

1 **The Use of Informativity in the Development of Robust Metaviromics-based Examinations**

2

3 Siobhan C. Watkins¹, Thomas Hatzopoulos², and Catherine Putonti^{1,2,3*}

4

5 ¹ Department of Biology, Loyola University Chicago, Chicago, IL, United States of America

6 ² Department of Computer Science, Loyola University Chicago, Chicago, IL, United States of America

7 ³ Bioinformatics Program, Loyola University Chicago, Chicago, IL, United States of America

8

9 * Corresponding author

10 E-mail: cputonti@luc.edu (CP)

11

12 *Short Title: Integrating Informativity in Metaviromics Analysis*

13

14 **Abstract**

15 The field of metagenomics has developed insight into many of the complex microbial communities
16 responsible for maintaining life on this planet. Sequencing efforts often uncover novel genetic content;
17 this is most evident for viral metagenomics, in which upwards of 90% of all sequences demonstrate no
18 sequence similarity with present databases. For the small fraction which can be identified, the top BLAST
19 hit is often posited as being representative of the phage taxon. However, as previous research has
20 shown, the top BLAST hit is sometimes misinterpreted. Furthermore, the appearance of a particular
21 gene homolog is frequently not representative of the presence of the particular taxon in question. To
22 circumvent these limitations, we have developed a new method for the analysis of metaviromic
23 datasets. BLAST hits are weighted, integrating the sequence identity and length of alignments as well as
24 a phylogenetic signal. A genic rather than genomic approach is presented in which each gene is
25 evaluated with respect to its information content. Through this quantifiable metric, predictions of viral
26 community structure can be made with greater confidence. As a proof-of-concept, the approach
27 presented here was implemented and applied to seven metaviromes. While providing a more robust
28 means of evaluating metaviromic data, the tool is versatile and can easily be customized to
29 investigations of any environment or biome.

30

31

32 Background

33 Bacterial viruses (bacteriophages) play an essential role in shaping microbial populations. They drive
34 community structure through the mediation of mortality, and shape diversity – fundamentally – through
35 their role as agents of genetic mobility (Wilhelm & Suttle, 1999; Canchaya et al., 2003; Beredjeb et al.,
36 2011; Clokie et al. 2011; Winget et al., 2011; Brum et al. 2016). Their impact has been described at
37 higher trophic levels (Rohwer & Thurber, 2009; Jover et al., 2014); phages affect microbial processes on
38 a global scale. In addition to their influence in the environment, evidence has uncovered that phages can
39 contribute to human disease (e.g. Holmes 2000) and may play a role in human health as part of the
40 human microbiome (e.g. Willner et al., 2012). Whole genome sequencing (WGS) inquiries of complex
41 viral communities (metaviromics) have been pivotal in ascertaining both the ubiquity of phages as well
42 as the sheer number of phages on Earth (Edwards & Rohwer, 2005). As such, a wide variety of
43 environments have been probed, from the world’s oceans (Hurwitz & Sullivan, 2013) to extreme
44 environments (Gudbergssdottir et al., 2015); from deserts (Fancello et al., 2013), to the human gut
45 (Minot et al., 2013).

46 In contrast with cellular organisms, no conserved coding regions are ubiquitous among all viral species.
47 Efforts to utilize genes coding for structural proteins have given limited insight into the diversity of
48 defined communities of phages (Dorigo, Jacquet & Humbert, 2004; Wilhelm et al., 2006). Similarly, DNA
49 polymerases have been used as markers for specific groups of phages (Breitbart, Miyake & Rohwer,
50 2004). However, the study of viral communities based on the examination of whole genomes is widely
51 considered to be the most robust approach to exploring phage diversity in the environment. The
52 approach taken for analyzing WGS data sets within metaviromics has paralleled that of metagenomics of
53 bacterial and archaeal populations – reads or contigs are compared to known, characterized sequences
54 within public data repositories. Although a powerful tool, the generation of metaviromic surveys, a
55 literal “who’s who” of the communities present, is confounded by bioinformatic challenges unique to
56 the examination of phages. Currently, only a small fraction of the genetic diversity that phages represent
57 is characterized – and it is certainly likely that the large gaps in our knowledge define key processes.
58 However these general gaps are translated directly from the genome level; most characterized phages
59 contain a surfeit of genes for which there are no known homologs (Hatfull, 2008). In addition, the
60 current collection of characterized genomes is sparse; presently, there are just over 2000 phage
61 genomes deposited in RefSeq, and strains that infect laboratory bacterial models are overrepresented.
62 Therefore, phages represent a remarkable reservoir of undiscovered genetic diversity (Suttle, 2007).

63 For the few viral species which can be identified, typically via BLAST searches, the single best hit is often
64 posited as being representative of the phage taxon containing the homologous region: a method
65 employed by many metagenome studies and analysis tools (e.g., Huson & Weber, 2013; Wommack et
66 al., 2012; Keegan et al., 2016; Roux et al., 2014). This approach, however, can be misleading; genes
67 present within annotated phage genomes may not be true indicators of the phage species. For instance,
68 such genes may be bacterial in origin (e.g. Mann et al., 2003; Thompson et al., 2011; Thompson et al.,
69 2011; Lindell et al., 2005; Gao, Gui & Zhang, 2012). Thus hits to such genes would be indicative of either
70 bacterial DNA within the sample sequenced or acquisition of the bacterial genome (which need not be
71 exclusive to the taxa represented in the sequence data repositories). In a recent metaviromic survey of

72 the nearshore waters of Lake Michigan, further investigation of viral species with the most “hits”
73 revealed that the matches were localized to a particular gene(s) within the genome, and therefore
74 indicative of the presence of a specific gene rather than that of the species (Watkins et al., 2015).
75 Moreover, as was the case with one of these phages – Planktothrix phage PaV-LD, BLAST results were
76 indicative of the presence of bacterial genes. Several Planktothrix phage genes exhibit greater sequence
77 similarity to bacterial proteins rather than other phage sequences (Gao, Gui & Zhang, 2012). Over half of
78 the publicly available datasets in the viral metagenomic sequence web server MetaVir (Roux et al., 2014)
79 include hits to this phage (including samples unlikely to harbor the phage’s cyanobacterial host species),
80 indicating that misreporting is widespread. Thus, a “BLAST and go” approach for species identification
81 must be replaced by a more rigorous assessment of each individual BLAST hit result.

