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

Elsevier Editorial System(tm) for Infection, Genetics and Evolution

Manuscript Draft

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 desserves.

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 intepretation 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 persistance 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.

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

Camille Jacqueline[1,2§], Jessica L. Abbate[2,3], Gabriele Sorci[4], Jean-François Guégan[2], Fréderic Thomas[1,2], Benjamin Roche[1,3]

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

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

Short title: Macroecology of cancer and endemic infections

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.

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

Introduction

While cancer remains one of the main causes of death in Western countries (Ferlay et al., 2010), its burden is increasing in low- and middle-income countries (Magrath et al., 2013). Although treatments, including chemo-, radio- and/or immunotherapy, have resulted in a slight decline in death rates worldwide, cancer incidence rates have remained stable over the past 10 years (Siegel et al., 2013). In this context, prevention seems currently the most efficient strategy to reduce the impact of cancer on populations. 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). For example, current evidence links Epstein–Barr virus (EBV), Hepatitis B and C viruses (HBV, HCV), the bacteria *Helicobacter pylori*, human papillomavirus (HPV), and the trematode *Schistosoma haematobium* to cancers of the lymph nodes, liver, stomach, cervix and bladder respectively. Consequently, vaccination against oncogenic agents is currently being considered as a potential key tool to prevent these cancers, such as the HPV vaccine which offers a relative protection against cervical cancer (Paavonen et al., 2009).

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

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

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

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

Material & methods

Data overview and inclusion criteria

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

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

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

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

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

Assessment of infectious agent assemblages

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

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

We conducted separate analyses for each guild (viruses, bacteria, helminths) within each biome (tropical and temperate) and we transformed cancer incidence data (ASI) to obtain Gaussian distributions (transformations are detailed in supplementary materials Table S2). 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 will henceforth refer to this statistical analysis as the "principal analysis". 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 analysis (using the 'accuracy' function in the package *rcompanion*; see Table S3).

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

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

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

Robustness assessment

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

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

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

Results

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

Bladder cancer

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

Liver cancer

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

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

Stomach cancer

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

Discussion

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

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

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

where the bacterium *Bordetella bronchiseptica* was shown to enhance helminth intensity through immune mediated effects (Pathak et al., 2012). Thus, our study may highlight the role of bacteria in cancer associated with helminth infections. 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.

Regarding liver cancer, we observed a specific and robust association between helminths and cancer incidences in both biomes. The interactions between certain helminths species and HCV/HBV have been reported in experimental studies and may be either protective or facilitator. It has been suggested that co-infection with *S. mansoni* could increase the persistence and severity of HCV infection because the helminth prevent the production of an HCV-specific CD4+/Th1 T cell response (Kamal et al., 2001). Conversely, a recent study has demonstrated the immunostimulant effect of a protein derived from *Onchocerca volvulus* which increases the interferon-γ response to HCV in vitro (MacDonald et al., 2008). Even if the identified component communities do not highlight these two particular species, it seems worth to expect that similar mechanisms could apply to other helminth species (McSorley and Maizels, 2012). 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.

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. This was the case, with the largest number and diversity of species across guilds associating both positively and negatively with this cancer. The principal result for stomach cancer is that component communities of viruses have an opposite effect between the two biomes. 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). 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).

As for any statistical approach dealing with large scale database, our study is susceptible to a number of issues which need to be discussed. First, some diseases based on observation of symptoms have multiple potential causative agents. When this was the case, we included only the agent responsible for the majority of cases. Second, a principal weakness of our study is that infectious assemblages were based on presence/absence data where even imported and isolated cases (observed rarely) can result in a presence assignation for the whole country. In addition, all worldwide-distributed infectious agents have been removed from the database because they would not be discriminating at all. These two last points suggested that it would be helpful to consider prevalence data in order to assess the proportion of the population which is really in contact with the infectious agent. 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. This consideration, however, do not necessarily negate the indirect impact of the communities we have identified. Finally, we defined guilds according to their taxonomic classifications, which could be not biologically meaningful, but which allows to deal with the high number of infectious agents considered here. In fact, it has been recommended that guilds should be based on functional similarity of species or based on their life-cycle categories instead of their taxonomy (McGill et al., 2006).

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

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.

