

# DNA METHYLATION ANALYSIS OF BICUSPID AORTIC VALVE IN TURNER SYNDROME

by

Jacob Gutierrez

## **A Thesis**

Submitted to Oregon Health & Science University  
in partial fulfillment of the requirements for the degree of

## **Master of Science**

School of Medicine  
Department of Medical Informatics and Clinical Epidemiology  
Portland, Oregon  
June 2021

# CERTIFICATE OF APPROVAL

This is to certify that the Master's Thesis of

**Student Name**

has been approved

\_\_\_\_\_  
Lucia Carbone

Mentor

\_\_\_\_\_  
Guanming Wu

Member

\_\_\_\_\_  
Cheryl Maslen

Member

## Acknowledgements

I would like to thank my mentor and advisor Cheryl Maslen for all of her guidance and mentorship throughout this thesis project. This project has given me valuable first hand experience of watching a project be taken from proposal to completion which has given me a greater appreciation for the work that goes into basic research science. I would like to express my gratitude to my thesis advisory committee mentor Lucia Carbone for her support and the support of her team including, Brett Davis and Jake VanCampen, for without their guidance and expertise this project would not have reached success. Many thanks to Guanming Wu for being a leader in knowledge bases such as reactome by providing guidance and assistance in the interpretation of results from these resources.

I am grateful to the Department of Medical Informatics and Clinical Epidemiology including the students, faculty, and staff. I am especially thankful for Diane Doctor for working with me to complete this program through the difficulty of this pandemic.

Many thanks to the GenTAC alliance and to the study participants with Turner Syndrome without whose time and contribution this project would not be possible.

Finally, I would like to thank my parents and friends for their support and encouragement which pushed me to success during these difficult times.

# Table of Contents

<b>Table of Contents</b>	<b>1</b>
<b>Chapter 1: Introduction and Background</b>	<b>4</b>
Turner Syndrome and Bicuspid Aortic Valve	4
<b>Genetics and Epigenetics of Turner Syndrome</b>	<b>5</b>
<b>Genetics and Epigenetics of Bicuspid Aortic Valve</b>	<b>6</b>
<b>Illumina Methyl Capture Sequencing</b>	<b>7</b>
<b>Motivation and Specific Aims</b>	<b>8</b>
<b>Chapter 2: Performance of Illumina Methyl Capture Sequencing</b>	<b>10</b>
Abstract	10
Introduction	10
Results	11
Samples, Data Quality, and Enrichment Performance	11
Methyl Capture Seq Target Performance	13
Failed Target Assessment	17
Discussion	19
Methods	20
Bisulfite Sequencing	20
Read Alignment and Analysis	20
<b>Chapter 3: DNA Methylation Analysis of Turner Syndrome BAV</b>	<b>22</b>
Abstract	22
Introduction	23
Methods	24
Samples	24
Study Design	24
Illumina Methyl Capture Sequencing	24
Data Processing and QC	25
Differential Methylation Analysis	25
Results	26

Data Quality	26
TS BAV Methylation	27
TS Methylation Alterations Support Previous Findings	31
Discussion	35
<b>Chapter 4: Discussion and Conclusion</b>	<b>38</b>
<b>References</b>	<b>40</b>
<b>Supplemental Tables and Figures</b>	<b>51</b>



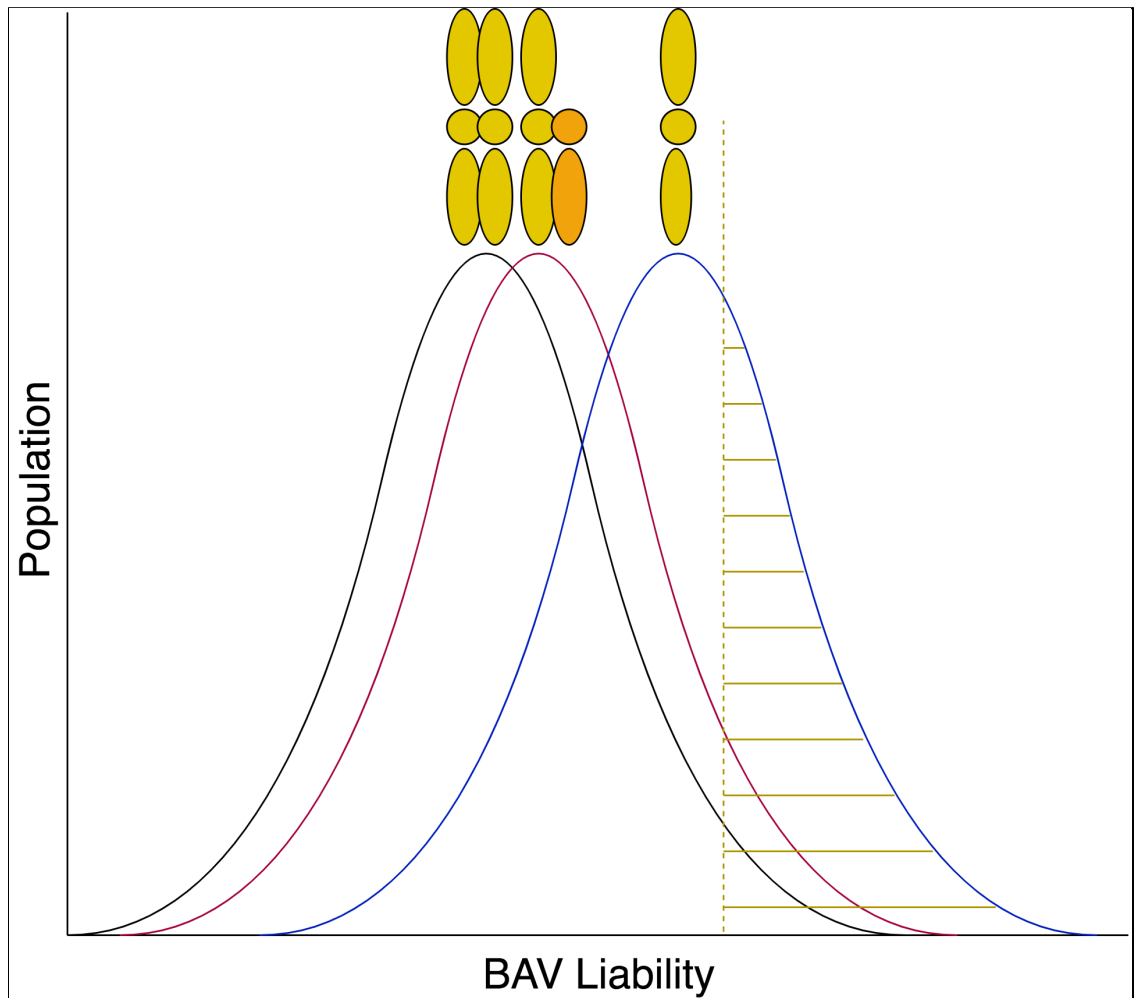
## I. Chapter 1: Introduction and Background

### I.i Turner Syndrome and Bicuspid Aortic Valve

Turner Syndrome (TS) is a rare cytogenetic disorder caused by the partial or complete loss of a second sex chromosome, which occurs in 1 in 2,000 female live births <sup>1</sup>. Approximately 50% of TS cases occur with the 45,X karyotype and the remaining 50% is made of other combinations of 45,X with additional chromosomal material (e.g. 45,X,ringX), or as mosaics e.g. 45,X/46,XX <sup>2</sup>. Girls with TS show a variety of clinical manifestations including short stature, premature ovarian failure, webbed neck, specific cognitive/visual spatial disabilities, hearing loss, thyroid dysfunction, scoliosis, endocrine disorders, autoimmune disorders, and cardiovascular disease. The most common cause of early mortality in TS is due to congenital heart defects, where patients with the most common 45,X karyotype having the highest burden of congenital defects and negative outcomes <sup>3-5</sup>. In addition to the increased post-natal cardiovascular defect related mortality risk, it is thought that over 99% of 45,X embryos are lost in utero with an increased prevalence for left-sided obstructive lesions otherwise known as Left Sided Heart Lesions (LSHL)<sup>6-8</sup>.

Bicuspid Aortic Valve (BAV) is the most common congenital heart defect in the general population with a prevalence of 0.5-2% and represents a mild form of LSHL <sup>9,10</sup>. BAV is where the aortic valve consists of two leaflets as opposed to the normal three leaflet configuration. There is a significant sex bias, where males account for approximately 75% of all BAV cases <sup>11</sup>. The specific negative cardiovascular outcomes of BAV include valve calcification, stenosis, aortic endocarditis, aortic dilation, and aortic aneurysm, collectively known as aortopathy. BAV is clinically relevant as approximately 40% of patients go on to develop some form of aortopathy in their lifetime <sup>11</sup>.

TS patients with the 45,X karyotype have the highest burden of BAV with a prevalence around 30% with near complete penetrance of aortopathy <sup>12</sup>. The high incidence of BAV in TS females with the 45,X karyotype and the high burden of BAV in karyotypically normal 46,XY males suggests the genetics of a second X chromosome offers protection against the development of BAV and BAV associated aortopathy in 46,XX females. Conversely, it can be thought that 45,X TS individuals are predisposed to develop BAV due to lacking a second sex chromosome leading to an increased rate of BAV in this population (Figure 1). Due to the obvious link of BAV in TS and the X chromosome, this is an instance of studying rare disease, TS, to inform common disease, BAV.



**Figure 1: BAV liability distribution for 46,XX females (black), 46,XY males (red) and 45,X TS females (blue) indicating that lack of X chromosome material predisposes individuals to higher rates of BAV.**

## I.ii Genetics and Epigenetics of Turner Syndrome

Due to the high prevalence of TS patients having different proportions of missing X chromosome material, cytogenetic mappings have been done in order to associate X chromosome regions to phenotypes<sup>13</sup>. Although this may allow researchers to identify larger regions of interest, very often specific genes associated with TS phenotypes cannot be identified with the only widely accepted genotype-phenotype correlation found in TS is the *SHOX* gene, found on pseudoautosomal region 1, linked with short stature and skeletal anomalies<sup>14-16</sup>. Early studies attempted to identify X chromosome loci associated with BAV, these approaches were hampered by the lack of study subjects with X chromosome deletions in order to robustly identify these associations<sup>17</sup>. Additional studies have identified that TS patients with short arm X chromosome (Xp) deletions have higher risk of BAV suggesting the lack of genes (haploinsufficiency) residing on Xp contribute to congenital heart defects



<sup>18</sup>. To expand on these observations, a TS BAV Copy Number Variation (CNV) study was able to confirm the BAV and Xp deletion association <sup>19</sup>. This study also identified various genomic loci and potential genes that were associated with BAV which includes a CNV on 12p13.31 overlapping the genes *SLC2A3*, *SLC2A14* and *NANOGP1* in addition to other rare CNVs which notably overlaps the *HOXA* gene cassette. A recent study from the Maslen lab has identified *TIMP1* and *TIMP3* as being associated with BAV and aortopathy through whole exome sequencing, which identified common variants in *TIMP3* to be associated with BAV in TS subjects <sup>20</sup>. This observation prompted an investigation of the *TIMP3* homologue, *TIMP1*, which resides on the long arm of the X chromosome. This led to the discovery that if only 1 copy of *TIMP1* is present, as in the case of 45,X karyotype, a patient with TS is 4 times more likely to have BAV. Variants in *TIMP3*, located on the long arm of chromosome 22, were found to sensitize TS patients with BAV to have a higher likelihood of thoracic aortic aneurysm. Both *TIMP1* and *TIMP3* play a critical role in extracellular matrix homeostasis in the aorta, and in aortic valve development <sup>21</sup>. The association of BAV and aortopathy with genes that are involved in the maintenance of the aortic wall and development of the aortic valve provides biological evidence for a causal effect of *TIMP1/3* deficiency. Although these genes show a very significant association with BAV, *TIMP1/3* deficiency only explains roughly 20% of the occurrence of BAV in TS. The large amount of unexplained risk necessitates the search for additional BAV risk factors in TS.

Epigenetic analysis of TS patients revealed genome-wide hypomethylation when compared to healthy 46,XX female and 46,XY males <sup>22,23</sup>. Differentially methylated genes are enriched in pathways that are known to be altered in TS patients having symptoms such as cardiovascular risk, short stature, autoimmune diseases, gonadal dysgenesis, metabolic syndromes, and epigenetic regulatory mechanisms <sup>23</sup>. When pairing the epigenetic analysis with mRNA sequencing, it was found that many genes not detected to be differentially methylated were differentially expressed. The authors of this study interpreted the lack of a strong correlation between methylation and gene expression to suggest that a simple gene dosage relationship (lack of X chromosome gene function directly causing TS phenotypes) is not present in TS. Rather than a simple genotype-phenotype relationship, the missing chromosomal material causes profound changes in regulatory pathways triggering a complex cascade of cellular events creating the observed phenotypes of individuals with TS. The authors suggest that the detected methylation changes found could represent epigenetic memory, where changes are accumulated during embryonic development but may remain dormant within adult tissue or have functional consequences in others <sup>24,25</sup>. Evidence for this interpretation is growing with functional blood DNAm alterations related to gestational age which may be associated with increased susceptibility to chronic diseases in these individuals <sup>26</sup>. In addition, identification of DNAm alterations in many genes relevant in heart development being found, such as *NOX5*, *PRDM16*, and *TBX20* in various epigenetic studies of congenital heart disease with some DNAm alterations correlated to changes in gene expression in primary heart tissue which suggests that these changes in genes important to heart development are found in adult tissues <sup>27-30</sup>. These studies highlight the possible deregulation of epigenetic mechanisms in TS individuals which can lead to the observed clinical features and high incidence of adverse outcomes.

### I.iii Genetics and Epigenetics of Bicuspid Aortic Valve

BAV has two basic etiologic classifications, syndromic and non-syndromic BAV <sup>9</sup>. Syndromic BAV is where BAV occurs in individuals who have the heart defect as part of a clinically defined syndrome

such as TS, Marfan syndrome, or Loeys-Dietz syndrome. Non-syndromic BAV occurs in individuals as an isolated trait with no accompanying features. However, unlike TS, BAV occurs infrequently in Marfan and Loeys-Dietz syndromes which are rare disorders and therefore contribute little to our understanding of BAV genetics<sup>31,32</sup>. Despite the high prevalence in the general population, most of the etiology of BAV is not known although a genetic component has been identified as ~10% of BAV is familial<sup>33</sup>. Mutations in *NOTCH1*, *GATA5*, *NKX2.5*, and *ROBO4* are known to cause non-syndromic BAV in some families, but the majority of BAV cases are unexplained<sup>34-37</sup>. Contributing genes with both types of BAV are highlighted in Table 1. Overall, these studies have shown that BAV is a complex and heterogeneous condition with multiple genes contributing only a small number of cases.

The epigenetic associations within BAV have also been explored due to the predictive value of epigenetics found in the larger cardiovascular disease context. In one study, the methylation profiles of BAV aortic wall tissue revealed significant changes in genes associated with heart development<sup>38</sup>. Another methylation profile study of the bicuspid aortic valve itself showed differences in genes that regulate cell proliferation, a known aspect of aneurysm formation<sup>39</sup>. In the realm of non-coding RNAs, there have been many non-coding RNA that have altered expression in BAV individuals whose functions have been validated to interact with critical pathways known to affect development and aneurysm formation<sup>40</sup>. Overall, there is strong evidence of an epigenetic contribution to BAV and congenital heart disease as a whole that has not been fully explored in the general population, let alone in a rare disease context with a high burden of comorbidity associated with BAV<sup>30,41,42</sup>.

<b>Non-Syndromic BAV</b>	<b>Gene Name</b>	<i>NOTCH1</i>	<i>GATA5</i>	<i>NKX2.5</i>	<i>ROBO4</i>
<b>Syndromic BAV</b>	<b>Gene Name</b>	<i>FBNI</i>	<i>TGFBR1/2</i>	<i>COL3A1</i>	<i>TIMP1/3</i>
	<b>Syndrome Name</b>	Marfan syndrome	Loeys-Dietz syndrome	Vascular Ehlers-Danlos syndrome	Turner syndrome

**Table 1: Genes known to contribute to syndromic and non-syndromic BAV**

#### 1.iv Illumina Methyl Capture Sequencing

Illumina TruSeq Methyl Capture EPIC Sequencing (Methyl Capture Seq) represents a new NGS method for genome-wide methylation studies<sup>43</sup>. This approach generates sequencing libraries enriched for functional regions of the genome (e.g. CpG islands, promoters, enhancers) for a fraction of the cost of whole-genome bisulfite sequencing (WGBS). In contrast, microarray technology is a cost-effective approach to large scale studies of DNA methylation (DNAm), but with limited coverage of total CpG sites in the genome with low resolution for methylation heterogeneity over larger regions when compared to sequencing based approaches. Together, this makes Methyl Capture seq an attractive option to compromise between high resolution and relatively low-cost genome-wide methylome assessment.

Methyl Capture Seq is a 4-plex library preparation strategy where 4 samples are processed through library prep and probe capture as one pool. Three capture pools are then combined with Illumina

adapters that are designed to scale to a 12-plex sequencing pool. Capture library preparation is performed by fragmenting high quality genomic DNA into ~200bp fragments. These fragments are end-repaired, A-tailed, and ligated to methylated Illumina adapters generating pre-capture libraries in pools of 4 samples. These pre-capture libraries are then hybridized with Illumina EPIC oligo probes to enrich these libraries for targeted regions. These enriched libraries are then bisulfite converted and PCR amplified to generate sequencing-ready libraries.

Due to the recent development of this technology, there are few published studies using this method. A literature search for this technology has returned a total of 10 peer reviewed publications as of April 2021, 5 more than the previous year. Earlier studies focused on comparing this method to other DNAm platforms including WGBS, Reduced Representation Bisulfite Sequencing, Agilent Methyl Capture Sequencing, Roche Methyl Capture Sequencing, and the Illumina EPIC methylation microarray<sup>44,45</sup>. These studies have found that this Methyl Capture seq is largely comparable to existing methods and performs robustly yielding comparable results to existing methods across replicate samples. A recent study by Lin et al. 2020 directly compare Methyl Capture seq to the Illumina EPIC DNAm microarray and found that the targeted sequencing approach offered a greater dynamic range and higher quantitative sensitivity to CpGs with invariant methylation levels (close to 0 or 1)<sup>46</sup>. Together these studies highlight that this platform offers benefits compared to existing platforms by detecting more CpGs with a higher sensitivity than microarrays, at a lower cost than WGBS. All previous characterization studies utilized replicate samples in order to evaluate the performance of methylation detection of this kit. However, there are no studies characterizing the performance of Methyl Capture seq against the regions stated to be targeted by the manufacturer across multiple samples from independent experiments. Such an analysis would give insight into the general performance of this kit and inform individuals interested in using this technique what data is to be expected for downstream methylation analysis and about the benefits and caveats of this approach.

#### I.v Motivation and Specific Aims

The target enrichment efficiency of the Illumina Methyl Capture Seq kit has not been explored thus far. Documentation for this platform provided by the manufacturer describes high correlation of replicate samples but there are no details on the performance across many samples in diverse tissues. Specifically, the background level and characterization of targeted regions that are consistently missed by this assay have not been reported.

