Genome-wide analysis of transcription factor binding sites

Anh Thu Phan[1], Chris M. Brown[1,2], Augustine Chen[1], Chun Shen Lim[1]

[1] Department of Biochemistry, University of Otago, Dunedin, Otago, 9054,New Zealand
[2] Genetics Otago, University of Otago, Dunedin, Otago, 9054,New Zealand

Anhthuphan@postgrad.otago.ac.nz

INTRODUCTION

Cis regulatory elements (CREs) are non-coding DNA sequences that are required for proper temporal and spatial control of gene expression. CREs often regulate gene expression by physically interacting with proteins at distinct sequences, referred to as protein-binding motifs. Genetic variations within CREs can disrupt protein-binding motifs, subsequently altering the interface of CRE–protein interaction, thus impacting gene expression. Certain sequence variations are linked to dysregulated gene expression and elevated susceptibility to human genetic diseases.

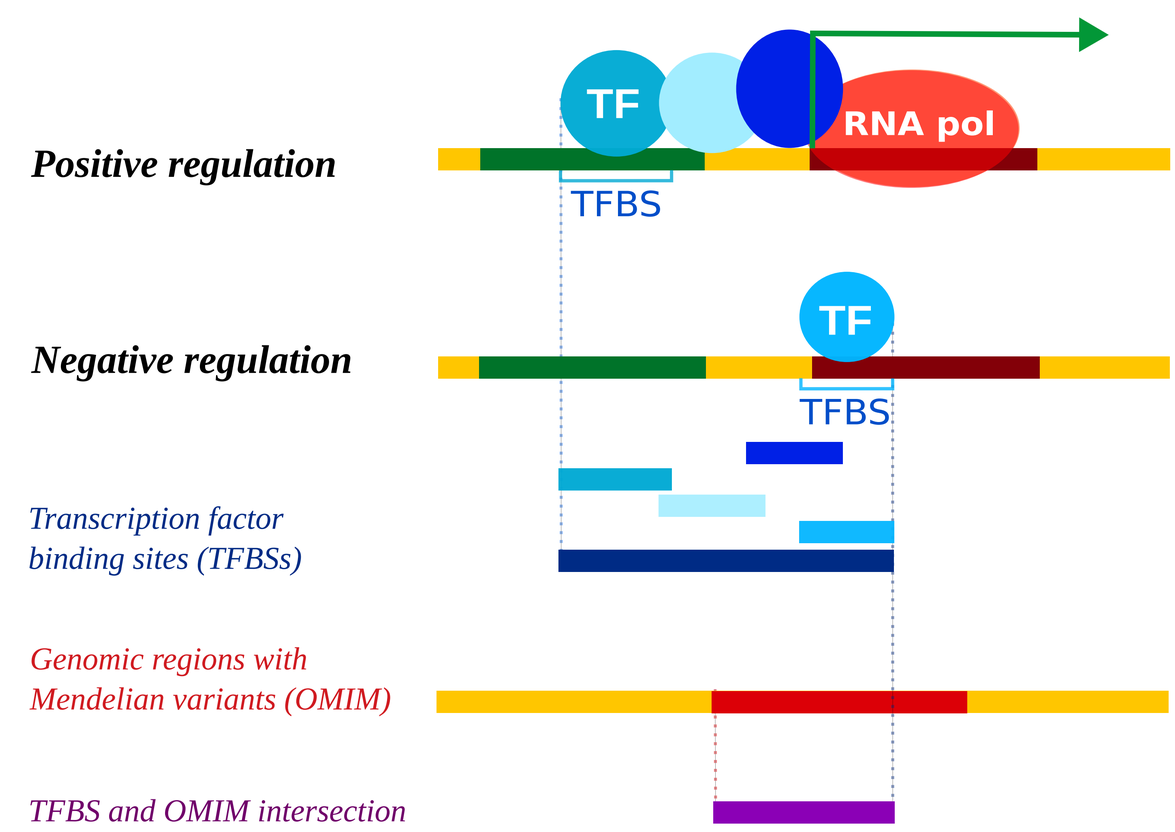

Figure 1. *Transcription factors (TFs) are proteins that target specific cis regulatory element termed transcription factor binding sites (TFBSs) to control transcription.* TFs can activate or suppress transcription by directly binding to the promoters of downstream genes in a sequence-specific manner.

OBJECTIVES

To investigate the association between sequence conservation in TFBS and disease associated changes, we generated a bioinformatic pipeline to:
(i) analyse the occurrence of experimentally determined TFBSs in regions overlapping variants associated with disease (OMIM).
(ii) evaluate the correlation between the clinical significance of genetic variation (OMIM) and sequence conservation of TFBSs (PhyloP).

