Inferring crucial pathways needed for differentiation of stem cells to required lineage

by

Haseena A

MT23246

Under the Supervision of Dr. Vibhor Kumar

Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi - 110020

18th June 2025

June 2025

Inferring crucial pathways needed for differentiation of stem cells to required lineage

A Thesis Report

submitted by

Haseena A

MT23246

in partial fulfilment of the requirements
for the award of the degree of

Master of Technology

to

Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi - 110020

18th June 2025

THESIS CERTIFICATE

This is to certify that the thesis titled "**Inferring crucial pathways needed for differentiation of stem cells to required lineage**" being submitted by **Haseena A**, to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of **Master of Technology**, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Vibhor Kumar
Associate Professor
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Place : New Delhi

Date: 18th June 2025

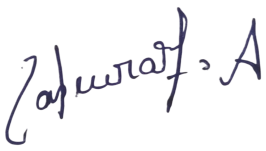
ACKNOWLEDGEMENTS

I extend my heartfelt appreciation to **Dr. Vibhor Kumar** for his invaluable guidance and support during my M.Tech thesis. His expertise, encouragement, and constructive feedback have been crucial in shaping my work, inspiring me to strive for excellence and overcome challenges. Dr. Vibhor Kumar's commitment to creating a supportive research atmosphere and his confidence in my capabilities have been incredibly inspiring. It has been a privilege to work under his guidance, and I am grateful for his significant time and contributions to my academic journey.

I am also grateful to Madhu ma'am, a PhD student under Dr. Vibhor Kumar, whose unwavering support and invaluable assistance played a pivotal role in completing my thesis. Her expertise, guidance, and encouragement were essential throughout this process, and I am truly grateful for her input. I would also like to thank Ariba ma'am for her helpful insights and kind guidance during critical stages of the project, which added clarity and direction to my work.

Furthermore, I extend my thanks to my friends for their continuous encouragement and support. Additionally, I acknowledge the contributions of the broader research community, whose insights shared through publications, conferences, and online platforms have been indispensable in shaping my research.

Lastly, I am profoundly thankful to my family for their unwavering love and support, which has been a constant source of strength throughout.



Haseena
MT23246

ABSTRACT

Regenerative medicine relies on the precise control of stem cell differentiation. While mesenchymal stem cells (MSCs) and human embryonic stem cells (hESCs) hold great promise, current differentiation methods struggle with efficiency, reproducibility, and a limited understanding of complex regulatory networks. Traditional genetic modification often yields unpredictable outcomes, and wet-lab methods are time and resource-intensive. This thesis presents a novel computational framework that systematically guides stem cell differentiation towards specific lineages without genetic modification.

By integrating single-cell RNA sequencing (scRNA-seq) and RNA velocity, the framework estimates the "poising levels" of MSCs and hESCs by capturing gene expression dynamics. Pathway enrichment scores from UniPath (a normalization-free gene-set enrichment tool) are combined with probabilistic graphical models to identify key signaling pathways influencing lineage decisions. A unique feature includes modeling bifurcations using relative RNA velocities of marker genes, enabling a pathway-centric view that accounts for cell variability.

We applied this framework to analyze human gastrulation using public scRNA-seq datasets, mapping developmental trajectories and identifying critical pathways (e.g., Wnt, BMP, TGFβ, FGF, Retinoic Acid) and transcription factors (e.g., ZSCAN10, STAT3, OTX2, SOX5, RUNX2) involved in ectoderm, mesoderm, and endoderm differentiation. The framework also revealed regulatory networks in endoderm-derived liver/pancreas and MSC-derived adipocyte, cartilage, and osteocyte differentiation. Bayesian Network inference and Random Forest analysis uncovered causal links between pathway activities and cell fates. Consistency with established developmental biology supports the validity of our computational predictions. This work offers a scalable and reproducible approach for stem cell engineering, advancing regenerative medicine.

Keywords: Stem Cell Differentiation, Mesenchymal Stem Cells (MSCs), Human Embryonic Stem Cells (hESCs), Single-cell RNA-seq, RNA Velocity, Poising, Lineage Commitment, Pathway Enrichment, UniPath, Bayesian Network, Random Forest, Computational Biology, Regenerative Medicine

Contents

List of Tables

List of Figures

Abbreviations

A-P: Anterior-Posterior
ACTH: Adrenocorticotropic Hormone
ADSC: Adipose-Derived Stem Cells
AI: Artificial Intelligence
AKT: Protein Kinase B
ALDH: Aldehyde Dehydrogenase
ALK1: Activin Receptor-Like Kinase 1
AM: Advanced Mesoderm
AMPK: AMP-activated Protein Kinase
ANPEP: Alanyl Aminopeptidase
ANOS1: Anosmin 1
AP-1: Activator Protein 1
APS: Anterior Primitive Streak
AR: Androgen Receptor
ARRB1: Arrestin Beta 1
ASB8: Ankyrin Repeat And SOCS Box Containing 8
ASH2L: ASH2 Like Histone Lysine Methyltransferase Complex Subunit
ATAC-seq: Assay for Transposase-Accessible Chromatin using sequencing
ATF6: Activating Transcription Factor 6
ATP: Adenosine Triphosphate
ATP6V0A2: ATPase H+ Transporting V0 Subunit A2
BAP31: B-cell Associated Protein 31
BBS4: Bardet-Biedl Syndrome 4
BCSC: Breast Cancer Stem Cells
BMP: Bone Morphogenetic Protein
BRCA: Breast Cancer
CAF: Cancer-Associated Fibroblasts
CCL: C-C Motif Chemokine Ligand
CDC: Cell Division Cycle
CDC20: Cell Division Cycle 20
CDKN3: Cyclin Dependent Kinase Inhibitor 3
CD-MSC: Cartilage-Derived Mesenchymal Stem Cell
CEBPZ: CCAAT/Enhancer Binding Protein Zeta
CENPF: Centrosome Protein F
CENPU: Centrosome Protein U
CENPX: Centrosome Protein X
CGNL1: Cingulin Like 1
CITED2: Cbp/P300 Interacting Transactivator With ED-Rich Tail 2
CLDN: Claudin
CLPB: Caseinolytic Peptidase B
CM: Cardiomyocyte
CRISPR/Cas9: Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated protein 9
CSCs: Cancer Stem Cells
CSF: Colony-Stimulating Factor
CSF3R: Colony Stimulating Factor 3 Receptor
CSV: Comma Separated Values
CTGF: Connective Tissue Growth Factor
CTHRC1: Collagen Triple Helix Repeat Containing 1
CTS: Cysteine-rich Transmembrane Spinal Cord 1
CTSK: Cathepsin K
CXCR4: C-X-C Motif Chemokine Receptor 4
DC: Diffusion Component
DCN: Decorin
DE: Definitive Endoderm
DEG: Differentially Expressed Genes
DES: Desmin
DHCR7: 7-Dehydrocholesterol Reductase

DLL3: Delta Like Canonical Notch Ligand 3
DNA: Deoxyribonucleic Acid
DNMT3B: DNA Methyltransferase 3 Beta
DRAM1: DNA Damage Regulated Autophagy Modulator 1
DSCAM: Down Syndrome Cell Adhesion Molecule
DST: Dystonin
ECDF: Empirical Cumulative Distribution Function
ECM: Extracellular Matrix
ECTO: Ectoderm
EGF: Epidermal Growth Factor
EGFR: Epidermal Growth Factor Receptor
EIF2D: Eukaryotic Translation Initiation Factor 2D
EM: Emergent Mesoderm
EMT: Epithelial-Mesenchymal Transition
ENA: European Nucleotide Archive
ENPP1: Ectonucleotide Pyrophosphatase/Phosphodiesterase 1
ENO3: Enolase 3
EOMES: Eomesodermin
Epi: Epiblast
EPO: Erythropoietin
ERK: Extracellular Signal-Regulated Kinase
ESR1: Estrogen Receptor 1
ESR2: Estrogen Receptor 2
ESRRA: Estrogen Related Receptor Alpha
ET-1: Endothelin 1
ETS: E26 Transformation-Specific
FABP: Fatty Acid Binding Protein
FAK: Focal Adhesion Kinase
FCGR: Fc Gamma Receptor
FDXACB1: Ferredoxin 1
FGF: Fibroblast Growth Factor
FGFR: Fibroblast Growth Factor Receptor
FHL: Four And A Half LIM Domains
FLI1: Friend Leukemia Virus Integration 1
FLT1: Fms Related Tyrosine Kinase 1
FN1: Fibronectin 1
FOXC1: Forkhead Box C1
FOXF1: Forkhead Box F1
FPKM: Fragments Per Kilobase of transcript per Million mapped reads
GATA: GATA Binding Protein
GDF: Growth Differentiation Factor
GEM: Gene Expression Matrix
GFI1B: Growth Factor Independent 1B Transcriptional Repressor
GLI: GLI Family Zinc Finger
GMPR: Guanosine Monophosphate Reductase
GSC: Goosecoid
GSN: Gelsolin
GSS: Glutathione Synthetase
GSTM5: Glutathione S-Transferase Mu 5
GTPASE: Guanosine Triphosphatase
GYG1: Glycogenin 1
GYG2: Glycogenin 2
GYPC: Glycophorin C
GYPB: Glycophorin B
HAPLN1: Hyaluronan And Proteoglycan Link Protein 1
HC: Hill-Climbing
HDAC: Histone Deacetylase
HDE1: High Density Endonuclease 1
HEDGEHOG: Hedgehog Signaling Pathway
HEMGN: Hemogen
HENMT1: HEN Methyltransferase 1
HEP: Hematopoietic Progenitors
hESC: Human Embryonic Stem Cell
HIF: Hypoxia-Inducible Factor
HJURP: Holliday Junction Recognition Protein
HLX: H2.0 Like Homeobox

