

mgr inż. Maciej Migdał

**Modeling of transcription factors influence on gene
expression based on data obtained using next-generation
sequencing methods**

**Rozprawa na stopień doktora nauk medycznych i nauk o zdrowiu
w dyscyplinie nauki medyczne**

Promotor: dr hab. Cecilia Winata

Laboratorium Genomiki Rozwoju Danio Pręgowanego,
Międzynarodowy Instytut Biologii Molekularnej i Komórkowej
w Warszawie



Obrona rozprawy doktorskiej przed Radą Dyscypliny Nauk Medycznych
Warszawskiego Uniwersytetu Medycznego

Warszawa 2023 r.

Keywords: ATAC-seq; Bioinformatics; Cardiomyocytes; Epigenomics; Epithelial–mesenchymal transition; Genomics; Gene expression; Gene regulation; Heart development; Liver injury; RNA-seq; Transcriptomics; Transcription factors.

Acknowledgements

I am deeply grateful for the opportunity that I was given by Cecilia Winata to join her lab for an internship and later to pursue a PhD degree under her supervision. I am thankful to Norber Dojer for helping to supervise my project at its early stages, and for the opportunity to present my research at the Warsaw University. The current and former members of ZDG laboratory have been a source of support and knowledge. I am grateful for all the discussions, support and time spent together, many of you have become my friends and I hope to keep in touch with you! #ZDGBazaPro. Special thanks to Karim Abu Nahia for sharing a desk with me all these years and for all the late nights we spent working together. I am much obliged to Maciej Łapiński and Leszek Prysycz, thank you for all the lunches, knowledge, enthusiasm and opportunities you have shared with me. I can't be more thankful to Michał Pawlak for all the projects we worked on and conferences we attended together. Thank you Agata Sulej for morning coffee and help with my lab work. Thank you Basia Uszczyńska-Ratajczak for a helping hand when I needed it. Thank you Adrianna Pakuła for all the cloning and injections. My deepest thanks to Erik Arner and Bogumił Kaczkowski for having me for the first ever remote internship at RIKEN. Finally, I want to express my gratitude to all the kind and supportive people that I had the pleasure of meeting over the course of my PhD studies. Though I am not able to list each and every person, please know that each of you have had a hand in making my journey successful and have shaped me into the person I am today. Thanks!

List of included articles

1. Pawlak M, Kedzierska KZ, Migdal M, Nahia KA, Ramilowski JA, Bugajski L, Hashimoto K, Marconi A, Piwocka K, Carninci P, Winata CL. Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development. *Genome Res.* 2019 Mar;29(3):506-519. doi: 10.1101/gr.244491.118.
2. Migdał M, Tralle E, Nahia KA, Bugajski Ł, Kędzierska KZ, Garbicz F, Piwocka K, Winata CL, Pawlak M. Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift. *BMC Genomics.* 2021 Dec18;22(1):904. doi: 10.1186/s12864-021-08173-1.
3. Migdał M, Arakawa T, Takizawa S, Furuno M, Suzuki H, Arner E, Winata CL, Kaczkowski B. xcore: an R package for inference of gene expression regulators. *BMC Bioinformatics.* 2023 Jan 11;24(1):14. doi: 10.1186/s12859-022-05084-0.

Table of contents

Acknowledgements	3
List of included articles	4
Table of contents	5
List of abbreviations	6
Abstract in Polish	7
Abstract in English	8
Introduction	9
References	17
Assumptions and aim of the work	22
Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development	23
Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift	37
xcore: an R package for inference of gene expression regulators	53
Summary and conclusions	65
Authors contributions statements	67

List of abbreviations

ATAC-seq	assay for transposase-accessible chromatin with high-throughput sequencing
CAGE	cap analysis gene expression
ChIP-seq	chromatin immunoprecipitation followed by sequencing
DNA	deoxyribonucleic acid
DNN	deep neural networks
EC	endothelial cells
EMT	epithelial-mesenchymal transition
FACS	fluorescence-activated cell sorting
LSEC	liver sinusoidal endothelial cells
NGS	next-generation sequencing
PWM	position weight matrix
RE	DNA regulatory element
RNA	ribonucleic acid
RNA-seq	RNA sequencing
SNP	single nucleotide polymorphism
SOM	self-organizing map
TAA	thioacetamide
TF	transcription factor
TFBS	transcription factor binding sites
TGF β	transforming growth factor beta

Abstract in Polish

Zbiór artykułów stanowiący podstawę niniejszej rozprawy doktorskiej składa się z trzech publikacji: Pawlak et al. „Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development”, *Genome Res.*, 2019; Migdał et al. „Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift”, *BMC Genomics*, 2021; oraz Migdał et al. „xcore: an R package for inference of gene expression regulators”, *BMC Bioinformatics*, 2022. Celem prac było poznanie mechanizmu regulacji transkrypcji genów, w szczególności identyfikacja czynników transkrypcyjnych (TF) i elementów regulatorowych DNA (RE) budujących układy kontroli transkrypcji genów leżące u podstaw różnorodnych procesów komórkowych. W tym celu, wykorzystałem dane eksperymentalne uzyskane metodą sekwencjonowania następnej generacji z organizmów na różnych poziomach złożoności biologicznej, w tym *in vivo* z Danio pręgowanego (*Danio rerio*) oraz *in vitro* z unieśmiertelnionych linii komórkowych człowieka. Do przetwarzania zebranych danych wykorzystałem publicznie dostępne oraz opracowane przeze mnie narzędzia bioinformatyczne służące do: przetwarzania surowych danych genomicznych, analizy wzbogaceń motywów TF czy uczenia maszynowego z zastosowaniem penalizowanych modeli liniowych. Głównym założeniem pracy jest związek przyczynowo-skutkowy pomiędzy TF, RE a transkrypcją kontrolowanych przez nie genów. Bazując na powyższym założeniu wybrane prace badają mechanizm regulacji transkrypcji wykorzystując informacje o poziomach ekspresji genów i aktywności RE.

Rozdział “Introduction” zawiera krótkie wprowadzenie z zakresu regulacji ekspresji genów oraz stosowanych przeze mnie metod bioinformatycznych wykorzystywanych do analizy danych genomicznych. Zawarte w nim trzy podrozdziały streszczają wybrane prace naukowe, z podkreśleniem wspólnego tematu przewodniego regulacji transkrypcji oraz różnic w wykorzystywanych metodach analizy danych. Kolejne trzy rozdziały zawierają kopie wybranych artykułów opublikowanych w czasopiśmie naukowym. Rozprawa kończy się rozdziałem „Summary and conclusions”, który podsumowuje uzyskane wyniki. Oświadczenia współautorów dotyczące każdej publikacji znajdują się na końcu rozprawy doktorskiej.

Abstract in English

The presented collection of articles providing the basis for this doctoral dissertation consists of three publications: Pawlak et al. „Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development”, *Genome Res.*, 2019; Migdał et al. „Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift”, *BMC Genomics*, 2021; and Migdał et al. „xcore: an R package for inference of gene expression regulators”, *BMC Bioinformatics*, 2023. The selected publications aimed to elucidate the principles of gene regulation, emphasizing on the identification of transcription factors (TFs) and DNA regulatory elements (REs) which constitute gene regulatory networks underlying various cellular processes. To achieve this principal aim, we utilized next-generation sequencing (NGS) data collected from organisms at various levels of biological complexity, including *in vivo* data from zebrafish (*Danio rerio*) and *in vitro* human cell lines. These were analyzed using either established bioinformatic algorithms or those which I developed, including NGS data processing, motif enrichment analysis, and machine learning using penalized linear models. The key assumption of the work is the causal relationship between TF, RE, and the transcriptional outcome of their target genes. Based on this assumption, the analytical frameworks exemplified in this collection of articles approaches the problem of transcriptional regulation mechanism using information on gene expression and the activity of RE.

The “Introduction” chapter provides a brief introduction to the topic of gene regulation and the approaches I used in the bioinformatic analysis of the experimental data obtained from experiments employing NGS. Its three subsections give an overview of the included articles, emphasizing on the common gene regulation theme of these studies and the differences in the employed analytical methodologies. The following three chapters contain the copies of the included articles; the associated supplementary materials can be accessed in their on-line forms. Finally, the dissertation is concluded with a “Summary and conclusions” chapter. The co-authors contribution statements for each publication can be found attached following the last chapter.

Introduction

Multicellular organisms are made up of many types of cells with specialized morphology and function. These cells constitute various organ systems which perform critical functions for the organism's survival. The identity of a cell is mostly dictated by the cell's specific protein composition. All the proteins that a cell can produce are encoded by the genes in the DNA (deoxyribonucleic acid). However, the DNA is not a direct template for protein synthesis. Instead, genes are first transcribed into ribonucleic acid (RNA) in a process called transcription. Next, instructions written in the RNA are used to synthesize proteins in the process of translation. This flow of genetic information, from DNA through RNA to protein, is called gene expression and constitutes the central dogma of molecular biology (Crick 1970). Remarkably, all cells in an organism carry an identical copy of the DNA, and yet cells can take a variety of different states. This is achieved thanks to the regulation of gene expression: while all cells carry the complete set of genetic information, not all of that information is actively expressed. Cells of different types express different sets of genes which underlie their different characteristics (Alberts et al. 2002). The dynamic character of gene expression does not only differentiate cell types; gene expression of a particular cell changes during its development, cell cycle or in response to external stimuli such as toxic chemicals.

The dynamic usage of genetic information is achieved through gene expression regulation. In eukaryotic cells, gene expression regulation can occur at multiple levels throughout the gene expression process. Three of the most recognized modes of regulation are: transcriptional regulation, post-transcriptional modifications and translational regulation (Stryer et al. 2018). The three publications included in this dissertation focus solely on the first layer of gene regulation - transcriptional regulation - which is thought to play a fundamental role in establishing cell type diversity. Transcriptional regulation is implemented by means of a complex system of interactions between TFs and REs. TFs are a class of proteins that participate in the assembly and regulation of the basal transcriptional machinery. Crucial to their function is their ability to bind DNA in a sequence specific manner (Latchman 1997). TFs commonly exert their regulatory role through binding with cofactors that contribute to transcription regulation. A well-known example is the Mediator complex which binds TFs to stimulate or repress the phosphorylation of polymerase II, effectively facilitating transcription initiation (Cramer 2019). REs contain regulatory signals encoded in the DNA sequence, together with genes, in *cis*. They serve as binding sites for TFs which bind to those sites by

recognizing specific DNA sequences. The TF binding specificity has been shown to be evolutionary conserved between distantly related organisms, such as insects and mammals (Nitta et al. 2015). Promoters and enhancers are the two classes of REs that play essential roles in gene transcription (Alberts et al. 2002). Promoters are defined as the DNA sequence located directly upstream of the transcription start site. They contain the signals necessary to stimulate the transcription initiation complex assembly as well as the binding sites for regulatory TFs (Cramer 2019). Enhancers are traditionally defined as sequences able to enhance gene transcription and are not restricted by the distance from the target gene they regulate. They are brought into close proximity to the promoter of their target genes through DNA looping and participate in gene expression regulation by providing binding sites to TFs that can upregulate or downregulate target gene transcription (Pennacchio et al. 2013). Each gene has its promoter and associated enhancers that provide binding sites for the TFs. In this way RE sequence defines which TFs can participate in the regulation of a given gene's expression.

Transcriptional regulation is further fine-tuned by the dynamic character of chromatin accessibility and TF activity, which provides the mechanism for spatio-temporal gene expression outcome. Chromatin accessibility expresses a physical property of DNA structure that defines whether a particular DNA region is accessible to proteins, such as TFs. In a simplified picture, we can consider the DNA as organized into regions of differing accessibility. Of particular significance is the accessibility of DNA regions harboring REs as they actively participate in gene regulation (ENCODE Project Consortium 2012). Chromatin accessibility is determined by many reversible processes, such as histone modification and DNA methylation (Klemm, Shipony, and Greenleaf 2019). By regulating the accessibility of specific REs cells can regulate their gene expression in a complex manner. Additionally, transcription is modulated by TF activity, which expresses the character and strength of TF effect on the genes they regulate. TF activity stems not only from the intrinsic properties of the particular TF, but also from other factors such as TF abundance.

The concept of transcriptional control was first established over 60 years ago (Jacob and Monod 1961). The large number of cellular, molecular, and structural studies conducted since then established a detailed knowledge on the factors involved in the process of transcription and their structural organization, providing mechanistic insights into gene transcription regulation (Cramer 2019). The complex interplay between these factors allow flexible use of genetic information. For instance, it allows cells to differentiate into a variety of specialized cell types such as cardiomyocytes or liver

sinusoidal endothelial cells (LSEC). Cardiomyocytes are one of the heart's main building blocks playing a fundamental role in heart contraction. Their cellular structure highly underlines their functional properties, which can be seen in the presence of Ca^{2+} storing sarcoplasmic reticulum and contractile apparatus composed of sarcomeres driving their contraction, leading to beating of the heart. The key determinants of the cardiomyocytes expression program are the NKX2-5, GATA4, and TBX5 transcription factors (Lyons et al. 1995; Durocher et al. 1997; Bruneau et al. 2001). On the other hand, LSECs constitute the walls of the liver's sinusoidal blood vessels and provide the interface between hepatocytes or hepatic stellate cells and the blood vessel lumen. They are characterized by a lack of basement membrane and the presence of a permeable fenestrae facilitating their role as a selective barrier (De Leeuw, Brouwer, and Knook 1990). At the regulatory level, a combination of TFs is involved in establishing LSEC identity, such as GATA4, C-MAF or TCFEC (de Haan et al. 2020). The commonly accepted view is that transcriptional regulation through combinatorial TF binding, is the key mechanism underlying the specification of cell identity in eukaryotes (Takahashi and Yamanaka 2006). While the main drivers of many cell expression programs are known, establishing the complete knowledge on the TFs and REs underlying gene regulatory networks at play in various cellular processes is still an open challenge. The works presented in this doctoral dissertation attempts to address this problem using computational analysis of transcriptome and epigenome data in different biological contexts, and further develop a computational framework for predictive modeling of TF regulatory activity.

Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development (Pawlak et al., 2019)

The first work included in this collection of articles, (Pawlak et al. 2019) aims to elucidate the gene regulatory network underlying heart development using *Danio rerio* as a model organism. The principal design of the study involved the characterization of the transcriptome and epigenome landscape at the early stages of heart formation, focusing on the identification of putative TFs associated with the observed variability in gene expression. To this end, the study produced assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) and RNA sequencing (RNA-seq) data from a population of cardiomyocytes obtained by fluorescence-activated cell sorting (FACS). To interpret the gene expression dynamics, the expression information collected at different time points and from wild-type and mutant conditions were

subjected to clustering analysis. Cluster analysis refers to a broad variety of techniques aiming at detecting subgroups of similar objects, or clusters, in the dataset. They define a similarity measure, such as Euclidean distance, which can be used to measure the distances between various data points (reviewed in James et al. 2021). In the case of gene regulation studies, clustering analysis is applied to identify groups of co-expressed genes based on gene expression data across different conditions or time points in a given biological process (eg. cell cycle, organ development). Co-expressed genes are defined as genes sharing expression patterns across these points. Such groups of genes suggest functional significance as they likely represent interconnected genes involved in common biological processes. In (Pawlak et al. 2019) we have identified several clusters of co-expressed genes enriched in genes associated with the “heart development” Gene Ontology term. These clusters included the genes *gata5* and *nkx2.5* which encode for TFs implicated in the specification of cardiac cell identity (Reiter et al. 1999, 5; Lyons et al. 1995; Durocher et al. 1997). Since gene expression is controlled by its REs, it is reasonable to expect that co-expressed genes share their underlying regulatory grammar in the form of a combination of active TF binding events. This notion is strongly supported by numerous observations (Allocco, Kohane, and Butte 2004; Br̄azma et al. 1998), and recently demonstrated in simulations using synthetic gene regulatory networks (Yin et al. 2021). TFs interact with DNA sequences by binding to short subsequences which in vertebrates are usually between 10-14 nucleotides long. The sequences recognized by TFs can be generalized and represented in the form of a motif. One of the simplest motif representations is the position weight matrix (PWM) that describes TF binding preferences by giving the likelihood of a particular nucleotide occurring at each position in the sequence (reviewed in Rzeszowska-Wolny and Jaksik 2010). These motifs can also be used to predict putative TF binding sites in the whole genome of an organism by scanning for genome subsequences that are similar enough with the motif. However, TF motifs do not provide enough information to accurately predict true transcription factor binding sites (TFBS) *in-vivo*, resulting in high false positives rates. Motif enrichment analysis aims to overcome this drawback by looking for motifs shared by a group of co-expressed genes in order to identify TFs with a putative regulatory role. A common approach is to perform motif enrichment analysis on a group of co-expressed genes discovered using cluster analysis (Frith et al. 2004). In (Pawlak et al. 2019), I used the corresponding ATAC-seq data to identify accessible REs within the promoters of co-expressed genes. To identify active REs, I implemented a bioinformatic pipeline which takes the raw sequencing reads to infer the genomic locations of accessible chromatin

regions. Such regions are commonly associated with different transcription related activities, such as active promoters and enhancers. Using the motif enrichment analysis tool HOMER (Heinz et al. 2010), I complemented the obtained co-expression gene clusters with TF motifs found to be enriched in the identified promoter-proximal open chromatin regions of co-expressed genes. Together, co-expression clusters and their motif description illustrate the dynamics of cardiomyocytes transcriptomic and gene regulatory landscape. Our results suggest a major transcriptomic and epigenomic shift towards more cell type specific expression patterns between linear heart tube formation (24 hpf) and heart looping (48 hpf). Analysis of data collected using *gata5*, *hand2*, and *tbx5* mutants, in which heart development is affected, revealed only minor changes in the identified promoter-proximal REs, suggesting the predominant role of distal regulatory elements in cardiomyocytes maturation.

Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift (Migdał et al., 2021)

The second work included in the collection of articles, (Migdał et al. 2021) studies the transcriptomic and epigenomic response to early hepatotoxic liver injury in the zebrafish as a model organism. Using thioacetamide (TAA) injections, we induced liver injury in adult zebrafish. We then collected cell-type specific transcriptomic and epigenomic data from untreated and treated animals using RNA-seq and ATAC-seq techniques on populations of hepatocytes, endothelial cells (EC) and hepatic stellate cells isolated by FACS. While focusing on a different biological context, the main analytical themes follow the ones explored in (Pawlak et al. 2019). The transcriptomic analysis employs gene co-expression clustering based on RNA-seq data collected across different cell types and treatment. This was obtained using a clustering technique called self organizing map (SOM) (Löffler-Wirth, Kalcher, and Binder 2015). In addition to clustering, SOM provides a lower-dimensional representation of a complex gene expression dataset that can be presented as a 2-D image of co-expressed genes and their expression levels (Migdał et al. 2021; Fig. 1e). Analysis of the obtained clustering representation indicated that the first liver cell population exposed to hepatotoxin are EC as they were the most affected at the transcriptomic level. Among the identified co-expression clusters, one in particular (cluster B) contained genes which showed the highest upregulation in response to TAA treatment in EC. Interestingly, this cluster contains genes related to metabolic and redox processes, including 20 members of the

cytochrome p450 superfamily. Similarly as in (Pawlak et al. 2019), we complemented the transcriptional picture with epigenomic information obtained from the ATAC-seq data. In this study we generated three ATAC-seq replicates per condition. Surprisingly, only few peak calling tools provide an option to include replicates, and most implementations are limited to only two replicates. To overcome this limitation, I updated our ATAC-seq pipeline so that the information from any number of replicates can be combined at the peak calling step using Fisher's method (Mosteller and Fisher 1948). Analysis of the dynamics of promoter accessibility revealed that it followed the patterns observed at the transcriptome level. We observed the largest change in promoter accessibility in EC. Moreover, we found the largest number of gene promoters with an increased accessibility in cluster B. These results suggest promoter accessibility remodeling as an important mechanism driving the transcriptional response to liver injury. To infer the mechanism of gene regulation in response to TAA treatment, I performed motif enrichment analysis at the differentially accessible chromatin regions within the promoter regions of co-expressed gene clusters. This analysis revealed enrichment of motifs recognized by known transcriptional activators, including pioneer factors *FOXA1* and *FOXA3* (Zaret and Carroll 2011), revealing potential players behind the observed promoter accessibility dynamics and a hypothetical mechanism underlying liver injury response.

xcore: an R package for inference of gene expression regulators (Migdał et al. 2022)

The last publication included in this collection of articles, (Migdał et al. 2023) takes advantage of the methodological experiences of the previous two and applies a predictive modeling approach to the gene expression regulation problem. Motif enrichment analysis results can be assessed through cross-referencing with the current state of knowledge on the studied biological process and known TFs interactions. However, creating a reliable cross-reference requires comprehensive annotation and an extensive knowledge about the mechanism of the TFs of interest, which is often sparse or non-existing (Tompa et al. 2005). An alternative approach to this problem is offered by predictive modeling methodology. Predictive modeling uses mathematical models to describe the problem at hand using collected experimental data. By expressing the problem in mathematical terms, it is possible to measure how well the model explains the observed data. Additionally, mathematical models can be used both to understand the investigated problem (eg. identify the putative regulatory motifs) and to make predictions

about the data that was not yet observed (eg. predict the regulatory effect of a mutation). (Migdał et al. 2023) builds on the concepts of gene expression prediction modeling methodology that already exists in the literature (FANTOM Consortium et al. 2009; Ouyang, Zhou, and Wong 2009; Natarajan et al. 2012; McLeay et al. 2012; Balwierz et al. 2014; Schmidt et al. 2017). Here, I developed the *xcORE* R package that implements a flexible gene expression prediction framework. This tool allows modeling of gene expression directly from chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments, instead of motifs, capitalizing on the large ChIP-seq databases available for mouse and human. To this end, each gene is described by its expression and the presence or absence of specific TF binding sites in its promoter. Using penalized linear regression, *xcORE* estimates the activities of TFs. Such information can be used to generate testable hypotheses about the studied system. To validate and test the performance of the *xcORE*, we applied it to the new cap analysis gene expression (CAGE) dataset (GSE17708) from transforming growth factor beta (TGFβ)-induced epithelial-mesenchymal transition (EMT) time-series experiment. The analyses revealed that *xcORE* could identify a larger number of EMT regulators as compared to the state-of-the-art motif based ISMARA tool (Balwierz et al. 2014). The *xcORE* R package and its user guide are publicly available on Bioconductor and GitHub (<https://github.com/bkaczkowski/xcORE>).

The discipline of gene regulation continues to shape itself as an exciting interdisciplinary field combining different areas of research in biology, mathematics and engineering. Progressive expansion in the knowledge of mechanisms underlying the dynamic expression of genes raises hopes to further advance medicine and other areas of life-sciences. Based on the available evidence, it is now established that many diseases, such as cancer or congenital defects, might be explained by underlying mutations in genes encoding transcription regulation machinery. As an example, the oncogenic transcription factor TAL1 is implicated in a large number of T-cell acute lymphoblastic leukemia cases. It acts in tandem with several other transcription factors to activate the TAL-1 regulated oncogenic programs (Sanda et al. 2012). The presence of single nucleotide polymorphisms (SNPs) in RE have also been identified as a causative factor for various human diseases, including cancer or congenital heart diseases (Maurano et al. 2012). The three publications included in this dissertation studied gene expression regulation using two approaches: classic descriptive bioinformatic analysis and predictive modeling. Recent work shows that a more accurate representation of gene regulation can be provided by deep neural networks (DNN) methods. DNN are mathematical models consisting of

connected units called artificial neurons organized in multiple layers. Key features of DNN is their ability to model complex non-linear relationships and to automatically discover relevant representations from the raw data (LeCun, Bengio, and Hinton 2015). Future development of these approaches and their applications are expected to revolutionize the fields of genomics and personalized medicine. Indeed, DNN-based studies have already provided interesting insights into the system level organization of transcriptional regulation. To provide a few examples, deep neural networks using only 125 TFs could explain around 80% of the transcriptomics variation in a broad range of experimental datasets (Magnusson, Tegnér, and Gustafsson 2022). Recently, DeepMind published a neural network that helps to understand the mechanism of disease associated SNPs (Avsec et al. 2021). These examples show the great potential of deep neural networks for advancing our understanding of gene regulation mechanisms and to create tools that could help guide further experimental design.

References

- Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. "Chapter 7. Control of Gene Expression." In *Molecular Biology of the Cell. 4th Edition*. Garland Science.
- Allocco, Dominic J, Isaac S Kohane, and Atul J Butte. 2004. "Quantifying the Relationship between Co-Expression, Co-Regulation and Gene Function." *BMC Bioinformatics* 5 (18). <https://doi.org/10.1186/1471-2105-5-18>.
- 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. <https://doi.org/10.1038/s41592-021-01252-x>.
- Balwierz, Piotr J., Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan, and Erik van Nimwegen. 2014. "ISMARA: Automated Modeling of Genomic Signals as a Democracy of Regulatory Motifs." *Genome Research* 24 (5): 869–884. <https://doi.org/10.1101/gr.169508.113>.
- Brāzma, Alvis, Inge Jonassen, Jaak Vilo, and Esko Ukkonen. 1998. "Predicting Gene Regulatory Elements in Silico on a Genomic Scale." *Genome Research* 8 (11): 1202–15.
- Bruneau, B. G., G. Nemer, J. P. Schmitt, F. Charron, L. Robitaille, S. Caron, D. A. Conner, et al. 2001. "A Murine Model of Holt-Oram Syndrome Defines Roles of the T-Box Transcription Factor Tbx5 in Cardiogenesis and Disease." *Cell* 106 (6): 709–21. [https://doi.org/10.1016/s0092-8674\(01\)00493-7](https://doi.org/10.1016/s0092-8674(01)00493-7).
- Cramer, Patrick. 2019. "Organization and Regulation of Gene Transcription." *Nature* 573 (7772): 45–54. <https://doi.org/10.1038/s41586-019-1517-4>.
- Crick, F. 1970. "Central Dogma of Molecular Biology." *Nature* 227 (5258): 561–63. <https://doi.org/10.1038/227561a0>.
- De Leeuw, A. M., A. Brouwer, and D. L. Knook. 1990. "Sinusoidal Endothelial Cells of the Liver: Fine Structure and Function in Relation to Age." *Journal of Electron Microscopy Technique* 14 (3): 218–36. <https://doi.org/10.1002/jemt.1060140304>.
- Di Tommaso, Paolo, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35 (4): 316–19. <https://doi.org/10.1038/nbt.3820>.

- Durocher, D, F Charron, R Warren, R J Schwartz, and M Nemer. 1997. “The Cardiac Transcription Factors Nkx2-5 and GATA-4 Are Mutual Cofactors.” *The EMBO Journal* 16 (18): 5687–96. <https://doi.org/10.1093/emboj/16.18.5687>.
- ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74. <https://doi.org/10.1038/nature11247>.
- FANTOM Consortium, Harukazu Suzuki, Alistair R. R. Forrest, Erik van Nimwegen, Carsten O. Daub, Piotr J. Balwiercz, Katharine M. Irvine, et al. 2009. “The Transcriptional Network That Controls Growth Arrest and Differentiation in a Human Myeloid Leukemia Cell Line.” *Nature Genetics* 41 (5): 553–62. <https://doi.org/10.1038/ng.375>.
- Frith, Martin C., Yutao Fu, Liqun Yu, Jiang-Fan Chen, Ulla Hansen, and Zhiping Weng. 2004. “Detection of Functional DNA Motifs via Statistical Over-Representation.” *Nucleic Acids Research* 32 (4): 1372–81. <https://doi.org/10.1093/nar/gkh299>.
- Haan, Willeke de, Cristina Øie, Mohammed Benkheil, Wouter Dheedene, Stefan Vinckier, Giulia Coppiello, Xabier López Aranguren, et al. 2020. “Unraveling the Transcriptional Determinants of Liver Sinusoidal Endothelial Cell Specialization.” *American Journal of Physiology - Gastrointestinal and Liver Physiology* 318 (4): G803–15. <https://doi.org/10.1152/ajpgi.00215.2019>.
- Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. “Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities.” *Molecular Cell* 38 (4): 576–89. <https://doi.org/10.1016/j.molcel.2010.05.004>.
- Jacob, François, and Jacques Monod. 1961. “Genetic Regulatory Mechanisms in the Synthesis of Proteins.” *Journal of Molecular Biology* 3 (3): 318–56. [https://doi.org/10.1016/S0022-2836\(61\)80072-7](https://doi.org/10.1016/S0022-2836(61)80072-7).
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. “Unsupervised Learning.” In *An Introduction to Statistical Learning: With Applications in R*, edited by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 497–552. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-1418-1_12.
- Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. “Chromatin Accessibility and the Regulatory Epigenome.” *Nature Reviews Genetics* 20 (4):

- 207–20. <https://doi.org/10.1038/s41576-018-0089-8>.
- Latchman, David S. 1997. “Transcription Factors: An Overview.” *The International Journal of Biochemistry & Cell Biology* 29 (12): 1305–12. [https://doi.org/10.1016/S1357-2725\(97\)00085-X](https://doi.org/10.1016/S1357-2725(97)00085-X).
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” *Nature* 521 (7553): 436–44. <https://doi.org/10.1038/nature14539>.
- Löffler-Wirth, Henry, Martin Kalcher, and Hans Binder. 2015. “OposSOM: R-Package for High-Dimensional Portraying of Genome-Wide Expression Landscapes on Bioconductor.” *Bioinformatics (Oxford, England)* 31 (19): 3225–27. <https://doi.org/10.1093/bioinformatics/btv342>.
- Lyons, I., L. M. Parsons, L. Hartley, R. Li, J. E. Andrews, L. Robb, and R. P. Harvey. 1995. “Myogenic and Morphogenetic Defects in the Heart Tubes of Murine Embryos Lacking the Homeo Box Gene Nkx2-5.” *Genes & Development* 9 (13): 1654–66. <https://doi.org/10.1101/gad.9.13.1654>.
- Magnusson, Rasmus, Jesper N. Tegnér, and Mika Gustafsson. 2022. “Deep Neural Network Prediction of Genome-Wide Transcriptome Signatures – beyond the Black-Box.” *Npj Systems Biology and Applications* 8 (1): 1–8. <https://doi.org/10.1038/s41540-022-00218-9>.
- Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.” *Science (New York, N.Y.)* 337 (6099): 1190–95. <https://doi.org/10.1126/science.1222794>.
- McLeay, Robert C., Tom Lesluyes, Gabriel Cuellar Partida, and Timothy L. Bailey. 2012. “Genome-Wide in Silico Prediction of Gene Expression.” *Bioinformatics* 28 (21): 2789–96. <https://doi.org/10.1093/bioinformatics/bts529>.
- Migdał, Maciej, Takahiro Arakawa, Satoshi Takizawa, Masaaki Furuno, Harukazu Suzuki, Erik Arner, Cecilia Lanny Winata, and Bogumił Kaczkowski. 2023. “Xcore: An R Package for Inference of Gene Expression Regulators.” *BMC Bioinformatics* 24 (1): 14. <https://doi.org/10.1186/s12859-022-05084-0>.
- Migdał, Maciej, Eugeniusz Tralle, Karim Abu Nahia, Łukasz Bugajski, Katarzyna Zofia Kędzińska, Filip Garbicz, Katarzyna Piwocka, Cecilia Lanny Winata, and Michał Pawlak. 2021. “Multi-Omics Analyses of Early Liver Injury Reveals Cell-Type-Specific Transcriptional and Epigenomic Shift.” *BMC Genomics* 22 (1): 904. <https://doi.org/10.1186/s12864-021-08173-1>.
- Mosteller, Frederick, and R. A. Fisher. 1948. “Questions and Answers.” *The American*

- Statistician* 2 (5): 30–31. <https://doi.org/10.2307/2681650>.
- Natarajan, Anirudh, Galip Gürkan Yardımcı, Nathan C. Sheffield, Gregory E. Crawford, and Uwe Ohler. 2012. “Predicting Cell-Type-Specific Gene Expression from Regions of Open Chromatin.” *Genome Research* 22 (9): 1711–22. <https://doi.org/10.1101/gr.135129.111>.
- Nitta, Kazuhiro R, Arttu Jolma, Yimeng Yin, Ekaterina Morgunova, Teemu Kivioja, Junaid Akhtar, Korneel Hens, et al. 2015. “Conservation of Transcription Factor Binding Specificities across 600 Million Years of Bilateria Evolution.” Edited by Bing Ren. *ELife* 4 (March): e04837. <https://doi.org/10.7554/eLife.04837>.
- Ouyang, Zhengqing, Qing Zhou, and Wing Hung Wong. 2009. “ChIP-Seq of Transcription Factors Predicts Absolute and Differential Gene Expression in Embryonic Stem Cells.” *Proceedings of the National Academy of Sciences* 106 (51): 21521–26. <https://doi.org/10.1073/pnas.0904863106>.
- Pawlak, Michal, Katarzyna Z. Kedzierska, Maciej Migdal, Karim Abu Nahia, Jordan A. Ramilowski, Lukasz Bugajski, Kosuke Hashimoto, et al. 2019. “Dynamics of Cardiomyocyte Transcriptome and Chromatin Landscape Demarcates Key Events of Heart Development.” *Genome Research* 29 (3): 506–19. <https://doi.org/10.1101/gr.244491.118>.
- Pennacchio, Len A., Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. 2013. “Enhancers: Five Essential Questions.” *Nature Reviews. Genetics* 14 (4): 288–95. <https://doi.org/10.1038/nrg3458>.
- Reiter, Jeremy F., Jonathan Alexander, Adam Rodaway, Deborah Yelon, Roger Patient, Nigel Holder, and Didier Y. R. Stainier. 1999. “Gata5 Is Required for the Development of the Heart and Endoderm in Zebrafish.” *Genes & Development* 13 (22): 2983–95.
- Rzeszowska-Wolny, Joanna, and Roman Jaksik. 2010. “Position Weight Matrix Model as a Tool for the Study of Regulatory Elements Distribution across the DNA Sequence.” *Archives of Control Sciences* 20 (LVI): 491–501.
- Sanda, Takaomi, Lee N. Lawton, M. Inmaculada Barrasa, Zi Peng Fan, Holger Kohlhammer, Alejandro Gutierrez, Wenxue Ma, et al. 2012. “Core Transcriptional Regulatory Circuit Controlled by the TAL1 Complex in Human T Cell Acute Lymphoblastic Leukemia.” *Cancer Cell* 22 (2): 209–21. <https://doi.org/10.1016/j.ccr.2012.06.007>.
- Schmidt, Florian, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K. Polansky, Peter Ebert, et al. 2017. “Combining Transcription

- Factor Binding Affinities with Open-Chromatin Data for Accurate Gene Expression Prediction.” *Nucleic Acids Research* 45 (1): 54–66. <https://doi.org/10.1093/nar/gkw1061>.
- Stryer, Lubert, Jeremy M. Berg, Tymoczko John L., and Gregory J. Gatto. 2018. *Biochemia*. Wydawnictwo Naukowe PWN.
- Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” *Cell* 126 (4): 663–76. <https://doi.org/10.1016/j.cell.2006.07.024>.
- Tompa, Martin, Nan Li, Timothy L. Bailey, George M. Church, Bart De Moor, Eleazar Eskin, Alexander V. Favorov, et al. 2005. “Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites.” *Nature Biotechnology* 23 (1): 137–44. <https://doi.org/10.1038/nbt1053>.
- Yin, Wencheng, Luis Mendoza, Jimena Monzon-Sandoval, Araxi O. Urrutia, and Humberto Gutierrez. 2021. “Emergence of Co-Expression in Gene Regulatory Networks.” *PLOS ONE* 16 (4): e0247671. <https://doi.org/10.1371/journal.pone.0247671>.
- Zaret, Kenneth S., and Jason S. Carroll. 2011. “Pioneer Transcription Factors: Establishing Competence for Gene Expression.” *Genes & Development* 25 (21): 2227–41. <https://doi.org/10.1101/gad.176826.111>.

