Prioritisation of Genes and Variants Underlying Hypertrophic Cardiomyopathy GWAS Loci Using Computational Approaches

Dominique May West

St Anne's College

Division of Cardiovascular Medicine, Radcliffe Department of Medicine

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Abstract

Hypertrophic cardiomyopathy (HCM) is a common heart condition, affecting at least 1 in 200 to 500 individuals, and a leading cause of sudden cardiac death. HCM has traditionally been recognised as a monogenic disorder caused by rare variants in sarcomere genes. This scenario does however only account for ~40% of HCM cases and so the majority of patients' causal variants do not get identified and the current applied clinical protocol needs reviewing.

In this thesis, the polygenic aspects of HCM are explored with the central aim being to identify plausible causal variants and their likely target gene(s). Prior to this thesis, HCM GWASs comprising of data from seven cohorts were combined into an overarching meta-analysis and a multi-trait analysis of genome-wide association summary statistics (MTAG) which identified 68 (50 novel) loci. Following this availability of HCM genome wide association data, a two-fold approach of gene prioritisation followed by functional variant prioritisation ensued. The former entailed employing existing gene prioritisation tools, leveraging expression data, and conducting case-control rare variant burden analyses all for the identification of plausible (and tractable) genes. The latter utilised machine learning approaches to gain insight into the chromatin landscape of the heart so as to pinpoint significant common susceptibility variants at a regulatory level. The final stage of this thesis attempted to draw a link between the shortlist of both prioritised genes and variants to best define probable routes of pathogenicity. Twenty-six of the 68 HCM loci were resolved to likely causal gene(s). Machine learning prioritised 53 variants, 24 of which were shown to open up and 29 to close the chromatin. Case studies linking prioritised variants to their genes computationally highlighted the detrimental effects of *MTSS1* in the context of HCM; identified an inversion disrupting interactions between *KANSL1* and other genes in an otherwise puzzling locus; fine-tuned the *TBX3* locus with this gene being the sole likely target; added weight to *PXN* through its functional common variant being assigned the most significant impact score.

Collectively, these analyses put forward a list of implicated genes and functional common variants to be investigated *in vitro* for greater insight into the disease mechanisms associated with common variants and HCM.

Prioritisation of Genes and Variants Underlying Hypertrophic Cardiomyopathy GWAS Loci Using Computational Approaches



Dominique May West

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

*Dedicated to my sister, **Alexandra West**, who passed away from HCM.*

[1992-1998]

Our girl with the purest soulbird dancing within,

Beautiful, loving, and rare –

An angel on Earth.

She taught us of the impermanence of life

And how to live with nothing but compassion.

She lives on in our hearts and in everything around us.

Acknowledgements

My DPhil undertakings reflect not only my own efforts but rather a wider team effort whereby many other individuals have contributed hugely to my understanding of HCM and scientific analyses, and without whom this would not have been possible.

On that note, I am deeply indebted to my sponsors, Sir James Lancaster as well as Professor Hugh Watkins (and the British Heart Foundation in tow) for their financial support spanning the duration of my project (which covered the necessary resources and my living costs) ultimately making my DPhil endeavours possible.

I am endlessly thankful for the support and guidance my supervisors, Professor Hugh Watkins and Dr Anuj Goel, have provided me.

My initial interaction with Hugh was before I had even considered applying for a DPhil in his research group and he was incredibly helpful and prompt upon realising my passion for this field. Since beginning my DPhil journey, Hugh has handled my personal connection to HCM both actively and delicately. I am in awe of his extensive knowledge in the field of cardiac genetics. Although his intellect is far beyond someone like myself, his humility and keenness in helping me learn throughout this process has been shown in his patience when translating thoughts and suggestions. I am grateful for his guidance, attentiveness, and constructive criticism.

Anuj's experience and knowledge in the computational biology realm has been hugely beneficial to me. His availability and thoroughness – always being there to troubleshoot with me and to lend his advice has been so appreciated. I was warned about feeling 'lost' during a DPhil and that has quite simply never been the case thanks to Anuj's support, reassurance, and constant reminders to assess my work from a bird's eye view (with less tunnel vision!). I am very grateful for the guiding hand I have received from Anuj and overall for his time and kindness.

The University of Oxford, specifically the Wellcome Centre for Human Genetics, MRC Weatherall Institute of Molecular Medicine, and Radcliffe Department of Medicine, provided an unprecedented research infrastructure that has been at the heart of my growth as a researcher and to the analyses I have been able to perform. Within the Watkins group, I have been very fortunate for Dr Chris Grace's support as he has helped me navigate the complexities of extensive bioinformatic data, as well as both Dr Kate Thomson and Dr Liz Ormondroyds' insights into the clinical aspects of HCM and variant classification. The Hughes group, specifically Simone Riva, Dr Matthew Baxter, Dr Edward Sanders, and Prof Jim Hughes have played an integral role in bridging my understanding of the regulatory landscape workings and have been invaluable in assisting me with machine learning implementations.

A special thank you to Dr Sarah Kraus and her team in South Africa for enrolling my family in IMHOTEP and for opening up Alex's case with hopes of finding a genetic diagnosis – for us to get closer to some kind of closure whilst also having our hearts examined too. My family is so grateful.

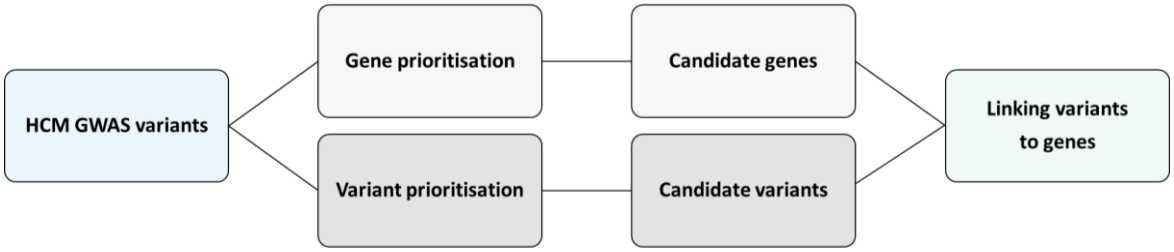
I am especially appreciative of my fiancé and my family for the unparalleled emotional support throughout this journey. Having a supportive partner by my side at each step of this journey – for all the small celebrations yet also always there to help pick up the pieces, and ultimately for believing in me whilst I pursued my lifelong dream. My parents and brother in South Africa, who have sacrificed a great deal to make this dream of mine (ours) possible and who urged me to move across the world for this opportunity, have endured facetimes each and every day, listened to my scientific waffling, and showered me with love from all the way across the world. It is with the love and support from these four people that I am where I am today.

Published works

Peer-reviewed publications

Tadros, R., Zheng, S. L., Grace, C., Jordà, P., Francis, C., **West, D. M.**, Jurgens, S. J., Thomson, K. L., Harper, A. R., Ormondroyd, E., Xu, X., Theotokis, P. I., Buchan, R. J., McGurk, K. A., Mazzarotto, F., Boschi, B., Pelo, E., Lee, M., Noseda, M., Varnava, A., … Watkins, H. (2025). **Large-scale genome-wide association analyses identify novel genetic loci and mechanisms in hypertrophic cardiomyopathy**. Nature genetics, 57(3), 530–538. https://doi.org/10.1038/s41588-025-02087-4

Abstract



Hypertrophic cardiomyopathy (HCM) is a common heart condition, affecting at least 1 in 200 to 500 individuals, and a leading cause of sudden cardiac death. HCM has traditionally been recognised as a monogenic disorder caused by rare variants in sarcomere genes. This scenario does however only account for ~40% of HCM cases and so the majority of patients' causal variants do not get identified and the current applied clinical protocol needs reviewing.

In this thesis, the polygenic aspects of HCM are explored with the central aim being to identify plausible causal variants and their likely target gene(s). Prior to this thesis, HCM GWASs comprising of data from seven cohorts were combined into an overarching meta-analysis and a multi-trait analysis of genome-wide association summary statistics (MTAG) which identified 68 (50 novel) loci. Following this availability of HCM genome wide association data, a two-fold approach of gene prioritisation followed by functional variant prioritisation ensued. The former entailed employing existing gene prioritisation tools, leveraging expression data, and conducting case-control rare variant burden analyses all for the identification of plausible (and tractable) genes. The latter utilised machine learning approaches to gain insight into the chromatin landscape of the heart so as to pinpoint significant common susceptibility variants at a regulatory level. The final stage of this thesis attempted to draw a link between the shortlist of both prioritised genes and variants to best define probable routes of pathogenicity. Twenty-six of the 68 HCM loci were resolved to likely causal gene(s). Machine learning prioritised 53 variants, 24 of which were shown to open up and 29 to close the chromatin. Case studies linking prioritised variants to their genes computationally highlighted the detrimental effects of *MTSS1* in the context of HCM; identified an inversion disrupting interactions between *KANSL1* and other genes in an otherwise puzzling locus; fine-tuned the *TBX3* locus with this gene being the sole likely target; added weight to *PXN* through its functional common variant being assigned the most significant impact score.

Collectively, these analyses put forward a list of implicated genes and functional common variants to be investigated *in vitro* for greater insight into the disease mechanisms associated with common variants and HCM.

Contents

ACKNOWLEDGEMENTS 4

PUBLISHED WORKS 5

ABSTRACT 6

CONTENTS 7

LIST OF FIGURES 10

LIST OF TABLES 13

INTRODUCTION 14

1.1 INCIDENCE 14
1.2 MANIFESTATION & PRESENTATION 14
1.3 GENETIC ARCHITECTURE 16
1.3.1 THE SARCOMERE 16
1.3.2 CORE SARCOMERIC GENES 17
1.3.3 CAUSAL VARIANTS AND THEIR MOLECULAR EFFECTS 18
1.3.4 PRESENT DAY HCM GENES & GENETIC TESTING 19
1.3.5 SHIFT FROM MONOGENIC MINDSET 20
1.4 POLYGENIC HCM 21
1.4.1 GWAS DISCOVERY 22
1.4.2 ADVANCEMENT IN TECHNOLOGIES 24
1.4.3 DEEP LEARNING APPLIED TO THE REGULATORY LANDSCAPE 26
HYPOTHESES AND KEY OBJECTIVES 29

MATERIALS AND METHOD 30

2.1 INTRODUCTION 30
2.2 GENE PRIORITISATION 32
2.2.1 OVERVIEW 32
2.2.2 DEFINING HCM LOCI 32
2.2.3 DEVISED EXPRESSION-BASED QUADRANT APPROACH 37
2.2.4 ASSESSING REGULATORY ROLES THROUGH COLOCALISATION OF COMMON VARIANTS AND EQTL DATA 46
2.2.5 ASSESSING THE BURDEN OF RARE VARIANTS IN PRIORITISED GENES 49
2.2.6 EMPLOYING THE POLYGENIC PRIORITY SCORE TOOL FOR GENE PRIORITISATION 53
2.3 VARIANT PRIORITISATION 56
2.3.1 OVERVIEW 56
2.3.2 CELL TYPE ENRICHMENT 56
2.3.3 CHROMATIN DATA QC AND STANDARDISATION VIA CATCH-UP 57
2.3.4 USING MACHINE LEARNING TO PRIORITISE FUNCTIONAL VARIANTS 60
2.4 LINKING PRIORITISED VARIANTS TO TARGET GENES 77
2.4.1 OVERVIEW 77
2.4.2 ANNOTATING REGULATORY ELEMENTS 77
2.4.3 VARIANT-TO-GENE ASSIGNMENT 78

2.4.4 Enhancer-to-gene Mapping 79
2.4.5 Peak-to-gene Linkage 79
2.4.6 Gene Ontology Exploration 81
2.4.7 Motif Analysis 82
2.4.8 Case Studies 84
2.4.9 Summarising implicated genes 85
2.4.10 Performing a PheWAS lookup for prioritised functional variants 85

GENE PRIORITISATION 87

3.1 Introduction 87
3.1.1 Gene prioritisation tools 87
3.1.2 Leveraging RNA-Seq data for gene prioritisation 90
3.1.3 Key objectives 92
3.1.4 Computational approaches employed 92
3.2 Defining HCM loci 93
3.2.1 Proximity-based approach 93
3.2.2 TAD-based approach 96
3.3 Devised Expression-Based Quadrant Approach 103
3.3.1 Bulk proximity-based quadrant approach 103
3.3.2 Single-cell TAD-based quadrant approach 107
3.3.3 Comparison of the two quadrant approaches 111
3.4 Assessing Regulatory Roles Through Colocalisation of Common Variants and eQTL Data 117
3.4.1 Coloc abf 117
3.5 Assessing the Burden of Rare Variants in Prioritised Genes 119
3.5.1 LoF burden analysis 119
3.5.2 Missense burden analysis 121
3.6 Employing the Polygenic Priority Score tool for gene prioritisation 127
3.7 Discussion & Limitations 131

VARIANT PRIORITISATION 135

4.1 Introduction 135
4.1.1 The regulatory landscape 135
4.1.2 Chromatin data and its uses 138
4.1.3 Chromatin data cell type / tissue selection 139
4.1.4 Chromatin Data QC and Standardisation Via CATCH-UP 140
4.1.5 Using Machine Learning to Prioritise Functional Variants 140
4.1.6 Key objectives 144
4.1.7 Computational approaches employed 145
4.2 Cell type enrichment 145
4.3 Chromatin Data QC and Standardisation Via CATCH-UP 148
4.4 Using Machine Learning to Prioritise Functional Variants 149
4.4.1 DeepHaem 149
4.4.2 Deepmind 177
4.5 Discussion & Limitations 187

LINKING PRIORITISED VARIANTS TO TARGET GENES 190

5.1 Introduction 190
5.1.1 Annotating Regulatory Elements 190

5.1.2 VARIANT-TO-GENE ASSIGNMENT 191
5.1.3 ENHANCER-TO-GENE MAPPING 192
5.1.4 PEAK-TO-GENE LINKAGE 192
5.1.5 MOTIF ANALYSIS 193
5.1.6 KEY OBJECTIVES 194
5.1.7 COMPUTATIONAL APPROACHES EMPLOYED 194
5.2 ANNOTATING REGULATORY ELEMENTS 195
5.3 VARIANT-TO-GENE ASSIGNMENT 200
5.4 ENHANCER-TO-GENE MAPPING 202
5.5 PEAK-TO-GENE LINKAGE 204
5.6 GENE ONTOLOGY EXPLORATION 212
5.6.1 DAVID 212
5.6.2 GPROFILER 216
5.7 MOTIF ANALYSIS 218
5.8 DISCUSSION & LIMITATIONS 224
5.9 CASE STUDIES 225
5.9.1 CONSIDERATIONS/ASSUMPTIONS 226
5.9.2 CASE STUDY LOCI 228
5.10 SUMMARY TABLES 252

CONCLUSIONS 262

6.1 INTRODUCTION 262
6.2 GENE PRIORITISATION GIVEN HCM GWAS COMMON VARIANTS 262
6.3 VARIANT PRIORITISATION GIVEN HCM GWAS COMMON VARIANTS 264
6.4 BRINGING THE STORY TOGETHER: CONNECTING PRIORITISED VARIANTS TO GENES 264
6.5 FUTURE DIRECTIONS 265

REFERENCES 267

List of figures

Figure 1 | HCM vs normal heart echo 15
Figure 2 | The cardiac sarcomere 16
Figure 3 | A neural network with fully connected layers 27
Figure 4 | Thesis methodologies summarised 31
Figure 5 | Defining HCM loci 32
Figure 6 | Assigning TADs in Juicebox 34
Figure 7 | Intersecting HCM variants & TADs for loci demarcation 36
Figure 8 | Quadrant approach pipeline 38
Figure 9 | Quadrant approach visualisation 40
Figure 10 | Colocalisation hypotheses 47
Figure 11 | Rare variant burden analysis overview 51
Figure 12 | Rare variant burden analysis pipeline 52
Figure 13 | PoPs gene prioritisation pipeline 55
Figure 14 | Chromatin data QC tool (CATCH-UP) pipeline 58
Figure 15 | DeepHaem rolling window predictions and chromatin accessibility 63
Figure 16 | DeepHaem rolling window predictions: forward, reverse, and combined DNA strands 64
Figure 17 | DeepHaem damage score prediction for variants 67
Figure 18 | DeepHaem visual inspections involving sequence modification: peak erasure visualisation 72
Figure 19 | Sigmoid vs SoftMax activation function 73
Figure 20 | Peak vs variant centering in DeepHaem's rolling window predictions 74
Figure 21 | Motif discovery via HOMER pipeline 83
Figure 22 | Gene prioritisation approach 92
Figure 23 | BAG3 quadrant approach visualisation as prioritised with bulk LV data 106
Figure 24 | SVIL quadrant approach visualisation as prioritised with single-cell cardiomyocyte data 110
Figure 25 | ALPK3 AlphaFold protein structure with HCM missense variants 125
Figure 26 | RAF1 AlphaFold protein structure with HCM missense variants 126
Figure 27 | Regulatory elements 135
Figure 28 | The Regulatory Landscape 138
Figure 29 | Transformers for natural language processing 143
Figure 30 | Variant prioritisation approach 145
Figure 31 | Chromatin QC (CATCH-UP) visualisation before and after 149
Figure 32 | DeepHaem training & validation auc/loss plots with 24 classifiers and sigmoid final activation layer 150
Figure 33 | DeepHaem training and validation ROC curves for 24 classifiers and sigmoid final activation layer 152
Figure 34 | DeepHaem visual inspection: region1 in ventricular cardiomyocyte classifier 154
Figure 35 | DeepHaem visual inspection: region2 in ventricular cardiomyocyte classifier 154
Figure 36 | DeepHaem visual inspection with modified DNA sequence: region 3 peak creation in ventricular cardiomyocyte classifier 155
Figure 37 | DeepHaem visual inspection with modified DNA sequence: region1 peak erasure in ventricular cardiomyocyte classifier 156
Figure 38 | DeepHaem's rolling window prediction visualisation: forward, reverse, and combined strands 157
Figure 39 | DeepHaem damage score distribution per 24 classifiers for HCM variants 158
Figure 40 | DeepHaem's damage score distribution per HCM positive control set in ventricular cardiomyocyte classifier 159
Figure 41 | DeepHaem's damage score distribution for HCM GWAS variants in the ventricular cardiomyocyte classifier 160

Figure 42 | DeepHaem training & validation auc/loss plots with 65 classifiers and sigmoid final activation layer 161
Figure 43 | DeepHaem training and validation ROC curves for 65 classifiers and sigmoid final activation layer 162
Figure 44 | DeepHaem damage score distribution per 65 classifiers for erythroid enhancer positive control set 163
Figure 45 | DeepHaem training & validation auc/loss plots with 65 classifiers and SoftMax final activation layer 164
Figure 46 | DeepHaem training and validation ROC curves for 65 classifiers and SoftMax final activation layer 165
Figure 47 | DeepHaem training & validation auc/loss plots with 1 erythroid & 1 heart classifier and sigmoid final activation layer 166
Figure 48 | DeepHaem training and validation ROC curves for 1 erythroid & 1 heart classifier and sigmoid final activation layer 167
Figure 49 | DeepHaem training & validation auc/loss plots with 3 erythroid & 3 heart classifiers and sigmoid final activation layer 167
Figure 50 | DeepHaem training and validation ROC curves for 3 erythroid & 3 heart classifiers and sigmoid final activation layer 168
Figure 51 | DeepHaem training & validation auc/loss plots with 1 erythroid & 12 heart classifiers and sigmoid final activation layer 168
Figure 52 | DeepHaem training and validation ROC curves for 1 erythroid & 12 heart classifiers and sigmoid final activation layer 169
Figure 53 | DeepHaem training & validation auc/loss plots with 1 erythroid & 28 heart classifiers and sigmoid final activation layer 169
Figure 54 | DeepHaem training and validation ROC curves for 1 erythroid & 28 heart classifiers and sigmoid final activation layer 170
Figure 55 | DeepHaem visual inspection (peak centered): variant rs198415 in Ery Don003 classifier 171
Figure 56 | DeepHaem visual inspection (variant centered): variant rs198415 in Ery Don003 classifier 172
Figure 57 | DeepHaem visual inspection: region1 in ventricular cardiomyocyte classifier 173
Figure 58 | DeepHaem visual inspection: region1 in Ery Don001 classifier 174
Figure 59 | DeepHaem visual inspection: region4 in ventricular cardiomyocyte classifier 175
Figure 60 | DeepHaem visual inspection: region4 in Ery Don001 classifier 175
Figure 61 | Process of employing DeepHaem for variant prioritisation 176
Figure 62 | Deepmind Enformer's damage score distribution per 5 classifiers for HCM variants 178
Figure 63 | Deepmind Enformer's damage score distribution of HUVEC gain, neutral, and loss empirical variants in both HUVEC classifiers 179
Figure 64 | Deepmind Enformer's damage score distribution of HUVEC extended empirical neutral variants in both HUVEC classifiers 180
Figure 65 | Overview of Deepmind Enformer's findings regarding HCM variant prioritisation 182
Figure 66 | Linking variants to genes approach 194
Figure 67 | Assigned regulatory elements (via REgulamentary) to cardiac muscle cell data 196
Figure 68 | Frequencies of assigned regulatory elements (via REgulamentary) to cardiac muscle cell data 197
Figure 69 | Frequencies of assigned regulatory elements (via REgulamentary) to cardiac muscle cell data for DNase+CTCF, DNase, and CTCF post-intersection with the 4395 HCM GWAS lead and proxy variants 198
Figure 70 | Frequencies of assigned regulatory elements (via REgulamentary) to cardiac muscle cell data intersected with prioritised HCM variants for DNase+CTCF, DNase, and CTCF post-intersection with the 53 prioritised HCM variants 198
Figure 71 | The cytoskeleton in muscle cells pathway 215
Figure 72 | GO ontology functional annotation clustering via gProfiler given prioritised genes list 216

Figure 73 | Clustering of HP terms via gProfiler given prioritised genes list 217
Figure 74 | rs35006907 Locus zoom plot 231
Figure 75 | rs116904997 STRING database interactions 235
Figure 76 | rs116904997 Locus zoom plot 235
Figure 77 | rs11748963 Locus zoom plot 241
Figure 78 | rs28768976 Locus zoom plot 247
Figure 79 | rs7487962 STRING database interactions 250
Figure 80 | rs7487962 Locus zoom plot 251
Figure 81 | PheWAS look-up for prioritised HCM variants 261

List of tables

Table 1 | Single-cell RNA-Seq data samples & labs 43
Table 2 | Applying one-hot encoding to DNA sequences 61
Table 3 | 24 classifiers for input to DeepHaem 62
Table 4 | DeepHaem visual inspections: regions selected 65
Table 5 | RegulomeDB HCM positive control set of variants 68
Table 6 | Putative regulatory HCM positive control set of variants 69
Table 7 | Proximity-based loci coordinates 93
Table 8 | Intersection of HCM variants with various sized TADs 97
Table 9 | TAD-based loci coordinates 100
Table 10 | Proximity-defined loci with bulk LV quadrant approach gene prioritisation 103
Table 11 | TAD-defined loci with single-cell cardiomyocyte quadrant approach gene prioritisation 107
Table 12 | HCM loci whereby the same gene was prioritised via both bulk LV and single-cell cardiomyocyte quadrant approach 112
Table 13 | HCM loci whereby a different gene was prioritised via bulk LV vs single-cell cardiomyocyte quadrant approach; single-cell gene most expressed in cardiomyocyte 113
Table 14 | HCM loci whereby a different gene was prioritised via bulk LV vs single-cell cardiomyocyte quadrant approach; single-cell gene most expressed in non-cardiomyocyte cell type 115
Table 15 | Colocalisation (coloc.abf) result 117
Table 16 | GeL LoF burden result 119
Table 17 | BRRD LoF burden result 120
Table 18 | Meta-analysis of GeL and BRRD LoF burden result 120
Table 19 | GeL missense burden result 121
Table 20 | BRRD missense burden result 122
Table 21 | Meta-analysis of GeL and BRRD missense burden result 122
Table 22 | PoPs gene prioritisation results 128
Table 23 | Cell type enrichment results 146
Table 24 | Deepmind Enformer HUVEC classifiers' derived prioritisation thresholds and associated counts 181
Table 25 | Detailed look at Deepmind Enformer's findings regarding HCM variant prioritisation 183
Table 26 | Regulatory elements assigned to HCM prioritised variants 199
Table 27 | Open target gene prioritisation given the prioritised HCM variants 200
Table 28 | ABC model enhancer-to-gene mapping intersected with prioritised HCM variants 203
Table 29 | ArchR peak-to-gene linkage intersected with prioritised HCM variants 205
Table 30 | GO ontology functional annotation clustering via DAVID given prioritised genes list 212
Table 31 | Kegg pathway functional annotation clustering via DAVID given prioritised genes list 214
Table 32 | Motif exploration surrounding prioritised variants via HOMER 219
Table 33 | rs35006907 Locus evidence table 228
Table 34 | rs116904997 Locus evidence table 232
Table 35 | rs11748963 Locus evidence table 237
Table 36 | rs28768976 Locus evidence table 243
Table 37 | rs7487962 Locus evidence table 248
Table 38 | Concluding target gene(s) per locus wherein variants were prioritised 252
Table 39 | PheWAS look-up for prioritised HCM variants 257

1

Introduction

1.1 Incidence

Hypertrophic cardiomyopathy (HCM), although previously classed as a rare disease, is the most common heart disorder (Maron *et al.*, 1995) with a global presence. HCM affects approximately 1 in 200 to 1 in 500 individuals resulting in ~15-20 million cases globally (Maron *et al.*, 2022).

The occurrence of HCM is considered equal for both genders. Nevertheless, research often indicates a higher prevalence among males, with reported figures varying between 55 and 70% (Dimitrow *et al.*, 1997).

1.2 Manifestation & presentation

HCM is a hereditary condition affecting cardiomyocytes of the heart and is distinguished by cardiac hypertrophy independent of loading conditions, a contracted left ventricle, and a normal to elevated ejection fraction (Marian and Braunwald, 2017).

HCM hearts' typically show a thickening of the left ventricle which results in a reduced chamber size and consequently, obstruction of outflow and mitral valve regurgitation (Geske *et al.*, 2018), diastolic dysfunction (Gwathmey *et al.*, 1991), and myocardial ischemia (Raphael *et al.*, 2016). Elongated mitral valve leaflets, abnormal intramural coronary arterioles, fibrosis, and left ventricular apical aneurysms accompanied by regional scarring may also result (Rowin *et al.*, 2020).

Echocardiograms, which are commonly used to detect HCM, can inspect systolic anterior motion (mitral-septal contact), mechanical resistance (Maron *et al.*, 2022), and mechanisms of left ventricular (LV) outflow obstruction, including mid-cavity muscle. Through echocardiograms, one can detect the LV outflow gradient, extent of mitral regurgitation, aortic/mitral valve issues, and systolic pulmonary pressures (normal versus HCM heart example shown in figure 1). In adulthood,

a maximum LV wall thickness of 15 mm or greater at any location within the chamber, not explained by loading conditions, is indicative of HCM (Zamorano *et al.*, 2014).

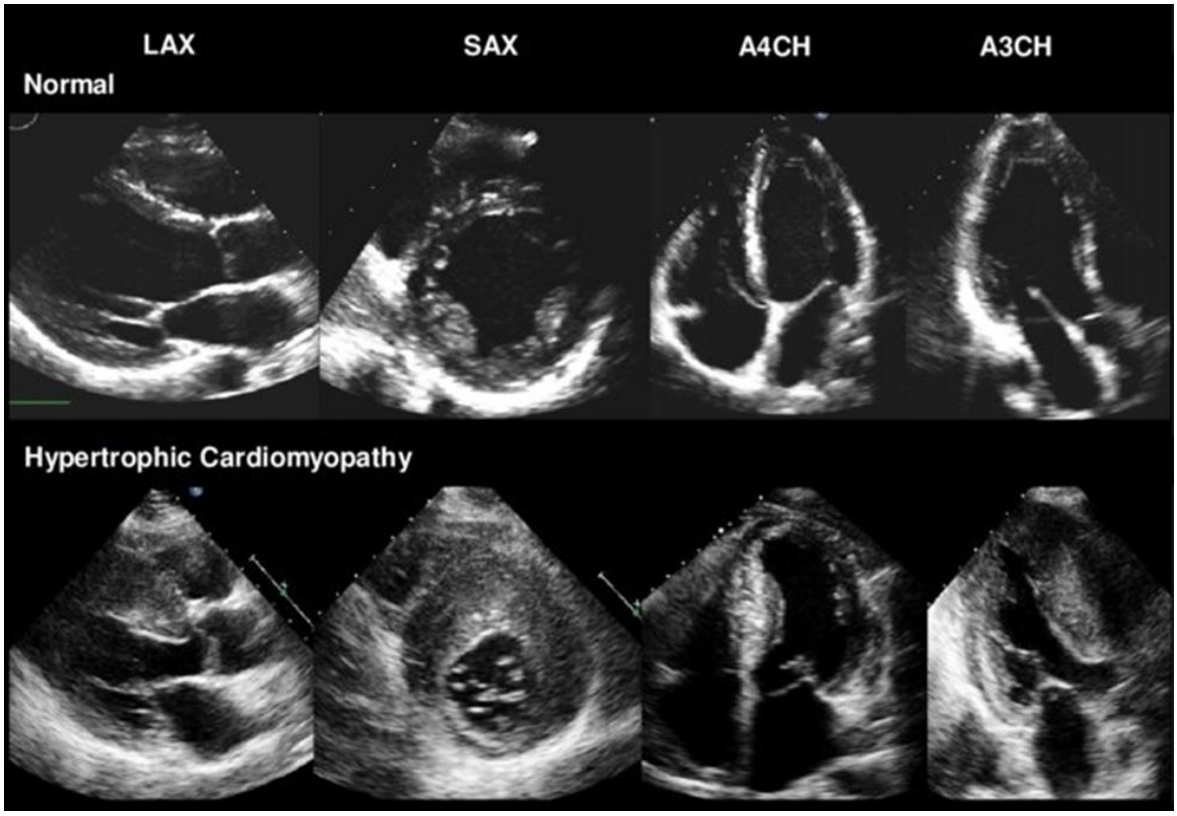


Figure 1 | HCM vs normal heart echo | A figure adapted from Kelshiker et al. (2014) showing the comparison of normal (at the top) versus HCM (at the bottom) myocardium as seen through conventional two-dimensional echocardiography. PLAX: parasternal long axis view; SAX: short axis view; A4CH: apical 4-chamber view; A3C: apical 3-chamber view.

Variability in the phenotype exists among cases regarding, for example, the extent and distribution of hypertrophy (Klues *et al.*, 1995). Marian (2002) contradicted the suggestion that specific clinical phenotypes are associated with a particular variant and instead suggested that variability arises from the diversity of causal genes and variants, as well as the influence of modifier genes and environmental factors. Additionally, the occurrence of unexplained asymptomatic hypertrophy among young adults in the United States (US) represents the recognised occurrence of HCM (1 in 200 to 1 in 500) while symptomatic hypertrophy affects fewer than 1 in 3000 adults (Semsarian *et al.*, 2015).

Sudden cardiac death (SCD) is a major cause of mortality in HCM cases and is often seen to impact younger as well as asymptomatic individuals (Maron *et al.*, 2000). Although the annual incidence of SCD is less than 1%, certain subgroups within the broader HCM population display elevated rates (Elliott *et al.*, 2000). Since 2014, the European Society of Cardiology (ESC) has advocated for the HCM Risk-SCD risk assessment tool to identify individuals with high SCD risk for whom the insertion of a prophylactic implantable cardioverter defibrillator would be beneficial (Sabrina *et al.*, 2025). Heart failure represents an even bigger cause of mortality from HCM in later life (Ho *et al.*, 2018).

1.3 Genetic architecture

1.3.1 The sarcomere

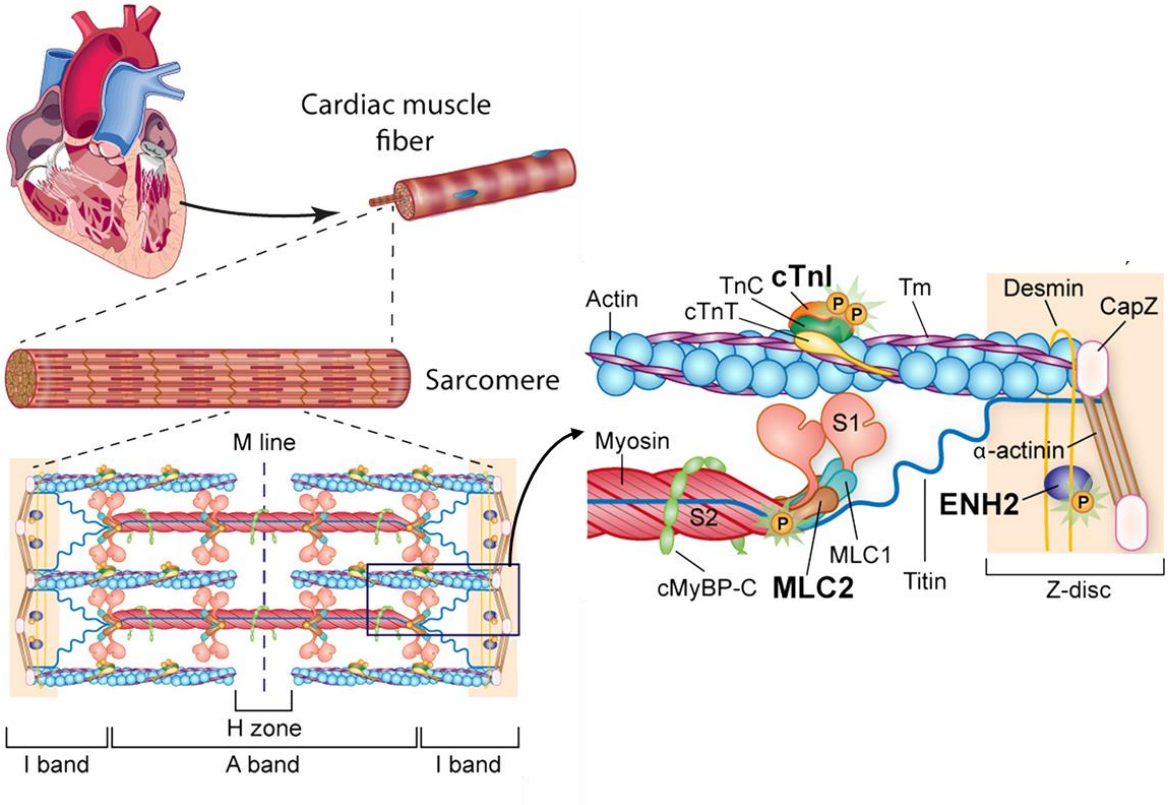


Figure 2 | The cardiac sarcomere | A figure adapted from Peng et al. (2014) which showcases the heart, the cardiac sarcomere, and its components (various bands) on the left as well as its myofilament proteins (actin, myosin and its light chain 2 (MLC2), tropomyosin (Tm), and the

troponin complex - troponin T (TnT), troponin C (TnC), and troponin I (TnI)) and Z-disc proteins (α-actinin, titin, and the Enigma homolog isoform 2 (ENH2)) in more detail on the right.

In cardiomyocytes of the heart, the sarcomere (depicted in figure 2) is the basic unit responsible for contraction. This structure is composed of a series of repeating contractile and regulatory proteins which assemble into thin and thick filaments. In addition to myosin, other thick filament genes include *MYBPC3*, *TTN*, *MYOM1*, and *OBSCN* (Wang *et al.*, 2018) while the thin filament includes *ACTC1*, *TPM1*, and the troponin (Tn) complex (Keyt *et al.*, 2022). Regulatory components essential for contraction are housed in the thin filament, while the thick filament, which is primarily composed of myosin, is where the force originates (Gordon *et al*, 2000).

1.3.2 Core sarcomeric genes

Research conducted in the 1980s and 1990s resulted in groundbreaking revelations that variants in the sarcomere are the underlying cause of HCM (Geisterfer-Lowrance *et al.*, 1990). Rare variants in sarcomeric genes are estimated to be responsible for 40% of HCM cases by Neubauer *et al.* (2019), Thomson *et al.* (2019), Alfares *et al.* (2015), Walsh *et al.* (2017), and de Marvao *et al.* (2021).

One of the first genes associated with HCM was *MYH7* (beta-myosin heavy chain protein) (Geisterfer-Lowrance *et al.*, 1990) which is a prime component of the thick filament within the sarcomere. Subsequent familial studies revealed pathogenic variants in additional genes, encoding other components of the sarcomere, such as *MYBPC3* , *TNNT2* , and *TNNI3* (Watkins *et al.*, 1995)(Thierfelder *et al.*, 1994). With the additional identification of *TPM1*, *ACTC1*, *MYL2*, and *MYL3*, the first eight HCM genes were established. These eight genes continue to provide the most evidence regarding their role in HCM – being responsible for over 90% of genotype-positive cases (Ingles *et al.*, 2019).

1.3.3 Causal variants and their molecular effects

Previously, HCM as a result of sarcomeric variants was attributed to myocyte contractility dysfunction (Thierfelder *et al.*, 1994); however, conflicting mechanisms in force generation were observed (Redwood *et al.*, 1999). Impaired cycling of actomyosin cross-bridges has a central role in the development of HCM (Spudich *et al.*, 2014). A widely accepted understanding regarding the pathogenesis of HCM is that variants lead to an overall increase in contractility, with thick filament variants favouring the transition of myosin molecules from a super-relaxed to an actin-bound state (McNamara *et al.*, 2016) and thin filament variants leading to increased Ca2+ sensitivity (Redwood *et al.*, 1999)(Watkins *et al.*, 2011).

Most causal variants associated with HCM are missense variants. However, variants in the *MYBPC3* gene involve insertion/deletion events and frameshifts resulting in premature truncations (Millat *et al.*, 2010)(Erdmann *et al.*, 2003). Missense variants can alter the amino acid sequence, thus modifying the protein structure and functionality (Marian and Braunwald, 2017). The fate of insertion/deletion variants which result in frameshifts and premature truncations is often nonsense-mediated decay and similarly, misfolded or damaged proteins are processed by the ubiquitin-proteasome system (UPS) resulting in haploinsufficiency (one functional copy of a gene is not sufficient for normal functioning) (Marian and Braunwald, 2017).

Helms *et al.* (2020) examined the phenotypic differences between truncating (91%) and non-truncating *MYBPC3* variants and found that both resulted in comparable levels of hypertrophy with similar clinical effects. While truncating variants appeared to induce haploinsufficiency regardless of their specific location, missense variants were predominantly located in the C3, C6, and C10 domain, with only those in the C10 domain causing haploinsufficiency. Additionally, mutations within the same gene illustrate the ying-yang HCM:DCM relationship (Watkins *et al.*, 2011) which are opposite in phenotype and can be allelic – those affecting adjacent amino acids in the β-myosin heavy chain can result in either HCM (where contractility is enhanced) or DCM (whereby force generation is hindered). This work reinforces the phenomenon whereby variants from distinct regions of the same protein can lead to varying phenotypes (Monserrat, 2018).

Molecular phenotypes vary depending on the gene in which variants are found. For instance, variants in the *MHY7* gene have a greater effect on myofibrillar ATPase activity compared to those in the *MYBPC3* gene (Witjas-Paalberends *et al.*, 2014). The heightened energy demand for tension generation associated with mutant sarcomere proteins leads to a reduced ratio of cardiac phosphocreatine to ATP and consequently, myocardial bioenergetic deficits in HCM cases (Crilley *et al.*, 2003). Additionally, variants affecting thin filament proteins, unlike those in thick filament proteins, lead to an increased sensitivity of myofibrillar ATPase activity to Ca2+ and enhance the force generation capacity which also plays a role in HCM pathogenesis (Lombardi *et al.*, 2008).

1.3.4 Present day HCM genes & genetic testing

At present, the HCM gene panel (as determined through the R131 clinical indication for adults/teenagers which investigates single-gene causes and includes whole exome sequencing (WES) or medium gene panel sequencing) includes 20 genes: *ACTC1*, *ACTN2*, *CSRP3*, *FHL1*, *FHOD3*, *FLNC*, *GLA*, *JPH2*, *LAMP2*, *MYBPC3*, *MYH7*, *MYL2*, *MYL3*, *PLN*, *PRKAG2*, *TNNC1*, *TNNI3*, *TNNT2*, *TPM1*, and *TTR*, along with two variants: p.Arg518 in *CACNA1C* and the mitochondrial variant MT-TI m.4300A>G.

The majority of cases with an identified disease-causing variant are found in one of the eight core sarcomere genes - *ACTC1*, *MYBPC3*, *MYH7*, *MYL2*, *MYL3*, *TNNT2*, *TNNI3* or *TPM1* – these genes were initially identified through linkage analysis and then confirmed through burden testing (Walsh *et al.*, 2017)(Mazzarotto *et al.*, 2019).

Additional genes such as *ACTN2*, *ALPK3*, *CSRP3*, *FHOD3*, *FLNC*, *JPH2*, *PLN*, *PRKAG2*, *TNNC1*, *TTR*, *GLA*, *FHL1*, *LAMP2*, and *SVIL* have also been associated with HCM but variants causing disease in these genes either account for a minority of cases or are difficult to interpret clinically due to a high proportion of variants of unknown significance.

Whole exome sequencing (WES) was employed for exploratory research purposes in cardiomyopathy patients to identify and sequence all protein-coding regions within the genome (Mak

et al., 2018). However, Bagnall *et al.* (2018) found whole genome sequencing (WGS) to reveal additional variants associated with HCM – beyond what targeted gene sequencing methods have previously uncovered. Furthermore, by broadening the scope to include deep intronic regions, pathogenic variants were identified in 9% of cases without a known casual gene (Bagnall *et al.*, 2018).

Advancements in next-generation sequencing (NGS) technologies have improved the availability of extensive whole genome and exome data and have allowed for the testing of gene panels associated with HCM (Gersh *et al.*, 2011). The growing pool of genetic data in addition to the increasing number of HCM families undergoing testing can lead to better classification of pathogenicity for previously identified HCM variants (Jipin *et al.* 2014) as well as better gene-therapy methods to prevent and/or lessen the impact of HCM in affected individuals. One of the most significant and persistent challenges, however, is the insufficient representation of groups with non-European ancestry – both in research cohorts focused on diseases and in population datasets – which restricts the capacity to accurately interpret genetic test results for a considerable part of the global population (Dunn, 2024).

MYBPC3 and *MYH7* are still the most common genes accounting for more than 70% of identified HCM variants while *TNNT2*, *TNNI3*, *TPM1*, *MYL2*, *MYL3*, and *ACTC1* account for between 1 and 5% each (Richard *et al.*, 2003)(Maron *et al.*, 2012). A combination of more than 2000 private and either pathogenic or likely pathogenic variants have been associated with HCM (Cirino *et al.*, 2017). Upon a positive pathogenic sarcomere variant genetic test, cascade screening ensues for the proband's family members at risk (Burke *et al.*, 2016).

1.3.5 Shift from monogenic mindset

The phenotypic characteristics of single gene disorders, especially autosomal dominant conditions, are influenced by genetic elements beyond the primary variant (Marian, 2002). The genetic background, commonly known as modifier genes, plays a role in determining the HCM morphology. A hypothesis was formed and tested by Li *et al.* (2017) to account for the 'missing' or otherwise unidentified variants and genes in a significant portion of HCM cases – whereby some cases (in

smaller families or sporadic instances) may be oligogenic arising from multiple variants, each with small effect sizes, failing to co-segregate with the phenotype. To support their proposition, Li *et al.* (2017) examined a small family in which affected members possessed two or three pathogenic variants in *TTN*, *MYL2*, and *ALPK3* (all previously associated with HCM). They found that the siblings containing three pathogenic variants exhibited a severe phenotype, while those with only one or two displayed either a mild or ambiguous phenotype confirming a collective effect.

Diseases with a genetic component may have various architectures, including monogenic, polygenic, oligogenic, and multifactorial (Tayal *et al.*, 2021). It is estimated that ~5% of HCM cases are digenic or oligogenic (Marian, 2021).

1.4 Polygenic HCM

Around 60% of patients diagnosed with HCM do not have an identified genetic cause (Walsh *et al.*, 2017) which highlights the need for elucidating additional causal HCM genes and/or variants.

Although linkage analyses are incapable of pinpointing complex-trait loci (Visscher *et al.*, 2012), GWASs have unearthed numerous complex-trait loci with strong disease associations (Visscher *et al.*, 2017) and thousands of corresponding common variants (Gibson, 2012). Additionally, some of the variance may be explained by intermediate-effect variants which are too rare to be detected in GWASs yet unnoticed in genetic studies due to their widespread occurrence. For example, two intermediate variants are known to be associated with HCM, p.Arg278Cys in *TNNT2* and p.Glu441Lys in *MYBPC3* (García-Hernández *et al.*, 2024).

Individuals diagnosed with HCM but lacking identified causal variant(s) in sarcomere gene(s) represent the majority of HCM cases (Neubauer *et al.*, 2019) and these scenarios have since been termed “sarcomere-negative”. US and European guidelines advocate for cascade screening accompanied by long-term monitoring of unaffected first-degree relatives. Genetic testing of sarcomere-negative individuals would not provide any insight and so this reflects how the current management of HCM is irrespective of sarcomere status. Additionally, because most research in the

field has been focused on the sarcomere, existing therapies are more applicable to those with a sarcomere positive status (Watkins, 2021). Watkins (2021) has highlighted the necessity for a shift in the rationale and handling of HCM.

Wooten *et al.* (2013) first brought attention to the polygenic aspect of HCM by revealing a common, intronic variant in *FHOD3* which increases the risk of HCM. This finding supported prior research which uncovered common variants linked to dilated cardiomyopathy (DCM) (Villard *et al.*, 2011) – thus reinforcing the significance of common variants in cardiomyopathies.

1.4.1 GWAS discovery

Genome-wide association studies (GWAS) examine multiple variants across diverse genomes with the aim of identifying those linked to a particular disease (Uffelmann *et al.*, 2021). However, their effect sizes range from small to moderate and require sufficient sample sizes for adequate power. Meta analyses of multiple GWASs allow for the detection of increased associations as well as evaluation cross-datasets and/or -populations (Zeggini and Ioannidis, 2009). Multi-trait analysis of GWAS (MTAG) combines GWAS output across various traits while accommodating for sample overlap; it improves the identification of loci as well as the variance explained by polygenic scores, and provides greater insight into biological annotation (Turley *et al.*, 2018).

Polygenic, sarcomere-negative HCM transformed from a mere hypothesis to an experimentally validated form of HCM in 2021 as a result of GWAS findings (Harper *et al.*, 2021)(Tadros *et al.*, 2021). Harper *et al.* (2021) divulged a modest influence of common variants on sarcomere-positive HCM with the variability in penetrance and expressivity being partly owed to such variants spanning multiple loci. In contrast, sarcomere-negative HCM exhibited a more pronounced polygenic inheritance. This explains how ~0.5% of the population have marked polygenic risk of HCM, even in the absence of rare variants, and why high polygenic heritability was seen for cases with affected relatives despite negative gene-panel testing results. The polygenic attribute of HCM also clarifies why there is low recurrence risk observed in first-degree relatives for the bulk of sarcomere-negative individuals.

To explore the role of common variants in HCM risk, Tadros *et al.* (2025) integrated Harper *et al.* (2021) and Tadros *et al.* (2021) HCM GWASs totalling data from seven cohorts (HCMR, BRRD, GeL, Canada, Italy, Netherlands, and Royal Bromptom Hospital) into an overarching meta-analysis and MTAG. The meta-analysis identified 34 loci (of which 15 were novel) and 68 loci via MTAG (of which 50 were novel). Among the common variant HCM loci, *SVIL* was pinpointed as a novel HCM gene in which rare variants also cause the condition. Additionally, Mendelian randomisation (MR) identified an association between an increase in LV contractility and both obstructive and non-obstructive HCM. Together, these findings revealed 68 HCM loci of interest which formed the basis and starting point of this thesis.

Limitations of GWAS

Noncoding genomic variations account for the majority of diseases and trait-associated SNPs. Understanding their functional impact is however challenging (Zhou and Troyanskaya, 2015). Our current inability to interpret noncoding variants hinders any advancements towards using personal genomes in medicine (Kelley *et al.*, 2016). GWAS & meta-analysis results highlight which variants are associated with disease but they do not indicate which variants are likely causal (Walker *et al.*, 2022). Despite the insight that has emerged from GWASs, the results thereof have thus not translated into substantial clinical insights primarily due to the challenges of interpretation.

The majority of the 68 lead HCM variants are located in noncoding regions suggesting that they modulate gene expression levels. Most of these variants are located outside of genes and are therefore expected to affect regulatory elements. Such regulatory elements can include promoters, enhancers, silencers, insulators, untranslated regions (5’ or 3’ UTRs), sites for the synthesis of RNA molecules, and CpG islands (Vitsios *et al.*, 2021). However, some of these variants are situated within noncoding regions within a particular gene, such as its introns. These variants could affect said gene’s splicing mechanisms or internal machinery in addition to the described regulatory elements.

On both accounts, regardless of whether a GWAS variant resides within or outside of a gene, along with the consideration of variants being cis- or trans- acting and the nature of these regions being

noncoding – drawing associations between GWAS variants and disease (or a trait) does not help to illuminate plausible target genes or the underlying mechanisms which result in phenotype changes (Visscher *et al.*, 2017).

Additionally, the assigned lead variants post-GWAS are not necessarily casual despite showing the most significant association with the trait of interest. In a specific locus, a variant's assigned significance may actually reflect that of the true causal variant which just happens to reside nearby (transferred signals). Fine-mapping approaches applied to GWAS data struggle to identify the causal variant due to the distribution of causal effect sizes among numerous variants in high linkage disequilibrium (LD) – collectively these form an LD block whereby specific alleles co-occur more frequently than would be expected by chance (Long *et al.*, 2025). Likewise, variation in the quality of variants' genotype imputation can give rise to altered association signals of variants in LD (Uffelmann *et al.*, 2021). Consequently, deducing causal variants poses a challenge too.

1.4.2 Advancement in technologies

Regulatory variants have a high evolutionary turnover and as a result the consequence is not binary; regulatory variants can sway the expression level of gene(s) in either direction. These variants' effects have been shown to vary depending on the tissue or cell, the stage of development, and across different individuals.

In light of the influx of NGS data, chromatin features capturing regulatory landscapes can be explored in cells/tissues of interest (Ritchie *et al.*, 2014).

Regulatory variant prioritisation using GWAS variants as a starting point generally entails the intersection of GWAS signal with genomic features such as eQTLs (expression quantitative trait loci), transcription factor binding sites, DNase hypersensitive sites, and histone modifications (Hou and Zhao, 2013). This is explained by the findings that cis-eQTL SNPs are enriched in known disease-associated SNPs, trait-associated SNPs in the GWAS catalogue are more frequent within

DNase Hypersensitive sites, and condensed genomic regions annotated with little expression are deficient in trait-associated SNPs.

Various databases were created to elucidate the regulatory landscape and how noncoding variants may affect these (Ritchie *et al.*, 2014). ENCODE was established in 2003 in an effort to identify all functional elements (both coding and noncoding regions, regulatory elements, and sequences that are important for chromosomal structure) and similarly in 2008 the NIH Roadmap Epigenomics made human epigenomic data available. Web servers were set up to perform SNP prioritisation post-GWAS such as FASTSNP (Yuan *et al.*, 2006) which employs a decision tree framework to examine the genomic location and functional effects, as well as FitSNPs (Chen *et al.*, 2008b) which makes use of associated genes' differential expression ratios. A web-based tool called SPOT (Saccone *et al.*, 2010) show-cased another approach of scoring candidate SNPs via multiple evidence integration from biological databases such as dbSNP, ECRBase, HapMap, and PolyPhen into a genomic information network (GIN) model. More advanced classifier-based approaches which retrieve data from a larger set of databases and with more up to date data later arose such as MutationTaster (Schwarz *et al.*, 2010) which can provide predictions for noncoding variants that can be mapped to a transcript model, such as those in untranslated regions and introns via a naive Bayes classifier and then GWAVA (Shihab *et al.*, 2015) which also provides predictions for noncoding variants via a random forest algorithm. Advanced mining tools such as HaploReg and RegulomeDB are aimed at assigning noncoding variant annotation that can aid in SNP prioritisation through the integration of evidences from ENCODE, the NIH Roadmap Epigenomics, GTEx, 1000 Genomes Project, GEO, and also from scanning the literature.

Annotated regulatory regions allow for the identification of loss of function variants; however, deep learning algorithms also aid in identifying *de novo* gain of function variants as they offer an alternative, more advanced means of predicting the effect of noncoding variants due to their being able to capture the biological relationship between DNA and its features (Lan and Corces, 2022).

1.4.3 Deep learning applied to the regulatory landscape

Deep learning is a type of artificial intelligence that uses artificial neural networks (ANNs) to train computers to replicate the human nervous system and its neuron functioning (Vaz and Balaji, 2021). ANNs operate via neurons in the form of hidden layers for input data processing to produce results. The input is received by the input nodes which then calculate a nonlinear relationship between the input and output and generate a solution for the task at hand. ANNs can be convolutional or recurrent. In convolutional neural networks (CNNs) and as shown in figure 3, each processing unit i.e. node (which exist within model layers) identify local features in the input data. Ultimately the contribution of all features is then determined (across all layers and their individual nodes) and output in a final layer. CNNs are made up of convolutional, pooling, and fully connected layers. Simple patterns are recognised at each CNN's core – leading to more complex patterns as depth increases. The presence of numerous hidden layers avoids having to use increasingly large parameters. Two main attributes at the core of CNNs are weight sharing and local connectivity. The former infers that each node at a specific layer shares the same weight and the latter that each node will acquire input from only a handful of local values so each node's output relates only to a fraction of the output.

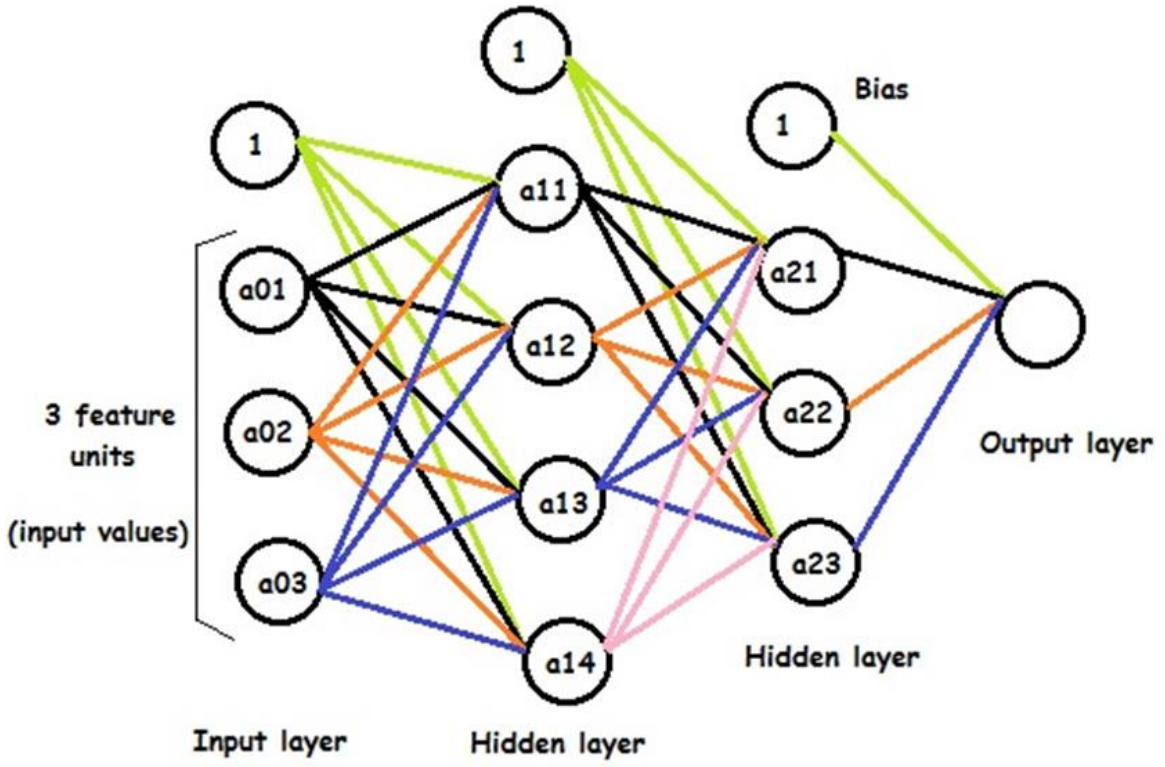


Figure 3 | A neural network with fully connected layers | A figure adapted from Singh (2019) illustrating a neural network with fully connected layers whereby the initial input layer comprises of 3 feature units (nodes) whereby a01, a02 and a03 are the input values, the next hidden layer has 4 activation units, and the 4 activation units of the first hidden layer are connected to 3 activation units of a second hidden layer.

CNNs can learn features from data used during training as opposed to needing prior manual selection. Applications of CNNs to DNA sequence analysis have surpassed more established algorithms, such as support vector machines or random forests, at predicting protein binding and accessibility from DNA sequence (Alipanahi *et al.* 2015; Zhou and Troyanskaya 2015). After training the model, the effect of a genetic variant is predicted through testing both alleles separately and inspecting any differences in the model's prediction based on the single-base change (Kelley *et al.*, 2016). Large effects are presumed to correlate with functional alterations of transcription factor (TF) binding which could theoretically influence gene expression and downstream cellular phenotypes. This approach allows one to home into a handful of variants at a finer resolution to prioritise those predicted to drive the regulatory activity while omitting others in a locus of interest. This also paved the way for understanding gain of function regulatory variants – which do not often overlap with

regulatory annotation and are largely undocumented in so-called ‘dead’ chromatin regions (as opposed to regulatory variants which disrupt known transcription factor motifs, for example (Currin *et al.*, 2021)).

Noncoding regulatory variants can affect different aspects of gene regulation such as chromatin interactions, TF binding site affinity, transcriptional activity of genes, post-transcriptional regulation, and the initiation of translation. All such interferences could either increase or decrease the chromatin accessibility, or have no effect whatsoever (Peña-Martínez and Rodríguez-Martínez, 2024).

To understand the possible impact of a noncoding variant on the regulatory landscape, scoring methods have been devised. *In silico* mutagenesis (ISM) is one process of scoring the change in prediction after altering a variant; the reference allele’s score is compared to the alternate allele’s score in the form of a maximum difference. Higher scores indicate a greater functional impact while lower scores suggest there would be little difference in effect (Lan and Corces, 2022). ISM is computationally intensive as all 3 possible mutations (homozygous dominant, homozygous recessive, and heterozygous) need to be assessed per SNP in a particular sequence. Additionally, noncoding sequences can possess redundant marks where only one may be important for gene expression and ISM would not identify the important feature (Movva *et al.*, 2019).

Consequently, when assessing variants with a notably high (‘gain of function’) or low (‘loss of function’) ISM score, one can infer that the chromatin’s accessibility will be affected at these regions (Kelley *et al.*, 2016). Significant loss of function variants point to an open region of chromatin / an existing functional motif where a variant’s presence may disrupt the motif and consequently decrease accessibility. In contrast, gain of function variants can reside in either a closed or open region of chromatin where the introduction of a variant can often result in, or enhance, a functional motif and therefore lead to an increase in accessibility.

Hypotheses and Key objectives

The aims of this thesis are rooted in the recent discovery and expansion of the polygenic basis of HCM derived from the most recent GWAS. Below is a list of core objectives, which contribute towards unveiling the effects and consequential mechanisms elicited by common variants in causing HCM.

- Defining the locus and leveraging bulk and single cell expression data to prioritise gene(s) within each locus (Chapter 3)

  This approach entailed using the GWAS dataset in addition to regulatory data (topologically associating domains (TAD) data specifically) to guide the demarcation. Thereafter, various gene prioritisation methods were employed to rank genes according to their likelihood of HCM association.
- Performing variant prioritisation via machine learning methods using the original set of HCM lead variants in addition to their proxy variants (Chapter 4)

  Because the majority of these HCM associated common variants reside in noncoding parts of the genome whereby they have an impact on the chromatin landscape – machine learning models trained on chromatin data were used to identify functional variants and assess their consequences.
- Gaining greater insight into the prioritised variants and computationally linking them to plausible target genes (Chapter 5)

  Once I had finalised a list of likely prioritised genes in addition to likely functional variants, methods were employed to gain perspective around the likely mechanisms involved. This was covered in more detail for loci where specific variants had been shown to be likely functional.

2

Materials and Method

2.1 Introduction

This chapter outlines the materials and methodologies employed in this thesis (as illustrated in figure 4). It includes information about the datasets used and bioinformatic tools and pipelines executed.

See the GitHub repository (https://github.com/domwest/DPhil-thesis) for supplementary material and data tracking referred to throughout the thesis.

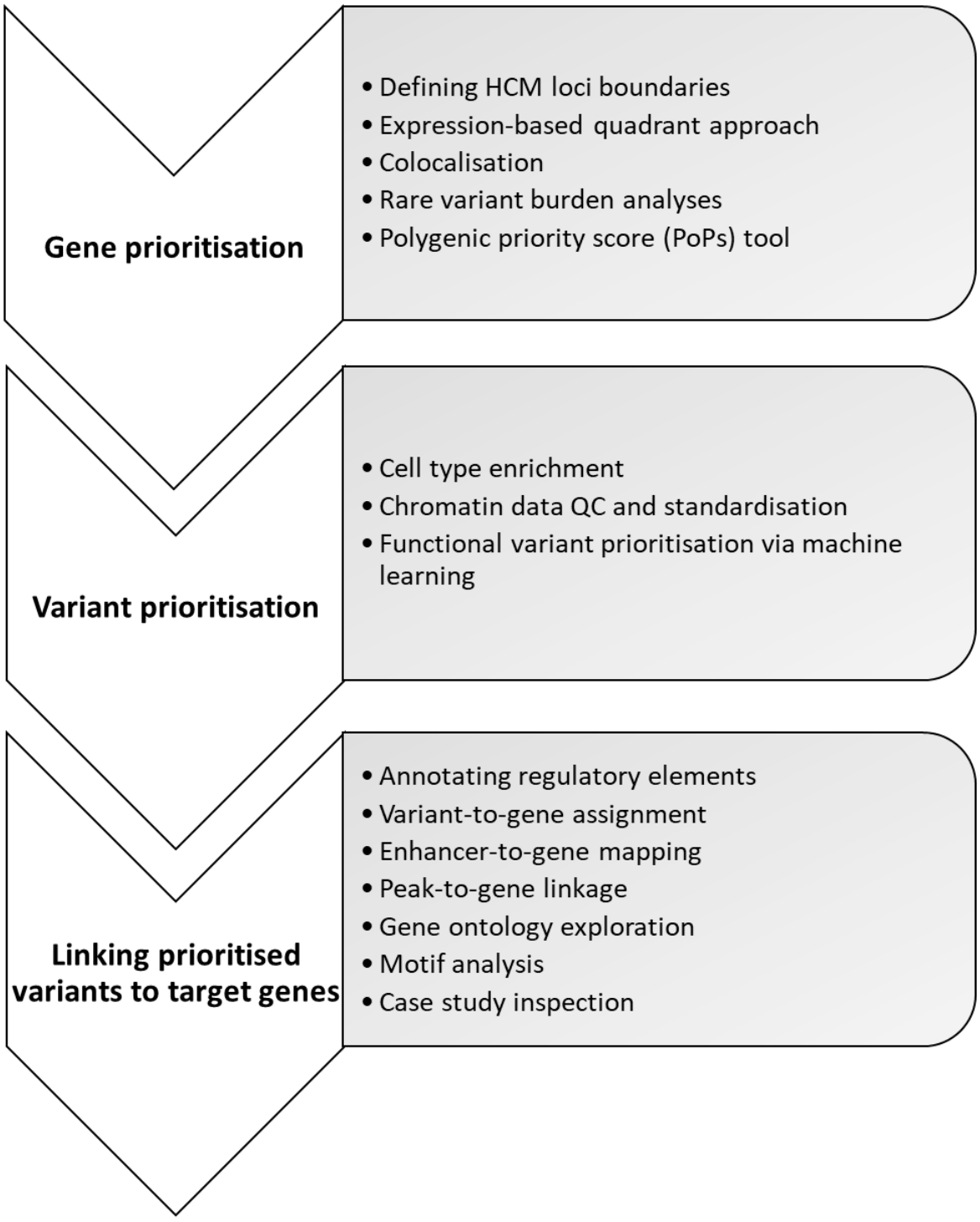


Figure 4 | Thesis methodologies summarised | A flow diagram illustrating all methodologies applied throughout this thesis in an attempt to tackle the three key objectives (gene prioritisation, variant prioritisation, and linking prioritised variants to target genes).

2.2 Gene Prioritisation

2.2.1 Overview

This thesis utilised the lead variants for each of the 68 loci from the HCM GWAS summary input data (Tadros *et al.*, 2025) which is provided at GitHub (https://github.com/domwest/DPhil-thesis/blob/main/HCM_GWAS_summary_stats.xlsx). Throughout this thesis, the "HCM lead variants" refers to those from the 68 MTAG loci as reported by Tadros *et al.* (2025). Intact Hi-C data was downloaded from the Encyclopedia of DNA Elements (ENCODE), both bulk expression data and eQTL data from the Genotype-Tissue Expression (GTEx) v7 database, and single-cell expression data from Reichart *et al.* (2022).

2.2.2 Defining HCM loci

My analyses utilised the HCM genome wide association summary statistics (representing 68 lead variants at genome-wide significant loci) as the starting point from which to define regions associated with the signal.

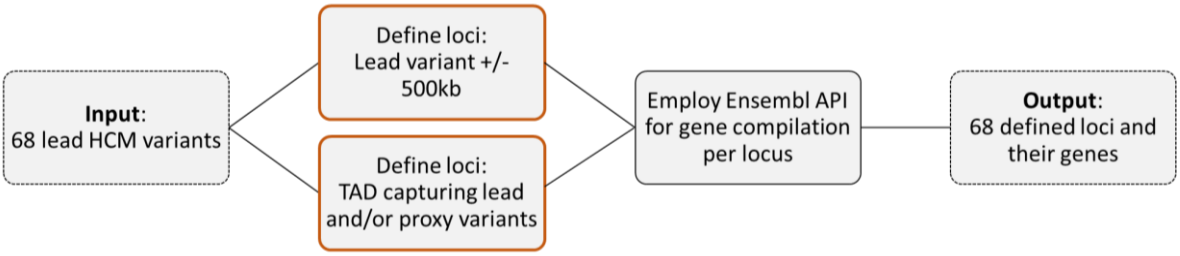


Figure 5 | Defining HCM loci | A flow diagram displaying the process of defining HCM loci regions given a set of HCM lead variants as input. Both a proximity-based and topologically associating domain (TAD)-based approach were employed.

Proximity-based approach

For each of the 68 lead variants identified in the HCM GWAS, loci regions were defined by scanning 500 kilobase pairs (kbp) both upstream and downstream of the lead variant. The chosen span of 1 megabase pair (Mbp) per locus was an arbitrary selection made to ensure that the likely causal gene was included.

The Ensembl application programming interface (API) was utilised to query these search regions, resulting in a compilation of Ensembl stable gene IDs residing in each locus.

TAD-based approach

For each of the 68 lead variants identified in the HCM GWAS, loci regions were defined through the use of topologically associating domains (TADs).

Intact Hi-C *Homo sapiens* heart left ventricle tissue (ENCODE ID: ENCFF745WPP) contact domains were downloaded in .bedpe format from ENCODE. TADs are annotated in Hi-C maps via the identification of 'corner dots' – a cluster of neighbouring pixels with an increased interaction frequency compared to the surrounds, representative of long-range looping interactions. Hi-C mammalian maps have revealed megabase-scale nested TADs (those contained within another) as well as unnested TADs (those of the highest order) loop domains. The majority are anchored by motifs associated with the CCCTC-binding factor (CTCF) (Beagan and Phillips-Cremins, 2020). Corner dots can be seen in the Juicebox plot (a visualisation software for hi-c data (Robinson *et al.*, 2018)) shown in figure 6 in the form of concentrated red pixels.

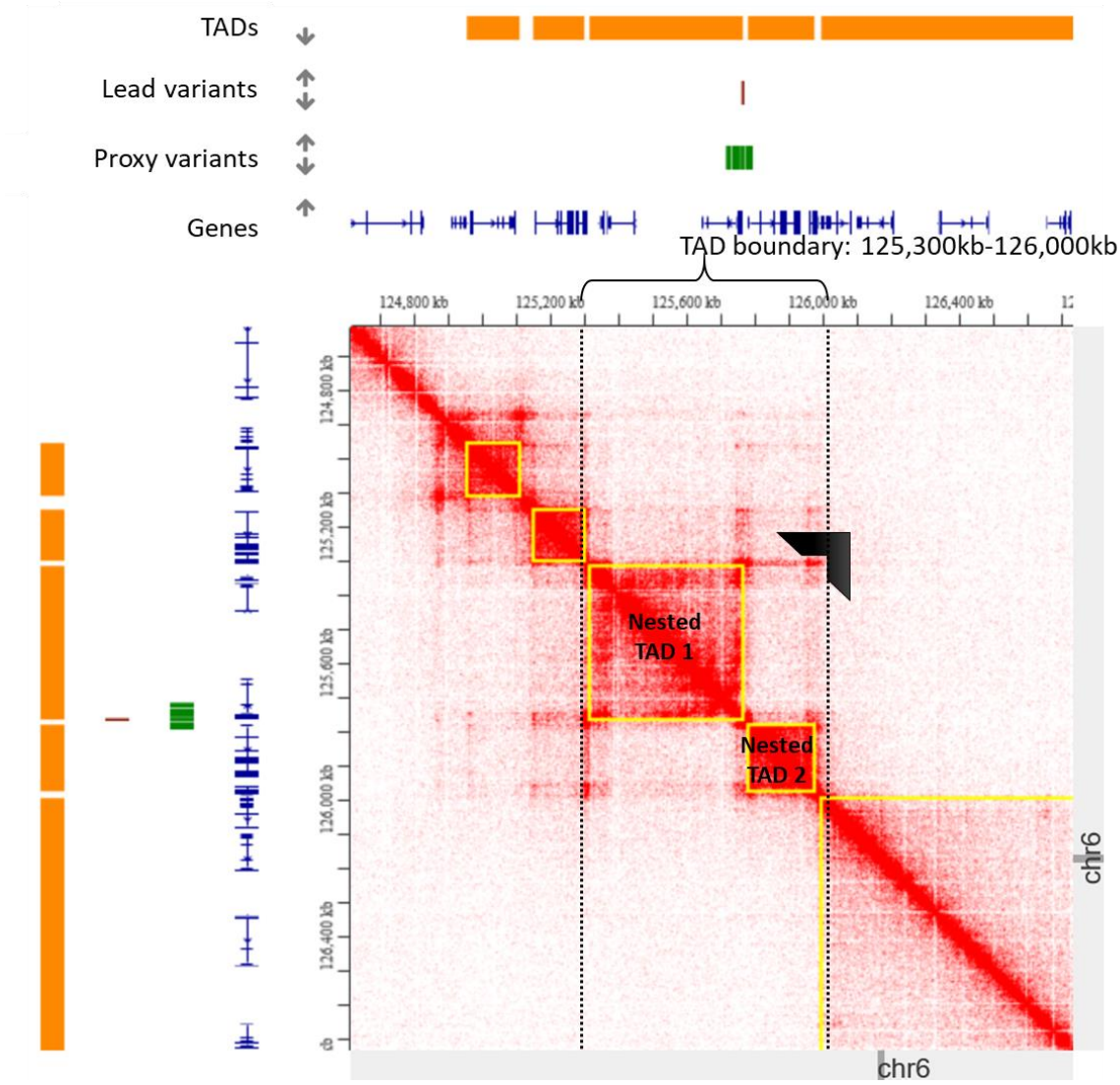


Figure 6 | Assigning TADs in Juicebox | A Juicebox visualisation illustrating the manual assignment/assessment of topologically associating domains (TADs) derived from heart left ventricle contact domain data. The top and left-hand panels show the following tracks: TAD regions in orange, HCM GWAS lead variants in red, proxy variants in green, and genes in navy blue. The intact hi-c contact domain matrix (in the main grid) also shows the assigned TADs in yellow boxes. The black half frame shape highlights a corner dot which represents an un-nested TAD capturing two nested sub-TADs within.

The 68 HCM lead variant positions were intersected on the TAD domains resulting in only 40 loci being captured within TADs (figure 7A). In light of lead variants being excluded from TADs and to ensure the delineation of all 68 loci regions, proxy variants were leveraged; The R package,

LDLinkR (using the genome build37 database), was employed to retrieve variants in LD with the 68 lead variants.

R2 is a measure of the association between alleles and the larger this value the stronger the association between 2 alleles. If R2 ≥ 0.8 then this indicates that two specific SNPs are in linkage disequilibrium (LD) i.e. inherited together more frequently than expected by chance and are therefore not independent from each other (Mu and Zhang, 2013). This threshold was chosen as a result to establish which variants were likely being inherited along with the GWAS lead variants identified, so as to increase the search space and also ensure possible causal variants were not being excluded.

This resulted in a total of 76104 variants annotated by chromosome, position, reference allele, and alternative allele (snp_test_ids) in the output – proxy variants were captured across 67 of the 68 loci; however, none for rs28768976. The variants were then filtered according to their extent of LD (those with an R2 score greater than 0.8 were retained) resulting in 2177 total snp_test_ids (figure 7B). The intersection of 2177 lead and proxy variants with the TADs resulted in 1013 being captured; some variants were captured in more than 1 assigned TAD. The 1013 captured variants corresponded to 46 of the lead rsids (figure 7C).

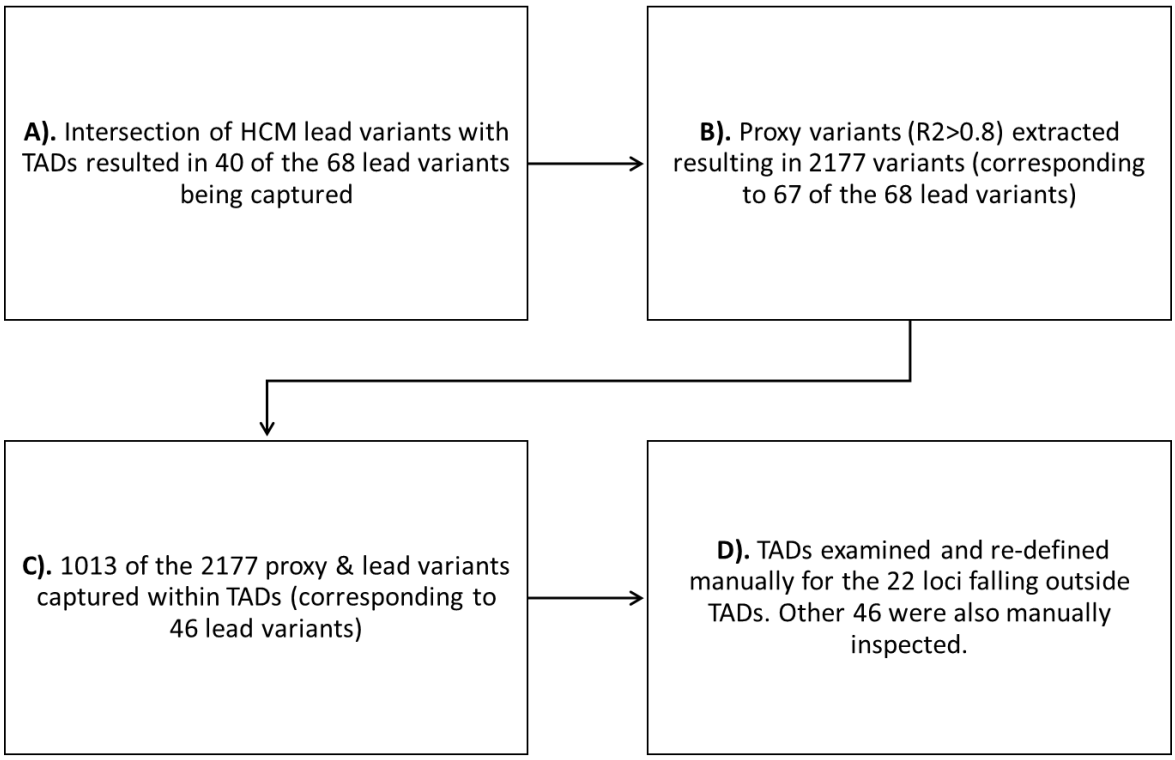


Figure 7 | Intersecting HCM variants & TADs for loci demarcation | A flow diagram showing the process and counts regarding loci demarcation via the intersection of lead and proxy HCM variants with heart left ventricle contact domain data containing topologically associating domains (TADs).

For each of the 46 loci where either the lead or proxy variants were captured in TAD(s), in the event of duplicated (overlapping) TADs for a particular variant, the most gene-inclusive and thus the largest TAD was chosen via visual inspection in Juicebox. For those of the 46 loci whereby only 1 TAD was assigned, visual inspection still ensued so that in the event this being a nested TAD with a further corner dot being present (without being annotated as a TAD), the designated locus could be extended. An example of this can be seen in figure 6 whereby the lead and proxy variants are contained within two TADs which both form nested TADs within a greater TAD – the greater TAD was chosen to demarcate this locus. For each of the 22 loci where neither lead nor proxy variants were captured within TAD(s), the variants were pinpointed in Juicebox and the most likely TAD derived manually.

Fetch genes for each HCM locus

After the 68 loci coordinates were determined, the Ensembl API was utilised to query these search regions. This resulted in a compilation of Ensembl stable gene IDs per locus.

2.2.3 Devised Expression-Based Quadrant Approach

Overview

I devised a 'quadrant approach' method aimed to inspect both tissue/cell type expression and specificity of genes compiled within each of the 68 HCM loci. Due to the fact that this tool was performed twice (first for loci defined via a proximity-based method with the use of bulk heart left ventricle expression data and then for loci defined via TADs using single-cell cardiomyocyte expression data), a flow chart detailing the common overarching steps as well as those specific to the two analyses is shown in figure 8. Additionally, a scatter plot example of how this tool's output was visualised per locus is shown in figure 9.

After performing the quadrant analyses, HCM positive control genes were defined and their loci used for evaluation. As genes with an empirical connection to HCM regarding regulatory effects have yet to be established, the HCM positive control genes selected were BAG3, SVIL, HSPB7, FHOD3, and ALPK3. These were chosen on the premise that they achieved genome-wide significance regarding common variants in addition to rare variant findings having been established previously.

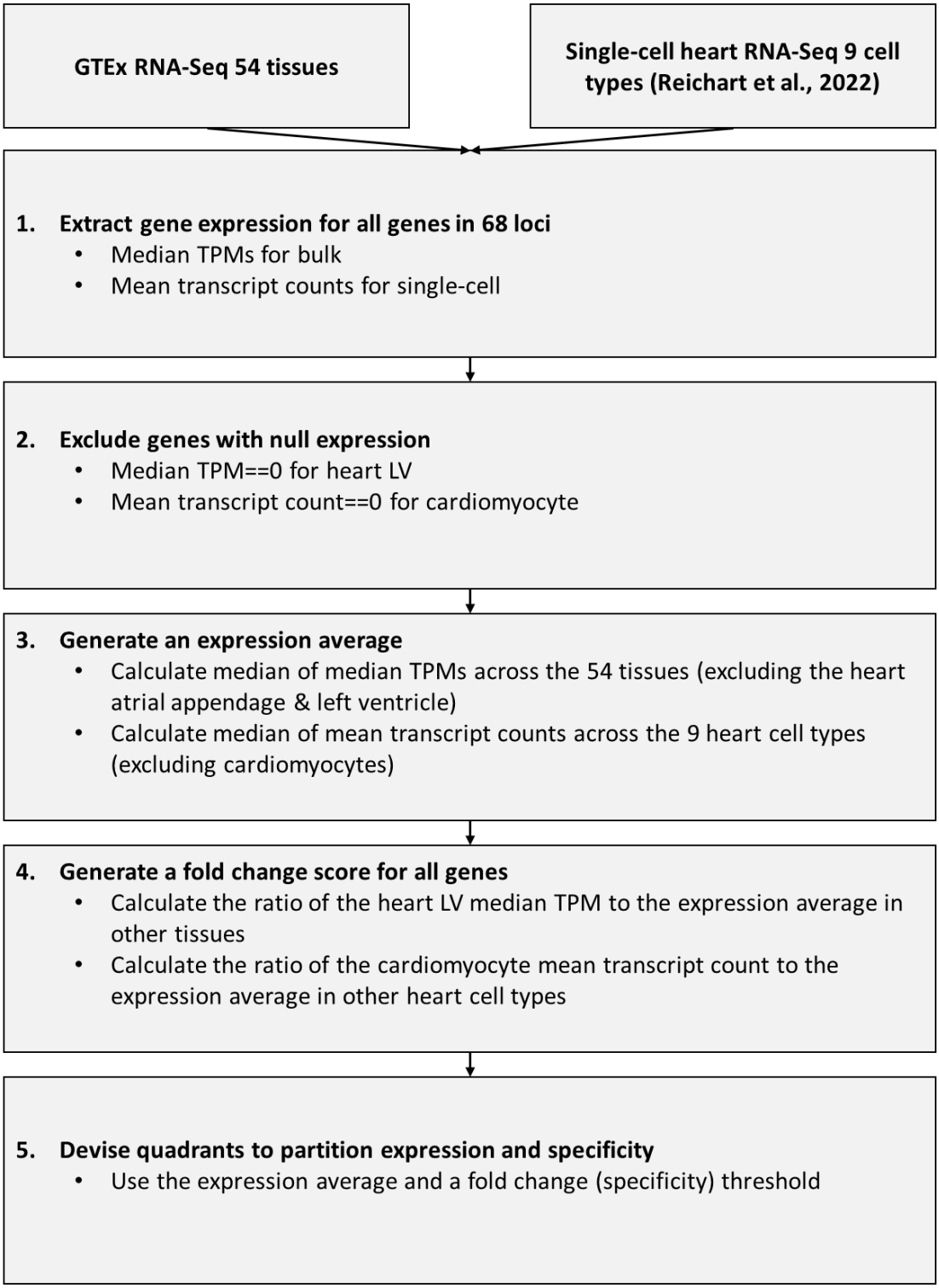


Figure 8 | Quadrant approach pipeline | A flow diagram depicting the devised expression-based gene prioritisation ('quadrant approach') pipeline. The pipeline contains 5 overarching steps explained according to use of bulk tissue heart left ventricle proximity-based approach versus single-cell cardiomyocyte topologically associating domain (TAD)-based approach.

Step 3 (figure 8) entailed the inclusion of all genes in the heart LV as well as the cardiomyocyte (with non-null expression) for the generation of an overall average expression value. Regarding step 4, resulting infinite fold change values (for genes in the heart left ventricle (LV)/cardiomyocyte with an expression value > zero but with an average expression across other tissues/cell types equal to zero) needed to be addressed. In these scenarios, the fold change value was replaced with the maximum fold change value assigned to a given gene within the particular locus. Regarding step 5, the fold change threshold was determined from prior lab experience to be 1.3 and the expression threshold was calculated from GTEx expression data whereby all genes in the heart were included in the calculation (not only those residing within the 68 loci); however, genes with null expression values (equal to zero) were excluded.