We have based our interpretations here on the hypothesis that infectious agents may modify the circulation or the persistence of oncogenic agents. 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. Finally, the inclusion of a cancer for which no infectious agent is suspected to play a role would have served as an evidence-boosting negative control. However, the assemblage identified in these controls could also have biological relevance as they could include species that indirectly alter immunosurveillance as suggested in (Jacqueline et al., 2017).

Conclusion

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

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

Acknowledgments

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

Competiting interests: No conflict of interests to declare.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Husson, F., Josse, J., Le, S., Maintainer, J.M., 2017. Package "FactoMineR" Title 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-
482 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

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 â€TM s Lymphoma : Implications for Preventive Trials Epstein-Barr
498 Virus-Malaria Interaction Models for Burkitt â€TM 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

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

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

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

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

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

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

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

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

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

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

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

Zur Hausen, H., Villiers, E. De, 2015. Cancer "Causation" by Infections—Individual Contributions and Synergistic Networks. Semin. Oncol. 41, 860–875. https://doi.org/10.1053/j.seminoncol.2014.10.003

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

	Temperate countries (TE)					Tropical countries (T)				
Type of cancer	Dimensions	*P*-value	Sign of association	Range of significance	Number of IA	Dimensions	*P*-value	Sign of association	Range of significance	Number of IA
Bladder	Bact TE1	0.008	-	60%-100%	10	Bact T4	0 .017	+	70%-100%	13
	Bact TE4	0.02	+	60%-100%	8	*Helm T9*	*0.04*	*+*	*100%*	*6*
	Helm TE8	0.019	-	80%-100%	7					
Total	**3**				**25**	**2**				**19**
Liver	Bact TE1	0.009	+	80%-100%	10	*Bact T6*	*0.046*	*+*	*100%*	*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
	Helm TE3	*0.049*	*+*	*100%*	*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)
H. pylori	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)
S. haematobium	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)

