High-throughput Phage Screening to Predict Pathogenicity of *E.coli* strains

Tatiana Lenskaia[1,*], Daniel Boley[1]
[1] Department of Computer Science and Engineering, University of Minnesota-Twin Cities Campus,
4-192 Keller Hall, 200 Union Street SE, Minnesota, MN 55455

[*]To whom correspondence should be addressed: lensk010@umn.edu

Abstract

Finding shared fragments between genomes is important for solving many biological challenges. Such fragments in microbial genomes suggest interactions between host bacteria and viral parasites helping to identify host-parasite associations and answer other critical questions about the organisms involved. Current methods can be supplemented by new computational technologies for versatile analysis of unannotated genomic string interactions. The goal of this study is to determine statistically significant genomic intersections that can imply important biological meaning. We explore how these intersections can be used to predict pathogenicity and distinguish between different *E.coli* strains.

We show the feasibility and usefulness of scalable computational algorithms to find pairs of organisms that interact with each other, such as bacteria-phage or host-parasite pairs, based on collected unannotated genome data. The statistical significance of the occurrence of matching strings is used to filter out matches possible due to chance. Our method, supplemented with machine learning techniques, can predict pathogenicity of bacterial strains using phage screening and profiling based on sequenced genomes without the need of annotation. We applied the algorithms to find "fingerprint" of phages interacting with bacterial hosts by analyzing 2,480 phage genomes from European Nucleotide Archive (ENA). The methods have adjustable sensitivity and specificity in identifying phages and provide bacterial "fingerprints" in terms of phage presence in microbial genomes with the desired level of resolution for evaluating pathogenicity of *E.coli* strains.

1 Introduction

1.1 Pathogenicity Prediction

Escherichia coli (*E.coli* for short) comes in many varieties. It can be a commensal bacterium that is a part of normal human intestinal microflora [Huttenhower *et al.*, 2012] or it can be highly pathogenic and cause severe infections in animals and humans [Kaper *et al.*, 2004]. It is also one of the most well-studied microbes in laboratory settings [Escherich, 1988; Raetz, 1996; Dunne *et al.*, 2017]. It is an important bacterium in biotechnology that can produce insulin [Goeddel *et al.*, 1979], biodiesel [Kalscheuer *et al.*, 2006] and other compounds.

Assessment of pathogenicity is very important for epidemiological [Rangel *et al.*, 2005; Grad *et al.*, 2012], food safety [Besser *et al.*, 1993; Scallan *et al.*, 2011], veterinary [Blanco *et al.*, 2001] and other health-related studies. As the cost of whole genome sequencing is decreasing, the availability of complete sequence genomes increases rapidly. Being able to quickly estimate a strain potential pathogenicity based just on its raw genome sequence would be an important advantage in diagnostics since it would save time and resources that otherwise would be necessary for wet-lab experiments and other resource-consuming techniques like multiple alignments.

The evolutionary transition to pathogenicity can result from acquiring different virulence factors when new genes responsible for producing toxins and other pathogenic components are incorporated. Although it is easy to determine presence of known genes in newly sequenced bacteria, some genes remain as unknown function. Bartoszek *et al.* (2018) created a model that was able to trace virulence factors based on persistence of trinucleotide repeats within clinical isolates. However, the presence of virulence genes itself does not necessarily result in pathogenicity [Wassenaar and Gunzer, 2015].

Bacteriophages (phage for short) are known for their contribution to the pathogenicity of bacteria [Penadés *et al.*, 2015] as well as to adaptive traits and diversification of bacterial strains. Since their survival depends on their success in infecting bacteria and getting viable progeny, phages must find and examine every possible flaw in a bacterial cell. By exploring existing remnants from many phages in bacterial genomes, we can evaluate the overall picture of the bacterial genome state.

Touchon *et al*. (2016) investigated associations of genetic and life-history traits in bacteria with distribution of prophages. They found slight correlation between pathogenicity and the number of prophages across different species. Presence of different types of pathogenicity patterns may blur the overall picture across different species of bacteria [Brüssow *et al.*, 2004]. In this research, we focus on evaluating the contribution of prophages to life traits within *E.coli* strains and exploring its predictive power on potential pathogenicity of individual strains of *E.coli*.

1.2 Identification of Shared Fragments between Host and Parasite Genomes

Genomes of different organisms might share extensive string fragments due to some biological reasons (e.g. temperate phages [Howard-Varona *et al.*, 2017], prophages and phage remnants [Touchon *et al.*, 2016]). Long shared fragments are a sign of a biological relationship between a host organism and a parasite. This may be a sign of an attack mechanism on the part of the parasite or a defense mechanism on the part of the host. By extracting and identifying shared genetic sequence fragments from bacteria and viruses, one can quickly distinguish between similar bacteria based on their functional behavior in the presence of phages, or quickly identify which viruses might be suitable as vectors with which to attack and destroy bacteria (phage therapy). This can also provide evidence of the historical evolution of bacteria and associated viruses. CRISPR-Cas defenses and other mechanisms based on recognizing substrings in viral genomes on the part of bacteria, and the mechanisms used by viruses to avoid recognition are still under investigation [Arber, 1978; Barrangou and van der Oost, 2013].

Identification of long genome fragments is a challenging task. Existing methods for detecting shared fragments between host and parasite genomes can be roughly divided into two categories: 1) "wet lab" *in-vitro* microbiological methods (hybridization capture sequencing, microarrays, etc.) [Kim *et al.*, 2012] and 2) "dry lab" *in-silico* computational methods (searching for the longest common substring, alignments, etc.).

"Wet lab" methods are technologically intensive, time-consuming, and have limited throughput. They may consist of hybridization assays, microarrays, polymerase chain reactions and work by detecting complementary base pairs between short segments of host genome and viral fragments (e.g. between human genome and retroviruses [Escalera-Zamudio and Greenwood, 2016], koala genome and retroviruses [Tsangaras *et al.*, 2014]). These methods often depend on the use of specific primers to locate and amplify target fragments.

