Machine learning for large-scale translation-wide association studies

Fátima González González[1], Musa A. Hassan[2] and James G. D. Prendergast[2]

[1]MSc Graduate the University of Edinburgh MSc Systems and Synthetic Biology, [2] The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, UK (musa.hassan@roslin.ed.ac.uk, james.prendergast@roslin.ed.ac.uk)

fatimagonzalezglez@gmail.com
in/fatima-gonzalez-gonzalez
07469430755
London, UK

ABSTRACT

This work aims to investigate the feasibility of using associations between the rate of protein synthesis and heritable phenotypes to gain functional insights into results from genome-wide-association studies (GWAS), given proteins are the ultimate determinants of cellular function and the functional links between the genome and phenome. Traditional transcriptome-centered studies may leave behind genes regulated at translational level. There is a fundamental deficiency in our knowledge of complex phenotypes and how genetic variants drive the proteome to influence complex phenotypes, probably due to experimental costs. Our aim is to impute missing translation efficiency (TE) data, as a proxy of protein synthesis, into a GWAS cohort to perform TE-phenotype association studies (TEAS) using machine learning models. We adapted the PredictDB pipeline developed to impute gene expression for TE data. We found significant models to impute TE data that have the potential to be used for TEAS to enlighten the mechanisms that cause phenotypes.

BACKGROUND

Proteins are the ultimate determinants of **cellular function** and the functional links between the genome and phenome. Therefore, protein levels contribute causally to **heritable phenotypes**. Understanding the multistage process of **genetic regulation** that control steady state protein levels would therefore likely help us understand the variation of phenotypic traits. However, some of the intricacies of these processes remain **unknown**, but efforts to comprehend phenotypic variation.

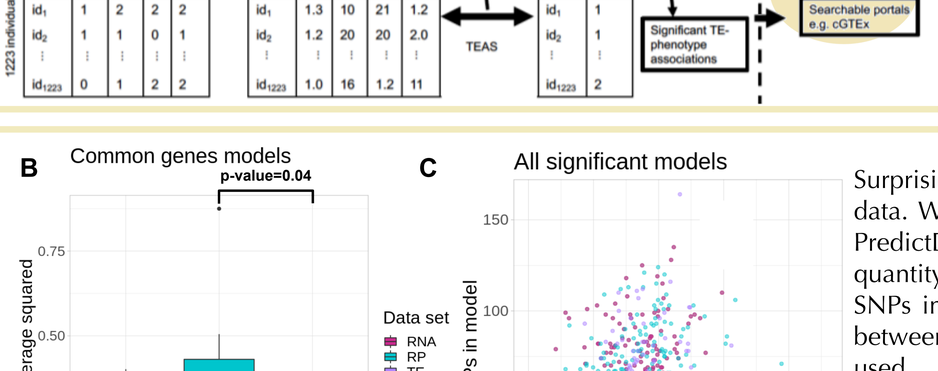
GenomeWide Association Studies (**GWAS**) map the links between genotype (SNPs) and phenotype.

TranscriptomeWide Association Studies (**TWAS**) link genes, rather than variants, to phenotypes. However, it cannot show translational and post translational impact on the genotype-phenotype relationship. Furthermore, protein levels often correlate poorly with RNA expression.Therefore, ideally we would like to work with protein level data, but this is extremely costly.