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

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

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

Figure 1

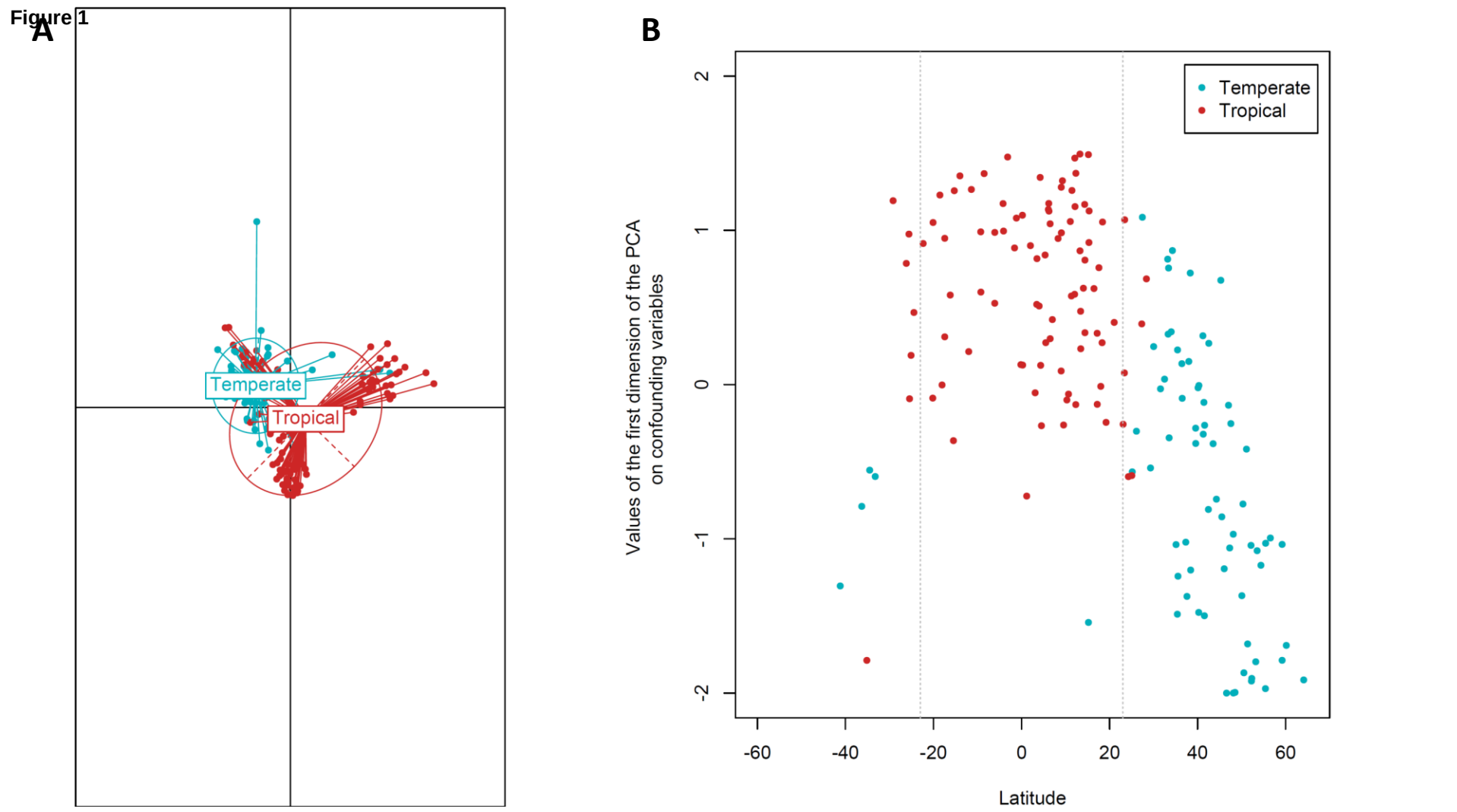

A
Temperate
Tropical
B
Temperate
Tropical
Values of the first dimension of the PCA on confounding variables
2
1
0
-1
-2
-60
-40
-20
0
20
40
60
Latitude

