Modeling and integration of multi-omics data to study regulatory landscapes governing placenta development

by

Ha T.H. Vu

(Thi Hong Ha Vu)

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Geetu Tuteja, Major Professor
Karin Dorman
Claus Kadelka
James Koltes
Justin Walley

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2023

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

NOMENCLATURE

TE	Trophectoderm
EPC	Ectoplacental cone
TGC	Trophoblast giant cells
GO	Gene ontology
MGI	Mouse Genome Informatics
TPM	Transcripts per million
EVT	Extravillous trophoblast
SCT	Syncytiotrophoblast
VCT	Villous trophoblast
FDR	False discovery rate
sc	Single-cell
sn	Single-nucleus
TSS	Transcription start site
FPKM	Fragments per kilobase of transcript per million of mapped reads

ACKNOWLEDGMENTS

This journey at times seemed impossible, but I made it to the finishing line thanks to wonderful people in my life.

I would like to thank my major professor, Dr. Geetu Tuteja, for being one of the best mentors I could imagine. Thank you for seeing my potential and encouraging me to reach beyond it, ever since day one when I had the biology knowledge of a 9^{th} grader, and my programming skills were at `print("Hello World!")`. Thank you for giving me not only freedom but also reassurance in times of need. I grow to be a sharper and more confident scientist thanks to you. I would also like to thank my co-major professor, Dr. Karin Dorman, for always encouraging me to explore new ideas and think deeply about it. Thank you for checking on me every semester until I mustered up enough courage to pursue a difficult idea that I never thought I could have done. I would like to thank my committee members, Dr. Claus Kadelka, Dr. James Koltes, and Dr. Justin Walley, for their great support and guidance; especially Dr. Claus Kadelka for helping me feel so much more confident in my mathematical ability. I have come a long way thanks to you.

I would like to thank our collaborators, especially Dr. Karin Dorman and Dr. Yudi Zhang at Iowa State University (ISU), and Dr. Michael Soares and Regan Scott at University of Kansas Medical Center. A great portion of this thesis was made possible thanks to their work. I also would like to thank many ISU staff members for their help. Thanks to Trish Stauble for giving me the warmest welcome when I first got accepted to ISU, and making me feel like home although I was $8,600$ miles away from home. Thank you, Carla Harris and Danise Jones, for helping me through every process and making sure I am on the right track, for chatting with me and making our interactions the most heartwarming ones. My work required a great amount of tech support from the Research IT and Biology IT Departments - thank you all for your help.

Tuteja Lab members, both current and former, are the kindest co-workers I could ever ask for. I'd love to extend my special thanks to Dr. Haninder Kaur for her immense help in my biology learning journey, Dr. Rebekah Starks for being a great mentor and a big sister who answered all of my questions even when she was busy, and to Ashwini Rangaraj and Kelby Kies for not only being my lab mates but also my close friends.

I have the most wonderful years living in Ames thanks to my friends in and outside of Iowa State. I would like to thank my American parents, Bruce and Brenda Tracy, for being my closest friends, for loving me and caring for me as your own daughter. Thanks to you, Iowa is now my second home, because here I have you - my "adoptive" family. Thanks to the Wednesday wing night group (Jia Liu, Rebekah Starks, Priyanka Bhandary, Ashiwini Rangaraj and Kelby Kies) who have always been there for me through the ups and downs of my adulting journey. Thank you, Ella Faulhaber, Austin Sympson, View from the Margins book club, and Central Iowa Ukulele group - you kept me sane and reminded me life has so much more beyond school. Thank you to my cohort and many other BCB students, for being the best study buddies and being such a friendly community. Deepest thanks to my friends from Vietnam for their life-long friendships no matter where we are in the world. I am forever grateful for you.

Last, I would like to express my deepest gratitude to my immediate and extended family for their support in my journey. Thank you, Mom and Dad, for raising me to be a strong independent woman, so that I had the courage to uproot myself from Vietnam to a new country and explore my new career. Thank you my sister for always looking out for me and inspiring me every single day to try harder and become better. I would like to thank my extended family, many of whom are scientists and have inspired me to become one. Thank you Francis for your snuggles and biscuits - you are one of the best parts in my life.

Thank you all for making this happen. I dedicate this thesis to you.

ABSTRACT

The placenta is a transient organ that is crucial during pregnancy. It has multiple functions to ensure optimal fetal growth, including nutrient transport, oxygen exchange and immune protection. The placenta develops in a stage-wise manner and requires precise regulation of gene expression. Abnormalities in placental gene regulation can lead to pregnancy disorders such as preeclampsia, placenta accreta and placental abruption, which can be detrimental to the short and long-term health of both the mother and the fetus. However, the regulatory mechanisms governing placental development, especially with respect to gene regulatory networks, are poorly understood.

In this dissertation, we aimed to identify regulatory networks associated with placental development by developing computational methods, and by analyzing and integrating various sequencing data at both the bulk and single-cell level. First, we generated and analyzed transcriptomic data from mouse fetal placenta tissues at embryonic day (e) 7.5, e8.5 and e9.5 to identify groups of genes that regulate placenta-specific developmental processes using cluster analysis, differential expression analysis, and network analysis. Second, we developed a deep learning framework to identify genome-wide chromatin accessibility regions. This framework is applicable for not only for placenta-derived data but also data generated in other tissues. Third, we integrated single-cell transcriptome and single-nucleus chromatin accessibility data generated from the rat uterine-placental interface to identify conserved gene regulatory networks governing rat and human placenta development.

The completion of these studies has led to a better understanding of the gene – gene, gene – transcription factor, and transcription factor – *cis*-regulatory element interactions regulating placental development. Furthermore, the pipelines and tools developed, including the novel deep

learning framework for chromatin accessibility analysis, are not limited to rodent and human placenta, but can be used to analyze data generated in any tissue or organism.

CHAPTER 1. GENERAL INTRODUCTION

The placenta is a transient organ established during pregnancy that is a crucial point of contact between the mother and the fetus. The placenta is required for fetal development and maintenance of pregnancy, as it carries out multiple functions including immune protection, nutrient transport and oxygen exchange, each described in more detail below.

During pregnancy, the placenta acts as an immune barrier to protect the fetus from harmful pathogens such as bacteria and viruses (Hoo et al., 2020). For example, it has been found that the outer cell layer of the placenta is resistant to *Listeria monocytogenes* due to its elasticity property, which helps reduce infection susceptibility (Zeldovich et al., 2013). Moreover, the placenta recognizes and sends signals to the fetus to regulate its immune responses (Ding et al., 2022). Failures in placental immune protection can lead to serious infections, which have been associated with preterm birth (Kiefer et al., 2009; Peltier et al., 2012).

Nutrient transport is one of the well appreciated roles of the placenta. Molecules and substrates are transported within the blood across the placenta, under the influence of several factors such as the maternal-fetal circulation (Jensen and Chernyavsky, 2019), placental metabolism, and transporter proteins in the placental barrier (Lager and Powell, 2012). For example, reduced activities of amino acid transporters in the placenta are observed in pregnancies with intrauterine growth restriction (IUGR), a condition where the fetus has abnormally low birth weight (Jansson and Powell, 2007; Gaccioli and Lager, 2016). On the other hand, increased placental nutrient transports in pregnancies with gestational diabetes are reported to be associated with fetal overgrowth (Gaccioli et al., 2013; Hulme et al., 2019).

Oxygen exchange between the mother and the fetus depends on the blood flow through the dense network of vasculature within the placenta. This exchange process includes the following steps: i) oxygen is carried through the maternal uterine arteries; ii) oxygen is transferred across

the placental membrane with umbilical veins; iii) the fetus uses the oxygen; and iv) deoxygenated blood returns to the maternal side via the umbilical arteries (Saini et al., 2020). Abnormalities in oxygen exchange processes such as increased fetal oxygen demands through the placenta are associated with pregnancy disorders such as IUGR and chronic fetal hypoxaemia (Saini et al., 2020).

Despite being short-lived by design, defects in placental development can lead to long-term impacts on the health of both the mother and fetus. Pregnancy complications potentially caused by placental disorders, some mentioned above, such as preeclampsia and fetal growth restriction affect $5 - 10\%$ of pregnancies (Rana et al., 2019; Bamfo and Odibo, 2011). Children exposed to preeclampsia run higher risk of developing cardiovascular, metabolic, and neurological diseases (Lu et al., 2019). Currently, the diagnosis and treatment of placental diseases are limited because understanding of early placental development is lacking. Therefore, it is urgently needed to understand various aspects of placentogenesis in order to detect and prevent these disorders.

1.1 Human placental development

The first steps of the human placenta formation can be divided into three phases: pre-lacunar, lacunar and primary villous stage, starting from $\sim$5 to $\sim$18 days post fertilization (dpf) (Turco and Moffett, 2019). During the pre-lacular stage ($\sim$5 to $\sim$7 dpf), the trophectoderm (TE), the outer layer of a blastocyst, attaches to the endometrium, and the blastocyst implants into the uterus. In the lacunar stage ($\sim$14 dpf), fluid-filled spaces (lacunae) are formed to serve as a direct connection to the maternal blood. This direct contact between placental tissues and the maternal blood defines the human placentation to be hemochorial (Soares et al., 2018). Next, during the primary villous stage, the trophoblast cells underlying the syncytium (cytoptrophoblast, CTB) experience rapid proliferation to form a primary villous structure where CTB is in the core and syncytiotrophoblast (STB) is in the outer layer. From 18 dpf, CTB, STB and extraembryonic mesoderm also together invade the maternal tissues, expand and form branching villi to establish the placenta (Turco and Moffett, 2019).

During the first trimester, further proliferation and branching of primary villi then form the villous trees, and the lacunae become the intervillous space (Turco and Moffett, 2019). The intervillous space has a dense network of villi, each of which can be divided into three sub-layers: fetal blood capillaries at the core, a layer of CTB cells and an outer layer of STB cells (Hemberger et al., 2020). First, fetal blood capillaries at the core have complex structures with multiple cell types such as mesenchymal cells, mesenchymal derived macrophages (Hofbauer cells), fetal vascular smooth muscle cells and endothelial cells (Wang and Zhao, 2010). These fetal capillaries connect to the umbilical arteries and veins, establishing the circulation necessary for nutrient and waste exchange. Second, the layer of CTB cells, also known as villous cytotrophoblast (VCT), are highly proliferative and give rise to multiple trophoblast subtypes such as extravillous trophoblast cells (EVT) and STB. Throughout the first trimester, EVT actively invade into maternal tissues in a process called trophoblast invasion, which is crucial to establish the maternal-fetal circulation. The EVT migrate into the decidua then remodel the maternal spiral arteries by incorporating themselves into the blood vessel walls. This transformation turns maternal spiral arteries into highly flexible, pliable, and thin-walled vessels, enabling a continuous and sufficient blood supply to the placenta (Silva and Serakides, 2016). Last, the outer STB layer plays crucial roles in nutrient transport and immune protection for the fetus as it covers the entire surface of villous trees and are directly connected with maternal blood (Turco and Moffett, 2019).

By the end of the first trimester ($\sim$week 20), the placenta is considered definitive with two additional layers to the intervillous space, namely chorionic plate and basal plate (Jansen et al., 2020). The chorionic plate is lined with an amniotic layer consisting of an ectodermal epithelium layer completely surrounding the embryo. Last, basal plate is the direct interface between the maternal and fetal tissues. It consists of multiple cell types such as fibroblasts, natural killer cells and EVT migrating from the chorion (Hoo et al., 2020). By this time, the placenta is established enough to take over crucial functions such as nutrient transport; however, it continues to grow throughout pregnancy, and is considered fully mature by week 34.

1.2 Using mouse and rat models to study human placental development

There are a limited number of mechanistic studies on human placenta due to ethical concerns. Because of this, rodent models have proven to be useful to shed light on different stages of placental development.

The mouse and rat placenta start developing upon implantation when the blastocyst attaches to the uterine wall and the trophectoderm proliferates to form the extra-embryonic ectoderm (ExE) and the ectoplacental cone (EPC) (Hemberger et al., 2020; Soares et al., 2012). In both species, the activation of the extra-embryonic mesoderm lineage is responsible for the formation of the allantois and the extra-embryonic mesodermal layers of the amnion and chorion (Hemberger et al., 2020; Woods et al., 2018). At ∼e8.5 in mouse (gestational day 10 to 11 in rat), the amnion and chorion fuse together in a process called chorioallantoic attachment, which enables the extra-embryonic mesoderm-derived blood vessels to invaginate into the chorionic trophoblast layer (Hemberger et al., 2020; Cross et al., 2003b; Furukawa et al., 2019). Fetal blood vessels and trophoblast-lined maternal blood sinuses together are the main components of the labyrinth layer in the mouse and rat placenta (Hemberger et al., 2020; Soares et al., 2012), which is a crucial site for blood circulation between the mother and the fetus.

Placental development in the mouse and rat share several similarities to that in human. First, they are all hemochorial (Hemberger et al., 2020), unlike other models such as horses and sheep that have an epitheliochorial placenta and carnivores that have an endotheliochorial placenta (Furukawa et al., 2011). Second, they have overall similar placental structures with three main layers, although each layer may have different characteristics specific to the species (Furukawa et al., 2019; Cross et al., 2003a). For example, the labyrinth zone in mouse and rat placenta is analogous to human villous tree and plays a role in gas exchange, nutrient transport for the fetus and removing waste products (Furukawa et al., 2019; Soares et al., 2012; Rossant, 2001). Third, placental development in the three organisms is regulated by several common signaling pathways, including those involving the transcription factors HIF-1α and ASCL2 (Rossant, 2001; Robb et al., 2017). Last, trophoblast subtypes in the three species share functional and structural

commonalities. For example, the trophoblast giant cells in mouse and the interstitial invasive trophoblast cells in rat are considered analogous to human EVT because of their invasive characteristics, although the degrees of invasiveness vary between species (Soncin et al., 2015; Soares et al., 2012).

It is important to note differences in placental development among species, in order to choose a suitable model for the research question of interest. For example, structurally, rodent (mouse and rat) placenta is trichorial, meaning that maternal and fetal blood flow are separated by three layers of trophoblast, while human placenta is monochorial (having one trophoblast layer as separation) (Marinello and Patisaul, 2021; Schmidt et al., 2018). Further, the mouse uses the visceral yolk sac to maintain early pregnancy before the placenta is formed, which is completely absent in humans (Schmidt et al., 2018). Finally, the depth of trophoblast invasion in the three species is not identical. In human, trophoblast cells invade deeply into the inner third of the myometrium, while in mice, trophoblast invasion is restricted to the decidua (Schmidt et al., 2018). The rat, however, has deeper invasion than the mouse, where the interstitial trophoblast invade beyond the decidua into the mesometrial triangle (Carter et al., 2006). Given these differences, it is important to integrate insights from mouse and rat models with those from human to identify conserved elements, such as genes, regulatory elements, cell lineages, and signaling pathways, and to validate findings from animal models in human cell lines (Grigsby, 2016; Schmidt et al., 2018), in order to better understand human placental development.

1.3 Regulation of gene expression

Gene expression is tightly controlled so that the right genes are expressed at the appropriate time and at a proper level to maintain normal cellular function. Gene regulation involves multiple mechanisms, such as transcriptional and epigenetic regulation, post-transcriptional regulation, translational and post-translational regulation.

Transcriptional regulation describes the regulation of the rate of synthesizing RNA from DNA (transcription). Transcriptional activation or repression relies on several factors such as activities

and functions of transcription factors (TFs), and the accessibility of regulatory elements. TFs are proteins that bind to specific DNA sequences in the regulatory elements of genes and can either activate or repress transcription. The specific DNA sequences that TFs bind to are known as motifs. In order for TFs to recognize and bind to motifs, the nucleosomal structures around the regulatory region must be loosened and become accessible. There are multiple mechanisms governing the opening or closeness of nucleosomal structures, one of which is histone modification (Annunziato, 2008). Histone modification refers to a process where proteins and complexes with specific enzymatic activities such as the acetyl, methyl, or phosphate groups are recruited and added to the histones (Bannister and Kouzarides, 2011; Fischle et al., 2003). The addition of these groups can directly perturb chromatin structure (Bannister and Kouzarides, 2011) and is often associated to different nucleosomal states. For example, histone acetylation such as H3K27 acetylation are often found at active enhancer elements, allowing for chromatin accessibility and protein binding (Shlyueva et al., 2014). On the other hand, histone methylation such as H3K9me3 are often associated with closed chromatin and repression of gene expression (Bannister and Kouzarides, 2011; Shlyueva et al., 2014).

Regulatory regions such as enhancers and repressive elements can be located far away from gene promoters (Lin et al., 2022; Courey and Jia, 2001; Li and Arnosti, 2011). These distal regions can interact with promoter regions through chromatin looping mechanisms (Whalen et al., 2016; Sanborn et al., 2015; Kadauke and Blobel, 2009; Fulco et al., 2019; Courey and Jia, 2001), and can thereby contribute to the formation of protein complexes between TFs binding at distal regions and those at promoter regions to activate or repress genes. Moreover, regardless of chromatin looping events, multiple TFs localized to one regulatory region can work in a combinatorial manner, or they can act individually (Allan and Thor, 2015; Spitz and Furlong, 2012). Given the complex nature of gene regulation, to study the regulatory mechanisms effectively, it is important to identify regulatory elements, the TFs binding to these locations, and how they interact with their target genes.

After DNA is transcribed into RNA and before RNA is translated into a protein, post-transcriptional regulation occurs. Post-transcriptional regulation includes alternative splicing, RNA editing, and RNA degradation, all of which can affect the quantity or quality of the final protein product. Next, proteins are synthesized from messenger RNA (mRNA), then they experience post-translational regulation, which can involve different modifications such as phosphorylation, acetylation, and glycosylation, which can affect the function, stability, or localization of the protein.

In this dissertation, we focus on studying transcriptional regulation and the identification of the accessible regions of the genome.

1.3.1 Using next generation sequencing (NGS) experiments to investigate transcriptional regulation and the chromatin lansdscape

RNA sequencing (RNA-seq) can be used to determine gene expression levels in different tissues or stages of development. The assay can be carried out using whole tissue (bulk RNA-seq) or at the single cell (sc)/single nucleus (sn) level. A standard protocol for RNA-seq experiments involves RNA isolation, complementary DNA (cDNA) conversion, sequencing library preparation, and sequencing using an NGS platform (Kukurba and Montgomery, 2015). Each of these steps requires several considerations and method selection depending on the research question. For example, if sequencing mRNA is the research focus, the sequencing library can be prepared with poly-A selection. If one would like to sequence mRNA, pre-mRNA and noncoding RNA, the library can be designed with ribo-depletion to remove ribosomal RNA (Kukurba and Montgomery, 2015). On the sc/sn level, snRNA-seq is more suitable for cells that are difficult to isolate such as brain, skeletal muscle and adipose (Ding et al., 2020).

For bulk RNA-seq, gene expression levels are generally considered averaged from a population of cells (Hegenbarth et al., 2022); therefore, it is suitable to detect global changes between conditions of interest. However, single cell resolution is not achieved with bulk RNA-seq. For sc/snRNA-seq, the expression is measured in each individual cell/nucleus, and this data type is

particularly useful to characterize expression profiles of individual cell populations, especially rare cell types. However, sc/snRNA-seq data suffers from various technical challenges such as sparsity, the amount of starting materials, cell size, or cell death (Lähnemann et al., 2020). Therefore, the combination of bulk and sc/snRNA-seq is beneficial for the study of biological processes that involve multiple cell types.

In the context of placental studies, several bulk and sc/sn RNA-seq data have been generated and used to investigate specific characteristics of placental tissues and cell populations under different conditions and time points in multiple experimental models. For example, RNA-seq analyses were carried out using data generated from endothelial cells isolated from e16.0 labyrinth zone of *Igf2* mouse mutants, revealing the direct roles of *Igf2* in placental development (Sandovici et al., 2022). Comparison amongst RNA-seq data from trophoblast cells differentiated in cell culture using different *in vitro* protocols has been done to provide justification for choosing a suitable experimental model (Seetharam et al., 2022). Additionally, scRNA-seq in the rat and human uterine-placental interface, as well as snRNA-seq data from mouse labyrinth zone, have been generated, providing valuable resources to study expression profiles of complex cell populations in the placenta (Vento-Tormo et al., 2018; Marsh and Blelloch, 2020; Marsh et al., 2022; Scott et al., 2022).

To study genome-wide chromatin accessibility (also called openness), several NGS assays can be used, such as MNase-seq, DNase-seq, and ATAC-seq (MNase: micrococcal nuclease; DNase I: deoxyribonuclease I; ATAC: assay for transposase-accessible chromatin). These three assays involve using different DNA cleavage agents such as the endo-exonuclease MNase (Johnson et al., 2006), non-specific double-strand endonuclease DNase I (Song and Crawford, 2010) and Tn5 transposase (Buenrostro et al., 2015), respectively, to detect and digest open chromatin regions. Consequently, the accessible regions can be subjected to PCR amplification and sequencing. Out of the three assays mentioned here, ATAC-seq has quickly become one of the more popular methods as it requires low starting cell numbers and has a simple and fast protocol (Tsompana and Buck, 2014). Similar to RNA-seq, ATAC-seq can be carried out on whole tissue or at the

sc/sn level to provide information about chromatin accessibility landscapes in multiple tissues or developmental stages. Moreover, utilization of chromatin accessibility sequencing assays like ATAC-seq enables inferences of nucleosome spacing, positioning and occupancy, and TF binding footprints, which are crucial in the study of chromatin architectures. ATAC-seq has proven to be useful to study the development of the placenta, such as to identify repressed gene networks in e9.5 mouse placenta (Starks et al., 2019) and to characterize the chromatin accessibility underlying trophoblast cell development (Nelson et al., 2017; Dong et al., 2020; Arutyunyan et al., 2023).

Chromatin immunoprecipitation followed by high-through put sequencing (ChIP-seq) can also provide information about chromatin accessibility, but can also be used to annotate specific kinds of regulatory regions. For example, mono-methylation of lysine 4 on histone H3 (H3K4me1) marks enhancers (O'Geen et al., 2011; Kimura, 2013), which are regulatory DNA sequences that can enhance associated genes' transcription. Trimethylation of lysine 4 on histone H3 (H3K4me3), on the other hand, marks gene promoters (Kimura, 2013). Trimethylation of lysines 9 and 27 on histone H3 (H3K9me3 and H3K27me3, respectively) mark compacted chromatin regions, associated with gene repression (Kimura, 2013). Assaying multiple histone modification via ChIP-seq, and including ATAC-seq data, can deepen our understanding of gene regulation. For example, in combining H3K27me3, H3K4me1 and H3K4me3, Starks et al. determined sets of regulatory regions associated with house-keeping genes or actively repressed genes with neural functions, and active enhancers associated to placental-specific genes (Starks et al., 2021). Upon integrating the region sets with ATAC-seq data, Starks et al. also observed a significant ratio of openness in house-keeping genes' regulatory regions, which agreed with the biological mechanisms of genes essential for basic cellular functions. In human EVT, it was found that H3K27ac marks had a high overlap with ATAC-Seq regions more accessible in the EVT cell state compared to trophoblast stem cells (Varberg et al., 2022).

An important aspect of studying chromatin is determining the 3-D organization of the genome, including the long-range interactions between regulatory elements and genes. After the

development of Chromosome Conformation Capture (3C) technique in 2002 (Dekker et al., 2002) which aims to detect the interaction frequency between two genomic loci, multiple assays with similar purposes have developed such as 4C (Simonis et al., 2006), 5C (Dostie et al., 2006), and Hi-C (Lieberman-Aiden et al., 2009). Out of these assays, Hi-C is one of the most popular as it utilizes NGS techniques such as high-throughput and parallel sequencing, which allow for interaction profiling between fragments within a chromosome or between different chromosomes (Oluwadare et al., 2019). In the absence of data from 3C-based assays, computational methods can be used to infer the interactions between regulatory elements and genes. These computational strategies will be discussed in the next section.

1.3.2 Computational methods for next generation sequencing methods

While specific analysis workflows of several next-generation sequencing (NGS) may vary depending on the experimental design and data type, they often have a similar general processing pipeline, involving basic steps such as quality control, read alignment, read quantification, and downstream analyses. This general pipeline is summarized below.

1.3.2.1 Pre-processing steps

Pre-processing steps are necessary to exclude low quality data and help to produce reliable and high-quality downstream results, as well as to save computational time. After sequencing files are obtained, the first step is to assess the quality of the raw sequencing data, also known as "reads", using various quality control (QC) metrics. These metrics include but are not limited to per base quality scores, distribution of read lengths, sequence duplication levels and over-represented sequences (which signal adapter contamination or sequencing artifacts). The most common program for this purpose is FastQC (Andrews, 2010), which is a Java toolkit. Since the development of FastQC, several versions of FastQC in other programming languages such as R (fastqcr (Kassambara, 2023)) and Python (Falco (de Sena Brandine and Smith, 2019)) have been implemented, allowing for faster run time and flexibility in pipeline development. To

compile many QC reports for different samples into one report, researchers can use MultiQC (Ewels et al., 2016).

Next, the raw sequencing reads are preprocessed to remove low-quality bases, trim adapters, and filter out reads that do not meet certain quality criteria. Popular tools for read preprocessing are Trimmomatic (Bolger et al., 2014), bbduk (Bushnell, 2023), Cutadapt (Martin, 2011), and Trim Galore! (Krueger, 2015). No matter which tool is used, trimming adapters and low-quality base pairs, as well as eliminating low-quality reads, in general improve the quality of downstream analyses (Del Fabbro et al., 2013). After the read pre-processing step, the remaining high-quality reads are then aligned or mapped to a reference genome. Depending on the type of NGS data, researchers can choose appropriate aligners. To illustrate, for RNA-seq data, it is advisable to employ aligners that can account for gene splicing events and align reads to the genome despite intron gaps. For instance, HISAT2 (Kim et al., 2015) and STAR (Dobin et al., 2013) are splice-aware aligners that are well-suited for this purpose. For ATAC-seq, ChIP-seq and Hi-C data, popular alignment software are BWA (Li, 2013) and Bowtie 2 (Langmead and Salzberg, 2012) which utilize the Burrows-Wheeler Transform (BWT) algorithm (Burrows, 1994) for efficient data compression and fast pattern matching.

Following sequence alignment is read quantification. For RNA-seq data, popular quantification tools are HTSeq (Anders et al., 2015), featureCounts (Liao et al., 2014) and RSEM (Li and Dewey, 2011). Recently, several algorithms capable of simultaneous pseudo-alignment and quantification have been developed such as Kallisto (Bray et al., 2016) and Salmon (Patro et al., 2017) which help speed up the analysis run time considerably. For ATAC-seq and ChIP-seq, read counting is followed by statistical testing of the read pile-ups in a process often referred to as "peak calling", which aims to determine the significance of regions of interest. There are many software to call peaks for ATAC-seq and ChIP-seq; they can be classified into two groups of methods: traditional statistical modeling (e.g., MACS2 (Zhang et al., 2008), ZINBA (Rashid et al., 2011), and HMMRATAC (Tarbell and Liu, 2019)), and supervised machine learning methods (e.g., CNN-peaks (Oh et al., 2020) and LanceOtron (Hentges et al., 2021)) (Yan et al.,

2020). Depending on the experiments, peak widths and shapes can vary greatly (Barth and Imhof, 2010), requiring researchers to choose the suitable statistical models for calling peaks (Thomas et al., 2017). With Hi-C data, the quantification step is different from that of RNA-seq, ATAC-seq or ChIP-seq. Briefly, after genome alignment and before interaction counting, sequencing reads are further filtered to obtain true pairs of reads which indicate valid interactions, then interactions are quantified and normalized for several factors such as mappability, GC content and fragment length (Lajoie et al., 2015). The quantification step are included in popular processing tools for Hi-C such as HiC-Pro (Servant et al., 2015) and HiCUP (Wingett et al., 2015).

1.3.2.2 Downstream analyses

Upon obtaining quantified features (expression level for RNA-seq, peaks for ATAC-seq and ChIP-seq, interaction counts for Hi-C), there are several downstream analyses one can do depending on their research questions. For example, a common analysis is identifying differentially expressed genes (RNA-seq), differentially bound regions (transcription factor ChIP-seq), regions with differential activity level (histone modification ChIP-seq), or differentially accessible regions (ATAC-seq). These types of tests often rely on popular algorithms such as EdgeR (Robinson et al., 2009) and DESeq2 (Love et al., 2014) which model feature counts with negative binomial distribution, built for RNA-seq data. When applying EdgeR or DESeq2 models to ChIP-seq and ATAC-seq data, normalization requires special attention (Wu et al., 2015; Reske et al., 2020) since: i) ChIP-seq/ATAC-seq peaks can be present at more varied locations in the genome than RNA-seq, which is generally limited to exons, ii) ChIP-seq and ATAC-seq have different technical variability from RNA-seq, and iii) ChIP-seq/ATAC-seq read counts in peaks may not follow the assumptions underlying normalizing methods for RNA-seq data. If analyses are carried out specifically with RNA-seq data that has been quantified using Kallisto (Bray et al., 2016), one can also use Sleuth (Yi et al., 2018) which can model the differences in both gene and transcript-level counts. Once obtaining a gene list or region list of interest, researchers can carry out motif enrichment analysis to predict transcription factor binding within regulatory

elements, gene ontology enrichment analysis to predict biological functions of genes, or infer interaction networks amongst genes and between transcription factors and genes. Popular network inference methods using gene expression level includes WGCNA (Weighted Gene Co-expression Network Analysis) (Langfelder and Horvath, 2008) and GENIE3 (GEne Network Inference with Ensemble of trees) (Huynh-Thu et al., 2010).

Inferring interactions between regulatory elements such as distal enhancers and genes is an on-going challenge, especially when Hi-C or other data measuring locus interactions are not readily available. Several computational methods have been developed, such as EpiTensor (Zhu et al., 2016), which uses high-order tensor decomposition to predict the interaction space between regulatory elements, IM-PET (Integrated Method for Predicting Enhancer Targets) (He et al., 2014), which relies on the correlation between activities at distal enhancers and gene promoters, and GREAT (Genomic Regions Enrichment of Annotations Tool) (McLean et al., 2010), which associates enhancers to genes based on genomic distances. On sc/sn level, Cicero (Pliner et al., 2018) is a useful method which predicts co-accessibility networks based on the covariance between regulatory elements using graphical LASSO (Friedman et al., 2008). Among the computational approaches for bulk data that do not involve locus contact information, genomic distances seem to perform relatively well (Fulco et al., 2019), and can be of great use in associating regulatory regions to genes.

1.4 Dissertation Organization

Chapter 1 serves as a general introduction to placental biology and computational techniques related to the work in this dissertation. It includes an introduction to placental development in human, mouse and rat models, the mechanisms of gene regulation, some relevant next generation sequencing assays and general approaches for the analyses of each data type.

Chapter 2 consists of a published manuscript titled "Identifying novel regulators of placental development using time-series transcriptome data" published in *Life Science Alliance* (Vu et al., 2023a). In this report, we generated RNA-seq data from mouse fetal placental tissues at e7.5, e8.5

and e9.5, and used clustering and network analyses to predict regulators of placenta-specific processes. First, with clustering analysis, we identified gene groups with timepoint-specific patterns of expression. Next, with network and deconvolution analysis, we determined gene network modules in mouse tissues that have similar expression profiles to different cell types in the human placenta. This is a crucial analysis to identify cell type-specific markers with bulk data because it helps to overcome confounding factors when averaging expression from multiple cell types. Last, we validated the biological roles of our predicted regulators *in vitro* with a human cell line, HTR8-SVneo cells, providing evidence for a potential role for the identified candidates across species (Fig 1.1).

Chapter 3 consists of a published manuscript titled "Unsupervised Contrastive Peak Caller for ATAC-seq" published in *Genome Research* (Vu et al., 2023b). In this work, we developed a novel deep learning framework to identify chromatin accessible regions, also known as "calling peaks", with ATAC-seq data. Our tool, named RCL (Replicated Contrastive Learning), is the first-ever framework developed to specifically integrate shared signals across biological replicates to identify high confident open genomic regions. We demonstrated that RCL performed superior to multiple existing tools in not only mouse placenta tissue data but also several human cell lines (Fig 1.2).

Chapter 4 consists of a manuscript named "Core conserved transcriptional regulatory networks define the invasive trophoblast cell lineage" published in *Development* (Vu et al., 2023c). In this report, we integrated single-cell (sc) RNA-seq and single-nucleus (sn) ATAC-seq data generated from the rat uterine-placental interface at gestational day 15.5 and 19.5, and bulk ATAC-seq data from human EVT cells. As a result, we identified invasive trophoblast-specific chromatin accessible regions in rat, and conserved regions between rat and human. Using motif enrichment and network analysis, we also determined a network of transcription factors and genes that likely play functional roles in the trophoblast lineage in both the rat and human uterine-placental interface (Fig 1.3).

Chapter 5 consists of the conclusions of the thesis and discusses the future directions of the work.

1.5 Main figures

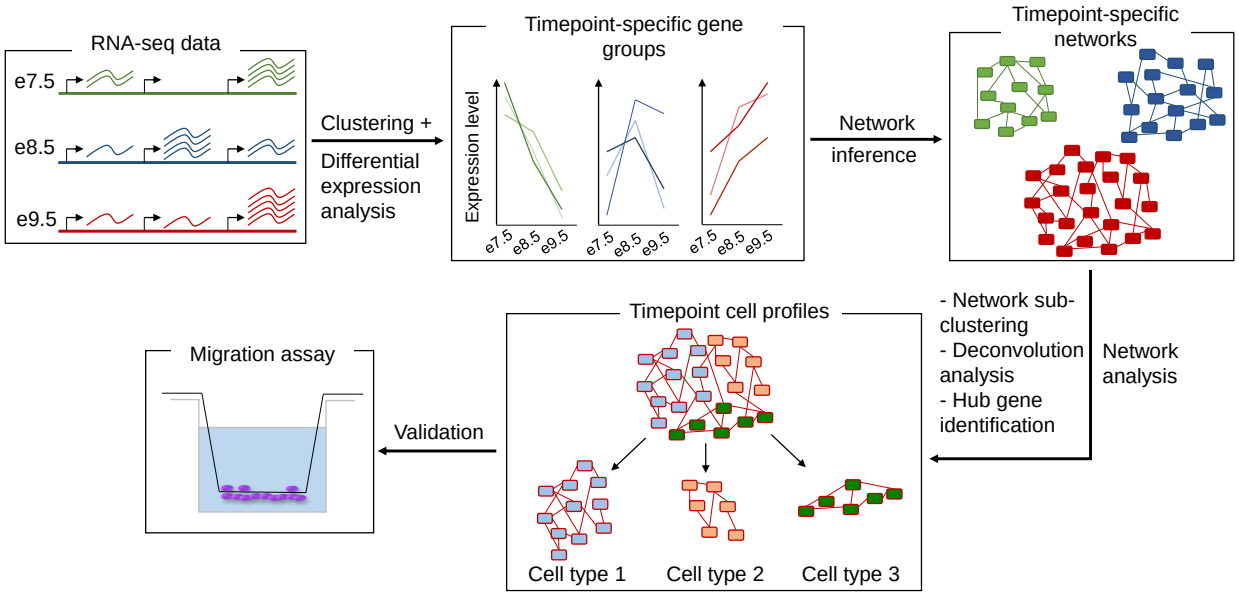

Figure 1.1: RNA-seq data from embryonic day (e)7.5, e8.5 and e9.5 mouse placenta was analyzed to identify novel biomarkers of placenta developmental processes. Selected markers were validated in the human cell line HTR-8/SVneo to show their potential roles in cell migration regulation.

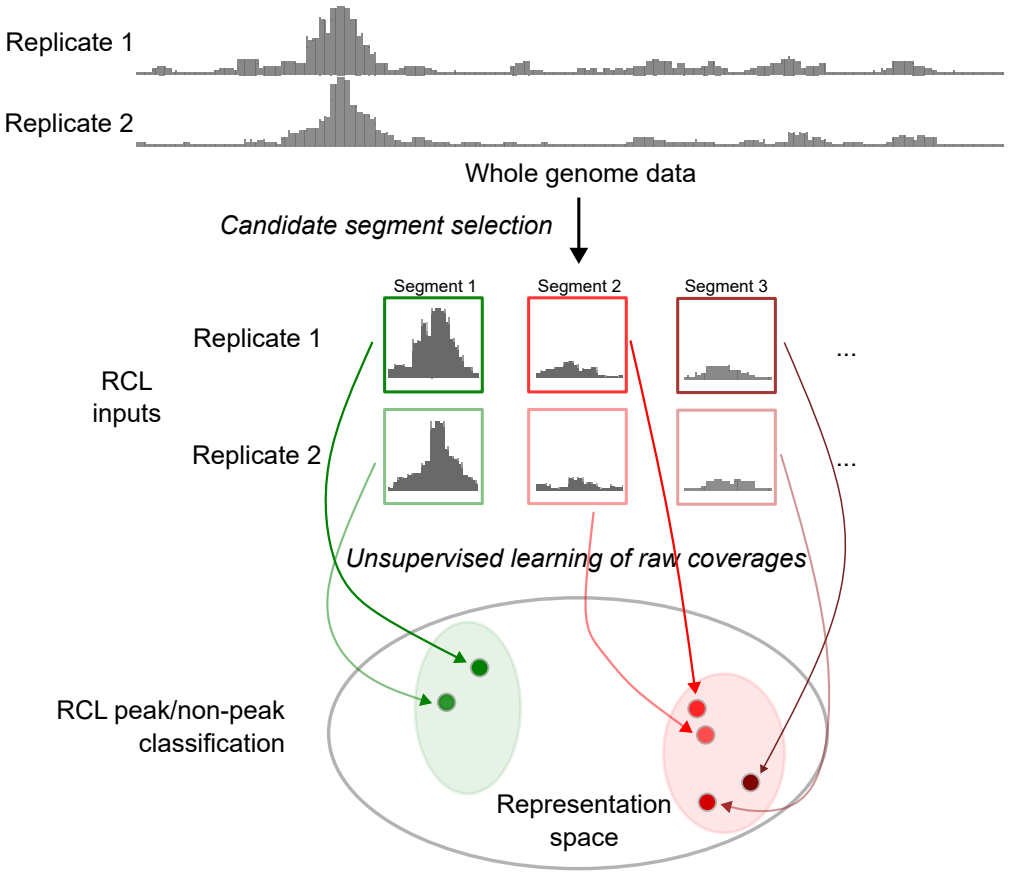

Figure 1.2: Overview of RCL (Replicated Contrastive Learning) framework where shared signals from biological replicates in ATAC-seq experiments are used to identify chromatin accessible regions. The framework involves two main steps: candidate region selection based on whole genome coverage data, and unsupervised contrastive learning to classify candidate regions into peak and non-peak classes.

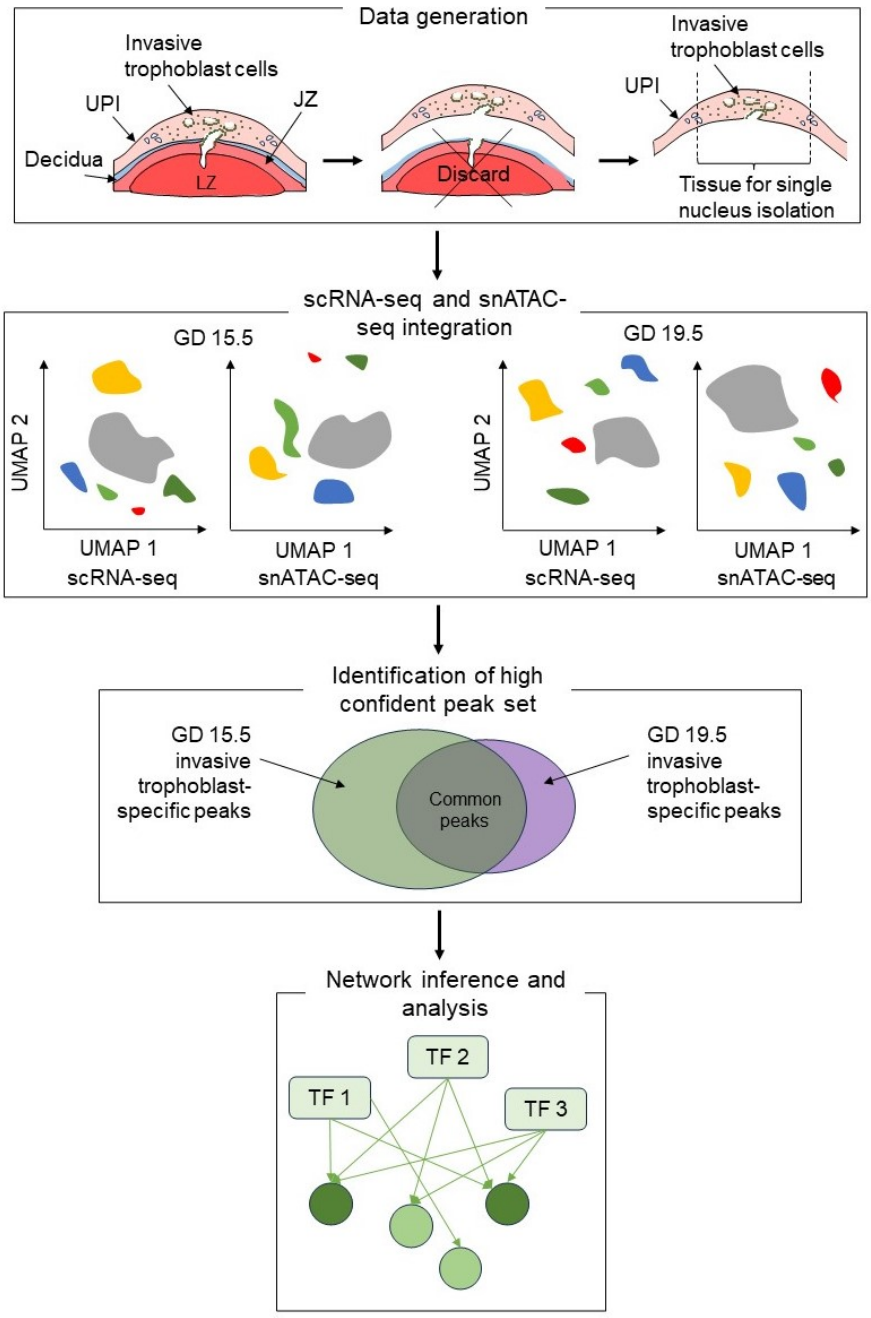

Figure 1.3: Single nucleus ATAC-seq data from rat uterine-placental interface on gestational day 15.5 and 19.5 were generated and integrated with single cell RNA-seq data with matching conditions.