Overarching steps following 1-5 are common to both approaches. They involved devising a score which combines both expression and specificity so as to rank genes within each quadrant (figure 9). The diagonal distance within each quadrant was calculated (which considers both the expression and fold change score). Genes were then prioritised within each locus – first based on the quadrant in which they resided (in the order of quadrants 2, 1, 4, and 3) and then based on their diagonal distance within a quadrant. The genes' expression and specificity scores were then visualised per locus whereby specificity values were plotted on the y-axis and expression values on the x-axis (forming a scatter plot). For the quadrant approach with single-cell cardiomyocyte expression data, each gene's expression was compared to that in the other heart cell types. The mean transcript counts for a particular gene across all 9 cell types was assessed and that with the highest value was noted to be the top cell type.

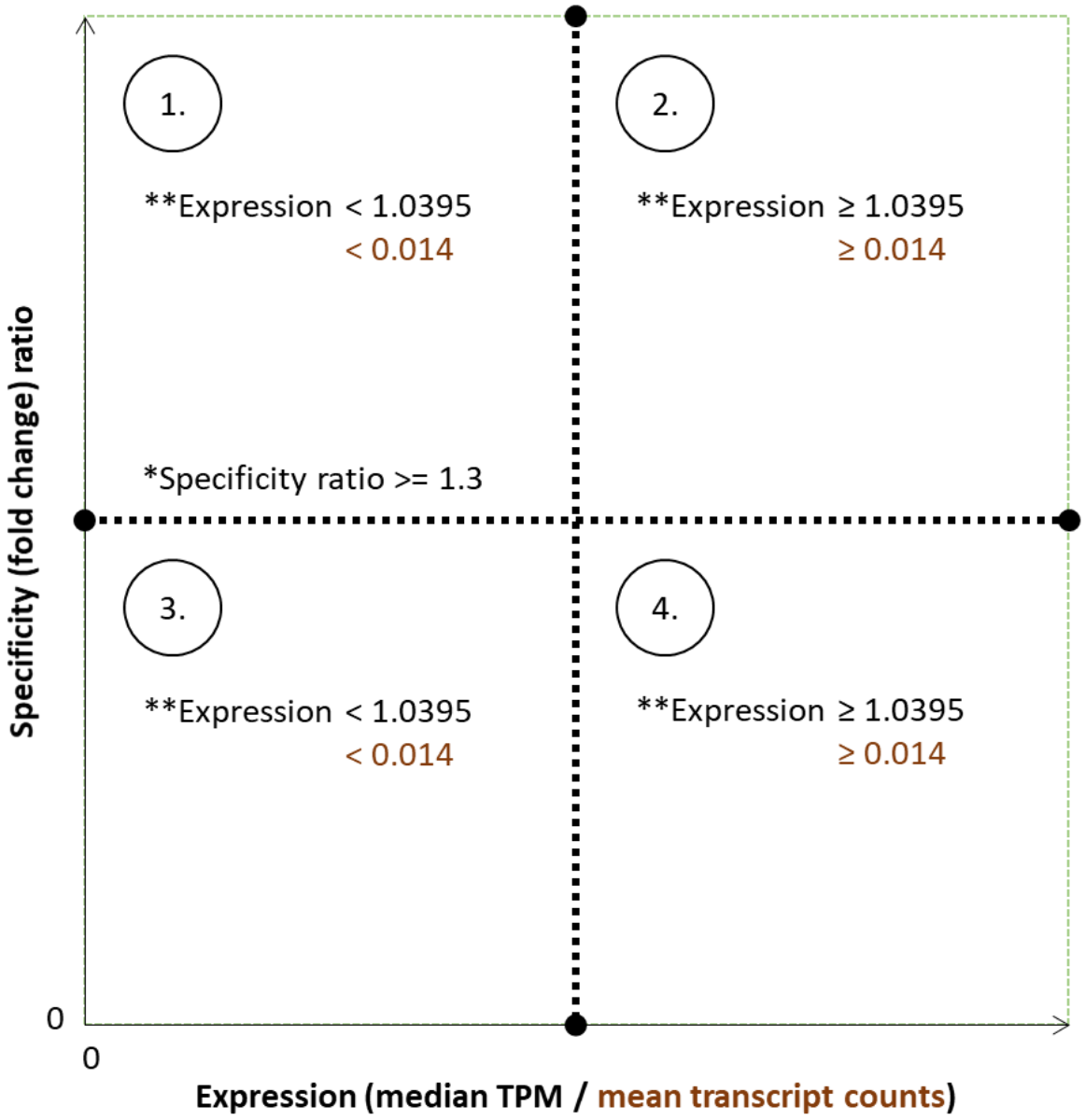


Figure 9 | Quadrant approach visualisation | A scatter plot illustrating the devised expression-based gene prioritisation ('quadrant approach') visualisation whereby the specificity (fold change) ratio is plotted against expression (in either the heart left ventricle (LV) (where the expression threshold is shown in black) or cardiomyocyte (expression threshold in dark orange) one locus at a time. Genes are prioritised in the order of quadrants 2, 1, 4, 3.

**Specificity ratio cutoff of 1.3 in both A and B based on prior lab experience.*

***Expression cutoff of 1.0395 determined by median of all expressed genes' median TPMs in the heart LV and 0.014 by the median of all expressed genes mean transcript counts in cardiomyocytes.*

Bulk proximity-based quadrant approach

GTEx bulk heart left ventricle expression data

The exploration of bulk tissue expression profiles has been facilitated by resource databases such as the GTEx portal whose aim is to clarify tissue-specific gene expression and the regulation thereof. Regarding expression data, the GTEx database contains RNA sequencing (RNA-Seq) data, cis- and trans-QTL data, as well as histology images. The v7 release utilised in my analysis comprises 11688 RNA-Seq samples spanning 53 tissues (from 714 healthy donors) and capturing 56202 unique gene IDs. Among the 53 tissues is heart LV as well as heart atrial appendage (AA) data, of note.

Analysts at the Broad Institute performed the quality-control (QC) of GTEx expression data before making it publicly available. RNA-Seq expression outliers were excluded via a multidimensional adaptation of the method employed by Wright *et al.* (2014). Samples with less than 10 million mapped reads were discarded and samples capturing the highest number of reads were retained when faced with sample replicates. For each tissue, expression values were normalised by filtering genes for expression values greater than 0.1 transcripts per million (TPM) as well as more than 6 reads in at least 20 percent of samples. Between-sample normalisation was carried out via the trimmed mean of m-values (TMM) and then per gene via an inverse normal transformation. Given the between-sample normalisation having already been performed, median TPMs per gene in a particular tissue were downloaded from GTEx for this analysis.

Lovén *et al.* (2012) stated that no computational method is applicable for reliable between-sample normalisation across the diverse tissues in GTEx and that only qualitative measures of relative expression levels can be deduced as a result. Therefore, in this analysis the median of median TPMs was used to represent the expression of a particular gene across numerous tissues (to account for between-tissue TPM differences). The median was chosen as opposed to the mean to avoid extreme TPMs in particular tissues from skewing the data.

Bulk heart left ventricle quadrant approach pipeline

The intersection of the gene IDs in the Ensembl API output (1977 genes) with GTEx gene expression data resulted in 1895 genes across the 68 loci. Regarding steps 1 and 2 (figure 8), the exclusion of genes with a median TPM in the heart LV of zero filtered this genes list down to 1115 gene IDs.

For step 5 in figure 8, the total set of 56202 unique gene IDs present in the GTEx heart LV dataset were reduced to 25202 after removing genes with null expression. The median of median TPMs across all 25202 genes was 1.039.

Single-cell TAD-based quadrant approach

Single-cell cardiomyocyte expression data

Reichart *et al.* (2022) performed single nucleus RNA sequencing (snRNA-Seq) of full-thickness LV and right ventricle (RV) samples. Their aim was to characterise the transcriptome of 880000 nuclei from 18 control and 61 failing, non-ischemic human hearts (with pathogenic variants in DCM and arrhythmogenic cardiomyopathy (ACM) genes or idiopathic disease). Because the failing hearts are not representative of HCM in addition to the fact that this gene prioritisation approach is based on which genes are typically expressed/significant in the cardiomyocyte, the control data (as opposed to diseased) was utilised. The metadata was consulted for linking control clinical metadata IDs to their donor IDs with a final tracking sheet shown below in table 1.

Mean transcript counts were calculated for a particular cell type and sample (which represented a specific individual, lab, and tissue sub-section). This was carried out to account for within-sample transcript count differences. Thereafter, the median of mean transcript counts was calculated to represent the expression of a particular gene across numerous cell-types (accounting for between-cell type transcript count differences).

Table 1 | Single-cell RNA-Seq data samples & labs | A table depicting the associated donor IDs, RNA-Sequencing (RNA-Seq) file names (H5 file), and the corresponding samples and laboratories (labs) applicable to each of the 18 control (non-failing heart) individuals included in the single cell topologically associating domain (TAD)-based gene prioritisation quadrant approach. The control single-cell RNA-Seq metadata as well as datasets were sourced from Reichart et al. (2022). Lab center: BO, Bad Oeynhausen; BS, Boston; HCA: Heart Cell Atlas

Donor ID	**Sample**	**Lab**
HCAHeart7664652	LV	HCA
HCAHeart7664653	LV	HCA
HCAHeart7829978	LV	HCA
HCAHeart7835148	LV	HCA
HCAHeart7880860	LV	HCA
HCAHeart7888922	LV	HCA
BS_H26	LV0	BS
BS_H25	LV0	BS
BS_H37	LV0	BS
BS_H15	LV0	BS
BS_H20	LV0	BS
BS_H35	LV0	BS
H46	LV0	BO
H49	LV0	BO
H49	LV1	BO
H51	LV0	BO
H51	LV1	BO
H53	LV0	BO
H55	LV0	BO
H67	LV0	BO

Reichart *et al.* (2022) generated scRNA-Seq data for a particular individual in multiple labs as well as for various replicates due to plenty of tissue being available. Regarding the former, this explains

why table 1 shows 3 different labs (HCA, BS, and BO) for a particular ID and sample at times and, regarding the latter, why different left ventricle samples are shown (LV0 as well as LV1) for the same ID and lab. Important to note that the authors did not have all regions available for all samples; after performing regional analysis to determine how comparable heart left ventricle free wall axis (LVFW-AX) or LV-Septum were, they found similar effects and thus jointly report free wall, apex and septum as 'LV'. The 20 files shown in table 1 were utilised for the single-cell TAD-based quadrant approach.

In light of the control individuals' data being collected from various labs and for multiple replicates, choosing to run the downstream analysis using all controls' data or a subset of these files instead was investigated. The mean transcript count distribution across the various heart cell types for a specific individual (and LV sample) but collected from different labs was inspected as well as the mean transcript counts for a specific individual (and lab) yet different LV samples. Although there were differences between the samples for a particular individual in either different labs or different types of LV tissue – this was not extreme and with the aim being to use averages across the control group of individuals, the decision was to utilise all 20 control samples in the analysis.

A Global_all.h5ad meta dataset was provided for linking cell barcodes in each of the individuals' RNA-Seq filtered matrices to their corresponding cell types (where the unique cell types captured were Mural, endothelial cells (EC), fibroblasts (FB), Lymphoid, neuronal cells (NC), Myeloid, Mast, and adipocytes (AD)) and this was leveraged to create a dictionary for mapping. I then concatenated each of the 20 controls' file matrices into a single matrix along with the relevant cell type barcodes.

Single-cell cardiomyocyte quadrant approach pipeline

The intersection of genes in the Ensembl API output (1636 genes) with scRNA-Seq data resulted in 1034 genes being captured across the 68 loci. Regarding steps 1 and 2, the exclusion of genes with a mean transcript count of zero for cardiomyocytes filtered this genes list down to 892 gene IDs.

Step 2 in figure 8 highlights the use of mean counts as opposed to the median (where the latter was utilised in the bulk tissue quadrant approach analysis). This is primarily due to most fold changes being null; the values were null because the median transcript counts for the heart cell types per gene were mainly zero. This can be explained by the nature of scRNA-Seq data being sparse (and with the mean being more susceptible to extreme values it is more likely to veer away from zero). Haque *et al.* (2017) describe the challenges associated with scRNA-Seq data whereby zero counts arise from true biological non-expression as well as from "dropout" events (expression being missed due to technical limitations). Technical limitations arise regardless of the sensitivity of the platform due to the sparse nature of the data whereas biological limitations are rooted in the on/off nature of expression. Additionally, transcriptional bursting also contributes to the frequency of zero observations. Thus, mean transcript counts were utilised for this analysis with single-cell data.

In step 5 of figure 8, the total set of 33538 unique gene IDs present in the cardiomyocyte scRNA-Seq data were reduced to 28072 after removing genes with null expression. The median of mean counts across all 28072 genes was 0.014.

In addition to generating cardiomyocyte expression and specificity scores per gene, the cell type in which each gene was most highly expressed (highest mean transcript count) was accounted for.

Combining bulk and single-cell quadrant approach output

Before comparing prioritised genes per locus via the two different quadrant approaches, genes were refined to include only those captured in the curated RefSeq genes list (Pruitt *et al.*, 2007). This was done to ensure the inclusion of reliably annotated genes only.

The final prioritised genes in each locus were investigated. Those of note were loci where the same gene was prioritised in both approaches, as well as those where different genes were prioritised and the single cell prioritised gene was most expressed in cardiomyocytes or alternatively in a different heart cell type.

2.2.4 Assessing Regulatory Roles Through Colocalisation of Common Variants and eQTL Data

Overview

Colocalisation analyses were performed for the genes prioritised thus far in the thesis. In particular, genes prioritised by Open Target and FUMA (for the lead variant in each HCM locus previously performed in Tadros *et al.*, (2025)) in addition to the top prioritised gene via the quadrant approach (using both bulk heart left ventricle tissue as well as single-cell cardiomyocyte data). This resulted in an input list of 168 gene IDs for the 68 loci.

In order to perform colocalisation, both eQTL and GWAS data was required. These input files contained the following columns: 'beta', 'varbeta', 'snp', 'position', 'MAF', 'pval_nominal'. Each genes' eQTL input file was filtered to include only those captured within the TAD for the relevant locus. Likewise, the HCM GWAS dataset was split into 68 separate input files per TAD-derived locus. In both datasets, variants lacking GWAS data were dropped and in the event of rsid duplicate entries, the entry with the highest minor allele frequency (MAF) was kept. Variants captured within both the eQTL and GWAS data with flipped alleles in the eQTL dataset had their slope (beta) signs inverted. Both a single eQTL and GWAS dataset was required as input for a particular run – with each run corresponding to a particular gene. Multiple runs were conducted for each locus to accommodate the eQTL data for each gene within that locus.

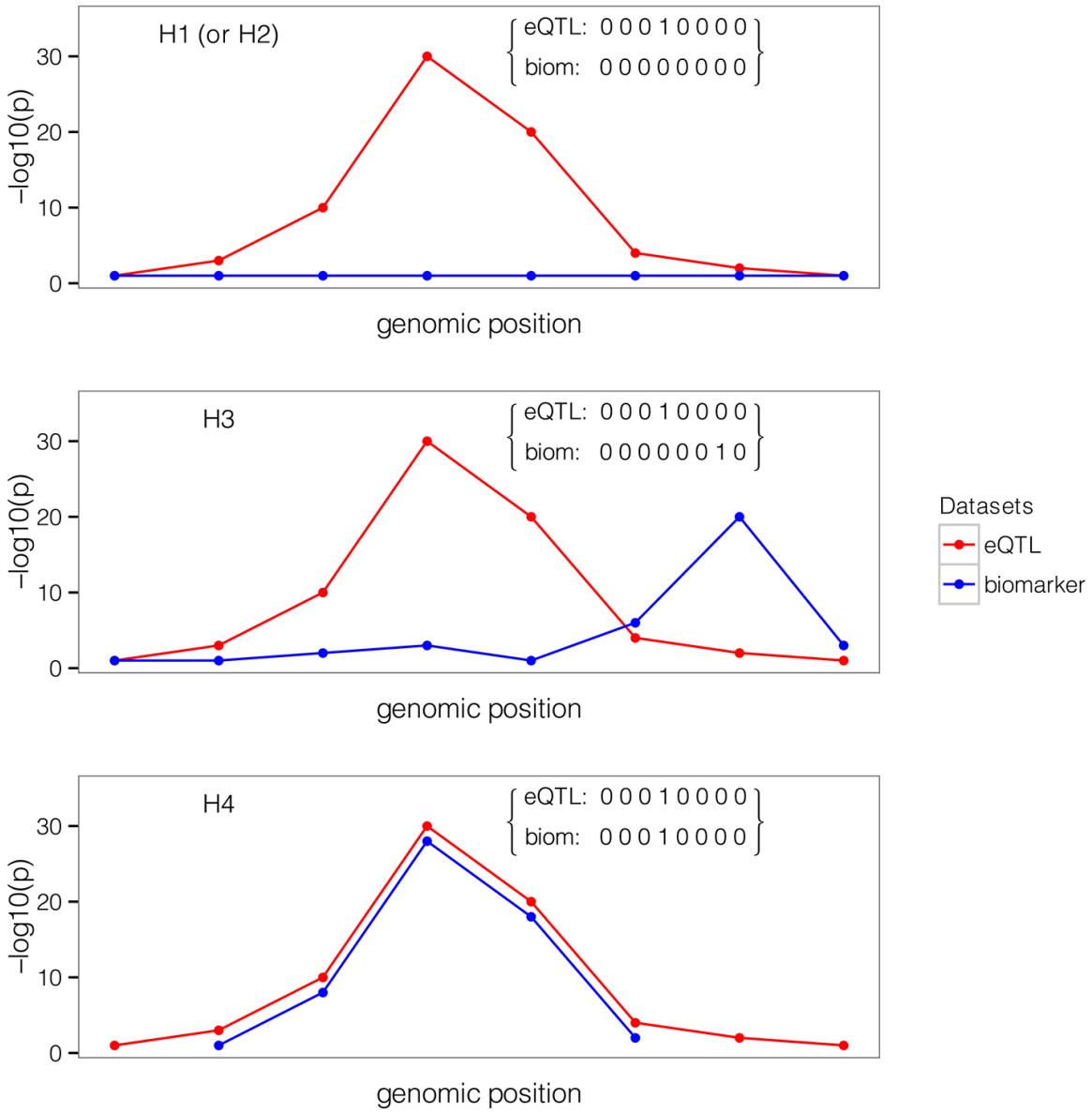


Figure 10 | Colocalisation hypotheses | A diagram adapted from Giambartolomei et al. (2014) which illustrates four potential outcomes when integrating GWAS and eQTL data in a colocalisation analysis. The upper subplot (H1/H2) depicts a situation where either the GWAS or eQTL trait exhibits a genetic association in the specified region. The central subplot (H3) illustrates where both traits possess a genetic association in the region, albeit with distinct causal variants. The concluding subplot (H4) shows both traits as genetically associated in the region while sharing a common causal variant.

Note: a scenario which is not shown above (termed H0) can occur when neither the GWAS nor eQTL trait show a genetic association in the region.

eQTL dataset preparation

Left ventricle eQTL data was obtained from the GTEx database (version 7) QTL tissue-specific all snp gene associations section. Given the list of prioritised genes per lead variant (and locus), the corresponding variant eQTLs were extracted from the overarching eQTL GTEx dataset and saved to a new file labelled with its gene name. The 'GTEx Analysis 2016-01-15 v7 WholeGenomeSeq 635Ind PASS AB02 GQ20 HETX MISS15 PLINKQC lookup table' file was merged onto each eQTL input file so as to map the GTEx's variant ID to rsid format. After pre-processing the datasets, 8 genes (*MYBPC3*, *FAM167A*, *ACTBL2*, *OR9Q1*, *MLIP-AS1*, *PKN2-AS1*, *SPON1-AS1*, and *ANKRD34C-AS1*) lacked any eQTL data and were not included in the colocalisation analysis.

For the colocalisation, the sample size for the eQTL datasets was derived from the laboratory methods of GTEx (release 7) which stated that the Affymetrix Human Gene 1.1 ST Expression Array (version 7) included 837 samples (GTEx, 2017). Additionally, the eQTL dataset input was classified as type 'quant' to represent a quantitative trait.

GWAS data preparation

Summary statistics from the HCM GWAS dataset were sourced from Tadros *et al.* (2025). After pre-processing the datasets, there was GWAS output for each locus which went through to the colocalisation analysis.

For the colocalisation, the effective sample size was derived from the MTAG analysis (N=28106). They were categorised as type 'cc' to denote case-control datasets. The proportion of cases was calculated as $\frac{cases}{(cases+controls)}$ using the raw counts of 5900 HCM cases and 68359 controls to result in a value of 0.086.

Running coloc

Coloc version 5.2.3 was run on the BMRC server. When assessing the colocalisation output, five potential outcomes could be observed, as illustrated in figure 10: (H0) neither trait shared a genetic association in the region; (H1) only trait 1 had a genetic association; (H2) only trait 2 had a genetic association; (H3) both traits had associations but with different causal variants; and (H4) both traits shared a single causal variant (Giambartolomei *et al.*, 2014). For coloc abf, significant associations were determined as H4≥0.5 based on our group's colocalisation analyses.

2.2.5 Assessing the Burden of Rare Variants in Prioritised Genes

Overview

Similarly to that seen in the colocalisation analyses, rare burden analyses were performed for the genes prioritised thus far in the thesis which resulted in an input list of 168 gene IDs across the 68 loci. The 168 genes list included six genes present in the HCM R131 panel (ACTN2, FHOD3, FLNC, MYBPC3, PLN, and TNNC1).

An overview of the rare variant burden analysis methodology is illustrated in figure 11. These analyses were performed within the Genomics England platform (GeL) as well as the National Institute for Health and Care Research BioResource (NIHR BRRD) environments. Case and control selection (step B in figure 11) was conducted as follows: In GeL, the LabKey Python API (version 3.8.1) was accessed and the main programme database (version 15) queried; the case set comprised patients diagnosed with primary HCM at the time of enrollment (as reflected in the 'rare diseases participant disease' table) and controls were selected on the basis that they had not been diagnosed with a primary cardiovascular condition (no such entry in the 'rare diseases participant disease' table) and were also not recruited to any cardiovascular disease panels at any stage thereafter (in the 'domain assignment' table). In BRRD, all cases were extracted from the 'HCM' directory in the rare disease domain and controls from all other directories (excluding those corresponding to HCM or GeL sourced directories) resulting in individuals from thrombogenesis and/or borderline personality

disorder with autism spectrum disorder (BPD), cerebral small vessel disease (CSVD), Ehler Danlos and ED-like syndromes (EDS), intrahepatic cholestasis of pregnancy (ICP), inherited retinal dystrophy (IRD), leber hereditary optic neuropathy (LHON), malignant primary malignant tumours (MPMT), neuro developmental disorders (NDD), neuropathic pain disorders (NPD), pulmonary arterial hypertension (PAH), primary immune disorders (PID), primary membranoproliferation glomerulonephritis (PMG), stem cell and myeloid disorders (SMD), steroid resistant nephrotic syndrome (SRNS), controls for WGS protocol testing (CNTRL), and UK Biobank (UKBio). This resulted in 713 cases and 25040 controls from GeL, and 240 cases and 8358 controls from BRRD.

Steps relating to the rare variant burden analysis pipeline are shown in figure 12 in more detail, for both the LoF and missense burden analyses.

As detailed in figure 12, the loss-of-function transcript effect estimator (LOFTEE) variant effect predictor (VEP) plugin was employed for the LoF burden analysis. Variants predicted to result in loss of function are associated with annotation inaccuracies and observed at low frequencies thus requiring meticulous variant annotation and substantial sample sizes; LOFTEE was designed for this purpose (Karczewski *et al.*, 2020) – implementing rigorous filtering criteria and with the awareness of enriched annotation artefacts (MacArthur *et al.*, 2012). LOFTEE is able to identify terminal truncation variants and rescued splice variants (anticipated to evade nonsense-mediated decay) and focuses instead on high-confidence pLoF variants.

Figure 12 illustrates that the rare exome variant ensemble learner (REVEL) plugin was employed in the missense burden analysis. REVEL is a collective approach for calculating the pathogenic potential of rare missense mutations. It integrates predictions from 18 different scores, which include 8 conservation and 10 functional scores (Ioannidis *et al.*, 2016). In comparative performance assessments involving REVEL, GAVIN, and ClinPred against SIFT and PolyPhen-2, REVEL emerged as the most effective predictor (Gunning *et al.*, 2021).

The fisher.test() package in R was employed to obtain a resulting p-value, odds ratio, and 95% confidence interval per gene (the final step in figure 12). The results considered in this thesis are

those pertaining to an allele frequency threshold of 0.0001 and without feature (transcript) selection. Typically, p-values of 0.05 or lower would be deemed statistically significant; however, due to the input list of prioritised genes containing 168 genes, multiple correction needed to be considered. For this reason, the Bonferroni-corrected p-value threshold of 0.05/168 was calculated. Thereafter, a meta-analysis was conducted using METAL on the biomedical research computing (BMRC) server to increase the power through combining results from both analyses (performed in GeL and BRRD).

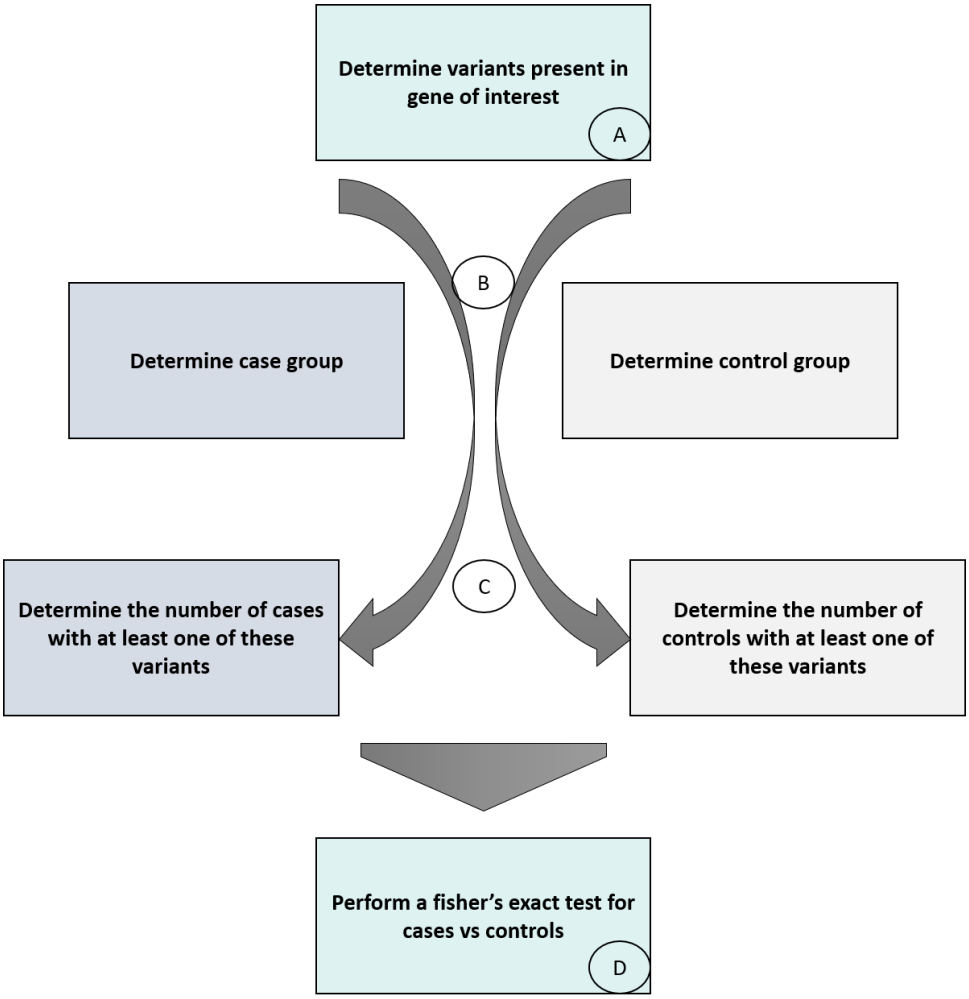


Figure 11 | Rare variant burden analysis overview | A schematic showing a summary of the adopted process of rare variant burden analysis for a particular gene. In step A) all variants of the consequence type of interest (for example, loss of function or missense variants) are compiled for the gene, step B) involves the selection of cases (for a particular trait of interest) and controls, C) entails counting the number of cases and controls with at least one variant present, and step D) encompasses a Fisher's exact test (given the counts) to determine if there is a significant rare variant burden associated with the trait.

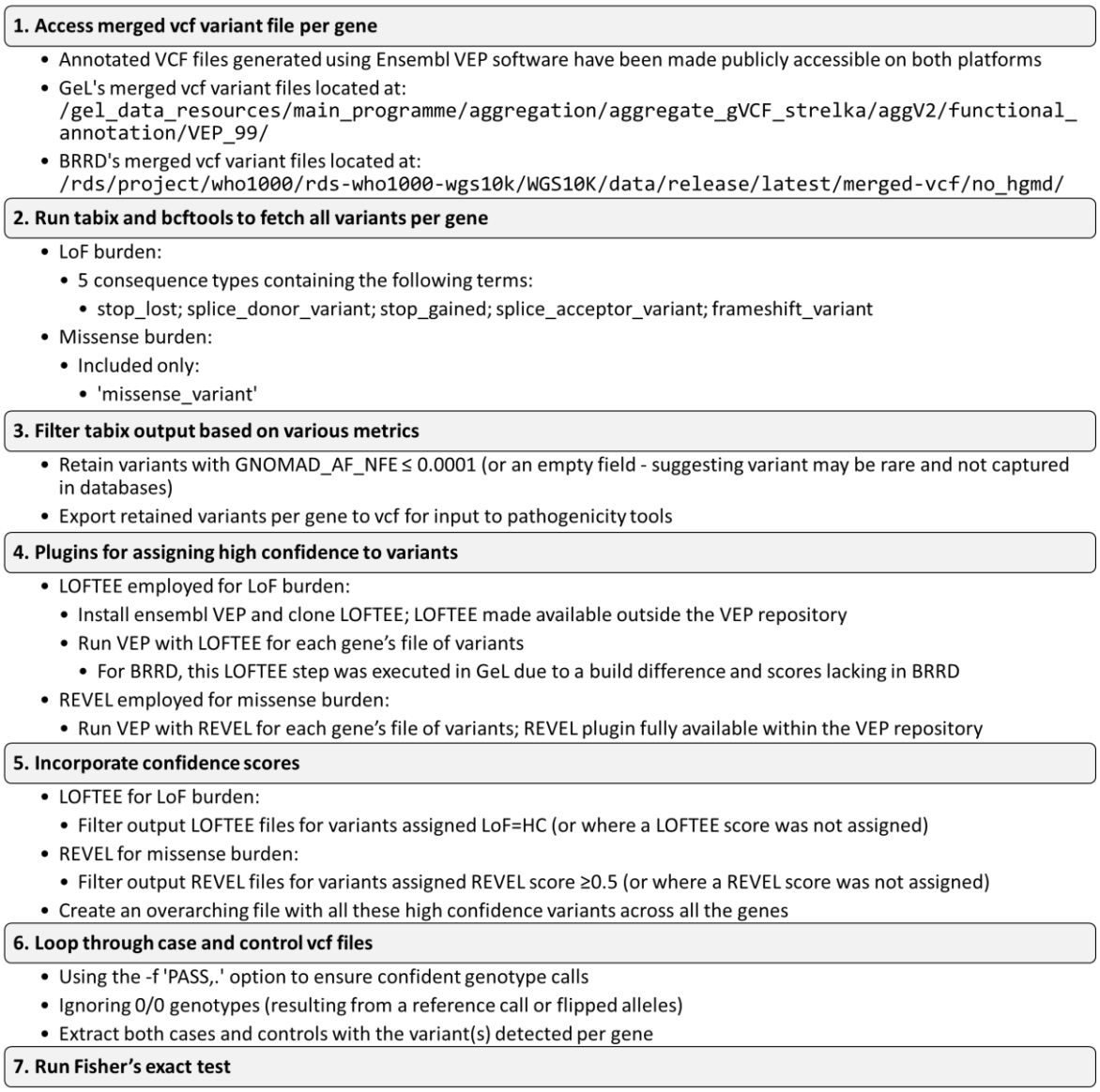


Figure 12 | Rare variant burden analysis pipeline | A detailed flowchart showing the rare variant burden analysis pipeline (both for loss of function as well as missense variants) as was conducted in GeL and BRRD. The steps outline how a gene of interest and its coordinates were used for variant extraction, the filters (regarding allele frequencies and confidence scores) applied, as well as the software required for the analysis.

Further exploration and plotting missense variants in protein structure

Genes with a significant rare missense variant burden were explored further from a protein structure perspective. The missense variants detected in cases and controls were first explored in the genome aggregation database (gnomAD) v4.1 to get the human genome variation society (HGVS) p.

annotation wherever possible as well as their pathogenicity score derived from the Alpha Fold Protein Structure Database. They were then plotted in their protein data bank (PDB) structure and visualised through PyMOL; the executable build of PyMOL was installed locally. Different colours were chosen for shading in case versus control variants to assess differences and/or similarities in their clustering and localisation throughout the protein and its domains. In light of the protein lacking a well-established, resolved structure, the Alpha Fold database was consulted. The Alpha Fold database predicts a protein's 3D structure given its amino acid sequence and so it may not mimic the exact structure as seen *in vivo*. Regardless, Alpha fold highlights likely pathogenic/benign parts of the structure which is helpful in this regard.

2.2.6 Employing the Polygenic Priority Score tool for gene prioritisation

Overview

A conda environment with python v3.7 was set up prior to running the polygenic priority score (PoPs) tool. The PoPs directory was cloned from GitHub (https://github.com/FinucaneLab/pops) and scripts run on the BMRC server; the workflow integrated was as described by the authors with the default (and recommended) settings. In summary, gene-level association statistics were generated through the multi-marker analysis of genomic annotation (MAGMA) (de Leeuw *et al.*, 2015) which were used as input to PoPs for gene score generation.

Selecting and preparing feature sets

Feature sets of interest were downloaded for the mouse heart control comprising of 7 datasets (average expression, projected icaloadings, projected pcaloadings, projected pcaloadings clusters, diffexprs genes clusters, diffexprs down genes clusters, and diffexprs tstat clusters) from the Finucane Lab gene features GitHub directory. These datasets encompassed public bulk and single-cell expression datasets, curated biological pathways, and predicted protein-protein interactions. As there was no human heart data available, mouse data was included as the next best alternative. As

described in figure 13, step 1 entailed providing PoPs with the directory of feature files so that they could be processed for downstream use (referred to as the munged feature files).

MAGMA for annotation and score generation

As shown in figure 13, step 2.1 served as a preliminary processing step prior to the PoPs analysis – this entailed input variants being aligned to their corresponding genes. This MAGMA alignment was determined by genomic positioning whereby variants captured within the designated region (defined by transcription start and stop sites) for a particular gene were paired. Step 2.2 entailed the gene analysis phase whereby p-values and other gene-level metrics were calculated for the associations. Additionally, correlations among adjacent genes were assessed in anticipation of the subsequent gene-level analysis.

Running PoPs

In figure 13, step 2 describes the actual running of PoPs for gene prioritisation given the integrated GWAS summary statistics, feature set data from the munged mouse heart control feature sets, and MAGMA gene-level statistics. The output predictions file was merged with TAD-defined locus coordinates to compile all genes captured per locus (sorted by PoPs score). The top-scoring gene was extracted per locus and marked as the prioritised gene.

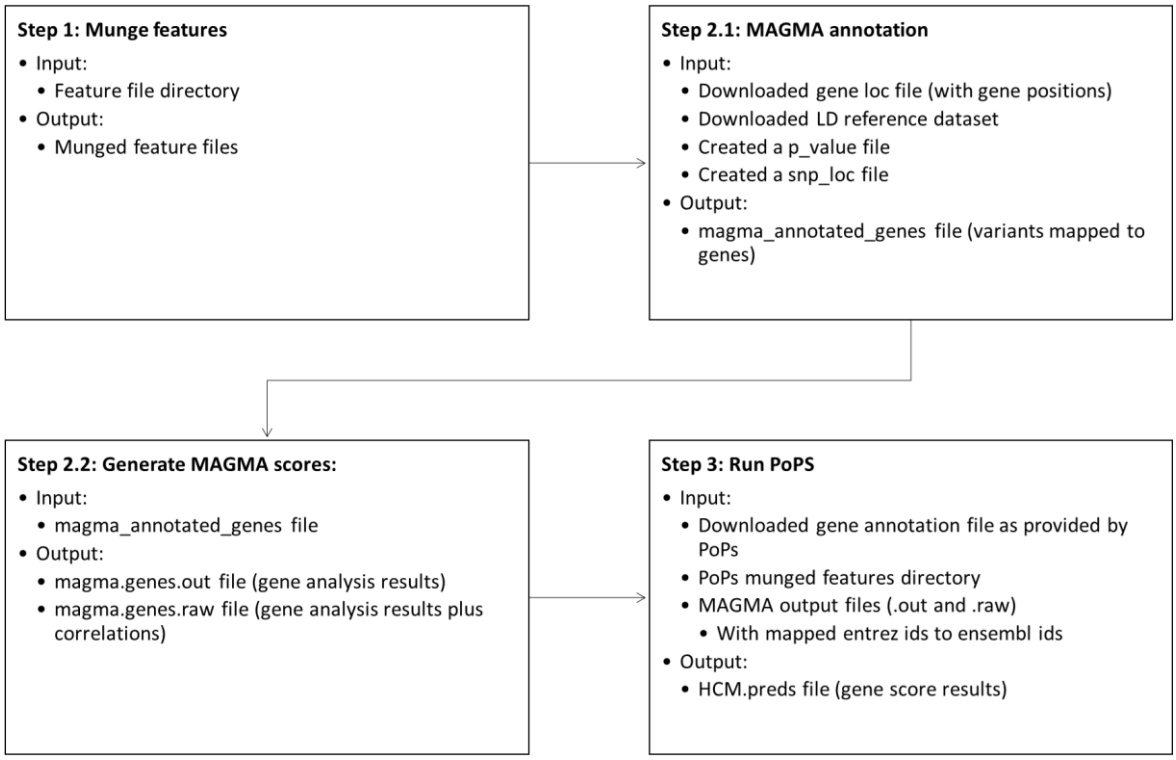


Figure 13 | PoPs gene prioritisation pipeline | A flowchart illustrating the process of employing the gene prioritisation method, polygenic priority score (PoPs). Step 1) describes munging feature datasets (as downloaded from https://github.com/FinucaneLab/gene_features/tree/master/features) which entailed the mouse heart datasets for this analysis; step 2.1) calls on another software, MAGMA, for annotation of the input HCM GWAS summary data, the gene loc file was in build37 as taken from the magma datasets website - filtered to contain only genes captured in gene_annot_jun10.txt provided by PoPs and duplicate entries for a particular EntrezID were dropped, the LD reference dataset was in build37for the European population as taken from the magma datasets website, the p-value file was created from the HCM GWAS dataset (subsetted for rs_id and p-value), and the snp_loc file was created from the HCM GWAS dataset (subsetted for rs_id, chromosome, and base_pair_location); step 2.2) entails the calculation of MAGMA scores based on the annotations provided in the previous step; and step 3) is where PoPs is employed, given the feature sets and MAGMA scores, to generate final gene prioritisation scores. The necessary input files as well as those generated as output are shown for each step.

2.3 Variant Prioritisation

2.3.1 Overview

This chapter leveraged chromatin datasets (representing ATAC-Seq, DNase-Seq, Histone ChIP, and ChIP TF) for input to deep learning. A detailed dataset tracking sheet is available at GitHub (github.com/domwest/DPhil-thesis/blob/main/data_tracking_sheet_for_thesis.xlsx). The tracking sheet outlines which tissues and cell types were captured and in which sections of the thesis they were used.

2.3.2 Cell type enrichment

The cell type enrichment analysis was conducted in Jupyter notebook using Python3 on the BMRC compute cluster. This workflow was provided by the Hughes' lab (Riva *et al.*, 2025) and I had privileged access to the code.

The aim of this pipeline is to give an enrichment score so that cell types can not only be prioritised but can also be ranked. Additionally, each cell type's background is independent and so the enrichment score will not change depending on whether the user chooses to investigate a cell type independently or in combination with others.

Only ATAC/DNase data was required and so a combination of ATAC and DNase bed and bigwig peak files across the cell types of interest were used as input. The appropriate input files had already been filtered for peaks with a LanceOtron (Hentges *et al.*, 2022) overall peak score > 0.5. The peak calling framework, LanceOtron, integrates deep learning techniques for identifying peak shapes with comprehensive enrichment calculations to evaluate the significance thereof – while performing peak calling and assigning scores to each peak representing the strength or the likelihood of the peak being real. This resulted in a total of 46 open chromatin datasets each representing a different cell type.

Proxy variants were detected using LDLinkR proxy batch with genome build38 specified (Myers *et al.*, 2020) while filtering for $R^2 \geq 0.8$ which resulted in a final set of 4395 variants; as the proxy variants were returned in rsid format, I used dbSNP to fetch the corresponding reference and alternate alleles. The 4395 (3548 bi- and 847 multi-allelic) HCM GWAS lead and proxy variants were used as input to the pipeline for variant enrichment containing the following fields: chromosome, position (in build38), and rsid. Proxy variants, in addition to the lead variants, were leveraged in the prioritisation analysis for inclusivity in addition to the consideration that lead variants are not always the causal variants.

2.3.3 Chromatin Data QC and Standardisation Via CATCH-UP

This pipeline was run as a precursor to DeepHaem with the aim to standardise and QC the data beforehand; the pipeline's process is illustrated in figure 14.

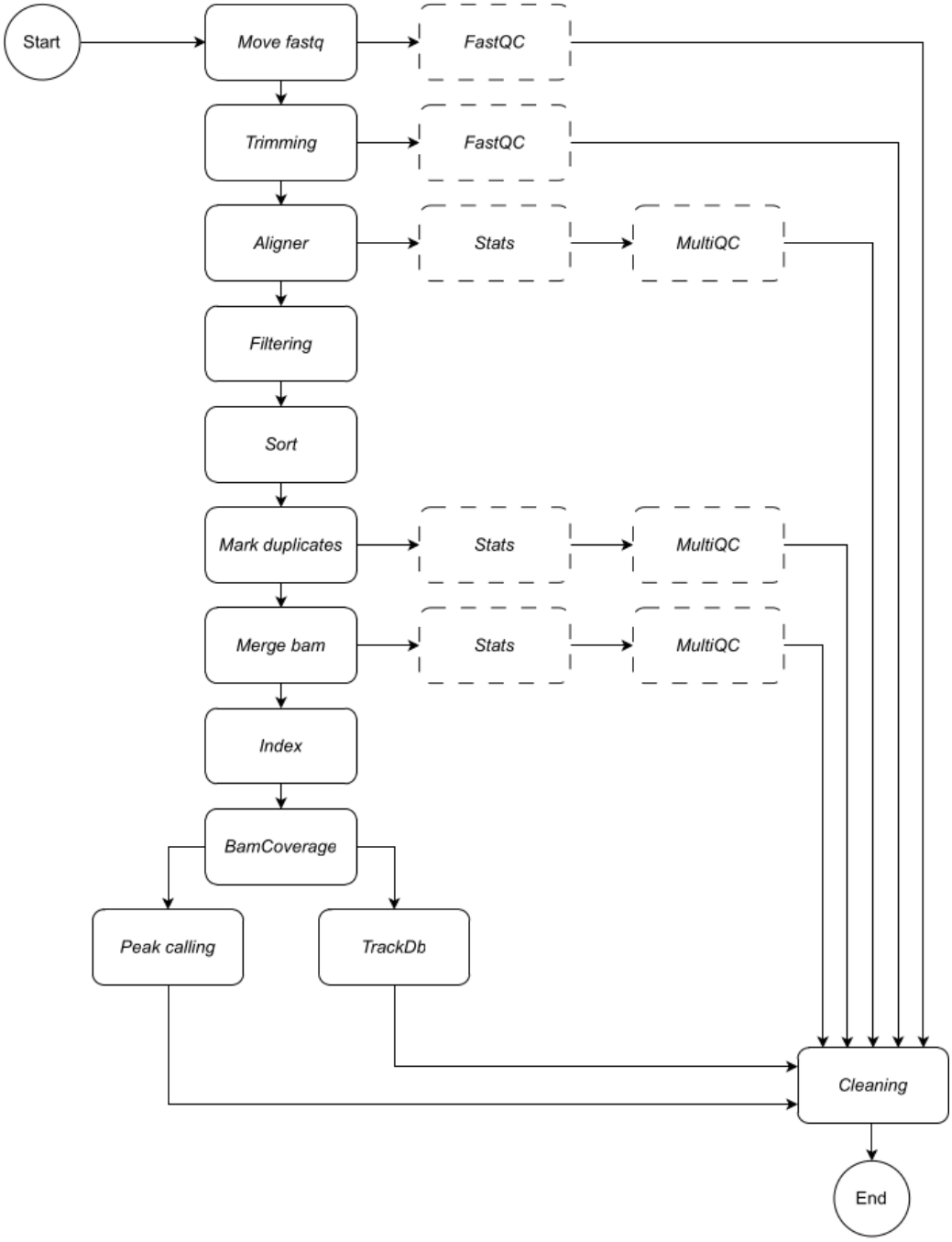


Figure 14 | Chromatin data QC tool (CATCH-UP) pipeline | A schematic, adopted from Riva et al. (2023), illustrating the upstream processing CATCH-UP tool's pipeline. The illustration outlines how CATCH-UP receives raw sequencing reads from ATAC-Seq, DNase-Seq, and ChIP-seq experiments, QCs the datasets, and then maps them to a reference genome. Mapped reads then undergo filtering for the removal of those showing poor quality as well as PCR duplicates. To aid in visualisation, coverage is calculated across reads and either LanceOtron or MACS2 is leveraged for peak calling.

Open chromatin data from bulk tissue heart left ventricle as well as single-cell heart datasets (including cardiac muscle cell, fibroblast, lymphocyte, macrophage, ventricular cardiomyocyte, atrial cardiomyocyte, adipocyte, smooth muscle cell, and endothelial cell) were downloaded from various databank sources. Most of the data was downloaded from ENCODE (Dunham *et al.*, 2012), some from the cardiac atlas of regulatory elements (CARE) portal, and one dataset from the national center for biotechnology information sequence read archive (NCBI SRA). The following methodologies were captured across the various datasets: ATAC-Seq, DNase-Seq, TF ChIP-seq, and Histone ChIP-seq (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3). All TF ChIP-seq datasets were CTCF transcription factor site data. See the tracking sheet at github.com/domwest/DPhil-thesis/blob/main/data_tracking_sheet_for_thesis.xlsx for dataset related details.

Both the command-line and web version of CATCH-UP were employed depending on the dataset file formats. Input files need to be in the form of fastq files for the command-line version to run. A combination of single and paired-end fastq files were downloaded wherever possible and used as input to the command-line version of this pipeline. In the event that fastq files were not available but bed and bigwig files were, LanceOtron web for peak-calling was run on the multi-locus view (MLV) interface (Sergeant *et al.*, 2021) instead. The command line version was carried out on the BMRC cluster by accessing the Upstream pipeline conda environment where all relevant packages were pre-installed.

Before running CATCH-UP, the reference genome (build 38) was downloaded directly from UCSC and indexed by Bowtie2 (integrated in CATCH-UP). When executing the CATCH-UP pipeline, certain parameters needed to be introduced to inform the pipeline as to whether concatenation of lanes, merging of samples, and/or adapter trimming was required. Lane concatenation was not implemented as there were no sample replicates in the input and adapter trimming was not performed due to the use of the reliable aligner, bowtie. The files were merged on single versus paired end sequencing, bulk tissue versus individual cell type data, and according to the methodology employed (ATAC versus DNase etc.). In addition, the files were merged across donors. The pipeline output bed, bigwig, and bam files per merged dataset.

2.3.4 Using Machine Learning to Prioritise Functional Variants

DeepHaem

The core model's code (defining the model parameters and layers) was made available to me from the Hughes' lab with the aim of re-training it on heart data as opposed to blood cell data.

Output .bed peak files from the CATCH-UP pipeline were used as input to DeepHaem (Schwessinger *et al.*, 2020). The various datasets (categorised by tissue/cell, methodology, and sequencing run type) became the defined classifiers for DeepHaem; this comprised of a combination of chromatin datasets representing either bulk heart left ventricle data or that from various heart cell types.

Before the peak files were used as input to DeepHaem, they were filtered to contain only peaks with an overall LanceOtron peak score greater than 0.5 and all blacklisted regions for build38 were excluded (Amemiya *et al.*, 2019). Blacklisted regions of the genome are problematic areas which often generate anomalous signals or errors throughout sequencing and analysis. After these steps, files containing a small number of peaks were removed. This threshold, as suggested by the Hughes' lab was 1000.

Across all datasets, overlapping peaks were merged into all-encompassing peaks based on the left- and right-most coordinates via bedtools. Two types of input were prepared for DeepHaem: an object capturing whether or not a particular peak was present in each classifier (in binary format) and one-hot encode sequences of the 1000bp positioned at the halfway point of each peak. An example of converting a DNA sequence to a one-hot encode sequence is shown in table 2. For the one-hot encode sequences, the following bases were initially accounted for: 'A', 'C', 'G', 'T', 'N' and arrays corresponding to 'N' were assigned zeros as this could represent any nucleotide and is therefore uninformative (Brown, 2002).

Table 2 | Applying one-hot encoding to DNA sequences | A table illustrating the conversion of DNA sequences to a numerical format via one-hot encoding prior to their use as input to machine learning methods. DNA sequences and the four bases they constitute (A, C, T, and G) are shown in the first column in red. The top row depicts a scoring per base. The matrix values depict an example of one-hot encoding conversion for the string of bases in column 1 resulting in an array for this DNA sequence (as shown in the grid in binary format).

		Base conversion to one-hot encoding			
		A	**C**	**T**	**G**
Input DNA sequence	**A**	1	0	0	0
	C	0	1	0	0
	N	0	0	0	0
	G	0	0	0	1
	…	…	…	…	…
	C	0	1	0	0
	T	0	0	1	0
	A	1	0	0	0

DeepHaem is a sequential model with a linear stack of layers, where each layer's output acts as input to the next layer. Regarding parameters, the number of samples processed at a time during training (batch size) was set to be 100, the number of times the complete training dataset was passed through the network (epoch) was 100, and early bird stopping was integrated with a patience of 10. Early bird stopping serves to prevent overfitting whereby the patience value informs the number of epochs prior to an early stop when there has been no progress on the validation set.

Before training could ensue, the datasets were split into training, validation, and test sets. 60% of the data was used for training and 20% for validation as well as 20% for testing. The DeepHaem model was implemented on the Biomedical Research Computing (BMRC) facility in the Medical Sciences Division at the University of Oxford's server and the machine learning platform, TensorFlow version 2.13.0, was used with the python interface Keras.

Initially the model was trained with 24 classifiers (shown in table 3). For computational validation of how the model was trained, AUC (area under curve) and loss plots were generated. Receiver operating characteristic (ROC) curves were also generated for the training and validation data per classifier to assess their individual performance.

Table 3 | 24 classifiers for input to DeepHaem | A table depicting the 24 classifiers input to DeepHaem for training, validation, and testing. Associating information such as the tissue/cell type of relevance (as well as the type thereof), the dataset methodology, histone target (if applicable), and the databank where the dataset was sourced from is shown. All TF ChIP datasets were for the TF CTCF.

Tissue or cell type name	**Methodology**	**Target**	**Databank**
Cardiac muscle cell	Histone ChIP-seq	H3K27ac	ENCODE
Cardiac muscle cell	Histone ChIP-seq	H3K27me3	ENCODE
Cardiac muscle cell	Histone ChIP-seq	H3K4me3	ENCODE
Fibroblast	ATAC-Seq		CARE Portal
Lymphocyte	ATAC-Seq		CARE Portal
Macrophage	ATAC-Seq		CARE Portal
Ventricular cardiomyocyte	ATAC-Seq		CARE Portal
Adipocyte	ATAC-Seq		CARE Portal
Atrial cardiomyocyte	ATAC-Seq		CARE Portal
Smooth_muscle	ATAC-Seq		CARE Portal
Endothelial	ATAC-Seq		CARE Portal
Cardiac muscle cell	DNase-Seq		ENCODE
Cardiac muscle cell	TF ChIP-seq		ENCODE
Cardiac muscle cell	DNase-Seq		ENCODE
Heart left ventricle	Histone ChIP-seq	H3K27ac	ENCODE
Heart left ventricle	Histone ChIP-seq	H3K27me3	ENCODE
Heart left ventricle	Histone ChIP-seq	H3K36me3	ENCODE
Heart left ventricle	Histone ChIP-seq	H3K4me1	ENCODE
Heart left ventricle	Histone ChIP-seq	H3K4me3	ENCODE

Heart left ventricle	Histone ChIP-seq	H3K9me3	ENCODE
Heart left ventricle	ATAC-Seq		ENCODE
Heart left ventricle	DNase-Seq		ENCODE
Heart left ventricle	TF ChIP-seq		ENCODE
Heart left ventricle	TF ChIP-seq		ENCODE

Visual inspection was then performed whereby the model's predictions of the underlying chromatin state (at random regions) were inspected and compared to that depicted in the actual raw data. DeepHaem uses a rolling window method for prediction. Five hundred bases both before and after the sequence of interest needed to be provided for the rolling window predictions. Due to the nature of rolling window scanning, each base's prediction score is generated through the average of each window's score at that particular base. The prediction score represents the predicted chromatin state at the specified base, where the x-axis shows the base position and the y-axis the probability score (of the chromatin being open) at each base. The probability scores are continuous values derived from the binary input of a peak being present or not, representing the likelihood of there being a peak (if this value is leaning towards 1) or not (closer to zero). The relationship between a high probability score and the chromatin being open (represented by a peak) versus a low probability score and closed chromatin is illustrated in figure 15.

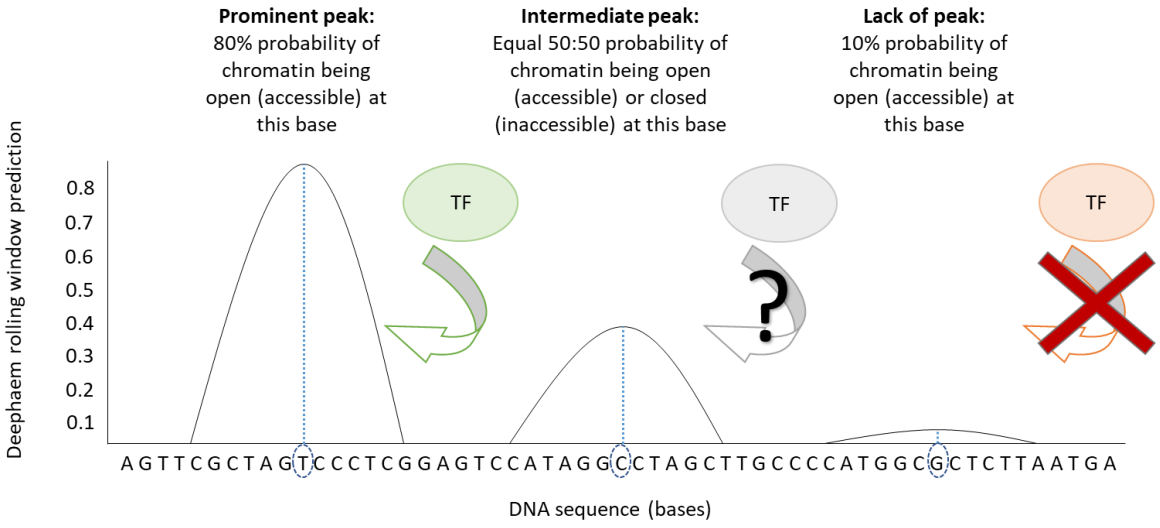


Figure 15 | DeepHaem rolling window predictions and chromatin accessibility | A figure illustrating the rolling window predictions performed by the machine learning model, DeepHaem. A genomic

interval with a specified region length needs to be input to the model, whereby the DNA is extracted and a prediction of the underlying chromatin regarding its openness/accessibility versus closedness/inaccessibility is performed. 'Open' chromatin in this instance refers to DNA that can be accessed/bound by transcription factors for gene expression/regulation to commence whereas 'closed' chromatin represents a scenario where such elements cannot bind and thus regulatory activity is not instigated.

To allow for capturing both the forward and reverse DNA strands, one-hot encode sequences of both the forward and reverse sequences were passed to the model. A combination of the forward and reverse predictions was showcased by taking the minimum score seen across both at a specific point (as shown in figure 16). This was aimed at eliminating as much background noise as possible.

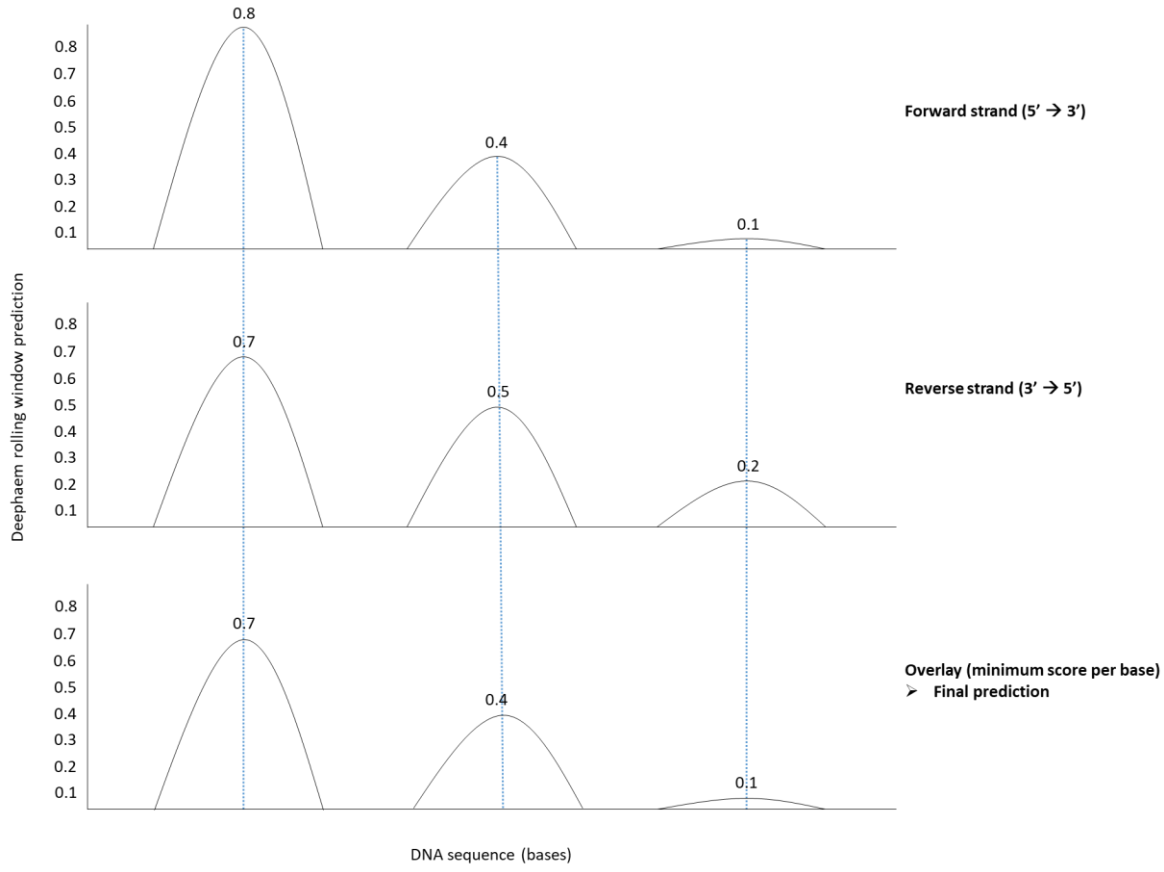


Figure 16 | DeepHaem rolling window predictions: forward, reverse, and combined DNA strands | A figure illustrating the rolling window predictions performed by the machine learning model, DeepHaem. A genomic interval with a specified region length needs to be input to the model,

whereby the DNA is extracted and a prediction of the underlying chromatin is performed. The DNA sequence extracted for the genome region specified by the user is extracted in its forward strand (5' to 3') read out, as well as its reverse strand (3' to 5'), and finally a combination of the two is performed (by picking the minimum prediction probability score when comparing the forward and reverse strands per base).

Three regions were chosen somewhat randomly for rolling window predictions as model checks. These 3 regions came from different chromosomes and differed in length as shown in table 4; this table also shows an additional region (region4) which was utilised in a later troubleshooting stage. Additionally, they were chosen based on how many classifiers within which they were detected – to check the model's accuracy in predicting the peaks regardless of whether the peak was well-represented across most classifiers or detected in only a few. The predictions were compared to the raw chromatin data as a means of validating the model predictions.

Table 4 | DeepHaem visual inspections: regions selected | A table depicting genomic interval regions chosen for performing visual inspection. The regions were chosen based on presence of a peak (detected in varying proportions of classifiers) or otherwise a closed/flat region selected for validations post-modification. The purpose column specifies the analysis stage during which the region was used. The characteristic column specifies whether a peak or closed region was of interest within a particular region. The chromosome, start, and end coordinates as well as their sizes are shown alongside for the peak/closed characteristic as well as the chosen region with which to use as input to the model's prediction. Positions shown according to genome build 38.

Purpose	**Genomic interval name**	**Characteristic**	**Peak presence in classifiers %**	**Chromosome**	**Characteristic start position**	**Characteristic end position**	**Characteristic size**	**Region start position**	**Region end position**
Visual inspection	region1	Peak	79	1	1396946	1401081	4135	1379046	1418981
	region2	Peak	12	9	175817	176695	878	163756	188756
	region3	Peak	4	2	242080657	242081103	446	242067380	242094380

		Flat	N/A	2	242067680	242071815	4135	242067380	242094380
Troubleshooting	region4	Peak	5	2	5074031	5075298	1267	5064031	5085298

Other means of visual inspection were also carried out. One approach was to modify the underlying sequence of a particular peak into that of a closed region, and another was to insert a peak sequence into a region that was otherwise closed. The former entailed extracting the exact peak coordinates from region1 and inserting this into the initial, closed part of region3; the latter entailed replacing the first peak sequence in region1 with the same length of a randomly shuffled sequence containing only 'A', 'G', 'T', and 'C'. These two approaches aimed to test whether the model could infer both the erasure and insertion of a peak pattern, respectively.

After computational validation and visual inspection steps were performed, the list of 4395 HCM GWAS lead and proxy variants were prepared. This was for the intention of running a post-analysis with the specific aim of leveraging DeepHaem's chromatin predictions to perform functional variant prioritisation.

For each variant's prediction, DeepHaem was given the variant plus 500bp both up- and down-stream due to the rolling window process resulting in tapering ends where less window predictions are accounted for. This led to sequences of 1001bp for each variant being passed as input to DeepHaem. For each variant, 2 different sequences were provided – the sequence containing the reference allele as well as that containing the alternate allele. For bi-allelic variants, only one prediction needed to be computed for the alternate; however, for multi-allelic variants multiple predictions were computed to account for all possible changes. To measure the plausible impact inflicted from a variant's alternate allele(s), the reference and alternate allele's scores were compared – producing what is referred to as a 'damage score' in this setting (as shown in figure 17). The damage score is calculated by taking the alternate allele's probability score and subtracting the

reference allele's score. The magnitude instructs the level of impact whereas the sign indicates whether there is an increase or decrease in the chromatin accessibility.

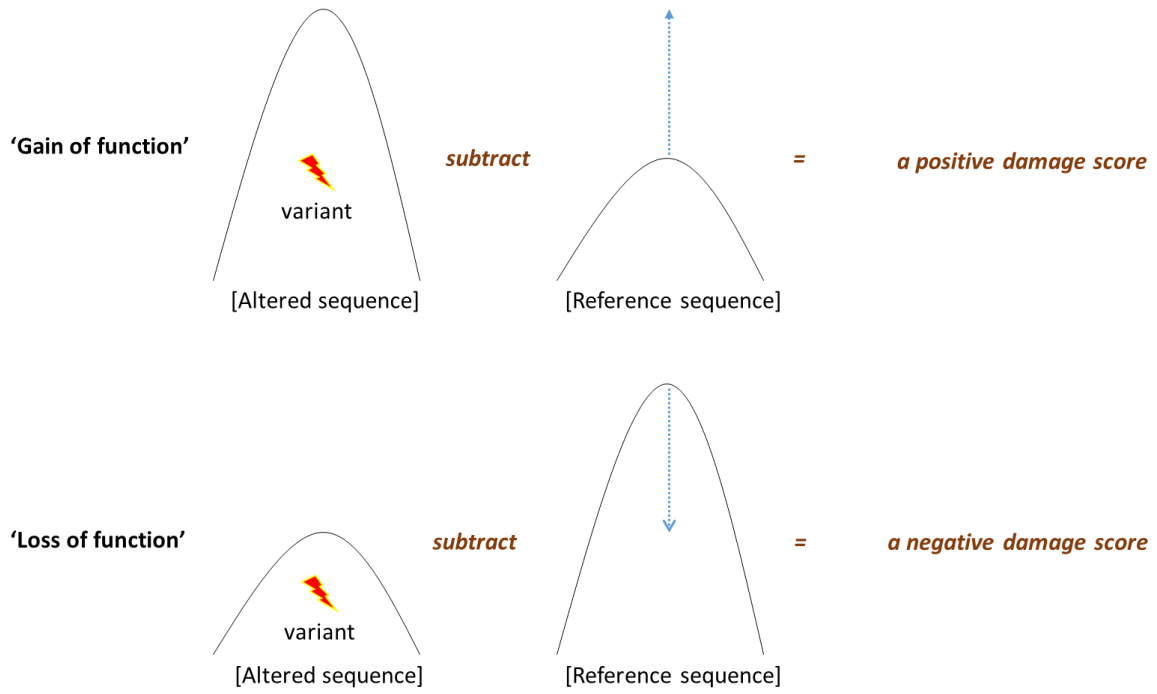


Figure 17 | DeepHaem damage score prediction for variants | A diagram illustrating what constitutes a 'gain of function' versus a 'loss of function' regulatory variant with regards to their effects on the underlying chromatin. In both scenarios, the effect of a particular variant (which contributed to an altered underlying sequence) inflicts change on the reference sequence (without the presence of a variant). The 'gain of function' mechanism is shown at the top whereby the variant induces the chromatin to become more open and therefore accessible. The 'loss of function' mechanism is shown at the bottom whereby the variant induces the chromatin to become more closed and therefore inaccessible. The base change for both scenarios is shown in the form of diagrams and in the layout of a mathematical equation whereby a score for the reference base sequence is subtracted from the altered sequence base score to equal either a positive or negative damage score (a score depicting the extent of impact on the chromatin).

The authors of DeepHaem, which was previously trained on blood cell data, determined an 'empirical threshold' of ±0.1 for variant prioritisation. After damage scores were generated for all

4395 variants, these scores were plotted across all classifiers to visualise the damage score distributions.

I sought a set of positive control HCM variants to act as a reference of what damage score thresholds should be set for the 4395 variants; however, due to a lack of empirical data linking common variants to HCM a true positive control set could not be obtained. I then attempted to gauge the next best 'positive control' set by looking into RegulomeDB for variants with a score of 1A (representing Eqtl + TF binding + matched TF motif + matched DNase footprint + DNase peak evidences) also showing a tissue-specific score in the heart greater than or equal to the mean±2SD across all other tissues. Ten variants which passed these thresholds and hence my positive controls were chosen (shown in table 5).

Table 5 | RegulomeDB HCM positive control set of variants | A table depicting the initial set of HCM positive control regulatory variants. This includes 10 randomly-chosen variants extracted from RegulomeDB with a score of 1A (representing Eqtl + TF binding + matched TF motif + matched DNase footprint + DNase peak evidences) which also showed a tissue-specific score in the heart ≥ mean±2SD scores across all other tissues. Each variant's identifiers are shown including its rsid, chromosome (chr), build 38 position (pos_b38), reference allele (ref), and alternative allele (alt) as well as its ranking in RegulomeDB (based on associated evidences), mean tissue specific score, and its heart specific score.

Variant					**RegulomeDB**		
rsid	**chr**	**pos_38**	**REF**	**ALT**	**Ranking**	**Mean_tissue_specific_scores**	**Heart_specific_score**
rs1002706	1	24724662	G	A	1a	0.362007	0.58714
rs13017338	2	86274430	T	C	1a	0.382073	0.61182
rs1888115	1	1.54E+08	G	A	1a	0.345035	0.53038
rs3768324	1	39026790	C	T	1a	0.265857	0.58826
rs4660652	1	38992187	A	G	1a	0.359533	0.55467
rs4672016	2	55097824	A	C	1a	0.345226	0.618
rs6685037	1	2.33E+08	C	T	1a	0.30986	0.48466
rs749579	2	86221285	A	G	1a	0.368444	0.57474

rs7523070	1	28314262	C	A	1a	0.366023	0.57076
rs7535313	1	58539719	C	A	1a	0.330305	0.51867

To act as another next best 'positive control' set, I obtained a list of variants that have been defined as putative regulatory variants for HCM by Lesurf *et al.* (2022) who suggest they found high-risk regulatory variants in HCM gene elusive cases (shown in table 6).

Table 6 | Putative regulatory HCM positive control set of variants | A table depicting another set of HCM positive control regulatory variants. This includes 41 putative regulatory variants identified by Lesurf et al. (2022) across 17 unique genes from gene elusive cases. This list has been segregated into prioritised high-risk variants as well as candidate variants. For each variant, the putatively-affected gene is shown, the variant's tier, category of gene, position (in both build 19 and 38), as well as the inflicted nucleotide change is shown.

Prioritised high-risk regulatory variants in gene elusive cases				
Gene	**Tier**	**Gene category**	**Genomic Position (hg38)**	**Nucleotide Change**
BRAF	Tier 1: Secondary	RAS	chr7:140924486	C>T
BRAF	Tier 1: Secondary	RAS	chr7:140924893	C>T
DSC2	Tier 1: Primary	Desmosomal/ACM	chr18:31101515	G>C
DSG2	Tier 1: Primary	Desmosomal/ACM	chr18:31497905	C>A
DSP	Tier 1: Primary	Desmosomal/ACM	chr6:7541543	G>A
DTNA	Tier 2: Primary	Dystroglycan	chr18:34493213	T>C
DTNA	Tier 2: Primary	Dystroglycan	chr18:34493326	C>T
DTNA	Tier 2: Primary	Dystroglycan	chr18:34493332	C>G
DTNA	Tier 2: Primary	Dystroglycan	chr18:34593068	G>A
EYA4	Tier 2: Primary	Signaling	chr6:133241280	G>A
EYA4	Tier 2: Primary	Signaling	chr6:133502899	A>C
FKTN	Tier 1: Secondary	Dystroglycan	chr9:105557710	A>C
FKTN	Tier 1: Secondary	Dystroglycan	chr9:105558049	G>A
JPH2	Tier 1: Primary	Structural	chr20:44137240	C>G
LAMP2	Tier 1: Secondary	Lysosomal	chrX:120469070	G>A

MYOZ2	Tier 2: Primary	Z disc	chr4:119046577	C>A
NEXN	Tier 1: Primary	Structural	chr1:77888572	A>G
NEXN	Tier 1: Primary	Structural	chr1:77889313	C>G
NEXN	Tier 1: Primary	Structural	chr1:77889430	A>T
PRKAG2	Tier 1: Primary	Metabolic	chr7:151695095	A>C
PRKAG2	Tier 1: Primary	Metabolic	chr7:151783095	G>A
PRKAG2	Tier 1: Primary	Metabolic	chr7:151876032	C>G
SOS1	Tier 1: Secondary	RAS	chr2:39119972	G>A
SOS1	Tier 1: Secondary	RAS	chr2:39121003	C>T
TCAP	Tier 2: Primary	Z disc	chr17:39663715	C>T
TGFB3	Tier 2: Primary	Signaling/ACM	chr14:75822875	A>G
TGFB3	Tier 2: Primary	Signaling/ACM	chr14:75823021	G>A
TGFB3	Tier 2: Primary	Signaling/ACM	chr14:75823544	C>G
TGFB3	Tier 2: Primary	Signaling/ACM	chr14:75905806	C>T
TGFB3	Tier 2: Primary	Signaling/ACM	chr14:75981836	T>A
VCL	Tier 1: Primary	Z disc	chr10:73997807	C>A
Candidate regulatory variants				
BRAF	Tier 1: Secondary	RAS	chr7:140924423	G>A
DSP	Tier 1: Primary	Desmosomal/ACM	chr6:7541235	T>C
DSP	Tier 1: Primary	Desmosomal/ACM	chr6:7541614	C>G
DSP	Tier 1: Primary	Desmosomal/ACM	chr6:7541882	C>T
DTNA	Tier 2: Primary	Dystroglycan	chr18:34492902	A>G
DTNA	Tier 2: Primary	Dystroglycan	chr18:34492979	C>A
DTNA	Tier 2: Primary	Dystroglycan	chr18:34593103	C>T
FKTN	Tier 1: Secondary	Dystroglycan	chr9:105558064	G>T
LARGE1	Tier 1: Secondary	Dystroglycan	chr22:33920428	C>T
LARGE1	Tier 1: Secondary	Dystroglycan	chr22:33920699	G>A

Having had to abandon the use of high confidence regulatory variants to act as positive controls and help define prioritisation thresholds, the next attempt at prioritisation was rooted in defining

thresholds based on the distribution of damage score values seen across the 4395 variants. This involved extracting those that passed the mean±2SD threshold. This approach was chosen on the basis that data points in a normal distribution are more likely to fall close to the mean with 68% of all data points falling within 1 standard deviation of the mean, 95% within 2 standard deviations of the mean, and 99% within 3.

Lacking a true positive control set was an obstacle to prioritisation attempts until a set of 821 empirical erythroid enhancer variants became available from the Hughes' lab. In order to use these variants as a positive control, I needed to include erythroid-specific datasets as new classifiers in the model. I obtained 3 erythroid tracks from the Hughes' lab (2 of which were from the actual donors in which these enhancer variants were identified) in addition to other heart-relevant ATAC datasets that were also made available.

65 total datasets were now prepared for re-training DeepHaem. After training, the computational validation and visual inspection checks ensued. There was a slight difference in approach for visual inspection regarding the erasure of a peak – the peak sequence was not replaced with a randomly shuffled sequence of nucleotides but instead, the same length sequence was taken from an otherwise closed section of chromatin (in region3) and inserted (flat characteristic from region 3 detailed in table 4 and as illustrated in figure 18).

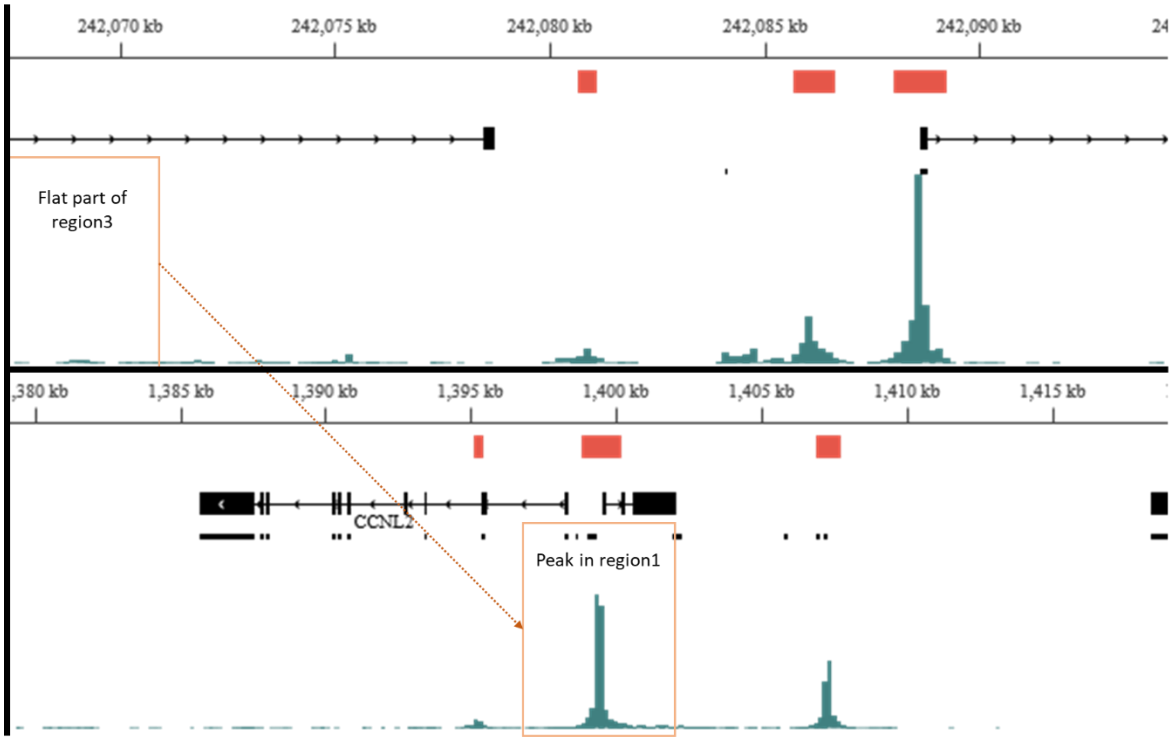


Figure 18 | DeepHaem visual inspections involving sequence modification: peak erasure visualisation | A figure illustrating how a closed/flat sequence was extracted from region3 and used to replace the same length peak as found in region1 for visual inspections post-modification. This was termed 'peak erasure' and performed to test whether the model's predictions could accurately predict when the underlying chromatin patterns had changed.

The positive control erythroid enhancer variant set's damage scores could now be plotted across all classifiers – to see if the variants showed more extreme damage scores in the 3 erythroid tracks when compared to the other heart tracks.

Given an unexpected result for the erythroid enhancer damage score distributions (an actual positive control set), exploration into the model's architecture commenced. Neurons in CNN models exist as units which perform calculations on the input data and an activation (mathematical) function is applied to the output of each neuron introducing non-linearity into the model. This led to the theory that perhaps a SoftMax activation function should have been used in the final layer instead of a sigmoidal one. SoftMax activation is for multi-class scenarios whereby probabilities are assigned to each class and the assigned values are dependent on other classifiers; sigmoid activation on the other hand is used for binary classification tasks which focuses on distinguishing between two exclusive

outcomes (figure 19). After tweaking the final layer to a SoftMax activation function, the model was re-trained and auc and loss plots as well as the classifiers' ROC curves observed.

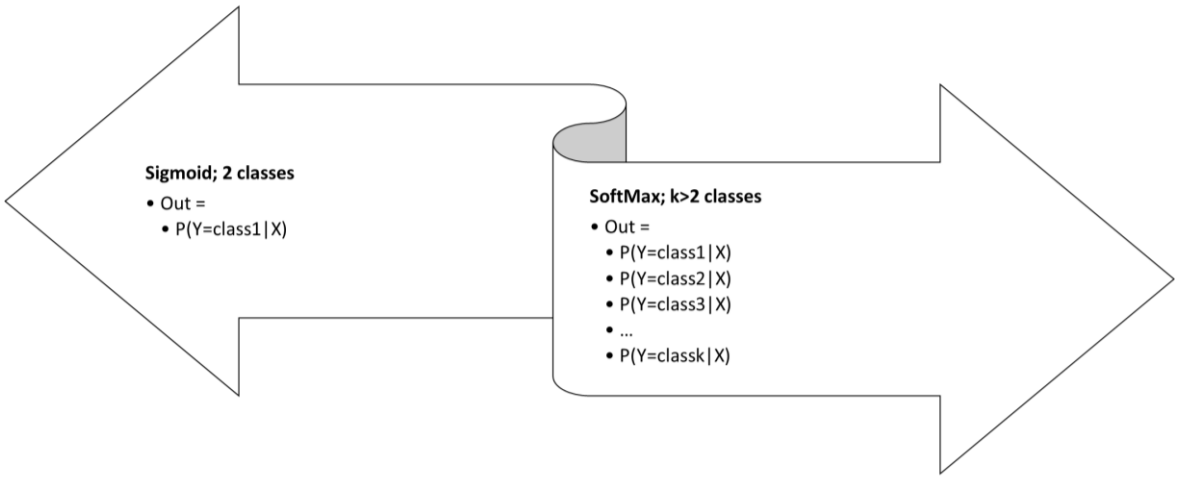


Figure 19 | Sigmoid vs SoftMax activation function | A figure illustrating the two activation functions employed for DeepHaem in this thesis: a sigmoid function (also called the logistic function) as well as the SoftMax function. Both functions output a probability (number between 0 and 1) of X being assigned to a specific class. The sigmoid function receives only one input and outputs a single probability; it performs binary classification (for 2 classes) while SoftMax receives and outputs a vector (with entries reflecting the relevant number of classes) and thus performs multiclass assignment.

Another attempt at improving the model's performance was to re-train different models with varying datasets. These models were as follows: 1 erythroid track & 1 heart track, 3 erythroid tracks & 3 heart tracks, 1 erythroid track & 12 dissimilar heart tracks, and 1 erythroid track & 29 dissimilar heart tracks.

Another suspicion arose that perhaps the model was performing adequately but the erythroid enhancer variants were not showing the expected damage scores as a direct result of their location within the peaks. This speculation was rooted in the fact that variants falling in the centre of a peak would be assigned more extreme damage scores compared to those near the edge of a peak (figure 20).

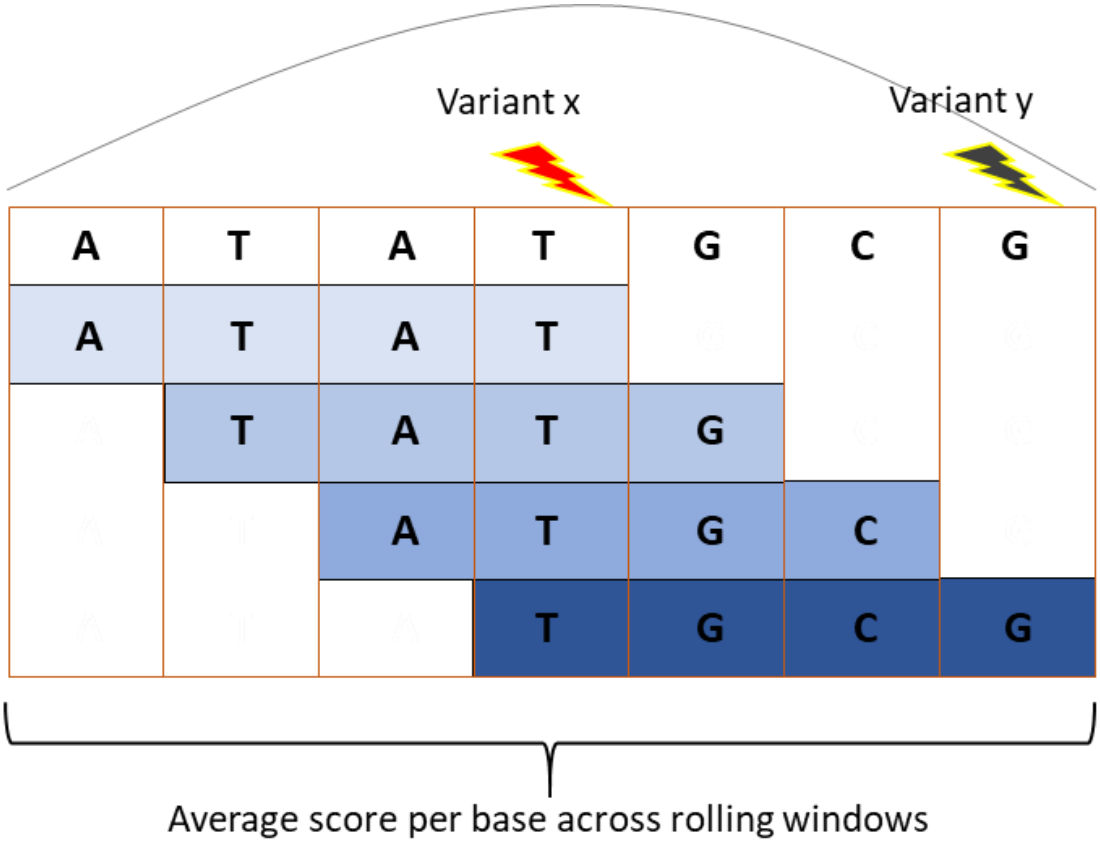


Figure 20 | Peak vs variant centering in DeepHaem's rolling window predictions | A diagram illustrating the implementation of a rolling window given an input DNA sequence, accompanied by 4 rolling window examples; each base receives a score as to how open or closed the chromatin is at that particular base and a final score is generated per base across all rolling windows. It also illustrates the hypothesis generated during troubleshooting stipulating that damage scores vary on the basis of where a particular variant is located within a given peak. X represents a variant residing near the middle of a peak whereas Y shows a variant residing near the edge of the peak; in line with this hypothesis, a more extreme damage score would be assigned to variant X when compared to variant Y as a result of the confidence for chromatin being open in the centre of the peak being higher in addition to more confidence across rolling windows.

