



HAL
open science

Increase in taxonomic assignment efficiency of viral reads in metagenomic studies

Sarah François, Denis Filloux, Marie Frayssinet, Philippe Roumagnac, D. P. Martin, Marie Helene Ogliastro, Remy Froissart

► To cite this version:

Sarah François, Denis Filloux, Marie Frayssinet, Philippe Roumagnac, D. P. Martin, et al.. Increase in taxonomic assignment efficiency of viral reads in metagenomic studies. *Virus Research*, 2018, 244, pp.230-234. 10.1016/j.virusres.2017.11.011 . hal-01837312

HAL Id: hal-01837312

<https://hal.science/hal-01837312>

Submitted on 13 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Increase in taxonomic assignment efficiency of viral reads**
2 **in metagenomic studies**

3
4 S. François ^a, D. Filloux ^b, M. Frayssinet ^a, P. Roumagnac ^b, D. P. Martin ^c, M. Ogliastro ^{a #}
5 and R. Froissart ^{d #}

6
7 ^a Laboratory « Diversité, Génomes et Interactions Microorganismes Insectes » (DGIMI)
8 UMR 1333, INRA, Université Montpellier, 34095 Montpellier, France.

9 ^b Laboratory « Biologie et Génétique des Interactions Plante-Parasite » (BGPI)
10 UMR 54/385, CIRAD, INRA, SUPAGRO, Campus International de Baillarguet, 34398 Montpellier,
11 France

12 ^c Computational Biology Group, Institute of Infectious Disease and Molecular Medicine,
13 Faculty of Health Sciences, University of Cape Town, Observatory, South Africa.

14 ^d Laboratory « Maladies infectieuses et vecteurs: écologie, génétique, évolution et contrôle »
15 (MIVEGEC) UMR 5290, CNRS, IRD, Université Montpellier, 911 avenue Agropolis, 34394
16 Montpellier, France.

17 # **corresponding authors:** remy.froissart@cnsr.fr, marie-helene.ogliastro@inra.fr

18
19 **Running Title :** Illuminating viral dark matter

20 **ABSTRACT**

21 Metagenomics studies have revolutionized the field of biology by revealing the presence of
22 many previously unisolated and uncultured micro-organisms. However, one of the main problems
23 encountered in metagenomic studies is the high percentage of sequences that cannot be assigned
24 taxonomically using commonly used similarity-based approaches (e.g. BLAST or HMM). These
25 unassigned sequences are allegorically called « dark matter » in the metagenomic literature and are
26 often referred to as being derived from new or unknown organisms. Here, based on published and
27 original metagenomic datasets coming from virus-like particle enriched samples, we present and
28 quantify the improvement of viral taxonomic assignment that is achievable with a new similarity-
29 based approach. Indeed, prior to any use of similarity based taxonomic assignment methods, we
30 propose assembling contigs from short reads as is currently routinely done in metagenomic studies,
31 but then to further map unassembled reads to the assembled contigs. This additional mapping step
32 increases significantly the proportions of taxonomically assignable sequence reads from a variety -
33 plant, insect and environmental (estuary, lakes, soil, feces) - of virome studies.

34

35 **Keywords:** Dark Matter, Viral metagenomics, BLAST, Mapping

36 **1. Introduction**

37 The advent of high throughput sequencing has enabled the cataloguing and enumeration of
38 microbial species without a priori information on their life cycles. When specifically focusing on
39 viruses, this so-called viral metagenomics approach, has so-far revealed the extraordinary diversity
40 and prevalence of viruses in aquatic and terrestrial ecosystems, highlighting the key contributions of
41 these microbes to all ecosystems on Earth (Brum and Sullivan, 2015; Mokili et al., 2012; Suttle,
42 2007).

43 One simple but important insight yielded by these astonishing discoveries is that we
44 probably currently know far less than 1% of all viral species that are circulating on Earth (Anthony
45 et al., 2013; Mokili et al., 2012). It is sobering to consider that despite the large numbers of viromes
46 that have been examined over the past 20 years, almost every new viromics project yields large
47 numbers of sequences that have no significant degree of similarity with those referenced in
48 databases. These sequences are often referred to as "dark matter". Our inability to properly
49 categorize the latter sequences has the potential to strongly bias our view of both the actual diversity
50 of viruses in a given environment and their ecological roles (Krishnamurthy and Wang, 2017;
51 Roossinck et al., 2015; Rosario and Breitbart, 2011).

52 When attempting to characterize any virome from metagenomic datasets, researchers face
53 two main challenges: i) purifying viral genomes present in heterogeneous materials or biological
54 tissues without introducing biases due to technical processes and ii) accurately assign sequence
55 reads. Whereas solutions to the first of these challenges will vary from environment to environment,
56 the second challenge could be met both with improved computational methods that are capable of
57 accounting for compositionally biased databases, and by vastly increasing the diversity of viral

58 genome sequences within public databases. For instance in most viral metagenomics projects, only
59 approximately 10 to 20% of sequence reads can be confidently attributed to viruses and, in most
60 cases, the remaining sequence reads are treated as unanalyzable dark matter (Krishnamurthy and
61 Wang, 2017; Rosario and Breitbart, 2011).

62 In viral metagenomic studies, the classical bioinformatical workflow consists of de novo
63 assembling contigs from short reads generated by high throughput sequencing and then performing
64 homology inferences via alignments of sequences (both reads and contigs) to reference databases
65 using a tool such as BLAST (Allander et al., 2001; Angly et al., 2006; Breitbart et al., 2002).
66 However, this method usually yields low quality taxonomic assignments due, at least in part, to both
67 the length of sequence reads generally being <500nts, and the low degrees of sequence identity that
68 are commonly shared between query sequences and the virus genomic sequences present in public
69 databases (Tangherlini et al., 2016). Moreover, the classical BLAST workflow most often leads to a
70 high number of reads that cannot be attributed with high confidence to related sequences and are
71 thus considered as unknown sequences.

