2025

Predicting the effect of human, heart-related regulatory SNPs in their native and orthologous mouse genomic contexts

ZENTRUM FÜR KREBSFORSCHUNG
OSCAR GAMARRA SIAPO

Table of Contents

1. Abstract

Congenital heart disease (CHD) is the most common developmental malformation in newborns. Recent genetical studies have raised attention at the role of non-coding regions in CHD. To study these variants, researchers relied on mouse and human models. However, with over a billion known single nucleotide polymorphisms (SNPs) and the lack of scalable assays, experimental validation remains largely unfeasible.

To address these challenges, we trained ChromBPNet base-resolution models using single cell ATAC-seq data from human and mouse fetal cardiac tissue. We retrieved SNPs from human GWAS, obtained mouse orthologues, and used our trained models to predict variant effects. Using these predictions, we aimed to study cross-species concordance in variant effect. This analysis showed that variant effect predictions were cell type and trait dependent and highly correlated in early developmental cell types.

2. Plain language summary

Congenital heart disease (CHD) is the most common birth defect in newborns. This disease happens when the heart does not grow appropriately during embryonical development and is life-threatening. Scientists have discovered some mutations in genes that code for proteins that cause CHD, but most of the other mutations fall in the non-coding genome. The non-coding genome is the part of the genome that does not create proteins. However, these regions often play a role in deciding which proteins are being made or not. These regions achieve this regulation by modifying the chromosomes, changing the accessibility of the genome to the machinery needed to create the proteins.

The non-coding genome makes up to a 98% of our genome and there are billions of possible changes that may occur in such regions. Because there are so many variants, it is almost impossible to do experiments and check the effects of each single variant. In this project we focused on single nucleotide variants (SNVs).

On top of this, to run these experiments we need to use models. These models can come from different species. Since we are interested in human CHD, we want to use models that resemble as much as possible the human development of the heart. One of the most popular models used is the mouse. But this comes with an extra challenge, which is to see to what extent the findings done in mouse are applicable to human.

To overcome these problems, scientists have instead started relying on machine learning. This term stands for tools that train on data from experiments and learn from them. Using these tools, we can use the current datasets available, learn patterns from DNA sequences, and predict the impact of any possible variant with our computers. More specifically, we used a tool named ChromBPNet. This tool was very useful for us because it had base-resolution, meaning that we could introduce our SNVs and investigate their impact.

To train these models we used human fetal heart sequenced with single-cell ATAC experiments. This assay detects the areas of the genome that are more accessible (and thus more likely to be translated into proteins) with a single cell resolution. It is very important to see the accessibility per each cell, because then we can also annotate their cell type. Having the cell type annotation allows us to detect patterns of accessibility specific to each cell type.

To make the comparison with mouse, we trained another set of models on single-nuclei ATAC-seq data from heart embryonic mouse.

With our models trained, we predicted the impact of SNVs linked to several traits (like CHD, asthma or rheumatoid arthritis) in gene regulation.

Overall, we demonstrate the potential of machine learning in studying gene regulation using non-coding variants. Our findings show that the variant effects predicted depend on the cell type and trait analyzed. We also found that the correlation between mouse and human predicted signal correlates better in early developmental cell types.

3. Introduction

3.1. Congenital heart disease

Congenital heart disease (CHD) refers to structural or functional abnormalities occurring during cardiac development (Zhao et al., 2024). Heart conditions are the most common type of congenital disease, occurring in 1% of newborns and accounting for 30% of fetal deaths. CHD is also associated with a high perinatal morbidity and mortality (Meller et al., 2020). Medical and surgical advances have significantly improved the survival of individuals with CHD, leading to an increased prevalence among older children and adults (Pierpont et al., 2018).

To better understand the molecular mechanisms of CHD, extensive genetic studies have been performed. Studies focused on the genetics of CHD found only 10% of all cases to be explained by single gene mutations of morphogens, transcription factors (TF) or other genes involved in cardiogenesis (Wang et al., 2022). These findings suggest that non-coding genomic regions may play a substantial role in CHD etiology.

A clear example of how non-coding genomic regions can influence gene expression patterns and cell identity are cis-regulatory elements (cREs). cREs are sections of DNA to which transcription factors bind to regulate gene expression, often having long-range interactions across the genome. Examples of cREs include promoters, enhancers, and silencers (Kim & Wysocka, 2023). Several studies have linked variants within cREs to CHD predisposition (J. Ma et al., 2021; Song et al., 2021). However, these findings were experimentally validated *in vitro*, which poses a substantial problem considering that non-coding genome comprises around 98% of the human genome. The vast number of possible non-coding variants makes experimental validation difficult. Moreover, current validation assays are relatively low throughput compared to the potential variants. In summary, there is a lack of understanding of the impact of variants in non-coding genomic regions in CHD, as well as a challenge to experimentally validate all potential candidates.

3.2. Single-cell technologies to map regulatory activities

Single-cell (sc) technologies are a well-consolidated class of high-throughput that allows us to characterize biological processes at different levels of the central dogma of biology (DNA – RNA – protein). Single-cell analysis helps us study complex cellular phenotypes at a single-cell resolution, and it has become the gold standard to study tissue heterogeneity (Tian et al., 2020). Cellular heterogeneity is especially important in developmental biology. Single-cell analysis has helped to better understand key stages of human development (Aissa et al., 2021; Fawkner-Corbett et al., 2021; Liang et al., 2022).

Single-cell technologies can be coupled to ATAC-seq, which stands for "assay for transposase-accessible chromatin sequencing" (Buenrostro et al., 2015). This assay locates open chromatin regions within each cell's genome using biologically engineered hyperactive Tn5 transposases (Grandi et al., 2022). These regions are more likely to be regulated via cREs and transcription factors. Maintaining single-cell resolution is crucial, as it helps us identify cell-type specific gene expression regulation patterns (S. Ma & Zhang, 2020). These technologies help elucidate the impact of non-coding genomic regions in both disease and developmental contexts, as shown in recent publications (Hocker et al., 2021; K. Zhang et al., 2021).

3.3. Comparative modeling of heart development between human and mouse

To run scATAC-seq assays, we need models that mimic as closely as possible the *in vivo* human conditions of CHD. Mouse models have been the foundation for several genetic and developmental studies on CHD. For instance, a study showed the importance of TGFß signaling in correct development of the mesenchymal cushions, the aortic arch, and vascular smooth muscle cells (Arthur & Bamforth, 2011). Similarly, the first evidence of the importance of cilia-related genes in CHD came from chemically mutagenized mouse models (Zhu et al., 2023). However, despite their utility, these models face significant limitations when studying non-coding variants, as large numbers of mice are needed for validation, maintenance costs are high, and experiments are time-consuming (Majumdar et al., 2021).

On the other hand, human models are unusual for CHD characterization. Human samples are extremely difficult to obtain at different stages of development. A recent publication analyzed fetal heart tissue from three different individuals at three different stages of development (Ameen et al., 2022). Given the challenge of retrieving human samples, other human models have been developed to model heart development and CHD using other methods. An instance of this is human induced pluripotent stem cells (hiPSC) derived cardiomyocytes (Doyle et al., 2015). In the same study, the authors highlight the benefits of hiPSC models because of direct applicability to human disease and uniqueness of human phenotypes; the ability to easily manipulate the cell lines; and the ability to study tissue organization. More recently, cardiac organoids, or cardioids, have been developed (Hofbauer et al., 2021; Schmidt et al., 2023). These cardioids are 3D self-organized heart-like structures. They have the benefit of being more heterogeneous and closer to human hearts regarding cell type populations and tissue architecture.

As we have seen, there are several models to study cardiac development. Several efforts have been carried out to model human non-coding variants in mouse. For instance, in a recent publication, researchers studied the role of Human Accelerated Regions (HARs), which comprise several human-specific enhancers. To study HARs, they used genetically humanized mouse models (Dutrow et al., 2022). They describe how HAR2 maintains enhancer activity specific to humans in mouse embryo. More recently, scientists also developed dual-enSERT, an assay that enables comparison of enhancer allele activity in live mice (Hollingsworth et al., 2025). In the same study, they identify variants that modify OXT2 and MIR9-2 activity, two brain enhancers linked to autism. However, as we can see, these studies focus on few variants and are low throughput compared to the number of potential variants in non-coding regions.

Additionally, we also encounter a challenge which is how can we compare human and mouse models among each other, considering also that they come from different species. Few studies have researched about the differences and similarities in human and mouse heart development. An image-based comparison reported that human and mouse early healthy heart development was similar, with minor differences only in atrial and venous morphology (Krishnan et al., 2014). Another key publication analyzed the gene expression patterns across cardiomyocytes and endothelial cells development in human and mouse embryos (Liu & Shao, 2024). In this study, researchers made correspondences between human and mouse developmental stages, as well as discovering important transcription factors relevant for interspecies cardiomyocytes differences. However, to our knowledge, there have been no clear attempts to quantify how many genetic variants identified in human genetic studies can be modeled in the context of the matching regulatory landscape in the mouse.

3.4. Computational modeling of regulatory variants using Deep Learning

As we have seen in the previous sections, we found two major challenges in current literature. First, the problem posed by the high number of non-coding variants and its unfeasibility to experimentally validate their effects via *in vitro* or *in vivo* models. Secondly, a gap in understanding the similarities among regulatory landscapes of different species during cardiac development.

To address both challenges, several deep learning tools that model and predict variant effects have been developed. *In silico* approaches enables computational validation at a large scale, reducing the time and expenses needed using *in vivo* models. An instance of such tools can be seen in VEGN (Cheng et al., 2021). VEGN employs a graph neural network (GNN) and creates a graph by assigning a variant to a gene and then builds a gene-gene interaction network. The model is then trained to add gene-gene information, as well as gene-variant, depending on the genes each variant is connected to. Another example is DREAM (Li et al., 2024), a deep learning tool built to predict sequence-based enhancer activity and design synthetic enhancers across several species.

In the end, we used ChromBPNet (Pampari et al., 2024). This tool is a deep learning DNA sequence model of accessibility profiles at base-pair resolution, making it a great candidate to work with SNVs. A great advantage of ChromBPNet is the way it predicts the profiles. First, a convolutional neural network (CNN) is trained to model unstranded, base-resolution ATAC-seq accessibility profiles in 1Kb bins as a function of local DNA sequences (2114bp). Base-resolution is essential for our project, as we want to evaluate the impact of single-nucleotide, non-coding variants across human and mouse species. To evaluate this impact, we: 1. trained human and mouse models with ChromBPNet; 2. retrieved single nucleotide polymorphisms (SNPs) linked to several traits obtained through human GWAS studies; 3. obtained their mouse orthologues; and 4. compared the predictions for each set of SNPs.

Importantly, ChromBPNet corrects for bias introduced by the mutant hyperactive Tn5 enzymes used in ATAC-seq assays. To achieve this, ChromBPNet trains a model that captures the enzyme effect on the chromatin background. This model is then used to regress out the Tn5 enzyme effect from the chromatin profiles.

3.5. Aim of the project

We have covered the challenges we want to address. Our intention in this project is to study the human regulome by predicting variant effect of cREs using ChromBPNet. We want to study to what extent can mouse cistrome can be used to model variants found in human regulatory regions active in heart development.

4. Methods

4.1. Data acquisition

We trained our ChromBPNet (v.0.1.7) models using publicly available data from human fetal cardiac tissue scATAC-seq experiments (Ameen et al., 2022). The data is available on the Gene Expression Omnibus database with the reference code number of GSE181346. This publication has scATAC-seq data from samples from heart fetal tissue at 3 different postconception weeks (PCW): 6, 8 and 19. In this publication only the raw data (fragments in a tsv format) is available. The authors found a total of 20 different cell types across all samples. We used the cell type annotation for each cell provided by the publication. The cell types found were:

Cell type	**Abbreviation**	**Cell type**	**Abbreviation**
Atrial cardiomyocytes	aCM	Fibroblast-like cells 2**	FB2
Arterial endothelial cells	aEC	Lymphatic endothelial cells	lEC
Capillaries	Cap	Neural crest	NC
Cardiac fibroblast	CF	Outflow tract	OFT
Cardiac fibroblast progenitors	CFP	Pericytes	PC
Early cardiomyocytes	eCM	Pre-cardiac fibroblast	preCF
Endocardium*	Endo1	Pre-smooth muscle cells	preSMC
Endocardium-like cells*	Endo2	Smooth muscle cells	SMC
Epicardial cells	EPC	Ventricular cardiomyocytes	vCM
Fibroblast-like cells 1**	FB1	Venous endothelial cells	vEC

*Table 1. Full description of clusters provided by Ameen et al. *Authors report that there were 2 different phenotypes across endocardial cells: Endo1 had high CDH11 high GA-score, whereas Endo2 resembled endocardial-like transitional cells. ** Authors report that FB1 had high TBX18, TCF21 and low WT1 GA-scores, whereas FB2 had high CNN1 and COL9A2 GA-scores.*

4.2. Peak calling with MACS2

Using the scATAC-seq data from Ameen et al., we merged fragments belonging to the same cell type across all PCW, which we will refer as "merged fragments". We ended up with a total of 20 merged fragments files. Afterwards, we called peaks across all merged fragments files using MACS2 (v2.2.9.1) using its *callpeak* function, setting a p-value $\leq$ 0.01 as recommended on the ChromBPNet Github repository. Following Ameen et al. peak calling strategy, we:

1. Sorted the peaks by q-value.
2. Removed overlapping peaks, keeping only the most significant peaks. Without this filter, only overrepresented high-quality regions would have been kept. This would drastically undermine underrepresented regions, thus worsening prediction for regions with low number of counts.
3. Removed peaks within blacklisted regions, as these regions are anomalous, unstructured and result often in high signal regardless of the experiments performed (Amemiya et al., 2019).
4. Removed peaks mapped to chrY to reduce bias produced by sex of sample of origin.
5. Removed peaks with "N" nucleotides in the sequence to avoid ambiguous regions.
6. Capped all peaks called per cell type to the top 215k most significant peaks due to computational limits during the training of the models.

If the number of peaks was lower than this, no imputation was performed.

Cell type	**Total number of peaks called**	**Final number of peaks**	**Cell type**	**Total number of peaks called**	**Final number of peaks**
aCM	1364942	215000	FB2	1271670	215000
aEC	207279	206906	lEC	558680	215000
Cap	356763	215000	NC	903413	215000
CF	369255	215000	OFT	1365858	215000
CFP	222265	215000	PC	1768096	215000
eCM	1490387	215000	preCF	875074	215000
Endo1	1293347	215000	preSMC	1382688	215000
Endo2	136545	136333	SMC	1365780	215000
EPC*	549542	215000	vCM	298955	215000
FB1	1361500	215000	vEC	938242	215000

Table 2. Number of peaks after callpeaks *(Total) and after filtering (Final). As stated, all the cell types were capped to the 215k peaks with highest q-values. Only two cell types did not reach this number (aEC and Endo2). *EPC model training failed due to unresolved issues.*

After applying this filtering to all peaks, we generated non-peaks for each cell type. Non-peaks refer to background regions whose GC regions match with the peak regions. This step is necessary to train the models and are used as background regions.

4.3. Model training with ChromBPNet

Once we generated both peaks and non-peaks, we proceeded with the per-cell type model training. To obtain the trained models we used the ChromBPNet function named *pipeline*. This function requires the peaks, the non-peaks, chromosome sizes, a chromosome fold file, and a bias model. We downloaded the human chromosome sizes (GRCh38) from the Encode Project website. The chromosome fold file is required to divide the dataset in training, validation, and testing. To obtain this file we used ChromBPNet's *prep splits* function, using their "fold 0". This fold divides the dataset in the following chromosomes:

- Training: 2, 4, 5, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, X.
- Validation: 8, 20.
- Testing: 1, 3, 6.

Finally, the developers of ChromBPNet also provide a bias model ready to use. Because the bias model captures enzyme bias, we would need to train a model depending on the type of the assay used. The bias model they provide is trained already on scATAC-seq data, which means that this model captures Tn5 bias. Thus, the model provided is also suitable for our data as well, as we are using data obtained from the same type of assay. Only then we proceeded to use the *pipeline* function from ChromBPNet to obtain the models for each cell type. To visualize the predicted regulatory landscapes of each model, we loaded the bigwig files provided by ChromBPNet's *pipeline* function on IGV (v2.19.4).

4.4. Model evaluation

One of the features of ChromBPNet is that we obtain an evaluation report for each model trained. In these reports we find metrics that evaluate both the bias and the ChromBPNet model performance, such as Pearson correlation and Jensen-Shannon Divergence (JSD) values. Pearson correlation calculates measures the linear correlation between two variables, whereas JSD calculates the difference between to probability distributions (with 0 being exactly similar distributions, and 1 being completely different). These values are obtained only if a reference file, namely, the merged fragments files for each cell type, is provided during the training of the model. These files are then used as ground truth, and the metrics calculate the differences between observed values and predicted values.