To investigate this theory, rolling window predictions were performed in regions containing a handful of erythroid enhancer variants – first centering on the peak and then centering on the variant in the erythroid classifiers. In doing so, the predictions highlighted that the actual peaks were not always being detected regardless of the centering. This motivated good cause for going back to peak validation except this time with a focus on erythroid datasets specifically.

The constructed plan to investigate why erythroid classifier peaks were not being predicted accurately involved two different scenarios. Scenario 1 entailed checking if the model could predict a peak region that is captured in roughly 90% of all the classifiers (including erythroid tracks) and scenario 2 entailed checking if the model could predict a peak region that is only captured in the erythroid datasets and not the heart.

Deepmind

Preparing input files for Deepmind's Enformer (Avsec *et al.*, 2021) was not necessary, as the pre-trained and publicly available model already included datasets of interest for this analysis. Enformer includes a total of 5313 datasets pertaining to human data and 1643 to mouse. For the human genome, the 5213 datasets comprises 2131 TF ChIP-seq, 1860 histone ChIP-seq, 684 DNase-Seq or ATAC-Seq, and 638 CAGE tracks. Datasets of interest for this analysis were both DNase datasets from cardiac muscle cells in addition to the heart left ventricle female adult (53 years), heart left ventricle female embryo (136 days), heart left ventricle female embryo (101 day), and heart left ventricle female embryo (103 days) tracks. These datasets were originally sourced from ENCODE and pre-processing entailed clipping and scaling accordingly. See the data tracking sheet at github.com/domwest/DPhil-thesis/blob/main/data_tracking_sheet_for_thesis.xlsx for information regarding these datasets.

For each dataset, Enformer processes input in the form of 393216 base pair sequences and produces 128 base pair window predictions for the 114688 middle base pairs of 196608 bp input sequences; the authors of Enformer chose 128bp windows due to the representation of a well-researched length of regulatory elements containing motifs (Avsec *et al.*, 2021). The input undergoes one-hot encoding with 'N' values being converted to zeros.

The model architecture, including parameters, were pre-defined. Because Enformer has already been validated regarding its model performance and predictions by the authors (Avsec *et al.*, 2021), computational and visual inspection was not necessary. Regarding computational validation, the authors trained, evaluated, and tested the model on identical targets, genomic intervals, and with the

same Poisson negative log-likelihood loss function as carried out for Basenji2. For prediction validation, the contribution scores from Enformer emphasised validated enhancer-gene pairs with greater accuracy compared to Basenji2 and GTEx eQTLs (across numerous human tissues) were also leveraged to confirm the predictions made by the model.

To perform variant prioritisation, a positive control set first needed to be established. Using the erythroid enhancer variants as a positive control, as was done in DeepHaem's post analysis, was not possible as erythroid datasets are not included in the model. As an alternative, I received a set of 400 empirical HUVEC regulatory variants from the Hughes' lab comprising 138 gain variants, 149 loss variants, and 113 neutral variants. The neutral definition here refers to variants that were not shown to have an impact on the underlying chromatin. Enformer contains 2 datasets pertaining to 'endothelial cell of umbilical vein new-born' – this made it possible to test the HUVEC regulatory variants as a positive control set.

To prioritise likely functional HCM variants, damage score thresholds needed to be defined for both loss and gain of function variants. Due to the nature of gain, neutral, and loss variants being grouped together without clear definition between the three sets and the lack of a normal distribution, statistical tests on these datasets would not prove helpful in determining a cut-off. For this reason, interrogating the distribution of a positive control set – particularly capturing neutral variants which are known to have no effect on the underlying chromatin would help in establishing these thresholds. 2411 empirical HUVEC neutral variants were next obtained from the Hughes' lab to aid in determining variant prioritisation score cut-offs; the distribution of a neutral set could define an upper and lower prioritisation cut-off through the calculation of empirical p-values. Monte Carlo approach was employed to obtain an empirical p-value approximating the exact p-value using the equation: $\frac{(r+1)}{(n+1)}$. Thresholds were defined using this formula for a p-value of 0.05 in both endothelial classifiers. The most extreme thresholds were chosen as the finalised cut-off values.

The 4395 HCM GWAS lead and proxy variants were input to Enformer and predictions generated across each of the 5 chosen heart datasets. The prioritisation cut-offs were marked in all 5 heart datasets and variants scoring above the upper or below the lower threshold were obtained.

2.4 Linking Prioritised Variants to Target Genes

2.4.1 Overview

This chapter utilised cardiac muscle cell chromatin datasets (H3K27ac, H3K4me1, H3K4me3, ChIP TF, and DNase) for assigning regulatory elements to the regions in which prioritised variants were shown to reside; cardiac muscle cell H3K27ac and DNase datasets, in addition to intact Hi-C from the heart LV for enhancer-to-gene mapping; and the Human Heart Cell Atlas scRNA-Seq and scATAC-Seq datasets for peak-to-gene linkage. A handful of the loci within which variants were prioritised (in the previous chapter) are outlined in more detail in 2.4.8.

2.4.2 Annotating Regulatory Elements

REgulamentary was developed by Riva *et al.* (2024) – a tool aimed at *de novo* genome-wide annotation of cis-regulatory elements per cell-type of interest (any supplied eukaryotic cell type). It optimises on the histone marks H3K4me3 and H3K27ac as being associated with active promoters and H3K4me1 and H3K27ac with active enhancers. CTCF binding factor marks are also considered for delimiting boundary domains. This tool is integrated as a rule-based model rather than a supervised/unsupervised learning approach.

The REgulamentary pipeline was accessed on the BMRC server and its conda environment set up with all relevant packages pre-installed.

Output from the CATCH-UP pipeline was used as input to the REgulamentary pipeline. Specifically, the cardiac muscle cell .bed, .bam, and .bw files comprising chromatin accessibility data (DNase), histone mark ChIP-seq datasets (H3K4me1, H3K4me3, and H3K27ac), and ChIP-seq for the CTCF boundary domain.

REgulamentary assigned peaks in the input data to their corresponding regulatory elements. An all-encompassing 'RE' annotation was assigned to each peak representing the most likely regulatory

element when considering all available data. Other annotations also assign the relevant regulatory element but based on different aspects, for example, TSS_RE assigns the element given the proximity data (where the transcription start site is located).

The 53 prioritised variants post Deepmind Enformer were intersected with the peaks in which they resided as well as the corresponding regulatory element. This was employed to gain insight into the plausible causal mechanisms seen for each variant – whereby the 'loss of function' variants are presumed to reduce/eradicate a peak as opposed to the 'gain of function' variants which create/increase peaks. However, the gain of function variants may exist in completely closed regions of chromatin and in that case, this tool would not add much insight.

2.4.3 Variant-to-gene Assignment

The 53 prioritised variants, corresponding to 49 unique rsids (due to some variants being multi-allelic), were used as input to a Perl script (written by Chris Grace) to enable querying the Open Target database. This script pulls information from the Open Targets Genetics GraphQL API by sending a .json file per variant of interest and decoding the returned .json file into readable output with each variant's associated genes.

Not all of the rsids were present in the Open Targets database and so results were only generated for 47 out of the 49 rsids; results were missing for rs66925847 and rs149015746. The top gene (with the highest variant-to-gene (V2G) score) was extracted per variant and the results tabulated.

2.4.4 Enhancer-to-gene Mapping

To act as an extra layer of assignment regarding enhancer element sites specifically (and with the intention of uncovering additional target genes), the activity-by-contact (ABC) model was employed.

Code was made available by the authors on GitHub (Fulco *et al.*, 2019) with instructions on how to set up and run the model; the model was cloned and implemented in BMRC.

A cardiac muscle cell DNase-Seq bam file, H3K27ac histone chip file, and heart left ventricle intact hic file (resolution: 5000) were provided as input to the pipeline.

The ABC model is capable of creating prediction files for subsequent variant overlap analysis – via the use of a less stringent filter on the ABC score as well as a reduction of regions for overlap with variants. These steps were introduced as a result of Nasser *et al.* (2021) and their attempts to integrate ABC predictions with variants. Considering that DNA sequences critical for enhancer functions (such as TF footprints) are predominantly located in the central nucleosome-free area, the authors also reduced the ~500bp regions by 150bp on each side, resulting in a region of ~200bp centred on the peak.

This prediction output file (for variant overlap) was filtered to include only enhancer coordinates which intersected with the 53 prioritised variant positions. The chosen ABC score threshold for enhancer-gene links was the advised threshold of 0.027 which represents that at 70% recall with the precision of 0.56 (Fulco *et al.*, 2019).

2.4.5 Peak-to-gene Linkage

Input data for the analysis of region-specific chromatin accessibility (ArchR) pipeline (Granja *et al.*, 2021) was downloaded from the Human Heart Cell Atlas. In particular, the Heart Global dataset which comprised of scRNA-Seq and scATAC-Seq data. The scATAC-Seq data contained 47 sample fragment files which were subsequently loaded into ArrowFiles within an ArchR project on the

BMRC server. The reference genome build 38 was specified for the project so as to align with the data.

Per-cell quality control (QC) steps were conducted for the scATAC-Seq data which included evaluating the number of unique nuclear fragments, assessing signal-to-background ratio, computing TSS enrichment scores, and checking the distribution of fragment sizes. Doublet inference was also performed; this entails the removal of doublets (single droplets which receive more than one nucleus) as these result in multiple reads (from more than one cell) being captured as a single cell. Default parameters, optimised for human data, were employed during the QC steps.

Subsequently, a technique called latent semantic indexing (LSI) was employed to reduce the dimensionality of sparse insertion counts in the scATAC-Seq data in addition to a batch effect correction tool called Harmony. After dimensionality reduction, clustering of single cells in the scATAC-Seq data ensued. ArchR is capable of assigning clusters with confidence; however, further steps are required to determine which cell type(s) are represented by each cluster. An initial attempt, using the scATAC-Seq data alone, entailed the use of cell type-specific marker genes and their expression levels (estimated based on chromatin accessibility data). Thereafter, the scATAC-Seq cells were aligned with scRNA-Seq cells to enhance cluster identification.

The integration was first performed in an unconstrained manner followed by constrained. Unconstrained integration entailed attempting to align all cells in the scATAC-Seq data to cells in the scRNA-Seq data. However, constrained integration uses existing knowledge of cell types to improve the alignment; the group to constrain was specified to be atrial and ventricular cardiomyocytes versus all other heart cell types.

Pseudo-bulk replicates were generated whereby data from each cell in the scATAC-Seq dataset was combined into a single pseudo-sample mimicking a bulk ATAC-Seq experiment. Next, the iterative overlap peak merging procedure was employed to define peak sets and the MACS2 algorithm for peak calling. The ArchR peak2gene linkage analysis was then implemented. Ultimately, peaks

assigned to ventricular cardiomyocytes were extracted and intersected with the 53 prioritised variants to find plausible target genes for these variants.

2.4.6 Gene Ontology Exploration

Tadros *et al.* (2025) employed MAGMA for gene set enrichment of GWAS loci associated with HCM. Their MAGMA analysis encompassed 9987 Gene Ontology (GO) gene sets from the Molecular Signatures Database (MSigDB, v.6.2). Their findings showed several significant gene sets associated with muscle function, contractility, and sarcomeric activity.

DAVID and gProfiler were employed so that gene ontology terms and pathways could be enriched given a set of prioritised genes as the input rather than a set of GWAS variants (as was performed by Tadros *et al.* (2025)).

The input genes were prioritised through various approaches. Such approaches included the Open Targets and FUMA assigned gene per locus (given the lead SNP), the top-performing gene in the quadrant analyses (using both the bulk heart LV proximity-based approach as well as the single-cell cardiomyocyte TAD-based approach), as well as genes identified through peak-to-gene linkage via ArchR, the ABC model's enhancer-gene links, and also the Open Targets assigned gene given each of the prioritised variants.

The input genes list included 203 unique HGNC gene symbols captured across the 68 HCM loci.

DAVID

Given the input genes list, 195 gene Ensembl IDs were captured in the DAVID database.

Before performing functional annotation clustering, the data sources to be consulted during the analysis needed to be selected. Only the gene ontology (GO) 5 terms (biological process (BP), cellular compartment (CC), and molecular function (MF)) were selected for functional annotation

clustering. GO 5 terms pertain to a particular classification at the fifth level of the GO hierarchy suggesting it represents a highly detailed biological function. These terms were selected to try and enrich for detailed processes as opposed to vague, broader functions. The default stringency option of medium was changed to high to limit the number of clusters being output. Next, functional annotation clustering was performed using only Kegg pathway annotations.

gProfiler

All the genes in the input list were captured in the gProfiler output.

gProfiler automatically implements multiple testing correction via its custom in-house algorithm called SCS (set counts and sizes); a significance threshold of 0.01 was applied given the gProfiler tool's default SCS multiple correction implementation. Despite Bonferroni correction and Benjamini-Hochberg false discovery rate (FDR) also being available, SCS was chosen in this analysis as it has been seen to account for the uneven distribution of gene set annotations and also performed better in distinguishing between significant and trivial results compared to other methods (Reimand *et al.*, 2007).

2.4.7 Motif Analysis

Homer v5.1 was executed on the BMRC server. This motif discovery tool required input in the form of a peak or bed file containing genomic coordinates. The former entails the execution of findMotifs.pl while the latter, findMotifsGenome.pl. Thus, findMotifsGenome.pl was applicable for these analyses. The operation of findMotifsGenome.pl involves several systematic steps to identify high-quality motifs which are detailed in figure 21.

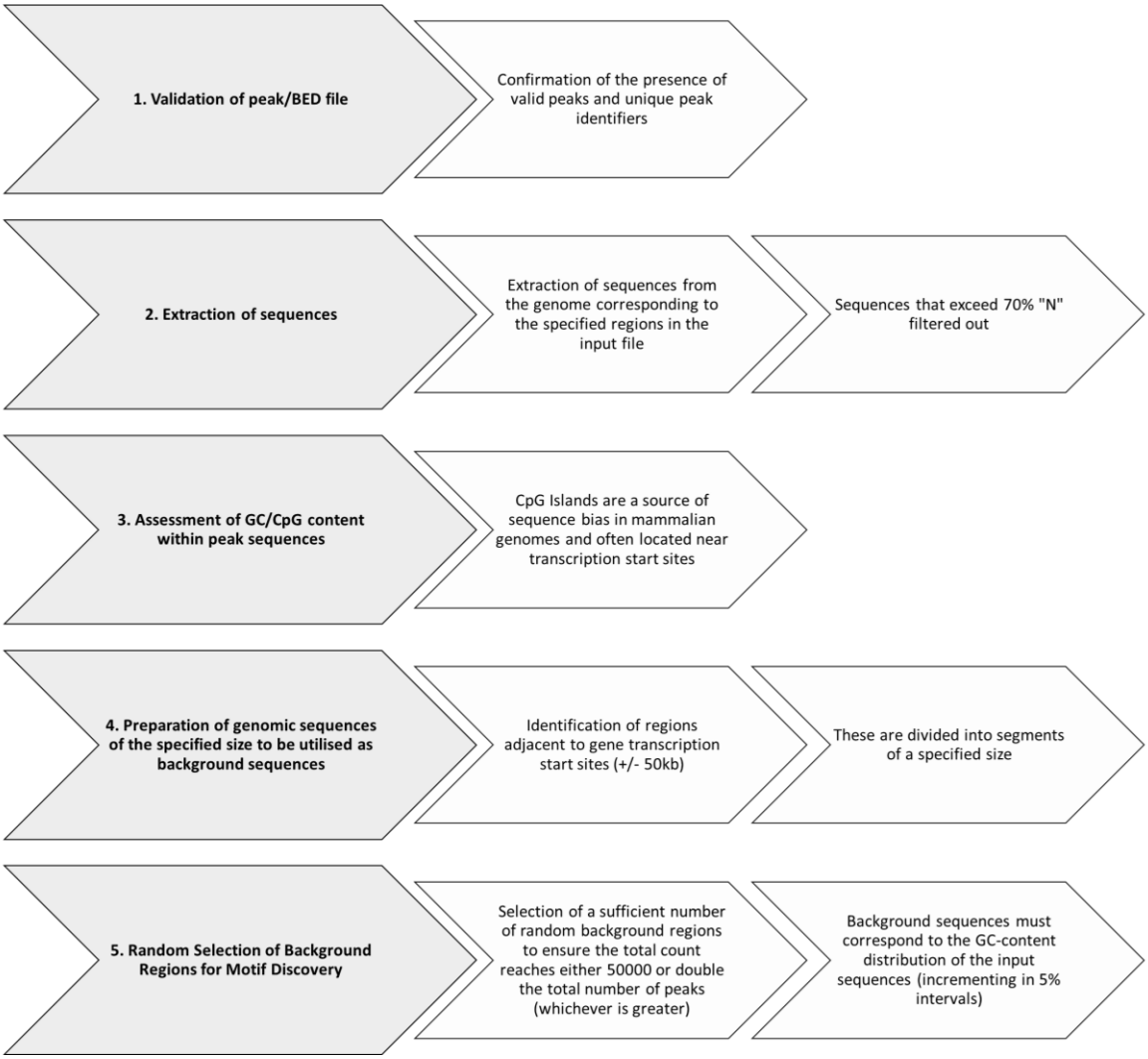


Figure 21 | Motif discovery via HOMER pipeline | A flowchart displaying the overarching steps employed by HOMER's findMotifsGenome Perl script. Additional insight is provided for each step along the process of peak inspection, sequence extraction, GC/CpG content assessment, background sequence preparation, and finally motif discovery.

Homer was executed for each of the prioritised 53 variants. Their genomic positions were specified with an added 20bp either side of the variant resulting in 40bp sized regions.

40bp sized regions were chosen due to the variants' positions and their immediate surroundings being of particular interest. Furthermore, for a motif to be disrupted by any of the 53 variants one would expect them to overlap with the motif. Consequently, motifs residing within close/immediate proximity to the variant would prove to be insightful as opposed to those residing further upstream or downstream.

HOMER output was generated for each of the 53 variants. HOMER outputs two different types of results – one for *de novo* motif identification and another for known motifs. The latter were of interest in this analysis, with the aim being to check for enrichment of a well-elucidated motif that may be affected and prove causal for HCM. Each variant's output file was read in, filtered to contain only the most significant motif identified (based on the p-value as well as the number of target versus background sequences containing the motif). These top hits per variant were then concatenated into an overarching output file and inspected.

2.4.8 Case Studies

For each of the 53 prioritised variants corresponding to 29 loci, evidences collected over the course of this thesis were collated and each locus (i.e. case study) is discussed in more detail in this subsection. Within each case study, the evidences are explored in more depth with plausible target gene(s) highlighted.

The case study approach entailed identifying possible causal genes and explaining the likely mechanisms per 53 variants prioritised by Deepmind Enformer. For each of the 29 case studies, they were grouped by lead variant whereby each of the prioritised variants per lead variant (/locus) and the supporting evidences were examined. This process was split into two broader categories. First, cross-loci exploration was performed whereby the spread of heart left ventricle (LV)/cardiomyocyte expression values as well as the specificity thereof was inspected across all 29 prioritised loci. Secondly, each locus was unpacked in isolation to uncover plausible causal targets and mechanisms. This was done via literature scanning, assessing STRING interactions for prominent gene products, locus zoom plot visualisations to assess the proximity of prioritised variants to the genes (as well as the associated signal) and ultimately concluding target gene(s) were determined per locus – in accordance with evidences collated throughout this thesis.

2.4.9 Summarising implicated genes

Each of the prioritised functional variants and their loci were shown in the form of an overarching summary table with their assigned implicated gene(s) as determined throughout the case studies.

A ‘gene weight’ column was added to this table indicating the number of evidences supporting the implicated genes; evidences were extracted from the case study summary tables, including: the Tadros (2025) paper, quadrant approach, other gene prioritisation tools, and element-based identification of plausible target genes. A score was assigned for a particular gene if it appeared in any of the category’s sub-fields. The score (out of 4) tallied per gene was shown. Per locus, those with a gene weight ≥ 2 were retained and placed in a final ‘implicated gene(s)’ column.

Thereafter, the Mouse Genome Informatics (MGI) database (Baldarelli *et al.*, 2024) was consulted (regarding the assigned mouse phenotypes) for loci with more than one implicated gene. In the event that the phenotype data was able to discern which was the likely causal gene, this was detailed following the summary table. For genes with cardiac phenotypes listed in MGI, the gene weight score was also increased by 1.

2.4.10 Performing a PheWAS lookup for prioritised functional variants

The Open Targets Genetics database contains GWAS credible sets with disease/phenotype associations. A credible set refers to a set of genetic variants residing near a signal which is assumed (with a certain probability) to encompass the causal variant. Post-fine-mapping, a posterior probability (PP) of causality is assigned to each variant in the region and those forming the top 95% likelihood of containing the causal variant delineate the credible sets within the Open Targets Platform.

Each of the prioritised functional variants from Deepmind’s Enformer output were queried against this database to extract available disease/phenotype associations. The results were first sorted by the assigned PP scores – five entries representing those with the top PPs were extracted. If any of the

five rows corresponded to a PP < 0.05 (5% likelihood of the hypothesis being true) these were discarded.

Across all prioritised variants' disease/phenotype associations, over-represented terms were ranked to gauge which phenotypes were most frequently being shown and likely key for HCM.

3

Gene prioritisation

3.1 Introduction

3.1.1 Gene prioritisation tools

Candidate gene prioritisation strategies can be classified into two main categories (Weeks *et al.*, 2023). One category is locus-based methods, which use GWAS data to link causal variants to their corresponding genes through the incorporation of protein-coding variants, genomic distance, enhancer-gene maps, or expression quantitative trait loci (eQTLs). The other category is similarity-based methods which make use of patterns detected among trait-associated genes focusing on shared functions, pathways, and networks.

Locus-based methods

Locus-based methods can prioritise genes via statistical fine mapping whereby marginal associations are resolved to causal variants to implicate target gene(s) directly. Alternatively, locus-based methods integrate other regulatory aspects such as epigenetic information (enhancer-promoter correlation, promoter capture hi-c, and activity by contact) with GWAS signal. Similarly, eQTL information in addition to GWAS signal may be leveraged as is seen in colocalisation (COLOC), colocalisation posterior probability (CLPP), transcriptome-wide association studies (TWAS), and summary data-based Mendelian randomisation (SMR). Locus-based methods may also simply involve GWAS signal with proximity awareness whereby the closest gene to the index variant is prioritised – an approach adopted in MAGMA analyses.

Locus-based method: colocalisation

Quantitative trait loci mapping (QTL) involves the identification of genetic or molecular markers that are linked to quantitative traits. Such traits may include gene expression (eQTLs), protein expression (pQTLs), or biometric traits such as low-density lipoprotein (LDL), high-density lipoprotein (HDL), blood pressure, and body mass index (BMI) (Cano-Gamez and Trynka, 2020). The GTEx project is recognised as one of the most comprehensive sources of eQTL data (Melé *et al.*, 2015).

Colocalisation employs quantitative trait loci (QTL) mapping and integrates characteristics such as gene expression to calculate the correlation of GWAS variants with intermediate traits. Colocalisation has proven to be especially insightful in the study of coronary artery disease as well as other common, complex diseases (Cano-Gamez and Trynka, 2020). Miller *et al.* (2015) found sixty-six CAD lead variants to have eQTL associations with genes situated within 1 Mb of their location – also reinforcing that most noncoding variants at CAD loci affect local gene expression.

Two popular tools in this domain are coloc.abf and coloc.susie. The former is a simple Bayesian approach (Giambartolomei *et al.*, 2014) and does not require linkage disequilibrium (LD) estimates thus proving applicable for large-scale analyses (Wallace, 2020) while the latter incorporates Sum of Single Effects (SuSiE) regression allowing for the consideration of multiple causal variants within a genomic region (Wallace, 2021) and improves accuracy in complex regions with multiple signals. It is advised that researchers utilise both coloc.abf and coloc SuSiE when multiple causal variants are present, resorting to a single-variant assumption approach only if coloc SuSiE fails to identify any colocalisations.

Locus-based method: rare variant burden analyses

A general assumption regarding the genetic framework of rare versus common disease is that rare variants are linked to Mendelian phenotypes while common variants are associated with an increased risk for common disease phenotypes. However, rare variants can also contribute to common diseases, and common variants can influence the expression of rare, Mendelian conditions.

Most HCM causal mutations are missense, with the exception of those in *MYBPC3*, which exhibit loss-of-function (LoF) effects (Marian and Braunwald, 2017). Loss-of-function variants can present in various forms (Glicksberg *et al.*, 2019); for instance, frameshift variants disrupt the reading frame (including insertions and deletions as well as splice junction variants), while some variants may result in the gain or loss of stop codons, and others may occur at splice sites, leading to the exclusion of exons or inclusion of introns. Most loss-of-function variants result in nonsense-mediated decay triggered by presence of a premature stop; hence they have an equivalent effect on the gene and can be analysed as a group.

Testing for rare LoF variant burden is generally more straightforward than for missense variants as LoF variants are rare, are clearly defined in the general population, and can be pooled to provide statistical power. Approximately 2% of the population carries a missense variant in any given gene and so such variants may appear relevant to disease phenotypes by coincidence. Missense variants can have opposite effects on phenotype and so cannot reliably be pooled. Additionally, being able to differentiate between pathogenic and benign missense variants remains a significant challenge (Gerasimavicius *et al.*, 2022). Walsh *et al.* (2017) categorised rare variants associated with hypertrophic cardiomyopathy (HCM) as those with a minor allele frequency (MAF) threshold of 0.0001, based on the prevalence of the most common pathogenic variants. Additionally, protein structures can be scrutinised regarding case versus control variants and their clustering so as to test the likelihood of case variants being pathogenic. AlphaFold along with Google Deepmind's AlphaMissense tool, for example, annotates missense variants as pathogenic or benign through the analysis of their structural context as well as evolutionary conservation within proteins – utilising predicted protein structures to identify residues that are critical for protein function (Minton, 2023).

Similarity-based methods

Similarity-based method: PoPs

The Polygenic Priority Score tool (PoPs) was built with the aim of improving gene prioritisation approaches through the integration of many different data sources.

PoPs integrates gene features from the following sources: 8479 pathway features (DEPICT curated from GO, KEGG, REACTOME, Mouse genome database), 8717 protein-protein interaction features (PPI neighbours from InWeb), and 40546 gene expression-based features (77 datasets which are mostly single cell). The gene expression datasets were used to compute whole dataset features (gene loadings from PCA and ICA), within-cluster features (average expression and gene loadings from PCA), as well as between-cluster comparison features (t-statistic for differential expression as well as differentially expressed genes).

3.1.2 Leveraging RNA-Seq data for gene prioritisation

RNA sequencing (RNA-Seq) has facilitated comprehensive exploration of the transcriptome and its relationship to disease, particularly regarding those arising from altered states of gene expression (Wang *et al.*, 2009). The RNA-Seq protocol begins with the transformation of RNA to cDNA which then undergoes fragmentation, adapter ligation, and index ligation before each fragment can be sequenced (Koch *et al.*, 2018). The resulting raw read data is aligned and mapped to genes resulting in a table of read counts per gene. Analysts typically obtain the data at this stage whereby normalisation ensues as well as downstream analyses.

Normalisation metrics

Various normalisation metrics exist for RNASeq count data which can be distinguished by the factors for which they normalise (Introduction to DGE – ARCHIVED, 2022). As outlined by Zhao *et al.* (2020), for two genes being expressed at the same level, an increased sequencing depth for only one

would lead to a proportional rise in read counts. Likewise, a gene with a shorter length will yield fewer read counts compared to a longer gene even if the sequencing depth is constant. Counts per million (CPM) account for sequencing depth alone while transcripts per million (TPM), reads per kilobase million (RPKM), and fragments per kilobase million (FPKM) account for both sequencing depth as well as gene length. Because CPM does not account for gene length, it is only applicable for gene count comparisons between replicates of a specific sample group. Contrastingly, TPM, RPKM, and FPKM are applicable to within-sample analyses as well as between samples of the same group. Unlike RPKM and FPKM, the total of all TPM values remain constant across samples, allowing for easier comparison of the read counts associated with genes in each sample.

Exploiting different pipelines

Not only have a diverse array of gene prioritisation methods been made available, varying pipelines have been employed across different research groups. Some groups have incorporated both similarity- and locus-based methods. Aragam *et al.* (2022) devised a consensus-based prioritisation framework which honed in on genes closest to the sentinel variant, associated with a relevant monogenic disorder, implicated in drug trials or MR, prioritised by PoPs, annotated as an eQTL in a CAD relevant tissue, showed a relevant phenotype in knockout mice, contained a protein-altering variant, or those with a previously identified CAD-associated rare variant. Genes being prioritised by two or more concordant predictors were marked as likely causal while those prioritised by three or more were deemed high confidence causal genes.

Other groups have leveraged expression data alone to perform gene prioritisation. Kosti *et al.* (2016) sought to elucidate the relationship between gene expression and protein abundance through the differential correlation of genes (mrna) and proteins in cancerous tissues compared to normal tissues; the authors found marked differential protein-gene expression correlation in cancer. Similarly, Ahn *et al.* (2020) conducted a median-based analysis of tissue-specific gene expression using data from GTEx identifying genes with a heart-specific or elevated expression level in the heart.

3.1.3 Key objectives

The key objectives for this Chapter are to:

1. Resort to the basics (similar to that described by Ahn *et al.* (2020) and Kosti *et al.* (2016)) through use of gene expression data alone to prioritise genes in the tissue and cell type of interest. Resorting to a basic RNASeq analysis (to investigate small expression changes locally) was the chosen route when considering that colocalisation approaches require eQTL data from heart tissue – which is extracted from only a few samples and has less power as a result.
2. Employ COLOC due to its effectiveness in prioritising genes post-GWAS in addition to the relevance of eQTLs when assessing a complex, polygenic disorder.
3. Run PoPs due to its consideration of the full polygenic signal (not only genome-wide significant variants).
4. Conduct rare variant burden analyses for assessing a possible monogenic component in the association with HCM as well as to add biological tractability.

3.1.4 Computational approaches employed

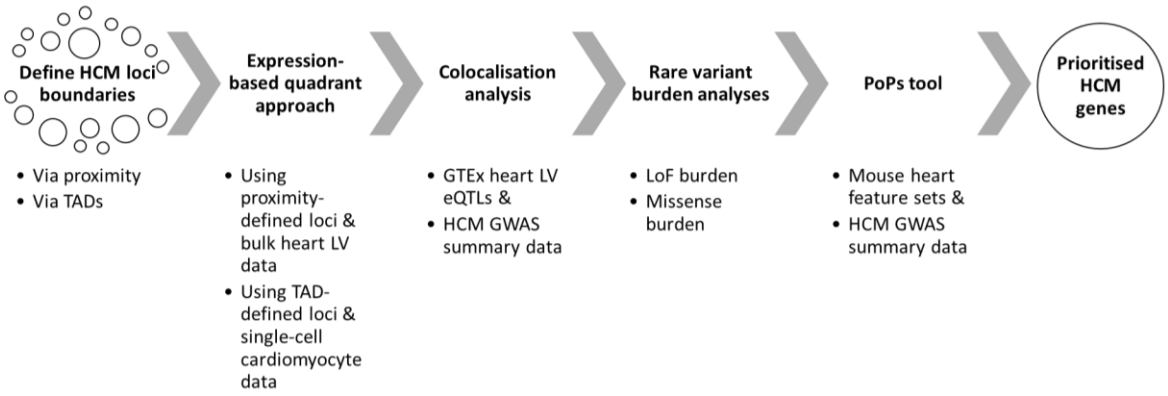


Figure 22 | Gene prioritisation approach | A flowchart illustrating the computational tools implemented in this chapter to obtain a final set of prioritised HCM genes. HCM loci were first defined via either a proximity method or a topologically-associating domain (TAD) approach; an expression-based quadrant approach was devised to assess expression and specificity values across genes; colocalisation was performed given the HCM GWAS summary data as well as heart LV

eQTLs from GTEx; a summary-based gene prioritisation tool, polygenic priority score (PoPs), was also leveraged.

Figure 22 summarises the computational approaches employed in this chapter: defining HCM loci followed by employing gene prioritisation tools (expression based, colocalisation, rare variant burden analyses, and the polygenic priority score (PoPs) tool) to output a set of prioritised genes.

3.2 Defining HCM loci

3.2.1 Proximity-based approach

The defined 1Mbp loci given each of the 68 HCM lead variants delimited via a scanning approach are shown in table 7.

Table 7 | Proximity-based loci coordinates | A table depicting the assigned proximity-based start and end coordinates (build 37) for each of the 68 HCM loci. The HCM GWAS lead variant rsid (from which the 500kbp scanning both up- and down-stream commenced) is shown alongside the relevant chromosome and final start and end locus coordinates. The loci are referred to by the 'lead rsid' or the 'opentarget/fuma gene symbol'; the former depicting the most significant rsid in a particular locus and the latter, the likely target gene(s) as output by Open Target and FUMA when given the lead rsid as input (Tadros et al., 2025) which also acts as a helpful comparator.

Lead rsid	Chromosome	Locus start (bp; b37)	Locus end (bp; b37)	Opentarget/fuma gene symbol
rs2503715	1	1644107	2644107	*FAAP20*
rs11121483	1	5763792	6763792	*ICMT, RNF207*
rs1048302	1	15840879	16840879	*HSPB7*
rs6699769	1	78004264	79004264	*DNAJB4*
rs2810883	1	88832165	89832165	*KYAT3*
rs74139614	1	213785221	214785221	*PROX1, SMYD2*
rs547048	1	236301085	237301085	*ACTN2, HEATR1*
rs11687178	2	11084197	12084197	*ROCK2*

rs3845778	2	36633127	37633127	*STRN*
rs1009358	2	64776452	65776452	*CEP68*
rs2540277	2	102926177	103926177	*MFSD9*
rs6747402	2	179259434	180259434	*CCDC141*
rs295114	2	200695602	201695602	*TYW5, SPATS2L*
rs1699340	3	12081416	13081416	*TSEN2, MKRN2*
rs7612736	3	13776238	14776238	*SLC6A6, LSM3*
rs6796333	3	51859678	52859678	*GLYCTK*
rs79502300	3	69356753	70356753	*MITF*
rs4894803	3	171300256	172300256	*PLD1, FNDC3B*
rs11730129	4	15532948	16532948	*TAPT1, PROM1*
rs75096272	5	36130209	37130209	*SLC1A3*
rs2191446	5	56511404	57511404	*ACTBL2*
rs11748963	5	138230037	139230037	*PROB1*
rs6914805	6	15781187	16781187	*GMPR*
rs3176326	6	36147289	37147289	*CDKN1A*
rs816379	6	53499025	54499025	*MLIP*
rs12210733	6	118153075	119153075	*PLN*
rs1159974	6	125590277	126590277	*HEY2*
rs2184370	6	149664378	150664378	*PPIL4, LRP11*
rs66520020	7	127938284	128938284	*CCDC136*
rs2945236	8	7668394	8668394	*CLDN23, PRAG1*
rs12216858	8	11283118	12283118	*FDFT1*
rs7824244	8	21302432	22302432	*XPO7, DOK2*
rs7835298	8	124052303	125052303	*FBXO32*
rs35006907	8	125359817	126359817	*MTSS1*
rs13297385	9	134044226	135044226	*RAPGEF1*
rs2645210	10	3598453	4598453	*KLF6*
rs2182400	10	29207551	30207551	*SVIL*
rs2177843	10	74909877	75909877	*MRPS16, MYOZ1*

rs11196085	10	114005037	115005037	*TCF7L2*
rs17617337	10	120926884	121926884	*BAG3*
rs12270374	11	13875079	14875079	*RRAS2*
rs182427065	11	57252909	58252909	*OR6Q1*
rs11571171	11	100474887	101474887	*PGR*
rs10841524	12	20089183	21089183	*PDE3A*
rs58747679	12	25848304	26848304	*SSPN*
rs7487962	12	114861930	115861930	*TBX3*
rs116904997	12	120168534	121168534	*GCN1, PXN*
rs9515201	13	110540798	111540798	*COL4A2*
rs41306688	13	113578558	114578558	*ADPRHL1*
rs113907726	14	52816867	53816867	*ERO1A*
rs8019721	14	71334210	72334210	*SIPA1L1*
rs140763915	14	88474767	89474767	*GPR65, PTPN21*
rs8006225	14	94719657	95719657	*GSC*
rs2002854	15	78553014	79553014	*ADAMTS7*
rs8033459	15	84753258	85753258	*NMB*
rs4984741	16	525464	1525464	*LMF1*
rs11860560	16	1738931	2738931	*MLST8*
rs28575249	16	3766062	4766062	*SRL*
rs28768976	17	43188317	44188317	*CRHR1, PLEKHM1*
rs17817857	17	52876942	53876942	*HLF*
rs7210446	17	63807014	64807014	*PRKCA*
rs497999	18	32172836	33172836	*MAPRE2*
rs2644262	18	33723566	34723566	*TPGS2*
rs6566955	18	55422789	56422789	*NEDD4L*
rs12460541	19	45812077	46812077	*DMWD*
rs62222424	21	30030131	31030131	*CCT8*
rs5760054	22	23661717	24661717	*MMP11, SMARCB1*
rs139920	22	40226749	41226749	*ADSL*

A total of 1977 gene IDs were captured across all 68 loci (output by the Ensembl API). At any given locus, the minimum number of identified genes was 9, while the maximum was 100.

Brodie *et al.* (2016) explored the distance between a causal variant and its associated disease gene to find that affected genes are typically located within 2 Mbp of the variant as opposed to simply being the nearest gene. Complementarily, Wells *et al.* (2019) identified a 2 Mbp pattern in Hi-C data in which the most constrained regulatory elements were seen to target the most critical genes – implying that the 1 Mbp search radius used in the proximity-based approach could have been broadened to avoid the exclusion of potential candidate genes. However, expanding the search areas would also lead to the inclusion of more genes thereby increasing the burden of multiple testing.

3.2.2 TAD-based approach

As described in chapter 2.2.2, the intersection of hic heart left ventricle contact domain data was intersected with the HCM lead and proxy variants for the derivation of newly defined loci regions (as an improvement on proximity-based methods). Table 8 depicts each locus post-intersection and whether or not its lead/proxy variant(s) were present in a TAD (sized small, moderate, or large).

Table 8 | Intersection of HCM variants with various sized TADs | A table showing which loci (labelled by their lead variant in the first column) could be defined following intersection of lead/proxy variants with topologically associating domains (TADs). Columns 2-4 show which loci had their lead variant residing in an assigned TAD (in addition to whether this was greater than 500kbp, less than 200kbp, or somewhere between). Column 5 indicates which loci did not have their lead variant captured within an assigned TAD but their proxy variants did. Column 6 shows which loci could not be intersected with TADs given their lead or proxy variants and thus needed to be assigned manually.

Lead rsid	**Lead variant in big TAD(>=500kbp)**	**Lead variant in small TAD(<=200kbp)**	**Lead variant in moderate TAD(200kbp<=x <=500kbp)**	**Only proxy variant in TADs not lead variant)**	**Neither lead nor proxy variant in TADs**
rs2503715		Y			
rs11121483					Y
rs1048302		Y			
rs6699769		Y			
rs2810883					Y
rs74139614		Y			
rs547048		Y			
rs11687178				Y	
rs3845778					Y
rs1009358				Y	
rs2540277					Y
rs6747402		Y			
rs295114		Y			
rs1699340					Y
rs7612736			Y		
rs6796333			Y		
rs79502300		Y			
rs4894803			Y		

rs11730129			Y		
rs75096272		Y			
rs2191446					Y
rs11748963				Y	
rs6914805		Y			
rs3176326		Y			
rs816379			Y		
rs12210733					Y
rs1159974				Y	
rs2184370		Y			
rs66520020		Y			
rs2945236			Y		
rs12216858					Y
rs7824244					Y
rs7835298		Y			
rs35006907					Y
rs13297385		Y			
rs2645210					Y
rs2182400		Y			
rs2177843					Y
rs11196085					Y
rs17617337					Y
rs12270374		Y			
rs182427065	Y				
rs11571171		Y			
rs10841524					Y
rs58747679			Y		
rs7487962					Y
rs116904997		Y			
rs9515201			Y		

rs41306688					Y
rs113907726		Y			
rs8019721	Y				
rs140763915		Y			
rs8006225			Y		
rs2002854				Y	
rs8033459		Y			
rs4984741			Y		
rs11860560					Y
rs28575249					Y
rs28768976					Y
rs17817857		Y			
rs7210446			Y		
rs497999		Y			
rs2644262					Y
rs6566955		Y			
rs12460541					Y
rs62222424				Y	
rs5760054		Y			
rs139920			Y		
total	2	26	12	6	22

The defined loci for each of the 68 HCM lead variants given the TAD-based approach are shown in table 9.

Table 9 | TAD-based loci coordinates | A table depicting the assigned topological-associating domain (TAD)-based start and end coordinates (build 38) for each of the 68 HCM loci. The HCM GWAS lead variant rsid (from which the intersection of this variant or its proxy variants with TAD data pinpointed locus coordinates) is shown alongside the relevant chromosome and final start and end locus coordinates. The loci are referred to by the 'lead rsid' or the 'opentarget/fuma gene symbol'; the former depicting the most significant rsid in a particular locus and the latter, the likely target gene(s) as output by Open Target and FUMA when given the lead rsid as input (Tadros et al., 2025) which also acts as a helpful comparator.

Lead rsid	**Chromosome**	**Locus start (bp; b38)**	**Locus end (bp; b38)**	**Locus size (kbp)**	**Opentarget/fuma gene symbol**
rs1009358	2	64900000	65185000	285	*CEP68*
rs1048302	1	15755000	16065000	310	*HSPB7*
rs10841524	12	19700000	20800000	1100	*PDE3A*
rs11121483	1	6000000	6700000	700	*ICMT, RNF207*
rs11196085	10	112400000	113600000	1200	*TCF7L2*
rs113907726	14	52520000	53300000	780	*ERO1A*
rs11571171	11	100500000	101430000	930	*PGR*
rs1159974	6	124880000	126000000	1120	*HEY2*
rs11687178	2	11125000	11575000	450	*ROCK2*
rs116904997	12	120140000	120290000	150	*GCN1, PXN*
rs11730129	4	15500000	16300000	800	*TAPT1, PROM1*
rs11748963	5	139280000	139515000	235	*PROB1*
rs11860560	16	2150000	2550000	400	*MLST8*
rs12210733	6	117950000	118600000	650	*PLN*
rs12216858	8	11400000	12000000	600	*FDFT1*
rs12270374	11	13750000	14450000	700	*RRAS2*
rs12460541	19	45350000	45900000	550	*DMWD*

rs13297385	9	131320000	131720000	400	*RAPGEF1*
rs139920	22	39500000	40800000	1300	*ADSL*
rs140763915	14	87910000	88550000	640	*GPR65, PTPN21*
rs1699340	3	12100000	12850000	750	*TSEN2, MKRN2*
rs17617337	10	119600000	119875000	275	*BAG3*
rs17817857	17	54970000	55830000	860	*HLF*
rs182427065	11	57300000	58520000	1220	*OR6Q1*
rs2002854	15	78600000	79200000	600	*ADAMTS7*
rs2177843	10	73475000	73650000	175	*MRPS16, MYOZ1*
rs2182400	10	28600000	29460000	860	*SVIL*
rs2184370	6	149550000	149925000	375	*PPIL4, LRP11*
rs2191446	5	57300000	58500000	1200	*ACTBL2*
rs2503715	1	2125000	2400000	275	*FAAP20*
rs2540277	2	102450000	102900000	450	*MFSD9*
rs2644262	18	36275000	36825000	550	*TPGS2*
rs2645210	10	3150000	4800000	1650	*KLF6*
rs2810883	1	88100000	88900000	800	*KYAT3*
rs28575249	16	3940000	4400000	460	*SRL*
rs28768976	17	45550000	46300000	750	*CRHR1, PLEKHM1*
rs2945236	8	8200000	9000000	800	*CLDN23, PRAG1*
rs295114	2	200000000	200800000	800	*TYW5, SPATS2L*
rs3176326	6	36415000	36750000	335	*CDKN1A*
rs35006907	8	124575000	125000000	425	*MTSS1*
rs3845778	2	36750000	38000000	1250	*STRN*
rs41306688	13	113250000	113575000	325	*ADPRHL1*
rs4894803	3	171600000	172350000	750	*PLD1, FNDC3B*
rs497999	18	34880000	35350000	470	*MAPRE2*
rs4984741	16	650000	2500000	1850	*LMF1*
rs547048	1	236100000	236800000	700	*ACTN2, HEATR1*
rs5760054	22	23500000	23850000	350	*MMP11, SMARCB1*

rs58747679	12	25400000	26410000	1010	*SSPN*
rs62222424	21	29070000	39840000	10770	*CCT8*
rs6566955	18	57500000	59050000	1550	*NEDD4L*
rs66520020	7	128400000	128880000	480	*CCDC136*
rs6699769	1	77125000	78115000	990	*DNAJB4*
rs6747402	2	178785000	179125000	340	*CCDC141*
rs6796333	3	51900000	53350000	1450	*GLYCTK*
rs6914805	6	16220000	17100000	880	*GMPR*
rs7210446	17	65200000	66675000	1475	*PRKCA*
rs74139614	1	213995000	214125000	130	*PROX1, SMYD2*
rs7487962	12	114650000	115900000	1250	*TBX3*
rs75096272	5	36150000	36800000	650	*SLC1A3*
rs7612736	3	14100000	14650000	550	*SLC6A6, LSM3*
rs7824244	8	21900000	22175000	275	*XPO7, DOK2*
rs7835298	8	123140000	123750000	610	*FBXO32*
rs79502300	3	69100000	70000000	900	*MITF*
rs8006225	14	94685000	96000000	1315	*GSC*
rs8019721	14	71130000	72415000	1285	*SIPA1L1*
rs8033459	15	84600000	85045000	445	*NMB*
rs816379	6	53980000	55230000	1250	*MLIP*
rs9515201	13	110100000	110700000	600	*COL4A2*

A total of 1636 gene IDs were captured across the 68 loci and output by the Ensembl API. At any given locus, the minimum number of identified genes was 3, while the maximum reached 268.

Although the contact domain data utilised in this analysis are specific to the Heart Left Ventricle as opposed to the cell type of interest (cardiomyocytes), TADs are primarily independent of cell type suggesting the insight can prove useful regardless of data availability for cardiomyocytes (Way *et al.*, 2017)). This paired with the fact that single-cell Hi-C and super resolution microscopy contact domain data is scarce (Bintu *et al.*, 2018) and cardiomyopathy-relevant cardiomyocyte Hi-C data is difficult to obtain (Nakato *et al.*, 2023), TADs presented a reasonable interim solution for defining loci.

3.3 Devised Expression-Based Quadrant Approach

3.3.1 Bulk proximity-based quadrant approach

Each of the 68 loci's prioritised gene is shown in table 10 along with the associated median expression and fold change (based on heart LV data), the quadrant in which this gene resides, and the diagonal distance within said quadrant (given the expression and fold change).

Table 10 | Proximity-defined loci with bulk LV quadrant approach gene prioritisation | A table depicting the prioritised gene per 68 loci following proximity-based locus demarcation and implementation of the quadrant approach (given bulk tissue heart left ventricle expression data). The lead rsid relevant to locus is shown alongside the prioritised gene symbol and associated values output by the quadrant approach (a median expression column representing the expression in heart LV tissue, the fold change LV value (the heart LV-specificity when compared to all 52 other tissues in GTEx which excluded heart atrial appendage (AA) and heart LV tissues), the quadrant within which the gene resides (1, 2, 3, or 4) and finally a diagonal distance within its residing quadrant (a combined score for its extent of heart expressivity as well as specificity).

Lead rsid	Gene symbol	Median expr	Fold change LV	Quadrant	Diagonal distance
rs6796333	*TNNC1*	3598.0	1075.233	2	3755.227
rs12460541	*CKM*	3002.0	737.773	2	3091.329
rs1048302	*HSPB7*	1712.0	109.045	2	1715.469
rs12216858	*GATA4*	45.94	1068.0	2	1068.987

rs12210733	*PLN*	827.7	73.541	2	830.961
rs547048	*ACTN2*	614.8	403.412	2	735.337
rs11860560	*RPL3L*	101.6	529.442	2	539.103
rs2177843	*SYNPO2L*	84.88	410.395	2	419.081
rs6566955	*ALPK2*	23.62	317.9	2	318.777
rs66520020	*FLNC*	297.7	30.208	2	299.229
rs28575249	*SRL*	174.6	185.38	2	254.658
rs5760054	*CHCHD10*	245.6	5.405	2	245.659
rs35006907	*NDUFB9*	187.0	2.037	2	187.011
rs816379	*MLIP*	45.85	158.377	2	164.88
rs6699769	*NEXN*	158.9	11.615	2	159.324
rs41306688	*ADPRHL1*	133.3	57.818	2	145.299
rs6747402	*TTN*	67.69	115.099	2	133.528
rs4984741	*SSTR5-AS1*	1.931	125.41	2	125.425
rs17617337	*BAG3*	88.54	1.426	2	88.551
rs74139614	*SMYD2*	77.04	6.139	2	77.284
rs6914805	*GMPR*	70.67	3.979	2	70.782
rs7835298	*KLHL38*	15.25	65.676	2	67.423
rs2644262	*FHOD3*	64.41	11.105	2	65.36
rs116904997	*ACADS*	58.43	2.496	2	58.483
rs2184370	*PPP1R14C*	47.37	26.042	2	54.056
rs13297385	*PPAPDC3*	50.94	6.673	2	51.375
rs2182400	*SVIL*	49.0	2.86	2	49.083
rs8033459	*ALPK3*	46.09	14.009	2	48.172
rs2002854	*DNAJA4*	43.88	2.296	2	43.94
rs7824244	*SLC39A14*	36.31	1.5	2	36.341
rs10841524	*PDE3A*	32.52	10.296	2	34.111
rs2540277	*TMEM182*	21.44	23.306	2	31.668
rs3176326	*PPIL1*	30.42	1.486	2	30.456
rs11121483	*RNF207*	27.61	2.916	2	27.764

rs75096272	*SLC1A3*	21.24	2.309	2	21.365
rs1159974	*HEY2*	19.83	4.146	2	20.259
rs1699340	*CAND2*	15.95	2.994	2	16.229
rs11748963	*PROB1*	12.56	6.528	2	14.155
rs497999	*DTNA*	13.73	2.114	2	13.892
rs7612736	*GRIP2*	9.989	5.0	2	11.17
rs79502300	*MITF*	9.495	1.306	2	9.584
rs3845778	*VIT*	6.413	2.07	2	6.739
rs11730129	*FGFBP2*	2.691	2.914	2	3.966
rs139920	*UQCRFS1P1*	2.259	1.338	2	2.626
rs1009358	*AC007318.5*	2.01	1.365	2	2.429
rs11571171	*PGR*	1.726	1.708	2	2.428
rs2645210	*LINC00702*	1.407	1.421	2	2.0
rs62222424	*LINC00189*	0.301	3.011	1	3.026
rs7487962	*RP11-116D17.3*	0.684	2.866	1	2.947
rs17817857	*RP11-515O17.2*	0.218	2.417	1	2.426
rs4894803	*RP11-408H1.3*	0.007	2.223	1	2.223
rs182427065	*SERPING1*	79.21	0.376	4	79.211
rs2503715	*GNB1*	54.99	0.279	4	54.991
rs9515201	*COL4A2*	46.63	0.765	4	46.636
rs12270374	*PSMA1*	43.65	0.613	4	43.654
rs28768976	*PLCD3*	18.33	0.94	4	18.354
rs113907726	*PSMC6*	14.55	0.604	4	14.563
rs8006225	*SERPINA3*	14.15	0.645	4	14.165
rs2191446	*GPBP1*	14.16	0.374	4	14.165
rs295114	*C2orf47*	12.97	0.822	4	12.996
rs11687178	*LPIN1*	10.59	0.444	4	10.599
rs11196085	*ZDHHC6*	10.43	0.326	4	10.435
rs140763915	*PTPN21*	7.947	0.73	4	7.98
rs2810883	*GTF2B*	5.656	0.257	4	5.662

rs8019721	*PCNX*	5.521	0.406	4	5.536
rs2945236	*MFHAS1*	4.715	0.55	4	4.747
rs58747679	*BHLHE41*	4.587	0.278	4	4.595
rs7210446	*CEP112*	2.095	0.716	4	2.214

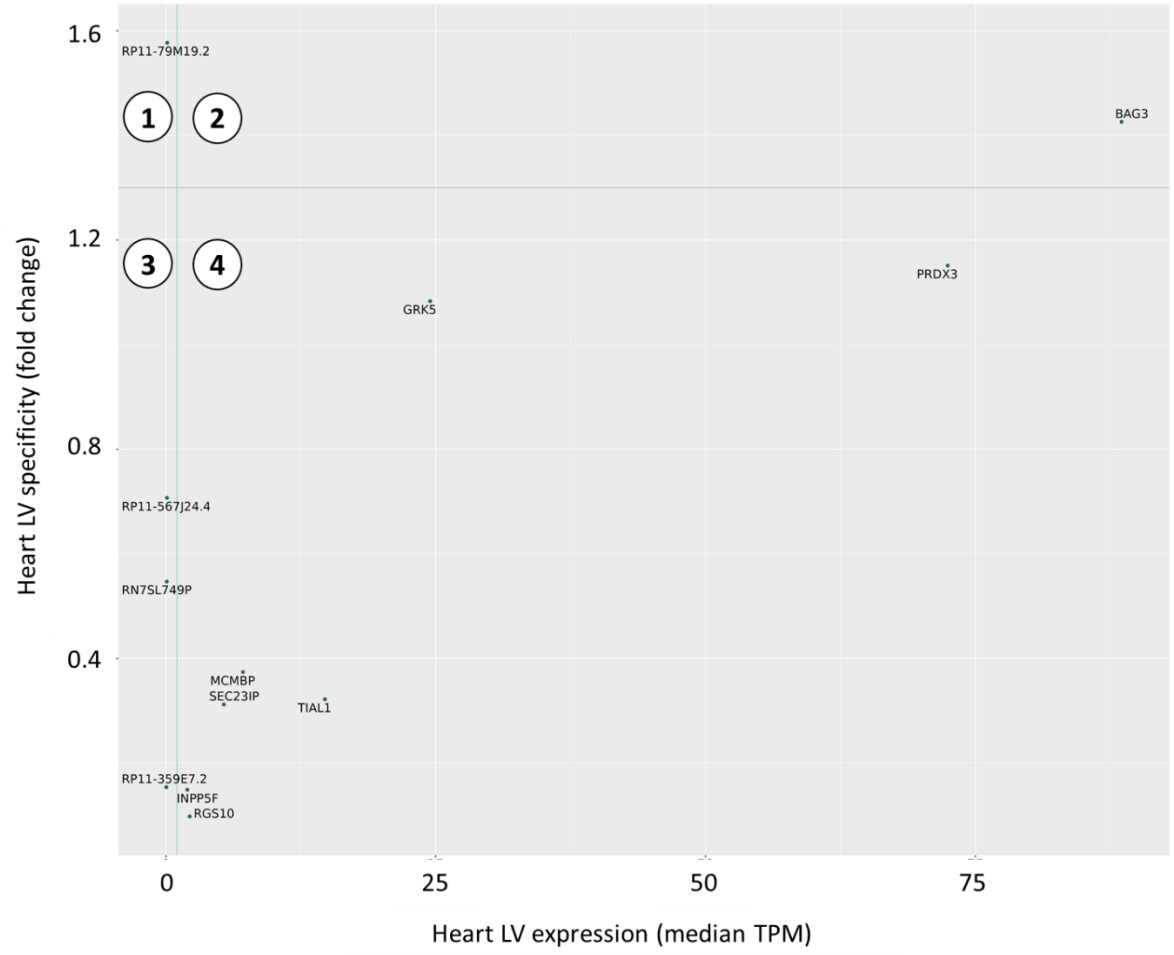


Figure 23 | BAG3 quadrant approach visualisation as prioritised with bulk LV data | A quadrant scatter plot with the heart LV specificity ratio plotted against the median TPM in the heart left ventricle (LV). A proximity-based quadrant approach, whereby 500kbp were scanned both upstream and downstream of the lead variant, was employed to define the locus and its residing genes; bulk heart LV expression data from GTEx was utilised in the quadrant analysis. BAG3 is the prioritised gene in this locus, being the only gene to reside in quadrant 2, with a median TPM value of ~89 and specificity ~1. Genes are prioritised in the order of quadrants 2, 1, 4, 3 (quadrants circled).

**Specificity ratio cut-off of 1.3 based on lab experience.*

***Expression cut-off of 1.04 determined by median of all heart genes' median TPMs in the heart LV.*

The prioritisation of *BAG3* in the rs17617337 locus is shown in figure 23 as an example of this approach where *BAG3* (HCM positive control gene) was clearly the top performing gene – residing in quadrant 2 with both high heart LV specificity as well as expression. All other genes in this locus are scattered throughout quadrants 1, 3, and 4.

All 68 loci bulk LV quadrant plots are provided at GitHub (https://github.com/domwest/DPhil-thesis/blob/main/proximity_based_bulk_LV_qa_plots.zip).

Single-cell RNA sequencing provides more in-depth insight regarding gene expression at the individual cell level thus proving suitable for investigating cellular diversity and rare cell populations. For this reason, single-cell RNA sequencing data was leveraged in the next sub-section.

3.3.2 Single-cell TAD-based quadrant approach

Each of the 68 loci's prioritised gene is shown in table 11 along with the associated mean transcript count and fold change (based on cardiomyocyte data), the quadrant in which this gene resides, and the diagonal distance within said quadrant (given the expression and fold change).

Table 11 | TAD-defined loci with single-cell cardiomyocyte quadrant approach gene prioritisation | A table depicting the prioritised gene per 68 loci following topologically associating domain (TAD)-based locus demarcation and implementation of the quadrant approach (given single cell cardiomyocyte expression data). The lead rsid relevant to locus is shown alongside the prioritised gene symbol and associated values output by the quadrant approach (a mean count column representing the expression in single cell cardiomyocytes, the fold change LV value (the cardiomyocyte-specificity when compared to all other 8 heart cell types i.e. Mural, endothelial cells (EC), fibroblasts (FB), Lymphoid, neuronal cells (NC), Myeloid, Mast, and adipocytes (AD)), the quadrant within which the gene resides (1, 2, 3, or 4) and finally a diagonal distance within its residing quadrant (a combined score for its extent of heart expressivity as well as specificity).

Lead rsid	Gene symbol	Mean count	Fold change CM	Quadrant	Diagonal distance
rs4984741	*RPL3L*	1.008	99.169	2	99.174

rs28575249	*SRL*	1.813	95.157	2	95.174
rs41306688	*ADPRHL1*	1.309	85.019	2	85.029
rs6796333	*TNNC1*	1.778	8.602	2	8.784
rs17617337	*BAG3*	0.164	8.132	2	8.134
rs1159974	*TPD52L1*	0.771	77.791	2	77.794
rs8019721	*RGS6*	2.921	75.220	2	75.277
rs3176326	*CPNE5*	1.175	73.007	2	73.016
rs7835298	*AC090193.1*	0.098	72.731	2	72.731
rs58747679	*BHLHE41*	0.043	7.412	2	7.412
rs12270374	*SPON1-AS1*	0.036	7.371	2	7.371
rs497999	*DTNA*	4.524	66.795	2	66.948
rs62222424	*CLIC6*	0.181	63.085	2	63.085
rs11860560	*ECI1*	0.285	6.848	2	6.854
rs7824244	*DMTN*	0.039	6.849	2	6.849
rs7487962	*TBX3*	0.048	6.366	2	6.366
rs17817857	*PCTP*	0.199	6.284	2	6.287
rs182427065	*OR9Q1*	0.024	6.281	2	6.282
rs2184370	*LRP11*	0.087	6.202	2	6.202
rs74139614	*PROX1*	0.497	57.609	2	57.611
rs3845778	*LINC00211*	0.019	52.900	2	52.900
rs8006225	*DICER1-AS1*	0.095	5.507	2	5.508
rs4894803	*PLD1*	1.546	5.236	2	5.459
rs11121483	*RNF207*	1.217	47.014	2	47.029
rs2182400	*SVIL*	7.160	46.312	2	46.862
rs12460541	*PPP1R13L*	0.940	46.707	2	46.717
rs11730129	*PROM1*	0.047	44.809	2	44.809
rs8033459	*ALPK3*	1.116	41.819	2	41.834
rs1048302	*HSPB7*	2.634	41.668	2	41.751
rs6914805	*GMPR*	0.728	41.407	2	41.414
rs2002854	*ANKRD34C-AS1*	0.052	41.254	2	41.254

rs2177843	*SYNPO2L*	0.193	40.925	2	40.925
rs9515201	*NAXD*	0.091	4.534	2	4.535
rs140763915	*AL162171.2*	0.049	4.521	2	4.521
rs12210733	*PLN*	4.320	38.275	2	38.518
rs2191446	*PLK2*	0.118	31.008	2	31.008
rs79502300	*LMOD3*	0.471	30.840	2	30.843
rs2540277	*TMEM182*	0.887	30.463	2	30.476
rs295114	*AC007163.1*	0.122	3.550	2	3.552
rs1009358	*SLC1A4*	0.099	3.463	2	3.464
rs2945236	*MFHAS1*	0.316	3.256	2	3.271
rs7210446	*AXIN2*	0.030	28.935	2	28.935
rs11687178	*GREB1*	0.234	27.135	2	27.136
rs12216858	*GATA4*	2.388	26.996	2	27.101
rs816379	*MLIP-AS1*	2.448	257.066	2	257.078
rs2644262	*FHOD3*	7.406	249.796	2	249.906
rs13297385	*PLPP7*	0.569	24.396	2	24.403
rs35006907	*LINC00964*	0.075	22.059	2	22.059
rs113907726	*TXNDC16*	0.145	2.982	2	2.985
rs116904997	*PXN-AS1*	0.062	2.711	2	2.712
rs28768976	*MAPT*	0.319	19.305	2	19.308
rs11571171	*PGR*	0.430	19.198	2	19.203
rs66520020	*FLNC-AS1*	0.611	180.624	2	180.625
rs1699340	*CAND2*	0.349	18.710	2	18.713
rs2810883	*PKN2-AS1*	0.541	18.515	2	18.523
rs7612736	*GRIP2*	1.082	170.300	2	170.304
rs139920	*FAM83F*	0.018	17.513	2	17.513
rs5760054	*CHCHD10*	1.074	16.995	2	17.029
rs6747402	*TTN*	72.583	140.773	2	158.384
rs11196085	*NRAP*	3.390	147.485	2	147.524
rs75096272	*AC008957.2*	0.673	142.451	2	142.453

rs11748963	*PROB1*	0.031	14.059	2	14.059
rs547048	*ACTN2*	8.646	134.597	2	134.874
rs10841524	*AC129102.1*	0.134	127.851	2	127.851
rs6699769	*NEXN*	11.001	127.006	2	127.482
rs6566955	*ALPK2*	1.770	126.029	2	126.041
rs2503715	*PRKCZ*	0.083	11.582	2	11.582
rs2645210	*LINC00702*	0.297	100.111	2	100.111

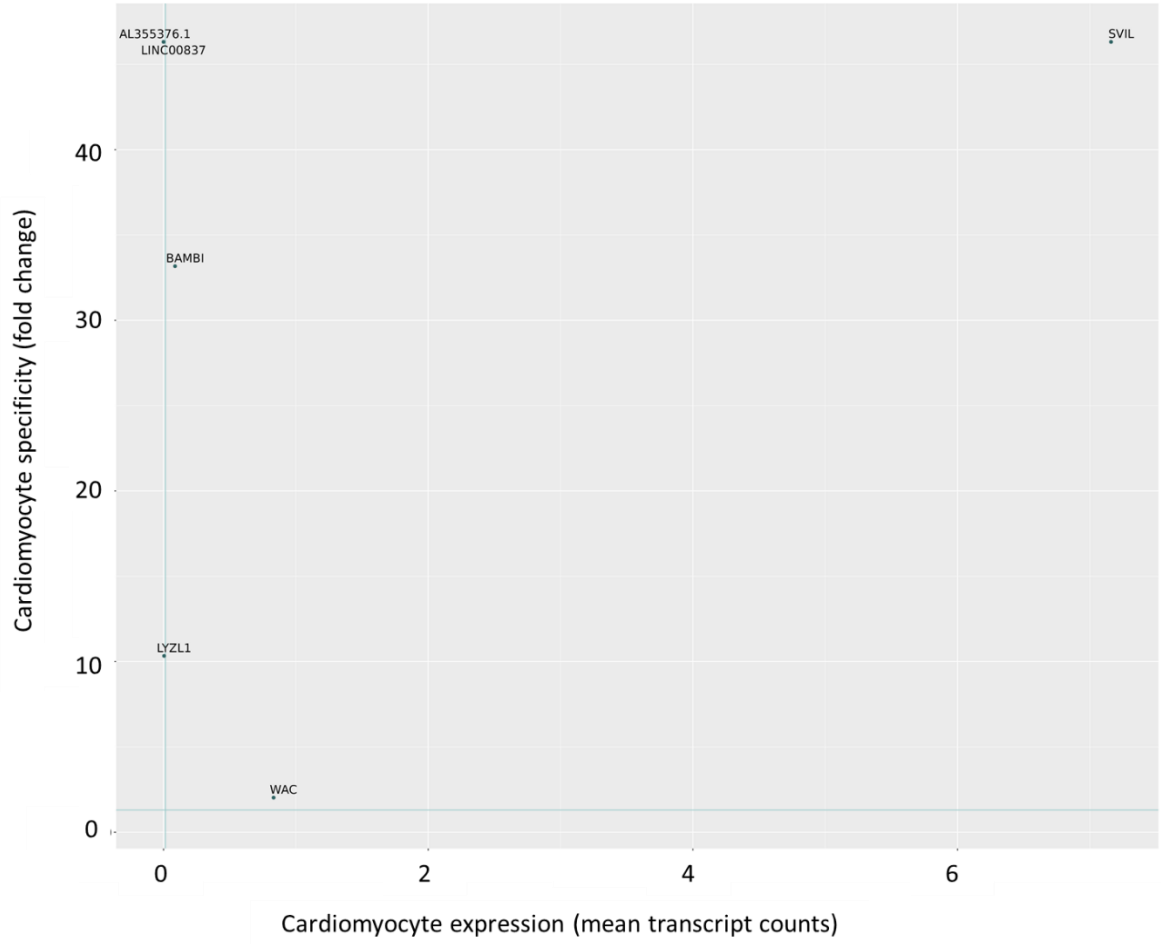


Figure 24 | SVIL quadrant approach visualisation as prioritised with single-cell cardiomyocyte data | A quadrant scatter plot with the cardiomyocyte specificity ratio plotted against the mean transcript counts in the cardiomyocyte. A TAD-based quadrant approach, whereby TADs containing the lead or proxy variant(s), were identified to define the locus and its residing genes; single-cell cardiomyocyte expression data generated by Reichart et al. (2022) was utilised in the quadrant analysis. SVIL is the prioritised gene with convincing expression values (~7 mean transcript counts)

in cardiomyocytes compared to the other genes which show less than 1 as well as a high cardiomyocyte specificity (~46). Genes are prioritised in the order of quadrants 2, 1, 4, 3.

**Specificity ratio cut-off of 1.3 based on lab experience.*

***Expression cut-off of 0.014 determined by median of all cardiomyocyte genes' mean transcript counts.*

Another example of the quadrant approach output can be seen in the prioritisation of *SVIL* (a novel HCM gene) in the rs2182400 locus (as is shown in figure 24) where *SVIL* was clearly the top performing gene – residing in quadrant 2 with both high cardiomyocyte specificity as well as expression.

All 68 loci single-cell cardiomyocyte quadrant plots are provided at GitHub (https://github.com/domwest/DPhil-thesis/blob/main/tad_based_sc_CM_qa_plots.zip).

3.3.3 Comparison of the two quadrant approaches

Loci whereby the same gene was prioritised given both quadrant approaches are shown in table 12, different genes but with the single-cell approach prioritised gene showing the highest expression in cardiomyocytes in table 13, and different genes with the single-cell approach prioritised gene being most expressed in a non-cardiomyocyte heart cell type in table 14.

Table 12 | HCM loci whereby the same gene was prioritised via both bulk LV and single-cell cardiomyocyte quadrant approach | A table depicting 32 of the 68 HCM GWAS loci whereby the prioritised gene following both the proximity-based locus demarcation with bulk tissue heart left ventricle expression data and the topologically associating domain (TAD)-based locus demarcation with single-cell cardiomyocyte expression data are the same. An additional column, 'Most expressed in CM', shows whether the prioritised gene was most expressed in the cardiomyocyte compared to all 8 other heart cell types (Mural, endothelial cells (EC), fibroblasts (FB), Lymphoid, neuronal cells (NC), Myeloid, Mast, and adipocytes (AD)) during the single-cell analysis.

Lead rsid	**Bulk and SC gene symbol**	**Most expressed in CM**
rs1048302	*HSPB7*	Y
rs10841524	*PDE3A*	Y
rs11121483	*RNF207*	Y
rs11571171	*PGR*	Y
rs11748963	*PROB1*	Y
rs12210733	*PLN*	Y
rs12216858	*GATA4*	Y
rs13297385	*PLPP7*	Y
rs1699340	*CAND2*	Y
rs17617337	*BAG3*	Y
rs2177843	*SYNPO2L*	Y
rs2182400	*SVIL*	Y
rs2540277	*TMEM182*	Y
rs2644262	*FHOD3*	Y
rs2645210	*LINC00702*	Y
rs28575249	*SRL*	Y
rs2945236	*MFHAS1*	
rs295114	*MAIP1*	Y
rs41306688	*ADPRHL1*	Y
rs497999	*DTNA*	Y
rs547048	*ACTN2*	Y

rs5760054	*CHCHD10*	Y
rs58747679	*BHLHE41*	
rs6566955	*ALPK2*	Y
rs6699769	*NEXN*	Y
rs6747402	*TTN*	Y
rs6796333	*TNNC1*	Y
rs6914805	*GMPR*	Y
rs7487962	*TBX3*	
rs75096272	*SLC1A3*	
rs7612736	*GRIP2*	Y
rs8033459	*ALPK3*	Y

Table 13 | HCM loci whereby a different gene was prioritised via bulk LV vs single-cell cardiomyocyte quadrant approach; single-cell gene most expressed in cardiomyocyte | A table depicting 30 of the 68 HCM GWAS loci whereby the prioritised gene following the proximity-based locus demarcation with bulk tissue heart left ventricle expression data versus the topologically associating domain (TAD) -based locus demarcation with single-cell cardiomyocyte expression data was different – but also where the gene prioritised via the latter was most expressed in cardiomyocytes. An additional column, 'Most expressed in CM', shows whether the prioritised gene was most expressed in the cardiomyocyte compared to all 8 other heart cell types (Mural, endothelial cells (EC), fibroblasts (FB), Lymphoid, neuronal cells (NC), Myeloid, Mast, and adipocytes (AD)) during the single-cell analysis.

Lead rsid	**Bulk gene symbol**	**SC gene symbol**	**Most expressed in CM**
rs1009358	*RAB1A*	*SLC1A4*	Y
rs11196085	*ZDHHC6*	*NRAP*	Y
rs113907726	*PSMC6*	*TXNDC16*	Y
rs1159974	*HEY2*	*TPD52L1*	Y
rs11687178	*LPIN1*	*GREB1*	Y
rs116904997	*ACADS*	*PXN-AS1*	Y
rs11860560	*RPL3L*	*ECI1*	Y

rs12460541	*CKM*	*PPP1R13L*	Y
rs139920	*UQCRFS1P1*	*FAM83F*	Y
rs17817857	*COX11*	*PCTP*	Y
rs182427065	*SERPING1*	*OR9Q1*	Y
rs2002854	*DNAJA4*	*ANKRD34C-AS1*	Y
rs2191446	*GPBP1*	*PLK2*	Y
rs2503715	*GNB1*	*PRKCZ*	Y
rs2810883	*GTF2B*	*PKN2-AS1*	Y
rs28768976	*PLCD3*	*MAPT*	Y
rs3176326	*PPIL1*	*CPNE5*	Y
rs35006907	*NDUFB9*	*LINC00964*	Y
rs3845778	*VIT*	*CDC42EP3*	Y
rs4894803	*TNFSF10*	*PLD1*	Y
rs4984741	*SSTR5-AS1*	*RPL3L*	Y
rs62222424	*LINC00189*	*CLIC6*	Y
rs66520020	*FLNC*	*FLNC-AS1*	Y
rs74139614	*SMYD2*	*PROX1*	Y
rs7835298	*KLHL38*	*FBXO32*	Y
rs79502300	*MITF*	*LMOD3*	Y
rs8006225	*SERPINA3*	*DICER1-AS1*	Y
rs8019721	*PCNX1*	*RGS6*	Y
rs816379	*MLIP*	*MLIP-AS1*	Y
rs9515201	*COL4A2*	*NAXD*	Y

Table 14 | HCM loci whereby a different gene was prioritised via bulk LV vs single-cell cardiomyocyte quadrant approach; single-cell gene most expressed in non-cardiomyocyte cell type | A table depicting 6 of the 68 HCM GWAS loci whereby the prioritised gene following the proximity-based locus demarcation with bulk tissue heart left ventricle expression data versus the topologically associating domain (TAD)-based locus demarcation with single-cell cardiomyocyte expression data was different – and where the gene prioritised via the latter was most expressed in non-cardiomyocyte heart cell types. An additional column, 'Most expressed cell type', shows which heart cell type (Mural, endothelial cells (EC), fibroblasts (FB), Lymphoid, neuronal cells (NC), Myeloid, Mast, or adipocytes (AD)) the prioritised gene was most expressed in during the single-cell analysis.

Lead rsid	Bulk gene symbol	SC gene symbol	Most expressed cell type
rs11730129	*FGFBP2*	*PROM1*	ec
rs12270374	*PSMA1*	*SPON1-AS1*	ad
rs140763915	*PTPN21*	*SPATA7*	myeloid
rs2184370	*PPP1R14C*	*LRP11*	fb
rs7210446	*CEP112*	*AXIN2*	ad
rs7824244	*SLC39A14*	*DMTN*	ad

Table 12 catalogues 32 genes which have higher associated confidence regarding their plausible impact (seen by their expression) due to the same gene being prioritised in both quadrant approaches. Some of the positive control genes (already associated with HCM) in this table were: *HSPB7*, *BAG3*, *SVIL*, *FHOD3*, *ADPRHL1*, and *ALPK3*.

The majority of the prioritised genes in table 12 were most expressed in the cardiomyocyte as opposed to other heart cell types with the exception of four genes (*MFHAS1*, *BHLHE41*, *SLC1A3*, and *TBX3*). Despite *MFHAS1*, *BHLHE41*, and *SLC1A3* being most expressed in non-cardiomyocyte cell types – this increased expression is only marginal when compared to that in the cardiomyocyte. However, *TBX3* is expressed to a greater extent by a factor of 4 in endothelial cells as opposed to cardiomyocytes. Recent clinical data trends emphasise HCM (among other cardiovascular diseases) and its pathogenesis being linked to CMD (coronary microvascular dysfunction) which has been attributed to endothelial cell dysfunction (ECD) (Gao *et al.*, 2022); however, *TBX3* is still being

expressed in the cardiomyocyte to some extent and so a disease mechanism could be possible in both.

Table 13 highlights genes that were not unanimously prioritised across both prioritisation approaches, but which were most expressed in cardiomyocytes and thus have an increased plausibility of association. Table 14 depicts the same scenario but where the single-cell expression approach prioritised a gene most expressed in a different heart cell type (not cardiomyocytes). Although it is known that HCM is a disease of the cardiomyocyte, it is not inconceivable for other heart cell types to be implicated/involved in the polygenic aspect of the condition.

Genes prioritised in the quadrant approach analysis which have a prior association with cardiomyopathy were *RNF207*, *ACTN2*, *RPL3L*, and *FBXO32*. A missense mutation in the *RNF207* gene has been reported in connection with DCM within a cohort of European Dobermanns (Dutton *et al.*, 2025). Four missense mutations in the *ACTN2* gene (p.Gly111Val, p.Ala119Thr, p.Met228Thr, and p.Thr247Met) have been detected within the actin-binding domain of patients diagnosed with HCM (Chiu *et al.*, 2010)(Girolami *et al.*, 2014)(Prondzynski *et al.*, 2019)(Theis *et al.*, 2006). The missense mutation (p.Met228Thr) compromises *ACTN2*'s stability causing the ubiquitin-proteasomal system to be triggered – a process previously associated with cardiomyopathies. Additionally, the absence of functional *ACTN2* results in energy deficits due to mitochondrial dysfunction (Broadway-Stringer *et al.*, 2023). *RPL3L*, a paralog of *RPL3*, plays a significant role in the dynamics of translation elongation; the absence of *RPL3L*-containing ribosomes in male mice lacking *RPL3L* led to compromised cardiac contractility with the biggest impact for transcripts associated with cardiac muscle contraction and DCM (Shiraishi *et al.*, 2023). Mutated *FBXO32* has been associated with DCM as a result of the up-regulation of ER-stress mediated apoptosis (Al-Yacoub *et al.*, 2021).

Through the prioritisation of genes via the quadrant approach, the following cardiac mechanisms also came to light: altered cardiac development, electrophysiology, contractility, fibrosis, hypertrophy, myofibril assembly, cardiac edema, left ventricular noncompaction cardiomyopathy (LVNC), atrial fibrillation (AF), and heart failure (HF). Of top interest in this regard, were 3 genes

with a known role in the heart, *SRL*, *ADPRHL1*, and *SYNPO2L* – all of which are discussed at the end of this chapter.

3.4 Assessing Regulatory Roles Through Colocalisation of Common Variants and eQTL Data

3.4.1 Coloc abf

Among the 68 HCM GWAS loci, 20 genes prioritised via colocalisation (utilising coloc.abf) are shown in table 15.

Table 15 | Colocalisation (coloc.abf) result | A table illustrating the coloc.abf results given an input list of 161 genes representing the most prioritised genes across the 68 loci (from Open Target, FUMA, the bulk quadrant approach, and the SC quadrant approach). Each locus is represented by its lead SNP and the search space as marked by its topologically associating domain (TAD); results are shown for the 20 loci for which the colocalisation was found to be significant regarding the H4 score (PP.H4.abf≥0.5) for a designated gene (referred to as coloc gene) are shown. Both the H3 score (where both traits possess a genetic association in the region, albeit with distinct causal variants) as well as the H4 score (where both traits are genetically associated in the region while sharing a common causal variant) are shown for the locus.*

**The H4≥0.5 threshold was selected arbitrarily, based on colocalisation analyses conducted by the Watkins group.*

Lead rsid	Coloc gene	Nsnps	PP.H3.abf	PP.H4.abf
rs8006225	*GSC*	3695	0.0	1.0
rs2540277	*TMEM182*	928	0.0	0.998
rs116904997	*PXN*	181	0.001	0.996
rs11687178	*E2F6*	1154	0.015	0.985
rs17817857	*HLF*	1904	0.009	0.981
rs11730129	*PROM1*	1981	0.021	0.979
rs58747679	*SSPN*	2810	0.008	0.976

rs8033459	*ALPK3*	1003	0.018	0.954
rs1009358	*CEP68*	754	0.048	0.952
rs4984741	*LMF1*	4678	0.106	0.893
rs6699769	*DNAJB4*	1942	0.09	0.85
rs2191446	*PLK2*	2922	0.063	0.807
rs12460541	*SYMPK*	1237	0.057	0.803
rs2503715	*SKI*	702	0.023	0.75
rs295114	*SPATS2L*	1634	0.305	0.694
rs3845778	*STRN*	3321	0.355	0.645
rs3176326	*CDKN1A*	915	0.035	0.565
rs497999	*MAPRE2*	798	0.172	0.546
rs11748963	*SPATA24*	205	0.466	0.534
rs11571171	*PGR*	2812	0.298	0.508

There was a significant colocalisation for the HCM positive control gene *ALPK3* and a handful of other genes are also plausible regarding a possible association with HCM. The *E2F6* repressor safeguards cardiomyocytes from apoptosis and boosts cell survival (Major *et al.*, 2017); however, it can cause abnormal expression in the myocardium and consequently DCM when dysregulated (Westendorp *et al.*, 2012). *SSPN* modulates the expression of complexes at the sarcolemma and provides protection against injuries such as cardiomyopathy (Parvatiyar *et al.*, 2015); this has been confirmed by a heightened susceptibility to arrhythmias in *SSPN* knockout hearts (Hwang *et al.*, 2023). *CDKN1A* was also recently established to reciprocally regulate cardiomyocyte death in a multi-omics analysis and is therefore a plausible target for heart failure drug development (Bian *et al.*, 2024).

PGR, which functions as a progesterone regulator, and progesterone plays a vital role in sustaining normal female reproductive health (Graham and Clarke, 2002) demonstrated a significant colocalisation as well. Research indicates that females with HCM exhibit less hypertrophy compared to males (O'Mahony and Elliott, 2017). The lead SNP for the locus in which *PGR* resides is rs11571171 with effect allele A and non-effect allele G. The presence of the A allele was associated

with an increased risk of HCM and a reduction in *PGR* expression. This observation supports the hypothesis that females with elevated *PGR* may experience a lower risk of HCM, potentially accounting for the observed decrease in hypertrophy. *SPATA24*, the Spermatogenesis Associated 24 protein, is primarily involved in sperm formation in male germ cells. The corresponding lead SNP is rs11748963 (effect allele T and non-effect allele C). The T allele was associated with an increased risk of HCM and a corresponding decrease in *SPATA24* expression. This may support the hypothesis that men with diminished *SPATA24* levels are more prone to HCM risk and thus increased levels of hypertrophy.

