Script P8: Phage Replication Cycle

HANNIGAN GD, GRICE EA, ET AL.

Abstract

This protocol provides a method for predicting the proportions of phage replication cycles within the skin virome. We will be looking at three markers for temperate phages: 1) presence of integrase genes, 2) presence of ACLAME database prophage elements, and 3) similarity to bacterial genomes. Based on the methods found in the following publication:

Hannigan, Geoffrey D., et al. "The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome." *mBio* 6.5 (2015): e01578-15.

Citation: HANNIGAN GD, GRICE EA, ET AL. Script P8: Phage Replication Cycle. **protocols.io** dx.doi.org/10.17504/protocols.io.egwbbxe
Published: 10 Mar 2016

Guidelines

Required Software:

- UniProt
- ACLAME
- NCBI Bacterial Genomes
- NCBI's BLAST+ v2.2.0

Relevent Files

Output:

- Phage_replication_cycle/end_contig_counts_final.tsv
- Phage_replication_cycle/final_contig_quant_annotation_ncbi.tsv
- Phage_replication_cycle/phage_lifecycle_otu_table_for_rel_abund.tsv

Perl script: remove_block_fasta_format.pl
R script: R11

Before start

Perl scripts and other supplementary information available at:

https://figshare.com/articles/The_Human_Skin_dsDNA_Virome_Topographical_and_Temporal_Diversity_Genetic_Enrichment_and_Dynamic_Associations_with_the_Host_Microbiome/1281248

Protocol

Database Preparation

Step 1.

First go to the UniProt website and search fro all of the phage integrases using the following search terms: organisms: phage AND integrase which gave 32 Swiss-Prot genes and 1,120 TrEMBL genes (accessed Sep 02, 2014).

LINK:
http://beta.uniprot.org/uniprot/?query=organism%3aphage+AND+integrase&columns=id%2centry+name%2creviewed%2cprotein+names%2cgenes%2corganism%2clength&offset=0&sort=score

NOTES

Geoffrey Hannigan 28 Jan 2016

Because of the way the downloads were setup, I just had to download these reference fastas onto my local computer using my internet browser. These could then be easily transfered to a server for remote analysis as './references/UniProt_Phage_Integrase/uniprot-organism_3Aphage_AND_integrase.fasta'.

Database Preparation

Step 2.

Remove block formatting from the integrase reference database.

COMMAND

```
echo Removing block format from integrase reference database...
perl remove_block_fasta_format.pl ./references/UniProt_Phage_Integrase/uniprot-organism_3Aphage_AND_integrase.fasta ./references/UniProt_Phage_Integrase/uniprot-organism_3Aphage_AND_integrase_no_block.fasta
```

Database Preparation

Step 3.

Create blast database from the integrase reference fasta and store it in the same reference directory.

COMMAND

```
echo Creating blast database from integrase reference fasta...
makeblastdb -dbtype prot -in ./references/UniProt_Phage_Integrase/uniprot-organism_3Aphage_AND_integrase_no_block.fasta -
out ./references/UniProt_Phage_Integrase/uniprot_phage_integrase_db
```

NOTES

Geoffrey Hannigan 28 Jan 2016

We are now able to use this reference dataset to calculate the numbers of contigs that were annotated as bacteriophages (see phage taxonomy script) as well as integrase genes. In the end we want to append the numbers of contigs that had integrase, were phages, or both, to the master data list. This will be used for the end result Euler diagram.

Blastx the contig OTUs against the integrase database

Step 4.

Make dir for the integrase output.

COMMAND

```
mkdir ./phage_lifecycle
mkdir ./phage_lifecycle/integrase
```

Blastx the contig OTUs against the integrase database

Step 5.

The query ORFs were generated in a previous script but I will use them again here.

cmd COMMAND

```
echo Blastxing ORFs to the integrase reference database...
blastx -
query /home/ghanni/Analysis/Human_virome_analysis/glimmer3/output/Contigs_no_block_with_nam
es_glimmer_output_final.fa -out ./phage_lifecycle/integrase/blastx_ORFs_against_int.tsv -
db /project/egricelab/references/UniProt_Phage_Integrase/uniprot_phage_integrase_db -
outfmt 6 -num_threads 16 -max_target_seqs 1 -evalue 1e-5
```

Blastx the contig OTUs against the integrase database

Step 6.

Pull out the contig IDs from the blastx output. This is the list of all contigs that contain an integrase gene.

cmd COMMAND

```
echo Getting list of contigs containing an integrase gene...
cut -
f 1 ./phage_lifecycle/integrase/blastx_ORFs_against_int.tsv | sed 's/\.orf.*//' | sort | un
iq | sed 's/^/>/' > ./phage_lifecycle/integrase/int_contig_hit_list.tsv
```

Blastx the contig OTUs against the integrase database

Step 7.

