From the Institute of Human genetics
of the University of Lübeck
Director: Prof. Dr. Malte Spielmann

“Single cell sequencing in development and disease”

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Natural Sciences

Submitted by

Saranya Balachandran
from Coimbatore
Lübeck 2024.

First referee: Prof. Dr. Malte Spielmann

Second referee: Prof. Dr. Tanja Zeller

Date of oral examination: May 16, 2025

Approved for printing. Lübeck, May 20, 2025

Declaration of Independence

I herewith certify that this thesis is my own work and I have formulated and written this thesis independently and I have documented all the sources used. Moreover, I have not applied for an examination procedure anywhere else nor submitted the dissertation in this or any other form to any other faculty.

Table of Content

Summary

To understand congenital diseases, it's essential to first understand the normal development of a wild-type organism. With advancements in clinical sequencing methods, several variants in key genes have been identified and characterized as potential candidate disease genes. However there are genes that have been uncharacterized thus classifying the variants as variants of unknown significance and the patients do not get a diagnosis. Techniques like in situ hybridisation have been used to evaluate the location of expression of genes in model organisms. This requires understanding of the similarities and differences between model organisms and humans. Single-cell sequencing (scRNA-seq) enabled cellular-level resolution, providing insights into the organism under study.

In this thesis, I have focused on two key aspects of the use of scRNA-seq in understanding development and disease. First, we analyzed the similarities and differences between the model organisms and humans. Second, we developed a machine learning model that captures tissue-specific development and prioritizes genes critical for the development of specific tissues.

In the first project, publicly available scRNA-seq datasets from the hearts of humans, macaques (*Macaca fascicularis*), mice, and zebrafish were integrated to analyze cross-species cell type-specific differences. This comparative approach highlighted gene expression differences at the cellular level, providing insights into species-specific variations in heart development and function. Notably, these expressional differences align with known physiological and structural distinctions between species.The findings emphasize the importance of thoroughly understanding these species-specific differences when selecting an appropriate model organism for studying human diseases.

In the second project, I developed STIGMA (*Single-cell Tissue-specific Gene prioritization using Machine Learning*), a computational framework designed to prioritize candidate disease genes using tissue-specific scRNA-seq datasets sequenced across

developmental timelines. STIGMA identifies candidate genes by evaluating the temporal dynamics of gene expression across cell types and comparing them to a known disease gene panel, leveraging their expressional similarity. We applied this method to a human fetal heart scRNA-seq dataset and a mouse limb scRNA-seq dataset. The prioritized genes demonstrated strong evidence, as several were found to be mutated in patients with similar phenotypes. Additionally, mouse model studies supported these findings, showing phenotypic abnormalities associated with the mutated genes. Thus it shows the ability of STIGMA to uncover candidate disease genes in congenital diseases and shows how this approach can be extended to other diseases.

Zusammenfassung

Um angeborene Krankheiten zu verstehen, ist es wichtig, zunächst die normale Entwicklung eines Wildtyp Organismus zu verstehen. Mit Fortschritten bei klinischen Sequenzierungsmethoden wurden mehrere Varianten in Schlüsselgenen identifiziert und als potenzielle Kandidaten für Krankheitsgene charakterisiert. Es gibt jedoch Gene, die nicht charakterisiert wurden, sodass die Varianten als Varianten von unbekannter Bedeutung klassifiziert werden und die Patienten keine Diagnose erhalten. Techniken wie In-situ-Hybridisierung wurden verwendet, um diese Varianten in Modellorganismen zu bewerten. Dies erfordert ein Verständnis der Ähnlichkeiten und Unterschiede zwischen Modellorganismen und Menschen. Die Einzelzell Sequenzierung (scRNA-seq) ermöglichte eine Auflösung auf Zellebene und lieferte Einblicke in den untersuchten Organismus.

In dieser Arbeit habe ich mich auf zwei Schlüsselaspekte der Verwendung von scRNA-seq zum Verständnis von Entwicklung und Krankheit konzentriert. Erstens haben wir die Ähnlichkeiten und Unterschiede zwischen den Modellorganismen und dem Menschen analysiert. Zweitens haben wir ein maschinelles Lernmodell entwickelt, das die gewebespezifische Entwicklung erfasst und Gene priorisiert, die für die Entwicklung bestimmter Gewebe entscheidend sind.

Im ersten Projekt wurden öffentlich verfügbare scRNA-Sequenz Datensätze aus den Herzen von Menschen, Makaken (*Macaca fascicularis*), Mäusen und Zebrafischen integriert, um artenübergreifende zelltyp spezifische Unterschiede zu analysieren. Dieser vergleichende Ansatz hob Unterschiede in der Genexpression auf zellulärer Ebene hervor und lieferte Einblicke in artspezifische Variationen in der Herzentwicklung und -funktion. Insbesondere stimmen diese Expression-Unterschiede mit bekannten physiologischen und strukturellen Unterscheidungen zwischen Arten überein. Die Ergebnisse unterstreichen die Bedeutung eines gründlichen Verständnisses dieser artspezifischen Unterschiede bei der Auswahl eines geeigneten Modellorganismus zur Untersuchung menschlicher Krankheiten.

Im zweiten Projekt entwickelte ich STIGMA (Single-cell Tissue-specific Gene prioritization using Machine Learning), ein Computer Rahmenwerk zur Priorisierung von Krankheits Kandidatengenen unter Verwendung gewebespezifischer scRNA-Sequenz Datensätze, die über Entwicklungsleitlinien hinweg sequenziert wurden. STIGMA identifiziert Kandidatengene, indem sie die zeitliche Dynamik der Genexpression über Zelltypen hinweg auswertet und sie mit einem bekannten Krankheitsgenpanel vergleicht, wobei ihre Expression Ähnlichkeit hat. Wir haben diese Methode auf einen scRNA-Sequenz Datensatz des menschlichen fetalen Herzens und einen scRNA-Sequenz Datensatz der Gliedmaßen einer Maus angewendet. Die priorisierten Gene zeigten starke Hinweise auf Relevanz, da mehrere davon bei Patienten mit ähnlichen Phänotypen mutiert waren. Darüber hinaus stützen Studien an Mausmodellen diese Ergebnisse und zeigten phänotypische Anomalien im Zusammenhang mit den mutierten Genen. Dies zeigt die Fähigkeit von STIGMA, Kandidatengene für angeborene Krankheiten aufzudecken, und zeigt, wie dieser Ansatz auf andere Krankheiten ausgeweitet werden kann.

1. Introduction

With the advent of Next generation sequencing technologies (NGS) the study of the genome and genetic variations has gained popularity (Shendure and Ji 2008). Deciphering the sequence of nucleotides before the era of bioinformatics was a highly tedious and time-consuming process (Larson et al. 2023; Mardis 2011). Without computational tools to automate analysis, organizing, interpreting sequence data, storing and sharing data, results acquisition was slow and prone to errors (Yanai and Chmielnicki 2017). The advent of bioinformatics revolutionized this process, enabling rapid and large-scale sequencing (Marx 2013).

With the emerging sequencing methods, new tools are developed to analyze the data and make biological predictions with the data (Bansal and Boucher 2019). Computational methods used in predicting the pathogenicity of variants are also part of the diagnosis filter strategy by the American College of Medical Genetics and Genomics (Pejaver et al. 2022). Methods are developed even to predict the expression of genes based on sequencing data (Avsec et al. 2021). Further gene tolerance scores from gnomAD have played a crucial role in the diagnosis of rare diseases (Lek et al. 2016).

The following introduction briefs about the state of the art of diagnosis in clinics and the approaches that would aid in better diagnosis. The section (1.1) discusses the use of sequencing methods in clinical diagnosis and possible reasons for low diagnostic yield. Section (1.2) talks about the need for machine learning models as a diagnostic tool, and also briefs on two machine learning models that were used in the thesis. Section (1.3). describes the available methods for gene prioritization and why gene prioritization is needed. Last section (1.4) introduces single cell sequencing methods and computational tools available to decipher the sequencing output. It also provides a brief introduction of how single cell sequencing is used in the study of congenital diseases.

1.1 Sequencing as a diagnostic tool

Next Generation Sequencing (NGS) has rapidly become faster and more affordable, significantly boosting the efficiency of genetic testing (Elsner et al. 2021). In recent

years, NGS has been used in clinics as a diagnostic tool for rare Mendelian diseases (Wojcik et al. 2024). This has significantly advanced our capabilities in gene discovery, functional annotation, and the mapping of disease-related genes (Figure 1.1) (Thorpe et al. 2024). Resources such as the Human Phenotype Ontology (HPO) (Gargano et al. 2024) and the Online Mendelian Inheritance in Man (OMIM) (Amberger et al. 2015, 2019) database have greatly benefited from these advancements. NGS has also driven the development of sophisticated tools and resources designed to call, annotate, prioritize, and filter genetic variants (Figure 1.1). Examples of such resources include the Genome Aggregation Database (gnomAD) (Karczewski et al. 2020) and the Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER) (Firth et al. 2009), which have been instrumental in enhancing our understanding of genetic variation and its implications in human health and disease. As per 30th July 2024, OMIM reported 7538 phenotypes with a known molecular basis and 4919 genes with a phenotype-causing mutation ("Gene Map Statistics - OMIM," n.d.; Amberger et al. 2019).

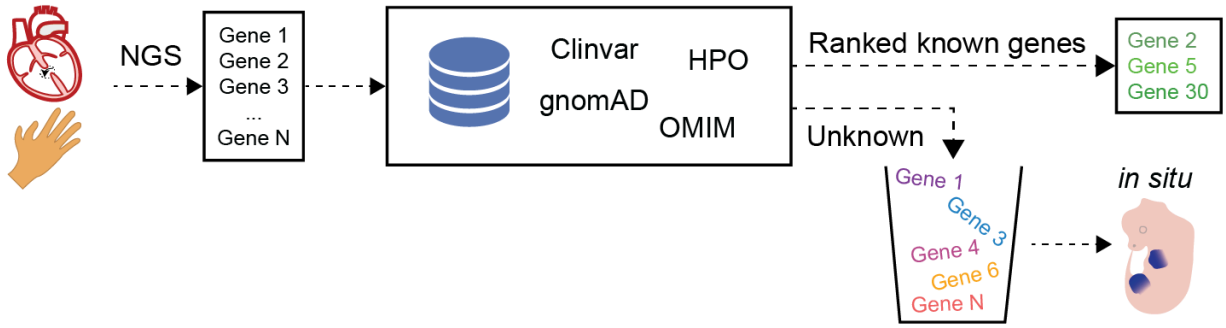


Figure 1.1 Current molecular diagnostic workflow. The genome or exome is sequenced to detect gene variants, causative of congenital malformation (eg Limb or heart malformation). These genes are then functionally annotated using HPO and OMIM to pinpoint the causal gene. When known genes do not harbor pathogenic variants, genes with unknown functions are validated through experimental techniques such as in situ hybridization.

Despite these advancements, our current diagnostic yield for rare diseases stands at only 41% (100,000 Genomes Project Pilot Investigators et al. 2021). This is due to the fact that many potentially pathogenic variants lie on genes that are not functionally

annotated and are labeled as variance of unknown significance (VUS) (Figure 1.1) (Joynt et al. 2022). Techniques like *in situ* hybridization, stains one single gene at a time to experimentally validate VUS (Figure 1.1) (Jin and Lloyd 1997). It is experimentally laborious and time consuming, to validate all genes that carry a variant. This limitation has spurred the search for new strategies to functionally annotate novel disease genes. Enhanced functional annotation is crucial for improving diagnostic accuracy and understanding the underlying mechanisms of rare diseases (Steward et al. 2017).

1.2 Machine learning as a tool diagnosis.

Machine learning (ML) is a field that uses statistical and algorithmic methods to model patterns in data (Jordan and Mitchell 2015). These models can then be applied to perform tasks or make predictions on new, unseen data (Duda, Hart, and Stork 2012). Broadly the methods can be divided into supervised, unsupervised and semi-supervised learning. Supervised learning is used in a setting where the model is trained on labeled data and the predictions are performed in unknown data (Hastie, Tibshirani, and Friedman 2001). In unsupervised, the model learns the relationship between the input features, where no label is provided, like cluster analysis (Berry, Mohamed, and Yap 2019). In semi-supervised, the input has a part of the data with labels and also uses the unlabeled data to enhance the predictions (Libbrecht and Noble 2015). The invention and application of various ML methods date back to the 1800s with the development of linear regression for astronomical data, gaining widespread popularity in the recent ages with its application in various fields like medical, image analysis, language processing, finance (James et al. 2021). Machine learning is a broad term that comprises methods like, linear, logistic regression, dimensionality reduction methods like principal component analysis, clustering methods like Lovain, k-means, ensemble methods like random forest, support vector machine and neural networks (James et al. 2021). The selection of a machine learning model is based on the tradeoff between flexibility and interpretability. Models with a large number of parameters, making them flexible, are very difficult to interpret like deep learning models and are only suitable for input which have a larger dataset (James et al. 2021). The parameters of the selected

model are optimized based on cross validation approaches, where the error is calculated for each combination of parameters on the training dataset, either by random search, grid search. The selected model is then trained with the optimized parameters, to predict or classify data.

In the real world, most often the labeled classes are not balanced leading to a bias in the prediction towards the major class (Fernández et al. 2018). In order to balance the classes, resampling approaches, like oversampling and undersampling are used (Benkendorf et al. 2023). In oversampling the minor class is multiplicated until the class balance is achieved. In undersampling the major class is subsampled until the class balance is achieved (Thölke et al. 2023). In the first approach, there is no new learning, and in the second there is a loss of information. Thus data augmentation methods like Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2011), are used to synthetically create data from the minor class.

Applications of ML have been extended to the field of genetics, due to the vast amount of information available, necessitating the use of statistical models to understand the underlying structure (Ghosh and Dasgupta 2022). Computational predictions have been used widely in the analysis of the genomic data, like DNA in the prediction of transcription start site regions, enhancers, promoters element and RNA expression data in the predictions of disease phenotypes and identifying genetic markers (Libbrecht and Noble 2015). In the following subsections, I will brief about two methods that were used in this thesis.

1.2.1 Random forest model

Decision trees are hierarchical structures that recursively split the input features based to make predictions or classifications (Fürnkranz 2011). The splitting of features is done in an optimized way over the gini index or entropy to calculate the gain of information. Gini index measures the probability of randomly chosen samples to be misclassified, based on the distribution of the classes and Entropy measures the randomness in the samples (Mitchel 1997). Entropy is highly sensitive to the data, but possesses computational burden due to the logarithmic calculation.

$$Gini\ Index = \sum_j p_j \left(1 - p_j\right)$$

$$Entropy = -\sum_j p_j . log_2 . p_j$$

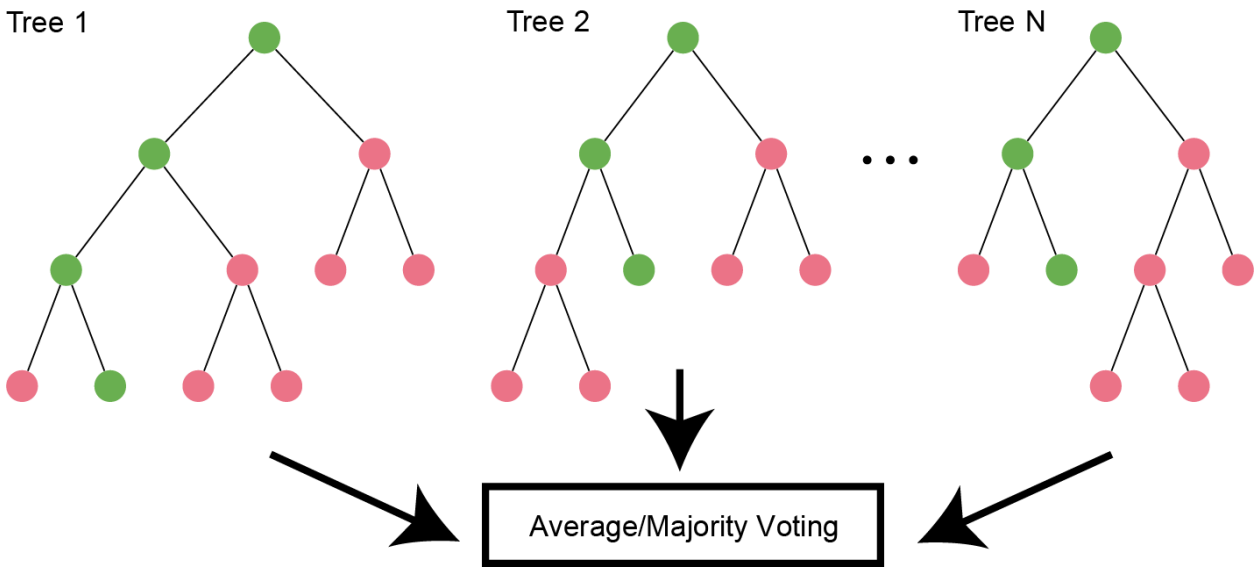


Figure 1.2.1 Random forest model. The model consists of an ensemble of decision trees, each processing different features and different samples. Majority voting of the trees are taken to classify the data.

Random Forest is an ensemble machine learning model comprising a collection of decision trees (Figure 1.2.1) (Breiman 2001). Each tree in the forest is trained on a different random subset of the training data, created using a method called bootstrapping, where samples are selected with replacement. Additionally, at each split within a decision tree, only a random subset of features is considered, ensuring that no two trees are exactly alike. This approach results in a diverse set of trees, each capturing different patterns in the data (Breiman 2001). For classification tasks, the Random Forest aggregates the predictions of all trees through majority voting, while for regression tasks, it takes the average of all predictions (Figure 1.2.1). The model has gained popularity in the field of genomics due to its interpretability and its wide range of applications in classification of disease vs non-diseased samples, based on microarray data (Statnikov, Wang, and Aliferis 2008), in pathway based analysis and GWAS (Chen and Ishwaran 2012).

1.2.2 Support vector machine

Support Vector Machine (SVM) is a popular algorithm used for classification tasks, aiming to find a hyperplane that separates data into different classes (Figure 1.2.2 A) (Cortes and Vapnik 1995). A hyperplane serves as a decision boundary, and in its linear form, SVM works well when data is linearly separable. However, biological datasets often exhibit nonlinear patterns, making it difficult for a linear SVM to classify them effectively. To address this, SVM uses kernel functions, which map the data into a higher-dimensional space, where it becomes more separable. Common kernel functions, like the Radial Basis Function (RBF) and polynomial kernels, enable SVM to handle non-linear relationships by transforming the data without explicitly computing the higher dimensions (Figure 1.2.2 B) (James et al. 2021). These kernels allow SVM to classify complex patterns in biological or other challenging datasets by creating a decision boundary in this transformed space. This versatility, combined with effective classification in high-dimensional settings, makes SVM with kernels a powerful tool for tackling non-linear classification problems. SVM's are applied in the classification of cancer (Shujun Huang et al. 2018).

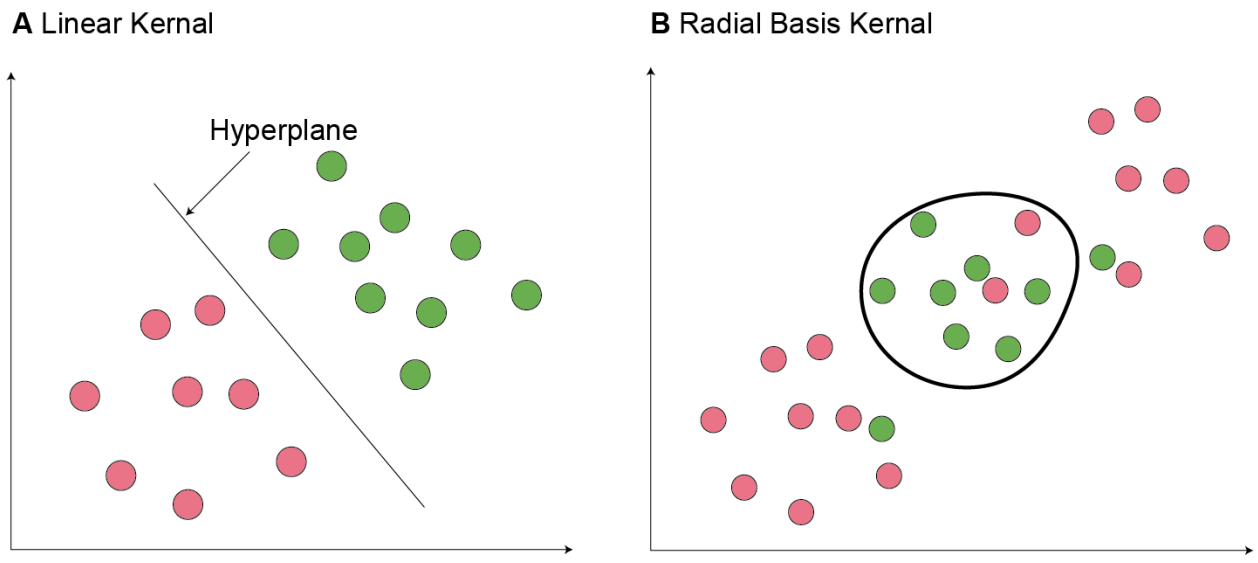


Figure 1.2.2 Support vector machine. (A) Linear separable data, classified by a hyperplane. (B) Radial basis Kernel separating nonlinear data.

1.2.3 Performance metrics

In order to validate if the designed model is performing well, we need to evaluate the model (Rainio, Teuho, and Klén 2024). In a classification task, more often a confusion matrix is constructed to compare the predicted classes vs the actual true classes. These metrics help understanding the accuracy and reliability of the model. For a classification task a confusion matrix consists of four groups, true positive (TP), where a true positive class is predicted as positive class, true negative (TN), where a true negative class is predicted to be negative, false positive (FP, Type I error), where a true negative class is predicted to be positive and false negative (FN, Type II error), where a true positive class is predicted to be negative (James et al. 2021) (Figure 1.2.3 A)

A Confusion Matrix

		Predicted	
		Positive Class	Negative Class
Actual	Positive Class	TP	FN
	Negative Class	FP	TN

B ROC curve

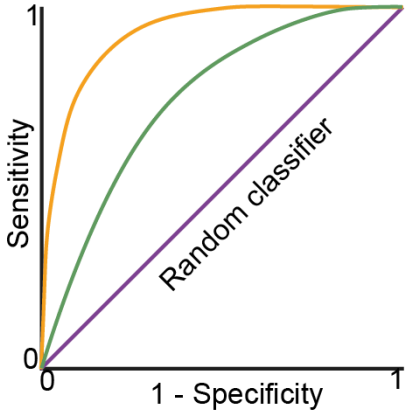


Figure 1.2.3 Performance metrics. (A) Confusion matrix. (B) ROC curve.

From these values, we calculate the following:

$$Sensitivity = \frac{TP}{(TP + FN)}, \; Specificity = \frac{TN}{(TN + FP)},$$

$$Precision = \frac{TP}{(TP + FP)}, \; Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

The values of the confusion matrix change based on the threshold, which affects metrics like sensitivity and specificity. At a higher threshold, specificity will be high,

meaning fewer negative classes are incorrectly predicted as positive. On the contrary, at a lower threshold, sensitivity increases, so fewer positive classes are misclassified as negative. To understand how sensitivity and specificity change with the threshold and to select the optimal threshold, the ROC (Receiver Operating Characteristic) curve is plotted with sensitivity (true positive rate) against 1 - specificity (false positive rate) (Figure 1.2.3 B) (Rainio, Teuho, and Klén 2024; Flach 2011). The ROC curve represents the trade-offs between these two metrics at various thresholds. This metric sums the ability of the model to provide a discrimination between classes across all thresholds. The area under the curve (AUC) represents the model's ability to distinguish between classes across all thresholds. The higher AUC a model has, the better the model performs. An AUC value close to 1 represents perfect discrimination while an AUC value close to 0.5 represents random classifier performance. The threshold which best balances the two values, sensitivity and specificity for that given problem, can be identified with the help of an ROC curve and AUC.

1.3 Gene prioritization

Not all candidate genes at a locus are associated with the phenotype (Zolotareva and Kleine 2019). Identifying the key candidate gene that is causative is experimentally tedious. Gene prioritization is a computational approach to rank candidate genes based on their probability to be associated with a disease (Azadifar and Ahmadi 2022). It helps in cutting down the list of candidate genes under consideration for experimental validation. To prioritize genes, computational methods use information about genes like expression, ontologies (Peng et al. 2021), pathways, co-expression, gene intrinsic properties (Piro and Di Cunto 2012) and protein protein interaction networks (Azadifar and Ahmadi 2022; Schlüter et al. 2023), to generate a score that defines the likelihood of a gene being causally related to a disease (Figure 1.3).

There are several approaches to associate genes to diseases, like direct association, guilt by association or by text mining approaches (Zolotareva and Kleine 2019). In the first approach, gene-disease associations are identified through case-control experiments, using genetic studies. There are several diseases where a single gene is

associated with the disease and segregates in the family. OMIM reports 6,484 phenotypes and 4,557 with single gene disorders and traits (Amberger et al. 2019). But for complex disorders, difficulty arises in mapping diseases to specific genes. Differential analysis of patients and control samples is performed to excavate the genes which behave differently in a disease setting.

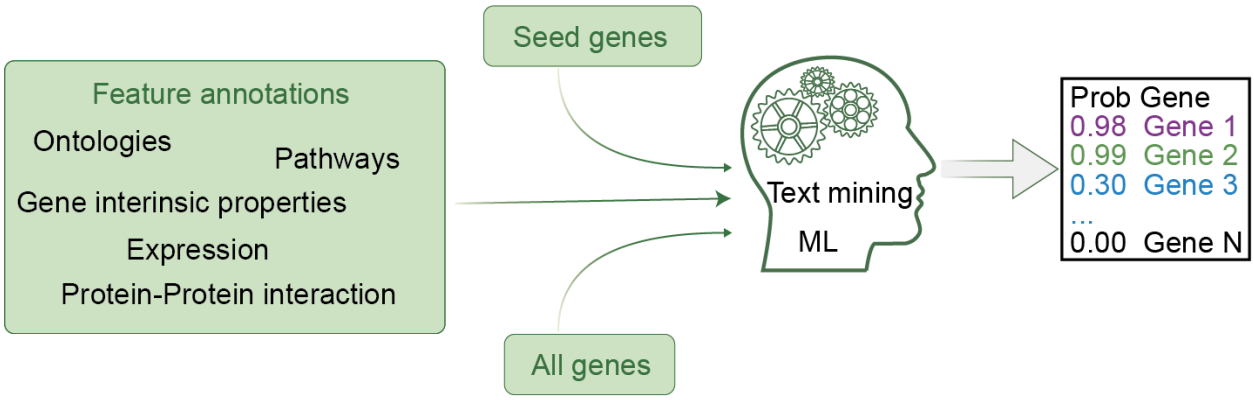


Figure 1.3 Gene prioritization strategies. The genome or exome is sequenced to detect gene variants. These genes are then functionally annotated using HPO and OMIM to pinpoint the causal gene. Given a starting seed and the known annotation for all genes, prioritization methods use machine learning (ML) methods or text mining approaches to score genes based on the disease causing probability.

The most common computational approach to prioritize genes is based on guilt by association or indirect association. This is based on the principle that genes that are expressed together cause similar diseases. This method compares a known set of genes to other genes, to see similar behavioral patterns in the biological data. The biological data source used by prioritization methods also plays a crucial role. Certain tools are based on annotation resources like HPO (Smedley and Robinson 2015), as discussed in chapter 1.1, not all genes are functionally well annotated, thus methods which use this approach would be highly biased towards genes that are well annotated and lead to false positive predictions (Moreau and Tranchevent 2012). On the other hand, tools like GADO (Deelen et al. 2019), GeneFriends (Raina et al. 2023), geneTier (Antanaviciute et al. 2015) use unbiased information like gene expression, gene intrinsic properties, gene tolerance metrics. Text mining approaches like Open targets (Koscielny et al. 2017), Genie (Fontaine et al. 2011) aggregate scores for genes based on several

pieces of evidence given an input phenotype. However these methods do not accurately detect the true gene associations (Zolotareva and Kleine 2019).

Gene prioritization based on the type of assumption can be broadly classified as direct association or indirect association (Zolotareva and Kleine 2019). In direct associations, genes are prioritized based on their differential expression in the wildtype vs diseased tissue. Tools like geneTIER are based on the hypothesis that genes that are associated with the disease are highly expressed in the disease tissue (Antanaviciute et al. 2015). In indirect association, genes associated with known disease associated genes through the guilt by association approach to prioritize potential candidate genes (Zolotareva and Kleine 2019).