HNF: Hepatocyte Nuclear Factor
HOXA1: Homeobox A1
HOXC: Homeobox C
HSC: Hematopoietic Stem Cell
ID2: Inhibitor Of DNA Binding 2
IFI: Interferon Induced Protein
IGF: Insulin-like Growth Factor
IGFBP: Insulin Like Growth Factor Binding Protein
IHSC: Induced Hematopoietic Stem Cells
IL: Interleukin
INPP: Inositol Polyphosphate-5-Phosphatase
INSIG1: Insulin Induced Gene 1
iPSC: Induced Pluripotent Stem Cell
IRF: Interferon Regulatory Factor
ISL1: ISL LIM Homeobox 1
ITGA: Integrin Alpha
ITGB: Integrin Beta
JAML: Junctional Adhesion Molecule Like
JDP2: Jun Dimerization Protein 2
JNK: Jun N-terminal Kinase
KDR: Kinase Insert Domain Receptor
KEGG: Kyoto Encyclopedia of Genes and Genomes
KIF: Kinesin Family Member
KIT: KIT Proto-Oncogene, Receptor Tyrosine Kinase
KLF: Kruppel-like Factor
KLHL: Kelch Like Family Member
KNN: K-Nearest Neighbors
KRT: Keratin
LAMA1: Laminin Subunit Alpha 1
LCP: Lymphocyte Cytosolic Protein
LDLRAD4: Low Density Lipoprotein Receptor Related Protein Adaptor Protein 4
LGR5: Leucine Rich Repeat Containing G Protein-Coupled Receptor 5
LIF: Leukemia Inhibitory Factor
LINC: Long Intergenic Non-Coding RNA
LIPA: Lipase A, Lysosomal Acid
LPL: Lipoprotein Lipase
LRP2: LDL Receptor Related Protein 2
LRRIQ1: Leucine Rich Repeat And IQ Motif Containing 1
LTBP1: Latent TGF Beta Binding Protein 1
LUM: Lumican
MAP2: Microtubule Associated Protein 2
MAPK: Mitogen-Activated Protein Kinase
MAZ: MYC Associated Zinc Finger Protein
MCAM: Melanoma Cell Adhesion Molecule
MCM: Minichromosome Maintenance Complex Component
MCOLN3: Mucolipin 3
MECOM: MDS1 And EVI1 Complex Locus
MEF2C: Myocyte Enhancer Factor 2C
MEPE: Matrix Extracellular Phosphoglycoprotein
MESP: Mesoderm Posterior BHLH Transcription Factor
MESDC2: Mesoderm Development Candidate 2
MESO: Mesoderm
MET: MET Proto-Oncogene, Receptor Tyrosine Kinase
METTL3: Methyltransferase Like 3
MHG: Mid-Hindgut
MICAL2: Microtubule Associated Monooxygenase, Calponin And LIM Domain Containing 2
MID1: Midline 1
MILR1: Myeloid LincRNA 1
MIXL1: Mix Paired-Like Homeobox 1
MME: Neprilysin
MMP: Matrix Metalloproteinase
MOXD1: Monooxygenase DBH-Like 1
MPL: MPL Proto-Oncogene,

Thrombopoietin Receptor
MPP1: Membrane Palmitoylated Protein 1
MPP: Membrane-Palmitoylated Protein
MPZL1: Myelin Protein Zero Like 1
MRC1: Mannose Receptor C-Type 1
MSC: Mesenchymal Stem Cell
MSI1: Musashi RNA Binding Protein 1
MSN: Moesin
MSX: Msh Homeobox
MTUS1: Mitochondrial Tumor Suppressor 1
MYB: MYB Proto-Oncogene, Transcriptional Regulator
MYC: MYC Proto-Oncogene, BHLH Transcription Factor
MYL4: Myosin Light Chain 4
MYLK: Myosin Light Chain Kinase
NAALAD2: N-Acetyl-Alpha-Linked Acidic Dipeptidase 2
NCAM1: Neural Cell Adhesion Molecule 1
NCL: Nucleolin
NCKAP1L: NCK Associated Protein 1 Like
NDC80: NDC80 Kinetochore Complex Component
NEBL: Nebulin Like
NES: Nestin
NEUROD1: Neuronal Differentiation 1
NFATC2: Nuclear Factor Of Activated T-Cells 2
NFIC: Nuclear Factor I C
NF-κB: Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B Cells
NGFR: Nerve Growth Factor Receptor
NID2: Nidogen 2
NKX: NK2 Homeobox
NM: Nascent Mesoderm
NNMT: Nicotinamide N-Methyltransferase
NODAL: Nodal Signaling Pathway
NOG: Noggin
NPR: Natriuretic Peptide Receptor
NPRL3: NPR3 Like, GATOR1 Complex Subunit
NR2A1: Nuclear Receptor Subfamily 2 Group A Member 1
NRG4: Neuregulin 4
NRIP1: Nuclear Receptor Interacting Protein 1
NRP: Neuropilin
NT5E: 5'-Nucleotidase Ecto
NTRK3: Neurotrophic Receptor Tyrosine Kinase 3
NUSAP1: Nucleolar And Spindle Associated Protein 1
ODAM: Odontogenic Ameloblast-Associated Protein
OPN: Osteopontin
ORC1: Origin Recognition Complex Subunit 1
OSGEP: O-Sialoglycoprotein Endopeptidase
OSTF1: Osteoclast Stimulating Factor 1
OTX: Orthodenticle Homeobox
P3H: Prolyl 3-Hydroxylase
PAX: Paired Box Gene
PBX1: Pre-B-Cell Leukemia Transcription Factor 1
PCA: Principal Component Analysis
PCAT14: Prostate Cancer Associated Transcript 14
PCDH: Protocadherin
PCLAF: PCNA Clamp Associated Factor
PDGF: Platelet-Derived Growth Factor
PDGFRA: Platelet Derived Growth Factor Receptor Alpha
PDGFB: Platelet Derived Growth Factor Subunit B
PDLIM4: PDZ And LIM Domain Protein 4