"Dry lab" methods are computational and hence often very scalable and flexible. Computational approaches have been often successfully applied to analyze and distinguish between biological sequences [Grau *et al.*, 2012]. Currently, the main *in-silico* sources of information about phage incorporation into microbial genomes are annotated databases and software that make use of the annotations to identify bacteria-phage pairs. Examples include (1) special databases, e.g., ACLAME database [Leplae *et al.,* 2004] and PhAnToMe [http://www.phantome.org] and (2) computational tools that depend on annotated databases, e.g. Phage Finder [Fouts, 2006], Phaster [Arndt *et al.,* 2016], VirSorter [Roux *et al.,* 2015]. Many existing computational methods for bacteria-phage interaction depend on meta-data (e.g., coding regions, protein sequences, etc.) and external software solutions for localization of prophage regions, e.g. FASTA33 [Pearson, 1990], NCBI BLASTALL [Altschul *et al.,* 1997], HMMSEARCH [Eddy, 1998], MUMMER [Delcher *et al.,* 1999]. Unfortunately, the annotations are limited to those locations which have been explicitly identified to be of interest, making it difficult to identify possible new locations with unidentified segments.

An ever-growing quantity of genetic sequence data is being accumulated that is yet to be annotated or whose function is unknown. Existing annotations vary depending on goals of individual databases. Attempts to standardize annotation exist, such as NCBI Prokaryotic Genome Annotation Pipeline, but annotations may change as new discoveries are made. Touchon *et al.* (2016) used PhageFinder [Fouts, 2006] to predict phage incorporation. However, they mentioned that it was hard to distinguish phages and other mobile elements. To avoid ambiguity in identification of the origin of mobile elements, we use exact matching to known phage genomes. As reviewed in [Edwards *et al.*, 2015], substring matching is the most accurate method in terms of predicting host-parasite associations.

1.3 Exact Matching for Host-Parasite Interactions

We seek a computational screening method that works on raw genome assemblies and gives a common picture of candidate string interactions, without using any meta-data annotations. Such a method could be used on newly discovered bacterial variants, or on genome regions of unknown function. It needs to be scalable, sensitive to capture many interesting interactions, while specific enough to avoid false positives.

The method of "all common subsequences" (ACS) [Wang, 2007] is a computationally effective method that measures similarity relationship between sequences by extracting many common subsequences. Because the subsequences are not necessarily contiguous, this can suffer from ambiguities similar to those found in alignments. Although improved alignment methods are effective [Morgenstern, 1999] and currently developed alignment methods are very efficient [https://github.com/knights-lab/BURST], the statistical significance of finding a particular pattern with this method is not trivial to estimate.

Seeking "all common [contiguous] substrings" avoids this alignment ambiguity, and can be implemented efficiently using suffix trees and string kernels. [Leslie *et al.*, 2002] used a string kernel based on counts of short common substrings. Here we use a similar search for common substrings but consider much longer substring lengths to distinguish between biologically related and unrelated genomes.

1.4 Adjustment of Sensitivity and Specificity

Many newly created methods for detecting host-parasite associations based on string content demonstrate good performance. An in-depth review of existing methods to find host-parasite associations was done by [Edwards *et al.*, 2015].

Many such methods have been trained on specific datasets and locally optimized, however they sometimes suffer from a lack of specificity for large-scale screening research since they depend on statistical models on short substrings. For example, HostPhinder [Villarroel *et al.*, 2016] predicts based on 16-mers; WIsH [Galiez *et al.*, 2017] uses Markov models of order 8; [Zhang *et al.*, 2017] uses frequencies of 6-lettered words; VirFinder [Ren *et al.*, 2017] utilizes 8-mers. Although this lack of specificity can be compensated to some extent by considering additional factors (annotation, metadata, and biological knowledge etc.) to distinguish from true biologically relevant and random (biologically non-relevant) matches, it substantially limits the ability of tools to work on raw genomic data *en masse*.

We compute the common substrings for a variety of fixed lengths between a host genome and a phage genome. The lengths chosen are long enough so that unrelated organisms are very unlikely to show commonality, while short enough to occur often among biological organisms of interest (Section 2.2). The use of fixed length strings makes it easy to get good estimates of the statistical significance of the computed results based on simulation of a simple statistical model. Our computational strategy leads to a screening technique for fast and resource-effective preliminary analysis of host-parasite interactions in unannotated databases of complete genomes. It also allows the pairing of a given new bacterium with many phages to produce a sort of fingerprint for the bacterium, permitting rapid identification of new bacteria based on their functional interactions with phages. The strings lengths can be adjusted to yield a variety of levels of resolution, sensitivity, and specificity in the results.

These computational methods yield important information about statistically significant intersections between bacterial and viral genomes. These data can be analyzed by machine learning techniques to obtain important patterns of phage contribution to properties of bacterial strains.

We hypothesize that it is possible to estimate host pathogenicity based on genetic sequence overlap with a library of phages. We develop an efficient computational tool to test this hypothesis and validate it on a set of *E.coli* strains and associated phages. Our algorithms have been implemented in Python, eventually to be collected into a library of tools called "*PhageScreen*". For an individual host, the measures of overlap with a large collection of phages can be considered as a sort of functional "fingerprint" of the bacterial host, which can be assembled from raw genome sequence data. In this report, we demonstrate that the fingerprints (assembly of interaction levels with many phages) can distinguish between benign and pathogenic strains of *E.coli* and that these methods are then followed by machine learning can be used to predict pathogenicity in *E.coli*.

2 Methods

For the input, the algorithm takes complete genome sequences in FASTA format. The first step in our analysis is to assemble the dictionary of strings representing each individual organism and then to compute indices of pairwise overlap between each pair of organisms in question. These indices are then used as predictors of certain functional properties of bacterial hosts, specifically their pathogenicity. As results, the algorithm produces a classifier and returns classification results. These results are used to identify a list of phages that are most capable to distinguish between pathogenic and other strains. These indicator phages allow to reduce feature space without loosing accuracy of the classifier's performance.

Experimental procedure steps:

0. Choose appropriate string length *n* based on statistical simulations
1. Construct phage fingerprints (*PhageScreen*: pairwise indices between a host and phages):
    a. Assemble dictionary of all substrings of length *n* for each given raw genome.
    b. Compute intersection indices between bacterial and phage dictionaries.
2. Apply machine learning classifiers:
    a. Divide dataset into train and test sets for 10 fold cross-validation
    b. Train classifiers
    c. Test classifiers
3. Determine a set of "indicator" phages

2.1 Dictionary Assembly and Computation of Indices

As an input file, the algorithm takes unannotated genomes in FASTA format. For a given genome *G*, we obtain its dictionary D_n (*G*) by scanning the genome with a sliding window of length *n,* sequentially shifting it one nucleotide at a time. The results are assembled into a table of unique strings ("keys"). The **dictionary** D_n (*G*) consists of all distinct contiguous substrings of length *n* present in *G*. The size of the dictionary is the number of distinct substrings. We could also store the number of occurrences of each substring, but this information was not used in the computations reported in this paper.