Figure 2

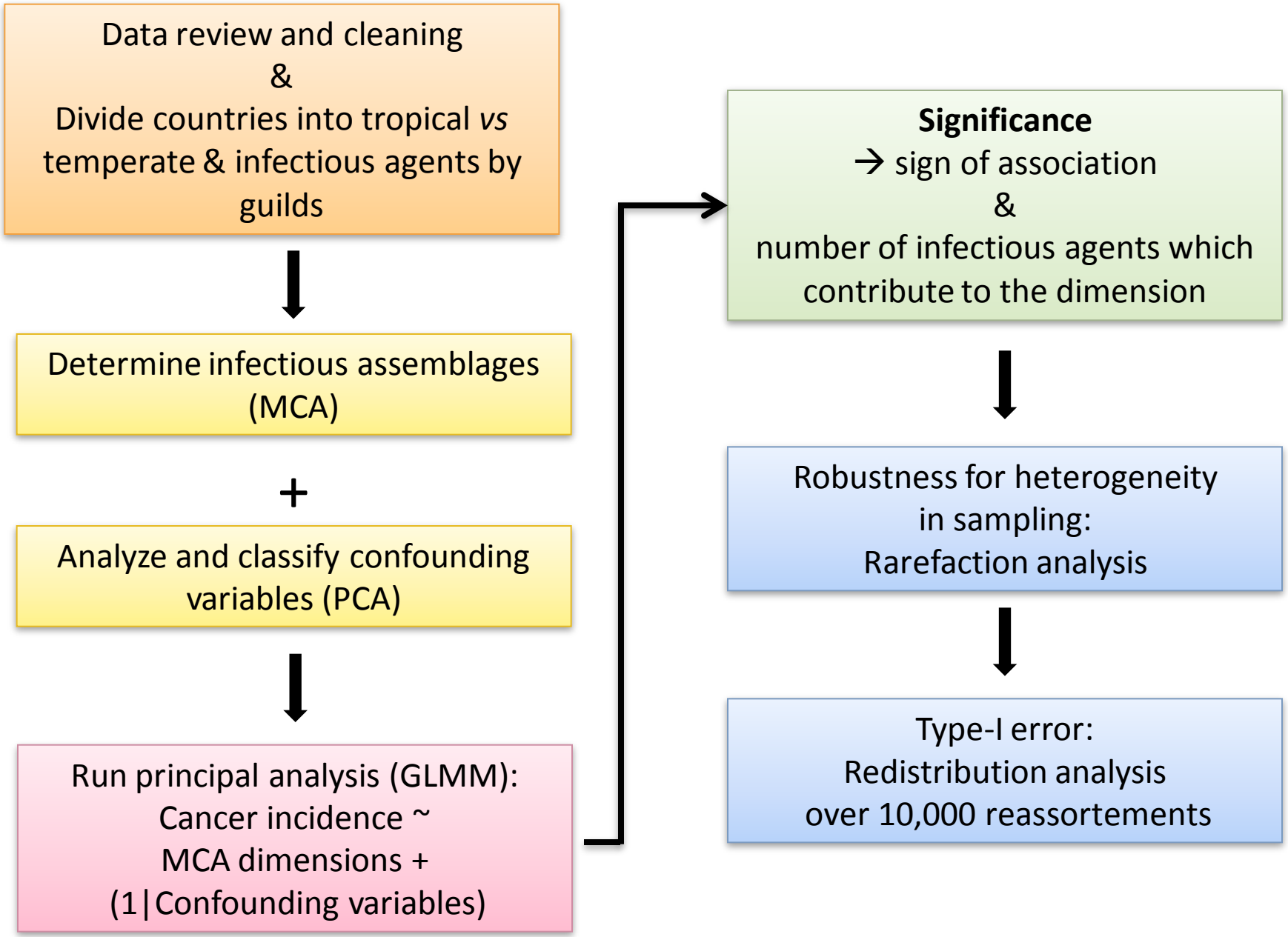
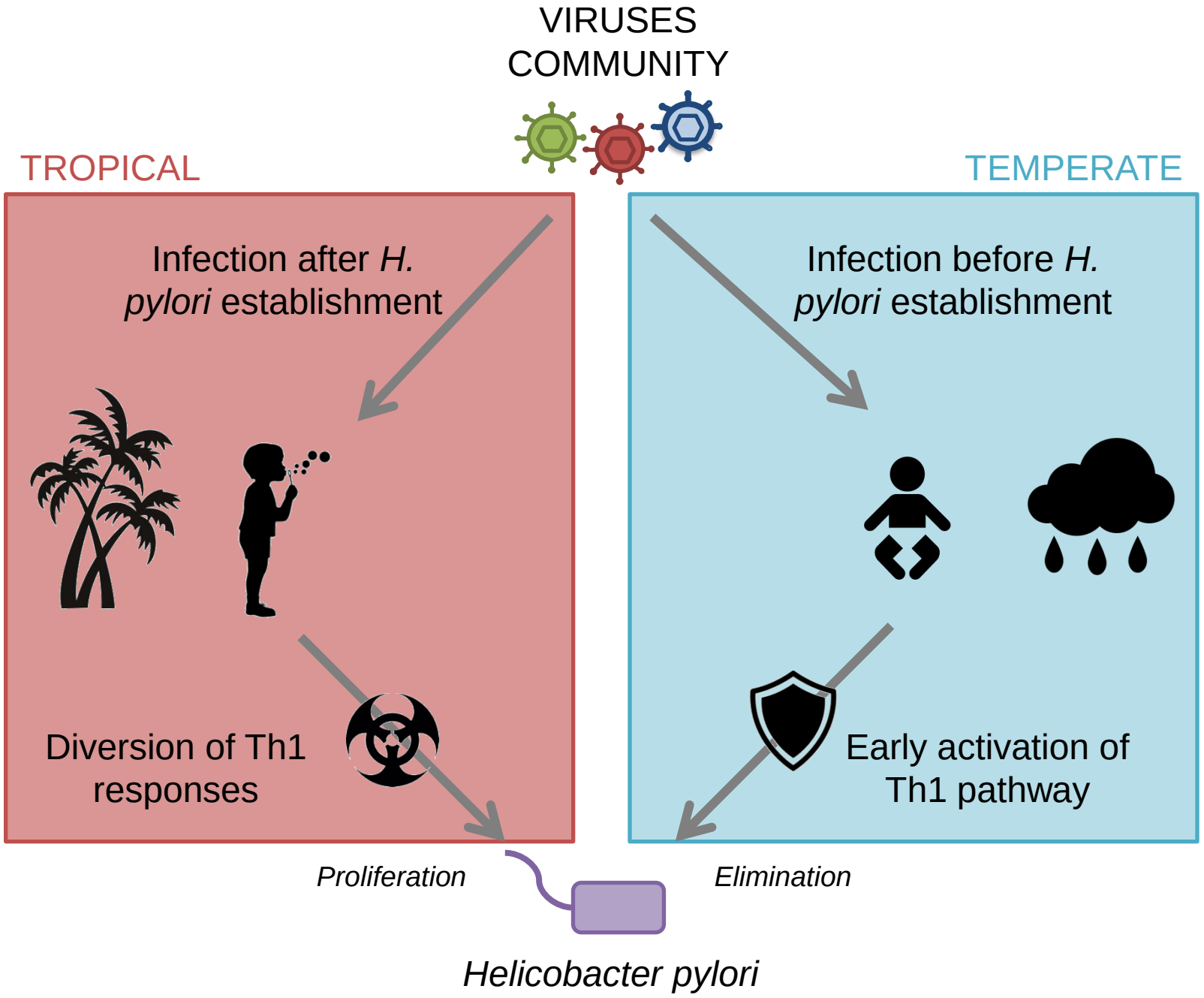