82 Herein we present a new, quantifiable, method for assessment of BLAST results, in an attempt to
83 address the aforementioned challenges. This approach can be applied to all studies, regardless of the
84 niche under investigation, as sequence similarity to databases is weighted. Weighting takes into
85 consideration not only the sequence identity between the metavirome contig and the database record,
86 but also the length of the alignment, and more importantly the informativity of the match. This latter
87 metric captures the taxonomic signal within sequence similarity results. Thus, a species’ presence or
88 absence within a population can be determined with greater confidence. As a proof-of-concept, we
89 examined seven publicly available freshwater DNA metagenomic datasets.

90 **Materials and Methods**

91 *Viral gene datasets.* Sequence data were retrieved from NCBI in January 2016. For the analysis of
92 Pbnalikeviruses, amino acid and nucleotide sequences for the Pbnalikeviruses Pseudomonas phage
93 PB1 (Accession Number: NC_011810) and Burkholderia phage BcepF1 (Accession Number: NC_009015).
94 All phage nucleotide sequences (omitting those belonging to the Pbnalikevirus genomes listed in
95 Supplemental Table 2) were retrieved through an advanced search via the NCBI website with the
96 following query: PHG[Division] NOT (txid538398[Organism] AND ...) in which the list of Pbnalikeviruses
97 were removed from the search by their taxonIDs (as indicated by “...”). In total over 500000 individual
98 records were retrieved.

99 *Metaviromic datasets.* SRA records were collected from the SRA database. Supplemental Table 1 lists all
100 of the datasets included in the proof-of-concept study. Each SRA record (line listed in the Supplemental
101 Table 1) was considered as an individual sample. (Note, two samples are aggregates of more than one
102 SRA record, both belonging to Metavirome IV, as they were combined in the downloadable file from
103 SRA.) Each individual sample was next assembled using Velvet (Zerbino & Birney, 2008) with a hash size
104 of 31. PB1 protein sequences were directly compared to these assembled contigs, rather than raw
105 reads, via blastx.

106

107

108

109 **Table 1.** Freshwater DNA metaviromic studies retrieved from NCBI's SRA database.

Metavirome	Environmental Niche	Number of Samples	Sequencing Technology	Mbp Total	Reference
I	Lake Michigan nearshore	40	Illumina	6 909	Watkins et al., 2015; Sible et al., 2015
II	Lake Bourget	2	454	698	Roux et al., 2012
III	Kent SeaTech tilapia pond	3	454	47	Dinsdale et al., 2008
IV	Lake Limnopolar	2	454	18	López-Bueno et al., 2009
V	Reclaimed water samples	6	454	364	Rosario et al., 2009
VI	Lake Ontario	3	454	223	n/a
VII	Feitsui Reservoir	5	454	86	Tseng et al., 2013

110

111 **Results and Discussion**

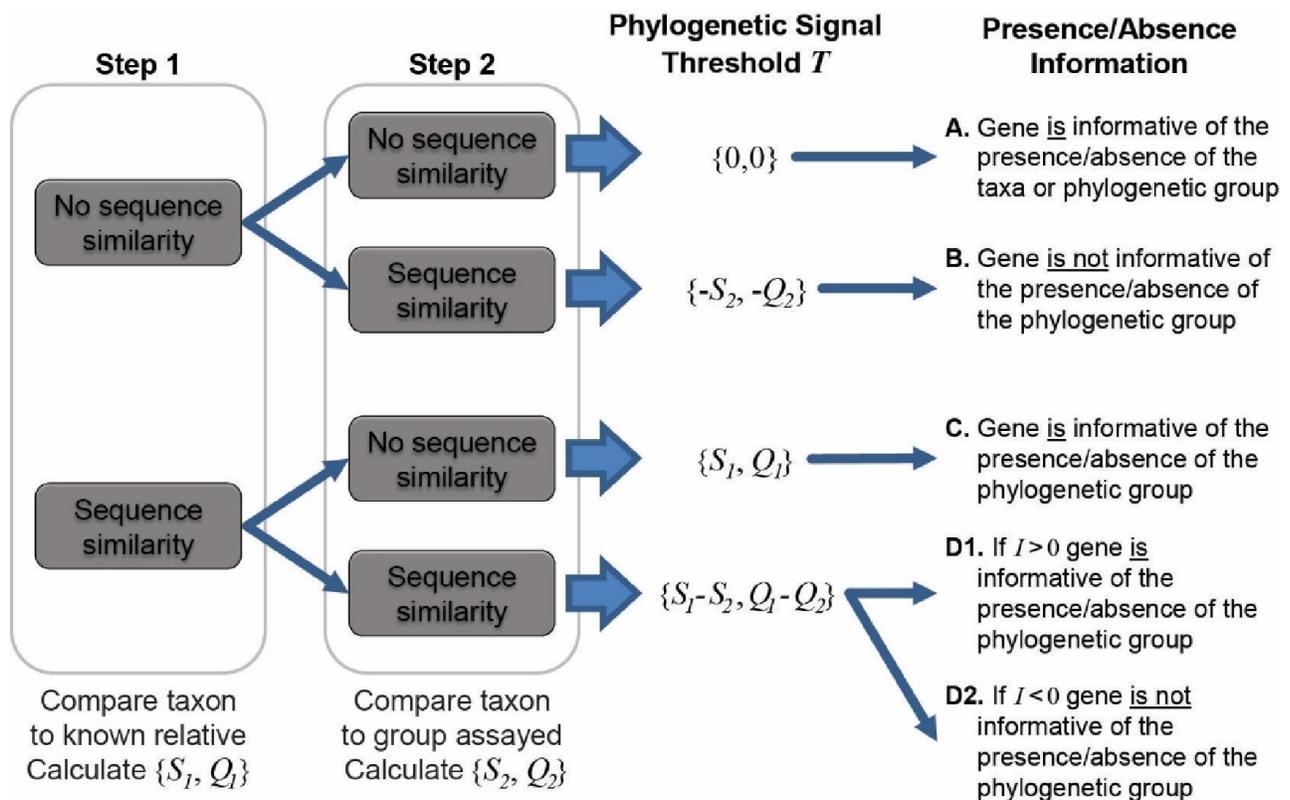
112 **Determination of Informativity Metric for Quantifying Hits**