4.5. Human SNPs collection

As mentioned in the introduction, we aim to study the impact of regulatory SNPs on heart development. To achieve this, we started by selecting sets of SNPs to evaluate with a focus on heart-related traits but also including potentially unrelated traits. We downloaded cureated top association SNPs linked to different traits from the GWAS catalog. As these SNPs were retrieved from a human catalog, we used LiftOver to obtain the mouse orthologues. As original assembly we used hg38, and for the mouse assembly we used mm39. To ensure that the orthologue positions were not artifacts, we used LiftOver again, this time from mm39 to hg38. We kept only the SNPs that identically matched the original positions, namely, the SNPs that we could LiftOver twice.

Trait	**Clinical significance**	**Total number of SNPs**	**LiftOver number of SNPs**	**Percentage of SNPs kept**
Asthma	Disease	3543	1466	41.38%
Atrial fibrillation	Disease	965	490	50.78%
BMI	Benign	15897	7327	46.09%
Body height	Benign	23246	6734	28.97%
Brain disease	Disease	24733	9508	38.44%
Congenital heart disease	Disease	331	125	37.76%
Coronary artery disease	Disease	3469	1436	41.40%
Diastolic blood pressure	Benign	4228	2017	47.71%
HDL	Benign	14038	6147	43.79%
Heart failure	Disease	451	163	36.14%
Heart rate	Benign	903	407	45.07%
LDL	Benign	9970	4332	43.45%
PR interval	Benign	612	324	52.94%
Rheumatoid arthritis	Disease	3439	1305	37.95%
Schizophrenia	Disease	5722	2103	36.75%

Systolic blood pressure	Benign	5653	2617	46.29%

Table 3. Traits selected for our study. Total number of SNPs refers to the number of SNPs available from GWAS catalog. LiftOver number of SNPs refers to number of SNPs after filtering (double LiftOver). As we can see in the last column, all traits kept around half of the original SNPs.

4.6. Variant effect prediction

Once we had all the sets of SNPs filtered, we proceeded to test their impact using our trained models. We first started by obtaining the base of reference at each position using bedtools (v2.28.0) in combination with the human (hg38) and mouse (mm39) reference genomes. We also curated a dataset where we added the genomic context (±50 nucleotides from the genomic location of the SNP), as well as the phyloP score of each SNP for all traits. The phyloP score reflects the evolutionary constraint at a specific genomic location, where positive values indicate slower evolution than expected, and negative values indicate faster evolution than expected.

Once we had the reference base, we retrieved the alternative allele of each SNP using the human reference VCF. Because the mouse reference VCF is not as well-characterized as the human, many of our alternative mouse alleles were not found in the mouse reference VCF. To overcome this issue, we assigned to each mouse SNP the corresponding orthologue human alternative allele. We then applied a last filtering step, which consisted of keeping only the SNPs where human and mouse had the same reference base. Due to this restriction, most SNPs were dropped. This step was added to maintain only SNPs with similar context across both species. Additionally, to simplify downstream analysis, we selected only a few traits. To select the traits, we wanted to include a non-heart related trait to compare to CHD, as well as benign and disease-related traits to use as control.

Trait	**Part of the body affected**	**Clinical significance**	**LiftOver number of SNPs**	**Final number of SNPs**	**Percentage of SNPs kept**
Atrial fibrillation	Heart	Disease	490	67	13.67%
Congenital heart disease	Heart	Disease	125	21	16.8%
Heart failure	Heart	Disease	163	38	23.31%
PR interval	Heart	Benign	324	65	20.06%

Rheumatoid arthritis	Joints	Benign	1305	201	15.41%

Table 4. Selected traits for downstream analysis. Here we see a greater loss of SNPs after filtering, mostly due to the matching reference base criteria and insufficient SNP characterization in VCF.

Finally, we used ChromBPNet's variant scorer to predict the impact of our selected SNPs. For human models, we used the 19 models that we trained on human fetal cardiac tissue. For mouse models, we used models trained on unpublished embryonic heart tissue obtained by Roland et al, obtained using snATAC-seq. This data was obtained from three different timepoints: E9.5, E10.5 and E11.5. The mouse snATAC-seq data includes the following cell types:

- Cardiomyocytes (CM)
- Anterior second heart field-derived cells (aSHF)/ outflow tract (OFT)
- Posterior second heart field-derived cells (pSHF)
- Cardiopharyngeal mesoderm progenitors (CPM)
- Endocardial cells (EC)
- Vascular endothelial cells (VEC)
- Endocardial cushion mesenchymal transition cells (EndMT)
- Epicardial cells (EP)
- Neural crest cells (NCC)

As we can observe, the mouse data is not as finely annotated as the human data. To overcome this issue, we grouped cell types from both human and mouse data into "superclusters", which are groups of cell types with similar characteristics across both species. This also helped with biologically interpretation downstream. The superclusters were created as follows:

Supercluster name	Supercluster abbreviation	Human cell type	Mouse cell type
Cardiomyocytes	CM	eCM, aCM, vCM	CM
Anterior second heart field & Outflow tract	aSHF/OFT	OFT	aSHF/OFT
Smooth muscle & posterior second heart field	pSHF	preSMC, SMC	pSHF
Endothelial cells	EC	aEC, vEC, Cap	EC
Lymphatic and venous ECs	VEC	lEC	VEC

Endocardium/Endocardial cusion mesenchymal transition cells	EndMT	Endo1, Endo2	EndMT
Epicardial/Fibroblast-like	EP	EPC, CF, preCF, CFP, FB1, FB2	EP
Neural crest	NCC	NC	NCC
Others	PC & CPM	PC	CPM

4.7. Evaluation of interspecies differences

To assess the differences, we loaded the output of ChromBPNet on RStudio (v4.5.1). All plots later presented in the Results section were generated using ggplot2 (v3.5.2).

5. Results

5.1. Prediction of cell-specific regulatory landscape

With our models trained, we wanted to have a first glance at the prediction of regulatory landscapes. In Figure 1, we see the genome tracks of all cell types. We selected only the regions corresponding chromosome 12 (Chr12) for memory reasons.

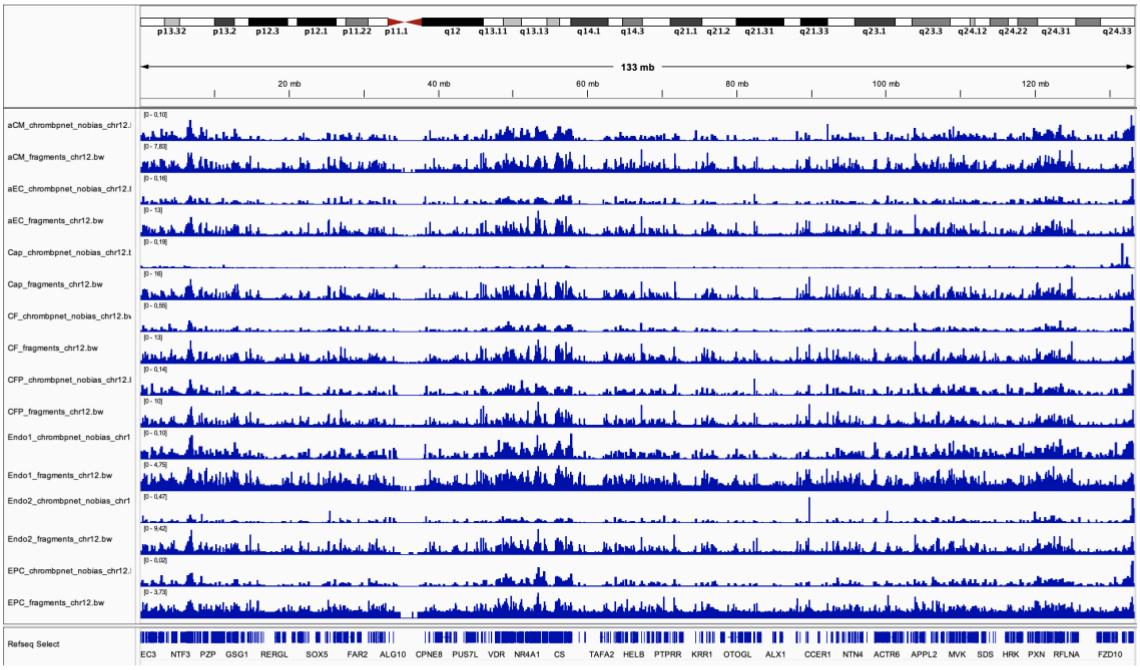


Figure 1. Genome tracks of top 10 cell types by alphabetical order, grouped in couples. Per cell type, the top track, which ends in "chrombpnet_nobias_chr12", is predicted. The bottom track, which ends in "fragments_chr12", is the reference, namely, the raw reads from merged fragments files. Above each track between brackets, we see the scale of the bands.

We can see how some models (corresponding to the Cap, CF and Endo2 cell types) show poor predictive performance. In these cases, the predicted track recovers virtually no signal compared to the reference values, especially for Cap. This happens likely due to the low coverage for these cell types, as can be seen in section 4.2. (Peak calling with MACS2). ChromBPNet produces more accurate models with higher coverages, as seen with the rest of the cell types.

If we compare the scale of the predicted tracks with the reference, we see that the reference tends to show a higher signal-to-noise ratio. Some factors that may explain this difference are the filtering and capping steps before training each model.

To have a more detailed summary of the accuracy of the models, we used ChromBPNet's evaluation reports. In these reports two main metrics are reported: correlation scores and

JSD, along with scatter plots of observed against predicted counts, and a density plot of JSD values of predicted and shuffled values against observed values.

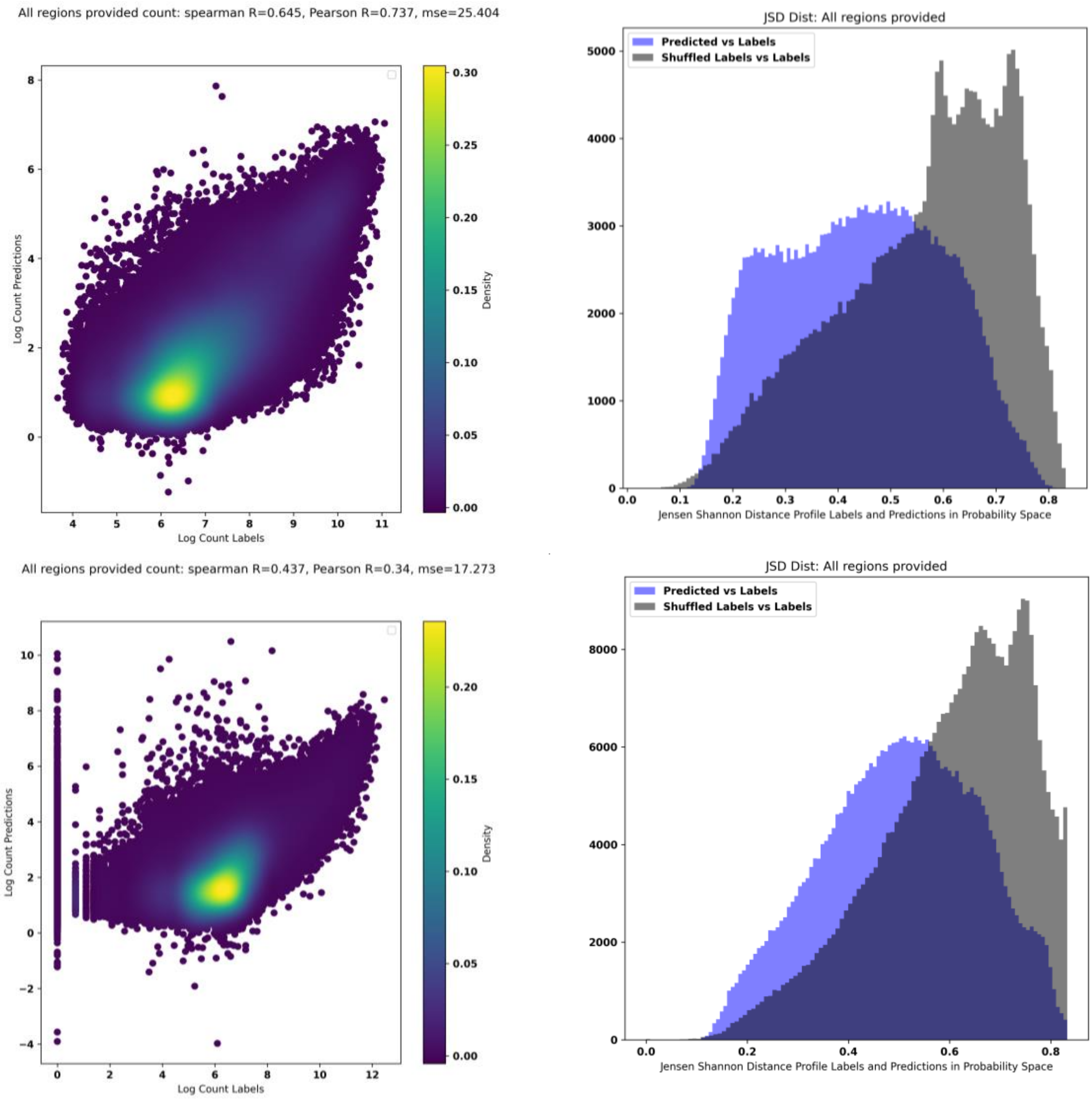


Figure 2. Example of plots provided by ChromBPNet's evaluation. Top belongs to aCM's model, bottom belongs to Cap. Left, scatter plot of log observed against log predicted counts, colored by point density where yellow is higher and blue is lower, and titled with both Spearman and Pearson correlations, as well as mean squared error (mse). Right, density plot of JSD of predicted against observed in blue and random shuffle against observed in grey.

In Figure 2, the scatter plots also display a scale difference between observed and predicted log counts. We observe that the regions with highest observed counts fall between the X=10 -11 in aCM, and 10-12 in Cap. In the aCM model, the predicted values fall between Y=6-7, whereas in Cap they reach Y=10. We hypothesized two reasons for the disparity between predicted and reference scales. First, that the bias model is regressing out background noise from the assay. Second, because we used as reference the unfiltered merged fragments files, we also hypothesized that this scale difference could be because of the filtering steps applied before training. As we reminded when looking at the genomic tracks, one of the filtering steps was the removal of overlapping peaks.

In the Cap scatter plot we see a set of values for which we find no reference values (X=0), but the model does predict signal. This can be explained due to the low read depth for Cap, which resulted in a lower number of peaks called for this cell type. The model could have learned sequence features associated with high chromatin accessibility regions and transfer them to other regions even if there is no detectable signal.

Shifting the attention now to the density plots, we observe that the aCM's predicted vs labels (leaning towards the right) is clearly separated from the random shuffle vs labels distributions (leaning towards the right). For Cap, however, the overlap between both distributions is higher. This further demonstrates the need for a higher read depth, as aCM's model achieves great results in both correlation between predicted and reference, as well as a shorter distance between predicted against reference distributions.

To summarize, these results highlight the importance of sequencing depth to train accurate models using ChromBPNet, as well as the utility of the evaluation reports to further evaluate each model individually.

5.2. Examining cross-species conservation of human GWAS SNPs across traits

Prior to the examination of the predicted variant effects, we ran a preliminary conservation analysis of the GWAS we selected. This part of the study focused on gaining a broad perspective on base conservation between human SNPs and their respective mouse orthologues across all the traits represented in our dataset.

We started this analysis with the hypothesis that, because conservation typically indicates functional importance of such region, the variants associated to disease-related traits would have overall a higher conservation score. As we saw in section 4.6. (Variant effect prediction), we annotated each SNP with phyloP scores, thus enabling us to link a degree of conservation to each genomic location.

Our first step then was to explore base conservation between human and mouse. To do so, we simply identified the SNPs where the reference nucleotide was the same one both in human and mouse.