The computational complexity for constructing a dictionary of a string length *n* for a genome *G* is *O*(n|G|). The dictionary is implemented using a hash table. We extract the string of length *n* at each position within a genome (we treat a genome as circular) using a sliding window, calculating its corresponding hash value. One could use a 'rolling hash' to compute all these hash values (after the first one) in time independent of *n* [Karp and Rabin, 1987], though this optimization was not needed for experiments reported here. The observed time complexity was found to be dominated by the lengths of the entire genomes, independent of *n*.

To find strings of a given length shared between two genomes (H – host, P – parasite), we need to compute a measure of the degree of intersection between two sets of dictionary keys by filtering out entries present in both dictionaries. We define the **index** $I_n^P(H)$ of relative presence of parasite genome *P* within host genome *H* to be the number of distinct common substrings of length *n* divided by the size of P's dictionary: $I_n^P(H) = (|D_n(H) \cap D_n(P)|)/|D_n(P)|$. We scan all the unique strings in the smaller genome, marking those that also appear in the larger genome. Using hash-tables, the computational complexity of this process is linear in the size of the smaller dictionary.

We use dictionaries, ordered pairs of keys (strings) and values (frequencies) as the primary data structure for implementation of our methods. This data structure is flexible, easy to implement and modify. It also allows us to keep track of string diversity for a variety of string lengths for common substrings between genomes. Dictionaries provide direct access to all substrings of a given length. For the results reported in this paper, it sufficed to implement the dictionaries as flat hash tables for easy access in time independent of the dictionary size. We have found it is simpler to store them individually as opposed to using a more sophisticated encoding such as suffix trees or some advanced methods of substring indexation. However, if we were to extend these methods to larger genomes (say human-scale), it may be necessary to modify the data structures. Using dictionaries, we can easily compute statistics and quickly obtain various counts (dictionary size, level of diversity, etc.) at intermediate stages of processing.

2.2 Determine Appropriate Window Size

We use a simple statistical model to determine the range of appropriate window sizes. An appropriate window size must be long enough to avoid string overlaps by pure random chance [specificity], but short enough to capture relations between organisms [sensitivity].

2.2.1 Statistical modeling to Determine Appropriate Window Size

We first estimate the probability of obtaining a non-empty intersection between two random genomes. To obtain estimates of the occurrence of non-empty intersections by chance, we repeatedly simulate the generation of dictionaries from randomly generated "genomes" 1024 times using IID with uniform distribution of nucleotides. This process is carried out for each length *n* of interest. Since the substrings of length *n* arise from a sliding window, they are not statistically independent, so they cannot be modelled by a simple statistical distributions over independent individual *bp*s, like a multinomial distribution. Hence we use a numerical simulation.

According to the results of numerical simulation, for any value of n up to 16, there is always some entry in the intersection. For any value of n ≥ 25, the observed probability of anything in the intersection is no greater than 0.0001. This gives us a threshold for a non-specific area. For values of n in a range from 17 to 24, intersection may be present or absent. The observed thresholds remain invariant (stable) within the range of analyzed *E.coli* genomes (4-6 Mbp) and viral genomes (2-500Kbp) (Fig.1) even as the GC content of the latter varied from 25% to 75%. According to the results of computational experiments, GC-content of phage genome has little effect on shifting the threshold. *E.coli* strains have GC-content close to 50% and the distribution of single nucleotide within *E.coli* genome is close to uniform. Thus, to make computations as simple as possible, we use a uniform distribution of nucleotides for both bacterial and viral genome to model the corresponding intersections.

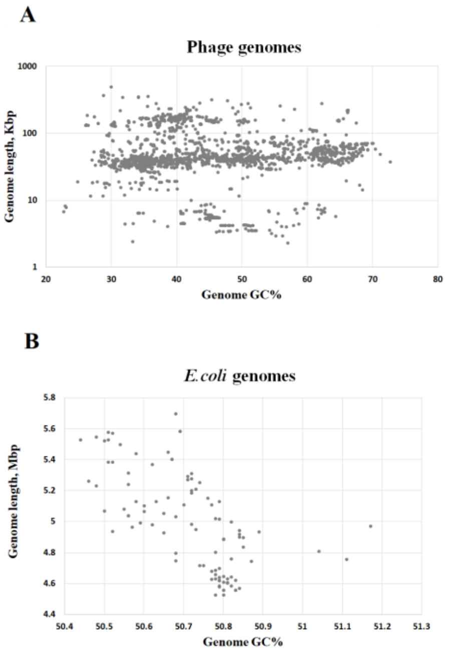

Fig.1. These diagrams represent the distribution of lengths and GC% for the analyzed genomes: (A) Phages; (B) *E.coli* strains.

2.2.2 Screening Sensitivity and Specificity

To demonstrate how sensitivity of the screening method varies as a function of string length, we counted the number of phages having non-empty intersection with two representative hosts, one for each class (*E.coli O157:H7* for pathogenic strains, *E.coli K12* for other strains), for different string lengths *n* (Fig.2). There are three areas: (I) non-specific area with high sensitivity; (II) “stable” sensitive and specific area; (III) specific area with low sensitivity. The string length must be above 25 bp to distinguish from random, but over 50 bp tends to lose sensitivity to some phages, hence the choice of *n* = 40 bp is appropriate. To evaluate specificity of our method we screened the two strains of *E.coli* against 4,743 viruses of eukaryotes (plant, animal, human, etc.) available in ENA http://www.ebi.ac.uk/ genomes/virus.html using the same string length 40 bp. We found only two viruses having non-empty intersection, albeit with small index values: *Vaccinia virus GLV-1h68* (EU410304) with *E.coli K12 (*index value $I_n^\mu(H) =$ 0.008825) and with *E.coli O157:H7* (index value .005842); and *Cyprinid herpesvirus 1 strain NG-J1* (JQ815363) with just *E.coli O157:H7* (index value .000024), in all cases with small index values. The presence of common strings of length as long as 40 bp even in such small amounts between these eukaryotes’ viruses and *E.coli* strains is very interesting and deserves further investigation. Such a small number (2 out of 4,743) of false positives for string length 40 bp is a high degree of specificity for the developed method.