72 To decrease the amount of this dark matter, it has been recently proposed to integrate a new
73 step in the computational workflow: a recruitment process consisting of the mapping of
74 unassembled short sequence reads onto assembled contigs prior BLASTx requests (Krishnamurthy
75 and Wang, 2017), a workflow that we will referred to as assembly-mapping-BLAST (AM-BLAST
76 for short) as opposed to the classical BLAST workflow. Although this methodology is used in viral
77 metagenomics (Cotten et al., 2014), no comparative study has ever been made to evaluate how
78 efficiently the use of AM-BLAST reduces the amount of dark matter relative to the classical
79 BLAST workflow.

80 Alternatives to BLAST have been developed to improve taxonomic assignments of query
81 sequences being compared to a database of reference sequences. One of the most used alternative
82 approaches involves a hidden Markov model (HMM) based classifier where position-specific
83 information on nucleotide variation across a set of related sequences is taken into account when
84 determining whether there are statistically significant matches within a database to query sequences.
85 This approach outperformed BLAST when attempting to find database matches to divergent viral
86 sequences, although it remained less accurate than BLAST with respect to taxonomic assignment
87 (Fancello et al., 2012; Remmert et al., 2012; Skewes-Cox et al., 2014).

88 The aim of the present study was to quantify improvement in the taxonomic assignment of
89 viral sequences after the use of AM-BLAST relative to classical-BLAST workflows. We thus
90 compared the number of unassigned reads after running these two workflows on fifteen datasets
91 consisting of samples enriched for virus-like particles (VLP). Our results indicate that the AM-
92 BLAST workflow reduced significantly the number of unassigned viral reads compared to the
93 classical-BLAST workflow.

94

95 **2. Materials and methods**

96 **2.1. Sampling, virome preparation and sequencing**

97 Three insect species (*Hypera postica*, *Acyrtosiphon pisum* and *Coccinella septempunctata*)
98 and one plant species (*Medicago sativa*) were collected in the Montpellier area of Southern France
99 (domaine de Restinclières, Prades le Lez, France, N 43°42'54.362" EO 3°51'31.749"); for each
100 species, several individuals were pooled and constituted one sample. Samples were stored at -80°C
101 without addition of any preservative solutions. One gram of insect or plant material was processed

102 using a virion-associated nucleic acids (VANA) based metagenomics approach to screen for the
103 presence of viruses (Palanga et al., 2016). Amplified and tagged DNA products of the VANA
104 approach were sequenced using an Illumina platform (MiSeq sequencing: 2 x 300 nt paired-end
105 sequencing with V3 chemistry, Beckman Coulter Genomics, USA).

106 **2.2. Bioinformatics analysis: virome cleaning, read assembly, taxonomic assignment** 107 **and clustering**

108 Raw reads were first demultiplexed using `agrep` (Wu and Manber, 1992). Illumina adaptors
109 were removed and we selected reads based on their quality ($\geq q30$ and length elimination of reads $<$
110 45 nt) using `Cutadapt 1.9` (Martin, 2011). The remaining reads will be hereafter referred to as
111 “cleaned reads”. Paired cleaned reads were merged using `FLASH 1.2.11` (Magoc and Salzberg,
112 2011). Then, random subsets of two hundred thousand cleaned reads per virome were used for all
113 the following steps.

114 First, we performed the classical-BLASTx workflow which involved taxonomically
115 assigning reads using BLASTx searches against the non-redundant GenBank viral protein
116 sequences database for taxonomic attribution (e-value cutoff of $< 10^{-3}$) (Altschul et al., 1990) on
117 200,000 randomly chosen “cleaned reads” (Fig. 1A).

118 Second, we performed the AM-BLAST workflow which involved subjecting the cleaned
119 reads to assembly using SPAdes (different kmer sizes: 21, 33, 55, 77, 125) (Bankevich et al., 2012).
120 Contigs and unassembled reads were then assembled using CAP3 with default parameters (Huang,
121 1999). It is to notice that CAP3 was only used to recruit reads; it should not be used for
122 identification of genomes because this software can result in creation of chimaeras. Mapping of the
123 remaining reads both (i) onto the new contigs obtained after *de novo* assembly and (ii) to the

124 remaining unassembled reads was performed using Bowtie 2.1.0 (using the local and very sensitive
125 option) (Langmead, 2010; Toland et al., 2013). All contigs and unassembled reads were then
126 subjected to BLASTx searches against the non-redundant GenBank viral protein sequences
127 database for taxonomic attribution (e-value cutoff of $< 10^{-3}$) (Altschul et al., 1990) (the whole
128 procedure is summarized in Fig. 1 B).

129 To obtain an overview of genetic diversity across all the metagenomic datasets, 10,000 reads
130 were randomly chosen (3 replicates) and subjected to BLASTx searches against the NCBI non-
131 redundant protein sequences database.

132 Seven publicly available metagenomic datasets were also analyzed in this study, originating
133 from five independent datasets after enrichment for virus-like particles from mosquitoes (Ng et al.,
134 2011), a human fecal sample (Kim et al., 2011), an estuary sample (McDaniel et al., 2008), two lake
135 samples (Roux et al., 2012), and two Antarctic ecosystem samples (Zablocki et al., 2014) (Table 1).
136 These seven datasets were *de novo* assembled and analyzed as described above.

