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]. For the published papers indexed by PubMed, the number of consortia-related papers have increased much faster than all papers in recent years (Fig. 1), which also implies the number of researchers have participated in scientific consortia is rapidly increasing, because the number of authors on consortia-related papers is typically much more than that of other papers. Recent scientific consortium examples include the international collaboration leading to the discovery of Higgs boson in CMS and ATLAS consortia [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, many fields in science indeed benefit 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 that 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 and modENCODE consortia as case studies.

Using publication data related to the ENCODE consortium [6], we identified 1,786 members and 8,211 non-members (Fig. 2). We constructed temporal co-authorship networks for ENCODE members and non-members cumulatively from 2004 to 2014 (Fig. 2A). The networks visualized how the information from the consortium has diffused out through specific individuals. Fig. 2B 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 (i.e., modularity dropped in 2007), and then broke up a little, but still steadily retained a unified modular structure till 2014 for their subsequent publication rollout in 2012 (i.e., low modularity after 2007). 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 (i.e., high modularity across years). Of particular interest are a number of key individuals that joined at least one ENCODE member to 40 non-members (Fig. 2C). These individuals, having strong connectivity between members and non-members, serve as brokers between the consortium and outside researchers. We didn't see that the random co-authorship network, whose members are the biomedical researcher randomly selected from Pubmed, has such network characteristics; i.e., it keeps very high modularity across years (Fig. 2B).

We also analyzed another 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. We identified 716 members and 959 non-members. We constructed temporal co-authorship networks for the modENCODE members and non-members cumulatively for the years from 2007 to 2014 (Fig. 3A). As before, the networks show how the information from the consortium diffused out through specific individuals. We found that the consortium has the similar network characterizes as ENCODE's (Fig. 3B); i.e., initially, the consortium members formed a tightly-connected single module in the years 2007 to 2010 in the first few years, 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 the modENCODE brokers between the consortium and outside researchers (Figs. 3C).

In summary, our analysis revealed that the consortium 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, thus we suggest that the large scientific consortia set up formal outreach groups or individuals to communicate with outside researchers. The outside researchers are encouraged to contact those outreach groups to find potential collaborators in consortia. The consortia members also should establish strong connections; e.g., together publishing consortia-related papers, to facilitate collaborations with both inside and outside communities. From the trends observed in both Fig. 2B and Fig. 3B, we can see the consortium structures from the publication patterns of individuals. Large collaborative efforts and traditional collaborations will continue to complement each other, benefiting the scientific community as a whole.

Fig. 1. Numbers of PubMed indexed papers from 1993-2013. The dashed curve with blue-diamond marks displays the numbers of consortia papers indexed by PubMed (i.e., the paper's author names appear "consortium" or "consortia") per year from 1993-2013. The solid curve with black-square marks displays the numbers of total papers indexed by PubMed per year from 1993-2013 in PubMed.

Fig. 2. 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. Green nodes are brokers in ENCODE members. Dark-red nodes are brokers in non-members. (**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. 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. **(C)** Number of ENCODE member neighbors (y-axis) vs. the number of non-member neighbors (x-axis) for all authors up to 2014. Brokers (dark-red, green) have at least one ENCODE member neighbor and 40 non-member neighbors.

Fig. 3. 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. Green nodes are brokers among the modENCODE members, and dark-red nodes are brokers among the non-members. **(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. 3A. **(C)** Number of modENCODE member neighbors (y-axis) vs. the number of non-member neighbors (x-axis) for all authors up to 2014. Brokers (dark-red, green) have at least one modENCODE member neighbor and 10 non-member neighbors.

