source")

#安装VirFinder，<path\_to\_the\_file>为下载VirFinder\_1.1.zip时保存的路径

library(VirFinder)

**使用实例：**

VirFinder实际为一R包，在R环境中运行分析，在此使用官方提供的序列文件进行实例讲解。

R #进入R，如为Windows系统则直接进入R

library(VirFinder) #加载VirFinder包

inFaFile <- system.file("data", "contigs.fa", package="VirFinder")

#指定需要识别病毒信号的序列文件，此处以官网给出的contigs.fa为例

#用户使用自己的文件只需将此句替换为 inFaFile <- "<path\_to\_the\_fasta\_file>"

predResult <- VF.pred(inFaFile)

#VirFinder执行病毒序列识别预测

predResult[order(predResult$pvalue),]

#升序排列p-value

predResult$qvalue <- VF.qvalue(predResult$pvalue)

#根据p-values估算q-value

predResult[order(predResult$qvalue),]

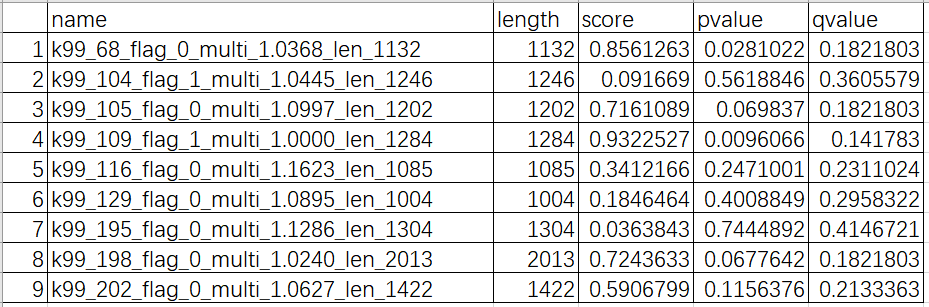
#升序排列q-value

write.table(predResult,file = "<path to output file>",sep = "\t")

#将predResult结果输出，<path to output file>为用户指定的输出路径

核心输出：

使用实例中最后输出的结果即为VirFinder核心输出。



结果储存了各条序列的名称、长度、打分、p-value和q-value。

用户可根据打分和p-value进一步筛选得到的病毒长序列。

**参考文献：**

1 Chen, L. X. *et al.* Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat Microbiol* **5**, 1504-1515, doi:10.1038/s41564-020-0779-9 (2020).

2 Sieradzki, E. T., Ignacio-Espinoza, J. C., Needham, D. M., Fichot, E. B. & Fuhrman, J. A. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat Commun* **10**, 1169, doi:10.1038/s41467-019-09106-z (2019).

3 Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425-430, doi:10.1038/nature19094 (2016).

4 <47.Patterns and ecological drivers of ocean viral communities.pdf>.

5 Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689-693, doi:10.1038/nature19366 (2016).

6 Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123 e1114, doi:10.1016/j.cell.2019.03.040 (2019).

7 Dzunkova, M. *et al.* Defining the human gut host-phage network through single-cell viral tagging. *Nat Microbiol* **4**, 2192-2203, doi:10.1038/s41564-019-0526-2 (2019).

8 Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817, doi:10.7717/peerj.3817 (2017).

9 Coutinho, F. H. *et al.* Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* **8**, 15955, doi:10.1038/ncomms15955 (2017).

10 Moniruzzaman, M. *et al.* Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat Commun* **8**, 16054, doi:10.1038/ncomms16054 (2017).