137

138 **3. Results**

139 The aim of our study was to compare the efficiency with which classical-BLASTx (Fig. 1 A)
140 and AM-BLAST (Fig. 1 B) workflows taxonomically assign reads from metagenomic sequencing
141 datasets. These datasets were obtained from samples of various origins and enriched for virus-like
142 particles using different procedures (Table 1): (i) eight insects and plants processed for the purpose
143 of the present study (hereafter referred to as viromes 1 to 8) and (ii) seven datasets from published
144 studies originating from environmental and insect samples (hereafter referred to as viromes 9 to 15)

145 (Table 1). The viromes represented by these two set will be hereafter referred to as original and
146 published viromes, respectively.

147 The classical BLASTx workflow was able to assign, with high level of confidence
148 (according to E-value of BLASTx, see M&M section), between 59% and only 1.3% of reads from
149 the original viromes (Insects 6 and Insects 4, respectively) and between 7.5% and 0.15% of reads
150 for the published viromes (Lake Bourget and Hypolith, respectively), in agreement with published
151 results (Fig. 2). The AM-BLASTx workflow on the other hand, allowed the assignment of 89.4%
152 and 42.5% of reads for the original viromes (Insects 4 and Insects 5, respectively) and between
153 18.6% and 1.4% of reads for the published viromes (Lake Pavin and Estuary, respectively) (Fig. 2).
154 AM-BLASTx workflow yielded a significant improvement in overall taxonomic assignment
155 efficiency ($P = 6.1 \times 10^{-5}$, Wilcoxon comparison test) (Fig. 2, Supplemental Table 1). This
156 improvement was particularly notable for insect virome 4 where the AM-BLASTx workflow
157 yielded a 70-fold improvement in the proportion of taxonomically assignable reads.

158 Proportions of assignable reads varied markedly between the analyzed viromes. For the
159 original datasets an average of 21% of reads were assignable by classical-BLASTx and 62% by
160 AM-BLASTx. For the published datasets an average of only 3% of reads were assignable by
161 classical-BLASTx and 13% by AM-BLASTx (Fig. 2 and Supplemental Table 1). On one hand, the
162 aquatic, marine and fecal environmental viromes were dominated by large dsDNA bacteriophages
163 (> 250 kb) belonging to the *Myoviridae* and *Siphoviridae* families. On the other hand, insect and
164 plant viromes were dominated by small RNA and DNA viruses (< 10 kb) belonging to the
165 *Iflaviridae*, *Dicistroviridae*, *Parvoviridae*, *Amalgaviridae* and *Partitiviridae* families (Supplemental
166 Table 2).

167 In order to assign the remaining unclassified reads to cellular origin or to dark matter, we

168 taxonomically assigned a subset of ten thousand reads that were randomly sampled from each
169 dataset (i.e. 5% of the total number of reads per dataset) using BLASTx. Despite the datasets all
170 being derived from samples that were processed to enrich for viral-like particles, from 1% to 55%
171 of the reads in both the original and published datasets were most likely of cellular origin (Fig. 3).
172 On the one hand, for the original viromes, up to 22% and 53% of the reads were respectively
173 assigned to bacteria and eukaryotes. From 47% and 53% of reads from the two plant viromes were
174 assignable to plant genomic sequences, while 18% and 22% of the reads from two of the insect
175 viromes (viromes 3 and 4 from the aphid *A. pisum*) were likely derived from *Candidatus*
176 *Hamiltonella defensa*, the aphid's endosymbiotic bacteria (Fig. 3). On the other hand, for the
177 published viromes, 3% to 39% and 0 to 1% of reads were assigned to bacterial and eukaryotic
178 organisms respectively, in agreement with published results (Fig. 3, Supplemental Table 3).
179 Specifically, lake datasets (viromes 11 to 14), contained similar proportions (about 25%) of
180 bacterial and bacteriophage sequences, indicating the presence of bacteria, bacteriophage particles
181 and prophage nucleic acids as already reported (Enault et al., 2016; Roux et al., 2013). Moreover,
182 the human feces and soil viromes contained a higher proportion of reads assigned to bacteria (from
183 15% to 39%) than those from other sources (Fig. 3).

184 **4. Discussion**

185 In this study, we propose a modification of the classical BLASTx-based workflow that
186 improves the taxonomic assignment of sequences from metagenomic virome studies. Based on the
187 statement that increasing the length of query sequences could improve the accuracy with which they
188 could be taxonomic assigned using BLASTx, we introduced a recruitment step of remapping
189 unassembled reads onto assembled contigs prior to BLASTx searches (a workflow that we called
190 “assembly-mapping BLASTx” or AM-BLASTx for short) and tested this on viral metagenomic

191 datasets. These datasets were obtained after different technical procedures, both prior to sequencing
192 (i.e. use of rolling-circle amplification or random PCR amplification) and during the sequencing
193 process (i.e. MiSeq Illumina or 454 Pyrosequencing ; Table 1). We found that, when applied to each
194 datasets, the AM-BLASTx workflow systematically and substantially increased the numbers of
195 virus-derived sequences that could be taxonomically assigned relative to the numbers that were
196 assignable using the classical BLASTx workflow. Analyses made on fifteen metagenomic datasets
197 lead to an average five-fold increase in the number of assignable reads.

198 Our analyses thus revealed that one major parameter to improve the performance of
199 BLASTx-based approaches for taxonomically assigning viral reads is likely the lengths of the
200 sequences that will be analyzed by BLASTx. Indeed, viral genomes are more variable than those
201 cellular organisms because of high mutation rates, so longer reads and contigs decrease the impact
202 of point mutations that decrease the degrees of similarity between query and reference sequences
203 within the database that is being searched by BLAST or HMM-based approaches. The lengths of
204 query sequences can be increased both by computational processing of the sequence data prior to
205 performing blast searches (as is done in the AM-BLASTx workflow), and by technical procedures
206 during virus-like particle enrichment. Specifically, it is desirable to lengthen the query sequences,
207 either by using sequencing technologies that enable long reads (such as 454 that is no longer used or
208 Pacific BioScience) or by increasing sequencing depth (such as with Illumina) so as to enable the
209 assembly of longer contigs. In fact, with simulated metagenomic datasets, a positive correlation has
210 been found between sequencing depth and the proportions of reads that could be taxonomically
211 assigned (García-López et al., 2015). Interestingly, our analyses did not reveal differences in the
212 degrees of taxonomic improvement between studies using 454 and Illumina sequencing
213 technologies, suggesting that large sequencing depth can compensate shorter read lengths.