Figure 1.3: (Continued)

High confident conserved chromatin accessible regions in the invasive trophoblast cell (iTB) populations were then identified, and used for transcription factor (TF) - gene network inference. Network analysis identified combinations of TFs that regulate iTB. Panel 1 ("Data generation") was adapted from Scott et al. (Scott et al., 2022). Abbreviations: UPI, uterine-placental interface; JZ, junctional zone; LZ: labyrinth zone; GD, gestational day; sc, single cell; sn, single nucleus; TF, transcription factor.

1.6 References

Allan, D. W. and Thor, S. (2015). Transcriptional selectors, masters, and combinatorial codes: regulatory principles of neural subtype specification. *Wiley Interdisciplinary Reviews. Developmental Biology*, 4(5):505.

Anders, S., Pyl, P. T., and Huber, W. (2015). Htseq—a python framework to work with high-throughput sequencing data. *bioinformatics*, 31(2):166–169.

Andrews, S. (2010). FastQC - A quality control tool for high throughput sequence data. *Babraham Bioinformatics*.

Annunziato, A. (2008). Dna packaging: nucleosomes and chromatin. *Nature education*, 1(1):26.

Arutyunyan, A., Roberts, K., Troulé, K., Wong, F. C., Sheridan, M. A., Kats, I., Garcia-Alonso, L., Velten, B., Hoo, R., Ruiz-Morales, E. R., Sancho-Serra, C., Shilts, J., Handfield, L. F., Marconato, L., Tuck, E., Gardner, L., Mazzeo, C. I., Li, Q., Kelava, I., Wright, G. J., Prigmore, E., Teichmann, S. A., Bayraktar, O. A., Moffett, A., Stegle, O., Turco, M. Y., and Vento-Tormo, R. (2023). Spatial multiomics map of trophoblast development in early pregnancy. *Nature 2023 616:7955*, 616(7955):143–151.

Bamfo, J. E. and Odibo, A. O. (2011). Diagnosis and management of fetal growth restriction. *Journal of pregnancy*, 2011:640715.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395.

Barth, T. K. and Imhof, A. (2010). Fast signals and slow marks: the dynamics of histone modifications. *Trends in Biochemical Sciences*, 35(11):618–626.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5):525–7.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109(1):21.29.1–21.29.9.

Burrows, M. (1994). A block-sorting lossless data compression algorithm. *SRC Research Report, 124*.

Bushnell, B. (2023). Bbtools software package. 2014. *Available online: http:/sourceforge. net/projects/bbmap (accessed on 11 June 2021)*.

Carter, A., Enders, A., Jones, C., Mess, A., Pfarrer, C., Pijnenborg, R., and Soma, H. (2006). Comparative placentation and animal models: Patterns of trophoblast invasion – a workshop report. *Placenta*, 27:30–33. Trophoblast Research Placenta: From Molecule to Morphology to Mother.

Courey, A. J. and Jia, S. (2001). Transcriptional repression: the long and the short of it. *Genes & development*, 15(21):2786–2796.

Cross, J. C., Baczyk, D., Dobric, N., Hemberger, M., Hughes, M., Simmons, D. G., Yamamoto, H., and Kingdom, J. C. (2003a). Genes, development and evolution of the placenta. *Placenta*, 24(2-3):123–130.

Cross, J. C., Simmons, D. G., and Watson, E. D. (2003b). Chorioallantoic morphogenesis and formation of the placental villous tree. *Annals of the New York Academy of Sciences*, 995(1):84–93.

de Sena Brandine, G. and Smith, A. D. (2019). Falco: high-speed fastqc emulation for quality control of sequencing data. *F1000Research*, 8.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science*, 295(5558):1306–1311.

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on illumina ngs data analysis. *PloS one*, 8(12):e85024.

Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., Hughes, T. K., Wadsworth, M. H., Burks, T., Nguyen, L. T., et al. (2020). Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nature biotechnology*, 38(6):737–746.

Ding, J., Maxwell, A., Adzibolosu, N., Hu, A., You, Y., Liao, A., and Mor, G. (2022). Mechanisms of immune regulation by the placenta: Role of type I interferon and interferon-stimulated genes signaling during pregnancy*. *Immunological Reviews*, 308(1):9–24.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

Dong, C., Beltcheva, M., Gontarz, P., Zhang, B., Popli, P., Fischer, L. A., Khan, S. A., Park, K.-m., Yoon, E.-J., Xing, X., Kommagani, R., Wang, T., Solnica-Krezel, L., and Theunissen, T. W. (2020). Derivation of trophoblast stem cells from naïve human pluripotent stem cells. *eLife*, 9:e52504.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048.

Fischle, W., Wang, Y., and Allis, C. D. (2003). Histone and chromatin cross-talk. *Current opinion in cell biology*, 15(2):172–183.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., and Engreitz, J. M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics 2019 51:12*, 51(12):1664–1669.

Furukawa, S., Hayashi, S., Usuda, K., Abe, M., Hagio, S., and Ogawa, I. (2011). Toxicological pathology in the rat placenta. *Journal of Toxicologic Pathology*, 24(2):95–111.

Furukawa, S., Tsuji, N., and Sugiyama, A. (2019). Morphology and physiology of rat placenta for toxicological evaluation. *Journal of Toxicologic Pathology*, 32(1):1–17.

Gaccioli, F. and Lager, S. (2016). Placental Nutrient Transport and Intrauterine Growth Restriction. *Frontiers in Physiology*, 7:40.

Gaccioli, F., Lager, S., Powell, T., and Jansson, T. (2013). Placental transport in response to altered maternal nutrition. *Journal of developmental origins of health and disease*, 4(2):101–115.

Grigsby, P. L. (2016). Animal models to study placental development and function throughout normal and dysfunctional human pregnancy. In *Seminars in reproductive medicine*, volume 34, pages 011–016. Thieme Medical Publishers.

He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer–promoter interactome in human cells. *Proceedings of the National Academy of Sciences*, 111(21):E2191–E2199.

Hegenbarth, J.-C., Lezzoche, G., De Windt, L. J., and Stoll, M. (2022). Perspectives on Bulk-Tissue RNA Sequencing and Single-Cell RNA Sequencing for Cardiac Transcriptomics. *Frontiers in Molecular Medicine*, 2:2.

Hemberger, M., Hanna, C. W., and Dean, W. (2020). Mechanisms of early placental development in mouse and humans. *Nature Reviews Genetics*, 21(1):27–43.

Hentges, L. D., Sergeant, M. J., Downes, D. J., Hughes, J. R., and Taylor, S. (2021). LanceOtron: A deep learning peak caller for ATAC-Seq, ChIP-Seq, and DNase-Seq. *Bioinformatics*, 38(18):4255–4263.

Hoo, R., Nakimuli, A., and Vento-Tormo, R. (2020). Innate Immune Mechanisms to Protect Against Infection at the Human Decidual-Placental Interface. *Frontiers in Immunology*, 11:2070.

Hulme, C. H., Nicolaou, A., Murphy, S. A., Heazell, A. E., Myers, J. E., and Westwood, M. (2019). The effect of high glucose on lipid metabolism in the human placenta. *Scientific reports*, 9(1):1–9.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9).

Jansen, C. H., Kastelein, A. W., Kleinrouweler, C. E., Van Leeuwen, E., De Jong, K. H., Pajkrt, E., and Van Noorden, C. J. (2020). Development of placental abnormalities in location and anatomy. *Acta Obstetricia et Gynecologica Scandinavica*, 99(8):983–993.

Jansson, T. and Powell, T. L. (2007). Role of the placenta in fetal programming: underlying mechanisms and potential interventional approaches. *Clinical science*, 113(1):1–13.

Jensen, O. E. and Chernyavsky, I. L. (2019). Blood flow and transport in the human placenta. *Annual Review of Fluid Mechanics*, 51:25–47.

Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P., and Fire, A. Z. (2006). Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Research*, 16(12):1505.

Kadauke, S. and Blobel, G. A. (2009). Chromatin loops in gene regulation. *Biochimica et biophysica acta*, 1789(1):17.

Kassambara, A. (2023). *fastqcr: Quality Control of Sequencing Data*. R package version 0.1.3.

Kiefer, D. G., Keeler, S. M., Rust, O. A., Wayock, C. P., Vintzileos, A. M., and Hanna, N. (2009). Is midtrimester short cervix a sign of intraamniotic inflammation? *American Journal of Obstetrics and Gynecology*, 200(4):374.e1–374.e5.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357–360.

Kimura, H. (2013). Histone modifications for human epigenome analysis. *Journal of human genetics*, 58(7):439–445.

Krueger, F. (2015). Trim galore!: A wrapper around cutadapt and fastqc to consistently apply adapter and quality trimming to fastq files, with extra functionality for rrbs data. *Babraham Institute*.

Kukurba, K. R. and Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 2015(11):951.

Lager, S. and Powell, T. L. (2012). Regulation of Nutrient Transport across the Placenta. *Journal of Pregnancy*, 2012:14.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35.

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The hitchhiker's guide to hi-c analysis: practical guidelines. *Methods*, 72:65–75.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12:1–16.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Li, L. M. and Arnosti, D. N. (2011). Long-and short-range transcriptional repressors induce distinct chromatin states on repressed genes. *Current Biology*, 21(5):406–412.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293.

Lin, X., Liu, Y., Liu, S., Zhu, X., Wu, L., Zhu, Y., Zhao, D., Xu, X., Chemparathy, A., Wang, H., Cao, Y., Nakamura, M., Noordermeer, J. N., La Russa, M., Wong, W. H., Zhao, K., and Qi, L. S. (2022). Nested epistasis enhancer networks for robust genome regulation. *Science (New York, N.Y.)*, 377(6610):1077–1085.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21.

Lu, J., Wu, W., Xin, Q., Zhou, C., Wang, J., Ni, Z., Liu, D., Xu, Y., Yu, Y., Yang, N., Sun, Y., He, B., Kong, S., Wang, S., Wang, C., and Wang, H. (2019). Spatiotemporal coordination of trophoblast and allantoic Rbpj signaling directs normal placental morphogenesis. *Cell Death & Disease 2019 10:6*, 10(6):1–14.

Marinello, W. P. and Patisaul, H. B. (2021). Chapter nine - endocrine disrupting chemicals (edcs) and placental function: Impact on fetal brain development. In Vandenberg, L. N. and Turgeon, J. L., editors, *Endocrine-Disrupting Chemicals*, volume 92 of *Advances in Pharmacology*, pages 347–400. Academic Press.

Marsh, B. and Blelloch, R. (2020). Single nuclei RNA-seq of mouse placental labyrinth development. *eLife*, 9:1–27.

Marsh, B., Zhou, Y., Kapidzic, M., Fisher, S., and Blelloch, R. (2022). Regionally distinct trophoblast regulate barrier function and invasion in the human placenta. *eLife*, 11.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501.

Nelson, A. C., Mould, A. W., Bikoff, E. K., and Robertson, E. J. (2017). Mapping the chromatin landscape and Blimp1 transcriptional targets that regulate trophoblast differentiation. *Scientific Reports 2017 7:1*, 7(1):1–15.

Oh, D., Strattan, J. S., Hur, J. K., Bento, J., Urban, A. E., Song, G., and Cherry, J. M. (2020). CNN-Peaks: ChIP-Seq peak detection pipeline using convolutional neural networks that imitate human visual inspection. *Scientific Reports*, 10(1):7933.

Oluwadare, O., Highsmith, M., and Cheng, J. (2019). An overview of methods for reconstructing 3-d chromosome and genome structures from hi-c data. *Biological procedures online*, 21(1):1–20.

O'Geen, H., Echipare, L., and Farnham, P. J. (2011). Using chip-seq technology to generate high-resolution profiles of histone modifications. *Epigenetics Protocols*, pages 265–286.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419.

Peltier, M. R., Klimova, N. G., Arita, Y., Gurzenda, E. M., Murthy, A., Chawala, K., Lerner, V., Richardson, J., and Hanna, N. (2012). Polybrominated Diphenyl Ethers Enhance the Production of Proinflammatory Cytokines by the Placenta. *Placenta*, 33(9):745.

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., and Trapnell, C. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell*, 71(5):858–871.e8.

Rana, S., Lemoine, E., Granger, J., and Karumanchi, S. A. (2019). Preeclampsia: Pathophysiology, Challenges, and Perspectives. *Circulation Research*, 124(7):1094–1112.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D. (2011). Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, 12:1–20.

Reske, J. J., Wilson, M. R., and Chandler, R. L. (2020). Atac-seq normalization method can significantly affect differential accessibility analysis and interpretation. *Epigenetics & chromatin*, 13(1):1–17.

Robb, K. P., Cotechini, T., Allaire, C., Sperou, A., and Graham, C. H. (2017). Inflammation-induced fetal growth restriction in rats is associated with increased placental hif-1α accumulation. *PLoS One*, 12(4):e0175805.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rossant, J. (2001). Stem cells from the Mammalian blastocyst. *Stem cells (Dayton, Ohio)*, 19(6):477–482.

Saini, B. S., Morrison, J. L., and Seed, M. (2020). *Gas Exchange across the Placenta*, page 34–56. Cambridge University Press.

Sanborn, A. L., Rao, S. S., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):E6456–E6465.

Sandovici, I., Georgopoulou, A., Pérez-García, V., Hufnagel, A., López-Tello, J., Lam, B. Y., Schiefer, S. N., Gaudreau, C., Santos, F., Hoelle, K., Yeo, G. S., Burling, K., Reiterer, M., Fowden, A. L., Burton, G. J., Branco, C. M., Sferruzzi-Perri, A. N., and Constância, M. (2022). The imprinted igf2-igf2r axis is critical for matching placental microvasculature expansion to fetal growth. *Developmental Cell*, 57(1):63–79.e8.

Schmidt, F., List, M., Cukuroglu, E., Köhler, S., Göke, J., and Schulz, M. H. (2018). An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics*, 34(17):i908–i916.

Scott, R. L., Vu, H. T., Jain, A., Iqbal, K., Tuteja, G., and Soares, M. J. (2022). Conservation at the uterine-placental interface. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41):e2210633119.

Seetharam, A. S., Vu, H. T. H., Choi, S., Khan, T., Sheridan, M. A., Ezashi, T., Roberts, R. M., and Tuteja, G. (2022). The product of BMP-directed differentiation protocols for human primed pluripotent stem cells is placental trophoblast and not amnion. *Stem Cell Reports*.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16(1):1–11.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics 2014 15:4*, 15(4):272–286.

Silva, J. F. and Serakides, R. (2016). Intrauterine trophoblast migration: A comparative view of humans and rodents. *Cell Adhesion and Migration*, 10(1-2):88–110.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11):1348–1354.

Soares, M. J., Chakraborty, D., Karim Rumi, M. A., Konno, T., and Renaud, S. J. (2012). Rat placentation: An experimental model for investigating the hemochorial maternal-fetal interface. *Placenta*, 33(4):233.

Soares, M. J., Varberg, K. M., and Iqbal, K. (2018). Hemochorial placentation: development, function, and adaptations. *Biology of Reproduction*, 99(1):196–211.

Soncin, F., Natale, D., and Parast, M. M. (2015). Signaling pathways in mouse and human trophoblast differentiation: A comparative review. *Cellular and Molecular Life Sciences*, 72(7):1291–1302.

Song, L. and Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols*, 2010(2):pdb.prot5384.

Spitz, F. and Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics 2012 13:9*, 13(9):613–626.

Starks, R. R., Biswas, A., Jain, A., and Tuteja, G. (2019). Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics and Chromatin*, 12(1):1–16.

Starks, R. R., Kaur, H., and Tuteja, G. (2021). Mapping cis-regulatory elements in the midgestation mouse placenta. *Scientific Reports 2021 11:1*, 11(1):1–13.

Tarbell, E. D. and Liu, T. (2019). HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Research*.

Thomas, R., Thomas, S., Holloway, A. K., and Pollard, K. S. (2017). Features that define the best chip-seq peak calling algorithms. *Briefings in bioinformatics*, 18(3):441–450.

Tsompana, M. and Buck, M. J. (2014). Chromatin accessibility: A window into the genome.

Turco, M. Y. and Moffett, A. (2019). Development of the human placenta. *Development (Cambridge)*, 146(22).

Varberg, K. M., Dominguez, E. M., Koseva, B., McNally, R. P., Moreno-Irusta, A., Wesley, E. R., Iqbal, K., Cheung, W. A., Okae, H., Arima, T., Lydic, M., Holoch, K., Marsh, C., Soares, M. J., Grundberg, E., and Varberg or Michael J Soares, K. M. (2022). Active remodeling of the chromatin landscape directs extravillous trophoblast cell lineage development. *medRxiv*, page 2022.05.25.22275520.

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J. E., Stephenson, E., Polański, K., Goncalves, A., Gardner, L., Holmqvist, S., Henriksson, J., Zou, A., Sharkey, A. M., Millar, B., Innes, B., Wood, L., Wilbrey-Clark, A., Payne, R. P., Ivarsson, M. A., Lisgo, S., Filby, A., Rowitch, D. H., Bulmer, J. N., Wright, G. J., Stubbington, M. J., Haniffa, M., Moffett, A., and Teichmann, S. A. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731):347–353.

Vu, H. T., Kaur, H., Kies, K. R., Starks, R. R., and Tuteja, G. (2023a). Identifying novel regulators of placental development using time-series transcriptome data. *Life Science Alliance*, 6(2).

Vu, H. T., Zhang, Y., Tuteja, G., and Dorman, K. S. (2023b). Unsupervised contrastive peak caller for atac-seq. *Genome Research*.

Vu, H. T. H., Scott, R. L., Iqbal, K., Soares, M. J., and Tuteja, G. (2023c). Core conserved transcriptional regulatory networks define the invasive trophoblast cell lineage. *Development*.

Wang, Y. and Zhao, S. (2010). Cell Types of the Placenta. In *Vascular Biology of the Placenta*, chapter 4. Morgan & Claypool Life Sciences.

Whalen, S., Truty, R. M., and Pollard, K. S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics 2016 48:5*, 48(5):488–496.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., and Andrews, S. (2015). Hicup: pipeline for mapping and processing hi-c data. *F1000Research*, 4.

Woods, L., Perez-garcia, V., and Hemberger, M. (2018). Regulation of Placental Development and Its Impact on Fetal Growth — New Insights From Mouse Models. *Frontiers in Endocrinology*, 9(September):1–18.

Wu, D.-Y., Bittencourt, D., Stallcup, M. R., and Siegmund, K. D. (2015). Identifying differential transcription factor binding in chip-seq. *Frontiers in genetics*, 6:169.

Yan, F., Powell, D. R., Curtis, D. J., and Wong, N. C. (2020). From reads to insight: A hitchhiker's guide to ATAC-seq data analysis.

Yi, L., Pimentel, H., Bray, N. L., and Pachter, L. (2018). Gene-level differential analysis at transcript-level resolution. *Genome Biology*, 19(1).

Zeldovich, V. B., Clausen, C. H., Bradford, E., Fletcher, D. A., Maltepe, E., Robbins, J. R., and Bakardjiev, A. I. (2013). Placental Syncytium Forms a Biophysical Barrier against Pathogen Invasion. *PLoS Pathogens*, 9(12):1–10.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137.

Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., Ding, B., Li, N., Zheng, L., and Wang, W. (2016). Constructing 3d interaction maps from 1d epigenomes. *Nature communications*, 7(1):10812.

CHAPTER 2. IDENTIFYING NOVEL REGULATORS OF PLACENTAL DEVELOPMENT USING TIME SERIES TRANSCRIPTOMIC DATA AND NETWORK ANALYSES

Ha T. H. Vu[1,2], Haninder Kaur[1], Kelby R. Kies[1,2], Rebekah R. Starks[1,2], Geetu Tuteja[1,2]

[1]Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA

[2]Bioinformatics and Computational Biology, Iowa State University, Ames, IA 50011, USA

Modified from a manuscript published in *Life Science Alliance*

2.1 Abstract

The placenta serves as a connection between the mother and the fetus during pregnancy, and provides the fetus with oxygen, nutrients, and growth hormones. However, the regulatory mechanisms and dynamic gene interaction networks underlying early placental development are understudied. Here, we generated RNA sequencing (RNA-seq) data from mouse fetal placenta tissues at embryonic day (e) 7.5, e8.5 and e9.5 to identify genes with timepoint-specific expression, then inferred gene interaction networks to analyze highly connected network modules. We determined that timepoint-specific gene network modules associated with distinct developmental processes, and with similar expression profiles to specific human placental cell populations. From each module, we obtained hub genes and their direct neighboring genes, which were predicted to govern placental functions. We confirmed that four novel candidate regulators identified through our analyses regulate cell migration in the HTR-8/SVneo cell line. Upon conclusion of this study, we were able to predict several novel regulators of placental development using network analysis of bulk RNA-seq data. Our findings and analysis approaches will be valuable for future studies investigating the transcriptional landscape of early placental development.

2.2 Introduction

The placenta is a transient organ that has critical roles during pregnancy, such as the transportation of oxygen and nutrients to the fetus, waste elimination, and the secretion of growth hormones. Placental defects are associated with devastating complications including preeclampsia and fetal growth restriction, which can lead to maternal or fetal mortality (Bamfo and Odibo, 2011; Rana et al., 2019). Therefore, it is fundamental to understand the mechanisms of placental development.

Due to ethical considerations as well as the opportunity for genetic manipulation, mouse models are frequently used when investigating early placental development. Like humans, mice have a hemochorial placenta (Hemberger et al., 2020), meaning that maternal blood directly comes in contact with the chorion. Although there are certain differences between the mouse and human placenta (Hemberger et al., 2020; Soncin et al., 2015), they do express common genes during gestation, including common regulators and signaling pathways involved in placental development (Cox et al., 2009; Soncin et al., 2018; Watson and Cross, 2005). For example, *Ascl2/ASCL2* and *Tfap2c/TFAP2C* are required for the trophoblast (TB) cell lineage in both mouse and human models (Guillemot et al., 1994; Kuckenberg et al., 2012; Varberg et al., 2021). Another example is the HIF signaling pathway, which regulates TB differentiation in both mouse and human placenta (Soncin et al., 2015).

Mouse placental development begins around embryonic day (e) 3.5 when the trophectoderm (TE) layer forms (Watson and Cross, 2005). The TE differentiates into different TB populations at e4.5, which eventually leads to the formation of the ectoplacental cone (EPC) (Bevilacqua et al., 2014). Between e7.5 and e9.5, the establishment of blood flow to the fetus begins, and highly dynamic changes in placental cell composition occur. At e7.5, the EPC is comprised of TB cells (Hemberger et al., 2020), organized into the inner and peripheral populations, with the inner cells actively proliferating and differentiating, while the outer cells can be invasive and interact with the decidua (Bevilacqua et al., 2014). Around e8.5, chorioallantoic attachment occurs, during which the chorion layer joins with the allantois (Cross et al., 2003). As a result, the e8.5

mouse fetal placenta includes cells from the EPC, chorion, and allantois (Cross et al., 2006). From e9.5 onwards, the mouse fetal placenta is composed of distinct layers, the trophoblast giant cell (TGC) layer, the junctional zone (spongiotrophoblast and glycogen TB cells), and the labyrinth zone (chorion TB cells, syncytiotrophoblast I and II cells, fetal endothelium, and spiral artery TGCs) (Simmons, 2014; Walentin et al., 2016). Within the labyrinth layer, there is a dense network of vasculature where nutrients and oxygen are transported and exchanged. Although the structure of the placenta is not identical between mouse and human, certain mouse placental cell types are thought to be equivalent to human placental cell types (Soncin et al., 2015). For example, parietal TGCs and glycogen TBs have been described as equivalent to human extravillous trophoblasts (EVTs) (Soncin et al., 2015). Mouse TGCs are not as invasive as human EVTs (Soncin et al., 2015), and they have different levels of polyploidy and copy number variation (Morey et al., 2021); however, both EVTs and TGCs are able to degrade extracellular matrix to enable TB migration into the decidua (Silva and Serakides, 2016).

Several individual regulators of the processes active between e7.5 and e9.5 have been identified, as reviewed in (Cross, 2005; Hemberger and Cross, 2001; Hu and Cross, 2010; Rossant, 2001; Watson and Cross, 2005). In addition, it is important to determine how these regulators potentially interact with other genes as networks. To identify novel regulators or infer gene interactions underlying developmental processes, unbiased whole genome transcriptomic data can be used. Previous studies that utilized transcriptomics in the developing mouse placenta were either focused on analysis of one timepoint, or focused on analysis of multiple -omics data (Abdulghani et al., 2019; Starks et al., 2020, 2019; Tuteja et al., 2016). Other studies of gene expression in human placenta across trimesters did not infer full gene interaction networks and instead focused on transcription factors (Morey et al., 2021; Prater et al., 2021). Single-cell analysis has been used to investigate cell-type specific gene expression in the placenta; however, these studies do not predict regulators underlying specific placental development processes (Marsh and Blelloch, 2020; Vento-Tormo et al., 2018).

Here, we generated RNA sequencing (RNA-seq) data from mouse fetal placental tissues at e7.5, e8.5, and e9.5. We then carried out clustering, differential expression, and network analyses to infer gene interactions and predict novel regulators of placental development. We further demonstrated that our network constructions could be used to infer cell populations in the mouse placenta at the three timepoints. Finally, we conducted in vitro validation experiments and confirmed that several genes we identified have a role in regulating TB cell migration.

2.3 Results

2.3.1 Genes associated with distinct placental processes show timepoint-specific expression

We generated and analyzed transcriptomic data from fetal placental tissues at e7.5, e8.5 and e9.5 to identify genes regulating distinct processes during placental development. Based on the stages of placental development and the cell types present at each stage, we predicted that genes with highest expression at e7.5 would be involved in TB proliferation or differentiation; genes with highest expression at e8.5 would have a role in chorioallantoic attachment; and genes with highest expression at e9.5 would have a role in the establishment of nutrient transport. Indeed, we observed that previously identified regulators of TB proliferation and differentiation (e.g., *Ascl2* (Guillemot et al., 1994; Tanaka et al., 1997), *Gjb5* (Kibschull et al., 2014)), chorioallantoic attachment (e.g., *Ccnf* (Tetzlaff et al., 2004), *Itga4* (Yang et al., 1995)), and nutrient transport (e.g., *Gjb2* (Gabriel et al., 1998), *Igf2* (Sferruzzi-Perri et al., 2011)) showed timepoint-specific patterns that matched with our predictions (Fig 2.1A). Next, we performed hierarchical clustering to determine if protein-coding transcripts would cluster into groups that displayed timepoint-specific expression. From this analysis, we obtained three groups of transcripts in which the median expression was highest at e7.5 (8242 transcripts, equivalent to 5566 genes), e8.5 (8091 transcripts, equivalent to 5536 genes) and e9.5 (7238 transcripts, equivalent to 5347 genes) (Fig 2.1B, Supplementary Table S1). Hereafter, these groups are referred to as hierarchical clusters.

To evaluate the computational robustness and biological significance of the hierarchical clusters, we carried out additional analyses. First, we used three different algorithms, K-means clustering, self-organizing maps, and spectral clustering, to validate the trends of the expression levels in hierarchical groups, as well as the number of transcript groups (k = 3, 4 and 5). Only with k = 3 did we obtain groups with median expression level trends consistent in all four algorithms (Fig S1). Moreover, with k = 3, the maximum percent of agreement (see Materials and methods) between hierarchical clusters and clusters obtained using each of the different algorithms was 70.34-87.26% (Fig S1), while the maximum percent of agreement between hierarchical clusters and clusters obtained from other algorithms decreases to between 55.67 - 65.72% with k = 4, and 54.81 - 59.19% with k = 5. Second, we compared our hierarchical groups with previously published mouse and human placental microarray time course data from Soncin et al., 2018 (Soncin et al., 2018). Despite the technical differences between the datasets, we observed a consensus that our e7.5 hierarchical cluster had the highest percent of overlap with Soncin et al. gene groups that are downregulated over time, and our e9.5 hierarchical cluster had the highest percent of overlap with Soncin et al. gene groups that either have highest expression at e9.5 - e12.5 or genes that are upregulated over time (Supplementary Table S1).

Lastly, we determined how the genes in each cluster relate to processes of placental development. From previously published review articles (Cross, 2005; Hemberger and Cross, 2001; Hu and Cross, 2010; Rossant, 2001; Watson and Cross, 2005), we acquired gene sets associated with distinct processes, namely ectoplacental cone and/or spongiotrophoblast maintenance (expected to be most active at e7.5, when the EPC is still in a highly proliferative state (Bevilacqua et al., 2014)), trophoblast giant cell differentiation (expected to be more active at e7.5 because the mouse placenta at e8.5 and e9.5 includes more differentiated TB subtypes (Cross et al., 2006; Hemberger et al., 2020; Simmons, 2014; Walentin et al., 2016)), chorioallantoic attachment (expected to be most active at e8.5 (Cross et al., 2003)), and labyrinth branching, vascularization and syncytiotrophoblast differentiation (expected to be most active at e9.5, after these processes have initiated (Rossant, 2001)) (Supplementary Table S1). Indeed, we observed

that the e7.5 hierarchical cluster captured the most genes in the ectoplacental cone and spongiotrophoblast maintenance and trophoblast giant cell differentiation group; the e8.5 hierarchical cluster included the most genes in the chorioallantoic attachment group; and the e9.5 hierarchical cluster included the most genes in the labyrinth branching, vascularization and syncytiotrophoblast differentiation group (Fig 2.1C, Supplementary Table S1). Together, these data demonstrate that hierarchical clustering can be used to obtain transcript groups that are associated with relevant biological processes at each timepoint, but is not sufficient to fully distinguish processes that may have varied activity levels throughout time.

To this end, and because hierarchical clustering is sensitive to small perturbations in the datasets (Jiang et al., 2004), we carried out differential expression analysis (DEA) and identified transcripts and genes with the strongest changes over time (Fig S2, Supplementary Table S2). After combining results from hierarchical clustering and DEA, we defined timepoint-specific gene groups (see Materials and methods for gene group definitions, Fig S2) and obtained 922 e7.5-specific genes, 915 e8.5-specific genes, and 1952 e9.5-specific genes (Supplementary Table S3). Gene ontology (GO) analyses showed that the timepoint-specific gene groups were enriched for highly relevant biological processes such as "trophoblast giant cell differentiation" (e7.5-specific genes), "labyrinthine layer development" (e8.5- and e9.5-specific genes), "blood vessel development" (e7.5- and e9.5-specific genes) and "response to nutrient" (e9.5-specific genes) (Supplementary Table S3).

It is possible that timepoint-specific groups share genes that have timepoint-specific transcripts. Indeed, we identified 37 genes shared between e7.5 and e8.5, 5 genes shared between e7.5 and e9.5, and 109 genes shared between e8.5 and e9.5 (Supplementary Table S3). We found that genes only present at one timepoint (timepoint-unique genes) were generally enriched for similar terms as the full group of timepoint-specific genes (Supplementary Table S3). However, terms related to the development of labyrinth layer like "labyrinthine layer morphogenesis" and "labyrinthine layer blood vessel development" were only enriched when using all e8.5-specific genes but not when using e8.5 timepoint-unique genes. Moreover, we found that, unlike genes

shared between e9.5 and e7.5, genes shared between e9.5 and e8.5 were enriched for processes such as “blood vessel development” and “insulin receptor signaling pathway”. This observation may indicate that different transcripts of the same genes could be expressed at different timepoints for the continuation of certain biological processes.

2.3.2 Network analysis reveals potential regulators of developmental processes in the placenta

To predict interactions amongst timepoint-specific genes and subset timepoint-specific genes into regulatory modules, we used the STRING database (Szklarczyk et al., 2019) and GENIE3 (Huynh-Thu et al., 2010) (see Materials and methods). With the two approaches of network inference, we were able to predict networks of genes by means of previously published experimental results and text-mining of available publications (STRING), as well as de novo computational analysis with random forest-based methods (GENIE3). We then carried out network sub-clustering with the GLay algorithm (Su et al., 2010) (see Materials and methods) and identified four network modules at e7.5, six at e8.5, and eight at e9.5 (Supplementary Table S4, Fig S3). To determine if the networks were associated with distinct processes of placental development, we used GO enrichment analysis.

Compared to e8.5 and e9.5 networks, e7.5 networks had a higher rank or fold change and were significantly enriched for the GO terms “inflammatory response” (e7.5_1_STRING: $-\log_{10}$(q-value) = 22.82 and e7.5_2_GENIE3: $-\log_{10}$(q-value) = 3.95) and “female pregnancy” (e7.5_2_GENIE3: $-\log_{10}$(q-value) = 4.1) (Fig 2.2A, Supplementary Table S5). The term “morphogenesis of a branching structure”, which can be expected following chorioallantoic attachment around e8.5, was not enriched at e7.5, but was enriched in multiple e8.5 and e9.5 networks (e8.5_1_STRING: $-\log_{10}$(q-value) = 1.73, e8.5_2_GENIE3: $-\log_{10}$(q-value) = 1.72, e9.5_1_STRING: $-\log_{10}$(q-value) = 4.01, e9.5_1_GENIE3: $-\log_{10}$(q-value) = 1.54, e9.5_2_STRING: $-\log_{10}$(q-value) = 14.33, and e9.5_2_GENIE3: $-\log_{10}$(q-value) = 2.2). After chorioallantoic attachment is complete, nutrient transport is being established. Accordingly, we

observed the following enrichments: "endothelial cell proliferation" (highest ranked in e9.5_2_STRING: $-\log_{10}$(q-value) = 15.91), "lipid biosynthetic process" (only significant after e7.5, highest ranked in e9.5_3_STRING: $-\log_{10}$(q-value) = 17.63), "cholesterol metabolic process" (only significant after e7.5, highest ranked in e9.5_2_GENIE3: $-\log_{10}$(q-value) = 2.76 and e9.5_3_STRING: $-\log_{10}$(q-value) = 7.79), and "response to insulin" (only significant after e7.5, highest ranked in e9.5_1_GENIE3: $-\log_{10}$(q-value) = 1.67). "Placenta development", "vasculature development" and cell migration related terms ("positive regulation of cell migration" and "epithelium migration") are observed in networks at all timepoints, although these terms are not consistently ranked in all networks. Using randomization tests, we observed the majority of these GO terms (10 out of 11 terms) were significantly enriched when using the network genes but not random gene sets (significance level of 0.05; the term "vasculature development" having p-value = 0.0549 and 0.0575 with subnetwork e9.5_1_GENIE3 and e9.5_3_GENIE3, respectively) (see Materials and methods, Fig S3). This analysis demonstrates that the network genes were highly relevant to the biological functions of interest. Moreover, the observed GO terms strongly aligned with the processes enriched when using the full lists of timepoint-specific genes (Supplementary Table S3), indicating the representative characteristics of the network genes. While the current analysis focuses on the biological processes related to placental development, there are other terms significantly enriched, which can be found in Supplementary Table S5. In summary, we identified 18 subnetworks across three timepoints for downstream analyses, some of which were enriched, according to GO analysis and randomization tests, for specific terms relating to placental development (Fig 2.2A).

We predicted that hub genes, defined to be nodes with high degree, closeness, and shortest path betweenness centrality in the networks (see Materials and methods), could be potential regulators of developmental processes in the placenta. We first determined if the hub genes from each network with enriched GO terms described in the previous paragraph were directly annotated or possibly related to placental functions using the Mouse Genome Informatics (MGI) database (Smith and Eppig, 2009) (see Materials and methods, Table 2.1, Supplementary Table

S6). Briefly, genes annotated under any GO or MGI phenotype terms related to placenta, TB cells, TE and the chorion layer are considered as having a "known" role in the placenta. Genes annotated under terms related to embryo are considered as having a "possible" role in the placenta, because embryonic lethal mouse knockout lines frequently have placentation defects, and because defects in placental development can be associated with the development of other embryonic tissues (Brown and Hay, 2016; Perez-Garcia et al., 2018; Woods et al., 2018). Hereafter, such genes are referred to as "known/possible genes". In the e7.5 networks, there were 17 hub genes in which six genes were known/possible. The number of hub genes that are labelled as known/possible is statistically significant when comparing to random gene sets selected from the e7.5 timepoint-specific group (Fig S3). In the e8.5 and e9.5 networks, 17 out of 28 and 48 out of 127 hub genes were known/possible, respectively. Similar to e7.5, the number of hub genes labelled as known/possible in e8.5 networks and e9.5 networks were both statistically significant when comparing to random gene sets selected from the corresponding timepoint-specific groups (Fig S3). These results indicate that the gene sets we identified are significantly associated with relevant phenotypes in the mouse.

In the network e7.5_1_STRING, we identified seven hub genes (Table 2.1), one of which (*Mmp9*) was considered as possibly related to placenta development according to the MGI database. Although *Mmp9* was not annotated directly to placenta using our definition of a placenta term on the MGI database, it has been shown to be required for proper implantation, TB differentiation and invasion (Plaks et al., 2013). In the network e7.5_2_GENIE3 (Fig 2.2B), ten hub genes were identified, five of which were annotated as regulating or having possible roles in placental development in the MGI database. Four of the genes are required for TB proliferation, differentiation, migration or invasion, namely *Nr2f2* (Hubert et al., 2010), *Prdm1* (Mould et al., 2012), *Hmox1* (Zenclussen et al., 2014) and *Ctbp2* (Hildebrand and Soriano, 2002) (Table 2.1, Supplementary Table S6). Other hub genes could be novel regulators of placental functions. One example is *Frk*, a hub gene of the e7.5_2_GENIE3 network, which has been suggested to inhibit cell migration and invasion in human glioma (Shi et al., 2015) and retinal

carcinoma cells (Yamada et al., 2020), but has not been studied in early placental development. Although the networks inferred by the two methods did not share any hub genes, hub genes identified with one method could be members of the other method's networks. These hub genes are *Mmp9* (e7.5_1_STRING), *Frk*, *Hmox1*, and *Nr2f2* (e7.5_2_GENIE3) (Table 2.1). This observation strengthens the potential roles of Frk gene in placental development.

At e8.5, hub genes included both novel and known genes in placental development or chorioallantoic attachment. For example, in the network e8.5_1_STRING, 11 hub genes were identified, eight of which were known/possibly associated with placental development according to the MGI database. For example, the genes *Akt1*, *Mapk1*, and *Mapk14* have a role in placental vascularization (Adams et al., 2000; Hatano et al., 2003; Yang et al., 2003) (Table 2.1, Supplementary Table S6). For the network e8.5_2_GENIE3 (Fig 2.2B), there were 17 hub genes identified (Table 2.1, Supplementary Table S6), with nine genes having known or possible functions. E8.5_2_GENIE3 network's hub genes include genes that are not known/possible, but have been studied in the context of placental development such as *Vgll1* (Soncin et al., 2018). Hub genes identified with one method and present in the other method's networks are *Hsp90aa1*, *Akt1*, and *Mapk14* (e8.5_1_STRING), *Dvl3* and *Msx2* (e8.5_2_GENIE3) (Table 2.1). An example of a novel gene is *Jade1* (hub node of e8.5_2_GENIE3), which has been found to have high expression in extraembryonic ectoderm and TB cells and hence may play roles in placental vascularization by interacting with VHL (Tzouanacou et al., 2003), but has not been tested functionally in placental tissues.

From the e9.5 networks, we identified 127 hub genes of which 48 have known/possible functions in placental development in the MGI database (Table 2.1, Supplementary Table S6). For instance, in e9.5_1_GENIE3, e9.5_2_STRING, and e9.5_4_STRING, hub genes that regulate labyrinth layer development include *Egfr* (Lee and Threadgill, 2009) and *Rb1* (Sun et al., 2006), and hub genes that regulate placental vasculature development include *Fn1* (George et al., 1993) and *Vegfa* (Li et al., 2014). Hub genes identified with one method and present in the other method's networks include important genes such as Rb1 (Sun et al., 2006), *Yap1* (Meinhardt

et al., 2020) (e9.5_1_GENIE3) and *Vegfa* (e9.5_2_STRING) (Table 2.1). Notably, *Vegfa* is the only hub gene identified with both network inference methods. There are also hub genes known to be important for placental nutrient transport such as *Igf2* (Sferruzzi-Perri et al., 2011), and other genes that could be novel regulators. For example, *Lhx2* is part of the mTOR signaling pathway in osteosarcoma (Song et al., 2019), but has yet to be studied in placenta although the mTOR signaling pathway is known to be involved in nutrient transport in the placenta (Lager and Powell, 2012).

In summary, we have identified hub genes in networks at each timepoint. Analyzing the annotations of hub genes using the MGI database demonstrated that the hub genes are biologically relevant to mouse development and will be strong candidates for future investigation.