##### **SA1: Methyl Capture Sequencing Targeting Enrichment Assessment.**

To assess the efficiency of Methyl Capture sequencing target enrichment, the published Illumina TruSeq Methyl Capture EPIC manifest will be compared to generated Methyl Capture Sequencing data. Previously generated datasets (n = 44) and a newly generated WGBS data set (n = 12) will be used for this analysis. The level of background from the capture experiments will be assessed and a capture efficiency metric will be computed for each sample. Underperforming targets will be detected and assessed on the UCSC genome browser to identify their genomic composition and determine if they fall in “problematic” genomic regions (e.g. repeat elements, alternative haplotypes, segmental duplications, high GC, low mappability).

There have been many studies into the epigenetic alterations of BAV within karyotypically normal individuals. DNA methylation alterations have been detected in primary aortic wall tissue and

non-coding RNA expression changes have been detected in blood<sup>38,40</sup>. These two findings suggest there is a strong epigenetic component to BAV. Paired with the fact TS individuals with the 45,X karyotype have a 30% increased prevalence of BAV, 45,X TS represents a BAV sensitized group whose epigenetic changes may be larger than the general population. In addition to this, studies suggest that epigenetic mechanisms are deregulated in TS and that these changes in DNAm could represent epigenetic memory that is detectable in blood<sup>23</sup>. Due to the obvious link of BAV in TS this is an instance of studying rare disease (TS) to inform common disease (BAV)

**SA2: Differential Methylation of 45,X TS BAV v. TAV and TS BAV v. 46,XX BAV.**

In order to test the hypothesis that there are DNAm differences between BAV and non-BAV in TS. Illumina TruSeq-Methyl Capture EPIC methylation sequencing (Methyl Capture Seq) will be performed on 12 BAV and 13 TAV whole blood genomic DNA from non-smoking TS individuals with 45,X karyotype. An additional 6 euploid 46,XX BAV are included to analyze DNAm differences associated with TS alone and followed the same study inclusion criteria. Differentially methylated regions (DMR) will be detected by linear regression using Limma adjusted for age and cell type composition via surrogate variable analysis followed by DMR detection using Comb-P<sup>47-49</sup>. These DMRs will then be assessed for biological significance using genomic feature enrichment, transcription factor motif enrichment, and the use of databases such as DAVID, GREAT, and STRING<sup>50-52</sup>.

## II. Chapter 2: Performance of Illumina Methyl Capture Sequencing

### II.i Abstract

Illumina TruSeq Methyl Capture EPIC Sequencing (Methyl Capture Seq) represents a new NGS method for genome-wide methylation studies. There are no studies that characterize the performance of this method across multiple experiments, to assess the efficiency in enrichment from the probes and background levels. In this study, Methyl Capture Seq data generated from multiple studies is compared to the published targeted regions to characterize the performance of the oligo probes used to enrich genomic regions of interest and identify targets that consistently lack data across multiple experiments. We find that over 90% of targeted regions are captured across all samples, another 5% of targeted regions consistently lacking read coverage, and the remaining 5% of targets having sporadic performance. These failed targeted regions show significant enrichment for genomic loci that overlap alternative haplotypes. Comparing Methyl Capture Seq to WGBS data, both methods show lack of coverage for these alternative haplotype loci, suggesting the lack of coverage for consistently failing targets is caused by the read alignment step, rather than the pull-down. Overall, Methyl Capture Seq offers a high-performance targeted bisulfite sequencing approach to analyze known functional elements for studies.

### II.ii Introduction

Illumina TruSeq Methyl Capture EPIC Sequencing (Methyl Capture Seq) represents a new NGS method for genome-wide methylation studies<sup>43</sup>. This approach generates sequencing libraries enriched for functional regions of the genome (e.g. CpG islands, promoters, enhancers) for a fraction of the cost of whole-genome bisulfite sequencing (WGBS). In contrast, microarray technology is a cost-effective approach to large scale studies of DNA methylation (DNAm), but with limited coverage of total CpG sites in the genome when compared to sequencing based approaches. Together this makes Methyl Capture seq an attractive option to compromise between high resolution and relatively low-cost genome-wide methylome assessment.

Methyl Capture Seq is a 4-plex library preparation strategy where 4 samples are processed through library prep and probe capture as one pool. Three capture pools are then combined with Illumina adapters that are designed to scale to a 12-plex sequencing pool. Capture library preparation is performed by fragmenting high quality genomic DNA into ~200bp fragments. These fragments are end-repaired, A-tailed, and ligated to methylated Illumina adapters generating pre-capture libraries in pools of 4 samples. These pre-capture libraries are then hybridized with Illumina EPIC oligo probes to enrich these libraries for targeted regions. These enriched libraries are then bisulfite converted and PCR amplified to generate sequencing-ready libraries.

Studies have focused on comparing this method to other DNAm platforms including WGBS, Reduced Representation Bisulfite Sequencing, Agilent Methyl Capture Sequencing, Roche Methyl Capture Sequencing, and the Illumina EPIC methylation microarray<sup>44,45</sup>. One reported strength of Methyl Capture seq compared to EPIC methylation microarrays is a larger dynamic range for methylation estimates, the difference between the largest and smallest methylation values, which could lead to greater number of detected differentially methylated positions when comparing sites found across both

platforms<sup>46</sup>. Although methylation microarrays are the most cost-effective approach for large cohort studies, the increasing availability and lowering cost of NGS is pushing researchers to utilize these more cost-effective sequencing applications for smaller studies.

The target enrichment efficiency of the Illumina DNA oligo probes for the Methyl Capture Seq platform has not been explored. Documentation for this platform provided by Illumina describes over 97% correlation for replicate samples. However, the performance of this kit compared across many samples in diverse tissues from separate experiments has yet to be characterized. Specifically, the level of background and characterization of targeted regions that are consistently missed have not been reported.

## II.iii Results

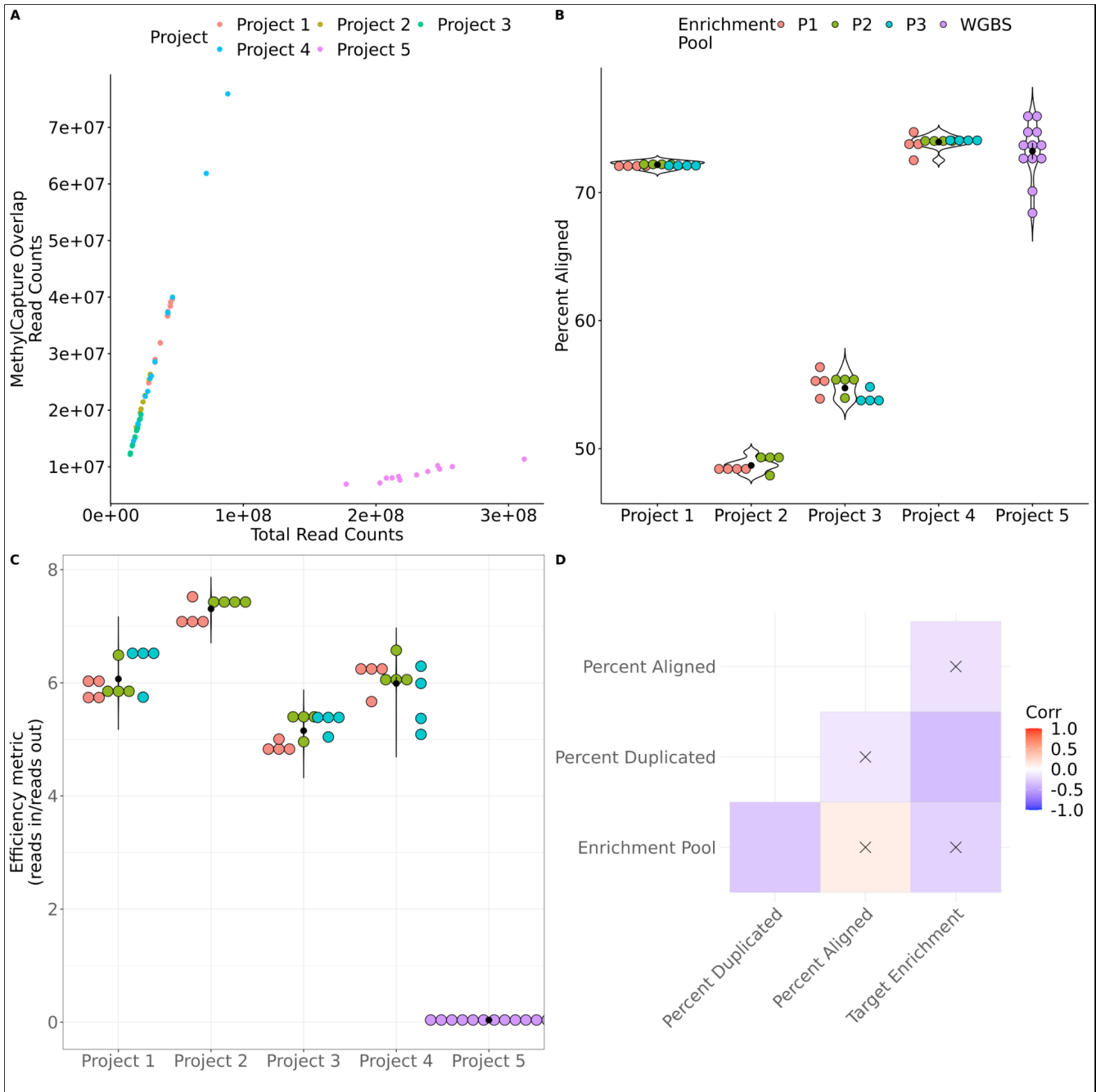
### II.iii.i Samples, Data Quality, and Enrichment Performance

A total of 56 samples from 4 separate MethylCapture seq experiments and 1 WGBS experiment were used to assess the performance of this assay across multiple runs (Table 1). Enrollment and studies for all projects have Internal Review Board approval and all study subjects had informed consent for participation. All project samples were processed through the Oregon Health & Science University KCVI Epigenetics Core. Project 1 and 4 assessed DNAm changes associated with gestational diabetes mellitus, project 2 assessed DNAm changes during neuronal differentiation, and project 3 and 5 assessed DNAm changes associated with bicuspid aortic valve in Turner syndrome. All projects also contain treatment groups and disease cases/controls that are not shown. Samples come from a diverse set of tissues, which allows identification of potential tissue specific effects. The nature of SA1 analysis is to assess the DNA capture step, which should not be affected by DNAm variation between these samples.

Project Name	Project 1		Project 2	Project 3	Project 4	Project 5
Total Sample	12		8	12	12	12
Tissue (n)	Placenta (10)	HeLa (2)	Stem Cells (8)	Blood (12)	Placenta (12)	Blood (12)
Library Preparation	Methyl Capture Sequencing		Methyl Capture Sequencing	Methyl Capture Sequencing	Methyl Capture Sequencing	Whole Genome Bisulfite Sequencing
Sequencing Platform	Illumina NextSeq 500		Illumina NextSeq 500	Illumina NextSeq 500	Illumina NextSeq 500	Illumina NovaSeq 6000
Sequencing Type	150bp Paired End		150bp Paired End	120bp Paired End	140bp Paired End	150bp Paired End

**Table 1: Summary Statistics for Available Methyl Capture Sequencing and Whole Genome Bisulfite Sequencing Data.**

All samples have sufficient read coverage with each sample having at least 8 million reads allowing all samples to be used for differential methylation analysis. Methyl Capture Seq samples show a linear relationship between the total amount of sequencing done and the number of reads overlapping MethylCapture seq targeted regions (Figure 2A). As expected, WGBS data does not display any such relationship due to the lack of enrichment during library preparation. Bisulfite sequencing data is known to suffer from reduced read alignment due to reduced library complexity following bisulfite conversion<sup>53,54</sup>. Each project shows alignment rates that are highly comparable within each project but has high variability across projects which could be influenced by a variety of technical factors including DNA source and DNA quality (Figure 2B). To compare Methyl Capture seq performance across all projects, an efficiency metric was computed which indicates the fold-enrichment of reads overlapping targeted regions within each sample. All Methyl Capture seq samples show ~6-fold read enrichment in targeted regions (Figure 2C); project 2 displayed the lowest alignment rate but also the most enrichment (~7-fold) which indicates robust performance of the oligo probe pulldown step used in this method. Batch effects are of particular concern within DNA methylation analysis, to assess the contribution of the 4-plex enrichment step correlation analysis between the enrichment pools used and alignment features was done. The enrichment pool used shows significant correlation (p-value < .05) to both the alignment rate for each sample and the level of target enrichment indicating the presence of batch effects derived from the enrichment step (Figure 2D).



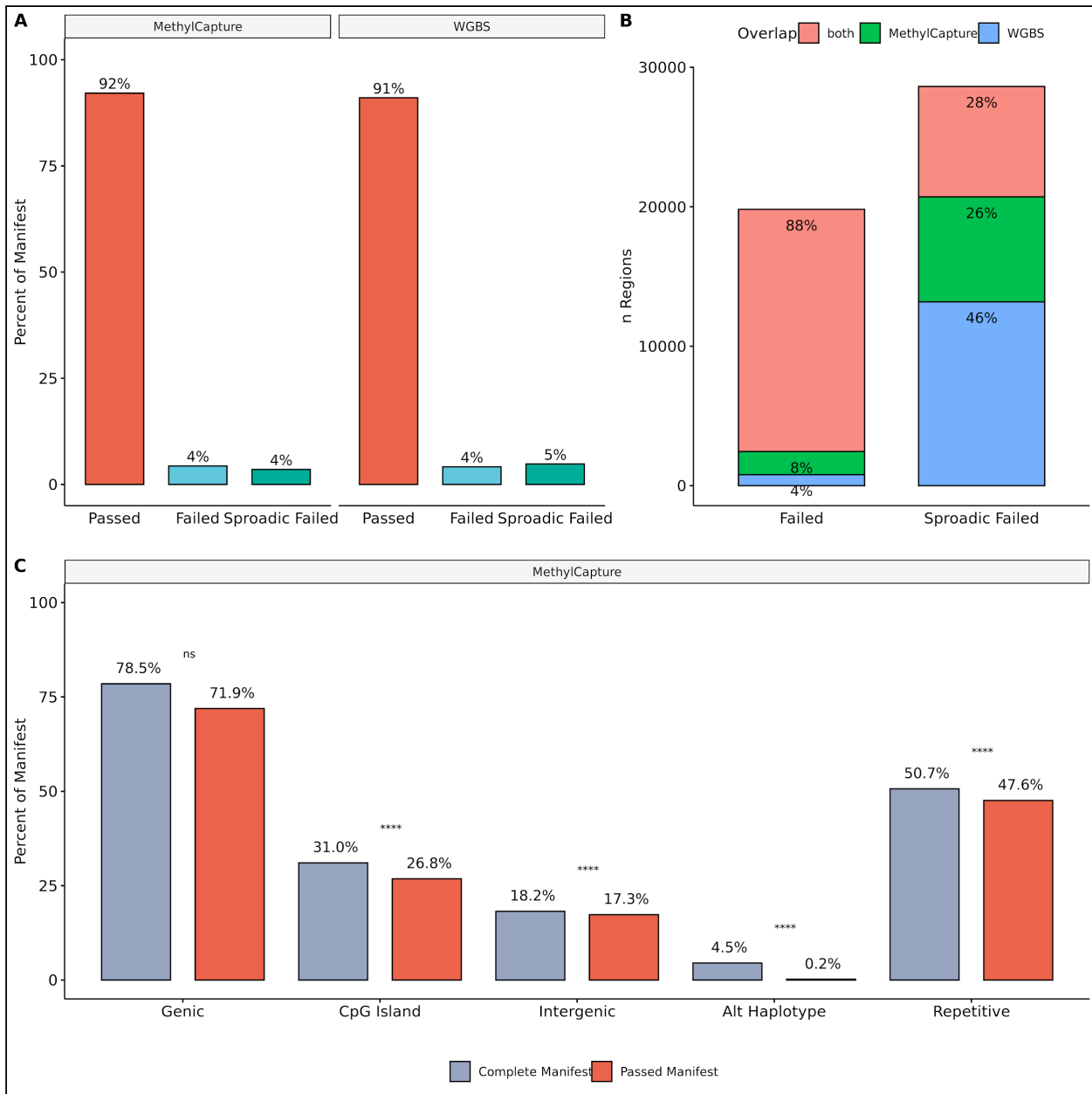
**Figure 2: MethylCapture seq samples display consistent mapping and target enrichment within the same project and enrichment pool.**  
**A.** Scatter plot of total read counts by read counts overlapping targeted regions colored by project. **B.** Violin-Dotplot of the percent of aligned reads for each project colored by enrichment pool. **C.** Dotplot of project efficiency metric with samples colored by enrichment pool. **D.** Correlation plot of Methyl Capture seq samples.

