



HAL
open science

The macroecology of cancer incidences in humans is associated with large-scale assemblages of endemic infections.

Camille Jacqueline, Jessica Lee Abbate, Gabriele Sorci, Jean-François Guégan, Frédéric Thomas, Benjamin Roche

► To cite this version:

Camille Jacqueline, Jessica Lee Abbate, Gabriele Sorci, Jean-François Guégan, Frédéric Thomas, et al.. The macroecology of cancer incidences in humans is associated with large-scale assemblages of endemic infections.. *Infection, Genetics and Evolution*, 2018, 61, pp.189-196. 10.1016/j.meegid.2018.03.016 . hal-01785931

HAL Id: hal-01785931

<https://hal.science/hal-01785931>

Submitted on 21 Feb 2024

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.



Distributed under a Creative Commons Attribution 4.0 International License

Manuscript Number: MEEGID-D-17-00719R1

Title: The macroecology of cancer incidences in humans is associated with large-scale assemblages of endemic infections

Article Type: Research paper

Keywords: neglected diseases; biomes; human cancer incidences; data mining; public health strategies; pathogen-cancer interactions

Corresponding Author: Dr. Camille Jacqueline,

Corresponding Author's Institution: IRD

First Author: Camille Jacqueline

Order of Authors: Camille Jacqueline; Jessica L Abbate; Gabriele Sorci; Jean-François Guégan; Frédéric Thomas; Benjamin Roche

Abstract: It is now well supported that 20% of human cancers have an infectious causation (i.e., oncogenic agents). Accumulating evidence suggests that aside from this direct role, other infectious agents may also indirectly affect cancer epidemiology through interactions with the oncogenic agents within the wider infection community. Here, we address this hypothesis via analysis of large-scale global data to identify associations between human cancer incidence and assemblages of neglected infectious agents. We focus on a gradient of three widely-distributed cancers with an infectious cause: bladder (~2% of recorded cancer cases are due to *Shistosoma haematobium*), liver (~60% consecutive to Hepatitis B and C infection) and stomach (*Helicobacter pylori* is associated with ~70% of cases). We analyzed countries in tropical and temperate regions separately, and controlled for many confounding social and economic variables. First, we found that particular assemblages of bacteria are associated with bladder cancer incidence. Second, we observed a specific and robust association between helminths and liver cancer incidences in both biomes. Third, we show that certain assemblages of viruses may facilitate stomach cancer in tropical area, while others protect against its development in temperate countries. Finally, we discuss the implications of our results in terms of cancer prevention and highlight the necessity to consider neglected diseases, especially in tropics, to adapt public health strategies against infectious diseases and cancer.

Submission of revised version of manuscript
The macroecology of cancer incidences in humans is associated with large-scale assemblages of endemic infections

Dear Editor,

We would like to thank you for your interest in our article and for the insightful comments from the reviewers, which have helped improving significantly this manuscript through their very positive reviews.

In short, the main concerns of referee #2 were that we did not sufficiently discuss our result and interpret the findings perhaps a bit too generously in one occasion. We have therefore moderated our conclusions and expanded our discussion to potential bias in all the datasets. As a consequence, our conclusions are now much more connected to the evidence.

In addition, the referee #2 suggested a very interesting additional analysis concerning the accuracy of the models generated in our study. We have therefore expanded our material and methods section to include this new analysis. The results are available in the supplementary materials of our manuscript.

We have also addressed all the minor points risen by the referees, which have greatly improved the clarity of our manuscript.

Please do not hesitate to contact me if your need further information.

Sincerely yours,

Camille JACQUELINE
911 Avenue Agropolis
34090 Montpellier, FRANCE
Phone:+33687748572

We would like to thank the reviewers for these constructive comments. Comments are reproduced in bold and we answer in italics.

Reviewer #2 :

A very important study on the worldwide association of cancer incidences and infectious diseases incidences in humans. The manuscript is well written, and with good statistical analyses leading to robust results. This paper merits publication in IGE.

I have only a minor comment.

I suggest that the authors can move some of the figures given in the Annex to the main manuscript. This will attract a broader readership that this study deserves.

We thank the reviewer for this enthusiasm about our manuscript.

To follow referee's comment, we moved the Figure S2 (Infectious communities between the two temperate and tropical biomes) and S3 (Values of the first dimension of the PCA on the confounding variables at worldwide scale) to the main manuscript (see Fig 1A and B).

Reviewer #3:

I read the current manuscript with great interest. Finding a link between parasite community structure and cancer incidence sounds like an incredibly interesting prospect, and this manuscript represents one of the first steps towards this effort.

We thank the referee for this enthusiasm about our manuscript.

However, I find many issues with the current manuscript that I believe preclude it from publication in the current form. I detail these issues below, as well as some other more minor concerns.

Major comments

The authors 're-distribution' analysis found that type I errors were rather high (40% in one case) suggesting that any observed relationship may simply be artefactual. The authors acknowledge this, but then go on to interpret the findings perhaps a bit too generously.

We thank the reviewer for pointing this out. First, it is important mentioning that we found a type I error of 40% in only two cases, the other cases have proportion of type I error much lower. Nevertheless, we agree that we did not underline enough the risk of error in these cases. We have corrected this in the discussion (lines 316-322):

“Second, viral species were not observed among component communities linked with bladder cancer, which could be seen at odds with what has been suggested in the literature (e.g. (Kamal and El Sayed Khalifa 2006), (Oldstone 2006)). As suggested by the redistribution analysis, this result needs to be taken with precaution. Indeed, the absence of viral implication, in our study, could be due to the scarcity of *S. haematobium* infections in wealthy temperate countries as well as on the low direct contribution of the parasite to bladder cancer.”

And (lines 333-337):

“Furthermore, our results suggest the absence of protective component communities (all guilds confounded) in temperate countries. However, the redistribution analysis showed that this last result may frequently occur by chance (in about 40% of redistribution simulations). Therefore, the interpretation of this finding in the context of liver cancer epidemiology is highly speculative.”

Perhaps this is not a large issue, but the scale of the data make the messages of the manuscript a bit unclear. That is, parasite community data was at the species-level, while cancer affects individuals. Placing cancer incidence in a species-level context is a bit odd, and the interpretation of results suggesting that parasite communities at the species level correspond to enhanced cancer incidence is a bit unfounded. Parasite community level data from Gideon are based on reports of a single individual being infected. There is certainly huge variation in coinfection at the individual level (not all individuals in a given country will have all the parasites that have ever occurred in the country), so cancer incidence is potentially driven by the parasite community, or those individuals with cancer could have 0 parasites. The paper is unable to distinguish between these two possibilities.

We totally agree that the scale of our data does not allow taking into account the levels of co-infection, and therefore does not allow us to make inference. Obviously, individual data would require huge cohort with personal infection histories and cancer development status that does not exist today. Accordingly, this study does not claim to provide a definitive answer, but rather to emphasize the opportunity to study such a link. We have now acknowledged this weakness in the following paragraph (lines 386-393):

“However, our population-scale data do not reveal whether human individuals with cancer are actually infected or have been in contact with the identified diseases. In addition, our data do not allow assessing the variation in co-infection by infectious species at the individual level. While our study does not claim to provide a definitive answer, it calls for designing cohort studies of cancer patients considering personal history of infection to determine the causality, and potentially the mechanism, of such indirect link between non-oncogenic infectious agents and cancer development.”

The authors interpret negative relationships with the axes from the correspondance analysis as being indicative of a "protective component community", but the axes are orthogonal to one another, and the directionality assigned by the researchers does not really hold. The authors should perhaps focus on the underlying parasites whose distributions are best captured by the relevant axes. For example, what parasites are associated with Helm TE8? It is unlikely that Helm TE8 is explaining much variation in the parasite component communities, so why is it coming out as important?

We apologize for having not been clear enough here. Actually, we performed two distinct analyses: 1) the MCA that decomposes the variation among our parasite dataset in a constrained number of dimensions and 2) GLMMs that use the standard coordinates of the dimensions obtained with MCA to explain cancer incidences.

Then, we have identified protective component communities by dimensions that are significantly, and negatively, associated to cancer incidences in these GLMMs. To clarify all these aspects, we have added all relevant information in the "Material and methods" section (lines 172-181):

*"Within a generalized linear mixed model (GLMM) implemented in the package *lme4* (Bolker et al. 2009), transformed ASI was used as the response variable, the standard coordinates of the MCA dimensions, representing infectious assemblages, were used as fixed factors and the classes of confounding variables as a random factor. We did not consider interactions between fixed factors because of the excessively high number of potential interactions and the difficulty to interpret them. Variable selection was conducted through analysis of variance (ANOVA) (using the 'anova' function in package *car*, with test specified as "type III" that quantifies the effect of each variable after all other factors have been accounted for (Fox and Weisberg 2011)). For each significant dimension, we assessed the sign of the association with cancer incidences through coefficient values."*

We had looked at the infectious agents associated with each dimension (supplementary materials Table S5, S6 and S7). However, as the information at species level may appear biased because only non-worldwide pathogens are included in our analyses, we have preferred to discuss our results at the community level in the main manuscript.

The rarefaction analysis found a minimum bound of data percentage required before a significant association was observed based on random sampling of the existing data. For some axes, this percentage was quite high (or absolute in some cases ... 100% of the data was needed before a significant effect was observed).

Is it possible to assess model accuracy? That is, could the authors use some error (mean squared error) or accuracy (percent cases correctly classified, sensitivity, specificity, AUC, etc.) to quantify how well the models actually perform. This would give more

support to the argument that the parasite community is truly related to cancer incidence at the population scale.

We thank the reviewer for highlighting this point and we agree that assessing model accuracy would be relevant in our study. Consequently, we have run more analyses and added the following paragraphs in the method section (lines 183-185):

“In order to test the accuracy of our GLMMs to fit population data, we calculated several error and accuracy statistics for all the models generated during the principal analyses (using the ‘accuracy’ function in the package *rcompanion*; see Table S3).”

And in the result’s section (lines 231-232):

“We found that GLMMs fit the actual data relatively well with an accuracy between 70% and 90% (see Table S3).”

Minor comments

We thank the reviewer for the careful proofreading and all the minor changes have been taken into account.

keywords are overgeneral (infections, communities, cancer, statistics)

We changed them:

“neglected diseases, biomes, human cancer incidence, data mining, public health strategies, pathogen-cancer interactions”

lines 47-53: break into 2 sentences

It has been done (lines 48-54):

“The term “infectious agents” describes the transmissible organisms that require living in or on another organism (the “host”) to complete its lifecycle – a process which decreases host fitness – and includes virus, bacteria, fungi, protozoans, helminths, etc... Since the 20th century, numerous infectious agents have been recognized as risk factors for the development of several cancers (i.e., acting through inflammation or introduction of foreign DNA into cells; henceforth referred to “oncogenic agents”) (Zur Hausen and Villiers 2015).”

line 141: include a brief discussion of potential reporting bias. The authors do a good job addressing the inherent sampling biases of parasites, but don’t spend an equivalent time addressing other potential biases in either dataset.

We have extended this point in the discussion addressing the question of potential sampling bias in cancer incidence data and confounding variables (lines 376-384):

“Furthermore, the detection of cancer cases could be better in temperate countries because of the higher investment in health and disease surveillance. However, we assessed this disparity using two summarizing indices provided by the IARC, which considers the methods and data quality. These indices have been considered as confounding variables (see supplementary materials). Finally, reporting bias could also affect other confounding variables. However, we have selected data from reference organizations such as the WHO and the World Bank which are susceptible to be the more accurate. Even if it could quantitatively impact our results, it is unlikely to influence them qualitatively as the information has been compiled in a unique variable used as random factor.”

line 192 "an ad+hoc rarefaction"

Done.

line 203: the word ‘signification’ sounds really awkward. This is stylistic though, and the authors don’t have to change it if they don’t want to.

We have changed it.

“Finally, as caveat of large-data analysis can be to find associations without biological meaning, we calculated the risk to observe significant combination of infectious assemblages in association with cancer incidence by random chance.”

line 337: 'associated with large component communities'. I don’t understand this. Large meaning species-rich?

We agree that the term was confusing, we have modified it.

“As the oncogenic agent *H. pylori* is highly prevalent worldwide (Atherton and Blaser 2009), stomach cancer had a particular likelihood to be associated with species-rich component communities in both biomes”.

linea 340-342: more discussion of the potential mechanisms underlying this difference between biomes is warranted.

We agree that it is an interesting finding and thus we have expanded our discussion about the contrasting role of viruses between the two biomes (lines 343-351):

“We hypothesize that it could rely on different sequences of infections between these two regions. In temperate countries, the highest incidence of viral diseases occurs among children younger than 10 years (Seward *et al.* 2002). However, *H. pylori* is also acquired at a very young age and the risk of infection declines rapidly after 5 years of age (Rowland *et al.* 2006). In temperate biomes, early activation of Th1 responses by viruses, even before *H. pylori* infection, could thus bring protection against *H. pylori* through cross immunity (Quiding-Järbrink *et al.* 2001). Conversely, viruses are acquired at older ages in tropical countries with a higher proportion of cases and higher susceptibility among adults (Lee 1998). In this situation, viral infections may temporarily divert the Th1 responses from *H. pylori* and allow it to proliferate (Figure 3).”

line 360: 'open date area' is a typo, and could potentially better be stated as 'facilitated by future development of openly available data sources...' or something similiar.

We thank the reviewer for pointing this out. We have modified the sentence according to the referee's suggestion (lines 361-364):

“Such approach should be facilitated by future development of openly available data sources and may allow assessing the global component community, including worldwide-distributed infectious agents, which may actually have a clearer impact on cancer incidence.”

Highlights :

- Specific assemblages of bacteria are associated with higher incidences of bladder cancer at country scale.
- Virus species could modify the persistence of oncogenic agent responsible of stomach cancer.
- The force of interaction between communities of infectious agents and cancer incidences is linked to the prevalence of the oncogenic agent and its causality in associated cancer.

19 **Abstract:**

20 It is now well supported that 20% of human cancers have an infectious causation (i.e.,
21 oncogenic agents). Accumulating evidence suggests that aside from this direct role, other
22 infectious agents may also indirectly affect cancer epidemiology through interactions with the
23 oncogenic agents within the wider infection community. Here, we address this hypothesis via
24 analysis of large-scale global data to identify associations between human cancer incidence
25 and assemblages of neglected infectious agents. We focus on a gradient of three widely-
26 distributed cancers with an infectious cause: bladder (~2% of recorded cancer cases are due to
27 *Shistosoma haematobium*), liver (~60% consecutive to Hepatitis B and C infection) and
28 stomach (*Helicobacter pylori* is associated with ~70% of cases). We analyzed countries in
29 tropical and temperate regions separately, and controlled for many confounding social and
30 economic variables. First, we found that particular assemblages of bacteria are associated with
31 bladder cancer incidence. Second, we observed a specific and robust association between
32 helminths and liver cancer incidences in both biomes. Third, we show that certain
33 assemblages of viruses may facilitate stomach cancer in tropical area, while others protect
34 against its development in temperate countries. Finally, we discuss the implications of our
35 results in terms of cancer prevention and highlight the necessity to consider neglected
36 diseases, especially in tropics, to adapt public health strategies against infectious diseases and
37 cancer.

