Description (1000 words)

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 DNA methylation (DNAm).

This project has 3 key aims: 1) using a variety of AI/ML approaches, determine which DNAm sites can be imputed across arrays, tissues and ancestries; 2) determine if these sites are enriched for functional characteristics (e.g., genomic location) or disease and lifestyle associations; 3) build a front-end imputation server to facilitate dataset augmentation for the global research community.

To do this, we will analyse data from what is currently the world's largest published blood-based DNAm dataset (Generation Scotland, n_{pepole}=19,000, n_{CpGs}=800,000) with an ongoing recruitment wave representing the world's largest saliva-based DNAm dataset (current n_{people}=10,000, n_{CpGs}=270,000 – the newest array). We will also use our extensive network of inter/national collaborators to test our pipelines.

This project will have an impact for the global research community by saving cohorts £millions from generating new data that will also lead to downstream discoveries in biomedical science.

Introduction

Genetic variation in the human genome, specifically single nucleotide polymorphisms (SNPs), underpins phenotypic variation in complex traits. SNPs that lie close to each other tend to be correlated. Therefore, while it is now common to profile ~1 million SNPs via inexpensive array technology, their correlation structure, in tandem with reference panels for the entire human genome, mean that upwards of 30 million SNPs can be imputed. Publicly-available web tools [1] implement this, even for older array data where only a few hundred thousand SNPs were directly measured.

Currently, there is no analogue for other commonly assessed omics layers, such as the methylome. DNA methylation is a dynamic, chemical (epigenetic) modification to DNA that is tissue- and cell-type-specific. It can be thought of as acting like a dimmer switch, turning genes up/down. This results in protein production and downstream health outcomes.

Methylation patterns are influenced by genetic and environmental factors. Regarding the latter, we have developed predictive signatures that accurately characterise smoking and drinking behaviours, predict chronological age and augment clinical predictors of future disease risk [2-9].

There is an extremely strong correlation structure (including long-range patterns) across the methylome. Presently, over 850,000 methylation marks (termed CpG sites) can be profiled on the market-leading array. Older versions of the array, which have been generated on tens of thousands of individuals from dozens of studies, characterise as few as 27,000 CpGs. Nearly all of these sites are contained on the larger arrays. However, no imputation panels exist to retrofit existing datasets. The latest market offering from the array-provider contains ~145,000 sites that overlap with previous arrays in addition to new content for ~125,000 sites thought to be functionally relevant to gene regulation. These arrays are up to 6-fold more expensive to run than SNP-arrays.

Research Challenge

There are 5 key challenges for this project:

1. Can DNAm content be accurately imputed?
2. What methods work best for this imputation both in terms of performance and scalability?
3. Do these methods extend across tissues, specifically blood and saliva-based DNAm?
4. Do the imputed DNAm estimates recapitulate findings between the observed values and health/disease outcomes?
5. Can we build an efficient, user-friendly and GDPR-compliant front-end server to enable users to safely and robustly impute DNAm data?

Data & Methodology

Data

The project will utilise what is currently the world's largest blood-based DNA methylation resource, Generation Scotland. DNAm has been assessed at ~800,000 CpG sites in over 19,000 individuals [2]. These volunteers also have extensive health questionnaire data recorded at the time of blood draw (2006 to 2011 when volunteers were aged 18 to 99 years). Furthermore, the volunteers provided consent for the integration of electronic health records, both before and after the blood draw, including GP and hospital codes. Recently, a new wave of recruitment began in Generation Scotland (2022 – present with volunteers aged 12 to 99 years) [10]. This time, online questionnaires were supplemented with a postal 'at-home' spit kit from which salivary DNA has been extracted and profiled for genotype and DNAm (~270,000 CpG sites in 10,000 individuals).

CpG Imputation

Multiple approaches will be taken to maximise the accuracy and scalability of the CpG imputation process. We have shown that in ultra-high dimension contexts that feature pre-selection improves predictive accuracy [11]. Here, this could include filtering to CpGs: on the same chromosome as the target CpG; with high variance; or to subsets known to track common influences on the blood methylome, such as age, sex, smoking, white cell proportions. 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.

Correlation coefficient will be used to assess predictive performance. We will also compare the performance of the imputed versus observed CpGs in association analyses with health outcomes e.g., how well do our imputed CpGs recapitulate associations with complex traits. To account for possible batch effects between train and test sets, we will also standardise data via rank inverse Normal transformation (within sets) as well as running the analyses on the original scale data. Given that methylation is dynamic, tissue specific and also under the influence of genetic variation, we will test the imputation accuracy in diverse datasets based on ancestry, age, tissue etc.

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. There will be GDPR/data security considerations when developing the online server for new users to upload their datasets.