113 **Establishing a Phylogenetic Signal Threshold.** To ascertain the presence/absence of specific taxon
 114 within a metagenome, we suggest a threshold to differentiate between informative and uninformative
 115 hits. The phylogenetic signal threshold T is determined through a two-step process prior to evaluation of
 116 the metagenomic data. Firstly, for a given taxon of interest, each annotated coding region is compared
 117 to all annotated sequences within the genome of a known relative. Thus, each coding region's sequence
 118 $x (x \in X$, where X is the set of sequences for all coding regions annotated within the genome of the taxon
 119 of interest) is compared to each coding region's sequence $g (g \in G$, where G is the set of sequences for
 120 all coding regions annotated within the genome of a known relative). The use of a known relative
 121 genome establishes if and how conserved the coding region is between known, related strains/species.
 122 Where sequence homology is detected, the sequence identity and query coverage of the match is
 123 recorded: S_1 and Q_1 , respectively.

124 In the second step, each coding region's sequence is compared again, this time to the sequences for all
 125 annotated coding regions for the group assayed by the metagenomic study (e.g. phages, all viruses,
 126 bacteria, archaea, etc.), however, those belonging to the phylogenetic group containing the taxon of
 127 interest and the known relative considered in step one are omitted. Many hits may be recorded for a
 128 particular gene x . Thus the best hit, both with respect to the sequence identity and the query coverage
 129 of the match, is selected; S_2 and Q_2 denote this best match's sequence identity and query coverage,

130 respectively. A phylogenetic signal threshold T is defined as $T=\{S_1-S_2, Q_1-Q_2\}$ where the subscripts 1 and
 131 2 represent the sequence identity and query coverage of the match detected from steps one and two,
 132 respectively. Figure 1 illustrates the two-step process, the T values produced.

133 It is important to note, that the phylogenetic group used for comparison is user defined. For instance, in
 134 order to ascertain if a gene can be used to distinguish between the presence/absence of a particular
 135 species, one may consider the phylogenetic group to be inclusive only of strains of the species.
 136 Therefore in this case, the most distant relative belonging to the phylogenetic group in step one would
 137 be the closest related species. If a more distant relative, say the most distantly related species of the
 138 same genus, were to be investigated, then the phylogenetic signal threshold T would serve as a means
 139 to distinguish between the presence/absence of a subset of the species (inclusive of the taxon of
 140 interest) within the genus. This flexibility enables the researcher to define and control the granularity of
 141 his/her analyses. In addition to the intended purpose of establishing the phylogenetic signal threshold,
 142 the two-step process can provide insight into putative horizontally acquired elements and gene loss
 143 events within a phylogenetic group. For example, instances in which the gene did not include a homolog
 144 in the most distant relative but did exhibit sequence similarity to a gene within the genome of another
 145 phylogenetic group. Furthermore, the two-step process can identify genomes which have been
 146 taxonomically misclassified - such instances would result in high S_2 and Q_2 scores for a large majority of
 147 the genes.



