Supplemental Materials

Figure Legends

Supplemental Figure 1: Flowchart of *phiSpy*.

Supplemental Figure 2: An example of how to calculate the parameter *transcriptional strand orientation*.

Supplemental Figure 3: Permutation distribution for four different test statistics. The blue line indicates the observed difference of the mean/median of the two distribution of skew. The permutation Achieved Significant Level (ASL_{perm}) leads to rejection of the null hypothesis for all four statistics. (A) the distribution for customized AT skew and the observed differences of mean. (B) the distribution for customized AT skew and the observed differences of median. (C) the distribution for customized GC skew and the observed differences of mean. (D) the distribution for customized AT skew and the observed differences of median.

Supplemental Figure 4: Median protein length difference for bacteria (■) and phages (□). For bacteria, the difference is the median length of all proteins in the genome and the median of all bacterial proteins in the genome. For phages, the difference is the median length of all proteins in the genome and the median of all phage proteins in the genome. The median difference is higher for phage proteins than bacterial proteins.

Supplemental Figure 5: Flowchart of performance analysis.

Supplimental Table 1: List of 41 bacterial genomes, which have manually annotated prophages.

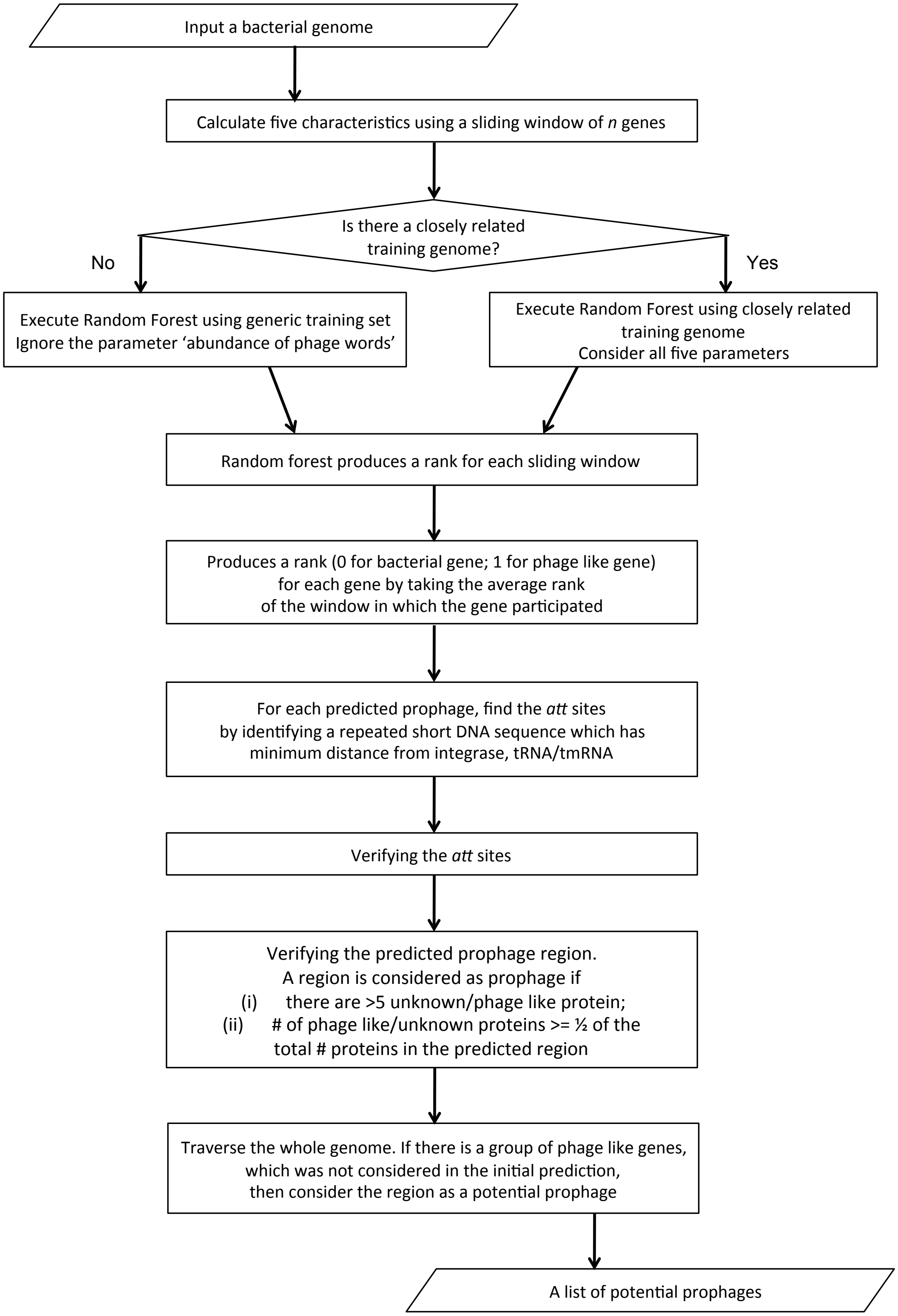
Bacillus halodurans C-125
Bacillus subtilis subsp. subtilis str. 168
Bifidobacterium longum NCC2705
Brucella melitensis 16M
Caulobacter crescentus CB15
Clostridium perfringens str. 13
Clostridium tetani E88
Deinococcus radiodurans R1
Escherichia coli CFT073
Escherichia coli K12
Escherichia coli O157:H7
Escherichia coli O157:H7 EDL933
Haemophilus influenzae Rd KW20
Lactococcus lactis subsp. lactis Il1403
Listeria innocua Clip11262
Listeria monocytogenes EGD-e
Mesorhizobium loti MAFF303099
Mycobacterium tuberculosis CDC1551
Mycobacterium tuberculosis H37Rv
Neisseria meningitidis MC58
Neisseria meningitidis Z2491
Pasteurella multocida subsp. multocida str. Pm70
Pseudomonas aeruginosa PAO1
Pseudomonas putida KT2440
Ralstonia solanacearum GMI1000
Salmonella enterica subsp. enterica serovar Typhi str. CT18
Shewanella oneidensis MR-1
Shigella flexneri 2a str. 301
Staphylococcus aureus subsp. aureus Mu50
Staphylococcus aureus subsp. aureus MW2
Streptococcus agalactiae 2603V/R
Streptococcus pyogenes M1 GAS
Streptococcus pyogenes MGAS315
Streptococcus pyogenes MGAS8232
Streptomyces coelicolor A3(2)
Vibrio cholerae O1 biovar eltor str. N16961
Xanthomonas axonopodis pv. citri str. 306
Xylella fastidiosa 9a5c
Xylella fastidiosa Temecula1
Yersinia pestis CO92
Yersinia pestis KIM

Supplemental Table 2: Prophage prediction in 45 complete bacterial genomes, which has a closely related training organism