PDX1: Pancreatic And Duodenal Homeobox 1
PECAM1: Platelet Endothelial Cell Adhesion Molecule 1
PFG: Posterior Foregut
PHF19: PHD Finger Protein 19
PI3K: Phosphoinositide 3-Kinase
PID: Pathway Interaction Database
PIN1: Peptidylprolyl Cis/Trans Isomerase NIMA-Interacting 1
PITX: Paired-Like Homeodomain 2
PKIB: Protein Kinase Inhibitor Beta
PKP2: Plakophilin 2
PLAUR: Plasminogen Activator, Urokinase Receptor
PLC: Phospholipase C
PLXNA2: Plexin A2
PMP22: Peripheral Myelin Protein 22
PODXL: Podocalyxin Like
PPAR: Peroxisome Proliferator-Activated Receptor
PPARG: Peroxisome Proliferator Activated Receptor Gamma
PPP1R: Protein Phosphatase 1 Regulatory Subunit
PPP3CC: Protein Phosphatase 3 Catalytic Subunit Gamma
PPS: Posterior Primitive Streak
PRKAR2B: Protein Kinase CAMP-Dependent Type II Regulatory Subunit Beta
PRKCB: Protein Kinase C Beta
PRKCD: Protein Kinase C Delta
PRKCQ: Protein Kinase C Theta
PROM1: Prominin 1
PRRX1: Paired Related Homeobox 1
PS: Primitive Streak
PSTPIP2: Proline Serine Threonine Phosphatase Interacting Protein 2
PTAFR: Platelet Activating Factor Receptor
PTEN: Phosphatase And Tensin Homolog
PTN: Pleiotrophin
PTPRC: Protein Tyrosine Phosphatase, Receptor Type C
PTPRE: Protein Tyrosine Phosphatase, Receptor Type E
PTPN: Protein Tyrosine Phosphatase, Non-Receptor Type
PTPRM: Protein Tyrosine Phosphatase Receptor Type M
PTPRZ1: Protein Tyrosine Phosphatase Receptor Type Z1
PXK: PX Domain Containing Kinase
RA: Retinoic Acid
RAB: RAS Oncogene Family Member
RAC1: Rac Family Small GTPase 1
RAS: RAS Oncogene Family
RASL: RAS Like
RASSF4: Ras Association Domain Family Member 4
RCBTB2: Regulator Of Cell Cycle B And T Box 2
RCSD1: Rhizoclonium Sp. CSD1 Domain Containing
REACTOME: Reactome Pathway Database
RELN: Reelin
REST: RE1 Silencing Transcription Factor
REX: REXO5 Exonuclease
RHAG: Rh Associated Glycoprotein
RHOA: Ras Homolog Family Member A
RHOJ: Ras Homolog Family Member J
RHOG: Ras Homolog Family Member G
RGS: Regulator Of G Protein Signaling
RIN2: Ras And Rab Interacting Protein 2
RIPK4: Receptor Interacting Serine/Threonine Kinase 4

RIPPLY2: RIPPLY C-Terminal Domain Containing 2
RNA: Ribonucleic Acid
RNA-seq: RNA Sequencing
RND3: Rho Family GTPase 3
RNF24: Ring Finger Protein 24
ROR1: Receptor Tyrosine Kinase Like Orphan Receptor 1
RPKM: Reads Per Kilobase of transcript per Million mapped reads
RPS: Ribosomal Protein S
RUNX2: RUNX Family Transcription Factor 2
RXRA: Retinoid X Receptor Alpha
S1P: Sphingosine-1-Phosphate
SALL4: Spalt Like Transcription Factor 4
SAMD3: Sterile Alpha Motif Domain Containing 3
SAMSN1: SAM And SH3 Domain Containing 1
SCARA5: Scavenger Receptor Class A Member 5
scRNA-seq: Single-Cell RNA Sequencing
SDRF: Sample and Data Relationship Format
SDHA: Succinate Dehydrogenase Complex Flavoprotein Subunit A
SEL: Selectin
SELP: Selectin P
SEMA: Semaphorin
SERINC5: Serine Incorporator 5
SFRP1: Secreted Frizzled Related Protein 1
SGK1: Serum Glucocorticoid Regulated Kinase 1
SHH: Sonic Hedgehog
SHISAL2B: Shisa Like 2B
SIPA1L2: Signal-Inducing Adapter Molecule 1 Like 2
SLC: Solute Carrier Family
SLCO2A1: Solute Carrier Organic Anion Transporter Family Member 2A1
SLUG: SNAI2
SMAD: Mothers Against Decapentaplegic Homolog
SMIM1: Small Integral Membrane Protein 1
SMN1: Survival Of Motor Neuron 1, Telomeric
SMO: Smoothened Frizzled Class Receptor
SMYD3: SET And MYND Domain Containing 3
SNAI: Snail Family Transcriptional Repressor
SNX8: Sorting Nexin 8
SOX: SRY-Box Transcription Factor
SP100: SP100 Nuclear Antigen
SP7: Sp7 Transcription Factor
SPARC: Secreted Protein Acidic And Cysteine Rich
SPOCK3: SPARC/Osteonectin, Cwcv And Kazal Like Domains Proteoglycan 3
SPTA1: Spectrin Alpha, Non-Erythrocytic 1
STAT: Signal Transducer And Activator Of Transcription
STAB1: Stabilin 1
STAR: Spliced Transcripts Alignment to a Reference
STX: Syntaxin
SULT1C4: Sulfotransferase Family 1C Member 4
SVEP1: Sushi, Von Willebrand Factor A And EGF Domain Containing Protein 1
SYK: Spleen Tyrosine Kinase
SYT: Synaptotagmin
TAGLN3: Transgelin 3
TAL1: TAL BHLH Transcription Factor 1
TANGO2: Transport And Golgi Organization 2 Homolog

TBX: T-Box Transcription Factor
TBXA: Thromboxane A Synthase
TBXT: T-Box Transcription Factor T (Brachyury)
TCF: Transcription Factor
TCR: T Cell Receptor
TDGF1: Teratocarcinoma Derived Growth Factor 1
TEAD1: TEA Domain Transcription Factor 1
TEK: TEK Receptor Tyrosine Kinase
TENM4: Teneurin Transmembrane Protein 4
TESC: Tescalcin
TF: Transcription Factor
TFR2: Transferrin Receptor 2
TFRC: Transferrin Receptor
TGFβ: Transforming Growth Factor Beta
TGFβ: Transforming Growth Factor Beta
TGFβR: Transforming Growth Factor Beta Receptor **THBS1**: Thrombospondin 1
THY1: Thymus Cell Antigen 1
TIE1: Tyrosine Kinase With Immunoglobulin And EGF Homology Domains 1
TLR4: Toll Like Receptor 4
TMEM: Transmembrane Protein
TMPRSS4: Transmembrane Serine Protease 4
TNF: Tumor Necrosis Factor
TNFAIP1: TNF Alpha Induced Protein 1
TNFSF: TNF Superfamily Member
TNNT2: Troponin T2, Cardiac Type
TNC: Tenascin C
TOP2A: Topoisomerase II Alpha
TPO: Thrombopoietin
TPX2: TPX2 Microtubule Nucleation Factor
TRAF3IP3: TRAF3 Interacting Protein 3
TRAP1: TNF Receptor Associated Protein 1
TRPC6: Transient Receptor Potential Cation Channel Subfamily C Member 6
TRPV2: Transient Receptor Potential Cation Channel Subfamily V Member 2
TTC: Tetratricopeptide Repeat Domain
TUBB3: Tubulin Beta 3 Class III
TWIST: Twist Family BHLH Transcription Factor
TYMP: Thymidine Phosphorylase
UBA7: Ubiquitin Like Modifier Activating Enzyme 7
UBAC1: Ubiquitin Associated And Coiled-Coil Domain Containing 1
UBE2: Ubiquitin Conjugating Enzyme E2
UGP2: UDP-Glucose Pyrophosphorylase 2
UMI: Unique Molecular Identifier
UMAP: Uniform Manifold Approximation and Projection
UNC5C: UNC5 Netrin Receptor C
UniPath: Normalization-free Gene-set Enrichment for Single-cell Expression
URAD: Uridine Phosphorylase
USP4: Ubiquitin Specific Peptidase 4
VAV1: Vav Guanine Nucleotide Exchange Factor 1
VAV3: Vav Guanine Nucleotide Exchange Factor 3
VCAM1: Vascular Cell Adhesion Molecule 1
VCAN: Versican
VEGF: Vascular Endothelial Growth Factor
VEGFR: Vascular Endothelial Growth Factor Receptor
VSNL1: Visonin Like 1
WDR35: WD Repeat Domain 35
WISP2: WNT1 Inducible Signaling Pathway Protein 2
WNT: Wingless-Type MMTV Integration Site Family
WT1: WT1 Transcription Factor
ZADH2: Zinc Binding Alcohol

Dehydrogenase 2

ZEB: Zinc Finger E-Box Binding Homeobox

ZIC1: Zic Family Member 1

ZNF: Zinc Finger Protein

ZSCAN10: Zinc Finger And SCAN Domain Containing 10

Notation

β	Beta
κ	Kappa

CHAPTER 1

Introduction

Regenerative medicine has emerged as a transformative field with vast potential to repair, replace, and regenerate damaged tissues and organs. At the core of this field, mesenchymal stem cells (MSCs), particularly those derived from umbilical cord sources, have garnered significant attention due to their multipotency, immunomodulatory properties, and non-invasive harvesting potential. Similarly, human embryonic stem cells (hESCs), with their ability to differentiate into virtually all cell types, offer immense promise for regenerative applications. However, despite substantial progress in the field, challenges persist in achieving precise and efficient differentiation of MSCs and hESCs into specific cell lineages [1].