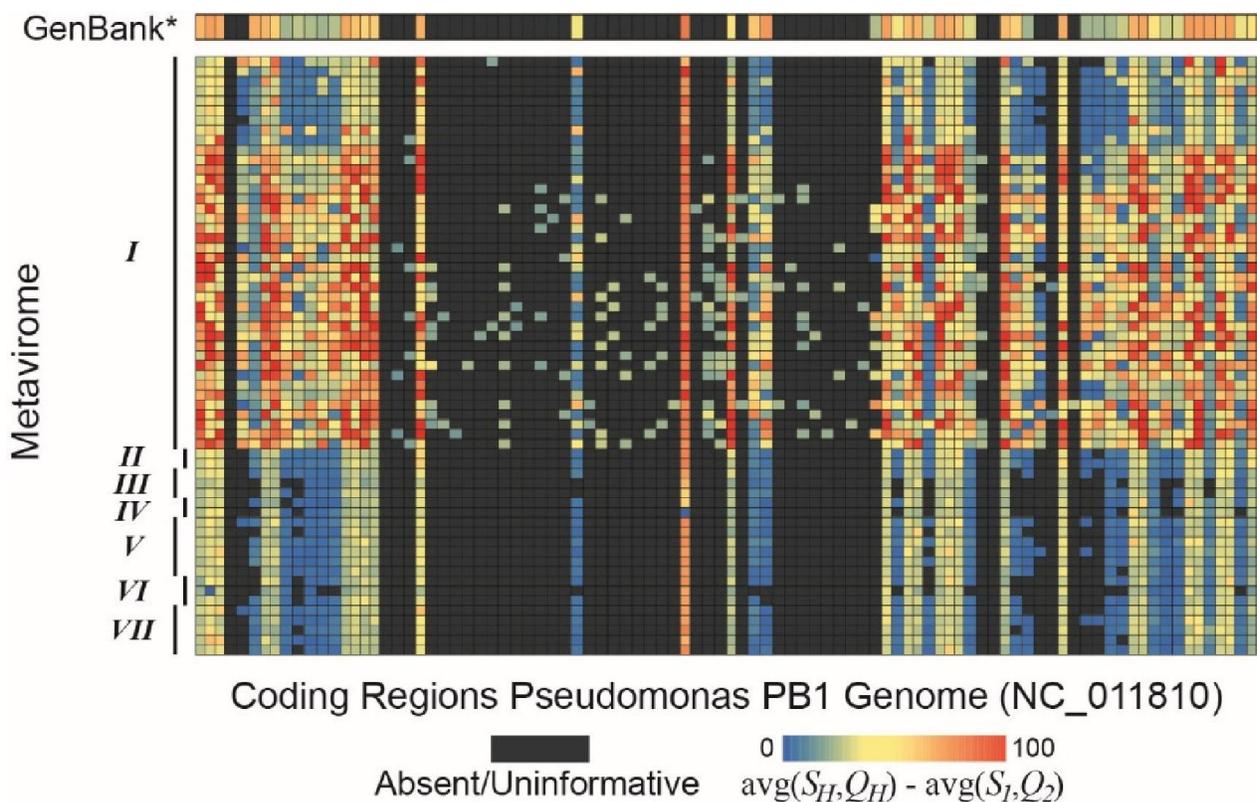
148

149 **Figure 1.** Two-step process for determining the phylogenetic signal threshold T and the information
 150 which can be gained regarding the presence/absence of a taxon's phylogenetic group. S_1 and S_2

151 represent the sequence identity of homologies identified in step 1 and 2, respectively. Likewise, Q_1 and
152 Q_2 refer to the query coverage of the match detected in step 1 and 2, respectively.

153

154 **Using Informativity to Ascertain Confidence in OTU Calls.** As indicated in Figure 1, when the set T is
155 greater than or equal to zero (outcomes A, C, and D1), the presence of a specific gene can provide
156 insight. OTU calls are informed by this threshold to decipher BLAST analyses of metaviromic datasets as
157 some hits may be to genes which are conserved and thus poor indicators of a species' or taxa's presence
158 or absence. For a given "hit" within a metaviromic dataset, the sequence identity and query coverage,
159 S_H and Q_H respectively, is assessed relative to the phylogenetic signal threshold T for the gene
160 producing the match. Genes in which $T < 0$ have already been classified as uninformative (Figure 2).
161 Now hits which fall below the gene's threshold, $\{S_H, Q_H\} - T < 0$, are also classified as uninformative. Hits
162 which are above the threshold are considered informative. The informativity I of each hit is quantified
163 based upon deviation from this threshold T such that $I = \{S_H, Q_H\} - T$. I can range from 0 (equivalent to the
164 threshold T) to 100 ($T = \{0, 0\}$, $S_H = 100\%$ sequence identity and $Q_H =$ query coverage of the gene). Thus
165 genes with a large value of I are strong indicators of the presence of a particular taxon.



166

167 **Figure 2.** BLAST hits to PB1 genes within both the set of non-Pbunlikevirus viral genomes and seven
168 freshwater DNA metaviromic datasets (Table 1). Hits (S_H and Q_H) are qualified relative to the sequence
169 similarity shared between PB1 and its distant Pbunlikevirus relative, Burkholderia phage BcepF1 (S_I and
170 Q_2).

171 **Implementation**

172 The posited method for assessing the informativity of metagenomic hits was implemented using a series
173 of BLAST databases and BLAST searches. First, a collection of all coding regions (either nucleotide or
174 amino acid sequences) were retrieved for the taxon of interest (X) as well as all genes annotated within
175 the user defined genome of the selected relative (G). A local BLAST database was created for G , and the
176 genes belonging to X were queried against the local database. The sequence identity and query
177 coverage of the match detected for the best hit for each gene was then parsed from the BLAST results
178 quantifying each gene's S_1 and Q_1 values. Next, a BLAST database was created using all characterized,
179 annotated sequences other than those associated with the phylogenetic group. Each of the genes for
180 the taxon of interest X was queried against the second local database; the results were again parsed for
181 each gene's S_2 and Q_2 values so that the phylogenetic signal threshold T could be calculated.

182 A metagenomic dataset was next evaluated, comparing each read or contig against a collection of
183 annotated gene sequences. While we implemented this step locally, users with limited computational
184 resources can utilize a resource such as MG-RAST (Keegan, Glass & Meyer, 2016), MEGAN (Huson &
185 Weber, 2013), VIROME (Wommack et al., 2012), or MetaVir (Roux et al., 2014) and use the remotely
186 generated BLAST results produced for further analysis here. Each BLAST hit was next assessed with
187 respect to its scores $\{S_H, Q_H\}$ relative to that of the gene's threshold T . Informative results were written
188 out to file, including the values of I , T , and $\{S_H, Q_H\}$. The user can then evaluate the likelihood of a
189 particular taxon or phylogenetic group's presence within the metagenomic sample based upon the I
190 values for informative genes, as described. Taking into consideration the number of informative genes
191 detected within a metagenomic sample and their individual I values can leverage additional confidence in
192 calling OTUs.