38
39 **Keywords:** neglected diseases, biomes, human cancer incidences, data mining, public health
40 strategies, pathogen-cancer interactions.

41 **Introduction**

42 While cancer remains one of the main causes of death in Western countries (Ferlay et
43 al., 2010), its burden is increasing in low- and middle-income countries (Magrath et al.,
44 2013). Although treatments, including chemo-, radio- and/or immunotherapy, have resulted in
45 a slight decline in death rates worldwide, cancer incidence rates have remained stable over the
46 past 10 years (Siegel et al., 2013). In this context, prevention seems currently the most
47 efficient strategy to reduce the impact of cancer on populations. The term “infectious agents”
48 describes the transmissible organisms that require living in or on another organism (the
49 “host”) to complete its lifecycle – a process which decreases host fitness – and includes virus,
50 bacteria, fungi, protozoans, helminths, etc... Since the 20th century, numerous infectious
51 agents have been recognized as risk factors for the development of several cancers (i.e., acting
52 through inflammation or introduction of foreign DNA into cells; henceforth referred to
53 “oncogenic agents”) (Zur Hausen and Villiers, 2015). For example, current evidence links
54 Epstein–Barr virus (EBV), Hepatitis B and C viruses (HBV, HCV), the bacteria *Helicobacter*
55 *pylori*, human papillomavirus (HPV), and the trematode *Schistosoma haematobium* to cancers
56 of the lymph nodes, liver, stomach, cervix and bladder respectively. Consequently,
57 vaccination against oncogenic agents is currently being considered as a potential key tool to
58 prevent these cancers, such as the HPV vaccine which offers a relative protection against
59 cervical cancer (Paavonen et al., 2009).

60 Accumulating evidence suggests that aside from this direct role, infectious agents
61 could also be indirectly involved in cancer epidemiology through interactions with oncogenic
62 agents in infra-communities (reviewed in (Jacqueline et al., 2017)). Indeed, hosts are
63 commonly infected by multiple infectious species simultaneously (Read and Taylor, 2001),
64 and within-host competition for resources and through immunity has been fairly well-
65 established (reviewed in (Mideo, 2009)). A number of infectious organisms are known to
66 impair immune system homeostasis with both positive and negative consequences for their
67 competitors, especially through the well-known trade-off between the Th1/Th17 and Th2
68 immune pathways (Pedersen and Fenton, 2007). For example, helminth infections are often
69 responsible for an up-regulation of the Th2 response, which has been linked to tuberculosis
70 reactivation and increase in likelihood of HIV infection (Borkow et al., 2001). Furthermore,
71 these interactions at individual scale are susceptible to have consequences on persistence and

72 transmission dynamics of the infectious agents at the population level (Ezenwa and Jolles,
73 2014).

74 A few studies have explicitly considered oncogenic agents as part of an infectious
75 community. The majority of the literature available on interactions between oncogenic and
76 non-oncogenic agents concerns species that are widely distributed and with high prevalence.
77 For instance, *Plasmodium falciparum*, the agent of malaria, increases the replication of EBV,
78 an oncogenic virus associated with Burkitt Lymphoma, through its impairment of an effective
79 immune response (Chêne et al., 2007; Morrow et al., 1976). Epidemiological studies have
80 also shown that infection with *Chlamydia trachomatis* increases the risk for persistence of
81 HPV infection, some lineages of which can lead to cervical cancer (Silins et al., 2005).
82 However, these non-oncogenic agents, with high prevalence, may not be representative of the
83 whole diversity of infectious organisms, especially in tropical countries that are affected by a
84 high number of endemic and rare diseases that are relatively under-studied (Hotez et al.,
85 2007).

86 Here, we explore the hypothesis that co-infections with endemic infections, in both
87 tropical and temperate latitudes, modify the persistence of oncogenic agents and thus interact
88 with the development of some infectious cancer in human population. For most countries,
89 prevalence of oncogenic agents is not available for the whole population and thus we
90 postulate that infectious agents favoring persistence of oncogenic agents should co-localize in
91 countries with higher cancer incidences. We focus on three widely-distributed cancers
92 (bladder, liver and stomach), for which an infectious causation is widely recognized and for
93 which sufficient large-scale data are available. These three examples allow considering
94 oncogenic agents that belong to different taxonomic guilds (helminthes, viruses and bacteria)
95 as well as a gradient in the frequency with which each cancer is known to have an infectious
96 origin (from 2% of recorded cases for bladder cancer, to 60% for liver cancer and 70% for
97 stomach cancer). By compiling and analyzing a global database, we identify associations
98 between the country-level incidence of cancer and the presence or absence of infectious
99 agents in each country. We then present hypotheses regarding the mechanism behind each
100 association, and discuss the potential consequences of our findings in terms of cancer
101 prevention strategies.

102

103 **Material & methods**

104 *Data overview and inclusion criteria*

105 Our cancer data were obtained from the International Agency for Research on Cancer
106 (IARC GLOBOCAN project, 2012, <http://globocan.iarc.fr/>). Among 48 human cancers in
107 women and men across 184 countries, we selected three cancers of interest (bladder, liver and
108 stomach) because they are recognized to have an infectious causation. First, the bacterium *H.*
109 *pylori* is responsible for 80% of stomach adenocarcinomas (Zur Hausen, 2009), which
110 represent 90% of stomach cancers (Brenner et al., 2009). Second, 75% of liver cancers are
111 hepatocellular carcinomas (HCCs) (Parkin, 2001) which have been linked in more than 80%
112 of cases to HCV and HBV infection (Parkin, 2006). Finally, squamous cell carcinomas
113 (SCCs), representing 5% of bladder cancers (Kantor et al., 1988), have been associated with
114 the helminth *Schistosoma haematobium* in approximately 30% of cases (Mostafa and
115 Sheweita, 1999). In our study, we used age-standardized incidence (ASI), which represents
116 the raw cancer incidence when country age structure is extrapolated to a standard one (World
117 Standard Population (Doll et al., 1996)).

118 Our analyses on infectious species relied on the GIDEON database (Global Infectious
119 Diseases and Epidemiology Network, <http://web.gideononline.com/>). This database contained
120 a presence/absence matrix for a total of 370 human infectious agents across 224 countries
121 (2004 update). We removed all infectious agents that were present or absent in all countries,
122 because they did not add any discriminating information, as well as infections caused by
123 fungi, protozoans and arthropods guilds which were represented by less than 10 infectious
124 agents respectively. This yielded a data subset with infectious agents belonging to three main
125 guilds (viruses, bacteria and helminths). It included very rare diseases such as Buruli ulcer
126 (3,000 cases/y worldwide; www.who.int/gho/neglected_diseases/buruli_ulcer) and tick-borne
127 encephalitis (11,000 cases/y worldwide; www.who.int/immunization/topics/tick_encephalitis)
128 to more widespread diseases affecting a higher number of persons such as leishmaniasis and
129 echinococcosis (1 million cases/y worldwide; <http://www.who.int/echinococcosis/>).

130 Finally, we constituted a database of 24 potential confounding variables for 167
131 countries from the Food and Agriculture Organization (FAO), World Health Organization

132 (WHO) and the World Bank (WB) to consider the diversity of social factors and economic
133 levels (Supplementary Table S1 for a description).

134 The dataset was analyzed separately between tropical (geographically situated between
135 the two tropics) and temperate (above or below tropics) countries in order to better control for
136 the huge environmental (in addition to social and economical) disparities between these two
137 biomes (distinct biological communities formed in response to a shared physical climate
138 (Cain et al., 2011)). In fact, tropical countries have a higher abundance/richness of infectious
139 species (Guernier et al., 2004) and are associated on average with lower socio-economic
140 wealth (World Bank 2016) compared to temperate countries. We confirmed that these two
141 groups of countries, made on geographical assumptions, correspond to two distinct biomes
142 which present distinct patterns of infectious species richness (Supplementary Fig. S1) and
143 composition (Fig.1A) as well as socio-economical disparities (Fig. 1B).

144

145 *Assessment of infectious agent assemblages*

146 Our database of infectious agents, after removing non-relevant species as previously
147 described, contained presence/absence values for 101 human infectious agents across 103
148 tropical countries and 88 infectious agents across 74 temperate countries. We first aimed to
149 characterize infectious assemblages within each guild (i.e. viruses, bacteria and helminths).
150 To do so, we ran Multiple Correspondence Analysis (MCA; FactomineR Lê *et al.* 2008) in
151 order to reduce the number of variables to a set of dimensions representing the
152 presence/absence data of all infectious agents within each guild. The composition of these
153 dimensions was then examined as a synthetic measure of assemblage structure. The number
154 of dimensions considered was calculated according to the Kaiser's criterion (dimensions for
155 which the percentage of variance explained was superior to a threshold calculated as follows:
156 $100\% / \text{total number of dimensions estimated by MCA analysis (Kaiser, 1958)}$). For each of
157 these selected dimensions, we calculated standard coordinates for each country by dividing
158 the principal coordinates (i.e. loadings) by the square root of the dimension's eigenvalue
159 (Husson et al., 2017). This standardization allowed considering each dimension as
160 independent variables.

161 We followed a similar methodology for the 24 confounding variables described
162 previously (substituting the Multiple Correspondence Analysis with a Principal Component
163 Analysis, as our confounding variables contained quantitative values rather than categorical
164 presence/absence). Given that the first dimension of the PCA on confounding variables
165 explained more than 40% of the variance within each biome (Supplementary materials Fig.
166 S2), we classified the values of this dimension into five evenly distributed classes of
167 countries. These classes were then used as a random factor in statistical models to consider
168 intra-biome variability in social and economic level (see below).

169 We conducted separate analyses for each guild (viruses, bacteria, helminths) within
170 each biome (tropical and temperate) and we transformed cancer incidence data (ASI) to obtain
171 Gaussian distributions (transformations are detailed in supplementary materials Table S2).
172 Within a generalized linear mixed model (GLMM) implemented in the package *lme4* (Bolker
173 et al. 2009), transformed ASI was used as the response variable, the standard coordinates of
174 the MCA dimensions, representing infectious assemblages, were used as fixed factors and the
175 classes of confounding variables as a random factor. We did not consider interactions between
176 fixed factors because of the excessively high number of potential interactions and the
177 difficulty to interpret them. Variable selection was conducted through analysis of variance
178 (ANOVA) (using the ‘anova’ function in package *car*, with test specified as “type III” that
179 quantifies the effect of each variable after all other factors have been accounted for (Fox and
180 Weisberg, 2011)). For each significant dimension, we assessed the sign of the association
181 with cancer incidences through coefficient values. We will henceforth refer to this statistical
182 analysis as the “principal analysis”. In order to test the accuracy of our GLMMs to fit
183 population data, we calculated several error and accuracy statistics for all the models
184 generated during the principal analysis (using the ‘accuracy’ function in the package
185 *rcompanion*; see Table S3).

186 After having considered each guild independently (one GLMM by guild/by cancer),
187 we ran a GLMM for each cancer (as described previously) considering all guilds together. In
188 this relative contribution analyses, we used only the significant dimensions identified in the
189 principal analysis as fixed factors (see Table S4).

190 Finally, for each significant MCA dimensions involved in the principal analysis, we
191 identified the infectious agents that explained more than 5% of the variability of the significant
192 dimensions (represented by “eta²” values in the MCA). For each cancer, we grouped all of

193 these infectious agents (separately for those with positive versus negative correlation with
194 cancer incidences), and we referred to each of these groups as an “infectious assemblage”. As
195 the agents in each assemblage may not necessarily interact with one-another, we distinguished
196 the term “assemblage” from “community”. We described the size and the composition of the
197 assemblages for each cancer in Tables S5, S6 and S7 of the Supplementary Materials.

198

199 *Robustness assessment*

200 The robustness of our results to heterogeneities in sampling was tested using ad+hoc
201 rarefaction analysis. We generated random samples of countries containing from 20% to
202 100% (in increments of 10% for each new random sample) of the entire database. For each
203 percentage, we repeated the random sampling 10 times and ran the primary analysis (for each
204 biomes, cancer and guilds) on each of these repetitions. We reported the percentage of the
205 database for which the median p-value (calculated on the 10 repetitions) becomes significant.
206 By doing this, we can “score” the robustness of each result and exclude dimensions that were
207 significant due only to outlier countries. It is worth pointing out that this score aims
208 quantifying how our different conclusions are relatively robust each against others. Therefore,
209 there is no significance threshold for this score.

210 Finally, as caveat of large-data analysis can be to find associations without biological
211 meaning, we calculated the risk to observe significant combination of infectious assemblages
212 in association with cancer incidence by random chance. This is the equivalent of type-I
213 statistical error and we aimed to assess the percentage of chance to incorrectly reject the null
214 hypothesis. We conducted a redistribution analysis by simulating normally-distributed
215 incidence of cancer. For each cancer of interest, we generated 10,000 random reassortments
216 of cancer incidences across countries with the same mean and variance as in the original
217 database in each biome. GLMMs were conducted as described previously for each guild of
218 infectious agents. We reported the percentage of reassortments for which we observe the same
219 combination of significant dimensions, across the three guilds, detected by the principal
220 analysis. A low percentage suggests that the risk of detecting a combination with no
221 biological meaning is low. As for the rarefaction analysis, this percentage quantifies relatively
222 how our conclusion can be found just by chance, which does not call for a threshold.

223 The statistical approach is summarized in Fig. 2 and all the analyses have been
224 conducted using R v3.1.2 statistical software (R Development Core Team).

225

226 **Results**

227 After applying MCA on each guild and removing the non-explicative dimensions
228 according to the Kaiser criterion, we obtained 7 dimensions explaining variance for bacterial
229 species occurrence, 9 dimensions for helminths and 8 for viruses in temperate countries. For
230 tropical countries, 8, 10 and 11 dimensions were kept for bacteria, helminths and viruses,
231 respectively. We found that GLMMs fit the actual data relatively well with an accuracy
232 between 70% and 90% (see Table S3). Because our results were similar for both sexes (see
233 Table S8), we report here only the results for females (Table 1).

234 *Bladder cancer*

235 In temperate biome, bladder cancer was negatively associated with one dimension of
236 bacteria (BACT TE1) and one dimension of helminths (HELM TE8), but positively
237 associated with a second dimension of bacteria (BACT TE4) (Table 1). Two dimensions (one
238 for bacteria (BACT T4) and one for helminths (HELM T9)) were positively associated with
239 bladder cancer incidence in tropical countries. The associations found with bacteria were well
240 supported in both biomes in the relative contribution analysis (see Table S4), as well as with
241 the rarefaction analysis (Table 1). Such associations were found in 12% of simulated
242 normally-distributed incidences (Table 2). In addition, the redistribution analysis showed that
243 the absence of virus dimensions in both temperate and tropical countries could be expected in
244 36% of trials by chance. Regarding assemblage composition two biomes confounded, we
245 noticed that the negatively associated assemblage is composed of 10 bacteria and seven
246 helminths species. The positively associated assemblage contains 17 bacteria and 6 helminths
247 species (see Table S5).

