# Point-by-Point response letter for revision

**Reviewer 1**

-- Ref 1.1 –value of the paper--

|  |  |
| --- | --- |
| Reviewer  Comment | My biggest concern is that it is not readily apparent what the value of this publication actually is. Other than that overall comment, I was asked to evaluate the technical accuracy of the piece. |
| Author  Response | We thank the reviewer. Rapidly inncreasing reseachers participating in large scientific consortia, it is valuable to study the large scientific enterprise to benefit the whole commutity. In this manuscript, we studied the temporal structural dynamics of collaborative networks driven by large scientific consortia, and the researchers who play broker roles in collaborations. The value of our study is to give suggestions on how the large scientific consortia should be better oragnized to facilitate the collaborations. For example, a formal outreach group will very likely benefit the both inside and outside researchers of the consortium. |

-- Ref 1.2 –missing methods--

|  |  |
| --- | --- |
| Reviewer  Comment | I think there are some missing methods that need to be spelled out to understand fully what was done. In particular, what are the edges in fig. 1A? I think they are related to co-publications but this should be explicitly defined. What visualization software and layout was used? How were the modules defined? There is a paper reference given but there are surely parameters, etc., to be described. |
| Author  Response | Thank the reviewer’s questions. We added the information on network nodes, edges, modules, modularities and layouts in a supplement file as well as figure captions. |
| Excerpt From Revised Manuscript | We constructed temporal co-authorship networks for these two groups cumulatively over a decade from 2004 to 2014 (Fig. 1A, Supplemental network methods).  In Figure captions: ” Temporal co-authorship networks for ENCODE members (yellow, green) and non-members (red, dark-red) cumulatively from 2004 to 2014. To obtain the set of ENCODE members, we first obtained the set of authors, S\_1, who have co-authored at least one of the major ENCODE consortium papers. We also obtained the set of authors, S\_2, who have co-authored at least one paper in which the corresponding author was part of S\_1. The set of members is then defined as S\_1 ⋃ S\_2 . The non-members are thus defined as those who have co-authored papers using ENCODE data, but are not in the set of members. Nodes are authors who were connected by number of co-authored publications; i.e., edge weights. Green nodes are brokers in ENCODE members. Dark-red nodes are brokers in non-members. The networks were visualized using ‘igraph’ R package with the fruchterman reingold layout [9]…., We used the walktrap community algorithm to detect network modules in [9]…”  [9] <http://igraph.org>  Supplemental file for the network methods:  **Node** – Author  **Edge** – Number of publications that two nodes co-authored  **Consortium non-member** – Authors who have never co-authored in the consortium papers and the papers with corresponding authors from the consortium papers  **Consortium member** – Authors who are not non-members; i.e., authors who co-authored in at least either one consortium paper or one paper with corresponding authors from the consortium papers  **Layout** - We used the fruchterman reingold layout to visualize the networks by the R package, ‘igraph’ (http://igraph.org)  **Module** - We clustered the network modules using the walktrap community algorithm by the ‘igraph’ function, walktrap.community()  **Modularity** - We calculated the network modularity using the ‘igraph’ function, modularity().  **Broker** - Authors connecting many non-members to members in the 2014 network (>=40 non-members for ENCODE brokers; >=10 for modENCODE brokers) |

-- Ref 1.3 –modular structure not supported by figure--

|  |  |
| --- | --- |
| Reviewer  Comment | I don't understand the basis of the claim that "One can see how initially the consortium members coalesced into a tightly connected single module from 2004 to 2007 for the initial ENCODE publication and then broke up a little but still steadily retained a unified modular structure for their subsequent publication rollout in 2012." This sentence could be tightened up, but even by the loose statement made Fig 1b doesn't seem to support this claim as far as I can tell. |
| Author  Response | Thank the reviewer pointing out this confusion. The structures were supported by the modularity changes across years. We added text explaining how the modularity changed across years to indicate this type of network modular structure on Figure 1B and its caption. |
| Excerpt From Revised Manuscript | Figure 1B caption: “…The modularity dropped in 2007 because the first sets of ENCODE consortium papers were published in 2007 so that the members coalesced into a single module. The members still retained a unified modular structure shown as the relatively low modularity levels from 2007 to 2014, in contrast to non-member modularity….” |