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

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

Figure 1A

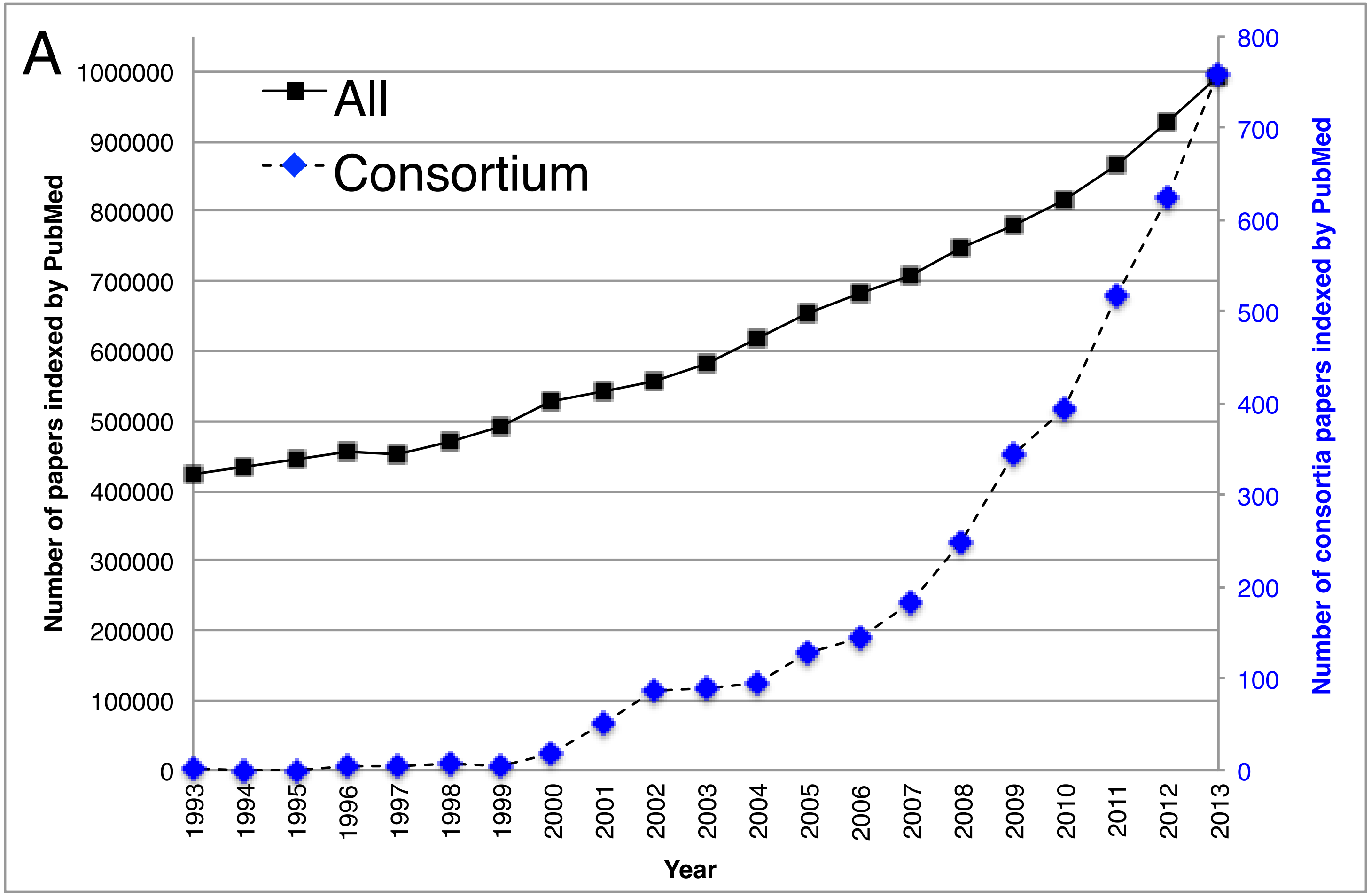

A
All
Consortium
Number of papers indexed by PubMed
1000000
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
Number of consortia papers indexed by PubMed
800
700
600
500
400
300
200
100
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Year

