Middle of the (by)line: Examining hyperauthorship networks in the Human Genome Project

Ly Dinh
School of Information Sciences, University of Illinois at Urbana-Champaign, USA. dinh4@illinois.edu

Yi-Yun Cheng
School of Information Sciences, University of Illinois at Urbana-Champaign, USA. yiyunyc2@illinois.edu

ABSTRACT

The era of big science promises rapid growth of scientific innovations and complex problem-solving, bringing forth the practice of doing science in large-scale collaborative effort rather than single author, solitary work. In disciplines such as genomics and high-energy physics, it is not uncommon that hyperauthorship phenomenon with the amount of authors soaring high from hundredth to thousandth. The purpose of this research is to explore the collaboration dynamics and the partial alphabetical author byline in one of the primary research article within the Human Genome Project (HGP). Using co-authorship network analysis, we find that middle authors play crucial roles in facilitating collaboration among previously unconnected authors as well as contributing to genetic sequencing efforts. Future work entails network analysis of all published works by HGP to comprehensively capture collaboration dynamics among multiple research centers.

KEYWORDS

Hyperauthorship, human genome project, co-authorship networks, author byline

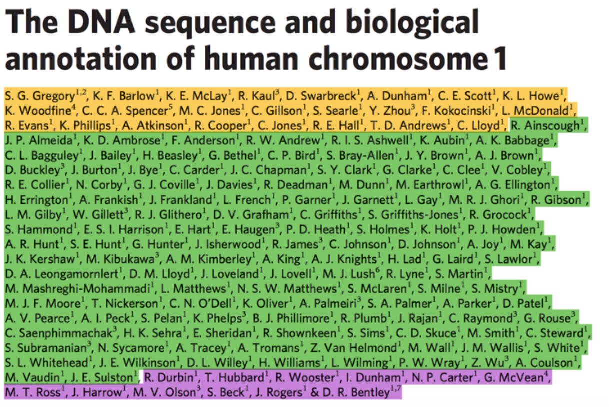
INTRODUCTION

As large-scale scientific endeavors such as the Human Genome Project (HGP) or the CERN Large Hadron Collider are on the rise, the number of authors associated with these research projects substantially increase as well (Ioannidis, 2008). In fact, co-authorship dynamics has evolved from multiple authors to ‘mega-’ or ‘hyper-’ amount of authors (Cronin, 2001), ranging from 80 to 200 authors per paper. Evidently, hyper-authored papers are more of a convention in scientific domains such as biology, high energy physics, and medicine that requires large and complex coordination of tasks (Birnholtz, 2006). Consistent with the current author-ordering tradition in the field of genomics, the research on chromosome sequencing often listed its authors in three parts: first, middle, and last. First authors are usually the main contributors, last authors are senior researchers who supervise the research, while middle authors usually take on tasks such as data collection or annotation, and are assumed to make smaller contributions (Mongeon et al, 2017). Given the complex contribution structures, there is a more pressing need to sort out questions of authorship order and credit allocations.

This study aims to examine the collaboration dynamics and the three-parts-structure of partial alphabetical author byline (some authors are listed alphabetically, some not) using co-authorship network analysis in the case of the HGP. In particular, we investigate the hyperauthorship phenomenon of one major research publication (Gregory et al, 2004) from the Wellcome Trust Sanger Institute, the biggest contributor to the sequencing of eight human chromosomes.

METHOD

With the HGP as a research setting and the Wellcome Trust Sanger Institute as the primary research center, we collect co-authorship data on articles published by the center. Given that the Sanger institute contributed efforts to sequencing chromosomes 1, 6, 9, 10, 13, 20, 22, and X, we only include published articles on the results of these eight chromosomes. The 8 papers identified were published between the years of 1999 to 2006, and were all highly-cited within the field. In this study, we focus specifically on the published work on chromosome 1 (Gregory et al, 2004), the largest human chromosome, thus requiring the most coordinated efforts to annotate and sequence.