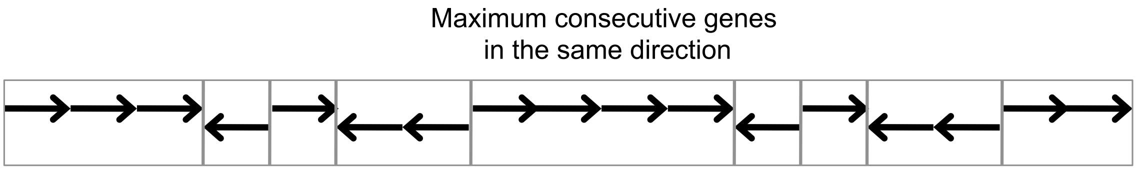
Training Organism	Organism	Known Prophage	Probable Prophage	Undefined
Brucella melitensis	*Brucella suis 1330*	0	0	0
Caulobacter crescentus	*Caulobacter sp. K31*	1	0	0
Escherichia coli K12	*Escherichia coli ATCC 8739*	3	0	0
Escherichia coli K12	*Escherichia coli CFT073*	5	0	0
Escherichia coli K12	*Escherichia coli E24377A*	6	1	0
Escherichia coli K12	*Escherichia coli O157:H7*	16	0	0
Escherichia coli K12	*Escherichia coli W3110*	3	0	0
Listeria innocua	*Listeria monocytogenes EGD-e*	1	0	0
Mycobacterium tuberculosis H37Rv	*Mycobacterium tuberculosis CDC1551*	2	0	0
Mycobacterium tuberculosis H37Rv	*Mycobacterium tuberculosis H37Ra*	2	0	0
Pseudomonas putida KT2440	*Pseudomonas putida W619*	3	0	0
Pseudomonas putida KT2440	*Pseudomonas syringae pv. phaseolicola 1448A*	3	0	0
Pseudomonas putida KT2440	*Pseudomonas syringae pv. tomato str. DC3000*	4	0	0
Escherichia coli K12	*Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67*	4	1	0
Escherichia coli K12	*Salmonella enterica subsp. enterica serovar Typhi str. CT18*	11	0	0
Escherichia coli K12	*Salmonella typhimurium LT2*	6	0	0
Shewanella oneidensis	*Shewanella baltica OS185*	4	0	0
Shewanella oneidensis	*Shewanella baltica OS195*	3	0	0
Shewanella oneidensis	*Shewanella sp.ANA-3*	0	0	0
Escherichia coli K12	*Shigella flexneri 2a str. 2457T*	20	0	0
Escherichia coli K12	*Shigella flexneri 2a str. 301*	8	0	0
Staphylococcus aureus subsp. aureus Mu50	*Staphylococcus aureus RF122*	2	0	0
Staphylococcus aureus subsp. aureus Mu50	*Staphylococcus aureus subsp. aureus COL*	1	0	0
Staphylococcus aureus subsp. aureus Mu50	*Staphylococcus aureus subsp. aureus MRSA252*	3	0	0
Staphylococcus aureus subsp. aureus MW2	*Staphylococcus aureus subsp. aureus Mu50*	3	0	0
Staphylococcus aureus subsp. aureus Mu50	*Staphylococcus aureus subsp. aureus NCTC 8325*	3	0	0
Staphylococcus aureus subsp. aureus Mu50	*Staphylococcus aureus subsp. aureus USA300*	2	0	0
Staphylococcus aureus subsp. aureus Mu50	*Staphylococcus epidermidis ATCC 12228*	0	0	0
Staphylococcus aureus subsp. aureus Mu50	*Staphylococcus haemolyticus JCSC1435*	2	0	0
Staphylococcus aureus subsp. aureus Mu50	*Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305*	0	0	0
Xanthomonas axonopodis	*Stenotrophomonas maltophilia K279a*	3	0	0
Streptococcus agalactiae 2603V/R	*Streptococcus agalactiae NEM316*	0	0	3
Streptococcus pyogenes MGAS8232	*Streptococcus pyogenes MGAS10394*	6	1	0
Streptococcus pyogenes MGAS8232	*Streptococcus uberis 0140J*	0	1	0
Streptomyces coelicolor	*Streptomyces avermitilis MA-4680*	1	0	0
Xanthomonas axonopodis	*Xanthomonas campestris pv. campestris ATCC 33913*	1	1	0
Xanthomonas axonopodis	*Xanthomonas campestris pv. campestris str. 8004*	0	2	0
Xanthomonas axonopodis	*Xanthomonas campestris pv. vesicatoria str. 85-10*	1	0	0
Xylella fastidiosa 2a str. 301	*Xylella fastidiosa M12*	3	0	0
Xylella fastidiosa 2a str. 301	*Xylella fastidiosa Temecula1*	7	0	0
Yersinia pestis CO92	*Yersinia enterocolitica 8081*	3	1	0
Yersinia pestis CO92	*Yersinia pestis biovar Medievalis str. 91001*	5	0	0
Yersinia pestis CO92	*Yersinia pestis KIM*	4	0	0
Yersinia pestis CO92	*Yersinia pseudotuberculosis IP 32953*	5	1	0
Yersinia pestis CO92	*Yersinia pseudotuberculosis YPIII*	6	1	0
	Total	**166**	**10**	**3**

