Reviewer 1

The authors discuss the nature of virus protein-protein interaction information that is currently available in various databases, and how these large-scale data might be mined. I have several issues with this manuscript.

Firstly, according to the request to review that I received, it is my understanding that this paper is an invited review article, yet it appears to contain primary research material – namely the analyses presented in figure 1,2 and 3. The structure of the paper is consistent with a review article, but without a fully described method section, I feel it would be inappropriate to publish these analyses in the current format.

The writing is often less formal than is found in many scientific manuscripts. In most cases, I enjoy this because it improves the readability. However, there are some phrases, I would consider revising: ‘It turns out…’ and ‘Virus proteins attack…’ are two I find too colloquial.

How do we respond to that? That sounds slight anal …

I also found that acronyms are frequently not defined, and are sometimes defined on the second use. In particular, PPI in the abstract is not defined anywhere. As many viruses form protein-protein interactions with peptidyl-prolyl isomerases, PPI as an acronym is confusing if not defined. I would suggest, as it is not used consistently throughout the manuscript, that PPI be removed.

*We corrected this now and PPIs is defined. Given that PPI is used in the manuscript 15 times, we prefer to keep it, given that it is a commonly used abbreviation in the field.*

I also find some of the comments in the introduction and conclusions to be generic and unhelpful (if not inaccurate). The assertion in the introduction that viruses are more difficult to control than bacteria due to the use of antibiotics is a very debatable statement, and given that the paper is not about bacteria, this statement is unnecessary.

*The reviewer is right in that this statement is not absolutely necessary but it puts the role of viruses into perspective, not the least because PPIs are a potentially important target for virus treatment, as opposed to antibiotics many of which target metabolic enzymes in bacteria. We also added a statement to explain this better.*

In the conclusion: ‘Viruses evolve much quicker than their hosts, especially RNA viruses, hence viruses can also adapt their virus-host interactions faster than a host can react by mutating its target proteins, including its immune system.’ This statement is at odds with the Red Queen hypothesis – that virus and host co-evolve to maintain mutual survival. Indeed the adaptive arm of the host immune system can, in many cases, actively evolve to eliminate infection faster than the virus can respond. Again, I’m not sure of the value of this statement in the context of the conclusions.

We have clarified this statement now. We agree that the adaptive immune system may keep up with an infection, a population of hosts can never keep up with a (much larger) virus population.

I am also concerned about the oversimplification of some of the biology in this paper. In Table 1, viruses are classified as either DNA or RNA (or ‘anim’ or ‘plant’ is some cases – why?). The Baltimore system of classification of viruses defines seven types of genome: dsDNA, ssDNA, dsRNA, ssRNA(+ve), ssRNA(-ve), ssRNA-RT, and dsDNA-RT. These different groups have different replication requirements which impacts on their mutation rate (amongst other things). In table 1, hepdnaviridae are labeled as DNA viruses – which isn’t strictly true as they have an RNA intermediate. Likewise, retroviridae have a DNA intermediate. So again, I am unsure as to the value of the describing the virus families in terms of being RNA or DNA, as is done in Table 1.

*We agree with this reviewer. The classification in Table 1 reflected the usage in Genbank, which is not entirely consistent. We have now removed the labels for plants and animals. Norman, can you look into this?*

In Figure 2 (and the accompanying section ‘The topology of human-virus interactions’) the authors describe critical nodes in interaction networks as statistically more likely to be targets from protein interactions. Intuitively, this seems reasonable, and I like it as a concept. But could it also be that these critical nodes are also overrepresented in false-positives in large-scale interactome analyses?

Stefan – can you address this?

In Figure 3, I am intrigued to see that retroviridae and orthomyxoviridae have a negative fold change with respect to cell wall proteins, while herpesviridae and papillomarviridae have a positive fold change. None of these families of virus infect an organism with a cell wall. This calls into question whether the authors understand the systems they are studying. Also, there appears to be a phylogenetic analysis on both axes of the heat map, but this is not described or mentioned.

Stefan?