Data review and cleaning
&
Divide countries into tropical *vs* temperate & infectious agents by guilds
Determine infectious assemblages (MCA)
+
Analyze and classify confounding variables (PCA)
Run principal analysis (GLMM):
Cancer incidence ~
MCA dimensions +
(1|Confounding variables)
Significance
→ sign of association
&
number of infectious agents which contribute to the dimension
Robustness for heterogeneity in sampling:
Rarefaction analysis
Type-I error:
Redistribution analysis
over 10,000 reassortements

Figure 3


VIRUSES
COMMUNITY
TROPICAL
TEMPERATE
Infection after H. pylori establishment
Infection before H. pylori establishment
Diversion of Th1 responses
Early activation of Th1 pathway
Proliferation
Elimination
Helicobacter pylori

Supplementary Material

Click here to download Supplementary Material: Annexes VF3-revised.docx

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

Camille Jacqueline[1,2§], Jessica L. Abbate[2,3], Gabriele Sorci[4], Jean-François Guégan[2], Fréderic Thomas[1,2], Benjamin Roche[1,3]

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

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

Short title : Macroecology of cancer and endemic infections

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, infectious assemblages found to be protective against liver cancer in the tropics were not associated with a protective effect among temperate countries. 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.

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

Introduction

While cancer remains one of the main causes of death in Western countries (Ferlay et al. 2010), its burden is increasing in low- and middle-income countries (Magrath et al. 2013). Although treatments, including chemo-, radio- and/or immunotherapy, have resulted in a slight decline in death rates worldwide, cancer incidence rates have remained stable over the past 10 years (Siegel et al. 2013). In this context, prevention seems currently the most efficient strategy to reduce the impact of cancer on populations. 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, ~~N~~numerous infectious agents ~~(for the sake of clarity, the term "infectious agent" is used throughout the manuscript to describe 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…)~~ have been recognized as ~~a~~ 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). For example, current evidence links Epstein–Barr virus (EBV), Hepatitis B and C viruses (HBV, HCV), the bacteria *Helicobacter pylori*, human papillomavirus (HPV), and the trematode *Schistosoma haematobium* to cancers of the lymph nodes, liver, stomach, cervix and bladder respectively. Consequently, vaccination against oncogenic agents is currently being considered as a potential key tool to prevent these cancers, such as the HPV vaccine which offers a relative protection against cervical cancer (Paavonen et al. 2009).

Formatted: Superscript

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

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

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

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

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

Material & methods

Data overview and inclusion criteria

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

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

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

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

Assessment of infectious agent assemblages

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

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

We followed a similar methodology for the 24 confounding variables described previously (substituting the Multiple Correspondence Analysis with a Principal Component Analysis, as our confounding variables contained quantitative values rather than categorical presence/absence). Given that the first dimension of the PCA on confounding variables explained more than 40% of the variance within each biome (Supplementary materials ~~Figure~~ Fig. S2~~4~~), we classified the values of this dimension into five evenly distributed classes of countries. These classes were then used as a random factor in statistical models to consider intra-biome variability in social and economic level (see below).