Supplemental Figures:

Supplemental Figure 1

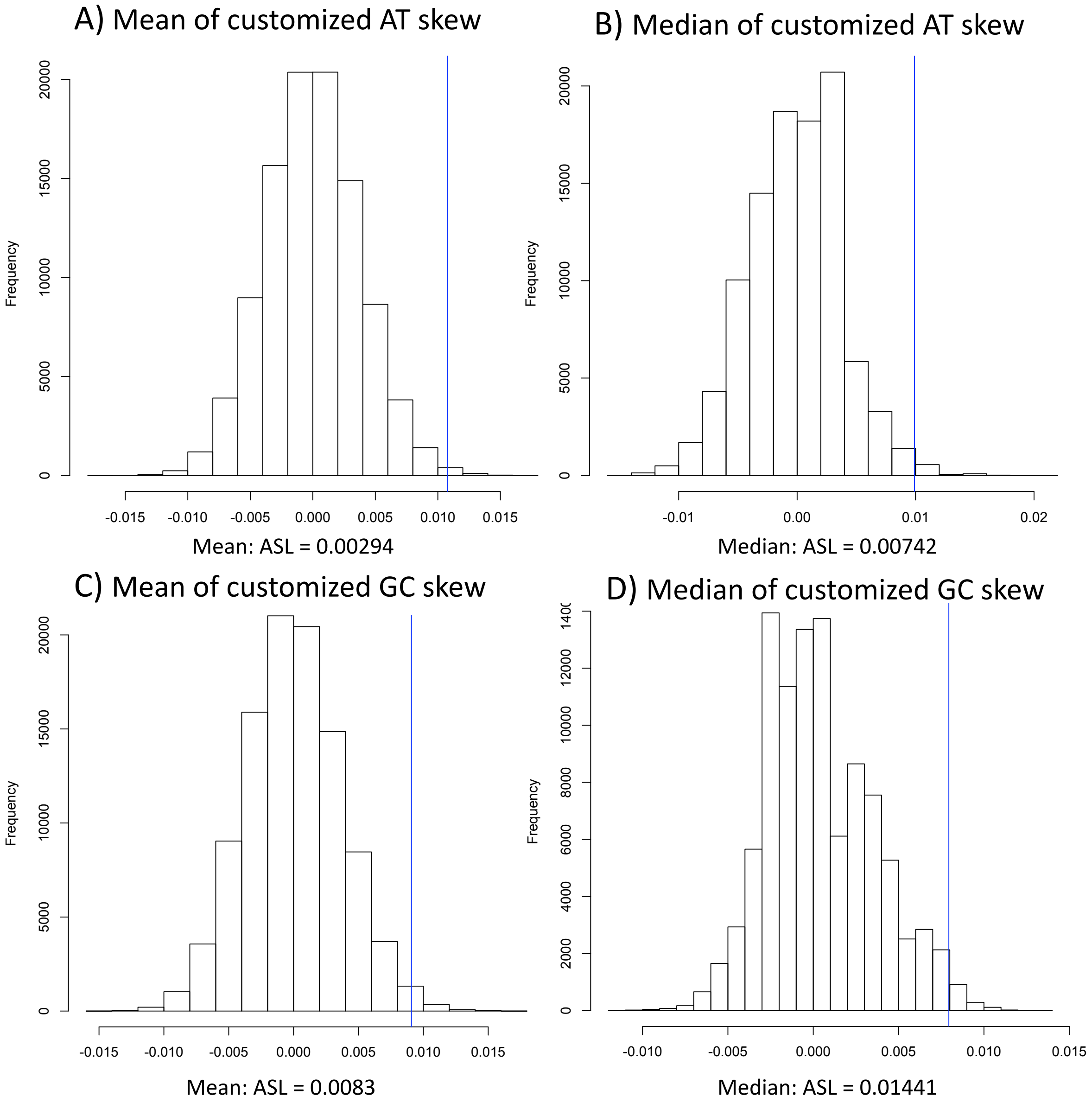

Input a bacterial genome
Calculate five characteristics using a sliding window of *n* genes
Is there a closely related training genome?
No
Yes
Execute Random Forest using generic training set Ignore the parameter 'abundance of phage words'
Execute Random Forest using closely related training genome Consider all five parameters
Random forest produces a rank for each sliding window
Produces a rank (0 for bacterial gene; 1 for phage like gene) for each gene by taking the average rank of the window in which the gene participated
For each predicted prophage, find the *att* sites by identifying a repeated short DNA sequence which has minimum distance from integrase, tRNA/tmRNA
Verifying the *att* sites
Verifying the predicted prophage region. A region is considered as prophage if
(i) there are >5 unknown/phage like protein;
(ii) # of phage like/unknown proteins >= ½ of the total # proteins in the predicted region
Traverse the whole genome. If there is a group of phage like genes, which was not considered in the initial prediction, then consider the region as a potential prophage
A list of potential prophages

Supplemental Figure 2



Total # of genes = 17
# of adjacent genes on opposite strands = 8
% of adjacent genes on opposite strands = 8/17 %
% of the maximum consecutive genes in the same direction = 4/17 %