Determine what contigs had hits in each anatomical location. Use bowtie2 alignments that have been calculated in other scripts to get a list of contig hits.

cmd COMMAND

```
mkdir ./phage_lifecycle/contig_presence_lists_per_sample_for_integrase
echo Getting lists of contigs present in each negative-cleaned sample...
for file in $(ls ./uniprot_contig_virome_trembl_rel_abund/abundance_from_sam); do
    sed 's/\t.*//' ./uniprot_contig_virome_trembl_rel_abund/abundance_from_sam/${file} | ta
il -
n +2 | sort | uniq | sed 's/^/>/' > ./phage_lifecycle/contig_presence_lists_per_sample_for_
integrase/${file}
done
```

Overall Integrase Instead of By-Sample (This is ultimately what we use)

Step 8.

First get a list of the contigs that are found in the sites of interest.

cmd COMMAND

```
echo Get list of contigs in overall non-negative sites...
# Get list of SampleIDs of the bacteriophage contigs in the specific sites
awk '$3 != "NA" && $7 != "Neg" { print $3 }' ./SkinMet_and_Virome_001_metadata.tsv > ./phag
e_lifecycle/integrase/specific_site_sampleID_list.txt
```

Overall Integrase Instead of By-Sample (This is ultimately what we use)

Step 9.

Get list of the contigs present in each of those samples listed.

cmd COMMAND

```
for name in $(cat ./phage_lifecycle/integrase/specific_site_sampleID_list.txt); do
    cat ./phage_lifecycle/contig_presence_lists_per_sample_for_phage_genes/${name}_R1.txt >
> ./phage_lifecycle/integrase/phage_contigs_no_negs.txt
done
```

Overall Integrase Instead of By-Sample (This is ultimately what we use)

Step 10.

Get the uniq contigs in this list.

cmd COMMAND

```
sort ./phage_lifecycle/integrase/phage_contigs_no_negs.txt | uniq > ./phage_lifecycle/integ
rase/phage_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/integrase/phage_contigs_no_negs_uniq.txt | sed 's/^ *//' | sed 's/ /\t/
```

```
' | sed 's/phage_contigs_no_negs_uniq\.txt/Phage_Contig_Count/'>> ./phage_lifecycle/end_con
tig_counts.tsv
```

Overall Integrase Instead of By-Sample (This is ultimately what we use)

Step 11.

Perform the same thing with the integrase genes.

cmd COMMAND

```
for name in $(cat ./phage_lifecycle/integrase/specific_site_sampleID_list.txt); do
    cat ./phage_lifecycle/int_contig_presence_lists_per_sample_for_integrase/${name}_R1.txt
 >> ./phage_lifecycle/integrase/int_contigs_no_negs.txt
done
sort ./phage_lifecycle/integrase/int_contigs_no_negs.txt | uniq > ./phage_lifecycle/integra
se/int_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/integrase/int_contigs_no_negs_uniq.txt | sed 's/^ *//' | sed 's/ /\t/'
| sed 's/int_contigs_no_negs_uniq\.txt/Integrase_Contig_Count/'>> ./phage_lifecycle/end_con
tig_counts.tsv
```

Overall Integrase Instead of By-Sample (This is ultimately what we use)

Step 12.

Get the number of shared contigs between the phage and int contig sets using awk.

cmd COMMAND

```
awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/integrase/phage_con
tigs_no_negs_uniq.txt ./phage_lifecycle/integrase/int_contigs_no_negs_uniq.txt > ./phage_li
fecycle/integrase/int_phage_shared_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/integrase/int_phage_shared_contigs_no_negs_uniq.txt | sed 's/^ *//' | s
ed 's/ /\t/' | sed 's/int_phage_shared_contigs_no_negs_uniq\.txt/Phage_and_Integrase_Contig
_Count/'>> ./phage_lifecycle/end_contig_counts.tsv
```

Temperate Phage Prediction by ACLAME Detection

Step 13.

Download the prophage protein database (fasta) of ACLAME version 0.4 checking the following boxes: Sequence length; Original NCBI annotation; ACLAME family assignment (only first box); Cross-references; ACLAME function annotations; Sequences; Export in FASTA format; gzipped data.

LINK:
http://aclame.ulb.ac.be/perl/Aclame/Tools/exporter.cgi?id=all&source=proteins&entry_id=1&length=on&ncbi_desc=on&family=on&xrefs=on&funct=on&sequence=on&fmt=fasta&format=gzip&x=145&y=23

Temperate Phage Prediction by ACLAME Detection

Step 14.

Unzip the downloaded file.

cmd COMMAND

```
gunzip ./references/aclame_protein_prophages_ref/aclame_proteins_prophages_0.4.fasta.gz
```

Temperate Phage Prediction by ACLAME Detection

Step 15.