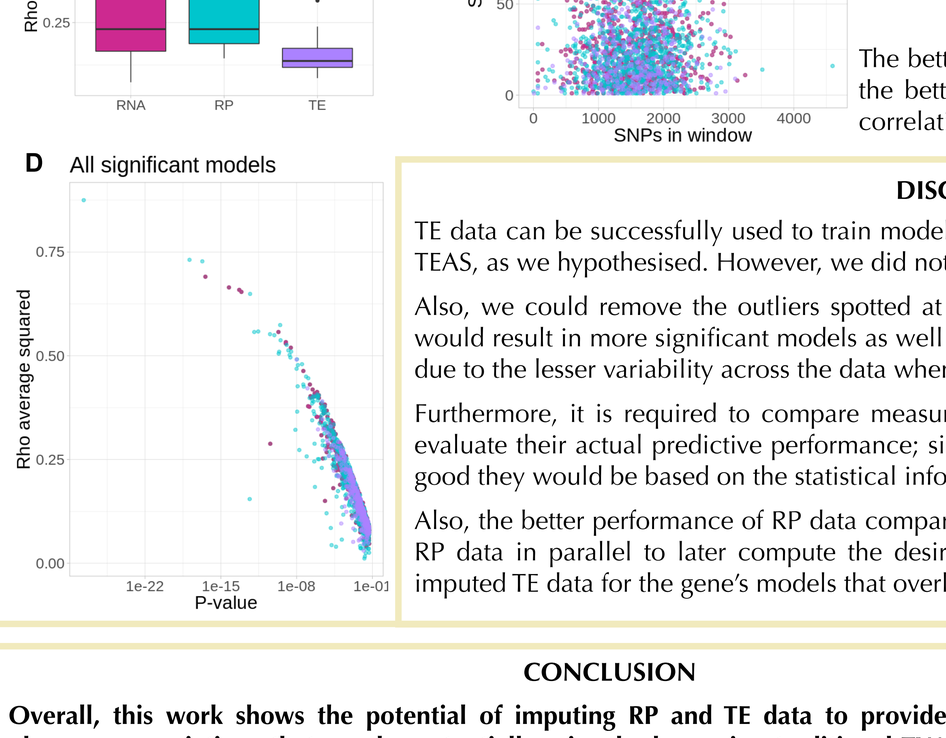
OBJECTIVE

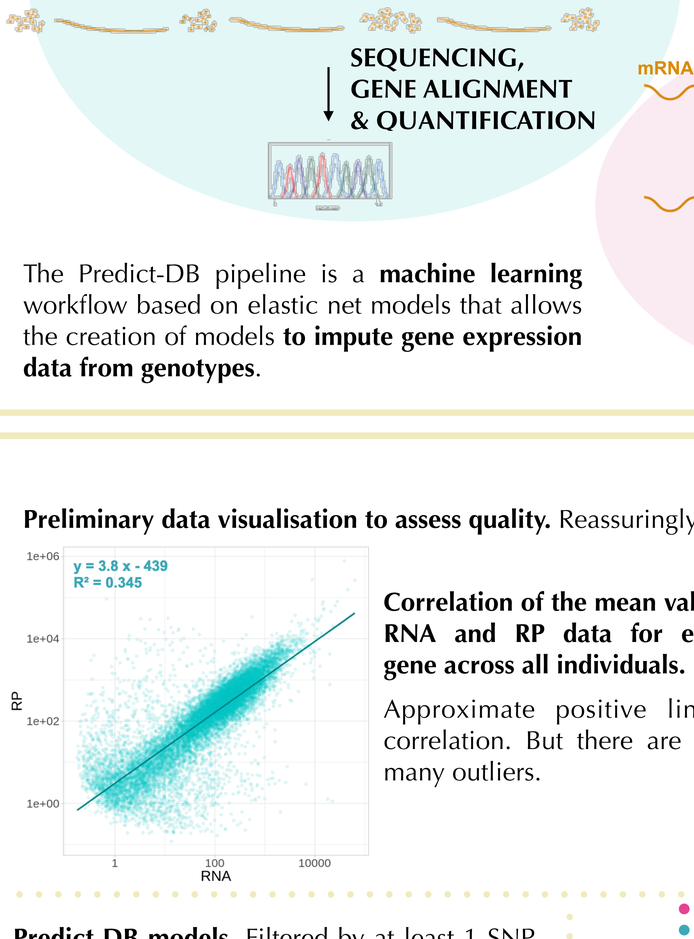
Investigate the feasibility of using associations between the rate of protein synthesis and heritable phenotypes to gain functional insights into results from Genome-Wide Association Studies (GWAS).

Overview of TWAS



Proposal of TEAS





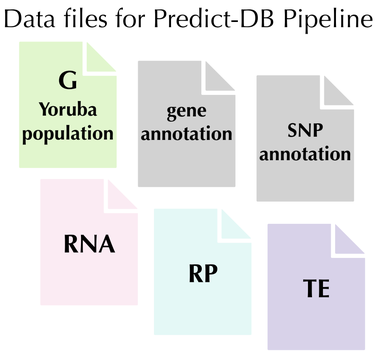
The Predict-DB pipeline is a **machine learning** workflow based on elastic net models that allows the creation of models **to impute gene expression data from genotypes**.

DATA COLLECTION AND PREPARATION

1. Normalisation
2. Compute TE for each gene.
3. Process with PEER tools to reduce non-genetic sources of variation.

Ribosome profiling data / RNA-seq data = **TRANSLATION EFFICIENCY**
The number of proteins synthesised per mRNA

Data files for Predict-DB Pipeline

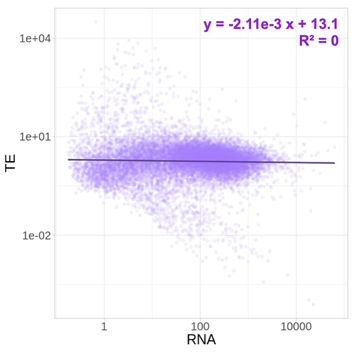


RESULTS

Preliminary data visualisation to assess quality. Reassuringly, the general trend of the data is consistent with previous observations.

Correlation of the mean values RNA and RP data for each gene across all individuals.

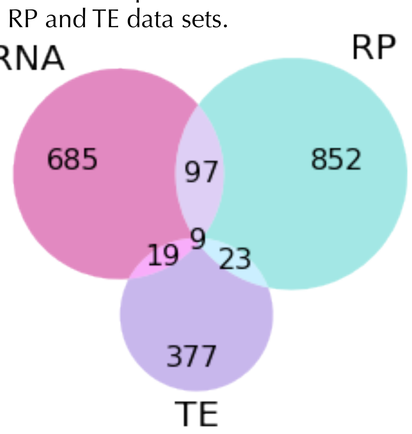
Approximate positive linear correlation. But there are still many outliers.