1.4 Single-cell sequencing

Cells are the fundamental structural and functional units of organisms. They are crucial in forming tissues, organs, and the entire organism. The organization of cells within an organism is essential for its proper functioning. Cellular organization involves the differentiation and arrangement of organ-specific cells to form structural units that carry out specific functions (Otani et al. 2010). Recent advancements in single-cell sequencing (sc-seq) technologies have revolutionized the profiling of genetic, epigenetic, transcriptional, and proteomic variations within individual cells. Recognized as the "Method of the Year" by Nature Methods in 2013 and the "Breakthrough of the Year" by Science in 2018 (Crespi 2018). sc-seq techniques have facilitated the identification of new cell types, the construction of cellular atlases in humans and model organisms (Domcke et al. 2020; Cao et al. 2020, 2019; Han et al. 2022). Through initiatives like the Human Cell Atlas, sc-seq technologies contribute to mapping the entire human body, enabling the deciphering of cellular-level changes during development, disease mechanisms, the tracking of cancer progression and therapy (Reed et al. 2024; Smajić et al. 2022; Sikkema et al. 2023; Kong et al. 2023; Z. L. Liu et al. 2024; Shengkang Huang et al. 2023). Organismal-level atlases offer insights into development, mature organ cell type constitution, and the spatial context of every cell, aiding investigations into developmental disorders and pleiotropism across multiple organs (X. Huang et al. 2023). Additionally, cross-species comparison atlases that

integrate mouse and human data provide valuable insights into the evolutionary aspects of organs, organism-specific cell types, and differential gene expression across species. These atlases help establish connections between species, enhancing our understanding of human diseases and how model organisms, like mice, can be effectively used to study these conditions. This comparative approach aids in translating findings from animal models to human biology and disease mechanisms (T. Liu et al. 2021).

1.4.1 Single cell sequencing approaches

The experimental steps involve dissociating the tissue, capturing and labeling the molecule and sequencing. The sample of interest is dissociated into single cells or single nuclei (Figure 1.4.1 A). Nuclei is extracted in cases where the cell membrane is not intact during dissociation (The manuscript uses the term cells to keep the consistency, but it could be nuclei or cell). The molecule is then barcoded using microdroplet (Klein et al. 2015) or microwell (Hashimshony et al. 2012) based methods (Figure 1.4.1 B, C). This technique helps to track back the cells/ nuclei of origin during the downstream analysis. The number of cells barcoded per experiment has scaled over the years, with microwell based methods are deeper sequenced with the capture of fewer cells and microdroplet based methods capture more cells with lower transcripts (Sreenivasan, Balachandran, and Spielmann 2022). Methods like combinatorial indexing aim to capture several million cells, with fewer transcripts per cell, which are generally used in whole organism sequencing as opposed to other methods used in sequencing a single organ (Cao et al. 2019). Though scRNA seq is a very common approach, there are several methods for capturing other modalities of a single cell, like genome, epigenome, proteome and combination of modalities (sc-multiome) (Stoeckius et al. 2017). For genome and epigenome, the DNA is fragmented and barcoded, the transcriptomes are captured based on the poly-A tail and the proteome based on the oligonucleotide tags. Which are then PCR amplified and sequenced (Kashima et al. 2020). The prepared samples are then often sequenced using paired end short read

sequencing methods (Figure 1.4.1 D). In case of transcriptome, full length or 3’ and 5’ end can be sequenced whereas there are studies which have also performed long read sequencing which are of primary interest in cancer studies due to the ability to capture fusion genome, splicing and alternative transcripts (Figure 1.4.1 E). There are several commercial kits available to prepare the samples and sequence them (De Simone et al. 2024).

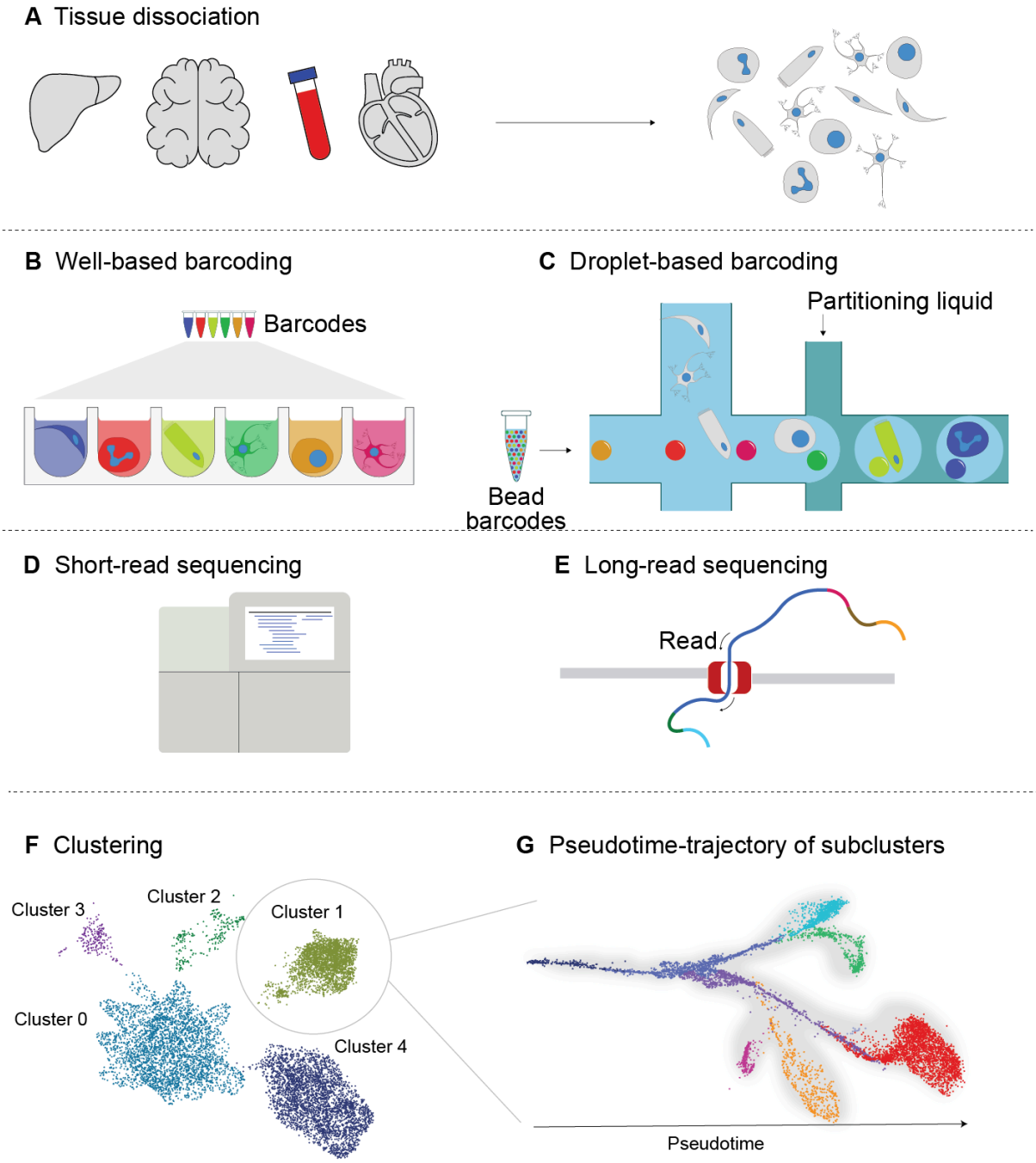


Figure 1.4.1 Workflow of sc-seq. (A) Tissue of interest is dissociated (B-C) Barcoding strategy using well or droplet based to capture each single cell for sequencing. (D-E) Short read sequencing commonly used and Long read sequencing in case of fusion gene or alternate transcript studies. (F) Clustering approach to group cells of similar type together. (G) Trajectory analysis to study the temporal dynamics within cell types. Note: Figure adapted from (Sreenivasan, Balachandran, and Spielmann 2022)

The sequencer outputs the data in binary base call (bcl) format which contains the base and the quality for every sequencing cycle. This file is converted to fastq format by bcl2fastq for the downstream analysis, which contains reads per sample. The sequenced reads are then demultiplexed to trace each reads cell of origin, which are then aligned to the reference genome. The aligned reads are quantified into a feature-barcode matrix, where each column contains the barcode identifying the cell of origin and each row represents a feature (Zheng et al. 2017). In the case of transcriptome sequencing, each row corresponds to a sequenced transcript, and the matrix quantifies the expression of each gene in its respective cell of origin. The raw barcode matrix is first filtered to remove cells which have captured a low number of features and high ribosomal or mitochondrial counts to remove the low quality and dead cells (Andrews et al. 2021). The data is further filtered to remove doublets (multiple cells are captured by one bead), as they can lead to false interpretation of the data. Methods like scrublet (Wolock, Lopez, and Klein 2019), simulate doublets by combining transcripts and use the nearest neighbor algorithm to detect doublets. The filtered data is highly heterogeneous because of biological and experimental variations, thus every single cell is different, therefore normalization is necessary to make them comparable. For normalization the data is scaled to a scale factor (10000) and log transformed to reduce the skewness in the data (Heumos et al. 2023).

In order to understand the biology, genes that can distinguish the underlying biological variation are selected (highly variable genes (hvg)), which are then dimensionality reduced by preserving the variance in the data, using methods like Principal Component Analysis (PCA). With the reduced data, nearest neighbors are computed which are then used to identify clusters based on network based clustering methods (Figure 1.4.1 F). Dimensional reduction methods like Uniform Manifold Approximation Projection (UMAP) (McInnes, Healy, and Melville 2018) and t-distributed stochastic neighbor embedding (t-SNE) are used for visualization of the data (van der Maaten and Hinton 2008). Both the projection algorithms are scatter plot representation of cells, t-SNE is based on gaussian probability function, where it finds the probability of which cells are likely to be neighbors, whereas UMAP creates a rough estimate graph from the high dimensional data, from which the low dimensional graph is created, thus preserving the global

structure. The cell types are annotated based on the top genes that are differentially expressed in the cluster as opposed to other clusters. The cell types are either manually annotated using various literature or tools which annotate based on a reference single cell dataset (Abdelaal et al. 2019). The most widely used tool for the analysis is Seurat (Butler et al. 2018) which is based on R and Scanpy (Wolf, Angerer, and Theis 2018) which is based on python. With the growth of the technology, various computational tools are available to analyze the data (Heumos et al. 2023; Davis et al. 2024).

1.4.2 Developmental Trajectory inference

Developing cells proliferate and differentiate to specialized cells that form the tissue. This process is governed by various gene regulatory programs and signaling processes, where each cell determines its fate causing the variability within a tissue (Trapnell et al. 2014). Understanding the different mechanisms that drive this decision provides insights into the ontogenesis of the organism.

In order to understand the changes cells undergo during differentiation and development, cells need to be organized in the order of development. Capturing cellular heterogeneity is the key advantage of sc-seq, thus providing the transcriptomic profile of individual cells. Trajectory analysis in sc-seq inferes the order of each cell across their differentiation time (termed as pseudotime as they are different from the sample's physical age), to understand the differences in cells during maturation. The cells are organized in the gradient of differentiation, based on the transcriptional changes, providing information of the dynamics of expression within the same cell type during development (X. Qiu et al. 2017) (Figure 1.4.1 G). There are several benchmarking articles comparing different methods of trajectory inference in sc-seq (Saelens et al. 2019).

The tool used in this thesis is based on reversed graph embedding (Qi Mao et al. 2017). To construct the trajectory the method identifies genes that are expressed in 5% of the cells, which are then projected onto principal components (PC). K-means clustering algorithm is used to select a centroid of the cells in the projected dimension. A graph is constructed through this centroid as the node. The cells are moved towards this centroid, which is then used to calculate the new centroid, iteratively until optimization.

The optimized graph is termed as the principal graph. The root node is selected based on the input from the user, and pseudo time is calculated as the distance from this root node to the corresponding node (X. Qiu et al. 2017).

Another measure of dynamics in cells is the changes in mRNA splicing over time, defined by RNA velocity (La Manno et al. 2018). This measure predicts the future state of the cells. The differentiation pathway is measured as the transition probability from one cell state to the other.

1.4.3 Integration methods for single cell data

Analysis in a single cell generally consists of several sequencing rounds or integrating with the publicly available datasets. Often technical variability exists in the datasets due to differences in sample, experimental methods and sequencing platforms known as batch effects (Luecken et al. 2022). Thus there is a need to develop computational methods that can integrate datasets from different sequencing batches. Integration methods aim to remove technical variability keeping intact the biological variability, thus enabling the comparison of multiple samples (Figure 1.5). Integration methods have been used in the comparison of samples sequenced at various conditions like control and mutant, drug response, samples at different time points of development and evolutionary (Sreenivasan et al. 2023; C. Qiu et al. 2022, 2024; Mah and Dunn 2024). There is several benchmarking literature that compares the different batch correction methods (Tran et al. 2020), this thesis elaborates two tools that I have used for the project, namely Harmony (Korsunsky et al. 2019) and reciprocal PCA from Seurat.

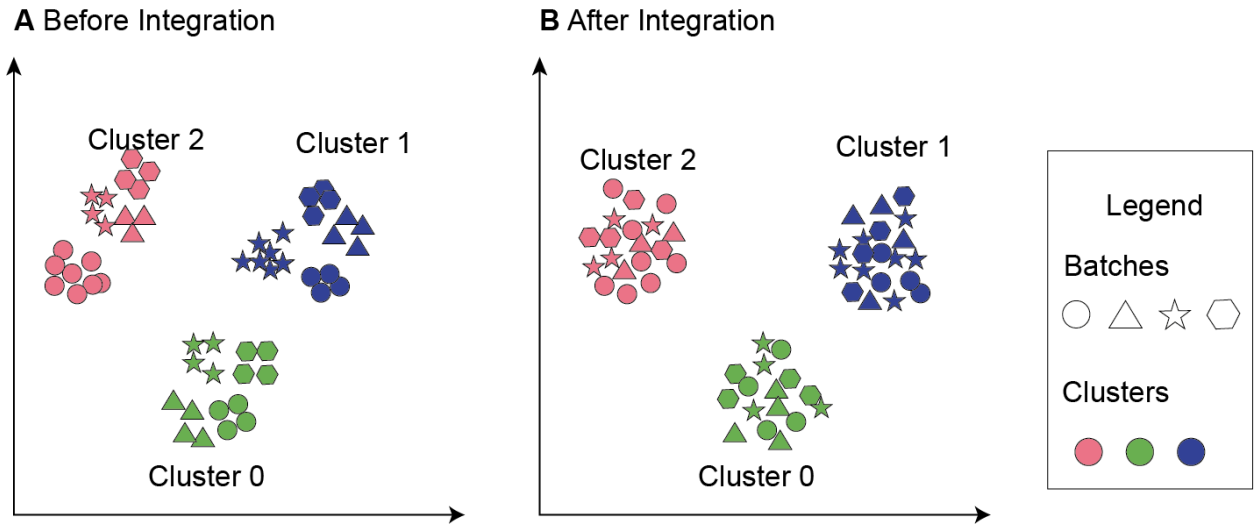


Figure 1.4.2 Illustration of single cell integration. (A) Shows sample clustering based on batch, where each symbol represents a batch and each color indicates the corresponding cell type. (B) Demonstrates how batch integration algorithms reduce technical differences between batches.

Harmony based integration strategy has been a recommended approach in the field, due to its efficiency to effectively integrate similar cells, sequenced by different technologies and has a low runtime (Tran et al. 2020). It is based on soft k-means clustering methods, where the cells are assigned to a preliminary cluster in the low dimensional space. Then the algorithm tries to find the centroid of each batch within the cluster and the global centroid of the cluster. This is then used to calculate the correction factor for each of the clusters, taking into account that the cluster also contains different states of the cell. Then a cell specific factor is calculated based on the linear combination of correction factor and the preliminary cluster assignment factor, which is used to move each cell towards the global centroid within the cluster. This process is performed iteratively until convergence (Korsunsky et al. 2019).

Seurat (Stuart et al. 2019) based integration has been recommended for large dataset integration (Tran et al. 2020). The reciprocal PCA based method, projects each dataset/batches into another datasets PCA space to calculate the anchors based on mutual neighbors. They use a shared nearest neighbor approach to avoid the capture of anchors between different cell states. The anchors are then weighted based on the associated between the cells of the datasets, which is used to calculate the correction

vector. For multiple datasets, a tree based approach is used where the pairwise similarity is calculated, and a recursive correction is made (Butler et al. 2018).
The choice of integration strategy depends on the type of datasets, taking into advantage of different approaches available. The efficiency of the methods is calculated based on its capability to bring similar cell types together explaining the biological similarity.

1.4.4 Study of Congenital diseases by single cell approaches

Congenital diseases are anomalies that are present at birth. They are a result of mutations in genes or non-coding elements that regulate the genes, with a prevalence in 2-5% of children (Deciphering Developmental Disorders Study 2017). Critical to human geneticists is identifying genes, pathways, or mechanisms underlying disease phenotypes, as well as detecting, monitoring, and predicting disease, its progression, or response to interventions. It is often seen that candidate disease genes exhibit preferential expression in particular cell types and/or at specific time points during embryonic development (Maj et al. 2024). Understanding the pathways where the altered gene functions, with bulk sequencing techniques is difficult due to the averaging of the measure of individual cells. There are several studies which have compared the wildtype and disease tissue at single cell level to elucidate the disease mechanisms and the cell types involved (Nomura et al. 2018; Manivannan et al. 2021; Maj et al. 2024). X.Huang et al. profiled 22 mutants ranging in phenotypically severity with sc-RNA seq in 101 embryos to study a range of phenotypes associated with developmental disorders, characterizing the molecular phenotypes at cellular resolution (X. Huang et al. 2023). Sc-seq has provided deeper insights into the molecular and cellular mechanisms underlying congenital diseases, allowing for a more precise understanding of how genetic variations at the individual cell level contribute to disease development and progression.

2. Aim of this study

The thesis focuses on utilizing single-cell sequencing technologies to investigate development and genetic diseases. To achieve this, two projects were undertaken. The first project aimed to explore evolutionary differences across various model organisms commonly used in disease research, leveraging single-cell datasets from adult organisms. The second project applied a machine learning approach to identify and prioritize novel disease genes within the context of congenital diseases. Detailed descriptions of these project aims are provided below:

Comparative single-cell analysis of the adult heart and coronary vasculature

In the first part of the thesis, we aimed to use single-cell RNA sequencing (scRNA-seq) data to compare the adult heart and coronary vasculature of three organisms: human, macaque, and mouse. Our goal was to determine whether the technology could identify organism-specific features of the heart while also highlighting conserved characteristics. Additionally, we sought to compare fibroblasts from the human heart and zebrafish, given the latter's remarkable ability to regenerate after injury. To achieve these objectives, we utilized publicly available scRNA-seq data and applied an integration analysis workflow using Seurat and Harmony to compare the model organisms to the human dataset.

STIGMA: Single-cell tissue-specific gene prioritization using machine learning

In the second part of this thesis, we aimed to use scRNA-seq data to prioritize potential candidate genes for congenital diseases. To achieve this, we developed a machine learning model that incorporates single-cell features, such as gene expression across cell types and gene expression over pseudotemporal trajectories, along with gene-intrinsic properties. The model ranks genes based on their probability of being associated with the disease. We applied our model to two different disease contexts: congenital limb disorders and congenital heart disease (CHD). Additionally, we

evaluated its performance on both human and mouse single-cell data for congenital heart disease, demonstrating that mouse single-cell data can serve as an effective proxy for studying human diseases.

3. Results

The results section focuses on the key findings of the thesis to understand development and congenital diseases. This was divided into two papers, where the first paper (Section 3.1) studies the evolutionary similarities and differences between adult hearts in vertebrates. The second paper (Section 3.2) studies the use of single cell transcriptome data to prioritize candidate disease genes for congenital diseases.

3.1 Comparative single-cell analysis of the adult heart and coronary vasculature.

Balachandran, Saranya, Jelena Pozojevic, Varun K. A. Sreenivasan, and Malte Spielmann. Mammalian Genome 34, no. 2 (June 2023): 276–84. https://doi.org/10.1007/s00335-022-09968-7.

Heart has been a key organ which pumps blood, though its function is common across species, it possesses anatomical differences, like shape, size, regeneration capacity. In order to further understand these differences, we used scRNA-seq to compare the adult hearts of human, mouse, macaque (*Macaca fascicularis*) and zebrafish. For all the species we used publicly available datasets. To measure the similarities and to remove the technical effects due to the different sequencing technology used, the data was integrated using Seurat's reciprocal PCA method. The cell types were annotated based on the original publications. We identified 9 major clusters, ventricular cardiomyocytes (vCM), atrial cardiomyocytes (aCM), pericytes and smooth muscle cells, fibroblasts, adipocytes, endothelial cells, mesothelial and epicardial cells, neuronal cells and immune cells. We compared the sub cell types from the model organisms to the human dataset, and observed no significant differences in aCM. We observed mouse specific cell types in vCM, which had a high expression of *Prune2* and a lower expression of *Myh7* and *Plcl1.* We also observed genes that encode mitochondrial respiratory chain enzymes like *Ndufa4*, *Ndufb11 and Cox7c* showed higher expression in mice than in human cell types. We saw macaque specific smooth muscle cell cluster which

expressed *POSTN*, *DES*, *ACTC1*, which are the marker genes of aorta and coronary arteries (Zhang et al. 2020). We further wanted to investigate the differences between fibroblasts of humans and zebrafish, due to the capability of cardiac regeneration (Hu et al. 2022). The genes responsible for regeneration, had a very low expression in very less percentage of cells in humans compared to zebrafish. The study showed differences that suggested functional and sampling specific differences. The study showed the capability of scRNA-seq to identify species specific differences and similarities.

3.1.1 Project contribution

Malte Spielmann and I designed the research. I performed the computational analysis. Jelana Pozojevic, Varun. K.A. Sreenivasan, Malte Spielmann and I interpreted the results. Jelana Pozojevic and I drafted the manuscript. Jelana Pozojevic, Varun. K.A. Sreenivasan, Malte Spielmann and I revised and approved the final manuscript.

Mammalian Genome (2023) 34:276–284
https://doi.org/10.1007/s00335-022-09968-7

Comparative single-cell analysis of the adult heart and coronary vasculature

Saranya Balachandran[1] · Jelena Pozojevic[1] · Varun K. A. Sreenivasan[1] · Malte Spielmann[1,2,3]

Received: 11 August 2022 / Accepted: 7 November 2022 / Published online: 19 November 2022

Abstract
The structure and function of the circulatory system, including the heart, have undergone substantial changes with the vertebrate evolution. Although the basic function of the heart is to pump blood through the body, its size, shape, speed, regeneration capacity, etc. vary considerably across species. Here, we address the differences among vertebrate hearts using a single-cell transcriptomics approach. Published datasets of macaque (*Macaca fascicularis*), mouse, and zebrafish hearts were integrated and compared to the human heart as a reference. While the three mammalian hearts integrated well, the zebrafish heart showed very little overlap with the other species. Our analysis revealed a mouse-specific cell subpopulation of ventricular cardiomyocytes (CM), represented by strikingly different expression patterns of specific genes related to high-energy metabolism. Interestingly, the observed differences between mouse and human CM coincided with actual biological differences between the two species. Smooth muscle and endothelial cells (EC) exhibited species-specific differences in clustering and gene expression, respectively, which we attribute to the tissues selected for sequencing, given different focuses of the original studies. Finally, we compared human and zebrafish heart-specific fibroblasts (FB) and identified a distinctively high expression of genes associated with heart regeneration following injury in zebrafish. Together, our results show that integration of numerous datasets of different species and different sequencing technologies is feasible and that this approach can identify species-specific differences and similarities in the heart.

Introduction

The heart is a complex organ at the center of the circulatory system, which pumps blood through the body, enabling the exchange of nutrients, respiratory gasses, metabolic waste, etc. It has evolved over millions of years, from simple structures like those seen in insects and worms to powerful four-chambered mammalian hearts. In vertebrates, a multi-chambered heart exists together with a closed vascular system composed of arteries, veins, and capillaries. The simplest vertebrate heart belongs to fish and consists of two chambers, while most reptiles (except for crocodiles and alligators) have a three-chambered heart, consisting of two atria and a ventricle (Stephenson et al. 2017). Finally, mammals and birds have a four-chambered heart that consists of two atria and two ventricles, where the right ventricle pumps deoxygenated blood to the lungs, while the left ventricle pumps blood rich in oxygen to the rest of the body. The four chambers of the heart are attached to major veins and arteries that bring blood into (e.g., vena cava) or carry blood away (e.g., aorta) from the heart, while the coronary arteries (coming out of the aorta) supply blood to the heart muscle itself. The heart is composed of multiple cell types, including cardiomyocytes (CM) that generate contractile forces, smooth muscle cells (SMC) and pericytes (PC) that form blood vessels and play key roles in vascular contraction, tone, and integrity, endothelial cells (EC) that regulate exchange between the bloodstream and the surrounding tissue, fibroblasts (FB) that produce connective tissue and other cell types such as neuronal-, lymphoid-, myeloid cells and adipocytes (Alberts et al. 2002; Litviňuková et al. 2020).

Single-cell sequencing technologies (sc-seq) have enabled the detailed characterization of these cell types based on their gene expression profiles (Sreenivasan et al. 2022).

✉ Malte Spielmann
malte.spielmann@uksh.de

[1] Institute of Human Genetics, University Hospital Schleswig-Holstein, University of Lübeck and Kiel University, Lübeck and Kiel, Germany

[2] Human Molecular Genetics Group, Max Planck Institute for Molecular Genetics, Berlin, Germany

[3] DZHK e.V. (German Center for Cardiovascular Research), Partner Site Hamburg, Kiel, Lübeck, Germany

The technology has facilitated the understanding of development, differentiation, homeostasis and diseases at cellular resolution (Smajić et al. 2022; Huang et al. 2022). Within the last few years, breakthrough sc-seq methods have been applied to analyze the cellular composition of various organisms and more specifically organs, including the heart (Cao et al. 2020). Moreover, the Human Cell Atlas initiative was established with the aim to map the entire human body in adults and in embryonic stages (https://www.humancellatlas.org) (Cao et al. 2019, 2020).

With so many rich datasets spanning even evolutionarily distant species available at our fingertips, here we set out to integrate the sc-seq data from the hearts of adult mouse, crab-eating monkey (*M. fascicularis*), and human, as well as the recently published zebrafish (*Danio rerio*) adult heart dataset (Vidal et al. 2019; Zhang et al. 2020; Litviňuková et al. 2020). The data from the two-chambered zebrafish heart was of particular interest, since it possesses the ability to regenerate upon injury, which was recently attributed to the transient cell states with fibroblast-like characteristics (Hu et al. 2022). In contrast, mammalian hearts cannot regenerate after an injury, instead leaving scar tissue with decreased functionality.

Methods

Single-cell feature barcode matrices of heart tissue were obtained from ERP123138 for adult humans (Litviňuková et al. 2020), E-MTAB-7869 for adult mice (Vidal et al. 2019) and GSE117715 for adult macaques (Zhang et al. 2020). Orthologous genes from the BioMart (Ensembl Genes 107) were used to create the feature barcode matrices of the three species. Individual species were processed separately prior to integration. Cells with more than 1500 UMIs and 1000 genes and less than 10% of mitochondrial genes and 10% of ribosomal gene counts were used. Genes detected in more than three barcodes were retained.

Seurat v4 was used for the downstream analysis. First, each species was log normalized using the standard workflow (as described in the Seurat—Guided Clustering Tutorial) with 2500 highly variable features. Principal component analysis (PCA) was done on the normalized and scaled expression matrix. For integration, the common features between the datasets were found using the Select Integration Features function in Seurat, which was then used to identify the anchors based on the reciprocal PCA method, where the human dataset was used as the reference. The datasets were integrated with Seurat due to its enhanced performance with huge data (Tran et al. 2020). The integrated data was then scaled and PCA was performed. The nearest neighbor graph was built with 30 PCs, which was then clustered using the Seurat *FindClusters* function based on the Louvain algorithm. The Wilcox algorithm of the Seurat *FindAllMarkers* function was used to identify the differentially expressed genes in each of the clusters. The gene markers provided in the three studies were compared to those identified by us to annotate the major clusters.