-- Ref 1.4 –compare with random expectation--

|  |  |
| --- | --- |
| Reviewer  Comment | It would seem critical to understand how this analysis compares to random expectation? |
| Author  Response | This is a good suggestion. In the revised version of the manuscript we have tried to put in a number of random controls. In particular, we have made up some random networks composed of randomly selected biomedical researchers from Pubmed and found that their network modularity values keep high and flat over time (e.g., Fig 1B blue line).  What would figs 1a, b, c look like if we scrambled the four colors on the node labels?  What would figs 1a, b, c look like if we scrambled the four colors on the node labels?  Also, the reviewer here asks for a very specific type of randomization. We did exactly what the reviewer requested in the following Figure R1; i.e., we simply randomized between ENCODE members and non-members, and compared the modularity patterns with true networks. We can see the both randomly assigned members and non-members coalesced into a single group in 2007 but no longer see the modularity difference between the groups as we’d expect from the randomization; i.e., both of they have high modularity before 2007, a low modularity in 2007 due to the consortium papers published in 2007, and fast increasing modularity from 2007 to 2014.  It is not clear from the plots, i.e. Fig. 1c is circular in this regard in that the axes and the data colors draw from the same labeling.  It is not clear from the plots, i.e. Fig. 1c is circular in this regard in that the axes and the data colors draw from the same labeling.  We agree the review’s point about the way we’ve colored nodes in Figure 1C. The member/non-member and broker/non-broker coloring is according to the node’s axes positions, which is purely to help one’s eye.    Figure R1. The co-authorship network modularity for the made-up ENCODE consortium. blue – modularity of sub-network composed of randomly selected members; red – modularity of sub-network composed of others. The randomly assigned members and non-members coalesced into a single group in 2007 but no longer see the modularity difference between the groups as we’d expect from the randomization. |

-- Ref 1.5 –the hub and broker authors --

|  |  |
| --- | --- |
| Reviewer  Comment | Who are the main ENCODE hubs? Who are some examples of brokers? Would be interesting to know. |
| Author  Response | Thank the reviewer for the questions. We provided the co-authorship network and the broker list for the ENCODE in the supplemental files:  1. encode\_network\_2014.txt – the co-authorship network edge list in 2014  2. encode\_brokers\_2014.txt – the broker list |

-- Ref 1.6 –citation cleanup--

|  |  |
| --- | --- |
| Reviewer  Comment | Remove space between text and numbered citations -- residual from previous submission of this manuscript elsewhere. |
| Author  Response | Thank the reviewer. We have removed the spaces between text and citations. |

**Reviewer 2**

-- Ref 2.1 –definition of members --

|  |  |
| --- | --- |
| Reviewer  Comment | The paper by Gerstein et al. purports to examine different patterns of use of ENCODE consortium data by members and non-members of the ENCODE consortium, over time. However, there are some fundamental problems with the paper. Perhaps the most basic problem is that the paper is written as if membership in the consortium is fixed, and the only change over time is that new research is published by member or non-member groups. It seems that the authors´ definition of membership is ¨co-authored at least one publication funded directly by the consortium¨ but if this is the definition used, then someone who joined the consortium in 2012, is still considered to have been a member in 2007, even if at that point they were a non-member. (And membership in ENCODE is specifically designed not to be static: in fact, the specific instructions for how to join ENCODE are posted at http://www.genome.gov/12513439). Not considering membership as dynamic makes the paper circular at best: do people who are members co-author many collaborative papers with other members, or do people who are non-members who publish many collaborative papers with members then find it in their interest to apply for consortium membership? This is a similar problem to measuring stock market performance by the NASDAQ 100 index, when stocks are deliberately added and thrown out of the index based on their performance in the recent past. |
| Author  Response | The reviewer is making a reasonable point.  However, it is difficult to precisely track when the researchers joined and left the ENCODE because such information is a dynamic process and too complex to be precise.  At the other hand, the definition for non-members is much more clear; i.e., non-members are the researchers that have never coauthered the papers directly related to ENCODE. Thus, we redefined that the non-members of ENCODE are the authors who have never co-authored in the consortium papers and the papers with corresponding authors from the consortium papers, and rest of authors are ENCODE members (Supplemental network methods).  We realize this an approx…. we add some text to the paper  The new defintion still shows the similar network modularity dynamics with the old definition (Figure 1B); i.e., the consortium members initially coalesced into a tightly-connected single module from 2004 to 2007 for the initial ENCODE publication, and then broke up a little, but still steadily retained a unified modular structure until 2014 for their subsequent publication rollout in 2012. In the supplemental Figure S1, we keep the network modularity changes for the old definition where ENCODE members were the authors that have coauthored at least one paper funded by ENCODE. |
| Excerpt From Revised Manuscript | In figure captions, “… To obtain the set of ENCODE members, we first obtained the set of authors, , who have co-authored at least one of the major ENCODE consortium papers. We also obtained the set of authors, , who have co-authored at least one paper in which the corresponding author was part of . The set of members is then defined as . The non-members are thus defined as those who have co-authored papers using ENCODE data, but are not in the set of members…”  Supplemental file for the network methods:  **…**  **Consortium non-member** – Authors who have never co-authored in the consortium papers and the papers with corresponding authors from the consortium papers  **Consortium member** – Authors who are not non-members; i.e., authors who co-authored in at least either one consortium paper or one paper with corresponding authors from the consortium papers |