Figure 2A

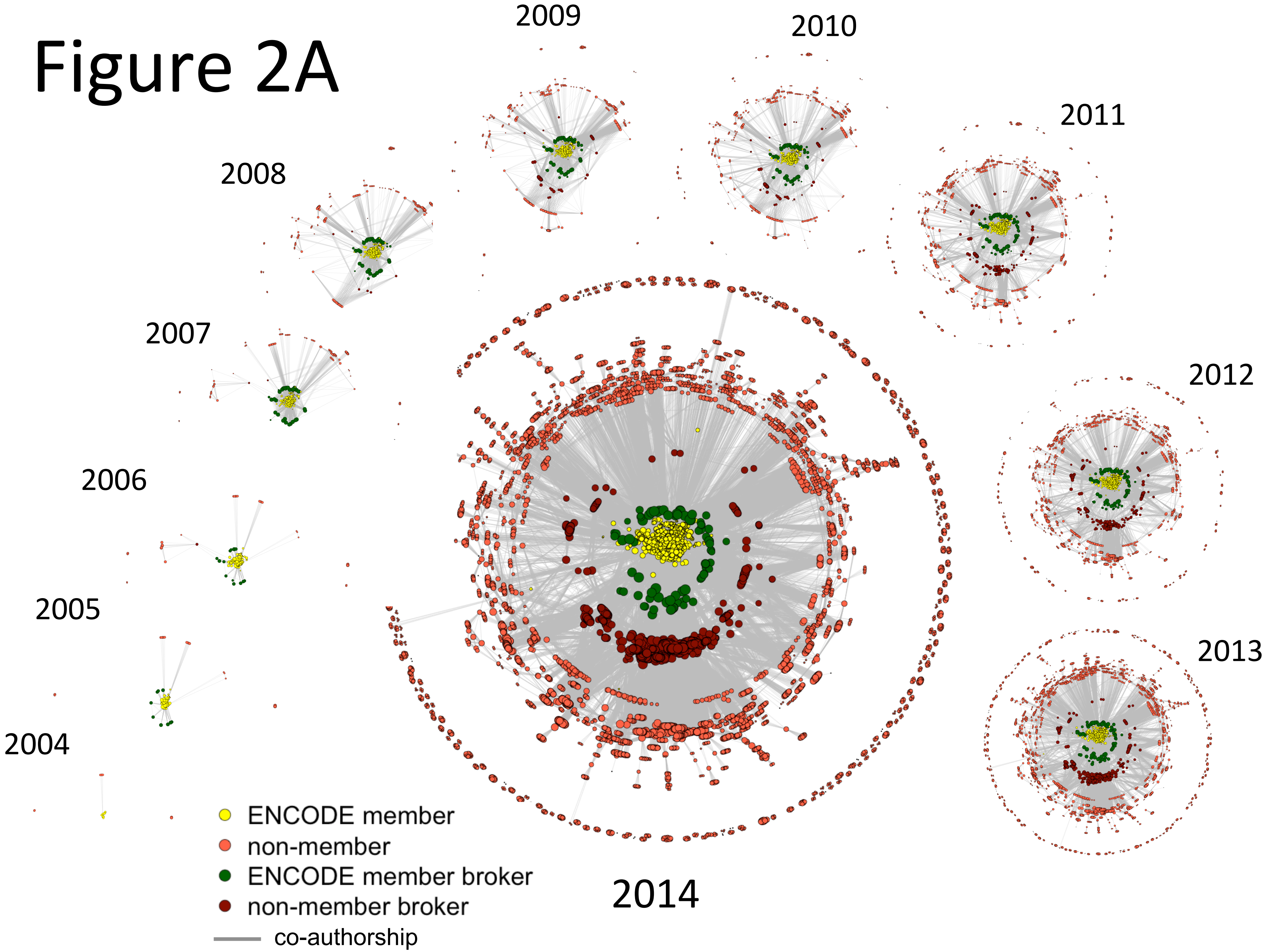

2009
2010
2008
2011
2007
2012
2006
2005
2013
2004
ENCODE member
non-member
ENCODE member broker
non-member broker
co-authorship
2014