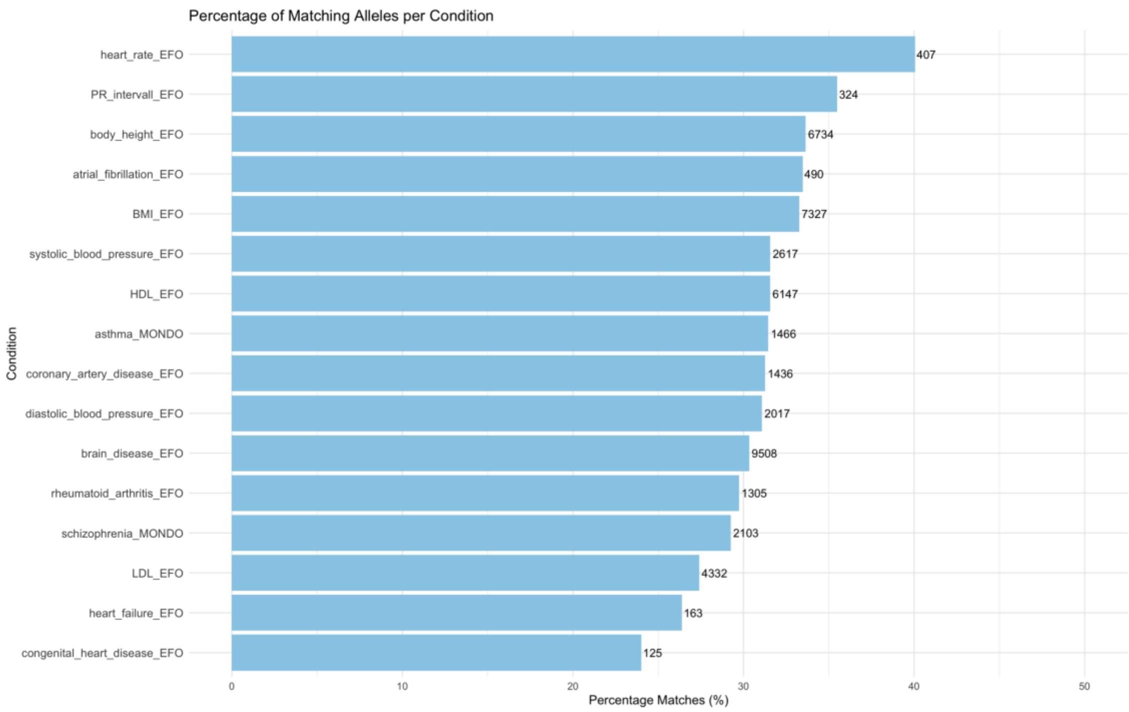


Figure 3. Horizontal bar plot of percentage of matching alleles per condition. On the X axis we see the percentage of nucleotides where reference human and mouse match. On the Y axis we see all the traits selected, ordered from highest to lowest. Right next to the bar we find the number of SNPs after double LiftOver.

From the bar plot of Figure 3, we see that there are few matching reference alleles between human and mouse, with heart rate having the highest value at around 40%. We find the lowest value in congenital heart disease, which also has the least number of SNPs after filtering. However, there seems to be no correlation between total number of SNPs and percentage of matches. We further confirm this observation by comparing heart rate, which tops the plot having 407 SNPs, to LDL, which lies third-to-last with 4,332 SNPs. More importantly, it is very curious to see that on the upper half of the list we find mostly benign traits (heart rate, PR interval, body height, BMI, systolic blood pressure, HDL). Contrarily, on the bottom half we find mostly disease-related traits, such as brain disease, rheumatoid arthritis, schizophrenia and heart failure. More traits could be added to potentially elucidate whether this occurs on a larger scale as well.

To further explore the evolutionary constraints of each trait, we used the phyloP scores we added to our dataset. As a first exploration, we wondered whether some SNPs tended to have higher conservation in one species over the other. This would lead to differences between the distributions of phyloP scores between human and mouse. To visualize the distributions, we represented in a density plot the phyloP scores from all SNPs altogether, separated by species, as can be seen in Figure 4.

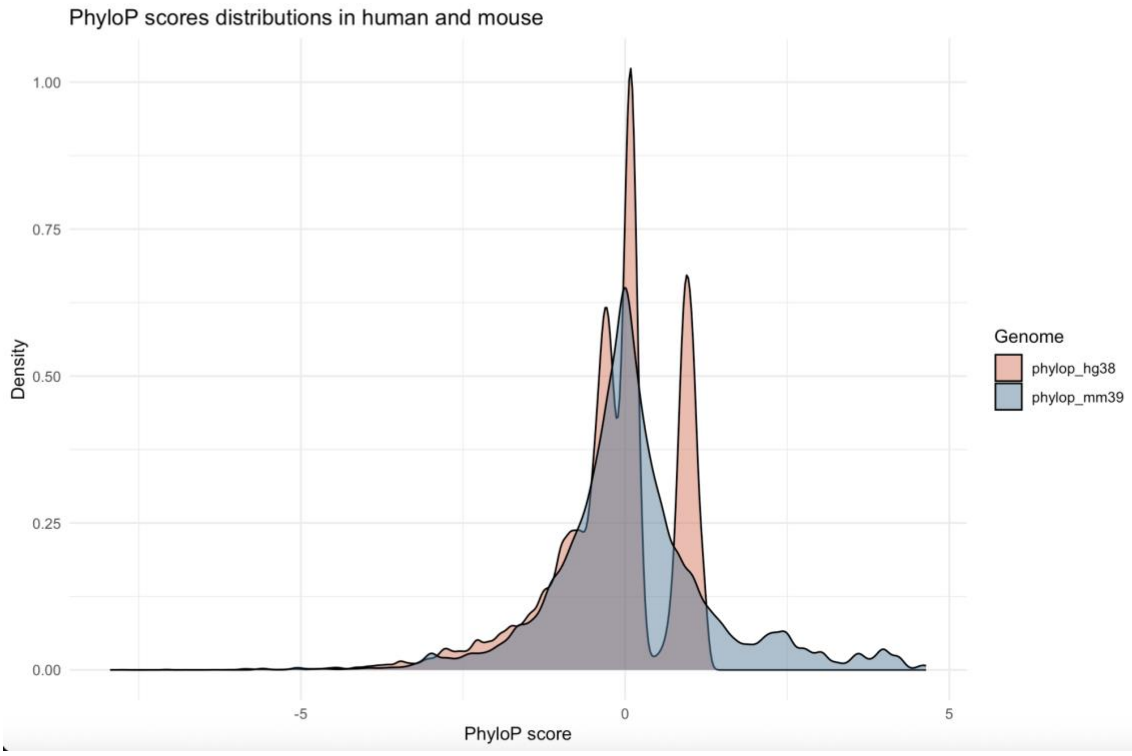


Figure 4. Density plot of phyloP scores in human and mouse for all traits. In the X axis we find the phyloP score. In the Y axis we find the proportion of values in each range. In orange we see the human phyloP scores distribution, whereas in blue we find the mouse phyloP scores distribution.

In these density plots we observe that there is a clear difference between human and mouse phyloP scores distributions. The human distribution seems to be bimodal, with one peak falling around X=0 and another to be around X=1. The highest peak reaches Y=1, and also has another small peak with phyloP scores smaller than 0. We can also see that the maximum X values fall around 1.5. As for the mouse distribution, we see that it is rather skewed towards the positive values, in contrast with the human distribution. There seems to be a greater set of SNPs with high conservation scores in mouse, reaching values phyloP scores slightly below 5. This could be generated due to the way the phyloP scores are generated, as the human values are capped to a maximum value of 3, as it is stated here.

We also wanted to explore why is there a bimodal distribution in human phyloP scores, and why is it different from the mouse's distribution. Our hypothesis was that disease-related traits clustered around the X=1 peak in human, as probably they are variants that alter more gene expression. Thus, changes in these bases would tend to have higher phyloP scores. To see whether the clinical significance was related to the phyloP score, we separated the traits following the Clinical significance classification from Table 3. Then, we created a violin plot to visualize the distributions across both levels, benign and disease.

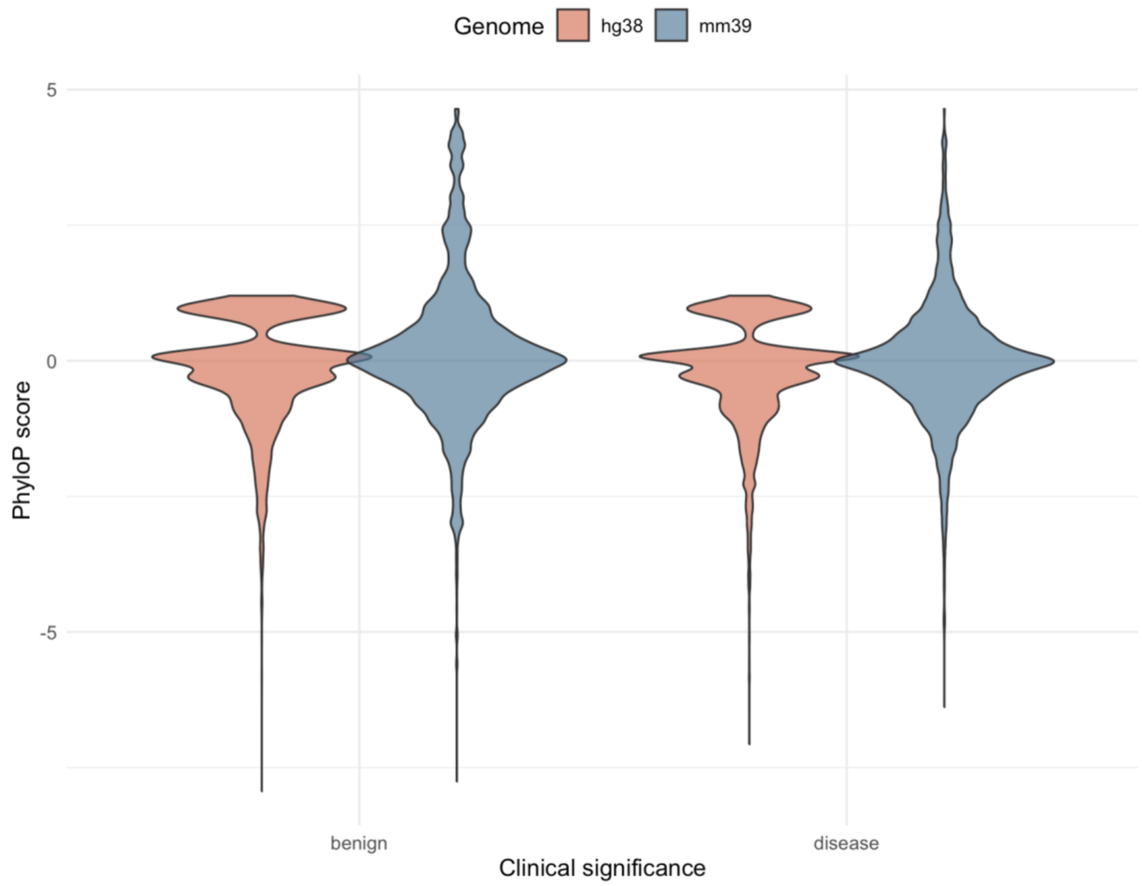


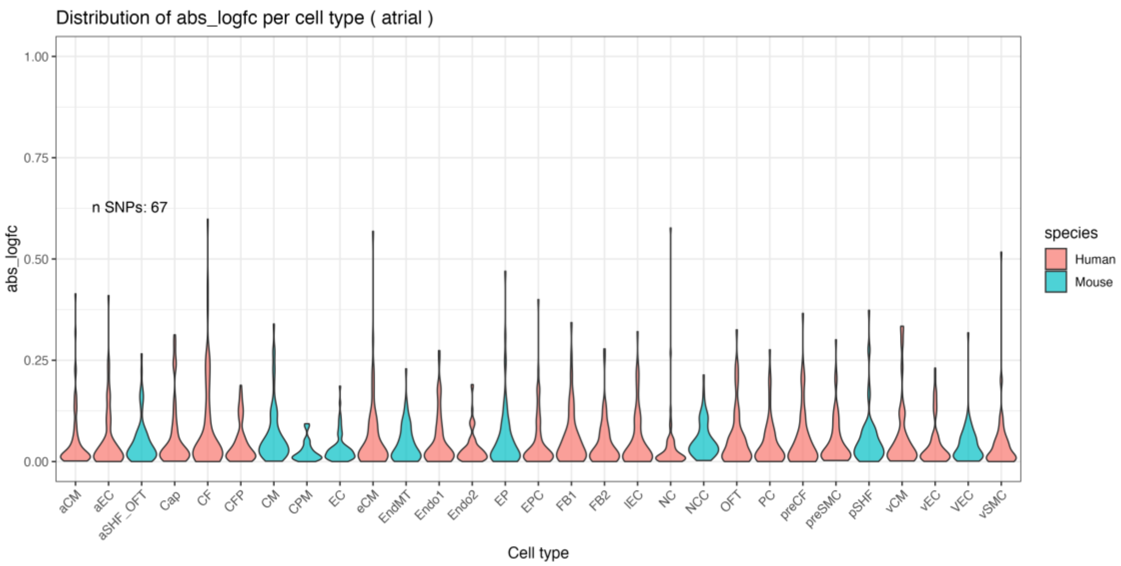
Figure 5. Box plot of phyloP scores grouped by clinical significance. Left values are for benign traits; right values are for disease-related traits. Y axis represents phyloP scores. In orange we see phyloP scores for human, in blue we see phyloP scores for mouse.

As we can see in Figure 5, there are again 2 peaks where most SNPs fall for human scores, exactly in the same values we saw in Figure 4, and for both benign and disease. This plot demonstrates that clinical significance does not fully explain the bimodal distribution, and thus we should look at something else. Other possibilities that might explain this are genomic location of SNP. If the SNPs fall in cREs closer to genes, or if they fall in coding regions, the impact might be greater, and thus tend more to be preserved, resulting in higher phyloP scores. As for the mouse, distributions do seem different from benign to disease. If we look at the Y positive values, the benign group has a tends to accumulate more variants. Once again, we see a difference between how human and mouse phyloP scores are distributed. However, we did not look further into this difference due to time constraints.

To summarize this section, we want to emphasize that, although same base conservation is generally low between human and mouse, polymorphic sites identified as correlated with differences not related to diseases (benign traits) seem to have overall higher percentage of conservation. When looking at conservation scores, we observed generally higher values in mouse, although probably because of how phyloP scores are calculated. We also discovered a bimodal distribution in human phyloP scores that was not present in mouse and suggested some potential explanations for this difference.

5.3. Predicted variant effects reveal cell type and trait-specific changes in chromatin accessibility

Having explored the cross-species conservation of SNPs from human GWAS, we continued studying the effects of these sets of variants on chromatin accessibility. To study the impacts of these variants, we used ChromBPNet's variant effect scorer, which gives us two metrics: the log fold change (logFC) of allele 2 (alternative) against allele 1 (reference); and JSD, which captures shifts on the profile shape. Thus, we have the logFC and JSD of each SNP for the traits mentioned in 4.6. Variant effect prediction for all models (19 for human, 9 for mouse). Our first approach at this data was to visualize per the effects of the set of SNPs belonging to CHD. We decided to compare it to atrial fibrillation, as it is another disease-related trait.



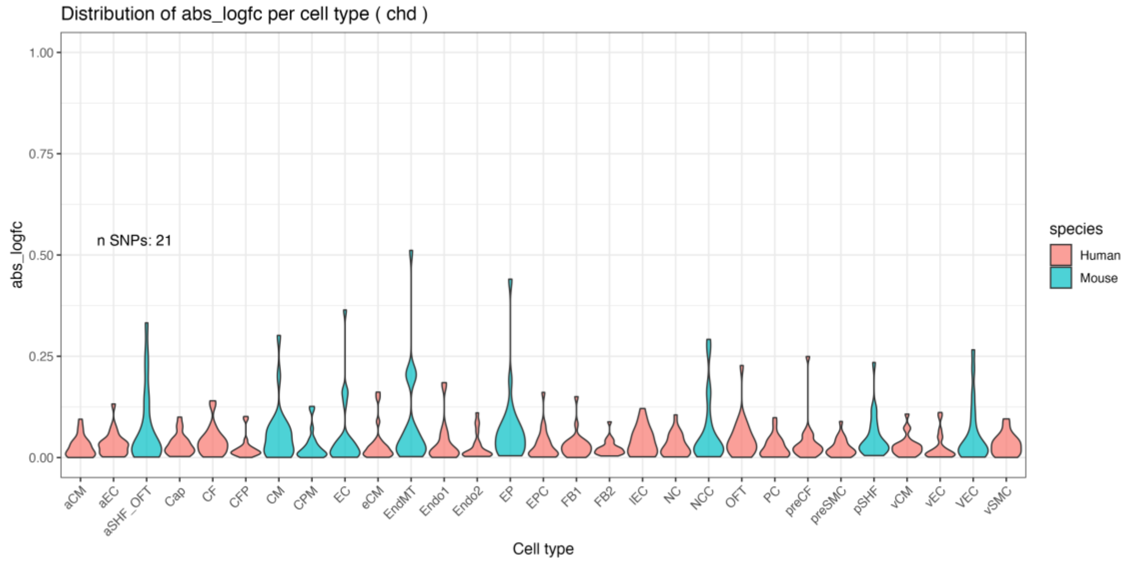


Figure 6. Two violin plots of distribution of absolute logFC across all our models colored by species (human in orange, mouse in blue). On top, for atrial fibrillation (with 67 SNPs); on the bottom, for CHD (with 21 SNPs). Across the X axis we see the cell types on which the models were trained; on the Y axis we find the absolute logFC.

