Temporal dynamics of collaborative networks driven by 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 creative enterprise is a unique feature in modern scientific research [1, 2]. Recent examples include the international collaboration leading to the discovery of Higgs boson [3, 4], and the ENCyclopedia Of DNA Elements (ENCODE) consortium aiming for annotating the human genome [5]. Though the scientific community should not be entirely dominated by consortium projects, most fields in science indeed are facilitated by such large collaborative efforts. For instance, the ENCODE consortium has generated an extensive amount of data and developed uniform annotations [5] for the genomics community. To ensure the scientific community can greatly benefit from various consortium 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 consortium as a case study.

Using publication data related to the ENCODE consortium [6], we identified 2940 members that have co-authored at least one publication funded directly by the consortium, and 2268 non-members that have used the data and annotation sets developed by the consortium in their publications. We constructed temporal co-authorship networks for ENCODE members and non-members cumulatively from 2004 to 2013 (Fig. 1A). The networks visualized how the information from the consortium has diffused out through specific individuals. Fig. 1B shows the number of co-authorship modules (right y-axis) along with network modularity over time (left y-axis) [7]. 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. 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 that joined at least one ENCODE member to many non-members (Fig. 1C). These individuals, having strong connectivity between members and non-members, serve as brokers between the consortium and outside researchers.

In summary, our analysis revealed that the ENCODE members work closely as a community whereas non-members collaborate in 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. From the trends observed in Fig. 1B, we believe that stronger links between two sides will be established in the near future. Large collaborative efforts and traditional collaborations will continue to complement each other, benefiting the scientific community as a whole.

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 2013. Nodes are authors who were connected by number of co-authored publications. Green nodes are brokers in ENCODE members. Dark-red nodes are brokers in non-members. (**B**) Number of co-authorship modules (dashed, right y-axis) and network modularity over time (solid, left y-axis) for temporal networks in Figure A. Note that the modularity, which is defined relative to the overall size of the network, decreases in 2007 even though the absolute number of modules increases. (**C**) Number of ENCODE member neighbors (y-axis) vs. number of non-member neighbors (x-axis) for all authors until 2013. Brokers (dark-red, green) have at least one ENCODE member neighbors and 30 non-member neighbors.

References:

1. Guimera, R., Uzzi, B., Spiro, J. & Amaral, L.A. Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697-702 (2005).

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

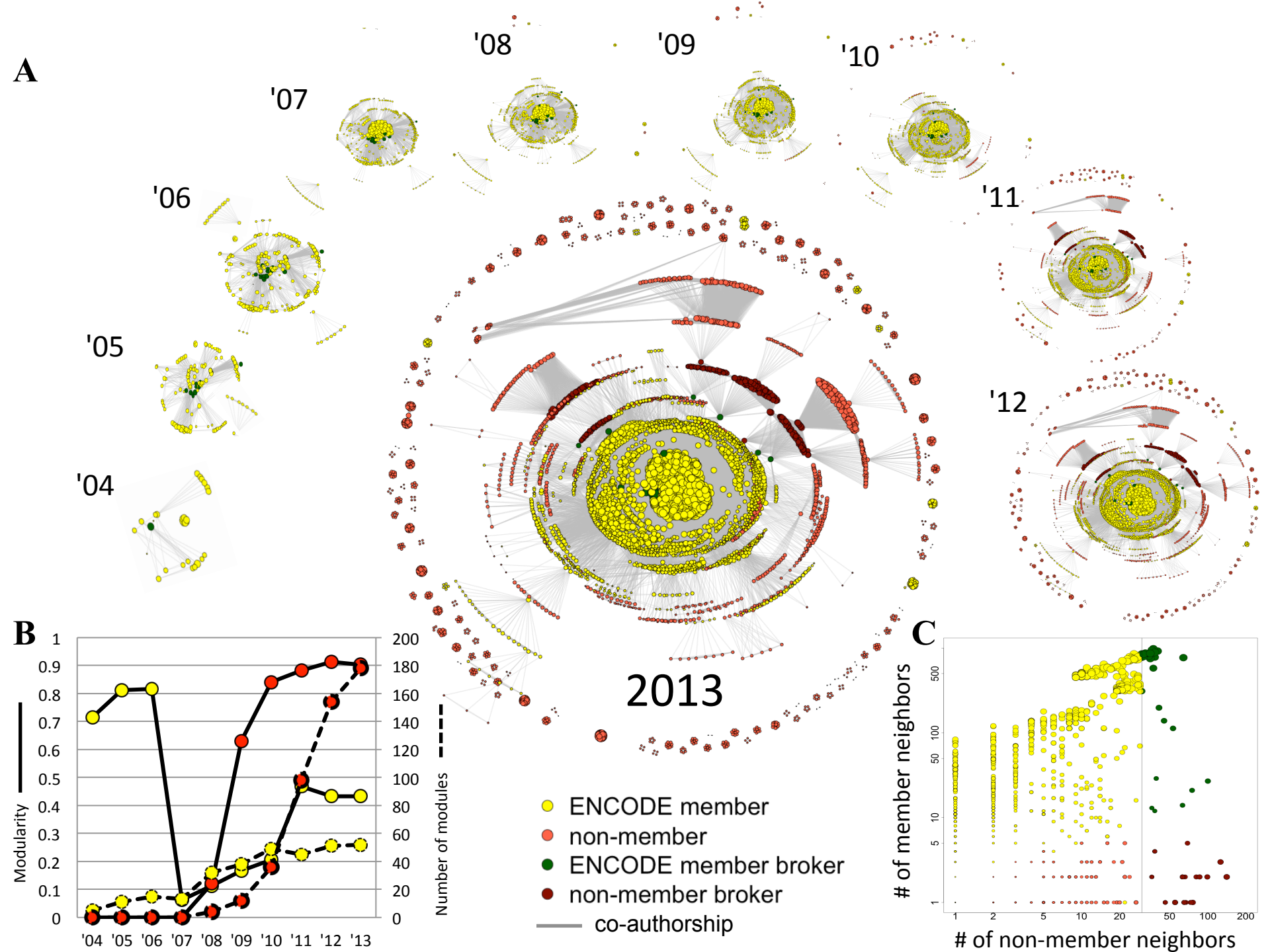
3. CMS Collaboration. A new boson with a mass of 125 GeV observed with the CMS experiment at the Large Hadron Collider. *Science* **338**, 1569-1575 (2012).

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

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

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

7. Clauset, A., Newman, M.E.J. & Moore, C. Finding community structure in very large networks. *Physical Review E* **70**, 066111 (2004).


A
'04
'05
'06
'07
'08
'09
'10
'11
'12
2013
B
Modularity
Number of modules
'04 '05 '06 '07 '08 '09 '10 '11 '12 '13
C
# of member neighbors
# of non-member neighbors
ENCODE member
non-member
ENCODE member broker
non-member broker
co-authorship