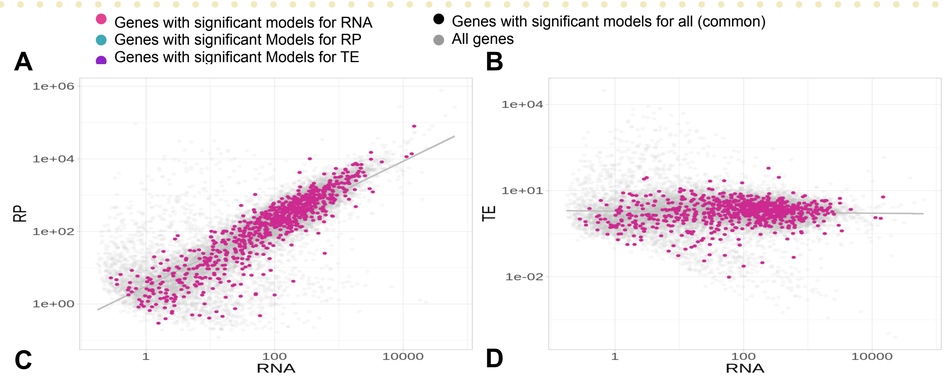
Correlation of the mean values RNA and TE data for each gene across all individuals.

Most data lies around the TE value of 1 showing a rough linear relationship between ribosome occupancy and transcript abundance.

Predict-DB models. Filtered by at least 1 SNP per model and a p-value smaller than 0.05 for RNA, RP and TE data sets.



We are surprised about the little overlapping we find, given the data sets is fairly correlated.



Surprisingly, we obtained the most significant models for RP data. We had expected it to be RNA, given this was what PredictDB has been developed for. However, other than in quantity, there is no significant differences between the SNPs in model, values of *rho average squared* or *p-values* between the models obtained for each of the data types used.

The better the correlation coefficient (*rho average squared)*, the better the *p-value*, which translates into the better the correlation in general.

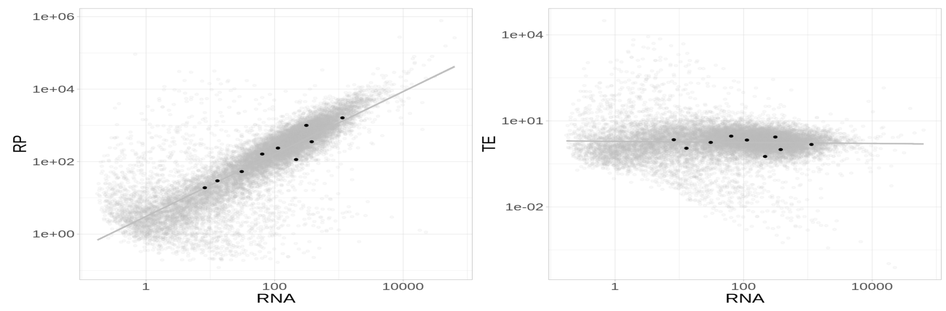
DISCUSSION

TE data can be successfully used to train models to further predict this type of data to integrate into TEAS, as we hypothesised. However, we did not perform an external validation of the models.

Also, we could remove the outliers spotted at the preliminary visualisation. Enhanced correlation would result in more significant models as well as more overlapping between the different data sets, due to the lesser variability across the data when performing the internal nested cross-validation.

Furthermore, it is required to compare measured TEs with imputed TEs through these models, to evaluate their actual predictive performance; since in this work we are only able to anticipate how good they would be based on the statistical information obtained.

Also, the better performance of RP data compared to TE, brings the possibility of imputing RNA and RP data in parallel to later compute the desired TEs. However, that alternative would only offer imputed TE data for the gene's models that overlap for the two data types.



Further, we looked for genes that obtained significant models in the correlation plots for RNA and RP or TE, to see if there is any pattern underlying the genes with the best models We do not see a preferred location for these genes, so we cannot infer any rule to establish if the data will be successful producing predictive models prior to the training of them.

CONCLUSION

Overall, this work shows the potential of imputing RP and TE data to provide insights into genome-phenome associations that can be potentially missed when using traditional TWAS, based on transcript levels. The usage of ribosome occupancy data to obtain translation efficiencies is a better approximation to protein levels, which are the actual effectors of cellular functions and therefore the responsible for phenotypes.

Based on the findings of this work, using these morels with TE and RP data is a promising approach to be further explored.

REFERENCES

1. Alexis Battle, et al. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667, 2015.
2. Alexander Gusev, et al. Integrative approaches for larges-cale transcriptome wide association studies. *Nature genetics*, 48(3):245–252, 2016.
3. Michael Wainberg, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599, 2019.
4. Hila Gingold, et al. Determinants of translation efficiency and accuracy. *Molecular systems biology*, 7(1):481, 2011.
5. Can Cenik, et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome research*, 25(11):1610–1621, 2015.
6. Eric R Gamazon, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.