3.5 Assessing the Burden of Rare Variants in Prioritised Genes

3.5.1 LoF burden analysis

Genomics England (GeL) platform

Table 16 | GeL LoF burden result | A table showing rare loss-of-function (LoF) burden analysis results for the genes with a significant burden (considering multiple testing given an input list of 168 gene ids – resulting in a Bonferroni-corrected threshold of 2.98E-04). The pipeline was conducted within the Genomics England Research Environment and LOFTEE was implemented so as to ascertain high confidence pathogenic variants.

Gene symbol	Case count	Control count	P val	OR	Lower CI	Upper CI
MYBPC3	17	13	6.55E-19	44.062	22.540	105.6
ALPK3	16	41	4.92E-12	13.159	7.604	25.6

Both *MYBPC3* and *ALPK3* showed a significant LoF burden after multiple testing. This is unsurprising given that *MYBPC3* was one of the first core HCM genes to be identified as a result of causal LoF variants (Watkins *et al.*, 1995) and *ALPK3* LoF variants have more recently been associated with HCM (Herkert *et al.*, 2020). Although insignificant, SVIL (relevant for table 18) achieved a p-value 0.01207 and OR 7.6 (1.4 to 27.1).

BioResource (BRRD) platform

Table 17 | BRRD LoF burden result | A table showing rare loss-of-function (LoF) burden analysis results for the relevant genes shown in tables 16 and 18 albeit lacking a significant burden (considering multiple testing given an input list of 168 gene ids – resulting in a Bonferroni-corrected threshold of 2.98E-04) given data in BRRD. The pipeline was conducted within the BioResource cluster and LOFTEE was implemented so as to ascertain high confidence pathogenic variants.

Gene symbol	Case count	Control count	P val	OR	Lower CI	Upper CI
ALPK3	3	6	0.00165	17.4	2.8	81.8
SVIL	2	2	0.00456	34.8	2.5	476
MYBPC3	0	2	1	0	0	185.9

There was no significant rare LoF variant burden for any of the 168 input gene ids in BRRD; however, results are shown for *ALPK3*, *SVIL*, and *MYBPC3* in table 17 to provide context for the meta-analysis result in table 18.

Meta-analysis

Table 18 | Meta-analysis of GeL and BRRD LoF burden result | A table showing rare loss-of-function (LoF) burden meta-analysis results for the genes with a significant burden (considering multiple testing given an input list of 168 gene ids – resulting in a Bonferroni-corrected threshold of 2.98E-04); both the burden results from GeL and BRRD were combined in this meta-analysis.

Gene symbol	P val	OR	Lower CI	Upper CI
ALPK3	1.28E-21	13.722	8.020	23.477
SVIL	6.23E-06	11.359	3.957	32.605

Both *ALPK3* and *SVIL* showed a significant LoF burden in the meta-analysis result (table 18). *SVIL*, albeit a non-sarcomeric gene, was recently shown to be a novel HCM gene in light of its rare LoF variants (Tadros *et al.*, 2025).

3.5.2 Missense burden analysis

GeL platform

Table 19 | GeL missense burden result | A table showing rare missense burden analysis results for the genes with a significant burden (considering multiple testing given an input list of 168 gene ids – resulting in a Bonferroni-corrected threshold of 2.98E-04). The pipeline was conducted within the Genomics England Research Environment and REVEL was implemented so as to ascertain high confidence pathogenic variants.

Gene symbol	Case count	Control count	P val	OR	Lower CI	Upper CI
RAF1	9	28	8.56E-07	10.838	4.482	23.721
MYBPC3	28	136	6.97E-14	6.943	4.419	10.569
ALPK3	6	10	3.65E-06	20.229	6.025	61.527

Table 19 shows that *RAF1*, *MYBPC3*, and *ALPK3* all had a significant missense burden in GeL. About 50% of *MYBPC3* mutations result in truncated protein products, whereas the other variants are those pertaining to insertion/deletion, frameshift, or missense mechanisms (Tudurachi *et al.*, 2023). *ALPK3* is well-known regarding its association with HCM as a result of LoF variants; however, an association regarding missense variants is not well-established or recognised. Herkert *et al.* (2020) did however identify missense variants (seven compound heterozygous and one homozygous deleterious) underlying pediatric-onset cardiomyopathy evidenced by DCM progressing to hypertrophy, though without statistical evidence of an excess burden. The associated clinical features resembled those seen for other deleterious *ALPK3* variants. The authors hypothesised that the missense variants would result in a conformational alteration affecting protein folding/flexibility, protein-protein/DNA interactions, or the functionality of the alpha-kinase domain.

BRRD platform

Table 20 | BRRD missense burden result | A table showing rare missense burden analysis results for the relevant genes shown in tables 19 and 21 albeit lacking a significant burden (considering multiple testing given an input list of 168 gene ids – resulting in a Bonferroni-corrected threshold of 2.98E-04). The pipeline was conducted within the BioResource cluster and REVEL was implemented so as to ascertain high confidence pathogenic variants.

Gene symbol	Case count	Control count	P val	OR	Lower CI	Upper CI
ALPK3	1	2	0.0817	17.392	0.294	334.167
RAF1	0	14	1	0	0	10.565

As was seen in the LoF burden analysis, there was no significant rare missense variant burden seen for any of the 168 gene ids in BRRD; however, results are shown for *ALPK3* and *RAF1* in table 20 to provide context for the meta-analysis result in table 21.

Meta-analysis

Table 21 | Meta-analysis of GeL and BRRD missense burden result | A table showing rare missense burden meta-analysis results for the genes with a significant burden (considering multiple testing given an input list of 168 gene ids – resulting in a Bonferroni-corrected threshold of 2.98E-04); both the burden results from GeL and BRRD were combined in this meta-analysis.

Gene symbol	P val	OR	Lower CI	Upper CI
MYBPC3	4.13E-20	6.945	4.593	10.502
RAF1	6.06E-10	10.837	5.096	23.048
ALPK3	6.34E-09	20.227	7.328	55.828

MYBPC3, *RAF1*, and *ALPK3* showed a significant missense burden in the meta-analysis result (table 21). In summary, this was expected for *MYBPC3*, somewhat novel for *ALPK3*, and a finding to be interpreted with caution for *RAF1* due to the link between RASopathies and HCM.

RASopathies are known to cause HCM – accounting for ~12.5% of pediatric HCM cases (Aljeaid *et al.*, 2019; Calcagni *et al.*, 2018; Zamorano *et al.*, 2014) and ~33% of cases in infants (Norrish *et al.*, 2021). Such disorders are characterised by abnormal intracellular signaling pathways involving *RAS* and heightened *MAPK* activity (Calcagni *et al.*, 2018). In comparison to nonsyndromic, isolated primary HCM, RAS-HCM individuals exhibit an earlier onset and more severe clinical course (Wilkinson *et al.*, 2012). Noonan syndrome (NS) is a type of RASopathy marked by unique facial characteristics, reduced stature, congenital heart anomalies, and varying levels of developmental delays (Gelb *et al.*, 2015). HCM occurs in 10-29% of NS patients and usually presents at a young age (Gelb *et al.*, 2015; Hickey *et al.*, 2011; Wilkinson *et al.*, 2012). *RAF1* mutations are implicated in 5-15% of NS cases (Pandit *et al.*, 2007; Razzaque *et al.*, 2007). For ~95% of NS individuals, *RAF1* is the predominant gene linked to RAS-HCM (Kobayashi *et al.*, 2010; Pandit *et al.*, 2007; Razzaque *et al.*, 2007).

Plotting missense variants

A PDB structure has not yet been made available for *ALPK3* in RCSB PDB. However, the Alpha Fold Protein Structure Database has made a structure available (for *Homo sapiens*). The predicted structure was utilised for plotting variants (identifier: AF-Q96L96-F1 and positions 1-1907). The structure's average predicted local distance difference test (pLDDT) scored 45.37 which is classed as 'very low'; the average pLDDT is a measure of confidence in the prediction at the residue level whereby higher scores (nearing 100) reflect a greater accuracy in the established structure and improved reliability.

Because HCM is often concomitant to Noonan's Syndrome, I also decided to investigate these cases to check which other conditions they had been diagnosed with. Although the GeL Labkey database version 15 was used in prior analyses (for collecting the case/control sets), the case individuals were interrogated in the most recent version 19 database. PDB structures have been released for *RAF1*; however, not for the entire structure. The Alpha Fold Database has predicted its whole structure (for *Homo sapiens*) and so this was used for plotting variants (identifier: AF-Q96L96-F1 and positions 1-648). The structure's average PLDDT scored 67.94 ('low').

Missense variants plotted in both *ALPK3* and *RAF1* are provided at GitHub (https://github.com/domwest/DPhil-thesis/blob/main/ALPK3_RAF1_missense_variants_for_PyMOL_visualisations.xlsx).

ALPK3

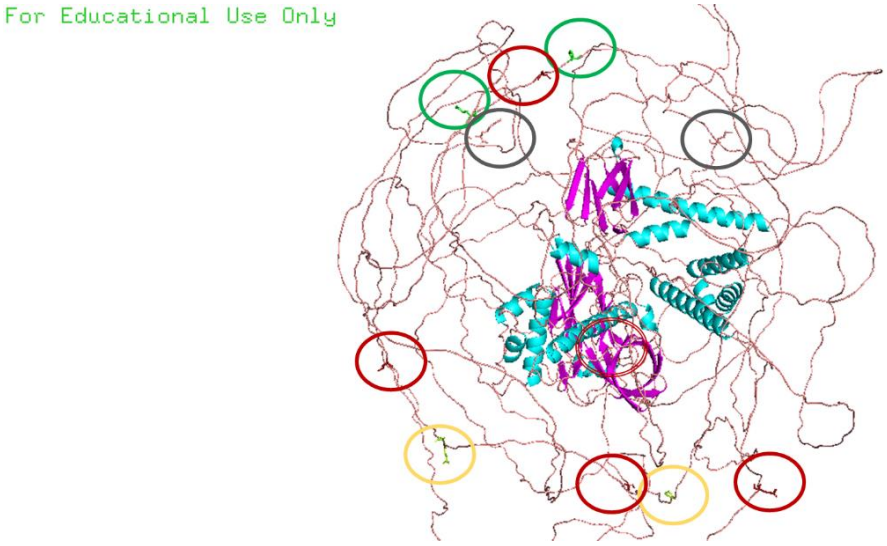


Figure 25 | ALPK3 AlphaFold protein structure with HCM missense variants | ALPK3 Alpha fold structure with GeL case and control missense variants plotted within. The N and C termini are positioned at the top and with the former on the left- and latter on the right-hand side.
Dark grey represents the N and C terminus residues; yellow for case and control shared variants; green for cases; red for controls.
The double-lined red circle highlights a control variant which sits behind the structure.

Figure 25 showcases the difficulty in detecting clusters or patterns for case versus control variants in *ALPK3*. Firstly, all of the variants reside in regions where the secondary structure has yet to be predicted (low confidence prediction regions). Secondly, there is no clear segregation seen for cases versus controls regarding their localisation even within these unresolved areas. However, both case variants are localised to a particular surface of the protein (seen to be near the N-terminus) although there is also a control residing in this region.

RAF1

Due to the link between NS and HCM, the 'panels applied' GeL Labkey table was consulted to see which secondary conditions were present for the 9 cases. 3 of the 9 cases had been recruited for RASopathies and 6 of the 9 for mitochondrial disorders. This confirmed that the majority of the

individuals were known to have syndromic features and hence likely to have RAS-HCM concomitant to NS but that some may not have been clinically diagnosed with NS. Plotting the variants in the structure ensued for possible detection of novel patterns (even if these would be attributed to RAS-HCM as opposed to P-HCM).

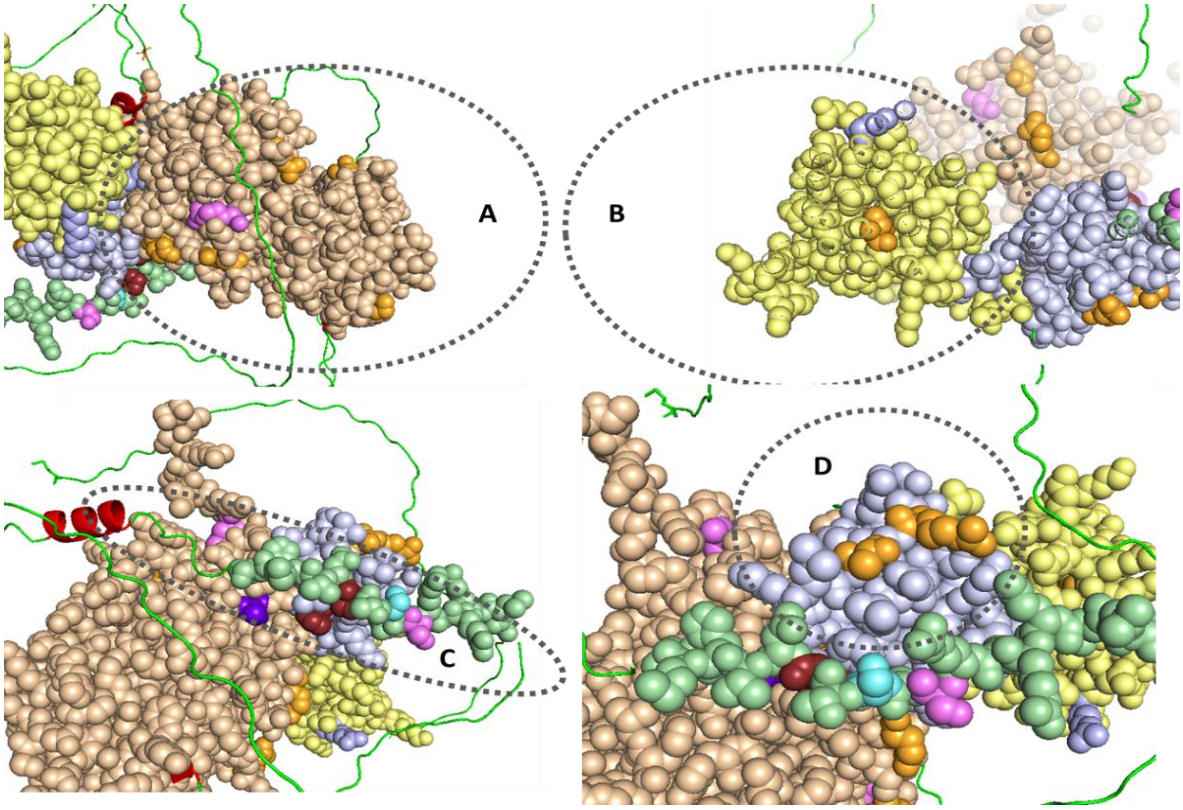


Figure 26 | RAF1 AlphaFold protein structure with HCM missense variants | RAF1 Alpha fold structure with GeL case and control missense variants plotted within. The CR3 domain (in A), RBD (in B), CR2, CRD, and CR3 domains' interacting surfaces (in C), and CRD domain (in D) are circled in dotted dark grey lines.

Regarding domains: the CR3 domain (aa 340-621; Daum et al., 1994) is shown in wheat; RBD (aa 51-139; Anton and Wennogle, 1998) in light yellow; CRD (aa 139-184; Anton and Wennogle, 1998) in light blue; CR2 (aa 250-265; Chan et al., 2002) in light green.

Regarding variants: light blue represents case and control shared variants; only case is violet pink; only controls is bright orange; phosphorylation hotspots are aquamarine; HCM and NS associated variants are ruby red; established DCM variants are purple blue

Figure 26A illustrates that the CR3 domain mainly comprises of control variants except for 1 HCM case variant which resides on a surface facing the RBD domain while the control variants are scattered at every other surface. Figure 26B shows the RBD domain and the 2 variants captured within whereby the control variant is internal to the domain and the variant seen in both a case and control localises on the outer surface facing the CR3 domain. Figure 26C highlights 2 DCM variants and an HCM case variant from this analysis residing where CR2, CR3, and CRD all interact. Figure 26D shows the CRD domain from a different angle (where it is facing away from the other domains in the structure) and how only control variants cluster on this surface.

RAF1 missense variants linked to NS and HCM have predominantly been shown to reside in the CR2 domain as well as the C terminus causing a heightened *RAF1* activity (Fullenkamp *et al.*, 2023)(Nakhaei-Rad *et al.*, 2023)(Zheng *et al.*, 2024). Fullenkamp *et al.* (2023) also identified the variants, F151L and T145M, residing in the CRD domain to cause HCM and NS. The PyMOL visualisations showcased how case variants appeared to cluster within domain binding surfaces as opposed to the outward-facing surfaces (as seen for the control variants). Although this finding is not specific to p-HCM, such investigations could help disclose some of the mechanisms in RAS-HCM manifestation. Additionally, most of the case variants identified in this analysis have not yet been reported in the literature. Likewise, this work clarifies that the majority of these cases should be evaluated by a clinician for a possible NS diagnosis.

3.6 Employing the Polygenic Priority Score tool for gene prioritisation

For each of the 68 HCM GWAS loci, the prioritised gene with the highest PoPs score within its locus was extracted. Table 22 shows the prioritised genes per 68 loci – loci are sorted by descending PoPs score.

Table 22 | PoPs gene prioritisation results | A table illustrating the PoPs results following a comprehensive analysis of 161 genes, which includes the most prioritised genes for each locus (from Open Target, FUMA, bulk quadrant approach, and SC quadrant approach). Only the top-scoring gene (with the highest PoPs score) is shown per 68 loci and the loci are sorted according to the highest assigned scores.

Lead rsid	Gene symbol	PoPs score
rs11121483	*RNF207*	1.200
rs6699769	*NEXN*	1.023
rs2644262	*FHOD3*	0.906
rs547048	*ACTN2*	0.892
rs6747402	*CCDC141*	0.796
rs6796333	*TNNC1*	0.793
rs8033459	*ALPK3*	0.778
rs62222424	*SON*	0.756
rs12210733	*PLN*	0.638
rs12216858	*GATA4*	0.567
rs12460541	*DMPK*	0.561
rs28768976	*MAPT*	0.485
rs28575249	*SRL*	0.482
rs1048302	*FBLIM1*	0.464
rs7612736	*SLC6A6*	0.454
rs75096272	*SLC1A3*	0.453
rs5760054	*CHCHD10*	0.448
rs41306688	*ADPRHL1*	0.414
rs79502300	*MITF*	0.408
rs497999	*MAPRE2*	0.408
rs6566955	*MALT1*	0.369
rs6914805	*ATXN1*	0.346
rs113907726	*GPR137C*	0.310
rs139920	*TNRC6B*	0.292

rs66520020	*FLNC*	0.285
rs4984741	*PKD1*	0.277
rs3845778	*NDUFAF7*	0.274
rs11196085	*TCF7L2*	0.253
rs1699340	*RAF1*	0.251
rs58747679	*RASSF8*	0.247
rs11860560	*ECI1*	0.243
rs2945236	*MFHAS1*	0.241
rs10841524	*PDE3A*	0.241
rs2191446	*PLK2*	0.239
rs17817857	*MMD*	0.238
rs7210446	*CEP112*	0.235
rs182427065	*SERPING1*	0.210
rs4894803	*PLD1*	0.209
rs116904997	*PXN*	0.205
rs2184370	*PCMT1*	0.195
rs17617337	*BAG3*	0.180
rs295114	*SPATS2L*	0.178
rs13297385	*PRRC2B*	0.173
rs3176326	*SRSF3*	0.169
rs2540277	*TMEM182*	0.154
rs8006225	*DICER1*	0.146
rs11748963	*PAIP2*	0.142
rs11687178	*ROCK2*	0.140
rs1159974	*HEY2*	0.139
rs12270374	*RRAS2*	0.130
rs11730129	*TAPT1*	0.125
rs1009358	*RAB1A*	0.115
rs2002854	*ADAMTS7*	0.110
rs7487962	*TBX3*	0.109

rs2810883	*PKN2*	0.109
rs7824244	*XPO7*	0.097
rs2177843	*USP54*	0.096
rs7835298	*FBXO32*	0.092
rs8019721	*SIPA1L1*	0.090
rs9515201	*COL4A2*	0.089
rs2503715	*SKI*	0.048
rs2182400	*BAMBI*	0.023
rs140763915	*SPATA7*	0.007
rs816379	*TINAG*	-0.027
rs11571171	*ARHGAP42*	-0.031
rs35006907	*ZNF572*	-0.083
rs2645210	*KLF6*	-0.089

The authors of PoPs categorise a gene as being prioritised if it resides in a 1 Mb locus (centered on a GWAS variant) and displays the highest PoPs score among other genes in the locus. For this reason, there is no threshold with which to subset significant target genes.

Out of all the gene prioritised through PoPs, those marked as plausible candidates due to an association with cardiomyopathy were *SON*, *SLC6A6*, *DICER1*, and *USP54*. The *de novo* heterozygous frameshift *SON* mutation c.3852_3856delGGTAT, which is essential for heart development and mitochondrial function, may be linked to a heightened vulnerability to myocardial injury or cardiomyopathy (Na *et al.*, 2024). A deficiency in taurine (product of *SLC6A6*) results in DCM in both cats and foxes (Ito and Murakami, 2024) and using iPSC-derived cardiomyocytes, *SLC6A6* was identified to be the most probable gene associated with DCM at the 3p25.1 locus (which also contains the HCM GWAS lead variant rs7612736) (Garnier *et al.*, 2021). *DICER1* plays an essential role in cardiac contraction and the specific removal of Dicer in the cardiac tissue results in DCM and subsequent heart failure (Chen *et al.*, 2008a). Ubiquitin-proteasome systems (UPSs) are capable of modulating fibrosis, lipid metabolism, glucose homeostasis, hypertrophic responses, post-ischemic recovery, and cell death mechanisms such as apoptosis and ferroptosis in cardiomyocytes;

given that the loss of myocardial cells is a significant factor in cardiomyopathy, UPSs are being increasingly recognised for their roles in this condition (Li and Ma, 2025).

Multiple genes in the PoPs prioritisation output have also been associated with cardiac phenotypes such as heart failure, hypertrophy, cardiac stress and remodelling, impaired cardiac function and development, myocardial injury, altered contraction, and cardiac fibrosis. Despite these mechanisms often being concomitant features in HCM individuals, it does not imply that these genes are directly involved in eliciting HCM. For example, *ARHGAP42* and *PDE3A* have been associated with hypertension and although common cardiovascular risk factors (including hypertension) have been linked to a more severe phenotype of HCM (Lopes *et al.*, 2023), this does not imply that these genes are associated with HCM.

3.7 Discussion & Limitations

The core input to this project was the HCM GWAS summary data. It's important to preface that this can always be improved through the addition of more data to enable the discovery of more loci. Additionally, ancestry differences are not considered due to the data being obtained from European populations.

TADs are typically ~1mbp in size (Dixon *et al.*, 2012); despite having manually extended loci TAD coordinates after visual inspection as well as providing manual assignment for those that were simply unassigned to a TAD, the majority were still smaller than 1 Mbp. This poses the question as to whether significant genes may have been excluded across the loci? Most Hi-C datasets are low resolution (25 kbp to 1 Mbp); insight into the 3D architecture of genomes is limited because the scope is limited to a kbp/Mbp scale (Liu *et al.*, 2019b). Due to the low resolution of Hi-C data, the identification of TADs and sub-TADs may be affected (Rao *et al.*, 2014). High resolution datasets necessitate extensive sequencing (and the depth thereof) and thus incur substantial costs – they only exist for a handful of tissues or cell lines. Additionally, through the re-examination of TADs derived from gold-standard mammalian Hi-C datasets, Chang *et al.* (2020) coined TADs as context

dependent regions that are only moderately insulated from adjacent areas. These findings suggest the necessary level of caution when linking HCM variants in a locus to a set of plausible genes.

Despite results from the quadrant approaches being insightful, gene expression levels change with different stimuli and because this data is from normal participants (in absence of the disease environment) the exact expression profiles may not have been captured. Additionally, one cannot infer purely based on expression data that a specific gene will likely be involved in disease because it is highly expressed in a cell or tissue type. However, García-González *et al.* (2025) integrated GWAS summary statistics and RNA-Seq gene expression data and found that tissues and cell types with a well-documented association to disease generally showed higher absolute and relative expression levels of prioritised genes compared to control genes; Barshir *et al.* (2014) also suggested that although disease-related genes may exhibit widespread expression, they are characterised by elevated transcript levels and enhanced protein-protein interactions in tissues linked to disease when contrasted with unaffected tissues. On the contrary, Feiglin *et al.* (2017) suggested that while the association between highly expressed genes and disease is evident, certain genes may be expressed across numerous tissues yet only lead to disease in one or a limited number of them. Additionally, a caveat of the quadrant approach is rooted in the understanding that causal genes associated with polygenic HCM are dosage-sensitive genes whereby their altered expression causes disease. Similarly to how the altered expression may be minimal and therefore the extent of protein product, the abundance of a candidate gene may be low in normal circumstances – with only a small amount needed to perform its function. For this reason, focusing on genes displaying high expression levels may not be applicable in all scenarios.

Extreme fold changes witnessed through the quadrant approach analyses are owed to biological differences in tissue- or cell-specific gene expression patterns (such as the relative abundance of cell types between different tissues and the diverse regulatory mechanisms), as well as technical factors (such as library size and composition, read depth, and batch effects).

Pseudobulk approaches have also become favourable in scRNA-Seq analyses (particularly with regards to aggregating expression data across cells) and the approach taken during this thesis during the quadrant approach was similar. Pseudobulk aggregation entails summing or averaging expression counts across cell types within a condition or subject; it has been shown to reduce the complexity of single-cell data and improve statistical power (Crowell *et al.*, 2020)(Junttila *et al.*, 2022).

The colocalisation analyses on the whole might have been flawed due to several factors (Hormozdiari *et al.*, 2016). Linkage disequilibrium (LD) may have introduced ambiguity, tissue heterogeneity might have obscured the eQTLs pertinent to cardiomyocytes (Heinig, 2018), the lack of relevant cell/tissue type (heart and cardiomyocyte-specific) data and sample sizes thereof to reach adequate power is limiting, a gene might only act as an eQTL under certain conditions or developmental stages, GWAS variants might have associated with eQTLs by chance (Heinig, 2018), variants may affect splicing or non-transcriptional regulation as opposed to expression, and QTL studies are often constrained to cis-eQTLs (trans-eQTLs mostly being available for blood samples only (Liu *et al.*, 2019a)). eQTLs also explain only a small fraction of the GWAS signal and have been seen to cluster near transcription start sites while GWAS hits do not; genes near GWAS hits are enriched in functional annotations whereas those near eQTLs are depleted of these (Mostafavi *et al.*, 2023). In spite of these issues, a highly significant enrichment of trait associated SNPs can be observed among eQTLs as demonstrated for heart related traits (Heinig, 2018) and these colocalisations may have uncovered even a few associations of note.

ADPRHL1, *SYNPO2L*, and *SRL* were prioritised in the quadrant approach analysis. *ADPRHL1* is of significance considering that missense variants in this gene have been associated with cardiac phenotypes in addition to the fact that one missense variant which causes conduction abnormalities also happens to be the lead HCM GWAS variant in this locus. Ultimately, although the coding variant model for *ADPRHL1* and its link to HCM is plausible, the expression level may indeed affect susceptibility to HCM. FinnGen variants intersected with 134 loci revealed from a meta-analysis of GWAS on AF (Roselli *et al.*, 2018) resulted in rs766868752 predicted to be a loss-of-function variant within *SYNPO2L*. Clausen *et al.* (2021) found rs60212594 to align with a reduced expression of *SYNPO2L*. The authors also found *SYNPO2L* to interact and be co-expressed with multiple genes

(*TTN*, *MYOM2*, *ANKRD1*, *CMYA5*, *MYOZ1*, *CASQ2*, *TRDN*, *LDB3*, *SMPX*, and *C9orf3*) associated with sarcomere structure, calcium signalling, and cardiomyopathies. The HCM GWAS lead variant associated with *SYNPO2L* also falls within this locus and is in high LD with rs60212594 (R2 of 1); however, the HCM GWAS lead variant's eQTL shows the opposite direction of effect on *SYNPO2L*. *SRL* could plausibly cause sex differences in the HCM phenotype manifestation. The X-chromosome microRNA (miRNA) locus, predominantly expressed in cardiac tissue and enriched in females, is important for heart development and function. *MiR-871* introduction mimics the effects of *SRL* loss; the *SRL* protein is twice as abundant in male hearts compared to female hearts. Emerson *et al.* (2025) investigated the average QTc interval and calcium reuptake/repolarisation times and concluded that there is an independent mechanism managing the reduction of *SRL* more effectively in females as opposed to males. The HCM GWAS lead variant in the *SRL* locus eQTL leads to a reduced expression of *SRL* which may prove problematic in males given Emerson *et al.* (2025) analysis.

ALPK3 (alpha-kinase 3), *PLK2* (polo-like kinase 2), and *TMEM182* (cell membrane transmembrane protein 182) were prioritised in the quadrant approach, colocalisation, as well as PoPs analyses. *ALPK3* was prioritised further due to its rare LoF as well as missense burden; *ALPK3* has been associated with HCM through a loss of function mechanism; however, its missense mechanism should be explored further. A downregulation of *PLK2* has been associated with AF and fibrosis (Künzel *et al.*, 2021); and *TMEM182* serves as a negative regulator of myogenic differentiation and muscle regeneration (Luo *et al.*, 2021).

4

Variant Prioritisation

4.1 Introduction

After attempting to annotate candidate genes across each of the 68 HCM GWAS loci, the approaches detailed in this chapter aimed to tackle variant prioritisation to identify functional variants. Post-GWAS, variants annotated as lead variants (the most significant associations with the trait of interest) are often expected to be the most likely causal variants. However, sometimes this is not the case and they may actually reflect the effect of a nearby causal variant in LD. The methods employed in this chapter leveraged chromatin data and various computational tools to highlight plausibly associated variants based on a predicted effect on gene expression.

4.1.1 The regulatory landscape

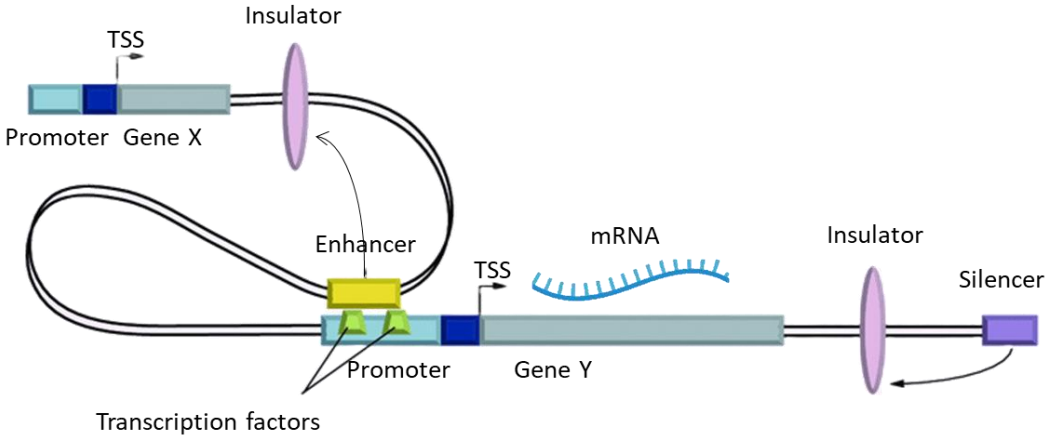


Figure 27 | Regulatory elements | A figure depicting the different types of regulatory elements as illustrated by Luizon and Ahituv (2015). The role of enhancers, insulators, and silencers is illustrated with regards to gene expression through their effect on promoters and various transcription factors.

Promoters

Promoters are short sequences situated close to transcription start sites (TSSs) and are 'active' when they correspond to actively transcribed genes; active promoters are found in open chromatin with the distinguishing tri-methylation of Histone H3 at Lysine 4 (H3K4me3) (Shilatifard, 2012). Transcription factor (TF) binding patterns at promoters exhibit significant predictive power for functionality; numerous TFs can bind to a particular promoter and enable transcription (Chandra *et al.*, 2023).

According to Li *et al.* (2021b), most developmental genes are controlled by several core promoters that produce analogous transcripts. Consequently, multiple promoters may exist for a particular gene due to the preference of enhancers for a specific promoter or otherwise due to a redundancy mechanism.

Enhancers

Enhancers contain cell type-specific TF binding sites (Panigrahi and O'Malley, 2021) and unlike promoters, can be situated further away from their target gene(s) and can work in either orientation. As a result, genomic position cannot be used as a standalone method to determine an enhancer's target gene. The described regulation over broad distances is mirrored in the conservation of synteny (Chua *et al.*, 2022). Enhancers can interact with more than one gene within a defined space (Karnuta and Scacheri, 2018). Active enhancers show mono-methylation of Histone H3 at Lysine 4 (H3K4me1) and acetylation of the lysine residue at N-terminal position 27 of the histone H3 (H3K27ac) (Calo and Wysocka, 2013).

Typically, an inactive enhancer is marked by H3K4me1 while an active enhancer by both H3K4me1 and H3K27ac (Kang *et al.*, 2021). H3K4me1 has been seen to precede H3K27ac at primed (pre-active) enhancers. In this way, the presence/absence of H3K27ac can differentiate between active enhancers and those that are either inactive or poised. However, it has been suggested that distal

H3K27ac regions lacking substantial H3K4me1 enrichment should also be regarded as potential enhancer elements (Creyghton *et al.*, 2010).

Silencers

Silencers diminish transcriptional activity and are therefore the repressive equivalent of enhancers; however, they exhibit bifunctionality with the ability to function as enhancers too depending on the cellular environment (Segert *et al.*, 2021). Silencers have received less attention compared to enhancers as explained by Segert *et al.* (2021) and consequently, there is no established "silencer chromatin signature" characterised by a uniform distribution of histone modifications or related proteins applicable to all silencers. The absence of histone marks exclusively associated with silencers might suggest that they are marked by a combinatorial code of chromatin marks that are also present in other elements instead.

Insulators

While cis-regulatory elements affect gene transcription across varying distances, insulators restrict the influence of enhancers and silencers. They do so by inhibiting communication between enhancers and promoters or obstructing the extension of repressive chromatin (Brasset and Vaury, 2005). CCCTC-binding factor (CTCF) is the established protein associated with the formation of insulators in vertebrate organisms (Felsenfeld *et al.*, 2004). *CTCF* encodes a 82-kDa protein with 11 zinc fingers. It can arrange chromosomal architecture by binding to a wide collection of sequences and controlling gene expression via the activation or repression of promoters, insulation of enhancers, and regulation of distant chromatin interactions (Kim *et al.*, 2015).

4.1.2 Chromatin data and its uses

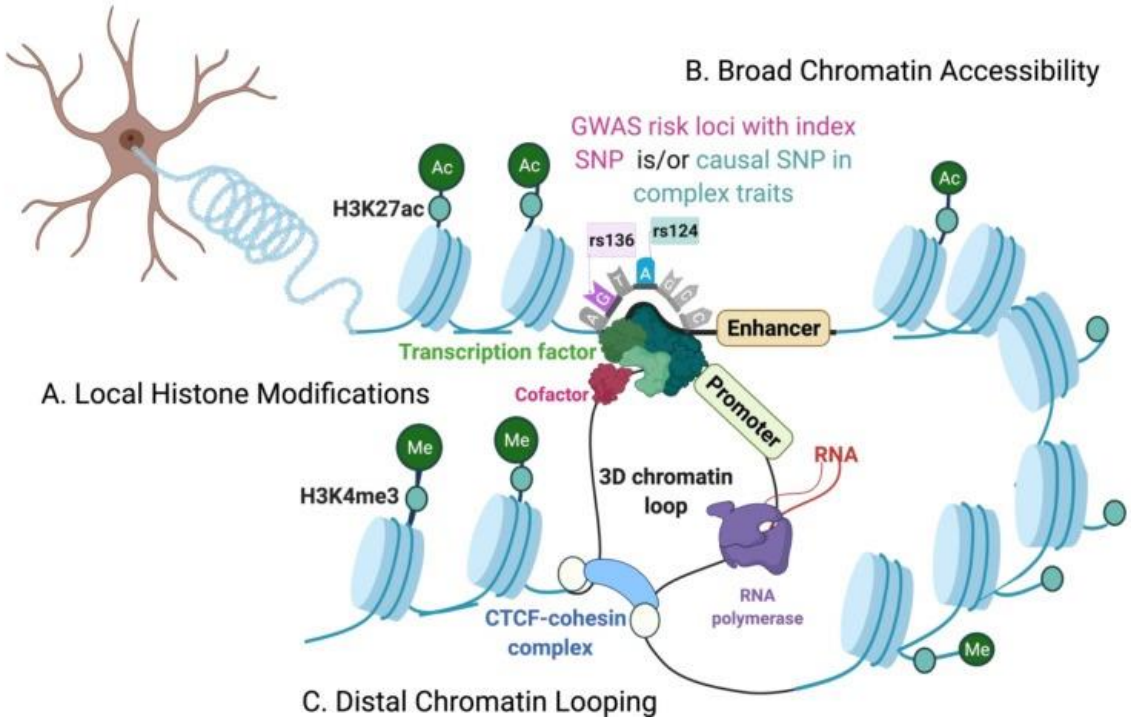


Figure 28 | The Regulatory Landscape | A figure illustrating the regulatory landscape and applicable techniques for uncovering its workings as illustrated by Chawla et al. (2021) which shows the local histone modifications (A) such as methylation and acetylation, broad chromatin accessibility (B) giving insight into whether proteins can bind the DNA at particular regions (transcriptions factors, cofactors, and enzymes, for example), and distal chromatin looping (C) showcasing distant elements, such as enhancers, being brought into close contact with genes and their promoters.

Open chromatin comprises of both noncoding and coding regions. There is interplay between regulatory elements and their target genes in both local and remote manners and this is conveyed experimentally through chromatin profiling. Both cis-regulatory elements and the nucleosome distribution are inspected through ATAC-Seq, DNase-Seq, MNase-seq or FAIRE-seq, while histone modifications and TF binding through ChIP-seq (figure 28B), Cleavage Under Targets and Release Using Nuclease (CUT&RUN), or Cleavage Under Targets and Tagmentation (CUT&TAG) (figure 28A). 3D chromatin conformation capturing long-range interactions is examined by 3C, 4C, 5C, or HI-C mapping (figure 28C) (Chawla *et al.*, 2021). Chromosome conformation capture techniques assess the closeness of DNA elements within the three-dimensional environment of the nucleus.

Such methods involve chromatin digestion and re-ligation whereby ligation junctions are subsequently analysed (Downes *et al.*, 2022).

Efforts to clarify the 3D genome structure via an array of technologies has resulted in genome-wide maps of cell-type-specific chromatin interactions and loops. Such datasets have been generated for many different cell types and conditions. Overall, they have highlighted the significance of cell-type specific physical proximity between regulatory elements and the genes they target. They have also revealed the impact even minor changes in proximity can have on gene expression. Many studies have identified enhancer regions containing causal variants – which regulate their target genes via looping (despite their linear remoteness). Thus, understanding the regulatory workings of the genome can help elucidate which mechanisms are disrupted (or even amplified) by identifying the elements in which such variants reside (Bhattacharyya and Ay, 2024).

4.1.3 Chromatin data cell type / tissue selection

The identification of disease-relevant genomic regions is frequently achieved through GWAS analyses. Knowing the cell type and possibly even state in which the disease manifests is key to understanding the relevant interactions involved in the disease trajectory (Jagadeesh *et al.*, 2022).

Cardiomyocytes are not only physically linked in the heart's myocardium, but they also communicate through a combination of direct contact and paracrine signalling. During heart-related disease, cells of the heart portray maladaptive responses which together result in a decline in function. Fibroblasts disease responses lead to fibrosis which in turn plays a significant role in heart failure, and certain interactions between fibroblasts and cardiomyocytes elicit a response contributing to arrhythmogenesis (Talman and Kivelä, 2018).

Executing a cell type enrichment pipeline prior to DeepHaem also helps to elucidate the most likely cell type(s) and to dictate the ratios thereof. Knowing the cell type(s) involved also makes testing a possible causal variant and/or its gene more biologically tractable. To prioritise sarcomere-negative HCM regulatory variants during prioritisation approaches, the inclusion of numerous heart cell type

datasets as input avoids the assumption that this form of HCM is solely associated with cardiomyocytes.

4.1.4 Chromatin Data QC and Standardisation Via CATCH-UP

Although there has been a rise in parallel sequencing methods, the comparison of chromatin datasets is difficult due to the differences in settings applied during generation, the cells in which each experiment was performed, and the antibodies used (Mendoza-Parra *et al.*, 2013).

There are many components that may affect the quality of next generation sequencing (NGS) profiles (Mendoza-Parra *et al.*, 2013) which is especially true when working with immunoprecipitation-based methods. Additionally, NGS platforms can have varying capacities (reads per flow cell) as well as differing sequencing depth; consequently, the number of binding sites being uncovered varies. There is also a lack of gold standard archives and the nature of functional validation is laborious (Han *et al.*, 2022).

Quality assessment of chromatin data typically includes visual assessment followed by the execution of bioinformatic tools which perform peak-calling among other tasks. Statistical methods have been devised to assess the signal-to-noise ratio by determining the proportion of reads located within peaks or tag-enriched regions. To allow for the accurate comparison of findings across experiments, the standardisation of these approaches is essential. CATCH-UP (Riva *et al.*, 2023) was employed in this thesis due to its reproducibility across research groups.

4.1.5 Using Machine Learning to Prioritise Functional Variants

Machine learning (deep learning) models are capable of learning from one's data to make predictions based on non-linear patterns detected. Such models can detect patterns in open chromatin data and predict whether changes to the nucleotide sequence open up (increase the accessibility thereof) or close (inhibit) the chromatin. Genetic discovery combined with chromatin data was used as input to machine learning models for the functional variant prioritisation performed in this chapter.

Introducing DeepHaem and the likes

One of the initial deep learning CNNs (called Basset, devised in 2016 (Lan and Corces, 2022)) was developed on the elusive nature of eukaryotic gene expression as well as the mechanisms behind noncoding variants associated with disease. The model utilises CNNs for learning the functional activity of DNA sequences and ultimately provides annotations for each mutation in the genome regarding its effect on accessibility.

Riding on these insights, the next approaches were rooted in the realisation that a model able to predict the effects of noncoding variants given DNA sequence alone could improve the prediction (Zhou and Troyanskaya, 2015). Reasons being that motifs hold less power than evolutionary features and chromatin annotation in variant prediction; *in vivo* TF binding is dependent on the extended sequence (not only at motif sites); DNase I–hypersensitive sites and histone marks display complex mechanisms involving multiple chromatin proteins. In 2015, the DeepSEA (deep learning–based sequence analyser) model, a multitask, multilayer, hierarchically structured model, was created with this approach in mind for TF binding prediction. DeepSEA receives ENCODE and Roadmap Epigenomics *de novo* sequence data (DNase sensitivity, transcription factor binding, and histone mark data specifically) as its input and then predicts the regulatory landscape. DeepSEA sidesteps the incorporation of annotation and instead learns directly from sequence data. This model demonstrates three overarching attributes: sequences used as input capture broad ranges, varying lengths of chromatin is learnt at several junctures, and multitask joint learning allows for patterns across all tracks to be predicted concurrently given one input sequence. DeepSEA performance surpassed that of previous methods in prioritising Human Gene Mutation Database (HGMD) regulatory variants, eQTLs, and GWAS variants and it could also distinguish between non-trait and trait-associated variants.

DeepSEA and Basset share similar features; however, DeepSEA reads in 1000 base pair sequences as opposed to 600bp and it leverages cell type-specific data (from 164 different cell types) (Lan and Corces, 2022). Additionally, DeepSEA combines *de novo* sequence chromatin data with conservation information (unlike Basset) contributing to an evolutionary-derived score. An overall

functional significance score is generated through the product of the geometric mean value for predicted chromatin effects and the geometric mean value for evolutionary conservation features (Zhou and Troyanskaya, 2015).

DeepHaem is a flexible TensorFlow application of a CNN model with max pooling. Its underlying architecture was built on inspiration from both Basset and DeepSEA. It consists of several layers of convolutional filters with subsequent Rectified Linear Unit (ReLU) non-linear activation functions (aimed at tackling the vanishing gradient problem whereby learning can be impeded during backpropagation) as well as max pooling and concluding with a fully connected layer. DeepHaem deployment can involve flexible assembly in light of the provided hyperparameters (regarding architecture stipulations and the learning process parameters). The model's overarching aim is to predict the chromatin feature class a given sequence belongs to and/or to predict the impact of a variant on these classes.

Deepmind advancements

A more recent advancement regarding uncovering the chromatin landscape and predicted impact of variants therein has been illustrated by Google Deepmind's Enformer model. Enformer identifies the main transcription start site (TSS) of a gene by calculating the highest predicted Cap Analysis of Gene Expression (CAGE) values (where CAGE is a technique used to identify and quantify transcription start sites (TSSs) and assess gene expression levels) exposing the DNA sequence centered at this site and ultimately devises a gradient x input, attention, and ISM score (Avsec *et al.*, 2021).

Enformer predicts gene expression and the regulatory landscape given both human and mouse genome data. The model shows an enhanced correlation between raw data and its predictions when compared to other models in the field; it can be distinguished by its self-attention mechanisms. As shown by CRISPRi enhancer assay evaluations, Enformer better assimilates long-range data and thus increases the receptive field. In a neural network, each neuron obtains input solely from a restricted region known as the receptive field. It can be inferred that a larger receptive field correlates

with an enhanced capacity to extract global features and mitigate interference. Additionally, Enformer has shown improved accuracy in predicting variant impact as benchmarked by direct mutagenesis assays and eQTL data (Avsec *et al.*, 2021).

Enformer utilises a transformer derived architecture. Major advancements have been seen for models that utilise transformers – illustrated in the recent transformation of natural language processing (summarised in figure 29). Transformers consist of multiple attention layers which modify each input position through weighted sums (representing other portions of the input sequence). For example, a prediction at a TSS can be fine-tuned through the consideration of surrounding sites like enhancers which affect the relevant gene. Unlike CNNs, transformers do not need many convolutional layers to reach distant elements. For example, two CNNs (Basenji2 and ExPecto) cannot predict elements further than 20kb away. Considering that many regulatory elements (such as enhancers, repressors, and insulators) can influence gene expression more than 20kb away, transformers tackle this limitation by providing the broader context (Avsec *et al.*, 2021).

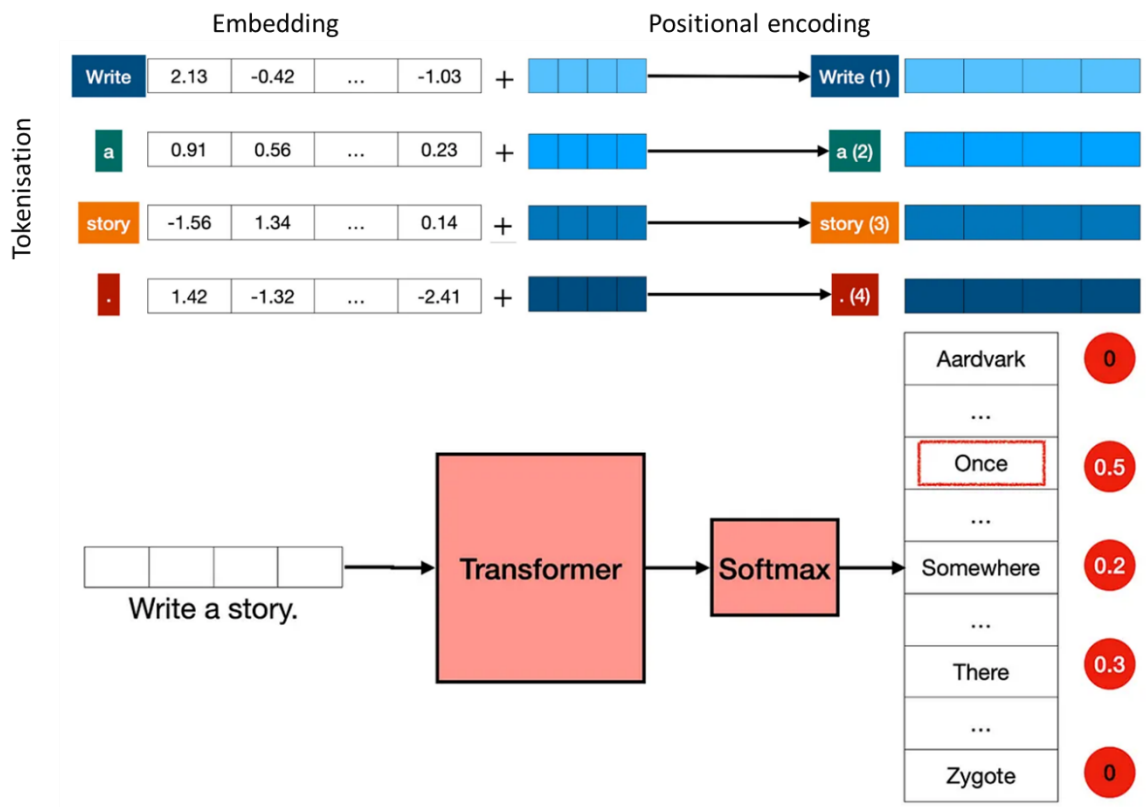


Figure 29 | Transformers for natural language processing | A figure adapted from Lee (2023) illustrating the use of transformer models in natural language processing (NLP). Initially

tokenisation occurs whereby raw text is converted into 'tokens' (a numerical representation of each word). Embedding follows whereby tokens are converted to vectors (the required format for input to the model). Positional encoding then ensues for each token to capture the order of words. Thereafter, the transformer (comprising of an attention and feedforward block) assigns the next word in the sequence; attention adds context and feedforward provides the prediction. Finally, a SoftMax function generates a probability distribution regarding next plausible words from which to sample.

Enformer outperformed Basenji2 in RNA expression prediction by an increase in correlation from 0.81 to 0.85. This improved performance better equips Enformer to detect enhancer-promoter interactions as well as the effect of noncoding variants. In addition to enhancers and promoters, the role of insulators is also learnt during prediction. Overall, Enformer can detect both activating and repressing variants, it does not depend on conservation data (an added bonus considering conserved enhancers make up a very small fraction), and it can also produce predictions for random given sequences (Avsec *et al.*, 2021).

4.1.6 Key objectives

The key objectives for this Chapter are to:

1. Perform cell type enrichment using the HCM GWAS summary data to instruct the necessary heart cell type datasets to be included in the variant prioritisation analysis.
2. Apply a quality control and standardisation pipeline to all chromatin datasets to be utilised in the variant prioritisation analysis to ensure the data is of high quality, minimise noise, artifacts, as well as technical and experimental biases, and to add conformity across datasets.
3. Employ machine learning methods using chromatin data to learn the underlying chromatin landscape of various heart cell types and predict the impact of noncoding HCM GWAS variants.

4.1.7 Computational approaches employed

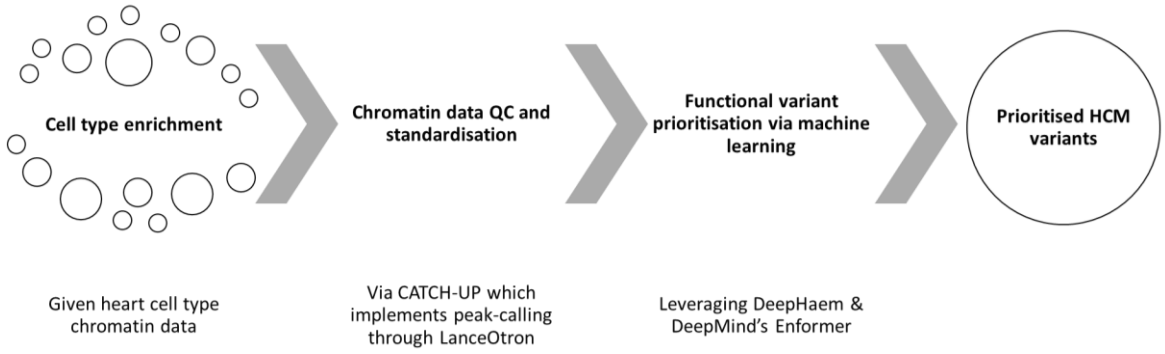


Figure 30 | Variant prioritisation approach | A flowchart illustrating the computational tools implemented in this chapter to obtain a final set of prioritised HCM variants. Cell type enrichment was first performed to confirm the likely cell type(s) of interest for inclusion in machine learning approaches; quality control (QC) and standardisation of the heart chromatin data was performed before input to machine learning; finally, two different machine learning methods were leveraged for functional variant prioritisation – the convolutional neural network, DeepHaem, followed by a transformer approach via Deepmind's Enformer.

Figure 30 summarises the computational approaches employed in this chapter: cell type enrichment, chromatin data quality control and standardisation, and ultimately the application of machine learning models to score and prioritise a set of HCM GWAS lead variants and their proxy variants.

4.2 Cell type enrichment

GWAS variants (of which, the majority are noncoding) tend to cluster in regulatory elements and using chromatin data from different cell types, one can tell where these variants cluster most and in which cell type (as shown in table 23).

Table 23 | Cell type enrichment results | A table depicting the output enrichment scores as generated by the cell type enrichment pipeline provided by Riva et al. The HCM GWAS lead and proxy variants were considered whilst generating enrichment scores and the enrichments were compared to 1). background sequence and 2). 1000G sets of variants. Each heart cell type/tissue has an associated mean enrichment score for the HCM GWAS variants (highest scores shown first).

Cell type	**Mean enrichment score**
Ventricular Cardiomyocyte 1 (ATAC)	2.77
Ventricular Cardiomyocyte 2 (ATAC)	2.58
Atrial Cardiomyocyte 1 (ATAC)	1.80
Atrial Cardiomyocyte 2 (ATAC)	1.61
Smooth Muscle (ATAC)	1.06
Pericyte General 2 (ATAC)	1.04
Adipocyte (ATAC)	0.97
Pericyte Muscularis (ATAC)	0.94
Cardiac Pericyte 2 (ATAC)	0.94
Fetal Endocardial (ATAC)	0.91
Pericyte General 4 (ATAC)	0.81
Cardiac Pericyte 4 (ATAC)	0.80
Cardiac Pericyte 1 (ATAC)	0.79
Fetal Ventricular Cardiomyocyte (ATAC)	0.79
Endocardial (ATAC)	0.78
Pericyte General 3 (ATAC)	0.74
Bulk Heart LV Tissue 2 (DNase)	0.73
Erythroid Donor 001 (ATAC)	0.64
Fibroblast (ATAC)	0.64
Cardiac Muscle Cell 2 (DNase)	0.61
Bulk Heart LV Tissue 2 (ATAC)	0.61
Cardiac Fibroblast (ATAC)	0.60
Erythroid Donor 002 (ATAC)	0.59
Lymphocyte (ATAC)	0.56

Fetal Atrial Cardiomyocyte (ATAC)	0.56
Pericyte General 1 (ATAC)	0.56
Vascular Smooth Muscle 1 (ATAC)	0.55
Fibroblast (ATAC)	0.54
Endothelial (ATAC)	0.54
Erythroid Donor 003 (ATAC)	0.53
Endothelial Myocardial (ATAC)	0.53
Macrophage (ATAC)	0.52
Vascular Smooth Muscle 2 (ATAC)	0.51
Pericyte (ATAC)	0.51
Endothelial (ATAC)	0.50
Mast 1 (ATAC)	0.49
Mast 2 (ATAC)	0.48
Smooth Muscle Cell 2 (ATAC)	0.48
Mesothelial (ATAC)	0.48
Cardiac Muscle Cell 1 (DNase)	0.47
Cardiac Pericyte 3 (ATAC)	0.46
Fetal Mesothelial (ATAC)	0.46
Macrophage (ATAC)	0.46
Fetal Cardiac Fibroblast (ATAC)	0.46
Adipocyte (ATAC)	0.44
Smooth Muscle Cell 1 (ATAC)	0.44

The data aligns with what one would expect with atrial and ventricular cardiomyocytes showing the most enrichment compared to the other cell types. The mean values provided by this pipeline show values that are ~2 fold higher in the ventricular cardiomyocytes compared to that shown by atrial cardiomyocytes which suggests that ventricular cardiomyocytes are likely associated with HCM in the presence of common variants. Smooth muscle cells and pericytes were the next most significant cell types.

Cardiac pericytes have recently emerged with regard to their therapeutic potential in the heart and they are critical with regards to the heart's response to myocardial infarction, fibrosis, and thrombosis (Alex *et al.*, 2023). Smooth muscle cells are important for regulating blood pressure and distribution as well as for structural integrity (Zhuge *et al.*, 2020). It is not inconceivable for pericytes and smooth muscle cells to be involved in HCM; however, cardiomyocytes show more significant enrichment values.

4.3 Chromatin Data QC and Standardisation Via CATCH-UP

Having selected the heart tissue and cell type datasets for inclusion as input to machine learning (post-cell type enrichment), these datasets needed to undergo quality control and standardisation by an upstream pipeline called CATCH-UP (with LanceOtron integrated) before being fed to machine learning for functional variant prioritisation. This was essential for noise filtering as well as peak inspection – ensuring accurate downstream DeepHaem predictions.

The quality control and standardisation of ATAC-Seq, ChIP-seq and DNase-Seq data found LanceOtron to surpass the established gold standard tools through enhancing selectivity and obtaining flawless sensitivity (Hentges *et al.*, 2022). For this reason, LanceOtron was implemented for peak calling across the datasets.

Running CATCH-UP led to more refined and therefore accurate chromatin datasets as shown in the example below. In figure 31, the cardiac muscle cell DNase peaks showed a cleaner track with more stringently assigned peaks after running CATCH-UP with LanceOtron (overall peak scores ≥ 0.5) compared to before. Overall peak scores ≥ 0.5 were chosen as this is the widely accepted threshold (as seen by Gao *et al.* (2023) and Crump *et al.* (2023)).

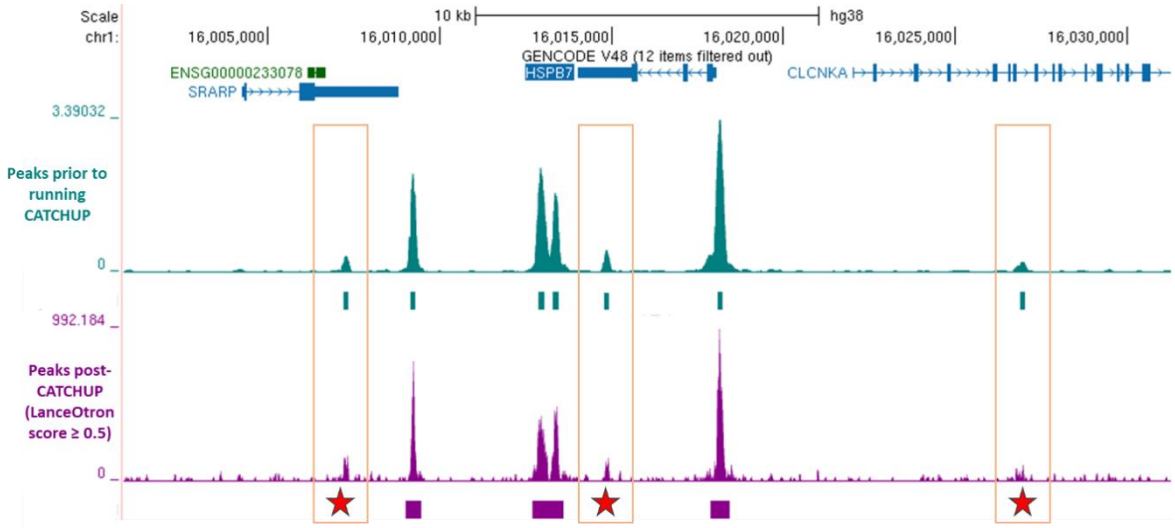


Figure 31 | Chromatin QC (CATCH-UP) visualisation before and after | A figure depicting cardiac muscle cell DNase peaks (specifically bigwig file peaks and bed file bars) first prior to running CATCH-UP by Riva et al. (with LanceOtron integrated for peak calling) in teal followed by post-CATCH-UP in purple. Three regions are shown in orange boxes depicting cleaner tracks post-CATCH-UP with more stringently assigned peaks (LanceOtron score > 0.5 resulting in peaks not being called as shown by red stars) as compared to beforehand (where the peaks are still being included as bars).

4.4 Using Machine Learning to Prioritise Functional Variants

4.4.1 DeepHaem

DeepHaem was employed, given various chromatin datasets as input (from heart tissue and cell types as described in github.com/domwest/DPhil-thesis/blob/main/data_tracking_sheet_for_thesis.xlsx), for predicting the presence or absence of open chromatin peaks. The model was trained using these datasets and its performance assessed through both computational validation and visual inspection. Ultimately, in the event of optimal model performance, the impact of HCM variants could be predicted.

Model trained with 24 classifiers

The first DeepHaem model was trained with 24 classifiers representing those from heart tissue as well as heart cell type level datasets; these datasets included ATAC, DNase, ChIP Histone, and ChIP TF methodologies.

Computational validation

Post-training, DeepHaem was validated computationally through the assessment of auc and loss plots (per training and validation sets) to assess the model's performance. Additionally, ROC curves representing the performance per classifier/dataset were interrogated to ensure optimal performance globally without underperforming classifiers.

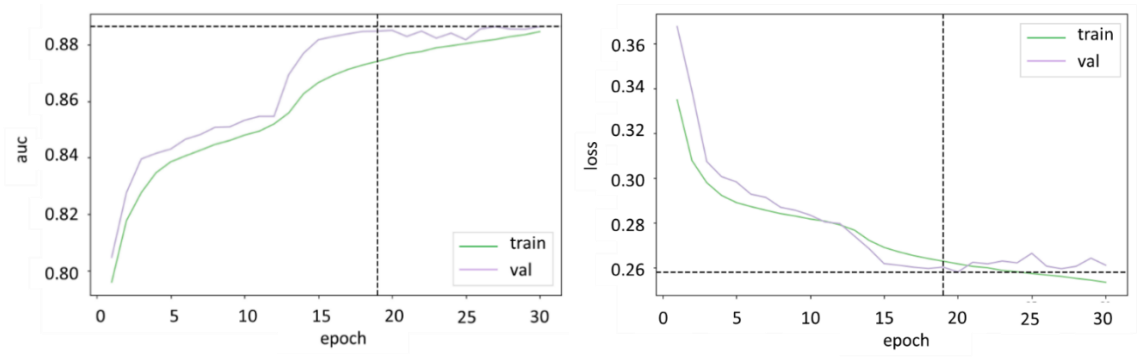


*Figure 32 | DeepHaem training & validation auc/loss plots with 24 classifiers and sigmoid final activation layer| A figure depicting the area under the curve (auc) plot on the left-hand side as well as the loss plot on the righthand side as a means of showing DeepHaem's performance (**given a sigmoid activation in the final layer**) with **24** classifiers/datasets.*

The auc plot in figure 32 illustrates that the final auc scores surpassed 0.88 and does not show any signs of overfitting or underfitting. A score > 0.7 (70%) generally reflects adequate model performance. The effect of the number of epochs on accuracy is influenced by several factors such as datasets, model architecture and training hyperparameters (settings defined before model learning). In general, one expects the model accuracy to increase with the number of epochs. Overfitting may lead the validation accuracy to decrease after a given number of epochs. This is

caused by the network being unable to generalise well to unseen data because of the noise and randomness being modelled from the training set. Using tactics like early stopping and learning rate schedule (which both demand rigorous experimentation) can help achieve strong generalisation performance. Underfitting, where patterns in the training dataset are too complex for the model to capture leading to poor performance in both training and validation, can also be seen in the auc plots with all-round poor auc scores being obtained for both training and validation sets.

Model loss plots were generated where loss represents binary-cross entropy (figure 32). This measures the dissimilarity between the actual labels and the predicted probabilities of the data points being in the positive class; predictions that are confident but wrong are penalised. Obtaining final loss scores of less than 0.05 is considered on the right track, 0.2 just fine, and anything above 0.3 not great while above 1 is terrible. Overfitting is suggested if there is a sudden increase in the loss for the validation set and underfitting when there is a large gap between the training and validation curves (highlighting that the model is missing information). Additionally, while training occurs, if there is no significant progress being made after multiple epochs then the training loss curve will end in a plateau. Since this is not the case, the number of epochs seemed to be fitting.

The ROC curves' true positive rates versus false positive rates were also compared. In each scenario, the area under the curve was inspected to assess the probability that if given a randomly chosen positive and negative case the model would rank the positive higher than the negative.

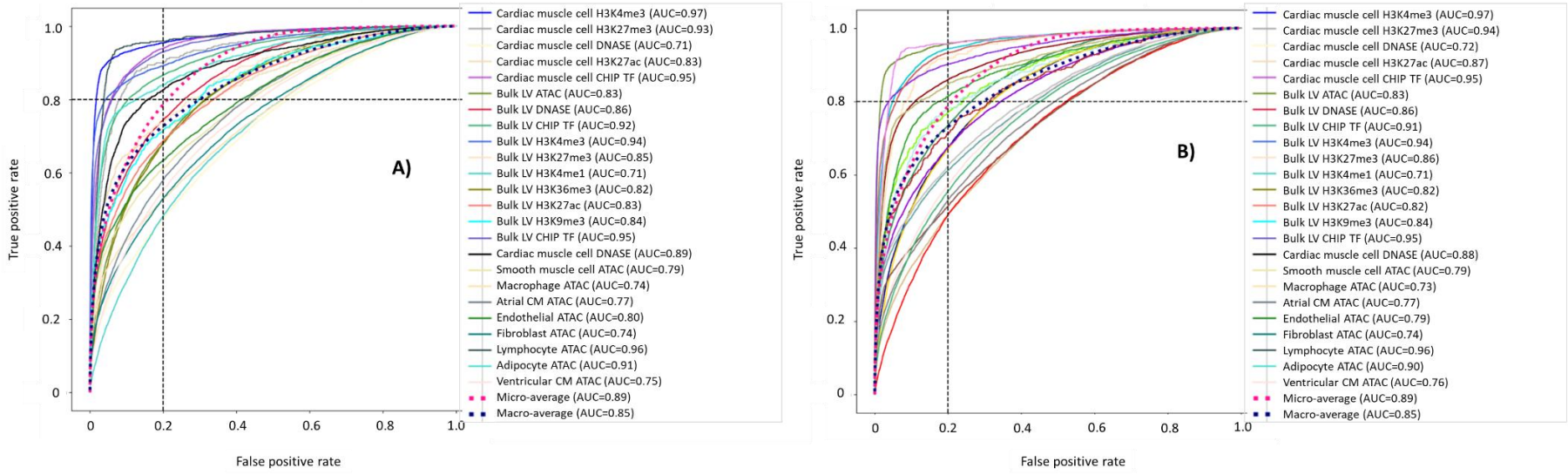


*Figure 33 | DeepHaem training and validation ROC curves for 24 classifiers and sigmoid final activation layer | A figure depicting receiver operating characteristic (ROC) curves for all **24** classifiers/datasets of true versus false positive rates to evaluate classifier performance during DeepHaem **testing in A) and validation in B)**. All classifiers and their associated area under curve (auc) scores are shown in the legend.*

An auc score of 1 would suggest perfect performance and 0.5 would suggest that the classifier's prediction is no better than random guessing. All auc scores for these classifiers (in figure 33) showed an auc score above 0.7 (70%) with some being very high (0.97, for example). This was consistent for both the testing data (figure 33A) as well as the validation (figure 33B) set. Additionally, a micro- and macro- average was calculated. The micro-average combines all classes into a single binary classification step before calculating the auc score and so proportions are considered (e.g. greater weight for larger classes); this is a useful metric when there is imbalance across classes and performance should reflect the bigger classes. The macro-average calculates the metric for each class separately before calculating the unweighted mean – in this way, all classes are treated equally regardless of their size. Both micro- and macro- averages (> 0.8) showed excellent performance (Mandrekar, 2010) and so these results confirmed that it was not necessary to troubleshoot any of the classifiers' performance.

Visual inspection

DeepHaem was also validated through the comparison of its prediction (of the underlying chromatin in particular regions) with that seen in the raw data. Additionally, underlying sequences were tweaked (through the insertion of an artificial peak or otherwise removal of a peak) before running the prediction to ascertain whether the model could infer a change in pattern.

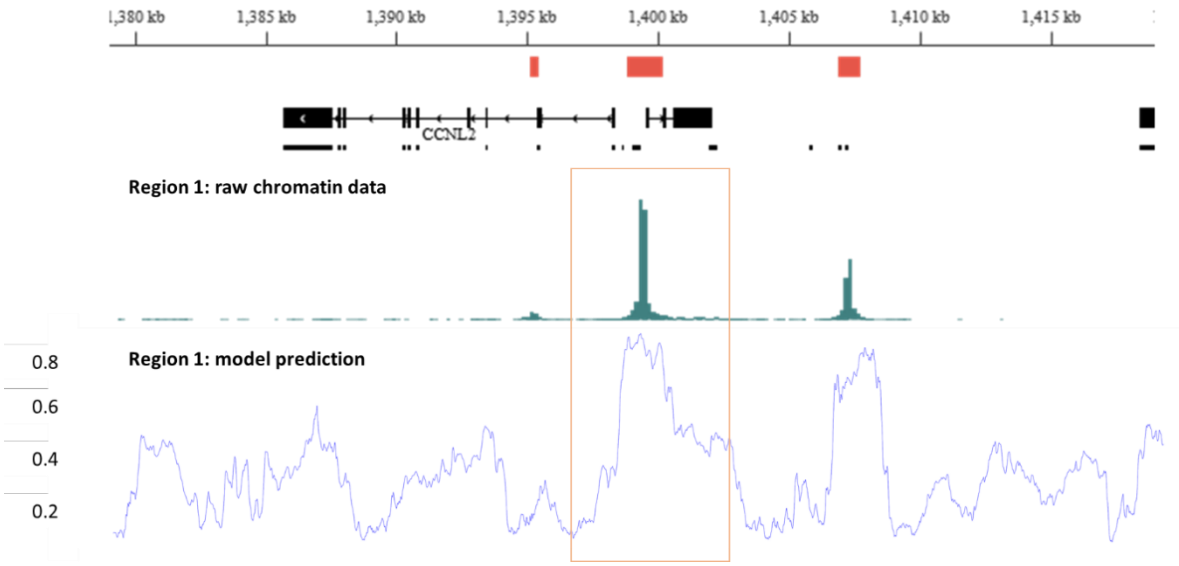


Figure 34 | DeepHaem visual inspection: region1 in ventricular cardiomyocyte classifier | DeepHaem's rolling window prediction for a particular region (region 1; detailed in table 4) in the ventricular cardiomyocyte classifier/dataset. The values on the y-axis represent a probability score as to whether the chromatin is likely open at a particular base position (value nearing 1) or closed (nearing zero). The x-axis represents genomic coordinates. The orange box highlights which peak is being examined.

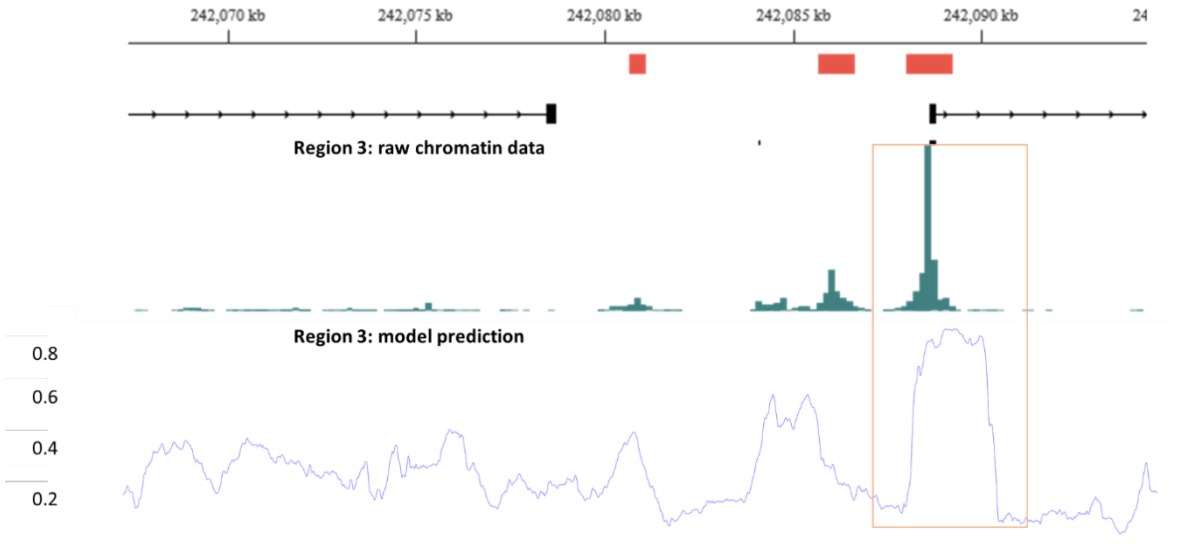


Figure 35 | DeepHaem visual inspection: region2 in ventricular cardiomyocyte classifier | DeepHaem's rolling window prediction for a particular region (region 3; detailed in table 4) in the ventricular cardiomyocyte classifier/dataset. The orange box highlights which peak is being examined.

These validations proved successful with the rolling window prediction closely resembling the raw chromatin at the specified regions. In region 1 (shown in figure 34) there are 2 prominent peaks while in region 3 (figure 35) there is 1. In both cases, the bases at these peak regions achieve probability scores ≥0.8. This suggests the machine learning model is predicting open chromatin accurately.

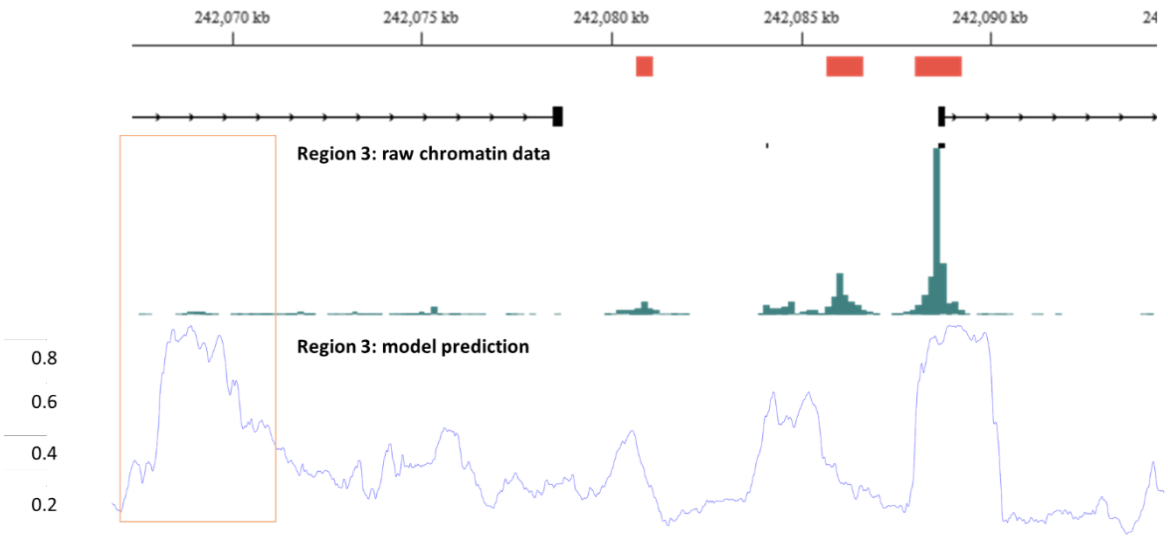


Figure 36 | DeepHaem visual inspection with modified DNA sequence: region 3 peak creation in ventricular cardiomyocyte classifier | Before testing DeepHaem's rolling window prediction, a modification was applied to the underlying DNA sequence whereby a peak sequence was extracted from region1 (detailed table 4) and inserted into an otherwise flat site of region3 (detailed table 4). The orange box highlights where the peak creation occurred.

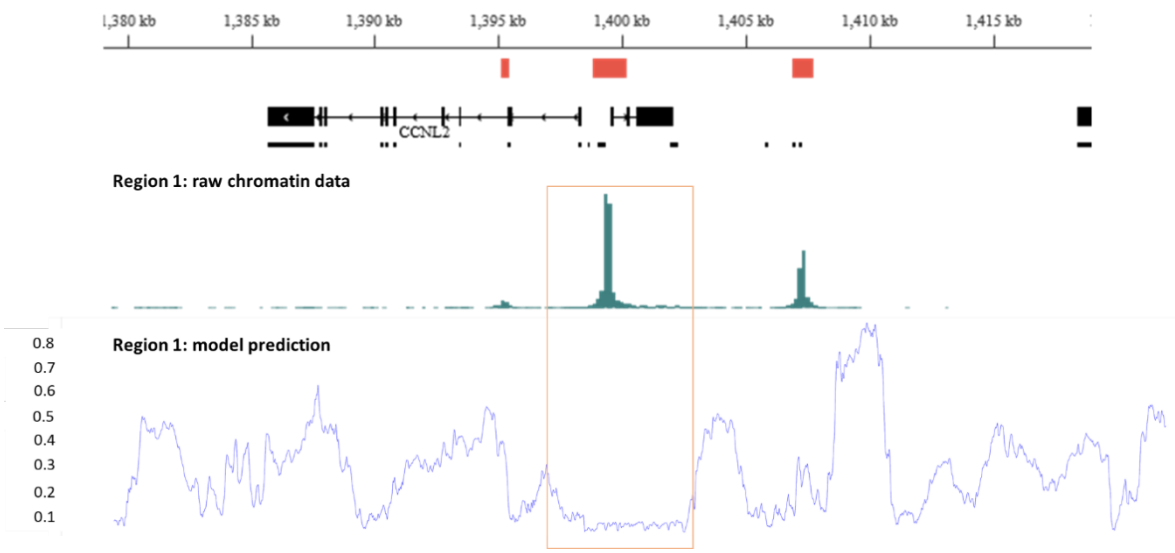


Figure 37 | DeepHaem visual inspection with modified DNA sequence: region1 peak erasure in ventricular cardiomyocyte classifier | Before testing DeepHaem's rolling window prediction, a flat sequence was extracted from region3 (detailed table 4) and inserted into the peak site of region1 (detailed table 4). The orange box highlights where the peak erasure occurred.

The model passed these additional validation checks which meant that it was able to detect peak patterns (even if they were being edited in). A clear peak was introduced at the beginning of region3 (figure 36) and there was a clear loss of one peak in region 1 (figure 37). The only peak in region 3 as well as the other peak in region 1 remained and so the model could accommodate the change in peak patterns without this affecting the entire region's prediction.

Being able to recognise both the forward and reverse strands of a non-palindromic motif in a CNN is difficult and explains why the different strands will be assigned different predictions or scores (Zhou *et al.*, 2021). In these scenarios, motifs could be missed if the model only has 1 strand to base its predictions on. There have been different approaches to assigning an overarching score which encompasses the combination of both strands' scores. DeepHaem takes the minimum score at each base with the aim of minimising the noise in a non-peak region albeit with the consequence of losing some significance when the region is in fact a peak; figure 38 illustrates an example of the forward, reverse, and minimum prediction for region 1 (raw peaks in figure 37) which shows the accurate prediction of two peaks (surpassing a probability of 0.75).

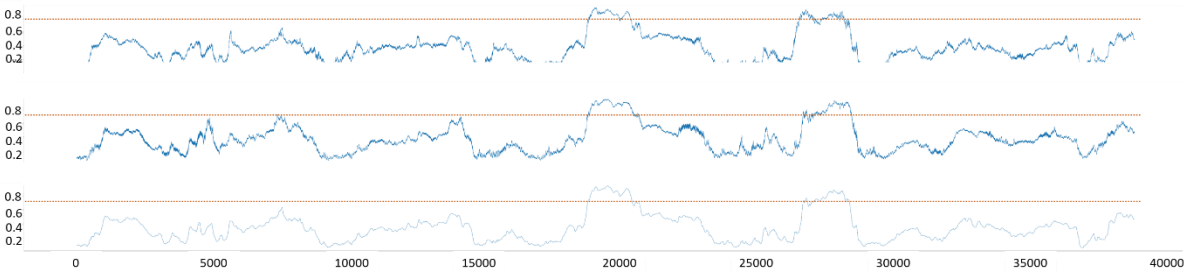


Figure 38 | DeepHaem's rolling window prediction visualisation: forward, reverse, and combined strands | Three rolling window predictions are shown (region 1; detailed in table 4): the forward strand prediction is shown at the top, the reverse strand prediction in the middle, and the combined prediction (minimum score per base considering both strands) at the bottom. Orange horizontal lines represent the 0.75 probability threshold above which peaks are predicted by the model.

Post-analysis: variant prioritisation first attempt

In addition to DeepHaem being able to predict the chromatin landscape at a particular region in the genome, per-base prediction scores can be leveraged to generate a 'damage score' for a particular variant. DeepHaem damage scores are generated by first performing a prediction with the reference allele inserted at the variant's position followed by that with the alternate allele and ultimately by combining both scores to determine what effect (and magnitude thereof) a variant likely has.

Authors of DeepLIFT plotted the distribution of LD-adjusted ISM scores (i.e. the maximum ISM score found across all variants in a particular LD block) for the lead versus proxy variants compared to unadjusted scores of both (Movva *et al.*, 2019). The resulting distributions did not differ and they concluded that proxy SNPs should be equally considered regarding possible causality. For that reason, each of the 4395 variant's individual damage scores from DeepHaem were plotted without any LD re-assignment.

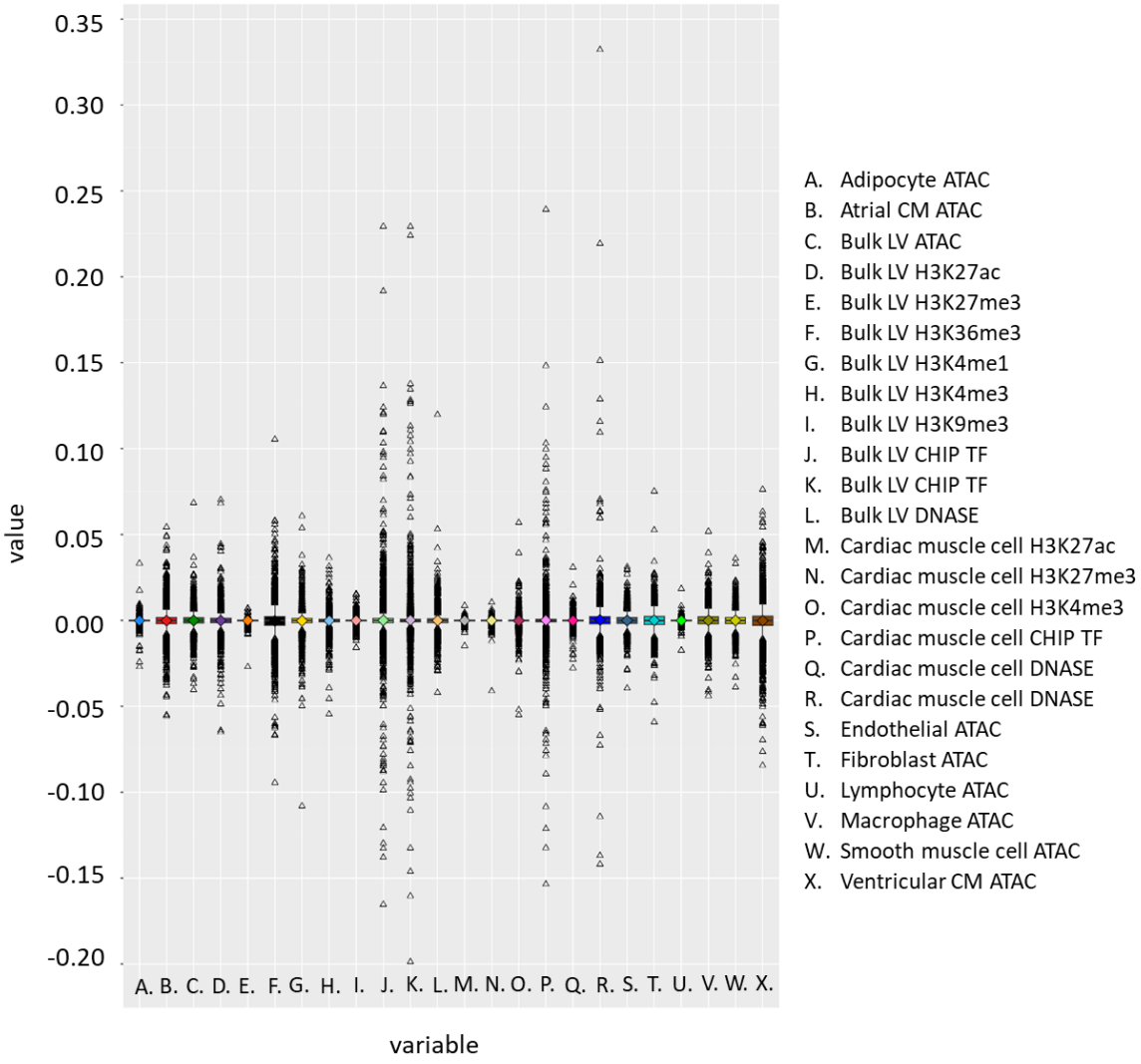


Figure 39 | DeepHaem damage score distribution per 24 classifiers for HCM variants | Figure illustrating boxplots of damage scores generated for all 4395 lead and proxy HCM GWAS variants across the classifiers. The y-axis shows the assigned damage scores and the x-axis shows the 24 different classifiers/datasets.