I also question the value of putting ‘other’ in the heat map. Given the assertion that different viruses have different interaction preferences, surely grouping everything that is not of the aforementioned families together would result in a largely featureless map, averaged around no change – which is close to what is shown.

We agree and have removed the “other” category. (Stefan?)

On the whole, I find this paper to be strangely constructed. It is a hybrid review/original research piece, with a smattering of ideas and concepts, with limited flow between sections.

*We have tried in the revision to improve the structure and flow between sections. Check at the end of editing.*

The conclusions of the paper are particularly unremarkable. The statement ‘…suggesting that different viral families use different strategies to invade a human host cell’ to be one of the most unprofound statements I have seen in a manuscript.

*This statement has been changed.*

As the conclusions read, the ‘take home message’ appears to be that: there are many interactome databases; some viruses are overrepresented and full of false positives; other viruses are underrepresented; and viruses are different from each other. So I would question what the value of such a paper would have to the scientific community.

*The intention of the paper was to provide a snapshot of our knowledge of virus-virus and virus-host PPIs, which is what we have done. The nature of the topic doesn't really allow an in-depths discussion of how these PPIs affect virus infections and their pathobiology, which may be unfortunate but way beyond the scope of this paper.*

Reviewer 2

In this manuscript, Goodacre and colleagues broadly examine protein-protein interactions (PPIs) in the context of human virus infections and discuss the physiological relevance of these interactions. While a large number of PPIs have been identified, the authors raise the question of how many interactions are physiological and required for viral infection. Interestingly, there is considerable overlap in the human proteins targeted by different virus families. The authors describe the databases available for the study of human-virus PPIs and provide a useful summary of the strengths and weaknesses of the databases and their potential uses. Overall, this work is likely to be of interest to a broad readership in virology and addresses some critical questions as the use of systems biology approaches becomes more prevalent.

Specific comments:

1. When questioning the reliability of published PPIs, the authors describe cases where the predicted interactions were validated by co-immunoprecipitation and GST pull-down experiments. However, another critical question is whether these interactions been shown to have a functional effect on viral infection (i.e. by examining viral replication when the interaction is blocked by knockdown, small molecule inhibitors, or genetics approaches). Extending this section to describe and provide examples of functional effects would strengthen the manuscript.

*We agree with this reviewer that such information would be highly interesting. Unfortunately, to our knowledge, there are no systematic attempts to investigate the functional role of host-virus (or even virus-virus PPIs). However, we have added a statements relating PPIs to the importance of host proteins for virus infections as found in genetic screens.*

2. Could the authors comment on potential effects of different cell lines in terms of PPIs? One problem is that the cell lines used to study viral infection are not always physiological (e.g., the use of cancer cell lines, where signaling pathways and protein expression levels may be disregulated). One can imagine that PPIs identified in such cell lines may be even less likely to accurately reflect the physiological situation. Some discussion of this, or providing some examples, would be appreciated.

*We completely agree with this reviewer. Again there, aren’t many studies addressing this but we have now added a few statements explaining this problem as well as some related references.*

3. Many viruses infect different types of hosts (e.g., Dengue is spread by mosquitoes, Influenza infects birds and pigs). Have any studies compared PPIs in different hosts for viruses such as these? While the reviewer appreciates this manuscript is focused on human disease, it may be interesting to include a brief section about PPIs in other hosts for this group of viruses, many of which are medically important human pathogens.

*This is an important question and we agree with the reviewer that this was not adequately addressed in the previous version. We have now added a discussion of this problem in the new section “How are protein-protein interactions related for viruses with multiple hosts”. Norman, can you check Intact for PPIs among one virus but different host species?*

4. More discussion of the outlook and future directions would be useful for researchers in the field. How can we now exploit this collection of PPIs to target viral infections or diseases associated with the viral interactome? How could databases be improved? How can we better ensure that identified virus-human PPIs are biologically relevant?