For sub-clustering, the raw count data for endothelial, smooth muscle cells, PC, ventricular CM, atrial CM, and fibroblast cell types were individually subset and 1500 highly variable features were used for principal component analysis. These subsets of cells were integrated across the species based on Harmony at the reduced PCA space. Due to the efficiency of harmony in integrating identical cell types sequenced by different technology, it was a method of choice for integrating the cellular subpopulation (Tran et al. 2020). The nearest neighbor graphs were built on the harmony-based reduction. The Louvain clustering approach was performed based on the top 10 reduction components of the integrated datasets. The Wilcox algorithm of the Seurat *FindAllMarkers* function was used to identify the differentially expressed genes in each of the clusters. The cell type markers from Litviňuková et al. 2020 were used for identifying the cellular sub-clusters.

For zebrafish heart tissue, the single-cell-barcode matrix was downloaded from GSE159032 (Hu et al. 2022). The orthologous genes in humans to the gene identifiers of zebrafish were compiled. The cell-barcode matrix of human and zebrafish were subset to contain only the orthologous genes. The raw count data for FB were obtained from both human and zebrafish datasets. The datasets were then log normalized and scaled before merging. The merged dataset was integrated using Harmony, correcting for the species and samples. We then built a hierarchical cluster tree based on harmony reduction using the BuildClusterTree function in Seurat. The genes associated with the regeneration of FB in the zebrafish dataset were visualized in the integrated data to quantify the differential expression in humans and zebrafish.

Gene ontology analysis was performed for the ventricular CM cluster that was specific to mice using g: Profiler web server (Raudvere et al. 2019). The top 500 markers with adjusted p-value < 0.05 and a higher percentage of cells expressing the gene in the cluster were provided as input. The gene ontological biological process terms that had a negative log adjusted p-value > 10 were plotted.

Results

The adult hearts from the three species, namely human, macaque (*M. fascicularis*) and mouse have been characterized using single-cell RNA sequencing (Vidal et al. 2019; Litviňuková et al. 2020; Zhang et al. 2020). Human and mouse datasets were sequenced by Chromium Single-Cell

3′ protocol (10 × Genomics) and macaque by STRT-seq. The adult human heart and mouse heart datasets consist of 451,513 cells and 12,710 cells, respectively, whereas the macaque heart dataset generated only from the aorta and coronary arteries consists of 7989 cells. To account for the variation in sequencing technologies and the cell types sequenced, we integrated the three species with Seurat's reciprocal PCA-based method followed by clustering (Fig. 1a, Methods). The marker genes provided in the published datasets were used to annotate the clusters (Supp Fig. 1a). We were able to identify the 9 major cell types composing the adult mammalian heart (Fig. 1b). Ventricular CM, atrial CM and adipocytes consisted of only human and mouse tissues due to the sequencing focus of the macaque heart tissue (Fig. 1c, Supp Fig. 1b). Endothelial cells, PC, and SMC had a higher percentage of cells originating from the macaque tissue. 50% of the FB were composed of mouse tissue and the rest were equally represented by humans and macaques.

We then investigated the cell subpopulations of atrial and ventricular CM, endothelial cells, PC and SMC and FB. We compared the identified cell subtypes with the adult human heart study (Litviňuková et al. 2020) and saw that FB, atrial CM and EC were in agreement with the cell subpopulations of the human heart atlas, though the approach to integrate and sub-cluster the cell types was not identical to our study (Supp Fig. 2). Although the cells integrated well, we observed the segregation of macaque cells in the FB3 cluster (Supp Fig. 2b).

On sub-clustering of the ventricular CM, we identified 6 cell subtypes (Fig. 2a). By comparing the cell type compositions, we noticed a cluster consisting exclusively of mouse cells that did not integrate into the human ventricular CM (Fig. 2c, d). We sought to analyze the marker genes of this mouse-specific cell type and saw that *Myh7* and *Plcl1* had a lower expression compared to other clusters, while *Prune2* had a higher expression comparatively. On the other hand, *Ndufa4, Ndufb11 and Cox7c*, marker genes of human vCM4 that encode mitochondrial respiratory chain enzymes, showed a slightly higher expression in mice than in any human cell types (Fig. 2b, Supp Fig. 3).

We identified 7 subclusters of the PC and SMC (Fig. 3a). Five of these cell types consisted of human and mouse samples (Fig. 3c, d). PC4 had only cells from the human dataset but it could represent unknown cell states or doublets, as stated by the authors of the study (Litviňuková et al. 2020). Interestingly, we also saw a new smooth muscle cell type SMC_macaque (Fig. 3c, d) which expressed the SMC markers of humans such as *MYH11, ACTA2, TAGLN* and *CNN1* as well as other marker genes such as *POSTN, DES, ACTC1, SORBS2* which were previously described as aortic and coronary artery-specific (Zhang et al. 2020) (Fig. 3b). In addition, they expressed *CSPG4*, a marker gene for PC (Fig. 3b).

Next, we investigated the EC and could show that they were composed of 10 cell subpopulations (Fig. 4a), defined by the marker genes as described by human heart atlas study (Litviňuková et al. 2020). The cells of the three species integrated well and we did not observe any species-specific clusters. However, EC7_atria and EC8_ln were mainly composed of macaque and mouse cells (Fig. 4c, d). Additional genes highly expressed in the EC7_atria cluster include *IL13RA2* and *WIF1*, associated with aortic artery EC in macaque (Zhang et al. 2020). In contrast, lymphatic EC8_ln cells showed high expression of *RELN*, a gene expressed in macaque lymphatic endothelial cells, consistent with the cell type specificity in both macaque and human. Furthermore, we observed that *CLDN5*, previously associated with coronary artery-specific EC in macaques, was expressed in all clusters (Fig. 4b).

Finally, we wanted to study the differences between human and zebrafish hearts. Humans and zebrafish are evolutionarily distant, resulting in only 500 orthologous genes, thus making it challenging to integrate the datasets compared to integration of the mammalian species (Qiu et al.

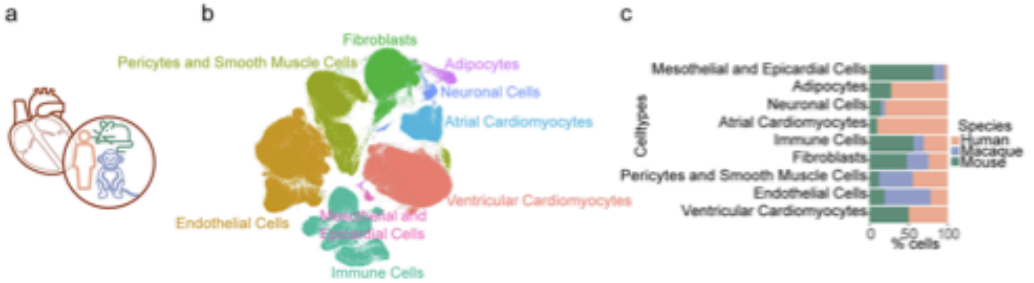


Fig. 1 Cell types and composition of adult mammalian hearts. **a** Integration of single-cell heart transcriptome data from human, macaque, and mouse. **b** 2D UMAP embedding of the major cell types after integration. **c** Percentage contribution of each species to the cells in the major cell types

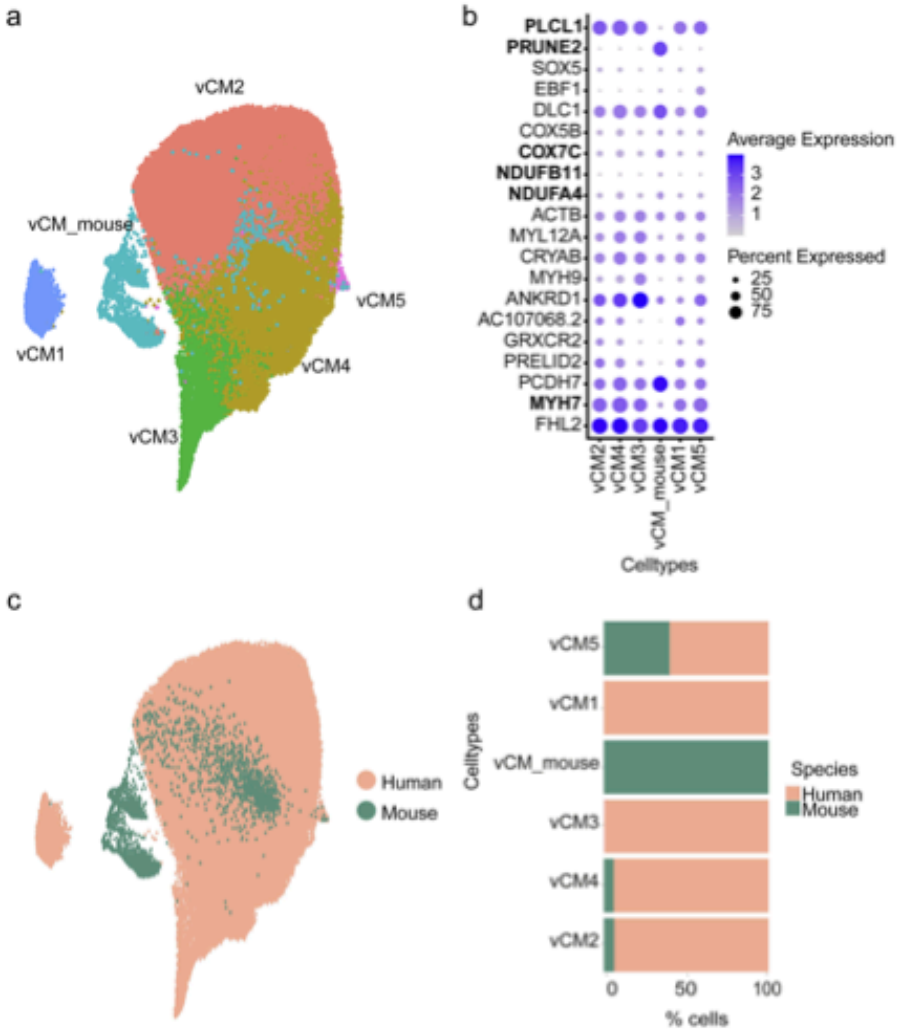


Fig. 2 Ventricular CM. **a** 2D UMAP embedding of the cell subpopulations of ventricular CM after integrating the cells from the two species. **b** Dot plot of the cell type-specific marker genes. The key genes showing differential expression between species are written in bold. **c** 2D UMAP embedding of the integration of the cells from the two species. **d** Contribution of each species to the percentage of cells in the ventricular CM

2022). The low heterogeneity in integrating an organ datasets versus the whole organism datasets adds another level of complexity due to the low number of integration anchor points (Lähnemann et al. 2020; Argelaguet et al. 2021).

Therefore, we focused on the integrating cardiac FB of zebrafish and human datasets, since zebrafish FB have recently been shown to play a major role in cardiac regeneration (Fig. 5a, Supp Fig.4a) (Hu et al. 2022). The zebrafish FB consisted of 11 cell subpopulations, whereas the human FB had 7 subpopulations (Fig. 5b). The regenerating FB have been shown to form a new cell state 3 days post-injury, which consists of fibroblast (proliferating), fibroblast (*nppc*), fibroblast (*col11a1a*) and fibroblast (*col12a1a*) (Hu et al. 2022) (Supp Fig. 4b).

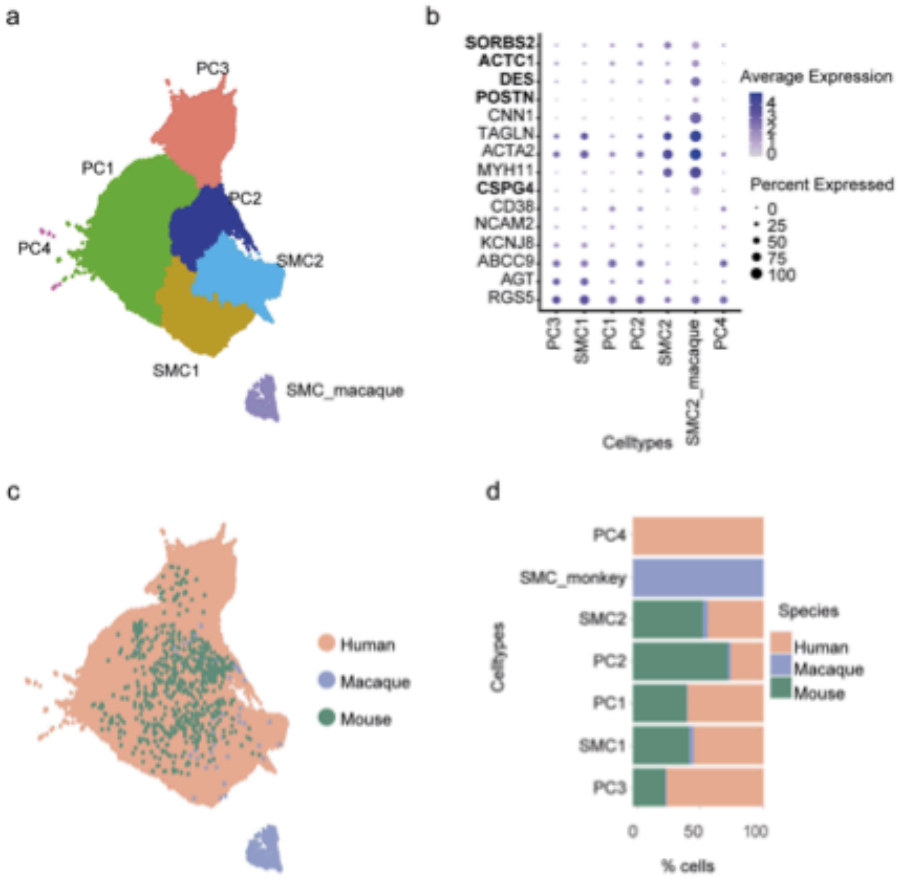


Fig. 3 Pericytes and smooth muscle cells. **a** 2D UMAP embedding of the cell subpopulations of PC and SMC after integrating the cells from the two species. **b** Dot plot of the cell type-specific marker genes. **c** 2D UMAP embedding of the integration of the cells from the three species. **d** Contribution of each species to the percentage of cells in the PC and smooth muscle cells

On hierarchical cluster tree analysis after integration, the cell subpopulations from human datasets were distant from the 4 zebrafish-specific cell states known for regeneration (Fig. 5c). Differential expression analysis of the genes involved in zebrafish heart regeneration (Hu et al. 2022) revealed that *COL12A1* was expressed in all the subpopulations of human FB, while FB4 and FB7 expressed *POSTN* and *DKK3*, respectively (Fig. 5d). All these genes were expressed by a very low percentage of cells and had an expression level lower in the human dataset than the regenerating zebrafish FB (Fig. 5d). Taken together, our data indicate that the human heart in contrast to the zebrafish does not seem to contain FB with a regenerative potential.

Discussion

Recent years have brought an enormous development of new single-cell technologies, enabling the discovery of subtle differences in species-specific molecular programs or the relative proportions of specific cell types at unprecedented resolution. The ever-growing collection of single-cell transcriptomic data deepens our knowledge of how organisms or more specifically particular organs are built and how they function. Here, we focused on the heart, a complex organ that pumps blood in various organisms developed throughout evolution, from fish to mammals. The obvious structural and functional differences that exist among different species were addressed from the sc-seq point of view, using

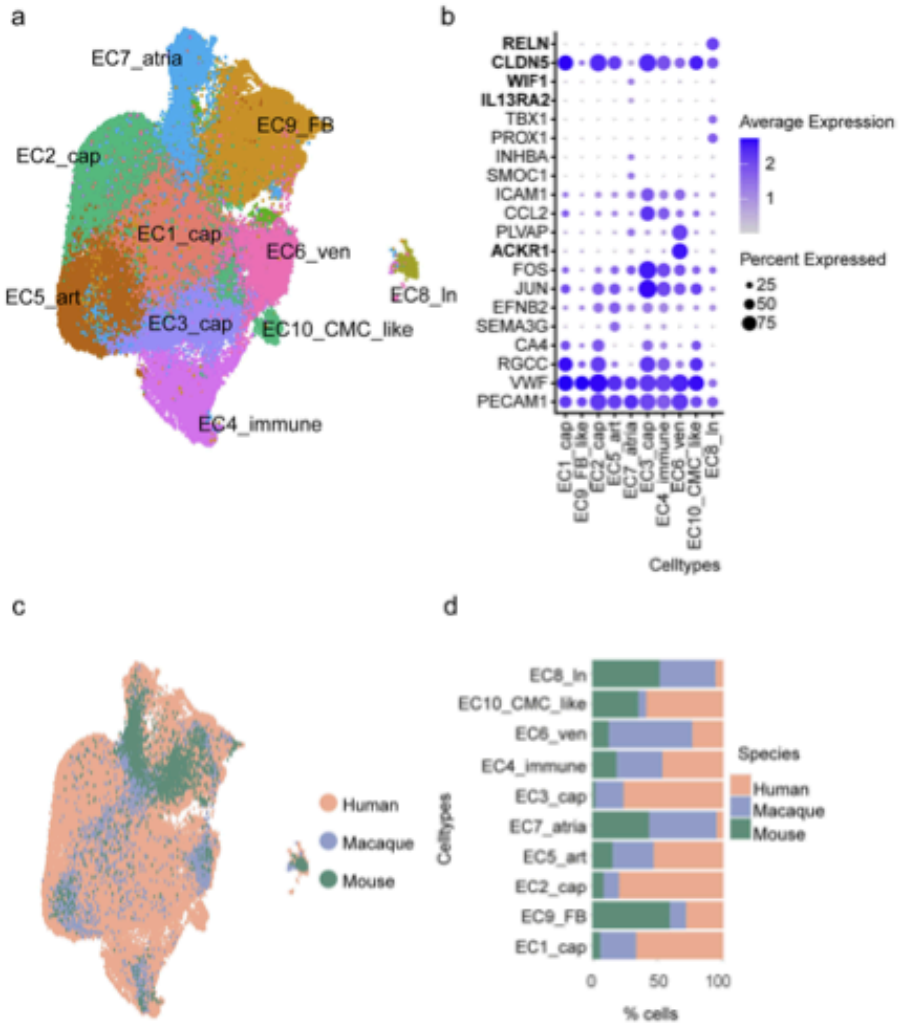


Fig. 4 Endothelial cells. **a** 2D UMAP embedding of the cell subpopulations of EC after integrating the cells from the 2 species. **b** Dot plot of the cell type-specific marker genes. EC_cap—capillary, EC_FB—fibroblast, EC_art—arterial, EC_atria—atrial, EC_ven—venous, EC_CMC_like—cardiomyocyte_like, EC_ln—lymphatic. **c** 2D UMAP embedding of the integration of the cells from the three species. **d** Contribution of each species to the percentage of cells in the endothelial cells

the human heart as a reference. By integrating single-cell transcriptomic data from adult human, macaque, mouse and zebrafish hearts considering the orthologous genes, i.e., those shared among these species, we aimed to find species-specific differences in molecular programs and cell type proportions.

While there were no obvious differences between humans and mice in the clustering of atrial CM, a mouse-specific cluster was observed for ventricular CM (vCM_mouse). Given that in four-chambered mammalian hearts the left ventricle pumps blood into the systemic circulation and that the resting heart rate in mice is approximately 10 times faster than in humans, we reason that these clustering differences could be attributed to the actual biological/functional differences (Wessels and Sedmera 2003). We observed a slightly higher expression of *Ndufa4*, *Ndufb11 and Cox7c* in the

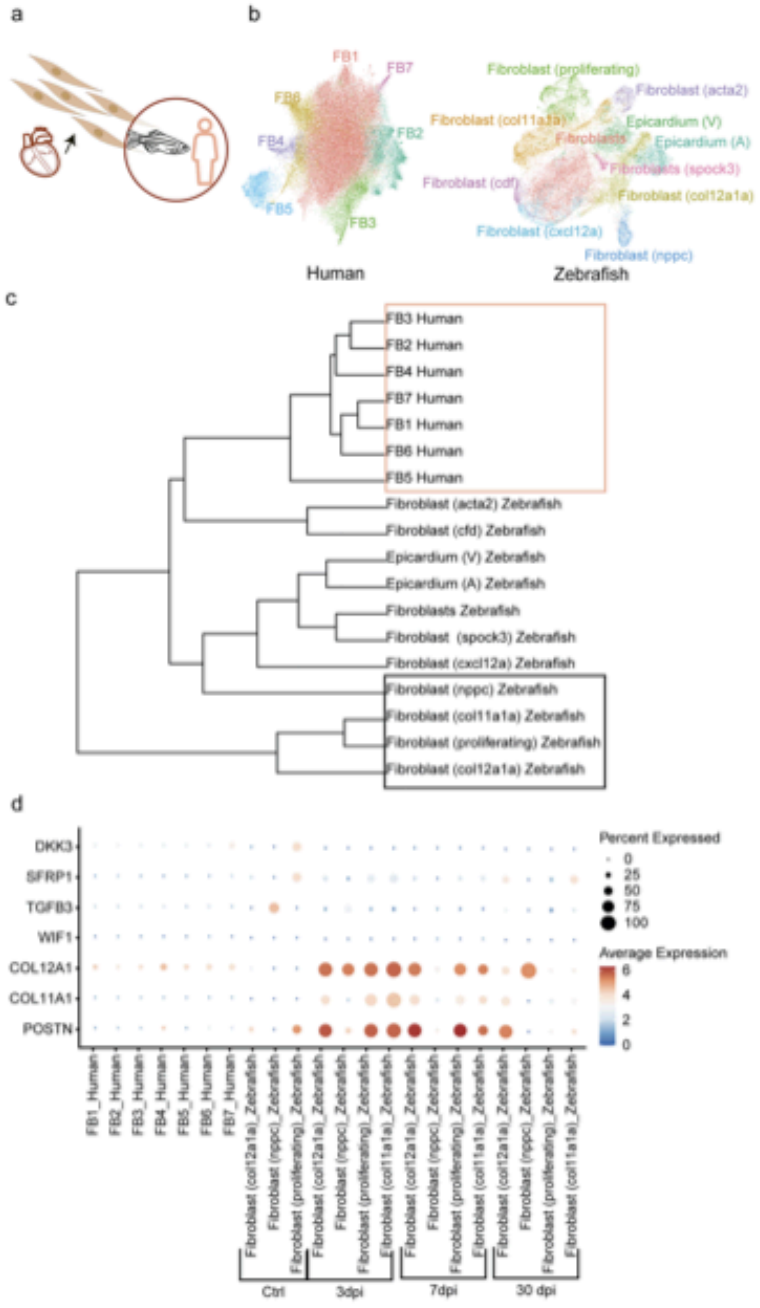


Fig. 5 Integration of human and zebrafish cardiac FB. **a** Single-cell integration of FB of heart datasets from human and zebrafish. **b** UMAP embedding of the integration of FB from human and zebrafish. **c** Hierarchical cluster tree analysis of the fibroblast subpopulations. The orange box encircles the populations of the human FB, while the black box encircles the key cell states involved in the regeneration of FB post-injury in zebrafish. **d** Dotplot representation of the expression of genes essential for regeneration of FB. The expression in zebrafish is split by control and number of days post-injury (dpi)

vCM_mouse cluster, which are nuclear-encoded mitochondrial genes and were previously associated with the vCM4 cluster in the human heart, characterized by a high energetic state (Litviňuková et al. 2020). Gene ontology reactome pathway analysis of the markers expressed in this cluster revealed that most of them participate in cardiac conduction (e.g., *Myh6, Pln, Kcnd2, Fgf13*). Furthermore, a mouse-specific gene *Prune2* was highly expressed in this cluster and besides showing heart-specific expression, this gene has been associated with energy metabolism in the mouse heart (Song et al. 2013). Interestingly, we also observed a low expression of *Myh7* in the mouse heart, a sarcomere gene specifically expressed in human ventricular CM. In rodents, expression of this gene is ventricular-specific during embryogenesis, but downregulated postnatally, so that in the adult mouse heart *Myh6* is the main myosin heavy chain gene, expressed both in the atria and ventricles (England and Loughna 2013). Mouse-specific cells (vCM_mouse) also expressed higher levels of *Pcdh7,* calcium-dependent adhesion molecule, as well as *Dlc1*, a Rho GTPase activating protein with tumor suppressor function that is essential for embryonic development (Durkin et al. 2005). Another striking difference between the two species is the expression of *PLCL1*, which is present in all vCM subpopulations of the human heart, but not in the mouse-specific cluster. This gene encodes a protein involved in inflammation, and an intronic variant has been associated with myocardial infarction (Lin et al. 2015; Hahn et al. 2020). Of note, cardiomyocyte clusters contain only human and mouse integration data, as the macaque dataset includes sequencing data of aortas and coronary arteries only.

Integration of PC and SMC was performed on datasets from all three mammalian species and our results show that the macaque-specific cluster (SMC_macaque) separates from the others. This cluster shows high expression of genes specific to the human SMC2 cluster (*MYH11, ACTA2, TAGLN, CNN1*), as well as *CSPG4*, a pericyte marker found also in SMCs (Murfee et al. 2005). Furthermore, this cluster expresses SMC genes specific to coronary arteries (*DES, ACTC1, SORBS2*) and the aortic arch (*POSTN*) in macaque, as described by Zhang et al., which might partially explain the separate clustering. Of note, the SMC_macaque cluster contains numerous ribosomal and mitochondrial genes, but even upon regressing them the cluster stands out.

Endothelial cells integrated well for all three mammalian species and we have not observed any species-specific clusters. However, clusters EC7_atria and EC8_ln consisted mainly of macaque and mouse cells (even though they were annotated based on the human data). In addition to the cluster-specific genes (*SMOC1, INHBA, NPR3*), EC7_atria cells were also found to express *IL13RA2* and *WIF1*, genes previously described to be specific for *M. fascicularis* EC originating from the aortic arch. In addition to the cluster-specific genes (*PROX1, TBX1, PDPN*), EC8 lymphatic cells showed high expression of *RELN*, in accordance with its specificity for the lymphatic ECs in *M. fascicularis* (Zhang et al. 2020). We have also observed that one of the marker genes for venous EC6_ven human cells—*ACKR1*—was characterized as a coronary artery-specific EC gene in the macaque dataset. Although our results show distinctively higher expression of this gene in the EC6_ven cluster as compared to other EC clusters, revisiting the single molecule fluorescent in situ hybridization (smFISH) image from the original publication showed its weak expression in arteries as well (Litviňuková et al. 2020).

Human and mouse FB integrated well, while macaque cells clumped together with the human FB3 cluster probably due to the sampling bias. The FB3 subtype of human FB was characterized by expression of the cytokine receptor genes (e.g., *OSMR, IL6ST*) and it was reported to be less abundant in the left ventricle as compared to other human fibroblast subtypes (Litviňuková et al. 2020). Consistent with this, the macaque heart dataset contains adventitial FB that make up the outermost layer of a blood vessel, composed mainly of collagen and elastic fibers secreted by FB.

Because of the evolutionary proximity, the three mammalian heart datasets integrated well, whereas the integration of the zebrafish heart dataset was challenging due to a low number of orthologous genes with the other three species, resulting in poor clustering. Therefore, we focused only on the heart FB, since they were recently shown to play a role in the heart regeneration process following injury in zebrafish (Hu et al. 2022). The regeneration-specific cell states in zebrafish (e.g., proliferating, *nppc*-, *col11a1a*-, and *col12a1a*-expressing FB) were more distant to the human cells than other zebrafish cell types, as shown by hierarchical cluster tree analysis. Importantly, these cell states exist only upon injury, and their gene expression patterns should be viewed in light of these events. Thus, the most accurate comparison of the two species would include only the non-injured heart samples, and splitting the zebrafish dataset (into control and different timepoints post-injury) enabled a better comparison to the human dataset. Our results show that all the genes relevant for regeneration in the zebrafish heart are weakly expressed in the non-injured human heart. Striking differences among the two species include expression levels of *POSTN, TGFB3, SFRP1,* and *DKK3*, all higher expressed in zebrafish. All these genes encode proteins that belong to TGFβ and WNT signaling pathways, which mutually interact and play key roles in fibrotic response (Akhmetshina et al. 2012).

Our study also has limitations: first, the heart tissue was sampled quite differently between mouse, human, and zebrafish compared to the macaque. Since other datasets do not sample aorta, the differences seen in the EC could have arised due to this cell population bias. Due to this sampling

bias towards aorta in the macaque, comparisons with other cell types should be interpreted with caution. Second, there are strong differences in cell numbers between the species. Third, no validation experiments were performed due to very limited sample access.

In summary, we have integrated single-cell transcriptomic heart datasets of four species, and while it was straightforward for those that are evolutionarily close, divergent species could also be compared by focusing on a selected cell type. The observed species-specific differences could be explained by taking into consideration multiple factors such as functional differences, origin of tissue, and gene nomenclature.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00335-022-09968-7.