2.3.3 Timepoint-specific genes can be associated with cell-specific expression profiles of human placenta

To determine if timepoint-specific genes could capture different placental cell populations, we carried out deconvolution analysis with LinSeed (Zaitsev et al., 2019) and inferred the cell type profiles. Briefly, LinSeed takes advantage of the mutual linearity relationships between cell-specific genes and their corresponding cells to infer the topological structures underlying cell populations of tissues. This approach would enable us to use bulk RNA-seq data to predict proportions of cell types in the mouse placenta without prior knowledge of cell type markers or matching single-cell datasets. As input to LinSeed, we used the 5000 most highly expressed genes across all timepoints (expression in TPM), from which 1413 genes were found to be statistically significant for the inference models and thus used to conduct the deconvolution analysis (see Materials and methods, Fig S4). As a result, we observed five cell groups which captured 99% of the variance in the placenta tissue samples (Fig S4). Amongst these groups, e7.5 samples had the highest proportion of cell group 3, e8.5 samples had highest proportion of cell group 2, and e9.5 samples had highest proportion of cell group 5 (Fig 2.3A – left panel, Supplementary Table S7). Cell group 1 and cell group 4 did not have consistent cell proportions across biological replicates of a

single timepoint. The identification of these cell groups could have resulted from noise introduced by both biological and technical variation, which is challenging to overcome when using a small sample size or analyzing without prior knowledge in the deconvolution analysis. Therefore, we focused on cell groups 3, 2 and 5. We identified 100 markers (see Materials and methods) for cell group 3, 100 markers for cell group 2, and 41 markers for cell group 5. Interestingly, 95 of the 100 markers of cell group 3 are e7.5-specific genes, 45 out of 100 markers of cell group 2 are e8.5-specific genes, and 40 in the 41 markers of cell group 5 are e9.5-specific genes (Fig 2.3A – right panel, Supplementary Table S7). This indicates that the independent timepoint-specific gene analysis we performed in Section 2 could represent gene profiles of distinct cell populations.

To this end, we used the PlacentaCellEnrich webtool to annotate timepoint-specific genes with human placental cell types (Jain and Tuteja, 2021). At all timepoints, we observed enrichment suggesting the presence of TB cells. Specifically, the e7.5-specific genes were most significantly enriched for genes with EVT-specific expression ($\log_2$(fold) = 1.75, $-\log_{10}$(adj. p-value)= 4.18), but also had enrichment for syncytiotrophoblast (SCT) ($\log_2$(fold) = 1.1, $-\log_{10}$(adj. p-value)= 2.09); the e8.5-specific group was only enriched for genes that had villous cytotrophoblast (VCT)-specific expression ($\log_2$(fold) = 1.51, $-\log_{10}$(adj. p-value)= 2.36), and the e9.5-specific group had the highest enrichment for genes with fetal fibroblast-specific expression ($\log_2$(fold) = 2.04, $-\log_{10}$(adj. p-value)= 22.04) (Fig 2.3B, Fig S5). We note that the e9.5-specific group had enrichment for genes with cell-type specific expression in multiple cells, including endothelial cells ($\log_2$(fold) = 2.02, $-\log_{10}$(adj. p-value)= 18.66), VCT ($\log_2$(fold) = 1.5, $-\log_{10}$(adj. p-value)= 7.38), SCT ($\log_2$(fold) = 1.23, $-\log_{10}$(adj. p-value)= 6.93), and EVT ($\log_2$(fold) = 1.05, $-\log_{10}$(adj. p-value)= 3.05) (Fig 2.3B, Fig S5). Together, this demonstrates that our analysis is picking up on the diverse cell populations present at e9.5 compared to e7.5.

Motivated by the fact that cell-specific expression profiles for multiple human placental cell types are enriched at e7.5 and e9.5, we hypothesized that the gene network modules at each timepoint could capture specific cell populations. Indeed, PlacentaCellEnrich analysis on e7.5_2_GENIE3 network genes was significantly enriched for genes with EVT-specific expression

($\log_2$(fold) = 2.32, $-\log_{10}$(adj. p-value)= 1.67) (Fig 2.3C, Fig S5), but no longer with genes that have SCT-specific expression. E8.5_1_STRING and e8.5_2_GENIE3 were both enriched for genes with VCT-specific expression ($\log_2$(fold) = 2.35 and 3, $-\log_{10}$(adj. p-value)= 1.43 and 5.41, respectively). In addition to VCT-specific expression, e8.5_2_GENIE3 had enrichment for genes that had SCT-specific expression ($\log_2$(fold) = 2.19, $-\log_{10}$(adj. p-value)= 2.93) (Fig 2.3C, Fig S5). At e9.5, genes in the networks e9.5_1_GENIE3 and e9.5_3_STRING showed strong enrichment for TB-specific expression, such as in SCT and VCT. On the other hand, e9.5_2_GENIE3, e9.5_2_STRING, e9.5_3_GENIE3 and e9.5_4_STRING had strong enrichment for fetal fibroblast and endothelium expression profiles (Fig 2.3C, Fig S5). Importantly, randomization tests showed that the enrichment of cell type-specific genes were only significant in these subnetworks but not in random gene sets selected from corresponding timepoint hierarchical groups (Fig S6), which highlights the biological relevance of the gene network modules.

For genes in networks e7.5_1_STRING, e9.5_1_STRING, and e9.5_3_GENIE3, we did not observe any enrichment for fetal placental cells, possibly because not all genes in the networks are annotated in the 1st trimester dataset (Vento-Tormo et al., 2018) used when calculating cell enrichments in PlacentaCellEnrich. Therefore, we also used Placenta Ontology (Naismith and Cox, 2021), which carries out enrichment tests based on different datasets than those used in PlacentaCellEnrich. With e7.5_1_STRING, in agreement with previous analyses on e7.5-specific genes or genes in e7.5_2_GENIE3 network, we observed annotations related to EVT cells being enriched, such as "EVT > side population" ($\log_2$(fold) = 1.99 and false discovery rate (FDR) = 0.027), and "EVT > CT" ($\log_2$(fold) = 1.96, FDR = 0.028) (Supplementary Table S8). With e9.5_1_STRING, the term "EGFR+ VCT > ITGA2+ TB niche" was enriched ($\log_2$(fold) = 1.89, FDR = 0.023), meaning there are a significant number of genes in this network that were upregulated in EGFR+ VCT compared to the ITGA2+ proliferative TB niche in 1st trimester placenta. Similarly, with e9.5_3_GENIE3, we found the term "EGFR+ VCT > HLA-G+ EVCT" enriched ($\log_2$(fold) = 1.5, FDR = 0.043), which means there is a significant number of genes in this network that were upregulated in EGFR+ VCT compared to HGL-A+ proximal column

extravillous cytotrophoblast in 1st trimester placenta. In the other networks, Placental Ontology enrichment results generally agreed with PlacentaCellEnrich (Supplementary Table S8). Together, the PlacentaCellEnrich and Placenta Ontology analyses provide evidence that network analysis can be used to identify genes more likely associated with specific placental cell types.

In summary, we have demonstrated that the identification of timepoint-specific gene groups and densely connected network modules can be used to infer the cellular composition of bulk RNA-seq samples. We used independent human datasets from different sources to annotate the cell types in each timepoint's samples. As a result, from the bulk RNA-seq data we were able to observe that at e7.5 and e8.5, there was a high proportion of different TB populations, whereas at e9.5, the placental tissues consisted of multiple cell types such as TB, endothelial and fibroblast cells.

2.3.4 Gene knockdown provides further evidence for a role of network genes in the placenta

As described in Section 2, we identified hub nodes, and as a result also obtained genes directly connected to the hub nodes (Supplementary Table S6). Many of the genes (23 genes at e7.5, 208 genes at e9.5) had drastic expression changes over time (having at least one transcript with fold change ≥ 5 between e7.5 and e9.5) (Supplementary Table S9), which may be more likely to have regulatory roles specific to processes or cell types associated to each timepoint. However, there were several hub genes and genes directly connected to the hub nodes that were differentially expressed but had lower fold changes and showed high expression across all timepoints. We predict these highly expressed genes to be generally important for TB function and processes such as cell migration, a term that was associated with multiple timepoint specific networks (Fig 2.2A).

To investigate this further, we performed gene knockdown and migration assays for four candidate genes from four different networks in the HTR-8/SVneo cell line, an established model for studying TB migration (Graham et al., 1993; Hirschberg et al., 2021; Wang et al., 2021). From the lists of hub genes and their directly connected nodes (Supplementary Table S6), we obtained

genes that met the following criteria: having expression levels > 5 TPM in the mouse placenta transcriptome data we generated, having expression levels > 5 FPKM (fragments per kilobase of transcript per million of mapped reads) in human TB cell lines (Okae et al., 2018) and having expression levels > 20 TPM in HTR-8/SVneo cell line (Starks et al., 2020) (Supplementary Table S6). From this list, we selected four genes: *Mtdh* and *Siah2* (from the e7.5_1_STRING and e7.5_2_GENIE3 network, respectively), *Hnrnpk* (from the e8.5_2_GENIE3), and *Ncor2* (from the e9.5_3_GENIE3), all of which were nodes in networks annotated as TB subtypes (see Section 3).

For each of the four genes we transfected two different siRNAs, and all eight siRNAs resulted in high knockdown efficiencies (74 – 93%, Fig 2.4A). Each pair of siRNAs similarly reduced target protein levels (Fig S7). Next, we performed cell migration assays and visually observed a reduction in cell migration capacity for all four genes (Fig 2.4B, Fig S7). To determine if the observed reduction in cell migration was statistically significant, we further quantified the integrated cell densities (Fig 2.4C, Supplementary Table S10). For *Siah2* and *Hnrnpk*, integrated densities of cells were significantly decreased upon knockdown with both siRNAs using a p-value ≤ 0.05. Specifically, for *Siah2*, the densities reduced by 98.57% ± 0.42% (mean ± standard error) and 83.87% ± 12.1% with siRNA #1 and siRNA #2, respectively. For *Hnrnpk*, the densities reduced by 99.55% ± 0.09% with siRNA #1 and 98.68% ± 0.2% for siRNA #2. For *Mtdh* and *Ncor2*, the reductions were significant for one siRNA (*Mtdh*, siRNA #2, 98.55% ± 0.86%; *Ncor2*, siRNA #1, 98.11% ± 0.09%), and were fair for the other siRNA, possibly due to the variable results between biological replicates (*Mtdh*, siRNA #1, 55.28% ± 17.22%; *Ncor2*, siRNA #2, 81.27% ± 14.04%). When comparing the number of cells 48 hours post-transfection for cells treated with target gene siRNA to cells treated with negative control siRNA, we determined that none of the target gene siRNA treatments resulted in significant changes in cell counts. We do note that *Siah2* siRNA #1 has some decrease in cell counts (p-value = 0.081), and *Ncor2* siRNA #1 and *Ncor2* siRNA #2 have some increase in cell counts (p-value = 0.081 and p-value = 0.077) compared to negative control treated samples (Fig S7). This provides evidence that, in general, the reduction in cell migration capacity was likely not due to the target gene impacting the rate

of cell death. Overall, these results confirm that network analysis and gene filtering based on defined criteria can identify genes important for TB function.

2.4 Discussion

Placental development involves multiple processes that are active during different stages of gestation. Using transcriptomic data generated from mouse placenta at e7.5, e8.5 and e9.5, we identified timepoint-specific gene groups that can be used for gene network inferences and analyses, as well as cell population annotations. Importantly, we were able to infer cell populations at different timepoints without known marker genes or reference dataset from the same species. The cell proportion inferences were necessary to bypass the confounding factors from cell heterogeneity, and thus predict more accurate novel regulators of cell-specific processes such as TB cell migration. This computational pipeline could be used to infer and analyze gene networks governing the development of placenta at other timepoints or to study developmental processes in other tissues.

Combining hierarchical clustering with differential expression analysis, we were able to identify gene groups using an unsupervised approach. It has also been shown that for times-series analyses with fewer than eight timepoints, pairwise differential expression analysis combined with additional methods identifies a more robust set of genes (Spies et al., 2019). Alternatively, model-based clustering using RNA-seq profiles (Si et al., 2014) could also be useful for gene group identification. However, it is still important to evaluate the robustness and functional relevance of the fitted models by carrying out additional downstream analyses.

We carried out DEA across the three timepoints on both the transcript and the gene level. These analyses revealed that a gene may have transcripts that are differentially expressed at different timepoints. For example, Igf2, a placental nutrient transport marker (Constância et al., 2002), has different transcripts grouped to e8.5 and e9.5 (Supplementary Table S1). This observation also aligned with a recent study which showed in 6–10 weeks' and 11–23 weeks' human placenta, differentially expressed genes, transcripts or differential transcript usage could

all assist in the understanding of placental development (Alfaidy et al., 2022). Therefore, in future studies, investigating roles of both genes and their transcripts could give a more complete functional profile at each timepoint. Moreover, our results, together with previous studies in human placenta (Alfaidy et al., 2022; Morey et al., 2021; Prater et al., 2021), suggest that time series transcriptomic analyses could be a useful approach to identify genes governing the development of the placenta. It will be beneficial to integrate these time series datasets to determine species-specific biomarkers of placental development.

We identified hub genes and their immediate neighboring genes which could regulate placental development and confirmed the roles of four novel genes (Mtdh, Siah2, Hnrnpk and *Ncor2*) in regulating cell migration in the HTR-8/SVneo cell line. These genes were selected primarily based on the network analyses, but also based on expression data from human cells to account for possible differences between mouse and human placental gene expression. Previous studies suggested these four candidates are functionally important in mouse. Mtdh has been suggested to regulate cell proliferation in mouse fetal development (Jeon et al., 2010). The Siah gene family is important for several functions (Qi et al., 2013). Of relevance to the placenta, Siah2 is an important regulator of HIF1α during hypoxia both in vitro and in vivo (Qi et al., 2008). Moreover, while Siah2 null mice exhibited normal phenotypes, combined knockouts of Siah2 and Siah1a showed enhanced lethality rates, suggesting the two genes have overlapping modulating roles (Frew et al., 2003). *Hnrnpk*$^{-/-}$ mice were embryonic lethal, and *Hnrnpk*$^{+/-}$ mice had dysfunctions in neonatal survival and development (Gallardo et al., 2015) . *Ncor2*$^{-/-}$ mice were embryonic lethal before e16.5 due to heart defects (Jepsen et al., 2007). According to the International Mouse Phenotyping Consortium database (Dickinson et al., 2016), *Ncor2* null mice also showed abnormal placental morphology at e15.5. However, none of these genes have been studied in the context of TB migration. We observed that while all siRNAs were able to decrease cell migration capacity, there was variability in the amount of decrease, even when comparing two siRNAs targeting the same gene. This observation did not seem to be associated with differences in transcript or protein knockdown levels and could be due to different off-target effects for

different siRNAs. Moreover, we observed that cell counts generally were not decreased upon target gene knockdown compared to negative control knockdown. However, more detailed analysis and process specific assays are needed. For example, future studies assessing each gene's role in cell adhesion, cell-cell fusion, cell proliferation and cell apoptosis can be done to better understand their roles in placental development. We also acknowledge the HTR-8/SVneo cell line bears certain differences to TB cells such as in their miRNA expression profiles (Donker et al., 2012). Therefore, in order to determine the exact roles of these genes in the placenta, future experiments in human TB stem cells derived with the Okae protocol (Okae et al., 2018) or gene knockout experiments in vivo are necessary.

Interestingly, all four genes have been shown to have roles in cancer cells: Siah2 was shown to promote cell invasiveness in human gastric cancer cells by interacting with ETS2 and TWIST1 (Das et al., 2016); Mtdh regulates proliferation and migration of esophageal squamous cell carcinoma cells (Yang et al., 2017); absence of Hnrnpk reduces cell proliferation, migration and invasion ability in human gastric cancer cells (Zhao Peng et al., 2019); and repression of NCOR2 and ZBTB7A increased cell migration in lung adenocarcinoma cells (Alam et al., 2018). This result further supports previous studies that show the comparability between placental cell migration and invasion, and tumor cell migration and invasion (Costanzo et al., 2018; Naismith and Cox, 2021), although specific genes may have different impacts on migration/invasion capacity such as with the *Ncor2* gene.

In our analyses, we observed that timepoint-specific genes and their networks represented expression profiles for specific placental cell populations at the three timepoints. In particular, analysis of e7.5-specific and e8.5-specific genes and networks showed that placental tissues at e7.5 and e8.5 contain different populations of TB cells, while e9.5-specific genes and networks showed multiple cell types including TB, endothelial and fibroblast cells. The significant overlap between e7.5-specific genes and genes of EVT cells yielded an interesting suggestion that the TB cell populations in e7.5 mouse placenta may share similarity in gene profiles to human EVT, although mouse TB and human EVT have certain differences such as their invasiveness levels (Soncin

et al., 2015) and levels of polyploidy and copy number variation (Morey et al., 2021). Examples of EVT genes present in e7.5-specific gene group include FSTL3 (downregulation decreased TB migration and invasion in JAR cell line (Xie et al., 2018)), ADM (increased TB migration and invasion in JAR and HTR-8/SVneo cell line (Zhang et al., 2005)), and ASCL2 (regulates TB differentiation (Guillemot et al., 1994)). Moreover, hub genes could be used to identify potential novel markers for the cell types corresponding to their subnetworks. For example, hub genes of subnetworks enriched for SCT-specific genes such as Dvl3 (e8.5_2_GENIE3) and Olr1 (e9.5_3_STRING) are not established SCT marker genes, but are in fact differentially expressed in SCT compared to human trophoblast stem cells, EVT (Sheridan et al., 2021) or endovascular TB (Gormley et al., 2021). In general, combining network analysis with existing gene expression data from single cell or pure cell populations will allow identification of novel cell-specific marker genes to help future studies focused on different TB populations.

While it is true that data at single-cell (sc) resolution is necessary to gain more insight into cell populations in heterogeneous tissues, these results showed strong evidence that bulk RNA-seq data could be used to infer the cell type composition. In addition, scRNA-seq assays could be noisier than bulk RNA-seq due to various technical aspects such as the amount of starting materials, cell size, cell cycle, and batch effects (Chen et al., 2019; Wang et al., 2018), which are difficult to correct (Andrews and Hemberg, 2018). Therefore, bulk RNA-seq, ideally in conjunction with scRNA-seq, is beneficial for the study of biological processes that involve multiple cell types. Nevertheless, we acknowledge that our deconvolution analysis and cell type annotations were limited due to the absence of matching scRNA-seq data, data from pure cell populations, or extensive cell marker lists. As these types of information become more available, deconvolution analysis can be used to identify species-specific cell types or correcting for confounding effects prior to DEA (Sutton et al., 2022).

In our network analysis, we observed that the GO term “inflammatory response” was enriched in e7.5_1_STRING (q-value = 1.52E-23), e7.5_2_GENIE3 (q-value = 0.00012) and e9.5_2_STRING (q-value = 4.17E-10) (Supplementary Table S5). The inflammatory process could be happening

in the placenta during e7.5 to e9.5 when TB cells actively invade the decidua (Woods et al., 2018) and create a pro-inflammatory environment (Mor et al., 2011). Another possibility is contamination from decidual cells, which could be detected when combining bulk and scRNA-seq (Suryawanshi et al., 2022). This further demonstrates the benefits of bulk and scRNA-seq data integration.

Upon conclusion of this study, we have shown that in the mouse placenta at e7.5, e8.5 and e9.5, genes with timepoint-specific expression patterns can be associated with distinct processes and cell types. The genes identified by timepoint-specific gene-network analysis could be interesting candidates for future studies focused on the understanding of placental development and placenta associated pregnancy disorders.

2.5 Materials and methods

2.5.1 RNA-seq library preparation and sequencing

Placenta tissue was collected from timed-pregnant CD-1 mice (Charles Rivers Labs) following the guidelines and protocol approved by Iowa State University Institutional Animal Care and Use Committee (IACUC), protocol number 18–350. Placenta samples were collected as previously described (Starks et al., 2021; Tuteja et al., 2016) at e7.5, e8.5, and e9.5 and the age of the embryo was determined by following the embryonic development guidelines (Theiler, 1989). Briefly, tissues from the ectoplacental cone (EPC) and chorion were separated from the decidua, yolk sac, umbilical cord, and embryo, and then collected. For e7.5, 12 EPCs were collected and pooled into one replicate, as described in (Starks et al., 2019). For e8.5, five placentas were collected per replicate, and for e9.5, one placenta was collected per replicate. Each timepoint had a total of 6 biological replicates.

Tissues were processed for RNA isolation immediately after collection using the Purelink RNA micro scale kit (Thermofisher, 12183016). RNA concentration and RIN values were measured using the RNA 6000 Nano assay kit on the Agilent 2100 Bioanalyzer (GTF facility, ISU), and all samples had a RIN score $\geq$ 7.7 (Supplementary Table S11). Further processing of the samples,

library preparation and sequencing was performed by the DNA facility at Iowa State University. Libraries were sequenced using the Illumina HiSeq 3000 with single-end 50 base pair reads. The pooled library sample was run over two sequencing lanes (technical replicates for each sample).

2.5.2 RNA-seq data processing

The quality and adapter content were assessed using FastQC (version 0.11.7) (Andrews, 2010). Low quality reads and adapters were trimmed with Trimmomatic (version 0.39) (Bolger et al., 2014).

Technical replicates were then merged, and the reads were pseudo-aligned and quantified (in TPM) using Kallisto (version 0.43.1; l = 200, s = 30; b = 100) (Bray et al., 2016). Transcript sequences on autosomal and sex chromosomes of the mouse genome (GRCm38.p6) from Ensembl release 98 (Cunningham et al., 2019) were used to build the Kallisto index.

For further quality control, we carried out hierarchical clustering and principal component analysis (PCA) of samples. First, from the transcripts with raw counts ≥ 20 in ≥ 6 samples, we obtained the top 50% most variable transcripts, then centered and scaled their expression. Next, we implemented hierarchical clustering with the hclust() function in R (package stats (R Core Development Team, 2013), version 3.6.3), using the agglomerative approach with Euclidean distance and complete linkage. To implement PCA, we used the prcomp() function in R (package stats, version 3.6.3). We observed samples of each timepoint cluster close to each other and away from other timepoints. Outlier samples, which did not cluster with their respective timepoint groups, were removed prior to carrying out downstream analyses (Fig S8).

2.5.3 Cluster analysis

Before performing all clustering procedures, transcripts with low raw counts (mean raw counts < 20 in all timepoints) were filtered out, and expression data (in TPM) was scaled and re-centered. Hierarchical clustering, k-means clustering, self-organizing map and spectral clustering were performed on the top 75

We implemented hierarchical clustering with the hclust() function in R (package stats (R Core Development Team, 2013), version 3.6.3), using the agglomerative approach with Euclidean distance and complete linkage. The resulting dendrogram was cut at the second highest level to obtain three clusters. To test the robustness of the clustering assignments, we also carried out clustering with the number of clusters as four and five. K-means clustering was carried out using the R function kmeans() (centers = 3, 4 and 5, other parameters: default; package stats, version 3.6.3).

Self-organizing map clustering was performed with the R function som() with rectangular 3×1, 4×1 and 5×1 grid (other parameters: default; package kohonen (Wehrens and Kruisselbrink, 2018), version 3.0.10).

To implement spectral clustering, we utilized the following functions in R: computeGaussianSimilarity() (sigma = 1) to compute similarity matrix, and spectralClustering() (K = 3, 4 and 5, other parameters: default; package RclusTool (Wacquet et al., 2013), version 0.91.3) to cluster.

The percent agreement between cluster assignments of different methods was quantified as
(number of transcripts in common between two clusters)/
(total number of transcripts in two clusters) $\times$ 100.

To determine how the genes in each cluster relate to specific processes of placental development, we obtained gene lists from previously published review articles (Cross, 2005; Hemberger and Cross, 2001; Hu and Cross, 2010; Rossant, 2001; Watson and Cross, 2005), then calculated the percentage of markers in hierarchical clusters as (number of markers in a cluster)/ (total number of markers of the process) $\times$ 100.

2.5.4 Differential expression analysis (DEA)

DEA at transcript and gene levels were carried out with Sleuth (version 0.30.0) (Pimentel et al., 2017) using the likelihood ratio test (default basic filtering) and the p-value aggregation process (Yi et al., 2018). Fold change of a transcript was calculated using its average raw TPM

across all samples. A transcript was considered differentially expressed (DE) if it had a fold change ≥ 1.5 and a q-value ≤ 0.05. A gene was considered DE if its q-value was ≤ 0.05 and had at least one protein-coding DE transcript. For lists of DE protein-coding transcripts that had at least one DE gene, and lists of DE genes with at least one DE protein-coding transcripts, see Supplementary Table S2.

2.5.5 Definition of timepoint-specific genes

Timepoint-specific gene groups are defined as the following:

- e8.5-specific transcripts: transcripts in e8.5 hierarchical cluster, are up-regulated at e8.5 (compared to e7.5) or are up-regulated at e8.5 (compared to e9.5). E8.5-specific genes are ones associated with e8.5-specific transcripts.
- e7.5-specific transcripts: transcripts in e7.5 hierarchical cluster, are up-regulated at e7.5 (compared to e9.5), and are not in e8.5-specific group. E7.5-specific genes are ones associated with e7.5-specific transcripts.
- e9.5-specific transcripts: transcripts in e9.5 hierarchical cluster, are up-regulated at e9.5 (compared to e7.5), and are not in e8.5-specific group. E9.5-specific genes are ones associated with e9.5-specific transcripts.

2.5.6 Network construction and analysis

The STRING database (version 11.0b) (Szklarczyk et al., 2019) was used to build protein – protein interaction networks at each timepoint. Edges from evidence channels: experiments, databases, text-mining and co-expression with confidence score ≥ 0.55 were chosen for further analyses.

Gene regulatory networks at each timepoint were constructed with GENIE3 (version 1.16.0) (Huynh-Thu et al., 2010). At each timepoint, as inputs for GENIE3, timepoint-specific transcripts with average TPM at the timepoint ≥ 5 were aggregated to obtain gene counts with

the R package tximport (version 1.14.2; countsFromAbundance = lengthScaledTPM) (Soneson et al., 2016). Genes that encode transcription factors (TFs) and co-TFs, downloaded from AnimalTFDB (version 3.0) (Hu et al., 2019), were treated as candidate regulators. Then, edges with weight $<$ the 90th percentile were filtered out.

Largest connected components of the networks were analyzed using Cytoscape (version 3.7.2) (Shannon et al., 2003). All networks were treated as undirected, and network sub-clustering was performed using the GLay plug-in (default parameters) (Su et al., 2010). Networks with $\geq$ 100 nodes were used for further analyses. Hub genes were defined as nodes that have degree, betweenness and closeness centralities in the 10th percentile of their networks. A gene was determined to have an annotated role in placental development if it was annotated under all GO and MGI Phenotype terms related to placenta, TB cells, TE and chorion layer. A gene was categorized as having possible roles in placental development if it was annotated under all GO and MGI Phenotype terms related to embryo. GO terms, MGI Phenotype terms and gene annotations were downloaded from MGI (http://www.informatics.jax.org/) (version 6.19) (Smith and Eppig, 2009). For lists of terms used, see Supplementary Table S6.

Randomization tests were carried out to determine if the number of known/possible hub genes at a timepoint is significant. For each timepoint, from the respective timepoint-specific groups, 10,000 gene sets of the same number as the hub gene numbers were sampled. Then the number of known/possible genes in each set were counted. A p-value was calculated as the number of times a random gene set has $\geq$ known/possible genes than the observed number, divided by 10,000.

2.5.7 Gene ontology (GO) analyses

To determine the relevant functions of the gene lists, we used gene ontology (GO) analysis. ClusterProfiler (version 4.0.5) (Yu et al., 2012) was used, with the mouse annotation from the org.Mm.eg.db R package (version 3.13.0) (Carlson, 2019), the maximum size of genes = 1000, and a q-value cut-off = 0.05. Next, a fold change for each term was calculated as GeneRatio/BgRatio. A GO term was considered enriched when its q-value $\leq$ 0.05, fold change $\geq$ 2, and the number of

observed genes ≥ 5. Hypergeometric test was used for enrichment following the suggestions in Rivals et al., 2007 (Rivals et al., 2007).

Randomization analysis was carried out to determine if a GO term is statistically significant for a subnetwork's genes. For each subnetwork, from the respective timepoint-hierarchical groups, 10,000 gene sets with the same size as the subnetwork were sampled. For each of the random sets, the q-value of a specific term with ClusterProfiler (same settings as above) was obtained. Then, the p-value of the randomization test was calculated as the number of random gene sets with q-values lower than the q-value of that term in the original subnetwork, divided by 10,000.

2.5.8 Deconvolution analysis

To infer the proportion of cell types across timepoints, we carried out deconvolution analysis using the R package LinSeed (version 0.99.2) (Zaitsev et al., 2019). Gene abundances (in TPM) used as inputs for the analysis were obtained using tximport (version 1.14.2; countsFromAbundance = lengthScaledTPM) (Soneson et al., 2016). Then, we used the top 5000 most expressed genes across timepoints, and sampled 100,000 times to test for the significance of the genes to be used for deconvolution analysis. A significant gene was one with p-value ≤ 0.05. The number of cell groups was determined after examining the singular value decomposition (SVD) plot, generated with the svdPlot() function in LinSeed. Cell markers were defined as the top 100 genes closest to the cell group's corner, and closer to the corner than any other corners.

2.5.9 Placenta Cell Enrichment and Placenta Ontology analysis

The PlacentaCellEnrich webtool (Jain and Tuteja, 2021) and Placenta Ontology (Naismith and Cox, 2021) were used to infer the relevant cell types using gene lists. For PlacentaCellEnrich, cell-type specific groups were based on single-cell transcriptome data of the first trimester human maternal-fetal interface from Vento-Tormo et al. (Vento-Tormo et al., 2018). A enrichment was considered significant if its adj. p-value is ≤ 0.05, fold change ≥ 2, and the number of associated genes found is ≥ 5. For Placenta Ontology, we obtained placenta ontology GMT file from

Naismith et al. and uploaded the file to the WEB-based GEne SeT AnaLysis Toolkit (www.webgestalt.org) (Liao et al., 2019) as a functional database. An ontology with FDR ≤ 0.05, fold change ≥ 2 and the number of observed genes ≥ 5 was considered enriched. To avoid duplication while sampling, only genes with one-to-one pairwise orthology were considered for the enrichment tests.

Randomization analysis was carried out to determine if the enrichment of human first trimester placenta cell type-specific genes is statistically significant for a subnetwork's genes. For each subnetwork, from the respective timepoint-hierarchical groups, 10,000 gene sets with the same size as the subnetwork were sampled. For each of the random sets, the adjusted p-value of a specific cell type enrichment with PlacentaCellEnrich (same settings as above) was obtained. Then, the p-value of the randomization test was calculated as the number of random gene sets with adjusted p-values lower than the adjusted p-value of that cell type in the original subnetwork, divided by 10,000.

2.5.10 In vitro validation experiments

Cell culture – HTR-8/SVneo (ATCC CRL3271) were cultured as recommended by ATCC and as done by others (Canfield et al., 2019). Briefly, cells were grown in RPMI-1640 media (ATCC 302001) supplemented with 5% FBS (VWR, 97068-085) without antibiotics. Cells were split every 3 to 4 days, at 80-90% confluency.

siRNA knockdown – HTR-8/SVneo cells were transfected with two different siRNAs for each target gene knockdown (KD). Cells were split at 80% confluency, and siRNA transfection was performed in 6-well plates; 150,000 cells/well were seeded (Starks et al., 2020). After 24 hours, cells were transfected with 30nM siRNA using RNAiMax 3000 (Thermofisher, 13778150). Media was replaced after 24 hours of transfection, then cells were collected after 48 hours of transfection, counted using the TC20 Automated Cell Counter (Bio-rad), and seeded for migration assays. The remaining cells were pelleted to isolate RNA using the Invitrogen RNA mini kit (Fisher Scientific, 12183018A). The RNA concentration was determined using a

Nanodrop, and 200ng of the RNA was converted to cDNA (Thermofisher, 4368814). KD efficiencies were checked by qPCR using primers listed in Supplementary Table S11. GAPDH was used for normalization of all four genes' expression (ΔCT). Percent KD were calculated with the $\Delta\Delta$CT method. SiRNA and primer information can be found in Supplementary Table S11.

Migration assays – Migration assays were performed using Costar inserts (Corning, 3464). The inserts were placed in a 24 well plate and 75,000 cells in serum-free RPMI media (ATCC, 30-2001) were directly seeded in the top chamber of the insert. The bottom chamber was filled with 600 µL of RPMI media supplemented with 10% FBS as a chemoattractant. The cells were allowed to migrate for 24 hours at 37°C. The cells on the bottom of the inserts were fixed in 4% PFA (Fisher Scientific, AAJ61899AK) for 5 min and then washed for 1 min with PBS twice. The cells in the top chamber were scraped off using a wet q-tip (Fisher Scientific, 22029488) and the cells on the bottom of the inserts were stained with Hematoxylin (Fisher Scientific, 23245677) for 24 hours. The inserts were washed twice in distilled water. The membrane was cut using a scalpel (Fisher Scientific, 1484002) and mounted on a clean glass slide in Vectamount mounting medium (Fisher Scientific, NC9354983). The cells were observed under a dissection microscope and imaged at 12.5X magnification. The images were analyzed using the ImageJ tool, and the integrated density was obtained for each image.

Western blot - Following siRNA KD, whole cell lysate (4x Laemmli protein sample buffer, BioRad, 1610747) or cytoplasmic extract (NE-PER extraction kit, Thermofisher, 78833) were resolved using Sodium dodecyl-sulfate polyacrylamide gel electrophoresis (SDS-PAGE) gel and transferred to the nitrocellulose membrane (BioRad, 1620113) using Trans-Blot Turbo transfer system (BioRad, 1704150). After protein transfer, membranes were blocked and probed with antibodies as listed in Supplementary Table S11.

Statistical analysis – Experiments were performed with three replicates per condition (negative control or knockdown) per gene. P-values were calculated with the one-sided Wilcoxon rank sum test to test for a significant decrease in cell migration, and the two-sided Wilcoxon rank sum test for cell count comparisons.

2.6 Main figures and table

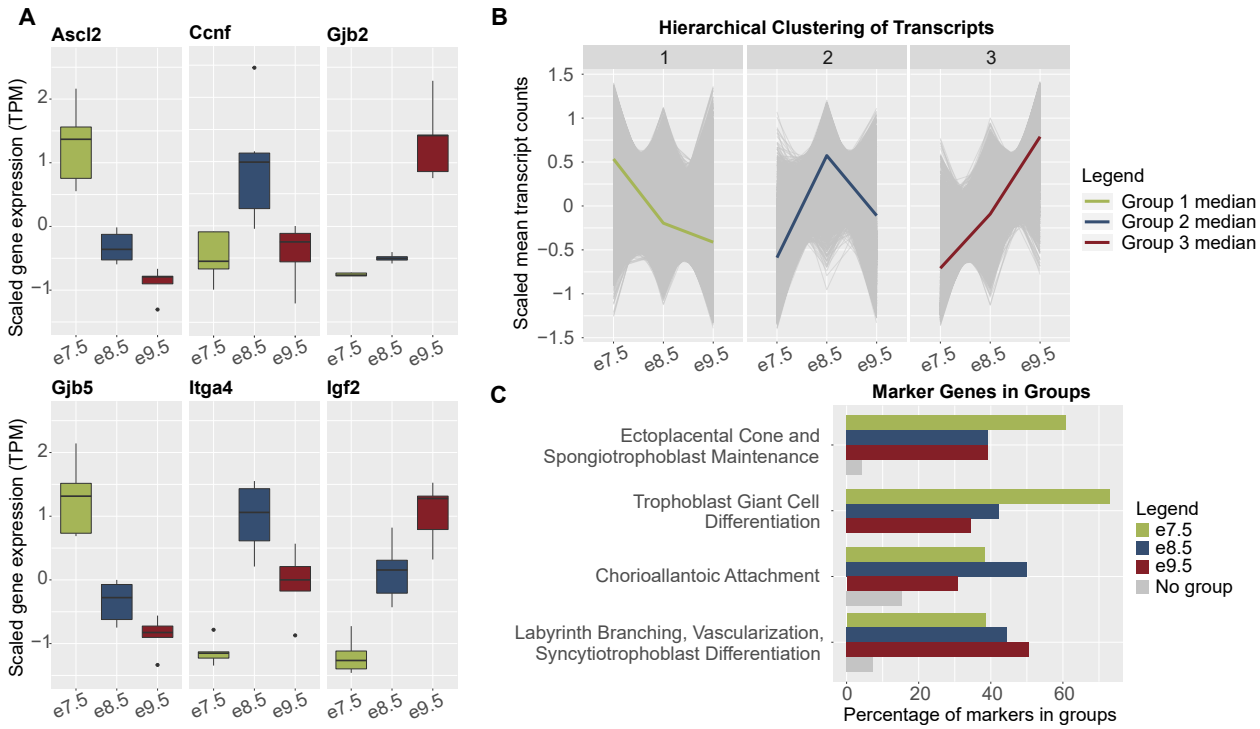

Figure 2.1: Gene associated with distinct placental processes show timepoint-specific expression. (A) Boxplots of scaled mean expression (in transcripts per million, TPM) of marker genes showing timepoint-specific patterns. *Ascl2* and *Gjb5*, expected to peak at e7.5, markers of trophoblast proliferation and differentiation (Guillemot et al., 1994; Kibschull et al., 2014); *Ccnf* and *Itga4*, expected to peak at e8.5, markers of chorioallantoic attachment (Tetzlaff et al., 2004; Yang et al., 1995); *Gjb2* and *Igf2*, expected to peak at e9.5, markers of nutrient transport (Gabriel et al., 1998; Sferruzzi-Perri et al., 2011). (B) Line charts of scaled mean raw counts of transcripts in hierarchical clusters showing group median expression levels peak at each timepoint.

Figure 2.1: (Continued)

(C) Bar plots showing that timepoint-associated hierarchical clusters captured most genes underlying distinct placental processes. Markers of timepoint-associated placental processes were obtained from previously published review articles (Cross, 2005; Hemberger and Cross, 2001; Hu and Cross, 2010; Rossant, 2001; Watson and Cross, 2005). Green, markers in hierarchical cluster with median expression level highest at e7.5; blue, markers in hierarchical cluster with median expression level highest at e8.5; dark red, markers in hierarchical cluster with median expression level highest at e9.5; grey, markers in no hierarchical clusters.

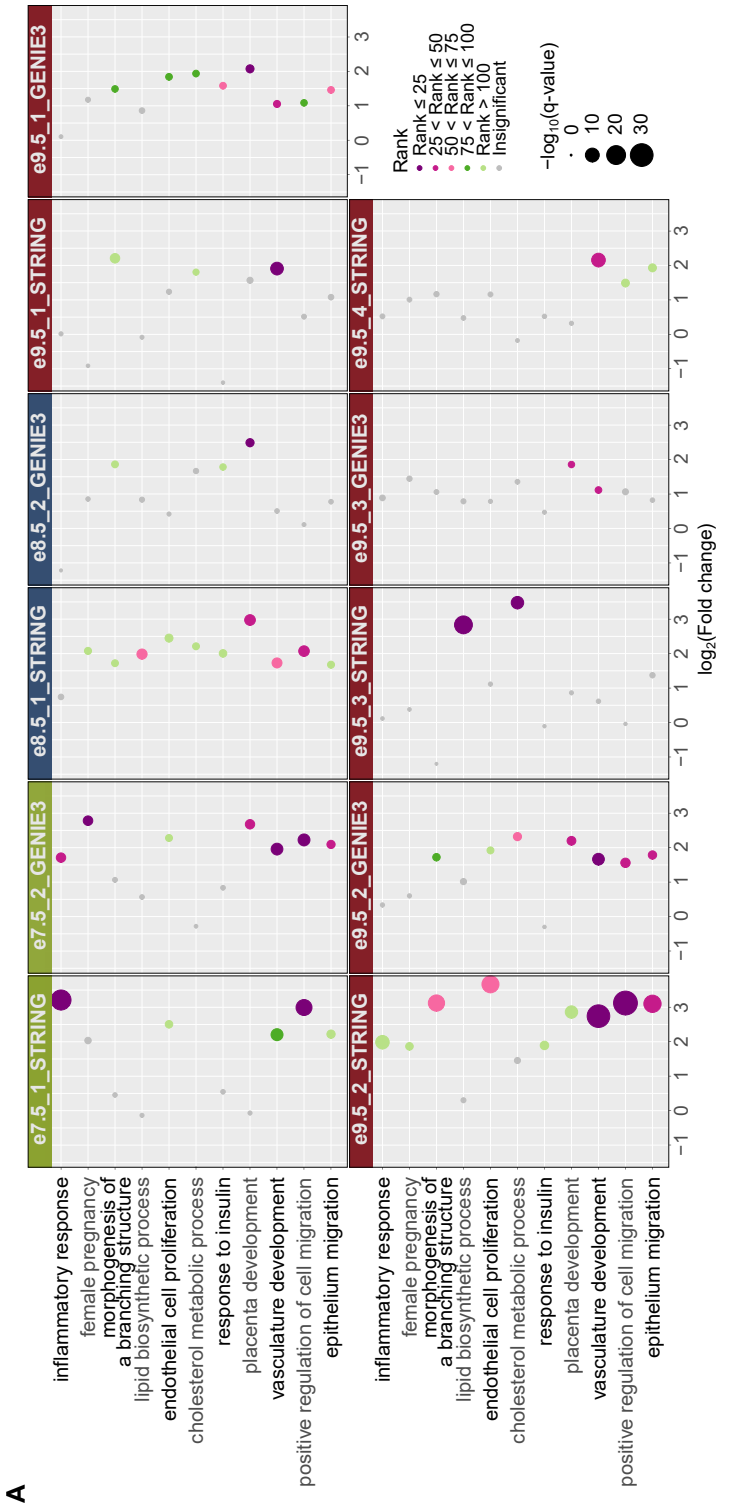

Figure 2.2: Network analysis identifies gene modules with relevant functions and reveals potential regulators of placental development.