The DNA sequence and biological annotation of human chromosome 1

S. G. Gregory[1,2], K. F. Barlow[1], K. E. McLay[1], R. Kaul[3], D. Swarbreck[1], A. Dunham[1], C. E. Scott[1], K. L. Howe[1], K. Woodfine[4], C. C. A. Spencer[5], M. C. Jones[1], C. Gillson[1], S. Searle[1], Y. Zhou[3], F. Kokocinski[1], L. McDonald[1], R. Evans[1], K. Phillips[1], A. Atkinson[1], R. Cooper[1], C. Jones[1], R. E. Hall[1], T. D. Andrews[1], C. Lloyd[1], R. Ainscough[1], J. P. Almeida[1], K. D. Ambrose[1], F. Anderson[1], R. W. Andrew[1], R. I. S. Ashwell[1], K. Aubin[1], A. K. Babbage[1], C. L. Bagguley[1], J. Bailey[1], H. Beasley[1], G. Bethel[1], C. P. Bird[1], S. Bray-Allen[1], J. Y. Brown[1], A. J. Brown[1], D. Buckley[3], J. Burton[1], J. Bye[1], C. Carder[1], J. C. Chapman[1], S. Y. Clark[1], G. Clarke[1], C. Clee[1], V. Cobley[1], R. E. Collier[1], N. Corby[1], G. J. Coville[1], J. Davies[1], R. Deadman[1], M. Dunn[1], M. Earthrowl[1], A. G. Ellington[1], H. Errington[1], A. Frankish[1], J. Frankland[1], L. French[1], P. Garner[1], J. Garnett[1], L. Gay[1], M. R. J. Ghori[1], R. Gibson[1], L. M. Gilby[1], W. Gillett[3], R. J. Glithero[1], D. V. Grafham[1], C. Griffiths[1], S. Griffiths-Jones[1], R. Grocock[1], S. Hammond[1], E. S. I. Harrison[1], E. Hart[1], E. Haugen[3], P. D. Heath[1], S. Holmes[1], K. Holt[1], P. J. Howden[1], A. R. Hunt[1], S. E. Hunt[1], G. Hunter[1], J. Isherwood[1], R. James[3], C. Johnson[1], D. Johnson[1], A. Joy[1], M. Kay[1], J. K. Kershaw[1], M. Kibukawa[3], A. M. Kimberley[1], A. King[1], A. J. Knights[1], H. Lad[1], G. Laird[1], S. Lawlor[1], D. A. Leongamornlert[1], D. M. Lloyd[1], J. Loveland[1], J. Lovell[1], M. J. Lush[6], R. Lyne[1], S. Martin[1], M. Mashreghi-Mohammadi[1], L. Matthews[1], N. S. W. Matthews[1], S. McLaren[1], S. Milne[1], S. Mistry[1], M. J. F. Moore[1], T. Nickerson[1], C. N. O’Dell[1], K. Oliver[1], A. Palmeiri[3], S. A. Palmer[1], A. Parker[1], D. Patel[1], A. V. Pearce[1], A. I. Peck[1], S. Pelan[1], K. Phelps[3], B. J. Phillimore[1], R. Plumb[1], J. Rajan[1], C. Raymond[3], G. Rouse[3], C. Saenphimmachak[3], H. K. Sehra[1], E. Sheridan[1], R. Shownkeen[1], S. Sims[1], C. D. Skuce[1], M. Smith[1], C. Steward[1], S. Subramanian[3], N. Sycamore[1], A. Tracey[1], A. Tromans[1], Z. Van Helmond[1], M. Wall[1], J. M. Wallis[1], S. White[1], S. L. Whitehead[1], J. E. Wilkinson[1], D. L. Willey[1], H. Williams[1], L. Wilming[1], P. W. Wray[1], Z. Wu[3], A. Coulson[1], M. Vaudin[1], J. E. Sulston[1], R. Durbin[1], T. Hubbard[1], R. Wooster[1], I. Dunham[1], N. P. Carter[1], G. McVean[4], M. T. Ross[1], J. Harrow[1], M. V. Olson[3], S. Beck[1], J. Rogers[1] & D. R. Bentley[1,7]

Figure 1. Three-parts author structure of chromosome 1 paper

The paper contains 166 authors, of which 24 are first authors, 127 are middle authors, and 15 are last authors (Figure 1). Data on authors’ affiliations, general demographics as well as the list of co-authors are obtained from the ISI Web of Science and Scopus. We conduct a co-authorship network analysis to capture collaboration patterns among all the authors involved in the chromosome 1 study. In this network, node i (ego) represents each author, and an edge a_{ij} between node i (ego) and j (alter) denotes that these two authors have worked on at least 5 papers together. The network is undirected because co-authorship is mutual between two researchers. We use R's *igraph* package as the tool for constructing the network matrix and calculating network measures. Visualizations of the whole network along with induced subgraphs are then created on *Gephi*.