248

249 *Liver cancer*

250 One dimension for bacteria (Bact TE1 and Bact T6) and one for helminths (Helm TE1
251 and Helm T5) were positively associated with liver cancer incidence in both temperate and
252 tropical countries. However, the presence of negatively associated dimensions (Helm T6 and

253 Virus T6) was observed only in tropical countries. Helminths represented the only guild
254 which was both positively and negatively correlated to this cancer incidence. The distinct
255 pattern of associations between the two biomes for negatively associated dimensions was
256 found in 40% of simulations in the redistribution analysis (Table 2). In addition, rarefaction
257 analysis and relative contribution of dimensions only highlighted the negative association
258 between helminths and liver cancer in tropical countries. In accordance with this result, the
259 association between any helminth dimensions and liver cancer was detected in 16% of
260 random trials. Assemblages differed greatly depending on the sign of the association with
261 liver cancer incidence; while positive associated assemblage contained 12 bacteria and 25
262 helminths, the negative associated assemblage was formed by 9 species of helminths and 14
263 viruses (See Table S6 in supplementary materials).

264

265 *Stomach cancer*

266 Compared to bladder and liver cancers, stomach cancer showed the highest number of
267 associated dimensions with two for bacteria (Bact TE1 and TE4), four for helminths (Helm
268 TE1, TE2, TE3, TE7) and two for viruses (Virus TE3 and TE8) in temperate countries and 6
269 dimensions (Bact T2 and T4, Helm T2 and Helm T3, Virus T1 and T5) in tropical countries
270 (Table 1). We found more dimensions and a higher number of species representing helminth
271 guild in temperate countries. In addition, we noticed that associations between viruses and
272 stomach cancer were different between the two biomes. While some dimensions of viruses
273 were positively associated with incidence in tropical countries, others were negatively
274 correlated in temperate countries. We showed that this last combination should occur by
275 chance in just 3% of cases. While most of the dimensions were well supported by rarefaction
276 analysis, the relative contribution of the dimension showed that only dimensions of helminths
277 and bacteria in temperate countries were retained. Finally, the assemblage negatively
278 associated with stomach cancer incidence (two biomes confounded) showed a lower diversity
279 overall with 69 species than positively associated assemblage which were composed of 101
280 human infections (See Table S7 in supplementary materials).

281

282 **Discussion**

283 Through a large-scale statistical analysis of global presence/absence of infectious
284 agents and cancer incidence data, we found that three well-distributed human cancers with
285 well-accepted infectious causation were associated with different species assemblages, as
286 described in Table 1. Overall, no common patterns were identified across the three cancers
287 but rather specific associations between each biome's composition of infectious agents and
288 cancer incidences. However, cancers that we found to be associated with higher number of
289 dimensions as well as higher number of infectious agents are also those known to have high
290 level of infectious causality by highly prevalent oncogenic agents (Table 3). This is
291 particularly striking for cancers caused by oncogenic agents at the extremes of the causality
292 gradient such as *H. pylori* and *S. haematobium*.

293 We postulate that the assemblages negatively associated with cancer incidences
294 formed a potential "protective component community" which includes all of the
295 infracommunities within a host population and begs further investigation of an underlying
296 mechanism. Alternatively, positively associated assemblages will be referred as "facilitating
297 component community" in the following sections. Our results emphasize the possibility that
298 certain infectious agents influence the persistence or the circulation of oncogenic agents, and
299 the urgent need to investigate the details of infectious species interactions in the context of
300 cancer prevention.

301 We noticed two specific patterns in component community associated with bladder
302 cancer, of which 2% of cases are thought to be due to infection with the trematode
303 *Schistosoma haematobium*. It is worth mentioning that this trematode has an heterogeneous
304 transmission pattern, with an extremely high prevalence in tropical countries, but circulating
305 only in several countries within the temperate biome. First, we found that bacteria may have a
306 preponderant role in both biomes. Interestingly, studies have shown that bacteria-helminth
307 interactions have already been observed on the field and could impact cancer risk. Notably, a
308 high percentage of co-infection with *S. haematobium* and some unspecified bacteria in the
309 urinary tract has been reported (Adeyeba and Ojeaga, 2002; Ossai et al., 2014). These co-
310 infections could increase the risk of bladder cancer as bacteria in the urinary tract produce
311 nitrosamines, which are carcinogenic compounds (Davis et al., 1984). In addition, bacteria
312 may bias the Th1/Th2 balance away from protection against helminths by down-regulating
313 Th2-mediated responses. This kind of indirect interaction has already been seen in rabbits

314 where the bacterium *Bordetella bronchiseptica* was shown to enhance helminth intensity
315 through immune mediated effects (Pathak et al., 2012). Thus, our study may highlight the role
316 of bacteria in cancer associated with helminth infections. Second, viral species were not
317 observed among component communities linked with bladder cancer, which could be seen at
318 odds with what has been suggested in the literature (e.g. (Kamal and El Sayed Khalifa, 2006),
319 (Oldstone 2006)). As suggested by the redistribution analysis, this result needs to be taken
320 with precaution. Indeed, the absence of viral implication, in our study, could be due to the
321 scarcity of *S. haematobium* infections in wealthy temperate countries as well as on the low
322 direct contribution of the parasite to bladder cancer.

323 Regarding liver cancer, we observed a specific and robust association between
324 helminths and cancer incidences in both biomes. The interactions between certain helminths
325 species and HCV/HBV have been reported in experimental studies and may be either
326 protective or facilitator. It has been suggested that co-infection with *S. mansoni* could increase
327 the persistence and severity of HCV infection because the helminth prevent the production of
328 an HCV-specific CD4+/Th1 T cell response (Kamal et al., 2001). Conversely, a recent study
329 has demonstrated the immunostimulant effect of a protein derived from *Onchocerca volvulus*
330 which increases the interferon- γ response to HCV in vitro (MacDonald et al., 2008). Even if
331 the identified component communities do not highlight these two particular species, it seems
332 worth to expect that similar mechanisms could apply to other helminth species (McSorley and
333 Maizels, 2012). Furthermore, our results suggest the absence of protective component
334 communities (all guilds confounded) in temperate countries. However, the redistribution
335 analysis showed that this last result may frequently occur by chance (in about 40% of
336 redistribution simulations). Therefore, the interpretation of this finding in the context of liver
337 cancer epidemiology is highly speculative.

338 As the oncogenic agent *H. pylori* is highly prevalent worldwide (Atherton and Blaser,
339 2009), stomach cancer had a particular likelihood to be associated with species-rich
340 component communities in both biomes. This was the case, with the largest number and
341 diversity of species across guilds associating both positively and negatively with this cancer.
342 The principal result for stomach cancer is that component communities of viruses have an
343 opposite effect between the two biomes. We hypothesize that it could rely on different
344 sequences of infections between these two regions. In temperate countries, the highest

345 incidence of viral diseases occurs among children younger than 10 years (Seward et al.,
346 2002). In temperate biomes, early activation of Th1 responses by viruses, even before *H.*
347 *pylori* infection, could thus bring protection against *H. pylori* through cross immunity
348 (Quiding-Järbrink et al., 2001). Conversely, viruses are acquired at older ages in tropical
349 countries with a higher proportion of cases and higher susceptibility among adults (Lee,
350 1998). In this situation, viral infections may temporarily divert the Th1 responses from *H.*
351 *pylori* and allow it to proliferate (Figure 3).

352 As for any statistical approach dealing with large scale database, our study is
353 susceptible to a number of issues which need to be discussed. First, some diseases based on
354 observation of symptoms have multiple potential causative agents. When this was the case,
355 we included only the agent responsible for the majority of cases. Second, a principal
356 weakness of our study is that infectious assemblages were based on presence/absence data
357 where even imported and isolated cases (observed rarely) can result in a presence assignment
358 for the whole country. In addition, all worldwide-distributed infectious agents have been
359 removed from the database because they would not be discriminating at all. These two last
360 points suggested that it would be helpful to consider prevalence data in order to assess the
361 proportion of the population which is really in contact with the infectious agent. Such
362 approach should be facilitated by future development of openly available data sources and
363 may allow assessing the global component community, including worldwide-distributed
364 infectious agents, which may actually have a clearer impact on cancer incidence. This
365 consideration, however, do not necessarily negate the indirect impact of the communities we
366 have identified. Finally, we defined guilds according to their taxonomic classifications, which
367 could be not biologically meaningful, but which allows to deal with the high number of
368 infectious agents considered here. In fact, it has been recommended that guilds should be
369 based on functional similarity of species or based on their life-cycle categories instead of their
370 taxonomy (McGill et al., 2006).

371 Considering cancer incidence data, the Globocan database (compiled by the IARC) do
372 not account for the different cancer subtypes. As oncogenic agents are implied in subtypes
373 that are only a proportion of the total number of cases, our conclusions may be more robust
374 when infection is associated with a prevalent cancer (stomach adenocarcinoma and
375 hepatocellular carcinoma) as opposed to rarer subtypes (e.g., SCCs of the bladder).

376 Furthermore, the detection of cancer cases could be better in temperate countries because of
377 the higher investment in health and disease surveillance. However, we assessed this disparity
378 using two summarizing indices provided by the IARC, which considers the methods and data
379 quality. These indices have been considered as confounding variables (see supplementary
380 materials). Finally, reporting bias could also affect other confounding variables. However, we
381 have selected data from reference organizations such as the WHO and the World Bank which
382 are susceptible to be the more accurate. Even if it could quantitatively impact our results, it is
383 unlikely to influence them qualitatively as the information has been compiled in a unique
384 variable used as random factor.

385 We have based our interpretations here on the hypothesis that infectious agents may
386 modify the circulation or the persistence of oncogenic agents. However, our population-scale
387 data do not reveal whether human individuals with cancer are actually infected or have been
388 in contact with the identified diseases. In addition, our data do not allow assessing the
389 variation in co-infection by infectious species at the individual level. While our study does not
390 claim to provide a definitive answer, it calls for designing cohort studies of cancer patients
391 considering personal history of infection to determine the causality, and potentially the
392 mechanism, of such indirect link between non-oncogenic infectious agents and cancer
393 development. Finally, the inclusion of a cancer for which no infectious agent is suspected to
394 play a role would have served as an evidence-boosting negative control. However, the
395 assemblage identified in these controls could also have biological relevance as they could
396 include species that indirectly alter immunosurveillance as suggested in (Jacqueline et al.,
397 2017).

398

399 **Conclusion**

400 Despite the further steps necessary to validate the biological relevance of our findings,
401 these results draw attention to the need and potential benefits of a “pathocenosis approach”
402 for the management of infection-derived cancers, in a global health perspective. Indeed,
403 acknowledging the indirect role of infectious agents may allow for the adaptation of public
404 health strategies (such as vaccination) to improve prevention against cancer as well as the
405 targeted infectious diseases. As cancers with an infectious origin are predominant in tropical

406 countries, our study raises the perspective to used current strategies of infectious diseases
407 control (vaccination, antibiotics...) to decrease cancer burden in this region. Our approach,
408 which serves to determine associations at the largest scale (Khoury and Ioannidis, 2014), is a
409 necessary first step to motivate further experimental and cohort-based studies to confirm our
410 findings, to identify the mechanisms implicated and thus to reveal new therapeutic
411 opportunities.

412

413 **Acknowledgments**

414 The authors thank CREEC sponsors CNRS and André HOFFAMNN (Fondation
415 Mava). This paper is a contribution of the EVOCAN and STORY projects funded by the
416 Agence Nationale de la Recherche. Post-doctoral support for JLA was provided by ANR JC
417 ‘STORY’ granted to Benjamin Roche. JFG and BR were funded by an “Investissement
418 d’Avenir” Laboratoire d’Excellence Centre d’Etude de la Biodiversité Amazonienne Grant
419 (ANR-10-LABX-25-01).

420

421 **Competing interests:** No conflict of interests to declare.

422

423 **References**

424 Adeyeba, O.A., Ojeaga, S.G.T., 2002. URINARY SCHISTOSOMIASIS AND
425 CONCOMITANT URINARY TRACT PATHOGENS AMONG SCHOOL CHILDREN
426 IN METROPOLITAN IBADAN ., Afr J Biomed Res 5, 103–108.

427 Atherton, J.C., Blaser, M.J., 2009. Review series Coadaptation of Helicobacter pylori and
428 humans : ancient history , modern implications. J. Clin. Invest. 119.
429 <https://doi.org/10.1172/JCI38605DS1>

430 Borkow, G., Weisman, Z., Leng, Q., Stein, M., Kalinkovich, A., Wolday, D., Bentwich,
431 Z.V.I., 2001. Helminths , Human Immunode ciencia Virus and Tuberculosis 568–571.

432 Brenner, H., Rothenbacher, D., Arndt, V., 2009. Epidemiology of Stomach Cancer. pp. 467–
433 477. https://doi.org/10.1007/978-1-60327-492-0_23

434 Cain, M.L. (Michael L., Bowman, W.D., Hacker, S.D., 2011. Ecology. Sinauer Associates.

435 Chêne, A., Donati, D., Guerreiro-Cacais, A.O., Levitsky, V., Chen, Q., Falk, K.I., Orem, J.,
436 Kironde, F., Wahlgren, M., Bejarano, M.T., 2007. A molecular link between malaria and
437 Epstein-Barr virus reactivation. PLoS Pathog. 3, e80.
438 <https://doi.org/10.1371/journal.ppat.0030080>

439 Chitsulo, L., Engels, D., Montresor, A., Savioli, L., 2000. The global status of schistosomiasis
440 and its control. Acta Trop. 77, 41–51. [https://doi.org/10.1016/S0001-706X\(00\)00122-4](https://doi.org/10.1016/S0001-706X(00)00122-4)

441 Davis, C.P., Cohen, M.S., Gruber, M.B., Anderson, M.D., Warren, M.M., 1984. Urothelial
442 hyperplasia and neoplasia: a response to chronic urinary tract infection in rats. J Urol
443 132, 1025–31.

444 Doll, R., Payne, P., Waterhouse, J.A., 1996. Cancer Incidence in Five Continents, Vol. I
445 Union Internationale Contre le Cancer. Geneva.

446 Ezenwa, V.O., Jolles, A.E., 2014. Opposite effects of anthelmintic treatment on microbial
447 infection at individual versus population scales 8–11.

448 Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., Parkin, D.M., 2010. Estimates of
449 worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer 127, 2893–2917.
450 <https://doi.org/10.1002/ijc.25516>

451 Fox, J., Weisberg, S., 2011. An {R} Companion to Applied Regression, Second Edition.,
452 Thousand O. ed.

453 Guernier, V., Hochberg, M.E., Guégan, J.-F., 2004. Ecology drives the worldwide distribution
454 of human diseases. PLoS Biol. 2, e141. <https://doi.org/10.1371/journal.pbio.0020141>

455 Hotez, P.J., Molyneux, D.H., Fenwick, A., Kumaresan, J., Sachs, S.E., Sachs, J.D., Savioli,
456 L., 2007. Control of Neglected Tropical Diseases 1018–1027.

457 Husson, F., Josse, J., Le, S., Maintainer, J.M., 2017. Package “FactoMineR” Title
458 Multivariate Exploratory Data Analysis and Data Mining.