Figure 2B

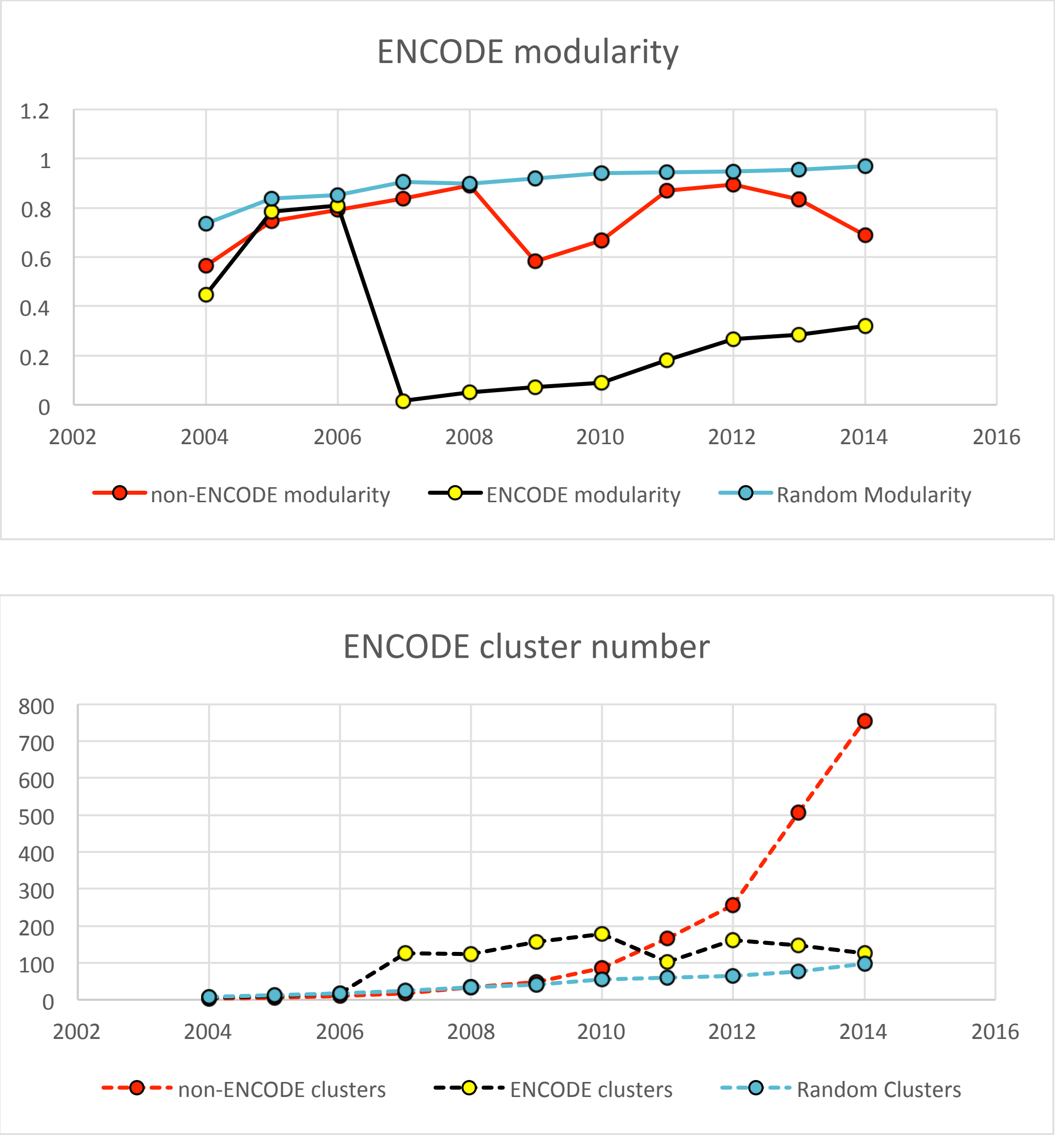