214 Our analyses also revealed that viral metagenomic dataset obtained for the purpose of this
215 study from arthropods and plants (our so called “original viromes”) seemed to be dominated by
216 small viruses (<10 kb), while published environmental viromes contained a high number of reads
217 assigned to prophages and genomic bacterial DNA. The generality of such differential viral
218 communities according to different environments is, however, questionable because only very few
219 studies have reported insect and plant viromes (Junglen and Drosten, 2013) and we can thus not
220 compare our results with those of others. Moreover, technical procedures during the preparation of
221 the original and published viromes differed in that the latter were obtained by rolling-circle
222 amplification, a technique known to induce amplification biases toward circular genomes, while the
223 former viromes were obtained after random PCR, a technique that is not known to have this bias.

224 Altogether, the AM-BLASTx workflow represents a simple and rapid way to improve the
225 taxonomic assignment of viral sequences from metagenomic datasets independently of the origin of
226 the samples. Our results indicate that the proportion of unassigned reads (i.e. the “dark matter”) in
227 virome datasets can be significantly reduced by combining the following approaches: (i) the use of
228 purification techniques that rigorously enrich samples for virus-like particles in order to minimize
229 amounts of cellular genomic DNA, (ii) use of sequencing technologies that maximize the number of
230 reads, and (iii) use of computational workflows that include steps of mapping of reads to *de novo*
231 assembled contigs prior to BLAST searches.

232 **Acknowledgments**

233 We are particularly grateful to the Conseil General de l'Hérault for providing us the
234 opportunity to collect insects and plants in the Domaine de Restinclières. We warmly thank
235 Francois Enault and the reviewers for their insightful comments on the manuscript. S. F. is a

236 doctoral fellow from the University of Montpellier and was supported by a scholarship from Institut
 237 National de la Recherche Agronomique (INRA). The authors declare no competing financial
 238 interests.

239

240 **Author Contributions:** Data acquisition (S.F., D.F., M.F.); Analysis and interpretation of data (S.F.,
 241 M.O. and R.F.); Manuscript preparation (S.F., M.O, D.M., D.F. and R.F.); Study supervision (S.F.,
 242 M.O. and R.F.).

243 **References**

- 244 Allander, T., Emerson, S.U., Engle, R.E., Purcell, R.H., Bukh, J., 2001. A virus discovery method
 245 incorporating DNase treatment and its application to the identification of two bovine
 246 parvovirus species. *Proc. Natl. Acad. Sci.* 98, 11609–11614. doi:10.1073/pnas.211424698.
- 247 Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., 1990. Basic local alignment search tool. *J.*
 248 *Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2.
- 249 Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. a, Carlson, C., Chan, A.M., Haynes,
 250 M., Kelley, S., Liu, H., Mahaffy, J.M., Mueller, J.E., Nulton, J., Olson, R., Parsons, R.,
 251 Rayhawk, S., Suttle, C. a, Rohwer, F., 2006. The marine viromes of four oceanic regions. *PLoS*
 252 *Biol.* 4, e368. doi:10.1371/journal.pbio.0040368.
- 253 Anthony, S.J., Epstein, J.H., Murray, K.A., Navarrete-macias, I., Zambrana-torrel, C.M.,
 254 Solovyov, A., Ojeda-flores, R., Arrigo, N.C., Islam, A., Khan, A., Hosseini, P., Bogich, T.L.,
 255 Mazet, J.A.K., Daszak, P., Lipkin, W., 2013. A strategy to estimate unknown viral diversity in
 256 mammals. *MBio* 4, 1–15. doi:10.1128/mBio.00598-13.
- 257 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M.,
 258 Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler,
 259 G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its
 260 applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–77.
 261 doi:10.1089/cmb.2012.0021.
- 262 Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., Rohwer,
 263 F., 2002. Genomic analysis of uncultured marine viral communities 99, 14250–134255.
 264 doi:10.1073/pnas.202488399.