We conducted separate analyses for each guild (viruses, bacteria, helminths) within each biome (tropical and temperate) and we transformed cancer incidence data (ASI) to obtain Gaussian distributions (transformations are detailed in supplementary materials Table S2). 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 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.~~In a generalized linear mixed model (GLMM) implemented in the package *lme4* (Bolker et al. 2009), transformed ASI were used as the response variable, the MCA dimensions representing infectious assemblages were used as fixed factors and the classes of confounding variables as random factors. We did not consider interactions between fixed factors because of combinatory explosion. Variable selection was conducted through analysis of variance (ANOVA) (using the 'anova' function in package *car*, with test specified as "type III" that tests the effect of each variable after all other factors have been accounted for (Fox and Weisberg 2011)).~~ We will henceforth refer to this statistical analysis as the "principal analysis".

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). After having considered each guild independently (one GLMM by guild/by cancer), we ran a GLMM for each cancer (as described previously) considering all guilds together. In this relative contribution analyses, we used only the significant dimensions identified in the principal analysis as fixed factors (see Table ~~S3~~S4).

Finally, for each significant MCA dimensions involved in the principal analysis, we identified the infectious agents that explained more that 5% of the variability of the significant dimensions (represented by "eta2" values in the MCA). For each cancer, we grouped all of these infectious agents (separately for those with positive versus negative correlation with cancer incidences), and we refered to each of these groups as an "infectious assemblage". As the agents in each assemblage may not necessarily interact with one-another, we distinguished the term "assemblage" from "community". We described the size and the composition of the assemblages for each cancer in Tables ~~S4~~S5, S6~~5~~ and S7~~6~~ of the Supplementary Materials.

Robustness assessment

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

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

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

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

Results

After applying MCA on each guild and removing the non-explicative dimensions according to the Kaiser criterion, we obtained 7 dimensions explaining variance for bacterial species occurrence, 9 dimensions for helminths and 8 for viruses in temperate countries. For tropical countries, 8, 10 and 11 dimensions were kept for bacteria, helminths and viruses, respectively. We found that GLMMs fit the actual data relatively well with an accuracy between 70% and 90% (see Table S3). ~~-~~Because our results were similar for both sexes (see ~~Supplementary data~~ Table ~~S7~~S8), we report here only the results for females (Table 1).

Bladder cancer

In temperate biome, bladder cancer was negatively associated with one dimension of bacteria (BACT TE1) and one dimension of helminths (HELM TE8), but positively associated with a second dimension of bacteria (BACT TE4) (Table 1). Two dimensions (one for bacteria (BACT T4) and one for helminths (HELM T9)) were positively associated with bladder cancer incidence in tropical countries. The associations found with bacteria were well supported in both biomes in the relative contribution analysis (~~Supplementary materials~~see Table ~~S3~~S4), as well as with the rarefaction analysis (Table 1). ~~and s~~ Such association were found in 12% of simulated normally-distributed incidences (Table 2). In addition, the redistribution analysis showed that the absence of virus dimensions in both temperate and

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

Liver cancer

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

Stomach cancer

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

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

Discussion

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

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

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

transmission pattern, with an extremely high prevalence in tropical countries, but circulating only in several countries within the temperate biome. First, we found that bacteria may have a preponderant role in both biomes. ~~Among component communities associated to bladder cancer, only four bacterial species have a congruent role between the two biomes (*Rickettsia felis, Rickettsia japonica, Rickettsia sibirica, Fusobacterium necrophorum*).~~ Interestingly, studies have shown that bacteria-helminth interactions have already been observed on the field and could impact cancer risk. Notably, a high percentage of co-infection with *S. haematobium* and some unspecified bacteria in the urinary tract has been reported (Adeyeba and Ojeaga 2002; Ossai et al. 2014). These co-infections could increase the risk of bladder cancer as bacteria in the urinary tract produce nitrosamines, which are carcinogenic compounds (Davis et al. 1984). In addition, bacteria may bias the Th1/Th2 balance away from protection against helminths by down-regulating Th2-mediated responses. This kind of indirect interaction has already been seen in rabbits where the bacterium *Bordetella bronchiseptica* was shown to enhance helminth intensity through immune mediated effects (Pathak et al. 2012). Thus, our study may highlight the role of bacteria in cancer associated with helminth infections. 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.~~. In fact, helminths have been shown to favor viral infection due to their immunoregulatory role~~ (Kamal and El Sayed Khalifa 2006)~~, , and it is likely that infections with certain viruses known to immunosuppress their host~~ (~~reviewed by~~ Oldstone 2006))~~ could have consequences for subsequent helminth infections~~. As suggested by the redistribution analysis, this result needs to be taken with precaution. Indeed, t~~T~~he 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~~, thus our results need to be taken with precaution~~.