Assumptions and aim of the work

The research presented in this doctoral dissertation was based on the following assumptions:

- transcriptional regulation through combinatorial TF binding is the key mechanism underlying the specification of cell identity,
- co-expressed genes are likely to share their underlying regulatory grammar in the form of a combination of active TF binding events,
- chromatin accessibility is the key determinant of TF binding and motif-based predictions of TFBS in accessible chromatin regions can serve as an approximation to TF binding data,
- TF binding specificity is evolutionary conserved between mammals and zebrafish.

The research presented in this doctoral dissertation aimed to:

- characterize the cardiomyocytes' transcriptome and epigenome landscape at early stages of heart development using zebrafish as a model organism,
- identify gene regulatory network governing cardiomyocytes' gene expression program at early stages of heart development in zebrafish,
- characterize the transcriptomic and epigenomic response to hepatotoxic liver injury at the cell type level in endothelial cells, hepatocytes and hepatic stellate cells *in vivo*,
- develop bioinformatic pipeline for ATAC-seq data processing and open chromatin regions identification,
- develop bioinformatic tools for gene expression modeling and transcription factor activity prediction in a complex biological processes.

Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development

Michal Pawlak,¹ Katarzyna Z. Kedzierska,¹ Maciej Migdal,¹ Karim Abu Nahia,¹ Jordan A. Ramilowski,² Lukasz Bugajski,³ Kosuke Hashimoto,² Aleksandra Marconi,¹ Katarzyna Piwocka,³ Piero Carninci,² and Cecilia L. Winata^{1,4}

¹International Institute of Molecular and Cell Biology in Warsaw, Laboratory of Zebrafish Developmental Genomics, 02-109 Warsaw, Poland; ²RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045 Japan; ³Nencki Institute of Experimental Biology, Laboratory of Cytometry, 02-093 Warsaw, Poland; ⁴Max Planck Institute for Heart and Lung Research, 61231 Bad Nauheim, Germany

Organogenesis involves dynamic regulation of gene transcription and complex multipathway interactions. Despite our knowledge of key factors regulating various steps of heart morphogenesis, considerable challenges in understanding its mechanism still exist because little is known about their downstream targets and interactive regulatory network. To better understand transcriptional regulatory mechanism driving heart development and the consequences of its disruption *in vivo*, we performed time-series analyses of the transcriptome and genome-wide chromatin accessibility in isolated cardiomyocytes (CMs) from wild-type zebrafish embryos at developmental stages corresponding to heart tube morphogenesis, looping, and maturation. We identified genetic regulatory modules driving crucial events of heart development that contained key cardiac TFs and are associated with open chromatin regions enriched for DNA sequence motifs belonging to the family of the corresponding TFs. Loss of function of cardiac TFs *Gata5*, *Tbx5a*, and *Hand2* affected the cardiac regulatory networks and caused global changes in chromatin accessibility profile, indicating their role in heart development. Among regions with differential chromatin accessibility in mutants were highly conserved noncoding elements that represent putative enhancers driving heart development. The most prominent gene expression changes, which correlated with chromatin accessibility modifications within their proximal promoter regions, occurred between heart tube morphogenesis and looping, and were associated with metabolic shift and hematopoietic/cardiac fate switch during CM maturation. Our results revealed the dynamic regulatory landscape throughout heart development and identified interactive molecular networks driving key events of heart morphogenesis.

[Supplemental material is available for this article.]

The myocardium makes up most of the heart tissues and is mainly responsible for its function. Upon completion of gastrulation, heart muscle cells or cardiomyocytes (CMs) are specified from a pool of mesodermal progenitors at the anterior portion of the embryonic lateral plate mesoderm (Stainier et al. 1993; Stainier and Fishman 1994; Kelly et al. 2014). These progenitors migrate to the midline and form the primitive heart tube (Stainier et al. 1993), which subsequently expands through cell division and addition of more cells from the progenitor pool (Knight and Yelon 2016). Looping of the heart tube gives rise to the atria and ventricles. Although the vertebrate heart can have between two and four chambers, the stepwise morphogenesis of progenitors specification, migration, tube formation, and looping, are highly conserved between species (Jensen et al. 2013).

CMs are specified early during embryogenesis and undergo proliferation, migration, and differentiation, which collectively give rise to a fully formed and functioning heart. Crucial to regulating each step of heart morphogenesis are cardiac transcription factors (TFs) *NKX2-5*, *GATA5*, *TBX5*, and *HAND2* (Nemer 2008).

These TFs are known to play a role in establishing CM identity, regulating the formation and looping of the heart tube and specification of atrial and ventricular CMs. Members of the GATA family of TFs (*GATA4/5/6*) are responsible for the earliest step of cardiac progenitor specification (Reiter et al. 1999; Singh et al. 2010; Lou et al. 2011; Turbendian et al. 2013). They activate the expression of *Nkx2-5* (Lien et al. 1999), which is responsible for initiating the expression of many cardiac genes (Targoff et al. 2008). *Hand2*, another TF expressed in CM progenitors, is responsible for proliferation of ventricular progenitors (Yelon et al. 2000) and regulating the expression of *Tbx5*, which is necessary for atrial specification (Liberatore et al. 2000; Bruneau et al. 2001).

Despite established knowledge, little is known about the molecular mechanism and downstream targets of cardiac TFs. Transcription is modulated by *cis*-regulatory elements in noncoding regions of the genome, which serve as binding sites for TFs (Shlyueva et al. 2014). Although these regulatory elements equally contribute to the molecular mechanism controlling development, there is still a lack of systematic resources and understanding of their roles in heart development. Moreover, cardiac TFs have been shown to interact with chromatin-modifying factors, and

Corresponding author: cwinata@iimcb.gov.pl

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.244491.118>. Freely available online through the *Genome Research* Open Access option.

© 2019 Pawlak et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

the loss of function of several histone-modifying enzymes has been found to affect various aspects of cardiac development (Miller et al. 2008; Nimura et al. 2009; Takeuchi et al. 2011). Therefore, the chromatin landscape is another factor that needs to be considered when studying heart development.

The zebrafish is an ideal model organism to study heart development because it allows for accessing developing embryos immediately after fertilization, and it can survive without a functioning heart up to late developmental stages (Stainier 2001). To elucidate the dynamics of the transcriptional regulatory landscape during heart development, we isolated CMs directly from the developing wild-type zebrafish heart at three key stages of morphogenesis: linear heart tube formation (24 h post-fertilization [hpf]), chamber formation and differentiation (48 hpf), and heart maturation (72 hpf). Similarly, we isolated CMs from cardiac TF mutants of *gata5*, *tbx5a*, and *hand2* at 72 hpf. We combined transcriptome profiling (RNA-seq) with an assay for chromatin accessibility (ATAC-seq) (Buenrostro et al. 2013) to capture the dynamics of regulatory landscape throughout the progression of heart morphogenesis in vivo therefore unravelling the gene regulatory network driving key processes of heart development.

Results

CM transcriptome reveals strong dynamics at early stages of heart morphogenesis

Two of the earliest markers of cardiac lineage are *nkx2.5* and myosin light chain 7 (*myl7*), which are expressed in cardiac precursor cells in the anterior lateral plate mesoderm (George et al. 2015) and in differentiated myocardial cells (Chen et al. 2008), respectively. To study gene regulatory networks underlying zebrafish heart development, we isolated CMs from zebrafish transgenic lines Tg(*nkx2.5*:GFP) (Witzel et al. 2012) and Tg(*myl7*:EGFP) (D'Amico et al. 2007) using fluorescence-activated cell sorting (FACS) (Fig. 1A). Cells were collected at 24, 48, and 72 hpf (Fig. 1B). Because of its earlier onset of CM-specific GFP expression, Tg(*nkx2.5*:GFP) was used to sort CM at 24 hpf, whereas Tg(*myl7*:EGFP) was used for subsequent stages (48 hpf and 72 hpf) (Houk and Yelon 2016). The average fraction of FACS-yielded GFP+ events obtained were between 1.37% and 2.56% of total singlet events (Supplemental Fig. 1A). To monitor the purity of FACS and establish the identity of the isolated cells, we measured mRNA levels of *nkx2.5*, *myl7*, and GFP in both GFP+ and GFP- cells. The expression of the CM markers and GFP were significantly enriched (P -value ≤ 0.05) in GFP+ as compared to GFP- fraction (Supplemental Fig. 1B). In contrast, mRNA levels of *neurogenin1* (*ngn1*), a neuronal-specific gene, were higher in GFP- cells. In line with that, RNA-seq expression of *nkx2.5*, *myl7*, and *myh6* was significantly enriched (adjusted P -value ≤ 0.05) in GFP+ as compared to GFP- cells, whereas expression of non-CM markers such as skeletal muscle (*myog*), pancreas (*ins*), pharyngeal arch (*frem2a*), retina (*arr3b*, *otx5*), skin (*tp63*, *col16a1*), neural system (*neurog1*, *zic3*, *otx1*), and eye (*pou4f2*) were higher in GFP- (Supplemental Fig. 2). Gene Ontology (GO) enrichment analysis of differentially expressed genes between GFP+ and GFP- across all three stages of heart development revealed the overrepresentation of CM-specific biological processes such as cell migration, cardiac development, and heart function (Fig. 1C; Supplemental Table 1). Among 50 genes with the highest average expression across all developmental stages, 35 have specific functions in CM according to the ZFIN database (<https://zfin.org>) and eight are associated with human cardiac diseases including car-

diomyopathy (*ttn.1*, *mybpc3*, *ttn.2*, *acta1b*, *actn2b*), atrial septal defects (*actc1a*, *myh6*), and Laing distal myopathy (*vmhc*) (Fig. 1D) according to the Online Mendelian Inheritance in Man database (<https://www.omim.org/>).

To determine the dynamics of CM transcriptome throughout development, we applied principal component analysis (PCA) and clustering based on Euclidean distance. Both analyses revealed strong dissimilarity in transcriptome profiles between CM at 24 hpf and later stages of heart development as compared to those between 48 and 72 hpf (Fig. 1E,F). This suggests that major gene expression profile changes of developing CMs occur between 24 and 48 hpf and correspond to heart tube formation and looping.

Taken together, transcriptome analyses identified CM-specific gene expression signatures among highly abundant transcripts and revealed the dynamic nature of gene expression profiles during heart morphogenesis.

Chromatin accessibility is correlated with CM gene expression levels during heart development

To characterize chromatin dynamics throughout heart development, we used an assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) and profiled chromatin accessibility at three developmental stages matching our transcriptome analyses (Buenrostro et al. 2013). To identify genome-wide nucleosome free regions (NFR), ATAC-seq read fragments were partitioned into four populations (Fig. 2A) based on exponential function for fragment distribution pattern at insert sizes below one nucleosome (123 bp) and Gaussian distributions for 1, 2, and 3 nucleosomes as previously described (Buenrostro et al. 2013). PCA analysis (Fig. 2B) and clustering based on Euclidean distances between ATAC-seq samples based on their NFR profiles (Fig. 2C) revealed that biological replicas clustered together, whereas the largest changes in chromatin accessibility were observed between 24 and 48 hpf stages, in agreement with observed transcriptome changes of CMs during heart development. We observed a large number of NFRs common to all stages (16,055), as well as those specific to a single developmental stage. The most stage-specific NFRs were found at 24 hpf (22,656) (Fig. 2D). The highest fraction of NFRs was localized within promoter (within ± 3 kb of transcription start site, TSS; $\sim 30\%$ of total NFRs), followed by intergenic ($\sim 25\%$) and intronic (20%) regions (Fig. 2E; Supplemental Table 2). These ratios remained comparable across all three developmental stages. NFR consensus heatmaps within transcription start site (TSS) proximal promoter regions (± 3 kb) (Fig. 2F) compared to distal promoter regions (more than ± 3 kb of TSS) (Fig. 2G), as well as ATAC-seq read density over the gene bodies of 1000 genes most highly expressed in CMs at all three stages of heart development (Fig. 2H), revealed the enrichment of NFRs around TSS regions. We further observed chromatin accessibility reflected by the presence of NFR in gene promoter regions was significantly correlated with the expression levels of the corresponding genes to which the promoter belonged to (Spearman's rho 0.46–0.48) at each stage of heart development (Fig. 2I). Our observations revealed a strong link between chromatin accessibility of promoter regions and gene expression levels.

Coexpression network analysis identifies CM regulatory modules

To better understand the relationship and functionality of cardiac genes involved in heart morphogenesis, we identified gene regulatory networks in an unsupervised and unbiased manner using the weighted gene correlation network analysis (WGCNA) based on

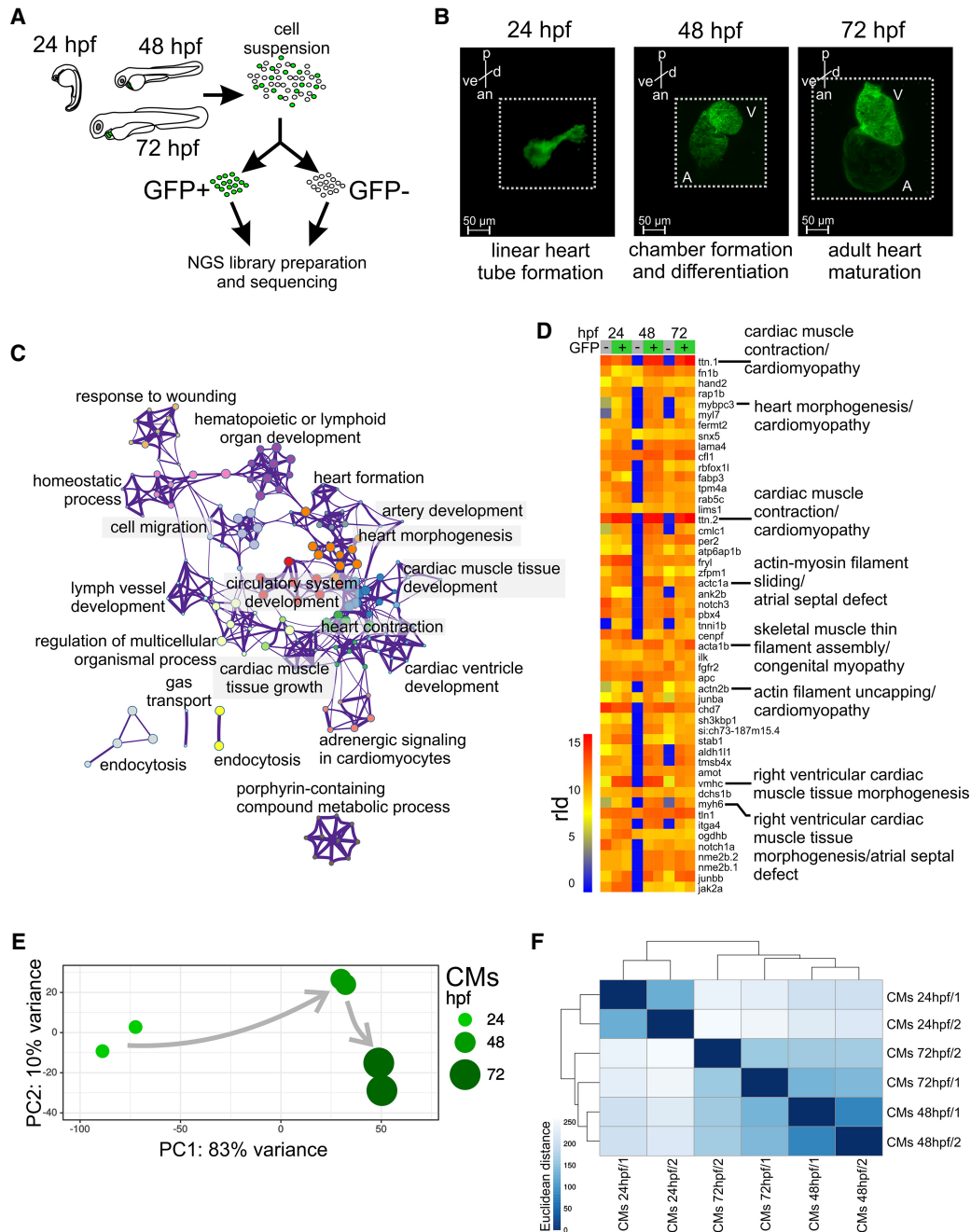


Figure 1. CM transcriptome landscape during heart development. (A) Schematics of experimental design. (B) Light sheet fluorescence microscope (LSFM) images of GFP-labeled CMs of developing zebrafish heart: (p) posterior; (an) anterior; (v) ventral; (d) dorsal. The dotted line indicates exact area of the LSFM image. (C) Network of 20 top-score GO clusters enriched in genes commonly up-regulated in GFP+ across heart development. Size nodes refer to the number of genes contributing to the same GO and nodes that share the same cluster ID are close to each other, adjusted *P*-value ≤ 0.05 . (D) Heatmap of top 50 highly expressed genes between 24 and 72 hpf based on normalized expression value (regularized log [rld]). (E) Graphical representation of PCA of CM RNA-seq data. (F) Heatmap and clustering of RNA-seq sample-to-sample Euclidean distances.

RNA-seq expression profiles (Langfelder and Horvath 2008). Hierarchical clustering of the similarity/dissimilarity matrix across the entire set of transcriptome samples distinguished 37 gene modules (Fig. 3A; Supplemental Table 3), five of which were enriched in functional terms related to cardiovascular system development and function (Fig. 3B; Supplemental Table 4): turquoise (4085 genes), brown (2156 genes), green (1166 genes), salmon (756 genes), and sienna3 (75 genes). We refer to these modules

as “cardiac modules.” Functional terms enriched in these cardiac modules included “embryonic heart tube development” (brown, green, and sienna3), “cardioblast differentiation” (green), “heart valve development” (salmon), “heart process” and “heart formation” (turquoise). The relatively small sienna3 module was enriched in GO terms associated with multiple cardiac developmental processes including “heart tube development,” “cardioblast migration,” and “heart rudiment development.”

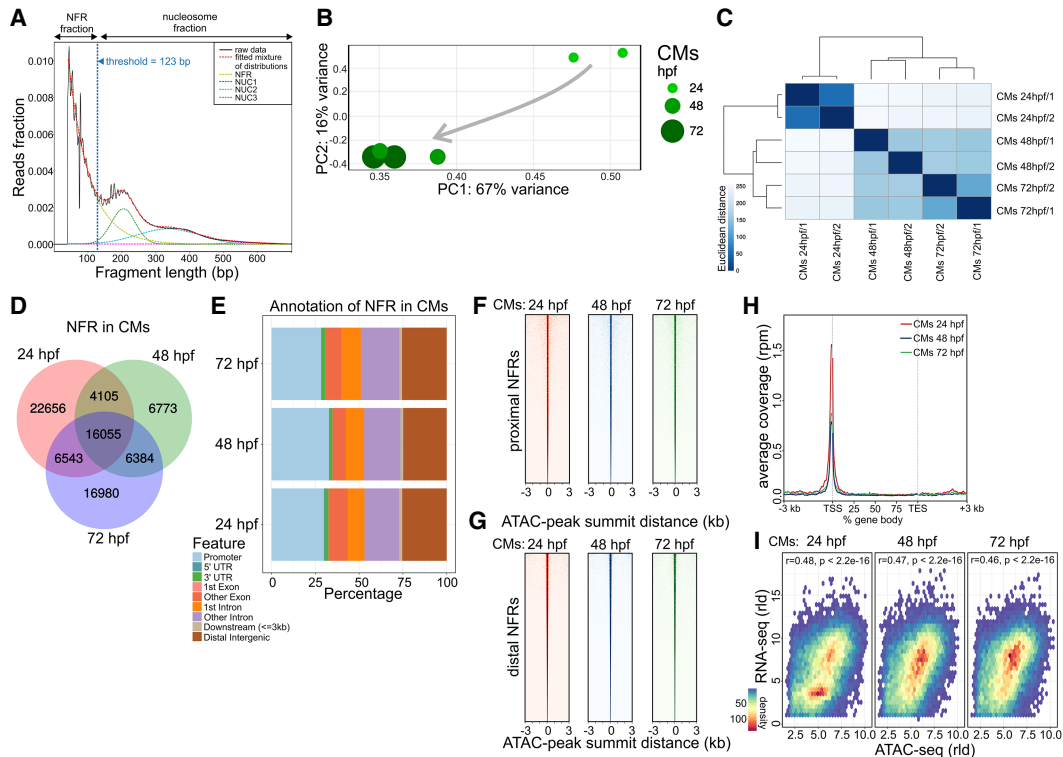


Figure 2. Cross-talk between transcriptome and chromatin accessibility profile across stages of cardiac development. (A) ATAC-seq read distribution and characterization of NFR fractions. (B) PCA of NFR chromatin accessibility during heart development. (C) Euclidean distances between chromatin accessibility within NFR. (D) Comparison of NFR presence and overlap across stages of heart development. (E) Genomic annotation of CM NFR consensus at different stages of heart development. (F) CM NFR consensus coverage heatmap of TSS proximal (± 3 kb of TSS) regions centered on ATAC-seq peak summits. (G) CM NFR consensus coverage heatmap of TSS distal (more than ± 3 kb of TSS) regions centered on ATAC-seq peak summits. (H) Metaplot of ATAC-seq read density over the gene bodies of the 1000 genes most highly expressed in CMs at each developmental stage. (TES) transcription end site. (I) Spearman's correlation of normalized log (rd) RNA-seq gene expression and ATAC-seq chromatin accessibility in corresponding NFR regions (± 3 kb of TSS).

To reveal potential driver genes with regulatory roles in each cardiac module, we searched for TFs and calculated their connectivity to other genes within a given module (normalized kDiff), as well as how their expression is affected by CM phenotypic traits (CM correlation) (Fig. 3C). Most of the cardiac modules contained TFs previously implicated in heart development, such as *gata1* (brown), *tbx5a*, *sox10* (turquoise), *hand2*, *smad7* (green), as well as *gata5*, *nkx2.5*, and *tbx20* (sienna3) (Ahn et al. 2000; Montero et al. 2002; Holtzinger and Evans 2007; Targoff et al. 2008; Moskowitz et al. 2011). Each of the modules exhibited different expression profile dynamics (eigengene expression) across three developmental stages, in GFP+ and GFP- fractions, further called CM+ and CM-, respectively (Fig. 3D). Two broad patterns of eigengene expression could be observed: modules with decreasing cardiac gene expression during heart development (brown and green) and modules in which expression increases between 24 and 48 hpf and then decreases between 48 and 72 hpf (salmon, sienna3, and turquoise). In addition, CM+ eigengene expression in the sienna3 module was consistently higher than in CM- samples at all stages of development, further suggesting the specificity of this module to CM.

The presence of key cardiac TFs in each module prompted us to identify specific functional patterns related to cardiovascular development. The sienna3 module, which contained cardiac TFs *nkx2.5*, *gata5*, *gata6*, and *tbx20*, also contained many other genes implicated in various aspects of heart morphogenesis, including CM migration and differentiation, and heart looping including *popdc2*, *apobec2a*, and *tdgf1* (Xu et al. 1999; Etard et al. 2010; Wang et al.

2011; Kirchmaier et al. 2012; Sakabe et al. 2012). Additionally, the module also contained many genes involved in cell adhesion and structural constituents of the heart muscle, which were previously implicated in cardiomyopathy. These included *act1a*, *myl7*, *myh7ba*, *myh7bb*, *vmhc*, and *ttn.2* (Olson et al. 1998; Xu et al. 2002; Shih et al. 2015). In support of this network, *Popdc2* and *Gata6* were previously shown to be a direct transcriptional target of NKX2-5 in mouse embryonic heart (Molkentin et al. 2000; Dupays et al. 2015). Additional evidence supports the cardiac-specific transcriptional activation of *nkx2.5* by Gata TFs (Lien et al. 1999).

Genes of the Wnt, Notch, TGFB, and FGF pathways were highly represented in all modules except sienna3, which consisted of mostly specialized CM genes. In particular, genes of both canonical and noncanonical Wnt signaling pathways involved in cardiogenesis (Ueno et al. 2007; Piven and Winata 2017) were almost exclusively distributed between the green and salmon modules. Finally, another cardiac TF, Hand2, was present in the green module, suggesting that it might control these pathways. Altogether, we identified gene modules exhibiting unique expression patterns throughout heart development, representing potential regulatory networks underlying various processes of heart development.

Integrative analysis of RNA-seq and ATAC-seq identifies regulatory networks of CM maturation

To further explore the relationship between chromatin state and transcriptional regulation of heart development, we integrated

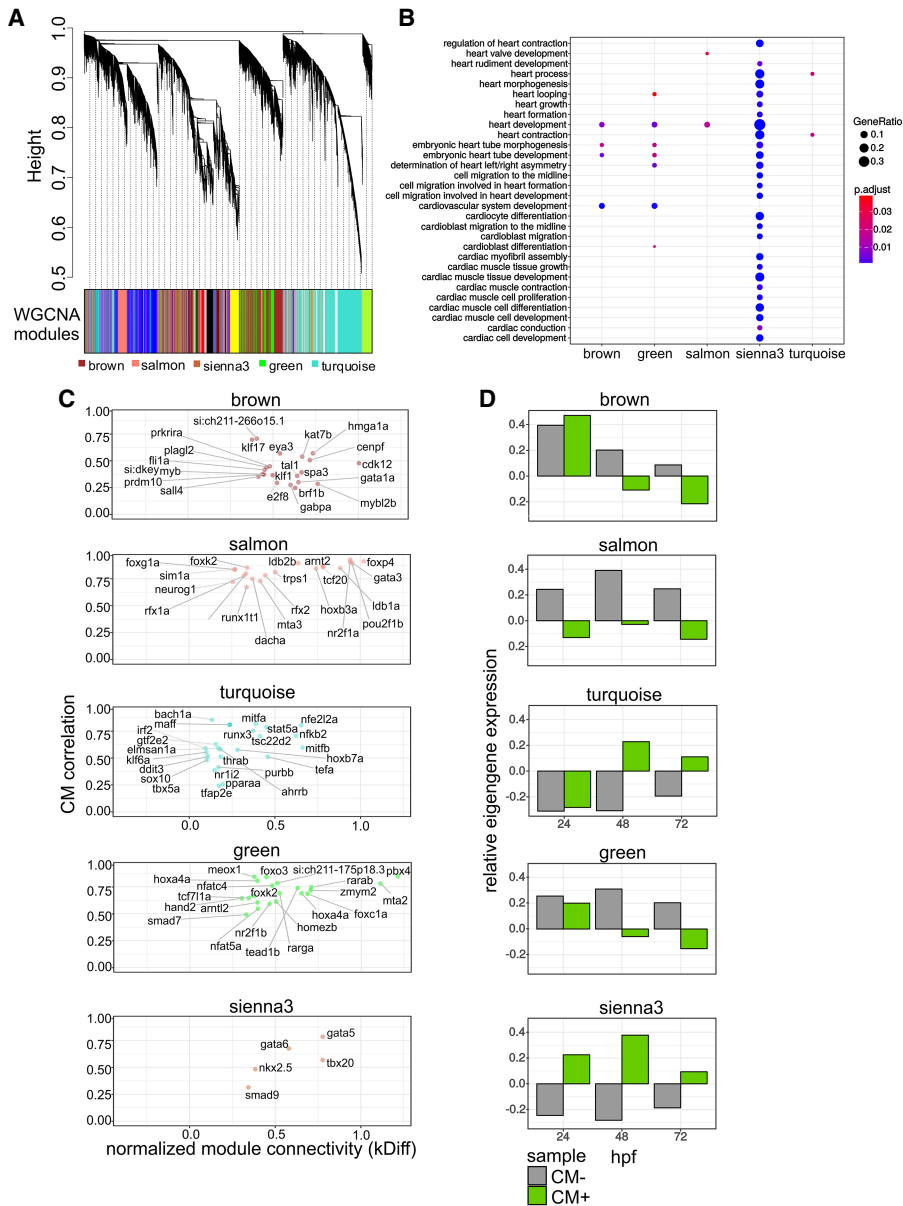


Figure 3. Cardiac coexpression regulatory networks. (A) Hierarchical clustering of gene expression similarity/dissimilarity matrix. (B) Cardiovascular-related GO enrichment in five cardiac modules. (C) Module gene connectivity plot of selected TFs. Twenty TFs with the highest normalized kDiff are shown. (D) Cardiac module eigengene expression during heart development.

coexpression networks generated from RNA-seq with accessible chromatin regions identified by ATAC-seq. We examined NFRs localized within ± 3 kb of the TSS of genes assigned to the same module for the presence of TF binding motifs (Table 1; Supplemental Fig. 3). NFRs associated with genes within the module sienna3 (which contained *gata5/6*, *nkx2.5*, and *tbx20*) were also enriched in motifs belonging to these families of TFs (Gata family [Gata1/2/3/4/6], Nkx family [Nkx2.2, Nkx2.5], Smad3 and T-box family [Tbr1]), whereas the salmon module containing the *sox3* gene showed overrepresentation of Sox3 motif. Similarly, in the turquoise and green modules (containing *tbx5*, *hand2*, and *smad7*), we found a wide range of significantly enriched (P -value ≤ 0.05) TF motifs including Tbx5 and Smad2/4, respectively. The presence

of the TFs together with the enrichment of their respective recognition motifs suggests their regulatory role within each module. Moreover, we observed an overrepresentation of motifs for TFs with profound roles in heart development, such as Sox10 in the salmon module and Tgif1/2 in both the sienna3 and turquoise modules (Montero et al. 2002; Powers et al. 2010), although their corresponding TFs were not present in the matching modules.