Remove the block fasta format using the perl script remove_block_fasta_format.pl.

cmd COMMAND

```
perl ./remove_block_fasta_format.pl ./references/aclame_protein_prophages_ref/aclame_protei
ns_prophages_0.4.fasta ./references/aclame_protein_prophages_ref/aclame_proteins_prophages_
no_block.fasta
```

Temperate Phage Prediction by ACLAME Detection

Step 16.

Create blast database from the ACLAME reference fasta and store it in the same reference directory.

cmd COMMAND

```
echo Creating blast database from ACLAME reference fasta...
makeblastdb -dbtype prot -
in ./references/aclame_protein_prophages_ref/aclame_proteins_prophages_no_block.fasta -
out ./references/aclame_protein_prophages_ref/aclame_proteins_prophages_db
```

Temperate Phage Prediction by ACLAME Detection

Step 17.

Make dir for the ACLAME output.

cmd COMMAND

```
mkdir ./phage_lifecycle/ACLAME
```

Temperate Phage Prediction by ACLAME Detection

Step 18.

The query ORFs were generated in a different script but I will use them again here.

cmd COMMAND

```
echo Blastxing ORFs to the ACLAME reference database...
blastx -query ./glimmer3/output/Contigs_no_block_with_names_glimmer_output_final.fa -
out ./phage_lifecycle/ACLAME/blastx_ORFs_against_ACLAME.tsv -
db ./references/aclame_protein_prophages_ref/aclame_proteins_prophages_db -outfmt 6 -
num_threads 16 -max_target_seqs 1 -evalue 1e-5
```

Temperate Phage Prediction by ACLAME Detection

Step 19.

Pull out the contig IDs from the blastx output, filter out the contigs that did not have at least one orf match per 10kb, return the numbers of orfs that had assigned taxonomy to each contig.

cmd COMMAND

```
echo Getting list of contigs containing an ACLAME gene once every 10kb...
cut -
f 1 ./phage_lifecycle/ACLAME/blastx_ORFs_against_ACLAME.tsv | sed 's/\.orf.*//' | sort | un
iq -
c | sed 's/^ *//' | sed 's/ /\t/' > ./phage_lifecycle/ACLAME/ACLAME_contig_filtered_count_l
ength_list.tsv
awk 'FNR==NR { a[$1]=$2; next } $2 in a { print $2"\t"$1"\t"a[$2]"\t"10000*$1/a[$2] }' ./co
ntig_stats/contig_length_without_greater_sign.txt ./phage_lifecycle/ACLAME/ACLAME_contig_fi
ltered_count_length_list.tsv | awk '$4 > 1' | cut -
f 1 > ./phage_lifecycle/ACLAME/ACLAME_contig_filtered_hit_list.tsv
```

Temperate Phage Prediction by ACLAME Detection

Step 20.

This is the list of all contigs that contain an ACLAME gene at least once every 10kb.

cmd COMMAND

```
sort ./phage_lifecycle/ACLAME/ACLAME_contig_filtered_hit_list.tsv | uniq | sed 's/^/>/' > .
/phage_lifecycle/ACLAME/ACLAME_contig_hit_list.tsv
```

Temperate Phage Prediction by ACLAME Detection

Step 21.

Now that you have both lists, use awk to generate a list of the shared contigs. This is the list of contigs, per sample, that contain an ACLAME gene and are present after negative control cleaning.

cmd COMMAND

```
echo Generating lists of ACLAME gene containing contigs...
mkdir ./phage_lifecycle/ACLAME_contig_presence_lists_per_sample_for_ACLAME
for file in $(ls ./phage_lifecycle/contig_presence_lists_per_sample_for_integrase); do
    awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/contig_presence
_lists_per_sample_for_integrase/${file} ./phage_lifecycle/ACLAME/ACLAME_contig_hit_list.tsv
 > ./phage_lifecycle/ACLAME_contig_presence_lists_per_sample_for_ACLAME/${file}
done
```

Temperate Phage Prediction by ACLAME Detection

Step 22.

First get a list of the contigs that are found in the sites of interest.

cmd COMMAND

```
echo Get list of contigs in overall non-negative sites...
# Get list of ACLAME hit contigs that are found in the non-negative control samples
for name in $(cat ./phage_lifecycle/integrase/specific_site_sampleID_list.txt); do
    cat ./phage_lifecycle/ACLAME_contig_presence_lists_per_sample_for_ACLAME/${name}_R1.txt
 >> ./phage_lifecycle/ACLAME/ACLAME_contigs_no_negs.txt
done

sort ./phage_lifecycle/ACLAME/ACLAME_contigs_no_negs.txt | uniq > ./phage_lifecycle/ACLAME/
ACLAME_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/ACLAME/ACLAME_contigs_no_negs_uniq.txt | sed 's/^ *//' | sed 's/ /\t/'
| sed 's/ACLAME_contigs_no_negs_uniq\.txt/ACLAME_Contig_Count/'>> ./phage_lifecycle/end_con
tig_counts.tsv
```

Temperate Phage Prediction by ACLAME Detection

Step 23.