Supplemental Figure 3


A) Mean of customized AT skew

Mean: ASL = 0.00294

B) Median of customized AT skew

Median: ASL = 0.00742

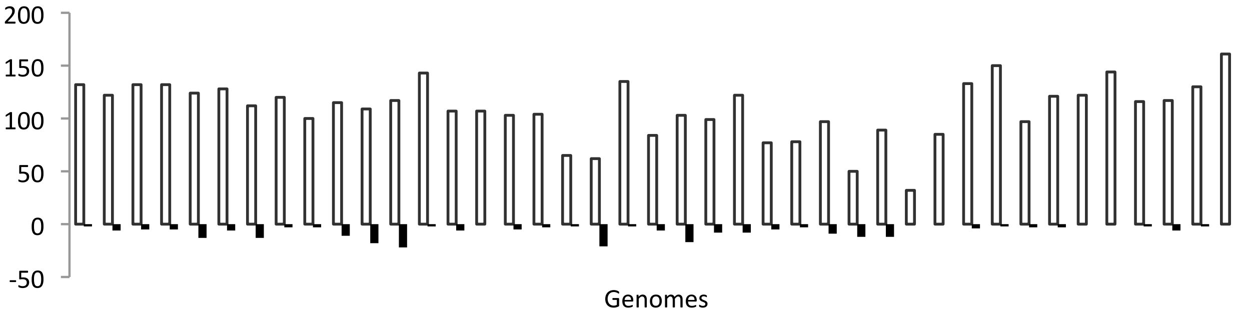
C) Mean of customized GC skew

Mean: ASL = 0.0083

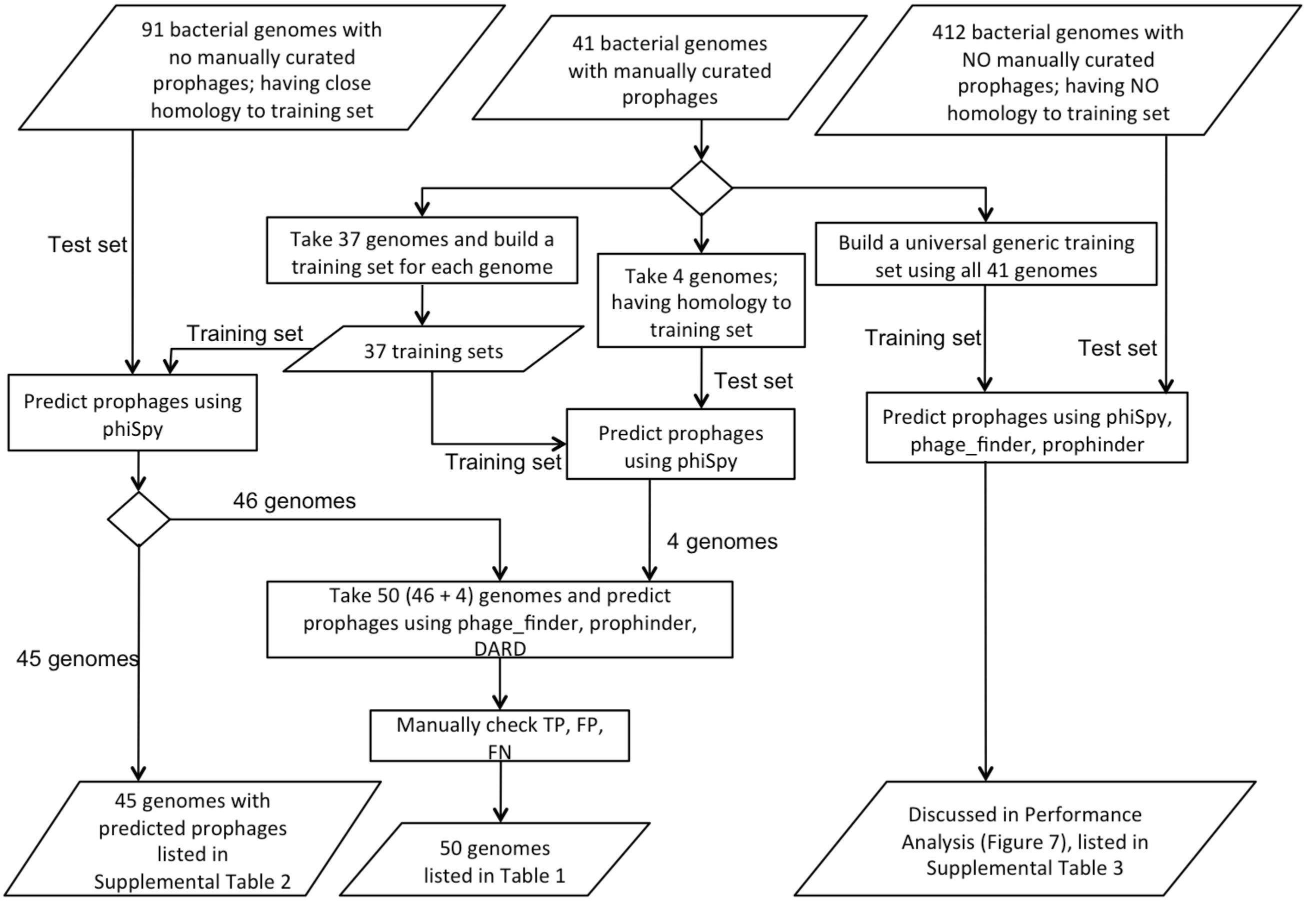
D) Median of customized GC skew

Median: ASL = 0.01441

Supplemental Figure 4



Supplemental Figure 5


91 bacterial genomes with no manually curated prophages; having close homology to training set
41 bacterial genomes with manually curated prophages
412 bacterial genomes with NO manually curated prophages; having NO homology to training set
Test set
Take 37 genomes and build a training set for each genome
Take 4 genomes; having homology to training set
Build a universal generic training set using all 41 genomes
Training set
37 training sets
Training set
Test set
Predict prophages using phiSpy
Test set
Predict prophages using phiSpy
Training set
Predict prophages using phiSpy, phage_finder, prophinder
46 genomes
4 genomes
Take 50 (46 + 4) genomes and predict prophages using phage_finder, prophinder, DARD
45 genomes
Manually check TP, FP, FN
45 genomes with predicted prophages listed in Supplemental Table 2
50 genomes listed in Table 1
Discussed in Performance Analysis (Figure 7), listed in Supplemental Table 3

Permutation test for customized AT/GC skew:

For permutation test, two samples (F and G) were created. To create sample F, 190 prophages in 41 complete bacterial genomes were considered. Sample F consists of the absolute difference between the customized AT/GC skew of prophage genes and the customized AT/GC skew of prophages' flanking genes (same length of corresponding prophage). The size of sample F is 190.