To understand the relationship between chromatin accessibility and gene expression and to link TFs to their effector genes, we identified genes that were dynamically regulated and associated with regions of differential chromatin accessibility (within ± 3 kb of TSS) throughout heart development (Fig. 4A). We compared normalized changes of gene expression to those of the corresponding NFRs between 24 and 48 hpf as well as between 48 and 72 hpf (Fig. 4B; Supplemental Tables 1, 5). We observed strong up-regulation of a large number of genes within the turquoise and salmon modules and down-regulation of genes in the brown module and for most genes belonging to the green module. This was generally consistent with the direction of changes in chromatin accessibility, for example, *gpd2*, *sox10* in the turquoise module, *commd5* in the salmon, *tbx16l*, *pappa2* in the brown, and *tfr1a*, *aff2* in the green; yet we also observed genes with the opposite behavior, including *klf6a*, *irf2bp2a* in the turquoise module, *sema4ab* in the brown, and *serinc2* in the green. No significant changes in gene expression and NFR were observed between 48 and 72 hpf (Supplemental Tables 1, 5), suggesting that both gene expression and chromatin accessibility were more stable by heart chamber formation.

GO and pathway analysis of the turquoise module revealed genes involved in mitochondrial oxidation (*mdh2*, *gpd2*), carbohydrate metabolism (*rdh8a*), and ketone body metabolism (*bdh2*) (Fig. 4C,D; Supplemental Table 6). We identified *sox10*, *klf6a*, and *irf2bp2a*, which were previously linked to zebrafish heart morphogenesis (Hill et al. 2017), as hub genes linked to their effector genes containing corresponding binding motifs in NFRs within their proximal promoter regions. Because the vast majority of genes within the turquoise module exhibited significant increase in expression and chromatin accessibility within associated NFRs between 24 and 48 hpf, it suggests the presence of a metabolic switch that takes place in CM between those developmental stages. This agrees with previous reports showing that mitochondrial oxidative capacity and fatty acid oxidation potential increase along with CM maturation (Lopaschuk and Jaswal 2010).

Table 1. HOMER-identified TF motifs found in NFR of cardiac coexpression modules

Brown				Green					
Motif name	P-value	Target sequences with motif (of 4698)	Target sequences with motif (%)	Background sequences with motif (%)	Motif name	P-value	Target sequences with motif (of 4698)	Target sequences with motif (%)	Background sequences with motif (%)
Smad3	1×10^{-2}	1172	24.94	23.07	Smad4	1×10^{-4}	545	19.93	16.96
Bhlh	1×10^{-2}	904	19.24	17.83	Smad2	1×10^{-2}	495	18.11	15.85
Fli1	1×10^{-9}	786	16.73	13.44	Sox3	1×10^{-2}	420	15.36	13.41
Etv1	1×10^{-5}	729	15.51	12.97	Sox6	1×10^{-2}	407	14.89	12.87
Nfy	1×10^{-6}	702	14.94	12.41	Sox10	1×10^{-2}	387	14.16	12.41
Sox3	1×10^{-2}	700	14.90	13.63	Gata4	1×10^{-2}	284	10.39	8.84
Erg	1×10^{-3}	672	14.30	12.46	Gata6	1×10^{-2}	255	9.33	7.79
Gata3	1×10^{-3}	664	14.13	12.30	Sox2	1×10^{-2}	218	7.97	6.7
Ets1	1×10^{-7}	619	13.17	10.60	Sox4	1×10^{-3}	212	7.75	6.12
Ehf	1×10^{-2}	581	12.36	11.16	Gata2	1×10^{-3}	201	7.35	5.75

Salmon				Sienna3					
Motif name	P-value	Target Sequences with motif (of 4698)	Target sequences with motif (%)	Background sequences with motif (%)	Motif name	P-value	Target Sequences with motif (of 4698)	Target sequences with motif (%)	Background sequences with motif (%)
Sox3	1×10^{-8}	332	18.24%	13.21%	Tgif1	1×10^{-2}	77	45.03%	33.41%
Sox10	1×10^{-7}	307	16.87%	12.34%	Tgif2	1×10^{-2}	77	45.03%	34.57%
Neurog2	1×10^{-2}	299	16.43%	14.17%	Meis1	1×10^{-2}	46	26.90%	18.63%
Sox6	1×10^{-4}	293	16.10%	12.60%	Nkx2.5	1×10^{-2}	44	25.73%	17.61%
Atoh1	1×10^{-2}	215	11.81%	9.78%	Bapx1	1×10^{-2}	42	24.56%	16.10%
Sox15	1×10^{-6}	202	11.10%	7.65%	Nkx2.2	1×10^{-2}	41	23.98%	16.70%
Sox2	1×10^{-6}	178	9.78%	6.64%	Gata3	1×10^{-3}	38	22.22%	13.17%
Neurod1	1×10^{-2}	159	8.74%	7.17%	Mef2b	1×10^{-9}	33	19.30%	5.66%
Sox4	1×10^{-4}	157	8.63%	6.08%	Tbr1	1×10^{-2}	33	19.30%	12.02%
Maz	1×10^{-2}	141	7.75%	6.03%	Gata6	1×10^{-3}	29	16.96%	8.61%

Turquoise				
Motif name	P-value	Target sequences with motif (of 4698)	Target sequences with motif (%)	Background sequences with motif (%)
Scl	1×10^{-6}	3906	45.22	42.29
Tgif2	1×10^{-20}	3467	40.14	34.68
Tgif1	1×10^{-20}	3353	38.82	33.42
Nanog	1×10^{-4}	3311	38.34	35.88
Pitx1	1×10^{-8}	3055	35.37	31.99
Thrb	1×10^{-3}	2376	27.51	25.78
Tbx5	1×10^{-6}	2270	26.28	23.71
Nkx6.1	1×10^{-2}	2166	25.08	23.66
Ar	1×10^{-2}	2120	24.55	23.18
Smad3	1×10^{-2}	2076	24.04	22.37

HOMER-identified motifs with the highest prevalence in NFRs localized ± 3 kb around the TSSs of selected cardiac module genes are listed. *P*-value < 0.05. Known vertebrate TF motifs were used for analysis.

Conversely, most of the genes assigned to the brown module were down-regulated from 48 hpf onward along with the associated NFR chromatin accessibility (Fig. 4C). Pathway and GO analysis of the brown module (Supplemental Table 7) revealed the presence of genes implicated in embryonic hematopoiesis. Notably, we have identified a number of hub TFs including *myb* (*v-myb*) and *prdm1a*, *mybl2*, *tbx16l*, *e2f8*, *klf17* as well as their effector genes, such as *lmo2*, *tal1*, *alas2*, *slc4a1a* with profound roles in hematopoiesis (Fig. 4E; Gering et al. 2003; Paw et al. 2003; Chan et al. 2009; Soza-Ried et al. 2010; Kotkamp et al. 2014). Moreover, ATAC-seq analyses revealed the enrichment of Gata, Fli1, Ets, Erg, and Etv motifs (Table 1), which belong to the regulatory network underlying hematopoietic/vascular lineage specification (Gottgens et al. 2002; Pimanda et al. 2007; Loughran et al. 2008;

Kaneko et al. 2010). The brown module possibly represents the regulatory network leading to hematopoietic fate, whose suppression promotes the development of CM identity. Altogether, we identified regulatory networks leading to metabolic and cardiac/hematopoietic changes occurring in CMs during early heart morphogenesis (Supplemental Table 8), which are regulated at both gene expression and chromatin levels.

Disruption of cardiac TFs affects regulatory networks driving CM maturation

To further explore cardiac regulatory modules and validate their importance in normal heart development, we used zebrafish mutants deficient in cardiac-related TFs (Gata5, Hand2, and Tbx5a),

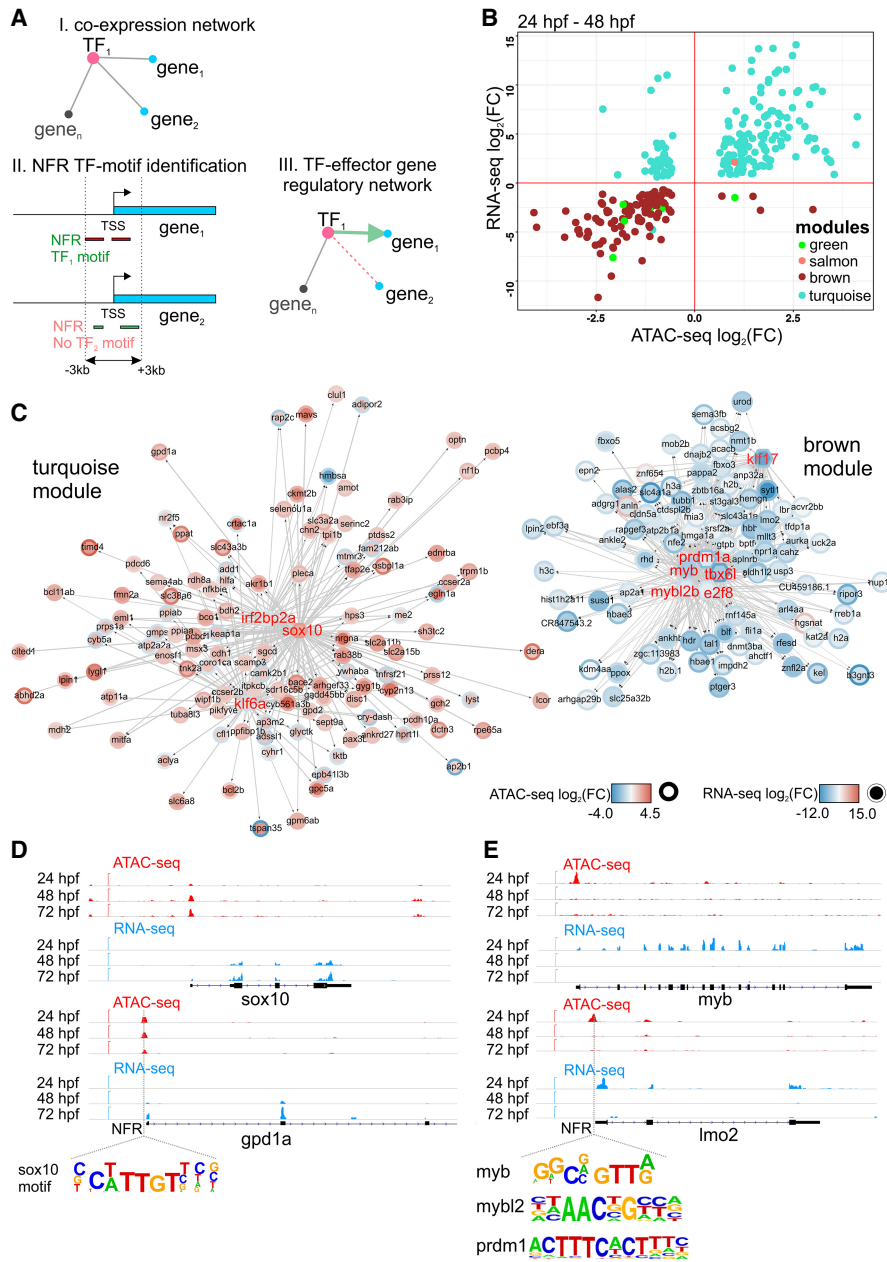


Figure 4. Dynamic regulatory networks of differentiating CMs. (A) Strategy used to establish gene-chromatin regulatory network. (B) Changes (log₂FC) of gene expression compared to those in chromatin accessibility of cardiac module genes during heart development. Only significant (FDR < 0.05) genes are shown. (C) Regulatory networks of heart development. Arrows indicate the direction of interaction. Colors and the intensity of the circle edges indicate changes of chromatin accessibility, whereas those inside the circle show expression changes. Only significant (adjusted *P*-value ≤ 0.05) genes are shown. Hub TFs are indicated in red font. (D) Visualization of ATAC-seq and RNA-seq read coverage of selected genomic regions related to the turquoise module. (E) Visualization of ATAC-seq and RNA-seq read coverage of selected genomic regions related to the brown module. Time points, NFRs, and TF binding motifs within NFRs are indicated.

the disruption of which was linked to impaired migration of the cardiac primordia to the embryonic midline, reduced number of myocardial precursors, and failure of heart looping, respectively (Reiter et al. 1999; Yelon et al. 2000; Garrity et al. 2002). RNA-seq and ATAC-seq were performed on CMs isolated from homozygous *gata5*^{tm236a/tm236a}, *tbx5a*^{m21/m21}, *hand2*^{s6/s6} mutant 72 hpf embry-

os in Tg(*myl7*:EGFP) genetic background. Homozygous mutant embryos were selected based on their phenotypes of *cardia bifida* (*gata5*^{tm236a/tm236a}, *hand2*^{s6/s6}) or *heart-string* (*tbx5a*^{m21/m21}) (Fig. 5A).

A number of genes were dysregulated (absolute[log₂FC] > 0, adjusted *P*-value ≤ 0.05) in response to disruption of *gata5* (287 down-regulated, 739 up-regulated), *hand2* (288 down-regulated, 618 up-regulated), and *tbx5a* (255 down-regulated, 584 up-regulated) (Fig. 5B; Supplemental Table 9). Only a small overlap was observed between genes down-regulated in the three mutants (14 genes including *vcamb*, *bmp3*, and *col18a1b*), whereas up-regulated genes showed larger overlap (307 genes, e.g., *trim46*, *map4k6*, *mtf1*) between all three mutants. GO enrichment analysis of all down-regulated genes revealed the presence of biological processes related to muscle development, muscle function, heart process, and sensory perception signaling; up-regulated genes were enriched in biological processes related to ion transport and inflammatory response (Supplemental Table 10).

Changes in chromatin accessibility of NFRs localized in proximal promoter regions (±3 kb of TSS) of mutants and wild-type embryos were generally less pronounced than changes in gene expression (Fig. 5B; Supplemental Table 9). Moreover, loss of different TFs affected the chromatin to a variable extent, the largest of which occurred in *gata5*^{tm236a/tm236a} mutants (335 differentially accessible NFRs associated with genes enriched in cardiac muscle development processes) (Fig. 5B; Supplemental Table 10). In *hand2*^{s6/s6} mutants, 53 NFRs were down-regulated. Lesser pronounced chromatin changes were identified in *tbx5a*^{m21/m21} mutant (17 NFRs). Seven down-regulated NFRs associated with *nkx1.21a*, *dmd*, *frzb*, *gpr4*, and *vap* were common between *gata5*^{tm236a/tm236a} and *hand2*^{s6/s6} mutants, whereas 246 up-regulated ones were localized in the proximity of *nr4a1*, *mycbp2*, *irf2bp2a*, *rpl3*. No differentially regulated proximal NFRs were shared between all three mutants.

We further explored which fraction of mutant down-regulated genes contributed to the cardiac regulatory modules identified in wild-type analyses. We found that 31% (91 genes), 24% (71 genes), and 31% (79 genes) of total down-regulated genes in *gata5*^{tm236a/tm236a}, *hand2*^{s6/s6}, and *tbx5a*^{m21/m21} mutants were present in cardiac modules, mainly in the brown and green modules (Fig. 5C). Among the 14 genes that were commonly down-regulated in all three mutants,

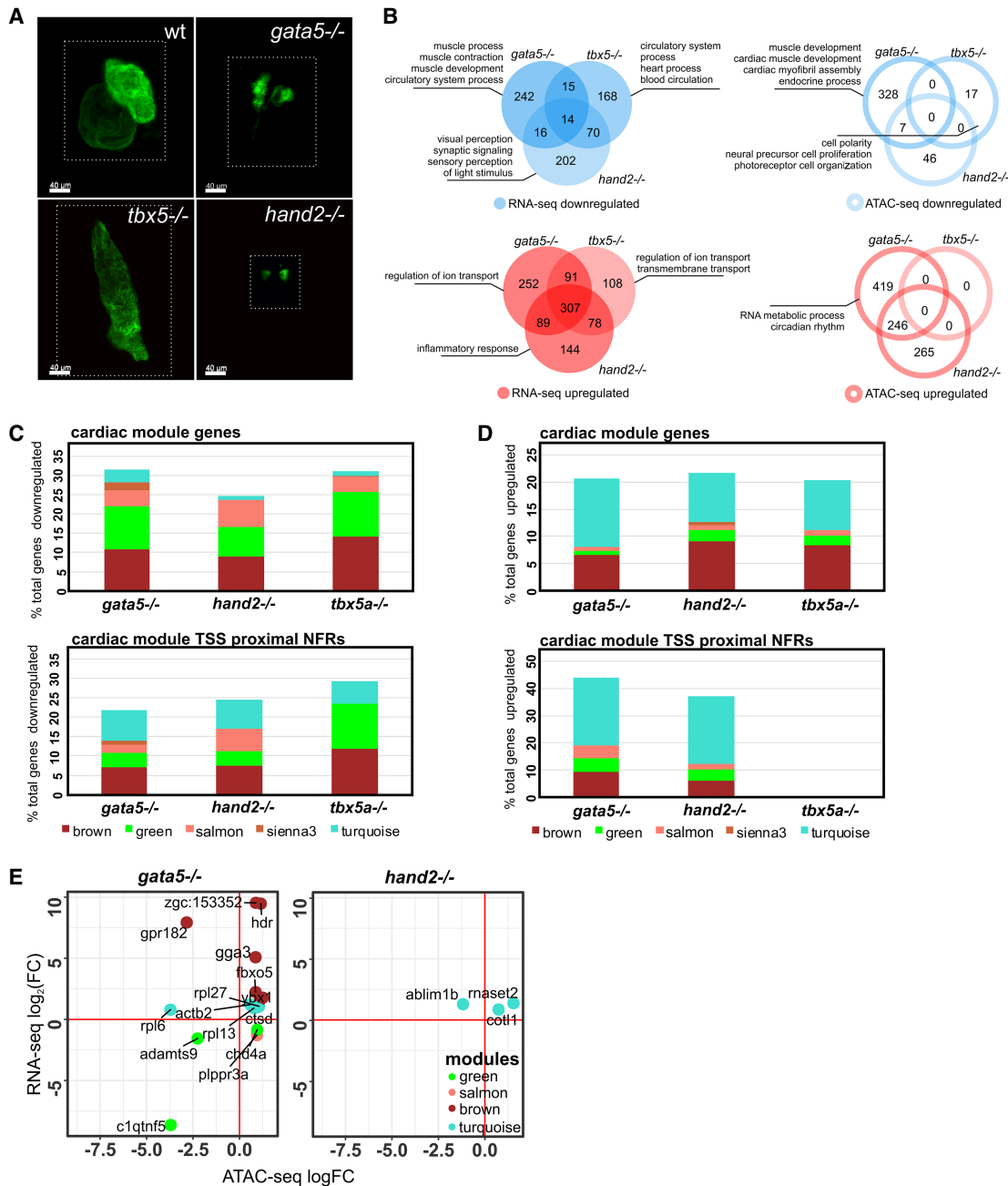


Figure 5. Loss-of-function mutations of cardiac TFs alters regulatory networks involved in heart development. (A) LSFM images of GFP-labeled CMs of wild-type and TF mutant zebrafish hearts at 72 hpf. The dotted line indicates exact area of the LSFM image. (B) Venn diagrams and GO enrichment analysis of TF mutant down-regulated (blue) and up-regulated (red) genes and chromatin accessibility of proximal promoter NFRs (± 3 kb of TSS), adjusted P -value ≤ 0.05 . (C) Percent distribution of cardiac module down-regulated genes/proximal NFR chromatin accessibility as compared to total number of TF mutants down-regulated genes/proximal NFR chromatin accessibility. (D) Percent distribution of cardiac module up-regulated genes/proximal NFR chromatin accessibility as compared to total number of TF mutants up-regulated genes/proximal NFR chromatin accessibility. (E) Cardiac module genes with differentially regulated expression and chromatin accessibility of proximal promoter NFRs (± 3 kb of TSS) in *gata5*, *hand2*, and *tbx5a* mutants.

four belonged to green (*nid1b*, *papss2b*, *vcanb*, *bmp3*) and two to salmon (*plppr3a*, *spn1b*) modules. Genes including *vcan*, *plppr3a*, and *Bmp* family were previously found to play a crucial role in heart morphogenesis and function (Marques and Yelon 2009; Kern et al. 2010; Chandra et al. 2018). Similarly, comparing chromatin accessibility data revealed that 21% (73 regions), 24% (13 regions), and 29% (five regions) of proximal NFRs that showed decreased acces-

sibility in *gata5*^{tm236a/tm236a}, *hand2*^{s6/s6}, and *tbx5a*^{m21/m21} mutants were located within the proximal promoters of genes belonging to cardiac modules (Fig. 5C). We also explored mutant up-regulated genes and proximal NFRs associated with cardiac modules (Fig. 5D). It showed that 20% (153 genes), 21% (134 genes), and 20% (119 genes) of total up-regulated genes in *gata5*^{tm236a/tm236a}, *hand2*^{s6/s6}, and *tbx5a*^{m21/m21} mutants were present in cardiac

modules, predominantly in the brown and turquoise modules. Consequently, the most prominent changes were observed for proximal NFRs in the brown and turquoise modules, and 43% (292 regions) and 37% (229 regions) of total up-regulated NFRs contributed to cardiac modules in *gata5*^{tm236a/tm236a} and *hand2*^{s6/s6}. No changes were observed in *tbx5a*^{m21/m21} mutants. The vast majority of either down-regulated or up-regulated cardiac module genes did not exhibit a similar regulation of NFR chromatin accessibility within their promoter regulatory regions (Supplemental Fig. 4). We observed that the decrease in proximal promoter NFRs was not correlated with the gene expression down-regulation, except for *c1qtnf5* and *adams9*, the latter being a vcan-degrading protease required for normal heart development and cardiac allostasis (Supplemental Fig. 5A,B; Kern et al. 2010). Similarly, only 10 genes including *hdr*, *gga3*, *fbxo5*, *rpl27*, *ybx1*, *actb2*, *cotl1*, *maset2* showed an increase in both gene expression and NFR chromatin accessibility (Fig. 5E). Only 15 genes showed changes both in expression level and chromatin accessibility (either increasing or decreasing) in *gata5* mutant and three genes in *hand2* mutant, whereas no such genes were found in *tbx5a* mutant.

Taken together, we identified genes that were responsive to loss of Gata5, Hand2, and Tbx5a functions, including those belonging to cardiac modules, therefore providing a strong validation of the cardiac regulatory networks controlling specific processes of heart development.

Evolutionarily conserved enhancers ensure proper heart development

Gene expression changes in all three mutants were, to a large extent, uncorrelated with changes in chromatin accessibility, at least in proximal promoter regulatory regions. This led us to question whether loss of Gata5, Hand2, and/or Tbx5a cardiac TFs may cause global chromatin changes at genomic sites other than proximal gene promoters, and whether the observed changes in gene expression could be attributed to distal regulatory elements. To this end, we have identified distal NFRs (more than ± 3 kb of TSS) and their differential accessibility between wild type at 72 hpf and the mutants. We identified 59, 14, and 33 down-regulated and 551, 321, and 2 regions up-regulated (adjusted *P*-value ≤ 0.05) in *gata5*^{tm236a/tm236a}, *hand2*^{s6/s6}, and *tbx5a*^{m21/m21} mutants, respectively (Fig. 6A). Among down-regulated regions, one was in common between *gata5*^{tm236a/tm236a} and *tbx5a*^{m21/m21} mutants (Fig. 6B). On the other hand, much stronger overlap was observed between *gata5*^{tm236a/tm236a} and *hand2*^{s6/s6} mutants for up-regulated regions (183 regions), whereas no overlap was found between *gata5*^{tm236a/tm236a} and *tbx5a*^{m21/m21}. One region at Chr 21: 15,013,048–15,013,154 was commonly up-regulated in all three mutants. To further explore genomic locations of differentially regulated distal NFRs and identify evolutionary conserved putative enhancers, we compared them with the database of highly conserved noncoding elements (HCNEs) between zebrafish and human (Fig. 6C; Supplemental Table 11; Engström et al. 2008). We found a total of 22 regions to be conserved between zebrafish and human genomic sequences among which three were down-regulated in *tbx5a* and *hand2* mutants, whereas 19 of them showed significantly increased accessibility in *hand2* and *gata5* mutants. The three most down-regulated HCNEs were localized on Chromosome 1, between *hand2* and *fbxo8* loci (Chr 1: 37,584,384–37,584,724) as well as those localized in the introns of *ppp3ccb* (Chr 10: 20,246,264–20,246,845) and *akt7a* (Chr 20: 4,714,760–4,715,050) genes (Fig. 6D). We also identified HCNE-NFRs with increased chromatin ac-

cessibility in *gata5* mutant (Chr 1: 8,598,642–8,598,893) and in the genomic region at Chr 10: 8,580,509–8,581,153, which was commonly up-regulated in *hand2* and *gata5* mutants (Fig. 6E). Therefore, we have determined a number of distal NFRs whose accessibility is affected by mutations of cardiac TFs, among which we identified highly conserved NFRs serving as potential enhancers that may play key roles in heart development.

Discussion

Heart development involves multiple layers of interactions at molecular, cellular, and tissue levels. These processes are regulated by various TFs, signaling proteins, as well as epigenetic factors such as histone and DNA modifications, chromatin remodeling, and transcriptional enhancers. We obtained CM-enriched cell fractions from developing heart during crucial events of heart morphogenesis. GFP-positive cells were sorted from transgenic Tg(*rxk2.5*:EGFP), Tg(*myl7*:EGFP) zebrafish embryos. In zebrafish, at 6–9 somite stage (~12–14 hpf), *nkx2.5* expression only partially overlaps the anterior lateral plate mesoderm (ALPM) in its medial part (Schoenebeck et al. 2007), whereas at 17 somite stage (~17–18 hpf), the most posterior *nkx2.5*+ cells of the bilateral cardiac primordia do not express *myl7*, a marker of terminal myocardial differentiation, suggesting the presence of *nkx2.5*+ cells that do not contribute to the myocardium (Yelon et al. 1999). This agrees with other studies in zebrafish, suggesting the presence of specific *nkx2.5*+ second heart field (SHF) progenitors that give rise to the fraction of ventricular myocardium and outflow tract (OFT) (Guner-Ataman et al. 2013). Nevertheless, it has been shown that at prim-5 stage (24–30 hpf), *nkx2.5* is expressed both in ventricular and atrial myocardium exactly overlapping the expression of *myl7* (Yelon et al. 1999). The most prominent changes in gene expression and chromatin accessibility occurred between linear heart tube formation (24 hpf) and looping (48 hpf). This major shift in molecular profile likely reflects the continuous process of CM differentiation throughout which progenitors migrate and differentiate into CMs once they are incorporated into the growing heart tube (Kelly et al. 2014). Importantly, genes in the sienna3 and turquoise modules showed significant increase in expression between the two developmental stages. In particular, sienna3 genes were enriched in the largest number of GO terms related to cardiac function and contained at least three TFs known for their crucial roles in specification of CMs and their function in heart contraction (Singh et al. 2005, 2010), which suggests the prominent role of this network in CM differentiation and heart tube formation during this developmental period. Concordantly, we observed that chromatin landscape changed most significantly between 24 and 48 hpf, suggesting that the changes in gene expression profiles during this stage were likely regulated at the chromatin level. Besides validating the biological relevance of our ATAC-seq data set, this observation suggests that active chromatin remodeling occurs throughout development, and the regions with differential accessibility represent *cis*-regulatory hubs driving the biological processes associated with differentiating CMs.