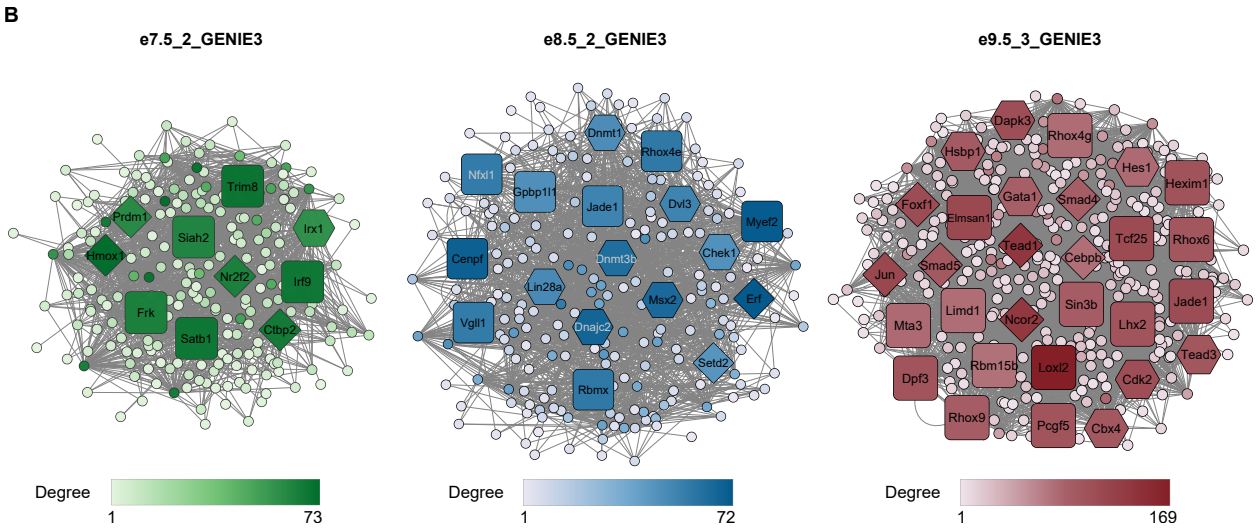

Figure 2.2: (Continued)

(A) Gene ontology (GO) analysis of networks demonstrates the association of gene sets with placental development processes. Only selected terms are shown. Dot colors correspond to ranks of the terms in each analysis; dot sizes correspond to $-\log_{10}$(q-value). A GO term is considered enriched if its q-value ≤ 0.05, fold change ≥ 2, and the number of observed genes ≥ 5. For full GO enrichment analysis, see Supplementary Table S5. (B) Network analysis highlights potential regulators of placental development. Only a subset of networks with enriched terms from (A) are shown. Diamond shape – hub genes with known roles in placenta; hexagon – hub genes with possible roles in placenta; rounded square – hubs without related annotation. Color: the darker the color is, the higher the node's degree centrality is. For visualization of all other subnetworks, see Fig S3.

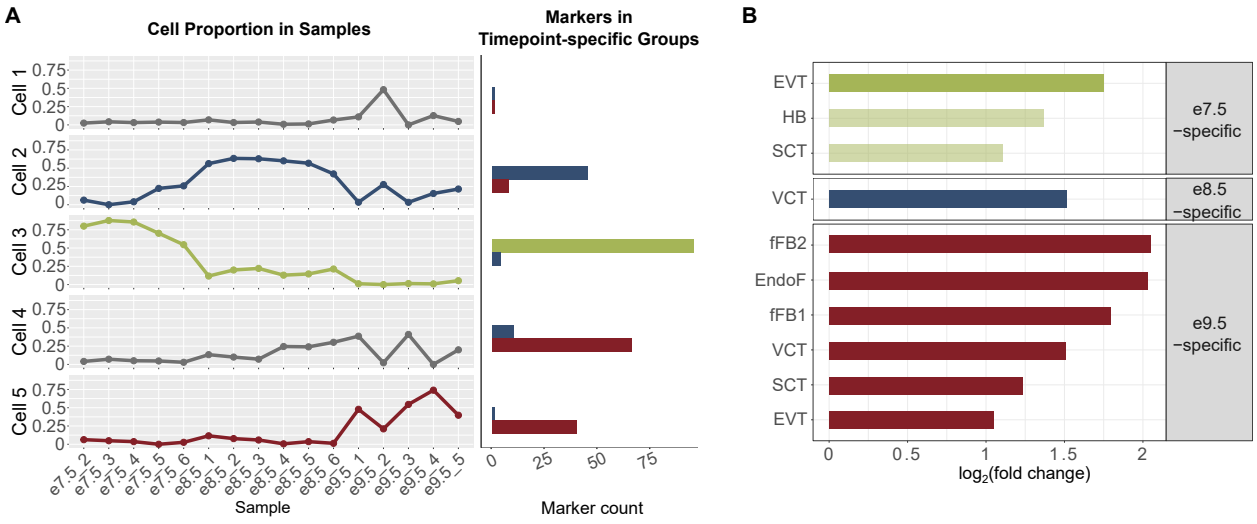

Figure 2.3: Timepoint-specific gene groups can be associated with human placenta cell-specific expression profiles.

(A) Deconvolution analysis using LinSeed showed five cell groups, three of which had highest proportions in e7.5 samples (group 3), e8.5 samples (group 2) and e9.5 samples (group 5). Also using LinSeed, we identified markers of each cell group and observed a high number of genes in common with timepoint-specific genes (cell group 3 with e7.5-specific genes, cell group 2 with e8.5-specific genes, cell group 5 with e9.5-specific genes). Left panel: line charts showing cell proportions in each sample; right panel: bar plots showing the number of cell markers in each timepoint-specific gene group. (B) Bar plots showing that timepoint-specific genes share similar profiles to these of human placental cell populations. Enrichment analysis was carried out with PlacentaCellEnrich using 1st trimester human placenta single-cell RNA-seq data to determine gene groups with cell-type specific expression. A significant enrichment has adj. p-value ≤ 0.05, fold change ≥ 2, and number of observed genes ≥ 5. The lightness of the colors corresponds to adj. p-value; the lighter colors, $0.005 <$ adj. p-value ≤ 0.05; the darker colors, adj. p-value ≤ 0.005. Only enrichments for cells of fetal origin are shown. Full enrichment results (including both maternal and fetal cells) are shown in Fig S5.

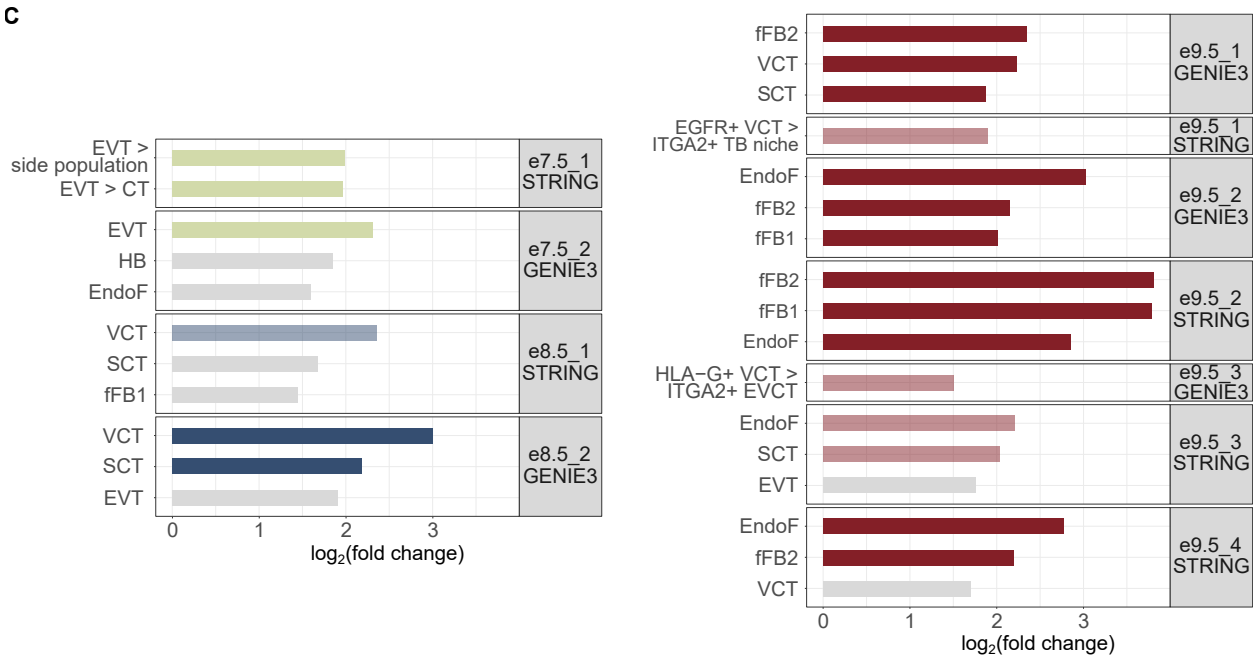

Figure 2.3: (Continued)

(C) Bar plots showing that network genes share similar profiles of specific human placental cell populations. Enrichment analysis was carried out with PlacentaCellEnrich as in (B) and Placental Ontology. Grey, adj. p-value > 0.05; the lighter colors, $0.005 <$ adj. p-value ≤ 0.05; the darker colors: adj. p-value ≤ 0.005. For PlacentaCellEnrich, three fetal cell types with the lowest adj. p-values are shown. For Placenta Ontology, selected enrichments are shown. Full enrichment results (including both maternal and fetal cells and for every network) of PlacentaCellEnrich are shown in Fig S5. Full enrichment results (for every network) of Placenta Ontology are in Supplementary Table S8. Abbreviations: SCT, syncytiotrophoblast, HB, Hofbauer cells, EVT, extravillous trophoblast, VCT, villous cytotrophoblast, EndoF, fetal endothelium, fFB1, fetal fibroblast cluster 1, fFB2, fetal fibroblast cluster 2, EVT > side population, GSE57834_extravillous_trophoblast_UP_side_population (genes upregulated in EVT compared to side population – original data from GSE57834),

Figure 2.3: (Continued)

EVT > CT, GSE57834_extravillous_trophoblast_UP_cytotrophoblast (genes upregulated in EVT compared to cytotrophoblast – original data from GSE57834), EGFR+ VCT > ITGA2+ TB niche, GSE106852_EGFR+_UP_ITGA2+ (genes upregulated in EGFR+ villous cytotrophoblast compared to ITGA2+ proliferative trophoblast niche, original data from GSE106852), EGFR+ VCT > HLA-G+ EVCT, GSE80996_EGFR+_villous_cytotrophoblast_UP_HLA_G+_proximal_column_extravillous_cytotrophoblast (genes upregulated in EGFR+ villous cytotrophoblast compared to HLA-G+ proximal column extravillous cytotrophoblast, original data from GSE80996).

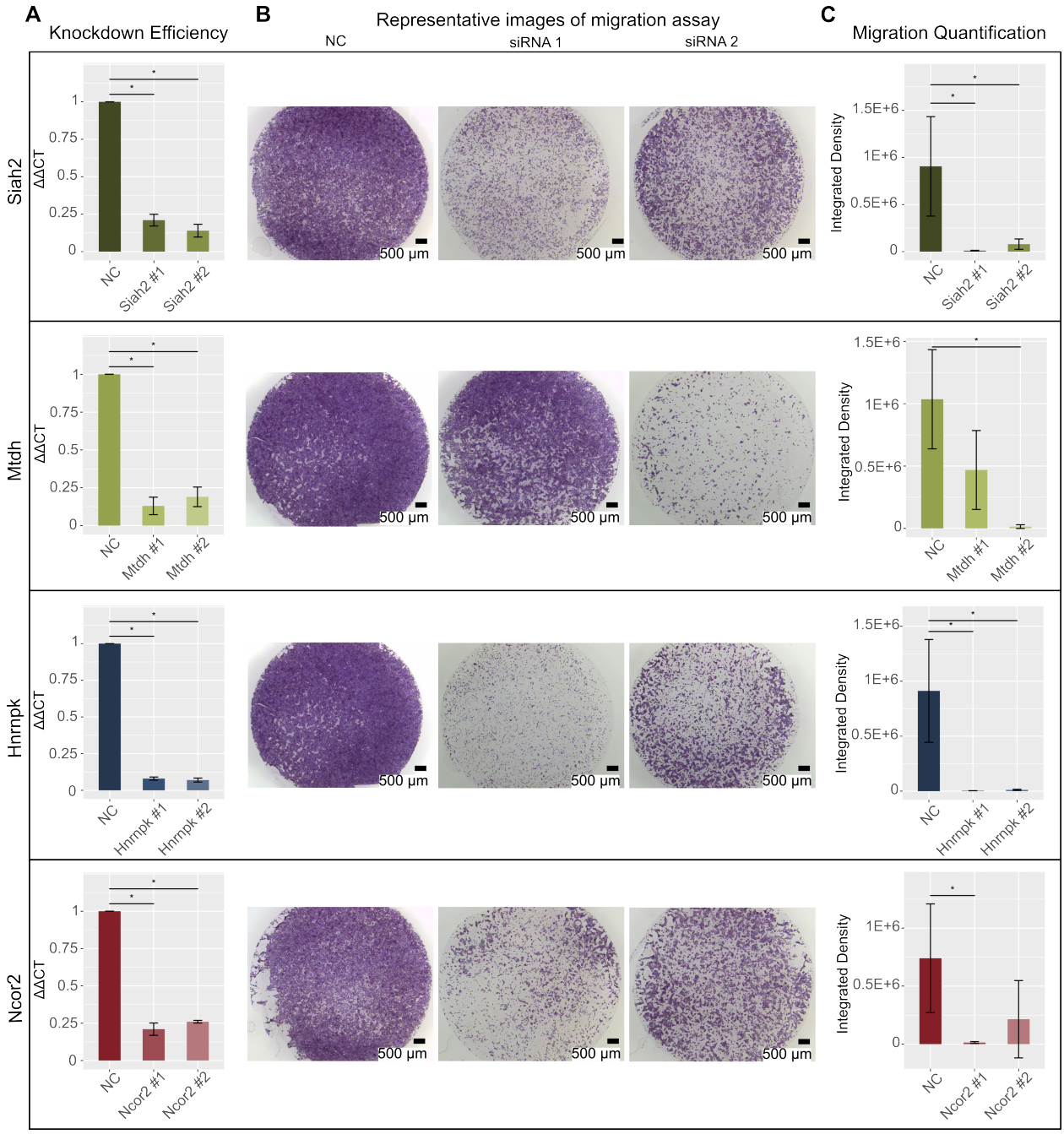

Figure 2.4: Gene knockdown of selected network genes showing reduction in cell migration capacity. Panels correspond to four genes, *Siah2*, *Mtdh*, *Hnrnpk* and *Ncor2*. Each condition, negative control (NC), siRNA #1 and siRNA #2, had three biological replicates. Error bars show standard deviation.

Figure 2.4: (Continued)

(A) Bar plots showing that gene expression was significantly reduced after knockdown (KD) compared to NC. GAPDH was used for normalization of all four genes' expression (ΔCT). Percent KD were calculated with the $\Delta\Delta$CT method. Values shown were normalized to the NC siRNAs. Y-axis shows $\Delta\Delta$CT value. Details of KD efficiencies, siRNAs, and primer sequences can be found in Supplementary Table S10 and Supplementary Table S11. (*) indicates p-value $<$ 0.05 (one-sided Wilcoxon rank sum test, n = 3). (B) Representative images of migration assays. Left, NC samples; middle, siRNA #1 samples; right, siRNA #2 samples. Scale bar: 500 µm. (C) Bar plots showing significant reduction in the integrated density of cells after knockdown (KD) compared to NC samples. Y-axis shows integrated densities of cells in NC samples, samples KD with siRNA #1 of each gene, and samples KD with siRNA #2 of each gene. Details of integrated densities can be found in Supplementary Table S10. (*) indicates p-value $<$ 0.05 (one-sided Wilcoxon rank sum test, n = 3).

Table 2.1: Hub genes associated with each network.

Timepoint	Network	Number of hub genes	Hub genes
e7.5	e7.5_1_STRING	7	***Mmp9***, *Ptprc*, *Tlr2*, *Cd68*, *Ctss*, *Cybb*, *Itgb2*
	e7.5_2_GENIE3	10	***Nr2f2*** (Hubert et al., 2010), ***Hmox1*** (Zenclussen et al., 2014), *Prdm1* (Mould et al., 2012), *Ctbp2* (Hildebrand and Soriano, 2002), *Irx1*, ***Frk***, *Siah2*, *Satb1*, *Trim8*, *Irf9*
e8.5	e8.5_1_STRING	11	***Akt1*** (Yang et al., 2003), ***Mapk14*** (Adams et al., 2000), *Mapk1* (Hatano et al., 2003), *Adam10*, *Creb1*, *Apob*, *Cdh2*, *Cttn*, ***Hsp90aa1***, *Apoe*, *Casp3*
	e8.5_2_GENIE3	17	*Erf* (Papadaki et al., 2007), *Setd2* (Hu et al., 2010), ***Msx2***, ***Dvl3***, *Dnmt1*, *Dnmt3b*, *Lin28a*, *Chek1*, *Dnajc2*, *Vgll1*, *Gpbp1l1*, *Jade1*, *Myef2*, *Nfxl1*, *Rbmx*, *Rhox4e*, *Cenpf*
e9.5	e9.5_1_STRING	7	***Fbxl19***, ***Smurf1***, ***Ubc***, *Wnt5a*, ***Ube2d1***, ***Mgrn1***, ***Nedd4***
	e9.5_1_GENIE3	34	***Rb1*** (Sun et al., 2006), ***Yap1*** (Meinhardt et al., 2020), *Esx1* (Li and Behringer, 1998), ***Ncoa3***, ***Ski***, ***Pitx1***, ***Zfx***, *Peg3*, *Ash1l*, ***Arid1b***, ***Arrb1***, ***Prmt2***, ***Tulp1***, ***Vgll4***, ***Creg1***, ***Foxo3***, ***Hif1an***, *Apbb1*, *2700081O15Rik*, *Ankrd2*, *Bbx*, *Bbx*, *Cdk5*, *Hdac6*, *Mllt3*, *Calcoco1*, *Cavin1*, *Cenpb*, *Cited4*, *Dtx1*, *Fam129b*, *Hcfc2*, *Mlxip*, *Phf8*, *Tsc22d1*
	e9.5_2_STRING	15	***Cdh1*** (Stemmler and Bedzhov, 2010), ***Fn1*** (George et al., 1993), ***Igf2*** (Fowden, 2003), ***Tgfb1*** (Graham et al., 1993), ***Vegfa*** (Ferrara et al., 1996), *Egfr* (Lee and Threadgill, 2009), ***Col1a1***, ***Csf1***, ***Timp1***, *Igf1*, ***App***, ***Spp1***, ***Itpkb***, ***Qsox1***, *Gas6*
	e9.5_2_GENIE3	27	***E2f8*** (Ouseph et al., 2012), ***Vegfa*** (Ferrara et al., 1996), ***Tead2*** (Sawada et al., 2008), ***Ets1***, ***Orc2***, ***Sox18***, *Klf3*, *Mrtfb*, *Trip6*, ***Cbx7***, ***Prnp***, ***Arhgef5***, ***Pias1***, ***Pias3***, ***Rasd1***, ***Txnip***, ***Zfp362***, *Plagl1*, *5730507C01Rik*, *BC004004*, *Bhlhe40*, *Ctdsp1*, *Grhl1*, *Ell2*, *Phf2*, *Fam83g*, *Rhox12*

Table 2.1: (Continued)

Timepoint	Network	Number of hub genes	Hub genes
e9.5	e9.5_3_STRING	5	***Gaa***, ***Lpcat1***, ***Olr1***, ***Cd59a***, *Stom*
	e9.5_3_GENIE3	29	***Tead1*** (Sawada et al., 2008), ***Smad4*** (Yang et al., 1998), ***Smad5*** (Yang et al., 1999), *Foxf1* (Ren et al., 2014), ***Cebpb***, ***Jun***, ***Hes1***, ***Tead3***, ***Cbx4***, ***Cdk2***, ***Dapk3***, ***Gata1***, *Hsbp1*, *Ncor2*, ***Dpf3***, ***Limd1***, ***Loxl2***, ***Pcgf5***, *Elmsan1*, *Hexim1*, *Lhx2*, *Sin3b*, *Mta3*, *Jade1*, *Rbm15b*, *Rhox4g*, *Rhox6*, *Rhox9*, *Tcf25*
	e9.5_4_STRING	10	*Lpar3* (Ye et al., 2005), ***Gna12***, ***Gnas***, ***Acta2***, ***Gcgr***, ***Pik3r3***, ***Rhoc***, ***Rhog***, *Rhoj*, *Adcy4*

Hub genes associated with each network. Colored genes are ones that have annotated or possible roles in placental development (see Materials and methods); green, e7.5-specific genes; blue, e8.5-specific genes; brown, e9.5-specific genes. Genes in bold are hub genes in one network inference method and nodes in the other method's networks.

2.7 References

Abdulghani, M., Song, G., Kaur, H., Walley, J. W., and Tuteja, G. (2019). Comparative Analysis of the Transcriptome and Proteome during Mouse Placental Development. *Journal of Proteome Research*, 18(5):2088–2099.

Adams, R. H., Porras, A., Alonso, G., Jones, M., Vintersten, K., Panelli, S., Valladares, A., Perez, L., Klein, R., and Nebreda, A. R. (2000). Essential Role of p38α MAP Kinase in Placental but Not Embryonic Cardiovascular Development. *Molecular Cell*, 6(1):109–116.

Alam, H., Li, N., Dhar, S. S., Wu, S. J., Lv, J., Chen, K., Flores, E. R., Baseler, L., and Lee, M. G. (2018). HP1γ Promotes Lung Adenocarcinoma by Downregulating the Transcription-Repressive Regulators NCOR2 and ZBTB7A. *Cancer research*, 78(14):3834–3848.

Alfaidy, N., Bogias, K. J., Pederson, S. M., Leemaqz, S., Smith, M. D., Mcaninch, D., Jankovic-Karasoulos, T., Mccullough, D., Wan, Q., Bianco-Miotto, T., Breen, J., Roberts, C. T., and Au, T. (2022). Placental Transcription Profiling in 6–23 Weeks' Gestation Reveals Differential Transcript Usage in Early Development. *International Journal of Molecular Sciences 2022, Vol. 23, Page 4506*, 23(9):4506.

Andrews, S. (2010). FastQC - A quality control tool for high throughput sequence data. *Babraham Bioinformatics*.

Andrews, T. S. and Hemberg, M. (2018). Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 59:114–122.

Bamfo, J. E. and Odibo, A. O. (2011). Diagnosis and management of fetal growth restriction. *Journal of pregnancy*, 2011:640715.

Bevilacqua, E., Lorenzon, A. R., Bandeira, C. L., and Hoshida, M. S. (2014). *Biology of the Ectoplacental Cone*. Elsevier.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5):525–7.

Brown, L. D. and Hay, W. W. (2016). Impact of placental insufficiency on fetal skeletal muscle growth. *Molecular and cellular endocrinology*, 435:69.

Canfield, J., Arlier, S., Mong, E. F., Lockhart, J., VanWye, J., Guzeloglu-Kayisli, O., Schatz, F., Magness, R. R., Lockwood, C. J., Tsibris, J. C., Kayisli, U. A., and Totary-Jain, H. (2019). Decreased LIN28B in preeclampsia impairs human trophoblast differentiation and migration. *The FASEB Journal*, 33(2):2759.

Carlson, M. (2019). org.Mm.eg.db: Genome wide annotation for Mouse.

Chen, G., Ning, B., and Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics*, 10(APR):317.

Constância, M., Hemberger, M., Hughes, J., Dean, W., Ferguson-Smith, A., Fundele, R., Stewart, F., Kelsey, G., Fowden, A., Sibley, C., and Reik, W. (2002). Placental-specific IGF-II is a major modulator of placental and fetal growth. *Nature*, 417(6892):945–948.

Costanzo, V., Bardelli, A., Siena, S., and Abrignani, S. (2018). Exploring the links between cancer and placenta development. *Open Biology*, 8(6).

Cox, B., Kotlyar, M., Evangelou, A. I., Ignatchenko, V., Ignatchenko, A., Whiteley, K., Jurisica, I., Adamson, S. L., Rossant, J., and Kislinger, T. (2009). Comparative systems biology of human and mouse as a tool to guide the modeling of human placental pathology. *Molecular Systems Biology*, 5:279.

Cross, J. C. (2005). How to Make a Placenta: Mechanisms of Trophoblast Cell Differentiation in Mice – A Review. *Placenta*, 26:S3–S9.

Cross, J. C., Baczyk, D., Dobric, N., Hemberger, M., Hughes, M., Simmons, D. G., Yamamoto, H., and Kingdom, J. C. (2003). Genes, development and evolution of the placenta. *Placenta*, 24(2-3):123–130.

Cross, J. C., Nakano, H., Natale, D. R., Simmons, D. G., and Watson, E. D. (2006). Branching morphogenesis during development of placental villi. *Differentiation*, 74(7):393–401.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Marugán, J. C., Maurel, T., McMahon, A. C., Moore, B., Morales, J., Mudge, J. M., Nuhn, M., Ogeh, D., Parker, A., Parton, A., Patricio, M., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sparrow, H., Stapleton, E., Szuba, M., Taylor, K., Threadgold, G., Thormann, A., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Yates, A. D., Zerbino, D. R., and Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, 47(D1):D745–D751.

Das, L., Kokate, S., Rath, S., Rout, N., Singh, S., Crowe, S., Mukhopadhyay, A., and Bhattacharyya, A. (2016). ETS2 and Twist1 promote invasiveness of Helicobacter pylori-infected gastric cancer cells by inducing Siah2. *Biochemical Journal*, 473(11):1629.

Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., Meehan, T. F., Weninger, W. J., Westerberg, H., Adissu, H., Baker, C. N., Bower, L., Brown, J. M., Brianna Caddle, L., Chiani, F., Clary, D., Cleak, J., Daly, M. J., Denegre, J. M., Doe, B., Dolan, M. E., Edie, S. M., Fuchs, H., Gailus-Durner, V., Galli, A., Gambadoro, A., Gallegos, J., Guo, S., Horner, N. R., wei Hsu, C., Johnson, S. J., Kalaga, S., Keith, L. C., Lanoue, L., Lawson, T. N., Lek, M., Mark, M., Marschall, S., Mason, J., McElwee, M. L., Newbigging, S., Nutter, L. M., Peterson, K. A., Ramirez-Solis, R., Rowland, D. J., Ryder, E., Samocha, K. E., Seavitt, J. R., Selloum, M., Szoke-Kovacs, Z., Tamura, M., Trainor, A. G., Tudose, I., Wakana, S., Warren, J., Wendling, O., West, D. B., Wong, L., Yoshiki, A., MacArthur, D. G., Tocchini-Valentini, G. P., Gao, X., Flicek, P., Bradley, A., Skarnes, W. C., Justice, M. J., Parkinson, H. E., Moore, M., Wells, S., Braun, R. E., Svenson, K. L., Hrabe de Angelis, M., Herault, Y., Mohun, T., Mallon, A. M., Mark Henkelman, R., Brown, S. D., Adams, D. J., Kent Lloyd, K. C., McKerlie, C., Beaudet, A. L., Bucan, M., Murray, S. A., McKay, M., Urban, B., Lund, C., Froeter, E., LaCasse, T., Mehalow, A., Gordon, E., Donahue, L. R., Taft, R., Kutney, P., Dion, S., Goodwin, L., Kales, S., Urban, R., Palmer, K., Pertuy, F., Bitz, D., Weber, B., Goetz-Reiner, P., Jacobs, H., Le Marchand, E., El Amri, A., El Fertak, L., Ennah, H., Ali-Hadji, D., Ayadi, A., Wattenhofer-Donze, M., Jacquot, S., André, P., Birling, M. C., Pavlovic, G., Sorg, T., Morse, I., Benso, F., Stewart, M. E., Copley, C., Harrison, J., Joynson, S., Guo, R., Qu, D., Spring, S., Yu, L., Ellegood, J., Morikawa, L., Shang, X., Feugas, P., Creighton, A., Penton, P. C., Danisment, O., Griggs, N., Tudor, C. L., Green, A. L., Icoresi Mazzeo, C., Siragher, E., Lillistone, C., Tuck, E., Gleeson, D., Sethi, D., Bayzetinova, T., Burvill, J., Habib, B., Weavers, L., Maswood, R., Miklejewska, E., Woods, M., Grau, E., Newman, S., Sinclair, C., Brown, E., Ayabe, S., Iwama, M., and Murakami, A. (2016). High-throughput discovery of novel developmental phenotypes. *Nature 2016 537:7621*, 537(7621):508–514.

Donker, R. B., Mouillet, J. F., Chu, T., Hubel, C. A., Stolz, D. B., Morelli, A. E., and Sadovsky, Y. (2012). The expression profile of C19MC microRNAs in primary human trophoblast cells and exosomes. *Molecular Human Reproduction*, 18(8):417–424.

Ferrara, N., Carver-Moore, K., Chen, H., Dowd, M., Lu, L., O'Shea, K. S., Powell-Braxton, L., Hillan, K. J., and Moore, M. W. (1996). Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature*, 380(6573):439–442.

Fowden, A. L. (2003). The Insulin-like Growth Factors and feto-placental Growth. *Placenta*, 24(8-9):803–812.

Frew, I. J., Hammond, V. E., Dickins, R. A., Quinn, J. M. W., Walkley, C. R., Sims, N. A., Schnall, R., Della, N. G., Holloway, A. J., Digby, M. R., Janes, P. W., Tarlinton, D. M., Purton, L. E., Gillespie, M. T., and Bowtell, D. D. L. (2003). Generation and Analysis of Siah2 Mutant Mice. *Molecular and Cellular Biology*, 23(24):9150.

Gabriel, H. D., Jung, D., Bützler, C., Temme, A., Traub, O., Winterhager, E., and Willecke, K. (1998). Transplacental uptake of glucose is decreased in embryonic lethal connexin26-deficient mice. *Journal of Cell Biology*, 140(6):1453–1461.

Gallardo, M., Lee, H. J., Zhang, X., Bueso-Ramos, C., Pageon, L. R., McArthur, M., Multani, A., Nazha, A., Manshouri, T., Parker-Thornburg, J., Rapado, I., Quintas-Cardama, A., Kornblau, S. M., Martinez-Lopez, J., and Post, S. M. (2015). HnRNP K Is a Haploinsufficient Tumor Suppressor that Regulates Proliferation and Differentiation Programs in Hematologic Malignancies. *Cancer Cell*, 28(4):486–499.

George, E. L., Georges-Labouesse, E. N., Patel-King, R. S., Rayburn, H., and Hynes, R. O. (1993). Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin. *Development*, 119(4):1079–1091.

Gormley, M., Oliverio, O., Kapidzic, M., Ona, K., Hall, S., and Fisher, S. J. (2021). RNA profiling of laser microdissected human trophoblast subtypes at mid-gestation reveals a role for cannabinoid signaling in invasion. *Development (Cambridge, England)*, 148(20).

Graham, C. H., Hawley, T. S., Hawley, R. G., Macdougall, J. R., Kerbel, R. S., Khoo, N., and Lala, P. K. (1993). Establishment and Characterization of First Trimester Human Trophoblast Cells with Extended Lifespan. *Experimental Cell Research*, 206(2):204–211.

Guillemot, F., Nagy, A., Auerbach, A., Rossant, J., and Joyner, A. L. (1994). Essential role of Mash-2 in extraembryonic development. *Nature*, 371(6495):333–336.

Hatano, N., Mori, Y., Oh-hora, M., Kosugi, A., Fujikawa, T., Nakai, N., Niwa, H., Miyazaki, J., Hamaoka, T., and Ogata, M. (2003). Essential role for ERK2 mitogen-activated protein kinase in placental development. *Genes to Cells*, 8(11):847–856.

Hemberger, M. and Cross, J. C. (2001). Genes governing placental development. *Trends in Endocrinology & Metabolism*, 12(4):162–168.

Hemberger, M., Hanna, C. W., and Dean, W. (2020). Mechanisms of early placental development in mouse and humans. *Nature Reviews Genetics*, 21(1):27–43.

Hildebrand, J. D. and Soriano, P. (2002). Overlapping and Unique Roles for C-Terminal Binding Protein 1 (CtBP1) and CtBP2 during Mouse Development. *Molecular and Cellular Biology*, 22(15):5296.

Hirschberg, A. L., Jakson, I., Graells Brugalla, C., Salamon, D., and Ujvari, D. (2021). Interaction between insulin and androgen signalling in decidualization, cell migration and trophoblast invasion in vitro. *Journal of Cellular and Molecular Medicine*, 25(20):9523–9532.

Hu, D. and Cross, J. C. (2010). Development and function of trophoblast giant cells in the rodent placenta. *International Journal of Developmental Biology*, 54(2-3):341–354.

Hu, H., Miao, Y. R., Jia, L. H., Yu, Q. Y., Zhang, Q., and Guo, A. Y. (2019). AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Research*, 47(D1):D33–D38.

Hu, M., Sun, X. J., Zhang, Y. L., Kuang, Y., Hu, C. Q., Wu, W. L., Shen, S. H., Du, T. T., Li, H., He, F., Xiao, H. S., Wang, Z. G., Liu, T. X., Lu, H., Huang, Q. H., Chen, S. J., and Chen, Z. (2010). Histone H3 lysine 36 methyltransferase Hypb/Setd2 is required for embryonic vascular remodeling. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2956–2961.

Hubert, M. A., Sherritt, S. L., Bachurski, C. J., and Handwerger, S. (2010). Involvement of Transcription Factor NR2F2 in Human Trophoblast Differentiation. *PLOS ONE*, 5(2):e9417.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9).

Jain, A. and Tuteja, G. (2021). PlacentaCellEnrich: A tool to characterize gene sets using placenta cell-specific gene enrichment analysis. *Placenta*, 103:164–171.

Jeon, H. Y., Choi, M., Howlett, E. L., Vozhilla, N., Yoo, B. K., Lloyd, J. A., Sarkar, D., Lee, S. G., and Fisher, P. B. (2010). Expression patterns of astrocyte elevated gene-1 (AEG-1) during development of the mouse embryo. *Gene expression patterns : GEP*, 10(7-8):361.

Jepsen, K., Solum, D., Zhou, T., McEvilly, R. J., Kim, H. J., Glass, C. K., Hermanson, O., and Rosenfeld, M. G. (2007). SMRT-mediated repression of an H3K27 demethylase in progression from neural stem cell to neuron. *Nature*, 450(7168):415–419.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386.

Kibschull, M., Colaco, K., Matysiak-Zablocki, E., Winterhager, E., and Lye, S. J. (2014). Connexin31.1 (Gjb5) deficiency blocks trophoblast stem cell differentiation and delays placental development. *Stem Cells and Development*, 23(21):2649–2660.

Kuckenberg, P., Kubaczka, C., and Schorle, H. (2012). The role of transcription factor Tcfap2c/TFAP2C in trophectoderm development. *Reproductive BioMedicine Online*, 25(1):12–20.

Lager, S. and Powell, T. L. (2012). Regulation of Nutrient Transport across the Placenta. *Journal of Pregnancy*, 2012:14.

Lee, T. C. and Threadgill, D. W. (2009). Generation and validation of mice carrying a conditional allele of the epidermal growth factor receptor. *Genesis (New York, N.Y. : 2000)*, 47(2):85–92.

Li, H., Qu, D., McDonald, A., Isaac, S. M., Whiteley, K. J., Sung, H.-K., Nagy, A., and Adamson, S. L. (2014). Trophoblast-Specific Reduction of VEGFA Alters Placental Gene Expression and Maternal Cardiovascular Function in Mice. *Biology of Reproduction*, 91(4):1–12.

Li, Y. and Behringer, R. R. (1998). Esx1 is an X-chromosome-imprinted regulator of placental development and fetal growth. *Nature Genetics 1998 20:3*, 20(3):309–311.

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, 47(W1):W199–W205.

Marsh, B. and Blelloch, R. (2020). Single nuclei RNA-seq of mouse placental labyrinth development. *eLife*, 9:1–27.

Meinhardt, G., Haider, S., Kunihs, V., Saleh, L., Pollheimer, J., Fiala, C., Hetey, S., Feher, Z., Szilagyi, A., Than, N. G., and Knöfler, M. (2020). Pivotal role of the transcriptional co-activator YAP in trophoblast stemness of the developing human placenta. *Proceedings of the National Academy of Sciences of the United States of America*, 117(24):13562–13570.

Mor, G., Cardenas, I., Abrahams, V., and Guller, S. (2011). Inflammation and pregnancy: the role of the immune system at the implantation site. *Annals of the New York Academy of Sciences*, 1221(1):80.

Morey, R., Farah, O., Kallol, S., Requena, D. F., Meads, M., Moretto-Zita, M., Soncin, F., Laurent, L. C., and Parast, M. M. (2021). Transcriptomic Drivers of Differentiation, Maturation, and Polyploidy in Human Extravillous Trophoblast. *Frontiers in Cell and Developmental Biology*, 9:2269.

Mould, A., Morgan, M. A., Li, L., Bikoff, E. K., and Robertson, E. J. (2012). Blimp1/Prdm1 governs terminal differentiation of endovascular trophoblast giant cells and defines multipotent progenitors in the developing placenta. *Genes & Development*, 26(18):2063.

Naismith, K. and Cox, B. (2021). Human placental gene sets improve analysis of placental pathologies and link trophoblast and cancer invasion genes. *Placenta*, 112:9–15.

Okae, H., Toh, H., Sato, T., Hiura, H., Takahashi, S., Shirane, K., Kabayama, Y., Suyama, M., Sasaki, H., and Arima, T. (2018). Derivation of Human Trophoblast Stem Cells. *Cell Stem Cell*, 22(1):50–63.e6.

Ouseph, M. M., Li, J., Chen, H. Z., Pécot, T., Wenzel, P., Thompson, J. C., Comstock, G., Chokshi, V., Byrne, M., Forde, B., Chong, J. L., Huang, K., Machiraju, R., de Bruin, A., and Leone, G. (2012). Atypical E2F Repressors and Activators Coordinate Placental Development. *Developmental Cell*, 22(4):849–862.

Papadaki, C., Alexiou, M., Cecena, G., Verykokakis, M., Bilitou, A., Cross, J. C., Oshima, R. G., and Mavrothalassitis, G. (2007). Transcriptional repressor erf determines extraembryonic ectoderm differentiation. *Molecular and cellular biology*, 27(14):5201–5213.

Perez-Garcia, V., Fineberg, E., Wilson, R., Murray, A., Mazzeo, C. I., Tudor, C., Sienerth, A., White, J. K., Tuck, E., Ryder, E. J., Gleeson, D., Siragher, E., Wardle-Jones, H., Staudt, N., Wali, N., Collins, J., Geyer, S., Busch-Nentwich, E. M., Galli, A., Smith, J. C., Robertson, E., Adams, D. J., Weninger, W. J., Mohun, T., and Hemberger, M. (2018). Placentation defects are highly prevalent in embryonic lethal mouse mutants. *Nature*, 555(7697):463.

Pimentel, H., Bray, N. L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14(7):687–690.

Plaks, V., Rinkenberger, J., Dai, J., Flannery, M., Sund, M., Kanasaki, K., Ni, W., Kalluri, R., and Werb, Z. (2013). Matrix metalloproteinase-9 deficiency phenocopies features of preeclampsia and intrauterine growth restriction. *Proceedings of the National Academy of Sciences of the United States of America*, 110(27):11109.

Prater, M., Hamilton, R. S., Yung, H. W., Sharkey, A. M., Robson, P., Hamid, N. E. A., Jauniaux, E., Charnock-Jones, D. S., Burton, G. J., and Cindrova-Davies, T. (2021). RNA-Seq reveals changes in human placental metabolism, transport and endocrinology across the first–second trimester transition. *Biology Open*, 10(6).

Qi, J., Kim, H., Scortegagna, M., and Ronai, Z. A. (2013). Regulators and effectors of Siah ubiquitin ligases. *Cell biochemistry and biophysics*, 67(1):15.

Qi, J., Nakayama, K., Gaitonde, S., Goydos, J. S., Krajewski, S., Eroshkin, A., Bar-Sagi, D., Bowtell, D., and Ronai, Z. (2008). The ubiquitin ligase Siah2 regulates tumorigenesis and metastasis by HIF-dependent and -independent pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 105(43):16713.

R Core Development Team (2013). R: A Language and Environment for Statistical Computing.

Rana, S., Lemoine, E., Granger, J., and Karumanchi, S. A. (2019). Preeclampsia: Pathophysiology, Challenges, and Perspectives. *Circulation Research*, 124(7):1094–1112.

Ren, X., Ustiyan, V., Pradhan, A., Cai, Y., Havrilak, J. A., Bolte, C. S., Shannon, J. M., Kalin, T. V., and Kalinichenko, V. V. (2014). FOXF1 Transcription Factor Is Required for Formation of Embryonic Vasculature by Regulating VEGF Signaling in Endothelial Cells. *Circulation research*, 115(8):709.

Rivals, I., Personnaz, L., Taing, L., and Potier, M. C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–407.

Rossant, J. (2001). Stem cells from the Mammalian blastocyst. *Stem cells (Dayton, Ohio)*, 19(6):477–482.

Sawada, A., Kiyonari, H., Ukita, K., Nishioka, N., Imuta, Y., and Sasaki, H. (2008). Redundant Roles of Tead1 and Tead2 in Notochord Development and the Regulation of Cell Proliferation and Survival. *MOLECULAR AND CELLULAR BIOLOGY*, 28(10):3177–3189.

Sferruzzi-Perri, A. N., Vaughan, O. R., Coan, P. M., Suciu, M. C., Darbyshire, R., Constancia, M., Burton, G. J., and Fowden, A. L. (2011). Placental-Specific Igf2 Deficiency Alters Developmental Adaptations to Undernutrition in Mice. *Endocrinology*, 152(8):3202–3212.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

Sheridan, M. A., Zhao, X., Fernando, R. C., Gardner, L., Perez-Garcia, V., Li, Q., Marsh, S. G., Hamilton, R., Moffett, A., and Turco, M. Y. (2021). Characterization of primary models of human trophoblast. *Development (Cambridge, England)*, 148(21).

Shi, Q., Song, X., Wang, J., Gu, J., Zhang, W., Hu, J., Zhou, X., and Yu, R. (2015). FRK inhibits migration and invasion of human glioma cells by promoting N-cadherin/β-catenin complex formation. *Journal of molecular neuroscience : MN*, 55(1):32–41.

Si, Y., Liu, P., Li, P., and Brutnell, T. P. (2014). Model-based clustering for RNA-seq data. *Bioinformatics*, 30(2):197–205.