- 265 Brum, J.R., Sullivan, M.B., 2015. Rising to the challenge: accelerated pace of discovery transforms
266 marine virology. *Nat. Rev. Microbiol.* 13, 147–159. doi:10.1038/nrmicro3404.
- 267 Cotten, M., Oude Munnink, B., Canuti, M., Deijs, M., Watson, S.J., Kellam, P., Van Der Hoek, L.,
268 2014. Full genome virus detection in fecal samples using sensitive nucleic acid preparation,
269 deep sequencing, and a novel iterative sequence classification algorithm. *PLoS One* 9, e93269.
270 doi:10.1371/journal.pone.0093269.
- 271 Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M.B., Petit, M.-A., 2016. Phages rarely
272 encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* 1–11.
273 doi:10.1038/ismej.2016.90.
- 274 Fancello, L., Raoult, D., Desnues, C., 2012. Computational tools for viral metagenomics and their
275 application in clinical research. *Virology* 434, 162–74. doi:10.1016/j.virol.2012.09.025.
- 276 García-López, R., Vázquez-Castellanos, J.F., Moya, A., 2015. Fragmentation and Coverage
277 Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations. *Front.*
278 *Bioeng. Biotechnol.* 3, 141. doi:10.3389/fbioe.2015.00141.
- 279 Huang, X., 1999. CAP3: A DNA Sequence Assembly Program. *Genome Res.* 9, 868–877.
280 doi:10.1101/gr.9.9.868.
- 281 Junglen, S., Drosten, C., 2013. Virus discovery and recent insights into virus diversity in arthropods.
282 *Curr. Opin. Microbiol.* 16, 507–513. doi:10.1016/j.mib.2013.06.005.
- 283 Kim, M.-S., Park, E.-J., Roh, S.W., Bae, J.-W., 2011. Diversity and abundance of single-stranded
284 DNA viruses in human feces. *Appl. Environ. Microbiol.* 77, 8062–70.
285 doi:10.1128/AEM.06331-11.
- 286 Krishnamurthy, S.R., Wang, D., 2017. Origins and challenges of viral dark matter. *Virus Res.* 239,
287 136–142. doi:10.1016/j.virusres.2017.02.002.
- 288 Langmead, B., 2010. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinforma.* 11.
- 289 Magoc, T., Salzberg, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome
290 assemblies. *Bioinformatics* 27, 2957–2963. doi:10.1093/bioinformatics/btr507.
- 291 Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
292 *EMB* 17, 10–12. doi:10.14806/ej.17.1.200.
- 293 McDaniel, L., Breitbart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F., Paul, J.H., 2008.
294 Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression.
295 *PLoS One* 3, e3263. doi:10.1371/journal.pone.0003263.
- 296 Mokili, J.L., Rohwer, F., Dutilh, B.E., 2012. Metagenomics and future perspectives in virus
297 discovery. *Curr. Opin. Virol.* 2, 63–77. doi:10.1016/j.coviro.2011.12.004.
- 298 Ng, T.F.F., Willner, D.L., Lim, Y.W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y.,
299 Rohwer, F., Breitbart, M., 2011. Broad surveys of DNA viral diversity obtained through viral
300 metagenomics of mosquitoes. *PLoS One* 6, e20579. doi:10.1371/journal.pone.0020579.

- 301 Palanga, E., Filloux, D., Martin, D.P., Fernandez, E., Bouda, Z., Gargani, D., Ferdinand, R., Zabre,
 302 J., Neya, B., Sawadogo, M., Traore, O., Peterschmitt, M., Roumagnac, P., 2016. Metagenomic-
 303 Based Screening and Molecular Characterization of Cowpea- Infecting Viruses in Burkina
 304 Faso. *PLoS One* 11, e0165188. doi:10.1371/journal.pone.0165188.
- 305 Remmert, M., Biegert, A., Hauser, A., Soding, J., 2012. HHblits: lightning-fast iterative protein
 306 sequence searching by HMM-HMM alignment. *Nat Meth* 9, 173–175.
 307 doi:10.1038/nmeth.1818.
- 308 Roossinck, M.J., Martin, D.P., Roumagnac, P., 2015. Plant Virus Metagenomics : Advances in Virus
 309 Discovery. *Phytopathology* 105, 716–727. doi:10.1094/PHYTO-12-14-0356-RVW.
- 310 Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. *Curr. Opin.*
 311 *Virol.* 1, 289–297. doi:10.1016/j.coviro.2011.06.004.
- 312 Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T.,
 313 Debros, D., 2012. Assessing the diversity and specificity of two freshwater viral communities
 314 through metagenomics. *PLoS One* 7, e33641. doi:10.1371/journal.pone.0033641.
- 315 Roux, S., Krupovic, M., Debros, D., Forterre, P., 2013. Assessment of viral community functional
 316 potential from viral metagenomes may be hampered by contamination with cellular sequences.
 317 *Open Biol* 3, 130160. doi:10.1038/ismej.2016.90.
- 318 Skewes-Cox, P., Sharpton, T.J., Pollard, K.S., DeRisi, J.L., 2014. Profile hidden Markov models for
 319 the detection of viruses within metagenomic sequence data. *PLoS One* 9, e105067.
 320 doi:10.1371/journal.pone.0105067.
- 321 Suttle, C.A., 2007. Marine viruses (mdash) major players in the global ecosystem. *Nat. Rev.*
 322 *Microbiol.* 5, 801–812. doi:10.1038/nrmicro1750.
- 323 Tangherlini, M., Dell’Anno, A., Zeigler Allen, L., Riccioni, G., Corinaldesi, C., 2016. Assessing
 324 viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of
 325 different bioinformatic tools for viral metagenomic analyses. *Sci. Rep.* 22, 28428.
 326 doi:10.1038/srep28428.
- 327 Toland, A.E., Çatalyürek, Ü. V, Hatem, A., Bozda, D., 2013. Benchmarking short sequence mapping
 328 tools. *BMC Bioinformatics* 7, 184. doi:10.1186/1471-2105-14-184.
- 329 Wu, S., Manber, U., 1992. A fast approximate pattern-matching tool. *Usenix Winter 1992 Tech.*
 330 *Conf.*
- 331 Zablocki, O., van Zyl, L., Adriaenssens, E.M., Rubagotti, E., Tuffin, M., Cary, S.C., Cowan, D.,
 332 2014. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like
 333 elements in the metaviromes of antarctic soils. *Appl. Environ. Microbiol.* 80, 6888–6897.
 334 doi:10.1128/AEM.01525-14.

335

336

337 **Table**

338 **Table 1.** Recapitulative history of the original (our data) and already published metagenomic
 339 datasets that were used in this study.