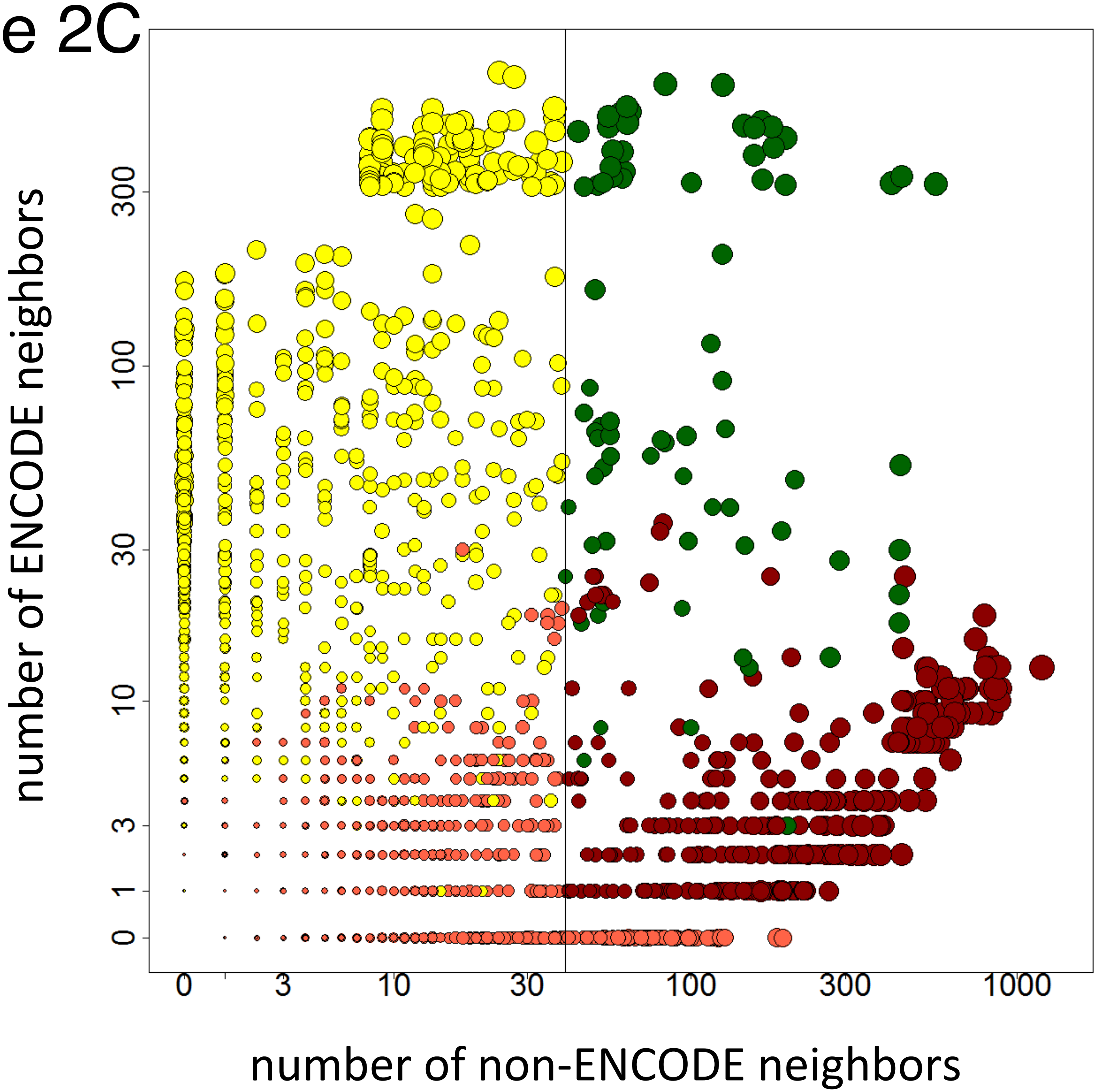