In figure 6 we see that we had more SNPs for atrial fibrillation than for CHD. From the violin plots we observe that the maximum values in atrial fibrillation are overall higher than CHD, particularly in human CF, ECM and NC. CHD has values that reach 0.5, but in the mouse-trained models, particularly in EndMT. If we look closer at all the distributions, we find that atrial fibrillation finds many of the SNPs around values of Y=0.05, whereas CHD has more spread distributions, especially in mouse aSHF_OFT, CM, EP, or human CF, lEC and OFT. On CHD we also find that there are some SNPs that cluster towards maximum values as represented by the small dot on top of the violin plots. Examples of models that predict a high impact for these SNPs are human Endo1, OFT, preCF or mouse EC, EndMT or NCC. Contrarily, in atrial fibrillation the SNPs with higher values seem to be outliers.

After this overview across all models, we wanted to explore whether the effects of SNPs were similar across all cell types. We first focused on our human models and hypothesized that there should be specific cell types where the effect of SNPs should be greater than on the rest of the other ones. To evaluate this idea, we grouped all the trained models by species (human and mouse), compared the absolute logFC of one model against the mean absolute logFC of the rest of the models.

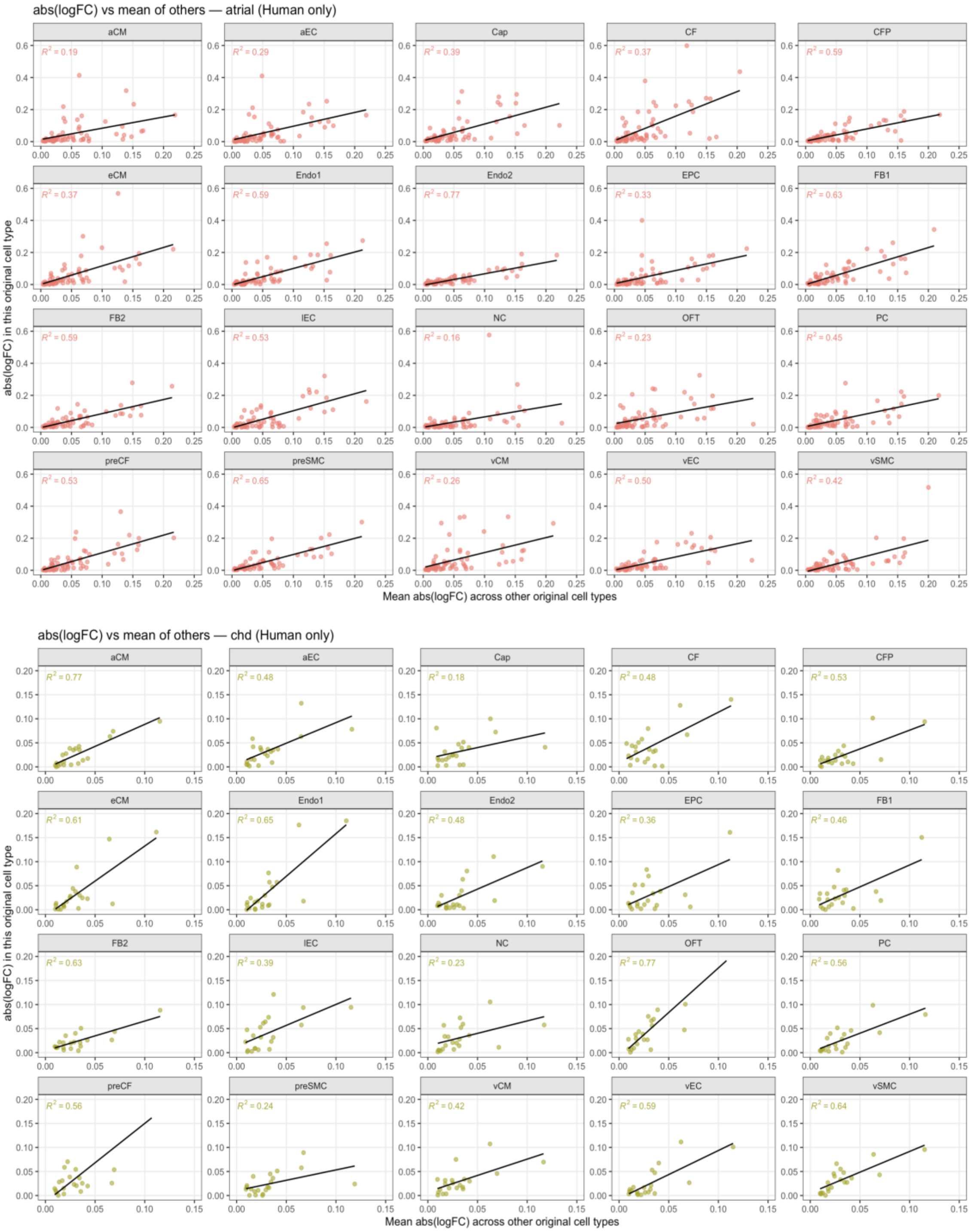


Figure 7. Scatter plots of absolute logFC per cell type (named on the header; Y axis) against mean absolute logFC of the rest of the cell types (X axis). On top, for atrial fibrillation, colored in red; on the bottom, for CHD, colored in green. We can also find the R^2 of each plot and the regression line.

In Figure 7 we can see that in CHD there are some cases where the Y value (queried cell type) has a different scale than the X value (rest of cell types). A clear example of this is Endo1, where there's a SNP (chr9:136404140-136404141) with a maximum Y value close to 0.2, but with a X value of around 0.1. We also see that in Endo1 we achieve a high correlation of both axes, suggesting that this difference in scale is not particular to that SNP, but a general tendency of the model. Shifting our attention to the Endo1 model in atrial fibrillation, we find that the difference in scale is not present anymore. Instead, we observe that the scale of both X and Y axes is rather similar, with a R^2 of 0.59. The outmost SNP falls around 0.2 for both X and Y variables. The same tendency applies to OFT, where there is a high correlation in CHD (0.77), but not in atrial fibrillation (0.23).

Overall, we have seen that in our human models CHD-related SNPs seems to have overall greater predicted absolute logFC when compared to atrial fibrillation, a heart-related disease. When focusing on the impact of SNP at each cell type, we saw that Endo1 and OFT were affected to a higher degree than the rest of the human cell types. This effect could not be found in atrial fibrillation, suggesting cell type and trait specific effects.

5.4. Cross-species variant effect predictions reveal trait and cell type-specific concordance

On the last section we focused only on the predicted variant effect using human models. Our main aim was to evaluate the similarities in chromatin accessibility prediction between human and mouse using deep learning models. To achieve this aim, we acquired sets of SNPs from human GWAS, obtained the mouse orthologues and predicted the variant effect of each SNP using machine-learning models. Now we can finally compare the predicted effect of the variants.

To compare the impact of such variants between both species, we first merged human and mouse cell types into "superclusters". Because both datasets were derived from fetal cardiac tissue, we also expected similar cell types. These superclusters reflect a broader group that include cell types from human and mouse origin that are similar among each other. We previously found that variant effect affects some cell types over the others depending on the trait. We now wanted to see how variant effect correlates between human and mouse across all traits. We hypothesized that, because we are dealing with cardiac tissue and following the last section, there is a specific supercluster where the correlation is greater than the rest. We wanted to explore too whether this applies to only CHD or can be extended to the rest of traits we selected. To validate this idea, we calculated the R^2 values between the human and mouse cell types absolute logFC for each trait, as we can see in Figure 8.

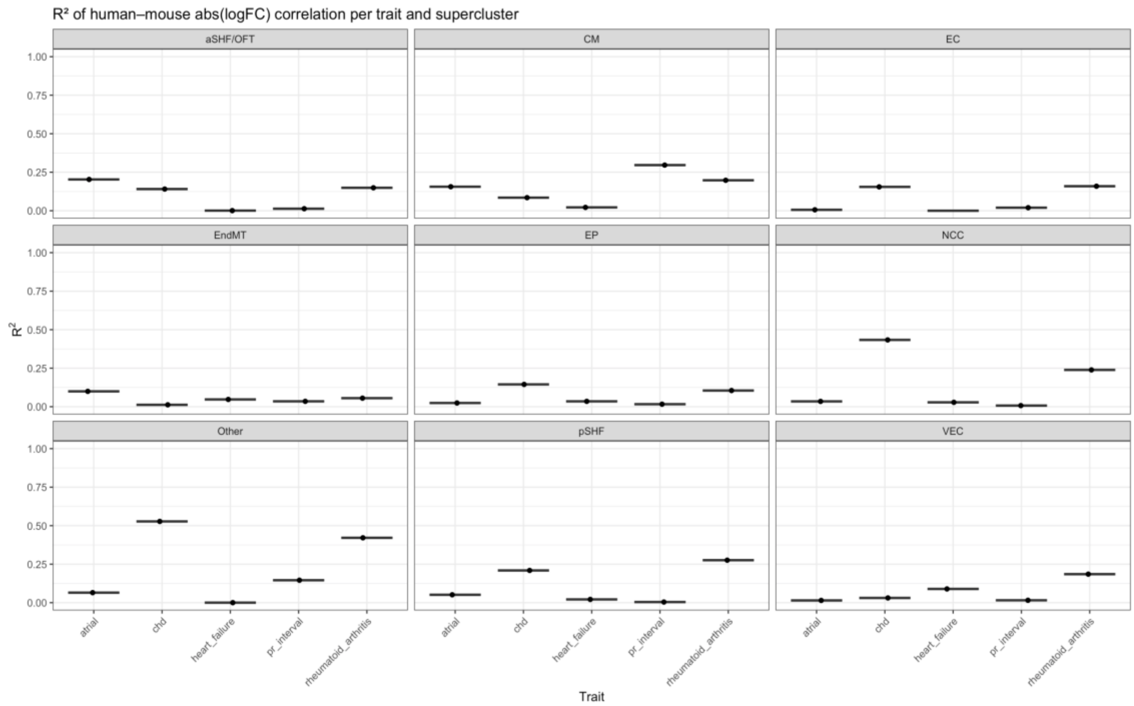


Figure 8. Nine box plots of human-mouse absolute logFC correlation per trait and supercluster. On the X axis we find all selected traits, and on the Y axis we find the R^2 values of their absolute logFC between human and mouse.

In Figure 8 we find how CHD's correlation stands out specifically in two superclusters: NCC and "Other" (which contains mouse CPM and PC), achieving around 0.5 in both. Rheumatoid arthritis surprisingly also stands out in Other (although it has a lower R^2 than CHD), pSHF and VEC. The rest of traits do not achieve such high correlations, and all of them fall below the 0.25 mark. These findings indicate that higher cross-species correlation of predicted variant effect depends on the trait and the cell identity.

Additionally, we also hypothesized cross-species correlation depends on the cell types we are comparing. For instance, correlation of very distinct cell types should be lower than correlation of similar cell types. To explore this idea, we created a per trait cross-species correlation heatmap where we compared human cell types from each supercluster against all the other mouse cell types, and vice versa.

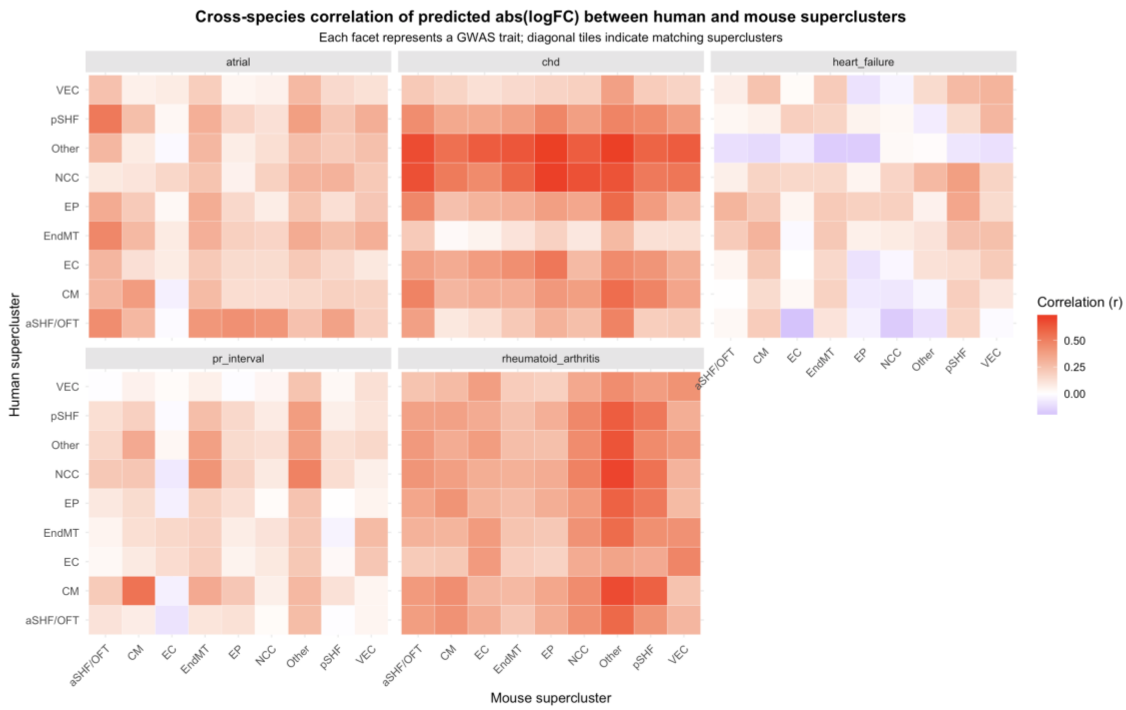


Figure 9. Five heatmaps of cross-species correlations of predicted absolute logFC between human and mouse superclusters grouped by traits. On the X axis, we find the mouse cell types corresponding to each supercluster, and on the Y axis we see the human counterparts of each supercluster. Each tile is colored by its correlation value, with blue indicating low values and red indicating high values.

First, in Figure 9 we observe that out of the 5 traits, only CHD and rheumatoid arthritis tend to have higher correlation scores. Moreover, heart failure has overall poor correlations compared to the rest. If we focus now on CHD, we find that the human cell types from CPM and NCC tend to have high correlation with the mouse counterparts of aSHF/OFT, CPM and EP. On the contrary, VEC and EndMT seem to have poor correlations overall with the mouse cell types. Looking now at the the mouse cell types in CHD, we find that aSHF/OFT and Other (which contains mouse CPM and human PC) stand out, as they have high correlations across all human cell types. Surprisingly, we find that no mouse cell type has a low correlation (colored as white or blue).

In a nutshell, cross-species variant effect prediction correlation shows to some extent cell-type- and trait- dependency. We observed CHD standing out, even though this was the trait with the lowest number of SNPs. Two human superclusters seemed to perform exceptionally well in variant effect prediction correlation, namely, NCC and the PC from the "Other" group. The high correlation for PC suggests high similarity of SNPs effect across most mouse cell types. Our findings align with a recent publication, where researches revealed a high gene regulation conservation in neural lineages across mammals, including humans and mouses (G. Zhang et al., 2024). High neural cell crest gene regulation conservation was also described in previous publications (Sauka-Spengler et al., 2007; Sauka-Spengler & Bronner-Fraser, 2008). It is also worth noting that conservation across cREs usually conserve function rather than sequence, adding an extra challenge to these comparative studies (Phan et al., 2024).

6. Discussion and conclusion

CHD affects up to 1% of the newborns. Current literature has used several models, both mouse and human, to study CHD's etiology. Researchers highlight the importance of non-coding genomic regions in CHD. With over 1 billion of known SNPs, it is unfeasible to experimentally validate each variant.

In this project, we aimed to evaluate human variant effect prediction using machine learning and compared it to their mouse orthologues.

We demonstrated how we could achieve accurate chromatin accessibility profiles prediction from sequence, in both human and mouse. We trained and tested ChromBPNet using human fetal cardiac tissue. We explored the regulatory landscapes of a set of human cell types. We also found that sequencing depth and library complexity is extremely relevant for model training, as the models with lower accuracy also had a low number of peaks called. A possible line of further research would be to test a minimal number of high-quality reads and sufficient complexity.