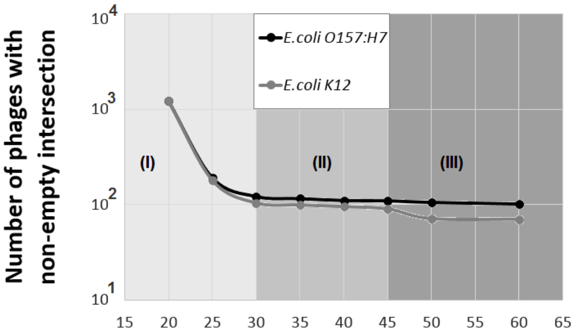

Fig. 2. Number of phages with non-empty intersection for *E.coli O157:H7* and *E.coli K12* for different string lengths: (I) high sensitivity; (II)” sensitivity plateau”; (III) decreasing sensitivity. According to the results of statistical modeling, area (I) is non-specific with high number of false positives; areas (II) and (III) are specific. After the threshold for *n*, specificity does not increase.

The choice of string length *n* represents a trade-off between sensitivity and specificity and depends on the specific genomes under study. Filtering phages based on intersections computed with multiple values of *n* might be appropriate for different host-parasite pairs with genomes of different lengths. This is a direction for future investigation. This choice of *n=40*bp provided a suitable balance between sensitivity and specificity for *E.coli* and associated phages. Our goal is to identify possible interactions of phages within host bacteria beyond simple defense mechanisms. Hence, we avoid false positives by choosing lengths above the typical spacer length in CRISPR-Cas loci of the bacterial hosts. According to the information retrieved from CRISPRdb http://crispr.i2bc.paris-saclay.fr/ accessed on May 8, 2018 more than 90% of known spacers in microorganisms have length below 40 bp.

2.3 Data Preparation

We computed phage profiles (*n* = 40 > 25 bp) for 2480 phage genomes with respect to 101 *E.coli* genomes available in ENA. We keep only those phages that have the non-empty statistically significant intersection with *E.coli* genomes. We found 172 phages that have their fragments inserted in at least one *E.coli* genome of interest. Within each *E.coli* genome, we found remnants of no less than 30 phages. Maximum number of phages which remnants were identified within a genome of sequenced *E.coli* strain using our screening method was 127. Using these profiles it is possible to compare phage contribution to *E.coli* genome and their impact on pathogenicity of different strains.

For each strain, we obtained information about its pathogenicity from the literature. We treat specific strain as potentially pathogenic if they were indicated to cause infection in animals or humans. Other strains include biotechnological strains, commensal strains obtained from healthy individuals, and laboratory strains.

2.4 Machine Learning Classifiers

We apply machine learning methods to investigate the possibility of inferring pathogenicity based on phage fingerprints. We use random forests [Breiman and Cutler, 2007] as a classifier since this algorithm has embedded feature selection, keeps only important features, and handles dimensionality well. It takes a bootstrap sample from the data and fits a classification tree. At each node, it randomly selects *m* features (mtry parameter) from all features in the data, finds the best possible split considering these *m* features, and grows the tree further. It uses voting for determining the best decision path based on the constructed trees. It provides out-of-bag (OOB) error to estimate the generalization error and evaluate future performance. OOB is computed based on analysis of a confusion matrix using permutation of features.

We used randomForest [Breiman *et al.*, 2018] and caret [Kuhn, 2018] packages in R. As input, the algorithm used phage profiles for 101 sequenced *E.coli* strains computed using *PhageScreen*. A phage profile is a vector that stores pairwise intersection index values for a phage genome and each of the selected *E.coli* genomes. Thus, the size of each feature vector is (101x1) and we have 172 such vectors (one for each phage). It constitutes our set of predictors. For each *E.coli* strain we have a pathogenicity label which takes value of 1 for pathogenic strains and 0 for other strains (commensal, laboratory, biotechnological).

The random forest classifier predicts the pathogenicity of a bacterial strain based on the fingerprint of phage remnants in its genome. To avoid overfitting and get a reasonable estimate of model performance on phage profile data, we used 10-fold cross-validation. We tried different cut-off for the *m* parameter, the number of features at each split, including the default value equal to the square root of the total number of features (*m*=13), half the default value (*m*=6), twice the default value (*m*=26), and the total number of features (m=172), to evaluate the relationship between prediction accuracy and the number of features necessary and sufficient to do the effective separation without overfitting.

To evaluate the contribution of features to the purity of separation on each step, the random forest algorithm [Breiman *et al.*, 2018] can compute the mean decrease in Gini coefficient closely related to AUC [Hand and Till, 2001] as a measure of information gain. We use the average value of this measure in 10 folds to get the list of phages arrange in decreasing order of their importance with respect to the purity of separation between pathogenic and other *E.coli* strains. The number of folds were experimentally determined as the one that provided better estimates for model parameters on this data.

We set different cut-offs for this list to determine the critical number of phages sufficient for proper classification of *E.coli* strains. Then we rebuild a prediction model on this reduced set of features and evaluate prediction accuracy. Finally, we used the cut-off that provides the highest level of accuracy as a reasonable estimate for the number of "indicator" phages.

3 Results

3.1 Window Size and String Length

We found a threshold on the string length n (p < 0.001 where p is a probability of finding non-empty intersection between host and parasite genomes by chance) to distinguish between random and biologically related shared fragments for genomes. For the reported strains of *E.coli*, this threshold equals 25 bp. Thus, we find a range of lengths starting at the threshold and extending to the length of a phage genome (at maximum) that allows us to analyze biologically important intersections between host and parasite genomes. We can vary the string length in this range for screening to obtain a desired level of specificity and sensitivity while analyzing shared fragments between genomes.

3.2 Prediction of Functional Properties: Pathogenicity

To estimate a possible difference in phage remnants between pathogenic and other strains of *E.coli*, we investigated two well-studied representatives of *E.coli* with available reference genomes: pathogenic – *O157:H7*, benign – *K12*. Fig.3 shows the phage presence in these two strains. We found 115 phages that have non-empty intersection with at least one of the two selected strains. Interestingly, we found that 91 of 115 (80%) phages were common for both bacteria. Spearman's correlation for the contributions of common phages between these two bacteria is 0.6 which suggest a strong positive correlation. The remaining 24 of 115 (20%) of phages were present only in one of the two bacteria (20 in *O157:H7* only, 4 in *K12* only).

Although the lists of found shared phage components were very similar, the amount of actual insertion for common phages were significantly different between the two strains. To estimate the difference between these amounts, we create distance metrics based on the sum of absolute differences between the intersection index values. On average, the pathogenic strain of *E.coli* has 60 times large values of the intersection index for common phages. This observation suggests positive association between pathogenicity and the amount of phage remnants within the host genome. We then investigated its predictive power over the entire set of bacterial hosts by machine learning.