Modules of coregulated genes represent subnetworks underlying specific biological processes associated with heart development. Further integration of these networks with ATAC-seq data allowed us to link TFs to their putative target genes, which was supported by the enrichment of DNA binding motif for specific TFs within NFRs in proximal promoters of the genes within each particular module. Collectively, our analyses of the regulatory networks and their representative expression patterns revealed

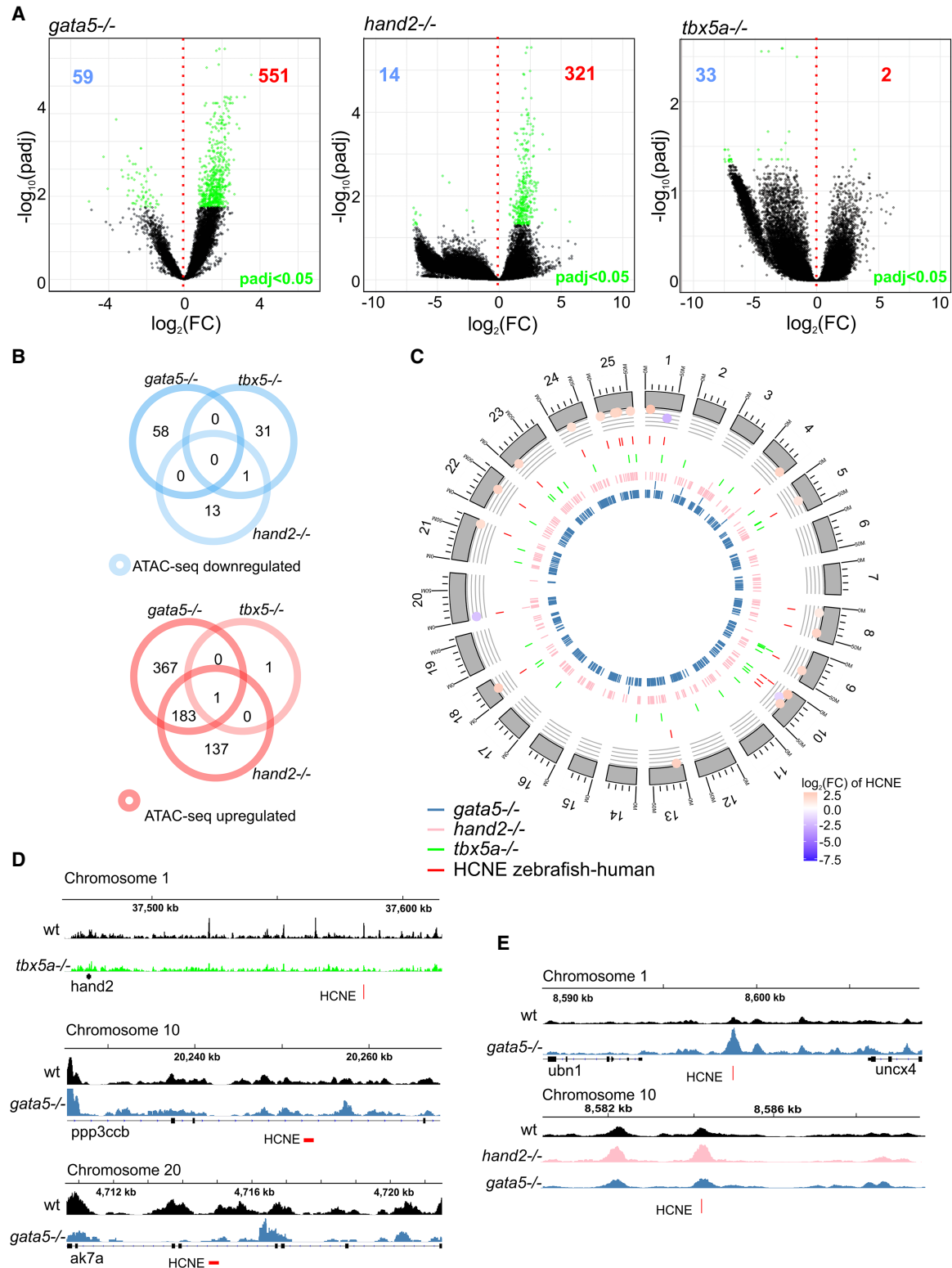


Figure 6. Identification of putative cardiac enhancers. (A) A volcano plot of differentially accessible distal NFRs between wild-type and TF mutants at 72 hpf. Adjusted P -value ≤ 0.05 are indicated in green, the number of down-regulated NFRs is indicated in blue, and up-regulated NFRs in red. (B) Venn diagram of mutant down- and up-regulated distal NFRs (more than ± 3 kb of TSS); adjusted P -value ≤ 0.05 . (C) Graphical representation of differentially accessible distal NFRs genomic localization onto zebrafish chromosomes. NFRs overlapping with HCNE (± 500 bp) and their accessibility \log_2 FC in comparison to wild type is indicated; adjusted P -value < 0.05 . (D) Genome track of ATAC-seq peaks for wild type (black), *tbx5a*^{-/-} (green), and *gata5*^{-/-} (blue) for the three most down-regulated NFRs overlapping with HCNE (± 500 bp). (E) Genome track of ATAC-seq peaks for wild type (black), *hand2*^{-/-} (pink), and *gata5*^{-/-} (blue) of the three most up-regulated NFRs overlapping with HCNE (± 500 bp).

increased expression of genes defining CM structure and function, whereas the expression and proximal promoter chromatin accessibility of hematopoietic genes were suppressed during CM differentiation. Sorted GFP-positive cells also expressed hematopoietic determinants at the earliest stage observed (24 hpf). These were grouped into a single expression module (brown) and correlated between gene expression dynamics and chromatin accessibility in proximal promoters that decreased between 24 and 48 hpf. One explanation is that the expression of hemato-vascular genes was contributed by cells giving rise to the pharyngeal arch mesoderm which also express *nkx2.5* used as our selection marker. Another equally plausible hypothesis is that a group of cells that possess alternative potential to become the blood or vascular lineage exist within the pool of CM progenitors. Numerous evidences from mouse studies suggested the presence of bipotential cardiac progenitor populations which coexpressed cardiac and hematopoietic markers in the developing heart tube (Caprioli et al. 2011; Nakano et al. 2013; Zamir et al. 2017). To distinguish between these possibilities, it would be necessary to obtain molecular profiles of individual cells to determine whether hemato-vascular progenitors exist as a separate population expressing specific markers or rather, as a common progenitor population expressing both CM and hemato-vascular markers. This also highlights the limitations of currently available marker genes and calls for higher resolution analyses of gene expression in specific cell types, which is possible with the single cell sequencing technology.

Finally, comparing wild-type CMs to that of *Gata5*, *Hand2*, and *Tbx5* mutants, we observed only a minor correlation between changes in gene expression and chromatin accessibility within proximal promoter NFRs, suggesting that transcriptional regulation of genes involved in heart development might be affected by distal regulatory elements. Alternatively, changes in gene expression between wild-type and TF mutants could be related to impaired TF binding to constitutively accessible proximal NFRs. Because we could only perform mutant analyses at 72 hpf, we could not rule out the possibility of observing secondary effects arising from changes in earlier developmental stages. Moreover, the lack of chromatin interaction data prevents the inference of definitive associations between distal regulatory elements and their target genes. Regardless, we identified a substantial number of gene-distal-located NFRs that were altered in accessibility in mutants that may serve as potential distal transcriptional regulatory elements, some of which were highly conserved between zebrafish and human, suggesting that they might be critical developmental enhancers (Polychronopoulos et al. 2017).

Altogether, we characterized the dynamics of gene expression and chromatin landscape during heart development and identified genetic regulatory hubs driving biological processes in CMs at different stages of heart morphogenesis. We elucidated the alterations in global transcriptional regulatory landscape resulting from disruptions to the developmental program caused by the loss of cardiac TFs. Collectively, our study identified potential target genes and regulatory elements implicated in heart development and disease.

Methods

CM collection by fluorescence-activated cell sorting (FACS)

Zebrafish transgenic lines *Tg(nkx2.5:EGFP)*, *Tg(myf7:EGFP)* in AB wild type and *gata5^{tm236a/+}* (Reiter et al. 1999), *tbx5a^{m21/+}* (Garrity et al. 2002), *hand2^{s6/+}* (Yelon et al. 2000) mutant back-

ground were maintained in the International Institute of Molecular and Cell Biology (IIMCB) zebrafish facility (License no. PL14656251) according to standard procedures. Cell suspension was prepared from 500 embryos and larvae as previously described (Winata et al. 2013), omitting the fixation step and directly resuspending cells in FACSmax Cell Dissociation Solution (AMS Biotechnology) for cell sorting. Fluorescent (GFP+) and nonfluorescent cells (GFP-) were sorted by using FACSaria II cytometer (BD Biosciences).

qPCR

Total RNA was extracted from 100,000 GFP+ and GFP- cells using TRIzol LS (Thermo Fisher Scientific) according to the manufacturer's protocol and followed by DNase I (Life Technologies) treatment. Transcript first strand cDNA synthesis kit (Roche Life Science) was used to obtain cDNA. Relative mRNA expression was quantified by using FastStart SYBR Green master mix on the Light Cycler 96 instrument (Roche Life Science) with specific primer sets (Supplemental Table 12).

RNA-seq

For RNA-seq, 100,000 GFP+ and GFP- cells were sorted directly to TRIzol LS (Thermo Fisher Scientific). cDNA synthesis for next-generation sequencing (NGS) was performed by SMARTer Universal Low Input RNA Kit (Clontech Laboratories) as recommended by the manufacturer. Paired-end sequencing (2 × 75 bp reads) was performed with NextSeq 500 (Illumina). Pearson correlation of biological replicates and read distribution over zebrafish genome features were performed (Supplemental Fig. 6A,B).

Assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq)

For ATAC-seq, 60,000 GFP+ cells from zebrafish embryos were sorted to Hank's solution (1 × HBSS, 2 mg/mL BSA, 10 mM HEPES at pH 8.0), centrifuged for 5 min at 500g, and prepared for chromatin tagmentation as previously described (Buenrostro et al. 2015). Paired-end sequencing (2 × 75 bp reads) was performed with NextSeq 500 (Illumina).

Light sheet fluorescence microscopy (LSFM)

Embryos were maintained in medium containing 0.003% 1-phenyl-2-thiourea. For LSFM (Zeiss), embryos were collected and mounted in 1% low-melting agarose (Sigma-Aldrich). Images were analyzed with Imaris 8 software (Bitplane).

Bioinformatics analysis

Raw RNA-seq and ATAC-seq reads were quality checked using FastQC (0.11.5) (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Adapters were removed using Trimmomatic (0.36) (Bolger et al. 2014). Reads matching ribosomal RNA were removed using rRNA dust (Hasegawa et al. 2014). Reads quality filtering was performed using SAMtools (1.4) (Li et al. 2009). RNA-seq reads were aligned to the zebrafish reference genome (GRCz10) using STAR (2.5) (Supplemental Fig. 7; Dobin et al. 2013). Bowtie 2 (2.2.9) (Langmead and Salzberg 2012) was used to map ATAC-seq reads to the GRCz10 genome (Supplemental Fig. 8). Read distribution was assessed with Picard (2.10.3). NFR regions were identified as previously described (Buenrostro et al. 2013). Peaks of open chromatin regions were called using MACS2 (2.1.0) (Zhang et al. 2008). Enriched motifs in NFRs were identified using HOMER (Heinz et al. 2010). Downstream bioinformatics analysis were performed in R 3.4 using following Bioconductor and CRAN (Huber et al. 2015)

packages as indicated in Supplemental Data. RNA-seq gene counts and ATAC-seq NFR read counts for all samples were transformed to regularized log (rld) (Supplemental Tables 13, 14). Differentially accessible ATAC-seq peaks were quantified by DESeq2 (Supplemental Table 15). Gene network visualization and statistical analysis of gene networks was performed using Cytoscape (Cline et al. 2007). Metascape was used to visualize the output of GO enrichment analysis (Tripathi et al. 2015).

Data access

RNA-seq and ATAC-seq data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE120238.

Acknowledgments

We thank D. Garrity for sharing *tbx5a*^{m21/+} and D. Yelon for *hand2*^{s61/+} zebrafish mutant lines and D. Stainier for sharing Tg (*nkx2.5:GFP*) and Tg (*myl7:GFP*). We thank V. Korzh and members of the Winata laboratory for fruitful discussions. This work was supported by EU/FP7: Research Potential FISHMED, grant number 316125, National Science Centre, Poland, OPUS grant number 2014/13/B/NZ2/03863. M.P. is supported by Foundation for Polish Science and Ministry of Science and Higher Education, Poland and National Science Centre, Poland, SONATA grant number 2014/15/D/NZ5/03421. M.M. is a recipient of the Postgraduate School of Molecular Medicine doctoral fellowship for the program "Next generation sequencing technologies in biomedicine and personalized medicine." This work was also supported by Research Grants from MEXT Japan to the RIKEN Center for Life Science Technologies (RIKEN CLST) and to the RIKEN Center for Integrative Medical Sciences (RIKEN IMS).

Author contributions: M.P., K.Z.K., M.M., and J.A.R. performed bioinformatics and statistical analysis. M.P., K.Z.K., and A.M. collected embryos, performed in vivo experiments, and collected biological material. M.P. and K.A.N. prepared NGS libraries and performed RNA-seq and ATAC-seq. M.P. performed LFSM. L.B. and K.P. performed FACS. M.P., C.L.W., J.A.R., and K.H. contributed to genomic data analysis. M.P., K.Z.K., M.M., J.A.R., P.C., and C.L.W. contributed to the design of the study and interpreted data. M.P. prepared the figures. M.P. and C.L.W. conceived the study and wrote the manuscript. C.L.W. is the corresponding senior author.

References

- Ahn DG, Ruvinsky I, Oates AC, Silver LM, Ho RK. 2000. *tbx20*, a new vertebrate T-box gene expressed in the cranial motor neurons and developing cardiovascular structures in zebrafish. *Mech Dev* **95**: 253–258. doi:10.1016/S0925-4773(00)00346-4
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. doi:10.1093/bioinformatics/btu170
- Bruneau BG, Nemer G, Schmitt JP, Charron F, Robitaille L, Caron S, Conner DA, Gessler M, Nemer M, Seidman CE, et al. 2001. A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* **106**: 709–721. doi:10.1016/S0092-8674(01)00493-7
- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218. doi:10.1038/nmeth.2688
- Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* **109**: 21.29.1–29.29.9. doi:10.1002/0471142727.mb2129s109
- Caprioli A, Koyano-Nakagawa N, Iacovino M, Shi X, Ferdous A, Harvey RP, Olson EN, Kyba M, Garry DJ. 2011. Nkx2-5 represses *Gata1* gene expression and modulates the cellular fate of cardiac progenitors during embryogenesis. *Circulation* **123**: 1633–1641. doi:10.1161/CIRCULATIONAHA.110.008185
- Chan YH, Chiang MF, Tsai YC, Su ST, Chen MH, Hou MS, Lin KI. 2009. Absence of the transcriptional repressor Blimp-1 in hematopoietic lineages reveals its role in dendritic cell homeostatic development and function. *J Immunol* **183**: 7039–7046. doi:10.4049/jimmunol.0901543
- Chandra M, Escalante-Alcalde D, Bhuiyan MS, Orr AW, Kevil C, Morris AJ, Nam H, Dominic P, McCarthy KJ, Miriyala S, et al. 2018. Cardiac-specific inactivation of LPP3 in mice leads to myocardial dysfunction and heart failure. *Redox Biol* **14**: 261–271. doi:10.1016/j.redox.2017.09.015
- Chen Z, Huang W, Dahme T, Rottbauer W, Ackerman MJ, Xu X. 2008. Depletion of zebrafish essential and regulatory myosin light chains reduces cardiac function through distinct mechanisms. *Cardiovasc Res* **79**: 97–108. doi:10.1093/cvr/cvn073
- Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campillo I, Creech M, Gross B, et al. 2007. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**: 2366–2382. doi:10.1038/nprot.2007.324
- D'Amico L, Scott IC, Jungblut B, Stainier DY. 2007. A mutation in zebrafish *hmgcr1b* reveals a role for isoprenoids in vertebrate heart-tube formation. *Curr Biol* **17**: 252–259. doi:10.1016/j.cub.2006.12.023
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21. doi:10.1093/bioinformatics/bts635
- Dupays L, Shang C, Wilson R, Kotecha S, Wood S, Towers N, Mohun T. 2015. Sequential binding of MEIS1 and NKX2-5 on the Popdc2 gene: a mechanism for spatiotemporal regulation of enhancers during cardiogenesis. *Cell Rep* **13**: 183–195. doi:10.1016/j.celrep.2015.08.065
- Engström PG, Fredman D, Lenhard B. 2008. Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol* **9**: R34. doi:10.1186/gb-2008-9-2-r34
- Etard C, Roostalu U, Strähle U. 2010. Lack of Apobec2-related proteins causes a dystrophic muscle phenotype in zebrafish embryos. *J Cell Biol* **189**: 527–539. doi:10.1083/jcb.2009.12.125
- Garrity DM, Childs S, Fishman MC. 2002. The *heartstrings* mutation in zebrafish causes heart/fin Tbx5 deficiency syndrome. *Development* **129**: 4635–4645.
- George V, Colombo S, Targoff KL. 2015. An early requirement for *nkx2.5* ensures the first and second heart field ventricular identity and cardiac function into adulthood. *Dev Biol* **400**: 10–22. doi:10.1016/j.ydbio.2014.12.019
- Gering M, Yamada Y, Rabbitts TH, Patient RK. 2003. Lmo2 and Scf/Tal1 convert non-axial mesoderm into haemangioblasts which differentiate into endothelial cells in the absence of Gata1. *Development* **130**: 6187–6199. doi:10.1242/dev.00875
- Gottgens B, Nastos A, Kinston S, Piltz S, Delabesse EC, Stanley M, Sanchez MJ, Ciau-Uitz A, Patient R, Green AR. 2002. Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. *EMBO J* **21**: 3039–3050. doi:10.1093/emboj/cdf286
- Guner-Ataman B, Paffett-Lugassy N, Adams MS, Nevis KR, Jahangiri L, Obregon P, Kikuchi K, Poss KD, Burns CE, Burns CG. 2013. Zebrafish second heart field development relies on progenitor specification in anterior lateral plate mesoderm and *nkx2.5* function. *Development* **140**: 1353–1363. doi:10.1242/dev.088351
- Hasegawa A, Daub C, Carninci P, Hayashizaki Y, Lassmann T. 2014. MOIRAI: a compact workflow system for CAGE analysis. *BMC Bioinformatics* **15**: 144. doi:10.1186/1471-2105-15-144
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576–589. doi:10.1016/j.molcel.2010.05.004
- Hill JT, Demarest B, Gorski B, Smith M, Yost HJ. 2017. Heart morphogenesis gene regulatory networks revealed by temporal expression analysis. *Development* **144**: 3487–3498. doi:10.1242/dev.154146
- Holtzinger A, Evans T. 2007. Gata5 and Gata6 are functionally redundant in zebrafish for specification of cardiomyocytes. *Dev Biol* **312**: 613–622. doi:10.1016/j.ydbio.2007.09.018
- Houk AR, Yelon D. 2016. Strategies for analyzing cardiac phenotypes in the zebrafish embryo. *Methods Cell Biol* **134**: 335–368. doi:10.1016/bs.mcb.2016.03.002
- Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, et al. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**: 115–121. doi:10.1038/nmeth.3252
- Jensen B, Wang T, Christoffels VM, Moorman AF. 2013. Evolution and development of the building plan of the vertebrate heart. *Biochim Biophys Acta, Mol. Cell* **1833**: 783–794. doi:10.1016/j.bbamcr.2012.10.004

- Kaneko H, Shimizu R, Yamamoto M. 2010. GATA factor switching during erythroid differentiation. *Curr Opin Hematol* **17**: 163–168. doi:10.1097/MOH.0b013e32833800b8
- Kelly RG, Buckingham ME, Moorman AF. 2014. Heart fields and cardiac morphogenesis. *Cold Spring Harb Perspect Med* **4**: a015750. doi:10.1101/cshperspect.a015750
- Kern CB, Wessels A, McGarity J, Dixon LJ, Alston E, Argraves WS, Geeting D, Nelson CM, Menick DR, Apte SS. 2010. Reduced versican cleavage due to *Adamts9* haploinsufficiency is associated with cardiac and aortic anomalies. *Matrix Biol* **29**: 304–316. doi:10.1016/j.matbio.2010.01.005
- Kirchmaier BC, Poon KL, Schwerte T, Huisken J, Winkler C, Jungblut B, Stainier DY, Brand T. 2012. The Popeye domain containing 2 (*popdc2*) gene in zebrafish is required for heart and skeletal muscle development. *Dev Biol* **363**: 438–450. doi:10.1016/j.ydbio.2012.01.015
- Knight HG, Yelon D. 2016. Utilizing zebrafish to understand second heart field development. In *Etiology and morphogenesis of congenital heart disease: from gene function and cellular interaction to morphology* (ed. Nakanishi T, et al.), pp. 193–199. Springer, Tokyo. doi:10.1007/978-4-431-54628-3_25
- Kotkamp K, Mössner R, Allen A, Onichtchouk D, Driever W. 2014. A Pou5f1/Oct4 dependent Klf2a, Klf2b, and Klf17 regulatory sub-network contributes to EVL and ectoderm development during zebrafish embryogenesis. *Dev Biol* **385**: 433–447. doi:10.1016/j.ydbio.2013.10.025
- Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559. doi:10.1186/1471-2105-9-559
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359. doi:10.1038/nmeth.1923
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. doi:10.1093/bioinformatics/btp352
- Liberatore CM, Searcy-Schrick RD, Yutzey KE. 2000. Ventricular expression of *tbx5* inhibits normal heart chamber development. *Dev Biol* **223**: 169–180. doi:10.1006/dbio.2000.9748
- Lien CL, Wu C, Mercer B, Webb R, Richardson JA, Olson EN. 1999. Control of early cardiac-specific transcription of *Nkx2-5* by a GATA-dependent enhancer. *Development* **126**: 75–84.
- Lopaschuk GD, Jaswal JS. 2010. Energy metabolic phenotype of the cardiomyocyte during development, differentiation, and postnatal maturation. *J Cardiovasc Pharmacol* **56**: 130–140. doi:10.1097/FJC.0b013e3181e74a14
- Lou X, Deshwar AR, Crump JG, Scott IC. 2011. Smarcd3b and *Gata5* promote a cardiac progenitor fate in the zebrafish embryo. *Development* **138**: 3113–3123. doi:10.1242/dev.064279
- Loughran SJ, Kruse EA, Hacking DF, de Graaf CA, Hyland CD, Willson TA, Henley KJ, Ellis S, Voss AK, Metcalf D, et al. 2008. The transcription factor Erg is essential for definitive hematopoiesis and the function of adult hematopoietic stem cells. *Nat Immunol* **9**: 810–819. doi:10.1038/ni.1617
- Marques SR, Yelon D. 2009. Differential requirement for BMP signaling in atrial and ventricular lineages establishes cardiac chamber proportionality. *Dev Biol* **328**: 472–482. doi:10.1016/j.ydbio.2009.02.010
- Miller SA, Huang AC, Miazgowski MM, Brassil MM, Weinmann AS. 2008. Coordinated but physically separable interaction with H3K27-demethylase and H3K4-methyltransferase activities are required for T-box protein-mediated activation of developmental gene expression. *Genes Dev* **22**: 2980–2993. doi:10.1101/gad.1689708
- Molkentin JD, Antos C, Mercer B, Taigen T, Miano JM, Olson EN. 2000. Direct activation of a *GATA6* cardiac enhancer by *Nkx2.5*: evidence for a reinforcing regulatory network of *Nkx2.5* and GATA transcription factors in the developing heart. *Dev Biol* **217**: 301–309. doi:10.1006/dbio.1999.9544
- Montero JA, Giron B, Arrechedera H, Cheng YC, Scotting P, Chimal-Monroy J, Garcia-Porrero JA, Hurlle JM. 2002. Expression of *Sox8*, *Sox9* and *Sox10* in the developing valves and autonomic nerves of the embryonic heart. *Mech Dev* **118**: 199–202. doi:10.1016/S0925-4773(02)00249-6
- Moskowitz IP, Wang J, Peterson MA, Pu WT, Mackinnon AC, Oxburgh L, Chu GC, Sarkar M, Berul C, Smoot L, et al. 2011. Transcription factor genes *Smad4* and *Gata4* cooperatively regulate cardiac valve development. [Corrected]. *Proc Natl Acad Sci* **108**: 4006–4011. doi:10.1073/pnas.1019025108
- Nakano H, Liu X, Arshi A, Nakashima Y, van Handel B, Sasidharan R, Harmon AW, Shin JH, Schwartz RJ, Conway SJ, et al. 2013. Haemogenic endocardium contributes to transient definitive haematopoiesis. *Nat Commun* **4**: 1564. doi:10.1038/ncomms2569
- Nemer M. 2008. Genetic insights into normal and abnormal heart development. *Cardiovasc Pathol* **17**: 48–54. doi:10.1016/j.carpath.2007.06.005
- Nimura K, Ura K, Shiratori H, Ikawa M, Okabe M, Schwartz RJ, Kaneda Y. 2009. A histone H3 lysine 36 trimethyltransferase links *Nkx2-5* to Wolf-Hirschhorn syndrome. *Nature* **460**: 287–291. doi:10.1038/nature08086
- Olson TM, Michels VV, Thibodeau SN, Tai YS, Keating MT. 1998. Actin mutations in dilated cardiomyopathy, a heritable form of heart failure. *Science* **280**: 750–752. doi:10.1126/science.280.5364.750
- Paw BH, Davidson AJ, Zhou Y, Li R, Pratt SJ, Lee C, Trede NS, Brownlie A, Donovan A, Liao EC, et al. 2003. Cell-specific mitotic defect and dyserythropoiesis associated with erythroid band 3 deficiency. *Nat Genet* **34**: 59–64. doi:10.1038/ng1137
- Pimanda JE, Ottersbach K, Knezevic K, Kinston S, Chan WY, Wilson NK, Landry JR, Wood AD, Kolb-Kokocinski A, Green AR, et al. 2007. *Gata2*, *Fli1*, and *Scl* form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proc Natl Acad Sci* **104**: 17692–17697. doi:10.1073/pnas.0707045104
- Piven OO, Winata CL. 2017. The canonical way to make a heart: β -catenin and plakoglobin in heart development and remodeling. *Exp Biol Med* **242**: 1735–1745. doi:10.1177/1535370217732737
- Polychronopoulos D, King JWD, Nash AJ, Tan G, Lenhard B. 2017. Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic Acids Res* **45**: 12611–12624. doi:10.1093/nar/gkx1074
- Powers SE, Taniguchi K, Yen W, Melhuish TA, Shen J, Walsh CA, Sutherland AE, Wotton D. 2010. *Tgif1* and *Tgif2* regulate Nodal signaling and are required for gastrulation. *Development* **137**: 249–259. doi:10.1242/dev.040782
- Reiter JF, Alexander J, Rodaway A, Yelon D, Patient R, Holder N, Stainier DY. 1999. *Gata5* is required for the development of the heart and endoderm in zebrafish. *Genes Dev* **13**: 2983–2995. doi:10.1101/gad.13.22.2983
- Sakabe NJ, Aneas I, Shen T, Shokri L, Park SY, Bulyk ML, Evans SM, Nobrega MA. 2012. Dual transcriptional activator and repressor roles of *TBX20* regulate adult cardiac structure and function. *Hum Mol Genet* **21**: 2194–2204. doi:10.1093/hmg/dds034
- Schoenebeck JJ, Keegan BR, Yelon D. 2007. Vessel and blood specification override cardiac potential in anterior mesoderm. *Dev Cell* **13**: 254–267. doi:10.1016/j.devcel.2007.05.012
- Shih YH, Zhang Y, Ding Y, Ross CA, Li H, Olson TM, Xu X. 2015. Cardiac transcriptome and dilated cardiomyopathy genes in zebrafish. *Circ Cardiovasc Genet* **8**: 261–269. doi:10.1161/CIRCGENETICS.114.000702
- Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**: 272–286. doi:10.1038/nrg3682
- Singh MK, Christoffels VM, Dias JM, Trowe MO, Petry M, Schuster-Gossler K, Burger A, Ericson J, Kispert A. 2005. *Tbx20* is essential for cardiac chamber differentiation and repression of *Tbx2*. *Development* **132**: 2697–2707. doi:10.1242/dev.01854
- Singh MK, Li Y, Li S, Cobb RM, Zhou D, Lu MM, Epstein JA, Morrisey EE, Gruber PJ. 2010. *Gata4* and *Gata5* cooperatively regulate cardiac myocyte proliferation in mice. *J Biol Chem* **285**: 1765–1772. doi:10.1074/jbc.M109.038539
- Soza-Ried C, Hess I, Netuschil N, Schorpp M, Boehm T. 2010. Essential role of *c-myb* in definitive hematopoiesis is evolutionarily conserved. *Proc Natl Acad Sci* **107**: 17304–17308. doi:10.1073/pnas.1004640107
- Stainier DY. 2001. Zebrafish genetics and vertebrate heart formation. *Nat Rev Genet* **2**: 39–48. doi:10.1038/35047564
- Stainier DY, Fishman MC. 1994. The zebrafish as a model system to study cardiovascular development. *Trends Cardiovasc Med* **4**: 207–212. doi:10.1016/1050-1738(94)90036-1
- Stainier DY, Lee RK, Fishman MC. 1993. Cardiovascular development in the zebrafish. I. Myocardial fate map and heart tube formation. *Development* **119**: 31–40.
- Takeuchi JK, Lou X, Alexander JM, Sugizaki H, Delgado-Olguín P, Holloway AK, Mori AD, Wylie JN, Munson C, Zhu Y, et al. 2011. Chromatin remodelling complex dosage modulates transcription factor function in heart development. *Nat Commun* **2**: 187. doi:10.1038/ncomms1187
- Targoff KL, Schell T, Yelon D. 2008. *Nkx* genes regulate heart tube extension and exert differential effects on ventricular and atrial cell number. *Dev Biol* **322**: 314–321. doi:10.1016/j.ydbio.2008.07.037
- Tripathi S, Pohl MO, Zhou Y, Rodriguez-Frandsen A, Wang G, Stein DA, Moulton HM, DeJesus P, Che J, Mulder LC, et al. 2015. Meta- and orthogonal integration of influenza “OMICs” data defines a role for UBR4 in virus budding. *Cell Host Microbe* **18**: 723–735. doi:10.1016/j.chom.2015.11.002
- Turbendian HK, Gordillo M, Tsai SY, Lu J, Kang G, Liu TC, Tang A, Liu S, Fishman GI, Evans T. 2013. GATA factors efficiently direct cardiac fate from embryonic stem cells. *Development* **140**: 1639–1644. doi:10.1242/dev.093260
- Ueno S, Weidinger G, Osugi T, Kohn AD, Golob JL, Pabon L, Reinecke H, Moon RT, Murry CE. 2007. Biphasic role for *Wnt/β-catenin* signaling in cardiac specification in zebrafish and embryonic stem cells. *Proc Natl Acad Sci* **104**: 9685–9690. doi:10.1073/pnas.0702859104

- Wang B, Yan J, Peng Z, Wang J, Liu S, Xie X, Ma X. 2011. Teratocarcinoma-derived growth factor 1 (*TDGF1*) sequence variants in patients with congenital heart defect. *Int J Cardiol* **146**: 225–227. doi:10.1016/j.ijcard.2009.08.046
- Winata CL, Kondrychyn I, Kumar V, Srinivasan KG, Orlov Y, Ravishankar A, Prabhakar S, Stanton LW, Korzh V, Mathavan S. 2013. Genome wide analysis reveals *Zic3* interaction with distal regulatory elements of stage specific developmental genes in zebrafish. *PLoS Genet* **9**: e1003852. doi:10.1371/journal.pgen.1003852
- Witzel HR, Jungblut B, Choe CP, Crump JG, Braun T, Dobrev G. 2012. The LIM protein *Ajuba* restricts the second heart field progenitor pool by regulating *Isl1* activity. *Dev Cell* **23**: 58–70. doi:10.1016/j.devcel.2012.06.005
- Xu C, Liguori G, Persico MG, Adamson ED. 1999. Abrogation of the *Cripto* gene in mouse leads to failure of postgastrulation morphogenesis and lack of differentiation of cardiomyocytes. *Development* **126**: 483–494.
- Xu X, Meiler SE, Zhong TP, Mohideen M, Crossley DA, Burggren WW, Fishman MC. 2002. Cardiomyopathy in zebrafish due to mutation in an alternatively spliced exon of titin. *Nat Genet* **30**: 205–209. doi:10.1038/ng816
- Yelon D, Horne SA, Stainier DY. 1999. Restricted expression of cardiac myosin genes reveals regulated aspects of heart tube assembly in zebrafish. *Dev Biol* **214**: 23–37. doi:10.1006/dbio.1999.9406
- Yelon D, Ticho B, Halpern ME, Ruvinsky I, Ho RK, Silver LM, Stainier DY. 2000. The bHLH transcription factor *Hand2* plays parallel roles in zebrafish heart and pectoral fin development. *Development* **127**: 2573–2582.
- Zamir L, Singh R, Nathan E, Patrick R, Yifa O, Yahalom-Ronen Y, Arraf AA, Schultheiss TM, Suo S, Han JJ, et al. 2017. *Nkx2.5* marks angioblasts that contribute to hemogenic endothelium of the endocardium and dorsal aorta. *eLife* **6**: e20994. doi:10.7554/eLife.20994
- Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137. doi:10.1186/gb-2008-9-9-r137

Received September 21, 2018; accepted in revised form January 9, 2019.