METHODS

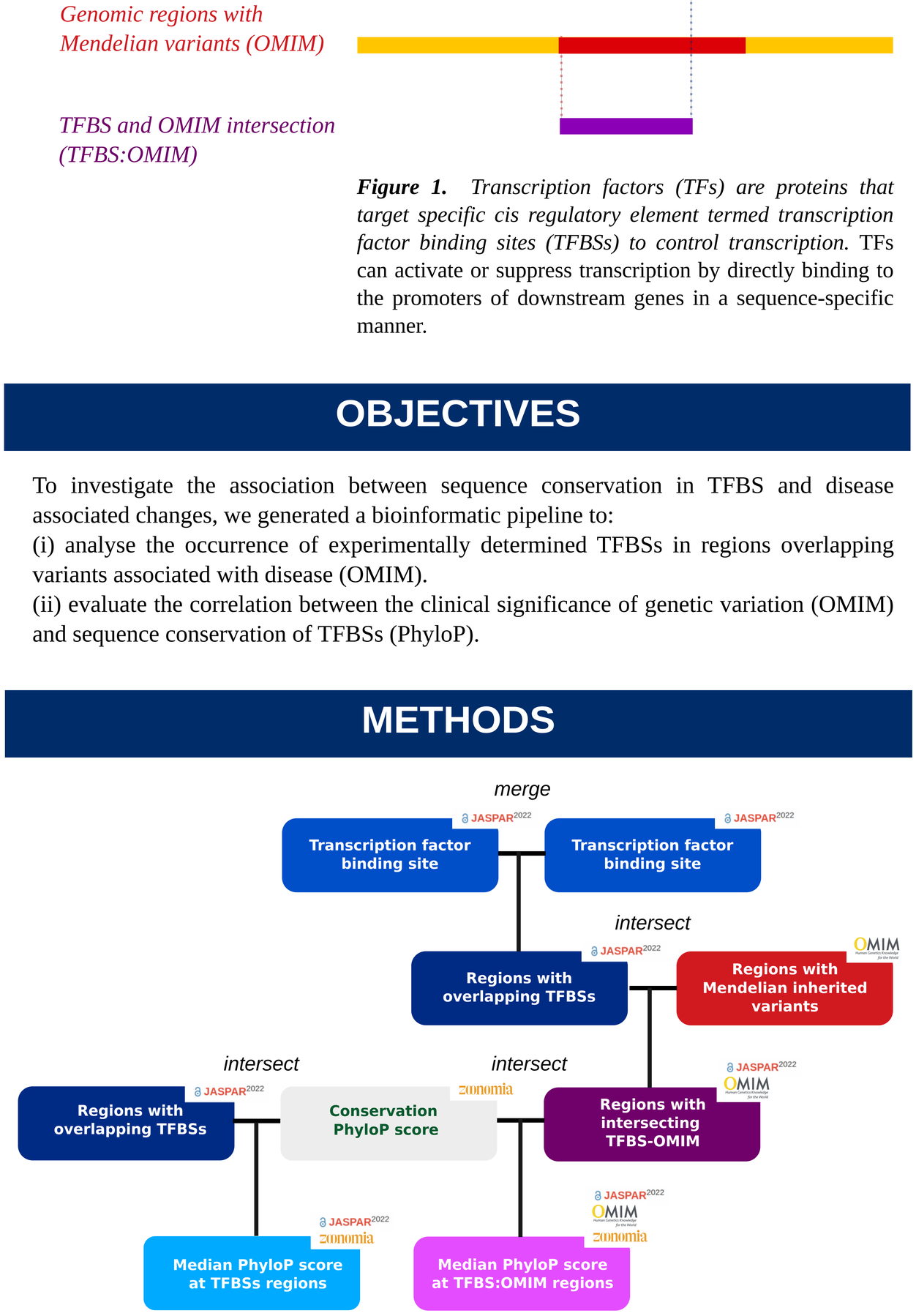

Figure 2. *An overview of the bioinformatics workflow*

RESULTS

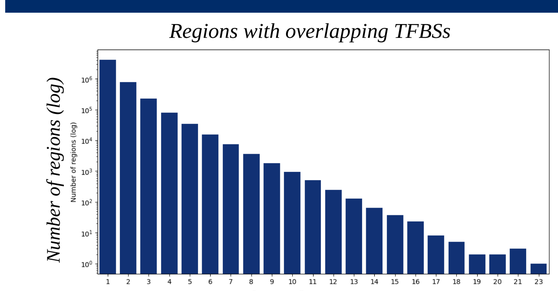

Figure 3. *Genomic regions containing multiple TFBSs were determined by merging overlapping TFBSs.*