193 The described process has been automated via a Python script and calls to commands within the BLAST+
194 command line application. Users must supply or specify the fasta format files for the taxon of interest
195 (X), the genome of a known relative (G), and the group assayed (less the taxonomic group of interest). If
196 metagenomic comparisons are to be conducted locally, the user must also supply the metagenomic
197 dataset. The script has been designed for both ease of use as well as flexibility, such that analyses can be
198 tailored to the environmental niche and/or hypothesis under investigation. Most importantly, this script
199 is a light-weight solution which can be integrated into the standard method of metaviromic analyses.

200 The script and documentation are publicly available through [http://www.putonti-](http://www.putonti-lab.com/software.html)
201 [lab.com/software.html](http://www.putonti-lab.com/software.html).

202

203 **Proof-of-Concept**

204 Our group previously isolated and characterized phages similar to the Pseudomonas phage PB1 (Malki et
205 al., 2015), therefore we sought to examine populations of PB1 within other freshwater environments.
206 Thus, each gene annotated for the PB1 genome (Accession Number: NC_011810) (Ceyssens et al., 2009)
207 was compared first to the set of genes for the most distant relative of PB1 within its genus
208 Pbnalikeviruses, Burkholderia phage BcepF1 (Accession Number: NC_009015). For each gene the S_I

209 and Q_I values were computed. Next each gene annotated for the PB1 genome was compared via blastx
210 to all genes from viral species other than those annotated as Pbnalikevirus in GenBank (see Methods),
211 determining the values of S_2 and Q_2 .

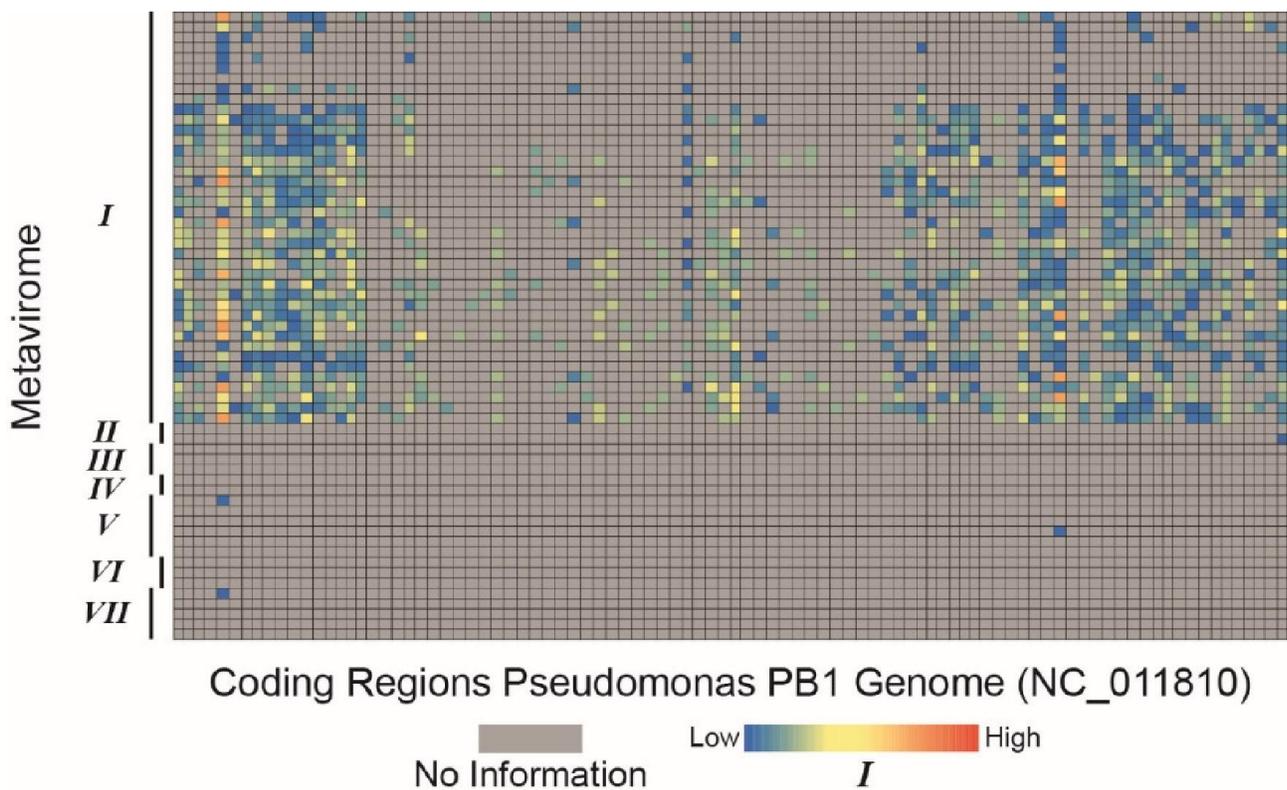
212 Surprisingly the majority of the PB1 genes exhibited greater sequence similarity to sequences within this
213 collection than they did to the Burkholderia phage BcepF1. This led us to manually inspect the genomes
214 producing these hits. In doing so, we identified a number of viral strains assigned to the taxonomic level
215 of “unclassified Myoviridae” within NCBI, rather than “Pbnalikeviruses”. These genomes were thus
216 removed from the collection of non-Pbnalikevirus viral gene sequences (as they are in fact
217 Pbnalikeviruses) and blastx was run again. (See Supplemental Table 2 for a list of the genomes
218 reclassified here as Pbnalikeviruses.) Threshold T was then calculated for all 93 annotated PB1 genes.
219 This threshold is visually represented in Figure 2 in the row marked as “GenBank*”. This variation is
220 represented as a single measure, the average of S_H and Q_H (S_2 and Q_2 in this case) less the average of S_I
221 and Q_I . Here we can see that several gene sequences (as indicated by the color scale) had better “hits”
222 to records within the GenBank collection queried than they did to the Burkholderia phage BcepF1; gray
223 blocks signify that no or weaker homology was detected ($T \leq 0$).