The challenges that hinder stem cell differentiation are multifaceted. These include variations in the initial potency of stem cells, a limited understanding of the complex signaling pathways that regulate differentiation, and the inherent inefficiencies of traditional differentiation methods. Conventional techniques, which often rely on genetic modifications through plasmids, can lead to unpredictable outcomes in gene expression, resulting in variability and imprecise lineage commitment. Furthermore, the reliance on trial-and-error strategies in wet-lab experiments to determine the optimal differentiation conditions is not only time-consuming but also resource-intensive, limiting the scalability and reproducibility of differentiation protocols.

Another significant hurdle is the identification of key cell-type markers and the correct signaling pathways required for directing stem cell differentiation. Without a structured computational framework to predict these pathways, researchers must experiment with numerous combinations of signaling factors and pathway perturbations, an approach that is highly inefficient. In this context, the field lacks an integrated method to systematically guide stem cell differentiation toward specific lineages.

This research introduces a novel computational framework designed to address these challenges. By leveraging single-cell RNA sequencing (scRNA-seq) data, we estimate the poising levels of MSCs and hESCs toward different cell lineages through RNA velocity analysis. RNA velocity [2], which captures the temporal dynamics of gene expression at the single-cell level, provides insights into the differentiation trajectory of stem cells. Using probabilistic graphical models, we model RNA velocity distributions across different lineages and combine them with enrichment scores derived from known signaling pathways. This enables the identification of key signaling pathways that directly influence lineage

commitment and guides the efficient differentiation of stem cells. Crucially, our approach eliminates the need for genetic modification, as we focus on small molecules to drive differentiation, a more precise and less invasive method.

Additionally, we introduce an innovative framework for modeling differentiation bifurcations. By analyzing the relative RNA velocities of marker genes from multiple potential lineages, we gain a better understanding of the decision-making process that drives stem cells to commit to a particular lineage. This method allows us to predict and control differentiation outcomes with higher precision, providing a pathway-centric protocol that accounts for the intrinsic variability among stem cells derived from different sources. This improved understanding of stem cell dynamics could lead to more reproducible and efficient differentiation protocols.

The findings of this research have significant implications for both stem cell engineering and regenerative medicine. By providing a computational approach to guide the differentiation process, this work opens the door to more precise tissue-specific differentiation protocols. It also offers a promising alternative to traditional genetic modification techniques, potentially reducing the risks associated with genetic alterations in clinical applications. Ultimately, this research lays the foundation for developing more efficient, scalable, and reproducible methods for stem cell-based therapies, thereby advancing the field of regenerative medicine.

This thesis outlines the methodology, validation, and application of the proposed computational approach, illustrating its potential to revolutionize stem cell differentiation protocols. Through systematic pathway identification and lineage prediction, our approach offers a step toward more controlled and precise stem cell differentiation, addressing longstanding challenges in the field and fostering advancements in regenerative medicine.

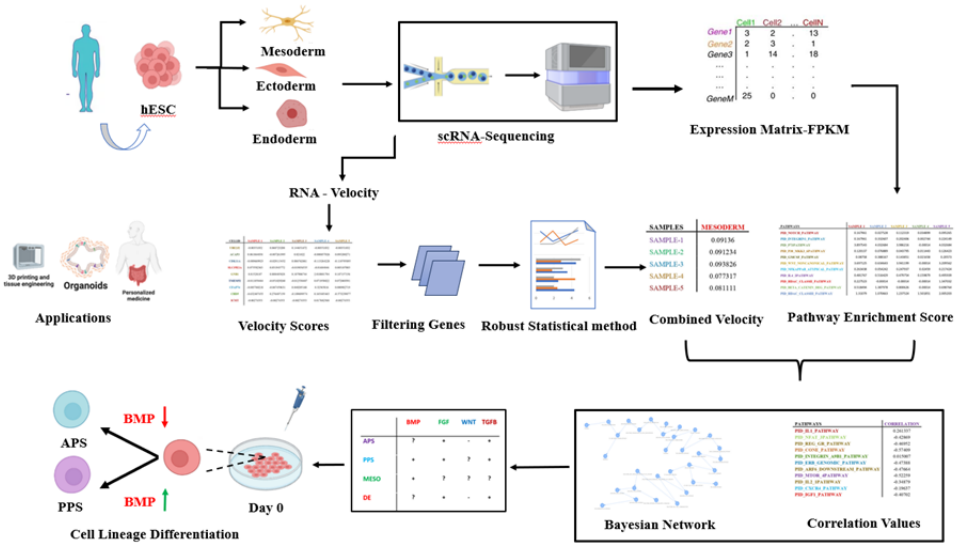


Figure 1.1: Pipeline Description: Single-Cell Transcriptome-Based Method and Markers for Differentiating Umbilical Cord Mesenchymal Stem Cells to Desired Lineages.

This pipeline describes a computational methodology for identifying pathways that directly

influence the differentiation potential of umbilical cord mesenchymal stem cells (UC-MSCs) or progenitor cells toward specific lineages. Single-cell transcriptomic data from progenitor cells are preprocessed to ensure quality and uniformity. RNA velocity analysis is employed to predict the future cellular states, offering insights into the dynamic trajectories of differentiation. Marker genes relevant to the desired lineage are extracted and combined using robust statistical techniques. A pathway enrichment analysis is conducted on single-cell expression data to calculate enrichment scores, which highlight critical biological pathways associated with the differentiation process. Advanced AI-based algorithms are utilized to integrate RNA velocity and pathway enrichment scores, enabling the identification of key pathways that regulate lineage specification. This integrated approach is designed to uncover novel pathways that can be leveraged in the development of organoids and personalized medicine applications. The identified pathways and marker genes provide a robust framework for directing stem cell differentiation, with potential applications in regenerative medicine, drug discovery, and therapeutic development.

CHAPTER 2

Methodology

This chapter describes the data collection, preprocessing, alignment, and quality control procedures used for single-cell RNA sequencing (scRNA-seq) data analysis. The workflow was designed to enable integrated analysis of lineage-specific transcriptional dynamics with an emphasis on RNA velocity.

2.1 Data Acquisition

Publicly available scRNA-seq datasets were obtained from the Gene Expression Omnibus (GEO), hosted by the National Center for Biotechnology Information (NCBI) [3]. Selection criteria included availability of raw FASTQ files, annotation of differentiation stages, and compatibility with RNA velocity analysis.

For this analysis, five scRNA-seq datasets were selected from GEO, each offering distinct biological insights into stem cell differentiation and mesenchymal development. The selected datasets are

Datasets

The following GEO datasets were included:

- **GSE75748**: Profiles human embryonic stem cells (hESCs) differentiating toward definitive endoderm. Time-series sampling enables dynamic analysis of lineage commitment. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75748
- **GSE85066**: Captures transcriptional changes during osteogenic and adipogenic differentiation of human umbilical cord-derived mesenchymal stem cells (hUC-MSCs). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85066
- **GSE199071**: Contains single-cell profiles of early mesodermal differentiation, providing insight into lineage bifurcation and regulatory networks. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE199071
- **GSE241130**: Offers high-resolution scRNA-seq data from human embryoid bodies at various stages, enabling analysis of multilineage potential. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE241130
- **GSE107552**: Profiles cells undergoing neural differentiation, contributing to velocity analysis in ectodermal contexts. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107552

Raw sequencing files were retrieved from the Sequence Read Archive (SRA) using prefetch and converted to FASTQ using fasterq-dump [4]. Metadata and cell annotations were parsed to

retain only relevant samples corresponding to undifferentiated or early differentiation stages.

2.2 Quality Control

Initial read-level quality control was conducted using FastQC, which assessed per-base sequence quality, GC content, duplication levels, and adapter contamination[5]. Summary reports were aggregated using MultiQC [6] for batch comparison. Samples showing excessive adapter content, high duplication rates, or low read quality were excluded from further processing.