RESEARCH

Open Access



Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift

Maciej Migdał^{1†}, Eugeniusz Tralle^{1†}, Karim Abu Nahia¹, Łukasz Bugajski², Katarzyna Zofia Kędzierska¹, Filip Garbicz^{3,4}, Katarzyna Piwocka², Cecilia Lanny Winata^{1*} and Michał Pawlak^{1,3*}

Abstract

Background: Liver fibrosis is a wound-healing response to tissue injury and inflammation hallmarked by the extra-cellular matrix (ECM) protein deposition in the liver parenchyma and tissue remodelling. Different cell types of the liver are known to play distinct roles in liver injury response. Hepatocytes and liver endothelial cells receive molecular signals indicating tissue injury and activate hepatic stellate cells which produce ECM proteins upon their activation. Despite the growing knowledge on the molecular mechanism underlying hepatic fibrosis in general, the cell-type-specific gene regulatory network associated with the initial response to hepatotoxic injury is still poorly characterized.

Results: In this study, we used thioacetamide (TAA) to induce hepatic injury in adult zebrafish. We isolated three major liver cell types - hepatocytes, endothelial cells and hepatic stellate cells - and identified cell-type-specific chromatin accessibility and transcriptional changes in an early stage of liver injury. We found that TAA induced transcriptional shifts in all three cell types hallmarked by significant alterations in the expression of genes related to fatty acid and carbohydrate metabolism, as well as immune response-associated and vascular-specific genes. Interestingly, liver endothelial cells exhibit the most pronounced response to liver injury at the transcriptome and chromatin level, hallmarked by the loss of their angiogenic phenotype.

Conclusion: Our results uncovered cell-type-specific transcriptome and epigenome responses to early stage liver injury, which provide valuable insights into understanding the molecular mechanism implicated in the early response of the liver to pro-fibrotic signals.

Keywords: Liver, Hepatocytes, Stellate cells, Endothelial cells, Chromatin, Transcriptomics, ATAC-seq, RNA-seq, Genomics, Epigenomics, Zebrafish

Background

Liver injury is a rising public health concern, especially in European and North American countries. Its increasing prevalence leads to an expanding body of work regarding the molecular mechanisms present in advanced liver disease, however our knowledge about the earliest stages of liver injury is still limited. Liver injury is manifested by the formation of fibrous tissue as a result of ECM deposition at the site of injury [1]. Progressive fibrous scar formation may distort normal liver structure by formation of septa and nodules of regenerating hepatocytes (HEPs)

*Correspondence: cwinata@iimcb.gov.pl; mpawlak@ihit.waw.pl

†Maciej Migdał and Eugeniusz Tralle contributed equally to this work.

¹ International Institute of Molecular and Cell Biology in Warsaw, Laboratory of Zebrafish Developmental Genomics, 4 Ks. Trojdena Street, 02-109 Warsaw, Poland

³ Department of Experimental Hematology, Institute of Hematology and Transfusion Medicine, ul. Indiry Gandhi 14, 02-776 Warsaw, Poland

Full list of author information is available at the end of the article



leading to impaired portal blood flow and formation of cirrhotic architecture [2]. Liver cirrhosis is the end-stage of hepatic fibrosis affecting about 0.1% of the European population [1]. The most serious outcome of cirrhosis is hepatocellular carcinoma (HCC), constituting 70-90% of cases of primary liver cancer [1]. The predominant causes of liver fibrosis are chronic excessive alcohol consumption, viral hepatitis B and C and non-alcoholic fatty liver disease (NAFLD), the latter becoming a major concern with the increasing incidence of obesity in Europe and the USA [1].

Liver parenchymal cells, HEPs, are the most abundant cell subpopulation in this organ in mammals, constituting ca. 85% of the total liver cell mass [3]. Under physiological conditions, HEPs are responsible for a wide range of functions, including carbohydrate, fatty acid and protein metabolism as well as immune response [3]. Upon liver damage, HEPs are a source of reactive oxygen species, pro-inflammatory signals as well as cytokines, taking part in the activation of repair pathways [3].

Hepatic stellate cells (HSCs) comprise 8% of the total liver cell population [4]. Under normal physiological conditions, these mesenchymal cells reside in the space of Disse, maintaining a quiescent state, storing vitamin A in cytoplasmic lipid droplets [5]. Upon liver damage, HSCs are activated and transdifferentiate into myofibroblast-like cells. Their activation is triggered by multiple autocrine and paracrine signals, such as transforming growth factor (TGF β), SMAD3, protein platelet-derived growth factor (PDGF), vascular endothelial growth factor (VEGF) and connective tissue growth factor (CTGF) [6]. In an active state, HSCs are the primary ECM-producing cell population, resulting in the creation of a temporary scar tissue at the damaged site. Active HSCs produce cytokines and growth factors, promoting liver regeneration. In chronic liver disease, however, the reoccurring HSC activation may result in permanent scar formation, resulting in sections of non-functional liver tissue [5].

Endothelial cells in the liver are found mainly lining the inner walls of the sinusoidal blood vessels (liver sinusoidal endothelial cells - LSECs). LSECs are highly specialized, forming a permeable barrier by virtue of their fenestrae, between hepatocyte membranes and blood vessel lumen. The presence of fenestrae, combined with the absence of a basement membrane, contribute to making the LSECs the most endocytosis-capable cell population in the human body [7]. LSECs regulate the tone of hepatic blood vessels and maintain the quiescent state of HSCs [7].

In response to chronic hepatotoxic injury, various molecular and cellular factors interact with HEPs and LSECs, leading to sequential activation of HSCs [8]. This

in turn initiates the perpetuation phase, hallmarked by proliferative, contractile and inflammatory phenotype characterized by increased production of ECM proteins including collagens, fibronectin, decorin, elastin and proteoglycans [2, 9]. The understanding of molecular mechanisms of hepatic fibrosis has markedly increased due to the availability of liver fibrosis models such as cell culture systems, rodent model systems and biopsied human material [10]. However, our knowledge of cell-type-specific gene regulatory networks and epigenetic hallmarks associated with the initial response to hepatotoxic injury is still lacking, mainly due to the challenges of studying cell interactions and their behaviour in a living organism. Such knowledge is crucial for accurate diagnosis and development of new therapeutic approaches targeting liver fibrosis and related disorders.

The zebrafish (*Danio rerio*) has emerged as a useful model organism for studying the mechanism of liver disease in vivo, both in larvae and adult individuals [11–13]. Despite the distinct architecture between mammalian and zebrafish liver, they contain similar main cell types, including HEPs, endothelial cells (ECs) and HSCs, with conserved function and gene expression profiles [5, 14, 15]. To dissect the molecular mechanisms regulating the initiation of hepatic fibrosis and understand the interplay between genetic and epigenetic signals in this process, we utilized the model of thioacetamide-induced liver injury in adult zebrafish and characterized cell-type-specific changes at both transcriptome and epigenome level in three main liver cell types. Thioacetamide (TAA) is a potent hepatotoxin that has been widely used to induce acute and chronic liver injury in rodent models [16–18]. There is a wide variation in the administration routes and time of exposure between studies, but most commonly a regimen of intraperitoneal injections of 100-200 mg/kg of body mass 2-3 times per week for over 6 weeks has been used to induce liver fibrosis and cirrhosis [19]. TAA has also been utilized to induce liver injury in zebrafish larvae, establishing it as a model for steatohepatitis [13]. The larvae used in the cited study were exposed to 0.025% TAA for 10 days starting at 72 h post-fertilization (hpf), when the embryonic liver becomes functional. At 5 days post-fertilization the embryos exhibited molecular markers of apoptosis and steatohepatitis, which continued until the end of the treatment. TAA has also been used in juvenile zebrafish, where intraperitoneal injections of 300 mg/kg b.m. three times a week induced steatosis [20].

We employed three transgenic zebrafish lines to isolate the respective cell populations: HEPs (*Tg(fabp10a:dsRed)*), HSCs (*Tg(hand2:EGFP)*), and ECs (*Tg(kdrl:ras-mCherry)*). We implemented a machine learning technique known as self-organizing maps (SOMs) to generate whole genome expression profiles

of both physiological state and early response to liver injury from the three studied cell types [21]. The integration of this data with genome-wide open chromatin maps (ATAC-seq) from corresponding samples allowed to uncover specific gene and chromatin signatures of the studied cell populations. Our analysis revealed that early response of the liver to pro-fibrotic signals is manifested in cell-type specific transcriptome and epigenome rearrangements and identified molecular hallmarks of this process. This work provides a step towards understanding the initial stages of liver injury and may serve as a resource for further investigation aimed at developing new diagnostic and treatment tools.

Results

Identification of liver cell-type-specific transcriptional portraits under normal physiological condition

In order to characterize the molecular profiles representing the HEPs, HSCs, and ECs under physiological conditions, we utilized three transgenic lines *Tg(fabp10a:dsRed)*, *Tg(hand2:EGFP)* and *Tg(kdrl:Hsa.HRAS-mCherry)* which express red (dsRed, mCherry) or green fluorescent proteins (GFP) in the corresponding cell types [14, 22, 23]. Whole livers were dissected from adult zebrafish from each of the transgenic lines used in this study (Fig. 1A). Fluorescent microscopy of liver from the corresponding transgenic lines confirmed the fluorescence observed in the corresponding cell types (Fig. 1B). We prepared cell suspensions and performed FACS according to previously established protocols (See [Methods](#), Supp. Fig. 1). The number of RNA-seq reads corresponding to fluorescent reporters specific to each cell-type (Fig. 1B) was strongly enriched in fluorescent-positive samples, which confirmed the purity of FACS isolated samples (Fig. 1C). In order to ascertain the cell-type gene signatures, we performed differential expression comparisons between samples and identified the most enriched genes in each cell type (Fig. 2A, Supp. Table 2). The largest number of cell-specific genes were found in ECs (4553), then in HSCs (380) and in HEPs (126) (Supp. Table 2). These included known cell-specific markers for ECs (*sox18* [24], *sele* [25], *flt1* [26]) and HEPs (*soat2* [27]) (Fig. 2B). On the other hand, genes related to fatty acid metabolism (*fasn* [28], *fat3b*, *hmgcra* [29], *hmgcs1* [30], *elovl4a* [31]) and cholesterol biosynthesis (*cyp51*, *sc5d*, *hmgcra*, *msmo1*, *nsdhl*, *hmgcs1*, *dhcr7*) were upregulated in HSCs which are known to contain vitamin A lipid droplets [32] (Supplementary Table 2). Gene ontology (GO) analysis revealed the enrichment of genes related to angiogenesis in ECs, insulin-like growth factor receptor signalling genes and cellular phosphate ion homeostasis in HEPs and lipid transport and metabolism genes in HSCs (Fig. 2C). Taken together, the enrichment

of known markers and the relevant GO terms in ECs, HEPs, and HSCs support the identity of the respective cell types.

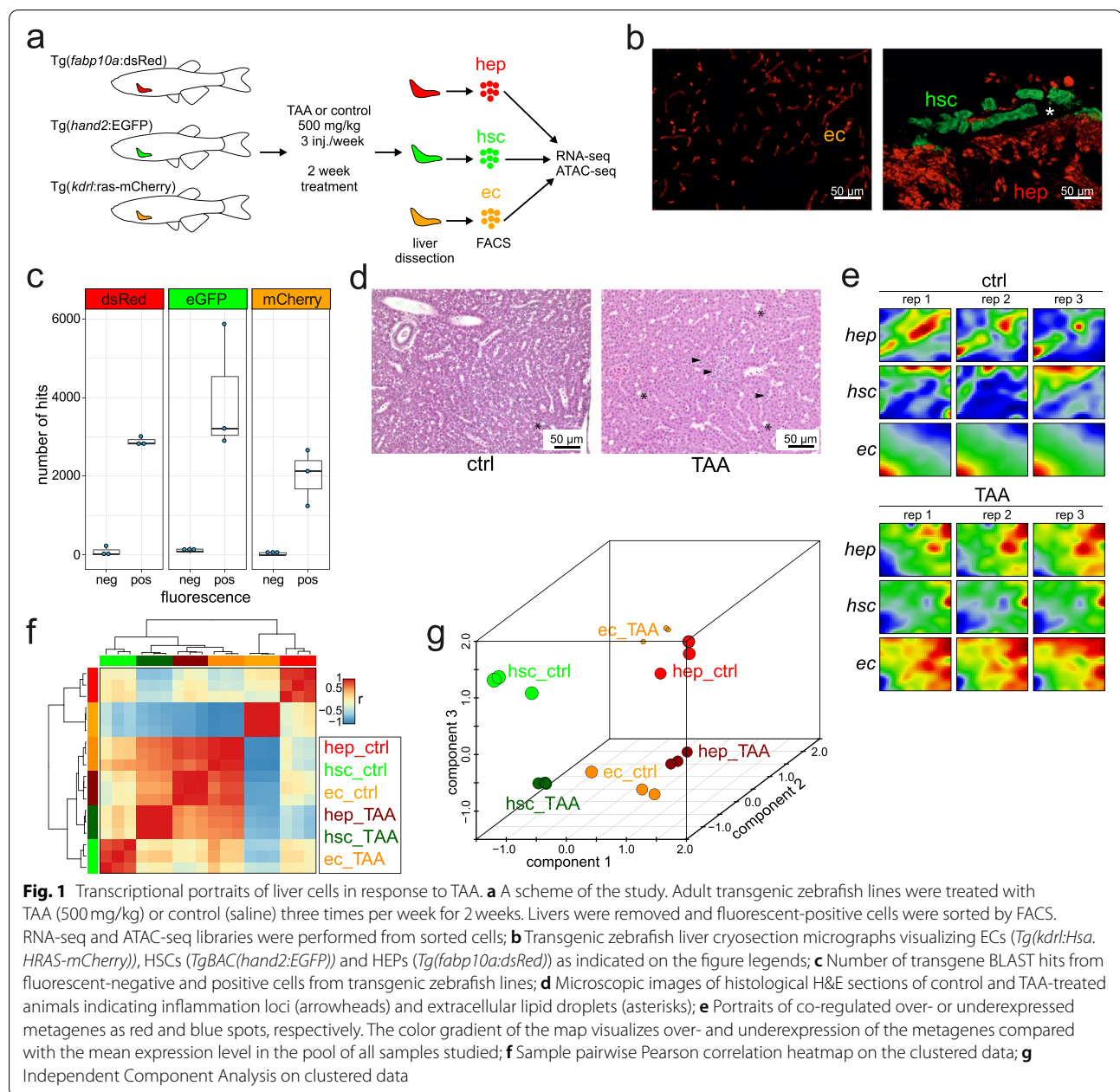
TAA metabolism is reflected in the transcriptional shift in liver cells

We then sought to determine the transcriptional signatures of early hepatotoxic injury response in each of the three liver cell types. We induced liver injury using TAA at a concentration of 500 mg/kg of body mass. The short term TAA treatment induced mild histological changes with observed inflammation (Fig. 1D). We then collected whole livers from TAA-treated *Tg(fabp10a:dsRed)*, *TgBAC(hand2:EGFP)* and *Tg(kdrl:Hsa.HRAS-mCherry)* fishes, isolated the corresponding cell types by FACS, and performed RNA-seq.

We evaluated cell-type-specific transcriptional response to TAA activation by looking at the expression of genes related to TAA metabolism and genes activated in response to liver injury and fibrogenesis (Fig. 2D, Supp. Table 3). The increased expression of genes related to cell redox homeostasis such as catalase (*cat*) [33], cytochromes (*cyp2y3*, *cyp2p6*) [34], superoxide dismutase 2 (*sod2*) [34], glutathione peroxidase 1a (*gpx1a*) [35] was observed in response to TAA, with the most striking response in ECs. Pro-fibrotic genes [8] including ECM proteins such as collagens (*col1a1a*, *col1a2*, *col5a2a*, *col5a1*, *col6a3*), decorin (*dcn*) as well as metalloproteinase inhibitor 2a (*timp2a*), integrin alpha V (*itgav*) and annexin 5b (*anxa5b*) were specifically upregulated in HSCs, in response to TAA (Fig. 2D).

TAA induces transcriptional reprogramming of hepatic endothelial cells

To provide a global view of the behaviour of correlated gene clusters in three hepatic cell types in response to TAA, we used self-organizing map based tool oposSOM R package [36]. The tool first constructed transcriptional portraits of all the samples, then a second unsupervised reduction step was performed, further reducing dimensionality to overexpression spots representing clusters (A-H, Supp. Table 4) of co-expressed metagenes which are highly expressed in, at minimum, one condition (Fig. 3A, B) [37]. To link overexpression with gene set overrepresentation in a sample- and spot-specific way, we visualized the metagene expression across samples on the heatmap (Fig. 3C) and performed the gene set overrepresentation analysis (Fig. 3D, E; Supp. Table 5). The gene expression portraits of both control and TAA-treated samples from each of the three cell types revealed that short-term TAA exposure induced strong changes in genome-wide expression landscapes between cell types in physiological state and upon TAA activation (Fig. 1E,



F). Interestingly, the most striking changes induced by TAA treatment were observed in ECs (Fig. 1G).

Analysis of the SOM clusters in ECs revealed an increase in expression of genes related to metabolic and redox processes as well as cellular transport (Fig. 3C, D - clusters B and F). We also observed downregulation of

genes related to vasculature development as well as activation of immune response in ECs after treatment with TAA (Fig. 3C, D - clusters G and H; Supp. Fig. 6).

In HEPs, TAA treatment induced an increase in the expression of gene sets associated with regulation of metabolic processes, namely carboxylic acid and hydroxy

(See figure on next page.)

Fig. 2 Liver cell signatures in quiescent and activated state. **a** Number of identified cell type specific genes at quiescent state in each cell type, $\log_{2}FC > 0$, $\text{padj} < 0.05$; **b** Heatmaps of top 25 cell type specific genes at quiescent state in each cell type, $\log_{2}FC > 0$, $\text{padj} < 0.05$; **c** GO over-representation analysis of identified cell type specific genes at quiescent state in each cell type; **d** Volcano plot of selected genes, involved in liver fibrosis and response to oxidative stress, under TAA treatment

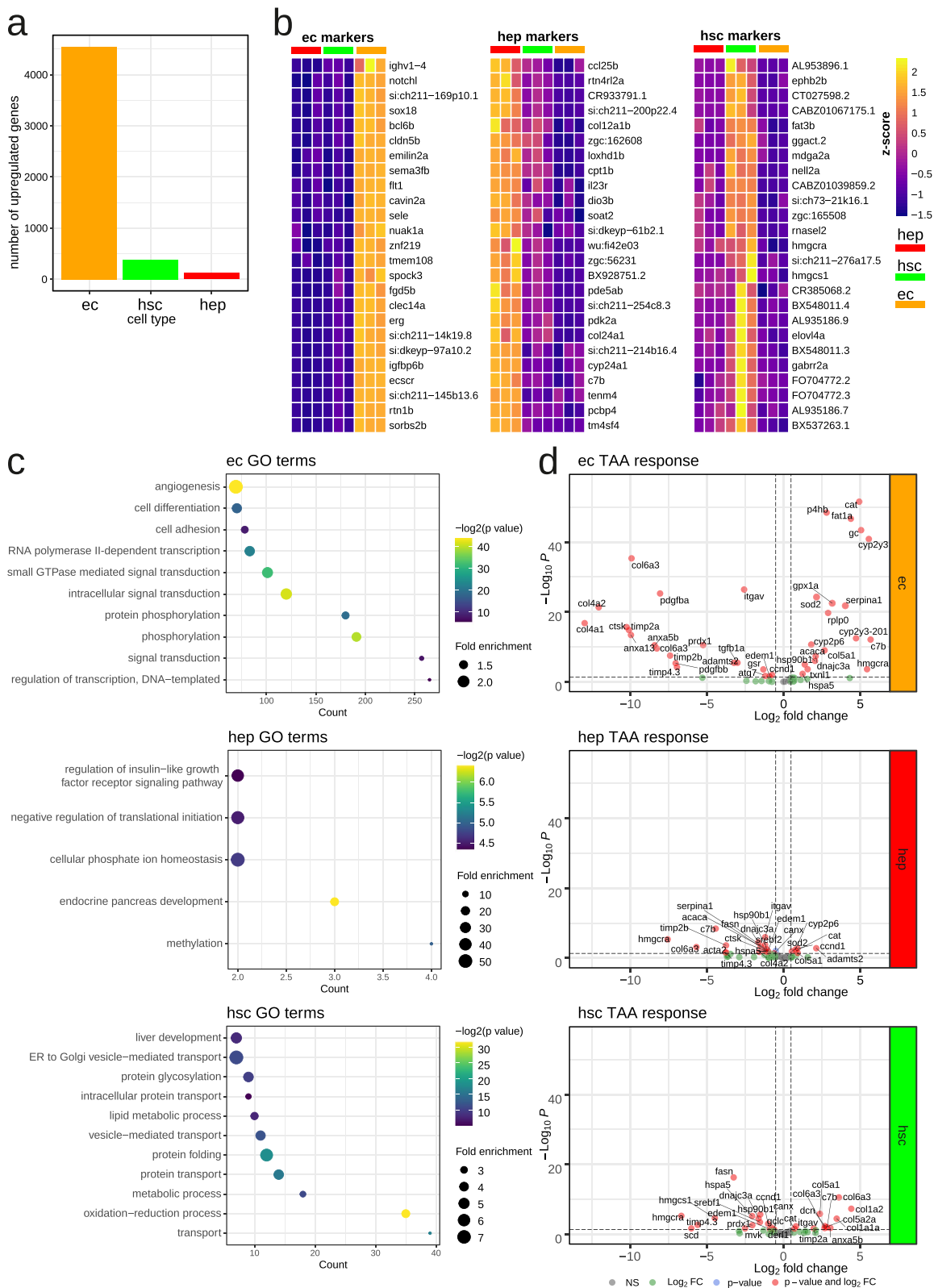
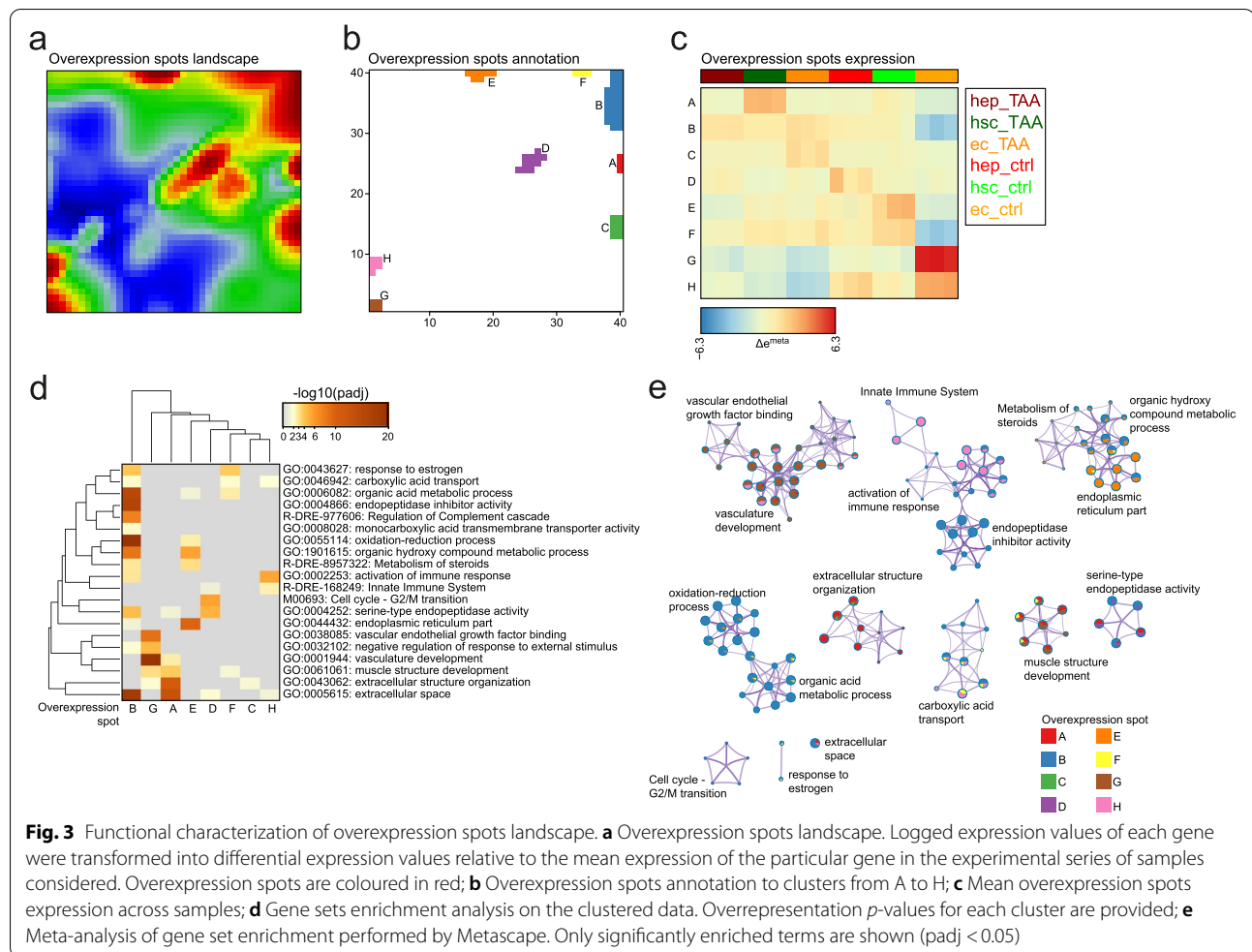


Fig. 2 (See legend on previous page.)



compound metabolism, as well as intra- and intercellular transport when compared to their control counterparts (Fig. 3C, D - cluster B). In contrast, we observed a decreased expression of gene sets associated with the formation and function of endoplasmic reticulum as well as negative regulation of various growth binding factors (Fig. 3C, D - clusters E and G). We also observed a relative reduction of expression of genes associated with the G2/M cell cycle transition in TAA-treated HEPs (Fig. 3C, D - cluster D; Supp. Fig. 5).

Modest changes in gene expression were observed in HSCs. Analysis of clusters revealed that upregulated gene sets were associated with extracellular space and structure organization as well as protein hydrolysis (Fig. 3C, D - cluster A), which reflects the known role of HSCs in ECM formation during liver damage response [9]. Conversely, we observed downregulation of genes associated with G2/M cell cycle transition, endoplasmic reticulum, estrogen response and immune activation (Fig. 3C, D - clusters G and H).

Altogether, cell-type-specific transcriptome profile revealed transcriptional response to short term TAA exposure. All of the analyzed cell types were subject to TAA-induced transcriptional shifts, with the highest change observed in ECs. These were hallmarked by decrease of vascular-specific genes and the increase of fatty acid and carbohydrate metabolism genes as well as in immune response-associated genes.

TAA leads to genome-wide changes in chromatin regions enriched in binding sites for transcription factors regulating fatty acid metabolism and angiogenesis

Epigenetics has been acknowledged as an important player in liver fibrosis and regeneration [38–40], with a prospect of the development of epigenetic biomarkers and therapies. To investigate this aspect of liver damage, we ask whether epigenetic changes are involved in the earliest stages of liver fibrosis. To determine whether and to what extent epigenetic landscape in each liver cell type is altered during early stage liver injury, we characterized

the changes in chromatin accessibility in HEPs, HSCs, and ECs upon TAA treatment.