- 459 Jacqueline, C., Tasiemski, A., Sorci, G., Ujvari, B., Maachi, F., Missé, D., Renaud, F., Ewald,
460 P., Thomas, F., Roche, B., 2017. Infections and cancer: the “fifty shades of immunity”
461 hypothesis. BMC Cancer in press, 1–11. <https://doi.org/10.1186/s12885-017-3234-4>
- 462 Kaiser, H.F., 1958. The varimax criterion for analytic rotation in factor analysis.
463 Psychometrika 23, 187–200. <https://doi.org/10.1007/BF02289233>
- 464 Kamal, S.M., Bianchi, L., Al Tawil, A., Koziel, M., El Sayed Khalifa, K., Peter, T., Rasenack,
465 J.W., 2001. Specific Cellular Immune Response and Cytokine Patterns in Patients
466 Coinfected with Hepatitis C Virus and *Schistosoma mansoni*. J. Infect. Dis. 184, 972–
467 982. <https://doi.org/10.1086/323352>
- 468 Kamal, S.M., El Sayed Khalifa, K., 2006. Immune modulation by helminthic infections:
469 Worms and viral infections. Parasite Immunol. 28, 483–496.
470 <https://doi.org/10.1111/j.1365-3024.2006.00909.x>
- 471 Kantor, A.F., Hartge, P., Hoover, R.N., Fraumeni, J.F., 1988. Epidemiological characteristics
472 of squamous cell carcinoma and adenocarcinoma of the bladder. Cancer Res. 48, 3853–
473 5.
- 474 Khoury, M.J., Ioannidis, J.P. a., 2014. Big data meets public health. Science (80-.). 346,
475 1054–1055. <https://doi.org/10.1126/science.aaa2709>
- 476 Lee, B.W., 1998. Review of varicella zoster seroepidemiology in India and South-east Asia.
477 Trop. Med. Int. Heal. 3, 886–890. <https://doi.org/10.1046/j.1365-3156.1998.00316.x>
- 478 MacDonald, A.J., Libri, N.A., Lustigman, S., Barker, S.J., Whelan, M.A., Semper, A.E.,
479 Rosenberg, W.M., 2008. A novel, helminth-derived immunostimulant enhances human
480 recall responses to hepatitis C virus and tetanus toxoid and is dependent on CD56+ cells
481 for its action. Clin. Exp. Immunol. 152, 265–273. <https://doi.org/10.1111/j.1365-2249.2008.03623.x>
- 483 Magrath, I., Steliarova-Foucher, E., Epelman, S., Ribeiro, R.C., Harif, M., Li, C.K., Kebudi,
484 R., Macfarlane, S.D., Howard, S.C., 2013. Paediatric cancer in low-income and middle-
485 income countries. Lancet Oncol. [https://doi.org/10.1016/S1470-2045\(13\)70008-1](https://doi.org/10.1016/S1470-2045(13)70008-1)
- 486 McGill, B.J., Enquist, B.J., Weiher, E., Westoby, M., 2006. Rebuilding community ecology

487 from functional traits. *Trends Ecol. Evol.* 21, 178–85.
488 <https://doi.org/10.1016/j.tree.2006.02.002>

489 McSorley, H.J., Maizels, R.M., 2012. Helminth infections and host immune regulation. *Clin.*
490 *Microbiol. Rev.* 25, 585–608. <https://doi.org/10.1128/CMR.05040-11>

491 Mideo, N., 2009. Parasite adaptations to within-host competition. *Trends Parasitol.* 25, 261–8.
492 <https://doi.org/10.1016/j.pt.2009.03.001>

493 Mohd Hanafiah, K., Groeger, J., Flaxman, A.D., Wiersma, S.T., 2013. Global epidemiology
494 of hepatitis C virus infection: New estimates of age-specific antibody to HCV
495 seroprevalence. *Hepatology* 57, 1333–1342. <https://doi.org/10.1002/hep.26141>

496 Morrow, R.H., Gutensohn, N., Smith, P.G., 1976. Epstein-Barr Virus-Malaria Interaction
497 Models for Burkitt's Lymphoma: Implications for Preventive Trials Epstein-Barr
498 Virus-Malaria Interaction Models for Burkitt's Lymphoma: Implications for
499 Preventive Trials 1 667–669.

500 Mostafa, M.H., Sheweita, S.A., 1999. Relationship between Schistosomiasis and Bladder
501 Cancer EVIDENCE SUPPORTING THE RELATIONSHIP BETWEEN
502 SCHISTOSOMIASIS AND BLADDER 12, 97–111.

503 Oldstone, M.B.A., 2006. Viral persistence: Parameters, mechanisms and future predictions.
504 *Virology* 344, 111–118. <https://doi.org/10.1016/j.virol.2005.09.028>

505 Ossai, O.P., Dankoli, R., Nwodo, C., Tukur, D., Nsubuga, P., Ogbuabor, D., Abonyi, G.,
506 Ezeanolue, E., Nguku, P., Nwagbo, D., Idris, S., Eze, G., 2014. Bacteriuria and urinary
507 schistosomiasis in primary school children in rural communities in Enugu State , 18, 4–8.

508 Paavonen, J., Naud, P., Salmerón, J., Wheeler, C.M., Chow, S.N., Apter, D., Kitchener, H.,
509 Castellsague, X., Teixeira, J.C., Skinner, S.R., Hedrick, J., Jaisamrarn, U., Limson, G.,
510 Garland, S., Szarewski, A., Romanowski, B., Aoki, F.Y., Schwarz, T.F., Poppe, W.,
511 Bosch, F.X., Jenkins, D., Hardt, K., Zahaf, T., Descamps, D., Struyf, F., Lehtinen, M.,
512 Dubin, G., 2009. Efficacy of human papillomavirus (HPV)-16/18 AS04-adjuvanted
513 vaccine against cervical infection and precancer caused by oncogenic HPV types
514 (PATRICIA): final analysis of a double-blind, randomised study in young women.
515 *Lancet* 374, 301–314. [https://doi.org/10.1016/S0140-6736\(09\)61248-4](https://doi.org/10.1016/S0140-6736(09)61248-4)

516 Parkin, D.M., 2006. The global health burden of infection-associated cancers in the year 2002.
517 Int. J. Cancer 118, 3030–44. <https://doi.org/10.1002/ijc.21731>

518 Parkin, D.M., 2001. Global cancer statistics in the year 2000. *Lancet Oncol.* 2, 533–543.
519 [https://doi.org/10.1016/S1470-2045\(01\)00486-7](https://doi.org/10.1016/S1470-2045(01)00486-7)

520 Pathak, A.K., Pelensky, C., Boag, B., Cattadori, I.M., 2012. Immuno-epidemiology of chronic
521 bacterial and helminth co-infections: observations from the field and evidence from the
522 laboratory. *Int. J. Parasitol.* 42, 647–55. <https://doi.org/10.1016/j.ijpara.2012.04.011>

523 Pedersen, A.B., Fenton, A., 2007. Emphasizing the ecology in parasite community ecology.
524 *Trends Ecol. Evol.* 22, 133–139. <https://doi.org/10.1016/j.tree.2006.11.005>

525 Peleteiro, B., Bastos, A., Ferro, A., Lunet, N., 2014. Prevalence of *Helicobacter pylori*
526 infection worldwide: A systematic review of studies with national coverage. *Dig. Dis.*
527 *Sci.* 59, 1698–1709. <https://doi.org/10.1007/s10620-014-3063-0>

528 Quiding-Järbrink, M., Lundin, B.S., Lönröth, H., Svennerholm, A.M., 2001. CD4+ and CD8+
529 T cell responses in *Helicobacter pylori*-infected individuals. *Clin. Exp. Immunol.* 123,
530 81–7.

531 Read, A.F., Taylor, L.H., 2001. *The Ecology of Genetically Diverse Infections* 292, 1099–
532 1103.

533 Seward, J.F., Watson, B.M., Peterson, C.L., Mascola, L., Pelosi, J.W., Zhang, J.X., Maupin,
534 T.J., Goldman, G.S., Tabony, L.J., Brodovicz, K.G., 2002. *Varicella Disease After*
535 *Introduction of Varicella Vaccine in the United States , 1995-2000* 287, 1995–2000.

536 Siegel, R., Naishadham, E., Jemal, A., 2013. *Cancer Statistics, 2013.* *CA Cancer J Clin* 37,
537 408–14. <https://doi.org/10.3322/caac.21166>.

538 Silins, I., Ryd, W., Strand, A., Wadell, G., Törnberg, S., Hansson, B.G., Wang, X., Arnheim,
539 L., Dahl, V., Bremell, D., Persson, K., Dillner, J., Rylander, E., 2005. *Chlamydia*
540 *trachomatis* infection and persistence of human papillomavirus. *Int. J. Cancer* 116, 110–
541 115. <https://doi.org/10.1002/ijc.20970>

542 Zur Hausen, H., 2009. The search for infectious causes of human cancers: where and why.
543 *Virology* 392, 1–10. <https://doi.org/10.1016/j.virol.2009.06.001>

544 Zur Hausen, H., Villiers, E. De, 2015. Cancer “Causation” by Infections—Individual
545 Contributions and Synergistic Networks. *Semin. Oncol.* 41, 860–875.
546 <https://doi.org/10.1053/j.seminoncol.2014.10.003>

547

548

549

550

551

552

553

554

555 **Table 1|** Results for the three cancers of interest in females for temperate and tropical countries. TE prefix represents temperate dimensions whereas the T
556 prefix refers to tropical dimensions. Statistical significance and sign of association was obtained from analysis of variance. Range of significance describes
557 results from the rarefaction analysis. Number of infectious agents (IA) was determined from the MCA analysis. Dimensions, *P*-value, sign of association and
558 range of significance in italics mean that statistics are marginally significant.

Type of cancer	Temperate countries (TE)					Tropical countries (T)					559
	Dimensions	<i>P</i> -value	Sign of association	Range of significance	Number of IA	Dimensions	<i>P</i> -value	Sign of association	Range of significance	Number of IA	560
Bladder	Bact TE1	0.008	-	60%-100%	10	Bact T4	0.017	+	70%-100%	13	
	Bact TE4	0.02	+	60%-100%	8	<i>Helm T9</i>	<i>0.04</i>	+	<i>100%</i>	6	
	Helm TE8	0.019	-	80%-100%	7						
Total	3				25	2				19	
Liver	Bact TE1	0.009	+	80%-100%	10	<i>Bact T6</i>	<i>0.046</i>	+	<i>100%</i>	6	
	Helm TE1	0.02	+	80%-100%	22	Helm T5	0.013	+	90%-100%	11	
						Helm T6	0.002	-	60%-100%	9	
						Virus T6	0.01	-	80%-100%	14	
Total	2			32	4				40		
Stomach	Bact TE1	0.0001	+	60%-100%	10	Bact T2	0.002	-	70%-100%	10	
	Bact TE4	0.001	-	70%-100%	8	Bact T7	0.007	-	70%-100%	5	
	Helm TE1	0.002	+	70%-100%	22	Helm T2	0.0002	+	40%-100%	10	
	Helm TE2	0.0002	-	70%-100%	20	Helm T3	0.006	-	60%-100%	9	
	<i>Helm TE3</i>	<i>0.049</i>	+	<i>100%</i>	13	Virus T1	0.00003	+	60%-100%	25	
	Helm TE7	0.0006	+	50%-100%	8	Virus T5	0.023	+	70%-100%	13	
	Virus TE3	0.02	-	90%-100%	10						
	Virus TE8	0.038	-	100%	5						
Total	8			96	6				72		

561 **Table 2| Description of the specific associations which are supported by the random redistribution**
 562 **analysis (type I error).** The combination observed in the principal analysis are described and the
 563 probability of finding them by chance was calculated over 10.000 reassortments normally distributed
 564 of cancer incidence data with the same mean and variance than in the original database in each
 565 biomes.

Interactions	Description	Probability
Bladder/Viruses	No significant association between bladder cancer incidence and dimension of viruses in both regions.	36%
Bladder/Bacteria	Significant association between bladder cancer incidence and dimension of bacteria in both regions.	12%
Liver/Negative associated dimension	No significant dimension of any guilds is negatively associated with liver cancer incidence in temperate countries. In addition, at least one significant dimension of at least one guild is negatively associated with liver cancer.	40%
Liver/Helminths	Significant association between liver cancer incidence and dimension of helminths in both regions.	16%
Stomach/Viruses	Virus dimensions negatively correlated to stomach incidence in temperate countries but positively correlated in tropical countries.	3%

566

567 **Table 3: Links between prevalence of oncogenic agents and the number of associated dimensions based on the scientific literature.** Mean prevalence for
 568 each region has been calculated according to data presented in the references.

569

Oncogenic agents	Causality by cancer subtype	Global causality	Temperate countries		Tropical countries	
			Mean prevalence	Number of associated dimensions (Number of infectious agents)	Mean prevalence	Number of associated dimensions (Number of infectious agents)
<i>H. pylori</i>	80% of adenocarcinoma	70% of stomach cancer	41%(Peleteiro et al., 2014)	8 (96)	53%(Peleteiro et al., 2014)	6 (72)
HCV/HBV	80% of hepatocellular carcinoma	60% of liver cancer	4.3%(Mohd Hanafiah et al., 2013)	2 (32)	4.2%(Mohd Hanafiah et al., 2013)	4 (40)
<i>S. haematobium</i>	30% of squamous cells carcinoma	2% of bladder cancer	0.000006%(Chitsulo et al., 2000)	3 (25)	4.5%(Chitsulo et al., 2000)	2 (19)

570 **Figure 1| Distinction between temperate and tropical countries.** A) Infectious
571 communities in temperate and tropical biomes assessed by MCA. Tropical countries show a
572 different community of infectious agents from temperate countries even if we can observe
573 some overlaps. B) Values of the first dimension of the PCA on the confounding variables at
574 worldwide scale. Confounding variables of tropical countries largely group together in a
575 cohesive range of the parameter space, overlapping with those of temperate countries only at
576 high values. Vertical lines represent the geographical location of the tropics.