Based on the computed overlaps between the 101 *E.coli* strains with the 172 phages (ignoring the 2308 phages that showed no overlaps at all), the random forest classifier yielded an average out-of-bag error rate in 10 folds of 12.84% ± 1.67%. The best average accuracy in 10 folds equaled 89.21% ± 10.68%, obtained with m=6. The average accuracy in 10 folds across different m values was 88.74% ± 0.04%.

3.3 Identification of Most Distinguishing Individual Phages

We seek for a small number of phages that is sufficient to do a complete classification. We called it "indicator" phages. To better understand the importance of features in making a decision on splits, we used mean decrease in Gini measure. Based on constructed random forests in 10 folds, we formed a list of the most important phages used by the algorithm to distinguish between pathogenic and other strains of *E.coli*. Considering the results for the random forests prediction model with the best accuracy achieved, we selected 6 most frequently used phages from this list. Then we reduced the list of features to those six phages and retrained the prediction model. The result is shown in Fig.3, with a slightly higher level of accuracy on the reduced set of features:

1) 91.94% ± 7.62% (6 phages, m = 3);
2) 89.21% ± 10.68% (all phages, m=6);

The results indicate that 6 is a suitable number of features to separate pathogenic and other strains. The six identified phages have similar genome length (51.44 ± 8.56 Kbp) and GC% (49.96% ± 1.30%). Five of the six phages belong to *Caudovirales*, they are dsDNA viruses: *Enterobacteria phage cdtI, Shigella phage Sf6, Stx2-converting phage 1717, Enterobacteria phage mEp460, Escherichia phage phi191.* The remaining virus is an unclassified bacterial virus: *Enterobacteria phage YYZ-2008*. It is worth noting that three of the identified phages have zero intersection (n=40) between each other indicating their mutual orthogonality in feature space. The rest three have overlaps that indicates their similarity. However, the existing differences between them make sufficient contribution to prediction accuracy (Fig.3).

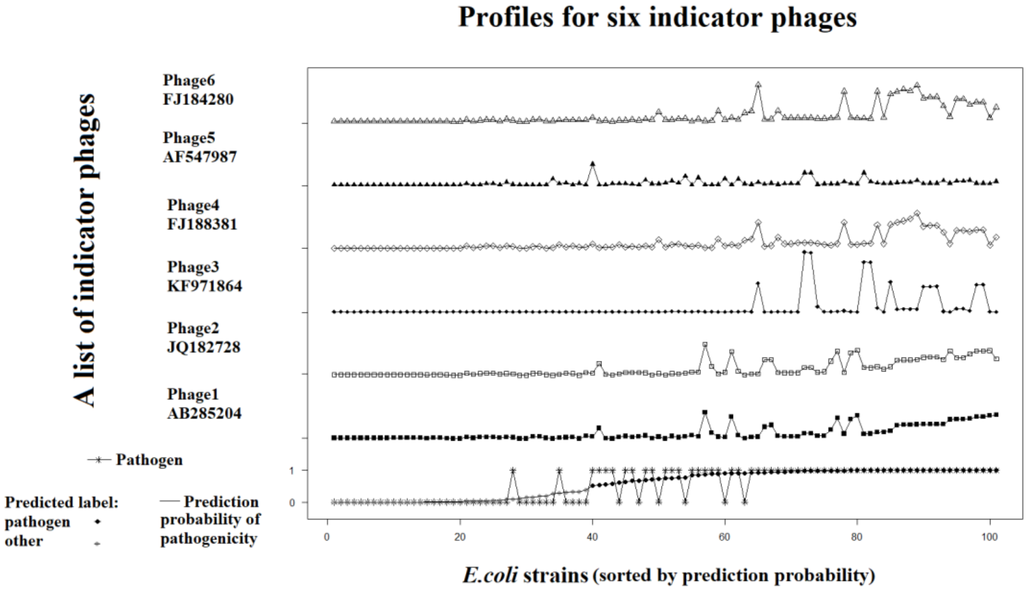

Fig. 3. Profiles for the top six phages and the results of random forests prediction on this reduced set of features. The *E.coli* strains have been sorted by predicted probability of pathogenicity (bottom line), based on these 6 phages. The bottom line also shows the ”true” pathogenicity (sec. 2.3). The remaining lines show the index values of each phage across the 101 *E.coli* strains.

Phages having statistically significant intersection with *E.coli* can help to distinguish between pathogenic and other strains using machine learning. Relationships between phages and *E.coli* hosts and between phages themselves are non-linear. Some phages have synergism and some exhibit antagonism. However, the number of sequenced phage genomes is sufficient for machine learning search of indicator phages to predict pathogenicity. It is possible to find a combination of phages among sequenced ones that provide a high prediction accuracy (>91%).

4 Discussion and Conclusion

4.1 Phage Fingerprints

We have shown that automated algorithms based on analysis of long unique shared strings applied to unannotated genome data can yield useful information about bacteria-phage interactions. We use statistical modeling for searching and investigating significant similarities between genomes in string diversity. It was possible to find a threshold above which it is virtually impossible to find unique strings common to two different unrelated genomes. Finding such pairs above the threshold strongly suggests an existing biological or evolutionary relationship between these genomes. Further investigation is needed to determine thresholds for genomes from other organisms, and to combine the filtering results from multiple string lengths.

We have used the screening method to construct a functional “viral fingerprint” for *E.coli* strains, where each fingerprint distinguishes between strains based on their evolutionary relationship to a wide variety of phages. Our analysis revealed the entire range of interactions between bacterial and phage genomes: no incorporation, partial incorporation, and almost complete incorporation of phages into microbial genomes. For example (Fig.4.), it was found that *Stx2 converting phage II* (AP005154) had index value I^{μ}_{40} above 0.9976 for *E.coli O157:H7* indicating almost complete incorporation of this phage into this host genome. We found that the index values for the pathogenic strain *E.coli O157:H7* are significantly higher than for the lab strain *E.coli K12*. It would suggest more active interactions between phages and pathogenic strains than laboratory strains. This might help in predicting pathogenicity of newly sequenced strains of *E.coli* based on phage occupancy of their genomes. The “wildness” of a bacterial strain (history of exposures to varied environments) might be predicted by a high degree of interaction with a variety of viruses, as indicated by high degree of virus incorporation. This warrants further investigation, including the possible use of overlap occurrence counts (not used in the present analysis).