### II.iii.ii Methyl Capture Seq Target Performance

Methyl Capture seq shows robust capture efficiency with over 92% of the targeted regions being captured across all enrichment pools (Figure 3A). Of the 437,447 total targets within the probe-set,



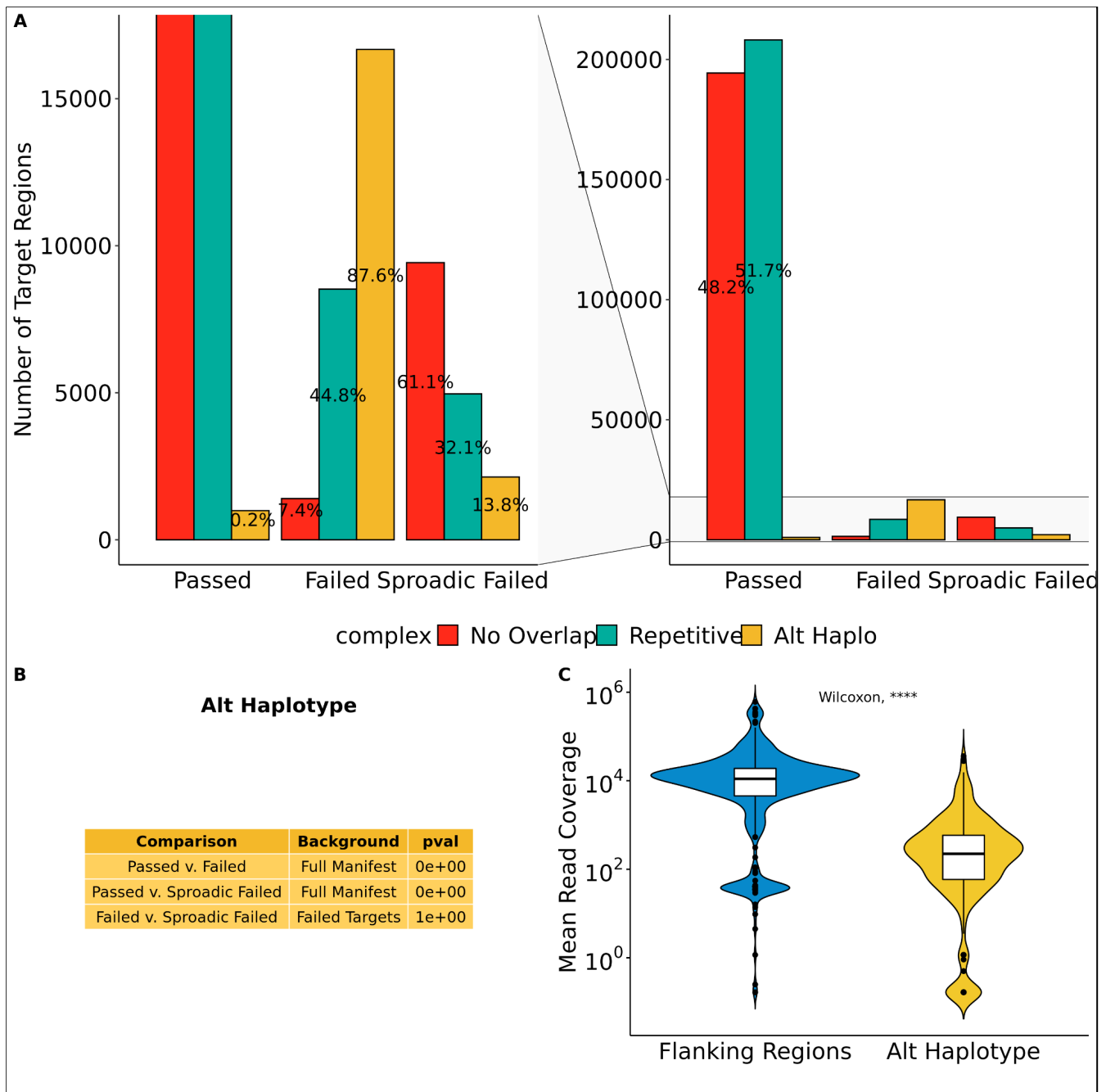
34475 (~ 8%) targets were not captured in at least one enrichment pool. Interestingly, we observed two types of failed capture regions: 1) those that fail in at least one enrichment pool (Sporadic Failed, n = 15443) and 2) those that failed across all enrichment pools (Failed, n = 19032). Of note, when looking at WGBS data we observed a similar read coverage for regions corresponding to the probes (Figure 3A). Moreover, the regions that fully failed in the capture experiments showed a 88% overlap with regions without coverage in the WGBS experiment, suggesting that the source of missing coverage is independent of the pull-down process (Figure 3B). To investigate genomic features that could have reduced coverage, we annotated the probe-set to identify overlap with genic (promoters, exons, introns), CpG island, intergenic, repetitive, and alternative haplotype (including both alternative haplotype and fix patches) features. Passed regions which always returned sequencing reads were compared to the full target regions in order to quantify the number of features available for downstream analysis. Significant change in the coverage of genomic features was found, generally of a small number of regions, however there was a striking decrease in the amount of alternative haplotype overlapping regions (Figure 3C).



**Figure 3: The pull-down in the Methyl Capture seq assay works for 90% of targeted regions. A. Percent of regions with read coverage across all enrichment pools or samples (Passed), regions consistently lacking read coverage (Failed), and regions without read coverage in at least one enrichment pool or sample (Sporadic Failed). B. Barplot of overlap between Failed or Sporadic Failed regions when Methyl Capture seq and WGBS data are compared. C. Barplot of Methyl Capture seq probe-set coverage for genomic features comparing the full probe-set to Passed regions that consistently have read coverage across all enrichment pools.**

Target regions appeared to have two distinct types of failure, with failed regions showing overlap within both Methyl Capture seq and WGBS data in addition to Passed regions having a significant decrease in the number of regions with alternative haplotype overlap. When this analysis is expanded across all target types, over 89% of Failed regions overlapped alternative haplotypes (Figure 4A). To

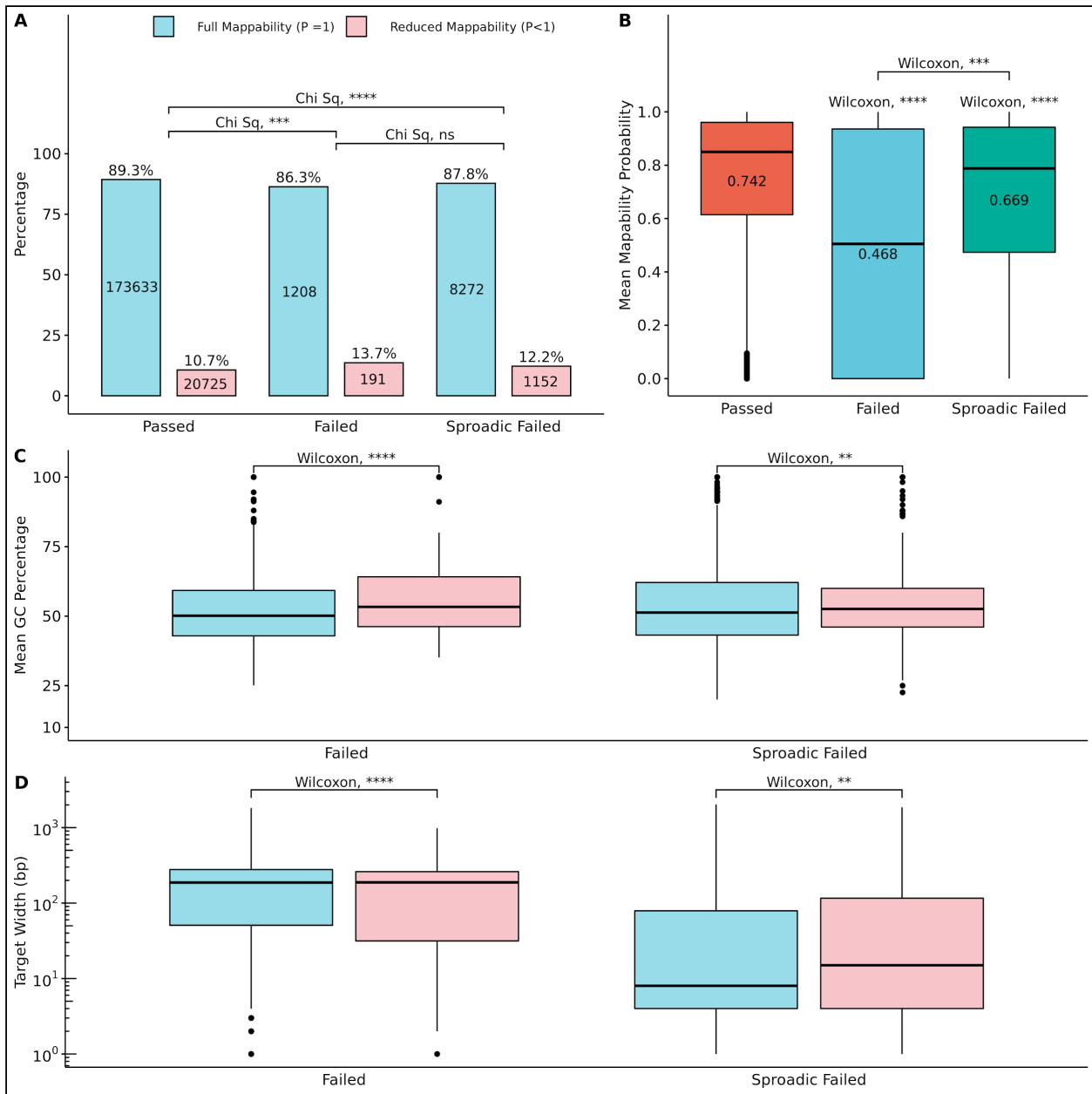
determine the significance of this enrichment, all targeted regions were assumed to be independent of genomic features and have an equal likelihood of being classified as Failed or Sporadic Failed at the rates observed. Permutation testing showed significant enrichment for alternative haplotype regions within both Failed and Sporadic Failed regions (Figure 4B). To extend this analysis to WGBS data, alternative haplotypes over the whole genome were obtained, together with a set of flanking regions of the same genomic size for each location. It was found that alternative haplotypes display a significant decrease in read coverage when compared to their adjacent flanking regions (Figure 4C). Together, these results implicate the read mapping step to be the largest contributor of poor performance in target regions that consistently lack read coverage in both Methyl Capture and WGBS.



**Figure 4: Regions overlapping Alternative Haplotype show a lack of read coverage across platforms. A. Bar chart of full Methyl Capture seq probe-set distribution for each target type colored by genomic feature. B. Summary table for permutation test for alternative haplotype regions, null hypothesis is no enrichment compared to random with two tailed hypothesis testing. C. Violin plot of WGBS read coverage for alternative haplotype regions compared to flanking regions normalized to the same number of basepairs for each region, for example a 1kb region would be compared to the 500bp flanking either side.**

### II.iii.iii Failed Target Assessment

Next we sought to characterize targets with poor performance, regions without consistent read coverage, to gain insight into potential mechanisms of failure outside of alternative haplotype overlap. All Methyl Capture seq target regions had predicted bisulfite sequencing mappability annotated with all target regions having 97% predicted mappability. The majority of targets (88.6%) had full predicted mappability ( $P = 1$ ), which led to the stratification of targets to be those with full mappability against those with reduced mappability ( $P < 1$ ). Both Failed and Sporadic Failed targets showed a greater proportion (13% and 12% respectively) of targets with reduced mappability than Passed regions (Figure 5A). Failed targets also displayed significantly reduced mean mappability (p-values  $3.6e-15$  and  $1.4e-04$  respectively) when compared to both Passed and Sporadic Failed regions (Figure 5B). To assess the contribution of known biases of Illumina short read sequencing, GC content and the genomic size, in bp, of all targets was computed. It was found that Failed targets with reduced mappability also display increased GC content compared to fully mappability regions (Figure 5C). Lastly, sporadic failed targets with full mappability had smaller target genomic size when compared to those targets with reduced mappability (Figure 5D). Together these results indicate that targets with poor performance could be caused by a variety of factors including read mappability, GC content, and genomic target size.



**Figure 5: Regions with no Alternative Haplotype overlap show reduced mappability. A: Barplots of targets with no overlap to alternative haplotypes colored by proportion of targets with reduced mappability. B: Boxplots of mappability within targets with reduced mappability. C: Boxplots of mean GC content for targets with no overlap to alternative haplotypes, GC content was computed using 5bp windows. D: Boxplots of target width in bp of targets with no overlap to alternative haplotypes.**

## II.iv Discussion

To our knowledge this study is the first to explore the technical performance of MethylCapture seq across many samples derived from separate studies. It was found that this targeted sequencing approach greatly enriches sequencing libraries for functionally relevant regions of the genome presenting with a ~6-fold read enrichment for these regions. Samples had consistent alignment rates and enrichment levels within the same project which could be influenced by a variety of factors including experimental conditions and DNA quality. Enrichment pools were found to be correlated to the alignment rate and the enrichment levels detected across studies although it is important to note that these correlations could be driven by differences across projects rather than individual library enrichment performance. This enrichment pool effect has not yet been described and could lead to confounding of analysis results in certain study designs. Batch effects such as this are also known to occur for the Illumina Methylation microarray studies where normalization and batch effect correction are required for robust results <sup>55,56</sup>.

To assess the performance of this method, detection of sequencing reads in target regions was investigated. This approach focuses on target regions which have zero read coverage in order to identify regions of the genome that are targeted by this approach but fail to return sequencing data. Such regions could lack read coverage either due to poor performance of the oligos used to target these areas or due to reduced read mapping to these areas. It was found that Methyl Capture seq target regions generally perform well with over 92% of targeted regions consistently retrieving read coverage across all enrichment pools. There was not a significant lack of coverage for genomic regions of interest indicating that these datasets should be highly comparable across multiple projects within these well performing regions. Surprisingly, both Methyl Capture seq and WGBS samples lacked read coverage for a consistent subset of targets that largely overlapped alternative haplotypes. Failed targets were enriched for alternative haplotypes within Methyl Capture seq data and there was a significant decrease in the mean read coverage of alternative haplotype regions found in WGBS data. When excluding regions with alternative haplotypes, both types of failed targets were found to show enrichment for regions with reduced predicted mappability. Stratifying by mappability, there does appear to be differences in GC content and genomic target width which may reflect known biases to Illumina short read sequencing <sup>57</sup>. It is important to note that these poorly performing regions represent less than 8% of the total manifest indicating that the majority of this enrichment method performs well across multiple experimental conditions.

Together, these results suggest that alternative haplotypes lead to reduced read coverage across bisulfite sequencing methods due to reduced mapping efficiency explaining the majority of failed Methyl Capture seq manifest targets. It is known that bisulfite sequencing data suffers from reduced mapping efficiency due to reduced library complexity following bisulfite conversion which may explain the reduced read coverage within these regions <sup>53,54</sup>. Alternative haplotypes could reduce read mapping efficiency at these regions through compounding the read alignment issue by providing mapping locations with high sequence similarity to the canonical genome for these low complexity reads. Only uniquely mapping reads are retained for downstream analysis and thus these regions may have a higher incidence of multi mapping reads leading to the observed lack of data in this analysis. In the context of Methyl Capture seq, there may be very robust oligo probe designs to enrich sequencing libraries for alternative haplotype regions but sequencing data is not returned for downstream use due to this read mapping issue. DNA probes which target regions that have greater mapping performance could be used to increase the amount of sequencing reads that could be used for downstream analysis in later iterations of this platform.

Limitations of this study include the classification of probe-set regions by the absence of reads in targeted regions. By thresholding this analysis to a binary presence or absence of reads, there is no quantitative assessment for the utility of probe-set regions for downstream differential methylation analysis. In addition, the DNA quality of all samples was not available so quality effects could not be explored in this analysis. One strength of this study is the inclusion of multiple samples across separate study conditions to explore the technical performance of this platform. Inclusion of WGBS samples allows for an independent methodology to differentiate the performance of the oligo enrichment from bisulfite sequencing and read mapping.

## II.v Methods

### II.v.i Bisulfite Sequencing

Genomic DNA samples were submitted for Methyl Capture Sequencing or Whole Genome Bisulfite Sequencing at the KCVI Epigenetics Core. The Illumina TruSeq Methyl Capture EPIC Library Prep Kit (TruSeq-Methyl Capture EPIC, cat # FC-151-1002, Illumina Inc., San Diego, CA) was used to generate these sequencing libraries. Briefly, 500-1000ng of high-quality of DNA is fragmented using the Bioruptor Pico sonicator (Diagenode). The captured fragments are then bisulfite converted followed by PCR amplification. Libraries were analyzed via Qubit (Invitrogen), TapeStation (Agilent), and qPCR (KAPA) prior to sequencing. WGBS libraries were prepared with the NEBNext Ultra II Modules (New England Biolabs, Ipswich, MA) and the NEBNext Methylated Adaptor (New England Biolabs). Approximately 100 ng genomic DNA was sheared to a target size of 200 bp using the Bioruptor Pico (Diagenode, Denville, NJ). The ligated DNA was size-selected for a 200 bp insert target using Sera-Mag Select magnetic beads (Cytiva, Marlborough, MA). Bisulfite conversion was performed with the EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, CA) before carrying out PCR amplification with the NEBNext Q5U polymerase and NEBNext Multiplex Oligos for Illumina (New England Biolabs) to barcode each library. The resulting libraries were normalized and multiplexed prior to sequencing. All sequencing was done at the OHSU Massively Parallel Sequencing Shared Resource using Nextseq 500 sequencer for MethylCapture seq samples and the NovaSeq 6000 for WGBS samples.

### II.v.ii Read Alignment and Analysis

All bisulfite sequencing data was aligned using the ENCODE WGBS standards. Raw sequencing reads were assessed for quality with FastQC v0.11.9, adapter trimmed with TrimGalore v0.6.6, aligned to the hg38 assembly using Bismark v0.23.0 and deduplicated<sup>58,59</sup>. The Illumina TruSeq Methyl Capture EPIC manifest was downloaded from Illumina and liftover was used to convert to hg38 assembly. CpG Island, Alternative Haplotypes, and BisMAP data was downloaded from UCSC genome browser. The ensemble hg38 genome file was used to identify exons, introns, promoters, and transcription start sites.

Read counts were computed using Rsamtools v.2.2.3 and manifest target region read coverage was assessed using bedtools v2.29.2 for all samples. MethylCapture seq manifest regions were labeled as failed if reads were not present within the majority of samples across all enrichment pools, sporadically failed if reads detected within some enrichment pools but not others, and passing regions had

sequencing reads detected in all enrichment pools. Plots were created using ggpubr v0.4.0 and ggplot2 v3.3.3 using R v3.6.2 <sup>60,61</sup>.



### III. Chapter 3: DNA Methylation Analysis of Turner Syndrome BAV

#### III.i Abstract

Turner Syndrome (TS) is a rare cytogenetic disorder caused by the partial or complete loss of a second sex chromosome, which occurs in about 1 in 2,000 female live births. The most common cause of early mortality in TS is due to congenital heart defects. Bicuspid Aortic Valve (BAV) is the most common congenital heart defect in the general population with a prevalence of 0.5-2%. TS patients have the highest burden of BAV with a prevalence around 30% with near complete penetrance of aortic disease. Little is known about why there is such a large increase of BAV in TS. TS is associated with genome wide hypomethylation when compared to karyotypically normal female and male controls. Epigenetic alterations in BAV have been found with changes identified in circulating miRNAs and in DNA methylation profiles of primary aortic tissue. We hypothesize that BAV is associated with DNA methylation alterations in TS.

The purpose of this study is to investigate DNA methylation alterations when comparing 1) BAV to non-BAV in TS and 2) TS BAV to 46,XX non-syndromic BAV. Illumina TruSeq-Methyl Capture EPIC methylation sequencing (Methyl Capture Seq) was performed on whole blood genomic DNA samples from 45,X TS BAV (n = 12), 45,X TS non-BAV (n = 13), and 46,XX non-syndromic BAV (n = 6). DMR detection was performed using Limma + Comb-p adjusted using Surrogate Variable Analysis, yielding 76 DMRs associated with TS BAV and 375 DMRs associated with TS alone. BAV DMRs showed overlap with ENCODE cis Regulatory Elements (cCRE) and directly overlapped genes associated with BAV in human and mice studies, namely *MYRF* and *ATP11A*. Transcription Factor Binding Site (TFBS) enrichment analyses identified known heart development regulators *RXR*, *PBX3*, and *PKNOX1* to be enriched within these DMRs suggesting their dysregulation in TS BAV. TS 46,XX DMRs were mostly hypomethylated and had significant enrichment for HOX genes in early development, consistent with previous studies. These DMRs were also enriched for overlap with ChIP-seq peaks targeting genes known to contribute to BAV, including *NOTCH1* and *MYH11*, in addition to known epigenetic regulators of heart development such as *KDM4A* and *HDAC2* using LOLA and the CODEX database. Finally, these DMRs show TFBS motif enrichment for the highly conserved heart development transcription factor *TBX20*. In summary, DNAm alterations are detected in TS BAV and TS alone with significant enrichment in TFBS known to regulate heart development.