Silva, J. F. and Serakides, R. (2016). Intrauterine trophoblast migration: A comparative view of humans and rodents. *Cell Adhesion and Migration*, 10(1-2):88–110.

Simmons, D. G. (2014). *Postimplantation Development of the Chorioallantoic Placenta*. Elsevier.

Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley interdisciplinary reviews. Systems biology and medicine*, 1(3):390–399.

Soncin, F., Khater, M., To, C., Pizzo, D., Farah, O., Wakeland, A., Rajan, K. A. N., Nelson, K. K., Chang, C. W., Moretto-Zita, M., Natale, D. R., Laurent, L. C., and Parast, M. M. (2018). Comparative analysis of mouse and human placentae across gestation reveals species-specific regulators of placental development. *Development (Cambridge)*, 145(2).

Soncin, F., Natale, D., and Parast, M. M. (2015). Signaling pathways in mouse and human trophoblast differentiation: A comparative review. *Cellular and Molecular Life Sciences*, 72(7):1291–1302.

Soneson, C., Love, M. I., and Robinson, M. D. (2016). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research*, 4:1521.

Song, H., Liu, J., Wu, X., Zhou, Y., Chen, X., Chen, J., Deng, K., Mao, C., Huang, S., and Liu, Z. (2019). LHX2 promotes malignancy and inhibits autophagy via mTOR in osteosarcoma and is negatively regulated by miR-129-5p. *Aging (Albany NY)*, 11(21):9794.

Spies, D., Renz, P. F., Beyer, T. A., and Ciaudo, C. (2019). Comparative analysis of differential gene expression tools for RNA sequencing time course data. *Briefings in Bioinformatics*, 20(1):288.

Starks, R. R., Alhasan, R. A., Kaur, H., Pennington, K. A., Schulz, L. C., and Tuteja, G. (2020). Transcription Factor PLAGL1 Is Associated with Angiogenic Gene Expression in the Placenta. *International Journal of Molecular Sciences 2020, Vol. 21, Page 8317*, 21(21):8317.

Starks, R. R., Biswas, A., Jain, A., and Tuteja, G. (2019). Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics and Chromatin*, 12(1):1–16.

Starks, R. R., Kaur, H., and Tuteja, G. (2021). Mapping cis-regulatory elements in the midgestation mouse placenta. *Scientific Reports 2021 11:1*, 11(1):1–13.

Stemmler, M. P. and Bedzhov, I. (2010). A Cdh1HA knock-in allele rescues the Cdh1-/- phenotype but shows essential Cdh1 function during placentation. *Developmental dynamics : an official publication of the American Association of Anatomists*, 239(9):2330–2344.

Su, G., Kuchinsky, A., Morris, J. H., States, D. J., and Meng, F. (2010). GLay: community structure analysis of biological networks. *Bioinformatics*, 26(24):3135–3137.

Sun, H., Chang, Y., Schweers, B., Dyer, M. A., Zhang, X., Hayward, S. W., and Goodrich, D. W. (2006). An E2F Binding-Deficient Rb1 Protein Partially Rescues Developmental Defects Associated with Rb1 Nullizygosity. *Molecular and Cellular Biology*, 26(4):1527.

Suryawanshi, H., Max, K., Bogardus, K. A., Sopeyin, A., Chang, M. S., Morozov, P., Castano, P. M., Tuschl, T., and Williams, Z. (2022). Dynamic genome-wide gene expression and immune cell composition in the developing human placenta. *Journal of Reproductive Immunology*, 151:103624.

Sutton, G. J., Poppe, D., Simmons, R. K., Walsh, K., Nawaz, U., Lister, R., Gagnon-Bartsch, J. A., and Voineagu, I. (2022). Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nature Communications 2022 13:1*, 13(1):1–18.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613.

Tanaka, M., Gertsenstein, M., Rossant, J., and Nagy, A. (1997). Mash2 Acts Cell Autonomously in Mouse Spongiotrophoblast Development. *Developmental Biology*, 190(1):55–65.

Tetzlaff, M. T., Bai, C., Finegold, M., Wilson, J., Harper, J. W., Mahon, K. A., and Elledge, S. J. (2004). Cyclin F Disruption Compromises Placental Development and Affects Normal Cell Cycle Execution. *Molecular and Cellular Biology*, 24(6):2487–2498.

Theiler, K. (1989). The House Mouse. *The House Mouse*.

Tuteja, G., Chung, T., and Bejerano, G. (2016). Changes in the enhancer landscape during early placental development uncover a trophoblast invasion gene-enhancer network. *Placenta*, 37:45–55.

Tzouanacou, E., Tweedie, S., and Wilson, V. (2003). Identification of Jade1, a Gene Encoding a PHD Zinc Finger Protein, in a Gene Trap Mutagenesis Screen for Genes Involved in Anteroposterior Axis Development. *MOLECULAR AND CELLULAR BIOLOGY*, 23(23):8553–8562.

Varberg, K. M., Iqbal, K., Muto, M., Simon, M. E., Scott, R. L., Kozai, K., Choudhury, R. H., Aplin, J. D., Biswell, R., Gibson, M., Okae, H., Arima, T., Vivian, J. L., Grundberg, E., and Soares, M. J. (2021). ASCL2 reciprocally controls key trophoblast lineage decisions during hemochorial placenta development. *Proceedings of the National Academy of Sciences of the United States of America*, 118(10).

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J. E., Stephenson, E., Polański, K., Goncalves, A., Gardner, L., Holmqvist, S., Henriksson, J., Zou, A., Sharkey, A. M., Millar, B., Innes, B., Wood, L., Wilbrey-Clark, A., Payne, R. P., Ivarsson, M. A., Lisgo, S., Filby, A., Rowitch, D. H., Bulmer, J. N., Wright, G. J., Stubbington, M. J., Haniffa, M., Moffett, A., and Teichmann, S. A. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731):347–353.

Wacquet, G., Poisson-Caillault, É., and Hébert, P. A. (2013). Semi-supervised K-way spectral clustering with determination of number of clusters. In *Studies in Computational Intelligence*, volume 465, pages 317–332. Springer Verlag.

Walentin, K., Hinze, C., and Schmidt-Ott, K. M. (2016). The basal chorionic trophoblast cell layer: An emerging coordinator of placenta development. *BioEssays*, 38(3):254–265.

Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M., and Zhang, N. R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 115(28):E6437–E6446.

Wang, M., Xu, Y., Wang, P., Xu, Y., Jin, P., Wu, Z., Qian, Y., Bai, L., and Dong, M. (2021). Galectin-14 Promotes Trophoblast Migration and Invasion by Upregulating the Expression of MMP-9 and N-Cadherin. *Frontiers in Cell and Developmental Biology*, 9:487.

Watson, E. D. and Cross, J. C. (2005). Development of structures and transport functions in the mouse placenta. *Physiology*, 20(3):180–193.

Wehrens, R. and Kruisselbrink, J. (2018). Flexible self-organizing maps in kohonen 3.0. *Journal of Statistical Software*, 87(7):1–18.

Woods, L., Perez-garcia, V., and Hemberger, M. (2018). Regulation of Placental Development and Its Impact on Fetal Growth — New Insights From Mouse Models. *Frontiers in Endocrinology*, 9(September):1–18.

Xie, J., Xu, Y., Wan, L., Wang, P., Wang, M., and Dong, M. (2018). Involvement of follistatin-like 3 in preeclampsia. *Biochemical and Biophysical Research Communications*, 506(3):692–697.

Yamada, Y., Kimura, N., ichi Takayama, K., Sato, Y., Suzuki, T., Azuma, K., Fujimura, T., Ikeda, K., Kume, H., and Inoue, S. (2020). TRIM44 promotes cell proliferation and migration by inhibiting FRK in renal cell carcinoma. *Cancer science*, 111(3):881–890.

Yang, C., Zheng, S., Liu, T., Liu, Q., Dai, F., Zhou, J., Chen, Y., Sheyhidin, I., and Lu, X. (2017). Down-regulated miR-26a promotes proliferation, migration, and invasion via negative regulation of MTDH in esophageal squamous cell carcinoma. *The FASEB Journal*, 31(5):2114–2122.

Yang, J. T., Rayburn, H., and Hynes, R. O. (1995). Cell adhesion events mediated by alpha 4 integrins are essential in placental and cardiac development. *Development*, 121(2):549–560.

Yang, X., Castilla, L. H., Xu, X., Li, C., Gotay, J., Weinstein, M., Liu, P. P., and Deng, C. X. (1999). Angiogenesis defects and mesenchymal apoptosis in mice lacking SMAD5. *Development*, 126(8):1571–1580.

Yang, X., Li, C., Xu, X., and Deng, C. (1998). The tumor suppressor SMAD4/DPC4 is essential for epiblast proliferation and mesoderm induction in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 95(7):3667–3672.

Yang, Z., Tschopp, O., Hemmings-Mieszczak, M., Feng, J., Brodbeck, D., Perentes, E., and Hemmings, B. (2003). Protein kinase B alpha/Akt1 regulates placental development and fetal growth. *The Journal of biological chemistry*, 278(34):32124–32131.

Ye, X., Hama, K., Contos, J. J., Anliker, B., Inoue, A., Skinner, M. K., Suzuki, H., Amano, T., Kennedy, G., Arai, H., Aoki, J., and Chun, J. (2005). LPA3-mediated lysophosphatidic acid signalling in embryo implantation and spacing. *Nature 2005 435:7038*, 435(7038):104–108.

Yi, L., Pimentel, H., Bray, N. L., and Pachter, L. (2018). Gene-level differential analysis at transcript-level resolution. *Genome Biology*, 19(1).

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287.

Zaitsev, K., Bambouskova, M., Swain, A., and Artyomov, M. N. (2019). Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nature Communications 2019 10:1*, 10(1):1–16.

Zenclussen, M. L., Linzke, N., Schumacher, A., Fest, S., Meyer, N., Casalis, P. A., and Zenclussen, A. C. (2014). Heme oxygenase-1 is critically involved in placentation, spiral artery remodeling, and blood pressure regulation during murine pregnancy. *Frontiers in Pharmacology*, 5(JAN).

Zhang, X., Green, K. E., Yallampalli, C., and Dong, Y. L. (2005). Adrenomedullin Enhances Invasion by Trophoblast Cell Lines. *Biology of Reproduction*, 73(4):619–626.

Zhao Peng, W., Xi Liu, J., Feng Li, C., Ma, R., and Zheng Jie, J. (2019). Hnrnpk promotes gastric tumorigenesis through regulating cd44e alternative splicing. *Cancer Cell International*, 19(1):1–11.

2.8 Appendix A: Notes

2.8.1 Data availability statement

All code for the analyses is available at `https://github.com/Tuteja-Lab/PlacentaRNA-seq`. All raw and processed data is available for download on NCBI Gene Expression Omnibus (GEO) Repository, accession number: GSE202243.

2.8.2 Acknowledgements

We acknowledge the Iowa State University DNA Facility for preparing and sequencing the RNA-seq libraries, and the Research IT group at Iowa State University (`http://researchit.las.iastate.edu`) for providing servers and IT support. We would like to thank Tuteja lab members for their discussion and support.

2.8.3 Author contributions

Conceptualization, H.V., G.T.; Methodology, H.V., G.T.; Data Generation, H.K., R.S.; Formal Analysis, H.V.; Data Interpretation, H.V., G.T.; Experimental Validation, H.K.; Analysis Validation, K.K.; Writing – Original Draft Preparation, H.V., G.T.; Writing – Review and Editing, H.V., H.K., K.K., R.S. and G.T.; Supervision, G.T.; Funding Acquisition, G.T.

2.8.4 Declaration of interests

The authors declare no competing interests.

2.8.5 Funding

This work was supported in part by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under award number R01HD096083 (to GT). Geetu Tuteja is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts. The views expressed are those of the author(s) and do not necessarily reflect the views of the funding agencies.

2.9 Appendix B: Supplementary tables and figures

All supplementary tables and figures can be found online at Vu et al., 2023: https://www.life-science-alliance.org/content/6/2/e202201788/tab-figures-data.

CHAPTER 3. UNSUPERVISED CONTRASTIVE PEAK CALLER FOR ATAC-SEQ

Ha T.H. Vu [*1,2], Yudi Zhang [*3], Geetu Tuteja [1,2], and Karin Dorman[1,2,3]

[1]Bioinformatics and Computational Biology, Iowa State University, Ames IA 50011, USA

[2]Genetics, Development and Cell Biology, Iowa State University, Ames IA 50011, USA

[3]Department of Statistics, Iowa State University, Ames IA 50011, USA

Modified from a manuscript published in *Genome Research*

3.1 Abstract

The assay for transposase-accessible chromatin with sequencing (ATAC-seq) is a common assay to identify chromatin accessible regions by using a Tn5 transposase that can access, cut, and ligate adapters to DNA fragments for subsequent amplification and sequencing. These sequenced regions are quantified and tested for enrichment in a process referred to as "peak calling". Most unsupervised peak calling methods are based on simple statistical models and suffer from elevated false positive rates. Newly developed supervised deep learning methods can be successful, but they rely on high quality labeled data for training, which can be difficult to obtain. Moreover, though biological replicates are recognized to be important, there are no established approaches for using replicates in the deep learning tools, and the approaches available for traditional methods either cannot be applied to ATAC-seq, where control samples may be unavailable, or are post-hoc and do not capitalize on potentially complex, but reproducible signal in the read enrichment data. Here, we propose a novel peak caller that uses unsupervised contrastive learning to extract shared signals from multiple replicates. Raw coverage data are encoded to obtain low-dimensional embeddings and optimized to minimize a contrastive

[*]These authors contributed equally to this work.

loss over biological replicates. These embeddings are passed to another contrastive loss for learning and predicting peaks and decoded to denoised data under an autoencoder loss. We compared our Replicative Contrastive Learner (RCL) method with other existing methods on ATAC-seq data, using annotations from ChromHMM genome and transcription factor ChIP-seq as noisy truth. RCL consistently achieved the best performance.

3.2 Introduction

The assay for transposase-accessible chromatin with sequencing (ATAC-seq) is widely used when studying chromatin biology (Grandi et al., 2022). ATAC-seq utilizes a hyperactive mutant Tn5 transposase to cleave double stranded DNA and to attach adapters for subsequent sequencing by high throughput technologies (Buenrostro et al., 2015). Since DNA is more easily cleaved where it is unwound and open, sequenced DNA fragments tend to arise from regions of open chromatin. A standard analysis for ATAC-seq starts with aligning the sequencing reads to a reference genome using BWA (Li, 2013), Bowtie2 (Langmead and Salzberg, 2012), or other short read aligner (Musich et al., 2021). Then peak calling methods will identify the open regions (peaks) in the genome where aligned reads are enriched. Downstream analyses include motif detection, differential binding analysis or footprint identification (Buenrostro et al., 2013; Grandi et al., 2022), all of which require accurate peak calls. Unfortunately, peaks of false enrichment may be called due to mapping errors or experimental noise (Park, 2009). Such errors can be reduced by masking repetitive regions and using control samples (Zhang et al., 2008), but input controls for ATAC-seq are typically not used due to high sequencing costs (Yan et al., 2020).

ATAC-seq peaks are often called with the most popular general-purpose peak caller, MACS (Zhang et al., 2008), and there is an ATAC-seq-specific method called HMMRATAC (Tarbell and Liu, 2019). MACS slides a fixed-width window across the genome to find candidate peaks. The number of reads aligned to the genome in the current window is modeled as a Poisson random variable, with a dynamic mean to capture local variation in background coverage rates. MACS calculates the p-value for each candidate peak as the

probability of obtaining coverage at or above the observed coverage given the current background rate. HMMRATAC (Tarbell and Liu, 2019) employs a hidden Markov model (HMM) with four-dimensional emissions of varying fragment sizes, nucleosome-free (NF), one nucleosome (1N), two nucleosome (2N) and three nucleosome (3N) fragments, from three possible hidden states: a "center" state (open chromatin), with high emissions in all four dimensions, a nucleosome state, with low NF fragment emission, and a background state, with low emissions in all dimensions. Once the HMM has been estimated, the Viterbi algorithm is used to classify every 10 base pair (bp) window in the genome into one of the three states.

Traditional modeling methods tend to predict many false positive peaks in ChIP-seq applications (Hocking et al., 2017), and some investigations have shown humans to be superior "peak callers" (Rye et al., 2011; Hocking et al., 2017). Inspired by such human performance and recent successes in artificial intelligence, two new peak callers, CNN-Peaks (Oh et al., 2020) and LanceOtron (Hentges et al., 2021), take a deep learning approach. CNN-Peaks (Oh et al., 2020) uses supervised convolutional neural networks (CNN) to call ChIP-seq peaks. In addition to the read count information obtained from BAM files, it uses genome annotation information, such as protein-coding transcripts, to improve estimation of peak locations. In their CNN architecture, filters of various sizes are used to extract diverse features and a weighted cross-entropy loss is adopted to account for the imbalanced labels. LanceOtron (Hentges et al., 2021) is another supervised CNN based deep learning method that can be used on ATAC-seq, ChIP-seq, and DNase-seq data. It feeds the output of a logistic regression, fit to 11 enrichment scores predicting labeled peaks, the output of a CNN, fit to fragment coverage in 2000 bp windows predicting labeled peaks, and the 11 enrichment scores to a multilayer perceptron to produce the overall peak score. Many of the false-positive peaks generated by other peak callers are filtered out by these supervised deep learners, increasing precision by about 18% (Hentges et al., 2021). Unfortunately, these supervised methods require labeled data for model training, which are often hard or costly to obtain.

None of these methods consider biological replicates, and in fact most peak calling methods assess biological replicates separately (Goren et al., 2018). HMMRATAC and some users of MACS recommend combining multiple replicates to increase signal, but joint analysis of multiple biological replicates could improve the power to distinguish actual transcription factor binding events (Newell et al., 2021), since some weak or highly variable peak signals may only become evident across multiple replicates (Zhang et al., 2014). One common approach for assessing reproducibility from replicates uses the Irreproducible Discovery Rate (IDR), which identifies reproducible peaks by measuring the consistency in peak ranks between replicates (Li et al., 2011). ChIP-R (Newell et al., 2021), which shows improvement over IDR and can handle more than two replicates, uses the rank product to evaluate the reproducibility across any number of ChIP-seq or ATAC-seq replicates.

We introduce a novel unsupervised learning method that uses contrastive learning (Le-Khac et al., 2020) across replicates to separate genomic regions into peaks and non-peaks. The proposed peak calling framework combines signals from multiple replicates to identify chromatin accessible regions with ATAC-seq data, and overcomes excess noise and lack of labels to make better inferences than existing methods.

3.3 Results

3.3.1 The RCL algorithm

In this study, we developed a peak calling tool (RCL), which contrasts biological replicates to identify the shared signals of replicates to identify high confident ATAC-seq peaks (Figure 3.1). Peak calling is a difficult task, where the genomic extent and significance of enrichment, together the *peak*, must be inferred. Our proposed method separates these tasks, first liberally identifying candidate regions of possible enrichment, and then learning how to score and classify data extracted from the regions. The learner makes no attempt to learn peak boundaries, so its predictions are passed back to the original candidate regions, which become peak predictions if sufficiently high scoring.

3.3.1.1 Prediction region selection.

In general, the RCL framework is applicable for replicated experiments. The individual BAM files of R replicates ($R \geq 2$) and a merged BAM file are required to identify candidate peak regions. Additionally, two user settable parameters, coverage threshold (t, default: "median", see Step 1 below) and input segment length (α, default: 1,000), affect the number and length of the candidate regions. Given these inputs, candidate peak regions are identified as follows:

Step 1: Retain genome positions with coverage $> t$ in all R individual BAM files. Threshold t defaults to a chromosome-specific value obtained from the input data. Specifically, the read coverage on each chromosome is calculated using BEDTools *genomecov* (Quinlan and Hall, 2010) for every replicate (bedtools genomecov -ibam bamFiles -pc -bga), then median coverage across all nonzero positions per chromosome is obtained. The minimum median observed across replicates for a chromosome is used as the threshold for that chromosome. Alternatively, t can be set as a single integer value to be used for every chromosome.

Step 2: Contiguous retained sites are aggregated into regions. Then, regions within 90 base pairs (bp) are merged, since DNA linkers are known to be 8–90 bp (Singh and Mueller-Planitz, 2021). All regions longer than 100 bp are retained for Step 3. Define this set of regions as $\mathcal{A}$.

Step 3.1: If a region in set $\mathcal{A}$ is shorter than α, an α bp long genomic segment is obtained by extending $\frac{\alpha}{2}$ bp up- and down-stream of its midpoint.

Step 3.2: For regions in set $\mathcal{A}$ longer than α bp, we first get positions with coverage summed across replicates ≥ 0.95 quantile of the region (obtained from the merged BAM file). Positions within α bp are merged, then we extended $\frac{\alpha}{2}$ bp up- and down-stream of each merged region's midpoint.

Hereafter, "segment" refers to these selected α bp genomic fragments. Any segment overlapping with a blacklist region (Amemiya et al., 2019) by at least 1 bp is removed. In the end, per-base coverage vectors for these length α bp segments from R replicates are the inputs to RCL.

3.3.1.2 Unsupervised learner.

Given the candidate peak regions, we use a neural network to assign a score to each segment. These segment scores are combined into a single score for the candidate region (see "RCL peak calling" in section Method comparison for details), and sufficiently high-scoring regions are called peaks. As illustrated in Figure 3.1, our method consists of three jointly learned components along with their respective losses, so the total minimized loss is

$$L = l_1 + l_2 + l_3,$$

shown without optional weights that can be tuned by standard cross-validation methods. The three components are a cross-replicate contrastive learner (Le-Khac et al., 2020), a segment class (peak/non-peak) learner (Zhong et al., 2020), and an autoencoder (Kramer, 1991). The input to the contrastive learner and segment class learner is the output of the encoder network that maps the α bp coverage data to a lower-dimensional representation space.

Encoder. With R replicates of observed coverage in S segments, the input data are per-base coverage vectors $\boldsymbol{m}_{ri}, r \in \{1, \ldots, R\}, i \in \{1, \ldots, S\}$. We use ResNET (He et al., 2016) as the backbone of our encoder network. A ResNET module is composed of three basic blocks followed by one residual block. A basic block is composed of a 1D convolutional layer (default: dilation 8 and kernel size 31), followed by a RELU activation function. Our whole encoder $e(\cdot)$ is made of five such ResNET modules, producing the lower dimensional (default dimension: 50) representation $\boldsymbol{x}_{ri} = e\left(\boldsymbol{m}_{ri}\right)$.

Replicate-wise contrastive learning. We use the latent space representations $\boldsymbol{x}_{ri}$ for computing the cross-replicate contrastive loss. We follow SimCLR (Chen et al., 2020), where the replicates are augmentations and the same segments across replicates are positive examples, otherwise they are negative examples. The pairwise replicate contrastive loss, l_1,

$$-\frac{1}{S \times \binom{R}{2}} \sum_{i=1}^{S} \sum_{1 \le r' \le r \le R} \log \frac{\exp\left(\frac{\boldsymbol{x}_{ri}^{\top} \boldsymbol{x}_{r'i}}{\|\boldsymbol{x}_{ri}\| \|\boldsymbol{x}_{r'i}\|} / \tau_1\right)}{\sum_{j \neq i}^{S} \exp\left(\frac{\boldsymbol{x}_{ri}^{\top} \boldsymbol{x}_{r'j}}{\|\boldsymbol{x}_{ri}\| \left\|\boldsymbol{x}_{r'j}\right\|} / \tau_1\right)}, \tag{3.1}$$

where τ_1 is the temperature hyperparameter (default: 0.5), aims to learn lower-dimensional representations such that positive examples are close and negative examples are distant in the new space.

Segment class learning. Assuming the actual peak/non-peak status of genomic segments is shared across replicates and there are underlying characteristics of coverage that define peaks and non-peaks, we expect the low-dimension representation of peak segments to cluster together and separate from the non-peak segments in the new space. Therefore, we also require the representations to match in discrete (classification) space, which we achieve by requiring peak probabilities for each segment to be similar across replicates. The embedded representations $\boldsymbol{x}_{ri}$ are reduced to two dimensions via a fully connected neural network (multilayer perceptron, MLP, in the following) with one hidden layer the same dimension as $\boldsymbol{x}_{ri}$, followed by the softmax function, together denoted as $g(\cdot)$. Letting $\boldsymbol{q}_{ri} = g(\boldsymbol{x}_{ri})$ be the peak/non-peak probabilities for segment i in replicate r and $\boldsymbol{p}'_{rk} = (q_{r1k}, q_{r2k}, \ldots, q_{rSk}), k \in \{1, 2\}$, vectors of peak/non-peak probabilities across segments for the rth replicate, we maximize similarity in peak calls among replicates using loss l_2

$$-\frac{1}{2 \times \binom{R}{2}} \sum_{k=1}^{2} \sum_{1 \le r' \le r \le R} \log \frac{\exp\left(\frac{\boldsymbol{p}_{rk}^{\top} \boldsymbol{p}_{r'k}}{\|\boldsymbol{p}_{rk}\| \|\boldsymbol{p}_{r'k}\|} / \tau_2\right)}{\exp\left(\frac{\boldsymbol{p}_{rk}^{\top} (\boldsymbol{1} - \boldsymbol{p}_{r'k})}{\|\boldsymbol{p}_{rk}\| \|\boldsymbol{1} - \boldsymbol{p}_{r'k}\|} / \tau_2\right)}, \tag{3.2}$$

where temperature hyperparameter $\tau_2 = \tau_1$ in our experiments. This loss strengthens the shared peak signal across replicates and provides a peak/non-peak prediction for each segment of each replicate.

Autoencoder learning. We also want to produce cleaner data in the original genomic space, useful for purposes such as visualization or replicate merging. Therefore, we use decoder $d(\cdot)$, with structure symmetric to the encoder $e(\cdot)$, to map $\boldsymbol{x}_{ri}$ back to predicted data $\hat{\boldsymbol{m}}_{ri}$ in the genomic space. An autoencoder has good embedded feature representation capability (Baldi,

2012), learned by minimizing the squared error loss l_3,

$$\frac{1}{S \times R} \sum_{i=1}^{S} \sum_{r=1}^{R} \text{MSE}(\boldsymbol{m}_{ri}, \hat{\boldsymbol{m}}_{ri}), \tag{3.3}$$

between the original data $\boldsymbol{m}_{ri}$ and the reconstructed data $\hat{\boldsymbol{m}}_{ri}$.

3.3.2 Performance benchmarking using ChromHMM annotations

We compared the performance of RCL to both unsupervised (MACS, ChIP-R and HMMRATAC) and pre-trained supervised (LanceOtron) peak callers, where we used data from four human cell lines, MCF-7, A549, K652 and GM12878, and one dataset generated from mouse placenta tissues at embryonic day 9.5. The datasets are summarized in Table 3.1.

The RCL method involves one important tunable parameter – the coverage threshold t (option $-t$) used to identify the candidate peak regions and segments for model training. By default, RCL uses a chromosome-specific threshold that depends on the median coverage (see section Prediction region selection). In all datasets, in addition to using this default setting, we also implemented RCL with *-t 2* to explore the impact of this tuning parameter. In datasets with higher library size (MCF-7, K562 and A549), chromosome-specific thresholds generally exceed two (Supplementary Figure S1); default thresholds for lower library size datasets (GM12878 and mouse placenta) for all chromosomes are one. We observed manual lowering of threshold t increases the number of candidate regions supplied to the RCL model.

Across all tested datasets of varying library size, RCL achieved the best overall performance (Tables 3.2, 3.3). Sporadically, HMMRATAC, LanceOtron or ChIP-R achieved higher precision at the cost of much lower recall. As the threshold t decreases, RCL predicts more peaks with lower precision and higher recall. Overall, the model trained with lower threshold (RCL-C2 for MCF-7, K562, and A549; RCL-MED for GM12878 and mouse placenta) achieved universally better F1 scores, suggesting that exposure to more low coverage regions can help RCL distinguish true peaks. In these comparisons, MACS and ChIP-R peaks were called with q-value 0.05, but HMMRATAC, LanceOtron, and RCL peaks were called without false discovery control.

HMMRATAC calls should be filtered by the score (Tarbell and Liu, 2019), typically a measure of coverage, and it is similarly advisable to filter RCL calls for higher precision.

Precision recall (PR) curves are useful for comparing methods across all false discovery rates (Figure 3.2, Supplementary Figure S2). We applied a relaxed q-value threshold (0.5) to generate candidate peaks for MACS and ChIP-R, and post-hoc thresholded to plot the curves. Since MACS and ChIP-R are usually run with smaller q-values and no post-hoc thresholding, we also plot precision and recall point estimates for typical choices of q (methods labeled "multiQ"). The linear portion of each PR curve from the black dot to 100% recall corresponds to the subset of ChromHMM-labeled regions with no score assigned by the method. The RCL PR curve, especially with lower threshold t, dominates the curves of other methods. RCL appears to use replicate information in the coverage data better than ChIP-R's post-hoc comparison of peak calls across replicates, which is generally better than naive aggregation of MACS calls. HMMRATAC achieves intermediate performance (Table 3.2, 3.3), with higher achievable recall than MACS and ChIP-R of weak peaks, but sometimes lower achievable precision on strong peaks (Figure 3.2). HMMRATAC also performs poorly on data with lower library size, probably because coverage data are too sparse, when partitioned by fragment length, to estimate this parameter-rich model. Despite the high number of predicted peaks for K562 and A549 data (Table 3.1), RCL maintained good precision out to much higher recall. In particular, RCL achieved nearly twice as many true predictions while maintaining higher precision than either MACS or LanceOtron. While ChIP-R, and sometimes HMMRATAC, can achieve near equal performance on the strongest peaks, only RCL can maintain high precision on the more difficult peaks. For lower library size GM12878 and mouse data, all methods called a limited number of peaks, with low achievable recall. Nevertheless, RCL still obtained better performance, except at the highest achieved recall, where LanceOtron had higher precision. The slightly lower PRAUC of RCL on these data (Table 3.2) may yet be overcome by allowing non-integer threshold values t on the average coverage across multiple sites.

3.3.3 Performance benchmarking using transcription factor ChIP-seq data

In addition to genome annotations obtained with ChromHMM, we used transcription factor (TF) ChIP-seq data to evaluate method performance. These data mark potential binding sites of various TFs, which bind where DNA is accessible and should coincide with ATAC-seq peaks. Due to the lack of TF ChIP-seq data generated in matching conditions, tool performance on the mouse placenta data was not evaluated using this metric. (Table 3.4). As observed with ChromHMM labels, RCL precision improved upon lowering the threshold t.

3.3.4 Gene ontology analysis

As ChromHMM and TF ChIP-seq labels do not cover the whole genome and all methods predicted peaks outside these label regions, we analyzed the biological functions of genes associated with peaks called by each method (see section Gene ontology analysis). We expect meaningful peaks to associate with genes that are related to the known functions of the cell types or tissues. For example, we expect MCF-7 peaks to be enriched for processes such as epithelial cell proliferation, migration and invasion, as well as angiogenesis (Comşa et al., 2015). We therefore checked for the enrichment of any gene ontology (GO) term containing words "epithelial", "epithelium", or "angiogenesis". The K562 cell line has antiapoptotic characteristics (Kuželová et al., 2004); therefore, we expect the enrichment of processes related to the negative regulation of apoptosis, and searched for terms that contained the words "apoptosis" or "apoptotic". The cell line A549, a type of lung carcinoma epithelial cell, is an alveolar type II (ATII) cell that secretes surfactant protein to maintain homeostasis (Lee et al., 2018). Hence, processes underlying this cell type are related to terms that include "epithelial", "epithelium" and "surfactant". GM12878 is a human lymphoblastoid cell line generated by transforming primary B cells from peripheral blood with Epstein-Barr virus (EBV) (Bird et al., 1981; Anderson and Gusella, 1984). Therefore, processes involving "B cell" should be enriched if biologically relevant peaks are supplied. Last, in the mouse placenta at day 9.5, the labyrinth layer is actively developing after chorioallantoic attachment finishes; as a result, a dense network of fetal blood vessels are forming within the layer

where nutrients are exchanged (Starks et al., 2021; Cross et al., 2003; Watson and Cross, 2005). In addition, the placenta is comprised mostly of trophoblast cells, which are epithelial-like cells. Thus, processes related to "placenta", "epithelium", "vasculature", "angiogenesis", "labybrinth" and "insulin" should be expected in meaningful peaks from day 9.5 mouse placenta tissue.

In general, we observed that only peaks uniquely called by RCL are enriched with relevant biological terms, with the exception of A549 data (Figure 3.3, Supplementary Figures S3–S7, Supplementary Tables S2–S6). For example, peaks that only RCL identified were associated with processes related to apoptosis in the K562 dataset (Figure 3.3). In case RCL benefitted from simply predicting a higher number of peaks, we randomly downsampled all peak sets and repeated the analysis. RCL continued to enrich on functionally relevant processes (Supplementary Figures S3–S7, Supplementary Tables S2–S6). For relevant terms, RCL peaks are often associated with at least five genes and have higher than two-fold enrichment (vertical line, Figure 3.3, Supplementary Figures S3–S7) unlike the other methods, suggesting they are more likely to be associated with relevant genes than peaks identified by competing methods. In summary, there is evidence that unique peaks predicted by RCL, not just those overlapping ChromHMM- or TF-derived labels, are biologically relevant.

3.4 Discussion

We propose RCL, an unsupervised peak caller for ATAC-seq data using contrastive learning across biological replicates. In our model, three losses: replicate similarity loss, class similarity loss and autoencoder loss, are learned simultaneously. We use ResNET as our backbone module with only five layers, making the network architecture shallow but efficient. On a server containing two Tesla V100 (16 GB) GPUs, the training time is 118 seconds when there are 4,828 1,000 bp regions and four replicates. Empirical results indicate RCL training time is roughly linear in the number of segments. In theory, training time is quadratic in the number of replicates because of the contrastive loss calculation, but replicate numbers remain quite low. Further investigation on timing is warranted, but total run times were acceptable on all datasets tested

here. For example, for the A549 dataset (the largest dataset and slowest to train), the training time (25 epochs) took about 62 minutes.

In practice, only a small proportion of the genome is accessible (Dunham et al., 2012). As a result, datasets for peak calling tend to be highly imbalanced, making it challenging to separate peak and non-peak regions. RCL showed no problems with class imbalance, probably because the region selection step effectively discards nonpeak regions and balances the data. If class imbalance proves to be a problem for calling datasets with sparser peaks or more widely across the genome in high coverage datasets, there are opportunities for improvement. For example, due to similarities to deep embedding clustering (Xie et al., 2016), cluster regularization methods proposed to avoid local optima or trivial solutions favoring predictions of the larger class (Zhong et al., 2020; Tao et al., 2018), may be applicable to contrastive learning and RCL.

Highly variable peaks or peaks in low coverage data may be difficult to find from single replicates, but their signal may become obvious when comparing across multiple replicates. HMMRATAC utilizes multiple replicates by combining them, which reduces the variance in the signal, but does not help the method learn what defines noise in a single replicate. ChIP-R, a post-hoc method to combine peaks called by another method, can improve performance over MACS, but only when used with a liberal q-value threshold followed by furthering filtering of ChIP-R-predicted peaks (red PR curves in Figure 3.2). Although both MACS and RCL make predictions for individual replicates, RCL predicts after learning from all replicates, while MACS predicts after learning from only the replicate in question. Currently, we combine the RCL prediction scores by taking the mean across replicates, but one can imagine more sophisticated approaches to combine predictions across replicates, possibly assessing the quality of prediction from each replicate and weighting the mean.

Replicates are, by design, an essential component of our method. To demonstrate the value of biological replicates we conducted an ablation study (Supplementary Text S3.1.2). Contrasting real biological replicates gave the best predictions across chromosomes, which is not surprising given that biological replicates are fundamental for reproducibility and false signal reduction (Yan

et al., 2020). In the absence of biological replicates, contrasting with an augmentation of the available data is better than contrasting with self. It could be that noise along the genome recapitulates some of the noise between biological replicates, but more study is necessary to understand RCL performance in the absence of replicates. Experiments varying the number of replicates available to RCL showed little effect on performance, even when the added replicate had substantially higher coverage (Supplementary Text S3.3). All the experimental data examined in the current study have used high quality data with minimal batch effects and samples mostly taken from cultured cells with likely little biological variation, all of which may explain the limited impact of additional replicates. It will be an interesting future direction to examine how contrastive learning and the RCL framework handle batch effects or the inclusion of low quality replicates.

We acknowledge that the labeled regions indicating the "ground truth" used for assessment are noisy. First, the annotations obtained using ChromHMM (Ernst and Kellis, 2017) applied to several ChIP-seq datasets and thus innately contain technical noise from data generation and model estimation. The TF ChIP-seq labels were specifically called by MACS2 (Zhang et al., 2008), which we know produces noisy, imperfect labels. Second, while we matched cell types and biological conditions, variation in the samples used to generate TF ChIP-seq or ChromHMM labels were not completely controlled. Third, our translations from ChromHMM states to open/closed regions were imperfectly determined to the best of our knowledge. There appear to be noisy truth labels in the MCF-7 data. Some negative ChromHMM regions were assigned high scores (logit-transformed scores > 10) (Supplementary Figure S9). Although these score assignments could be due to the shortcomings of RCL, it is also possible that some labels are wrong. Further investigation will be enabled when the quality of labels is improved.

When there is noise in the labels, the observed performance metrics (precision, recall, F1, and PR curve) are not equal to the true performance metrics evaluated against the truth (Jiang et al., 2014). Most worryingly, the observed recall is a function of true recall *and* the true false positive rate. Specifically, let $\hat{y}$ be predicted labels, y unobserved true labels, and z observed noisy labels.

Further, suppose the labeling error rates $P(z = 0 \mid y = 1) = P(z = 1 \mid y = 0) = \epsilon$ are constant and independent of any signal in the data. Then, the observed recall is

$$P(\hat{y} = 1 \mid z = 1) = P(\hat{y} = 1 \mid y = 1)P(y = 1 \mid z = 1) + P(\hat{y} = 1 \mid y = 0)P(y = 0 \mid z = 1), \quad (3.4)$$

where $P(\hat{y} = 1 \mid y = 0)$ is the true false positive rate (FPR). Thus, observed recall is a contaminated measure of recall, and methods compared via observed recall (or F1 or PR curve) may not reveal their actual ranking. Given this concern, it is possible to estimate method performance *in the context of label errors* (Raykar et al., 2009; Yan et al., 2014) or correct errored labels so traditional assessment metrics are more accurate (Sabetpour et al., 2021; Zheng et al., 2021). Alternatively, performance evaluation can be carried out with simulated data. However, there is no existing simulation method for ATAC-seq data, and the existing methods used for ChIP-seq, such as (Zheng et al., 2022), are not applicable for ATAC-seq.

RCL learns and predicts on fixed sized segments (length α, default 1,000 bp). We did not examine the impact of hyperparameter α on RCL performance, but it certainly complicates peak calling. We chose to transfer RCL prediction scores from the α bp segments to the variable-length candidate regions produced by the algorithm in section Prediction region selection, because it works well (Supplementary Text S3.4). Using these candidate regions with mean coverage as a simple score already does well in MCF-7, but RCL learns additional signals, perhaps peak shape, that further improve the performance (Supplementary Figure S11). Not only do the candidate regions work well, but they are not easily substituted. Using the α bp segments as peak predictions in A549 failed, probably because they lack the resolution to pinpoint narrow peaks, but a quick and dirty attempt to shrink the prediction regions to the relevant peak summit performed even worse (Supplementary Figure S11). A better solution may be to learn and predict directly on the variable-sized candidate regions. We could pad variable-sized inputs to the same length or we could add a Spatial Pyramid Pooling layer (He et al., 2015) before the first fully-connected layer to remove the fixed-size constraint of the network. On the other hand, such an approach would still require data preprocessing to choose the candidate regions. An even

better solution might be to predict at the nucleotide level, a one-step solution to identify peaks and their extent.

RCL can be extended and improved in other ways. First, we used simple read coverage as input, but HMMRATAC reports reproducible signal in the coverage of distinct fragment lengths around open regions (Tarbell and Liu, 2019). RCL could be extended to take coverage vectors for multiple fragment lengths, the fragments themselves, or even annotation information, as used by the supervised method CNN-Peaks (Oh et al., 2020). Second, multiple hyperparameters in both data processing and model training can be further tuned. For example in input preparation, regions longer than 100bp are kept in the current method. We have tried keeping regions longer than 147bp, and it resulted in fewer inputs and fewer called peaks; however, we still obtained good predictions. Last, we have focused on ATAC-seq data, where peak calling has been particularly difficult because of the lack of control samples and good truth labels. Nevertheless, our model assumes nothing particular to ATAC-seq data and can be applied to ChIP-seq, CUT&RUN (Skene et al., 2018) and other techniques requiring peak calling.

There is clearly much left to learn about how RCL works to extract useful signal from replicates, but we can offer some preliminary recommendations. First, we recommend users follow established data preprocessing and quality control steps for ATAC-seq data (Yan et al., 2020). Where we have tested, few hyperparameters and inputs of RCL had much impact on performance, other than the coverage threshold t, option `-t`, and the candidate regions. The default threshold ("median") identified highly confident peaks with excellent precision (Figure 3.2); therefore, this setting can be a good starting point for researchers to find the highest confident peaks. If a researcher wishes to predict more peaks accurately, it may be better to reduce threshold t and expose RCL to more and less obvious candidate peaks. This is a particularly good option for high coverage datasets, where RCL reproducibly outshines the competing methods. We recommend using all replicates under the assumption that replicates are still quite sparse because of cost. While additional replicates did not improve performance on datasets tested here, they also did not hurt performance. Finally, we provide no options to

explore alternative prediction regions, but users may have good ideas for choosing candidate regions and they can try them out with the RCL software package.

In summary, we have developed a novel peak calling framework for ATAC-seq data using contrastive learning techniques to extract signals shared across biological replicates and identify high confident open chromatin regions. Because RCL can predict more peaks with higher precision, it will facilitate future epigenome and chromatin accessibility studies in various biological contexts.