577

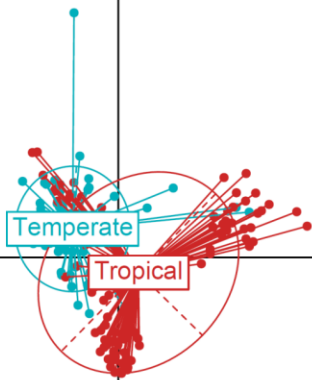
578 **Figure 2| Flow diagram summarizing the main statistical analyses undertaken in this**
579 **study.**

580

581 **Figure 3| Hypothesis on the contrasting role of viruses between the two biomes.**

Figure 1

A



B

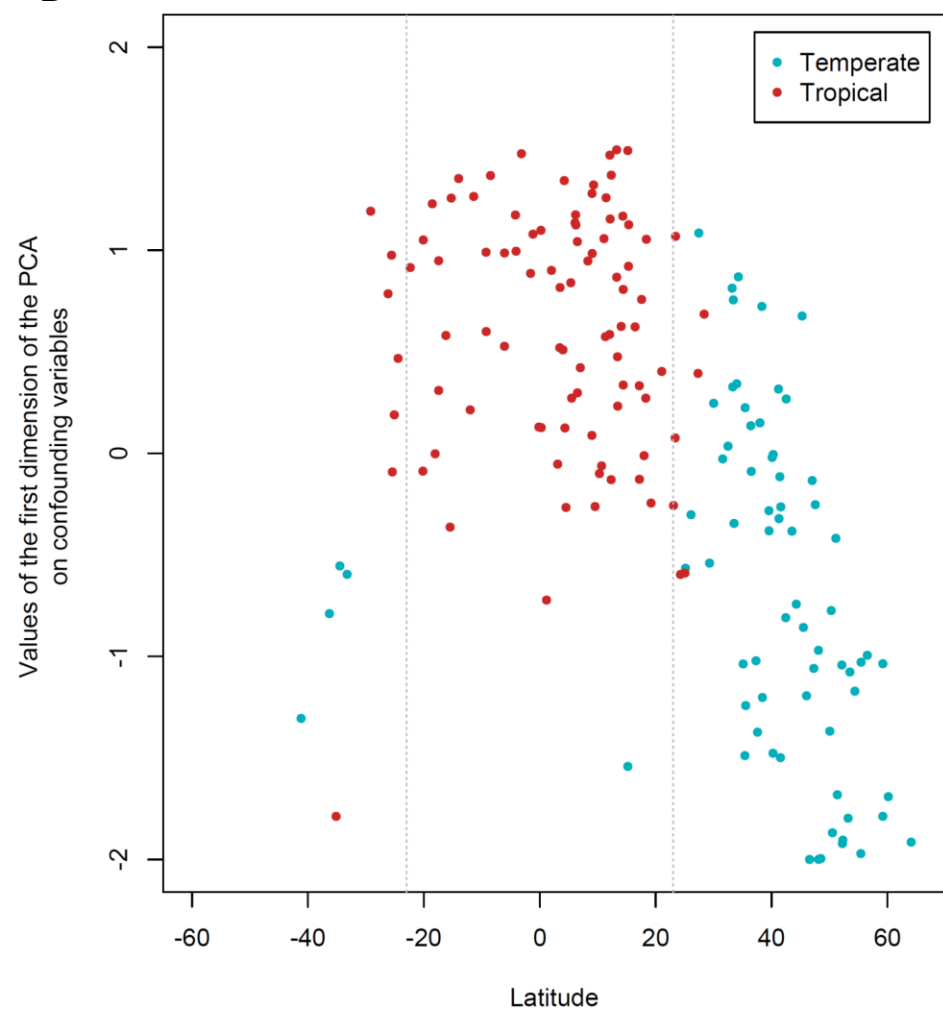


Figure 2

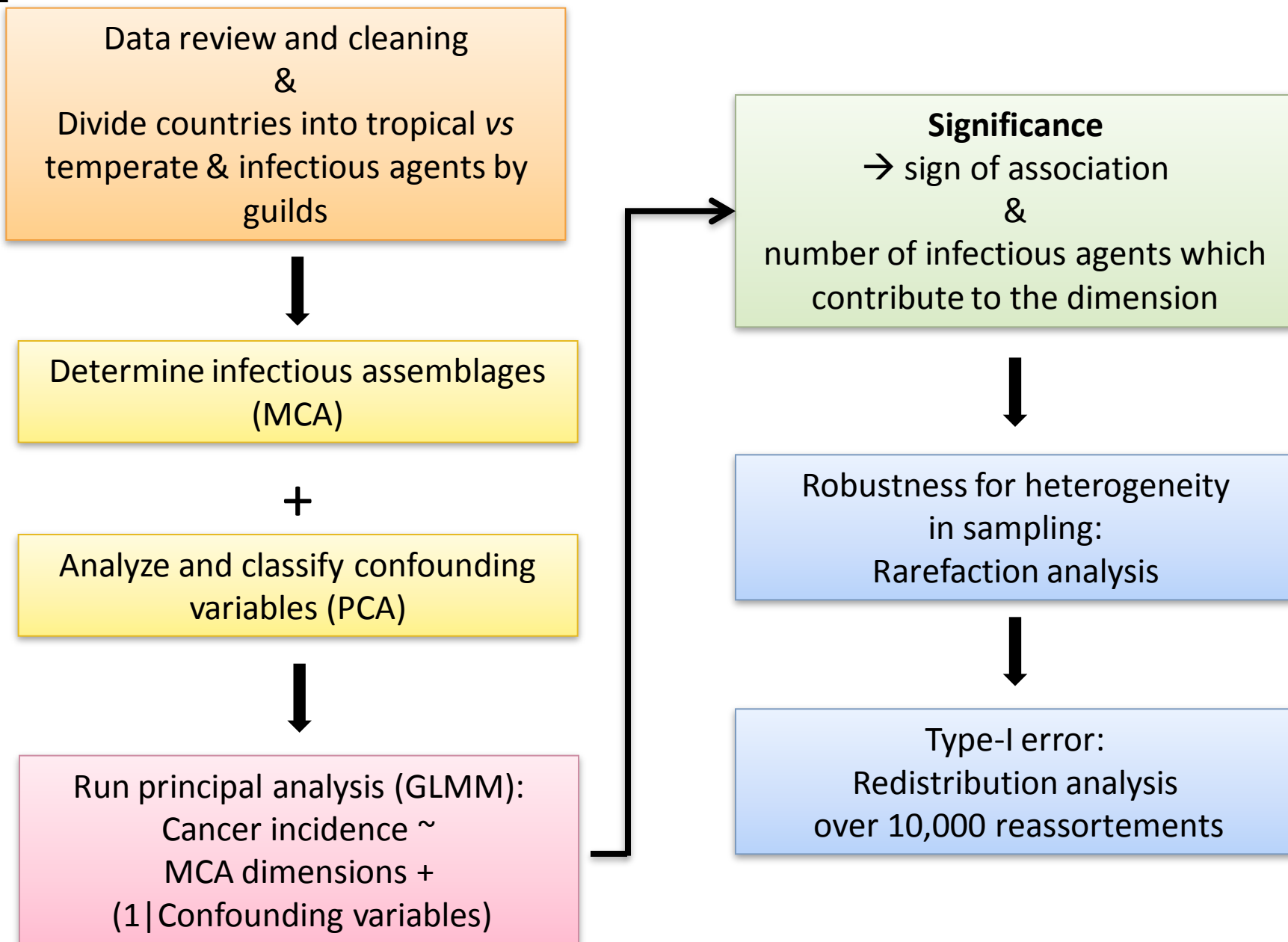
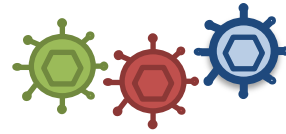
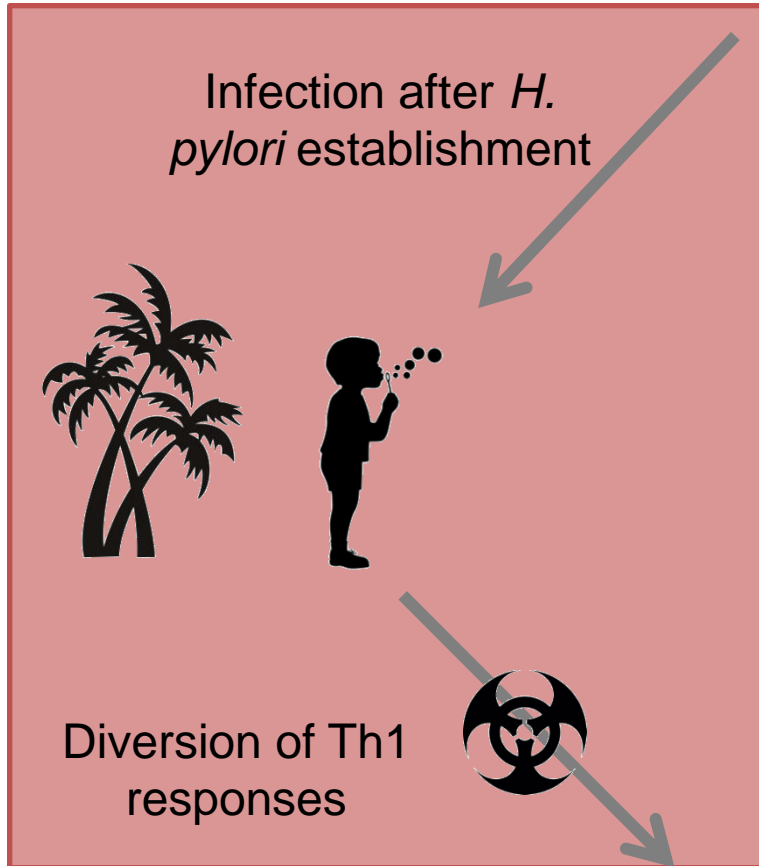


Figure 3

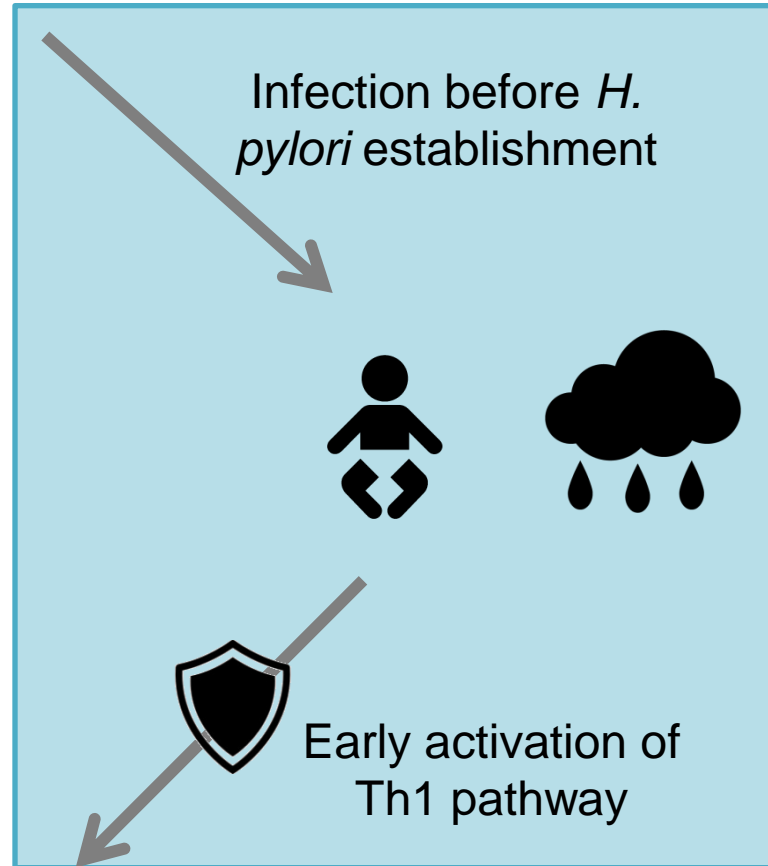
VIRUSES
COMMUNITY



TROPICAL



TEMPERATE



Proliferation

Elimination

Helicobacter pylori



Supplementary Material

[Click here to download Supplementary Material: Annexes VF3-revised.docx](#)

1 **The macroecology of cancer incidences in humans is associated with large-scale**
2 **assemblages of endemic infections**

3

4 Camille Jacqueline^{1,2§}, Jessica L. Abbate^{2,3}, Gabriele Sorci⁴, Jean-François Guégan², Frédéric
5 Thomas^{1,2}, Benjamin Roche^{1,3}

- 6 1. CREEC, 911 Avenue Agropolis, BP 64501, 34394 Montpellier Cedex 5, France
7 2. MIVEGEC, UMR IRD/CNRS/Université de Montpellier, 911 Avenue Agropolis, BP 64501, 34394
8 Montpellier Cedex 5, France
9 3. International Center for Mathematical and Computational Modeling of Complex Systems (UMI
10 IRD/UPMC UMMISCO), 32 Avenue Henri Varagnat, 93143 Bondy Cedex, France
11 4. BiogéoSciences, CNRS UMR 6282, Université de Bourgogne, 6 Boulevard Gabriel, 21000 Dijon,
12 France

13

14 § Corresponding author: 911 avenue Agropolis, 34000 Montpellier; +33687748572;
15 camille.jacqueline@ird.fr

16

17 **Short title** : Macroecology of cancer and endemic infections

18

19 **Abstract:**

20 It is now well supported that 20% of human cancers have an infectious causation (i.e.,
21 oncogenic agents). Accumulating evidence suggests that aside from this direct role, other
22 infectious agents may also indirectly affect cancer epidemiology through interactions with the
23 oncogenic agents within the wider infection community. Here, we address this hypothesis via
24 analysis of large-scale global data to identify associations between human cancer incidence
25 and assemblages of neglected infectious agents. We focus on a gradient of three widely-
26 distributed cancers with an infectious cause: bladder (~2% of recorded cancer cases are due to
27 *Shistosoma haematobium*), liver (~60% consecutive to Hepatitis B and C infection) and
28 stomach (*Helicobacter pylori* is associated with ~70% of cases). We analyzed countries in
29 tropical and temperate regions separately, and controlled for many confounding social and
30 economic variables. First, we found that particular assemblages of bacteria are associated with
31 bladder cancer incidence. Second, infectious assemblages found to be protective against liver
32 cancer in the tropics were not associated with a protective effect among temperate countries.
33 Third, we show that certain assemblages of viruses may facilitate stomach cancer in tropical
34 area, while others protect against its development in temperate countries. Finally, we discuss
35 the implications of our results in terms of cancer prevention and highlight the necessity to
36 consider neglected diseases, especially in tropics, to adapt public health strategies against
37 infectious diseases and cancer.

38
39 **Keywords:** neglected diseases, biomes, human cancer incidence, data mining, public health
40 strategies, pathogen-cancer interactions, infections, biomes, communities, cancer, statistics.

41

42 **Introduction**

43 While cancer remains one of the main causes of death in Western countries (Ferlay et
44 al. 2010), its burden is increasing in low- and middle-income countries (Magrath et al. 2013).
45 Although treatments, including chemo-, radio- and/or immunotherapy, have resulted in a
46 slight decline in death rates worldwide, cancer incidence rates have remained stable over the
47 past 10 years (Siegel et al. 2013). In this context, prevention seems currently the most
48 efficient strategy to reduce the impact of cancer on populations. The term “infectious agents”
49 describes the transmissible organisms that require living in or on another organism (the
50 “host”) to complete its lifecycle – a process which decreases host fitness – and includes virus,
51 bacteria, fungi, protozoans, helminths, etc... Since the 20th century, Numerous infectious
52 agents ~~(for the sake of clarity, the term “infectious agent” is used throughout the manuscript~~
53 ~~to describe transmissible organisms that require living in or on another organism (the “host”)~~
54 ~~to complete its lifecycle — a process which decreases host fitness — and includes virus,~~
55 ~~bacteria, fungi, protozoans, helminths, etc...)~~ have been recognized as a-risk factors for the
56 development of several cancers (i.e., acting through inflammation or introduction of foreign
57 DNA into cells; henceforth referred to “oncogenic agents”) (Zur Hausen and Villiers 2015).
58 For example, current evidence links Epstein–Barr virus (EBV), Hepatitis B and C viruses
59 (HBV, HCV), the bacteria *Helicobacter pylori*, human papillomavirus (HPV), and the
60 trematode *Schistosoma haematobium* to cancers of the lymph nodes, liver, stomach, cervix
61 and bladder respectively. Consequently, vaccination against oncogenic agents is currently
62 being considered as a potential key tool to prevent these cancers, such as the HPV vaccine
63 which offers a relative protection against cervical cancer (Paavonen et al. 2009).