Moreover, the collected indices of phages incorporated into bacteria can be considered as a fingerprint for the bacteria (Fig.4) in order to (1) classify distant strains with the help of common phages from area **(A)** of fig. 4; (2) identify and distinguish between closely related strains using the differences in their phage indices between areas **(B)** and **(C)**. The differences between **(B)** and **(C)** areas also can be applied for microbial typing as alternative to typing based on CRISPR loci [Briner and Barrangou, 2014]. Currently there is a trend in transiting from “wet lab” to “dry lab” methods to carry out voluminous tasks such as epidemiological studies [Chattaway *et al.*, 2017].

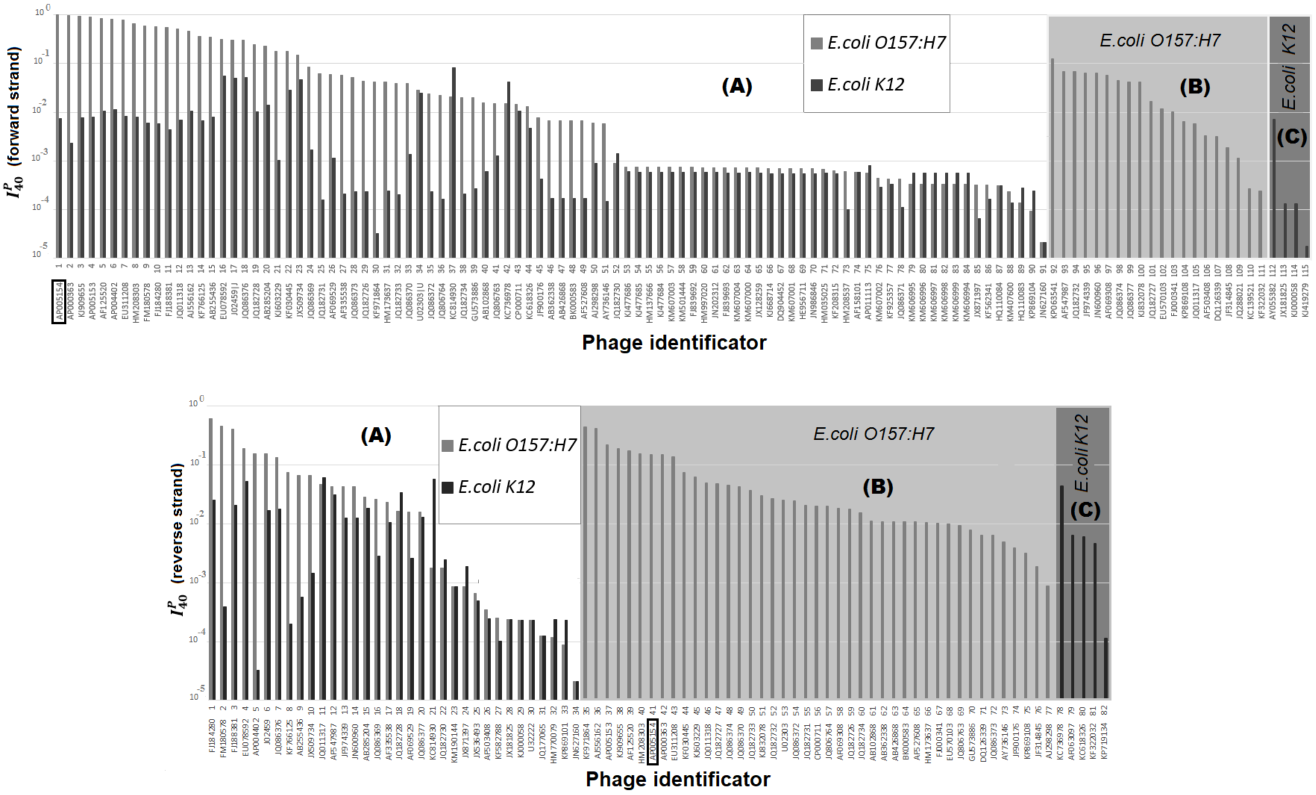

Fig. 4. Among 2,480 phages in ENA, 115 phages have non-empty intersection with *E.coli O157:H7* or *E.coli K12* on forward strand: 91 phage have common strings with both strains (area A), 20 phages have common strings only with *E.coli O157:H7* (area B), and 4 phages have common strings only with *E.coli K12* (area C). For reverse strand, 82 phages have non-empty intersection with *E.coli O157:H7* or *E.coli K12*: 34 phage have common strings with both strains (area A), 43 phages have common strings only with *E.coli O157:H7* (area B), and 5 phages have common strings only with *E.coli K12* (area C). Phages are arranged by decreasing order of the index of their presence in *E.coli O157:H7* (areas A and B) and in *E.coli K12* (area C). Black frames indicate a position *of Stx2 converting phage II* (AP005154) that is almost completely incorporated into *E.coli O157:H7* genome.

Based on screening results, this method can locate the most significant area(s) for fingerprints to fulfill specific research purposes. The potential range of capabilities for the algorithms proposed in this paper are limited primarily by the presence of virus and bacterial sequence data within databases.

In addition, such fingerprints allow us screen for phages with potentially high level of similarity (bars of equal length on the forward strand fingerprint in area **(A)**, Fig.4). Such phages deserve a close look at their mutual similarity. We can sort out these viruses based on the screening results and investigate the relationship in detail using similarity screening and alignment methods. The double impact of phages with resembling levels of similarities could be downweighted or excluded, depending on the variability. However, certain level of differences in content between highly similar phages might provide an important typing advantage.

4.2 Pathogenicity Prediction

We found that information about phage occupancy of host genomes is a good predictor of host potential pathogenicity. We applied machine learning techniques to predict pathogenicity of *E.coli* strains based on their phage spectra since currently databases contain sufficient amount of host and parasite genomes for this species. For each *E.coli* strain with sequenced genome available in ENA, we found at least 30 phages with sequenced genomes which fragments were identified in a host genome. With growing availability of other bacterial and viral sequenced genomes it is possible to expand this prediction approach to other species.

The presence of long common substrings of over a hundred phages in the *E.coli* strains indicates a significant pressure of those viruses on the host. The presence or absence of common substrings, computed in an automated way, can be used to distinguish bacteria phages from other viruses, identify particular phages that could be used as vectors against a very selective group of bacteria strains, and to distinguish between superficially similar bacteria based on differences between their evolutionary history and/or their putative functional interaction with their "viral environment".