To study the variant effects, we selected a set of traits from human GWAS. With these variants, we obtained their mouse counterparts. Additionally, we built a dataset with the genomic context and conservation scores of each SNP. We also explored evolutionary context of each trait, finding a tendency of benign traits to have higher base conservation between human and mouse. In this section, we found an overall low percentage of matching reference alleles between human and mouse SNPs. Considering that evolution tends to conserve function rather than sequence, as mentioned in section 5.4. (Cross-species variant effect predictions reveal trait and cell-type-specific concordance), this might explain the low base conservation observed. We deem interesting to add more benign and disease-related traits to explore whether this follows a general tendency or is rather an artifact.

We also described how mouse and human seem to have different conservation scores distributions. Because phyloP scores are calculated differently to each species, it is important to pay close attention to how these scores are obtained. We believe that looking deeper into this could explain such distinct distributions. Another variable to consider is the genomic context in which SNPs are located. Their impact may rely on their distance to the regulated gene. Nonetheless, further effort is required to accurately explain such differences.

Once we acquired the SNPs for all the traits and the mouse orthologues, we finally evaluated the variant effect prediction. We described how variant effect prediction relies on factors like cell type and trait. We saw how CHD derived SNPs had a higher impact in Endo1 cells than

in any other human cell. Because our aim was a cross-species comparison, we focused on the correlation of absolute logFC in both human and mouse. Concordantly with the previous findings, we observed how correlation was only high in a few sets of cell types, and only for CHD. We also observed uncharacteristic high correlation scores in rheumatoid arthritis.
To further elucidate these findings, we should also take into consideration the sign of the absolute logFC. If we also consider whether the variant increases or decreases accessibility, we might further refine our findings. Moreover, we focused correlations, but in the future the magnitudes of change among superclusters should also be taken into account.

7. Supplementary figures

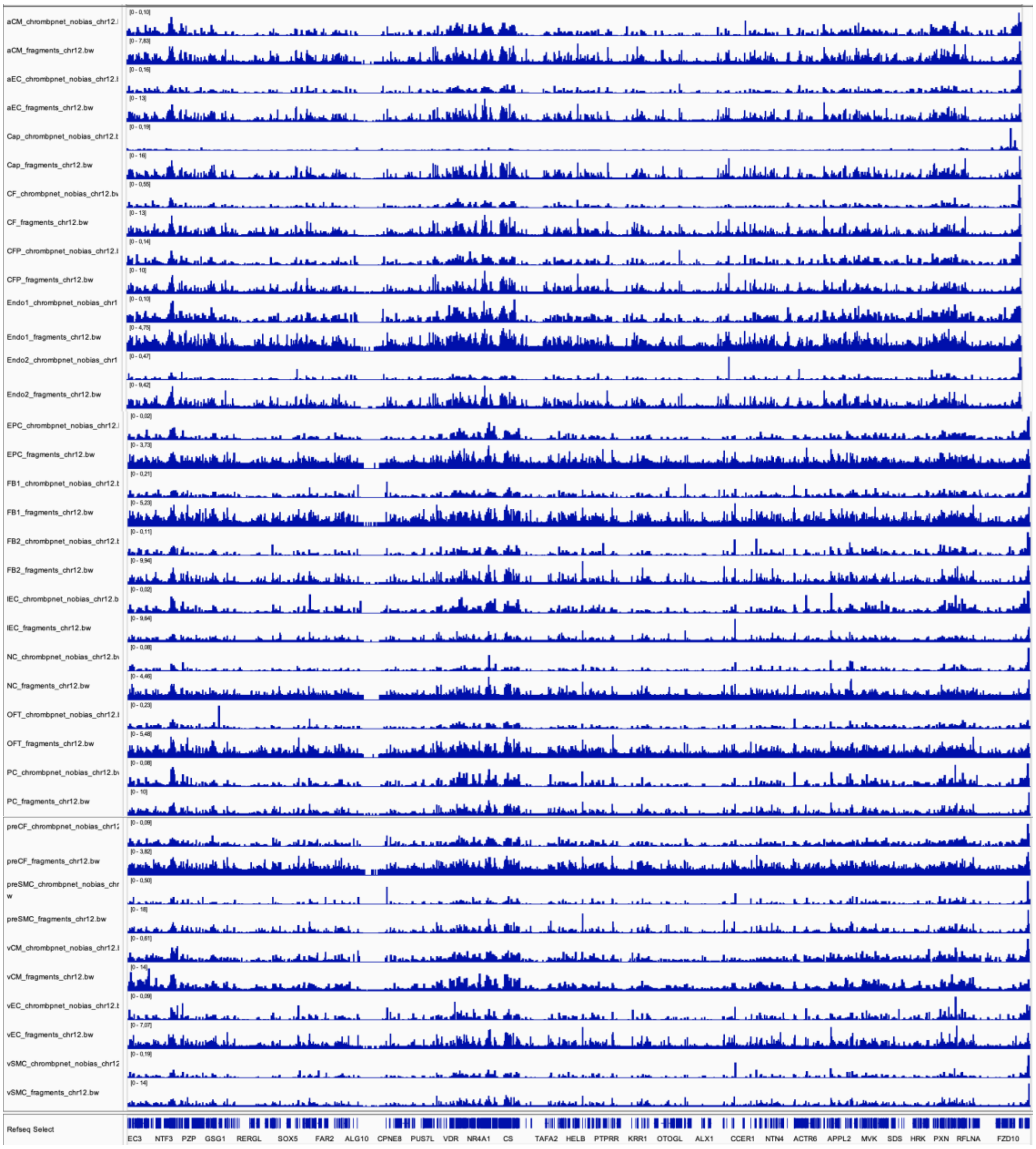
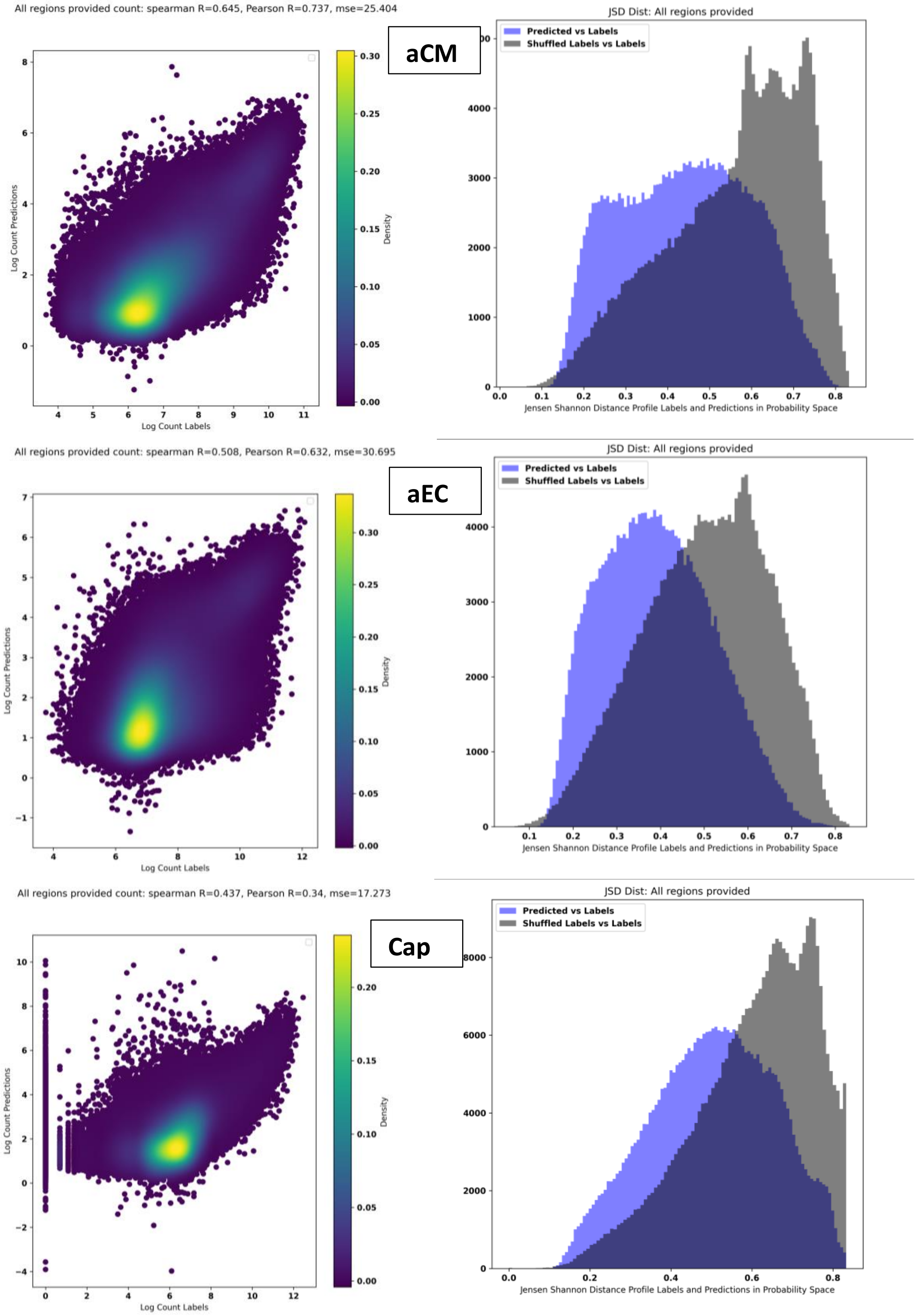
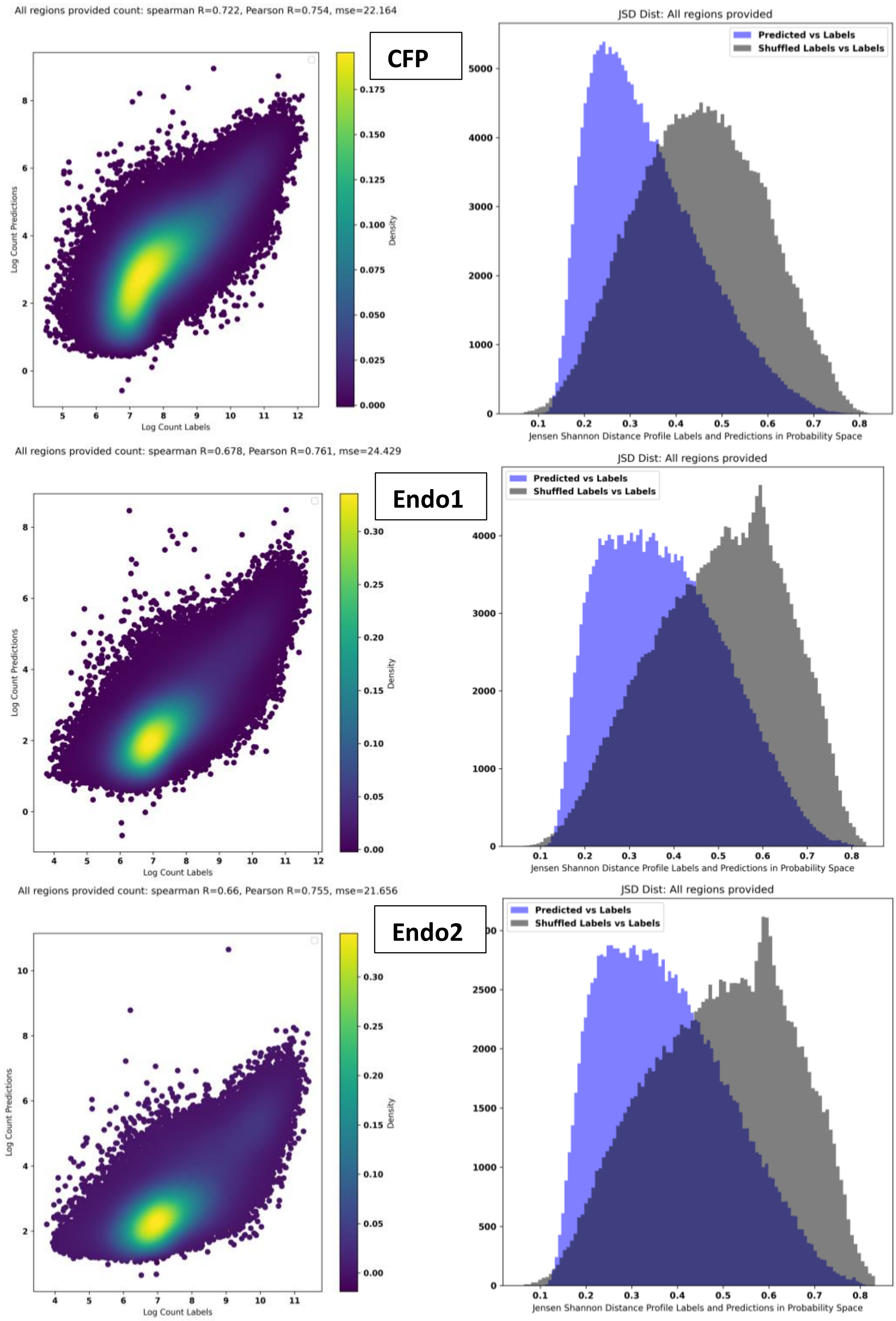
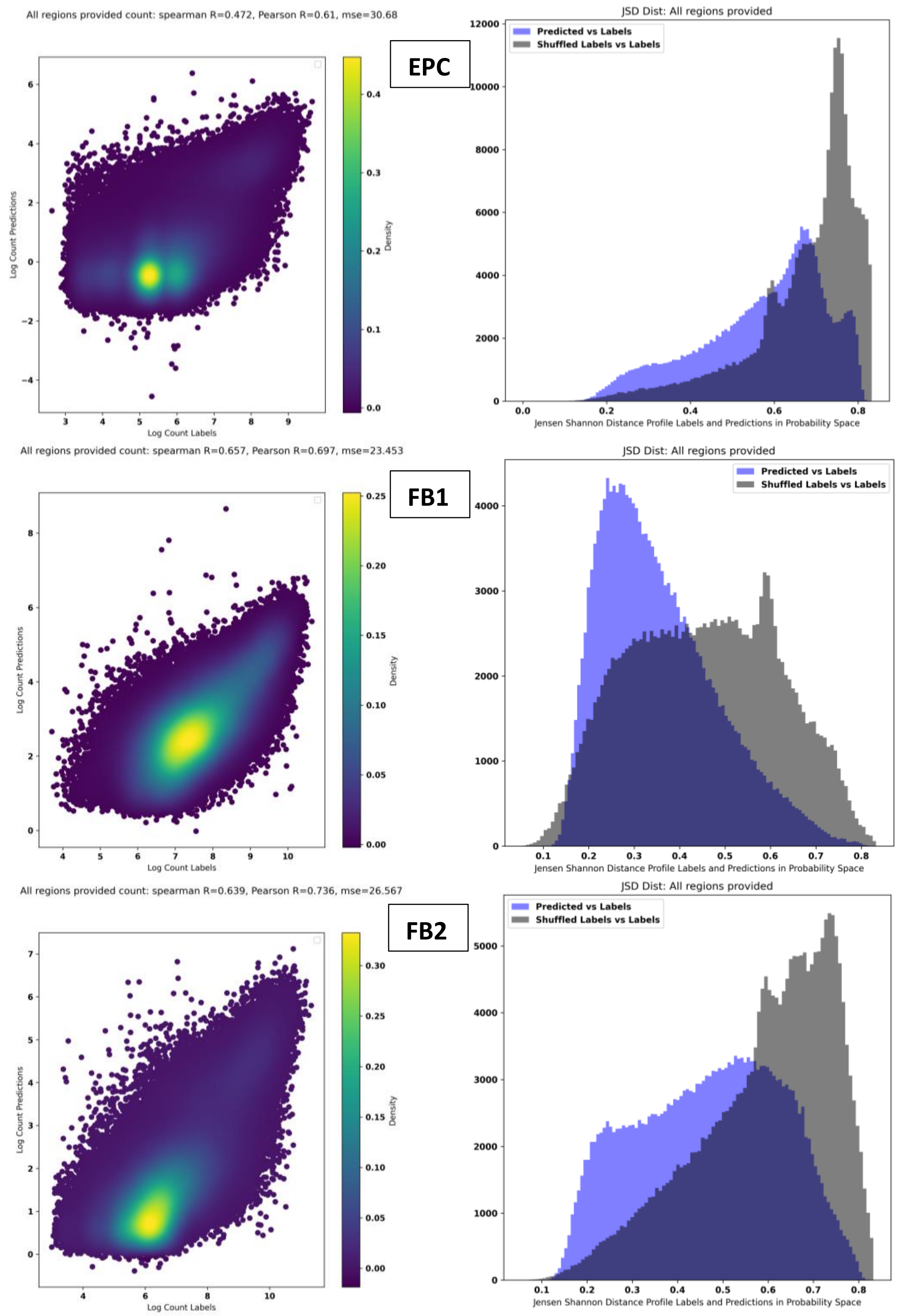
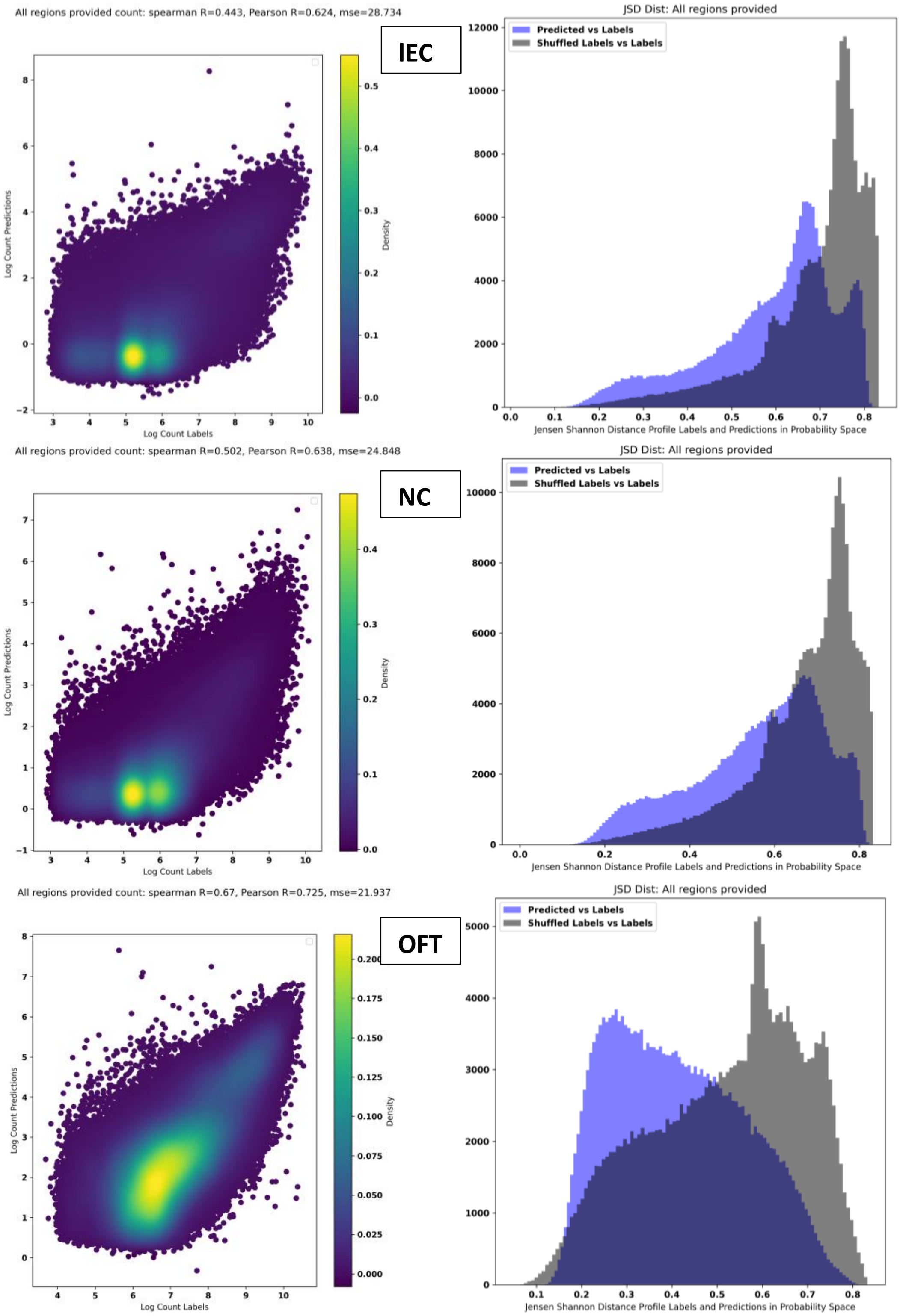