Get the number of shared contigs between the phage and ACLAME contig sets using awk.

cmd COMMAND

```
awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/integrase/phage_con
tigs_no_negs_uniq.txt ./phage_lifecycle/ACLAME/ACLAME_contigs_no_negs_uniq.txt > ./phage_li
fecycle/ACLAME/ACLAME_phage_shared_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/ACLAME/ACLAME_phage_shared_contigs_no_negs_uniq.txt | sed 's/^ *//' | s
ed 's/ /\t/' | sed 's/ACLAME_phage_shared_contigs_no_negs_uniq\.txt/Phage_and_ACLAME_Contig
_Count/'>> ./phage_lifecycle/end_contig_counts.tsv
```

Temperate Phage Prediction by ACLAME Detection

Step 24.

Get the number of contigs that are shared between integrase and ACLAME contigs.

cmd COMMAND

```
awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/integrase/int_conti
gs_no_negs.txt ./phage_lifecycle/ACLAME/ACLAME_contigs_no_negs_uniq.txt > ./phage_lifecycle
/ACLAME/ACLAME_int_shared_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/ACLAME/ACLAME_int_shared_contigs_no_negs_uniq.txt | sed 's/^ *//' | sed
 's/ /\t/' | sed 's/ACLAME_int_shared_contigs_no_negs_uniq\.txt/ACLAME_and_Integrase_Contig
_Count/'>> ./phage_lifecycle/end_contig_counts.tsv
```

Be careful in this awk line because the order of the input file matters and it must be in this order or else there will be duplicates.

Temperate Phage Prediction by ACLAME Detection

Step 25.

Clean up the end_contig_count row names.

cmd COMMAND

```
sed 's/\.\/.*\///' ./phage_lifecycle/end_contig_counts.tsv > ./phage_lifecycle/end_contig_c
ounts_final.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 26.

Finally we looked at predicting the temperate phage contigs by matching the contigs to bacterial genomes (downloaded directly from NCBI).

cmd COMMAND

```
wget ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz
```

Temperate Phage Prediction with Bacterial Genomes

Step 27.

Download the reference table.

cmd COMMAND

```
wget ftp://ftp.ncbi.nih.gov/genomes/Bacteria/summary.txt
```

Temperate Phage Prediction with Bacterial Genomes

Step 28.

Unzip the reference.

cmd COMMAND

```
tar -zxvf ./all.fna.tar.gz
rm all.fna.tar.gz
cat ../all_fasta/*/*.fna > ./ncbi_bacteria.fa
perl ./remove_block_fasta_format.pl ./ncbi_bacteria.fa ./ncbi_bacteria_no_block_chromosome.
fa
```

Temperate Phage Prediction with Bacterial Genomes

Step 29.

First make a blast+ database of the bacterial fasta file.

cmd COMMAND

```
echo Make blast database for bacteria genome reference...
makeblastdb -dbtype nucl -
in ./references/ncbi_bacteria_complete_genomes/ncbi_bacteria_no_block_chromosome.fa -
out ./references/ncbi_bacteria_complete_genomes/ncbi_bacteria_chromosome_db
```

NOTES

Geoffrey Hannigan 28 Jan 2016

The third and final method for predicting what contigs could be temerpate phages will be determining what contigs have significant matches to bacterial genomes. This metric, like the previous two metrics, was outlined in Minot S, et al, 2011.

Temperate Phage Prediction with Bacterial Genomes

Step 30.

Make dir for the bacteria hit output.

cmd COMMAND

```
mkdir ./phage_lifecycle/bacteria_hits
```

Temperate Phage Prediction with Bacterial Genomes

Step 31.

The query contigs were generated in a different script but I will use them again here.

cmd COMMAND

```
echo Blastxing contigs to the bacteria_hits reference database...
blastn -query ./glimmer3/contigs/Contigs_no_block_with_names.fasta -
out ./phage_lifecycle/bacteria_hits/blastx_ORFs_against_bacteria_hits_no_length_filter.tsv
-db ./references/ncbi_bacteria_complete_genomes/ncbi_bacteria_chromosome_db -
outfmt "6 qseqid sseqid pident qlen length mismatch" -num_threads 16 -max_target_seqs 1 -
perc_identity 90 -evalue 1e-3
```

Temperate Phage Prediction with Bacterial Genomes

Step 32.

Print only the lines with a length hit greater than 90% of the query length. Filter blastn hits for only hits similar to more than 90% of the query.

cmd COMMAND

```
awk ' (100 * $5 / $4) > 90 { print }' ./phage_lifecycle/bacteria_hits/blastx_ORFs_against_b
acteria_hits_no_length_filter.tsv > ./phage_lifecycle/bacteria_hits/blastx_ORFs_against_bac
teria_hits.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 33.