*I can do that, but need to think a bit about it…*

Minor comments:

1. Abstract, line 10: insert comma “…relevant, for many…”

2. Page 2, Introduction, paragraph 1, line 4: insert comma “…and structure, viruses…”

3. Page 3, Virus-host interactions in humans, line 9: change to “…more than 100 interactions…” instead of “more than a 100”

4. Page 6, line 2: “Betweenes centrality” should be “Betweenness centrality”?

5. Page 7, Virus interactions with the host metabolome, line 4: change to “…with the cancer protein…” instead of “th cancer”

*We thank the reviewer for spotting those and have fixed all typos and punctuation errors.*

Reviewer 3 (apparently Vincent Lotteau :S)

This manuscript provides a global survey of virus-host PPIs in literature and databases. Recent advances and limitations of these resources are discussed. This is an update of the current situation, but some paragraphs have significant weaknesses. Several points should be addressed.

Point 1) In page 2 (second paragraph), some sentences are misleading. Indeed, it could be wrongly understood that « medical interest » is somehow related to higher « infectivity », which could be correlated with higher « pathogenicity » (“In fact, …”). This is confusing and should be clarified.

*We are sorry about the confusion. The statement is further clarified in the next section with “medically important” being defined as number of “infected” and “sick” people as well as mortality, hence we believe this give a reasonable impression of the medical significance of these viruses.*

Point 2) Table 1. Which numbers were used to calculate the U/C ratio? For example, 1008 total genomes are available for Paramyxoviridae and clustered into 327 groups. However, the U/C ratio is 10.06.

Norman, can you clarify and add an explanation to the text or the table??

The reviewers are correct – the U/C ratios were off. This is because we originally drew the Table from a larger spreadsheet, wherein there were several U/C ratios. We accidentally transferred the U/C ratios for all sequences, unclustered/clustered, rather than just complete genomes. I have amended the table with the correct U/C ratios.

Point 3) Table 1: Arenaviridae are responsible for major human diseases such as Lassa fever, and should not be restricted to “animals”.

*We have corrected this mistake and added Lassa fever to Table 1.*

Point 4) In the legend of Table 1, it is said “Clustered sequenced were clustered at ≥98% sequence identity”. This is different from the main text (Page 2; sequences are clustered and those genomes combined which have less than 5% sequence…”).

Norman?

Table 1 is correct. The text from page 2 should read “sequences are clustered and those genomes combined which have less than 2% sequence identity… “

Point 5) Page 2: “We have thus compiled data on the diversity of human viruses by looking both at sequence diversity (Table 1)…”. Table 1 compiled data on sequence diversity inside viral families, not “viruses” or “human viruses”. This should be corrected.

*We have now corrected this with an expanded introduction one the* ***Diversity of human viruses*** *to also cover within-family diversity of viruses, and the effect of this on pathogenicity.*

Point 6) Table 2: The viral family should be indicated for each virus.

Norman, can you add those?

Table 2 now contains the viral families

Point 7) Page 3: “Coronaviruses, are highly pathogenic with a lot of victims, but have received rather scant attention”. This is contradictory with data presented in Table 2, since coronaviruses were responsible for a very small number of victims (<10.000 for SARS-CoV and MERS-CoV over several years). In fact, dengue virus, hRSV or rabies virus that are not cited in Table 2 have a much higher impact on human populations.

*The reviewer is correct in that this is somewhat contradictory and that Table 2 is incomplete. We have now expanded the table and added more detail to the text to make this more consistent and more comprehensive.*

Point 8) Table 2: “Unless otherwise indicated, figures are yearly”. Data presented for SARS-CoV and MERS-CoV are not yearly. This should be corrected.

Norman, can you ask Eunhae to correct this?

Point 9) Page 3. It is never specified if human-virus PPIs from databases correspond to direct or indirect physical interactions. Authors need to discriminate these two situations in their review. Indeed, a vast majority of PPIs identified by co-affinity purification followed by MS identification correspond to large protein complexes and indirect interactions. Furthermore, this is directly related to a key question raised by the authors on page 3: “…we need to ask if it is biologically meaningful or even possible that a protein has >100 interactions. ». The answer is yes if interactions were identified by co-AP.

Norman? Can you pull this out of Intact?