224 The methodology developed here was then applied to seven freshwater DNA metaviromic studies
225 (Table1); a list of the SRA datasets from each study is provided in Supplemental Table 1. Reads from all
226 seven metavirome datasets were first assembled (see Methods for details). The contigs were then
227 compared to the PB1 genome via blastx. Figure 2 graphically represents these results. Again, each gene’s
228 best hit within each metavirome sample was qualified (colored) with respect to its value relative to S_I ,
229 and Q_I . From Figure 2, one can readily identify that not all genes provide an equal signal as to the
230 presence or absence of PB1 within the sample, some serve as better markers. For instance, there are
231 several genes which have a greater sequence similarity to the PB1 genome than PB1 has to BcepF1;
232 these hits are represented within the heatmap. However non-Pbnalikevirus phage sequences may
233 exhibit equivalent or greater sequence similarity to the PB1 gene sequence (as shown in the GenBank*
234 row). The informativity metric provides a quantifiable confidence in assigning the presence/absence of a
235 taxon. Thus, the informativity I of each BLAST hit within the metaviromic samples was calculated. In
236 doing so, individual genes which provide a strong phylogenetic signal for the Pbnalikeviruses can
237 readily be identified. Figure 3 represents the results of this computation, in which each hit to a PB1 gene
238 is now assessed in light of the phylogenetic signal.

239 In an effort to assess the strength of the metric presented here, we evaluated the raw BLAST results of
240 the datasets and a BLAST score-based analysis. The BLAST results of Metaviromes II, IV, V, and VII are
241 publicly available through the web service MetaVir (Roux et al, 2014). Nine of the samples from
242 Metavirome I are also available through MetaVir. It is important to note that in contrast to the uniform
243 method in which the metavirome samples were preprocessed here (see Methods), the sequences
244 submitted to MetaVir may be assembled or raw sequences. Furthermore, MetaVir conducts BLAST
245 comparisons against the RefSeq viral database, whereas here we have included all partial and complete
246 phage sequences from GenBank which is several magnitudes of difference greater in size. Nevertheless,
247 hits to the Pbnalikeviruses (Supplemental Table 2) genomes were identified in all five MetaVir datasets;
248 the Lake Michigan and Lake Bourget samples (nine samples from Metavirome I and both samples from

249 Metavirome II) produced the most BLAST hits to the Pbnalikeviruses genomes (hundreds to
250 thousands). As MetaVir determines taxonomy based upon the best BLAST hit, these best hits were next
251 evaluated. All five datasets again included hits which were classified as Pbnalikeviruses.

252 As Figure 3 shows, Metavirome I (the Lake Michigan metaviromes generated by our group (Watkins et
253 al., 2015; Sible et al., 2015)) identifies many informative genes indicative of the presence of
254 Pseudomonas phage PB1. Metaviromes II, V, and VII contain informative hits to 1, 2, and 1 PB1 genes
255 respectively. Their informativity, however, is low, i.e. $\{S_H, Q_H\} \approx T$. This would suggest that PB1 is not
256 present within the sample: rather a homolog of the gene is present. The prevalence of informative
257 genes within several of the samples of Metavirome I and the lack thereof in the other metaviromes
258 suggests that PB1 and likewise other Pbnalikeviruses are not present (or at the least not prevalent) in
259 the other metaviromes. As viral sequence databases expand through the isolation and characterization
260 of additional viral strains, the threshold T is likely to change thus providing greater confidence in the
261 evaluation of BLAST hits for OTU calling.



263 **Figure 3.** Informativity of hits to PB1 genes within seven freshwater DNA metaviromic datasets (Table 1).

264 **Conclusions**

265 The presented method for extrapolating the presence/absence of microbial taxa is both robust and
266 versatile. Although specifically developed to tackle some of the challenges facing metavirome studies, it
267 can be applied to any WGS dataset. Specifically, the proof-of-concept investigation of seven freshwater
268 metavirome datasets can be applied in the effort to identify novel strains and species of phages with

269 confidence. Many of the prokaryote members of the human microbiome are undergoing examination,
270 but exploration of human viromes is the next frontier (Abeles & Pride, 2014; Ogilvie & Jones, 2015). As
271 such, these studies will face many of the same challenges that are detailed as part of the presented
272 study. Nevertheless, improved bioinformatic tools for mining metaviromic analyses, coupled with
273 further physical isolation and characterization of viral species have the potential to greatly expand our
274 knowledge of the viral diversity on Earth.

275

276 **Acknowledgements**

277 The authors would like to thank Ms. Katherine Bruder, Alexandria Cooper, Kema Malki, and Emily Sible
278 for their contributions to general research investigating Pbnalikeviruses.