Formatted: Superscript

64 Accumulating evidence suggests that aside from this direct role, infectious agents
65 could also be indirectly involved in cancer epidemiology through interactions with oncogenic
66 agents in infra-communities (reviewed in (Jacqueline et al. 2017)). Indeed, hosts are
67 commonly infected by multiple infectious species simultaneously (Read and Taylor 2001),
68 and within-host competition for resources and through immunity has been fairly well-
69 established (reviewed in (Mideo 2009)). A number of infectious organisms are known to
70 impair immune system homeostasis with both positive and negative consequences for their
71 competitors, especially through the well-known trade-off between the Th1/Th17 and Th2
72 immune pathways (Pedersen and Fenton 2007). For example, helminth infections are often

73 responsible for an up-regulation of the Th2 response, which has been linked to tuberculosis
74 reactivation and increase in likelihood of HIV infection (Borkow et al. 2001). Furthermore,
75 these interactions at individual scale are susceptible to have consequences on persistence and
76 transmission dynamics of the infectious agents at the population level (Ezenwa and Jolles
77 2014).

78 A few studies have explicitly considered oncogenic agents as part of an infectious
79 community. The majority of the literature available on interactions between oncogenic and
80 non-oncogenic agents concerns species that are widely distributed and with high prevalence.
81 For instance, *Plasmodium falciparum*, the agent of malaria, increases the replication of EBV,
82 an oncogenic virus associated with Burkitt Lymphoma, through its impairment of an effective
83 immune response (Chêne et al. 2007; Morrow et al. 1976). Epidemiological studies have also
84 shown that infection with *Chlamydia trachomatis* increases the risk for persistence of HPV
85 infection, some lineages of which can lead to cervical cancer (Silins et al. 2005). However,
86 these non-oncogenic agents, with high prevalence, may not be representative of the whole
87 diversity of infectious organisms, especially in tropical countries that are affected by a high
88 number of endemic and rare diseases that are relatively under-studied (Hotez et al. 2007).

89 Here, we explore the hypothesis that co-infections with endemic infections, in both
90 tropical and temperate latitudes, modify the persistence of oncogenic agents and thus interact
91 with the development of some infectious cancer in human population. For most countries,
92 prevalence of oncogenic agents is not available for the whole population and thus we
93 postulate that infectious agents favoring persistence of oncogenic agents should co-localize in
94 countries with higher cancer incidences. We focus on three widely-distributed cancers
95 (bladder, liver and stomach), for which an infectious causation is widely recognized and for
96 which sufficient large-scale data are available. These three examples allow considering
97 oncogenic agents that belong to different taxonomic guilds (helminthes, viruses and bacteria)
98 as well as a gradient in the frequency with which each cancer is known to have an infectious
99 origin (from 2% of recorded cases for bladder cancer, to 60% for liver cancer and 70% for
100 stomach cancer). By compiling and analyzing a global database, we identify associations
101 between the country-level incidence of cancer and the presence or absence of infectious
102 agents in each country. We then present hypotheses regarding the mechanism behind each

103 association, and discuss the potential consequences of our findings in terms of cancer
104 prevention strategies.

105

106 **Material & methods**

107 *Data overview and inclusion criteria*

108 Our cancer data were obtained from the International Agency for Research on Cancer
109 (IARC GLOBOCAN project, 2012, <http://globocan.iarc.fr/>). Among 48 human cancers in
110 women and men across 184 countries, we selected three cancers of interest (bladder, liver and
111 stomach) because they are recognized to have an infectious causation. First, the bacterium *H.*
112 *pylori* is responsible for 80% of stomach adenocarcinomas (Zur Hausen 2009), which
113 represent 90% of stomach cancers (Brenner et al. 2009). Second, 75% of liver cancers are
114 hepatocellular carcinomas (HCCs) (Parkin 2001) which have been linked in more than 80% of
115 cases to HCV and HBV infection (Parkin 2006). Finally, squamous cell carcinomas (SCCs),
116 representing 5% of bladder cancers (Kantor et al. 1988), have been associated with the
117 helminth *Schistosoma haematobium* in approximately 30% of cases (Mostafa and Sheweita
118 1999). In our study, we used age-standardized incidence (ASI), which represents the raw
119 cancer incidence when country age structure is extrapolated to a standard one (World
120 Standard Population (Doll et al. 1996)).

121 Our analyses on infectious species relied on the GIDEON database (Global Infectious
122 Diseases and Epidemiology Network, <http://web.gideononline.com/>). This database contained
123 a presence/absence matrix for a total of 370 human infectious agents across 224 countries
124 (2004 update). We removed all infectious agents that were present or absent in all countries,
125 because they did not add any discriminating information, as well as infections caused by
126 fungi, protozoans and arthropods guilds which were represented by less than 10 infectious
127 agents respectively. This yielded a data subset with infectious agents belonging to three main
128 guilds (viruses, bacteria and helminths). It included very rare diseases such as Buruli ulcer
129 (3,000 cases/y worldwide; www.who.int/gho/neglected_diseases/buruli_ulcer) and tick-borne
130 encephalitis (11,000 cases/y worldwide; www.who.int/immunization/topics/tick_encephalitis)
131 to more widespread diseases affecting a higher number of persons such as leishmaniasis and
132 echinococcosis (1 million cases/y worldwide; <http://www.who.int/echinococcosis/>).

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

133 Finally, we constituted a database of 24 potential confounding variables for 167
134 countries from the Food and Agriculture Organization (FAO), World Health Organization
135 (WHO) and the World Bank (WB) to consider also the diversity of social factors and
136 economic levels (Supplementary Table S1 for a description).

137 The dataset was analyzed separately between tropical (geographically situated between
138 the two tropics) and temperate (above or below tropics) countries in order to better control for
139 the huge environmental (in addition to social and economical) disparities between these two
140 biomes (distinct biological communities formed in response to a shared physical climate
141 (Cain et al. 2011)). In fact, tropical countries have a higher abundance/richness of infectious
142 species (Guernier et al. 2004) and are associated on average with lower socio-economic
143 wealth (World Bank 2016) compared to temperate countries. We confirmed that these two
144 groups of countries, made on geographical assumptions, correspond to two distinct biomes
145 which present distinct-regarding patterns of ~~both~~ infectious species richness (Supplementary
146 Fig. S1) and richness-composition (Fig.1A) as well as socio-economical disparities (Fig.
147 1B Supplementary materials Figure S1, S2, S3).

148

149 *Assessment of infectious agent assemblages*

150 Our database of infectious agents, after removing non-relevant species as previously
151 described, contained presence/absence values for 101 human infectious agents across 103
152 tropical countries and 88 infectious agents across 74 temperate countries. We first aimed to
153 characterize infectious assemblages within each guild (i.e. viruses, bacteria and helminths).
154 To do so, we ran Multiple Correspondence Analysis (MCA; FactomineR Lê *et al.* 2008) in
155 order to reduce the number of variables to a set of dimensions representing the
156 presence/absence data of all infectious agents within each guild. The composition of these
157 dimensions was then examined as a synthetic measure of assemblage structure. The number
158 of dimensions considered was calculated according to the Kaiser's criterion (dimensions for
159 which the percentage of variance explained was superior to a threshold calculated as follows:
160 $100\% / \text{total number of dimensions estimated by MCA analysis (Kaiser 1958)}$). For each of
161 these selected dimensions, we calculated standard coordinates for each country by dividing
162 the principal coordinates (i.e. loadings) by the square root of the dimension's eigenvalue

163 (Husson et al. 2017). This standardization allowed considering each dimension as independent
164 variables.

165 We followed a similar methodology for the 24 confounding variables described
166 previously (substituting the Multiple Correspondence Analysis with a Principal Component
167 Analysis, as our confounding variables contained quantitative values rather than categorical
168 presence/absence). Given that the first dimension of the PCA on confounding variables
169 explained more than 40% of the variance within each biome (Supplementary materials [Figure](#)
170 [Fig. S24](#)), we classified the values of this dimension into five evenly distributed classes of
171 countries. These classes were then used as a random factor in statistical models to consider
172 intra-biome variability in social and economic level (see below).

173 We conducted separate analyses for each guild (viruses, bacteria, helminths) within
174 each biome (tropical and temperate) and we transformed cancer incidence data (ASI) to obtain
175 Gaussian distributions (transformations are detailed in supplementary materials Table S2).
176 Within a generalized linear mixed model (GLMM) implemented in the package *lme4* (Bolker
177 et al. 2009), transformed ASI was used as the response variable, the standard coordinates of
178 the MCA dimensions, representing infectious assemblages, were used as fixed factors and the
179 classes of confounding variables as a random factor. We did not consider interactions between
180 fixed factors because the excessively high number of potential interactions and the difficulty
181 to interpret them. Variable selection was conducted through analysis of variance (ANOVA)
182 (using the ‘anova’ function in package *car*, with test specified as “type III” that quantifies the
183 effect of each variable after all other factors have been accounted for (Fox and Weisberg
184 2011)). For each significant dimension, we assessed the sign of the association with cancer
185 incidences through coefficient values.~~In a generalized linear mixed model (GLMM)~~
186 ~~implemented in the package *lme4* (Bolker et al. 2009), transformed ASI were used as the~~
187 ~~response variable, the MCA dimensions representing infectious assemblages were used as~~
188 ~~fixed factors and the classes of confounding variables as random factors. We did not consider~~
189 ~~interactions between fixed factors because of combinatory explosion. Variable selection was~~
190 ~~conducted through analysis of variance (ANOVA) (using the ‘anova’ function in package~~
191 ~~*car*, with test specified as “type III” that tests the effect of each variable after all other factors~~
192 ~~have been accounted for (Fox and Weisberg 2011)).~~ We will henceforth refer to this
193 statistical analysis as the “principal analysis”.

194 | In order to test the accuracy of our GLMMs to fit population data, we calculated
195 | several error and accuracy statistics for all the models generated during the principal analyses
196 | (using the ‘accuracy’ function in the package rcompanion; see Table S3). After having
197 | considered each guild independently (one GLMM by guild/by cancer), we ran a GLMM for
198 | each cancer (as described previously) considering all guilds together. In this relative
199 | contribution analyses, we used only the significant dimensions identified in the principal
200 | analysis as fixed factors (see Table S3S4).

201 | Finally, for each significant MCA dimensions involved in the principal analysis, we
202 | identified the infectious agents that explained more than 5% of the variability of the significant
203 | dimensions (represented by “eta2” values in the MCA). For each cancer, we grouped all of
204 | these infectious agents (separately for those with positive versus negative correlation with
205 | cancer incidences), and we refered to each of these groups as an “infectious assemblage”. As
206 | the agents in each assemblage may not necessarily interact with one-another, we distinguished
207 | the term “assemblage” from “community”. We described the size and the composition of the
208 | assemblages for each cancer in Tables S4S5, S65 and S76 of the Supplementary Materials.

209

210 | *Robustness assessment*

211 | The robustness of our results to heterogeneities in sampling was tested using ~~a~~, ad-
212 | hoc rarefaction analysis. We generated random samples of countries containing from 20% to
213 | 100% (in increments of 10% for each new random sample) of the entire database ~~for each of~~
214 | ~~the three guilds~~. For each percentage, we repeated the random sampling 10 times and ran the
215 | primary analysis (for each biomes, cancer and guilds) on each of these repetitions. We
216 | reported the percentage of the database for which the median p-value (calculated on the 10
217 | repetitions) becomes significant. By doing this, we can “score” the robustness of each result
218 | and exclude dimensions that were significant due only to outlier countries. It is worth
219 | pointing out that this score aims quantifying how our different conclusions are relatively
220 | robust each against others. Therefore, there is no significance threshold for this score.

221 | Finally, as caveat of large-data analysis can be to find associations without biological
222 | signification meaning, we calculated the risk to observe significant combination of infectious
223 | assemblages in association with cancer incidence by random chance. This is the equivalent of
224 | type-I statistical error and we aimed to assess the percentage of chance to incorrectly reject

225 the null hypothesis. We conducted a redistribution analysis by simulating normally-distributed
226 incidence of cancer. For each cancer of interest, we generated 10,000 random reassortments
227 of cancer incidences across countries with the same mean and variance as in the original
228 database in each biome. GLMMs were conducted as described previously for each guild of
229 infectious agents. We reported the percentage of reassortments for which we observe the same
230 combination of significant dimensions, across the three guilds, detected by the principal
231 analysis. A low percentage suggests that the risk of detecting a combination with no
232 biological meaning is low. As for the rarefaction analysis, this percentage quantifies relatively
233 how our conclusion can be found just by chance, which does not call for a threshold.

234 The statistical approach is summarized in [Figure Fig. 1-2](#) and all the analyses have
235 been conducted using R v3.1.2 statistical software (R Development Core Team).

236

237 **Results**

238 After applying MCA on each guild and removing the non-explicative dimensions
239 according to the Kaiser criterion, we obtained 7 dimensions explaining variance for bacterial
240 species occurrence, 9 dimensions for helminths and 8 for viruses in temperate countries. For
241 tropical countries, 8, 10 and 11 dimensions were kept for bacteria, helminths and viruses,
242 respectively. [We found that GLMMs fit the actual data relatively well with an accuracy
243 between 70% and 90% \(see Table S3\).](#) -Because our results were similar for both sexes ([see
244 Supplementary data-Table S7S8](#)), we report here only the results for females (Table 1).

245 *Bladder cancer*

246 In temperate biome, bladder cancer was negatively associated with one dimension of
247 bacteria (BACT TE1) and one dimension of helminths (HELM TE8), but positively
248 associated with a second dimension of bacteria (BACT TE4) (Table 1). Two dimensions (one
249 for bacteria (BACT T4) and one for helminths (HELM T9)) were positively associated with
250 bladder cancer incidence in tropical countries. The associations found with bacteria were well
251 supported in both biomes in the relative contribution analysis ([Supplementary materialssee
252 Table S3S4](#)), as well as with the rarefaction analysis (Table 1). ~~and s~~ Such association were
253 found in 12% of simulated normally-distributed incidences (Table 2). In addition, the
254 redistribution analysis showed that the absence of virus dimensions in both temperate and

255 tropical countries could be expected in 36% of trials by chance. Regarding assemblage
256 composition two biomes confounded, we noticed that the negatively associated assemblage is
257 composed of 10 bacteria and seven helminths species. The positively associated assemblage
258 contains 17 bacteria and 6 helminths species ([See see Table S4-S5 supplementary materials](#)).