Acknowledgements S.B. thank Chengxiang Qiu for the discussion regarding zebrafish integration.

Author contributions S.B. and M.S. designed the research. S.B. performed the computational analysis. S.B., J.P., V.K.A.S. and M.S. interpreted the results. S.B. and J.P. drafted the manuscript. S.B., J.P., V.K.A.S. and M.S. revised and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. M.S. is DZHK principal investigator and is supported by grants from the Deutsche Forschungsgemeinschaft (DFG) (Grant No. SP1532/3-2,SP1532/4-1 and SP1532/5-1), the Max Planck Society and the Deutsches Zentrum für Luft- und Raumfahrt (Grant No. DLR 01GM1925). J.P. is supported by a research grant from the University of Lübeck, Germany (Grant No. J14-2021).

Declarations

Conflict of interest The authors declare no competing interests.

References

Akhmetshina A, Palumbo K, Dees C et al (2012) Activation of canonical Wnt signalling is required for TGF-β-mediated fibrosis. Nat Commun 3:735

Alberts B, Johnson A, Lewis J et al (2002) Blood vessels and endothelial cells. In: Dries, David J. (eds) Molecular Biology of the cell, 4th edn. Garland Science

Argelaguet R, Cuomo ASE, Stegle O, Marioni JC (2021) Computational principles and challenges in single-cell data integration. Nat Biotechnol 39:1202–1215

Cao J, Spielmann M, Qiu X et al (2019) The single-cell transcriptional landscape of mammalian organogenesis. Nature 566:496–502

Cao J, O'Day DR, Pliner HA et al (2020) A human cell atlas of fetal gene expression. Science. https://doi.org/10.1126/science.aba7721

Durkin ME, Avner MR, Huh CG et al (2005) DLC-1, a Rho GTPase-activating protein with tumor suppressor function, is essential for embryonic development. FEBS Lett. https://doi.org/10.1016/j.febslet.2004.12.090

England J, Loughna S (2013) Heavy and light roles: myosin in the morphogenesis of the heart. Cell Mol Life Sci 70:1221

Hahn J, Fu YP, Brown MR et al (2020) Genetic loci associated with prevalent and incident myocardial infarction and coronary heart disease in the cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium. PLoS ONE. https://doi.org/10.1371/journal.pone.0230035

Hu B, Lelek S, Spanjaard B et al (2022) Origin and function of activated fibroblast states during zebrafish heart regeneration. Nat Genet. https://doi.org/10.1038/s41588-022-01129-5

Huang X, Henck J, Qiu C, et al (2022) Single cell whole embryo phenotyping of pleiotropic disorders of mammalian development. bioRxiv 2022.08.03.500325

Lähnemann D, Köster J, Szczurek E et al (2020) Eleven grand challenges in single-cell data science. Genome Biol 21:1–35

Lin YJ, Chang JS, Liu X et al (2015) Genetic variants in PLCB4/PLCB1 as susceptibility loci for coronary artery aneurysm formation in Kawasaki disease in Han Chinese in Taiwan. Sci. https://doi.org/10.1038/srep14762

Litviňuková M, Talavera-López C, Maatz H et al (2020) Cells of the adult human heart. Nature 588:466–472

Murfee WL, Skalak TC, Peirce SM (2005) Differential arterial/venous expression of NG2 proteoglycan in perivascular cells along microvessels: identifying a venule-specific phenotype. Microcirculation. https://doi.org/10.1080/10739680590904955

Qiu C, Cao J, Martin BK et al (2022) Systematic reconstruction of cellular trajectories across mouse embryogenesis. Nat Genet 54:328–341

Raudvere U, Kolberg L, Kuzmin I et al (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 47:W191–W198

Smajić S, Prada-Medina CA, Landoulsi Z et al (2022) Single-cell sequencing of human midbrain reveals glial activation and a parkinson-specific neuronal state. Brain 145:964–978

Song Y, Ahn J, Suh Y et al (2013) Identification of novel tissue-specific genes by analysis of microarray databases: a human and mouse model. PLoS ONE 8:e64483

Sreenivasan VKA, Balachandran S, Spielmann M (2022) The role of single-cell genomics in human genetics. J Med Genet. https://doi.org/10.1136/jmedgenet-2022-108588

Stephenson A, Adams JW, Vaccarezza M (2017) The vertebrate heart: an evolutionary perspective. J Anat 231:787–797

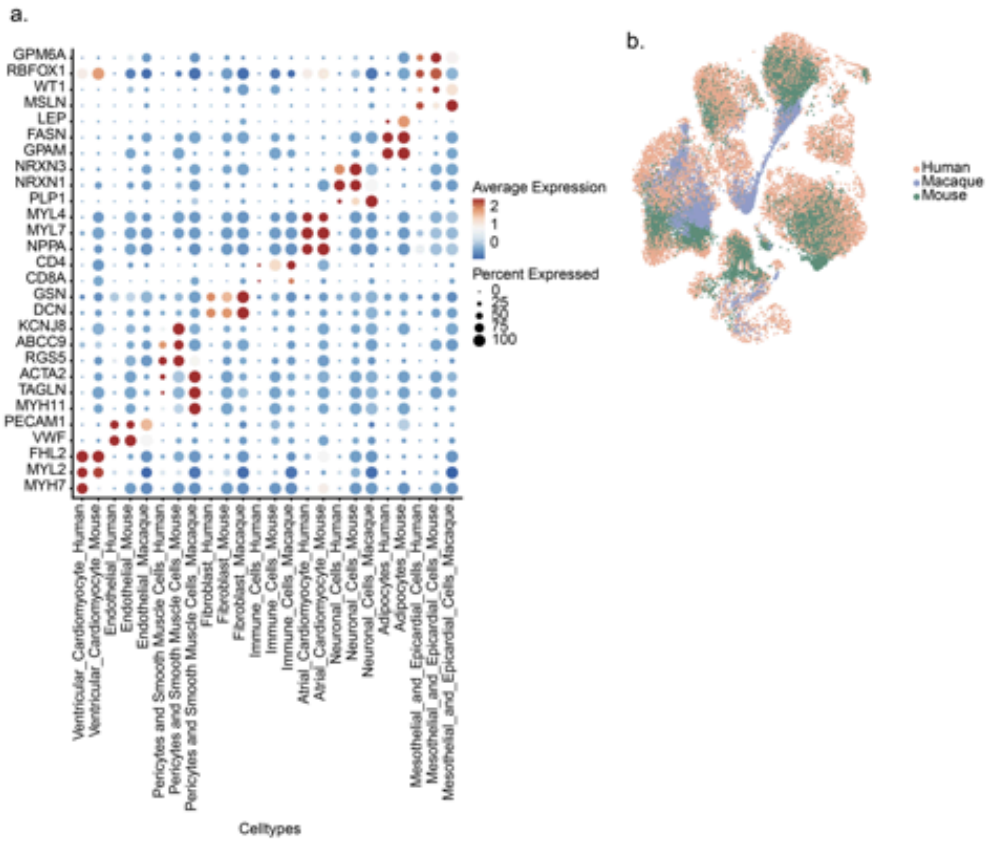
Tran HTN, Ang KS, Chevrier M et al (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 21:12

Vidal R, Wagner JUG, Braeuning C et al (2019) Transcriptional heterogeneity of fibroblasts is a hallmark of the aging heart. JCI Insight. https://doi.org/10.1172/jci.insight.131092

Wessels A, Sedmera D (2003) Developmental anatomy of the heart: a tale of mice and man. Physiol Genomics. https://doi.org/10.1152/physiolgenomics.00033.2003

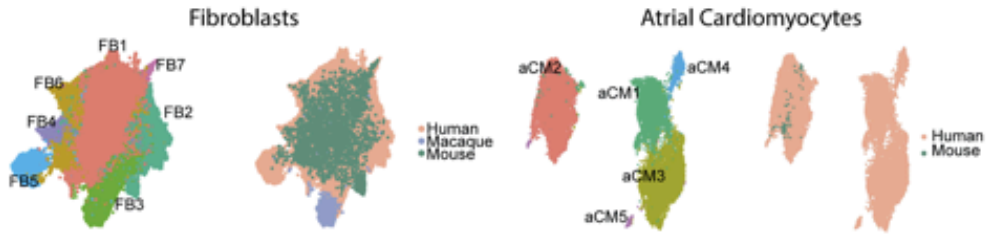
Zhang W, Zhang S, Yan P et al (2020) A single-cell transcriptomic landscape of primate arterial aging. Nat Commun 11:1–13

Supplementary material



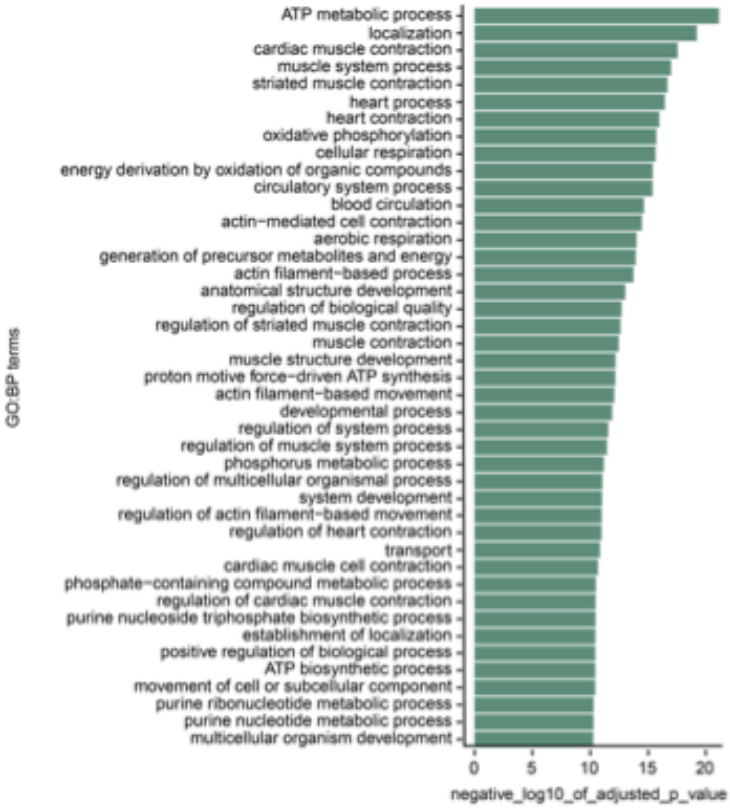
Supp Fig 1: Integration of single cell heart transcriptome data a) Dot plot of the cell type-specific marker genes, split across species. b) 2D UMAP embedding of the integration of the cells from the three species (human dataset downsampled to the size of mouse dataset).

a. b. c. d.

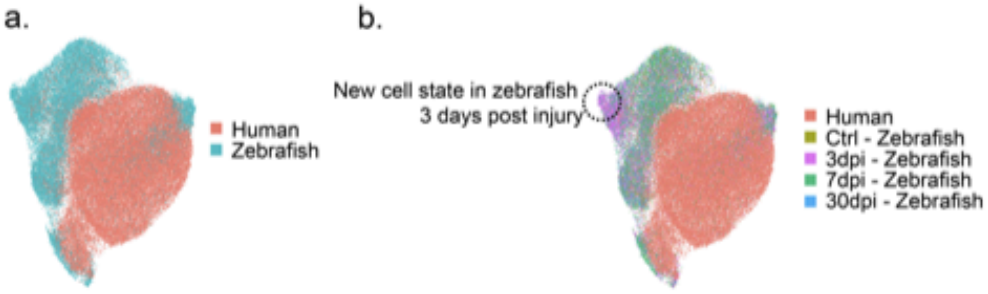


Supp Fig 2: Fibroblasts and atrial cardiomyocytes a) 2D UMAP embedding of the cell subpopulations of fibroblasts after integrating the cells from the three species. b) 2D UMAP embedding of the integration of the cells from the three species in fibroblasts. c) 2D UMAP embedding of the cell subpopulations of atrial cardiomyocytes after integrating the cells from

the three species. d) 2D UMAP embedding of the integration of the cells from the three species in atrial cardiomyocytes.



Supp Fig 3: Gene Ontology analysis of ventricular cardiomyocytes specific to mice.



Supp Fig 4: Integration of cardiac fibroblasts from human and zebrafish datasets a) 2D UMAP embedding of the after integrating the cells from human and zebrafish dataset. b) 2D UMAP embedding showing the new population of fibroblasts 3 days post injury.

3.2 STIGMA: Single-cell tissue-specific gene prioritization using machine learning.

Balachandran, Saranya, Cesar A. Prada-Medina, Martin A. Mensah, Naseebullah Kakar, Inga Nagel, Jelena Pozojevic, Enrique Audain, et al. American Journal of Human Genetics, 8 January 2024, S0002-9297(23)00443-3. https://doi.org/10.1016/j.ajhg.2023.12.011.

The study aimed at implementing a machine learning model which would learn gene expression signatures from single cell developmental datasets to prioritize candidate disease genes for congenital diseases. STIGMA uses feature input from scRNA-seq data and gene-intrinsic properties. From scRNA-seq data, we extracted gene-level metrics like mean, variance and fold change expression of gene and percentage of cells expressing the gene per cell type. Pseudo-temporal developmental trajectory was built per cell type to access the dynamics of expression along the maturation state of the cell type. We tested our approach using SVM and Random forest, of which we received a low precision with SVM, hence Random forest was selected as a model of choice. Random forest based supervised machine learning model was trained on known disease associated genes and tolerant housekeeping genes. This was based on the assumption that genes associated with congenital diseases have cell type specific expression (Tu et al. 2006). Thus STIGMA does not prioritize genes with syndromic phenotype. The model was trained on SMOTE-ADASYN based class balanced data on known seed genes associated with the disease (positive class) and tolerant housekeeping genes (negative class). A 5-fold cross validation approach was used to access the model performance. The model was implemented in two disease settings namely, congenital limb malformations and congenital heart disease.

STIGMA was trained on three mouse limb datasets, two were published and one was sequenced in the study. The model was trained on a positive class of known disease associated genes (n=87) from PanelApp and a negative class of tolerant housekeeping genes (n=643). Using SMOTE for class imbalance correction, resulted in both classes having 643 genes. The area under the curve (AUC) of receiver operator curve was 0.99

and at a threshold of 0.725, the model has a sensitivity of 0.9545 and a precision of 0.875, predicting 864 STIGMA candidate genes (SCGs). We saw enrichment of SCG's in limb associated phenotypes by the Monarch Initiative (Mungall et al. 2017) with a Fisher's exact test p-value of 5.5e-14 and 2.3e-6 in human and mouse respectively. In a cohort of 69 patients with congenital limb malformation harboring 7,082, potential non-structural loss of function variants, STIGMA prioritized 345 genes having 469 variants. Of these, *UBA2* had *de novo* pathogenic variants and *DUS2, MAPK1, F11R, PHIP* had *de novo* variants. Another interesting finding in the study was two genes *UBA2* and *PHIP* associated with similar disease profiles ectrodactyly and oligodactyly had similar temporal expression pattern in the mesenchymal-chondrocytes, -fibroblasts, ectodermal-sost, and muscle cells.

We wanted to implement STIGMA on congenital heart disease. Given the availability of human fetal atlas, we used this dataset to acquire the scRNA-seq feature input for the model. The model was trained on a positive class from a manually curated gene list (Audain et al. 2021) (n=36), which was the resultant of filtering ubiquitously expressed genes against the sc-seq mouse organogenesis dataset (Cao et al. 2019) and a negative class of tolerant housekeeping genes (n=643). As before the class was balanced with SMOTE. Upon cross validation, we obtained a ROC curve with AUC of 0.9972. At a threshold of 0.57, the sensitivity and precision were 0.8333 and 0.8421 respectively. We validated our predictions with the published literature. In a CHD cohort of 7,958 individuals harboring 4,190 *de novo* variants, we predicted 468 genes with 543 variants to be candidate disease associated genes. Of the predicted genes, 34 genes had non synonymous *de novo* variants in two or more individuals, among which 10 genes had heart phenotypes reported.

In both implementations of STIGMA, we saw features related to scRNA-seq were significant in the prediction of the candidate disease genes. Overall, our findings demonstrate that STIGMA effectively prioritizes tissue-specific candidate genes by utilizing scRNA-seq data.

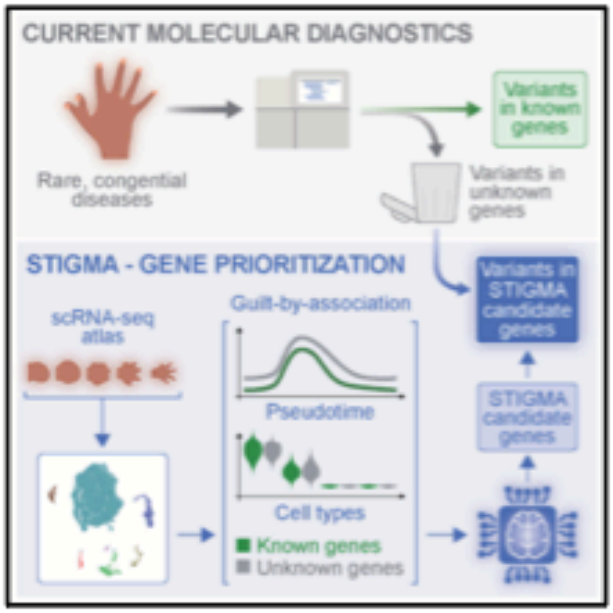
3.2.1 Project contribution

Malte Spielmann and I designed the research. Cesar A. Prada-Medina, Martin Kircher, and Varun K.A. Sreenivasan gave advice on the research design. Juliane Glaser generated the in-house limb gene expression data. Limb gene expression data were analyzed by Cesar A. Prada-Medina and me. I developed the gene prioritization machine learning model. Martin A. Mensah, Naseebullah Kakar, Inga Nage, Jelena Pozojevic, Enrique Audain, Marc-Phillip Hit,Varun K.A. Sreenivasan, Malte Spielmann and I interpreted the results. Varun K.A. Sreenivasan and I drafted the manuscript. Cesar A. Prada-Medina, Martin A. Mensah, Naseebullah Kakar, Inga Nage,Jelena Pozojevic, Enrique Audain, Marc-Phillip Hit, Martin Kircher, Varun K.A. Sreenivasan, Malte Spielmann and I revised and approved the final manuscript.

ARTICLE

STIGMA: Single-cell tissue-specific gene prioritization using machine learning

Graphical abstract



Authors

Saranya Balachandran, Cesar A. Prada-Medina, Martin A. Mensah, ..., Martin Kircher, Varun K.A. Sreenivasan, Malte Spielmann

Correspondence

varun.sreenivasan@uksh.de (V.K.A.S.), malte.spielmann@uksh.de (M.S.)

Single-cell tissue-specific gene prioritization using machine learning (STIGMA) is an approach to prioritize candidate genes for congenital diseases. STIGMA uses single-cell RNA-seq data to capture the dynamics of gene expression within cell populations across developmental time, making it a powerful tool for the discovery of disease-associated genes.

Balachandran et al., 2024, The American Journal of Human Genetics *111*, 338–349
February 1, 2024
https://doi.org/10.1016/j.ajhg.2023.12.011

CellPress

ARTICLE

STIGMA: Single-cell tissue-specific gene prioritization using machine learning

Saranya Balachandran,[1] Cesar A. Prada-Medina,[2,11] Martin A. Mensah,[3,4,5] Juliane Glaser,[10] Naseebullah Kakar,[1,6] Inga Nagel,[1] Jelena Pozojevic,[1] Enrique Audain,[7,8,9] Marc-Phillip Hitz,[7,8,9] Martin Kircher,[1] Varun K.A. Sreenivasan,[1,*] and Malte Spielmann[1,2,8,*]

Summary

Clinical exome and genome sequencing have revolutionized the understanding of human disease genetics. Yet many genes remain functionally uncharacterized, complicating the establishment of causal disease links for genetic variants. While several scoring methods have been devised to prioritize these candidate genes, these methods fall short of capturing the expression heterogeneity across cell subpopulations within tissues. Here, we introduce single-cell tissue-specific gene prioritization using machine learning (STIGMA), an approach that leverages single-cell RNA-seq (scRNA-seq) data to prioritize candidate genes associated with rare congenital diseases. STIGMA prioritizes genes by learning the temporal dynamics of gene expression across cell types during healthy organogenesis. To assess the efficacy of our framework, we applied STIGMA to mouse limb and human fetal heart scRNA-seq datasets. In a cohort of individuals with congenital limb malformation, STIGMA prioritized 469 variants in 345 genes, with *UBA2* as a notable example. For congenital heart defects, we detected 34 genes harboring nonsynonymous *de novo* variants (nsDNVs) in two or more individuals from a set of 7,958 individuals, including the ortholog of *Prdm1*, which is associated with hypoplastic left ventricle and hypoplastic aortic arch. Overall, our findings demonstrate that STIGMA effectively prioritizes tissue-specific candidate genes by utilizing single-cell transcriptome data. The ability to capture the heterogeneity of gene expression across cell populations makes STIGMA a powerful tool for the discovery of disease-associated genes and facilitates the identification of causal variants underlying human genetic disorders.

Introduction

The widespread introduction of next-generation sequencing approaches has rendered the analysis of genes a routine in the clinical setting. It has benefited the ongoing discovery, functional annotation, and disease mappings of genes (e.g., HPO,[1] OMIM[2])[3] as well as improvements in tools and resources to call, annotate, prioritize, and filter variants within these genes (e.g., gnomAD,[4] DECIPHER[5]). As a result, the diagnostic yield with genome or exome sequencing has been steadily increasing, recently reaching 41%.[3] However, to date, a causal disease link has been established for variants in only about 5,000 genes.[2,6] Consequently, many potentially deleterious variants in genes of unknown function are classified as variants of uncertain significance (VUSs) and do not contribute to a diagnosis of rare diseases, until further validated by experimental verification, e.g., using *in situ* hybridization. In other words, incomplete gene-disease associations remain a significant bottleneck in finding a molecular diagnosis in individuals with rare genetic diseases. Gene prioritization can help overcome this limitation.[7,8]

Gene prioritization refers to arranging genes in the order of probability of association with a disease. It can help narrow down the list of candidate genes under consideration. Gene prioritization usually requires prior knowledge about the genes, including (1) a list of seed genes that are known to be associated with the disease and (2) data on the genes/proteins, such as protein-protein interactions, gene expression profiles, known functional annotations (ontology, pathways etc.), disease-gene associations,[9] and intrinsic gene properties (genomic position, sequence, GC content, conservation, structure, etc.).[10] A computational model then assigns a "disease-causing" probability to every gene either based on existing annotations for that gene or based on "guilt by association" with known disease-associated genes in interacting networks or machine learning models.[7] The tools that rely on functional annotations or disease associations of the gene being prioritized are often heavily biased toward highly characterized genes.[8] Such methods have also been reported to yield false positive predictions due to evolving disease-gene associations.[11] In contrast, tools such as GeneFriends,[12] GADO,[13] EvoTol,[14] and GeneTIER[15] that rely exclusively on gene expression data,

[1]Institute of Human Genetics, University Hospital Schleswig-Holstein, University of Lübeck and Kiel University, Lübeck, Germany; [2]Human Molecular Genetics Group, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; [3]Institut für Medizinische Genetik und Humangenetik, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Augustenburger Platz 1, 13353 Berlin, Germany; [4]BIH Charité Digital Clinician Scientist Program, BIH Biomedical Innovation Academy, Anna-Louisa-Karsch-Strasse 2, 10178 Berlin, Germany; [5]RG Development & Disease, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; [6]Department of Biotechnology, BUITEMS, Quetta, Pakistan; [7]Institute of Medical Genetics, Carl von Ossietzky University, 26129 Oldenburg, Germany; [8]DZHK e.V. (German Center for Cardiovascular Research), Partner Site Hamburg/Kiel/Lübeck; [9]Department of Congenital Heart Disease and Pediatric Cardiology, University Hospital of Schleswig-Holstein, 24105 Kiel, Germany; [10]Development and Disease, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany
[11]Present address: Novo Nordisk Research Center Oxford, Innovation Building, Oxford OX3 7FZ, UK
*Correspondence: varun.sreenivasan@uksh.de (V.K.A.S.), malte.spielmann@uksh.de (M.S.)
https://doi.org/10.1016/j.ajhg.2023.12.011.

evolutionary intolerance, and intrinsic gene data are inherently unbiased. For example, GeneTIER is based on the hypothesis that "genes responsible for a tissue-specific phenotype are expected to be more highly expressed in affected than unaffected tissues."[15,16] Such annotation-agnostic tools can prioritize candidate genes lacking functional annotations.[10] However, most current gene expression-based prioritization tools use bulk-RNA sequencing (bulkRNA-seq) data (e.g., GTEx[17]) containing expression profiles at an organ-level resolution.[18] This introduces two major issues with regards to the specificity of gene expression. Firstly, the expression of cell type-specific genes are averaged out in these datasets. Secondly, such approaches do not explicitly consider the temporal dynamics of expression, which is crucial during organogenesis. In the context of diagnosing a rare congenital disease, this can lead to the current approaches being non-specific and insensitive. For instance, the inability to predict a known disease-gene association in the case of Parkinsonism-dystonia (MIM: 613135) was attributed by the authors to the highly cell type-specific expression of *SLC6A3* (MIM: 126455).[13] Using cell type- and developmental time-specific gene-expression data could improve the gene prioritization outcome.

The boom of single-cell sequencing (sc-seq) has enabled the creation of cell atlases of humans and model organisms, providing reference maps with cell types, cell states, their gene expression profiles, spatial location, and chromatin profiles throughout embryogenesis and adulthood.[19–22] The technology has enabled a more in-depth analysis of molecular mechanisms throughout a lifetime (i.e., from embryogenesis through birth to old age) in states of health and disease at cellular resolution and is transforming healthcare.[23–28] These cell atlases are already being used to prioritize variants or to establish variant-to-function mappings.[29] However, to the best of our knowledge, single-cell RNA sequencing (scRNA-seq) data have not yet been applied for gene prioritization, where the cell type-specific or developmental stage-specific expression profiles are taken into consideration. Arguably, the only exception is a recently published risk gene identification method, VBASS.[30] VBASS uses scRNA-seq data to identify disease-associated genes from *de novo* variant data from large cohorts. In contrast, the goal of gene prioritization as discussed here is to narrow down the list of candidate genes in rare disorders or a single individual.

Here, we introduce scRNA-seq data-based gene prioritization for congenital diseases by developing single-cell tissue-specific gene prioritization using machine learning (STIGMA). STIGMA predicts the disease-causing probability of genes based on their expression profiles across cell types, while considering the temporal dynamics during the embryogenesis of a healthy (wild-type) organism, as well as several intrinsic gene properties. We validate our approach by applying the model on mouse limb and human fetal heart scRNA-seq datasets, to prioritize genes for congenital limb malformations and congenital heart disease (CHD), respectively. STIGMA successfully predicted several gene-disease associations, such as *UBA2* (MIM: 613295), which was recently reported to be related to limb malformations,[31] as well as *ALDOB* (MIM: 612724) and *MMP9* (MIM: 120361) that have been associated with ventricular septal defect (MIM: 614429).[32,33] It also suggested *PRDM1* (MIM: 603423), the ortholog of which has been shown to be associated with hypoplastic left ventricle and hypoplastic aortic arch in mouse models (MGI: J:175213).

Material and methods

Preparation of mouse limb scRNA-seq data

All of the following steps were carried out using cellranger (v.3.0, 10× Genomics),[34] scrublet (v.3),[35] seurat (v.3),[36] biomaRt (v.2.46.3),[37] splines (v.4.0.0),[38] and monocle3[39] as well as standard packages for R (v.4.0.5) and python (v.3.7.4).

The wild-type mouse scRNA-seq data are a combination of a dataset generated in this study and published datasets. The generated data originate from forelimbs (E9.5 to E12.5) and hindlimbs (E11.5 to E12.5) and it was combined with published scRNA-seq datasets of the forelimb between time points E10.5 and E15.0 from ENCODE accession ENCSR713GIS[40] (fastq files) and of the hindlimb between time points E11.5 and E18.5 from GEO accession GEO: GSE142425[41] (gene-barcode UMI count matrices).