The set of 4395 HCM variants' damage score distribution can be seen for each of the 24 classifiers in figure 39. Some classifiers show more spread than others – with very few variants showing extreme damage scores (either a very high or low score relative to the others).

The vast majority of variants did not cross the previously determined 'empirical threshold' of ±0.1 for variant prioritisation. This cut-off was deemed inapplicable considering it was derived using blood cell data and thus having little relevance to HCM-related datasets.

Post-analysis: variant prioritisation via a positive control set

The next approach was to use two plausible sets of HCM causal noncoding variants to act as 'positive controls' and thus contribute to the derivation of a prioritisation threshold. One set of variants was extracted from a recent paper which suggested a list of putative regulatory HCM variants (Lesurf *et al.*, 2022) and the other set from RegulomeDB with high associated evidences.

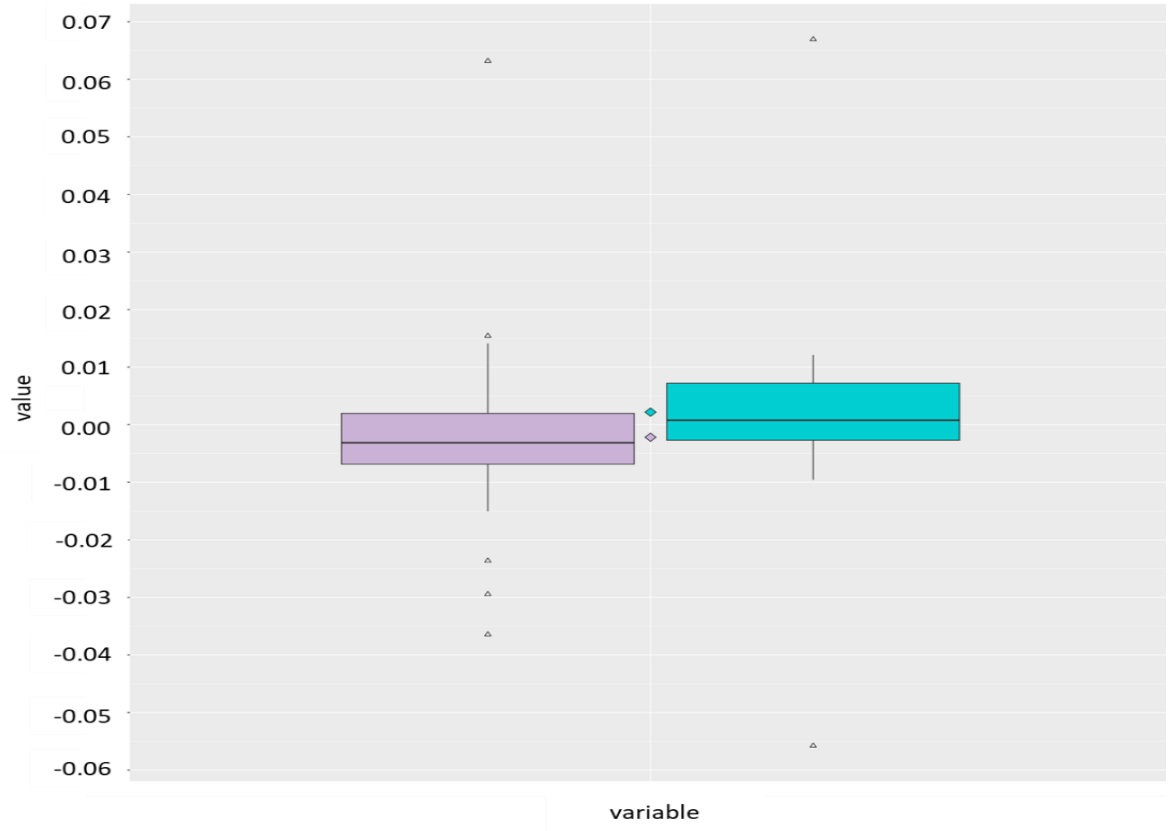


Figure 40 | DeepHaem's damage score distribution per HCM positive control set in ventricular cardiomyocyte classifier | A figure illustrating two boxplots for the ten variants from RegulomeDB (described in table 5) and 41 from Lesurf et al. (2022) (described in table 6) for DeepHaem's ventricular cardiomyocyte classifier. The y-axis shows the assigned damage scores and the x-axis is split by variant set as coloured by that specified in the legend on the right-hand side.

The majority of these 2 sets of variants scattered closely around zero (as shown in figure 40) which was not the expected result. Because these are supposedly high-confidence regulatory variants associated with HCM, one would expect all of them to achieve either very high or low damage scores

suggestive of their impact. In light of this result, gain and loss thresholds from either of these sets would not be insightful and therefore transferable for the prioritisation of HCM variants.

Post-analysis: variant prioritisation via the damage score distribution

Because HCM does not have an adequate set of positive control variants for benchmarking the damage score, the next approach involved plotting the overall distribution of damage scores as a means to determine a prioritisation cut-off (an example of the damage score distribution is shown for a classifier in figure 41).

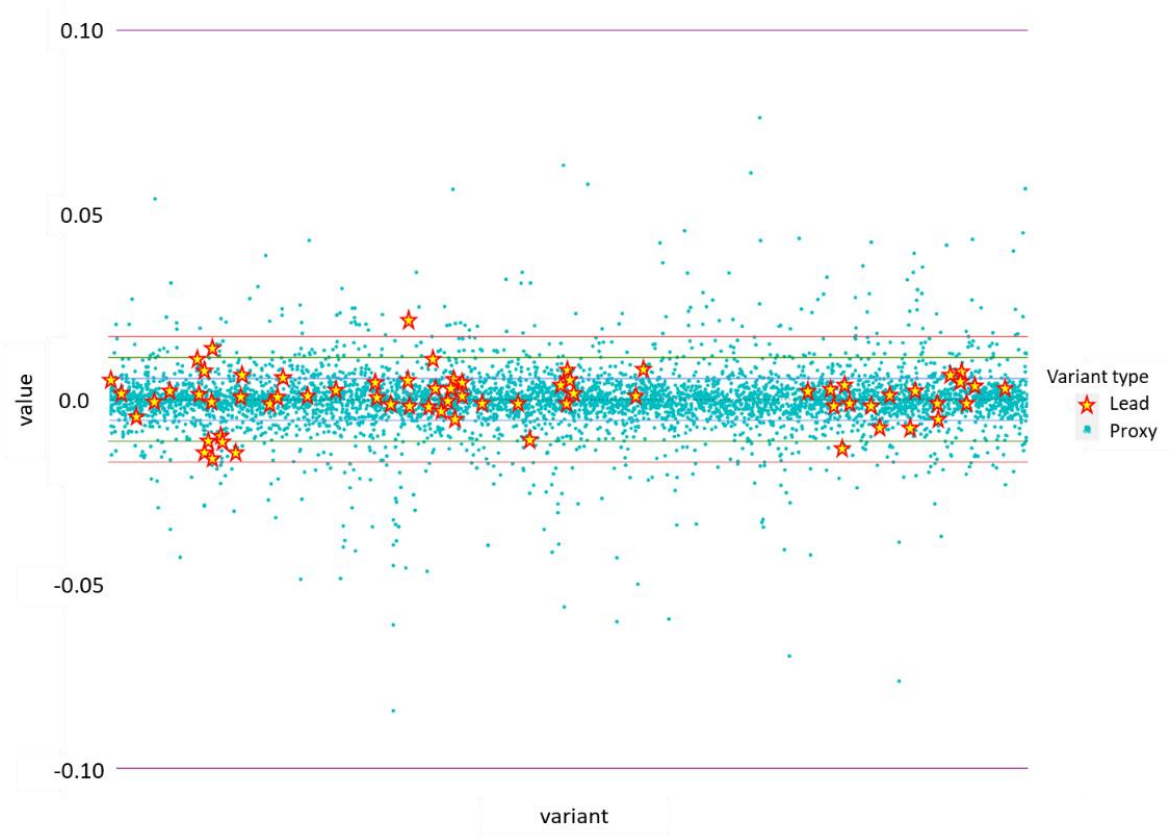


Figure 41 | DeepHaem's damage score distribution for HCM GWAS variants in the ventricular cardiomyocyte classifier | A figure illustrating the distribution of damage scores generated for all 4395 lead and proxy HCM GWAS variants. The y-axis shows the assigned damage scores and the x-axis has the variants plotted randomly. The empirical threshold shown in purple (±0.1) is used by the developers of DeepHaem, in addition to the mean±3SD (±0.0167) in red, mean±2SD (±0.0012)

in green, and the mean±1SD (±0.0005) in blue. The legend on the right also shows the colour code used to categorise variants that were either lead variants (red stars) or proxy variants (blue).

Extracting variants annotated with a damage score greater than the mean±2SD resulted in 1210 out of the 4395 variants being prioritised. This prioritisation method was abandoned considering the huge number of variants being prioritised. This approach may have been more effective if the underlying data depicted a normal distribution; however, it does not. Regardless, the fact that a positive control set was still lacking and therefore no real confidence could be assigned to 'extreme' damage scores – the exploration for a positive control group continued (except this time not just for HCM-related but rather a credible set of erythroid variants).

Trained model with 65 classifiers (including erythroid positive control tracks)

In addition to the previous heart tissue and cell type chromatin datasets (used as input for the DeepHaem model with 24 classifiers), a new model was also trained with erythroid tracks (from the Hughes' lab) as well as additional acquired heart ATAC datasets.

Computational validation

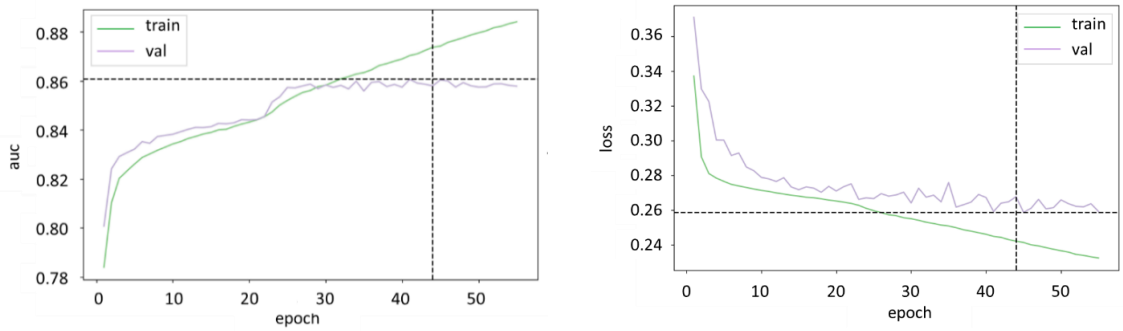


Figure 42 | DeepHaem training & validation auc/loss plots with 65 classifiers and sigmoid final activation layer | A figure depicting the area under the curve (auc) plot on the left-hand side as well as the loss plot on the righthand side as a means of showing DeepHaem's performance (given a ***sigmoid*** *activation in the final layer).*

Similar auc and loss scores were seen for this re-trained model with 65 classifiers (in figure 42) as was seen for the model with 24 classifiers. The final auc scores surpassed 0.86 and loss was around 0.26.

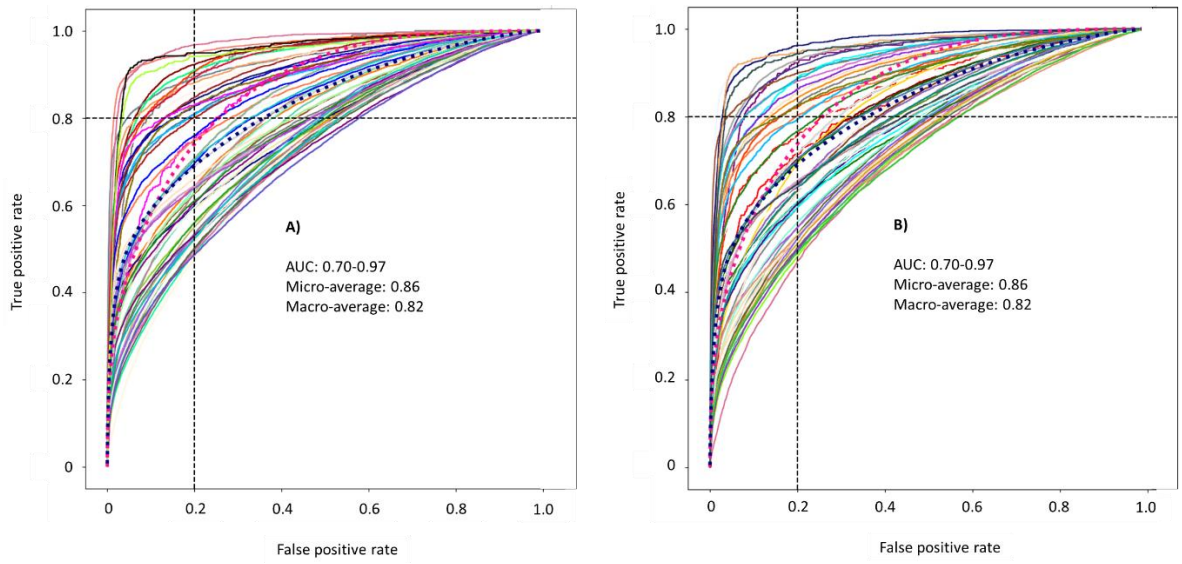


Figure 43 | DeepHaem training and validation ROC curves for 65 classifiers and sigmoid final activation layer | A figure depicting receiver operating characteristic (ROC) curves for all ***65*** *classifiers/datasets during DeepHaem* ***testing in A) and validation in B) (both given a sigmoid activation in the final layer)***.

None of the classifiers' ROC curve auc scores dipped below 0.7 (as suggested by the range of auc scores in figure 43) and so training was marked as a success.

Visual inspection

The visual inspections also proved successful with the newly trained model (and its 65 classifiers) accurately predicting these peak regions with probability scores ≥0.8. This model was also able to detect the introduction or removal of peak patterns.

Post-analysis: prioritisation via a positive control variant set

With 821 empirical erythroid enhancer variants being made available to me (from the Hughes' lab), the variants could be used as a positive control set for defining a prioritisation threshold.

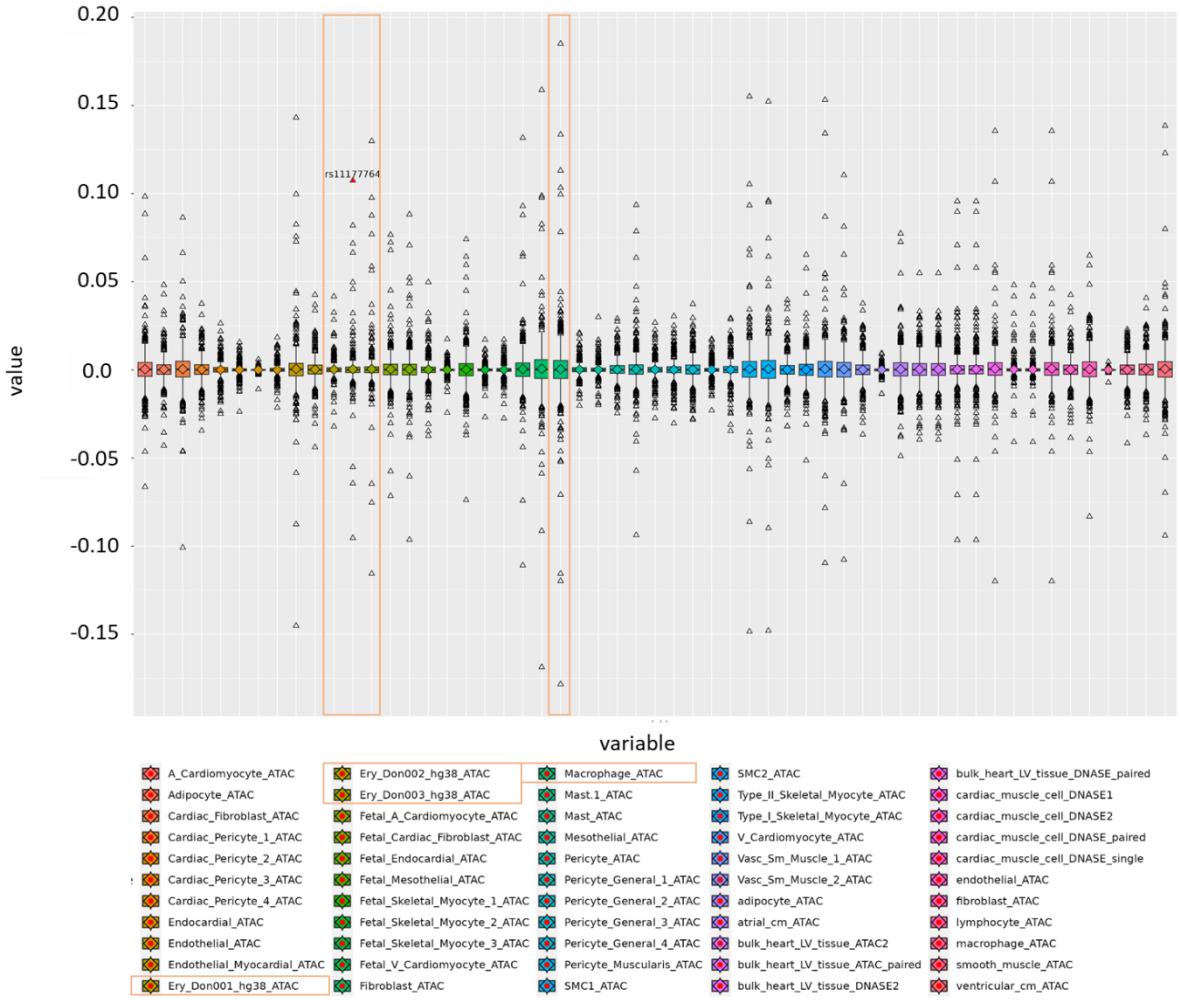


Figure 44 | DeepHaem damage score distribution per 65 classifiers for erythroid enhancer positive control set | A figure illustrating boxplots representing the distribution of damage scores generated for all 821 erythroid enhancer variants. The y-axis shows the assigned damage scores and the x-axis shows the 55 (out of 65) different classifiers/datasets which have been colour-coded in the legend below. The orange box highlights the three erythroid datasets as well as the macrophage dataset. Any variants within the 3 erythroid datasets achieving a damage score ≥ 0.1 or ≤ -0.1 were labelled in the diagram.

Unexpectedly and as shown in figure 44, these erythroid enhancer variants were not assigned extreme damage scores in the erythroid classifiers (Ery_Don001, Ery_Don002, and Ery_Don003) shown in the leftmost orange box, and they showed even greater damage scores in the heart classifiers. The highest damage score seen across the 3 erythroid classifiers was ~0.13 and the lowest ~-0.12 whereas the highest seen across all classifiers was ~0.185 and the lowest ~-0.2, both detected in the Macrophage_ATAC classifier (rightmost orange box).

Troubleshooting: tweaking the final layer's activation function

Because the empirical erythroid enhancer variants performed poorly regarding their damage scores (despite being positive control variants and thus expected to be assigned extreme damage scores), a shift occurred in the troubleshooting – from damage score distributions to the model's architecture and the optimality thereof (specifically its final layer activation function). For this reason, a SoftMax (as opposed to sigmoidal) activation function was implemented and the model performance checked through computational validation.

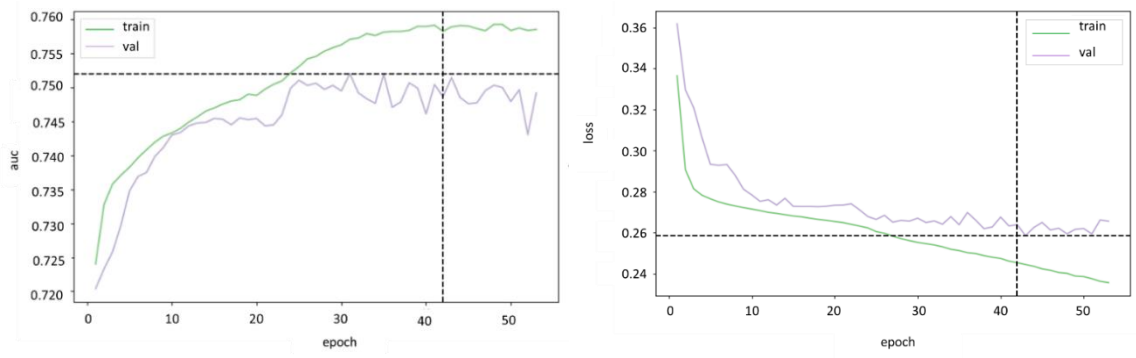


Figure 45 | DeepHaem training & validation auc/loss plots with 65 classifiers and SoftMax final activation layer | A figure depicting the area under the curve (auc) plot on the left-hand side as well as the loss plot on the righthand side as a means of showing DeepHaem's performance (given a ***SoftMax*** *activation in the final layer).*

The auc plot (figure 45) suggests that overfitting has occurred – this can be seen in the gap between the training and validation curves, as well as the fluctuating validation accuracy points. The validation loss curve also shows an early plateau suggesting that there is no significant progress

being made after 15 epochs. Overall, the final auc and loss thresholds are worse than previous training attempts.

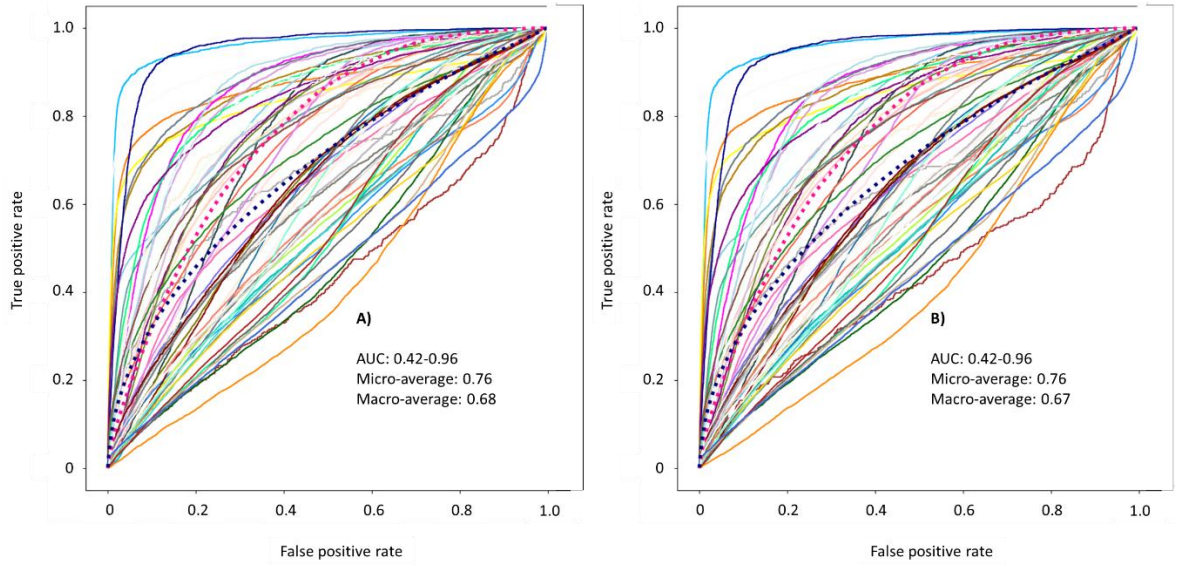


Figure 46 | DeepHaem training and validation ROC curves for 65 classifiers and SoftMax final activation layer | A figure depicting receiver operating characteristic (ROC) curves for all ***65*** *classifiers/datasets during DeepHaem* ***testing in A) and validation in B)****. Due to the vast number of classifiers, the range of associated area under curve (auc) scores, micro- and macro-average is shown.*

The model trained poorly and this is best captured in the ROC curves with some classifiers being assigned an auc less than 0.5 (as shown by the auc range in figure 46A and B); predictions seen by these classifiers are no better than random guessing. Whilst looking into this result, the decision to revert back to sigmoid was made. This was through the realisation that each classifier's output needed to be independent of the others especially given the split between erythroid and heart tracks but also due to possible cell type-specific differences in the chromatin landscape among the various heart tracks. With a SoftMax activation function, probabilities returned would sum to one which assumes the classes are mutually exclusive. Because each classifiers' prediction needs to be unaffected by the others, the sigmoidal activation function was decided to be better suited.

Troubleshooting: re-training the model with varying sets of classifiers

Under the same assumption that there was an architectural flaw in the model (as opposed to the erythroid enhancer variants plausibly having low to moderate damage scores in the erythroid classifiers), another attempt at improving the model's performance was to re-train different models with varying classifiers/datasets. Presuming that some of the datasets may have been too alike (e.g. multiple bulk LV chromatin datasets) or possibly even there being too much diversity across all the datasets (e.g. patterns detected across cardiac muscle cell, bulk LV, and erythroid datasets being contrasting), more select combinations of classifiers were used for training.

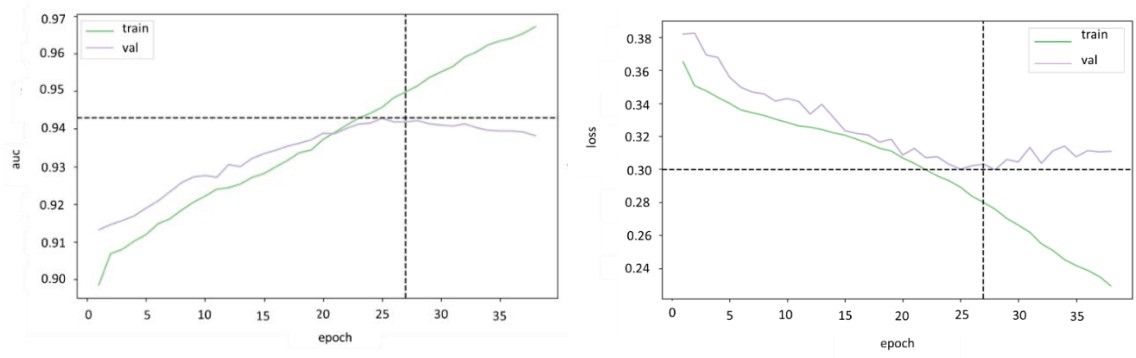


Figure 47 | DeepHaem training & validation auc/loss plots with 1 erythroid & 1 heart classifier and sigmoid final activation layer | A figure depicting the area under the curve (auc) plot on the left-hand side and loss plot on the righthand side as a means of showing DeepHaem's performance (given a ***sigmoid*** *activation in the final layer).*

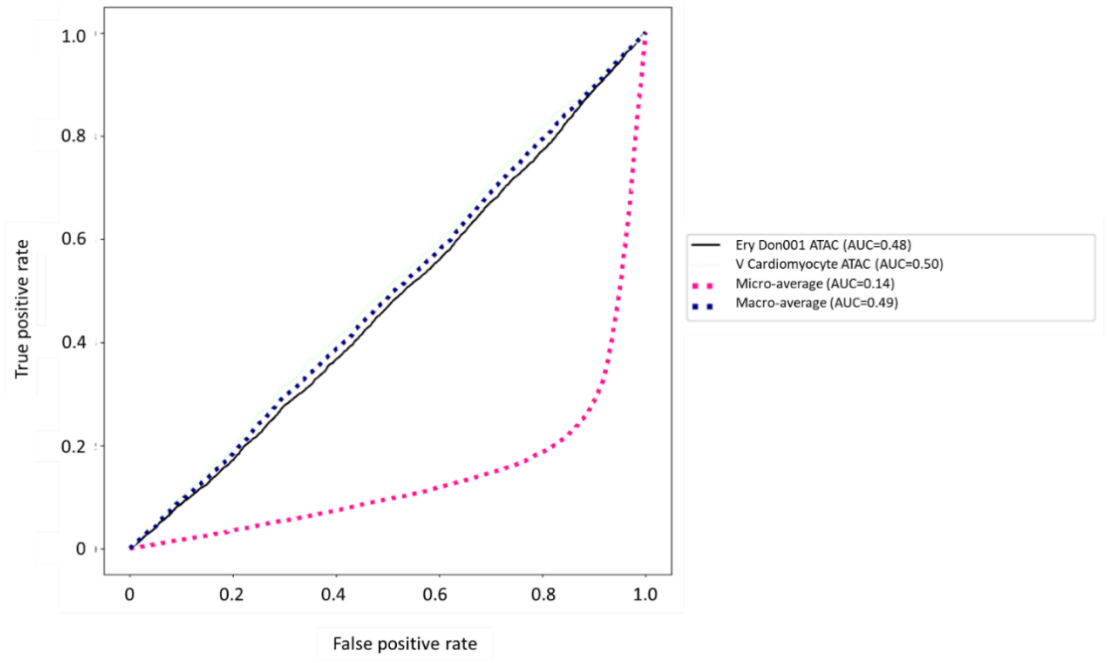


*Figure 48 | DeepHaem training and validation ROC curves for 1 erythroid & 1 heart classifier and sigmoid final activation layer | A figure depicting receiver operating characteristic (ROC) curves for **1 erythroid and 1 heart** (given a **sigmoid** activation in the final layer).*

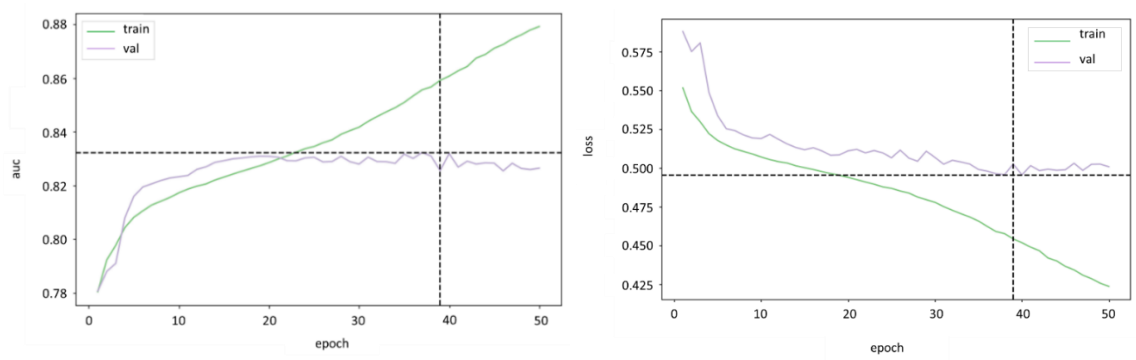


Figure 49 | DeepHaem training & validation auc/loss plots with 3 erythroid & 3 heart classifiers and sigmoid final activation layer | A figure depicting the area under the curve (auc) plot on the left-hand side and loss plot on the righthand side.

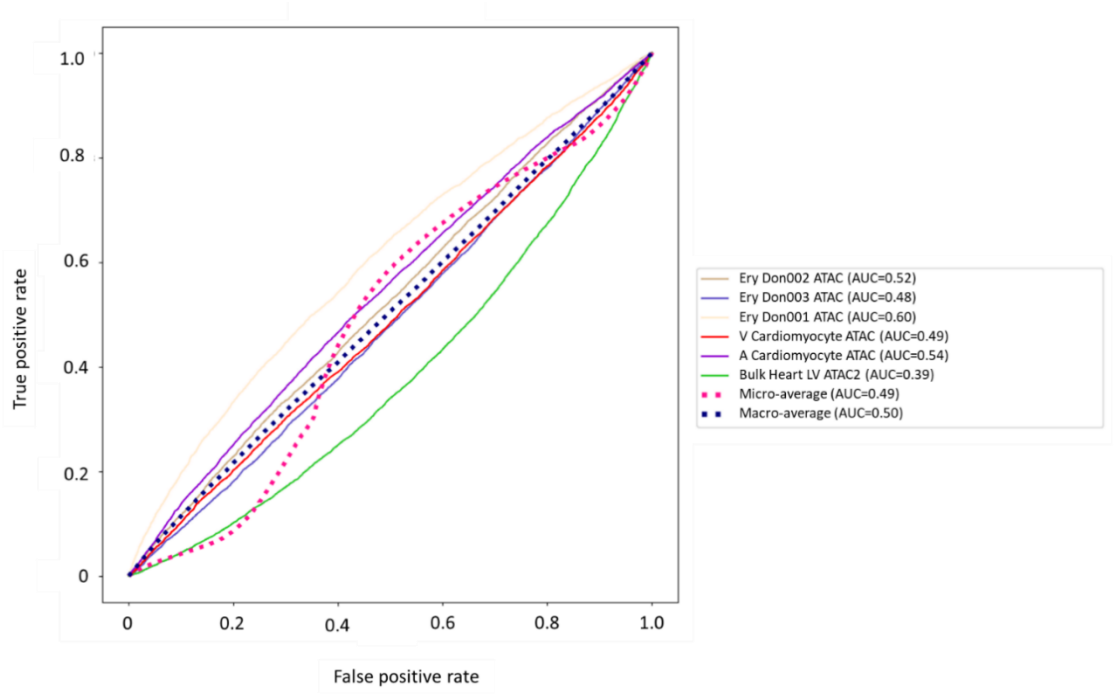


Figure 50 | DeepHaem training and validation ROC curves for 3 erythroid & 3 heart classifiers and sigmoid final activation layer | A figure depicting receiver operating characteristic (ROC) curves.

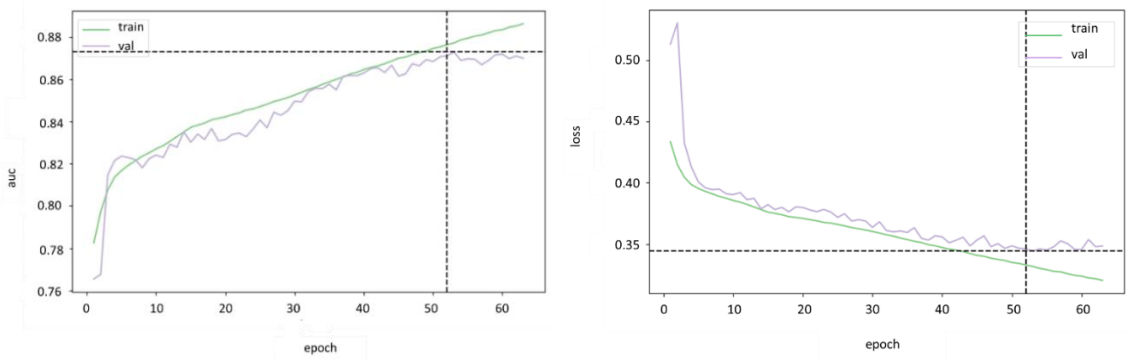


Figure 51 | DeepHaem training & validation auc/loss plots with 1 erythroid & 12 heart classifiers and sigmoid final activation layer | A figure depicting the area under the curve (auc) plot on the left-hand side and loss plot.

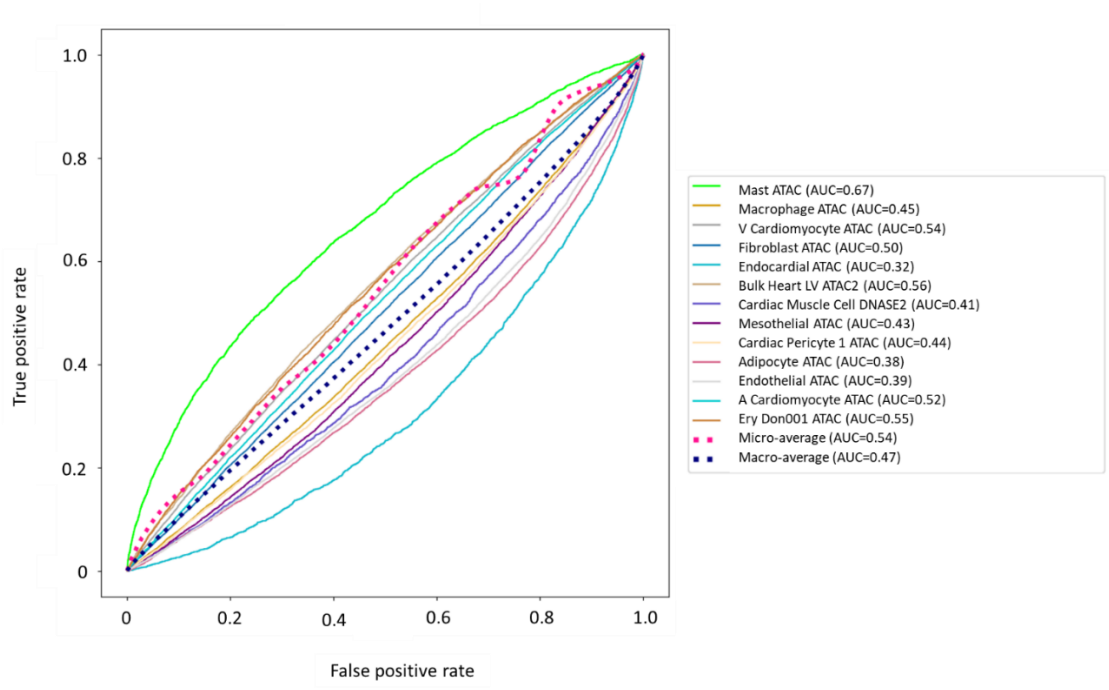


Figure 52 | DeepHaem training and validation ROC curves for 1 erythroid & 12 heart classifiers and sigmoid final activation layer | A figure depicting receiver operating characteristic (ROC) curves.

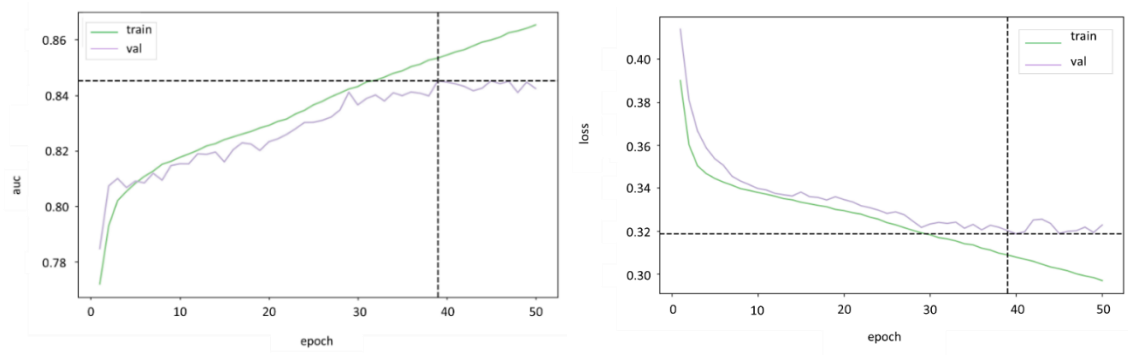


Figure 53 | DeepHaem training & validation auc/loss plots with 1 erythroid & 28 heart classifiers and sigmoid final activation layer | A figure depicting the area under the curve (auc) plot on the left-hand side and loss plot on the righthand side.

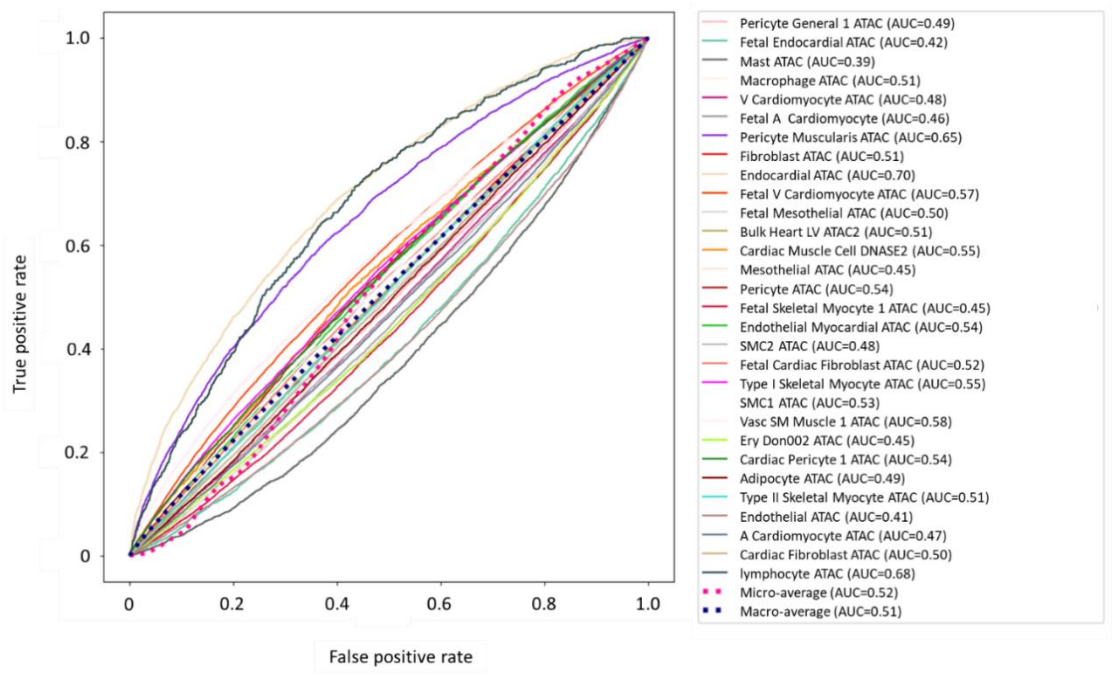


Figure 54 | DeepHaem training and validation ROC curves for 1 erythroid & 28 heart classifiers and sigmoid final activation layer | A figure depicting receiver operating characteristic (ROC) curves.

The first model including 1 erythroid track and 1 heart track (auc/loss plots in figure 47 and roc curves in figure 48) clearly did not have enough data to learn anything. The second model including 3 erythroid tracks and 3 heart tracks (auc/loss plots in figure 49 and roc curves in figure 50) suggest that the repetitive nature of the erythroid tracks may have hindered the learning. The 3 erythroid tracks all come from different donors, and so these could have been merged beforehand to avoid this issue. The third model including 1 erythroid track and 12 dissimilar heart tracks (auc/loss plots in figure 51 and roc curves in figure 52) aimed to rule out classifiers capturing repetitive data; however, this model trained poorly. Lastly a model including 1 erythroid track and 29 dissimilar heart tracks (auc/loss plots in figure 53 and roc curves in figure 54) which also aimed to tackle the same repetitive data issue (while simultaneously capturing a broad amount of heart data) still showed a poor performance. Overall, none of these models performed well as they were not able to generalise and so this attempt at improving model performance did not work. This resulted in the reversion back to the model with 65 classifiers and a sigmoid activation in the final layer.

Troubleshooting: investigating peak versus variant centering

Moving away from model architecture tweaking and representative classifiers, another plausible explanation for poor damage score annotation across the erythroid enhancer variants was rooted in DeepHaem's rolling window predictions. DeepHaem's prediction scores per base encompass an average of all overlapping window scores at that particular base. The new troubleshooting hypothesis was that a variant residing within the center of a peak would obtain a higher score (as a result of the pattern being more strongly detected given the surrounds) whereas those residing towards the end of said peak would be assigned lower scores. This could be supported by the nature of sliding windows contributing towards more windows being averaged for bases in the middle (and thus higher attributed confidence) as opposed to bases towards the end of the window being scored by fewer window predictions. As a result, peak-centering was implemented for the erythroid enhancer variants (as opposed to the standard approach of variant centering) to try and check if the damage scores would be improved because of this phenomenon; variant centering entails setting the window to 500 bases both upstream and downstream of the variant to get the exact prediction at the variant base whereas peak centering ensures that regardless of where the variant resides, the base at the peak's middle point would be scored.

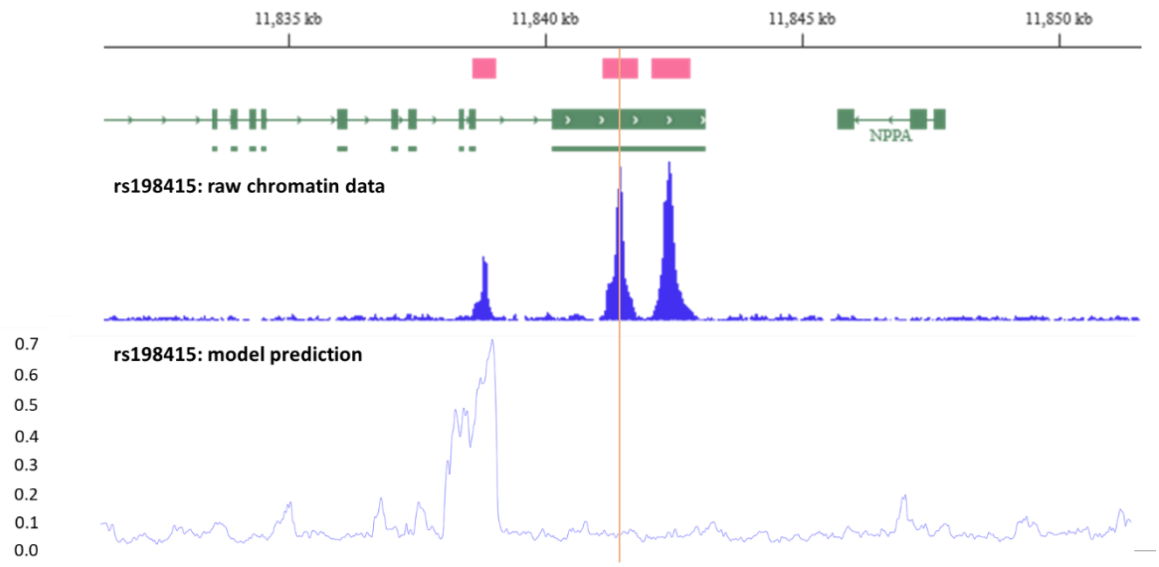


Figure 55 | DeepHaem visual inspection (peak centered): variant rs198415 in Ery Don003 classifier | A figure depicting a peak-centering prediction for comparison to variant-centering: DeepHaem's

rolling window prediction is shown for rs198415 (orange vertical line; peak coordinates: chr1:11841171-11841863) centered on the peak with the reference allele (i.e. variant not inserted) for the Ery_Don003 classifier/dataset.

Peak containing rs198415 (figure 55) is shown in the raw data for Ery_Don003 but not in the model's prediction (based on peak centering). There is a peak slightly to the left that is being captured but the 2 most prominent peaks are not.

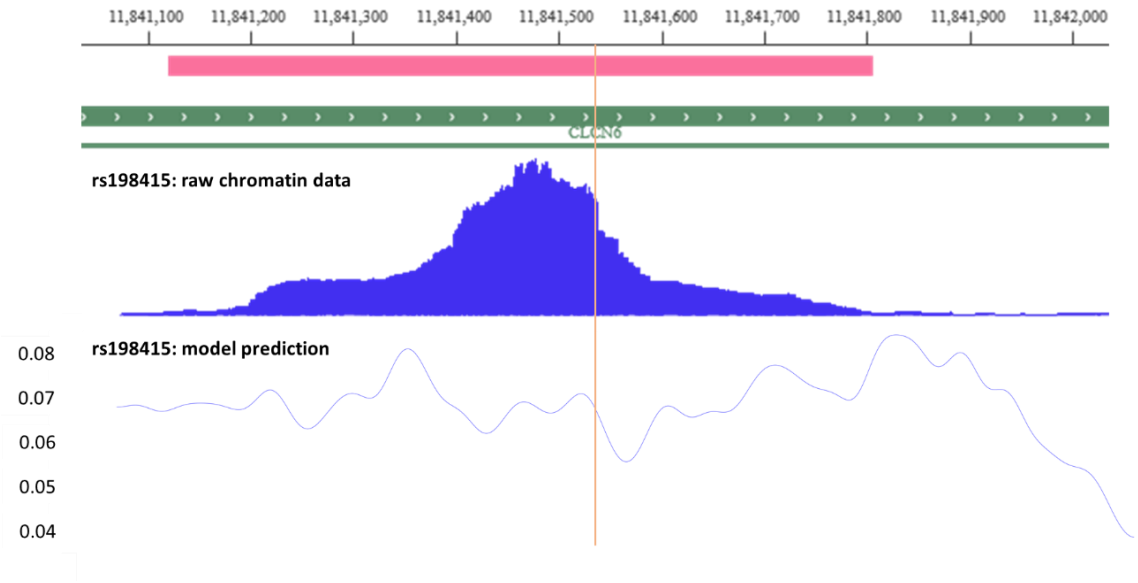


Figure 56 | DeepHaem visual inspection (variant centered): variant rs198415 in Ery Don003 classifier | A figure depicting a variant-centering prediction for comparison to peak-centering: DeepHaem's rolling window prediction is shown for rs198415 (orange vertical line; peak coordinates: chr1:11841171-11841863) centered on the variant with the reference allele (i.e. variant not inserted) for the Ery_Don003 classifier/dataset.

The same outcome is seen for the prediction based on variant centering (in figure 56). The y-axis probability score only reaches 0.06 in this region suggesting that the accessibility is completely flat across this peak. These predictions highlighted that the actual peaks were not always being detected regardless of the centering. This confirmed that DeepHaem was not able to predict the Erythroid peak regardless of whether the prediction was centered on the peak or the variant itself. Consequently, peak validation would need to be carried out again to ensure accurate predictions for

the erythroid datasets (considering previous approaches focused on confirming those in the heart datasets only).

Troubleshooting: visual inspection with a focus on erythroid datasets

The previous round of troubleshooting (which focused on variant versus peak centering) highlighted that peaks specific to the erythroid datasets were not being predicted in the first place – explaining why the variants were being assigned low damage scores (as they were not predicted to reside in peaks). For this reason, visual inspection (comparing the model's predictions versus the actual raw chromatin) was carried out again except this time with less focus on the heart datasets (which did not flag up issues previously) and rather on the erythroid datasets. Two scenarios were explored – one where a particular peak was present across many datasets, and another where the peak was only detected in erythroid datasets to see if the model could accurately predict both scenarios.

Scenario1: Peak present in all classifiers

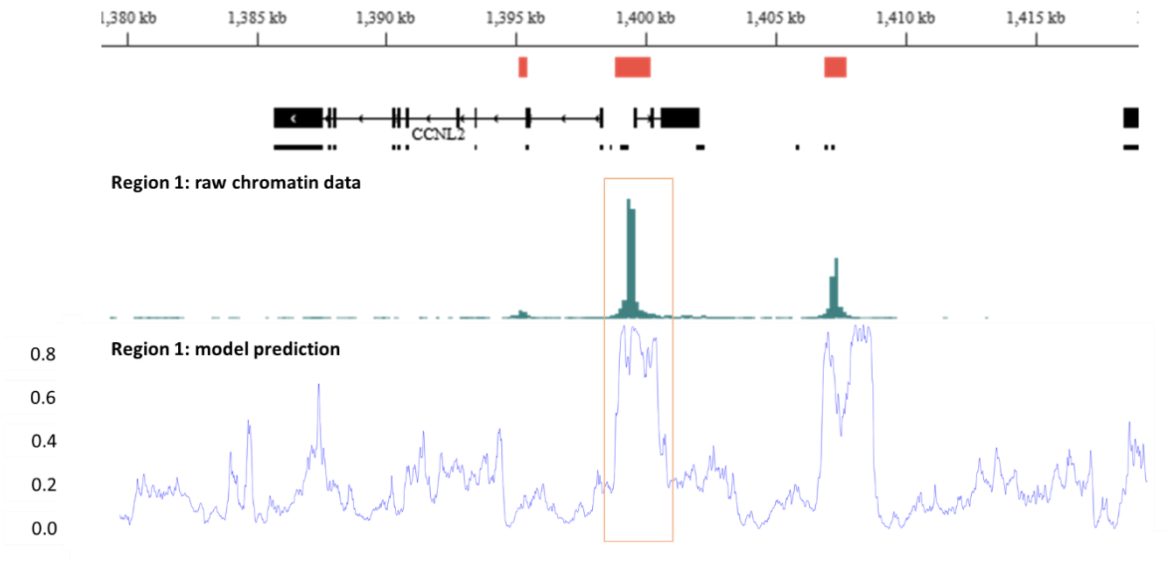


Figure 57 | DeepHaem visual inspection: region1 in ventricular cardiomyocyte classifier | DeepHaem's rolling window prediction for a particular region (region 1; detailed in table 4). The orange box highlights which peak is being examined.

The peak captured in the orange box (in figure 57) is captured across all 65 classifiers in the model and as shown above, the ventricular_cm classifier is accurately able to detect this peak pattern and assign 0.8 probability scores to the majority of bases in this peak.

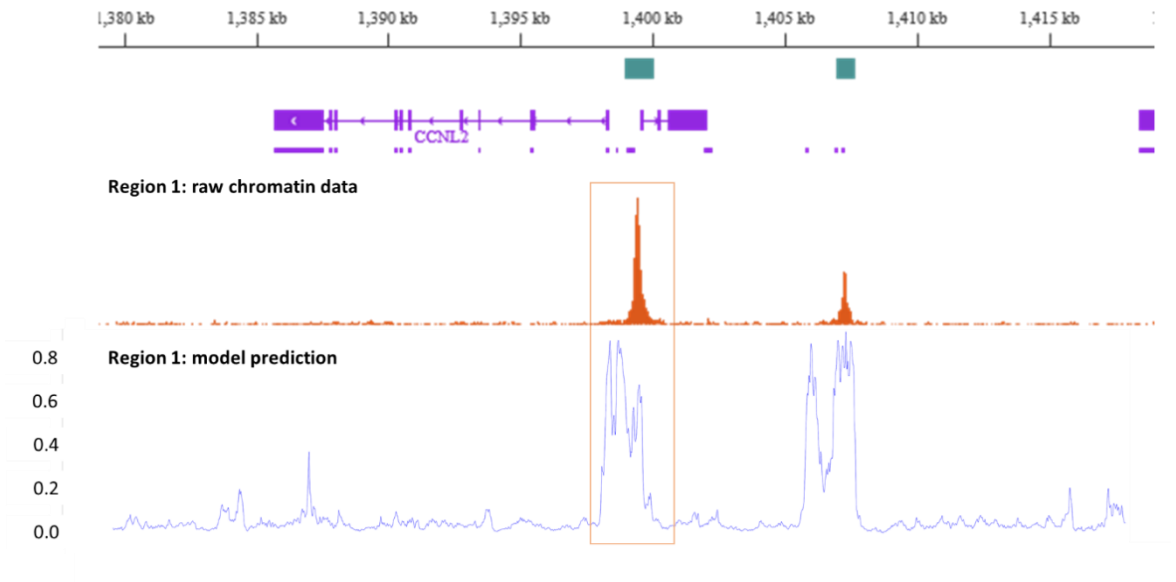


Figure 58 | DeepHaem visual inspection: region1 in Ery Don001 classifier | DeepHaem's rolling window prediction for a particular region (region 1; detailed in table 4) in the Ery_Don001 classifier/dataset. The orange box highlights which peak is being examined.

Likewise, the same peak as is shown in figure 57 is accurately being predicted for the Ery_Don001 classifier (with a probability score of 0.8 being obtained) in figure 58.

Scenario2: Peak only present in erythroid classifiers

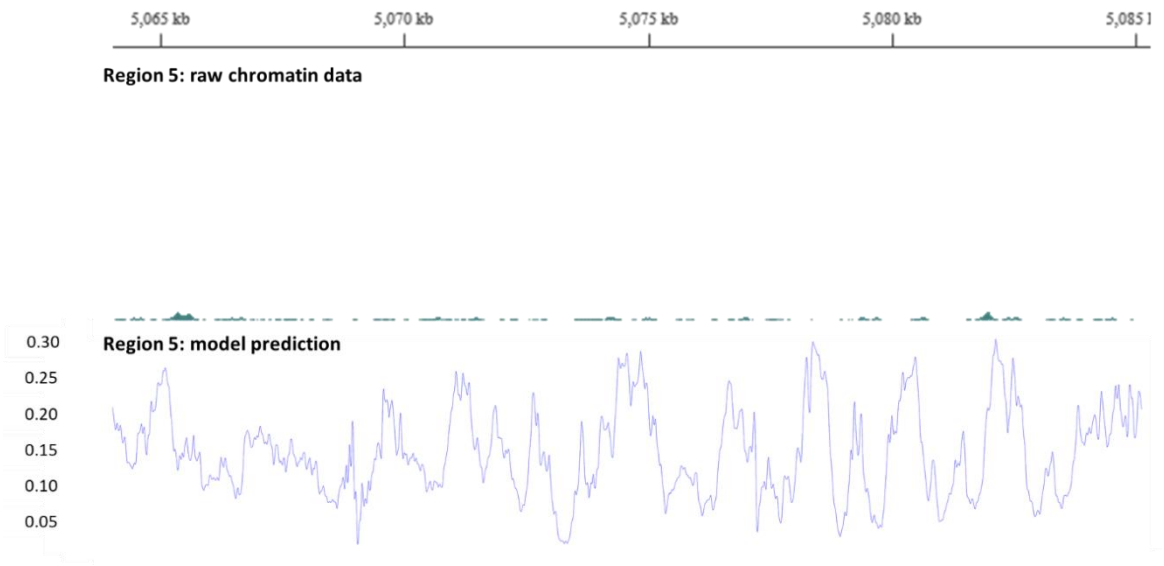


Figure 59 | DeepHaem visual inspection: region4 in ventricular cardiomyocyte classifier | DeepHaem's rolling window prediction for a particular region (region 4; detailed in table 4) in the ventricular cardiomyocyte classifier/dataset.

As expected, when predicting the chromatin landscape in a region that is otherwise closed (for ventricular_cm) the prediction is flat throughout this region (figure 59).

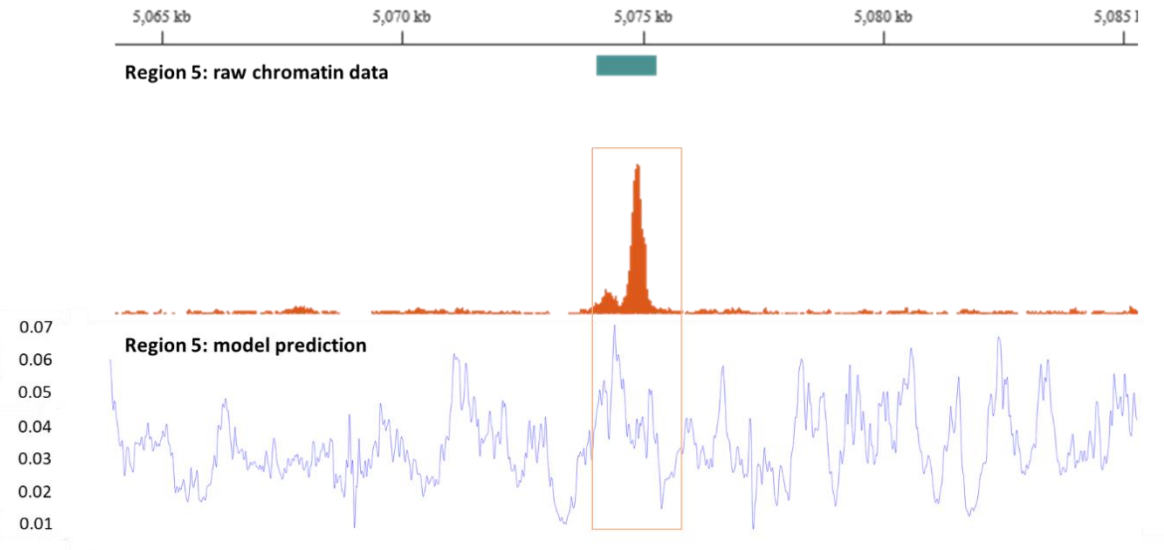


Figure 60 | DeepHaem visual inspection: region4 in Ery Don001 classifier | DeepHaem's rolling window prediction for region 4. The orange box highlights which peak is being examined.

When looking at the Ery_Don001 classifier in a region containing a peak (captured across all the erythroid classifiers but not the heart tracks), the prediction proves to be flat with a maximum probability score of 0.07 being obtained in the region (shown in figure 60). This prediction is not accurately predicting this peak.

These results accentuated how peaks present in both heart and erythroid classifiers were accurately predicted in the erythroid datasets; however, peaks specific to the erythroid datasets were not predicted. Due to the fact that these erythroid enhancer variants are present in peaks specific to erythroid datasets, this explains why the model was not predicting their impact effectively.

DeepHaem findings

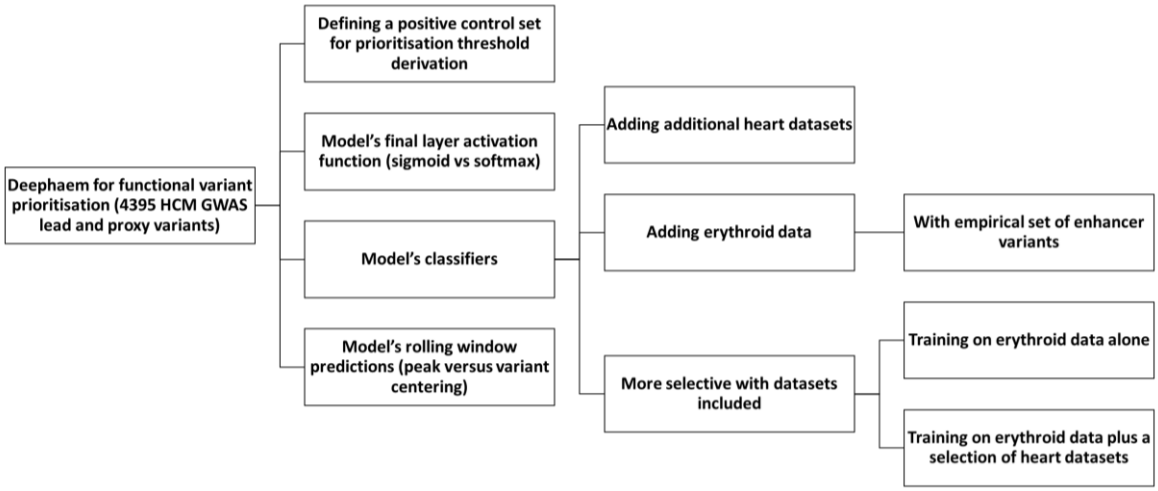


Figure 61 | Process of employing DeepHaem for variant prioritisation | A flowchart illustrating the process of prioritising functional variants (out of the HCM GWAS lead and proxy set) via DeepHaem. Each box represents a part of the troubleshooting process.

After exploring many different routes (summarised in figure 61) with the intention of optimal model performance and a compliant positive control set, the model was still not able to accurately predict the peaks – particularly in the positive control Erythroid datasets. It became clear that DeepHaem could not learn from each of the datasets, specifically the erythroid tracks, suggesting it could only generalise heart datasets which existed in bulk in comparison.

Ultimately, regardless of steps that could have been included for improvement – DeepHaem lacks the advanced architecture required to perform variant prioritisation due to its architectural design which relies on the use of convolutional layers to extract features hierarchically. CNNs, despite their successes in variant annotation, are not the most prevalent models since the introduction of transformer models which employ self-attention mechanisms to capture the global dependencies and relations across different regions.

4.4.2 Deepmind

I therefore decided to employ Google Deepmind's Enformer next (in light of the issues encountered with DeepHaem and also the fact that Enformer uses transformers as opposed to convolutional layers) so that functional variant prioritisation could still ensue. Unlike the approach with DeepHaem, computational validation and visual inspection was not carried out with Enformer as the model was pre-trained and tested by the authors. Likewise, chromatin datasets (like those used for DeepHaem) did not need to be prepared as input to the model due to the authors already having included various datasets (with 5 heart related datasets of interest). Similarly to that encountered with DeepHaem, Enformer provides predictions for a provided region (whereby damage scores are annotated per base). For this reason, the same overarching aim was applied using Enformer – to predict the impact of HCM variants on the underlying chromatin.

Employing model with 5 classifiers of interest

Post-analysis: exploring damage scores

First, the spread of 4395 HCM variants' damage score distribution was assessed per 5 classifiers in figure 62. The figure illustrates how variants were annotated scores as extreme as ~2 and ~-4 in the DNase:cardiac muscle cell 1 classifier and that across the 5 classifiers, the majority of HCM GWAS variants were scattered closely around zero with only a handful being notable outliers (with more extreme damage scores).

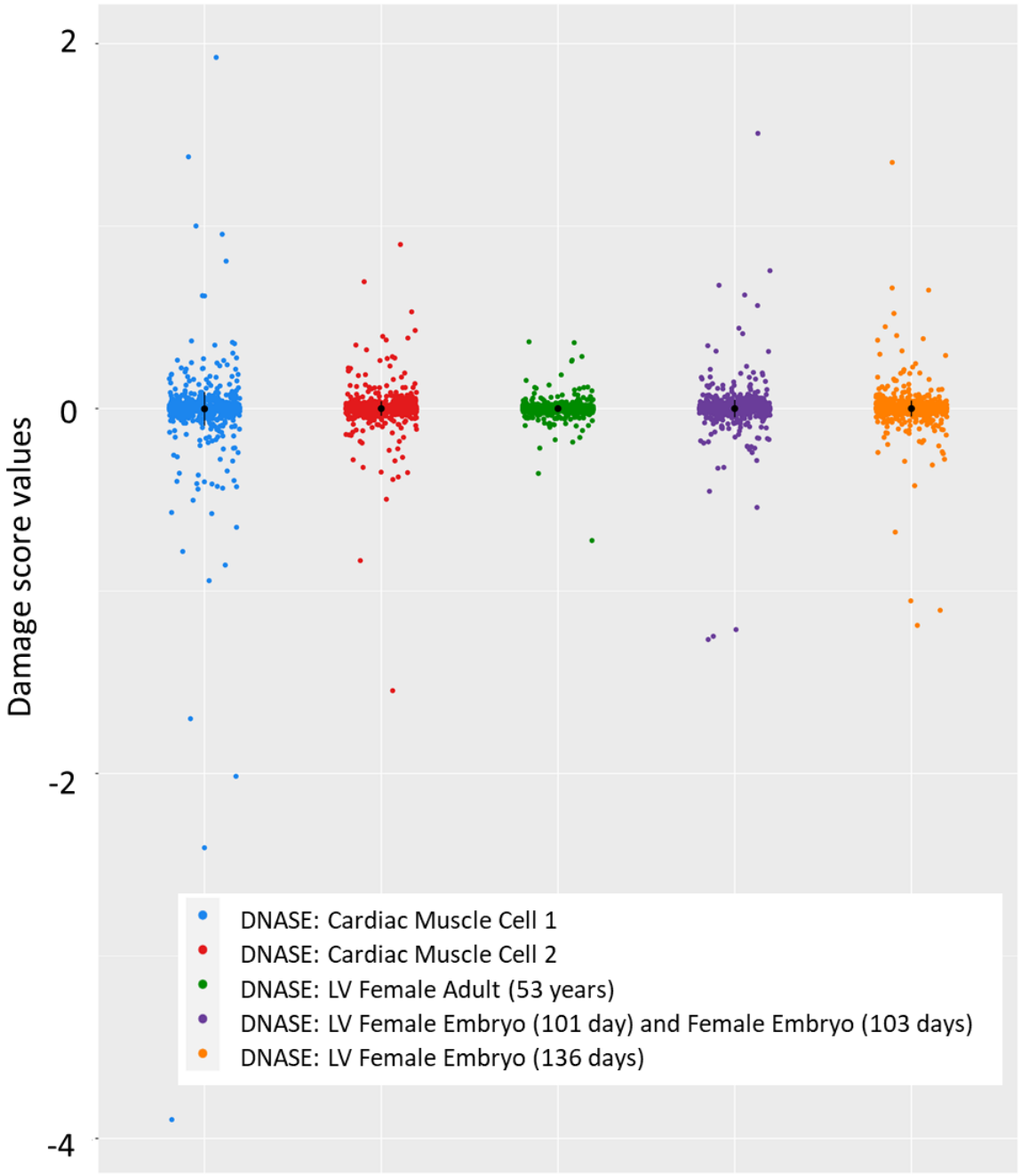


Figure 62 | Deepmind Enformer's damage score distribution per 5 classifiers for HCM variants | Scatter plots representing the distribution of damage scores generated for all 4395 lead and proxy HCM GWAS variants across the 5 classifiers. The y-axis shows the assigned damage scores and the x-axis shows the 5 different colour-coded classifiers/datasets.

Positive control variant set

Similarly to that performed when employing DeepHaem for functional variant prioritisation, a positive control set was chosen and used for damage score visualisation in the relevant classifiers. The erythroid enhancer set used in the DeepHaem analysis could not be used as the positive control set here because Enformer does not contain erythroid classifiers. As a result, an empirical set of gain of function, loss of function, and neutral HUVEC variants were used as input to the model considering the presence of 2 relevant classifiers (DNase_endothelial_cell_1 and DNase_endothelial_cell_2) to check they aligned with their expected distributions. The intention was still to leverage these variants to define a prioritisation threshold that could then be applied to the HCM GWAS variants.

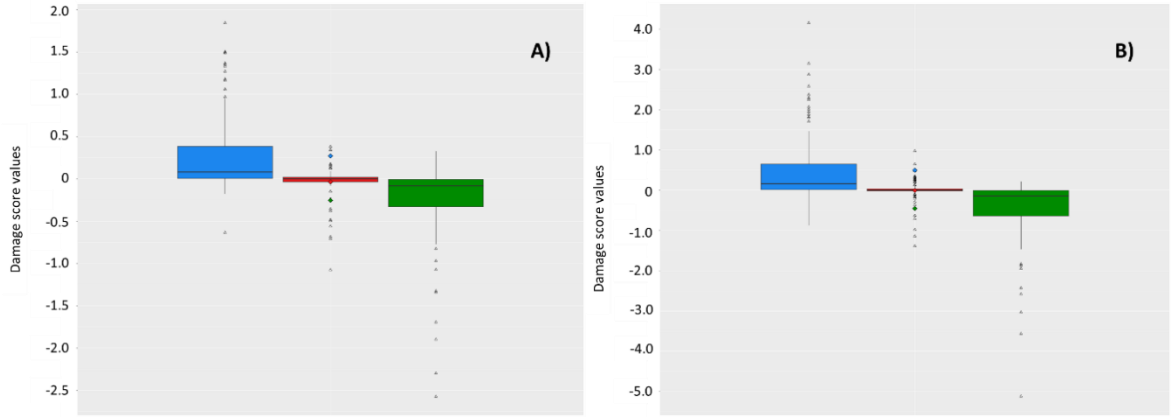


*Figure 63 | Deepmind Enformer's damage score distribution of HUVEC gain, neutral, and loss empirical variants in both HUVEC classifiers | A figure illustrating three boxplots per the three sets of HUVEC positive control variants (138 gain: blue, 113 neutral: red, and 149 loss: green) representing the distribution of damage scores generated respectively for DNase_endothelial_cell_1 classifier (**A**) and DNase_endothelial_cell_2 (**B**). The y-axis shows the assigned damage scores and the x-axis is split by variant set.*

As expected, the gain variants localised to positive damage score values, the neutrals scattered closely around zero, and the losses were in the negatives. The resulting damage scores across these 3 groups remained consistent for both endothelial classifiers (shown in figure 63). This showcased a

good positive control set and paved the way for variant prioritisation using HCM GWAS and proxy variants in the relevant datasets.

Using an extended set of positive control (neutral) variants distribution for prioritisation

Visualising the HUVEC gain, loss, and neutral set of variants' damage scores was reassuring in the sense that they were being annotated as one would expect for variants whose effects had already been experimentally established. This provided confidence in the positive control set and the idea of using an extended set of neutral variants became a desirable approach for defining prioritisation thresholds. This approach was rooted in the understanding that the neutral variants would show a normal distribution – whereby the lower bound could distinguish between loss of function and neutral variants and likewise with the higher bound expect for gain of function variants.

The calculation of p-values from Monte Carlo procedures was used to calculate the prioritisation thresholds for functional gain and loss variants (derived thresholds shown in figure 64). Reasons for using this approach included the understanding that many test statistics do not have a standard asymptotic distribution (and even if there was such a distribution, the ideal sample size may not be present), as well as the fact that calculating the sampling distribution through exhaustive enumeration can be computationally intensive (North *et al.*, 2002).

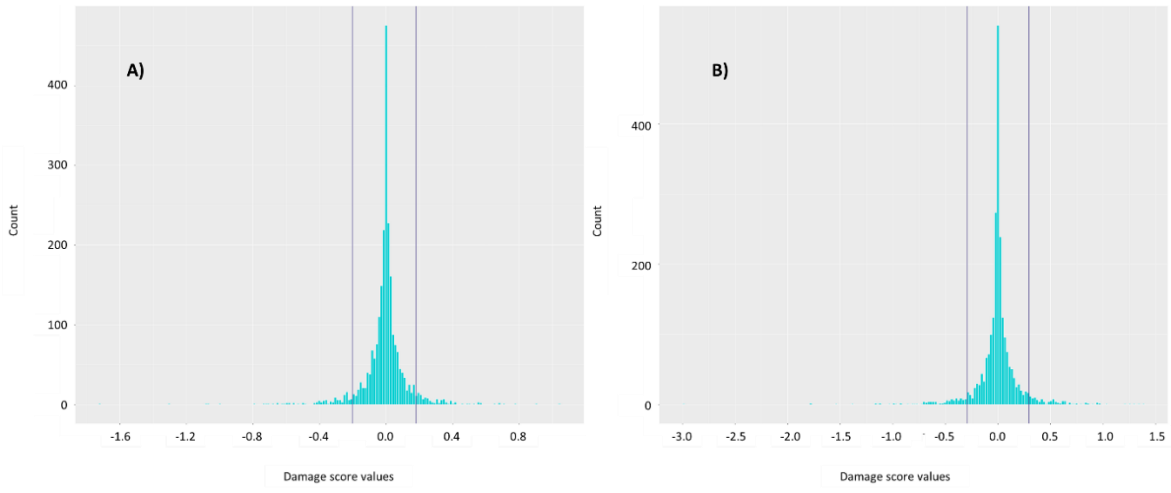


Figure 64 | Deepmind Enformer's damage score distribution of HUVEC extended empirical neutral variants in both HUVEC classifiers | A figure showing a histogram of 2411 neutral HUVEC positive

*control variants whereby the y-axis shows the count of variants per assigned damage score (on the x-axis) for Deepmind Enformer's DNase_endothelial_cell_1 classifier (**A**) and DNase_endothelial_cell_2 (**B**). The navy vertical lines represent thresholds used for prioritisation measures – exact values detailed in table 24.*

Both endothelial datasets showed a normal distribution when plotting the extended set of 2411 HUVEC neutral variants. Given the calculation of empirical p-value (0.05), an upper and lower threshold were defined for each dataset as shown in table 24. These thresholds were then applied to the set of 400 HUVEC variants (comprising 138 gain, 113 neutral, and 149 loss) to assess how many variants were being annotated damage scores higher than the upper threshold or otherwise lower than the lower threshold – these counts were tallied based on their variant type classification.

Table 24 | Deepmind Enformer HUVEC classifiers' derived prioritisation thresholds and associated counts | A table depicting the associated (derived) prioritisation upper and lower thresholds as obtained through the calculation of p-values (p=0.05) from Monte Carlo procedures as well as the number of gain, neutral, and loss variants passing said thresholds.

Classifier	Threshold	Threshold score	#Gain	#Neutral	#Loss
DNase_endothelial_cell_1	Upper	0.179	48	3	1
	Lower	-0.203	1	12	51
DNase_endothelial_cell_2	Upper	0.295	55	7	0
	Lower	-0.293	2	6	63

Enformer findings

Extracting variants annotated with a damage score ≥ 0.295 and ≤ -0.293 (which corresponded to endothelial classifier 2 prioritisation thresholds due to them being more extreme and thus likely more stringent) in any of the 5 heart datasets resulted in 53 unique variants being prioritised out of the initial set of 4395 variants. 24 of which are gain; 29 are loss, 9 are lead variants; 44 are proxy (detailed in figure 65). A more detailed breakdown of the 53 prioritised HCM GWAS variants is

shown in table 25. The associated VEP output for these variants is provided at GitHub (https://github.com/domwest/DPhil-thesis/blob/main/prioritised_variants_VEP_output.xlsx).

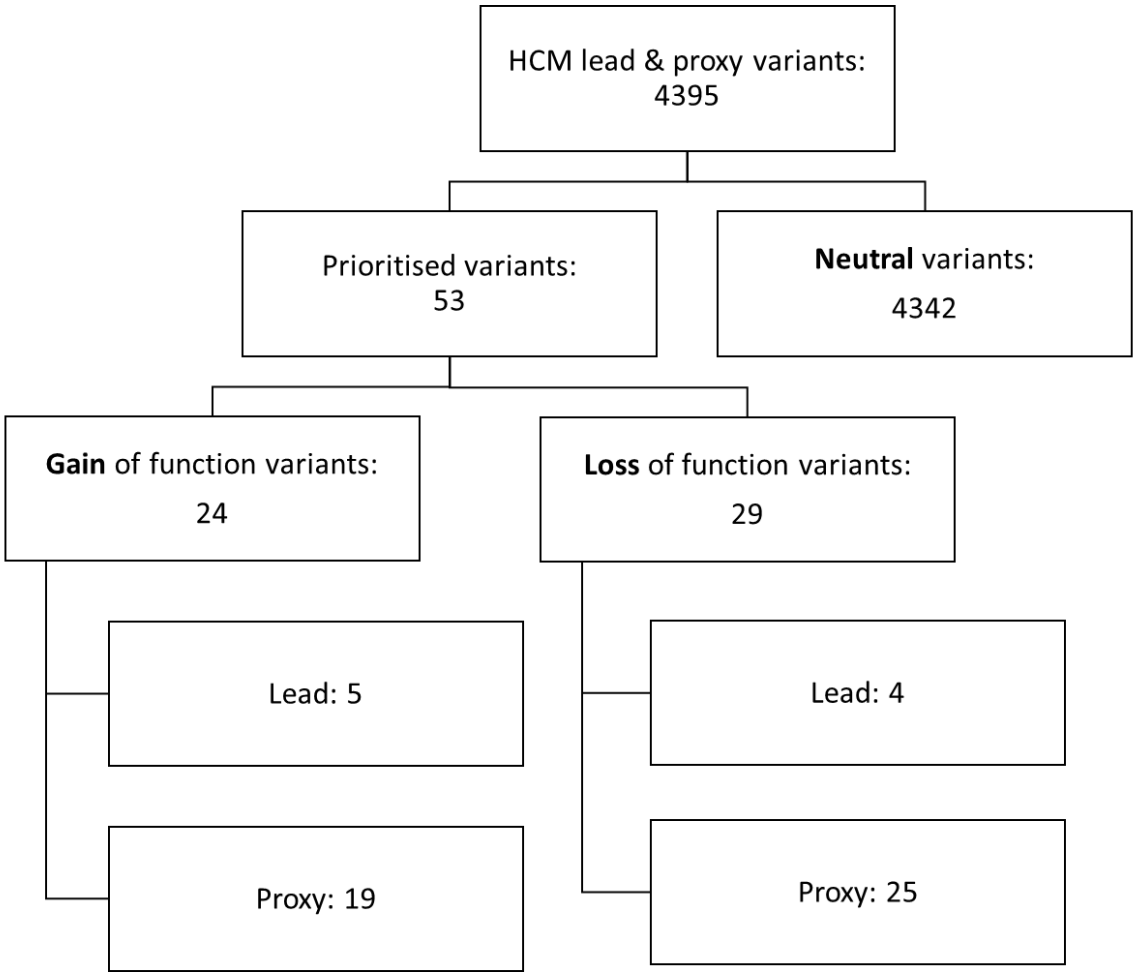


Figure 65 | Overview of Deepmind Enformer's findings regarding HCM variant prioritisation | A flowchart illustrating the breakdown of HCM GWAS variants and how they were scored by Google Deepmind's Enformer – how many were prioritised and of those, how many were gain and loss of function variants as well as lead and proxy.

Table 25 | Detailed look at Deepmind Enformer's findings regarding HCM variant prioritisation | A table showing a more detailed breakdown of the 53 HCM GWAS variants prioritised through Enformer. This includes details about the prioritised variant regarding its rsid, lead rsid, whether it's a lead or proxy variant, the classifiers in which it was prioritised (1: DNase:cardiac muscle cell 1; 2: DNase:cardiac muscle cell 2, 3: DNase:heart left ventricle female embryo (136 days), 4: DNase:heart left ventricle female embryo (101 day) and female embryo (103 days), and 5: DNase:heart left ventricle female adult (53 years)) as well as the corresponding damage scores, whether it is a gain or loss of function variant, and the locus prioritised genes (as assigned by Open Target and FUMA based on the lead variant).