3.5 Methods

3.5.1 ATAC-seq data acquisition

ATAC-seq data sets of the following human cell lines and mouse tissues were obtained from public databases: MCF-7, A549, K562, GM12878, and mouse placenta. The MCF-7 dataset, with two biological replicates, was accessed through the ENCODE experiment ID ENCSR422SUG (Dunham et al., 2012). The A549 dataset, with three biological replicates, was accessed through the ENCODE experiment ID ENCSR032RGS (Dunham et al., 2012). The K562 dataset, with three biological replicates, was accessed through the ENCODE experiment ID ENCSR868FGK (Dunham et al., 2012). The GM12878 dataset generated using $50,000$ cells was obtained from four replicates with the accession numbers SRR891268, SRR891269, SRR891270 and SRR891271 (Buenrostro et al., 2013). Last, the mouse data generated from mouse placenta at day 9.5, with three biological replicates, was accessed with the accession numbers SRR7912013, SRR7912014 and SRR7912015 (Starks et al., 2019).

3.5.2 ATAC-seq data processing

FASTQ files were assessed using FastQC (Andrews, 2010) (version 0.11.7) to identify samples with over-represented sequences or adapter contamination. Trimmomatic (Bolger et al., 2014) was used to remove adapter content and filter low quality base pairs and reads (ILLUMINACLIP:overpresentedSeq.fa:2:30:10:2:keepBothReads LEADING:3 TRAILING:3

MINLEN:36, other settings: default, version 0.39). Here, the overpresentedSeq.fa file contains the over-represented sequences and adapter content identified with FastQC. Reads were aligned to the autosomal and sex chromosomes of human reference genome GRCh38 or mouse reference genome GRCm38 (release 98) (Cunningham et al., 2019) using Bowtie2 (Langmead and Salzberg, 2012) (-X 1000 --no-discordant, other settings: default, version 2.3.4.1). The genome reference versions were matched with the label files used for performance assessment, downloaded from public databases (see Method comparison). As performance comparisons were carried within organisms with matching references, using different genome versions should not affect the conclusion of the study. Picard (Broad Institute, 2019) was used to remove duplicate reads (REMOVE_DUPLICATES=true, version 2.17.0). Reads with low quality mapping (MapQ < 20) were removed before merging, sorting, and indexing the resulting BAM files with SAMtools (Danecek et al., 2021). Last, to assess sample quality after preprocessing, ataqv (Orchard et al., 2020) (--ignore-read-groups, other settings: default, version 1.2.1) was used to check for fragment length distribution and transcription start site (TSS) enrichment. Samples used for downstream analyses must have a mononucleosome peak in the fragment length distribution, and TSS enrichment ≥ 1.5.

3.5.3 Tuning RCL

We used dilation 8 and kernel size 31 to train our model. Other hyperparameters are set to be default values (number of epochs = 25, batch size = 256, learning rate = 10^{-4} and temperature $\tau_1 = \tau_2 = 0.5$). Details regarding choosing dilation 8, kernel size 31 and model development were are discussed in Supplementary Text S3.1. Briefly, RCL was developed on the MCF-7 cell line data (see section ATAC-seq data acquisition) using different truth labels than those presented in results. We will demonstrate that a *roughly* tuned RCL is already substantially superior to existing methods, not only on MCF-7 with a distinct truth, but on additional holdout datasets as well.

3.5.4 Method comparison

We compared RCL to MACS (Zhang et al., 2008), ChIP-R (Newell et al., 2021), HMMRATAC (Tarbell and Liu, 2019), and LanceOtron (Hentges et al., 2021). Call performance was assessed using three analyses: comparisons using truth labels of genome annotation obtained with ChromHMM (Ernst and Kellis, 2017) from independent data collected on the same cell lines and tissues; comparisons using truth labels of transcription factor ChIP-seq data collected on the same cell lines and tissues; and the association of peak prediction to biologically relevant genes.

MACS peak calling. MACS (version 2.1.1) (Zhang et al., 2008; Gaspar, 2018) was used to call peaks with BAM files from individual replicates. Peaks were called with options -g hg -f BAMPE --bdg --keep-dup all, with the following cut-offs for the q-value q: 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00002, and 0.00001, and other settings: default. Any peaks overlapping with a blacklist region (Amemiya et al., 2019) by at least 1 bp are removed. MACS was originally developed for calling peaks on transcription factor ChIP-seq data, so the default settings and model assumptions may not apply for ATAC-seq data. We have some evidence that altering shift and window sizes can improve MACS performance in some aspects (Supplementary Text S3.5), but settings to consistently improve MACS performance were elusive and beyond the scope of this work. Given peak calls from individual replicates, the peak union method was used to combine peaks across replicates. Specifically, a consensus peak set is the union of peaks overlapping with each other by $\geq 50\%$ length in ≥ 2 replicates. Scores of consensus peaks were the mean $-\log_{10}(q\text{-value})$ at peak summit of the individual peaks observed in separate replicates. As MACS does not report scores of non-peak regions, replicates not calling a peak in the region are not used when calculating scores of consensus peaks.

ChIP-R peak calling. We used ChIP-R (version 1.1.0) (Newell et al., 2021) as an additional, independent method for combining peaks called from MACS. Peaks were first called with MACS as described above. Then, ChIP-R was run with the following setting: -m 2, -a 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001,0.00002, and 0.00001 (matching with $-q$

in MACS), other settings: default. Any peaks overlapping with a blacklist region (Amemiya et al., 2019) by at least 1 bp are removed. The reported *score* was used as scores of ChIP-R peaks.

HMMRATAC peak calling. HMMRATAC (version 1.2.4) (Tarbell and Liu, 2019) was used to call peaks with a merged BAM file from all replicates, options
-Xmx128G, --window 250000, other settings: default. A peak is a region in the open state with scores ≥ 0, reported by default with the --peaks option. By default, peak scores of HMMRATAC are the maximum read coverage of the called center state region. Any peaks overlapping with a blacklist region (Amemiya et al., 2019) by at least 1 bp were removed.

LanceOtron peak calling. To implement LanceOtron (Hentges et al., 2021) (version 1.0.8), the input bigwig files were obtained using deeptools (version 2.5.2) (Ramírez et al., 2016) with the following command:
bamCoverage -b bamFile -o bigwigFile --extendReads -bs 1 --normalizeUsingRPKM. The inputs were then used to call peaks with default settings and LanceOtron's pretrained model. Any resulting regions overlapping with a blacklist region (Amemiya et al., 2019) by at least 1 bp are removed. Then, the union of regions overlapping with each other by $\geq 50\%$ length in $\geq$ two replicates were obtained as a consensus region set. Scores of consensus regions were the mean overall_peak_score of the individual regions. Regions with scores > 0.5 were then defined as peaks by default.

RCL peak calling. RCL was used with coverage threshold t of "median" and 2, other settings: default. By default, any segment overlaps with blacklist regions was excluded due to the segment selection procedure. Let peak prediction score given by RCL be $\xi_{ri} = q_{ri1}$ for the ith α bp segment in the rth replicate. We obtain a final peak prediction score for each region in $\mathcal{A}$ (see Step 2 in 3.3.1.1 Prediction region selection) by averaging over ξ_{ri} for all replicates $r = 1, 2, \ldots, R$ and segments i extracted from the region. A region in $\mathcal{A}$ is predicted to contain at least one peak if this score is > 0.5.

Compilation of true positive and true negative labeled regions by ChromHMM. For human cell line data, we obtained genome annotations inferred with ChromHMM (Ernst and Kellis, 2017) from ENCODE. Specifically, genome annotation for MCF-7 data was accessed via the experiment ID ENCSR579CCH, the A549 data via ENCSR283FYU, and the GM12878 data via ENCSR988QYW. True positive regions are those marked "EnhA1", "EnhA2", "EnhG1", "EnhG2", "TssA", "TssFlnk", "TssFlnkD", "TssFlnkU", and "Tx". True negative regions are marked as "Het", "Quies", "ReprPC", and "ZNF/Rpts". Annotations not in these lists were not used, and regions overlapping with a blacklist region (Amemiya et al., 2019) by at least 1 bp were removed. For full definitions of the states, see Supplementary Table S7.
For the mouse placenta data, we obtained ChromHMM annotation from (Starks et al., 2021). True positive regions are those belonging to State 8, 9 and 10, and true negative regions are those belonging to State 2. Detailed biological characterization of these states were described in (Starks et al., 2021).

Compilation of true positives from transcription factor (TF) ChIP-seq data. For human cell line data, we obtained TF ChIP-seq data from matching cell lines from ENCODE (Dunham et al., 2012). Bed files of IDR thresholded peaks were downloaded from all datasets that passed all quality control criteria of ENCODE and had "released" status, and their bio-samples were not perturbed. For mouse placenta data, no TF ChIP-seq data from matching conditions was available. Therefore, this analysis was not carried out for mouse placenta data. True positive regions were defined as those with at least one TF ChIP-seq peak. Regions overlapping with a blacklist region (Amemiya et al., 2019) by at least 1 bp were removed. No true negative regions were defined using these datasets. For lists of data used, see Supplementary Table S7.

Calculation of evaluation metrics. Since labelled regions and called regions do not necessarily coincide, we defined a mapping function to transfer scores of called regions in the ATAC-seq data to predicted scores for annotated regions. Specifically, suppose there are n_i called

regions overlapping with the ith labeled region, and c_j $(1 \leq j \leq n_i)$ is the predicted score that overlaps by o_j base pairs with the ith labeled region. Then the weighted prediction score for the ith labeled region is $\sum_{j=1}^{n_i} \frac{o_j c_j}{\sum_{v=1}^{n_i} o_v}$. In case $n_i = 0$, we assign the lowest weighted score observed for that method as the predicted score.

We used point estimates of precision, recall and F1, as well as Precision-Recall (PR) curves to compare the performance of the methods. Specifically, for MACS and ChIP-R, we calculated precision and recall for each $-q$ $(-a)$ cut-off. For plotting PR curves, we used MACS or ChIP-R with q-value 0.5. Results from other q-value settings were presented as point estimates on the plots. RCL can call more peaks by lowering the threshold t, but changing t will also change the fitted model and the peak calls made. We always ran RCL with defaults, but we also selected non-default t so the *number* of candidate peaks roughly matched the number of peaks called by MACS (q-value 0.5), ChIP-R (q-value 0.5) and HMMRATAC.

Gene ontology analysis. To assess the potential functional roles of the called peaks by all methods, we used gene ontology analysis. We examined the following sets of peaks. For the MCF-7, K562 and A549 data, we used ChIP-R, HMMRATAC, LanceOtron, and RCL peaks called when $t = 2$ peaks; for the GM12878 and mouse placenta data, ChIP-R, HMMRATAC, LanceOtron, and RCL peaks called when $t = median$ peaks; and for all datasets, peaks identified uniquely by each of these methods. Specifically, a peak is uniquely assigned to a method if it does not overlap with peaks predicted by any other method, as assessed using BEDTools intersect -v (Quinlan and Hall, 2010).

We used the Genomic Regions Enrichment of Annotations Tool (GREAT) (version 4.0.4) (McLean et al., 2010; Tanigawa et al., 2022) implemented in R (Gu and Hübschmann, 2022) to carry out GO enrichment using either the human GRCh38 or mouse GRCm38 annotations and the default basal plus extension association rule. For each analysis, we randomly selected peaks so that the number of input regions for GREAT was the smallest or second smallest peak set size amongst all tools (see number of peaks from each tool in Supplementary Table S1). For unique peaks, we also analyzed all peaks without down-sampling. A biological

process term was considered enriched if its binomial q-value ≤ 0.05, binomial fold change ≥ 2, and the observed number of associated genes was ≥ 5.

3.6 Main figures and tables

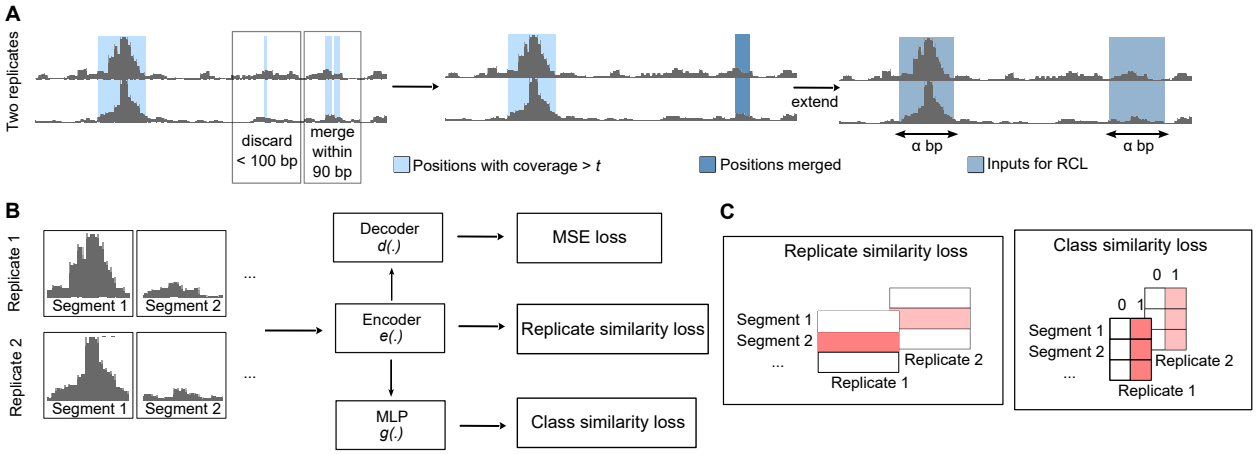

Figure 3.1: RCL model. Segment identifies the same genomic region for all replicates. (**A**) The raw input is processed to extract α-length input segments. (**B**) The α-length input segments are fed to encoder $e(\cdot)$ to compute the cross-replicate contrastive loss. Then the embedding is fed to a multilayer perceptron (MLP), specifically a fully connected neural network, for class similarity loss and a decoder for the autoencoder (MSE) loss. The encoder/decoder has five ResNET blocks. (**C**) Shaded red boxes represent the elements contrasted in the respective losses.

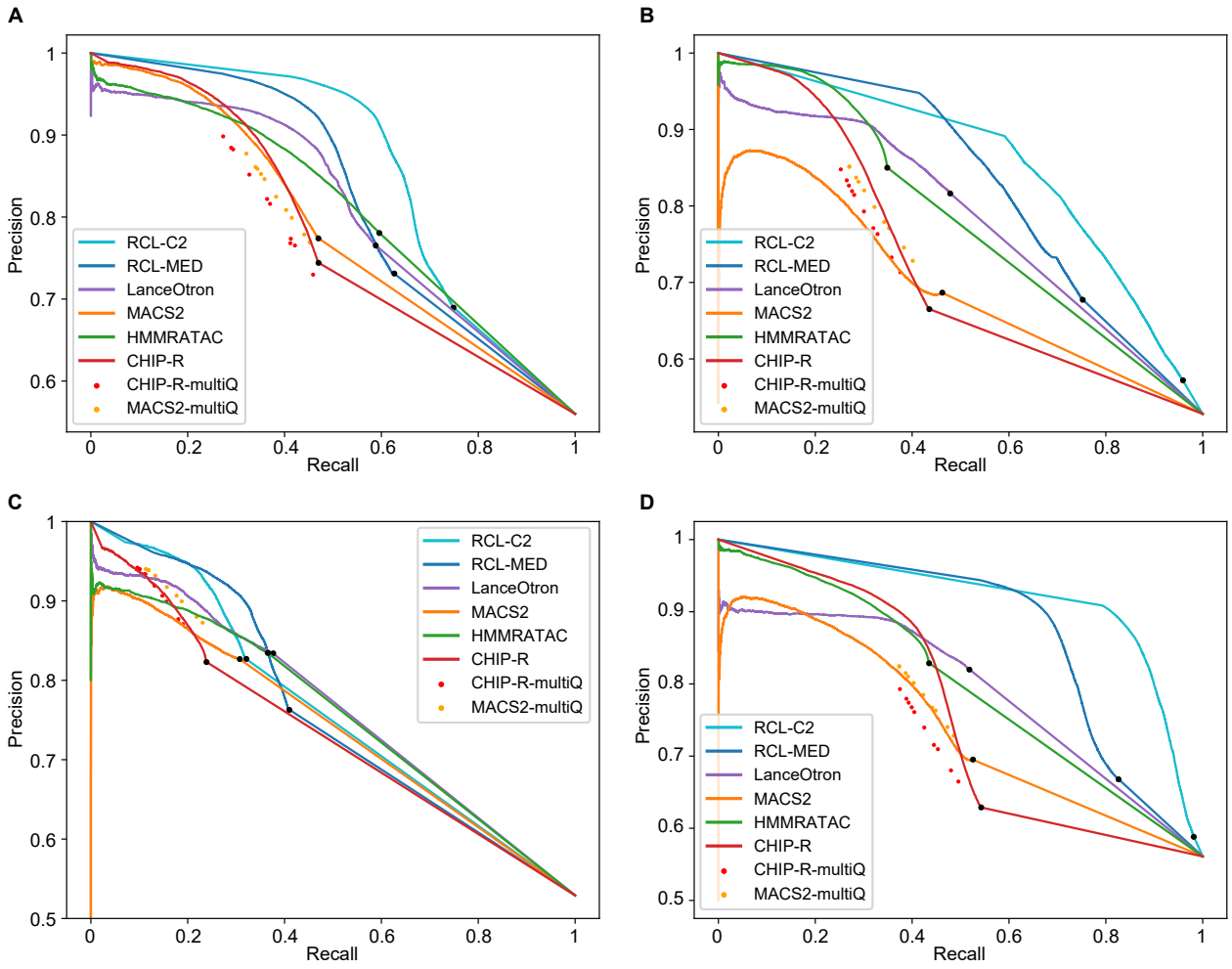

Figure 3.2: Precision-Recall (PR) curves for ChromHMM-labeled regions. Black dot in each curve denotes the region with lowest score; all remaining ChromHMM-labeled regions are not scored the method. RCL-C2, analysis with coverage threshold 2; RCL-MED, analysis with default "median" coverage threshold; MACS-multiQ and ChIP-R-multiQ dots are obtained by varying q-value cut-offs.

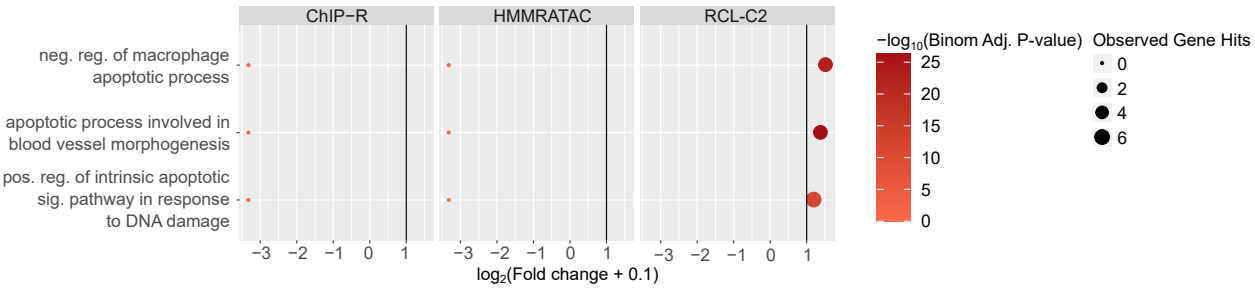

Figure 3.3: Gene ontology analysis using unique peaks called by each method in K562 data. Only relevant terms enriched with at least one peak set are plotted. Colors correspond to $-\log_{10}$(Binomial Adjusted P-value) where the adjustment was done following the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995); dot sizes correspond to the observed number of genes associated with the term; x-axis corresponds to $\log_2$(Fold change+0.1) and vertical line is fold change of two. LanceOtron was not plotted since there was no unique peak called by the tool. Abbreviations: reg., regulation; pos., positive; neg., negative; sig., signaling.

Table 3.1: Datasets used to compare methods on genome-wide annotation regions generated by ChromHMM.

						ChromHMM Labels		Number of Called Peaks in ChromHMM Labels					
					Map.					HMMR	RCL		Lance
Data	Src.	Org.	R	Size	Rate	+	−	MACS	ChIP-R	ATAC	MED	C2	Otron
MCF-7	(Dunham et al., 2012)	Human	2	37M	96%	148,531	116,729	65,440	93,860	113,362	92,333	116,367	90,050
A549	(Dunham et al., 2012)	Human	3	157M	98%	154,274	120,726	116,654	133,121	80,991	148,725	213,263	94,284
GM12878	(Buenrostro et al., 2013)	Human	4	31M	68%	156,817	139,567	58,341	45,498	47,315	59,130	38,637	33,056
K562	(Dunham et al., 2012)	Human	3	100M	96%	212,779	190,127	143,269	139,343	87,284	173,790	227,839	110,011
Placenta	(Starks et al., 2019)	Mouse	3	20M	70%	88,211	18,754	14,598	15,888	6,277	11,332	5,719	4,141

Src., literature source. Org., organism. R, number of biological replicates. Size, mean number of reads across replicates after filtering in the dataset. Map. Rate, median proportion of aligned reads to autosomal and sex chromosomes using Bowtie2. ChromHMM labels, number of positive and negative true regions annotated using ChromHMM. RCL MED (default threshold t based on

median coverage) and C2 (threshold $t = 2$) indicate two different coverage thresholds used to build training segments. Lower threshold results in larger input datasets and provides more predictions, specifically including predictions for lower coverage, harder-to-predict segments. For the total number of peaks including those outside of annotated regions, see Supplementary Table S1.

Table 3.2: Metrics on human cell line datasets

	MCF-7				**A549**				**GM12878**				**K562**			
	Precision	Recall	F1	PRAUC	Precision	Recall	F1	PRAUC	Precision	Recall	F1	PRAUC	Precision	Recall	F1	PRAUC
RCL-C2	0.848	**0.637**	**0.728**	**0.858**	0.791	**0.739**	**0.764**	**0.855**	**0.931**	0.229	0.368	0.763	0.686	**0.948**	**0.796**	**0.914**
RCL-MED	**0.909**	0.422	0.576	0.818	0.778	0.636	0.700	0.830	0.882	**0.333**	**0.483**	0.761	0.782	0.754	0.768	0.874
HMMRATAC	0.781	0.596	0.676	0.808	**0.850**	0.349	0.495	0.783	0.850	0.256	0.393	0.738	0.827	0.435	0.571	0.799
MACS	0.779	0.441	0.563	0.748	0.746	0.383	0.506	0.695	0.880	0.218	0.350	0.739	0.740	0.473	0.577	0.746
ChIP-R	0.768	0.412	0.536	0.781	0.733	0.358	0.481	0.725	0.877	0.181	0.300	0.734	0.680	0.480	0.563	0.762
LanceOtron	0.842	0.510	0.635	0.809	0.842	0.435	0.574	0.783	0.911	0.192	0.317	**0.764**	**0.828**	0.506	0.628	0.791

Table 3.3: Metrics on mouse placenta dataset

	Mouse Placenta			
	Precision	Recall	F1	PRAUC
RCL-C2	0.996	0.0648	0.122	0.926
RCL-MED	**0.999**	**0.128**	**0.227**	**0.927**
HMMRATAC	0.993	0.071	0.132	0.915
MACS	**0.999**	0.078	0.144	0.923
ChIP-R	**0.999**	0.090	0.166	0.923
LanceOtron	0.991	0.047	0.089	0.921

Table 3.4: Precision against ChIP-seq labels, using human and mouse data

	MCF-7	**A549**	**GM12878**	**K562**
RCL-C2	**0.656**	**0.929**	0.361	**0.649**
RCL-MED	0.541	0.753	**0.496**	0.562
HMMRATAC	0.593	0.560	0.357	0.227
MACS	0.571	0.682	0.422	0.329
ChIP-R	0.531	0.675	0.422	0.327
LanceOtron	0.547	0.549	0.269	0.257

Performance evaluation metrics. Table 3.2 and 3.3: Precision, recall, F1 scores and PRAUC (area under the PR curve). To compute precision, recall, and F1 scores for MACS and ChIP-R, a

q-value of 0.05 was used. To compute PRAUC for MACS and ChIP-R, a q-value of 0.5 was used and then post-hoc thresholded to obtain a PR curve. All HMMRATAC results were obtained using scores > 0. All RCL and LanceOtron results were obtained using average scores across replicates > 0.5. Table 3.4: Precision using transcription factor (TF) ChIP-seq as labels.

3.7 References

Amemiya, H. M., Kundaje, A., and Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, 9(1):1–5.

Anderson, M. A. and Gusella, J. F. (1984). Use of cyclosporin a in establishing epstein-barr virus-transformed human lymphoblastoid cell lines. *In Vitro*, 20.

Andrews, S. (2010). FastQC - A quality control tool for high throughput sequence data. *Babraham Bioinformatics*.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, ICML '12, pages 37–49, Edinburgh, Scotland. JMLR Workshop and Conference Proceedings.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bird, A. G., McLachlan, S. M., and Britton, S. (1981). Cyclosporin a promotes spontaneous outgrowth in vitro of epstein–barr virus-induced b-cell lines. *Nature*, 289.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Broad Institute (2019). Picard toolkit. `https://broadinstitute.github.io/picard/`.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109(1):21.29.1–21.29.9.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Comşa, Ş., Cîmpean, A. M., and Raica, M. (2015). The story of MCF-7 breast cancer cell line: 40 years of experience in research. *Anticancer Research*, 35(6):3147–3154.

Cross, J. C., Simmons, D. G., and Watson, E. D. (2003). Chorioallantoic morphogenesis and formation of the placental villous tree. *Annals of the New York Academy of Sciences*, 995(1):84–93.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Marugán, J. C., Maurel, T., McMahon, A. C., Moore, B., Morales, J., Mudge, J. M., Nuhn, M., Ogeh, D., Parker, A., Parton, A., Patricio, M., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sparrow, H., Stapleton, E., Szuba, M., Taylor, K., Threadgold, G., Thormann, A., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Yates, A. D., Zerbino, D. R., and Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, 47(D1):D745–D751.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2).

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature 2012 489:7414*, 489(7414):57–74.

Ernst, J. and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*, 12(12):2478–2492.

Gaspar, J. M. (2018). Improved peak-calling with MACS2. *bioRxiv*, page 496521.

Goren, E., Liu, P., Wang, C., and Wang, C. (2018). BinQuasi: a peak detection method for ChIP-sequencing data with biological replicates. *Bioinformatics*, 34(17):2909–2917.

Grandi, F. C., Modi, H., Kampman, L., and Corces, M. R. (2022). Chromatin accessibility profiling by ATAC-seq. *Nature Protocols*, 17(66):1518–1552.

Gu, Z. and Hübschmann, D. (2022). rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778, Las Vegas, NV.

Hentges, L. D., Sergeant, M. J., Downes, D. J., Hughes, J. R., and Taylor, S. (2021). LanceOtron: A deep learning peak caller for ATAC-Seq, ChIP-Seq, and DNase-Seq. *Bioinformatics*, 38(18):4255–4263.

Hocking, T. D., Goerner-Potvin, P., Morin, A., Shao, X., Pastinen, T., and Bourque, G. (2017). Optimizing ChIP-Seq peak detectors using visual labels and supervised machine learning. *Bioinformatics*, 33(4):491–499.

Jiang, Y., Clark, W., Friedberg, I., and Radivojac, P. (2014). The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. *Bioinformatics*, 30(17):i609–i616.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243.

Kuželová, K., Grebeňová, D., Pluskalová, M., Marinov, I., and Hrkal, Z. (2004). Early apoptotic features of k562 cell death induced by 5-aminolaevulinic acid-based photodynamic therapy. *Journal of Photochemistry and Photobiology B: Biology*, 73(1):67–78.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Lee, D. F., Salguero, F. J., Grainger, D., Francis, R. J., MacLellan-Gibson, K., and Chambers, M. A. (2018). Isolation and characterisation of alveolar type ii pneumocytes from adult bovine lung. *Scientific Reports*, 8.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, X. Y., Thomas, S., Sabo, P. J., Eisen, M. B., Stamatoyannopoulos, J. A., and Biggin, M. D. (2011). The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biology*, 12(4).

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501.

Musich, R., Cadle-Davidson, L., and Osier, M. V. (2021). Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Frontiers in Plant Science*, 12.

Newell, R., Pienaar, R., Balderson, B., Piper, M., Essebier, A., and Bodén, M. (2021). ChIP-R: Assembling reproducible sets of ChIP-seq and ATAC-seq peaks from multiple replicates. *Genomics*, 113(4):1855–1866.

Oh, D., Strattan, J. S., Hur, J. K., Bento, J., Urban, A. E., Song, G., and Cherry, J. M. (2020). CNN-Peaks: ChIP-Seq peak detection pipeline using convolutional neural networks that imitate human visual inspection. *Scientific Reports*, 10(1):7933.

Orchard, P., Kyono, Y., Hensley, J., Kitzman, J. O., and Parker, S. C. (2020). Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. *Cell Systems*, 10(3):298–306.e4.

Park, P. J. (2009). Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165.

Raykar, V., Yu, S., Zhao, L., Jerebko, A., Florin, C., Valadez, G., Bogoni, L., and Moy, L. (2009). Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 889–896.

Rye, M., Sætrom, P., and Drabløs, F. (2011). A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Research*, 39(4):e25.

Sabetpour, N., Kulkarni, A., Xie, S., and Li, Q. (2021). Truth discovery in sequence labels from crowds. In *2021 IEEE International Conference on Data Mining (ICDM)*, ICDM '21, pages 539–548, Aucklan, New Zealand.

Singh, A. K. and Mueller-Planitz, F. (2021). Nucleosome Positioning and Spacing: From Mechanism to Function. *Journal of Molecular Biology*, 433(6):166847.

Skene, P. J., Henikoff, J. G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature Protocols*, 13(5):1006–1019.

Starks, R. R., Biswas, A., Jain, A., and Tuteja, G. (2019). Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics and Chromatin*, 12(1):1–16.

Starks, R. R., Kaur, H., and Tuteja, G. (2021). Mapping cis-regulatory elements in the midgestation mouse placenta. *Scientific Reports 2021 11:1*, 11(1):1–13.

Tanigawa, Y., Dyer, E. S., and Bejerano, G. (2022). Whichtf is functionally important in your open chromatin data? *PLOS Computational Biology*, 18(8).

Tao, Y., Takagi, K., and Nakata, K. (2018). RDEC: integrating regularization into deep embedded clustering for imbalanced datasets. In *Asian Conference on Machine Learning*, ACML '18, pages 49–64, Beijing, China. PMLR.

Tarbell, E. D. and Liu, T. (2019). HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Research*.

Watson, E. D. and Cross, J. C. (2005). Development of structures and transport functions in the mouse placenta. *Physiology*, 20(3):180–193.

Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, ICML '16, pages 478–487, New York City, NY. PMLR.

Yan, F., Powell, D. R., Curtis, D. J., and Wong, N. C. (2020). From reads to insight: A hitchhiker's guide to ATAC-seq data analysis.

Yan, Y., Rosales, R., Fung, G., Subramanian, R., and Dy, J. (2014). Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327.

Zhang, Y., Lin, Y.-H., Johnson, T. D., Rozek, L. S., and Sartor, M. A. (2014). PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*, 30(18):2568–2575.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137.

Zheng, A., Lamkin, M., Qiu, Y., Ren, K., Goren, A., and Gymrek, M. (2022). A flexible chip-sequencing simulation toolkit. *BMC Bioinformatics*, 22:1518–1552.

Zheng, G., Awadallah, A., and Dumais, S. (2021). Meta label correction for noisy label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11053–11061.

Zhong, H., Chen, C., Jin, Z., and Hua, X. (2020). Deep robust clustering by contrastive learning. *arXiv*.

3.8 Appendix A: Notes

3.8.1 Software availability

The entire pipeline is released under the GNU Public License to the community as a package named RCL, for Replicative Contrastive Learner (GitHub `https://github.com/Tuteja-Lab/UnsupervisedPeakCaller`). The source code can also be found in Supplementary Code file.

3.8.2 Competing interest statement

The authors declared no competing interest.

3.8.3 Acknowledgements

We acknowledge the Research IT group at Iowa State University (`http://researchit.las.iastate.edu`) for providing servers and IT support. We would like to thank Dorman lab and Tuteja lab members for their discussion and support. This work was supported in part by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under award number R01HD096083 (to G Tuteja). G Tuteja is Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts. This work was supported in part by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Hatch project IOW03717. The findings and conclusions in this publication are those of the author(s) and should not be construed to represent any official USDA or U.S. Government determination or policy.

3.9 Appendix B: Supplementary materials

Full supplementary materials can be found online at Vu et al., 2023: https://genome.cshlp.org/content/33/7/1133.full.

3.10 Appendix C: Consent to include co-authored article in thesis/dissertation

THE PARTIES

Student Author (Full Name, Major, and Institution)	Ha T.H. Vu Bioinformatics and Computational Biology Iowa State University, Ames, IA, 50010
List other student co-authors and their institutions.	Yudi Zhang Iowa State University, Ames, IA, 50010
Title(s) of the co-authored section (Chapter, etc.)	Unsupervised contrastive peak caller for ATAC-seq Chapter 3
Journal Name, Book Title, etc. (if applicable)	Genome Research

DISTRIBUTION OF TASKS AND RESPONSIBILITIES

In this research publication, I, Ha T.H. Vu, was responsible for the following roles: (Select all roles that apply.)

☒Conceptualization
☒Data curation
☒Formal analysis
☐Funding acquisition
☒Investigation
☒Methodology
☐Resources
☒Software
☐Supervision
☒Validation
☒Visualization
☒Writing – original draft
☒Writing – review & editing
☐Other: Please describe briefly: Click or tap here to enter text.

The CRediT taxonomy is taken from https://credit.niso.org/. Go to the link to see the descriptions of contributor roles.

CHAPTER 4. CORE CONSERVED TRANSCRIPTIONAL REGULATORY NETWORKS DEFINE THE INVASIVE TROPHOBLAST CELL LINEAGE

Ha T. H. Vu[*,1,2], Regan L. Scott[*,3], Khursheed Iqbal[3], Michael J. Soares[3,4,5], Geetu Tuteja[1,2]

[1]Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, 50011

[2]Bioinformatics and Computational Biology, Iowa State University, Ames, IA 50011

[3]Institute for Reproductive and Developmental Sciences and Department of Pathology & Laboratory Medicine, University of Kansas Medical Center, Kansas City, KS, 66160

[4]Obstetrics and Gynecology, University of Kansas Medical Center, Kansas City, KS, 66160

[5] Center for Perinatal Research, Children's Mercy Research Institute, Children's Mercy, Kansas City, MO, 64108

Modified from a manuscript published in *Development*

4.1 Abstract

The invasive trophoblast cell lineage in rat and human share crucial responsibilities in establishing the uterine-placental interface of the hemochorial placenta. These observations have led to the rat becoming an especially useful animal model to study hemochorial placentation. However, our understanding of similarities or differences between regulatory mechanisms governing rat and human invasive trophoblast cell populations is limited. In this study, we generated single-nucleus (sn) ATAC-seq data from gestation day (gd) 15.5 and 19.5 rat uterine-placental interface tissues and integrated the data with single-cell RNA-seq data generated at the same stages. We determined the chromatin accessibility profiles of invasive trophoblast, natural killer, macrophage, endothelial, and smooth muscle cells, and compared invasive trophoblast chromatin accessibility to extravillous trophoblast (EVT) cell accessibility. In

[*]These authors contributed equally to this work.

comparing chromatin accessibility profiles between species, we found similarities in patterns of gene regulation and groups of motifs enriched in accessible regions. Finally, we identified a conserved gene regulatory network in invasive trophoblast cells. Our data, findings and analysis will facilitate future studies investigating regulatory mechanisms essential for the invasive trophoblast cell lineage.

4.2 Introduction

Hemochorial placentation is a reproductive strategy utilized by some mammals, including the mouse, rat, and human (Roberts et al., 2016). This type of placentation involves establishment of a uterine-placental interface characterized by trophoblast cells of extraembryonic origin breaching the maternal vasculature (Roberts et al., 2016). Trophoblast cells are the parenchymal cells of the placenta (Knöfler et al., 2019; Red-Horse et al., 2004; Soares et al., 2018). Their origins can be traced to the trophectoderm of the early embryo and the initial cell differentiation event during embryogenesis (Gardner and Beddington, 1988; Rossant, 2001). Trophoblast cells differentiate into a range of specialized lineages (Gardner and Beddington, 1988; Knöfler et al., 2019; Soares et al., 2018). Among the specialized trophoblast cell lineages are invasive trophoblast (generic term) or extravillous trophoblast (EVT, human/primate specific term). These cells exit the placenta and enter the uterine compartment where they transform the vasculature and immune environment into a structure ensuring placental and fetal viability and growth (Knöfler et al., 2019; Soares et al., 2018; Turco and Moffett, 2019). Failures in invasive trophoblast/EVT cell differentiation and function result in a range of pregnancy diseases such as preeclampsia, intrauterine growth restriction, and preterm birth (Brosens et al., 2011, 2019). Deep trophoblast cell invasion and uterine transformation are characteristic features of rat and human placentation sites (Pijnenborg et al., 2011; Shukla and Soares, 2022; Soares et al., 2012). Identification of potential regulatory mechanisms controlling cellular constituents of the rodent and human uterine-placental interface have emerged from single-cell RNA-sequencing (scRNA-seq) (Liu et al., 2018; Marsh et al., 2022; Nelson et al., 2016; Scott et al., 2022; Sun et al., 2020; Suryawanshi

et al., 2018; Vento-Tormo et al., 2018). Conserved sets of transcripts have been identified in rat invasive trophoblast and human EVT cells (Scott et al., 2022). These insights have led to the identification of candidate regulators of invasive trophoblast and EVT cell lineages and dissection of their biological relevance using trophoblast stem (TS) cells and rat models (Kuna et al., 2023; Muto et al., 2021; Varberg et al., 2021). Such experimentation has advanced the field but on its own is an inefficient strategy for defining gene regulatory networks driving invasive trophoblast/EVT cell lineage development and function. Gene regulatory networks can be accessed through genome-wide analysis of the chromatin landscape (Ong and Corces, 2012; Peñalosa-Ruiz et al., 2019; Tuteja et al., 2014; Yadav et al., 2018). Indeed, insights into the hierarchical regulation of rodent and human trophoblast cell development have been achieved through deep sequencing of histone modifications defining gene activation and repression states (Chuong et al., 2013; Kwak et al., 2019; Lee et al., 2019; Rugg-Gunn et al., 2010; Schoenfelder et al., 2018; Starks et al., 2021; Tuteja et al., 2016; Xu and Kidder, 2018; Zhang et al., 2021). The integration of transcriptome and chromatin accessibility datasets has also been used as an effective tool to elucidate gene regulatory networks in trophoblast tissue and cells (Nelson et al., 2017; Starks et al., 2019). In this report, we interrogated the chromatin landscape of invasive trophoblast cells isolated from the uterine-placental interface of the rat using single-nucleus assay for transposase-accessible chromatin-sequencing (snATAC-seq). These datasets were integrated with scRNA-seq datasets from rat and human invasive trophoblast/EVT cells (Scott et al., 2022), as well as ATAC-seq from EVT cells (Varberg et al., 2022), to identify conserved gene regulatory networks controlling the invasive trophoblast cell lineage.

4.3 Results

4.3.1 Identification of chromatin accessibility profiles in cell types of the rat uterine-placental interface

We generated snATAC-seq profiles from gestation day (gd) 15.5 and 19.5 uterine-placental interface tissue of the rat to determine chromatin accessibility of its cellular constituents. These

datasets were integrated with scRNA-seq profiles obtained from the same tissues (Scott et al., 2022).

Following quality control and preprocessing (Fig S1, Fig S2), we obtained 25,321 and 14,388 high quality nuclei in the gd 15.5 and gd 19.5 samples, respectively (Table S1). Next, snATAC-seq data was integrated with scRNA-seq data (Scott et al., 2022) to identify cell populations based on the relationship between accessibility and gene expression profiles (Stuart et al., 2019) (Fig 4.1A). Clusters and chromatin accessibility profiles of invasive trophoblast, natural killer, macrophage, endothelial, and smooth muscle cells were identified (Table S1).

These analyses are based on an assumption that there is a significant correlation between gene expression level (scRNA-seq data) and chromatin accessibility (snATAC-seq data) (Stuart et al., 2019). Therefore, as a quality control step for the snATAC-seq cluster labeling, we calculated the Spearman correlation between gene expression and chromatin accessibility profiles. We obtained moderate but significant correlations ($0.44 \leq \rho \leq 0.54$, p-value<2.2e-16) in all cell populations (Fig S3), which agrees with previous studies done at the both single cell and tissue levels (Merrill et al., 2022; Pervolarakis et al., 2020; Starks et al., 2019). Moreover, we observed that established marker gene expression for each cell population are generally more accessible in the respective cell population (Fig 4.1B), demonstrating we have obtained high quality clustering and cluster annotation.

We further performed differential accessibility analysis at both gestation days to identify the most accessible peaks in each cell type (defined as cell type-specific peaks). The distance distribution of cell type-specific peaks to the nearest gene transcription start site (TSS) showed, that in general, most of the cell type-specific peaks are distal to the TSS (>5 kb) (81.65% at gd 15.5, and 75.16% at gd 19.5) (Figs 4.1C, S4). Moreover, we observed that the invasive trophoblast cell population had the highest number of cell type-specific peaks of the major cell types analyzed, despite being of less abundance than some other cell types (Fig 4.1D). This result may indicate that the invasive trophoblast cell population had the most gene-associated accessible chromatin among the cell types identified.