Pull out the contig IDs from the blastn output.

cmd COMMAND

```
echo Getting list of contigs containing bacteria_hits gene...
cut -
f 1 ./phage_lifecycle/bacteria_hits/blastx_ORFs_against_bacteria_hits.tsv | sort | uniq | s
ed 's/^/>/' > ./phage_lifecycle/bacteria_hits/bacteria_hits_contig_hit_list.tsv
```

NOTES

Geoffrey Hannigan 28 Jan 2016

This is the list of all contigs that contain a bacteria_hits gene.

Temperate Phage Prediction with Bacterial Genomes

Step 34.

Now that I have both lists I can use awk to generate a list of only the shared contigs. This is the list of contigs, per sample, that contain a bacteria_hit and are present after negative control cleaning.

cmd COMMAND

```
echo Generating lists of bacteria_hits gene containing contigs...
mkdir ./phage_lifecycle/bacteria_hits_contig_presence_lists_per_sample_for_bacteria_hits
for file in $(ls ./phage_lifecycle/contig_presence_lists_per_sample_for_integrase); do
    awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/contig_presence
_lists_per_sample_for_integrase/${file} ./phage_lifecycle/bacteria_hits/bacteria_hits_conti
g_hit_list.tsv > ./phage_lifecycle/bacteria_hits_contig_presence_lists_per_sample_for_bacte
ria_hits/${file}
done
```

Temperate Phage Prediction with Bacterial Genomes

Step 35.

First get a list of the contigs that are found in the sites of interest.

cmd COMMAND

```
echo Get list of contigs in overall non-negative sites...
# Get list of bacteria_hits hit contigs that are found in the non-negative control samples
for name in $(cat ./phage_lifecycle/integrase/specific_site_sampleID_list.txt); do
    cat ./phage_lifecycle/bacteria_hits_contig_presence_lists_per_sample_for_bacteria_hits/
${name}_R1.txt >> ./phage_lifecycle/bacteria_hits/bacteria_hits_contigs_no_negs.txt
done
sort ./phage_lifecycle/bacteria_hits/bacteria_hits_contigs_no_negs.txt | uniq > ./phage_lif
ecycle/bacteria_hits/bacteria_hits_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/bacteria_hits/bacteria_hits_contigs_no_negs_uniq.txt | sed 's/^ *//' |
sed 's/ /\t/' | sed 's/bacteria_hits_contigs_no_negs_uniq\.txt/bacteria_hits_Contig_Count/'
>> ./phage_lifecycle/end_contig_counts.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 36.

Get the number of shared contigs between phage and bacteria_hits contig sets using awk.

cmd COMMAND

```
awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/integrase/phage_con
tigs_no_negs_uniq.txt ./phage_lifecycle/bacteria_hits/bacteria_hits_contigs_no_negs_uniq.tx
t > ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_shared_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_shared_contigs_no_negs_uniq.txt | sed
 's/^ *//' | sed 's/ /\t/' | sed 's/bacteria_hits_phage_shared_contigs_no_negs_uniq\.txt/Ph
age_and_bacteria_hits_Contig_Count/'>> ./phage_lifecycle/end_contig_counts.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 37.

Get the number of contigs that are shared between integrase and bacteria_hits contigs.

cmd COMMAND

```
awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/integrase/int_conti
```

```
gs_no_negs.txt ./phage_lifecycle/bacteria_hits/bacteria_hits_contigs_no_negs_uniq.txt > ./p
hage_lifecycle/bacteria_hits/bacteria_hits_int_shared_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/bacteria_hits/bacteria_hits_int_shared_contigs_no_negs_uniq.txt | sed '
s/^ *//' | sed 's/ /\t/' | sed 's/bacteria_hits_int_shared_contigs_no_negs_uniq\.txt/bacter
ia_hits_and_Integrase_Contig_Count/'>> ./phage_lifecycle/end_contig_counts.tsv
```

NOTES

Geoffrey Hannigan 02 Feb 2016

Be careful in this awk line because the order of the input file matters and it must be in this order or else there will be duplicates.

Temperate Phage Prediction with Bacterial Genomes

Step 38.

