Temporal dynamics of collaborative networks in large-scientific consortia

Daifeng Wang[1,2], Koon-Kiu Yan[1,2], Joel Rozowsky[1,2], Eric Pan[3], Mark Gerstein[1,2,3]*

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [3]Department of Computer Science, Yale University, New Haven, CT, USA. *Correspondence to: pi@gersteinlab.org

The emergence of collective creative enterprise is a unique feature in modern scientific research [1, 2]. Recent examples include large scientific consortia such as the international collaboration leading to the discovery of Higgs boson (the CMS and ATLAS consortia) [3, 4] and the ENCyclopedia Of DNA Elements (ENCODE) consortium annotating the human genome [5]. Though the scientific community should not be dominated by large projects, many fields in science benefit by such multi-investigator collaborative efforts. For instance, the 1000 Genomes consortium has generated an extensive amount of data and developed a catalog of uniformly called variants [6] for the biomedical community. To ensure that the scientific community can benefit from these efforts, it is important to understand the connections between consortium members and researchers outside of the consortium. To address the issue, we examined the ENCODE and modENCODE consortia as case studies.

Using publication data related to the ENCODE consortium [7], we identified 1,786 members and 8,263 non-members from 558 consortium papers supported by ENCODE funding and 702 community papers that used ENCODE data but were not supported by ENCODE funding (Fig. 1). We constructed temporal co-authorship networks for these two groups cumulatively over a decade from 2004 to 2014 (Fig. 1A, Supplemental network methods). The networks visualized how the information from the consortium has diffused out through the co-authorships relationships among specific individuals. Fig. 1B shows the number of co-authorship modules along with network modularity over time [8]. Based on this, one can see how initially the consortium members coalesced into a tightly connected single cluster 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. Conversely, the users of the ENCODE data and annotations (non-members) tended to form independent modules whose number was growing but without forming a unified structure. Of particular interest are a number of key individuals connecting many non-members to members (≥40) (Fig. 1C). These individuals serve as brokers between the consortium and outside researchers. 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 – in particular, it keeps very high modularity across years (Fig. 1B).

ROLLOUT

Daifeng Wang 1/25/16 11:02 AM
Deleted: 211 non-members (Fig.

Daifeng Wang 1/25/16 11:02 AM
Deleted:

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 revealed similar results to ENCODE even though the modENCODE consortium had independent membership and publications. In particular, we identified 716 members and 959 non-members from 162 consortium papers supported by modENCODE funding and 161 community papers that used modENCODE data but were not supported by modENCODE funding. We then constructed their temporal co-authorship networks cumulatively for the years from 2007 to 2014 (Fig. 2A). As before, the networks show how the information from the consortium diffused out through specific individuals. We found that the consortium had similar network characteristics as ENCODE's (Fig. 2B); i.e., initially, the consortium members formed a tightly connected single module in the first few years (2007 to 2010), and continued to maintain a generally unified modular structure in later years. On the other hand, the non-members tended to form independent modules whose numbers were increasing, but without forming a unified structure. We also found modENCODE brokers connecting no less than ten non-members between the consortium and outside researchers (Figs. 2C).

In summary, from the trends observed in both Fig. 1B and Fig. 2B, we can see the consortium structures from the publication patterns of individuals. Our analysis revealed that the consortium members work closely as a community whereas non-members collaborate on the scale of a few laboratories. We found that there are a few brokers playing an important role by initiating the connections between the consortium and non-members, thus we suggest that the large scientific consortia set up formal outreach groups or individuals to communicate with outside researchers. It is difficult to precisely track the timing when researchers are members or not, so we tried different definitions of membership, which gave very similar network characteristics (Supplemental files). In addition to the co-authorship networks, in future, ones can study the consortia impacts via other types of network connections such as citations [10].