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

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

As the oncogenic agent *H. pylori* is highly prevalent worldwide (Atherton and Blaser 2009), ~~S~~stomach cancer had a particular likelihood to be associated with ~~large~~ species-rich component communities in both biomes~~, as the oncogenic agent *H. pylori* is highly prevalent worldwide (Atherton and Blaser 2009)~~. This was the case, with the largest number and diversity of species across guilds associating both positively and negatively with this cancer. The principal result for stomach cancer is that component communities of viruses have an opposite effect between the two biomes. 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). 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). ~~Viruses have numerous mechanisms to escape immune system and especially they are able to down-regulate interferon signaling (Gale and Sen 2009). However, IFN γ responses produced by CD8+ cells contribute significantly to the elimination of *H.*~~

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

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

Formatted: Font: Not Bold

Considering cancer incidence data, the Globocan database (compiled by the IARC) do not account for the different cancer subtypes. As oncogenic agents are implied in subtypes that are only a proportion of the total number of cases, our conclusions may be more robust when infection is associated with a prevalent cancer (stomach adenocarcinoma and hepatocellular carcinoma) as opposed to rarer subtypes (e.g., SCCs of the bladder). ~~Furthermore, the detection of cancer cases could be better in temperate countries because of the higher investment in health and disease surveillance.~~ 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 our dataset on 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.~~However, we assess this disparity using indices provided by the IARC, which described the methods and data quality. These indices have been considered among the confounding variables used as random factors in our analyses (see supplementary materials).~~

We have based our interpretations here on the hypothesis that infectious agents may modify the circulation or the persistence of oncogenic agents. 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 this study does not claim to provide a definitive answer, it calls for designing cohort studies of cancer patients considering personal infection history to determine the causality, and potentially the mechanism, of such indirect link between non-oncogenic infectious agents and cancer development.~~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. One alternative, though not mutually-exclusive, hypothesis is that infectious agents belonging to the community could play a role at the cellular level by disturbing immunosurveillance, introducing viral genes, promoting mutation or providing an environment where cancer growth can go unchecked (Dalton-Griffin and Kellam 2009). Some of these oncogenic mechanisms require that infectious agents may have a specific tropism for cancer cells, i.e., stomach, bladder or liver cells, but to the best of our knowledge species identified by our analysis do not necessarily show such tropism and cohort studies of cancer patients at the individual scale would be necessary to detect this mechanism.~~ Finally, the inclusion of a cancer for which no infectious agent is suspected to play a role would have served as an evidence-boosting negative control. However, the assemblage identified in these controls

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

Conclusion

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

Acknowledgments

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

Competing interests: No conflict of interests to declare.

References

Adeyeba OA, Ojeaga SGT. 2002. URINARY SCHISTOSOMIASIS AND CONCOMITANT URINARY TRACT PATHOGENS AMONG SCHOOL CHILDREN IN METROPOLITAN IBADAN ,. Afr J Biomed Res 5: 103–108.

Atherton JC, Blaser MJ. 2009. Review series Coadaptation of Helicobacter pylori and humans : ancient history , modern implications. J. Clin. Invest. 119; doi:10.1172/JCI38605DS1.

Borkow G, Weisman Z, Leng Q, Stein M, Kalinkovich A, Wolday D, et al. 2001. Helminths , Human Immunode ciency Virus and Tuberculosis. 568–571.

Brenner H, Rothenbacher D, Arndt V. 2009. Epidemiology of Stomach Cancer. 467–477.

Cain ML (Michael L, Bowman WD, Hacker SD. 2011. *Ecology*. Sinauer Associates.

Chêne A, Donati D, Guerreiro-Cacais AO, Levitsky V, Chen Q, Falk KI, et al. 2007. A molecular link between malaria and Epstein-Barr virus reactivation. PLoS Pathog. 3:e80; doi:10.1371/journal.ppat.0030080.

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

Davis CP, Cohen MS, Gruber MB, Anderson MD, Warren MM. 1984. Urothelial hyperplasia and neoplasia: a response to chronic urinary tract infection in rats. J Urol 132: 1025–31.

Doll R, Payne P, Waterhouse JA. 1996. *Cancer Incidence in Five Continents, Vol. I Union Internationale Contre le Cancer*. Geneva.