2.3 Read Alignment

Reads were aligned to the GRCh38 human reference genome using the STAR aligner [7], which enables accurate mapping of spliced reads essential for RNA velocity estimation. Genome indices were generated based on GENCODE GTF annotation files. Key alignment parameters included the use of `--twopassMode Basic` to improve splice junction detection, `--outSAMtype BAM SortedByCoordinate` to produce coordinate-sorted BAM files, `--quantMode TranscriptomeSAM` to generate transcriptome-aligned BAMs compatible with velocity analysis, and `--outSAMattributes NH HI AS nM` to retain read-level alignment attributes. These configurations ensured effective capture of both exonic and intronic reads necessary for distinguishing unspliced and spliced transcripts.

2.4 Post-Alignment Processing

Post-alignment processing was performed using SAMtools [8] to compute alignment quality metrics and prepare data for downstream analyses. BAM files were indexed to facilitate efficient data access. Reads with low mapping quality or identified as PCR duplicates were filtered out. Further quality assessments included evaluation of mapping rates, insert size distributions, and overall alignment depth. Samples exhibiting poor alignment statistics or excessive dropout rates were excluded from subsequent RNA velocity analyses to ensure data reliability.

2.5 RNA Velocity Analysis

2.5.1 Concept of RNA Velocity

RNA velocity is a computational method that infers the future state of individual cells by analyzing the transcriptional dynamics of RNA at single-cell resolution [2]. Unlike static

transcriptomic snapshots, RNA velocity captures the direction and speed of cellular transitions by quantifying unspliced (nascent) and spliced (mature) mRNA within each cell. This dynamic information enables the reconstruction of developmental trajectories, lineage relationships, and prediction of cell fate decisions, thereby providing temporal context to single-cell RNA sequencing (scRNA-seq) data.

2.5.2 Velocyto: Foundational Framework for RNA Velocity

Velocyto is a pioneering tool that estimates RNA velocity by leveraging counts of spliced and unspliced transcripts derived from aligned scRNA-seq data [2]. It assumes a steady-state model wherein transcription, splicing, and degradation rates are balanced, and infers velocity vectors by comparing observed ratios of unspliced to spliced RNA against expected equilibrium levels. The output includes velocity vectors visualized as directional arrows overlaid on dimensionality reduction embeddings such as t-SNE or UMAP, indicating the trajectory of cellular transitions. Although Velocyto represents a significant advance, its steady-state assumption limits sensitivity to transient or rapidly changing gene expression states.

2.5.3 scVelo: Advanced Modeling of Transcriptional Dynamics

Building on Velocyto's framework, scVelo introduces more flexible modeling approaches that relax steady-state constraints to better capture dynamic transcriptional processes [9].

2.5.3.1 Steady-State Model in scVelo

The steady-state model in scVelo parallels Velocyto's approach, positing equilibrium between RNA species. It estimates RNA velocity by quantifying deviations of unspliced RNA levels from expected steady-state ratios relative to spliced RNA. This model provides computational efficiency and is appropriate for cells in relatively stable states but may fail to detect rapid transcriptional changes [9].

2.5.3.2 Stochastic (Dynamic) Model in scVelo

To address limitations of steady-state assumptions, scVelo's stochastic model employs ordinary differential equations (ODEs) that describe time-dependent changes in transcription and splicing rates without equilibrium constraints. Parameters vary with pseudotime, enabling reconstruction of transient gene expression dynamics and latent cellular ordering (latent time) [9]. This approach is particularly effective for differentiating or perturbed cells, offering biologically realistic insights into developmental trajectories and transient states. It is, however, more computationally demanding.

2.5.3.3 Likelihood-Based Dynamical Model

scVelo further incorporates a likelihood-based dynamical model that accounts for measurement noise and biological variability. By maximizing the likelihood of observed spliced and unspliced counts under the kinetic model, it produces robust RNA velocity estimates and facilitates hypothesis testing on gene-specific kinetics. This probabilistic approach enhances velocity inference accuracy, especially in noisy or sparse single-cell datasets [9].

2.5.4 Integration and Perspectives

RNA velocity analysis, originating from Velocyto and refined by scVelo, provides a powerful framework for incorporating temporal dynamics into scRNA-seq studies. The integration of steady-state, stochastic, and likelihood-based models allows for comprehensive characterization of cellular trajectories and regulatory mechanisms in both stable and dynamic systems. These methods enable detailed investigation of development, disease progression, and cellular responses at unprecedented temporal resolution.

2.5.5 Data Sources and Implementation

The RNA velocity analyses in this study utilized publicly available single-cell RNA-seq datasets accessed from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). Raw sequencing reads were processed and aligned as described in sections 2.3 and 2.4, with spliced and unspliced transcript quantification performed using the standard Velocyto pipeline [2]. Subsequent RNA velocity estimations and dynamic modeling were implemented with scVelo in Python, employing both steady-state and stochastic models to capture the transcriptional dynamics of the investigated cell populations [9].

2.6 Marker Selection for Lineage-Specific Analysis in RNA Velocity Studies

Marker selection is essential for accurately interpreting cellular differentiation trajectories in RNA velocity analyses. Lineage-specific genes serve as transcriptional signatures that reflect dynamic changes in cell states, enabling precise mapping of temporal progression and future states in differentiation processes.

Biologically, these markers include transcription factors, surface proteins, and signaling molecules uniquely expressed at specific differentiation stages. Their dynamic expression patterns form the basis for RNA velocity estimations by capturing shifts in both spliced and

unspliced transcripts, which provide insight into regulatory mechanisms and lineage commitment.

Selecting reliable markers involves reviewing literature and databases to identify genes with known functional roles in lineage specification. Emphasis is placed on genes exhibiting temporal expression changes consistent with early, intermediate, or terminal differentiation phases. Validation across datasets and experimental models ensures robustness and reproducibility of selected markers.

2.7 Marker Sets for Differentiation from hESCs and Mesenchymal Stem Cells

For human embryonic stem cell (hESC) differentiation, lineage markers reflect the commitment to the three germ layers:

Table 2.1: Comprehensive Gene Markers for Germ Layers

Germ Layer	**Associated Genes**
Mesoderm	INHBA, INHBB, INHBE, BMP2, BMP4, BMP6, BMP7, TGFβ1, TGFβ2, TGFβ3, WNT3A, WNT8A, WNT3, T, EOMES, FOXC1, FOXF1, GSC, HAND1, MIXL1, SNAI2, SNAI1, TBX6, TWIST1, TWIST2, HLX, MESP1, MESP2, MEST, NKX2-5, KDR, CDH2, CFC1, FABP4, FGF5, GDF1, GDF3, GSC2, INHBC, NCL, NODAL, SLUG, TBXT, A-FABP, BMP2A, MESDC2, ESR1, ESR2, BRA, CDX2, DCN, DES, GATA2, GATA4, IGF2, MSX1, NCAM1, PDGFB, PDGFRA, SMN1, TNNT2, WT1, WNT11, WNT4, WNT6, WNT5B, BMP1, BMP2K, BMP3, BMP5, BMP7-AS1, BMP8A, BMP8B, BMPER, BMPR1A, BMPR1B, BMPR1B-DT, TGFβI, ACVRC1, FGF7, FGF12, FGF17, GDF7, WNT2, WNT2B, WNT1, WNT5A
Ectoderm	BMP4, CHRD, FGF8, FOXJ3, GBX2, NES, NOG, OTX2, TP63, PAX2, PAX6, TUBB3, PAX7, SOX1, MSI1, NKX22, NKX61, FOXP2, FOXD3, SOX10, ZIC1, CDH1, DCX, FGF5, IGF1, NCAM1, NEUROD1, OTX1, ZIP2, NOTCH1, MAP2, FGFBP3, FGFR2, FGF3, FGF6, FGF14, FGF9, FGF11, PAX6-AS1, BMP, FGF17

Table 2.1 – continued from previous page

Germ Layer	Associated Genes
Endoderm	AFP, ALB, CLIC6, CTNNB1, CXCR4, CDX2, CLDN8, ECD, EOMES, FABP1, FABP2, FOXA1, FOXA2, GATA4, GATA6, GDF1, GDF3, GSC, HDE1, HNF1B, HNF3A, HNF3B, HNF4A, KIT, KRT19, MIXL1, NR2A1, PDX1, RIPK4, SOX17, SOX7, ST14, TCF2, TMPRSS4, FGF17, FGF6, FGFRL1, FGF23, FGF4, FGF8, FGF3, FGFBP2, FGFBP1, FGFR2, FGF10-AS1, FGF7, FGFR3, FGF10, FGF14, FGF9, FGF7P6, FGF2, FGF13-AS1, FGF19, TGFβR3L, TGFβ1I1, TGFβR3, TGFβ2-AS1, TGFβR2, FABP3, HNF4, CLDN, FOXP2, SOX10

The inherent capacity for self-renewal and pluripotency of Human Embryonic Stem Cells (hESCs) allows for their controlled differentiation into the three fundamental embryonic germ layers: the ectoderm, mesoderm, and endoderm. This developmental process, known as differentiation, is orchestrated through a precise interplay of signaling cascades and modulated expression patterns. As shown in the preceding table, specific lineage markers emerge during this commitment, providing definitive evidence of a cell's progression towards a particular germ layer. Therefore, these markers are indispensable tools in research, enabling the rigorous assessment and confirmation of cellular fate decisions within differentiation experiments.