340

Origin of samples	Technic of virus-like particles enrichment	Technic of sequencing	Cleaned reads median length	Total read number	Reference
Insects 1	0.45 µm filtration, DNA and RNA extraction, random PCR amplification	MiSeq Illumina	219	324 246	Our data
Insects 2			230	399 954	
Insects 3			228	428 089	
Insects 4			217	611 722	
Insects 5			225	203 015	
Insects 6			212	343 955	
Plants 1			195	224 890	
Plants 2			245	440 408	
Estuary	0.2 µm filtration, DNA extraction, RCA amplification	454 pyrosequencing	105	294 068	26
Lake Bourget			471	593 084	27
Lake Pavin			445	649 290	
Antarctic open soil		MiSeq Illumina	250	870 687	28
Antarctic hypolith			250	1 057 555	
Human feces		454	466	504 646	25
Mosquitoes		pyrosequencing	104	336 760	24

341

342 **Figures**

343 **Fig. 1** Comparative analyses of two BLASTx-based methods of taxonomic assignment. A: Direct
344 taxonomic assignment by: After the classical BLASTx-based approach; B: the *de novo* assembly,
345 mapping and BLASTx approach (AM-BLASTx).

346 **Fig. 2** Comparison on the performance of BLASTx searches. BLASTx searches were performed on
347 raw reads (blue) and after mapping on contigs (red).

348 **Fig. 3** Average proportions of reads in each virome according to their taxonomic assignment using
349 BLASTx. Query read sequences were assigned to viral (purple), bacterial (red), eukaryotic (green)
350 and unclassified (black) published sequences. Each analysis has been performed three times on a
351 random sample of 10,000 reads from each dataset.

Direct taxonomic attribution (A)



Taxonomic attribution



Dark Matter



Taxonomic attribution after de novo assembly and mapping (B)



de novo assembly



Mapping



Taxonomic attribution

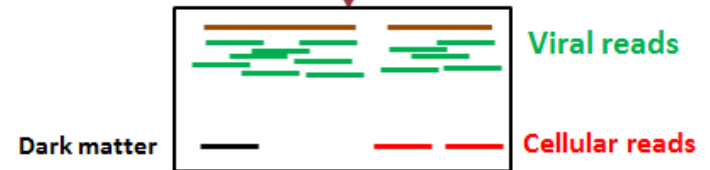


FIGURE 1 Comparative analyses of two BLASTx-based methods of taxonomic assignment. A: Direct taxonomic assignment by: After the classical BLASTx-based approach; B: the *de novo* assembly, mapping and BLASTx approach (AM-BLASTx).

Percentage of viral reads

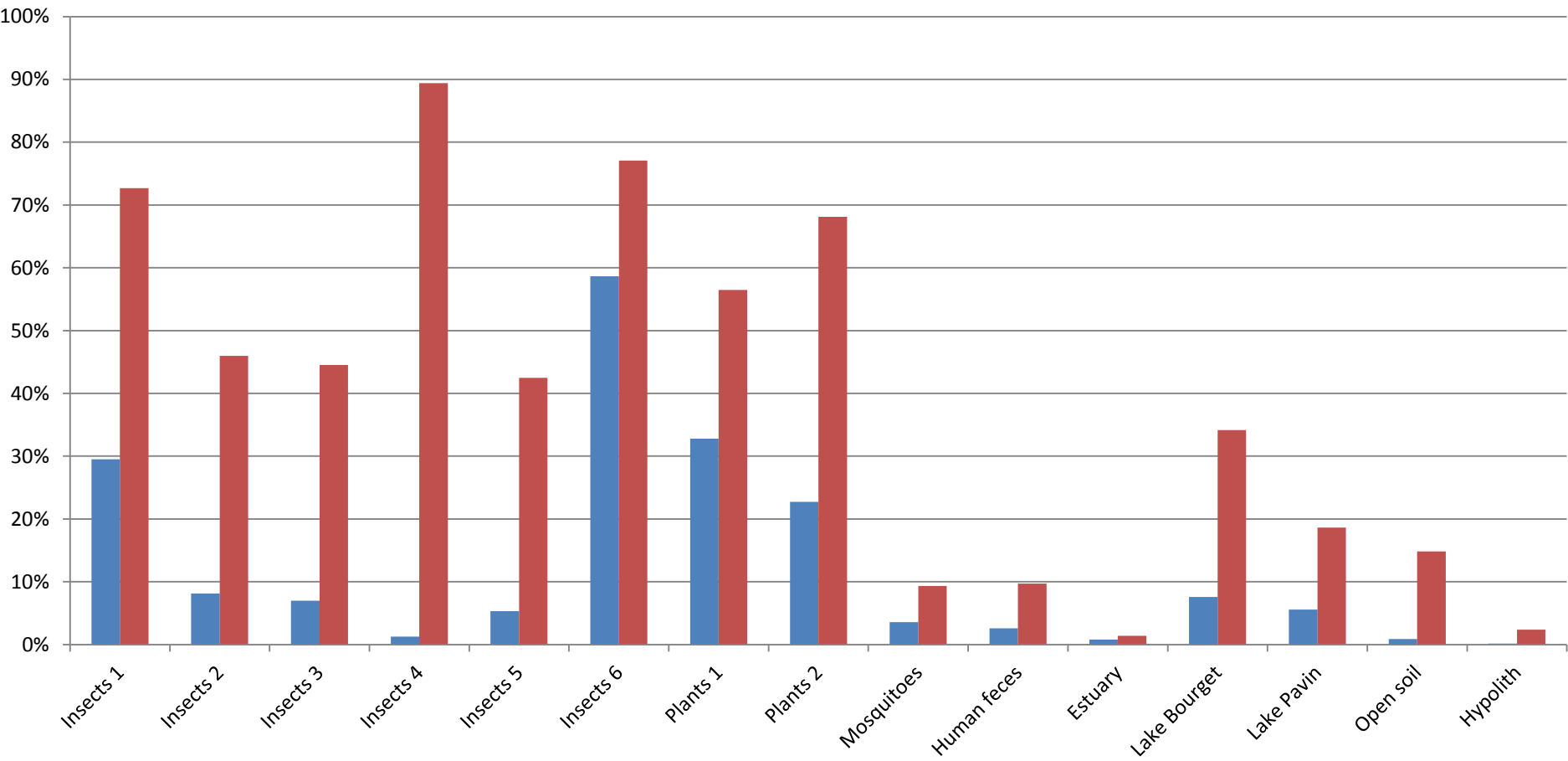


FIGURE 2 Comparison on the performance of BLASTx searches. BLASTx searches were performed on raw reads (blue) and after mapping on contigs (red).

