Open Research Awards 2026

Health and Life Sciences disciplines prize

Sophia McDowell (Clinical and Biomedical Sciences)

Prioritising Reproducibility in Large-Scale Genomic Analyses

Introduction to the case study

My PhD in genetics and genomics utilises large datasets, such as the UK Biobank with genetic data for over 500,000 individuals. I am exploring the genetic basis of proteins and their relationships with reproductive health. Given the scale and the human impact of the data I am using, I have prioritised transparency and reproducibility throughout my work. It is currently quite common to share genetic summary data; however, sharing our code is less common. I believe that sharing how we carried out our work, alongside our data, with other researchers is an integral part of producing high-quality research.

Detail the practices you have used to support openness/reproducibility

When starting my current project, exploring the role of a protein (glycodelin) in reproductive health, I created an R project (.Rproj) in a private GitHub repository to contain all the information relevant to my work. I structured my project by creating separate folders for processed and raw data, scripts, and output. My scripts had descriptive names with numbered prefixes to organise my analyses. I described my repository structure, content, and software packages used in the repository README.

My scripts had detailed comments to guide others, and myself, through the code. I used R Markdown scripts, which were useful to add chunks of code between text, so I could describe my logic. When updating my work, I committed my changes frequently, using short messages to describe each update.

Upon finishing my analyses, I uploaded my manuscript to a preprint server (medRxiv) and submitted it to a journal with the Gold Open Access publishing option for peer review. In both my pre-print and the journal article (currently under review), I have

included Data and Code availability statements and added ORCIDs for transparency. Finally, I added an MIT license and made my repository publicly available.

Outline any barriers or challenges you encountered, and how they were handled

The UK Biobank has strict data protection policies, and access is only granted after a successful application. Sensitive data is restricted to their platform, and only non-identifiable data is downloadable. I used gitignore protocols to ensure sensitive data were not committed to my repository automatically. Working across the UK Biobank platform, university servers and my local environment involved making sure my repository was accessible from all environments. Another challenge I faced was worrying about creating the 'perfect' system from the start. I overcame this by adapting my approach as I learned about reproducibility and open research practices.

What benefits have you identified from your case?

Sharing my code creates lots of opportunities for feedback, and for any errors to be picked up, which increases my confidence in my work. Throughout my PhD, I have used many publicly available resources, and I am very grateful to people who share their data, as it allowed me to conduct my own research and learn a lot from it. I am excited to be able to contribute my own resources, to hopefully help others in the same way, and keep driving progress in my field.

How do you plan to continue this case moving forward?

Upon publication of my work, I plan to upload my summary statistics to the GWAS Catalog, which is a large database that allows the sharing of summary genetic data. I would also like to employ other methods, e.g. using R environment manager (renv) to maintain a persistent snapshot of the project environment, ensuring my results remain replicable as R packages evolve and different versions are released. I have implemented the practices described in this case study in my other PhD projects, and plan to continue improving my approaches, alongside sharing this knowledge with peers, in my future work.

Conclusions

Writing this case study has been an enjoyable and satisfying validation of my progress. I recognise that adopting transparent practices requires constant learning and adaptation, and the importance of prioritising accessibility and reproducibility cannot

be overstated. I am looking forward to continuing my progress throughout the remainder of my PhD.

Links to any materials/resources relevant to the case

- **Preprint (medRxiv):** https://doi.org/10.64898/2026.03.06.26347586
- **GitHub repository:** https://github.com/sophiaq229/glycodelin_female_reproductive_traits