ENCODE modularity
1.2
1
0.8
0.6
0.4
0.2
0
2002
2004
2006
2008
2010
2012
2014
2016
non-ENCODE modularity
ENCODE modularity
Random Modularity
ENCODE cluster number
800
700
600
500
400
300
200
100
0
2002
2004
2006
2008
2010
2012
2014
2016
non-ENCODE clusters
ENCODE clusters
Random Clusters

Figure 2C


number of ENCODE neighbors
0
1
3
10
30
100
300
number of non-ENCODE neighbors
0
3
10
30
100
300
1000

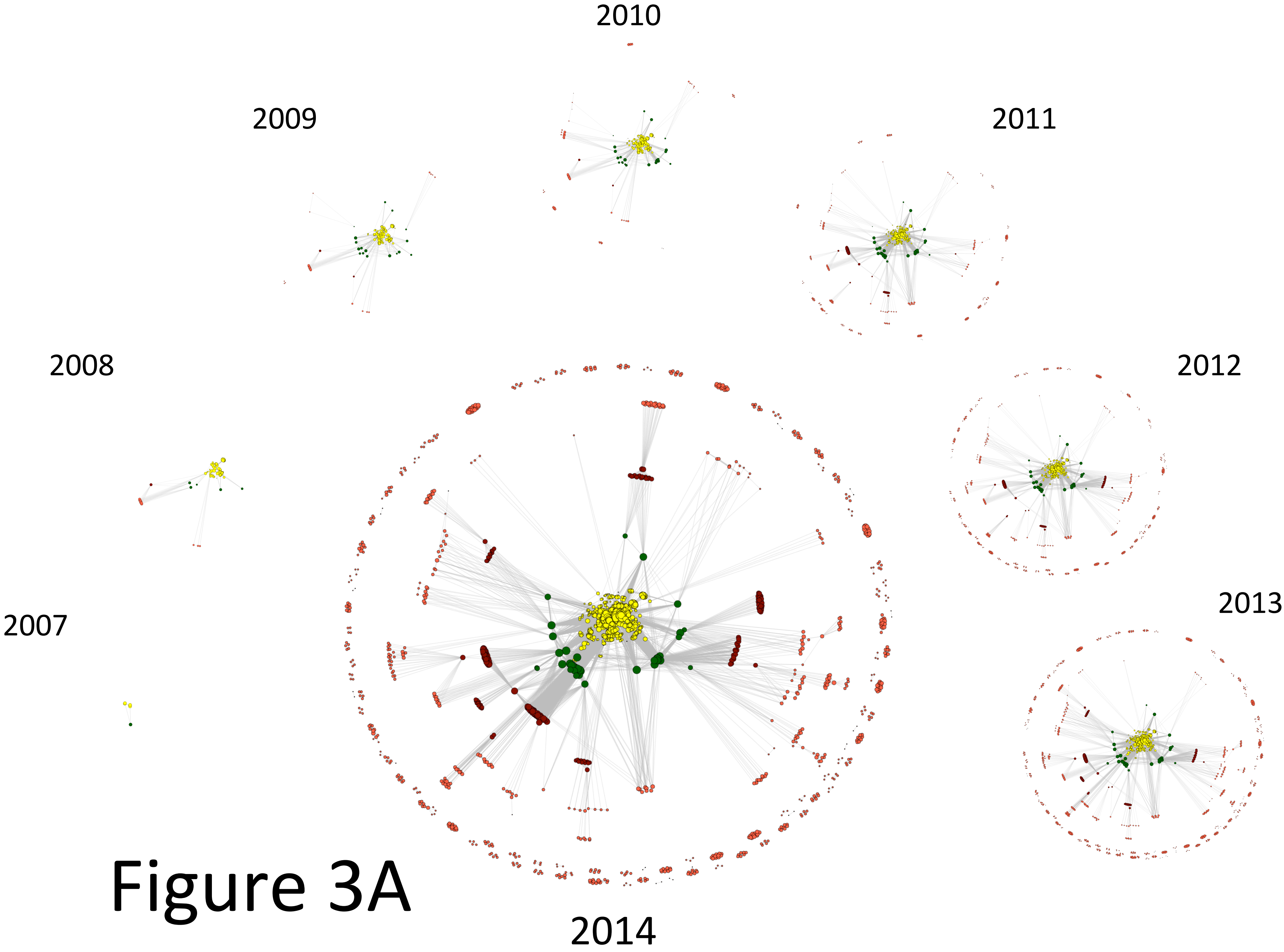
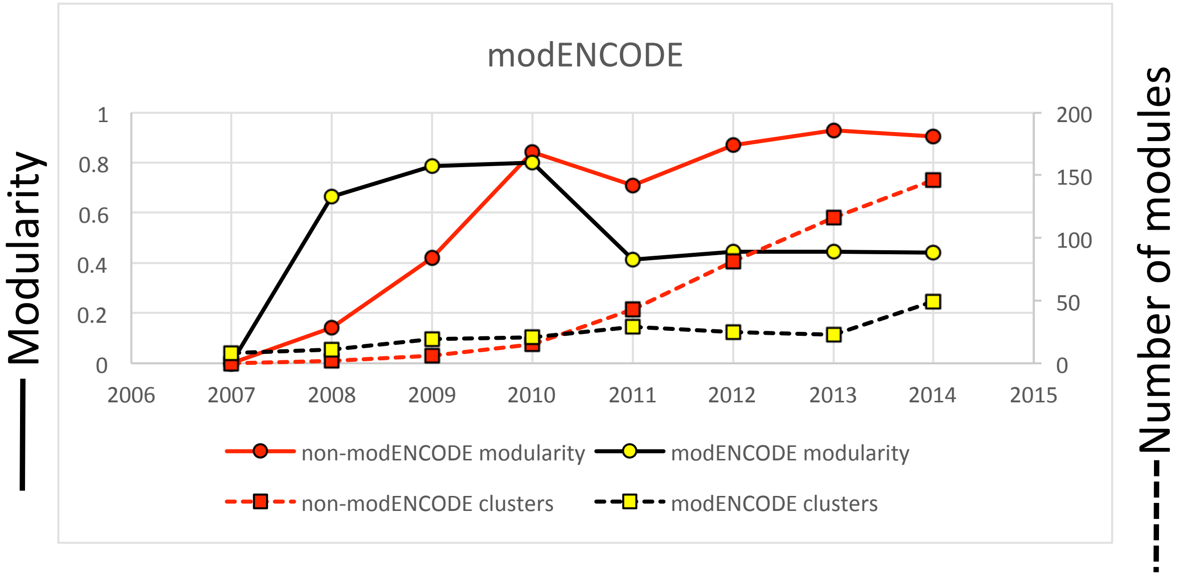