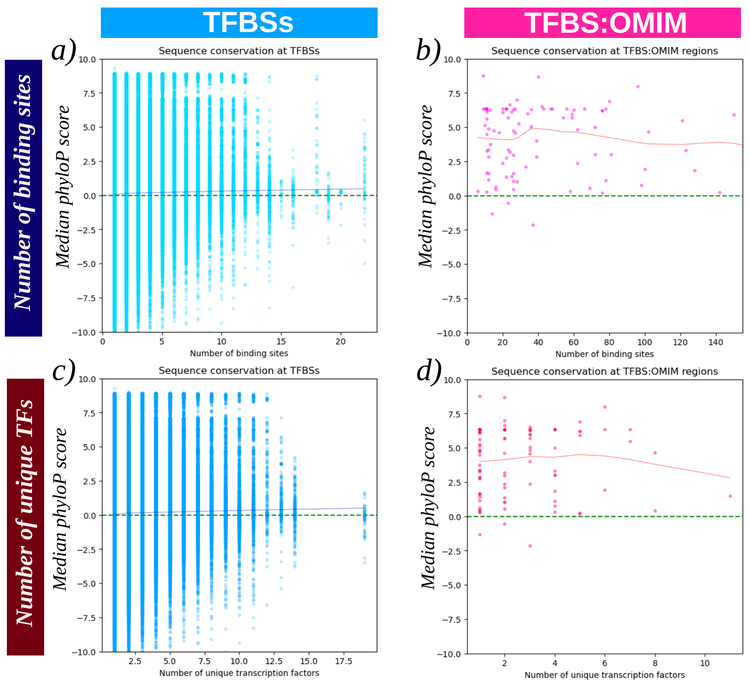

Figure 4. *Transcription factor binding sites in OMIM regions exhibit higher conservation levels than expected at random.*
a) and c) display the median phyloP score for 20,000 randomly selected TFBSs, b) and d) show the median phyloP score for TFBS that intersect with OMIM regions. A positive phyloP score denotes evolutionary sequence conservation, while a negative phyloP score indicates accelerated evolution.

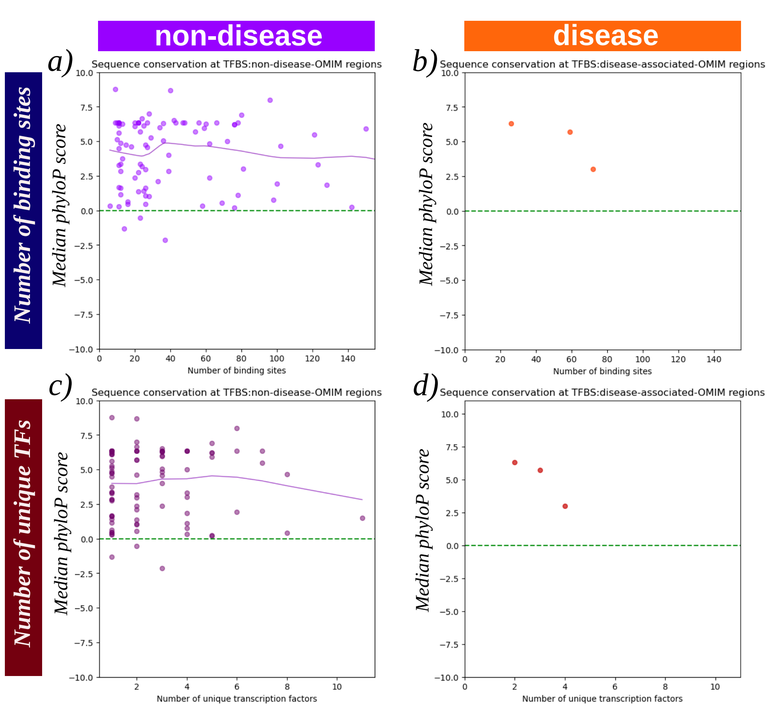

Figure 5. *Sequence conservation for TFBSs intersected with OMIM regions based on disease status.*
a) and c) depict the median phyloP score for TFBSs within OMIM regions not linked to any disease, b) and d) show the median phyloP score for TFBSs within disease-associated OMIM regions.

FUTURE DIRECTIONS

My PhD studies will:
(i) Further assess the genome-wide frequency of sequence variations in candidate and known TFBS and other cis regulatory elements.
(ii) Explore the functional consequences of variation in TFBS on disease development.
(iii) Test selected elements experimentally.

ACKNOWLEDGEMENTS

This research is funded by:
-University of Otago *Postgraduate Scholarship (to ATP)*
-Royal Society Te Aparangi *Marsden Fast-Start Grant (to CSL)*
-University of Otago *Research Grant (to CSL and CMB)*
-Otago Medical School Foundation Trust *Dean's Bequest (to CSL)*
Special thanks to *Gabrielle Chieng* and *Sofia Magalhães Moreira.*

REFERENCES

Albert & Kruglyak (2015). *Nature Reviews Genetics*, 16(4), 197-212.
Amberger et al. (2015). *Nucleic acids research*, 43(D1), D789-D798.
Andrews et al. (2023). *Science*, 380(6643), eabn7930.
Arbiza et al. (2013). *Nature genetics*, 45(7), 723-729.
Epstein (2009). *Briefings in Functional Genomics and Proteomics*, 8(4), 310-316.
Farnham (2009). *Nature Reviews Genetics*, 10(9), 605-616.
Khan et al. (2018). *Nucleic acids research*, 46(D1), D260-D266.