Mesenchymal stem cell (MSC) differentiation into osteoblasts, adipocytes, and cartilage is tracked using key transcriptional regulators:

Table 2.2: Markers for Adipocyte, Cartilage, and Osteocyte Differentiation

Cell Type	Associated Markers
Adipocyte	COL1A2, CFD, LUM, MGP, CCN5, THY1, APOD, CEBPA, EGFR, FGF10, GSC, MMP3, PPARG, TWIST2, GATA3, GLI1, SMO, STAB1, WISP2, WT1, SCARA5, HOXC8, TCF7L1, HOXC9, HOXC6, HOXC4, ZNF521, SULT1E1, EBF2, NRG4, COL1A1, APOH, APOLD1
Cartilage	CENPF, UBE2C, MCAM, CKAP2, PCLAF, CDKN3, CDC20, NGFR, ITGB1, NOTCH1, ITGA5, MCAM, MME, NCAM1, TFRC, THY1, VCAM1, TNFRSF1A, NT5E, PTPRC, KIT, ANPEP, ERBB2, CD19, PECAM1, ITGAV, CD34, CD44, NT5E, MCAM, CD9, THY1, MCAM, COL2A1, CENPX, CENPU

Table 2.2 – continued from previous page

Cell Type	Associated Markers
Osteocyte	ALPP, ALPI, COL1A1, COL2A1, DCN, MCAM, MEPE, SP7, RUNX2, TPO, ALPP, ALPI, BAP31, FN1, IGFBP3, BGLAP, OPN, SP7, SCUBE3, SPARC, TMEM119, UBE2D1, UBE2Q2P1, UBE2G2, UBE2QL1, UBE2J1, UBE2A, CDK10, CDK14, CDKN2A, CDKN2B, CDCA5

Characterized by their inherent multipotency, Mesenchymal Stem Cells (MSCs) undergo directed differentiation to form specialized cell types such as adipocytes (fat), chondrocytes (cartilage), and osteocytes (bone). This process of cellular commitment is driven by specific biological signals. As illustrated in the foregoing table, the emergence of distinct lineage-specific markers serves as a definitive confirmation of successful differentiation into these mature cell populations. For researchers, these markers are indispensable for the precise validation of differentiation outcomes in laboratory settings.

2.8 Application of Lineage Markers in RNA Velocity Analysis

Integrating lineage markers into RNA velocity analyses enhances biological interpretation by anchoring velocity vectors to known differentiation programs. This integration allows for refined trajectory mapping, resolving lineage bifurcations, and providing insights into dynamic gene regulation. Tracking velocity changes in key transcription factors highlights regulatory events that drive cell fate decisions, enabling a deeper understanding of developmental and pathological processes.

2.9 Integrated Strategies for RNA Velocity Analysis Using Lineage-Specific Markers

RNA velocity analysis, utilizing lineage-specific gene markers, employs three distinct methods to understand cellular differentiation and gene expression dynamics, particularly in mesoderm, ectoderm, endoderm, and mesenchymal lineages.

2.9.1 Method 1: Averaging RNA Velocity Across Lineage-Specific Gene Markers

This simple method computes an average RNA velocity for a set of pre-defined lineage-specific marker genes within each cell, providing a single, interpretable score per lineage per cell that summarizes overall transcriptional momentum. Its conceptual strengths lie in its simplicity, clarity, and suitability for visualizing broad lineage commitment trends and population-level patterns during initial exploratory analysis. However, it has limitations as it assumes uniform behavior and relevance across marker genes, potentially diluting meaningful differences in gene regulation or biological importance. Furthermore, it disregards variability in expression magnitude, statistical confidence, or gene covariance.

2.9.2 Method 2: Adjusting Combined p-Values Using a Null Model

This method introduces statistical control by comparing observed RNA velocity signals to a null distribution, which is often derived from permuted or background data. The process involves converting RNA velocity values to z-scores and then to adjusted p-values by comparing them to the null model's empirical distribution; log-transformed p-values are frequently used for easier interpretation. A key advantage of this approach is its statistical robustness, which enables filtering out background noise to highlight true biological signals. It offers fine-grained resolution at the gene and cell level and is applicable across various gene sets. However, its reliability heavily depends on the integrity and representativeness of the null model. There are also risks of overfitting if the null model is too similar to the query data, and the method can be computationally demanding for large datasets.

2.9.3 Method 3: Combining P-Values with Covariance Adjustment

This method aggregates individual gene p-values into a single p-value for a gene set, a crucial step that accounts for gene interdependencies. It employs two key statistical techniques for Combining P-Values: Fisher's and Brown's Methods. Fisher's Method, which assumes independent p-values, combines them by summing the negative logarithms of the p-values, following a chi-squared distribution, expressed as,

$$X^2 = -2 \sum_{i=1}^{k} \ln(p_i)$$

Brown's Method, on the other hand, is designed for situations where p-values are not independent, such as with co-expressed gene sets. It adjusts for the covariance structure by modifying the degrees of freedom and a scale factor for the chi-squared distribution, making the combined p-value more accurate.

The **Mathematical Justification of Covariance-Adjusted Methods** for Brown's method stems from its ability to correct for inaccuracies in combined p-values when gene expression levels are correlated. This is achieved by adjusting the degrees of freedom and scale factor based on the variance introduced by covariance. The formulas for the scale factor (c) and degrees of freedom (df) are given by:

$$c = \frac{E}{Var}$$

and

$$df = \frac{2E^2}{Var}$$

The analytical process for this method begins with Gene Expression Data Loading and Preprocessing. Raw gene expression data, typically organized with genes as rows and samples as columns, is loaded and then transposed. Each gene's expression across all samples is then normalized (by calculating mean and standard deviation) to ensure comparability. Following this, Empirical Cumulative Distribution Function (ECDF) Transformation is applied to the standardized gene expression values. This transformation generates a statistical measure that quantifies how extreme an observation is relative to the overall distribution of gene expression, which is essential for generating p-values under a null hypothesis.

A Covariance Matrix Calculation is subsequently performed on this ECDF-transformed data. This matrix captures the interdependencies between gene expression levels, playing a critical role in Brown's method by correcting for non-independent p-values. Before the final p-value combination, Thresholding and Filtering are applied. Genes with insufficient or irrelevant expression (those below a specified non-zero expression count threshold) are filtered out to reduce noise and focus on biologically meaningful genes. The process concludes with the generation of Final Combined P-Values and Output. These are the combined p-values for each gene set, reflecting their overall significance, and are typically saved to a CSV file for subsequent downstream analyses, such as pathway enrichment.

2.10 UniPath: Normalization-free Gene-set Enrichment for Single-cell Expression

2.10.1 Conceptual Overview

UniPath is a computational framework designed to assess gene-set and pathway enrichment in single-cell RNA sequencing (scRNA-seq) data without requiring inter-cell normalization [10]. This is particularly important given the challenges posed by technical variability, sequencing depth, and the high dropout rate common in single-cell datasets. UniPath addresses these issues by modeling each cell independently, thereby preserving within-cell statistical structure

and avoiding artefactual variation introduced by global normalization.

2.10.2 Probabilistic Modeling of Expression

The core of UniPath lies in the probabilistic modeling of gene expression within individual cells. Specifically, it models the log-transformed expression values (e.g., TPM, FPKM, RPKM, or UMI counts) using a bimodal distribution, where one mode captures zero expression (either due to biological inactivity or dropout) and the second mode follows a Gaussian distribution describing the distribution of log-transformed non-zero expression values [10].