We found that six "indicator" viruses are sufficient to distinguish between pathogenic and other *E.coli* strains (Fig.3). Since bacteria and viruses adapt quickly and they are able to change their genome rapidly (mutations, horizontal transfer, etc.), the detected indicator viruses have the best predictive power in relation to the current state of the analyzed bacterium genomes. However, the described approach allows to identify a relevant set of indicator viruses for genomes placed in different time frames and environmental conditions. This method is capable to reveal indicator phages for distinguishing between potentially pathogenic and other strains. It also might help to locate current pathogenicity hot spots in *E.coli* genomes.

In conclusion, we observed the interconnection between phage occupation of *E.coli* genomes and potential strain pathogenicity. We applied this to develop a computational "dry-lab" technology to predict pathogenicity of *E.coli* strains using phage screening of their unannotated sequenced genomes. The accuracy of the method will only increase with growing availability of sequenced viral and bacterial genomes in the databases.

4.3 Method possible applications

The methods proposed here do not depend on annotations. Due to exact matching, they are very specific and able to detect and distinguish between even closely related phages. This approach allows to accurately identify integration of phages into host genomes, but it has limited ability to detect interaction without such integration. It is currently optimized to detect integration of phages into host chromosomes, but it could be adapted to other types of genetic integration. The computational complexity of the methods is linearly related to the size of analyzed genomes which is an important advantage for a screening tool.

The methods here have potential in monitoring host-parasite interactions and tracking different trajectories of viral fragments inside microbial genomes: incorporation of certain fragments, further increment/decrement in a number of copies, and elimination of particular viral fragments. Currently our method works on unannotated complete genomes, but similar methods could potentially be developed to work on raw genome assemblies (scaffolds) and reads, to form a handy software screening tool for laboratory and medical applications, e.g. identifying prospective candidates for phage therapy and monitoring interactions between microbial and viral genomes during treatment.

Our approach allows us to detect a degree of viral incorporation into bacterial genomes with varied levels of resolution and with various goals. The resolution can vary in a range of string lengths above the threshold obtained by analysis of shared strings and evaluation of findings by statistical modeling. Differences in the values of the phage presence index suggest differences in the evolutionary history of genome interactions. For example, wild type *E.coli O157:H7* has generally higher values of the phage presence index I_{40}^{P} compared to artificial *E.coli K12* developed in a "sheltered" laboratory environment (Fig.4).

The selectivity will only grow as the databases grow and the methods are applied to ever wider classes of genomes. The methods can be used to screen genome sequence data semi-autonomously without any annotations. It can be useful as an early screening tool to find potential new biological interactions or selective interactions, as precursor to more in-depth validation *in-silico* with other meta-data or *in-vitro*.

Acknowledgments

This research was supported by NSF grant IIS 1319749. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NSF. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No other financial interests exist.

Conflict of Interest: none declared.

References

[Altschul *et al*., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402, 1997.

[Arber, 1978] Arber, W. Restriction endonucleases. Angewandte Chemie International Edition in English 17.2: 73-79, 1978.

[Arndt *et al*., 2016] Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart D.S. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research* 44.W1: W16-W21, 2016.

[Barrangou and van der Oost, 2013] Barrangou, R. and van der Oost, J., eds. CRISPR-Cas systems RNA-mediated adaptive immunity in Bacteria and Archaea. Springer-Verlag, Heidelberg,21-22, 2013.

[Bartoszek *et al.*, 2018] Bartoszek, K., Majchrzak, M., Sakowski, S., Kubiak-Szeligowska, A. B., Kaj, I., & Parniewski, P. Predicting pathogenicity behavior in Escherichia coli population through a state dependent model and TRS profiling. *PLoS computational biology*, *14*(1), e1005931, 2018.

[Besser *et al.*, 1993] Besser, R. E., Lett, S. M., Weber, J. T., Doyle, M. P., Barrett, T. J., Wells, J. G., & Griffin, P. M. An outbreak of diarrhea and hemolytic uremic syndrome from Escherichia coli O157: H7 in fresh-pressed apple cider. *Jama*, *269*(17), 2217-2220, 1993.

[Blanco *et al*., 2001] Blanco, J., Blanco, M., Blanco, J. E., Mora, A., Alonso, M. P., González, E. A., & Bernárdez, M. I. Epidemiology of verocytotoxigenic Escherichia coli (VTEC) in ruminants. *Verocytotoxigenic E. coli*, *113*, 2001.

[Breiman *et al*., 2018] Breiman, L., Cutler, A., Liaw, A., & Wiener, M. Breiman and Cutler's Random Forests for Classification and Regression. Retrieved from https://cran.rproject.org/web/packages/randomForest/ran domForest.pdf, 2018.

[Breiman and Cutler, 2007] Breiman, L., & Cutler, A. Random forests-classification description. *Department of Statistics, Berkeley, 2*, 2007.

[Briner and Barrangou, 2014] Briner, A.E. and Barrangou, R. Lactobacillus buchneri genotyping on the basis of clustered regularly interspaced short palindromic repeat (CRISPR) locus diversity. *Applied and environmental microbiology* 80.3: 994-1001, 2014.

[Brüssow *et al*., 2004] Brüssow, H., Canchaya, C., & Hardt, W. D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and molecular biology reviews*, *68*(3), 560-602, 2004.

[Chattaway *et al*., 2017] Chattaway, M. A., Schaefer, U., Tewolde, R., Dallman, T. J., and Jenkins, C. Identification of Escherichia coli and Shigella Species from Whole-Genome Sequences. *Journal of clinical microbiology*, 55(2), 616-623, 2017.

[Cowley *et al*., 2015] Cowley, L. A., Beckett, S. J., Chase-Topping, M., Perry, N., Dallman, T. J., Gally, D. L., and Jenkins, C. Analysis of whole genome sequencing for the Escherichia coli O157: H7 typing phages. *BMC genomics*, 16(1), 271, 2015.

[Delcher *et al*., 1999] Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., & Salzberg, S. L. Alignment of whole genomes. *Nucleic acids research*, *27*(11), 2369-2376, 1999.

[Dunne *et al*., 2017] Dunne, K. A., Chaudhuri, R. R., Rossiter, A. E., Beriotto, I., Browning, D. F., Squire, D., ... & Henderson, I. R. Sequencing a piece of history: complete genome sequence of the original Escherichia coli strain. *Microbial genomics*, *3*(3), 2017.

[Eddy, 1998] Eddy, S. R. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, *14*(9), 755-763, 1998.

[Edwards *et al*., 2015] Edwards, R. A., McNair, K., Faust, K., Raes, J., and Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS microbiology reviews*, 40(2), 258-272, 2015.