### III.ii Introduction

Turner syndrome (TS) is a rare cytogenetic disorder caused by the partial or complete loss of a second sex chromosome, which occurs in 1 in 2,000 female live births<sup>1</sup>. Girls with TS show a variety of clinical manifestations including short stature, premature ovarian failure, webbed neck, specific cognitive/visual spatial disabilities, hearing loss, thyroid dysfunction, scoliosis, endocrine disorders, autoimmune disorders, and cardiovascular disease. The most common cause of early mortality in TS is due to congenital heart defects, where patients with the most common 45,X karyotype have the highest burden of congenital defects and negative outcomes<sup>6</sup>.

Bicuspid Aortic Valve (BAV) is where the aortic valve consists of two leaflets as opposed to the normal three leaflet configuration of the Tricuspid Aortic Valve (TAV). BAV is the most common congenital heart defect in the general population with a prevalence of 0.5-2%<sup>9</sup>. There is a significant sex bias, where males account for approximately 75% of all BAV cases<sup>11</sup>. The specific negative cardiovascular outcomes of BAV include valve calcification, stenosis, aortic endocarditis, aortic dilation, and aortic aneurysm, collectively known as aortopathy. Approximately 40% of patients with BAV go on to develop some form of aortopathy in their lifetime<sup>11</sup>. TS patients with the 45,X karyotype have the highest burden of BAV with a prevalence around 30% with nearly all TS individuals with BAV developing aortopathy<sup>12</sup>. The high incidence of BAV in TS females with the 45,X karyotype and the high burden of BAV in karyotypically normal 46,XY males suggests the genetics of having one X chromosome predisposes individuals to the development of BAV and BAV associated aortopathy.

Despite the high prevalence in the general population, most of the etiology of BAV is not known. However, a genetic component of BAV has been identified as 10-40% of BAV is familial<sup>33</sup>. Mutations in *NOTCH1*, *GATA5*, *NKX2.5*, and *ROBO4* are known to cause BAV in some families, but the majority of BAV cases are unexplained<sup>34-37</sup>. In the case of BAV in TS, a recent whole exome sequencing study has identified copy number variation of the X chromosome escape gene *TIMP1* and common variants in *TIMP3* to be associated with BAV and aortopathy in TS subjects<sup>20</sup>. Although these genes show a very significant association with BAV, *TIMP1/3* deficiency only explains roughly 20% of the occurrence of BAV in TS. Taken together, these studies have shown that BAV is a complex and heterogeneous condition with multiple genes contributing only a small amount.

DNA methylation (DNAm) alterations associated with BAV have been detected in primary aortic wall tissue and within the aortic valve itself in addition to non-coding RNA expression differences detectable in blood samples of BAV subjects<sup>38-40</sup>. DNAm analysis of TS has identified genome-wide hypomethylation when compared to healthy 46,XX female and 46,XY males<sup>22,23</sup>. Together, these findings suggest a significant role for epigenetic regulation both in TS and BAV which have not been explored. This study attempts to fill this gap by exploring DNAm alterations associated with TS BAV as well as between TS and 46,XX euploid individuals in order to generate hypotheses associating genes or pathways to TS BAV related aortopathy for further study.

## I.i Methods

### I.i.i Samples

All blood samples were collected from the GenTAC consortium and supplied through the BioLINCC biorepository resource<sup>62</sup>. In order to control for known sources of variation which could confound DNAm studies, all samples included were of non-smoking individuals. Smoking status was determined by subject self-reporting at time of GenTAC enrollment. For TS subjects, karyotype information was primarily determined based on clinical information gathered at GenTAC enrollment. A subset of subjects had karyotype information confirmed via molecular karyotyping performed in a previous exome sequencing study<sup>20</sup>. All subjects were over 13 years of age in order to minimize adolescent age effects with both biological groups displaying large overlap in age ranges (Table 3). Enrollment and studies have Internal Review Board approval and all study subjects had informed consent for participation.

Study Groups	Samples (n)	Mean Age (Range)
ns BAV	6	52 (28-67)
ts TAV	13	35 (14-68)
ts BAV	12	42 (16-65)

**TABLE 3: Study Subject Summary. Summary statistics for all study participants with large overlap in age ranges between groups.**

### I.i.ii Study Design

A total of 36 Whole Blood DNA samples from 3 groups, TS BAV, TS TAV, and 46,XX ns BAV were analyzed using Illumina Methyl Capture sequencing. 3 samples did not yield enough reads to be included in downstream analysis following deduplication. Unexpectedly, 2 TS samples showed X chromosome methylation levels comparable with the 46,XX ns BAV samples indicating mosaicism of the X chromosome. The newly developed DAMEfinder allelic methylation analysis method was used to confirm X inactivation within these samples leading to their exclusion leaving a total of 31 samples used for differential methylation analysis.

### I.i.iii Illumina Methyl Capture Sequencing

Genomic DNA samples were submitted for Methyl Capture Sequencing at the KCVI Epigenetics Core using the Illumina TruSeq Methyl Capture EPIC Library Prep Kit (TruSeq-Methyl Capture EPIC, cat # FC-151-1002, Illumina Inc., San Diego, CA) as directed. Briefly, 500-1000ng of high-quality DNA is fragmented using the Bioruptor Pico sonicator (Diagenode). The captured fragments were then bisulfite converted followed by PCR amplification. Libraries were analyzed via Qubit (Invitrogen), TapeStation (Agilent), and qPCR (KAPA) prior to sequencing. All sequencing was done at the OHSU Massively Parallel Sequencing Shared Resource using the NovaSeq 6000.

#### I.i.iv Data Processing and QC

Bisulfite sequencing data was aligned using the ENCODE WGBS standard. Raw sequencing reads were assessed for quality with FastQC v0.11.9, adapter trimmed with TrimGalore v0.6.6, aligned to the hg38 assembly using Bismark v0.23.0 and deduplicated<sup>58,59</sup>. Bismark coverage reports were generated using BismarkMethylationExtractor command and processed in R v3.6.2 using MethylKit v1.12.0<sup>63</sup>. CpG data was filtered for 10X coverage for each sample, CpGs with majority coverage within each study group was used for differential methylation analysis.

Due to the some TS samples showing X chromosome features similar to euploid samples, X inactivation status for all samples was validated using the newly developed allelic methylation analysis tool DAMEfinder v1.2.0<sup>64</sup>. Briefly, MethTuple v1.5.3 was applied to the same aligned data and to detect di-CpG methylation status within the same molecule (read)<sup>65</sup>. These diCpG loci are then filtered for 10X coverage and loci with complete coverage across all available samples were retained. Following the original publication, mean allelic methylation scores from X chromosome gene promoters were extracted to distinguish samples that have bi-allelic methylation as a proxy for X inactivation.

Principal components analysis (PCA) was done using the R stats prcomp function. Principal Components Partial R Squared (PCPR2) analysis is an extension of PCA which allows for the assessment of technical factors across all principal components. Originally conceived for batch effect assessment for metabolomics data, PCPR2 has since been extended to DNA methylation microarray studies<sup>56,66</sup>. Briefly, PCPR2 is performed by using the previously computed PCs and their respective eigenvalues, i.e. how much variance is explained. Continuous and categorical variables are regressed onto each PC and their resulting partial R square is extracted and weighted by the previously found eigenvector. This is repeated across all PCs until the variance threshold is reached, in this case 80% of total variation. PCPR2 was performed using a custom R function following the original publication.

#### I.i.v Differential Methylation Analysis

Surrogate Variable Analysis v3.34.0 from the sva R package was used to adjust the model for known batch effects such as the enrichment pool or sequencing run in addition to unknown sources of variation including cell type heterogeneity<sup>48</sup>. DMRs were detected using a two step approach with differentially methylated CpGs being detected using Limma v3.42.2 adjusted for Age and Surrogate Variables followed by DMR detection using Comb-P v33.1.1 using default parameters<sup>47,49</sup>. The statistical significance threshold was set at  $<.1$  due to the hypothesis generating nature of this study. Significant DMRs were called with a sidak adjusted p value  $<.1$  and no difference in methylation threshold was used due to the phenotype of interest, BAV, occurring early in development which may not lead to a large difference in methylation states between our groups of interest. Due to 46,XX karyotype samples being subject to X inactivation and being incomparable to a single activated X chromosome, the TS v. 46,XX comparison had CpGs on the sex chromosomes excluded from comb-p DMR detection.

Genes overlapping these DMRs were annotated using Genomation v1.18.0 with DMRs being annotated to genes overlapping exon, intron, or promoter regions being deemed genic and all other DMRs

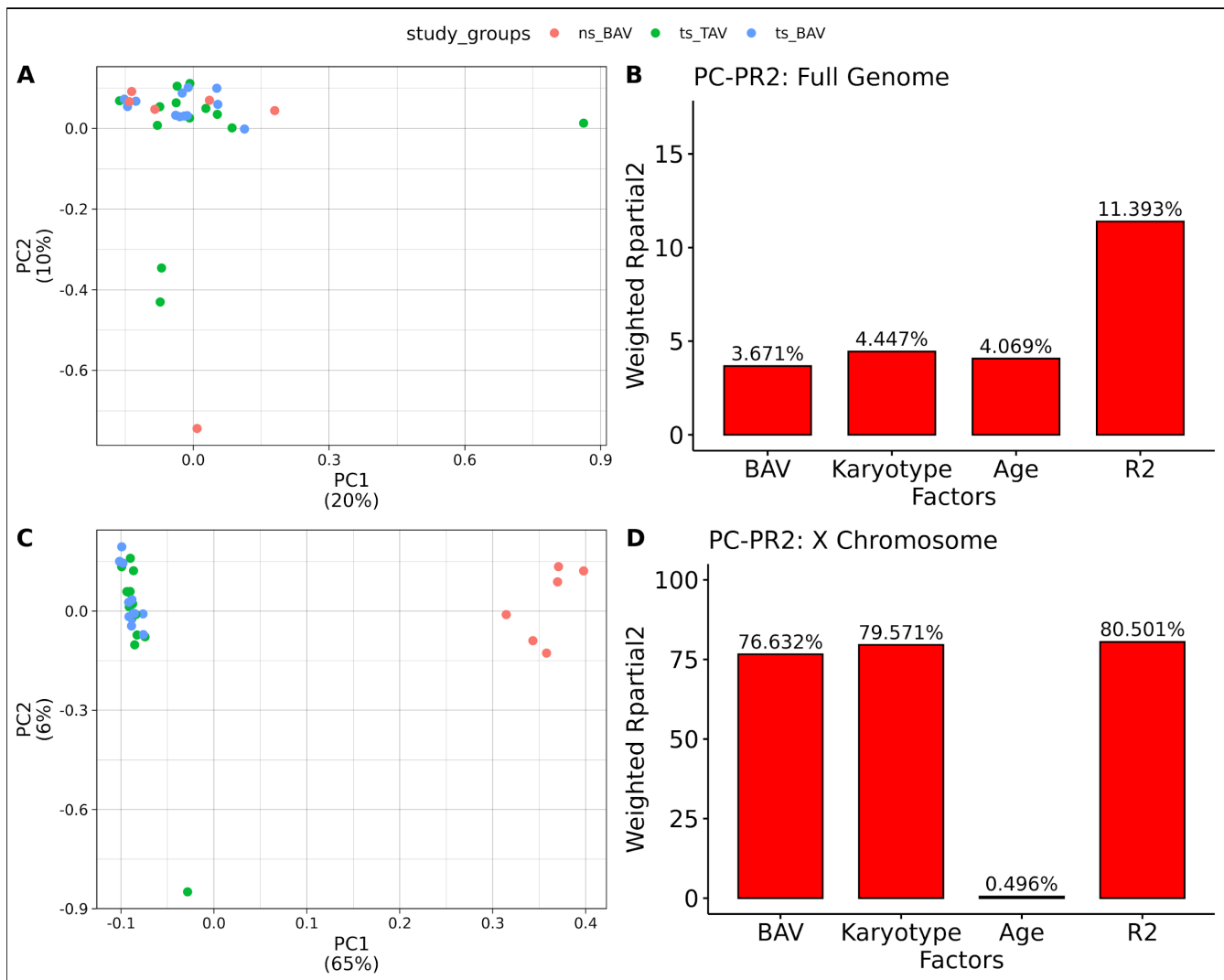
intergenic<sup>67</sup>. GeneHancer 2017 data was downloaded from GeneCards<sup>68</sup>. ENCODE cCRE regulatory regions were downloaded from the SCREEN ENCODE portal<sup>69</sup>. LOLA v1.16.0 was used to analyze enrichment for known genomic loci by comparing DMRs with their appropriate background regions to the available databases with cCRE and genehancer files processed into database collections for LOLA analysis using a custom script<sup>70</sup>. TFBS motif enrichment was performed using HOMER v4.11.1 to analyze TF networks which could be altered by DNAm alterations. TFBS sequence logos were generated by using motifs files produced by HOMER and were visualized with ggseqlogo v0.1<sup>71</sup>. DMRs were analyzed in bulk or subset by hypo/hyper methylation status, with background regions being defined as all tested regions extracted from Comb-P<sup>72</sup>. STRINGdb and ENRICHR were used to assess pathways contributing to the extracted gene lists<sup>52,73</sup>. Reactome Pathway analysis was performed using web based Analysis Tools<sup>74</sup>. Plots were created using ggpubr v0.4.0 and ggplot2 v3.3.3<sup>60,61</sup>.

## I.ii

## I.iii Results

### I.iii.i Data Quality

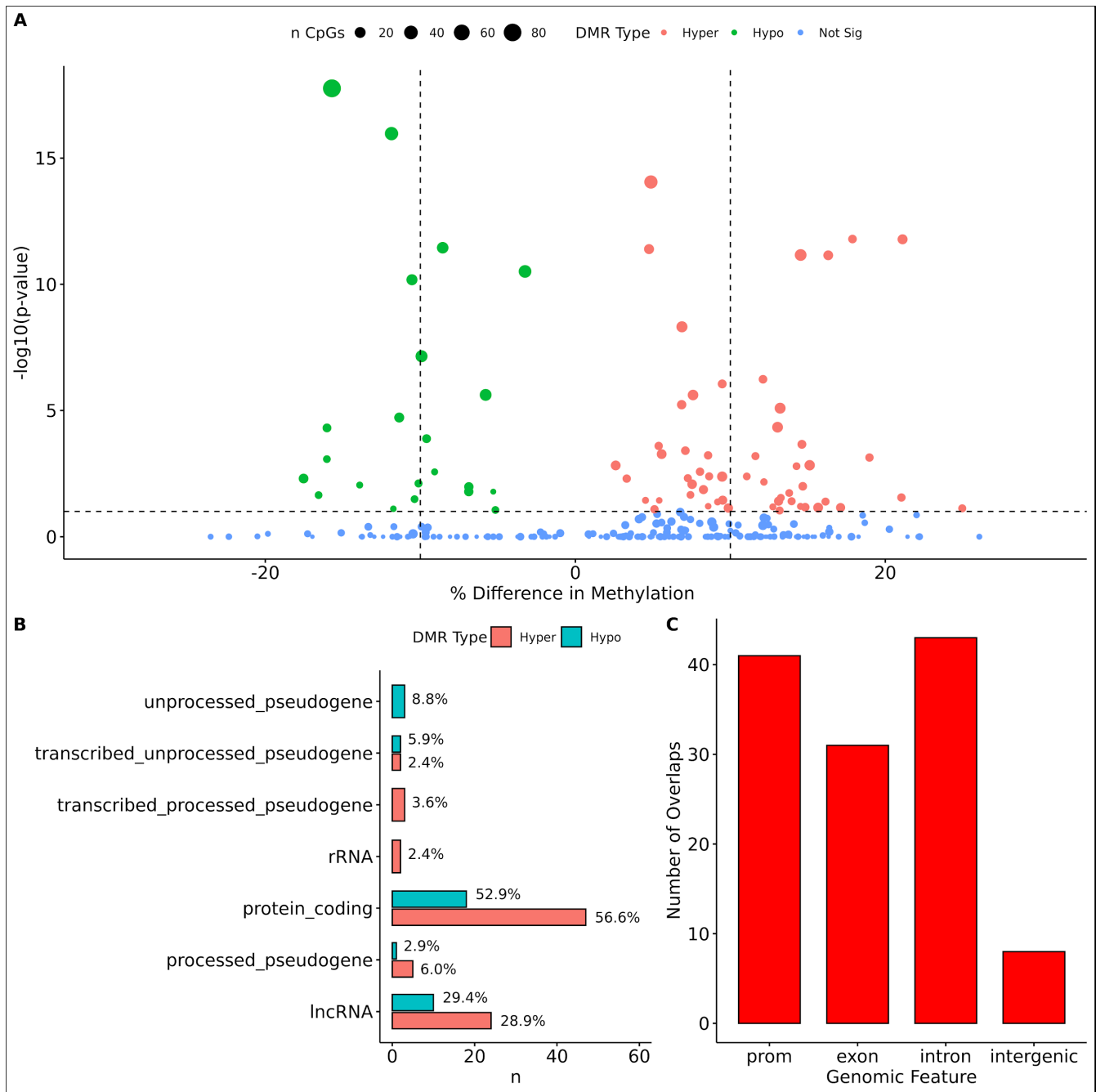
A total of 31 human blood samples collected from the GenTAC consortium were analyzed in this study. All samples showed robust bisulfite conversion with <1% nonCpG methylation for all samples, with a mean alignment rate of 81%. After filtering for CpGs with at least 10X coverage, all samples yielded an average of 3.1M CpGs, with a mean read depth of 30X. Once CpGs were filtered for majority coverage across each study group, there were approximately 2.7M CpGs used for downstream analysis with a mean read coverage of 36X. When performing Principal Component Analysis, samples did not separate by study group (Figure 6A), suggesting the absence of global DNAm differences and also the presence of high variability within each group. Such variability is expected due to the use of a cohort of human blood samples from a multi-site registry which could have differences in DNA extraction and storage. The contribution of both BAV and karyotype was inferred from Principal Component Partial R Squared (PC-PR2) analysis with the main variables of interest explaining roughly 11% of the variation (Figure 6B). PCA of the X chromosome CpGs shows clear separation based on karyotype (Figure 6C). Within the X chromosome, karyotype alone is the major contributor of the variation explaining roughly 80% of the variation (Figure 6D).



**Figure 6: Principal Components Analysis of Methylation data indicates TS samples are distinct from ns BAV on the X chromosome. A. PCA biplot of first 2 principal components for all autosomes. B. PCPR2 analysis of biological variables for autosomes. C. PCA biplot of X chromosome. D. PCPR2 analysis of biological variables for the X chromosome. Note, PCPR2 is sensitive to multicollinearity and overestimates BAV contribution due to unbalanced representation for each karyotype.**

### 1.iii.ii TS BAV Methylation

When comparing TS BAV against TS TAV, a total of 76 significant DMRs ( $q$ -value  $< .1$ ) were detected of which 44 showed a methylation difference  $> 10\%$  (Figure 7A). The average difference in methylation was low ( $\sim 11\%$ ) which might be due to 1) BAV being a defect which occurs very early in development, within 60 days post fertilization<sup>75</sup> and 2) the use of blood for these analyses, rather than heart tissue. The majority of ( $n = 54$ ) DMRs are hypomethylated in TAV which largely overlap with genic regions (promoter, exon, intron) overlapping protein coding genes followed by lncRNAs (Figure 7B,C). Gene Set enrichment analysis, Reactome pathway analysis, or STRINGdb analysis did not yield significant results for genes overlapping DMRs<sup>52,73,74</sup>.