Point 10) Table 3. DenHunt (http://proline.biochem.iisc.ernet.in/DenHunt/) or the most recent DenvInt (https://denvint.000webhostapp.com) are dedicated databases for dengue virus and should be included in Table 3.

*We have included these database in Table 3 now.*

Point 11) Page 3. “however small” should be deleted.

*This has been deleted in the revision.*

Point 12) Page 3. Authors cite their work on KSHV-human interactions as an example of “validation experiment”. This is irrelevant because interactions were predicted by computational analysis, and then tested by co-IP. This is not an example of PPIs identified by high-throughput screening, and validated with a secondary assay. Other works should be cited where large datasets of virus-host PPIs identified in a primary screen were validated using either biochemical (PMID:22898364; PMID:21994455 ; PMID:18985028 ; PMID:23816991) or functional (PMID:25464832) secondary assays. In addition, the overall quality of PPIs obtained by different methods or available in literature and databases has been evaluated (PMID:19060904; PMID:19116613). This should be discussed as the overall quality of virus-host PPIs is probably equivalent to host-host PPIs available in databases or literature.

We have added the studies this reviewer has suggested and have rewritten the statement on KSHV. [PU working on this …]

Point 13) “How many interactions does a virus require?”. It is surprising that authors used bacteriophages as a model for this estimation, as the review is essentially focusing on human viruses. There is no attempt to use the “large number” of host-interactions available for human viruses, as stated by the authors, to propose some estimation. Furthermore, the statement that “they are likely to contain a large number of false positives” is gratuitous and not supported. Finally, it should be mentioned that viral proteins are highly enriched for disordered regions that are known to favor interactions with many partners (PMID:19062293).

PU working on this …

Point 14) Page 5. « Of the over 17,000 HIV-1 – human PPIs reported in HIV-1db as of August, 2017, fewer than 7,000 are direct physical interactions. Many interactions are indirect, e.g. genetic interactions or inferred from mutagenesis experiments.». This is confusing because physical interactions can be direct or indirect (see Point 9). This needs to be rephrased.

PU can rephrase. Norman, can you look up direct vs. indirect numbers for HIV?

Precision is limited by the annotation of HIV-1db, which is based on text-mining. There are 129 interaction types in HIV-1db, denoted by phrases such as “binds”, “interacts with”, “complexes with”, etc that draw a connection between an HIV-1 and a human protein. However, we can give some upper and lower estimates for the true number of physical AND direct PPIs. Including only 16 interaction types – “binds”, as well as direct enzymatic activities like “phosphorylates”, “cleaves”, etc – cuts the total interactions from 17,103 to 1,592. **1,592 is therefore a lower estimate of the number of direct physical interactions.** Adding the keywords “interacts with” and “recruits”, “recruited by”, and “localized by” increased to what could be called a lenient estimate of 4,944 interactions. Probably many of the “interacts with” are not direct, however, we would like to note that the 4,944 does not include “complexes with” interaction types, so **it is in fact not known whether the interactions under “interacts with” are direct physical or indirect physical**. **The true number of direct physical interactions will lie between somewhere between 1,592 and 4,944.** Incorporating all interaction types that are physical, but omitting types such as “upregulates”, “down-regulates”, “co-localizes with”, etc. increases the number to 6,683. All numbers from August, 2017 HIV-1db.

We have modified the text to reflect the different numbers for direct physical (1,592), direct + unknown physical (4,944), or direct + indirect physical (6,683) interactions.

Point 15) Page 5. “Notably, we collected 5,495 interactions with human host proteins involving proteins of orthomyxoviridae.” Again, are these interactions essentially direct or indirect?

Norman, if this is from Intact, can you please double-check?

Point 16) Since Fig. 2A, 2C and 2D correspond to enrichment values, it is expected that values >1 correspond to enrichment, whereas values <1 correspond to an under representation. This is apparently not the case, and should be better explained in the figure legend. Did the authors used log-values in these three figures?

Stefan, can you take care of this?