Prioritised rsid	**lead rsid**	**Prioritised SNP details**	**classifier**	**damage scores**	**var type**	**OT**	**FUMA**
rs10872646	rs2184370	6 149746539 A G	[1]	[-0.412]	loss	NUP43	PCMT1
rs945418	rs1048302	1 16018235 T C	[1]	[-0.342]	loss	*HSPB7*	*HSPB7*
rs3734634	rs1159974	6 125790654 G A	[2 1]	[-0.388 -0.428]	loss	HEY2	HEY2
rs149015746	rs6699769	1 77683091 CGCCGTCGCA C	[1 2 3]	[0.619 0.4289 0.2989]	gain	DNAJB4	NEXN
rs3923060	rs11748963	5 139439612 C T	[2 1]	[-0.375 -0.575]	loss	PROB1	SPATA24
rs2297932	rs2184370	6 149718225 C G	[2 1]	[-0.322 -0.441]	loss	NUP43	PCMT1
rs1763603	rs1048302	1 16013314 T G	[1 2]	[0.371 0.322]	gain	*HSPB7*	*HSPB7*
rs2503715	rs2503715	1 2212668 A G	[1 2 3 4 5]	[1.925 0.899 0.676 0.661 0.361]	gain	FAAP20	SKI

rs116904997	rs116904997	12 120230731 G A	[1 2 3 4 5]	[-0.723 -1.188 - 1.248 - 1.546 - 3.897]	loss	PXN	PXN
rs11748963	rs11748963	5 139394348 T C	[1 2 3 4 5]	[-0.355 - 0.542 - 0.650 - 0.677 - 0.833]	loss	PROB1	SPATA24
rs2945236	rs2945236	8 8310872 G A	[1]	[-0.394]	loss	ERI1	ERI1
rs61991243	rs8019721	14 71380400 A G	[1]	[-0.783]	loss	SIPA1L1	SIPA1L1
rs4699910	rs2191446	5 57700364 A T	[2 3 4]	[1.508 1.349 0.695]	gain	ACTBL2	ACTBL2
rs66925847	rs13297385	9 131640601 ACT A	[1 5]	[1.380 0.366]	gain	RAPGEF1	RAPGEF1
rs11838776	rs9515201	13 110388334 G A	[1 3 4]	[1.001 0.400 0.345]	gain	COL4A2	COL4A2
rs11023178	rs12270374	11 14288736 G A	[1]	[0.955]	gain	COPB1	RRAS2
rs1112729	rs2184370	6 149766491 C A	[1]	[0.808]	gain	NUP43	PCMT1
rs566408	rs497999	18 35085797 C G	[4 3]	[0.756 0.649]	gain	MAPRE2	MAPRE2
rs5760061	rs5760054	22 23836092 G A	[4 3]	[0.623 0.521]	gain	VPREB3	SMARCB 1
rs616579	rs816379	6 54105134 G C	[1]	[0.618]	gain	MLIP	MLIP
rs4894803	rs4894803	3 172082466 A G	[4 3]	[0.565 0.449]	gain	FNDC3B	FNDC3B
rs80346216	rs28768976	17 45897110 G T	[2]	[0.530]	gain	PLEKHM1	MAPT

rs9515201	rs9515201	13 110388451 A C	[4 3 2]	[0.441 0.383 0.348]	gain	COL4A2	COL4A2
rs3808802	rs13297385	9 131624672 G A	[4 3]	[0.411 0.375]	gain	RAPGEF1	RAPGEF1
rs6003909	rs5760054	22 23839465 A G	[2]	[0.396]	gain	VPREB3	SMARCB1
rs9442216	rs1048302	1 16026905 T C	[2]	[0.387]	gain	*HSPB7*	*HSPB7*
rs4887092	rs2002854	15 78751333 T C	[2]	[0.376]	gain	ADAMTS7	MORF4L1
rs62061810	rs28768976	17 46101518 G C	[1]	[0.363]	gain	PLEKHM1	MAPT
rs2891537	rs7487962	12 114921147 T G	[1]	[0.356]	gain	TBX3	TBX3
rs3825719	rs8019721	14 71586380 G C	[1]	[0.348]	gain	SIPA1L1	SIPA1L1
rs6566955	rs6566955	18 58255557 A G	[3 4]	[0.316 0.313]	gain	NEDD4L	NEDD4L
rs295114	rs295114	2 200330879 C T	[4]	[0.315]	gain	SPATS2L	SPATS2L
rs3825719	rs8019721	14 71586380 G A	[1]	[0.304]	gain	SIPA1L1	SIPA1L1
rs62064665	rs28768976	17 46004161 T C	[3 4]	[-0.309 - 0.322]	loss	PLEKHM1	MAPT
rs7537908	rs6699769	1 78053255 A G	[4]	[-0.327]	loss	DNAJB4	NEXN
rs62439782	rs2184370	6 149566308 T C	[2 1]	[-0.348 - 0.401]	loss	NUP43	PCMT1
rs76667867	rs28768976	17 45722444 G C	[2]	[-0.351]	loss	PLEKHM1	MAPT
rs10873246	rs8019721	14 71686976 A G	[1]	[-0.354]	loss	SIPA1L1	SIPA1L1
rs11372647	rs2191446	5 57693080 T TA	[1]	[-0.364]	loss	ACTBL2	ACTBL2
rs10744836	rs7487962	12 114933937 T G	[1]	[-0.399]	loss	TBX3	TBX3
rs35006907	rs35006907	8 124847575 C A	[1]	[-0.413]	loss	MTSS1	MTSS1
rs13133886	rs11730129	4 16034172 A C	[3 4]	[-0.423 - 0.453]	loss	PROM1	PROM1
rs6450431	rs2191446	5 57718369 C T	[1]	[-0.426]	loss	ACTBL2	ACTBL2

rs11386443	rs4894803	3 172041620 C CCT	[1]	[-0.436]	loss	FNDC3B	FNDC3B
rs10457327	rs12210733	6 118370989 G C	[2 3 4]	[-0.497 - 1.105 - 1.265]	loss	PLN	SLC35F1
rs7183401	rs8033459	15 84828713 G A	[1]	[-0.502]	loss	ALPK3	NMB
rs7183401	rs8033459	15 84828713 G T	[1]	[-0.569]	loss	ALPK3	NMB
rs10873246	rs8019721	14 71686976 A T	[1]	[-0.857]	loss	SIPA1L1	SIPA1L1
rs888083	rs3845778	2 36889930 C T	[1]	[-0.943]	loss	STRN	STRN
rs476348	rs497999	18 35090057 T C	[3 4]	[-1.054 - 1.211]	loss	MAPRE2	MAPRE2
rs10402263	rs12460541	19 45810500 G C	[1]	[-1.699]	loss	DMPK	SYMPK
rs7148679	rs8019721	14 71693604 T C	[1]	[-2.015]	loss	SIPA1L1	SIPA1L1
rs7148679	rs8019721	14 71693604 T A	[1]	[-2.407]	loss	SIPA1L1	SIPA1L1

Deepmind Enformer has produced a putative causal set of 53 HCM regulatory variants in this analysis. Even though deep learning can lead to the prioritisation of noncoding variants, these approaches often end up prioritising a large set making it difficult to test each of these individually and tricky to further prioritise within this set. Because of this, high-throughput Massively Parallel Reporter Assay (MPRA) experiments could be used to test numerous variants in one single experiment (Mulvey *et al.*, 2021). However, reducing a set of 4395 variants to a final set of 53 prioritised variants is a step in the right direction and somewhat manageable in quantity to look through them case-by-case.

Of the 53 prioritised variants, there is roughly a 50:50 split between those assigned gain and loss of function. One would have expected most of these to be loss of function variants considering that the loss of a process appears to be more common compared to the creation of an entirely new process in biology. Given the fact that there is more gold standard data on loss of function variants, it is easier to infer what might happen if one removes a specific interaction rather than to create a novel interaction. For that reason, gain variants existing in closed areas of chromatin (that potentially improve chromatin accessibility) are difficult to predict. However, Deepmind Enformer helps one

gain insight into closed regions of chromatin where the introduction of a gain of function variant and its impact on HCM can be explored.

4.5 Discussion & Limitations

The biological data used to train both DeepHaem and Enformer is measured from human/mouse tissues and cell types of interest; however, they do not represent trait-specific information with regards to the diseased cell / cell state / developmental timing / age or lifestyle exposure. For example, even though DNase-Seq data in cardiomyocytes would be more informative than that from blood cells, HCM-specific epigenome mapping data is yet to be generated (Hawkins-Hooker *et al.*, 2023). There is also an opportunity for improvement by including more functional datasets, like Massively Parallel Reporter Assays (MPRA) data, to increase the detection of variants (Avsec *et al.*, 2021). MPRA is a high-throughput method which evaluates the regulatory activity of thousands of DNA sequences or variants which can prove helpful with understanding polygenic disease contribution in one single experiment.

Both DeepHaem and Enformer training is also based on 1D chromatin sequences input to the model; however, the genome exists in a three-dimensional space. The 3D conformation could add an additional layer of insight when considered in the model's prediction process. For example, noncoding variants can disrupt enhancer-promoter loop formation leading to altered gene expression. The Akita model (Fudenberg *et al.*, 2020) enables the prediction of interaction between sections of DNA with contact frequency maps in Hi-C and micro-C data – which allows for more personalised assessment of perturbation in the genome and circumstance of interest. Additionally, because deep learning approaches have tailored aims in their modelling, by combining a number of different models one may better be able to understand regulatory mechanisms. For example, one could first use a model with TF binding data to predict the functional effect of a specific noncoding variant. Thereafter, a model predicting the 3D conformation and gene expression could help detect enhancer-promoter interactions and potential target genes (Lan and Corces, 2022).

Although all chromatin data underwent QC and standardisation before input to machine learning, informatic challenges, data quality differences, and methods of integration could have resulted in spurious results despite taking care of all confounders.

Datasets from ENCODE that were used to train Enformer do not appear to have been run through a pipeline like CATCH-UP and so the confidence assigned to called peaks is uncertain. However, in a study benchmarking LanceOtron (which performed the best), ENCODE's processing pipeline (with MACS2 integrated) was one of the top contenders. ENCODE also performs post-processing after peak calling (Hentges *et al.*, 2022). There were only 5 heart datasets included in Enformer – stronger pattern detection may have resulted in the presence of additional heart tissue and cell type datasets.

There appear to be redundant classifiers in Enformer – different datasets capture the same tissue or cell type (from different donors or laboratories). This is potentially problematic; Li *et al.* (2023) deliberate the importance of informative classifier inclusion in prediction performance – questioning the "bigger is better" mindset and suggests the level of information should be considered over the amount of data. Enformer's performance could thus have been enhanced by first re-training with less redundant classifiers, excluding tissue or cell types irrelevant to this analysis, and including more heart cell types of interest. This would have been too timely to perform within the scope of this thesis.

It is possible that using prioritisation thresholds derived from endothelial data rather than heart data may have inflicted some biases. Additionally, all HUVEC neutral variants were extracted from within a peak suggesting bias against gain variants which result in peak creation rather than an increase in accessibility. Because the 1000 Genome SNP set are uniformly distributed genome wide with the majority being null and therefore expected to scatter around zero (like the HUVEC neutrals), using this set of variants could have side-stepped potential biases while also performing GWAS enrichment.

Additionally, a similar fine-tuning approach to that performed for DeepHaem could have been leveraged with Enformer – to train the model on only relevant datasets and in a step-wise manner try

improving the model's architecture and performance. This has been carried out by the Hughes' lab who have recently published their approach which has been shown to outperform Enformer (Riva *et al.*, 2025). Deepmind has now also released an updated model called AlphaGenome (Avsec *et al.*, 2025); both models will become of interest post-thesis to assess and compare results.

5

Linking Prioritised Variants to Target Genes

5.1 Introduction

To assign prioritised noncoding variants to their effector genes and understand their functional significance , insight into the cis-regulatory elements within these regions is essential. A plethora of regulatory elements (shown in figure 27) populate the noncoding genome and together they control gene expression.

5.1.1 Annotating Regulatory Elements

White (2015) alluded to the shortage of technologies capable of validating and annotating regulatory elements at the same rate as their discovery and suggested the need for such technologies to assess the regulatory activity of extensive collections of sequences. Due to the taxing task of directly measuring cis-regulatory element activity, the standard approach is to deduce activity based on chromatin accessibility, histone marks, and TF binding. However, this becomes challenging given the unification of these features among enhancers and promoters. For example, both are marked with H3K4me1 and H3K4me3 albeit in different ratios (Riva *et al.*, 2024).

Numerous bioinformatic tools have been devised with the purpose of classifying cis-regulatory elements. Initial approaches were primarily based on the presence/absence of histone marks. Some approaches employ supervised learning techniques to predict enhancers given histone modifications (like the ensemble predictor DEEP (Kleftogiannis *et al.*, 2015) or the random-forest based algorithm RFECS (Rajagopal *et al.*, 2013)) or P300 binding. However, this approach requires validated enhancers and because only a handful have been corroborated empirically, said approaches can be prejudiced towards a certain set of enhancers (Zacher *et al.*, 2017).

Other approaches employ unsupervised learning algorithms which model the genome as an arrangement of chromatin states (Zacher *et al.*, 2017). The majority of these are based on hidden Markov models (HMMs) or dynamic Bayesian networks. Variation exists across these methods due to the distribution of ChIP-seq signals assigned to each chromatin state.

5.1.2 Variant-to-gene Assignment

Variant annotations help outline which GWAS noncoding variants should be prioritised; however, due to the heterogenous nature of functional genomic datasets, integrative analyses are required for variant characterisation. Extensive projects responsible for primary annotation, such as ENCODE (https://www.encodeproject.org/), Roadmap Epigenomics (http://www.roadmapepigenomics.org), GTEx (https://www.gtexportal.org/home/) and FANTOM5 (https://fantom.gsc.riken.jp/5/), contain copious experimental datasets across multiple tissues, cell types and conditions. Additionally, population-level studies such as UK BioBank, Trans-Omics for Precision Medicine (TOPMed) as well as specific disease studies exist. Despite the availability of functional datasets, the data is often sparse or simply does not exist for the particular tissue, cell type, developmental phase, or condition of interest (Kuksa *et al.*, 2022). In light of the complexity of functional data and the need for its interpretation before reaching biological conclusions, annotation tools and variant databases have been developed to highlight the relevant features.

One of the variant annotation databases is Open Targets Genetics which aims to link a particular target (gene, transcript or protein) to a disease. Additionally, variants in Open Targets are assigned to plausible target gene(s) via the implementation of their Variant-to-Gene (V2G) pipeline (Ghoussaini *et al.*, 2021). The pipeline incorporates V2G evidence from molecular phenotype quantitative trait loci experiments, chromatin interaction experiments, *in silico* functional predictions, as well as the distance between the variant and gene's canonical transcription start site (TSS). This pipeline prioritises and ranks genes on the basis of an overall V2G score.

5.1.3 Enhancer-to-gene Mapping

Fulco *et al.* (2019) developed an experimental technique known as CRISPRi-FlowFISH to investigate over 3500 potential enhancer-gene interactions involving 30 genes. Additionally, they created the Activity-by-Contact (ABC) model to predict enhancers' target genes. In the ABC model, the influence of an element on a gene is determined by its strength as an enhancer (the activity) which is adjusted for the frequency of its 3D interactions with the gene's promoter (the contact). Additionally, the impact of a particular element on gene expression (measured by the reduction in expression following CRISPR inhibition) is described in its ratio with the collective effect of all elements involved.

The perturbation data Fulco *et al.* (2019) collected, which aligned with the ABC model's predictions, demonstrated that enhancers frequently regulate multiple genes, the majority of enhancers with measurable effects are situated within 100kb of their target promoters, and that enhancers are capable of affecting gene expression with varying orders of magnitude and direction. Their findings also highlighted that the activity of an enhancer as well as the frequency of its contact with promoters drive its specificity as opposed to the TFs located at the enhancer and promoter sites.

5.1.4 Peak-to-gene Linkage

ArchR is an R-based pipeline for the processing and analysis of scATAC-Seq data as well as its integration with scRNA-Seq data. This tool facilitates dataset quality control, clustering of cells, peak calling as well as peak co-accessibility and peak-to-gene linkage, motif analysis, and tracking of cell lineages. Consequently, ArchR significantly enhances the understanding of gene regulation at the level of individual cells.

ArchR's peak-to-gene linkage utilises integrated scRNA-Seq and scATAC-Seq data to identify significant relationships between peak accessibility and gene expression. In this way, regulatory elements can be linked to their plausible target gene(s). For each peak2gene link, ArchR assigns a correlation and FDR score. The correlation score represents the strength of the correlation between

chromatin peak accessibility (assessed by scATAC-Seq) and the expression level of an adjacent gene (evaluated through scRNA-Seq). The correlation score thus explains the extent to which the accessibility of a peak can predict the expression of a gene.

5.1.5 Motif Analysis

Sequence-specific TFs modulate the expression of target genes (and thus regulate biological processes) by interacting with regulatory elements. Most TFs can recognise up to hundreds of similar DNA sequences otherwise known as binding site motifs – each with different binding affinities. For this reason, TF binding is illustrated through binding site motifs which show the array of sequences that are preferentially bound (Inukai *et al.*, 2017).

Over one hundred and twenty tools for DNA motif discovery exist (Wei and Yu, 2007). However, identifying DNA motifs presents many challenges due to factors such as genomic complexity, a limited understanding of evolutionary influences on motifs, their diverse sizes, as well as co-occurrence and overlap. Motif discovery tools perform adequately when analysing small genomic regions, such as transcription factor binding sites (TFBS) within ChIP-sequencing data, where the regions are relatively uniform in length and short. However, less so as the regions increase in size as seen for regulatory elements, for example (Castellana *et al.*, 2021).

Heinz *et al.* (2010) developed HOMER which allows one to analyse various genomic positions for enriched motifs. More specifically, HOMER performs *de novo* motif discovery while also evaluating the enrichment of known motifs. HOMER employs an algorithm designed to analyse regulatory elements through the interrogation of DNA without the consideration of proteins. The algorithm exists as a differential tool whereby two sets of sequences are compared to identify elements significantly enriched in one compared to the other. It employs ZOOPS scoring (zero or one occurrence per sequence) and presumes that motif occurrences adhere to a hypergeometric or binomial distribution.

5.1.6 Key objectives

The key objectives for this Chapter are to:

1. Annotate regulatory elements and motifs residing at the prioritised functional HCM variants for insight into the possible mechanisms inflicted and genes affected.
2. Prioritise likely target genes:
    a. Via the Open Targets database
    b. Through enhancer-to-gene mapping with the ABC model
    c. By peak-to-gene linkage via ArchR
3. Explore gene ontology regarding enriched functional terms as well as Kegg pathways given genes prioritised in chapter 3.
4. Inspect 'case studies' i.e. loci whereby prioritised functional HCM variants have been highlighted – integrating all evidence layers obtained throughout the thesis to help single out likely target variants and their gene(s).

5.1.7 Computational approaches employed

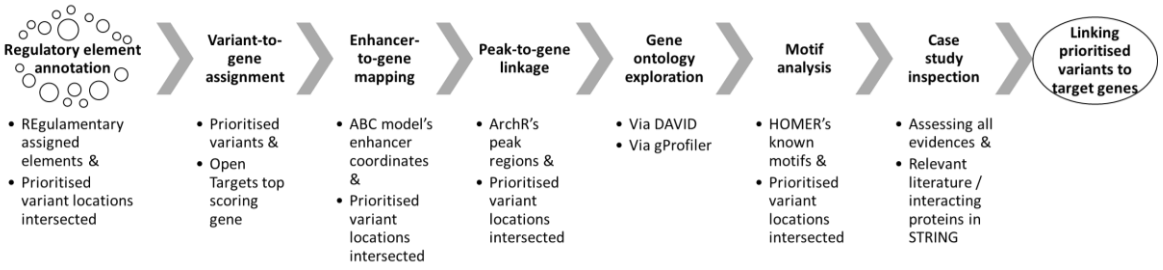


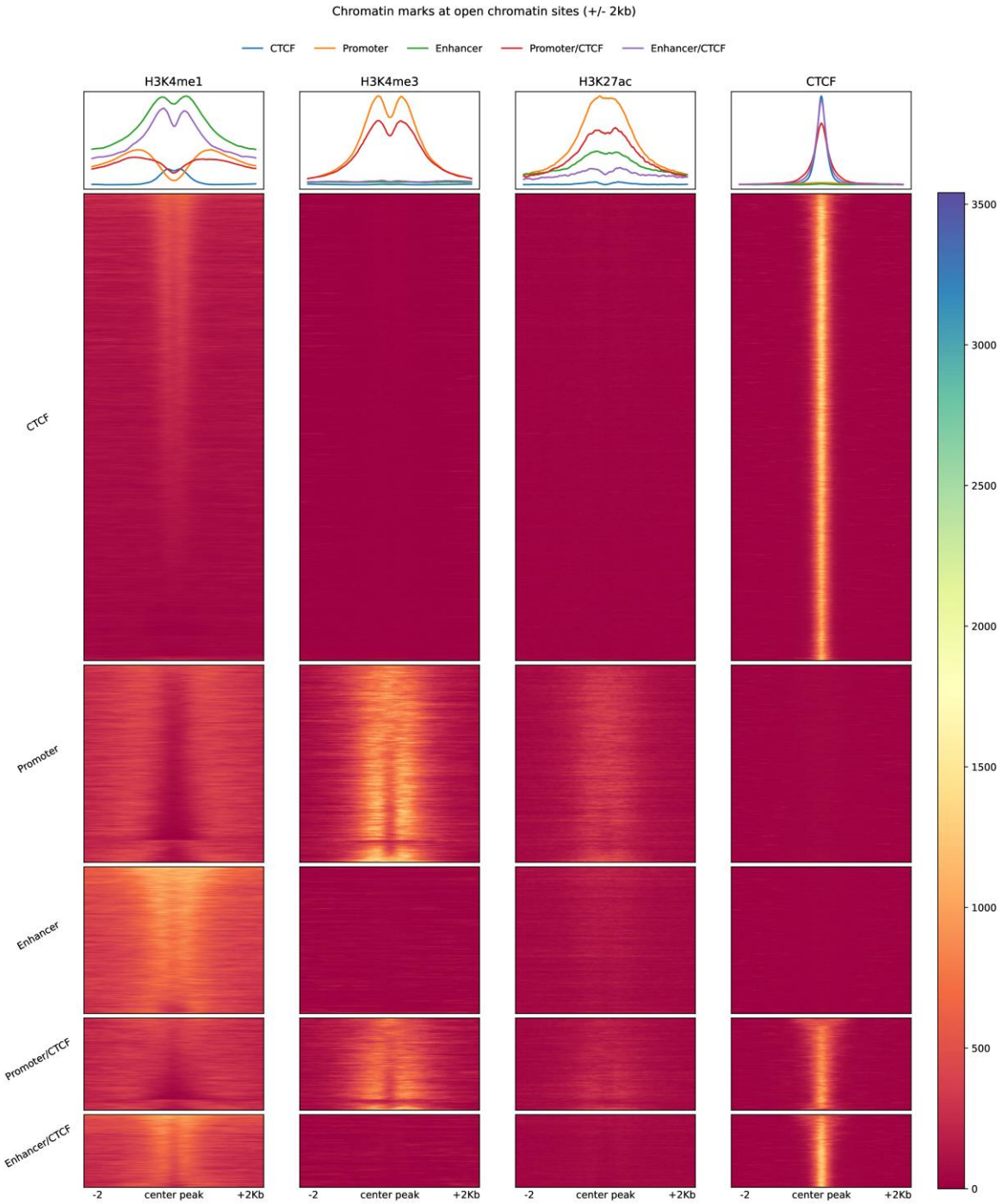
Figure 66 | Linking variants to genes approach | A flowchart illustrating the computational tools implemented in this chapter to link prioritised HCM variants to their target gene(s). REgulamentary was employed to assign regions wherein the prioritised variants resided to regulatory elements; Open Targets was queried to obtain the highest scoring genes given each of the prioritised HCM variants; the Activity-by-Contact (ABC) model was employed to assign enhancer elements and their target gene(s) post-intersection with prioritised HCM variants; ArchR's peak-to-gene linkage was intersected on the prioritised HCM variants to identify target gene(s); DAVID and gProfiler were used for ontology/pathway enrichment given the set of HCM prioritised genes from previous

chapters; the HOMER motif assignment tool was leveraged to assign motifs intersected on the prioritised HCM variants; a case study inspection per locus (containing prioritised HCM variants) was conducted to assign likely target gene(s) given the collated evidences.

Figure 66 summarises the computational approaches employed in this chapter: regulatory elements at each of the prioritised variants were assigned via REgulamentary, target genes per prioritised variant via Open Targets, the ABC model, and ArchR, insight into important ontologies and pathways (given a list of HCM prioritised genes) was obtained via DAVID and gProfiler, known motifs underlying each of the prioritised variants were explored through HOMER to identify potential disruptions and thus likely pathological mechanisms, and finally each of the prioritised variants were explored in a case study inspection to conclude the likely target genes and mechanisms.

5.2 Annotating Regulatory Elements

To explore the likely mechanisms in which the prioritised variants (from the previous chapter) are involved, gaining insight into the regulatory elements existing at each of these regions could aid in this. Knowing whether some of these causal variants likely reside in enhancers or promoters, for example, could point to the candidate target gene. REgulamentary was employed to predict these elements given the prioritised variants and their positions.



F*igure 67 | Assigned regulatory elements (via REgulamentary) to cardiac muscle cell data | A meta-plot showing the coverage ±2kb from the centre of the sorted regions of interest in cardiac muscle cells grouped by regulatory element (promoter, CTCF, enhancer, promoter/CTCF, and enhancer/CTCF) assigned by REgulamentary. The plot depicts the distribution of all peaks across the different cardiac muscle cell chromatin datasets (H3K4me1, H3K4me3, H3K27ac, and CTCF) in order of decreasing coverage from top to bottom.*

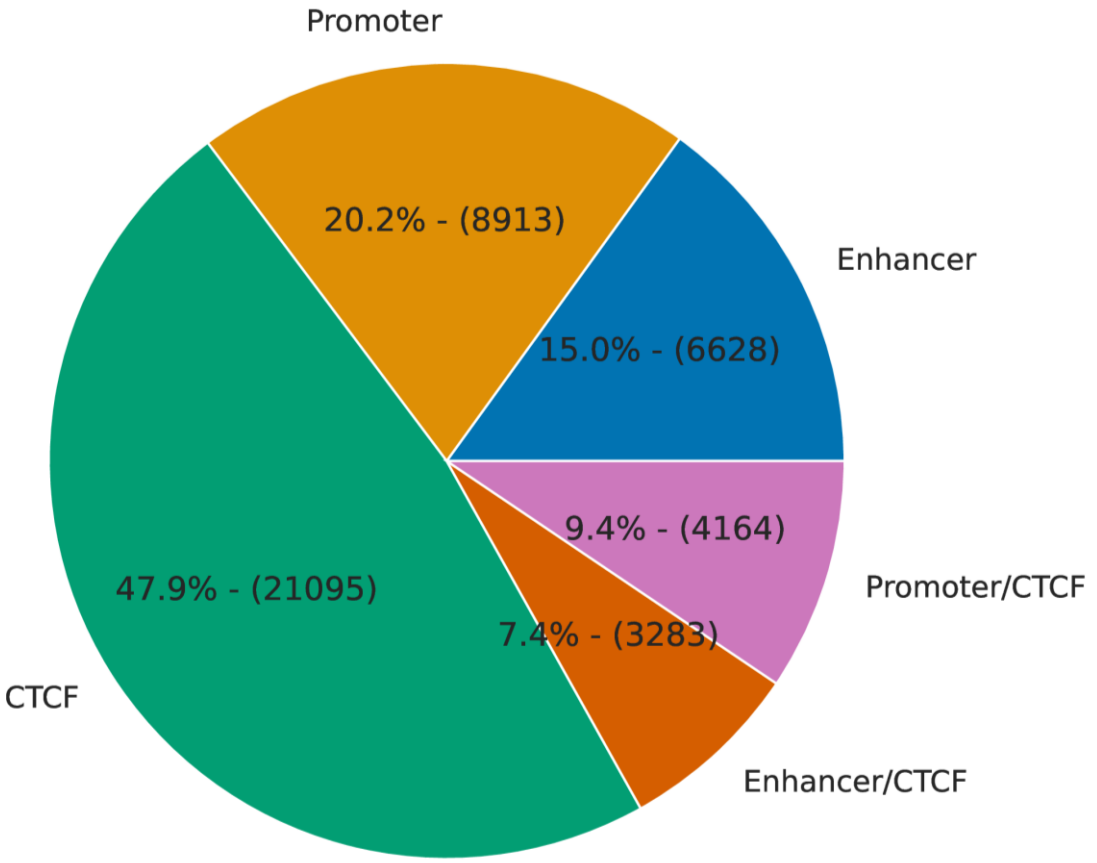


Figure 68 | Frequencies of assigned regulatory elements (via REgulamentary) to cardiac muscle cell data | A pie chart showing the percentage (%) and absolute count of regulatory elements (promoter, CTCF, enhancer, promoter/CTCF, and enhancer/CTCF) in cardiac muscle cell regions of interest.

Both figures 67 and 68 include the following elements: promoters, enhancers, CTCF, promoter/CTCF, and enhancer/CTCF. There are 5 distinct sites rather than just 3 due to the fact that CTCF not only binds at chromatin domain boundaries, but also at enhancers, gene promoters, and inside genes. CTCF's association with enhancer-like sequences and gene promoters helps to establish loops between these elements. Additionally, CTCF binding between enhancers and promoters can inhibit the interaction between an enhancer and its corresponding target promoter(s) (Holwerda and de Laat, 2013).

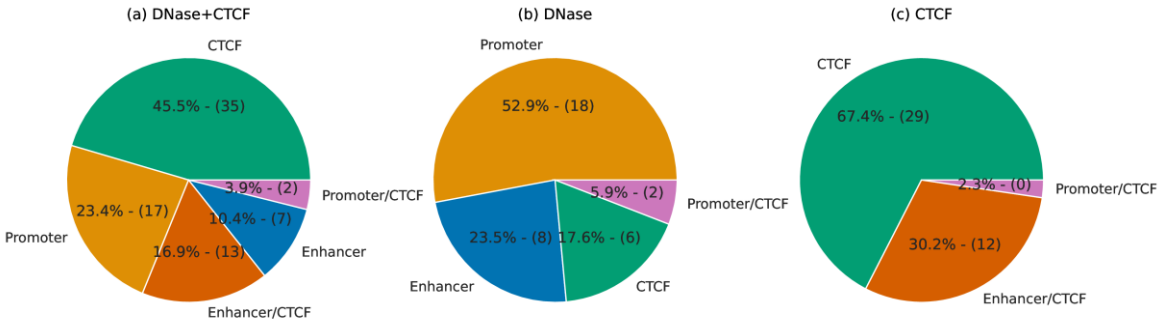


Figure 69 | Frequencies of assigned regulatory elements (via REgulamentary) to cardiac muscle cell data for DNase+CTCF, DNase, and CTCF post-intersection with the 4395 HCM GWAS lead and proxy variants | A pie chart showing the percentage (%) and absolute count of regulatory elements (promoter, CTCF, enhancer, promoter/CTCF, and enhancer/CTCF) in cardiac muscle cell for a). DNase+CTCF, b). DNase only, and c). CTCF only data.

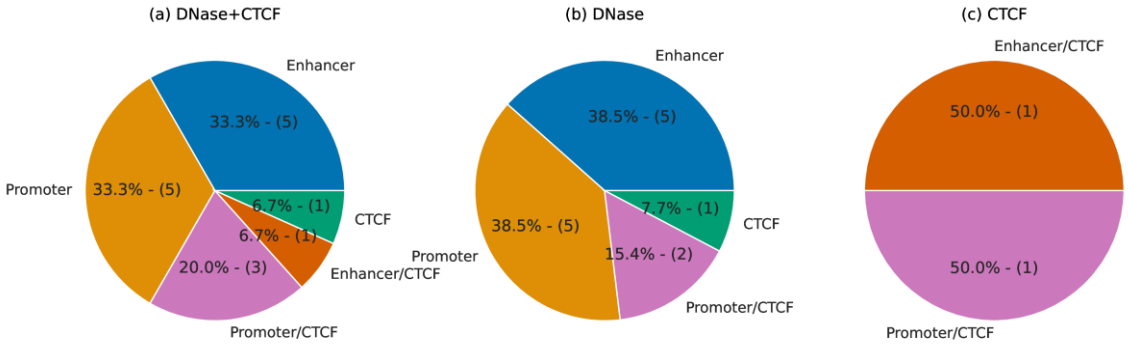


Figure 70 | Frequencies of assigned regulatory elements (via REgulamentary) to cardiac muscle cell data intersected with prioritised HCM variants for DNase+CTCF, DNase, and CTCF post-intersection with the 53 prioritised HCM variants | A pie chart showing the percentage (%) and absolute count of regulatory elements (promoter, CTCF, enhancer, promoter/CTCF, and enhancer/CTCF) in cardiac muscle cell for a). DNase+CTCF, b). DNase only, and c). CTCF only data.

As seen in figures 69 (regarding the entire set of HCM GWAS variants) and 70 (only intersected on the 53 prioritised variants), the REgulamentary pipeline outputs results for ATAC or DNase alone (b), CTCF alone (c), and a merged output of the two (a). This is because bulk ATAC-Seq data has been deemed incapable of correctly assigning CTCF boundaries unlike scATAC-Seq which performs better. Because scATAC-Seq data was used as input (as opposed to bulk tissue data), the

result depicted in figure 69 a). and b). as well as figure 70 a). and b). do not differ significantly and so the DNase+CTCF pie chart alone could be consulted.

4396 GWAS lead and proxy variants were intersected with REgulamentary assigned elements (figure 69) which yielded 73 unique rsids being assigned to a specific regulatory element while the rest were assumed to reside in otherwise closed chromatin. Figure 69 shows that ~46% of the assigned regulatory elements were represented by CTCF sites, ~23% promoters, and ~10% enhancers. Intersecting only the 53 prioritised variants with REgulamentary assigned regions (figure 70) yielded 14 rsids assigned to regulatory elements (1 rsid, rs4887092, has been annotated twice to both a possible enhancer and CTCF site; overall these sites comprise of ~33% enhancers, ~33% promoters, and 7% CTCF).

Table 26 | Regulatory elements assigned to HCM prioritised variants | The assignment of regulatory elements to 15 of the prioritised variants (as output in chapter 4) via REgulamentary. The regulatory element (RE) and corresponding coordinates (RE_coords in b38) are shown alongside the variant's rsid.

prioritised rsid	snp test id	RE	RE coords
rs945418	1 16018235 T C	Enhancer	chr1:16017887:16018441
rs2503715	1 2212668 A G	Enhancer	chr1:2212297:2213426
rs149015746	1 77683091 CGCCGTCGCA C	Promoter/CTCF	chr1:77682146:77683999
rs11386443	3 172041620 C CCT	Promoter	chr3:172041323:172041789
rs11748963	5 139394348 T C	Promoter/CTCF	chr5:139393872:139394647
rs3923060	5 139439612 C T	Promoter	chr5:139439026:139439848
rs11372647	5 57693080 T TA	Enhancer	chr5:57692930:57693589
rs3734634	6 125790654 G A	Promoter	chr6:125789554:125791173
rs62439782	6 149566308 T C	Promoter/CTCF	chr6:149565518:149567113
rs2297932	6 149718225 C G	Promoter	chr6:149717413:149718823
rs10872646	6 149746539 A G	Promoter	chr6:149746291:149746907
rs35006907	8 124847575 C A	Enhancer	chr8:124846918:124848899
rs116904997	12 120230731 G A	Enhancer	chr12:120230124:120231113

rs4887092	15 78751333 T C	Enhancer/CTCF	chr15:78750860:78751825
	15 78751333 T C	CTCF	chr15:78750931:78751463

This tool aided in assigning regulatory regions to 14 of the 53 prioritised variants (shown in table 26). Due to the abundance of TFs that can bind at promoter or enhancer elements in a spatio-temporal manner, integrating promoter- and enhancer-specific TF binding data in the near future may improve the accuracy of assignment and biological significance for cardiac muscle cells (in this context) and also at various differentiation stages (Riva *et al.*, 2024).

5.3 Variant-to-gene Assignment

The Open Targets database enables target gene prediction given a particular variant. At the beginning of the thesis, Open Target was employed given the list of 68 lead HCM GWAS variants to predict a likely target gene per locus; in this analysis, the candidate gene prediction was tailored to the set of prioritised variants.

The 53 prioritised variants corresponded to 49 unique rsids (some variants being multi-allelic); 47 of the 49 rsids were present in the Open Target database. 33 unique target genes were identified for the 47 rsids and their associated V2G scores are shown in table 27.

Table 27 | Open target gene prioritisation given the prioritised HCM variants | A table showing the top genes assigned by Open Targets database to the 47 prioritised HCM variants (rsids). Overall V2G scores are shown for the top gene.

SNP	snp_test_id	Gene	Overall V2G
rs1763603	1_16013314_T_G	*CLCNKA*	0.367
rs2503715	1_2212668_A_G	*FAAP20*	0.173
rs7537908	1_78053255_A_G	*DNAJB4*	0.354
rs9442216	1_16026905_T_C	*CLCNKA*	0.387
rs945418	1_16018235_T_C	***HSPB7***	0.387
rs295114	2_200330879_C_T	*SPATS2L*	0.206

rs888083	2_36889930_C_T	*STRN*	0.179
rs11386443	3_172041620_C_CCT	*PLD1*	0.120
rs4894803	3_172082466_A_G	*FNDC3B*	0.080
rs13133886	4_16034172_A_C	*PROM1*	0.233
rs11372647	5_57693080_T_TA	*ACTBL2*	0.040
rs11748963	5_139394348_T_C	*PROB1*	0.319
rs3923060	5_139439612_C_T	*SPATA24*	0.307
rs4699910	5_57700364_A_T	*ACTBL2*	0.040
rs6450431	5_57718369_C_T	*ACTBL2*	0.033
rs10457327	6_118370989_G_C	*PLN*	0.053
rs10872646	6_149746539_A_G	*NUP43*	0.513
rs1112729	6_149766491_C_A	*PCMT1*	0.407
rs2297932	6_149718225_C_G	*NUP43*	0.374
rs3734634	6_125790654_G_A	*NCOA7*	0.179
rs616579	6_54105134_G_C	*MLIP*	0.307
rs62439782	6_149566308_T_C	*NUP43*	0.239
rs2945236	8_8310872_G_A	*ERI1*	0.207
rs35006907	8_124847575_C_A	*MTSS1*	0.232
rs3808802	9_131624672_G_A	*RAPGEF1*	0.259
rs11023178	11_14288736_G_A	*COPB1*	0.246
rs10744836	12_114933937_T_G	*TBX3*	0.033
rs116904997	12_120230731_G_A	*PXN*	0.173
rs2891537	12_114921147_T_G	*TBX3*	0.033
rs11838776	13_110388334_G_A	*COL4A2*	0.100
rs9515201	13_110388451_A_C	*COL4A2*	0.120
rs10873246	14_71686976_A_G	*SIPA1L1*	0.280
rs10873246	14_71686976_A_T	*SIPA1L1*	0.280
rs3825719	14_71586380_G_C	*SIPA1L1*	0.300
rs3825719	14_71586380_G_A	*SIPA1L1*	0.300
rs61991243	14_71380400_A_G	*SIPA1L1*	0.300

rs7148679	14_71693604_T_C	*SIPA1L1*	0.280
rs7148679	14_71693604_T_A	*SIPA1L1*	0.280
rs4887092	15_78751333_T_C	*ADAMTS7*	0.313
rs7183401	15_84828713_G_A	***ALPK3***	0.280
rs7183401	15_84828713_G_T	***ALPK3***	0.280
rs62061810	17_46101518_G_C	*KANSL1*	0.440
rs62064665	17_46004161_T_C	*MAPT*	0.454
rs76667867	17_45722444_G_C	*MAPT*	0.380
rs80346216	17_45897110_G_T	*MAPT*	0.421
rs476348	18_35090057_T_C	*MAPRE2*	0.179
rs566408	18_35085797_C_G	*MAPRE2*	0.200
rs6566955	18_58255557_A_G	*NEDD4L*	0.060
rs10402263	19_45810500_G_C	*DMPK*	0.360
rs5760061	22_23836092_G_A	*VPREB3*	0.334
rs6003909	22_23839465_A_G	*VPREB3*	0.280

Table 27 highlights *ALPK3* and *HSPB7* – genes which have an established association with cardiomyopathy.

5.4 Enhancer-to-gene Mapping

The ABC model was developed with the intention of detecting enhancer sites (also linked to likely target genes). For the same reason that REgulamentary was employed to unpack which elements and mechanisms the prioritised variants may affect, this model was employed – homing in on enhancer elements specifically.

Among the 53 prioritised variants, 12 were found to overlap with an enhancer site. All 14 enhancer-gene(s) connections are shown in table 28 (with an ABC score ≥ 0.027 thus passing this significance threshold).

Table 28 | ABC model enhancer-to-gene mapping intersected with prioritised HCM variants | A table showing the enhancer-gene links intersecting with any of the 53 prioritised variants. The enhancer coordinates are shown for each overlapping variant, as well as the associated target gene(s) and genic region. An ABC score is supplied for each enhancer-gene link. Only associations passing the chosen significance score threshold of 0.027 are shown.

SNPS	ABC_enhancer_coords	TargetGene	Class	ABC.Score
rs149015746	1_77682558_77683557	*ZZZ3*	promoter	1.000
rs945418	1_16018050_16018250	*HSPB7*	promoter	1.000
rs945418	1_16018050_16018250	*CLCNKA*	promoter	0.266
rs1763603	1_16012834_16013472	*HSPB7*	intergenic	0.221
rs2503715	1_2212521_2213314	*SKI*	intergenic	0.190
rs149015746	1_77682558_77683557	*PIGK*	promoter	0.127
rs149015746	1_77682558_77683557	*USP33*	promoter	0.119
rs149015746	1_77682558_77683557	*MIGA1*	promoter	0.107
rs1763603	1_16012834_16013472	*CLCNKA*	intergenic	0.102
rs1763603	1_16012834_16013472	*C1orf64*	intergenic	0.090
rs1763603	1_16012834_16013472	*CLCNKB*	intergenic	0.079
rs2503715	1_2212521_2213314	*FAAP20*	intergenic	0.077
rs2503715	1_2212521_2213314	*TNFRSF14*	intergenic	0.066
rs2503715	1_2212521_2213314	*PEX10*	intergenic	0.049
rs2503715	1_2212521_2213314	*PRKCZ*	intergenic	0.049
rs2503715	1_2212521_2213314	*RER1*	intergenic	0.048
rs2503715	1_2212521_2213314	*MORN1*	intergenic	0.048
rs2503715	1_2212521_2213314	*HES5*	intergenic	0.042
rs2503715	1_2212521_2213314	*MMEL1*	intergenic	0.037
rs2503715	1_2212521_2213314	*GABRD*	intergenic	0.036
rs2503715	1_2212521_2213314	*PANK4*	intergenic	0.030
rs2503715	1_2212521_2213314	*FAM213B*	intergenic	0.029
rs3923060	5_139439145_139439677	*DNAJC18*	promoter	1.000
rs11748963	5_139393946_139394433	*MZB1*	genic	0.058

rs11748963	5_139393946_139394433	*PROB1*	genic	0.030
rs10872646	6_149746472_149746684	*NUP43*	promoter	1.000
rs2297932	6_149717596_149718737	*LATS1*	promoter	1.000
rs3734634	6_125790543_125791015	*NCOA7*	promoter	1.000
rs10872646	6_149746472_149746684	*PCMT1*	promoter	0.190
rs2297932	6_149717596_149718737	*KATNA1*	promoter	0.141
rs2297932	6_149717596_149718737	*NUP43*	promoter	0.139
rs2297932	6_149717596_149718737	*PCMT1*	promoter	0.137
rs2945236	8_8310560_8310984	*PRAG1*	intergenic	0.044
rs116904997	12_120230621_120230821	*PXN*	genic	0.090
rs116904997	12_120230621_120230821	*RPLP0*	genic	0.030
rs61991243	14_71380382_71380582	*SIPA1L1*	intergenic	0.037

5.5 Peak-to-gene Linkage

Given the fact that the prioritised noncoding variants exist in the regulatory space in addition to the aim of understanding which genes they possibly target, the ArchR peak-to-gene linkage analysis was employed – whereby peaks are regions of open chromatin which the software maps to likely target genes (given both chromatin and expression data). The ArchR pipeline involves numerous data processing steps prior to peak-to-gene linkage.

Table 29 shows peak-to-gene links output by ArchR within the ventricular cardiomyocyte cell type specifically. Some prioritised variants intersected with a peak which was assigned to more than one gene. Additionally, some peak-to-gene links showed a negative correlation score. In ArchR, a peak to gene negative correlation score indicates that when the accessibility of a chromatin peak increases, the expression of the associated gene tends to decrease. The ArchR manual recommends using a correlation cut-off of 0.45. Only 3 rsids intersected with a peak where the link scored above this significance threshold and these are highlighted green in table 29. These were for rs1763603 which linked to both *HSPB7* and *CLCNKA*, rs4699910 with *ACTBL2*, and rs945418 also with *HSPB7*.

Table 29 | ArchR peak-to-gene linkage intersected with prioritised HCM variants | A table showing the intersection of peak-to-gene links for the 53 prioritised variants. The correlation as well as FDR and -log10FDR are shown per link; those with a correlation ≥ 0.45 and FDR ≤ 0.05 are shaded green.

Prioritised_rsid	PeakName	GeneName	Correlation	FDR	-log10FDR
rs1763603	chr1_16012832_16013332	*CLCNKA*	0.524	0.0	33.53
rs945418	chr1_16017870_16018370	*HSPB7*	0.505	0.0	30.68
rs1763603	chr1_16012832_16013332	*HSPB7*	0.496	0.0	29.4
rs4699910	chr5_57700005_57700505	*ACTBL2*	0.473	0.0	26.34
rs945418	chr1_16017870_16018370	*CLCNKA*	0.449	0.0	23.43
rs116904997	chr12_120230472_120230972	*MSI1*	0.423	0.0	20.5
rs5760061	chr22_23835687_23836187	*VPREB3*	0.417	0.0	19.93
rs4699910	chr5_57700005_57700505	*LINC02225*	0.413	0.0	19.51
rs2503715	chr1_2212593_2213093	*PRKCZ*	0.38	0.0	16.2
rs9442216	chr1_16026699_16027199	*CLCNKA*	0.365	0.0	14.87
rs4887092	chr15_78750921_78751421	*MORF4L1*	0.36	0.0	14.44
rs11748963	chr5_139394081_139394581	*UBE2D2*	0.329	0.0	11.93
rs7183401	chr15_84828318_84828818	*ZSCAN2*	0.324	0.0	11.59
rs11748963	chr5_139394081_139394581	*MZB1*	0.322	0.0	11.44
rs6003909	chr22_23839075_23839575	*MIF-AS1*	0.314	0.0	10.79
rs6003909	chr22_23839075_23839575	*DERL3*	0.301	0.0	9.93
rs6003909	chr22_23839075_23839575	*SLC2A11*	0.295	0.0	9.47
rs3923060	chr5_139439306_139439806	*SNHG4*	0.274	0.0	8.14
rs4894803	chr3_172082269_172082769	*FNDC3B*	0.272	0.0	8.02
rs6003909	chr22_23839075_23839575	*RGL4*	0.271	0.0	7.96
rs116904997	chr12_120230472_120230972	*GATC*	0.269	0.0	7.85
rs6003909	chr22_23839075_23839575	*DRICH1*	0.263	0.0	7.5
rs11748963	chr5_139394081_139394581	*SLC23A1*	0.254	0.0	7.0
rs5760061	chr22_23835687_23836187	*DDTL*	0.253	0.0	6.91
rs6003909	chr22_23839075_23839575	*ZNF70*	0.251	0.0	6.81

rs5760061	chr22_23835687_23836187	*CHCHD10*	0.25	0.0	6.76
rs9442216	chr1_16026699_16027199	*HSPB7*	0.25	0.0	6.79
rs566408	chr18_35085548_35086048	*ZNF397*	0.247	0.0	6.63
rs62064665	chr17_46003888_46004388	*SPPL2C*	0.242	0.0	6.32
rs2503715	chr1_2212593_2213093	*PLCH2*	0.234	0.0	5.94
rs35006907	chr8_124847479_124847979	*SQLE*	0.23	0.0	5.74
rs1763603	chr1_16012832_16013332	*UQCRHL*	0.223	0.0	5.37
rs3808802	chr9_131624641_131625141	*RAPGEF1*	0.22	0.0	5.22
rs11372647	chr5_57693006_57693506	*LINC02225*	0.219	0.0	5.18
rs4887092	chr15_78750921_78751421	*ADAMTS7*	0.219	0.0	5.17
rs62064665	chr17_46003888_46004388	*CRHR1*	0.218	0.0	5.16
rs2503715	chr1_2212593_2213093	*PEX10*	0.212	0.0	4.87
rs6003909	chr22_23839075_23839575	*MMP11*	0.203	0.0	4.46
rs3808802	chr9_131624641_131625141	*POMT1*	0.199	0.0	4.3
rs1763603	chr1_16012832_16013332	*SPEN*	0.189	0.0	3.88
rs9442216	chr1_16026699_16027199	*SPEN*	0.187	0.0	3.8
rs62064665	chr17_46003888_46004388	*KANSL1*	0.182	0.0	3.6
rs3808802	chr9_131624641_131625141	*UCK1*	0.181	0.0	3.58
rs35006907	chr8_124847479_124847979	*WASHC5*	0.18	0.0	3.54
rs11748963	chr5_139394081_139394581	*DNAJC18*	0.173	0.001	3.27
rs566408	chr18_35085548_35086048	*MAPRE2*	0.17	0.001	3.18
rs476348	chr18_35089861_35090361	*ZNF397*	0.165	0.001	3.0
rs80346216	chr17_45896834_45897334	*CRHR1*	0.16	0.002	2.81
rs6003909	chr22_23839075_23839575	*CABIN1*	0.156	0.002	2.7
rs5760061	chr22_23835687_23836187	*ZNF70*	0.155	0.002	2.64
rs11748963	chr5_139394081_139394581	*PROB1*	0.149	0.003	2.46
rs80346216	chr17_45896834_45897334	*SPPL2C*	0.148	0.004	2.42
rs10873246	chr14_71686580_71687080	*RGS6*	0.146	0.004	2.37
rs945418	chr1_16017870_16018370	*SPEN*	0.144	0.005	2.3
rs1763603	chr1_16012832_16013332	*ZBTB17*	0.139	0.007	2.17

rs4894803	chr3_172082269_172082769	*TMEM212-AS1*	0.137	0.008	2.1
rs10872646	chr6_149746333_149746833	*RAET1E*	0.135	0.009	2.06
rs116904997	chr12_120230472_120230972	*PXN*	0.128	0.014	1.86
rs11372647	chr5_57693006_57693506	*ACTBL2*	0.127	0.014	1.84
rs2503715	chr1_2212593_2213093	*SKI*	0.122	0.02	1.69
rs945418	chr1_16017870_16018370	*UQCRHL*	0.122	0.02	1.7
rs76667867	chr17_45722120_45722620	*SPPL2C*	0.121	0.022	1.66
rs1763603	chr1_16012832_16013332	*ARHGEF19*	0.114	0.032	1.49
rs11838776	chr13_110388060_110388560	*RAB20*	0.111	0.037	1.44
rs9515201	chr13_110388060_110388560	*RAB20*	0.111	0.037	1.44
rs11748963	chr5_139394081_139394581	*SNHG4*	0.109	0.041	1.39
rs476348	chr18_35089861_35090361	*MAPRE2*	0.108	0.043	1.37
rs9442216	chr1_16026699_16027199	*UQCRHL*	0.107	0.046	1.34
rs10872646	chr6_149746333_149746833	*PPIL4*	0.096	0.077	1.12
rs9442216	chr1_16026699_16027199	*ZBTB17*	0.091	0.097	1.01
rs945418	chr1_16017870_16018370	*ARHGEF19*	0.085	0.125	0.9
rs66925847	chr9_131640395_131640895	*POMT1*	0.083	0.137	0.86
rs6003909	chr22_23839075_23839575	*CHCHD10*	0.081	0.148	0.83
rs3923060	chr5_139439306_139439806	*MZB1*	0.078	0.165	0.78
rs1763603	chr1_16012832_16013332	*CLCNKB*	0.076	0.175	0.76
rs945418	chr1_16017870_16018370	*ZBTB17*	0.075	0.184	0.73
rs76667867	chr17_45722120_45722620	*CRHR1*	0.074	0.189	0.72
rs9442216	chr1_16026699_16027199	*ARHGEF19*	0.07	0.217	0.66
rs3923060	chr5_139439306_139439806	*UBE2D2*	0.07	0.223	0.65
rs4887092	chr15_78750921_78751421	*CTSH*	0.069	0.224	0.65
rs62064665	chr17_46003888_46004388	*KANSL1-AS1*	0.064	0.265	0.58
rs116904997	chr12_120230472_120230972	*SRSF9*	0.064	0.273	0.56
rs7183401	chr15_84828318_84828818	*ALPK3*	0.06	0.301	0.52
rs9442216	chr1_16026699_16027199	*CLCNKB*	0.055	0.352	0.45
rs3923060	chr5_139439306_139439806	*DNAJC18*	0.053	0.371	0.43

rs10872646	chr6_149746333_149746833	*RAET1E-AS1*	0.05	0.403	0.4
rs10872646	chr6_149746333_149746833	*KATNA1*	0.045	0.459	0.34
rs11386443	chr3_172041480_172041980	*TMEM212-AS1*	0.04	0.521	0.28
rs4894803	chr3_172082269_172082769	*TMEM212*	0.033	0.603	0.22
rs945418	chr1_16017870_16018370	*CLCNKB*	0.032	0.62	0.21
rs4887092	chr15_78750921_78751421	*CHRNB4*	0.031	0.625	0.2
rs76667867	chr17_45722120_45722620	*MAPT*	0.031	0.627	0.2
rs62439782	chr6_149566038_149566538	*PPIL4*	0.027	0.678	0.17
rs11386443	chr3_172041480_172041980	*PLD1*	0.027	0.682	0.17
rs4887092	chr15_78750921_78751421	*HYKK*	0.023	0.727	0.14
rs3734634	chr6_125790622_125791122	*HEY2*	0.017	0.808	0.09
rs10872646	chr6_149746333_149746833	*LRP11*	0.016	0.815	0.09
rs7183401	chr15_84828318_84828818	*NMB*	0.015	0.821	0.09
rs116904997	chr12_120230472_120230972	*SIRT4*	0.014	0.833	0.08
rs80346216	chr17_45896834_45897334	*MAPT-AS1*	0.012	0.865	0.06
rs3923060	chr5_139439306_139439806	*PROB1*	0.011	0.88	0.06
rs80346216	chr17_45896834_45897334	*MAPT*	0.008	0.915	0.04
rs5760061	chr22_23835687_23836187	*SMARCB1*	0.007	0.924	0.03
rs11386443	chr3_172041480_172041980	*TMEM212*	0.006	0.929	0.03
rs3923060	chr5_139439306_139439806	*SLC23A1*	0.003	0.969	0.01
rs11748963	chr5_139394081_139394581	*MATR3*	-0.001	0.993	0.0
rs116904997	chr12_120230472_120230972	*DYNLL1*	-0.002	0.98	0.01
rs35006907	chr8_124847479_124847979	*LINC00964*	-0.01	0.892	0.05
rs76667867	chr17_45722120_45722620	*MAPT-AS1*	-0.011	0.877	0.06
rs3923060	chr5_139439306_139439806	*SPATA24*	-0.012	0.867	0.06
rs62064665	chr17_46003888_46004388	*MAPT*	-0.013	0.85	0.07
rs2945236	chr8_8310415_8310915	*PRAG1*	-0.015	0.83	0.08
rs6003909	chr22_23839075_23839575	*DDT*	-0.02	0.759	0.12
rs62439782	chr6_149566038_149566538	*PCMT1*	-0.025	0.696	0.16
rs62064665	chr17_46003888_46004388	*MAPT-AS1*	-0.028	0.664	0.18

rs11386443	chr3_172041480_172041980	*FNDC3B*	-0.033	0.608	0.22
rs3923060	chr5_139439306_139439806	*CXXC5*	-0.036	0.569	0.24
rs35006907	chr8_124847479_124847979	*NSMCE2*	-0.04	0.521	0.28
rs66925847	chr9_131640395_131640895	*RAPGEF1*	-0.044	0.467	0.33
rs62439782	chr6_149566038_149566538	*KATNA1*	-0.047	0.433	0.36
rs3734634	chr6_125790622_125791122	*TRMT11*	-0.048	0.427	0.37
rs10872646	chr6_149746333_149746833	*GINM1*	-0.049	0.42	0.38
rs7183401	chr15_84828318_84828818	*ZNF592*	-0.05	0.403	0.39
rs7183401	chr15_84828318_84828818	*SEC11A*	-0.05	0.407	0.39
rs2503715	chr1_2212593_2213093	*RER1*	-0.063	0.279	0.55
rs9442216	chr1_16026699_16027199	*FAM131C*	-0.063	0.279	0.55
rs11748963	chr5_139394081_139394581	*PAIP2*	-0.063	0.281	0.55
rs10872646	chr6_149746333_149746833	*RAET1G*	-0.067	0.243	0.61
rs10872646	chr6_149746333_149746833	*ULBP1*	-0.067	0.245	0.61
rs566408	chr18_35085548_35086048	*ZSCAN30*	-0.074	0.193	0.71
rs11748963	chr5_139394081_139394581	*SIL1*	-0.076	0.179	0.75
rs476348	chr18_35089861_35090361	*ZSCAN30*	-0.078	0.167	0.78
rs9442216	chr1_16026699_16027199	*SRARP*	-0.079	0.158	0.8
rs4887092	chr15_78750921_78751421	*PSMA4*	-0.081	0.145	0.84
rs2503715	chr1_2212593_2213093	*MORN1*	-0.081	0.147	0.83
rs3923060	chr5_139439306_139439806	*SIL1*	-0.084	0.13	0.88
rs116904997	chr12_120230472_120230972	*TRIAP1*	-0.086	0.121	0.92
rs10872646	chr6_149746333_149746833	*PCMT1*	-0.087	0.114	0.94
rs62439782	chr6_149566038_149566538	*GINM1*	-0.087	0.115	0.94
rs6566955	chr18_58255147_58255647	*NEDD4L*	-0.087	0.115	0.94
rs6003909	chr22_23839075_23839575	*DDTL*	-0.094	0.085	1.07
rs9442216	chr1_16026699_16027199	*EPHA2*	-0.107	0.045	1.35
rs116904997	chr12_120230472_120230972	*GCN1*	-0.111	0.036	1.44
rs6003909	chr22_23839075_23839575	*SMARCB1*	-0.113	0.033	1.49
rs116904997	chr12_120230472_120230972	*PXN-AS1*	-0.119	0.024	1.63

rs66925847	chr9_131640395_131640895	*UCK1*	-0.121	0.021	1.68
rs76667867	chr17_45722120_45722620	*PLEKHM1*	-0.121	0.021	1.68
rs80346216	chr17_45896834_45897334	*STH*	-0.123	0.019	1.72
rs6003909	chr22_23839075_23839575	*MIF*	-0.128	0.014	1.85
rs5760061	chr22_23835687_23836187	*CABIN1*	-0.132	0.011	1.97
rs1763603	chr1_16012832_16013332	*FAM131C*	-0.135	0.009	2.04
rs11838776	chr13_110388060_110388560	*COL4A1*	-0.141	0.006	2.22
rs9515201	chr13_110388060_110388560	*COL4A1*	-0.141	0.006	2.22
rs2503715	chr1_2212593_2213093	*FAAP20*	-0.143	0.005	2.27
rs10872646	chr6_149746333_149746833	*ULBP2*	-0.145	0.004	2.35
rs5760061	chr22_23835687_23836187	*MIF-AS1*	-0.145	0.005	2.34
rs3923060	chr5_139439306_139439806	*MATR3*	-0.157	0.002	2.71
rs1763603	chr1_16012832_16013332	*SRARP*	-0.159	0.002	2.79
rs62064665	chr17_46003888_46004388	*STH*	-0.159	0.002	2.78
rs945418	chr1_16017870_16018370	*FAM131C*	-0.161	0.001	2.85
rs116904997	chr12_120230472_120230972	*PLA2G1B*	-0.168	0.001	3.1
rs9442216	chr1_16026699_16027199	*CPLANE2*	-0.174	0.0	3.32
rs5760061	chr22_23835687_23836187	*DDT*	-0.174	0.001	3.29
rs1763603	chr1_16012832_16013332	*CPLANE2*	-0.183	0.0	3.63
rs35006907	chr8_124847479_124847979	*ZNF572*	-0.187	0.0	3.79
rs7183401	chr15_84828318_84828818	*SLC28A1*	-0.194	0.0	4.07
rs945418	chr1_16017870_16018370	*SRARP*	-0.2	0.0	4.34
rs3923060	chr5_139439306_139439806	*PAIP2*	-0.201	0.0	4.39
rs11838776	chr13_110388060_110388560	*COL4A2-AS2*	-0.203	0.0	4.46
rs9515201	chr13_110388060_110388560	*COL4A2-AS2*	-0.203	0.0	4.46
rs5760061	chr22_23835687_23836187	*DRICH1*	-0.21	0.0	4.78
rs11748963	chr5_139394081_139394581	*SPATA24*	-0.211	0.0	4.8
rs116904997	chr12_120230472_120230972	*COX6A1*	-0.219	0.0	5.19
rs2891537	chr12_114920788_114921288	*TBX3*	-0.22	0.0	5.23
rs5760061	chr22_23835687_23836187	*C22orf15*	-0.223	0.0	5.4

rs11838776	chr13_110388060_110388560	*COL4A2-AS1*	-0.24	0.0	6.25
rs6003909	chr22_23839075_23839575	*VPREB3*	-0.24	0.0	6.23
rs9515201	chr13_110388060_110388560	*COL4A2-AS1*	-0.24	0.0	6.25
rs5760061	chr22_23835687_23836187	*MIF*	-0.254	0.0	6.99
rs945418	chr1_16017870_16018370	*EPHA2*	-0.254	0.0	6.97
rs11838776	chr13_110388060_110388560	*NAXD*	-0.258	0.0	7.23
rs9515201	chr13_110388060_110388560	*NAXD*	-0.258	0.0	7.23
rs35006907	chr8_124847479_124847979	*MTSS1*	-0.259	0.0	7.27
rs3734634	chr6_125790622_125791122	*NCOA7*	-0.263	0.0	7.53
rs3734634	chr6_125790622_125791122	*HINT3*	-0.28	0.0	8.55
rs945418	chr1_16017870_16018370	*CPLANE2*	-0.28	0.0	8.56
rs5760061	chr22_23835687_23836187	*MMP11*	-0.285	0.0	8.87
rs6003909	chr22_23839075_23839575	*C22orf15*	-0.292	0.0	9.32
rs5760061	chr22_23835687_23836187	*RGL4*	-0.296	0.0	9.58
rs4887092	chr15_78750921_78751421	*CHRNA5*	-0.299	0.0	9.77
rs5760061	chr22_23835687_23836187	*DERL3*	-0.31	0.0	10.53
rs1763603	chr1_16012832_16013332	*EPHA2*	-0.313	0.0	10.76
rs7183401	chr15_84828318_84828818	*PDE8A*	-0.342	0.0	12.98
rs7183401	chr15_84828318_84828818	*WDR73*	-0.346	0.0	13.25
rs11838776	chr13_110388060_110388560	*COL4A2*	-0.368	0.0	15.11
rs9515201	chr13_110388060_110388560	*COL4A2*	-0.368	0.0	15.11
rs5760061	chr22_23835687_23836187	*SLC2A11*	-0.527	0.0	33.87

5.6 Gene Ontology Exploration

Gene ontology and pathway enrichment was explored given an input list of genes that were prioritised throughout the thesis (Open Targets and FUMA given the lead SNP, the quadrant analyses, peak-to-gene linkage, enhancer-gene links, and Open Targets given each of the prioritised variants).

5.6.1 DAVID

DAVID was employed for both gene ontology enrichment as well as pathway enrichment.

Table 30 | GO ontology functional annotation clustering via DAVID given prioritised genes list | A table showing 3 of the 17 clusters output from the DAVID functional annotation clustering tool given GO ontology terms 5 as the data source.

Annotation Cluster 1: Enrichment Score: 12.922					
Term	**Count**	**PValue**	**Genes**	**List Total**	**Bonferroni**
GO:0055002~striated muscle cell development	18	3.863E-15	*ACTN2 ADPRHL1 ALPK2 ALPK3 FHOD3 FLNC GATA4 HEY2 LMOD3 MYBPC3 MYOZ1 NRAP PROX1 SKI SYNPO2L TBX3 TMEM182 TTN*	153	4.919E-12
GO:0055001~muscle cell development	18	8.521E-14	*ACTN2 ADPRHL1 ALPK2 ALPK3 FHOD3 FLNC GATA4 HEY2 LMOD3 MYBPC3 MYOZ1 NRAP PROX1 SKI SYNPO2L TBX3 TMEM182 TTN*	153	1.078E-10
GO:0051146~striated muscle cell differentiation	18	5.210E-12	*ACTN2 ADPRHL1 ALPK2 ALPK3 FHOD3 FLNC GATA4 HEY2 LMOD3 MYBPC3 MYOZ1*	153	6.596E-9

			NRAP PROX1 SKI SYNPO2L TBX3 TMEM182 TTN		
Annotation Cluster 2: Enrichment Score: 8.632					
Term	**Count**	**PValue**	**Genes**	**List Total**	**Bonferroni**
GO:0055013~cardiac muscle cell development	11	1.510E-10	*ACTN2 ADPRHL1 ALPK2 ALPK3 FHOD3 GATA4 HEY2 NRAP PROX1 TBX3 TTN*	153	1.911E-7
GO:0055006~cardiac cell development	11	4.167E-10	*ACTN2 ADPRHL1 ALPK2 ALPK3 FHOD3 GATA4 HEY2 NRAP PROX1 TBX3 TTN*	153	5.275E-7
GO:0055007~cardiac muscle cell differentiation	11	5.620E-9	*ACTN2 ADPRHL1 ALPK2 ALPK3 FHOD3 GATA4 HEY2 NRAP PROX1 TBX3 TTN*	153	7.115E-6
GO:0035051~cardiocyte differentiation	11	8.414E-8	*ACTN2 ADPRHL1 ALPK2 ALPK3 FHOD3 GATA4 HEY2 NRAP PROX1 TBX3 TTN*	153	1.065E-4
Annotation Cluster 3: Enrichment Score: 2.078					
Term	**Count**	**PValue**	**Genes**	**List Total**	**Bonferroni**
GO:0003180~aortic valve morphogenesis	4	0.004	*AXIN2 GATA4 HEY2 ROCK2*	153	0.992
GO:0003176~aortic valve development	4	0.006	*AXIN2 GATA4 HEY2 ROCK2*	153	0.999
GO:0003170~heart valve development	4	0.026	*AXIN2 GATA4 HEY2 ROCK2*	153	1.000

Functional annotation clustering resulted in 17 clusters. Table 30 shows 3 clusters whose GO terms all had significant p values (≤0.05) and 2 of these produced significant Bonferroni corrected p-values (≤ 0.000256 (0.05/195)).

The cluster showing the highest enrichment score (12.922) was enriched for striated muscle cell development and differentiation; cardiomyopathy is an established disease of the cardiomyocyte which are known to exhibit a striated and branched structure. One of the genes enriched for these terms was the positive control gene for polygenic HCM, *ALPK3*. The cluster with the second highest enrichment score (8.632) was for cardiac muscle cell development and differentiation (also supported by ALPK3). The third highest cluster enrichment score (2.078) was for heart (aortic) valve development and morphogenesis. There is an established link between aortic valve abnormalities and cardiomyopathy – left ventricular noncompaction cardiomyopathy (LVNC) represents a significant contributor to DCM (Chin *et al.*, 1990); a potential association exists between LVNC and bicuspid aortic valve (BAV) (Nakanishi *et al.*, 2016).

Table 31 | Kegg pathway functional annotation clustering via DAVID given prioritised genes list | A table showing 1 of the 4 clusters output from the functional annotation clustering tool by DAVID given Kegg pathways as the data source.

Annotation Cluster 1: Enrichment Score: 3.077			
Term	**Count**	**PValue**	**Genes**
hsa04820:Cytoskeleton in muscle cells	12	1.206E-5	*ACTN2 CKM COL4A2 DTNA FLNC LMOD3 MYBPC3 MYOZ1 NRAP SSPN TNNC1 TTN*
hsa05414:Dilated cardiomyopathy	6	0.003	*DTNA MYBPC3 PLN SSPN TNNC1 TTN*
hsa05410:Hypertrophic cardiomyopathy	5	0.015	*DTNA MYBPC3 SSPN TNNC1 TTN*

Functional annotation clustering given Kegg pathways resulted in 4 clusters. Table 31 shows 1 of the 4 clusters – whereby all captured pathways in the clusters had a significant p-value (≤0.05). However, the terms in this cluster do not all show a Bonferroni correction p-value ≤ 0.000256 (0.05/195). Only 1 pathway passes this threshold (Cytoskeleton in muscle cells).

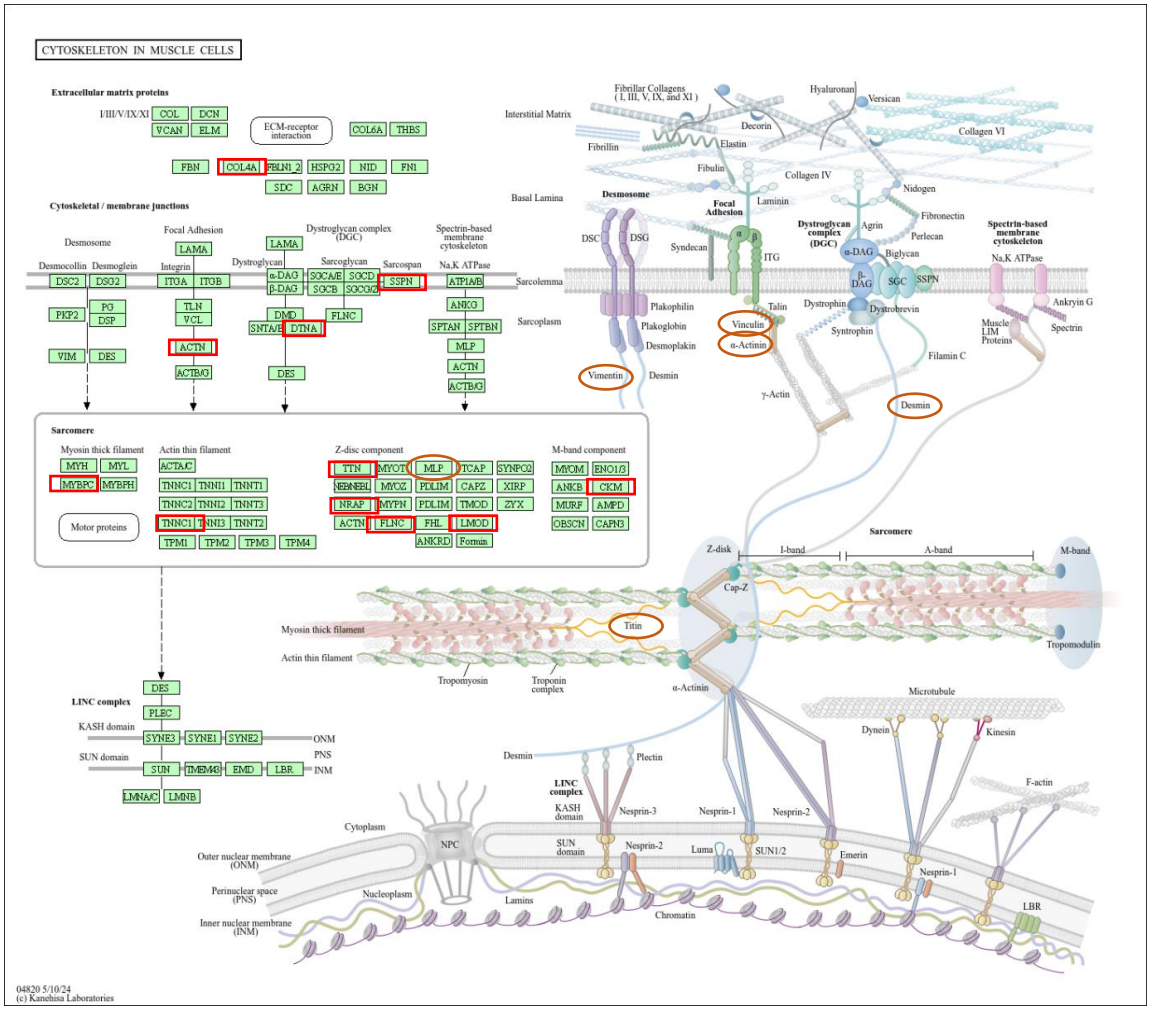
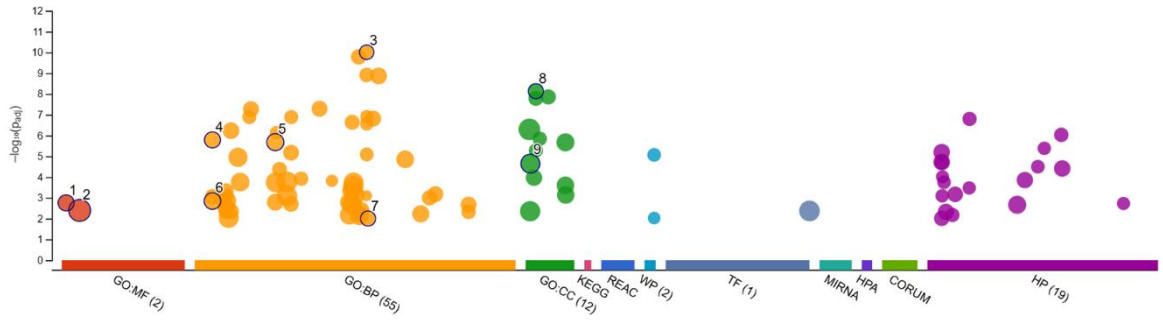


Figure 71 | The cytoskeleton in muscle cells pathway | Kegg pathway output by DAVID which shows the genes and interactions involved in the extracellular matrix, membrane junction, and sarcomere. Cytoskeletal proteins from the input list of genes are outlined in red boxes and those circled in orange have been associated with cardiomyopathy.

The cytoskeleton pathway depicted in figure 71 contains the following genes from the initial prioritised genes input list: *DTNA*, *SSPN*, *LMOD3*, *MYOZ1*, *NRAP*, *ACTN2*, *FLNC*, *CKM*, *TNNC1*, *COL4A2*, *MYBPC3*, and *TTN* (outlined in red boxes and detailed in table 31). These genes are situated in the sarcomere, cytoskeletal/membrane junction, as well as the extracellular matrix. Additionally, other cytoskeletal proteins (vinculin, desmin, tubulin, dystrophin, MLP, and myomesin) which have been associated with cardiomyopathy (mainly DCM) have been circled in orange.

5.6.2 gProfiler

gProfiler was employed as another method for gene ontology analysis; in contrast to the output seen for DAVID, gProfiler allowed for the visualisation across all input genes as to whether or not a particular term was enriched.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0003779	actin binding	1.747×10^{-3}
2	GO:MF	GO:0005515	protein binding	4.022×10^{-3}
3	GO:BP	GO:0055002	striated muscle cell development	9.765×10^{-11}
4	GO:BP	GO:0003012	muscle system process	1.621×10^{-6}
5	GO:BP	GO:0030029	actin filament-based process	2.081×10^{-6}
6	GO:BP	GO:0003013	circulatory system process	1.424×10^{-3}
7	GO:BP	GO:0060047	heart contraction	9.882×10^{-3}
8	GO:CC	GO:0030016	myofibril	7.329×10^{-9}
9	GO:CC	GO:0005856	cytoskeleton	2.252×10^{-5}

Figure 72 | GO ontology functional annotation clustering via gProfiler given prioritised genes list | A figure generated by gProfiler whereby GO driver terms are shown in the table and also numbered in the figure above according to their level of significance. All enriched terms passing the SCS correction threshold of 0.01 are plotted per data source.

Figure 72 shows that significant enrichments resulted for GO:MF, GO:BP, GO:CC, WP, TF, and HP, with no significant enrichment for KEGG, REAC, MIRNA, HPA, and CORUM terms. Only the GO driver terms are shown in the table segment of figure 72. As described by Kolberg *et al.* (2023), gProfiler assigns GO driver terms which represent the main GO terms driving the enrichment results. Given the driver terms and similarly to the result seen for DAVID, sarcomere and striated muscle cell development along with the cytoskeletal components and organisation thereof have been highlighted.

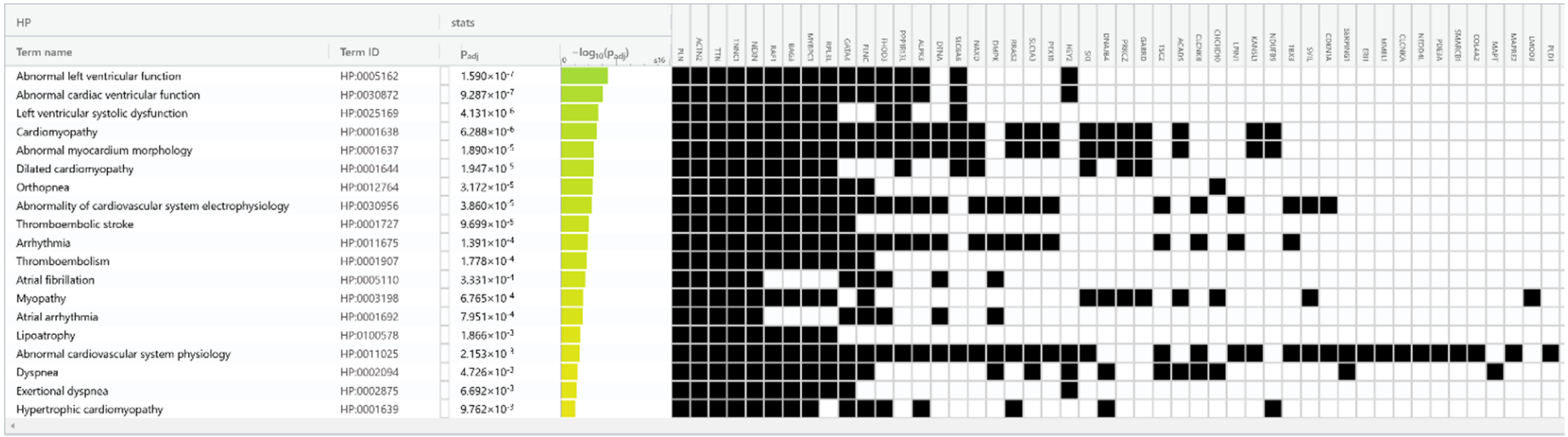


Figure 73 | Clustering of HP terms via gProfiler given prioritised genes list | A figure generated by gProfiler whereby the HP terms are shown in the detailed overview layout and the genes are ordered by the number of terms for which they are enriched. The associated significance per term is also shown via an adjusted p-value. Black boxes indicate the presence of human phenotype ontology enrichment (sequence homologs in other species).

gProfiler also provides a more detailed overview of all the enriched annotations passing the SCS threshold of 0.01 (not only the driver terms) across the input list of genes. Genes can be enriched for a specific term given an array of different evidences. These gProfiler results are confirmatory due to the enrichment terms being in alignment with HCM and its phenotype (as shown in figure 73). Most importantly, the enrichment of the cytoskeleton is concordant with that seen in the DAVID output – suggesting the importance of the cytoskeleton pathway in HCM pathogenesis.

5.7 Motif Analysis

In each of the regions where the prioritised variants reside, motif exploration was performed to see if any of the variants overlap with a known motif. The aim was to identify motifs likely being affected so as to investigate which mechanisms may be disrupted by the variants. The top motif per variant was determined through sorting by p-value significance (ascending), the number of target sequences containing the motif (descending), and by the number of background sequences containing the motif (ascending).

Upon inspecting the top known motifs across the variants, the minimum p-value was 0.001. When examining ChIP-seq or ChIP-Chip peaks, p-values provided by HOMER exceeding 1e-50 specify a significant motif enrichment (Heinz *et al.*, 2010). Heinz *et al.* (2010) suggest that if this expectation is not met, it may indicate issues in one's pipeline such as low-quality peaks which may be as a result of insufficient expression levels of the factor being studied.

The concatenated HOMER results for all 53 prioritised variants showing their top motif can be seen in table 32. Due to the p-values only being ~1e-3 as opposed to ~1e-50, these results were not assessed further. It is possible that some of these variants disrupt regulatory elements not yet identified or logged in databases (such as JASPAR), impact TF binding indirectly, or perhaps even alter chromatin accessibility rather than directly disrupting a TF motif.

Table 32 | Motif exploration surrounding prioritised variants via HOMER | A table depicting concatenated HOMER results showing the most significant 'known' motif detected per 53 prioritised variants. The motif name and consensus sequence is shown alongside the assigned p-value, percentage of either target or background sequences containing the motif, as well as the snp_test_id.