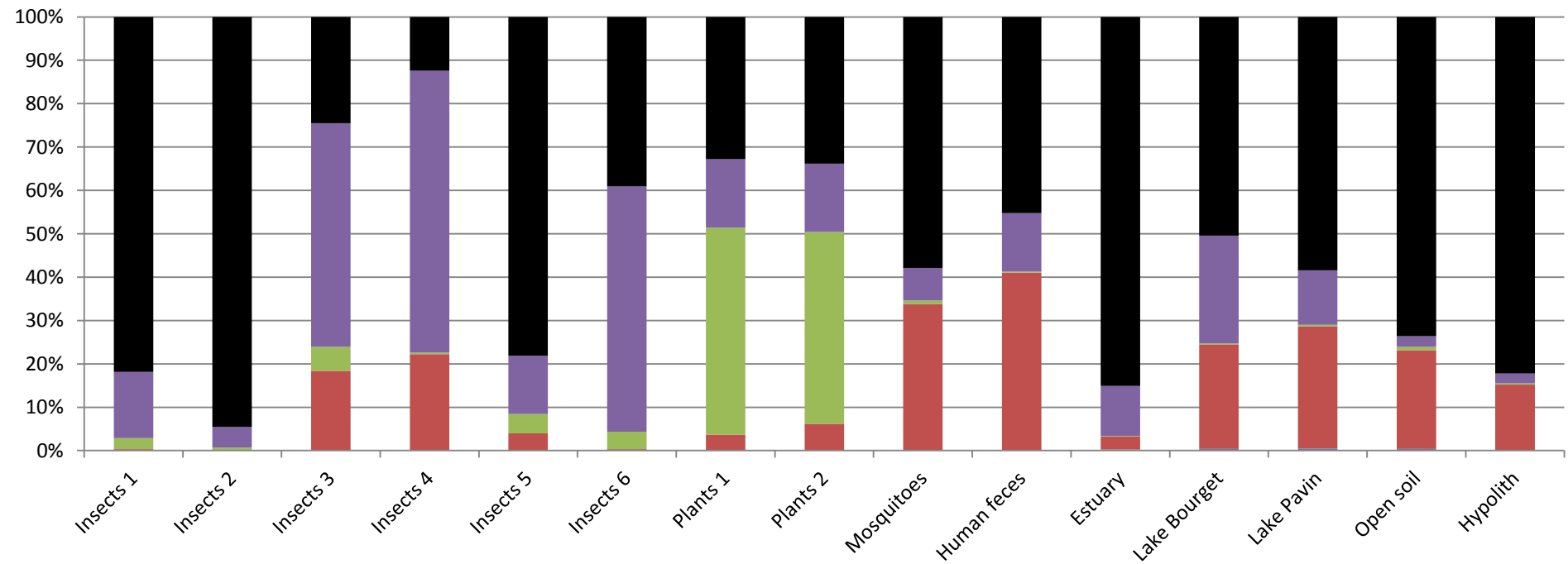


FIGURE 3 Average proportions of reads in each virome according to their taxonomic assignment using BLASTx. Query read sequences were assigned to viral (purple), bacterial (red), eukaryotic (green) and unclassified (black) published sequences. Each analysis has been performed three times on a random sample of 10,000 reads from each dataset.

352 **Supplementary materials**

353 **Supplemental Table 1:** Comparison of the performance of BLASTx searches against the viral
354 published sequence database with or without a prior read to contig mapping step. BLASTx searches
355 were performed on raw reads and after mapping of reads to contigs.

356 **Supplemental Table 2:** Five most abundant viral families found across the 15 viromes used in this
357 study

358 **Supplemental Table 3:** Global diversity in samples (BLASTx searches on 10.000 reads)

TABLE S1 Comparison of the performance of BLASTx searches against the viral published sequence database with or without a prior read to contig mapping step. BLASTx searches were performed on raw reads and after mapping of reads to contigs.

	Insects 1	Insects 2	Insects 3	Insects 4	Insects 5	Insects 6	Plants 1	Plants 2	Mosquitoes	Human feces	Estuary	Lake Bourget	Lake Pavin	Open soil	Hypolith
Number of viral reads classical BLASTx analyses	59028	16268	13983	2525	10608	117304	65600	45443	7169	5155	1599	15124	11187	1733	304
Percentage of viral reads classical BLASTx analyses	29.51%	8.13%	6.99%	1.26%	5.30%	58.65%	32.80%	22.72%	3.58%	2.58%	0.80%	7.56%	5.59%	0.87%	0.15%
Number of viral reads AM-BLASTx	145327	91962	89002	178828	84934	154178	112934	136233	18661	19426	2746	68340	37271	29688	4751
Percentage of viral reads AM-BLASTx	72.66%	45.98%	44.50%	89.41%	42.47%	77.09%	56.47%	68.12%	9.33%	9.71%	1.37%	34.17%	18.64%	14.84%	2.38%

TABLE S2 Five most abundant viral families found across the 15 viromes used in this study