Expected outcome and impact

This is a high impact project. SNP imputation panels are used by all studies that generate genetic data. As mentioned above, no analogue currently exists for DNAm. This project would therefore have an impact for the global research community by saving cohorts £millions from generating new data that would lead to discoveries in biomedical science.

References

1. *https://imputationserver.sph.umich.edu/index.html#!*
2. Bernabeu, E, McCartney, DL, Gadd, DA, Hillary, RF, Lu, AT, Murphy, L, Wrobel, N, Campbell, A, Harris, SE, Liewald, D, Hayward, C, Sudlow, C, Cox, SR, Evans, KL, Horvath, S, McIntosh, AM, Robinson, MR, Vallejos, CA, & Marioni, RE (2023). Refining epigenetic prediction of chronological and biological age. *Genome Med*, 15, 1:12.
3. McCartney, DL, Hillary, RF, Stevenson, AJ, Ritchie, SJ, Walker, RM, Zhang, Q, Morris, SW, Bermingham, ML, Campbell, A, Murray, AD, Whalley, HC, Gale, CR, Porteous, DJ, Haley, CS, McRae, AF, Wray, NR, Visscher, PM, McIntosh, AM, Evans, KL, Deary, IJ, & Marioni, RE (2018). Epigenetic prediction of complex traits and death. *Genome Biol*, 19, 1:136.
4. McCartney, DL, Stevenson, AJ, Hillary, RF, Walker, RM, Bermingham, ML, Morris, SW, Clarke, TK, Campbell, A, Murray, AD, Whalley, HC, Porteous, DJ, Visscher, PM, McIntosh, AM, Evans, KL, Deary, IJ, & Marioni, RE (2018). Epigenetic signatures of starting and stopping smoking. *EBioMedicine*, 37:214-220.
5. Stevenson, AJ, Gadd, DA, Hillary, RF, McCartney, DL, Campbell, A, Walker, RM, Evans, KL, Harris, SE, Spires-Jones, TL, McRae, AF, Visscher, PM, McIntosh, AM, Deary, IJ, & Marioni, RE (2021). Creating and Validating a DNA Methylation-Based Proxy for Interleukin-6. *J Gerontol A Biol Sci Med Sci*, 76, 12:2284-2292.
6. Gadd, DA, Hillary, RF, McCartney, DL, Zaghlool, SB, Stevenson, AJ, Cheng, Y, Fawns-Ritchie, C, Nangle, C, Campbell, A, Flaig, R, Harris, SE, Walker, RM, Shi, L, Tucker-Drob, EM, Gieger, C, Peters, A, Waldenberger, M, Graumann, J, McRae, AF, Deary, IJ, Porteous, DJ, Hayward, C, Visscher, PM, Cox, SR, Evans, KL, McIntosh,

AM, Suhre, K, & Marioni, RE (2022). Epigenetic scores for the circulating proteome as tools for disease prediction. *Elife*, 11:no page given.

7. Gadd DA, Smith HM, Mullin D, Chybowska O, Hillary RF, Kimenai DM,…, Harris SE, Welsh P, Sattar N, Cox SR, McCartney DL, & Marioni RE. DNAm scores for serum GDF15 and NT-proBNP levels associate with a range of traits affecting the body and brain. (2023) *medRxiv* https://doi.org/10.1101/2023.10.18.23297200
8. Hillary RF, Ng HK, McCartney DL, Elliott HR, Walker RM, Campbell A,…, Cox SR, Chambers JC, Loh M, Relton CR, Marioni RE*, Yousefi P*, Suderman M*. Blood-based epigenome–wide analyses of chronic low–grade inflammation across diverse population cohorts. (2023) [accepted at *Cell Genomics*]. *medRxiv*; 2023 doi: https://doi.org/10.1101/2023.11.02.23298000 *co-corresponding
9. Trejo Banos, D, McCartney, DL, Patxot, M, Anchieri, L, Battram, T, Christiansen, C, Costeira, R, Walker, RM, Morris, SW, Campbell, A, Zhang, Q, Porteous, DJ, McRae, AF, Wray, NR, Visscher, PM, Haley, CS, Evans, KL, Deary, IJ, McIntosh, AM, Hemani, G, Bell, JT, Marioni, RE, & Robinson, MR (2020). Bayesian reassessment of the epigenetic architecture of complex traits. *Nat Commun*, 11, 1:2865.
10. https://www.ed.ac.uk/generation-scotland
11. Cheng Y, Gieger C, Campbell A, McIntosh AM, Waldenberger M, McCartney DL, Marioni RE*, & Vallejos CA*. Feature pre-selection for the development of epigenetic biomarkers. (2024). *medRxiv*; 2024 doi: https://doi.org/10.1101/2024.02.14.24302694 *co-corresponding