Figure 3A

Figure 3B


modENCODE
Modularity
Number of modules
1
0.8
0.6
0.4
0.2
0
200
150
100
50
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
non-modENCODE modularity
modENCODE modularity
non-modENCODE clusters
modENCODE clusters

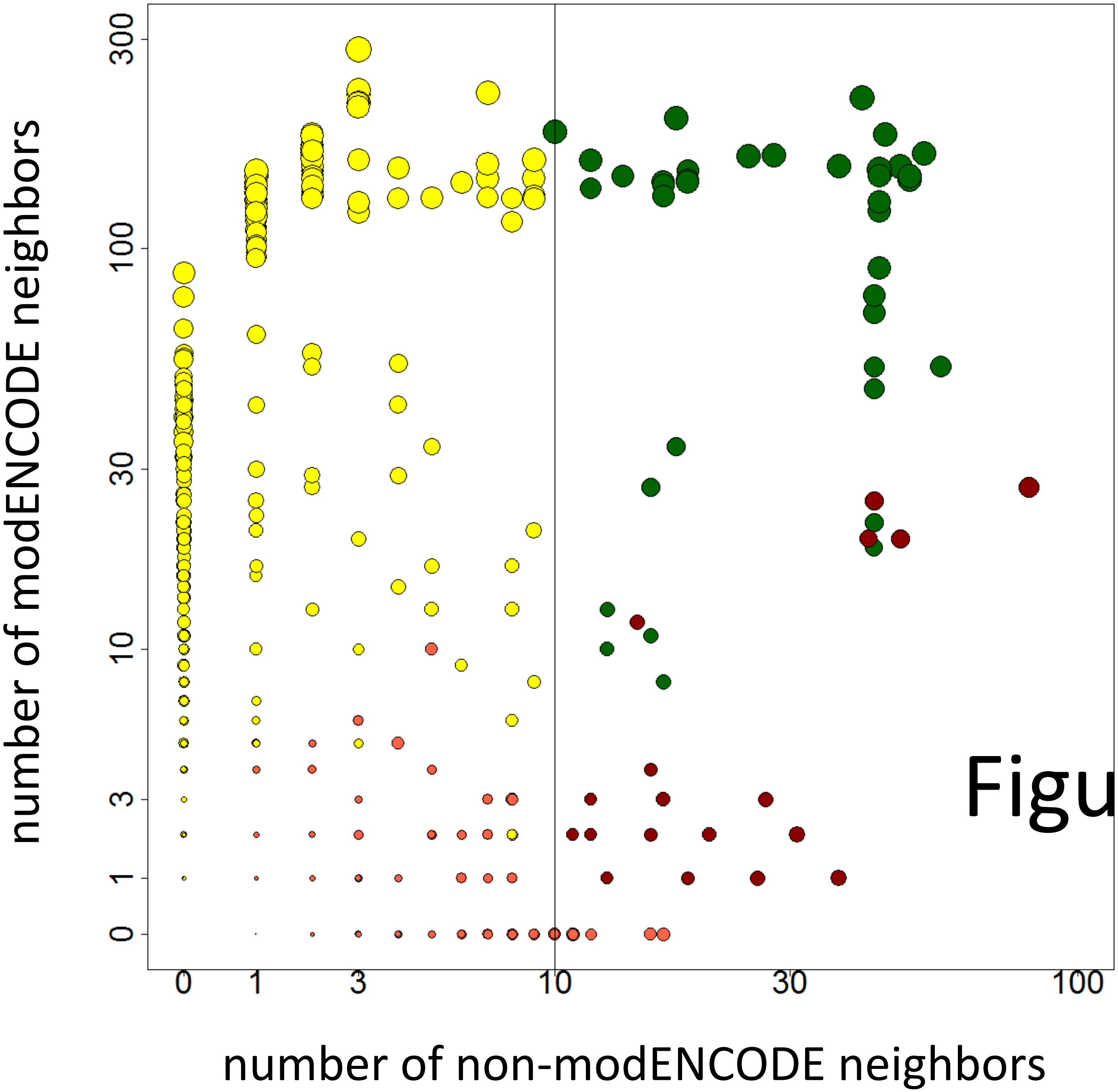

number of modENCODE neighbors
300
100
30
10
3
1
0
0
1
3
10
30
100
number of non-modENCODE neighbors

Figure 3C