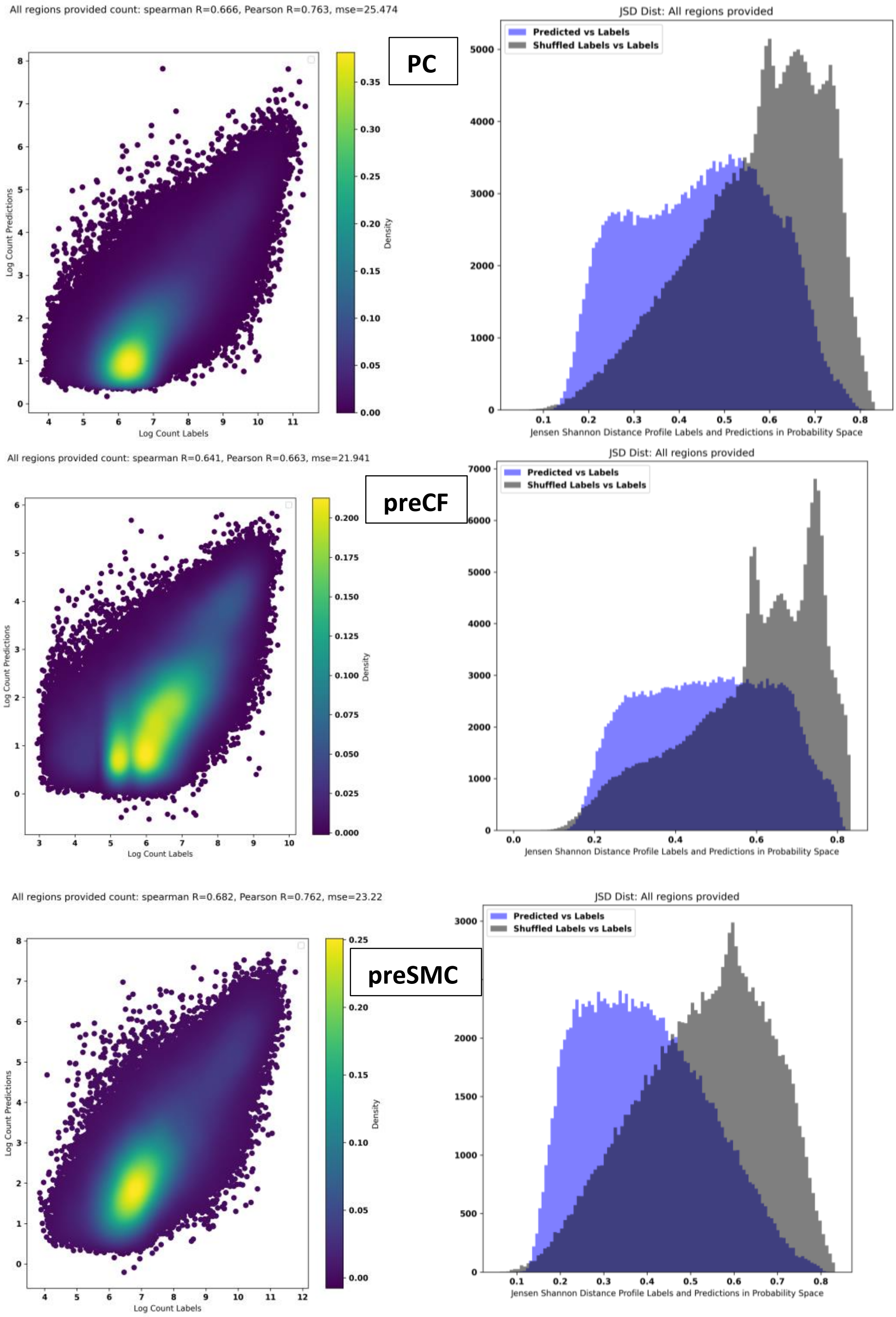
Figure S1. Genome tracks of all cell types by alphabetical order, grouped in couples. Per cell type, the top track, which ends in "chrombpnet_nobias_chr12", is predicted. The bottom track, which ends in "fragments_chr12", is the reference, namely, the raw reads from merged fragments files. Above each track between brackets, we see the scale of the bands


All regions provided count: spearman R=0.645, Pearson R=0.737, mse=25.404
aCM
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.508, Pearson R=0.632, mse=30.695
aEC
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.437, Pearson R=0.34, mse=17.273
Cap
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space


All regions provided count: spearman R=0.722, Pearson R=0.754, mse=22.164
CFP
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.678, Pearson R=0.761, mse=24.429
Endo1
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.66, Pearson R=0.755, mse=21.656
Endo2
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space


All regions provided count: spearman R=0.472, Pearson R=0.61, mse=30.68
EPC
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.657, Pearson R=0.697, mse=23.453
FB1
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.639, Pearson R=0.736, mse=26.567
FB2
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space


All regions provided count: spearman R=0.443, Pearson R=0.624, mse=28.734
IEC
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.502, Pearson R=0.638, mse=24.848
NC
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.67, Pearson R=0.725, mse=21.937
OFT
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space



All regions provided count: spearman R=0.666, Pearson R=0.763, mse=25.474
PC
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.641, Pearson R=0.663, mse=21.941
preCF
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space
All regions provided count: spearman R=0.682, Pearson R=0.762, mse=23.22
preSMC
Log Count Predictions
Log Count Labels
Density
JSD Dist: All regions provided
Predicted vs Labels
Shuffled Labels vs Labels
Jensen Shannon Distance Profile Labels and Predictions in Probability Space

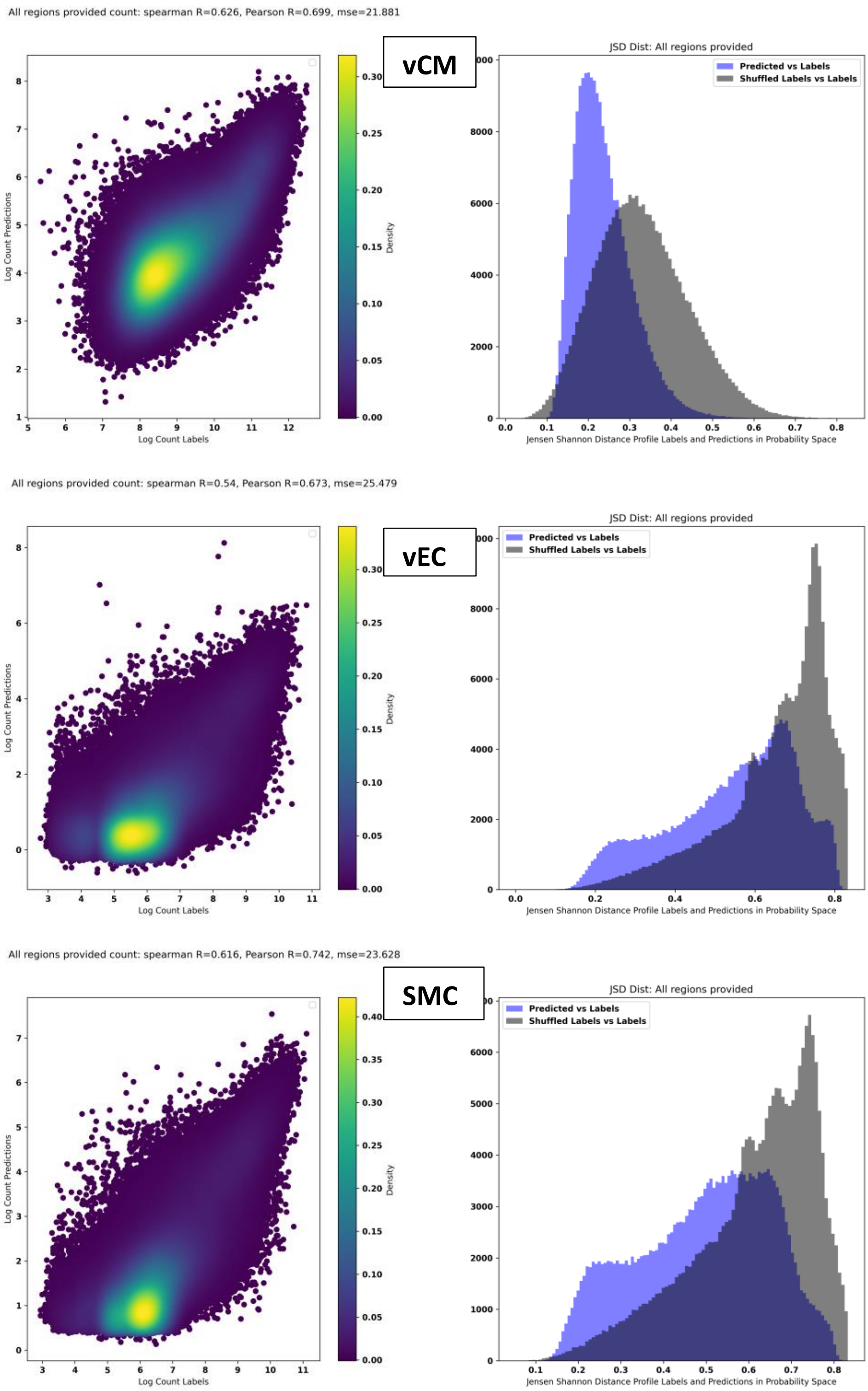
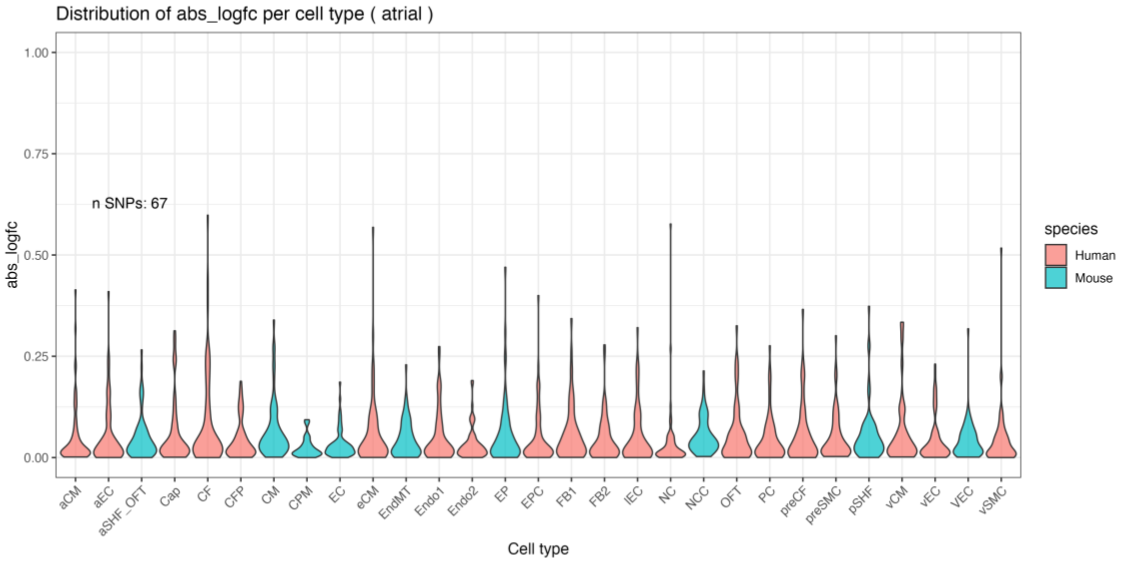
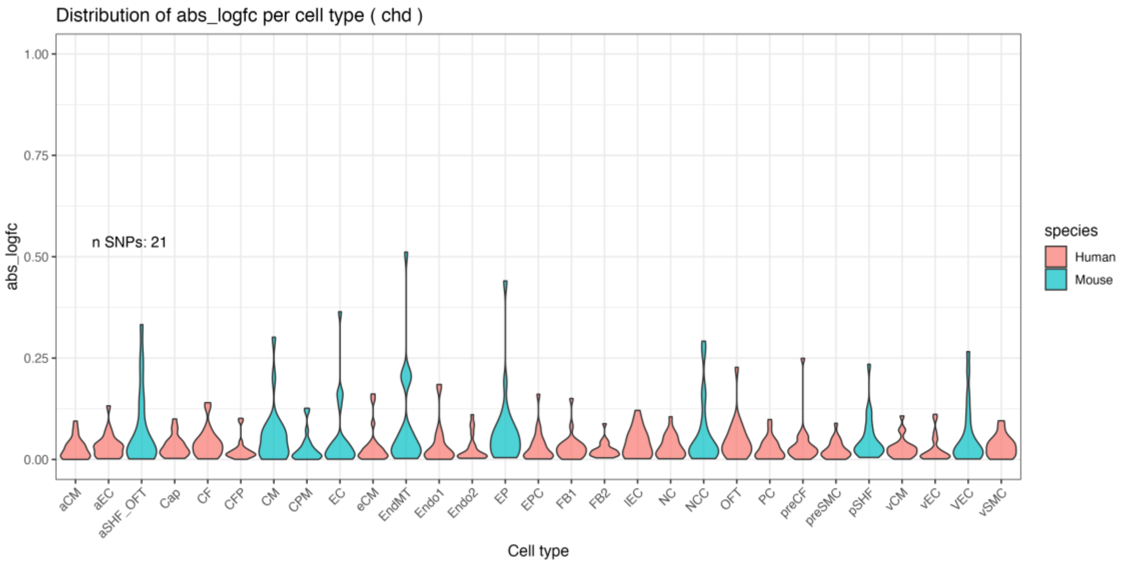
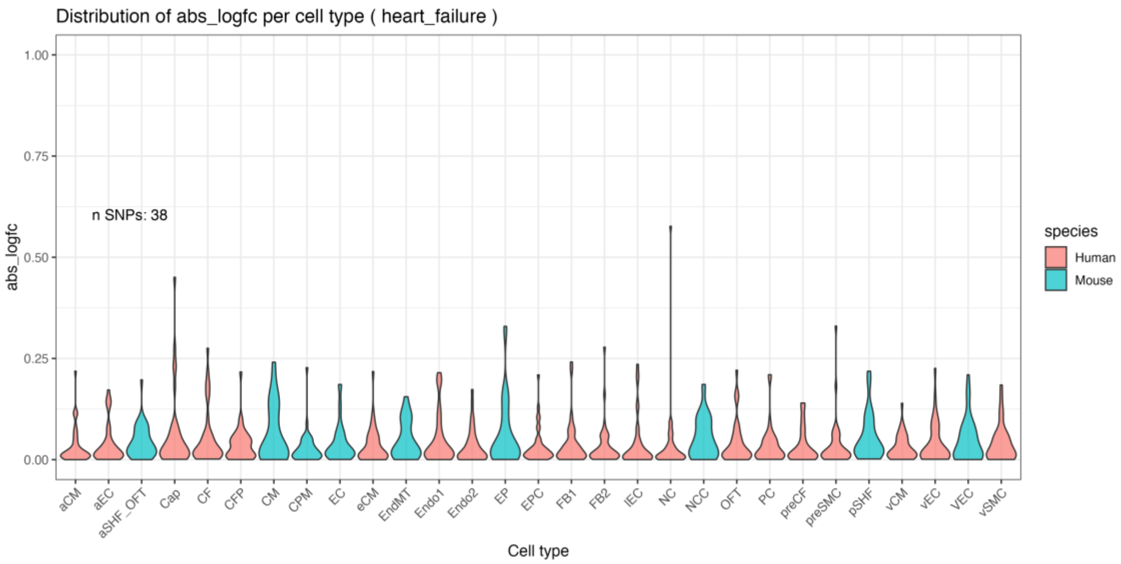