RESULTS

The co-authorship network created from the dataset above is the result of combining 166 different ego-networks of 166 original authors listed on the chromosome 1 paper byline (Figure 2). An ego-network of each focal author (ego) is composed of the edges that link that author with all co-authors (alters) that he/she has worked with on at least 5 papers. The whole network includes not only ego-alter relationships, but also alter-alter relationships (co-authors tied to the focal author are also tied to each other). Hence, the network size is larger with 1918 nodes, and 14088 edges. At the whole network level, density is relatively low (ρ=.008), with a mean degree of 14.69 edges per author. Though the network is sparse, clustering coefficient is very high at (C=.843) and with a relatively short average path length (ℓ=3.538) given the large network size. Thus, the topology of this network exhibits more desired properties of a small world network, rather than a scale-free network as with most co-authorship networks (Barabasi & Albert, 1999). In fact, our degree distribution is skewed, but does not entirely follow a power-law distribution (α=.582, R^2= .431) of a typical scale-free network.

At the subgraph level, it is interesting that we detected 6 communities (using Girvan-Newman modularity) beyond the three-part author structure. Since our network boundary is beyond the main authors of the chromosome 1 paper, we can infer that there are more communities of authors that exist beyond this research and perhaps give a holistic view to the general collaboration structure of the HGP. More interestingly, we find that central actors in this network are primarily from the middle author list (e.g. Leongamonert, D; Hunt, S; Lloyd, D). In fact, 60-70% of the most influential nodes are from these middle authors, suggesting that they are actively working with other scholars (high degree), connecting scholars of different institutions together (high betweenness), and also directly working with other well-established scholars (high eigenvector).

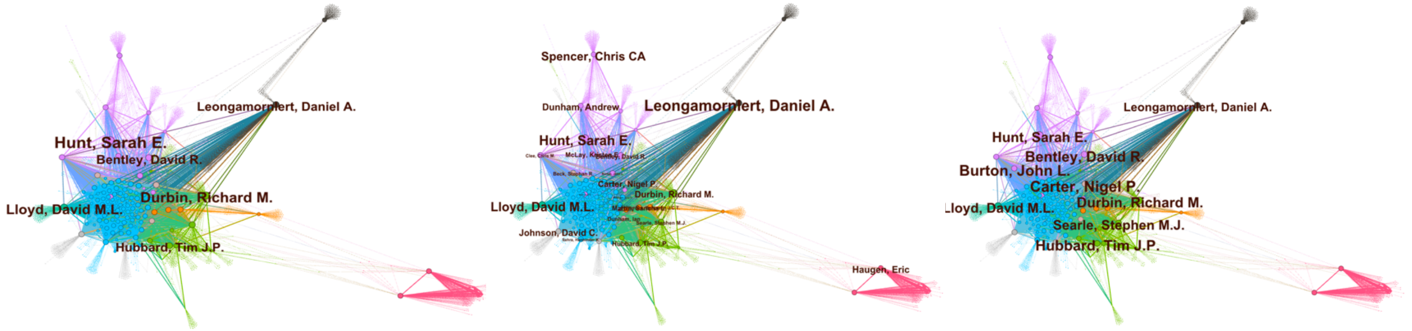


Figure 2. Top authors ranked by degree, betweenness, eigenvector centralities, respectively (by node label size)

CONCLUSION

The HGP is a leading example of big science that requires collaborative efforts from hundreds to thousands of researchers who are from a wide range of disciplines. Though determining where credit is due remains a conundrum in a hyper-authored setting, we find that middle authors may hold essential positions within the collaboration network in the chromosome 1 research. Recognizing that misallocation of credits may result in misallocation of funding opportunities and academic positions, we believe that providing a contribution list to describe *whodunwhat* in conjunction with the author byline is a necessary practice. Our future work includes examining in detail the collaboration dynamics among the first, middle, and last authors of all published works by a number of research centers affiliated with the HGP.

REFERENCES

Birnholtz, J. P. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. Journal of the Association for Information Science and Technology, 57(13), 1758-1770.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices?. Journal of the Association for Information Science and Technology, 52(7), 558-569.

Gregory, S. G., Barlow, K. F., McLay, K. E., Kaul, R., Swarbreck, D., Dunham, A., ... & Jones, M. C. (2006). The DNA sequence and biological annotation of human chromosome 1. Nature, 441(7091), 315.

Ioannidis, J. P. (2008). Measuring co-authorship and networking-adjusted scientific impact. PLoS One, 3(7).

Mongeon, P., Smith, E., Joyal, B., & Larivière, V. (2017). The rise of the middle author: Investigating collaboration and division of labor in biomedical research using partial alphabetical authorship. PloS one, 12(9).