We observed that in TAA-treated animals the most significant changes in chromatin state compared to control were observed in ECs, followed by HSCs and HEPs (Fig. 4A, B). ATAC-seq peaks distribution across the genome showed that the highest fraction of peaks (30-40%) was localized in the promoter (+/- 3kb) regions (Fig. 4C, Supp. Table 7). Interestingly, the most significant changes in chromatin accessibility was observed in ECs, with the largest number of upregulated peaks found

within the promoters of genes in clusters B (440 peaks) and F (74 peaks) and downregulated peaks in clusters G (120 peaks) and H (113 peaks) (Fig. 5A). The observed changes in chromatin accessibility correlates with changes observed in the transcriptional levels of genes within the corresponding clusters (increase in clusters B and F, and decrease in clusters G and H) (Fig. 4D). On the other hand, modest changes in chromatin accessibility were observed in the other two cell types. In HEPs, the highest change was observed in cluster B (30 up- and 18 downregulated). In HSC, 62 and 7 peaks were

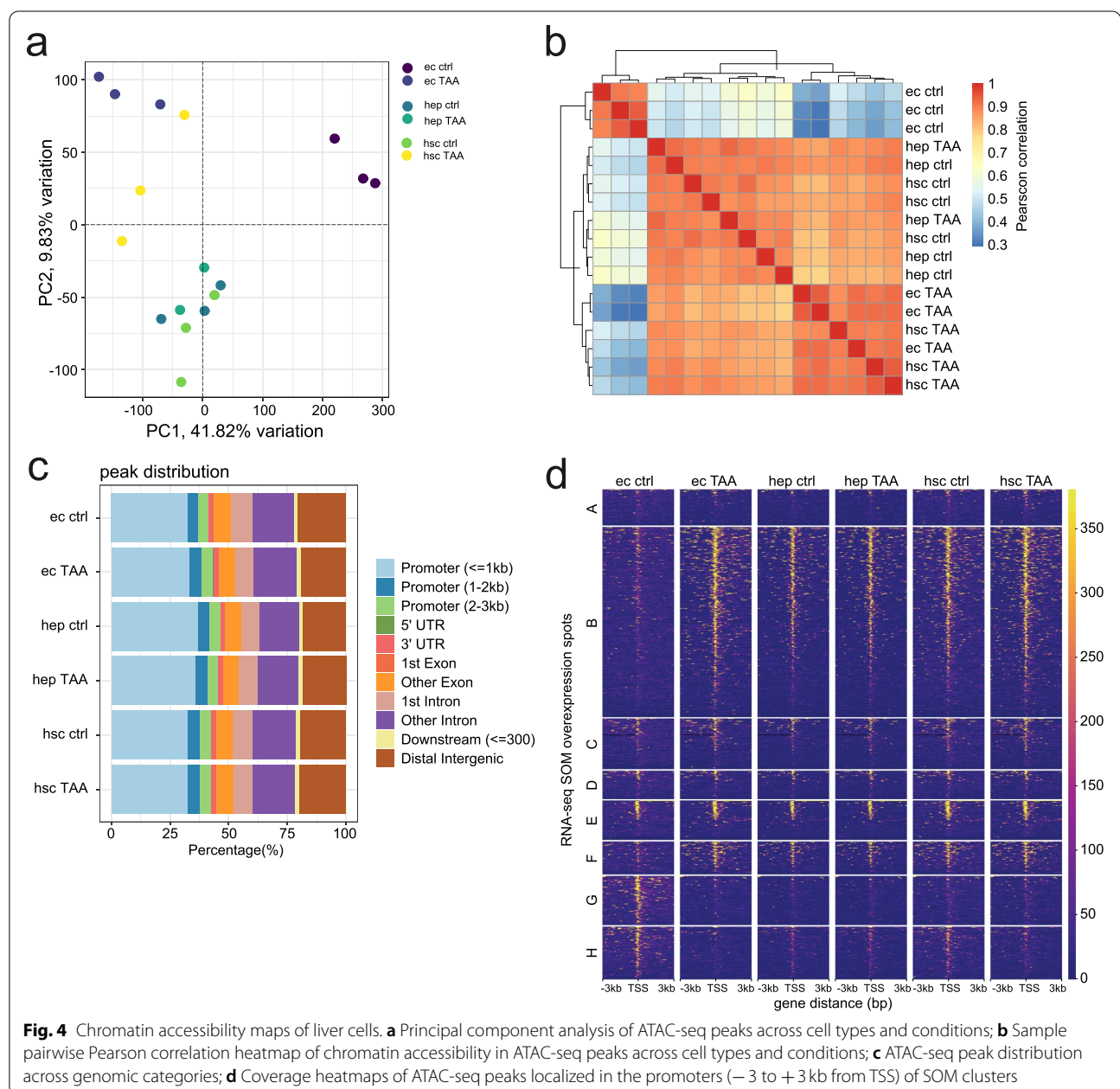
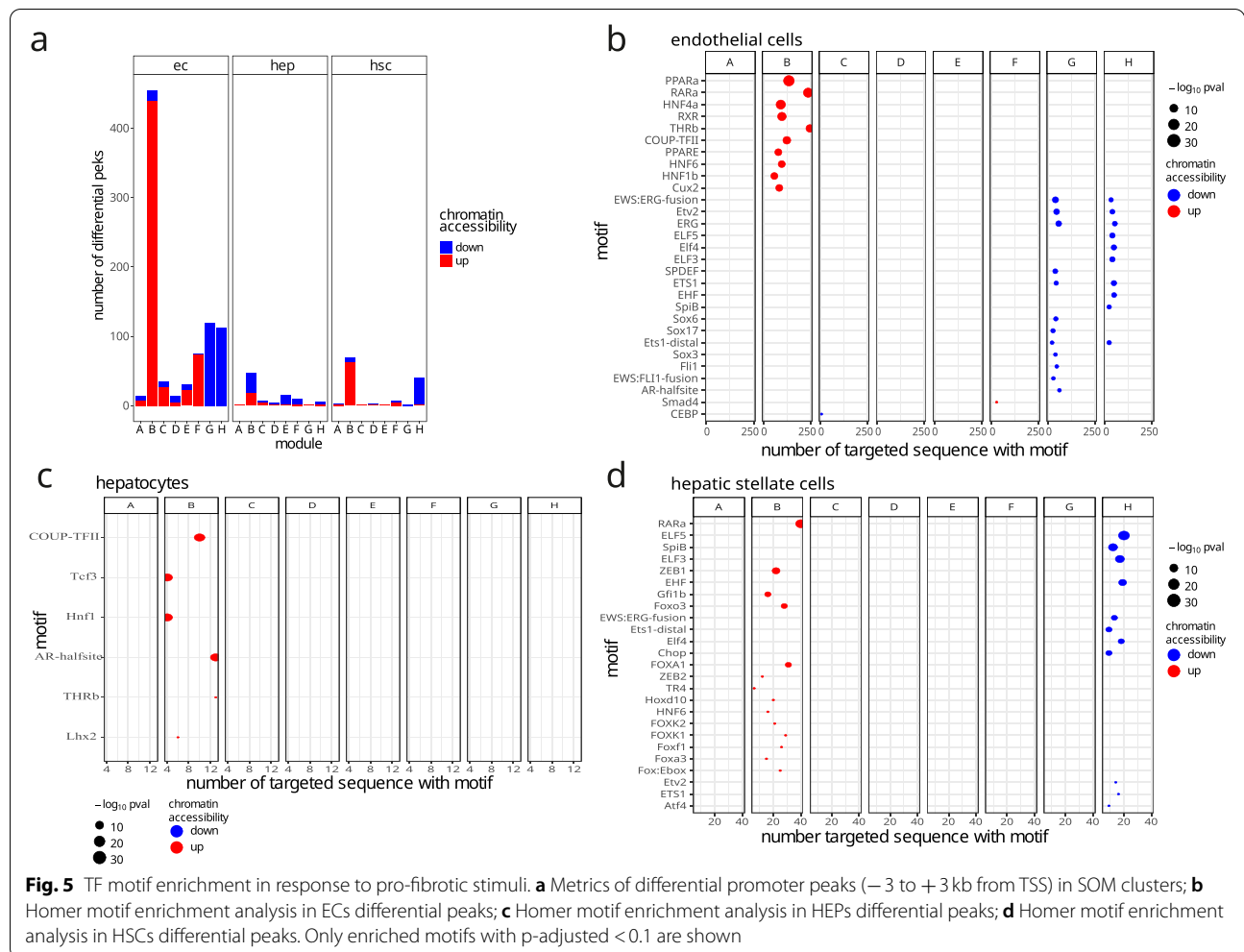


Fig. 4 Chromatin accessibility maps of liver cells. **a** Principal component analysis of ATAC-seq peaks across cell types and conditions; **b** Sample pairwise Pearson correlation heatmap of chromatin accessibility in ATAC-seq peaks across cell types and conditions; **c** ATAC-seq peak distribution across genomic categories; **d** Coverage heatmaps of ATAC-seq peaks localized in the promoters (−3 to +3 kb from TSS) of SOM clusters



upregulated or downregulated in cluster B, respectively and 39 downregulated in cluster H.

To identify potential regulators involved in TAA response in each cell type, we searched for transcription factor (TF) motifs enriched in differentially accessible promoter peaks from SOM cluster genes (Fig. 5B-D, Supp. Table 6). Significant enrichments (p-adjusted < 0.05, Supp. Table 6) were identified predominantly in five tested groups of regions: cluster B upregulated regions in ECs and HSCs, cluster G downregulated regions in ECs and cluster H downregulated regions in ECs and HSCs. In ECs, we observed significant enrichment in motifs of fatty acid metabolism nuclear receptors such as RXR [41], THRb [42], HNF4A [43] and PPARa [41] among peaks upregulated in cluster B. This is in accordance with the result of gene set overrepresentation analysis (Fig. 3D). A drop in chromatin accessibility was observed for ECs peaks located in the promoter of genes from cluster G. TFs motifs identified in this

cluster belong to ETS family (ETV2, ERG, SPDEF, ETS1) and Sox family (Sox6, Sox17, Sox3) involved in cell differentiation, migration and proliferation [44–46]. In HSCs, we found enriched motifs of TFs involved in cellular glucose homeostasis such as FOXA3 [47], FOXK1 [48], FOXK2 [49] and cell differentiation such as RARA, TR4, FOXA1, FOXA3 [50]. In cluster H downregulated regions, both in EC and HSC, we also found enriched motifs of ETS family including ETV2, ERG, ELF5, ELF3, ETS1, EHF, SPIB, ELF4. Additionally, in HSCs we found enrichment of ATF4 and Chop motifs, which are known to be involved in response to endoplasmic reticulum stress [51, 52]. Notably, ETS TFs also regulate endothelial function and homeostasis [53]. Altogether, our results show increased chromatin accessibility in the promoter regions of gene clusters associated with fatty acid metabolism, especially in ECs, and decrease of accessibility in clusters related to endothelial homeostasis and inflammatory response.

cluster A (Fig. 4D and Supp. Fig. 3A). Among the 5 genes within the top 25th percentile of accessibility changes and lower 25th percentile of read counts in control were *col4a6* and *elovl1a* (Supp. Fig. 4B, D, E).

Discussion

Liver fibrosis is a wound-healing response to tissue injury and inflammation hallmarked by the ECM protein deposition in the liver parenchyma and tissue remodelling [57]. The predominant causes of liver fibrosis are chronic excessive alcohol consumption, viral hepatitis B and C and non-alcoholic fatty liver disease (NAFLD), the latter becoming a major concern with the increasing incidence of obesity in Europe and the USA [1]. While these conditions have been widely studied [1], current knowledge of gene regulatory networks and epigenetic hallmarks associated with the early response to hepatotoxic injury is still lacking. It is crucial to study these primary changes in the cell types most affected by injury to improve the tools for diagnosis of early liver fibrosis and related disorders. In order to dissect the molecular mechanisms regulating the initiation of hepatic fibrosis and understand the interplay between genetic and epigenetic signals in this process, we utilized the model of TAA-induced liver injury in adult zebrafish and characterized cell-type-specific changes at both transcriptome and epigenome level in three main liver cell types: HEPs, HSCs and ECs.

The conservation of many metabolic pathways across vertebrate species renders the zebrafish a potent model organism in drug discovery studies. It has been extensively used to study liver development and injury [58, 59], and has been especially useful in establishing various toxicity models [60]. Many xenobiotics used to establish murine models of drug-induced liver injury have been found to be as effective in zebrafish, with an added advantage of the larvae being suitable for toxicological studies at 3 days post-fertilization, when mature liver parenchyma can be observed [60]. While the zebrafish liver architecture is distinct from its mammalian counterpart, the morphology, localization and gene expression profiles of HEPs, ECs and HSCs are similar [58, 60, 61].

The hepatotoxic properties of TAA in mice and rats induces oxidative stress resulting first in formation of intracellular lipid deposits in the liver parenchymal cells (hepatocyte ballooning), and later leading to HEPs damage and necrosis [62]. Bioactivation of TAA into its hepatotoxic counterpart, TASO₂ [63], requires proteins from the cytochrome p450 complex, functional orthologs for many of which exist in zebrafish, including proteins with >44.87% sequence similarity to CYP2E1, the protein thought to be directly responsible for TAA metabolism in humans [64]. Moreover, CYP2E1 function was

reproduced in zebrafish tissue homogenates, albeit without identifying the specific enzyme responsible for the process [65].

In line with previous reports [5, 66], we observed that gene expression profiles of respective cell populations are similar to those exhibited by their mammalian counterparts. Specifically, our sorted cell populations were enriched for known cell specific markers and relevant GO terms. These results are in agreement with the established existence of conserved molecular pathways between species [58]. Moreover, our analysis of cell-type-specific transcriptional response to TAA treatment highlighted known molecular components of the TAA metabolism pathway such as elements of the cytochrome p450 superfamily (Supp. Table 3). The most striking transcriptional response to TAA was observed in the ECs, highlighting those cells as the most affected by the treatment. This is likely a consequence of high permeability of ECs and also reflects their driving role in hepatotoxic injury response [67]. ECs, particularly LSEC, due to their exceptional permeability and intimate contact with the blood stream [68], are at the frontline of the toxic stimuli sensing. During liver damage, endothelial dysfunction occurs at early phases, before fibrosis initiation [69–71], under many liver etiologies such as non-alcoholic fatty liver disease (NAFLD) and alcoholic liver damage. Some evidence shows that LSEC dysfunction occurs before other liver injury early markers including Kupffer cell activation, nitric oxide content reduction or TNF α , IL-6 and ICAM-1 up-regulation [67, 70, 72]. To accompany their high toxins susceptibility ECs play a regulatory role in the liver cellular response to an injuring factor [67]. The main target of this regulation are the hepatic stellate cells (HSC), but evidence was shown on ECs involvement in control of HEPs proliferation [73]. In chronic models of liver injury, ECs, specifically LSEC, can generate a strong immune response and became highly proinflammatory, while secreting a vast range of cytokines and chemokines including TNF- α , IL-6, IL-1, CCL2 [67]. In response to those stimuli as well as the damaging toxin, other cells co-participate in the liver cellular response regulation. Injured hepatocytes and inflammatory cells secrete inflammatory mediators, which further stimulate LSEC and the inflammatory response.

To assess TAA-induced transcriptional changes in more detail, we applied SOM to identify clusters of co-expressed genes in our transcriptome data. We found eight clusters that showed greatest variability between conditions. The largest of these, cluster B, showed highest upregulation in response to TAA treatment in ECs. Interestingly, this cluster consists of genes related to metabolic and redox processes, including 20 members of the cytochrome p450 superfamily. This suggests that cluster

B represents the set of genes most directly responding to TAA treatment. The expression of CYP2E1 in LSECs was recently reported in the case of alcohol induced liver injury in mice [74]. Moreover, in agreement with the ability of ECs to regulate neighboring cells, eg. via angiocrine factors, we found many genes whose products are known to localize in the extracellular space in cluster B. This includes Apolipoprotein A-IV which has been recently identified as a potent liver fibrosis biomarker [54]. Conversely, clusters G and H showed strong downregulation upon TAA treatment. Of these, genes involved in extracellular structure organisation (cluster G) showed the strongest response in the ECs, while genes involved in immune response (cluster H) were commonly downregulated across all cell types. Contrary to previous reports [75, 76], we did not observe an upregulation of extracellular space-associated genes, especially matrix metalloproteinase genes (clusters A and C) in HEPs. This may be due to the differences in experimental design, as in contrast to the cited studies we investigated the earliest stages of liver injury. Other possible sources of divergent results may be the choice of hepatotoxin, as both cited studies employed CCl₄. This result could also highlight the differences in model organisms of choice, as the cited studies have employed mice, rats and human cell lines.

The observed gene expression upregulation in response to treatment is accompanied by increased promoter accessibility. In agreement with RNA-seq data, we observe the largest chromatin rearrangements in ECs. This result suggests that chromatin remodeling is an important mechanism driving gene expression response to liver injury. Indeed, our motif enrichment analysis identified known motifs of transcriptional activators, such as the pioneer factors *foxa1* and *foxa3*, to be enriched in the regions of increased accessibility. Curiously, the murine homolog of *foxa3* has been implicated in promoting liver regeneration [77], while *foxa1* is important for proper liver parenchyma development [78]. Changes in promoter accessibility in other cell types were less prominent, however the increase in chromatin accessibility was observed in HSCs' *col4a6* promoter region upon TAA treatment. This, taken together with the increased transcription of ECM genes in both ECs and HSCs can suggest that the initiation of ECM remodeling driven by both these cell types is triggered by hepatic injury.

Conclusions

We induced liver injury using TAA, an established potent hepatotoxin, in adult zebrafish. Using this system, we identified cell-type specific response to early hepatotoxic liver injury at the transcriptomic and regulatory level. We demonstrated that in zebrafish,

the first major liver cell population exposed to hepatotoxin - ECs - is also the most affected at both transcriptomic and chromatin accessibility level at this stage of liver injury. Importantly, genes known to be key players in ECM remodelling as well as metabolic and redox processes were observed to be responsive to TAA-mediated liver injury, including some which undergo chromatin re-arrangement at their promoter regions. Besides revealing the global transcriptome and epigenome landscape of early liver injury, this work provides insight into the molecular processes involved in early stages of liver damage. It also promises the viability of employing approaches providing even more specific, in-depth information, such as single cell sequencing or long read sequencing. These could potentially allow researchers to identify subpopulations of cells within major cell types that are responsible for distinct signals and injury response patterns, or assess transcript modifications triggered by early liver injury.

Methods

TAA dose-response assessment

Treatment of adult zebrafish individuals with TAA at a concentration of 300 mg/kg b.m. which was previously reported for female zebrafish [20] did not result in morphological changes compared to saline-injected controls (Supp. Fig. 2), thus suggesting that a higher concentration of TAA is required to induce liver injury in adult fish. In order to establish the optimal TAA concentration for adult zebrafish, we first performed a range-finding experiment to identify the working dose for zebrafish embryos, which we would then use as a guideline for establishing the higher dose in adults. By performing the toxicity assay in embryos instead of adults we bypassed the need to sacrifice large numbers of animals. Embryos at 48 hpf ($n=18$ for each concentration) were placed individually in 12-well plates. 5 concentrations were tested: 150 mg/l, 375 mg/l, 750 mg/l, 1500 mg/l and 3750 mg/l. The TAA solution was changed every 24h for 72h, at which point the embryo survival was estimated. A control group for each concentration was kept in E3 medium (5 mM NaCl, 0.17 mM KCl, 0.33 mM CaCl₂, 0.33 mM MgSO₄) and changed every 24h for the duration of the experiment. We found that treatment of embryos with 1500 mg/l of TAA for 72h resulted in ~50% mortality, thereby approximating the embryonic LC50 for TAA at this concentration. To ensure an adequate amount of TAA delivered to the adult liver, we adopted the intraperitoneal injection strategy repeated 6 times over the span of 2 weeks, with a dose of 500 mg/kg of body mass per injection.

TAA administration and isolation of liver cell populations by fluorescence-activated cell sorting (FACS)

Zebrafish transgenic lines *Tg(fabp10a:dsRed)*, *Tg(hand2:EGFP)* and *Tg(kdrl:ras-mCherry)* in AB wild-type background were maintained in the IIMCB zebrafish facility (License no. PL14656251) according to standard procedures. Adult females were anesthetized with MS-222 (Sigma-Aldrich, Germany) as previously described [79] and injected intraperitoneally with 500 mg/kg thioacetamide (TAA) or sterile water as a control 6 times over the course of 2 weeks. A single dose of TAA would not approach the estimated LC50 for embryos, but the overall exposure to the toxin would exceed the estimated LC50. Adult fish weighing less than 2 g prior to the injections were excluded due to welfare concerns. Prior to toxin administration, the injection spot was wiped down with 1% povidone iodine to further limit the risk of infection. Overall, 15 fishes were injected with TAA. An additional 6 were injected with saline as a control. Fishes injected with TAA survived to the end of the 2-week treatment with 20% mortality (n surviving = 12). All saline-injected fishes survived the procedure. Experimental protocol for the treatment of animals in this study follows the guidelines approved by First Warsaw Local Ethics Committee for Animal Experimentation (file 15/2015). Livers were dissected and digested in Hank's solution (1× HBSS, 2 mg/mL BSA, 10 mM Hepes pH 8.0) containing 0.05% trypsin (Sigma-Aldrich, Germany) and 2% collagenase (Sigma-Aldrich, Germany). Cell suspension was centrifuged at 500 g for 10 min at 4 °C. Cell pellet was resuspended in FACSmax (Amsbio, UK) and passed through a sterile 0.22 μm cell strainer (VWR, USA). Fluorescent cells were sorted by using FACSaria II cytometer (BD Biosciences, USA).

RNA-seq

For RNA sequencing 100,000 fluorescent liver cells were sorted directly to TRIzol LS (Thermo Fisher Scientific, USA). After ethanol precipitation RNA was depleted of DNA by using DNase I treatment and purified on columns by using RNA Clean & ConcentratorTM-5 (Zymo Research, USA). RNA integrity was measured by RNA ScreenTape on the Agilent 2200 TapeStation system (Agilent Technologies, USA). RNA Integrity Number (RIN) was in the range from 8.5 to 10 for all the samples used for RNA-seq. Ribosomal RNA removal from 10 ng of total RNA was performed using RiboGone Kit (Clontech Laboratories, USA). cDNA synthesis for next-generation sequencing (NGS) was performed by SMARTer Universal Low Input RNA Kit (Clontech Laboratories, USA) as recommended by the manufacturer. DNA libraries were purified with Agencourt AMPure XP PCR purification beads (Beckman Coulter, USA) and DNA fragment

distribution was assessed by using D1000 ScreenTape and Agilent 2200 TapeStation system (Agilent Technologies, USA). KAPA library quantification kit (Kapa Biosystems, USA) was used for qPCR-based quantification of the libraries obtained. Paired-end sequencing (2 × 75 bp reads) was performed with NextSeq 500 sequencing system (Illumina, USA).

ATAC-seq

For ATAC-seq 60,000 fluorescent liver cells were sorted to Hank's solution (1× HBSS, 2 mg/mL BSA, 10 mM Hepes pH 8.0), centrifuged for 5 min at 500×g and prepared for chromatin tagmentation as previously described [80]. NEBNext High-Fidelity 2 × PCR Master Mix (New England Biolabs, USA) and custom HPLC-purified primers containing Illumina-compatible indexes were used to prepare DNA sequencing libraries as previously described [81]. DNA libraries were purified with Agencourt AMPure XP PCR purification beads (Beckman Coulter, USA) and DNA fragment distribution was assessed by using D1000 ScreenTape and Agilent 2200 TapeStation system (Agilent Technologies, USA). KAPA library quantification kit (Kapa Biosystems, USA) was used for qPCR-based quantification of the libraries obtained. Paired-end sequencing (2 × 75 bp reads) was performed with NextSeq500 sequencing system (Illumina, USA).

Bioinformatics analysis

Raw RNA-seq and ATAC-seq reads were quality checked using Fastqc (0.11.8). Adapters were removed using Cutadapt (1.18) [82]. RNA-seq reads matching ribosomal RNA were removed using rRNA dust [83] and remaining reads were aligned to the zebrafish reference genome (GRCz11) using STAR (2.6) [84]. ATAC-seq reads were aligned to the zebrafish reference genome (GRCz11) using Bowtie2 (2.3.4.3) [85]. Reads quality filtering was performed using SAMtools (1.9) [86]. Read and alignment quality reports were prepared in Multiqc (1.6). To identify nucleosome free regions (NFRs) ATAC-seq reads originating from fragments not longer than 128 bp were retained and shifted by +4 / -5 bp depending on the alignment strand using alignmentSieve utility from deepTools suite (3.2.0) [87]. Those reads were further used for peak calling using Macs2 (2.1.0.2) [88] subcommands. Shortly for each of the three replicates per base enrichment p -value track was calculated using the Poisson test. Then p -values tracks from replicates were combined using Fisher method. After Benjamini - Hochberg multiple testing correction, peaks were called on obtained tracks with q -value cutoff of $1e-5$. Further obtained BED files were manipulated using Bedtools (2.27.1) [89] to discard NFRs overlapping low complexity regions as defined

in the Ensembl's [90] reference genome (GRCz11). Enriched motifs in NFRs were identified using Homer (4.10) [91]. Downstream bioinformatics analysis were performed in R 3.4.4 using several Bioconductor [92] packages. Cell type specific genes at quiescent state, were identified using DESeq2 [93] by comparing gene expression in specific cell type with gene expression in the other two. High-dimensional portraying of gene expression profiles was performed using oposSOM [36]. Differential gene expression analysis and differential accessibility analysis was performed using DESeq2 [93]. ATAC-seq peaks were processed and visualized using ChIPseeker [94], clusterProfiler [95], rtracklayer [96] and Gviz [97].

Histology and fluorescent microscopy

Adult females were sacrificed by overdosing MS-222 (Sigma-Aldrich, Germany) as previously described [98]. Samples were fixed in Dietrich's fixative [98], dehydrated in ethanol and embedded in JB-4 resin (Sigma-Aldrich, Germany) for 3 h at 4°C. Liver histology was examined microscopically in sections (4 µm thick) after hematoxylin and eosin (Sigma-Aldrich, Germany) staining using a modified protocol with increased staining and wash times to account for the lower staining efficiency in JB-4 resin. To detect fluorescence of GFP, mCherry and RFP, livers were fixed in 4% formaldehyde, incubated overnight in 20% sucrose, frozen in OCT solution (Leica Biosystems, France) and viewed under fluorescence microscope after sectioning (section thickness = 15 µm).

Abbreviations

ECM: Extracellular matrix; TAA: Thioacetamide; HEP: Hepatocyte; HCC: Hepatocellular carcinoma; NAFLD: Non-alcoholic fatty liver disease; HSC: Hepatic stellate cell; LSEC: Liver sinusoidal endothelial cell; EC: Endothelial cell; SOM: Self-organising map; FACS: Fluorescence-activated cell sorting; GO: Gene ontology; TF: Transcription factor.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-08173-1>.

Additional file 1.

Additional file 2.

Acknowledgments

We acknowledge the IIMCB Zebrafish Core Facility for service and fish maintenance. We thank E. Ober for the kind gift of *Tg(fabp10a:dsRed)*.

Authors' contributions

MP and ET performed in vivo experiments and collected biological material. KZK performed preliminary experiments and optimized the protocols. LB performed and KP supervised FACS analysis. ET performed histological staining and took microscopic images. KAN prepared NGS libraries and performed sequencing. MP and MM performed bioinformatics and statistical analysis. MP and MM contributed to the design of the study and interpreted data. FG analyzed and interpreted the data. MP, MM and ET prepared the figures. MP and CLW conceived the study. MP, ET, MM and CLW wrote the manuscript. MP and

CLW are senior corresponding authors. All authors have read and approved the manuscript.

Funding

This work has been supported by National Science Centre, Poland, SONATA grant number 2014/15/D/NZ5/03421. MP was supported by the Ministry of Science and Higher Education, Poland, and National Science Centre, Poland, OPUS grant number 2018/29/B/NZ2/01010 and Foundation For Polish Science TEAM grant number POIR.04.04.00-00-5C84/17. FG was supported by Polish National Science Centre grants 2018/31/N/NZ5/03214 and 2020/36/T/NZ5/00610. ET and MM are recipients of the Postgraduate School of Molecular Medicine doctoral fellowship for the program "Next generation sequencing technologies in biomedicine and personalized medicine". The project no. POIR.04.04.00-00-1AF0/16-00/ carried out within the First TEAM programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund supports CW and KAN.

Availability of data and materials

RNA-seq and ATAC-seq data have been submitted to the NCBI Gene Expression Omnibus database (<https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE145565.

Declarations

Ethics approval and consent to participate

All experimental protocols were approved by First Warsaw Local Ethics Committee for Animal Experimentation (file 15/2015). All methods were carried out in accordance with relevant guidelines and regulations and reported in accordance with ARRIVE guidelines for the reporting of animal experiments.

Consent for publication

Not applicable.

Competing interests

Authors declare no conflict of interest.

Author details

¹International Institute of Molecular and Cell Biology in Warsaw, Laboratory of Zebrafish Developmental Genomics, 4 Ks. Trojdena Street, 02-109 Warsaw, Poland. ²Nencki Institute of Experimental Biology, Laboratory of Cytometry, Warsaw, Poland. ³Department of Experimental Hematology, Institute of Hematology and Transfusion Medicine, ul. Indrzej Gandhi 14, 02-776 Warsaw, Poland. ⁴Department of Oncologic Pathology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA.

Received: 9 July 2021 Accepted: 10 November 2021

Published online: 18 December 2021

References

- Blachier M, Leleu H, Peck-Radosavljevic M, Valla D-C, Roudot-Thoraval F. The burden of liver disease in Europe: a review of available epidemiological data. *J Hepatol*. 2013;58:593–608.
- Baranova A, Lal P, Biredinc A, Younossi ZM. Non-invasive markers for hepatic fibrosis. *BMC Gastroenterol*. 2011;11:91.
- Tu T, Calabro SR, Lee A, Maczurek AE, Budzinska MA, Warner FJ, et al. Hepatocytes in liver injury: victim, bystander, or accomplice in progressive fibrosis? *J Gastroenterol Hepatol*. 2015;30:1696–704.
- Baratta JL, Ngo A, Lopez B, Kasabwalla N, Longmuir KJ, Robertson RT. Cellular organization of normal mouse liver: a histological, quantitative immunocytochemical, and fine structural analysis. *Histochem Cell Biol*. 2009;131:713–26.
- Yin C, Evason KJ, Asahina K, Stainier DYR. Hepatic stellate cells in liver development, regeneration, and cancer. *J Clin Invest*. 2013;123:1902–10.
- Gandhi CR. Hepatic stellate cell activation and pro-fibrogenic signals. *J Hepatol*. 2017;67:1104–5.
- Poisson J, Lemoine S, Boulanger C, Durand F, Moreau R, Valla D, et al. Liver sinusoidal endothelial cells: physiology and role in liver diseases. *J Hepatol*. 2017;66:212–27.