The probability density function $f(x)$ of log-transformed gene expression x in a single cell is given by:

$$f(x) = p_0 \cdot I(x = 0) + (1 - p_0) \cdot N(x; \mu, \sigma)$$

Here, $I(x = 0)$ is an indicator function for zero expression, p_0 represents the fraction of genes with zero expression, and $N(x; \mu, \sigma)$ denotes the normal distribution of non-zero log-transformed expression values with mean μ and standard deviation σ.

From this model, UniPath computes a right-tailed p-value for each gene with non-zero expression by evaluating its deviation from the fitted Gaussian distribution. These p-values reflect the relative expression strength of each gene within its own cell.

2.10.3 Gene Set Enrichment Using Brown's Method

To determine the significance of a gene set (or pathway) in a given cell, UniPath aggregates the p-values of genes in the set using Brown's method. This method accounts for the dependence between gene-level statistics, which is critical given that many genes in the same pathway are co-regulated and thus correlated [10].

Let P_i denote the right-tailed p-value for gene i (out of k non-zero genes in the set). Define the combined test statistic ψ as:

$$\psi = -2 \sum_{i=1}^{k} \log P_i$$

The combined p-value for the gene set is then computed as:

$$P_{\text{combined}} = 1.0 - \Phi_{2f}\left(\frac{\psi}{c}\right)$$

where Φ_{2f} is the cumulative distribution function of the chi-square distribution with $2f$ degrees of freedom, and c is a scaling factor that adjusts for gene-gene dependencies. The

degrees of freedom f and the scaling factor c are calculated as follows:

$$\mathbb{E}[\psi] = 2k, \quad \mathrm{Var}[\psi] = 4k + 2\sum_{i<j} \mathrm{Cov}(-2\log P_i, -2\log P_j)$$

$$f = \frac{(\mathbb{E}[\psi])^2}{\mathrm{Var}[\psi]}, \quad c = \frac{\mathrm{Var}[\psi]}{2\mathbb{E}[\psi]}$$

This procedure results in a single enrichment p-value for each pathway in each cell, enabling a robust, cell-specific characterization of functional activity.

2.10.4 Normalization and Transformation

To ensure comparability across cells, UniPath applies z-score normalization to the expression matrix, followed by an empirical cumulative distribution function (ECDF) transformation [10]. These steps standardize the data and quantify the extremity of gene expression values relative to the overall distribution, improving the robustness of p-value calculation.

2.10.5 Null Model for Statistical Adjustment

To mitigate false positives and normalize for inherent pathway biases (e.g., from highly expressed housekeeping genes), UniPath constructs a null distribution by generating pseudo-cells [10]. These are derived from randomly paired, transcriptomically distinct cells taken from publicly available datasets, and their expression profiles are averaged to simulate heterogeneity.

Each pseudo-cell is processed identically to real cells, and pathway-level p-values are computed as described above. The adjusted p-value for a pathway in a real cell is then defined as the empirical fraction of pseudo-cells in which the same pathway exhibits a more significant (i.e., lower) enrichment score. This null model acts as a robust reference, ensuring specificity of the reported enrichment.

2.10.6 Practical Implementation and Interpretation

In practice, UniPath requires a minimum of five non-zero genes in a gene set to reliably estimate enrichment. By producing cell-specific pathway p-values, UniPath enables a wide range of downstream analyses including trajectory inference, functional state classification, and integration with RNA velocity or lineage tracing methods.

Its core advantage lies in its normalization-free design, which respects the stochastic nature of single-cell data while delivering high-resolution insights into cellular function [10]. By focusing on within-cell distributions and adapting classical enrichment techniques to a

single-cell framework, UniPath provides an interpretable, statistically grounded approach to understanding complex biological processes at the resolution of individual cells.

2.11 Correlating RNA Velocity with UniPath Pathway Scores

This section examines the relationship between RNA velocity, a measure of transcriptional dynamics [2] and UniPath pathway scores, which summarize gene set activity at the single-cell level [10]. The integration of these two analytical outputs provides insights into how gene-level transcriptional kinetics correspond to broader biological pathway activities, facilitating the study of cellular state transitions and regulatory mechanisms.

2.11.1 Conceptual Basis

RNA velocity estimates the future state of a cell by modeling the ratio of unspliced to spliced mRNA, indicating the direction and speed of gene expression changes. In contrast, UniPath assigns pathway-level scores by evaluating gene-set enrichment using normalized gene expression and p-value aggregation. Correlating these two features enables the identification of pathways that are dynamically regulated during cellular progression.

2.11.2 Statistical Methods for Association Analysis

To evaluate the relationship between RNA velocity and UniPath scores across cells, correlation analysis is employed. Two commonly used methods are:

2.11.2.1 Pearson Correlation

Pearson correlation measures the strength of a linear association between two continuous variables. While effective for normally distributed and linearly related data, it may not perform well in capturing non-linear relationships commonly seen in single-cell transcriptomics [11].

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

2.11.2.2 Spearman Correlation

Spearman correlation is a non-parametric method based on rank-order values. It assesses monotonic relationships without assuming normality or linearity, making it more suitable for high-variance, non-linear single-cell data [11]. Given the distributional complexity of RNA

velocity and pathway scores, Spearman correlation is generally favored in this context.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

2.11.3 Interpretation and Relevance

Given the variability and non-linear structure of single-cell transcriptomic data, Spearman correlation emerges as a robust method for assessing the association between RNA velocity and UniPath pathway scores. Its ability to detect monotonic trends without strict distributional assumptions makes it well-suited for revealing biologically meaningful relationships between gene-level dynamics and pathway activity. This correlation framework supports further exploration of lineage progression, differentiation potential, and regulatory shifts in single-cell systems.

2.12 Integrative Modeling of RNA Velocity and Pathway Activity for Predicting Lineage Poising

Understanding early cell fate decisions requires both dynamic and functional insights at single-cell resolution. This section introduces a two-pronged computational framework that integrates RNA velocity estimates with pathway-level activity from UniPath to predict lineage poising and infer regulatory pathways influencing cell commitment during differentiation.

2.12.1 Estimating Lineage Poising from RNA Velocity

RNA velocity reflects the rate of gene expression change and serves as a proxy for future transcriptional states. For each cell, lineage-specific marker genes are selected, and their RNA velocities are standardized using z-scores to account for gene-level variability. The average of these z-scores across all markers of a lineage defines the cell's poising level toward that lineage. A high, positive score indicates transcriptional bias or early priming toward the lineage, enabling a quantitative view of lineage predisposition before phenotypic commitment occurs.

2.12.2 Method 1: Pathway-Poising Correlation via UniPath

In Method 1, we utilize UniPath to compute pathway enrichment scores for each cell based on its gene expression profile. The result is a matrix of size $P \times M$, where P is the number of known pathways and M is the number of single cells. These pathway scores represent the transcriptional activity of known gene sets.

To identify pathways that influence lineage poising, we compute the Spearman correlation between pathway scores and poising levels for a specific lineage across all cells. Pathways with an absolute correlation above a defined threshold are considered potentially influential. Subsequently, we apply probabilistic graphical models, such as Bayesian networks [12] or Markov blankets [13], to prune indirect associations and retain only those pathways that exhibit direct influence on lineage poising. These direct influencer pathways provide mechanistic insight into how certain signaling axes or regulatory programs drive commitment toward a particular lineage.

This method is particularly effective for identifying global pathway regulators that consistently correlate with commitment across cells and conditions.

2.12.3 Method 2: Resolving Bifurcations via Poising Ratios

When cells exhibit partial poising toward multiple lineages such as in early bifurcation events Method 2 provides resolution. A poising ratio is computed between the two lineages of interest (e.g., endoderm vs. mesoderm). This ratio captures the relative bias within individual cells. Correlation of pathway scores with this ratio highlights pathways that influence fate divergence, either by enhancing one trajectory or repressing the other. This method is particularly suited for analyzing progenitor states where binary fate choices are unresolved.

2.12.4 Application: Embryonic Stem Cell Differentiation

The proposed framework is applied to single-cell transcriptomic data from early embryonic stem cell differentiation. Using RNA velocity and UniPath, poising levels toward the ectoderm, mesoderm, and endoderm are computed. Method 1 identifies pathways such as WNT and NODAL as global enhancers of endodermal bias. Method 2 reveals FOXA2 and SMAD2/3 as bifurcation regulators that favor endoderm while repressing mesoderm. These findings validate the framework's ability to capture early lineage priming and pinpoint signaling determinants.