Figure S2. Plots provided by ChromBPNet's evaluation grouped by cell type (as indicated in the boxes). Left, scatter plot of log observed against log predicted counts, colored by point density where yellow is higher and blue is lower, and titled with both Spearman and Pearson correlations, as well as mean squared error (mse). Right, density plot of JSD of predicted against observed in blue and random shuffle against observed in grey.


Distribution of abs_logfc per cell type (atrial)
n SNPs: 67
abs_logfc
1.00
0.75
0.50
0.25
0.00
aCM aEC aSHF_OFT Cap CF CFP CM CPM EC eCM EndMT Endo1 Endo2 EP EPC FB1 FB2 IEC NC NCC OFT PC preCF preSMC pSHF vCM vEC VEC vSMC
Cell type
species
Human
Mouse


Distribution of abs_logfc per cell type (chd)
n SNPs: 21
abs_logfc
1.00
0.75
0.50
0.25
0.00
aCM aEC aSHF_OFT Cap CF CFP CM CPM EC eCM EndMT Endo1 Endo2 EP EPC FB1 FB2 IEC NC NCC OFT PC preCF preSMC pSHF vCM vEC VEC vSMC
Cell type
species
Human
Mouse


Distribution of abs_logfc per cell type (heart_failure)
n SNPs: 38
abs_logfc
1.00
0.75
0.50
0.25
0.00
aCM aEC aSHF_OFT Cap CF CFP CM CPM EC eCM EndMT Endo1 Endo2 EP EPC FB1 FB2 IEC NC NCC OFT PC preCF preSMC pSHF vCM vEC VEC vSMC
Cell type
species
Human
Mouse

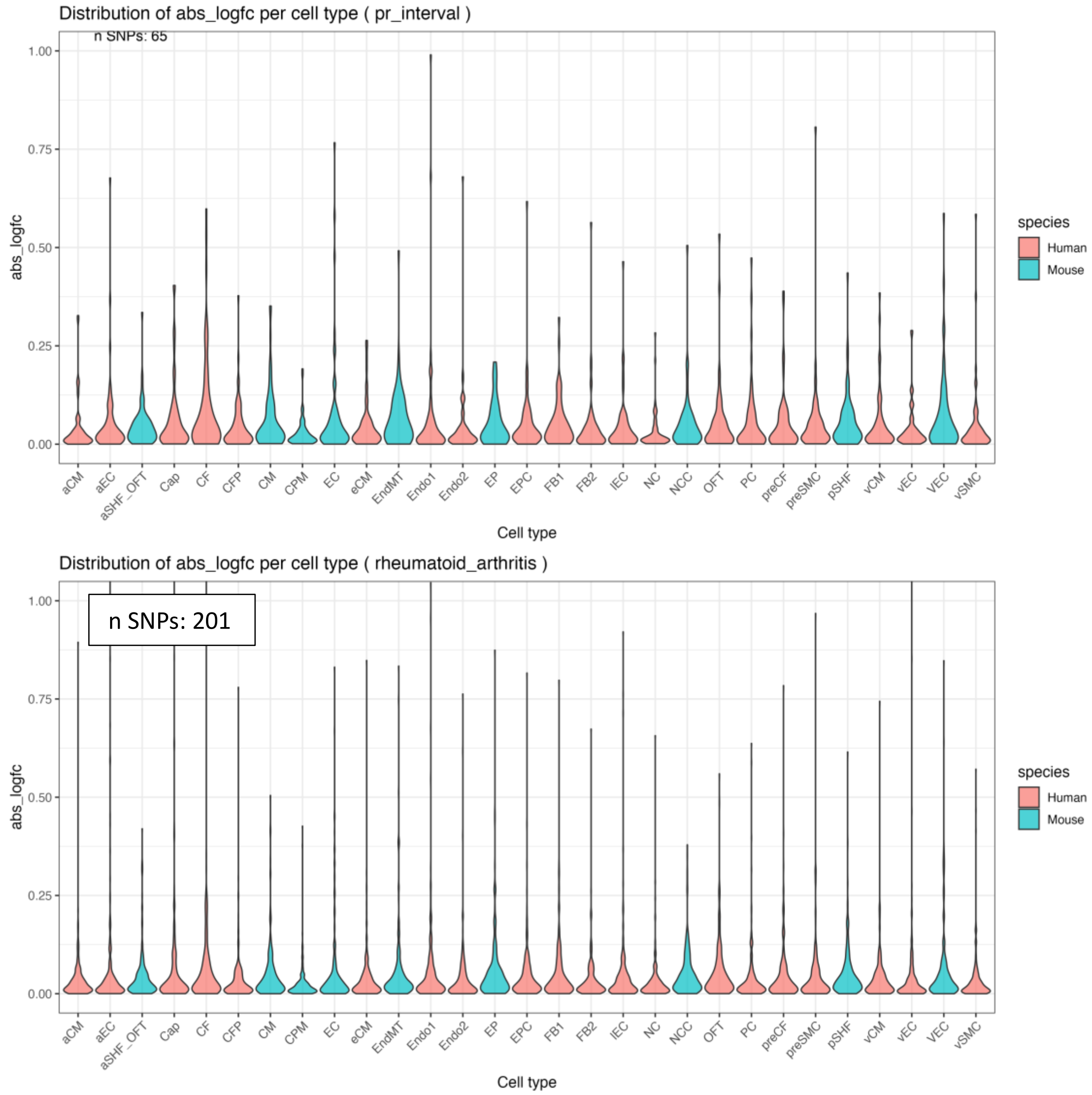
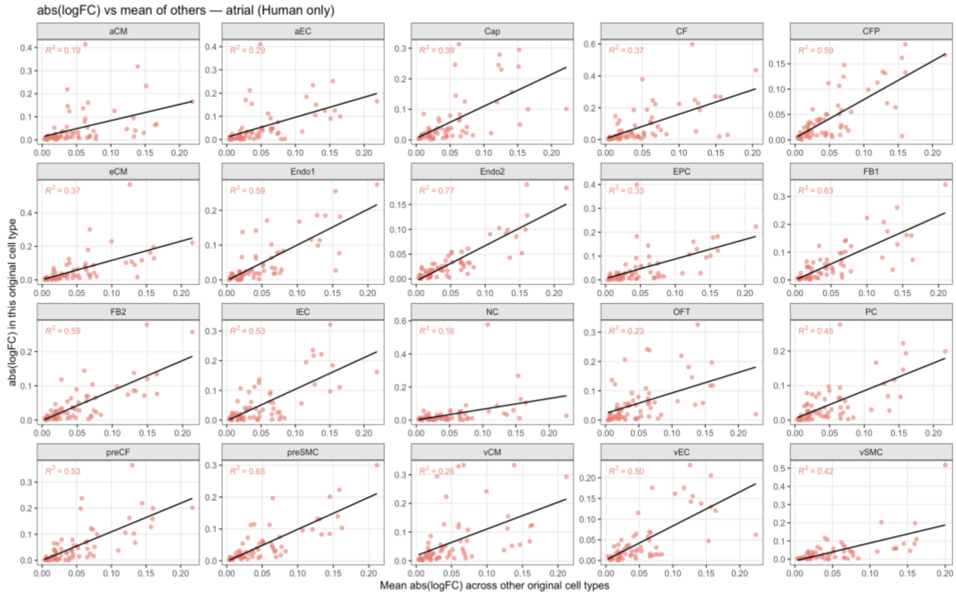
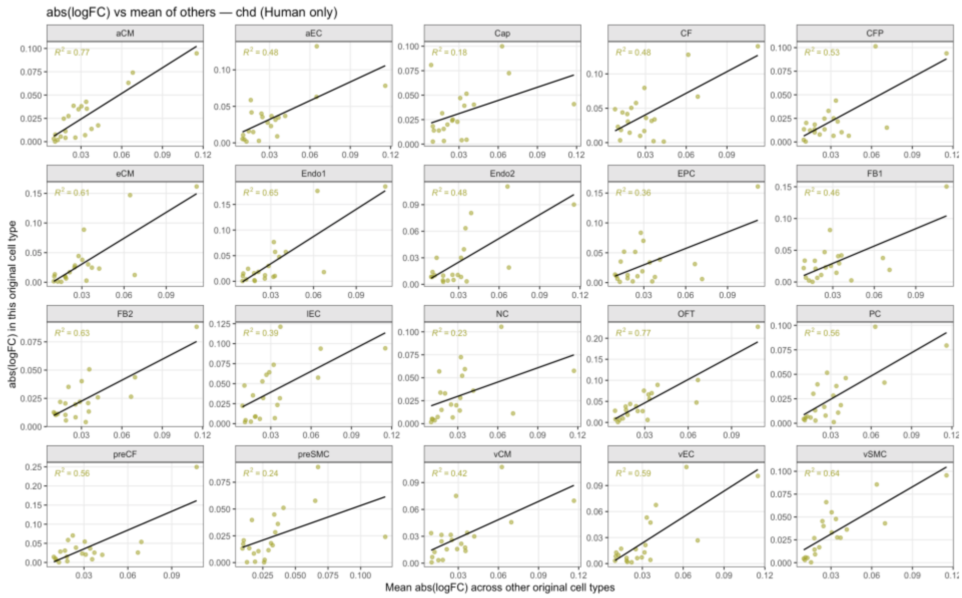


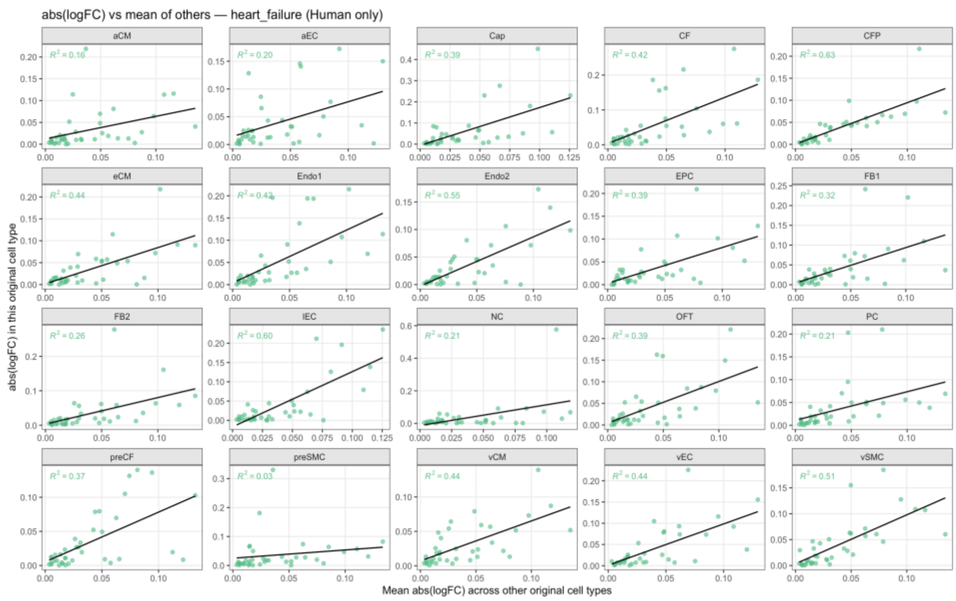
Figure S3. Violin plots of distribution of absolute logFC across all our models colored by species (human in orange, mouse in blue). Across the X axis we see the cell types on which the models were trained; on the Y axis we find the absolute logFC.In each plot we also find the number of SNPs per trait.



abs(logFC) vs mean of others — atrial (Human only)
aCM $R^2 = 0.19$
aEC $R^2 = 0.29$
Cap $R^2 = 0.39$
CF $R^2 = 0.37$
CFP $R^2 = 0.59$
eCM $R^2 = 0.37$
Endo1 $R^2 = 0.59$
Endo2 $R^2 = 0.77$
EPC $R^2 = 0.33$
FB1 $R^2 = 0.63$
FB2 $R^2 = 0.59$
IEC $R^2 = 0.53$
NC $R^2 = 0.16$
OFT $R^2 = 0.23$
PC $R^2 = 0.45$
preCF $R^2 = 0.53$
preSMC $R^2 = 0.65$
vCM $R^2 = 0.26$
vEC $R^2 = 0.50$
vSMC $R^2 = 0.42$
abs(logFC) in this original cell type
Mean abs(logFC) across other original cell types



abs(logFC) vs mean of others — chd (Human only)
aCM $R^2 = 0.77$
aEC $R^2 = 0.48$
Cap $R^2 = 0.18$
CF $R^2 = 0.48$
CFP $R^2 = 0.53$
eCM $R^2 = 0.61$
Endo1 $R^2 = 0.65$
Endo2 $R^2 = 0.48$
EPC $R^2 = 0.36$
FB1 $R^2 = 0.46$
FB2 $R^2 = 0.63$
IEC $R^2 = 0.39$
NC $R^2 = 0.23$
OFT $R^2 = 0.77$
PC $R^2 = 0.56$
preCF $R^2 = 0.56$
preSMC $R^2 = 0.24$
vCM $R^2 = 0.42$
vEC $R^2 = 0.59$
vSMC $R^2 = 0.64$
abs(logFC) in this original cell type
Mean abs(logFC) across other original cell types



abs(logFC) vs mean of others — heart_failure (Human only)
aCM $R^2 = 0.16$
aEC $R^2 = 0.20$
Cap $R^2 = 0.39$
CF $R^2 = 0.42$
CFP $R^2 = 0.63$
eCM $R^2 = 0.44$
Endo1 $R^2 = 0.42$
Endo2 $R^2 = 0.55$
EPC $R^2 = 0.39$
FB1 $R^2 = 0.32$
FB2 $R^2 = 0.26$
IEC $R^2 = 0.60$
NC $R^2 = 0.21$
OFT $R^2 = 0.39$
PC $R^2 = 0.21$
preCF $R^2 = 0.37$
preSMC $R^2 = 0.03$
vCM $R^2 = 0.44$
vEC $R^2 = 0.44$
vSMC $R^2 = 0.51$
abs(logFC) in this original cell type
Mean abs(logFC) across other original cell types