8. Lefebvre P, Lalloyer F, Baugé E, Pawlak M, Gheeraert C, Dehondt H, et al. Interspecies NASH disease activity whole-genome profiling identifies a fibrogenic role of PPAR α -regulated dermatopontin. *JCI Insight*. 2017;2:e92264.
9. Tsuchida T, Friedman SL. Mechanisms of hepatic stellate cell activation. *Nat Rev Gastroenterol Hepatol*. 2017;14:397–411.
10. Iredale JP. Models of liver fibrosis: exploring the dynamic nature of inflammation and repair in a solid organ. *J Clin Invest*. 2007;117:539–48.
11. Sapp V, Gaffney L, EauClaire SF, Matthews RP. Fructose leads to hepatic steatosis in zebrafish that is reversed by mechanistic target of rapamycin (mTOR) inhibition. *Hepatology*. 2014;60:1581–92.
12. Sadler KC, Amsterdam A, Soroka C, Boyer J, Hopkins N. A genetic screen in zebrafish identifies the mutants vps18, nf2 and foie gras as models of liver disease. *Development*. 2005;132:3561–72.
13. Amali AA, Rekha RD, Lin CJ-F, Wang W-L, Gong H-Y, Her G-M, et al. Thioacetamide induced liver damage in zebrafish embryo as a disease model for steatohepatitis. *J Biomed Sci*. 2006;13:225–32.
14. Yin C, Evason KJ, Maher JJ, Stainer D. The basic helix-loop-helix transcription factor, heart and neural crest derivatives expressed transcript 2, marks hepatic stellate cells in zebrafish: analysis of stellate cell entry into the developing liver. *Hepatology*. 2012;56:1958–70.
15. Langheinrich U. Zebrafish: a new model on the pharmaceutical catwalk. *BioEssays News Rev Mol Cell Dev Biol*. 2003;25:904–12.
16. Ramaiah SK, Apte U, Mehendrale HM. Cytochrome P450E1 induction increases thioacetamide liver injury in diet-restricted rats. *Drug Metab Dispos Biol Fate Chem*. 2001;29:1088–95.
17. Hajovsky L, Hu G, Koen Y, Sarma D, Cui W, Moore DS, et al. Metabolism and toxicity of thioacetamide and thioacetamide s-oxide in rat hepatocytes. *Chem Res Toxicol*. 2012;25:1955–63.
18. Akhtar T, Sheikh N. An overview of thioacetamide-induced hepatotoxicity. *Toxin Rev*. 2013;32:43–6.
19. Wallace M, Hamesch K, Lunova M, Kim Y, Weiskirchen R, Strnad P, et al. Standard operating procedures in experimental liver research: thioacetamide model in mice and rats. *Lab Anim*. 2015;49(1_suppl):21–9.
20. Rekha RD, Amali AA, Her GM, Yeh YH, Gong H-Y, Hu S-Y, et al. Thioacetamide accelerates steatohepatitis, cirrhosis and HCC by expressing HCV core protein in transgenic zebrafish *Danio rerio*. *Toxicology*. 2008;243:11–22.
21. Wirth H, von Bergen M, Binder H. Mining SOM expression portraits: feature selection and integrating concepts of molecular function. *BioData Min*. 2012;5:18.
22. Her GM, Chiang C-C, Chen W-Y, Wu J-L. In vivo studies of liver-type fatty acid binding protein (L-FABP) gene expression in liver of transgenic zebrafish (*Danio rerio*). *FEBS Lett*. 2003;538:125–33.
23. Chi NC, Shaw RM, De Val S, Kang G, Jan LY, Black BL, et al. Foxn4 directly regulates *tbx2b* expression and atrioventricular canal formation. *Genes Dev*. 2008;22:734–9.
24. Yao Y, Yao J, Boström KI. SOX transcription factors in endothelial differentiation and endothelial-mesenchymal transitions. *Front Cardiovasc Med*. 2019;6. <https://doi.org/10.3389/fcvm.2019.00030>.
25. Goncharov NV, Nadeev AD, Jenkins RO, Avdonin PV. Markers and biomarkers of endothelium: when something is rotten in the state. *Oxidative Med Cell Longev*. 2017;2017:e9759735.
26. Shay S, Ahuva I, Shira N-Y, Caryn G, Debra G-W, Simcha Y, et al. A novel human-specific soluble vascular endothelial growth factor receptor 1. *Circ Res*. 2008;102:1566–74.
27. Marshall SM, Gromovsky AD, Kelley KL, Davis MA, Wilson MD, Lee RG, et al. Acute Sterol O-Acyltransferase 2 (SOAT2) knockdown rapidly mobilizes hepatic cholesterol for fecal excretion. *PLoS One*. 2014;9:e98953.
28. Jayakumar A, Tai MH, Huang WY, al-Feel W, Hsu M, Abu-Elheiga L, et al. Human fatty acid synthase: properties and molecular cloning. *Proc Natl Acad Sci U S A*. 1995;92:8695–9.
29. Yeh Y-S, Jheng H-F, Iwase M, Kim M, Mohri S, Kwon J, et al. The mevalonate pathway is indispensable for adipocyte survival. *iScience*. 2018;9:175–91.
30. Rokosz LL, Boulton DA, Butkiewicz EA, Sanyal G, Cueto MA, Lachance PA, et al. Human cytoplasmic 3-hydroxy-3-methylglutaryl coenzyme a synthase: expression, purification, and characterization of recombinant wild-type and Cys129 mutant enzymes. *Arch Biochem Biophys*. 1994;312:1–13.
31. Yao Y, Sun S, Wang J, Fei F, Dong Z, Ke A-W, et al. Canonical Wnt signaling remodels lipid metabolism in Zebrafish hepatocytes following Ras oncogenic insult. *Cancer Res*. 2018;78:5548–60.
32. Hautekeete M, Geerts A. Limited evidence for redistribution of vitamin A from the liver to oesophageal mucosa in chronic liver disease in humans. *Leiden: Cells Hepatic Sinusoid*; 1997. p. 54–7.
33. Albadri S, Naso F, Thauvin M, Gauron C, Parolin C, Duroure K, et al. Redox signaling via lipid peroxidation regulates retinal progenitor cell differentiation. *Dev Cell*. 2019;50:73–89.e6.
34. Park K-H, Kim S-H. Low dose of chronic ethanol exposure in adult zebrafish induces hepatic steatosis and injury. *Biomed Pharmacother Biomedecine Pharmacother*. 2019;117:109179.
35. Timme-Laragy AR, Goldstone JV, Imhoff BR, Stegeman JJ, Hahn ME, Hansen JM. Glutathione redox dynamics and expression of glutathione-related genes in the developing embryo. *Free Radic Biol Med*. 2013;65. <https://doi.org/10.1016/j.freeradbiomed.2013.06.011>.
36. Löffler-Wirth H, Kalcher M, Binder H. oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on bioconductor. *Bioinforma Oxf Engl*. 2015;31:3225–7.
37. Wirth H, Löffler M, von Bergen M, Binder H. Expression cartography of human tissues using self organizing maps. *BMC Bioinformatics*. 2011;12:306.
38. Moran-Salvador E, Mann J. Epigenetics and liver fibrosis. *Cell Mol Gastroenterol Hepatol*. 2017;4:125–34.
39. Leung A, Parks BW, Du J, Trac C, Setten R, Chen Y, et al. Open chromatin profiling in mice livers reveals unique chromatin variations induced by high fat diet *. *J Biol Chem*. 2014;289:23557–67.
40. Wang AW, Wang YJ, Zahm AM, Morgan AR, Wangenstein KJ, Kaestner KH. The dynamic chromatin architecture of the regenerating liver. *Cell Mol Gastroenterol Hepatol*. 2020;9:121–43.
41. Kersten S. Peroxisome proliferator activated receptors and lipoprotein metabolism. *PPAR Res*. 2008;2008. <https://doi.org/10.1155/2008/132960>.
42. Pramfalk C, Pedrelli M, Parini P. Role of thyroid receptor β in lipid metabolism. *Biochim Biophys Acta (BBA) - Mol Basis Dis*. 1812;2011:929–37.
43. Reddy S, Yang W, Taylor DG, Shen X, Oxender D, Kust G, et al. Mitogen-activated protein kinase regulates transcription of the ApoCIII gene. Involvement of the orphan nuclear receptor HNF4. *J Biol Chem*. 1999;274:33050–6.
44. Sarkar A, Hochedlinger K. The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell Stem Cell*. 2013;12:15–30.
45. Oikawa T, Yamada T. Molecular biology of the Ets family of transcription factors. *Gene*. 2003;303:11–34.
46. Remy P, Baltzinger M. The Ets-transcription factor family in embryonic development: lessons from the amphibian and bird. *Oncogene*. 2000;19:6417–31.
47. Lin B, Morris DW, Chou JY. The role of HNF1 α , HNF3 γ , and cyclic AMP in glucose-6-phosphatase gene activation. *Biochemistry*. 1997;36:14096–106.
48. He L, Gomes AP, Wang X, Yoon SO, Lee G, Nagiec MJ, et al. mTORC1 promotes metabolic reprogramming by the suppression of GSK3-dependent Foxk1 phosphorylation. *Mol Cell*. 2018;70:949–960.e4.
49. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49:D480–9.
50. Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Brief Bioinform*. 2011;12:449–62.
51. Chami M, Oulès B, Szabadkai G, Tacine R, Rizzuto R, Paterlini-Bréchet P. Role of SERCA1 truncated isoform in the proapoptotic calcium transfer from ER to mitochondria during ER stress. *Mol Cell*. 2008;32:641–51.
52. Su N, Kilberg MS. C/EBP homology protein (CHOP) interacts with activating transcription factor 4 (ATF4) and negatively regulates the stress-dependent induction of the asparagine synthetase gene. *J Biol Chem*. 2008;283:35106–17.
53. Shah AV, Birdsey GM, Randi AM. Regulation of endothelial homeostasis, vascular development and angiogenesis by the transcription factor ERG. *Vasc Pharmacol*. 2016;86:3–13.
54. Wang P-W, Hung Y-C, Wu T-H, Chen M-H, Yeh C-T, Pan T-L. Proteome-based identification of apolipoprotein A-IV as an early diagnostic biomarker in liver fibrosis. *Oncotarget*. 2017;8:88951–64.
55. Bracht T, Schweinsberg V, Trippler M, Kohl M, Ahrens M, Padden J, et al. Analysis of disease-associated protein expression using quantitative

- proteomics—Fibulin-5 is expressed in association with hepatic fibrosis. *J Proteome Res.* 2015;14:2278–86.
56. Oh S-Y, Kim JY, Park C. The ETS factor, ETV2: a master regulator for vascular endothelial cell development. *Mol Cell.* 2015;38:1029–36.
 57. Bataller R, Brenner DA. Liver fibrosis. *J Clin Invest.* 2005;115:209–18.
 58. Wilkins BJ, Pack M. Zebrafish models of human liver development and disease. *Compr Physiol.* 2013;3:1213–30.
 59. Kim S-H, Wu S-Y, Baek J-I, Choi SY, Su Y, Flynn CR, et al. A post-developmental genetic screen for Zebrafish models of inherited liver disease. *PLoS One.* 2015;10. <https://doi.org/10.1371/journal.pone.0125980>.
 60. Goessling W, Sadler KC. Zebrafish: an important tool for liver disease research. *Gastroenterology.* 2015;149:1361–77.
 61. Wrighton PJ, Oderberg IM, Goessling W. There is something fishy about liver cancer: Zebrafish models of hepatocellular carcinoma. *Cell Mol Gastroenterol Hepatol.* 2019;8:347–63.
 62. Hou W, Syn W-K. Role of metabolism in hepatic stellate cell activation and fibrogenesis. *Front Cell Dev Biol.* 2018;6. <https://doi.org/10.3389/fcell.2018.00150>.
 63. Mehendale HM, Chilakapati J. 9.29 - Thioacetamide. In: CA MQ, editor. *Comprehensive toxicology*. 2nd ed. Oxford: Elsevier; 2010. p. 627–38. <https://doi.org/10.1016/B978-0-08-046884-6.01029-0>.
 64. Pritchard MT, Apte U. Chapter 2 - models to study liver regeneration. In: Apte U, editor. *Liver regeneration*. Boston: Academic Press; 2015. p. 15–40. <https://doi.org/10.1016/B978-0-12-420128-6.00002-6>.
 65. Hartman JH, Kozal JS, Di Giulio RT, Meyer JN. Zebrafish have an ethanol-inducible hepatic 4-nitrophenol hydroxylase that is not CYP2E1-like. *Environ Toxicol Pharmacol.* 2017;54:142–5.
 66. Chu J, Sadler KC. A new school in liver development: lessons from Zebrafish. *Hepatol Baltim Md.* 2009;50:1656–63.
 67. Lafoz E, Ruart M, Anton A, Oncins A, Hernández-Gea V. The endothelium as a driver of liver fibrosis and regeneration. *Cells.* 2020;9. <https://doi.org/10.3390/cells9040929>.
 68. Braet F, Spector I, De Zanger R, Wisse E. A novel structure involved in the formation of liver endothelial cell fenestrae revealed by using the actin inhibitor misakinolide. *Proc Natl Acad Sci U S A.* 1998;95:13635–40.
 69. DeLeve LD, Wang X, Kanel GC, Atkinson RD, McCuskey RS. Prevention of hepatic fibrosis in a murine model of metabolic syndrome with nonalcoholic steatohepatitis. *Am J Pathol.* 2008;173:993–1001.
 70. Horn T, Christoffersen P, Henriksen JH. Alcoholic liver injury: defenestration in noncirrhotic livers—a scanning electron microscopic study. *Hepatol Baltim Md.* 1987;7:77–82.
 71. Pasarín M, La Mura V, Gracia-Sancho J, García-Calderó H, Rodríguez-Villarrupla A, García-Pagán JC, et al. Sinusoidal endothelial dysfunction precedes inflammation and fibrosis in a model of NAFLD. *PLoS One.* 2012;7:e32785.
 72. Tateya S, Rizzo NO, Handa P, Cheng AM, Morgan-Stevenson V, Daum G, et al. Endothelial NO/cGMP/AVSP signaling attenuates Kupffer cell activation and hepatic insulin resistance induced by high-fat feeding. *Diabetes.* 2011;60:2792–801.
 73. Greene AK, Wiener S, Puder M, Yoshida A, Shi B, Perez-Atayde AR, et al. Endothelial-directed hepatic regeneration after partial hepatectomy. *Ann Surg.* 2003;237:530–5.
 74. Yang Y, Sangwung P, Kondo R, Jung Y, McConnell MJ, Jeong J, et al. Alcohol-induced Hsp90 acetylation is a novel driver of liver sinusoidal endothelial dysfunction and alcohol-related liver disease. *J Hepatol.* 2021;75:377–86.
 75. Calabro SR, Maczurek AE, Morgan AJ, Tu T, Wen VW, Yee C, et al. Hepatocyte produced matrix metalloproteinases are regulated by CD147 in liver fibrogenesis. *PLoS One.* 2014;9:e90571.
 76. del Carmen García de León M, Montfort I, Tello Montes E, López Vancell R, Olivos García A, González Canto A, et al. Hepatocyte production of modulators of extracellular liver matrix in normal and cirrhotic rat liver. *Exp Mol Pathol.* 2006;80:97–108.
 77. Wangenstein KJ, Zhang S, Greenbaum LE, Kaestner KH. A genetic screen reveals Foxa3 and TNFR1 as key regulators of liver repopulation. *Genes Dev.* 2015;29:904–9.
 78. Le Lay J, Kaestner KH. The fox genes in the liver: from organogenesis to functional integration. *Physiol Rev.* 2010;90:1–22.
 79. Matthews M, Varga ZM. Anesthesia and euthanasia in zebrafish. *ILAR J.* 2012;53:192–204.
 80. Pawlak M, Kedzierska KZ, Migdal M, Nahia KA, Ramilowski JA, Bugajski L, et al. Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development. *Genome Res.* 2019;29:506–19.
 81. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol.* 2015;109:21.29.1–9.
 82. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal.* 2011;17:10–2.
 83. Hasegawa A, Daub C, Carninci P, Hayashizaki Y, Lassmann T. MOIRAI: a compact workflow system for CAGE analysis. *BMC Bioinformatics.* 2014;15:144.
 84. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21.
 85. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods.* 2012;9:357–9.
 86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinforma Oxf Engl.* 2009;25:2078–9.
 87. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* 2014;42:W187–91.
 88. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9:R137.
 89. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–2.
 90. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. *Nucleic Acids Res.* 2021;49:D884–91.
 91. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010;38:576–89.
 92. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods.* 2015;12:115–21.
 93. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.
 94. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics.* 2015;31:2382–3.
 95. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J Integr Biol.* 2012;16:284–7.
 96. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinforma Oxf Engl.* 2009;25:1841–2.
 97. Hahne F, Ivanek R. Visualizing genomic data using Gviz and bioconductor. *Methods Mol Biol Clifton NJ.* 2016;1418:335–51.
 98. Ellis JL, Yin C. Histological analyses of acute alcoholic liver injury in Zebrafish. *JoVE J Vis Exp.* 2017;(123):55630.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions




SOFTWARE

Open Access



xcore: an R package for inference of gene expression regulators

Maciej Migdał¹, Takahiro Arakawa², Satoshi Takizawa², Masaaki Furuno², Harukazu Suzuki², Erik Arner^{2,3}, Cecilia Lanny Winata^{1*}  and Bogumił Kaczkowski^{2,4*}

*Correspondence:
cwinata@iimcb.gov.pl;
b.kaczkowski@gmail.com

¹ Laboratory of Zebrafish Developmental Genomics, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

² RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

³ Present Address: GSK, Gunnels Wood Rd, Stevenage SG1 2NY, UK

⁴ Present Address: Data Sciences and Quantitative Biology, Discovery Sciences, AstraZeneca R&D, Cambridge, UK

Abstract

Background: Elucidating the Transcription Factors (TFs) that drive the gene expression changes in a given experiment is a common question asked by researchers. The existing methods rely on the predicted Transcription Factor Binding Site (TFBS) to model the changes in the motif activity. Such methods only work for TFs that have a motif and assume the TF binding profile is the same in all cell types.

Results: Given the wealth of the ChIP-seq data available for a wide range of the TFs in various cell types, we propose that gene expression modeling can be done using ChIP-seq “signatures” directly, effectively skipping the motif finding and TFBS prediction steps. We present *xcore*, an R package that allows TF activity modeling based on ChIP-seq signatures and the user’s gene expression data. We also provide *xcoredata* a companion data package that provides a collection of preprocessed ChIP-seq signatures. We demonstrate that *xcore* leads to biologically relevant predictions using transforming growth factor beta induced epithelial-mesenchymal transition time-courses, rinderpest infection time-courses, and embryonic stem cells differentiated to cardiomyocytes time-course profiled with Cap Analysis Gene Expression.

Conclusions: *xcore* provides a simple analytical framework for gene expression modeling using linear models that can be easily incorporated into differential expression analysis pipelines. Taking advantage of public ChIP-seq databases, *xcore* can identify meaningful molecular signatures and relevant ChIP-seq experiments.

Keywords: Gene expression, Gene regulation, Regression, Transcription factors, ChIP-seq

Background

Gene expression profiling is often performed to elucidate the transcriptional regulators in a given system/perturbation. A common approach is to use transcription factor motifs to computationally predict the TFBS within promoter regions of known genes. The “motif activity” is then inferred based on gene expression profiles [1–3]. Although such methods are quite simplistic, they proved useful for the identification of key molecular regulators [1, 2, 4, 5]. The limitations are that many TFs do not have a defined motif and some binding events may be specific to a particular biological context.



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

ReMap [6] and ChIP-Atlas [7] provide a wealth of uniformly processed ChIP-seq data (genome-wide peaks) for TFs but also other transcriptional regulators including transcriptional coactivators and chromatin-remodeling factors. Currently, only a limited number of tools exist that tap into these databases. Two examples are Lisa [8] identifying the most likely transcriptional regulators in an experiment based on user-supplied gene expression information, and Virtual ChIP-seq program [9] that can predict the binding of individual TF in a cell type of interest based on gene expression information. However, to our knowledge, there are no published methods that take advantage of this data to directly model the activity of transcriptional regulators.

Here, we propose to use the publicly available ChIP-seq data to directly represent the genome-wide occupancy of regulators. We intersected the peaks with promoter regions and used linear ridge regression to infer the regulators associated with observed gene expression changes (Fig. 1A). The advantage of this approach is the direct integration of gene expression profiles with experimental TF binding data. We provide (a) processed and pre-computed, ChIP-seq based molecular signatures (*xcoredata*), and (b) methodology for activity modeling (*xcore*). The framework is implemented as an R package (available in Bioconductor) and integrates smoothly with commonly used differential expression workflows like edgeR [10] or DESeq2 [11].

Implementation

Expression data processing

Xcore takes promoter or gene expression counts matrix as input, the data is then filtered for lowly expressed features, normalized for the library size and transformed into counts per million (CPM) using edgeR [10]. Users need to designate the base-level samples by providing an experiment design matrix. These samples are used as a baseline expression when modeling changes in gene expression. *xcore* implements promoter- and gene-level analyses, using either promoter or gene expression data. In our experience we found promoter-level analysis to provide better results (Additional file 1: Fig. S1). Cap Analysis Gene Expression (CAGE) data is an input of choice for promoter level analysis. However, *xcore* can be used with other types of expression data such as microarray or RNA-seq data to perform gene-level analysis. Promoter-level analysis based on RNA-seq data is possible in principle but currently not implemented.

Molecular signatures

A second input consists of molecular signatures describing known transcription factors' binding preferences within the promoter's vicinity. We provide sets of pre-computed molecular signatures with *xcoredata*, the accompanying data package. The signatures were obtained by downloading all ChIP-seq data from ReMap2020 [6] and ChIP-Atlas [7] and intersecting it against ± 500 nt window of known promoter regions, defined based on FANTOM5's hg38 annotation [12]. The signatures can also be easily constructed using *xcore* by providing predicted TFBS or custom ChIP-seq peaks (see

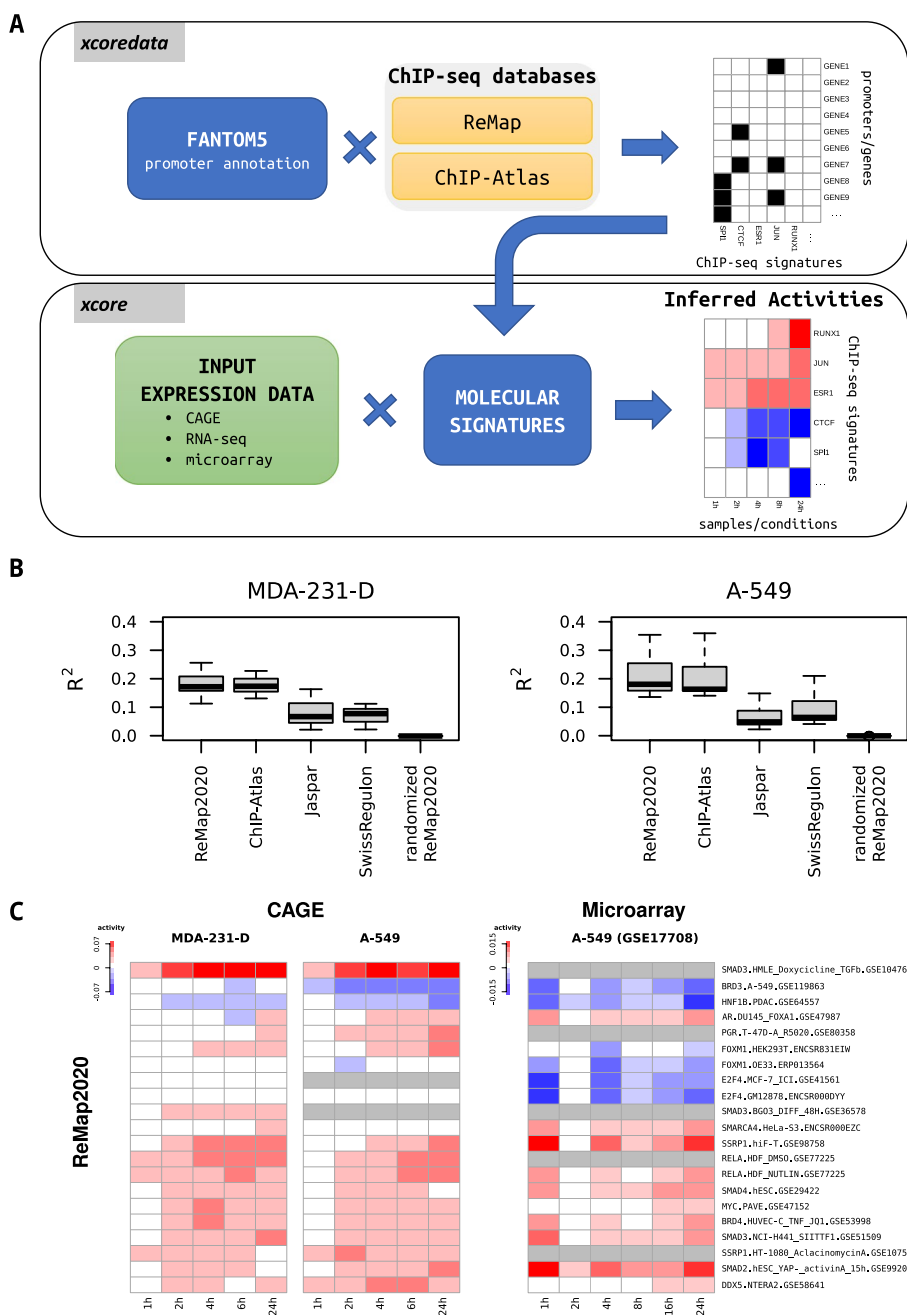


Fig. 1 Inferring transcription factors activities from gene expression during TGFβ induced EMT in A-549 and MDA-231-D cell lines. **A** Flowchart depicting *xcore* and *xcoredata* functionalities. **B** Boxplots showing R² values for gene expression prediction models constructed using different molecular signature sets: Motif-based (Jaspar, SwissRegulon) and ChIP-seq based (ReMap2020, ChIP-Atlas). Each boxplot shows R² values pooled across all the replicates. Models were trained and evaluated in tenfold cross-validation on individual replicates, using data on gene expression changes between 0 and 24 h after treatment in our newly generated TGFβ induced EMT experiment performed in A-549 and MDA-231-D cell lines. **C** Heatmap showing the dynamics of TF activities during TGFβ induced EMT. Heatmaps on the left present TF activities estimated using CAGE data from our newly generated TGFβ induced EMT experiment performed on A-549 and MDA-231-D cell lines. Heatmap on the right depicts TF activities estimated using previously published microarray data from the TGFβ induced EMT experiment performed on A-549 cell lines. The TF activities were calculated in the reference to 0 h time point. Only the top-scoring ReMap2020 signatures are shown. Grey color designates NA values

xcore user guide). Detailed information on the molecular signatures construction can be found in Extended Materials and Methods (Additional file 3).

Expression modeling

In *xcore* we describe the relationship between the expression (Y) and molecular signatures (X) using linear model formulation:

$$Y = \mu + \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

where Y is a sample expression level, μ is the basal expression level, β_0 is the intercept, β_j is a j -th molecular signature activity and X_j is a j -th molecular signature.

Here, we are interested in finding the unknown molecular signatures' activities (β) that describe the effect of molecular signature (X) on expression (Y). By including μ in the above equation we effectively model the change in expression between the basal expression level and the corresponding sample. Models are trained using penalized linear regression. In particular, we use ridge regression [13] as it allows us to take advantage of an existing significance testing methodology [14]. We observed ridge regression to work equally well to lasso and elastic net regression (Additional file 2: Fig. S2C). In practice, to fit our linear models we use the popular R package *glmnet* [15]. For each sample, that is for each time point and replicate, a separate model is trained using sample change in expression and molecular signatures shared at the experiment scale. In layman's terms, for each sample, we are seeking to find a combination of ChIP-seq based signatures that best explains the observed changes in gene expression. For each model, the ridge regression λ tuning parameter is found separately using the cross-validation technique (CV). By default tenfold CV is used, and λ value giving the smallest mean squared error is selected.

Next, the estimated molecular signatures' activities can be tested for significance. In short, using matrix formulation the ridge regression estimator is defined as

$$\hat{\beta}^\lambda = (X'X + \lambda I)^{-1} X'Y$$

where X is our molecular signatures matrix, λ is a ridge regression tuning parameter, and Y is a vector of our sample's changes in expression. Then, the estimate of β^λ standard error is calculated from the following:

$$\text{Var}(\hat{\beta}^\lambda) = \hat{\sigma}^2 (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1},$$

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta}^\lambda)'(Y - X\hat{\beta}^\lambda)}{\nu}$$

where ν is the residual effective degrees of freedom. The significance of the individual molecular signatures' activities can be then tested using a test of significance for ridge regression coefficients. For further details, we refer interested readers to [14].

To summarize the results from individual replicates, following the procedure described in [16], the obtained estimates and their standard errors are pooled across

the replicates by calculating their weighted mean with variance-defined weights and weighted mean error:

$$\bar{x} = \frac{\sum_{i=1}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}, \quad \sigma_{\bar{x}} = \sqrt{\frac{1}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}}$$

Using this result, we calculate a Z-score for each molecular signature and time-point.

Finally, molecular signatures are ranked based on their overall Z-score across the time-points calculated using Stouffer's Z method [17].

Linear regression models comparison

To compare different models, coefficients of determination (R^2) were calculated for models trained on individual replicates at selected time points using tenfold cross-validation and pooled across replicates. Additional information on this procedure is provided in Extended Materials and Methods (Additional file 3).

Results

We used *xcore* to perform gene expression modeling analysis in the context of three CAGE datasets: (a) newly generated transforming growth factor beta (TGF β) induced epithelial-mesenchymal transition (EMT) experiment performed in A-549 and MDA-231-D cell lines, (b) previously published FANTOM5's rinderpest infection time-course dataset performed in 293SLAM and COBL-a cell lines using native and recombinant rinderpest virus lacking accessory V and C proteins [12], (c) previously published FANTOM5's Human H3 embryonic stem cells differentiated to cardiomyocytes time-course dataset [12] and a microarray dataset: previously published TGF β induced EMT in A-549 cell line (GSE17708) [18]. Detailed information on the procedures used to process the raw CAGE data can be found in Extended Materials and Methods (Additional file 3).

ChIP-seq molecular signatures provides better model performance

We compared the models built using ChIP-seq signatures (ReMap2020 and ChIP-Atlas) vs motif-based signatures (Jaspar and SwissRegulon). The models based on ChIP-seq signatures showed on average higher R^2 values, which reflects the proportion of variance explained by the model and overall "goodness of fit". In particular, modeling expression between 0 and 24 h after TGF β treatment in our novel MDA-231-D dataset yielded an average R^2 of 0.179 for ChIP-seq signatures and 0.077 for motif signatures. For comparison the randomized version of ReMap2020 molecular signature yielded R^2 close to 0 (Fig. 1B, Additional file 2: Fig. S2B).

xcore recovers biologically relevant expression regulators

To investigate the biological relevance of the obtained results, we looked at the top-scoring signatures from ReMap2020 (Fig. 1C) and ChIP-Atlas (Additional file 2: Fig. S2A) in TGF β induced EMT datasets. Among those, we identified known key TFs involved in the TGF β pathway such as *SMAD2/3/4* [19], *SSRP1*, *HNF1B* [20], *DDX5* [21] or *RELA* [22]. Other well-known EMT-linked TFs also returned as significant including

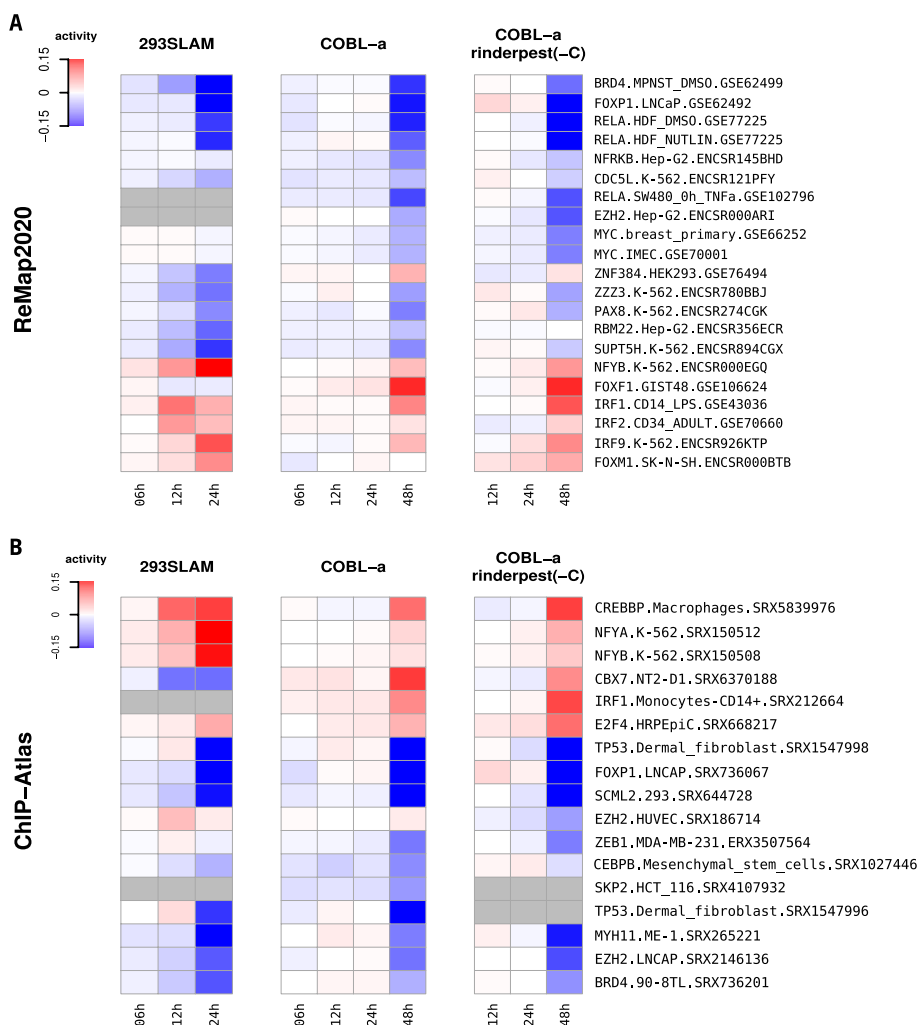


Fig. 2 Estimating transcription factors activities from gene expression during rinderpest infection in 293SLAM and COBL-a cell lines. **A, B** Heatmaps presenting TF activities of the most significant molecular signatures inferred using FANTOM5's rinderpest infection time-series dataset. The underlying experiments were performed in 293SLAM and COBL-a cell lines using native and recombinant rinderpest virus lacking accessory V and C proteins (rinderpest(-C)). Results obtained using ReMap2020 and ChIP-Atlas based molecular signatures are displayed on the top and bottom panels respectively

ZEB1, *SNAI2*, *TBX3*, *SOX4* (Additional file 4: Table S1, Additional file 5: Table S2, Additional file 6: Table S3). In case of FANTOM5's rinderpest infection dataset, top-scoring ReMap2020 and ChIP-Atlas signatures (Fig. 2, Additional file 7: Table S4) showed several TFs involved in the closely related measles infection pathway, including *RELA*, *IRF9*, *TP53* (KEGG PATHWAY:map05162) [23]. For human H3 embryonic stem cells differentiated to cardiomyocytes time-course dataset, a number of known heart development regulators were found among top-scoring ReMap2020 and ChIP-Atlas signatures (Additional file 8: Table S5), such as *JARID2*, *SMAD3*, *NKX2-5* (GO:0007507) [24].

