Mining of methodological choices in phylogenetic workflows and why your supervisor may not have all the answers

James M. Eales[1], John W. Pinney[1], Robert D. Stevens[2] and David L. Robertson[1,*]

[1]Faculty of Life Sciences and [2]School of Computer Science, University of Manchester, Oxford Road, Manchester, United Kingdom

The methods we choose to use in a phylogenetic analysis can influence the results significantly. Furthermore the number of software implementations to choose between when designing an experiment can be overwhelming. When we also consider that our methods should be replicable as well as communicable, we are left with some significant difficulties to resolve. This work is aiming to provide best practice information on experimental design to the phylogenetic community as a whole. Using experimental workflow capture techniques that operated on the full text of 21,866 scientific articles we have extracted workflows that describe the full range of methodologies used by a large group of phylogenetic practitioners. The analysis of these articles made use of a naïve Bayesian text classifier for targeted text analysis as well as a semantically annotated controlled vocabulary to enable reconstruction of workflows from text. We have used similarity networks of the extracted workflows to explore the changes in community practice within the field over the last 10 years. We also constructed an author collaboration network for the identification of 'best' practice information on the choice of phylogenetic methods.

Our workflow network has shown significant variety in methodological choice. Furthermore the pattern of choices is field-specific in nature, with 3 clear fields of researchers: evolutionary biologists, microbiologists and virologists. Interestingly the level of field-specificity (as inferred from network assortativity calculations) has increased significantly through time with the practice of researchers from evolutionary biology being most distinct from the other 2 groups (Figure 1).

Using collaboration data we have identified the 'valued core' of researchers from the community. These authors have made methodological choices very similar to the rest of the community and to their specific field. We therefore conclude that the work of highly-active and collaborative authors in the field is not a good proxy for best practice, although it is well cited and respected.

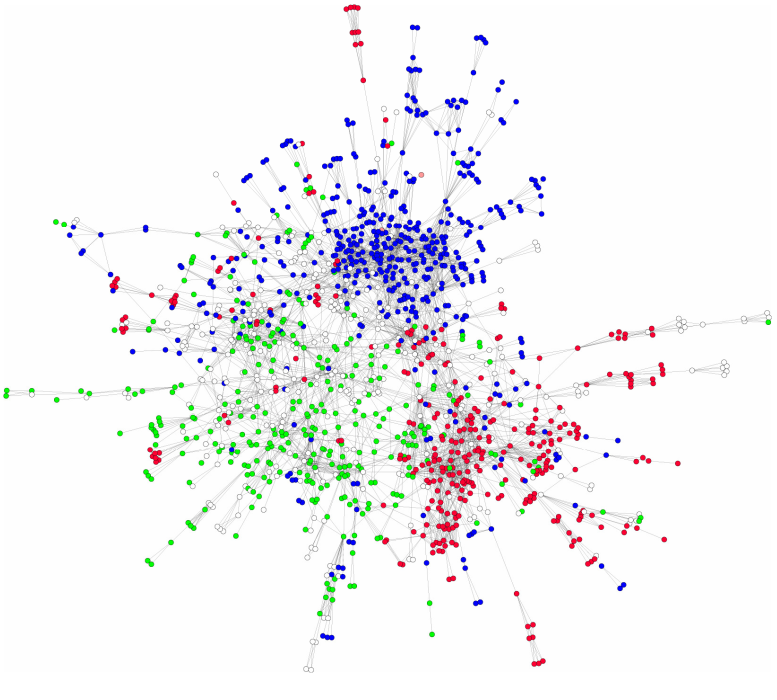

Figure 1. The 'valued core' of the collaboration network constructed from authors lists from 21,866 phylogenetics related articles. All authors who have collaborated with one or more other authors more than twice are represented as a node. Edges represent collaborations between authors. Nodes are coloured according to the field of the author. Blue represents microbiology/bacteriology, green evolutionary biology and red virology, white nodes represent authors who publish in more than one field.