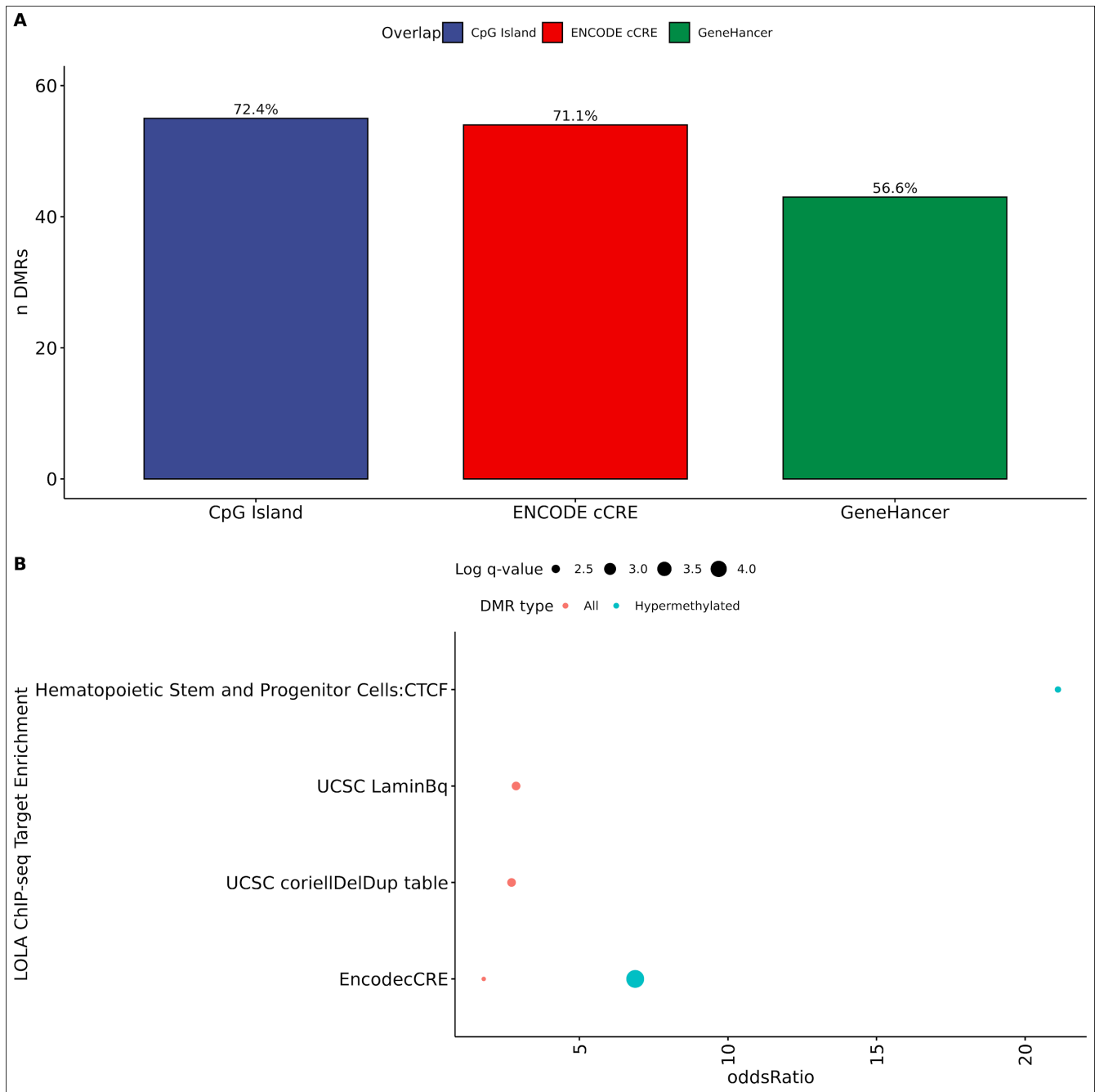
**Figure 7: 76 DMRs were detected in TS BAV mostly near protein coding regions. A. Volcano plot of all regions detected colored by DMR type, hypermethylated, hypomethylated, or not significant with dashed lines at  $-\log_{10}(.1)$  and  $\pm 10\%$  methylation difference. B. Barplot of gene type annotations for all genes overlapping DMRs or the nearest gene when DMRs lack overlap. C. Barplot of genomic features promoters, exon, intron, and intergenic regions overlapping DMRs.**

To assess the possible function of the DMRs, we overlapped them to known regulatory enhancers from GeneHancer, ENCODE cCRE, and CpG Islands. We found that the majority of DMRs overlap cCRE and CpG islands (72.4% and 71.1% respectively) suggesting that these DMRs reside in functionally

relevant regions of the genome (Figure 8A). Using a locus overlap enrichment analysis (LOLA) we then tested for enrichment among all DMRs, hypermethylated DMRs, and hypomethylated DMRs and found that only hypermethylated DMRs displayed significant enrichment for cCRE overlap (Figure 8B). Investigation of hypermethylated cCRE elements identified various genes associated with congenital heart defects (*DUSP22* and *MYOM2*) and changes in cardiovascular function in patients with congenital heart defects (*UTS2*)<sup>76-79</sup>. GeneHancer annotation for these regulatory cCRE was used to identify additional genes associated with these TS BAV DMRs. As before, enrichment was not observed for these gene sets even when subset by hypo/hyper methylation status.

There was special interest for DMRs present on the X chromosome as they could shed light on X chromosome dynamics that could predispose TS individuals to develop BAV and BAV associated aortopathy. Only 2 DMRs were detected on the X chromosome both of which overlapped CpG islands within pseudogenes *ANKRD11P2* and *FTHIP27*. No known regulatory elements were overlapping these X chr DMRs and there were no protein coding genes in the vicinity of these DMRs. Notable findings include *DUSP22*, which shows significant differences in methylation along most of the locus covering 3 separate DMRs with a 7.5% average difference in methylation. *DUSP22* is known to activate JNK signaling in T cells and aged knockout mice show increased autoimmunity implicating immune system response as well as CNV in this gene which has been linked to atrial septal defects suggesting this gene is important to the development and maintenance of<sup>76,80</sup>. Noteworthy genes directly overlapping DMRs include *MYRF* and *ATP11A* which reside in the intronic regions of these genes and overlap cCRE elements which may regulate the expression of these genes (Supplemental Figure 1). Interestingly, *MYRF* is a known regulatory factor of heart development with mutations found in patients with congenital heart defects including BAV<sup>81</sup>. Additionally, *ATP11A* was recently described to cause BAV in knockout mice screens<sup>82</sup>.








**Figure 8: Feature enrichment for genes overlapping TS BAV DMRs. A. Barplot of DMR overlap with CpG islands, ENCODE cCRE, and GeneHancer elements. B. Dotplot of LOLA feature overlap with enrichment for cCREs, CTCF ChIP-seq peaks, Lamin B1 nuclear lamina interactions, and Coriell Cell Line Copy Number Variants (coriellDelDup table).**

To analyze potential transcription factor networks that could be altered by changes to DNAm within these DMRs, HOMER was used to identify enrichment for known transcription factor binding site (TFBS) motifs. It was found that known regulators of heart valve development *PBX3* and *PKNOX1* were enriched with a 15 and 19-fold enrichment respectively which approached significance (qvalues = .1041) (Table 3). *PBX3* and *PKNOX1* binding sites were present in three DMRs and their binding

co-occurred with one another, which potentially suggests that their functions may be altered together. The three DMRs with these binding motifs also overlapped cCRE regulatory elements, suggesting that DNAm alterations could produce functional differences in genes regulated by these elements. To explore this, the nearest genes were extracted in order to investigate which pathways may be altered by changes in TF binding through DNAm alterations. A second Reactome pathway enrichment analysis for genes associated with *PBX3/PKNOX1* motif cCRE's revealed significant enrichment for signaling by hedgehog (*GNAS*) (FDR = 0.03), gene and protein expression by JAK-STAT signaling after interleukin-12 stimulation (*HNRNPF*) (FDR= 0.01), and metabolism of angiotensinogen to angiotensins (*CTS2*) (FDR = 0.05) (Supplemental Table 1). In addition, many well characterized 'late' HOX gene TFBS motifs, known to contribute to limb and heart development, were also enriched within these DMRs that approached statistical significance (Table 3). Specifically, *HOXA9* and *HOXA10* were found to be enriched and are known to regulate heart development through interactions with *NKX2-5* with mutations in this gene known to cause BAV<sup>83</sup>.

All DMRs								
Motif	Name	P-value	qvalue-FDR	n Targets	% Targets	n Background	% Background	Fold Enrichment
	Hoxd11(Homeobox) ChickenMSG-Hoxd11.Flag-ChIP-Seq(GSE86088)	1.00E-05	0.004	12	16.67%	5.9	4.20%	3.97
	c-Myc(bHLH) mES-cMyc-ChIP-Seq(GSE11431)	1.00E-04	0.005	9	12.50%	4	2.85%	4.39
	bHLHE41(bHLH) proB-Bhlhe41-ChIP-Seq(GSE93764)	1.00E-03	0.028	24	33.33%	22.1	15.88%	2.1
	Hoxa13(Homeobox) ChickenMSG-Hoxa13.Flag-ChIP-Seq(GSE86088)	1.00E-03	0.028	10	13.89%	5.1	3.68%	3.77
	HINFP(Zf) K562-HINFP.eGFP-ChIP-Seq(Encode)	1.00E-03	0.039	15	20.83%	12	8.60%	2.42

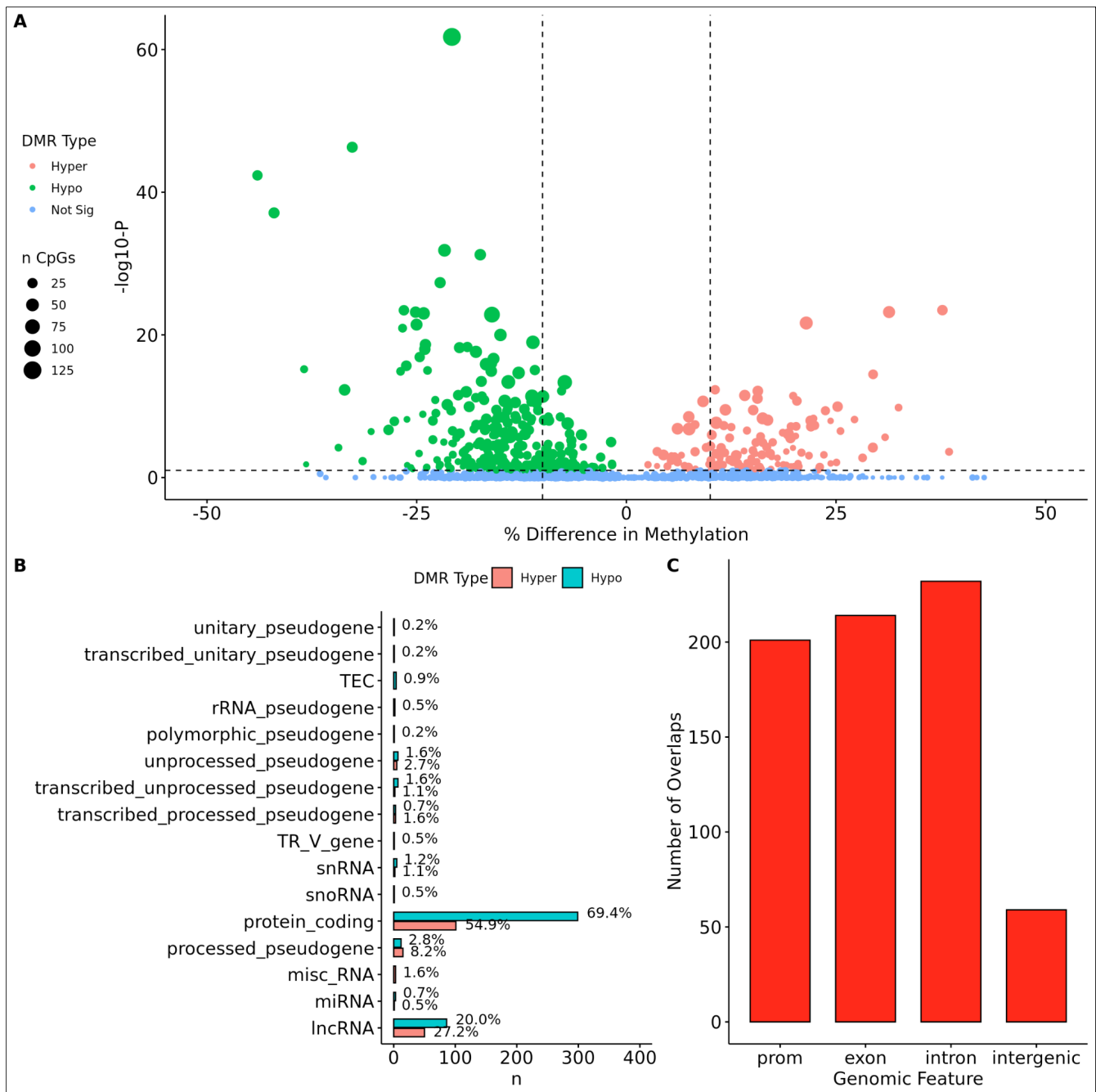
	Max(bHLH) K562-Max-ChIP-Seq(GSE31477)	1.00E-03	0.068	11	15.28%	7.8	5.57%	2.74
	BMAL1(bHLH) Liver-Bmal1-ChIP-Seq(GSE39860)	1.00E-02	0.104	19	26.39%	18.7	13.42%	1.97
	Hoxa10(Homeobox) ChickenMSG-Hoxa10.Flag-ChIP-Seq(GSE86088)	1.00E-02	0.104	4	5.56%	1.8	1.29%	4.31
	HOXB13(Homeobox) ProstateTumor-HOXB13-ChIP-Seq(GSE56288)	1.00E-02	0.104	4	5.56%	1.6	1.18%	4.71
	Hoxd13(Homeobox) ChickenMSG-Hoxd13.Flag-ChIP-Seq(GSE86088)	1.00E-02	0.104	4	5.56%	0.4	0.28%	19.86
	Pbx3(Homeobox) GM12878-PBX3-ChIP-Seq(GSE32465)	1.00E-02	0.104	4	5.56%	0.5	0.35%	15.89
	Pknox1(Homeobox) ES-Prep1-ChIP-Seq(GSE63282)	1.00E-02	0.104	4	5.56%	0	0.00%	Inf

**Table 4: Homer TFBS Motif enrichment results for all DMRs comparing TS BAV vs. TS TAV indicating *PBX3* and *PKNOX1* approach statistical significance (q value < .1).**

#### I.iii.iii TS Methylation Alterations Support Previous Findings

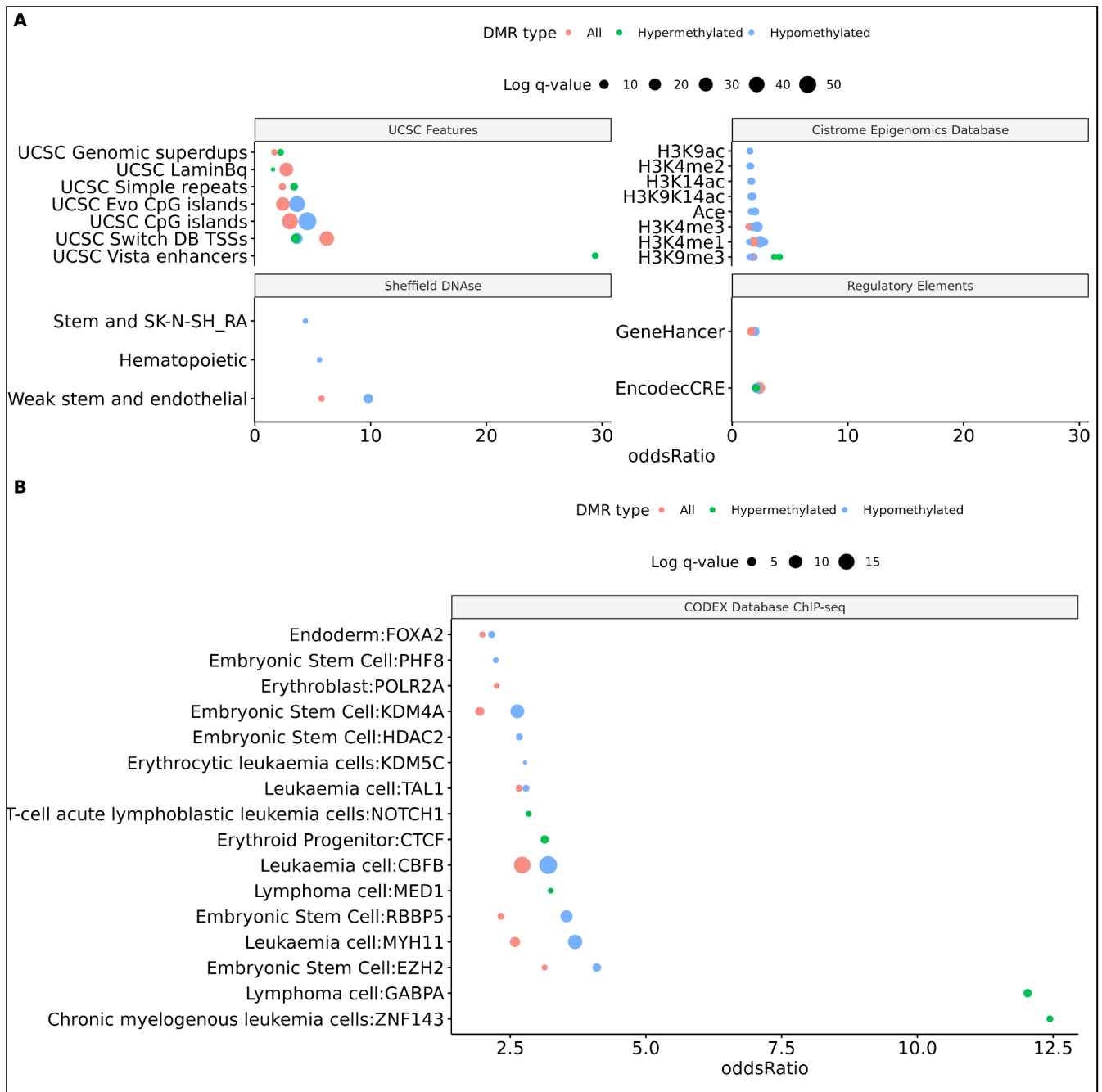
Within the TS v. 46,XX comparison a total of 414 DMRs were detected with an adjusted p value < .1, of which 329 had a methylation difference >10% the majority of which are hypomethylated (Figure 9A). These DMRs largely overlap with genic regions (promoter, exon, intron) largely overlapping protein coding genes followed by lncRNAs (Figure 9C,D). Genes overlapping DMRs were annotated as previously done to be used for Geneset, Reactome pathway, and STRING enrichment analysis. Reactome pathway analysis revealed significant enrichment (FDR = 0.003) for activation of anterior *HOX* genes in hindbrain development during early embryogenesis due to the presence of four *HOX*

genes within DMRs *HOXB3*, *HOXB6*, *HOXA3*, *HOXA4*, and *HOXC4*. Notably, *HOXA3* and *HOXB3* are known to contribute to cardiac development through their gene function<sup>84</sup>.