Point 17) Page 7. “The virus interactome-diseasome connection”. Authors should discuss in this chapter the paper from Rozenblatt-Rosen O. & al. (PMID:22810586) where this interactome-diseasome connection is directly addressed.

PU working on this …

Point 18) Page 7. “Virus interactions with the host metabolome”. NS5A from HCV was shown to bind and to activate the hexokinase 2, thus increasing glucose uptake and glycolysis. This is an example of a direct interaction between a viral protein and a metabolic enzyme (PMID:24390321).

PU working on this … (maybe I can find a few other examples too)

Reviewer 4

While the purpose of this article is to provide a thorough overview about virus–host protein-protein interactions (PPIs) and on relevant sources of biological data for studying virus–host PPIs, I find the description of the various sections incomplete, general, confusing and sometime irrelevant, without providing the current state of knowledge. Overall I find that the article unstructured, inaccurate with description of studies that seem to be selected with a weak rational for expressing an opinion.

*We agree with the reviewer that our manuscript cannot realistically provide the complete current state of knowledge, given the vast body of literature on the subject. We have now clarified this by stating explicitly that we used existing databases as our starting point to give an overview, as opposed to summarize thousands of published papers which is practically impossible.*

*We also tried to restructure the manuscript to make it less confusing and removed some sections that may appear to be irrelevant.*

Few section are correctly emphasizing conclusion of previous review studies with limited added value. The section on bacteriophages as model is adding to the confusion of better defining the number of physiological interactions per proteins from data of large scale studies. Indeed, it may not be a good model for human virus as highlighted by the high false negative rate due to protein post-translational modifications and high protein maturation required in bacteriophages.

*Phage may not be good models of human viruses but they provide a paradigm using a much simpler system. We have now modified and slightly shortened this section but we have also added a statement that many human viruses undergo extensive proteolytic processing (e.g. most hepadnaviruses or flaviviruses).*

The section is closing with a statement that no comprehensive analysis or even a review of published interactions among human and virus proteins has attempted to evaluate these interactions for their plausibility or physiological relevance. Finally the conclusion and outlook are hardly based on the description of the article.

More specifically,

 The abstract as it stands is incomprehensible with many sentences that are meaningless.

*Without specific examples we are not sure which sentences this reviewer is referring to.*

 The Introduction section is quite irrelevant without providing any current state of knowledge for the different topics of the review and updated references. While the introduction about virus diversity makes the point that for some viruses like Zika there are no extensive data, Table 1 highlights Zika in the 20 best-studied viruses of the Flaviviridae (and how about HCV?).

*Table 1 is about genome sequences not PPI, which explains the discrepancy. Many well surveyed viruses have thousands of genome sequences but few PPI studies.*

The following sections:

Diversity of human viruses and Virus-host interactions in humans

- Unstructured and no rational for the description of the text based mainly on ref. 1

- Does not take into consideration blood borne viral infections

*We have covered many blood-borne viral infections, including HIV, Hepatitis, etc., hence we are not sure why this reviewer thinks we have not taken them into consideration.*

- Doubt on the accuracy of Tables 1, 2 and 4 (why is the number of interactions per individual viruses is often higher when added than the number in the column HPIs?)

*We thank the reviewer for pointing that out! There are indeed a few cases in which the sum of individual PPI numbers is higher than the total, namely that for adenoviridae, filoviridae, and phenuiviridae. This has now been corrected.*

- The numbers of PPIs are not up-to-date (for Flaviviridae it is much higher than 575 HPI)

Norman, can you double-check the database(s)?

- There is no description of the various orthogonal approaches or limitations of proteomics studies

*This issue has been discussed briefly in the section “How reliable are published virus-host interactions?” We have now expanded this section and added more detail about orthogonal approaches or limitations of proteomics studies.*

How reliable are published virus-host interactions?

Authors discuss the value of homologous interactions to validate 68% PPIs and describe one PPI study but claimed a biased validation of Y2H data set. It is not clear what the authors want to achieve in this section and their conclusion. HTS screens generate unlikely data of identified PPIs.

*We have* *discussed this in more detail now and provided an assessment of the value of HTS data, including a few examples.*