Ezenwa VO, Jolles AE. 2014. Opposite effects of anthelmintic treatment on microbial infection at individual versus population scales. 8–11.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. 2010. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer 127:2893–2917; doi:10.1002/ijc.25516.

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

Formatted: English (United States)

Guernier V, Hochberg ME, Guégan J-F. 2004. Ecology drives the worldwide distribution of human diseases. PLoS Biol. 2:e141; doi:10.1371/journal.pbio.0020141.

Hotez PJ, Molyneux DH, Fenwick A, Kumaresan J, Sachs SE, Sachs JD, et al. 2007. Control of Neglected Tropical Diseases. 1018–1027.

Husson F, Josse J, Le S, Maintainer JM. 2017. Package "FactoMineR" Title Multivariate Exploratory Data Analysis and Data Mining.

Jacqueline C, Tasiemski A, Sorci G, Ujvari B, Maachi F, Missé D, et al. 2017. Infections and cancer: the "fifty shades of immunity" hypothesis. BMC Cancer in press:1–11; doi:10.1186/s12885-017-3234-4.

Kaiser HF. 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23:187–200; doi:10.1007/BF02289233.

Kamal SM, Bianchi L, Al Tawil A, Koziel M, El Sayed Khalifa K, Peter T, et al. 2001. Specific Cellular Immune Response and Cytokine Patterns in Patients Coinfected with Hepatitis C Virus and *Schistosoma mansoni*. J. Infect. Dis. 184:972–982; doi:10.1086/323352.

Kamal SM, El Sayed Khalifa K. 2006. Immune modulation by helminthic infections: Worms and viral infections. Parasite Immunol. 28:483–496; doi:10.1111/j.1365-3024.2006.00909.x.

Kantor AF, Hartge P, Hoover RN, Fraumeni JF. 1988. Epidemiological characteristics of squamous cell carcinoma and adenocarcinoma of the bladder. Cancer Res. 48: 3853–5.

Khoury MJ, Ioannidis JP a. 2014. Big data meets public health. Science (80-.). 346:1054–1055; doi:10.1126/science.aaa2709.

Lee BW. 1998. Review of varicella zoster seroepidemiology in India and South-east Asia. Trop. Med. Int. Heal. 3:886–890; doi:10.1046/j.1365-3156.1998.00316.x.

MacDonald AJ, Libri NA, Lustigman S, Barker SJ, Whelan MA, Semper AE, et al. 2008. A novel, helminth-derived immunostimulant enhances human recall responses to hepatitis C virus and tetanus toxoid and is dependent on CD56+ cells for its action. Clin. Exp. Immunol. 152:265–273; doi:10.1111/j.1365-2249.2008.03623.x.

Formatted: English (United States)

Formatted: English (United States)

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

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

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

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

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

Morrow RH, Gutensohn N, Smith PG. 1976. Epstein-Barr Virus-Malaria Interaction Models for Burkitt â€™s Lymphoma : Implications for Preventive Trials Epstein-Barr Virus-Malaria Interaction Models for Burkitt â€™s Lymphoma : Implications for Preventive Trials 1. 667–669.

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

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

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

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

Formatted: English (United States)

Formatted: English (United States)

6736(09)61248-4.

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

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

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

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

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

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

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

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

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

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

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

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

Formatted: English (United States)

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

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.

	Temperate countries (TE)					Tropical countries (T)				
Type of cancer	Dimensions	*P*-value	Sign of association	Range of significance	Number of IA	Dimensions	*P*-value	Sign of association	Range of significance	Number of IA
Bladder	Bact TE1	0.008	-	60%-100%	10	Bact T4	0 .017	+	70%-100%	13
	Bact TE4	0.02	+	60%-100%	8	*Helm T9*	*0.04*	*+*	*100%*	*6*
	Helm TE8	0.019	-	80%-100%	7					
Total	**3**				**25**	**2**				**19**
Liver	Bact TE1	0.009	+	80%-100%	10	*Bact T6*	*0.046*	*+*	*100%*	*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
	Helm TE3	*0.049*	*+*	*100%*	*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**

592
593

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

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)
H. pylori	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)
S. haematobium	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

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

Figure ~~1~~2| Flow diagram summarizing the main statistical analyses undertaken in this study.

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