Viromes	Viral Taxonomy	Genome length (kb)	Abundance rank	Host range
Insects 1	Iflaviridae	<10	1	Eukaryotic macroorganisms
	Partitiviridae	<10	2	Eukaryotic macroorganisms
	Alphaflexiviridae	<10	3	Eukaryotic macroorganisms
	Luteoviridae	<10	4	Eukaryotic macroorganisms
	Flexviridae	<10	5	Eukaryotic macroorganisms
Insects 2	Iflaviridae	<10	1	Eukaryotic macroorganisms
	Partitiviridae	<10	2	Eukaryotic macroorganisms
	Luteoviridae	<10	3	Eukaryotic macroorganisms
	Alphaflexiviridae	<10	4	Eukaryotic macroorganisms
	Tymoviridae	<10	5	Eukaryotic macroorganisms
Insects 3	Parvoviridae	<10	1	Eukaryotic macroorganisms
	Carmotetraviridae	<10	2	Eukaryotic macroorganisms
	Podoviridae	40	3	Bacteria
	Iflaviridae	<10	4	Eukaryotic macroorganisms
	Mesoniviridae	20	5	Eukaryotic macroorganisms
Insects 4	Podoviridae	40	1	Bacteria
	Parvoviridae	<10	2	Eukaryotic macroorganisms
	Iflaviridae	<10	3	Eukaryotic macroorganisms
	Tymoviridae	<10	4	Eukaryotic macroorganisms
	Luteoviridae	<10	5	Eukaryotic macroorganisms
Insects 5	Parvoviridae	<10	1	Eukaryotic macroorganisms
	Iflaviridae	<10	2	Eukaryotic macroorganisms
	Tymoviridae	<10	3	Eukaryotic macroorganisms
	Dicistroviridae	<10	4	Eukaryotic macroorganisms
	Partitiviridae	<10	5	Eukaryotic macroorganisms
Insects 6	Dicistroviridae	<10	1	Eukaryotic macroorganisms
	Nanoviridae	<10	2	Eukaryotic macroorganisms
Plants 1	Amalgaviridae	<10	1	Eukaryotic macroorganisms
	Partitiviridae	<10	2	Eukaryotic macroorganisms
	Luteoviridae	<10	3	Eukaryotic macroorganisms
	Bromoviridae	<10	4	Eukaryotic macroorganisms
	Iflaviridae	<10	5	Eukaryotic macroorganisms
Plants 2	Amalgaviridae	<10	1	Eukaryotic macroorganisms
	Partitiviridae	<10	2	Eukaryotic macroorganisms
	Tymoviridae	<10	3	Eukaryotic macroorganisms
Mosquitoes	Parvoviridae	<10	1	Eukaryotic macroorganisms
	Anelloviridae	<10	2	Eukaryotic macroorganisms
	Nudiviridae	90-230	3	Eukaryotic macroorganisms
	Microviridae	<10	4	Bacteria
	Circoviridae	<10	5	Eukaryotic macroorganisms
Human feces	Microviridae	<10	1	Bacteria
	Siphoviridae	50	2	Bacteria
	Podoviridae	40	3	Bacteria
	Myoviridae	30-250	4	Bacteria
	Phycodnaviridae	100-550	5	Eukaryotic microorganisms
Estuary	Phycodnaviridae	100-550	1	Eukaryotic microorganisms
	Myoviridae	30-250	2	Bacteria
	Circoviridae	<10	3	Eukaryotic macroorganisms
	Podoviridae	40	4	Bacteria
	Siphoviridae	50	5	Bacteria
Lake Bourget	Microviridae	<10	1	Bacteria
	Phycodnaviridae	100-550	2	Eukaryotic microorganisms
	Myoviridae	30-250	3	Bacteria
	Siphoviridae	50	4	Bacteria
	Podoviridae	40	5	Bacteria
Lake Pavin	Circoviridae	<10	1	Eukaryotic macroorganisms
	Phycodnaviridae	100-550	2	Eukaryotic microorganisms
	Siphoviridae	50	3	Bacteria
	Myoviridae	30-250	4	Bacteria
	Microviridae	<10	5	Bacteria
Open Soil	Phycodnaviridae	100-550	1	Eukaryotic microorganisms
	Podoviridae	40	2	Bacteria
	Myoviridae	30-250	3	Bacteria
	Siphoviridae	50	4	Bacteria
	Mimiviridae	1200	5	Eukaryotic microorganisms
Hypolith	Siphoviridae	50	1	Bacteria
	Myoviridae	30-250	2	Bacteria
	Phycodnaviridae	100-550	3	Eukaryotic microorganisms
	Podoviridae	40	4	Bacteria
	Mimiviridae	1200	5	Eukaryotic microorganisms

TABLE S3 Global diversity in samples (BLASTx searches on 10,000 reads)

		Taxonomy	Insects 1	Insects 2	Insects 3	Insects 4	Insects 5	Insects 6	Plants 1	Plants 2	Mosquitoes	Human feces	Estuary	Lake Bourget	Lake Pavin	Open soil	Hypolith
Number of reads	Classified	Archaea	0	0	0	0	0	0	0	1	0	11	29	50	36	28	6
		Bacteria	44	23	1848	2244	389	51	69	209	3304	3935	294	2507	2920	2356	1528
		Eukaryota	210	41	545	47	427	378	4705	5254	81	21	8	32	31	87	27
		Viruses	2000	714	5360	6515	1389	5695	3495	2143	749	1543	1241	2526	1278	231	232
		Unclassified	Unclassified	7746	9222	2247	1194	7795	3877	1732	2394	5866	4491	8427	4885	5736	7298
	Percentage of reads	Classified	Archaea	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
		Bacteria	0%	0%	18%	22%	4%	1%	1%	2%	33%	39%	3%	25%	29%	24%	15%
		Eukaryota	2%	0%	5%	0%	4%	4%	47%	53%	1%	0%	0%	0%	0%	1%	0%
		Viruses	20%	7%	54%	65%	14%	57%	35%	21%	7%	15%	12%	25%	13%	2%	2%
	Unclassified	Unclassified	77%	92%	22%	12%	78%	39%	17%	24%	59%	45%	84%	49%	57%	73%	82%