[Escalera-Zamudio and Greenwood, 2016] Escalera-Zamudio, M. and Greenwood, A.D. On the classification and evolution of endogenous retrovirus: human endogenous retroviruses may not be ‘human’after all. *Apmis* 124.1-2: 44-51, 2016.

[Escherich, 1988] Escherich, T. The intestinal bacteria of the neonate and breast-fed infant. *Clinical Infectious Diseases*, *10*(6), 1220-1225, 1988.

[Fouts, 2006] Fouts, D. E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic acids research* 34.20: 5839-5851, 2006.

[Galiez *et al*., 2017] Galiez, C., Siebert, M., Enault, F., Vincent, J., and Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, 33(19), 3113-3114, 2017.

[Goeddel *et al*., 1979] Goeddel, D. V., Kleid, D. G., Bolivar, F., Heyneker, H. L., Yansura, D. G., Crea, R., ... & Riggs, A. D. Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences*, *76*(1), 106-110, 1979.

[Grad *et al*., 2012] Grad, Y. H., Lipsitch, M., Feldgarden, M., Arachchi, H. M., Cerqueira, G. C., FitzGerald, M., ... & Sykes, S. Genomic epidemiology of the Escherichia coli O104: H4 outbreaks in Europe, 2011. *Proceedings of the national academy of sciences*, *109*(8), 3065-3070, 2012.

[Grau *et al*., 2012] Grau, J., Keilwagen, J., Gohr, A., Haldemann, B., Posch, S., and Grosse, I. Jstacs: a java framework for statistical analysis and classification of biological sequences. *Journal of Machine Learning Research*, 13(Jun), 1967-1971, 2012.

[Hand and Till, 2001] Hand, D. J., & Till, R. J.. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine learning, 45(2), 171-186.

[Howard-Varona *et al*., 2017] Howard-Varona, C., Hargreaves, K. R., Abedon, S. T., and Sullivan, M. B. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *The ISME Journal*, 11(7), 1511-1520, 2017.

[Huttenhower *et al*., 2012] Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., ... & Giglio, M. G. Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207, 2012.

[Kalscheuer *et al*., 2006] Kalscheuer, R., Stölting, T., & Steinbüchel, A. Microdiesel: Escherichia coli engineered for fuel production. *Microbiology*, *152*(9), 2529-2536, 2006.

[Kaper *et al*., 2004] Kaper, J. B., Nataro, J. P., & Mobley, H. L. Pathogenic escherichia coli. *Nature reviews microbiology*, *2*(2), 123, 2004.

[Karp and Rabin, 1987] Karp, R. M., and Rabin, M. O. Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, 31(2), 249-260, 1987.

[Kim *et al*., 2012] Kim, J. H., Kalitsis, P., Pertile, M. D., Magliano, D., Wong, L., Choo, A. and Hudson, D. F. Nucleic Acids: Hybridisation. *eLS*. Web, 2012.

[Kuhn, 2018] Kuhn, M. Classification and Regression Training. Retrieved from https://cran.r-project.org/web/packages/caret/caret.pdf, 2018.

[Leplae *et al*., 2004] Leplae, R., Hebrant, A., Wodak, S. J., & Toussaint, A. ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic acids research*, *32*(suppl_1), D45-D49, 2004.

[Leslie *et al*., 2002] Leslie, C.S., Eskin, E., and Noble, W.S. The spectrum kernel: A string kernel for SVM protein classification. Pacific symposium on biocomputing. Vol. 7,7, 2002.

[Morgenstern, 1999] Morgenstern, B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* (Oxford, England), 15(3), 211-218, 1999.

[Pearson, 1990] Pearson W.R. Rapid and sensitive sequence comparison with FASTP and FASTA *Methods Enzymol* 18363–98, 1990.

[Penadés *et al*., 2015] Penadés, J. R., Chen, J., Quiles-Puchalt, N., Carpena, N., & Novick, R. P. Bacteriophage-mediated spread of bacterial virulence genes. *Current opinion in microbiology*, *23*, 171-178, 2015.

[Raetz, 1996] Raetz, C. R. H. Escherichia coli and Salmonella cellular and molecular biology. *Escherichia coli and Salmonella: Cellular and Molecular Biology*, *1*, 1035-1063, 1996.

[Rangel *et al*., 2005] Rangel, J. M., Sparling, P. H., Crowe, C., Griffin, P. M., & Swerdlow, D. L. Epidemiology of Escherichia coli O157: H7 outbreaks, united states, 1982–2002. *Emerging infectious diseases*, *11*(4), 603, 2005.

[Rasko *et al*., 2008] Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., ... & Henderson, I. R. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of bacteriology*, *190*(20), 6881-6893, 2008.

[Ren *et al*., 2017] Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1), 69, 2017.

[Roux *et al*., 2015] Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3: e985, 2015.

[Scallan *et al*., 2011] Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., ... & Griffin, P. M. Foodborne illness acquired in the United States—major pathogens. *Emerging infectious diseases*, *17*(1), 7, 2011.

[Touchon *et al*., 2016] Touchon, M., Bernheim, A., and Rocha, E. Genetic and life-history traits associated with the distribution of prophages in bacteria. *The ISME journal* 10.11: 2744-2754, 2016.

[Tsangaras *et al*., 2014] Tsangaras, K., Siracusa, M.C., Nikolaidis, N., Ishida, Y., Cui, P., Vielgrader, H., Helgen, K.M., Roca, A.L., and Greenwood, A.D. Hybridization capture reveals evolution and conservation across the entire koala retrovirus genome. *PLoS One* 9.4 (2014): e95633, 2014.

[Villarroel *et al*., 2016] Villarroel, J., Kleinheinz, K. A., Jurtz,V.I., Zschach, H., Lund,O., Nielsen, M., and Larsen, M. HostPhinder: a phage host prediction tool. *Viruses*, 8(5), 116, 2016.

[Wang, 2007] Wang, H. (2007) All Common Subsequences. Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 635–640. Web. IJCAI'07, 2007.

[Wassenaar and Gunzer, 2015] Wassenaar, T. M., & Gunzer, F. The prediction of virulence based on presence of virulence genes in E. coli may not always be accurate. *Gut pathogens*, *7*(1), 15, 2015.

[Zhang *et al*., 2017] Zhang, M., Yang, L., Ren, J., Ahlgren, N. A., Fuhrman, J. A., and Sun, F. Prediction of virus-host infectious association by supervised learning methods. *BMC bioinformatics*, 18(3), 60, 2017.