Get the number of contigs that are shared between integrase, phage, and bacteria_hits contigs.

cmd COMMAND

```
awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/integrase/int_conti
gs_no_negs.txt ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_shared_contigs_no_negs_u
niq.txt > ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_integrase_shared_contigs_no_n
egs_uniq.txt
wc -
l ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_integrase_shared_contigs_no_negs_uniq
.txt | sed 's/^ *//' | sed 's/ /\t/' | sed 's/bacteria_hits_phage_integrase_shared_contigs_
no_negs_uniq\.txt/Phage_and_Integrase_and_bacteria_hits_Contig_Count/'>> ./phage_lifecycle/
end_contig_counts.tsv

awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/ACLAME/ACLAME_conti
gs_no_negs.txt ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_shared_contigs_no_negs_u
niq.txt > ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_ACLAME_shared_contigs_no_negs
_uniq.txt
wc -
l ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_ACLAME_shared_contigs_no_negs_uniq.tx
t | sed 's/^ *//' | sed 's/ /\t/' | sed 's/bacteria_hits_phage_ACLAME_shared_contigs_no_neg
s_uniq\.txt/Phage_and_ACLAME_and_bacteria_hits_Contig_Count/'>> ./phage_lifecycle/end_conti
g_counts.tsv

awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/ACLAME/ACLAME_conti
gs_no_negs.txt ./phage_lifecycle/bacteria_hits/bacteria_hits_int_shared_contigs_no_negs_uni
q.txt > ./phage_lifecycle/bacteria_hits/bacteria_hits_integrase_ACLAME_shared_contigs_no_ne
gs_uniq.txt
wc -
l ./phage_lifecycle/bacteria_hits/bacteria_hits_integrase_ACLAME_shared_contigs_no_negs_uni
q.txt | sed 's/^ *//' | sed 's/ /\t/' | sed 's/bacteria_hits_integrase_ACLAME_shared_contig
s_no_negs_uniq\.txt/Integrase_and_ACLAME_and_bacteria_hits_Contig_Count/'>> ./phage_lifecyc
le/end_contig_counts.tsv

awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/ACLAME/ACLAME_conti
gs_no_negs.txt ./phage_lifecycle/bacteria_hits/bacteria_hits_contigs_no_negs_uniq.txt > ./p
hage_lifecycle/bacteria_hits/bacteria_hits_ACLAME_shared_contigs_no_negs_uniq.txt
wc -
l ./phage_lifecycle/bacteria_hits/bacteria_hits_ACLAME_shared_contigs_no_negs_uniq.txt | se
d 's/^ *//' | sed 's/ /\t/' | sed 's/bacteria_hits_ACLAME_shared_contigs_no_negs_uniq\.txt/
ACLAME_and_bacteria_hits_Contig_Count/'>> ./phage_lifecycle/end_contig_counts.tsv

awk 'FNR==NR { a[$1]=$1; next } $1 in a { print $1 }' ./phage_lifecycle/ACLAME/ACLAME_conti
gs_no_negs.txt ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_integrase_shared_contigs
_no_negs_uniq.txt > ./phage_lifecycle/bacteria_hits/bacteria_hits_ACLAME_integrase_phage_sh
ared_contigs_no_negs_uniq.txt
wc -
```

```
l ./phage_lifecycle/bacteria_hits/bacteria_hits_ACLAME_integrase_phage_shared_contigs_no_negs_uniq.txt | sed 's/^ *//' | sed 's/ /\t/' | sed 's/bacteria_hits_ACLAME_integrase_phage_shared_contigs_no_negs_uniq\.txt/ACLAME_and_bacteria_hits_and_integrase_and_phage_Contig_Count/'>> ./phage_lifecycle/end_contig_counts.tsv
```

NOTES

Geoffrey Hannigan 02 Feb 2016

Be careful in this awk line because the order of the input file matters and it must be in this order or else there will be duplicates.

Temperate Phage Prediction with Bacterial Genomes

Step 39.

Clean up the end_contig_count row names.

cmd COMMAND

```
sed 's/\.\/.*\///' ./phage_lifecycle/end_contig_counts.tsv > ./phage_lifecycle/end_contig_counts_final.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 40.

Going to want to remove the intermediate contig count file in the end.

cmd COMMAND

```
rm ./phage_lifecycle/end_contig_counts.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 41.

Remove the appended files in case the script needs to be run again.

cmd COMMAND

```
rm ./phage_lifecycle/bacteria_hits/bacteria_hits_contigs_no_negs.txt
rm ./phage_lifecycle/ACLAME/ACLAME_contigs_no_negs.txt
```

Temperate Phage Prediction with Bacterial Genomes

Step 42.