4.3.2 Identification of invasive trophoblast cell regulated genes using cell type-specific chromatin accessibility profiles

Following the observation that invasive trophoblast cells had the most cell type-specific peaks, we next checked the number of peaks associated with each gene at each gestation day. At both gestational timepoints, there were many genes associated with at least two open regions (808 and 349 genes at gd 15.5 and 19.5, respectively) (Fig 4.2A). Next, we investigated the differences in expression levels of transcripts linked to 2, 3, 4, or 5 peaks using the average expression level obtained from the scRNA-seq data. In general, we observed an increasing trend of expression level when a transcript is associated with more peaks. Furthermore, we observed that while gd 15.5 expression levels were significantly different as the number of associated peaks increased, at gd 19.5, transcript expression profiles were not significantly different when more peaks were associated with a transcript after a cut-off of 3 (Fig S5). Therefore, we partitioned transcripts into two groups for the downstream analyses: $\geq$3 peaks or <3 peaks. At both gestation days, genes with more than three peaks had significantly higher expression than genes with less than three peaks (p-value=7.029e-09 and 1.374e-08 at gd 15.5 and 19.5, respectively) (Fig 4.2B), suggesting that, in general, genes with $\geq$3 trophoblast-specific peaks are more active within the cell population and could have important functional roles for trophoblast cells. However, there are notable exceptions, including *Prl7b1* (Table S2), which has <3 peaks, but whose expression is specific and among the highest in the invasive trophoblast cell lineage.

In addition, we compared transcripts with $\geq$3 open regions to transcripts with invasive trophoblast cell cluster-specific expression (invasive trophoblast cell marker transcripts), previously determined from the scRNA-seq data (Scott et al., 2022) at each gestation day. At gd 15.5, 57 of the 274 genes with $\geq$3 peaks were also markers of the invasive trophoblast cell cluster (p-value=5.29e-08), and at gd 19.5, 39 of the 103 genes with $\geq$3 peaks were markers of the invasive trophoblast cell cluster (p-value=6.79e-06). These markers included genes with known trophoblast functions (*Tfap2c* (Auman et al., 2002; Kuckenberg et al., 2012; Werling and Schorle, 2002), *Ets2* (Yamamoto et al., 1998), and *Cited2* (Imakawa et al., 2016; Kuna et al., 2023;

Withington et al., 2006)), and genes known to be prominently expressed in invasive trophoblast cells (*Prl5a1* (Ain et al., 2003)) (Fig 4.2C). Of note, while some of these markers (*Cited2* and *Ets2*) have similar activities around their promoter regions in all cell types, they had multiple associating peaks specific to the invasive trophoblast cell cluster.

To determine if transcripts that have multiple associated peaks in rat invasive trophoblast cells also possess multiple associated peaks in human EVT cells, we incorporated open regions (ATAC-seq peaks) identified in EVT cells into our analysis (Varberg et al., 2022). First, we associated the EVT cell open regions to genes. Then, we compared the number of EVT cell peaks associated with genes that have either $\geq$3 or <3 peaks in rat invasive trophoblast cells (Table S2). We observed that, at both time points, genes with $\geq$3 peaks in rat invasive trophoblast cells had significantly more peaks in human EVT cells than genes that had fewer than 3 peaks in rat invasive trophoblast cells (p-value<2.2e-16) (Fig 4.2D).

4.3.3 Identification of transcription factors (TFs) enriched in invasive trophoblast cell-specific peaks

We first defined multiple peak sets that TFs could bind: invasive trophoblast cell-specific peaks identified at both gestation days (named "common peaks", consisting of 1242 peaks) (Table S3), peaks differentially accessible at gd 15.5 (named "gd 15.5-specific peaks", 51 peaks), and peaks differentially accessible at gd 19.5 (named "gd 19.5-specific peaks", 194 peaks) (Table S4). Using gene ontology (GO) analysis, we observed that only common peaks were enriched for relevant biological functions such as "cell-cell adhesion" (FDR=1.16e-04), "positive regulation of cell migration" (FDR=0.012), and "female pregnancy" (FDR=0.017) (Fig. 4.3A, Table S3), while differentially accessible peaks at both gd were not enriched for any processes. This result suggests that in the invasive trophoblast cell population, regulatory elements are consistently accessible at gd 15.5 and 19.5 to regulate biologically important genes, using both stages to select common peaks may increase confidence, and that stage-specific peaks may be noise. We therefore proceeded with motif enrichment analysis in the common peaks.

Following the enrichment tests, filtering, and TF family grouping, we identified 11 TF families that were enriched in the common peaks, some of which have known roles in regulating trophoblast biology (Fig 4.3A, Table S3). For example, TFAP2C motifs were enriched with the highest fold change in the common peaks. TFAP2C is a member of the *AP-2* TF family and is a known regulator of the trophoblast cell lineage in both mouse and human (Kaiser et al., 2015; Kuckenberg et al., 2012; Soares et al., 2018). We further confirmed the enrichment of the TFAP2C binding sites by comparing the rat open regions with TFAP2C motifs to TFAP2C chromatin immunoprecipitation (ChIP)-seq peaks from differentiated mouse TS cells (Lee et al., 2019). We found that of the 439 rat peaks with TFAP2C motifs, 208 (47.38%) overlapped with TFAP2C peaks in differentiated mouse TS cells, which was significant (p-value=0.009). Additionally, all 11 TF families enriched in the rat peaks were enriched in EVT cell ATAC-seq peaks (Varberg et al., 2022) (Table S3). These comparisons provide evidence for the validity of the computationally based binding site predictions.

To determine if TF functions could be predicted using the binding sites, we carried out functional enrichment analysis on the genes associated with peaks where the TF families' binding sites were found. We observed four families with at least one term enriched (Table S3), two of which were enriched for important invasive trophoblast functions: "NR2F6, Pparg::Rxra" (*Thyroid hormone receptor-related factors – RXR-related receptors* family, *Nuclear receptors with C4 zinc fingers* class) enriched for "positive regulation of cell migration" and "vasculature development"; and "TFAP2C" (*AP-2* family, *Basic helix-span-helix factors* class) enriched for "cell-cell adhesion", "positive regulation of cell motility", and "vasculature development". Many of these observed terms agree with previous findings about roles of the families in trophoblast cell functions (Auman et al., 2002; Barak et al., 1999; Werling and Schorle, 2002).

Next, we investigated which TF families were associated with the same target genes. We observed multiple pairs of TF families that shared a significant number of overlapping target genes, such as: "TCF4" (*E2A-related factors* family, *Basic helix-loop-helix factors* class) and "SNAI1" (*More than 3 adjacent zinc finger factors* family, *C2H2 zinc finger factors* class)

(adjusted p-value=3.64e-55); "JUNB, FOSL2::JUN" (*FOS-related factors – JUN-related factors* family, *Basic leucine zipper factors* class) and "CREB3, Creb5" (*CREB-related factors* family, *Basic leucine zipper factors* class) (adjusted p-value=9.35e-41); and "TFAP2C" (*AP-2* family, *Basic helix-span-helix factors* class) and "TCF4" (*E2A-related factors* family, *Basic helix-loop-helix factors* class) (adjusted p-value=7.88e-06) (Fig 4.3B, blue scale). Overall, this analysis highlights TF families that share common target genes. We also checked if TF family pairs occurred in the same peaks more than expected by chance. We found six pairs of TF families significantly over-represented together, including: "TCF4" (*E2A-related factors*) and "SNAI1" (*More than 3 adjacent zinc finger factors* family) (adjusted p-value=3.09e-58), "CREB3, Creb5" (*CREB-related factors* family) and "JUNB, FOSL2::JUN" (*FOS-related factors – JUN-related factors* family) (adjusted p-value=1.36e-45), and "TFAP2C" (*AP-2* family) and "TCF4" (*E2A-related factors* family) (adjusted p-value=4.67e-04) (Fig 4.3B, dark red scale). Each TF family that is part of the over-represented pairs has been individually connected to the regulation of trophoblast cell function. For example, TCF4 and SNAI1 are regulators of trophoblast cell differentiation and motility (Meinhardt et al., 2014) and trophoblast invasion (E. Davies et al., 2016), respectively. Moreover, most of the peaks were bound by at least two TF families (Fig 4.3C). This analysis suggested that TF families can bind in the same locations to interact and regulate cell type-specific functions. TFs can also bind individually to act in their regulatory roles.

4.3.4 Identification of conserved, invasive trophoblast cell-specific regulatory regions using network analysis

We aimed to create a network that would provide information on TFs that activate gene expression through conserved cis-regulatory regions. We first determined that 264 of the common peaks in rat invasive trophoblast cells were conserved in human EVT based on overlap with EVT ATAC-seq data (Table S5). Of note, a significantly higher proportion of common peaks was conserved compared to gd 19.5-specific peaks (p-value=1.039e-05), and only 11 gd 15.5-specific peaks were conserved, further justifying the use of common peaks in the network analysis. Next,

because of the moderate correlation between ATAC-seq and RNA-seq signal (see Section 1), we determined if conserved common peaks also had signal for H3K27ac, a histone modification associated with active enhancers, by using ChIP-Seq data from Varberg et al. (Varberg et al., 2022). Indeed, we found there was a significant overlap (p-value¡2.2e-16) of conserved common peaks and EVT H3K27ac peaks (Table S5), indicating that the conserved common peaks are likely to contribute to activating associated gene expression. Finally, we established the network by compiling several datasets: i) conserved common peaks, ii) motifs enriched within these regions, and iii) conserved genes that exhibited invasive trophoblast cell-specific expression, according to the scRNA-seq analysis (Scott et al., 2022), at both gd 15.5 and 19.5. The resulting network had 11 source nodes, corresponding to 11 TF families, and 34 target genes (Fig 4.4A, Table S5).

In this network, there are multiple genes with high in-degree centrality ($\geq$5), meaning the genes were associated with invasive trophoblast cell-specific peaks predicted to be bound by TFs connected to $\geq$5 TF families. These genes were Plk2 (linked with five TFs), Scap (linked with five TFs), AABR07027306.1 (PHACTR1 human ortholog, linked with six TFs), Pcdh12 (linked with six TFs), Galnt6 (linked with six TFs), and Col4a1 (linked with seven TFs) (Fig 4.4A). Pcdh12, Plk2, Scap, and Col4a1 have previously been linked to the regulation of embryonic and placental development (Bouillot et al., 2006, 2011; Ma et al., 2003; Matsuda et al., 2001; Oefner et al., 2015; Okae et al., 2018; Schenke-Layland et al., 2007). Although, Phactr1 and Galnt6 have not been directly implicated in trophoblast cell biology, they have been shown to regulate migration and invasion of cancer cells (Gao and Zheng, 2022; Herman et al., 2021; Song et al., 2020). Further analysis of the involvement of these genes in the regulation of the invasive trophoblast cell lineage is merited. Regulatory elements and the enriched motifs associated with these genes as well as all other target genes in the network can be found in Table S4.

Moreover, other target genes in the network and their distal elements could also be important for regulating invasive trophoblast cell functions. For example, *Cited2*, a gene required for trophoblast cell differentiation, placental development, and regulation of invasive trophoblast/EVT cells (Imakawa et al., 2016; Kuna et al., 2023; Moreau et al., 2014; Withington

et al., 2006), was predicted to be regulated by a distal peak where TFAP2C and STAT3 motifs were found (Fig 4.4A and B). This peak (chr1:12808761-12809434) also overlapped with a TFAP2C ChIP-seq peak from differentiated mouse TS cells (Lee et al., 2019) (Fig 4.4C), suggesting that it may be bound *in vivo*.

Together, the target genes, regulatory elements, and TFs we identified will be candidates for future experiments to interrogate gene regulatory networks controlling invasive trophoblast cells.

4.4 Discussion

The invasive trophoblast cell lineage is an evolutionary adaptation facilitating viviparity in mammals possessing hemochorial placentation (Pijnenborg et al., 1981). Invasive trophoblast cells acquire migratory behavior, penetrate the uterine parenchyma, and serve a transformative role on cellular constituents ensuring a successful pregnancy outcome (Knöfler et al., 2019; Soares et al., 2018; Turco and Moffett, 2019). The root cause of many obstetric complications is predicted to be a failure in invasive trophoblast cell-guided uterine transformation (Brosens et al., 2011, 2019). Surprisingly, existing knowledge of gene regulatory networks controlling development and function of the invasive trophoblast cell lineage is modest. In this report, we sought to provide new insights into the regulation of the invasive trophoblast cell lineage. Our efforts focused on the rat, a species possessing deep intrauterine trophoblast cell invasion with similarities to human placentation and amenable to testing hypotheses pertaining to the invasive trophoblast cell lineage *in vivo* (Shukla and Soares, 2022; Soares et al., 2012). In this report, we integrated snATAC-seq and scRNA-seq (Scott et al., 2022) datasets from the rat uterine-placental interface with the goal of gaining insight into gene regulatory networks controlling the invasive trophoblast cell lineage. Chromatin accessibility profiles for each of the cellular constituents of the uterine-placental interface were determined. An in-depth analysis of invasive trophoblast cells led to the identification of invasive trophoblast cell specific genes, TFs, and TF target genes. A correlation was established between the presence of invasive trophoblast cell-specific open chromatin and gene expression. Using DNA motif binding enrichment and network analysis, we

predicted TF pairs and cis-regulatory elements linked to invasive trophoblast cell genes. The efforts led to the recognition of conservation between rat and human invasive trophoblast cell lineages and predictions of distal regulatory elements within the invasive trophoblast cell lineage.

Our approach of relating open chromatin to gene expression profiles is not perfect. Gene regulatory regions can regulate multiple genes (McLean et al., 2010) and can be located considerable distances from the gene they regulate (Lin et al., 2022). We observed that most open chromatin regions were distal to genes. Moreover, the open chromatin-gene association rule we used, together with the stringent requirement for conserved regulatory regions and genes, contributed to the inference of a relatively small and manageable network of TFs and target genes. This contributed to a straightforward network analysis that enabled the prediction of relevant interactions. Other computational methods such as co-accessibility analysis, which employs chromatin accessibility profiles to predict interactions of cis-elements (Pliner et al., 2018), represents a complementary approach. Although our network construction method involved using only conserved open regions and conserved target genes, this does not negate the merits of investigating TFs and target genes inferred with species-specific elements.

Candidate TFs driving gene regulation in invasive trophoblast cells were identified through their expression in invasive trophoblast cells and through the presence of corresponding TF DNA binding motifs associated with invasive trophoblast cell specific genes. The most striking TF families linked to the invasive trophoblast cell lineage exhibit conservation in human EVT cells (Varberg et al., 2022) and have been previously implicated in trophoblast cell biology (Hemberger et al., 2020; Rossant, 2001). Most interestingly, many of the invasive trophoblast cell relevant TFs are implicated in early phases of trophoblast cell lineage development or the differentiation of other trophoblast cell lineages. For example, mouse mutagenesis has demonstrated indispensable roles for Tfap2c, Cdx2, Ets2, and Pparg in trophoblast cells and placentation that precede the appearance of the invasive trophoblast cell lineage (Auman et al., 2002; Barak et al., 1999; Chawengsaksophak et al., 1997, 2004; Werling and Schorle, 2002; Yamamoto et al., 1998). Some of these TFs were predicted to regulate the same genes based on the motif enrichment analysis,

and all of these TFs had a high degree of connectivity with each other in the network we present. Previous studies have determined that TFs can work in combination to regulate trophoblast cell lineages, but different TF partnerships are implicated in the regulation of distinct processes (Hemberger et al., 2020; Latos et al., 2015; Latos and Hemberger, 2016). Re-use of trophoblast lineage associated TFs in the regulation of invasive trophoblast cells is intriguing but creates experimental challenges. Future *in vivo* investigation will necessitate the establishment of conditional mutagenesis rat models specific to the invasive trophoblast cell lineage. Such efforts will be facilitated by the integration of single-nucleus chromatin accessibility and single-cell gene expression profiles reported here. Unique TF combinations at gene regulatory domains and/or the recruitment of unique sets of co-regulators may prove crucial to invasive trophoblast cell biology.

The uterine-placental tissue used in generating the snATAC-seq and scRNA-seq contains invasive trophoblast cells that have exited the placenta and entered the uterus and thus represent a differentiated cell type. We did not observe any evidence for multiple types of differentiated invasive trophoblast cell types nor did we detect evidence for invasive trophoblast cell progenitors. This latter population of progenitor cells should reside in the junctional zone of the rat placenta or the EVT cell column of the human placenta. Thus, the present analysis is biased towards characterization of a mature invasive trophoblast cell population. Consequently, the invasive trophoblast cell gene signature, including TFs, may best represent requirements for maintenance of the invasive trophoblast cell state. Comparisons of these rat invasive trophoblast cell chromatin and gene expression profiles with human EVT cell populations isolated from first trimester tissues (Liu et al., 2018; Marsh et al., 2022; Scott et al., 2022; Sun et al., 2020; Suryawanshi et al., 2018; Varberg et al., 2022; Vento-Tormo et al., 2018) or derived from human TS cells (Okae et al., 2018; Varberg et al., 2022) have some inherent limitations. Elucidation of single cell multi-omic profiles for the junctional zone will provide valuable information regarding derivation of the invasive trophoblast cell lineage and further insights into conservation of this important developmental process.

The datasets and analyses presented in this report represent a framework for constructing hypotheses relevant to establishing a gene regulatory network controlling the invasive trophoblast cell lineage. A research approach can now proceed involving identification of candidate conserved regulatory pathways, evaluating the importance of the regulators using TS cell models, and testing critical hubs within the pathways using relevant *in vivo* rat models.

4.5 Materials and methods

4.5.1 Animals

Holtzman rats were originally purchased from Envigo. Rats were maintained on a 14 h light/10 h dark cycle with open access to food and water. Timed pregnancies were obtained by mating adult males (>10 weeks of age) and adult females (8-12 weeks of age). Pregnancies were confirmed the next morning by presence of sperm in a saline vaginal lavage and defined as gd 0.5. Protocols for research with animals were approved by the University of Kansas Medical Center (KUMC) Animal Care and Use Committee.

4.5.2 Cell isolation from tissue

Uterine-placental interface tissue (also called metrial glands) were dissected from gd 15.5 (n=3 pregnancies) and 19.5 rat placentation sites (n=3 pregnancies) as previously described (Ain et al., 2006; Scott et al., 2022) and put in ice cold Hank's balanced salt solution (HBSS). Tissues were minced into fine pieces with a razor blade and digested in Dispase II (1.25 units/mL, D4693, Sigma-Aldrich), 0.4 mg/mL collagenase IV (C5138, Sigma-Aldrich), and DNase I (80 units/mL, D4513, Sigma-Aldrich) in HBSS for 30 min. Red blood cells were lysed using ACK lysis buffer (A10492-01, Thermo-Fisher), rotating at room temperature for 5 min. Samples were washed with HBSS supplemented with 2% fetal bovine serum (FBS, Thermo-Fisher), and DNase1 (Sigma-Aldrich) and passed through a 100 µm cell strainer (100ICS, Midwest Scientific). Following enzymatic digestion, cell debris was removed using MACS Debris Removal Solution

(130-109-398, Miltenyi Biotec). Cells were then filtered through a 40 µm cell strainer (40ICS, Midwest Scientific) and cell viability was assessed, which ranged from 90 to 93%.

4.5.3 Nuclei isolation, library preparation, and sequencing

Cells were isolated from gd 15.5 and 19.5 uterine-placental interface tissue as described above, and nuclei were isolated from the cell suspension according to the 10X Genomics Nuclei Isolation protocol. Briefly, cells were washed with HBSS supplemented with 2% FBS (Thermo-Fisher) and cell number determined. Approximately 500,000 cells were centrifuged, and 100 µL 10X Genomics Nuclei Isolation Lysis Buffer was added. The suspension was incubated for 3 min, then 10X Genomics Nuclei Isolation Wash Buffer was added. Cells were passed through a 40 µm cell strainer and centrifuged. Cells were resuspended in 50 µL chilled 10X Genomics Nuclei Isolation Buffer. Single nuclei were captured using the Chromium Controller into 10X barcoded gel beads. Libraries were generated using Chromium Next GEM Single Cell ATAC Library & Gel Bead Kit v1.1 (10X Genomics) and sequenced in a NovaSeq6000 sequencer at the KUMC Genome Sequencing Core.

4.5.4 snATAC-seq preprocessing

Read alignment to the rat genome (Rnor 6.0, Ensembl 98 (Cunningham et al., 2019)), primary peak calling, and feature quantification were performed using Cell Ranger Software (version 4.0.0). Quality control steps and downstream analyses were performed using the R package *Signac* (version 1.1.1) (Stuart et al., 2019). Unless otherwise reported, default parameters were used. We identified accessible regions using the CallPeaks() function in *Signac*, which utilizes model-based analysis for ChIP-seq (MACS) (Zhang et al., 2008). Parameters used for the analyses were nuclei with a total number of fragments in peaks ranging from 1000 to 20000, percentage of reads in peaks >15%, and enrichment ratio at transcription start sites >1.5 (Fig S1). We normalized across samples and across peaks using term frequency-inverse document frequency, which is implemented through RunTFIDF() in Seurat. We used method =3, which

computes log(term frequency) × log(IDF), due to great sparsity in the feature matrix and strong count outliers (Fig S2). All features are retained to perform dimension reduction with singular value decomposition (SVD). Normalization with term frequency-inverse document frequency followed by SVD is also known as latent semantic indexing (LSI) (Cusanovich et al., 2015). We also investigated the correlations between sequencing depth and LSI components (using the DepthCor() function) as well as ranked the LSI components using the percentage of variance (using the ElbowPlot() function). As a result, we kept LSI components 2 to 20 for gd 15.5 replicates, and LSI components 2 to 10 for gd 19.5 replicates (Fig S2). Replicates for each time point were then merged using the Merge() function in Seurat.

4.5.5 snATAC-seq clustering

To identify cell clusters for each time point, we utilized K-nearest neighbor (KNN) graphs with retained significant LSI components and the smart local moving algorithm (Waltman and Van Eck, 2013), which was implemented through the Seurat functions FindNeighbors() and FindClusters(). The clusters were then visualized with uniform manifold approximation and projection (UMAP).

4.5.6 scRNA-seq and snATAC-seq integration – label transferring

To transfer cluster labels from our corresponding scRNA-seq data, we used the FindTransferAnchors() and TransferData() functions in the Seurat package (version 4.1.0) (Butler et al., 2018). Briefly, this process uses canonical correlation analysis for initial dimension reduction, then identifies cell neighborhoods with KNNs, and mutual nearest neighbors (MNN). The correspondences between cells were referred to as “anchors”. Next, the anchors were given scores and weights to eliminate incorrect correspondences and to define the association strengths between cells and anchors. Finally, anchor classification and anchor weights were used to transfer labels from scRNA-seq to snATAC-seq data. To check the correlation between snATAC-seq and scRNA-seq profiles in each cell population, we first estimated the chromatin accessibility profiles

around transcription start sites using the *Signac* function GeneActivity(). Then Spearman correlation and its statistical significance were calculated using the R function cor.test() (*stats* package version 4.0.2 (R Core Development Team, 2013)).

4.5.7 Analysis of cell population-specific peaks

The FindAllMarkers() function was used with cell identities transferred from scRNA-seq data and the fragment counts in peaks, to compare chromatin accessibility profiles between cell types for each gd. We used a logistic regression framework with a latent variable of the total number of fragments in peaks to account for the difference in sequencing depths. A peak is considered more accessible in a cell population (and hence specific) if it has an adjusted p-value $\leq$0.05 and an average $\log_2$(fold change) $\geq \log_2(1.5)$. Rat peaks were associated with the nearest gene (according to the start position) on the same chromosomes using the *Signac* function ClosestFeature() with the underlying genome annotation from Ensembl 98 (Cunningham et al., 2019). This association rule was also used when the distance distribution of peaks to transcription start sites was calculated with the R package ChIPseeker (Yu et al., 2015). ATAC-seq peaks in EVT cells (Varberg et al., 2022) were associated to the single nearest genes with the maximum distance of 1000 kb around the TSS using GREAT (Genomic Regions Enrichment of Annotations Tool) (McLean et al., 2010). Rat genes were mapped to their one-to-one human orthologs using gene mapping from Ensembl 98. To assess changes in the expression level of transcripts with different numbers of associated peaks, or differences in the numbers of EVT peaks between two gene groups, we used Wilcoxon rank sum test, implemented with the R function wilcox.test() (*stats* package version 4.0.2 (R Core Development Team, 2013)). To test the significance of overlap between genes with $\geq$3 peaks and invasive trophoblast cell markers, we used hypergeometric tests with the R function phyper() (*stats* package version 4.0.2 (R Core Development Team, 2013)) using options lower.tail = TRUE. In all tests, the significance level used was 0.05.

4.5.8 Common peaks, peak mapping across species, and conserved common peaks

Common invasive trophoblast cell-specific peaks between the two gd were obtained using BEDTools intersect (version 2.27.1) (Quinlan and Hall, 2010). Regions between the two gd were considered common if ≥50% of the base pairs overlapped. To compare peaks across species (rat, mouse and human), all peak sets were converted to human coordinates (hg38) using LiftOver (default settings) (Hinrichs et al., 2006). Bedtools intersect (version 2.27.1) (Quinlan and Hall, 2010) was used to identify conserved peaks, which were defined as peaks that overlapped with ATAC-seq peaks in EVT cells (Varberg et al., 2022) by ≥1 base pair (bp).

4.5.9 Motif analysis with common peaks

To identify enriched motifs in common peaks, we used the Homo sapiens, Mus musculus and Rattus norvegicus motif databases from JASPAR (version 2020) (Fornes et al., 2020). A BSgenome object for Rattus norvegicus, necessary to add motif information to Seurat objects, was built using the BSgenome R package (version 1.58.0) (Hervé, 2020) and genome sequences obtained from Ensembl 98 (Cunningham et al., 2019). We used the gd 19.5 coordinates of the common peak sets as input, then generated a set of 50,000 background sequences with matched length and GC content distribution using the Seurat function MatchRegionStats(). For each motif, we calculated a fold change as the percentage the motif is observed in the input sequences divided by the percentage it is observed in the background. A motif is considered enriched if its hypergeometric adjusted p-value is ≤0.05 and fold change ≥1.5. The p-values were adjusted with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

To identify motif groups, we first mapped enriched motifs for all three organisms to their corresponding TFs using TF – motif mapping information from the JASPAR database, then retained only TFs with expression level ≥0.5 at both gd using the scRNA-seq data. Next, we grouped TFs according to their protein families, also obtained from the JASPAR database. To compare the observed binding sites of the protein TFAP2C with previously published data from Lee et al. (Lee et al., 2019), we accessed the TFAP2C ChIP-seq data generated from

differentiated TS cells through the GEO ID GSM3019344. A rat peak with TFAP2C motifs was defined to agree with mouse TFAP2C ChIP-seq peaks if they overlapped by ≥1 bp as assessed with BEDTools intersect (version 2.27.1) (Quinlan and Hall, 2010). The significance of the overlapping was determined using Fisher's exact test, with the option alternative = "greater" and a significance level of 0.05. To carry out functional enrichment of target genes of the enriched TF families, we used Webgestalt (version 2019) (Liao et al., 2019) with the rat genome. A term was considered enriched if its FDR <0.05, enrichment rate ≥2, and number of observed genes is ≥5. To test for over-representation of shared genes and shared binding locations, we used hypergeometric tests with the R function phyper() (*stats* package version 4.0.2 (R Core Development Team, 2013)) using options lower.tail = TRUE. Correction for multiple testing was carried out using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Significance level was set at 0.05.

4.5.10 Network inferences and analyses with conserved common peaks

In our networks, an edge between a TF family and a gene means the gene is the nearest one to conserved common peaks with the enriched motifs of the family. Source nodes in the network were TF families named with representative motifs. Target genes were marker genes of the invasive trophoblast cell clusters at both gd and were conserved in EVT cells according to the scRNA-seq data (Scott et al., 2022). The network was visualized and analyzed with Cytoscape (Shannon et al., 2003).

4.6 Main figures

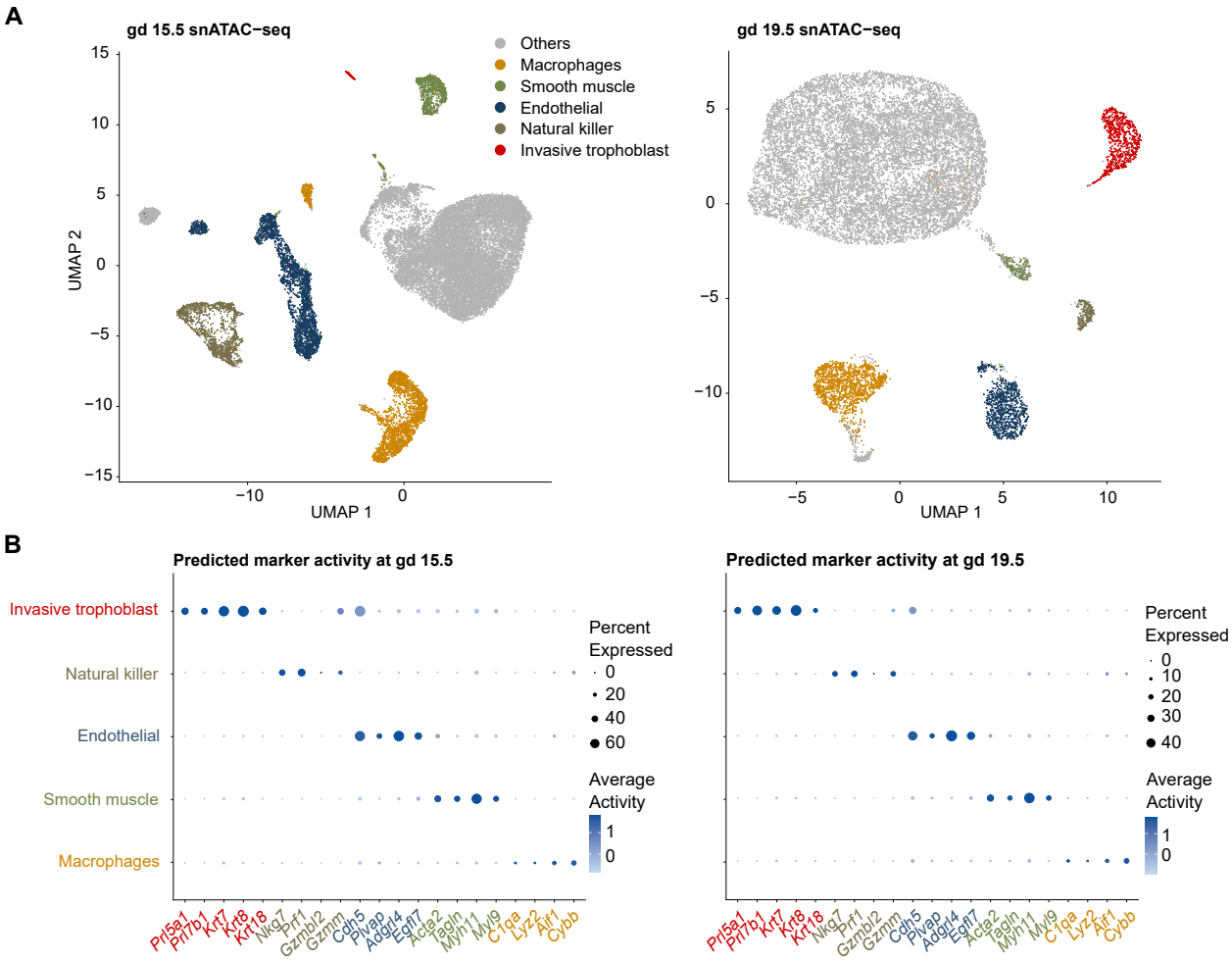

Figure 4.1: Chromatin accessibility profiles of cell populations at the uterine-placental interface. (A) UMAPs of snATAC-seq profiles of cell populations at gestation day (gd) 15.5 and 19.5 showing cell identities obtained by transferring labels from scRNA-seq data. (B) Dot plots showing known markers of cell types generally have higher accessibility within 2,000 base pairs (bp) of the transcription start sites (TSS) and in higher percent of nuclei than in other cell populations. Dot sizes correspond to the percent of nuclei in each cell population that were open around the TSS; colors correspond to the levels of predicted gene activity.

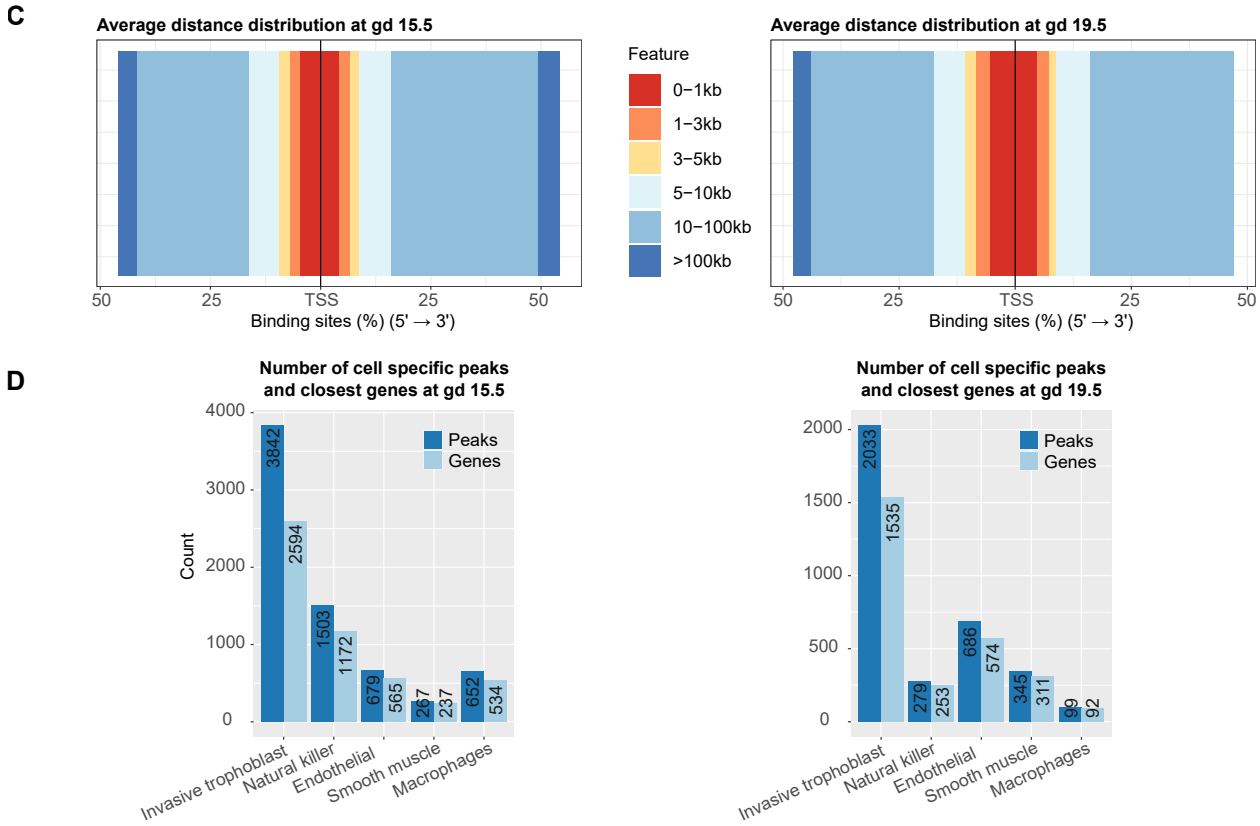

Figure 4.1: (Continued)

(C) Stack bar plots showing that cell type-specific open chromatin peaks were most often distal to the TSS. For distribution of distances for each individual cell type see Fig S4. (D) Bar plots showing the number of open chromatin peaks specific to a cell population, and the number of nearest genes to cell-specific open chromatin peaks. Cell specific open chromatin peaks are open chromatin peaks differentially accessible in the cell population compared to all other cell populations (adjusted p-value ≤ 0.05, average $\log_2$(fold change) $\geq \log_2(1.5)$.

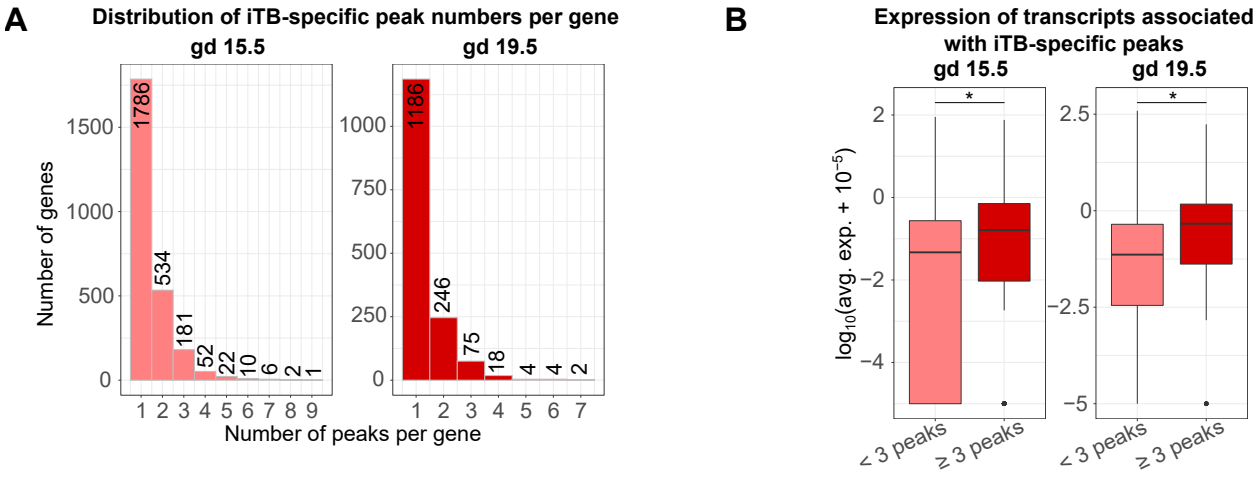

Figure 4.2: Analysis of chromatin accessibility profiles can identify regulatory regions for genes defining the invasive trophoblast cell population.

(A) Histograms of the number of invasive trophoblast-specific (iTB-specific) peaks per gene showing that many genes had $\geq$1 peaks. The x-axis shows the number of peaks per gene, and the y-axis shows the number of genes. (B) Boxplots of transcript expression associated with iTB-specific peaks showing that genes with $\geq$3 peaks had significantly higher expression than genes with fewer than 3 peaks. Expression was plotted in a $\log_{10}(\text{average expression} + 10^{-5})$ scale.

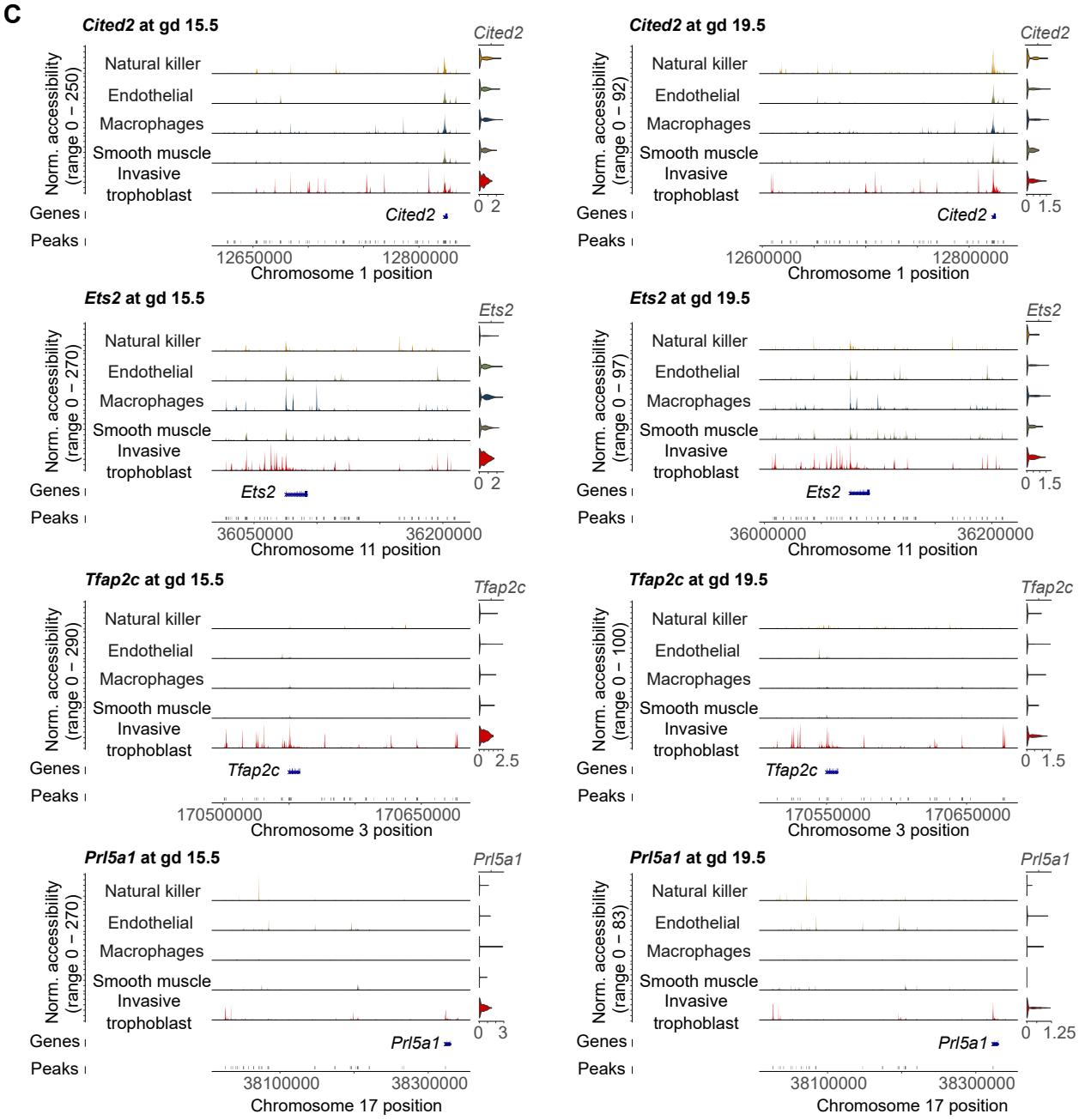

Figure 4.2: (Continued)

(C) Examples of iTB-specific genes associated with ≥3 peaks at both gd 15.5 and 19.5. For each subplot, the first section was composed of five tracks of normalized accessibility, corresponding to five cell types. The right-most column of the first section showed the predicted gene activity using chromatin accessibility within 2,000 bp of the TSS. The second and third section include two tracks corresponding to gene location and open chromatin peak locations, respectively.