279

280 **References**

281 Abeles SR, Pride DT. 2014. Molecular bases and role of viruses in the human microbiome. *J Mol Biol.*
282 426:3892-3906. doi: 10.1016/j.jmb.2014.07.002

283 Berdjeb L, Pollet T, Domaizon I, Jacquet S. 2011. Effect of grazers and viruses on bacterial community
284 structure and production in two contrasting trophic lakes. *BMC Microbiol.* 11: 88. doi:10.1186/1471-
285 2180-11-88

286 Breitbart M, Miyake JH, Rohwer F. 2004. Global distribution of nearly identical phage-encoded DNA
287 sequences. *FEMS Microbiol Lett.* 236: 249–256. doi:10.1016/j.femsle.2004.05.042

288 Brum JR, Hurwitz BL, Schofield O, Ducklow HW, Sullivan MB. 2016. Seasonal time bombs: dominant
289 temperate viruses affect Southern Ocean microbial dynamics. *ISME J.* 10: 437–449.
290 doi:10.1038/ismej.2015.125

291 Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüßow H. 2003. Phage as agents of lateral
292 gene transfer. *Curr Opin Microbiol.* 6: 417–424.

293 Ceysens P-J, Miroshnikov K, Mattheus W, Krylov V, Robben J, Noben J-P, Vanderschraeghe S, Sykilinda
294 N, Kropinski AM, Volckaert G, Mesyanzhinov V, Lavigne R. 2009. Comparative analysis of the widespread
295 and conserved PB1-like viruses infecting *Pseudomonas aeruginosa*. *Environ Microbiol.* 11: 2874–2883.
296 doi:10.1111/j.1462-2920.2009.02030.x

297 Clokie MR, Millard AD, Letarov AV, Heaphy S. 2011. Phages in nature. *Bacteriophage.* 1: 31–45.
298 doi:10.4161/bact.1.1.14942

299 Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L,
300 McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R,
301 Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. 2008. Functional metagenomic profiling of
302 nine biomes. *Nature.* 452: 629–632. doi:10.1038/nature06810

- 303 Dorigo U, Jacquet S, Humbert J-F. 2004. Cyanophage diversity, inferred from g20 gene analyses, in the
304 largest natural lake in France, Lake Bourget. *Appl Environ Microbiol.* 70: 1017–1022.
- 305 Edwards RA, Rohwer F. 2005. Opinion: Viral metagenomics. *Nat Rev Microbiol.* 3: 504–510.
306 doi:10.1038/nrmicro1163
- 307 Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, Raoult D, Desnues C. 2013. Viruses in the desert:
308 a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *ISME J.*
309 7: 359–369. doi:10.1038/ismej.2012.101
- 310 Gao E-B, Gui J-F, Zhang Q-Y. 2012. A novel cyanophages with a cyanobacterial nonbleaching protein A
311 gene in the genome. *J Virol.* 86: 236-245. doi: 10.1128/JVI.06282-11
- 312 Gudbergsdóttir SR, Menzel P, Krogh A, Young M, Peng X. 2015. Novel viral genomes identified from six
313 metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot
314 springs. *Environ Microbiol.* 18: 863–874. doi:10.1111/1462-2920.13079
- 315 Hatfull GF. 2008. Bacteriophage genomics. *Curr Opin Microbiol.* 11: 447–453.
316 doi:10.1016/j.mib.2008.09.004
- 317 Holmes RK. 2000. Biology and molecular epidemiology of diphtheria toxin and the *tox* gene. *J Infect Dis.*
318 181: S156–S167. doi:10.1086/315554
- 319 Hurwitz BL, Sullivan MB. 2013. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and
320 associated protein clusters for quantitative viral ecology. *PLoS ONE.* 8: e57355.
321 doi:10.1371/journal.pone.0057355
- 322 Huson DH, Weber N. 2013. Microbial community analysis using MEGAN. *Meth Enzymol.* 531: 465–485.
323 doi:10.1016/B978-0-12-407863-5.00021-6
- 324 Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles:
325 implications for marine biogeochemical cycles. *Nat Rev Microbiol.* 12: 519–528.
326 doi:10.1038/nrmicro3289
- 327 Keegan KP, Glass EM, Meyer F. 2016. MG-RAST, a Metagenomics Service for Analysis of Microbial
328 Community Structure and Function. *Methods Mol Biol.* 1399: 207–233. doi:10.1007/978-1-4939-3369-
329 3_13
- 330 Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses
331 yield proteins during host infection. *Nature.* 438: 86-89. doi: 10.1038/nature04111
- 332 López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. 2009. High diversity of the viral
333 community from an Antarctic lake. *Science.* 326: 858–861. doi:10.1126/science.1179287
- 334 Malki K, Kula A, Bruder K, Sible E, Hatzopoulos T, Steidel S, Watkins SC, Putonti C. 2015. Bacteriophages
335 isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virology.* 12:
336 164. doi:10.1186/s12985-015-0395-0
- 337 Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Marine ecosystems: bacterial photosynthesis