**Figure 9: 414 DMRs were detected in TS v. 46,XX which are largely hypomethylated and near protein coding regions. A. Volcano plot of all regions detected colored by DMR type, hypermethylated, hypomethylated, or not significant with dashed lines at  $-\log_{10}(.1)$  and  $\pm 10\%$  methylation difference. B. Barplot of DMR distance to Transcription Start Site (TSS) for all significant DMRs. C. Barplot of gene type annotations for all genes overlapping DMRs or the nearest gene when DMRs lack overlap. D. Barplot of genomic features promoters, exon, intron, and intergenic regions overlapping DMRs.**








Similar to the BAV the comparison, cCRE and enhancer regulatory elements were overlapped to DMRs. DMRs were analyzed using LOLA to investigate enrichment for experimentally determined regions of interest which includes UCSC genome browser features such as CpG Islands, ENCODE TFBS via chip-seq datasets, CODEX dataset, cistrome epigenetic features, and DNase hypersensitivity sites. All DMRs subsets were enriched for cCREs and only hypermethylated DMRs did not show enrichment for enhancer elements (Figure 10A). The enrichment for these functional elements suggests that these DMRs may have functional roles at some stages of development. It was found TS DMRs show enrichment for CpG islands and evolutionarily conserved CpG islands identified by Cohen et al. 2011 (Figure 10A)<sup>85</sup>. These DMRs show enrichment for hematopoietic cells and weak stem-epithelial cell DNase hypersensitivity sites derived from Sheffield et al. 2013 reflecting the use of blood DNA samples.






**Figure 10: Significant LOLA genome feature overlap enrichment for TS DMRs (<.1 qvalue). A. Dotplot of LOLA enrichment for UCSC features, Cistrome epigenomics database, Sheffield et al. 2013 DNase database, and Encode cCRE/GeneHancer overlap. B. Dotplot of LOLA enrichment for CODEX database ChIP-seq peak overlap for NOTCH1, MYH11, KDM4A, and HDAC2.**

DMRs were also enriched for *ZNF143*, *GABPA*, *EZH2*, *RBBP5*, *HDAC4*, and *KDM4A* via CODEX, all of which are known epigenetic regulators of gene expression and development. In addition there was enrichment for many known chromatin states such as heterochromatin (H3K9me3) and promoters and primed enhancers (H3K4me3 and H3K4me1). In addition, there was enrichment for TFBS overlap for gene expression regulators including *SIN3A*, *CTCF*, *YY1*, and *POL2* derived from ENCODE database

(Supplemental Figure 2). DMRs show TFBS enrichment for *NOTCH1* and the downstream *NOTCH* pathway gene *MYH11*. Both of these genes are known to contribute to congenital heart defects, including BAV (Figure 10B)<sup>34,86</sup>.

All DMRs								
Motif	Name	P-value	qvalue-FDR	n Targets	% Targets	n Background	% Background	Fold Enrichment
	Pit1+1bp(Homeobox) GCrat-Pit1-ChIP-Seq(GSE58009)	1.00 E-04	0.026	7	1.79%	2.5	0.30%	5.97
	OCT:OCT-short(POU,Homeobox) NPC-OCT6-ChIP-Seq(GSE43916)	1.00 E-03	0.139	11	2.82%	7.8	0.95%	2.97
	Znf263(Zf) K562-Znf263-ChIP-Seq(GSE31477)	1.00 E-03	0.139	106	27.18%	167.2	20.34%	1.34
	NF-E2(bZIP) K562-NFE2-ChIP-Seq(GSE31477)	1.00 E-02	0.158	4	1.03%	1.8	0.22%	4.68
	Tbx20(T-box) Heart-Tbx20-ChIP-Seq(GSE29636)	1.00 E-02	0.158	12	3.08%	10	1.21%	2.55
	Pknox1(Homeobox) ES-Prep1-ChIP-Seq(GSE63282)	1.00 E-02	0.158	11	2.82%	8.9	1.09%	2.59
	Isl1(Homeobox) Neuron-Isl1-ChIP-Seq(GSE31456)	1.00 E-02	0.158	49	12.56%	67.2	8.17%	1.54
Hypomethylated DMRs								
Motif	Name	P-value	qvalue-FDR	n Targets	% Targets	n Background	% Background	Fold Enrichment

	Pit1+1bp(Homeobox) GCrat-Pit1-ChIP-Seq(GSE58009)	1.00 E-06	0.001	6	2.76%	1.7	0.17%	16.24
	ZFX(Zf) mES-Zfx-ChIP-Seq(GSE11431)	1.00 E-04	0.007	68	31.34%	189	19.77%	1.59
	PRDM15(Zf) ESC-Prdm15-ChIP-Seq(GSE73694)	1.00 E-02	0.154	29	13.36%	69.5	7.27%	1.84

**Table 5: HOMER TFBS motif enrichment results for all DMRs and hypomethylated DMRs for TS v. 46,XX indicating significant enrichment for *PIT1* and *ZFX* (q value < .1).**

Homer TFBS motif enrichment was performed and it was found that all DMRs display enrichment for *PIT1* (qvalue = 0.00, a regulator of hormone expression, and *TBX20* approaching significance (qvalue = 0.15), a known regulator of heart development (Table 5)<sup>87,88</sup>.

Hypomethylated DMRs show enrichment for *TBX20* and *ZFX*. Interestingly, *ZFX* is an X escape gene that could contribute to the phenotypes seen in TS due to these individuals only having one X chromosome which has been previously detected to be differentially methylated in TS subjects<sup>23</sup>.

#### I.iv Discussion

TS individuals are at a 60-fold increased risk of BAV compared to the general population. Although BAV is the most common congenital heart defect, there is little understanding of the epigenetic alterations associated with this condition in the general population let alone in TS individuals. Considering BAV is a developmental disorder, exploring DNAm alterations gives greater insight into the genes or pathways that contribute to this condition. This study attempts to fill this gap by directly comparing 45,X TS individuals with BAV against TS individuals who have the normal tricuspid aortic valve.

We have found detectable DNAm differences between BAV and TAV within individuals with TS where most of these DMRs overlap regulatory elements. These DMRs show TFBS motif enrichment which approach significance for *PBX3*, known to contribute to BAV in mice models through interactions with the chromatin remodeling complex MEIS1<sup>89</sup>. The regulator of *PBX-MEIS* interactions, *PKNOX1*, was also found to have motifs the same DMRs suggesting their involvement in a complex multigenic regulation of development<sup>90</sup>.



An unexpected finding was enrichment for cholesterol biosynthesis and regulation within BAV DMRs. Studies of BAV in the euploid population have linked increased cholesterol levels in BAV patients with aortic stenosis and linear correlation of low density lipoprotein levels and ascending aorta diameter<sup>91,92</sup>. BAV patients undergoing statin treatment were observed to have reduced progression of aortopathy following heart surgery, which was not found in the TAV counterparts<sup>93,94</sup>. An interesting connection can be found when we consider that TS individuals are at an increased risk for dyslipidemia which presents at an early age, although cholesterol levels improve following hormone replacement therapy<sup>95,96</sup>. These findings suggest that there might be an unexplored connection between BAV, cholesterol regulation, and aortopathy which may contribute to TS BAV.

We identified DMRs overlapping *MYRF* and *ATP11A* which are both genes directly associated with congenital heart disease including BAV<sup>81,82</sup>. In humans, *MYRF* is a key regulator of myelin development and is required for development of oligodendrocytes and mutations in this gene are associated with a newly identified disorder, cardiac-urogenital syndrome, which is characterized by congenital heart defects including BAV<sup>81,97</sup>. Mutations in this gene are also associated with non-myelin disease and orthologs play important roles in organisms without myelin such as *C. Elegans* which provides additional evidence that *MYRF* contains important roles in development which have not yet been fully explored<sup>98</sup>. *ATP11A* is a ubiquitous expressed phospholipid flippase which could play important roles for cell-cell signaling through the cell membrane<sup>99-102</sup>. The Deciphering Mechanisms of Developmental Disorders (DMDD) project conducted a mouse knockout screen to identify genes which confer embryonic lethality found that *ATP11A* knockout mice had aortic defects indicating this gene is critical for normal heart development<sup>82</sup>. These two genes are interesting candidates for further analysis because they both have been independently associated with non-syndromic BAV and likely act outside of known pathogenic mechanisms of BAV development such as through *NOTCH* signaling<sup>82</sup>. Taken together, these findings support the hypothesis that there are DNAm alterations in genes and pathways relevant to BAV.

There were significantly more DMRs found when comparing TS BAV to 46,XX BAV subjects than the TS BAV to TS TAV comparison which is consistent with a larger impact of X chromosome monosomy on the epigenetic landscape. Over 99% of embryos with 45,X karyotype are not viable during development with most of these fetuses failing in utero due to LSHL, therefore it would not be expected that DNAm alterations on the same scale as monosomy X to be found in the TS BAV subjects who are seen in clinic<sup>8,103</sup>. Similar to previous findings, TS DMRs are hypomethylated across the full genome indicating that loss of a second sex chromosome leads to global changes to DNAm and potentially other epigenetic regulators. Consistent with these previous findings, *HOX* genes, critical for embryogenesis and hindbrain, were found to be hypomethylation in TS.

We have found that TS DMRs show significant overlap with genomic targets for *NOTCH1*, and the downstream *NOTCH* pathway gene *MYH11*. *NOTCH1* mutations are known to cause familial BAV<sup>34</sup>. However, *MYH11* mutations are known to cause familial thoracic aortic aneurysms<sup>104</sup>. Vascular smooth muscle cells (VSMC) derived from induced pluripotent stem cells from BAV subjects implicate

*NOTCH1* and *MYH11* expression in VSMC differentiation in aortopathy<sup>105</sup>. Together, *NOTCH1* and *MYH11* appear to contribute to both BAV development and BAV associated aortopathy. TS DMRs also show TFBS motif enrichment for *TBX20*; copy number variations involving this gene have been identified in BAV subjects with a prevalence ~1%<sup>106</sup>. *TBX20* is an ancient member of the *TBX* family which has been characterized to be essential for heart development and valvulogenesis in multiple animal models and mutations have been found in congenital heart disease probands<sup>88</sup>. Alterations in the function of this transcription factor could lead to heart defects especially in concert with dysregulation of other heart development pathways such as *NOTCH1*. In addition to these findings, the epigenetic regulators *KDM4A* and *HDAC2* were significantly enriched within TS DMRs and these genes have been linked to increased risk of congenital heart disease<sup>107</sup>. Overall, the presence of DMRs within these genes suggest dysregulation of known epigenetic pathways, *TBX20* mediated heart development regulation, and *NOTCH* signaling present in TS which could predispose these individuals to a higher risk of BAV.

These findings suggest that alterations in pathways directed by *TBX20* and *NOTCH1* pathways are altered in TS generally, with BAV individuals also showing DNAm alterations at *PBX3* and *PKNOX1* TFBS. These findings are not powered to distinguish whether these alterations are casual to BAV formation or a downstream effect of the X chromosome monosomy, which leads to BAV. Further studies to validate these findings as well as functional studies in the appropriate model systems are needed to elucidate the mechanisms behind the epigenetic basis of BAV formation in TS. Overall, these DNAm changes are most likely due to haploinsufficiency of X escape genes that lead to alterations in epigenetic programming causing the phenotypes associated with TS. This hypothesis is supported by previous epigenetic studies of other sex chromosome abnormalities (47,XXY or 47,XXX) where TS individuals have the largest change in DNAm compared to euploid controls<sup>23,108,109</sup>. It is important to note that, although X escape genes have been studied for many years, we still lack a complete map of all X escape genes and functional studies of their activity in normal development<sup>110</sup>. Considering the X chromosome has more non-coding RNA than expected and that known genes on the X chromosome have regulatory roles critical to development, there is still much to learn about the function of genes on the X chromosome<sup>110,111</sup>.

Strengths of this study include utilizing a high throughput sequencing approach to analyze DNAm changes and leveraging newly developed allelic methylation analysis techniques to validate biallelic DNAm expression to only analyze TS individuals with a lack of X inactivation within our comparison of interest. Limitations of this study include using whole blood DNA to probe DNAm alterations relevant to the heart in addition to limited study size for each group of interest. The diabetic and lipid status of study participants was not captured at time of enrollment which means we cannot exclude potential confounding due to participants with metabolic disease or dyslipidemia within our study. Diabetes and dyslipidemia could confound this analysis due to potential methylation alterations associated with these diseases being detected contributing as another unknown source of variation reducing statistical power.



## II. Chapter 4: Discussion and Conclusion

Girls with TS are at a 60-fold increased likelihood of having BAV and aortopathy is seen with near complete penetrance in this population <sup>112</sup>. There is strong evidence that TS is characterized by alterations of DNAm. Moreover, BAV is characterized by epigenetic dysregulation. Although BAV is the most common congenital heart defect in humans, there is very little understood about the development of this condition. Taken together, TS represents a sensitive population for BAV and by studying the epigenetic changes associated with this common heart defect we can attempt to gain insight into genes or regulatory pathways that are altered which could be translated into the general population. This project utilized the newly developed Methyl Capture seq bisulfite sequencing method to investigate DNAm alterations associated with BAV in TS in order to generate hypotheses for further study.

The Methyl Capture seq method uses DNA probes to enrich sequencing libraries for functionally relevant regions of the genome, followed by bisulfite conversion to distinguish methylated cytosines from unmethylated cytosines. In Chapter 2, we sought to first characterize Methyl Capture seq performance by utilizing previously generated sequencing data. It was found that the vast majority (92%) of targeted regions perform well. Surprisingly, 4% of targeted regions consistently lacked any sequencing data which largely overlapped regions with alternative haplotype overlap with another 4% of targets lacking data in some enrichment pools. Leveraging a separate whole genome bisulfite sequencing dataset, it was found that regions with alternative haplotypes have reduced sequencing depth compared to their adjacent regions. When we consider that bisulfite sequencing suffers from reduced read mapping efficiency due to reduced library complexity, the additional mapping locations offered by alternative haplotypes could lead to inflated multimapping rates at these locations <sup>53</sup>. Standard bisulfite sequencing data alignment pipelines remove all multi mapping reads in order to obtain confident estimates of DNAm genome wide and could explain the lack of sequencing data at these locations. In the context of Methyl Capture sequencing, these alternative haplotype targets could have well performing DNA probes capturing these sequences within the libraries that are subsequently lost in the read mapping step. These probes and their enriched sequences take up valuable sequencing space for each sample and future iterations of this method could remove these poorly performing DNA probes to target other regions of the genome that are more likely to provide usable sequencing data for downstream methylation analysis.

In Chapter 3, Methyl Capture seq was applied to 31 whole blood samples from 3 groups, TS BAV, TS TAV, and 46,XX ns BAV in order to detect differentially methylated regions across these groups. Two comparisons were made, first comparing TS BAV against TS TAV and TS BAV against 46,XX BAV. When comparing TS BAV to TS TAV a total of 76 significant DMRs were found which had significant transcription factor binding site motif enrichment for known regulators of heart and heart valve development, namely *PBX3*, *PKNOX1* <sup>89</sup>. Additionally there was DMR overlap with interesting genes, *MYRF* and *ATPIIA*, linked to BAV in humans <sup>81,82</sup>. These two genes are not associated with the *NOTCH* signaling pathway which hints that other pathways also contribute to BAV development <sup>82</sup>. Comparing TS BAV to 46,XX BAV, there were 414 DMRs found which were mostly hypomethylated. Genes overlapping these DMRs showed significant pathway enrichment for *HOX* genes in early development which is consistent with previous findings <sup>23</sup>. These DMRs also showed significant

enrichment for transcription factors directly associated with BAV and BAV aortopathy, namely *NOTCH1* and *MYH11*<sup>34,104</sup>. Additionally there was transcription factor motif binding enrichment for a highly conserved regulator of heart development *TBX20* which is known to contribute to BAV<sup>88,106</sup>. Together these results indicate that there are significant changes in DNAm that could lead to alterations to transcription factors or epigenetic networks that contribute to BAV.

Limitations of this analysis include the use of blood as a proxy tissue in order to investigate heart defects. The small number of samples used for each group of interest also limit the conclusions we can draw from these findings. Due to the lack of reported data for all subjects, we cannot rule out additional clinical conditions as confounders to this analysis such as diabetes or dyslipidemia, which is of higher prevalence in TS and are both known to be associated with DNAm alterations<sup>95,113,114</sup>. Future directions for this work including a replication study to confirm these findings within another cohort of TS subjects with much larger sample sizes. Additionally, future differential methylation analysis for these subjects could include investigation of cell type specific effects which were not explored in this work. Future basic biological studies could include characterization of non-canonical NOTCH signaling pathways for genes such as *MYRF* and *ATP11A*. Currently their function in normal heart development is not understood and perturbations by monosomy X may elucidate pathways that could be altered during development. These findings highlight the fact that BAV is a complex disorder with multiple genes and pathways contributing to BAV development and BAV aortopathy. By focusing efforts on understanding which genes or pathways contribute to normal development we may gain understanding on which genetic or epigenetic factors lead to congenital heart defects especially in the context of sex chromosome aneuploidy.

### III. References

---

1. Shankar, R. K. & Backeljauw, P. F. Current best practice in the management of Turner syndrome. *Ther. Adv. Endocrinol. Metab.* **9**, 33–40 (2018).
2. Prakash, S. *et al.* Single Nucleotide Polymorphism Array Genotyping is Equivalent to Metaphase Cytogenetics for Diagnosis of Turner Syndrome. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **16**, (2014).
3. L, M. *et al.* Heart disease in Turner's syndrome. *Helv. Paediatr. Acta* **43**, 25–31 (1988).
4. Gøtzsche, C. O., Krag-Olsen, B., Nielsen, J., Sørensen, K. E. & Kristensen, B. O. Prevalence of cardiovascular malformations and association with karyotypes in Turner's syndrome. *Arch. Dis. Child.* **71**, 433–436 (1994).
5. Silberbach Michael *et al.* Cardiovascular Health in Turner Syndrome: A Scientific Statement From the American Heart Association. *Circ. Genomic Precis. Med.* **11**, e000048 (2018).
6. Barr, M. & Oman-Ganes, L. Turner syndrome morphology and morphometrics: Cardiac hypoplasia as a cause of midgestation death. *Teratology* **66**, 65–72 (2002).
7. Surerus, E., Huggon, I. C. & Allan, L. D. Turner's syndrome in fetal life. *Ultrasound Obstet. Gynecol. Off. J. Int. Soc. Ultrasound Obstet. Gynecol.* **22**, 264–267 (2003).
8. Urbach, A. & Benvenisty, N. Studying Early Lethality of 45,XO (Turner's Syndrome) Embryos Using Human Embryonic Stem Cells. *PLoS ONE* **4**, e4175 (2009).
9. Giusti, B. *et al.* Genetic Bases of Bicuspid Aortic Valve: The Contribution of Traditional and High-Throughput Sequencing Approaches on Research and Diagnosis. *Front. Physiol.* **8**, (2017).
10. Parker Lauren E. & Landstrom Andrew P. Genetic Etiology of Left-Sided Obstructive Heart Lesions: A Story in Development. *J. Am. Heart Assoc.* **10**, e019006 (2021).