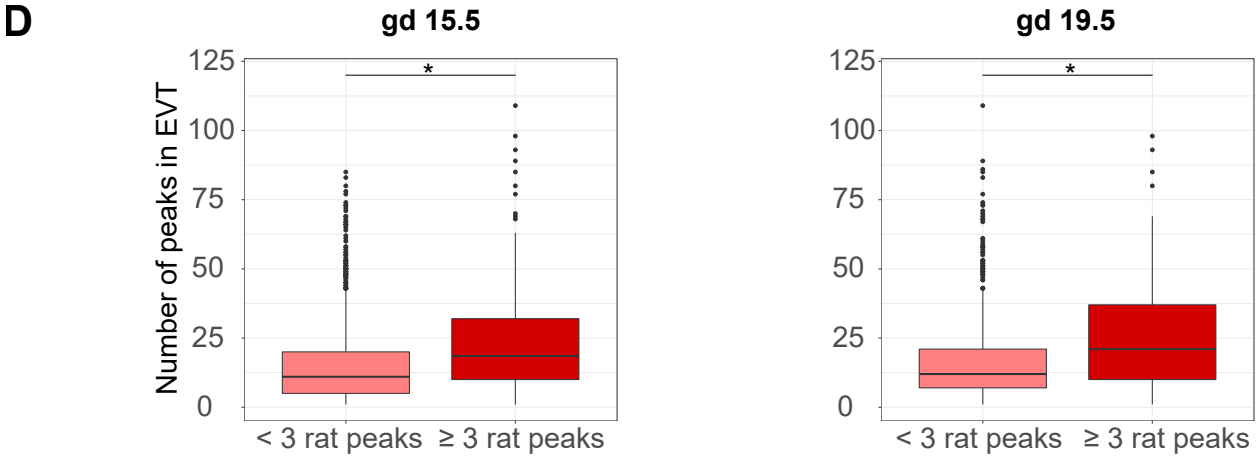

Figure 4.2: (Continued)

(D) Boxplots of the number of conserved ATAC-seq peaks in EVT cells and rat invasive trophoblast cells. Rat genes with ≥3 invasive trophoblast cell-specific peaks had significantly more EVT cell ATAC-seq peaks than rat genes with <3 invasive trophoblast cell-specific peaks. ATAC-seq peaks in EVT cells were obtained from Varberg et al. (Varberg et al., 2022). Statistical analyses were performed using Wilcoxon rank sum tests at a significance level of 0.05.

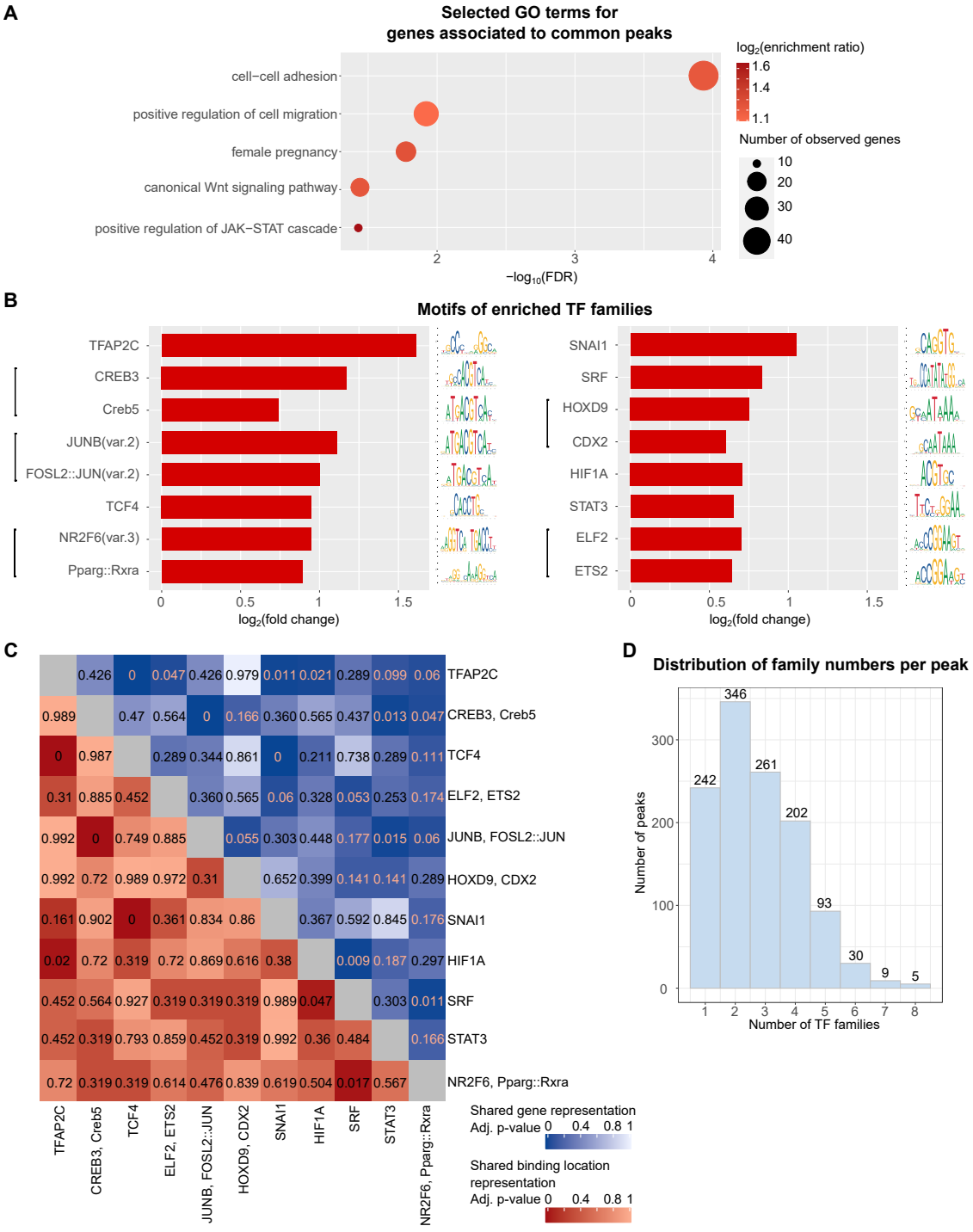

Figure 4.3: Motif analysis identified transcription factor (TF) combinations regulating invasive trophoblast cell functions.

Figure 4.3: (Continued)

(A) Representative motifs for enriched TF families found in common open chromatin peaks. Motifs for the top two most highly expressed TFs in each family are shown. In case multiple motifs are enriched that correspond to the same TF, motifs with the highest fold change are shown. See the mapping of motifs to TFs in Table S3. A motif is considered enriched if its hypergeometric adjusted p-value is $\leq$0.05 and fold change $\geq$1.5. The p-values were adjusted with the Benjamini-Hochberg procedure. Only motifs corresponding to genes with expression level $\geq$0.5 at both gd 15.5 and gd 19.5 were used in the downstream analysis. (B) Representative motifs for enriched TF families found in common open chromatin peaks. Motifs for the top two most highly expressed TFs in each family are shown. If multiple motifs are enriched that correspond to the same TF, then motifs with the highest fold change are shown. See the mapping of motifs to TFs in Table S3. A motif is considered enriched if its hypergeometric adjusted P $\leq$ 0.05 and it fold change $\geq$ 1.5. The P-values were adjusted with the Benjamini-Hochberg procedure. Only motifs corresponding to genes with expression levels of at least 0.5 at both gd 15.5 and gd 19.5 were used in the downstream analysis. (C) Heatmap of hypergeometric adjusted (adj.) P-values showing that some TF family pairs share a significant number of target genes and binding locations. The P-values were adjusted with the Benjamini-Hochberg procedure. Representative motif names (as in B) were used for TF family names. Significance level was 0.05. Blue scale, adj. P-values when testing for significance of shared genes; dark red scale, adj. P-values when testing for significance of shared binding locations. (D) Histogram for number of TF families per common open chromatin peak showing that most open chromatin peaks had at least two TF families predicted to be bound, while there were some common open chromatin peaks with only one TF family predicted to be bound. The x-axis shows the number of TF families per peak; the y-axis shows the number of peaks.

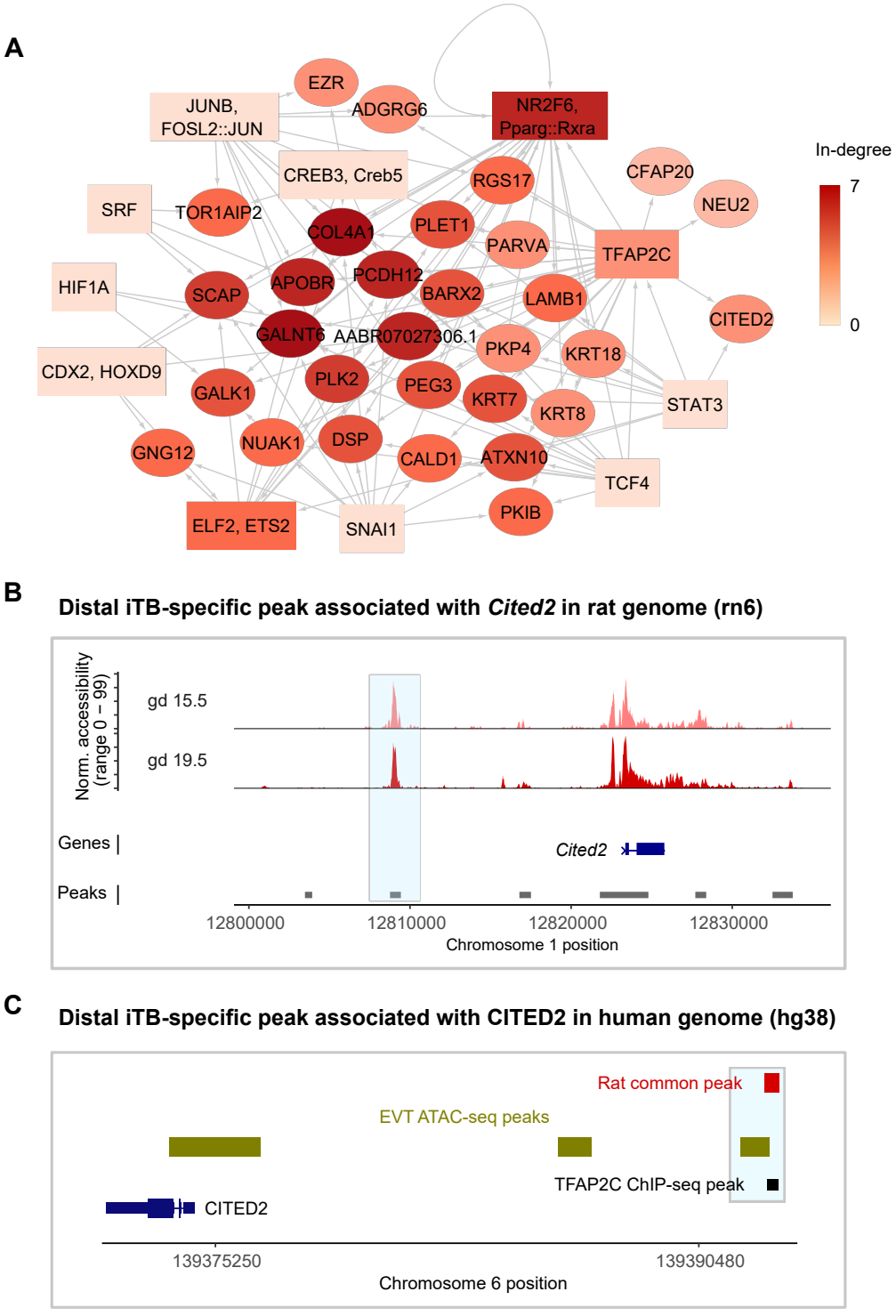

Figure 4.4: Network analysis predicted candidate genes and their distal regulatory elements that govern invasive trophoblast cell functions.

Figure 4.4: (Continued)

(A) Analysis of a network of TF families and target genes highlighting candidate genes and their distal regulatory elements underlying invasive trophoblast cell functions. Rectangular nodes: TF families with representative motif names (as in Fig 4.3A). Round nodes: target genes. Color: the darker the color, the higher the node in-degree centrality. Directed edges mean peaks with the predicted TF families were associated to the target genes. (B) Chromatin accessibility tracks of a candidate invasive trophoblast (iTB) cell-specific distal element associated with the *Cited2* gene in the rat genome (rn6). A region of interest was highlighted in light blue. (C) Locations of the candidate region, ATAC-seq peaks in EVT cells and TFAP2C ChIP-seq peaks in the human genome (hg38). A region of interest was highlighted in light blue.

4.7 References

Ain, R., Canham, L. N., and Soares, M. J. (2003). Gestation stage-dependent intrauterine trophoblast cell invasion in the rat and mouse: Novel endocrine phenotype and regulation. *Developmental Biology*, 260(1):176–190.

Ain, R., Konno, T., Canham, L. N., and Soares, M. J. (2006). Phenotypic analysis of the rat placenta. *Methods in molecular medicine*, 121:295–313.

Auman, H. J., Nottoli, T., Lakiza, O., Winger, Q., Donaldson, S., and Williams, T. (2002). Transcription factor AP-2γ is essential in the extra-embryonic lineages for early postimplantation development. *Development*, 129(11):2733–2747.

Barak, Y., Nelson, M. C., Ong, E. S., Jones, Y. Z., Ruiz-Lozano, P., Chien, K. R., Koder, A., and Evans, R. M. (1999). PPARγ Is Required for Placental, Cardiac, and Adipose Tissue Development. *Molecular Cell*, 4(4):585–595.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

Bouillot, S., Rampon, C., Tillet, E., and Huber, P. (2006). Tracing the Glycogen Cells with Protocadherin 12 During Mouse Placenta Development. *Placenta*, 27(8):882–888.

Bouillot, S., Tillet, E., Carmona, G., Prandini, M. H., Gauchez, A. S., Hoffmann, P., Alfaidy, N., Cand, F., and Huber, P. (2011). Protocadherin-12 Cleavage Is a Regulated Process Mediated by ADAM10 Protein: EVIDENCE OF SHEDDING UP-REGULATION IN PRE-ECLAMPSIA. *Journal of Biological Chemistry*, 286(17):15195–15204.

Brosens, I., Pijnenborg, R., Vercruysse, L., and Romero, R. (2011). THE "GREAT OBSTETRICAL SYNDROMES" ARE ASSOCIATED WITH DISORDERS OF DEEP PLACENTATION. *American Journal of Obstetrics and Gynecology*, 204(3):193.

Brosens, I., Puttemans, P., and Benagiano, G. (2019). Placental bed research: I. The placental bed: from spiral arteries remodeling to the great obstetrical syndromes. *American Journal of Obstetrics and Gynecology*, 221(5):437–456.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.

Chawengsaksophak, K., De Graaff, W., Rossant, J., Deschamps, J., and Beck, F. (2004). Cdx2 is essential for axial elongation in mouse development. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7641–7645.

Chawengsaksophak, K., James, R., Hammond, V. E., Köntgen, F., and Beck, F. (1997). Homeosis and intestinal tumours in Cdx2 mutant mice. *Nature 1997 386:6620*, 386(6620):84–87.

Chuong, E. B., Rumi, M. A., Soares, M. J., and Baker, J. C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genetics 2013 45:3*, 45(3):325–329.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Marugán, J. C., Maurel, T., McMahon, A. C., Moore, B., Morales, J., Mudge, J. M., Nuhn, M., Ogeh, D., Parker, A., Parton, A., Patricio, M., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sparrow, H., Stapleton, E., Szuba, M., Taylor, K., Threadgold, G., Thormann, A., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Yates, A. D., Zerbino, D. R., and Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, 47(D1):D745–D751.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914.

E. Davies, J., Pollheimer, J., Yong, H. E., Kokkinos, M. I., Kalionis, B., Knöfler, M., and Murthi, P. (2016). Epithelial-mesenchymal transition during extravillous trophoblast differentiation. *Cell Adhesion & Migration*, 10(3):310.

Fornes, O., Castro-Mondragon, J. A., Khan, A., Van Der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., and Mathelier, A. (2020). JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Research*, 48(D1):D87–D92.

Gao, F. and Zheng, G. (2022). RUNX3-Regulated GALNT6 Promotes the Migration and Invasion of Hepatocellular Carcinoma Cells by Mediating O-Glycosylation of MUC1. *Disease Markers*, 2022.

Gardner, R. L. and Beddington, R. S. (1988). Multi-lineage 'stem' cells in the mammalian embryo. *Journal of cell science. Supplement*, 10(SSUPL. 10):11–27.

Hemberger, M., Hanna, C. W., and Dean, W. (2020). Mechanisms of early placental development in mouse and humans. *Nature Reviews Genetics*, 21(1):27–43.

Herman, L., Legois, B., Todeschini, A. L., and Veitia, R. A. (2021). Genomic exploration of the targets of FOXL2 and ESR2 unveils their implication in cell migration, invasion, and adhesion. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 35(4).

Hervé, P. (2020). BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., Sultan-Qurraie, A., Thomas, D. J., Trumbower, H., Weber, R. J., Weirauch, M., Zweig, A. S., Haussler, D., and Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic acids research*, 34(Database issue).

Imakawa, K., Dhakal, P., Kubota, K., Kusama, K., Chakraborty, D., Rumi, M. A., and Soares, M. J. (2016). CITED2 MODULATION OF TROPHOBLAST CELL DIFFERENTIATION: INSIGHTS FROM GLOBAL TRANSCRIPTOME ANALYSIS. *Reproduction (Cambridge, England)*, 151(5):509.

Kaiser, S., Koch, Y., Kühnel, E., Sharma, N., Gellhaus, A., Kuckenberg, P., Schorle, H., and Winterhager, E. (2015). Reduced gene dosage of Tfap2c impairs trophoblast lineage differentiation and alters maternal blood spaces in the mouse placenta. *Biology of Reproduction*, 93(2):31–32.

Knöfler, M., Haider, S., Saleh, L., Pollheimer, J., Gamage, T. K., and James, J. (2019). Human placenta and trophoblast development: key molecular mechanisms and model systems. *Cellular and Molecular Life Sciences*, 76(18):3479.

Kuckenberg, P., Kubaczka, C., and Schorle, H. (2012). The role of transcription factor Tcfap2c/TFAP2C in trophectoderm development. *Reproductive BioMedicine Online*, 25(1):12–20.

Kuna, M., Dhakal, P., Iqbal, K., Dominguez, E. M., Kent, L. N., Muto, M., Moreno-Irusta, A., Kozai, K., Varberg, K. M., Okae, H., Arima, T., Sucov, H. M., and Soares, M. J. (2023). CITED2 is a conserved regulator of the uterine-placental interface. *Proceedings of the National Academy of Sciences of the United States of America*, 120(3):e2213622120.

Kwak, Y. T., Muralimanoharan, S., Gogate, A. A., and Mendelson, C. R. (2019). Human Trophoblast Differentiation Is Associated With Profound Gene Regulatory and Epigenetic Changes. *Endocrinology*, 160(9):2189.

Latos, P. A. and Hemberger, M. (2016). From the stem of the placental tree: Trophoblast stem cells and their progeny. *Development (Cambridge)*, 143(20):3650–3660.

Latos, P. A., Sienerth, A. R., Murray, A., Senner, C. E., Muto, M., Ikawa, M., Oxley, D., Burge, S., Cox, B. J., and Hemberger, M. (2015). Elf5-centered transcription factor hub controls trophoblast stem cell self-renewal and differentiation through stoichiometry sensitive shifts in target gene networks. *Genes and Development*, 29(23):2435–2448.

Lee, B. K., Jang, Y., Kim, M., LeBlanc, L., Rhee, C., Lee, J., Beck, S., Shen, W., and Kim, J. (2019). Super-enhancer-guided mapping of regulatory networks controlling mouse trophoblast stem cells. *Nature Communications*, 10(1).

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, 47(W1):W199–W205.

Lin, X., Liu, Y., Liu, S., Zhu, X., Wu, L., Zhu, Y., Zhao, D., Xu, X., Chemparathy, A., Wang, H., Cao, Y., Nakamura, M., Noordermeer, J. N., La Russa, M., Wong, W. H., Zhao, K., and Qi, L. S. (2022). Nested epistasis enhancer networks for robust genome regulation. *Science (New York, N.Y.)*, 377(6610):1077–1085.

Liu, Y., Fan, X., Wang, R., Lu, X., Dang, Y. L., Wang, H., Lin, H. Y., Zhu, C., Ge, H., Cross, J. C., and Wang, H. (2018). Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns of differentiation in the human placenta. *Cell Research 2018 28:8*, 28(8):819–832.

Ma, S., Charron, J., and Erikson, R. L. (2003). Role of Plk2 (Snk) in Mouse Development and Cell Proliferation. *Molecular and Cellular Biology*, 23(19):6936.

Marsh, B., Zhou, Y., Kapidzic, M., Fisher, S., and Blelloch, R. (2022). Regionally distinct trophoblast regulate barrier function and invasion in the human placenta. *eLife*, 11.

Matsuda, M., Korn, B. S., Hammer, R. E., Moon, Y. A., Komuro, R., Horton, J. D., Goldstein, J. L., Brown, M. S., and Shimomura, I. (2001). SREBP cleavage-activating protein (SCAP) is required for increased lipid synthesis in liver induced by cholesterol deprivation and insulin elevation. *Genes & development*, 15(10):1206–1216.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501.

Meinhardt, G., Haider, S., Haslinger, P., Proestling, K., Fiala, C., Pollheimer, J., and Knöfler, M. (2014). Wnt-Dependent T-Cell Factor-4 Controls Human Etravillous Trophoblast Motility. *Endocrinology*, 155(5):1908–1920.

Merrill, C. B., Montgomery, A. B., Pabon, M. A., Shabalin, A. A., Rodan, A. R., and Rothenfluh, A. (2022). Harnessing changes in open chromatin determined by ATAC-seq to generate insulin-responsive reporter constructs. *BMC Genomics*, 23(1).

Moreau, J. L., Artap, S. T., Shi, H., Chapman, G., Leone, G., Sparrow, D. B., and Dunwoodie, S. L. (2014). Cited2 is required in trophoblasts for correct placental capillary patterning. *Developmental Biology*, 392(1):62–79.

Muto, M., Chakraborty, D., Varberg, K. M., Moreno-Irusta, A., Iqbal, K., Scott, R. L., McNally, R. P., Choudhury, R. H., Aplin, J. D., Okae, H., Arima, T., Matsumoto, S., Ema, M., Mast, A. E., Grundberg, E., and Soares, M. J. (2021). Intersection of regulatory pathways controlling hemostasis and hemochorial placentation. *Proceedings of the National Academy of Sciences of the United States of America*, 118(50).

Nelson, A. C., Mould, A. W., Bikoff, E. K., and Robertson, E. J. (2016). Single-cell RNA-seq reveals cell type-specific transcriptional signatures at the maternal-foetal interface during pregnancy. *Nature Communications*, 7.

Nelson, A. C., Mould, A. W., Bikoff, E. K., and Robertson, E. J. (2017). Mapping the chromatin landscape and Blimp1 transcriptional targets that regulate trophoblast differentiation. *Scientific Reports 2017 7:1*, 7(1):1–15.

Oefner, C. M., Sharkey, A., Gardner, L., Critchley, H., Oyen, M., and Moffett, A. (2015). Collagen type IV at the fetal–maternal interface. *Placenta*, 36(1):59.

Okae, H., Toh, H., Sato, T., Hiura, H., Takahashi, S., Shirane, K., Kabayama, Y., Suyama, M., Sasaki, H., and Arima, T. (2018). Derivation of Human Trophoblast Stem Cells. *Cell Stem Cell*, 22(1):50–63.e6.

Ong, C. T. and Corces, V. G. (2012). Enhancers: emerging roles in cell fate specification. *EMBO Reports*, 13(5):423.

Peñalosa-Ruiz, G., Bright, A. R., Mulder, K. W., and Veenstra, G. J. C. (2019). The interplay of chromatin and transcription factors during cell fate transitions in development and reprogramming. *Biochimica et biophysica acta. Gene regulatory mechanisms*, 1862(9).

Pervolarakis, N., Nguyen, Q. H., Williams, J., Gong, Y., Gutierrez, G., Sun, P., Jhutty, D., Zheng, G. X., Nemec, C. M., Dai, X., Watanabe, K., and Kessenbrock, K. (2020). Integrated Single-Cell Transcriptomics and Chromatin Accessibility Analysis Reveals Regulators of Mammary Epithelial Cell Identity. *Cell reports*, 33(3):108273.

Pijnenborg, R., Robertson, W. B., Brosens, I., and Dixon, G. (1981). Review article: Trophoblast invasion and the establishment of haemochorial placentation in man and laboratory animals. *Placenta*, 2(1):71–91.

Pijnenborg, R., Vercruysse, L., and Brosens, I. (2011). Deep placentation. *Best practice & research. Clinical obstetrics & gynaecology*, 25(3):273–285.

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., and Trapnell, C. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell*, 71(5):858–871.e8.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Development Team (2013). R: A Language and Environment for Statistical Computing.

Red-Horse, K., Zhou, Y., Genbacev, O., Prakobphol, A., Foulk, R., McMaster, M., and Fisher, S. J. (2004). Trophoblast differentiation during embryo implantation and formation of the maternal-fetal interface. *Journal of Clinical Investigation*, 114(6):744.

Roberts, R. M., Green, J. A., and Schulz, L. C. (2016). The Evolution of the Placenta. *Reproduction (Cambridge, England)*, 152(5):R179.

Rossant, J. (2001). Stem cells from the Mammalian blastocyst. *Stem cells (Dayton, Ohio)*, 19(6):477–482.

Rugg-Gunn, P. J., Cox, B. J., Ralston, A., and Rossant, J. (2010). Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24):10783–10790.

Schenke-Layland, K., Angelis, E., Rhodes, K. E., Heydarkhan-Hagvall, S., Mikkola, H. K., and MacLellan, W. R. (2007). Collagen IV induces trophoectoderm differentiation of mouse embryonic stem cells. *Stem cells (Dayton, Ohio)*, 25(6):1529–1538.

Schoenfelder, S., Mifsud, B., Senner, C. E., Todd, C. D., Chrysanthou, S., Darbo, E., Hemberger, M., and Branco, M. R. (2018). Divergent wiring of repressive and active chromatin interactions between mouse embryonic and trophoblast lineages. *Nature Communications 2018 9:1*, 9(1):1–10.

Scott, R. L., Vu, H. T., Jain, A., Iqbal, K., Tuteja, G., and Soares, M. J. (2022). Conservation at the uterine-placental interface. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41):e2210633119.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

Shukla, V. and Soares, M. J. (2022). Modeling Trophoblast Cell-Guided Uterine Spiral Artery Transformation in the Rat. *International Journal of Molecular Sciences*, 23(6).

Soares, M. J., Chakraborty, D., Karim Rumi, M. A., Konno, T., and Renaud, S. J. (2012). Rat placentation: An experimental model for investigating the hemochorial maternal-fetal interface. *Placenta*, 33(4):233.

Soares, M. J., Varberg, K. M., and Iqbal, K. (2018). Hemochorial placentation: development, function, and adaptations. *Biology of Reproduction*, 99(1):196–211.

Song, J., Liu, W., Wang, J., Hao, J., Wang, Y., You, X., Du, X., Zhou, Y., Ben, J., Zhang, X., Ye, M., and Wang, Q. (2020). GALNT6 promotes invasion and metastasis of human lung adenocarcinoma cells through O-glycosylating chaperone protein GRP78. *Cell Death & Disease*, 11(5).

Starks, R. R., Biswas, A., Jain, A., and Tuteja, G. (2019). Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics and Chromatin*, 12(1):1–16.

Starks, R. R., Kaur, H., and Tuteja, G. (2021). Mapping cis-regulatory elements in the midgestation mouse placenta. *Scientific Reports 2021 11:1*, 11(1):1–13.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21.

Sun, T., Gonzalez, T. L., Deng, N., Di Pentino, R., Clark, E. L., Lee, B., Tang, J., Wang, Y., Stripp, B. R., Yao, C., Tseng, H. R., Karumanchi, S. A., Koeppel, A. F., Turner, S. D., Farber, C. R., Rich, S. S., Wang, E. T., Williams, J., and Pisarska, M. D. (2020). Sexually Dimorphic Crosstalk at the Maternal-Fetal Interface. *The Journal of Clinical Endocrinology and Metabolism*, 105(12):e4831.

Suryawanshi, H., Morozov, P., Straus, A., Sahasrabudhe, N., Max, K. E., Garzia, A., Kustagi, M., Tuschl, T., and Williams, Z. (2018). A single-cell survey of the human first-trimester placenta and decidua. *Science Advances*, 4(10).

Turco, M. Y. and Moffett, A. (2019). Development of the human placenta. *Development (Cambridge)*, 146(22).

Tuteja, G., Chung, T., and Bejerano, G. (2016). Changes in the enhancer landscape during early placental development uncover a trophoblast invasion gene-enhancer network. *Placenta*, 37:45–55.

Tuteja, G., Moreira, K. B., Chung, T., Chen, J., Wenger, A. M., and Bejerano, G. (2014). Automated discovery of tissue-targeting enhancers and transcription factors from binding motif and gene function data. *PLoS computational biology*, 10(1).

Varberg, K. M., Dominguez, E. M., Koseva, B., McNally, R. P., Moreno-Irusta, A., Wesley, E. R., Iqbal, K., Cheung, W. A., Okae, H., Arima, T., Lydic, M., Holoch, K., Marsh, C., Soares, M. J., Grundberg, E., and Varberg or Michael J Soares, K. M. (2022). Active remodeling of the chromatin landscape directs extravillous trophoblast cell lineage development. *medRxiv*, page 2022.05.25.22275520.

Varberg, K. M., Iqbal, K., Muto, M., Simon, M. E., Scott, R. L., Kozai, K., Choudhury, R. H., Aplin, J. D., Biswell, R., Gibson, M., Okae, H., Arima, T., Vivian, J. L., Grundberg, E., and Soares, M. J. (2021). ASCL2 reciprocally controls key trophoblast lineage decisions during hemochorial placenta development. *Proceedings of the National Academy of Sciences of the United States of America*, 118(10).

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J. E., Stephenson, E., Polański, K., Goncalves, A., Gardner, L., Holmqvist, S., Henriksson, J., Zou, A., Sharkey, A. M., Millar, B., Innes, B., Wood, L., Wilbrey-Clark, A., Payne, R. P., Ivarsson, M. A., Lisgo, S., Filby, A., Rowitch, D. H., Bulmer, J. N., Wright, G. J., Stubbington, M. J., Haniffa, M., Moffett, A., and Teichmann, S. A. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731):347–353.

Waltman, L. and Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, 86(11):1–14.

Werling, U. and Schorle, H. (2002). Transcription Factor Gene AP-2γ Essential for Early Murine Development. *Molecular and Cellular Biology*, 22(9):3149.

Withington, S. L., Scott, A. N., Saunders, D. N., Lopes Floro, K., Preis, J. I., Michalicek, J., Maclean, K., Sparrow, D. B., Barbera, J. P., and Dunwoodie, S. L. (2006). Loss of Cited2 affects trophoblast formation and vascularization of the mouse placenta. *Developmental Biology*, 294(1):67–82.

Xu, J. and Kidder, B. L. (2018). KDM5B decommissions the H3K4 methylation landscape of self-renewal genes during trophoblast stem cell differentiation. *Biology open*, 7(5).

Yadav, T., Quivy, J. P., and Almouzni, G. (2018). Chromatin plasticity: A versatile landscape that underlies cell fate and identity. *Science*, 361(6409):1332–1336.

Yamamoto, H., Flannery, M. L., Kupriyanov, S., Pearce, J., McKercher, S. R., Henkel, G. W., Maki, R. A., Werb, Z., and Oshima, R. G. (1998). Defective trophoblast function in mice with a targeted mutation of Ets2. *Genes & Development*, 12(9):1315.

Yu, G., Wang, L. G., and He, Q. Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383.

Zhang, B., Kim, M. Y., Elliot, G. N., Zhou, Y., Zhao, G., Li, D., Lowdon, R. F., Gormley, M., Kapidzic, M., Robinson, J. F., McMaster, M. T., Hong, C., Mazor, T., Hamilton, E., Sears, R. L., Pehrsson, E. C., Marra, M. A., Jones, S. J., Bilenky, M., Hirst, M., Wang, T., Costello, J. F., and Fisher, S. J. (2021). Human placental cytotrophoblast epigenome dynamics over gestation and alterations in placental disease. *Developmental Cell*, 56(9):1238–1252.e5.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137.

4.8 Appendix A: Notes

4.8.1 Data and resource availability

snATAC-seq datasets are available at the Gene Expression Omnibus website (`https://www.ncbi.nlm.nih.gov/geo/GSE227943`). All data generated and analyzed during this study are included in the published article and the online supporting files. All code used for the analyses are available at `https://github.com/Tuteja-Lab/MetrialGland-scATAC-seq`.

4.8.2 Acknowledgements

We would like to thank the Research IT group at Iowa State University (`http://researchit.las.iastate.edu`) for providing servers and IT support, and members of the Tuteja and Soares laboratories for their valuable discussions.

4.8.3 Funding

Supported by an NIH National Research Service, HD104495 (RLS), NIH grants: HD020676 (MJS), ES029280 (MJS), HD099638 (MJS), HD104033 (MJS, GT), HD105734 (MJS), HD096083 (GT) and the Sosland Foundation (MJS). Geetu Tuteja is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts. The views expressed are those of the author(s) and do not necessarily reflect the views of the funding agencies.

4.9 Appendix B: Supplementary tables and figures

All supplementary tables can be found online at Vu et al., 2023: https://journals.biologists.com/dev/article/150/15/dev201826/324977/Core-conserved-transcriptional-regulatory-networks.

4.10 Appendix C: Consent to include co-authored article in thesis/dissertation

THE PARTIES

Student Author A (Full Name, Major, and Institution)	Ha T.H. Vu, Bioinformatics and Computational Biology Iowa State University, Ames, IA, 50010
Student Author B (Full Name, Major, and Institution)	Regan L. Scott University of Kansas Medical Center, Kansas City, KS, 66160
Title of the co-authored section (Chapter, etc.)	Core conserved transcriptional regulatory networks define the invasive trophoblast cell lineage Chapter 4
Journal Name, Book Title, etc. (if applicable)	Journal: Development

(insert rows for additional graduate-student co-authors)

DISTRIBUTION OF TASKS AND RESPONSIBILITIES

In the table below, provide the overall percentage effort of each student and describe roles each author played, such as planning, executing, data analysis, writing, etc.

Names of Student Authors	Percentage	Description of Roles
Ha T.H. Vu	50%	Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing
Regan L. Scott	50%	Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing

The CRediT taxonomy is taken from https://credit.niso.org/. Go to the link to see the descriptions of contributor roles.

CHAPTER 5. GENERAL CONCLUSION

In this thesis, we utilized both bulk and single cell NGS datasets generated from rodent placenta, as well as developed a novel ATAC-seq peak calling framework, to better understand the molecular mechanisms of placenta development. First, we used transcriptome data from early to mid-gestation mouse placenta to infer and study gene regulatory networks specific to different developmental stages. We determined network modules with similar gene profiles to specific cell types in the human placenta, and as a result identified novel genes regulating placental development processes. Second, we developed a deep learning framework to identify chromatin accessible regions using shared signal from biological replicates, with higher precision and recall rates than other existing tools. Last, we integrated single-cell RNA-seq and single-nucleus ATAC-seq generated from the rat uterine-placental interface to identify a conserved transcription factor - gene network that may be active in the invasive trophoblast lineage in both rat and human. This chapter serves as a summary of the critical findings as well as potential future directions of the research in this dissertation.

5.1 Specific findings and contributions

5.1.1 Identifying novel regulators of placental development using time-series transcriptome data

In this report, we generated RNA-seq data from mouse fetal placental tissues at e7.5, e8.5 and e9.5. Next, we used hierarchical clustering and pair-wise differential expression analysis to identify three gene groups with timepoint-specific expression patterns. Using gene ontology analysis, we showed that the gene groups are significantly associated with key developmental functions in the placenta such as “trophoblast giant cell differentiation” and “labyrinthine layer development”. These timepoint-specific groups also have similar gene profiles to distinct cell

populations in the human placenta, which was determined using deconvolution analysis and cell-type enrichment tests. With network analysis, we identified sub-network modules with significantly similar gene profiles to human extravillous trophoblast (EVT), villous cytotrophoblast, syncytiotrophoblast, fibroblast and endothelial cells, then determined important genes using various network interaction metrics. Last, we validated the biological roles of four predicted regulators, *Siah2*, *Mtdh*, *Hnrnpk*, and *Ncor2* in the HTR8-SVneo cells, providing evidence for conserved functions of the identified candidates across species.

5.1.2 Unsupervised contrastive peak caller for ATAC-seq

In this work, we developed a novel deep learning framework called Replicated Contrastive Learning (RCL) to identify chromatin accessible regions, also known as "calling peaks", with ATAC-seq data. Here, we addressed an important problem of integrating shared signals across biological replicates to identify open genomic regions with higher precision and recall rates by using contrastive learning, an unsupervised deep learning techniques. Our model is unsupervised learning; therefore, it can handle higher dimension than traditional statistical learning, and does not require labels acquired from external sources like in supervised learning methods. Moreover, the framework provides researchers with an approach to integrate biological replicates without having to specify a hard cut-off of meaningful replicate numbers or carrying out post-hoc p-value correction. We have demonstrated that RCL performed superior to multiple existing tools based on precision, recall, F1 and area under curve metrics obtained using independent sets of labels in not only mouse placenta tissue data but also several human cell line data.

5.1.3 Core conserved transcriptional regulatory networks define the invasive trophoblast cell lineage

In this report, we generated single-nucleus (sn) ATAC-seq data from the uterine-placental interface in rat tissues at gestational day 15.5 and 19.5, then integrated with previously published single-cell (sc) RNA-seq with matching conditions and bulk ATAC-seq data from human EVT

cells. First, upon the integration of snATAC-seq and scRNA-seq, we determined the chromatin accessibility profiles of major cell populations in the uterine-placental interface, namely invasive trophoblast, natural killer, macrophage, endothelial, and smooth muscle cells. Specifically in the invasive trophoblast population, we observed that genes with multiple accessible regions tend to have crucial roles in the regulation of trophoblast functions. We also carried out motif enrichment analysis in important open regions in the rat invasive trophoblast, and predicted TF families that may bind in the same locations to interact and regulate cell type-specific functions. Last, we integrated the accessible peaks in rat invasive trophoblast with those in human EVT to identify a conserved network of transcription factors and genes that likely play functional roles in the trophoblast lineage. Upon the conclusion of this study, we have provided a valuable resource of novel regulatory regions in major cell types at the uterine-placental interface in rat, genomic locations for predicted transcription factor binding sites and their associated genes in the rat invasive trophoblast population, and conserved elements between rat invasive trophoblast and human EVT. This will be crucial for future hypothesis generation and to identify novel biomarkers of the trophoblast lineage.

5.2 Future directions

Despite our efforts, we acknowledge several factors that could be improved in our studies. First, we recognize that our findings could benefit from further experimental validation both *in vitro* and *in vivo*. For example, we tested roles of four candidates (*Siah2*, *Mtdh*, *Hnrnpk*, and *Ncor2*) in cell migration using the HTR8-SVneo cell line, which has certain differences to first trimester primary trophoblast cells such as in their miRNA expression profiles (Donker et al., 2012). It would also be beneficial to validate various novel targets identified when we integrated rat snATAC-seq data with human EVT ATAC-seq data. Useful models for future experiments are human TB stem cells derived with the Okae protocol (Okae et al., 2018) or gene knockout *in vivo*.

Second, our studies could be strengthened with the generation of matching-conditioned data in mouse, rat and human. For example, our original deconvolution analysis on mouse e7.5, e8.5

and e9.5 relied on human scRNA-seq data, which suffered from technical and biological noise of the differences in both the experimental assays and biological models. Since the publication of our study, scRNA-seq data in mouse placenta in matching timepoints has been generated (Jiang et al., 2023), which could be useful to integrate with our data to identify mouse-specific cell populations.

Last, our peak calling framework RCL is only one of the first steps to utilize shared signals across biological replicates. The model can be easily adapted for ChIP-seq and CUT&Tag peak calling with minor changes, which will be useful to predict high confident transcription factor binding sites or regions pulled down by histone modification ChIP-seq. We acknowledge that RCL is limited when determining open region boundaries, which is also a common challenge faced by all existing peak calling methods. Future extension of RCL models to enable varied-length inputs could serve as a potential solution to this disadvantage of RCL. Finally, RCL is currently implemented explicitly for bulk ATAC-seq data. It is an interesting future direction to extend RCL to call peaks for sc/snATAC-seq data, which is much sparser and noisier than bulk-level data.

5.3 References

Donker, R. B., Mouillet, J. F., Chu, T., Hubel, C. A., Stolz, D. B., Morelli, A. E., and Sadovsky, Y. (2012). The expression profile of C19MC microRNAs in primary human trophoblast cells and exosomes. *Molecular Human Reproduction*, 18(8):417–424.

Jiang, X., Wang, Y., Xiao, Z., Yan, L., Guo, S., Wang, Y., Wu, H., Zhao, X., Lu, X., and Wang, H. (2023). A differentiation roadmap of murine placentation at single-cell resolution. *Cell Discovery*, 9(1):30.

Okae, H., Toh, H., Sato, T., Hiura, H., Takahashi, S., Shirane, K., Kabayama, Y., Suyama, M., Sasaki, H., and Arima, T. (2018). Derivation of Human Trophoblast Stem Cells. *Cell Stem Cell*, 22(1):50–63.e6.