- 338 genes in a virus. *Nature*. 424: 741. doi: 10.1038/424741a
- 339 Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. 2013. Rapid evolution of the human gut
340 virome. *Proc Natl Acad Sci USA*. 110: 12450–12455. doi:10.1073/pnas.1300833110
- 341 Ogilvie LA, Jones BV. 2015. The human gut virome: a multifaceted majority. *Front Microbiol*. 6:918. doi:
342 10.3389/fmicb.2015.00918
- 343 Rohwer F, Thurber RV. 2009. Viruses manipulate the marine environment. *Nature*. 459: 207–212.
344 doi:10.1038/nature08060
- 345 Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. 2009. Metagenomic analysis of viruses in reclaimed
346 water. *Environ Microbiol*. 11: 2806–2820. doi:10.1111/j.1462-2920.2009.01964.x
- 347 Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D. 2012.
348 Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS*
349 *ONE*. 7: e33641. doi:10.1371/journal.pone.0033641
- 350 Roux S, Tournayre J, Mahul A, Debroas D, Enault F. 2014. Metavir 2: new tools for viral metagenome
351 comparison and assembled virome analysis. *BMC Bioinformatics*. 15: 76. doi:10.1186/1471-2105-15-76
- 352 Sible E, Cooper A, Malki K, Bruder K, Watkins SC, Fofanov Y, Putonti C. 2015. Survey of viral populations
353 within Lake Michigan nearshore waters at four Chicago area beaches. *Data Brief*. 5: 9–12.
354 doi:10.1016/j.dib.2015.08.001
- 355 Suttle CA. 2007. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol*. 5: 801–812.
356 doi:10.1038/nrmicro1750
- 357 Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. 2011. Phage auxiliary
358 metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl Acad Sci USA*.
359 108: E757-E764. doi: 10.1073/pnas.1102164108
- 360 Tseng C-H, Chiang P-W, Shiah F-K, Chen Y-L, Liou J-R, Hsu T-C, Maheswararajah S, Saeed I, Halgamuge S,
361 Tang SL. 2013. Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic
362 disturbances. *ISME J*. 7: 2374–2386. doi:10.1038/ismej.2013.118
- 363 Watkins SC, Kuehnle N, Ruggeri CA, Malki K, Bruder K, Elayyan J, Damisch K, Vahora N, O'Malley P,
364 Ruggles-Sage B, Romer Z, Putonti C. 2015. Assessment of a metaviromic dataset generated from
365 nearshore Lake Michigan. *Marine and Freshwater Research*. 2015; doi:10.1071/MF15172
- 366 Wilhelm SW, Carberry MJ, Eldridge ML, Poorvin L, Saxton MA, Doblin MA. 2006. Marine and freshwater
367 cyanophages in a Laurentian Great Lake: evidence from infectivity assays and molecular analyses of g20
368 genes. *Appl Environ Microbiol*. 72: 4957–4963. doi:10.1128/AEM.00349-06
- 369 Wilhelm SW, Suttle CA. 1999. Viruses and Nutrient Cycles in the Sea. *BioScience*. 49: 781.
370 doi:10.2307/1313569
- 371 Willner D, Haynes MR, Furlan M, Hanson N, Kirby B, Lim YW, Rainey PB, Schmieder R, Youle M, Conrad
372 D, Rohwer F. 2012. Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung.

- 373 *Am J Respir Cell Mol Biol.* 46: 127–131. doi:10.1165/rcmb.2011-0253OC
- 374 Winget DM, Helton RR, Williamson KE, Bench SR, Williamson SJ, Wommack KE. 2011. Repeating patterns
375 of virioplankton production within an estuarine ecosystem. *Proc Natl Acad Sci USA.* 108: 11506–11511.
376 doi:10.1073/pnas.1101907108
- 377 Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ.
378 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand*
379 *Genomic Sci.* 6: 427–439. doi:10.4056/sigs.2945050
- 380 Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
381 *Genome Res.* 18: 821–829. doi:10.1101/gr.074492.107
- 382

383 **Supplemental Data**

384 **Supplemental Table 1.** SRA datasets from each study.

Metavirome ID	SRA run datasets
I	SRR1301999
	SRR1302020
	SRR1302010
	SRR1296481
	Private data set
	SRR1974493
	SRR1974494
	SRR1974490
	SRR1974491
	SRR1974495
	SRR1974496
	SRR1974497
	SRR1974498
	SRR1915829
	SRR1915851
	SRR1974488
	SRR1974489
	SRR1974499
	SRR1974500
	SRR1974501
	SRR1974502
	SRR1974503
	SRR1974504
	SRR1974505
	SRR1974506
	SRR1974507
	SRR1974508
	SRR1974509
	SRR1974510
SRR1974511	
SRR1974512	
SRR1974513	
SRR1974514	
SRR1974515	
SRR1974516	

	SRR1974517
II	ERR019477
	ERR019478
III	SRR001047
	SRR001075
	SRR001076
IV	SRR013515, SRR013516, SRR013517
	SRR013520, SRR013521
V	SRR014584
	SRR014585
	SRR014586
	SRR014587
	SRR014588
	SRR014589
VI	SRR138365
	SRR155589
	SRR171296
VII	SRR371574
	SRR648311
	SRR648312
	SRR648313
	SRR648314

385

386 **Supplemental Table 2: Pbunalikevirus genomes**

Phage	Genome size	NCBI Assigned Taxonomy	GenBank Accession No.
PB1	65764	Pbunalikevirus	NC_011810
SN	66390	Pbunalikevirus	NC_011756
14-1	66238	Pbunalikevirus	NC_011703
LMA2	66530	Pbunalikevirus	NC_011166
LBL3	64427	Pbunalikevirus	NC_011165
F8	66015	Pbunalikevirus	NC_007810
BcepF1	72415	Pbunalikevirus	NC_009015
PaMx13	66450	Pbunalikevirus	JQ067083
pp DL52	65867	Pbunalikevirus	KR054028
pp SPM-1	65729	Pbunalikevirus	NC_023596
pp vB_PaeM_C1-14_Ab28	66181	unclassified Myoviridae	NC_026600
pp DL60	66103	unclassified Myoviridae	KR054030
pp KPP12	64144	unclassified Myoviridae	NC_019935
pp NH-4	66116	unclassified Myoviridae	NC_019451
pp vB_PaeM_PAO1_Ab27	66299	unclassified Myoviridae	NC_026586
pp vB_PaeM_PAO1_Ab29	66326	unclassified Myoviridae	LN610588
pp JG024	66275	unclassified Myoviridae	NC_017674
pp DL68	66111	unclassified Myoviridae	KR054033
S12-1	66257	unclassified Myoviridae	LC102730
R18	63560	unclassified Myoviridae	LC102729

387