Motif Name	Consensus	P-value	% of Target Sequences with Motif	% of Background Sequences with Motif	Snp
Mouse_Recombination_Hotspot(Zf)/Testis-DMC1-ChIP-seq(GSE24438)/Homer	ACTYKNATTCGTGNTACTTC	0.001	100.0	0.09	6_54105134_G_C
MyoD(bHLH)/Myotube-MyoD-ChIP-seq(GSE21614)/Homer	RRCAGCTGYTSY	0.01	100.0	0.64	14_71380400_A_G
STAT1(Stat)/HelaS3-STAT1-ChIP-seq(GSE12782)/Homer	NATTTCCNGGAAAT	0.01	100.0	0.62	18_58255557_A_G
bHLHE40(bHLH)/HepG2-BHLHE40-ChIP-seq(GSE31477)/Homer	KCACGTGMCN	0.01	100.0	0.56	8_124847575_C_A
PRDM9(Zf)/Testis-DMC1-ChIP-seq(GSE35498)/Homer	ADGGYAGYAGCATCT	0.01	100.0	0.47	11_14288736_G_A
IRF4(IRF)/GM12878-IRF4-ChIP-seq(GSE32465)/Homer	ACTGAAACCA	0.01	100.0	0.59	12_120230731_G_A
ETS(ETS)/Promoter/Homer	AACCGGAAGT	0.01	100.0	0.33	14_71693604_T_C

c-Myc(bHLH)/mES-cMyc-ChIP-seq(GSE11431)/Homer	VVCCACGTGG	0.01	100.0	0.5	1_77683091_CGCCGTCGCA_C
p63(p53)/Keratinocyte-p63-ChIP-seq(GSE17611)/Homer	NNDRCATGYCYNRRCATGYH	0.01	100.0	0.41	17_46004161_T_C
Tlx?(NR)/NPC-H3K4me1-ChIP-seq(GSE16256)/Homer	CTGGCAGSCTGCCA	0.01	100.0	0.71	15_78751333_T_C
Tbet(T-box)/CD8-Tbet-ChIP-seq(GSE33802)/Homer	AGGTGTGAAM	0.01	100.0	0.97	1_16026905_T_C
ETS(ETS)/Promoter/Homer	AACCGGAAGT	0.01	100.0	0.35	14_71693604_T_A
Zscan4c(Zf)/ESC-Zscan4c-ChIP-seq(GSE140619)/Homer	ATRTGTGCAC	0.01	100.0	0.48	2_36889930_C_T
Foxa3(Forkhead)/Liver-Foxa3-ChIP-seq(GSE77670)/Homer	BSNTGTTTACWYWGN	0.01	100.0	0.85	5_57718369_C_T
BMYB(HTH)/Hela-BMYB-ChIP-seq(GSE27030)/Homer	NHAACBGYYV	0.01	100.0	0.56	6_125790654_G_A
Prop1(Homeobox)/GHFT1-PROP1.biotin-ChIP-seq(GSE77302)/Homer	NTAATBNAATTA	0.01	100.0	0.84	6_149766491_C_A
Tbx20(T-box)/Heart-Tbx20-ChIP-seq(GSE29636)/Homer	GGTGYTGACAGS	0.01	100.0	0.29	1_78053255_A_G

Gata1(Zf)/K562-GATA1-ChIP-seq(GSE18829)/Homer	SAGATAAGRV	0.01	100.0	0.52	9_131624672_G_A
TEAD(TEA)/Fibroblast-PU.1-ChIP-seq(Unpublished)/Homer	YCWGGAATGY	0.01	100.0	0.39	1_16018235_T_C
NFIL3(bZIP)/HepG2-NFIL3-ChIP-seq(Encode)/Homer	VTTACGTAAYNNNNN	0.01	100.0	0.67	19_45810500_G_C
CTCF(Zf)/CD4+-CTCF-ChIP-seq(Barski_et_al.)/Homer	AYAGTGCCMYCTRGTGGCCA	0.01	100.0	0.17	13_110388334_G_A
Hnf1(Homeobox)/Liver-Foxa2-ChIP-seq(GSE25694)/Homer	GGTTAAWCATTAA	0.01	100.0	0.27	4_16034172_A_C
RARa(NR)/K562-RARa-ChIP-seq(Encode)/Homer	TTGAMCTTTG	0.01	100.0	0.78	1_16013314_T_G
Mef2d(MADS)/Retina-Mef2d-ChIP-seq(GSE61391)/Homer	GCTATTTTTAGC	0.01	100.0	0.54	1_2212668_A_G
FXR(NR),ER2/Liver-FXR-ChIP-seq(GSE133700)/Homer	NTGACCYMKRGGTCA	0.01	100.0	0.74	22_23839465_A_G
Zac1(Zf)/Neuro2A-Plagl1-ChIP-seq(GSE75942)/Homer	HAWGRGGCCM	0.01	100.0	0.29	5_57693080_T_TA
Gata1(Zf)/K562-GATA1-ChIP-seq(GSE18829)/Homer	SAGATAAGRV	0.01	100.0	0.56	6_118370989_G_C

n-Myc(bHLH)/mES-nMyc-ChIP-seq(GSE11431)/Homer	VRCCACGTGG	0.01	100.0	0.27	18_35090057_T_C
ZFX(Zf)/mES-Zfx-ChIP-seq(GSE11431)/Homer	AGGCCTRG	0.1	100.0	1.26	3_172041620_C_CCT
AMYB(HTH)/Testes-AMYB-ChIP-seq(GSE44588)/Homer	TGGCAGTTGG	0.1	100.0	2.97	8_8310872_G_A
ELF3(ETS)/PDAC-ELF3-ChIP-seq(GSE64557)/Homer	ANCAGGAAGT	0.1	100.0	1.5	14_71686976_A_G
Dlx3(Homeobox)/Keratinocytes-Dlx3-ChIP-seq(GSE89884)/Homer	NDGTAATTAC	0.1	100.0	1.16	14_71586380_G_A
MRE(NR)/Neuro2A-NR3C2-ChIPnexus(GSE115417)/Homer	GGAACAGAVTGTCCT	0.1	100.0	2.47	6_149746539_A_G
MITF(bHLH)/MastCells-MITF-ChIP-seq(GSE48085)/Homer	RTCATGTGAC	0.1	100.0	1.63	17_46101518_G_C
BORIS(Zf)/K562-CTCFL-ChIP-seq(GSE32465)/Homer	CNNBRGCGCCCCCTGSTGGC	0.1	100.0	1.69	5_139394348_T_C
ELF3(ETS)/PDAC-ELF3-ChIP-seq(GSE64557)/Homer	ANCAGGAAGT	0.1	100.0	1.45	14_71686976_A_T
TEAD(TEA)/Fibroblast-PU.1-ChIP-seq(Unpublished)/Homer	YCWGGAATGY	0.1	100.0	1.33	3_172082466_A_G

HIF-1b(HLH)/T47D-HIF1b-ChIP-seq(GSE59937)/Homer	RTACGTGC	0.1	100.0	2.1	22_23836092_G_A
ZNF341(Zf)/EBV-ZNF341-ChIP-seq(GSE113194)/Homer	GGAACAGCCG	0.1	100.0	1.57	17_45722444_G_C
Zac1(Zf)/Neuro2A-Plagl1-ChIP-seq(GSE75942)/Homer	HAWGRGGCCM	0.1	100.0	3.2	5_57700364_A_T
Zfp541(Zf)/Sperm-Zfp541-ChIP-seq(GSE163916)/Homer	NRSCWCGYGC	0.1	100.0	8.49	17_45897110_G_T
CEBP(bZIP)/ThioMac-CEBPb-ChIP-seq(GSE21512)/Homer	ATTGCGCAAC	0.1	100.0	1.51	15_84828713_G_A
AP-2alpha(AP2)/Hela-AP2alpha-ChIP-seq(GSE31477)/Homer	ATGCCCTGAGGC	0.1	100.0	3.44	6_149718225_C_G
FOXA1(Forkhead)/MCF7-FOXA1-ChIP-seq(GSE26831)/Homer	WAAGTAAACA	0.1	100.0	2.81	18_35085797_C_G
CEBP(bZIP)/ThioMac-CEBPb-ChIP-seq(GSE21512)/Homer	ATTGCGCAAC	0.1	100.0	1.53	15_84828713_G_T
WT1(Zf)/Kidney-WT1-ChIP-seq(GSE90016)/Homer	MCTCCCMCRCAB	0.1	100.0	2.2	5_139439612_C_T
ZNF467(Zf)/HEK293-ZNF467.GFP-ChIP-seq(GSE58341)/Homer	TGGGGAAGGGCM	0.1	100.0	4.96	6_149566308_T_C

CLOCK(bHLH)/Liver-Clock-ChIP-seq(GSE39860)/Homer	GHCACGTG	0.1	100.0	1.03	9_131640601_ACT_A
Dlx3(Homeobox)/Keainocytes-Dlx3-ChIP-seq(GSE89884)/Homer	NDGTAATTAC	0.1	100.0	1.18	14_71586380_G_C
DLX1(Homeobox)/BasalGanglia-Dlx1-ChIP-seq(GSE124936)/Homer	NSNNTAATTA	0.1	100.0	3.0	2_200330879_C_T
Oct6(POU,Homeobox)/NPC-Pou3f1-ChIP-seq(GSE35496)/Homer	WATGCAAATGAG	0.1	100.0	1.03	12_114921147_T_G
BORIS(Zf)/K562-CTCFL-ChIP-seq(GSE32465)/Homer	CNNBRGCGCCCCCTGSTGGC	1.0	0.0	0.0	12_114933937_T_G
DUX(Homeobox)/C2C12-Dux-ChIP-seq(GSE87279)/Homer	BCWGATTCAATCAAN	1.0	0.0	0.0	13_110388451_A_C

5.8 Discussion & Limitations

Riva *et al.* (2024) emphasised how the majority of regulatory annotation methods not only differ in their approach and assumptions but also produce diversified results. HMM approaches require manual annotation of elements after assigning arbitrary chromatin states and unsupervised learning methods are better for grouping elements as opposed to classification. REgulamentary was leveraged in this analysis as it has been shown to outperform GenoSTAN (which is otherwise the gold-standard tool for regulatory element assignment).

One would assume that the 53 prioritised variants exist in functional regions (also known as regulatory elements) if they are deemed capable of having an impact on the surrounding chromatin,

expression of genes, and susceptibility to HCM. This is especially true for loss of function variants as opposed to gain. There were, however, numerous examples where the underlying regulatory element was not assigned by REgulamentary and visual inspection failed to resolve the site. A regulatory site may only be established under certain conditions (specific environment, cell phase, cell type etc.) which might not have been captured in the data. Additionally, the quality of some datasets may have prevented accurate assignment.

Understanding and unpacking the dynamics surrounding a variant of interest without the chromatin landscape being annotated at this point is challenging. With the absence of element annotation, one might assume the site acts as a silencer – complicated by the lack of a histone mark signature to confirm this. Additionally, when residing within a gene, the internal machinery could be affected (such as alternative splicing) (Park *et al.*, 2018).

Nasser *et al.* (2021) examined all distal element–gene associations with an ABC model score of at least 0.015 with the aim of enhancing recall and to enhance the identification of gain-of-function variants. I should possibly have used the less stringent threshold of 0.015 as opposed to 0.027; however, the output enhancer-gene links will have been assigned higher confidence given the 0.027 cut-off. Regarding gene ontology and pathway enrichment, both DAVID and gProfiler were employed as Kolberg *et al.* (2023) confirmed that they are frequently updated regarding their data.

5.9 Case Studies

For each of the 53 prioritised variants, the relevant loci and all evidences collated throughout the thesis were investigated in more detail.

Five case studies are illustrated:

- A locus whereby the prioritised target gene has been confirmed in prior research
- A locus containing a prioritised variant which was assigned the most extreme damage score via Deepmind's Enformer
- A locus where the evidences prioritise different target genes and thus show a lack of consensus

- A locus deemed cautionary as the prioritised target genes differ and the literature provides an explanation as to why
- A locus where all evidences align and point to the same target gene (a consensus is reached)

The expression quadrant plots for each locus discussed can be found at GitHub (https://github.com/domwest/DPhil-thesis/blob/main/proximity_based_bulk_LV_qa_plots.zip; https://github.com/domwest/DPhil-thesis/blob/main/tad_based_sc_CM_qa_plots.zip).

5.9.1 Considerations/assumptions

Using eQTL Supporting Data to Link Variants to Target Genes

Transcriptional repressors attach to promoter or enhancer regions thereby inhibiting transcription. Similar to transcriptional activators, repressors also react to external signals albeit to obstruct (rather than recruit) the attachment of activating transcription factors (Wolffe, 1998). However, activators generally exhibit a stronger correlation with nucleosome sequence profiles compared to repressors (Charoensawan *et al.*, 2012).

In light of this, an assumption is made throughout the case studies that enhancers show more of an activating role and thus lead to an increase in the expression of target genes.

Enhancer/promoter interchangeability

Within each case study, an overview of all collated evidences is first shown, followed by the relevant established literature for prioritised genes in the locus, STRING database interactions in the locus (in the event of an interesting interaction), and a locus zoom plot paired with a conclusion highlighting the likely target gene(s) in the locus.

Kim and Shiekhattar (2015) reinforce the idea that enhancers and promoters not only exhibit numerous architectural similarities, such as nucleosome hypersensitivity and chromatin

modifications, but that they may also serve interchangeable functional roles. Even in the presence of a REgulamentary assignment, an assumption is made throughout the case studies that an assigned promoter could also act as an enhancer (and vice versa). Additionally, in the absence of a REgulamentary assignment, if enhancer activity is suggested by means of one/some of the enhancer histone mark dataset activity then this role is considered.

Linking SNP to gene

Pinning the SNP to gene, as was carried out in the case studies, introduces an element of subjectivity. This is due to the consideration of many prioritisation tools in combination with visual inspection / elements in which the variants reside. In most cases, the consensus gene(s) are put forward; however, in some cases, the established literature might have prompted a different conclusion.

5.9.2 Case study Loci

Confirmatory example; rs35006907 Locus

Table 33 | rs35006907 Locus evidence table | A table depicting an overview of relevant evidences for the rs35006907 locus and its prioritised variant(s).

Locus-to-gene	**Tadros paper**	**Lead rsid**	rs35006907
		OT	*MTSS1*
		FUMA	*MTSS1*
	Quadrant approach	**Bulk gene symbol**	*NDUFB9*
		Sc gene symbol	*LINC00964*
		Sc gene highest expr in CM	Y
	Gene prioritisation tools	**PoPs top gene**	*ZNF572*
		PoPs score	-0.083
		Significant colocalisation	N
		Coloc gene	
		Coloc H4	
		Coloc SNP	
		Coloc SNP H4	
	Deepmind	**Prioritised proxy rsid**	rs35006907

SNP-to-gene		**Prioritised SNP details**	8 124847575 C A
		MTAG p val 5e-8	N/A
		MTAG OR 1	N/A
		Var type	loss
		Top classifier's damage score	-0.413
		Top classifier	DNase:cardiac muscle cell 1
	Identification of plausible target genes	**REgulamentary element**	Enhancer
		Visual inspection of chromatin tracks	Enhancer within the RNA gene, ENSG00000255491
		Eqtl target genes	[*MTSS1, LINC00964, MTSS1, SQLE*]
		Eqtl heart tissue	[Heart - Left Ventricle, Heart - Left Ventricle, Heart - Atrial Appendage, Heart - Left Ventricle]
		Eqtl p value	[1.1e-44, 9.1e-19, 2.4e-9, 0.0000038]
		Eqtl NES	[-0.72, -0.36, -0.26, -0.19]
		ABC target genes	
		[SNP] archr top gene FDR<=0.05	*SQLE*
		[SNP] OT top gene	*MTSS1*
Implicated gene(s)			*MTSS1*

***MTSS1 (*Metastasis suppressor 1 gene); Cytoplasm, cytoskeleton**

Role in heart / myopathy

Choe *et al.* (2024) fine-mapped the *MTSS1* locus with CMR traits in UK Biobank individuals and found *MTSS1* enhancer variants to interact with biological sex in their associations with numerous of the traits – mainly in women and significantly less in men. *MTSS1*-lowering alleles and a corresponding increased LV ejection fraction and reduced LV volumes were observed for females. These findings suggest possible sex dimorphism in cardiac remodelling.

Association with CM

Morley *et al.*, (2019) interest was piqued regarding a (lead) variant, rs12541595, residing in the *MTSS1* locus after revealing that its minor allele associated with reduced left ventricular end-diastolic diameter (LVEDD) *MTSS1* expression. The authors found 2 common variants in high LD with the lead variant (rs35006907 and rs34866937) and found all three variants to be residing in a 1kb predicted enhancer; the enhancer's cardiac- and myocyte-specific activity was confirmed in zebrafish embryo ventricles. The 3 variants were shown to define a major (GCG, frequency 0.66) and minor haplotype (TAA, frequency 0.33) at this site. Deletion of the enhancer region via CRISPR was seen to reduce *MTSS1* expression. Additionally, the initiation of CRISPR aimed at the enhancer variants resulted in reduced activity with Cas9 recruited to rs35006907. In DCM cases, the authors observed an 18% risk reduction per copy of the lead variant's minor allele (P=1×10–4) as well as an increase in LV systolic function and decrease in LV mass. Ultimately the authors concluded that the enhancer variants which reduce *MTSS1* expression have cardioprotective potential.

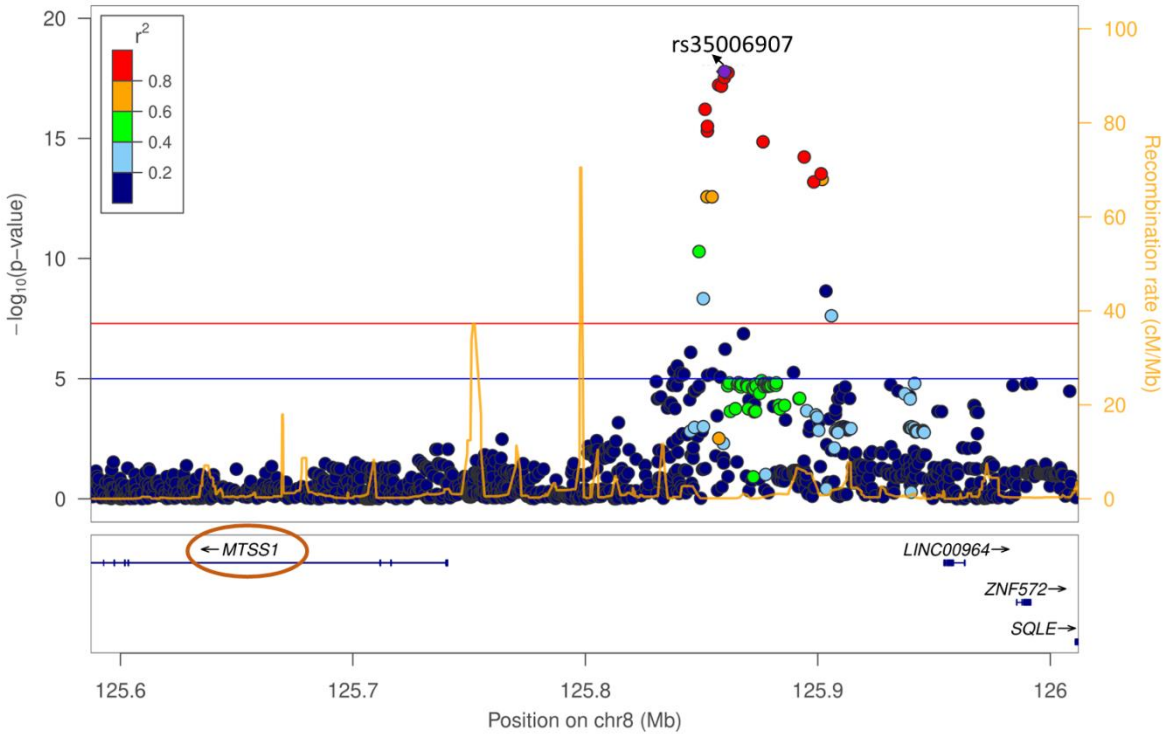


Figure 74 | *rs35006907 Locus zoom plot* | *A locus zoom plot for the TAD-defined locus depicting the variant(s) prioritised, rs35006907. Circled gene(s) represent plausible targets given all the supporting evidences.*

Although *LINC00964* was prioritised with regards to its specificity and expression in cardiomyocytes, *MTSS1* showed the highest expression in cardiomyocyte cells. The eQTL hits proposed *MTSS1* as a likely target gene (with the highest significance) along with *LINC00964* and *SQLE* all in heart tissue – with the expected direction of effect given a loss of function variant in an enhancer. ArchR assigned the peak overlapping with this variant to the target gene *SQLE*; however, this did not pass the significance threshold regarding its correlation score. Not only has *MTSS1* previously been associated with DCM, but the established causal variant is the same as was prioritised in my analysis. This variant has been shown to reside in an enhancer region whereby the deletion thereof resulted in reduced expression of *MTSS1*. *MTSS1* therefore presents an interesting potential target gene, especially when one considers the opposite effect of directionality in HCM and DCM (Harper *et al.*, 2021). The reduced expression of *MTSS1* appears to be cardioprotective for DCM yet causal for HCM.

Extreme damage score example; rs116904997 Locus

Table 34 | rs116904997 Locus evidence table | A table depicting an overview of relevant evidences for the rs116904997 locus and its prioritised variant(s).

Locus-to-gene	**Tadros paper**	**Lead rsid**	rs116904997
		OT	*PXN*
		FUMA	*PXN*
	Quadrant approach	**Bulk gene symbol**	*ACADS*
		Sc gene symbol	*PXN-AS1*
		Sc gene highest expr in CM	Y
	Gene prioritisation tools	**PoPs top gene**	*PXN*
		PoPs score	0.205
		Significant colocalisation	Y
		Coloc gene	*PXN*
		Coloc H4	0.996
		Coloc SNP	rs116904997
		Coloc SNP H4	0.846
SNP-to-gene	**Deepmind**	**Prioritised proxy rsid**	rs116904997
		Prioritised SNP details	12 120230731 G A

		MTAG p val 5e-8	N/A
		MTAG OR 1	N/A
		Var type	loss
		Top classifier's damage score	-3.897
		Top classifier	DNase:cardiac muscle cell 1
	Identification of plausible target genes	**REgulamentary element**	Enhancer
		Visual inspection of chromatin tracks	Enhancer within *PXN*
		Eqtl target genes	[*PXN*]
		Eqtl heart tissue	[Heart - Left Ventricle]
		Eqtl p value	[1.4e-11]
		Eqtl NES	[-0.64]
		ABC target genes	[*PXN, RPLP0*]
		ABC scores	[0.090102, 0.029797]
		[SNP] archr top gene FDR<=0.05	*MSI1*
		[SNP] OT top gene	*PXN*
Implicated gene(s)			*PXN*

***PXN* (Focal adhesion adapter protein called paxillin); Cytoplasm, cytoskeleton, cell junction, focal adhesion**

Association with CM

Pirruccello *et al.* (2020) highlighted the critical role of costameres (cytoskeletal structures linking the sarcomere to the cell membrane and extracellular matrix) within cardiomyocytes whilst showing that a disruption to their assemblies could result in cardiomyopathies. Costameres comprise various protein complexes, including the dystrophin glycoprotein complex (DGC) and the vinculin–talin–integrin system. While both the DGC and the vinculin–talin–integrin system are involved in mechanical anchoring, the latter also contributes to mechanosensation and signal transduction. The authors found *FLNC*, *PTK2*, *PXN*, and *SSPN* to reside closest to four lead variants; *SSPN* and *PXN* were further prioritised at their respective loci through transcriptome-wide association studies (TWAS). *PXN* interacts with vinculin and integrin, thereby forming a component of the vinculin–talin–integrin system.

Aung *et al.* (2023) conducted a GWAS on left ventricular mass wall thickness (LVMWT) to assess the genetic correlation between LVMWT and HCM which led to the identification of *PROX1*, *PXN*, and *PTK2* as novel genes with known functional roles in myocardial growth and sarcomere organisation.

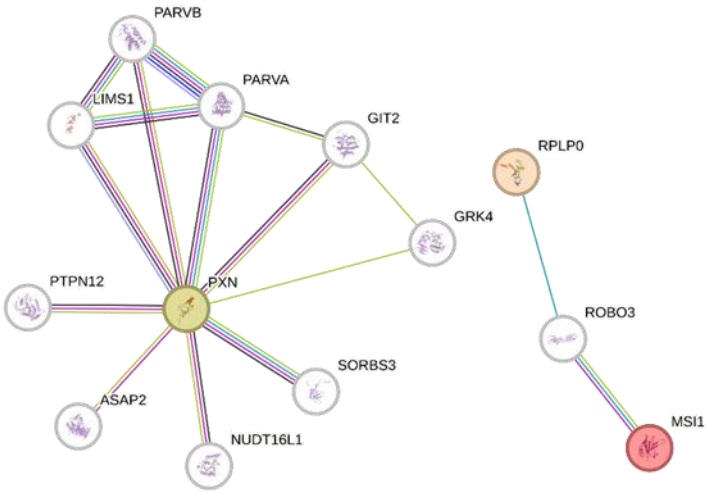


Figure 75 | rs116904997 STRING database interactions| A visualisation by STRING of noted functional relationships given the input list of all plausible genes within the rs116904997 locus whereby a function in the heart has been elucidated.

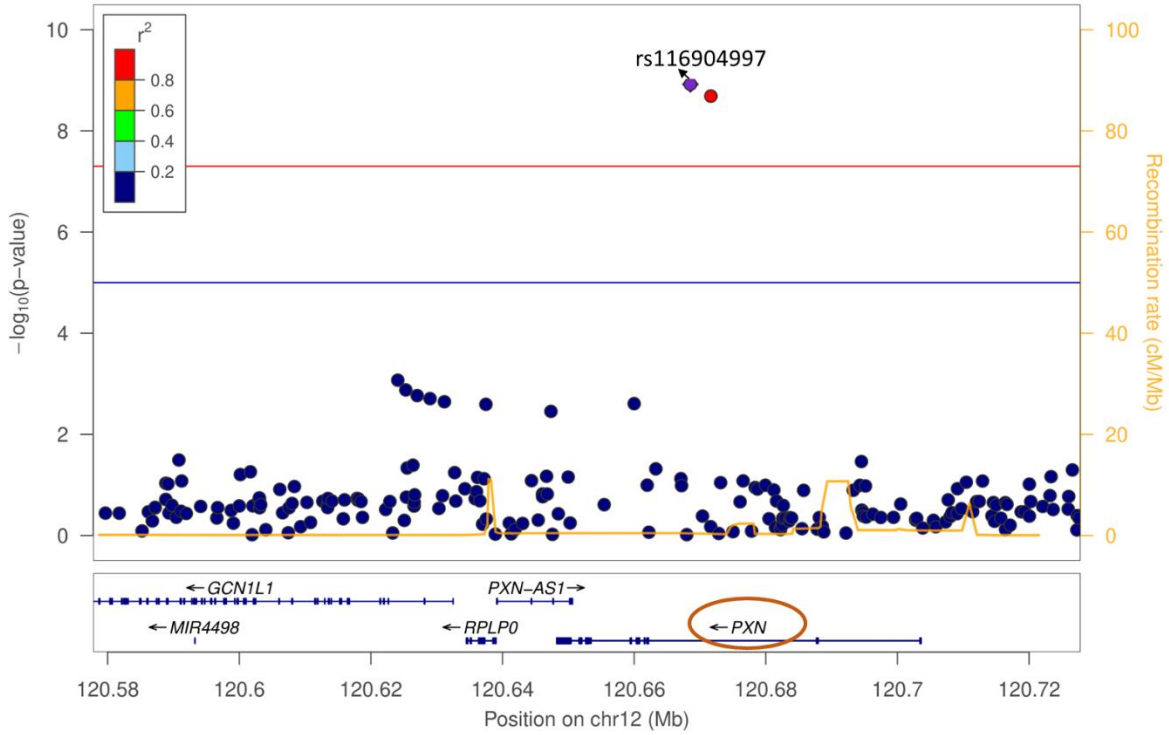


Figure 76 | rs116904997 Locus zoom plot | A locus zoom plot for the TAD-defined locus depicting the variant(s) prioritised, rs116904997. Circled gene(s) represent plausible targets given all the supporting evidences.

ACADS was prioritised given proximity-based locus demarcation with bulk heart LV data and *PXN-AS1* was prioritised using the TAD-based single-cell expression data approach. *ACADS* was excluded in the TAD-demarcation approach. In the TAD-based approach, the prioritisation of *PXN-AS1* was mostly as a result of its specificity as opposed to its expression – *PXN* showed a much higher expression albeit having a specificity value below the specificity threshold. The eQTL data suggested *PXN* to be the most likely target gene with the direction of effect in accordance with that of a loss of function variant operating in an enhancer site leading to its increased expression. The ABC model output data also highlighted *PXN* as the most likely target gene given the presence of an enhancer. ArchR linked the peak in which this variant resides to *MSI1*; however, the correlation was not significant and this gene did not fall within the TAD suggesting its implausibility. Much less evidence exists for *PXN-AS1* regarding its role in the heart when compared to *PXN*; however, *PXN-AS1* was shown to upregulate *PXN* in lung cancer patients (Zhang *et al.*, 2019). *PXN* has been shown to link functionally with *SORBS3* as a result of their presence in the curated integrin: cytoskeletal adhesion complex, among others. *SORBS2* knockout mice have exhibited fatal arrhythmogenic cardiomyopathy (Ding *et al.*, 2020) and left ventricular noncompaction (Li *et al.*, 2020). McLendon *et al.* (2022) demonstrated that Sorbs2 plays a structural role in cardiomyocytes through assisting in microtubule-cytoskeletal interaction. The authors also expressed uncertainty as to whether other SORBS proteins may compensate for the absence of Sorbs2 to preserve cardiac function. All three proteins are highly conserved, share localisation patterns, and interact suggesting that future research should focus on the roles of Sorbs1 and Sorbs3. Regardless, *PXN* is a prominent component of the vinculin–talin–integrin cytoskeletal system and was also prioritised while assessing the correlation between LVMWT and HCM and it is therefore the most likely target gene in this locus.

Lack of Consensus Across Evidences Example; rs11748963 Locus

Table 35 | rs11748963 Locus evidence table | A table depicting an overview of relevant evidences for the rs11748963 locus and its prioritised variant(s).

Locus-to-gene	**Tadros paper**	**Lead rsid**	rs11748963	
		OT	*PROB1*	
		FUMA	*SPATA24*	
	Quadrant approach	**Bulk gene symbol**	*PROB1*	
		Sc gene symbol	*PROB1*	
		Sc gene highest expr in CM	Y	
	Gene prioritisation tools	**PoPs top gene**	*PAIP2*	
		PoPs score	0.142	
		Significant colocalisation	Y	
		Coloc gene	*SPATA24*	
		Coloc H4	0.534	
		Coloc SNP	rs4385202	
		Coloc SNP H4	0.368	
SNP-to-gene	**Deepmind**	**Prioritised proxy rsid**	rs3923060	rs11748963
		Prioritised SNP details	5 139439612 C T	5 139394348 T C

		Mtag p val 5e-8	5.135000e-14	N/A
		Mtag OR 1	1.176424	N/A
		Var type	loss	loss
		Top classifier's damage score	-0.575	-0.833
		Top classifier	DNase:cardiac muscle cell 1	DNase:cardiac muscle cell 2
	Identification of plausible target genes	**REgulamentary element**	Promoter	Promoter/CTCF
		Visual inspection of chromatin tracks	Promoter & possible enhancer within *DNAJC18*	Promoter & possible enhancer within *PROB1*
		Eqtl target genes	[*DNAJC18, DNAJC18, SPATA24, PROB1, SPATA24, PROB1*]	[*DNAJC18, DNAJC18, SPATA24, PROB1, SPATA24, PROB1*]
		Eqtl heart tissue	[Heart - Atrial Appendage, Heart - Left Ventricle, Heart - Left Ventricle, Heart - Atrial Appendage, Heart - Atrial Appendage, Heart - Left Ventricle]	[Heart - Atrial Appendage, Heart - Left Ventricle, Heart - Left Ventricle, Heart - Atrial Appendage, Heart - Atrial Appendage, Heart - Left Ventricle]
		Eqtl p value	[1.8e-41, 2.5e-31, 3.9e-9, 1.8e-7, 0.0000036, 0.000019]	[5.7e-32, 4.2e-28, 2.5e-10, 3.7e-9, 0.0000010, 0.0000016]

		Eqtl NES	[-0.64, -0.48, -0.19, 0.14, -0.14, 0.094]	[0.59, 0.47, 0.21, -0.16, 0.15, -0.11]
		ABC target genes	[*DNAJC18*]	[*MZB1*, *PROB1*]
		ABC scores	[1.0]	[0.057833, 0.029738]
		[SNP] archr top gene FDR<=0.05	*SNHG4*	*UBE2D2*
		[SNP] OT top gene	*SPATA24*	*PROB1*
Implicated gene(s)			*PROB1/DNAJC18*	

***PROB1 (*Proline rich basic protein 1*);* Nucleoplasm**

Role in heart / myopathy

CEFIP contains a domain of unknown function (DUF4585, pfam15232) which is also present in PROB1.

Dierck *et al.* (2017) found the cardiac z-disc protein *CEFIP* to regulate cardiomyocyte hypertrophy through calcineurin signalling. Specifically, they found its overexpression to result in increased hypertrophy.

***DNAJC18 (*DnaJ heat shock protein family member C18); Endoplasmic reticulum membrane**

Role in heart / myopathy

Levin *et al.* (2022) performed colocalisation to uncover common genetic signals linked to gene expression in the heart, heart failure, and MRI characteristics. They found *DNAJC18* to demonstrate strong evidence of colocalisation.

Spielmann *et al.* (2022) revealed that mice with variants in *DNAJC18* develop cardiac structural abnormalities; human orthologs were categorised as variants of unknown relevance (VUR). The authors also validated the importance of *DNAJC18* for cardiac homeostasis by showing that its loss of function is associated with altered left ventricular systolic function.

Association with CM

Epigenetic marks at *DNAJC18* have previously been linked to DCM due to a haQTL (rs7445264) in its promoter whose alternative allele demonstrated diminished enhancer activity (Tan *et al.*, 2020). This variant is in LD (R2 0.775) with the prioritised variant in this case study residing within *DNAJC18*, rs3923060.

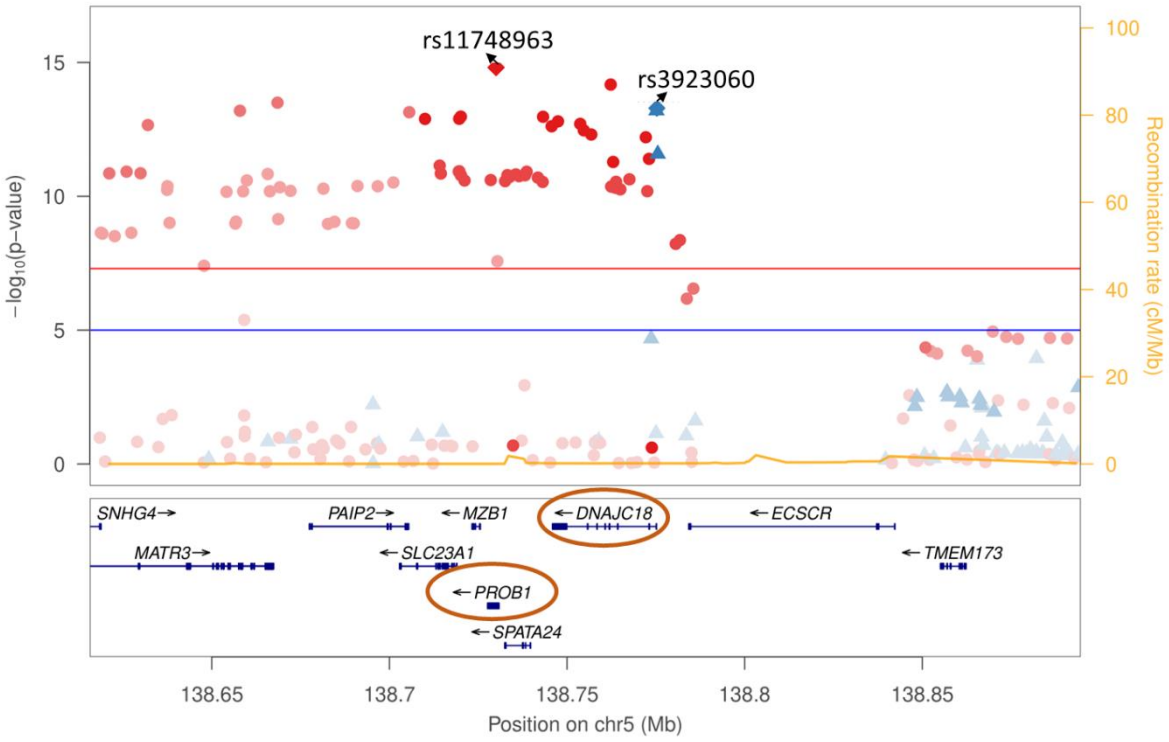


Figure 77 | rs11748963 Locus zoom plot | A locus zoom plot for the TAD-defined locus depicting the variant(s) prioritised (rs3923060 and rs11748963). Circled gene(s) represent plausible targets given all the supporting evidences.

PROB1 was prioritised given proximity-based locus demarcation with bulk heart LV data as well as through TAD-demarcation with single-cell cardiomyocyte expression data. Although *PROB1* was the best gene in cardiomyocyte-specific quadrant analysis, this is largely due to its high specificity as other genes showed higher expression values: *SPATA24*, *DNAJC18* and *TMEM173* (to name a few). For rs3923060, the eQTL data suggested *DNAJC18* and *SPATA24* as the most likely target genes with the direction of effect in accordance with the loss of function variant in either *DNAJC18*'s promoter or an enhancer site causing reduced expression thereof; however, the ABC model output narrowed the most likely target gene down to *DNAJC18*. For rs11748963, the eQTL data suggested *PROB1* which was supported by the ABC model output. Although REgulamentary assigned these two variant sites as promoter elements, the H3Kme1 and H3K27ac tracks suggest possible enhancer activity. Research conducted by Tan *et al.* (2020) confirms enhancer activity at the *DNAJC18* promoter site. Although *PROB1* has little evidence of its role in the heart or association with cardiomyopathy, it has been revealed to contain a domain of unknown function which is also found in *CEFIP*, a protein associated with cardiac hypertrophy. *DNJAC18* has been associated with DCM,

cardiac abnormalities, altered cardiac homeostasis, and modified left ventricular systolic function. Both genes constitute the most likely target genes in this locus; rs3923060 likely targets *DNAJC18* and rs11748963 *PROB1*.

Cautionary Example; rs28768976 Locus

Table 36 | rs28768976 Locus evidence table | A table depicting an overview of relevant evidences for the rs28768976 locus and its prioritised variant(s).

Locus-to-gene	**Tadros paper**	**Lead rsid**	rs28768976			
		OT	*PLEKHM1*			
		FUMA	*MAPT*			
	Quadrant approach	**Bulk gene symbol**	*PLCD3*			
		Sc gene symbol	*MAPT*			
		Sc gene highest expr in CM	Y			
	Gene prioritisation tools	**PoPs top gene**	*MAPT*			
		PoPs score	0.485			
		Significant colocalisation	N			
		Coloc gene				
		Coloc H4				
		Coloc SNP				
		Coloc SNP H4				
SNP-to-gene	**Deepmind**	**Prioritised proxy rsid**	rs80346216	rs62061810	rs62064665	rs76667867
		Prioritised SNP details	17 45897110 G T	17 46101518 G C	17 46004161 T C	17 45722444 G C

		MTAG p val 5e-8	N/A	N/A	N/A	N/A
		MTAG OR 1	N/A	N/A	N/A	N/A
		Var type	gain	gain	Loss	loss
		Top classifier's damage score	0.530	0.363	-0.322	-0.351
		Top classifier	DNase:cardiac muscle cell 2	DNase:cardiac muscle cell 1	DNase:heart left ventricle female embryo (101 day) and female embryo (103 days)	DNase:cardiac muscle cell 2
	Identification of plausible target genes	**REgulamentary element**				
		Visual inspection of chromatin tracks	Within *MAPT*	Within *KANSL1*	Within *MAPT*	Within *LINC02210-CRHR1*
		Eqtl target genes	[*LINC02210, LINC02210, LRRC37A4P, KANSL1-AS1, MAPK8IP1P2, MAPK8IP1P2*]		[*LINC02210, LINC02210, LRRC37A4P, KANSL1-AS1, MAPK8IP1P2, MAPK8IP1P2*]	[*LINC02210, LINC02210, LRRC37A4P, KANSL1-AS1, MAPK8IP1P2, MAPK8IP1P2*]

		Eqtl heart tissue	[Heart - Atrial Appendage, Heart - Left Ventricle, Heart - Atrial Appendage, Heart - Atrial Appendage, Heart - Left Ventricle, Heart - Atrial Appendage]		[Heart - Atrial Appendage, Heart - Left Ventricle, Heart - Atrial Appendage, Heart - Atrial Appendage, Heart - Atrial Appendage, Heart - Left Ventricle]	[Heart - Atrial Appendage, Heart - Left Ventricle, Heart - Atrial Appendage, Heart - Atrial Appendage, Heart - Atrial Appendage, Heart - Left Ventricle]
		Eqtl p value	[3.6e-135, 7.4e-101, 1.2e-91, 1.4e-88, 9.3e-84, 7.8e-83]		[3.0e-138, 7.4e-101, 1.2e-92, 5.4e-90, 2.7e-84, 9.3e-84]	[1.0e-137, 1.6e-99, 2.0e-93, 5.2e-89, 3.1e-84, 1.6e-83]
		Eqtl NES	[1.1, 1.1, -1.2, 1.2, 1.1, 1.2]		[1.1, 1.1, -1.2, 1.2, 1.2, 1.1]	[1.1, 1.1, -1.2, 1.2, 1.2, 1.2]
		ABC target genes				
		[SNP] archr top gene FDR<=0.05	*CRHR1*		*SPPL2C*	*SPPL2C*
		[SNP] OT top gene	*MAPT*	*KANSL1*	*MAPT*	*MAPT*
Implicated gene(s)			*KANSL1*			

***KANSL1 (*KAT8 regulatory NSL complex subunit 1 protein); Nucleus, mitochondrion**

Role in heart / myopathy

Genes involved in histone modification, such as the *WDR5-MLL1* complex, of which *KANSL1* is a part, are well-established contributors to cardiac development and various forms of congenital heart disease (Jin *et al.*, 2017)(Kulkarni *et al.*, 2018).

Yu *et al.* (2023) findings suggest that common variants which affect the dosage and transcription of *KANSL1* significantly influence the development and growth of the interventricular septum (IVS).

Association with CM

Because genetic risk has yet to be elucidated for cardiac disease involving the interventricular septum (IVS), including HCM, Yu *et al.* (2023) aimed to identify novel associated loci through IVS imaging. The authors found the inheritance of common variants (which either resulted in small versus large IVS) to dictate whether an individual would be at increased risk for ventricular septal defect or HCM. This led to the identification of the common variant, rs62063281, which represents a duplication flanked inversion (linked to IVS cross-sectional area at diastole (IVS.csad)) to result in significant chromatin disruption and so they employed the Akita model for 3D genome architecture to assess the impact. They found a loss of interactions between the *KANSL1* promoter and a few other genes such as *MAPT* and *KANSL1-AS1*. The variant appeared to function as a LV-specific splice QTL for *KANSL1*, aligning with the predicted disruption of the interaction between the *KANSL1* promoter and its regulatory elements as suggested by the AKITA model. Furthermore, genes that interact with *KANSL1-AS1* (including *ARL17B*, *LRRC37A*, *NSFP1*, *LRRC37A2*, *NSF*) were also anticipated to lose this interaction which could result in their altered expression.

Haploinsufficiency or truncating variants in *KANSL1* are also known to cause Koolen-de Vries syndrome which encompasses ventricular septal defects (VSDs).

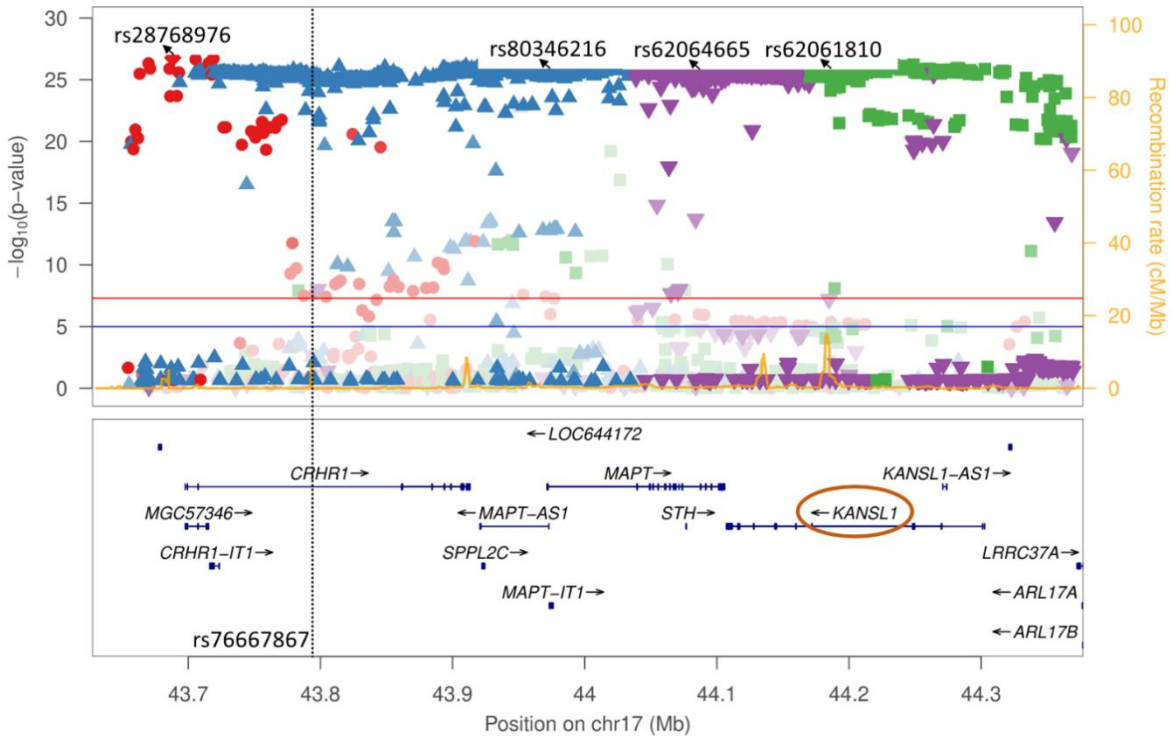


Figure 78 | rs28768976 Locus zoom plot | A locus zoom plot for the TAD-defined locus depicting the variant(s) prioritised (rs80346216, rs62061810, rs62064665, and rs76667867). Circled gene(s) represent plausible targets given all the supporting evidences.

PLCD3 was prioritised given proximity-based locus demarcation with bulk heart LV data and *MAPT* through TAD-demarcation with single-cell cardiomyocyte expression data. *PLCD3* was excluded from the TAD-derived locus and so this reduces the plausibility of it being a target gene. Even though *MAPT* was prioritised through the cardiomyocyte-specific quadrant approach, this was mostly due to its extreme specificity and less so its expression (both *ARL17B* and *KANSL1* showed higher expression with the highest shown by *KANSL1*). The eQTL data output for all variants were in line with one another suggesting numerous potential target genes. However, these were not able to be investigated in too much detail considering the unknown regulatory elements at these sites. The prioritised variants in this locus may affect the expression of multiple target genes – this scenario is alluded to by Yu *et al.* (2023) who identified a common variant residing in this TAD to represent an inversion which results in disrupted interactions (directly implicating *KANSL1* and consequentially other genes in the locus). Given all the supporting information, *KANSL1* is the most likely target gene in this locus, despite not being one of the initially prioritised target genes.

Consensus Example; rs7487962 Locus

Table 37 | rs7487962 Locus evidence table | A table depicting an overview of relevant evidences for the rs7487962 locus and its prioritised variant(s).

Locus-to-gene	**Tadros paper**	**Lead rsid**	rs7487962	
		OT	*TBX3*	
		FUMA	*TBX3*	
	Quadrant approach	**Bulk gene symbol**	*TBX3*	
		Sc gene symbol	*TBX3*	
		Sc gene highest expr in CM		
	Gene prioritisation tools	**PoPs top gene**	*TBX3*	
		PoPs score	0.109	
		Significant colocalisation	N	
		Coloc gene		
		Coloc H4		
		Coloc SNP		
		Coloc SNP H4		
SNP-to-gene	**Deepmind**	**Prioritised proxy rsid**	rs2891537	rs10744836
		Prioritised SNP details	12 114921147 T G	12 114933937 T G

		MTAG p val 5e-8	8.843000e-14	2.696000e-15
		MTAG OR 1	0.844808	0.835335
		Var type	gain	loss
		Top classifier's damage score	0.356	-0.399
		Top classifier	DNase:cardiac muscle cell 1	DNase:cardiac muscle cell 1
	Identification of plausible target genes	**REgulamentary element**		
		Visual inspection of chromatin tracks		
		Eqtl target genes		
		Eqtl heart tissue		
		Eqtl p value		
		Eqtl NES		
		ABC target genes		
		[SNP] archr top gene FDR<=0.05	*TBX3*	
		[SNP] OT top gene	*TBX3*	*TBX3*
Implicated gene(s)			*TBX3*	

***TBX3* (T-box 3 transcription factor); Nucleus**

Role in heart / myopathy

TBX3 is expressed in all elements of the adult cardiac conduction system (CCS), with the exception of the Purkinje fibre network (PFN) (Hoogaars *et al.*, 2004). It is expressed in a continuous myocardial domain ranging from the sinoatrial node to the atrioventricular region (Hoogaars *et al.*, 2004; Mommersteeg *et al.*, 2007). *TBX3* plays a pivotal role in inhibiting the chamber myocardial gene program while promoting the pacemaker gene program and phenotype in a dose-dependent manner (Bakker *et al.*, 2012, 2008; Frank *et al.*, 2012; Hoogaars *et al.*, 2007; Singh *et al.*, 2012). Mohan *et al.* (2018) found that the establishment of the CCS is achieved through the progressive fate restriction of a Tbx3+ cell population during early heart development.

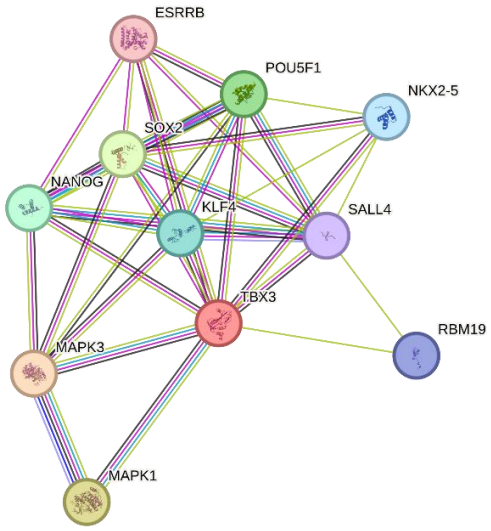


Figure 79 | rs7487962 STRING database interactions| A visualisation by STRING of noted functional relationships given the input list of all plausible genes within the rs7487962 locus whereby a function in the heart has been elucidated.

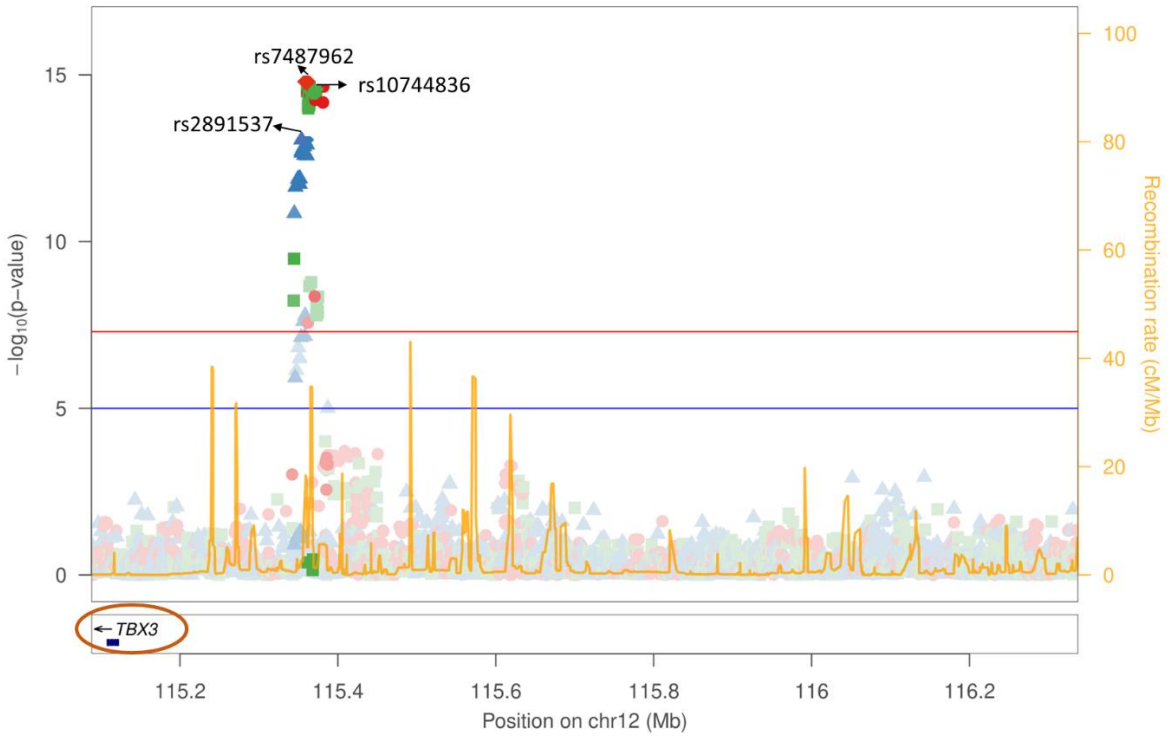


Figure 80 | rs7487962 Locus zoom plot | A locus zoom plot for the TAD-defined locus depicting the variant(s) prioritised (rs2891537 and rs10744836). Circled gene(s) represent plausible targets given all the supporting evidences.

TBX3 was prioritised given proximity-based locus demarcation with bulk heart LV data as well as through TAD-demarcation with single-cell cardiomyocyte expression data. Regarding both quadrant approach analyses, *TBX3* was the only protein coding gene to reside in the locus. *TBX3* has not been associated with cardiomyopathy; however, it has been proven essential in the cardiac conduction system of developing hearts. *TBX3* has been shown to link functionally with *ESRRβ* due to a protein-protein interaction with an average detection confidence of medium (albeit in *Mus musculus* not *Homo sapiens*). Rowe *et al.* (2017) showed knockout *ESRRβ* mice to develop DCM and experience sudden death as a result of increased calcium sensitivity and reduced contractility. Due to the lack of other candidate genes in this locus and due to its established role in the heart, *TBX3* is the most likely target gene in this locus.

5.10 Summary tables

*Table 38 | Concluding target gene(s) per locus wherein variants were prioritised | A table depicting the 53 prioritised HCM variants alongside the prioritised gene as assigned by Open Target and FUMA (based on the lead variant per locus) as well as the concluding target gene(s) per locus (given the prioritised variants residing within and findings per case study). A 'gene weight (/4)'score is assigned given the number of evidences supporting the implicated genes; those included the Tadros (2025) paper, quadrant approach, other gene prioritisation tools, and element-based identification. Those with a gene weight ≥ 2 are shown in the 'implicated gene(s)' column. *An asterisk indicates that the gene weight has been increased by 1 for a particular gene (in a locus with more than one implicated gene) due to the presence of a relevant cardiac phenotype in the Mouse Genome Informatics (MGI) database.*

Lead rsid	**Proxy rsid**	**Snp test id (b38)**	**Locus OT gene**	**Locus FUMA gene**	**Gene weight (/4)**	**Implicated gene(s)**
rs1048302	rs945418	1_16018235_T_C	*HSPB7*	*HSPB7*	*HSPB7: 3*	*HSPB7*
	rs1763603	1_16013314_T_G	*HSPB7*	*HSPB7*	*HSPB7: 3*	*HSPB7*
	rs9442216	1_16026905_T_C	*HSPB7*	*HSPB7*	*HSPB7: 3*	*HSPB7*
rs1159974	rs3734634	6_125790654_G_A	*HEY2*	*HEY2*	*NCOA7: 1*	-
rs116904997	rs116904997	12_120230731_G_A	*PXN*	*PXN*	*PXN: 3*	*PXN*
rs11730129	rs13133886	4_16034172_A_C	*PROM1*	*PROM1*	*PROM1: 4*	*PROM1*
rs11748963	rs3923060	5_139439612_C_T	*PROB1*	*SPATA24*	*PROB1: 3; DNAJC18: 1*	*PROB1*
	rs11748963	5_139394348_T_C	*PROB1*	*SPATA24*	*PROB1: 3; DNAJC18: 1*	*PROB1*
rs12210733	rs10457327	6_118370989_G_C	*PLN*	*SLC35F1*	*PLN: 4*	*PLN*
rs12270374	rs11023178	11_14288736_G_A	*COPB1*	*RRAS2*	*SPON1: 0; RRAS2: 3*	*RRAS2*

rs12460541	rs10402263	19_45810500_G_C	*DMPK*	*SYMPK*	*DMPK: 3*	*DMPK*
rs13297385	rs66925847	9_131640601_ACT_A	*RAPGEF1*	*RAPGEF1*	*RAPGEF1: 2*	*RAPGEF1*
	rs3808802	9_131624672_G_A	*RAPGEF1*	*RAPGEF1*	*RAPGEF1: 2*	*RAPGEF1*
rs2002854	rs4887092	15_78751333_T_C	*ADAMTS7*	*MORF4L1*	*ADAMTS7: 3*	*ADAMTS7*
rs2184370	rs10872646	6_149746539_A_G	*NUP43*	*PCMT1*	*NUP43: 2; LATS1: 1; PCMT1: 3; GINM1: 1*	*NUP43 /PCMT1*
	rs2297932	6_149718225_C_G	*NUP43*	*PCMT1*	*NUP43: 2; LATS1: 1; PCMT1: 3; GINM1: 1*	*NUP43 /PCMT1*
	rs1112729	6_149766491_C_A	*NUP43*	*PCMT1*	*NUP43: 2; LATS1: 1; PCMT1: 3; GINM1: 1*	*NUP43 /PCMT1*
	rs62439782	6_149566308_T_C	*NUP43*	*PCMT1*	*NUP43: 2; LATS1: 1; PCMT1: 3; GINM1: 1*	*NUP43 /PCMT1*
rs2191446	rs4699910	5_57700364_A_T	*ACTBL2*	*ACTBL2*	*PLK2: 2*	*PLK2*
	rs11372647	5_57693080_T_TA	*ACTBL2*	*ACTBL2*	*PLK2: 2*	*PLK2*
	rs6450431	5_57718369_C_T	*ACTBL2*	*ACTBL2*	*PLK2: 2*	*PLK2*
rs2503715	rs2503715	1_2212668_A_G	*FAAP20*	*SKI*	*PRKCZ: 2*	*PRKCZ*
rs28768976	rs80346216	17_45897110_G_T	*PLEKHM1*	*MAPT*	*KANSL1: 1*	-

	rs62061810	17_46101518_G_C	*PLEKHM1*	*MAPT*	*KANSL1: 1*	-
	rs62064665	17_46004161_T_C	*PLEKHM1*	*MAPT*	*KANSL1: 1*	-
	rs76667867	17_45722444_G_C	*PLEKHM1*	*MAPT*	*KANSL1: 1*	-
rs2945236	rs2945236	8_8310872_G_A	*ERI1*	*ERI1*	*MFHAS1: 2; PRAG1: 1*	*MFHAS1*
rs295114	rs295114	2_200330879_C_T	*SPATS2L*	*SPATS2L*	*SPATS2L: 3*	*SPATS2L*
rs35006907	rs35006907	8_124847575_C_A	*MTSS1*	*MTSS1*	*MTSS1: 2*	*MTSS1*
rs3845778	rs888083	2_36889930_C_T	*STRN*	*STRN*	*STRN: 3*	*STRN*
rs4894803	rs4894803	3_172082466_A_G	*FNDC3B*	*FNDC3B*	*FNDC3B: 2; PLD1: 3*	*FNDC3B/PLD1*
	rs11386443	3_172041620_C_CCT	*FNDC3B*	*FNDC3B*	*FNDC3B: 2; PLD1: 3*	*FNDC3B/PLD1*
rs497999	rs566408	18_35085797_C_G	*MAPRE2*	*MAPRE2*	*MAPRE2: 3; DTNA: 2**	*MAPRE2/DTNA*
	rs476348	18_35090057_T_C	*MAPRE2*	*MAPRE2*	*MAPRE2: 3; DTNA: 2**	*MAPRE2/DTNA*
rs5760054	rs5760061	22_23836092_G_A	*VPREB3*	*SMARCB1*	*VPREB3: 2; SMARCB1: 2*	*VPREB3/SMARCB1*
	rs6003909	22_23839465_A_G	*VPREB3*	*SMARCB1*	*VPREB3: 2; SMARCB1: 2*	*VPREB3/SMARCB1*

rs6566955	rs6566955	18_58255557_A_G	*NEDD4L*	*NEDD4L*	*ALPK2: 1; NEDD4L: 3**	*NEDD4L*
rs6699769	rs149015746	1_77683091_CGCCGTCGCA_C	*DNAJB4*	*NEXN*	*NEXN: 3; DNAJB4: 3; ZZZ3: 1*	*NEXN/DNAJB4*
	rs7537908	1_78053255_A_G	*DNAJB4*	*NEXN*	*NEXN: 3; DNAJB4: 3; ZZZ3: 1*	*NEXN/DNAJB4*
rs7487962	rs2891537	12_114921147_T_G	*TBX3*	*TBX3*	*TBX3: 4*	*TBX3*
	rs10744836	12_114933937_T_G	*TBX3*	*TBX3*	*TBX3: 4*	*TBX3*
rs8019721	rs61991243	14_71380400_A_G	*SIPA1L1*	*SIPA1L1*	*SIPA1L1: 3; RGS6: 3**	*SIPA1L1/RGS6*
	rs3825719	14_71586380_G_C	*SIPA1L1*	*SIPA1L1*	*SIPA1L1: 3; RGS6: 3**	*SIPA1L1/RGS6*
	rs3825719	14_71586380_G_A	*SIPA1L1*	*SIPA1L1*	*SIPA1L1: 3; RGS6: 3**	*SIPA1L1/RGS6*
	rs10873246	14_71686976_A_G	*SIPA1L1*	*SIPA1L1*	*SIPA1L1: 3; RGS6: 3**	*SIPA1L1/RGS6*
	rs10873246	14_71686976_A_T	*SIPA1L1*	*SIPA1L1*	*SIPA1L1: 3; RGS6: 3**	*SIPA1L1/RGS6*
	rs7148679	14_71693604_T_C	*SIPA1L1*	*SIPA1L1*	*SIPA1L1: 3; RGS6: 3**	*SIPA1L1/RGS6*
	rs7148679	14_71693604_T_A	*SIPA1L1*	*SIPA1L1*	*SIPA1L1: 3; RGS6: 3**	*SIPA1L1/RGS6*
rs8033459	rs7183401	15_84828713_G_A	*ALPK3*	*NMB*	*ALPK3: 4*	*ALPK3*
	rs7183401	15_84828713_G_T	*ALPK3*	*NMB*	*ALPK3: 4*	*ALPK3*
rs816379	rs616579	6_54105134_G_C	*MLIP*	*MLIP*	*MLIP: 3*	*MLIP*
rs9515201	rs11838776	13_110388334_G_A	*COL4A2*	*COL4A2*	*COL4A2: 4*	*COL4A2*
	rs9515201	13_110388451_A_C	*COL4A2*	*COL4A2*	*COL4A2: 4*	*COL4A2*

Table 38 summarises the concluding target gene(s) per prioritised locus – as determined in this chapter. This entailed assessing the collective evidence from gene prioritisation (chapter 3) as well as variant prioritisation (chapter 4) which both involved employing various bioinformatic tools aimed at addressing the limitations of GWAS discovery. The table depicts each of the 53 prioritised variants which correspond to 28 loci (wherein they reside). Of the 28 loci, a single candidate gene resulted for 18 loci, 2 genes for 8 loci, 3 genes for 1 locus, and 4 genes for a single locus after inspecting loci as demonstrated in the case studies. These genes are listed in the 'Gene weight (/4)' column whereby four evidences collated throughout the thesis (including whether the gene was prioritised in the Tadros (2025) paper, quadrant approach, other gene prioritisation tools, or element-based assessment) were considered for each of these genes. The 'Implicated gene(s)' column depicts the concluding candidate gene(s) for that locus.

The Mouse Genome Informatics (MGI) database and its documented phenotypes (Baldarelli *et al.*, 2024) were consulted for loci in table 38 showing more than one implicated gene to help narrow this down to one causal gene. For the majority of the 10 loci (with multiple implicated genes), the MGI phenotypes were not able to highlight a single likely causal gene due to there either being no evidence of heart-related traits or due to multiple candidates showing heart-related phenotypes. However, for 3 of these loci, a particular gene and its cardiac relevance was strengthened as a result of its phenotype and its gene weight score was increased by one (marked by an asterisk in table 38): DTNA in the rs497999 locus (cardiac fibrosis, cardiomyopathy, and myocardial fiber degeneration), NEDD4L in the rs6566955 locus (decreased cardiac muscle contractility, decreased heart rate, hypertension, and increased heart left ventricle weight), and RGS6 in the rs8019721 locus (abnormal sinoatrial node conduction and decreased heart rate).

Table 39 | PheWAS look-up for prioritised HCM variants | A table depicting 26 of the 53 prioritised HCM variants with genome-wide significant disease phenotype associations in Open Target; these were filtered for a posterior probability (PP) ≥ 0.05 and limited only to the top 5 (maximum) associations regarding their PP value.

Lead rsid	Proxy rsid	Snp test id	Disease phenotype	P-value	PP
rs1048302	rs945418	1_16018235_T_C	['left ventricular ejection fraction measurement', 'hypertrophic cardiomyopathy', 'type 2 diabetes mellitus']	['2.545E-17', '3.000E-17', '3.739E-10']	['7.386E-02', '6.529E-02', '6.159E-02']
rs1048302	rs9442216	1_16026905_T_C	['cardiomyopathy', 'right ventricular ejection fraction measurement']	['3.282E-11', '2.312E-09']	['1.427E-01', '9.787E-02']
rs1159974	rs3734634	6_125790654_G_A	['neuroimaging measurement', 'systolic blood pressure', 'Brugada syndrome']	['7.591E-12', '3.113E-12', '4.594E-57']	['7.584E-02', '7.523E-02', '6.200E-02']
rs116904997	rs116904997	12_120230731_G_A	['Left ventricular mass to end-diastolic volume ratio', 'hypertrophic cardiomyopathy', 'atrial fibrillation', 'cardiac arrhythmia']	['3.000E-10', '1.195E-09', '1.238E-11', '6.998E-09']	['6.863E-01', '6.253E-01', '4.496E-01', '4.416E-01']
rs11730129	rs13133886	4_16034172_A_C	['diastolic blood pressure change measurement', 'electrocardiography', 'pulse pressure measurement', 'QRS duration']	['6.699E-09', '2.000E-09', '1.000E-10', '2.000E-09']	['3.370E-01', '2.201E-01', '2.125E-01', '1.548E-01']

rs11748963	rs3923060	5_139439612_C_T	['cardiomyopathy', 'systolic blood pressure']	['1.853E-09', '2.000E-09']	['2.261E-01', '1.257E-01']
rs11748963	rs11748963	5_139394348_T_C	['electrocardiography', 'hypertrophic cardiomyopathy', 'heart failure']	['1.000E-14', '3.852E-11', '3.000E-09']	['2.807E-01', '1.636E-01', '6.101E-02']
rs12210733	rs10457327	6_118370989_G_C	['heart rate', 'QT interval', 'hypertrophic cardiomyopathy', 'heart valve disease']	['1.337E-88', '6.169E-49', '1.797E-17', '1.290E-15']	['4.189E-01', '3.864E-01', '2.408E-01', '1.162E-01']
rs12460541	rs10402263	19_45810500_G_C	['platelet volume', 'right ventricular systolic volume measurement']	['2.249E-22', '8.079E-11']	['3.332E-01', '2.131E-01']
rs2002854	rs4887092	15_78751333_T_C	['kidney failure']	['1.245E-12']	['5.073E-02']
rs2184370	rs62439782	6_149566308_T_C	['neuroimaging measurement']	['1.466E-12']	['1.353E-01']
rs2191446	rs4699910	5_57700364_A_T	['hypertrophic cardiomyopathy', 'JT interval']	['3.474E-11', '6.000E-27']	['7.247E-02', '5.059E-02']
rs2191446	rs11372647	5_57693080_T_TA	['QT interval']	['3.199E-15']	['1.000E+00']
rs2503715	rs2503715	1_2212668_A_G	['QRS duration', 'JT interval', 'QT interval', 'PR interval']	['8.000E-10', '5.000E-36', '3.000E-24', '1.000E-51']	['1.000E+00', '1.000E+00', '1.000E+00', '1.000E+00']

rs35006907	rs35006907	8_124847575_C_A	['cardiomyopathy', 'atrial fibrillation', 'cardiovascular measurement']	['1.464E-13', '2.320E-09', '1.332E-16']	['7.030E-01', '3.380E-01', '2.441E-01']
rs3845778	rs888083	2_36889930_C_T	['neuroimaging measurement', 'cortical thickness', 'HbA1c measurement']	['9.693E-11', '2.000E-12', '7.563E-11']	['2.823E-01', '1.280E-01', '9.777E-02']
rs4894803	rs4894803	3_172082466_A_G	['heart rate', 'neuroimaging measurement']	['8.397E-30', '3.149E-09']	['9.994E-01', '9.095E-01']
rs497999	rs566408	18_35085797_C_G	['glomerular filtration rate', 'left ventricular structural measurement', 'serum creatinine amount', 'intestinal disease']	['1.253E-08', '2.058E-13', '9.998E-09', '1.487E-08']	['1.293E-01', '1.271E-01', '1.001E-01', '8.949E-02']
rs497999	rs476348	18_35090057_T_C	['left ventricular structural measurement', 'Brugada syndrome', 'magnetic resonance imaging of the heart, electrocardiography', 'electrocardiography']	['6.163E-16', '5.389E-11', '2.000E-10', '2.000E-10']	['5.397E-01', '4.562E-01', '3.700E-01', '3.700E-01']
rs5760054	rs5760061	22_23836092_G_A	['left ventricular systolic function measurement', 'left ventricular ejection fraction measurement', 'heart failure']	['2.075E-13', '6.595E-12', '8.000E-11']	['1.636E-01', '1.584E-01', '1.422E-01']
rs5760054	rs6003909	22_23839465_A_G	['hypertrophic cardiomyopathy']	['2.000E-19']	['8.662E-01']

rs6699769	rs7537908	1_78053255_A_G	['HbA1c measurement', 'glomerular filtration rate', 'QT interval', 'protein measurement']	['2.399E-10', '4.236E-13', '2.000E-09', '6.000E-53']	['9.772E-02', '6.523E-02', '6.166E-02', '6.150E-02']
rs8019721	rs61991243	14_71380400_A_G	['QRS-T angle']	['2.454E-29']	['6.006E-02']
rs8033459	rs7183401	15_84828713_G_T	['PR interval']	['1.000E-17']	['5.320E-02']
rs9515201	rs11838776	13_110388334_G_A	['coronary artery disease', 'neuroimaging measurement', 'diastolic blood pressure', 'cardiovascular disease']	['1.268E-22', '1.049E-12', '1.347E-08', '3.994E-11']	['1.000E+00', '9.923E-01', '9.922E-01', '9.831E-01']
rs9515201	rs9515201	13_110388451_A_C	['stroke', 'hypertrophic cardiomyopathy', 'BMI-adjusted waist-hip ratio', 'small vessel stroke', 'waist-hip ratio']	['8.332E-10', '4.636E-09', '5.000E-15', '2.000E-11', '7.000E-11']	['9.497E-01', '8.230E-01', '6.400E-01', '5.596E-01', '5.443E-01']

Table 39 shows the results after performing a PheWAS look-up for each of the 53 prioritised variants from the Deepmind Enformer analysis. Only 26 of the 53 variants showed diseases/phenotypes passing the defined posterior probability threshold of 5%. To inspect which terms are likely significant in the underlying HCM mechanisms for these suggested functional variants, the frequency at which the terms were captured (across all variants) was explored (shown in figure 81). Annotations occurring most frequently were: hypertrophic cardiomyopathy, neuroimaging measurement, QT interval, cardiomyopathy, and electrocardiography. These terms assign some confidence to the prioritised set of variants considering HCM pathology.

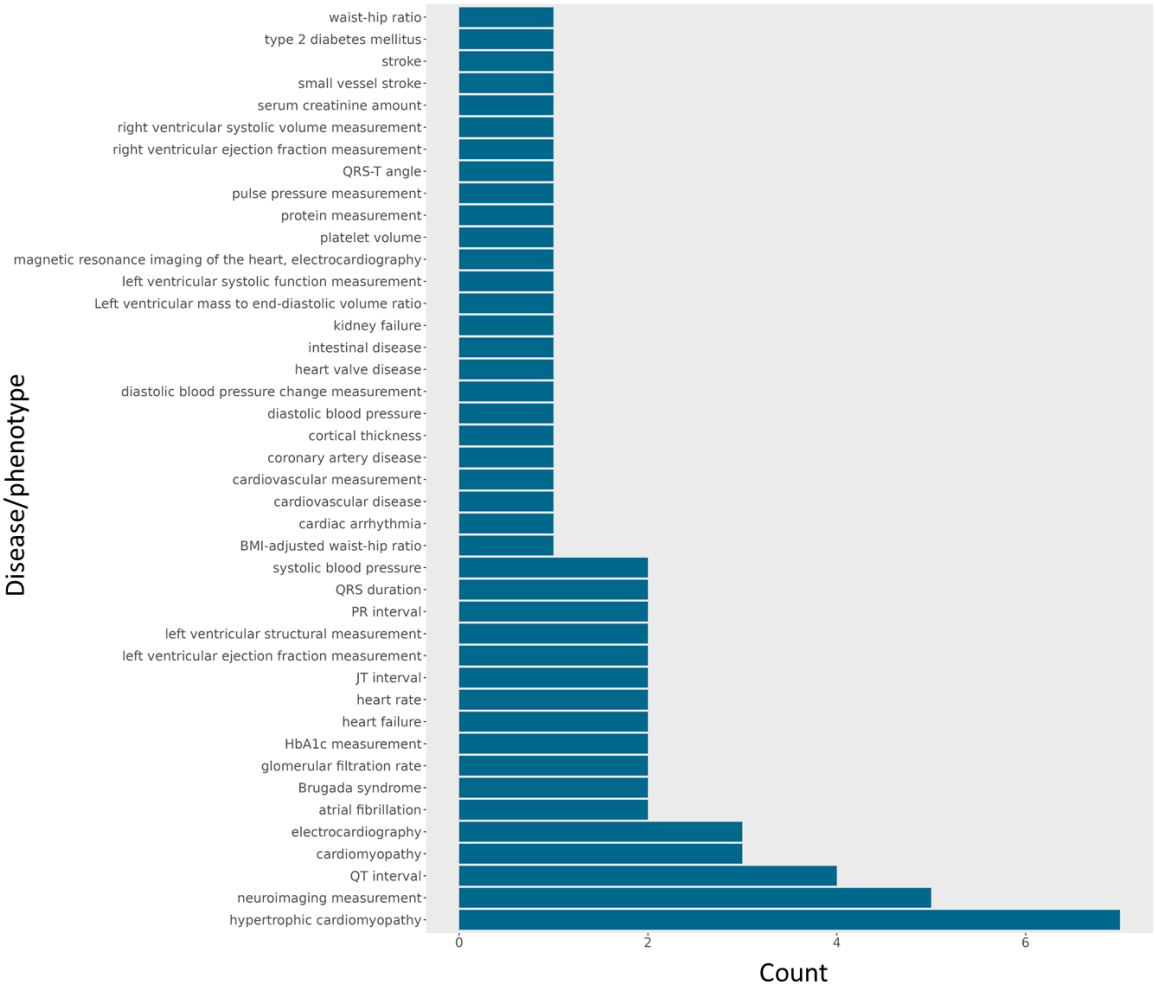


Figure 81 | PheWAS look-up for prioritised HCM variants | A figure illustrating the PheWAS disease/phenotypes extracted from Open Targets for the prioritised functional variants as output by Deepmind's Enformer. 26 of the 53 variants had significant associations and only these are captured in the count.

6

Conclusions

6.1 Introduction

The purpose of this thesis was to discover candidate genes and functional variants which are causal for polygenic HCM. Although GWAS has identified significant common variants, it is challenging to infer how exactly these variants affect disease risk and via which genes. This is especially true when considering that 90% of the genome-wide significant variants reside in noncoding parts of the genome. Noncoding variants are largely understood to regulate gene expression by altering promoter or enhancer activity, or through the disruption of transcription factor binding. Three key objectives were addressed, specifically: 1. prioritise genes (Chapter 3); 2. identify functional variants (Chapter 4); and 3. assess all evidences to define likely loci with prioritised functional variants associated with HCM and plausibly linked prioritised target genes (Chapter 5).

I will discuss the main findings for each of these objectives and consider possible future undertakings.

6.2 Gene prioritisation given HCM GWAS common variants

As presented in Chapter 3, the intersection of hic heart left ventricle contact domain data (and the marked topologically associating domains) was intersected with the HCM lead and proxy variants for the derivation of loci regions per 68 lead HCM variants.

The first attempt at gene prioritisation per locus was via a devised ‘quadrant approach’ method – whereby a combination of gene expression as well as tissue or cell type specificity were considered. Table 12 catalogues 32 genes whereby there was an agreement in the chosen prioritised gene for a particular locus across both quadrant approaches (using bulk heart left ventricle data as well as single-cell cardiomyocyte data). Most of these prioritised genes (28) were shown to be enriched in the cardiomyocytes as opposed to other heart cell types. Table 13 highlights 30 loci whereby the prioritised genes were not unanimously prioritised via both quadrant approaches but were enriched

for cardiomyocytes. 6 loci were shown whereby a different prioritised gene was implicated via both approaches but where the gene prioritised via the single-cell cardiomyocyte approach was most expressed in a non-cardiomyocyte heart cell type (in table 14). This quadrant approach provided insight into one candidate gene per locus given expression data alone.

A colocalisation analysis performed for the 68 loci revealed 20 exhibiting significant associations (whereby both traits were linked and shared a single causal variant with PP.H4.abf≥0.5) when utilising coloc.abf (as shown in table 15).

Both a LoF as well as missense burden test was carried out on two different datasets (GeL and BRRD), followed by a meta-analysis, given an input list of all genes prioritised within the gene prioritisation chapter (168 gene ids). The LoF burden testing revealed a significant burden (considering multiple testing) for *ALPK3* (p-value 1.280e-21) and *SVIL* (p-value 6.231e-06) as shown in table 18. *ALPK3* and *SVIL* have recently been associated with HCM and so these results were expected. The missense testing revealed *MYBPC3* (p-value 4.133e-20), *RAF1* (p-value 6.057e-10), and *ALPK3* (p-value 6.338e-09) as depicted in table 21. *MYBPC3* is known to be associated with HCM through missense variant pathogenicity while *ALPK3* and *RAF1* were investigated further. Plotting variants in the *ALPK3* predicted protein structure revealed little due to the poor structure prediction as well as marginal segregation between case and control variants; however, the case variants appeared to localise to the N-terminus surface. *RAF1* showed a clustering of case variants within domain binding surfaces as opposed to the outward-facing surfaces; however, little can be said about HCM specifically due to the frequently reported concomitance of RAS-HCM.

The existing gene prioritisation tool, PoPs, was employed for all annotated genes in the genome combined with the HCM GWAS variants resulting in a top-scoring gene per each of the 68 loci (depicted in table 22).

Collectively, these approaches aided in prioritising the most likely causal HCM gene in each of the 68 loci after which these genes could be interrogated further.

6.3 Variant prioritisation given HCM GWAS common variants

Chapter 4 entailed use of cell type enrichment prior to data standardisation and input to a machine learning model for impact prediction. Cell type enrichment showed atrial and ventricular cardiomyocytes to be most enriched compared to the other cell types with ventricular cardiomyocytes showing the highest enrichment (~2 fold higher than that shown by atrial cardiomyocytes). This suggested that cardiomyocytes are also responsible for the polygenic pathogenic mechanisms associated with HCM (in addition to monogenic).

Initially, a machine learning model called DeepHaem was employed for functional variant prioritisation. After many attempts, DeepHaem was not able to generalise and due to the rise in machine learning architecture, this approach (based on convolutional neural networks) was abandoned and a model with transformers used instead.

Google's Deepmind Enformer was leveraged for functional variant prioritisation and after a successful performance and validation of the model, 53 unique variants were prioritised out of the initial set of 4395 HCM GWAS lead and proxy variants. This comprised 24 gain and 29 loss variants whereby 9 were lead variants and 44 proxy.

6.4 Bringing the story together: connecting prioritised variants to genes

The underlying regulatory elements at each of the 53 prioritised variants' locations were annotated using REgulamentary – resulting in 14 of these rsids being assigned to known regulatory elements (~33% enhancers, ~33% promoters, and 7% CTCF sites).

Target genes for each of the 53 variants were pulled from the Open Targets database (wherever possible) and this resulted in 33 unique target genes being identified (for 47 of the 53 rsids present in the database).

Among the 53 prioritised variants, 14 were found to overlap with an enhancer site using ABC method. Of which, 12 showed significant connections linking with 36 genes.

Peak-to-gene links were determined using ArchR and ventricular cardiomyocyte data whereby peaks intersected with any of the 53 prioritised variants. Only 3 rsids were shown to intersect with a peak scoring above the recommended significance threshold.

Functional annotation clustering via DAVID, given an input list of all genes prioritised in chapters 3 and 4 (totalling 195 genes) in addition to Kegg pathways, resulted in 1 cluster whereby all captured pathways had a significant p-value (≤ 0.05). Additionally, one of the terms in this cluster proved significant given multiple testing: the Cytoskeleton in muscle cells pathway (p-value 0.003) suggesting that the cytoskeleton and genes involved should be inspected regarding their role in HCM (in adding to the already established sarcomere genes).

Ultimately, all of the information collated in chapters 3 (gene prioritisation), 4 (variant prioritisation), and 5 (linking variants to likely genes) was used to derive possible causal scenarios for the 28 loci in which the 53 variants resided. These implicated genes are shown in table 38. For the 28 loci, 2 of these could not be mapped to likely causal gene(s) due to a low gene weight being assigned to the residing genes. Overall, I linked 53 variants from 28 loci to 32 genes with 20 loci mapping to 1 gene and 6 to two genes.

6.5 Future directions

This thesis presents a body of work that has aimed to pinpoint genes and variants worthy of further interrogation regarding a plausible causal role in HCM pathogenesis. Hypertrophic cardiomyopathy (HCM) is a widespread, devastating genetic heart disorder frequently resulting in sudden cardiac death. Currently, best practise involves testing for rare, sarcomere variants – which only account for the minority of HCM patients. In this thesis, through both gene and variant prioritisation techniques, a handful of common variants and likely genes have been shortlisted for confirmatory testing regarding a causal role in HCM causality. Following rigorous testing, in the event some are

confirmed to be causal, these findings may influence clinical diagnostic approaches as well as target identification for development of novel therapies. The findings will also enhance our understanding of HCM pathology as influenced by common genetic variants.

It is evident that future computational work can be performed to further investigate these loci in addition to functionally testing whether the prioritised variants and genes are actually causal. It is foreseeable that future statistical genetic efforts will include the recruitment of additional HCM case/control series – stratified by sarcomere status – to generate additional discovery power so that meta-analyses can refine effect size estimates and identify additional loci.

Testing the role of prioritised genes and common variants by CRISPRing-in the causal variant on cardiomyocyte physiology and function is underway. Due to the fact that noncoding variants are highly specific to a given circumstance and cell type, functional validation of such variants requires a biologically faithful environment/experimental design. Also, given that common variants possess modest odds ratios, it is assumed that their biological effects would also be moderate and so quantitative approaches that can be scaled will be needed for assessing the effects of the prioritised common variants, for example in iPSCCM lines. An alternative strategy, given the anticipated small effect sizes of the common variants, is to knock down or over-express the implicated gene – to achieve bigger biological impacts. Where rare coding variants are also identified, functional analyses that explore whether the variants influence the quantity and/or functionality of the protein could give important insights (for example, for the newly recognised ALPK3 missense variant burden).