11. Liu, T. *et al.* Bicuspid Aortic Valve: An Update in Morphology, Genetics, Biomarker, Complications, Imaging Diagnosis and Treatment. *Front. Physiol.* **9**, (2019).
12. Miller, M. J. *et al.* Echocardiography reveals a high incidence of bicuspid aortic valve in Turner syndrome. *J. Pediatr.* **102**, 47–50 (1983).
13. Miguel-Neto, J., Carvalho, A. B., Marques-de-Faria, A. P., Guerra-Júnior, G. & Maciel-Guerra, A. T. New approach to phenotypic variability and karyotype-phenotype correlation in Turner syndrome. *J. Pediatr. Endocrinol. Metab. JPEM* **29**, 475–479 (2016).
14. Rao, E. *et al.* Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nat. Genet.* **16**, 54–63 (1997).
15. Fukami, M., Seki, A. & Ogata, T. SHOX Haploinsufficiency as a Cause of Syndromic and Nonsyndromic Short Stature. *Mol. Syndromol.* **7**, 3–11 (2016).
16. Gravholt, C. H., Viuff, M. H., Brun, S., Stochholm, K. & Andersen, N. H. Turner syndrome: mechanisms and management. *Nat. Rev. Endocrinol.* **15**, 601–614 (2019).
17. Prandstraller, D. *et al.* Turner’s Syndrome: Cardiologic Profile According to the Different Chromosomal Patterns and Long-Term Clinical Follow-Up of 136 Nonpreselected Patients. *Pediatr. Cardiol.* **20**, 108–112 (1999).
18. Bondy, C. *et al.* Bicuspid aortic valve and aortic coarctation are linked to deletion of the X chromosome short arm in Turner syndrome. *J. Med. Genet.* **50**, 662–665 (2013).
19. Prakash, S. K. *et al.* Autosomal and X chromosome structural variants are associated with congenital heart defects in Turner syndrome: The NHLBI GenTAC registry. *Am. J. Med. Genet. A.* **170**, 3157–3164 (2016).
20. Corbitt, H. *et al.* TIMP3 and TIMP1 are risk genes for bicuspid aortic valve and aortopathy in Turner syndrome. *PLOS Genet.* **14**, e1007692 (2018).

21. Dreger, S. A., Taylor, P. M., Allen, S. P. & Yacoub, M. H. Profile and localization of matrix metalloproteinases (MMPs) and their tissue inhibitors (TIMPs) in human heart valves. *J. Heart Valve Dis.* **11**, 875–880; discussion 880 (2002).
22. Sharma, A. *et al.* DNA methylation signature in peripheral blood reveals distinct characteristics of human X chromosome numerical aberrations. *Clin. Epigenetics* **7**, 76 (2015).
23. Trolle, C. *et al.* Widespread DNA hypomethylation and differential gene expression in Turner syndrome. *Sci. Rep.* **6**, 34220 (2016).
24. Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.* **45**, 1198–1206 (2013).
25. Jadhav, U. *et al.* Extensive Recovery of Embryonic Enhancer and Gene Memory Stored in Hypomethylated Enhancer DNA. *Mol. Cell* **74**, 542-554.e5 (2019).
26. Kashima, K. *et al.* Identification of epigenetic memory candidates associated with gestational age at birth through analysis of methylome and transcriptional data. *Sci. Rep.* **11**, 3381 (2021).
27. Zhu, C. *et al.* DNA hypermethylation of the NOX5 gene in fetal ventricular septal defect. *Exp. Ther. Med.* **2**, 1011–1015 (2011).
28. Wijnands, K. P. *et al.* Genome-wide methylation analysis identifies novel CpG loci for perimembranous ventricular septal defects in human. *Epigenomics* **9**, 241–251 (2017).
29. Gong, J., Sheng, W., Ma, D., Huang, G. & Liu, F. DNA methylation status of TBX20 in patients with tetralogy of Fallot. *BMC Med. Genomics* **12**, 75 (2019).
30. Lim, T. B., Foo, S. Y. R. & Chen, C. K. The Role of Epigenetics in Congenital Heart Disease. *Genes* **12**, 390 (2021).
31. Braverman Alan C. & Roman Mary J. Bicuspid Aortic Valve in Marfan Syndrome. *Circ. Cardiovasc. Imaging* **12**, e008860 (2019).



32. Gillis, E. *et al.* Corrigendum: Candidate Gene Resequencing in a Large Bicuspid Aortic Valve-Associated Thoracic Aortic Aneurysm Cohort: SMAD6 as an Important Contributor. *Front. Physiol.* **8**, (2017).
33. Silberbach, M. Bicuspid Aortic Valve and Thoracic Aortic Aneurysm: Toward a Unified Theory\*\*Editorials published in the Journal of the American College of Cardiology reflect the views of the authors and do not necessarily represent the views of JACC or the American College of Cardiology. *J. Am. Coll. Cardiol.* **53**, 2296–2297 (2009).
34. McKellar, S. H. *et al.* Novel NOTCH1 mutations in patients with bicuspid aortic valve disease and thoracic aortic aneurysms. *J. Thorac. Cardiovasc. Surg.* **134**, 290–296 (2007).
35. Shi, L.-M. *et al.* GATA5 loss-of-function mutations associated with congenital bicuspid aortic valve. *Int. J. Mol. Med.* **33**, 1219–1226 (2014).
36. Qu, X.-K. *et al.* A Novel NKX2.5 Loss-of-Function Mutation Associated With Congenital Bicuspid Aortic Valve. *Am. J. Cardiol.* **114**, 1891–1895 (2014).
37. Gould, R. A. *et al.* ROBO4 variants predispose individuals to bicuspid aortic valve and thoracic aortic aneurysm. *Nat. Genet.* **51**, 42–50 (2019).
38. Pan, S. *et al.* DNA methylome analysis reveals distinct epigenetic patterns of ascending aortic dissection and bicuspid aortic valve. *Cardiovasc. Res.* **113**, 692–704 (2017).
39. Björck, H. M. *et al.* Altered DNA methylation indicates an oscillatory flow mediated epithelial-to-mesenchymal transition signature in ascending aorta of patients with bicuspid aortic valve. *Sci. Rep.* **8**, (2018).
40. Pulignani, S., Borghini, A. & Andreassi, M. G. microRNAs in bicuspid aortic valve associated aortopathy: Recent advances and future perspectives. *J. Cardiol.* **74**, 297–303 (2019).
41. Fernández, B. *et al.* Bicuspid Aortic Valve in 2 Model Species and Review of the Literature. *Vet.*

*Pathol.* **57**, 321–331 (2020).

42. Cao, J. *et al.* The role of DNA methylation in syndromic and non-syndromic congenital heart disease. *Clin. Epigenetics* **13**, 93 (2021).
43. TruSeq Methyl Capture EPIC Library Prep Kit. 8.
44. Kacmarczyk, T. J. *et al.* “Same difference”: comprehensive evaluation of four DNA methylation measurement platforms. *Epigenetics Chromatin* **11**, (2018).
45. Heiss, J. A. *et al.* Battle of epigenetic proportions: comparing Illumina’s EPIC methylation microarrays and TruSeq targeted bisulfite sequencing. *Epigenetics* **15**, 174–182 (2020).
46. Lin, N. *et al.* Genome-wide DNA methylation profiling in human breast tissue by illumina TruSeq methyl capture EPIC sequencing and infinium methylationEPIC beadchip microarray. *Epigenetics* 1–16 (2020) doi:10.1080/15592294.2020.1827703.
47. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
48. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genet.* **3**, e161 (2007).
49. Pedersen, B. S., Schwartz, D. A., Yang, I. V. & Kechris, K. J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* **28**, 2986–2988 (2012).
50. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
51. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
52. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage,

- supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
53. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
  54. Grehl, C., Wagner, M., Lemnian, I., Glaser, B. & Grosse, I. Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants. *Front. Plant Sci.* **11**, 176 (2020).
  55. Price, E. M. & Robinson, W. P. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Front. Genet.* **9**, 83 (2018).
  56. Perrier, F. *et al.* Identifying and correcting epigenetics measurements for systematic sources of variation. *Clin. Epigenetics* **10**, 38 (2018).
  57. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
  58. Andrews, S. *FASTQC. A quality control tool for high throughput sequence data.* (2010).
  59. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
  60. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots.* (2020).
  61. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* (Springer-Verlag New York, 2016).
  62. Kroner, B. L. *et al.* The National Registry of Genetically Triggered Thoracic Aortic Aneurysms and Cardiovascular Conditions (GenTAC): Results from Phase I and Scientific Opportunities in Phase II. *Am. Heart J.* **162**, 627-632.e1 (2011).
  63. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
  64. Orjuela, S., Machlab, D., Menigatti, M., Marra, G. & Robinson, M. D. DAMEfinder: a method to

- detect differential allele-specific methylation. *Epigenetics Chromatin* **13**, 25 (2020).
65. Hickey, P. *PeteHaitch/methtuple*. (2020).
  66. Fages, A. *et al.* Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method. *Metabolomics* **10**, 1074–1083 (2014).
  67. Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E. & Schübeler, D. genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**, 1127–1129 (2015).
  68. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database J. Biol. Databases Curation* **2017**, (2017).
  69. The ENCODE Project Consortium *et al.* Perspectives on ENCODE. *Nature* **583**, 693–698 (2020).
  70. Sheffield, N. C. & Bock, C. LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* (2016).
  71. Wagih, O. *ggseqlogo: A 'ggplot2' Extension for Drawing Publication-Ready Sequence Logos*. (2017).
  72. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
  73. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
  74. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* gkz1031 (2019) doi:10.1093/nar/gkz1031.
  75. Schoenwolf, G. C., Larsen, W. J. (William J., Schoenwolf, G. C. & Larsen, W. J. (William J. *Larsen's human embryology*. (Philadelphia : Churchill Livingstone/Elsevier, 2009).
  76. Thorsson, T. *et al.* Chromosomal Imbalances in Patients with Congenital Cardiac Defects: A

Meta-analysis Reveals Novel Potential Critical Regions Involved in Heart Development.

*Congenit. Heart Dis.* **10**, 193–208 (2015).

77. Grunert, M. *et al.* Rare and private variations in neural crest, apoptosis and sarcomere genes define the polygenic background of isolated Tetralogy of Fallot. *Hum. Mol. Genet.* **23**, 3115–3128 (2014).
78. Auxerre-Plantié, E. *et al.* Identification of MYOM2 as a candidate gene in hypertrophic cardiomyopathy and Tetralogy of Fallot, and its functional evaluation in the Drosophila heart. *Dis. Model. Mech.* **13**, (2020).
79. Simpson, C. M., Penny, D. J. & Stocker, C. F. Urotensin II is raised in children with congenital heart disease. *Heart* **92**, 983–984 (2006).
80. Li, J.-P. *et al.* The phosphatase JKAP/DUSP22 inhibits T-cell receptor signalling and autoimmunity by inactivating Lck. *Nat. Commun.* **5**, 3618 (2014).
81. Rossetti, L. Z. *et al.* Review of the phenotypic spectrum associated with haploinsufficiency of MYRF. *Am. J. Med. Genet. A.* [ajmg.a.61182](https://doi.org/10.1002/ajmg.a.61182) (2019) doi:10.1002/ajmg.a.61182.
82. Szumska, D., Wilson, R., Weninger, W. & Mohun, T. Deciphering the Mechanisms of Developmental Heart Disease: Research from Embryonic Knockout Mice. in *Fetal Therapy* (eds. Kilby, M. D., Johnson, A. & Oepkes, D.) 133–145 (Cambridge University Press, 2019). doi:10.1017/9781108564434.015.
83. Behrens, A. N. *et al.* Nkx2-5 mediates differential cardiac differentiation through interaction with Hoxa10. *Stem Cells Dev.* **22**, 2211–2220 (2013).
84. Roux, M. & Zaffran, S. Hox Genes in Cardiovascular Development and Diseases. *J. Dev. Biol.* **4**, (2016).
85. Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG Islands Are Maintained by

- Heterogeneous Evolutionary Regimes Involving Minimal Selection. *Cell* **145**, 773–786 (2011).
86. Kerstjens-Frederikse, W. S. *et al.* Cardiovascular malformations caused by NOTCH1 mutations do not keep left: data on 428 probands with left-sided CHD and their families. *Genet. Med.* **18**, 914–923 (2016).
  87. Pfaffle, R., Blankenstein, O., Wüller, S. & Kentrup, H. Combined pituitary hormone deficiency: role of Pit-1 and Prop-1. *Acta Paediatr.* **88**, 33–41 (1999).
  88. Kirk, E. P. *et al.* Mutations in Cardiac T-Box Factor Gene TBX20 Are Associated with Diverse Cardiac Pathologies, Including Defects of Septation and Valvulogenesis and Cardiomyopathy. *Am. J. Hum. Genet.* **81**, 280–291 (2007).
  89. Stankunas, K. *et al.* Pbx/Meis Deficiencies Demonstrate Multigenetic Origins of Congenital Heart Disease. *Circ. Res.* **103**, 702–709 (2008).
  90. Schulte, D. & Geerts, D. MEIS transcription factors in development and disease. *Development* **146**, dev174706 (2019).
  91. Endo, M. *et al.* Differing relationship between hypercholesterolemia and a bicuspid aortic valve according to the presence of aortic valve stenosis or aortic valve regurgitation. *Gen. Thorac. Cardiovasc. Surg.* **63**, 502–506 (2015).
  92. Alegret, J. M., Masana, L., Martinez-Micaelo, N., Heras, M. & Beltrán-Debón, R. LDL cholesterol and apolipoprotein B are associated with ascending aorta dilatation in bicuspid aortic valve patients. *QJM* **108**, 795–801 (2015).
  93. Taylor, A. P. *et al.* Statin Use and Aneurysm Risk in Patients With Bicuspid Aortic Valve Disease. *Clin. Cardiol.* **39**, 41–47 (2016).
  94. Sequeira Gross, T. *et al.* Does statin therapy impact the proximal aortopathy in aortic valve disease? *QJM Int. J. Med.* **111**, 623–628 (2018).

95. Ross, J. L. *et al.* Lipid abnormalities in Turner syndrome. *J. Pediatr.* **126**, 242–245 (1995).
96. Mavinkurve, M. & O’Gorman, C. S. Cardiometabolic and vascular risks in young and adolescent girls with Turner syndrome. *BBA Clin.* **3**, 304–309 (2015).
97. Bujalka, H. *et al.* MYRF Is a Membrane-Associated Transcription Factor That Autoproteolytically Cleaves to Directly Activate Myelin Genes. *PLOS Biol.* **11**, e1001625 (2013).
98. An, H. *et al.* Functional mechanism and pathogenic potential of MYRF ICA domain mutations implicated in birth defects. *Sci. Rep.* **10**, 814 (2020).
99. Takatsu, H. *et al.* Phospholipid Flippase Activities and Substrate Specificities of Human Type IV P-type ATPases Localized to the Plasma Membrane \*. *J. Biol. Chem.* **289**, 33543–33556 (2014).
100. Segawa, K., Kurata, S. & Nagata, S. Human Type IV P-type ATPases That Work as Plasma Membrane Phospholipid Flippases and Their Regulation by Caspase and Calcium. *J. Biol. Chem.* **291**, 762–772 (2016).
101. Miyano, R., Matsumoto, T., Takatsu, H., Nakayama, K. & Shin, H.-W. Alteration of transbilayer phospholipid compositions is involved in cell adhesion, cell spreading, and focal adhesion formation. *FEBS Lett.* **590**, 2138–2145 (2016).
102. Hawkey-Noble, A., Umali, J., Fowler, G. & French, C. R. Expression of three P4-phospholipid flippases—*atp11a*, *atp11b*, and *atp11c* in zebrafish (*Danio rerio*). *Gene Expr. Patterns* **36**, 119115 (2020).
103. Mortensen, K. H., Andersen, N. H. & Gravholt, C. H. Cardiovascular phenotype in Turner syndrome--integrating cardiology, genetics, and endocrinology. *Endocr. Rev.* **33**, 677–714 (2012).
104. Takeda, N. *et al.* A deleterious MYH11 mutation causing familial thoracic aortic dissection. *Hum. Genome Var.* **2**, 15028 (2015).
105. Harrison, O. J. *et al.* Defective NOTCH signalling drives smooth muscle cell death and

- differentiation in bicuspid aortic valve aortopathy. *Eur. J. Cardiothorac. Surg.* **56**, 117–125 (2019).
106. MIBAVA Leducq Consortium *et al.* Copy number variation analysis in bicuspid aortic valve-related aortopathy identifies TBX20 as a contributing gene. *Eur. J. Hum. Genet.* **27**, 1033–1043 (2019).
107. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).
108. Zhang, X. *et al.* Integrated functional genomic analyses of Klinefelter and Turner syndromes reveal global network effects of altered X chromosome dosage. *Proc. Natl. Acad. Sci.* **117**, 4864–4873 (2020).
109. Skakkebaek, A. *et al.* DNA hypermethylation and differential gene expression associated with Klinefelter syndrome. *Sci. Rep.* **8**, (2018).
110. Di Palo, A. *et al.* What microRNAs could tell us about the human X chromosome. *Cell. Mol. Life Sci.* **77**, 4069–4080 (2020).
111. Guo, X., Su, B., Zhou, Z. & Sha, J. Rapid evolution of mammalian X-linked testis microRNAs. *BMC Genomics* **10**, 1–8 (2009).
112. Corbitt, H., Gutierrez, J., Silberbach, M. & Maslen, C. L. The genetic basis of Turner syndrome aortopathy. *Am. J. Med. Genet. C Semin. Med. Genet.* **181**, 101–109 (2019).
113. Ahmed, S. A. H., Ansari, S. A., Mensah-Brown, E. P. K. & Emerald, B. S. The role of DNA methylation in the pathogenesis of type 2 diabetes mellitus. *Clin. Epigenetics* **12**, 104 (2020).
114. The role of DNA methylation in dyslipidaemia: A systematic review. *Prog. Lipid Res.* **64**, 178–191 (2016).