When the UMI count matrix was not available, cellranger[34] was used with default parameters to generate it from the fastq files. Scrublet[35] was used to detect doublets and only cells with doublet scores below 0.2 were retained for the analysis. Further, only cells with more than 1,000 UMI and 500 genes and less than 10% mitochondrial DNA and 50% of ribosomal gene content were retained. Ribosomal and mitochondrial genes were removed for calculating cell embeddings. The data were normalized using *SCTransform* function in seurat[36] with 6,000 highly variable genes (hvgs). At this point, the datasets from the three sources were integrated to remove batch effects using the built-in integration pipeline in seurat[35,36,42] based on 1,000 genes as integration anchors. Principal component (PC) analysis based on the top 1,000 hvgs was performed on the integrated data to reduce the dimensionality. The nearest neighbors cell-cell graph built using the top 50 PCs was clustered using the Louvain algorithm,[43] with a resolution of 0.05. Cell-type marker genes were identified by differential expression (DE) analysis using the ROC approach implemented in the *FindAllMarker* function in seurat. DE analysis was performed on genes passing the cut offs of average fold change (|avg_logFC| > 0.25) and percentage of cells expressing the gene per cluster (min.pct > 0.1). The DE genes were used to annotate the main clusters. The clusters (immune cells, neuronal cells, vascular cells, and erythrocytes) that represented less than 4% of the data and those that were deemed not to generate limb-specific congenital malformations were removed. The remaining clusters were further sub-clustered (muscle cells: nhvg = 500, npcs = 20; ectoderm: nhvg = 500, npcs = 20; mesenchyme: nhvg = 1,000, npcs = 35) and annotated as before. Several characteristics of gene expression were also calculated using seurat to be used as STIGMA-classification features for each gene. These included mean expression in each sub-cluster (*AverageExpression*), variance in expression within each sub-cluster (*HVFInfo*), the percentage cells expressing the gene in each sub-cluster (*PrctCellExpringGene*), and the fold-change in expression between each sub-cluster and

the rest of the cells (*FoldChange*). Only genes that had an average expression greater than 0 in at least 1 of the cell types were retained.

Trajectory analysis to capture the gene expression dynamics was performed separately for each sub-cluster using the monocle3[39] workflow. The cells were ordered using *order_cells* with the earliest embryonic time point set as the root. The resulting pseudo time data were pooled into 20 bins and the average expression of the genes in each of these bins was calculated. To adapt this temporal data into a feature for random forest classification, it was fitted to a cubic spline function with 10 control points using the *bs* function of the splines package. The coefficients of the spline were obtained for each of the genes per cluster by solving the least squares fit and used as input features for the model.

Preparation of human fetal heart scRNA-seq data

Analyses were carried out using Seurat (v.4), splines (v.4.0.0), monocle3[39] as well as standard packages for R (v.4.0.5) and python (v.3.7.4). The human cell atlas of fetal gene expression consisted of 101,748 cells from 121 human fetal samples with data from the heart, ranging from 90 to 122 days post-conception.[19] Data were downloaded as a loom file and contained 16 annotated cell types. Only the cell types representing at least 1.5% of the data were retained. The remaining processing of the dataset, like calculating the gene features per cluster, was identical to that of mouse limb scRNA-seq data described above.

Intrinsic gene properties as features for classification

Processing steps were carried out using the R packages biomaRt (v.2.46.3), GenomicFeatures (v.3.10.0),[44] BSgenome.Hsapiens.UCSC.hg38 (v.1.4.3),[45] and Repitools v.1.36.0[46] for R (v.4.0.5). Gene constraints such as pLI, pNull, pRec, syn_Z, mis_Z, and lof_Z metrics for protein-coding genes were downloaded from gnomAD (v.2.1.1).[4] When absent and for non-coding genes, these metrics were imputed (see steps 1–3 in classification pipeline). To estimate the GC content of each gene and its upstream promoter region, the list of known genes for the human genome build hg38/GRCh38 was obtained using the BSgenome.Hsapiens.UCSC.hg38 library. For every gene, the promoter sequence, spanning 500 base pairs upstream and 100 base pairs downstream of the transcription start site, was obtained using the *promoters* function on the BSgenome.Hsapiens.UCSC.hg38 object. The percentage GC content in the gene and the promoter sequences were separately estimated using *gcContentCalc* and used as classifier features for the genes. Additionally for the limb dataset, mouse human ortholog confidence (BioMart) was included as input feature.[37,47]

Positive and negative classes for congenital limb malformation

The green list of genes associated with "Limb Disorders" (PanelApp v.2.0, downloaded on 23 June 2021)[48] was filtered to include only genes that show cell type specificity in expression. The average expression of the genes was quantified for each sub-trajectory within epithelial, hepatic, and mesenchyme trajectories in the mouse organogenesis cell atlas.[23] If a gene had the same expression (SD ± 1) in more than 10 sub-trajectories, they were filtered out. The negative training set was composed of housekeeping genes that were LoF tolerant based on gnomAD (pNull > pRec and pNull > pLI).[49,50]

Positive and negative classes for congenital heart disease

A curated list of genes known to be associated with congenital heart disease,[51] whose average expression was not ubiquitous across epithelial, hepatic, mesenchyme trajectories in the mouse organogenesis cell atlas,[23] was used for the positive class of the training set. As before for the predictions on the limb dataset, housekeeping genes that were LoF tolerant based on gnomAD (pNull > pRec and pNull > pLI) were used as the negative training set.[49,50]

Classification pipeline

The following steps were carried out using the sklearn (v.0.24.2)[52] package for python (v.3.7.4). A pipeline was set up using the *make_pipeline* function to optimize the parameters of the classifier. The classification workflow consisted of the following steps: (1) iterative imputing, (2) scaling, (3) synthetic oversampling, and (4) generating the random forest model. Missing data in the dataset were imputed using the *IterativeImputer* from scikit-learn with default parameters. The data were scaled using the *MinMaxScaler*. The class imbalance in the positive and the negative classes was corrected by synthetic minority over-sampling using an adaptive synthetic (SMOTE-ADASYN) algorithm.[53] This algorithm was chosen because it creates a synthetic representative dataset rather than simply duplicating the minor dataset. The best parameters for synthetic oversampling (n_neighbors) and the random forest model (n_estimators, max_depth, min_samples_split, min_samples_leaf) were optimized using *GridSearchCV* based on recall (for congenital limb malformations: adasyn: n_neighbor = 10, randomforest: n_estimators = 130, max_depth = 15, min_samples_split = 2, min_samples_leaf = 1 and for congenital heart disease: adasyn: n_neighbor = 5, randomforest: n_estimators = 90, max_depth = 30, min_samples_split = 5, min_samples_leaf = 1).

The final random forest model was built based on these optimized parameters and bootstrap resampling. Features that were significant for the performance of the model were obtained using the attribute *feature_importances_*. 5-fold cross-validation was used to calculate the out-of-bag error to validate the model and to avoid overfitting. The trained model was used to classify all genes. Those represented in the training classes were later removed from the predicted list. The area under the curve and other ROC metrics were calculated using the *roc_curve* function of sklearn.metrics. The threshold was chosen by plotting the density graph of the validation dataset (Figures 2F and 3D). The probability at which the negative class was at 0 density was chosen as the threshold. It is worth noting that the duplicated use of the same dataset for parameter optimization and validation likely leads to slightly inflated ROC metrics. The relatively small number of high-confidence positive class genes made the creation of a dedicated hold-out set for model validation impractical. However, this limitation can be overcome in the future as more genes acquire phenotypic annotations.

UMAP embedding of training classes based on input features

The input data were imputed, scaled, and class balanced as stated before. The UMAP object was constructed using the UMAP library of python. The *fit_transform* method of the UMAP class learns the embedding and transforms it to a numpy array, which is then plotted using the scatterplot method of plotly.

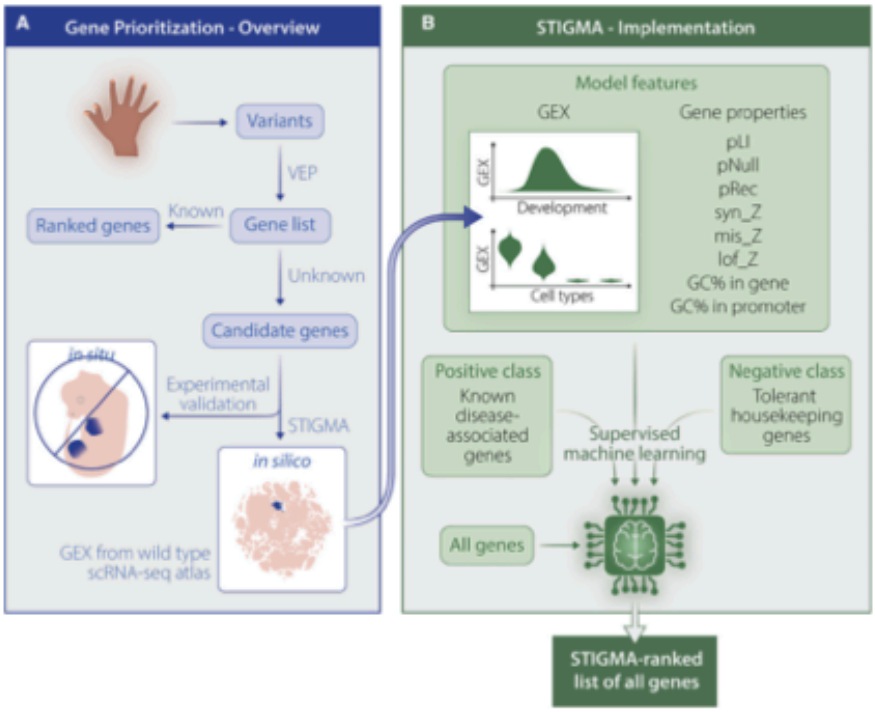


Figure 1. Implementation of gene prioritization within STIGMA for congenital diseases
(A) The genetic diagnostic workflow for congenital diseases (e.g., limb malformations) comprises the detection of variants and their prioritization, often resulting in many candidate genes that necessitate experimental validation. STIGMA enables the prioritization of the candidate genes with the use of development cell atlases of wild-type model organisms.
(B) In STIGMA, supervised machine learning is applied to the single-cell gene expression data as well as intrinsic gene properties (e.g., pLI, lof_z) on positive and negative classes. The probability of pathogenicity is then predicted for all genes (including genes lacking functional annotations) resulting in a ranked list of genes. GEX represents gene expression.

Explorative analysis based on Monarch Initiative

All gene phenotypes were downloaded from the Monarch Initiative Explorer.[54] Disease-specific ontology terms were downloaded from MouseMine.[55] Fisher's exact test was performed to verify the significance of the association between phenotype and STIGMA ranking.

Results

STIGMA model setup

Since we set out to predict the probability of every gene to be associated with the disease of interest, including those with little or no prior functional annotation, STIGMA was designed to use only scRNA-seq data and gene-intrinsic properties as features for model training and prediction (Figure 1A). The scRNA-seq data from wild-type samples during embryonic development were obtained from published datasets as well as datasets generated in this study. Gene expression in these datasets was encoded at the cell cluster-level to represent cell type specificity and developmental dynamics. Gene-level metrics per cluster included mean, variance, fold change compared to the rest of the cells, and the fraction of expressing cells. Developmental dynamics were captured by organizing the cells along a pseudo-temporal developmental trajectory and aggregating the gene expression along pseudo-time bins.

The gene-intrinsic properties included gene constraint metrics from gnomAD and GC content of the gene as well as its promoter. The gene constraint metrics used were related to the genes' (in)tolerance to LoF, synonymous, or missense variants, specifically the pLI (probability of being intolerant to LoF heterozygous variants), pRec (probability of being intolerant to LoF homozygous variants), pNull (probability of being tolerant to LoF variants), syn_Z (*Z* score of the number of synonymous variants in gene), mis_Z (*Z* score of the number of missense variants in gene), and lof_Z (*Z* score of the number of LoF variants in gene)[49] scores.

The supervised learning of these features in STIGMA was implemented using a random forest classifier[52,56] (Figure 1B). This machine learning algorithm has been widely used in disease classification due to its ensemble

property that allows combining predictions from multiple decision trees and due to its interpretability.[57,58] This choice was also based on our preliminary tests on other algorithms, such as support vector machines, which yielded suboptimal validation outcomes (e.g., precision = 0.413). The model was trained on the aforementioned features using two classes of genes: (1) a positive class, composed of genes known to be associated with the disease of interest and (2) a negative class, composed of housekeeping[50] genes that were more probable to be "tolerable" to LoF than being intolerant to homozygous or heterozygous LoF (i.e., pNull > pRec and pNull > pLI).[49] Due to the ubiquitous expression of housekeeping genes, STIGMA will likely not prioritize genes with syndromic phenotypes. Conversely, congenital diseases, which are the focus of STIGMA, are most likely caused by the LoF of genes crucial to the development of a distinctive organ and will likely exhibit increased temporal and/or tissue-level expression specificity.[59] The model performance in terms of accuracy, sensitivity, and precision was evaluated using a 5-fold cross-validation approach. Separate models were generated to prioritize genes for each of the two congenital disease groups, with disease-specific positive class and the associated model features.

STIGMA for congenital limb malformations

First, we trained STIGMA to predict genes associated with congenital limb malformations. Congenital limb malformations were chosen since the diagnostic yield is currently quite low, at less than 20%, and candidate genes are likely to have a distinct cell type-specific expression in the limb. scRNA-seq data were compiled from three mouse limb datasets, two published[40,41] and one from this study across embryonic days E9.5 to E18.5, spanning the period of limb development from the appearance of limb buds to interdigital separation and the completion of the limb outgrowth.[60] The data represented a total of 151,444 cells and 40,098 genes of which 19,571 had a human ortholog. Standard analysis including dimensionality reduction, clustering, and differential gene expression analysis revealed seven main cell types (Figures 2A and 2B), which were annotated based on marker genes (Figure 2C). Next, we reduced the dataset to contain only mesenchyme, ectoderm, and muscle cells by removing immune cells, neuronal cells, vascular cells, and erythrocytes, which have not been described to cause limb-specific congenital morphological malformations.[61] The final dataset contained 144,266 cells. Further sub-clustering to increase the cell type specificity of gene expression profiles led to two ectoderm sub-clusters and four mesenchyme sub-clusters, which were manually annotated (Figure S1). Pseudobulk gene expression of every gene was calculated per sub-cluster at several pseudo-time bins (Figure 2D).

The positive class of genes (n = 88) was a subset of the diagnostic-grade "green" list of genes in the panel "Limb Disorders" from the Genomics England PanelApp.[48] We removed genes that showed pervasive expression in all trajectories in the mouse organogenesis cell atlas (MOCA) (Figure S2),[23] resulting in 87 genes in the positive class. Tolerant housekeeping genes (643 genes) were used for the negative class (Table S1 containing gene lists of both classes). Class imbalance-correction by SMOTE resulted in a size of 643 for both classes. To verify whether positive and negative classes segregate based on the selected model features, we visualized the genes by projecting all the input features onto a 2D uniform manifold approximation and projection (UMAP), which showed a clear segregation between the two classes (Figure 2E), suggesting that a classification based on the features included in the model was appropriate.

Next, we used the positive and negative training classes to optimize the hyperparameters of the classifier using GridSearchCV. The hyperparameter-optimized model was trained using 5-fold cross validation. The receiver operator characteristic (ROC) curve, where the sensitivity (true positive rate) is plotted against 1-specificity (false positive rate), had an area under the curve (AUC) of 0.99 (Figures 2F and 2G). At a threshold STIGMA score (disease-causing probability) of 0.725, the sensitivity and the precision of the binary classifier reached 0.9545 and 0.875, respectively. Application of the final model trained on single-cell features on all genes resulted in 864 STIGMA-predicted candidate genes (SCGs) associated with congenital limb malformations with STIGMA scores greater than 0.725 (Table S2).

Since the random forest model lends itself to the analysis of relative importance of the various features that contribute to the classifier, we wondered to what extent the single-cell features influenced the model. Including pseudotime features, the single-cell features had a feature importance mean square of 3.25 to contrast with a value of 0.01 for gene-intrinsic properties (Figure S3A). In other words, the STIGMA score that each gene receives is based on the cell type-specific temporal dynamics in gene expression and, to a smaller extent, is based on the gene-intrinsic metrics, including the population-level constraint metrics. We also confirmed the importance of single-cell data for the performance of STIGMA by training on pseudobulkRNA-seq data generated from the same dataset. This resulted in a dramatic drop in performance, leading to the misclassification of nearly 40% of the positive class genes. Moreover, as can be expected from the feature importance plot, the cell type-specific expression alone is also insufficient to classify the genes (Figure S4). Together, these analyses indicated the combined importance of cell type-specific and pseudotime-specific gene expression information.

We verified these SCGs by several means. Firstly, we systematically explored the phenotypes reported for the SCGs and non-SCGs by the Monarch Initiative,[54] a portal for genotype-phenotype data across multiple species, with the rationale to expect enrichment of genes with limb-associated phenotypes in the top-ranking STIGMA genes. Indeed, this analysis (Figure 2H; Table S3) showed a

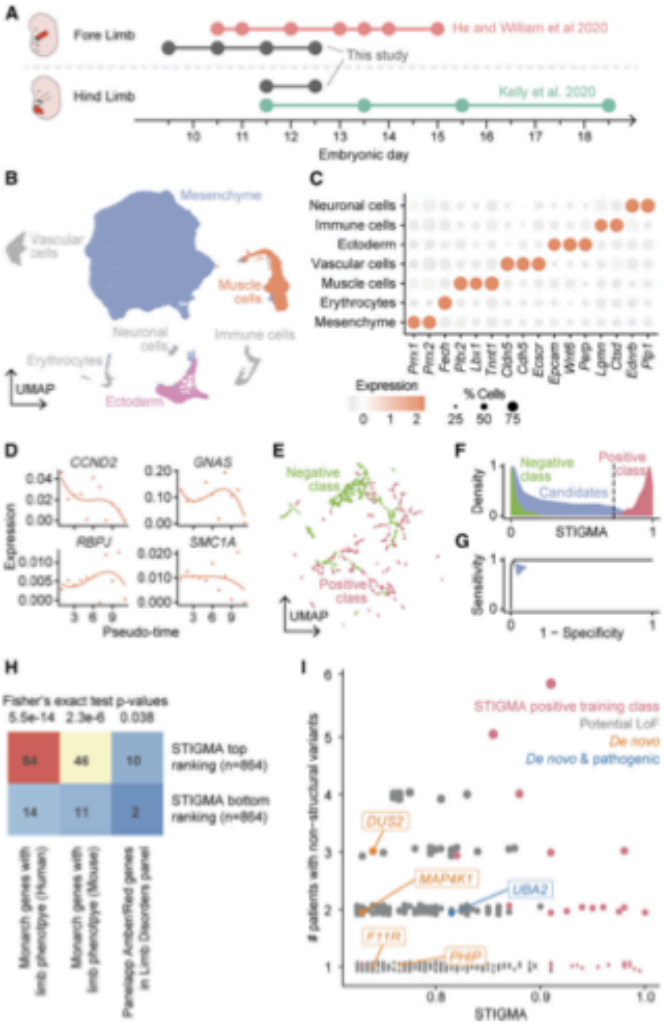


Figure 2. scRNA-seq dataset and performance of the disease-gene classifier for congenital limb malformations

(A) Embryonic time points represented across the scRNA-seq datasets.[40,41]

(B) 2D UMAP embedding of the major cell types after batch correction across the datasets, where points represent cells. Cell types not used for training STIGMA are grayed out.

(C) Marker genes corresponding to the cell types in (B).

(D) Dynamics in the gene expression of a representative set of positive class genes in muscle cells along the developmental pseudo-time. Points represent the spline knots and lines represent spline fits.

(E) 2D UMAP embedding of the genes in the training dataset, including those imputed for class balancing, based on the input features used for STIGMA.

(F) Distribution of STIGMA scores for training classes and candidate genes. Dotted line marks the threshold of 0.725.

(G) ROC curve (AUC 0.99) showing the performance of the model. The arrowhead marks the threshold of 0.725.

(H) Number of genes ranked top or bottom in STIGMA (excluding the training class), with at least one associated limb phenotype in Monarch or those being members of the Limb Disorder panel of PanelApp (classified Amber or Red). p values of Fisher's exact test are provided. The 95% confidence interval of the odds ratio did not cross unity for the three tests.

(I) STIGMA scores of genes predicted to be disease associated and featuring potential LoF variants in a previously published cohort of 69 individuals with limb malformations.[31] Genes containing *de novo* variants, including those identified to be pathogenic in the study, are highlighted.

significant enrichment of genes with at least one limb-associated phenotype in SCGs when compared to bottom ranking genes in both human (84 vs. 14 genes) and mouse (46 vs. 11), with Fisher's exact test p values of 5.5e−14 and 2.3e−6, respectively. Secondly, we checked the representation of genes labeled "Amber" (borderline evidence) and "Red" (low level of evidence) in the Limb Disorders panel of PanelApp in the STIGMA top and bottom ranking genes. This also showed a 5-fold enrichment (10 vs. 2), with a Fisher's exact test p value of 0.038.

As a final means of validation, we performed a manual search through the literature for reports where the SCGs were associated with limb disorders. This led to the identification of 112 SCGs, which were either genes with known association with congenital limb malformations, but not yet in the PanelApp green list (and by extension not in our positive training class), or genes that had nominal evidence in the literature (Table S3). For example, genes that were assigned a disease probability of greater than 0.9 included *HAS2* (MIM: 601636) and *FGFR3* (MIM: 134934), which are known to be associated with limb malformations.[62,63] While *FGFR3* is on the PanelApp green list (limb disorders), it was not included in STIGMA's positive class training list, because of its ubiquitous expression in MOCA. Another example is *UBA2*, which was ranked 309 by STIGMA with a probability of 0.81, and was recently reported to be associated with ectrodactyly (MIM: 619959).[31,64] In addition, as a means of further validation, we also identified several genes that carried potential LoF mutations in a cohort study of undiagnosed individuals with congenital limb malformations (Figure 2I).[31] Of the 7,082 potential rare non-structural

LoF variants identified by genome sequencing in 69 individuals with congenital limb defects,[31] 469 variants were found in 345 genes with STIGMA scores higher than the classification threshold of 0.725. These comprised eight of the nine genes found to carry likely pathogenic variants in the original study, including well-described genes with variants previously associated with limb disorders, such as *HOXD13* (MIM: 142989) and *GLI3* (MIM: 165240) from the positive training class as well as the STIGMA-predicted CG *UBA2*, described above, and missing only *HMGB1* (probably because of its ubiquitous expression, Figure S2). Notably, five genes implicated by STIGMA featured *de novo* variants in this dataset, which were not identified as potentially pathogenic in the original study (*DUS2* [MIM: 609707], *MAP4K1* [MIM: 601983], *F11R* [MIM: 605721], *PHIP* [MIM: 612870], and *LRP4* [MIM: 604270]). Only two of these genes have been previously associated with diseases: *LRP4* and *PHIP*. *LRP4* is associated with autosomal-recessive Cenani-Lenz syndactyly syndrome (CLS [MIM: 212780]) and was in our positive training class. *PHIP* was not part of the positive training class (absent in PanelApp) and is associated with autosomal-dominant Chung-Jansen syndrome (MIM: 617991), a phenotype comprising intellectual disability, obesity, dysmorphic facial features, notably tapering fingers, and clino- and syndactyly. The *PHIP* variant occurred in an individual with a complex malformation syndrome including renal agenesis, hypoplastic radii, oligodactyly of the hands, and polydactyly of the feet. Interestingly, *PHIP* and *UBA2*, which have been associated with similar disease profiles of oligodactyly and ectrodactyly, respectively, also showed similar temporal expression patterns in mesenchymal-chondrocytes, -fibroblasts, ectodermal-sost, and muscle cells (Figures S5B, S5D, S5E, and S5G). *DUS2*, *MAP4K1*, and *F11R*, which were not previously associated with any inheritable disease, were identified by STIGMA to be promising candidate genes from this cohort. *DUS2* variant was found in some individuals who also carried the LRP4 variant. Variants in *MAP4K1* and *F11R* were found in an individual with syndactyly of the hands and feet and in an individual with forearm reduction defects, respectively. Whether these genes are additionally associated with these phenotypes remains to be determined.

STIGMA for congenital heart diseases

Given the performance of STIGMA for congenital limb malformations, we extended the approach to predict genes associated with congenital heart diseases (CHDs).[51] After downloading and filtering, the scRNA-seq dataset[19] contained expression values of 63,561 genes in 101,749 cells, within 16 annotated cell types, of which the cardiomyocytes represented the largest cluster, containing 66% of cells in the dataset (Figure 3A). Removal of cell types such as lymphoid cells and visceral neurons, which have not been reported to lead to congenital heart disease,[65] resulted in 96,276 cells across 6 cell types. As before, the gene-expression values across these cell types and along the pseudo-time bins in addition to gene intrinsic features were used as input features for training STIGMA (Figure 3B).

A manually curated list of genes (n = 331) known to be associated with congenital heart disease was used as the positive class of the training set.[51] As before for the predictions on the limb dataset, 643 tolerant housekeeping genes were used as the negative training set. When the complete list of curated disease-causing genes were used to train and run the model, STIGMA predicted 12,012 genes potentially associated with CHD, with a precision of 0.8067. To improve the precision and to reduce the number of SCGs, we analyzed the positive class genes based on their expression pattern in other tissues in MOCA.[23] This revealed several ubiquitously expressed genes (UEGs) whose removal resulted in as few as 36 genes in the positive class (Table S1). As before, the positive and negative class genes for CHD demonstrated good separation based on the input training features, as visualized by a UMAP embedding, confirming compliance to random forest classification (Figure 3C). As in the limb, the single-cell features for the cardiac disease model were considered important for the model performance compared to the gene-intrinsic property, with a mean square value of 0.97 for single-cell features, including pseudotime features, and 0.03 for gene properties (Figure S3B).

The hyperparameters were optimized as before, resulting in a ROC curve with an AUC of 0.9972 (Figure 3E). A low number of genes in the positive training class resulted in a skewness in the distribution of the prediction probability, so a threshold of 0.57 was chosen to achieve a sensitivity above 0.8. At this chosen threshold, sensitivity and precision were 0.8333 and 0.8421, respectively (Figures 3D and 3E), predicting 3,715 SCGs to be potentially associated with CHD (Table S4).

We verified the STIGMA predictions by manually searching the literature. In an integrative study of genomic copy number variants (CNVs) and *de novo* intragenic variations (DNVs) of a CHD cohort with 4,190 DNVs (in 4,190 genes),[51] 468 genes were among the predicted SCGs, accounting for 543 variants. Furthermore, 34 of these genes had nonsynonymous *de novo* mutations in at least two individuals, nine of which had CADD scores over 30, and 10 of which were reported to be associated with heart phenotypes (Figure 3F). For example, in humans, *ALDOB* (MIM: 612724) and *MMP9* (MIM: 120361) have been found to be associated with ventricular septal defect.[32,33] *FLT4* (MIM: 136352) has been associated with pulmonary atresia with ventricular septal defect (MIM: 178370) at a prevalence of 0.2% and constituting 2% of the CHDs.[66] *MYH7B* (MIM: 609928) has been associated with left ventricular non-compaction cardiomyopathy (MIM: 604169), where the muscles extending from the left ventricle to the chamber gradually transform from sponge-like to smooth and solid.[67] Some of these genes have been implicated in heart phenotypes in mice. Namely, *Myh6* has been associated with dilated

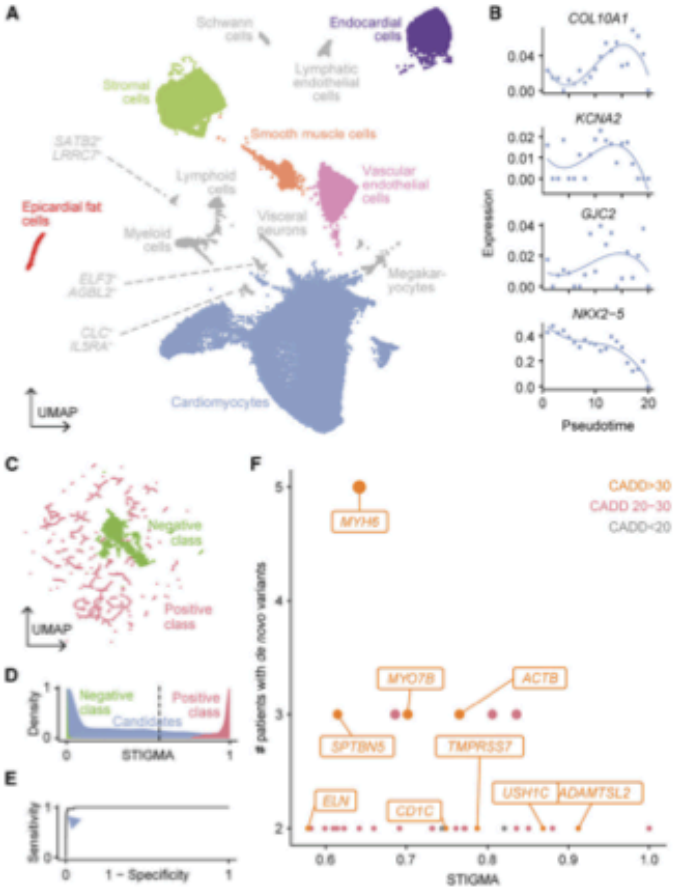


Figure 3. scRNA-seq dataset and performance of the disease-gene classifier for congenital heart disease

(A) 2D UMAP embedding of the cells, where the colors indicate the cell type annotations. Cell types not used for training STIGMA are grayed out.