References

- Ahn, J., Wu, H., & Lee, K. (2020). Integrative Analysis Revealing Human Heart-Specific Genes and Consolidating Heart-Related Phenotypes. Frontiers in Genetics, 11. https://doi.org/10.3389/fgene.2020.00777
- Alex, L., Tuleta, I., Hernandez, S. C., Hanna, A., Venugopal, H., Astorkia, M., Humeres, C., Kubota, A., Su, K., Zheng, D., & Frangogiannis, N. G. (2023). Cardiac Pericytes Acquire a Fibrogenic Phenotype and Contribute to Vascular Maturation after Myocardial Infarction. Circulation, 148(11). https://doi.org/10.1161/CIRCULATIONAHA.123.064155
- Alfares, A. A., Kelly, M. A., McDermott, G., Funke, B. H., Lebo, M. S., Baxter, S. B., Shen, J., McLaughlin, H. M., Clark, E. H., Babb, L. J., Cox, S. W., Depalma, S. R., Ho, C. Y., Seidman, J. G., Seidman, C. E., & Rehm, H. L. (2015). Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: Expanded panels offer limited additional sensitivity. Genetics in Medicine, 17(11). https://doi.org/10.1038/gim.2014.205
- Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 33(8). https://doi.org/10.1038/nbt.3300
- Aljeaid, D., Sanchez, A. I., Wakefield, E., Chadwell, S. E., Moore, N., Prada, C. E., & Zhang, W. (2019). Prevalence of pathogenic and likely pathogenic variants in the RASopathy genes in patients who have had panel testing for cardiomyopathy. American Journal of Medical Genetics, Part A, 179(4). https://doi.org/10.1002/ajmg.a.61072
- Al-Yacoub, N., Colak, D., Mahmoud, S. A., Hammonds, M., Muhammed, K., Al-Harazi, O., Assiri, A. M., Al-Buraiki, J., Al-Habeeb, W., & Poizat, C. (2021). Mutation in FBXO32 causes dilated cardiomyopathy through up-regulation of ER-stress mediated apoptosis. Communications Biology, 4(1). https://doi.org/10.1038/s42003-021-02391-9
- Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-45839-z
- Aragam, K. G., Jiang, T., Goel, A., Kanoni, S., Wolford, B. N., Atri, D. S., Weeks, E. M., Wang, M., Hindy, G., Zhou, W., Grace, C., Roselli, C., Marston, N. A., Kamanu, F. K., Surakka, I., Venegas, L. M., Sherliker, P., Koyama, S., Ishigaki, K., … von Scheidt, M. (2022). Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. Nature Genetics, 54(12). https://doi.org/10.1038/s41588-022-01233-6
- Aung, N., Lopes, L. R., van Duijvenboden, S., Harper, A. R., Goel, A., Grace, C., Ho, C. Y., Weintraub, W. S., Kramer, C. M., Neubauer, S., Watkins, H. C., Petersen, S. E., & Munroe, P. B. (2023). Genome-Wide Analysis of Left Ventricular Maximum Wall Thickness in the UK Biobank Cohort Reveals a Shared Genetic Background with Hypertrophic Cardiomyopathy. Circulation: Genomic and Precision Medicine, 16(1). https://doi.org/10.1161/CIRCGEN.122.003716
- Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18(10). https://doi.org/10.1038/s41592-021-01252-x

- Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., Taylor, K. R., Ward, T., Bycroft, C., Nicolaisen, L., Arvaniti, E., Pan, J., Thomas, R., Dutordoir, V., Perino, M., De, S., Karollus, A., Gayoso, A., Sargeant, T., Mottram, A., Wong, L. H., … Kohli, P. (2025). AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. BioRxiv, 2025.06.25.661532. https://doi.org/10.1101/2025.06.25.661532
- Bagnall, R. D., Ingles, J., Dinger, M. E., Cowley, M. J., Ross, S. B., Minoche, A. E., Lal, S., Turner, C., Colley, A., Rajagopalan, S., Berman, Y., Ronan, A., Fatkin, D., & Semsarian, C. (2018). Whole Genome Sequencing Improves Outcomes of Genetic Testing in Patients With Hypertrophic Cardiomyopathy. Journal of the American College of Cardiology, 72(4). https://doi.org/10.1016/j.jacc.2018.04.078
- Bakker, M. L., Boink, G. J. J., Boukens, B. J., Verkerk, A. O., van den Boogaard, M., den Haan, A. D., Hoogaars, W. M. H., Buermans, H. P., de Bakker, J. M. T., Seppen, J., Tan, H. L., Moorman, A. F. M., T Hoen, P. A. C., & Christoffels, V. M. (2012). T-box transcription factor TBX3 reprogrammes mature cardiac myocytes into pacemaker-like cells. Cardiovascular Research, 94(3). https://doi.org/10.1093/cvr/cvs120
- Bakker, M. L., Boukens, B. J., Mommersteeg, M. T. M., Brons, J. F., Wakker, V., Moorman, A. F. M., & Christoffels, V. M. (2008). Transcription factor Tbx3 is required for the specification of the atrioventricular conduction system. Circulation Research, 102(11). https://doi.org/10.1161/CIRCRESAHA.107.169565
- Baldarelli, R. M., Smith, C. L., Ringwald, M., Richardson, J. E., Bult, C. J., & Group, M. G. I. (2024). Mouse Genome Informatics: an integrated knowledgebase system for the laboratory mouse. Genetics, 227(1), iyae031. https://doi.org/10.1093/genetics/iyae031
- Barshir, R., Shwartz, O., Smoly, I. Y., & Yeger-Lotem, E. (2014). Comparative Analysis of Human Tissue Interactomes Reveals Factors Leading to Tissue-Specific Manifestation of Hereditary Diseases. PLoS Computational Biology, 10(6). https://doi.org/10.1371/journal.pcbi.1003632
- Holwerda, S. J., & de Laat, W. (2013). CTCF: The protein, the binding partners, the binding sites and their chromatin loops. In Philosophical Transactions of the Royal Society B: Biological Sciences (Vol. 368, Issue 1620). https://doi.org/10.1098/rstb.2012.0369
- Bhattacharyya, S., & Ay, F. (2024). Identifying genetic variants associated with chromatin looping and genome function. Nature Communications, 15(1), 8174. https://doi.org/10.1038/s41467-024-52296-4
- Bian, R., Zhang, L., Li, D., & Xu, X. (2024). CDKN1A as a target of senescence in heart failure: insights from a multiomics study. Frontiers in Pharmacology, Volume 15-2024. https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2024.1446300
- Bintu, B., Mateo, L. J., Su, J. H., Sinnott-Armstrong, N. A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A. N., & Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science, 362(6413). https://doi.org/10.1126/science.aau1783
- Brasset, E., & Vaury, C. (2005). Insulators are fundamental components of the eukaryotic genomes. In Heredity (Vol. 94, Issue 6). https://doi.org/10.1038/sj.hdy.6800669
- Broadway-Stringer, S., Jiang, H., Wadmore, K., Hooper, C., Douglas, G., Steeples, V., Azad, A. J., Singer, E., Reyat, J. S., Galatik, F., Ehler, E., Bennett, P., Kalisch-Smith, J. I., Sparrow, D. B., Davies, B., Djinovic-Carugo, K., Gautel, M., Watkins, H., & Gehmlich, K. (2023). Insights into the Role of a Cardiomyopathy-Causing Genetic Variant in ACTN2. Cells, 12(5). https://doi.org/10.3390/cells12050721
- Brodie, A., Azaria, J. R., & Ofran, Y. (2016). How far from the SNP may the causative genes be? Nucleic Acids Research, 44(13). https://doi.org/10.1093/nar/gkw500

- Brown, T. A. (2002). Chapter 7: Understanding a Genome Sequence. In Genomes (Vol. 7).
- Burke, M. A., Cook, S. A., Seidman, J. G., & Seidman, C. E. (2016). Clinical and Mechanistic Insights Into the Genetics of Cardiomyopathy. In Journal of the American College of Cardiology (Vol. 68, Issue 25). https://doi.org/10.1016/j.jacc.2016.08.079
- Calcagni, G., Adorisio, R., Martinelli, S., Grutter, G., Baban, A., Versacci, P., Digilio, M. C., Drago, F., Gelb, B. D., Tartaglia, M., & Marino, B. (2018). Clinical Presentation and Natural History of Hypertrophic Cardiomyopathy in RASopathies. In Heart Failure Clinics (Vol. 14, Issue 2). https://doi.org/10.1016/j.hfc.2017.12.005
- Calo, E., & Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? In Molecular Cell (Vol. 49, Issue 5). https://doi.org/10.1016/j.molcel.2013.01.038
- Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. In Frontiers in Genetics (Vol. 11). https://doi.org/10.3389/fgene.2020.00424
- Castellana, S., Biagini, T., Parca, L., Petrizzelli, F., Bianco, S. D., Vescovi, A. L., Carella, M., & Mazza, T. (2021). A comparative benchmark of classic DNA motif discovery tools on synthetic data. In Briefings in Bioinformatics (Vol. 22, Issue 6). https://doi.org/10.1093/bib/bbab303
- Chandra, O., Sharma, M., Pandey, N., Jha, I. P., Mishra, S., Kong, S. L., & Kumar, V. (2023). Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes. Computational and Structural Biotechnology Journal, 21. https://doi.org/10.1016/j.csbj.2023.07.014
- Chang, L. H., Ghosh, S., & Noordermeer, D. (2020). TADs and Their Borders: Free Movement or Building a Wall? In Journal of Molecular Biology (Vol. 432, Issue 3). https://doi.org/10.1016/j.jmb.2019.11.025
- Charoensawan, V., Janga, S. C., Bulyk, M. L., Babu, M. M., & Teichmann, S. A. (2012). DNA Sequence Preferences of Transcriptional Activators Correlate More Strongly than Repressors with Nucleosomes. Molecular Cell, 47(2). https://doi.org/10.1016/j.molcel.2012.06.028
- Chawla, A., Nagy, C., & Turecki, G. (2021). Chromatin profiling techniques: Exploring the chromatin environment and its contributions to complex traits. In International Journal of Molecular Sciences (Vol. 22, Issue 14). https://doi.org/10.3390/ijms22147612
- Chen, J. F., Murchison, E. P., Tang, R., Callis, T. E., Tatsuguchi, M., Deng, Z., Rojas, M., Hammond, S. M., Schneider, M. D., Selzman, C. H., Meissner, G., Patterson, C., Hannon, G. J., & Wang, D. Z. (2008a). Targeted deletion of Dicer in the heart leads to dilated cardiomyopathy and heart failure. Proceedings of the National Academy of Sciences of the United States of America, 105(6). https://doi.org/10.1073/pnas.0710228105
- Chen, R., Morgan, A. A., Dudley, J., Deshpande, T., Li, L., Kodama, K., Chiang, A. P., & Butte, A. J. (2008b). FitSNPs: Highly differentially expressed genes are more likely to have variants associated with disease. Genome Biology, 9(12). https://doi.org/10.1186/gb-2008-9-12-r170
- Chin, T. K., Perloff, J. K., Williams, R. G., Jue, K., & Mohrmann, R. (1990). Isolated noncompaction of left ventricular myocardium: A study of eight cases. Circulation, 82(2). https://doi.org/10.1161/01.cir.82.2.507
- Chiu, C., Bagnall, R. D., Ingles, J., Yeates, L., Kennerson, M., Donald, J. A., Jormakka, M., Lind, J. M., & Semsarian, C. (2010). Mutations in Alpha-Actinin-2 Cause Hypertrophic Cardiomyopathy. A Genome-Wide Analysis. Journal of the American College of Cardiology, 55(11). https://doi.org/10.1016/j.jacc.2009.11.016

- Choe, D., Burke, M., Brandimarto, J. A., Marti-Pamies, I., Yob, J., Yang, Y., Morley, M. P., Drivas, T. G., Day, S., Damrauer, S., Wang, X., & Cappola, T. P. (2024). Sex-Specific Effect of MTSS1 Downregulation on Dilated Cardiomyopathy. MedRxiv, 2024.02.28.24303451. https://doi.org/10.1101/2024.02.28.24303451
- Chua, E. H. Z., Yasar, S., & Harmston, N. (2022). The importance of considering regulatory domains in genome-wide analyses - the nearest gene is often wrong! In Biology open (Vol. 11, Issue 4). https://doi.org/10.1242/bio.059091
- Cirino, A. L., Harris, S., Lakdawala, N. K., Michels, M., Olivotto, I., Day, S. M., Abrams, D. J., Charron, P., Caleshu, C., Semsarian, C., Ingles, J., Rakowski, H., Judge, D. P., & Ho, C. Y. (2017). Role of genetic testing in inherited cardiovascular disease: A review. In JAMA Cardiology (Vol. 2, Issue 10). https://doi.org/10.1001/jamacardio.2017.2352
- Clausen, A. G., Vad, O. B., Andersen, J. H., & Olesen, M. S. (2021). Loss-of-Function Variants in the SYNPO2L Gene Are Associated With Atrial Fibrillation. Frontiers in Cardiovascular Medicine, 8. https://doi.org/10.3389/fcvm.2021.650667
- Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America, 107(50). https://doi.org/10.1073/pnas.1016071107
- Crilley, J. G., Boehm, E. A., Blair, E., Rajagopalan, B., Blamire, A. M., Styles, P., McKenna, W. J., Östman-Smith, I., Clarke, K., & Watkins, H. (2003). Hypertrophic cardiomyopathy due to sarcomeric gene mutations is characterized by impaired energy metabolism irrespective of the degree of hypertrophy. Journal of the American College of Cardiology, 41(10). https://doi.org/10.1016/S0735-1097(02)03009-7
- Crowell, H. L., Soneson, C., Germain, P. L., Calini, D., Collin, L., Raposo, C., Malhotra, D., & Robinson, M. D. (2020). muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-19894-4
- Crump, N. T., Smith, A. L., Godfrey, L., Dopico-Fernandez, A. M., Denny, N., Harman, J. R., Hamley, J. C., Jackson, N. E., Chahrour, C., Riva, S., Rice, S., Kim, J., Basrur, V., Fermin, D., Elenitoba-Johnson, K., Roeder, R. G., Allis, C. D., Roberts, I., Roy, A., … Milne, T. A. (2023). MLL-AF4 cooperates with PAF1 and FACT to drive high-density enhancer interactions in leukemia. Nature Communications, 14(1). https://doi.org/10.1038/s41467-023-40981-9
- Currin, K. W., Erdos, M. R., Narisu, N., Rai, V., Vadlamudi, S., Perrin, H. J., Idol, J. R., Yan, T., Albanus, R. D. O., Broadaway, K. A., Etheridge, A. S., Bonnycastle, L. L., Orchard, P., Didion, J. P., Chaudhry, A. S., Barnabas, B. B., Black, S., Bouffard, G. G., Brooks, S. Y., … Mohlke, K. L. (2021). Genetic effects on liver chromatin accessibility identify disease regulatory variants. American Journal of Human Genetics, 108(7). https://doi.org/10.1016/j.ajhg.2021.05.001
- de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Computational Biology, 11(4). https://doi.org/10.1371/journal.pcbi.1004219
- de Marvao, A., McGurk, K. A., Zheng, S. L., Thanaj, M., Bai, W., Duan, J., Biffi, C., Mazzarotto, F., Statton, B., Dawes, T. J. W., Savioli, N., Halliday, B. P., Xu, X., Buchan, R. J., Baksi, A. J., Quinlan, M., Tokarczuk, P., Tayal, U., Francis, C., … O'Regan, D. P. (2021). Phenotypic Expression and Outcomes in Individuals With Rare Genetic Variants

of Hypertrophic Cardiomyopathy. Journal of the American College of Cardiology, 78(11). https://doi.org/10.1016/j.jacc.2021.07.017

- Dierck, F., Kuhn, C., Rohr, C., Hille, S., Braune, J., Sossalla, S., Molt, S., van der Ven, P. F. M., Fürst, D. O., & Frey, N. (2017). The novel cardiac Z-disc protein CEFIP regulates cardiomyocyte hypertrophy by modulating calcineurin signaling. In Journal of Biological Chemistry (Vol. 292, Issue 37). https://doi.org/10.1074/jbc.M117.786764
- Dimitrow, P. P., Czarnecka, D., Jaszcz, K. K., & Dubiel, J. S. (1997). Sex differences in age at onset of symptoms in patients with hypertrophic cardiomyopathy. European Journal of Cardiovascular Prevention & Rehabilitation, 4(1). https://doi.org/10.1177/174182679700400106
- Ding, Y., Yang, J., Chen, P., Lu, T., Jiao, K., Tester, D. J., Giudicessi, J. R., Jiang, K., Ackerman, M. J., Li, Y., Wang, D. W., Lee, H. C., Wang, D. W., & Xu, X. (2020). Knockout of sorbs2 protein disrupts the structural integrity of intercalated disc and manifests features of arrhythmogenic cardiomyopathy. Journal of the American Heart Association, 9(17). https://doi.org/10.1161/JAHA.119.017055
- Graham, J. D., & Clarke, C. L. (2002). Expression and transcriptional activity of progesterone receptor A and progesterone receptor B in mammalian cells. In Breast Cancer Research (Vol. 4, Issue 5). https://doi.org/10.1186/bcr450
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398). https://doi.org/10.1038/nature11082
- Downes, D. J., Smith, A. L., Karpinska, M. A., Velychko, T., Rue-Albrecht, K., Sims, D., Milne, T. A., Davies, J. O. J., Oudelaar, A. M., & Hughes, J. R. (2022). Capture-C: a modular and flexible approach for high-resolution chromosome conformation capture. In Nature Protocols (Vol. 17, Issue 2). https://doi.org/10.1038/s41596-021-00651-w
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414). https://doi.org/10.1038/nature11247
- Dunn, K. E. (2024). Why Ancestral Diversity in Genetic Testing Matters: Lessons Learned From Hypertrophic Cardiomyopathy. Circulation: Heart Failure, 17(3), e011484. https://doi.org/10.1161/CIRCHEARTFAILURE.124.011484
- Dutton, L. C., Crosland, A., Dukes-McEwan, J., & Connolly, D. J. (2025). Association of the TTN, PDK4, and RNF207 mutations with dilated cardiomyopathy in Dobermanns from the United Kingdom. PLOS ONE, 20(3), e0319932-. https://doi.org/10.1371/journal.pone.0319932
- Elliott, P. M., Poloniecki, J., Dickie, S., Sharma, S., Monserrat, L., Varnava, A., Mahon, N. G., & McKenna, W. J. (2000). Sudden death in hypertrophic cardiomyopathy: Identification of high risk patients. Journal of the American College of Cardiology, 36(7). https://doi.org/10.1016/S0735-1097(00)01003-2
- Emerson, J. I., Shi, W., Paredes-Larios, J., Walker, W. G., Hutton, J. E., Cristea, I. M., Marzluff, W. F., & Conlon, F. L. (2025). X-Chromosome–Linked miRNAs Regulate Sex Differences in Cardiac Physiology. Circulation Research, 136(3), 258–275. https://doi.org/10.1161/CIRCRESAHA.124.325447
- Erdmann, J., Daehmlow, S., Wischke, S., Senyuva, M., Werner, U., Raible, J., Tanis, N., Dyachenko, S., Hummel, M., Hetzer, R., & Regitz-Zagrosek, V. (2003). Mutation spectrum in a large cohort of unrelated consecutive patients with hypertrophic

cardiomyopathy. Clinical Genetics, 64(4). https://doi.org/10.1034/j.1399-0004.2003.00151.x

- Feiglin, A., Allen, B. K., Kohane, I. S., & Kong, S. W. (2017). Comprehensive Analysis of Tissue-wide Gene Expression and Phenotype Data Reveals Tissues Affected in Rare Genetic Disorders. Cell Systems, 5(2). https://doi.org/10.1016/j.cels.2017.06.016
- Felsenfeld, G., Burgess-Beusse, B., Farrell, C., Gaszner, M., Ghirlando, R., Huang, S., Jin, C., Litt, M., Magdinier, F., Mutskov, V., Nakatani, Y., Tagami, H., West, A., & Yusufzai, T. (2004). Chromatin boundaries and chromatin domains. Cold Spring Harbor Symposia on Quantitative Biology, 69. https://doi.org/10.1101/sqb.2004.69.245
- Frank, D. U., Carter, K. L., Thomas, K. R., Burr, R. M., Bakker, M. L., Coetzee, W. A., Tristani-Firouzi, M., Bamshad, M. J., Christoffels, V. M., & Moon, A. M. (2012). Lethal arrhythmias in Tbx3-deficient mice reveal extreme dosage sensitivity of cardiac conduction system function and homeostasis. Proceedings of the National Academy of Sciences of the United States of America, 109(3). https://doi.org/10.1073/pnas.1115165109
- Fudenberg, G., Kelley, D. R., & Pollard, K. S. (2020). Predicting 3D genome folding from DNA sequence with Akita. Nature Methods, 17(11). https://doi.org/10.1038/s41592-020-0958-x
- Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., & Engreitz, J. M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. In Nature Genetics (Vol. 51, Issue 12). https://doi.org/10.1038/s41588-019-0538-0
- Fullenkamp, D. E., Jorgensen, R. M., Leach, D. F., Sinha, A., Salamone, I. M., Johnston, J. R., Dellefave-Castillo, L. M., Choudhury, L., McNally, E. M., & Wilsbacher, L. D. (2023). Hypertrophic Cardiomyopathy Secondary to RAF1 Cysteine-Rich Domain Variants. Circulation: Genomic and Precision Medicine, 16(6). https://doi.org/10.1161/CIRCGEN.123.004262
- Gao, J., Meng, T., Li, M., Du, R., Ding, J., Li, A., Yu, S., Li, Y., & He, Q. (2022). Global trends and frontiers in research on coronary microvascular dysfunction: a bibliometric analysis from 2002 to 2022. In European Journal of Medical Research (Vol. 27, Issue 1). https://doi.org/10.1186/s40001-022-00869-8
- Gao, Z., Smith, A. L., Scott, J. N. F., Bevington, S. L., & Boyes, J. (2023). Temporal analyses reveal a pivotal role for sense and antisense enhancer RNAs in coordinate immunoglobulin lambda locus activation. Nucleic Acids Research, 51(19). https://doi.org/10.1093/nar/gkad741
- García-González, J., Cote, A. C., Garcia-Gonzalez, S., Liou, L., & O'Reilly, P. F. (2025). The Gene Expression Landscape of Disease Genes. MedRxiv, 2024.06.20.24309121. https://doi.org/10.1101/2024.06.20.24309121
- García-Hernández, S., de la Higuera Romero, L., Ochoa, J. P., & McKenna, W. J. (2024). Emerging Themes in Genetics of Hypertrophic Cardiomyopathy: Current Status and Clinical Application. In Canadian Journal of Cardiology (Vol. 40, Issue 5). https://doi.org/10.1016/j.cjca.2024.01.011
- Garnier, S., Harakalova, M., Weiss, S., Mokry, M., Regitz-Zagrosek, V., Hengstenberg, C., Cappola, T. P., Isnard, R., Arbustini, E., Cook, S. A., van Setten, J., Calis, J. J. A., Hakonarson, H., Morley, M. P., Stark, K., Prasad, S. K., Li, J., O'Regan, D. P., Grasso, M., … Charron, P. (2021). Genome-wide association analysis in dilated cardiomyopathy

reveals two new players in systolic heart failure on chromosomes 3p25.1 and 22q11.23. European Heart Journal, 42(20). https://doi.org/10.1093/eurheartj/ehab030

- Geisterfer-Lowrance, A. A. T., Kass, S., Tanigawa, G., Vosberg, H. P., McKenna, W., Seidman, C. E., & Seidman, J. G. (1990). A molecular basis for familial hypertrophic cardiomyopathy: A β cardiac myosin heavy chain gene missense mutation. Cell, 62(5). https://doi.org/10.1016/0092-8674(90)90274-I
- Gelb, B. D., Roberts, A. E., & Tartaglia, M. (2015). Cardiomyopathies in Noonan syndrome and the other RASopathies. In Progress in Pediatric Cardiology (Vol. 39, Issue 1). https://doi.org/10.1016/j.ppedcard.2015.01.002
- Gerasimavicius, L., Livesey, B. J., & Marsh, J. A. (2022). Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. Nature Communications, 13(1). https://doi.org/10.1038/s41467-022-31686-6
- Gersh, B. J., Maron, B. J., Bonow, R. O., Dearani, J. A., Fifer, M. A., Link, M. S., Naidu, S. S., Nishimura, R. A., Ommen, S. R., Rakowski, H., Seidman, C. E., Towbin, J. A., Udelson, J. E., Yancy, C. W., Jacobs, A. K., Smith, S. C., Anderson, J. L., Albert, N. M., Buller, C. E., … Tarkington, L. G. (2011). 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy. Journal of Thoracic and Cardiovascular Surgery, 142(6). https://doi.org/10.1016/j.jtcvs.2011.10.020
- Geske, J. B., Ommen, S. R., & Gersh, B. J. (2018). Hypertrophic Cardiomyopathy: Clinical Update. In JACC: Heart Failure (Vol. 6, Issue 5). https://doi.org/10.1016/j.jchf.2018.02.010
- Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E. M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., Burdett, T., Hayhurst, J., Baker, J., Ferrer, J., Gonzalez-Uriarte, A., Jupp, S., Karim, M. A., Koscielny, G., MacHlitt-Northen, S., … Dunham, I. (2021). Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Research, 49(D1). https://doi.org/10.1093/nar/gkaa840
- Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genetics, 10(5). https://doi.org/10.1371/journal.pgen.1004383
- Gibson, G. (2012). Rare and common variants: Twenty arguments. In Nature Reviews Genetics (Vol. 13, Issue 2). https://doi.org/10.1038/nrg3118
- Girolami, F., Iascone, M., Tomberli, B., Bardi, S., Benelli, M., Marseglia, G., Pescucci, C., Pezzoli, L., Sana, M. E., Basso, C., Marziliano, N., Merlini, P. A., Fornaro, A., Cecchi, F., Torricelli, F., & Olivotto, I. (2014). Novel α-actinin 2 variant associated with familial hypertrophic cardiomyopathy and juvenile atrial arrhythmias: A massively parallel sequencing study. Circulation: Cardiovascular Genetics, 7(6). https://doi.org/10.1161/CIRCGENETICS.113.000486
- Glicksberg, B. S., Amadori, L., Akers, N. K., Sukhavasi, K., Franzén, O., Li, L., Belbin, G. M., Akers, K. L., Shameer, K., Badgeley, M. A., Johnson, K. W., Readhead, B., Darrow, B. J., Kenny, E. E., Betsholtz, C., Ermel, R., Skogsberg, J., Ruusalepp, A., Schadt, E. E., … Chen, R. (2019). Integrative analysis of loss-of-function variants in clinical and genomic data reveals novel genes associated with cardiovascular traits. BMC Medical Genomics, 12. https://doi.org/10.1186/s12920-019-0542-3
- Gordon, A. M., Homsher, E., & Regnier, M. (2000). Regulation of contraction in striated muscle. In Physiological Reviews (Vol. 80, Issue 2). https://doi.org/10.1152/physrev.2000.80.2.853

- Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., & Greenleaf, W. J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature Genetics, 53(3). https://doi.org/10.1038/s41588-021-00790-6
- Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L., & Wright, C. F. (2021). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. Journal of Medical Genetics, 58(8). https://doi.org/10.1136/jmedgenet-2020-107003
- Gwathmey, J. K., Warren, S. E., Maurice Briggs, G., Copelas, L., Feldman, M. D., Phillips, P. J., Callahan, M., Schoen, F. J., Grossman, W., & Morgan, J. P. (1991). Diastolic dysfunction in hypertrophic cardiomyopathy: Effect on active force generation during systole. Journal of Clinical Investigation, 87(3). https://doi.org/10.1172/JCI115061
- Han, S. K., Muto, Y., Wilson, P. C., Humphreys, B. D., Sampson, M. G., Chakravarti, A., & Lee, D. (2022). Quality assessment and refinement of chromatin accessibility data using a sequence-based predictive model. Proceedings of the National Academy of Sciences of the United States of America, 119(51). https://doi.org/10.1073/pnas.2212810119
- Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-Sequencing for biomedical research and clinical applications. In Genome Medicine (Vol. 9, Issue 1). https://doi.org/10.1186/s13073-017-0467-4
- Harper, A. R., Goel, A., Grace, C., Thomson, K. L., Petersen, S. E., Xu, X., Waring, A., Ormondroyd, E., Kramer, C. M., Ho, C. Y., Neubauer, S., Kolm, P., Kwong, R., Dolman, S. F., Desvigne-Nickens, P., Dimarco, J. P., Geller, N., Kim, D. Y., Zhang, C., … Watkins, H. (2021). Common genetic variants and modifiable risk factors underpin hypertrophic cardiomyopathy susceptibility and expressivity. Nature Genetics, 53(2). https://doi.org/10.1038/s41588-020-00764-0
- Hawkins-Hooker, A., Visonà, G., Narendra, T., Rojas-Carulla, M., Schölkopf, B., & Schweikert, G. (2023). Getting personal with epigenetics: towards individual-specific epigenomic imputation with machine learning. Nature Communications, 14(1). https://doi.org/10.1038/s41467-023-40211-2
- Heinig, M. (2018). Using Gene Expression to Annotate Cardiovascular GWAS Loci. In Frontiers in Cardiovascular Medicine (Vol. 5). https://doi.org/10.3389/fcvm.2018.00059
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell, 38(4). https://doi.org/10.1016/j.molcel.2010.05.004
- Helms, A. S., Thompson, A. D., Glazier, A. A., Hafeez, N., Kabani, S., Rodriguez, J., Yob, J. M., Woolcock, H., Mazzarotto, F., Lakdawala, N. K., Wittekind, S. G., Pereira, A. C., Jacoby, D. L., Colan, S. D., Ashley, E. A., Saberi, S., Ware, J. S., Ingles, J., Semsarian, C., … Day, S. M. (2020). Spatial and Functional Distribution of MYBPC3 Pathogenic Variants and Clinical Outcomes in Patients With Hypertrophic Cardiomyopathy. Circulation: Genomic and Precision Medicine, 13(5). https://doi.org/10.1161/CIRCGEN.120.002929
- Hentges, L. D., Sergeant, M. J., Cole, C. B., Downes, D. J., Hughes, J. R., & Taylor, S. (2022). LanceOtron: a deep learning peak caller for genome sequencing experiments. Bioinformatics, 38(18). https://doi.org/10.1093/bioinformatics/btac525
- Herkert, J. C., Verhagen, J. M. A., Yotti, R., Haghighi, A., Phelan, D. G., James, P. A., Brown, N. J., Stutterd, C., Macciocca, I., Leong, K. E., Bulthuis, M. L. C., van Bever, Y., van Slegtenhorst, M. A., Boven, L. G., Roberts, A. E., Agarwal, R., Seidman, J.,

Lakdawala, N. K., Fernández-Avilés, F., … van de Laar, I. M. B. H. (2020). Expanding the clinical and genetic spectrum of ALPK3 variants: Phenotypes identified in pediatric cardiomyopathy patients and adults with heterozygous variants. American Heart Journal, 225. https://doi.org/10.1016/j.ahj.2020.03.023

- Hickey, E. J., Mehta, R., Elmi, M., Asoh, K., McCrindle, B. W., Williams, W. G., Manlhiot, C., & Benson, L. (2011). Survival implications: Hypertrophic cardiomyopathy in noonan syndrome. Congenital Heart Disease, 6(1). https://doi.org/10.1111/j.1747-0803.2010.00465.x
- Ho, C. Y., Charron, P., Richard, P., Girolami, F., van Spaendonck-Zwarts, K. Y., & Pinto, Y. (2015). Genetic advances in sarcomeric cardiomyopathies: State of the art. In Cardiovascular Research (Vol. 105, Issue 4). https://doi.org/10.1093/cvr/cvv025
- Hoogaars, W. M. H., Engel, A., Brons, J. F., Verkerk, A. O., de Lange, F. J., Wong, L. Y. E., Bakker, M. L., Clout, D. E., Wakker, V., Barnett, P., Ravesloot, J. H., Moorman, A. F. M., Verheijck, E. E., & Christoffels, V. M. (2007). Tbx3 controls the sinoatrial node gene program and imposes pacemaker function on the atria. Genes and Development, 21(9). https://doi.org/10.1101/gad.416007
- Hoogaars, W. M. H., Tessari, A., Moorman, A. F. M., de Boer, P. A. J., Hagoort, J., Soufan, A. T., Campione, M., & Christoffels, V. M. (2004). The transcriptional repressor Tbx3 delineates the developing central conduction system of the heart. Cardiovascular Research, 62(3). https://doi.org/10.1016/j.cardiores.2004.01.030
- Hormozdiari, F., van de Bunt, M., Segrè, A. v., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. American Journal of Human Genetics, 99(6). https://doi.org/10.1016/j.ajhg.2016.10.003
- Hou, L., & Zhao, H. (2013). A review of post-GWAS prioritization approaches. In Frontiers in Genetics (Vol. 4, Issue DEC). https://doi.org/10.3389/fgene.2013.00280
- Hwang, H. S., Kahmini, A. R., Prascak, J., Cejas-Carbonell, A., Valera, I. C., Champion, S., Corrigan, M., Mumbi, F., & Parvatiyar, M. S. (2023). Sarcospan Deficiency Increases Oxidative Stress and Arrhythmias in Hearts after Acute Ischemia-Reperfusion Injury. International Journal of Molecular Sciences, 24(14). https://doi.org/10.3390/ijms241411868
- Ingles, J., Goldstein, J., Thaxton, C., Caleshu, C., Corty, E. W., Crowley, S. B., Dougherty, K., Harrison, S. M., McGlaughon, J., Milko, L. v., Morales, A., Seifert, B. A., Strande, N., Thomson, K., Peter Van Tintelen, J., Wallace, K., Walsh, R., Wells, Q., Whiffin, N., … Funke, B. (2019). Evaluating the Clinical Validity of Hypertrophic Cardiomyopathy Genes. Circulation: Genomic and Precision Medicine, 12(2). https://doi.org/10.1161/CIRCGEN.119.002460
- Inukai, S., Kock, K. H., & Bulyk, M. L. (2017). Transcription factor–DNA binding: beyond binding site motifs. In Current Opinion in Genetics and Development (Vol. 43). https://doi.org/10.1016/j.gde.2017.02.007
- Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., … Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. American Journal of Human Genetics, 99(4). https://doi.org/10.1016/j.ajhg.2016.08.016
- Ito, T., Oishi, S., Takai, M., Kimura, Y., Uozumi, Y., Fujio, Y., Schaffer, S. W., & Azuma, J. (2010). Cardiac and skeletal muscle abnormality in taurine transporter-knockout mice.

Journal of Biomedical Science, 17(SUPPL. 1). https://doi.org/10.1186/1423-0127-17-S1-S20

- Jagadeesh, K. A., Dey, K. K., Montoro, D. T., Mohan, R., Gazal, S., Engreitz, J. M., Xavier, R. J., Price, A. L., & Regev, A. (2022). Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-Sequencing and human genetics. Nature Genetics, 54(10). https://doi.org/10.1038/s41588-022-01187-9
- Jin, S. C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., Depalma, S. R., Zeng, X., Qi, H., Chang, W., Sierant, M. C., Hung, W. C., Haider, S., Zhang, J., Knight, J., Bjornson, R. D., Castaldi, C., Tikhonoa, I. R., Bilguvar, K., Mane, S. M., … Brueckner, M. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nature Genetics, 49(11). https://doi.org/10.1038/ng.3970
- Jipin Das, K., Ingles, J., Bagnall, R. D., & Semsarian, C. (2014). Determining pathogenicity of genetic variants in hypertrophic cardiomyopathy: Importance of periodic reassessment. Genetics in Medicine, 16(4). https://doi.org/10.1038/gim.2013.138
- Junttila, S., Smolander, J., & Elo, L. L. (2022). Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-Seq data. Briefings in Bioinformatics, 23(5). https://doi.org/10.1093/bib/bbac286
- Kang, Y., Kim, Y. W., Kang, J., & Kim, A. R. (2021). Histone H3K4me1 and H3K27ac play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers. FASEB Journal, 35(8). https://doi.org/10.1096/fj.202100488R
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … Xavier, R. J. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809). https://doi.org/10.1038/s41586-020-2308-7
- Karnuta, J. M., & Scacheri, P. C. (2018). Enhancers: Bridging the gap between gene control and human disease. In Human Molecular Genetics (Vol. 27, Issue R2). https://doi.org/10.1093/hmg/ddy167
- Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research, 26(7). https://doi.org/10.1101/gr.200535.115
- Kelshiker, M., Mayet, J., Unsworth, B., & Okonko, D. (2014). Basal Septal Hypertrophy. Current Cardiology Reviews, 9(4). https://doi.org/10.2174/1573403x09666131202125424
- Keyt, L. K., Duran, J. M., Bui, Q. M., Chen, C., Miyamoto, M. I., Silva Enciso, J., Tardiff, J. C., & Adler, E. D. (2022). Thin filament cardiomyopathies: A review of genetics, disease mechanisms, and emerging therapeutics. In Frontiers in Cardiovascular Medicine (Vol. 9). https://doi.org/10.3389/fcvm.2022.972301
- Kim, S., Yu, N. K., & Kaang, B. K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. In Experimental and Molecular Medicine (Vol. 47, Issue 6). https://doi.org/10.1038/EMM.2015.33
- Kim, T. K., & Shiekhattar, R. (2015). Architectural and Functional Commonalities between Enhancers and Promoters. In Cell (Vol. 162, Issue 5). https://doi.org/10.1016/j.cell.2015.08.008
- Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2015). DEEP: A general computational framework for predicting enhancers. Nucleic Acids Research, 43(1). https://doi.org/10.1093/nar/gku1058

- Klues, H. G., Schiffers, A., & Maron, B. J. (1995). Phenotypic spectrum and patterns of left ventricular hypertrophy in hypertrophic cardiomyopathy: Morphologic observations and significance as assessed by two-dimensional echocardiography in 600 patients. Journal of the American College of Cardiology, 26(7). https://doi.org/10.1016/0735-1097(95)00390-8
- Kobayashi, T., Aoki, Y., Niihori, T., Cavé, H., Verloes, A., Okamoto, N., Kawame, H., Fujiwara, I., Takada, F., Ohata, T., Sakazume, S., Ando, T., Nakagawa, N., Lapunzina, P., Meneses, A. G., Gillessen-Kaesbach, G., Wieczorek, D., Kurosawa, K., Mizuno, S., … Matsubara, Y. (2010). Molecular and clinical analysis of RAF1 in Noonan syndrome and related disorders: Dephosphorylation of serine 259 as the essential mechanism for mutant activation. Human Mutation, 31(3). https://doi.org/10.1002/humu.21187
- Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., & Winter, D. R. (2018). A beginner's guide to analysis of RNA sequencing data. In American Journal of Respiratory Cell and Molecular Biology (Vol. 59, Issue 2). https://doi.org/10.1165/rcmb.2017-0430TR
- Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). G:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Research, 51(W1). https://doi.org/10.1093/nar/gkad347
- Kosti, I., Jain, N., Aran, D., Butte, A. J., & Sirota, M. (2016). Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. Scientific Reports, 6. https://doi.org/10.1038/srep24799
- Kuksa, P. P., Greenfest-Allen, E., Cifello, J., Ionita, M., Wang, H., Nicaretta, H., Cheng, P. L., Lee, W. P., Wang, L. S., & Leung, Y. Y. (2022). Scalable approaches for functional analyses of whole-genome sequencing non-coding variants. In Human Molecular Genetics (Vol. 31, Issue R1). https://doi.org/10.1093/hmg/ddac191
- Kulkarni, S. S., & Khokha, M. K. (2018). WDR5 regulates left-right patterning via chromatin-dependent and-independent functions. Development (Cambridge), 145(23). https://doi.org/10.1242/dev.159889
- Künzel, S. R., Hoffmann, M., Weber, S., Künzel, K., Kämmerer, S., Günscht, M., Klapproth, E., Rausch, J. S. E., Sadek, M. S., Kolanowski, T., Meyer-Roxlau, S., Piorkowski, C., Tugtekin, S. M., Rose-John, S., Yin, X., Mayr, M., Kuhlmann, J. D., Wimberger, P., Grützmann, K., … El-Armouche, A. (2021). Diminished PLK2 Induces Cardiac Fibrosis and Promotes Atrial Fibrillation. Circulation Research, 129(8). https://doi.org/10.1161/CIRCRESAHA.121.319425
- Lan, A. Y., & Corces, M. R. (2022). Deep learning approaches for noncoding variant prioritization in neurodegenerative diseases. In Frontiers in Aging Neuroscience (Vol. 14). https://doi.org/10.3389/fnagi.2022.1027224
- Lee, D. (2023). The Transformer Model: Revolutionizing Natural Language Processing. Medium. https://medium.com/@lordmoma/the-transformer-model-revolutionizing-natural-language-processing-a16be54ddb1e
- Lesurf, R., Said, A., Akinrinade, O., Breckpot, J., Delfosse, K., Liu, T., Yao, R., Persad, G., McKenna, F., Noche, R. R., Oliveros, W., Mattioli, K., Shah, S., Miron, A., Yang, Q., Meng, G., Yue, M. C. S., Sung, W. W. L., Thiruvahindrapuram, B., … Mital, S. (2022). Whole genome sequencing delineates regulatory, copy number, and cryptic splice variants in early onset cardiomyopathy. Npj Genomic Medicine, 7(1). https://doi.org/10.1038/s41525-022-00288-y

- Levin, M. G., Tsao, N. L., Singhal, P., Liu, C., Vy, H. M. T., Paranjpe, I., Backman, J. D., Bellomo, T. R., Bone, W. P., Biddinger, K. J., Hui, Q., Dikilitas, O., Satterfield, B. A., Yang, Y., Morley, M. P., Bradford, Y., Burke, M., Reza, N., Charest, B., … Damrauer, S. M. (2022). Genome-wide association and multi-trait analyses characterize the common genetic architecture of heart failure. Nature Communications, 13(1). https://doi.org/10.1038/s41467-022-34216-6
- Li, C., Liu, F., Liu, S., Pan, H., Du, H., Huang, J., Xie, Y., Li, Y., Zhao, R., & Wei, Y. (2020). Elevated myocardial SORBS2 and the underlying implications in left ventricular noncompaction cardiomyopathy. EBioMedicine, 53. https://doi.org/10.1016/j.ebiom.2020.102695
- Li, D., & Ma, Q. (2025). Ubiquitin-specific protease: an emerging key player in cardiomyopathy. Cell Communication and Signaling, 23(1), 143. https://doi.org/10.1186/s12964-025-02123-0
- Li, K., Persaud, D., Choudhary, K., DeCost, B., Greenwood, M., & Hattrick-Simpers, J. (2023). Exploiting redundancy in large materials datasets for efficient machine learning with less data. Nature Communications, 14(1). https://doi.org/10.1038/s41467-023-42992-y
- Li, L., Bainbridge, M. N., Tan, Y., Willerson, J. T., & Marian, A. J. (2017). A potential oligogenic etiology of hypertrophic cardiomyopathy: a classic single-gene disorder. Circulation Research, 120(7). https://doi.org/10.1161/CIRCRESAHA.116.310559
- Li, L., Waymack, R., Gad, M., & Wunderlich, Z. (2021b). Two promoters integrate multiple enhancer inputs to drive wild-type knirps expression in the Drosophila melanogaster embryo. Genetics, 219(4). https://doi.org/10.1093/genetics/iyab154
- Liu, X., Li, Y. I., & Pritchard, J. K. (2019a). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell, 177(4). https://doi.org/10.1016/j.cell.2019.04.014
- Liu, Q., Lv, H., & Jiang, R. (2019b). HicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics, 35(14). https://doi.org/10.1093/bioinformatics/btz317
- Lombardi, R., Bell, A., Senthil, V., Sidhu, J., Noseda, M., Roberts, R., & Marian, A. J. (2008). Differential interactions of thin filament proteins in two cardiac troponin T mouse models of hypertrophic and dilated cardiomyopathies. Cardiovascular Research, 79(1). https://doi.org/10.1093/cvr/cvn078
- Long, E., Williams, J., Zhang, H., & Choi, J. (2025). An evolving understanding of multiple causal variants underlying genetic association signals. The American Journal of Human Genetics, 112(4), 741–750. https://doi.org/https://doi.org/10.1016/j.ajhg.2025.01.018
- Lopes, L. R., Losi, M. A., Sheikh, N., Laroche, C., Charron, P., Gimeno, J., Kaski, J. P., Maggioni, A. P., Tavazzi, L., Arbustini, E., Brito, D., Celutkiene, J., Hagege, A., Linhart, A., Mogensen, J., Garcia-Pinilla, J. M., Ripoll-Vera, T., Seggewiss, H., Villacorta, E., … Marques-Sule, E. (2023). Association between common cardiovascular risk factors and clinical phenotype in patients with hypertrophic cardiomyopathy from the European Society of Cardiology (ESC) EurObservational Research Programme (EORP) Cardiomyopathy/Myocarditis registry. European Heart Journal - Quality of Care and Clinical Outcomes, 9(1). https://doi.org/10.1093/ehjqcco/qcac006
- Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., Levens, D. L., Lee, T. I., & Young, R. A. (2012). Revisiting global gene expression analysis. In Cell (Vol. 151, Issue 3). https://doi.org/10.1016/j.cell.2012.10.012

- Luizon, M. R., & Ahituv, N. (2015). Uncovering drug-responsive regulatory elements. In Pharmacogenomics (Vol. 16, Issue 16). https://doi.org/10.2217/pgs.15.121
- Luo, W., Lin, Z., Chen, J., Chen, G., Zhang, S., Liu, M., Li, H., He, D., Liang, S., Luo, Q., Zhang, D., Nie, Q., & Zhang, X. (2021). TMEM182 interacts with integrin beta 1 and regulates myoblast differentiation and muscle regeneration. Journal of Cachexia, Sarcopenia and Muscle, 12(6). https://doi.org/10.1002/jcsm.12767
- MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., … Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science, 335(6070). https://doi.org/10.1126/science.1215040
- Major, J. L., Salih, M., & Tuana, B. S. (2017). E2F6 protein levels modulate drug induced apoptosis in cardiomyocytes. Cellular Signalling, 40. https://doi.org/10.1016/j.cellsig.2017.09.022
- Mak, T. S. H., Lee, Y. K., Tang, C. S., Hai, J. J. S. H., Ran, X., Sham, P. C., & Tse, H. F. (2018). Coverage and diagnostic yield of Whole Exome Sequencing for the Evaluation of Cases with Dilated and Hypertrophic Cardiomyopathy. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-29263-3
- Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 5(9). https://doi.org/10.1097/JTO.0b013e3181ec173d
- Marian, A. J. (2002). Modifier genes for hypertrophic cardiomyopathy. In Current Opinion in Cardiology (Vol. 17, Issue 3). https://doi.org/10.1097/00001573-200205000-00006
- Marian, A. J. (2021). Molecular Genetic Basis of Hypertrophic Cardiomyopathy. Circulation Research, 128(10). https://doi.org/10.1161/CIRCRESAHA.121.318346
- Marian, A. J., & Braunwald, E. (2017). Hypertrophic cardiomyopathy: Genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. Circulation Research, 121(7). https://doi.org/10.1161/CIRCRESAHA.117.311059
- Maron, B. J., Desai, M. Y., Nishimura, R. A., Spirito, P., Rakowski, H., Towbin, J. A., Rowin, E. J., Maron, M. S., & Sherrid, M. v. (2022). Diagnosis and Evaluation of Hypertrophic Cardiomyopathy: JACC State-of-the-Art Review. In Journal of the American College of Cardiology (Vol. 79, Issue 4). https://doi.org/10.1016/j.jacc.2021.12.002
- Maron, B. J., Gardin, J. M., Flack, J. M., Gidding, S. S., Kurosaki, T. T., & Bild, D. E. (1995). Prevalence of hypertrophic cardiomyopathy in a general population of young adults: Echocardiographic analysis of 4111 subjects in the CARDIA study. Circulation, 92(4). https://doi.org/10.1161/01.CIR.92.4.785
- Maron, B. J., Maron, M. S., & Semsarian, C. (2012). Genetics of hypertrophic cardiomyopathy after 20 years: Clinical perspectives. In Journal of the American College of Cardiology (Vol. 60, Issue 8). https://doi.org/10.1016/j.jacc.2012.02.068
- Maron, B. J., Olivotto, I., Spirito, P., Casey, S. A., Bellone, P., Gohman, T. E., Graham, K. J., Burton, D. A., & Cecchi, F. (2000). Epidemiology of hypertrophic cardiomyopathy-related death: Revisited in a large non-referral-based patient population. Circulation, 102(8). https://doi.org/10.1161/01.CIR.102.8.858
- Mazzarotto, F., Girolami, F., Boschi, B., Barlocco, F., Tomberli, A., Baldini, K., Coppini, R., Tanini, I., Bardi, S., Contini, E., Cecchi, F., Pelo, E., Cook, S. A., Cerbai, E., Poggesi, C., Torricelli, F., Walsh, R., & Olivotto, I. (2019). Defining the diagnostic effectiveness of genes for inclusion in panels: the experience of two decades of genetic testing for

hypertrophic cardiomyopathy at a single center. Genetics in Medicine, 21(2). https://doi.org/10.1038/s41436-018-0046-0

- McLendon, J. M., Zhang, X., Matasic, D. S., Kumar, M., Koval, O. M., Grumbach, I. M., Sadayappan, S., London, B., & Boudreau, R. L. (2022). Knockout of Sorbin And SH3 Domain Containing 2 (Sorbs2) in Cardiomyocytes Leads to Dilated Cardiomyopathy in Mice. Journal of the American Heart Association, 11(13). https://doi.org/10.1161/JAHA.122.025687
- McNamara, J. W., Li, A., Smith, N. J., Lal, S., Graham, R. M., Kooiker, K. B., van Dijk, S. J., Remedios, C. G. dos, Harris, S. P., & Cooke, R. (2016). Ablation of cardiac myosin binding protein-C disrupts the super-relaxed state of myosin in murine cardiomyocytes. Journal of Molecular and Cellular Cardiology, 94. https://doi.org/10.1016/j.yjmcc.2016.03.009
- Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. v., Djebali, S., Niarchou, A., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., … Guigó, R. (2015). The human transcriptome across tissues and individuals. Science, 348(6235). https://doi.org/10.1126/science.aaa0355
- Mendoza-Parra, M. A., van Gool, W., Saleem, M. A. M., Ceschin, D. G., & Gronemeyer, H. (2013). A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Research, 41(21). https://doi.org/10.1093/nar/gkt829
- Millat, G., Bouvagnet, P., Chevalier, P., Dauphin, C., Simon Jouk, P., da Costa, A., Prieur, F., Bresson, J. L., Faivre, L., Eicher, J. C., Chassaing, N., Crehalet, H., Porcher, R., Rodriguez-Lafrasse, C., & Rousson, R. (2010). Prevalence and spectrum of mutations in a cohort of 192 unrelated patients with hypertrophic cardiomyopathy. European Journal of Medical Genetics, 53(5). https://doi.org/10.1016/j.ejmg.2010.07.007
- Miller, C. L., Pjanic, M., & Quertermous, T. (2015). From locus association to mechanism of gene causality the devil is in the details. In Arteriosclerosis, Thrombosis, and Vascular Biology (Vol. 35, Issue 10). https://doi.org/10.1161/ATVBAHA.115.306366
- Minton, K. (2023). Predicting variant pathogenicity with AlphaMissense. In Nature Reviews Genetics (Vol. 24, Issue 12). https://doi.org/10.1038/s41576-023-00668-9
- Mohan, R. A., Mommersteeg, M. T. M., Domínguez, J. N., Choquet, C., Wakker, V., de Gier-de Vries, C., Boink, G. J. J., Boukens, B. J., Miquerol, L., Verkerk, A. O., & Christoffels, V. M. (2018). Embryonic Tbx3+ cardiomyocytes form the mature cardiac conduction system by progressive fate restriction. Development (Cambridge), 145(17). https://doi.org/10.1242/dev.167361
- Mommersteeg, M. T. M., Hoogaars, W. M. H., Prall, O. W. J., de Gier-De Vries, C., Wiese, C., Clout, D. E. W., Papaioannou, V. E., Brown, N. A., Harvey, R. P., Moorman, A. F. M., & Christoffels, V. M. (2007). Molecular pathway for the localized formation of the sinoatrial node. Circulation Research, 100(3). https://doi.org/10.1161/01.RES.0000258019.74591.b3
- Monserrat, L. (2018). Perspectives on current recommendations for genetic testing in HCM. Global Cardiology Science and Practice, 2018(3). https://doi.org/10.21542/gcsp.2018.23
- Morley, M. P., Wang, X., Hu, R., Brandimarto, J., Tucker, N. R., Felix, J. F., Smith, N. L., van der Harst, P., Ellinor, P. T., Margulies, K. B., Musunuru, K., & Cappola, T. P. (2019). Cardioprotective effects of mtss1 enhancer variants. In Circulation (Vol. 139, Issue 17). https://doi.org/10.1161/CIRCULATIONAHA.118.037939

- Mostafavi, H., Spence, J. P., Naqvi, S., & Pritchard, J. K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. Nature Genetics, 55(11). https://doi.org/10.1038/s41588-023-01529-1
- Movva, R., Greenside, P., Marinov, G. K., Nair, S., Shrikumar, A., & Kundaje, A. (2019). Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. PLoS ONE, 14(6). https://doi.org/10.1371/journal.pone.0218073
- Mu, W., & Zhang, W. (2013). Molecular Approaches, Models, and Techniques in Pharmacogenomic Research and Development. In Pharmacogenomics: Challenges and Opportunities in Therapeutic Implementation. https://doi.org/10.1016/B978-0-12-391918-2.00008-1
- Mulvey, B., Lagunas, T., & Dougherty, J. D. (2021). Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants Across Biological Contexts. In Biological Psychiatry (Vol. 89, Issue 1). https://doi.org/10.1016/j.biopsych.2020.06.011
- Myers, T. A., Chanock, S. J., & Machiela, M. J. (2020). LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. Frontiers in Genetics, 11. https://doi.org/10.3389/fgene.2020.00157
- Na, J., Cui, L., Zhen, Z., Chen, X., Li, Q., Gao, L., & Yuan, Y. (2024). Recurrent myocardial injury in a de novo SON mutation ZTTK syndrome patient: a case report. BMC Pediatrics, 24(1), 232. https://doi.org/10.1186/s12887-024-04703-4
- Nakanishi, T., Markwald, R. R., Baldwin, H. S., Keller, B. B., Srivastava, D., & Yamagishi, H. (2016). Etiology and morphogenesis of congenital heart disease: From gene function and cellular interaction to morphology. In Etiology and Morphogenesis of Congenital Heart Disease: From Gene Function and Cellular Interaction to Morphology. https://doi.org/10.1007/978-4-431-54628-3
- Nakato, R., Sakata, T., Wang, J., Nagai, L. A. E., Nagaoka, Y., Oba, G. M., Bando, M., & Shirahige, K. (2023). Context-dependent perturbations in chromatin folding and the transcriptome by cohesin and related factors. Nature Communications, 14(1). https://doi.org/10.1038/s41467-023-41316-4
- Nakhaei-Rad, S., Haghighi, F., Bazgir, F., Dahlmann, J., Busley, A. V., Buchholzer, M., Kleemann, K., Schänzer, A., Borchardt, A., Hahn, A., Kötter, S., Schanze, D., Anand, R., Funk, F., Kronenbitter, A. V., Scheller, J., Piekorz, R. P., Reichert, A. S., Volleth, M., … Ahmadian, M. R. (2023). Molecular and cellular evidence for the impact of a hypertrophic cardiomyopathy-associated RAF1 variant on the structure and function of contractile machinery in bioartificial cardiac tissues. Communications Biology, 6(1). https://doi.org/10.1038/s42003-023-05013-8
- Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., … Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature, 593(7858). https://doi.org/10.1038/s41586-021-03446-x
- Neubauer, S., Kolm, P., Ho, C. Y., Kwong, R. Y., Desai, M. Y., Dolman, S. F., Appelbaum, E., Desvigne-Nickens, P., DiMarco, J. P., Friedrich, M. G., Geller, N., Harper, A. R., Jarolim, P., Jerosch-Herold, M., Kim, D. Y., Maron, M. S., Schulz-Menger, J., Piechnik, S. K., Thomson, K., … Kramer, C. M. (2019). Distinct Subgroups in Hypertrophic Cardiomyopathy in the NHLBI HCM Registry. Journal of the American College of Cardiology, 74(19). https://doi.org/10.1016/j.jacc.2019.08.1057

- Norrish, G., Kolt, G., Cervi, E., Field, E., Dady, K., Ziółkowska, L., Olivotto, I., Favilli, S., Passantino, S., Limongelli, G., Caiazza, M., Rubino, M., Baban, A., Drago, F., Mcleod, K., Ilina, M., McGowan, R., Stuart, G., Bhole, V., … Kaski, J. P. (2021). Clinical presentation and long-term outcomes of infantile hypertrophic cardiomyopathy: a European multicentre study. ESC Heart Failure, 8(6). https://doi.org/10.1002/ehf2.13573
- North, B. v., Curtis, D., & Sham, P. C. (2002). A note on the calculation of empirical P values from Monte Carlo procedures [1]. In American Journal of Human Genetics (Vol. 71, Issue 2). https://doi.org/10.1086/341527
- O'Mahony, C., & Elliott, P. (2017). Affairs of the heart: Outcomes in men and women with hypertrophic cardiomyopathy. In European Heart Journal (Vol. 38, Issue 46). https://doi.org/10.1093/eurheartj/ehx639
- Pandit, B., Sarkozy, A., Pennacchio, L. A., Carta, C., Oishi, K., Martinelli, S., Pogna, E. A., Schackwitz, W., Ustaszewska, A., Landstrom, A., Bos, J. M., Ommen, S. R., Esposito, G., Lepri, F., Faul, C., Mundel, P., López Siguero, J. P., Tenconi, R., Selicorni, A., … Gelb, B. D. (2007). Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. Nature Genetics, 39(8). https://doi.org/10.1038/ng2073
- Panigrahi, A., & O'Malley, B. W. (2021). Mechanisms of enhancer action: the known and the unknown. In Genome Biology (Vol. 22, Issue 1). https://doi.org/10.1186/s13059-021-02322-1
- Park, E., Pan, Z., Zhang, Z., Lin, L., & Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. In American Journal of Human Genetics (Vol. 102, Issue 1). https://doi.org/10.1016/j.ajhg.2017.11.002
- Parvatiyar, M. S., Marshall, J. L., Nguyen, R. T., Jordan, M. C., Richardson, V. A., Roos, K. P., & Crosbie-Watson, R. H. (2015). Sarcospan regulates cardiac isoproterenol response and prevents duchenne muscular dystrophy-associated cardiomyopathy. Journal of the American Heart Association, 4(12). https://doi.org/10.1161/JAHA.115.002481
- Peña-Martínez, E. G., & Rodríguez-Martínez, J. A. (2024). Decoding Non-coding Variants: Recent Approaches to Studying Their Role in Gene Regulation and Human Diseases. In Frontiers in Bioscience - Scholar (Vol. 16, Issue 1). https://doi.org/10.31083/j.fbs1601004
- Peng, Y., Gregorich, Z. R., Valeja, S. G., Zhang, H., Cai, W., Chen, Y. C., Guner, H., Chen, A. J., Schwahn, D. J., Hacker, T. A., Liu, X., & Ge, Y. (2014b). Top-down proteomics reveals concerted reductions in myofilament and Z-disc protein phosphorylation after acute myocardial infarction. Molecular and Cellular Proteomics, 13(10). https://doi.org/10.1074/mcp.M114.040675
- Pirruccello, J. P., Bick, A., Wang, M., Chaffin, M., Friedman, S., Yao, J., Guo, X., Venkatesh, B. A., Taylor, K. D., Post, W. S., Rich, S., Lima, J. A. C., Rotter, J. I., Philippakis, A., Lubitz, S. A., Ellinor, P. T., Khera, A. v., Kathiresan, S., & Aragam, K. G. (2020). Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-15823-7
- Prondzynski, M., Lemoine, M. D., Zech, A. T., Horváth, A., di Mauro, V., Koivumäki, J. T., Kresin, N., Busch, J., Krause, T., Krämer, E., Schlossarek, S., Spohn, M., Friedrich, F. W., Münch, J., Laufer, S. D., Redwood, C., Volk, A. E., Hansen, A., Mearini, G., … Carrier, L. (2019). Disease modeling of a mutation in α-actinin 2 guides clinical therapy in hypertrophic cardiomyopathy. EMBO Molecular Medicine, 11(12). https://doi.org/10.15252/emmm.201911115

- Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 35(SUPPL. 1). https://doi.org/10.1093/nar/gkl842
- Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., & Ren, B. (2013). RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. PLoS Computational Biology, 9(3). https://doi.org/10.1371/journal.pcbi.1002968
- Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A., Machol, I., Omer, A. D., Lander, E. S., Lieberman Aiden, E., Designed, E. L. A., & Performed Hi-C Experiments, M. H. H. (2014). A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping eToC Blurb HHS Public Access. Cell, 159(7).
- Raphael, C. E., Cooper, R., Parker, K. H., Collinson, J., Vassiliou, V., Pennell, D. J., de Silva, R., Hsu, L. Y., Greve, A. M., Nijjer, S., Broyd, C., Ali, A., Keegan, J., Francis, D. P., Davies, J. E., Hughes, A. D., Arai, A., Frenneaux, M., Stables, R. H., … Prasad, S. K. (2016). Mechanisms of Myocardial Ischemia in Hypertrophic Cardiomyopathy: Insights From Wave Intensity Analysis and Magnetic Resonance. Journal of the American College of Cardiology, 68(15). https://doi.org/10.1016/j.jacc.2016.07.751
- Razzaque, M. A., Nishizawa, T., Komoike, Y., Yagi, H., Furutani, M., Amo, R., Kamisago, M., Momma, K., Katayama, H., Nakagawa, M., Fujiwara, Y., Matsushima, M., Mizuno, K., Tokuyama, M., Hirota, H., Muneuchi, J., Higashinakagawa, T., & Matsuoka, R. (2007). Germline gain-of-function mutations in RAF1 cause Noonan syndrome. Nature Genetics, 39(8). https://doi.org/10.1038/ng2078
- w, C. S., Moolman-Smook, J. C., & Watkins, H. (1999). Properties of mutant contractile proteins that cause hypertrophic cardiomyopathy. In Cardiovascular Research (Vol. 44, Issue 1). https://doi.org/10.1016/S0008-6363(99)00213-8
- Reichart, D., Lindberg, E. L., Maatz, H., Miranda, A. M. A., Viveiros, A., Shvetsov, N., Gärtner, A., Nadelmann, E. R., Lee, M., Kanemaru, K., Ruiz-Orera, J., Strohmenger, V., DeLaughter, D. M., Patone, G., Zhang, H., Woehler, A., Lippert, C., Kim, Y., Adami, E., … Seidman, C. E. (2022). Pathogenic variants damage cell composition and single-cell transcription in cardiomyopathies. Science, 377(6606). https://doi.org/10.1126/science.abo1984
- Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). G:Profiler-a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Research, 35(SUPPL.2). https://doi.org/10.1093/nar/gkm226
- Richard, P., Charron, P., Carrier, L., Ledeuil, C., Cheav, T., Pichereau, C., Benaiche, A., Isnard, R., Dubourg, O., Burban, M., Gueffet, J. P., Millaire, A., Desnos, M., Schwartz, K., Hainque, B., & Komajda, M. (2003). Hypertrophic cardiomyopathy: Distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy. Circulation, 107(17). https://doi.org/10.1161/01.CIR.0000066323.15244.54
- Ritchie, G. R. S., Dunham, I., Zeggini, E., & Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nature Methods, 11(3). https://doi.org/10.1038/nmeth.2832
- Riva, S. G., Georgiades, E., Gur, E. R., Baxter, M., & Hughes, J. R. (2023). CATCH-UP: A High-Throughput Upstream-Pipeline for Bulk ATAC-Seq and ChIP-seq Data. Journal of Visualized Experiments, 2023(199). https://doi.org/10.3791/65633
- Riva, S. G., Georgiades, E., Herrmann, J. C., Gür, E. R., Sanders, E., Sergeant, M., Baxter, M., & Hughes, J. R. (2024). Deciphering cis-regulatory elements using REgulamentary. BioRxiv, 2024.05.24.595662. https://doi.org/10.1101/2024.05.24.595662

- Riva, S. G., Sanders, E., Wilson, T., Stranieri, N., Gür, E. R., Baxter, M., & Hughes, J. R. (2025). REnformer, a single-cell ATAC-seq predicting model to investigate open chromatin sites. BioRxiv, 2025.06.04.657786. https://doi.org/10.1101/2025.06.04.657786
- Riva, S. G., Venkatesh, S. S., Gür, E. R., Stranieri, N., Sergeant, M., Baxter, M., & Hughes, J. R. (2025). CREST-GV: Cell types Ranking and Enrichment Score for selecTive Genetic Variants. MedRxiv, 2025.07.29.25332311. https://doi.org/10.1101/2025.07.29.25332311
- Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., & Aiden, E. L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. Cell Systems, 6(2). https://doi.org/10.1016/j.cels.2018.01.001
- Roselli, C., Chaffin, M. D., Weng, L. C., Aeschbacher, S., Ahlberg, G., Albert, C. M., Almgren, P., Alonso, A., Anderson, C. D., Aragam, K. G., Arking, D. E., Barnard, J., Bartz, T. M., Benjamin, E. J., Bihlmeyer, N. A., Bis, J. C., Bloom, H. L., Boerwinkle, E., Bottinger, E. B., … Ellinor, P. T. (2018). Multi-ethnic genome-wide association study for atrial fibrillation. Nature Genetics, 50(9). https://doi.org/10.1038/s41588-018-0133-9
- Rowe, G. C., Asimaki, A., Graham, E. L., Martin, K. D., Margulies, K. B., Das, S., Saffitz, J., & Arany, Z. (2017). Development of dilated cardiomyopathy and impaired calcium homeostasis with cardiac-specific deletion of ESRRβ. American Journal of Physiology - Heart and Circulatory Physiology, 312(4). https://doi.org/10.1152/ajpheart.00446.2016
- Rowin, E. J., Maron, B. J., & Maron, M. S. (2020). The Hypertrophic Cardiomyopathy Phenotype Viewed Through the Prism of Multimodality Imaging: Clinical and Etiologic Implications. In JACC: Cardiovascular Imaging (Vol. 13, Issue 9). https://doi.org/10.1016/j.jcmg.2019.09.020
- Sabrina, M., Emanuele, M., & Giuseppe, L. (2025). Risk Models for Sudden Cardiac Death in Hypertrophic Cardiomyopathy. JACC, 86(1), 67–70. https://doi.org/10.1016/j.jacc.2025.05.027
- Saccone, S. F., Bolze, R., Thomas, P., Quan, J., Mehta, G., Deelman, E., Tischfield, J. A., & Rice, J. P. (2010). SPOT: A web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. Nucleic Acids Research, 38(SUPPL. 2). https://doi.org/10.1093/nar/gkq513
- Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. In Nature Methods (Vol. 7, Issue 8). https://doi.org/10.1038/nmeth0810-575
- Schwessinger, R., Gosden, M., Downes, D., Brown, R. C., Oudelaar, A. M., Telenius, J., Teh, Y. W., Lunter, G., & Hughes, J. R. (2020). DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nature Methods, 17(11). https://doi.org/10.1038/s41592-020-0960-3
- Segert, J. A., Gisselbrecht, S. S., & Bulyk, M. L. (2021). Transcriptional Silencers: Driving Gene Expression with the Brakes On. In Trends in Genetics (Vol. 37, Issue 6). https://doi.org/10.1016/j.tig.2021.02.002
- Semsarian, C., Ingles, J., Maron, M. S., & Maron, B. J. (2015). New perspectives on the prevalence of hypertrophic cardiomyopathy. In Journal of the American College of Cardiology (Vol. 65, Issue 12). https://doi.org/10.1016/j.jacc.2015.01.019
- Sergeant, M. J., Hughes, J. R., Hentges, L., Lunter, G., Downes, D. J., & Taylor, S. (2021). Multi Locus View: an extensible web-based tool for the analysis of genomic data. Communications Biology, 4(1). https://doi.org/10.1038/s42003-021-02097-y
- Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., Gaunt, T. R., & Campbell, C. (2015). An integrative approach to predicting the functional effects of

non-coding and coding sequence variation. Bioinformatics, 31(10). https://doi.org/10.1093/bioinformatics/btv009

- Shilatifard, A. (2012). The COMPASS family of histone H3K4 methylases: Mechanisms of regulation in development and disease pathogenesis. Annual Review of Biochemistry, 81. https://doi.org/10.1146/annurev-biochem-051710-134100
- Shiraishi, C., Matsumoto, A., Ichihara, K., Yamamoto, T., Yokoyama, T., Mizoo, T., Hatano, A., Matsumoto, M., Tanaka, Y., Matsuura-Suzuki, E., Iwasaki, S., Matsushima, S., Tsutsui, H., & Nakayama, K. I. (2023). RPL3L-containing ribosomes determine translation elongation dynamics required for cardiac function. Nature Communications, 14(1). https://doi.org/10.1038/s41467-023-37838-6
- Singh, R., Hoogaars, W. M., Barnett, P., Grieskamp, T., Sameer Rana, M., Buermans, H., Farin, H. F., Petry, M., Heallen, T., Martin, J. F., Moorman, A. F. M., Hoen, P. A. C., Kispert, A., & Christoffels, V. M. (2012). Tbx2 and Tbx3 induce atrioventricular myocardial development and endocardial cushion formation. Cellular and Molecular Life Sciences, 69(8). https://doi.org/10.1007/s00018-011-0884-2
- Spielmann, N., Miller, G., Oprea, T. I., Hsu, C. W., Fobo, G., Frishman, G., Montrone, C., Haseli Mashhadi, H., Mason, J., Munoz Fuentes, V., Leuchtenberger, S., Ruepp, A., Wagner, M., Westphal, D. S., Wolf, C., Görlach, A., Sanz-Moreno, A., Cho, Y. L., Teperino, R., … Hrabe de Angelis, M. (2022). Extensive identification of genes involved in congenital and structural heart disorders and cardiomyopathy. Nature Cardiovascular Research, 1(2). https://doi.org/10.1038/s44161-022-00018-8
- Spudich, J. A. (2014). Hypertrophic and dilated cardiomyopathy: Four decades of basic research on muscle lead to potential therapeutic approaches to these devastating genetic diseases. In Biophysical Journal (Vol. 106, Issue 6). https://doi.org/10.1016/j.bpj.2014.02.011
- Singh, S. (2019). Fully Connected Layer: The brute force layer of a Machine Learning model. OpenGenus. https://iq.opengenus.org/fully-connected-layer/
- Tadros, R., Francis, C., Xu, X., Vermeer, A. M. C., Harper, A. R., Huurman, R., Kelu Bisabu, K., Walsh, R., Hoorntje, E. T., te Rijdt, W. P., Buchan, R. J., van Velzen, H. G., van Slegtenhorst, M. A., Vermeulen, J. M., Offerhaus, J. A., Bai, W., de Marvao, A., Lahrouchi, N., Beekman, L., … Bezzina, C. R. (2021). Shared genetic pathways contribute to risk of hypertrophic and dilated cardiomyopathies with opposite directions of effect. Nature Genetics, 53(2). https://doi.org/10.1038/s41588-020-00762-2
- Tadros, R., Zheng, S.L., Grace, C., Jorda, P., Francis, C., Jurgens, S. J., Thomson, K. L., Harper, A. R., Wilde, A. A., Olivotto, I., Adler, A., Goel, A., Ware, J. S., Bezzina, C. R., & Watkins, H. (2025). Large-scale genome-wide association analyses identify novel genetic loci and mechanisms in hypertrophic cardiomyopathy. Nature Genetics, 57, 530–538. https://doi.org/10.1038/s41588-025-02087-4
- Talman, V., & Kivelä, R. (2018). Cardiomyocyte—Endothelial Cell Interactions in Cardiac Remodeling and Regeneration. In Frontiers in Cardiovascular Medicine (Vol. 5). https://doi.org/10.3389/fcvm.2018.00101
- Tan, W. L. W., Anene-Nzelu, C. G., Wong, E., Lee, C. J. M., Tan, H. S., Tang, S. J., Perrin, A., Wu, K. X., Zheng, W., Ashburn, R. J., Pan, B., Lee, M. Y., Autio, M. I., Morley, M. P., Tam, W. L., Cheung, C., Margulies, K. B., Chen, L., Cappola, T. P., … Foo, R. S. Y. (2020). Epigenomes of Human Hearts Reveal New Genetic Variants Relevant for Cardiac Disease and Phenotype. Circulation Research, 127(6). https://doi.org/10.1161/CIRCRESAHA.120.317254

- Tayal, U., Ware, J. S., Lakdawala, N. K., Heymans, S., & Prasad, S. K. (2021). Understanding the genetics of adult-onset dilated cardiomyopathy: What a clinician needs to know. In European Heart Journal (Vol. 42, Issue 24). https://doi.org/10.1093/eurheartj/ehab286
- Theis, J. L., Martijn Bos, J., Bartleson, V. B., Will, M. L., Binder, J., Vatta, M., Towbin, J. A., Gersh, B. J., Ommen, S. R., & Ackerman, M. J. (2006). Echocardiographic-determined septal morphology in Z-disc hypertrophic cardiomyopathy. Biochemical and Biophysical Research Communications, 351(4). https://doi.org/10.1016/j.bbrc.2006.10.119
- Thierfelder, L., Watkins, H., MacRae, C., Lamas, R., McKenna, W., Vosberg, H. P., Seldman, J. G., & Seidman, C. E. (1994). α-tropomyosin and cardiac troponin T mutations cause familial hypertrophic cardiomyopathy: A disease of the sarcomere. Cell, 77(5). https://doi.org/10.1016/0092-8674(94)90054-X
- Thomson, K. L., Ormondroyd, E., Harper, A. R., Dent, T., McGuire, K., Baksi, J., Blair, E., Brennan, P., Buchan, R., Bueser, T., Campbell, C., Carr-White, G., Cook, S., Daniels, M., Deevi, S. V. V., Goodship, J., Hayesmoore, J. B. G., Henderson, A., Lamb, T., … Watkins, H. C. (2019). Analysis of 51 proposed hypertrophic cardiomyopathy genes from genome sequencing data in sarcomere negative cases has negligible diagnostic yield. Genetics in Medicine, 21(7). https://doi.org/10.1038/s41436-018-0375-z
- Tudurachi, B. S., Zăvoi, A., Leonte, A., Țăpoi, L., Ureche, C., Bîrgoan, S. G., Chiuariu, T., Anghel, L., Radu, R., Sascău, R. A., & Stătescu, C. (2023). An Update on MYBPC3 Gene Mutation in Hypertrophic Cardiomyopathy. In International Journal of Molecular Sciences (Vol. 24, Issue 13). https://doi.org/10.3390/ijms241310510
- Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., Nguyen-Viet, T. A., Wedow, R., Zacher, M., Furlotte, N. A., Magnusson, P., Oskarsson, S., Johannesson, M., Visscher, P. M., Laibson, D., Cesarini, D., Neale, B. M., Benjamin, D. J., Agee, M., … Pitts, S. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature Genetics, 50(2). https://doi.org/10.1038/s41588-017-0009-4
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. In Nature Reviews Methods Primers (Vol. 1, Issue 1). https://doi.org/10.1038/s43586-021-00056-9
- Vaz, J. M., & Balaji, S. (2021). Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics. Molecular Diversity, 25(3). https://doi.org/10.1007/s11030-021-10225-3
- Villard, E., Perret, C., Gary, F., Proust, C., Dilanian, G., Hengstenberg, C., Ruppert, V., Arbustini, E., Wichter, T., Germain, M., Dubourg, O., Tavazzi, L., Aumont, M. C., de Groote, P., Fauchier, L., Trochu, J. N., Gibelin, P., Aupetit, J. F., Stark, K., … Cambien, F. (2011). A genome-wide association study identifies two loci associated with heart failure due to dilated cardiomyopathy. European Heart Journal, 32(9). https://doi.org/10.1093/eurheartj/ehr105
- Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. In American Journal of Human Genetics (Vol. 90, Issue 1). https://doi.org/10.1016/j.ajhg.2011.11.029
- Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. In American Journal of Human Genetics (Vol. 101, Issue 1). https://doi.org/10.1016/j.ajhg.2017.06.005

- Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B., & Petrovski, S. (2021). Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-21790-4
- Walker, V. M., Zheng, J., Gaunt, T. R., & Smith, G. D. (2022). Phenotypic Causal Inference Using Genome-Wide Association Study Data: Mendelian Randomization and Beyond. In Annual Review of Biomedical Data Science (Vol. 5). https://doi.org/10.1146/annurev-biodatasci-122120-024910
- Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genetics, 16(4). https://doi.org/10.1371/journal.pgen.1008720
- Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genetics, 17(9). https://doi.org/10.1371/journal.pgen.1009440
- Walsh, R., Thomson, K. L., Ware, J. S., Funke, B. H., Woodley, J., McGuire, K. J., Mazzarotto, F., Blair, E., Seller, A., Taylor, J. C., Minikel, E. v., MacArthur, D. G., Farrall, M., Cook, S. A., & Watkins, H. (2017). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genetics in Medicine, 19(2). https://doi.org/10.1038/gim.2016.90
- Wang, L., Geist, J., Grogan, A., Hu, L. Y. R., & Kontrogianni-Konstantopoulos, A. (2018). Thick filament protein network, functions, and disease association. Comprehensive Physiology, 8(2). https://doi.org/10.1002/cphy.c170023
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. In Nature Reviews Genetics (Vol. 10, Issue 1). https://doi.org/10.1038/nrg2484
- Watkins, H. (2021). Time to Think Differently about Sarcomere-Negative Hypertrophic Cardiomyopathy. In Circulation (Vol. 143, Issue 25). https://doi.org/10.1161/CIRCULATIONAHA.121.053527
- Watkins, H., Conner, D., Thierfelder, L., Jarcho, J. A., MacRae, C., McKenna, W. J., Maron, B. J., Seidman, J. G., & Seidman, C. E. (1995). Mutations in the cardiac myosin binding protein–C gene on chromosome 11 cause familial hypertrophic cardiomyopathy. Nature Genetics, 11(4). https://doi.org/10.1038/ng1295-434
- Watkins, H., Ashrafian, H., & Redwood, C. (2011). Inherited Cardiomyopathies. New England Journal of Medicine, 364(17), 1643–1656. https://doi.org/10.1056/NEJMra0902923
- Way, G. P., Youngstrom, D. W., Hankenson, K. D., Greene, C. S., & Grant, S. F. A. (2017). Implicating candidate genes at GWAS signals by leveraging topologically associating domains. European Journal of Human Genetics, 25(11). https://doi.org/10.1038/ejhg.2017.108
- Weeks, E. M., Ulirsch, J. C., Cheng, N. Y., Trippe, B. L., Fine, R. S., Miao, J., Patwardhan, T. A., Kanai, M., Nasser, J., Fulco, C. P., Tashman, K. C., Aguet, F., Li, T., Ordovas-Montanes, J., Smillie, C. S., Biton, M., Shalek, A. K., Ananthakrishnan, A. N., Xavier, R. J., … Finucane, H. K. (2023). Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Nature Genetics, 55(8). https://doi.org/10.1038/s41588-023-01443-6
- Wei, W., & Yu, X. D. (2007). Comparative Analysis of Regulatory Motif Discovery Tools for Transcription Factor Binding Sites. Genomics, Proteomics and Bioinformatics, 5(2). https://doi.org/10.1016/S1672-0229(07)60023-0

- Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., & di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-13212-3
- Westendorp, B., Major, J. L., Nader, M., Salih, M., Leenen, F. H. H., & Tuana, B. S. (2012). The E2F6 repressor activates gene expression in myocardium resulting in dilated cardiomyopathy. The FASEB Journal, 26(6). https://doi.org/10.1096/fj.11-203174
- White, M. A. (2015). Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. Genomics, 106(3). https://doi.org/10.1016/j.ygeno.2015.06.003
- Wilkinson, J. D., Lowe, A. M., Salbert, B. A., Sleeper, L. A., Colan, S. D., Cox, G. F., Towbin, J. A., Connuck, D. M., Messere, J. E., & Lipshultz, S. E. (2012). Outcomes in children with Noonan syndrome and hypertrophic cardiomyopathy: A study from the Pediatric Cardiomyopathy Registry. American Heart Journal, 164(3). https://doi.org/10.1016/j.ahj.2012.04.018
- Witjas-Paalberends, E. R., Güclü, A., Germans, T., Knaapen, P., Harms, H. J., Vermeer, A. M. C., Christiaans, I., Wilde, A. A. M., DosRemedios, C., Lammertsma, A. A., van Rossum, A. C., Stienen, G. J. M., van Slegtenhorst, M., Schinkel, A. F., Michels, M., Ho, C. Y., Poggesi, C., & van der Velden, J. (2014). Gene-specific increase in the energetic cost of contraction in hypertrophic cardiomyopathy caused by thick filament mutations. Cardiovascular Research, 103(2). https://doi.org/10.1093/cvr/cvu127
- Wolffe, A. P. (1998). The Cell: A Molecular Approach. Geoffrey M. Cooper . In The Quarterly Review of Biology (Vol. 73, Issue 1). https://doi.org/10.1086/420083
- Wooten, E. C., Hebl, V. B., Wolf, M. J., Greytak, S. R., Orr, N. M., Draper, I., Calvino, J. E., Kapur, N. K., Maron, M. S., Kullo, I. J., Ommen, S. R., Bos, J. M., Ackerman, M. J., & Huggins, G. S. (2013). Formin homology 2 domain containing 3 variants associated with hypertrophic cardiomyopathy. In Circulation: Cardiovascular Genetics (Vol. 6, Issue 1). https://doi.org/10.1161/CIRCGENETICS.112.965277
- Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y. H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T. H., D'Ambrosio, D., Gallins, P., Ha, M. J., Hottenga, J. J., … Boomsma, D. I. (2014). Heritability and genomics of gene expression in peripheral blood. Nature Genetics, 46(5). https://doi.org/10.1038/ng.2951
- Yu, M., Harper, A. R., Aguirre, M., Pittman, M., Tcheandjieu, C., Amgalan, D., Grace, C., Goel, A., Farrall, M., Xiao, K., Engreitz, J., Pollard, K. S., Watkins, H., & Priest, J. R. (2023). Genetic Determinants of the Interventricular Septum Are Linked to Ventricular Septal Defects and Hypertrophic Cardiomyopathy. Circulation: Genomic and Precision Medicine, 16(3). https://doi.org/10.1161/CIRCGEN.122.003708
- Yuan, H. Y., Chiou, J. J., Tseng, W. H., Liu, C. H., Liu, C. K., Lin, Y. J., Wang, H. H., Yao, A., Chen, Y. T., & Hsu, C. N. (2006). FASTSNP: An always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Research, 34(WEB. SERV. ISS.). https://doi.org/10.1093/nar/gkl236
- Zacher, B., Michel, M., Schwalb, B., Cramer, P., Tresch, A., & Gagneur, J. (2017). Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. PLoS ONE, 12(1). https://doi.org/10.1371/journal.pone.0169249
- Zamorano, J. L., Anastasakis, A., Borger, M. A., Borggrefe, M., Cecchi, F., Charron, P., Hagege, A. A., Lafont, A., Limongelli, G., Mahrholdt, H., McKenna, W. J., Mogensen, J.,

Nihoyannopoulos, P., Nistri, S., Piepe, P. G., Pieske, B., Rapezzi, C., Rutten, F. H., Tillmanns, C., … Wolpert, C. (2014). 2014 ESC guidelines on diagnosis and management of hypertrophic cardiomyopathy: The task force for the diagnosis and management of hypertrophic cardiomyopathy of the European Society of Cardiology (ESC). In European Heart Journal (Vol. 35, Issue 39). https://doi.org/10.1093/eurheartj/ehu284

- Zeggini, E., & Ioannidis, J. P. A. (2009). Meta-analysis in genome-wide association studies. In Pharmacogenomics (Vol. 10, Issue 2). https://doi.org/10.2217/14622416.10.2.191
- Zhang, Z., Peng, Z., Cao, J., Wang, J., Hao, Y., Song, K., Wang, Y., Hu, W., & Zhang, X. (2019). Long noncoding RNA PXN-AS1-L promotes non-small cell lung cancer progression via regulating PXN. Cancer Cell International, 19(1). https://doi.org/10.1186/s12935-019-0734-0
- Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA, 26(8). https://doi.org/10.1261/RNA.074922.120
- Zheng, J., Peng, L., Cheng, R., Li, Z., Xie, J., Huang, E., Cheng, J., & Zhao, Q. (2024). RAF1 mutation leading to hypertrophic cardiomyopathy in a Chinese family with a history of sudden cardiac death: A diagnostic insight into Noonan syndrome. Molecular Genetics and Genomic Medicine, 12(1). https://doi.org/10.1002/mgg3.2290
- Zhou, H., Shrikumar, A., & Kundaje, A. (2021). Towards a Better Understanding of Reverse-Complement Equivariance for Deep Learning Models in Regulatory Genomics. Proceedings of Machine Learning Research, 165.
- Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nature Methods, 12(10). https://doi.org/10.1038/nmeth.3547
- Zhuge, Y., Zhang, J., Qian, F., Wen, Z., Niu, C., Xu, K., Ji, H., Rong, X., Chu, M., & Jia, C. (2020). Role of smooth muscle cells in cardiovascular disease. International Journal of Biological Sciences, 16(14). https://doi.org/10.7150/ijbs.49871

Words: 40303