259

260 *Liver cancer*

261 One dimension for bacteria (Bact TE1 and Bact T6) and one for helminths (Helm TE1
262 and Helm T5) were positively associated with liver cancer incidence in both temperate and
263 tropical countries. However, the presence of negatively associated dimensions (Helm T6 and
264 Virus T6) was observed only in tropical countries. Helminths represented the only guild
265 which was both positively and negatively correlated to this cancer incidence. The distinct
266 pattern of associations between the two biomes for negatively associated dimensions was
267 found in 40% of simulations in the redistribution analysis (Table 2). In addition, rarefaction
268 analysis and relative contribution of dimensions only highlighted the negative association
269 between helminths and liver cancer in tropical countries. In accordance with this result, the
270 association between any helminth dimensions and liver cancer was detected in 16% of
271 random trials. Assemblages differed greatly depending on the sign of the association with
272 liver cancer incidence; while positive associated assemblage contained 12 bacteria and 25
273 helminths, the negative associated assemblage was formed by 9 species of helminths and 14
274 viruses (See Table [S5-S6](#) in supplementary materials).

275

276 *Stomach cancer*

277 Compared to bladder and liver cancers, stomach cancer showed the highest number of
278 associated dimensions with two for bacteria (Bact TE1 and TE4), four for helminths (Helm
279 TE1, TE2, TE3, TE7) and two for viruses (Virus TE3 and TE8) in temperate countries and 6
280 dimensions (Bact T2 and T4, Helm T2 and Helm T3, Virus T1 and T5) in tropical countries
281 (Table 1). We found more dimensions and a higher number of species representing helminth
282 guild in temperate countries. In addition, we noticed that associations between viruses and
283 stomach cancer were different between the two biomes. While some dimensions of viruses
284 were positively associated with incidence in tropical countries, others were negatively

285 correlated in temperate countries. We showed that this last combination should occur by
286 chance in just 3% of cases. While most of the dimensions were well supported by rarefaction
287 analysis, the relative contribution of the dimension showed that only dimensions of helminths
288 and bacteria in temperate countries were retained. Finally, the assemblage negatively
289 associated with stomach cancer incidence (two biomes confounded) showed a lower diversity
290 overall with 69 species than positively associated assemblage which were composed of 101
291 human infections (See Table [S6-S7](#) in supplementary materials).

292

293 **Discussion**

294 Through a large-scale statistical analysis of global presence/absence of infectious agents and
295 cancer incidence data, we found that three well-distributed human cancers with well-accepted
296 infectious causation were associated with different species assemblages, as described in Table
297 1. Overall, no common patterns were identified across the three cancers but rather specific
298 associations between each biome's composition of infectious agents and cancer incidences.
299 However, cancers that we found to be associated with higher number of dimensions as well as
300 higher number of infectious agents are also those known to have high level of infectious
301 causality by highly prevalent oncogenic agents (Table 3). This is particularly striking for
302 cancers caused by oncogenic agents at the extremes of the causality gradient such as *H. pylori*
303 and *S. haematobium*.

304 We postulate that the assemblages negatively associated with cancer incidences
305 formed a potential “protective component community” which includes all of the
306 infracommunities within a host population and begs further investigation of an underlying
307 mechanism. Alternatively, positively associated assemblages will be referred as “facilitating
308 component community” in the following sections. Our results emphasize the possibility that
309 certain infectious agents influence the persistence or the circulation of oncogenic agents, and
310 the urgent need to investigate the details of infectious species interactions in the context of
311 cancer prevention.

312 We noticed two specific patterns in component community associated with bladder
313 cancer, of which 2% of cases are thought to be due to infection with the trematode
314 *Schistosoma haematobium*. It is worth mentioning that this trematode has an heterogeneous

315 transmission pattern, with an extremely high prevalence in tropical countries, but circulating
316 only in several countries within the temperate biome. First, we found that bacteria may have a
317 preponderant role in both biomes. ~~Among component communities associated to bladder~~
318 ~~cancer, only four bacterial species have a congruent role between the two biomes (*Rickettsia*~~
319 ~~*felis*, *Rickettsia japonica*, *Rickettsia sibirica*, *Fusobacterium necrophorum*).~~ Interestingly,
320 studies have shown that bacteria-helminth interactions have already been observed on the
321 field and could impact cancer risk. Notably, a high percentage of co-infection with *S.*
322 *haematobium* and some unspecified bacteria in the urinary tract has been reported (Adeyeba
323 and Ojeaga 2002; Ossai et al. 2014). These co-infections could increase the risk of bladder
324 cancer as bacteria in the urinary tract produce nitrosamines, which are carcinogenic
325 compounds (Davis et al. 1984). In addition, bacteria may bias the Th1/Th2 balance away from
326 protection against helminths by down-regulating Th2-mediated responses. This kind of
327 indirect interaction has already been seen in rabbits where the bacterium *Bordetella*
328 *bronchiseptica* was shown to enhance helminth intensity through immune mediated effects
329 (Pathak et al. 2012). Thus, our study may highlight the role of bacteria in cancer associated
330 with helminth infections. Second, viral species were not observed among component
331 communities linked with bladder cancer, which could be seen at odds with what has been
332 suggested in the literature ~~(e.g., In fact, helminths have been shown to favor viral infection~~
333 ~~due to their immunoregulatory role (Kamal and El Sayed Khalifa 2006), and it is likely that~~
334 ~~infections with certain viruses known to immunosuppress their host (reviewed by Oldstone~~
335 ~~2006)) could have consequences for subsequent helminth infections. As suggested by the~~
336 ~~redistribution analysis, this result needs to be taken with precaution. Indeed, t~~The absence of
337 viral implication, in our study, could be due to the scarcity of *S. haematobium* infections in
338 wealthy temperate countries as well as on the low direct contribution of the parasite to bladder
339 cancer, ~~thus our results need to be taken with precaution.~~

340 Regarding liver cancer, we observed a specific and robust association between
341 helminths and cancer incidences in both biomes. The interactions between certain helminths
342 species and HCV/HBV have been reported in experimental studies and may be either
343 protective or facilitator. It has been suggested that co-infection with *S. mansoni* could increase
344 the persistence and severity of HCV infection because the helminth prevent the production of
345 an HCV-specific CD4+/Th1 T cell response (Kamal et al. 2001). Conversely, a recent study
346 has demonstrated the immunostimulant effect of a protein derived from *Onchocerca volvulus*

347 which increases the interferon- γ response to HCV in vitro (MacDonald et al. 2008). Even if
348 the identified component communities do not highlight these two particular species, it seems
349 worth to expect that similar mechanisms could apply to other helminth species (McSorley and
350 Maizels 2012). Furthermore, our results suggest the absence of protective component
351 communities (all guilds confounded) in temperate countries. However, the redistribution
352 analysis showed that this last result may frequently occur by chance (in about 40% of
353 redistribution simulations). Therefore, the interpretation of this finding in the context of liver
354 cancer epidemiology is highly speculative.~~Furthermore, our results reveal the absence of a~~
355 ~~protective component community (all guilds confounded) in temperate countries.~~
356 ~~Redistribution analysis showed that this last result may frequently occur by chance (around~~
357 ~~40%); nevertheless, the interpretation of this result could give insights into our understanding~~
358 ~~of liver cancer epidemiology. Indeed, temperate countries benefit from high vaccination~~
359 ~~coverage (e.g., 76% for Europe according to the WHO) which may decrease the number of~~
360 ~~cancer cases attributable to HBV and thus the strength of the interaction between infectious~~
361 ~~agents and cancer incidences.~~

362 As the oncogenic agent *H. pylori* is highly prevalent worldwide (Atherton and Blaser
363 2009), ~~S~~stomach cancer had a particular likelihood to be associated with ~~large-species-rich~~
364 component communities in both biomes,~~as the oncogenic agent *H. pylori* is highly prevalent~~
365 ~~worldwide (Atherton and Blaser 2009).~~ This was the case, with the largest number and
366 diversity of species across guilds associating both positively and negatively with this cancer.
367 The principal result for stomach cancer is that component communities of viruses have an
368 opposite effect between the two biomes. We hypothesize that it could rely on different
369 sequences of infections between these two regions. In temperate countries, the highest
370 incidence of viral diseases occurs among children younger than 10 years (Seward et al. 2002).
371 In temperate biomes, early activation of Th1 responses by viruses, even before *H. pylori*
372 infection, could thus bring protection against *H. pylori* through cross immunity (Quiding-
373 Järbrink et al. 2001). Conversely, viruses are acquired at older ages in tropical countries with
374 a higher proportion of cases and higher susceptibility among adults (Lee 1998). In this
375 situation, viral infections may temporarily divert the Th1 responses from *H. pylori* and allow
376 it to proliferate (Figure 3). ~~Viruses have numerous mechanisms to escape immune system and~~
377 ~~especially they are able to down regulate interferon signaling (Gale and Sen 2009). However,~~
378 ~~IFN γ responses produced by CD8⁺ cells contribute significantly to the elimination of *H.*~~

379 | ~~pylori infection (Quiding-Järbrink et al. 2001). Thus, viruses could allow the persistence of H.~~
380 | ~~pylori and consequent pro-tumoral inflammation by protecting it from immune destruction.~~
381 | ~~According to this hypothesis, five viruses have a consistent facilitating role between biomes~~
382 | ~~(Cowpox virus, Western and Eastern equine encephalitis virus, St Louis encephalitis virus,~~
383 | ~~Hantavirus).~~

384 | As for any statistical approach dealing with large scale database, our study is
385 | susceptible to a number of issues which need to be discussed. First, some diseases based on
386 | observation of symptoms have multiple potential causative agents. When this was the case,
387 | we included only the agent responsible for the majority of cases. Second, a principal
388 | weakness of our study is that infectious assemblages were based on presence/absence data
389 | where even imported and isolated cases (observed rarely) can result in a presence assignment
390 | for the whole country. In addition, all worldwide-distributed infectious agents have been
391 | removed from the database because they would not be discriminating at all. These two last
392 | points suggested that it would be helpful to consider prevalence data in order to assess the
393 | proportion of the population which is really in contact with the infectious agent. Such
394 | approach should be facilitated by future development of openly available data sources by
395 | ~~open data area in the future~~ and may allow assessing the global component community,
396 | including worldwide-distributed infectious agents, which may actually have a clearer impact
397 | on cancer incidence. This consideration, however, do not necessarily negate the indirect
398 | impact of the communities we have identified. Finally, we defined guilds according to their
399 | taxonomic classifications, which could be not biologically meaningful, but which allows to
400 | deal with the high number of infectious agents considered here. In fact, it has been
401 | recommended that guilds should be based on functional similarity of species or based on their
402 | life-cycle categories instead of their taxonomy (McGill et al. 2006).

403 | Considering cancer incidence data, the Globocan database (compiled by the IARC) do
404 | not account for the different cancer subtypes. As oncogenic agents are implied in subtypes
405 | that are only a proportion of the total number of cases, our conclusions may be more robust
406 | when infection is associated with a prevalent cancer (stomach adenocarcinoma and
407 | hepatocellular carcinoma) as opposed to rarer subtypes (e.g., SCCs of the bladder).
408 | ~~Furthermore, the detection of cancer cases could be better in temperate countries because of~~
409 | ~~the higher investment in health and disease surveillance.~~ Furthermore, the detection of cancer

Formatted: Font: Not Bold

410 ~~cases could be better in temperate countries because of the higher investment in health and~~
411 ~~disease surveillance. However, we assessed this disparity using two summarizing indices~~
412 ~~provided by the IARC, which considers the methods and data quality. These indices have~~
413 ~~been considered as confounding variables (see supplementary materials). Finally, reporting~~
414 ~~bias could also affect our dataset on confounding variables. However, we have selected data~~
415 ~~from reference organizations such as the WHO and the World Bank which are susceptible to~~
416 ~~be the more accurate. Even if it could quantitatively impact our results, it is unlikely to~~
417 ~~influence them qualitatively as the information has been compiled in a unique variable used as~~
418 ~~random factor. However, we assess this disparity using indices provided by the IARC, which~~
419 ~~described the methods and data quality. These indices have been considered among the~~
420 ~~confounding variables used as random factors in our analyses (see supplementary materials).~~

421 We have based our interpretations here on the hypothesis that infectious agents may
422 modify the circulation or the persistence of oncogenic agents. ~~However, our population-scale~~
423 ~~data do not reveal whether human individuals with cancer are actually infected or have been~~
424 ~~in contact with the identified diseases. In addition, our data do not allow assessing the~~
425 ~~variation in co-infection by infectious species at the individual level. While our this study~~
426 ~~does not claim to provide a definitive answer, it calls for designing cohort studies of cancer~~
427 ~~patients considering personal infection history to determine the causality, and potentially the~~
428 ~~mechanism, of such indirect link between non-oncogenic infectious agents and cancer~~
429 ~~development. However, our population-scale data do not reveal whether human individuals~~
430 ~~with cancer are actually infected or have been in contact with the identified diseases. One~~
431 ~~alternative, though not mutually exclusive, hypothesis is that infectious agents belonging to~~
432 ~~the community could play a role at the cellular level by disturbing immunosurveillance,~~
433 ~~introducing viral genes, promoting mutation or providing an environment where cancer~~
434 ~~growth can go unchecked (Dalton-Griffin and Kellam 2009). Some of these oncogenic~~
435 ~~mechanisms require that infectious agents may have a specific tropism for cancer cells, i.e.,~~
436 ~~stomach, bladder or liver cells, but to the best of our knowledge species identified by our~~
437 ~~analysis do not necessarily show such tropism and cohort studies of cancer patients at the~~
438 ~~individual scale would be necessary to detect this mechanism.~~ Finally, the inclusion of a
439 cancer for which no infectious agent is suspected to play a role would have served as an
440 evidence-boosting negative control. However, the assemblage identified in these controls

441 could also have biological relevance as they could include species that indirectly alter
442 immunosurveillance as suggested in (Jacqueline et al. 2017).

443

444 **Conclusion**

445 Despite the further steps necessary to validate the biological relevance of our findings,
446 these results draw attention to the need and potential benefits of a “pathocenosis approach”
447 for the management of infection-derived cancers, in a global health perspective. Indeed,
448 acknowledging the indirect role of infectious agents may allow for the adaptation of public
449 health strategies (such as vaccination) to improve prevention against cancer as well as the
450 targeted infectious diseases. As cancers with an infectious origin are predominant in tropical
451 countries, our study raises the perspective to used current strategies of infectious diseases
452 control (vaccination, antibiotics...) to decrease cancer burden in this region. Our approach,
453 which serves to determine associations at the largest scale (Khoury and Ioannidis 2014), is a
454 necessary first step to motivate further experimental and cohort-based studies to confirm our
455 findings, to identify the mechanisms implicated and thus to reveal new therapeutic
456 opportunities.

457

458 **Acknowledgments**

459 The authors thank CREEC sponsors CNRS and André HOFFAMNN (Fondation
460 Mava). This paper is a contribution of the EVOCAN and STORY projects funded by the
461 Agence Nationale de la Recherche. Post-doctoral support for JLA was provided by ANR JC
462 ‘STORY’ granted to Benjamin Roche. JFG and BR were funded by an “Investissement
463 d’Avenir” Laboratoire d’Excellence Centre d’Etude de la Biodiversité Amazonienne Grant
464 (ANR-10-LABX-25-01).

465

466 **Competing interests:** No conflict of interests to declare.