(B) Dynamics in the gene expression of representative positive class genes along the developmental pseudo-time. Points represent the spline knots within a subcluster and lines represent cubic spline fits.

(C) 2D UMAP embedding of the genes in the training dataset, including those imputed for class balancing, based on the input features used for the STIGMA.

(D) Distribution of STIGMA scores for training classes and candidate genes. Dotted line marks the threshold of 0.57.

(E) ROC curve (AUC 0.9972) showing the performance of the model. The arrowhead marks the threshold of 0.57.

(F) STIGMA scores of genes featuring disruptive *de novo* variants in a previously published cohort of 2,489 trios with congenital heart disease.[51] Only SCGs with at least two *de novo* variants are plotted.

cardiomyopathy (MIM: 613252) and decreased contractile function[68] with hypoplastic left heart syndrome (HLHS [MIM: 241550]),[69] but is also associated with atrial septal defects (MIM: 614089) in humans.[70] *Scn10a* has been associated with sinus bradycardia phenotype and irregular RR interval upon scruffing.[71] *Eln* haploinsufficiency has been associated with aortic valve malformation.[72] Finally, another SCG, *Prdm1*, is associated with hypoplastic left heart syndrome and with hypoplastic aortic arch (MGI: J:175213).

Discussion

Exome and genome sequencing has become a valuable tool in understanding the genetic basis of human diseases, enabling the identification of genetic variants associated with various conditions.[73] However, the sheer volume of variants detected in a single individual poses a significant challenge in distinguishing pathogenic variants from benign ones.[3] Several approaches have been developed to aid this process, including searching through well-established databases such as 1000 Genomes, gnomAD, and ClinVar to determine the population frequencies of detected variants.[49] Additionally, the functional impact of variants is predicted using various computational approaches, enabling the identification of potentially relevant variants.[74]

While these initial filtering steps are valuable, they primarily focus on the variant level and may yield a substantial number of candidate variants in poorly understood genes that need further evaluation or painstaking experimental validation. Computational gene prioritization methods that do not rely on prior functional/disease annotations offer an alternative to shorten the list of these candidate variants further. However, all existing gene-prioritization methods based on gene expression data use bulkRNA-seq data. Indeed, a recently published method for risk gene identification, VBASS,[30] incorporated scRNA-seq data to improve upon previous methods[75,76] to identify disease-associated genes in *de novo* variant data within cohorts of affected and control individuals. However, these methods are not exactly gene-prioritization methods, because they do not globally prioritize all genes. That is, unlike STIGMA, VBASS is not designed to narrow down the list of candidate genes under consideration for an individual. Overall, STIGMA addresses some of the limitations of

traditional gene prioritization techniques. Specifically, STIGMA leverages recent developments of scRNA-seq to better understand the expression dynamics of genes across different cell types during organogenesis. By incorporating this information into the prioritization process, STIGMA provides a more comprehensive and tissue-specific assessment of candidate genes, making it a promising and cohort-independent tool for identifying variants in potentially disease-associated genes in an individual.

We implemented STIGMA in the context of two congenital disease groups—limb malformations and CHD. Since the genes in training classes and the features used to train the model directly influence model performance, we first verified that the features sufficiently discriminated the positive from the negative classes and then confirmed the results by cross-validation. STIGMA classified 864 and 3,678 genes to be SCGs for congenital limb malformations and heart disease, respectively.

We validated STIGMA predictions using multiple approaches. Automated analysis based on gene-phenotype data aggregated by the Monarch Initiative[54] as well as in the PanelApp[48] Amber/Red lists demonstrated the enrichment of genes with limb phenotypes in the top genes ranked by STIGMA. A manual search of the literature also revealed multiple lines of phenotypic evidence for the SCGs. For example, 469 LoF potential variants were found in 345 SCGs in a cohort study, with notable genes such as *UBA2*, *PHIP*, and *LRP4* not present in curated lists such as PanelApp (at the time of our download).[31] Similarly, CNVs and *de novo* variants were present in 468 SCGs in a CHD cohort, with many such as *ALDOB*, *FLT4*, *MYH7B*, *Scn10a*, and *Eln* associated with heart phenotypes in humans or mice.[51]

Although trained merely on murine scRNA-seq data, STIGMA was able to correctly suggest genes known to cause limb malformations in humans, confirming that it is able to prioritize human genes. Indeed, a direct comparison of predictions by STIGMA models for congenital heart disease trained with comparable murine and human scRNA-seq datasets revealed a statistically significant Pearson's correlation of 0.76 (Figure S6). Moreover, both models retrieved the same 34 genes that harbored *de novo* mutations in the cohort, confirming that a murine dataset can be a good approximation when a human dataset is unavailable. Nevertheless, the use of a future human scRNA-seq dataset is likely to improve the model predictions.[77]

Interestingly, temporal gene expression dynamics was more important in the STIGMA congenital limb malformation model than in the STIGMA CHD model. This is possibly because the murine limb scRNA-seq datasets spanning E9.5 to E18.5 match the embryonic stages most relevant to limb development (E9.5 to E14.5).[77,78] The human heart dataset invoked in STIGMA, however, spans days 90–122 after conception,[19] while cardiac organogenesis occurs earlier—from 26 to 56 days post conception.[79] This could have rendered the temporal dynamics in gene expression less relevant for the CHD model. A better matched heart development dataset could improve the model outcomes to levels obtained for the limb model. Nevertheless, as implemented currently, cell type-dependent gene expression values appear to facilitate clinically relevant gene prioritization.

The approach of STIGMA, as currently implemented, also has certain limitations: the choice of genes for training affects the prediction and accuracy. STIGMA assumes that genes crucial to the development of a distinctive organ (e.g., limb) are neither ubiquitously expressed nor expressed in all cell types within that organ. However, it is possible that the assumed expressional specificity occurs only at the transcript level, which most currently available atlas-level scRNA-seq data are insensitive to.[80,81] This could result in false negative predictions due to removal of "ubiquitously expressed genes" from the positive training class. Splice-sensitive scRNA-seq atlases that allow transcript counts rather than pooled gene counts could overcome these limitations. Additionally, the incomplete coverage of exonic LoF variants and the underrepresentation of several populations in gnomAD could have limited the functionality of STIGMA.[18,49] Moreover, like other expression-based annotation-agnostic gene prioritization methods, STIGMA too, is based on the principle of guilt by association. This could miss genes directly associated with a disease, if their molecular mechanisms differ from those used to train the classifier. Future STIGMA versions will require updating of the positive training class as more genes are phenotypically annotated.[7] Increased number of genes from the positive training class can also help remove biases introduced due to oversampling used to attain class balance. Contrariwise, STIGMA appears to perform reasonably well when trained with as few as ~10 genes in the positive training class based on the performance metrics alone (Figure S7). Techniques such as VBASS, which identify risk genes based on *de novo* variants in cohorts of affected individuals, could be utilized to expand the positive training class. Here, STIGMA can also help identify risk genes that may not feature any *de novo* variants in the cohort. STIGMA, as it is currently implemented, includes intolerance metrics (from gnomAD) as model features as a means of capturing genes based on these features as well. Consequently, it is possible that the predictions are biased against potential disease-associated genes that are not under selection pressure. Finally, while STIGMA will benefit from a more comprehensive validation of all the predicted SCGs, this will be possible only as phenotypic information on more genes becomes available.

We believe that STIGMA is a valuable tool for clinical gene prioritization. Efforts like the Human Cell Atlas to map every cell type in the human body will further enhance STIGMA and other comparable tools.[82]

Data and code availability

All the scripts used in this study for data preprocessing, parameter optimization, and building the random forest

classifier are available for download at our GitHub repository https://github.com/SpielmannLab/STIGMA.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.12.011.

Acknowledgments

We thank Prof. Dr. Dominik Seelow for his idea to use genes from PanelApp as a positive training class. M.S. is a DZHK principal investigator and is supported by grants from the Deutsche Forschungsgemeinschaft (DFG) (SP1532/3-2,SP1532/4-1 and SP1532/5-1), the Max Planck Society, and the Deutsches Zentrum für Luft- und Raumfahrt (DLR 01GM1925). J.P. is supported by a research grant from the University of Lübeck, Germany (J14-2021) and Else Kröner-Fresenius-Stiftung (2022_EKEA.55).

Author contributions

S.B., C.A.P.-M., M.K., M.S., and V.K.A.S. designed the research. J.G. generated the in-house limb gene expression data. Limb gene expression data were analyzed by C.A.P.-M. and S.B. S.B. performed the computational analysis. S.B., M.A.M., N.K., I.N., J.P., E.A., M.-P.H., V.K.A.S., and M.S. interpreted the results. S.B. and V.K.A.S. drafted the manuscript. S.B., C.A.P.-M., M.A.M., N.K., I.N., J.P., E.A., M.-P.H., M.K., V.K.A.S., and M.S. revised and approved the final manuscript.

Declaration of interests

The authors declare no competing interests.

Received: August 4, 2023
Accepted: December 7, 2023
Published: January 15, 2024; corrected online: February 5, 2024

References

1. Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al. (2020). The Human Phenotype Ontology in 2021. Nucleic Acids Res. *49*, D1207–D1217.
2. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789–D798.
3. Wright, C.F., Campbell, P., Eberhardt, R.Y., Aitken, S., Perrett, D., Brent, S., Danecek, P., Gardner, E.J., Chundru, V.K., Lindsay, S.J., et al. (2023). Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland. N. Engl. J. Med. *388*, 1559–1571.
4. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. Preprint at bioRxiv *1234*.
5. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. *84*, 524–533.
6. Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Bennett, R., et al. (2022). Ensembl 2022. Nucleic Acids Res. *50*, D988–D995.
7. Zolotareva, O., and Kleine, M. (2019). A Survey of Gene Prioritization Tools for Mendelian and Complex Human Diseases. J. Integr. Bioinform. *16*, 20180069.
8. Moreau, Y., and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat. Rev. Genet. *13*, 523–536.
9. Peng, C., Dieck, S., Schmid, A., Ahmad, A., Knaus, A., Wenzel, M., Mehnert, L., Zirn, B., Haack, T., Ossowski, S., et al. (2021). CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. NAR Genom. Bioinform. *3*, lqab078.
10. Piro, R.M., and Di Cunto, F. (2012). Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J. *279*, 678–696.
11. Tarailo-Graovac, M., Zhu, J.Y.A., Matthews, A., van Karnebeek, C.D.M., and Wasserman, W.W. (2017). Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. Genet. Med. *19*, 1300–1308.
12. van Dam, S., Cordeiro, R., Craig, T., van Dam, J., Wood, S.H., and de Magalhães, J.P. (2012). GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. BMC Genom. *13*, 535.
13. Deelen, P., van Dam, S., Herkert, J.C., Karjalainen, J.M., Brugge, H., Abbott, K.M., van Diemen, C.C., van der Zwaag, P.A., Gerkes, E.H., Zonneveld-Huijssoon, E., et al. (2019). Improving the diagnostic yield of exome- sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. Nat. Commun. *10*, 2837.
14. Rackham, O.J.L., Shihab, H.A., Johnson, M.R., and Petretto, E. (2015). EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. Nucleic Acids Res. *43*, e33.
15. Antanaviciute, A., Daly, C., Crinnion, L.A., Markham, A.F., Watson, C.M., Bonthron, D.T., and Carr, I.M. (2015). GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles. Bioinformatics *31*, 2728–2735.
16. Feiglin, A., Allen, B.K., Kohane, I.S., and Kong, S.W. (2017). Comprehensive Analysis of Tissue-wide Gene Expression and Phenotype Data Reveals Tissues Affected in Rare Genetic Disorders. Cell Syst. *5*, 140–148.e2.
17. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
18. Leitão, E., Schröder, C., Parenti, I., Dalle, C., Rastetter, A., Kühnel, T., Kuechler, A., Kaya, S., Gérard, B., Schaefer, E., et al. (2022). Systematic analysis and prediction of genes associated with monogenic disorders on human chromosome X. Nat. Commun. *13*, 6570.
19. Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*, eaba7721.
20. Luecken, M.D., Zaragosi, L.-E., Madissoon, E., Sikkema, L., Firsova, A.B., De Domenico, E., Kümmerle, L., Saglam, A., Berg, M., Gay, A.C.A., et al. (2022). The discovAIR project: a roadmap towards the Human Lung Cell Atlas. Eur. Respir. J. *60*, 2102057.

21. Caetano, A.J., Sequeira, I., Byrd, K.M.; and Human Cell Atlas Oral and Craniofacial Bionetwork (2022). A Roadmap for the Human Oral and Craniofacial Cell Atlas. J. Dent. Res. *101*, 1274–1288.
22. Suo, C., Dann, E., Goh, I., Jardine, L., Kleshchevnikov, V., Park, J.-E., Botting, R.A., Stephenson, E., Engelbert, J., Tuong, Z.K., et al. (2022). Mapping the developing human immune system across organs. Science *376*, eabo0510.
23. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.
24. Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., et al. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell *185*, 1777–1792.e21.
25. Meier, A.B., Zawada, D., De Angelis, M.T., Martens, L.D., Santamaria, G., Zengerle, S., Nowak-Imialek, M., Kornherr, J., Zhang, F., Tian, Q., et al. (2023). Epicardioid single-cell genomics uncovers principles of human epicardium biology in heart development and disease. Nat. Biotechnol., 1–14.
26. Sreenivasan, V.K.A., Balachandran, S., and Spielmann, M. (2022). The role of single-cell genomics in human genetics. J. Med. Genet. *59*, 827–839.
27. Rajewsky, N., Almouzni, G., Gorski, S.A., Aerts, S., Amit, I., Bertero, M.G., Bock, C., Bredenoord, A.L., Cavalli, G., Chiocca, S., et al. (2020). LifeTime and improving European healthcare through cell-based interceptive medicine. Nature *587*, 377–386.
28. Huang, X., Henck, J., Qiu, C., Sreenivasan, V.K.A., Balachandran, S., Amarie, O.V., de Angelis, M.H., Behncke, R.Y., Chan, W.-L., Despang, A., et al. (2023). Single-cell, whole-embryo phenotyping of mammalian developmental disorders. Nature *623*, 772–781.
29. Yu, F., Cato, L.D., Weng, C., Liggett, L.A., Jeon, S., Xu, K., Chiang, C.W.K., Wiemels, J.L., Weissman, J.S., de Smith, A.J., and Sankaran, V.G. (2022). Variant to function mapping at single-cell resolution through network propagation. Nat. Biotechnol. *40*, 1644–1653.
30. Zhong, G., Choi, Y.A., and Shen, Y. (2023). VBASS enables integration of single cell gene expression data in Bayesian association analysis of rare variants. Commun. Biol. *6*, 774.
31. Elsner, J., Mensah, M.A., Holtgrewe, M., Hertzberg, J., Bigoni, S., Busche, A., Coutelier, M., de Silva, D.C., Elçioglu, N., Filges, I., et al. (2021). Genome sequencing in families with congenital limb malformations. Hum. Genet. *140*, 1229–1239.
32. Huang, Q., Geng, Z., Chen, T., Cheng, X., Gu, H., Li, Q., Li, D., and Liu, R. (2019). Comparative proteomic analysis of plasma of children with congenital heart disease. Electrophoresis *40*, 1848–1854.
33. Cheng, K.-S., Liao, Y.-C., Chen, M.-Y., Kuan, T.-C., Hong, Y.-H., Ko, L., Hsieh, W.-Y., Wu, C.-L., Chen, M.-R., and Lin, C.-S. (2013). Circulating matrix metalloproteinase-2 and -9 enzyme activities in the children with ventricular septal defect. Int. J. Biol. Sci. *9*, 557–563.
34. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.
35. Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. *8*, 281–291.e9.
36. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.
37. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.
38. R Core Team (2021). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
39. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.
40. He, P., Williams, B.A., Trout, D., Marinov, G.K., Amrhein, H., Berghella, L., Goh, S.-T., Plajzer-Frick, I., Afzal, V., Pennacchio, L.A., et al. (2020). The changing mouse embryo transcriptome at whole tissue and single-cell resolution. Nature *583*, 760–767.
41. Kelly, N.H., Huynh, N.P.T., and Guilak, F. (2020). Single cell RNA-sequencing reveals cellular heterogeneity and trajectories of lineage specification during murine embryonic limb development. Matrix Biol. *89*, 1–10.
42. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.
43. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008.
44. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118.
45. The Bioconductor Dev Team (2020). BSgenome.Hsapiens.UCSC.hg38: Full Genome Sequences for Homo sapiens (UCSC version hg38, based on GRCh38.p12).
46. Statham, A.L., Strbenac, D., Coolen, M.W., Stirzaker, C., Clark, S.J., and Robinson, M.D. (2010). Repitools: an R package for the analysis of enrichment-based epigenomic data. Bioinformatics *26*, 1662–1663.
47. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.
48. Martin, A.R., Williams, E., Foulger, R.E., Leigh, S., Daugherty, L.C., Niblock, O., Leong, I.U.S., Smith, K.R., Gerasimenko, O., Haraldsdottir, E., et al. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat. Genet. *51*, 1560–1565.
49. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
50. Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends Genet. *29*, 569–574.
51. Audain, E., Wilsdon, A., Breckpot, J., Izarzugaza, J.M.G., Fitzgerald, T.W., Kahlert, A.-K., Sifrim, A., Wünnemann, F., Perez-Riverol, Y., Abdul-Khaliq, H., et al. (2021). Integrative analysis of genomic variants reveals new associations of candidate haploinsufficient genes with congenital heart disease. PLoS Genet. *17*, e1009679.
52. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R.,

Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

53. He, H., Bai, Y., Garcia, E.A., and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).
54. Mungall, C.J., McMurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. *45*, D712–D722.
55. Motenko, H., Neuhauser, S.B., O'Keefe, M., and Richardson, J.E. (2015). MouseMine: a new data warehouse for MGI. Mamm. Genome *26*, 325–330.
56. Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32.
57. Smedley, D., Schubach, M., Jacobsen, J.O.B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N.L., McMurry, J.A., et al. (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am. J. Hum. Genet. *99*, 595–606.
58. Goldstein, B.A., Polley, E.C., and Briggs, F.B.S. (2011). Random forests for genetic association studies. Stat. Appl. Genet. Mol. Biol. *10*, 32.
59. Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T., and Sun, F. (2006). Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genom. *7*, 31.
60. Wanek, N., Muneoka, K., Holler-Dinsmore, G., Burton, R., and Bryant, S.V. (1989). A staging system for mouse limb development. J. Exp. Zool. *249*, 41–49.
61. Warman, M.L., Cormier-Daire, V., Hall, C., Krakow, D., Lachman, R., LeMerrer, M., Mortier, G., Mundlos, S., Nishimura, G., Rimoin, D.L., et al. (2011). Nosology and classification of genetic skeletal disorders: 2010 revision. Am. J. Med. Genet. *155A*, 943–968.
62. Liu, J., Li, Q., Kuehn, M.R., Litingtung, Y., Vokes, S.A., and Chiang, C. (2013). Sonic hedgehog signaling directly targets Hyaluronic Acid Synthase 2, an essential regulator of phalangeal joint patterning. Dev. Biol. *375*, 160–171.
63. Eswarakumar, V.P., and Schlessinger, J. (2007). Skeletal overgrowth is mediated by deficiency in a specific isoform of fibroblast growth factor receptor 3. Proc. Natl. Acad. Sci. USA *104*, 3937–3942.
64. Schnur, R.E., Yousaf, S., Liu, J., Chung, W.K., Rhodes, L., Marble, M., Zambrano, R.M., Sobreira, N., Jayakar, P., Pierpont, M.E., et al. (2021). UBA2 variants underlie a recognizable syndrome with variable aplasia cutis congenita and ectrodactyly. Genet. Med. *23*, 1624–1635.
65. (2006). Making or Breaking the Heart: From Lineage Determination to Morphogenesis. Cell *126*, 1037–1048.
66. Xie, H., Hong, N., Zhang, E., Li, F., Sun, K., and Yu, Y. (2019). Identification of Rare Copy Number Variants Associated With Pulmonary Atresia With Ventricular Septal Defect. Front. Genet. *10*, 15.
67. Esposito, T., Sampaolo, S., Limongelli, G., Varone, A., Formicola, D., Diodato, D., Farina, O., Napolitano, F., Pacileo, G., Gianfrancesco, F., and Di Iorio, G. (2013). Digenic mutational inheritance of the integrin alpha 7 and the myosin heavy chain 7B genes causes congenital myopathy with left ventricular non-compact cardiomyopathy. Orphanet J. Rare Dis. *8*, 91.
68. Schmitt, J.P., Debold, E.P., Ahmad, F., Armstrong, A., Frederico, A., Conner, D.A., Mende, U., Lohse, M.J., Warshaw, D., Seidman, C.E., and Seidman, J.G. (2006). Cardiac myosin missense mutations cause dilated cardiomyopathy in mouse models and depress molecular motor function. Proc. Natl. Acad. Sci. USA *103*, 14525–14530.
69. Anfinson, M., Fitts, R.H., Lough, J.W., James, J.M., Simpson, P.M., Handler, S.S., Mitchell, M.E., and Tomita-Mitchell, A. (2022). Significance of α-Myosin Heavy Chain Variants in Hypoplastic Left Heart Syndrome and Related Cardiovascular Diseases. J. Cardiovasc. Dev. Dis. *9*, 144.
70. Ching, Y.-H., Ghosh, T.K., Cross, S.J., Packham, E.A., Honeyman, L., Loughna, S., Robinson, T.E., Dearlove, A.M., Ribas, G., Bonser, A.J., et al. (2005). Mutation in myosin heavy chain 6 causes atrial septal defect. Nat. Genet. *37*, 423–428.
71. Blasius, A.L., Dubin, A.E., Petrus, M.J., Lim, B.-K., Narezkina, A., Criado, J.R., Wills, D.N., Xia, Y., Moresco, E.M.Y., Ehlers, C., et al. (2011). Hypermorphic mutation of the voltage-gated sodium channel encoding gene Scn10a causes a dramatic stimulus-dependent neurobehavioral phenotype. Proc. Natl. Acad. Sci. USA *108*, 19413–19418.
72. Krishnamurthy, V.K., Opoka, A.M., Kern, C.B., Gullak, F., Narmoneva, D.A., and Hinton, R.B. (2012). Maladaptive matrix remodeling and regional biomechanical dysfunction in a mouse model of aortic valve disease. Matrix Biol. *31*, 197–205.
73. 100,000 Genomes Project Pilot Investigators, Smedley, D., Smith, K.R., Martin, A., Thomas, E.A., McDonagh, E.M., Cipriani, V., Ellingford, J.M., Arno, G., Tucci, A., et al. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N. Engl. J. Med. *385*, 1868–1880.
74. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.
75. Nguyen, H.T., Bryois, J., Kim, A., Dobbyn, A., Huckins, L.M., Munoz-Manchado, A.B., Ruderfer, D.M., Genovese, G., Fromer, M., Xu, X., et al. (2017). Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. Genome Med. *9*, 114.
76. Nguyen, T.-H., He, X., Brown, R.C., Webb, B.T., Kendler, K.S., Vladimirov, V.I., Riley, B.P., and Bacanu, S.-A. (2021). DECO: a framework for jointly analyzing de novo and rare case/control variants, and biological pathways. Brief. Bioinform. *22*, bbab067.
77. Petit, F., Sears, K.E., and Ahituv, N. (2017). Limb development: a paradigm of gene regulation. Nat. Rev. Genet. *18*, 245–258.
78. Zeller, R., López-Ríos, J., and Zuniga, A. (2009). Vertebrate limb bud development: moving towards integrative analysis of organogenesis. Nat. Rev. Genet. *10*, 845–858.
79. Hikspoors, J.P.J.M., Kruepunga, N., Mommen, G.M.C., Köhler, S.E., Anderson, R.H., and Lamers, W.H. (2022). A pictorial account of the human embryonic heart between 3.5 and 8 weeks of development. Commun. Biol. *5*, 226.
80. Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y., et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. *27*, 1759–1768.
81. Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., and Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat. Genet. *40*, 1416–1425.
82. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The Human Cell Atlas. Elife *6*, e27041.

Update

The American Journal of Human Genetics
Volume 111, Issue 3, 7 March 2024, Page 618

DOI: https://doi.org/10.1016/j.ajhg.2024.01.009

CORRECTION

STIGMA: Single-cell tissue-specific gene prioritization using machine learning

Saranya Balachandran, Cesar A. Prada-Medina, Martin A. Mensah, Juliane Glaser, Naseebullah Kakar, Inga Nagel, Jelena Pozojevic, Enrique Audain, Marc-Phillip Hitz, Martin Kircher, Varun K.A. Sreenivasan,* and Malte Spielmann*

(The American Journal of Human Genetics *111*, 338–349; February 1, 2024)

As a result of an author oversight in the originally published version of this article, the author Juliane Glaser was missing. This error has now been corrected in the article online. The authors apologize for the error and any inconvenience that may have resulted.

*Correspondence: varun.sreenivasan@uksh.de (V.K.A.S.), malte.spielmann@uksh.de (M.S.)
https://doi.org/10.1016/j.ajhg.2024.01.009.

The American Journal of Human Genetics, Volume *111*

Supplemental information

STIGMA: Single-cell tissue-specific gene prioritization using machine learning

Saranya Balachandran, Cesar A. Prada-Medina, Martin A. Mensah, Juliane Glaser, Naseebullah Kakar, Inga Nagel, Jelena Pozojevic, Enrique Audain, Marc-Phillip Hitz, Martin Kircher, Varun K.A. Sreenivasan, and Malte Spielmann

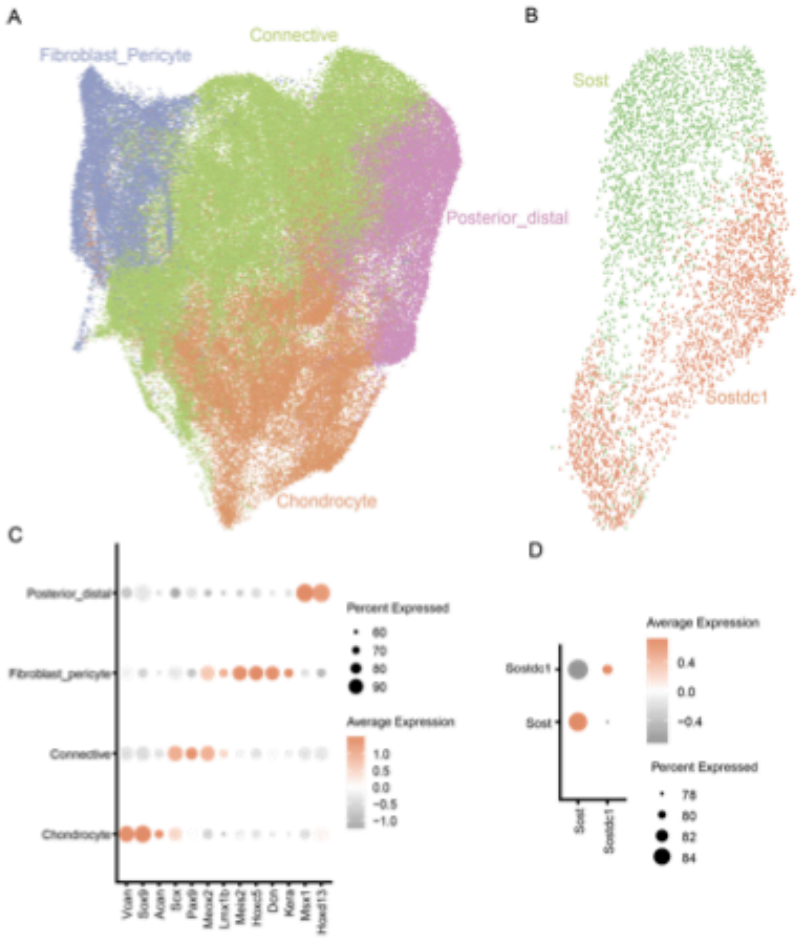


Figure S1. Sub-clustering of limb mesenchyme and ectoderm clusters. 2D UMAP embeddings of mesenchyme (**A**) and ectoderm (**B**) clusters coloured by sub-clusters, which were annotated based on the marker genes in **C**. and **D**. respectively.