Get the number of phage contigs that have at least one temperate phage marker.

cmd COMMAND

```
cat ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_shared_contigs_no_negs_uniq.txt ./phage_lifecycle/ACLAME/ACLAME_phage_shared_contigs_no_negs_uniq.txt ./phage_lifecycle/integrase/int_phage_shared_contigs_no_negs_uniq.txt | sort | uniq | wc -
l > ./phage_lifecycle/list_contigs_at_least_one_temperate_hit.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 43.

Because we were matching the phage contigs to bacterial genomes, we are interested to see what those hits actually were. To do this, we simply annotated the resulting blast output from above. We took these annotations (genus level) and manually added the order level annotations since there were relatively few unique hits.

Temperate Phage Prediction with Bacterial Genomes

Step 44.

Get a list of the names and accession numbers that matched the contigs after blastn (f1 = contig number, f2 = accession). This is the overall list of possible contig hits to an accession number, but have not yet been length filtered (fine here, but watch out in downstream processes)

cmd COMMAND

```
cut -
f 1,2 ./phage_lifecycle/bacteria_hits/blastx_ORFs_against_bacteria_hits_no_length_filter.tsv | sed 's/^/>/' | sed 's/gi.*ref|//' | sed 's/|//' | sort | uniq > ./phage_lifecycle/bacteria_hits/blastn_contigs_ncbi_accs.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 45.
Get these accession number for those contigs which were similar to both bacteria and phages.

cmd COMMAND
```
awk 'FNR==NR { a[$1]=$2; next } $1 in a { print $1"\t"a[$1] }' ./phage_lifecycle/bacteria_h
its/blastn_contigs_ncbi_accs.tsv ./phage_lifecycle/bacteria_hits/bacteria_hits_phage_shared
_contigs_no_negs_uniq.txt > ./phage_lifecycle/bacteria_hits/ncbi_accs_for_contigs_match_pha
ge_and_bactera.tsv
```

Temperate Phage Prediction with Bacterial Genomes

Step 46.
Get a list of only the unique accession numbers that matched the phage+bacteria contigs.

cmd COMMAND
```
cat ./phage_lifecycle/bacteria_hits/ncbi_accs_for_contigs_match_phage_and_bactera.tsv | cut
 -
f 2 | sort | uniq > ./phage_lifecycle/bacteria_hits/ncbi_accs_for_contigs_match_phage_and_b
actera_uniq.tsv
```

Step 47.
Get a reference list of the uniq phage+bacteria accession numbers with their taxonomic information. First format the reference.

cmd COMMAND
```
grep '>' ./references/ncbi_bacteria_complete_genomes/ncbi_bacteria_no_block_chromosome.fa |
 sed 's/| /|\t/' | sed 's/ /_/g' | sed 's/.*|ref|\(.*\)|\t\(.*\)$/\1\t\2/' | sed 's/_chromo
some,_complete_genome//' | sed 's/complete_chromosome_//' > ./references/ncbi_bacteria_comp
lete_genomes/summary_chromosome_from_fasta.txt
awk 'FNR==NR { a[$1]=$2; next } { print $1"\t"a[$1] }' ./references/ncbi_bacteria_complete_
genomes/summary_chromosome_from_fasta.txt ./phage_lifecycle/bacteria_hits/ncbi_accs_for_con
tigs_match_phage_and_bactera_uniq.tsv > ./phage_lifecycle/bacteria_hits/ncbi_taxonomy_for_c
ontigs_match_phage_and_bactera_uniq.tsv
```

Step 48.
Get a list of only the genera.

cmd COMMAND
```
sed 's/\([^C]\)_.*/\1/' ./phage_lifecycle/bacteria_hits/ncbi_taxonomy_for_contigs_match_pha
ge_and_bactera_uniq.tsv > ./phage_lifecycle/bacteria_hits/ncbi_genus_taxonomy_for_contigs_m
atch_phage_and_bactera_uniq.tsv
```

Step 49.
Annotate each contig, with genus level, that matched both phage and bacteria.

cmd COMMAND
```
awk 'FNR==NR { a[$1]=$2; next } { print $1"\t"a[$2] }' ./phage_lifecycle/bacteria_hits/ncbi
_genus_taxonomy_for_contigs_match_phage_and_bactera_uniq.tsv ./phage_lifecycle/bacteria_hit
s/ncbi_accs_for_contigs_match_phage_and_bactera.tsv > ./phage_lifecycle/bacteria_hits/ncbi_
genus_taxonomy_contigs_match_phage_and_bactera_uniq.tsv
```

Step 50.
Quantify how many contigs hit each genus.

cmd COMMAND
```
cut -
f 2 ./phage_lifecycle/bacteria_hits/ncbi_genus_taxonomy_contigs_match_phage_and_bactera_uni
q.tsv | sort | uniq -
c | sed 's/ *//' | sed 's/ /\t/' | sed '1 s/^/Number_Contigs\tBacterial_Genus\n/' > ./phage
_lifecycle/bacteria_hits/final_contig_quant_annotation_ncbi.tsv
```

NOTES

Geoffrey Hannigan 02 Feb 2016
This output can be used in R for graphing. There are not many categories so the phylum levels can just be manually entered. This is faster than trying to deal with a data base.