## IV. Supplemental Tables and Figures

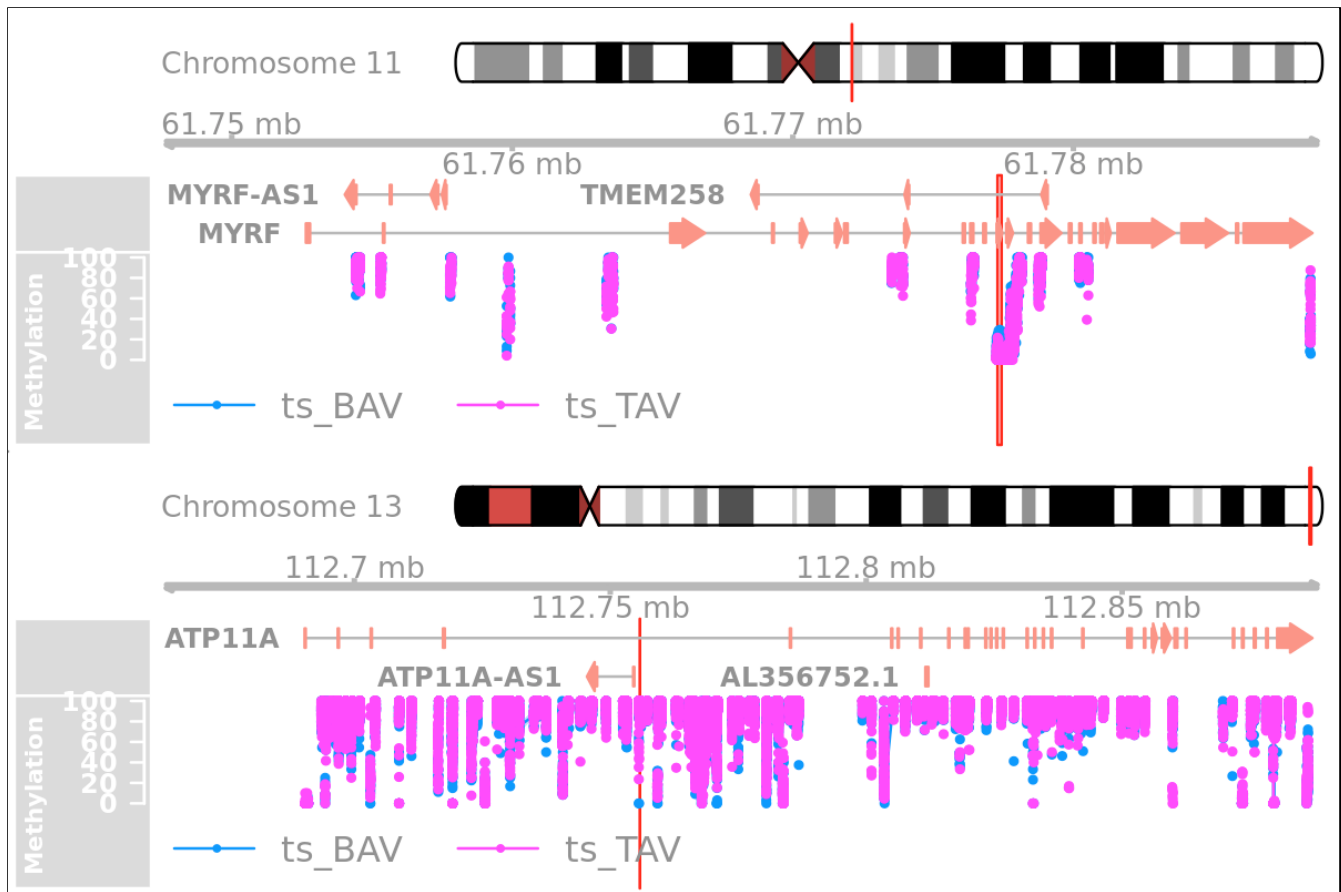
Pathway identifier	Pathway name	#Entities found	#Entities total	Entities ratio	Entities pValue	Entities FDR	#Reactions found	#Reactions total	Reactions ratio	Submitted entities found
<b>R-HSA-3 92851</b>	Prostacyclin signalling through prostacyclin receptor	2	23	0.001583 803884	1.37E-04	0.006559 635605	3	4	2.98E-04	GNAS
<b>R-HSA-1 64378</b>	PKA activation in glucagon signalling	2	23	0.001583 803884	1.37E-04	0.006559 635605	1	2	1.49E-04	GNAS
<b>R-HSA-4 20092</b>	Glucagon-type ligand receptors	2	35	0.002410 136345	3.15E-04	0.009850 621641	1	8	5.96E-04	GNAS
<b>R-HSA-1 63359</b>	Glucagon signaling in metabolic regulation	2	40	0.002754 441537	4.10E-04	0.009850 621641	4	6	4.47E-04	GNAS
<b>R-HSA-3 81676</b>	Glucagon-like Peptide-1 (GLP1) regulates insulin secretion	2	49	0.003374 190883	6.14E-04	0.011043 47825	5	11	8.19E-04	GNAS
<b>R-HSA-4 32040</b>	Vasopressin regulates renal water homeostasis via Aquaporins	2	52	0.003580 773998	6.90E-04	0.011043 47825	5	15	0.001117 235215	GNAS
<b>R-HSA-4 18597</b>	G alpha (z) signalling events	2	62	0.004269 384382	9.77E-04	0.012703 15558	1	13	9.68E-04	GNAS
<b>R-HSA-4 45717</b>	Aquaporin-mediated transport	2	68	0.004682 550613	0.001172 535829	0.013485 18359	5	25	0.001862 058692	GNAS
<b>R-HSA-8 950505</b>	Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation	2	73	0.005026 855805	0.001348 518359	0.013485 18359	1	36	0.002681 364517	HNRNPF
<b>R-HSA-9 020591</b>	Interleukin-12 signaling	2	84	0.005784 327228	0.001777 444264	0.015996 99838	1	56	0.004171 01147	HNRNPF
<b>R-HSA-4 47115</b>	Interleukin-12 family signaling	2	96	0.006610 659689	0.002310 078593	0.018480 62875	1	114	0.008490 987636	HNRNPF
<b>R-HSA-3 73080</b>	Class B/2 (Secretin family receptors)	2	99	0.006817 242804	0.002453 6714	0.019629 3712	1	20	0.001489 646954	GNAS
<b>R-HSA-4 22356</b>	Regulation of insulin secretion	2	106	0.007299 270073	0.002804 79915	0.019633 59405	5	34	0.002532 399821	GNAS
<b>R-HSA-4 18346</b>	Platelet homeostasis	2	117	0.008056 741496	0.003401 638102	0.020409 82861	3	30	0.002234 470431	GNAS
<b>R-HSA-5 610787</b>	Hedgehog 'off' state	2	124	0.008538 768765	0.003809 814308	0.022858 88585	1	32	0.002383 435126	GNAS
<b>R-HSA-1 63685</b>	Integration of energy metabolism	2	145	0.009984 850572	0.005164 504784	0.030870 60681	8	62	0.004617 905556	GNAS

<b>R-HSA-9 660821</b>	ADORA2B mediated anti-inflammatory cytokines production	2	159	0.010948 90511	0.006174 121361	0.030870 60681	5	12	8.94E-04	GNAS
<b>R-HSA-5 358351</b>	Signaling by Hedgehog	2	168	0.011568 65446	0.006867 280767	0.034336 40384	1	82	0.006107 55251	GNAS
<b>R-HSA-4 18555</b>	G alpha (s) signalling events	2	172	0.011844 09861	0.007186 30277	0.035931 51385	10	18	0.001340 682258	GNAS
<b>R-HSA-5 663205</b>	Infectious disease	4	1343	0.092480 3746	0.014175 30853	0.056701 23412	29	750	0.055861 76076	FXYD4;GNAS;NELFCD
<b>R-HSA-1 67242</b>	Abortive elongation of HIV-1 transcript in the absence of Tat	1	27	0.001859 248037	0.020262 66076	0.056803 34617	2	2	1.49E-04	NELFCD
<b>R-HSA-2 022377</b>	Metabolism of Angiotensinogen to Angiotensins	1	27	0.001859 248037	0.020262 66076	0.056803 34617	1	20	0.001489 646954	CTSZ
<b>R-HSA-9 664433</b>	Leishmania parasite growth and survival	2	297	0.020451 72841	0.020348 82726	0.056803 34617	5	40	0.002979 293907	GNAS
<b>R-HSA-9 662851</b>	Anti-inflammatory response favouring Leishmania parasite infection	2	297	0.020451 72841	0.020348 82726	0.056803 34617	5	40	0.002979 293907	GNAS
<b>R-HSA-6 803529</b>	FGFR2 alternative splicing	1	28	0.001928 109076	0.021005 90974	0.056803 34617	3	4	2.98E-04	HNRNPF
<b>R-HSA-1 98933</b>	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	2	316	0.021760 08814	0.022855 63256	0.056803 34617	1	44	0.003277 223298	SIGLECL1;CD33
<b>R-HSA-1 67158</b>	Formation of the HIV-1 Early Elongation Complex	1	37	0.002547 858422	0.027672 12078	0.056803 34617	2	5	3.72E-04	NELFCD
<b>R-HSA-1 13418</b>	Formation of the Early Elongation Complex	1	37	0.002547 858422	0.027672 12078	0.056803 34617	1	3	2.23E-04	NELFCD
<b>R-HSA-5 694530</b>	Cargo concentration in the ER	1	37	0.002547 858422	0.027672 12078	0.056803 34617	2	12	8.94E-04	CTSZ
<b>R-HSA-1 67243</b>	Tat-mediated HIV elongation arrest and recovery	1	39	0.002685 580499	0.029147 88548	0.056803 34617	3	3	2.23E-04	NELFCD
<b>R-HSA-1 67238</b>	Pausing and recovery of Tat-mediated HIV elongation	1	39	0.002685 580499	0.029147 88548	0.056803 34617	2	2	1.49E-04	NELFCD
<b>R-HSA-1 67287</b>	HIV elongation arrest and recovery	1	40	0.002754 441537	0.029885 00397	0.056803 34617	3	3	2.23E-04	NELFCD
<b>R-HSA-1 67290</b>	Pausing and recovery of HIV elongation	1	40	0.002754 441537	0.029885 00397	0.056803 34617	2	2	1.49E-04	NELFCD
<b>R-HSA-3 82551</b>	Transport of small molecules	3	958	0.065968 87481	0.031728 39373	0.056803 34617	7	441	0.032846 71533	FXYD4;GNAS

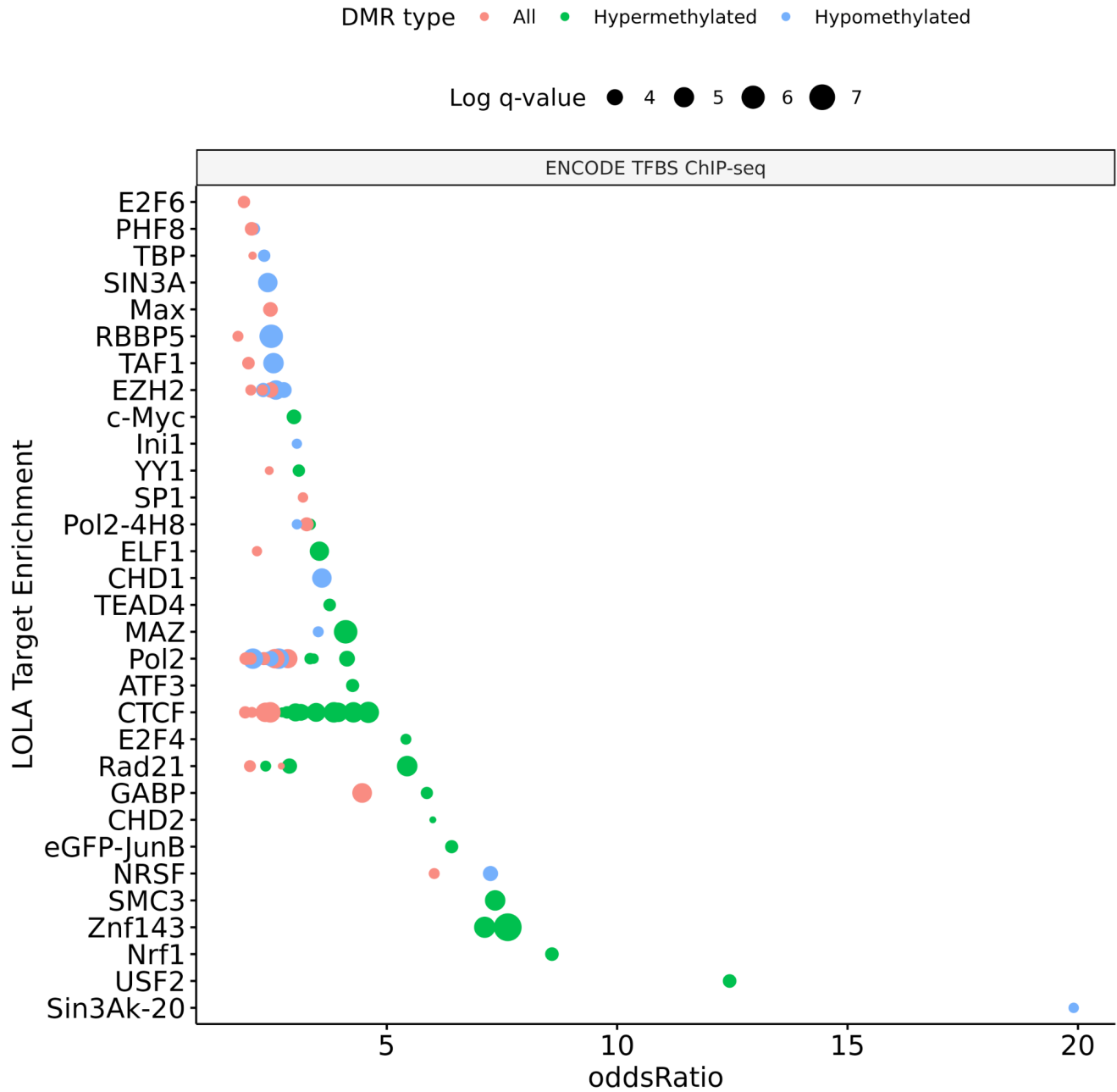
<b>R-HSA-4 32720</b>	Lysosome Vesicle Biogenesis	1	43	0.002961 024652	0.032093 30784	0.056803 34617	4	8	5.96E-04	CTSZ
<b>R-HSA-9 658195</b>	Leishmania infection	2	406	0.027957 5816	0.036352 50679	0.056803 34617	5	95	0.007075 82303	GNAS
<b>R-HSA-1 67200</b>	Formation of HIV-1 elongation complex containing HIV-1 Tat	1	49	0.003374 190883	0.036496 2117	0.056803 34617	4	5	3.72E-04	NELFCD
<b>R-HSA-1 68256</b>	Immune System	5	2681	0.184616 444	0.036907 60958	0.056803 34617	6	1621	0.120735 8856	SIGLECL 1;HNRN PF;CTS Z;CD33
<b>R-HSA-1 67152</b>	Formation of HIV elongation complex in the absence of HIV Tat	1	50	0.003443 051921	0.037228 25627	0.056803 34617	2	2	1.49E-04	NELFCD
<b>R-HSA-1 67246</b>	Tat-mediated elongation of the HIV-1 transcript	1	52	0.003580 773998	0.038690 82872	0.056803 34617	7	8	5.96E-04	NELFCD
<b>R-HSA-1 67169</b>	HIV Transcription Elongation	1	52	0.003580 773998	0.038690 82872	0.056803 34617	11	15	0.001117 235215	NELFCD
<b>R-HSA-4 18594</b>	G alpha (i) signalling events	2	421	0.028990 49718	0.038847 01647	0.056803 34617	1	74	0.005511 693729	GNAS
<b>R-HSA-1 12382</b>	Formation of RNA Pol II elongation complex	1	63	0.004338 245421	0.046698 93473	0.056803 34617	2	2	1.49E-04	NELFCD
<b>R-HSA-5 578775</b>	Ion homeostasis	1	64	0.004407 106459	0.047423 9286	0.056803 34617	2	16	0.001191 717563	FXYD4
<b>R-HSA-7 5955</b>	RNA Polymerase II Transcription Elongation	1	66	0.004544 828536	0.048872 41281	0.056803 34617	6	8	5.96E-04	NELFCD
<b>R-HSA-6 798695</b>	Neutrophil degranulation	2	480	0.033053 29844	0.049283 55073	0.056803 34617	4	10	7.45E-04	CTSZ;C D33
<b>R-HSA-9 36837</b>	Ion transport by P-type ATPases	1	71	0.004889 133728	0.052484 86739	0.056803 34617	2	15	0.001117 235215	FXYD4
<b>R-HSA-2 04005</b>	COPII-mediated vesicle transport	1	77	0.005302 299959	0.056803 34617	0.056803 34617	10	16	0.001191 717563	CTSZ
<b>R-HSA-1 99992</b>	trans-Golgi Network Vesicle Budding	1	80	0.005508 883074	0.058955 86872	0.058955 86872	4	19	0.001415 164606	CTSZ
<b>R-HSA-1 67172</b>	Transcription of the HIV genome	1	81	0.005577 744112	0.059672 38322	0.059672 38322	23	47	0.003500 670341	NELFCD
<b>R-HSA-9 679191</b>	Potential therapeutics for SARS	1	84	0.005784 327228	0.061818 95198	0.061818 95198	1	32	0.002383 435126	FXYD4
<b>R-HSA-6 74695</b>	RNA Polymerase II Pre-transcription Events	1	88	0.006059 771381	0.064674 11414	0.064674 11414	8	17	0.001266 199911	NELFCD
<b>R-HSA-5 654738</b>	Signaling by FGFR2	1	88	0.006059 771381	0.064674 11414	0.064674 11414	3	46	0.003426 187993	HNRNPF

<b>R-HSA-6 796648</b>	TP53 Regulates Transcription of DNA Repair Genes	1	89	0.006128 63242	0.065386 66932	0.065386 66932	3	17	0.001266 199911	NELFCD
<b>R-HSA-5 00792</b>	GPCR ligand binding	2	602	0.041454 34513	0.073715 66176	0.073715 66176	1	185	0.013779 23432	GNAS
<b>R-HSA-1 90236</b>	Signaling by FGFR	1	107	0.007368 131111	0.078128 56681	0.078128 56681	3	142	0.010576 49337	HNRNPF
<b>R-HSA-4 49147</b>	Signaling by Interleukins	2	643	0.044277 64771	0.082689 24288	0.082689 24288	1	493	0.036719 79741	HNRNPF
<b>R-HSA-1 643685</b>	Disease	4	2360	0.162512 0507	0.088624 84843	0.088624 84843	29	1591	0.118501 4152	FXJD4;G NAS;NEL FCD
<b>R-HSA-2 980736</b>	Peptide hormone metabolism	1	126	0.008676 490841	0.091406 844	0.091406 844	1	63	0.004692 387904	CTSZ
<b>R-HSA-5 576891</b>	Cardiac conduction	1	138	0.009502 823303	0.099703 29298	0.099703 29298	2	27	0.002011 023387	FXJD4

**Supplemental Table 1: Reactome pathway enrichment results for PBX3/PKNOX1 cCRE associated genes.**



**Supplemental Figure 1: Genome Browser tracks for MYRF and ATP11A.**



**Supplemental Figure 2: LOLA enrichment analysis for ENCODE TFBS.**