-- Ref 2.2 – compare with control --

|  |  |
| --- | --- |
| Reviewer  Comment | I don´t see an easy fix for this problem (unless it turns out that the set of people who joined ENCODE over time is very small and membership is indeed very stable; there is no way of determining this from the manuscript). But even if it were fixed, other questions about the manuscript remain. The authors make a lot of pretty pictures, and do not have a lot of conclusions, so it is hard to critique their conclusions, but it is at least clear that the authors are making some sort of claim that the clustering structure they observe has something to do with the existence of a consortium. But they examine one network only, so there is no control. One easy control that they should do at a minimum is the following: take 100 networks where in place of each consortium ¨member" node, you take another biomedical researcher at random with a similar h-index to that member in 2007 (before the start of the ENCODE project). Then look at the papers produced by these ¨members" over time  (starting in 2007), look at how citations to these papers diffuse over time to ¨members" and other ¨non-member¨ nodes. Do you see similar patterns to figure 1, or something quite different? Since there is no consortium here, hopefully there is a way to quantify the difference of what is observed in the control, random case. |
| Author  Response | This is a great suggestion. We constructed a random network as a control, which consisted of 438 randomly selected biomedical researchers (from 100 random biomedical PubMed papers) and calculated the network characteristics of their co-authorship networks from 2004 to 2014. We found that the random network always kept a high modularity across years (Figure 1B blue line), which is very different from ENCODE/modENCODE network characteristics. |
| Excerpt From Revised Manuscript | To evaluate our findings, we compared them to a random co-authorship network as a control, whose members are biomedical researchers randomly selected from Pubmed, and did not see that it has such network characteristics; i.e., it keeps very high modularity across years (Fig. 1B). |

-- Ref 2.3 – generalization of conclusion --

|  |  |
| --- | --- |
| Reviewer  Comment | The title, ¨Temporal dynamics of collaborative networks driven by large scientific consortia¨ is a bit grandiose, since the authors consider only one network, and can make no general conclusions. |
| Author  Response | This is a good point. To address this point, we studied another separate large scientific consortium, the Model Organism ENCyclopedia Of DNA Elements (modENCODE), and found that modENCODE had similar results to ENCODE even though the modENCODE consortium had independent memberships and publications from ENCODE. |
| Excerpt From Revised Manuscript | The third paragraph: “As a comparison, we also analyzed another separate large scientific consortium, the Model Organism ENCyclopedia Of DNA Elements (modENCODE), which studied the genomes of two model organisms, *D. melanogaster* and *C. elegans*. Our investigation of the modENCODE consortium had similar results to ENCODE even though the modENCODE consortium had independent memberships and publications. We identified 716 members and 959 non-members...” |

-- Ref 2.4 – conflict of interest --

|  |  |
| --- | --- |
| Reviewer  Comment | The corresponding author, Mark Gerstein, needs to disclose explicitly that he has received funding from the ENCODE project in the paper (as should any of his co-authors, if they have also received ENCODE funding). Even if ENCODE has not funded the present paper, it is well known that researchers are likely to think that funding programs that funded them personally are more likely to be good and wise programs, so the fact that he has received ENCODE funding is a potential source of bias, leading him to see the consortium as favorable. And there is even a potential for a real conflict of interest, should a paper such as this make it more likely that ENCODE funding would be renewed. |
| Author  Response | Thank the review for this fair point. Yes, we agree and are willing to disclose our roles in ENCODE consortium. We will consult the editor on what and where we should make this disclosure. |