Step 51.
Now we have the proportions of contigs that were annotated as potential temperate phages, but this is not a relative abundance of the samples, which would need to take into account the individual, un-assembled sequences. To calculate the relative abundance of temperate phages per site, we used the relative abundance table generated during our taxonomy analysis.

Step 52.
Get list of all of the predicted temperate phage contig IDs (predicted in the 'predict_temperate_phage.sh' script)

cmd COMMAND
```
mkdir ./phage_lifecycle
mkdir ./phage_lifecycle/temperate_phage_rel_abund
cat ./phage_lifecycle/integrase/int_phage_shared_contigs_no_negs_uniq.txt ./phage_lifecycle
/ACLAME/ACLAME_phage_shared_contigs_no_negs_uniq.txt ./phage_lifecycle/bacteria_hits/bacter
ia_hits_phage_shared_contigs_no_negs_uniq.txt | sort | uniq > ./phage_lifecycle/temperate_p
hage_rel_abund/temperate_phage_contig_id_list.txt
```

Step 53.
Get a list of all of the phage contig IDs that are not included in the temperate phage list.

cmd COMMAND
```
grep -v --
file=./phage_lifecycle/temperate_phage_rel_abund/temperate_phage_contig_id_list.txt ./phage
_lifecycle/integrase/phage_contigs_no_negs_uniq.txt  > ./phage_lifecycle/temperate_phage_re
l_abund/non_temperate_phage_contig_id_list.txt
```

Step 54.
Add annotation for whether the contig list is for temperate or lytic phages.

cmd COMMAND
```
sed 's/$/\tTemperate_Phage/' ./phage_lifecycle/temperate_phage_rel_abund/temperate_phage_co
ntig_id_list.txt > ./phage_lifecycle/temperate_phage_rel_abund/named_temperate_phage_contig
_id_list.txt
sed 's/$/\tNon-
Temperate_Phage/' ./phage_lifecycle/temperate_phage_rel_abund/non_temperate_phage_contig_id
_list.txt > ./phage_lifecycle/temperate_phage_rel_abund/named_non_temperate_phage_contig_id
_list.txt
```

Step 55.
Format the contig OTU table with greater-than signs.

cmd COMMAND
```
sed 's/^/>/' ./uniprot_contig_virome_trembl_rel_abund/contig_otu_table.txt > ./phage_lifecy
cle/temperate_phage_rel_abund/formatted_contig_otu_table.txt
```

Step 56.
Also get the header from this file.

cmd COMMAND
```
head -
n 1 ./uniprot_contig_virome_trembl_rel_abund/contig_otu_table.txt > ./phage_lifecycle/tempe
rate_phage_rel_abund/contig_otu_table_header.txt
```

Step 57.
In contig OTU table that was used for uniprot taxonomy, replace contig IDs with annotation of temperate status. Determine, of all of the contigs, how many have hits to temperate or non-temperate phages.

cmd COMMAND
```
awk 'FNR==NR { a[$1]=$2; next } $1 in a { print a[$1]"\t"$0 }' ./phage_lifecycle/temperate_
phage_rel_abund/named_temperate_phage_contig_id_list.txt ./phage_lifecycle/temperate_phage_
rel_abund/formatted_contig_otu_table.txt | cut -f 1,3-
 > ./phage_lifecycle/temperate_phage_rel_abund/temperate_contig_otu_table.txt
```

```
awk 'FNR==NR { a[$1]=$2; next } $1 in a { print a[$1]"\t"$0 }' ./phage_lifecycle/temperate_
phage_rel_abund/named_non_temperate_phage_contig_id_list.txt ./phage_lifecycle/temperate_ph
age_rel_abund/formatted_contig_otu_table.txt | cut -f 1,3-
 > ./phage_lifecycle/temperate_phage_rel_abund/non_temperate_contig_otu_table.txt
```

Step 58.

Put together the header, temperate phage abundance list, and the non-temperate abundance list.

cmd COMMAND

```
cat ./phage_lifecycle/temperate_phage_rel_abund/contig_otu_table_header.txt ./phage_lifecyc
le/temperate_phage_rel_abund/non_temperate_contig_otu_table.txt ./phage_lifecycle/temperate
_phage_rel_abund/temperate_contig_otu_table.txt > ./phage_lifecycle/temperate_phage_rel_abu
nd/phage_lifecycle_otu_table_for_rel_abund.tsv
```

NOTES

Geoffrey Hannigan 02 Feb 2016

In R, these relative abundance values can be summed based on their temperate status easily, and then stats can be run.