To make the sample G, 800 different bacterial regions were randomly selected from 41 bacterial genomes. The absolute differences of the customized AT/GC skew of these regions and the customized AT/GC skew of the flanking genes of these regions was calculated for sample G. The sample size of G is 800.

Null hypothesis, H_0: F = G

To test the null hypothesis we did the permutation test using both the difference of mean (mean of F – mean of G) and the difference of median (median of F – median of G). The test was done as follows:

1. The difference in means/medians between the two samples was calculated, which was the observed value of the test statistic.
2. Sample F and G were combined and randomly divided them into two groups (A and B) of size 190 and 800.
3. The difference in means/medians of group A and B was calculated and recorded.
4. Step 2 and 3 were repeated for 100,000 times.

<u>Customized AT skew (mean)</u>
Sampled permutation size, s = 100,000
Mean of Sample F = 0.06627
Mean of Sample G = 0.0555
The difference in mean between sample F and G, (say T) = 0.01076
The sampled permutation values where the difference in means is greater than T = 294
P value = 294/100000 = 0.00294 < 0.01
So we can reject the null hypothesis.

<u>Customized AT skew (median)</u>
Sampled permutation size, s = 100,000
Median of Sample F = 0.0539
Median of Sample G = 0.04408
The difference in median between sample F and G, (say T) = 0.0099
The sampled permutation values where the difference in medians is greater than T = 742
P value = 742/100000 = 0.00742 < 0.01
So we can reject the null hypothesis.

<u>Customized GC skew (mean)</u>
Sampled permutation size, s = 100,000
Mean of Sample F = 0.05445

Mean of Sample G = 0.04537
The difference in mean between sample F and G, (say T) = 0.00908
The sampled permutation values where the difference in means is greater than T = 830
P value = 830/100000 = 0.00830 < 0.01
So we can reject the null hypothesis.

Customized GC skew (median)
Sampled permutation size, s = 100,000
Median of Sample F = 0.04099
Median of Sample G = 0.03304
The difference in median between sample F and G, (say T) = 0.00794
The sampled permutation values where the difference in medians is greater than T = 1441
P value = 1441/100000 = 0.01441 < 0.05
So we can reject the null hypothesis.

T-test for the slope of the model of Shannon's index and the frequency of phage words:

There are two samples: bacteria and phages.
The sample size of bacteria, m = 401
The linear model of bacterial sample: F = 5.85 H + 0.014 … (1)
The sample size of phages, n = 600
The linear model of phage sample: F = 8.57 H + 0.047 … (2)

We want to test whether the slope of these two equations are significantly different or not.

H$_0$: ß$_b$ = ß$_p$
H$_A$: ß$_b$ ≠ ß$_p$
where, ß$_b$ is the slope of bacterial sample and ß$_p$ is the slope of phage sample.

T test for two independent unequal sample sizes:

$$t = \frac{\bar{\beta}_p - \bar{\beta}_b}{\sqrt{SE(\beta_p)^2 + SE(\beta_b)^2}}$$ where, SE is the standard error.

$$t = \frac{8.9288\ -\ 5.708284}{\sqrt{0.0224^2 + 0.02371^2}} = 9.87$$

Degree of freedom = m-2 + n -2 = 997

In t table, for degree of freedom 1000 and p = 0.001, the value is 3.3. Our *t* = 9.87 > 3.3. So we can reject the null hypothesis.

That means the slope of the bacterial and phage samples are different.