Comparison with the state-of-the-art tools

We compared our results with state-of-the-art motif-based gene expression prediction framework ISMARA [1] and Lisa program which predicts the most likely transcriptional regulators from gene expression data based on ChIP-seq and chromatin accessibility data available in Cistrome Data Browser [25]. While ISMARA is conceptually similar and was inspirational to *xcORE*, Lisa takes a different approach. Using a user supplied list of differentially expressed genes, Lisa first selects a subset of relevant experiments describing chromatin state (H3K27ac ChIP-seq or DNase-seq) using lasso regression. Next it identifies the most relevant TF using *in-silico* deletion technique [8]. To compare with our results, we used both tools on our novel TGFβ induced EMT, rinderpest infection and embryonic stem cells differentiated to cardiomyocytes datasets. We have run ISMARA in RNA-seq mode with a genome version hg38 and no miRNA using raw FASTQ files for our novel TGFβ induced EMT dataset and BAM files available in FANTOM5 study [12] mapped against genome version hg38 for the other datasets. To use Lisa we performed differential expression analysis using edgeR [10] between the most extreme time points in our time-course datasets. Then lists of 100 most significant up- ($\log_{2}FC > 0$) and 100 most significant down-regulated ($\log_{2}FC < 0$) genes were submitted to Lisa. Next, we compared the results from all tools with a list of related transcriptional regulators. We constructed lists of related transcriptional regulators for each dataset using Gene Ontology term *regulation of epithelial to mesenchymal transition* (GO:0010717), KEGG pathway *Measles* (map05162) and Gene Ontology term *heart development* (GO:0007507) by including only regulators available in the references of all tools. The number of EMT related transcriptional regulators recovered among the top-scoring signatures was higher for *xcORE* and Lisa than ISMARA (Table 1). In case of rinderpest infection (Table 2) Lisa recovered the highest number of related TF in 293SLAM cell line. In the case of COBL-a and COBL-a rinderpest(-C) analyzes *xcORE* found one more TF than ISMARA and Lisa. Finally, for embryonic stem cells differentiated to

Table 1 Recovering epithelial to mesenchymal transition transcriptional regulators

Top signatures	A-549				MDA-231-D			
	ISMARA	Lisa	<i>xcORE</i>		ISMARA	Lisa	<i>xcORE</i>	
			ReMap2020	ChIP-Atlas			ReMap2020	ChIP-Atlas
1–10	SMAD4	SMAD3, SMAD4, GATA3	SMAD3, SMAD2	SMAD3, SMAD2, SMAD4	SMAD4, SMAD3	SMAD3, SMAD4	SMAD3, SMAD4, SMAD2, GATA3	
11–50		FOXA2, FOXA1	EZH2, GATA3, SMAD4	EZH2, FOXA2	SMAD4	GATA3	SMAD2, EZH2, FOXA2, FOXA1	
51–100	GATA3, FOXA1	NKX2-1, TCF7L2	FOXA1, FOXA2, NKX2-1	FOXA1, GATA3	FOXA1, NKX2-1	EZH2, FOXA1		

Table summarizing EMT-related transcriptional regulators recovered by ISMARA, Lisa and *xcORE* among their top-scoring signatures based on TGFβ induced EMT CAGE datasets. The list of EMT-related transcriptional regulators used to assess the recovery was constructed using Gene Ontology term *regulation of epithelial to mesenchymal transition* (GO:0010717) by including only regulators available in the references of all tools

Table 2 Recovering measles infection transcriptional regulators

Top signatures	293SLAM			COBL-a			COBL-a rinderpest(-C)		
	ISMARA	Lisa	xcore	ISMARA	Lisa	xcore	ISMARA	Lisa	xcore
1–10	STAT2, IRF3, IRF9	TP53, STAT1	IRF9			RELA			RELA, IRF9
11–50	RELA, JUN	STAT3, IRF9, REL, STAT1, NFKB1, REL, STAT2	RELA, STAT1, NFKB1, REL, STAT2	FOS	TP53, STAT3	TP53, FOS	IRF3	TP53, STAT3	STAT5A, NFKB1, JUN
51–100	NFKB1, FOS, STAT1		FOS	STAT2, IRF3	JUN, REL	JUN, STAT5A	FOS, REL	JUN, REL	FOS, TP53

Table summarizing measles infection pathway-related transcriptional regulators recovered by ISMARA, Lisa and xcore among their top-scoring signatures based on rinderpest infection datasets. The list of measles infection pathway-related transcriptional regulators used to assess the recovery was constructed using KEGG pathway/Measles (map05162) by including only regulators available in the references of all tools

Table 3 Recovering heart development transcriptional regulators

Top signatures	ISMARA	Lisa	<i>xcORE</i>	
			ReMap2020	ChIP-Atlas
1–10		GATA6, SMAD3, GATA4		
11–50	SNAI2, MEF2A, SRF, GATA4	SMAD1, EOMES, GATA3, SMAD2	SMAD3, NKX2-5, ATF2, TBX5, RBPJ	RARA
51–100	MEF2C, WT1	TBX5, REST, MBD2, TP53, SMAD4	SNAI2	JUN, TP53

Table summarizing heart development-related transcriptional regulators recovered by ISMARA, Lisa and *xcORE* among their top-scoring signatures based on Human H3 embryonic stem cells differentiated to cardiomyocytes time-series dataset. The list of heart development-related transcriptional regulators used to assess the recovery was constructed using Gene Ontology term *heart development* (GO:0007507) by including only regulators available in the references of all tools

cardiomyocytes (Table 3) Lisa was able to find the highest number of related TF, while *xcORE* and ISMARA found the same number of related TF.

Conclusions

Xcore provides a flexible framework for integrative analysis of gene expression and publicly available TF binding data to unravel putative transcriptional regulators and their activities. Our analyses showed superior results when using ChIP-seq based signatures as compared to motifs-based ones. We attribute this difference to the presence of biotype-specific binding information which might be lost in motifs that describe more general transcription factor binding preferences. Despite high numbers of ChIP-seq signatures and redundancy, our machine learning framework is able to select biologically relevant signatures. In our comparison with motif-based ISMARA and ChIP-seq based Lisa, *xcORE* performed competitively with those tools. Especially, both *xcORE* and Lisa worked exceptionally well at recovering EMT-related transcriptional regulators. However, a comprehensive comparison of *xcORE* with other tools would require further benchmarking efforts. Such efforts are currently hindered by the lack of standard benchmarking datasets for transcriptional regulators' inference problems. In conclusion, *xcORE* is useful for generating testable hypotheses about the data and provides a novel way to connect gene expression data with relevant ChIP-seq experiments.

Methods

TGF- β 1 stimulation to A-549/MDA-231-D

A-549 Lung cancer cells (CCL-185, ATCC) and MDA-231-D highly metastatic human breast cancer cells [26] (gift from Dr. Kohei Miyazono, Tokyo Univ.) were cultured in Dulbecco's modified Eagle's medium (Thermo Fisher Scientific Inc., Waltham, MA, USA) supplemented with 10% fetal bovine serum, 1 mM sodium pyruvate (Thermo Fisher Scientific Inc., Waltham, MA, USA) and penicillin/streptomycin (100 U/mL, 100 μ g/mL; Thermo Fisher Scientific Inc., Waltham, MA, USA). TGF- β 1 (7754-BH, Recombinant Human TGF-beta 1, R&D Systems) was added at the final concentration of 1 ng/mL. At 0, 1, 2, 4, 6, and 24 h post stimulation, cells were harvested followed by RNA extraction using RNeasy mini kit (Qiagen, Valencia, CA, USA). Transcriptome data was produced by nAnT-iCAGE [27]. CAGE libraries were sequenced on Illumina HiSeq 2500 (50-nt single read).

Abbreviations

CAGE	Cap analysis gene expression
CPM	Counts per million
CV	Cross-validation
EMT	Epithelial-mesenchymal transition
TF	Transcription factor
TFBS	Transcription factor binding site
TGFβ	Transforming growth factor beta

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-022-05084-0>.

Additional file 1: Figure S1. (A) Boxplots showing R^2 values for gene expression prediction models constructed either on gene- or promoter-level expression data. Each boxplot shows R^2 values pooled across all the replicates. Models were trained and evaluated in tenfold cross-validation on individual replicates, using data on gene expression changes between 0 and 24 h after treatment in our newly generated TGFβ induced EMT experiment performed in A-549 and MDA-231-D cell lines. The models were constructed using ReMap2020 or ChIP-Atlas molecular signatures.

Additional file 2: Figure S2. (A) Heatmap showing the dynamics of TF activities during TGFβ induced EMT. Heatmaps on the left present TF activities estimated using CAGE data from our newly generated TGFβ induced EMT experiment performed on A-549 and MDA-231-D cell lines. Heatmap on the right depicts TF activities estimated using previously published microarray data from the TGFβ induced EMT experiment performed on A-549 cell lines. The TF activities were calculated in the reference to 0 h time point. Only the top-scoring ChIP-Atlas signatures are shown. Grey color designates NA values. (B) Boxplots showing R^2 values for gene expression prediction models constructed using different molecular signature sets: Motif based (Jaspar, SwissRegulon) and ChIP-seq based (ReMap2020, ChIP-Atlas). Each boxplot shows R^2 values pooled across all the replicates. Models were trained and evaluated in tenfold cross-validation on individual replicates, using data on gene expression changes between 0 and 24 h after the rinderpest infection treatment experiment performed in 293SLAM cell line. (C) Boxplots showing R^2 values for gene expression prediction models trained using lasso, elastic net or ridge regression method. Each boxplot shows R^2 values pooled across all the replicates. Models were trained and evaluated in tenfold cross-validation on individual replicates, using data on gene expression changes between 0 and 24 h after treatment in our newly generated TGFβ induced EMT experiment performed in A-549 and MDA-231-D cell lines. The models were constructed using ReMap2020 molecular signatures and promoter-level expression data.

Additional file 3: Extended Materials and Methods. Extended description of procedures used to process the raw CAGE data, construct molecular signatures, and assess the accuracy of used models.

Additional file 4: Table S1. Table provides the activities of ReMap2020 and ChIP-Atlas molecular signatures estimated using TGFβ induced EMT in A-549 cell line dataset.

Additional file 5: Table S2. Table provides the activities of ReMap2020 and ChIP-Atlas molecular signatures estimated using TGFβ induced EMT in MDA-231-D cell line dataset.

Additional file 6: Table S3. Table provides the activities of ReMap2020 and ChIP-Atlas molecular signatures estimated using TGFβ induced EMT in A-549 cell line dataset (GSE17708).

Additional file 7: Table S4. Table provides the activities of ReMap2020 and ChIP-Atlas molecular signatures estimated using rinderpest infection in 293SLAM and COBL-a cell lines datasets.

Additional file 8: Table S5. Table provides the activities of ReMap2020 and ChIP-Atlas molecular signatures estimated using Human H3 embryonic stem cells differentiated to cardiomyocytes time-course.

Acknowledgements

We thank Dr. Iga Jancewicz for insightful comments on the manuscript. We thank Dr. Daizo Koinuma and Dr. Kohei Miyazono (Department of Molecular Pathology, Graduate School of Medicine, The University of Tokyo, Japan) for their kindly providing MDA-231-D cells. We thank Dr. Norbert Dojer for consulting the manuscript revision.

Author contributions

BK and EA conceived the study. TA, ST, MF and HS generated the data on TGFβ induced EMT in A-549 and MDA-231-D cell lines. MM and BK contributed to the design of the study and analyzed the data. MM wrote the R package. MM, CLW, EA and BK wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by a research grant from the Ministry of Education, Culture, Sport, Science and Technology of Japan for the RIKEN Center for Integrative Medical Sciences. MM was supported by RIKEN's IMS Internship Program. MM is recipient of the Postgraduate School of Molecular Medicine doctoral fellowship financed by the European Union through the European Regional Development Fund under Knowledge Education Development programme within the project "Next generation sequencing technologies in biomedicine and personalised medicine". The founding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Xcore and xcoredata R packages are open-source and freely available on GitHub under <https://github.com/bkaczkowski/xcore> and <https://github.com/mcjmigdal/xcoredata>. xcore user guide is available https://bkaczkowski.github.io/xcore/articles/xcore_vignette.html. Official releases of xcore and xcoredata packages are also included in the Bioconductor, from where they can be easily installed using BiocManager functionality. The EMT datasets generated and/or analyzed during the current study are available in the NCBI SRA repository: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17708> and <https://www.ncbi.nlm.nih.gov/bioproject/879326>. The rinderpest infection dataset analyzed during the current study is available in the FANTOM5 catalog: https://fantom.gsc.riken.jp/5/sstar/Rinderpest_infection_series. The Human H3 embryonic stem cells differentiated to cardiomyocytes dataset analyzed during the current study is available in the FANTOM5 catalog: https://fantom.gsc.riken.jp/5/sstar/ES_to_cardiomyocyte.

Availability and requirements

Project name: xcore. Project home page: <https://github.com/bkaczkowski/xcore>. Archived version: 1.1.4. Operating system: Platform independent. Programming language: R. Other requirements: None. License: GPL-2. Any restrictions to use by non-academics: No restrictions.

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests. EA took a position with GSK during the submission of this manuscript. GSK was not involved at any stage of the work presented here and there is no conflict of interest related to GSK for this work. BK took a position with AstraZeneca R&D during the submission of this manuscript. AstraZeneca R&D was not involved at any stage of the work presented here and there is no conflict of interest related to AstraZeneca R&D for this work.

Received: 3 July 2022 Accepted: 25 November 2022

Published online: 11 January 2023

References

- Balwierz PJ, Pachkov M, Arnold P, Gruber AJ, Zavolan M, van Nimwegen E. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* 2014. <https://doi.org/10.1101/gr.169508.113>.
- Schmidt F, Gasparoni N, Gasparoni G, Gianmoena K, Cadenas C, Polansky JK, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.* 2017;45:54–66.
- Madsen JGS, Rauch A, Van Hauwaert EL, Schmidt SF, Winnefeld M, Mandrup S. Integrated analysis of motif activity and gene expression changes of transcription factors. *Genome Res.* 2018;28:243–55.
- FANTOM Consortium, Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, et al. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet.* 2009;41:553–62.
- Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.* 2012;22:1711–22.
- Chèneby J, Ménétrier Z, Mestdagh M, Rosnet T, Douida A, Rhalloussi W, et al. ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* 2020;48:D180–8.
- Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* 2018;19:e46255.
- Qin Q, Fan J, Zheng R, Wan C, Mei S, Wu Q, et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol.* 2020;21:32.
- Karimzadeh M, Hoffman MM. Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome. *Genome Biol.* 2022;23:126.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26:139–40.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.
- Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* 2015;16:22.
- Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics.* 1970;12:55–67.
- Cule E, Vineis P, De Iorio M. Significance testing in ridge regression for genetic data. *BMC Bioinform.* 2011;12:372.
- Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33:1–22.

16. Arner E, Mejhert N, Kulyté A, Balwiercz PJ, Pachkov M, Cormont M, et al. Adipose tissue microRNAs as regulators of CCL2 production in human obesity. *Diabetes*. 2012;61:1986–93.
17. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM Jr. *The American soldier: adjustment during army life. (Studies in social psychology in World War II)*, vol. 1. Oxford: Princeton University Press; 1949.
18. Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J, Wright Z, Karnovsky A, et al. ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics*. 2010;26:456–63.
19. Xu J, Lamouille S, Derynck R. TGF- β -induced epithelial to mesenchymal transition. *Cell Res*. 2009;19:156–72.
20. Lavin DP, Tiwari VK. Unresolved complexity in the gene regulatory network underlying EMT. *Front Oncol*. 2020;10:554.
21. Dardenne E, Polay Espinoza M, Fattet L, Germann S, Lambert M-P, Neil H, et al. RNA helicases DDX5 and DDX17 dynamically orchestrate transcription, miRNA, and splicing programs in cell differentiation. *Cell Rep*. 2014;7:1900–13.
22. Tian B, Widen SG, Yang J, Wood TG, Kudlicki A, Zhao Y, et al. The NF κ B subunit RELA is a master transcriptional regulator of the committed epithelial-mesenchymal transition in airway epithelial cells. *J Biol Chem*. 2018;293:16528–45.
23. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hiraoka M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*. 2010;38(Database issue):D355–360.
24. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. *Bioinformatics*. 2009;25:288–9.
25. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res*. 2019;47:D729–35.
26. Ehata S, Hanyu A, Fujime M, Katsuno Y, Fukunaga E, Goto K, et al. Ki26894, a novel transforming growth factor- β type I receptor kinase inhibitor, inhibits in vitro invasion and in vivo bone metastasis of a human breast cancer cell line. *Cancer Sci*. 2007;98:127–33.
27. Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M, Carninci P, Hayashizaki Y, Itoh M. Detecting expressed genes using CAGE. In: Miyamoto-Sato E, Ohashi H, Sasaki H, Nishikawa J, Yanagawa H, editors. *Transcription factor regulatory networks: methods and protocols*. New York: Springer; 2014. p. 67–85.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Summary and conclusions

In eukaryotes, the first layer of gene expression regulation is mediated by DNA-binding proteins called transcription factors. These regulatory proteins can enhance or inhibit expression by binding to cis-regulatory DNA elements and interacting with the basal transcription machinery. Despite substantial progress in our understanding of these processes, identification of the transcription factors causative to gene expression changes remain one of the key questions in the field of transcriptional regulation. In my studies I have explored several bioinformatics approaches to study gene regulation on the genome wide level by integrating transcriptomics and epigenomics data. Through gene expression clustering combined with transcription factor motif enrichment analysis, I have identified putative regulators of co-expressed gene clusters in the contexts of heart development and early liver injury. These experiences led me to develop a new bioinformatics tool for transcription factor activity modeling using linear models and known transcription factors molecular signatures. Collectively, the presented articles showcase how gene expression and transcription factors binding data can be jointly analyzed to gain deeper understanding of gene regulation mechanisms.

The main results of the works included in this doctoral dissertation are:

- Development of ATAC-seq processing pipeline that allows identification of open chromatin regions, with methods to handle replicated data. The pipeline is implemented in Nextflow framework (Di Tommaso et al. 2017) and is available at our lab's GitLab repository (https://gitlab.com/zdglab/atacseq_pipeline).
- Characterization of the cardiomyocytes' transcriptome and epigenome landscape at early stages of heart development using zebrafish as a model organism, by means of RNA-seq expression clustering and motif enrichment analysis. The analysis revealed major transcriptomic and epigenomic shifts towards more cell type specific expression patterns during development. Data collected from *gata5*, *hand2*, and *tbx5* mutants, in which heart development is affected, suggested the predominant role of distal regulatory elements in cardiomyocytes development.
- Characterization of the transcriptomic and epigenomic response to hepatotoxic liver injury of selected liver cell types: liver sinusoidal endothelial cells, hepatocytes and hepatic stellate cells, using zebrafish as a model organism. RNA-seq expression clustering together with ATAC-seq based motif enrichment

analysis of the co-expression clusters indicated the endothelial cells to be the first cell population to respond to liver injury. The molecular response involves activation of genes related to metabolic and redox processes, including opening chromatin at their promoters. Motif enrichment analysis suggested the transcription factors FOXA1 and FOXA3 as potential regulators of endothelial cells activation.

- Development of xcore R package, implementing a flexible gene expression prediction framework. Its key use case is for modeling gene expression regulators based on large ChIP-seq databases. The package and user guide is available at <https://bkaczkowski.github.io/xcore/>.

Authors contributions statements

Warsawa 26 10 2022

(city, date)

Michał Pawlak, Ph.D.

Author contribution statement

As a co-author of the work entitled "*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in conceiving the study, I participated in collecting embryos and performing in vivo experiments, I participated in collecting biological material, I participated in preparation of NGS libraries and performing RNA-seq and ATAC-seq, I performed LSM, I participated in performing bioinformatics and statistical analysis, I contributed to the design of the study and data interpretation, I contributed to genomic data analysis, I participated in writing the manuscript, I am the first author of the study, I read and approved the manuscript.

I estimate my contribution to this publication as 20%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

Michał Pawlak

(signature)

Oxford, 10/11/2022

(city, date)

Katarzyna Zofia Kędzierska, M.Sc.

Author contribution statement

As a co-author of the work entitled "*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in collecting embryos and performing in vivo experiments, I participated in collecting biological material, I participated in performing bioinformatics and statistical analysis, I contributed to the design of the study and data interpretation, I read and approved the manuscript.

I estimate my contribution to this publication as 10%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



(signature)

WARSAW, 19.05.2022

(city, date)

Karim Abu Nahia, M.Sc.

Author contribution statement

As a co-author of the work entitled "*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in preparation of NGS libraries and performing RNA-seq and ATAC-seq, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

AGU Nahia
.....
(signature)

Tokyo, 30-05-2022
.....
(city, date)

Jordan A. Ramilowski, Ph.D.

Author contribution statement

As a co-author of the work entitled "*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

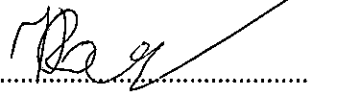
I participated in performing bioinformatics and statistical analysis, I contributed to the design of the study and data interpretation, I contributed to genomic data analysis, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

.....

(signature)

Warsaw, 08-11-2022

(city, date)

Łukasz Bugajski, M.Sc.

Author contribution statement

As a co-author of the work entitled "*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in performing FACS analysis, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

Łukasz Bugajski

(signature)

Osaka, Japan
17th November, 2022
(city, date)

Kosuke Hashimoto, Ph.D.

Author contribution statement

As a co-author of the work entitled "*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

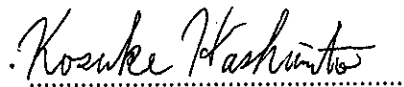
I contributed to genomic data analysis, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

.....

(signature)

Cambridge, UK, 23/05/2022

(city, date)

Aleksandra Marconi, M.Phil.

Author contribution statement

As a co-author of the work entitled "*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in collecting embryos and performing in vivo experiments, I participated in collecting biological material, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

A. Marconi

.....
(signature)

Warsaw, 09.11.2022

(city, date)

Katarzyna Piwocka, Ph.D.

Author contribution statement

As a co-author of the work entitled "***Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development***", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in performing FACS analysis.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



(signature)

Yokohama, 27th May 2022.

(city, date)

Piero Carninci, Ph.D.

Author contribution statement

As a co-author of the work entitled “*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*”, I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I contributed to the design of the study and data interpretation, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

A handwritten signature in black ink, appearing to read 'Piero Carninci', written over a horizontal dotted line.

(signature)

Warsaw, 26.05.2022

.....
(city, date)

Cecilia Lanny Winata, Ph.D.

Author contribution statement

As a co-author of the work entitled "*Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in conceiving the study, I contributed to the design of the study and data interpretation, I contributed to genomic data analysis, I participated in writing the manuscript, I am a senior corresponding author of the study, I read and approved the manuscript.

I estimate my contribution to this publication as 15%.

I estimate Maciej Migdał contribution to this publication as 10%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



.....
(signature)

Warsaw, 26/05/22

(city, date)

Eugeniusz Tralle, M.Sc., Eng.

Author contribution statement

As a co-author of the work entitled "*Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in performing *in vivo* experiments and collecting biological material, I performed histological staining and took microscopic images, I participated in preparing the figures, I participated in writing the manuscript, I read and approved the manuscript.

I estimate my contribution to this publication as 40%.

I estimate Maciej Migdał contribution to this publication as 40%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation, participation in preparing the figures and writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

Tralle

(signature)

Warsaw, 19.05.2022

(city, date)

Karim Abu Nahia, M.Sc.

Author contribution statement

As a co-author of the work entitled "*Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I prepared NGS libraries and performed sequencing, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 40%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation, participation in preparing the figures and writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

AGU Nahia

(signature)

Warsaw, 08-11-2022

(city, date)

Łukasz Bugajski, M.Sc.

Author contribution statement

As a co-author of the work entitled "*Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I performed FACS analysis, I read and approved the manuscript.

I estimate my contribution to this publication as 1%.

I estimate Maciej Migdał contribution to this publication as 40%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation, participation in preparing the figures and writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

Łukasz Bugajski

(signature)

Oxford, 10/11/2022.

(city, date)

Katarzyna Zofia Kędzierska, M.Sc.

Author contribution statement

As a co-author of the work entitled “*Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift*”, I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I performed preliminary experiments and optimized the protocols, I read and approved the manuscript.

I estimate my contribution to this publication as 2%.

I estimate Maciej Migdał contribution to this publication as 40%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation, participation in preparing the figures and writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



(signature)

Warsaw, 11/08/2022

.....

(city, date)

Filip Garbicz MD.

Author contribution statement

As a co-author of the work entitled "*Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I analyzed and interpreted histological data, I read and approved the manuscript.

I estimate my contribution to this publication as 1%.

I estimate Maciej Migdał contribution to this publication as 40%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation, participation in preparing the figures and writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



.....

(signature)

Warsaw, 03.11.2022

(city, date)

Katarzyna Piwocka, Ph.D.

Author contribution statement

As a co-author of the work entitled "**Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift**", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I supervised the FACS analysis, I read and approved the manuscript.

I estimate my contribution to this publication as 1%.

I estimate Maciej Migdał contribution to this publication as 40%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation, participation in preparing the figures and writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



.....
(signature)

Warsaw, 26.05.2022

.....
(city, date)

Cecilia Lanny Winata, Ph.D.

Author contribution statement

As a co-author of the work entitled "*Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in conceiving the study. I participated in writing the manuscript, I am a senior corresponding co-author of the study, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 40%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation, participation in preparing the figures and writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



.....
(signature)

Wrocław 26.05.2022

(city, date)

Michał Pawlak, Ph.D.

Author contribution statement

As a co-author of the work entitled "*Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

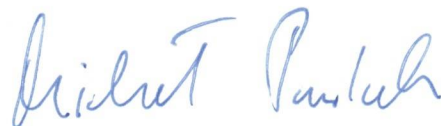
I participated in performing *in vivo* experiments and collecting biological material, I participated in performing bioinformatics and statistical analysis, I contributed to the design of the study and data interpretation, I participated in preparing the figures, I participated in conceiving the study, I participated in writing the manuscript, I am a senior corresponding author of the study, I read and approved the manuscript.

I estimate my contribution to this publication as 5%.

I estimate Maciej Migdał contribution to this publication as 40%.

It included: participation in performing bioinformatics and statistical analysis, contributing to the design of the study and data interpretation, participation in preparing the figures and writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



.....
(signature)

Yokohama, Dec 2nd 2022

(city, date)

Takahiro Arakawa, Ph.D

Author contribution statement

As a co-author of the work entitled "*xcore: an R package for inference of gene expression regulators using publicly available human ChIP-seq experiments*", I hereby declare that my that my own contribution to the research work and manuscript preparation constitutes:

I participated in generating the data on TGF β induced EMT in A-549 and MDA-231-D cell lines;
I read and approved the manuscript.

I estimate my contribution to this publication as 5 %.

I estimate Maciej Migdał contribution to this publication as 45 %.

It included: contributing to the design of the study and data analysis; writing the R software package; participation in writing the manuscript; reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.


.....

(signature)

Yokohama, 12/02/2022

Prof. Satoshi Takizawa

Author contribution statement

As a co-author of the work entitled "*xcORE: an R package for inference of gene expression regulators using publicly available human ChIP-seq experiments*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in generating the data on TGF β induced EMT in A-549 and MDA-231-D cell lines; I read and approved the manuscript.

I estimate my contribution to this publication as 5 %.

I estimate Maciej Migdał contribution to this publication as 45 %.

It included: contributing to the design of the study and data analysis; writing the R software package; participation in writing the manuscript; reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

Satoshi Takizawa

Yokohama, Feb. 9. 2023
..... (city, date)

Masaaki Furuno, Ph.D.

Author contribution statement

As a co-author of the work entitled "*xcore: an R package for inference of gene expression regulators using publicly available human ChIP-seq experiments*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in generating the data on TGF β induced EMT in A-549 and MDA-231-D cell lines;
I read and approved the manuscript.

I estimate my contribution to this publication as 5 %.

I estimate Maciej Migdał contribution to this publication as 45 %.

It included: contributing to the design of the study and data analysis; writing the R software package; participation in writing the manuscript; reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.

Masaaki Furuno
..... (signature)

Yokohama, Japan.

December 5, 2022

Harukazu Suzuki, Ph.D.

Author contribution statement

As a co-author of the work entitled "*xcORE: an R package for inference of gene expression regulators using publicly available human ChIP-seq experiments*", I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

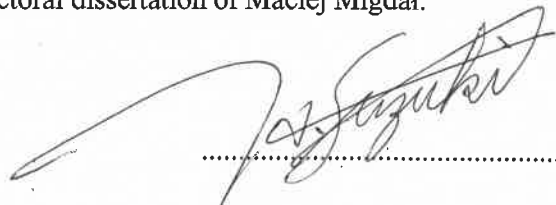
I participated in generating the data on TGF β induced EMT in A-549 and MDA-231-D cell lines; I read and approved the manuscript.

I estimate my contribution to this publication as 5 %.

I estimate Maciej Migdał contribution to this publication as 45 %.

It included: contributing to the design of the study and data analysis; writing the R software package; participation in writing the manuscript; reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.


..... (signature)

Erik Arner, Ph.D.

Author contribution statement

As a co-author of the work entitled “*xcore: an R package for inference of gene expression regulators using publicly available human ChiP-seq experiments*”, I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in conceiving the study and writing the manuscript; I read and approved the manuscript.

I estimate my contribution to this publication as 5 %.

I estimate Maciej Migdał contribution to this publication as 45 %.

It included: contributing to the design of the study and data analysis; writing the R software package; participation in writing the manuscript; reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



..... (signature)

Warsaw, 6.12.2022

(city, date)

Cecilia Lanny Winata, Ph.D.

Author contribution statement

As a co-author of the work entitled "*xcORE: an R package for inference of gene expression regulators using publicly available human ChiP-seq experiments*", I hereby declare that my that my own contribution to the research work and manuscript preparation constitutes:

I participated in writing the manuscript; I read and approved the manuscript.

I estimate my contribution to this publication as 5 %.

I estimate Maciej Migdał contribution to this publication as 45 %.

It included: contributing to the design of the study and data analysis; writing the R software package; participation in writing the manuscript; reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



(signature)

Cambridge, December 3rd, 2022

(city, date)

Bogumił Kaczkowski, Ph.D.

Author contribution statement

As a co-author of the work entitled “*xcore: an R package for inference of gene expression regulators using publicly available human ChiP-seq experiments*”, I hereby declare that my own contribution to the research work and manuscript preparation constitutes:

I participated in conceiving the study, contributed to designing the study and analyzing the data;

I participated in writing the manuscript, I read and approved the manuscript.

I estimate my contribution to this publication as 25 %.

I estimate Maciej Migdał contribution to this publication as 45 %.

It included: contributing to the design of the study and data analysis, writing the R software package, participation in writing the manuscript, reading and approving the manuscript.

I agree to include this publication in the doctoral dissertation of Maciej Migdał.



(signature)