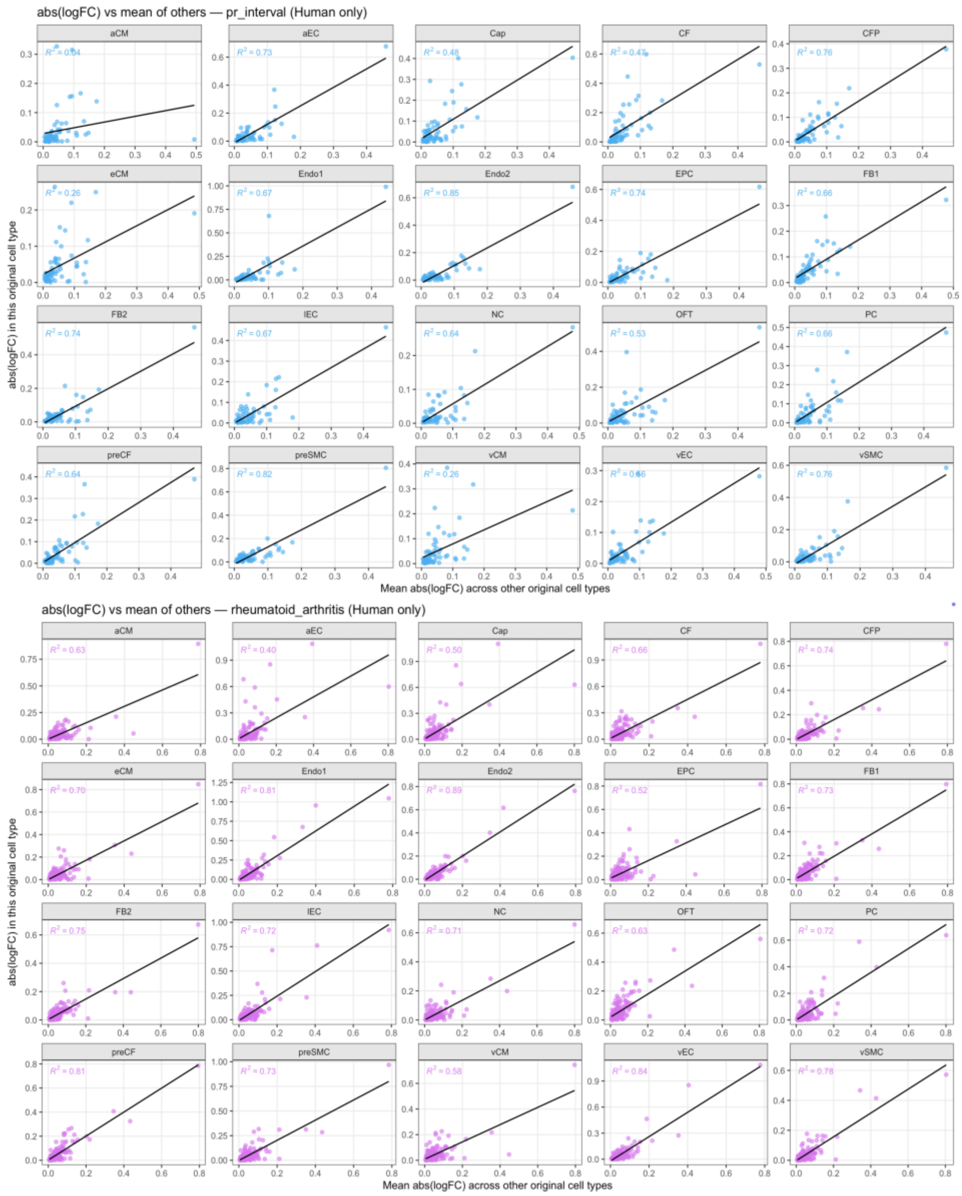


Figure S4. Scatter plots of absolute logFC per cell type (named on the header; Y axis) against mean absolute logFC of the rest of the cell types (X axis). We can also find the R^2 of each plot and the regression line.

8. References

GenAI statement: Generative AI has been used in this report to propose general ideas of improvement, like correcting grammar mistakes, conciseness or clarity of the writing.

Aissa, A. F., Islam, A. B. M. M. K., Ariss, M. M., Go, C. C., Rader, A. E., Conrardy, R. D., Gajda, A. M., Rubio-Perez, C., Valyi-Nagy, K., Pasquinelli, M., Feldman, L. E., Green, S. J., Lopez-Bigas, N., Frolov, M. V., & Benevolenskaya, E. V. (2021). Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature Communications*, *12*(1), 1628. https://doi.org/10.1038/s41467-021-21884-z

Ameen, M., Kundaje, A., & Quertermous, T. (2022). Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and noncoding mutations in congenital heart disease. *Cell*, *185*, 4937–4953. https://doi.org/10.1016/j.cell.2022.11.028

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, *9*(1), 9354. https://doi.org/10.1038/s41598-019-45839-z

Arthur, H. M., & Bamforth, S. D. (2011). TGFβ signaling and congenital heart disease: Insights from mouse studies. *Birth Defects Research Part A: Clinical and Molecular Teratology*, *91*(6), 423–434. https://doi.org/10.1002/bdra.20794

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, *109*(1). https://doi.org/10.1002/0471142727.mb2129s109

Cheng, J., Lawrence, C., & Niepert, M. (2021). *VEGN: Variant Effect Prediction with Graph Neural Networks* (No. arXiv:2106.13642). arXiv. https://doi.org/10.48550/arXiv.2106.13642

Doyle, M. J., Lohr, J. L., Chapman, C. S., Koyano-Nakagawa, N., Garry, M. G., & Garry, D. J. (2015). Human Induced Pluripotent Stem Cell-Derived Cardiomyocytes as a Model for Heart Development and Congenital Heart Disease. *Stem Cell Reviews and Reports*, *11*(5), 710–727. https://doi.org/10.1007/s12015-015-9596-6

Dutrow, E. V., Emera, D., Yim, K., Uebbing, S., Kocher, A. A., Krenzer, M., Nottoli, T., Burkhardt, D. B., Krishnaswamy, S., Louvi, A., & Noonan, J. P. (2022). Modeling uniquely human gene regulatory function via targeted humanization of the mouse genome. *Nature Communications*, *13*(1), 304. https://doi.org/10.1038/s41467-021-27899-w

Fawkner-Corbett, D., Antanaviciute, A., Parikh, K., Jagielowicz, M., Gerós, A. S., Gupta, T., Ashley, N., Khamis, D., Fowler, D., Morrissey, E., Cunningham, C., Johnson, P. R. V., Koohy, H., & Simmons, A. (2021). Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell*, *184*(3), 810-826.e23. https://doi.org/10.1016/j.cell.2020.12.016

Grandi, F. C., Modi, H., Kampman, L., & Corces, M. R. (2022). Chromatin accessibility profiling by ATAC-seq. *Nature Protocols*, *17*(6), 1518–1552. https://doi.org/10.1038/s41596-022-00692-9

Hocker, J. D., Poirion, O. B., Zhu, F., Buchanan, J., Zhang, K., Chiou, J., Wang, T.-M., Zhang, Q., Hou, X., Li, Y. E., Zhang, Y., Farah, E. N., Wang, A., McCulloch, A. D., Gaulton, K. J., Ren, B., Chi, N. C., & Preissl, S. (2021). Cardiac cell type–specific gene regulatory programs and disease risk association. *SCIENCE ADVANCES*.

Hofbauer, P., Jahnel, S. M., Papai, N., Giesshammer, M., Deyett, A., Schmidt, C., Penc, M., Tavernini, K., Grdseloff, N., Meledeth, C., Ginistrelli, L. C., Ctortecka, C., Šalic, Š., Novatchkova, M., & Mendjan, S. (2021). Cardioids reveal self-organizing principles of human cardiogenesis. *Cell*, *184*(12), 3299-3317.e22. https://doi.org/10.1016/j.cell.2021.04.034

Hollingsworth, E. W., Liu, T. A., Alcantara, J. A., Chen, C. X., Jacinto, S. H., & Kvon, E. Z. (2025). Rapid and quantitative functional interrogation of human enhancer variant activity in live mice. *Nature Communications*, *16*(1), 409. https://doi.org/10.1038/s41467-024-55500-7

Kim, S., & Wysocka, J. (2023). Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell*, *83*(3), 373–392. https://doi.org/10.1016/j.molcel.2022.12.032

Krishnan, A., Samtani, R., Dhanantwari, P., Lee, E., Yamada, S., Shiota, K., Donofrio, M. T., Leatherbury, L., & Lo, C. W. (2014). A detailed comparison of mouse and human cardiac development. *Pediatric Research*, *76*(6), 500–507. https://doi.org/10.1038/pr.2014.128

Li, Z., Zhang, Y., Peng, B., Qin, S., Zhang, Q., Chen, Y., Chen, C., Bao, Y., Zhu, Y., Hong, Y., Liu, B., Liu, Q., Xu, L., Chen, X., Ma, X., Wang, H., Xie, L., Yao, Y., Deng, B., … Liu, Y. (2024). A novel interpretable deep learning-based computational framework designed synthetic enhancers with broad cross-species activity. *Nucleic Acids Research*, *52*(21), 13447–13468. https://doi.org/10.1093/nar/gkae912

Liang, Y., Kaneko, K., Xin, B., Lee, J., Sun, X., Zhang, K., & Feng, G.-S. (2022). Temporal analyses of postnatal liver development and maturation by single-cell transcriptomics. *Developmental Cell*, *57*(3), 398-414.e5. https://doi.org/10.1016/j.devcel.2022.01.004

Liu, C., & Shao, N.-Y. (2024). The Differences in the Developmental Stages of the Cardiomyocytes and Endothelial Cells in Human and Mouse Embryos at the Single-Cell Level. *International Journal of Molecular Sciences*, *25*(6), 3240. https://doi.org/10.3390/ijms25063240

Ma, J., Chen, S., Hao, L., Sheng, W., Chen, W., Ma, X., Zhang, B., Ma, D., & Huang, G. (2021). Long non-coding RNA SAP30-2:1 is downregulated in congenital heart disease and regulates cell proliferation by targeting HAND2. *Frontiers of Medicine*, *15*(1), 91–100. https://doi.org/10.1007/s11684-020-0778-5

Ma, S., & Zhang, Y. (2020). Profiling chromatin regulatory landscape: Insights into the development of ChIP-seq and ATAC-seq. *Molecular Biomedicine*, *1*(1), 9. https://doi.org/10.1186/s43556-020-00009-w

Majumdar, U., Yasuhara, J., & Garg, V. (2021). In Vivo and In Vitro Genetic Models of Congenital Heart Disease. *Cold Spring Harbor Perspectives in Biology*, *13*(4), a036764. https://doi.org/10.1101/cshperspect.a036764

Meller, C., Grinenco, S., Aiello, H., Córdoba, A., Sáenz-Tejeira, M., Marantz, P., & Otaño, L. (2020). Congenital heart disease, prenatal diagnosis and management. *Archivos Argentinos de Pediatria*, *118*(2). https://doi.org/10.5546/aap.2020.eng.e149

Pampari, A., Shcherbina, A., Kvon, E. Z., Kosicki, M., Nair, S., Kundu, S., Kathiria, A. S., Risca, V. I., Kuningas, K., Alasoo, K., Greenleaf, W. J., Pennacchio, L. A., & Kundaje, A. (2024). *ChromBPNet: Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants*. Genetics. https://doi.org/10.1101/2024.12.25.630221

Phan, M. H., Zehnder, T. M., Puntieri, F., Lo, B.-W., Lenhard, B., Mueller, F., Vingron, M., & Ibrahim, D. M. (2024). *Conservation of Regulatory Elements with Highly Diverged Sequences Across Large Evolutionary Distances*. Genetics. https://doi.org/10.1101/2024.05.13.590087

Pierpont, M. E., Brueckner, M., Chung, W. K., Garg, V., Lacro, R. V., McGuire, A. L., Mital, S., Priest, J. R., Pu, W. T., Roberts, A., Ware, S. M., Gelb, B. D., Russell, M. W., & On behalf of the American Heart Association Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; and Council on Genomic and Precision Medicine. (2018). Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association. *Circulation*, *138*(21). https://doi.org/10.1161/CIR.0000000000000606

Sauka-Spengler, T., & Bronner-Fraser, M. (2008). A gene regulatory network orchestrates neural crest formation. *Nature Reviews Molecular Cell Biology*, *9*(7), 557–568. https://doi.org/10.1038/nrm2428

Sauka-Spengler, T., Meulemans, D., Jones, M., & Bronner-Fraser, M. (2007). Ancient Evolutionary Origin of the Neural Crest Gene Regulatory Network. *Developmental Cell*, *13*(3), 405–420. https://doi.org/10.1016/j.devcel.2007.08.005

Schmidt, C., Deyett, A., Ilmer, T., Haendeler, S., Torres Caballero, A., Novatchkova, M., Netzer, M. A., Ceci Ginistrelli, L., Mancheno Juncosa, E., Bhattacharya, T., Mujadzic, A., Pimpale, L., Jahnel, S. M., Cirigliano, M., Reumann, D., Tavernini, K., Papai, N., Hering, S., Hofbauer, P., & Mendjan, S. (2023). Multi-chamber cardioids unravel human heart development and cardiac defects. *Cell*, *186*(25), 5587-5605.e27. https://doi.org/10.1016/j.cell.2023.10.030

Song, Y., Chen, W., Huang, Z., Tian, G., Li, M., Zhao, Z., Feng, Z., Wu, F., Qian, M., Ma, X., Sheng, W., & Huang, G. (2021). A Non-coding HES1 Variant Predisposes Children to Congenital Heart Disease in Chinese Population. *Frontiers in Cell and Developmental Biology*, *9*, 631942. https://doi.org/10.3389/fcell.2021.631942

Tian, H., Liu, H., Zhu, Y., Xing, D., & Wang, B. (2020). The Trends of Single-Cell Analysis: A Global Study. *BioMed Research International*, *2020*(1), 7425397. https://doi.org/10.1155/2020/7425397

Wang, G., Wang, B., & Yang, P. (2022). Epigenetics in Congenital Heart Disease. *Journal of the American Heart Association*, *11*(7), e025163. https://doi.org/10.1161/JAHA.121.025163

Zhang, G., Fu, Y., Yang, L., Ye, F., Zhang, P., Zhang, S., Ma, L., Li, J., Wu, H., Han, X., Wang, J., & Guo, G. (2024). Construction of single-cell cross-species chromatin accessibility landscapes with combinatorial-hybridization-based ATAC-seq. *Developmental Cell*, *59*(6), 793-811.e8. https://doi.org/10.1016/j.devcel.2024.01.015

Zhang, K., Hocker, J. D., Miller, M., Hou, X., Chiou, J., Poirion, O. B., Qiu, Y., Li, Y. E., Gaulton, K. J., Wang, A., Preissl, S., & Ren, B. (2021). A single-cell atlas of chromatin accessibility in the human genome. *Cell*, *184*(24), 5985-6001.e19. https://doi.org/10.1016/j.cell.2021.10.024

Zhao, Y., Deng, W., Wang, Z., Wang, Y., Zheng, H., Zhou, K., Xu, Q., Bai, L., Liu, H., Ren, Z., & Jiang, Z. (2024). Genetics of congenital heart disease. *Clinica Chimica Acta*, *552*, 117683. https://doi.org/10.1016/j.cca.2023.117683

Zhu, W., W. Lo, C., 香港中文大学心血管基因组医学中心, 香港, 中国, 匹茨堡大学发育生物学系, 宾夕法尼亚 匹茨堡15201, 美国, Kunming Institute of Zoology-The Chinese University of Hong Kong (KIZ-CUHK) Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Hong Kong SAR, China, The Chinese University of Hong Kong, Hong Kong SAR, China, Department of Developmental Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, 15201 USA, & Kunming Institute of Zoology-The Chinese University of Hong Kong (KIZ-CUHK) Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Hong Kong SAR, China. (2023). Insights into the genetic architecture of congenital heart disease from animal modeling. *Zoological Research*, *44*(3), 577–590. https://doi.org/10.24272/j.issn.2095-8137.2022.463