Abstract

Multi-omic analyses are expensive and often capture a mere fraction of what is measurable. Imputation panels exist for genetic data but not for other omics types, such as the circulating proteome.

This project has 3 key aims: 1) using a variety of AI/ML approaches, determine which proteins can be robustly imputed; 2) determine if the newly imputed measures offer novel insights through genome-wide association studies or incident disease analyses; 3) build a front-end imputation server or R package to facilitate dataset augmentation for the global research community.

To do this, we will analyse data from UK Biobank, which is currently the largest proteomics resource in the world with 3,000 Olink proteins assessed on 50,000 individuals. This project will have an impact for the global research community by potentially saving cohorts £millions from generating new data and will also lead to downstream discoveries in biomedical science.

Intro

How much about human health can we learn from a blood sample? The creation of mega-cohorts, such as UK Biobank have enabled us to develop risk signatures for dozens of disease outcomes [1]. This is reflected through individual proteins as well as multi-protein signatures [1].

Omics measurement technologies have improved over time, meaning that we can now assess upwards of 11,000 proteins per sample. However, many cohorts have profiled measurements on older technologies – due to tech limitations at the point of analysis – or for fewer proteins due to cost implications. Given the highly correlated structure of proteins, it may be possible to create an imputation framework that helps to 1) augment historic datasets, leading to better powered meta-analyses; 2) de-noise protein measurement outliers, leading to improved inference and 3) design sparse protein panels that optimise the trade-off between cost and content.

Recently, we showed that imputation of GDF15, a marker of inflammation and multimorbidity [1], can be successfully imputed based on other protein measures (r=0.87) [2]. In a test set from UK Biobank, the imputed protein showed a slightly stronger and more statistically significant

association with incident dementia (Hazard Ratios of 1.42 versus 1.37 per SD change in the imputed and measured proteins, P = $3x10^{-15}$ and $2.5x10^{-17}$, respectively). This analysis is a proof of principle, using a basic elastic net approach to train the predictor. Here, we will extend the approach across all 3,000 proteins available in UK Biobank, using a variety of statistical methods to predict each protein.

Research Challenge

There are four key challenges for this project:

1. Which proteins can be accurately imputed?
2. What methods work best for this imputation both in terms of performance and scalability?
3. Do the imputed proteins recapitulate findings between the observed values and health/disease outcomes?
4. Can we build an efficient, user-friendly and GDPR-compliant front-end server (or R package) to enable users to safely and robustly impute proteins?

Data & Methodology

Data

The project will utilise what is currently the world's largest blood-based proteomic resource, UK Biobank. Around 3,000 proteins have been assessed in 50,000 individuals between 2006 and 2010 when they were aged between 40 and 70 years old. The volunteers provided consent for the integration of electronic health records, both before and after the blood draw, including GP and hospital codes.

Protein Imputation

Multiple approaches will be taken to maximise the accuracy and scalability of the protein imputation process. This could include filtering the feature set to proteins housed on a specific panel (https://olink.com/products/compare). We will split the cohort into training and test sets before benchmarking the predictions via penalised linear regression. Non-linear and more computationally expensive approaches can also be explored. We will also examine sex-specific effects and differences across age groups.

Correlation coefficient and RMSE will be used to assess predictive performance. We will also compare the performance of the imputed versus observed proteins in association analyses with health outcomes e.g., how well do our imputed proteins recapitulate associations with complex traits or do well-imputed proteins relate to certain diseases/lifestyle factors/biological pathways?

Finally, we will develop an online server and user-friendly front-end where researchers can securely upload their DNAm datasets for the imputation of new CpG content.

RRI/Ethical considerations

All data have been generated under existing ethics approvals.

Expected outcome and impact

This is a high impact project. Protein imputation approaches do not currently exist. Furthermore, designing a parsimonious panel of proteins could offer major savings to cohort studies, enabling them to profile additional samples. The de-noising of data will also boost the power to discover novel findings within existing datasets, such as UK Biobank.

References

1. Gadd DA, Hillary RF, Kuncheva Z, Mangelis T, Cheng Y, Dissanayake M, Admanit R, Gagnon J, Lin T, Ferber KL, Runz H; Biogen Biobank Team; **Foley CN*, Marioni RE*, Sun BB***. (2024), Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat Aging*. 4(7):939-948. *equal contribution
2. Richmond A, Mur J, **Harris SE**, Corley J, Elliott HR, **Foley CN**, Hannon E, **Kuncheva Z**, Min JL, Moqri M, Ndiaye M, **Sun BB**, Vallejos CA, Ying K, Gladyshev VN, **Cox SR, McCartney DL, Marioni RE**. (2024). Imputed DNA methylation outperforms measured loci associations with smoking and chronological age. *BioRxiv*. doi: https://doi.org/10.1101/2024.09.05.611501