Fig. 1. Visualization and analysis of co-authorship networks driven by ENCODE consortium. (**A**) 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 $S_1 \cup S_2$. 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]. (**B**) Number of co-authorship modules (circles + dashed line, right y-axis) and network modularity over time (circles + solid line, left y-axis) for temporal networks in Fig. 1A. The modularity

Daifeng Wang 1/25/16 11:02 AM
Deleted: and

Daifeng Wang 1/25/16 11:02 AM
Deleted: squares

Daifeng Wang 1/25/16 11:02 AM
Deleted: 2A

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. The random co-authorship network was constructed from 438 randomly selected biomedical researchers (from 100 random papers) and their co-authorship relationships in Pubmed in 2004-2014. We used the walktrap community algorithm to detect network modules in [9]. **(C)** Number of ENCODE member neighbors (y-axis) vs. the number of non-member neighbors (x-axis) for all authors up to 2014. Brokers have at least one ENCODE member neighbor and 40 non-member neighbors. There are 81 ENCODE member brokers (green) and 2417 non-member brokers (dark-red) in total. A node's size is equal to (1+2log(the sum of all node's edges weights)).

Daifeng Wang 1/25/16 11:02 AM
Deleted:

Daifeng Wang 1/25/16 11:02 AM
Deleted: (dark-red, green)

Fig. 2. Visualization and analysis of co-authorship networks driven by modENCODE consortium. (A) Temporal co-authorship networks for modENCODE members (yellow, green) and non-members (red, dark-red) cumulatively from 2007 to 2014. To get modENCODE members, we obtained the set of authors, S_1, who have co-authored at least one of the modENCODE consortium major papers published by the modENCODE consortium. 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 defined as $S_1 \cup S_2$. Nodes are authors connected by the number of co-authored publications; i.e., edge weights. Green nodes are brokers among the modENCODE members, and dark-red nodes are brokers among the non-members. The networks were visualized using 'igraph' R package with the fruchterman reingold layout [9]. **(B)** Number of co-authorship modules (squares + dashed line, right-y-axis) and network modularity over time (circles + solid line, left y-axis) for temporal networks in Fig. 2A. We used the walktrap community algorithm to detect network modules in [9]. **(C)** Number of modENCODE member neighbors (y-axis) vs. the number of non-member neighbors (x-axis) for all authors up to 2014. Brokers have at least one modENCODE member neighbor and 10 non-member neighbors. There are 45 modENCODE member brokers (green) and 184 non-member brokers (dark-red) in total. A node's size is equal to (1+2log(the sum of all node's edges weights)).

Daifeng Wang 1/25/16 11:02 AM
Deleted: 3A

Daifeng Wang 1/25/16 11:02 AM
Deleted: (dark-red, green)

References:

1 Guimera, R., *et al.* (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308, 697-702

2 Barabasi, A.L. (2005) Sociology. Network theory--the emergence of the creative enterprise. *Science* 308, 639-641

3 Collaboration, C.M.S. (2012) A new boson with a mass of 125 GeV observed with the CMS experiment at the Large Hadron Collider. *Science* 338, 1569-1575

4 Collaboration, A. (2012) A particle consistent with the Higgs boson observed with the ATLAS detector at the Large Hadron Collider. *Science* 338, 1576-1582

5 Consortium, E.P., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74

6 Khurana, E., *et al.* (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342, 1235587

7 ENCODE-related publication data are obtained from pages: http://genome.ucsc.edu/ENCODE/pubsEncode.html, http://encodeproject.org/ENCODE/pubsOther.html.

8 Clauset, A., *et al.* (2004) Finding community structure in very large networks. *Physical Review E* 70, 066111

9 http://igraph.org

10 Wang, F., *et al.* (2014) Towards a Scientific Impact Measuring Framework for Large Computing Facilities - a Case Study on XSEDE. In Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment (pp. 25:1-25:8). Atlanta, GA, USA: ACM.

Daifeng Wang 1/25/16 11:02 AM
Deleted: .