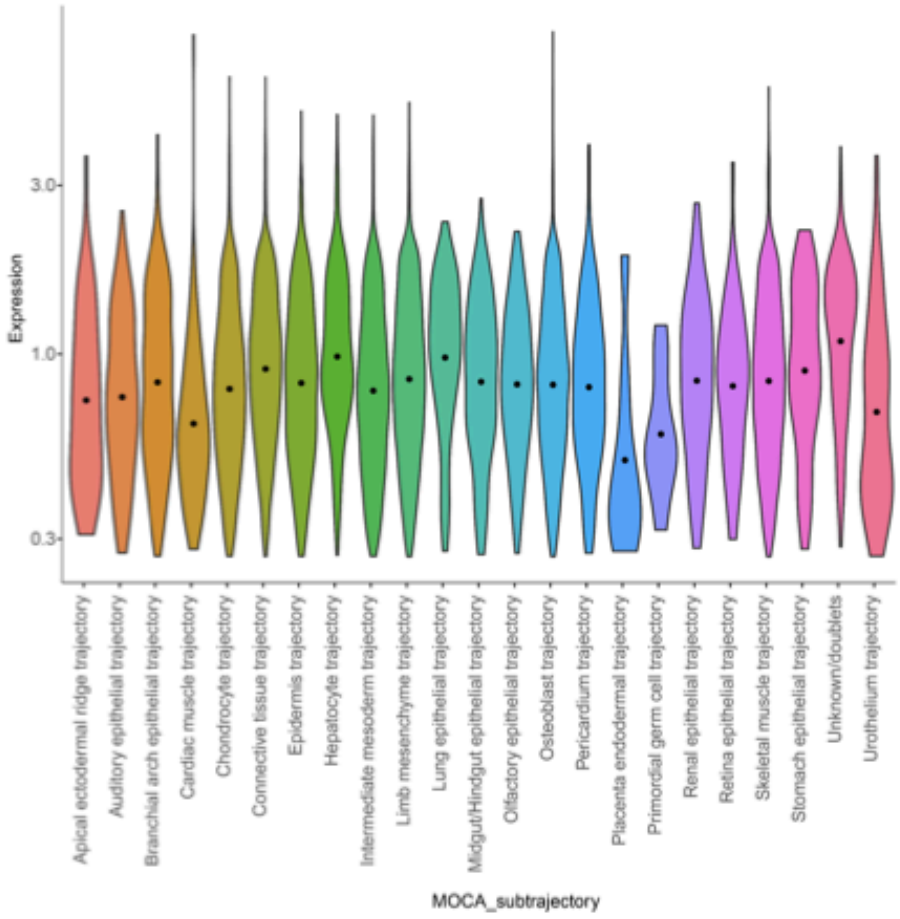


Figure S2. Some genes (e.g. *Hmgb1*) showed ubiquitous expression across all MOCA trajectories. The ubiquitous expression of *Hmgb1* across the sub trajectories of epithelial, hepatic, mesenchyme is shown as a representative example.

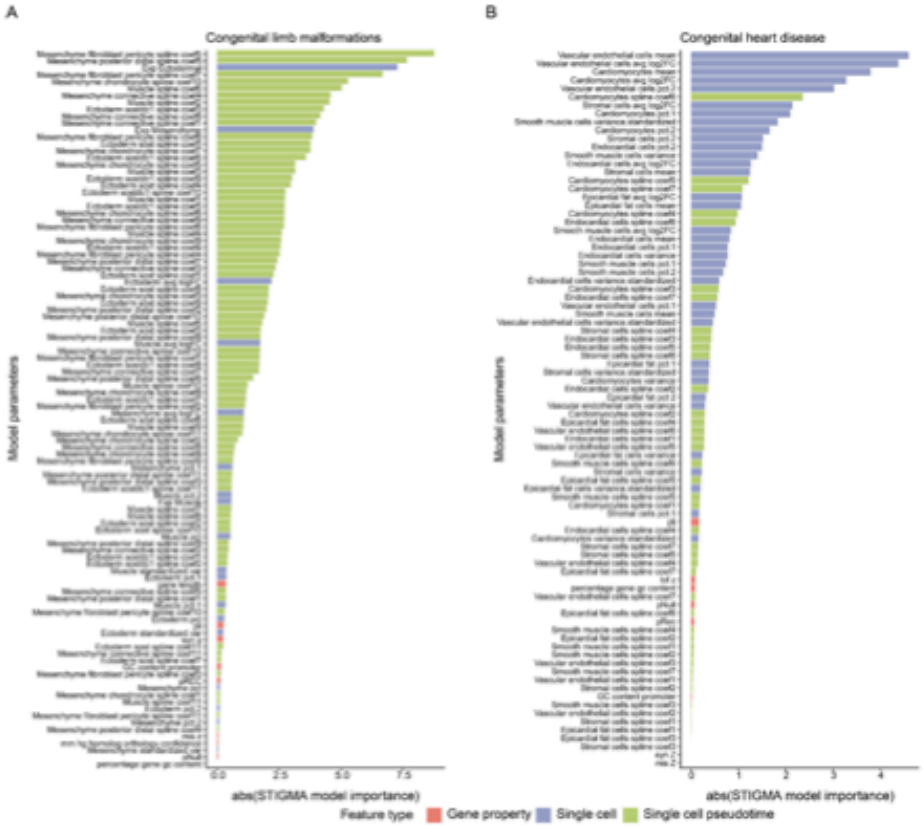


Figure S3. Importance of model features for STIGMA classification. The absolute values of the importance of model features for (**A**) congenital limb malformations and (**B**) congenital heart disease, where colors represent the feature type. The features labeled "Single cell" include cell type-specific features such as the percentage of cells expressing each gene and fold change in expression of a cell type compared to the rest of the cells, whereas "Single cell pseudotime" include spline coefficients of the temporal dynamics in expression per cell type.

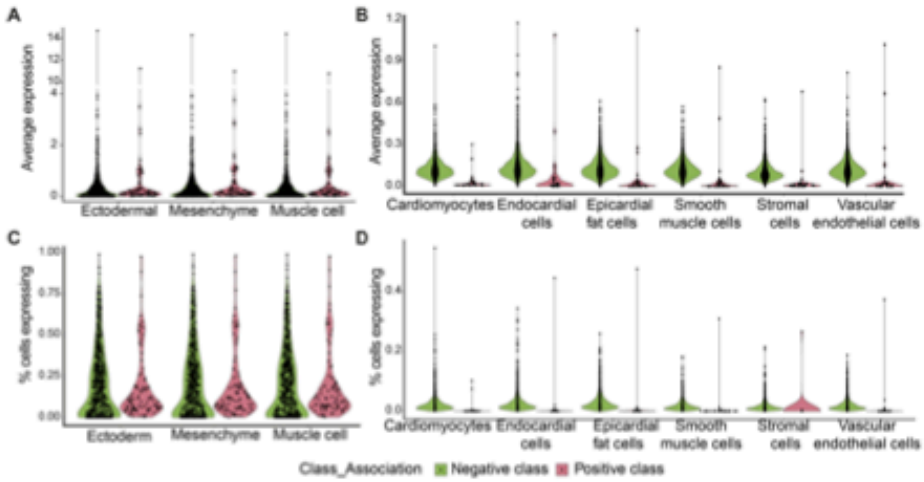


Figure S4. Expression of positive and negative class genes across the different cell types used in STIGMA. Average expression (**A**, **B**) and percentage of cells expressing (**C**, **D**) of the genes and in the limb (**A**, **C**) and heart (**B**, **D**) scRNA-seq datasets. The green and pink violins represent the distributions of the negative and positive class genes, respectively.

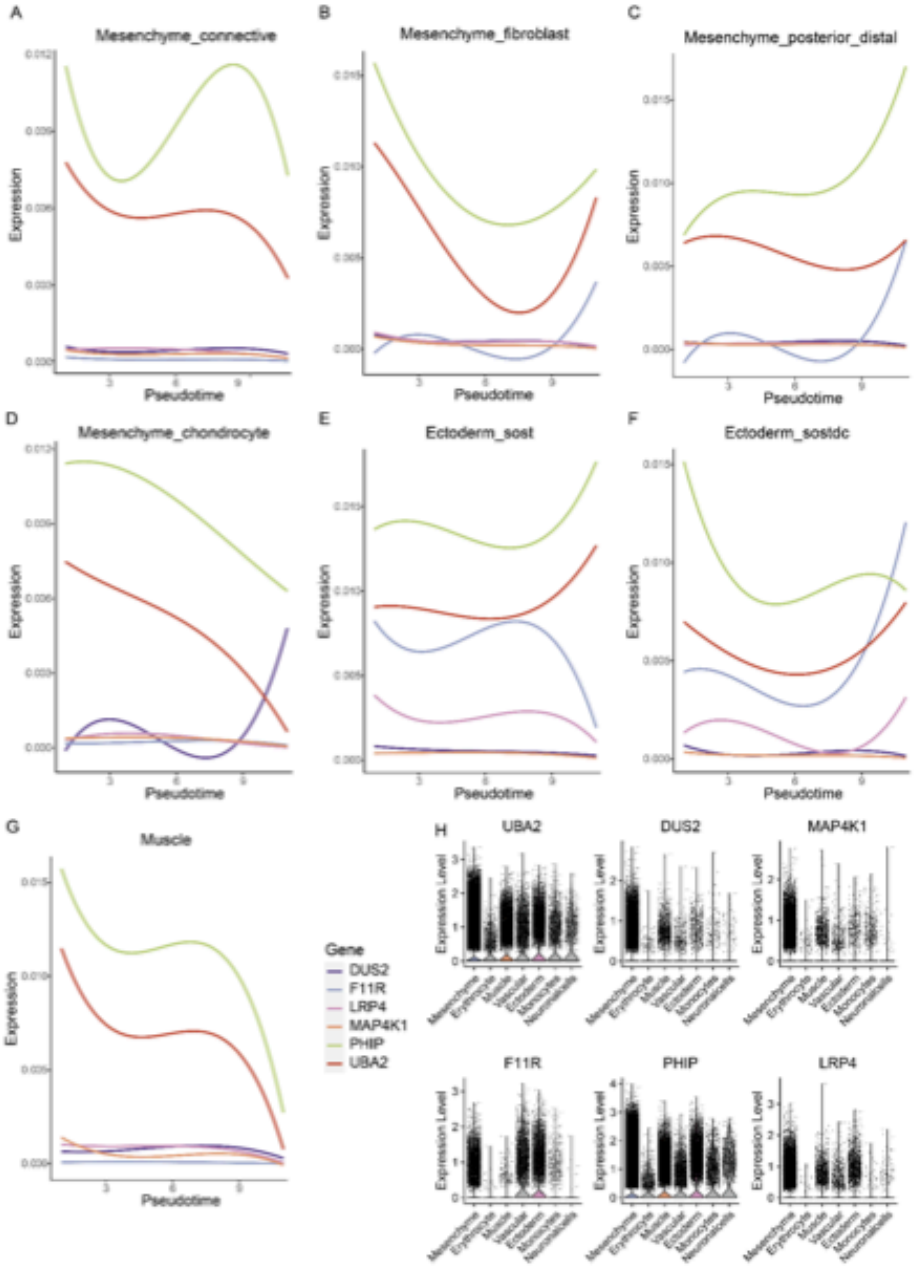

A
Mesenchyme_connective
B
Mesenchyme_fibroblast
C
Mesenchyme_posterior_distal
D
Mesenchyme_chondrocyte
E
Ectoderm_sost
F
Ectoderm_sostdc
G
Muscle
Expression
Pseudotime
Gene
DUS2
F11R
LRP4
MAP4K1
PHIP
UBA2
H
UBA2
DUS2
MAP4K1
F11R
PHIP
LRP4
Expression Level
Mesenchyme
Erythrocyte
Muscle
Vascular
Ectoderm
Monocytes
Neuronalcells

Figure S5. Temporal dynamics and cellular expression of genes with *de novo* mutations. A-G. Dynamics in the expression of genes with potential LoF identified in the cohort of congenital limb malformations along the developmental pseudo-time in limb sub-trajectories [30]. The lines represent spline fits. **H**. Average expression of the genes across cellular clusters. Cell types not used for training STIGMA are grayed out.

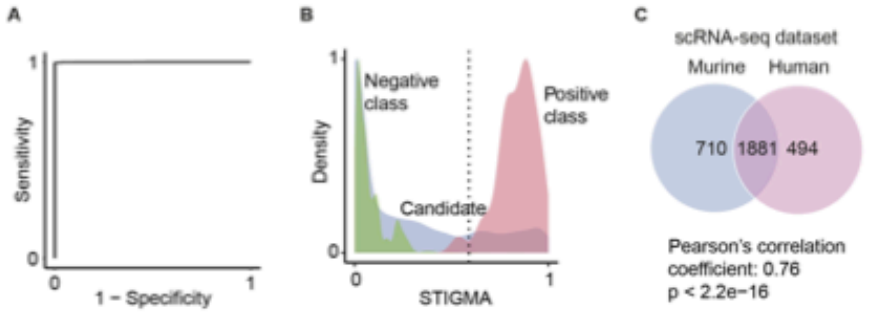


Figure S6: Performance of STIGMA for congenital heart disease trained on murine dataset and its comparison to that trained with human dataset. A. ROC curve showing the performance of the model. **B.** Distribution of STIGMA scores for training class and candidate genes. Dotted line marks the binary classification threshold of 0.59. **C.** Venn Diagram showing the number of genes predicted by using the murine and human scRNA-seq fetal heart datasets. Pearson's correlation coefficient, R, between the STIGMA scores of genes by training with murine and human datasets and the corresponding p-value are shown. scRNA-seq data for the murine dataset (E14.5 - E18) were downloaded from Feng et al [83] and processed as described in Methods. Note, 4622 genes in the human dataset were not present in the murine dataset, of which some are due to the absence of orthologs.

4. Discussion

Next-generation sequencing (NGS) has notably advanced the understanding of diseases and development in clinical settings. However, all candidate disease genes remain mostly poorly annotated and therefore categorized as genes of unknown significance. Such genes are often excluded from diagnostic considerations, posing a limitation to comprehensive disease diagnosis.

While bulk RNA-sequencing (RNA-seq) techniques can provide insights into differences between case and control groups, they are limited in their ability to capture the full complexity of gene expression (Velmeshev et al. 2019; Smajić et al. 2022; Rai et al. 2020). Many disease-related genes, especially those associated with congenital disorders, are not ubiquitously expressed throughout development. Instead, their expression is often highly specific to particular cell types or stages of development (Tu et al. 2006).

To address this challenge, we employed single-cell sequencing methods to offer a more powerful approach. These techniques can capture the dynamic and variable gene expression patterns across individual cells within a tissue, providing a more nuanced understanding of the molecular processes involved in both normal development and disease.

In our study, we explored two key aspects of the application of single-cell sequencing in disease and development. We assessed the power of single-cell sequencing to detect evolutionary differences in the same organ across human, mouse, macaque and zebrafish. This approach helps to highlight the conserved and non-conserved gene expression patterns that may underlie developmental processes and disease susceptibility. It also shows how model organisms can be used in the study of human diseases. Given that we could study the evolutionary aspect of the organism using sc-seq, we then wanted to study how the expression pattern during development determines the disease state of the organs from a wildtype sc-seq dataset. To achieve this we integrated sc-seq data from various stages of organogenesis to identify

disease-associated genes during development. By providing the features from this developmental dataset to a machine learning model, we aimed to predict and prioritize genes that may play a role in congenital diseases. This approach has the potential to uncover novel disease genes that are specific to particular developmental stages or cell types, which might otherwise be missed using traditional bulk sequencing methods.

The discussion section is divided into two parts based on two publications in peer reviewed journals. The first part provides insights into the findings of the article "Balachandran, S., Pozojevic, J., Sreenivasan, V. K. A. & Spielmann, M. Comparative single-cell analysis of the adult heart and coronary vasculature. *Mamm Genome* **34**, 276–284 (2023)." The second part discusses the article "Balachandran, S. *et al.* STIGMA: Single-cell tissue-specific gene prioritization using machine learning. *Am J Hum Genet* S0002-9297(23)00443–3 (2024) doi:10.1016/j.ajhg.2023.12.011."

4.1 Comparative single-cell analysis of the adult heart and coronary vasculature.

Over millions of years, the heart has evolved from simpler forms seen in insects and worms to the more advanced four-chambered hearts found in mammals. Heart plays a key role in the circulatory system by pumping blood to enable the exchange of oxygen, nutrients, and waste products between the blood and tissues. The anatomical differences in heart of human, mouse, macaque and zebrafish were studied using scRNA-seq.

We used the publicly accessible datasets of human, mouse sequenced by Chromium Single-Cell 3′ protocol (10 × Genomics) and macaque by STRT-seq. The adult human heart dataset includes 451,513 cells (Litviňuková et al. 2020), the mouse heart dataset contains 12,710 cells (Vidal et al. 2019), the macaque heart dataset, generated specifically from the aorta and coronary arteries, consists of 7,989 cells (Zhang et al. 2020) and the zebrafish heart consisted of 200,000 cells of healthy and regenerating heart (Hu et al. 2022). To account for technical variability we applied a Seurat based

integration strategy, then proceeded with clustering analysis. The cell types were identified based on the marker genes provided in the original publications. We identified cardiomyocytes (ventricular and atrial), adipocytes, endothelial cells, fibroblasts, pericytes, smooth muscle cells, endocardial cells, neuronal and immune cells. We then sub clustered to identify cell type specific differences. We observed a ventricular cardiomyocytes (vCM) subcluster specific to mice, consisting of cells with lower expression of *Myh7* and *Plc1* and higher expression of *Prune2* gene. In other clusters of vCM as well we observed genes related to high energetic state highly expressed, these genes were shown to be associated with cardiac conduction in gene ontology analysis. Species-specific differences in smooth muscle cell clustering and endothelial cell gene expression likely resulted from variations in the tissues selected for sequencing in the original studies. Due to the evolutionary difference between zebrafish and mammalian hearts, it did not integrate well. We then sought out to integrate the fibroblasts, due to its role in cardiac regeneration as shown in the original study. We performed hierarchical cluster tree analysis and saw that the regeneration specific cell types were more distant to human fibroblasts compared to the other fibroblast subclusters.
Our study was limited by the choice of datasets, as the tissues were sampled differently. The number of cells from each of the datasets were also high variable leading to the bias. We also did not validate our findings experimentally.

4.2 STIGMA: Single-cell tissue-specific gene prioritization using machine learning.

NGS has made gene analysis routine in clinics. Yet many variants in genes of unknown function are labeled as variants of unknown significance and do not contribute to the diagnosis of a rare congenital disease, until further validated by experimental approaches like in situ hybridization. Gene prioritization can help narrow down the list of candidate genes under consideration. Here we introduce Single-cell tissue-specific gene prioritization using machine learning (STIGMA), to prioritize disease genes for congenital malformations. STIGMA predicts the disease-causing probability of genes based on their expression profile considering the temporal dynamics during the

organogenesis of an healthy organism. We validated our approach by applying the model on fetal mouse limb (He et al. 2020; Kelly, Huynh, and Guilak 2020) and human fetal heart (Cao et al. 2020) single cell datasets to prioritize candidate genes for congenital limb malformation and congenital heart disease respectively.

In order to achieve the aim, we wanted to develop a machine learning model that was trained on single cell features like expression of genes across various cell types and the pseudo temporal expression of genes in each of the subcell types. Apart from single cell features we also used gene intrinsic properties like tolerance scores, and GC content. We tested our approach on two congenital diseases, namely congenital limb malformations and CHD. We used a cross validation approach to validate the models performance to distinguish the two classes. STIGMA predicted 864 genes to be associated with congenital limb malformations. We validated our findings on a patient cohort, and saw several of the candidate genes that we predicted, harbored *de novo* and potential loss of function mutations in patients. One of the notable examples is *UBA2*, which was recently reported to be associated with limb malformations (Elsner et al. 2021).

Similarly we applied the model on the human fetal heart dataset and identified 3678 candidate genes. Likely we validated our findings on a cohort of patients with congenital heart defect, and 34 of our candidate genes possessed two or more nonsynonymous de novo mutations. Several of the identified genes were also associated with heart phenotypes in humans and mice by the Monarch Initiative (Mungall et al. 2017), noble example is *Prdm1*, which is associated with hypoplastic left ventricle and hypoplastic aortic arch (MGI: J:175213). We also applied the model on the mouse fetal heart dataset (Feng et al. 2022), and the predictions from both the datasets had a correlation of 0.76, showing that mice can be a good approximation when human datasets are not available.

The current version of STIGMA has certain limitations, the accuracy of the model is highly dependent on the seed positive and negative class genes. Also the negative class was chosen based on the assumption that disease genes do not have ubiquitous expression like that of housekeeping genes. Also STIGMA is based on guilt by association which would miss genes that have a different mechanism of disease than

the seed genes provided. The genes found to be associated with the disease were not experimentally validated.

5. Conclusion

In conclusion, we demonstrate how single-cell sequencing can be leveraged to better understand the development of an organism and how applying machine learning models to single-cell data can significantly improve diagnostic outcomes by analyzing expression dynamics across various cell types and developmental stages.

With vertebrate evolution, there have been huge changes in the circulatory system, including the heart. Although the heart simply circulates blood throughout the body, size and shape, as well as speed and capacity for regeneration, differ greatly among species. Through comparative single-cell analysis, we showcased how species-specific changes within cell types can be captured by single-cell sequencing. We identified populations of cells that are common between species like endothelial cells and atrial cardiomyocytes. We also identified subpopulations of cells that differ between mice and human ventricular cardiomyocytes due to biological differences in the rate of heart beat. We identified certain differences in the population of cells due to different sampling techniques. These results demonstrate single cell sequencing methods capability to capture subtle cellular differences in organismal development and the potential to use model organisms to study human diseases.

In the STIGMA project, we present an approach to prioritize genes, particularly for rare congenital diseases, by integrating single-cell RNA-seq data with machine learning. This method stands out by learning the intricate temporal dynamics of gene expression across various cell types during healthy organ development, offering a tissue-specific view, through which candidate genes can be assessed. By applying STIGMA to both mouse limb and human fetal heart datasets, we were able to prioritize a significant number of variants and genes that are highly relevant to congenital malformations. For instance, the prioritization of 469 variants in 345 genes for limb malformations, including *UBA2*, demonstrates the precision and relevance of this framework. Additionally, the detection of 34 genes with nonsynonymous de novo variants (nsDNVs) in individuals with congenital heart defects, particularly the ortholog of *Prdm1*, underscores STIGMA's

capacity to identify genes tied to complex developmental disorders, such as hypoplastic left ventricle and aortic arch defects. The capability of STIGMA to capture the subtle changes in expression dynamics, makes it a valuable tool for advancing the discovery of causal candidate genes in human genetic disorders, ultimately contributing to more precise diagnostics and potential therapeutic interventions. We hope that by implementing approaches like STIGMA, the diagnostic yield for rare congenital diseases will improve, even by a few percent.

Nevertheless, this approach only captures the small 2% of the whole genome, thus we hope to use multiomics sequencing approaches of single cell to further uncover the non-coding regions to understand how gene regulatory elements affect congenital diseases.

6. Appendix

6.1 List of Abbreviations

NGS	Next generation sequencing technologies
HPO	Human Phenotype Ontology
OMIM	Online Mendelian Inheritance in Man
gnomAD	Genome Aggregation Database
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
VUS	variance of unknown significance
ML	Machine learning
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
RBF	Radial Basis Function
TP	true positive
TN	true negative
FP	false positive
FN	false negative
ROC	Receiver Operating Characteristic
AUC	area under the curve
sc-seq	single-cell sequencing
scRNA-seq	single cell rna sequencing
sc-multiome	single cell multiome seq
BCL	binary base call

HVG	highly variable genes
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation Projection
t-SNE	t-distributed stochastic neighbor embedding
vCM	Ventricular cardiomyocytes
SCG	STIGMA candidate gene
CHD	Congenital Heart disease

6.2 List of Figures

6.3 Code availability

All the scripts used in this study for data preprocessing, parameter optimization, and building the random forest classifier are available for download at our GitHub repository https://github.com/SpielmannLab/STIGMA.

7. References

100,000 Genomes Project Pilot Investigators, Damian Smedley, Katherine R. Smith, Antonio Martin, Ellen A. Thomas, Ellen M. McDonagh, Valentina Cipriani, et al. 2021. “100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report.” *The New England Journal of Medicine* 385 (20): 1868–80.

Abdelaal, Tamim, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J. T. Reinders, and Ahmed Mahfouz. 2019. “A Comparison of Automatic Cell Identification Methods for Single-Cell RNA Sequencing Data.” *Genome Biology* 20 (1): 194.

Amberger, Joanna S., Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. 2015. “OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders.” *Nucleic Acids Research* 43 (Database issue): D789–98.

Amberger, Joanna S., Carol A. Bocchini, Alan F. Scott, and Ada Hamosh. 2019. “OMIM.org: Leveraging Knowledge across Phenotype-Gene Relationships.” *Nucleic Acids Research* 47 (D1): D1038–43.

Andrews, Tallulah S., Vladimir Yu Kiselev, Davis McCarthy, and Martin Hemberg. 2021. “Tutorial: Guidelines for the Computational Analysis of Single-Cell RNA Sequencing Data.” *Nature Protocols* 16 (1): 1–9.

Antanaviciute, Agne, Catherine Daly, Laura A. Crinnion, Alexander F. Markham, Christopher M. Watson, David T. Bonthron, and Ian M. Carr. 2015. “GeneTIER: Prioritization of Candidate Disease Genes Using Tissue-Specific Gene Expression Profiles.” *Bioinformatics* 31 (16): 2728–35.

Audain, Enrique, Anna Wilsdon, Jeroen Breckpot, Jose M. G. Izarzugaza, Tomas W. Fitzgerald, Anne-Karin Kahlert, Alejandro Sifrim, et al. 2021. “Integrative Analysis of Genomic Variants Reveals New Associations of Candidate Haploinsufficient Genes with Congenital Heart Disease.” *PLoS Genetics* 17 (7): e1009679.

Avsec, Žiga, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. 2021. “Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions.” *Nature Methods* 18 (10): 1196–1203.

Azadifar, Saeid, and Ali Ahmadi. 2022. “A Novel Candidate Disease Gene Prioritization Method Using Deep Graph Convolutional Networks and Semi-Supervised Learning.” *BMC Bioinformatics* 23 (1): 422.

Bansal, Vikas, and Christina Boucher. 2019. “Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going?” *iScience* 18 (August):37–41.

Benkendorf, Donald J., Samuel D. Schwartz, D. Richard Cutler, and Charles P. Hawkins. 2023. “Correcting for the Effects of Class Imbalance Improves the Performance of Machine-Learning Based Species Distribution Models.” *Ecological Modelling* 483 (110414): 110414.

Berry, Michael W., Azlinah Mohamed, and Bee Wah Yap. 2019. *Supervised and Unsupervised Learning for Data Science*. Springer Nature.

Breiman, Leo. 2001. *Machine Learning* 45 (1): 5–32.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija.

