# Open Research Awards 2025

**HLS disciplines prize**
**Bethan Rimmer (Department of Clinical and Biomedical Sciences)**

**Title of case study**
**Re-annotating the EPICv2 manifest**

Introduction to the case study
The Illumina EPICv2 array is a technology which reads chemical marks (DNA methylation) at >950,000 sites on human DNA. A file provided by Illumina, the EPICv2 manifest, labels these sites with crucial information for researchers, e.g. which genes they are located in. After the Exeter Epigenetics group noticed the manifest was frequently inaccurate, I corrected/re-annotated the manifest using publicly-available data. The new file is available to download online; details of the correction/update are described in a preprint; and I am collaborating with researchers internationally to integrate the new manifest into existing analysis software, and make updates in response to feedback.

Detail the practices you have used to support openness/reproducibility
The new manifest was intended as an open-source tool to benefit the research community because inaccuracies in the original manifest pose a major challenge to researchers interpreting EPICv2 study results.

During development, I integrated feedback from another PhD student using EPICv2 data, my supervisor, and our lab lead, and collaborated with a postdoctoral researcher at Exeter who added part of the re-annotation.

I wrote a preprint describing: the method/protocol used for re-annotation, how the new manifest compares to the original, and a guide to the manifest file contents.

The re-annotated EPICv2 manifest has been uploaded publicly to Zenodo and currently has >100 downloads. In keeping with open-source principles, I included a version that compares our manifest to Illumina's allowing anyone to reproduce the findings included in the preprint.

The file and preprint were announced at Epigenomics of Common Disease conference 2025. I answered questions from other researchers and an Illumina representative, and started a collaboration with Dr Scott Waterfield to integrate the file into his widely-used DNA methylation analysis software wateRmelon.

This has led to an international collaboration with Professor Andrew Teschendorff, to adapt his existing software EpiSCORE for EPICv2 data, which involves updating the manifest in response to feedback.

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

Annotating the >950,000 sites posed multiple challenges. This included deciding how to annotate sites located in multiple overlapping genes. Additionally, certain features (such as introns - regions of genes which aren't read/expressed by the cell) were not present in the public reference data so I had to calculate their positions. Processing this amount of data was computationally demanding (the code I wrote to re-annotate the manifest was originally predicted to take ~4 days to run!) so I had to learn new methods to optimise efficiency, reducing the code runtime to 2 hours.

What benefits have you identified from your case?

The original Illumina manifest annotated ~215,000 out of ~950,000 sites read by the EPICv2 array as being in a gene, whereas the updated manifest labels ~700,400 sites as in a gene, a huge increase. This additional information will be useful to researchers using EPICv2 data, for identifying genes and biological processes prevalent in study results. Making the file open-source and reproducible supports transparency and accessibility. Its uptake, evidenced by >100 downloads and integration into widely-used software, demonstrates its practical value. Overall the new manifest could facilitate analysis for researchers worldwide using EPICv2 data.

How do you plan to continue this case moving forward?

The collaboration with Andrew Teschendorff, has resulted in updates to the manifest, including adding new features and adding more detail to already-annotated features. I plan to make this update and all future updates open-source via Zenodo. I also plan to update the preprint associated with the new manifest with details on any changes, and to submit the preprint to a journal (either Epigenetics or Bioinformatics). I will validate which features labelled in the new manifest are affecting gene expression, using two matched datasets (including EPICv2 DNA methylation data, and RNA data, from the same samples).

Conclusions

The re-annotated EPICv2 manifest significantly improves gene annotation accuracy, increasing coverage from ~215,000 to ~700,400 sites. It is open-source and already being integrated into key software, addressing a major barrier for researchers, supporting more reliable DNA methylation analysis and fostering international collaboration and reproducibility in epigenetics research.

Links to any materials/resources relevant to the case

Publicly uploaded re-annotated manifest file: https://zenodo.org/uploads/14933469
Associated preprint on BioRxiv: https://www.biorxiv.org/content/10.1101/2025.03.12.642895v2