467

468 **References**

- 469 Adeyeba OA, Ojeaga SGT. 2002. URINARY SCHISTOSOMIASIS AND CONCOMITANT
470 URINARY TRACT PATHOGENS AMONG SCHOOL CHILDREN IN
471 METROPOLITAN IBADAN ,. Afr J Biomed Res 5: 103–108.
- 472 Atherton JC, Blaser MJ. 2009. Review series Coadaptation of *Helicobacter pylori* and
473 humans: ancient history , modern implications. J. Clin. Invest. 119;
474 doi:10.1172/JCI38605DS1.
- 475 Borkow G, Weisman Z, Leng Q, Stein M, Kalinkovich A, Wolday D, et al. 2001. Helminths ,
476 Human Immunode ciency Virus and Tuberculosis. 568–571.
- 477 Brenner H, Rothenbacher D, Arndt V. 2009. Epidemiology of Stomach Cancer. 467–477.
- 478 Cain ML (Michael L, Bowman WD, Hacker SD. 2011. *Ecology*. Sinauer Associates.
- 479 Chêne A, Donati D, Guerreiro-Cacais AO, Levitsky V, Chen Q, Falk KI, et al. 2007. A
480 molecular link between malaria and Epstein-Barr virus reactivation. PLoS Pathog. 3:e80;
481 doi:10.1371/journal.ppat.0030080.
- 482 Chitsulo L, Engels D, Montresor A, Savioli L. 2000. The global status of schistosomiasis and
483 its control. Acta Trop. 77:41–51; doi:10.1016/S0001-706X(00)00122-4.
- 484 Davis CP, Cohen MS, Gruber MB, Anderson MD, Warren MM. 1984. Urothelial hyperplasia
485 and neoplasia: a response to chronic urinary tract infection in rats. J Urol 132: 1025–31.
- 486 Doll R, Payne P, Waterhouse JA. 1996. *Cancer Incidence in Five Continents, Vol. I Union*
487 *Internationale Contre le Cancer*. Geneva.
- 488 Ezenwa VO, Jolles AE. 2014. Opposite effects of anthelmintic treatment on microbial
489 infection at individual versus population scales. 8–11.
- 490 Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. 2010. Estimates of worldwide
491 burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer 127:2893–2917;
492 doi:10.1002/ijc.25516.
- 493 Fox J, Weisberg S. 2011. *An {R} Companion to Applied Regression, Second Edition*.
494 Thousand O.

Formatted: English (United States)

- 495 Guernier V, Hochberg ME, Guégan J-F. 2004. Ecology drives the worldwide distribution of
496 human diseases. *PLoS Biol.* 2:e141; doi:10.1371/journal.pbio.0020141.
- 497 Hotez PJ, Molyneux DH, Fenwick A, Kumaresan J, Sachs SE, Sachs JD, et al. 2007. Control
498 of Neglected Tropical Diseases. 1018–1027.
- 499 Husson F, Josse J, Le S, Maintainer JM. 2017. Package “FactoMineR” Title Multivariate
500 Exploratory Data Analysis and Data Mining.
- 501 Jacqueline C, Tasiemski A, Sorci G, Ujvari B, Maachi F, Missé D, et al. 2017. Infections and
502 cancer: the “fifty shades of immunity” hypothesis. *BMC Cancer in press*:1–11;
503 doi:10.1186/s12885-017-3234-4.
- 504 Kaiser HF. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*
505 23:187–200; doi:10.1007/BF02289233.
- 506 Kamal SM, Bianchi L, Al Tawil A, Koziel M, El Sayed Khalifa K, Peter T, et al. 2001.
507 Specific Cellular Immune Response and Cytokine Patterns in Patients Coinfected with
508 Hepatitis C Virus and *Schistosoma mansoni*. *J. Infect. Dis.* 184:972–982;
509 doi:10.1086/323352.
- 510 Kamal SM, El Sayed Khalifa K. 2006. Immune modulation by helminthic infections: Worms
511 and viral infections. *Parasite Immunol.* 28:483–496; doi:10.1111/j.1365-
512 3024.2006.00909.x.
- 513 Kantor AF, Hartge P, Hoover RN, Fraumeni JF. 1988. Epidemiological characteristics of
514 squamous cell carcinoma and adenocarcinoma of the bladder. *Cancer Res.* 48: 3853–5.
- 515 Khoury MJ, Ioannidis JP a. 2014. Big data meets public health. *Science* (80-.). 346:1054–
516 1055; doi:10.1126/science.aaa2709.
- 517 Lee BW. 1998. Review of varicella zoster seroepidemiology in India and South-east Asia.
518 *Trop. Med. Int. Heal.* 3:886–890; doi:10.1046/j.1365-3156.1998.00316.x.
- 519 MacDonald AJ, Libri NA, Lustigman S, Barker SJ, Whelan MA, Semper AE, et al. 2008. A
520 novel, helminth-derived immunostimulant enhances human recall responses to hepatitis
521 C virus and tetanus toxoid and is dependent on CD56+ cells for its action. *Clin. Exp.*
522 *Immunol.* 152:265–273; doi:10.1111/j.1365-2249.2008.03623.x.

Formatted: English (United States)

Formatted: English (United States)

523 Magrath I, Steliarova-Foucher E, Epelman S, Ribeiro RC, Harif M, Li CK, et al. 2013.
524 Paediatric cancer in low-income and middle-income countries. *Lancet Oncol.* 14;
525 doi:10.1016/S1470-2045(13)70008-1.

526 McGill BJ, Enquist BJ, Weiher E, Westoby M. 2006. Rebuilding community ecology from
527 functional traits. *Trends Ecol. Evol.* 21:178–85; doi:10.1016/j.tree.2006.02.002.

528 McSorley HJ, Maizels RM. 2012. Helminth infections and host immune regulation. *Clin.*
529 *Microbiol. Rev.* 25:585–608; doi:10.1128/CMR.05040-11.

530 Mideo N. 2009. Parasite adaptations to within-host competition. *Trends Parasitol.* 25:261–8;
531 doi:10.1016/j.pt.2009.03.001.

532 Mohd Hanafiah K, Groeger J, Flaxman AD, Wiersma ST. 2013. Global epidemiology of
533 hepatitis C virus infection: New estimates of age-specific antibody to HCV
534 seroprevalence. *Hepatology* 57:1333–1342; doi:10.1002/hep.26141.

535 Morrow RH, Gutensohn N, Smith PG. 1976. Epstein-Barr Virus-Malaria Interaction Models
536 for Burkitt's Lymphoma : Implications for Preventive Trials Epstein-Barr Virus-
537 Malaria Interaction Models for Burkitt's Lymphoma : Implications for Preventive
538 Trials 1. 667–669.

539 Mostafa MH, Sheweita SA. 1999. Relationship between Schistosomiasis and Bladder Cancer
540 EVIDENCE SUPPORTING THE RELATIONSHIP BETWEEN SCHISTOSOMIASIS
541 AND BLADDER. 12: 97–111.

542 Oldstone MBA. 2006. Viral persistence: Parameters, mechanisms and future predictions.
543 *Virology* 344:111–118; doi:10.1016/j.virol.2005.09.028.

544 Ossai OP, Dankoli R, Nwodo C, Tukur D, Nsubuga P, Ogbuabor D, et al. 2014. Bacteriuria
545 and urinary schistosomiasis in primary school children in rural communities in Enugu
546 State ., 18: 4–8.

547 Paavonen J, Naud P, Salmerón J, Wheeler CM, Chow SN, Apter D, et al. 2009. Efficacy of
548 human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine against cervical infection
549 and precancer caused by oncogenic HPV types (PATRICIA): final analysis of a double-
550 blind, randomised study in young women. *Lancet* 374:301–314; doi:10.1016/S0140-

Formatted: English (United States)

Formatted: English (United States)

551 6736(09)61248-4.

552 Parkin DM. 2001. Global cancer statistics in the year 2000. *Lancet Oncol.* 2:533–543;
553 doi:10.1016/S1470-2045(01)00486-7.

554 Parkin DM. 2006. The global health burden of infection-associated cancers in the year 2002.
555 *Int. J. Cancer* 118:3030–44; doi:10.1002/ijc.21731.

556 Pathak AK, Pelensky C, Boag B, Cattadori IM. 2012. Immuno-epidemiology of chronic
557 bacterial and helminth co-infections: observations from the field and evidence from the
558 laboratory. *Int. J. Parasitol.* 42:647–55; doi:10.1016/j.ijpara.2012.04.011.

559 Pedersen AB, Fenton A. 2007. Emphasizing the ecology in parasite community ecology.
560 *Trends Ecol. Evol.* 22:133–139; doi:10.1016/j.tree.2006.11.005.

561 Peleteiro B, Bastos A, Ferro A, Lunet N. 2014. Prevalence of *Helicobacter pylori* infection
562 worldwide: A systematic review of studies with national coverage. *Dig. Dis. Sci.*
563 59:1698–1709; doi:10.1007/s10620-014-3063-0.

564 Quiding-Järbrink M, Lundin BS, Lönroth H, Svennerholm AM. 2001. CD4+ and CD8+ T cell
565 responses in *Helicobacter pylori*-infected individuals. *Clin. Exp. Immunol.* 123: 81–7.

566 Read AF, Taylor LH. 2001. The Ecology of Genetically Diverse Infections. 292: 1099–1103.

567 Seward JF, Watson BM, Peterson CL, Mascola L, Pelosi JW, Zhang JX, et al. 2002. Varicella
568 Disease After Introduction of Varicella Vaccine in the United States , 1995-2000. 287:
569 1995–2000.

570 Siegel R, Naishadham E, Jema A. 2013. Cancer Statistics, 2013. *CA Cancer J Clin* 37:408–
571 14; doi:10.3322/caac.21166.

572 Silins I, Ryd W, Strand A, Wadell G, Törnberg S, Hansson BG, et al. 2005. Chlamydia
573 trachomatis infection and persistence of human papillomavirus. *Int. J. Cancer* 116:110–
574 115; doi:10.1002/ijc.20970.

575 Zur Hausen H. 2009. The search for infectious causes of human cancers: where and why.
576 *Virology* 392:1–10; doi:10.1016/j.virol.2009.06.001.

577 Zur Hausen H, Villiers E De. 2015. Cancer “Causation” by Infections—Individual

578 | Contributions and Synergistic Networks. *Semin. Oncol.* 41:860–875;
579 | doi:10.1053/j.seminoncol.2014.10.003.

580

581

582

583

584

585

586

587

588 **Table 1** | Results for the three cancers of interest in females for temperate and tropical countries. TE prefix represents temperate dimensions whereas the T
589 prefix refers to tropical dimensions. Statistical significance and sign of association was obtained from analysis of variance. Range of significance describes
590 results from the rarefaction analysis. Number of infectious agents (IA) was determined from the MCA analysis. Dimensions, *P*-value, sign of association and
591 range of significance in italics mean that statistics are marginally significant.

Type of cancer	Temperate countries (TE)					Tropical countries (T)					592
	Dimensions	<i>P</i> -value	Sign of association	Range of significance	Number of IA	Dimensions	<i>P</i> -value	Sign of association	Range of significance	Number of IA	593
Bladder	Bact TE1	0.008	-	60%-100%	10	Bact T4	0.017	+	70%-100%	13	
	Bact TE4	0.02	+	60%-100%	8	<i>Helm T9</i>	<i>0.04</i>	+	<i>100%</i>	6	
	Helm TE8	0.019	-	80%-100%	7						
Total	3				25	2				19	
Liver	Bact TE1	0.009	+	80%-100%	10	<i>Bact T6</i>	<i>0.046</i>	+	<i>100%</i>	6	
	Helm TE1	0.02	+	80%-100%	22	Helm T5	0.013	+	90%-100%	11	
						Helm T6	0.002	-	60%-100%	9	
						Virus T6	0.01	-	80%-100%	14	
Total	2			32	4				40		
Stomach	Bact TE1	0.0001	+	60%-100%	10	Bact T2	0.002	-	70%-100%	10	
	Bact TE4	0.001	-	70%-100%	8	Bact T7	0.007	-	70%-100%	5	
	Helm TE1	0.002	+	70%-100%	22	Helm T2	0.0002	+	40%-100%	10	
	Helm TE2	0.0002	-	70%-100%	20	Helm T3	0.006	-	60%-100%	9	
	<i>Helm TE3</i>	<i>0.049</i>	+	<i>100%</i>	13	Virus T1	0.00003	+	60%-100%	25	
	Helm TE7	0.0006	+	50%-100%	8	Virus T5	0.023	+	70%-100%	13	
	Virus TE3	0.02	-	90%-100%	10						
	Virus TE8	0.038	-	100%	5						
Total	8			96	6				72		

594 **Table 2 | Description of the specific associations which are supported by the random redistribution**
 595 **analysis (type I error).** The combination observed in the principal analysis are described and the
 596 probability of finding them by chance was calculated over 10.000 reassortments normally distributed
 597 of cancer incidence data with the same mean and variance than in the original database in each
 598 biomes.

Interactions	Description	Probability
Bladder/Viruses	No significant association between bladder cancer incidence and dimension of viruses in both regions.	36%
Bladder/Bacteria	Significant association between bladder cancer incidence and dimension of bacteria in both regions.	12%
Liver/Negative associated dimension	No significant dimension of any guilds is negatively associated with liver cancer incidence in temperate countries. In addition, at least one significant dimension of at least one guild is negatively associated with liver cancer.	40%
Liver/Helminths	Significant association between liver cancer incidence and dimension of helminths in both regions.	16%
Stomach/Viruses	Virus dimensions negatively correlated to stomach incidence in temperate countries but positively correlated in tropical countries.	3%

599

600 **Table 3: Links between prevalence of oncogenic agents and the number of associated dimensions based on the scientific literature.** Mean prevalence for
 601 each region has been calculated according to data presented in the references.

602

Oncogenic agents	Causality by cancer subtype	Global causality	Temperate countries		Tropical countries	
			Mean prevalence	Number of associated dimensions (Number of infectious agents)	Mean prevalence	Number of associated dimensions (Number of infectious agents)
<i>H. pylori</i>	80% of adenocarcinoma	70% of stomach cancer	41%(Peleteiro et al. 2014)	8 (96)	53%(Peleteiro et al. 2014)	6 (72)
HCV/HBV	80% of hepatocellular carcinoma	60% of liver cancer	4.3%(Mohd Hanafiah et al. 2013)	2 (32)	4.2%(Mohd Hanafiah et al. 2013)	4 (40)
<i>S. haematobium</i>	30% of squamous cells carcinoma	2% of bladder cancer	0.000006%(Chitsulo et al. 2000)	3 (25)	4.5%(Chitsulo et al. 2000)	2 (19)

Field Code Changed

603 | **Figure 1| Distinction between temperate and tropical countries. A) Infectious**
604 | **communities in temperate and tropical biomes assessed by MCA. Tropical countries show a**
605 | **different community of infectious agents from temperate countries even if we can observe**
606 | **some overlaps. B) Values of the first dimension of the PCA on the confounding variables at**
607 | **worldwide scale. Confounding variables of tropical countries largely group together in a**
608 | **cohesive range of the parameter space, overlapping with those of temperate countries only at**
609 | **high values. Vertical lines represent the geographical location of the tropics.**

610 |
611 | **Figure 12| Flow diagram summarizing the main statistical analyses undertaken in this**
612 | **study.**

613 |
614 | **Figure 3| Hypothesis on the contrasting role of viruses between the two biomes.**