2018. “Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species.” *Nature Biotechnology* 36 (5): 411–20.
Cao, Junyue, Diana R. O’Day, Hannah A. Pliner, Paul D. Kingsley, Mei Deng, Riza M. Daza, Michael A. Zager, et al. 2020. “A Human Cell Atlas of Fetal Gene Expression.” *Science* 370 (6518). https://doi.org/10.1126/science.aba7721.
Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, et al. 2019. “The Single-Cell Transcriptional Landscape of Mammalian Organogenesis.” *Nature* 566 (7745): 496–502.
Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2011. “SMOTE: Synthetic Minority over-Sampling Technique.” https://doi.org/10.48550/ARXIV.1106.1813.
Chen, Xi, and Hemant Ishwaran. 2012. “Random Forests for Genomic Data Analysis.” *Genomics* 99 (6): 323–29.
Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” *Machine Learning* 20 (3): 273–97.
Crespi, Sarah. 2018. “Video: 2018’s Breakthrough of the Year and Runners-Up.” *Science*, December. https://doi.org/10.1126/science.aaw4480.
Davis, Sean, Rintu Kutum, Hallvard A. Wæhler, Zhe Wang, Parashar, Atakan Ekiz, Luke Zappia, et al. 2024. *Seandavi/awesome-Single-Cell: 2024-07-08*. Zenodo. https://doi.org/10.5281/ZENODO.1117762.
Deciphering Developmental Disorders Study. 2017. “Prevalence and Architecture of de Novo Mutations in Developmental Disorders.” *Nature* 542 (7642): 433–38.
Deelen, Patrick, Sipko van Dam, Johanna C. Herkert, Juha M. Karjalainen, Harm Brugge, Kristin M. Abbott, Cleo C. van Diemen, et al. 2019. “Improving the Diagnostic Yield of Exome- Sequencing by Predicting Gene-Phenotype Associations Using Large-Scale Gene Expression Analysis.” *Nature Communications* 10 (1): 2837.
De Simone, Marco, Jonathan Hoover, Julia Lau, Hayley Bennet, Bing Wu, Cynthia Chen, Hari Menon, et al. 2024. “Comparative Analysis of Commercial Single-Cell RNA Sequencing Technologies.” *bioRxiv*. https://doi.org/10.1101/2024.06.18.599579.
Domcke, Silvia, Andrew J. Hill, Riza M. Daza, Junyue Cao, Diana R. O’Day, Hannah A. Pliner, Kimberly A. Aldinger, et al. 2020. “A Human Cell Atlas of Fetal Chromatin Accessibility.” *Science* 370 (6518). https://doi.org/10.1126/science.aba7612.
Duda, Richard O., Peter E. Hart, and David G. Stork. 2012. *Pattern Classification*. John Wiley & Sons.
Elsner, Jonas, Martin A. Mensah, Manuel Holtgrewe, Jakob Hertzberg, Stefania Bigoni, Andreas Busche, Marie Coutelier, et al. 2021. “Genome Sequencing in Families with Congenital Limb Malformations.” *Human Genetics* 140 (8): 1229–39.
Feng, Wei, Abha Bais, Haoting He, Cassandra Rios, Shan Jiang, Juan Xu, Cindy Chang, Dennis Kostka, and Guang Li. 2022. “Single-Cell Transcriptomic Analysis Identifies Murine Heart Molecular Features at Embryonic and Neonatal Stages.” *Nature Communications* 13 (1): 7960.
Fernández, Alberto, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from Imbalanced Data Sets*. Springer.
Firth, Helen V., Shola M. Richards, A. Paul Bevan, Stephen Clayton, Manuel Corpas,

Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M. Pettett, and Nigel P. Carter. 2009. “DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.” *American Journal of Human Genetics* 84 (4): 524–33.

Flach, Peter A. 2011. “ROC Analysis.” In *Encyclopedia of Machine Learning*, 869–75. Boston, MA: Springer US.

Fontaine, Jean-Fred, Florian Priller, Adriano Barbosa-Silva, and Miguel A. Andrade-Navarro. 2011. “Génie: Literature-Based Gene Prioritization at Multi Genomic Scale.” *Nucleic Acids Research* 39 (Web Server issue): W455–61.

Fürnkranz, Johannes. 2011. “Decision Tree.” In *Encyclopedia of Machine Learning*, 263–67. Boston, MA: Springer US.

Gargano, Michael A., Nicolas Matentzoglu, Ben Coleman, Eunice B. Addo-Lartey, Anna V. Anagnostopoulos, Joel Anderton, Paul Avillach, et al. 2024. “The Human Phenotype Ontology in 2024: Phenotypes around the World.” *Nucleic Acids Research* 52 (D1): D1333–46.

“Gene Map Statistics - OMIM.” n.d. Accessed July 31, 2024. https://omim.org/statistics/geneMap.

Ghosh, Shyamasree, and Rathi Dasgupta. 2022. *Machine Learning in Biological Sciences: Updates and Future Prospects*. Springer Nature.

Han, Lei, Xiaoyu Wei, Chuanyu Liu, Giacomo Volpe, Zhenkun Zhuang, Xuanxuan Zou, Zhifeng Wang, et al. 2022. “Cell Transcriptomic Atlas of the Non-Human Primate Macaca Fascicularis.” *Nature* 604 (7907): 723–31.

Hashimshony, Tamar, Florian Wagner, Noa Sher, and Itai Yanai. 2012. “CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification.” *Cell Reports* 2 (3): 666–73.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

He, Peng, Brian A. Williams, Diane Trout, Georgi K. Marinov, Henry Amrhein, Libera Berghella, Say-Tar Goh, et al. 2020. “The Changing Mouse Embryo Transcriptome at Whole Tissue and Single-Cell Resolution.” *Nature* 583 (7818): 760–67.

Heumos, Lukas, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, et al. 2023. “Best Practices for Single-Cell Analysis across Modalities.” *Nature Reviews. Genetics* 24 (8): 550–72.

Huang, Shengkang, Xinyu Wang, Yu Wang, Yajing Wang, Chenglong Fang, Yazhuo Wang, Sifei Chen, et al. 2023. “Deciphering and Advancing CAR T-Cell Therapy with Single-Cell Sequencing Technologies.” *Molecular Cancer* 22 (1): 80.

Huang, Shujun, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. 2018. “Applications of Support Vector Machine (SVM) Learning in Cancer Genomics.” *Cancer Genomics & Proteomics* 15 (1): 41–51.

Huang, Xingfan, Jana Henck, Chengxiang Qiu, Varun K. A. Sreenivasan, Saranya Balachandran, Oana V. Amarie, Martin Hrabě de Angelis, et al. 2023. “Single-Cell, Whole-Embryo Phenotyping of Mammalian Developmental Disorders.” *Nature* 623 (7988): 772–81.

Hu, Bo, Sara Lelek, Bastiaan Spanjaard, Hadil El-Sammak, Mariana Guedes Simões, Janita Mintcheva, Hananeh Aliee, et al. 2022. “Origin and Function of Activated

Fibroblast States during Zebrafish Heart Regeneration." *Nature Genetics* 54 (8): 1227–37.
James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in R*. Springer Nature.
Jin, L., and R. V. Lloyd. 1997. "In Situ Hybridization: Methods and Applications." *Journal of Clinical Laboratory Analysis* 11 (1): 2–9.
Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science (New York, N.Y.)* 349 (6245): 255–60.
Joynt, Alyssa C. M., Michelle M. Axford, Lauren Chad, and Gregory Costain. 2022. "Understanding Genetic Variants of Uncertain Significance." *Paediatrics & Child Health* 27 (1): 10–11.
Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581 (7809): 434–43.
Kashima, Yukie, Yoshitaka Sakamoto, Keiya Kaneko, Masahide Seki, Yutaka Suzuki, and Ayako Suzuki. 2020. "Single-Cell Sequencing Techniques from Individual to Multiomics Analyses." *Experimental & Molecular Medicine* 52 (9): 1419–27.
Kelly, Natalie H., Nguyen P. T. Huynh, and Farshid Guilak. 2020. "Single Cell RNA-Sequencing Reveals Cellular Heterogeneity and Trajectories of Lineage Specification during Murine Embryonic Limb Development." *Matrix Biology: Journal of the International Society for Matrix Biology* 89 (July):1–10.
Klein, Allon M., Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. 2015. "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells." *Cell* 161 (5): 1187–1201.
Kong, Lingjia, Vladislav Pokatayev, Ariel Lefkovith, Grace T. Carter, Elizabeth A. Creasey, Chirag Krishna, Sathish Subramanian, et al. 2023. "The Landscape of Immune Dysregulation in Crohn's Disease Revealed through Single-Cell Transcriptomic Profiling in the Ileum and Colon." *Immunity* 56 (12): 2855.
Korsunsky, Ilya, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. 2019. "Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony." *Nature Methods* 16 (12): 1289–96.
Koscielny, Gautier, Peter An, Denise Carvalho-Silva, Jennifer A. Cham, Luca Fumis, Rippa Gasparyan, Samiul Hasan, et al. 2017. "Open Targets: A Platform for Therapeutic Target Identification and Validation." *Nucleic Acids Research* 45 (D1): D985–94.
La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. 2018. "RNA Velocity of Single Cells." *Nature* 560 (7719): 494–98.
Larson, Nicholas Bradley, Ann L. Oberg, Alex A. Adjei, and Liguo Wang. 2023. "A Clinician's Guide to Bioinformatics for Next-Generation Sequencing." *Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer* 18 (2): 143–57.
Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks,

Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285–91.
Libbrecht, Maxwell W., and William Stafford Noble. 2015. "Machine Learning Applications in Genetics and Genomics." *Nature Reviews. Genetics* 16 (6): 321–32.
Litviňuková, Monika, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L. Worth, Eric L. Lindberg, Masatoshi Kanda, et al. 2020. "Cells of the Adult Human Heart." *Nature* 588 (7838): 466–72.
Liu, Tianbin, Jie Li, Leqian Yu, Hai-Xi Sun, Jing Li, Guoyi Dong, Yingying Hu, et al. 2021. "Author Correction: Cross-Species Single-Cell Transcriptomic Analysis Reveals Pre-Gastrulation Developmental Differences among Pigs, Monkeys, and Humans." *Cell Discovery* 7 (1): 14.
Liu, Z. L., X. Y. Meng, R. J. Bao, M. Y. Shen, J. J. Sun, W. D. Chen, F. Liu, and Y. He. 2024. "Single Cell Deciphering of Progression Trajectories of the Tumor Ecosystem in Head and Neck Cancer." *Nature Communications* 15 (1): 2595.
Luecken, Malte D., M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, et al. 2022. "Benchmarking Atlas-Level Data Integration in Single-Cell Genomics." *Nature Methods* 19 (1): 41–50.
Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9 (86): 2579–2605.
Mah, Jasmine L., and Casey W. Dunn. 2024. "Cell Type Evolution Reconstruction across Species through Cell Phylogenies of Single-Cell RNA Sequencing Data." *Nature Ecology & Evolution* 8 (2): 325–38.
Maj, Carlo, Antonia Eberts, Johannes Schumacher, and Pouria Dasmeh. 2024. "Single-Cell Analysis Reveals the Spatial-Temporal Expression of Genes Associated with Esophageal Malformations." *Scientific Reports* 14 (1): 3752.
Manivannan, Sathiyanarayanan, Corrin Mansfield, Xinmin Zhang, Karthik M. Kodigepalli, Uddalak Majumdar, Vidu Garg, and Madhumita Basu. 2021. "Single-Cell Transcriptomic Profiling Unveils Cardiac Cell-Type Specific Response to Maternal Hyperglycemia Underlying the Risk of Congenital Heart Defects." *bioRxiv*. bioRxiv. https://doi.org/10.1101/2021.05.28.446177.
Mardis, Elaine R. 2011. "A Decade's Perspective on DNA Sequencing Technology." *Nature* 470 (7333): 198–203.
Marx, Vivien. 2013. "Biology: The Big Challenges of Big Data." *Nature* 498 (7453): 255–60.
McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv [stat.ML]*. https://doi.org/10.48550/ARXIV.1802.03426.
Mitchel, Thomas. 1997. *Machine Learning (Mcgraw-Hill International Edit)*.
Moreau, Yves, and Léon-Charles Tranchevent. 2012. "Computational Tools for Prioritizing Candidate Genes: Boosting Disease Gene Discovery." *Nature Reviews. Genetics* 13 (8): 523–36.
Mungall, Christopher J., Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, et al. 2017. "The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species." *Nucleic Acids Research* 45 (D1): D712–22.
Nomura, Seitaro, Masahiro Satoh, Takanori Fujita, Tomoaki Higo, Tomokazu Sumida,

Toshiyuki Ko, Toshihiro Yamaguchi, et al. 2018. “Cardiomyocyte Gene Programs Encoding Morphological and Functional Signatures in Cardiac Hypertrophy and Failure.” *Nature Communications* 9 (1): 4435.
Otani, Hiroki, Jun Udagawa, Toshihisa Hatta, Yukiko Kagohashi, Ryuju Hashimoto, Akihiro Matsumoto, Fumio Satow, and Masayuki Nimura. 2010. “Individual Variation in Organ Histogenesis as a Causative Factor in the Developmental Origins of Health and Disease: Unnoticed Congenital Anomalies?” *Congenital Anomalies* 50 (4): 205–11.
Pejaver, Vikas, Alicia B. Byrne, Bing-Jian Feng, Kymberleigh A. Pagel, Sean D. Mooney, Rachel Karchin, Anne O’Donnell-Luria, et al. 2022. “Calibration of Computational Tools for Missense Variant Pathogenicity Classification and ClinGen Recommendations for PP3/BP4 Criteria.” *American Journal of Human Genetics* 109 (12): 2163–77.
Peng, Chengyao, Simon Dieck, Alexander Schmid, Ashar Ahmad, Alexej Knaus, Maren Wenzel, Laura Mehnert, et al. 2021. “CADA: Phenotype-Driven Gene Prioritization Based on a Case-Enriched Knowledge Graph.” *NAR Genomics and Bioinformatics* 3 (3): lqab078.
Piro, Rosario M., and Ferdinando Di Cunto. 2012. “Computational Approaches to Disease-Gene Prediction: Rationale, Classification and Successes.” *The FEBS Journal* 279 (5): 678–96.
Qi Mao, Li Wang, Ivor W. Tsang, and Yijun Sun. 2017. “Principal Graph and Structure Learning Based on Reversed Graph Embedding.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (11): 2227–41.
Qiu, Chengxiang, Junyue Cao, Beth K. Martin, Tony Li, Ian C. Welsh, Sanjay Srivatsan, Xingfan Huang, et al. 2022. “Systematic Reconstruction of Cellular Trajectories across Mouse Embryogenesis.” *Nature Genetics* 54 (3): 328–41.
Qiu, Chengxiang, Beth K. Martin, Ian C. Welsh, Riza M. Daza, Truc-Mai Le, Xingfan Huang, Eva K. Nichols, et al. 2024. “A Single-Cell Time-Lapse of Mouse Prenatal Development from Gastrula to Birth.” *Nature* 626 (8001): 1084–93.
Qiu, Xiaojie, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. 2017. “Reversed Graph Embedding Resolves Complex Single-Cell Trajectories.” *Nature Methods* 14 (10): 979–82.
Raina, Priyanka, Rodrigo Guinea, Kasit Chatsirisupachai, Inês Lopes, Zoya Farooq, Cristina Guinea, Csaba-Attila Solyom, and João Pedro de Magalhães. 2023. “GeneFriends: Gene Co-Expression Databases and Tools for Humans and Model Organisms.” *Nucleic Acids Research* 51 (D1): D145–58.
Rainio, Oona, Jarmo Teuho, and Riku Klén. 2024. “Evaluation Metrics and Statistical Tests for Machine Learning.” *Scientific Reports* 14 (1): 6086.
Rai, Vivek, Daniel X. Quang, Michael R. Erdos, Darren A. Cusanovich, Riza M. Daza, Narisu Narisu, Luli S. Zou, et al. 2020. “Single-Cell ATAC-Seq in Human Pancreatic Islets and Deep Learning Upscaling of Rare Cells Reveals Cell-Specific Type 2 Diabetes Regulatory Signatures.” *Molecular Metabolism* 32 (February):109–21.
Reed, Austin D., Sara Pensa, Adi Steif, Jack Stenning, Daniel J. Kunz, Linsey J. Porter, Kui Hua, et al. 2024. “A Single-Cell Atlas Enables Mapping of Homeostatic Cellular Shifts in the Adult Human Breast.” *Nature Genetics* 56 (4): 652–62.
Saelens, Wouter, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. 2019. “A

Comparison of Single-Cell Trajectory Inference Methods." *Nature Biotechnology* 37 (5): 547–54.
Schlüter, Agatha, Valentina Vélez-Santamaría, Edgard Verdura, Agustí Rodríguez-Palmero, Montserrat Ruiz, Stéphane Fourcade, Laura Planas-Serra, et al. 2023. "ClinPrior: An Algorithm for Diagnosis and Novel Gene Discovery by Network-Based Prioritization." *Genome Medicine* 15 (1): 68.
Shendure, Jay, and Hanlee Ji. 2008. "Next-Generation DNA Sequencing." *Nature Biotechnology* 26 (10): 1135–45.
Sikkema, Lisa, Ciro Ramírez-Suástegui, Daniel C. Strobl, Tessa E. Gillett, Luke Zappia, Elo Madissoon, Nikolay S. Markov, et al. 2023. "An Integrated Cell Atlas of the Lung in Health and Disease." *Nature Medicine* 29 (6): 1563–77.
Smajić, Semra, Cesar A. Prada-Medina, Zied Landoulsi, Jenny Ghelfi, Sylvie Delcambre, Carola Dietrich, Javier Jarazo, et al. 2022. "Single-Cell Sequencing of Human Midbrain Reveals Glial Activation and a Parkinson-Specific Neuronal State." *Brain: A Journal of Neurology* 145 (3): 964–78.
Smedley, Damian, and Peter N. Robinson. 2015. "Phenotype-Driven Strategies for Exome Prioritization of Human Mendelian Disease Genes." *Genome Medicine* 7 (1): 81.
Sreenivasan, Varun K. A., Saranya Balachandran, and Malte Spielmann. 2022. "The Role of Single-Cell Genomics in Human Genetics." *Journal of Medical Genetics* 59 (9): 827–39.
Sreenivasan, Varun K. A., Riccardo Dore, Julia Resch, Julia Maier, Carola Dietrich, Jana Henck, Saranya Balachandran, Jens Mittag, and Malte Spielmann. 2023. "Single-Cell RNA-Based Phenotyping Reveals a Pivotal Role of Thyroid Hormone Receptor Alpha for Hypothalamic Development." *Development* 150 (3). https://doi.org/10.1242/dev.201228.
Statnikov, Alexander, Lily Wang, and Constantin F. Aliferis. 2008. "A Comprehensive Comparison of Random Forests and Support Vector Machines for Microarray-Based Cancer Classification." *BMC Bioinformatics* 9 (July):319.
Steward, Charles A., Alasdair P. J. Parker, Berge A. Minassian, Sanjay M. Sisodiya, Adam Frankish, and Jennifer Harrow. 2017. "Genome Annotation for Clinical Genomic Diagnostics: Strengths and Weaknesses." *Genome Medicine* 9 (1): 49.
Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. 2017. "Simultaneous Epitope and Transcriptome Measurement in Single Cells." *Nature Methods* 14 (9): 865–68.
Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177 (7): 1888–1902.e21.
Thölke, Philipp, Yorguin-Jose Mantilla-Ramos, Hamza Abdelhedi, Charlotte Maschke, Arthur Dehgan, Yann Harel, Anirudha Kemtur, et al. 2023. "Class Imbalance Should Not Throw You off Balance: Choosing the Right Classifiers and Performance Metrics for Brain Decoding with Imbalanced Data." *NeuroImage* 277 (August):120253.
Thorpe, Erin, Taylor Williams, Chad Shaw, Evgenii Chekalin, Julia Ortega, Keisha

Robinson, Jason Button, et al. 2024. “The Impact of Clinical Genome Sequencing in a Global Population with Suspected Rare Genetic Disease.” *American Journal of Human Genetics* 111 (7): 1271–81.
Tran, Hoa Thi Nhu, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. 2020. “A Benchmark of Batch-Effect Correction Methods for Single-Cell RNA Sequencing Data.” *Genome Biology* 21 (1): 12.
Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. 2014. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells.” *Nature Biotechnology* 32 (4): 381–86.
Tu, Zhidong, Li Wang, Min Xu, Xianghong Zhou, Ting Chen, and Fengzhu Sun. 2006. “Further Understanding Human Disease Genes by Comparing with Housekeeping Genes and Other Genes.” *BMC Genomics* 7 (February):31.
Velmeshev, Dmitry, Lucas Schirmer, Diane Jung, Maximilian Haeussler, Yonatan Perez, Simone Mayer, Aparna Bhaduri, Nitasha Goyal, David H. Rowitch, and Arnold R. Kriegstein. 2019. “Single-Cell Genomics Identifies Cell Type-Specific Molecular Changes in Autism.” *Science (New York, N.Y.)* 364 (6441): 685–89.
Vidal, Ramon, Julian Uwe Gabriel Wagner, Caroline Braeuning, Cornelius Fischer, Ralph Patrick, Lukas Tombor, Marion Muhly-Reinholz, et al. 2019. “Transcriptional Heterogeneity of Fibroblasts Is a Hallmark of the Aging Heart.” *JCI Insight* 4 (22). https://doi.org/10.1172/jci.insight.131092.
Wojcik, Monica H., Gabrielle Lemire, Eva Berger, Maha S. Zaki, Mariel Wissmann, Wathone Win, Susan M. White, et al. 2024. “Genome Sequencing for Diagnosing Rare Diseases.” *The New England Journal of Medicine* 390 (21): 1985–97.
Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. “SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis.” *Genome Biology* 19 (1): 15.
Wolock, Samuel L., Romain Lopez, and Allon M. Klein. 2019. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data.” *Cell Systems* 8 (4): 281–91.e9.
Yanai, Itai, and Eva Chmielnicki. 2017. “Computational Biologists: Moving to the Driver’s Seat.” *Genome Biology* 18 (1): 223.
Zhang, Weiqi, Shu Zhang, Pengze Yan, Jie Ren, Moshi Song, Jingyi Li, Jinghui Lei, et al. 2020. “A Single-Cell Transcriptomic Landscape of Primate Arterial Aging.” *Nature Communications* 11 (1): 2202.
Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January):14049.
Zolotareva, Olga, and Maren Kleine. 2019. “A Survey of Gene Prioritization Tools for Mendelian and Complex Human Diseases.” *Journal of Integrative Bioinformatics* 16 (4). https://doi.org/10.1515/jib-2018-0069.

8. List of publications

1. Janjetovic, Snjezana, Juliane Hinke, **Saranya Balachandran**, Nuray Akyüz, Petra Behrmann, Carsten Bokemeyer, Judith Dierlamm, and Eva Maria Murga Penas. 2022. **'Non-Random Pattern of Integration for Epstein-Barr Virus with Preference for Gene-Poor Genomic Chromosomal Regions into the Genome of Burkitt Lymphoma Cell Lines'. Viruses 14(1):86.** doi: 10.3390/v14010086.

2. Smajić, Semra, Cesar A. Prada-Medina, Zied Landoulsi, Jenny Ghelfi, Sylvie Delcambre, Carola Dietrich, Javier Jarazo, Jana Henck, **Saranya Balachandran**, Sinthuja Pachchek, Christopher M. Morris, Paul Antony, Bernd Timmermann, Sascha Sauer, Sandro L. Pereira, Jens C. Schwamborn, Patrick May, Anne Grünewald, and Malte Spielmann. 2022. **'Single-Cell Sequencing of Human Midbrain Reveals Glial Activation and a Parkinson-Specific Neuronal State'. Brain 145(3):964–78.** doi: 10.1093/brain/awab446.

3. Sreenivasan, Varun K. A., **Saranya Balachandran**, and Malte Spielmann. 2022. **'The Role of Single-Cell Genomics in Human Genetics'. Journal of Medical Genetics 59(9):827–39.** doi: 10.1136/jmedgenet-2022-108588.

4. Sreenivasan, Varun K. A., Riccardo Dore, Julia Resch, Julia Maier, Carola Dietrich, Jana Henck, **Saranya Balachandran**, Jens Mittag, and Malte Spielmann. 2023. **'Single-Cell RNA-Based Phenotyping Reveals a Pivotal Role of Thyroid Hormone Receptor Alpha for Hypothalamic Development'. Development 150(3):dev201228.** doi: 10.1242/dev.201228.

5. Boschann, Felix, Ozgur Cogulu, Davut Pehlivan, **Saranya Balachandran**, Pedro Vallecillo-Garcia, Christopher M. Grochowski, Nils R. Hansmeier, Zeynep H. Coban Akdemir, Cesar A. Prada-Medina, Ayca Aykut, Björn Fischer-Zirnsak, Simon Badura, Burak Durmaz, Ferda Ozkinay, René Hägerling, Jennifer E. Posey, Sigmar Stricker, Gabriele Gillessen-Kaesbach, Malte Spielmann, Denise

Horn, Knut Brockmann, James R. Lupski, Uwe Kornak, and Julia Schmidt. 2023. **'Biallelic Variants in ADAMTS15 Cause a Novel Form of Distal Arthrogryposis'. Genetics in Medicine 25(5):100799.** doi: 10.1016/j.gim.2023.100799.

6. **Balachandran, Saranya**, Jelena Pozojevic, Varun K. A. Sreenivasan, and Malte Spielmann. 2023. **'Comparative Single-Cell Analysis of the Adult Heart and Coronary Vasculature'. Mammalian Genome 34(2):276–84.** doi: 10.1007/s00335-022-09968-7.

7. Huang, Xingfan, Jana Henck, Chengxiang Qiu, Varun K. A. Sreenivasan, **Saranya Balachandran**, Oana V. Amarie, Martin Hrabě De Angelis, Rose Yinghan Behncke, Wing-Lee Chan, Alexandra Despang, Diane E. Dickel, Madeleine Duran, Annette Feuchtinger, Helmut Fuchs, Valerie Gailus-Durner, Natja Haag, Rene Hägerling, Nils Hansmeier, Friederike Hennig, Cooper Marshall, Sudha Rajderkar, Alessa Ringel, Michael Robson, Lauren M. Saunders, Patricia Da Silva-Buttkus, Nadine Spielmann, Sanjay R. Srivatsan, Sascha Ulferts, Lars Wittler, Yiwen Zhu, Vera M. Kalscheuer, Daniel M. Ibrahim, Ingo Kurth, Uwe Kornak, Axel Visel, Len A. Pennacchio, David R. Beier, Cole Trapnell, Junyue Cao, Jay Shendure, and Malte Spielmann. 2023. **'Single-Cell, Whole-Embryo Phenotyping of Mammalian Developmental Disorders'. Nature 623(7988):772–81.** doi: 10.1038/s41586-023-06548-w.

8. **Balachandran, Saranya,** Cesar A. Prada-Medina, Martin A. Mensah, Juliane Glaser, Naseebullah Kakar, Inga Nagel, Jelena Pozojevic, Enrique Audain, Marc-Phillip Hitz, Martin Kircher, Varun K. A. Sreenivasan, and Malte Spielmann. 2024. **'STIGMA: Single-Cell Tissue-Specific Gene Prioritization Using Machine Learning'. American Journal of Human Genetics S0002-9297(23)00443-3**. doi: 10.1016/j.ajhg.2023.12.011.

9. Figueroa, Karla P., Caspar Gross, Elena Buena-Atienza, Sharan Paul, Mandi Gandelman, Naseebullah Kakar, Marc Sturm, Nicolas Casadei, Jakob Admard, Joohyun Park, Christine Zühlke, Yorck Hellenbroich, Jelena Pozojevic, **Saranya Balachandran**, Kristian Händler, Simone Zittel, Dagmar Timmann, Friedrich Erdlenbruch, Laura Herrmann, Thomas Feindt, Martin Zenker, Thomas Klopstock, Claudia Dufke, Daniel R. Scoles, Arnulf Koeppen, Malte Spielmann, Olaf Riess, Stephan Ossowski, Tobias B. Haack, and Stefan M. Pulst. 2024. **'A GGC-Repeat Expansion in ZFHX3 Encoding Polyglycine Causes Spinocerebellar Ataxia Type 4 and Impairs Autophagy'. Nature Genetics 56(6):1080–89.** doi: 10.1038/s41588-024-01719-5.

9.Acknowledgements

First and foremost, I would like to thank my supervisor Malte Spielmann for supporting me during my PhD. He has encouraged me to do good science, and provided a comfortable environment to learn and grow. Thank you for encouraging me to go to conferences, to build my network and explore new research ideas.
I would like to thank Martin for the discussions on dealing with biases and variability in the data. You provided a safe environment to all my questions. I am grateful for inviting me to your lab retreats and workshops which was a great learning opportunity. I also want to thank Varun Sreenivasan for all the discussions we had, for providing me ideas on visualizations, ways to improve my analysis and encouraging me to pursue my ideas. Thank you, Cesar, for providing me a smooth transition into the team and your guidance during my preliminary days of PhD. Thank you, Veronica, for proof-reading my thesis and to Naseeb and Jelena for bringing the biological insights to my computational work. I would like to thank all the members of Spielmann lab for creating a great atmosphere.
A special thanks to my family, who have been my constant source of strength. I want to thank my parents and sister for believing in me, and encouraging me to explore my career opportunities. First and foremost, I want to thank my husband for supporting me throughout. Thank you for always believing in me and encouraging me and supporting me emotionally. Thanks to my son for the love and bringing joy to our lives. Lastly, to my friends, thank you for always standing by me and uplifting me.
This journey would not be possible without your support. You all made me believe in myself and I am grateful to you all.