2.12.5 Future Extensions and Outlook

While correlation-based methods provide initial insights into pathway-lineage associations, they do not infer causality. Future work will involve Bayesian and causal inference models to unravel directional relationships between pathways and lineage poising. This approach will improve understanding of how pathways directly modulate differentiation decisions and identify molecular targets for modulating cell fate in regenerative applications. Ultimately, this integrative strategy offers a scalable framework for decoding dynamic transcriptional programs governing development and disease.

2.13 Modeling Regulatory Dependencies with Bayesian Networks

Understanding cell fate decisions requires more than just identifying correlations; it demands uncovering the direction and structure of regulatory influences between pathways and lineage commitment. Bayesian Networks [12] provide a powerful framework to model these complex, conditional dependencies using single-cell RNA velocity and pathway activity data. By representing pathway activities and lineage poising as interconnected nodes in a probabilistic graph, Bayesian Networks help identify which pathways directly influence lineage bias, distinguishing true regulatory drivers from indirect associations. This approach also accounts for noise and uncertainty inherent in single-cell data, offering a robust way to interpret the regulatory architecture controlling early differentiation. In essence, Bayesian Networks allow us to move beyond correlation toward modeling potential causal relationships, which is crucial for understanding how cells make fate decisions and for identifying key molecular targets to guide differentiation.

2.14 Ranking Pathway Importance Using Random Forests

Understanding the regulatory basis of cell fate commitment requires identifying not only causal relationships but also the predictive strength of each pathway. While Bayesian Networks capture dependencies among pathway scores and lineage poising, they do not rank individual pathways by predictive power. To address this, we used supervised machine learning, specifically Random Forests, to evaluate and rank pathway importance.

We tested several models, including XGBoost and Adaptive Boosting (ABF), but found Random Forests performed better in accuracy, robustness, and generalization. Their ability to handle high-dimensional, noisy biological data and model non-linear relationships made them ideal for our dataset.

The input included pathway enrichment scores (UniPath) and RNA velocity-derived poising. After preprocessing and cross-validation, the Random Forest model was trained to classify lineage outcomes and rank pathways based on their contribution to prediction.

Pathways with high importance scores often aligned with upstream regulators identified in the Bayesian analysis, reinforcing their role in lineage decisions. While Random Forests lack causal directionality, their feature importance outputs provided a complementary layer of interpretation to the Bayesian model.

Overall, this combined approach enabled a refined, predictive understanding of lineage transitions, highlighting pathways most relevant for experimental targeting.

2.15 Correlation of Transcription Factor Expression with Combined RNA Velocity Scores

To identify transcription factors (TFs) potentially involved in lineage regulation, we correlated their expression levels with combined RNA velocity-derived scores that capture overall differentiation dynamics. TF expression data (FPKM values) were aligned with RNA velocity scores across shared single-cell samples, and non-informative TFs with constant expression were excluded.

Spearman correlation was used to measure the association between each TF's expression and the combined RNA velocity scores. Positive or negative correlations reflect potential regulatory roles in directing cell state transitions. This approach provides a data-driven framework to prioritize TFs likely involved in controlling developmental trajectories.

CHAPTER 3

Single-Cell Transcriptomic Analysis of Human Gastrulation

This chapter presents the computational workflow applied to construct a spatially resolved single-cell atlas of human gastrulation using publicly available transcriptomic data. The study employed single-cell RNA sequencing (scRNA-seq) data from a human embryo at embryonic day 16 (accession: E-MTAB-9388), encompassing 1,195 cells. The analytical pipeline integrated several tools and libraries including Scanpy, Velocyto, UniPath, Seurat, and DESeq2 for data processing, clustering, RNA velocity estimation, and trajectory inference.

3.1 Data Acquisition and Preprocessing

The raw data for the study was downloaded from the European Nucleotide Archive (ENA), linked to the E-MTAB-9388 dataset. Metadata provided in the Sample and Data Relationship Format (SDRF) file was parsed to extract cell type annotations based on both ontology and author-curated labels. The transcriptomic profiles were aligned using the STAR aligner [7], and spliced/unspliced count matrices were generated via Velocyto [2] to enable downstream RNA velocity analysis.

An integrated .loom file representing all processed cells was used as input for further analysis in Python using Scanpy [14]. Initial preprocessing included normalization of transcript counts per cell to a fixed total (10,000 counts), followed by logarithmic transformation. High variability genes were identified using the Seurat v3 method, restricting further analysis to the top 4,000 most variable genes. Principal component analysis (PCA) was performed to reduce dimensionality, retaining the top 30 components.

3.2 Clustering and Visualization of Cell States

To identify discrete transcriptional cell states, a K-nearest neighbors (KNN) graph was constructed using the top principal components. The Leiden algorithm [15] was applied for community detection, with the resolution parameter optimized to yield exactly eleven distinct clusters, corresponding to known germ layer derivatives and intermediate stages of early human development. UMAP (Uniform Manifold Approximation and Projection) was employed for visualizing the cellular state space. The resulting embedding was annotated with cluster identities derived from the SDRF metadata. Cluster-specific marker genes were determined using the Wilcoxon rank-sum test [16], and their expression patterns visualized through heatmaps. The annotated UMAP plot served as a comprehensive overview of the

transcriptional landscape during gastrulation.

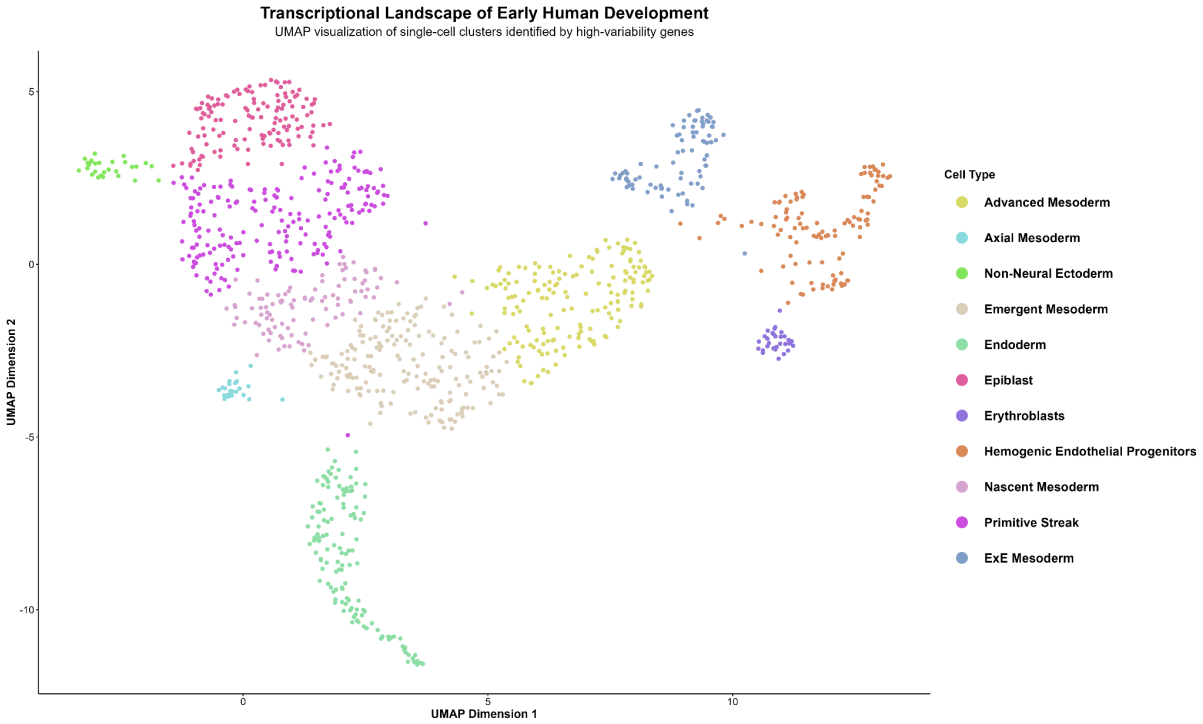


Figure 3.1: UMAP visualization of single-cell clusters representing distinct transcriptional cell states during early human development, annotated with known cell types.

This UMAP visualization illustrates the transcriptional landscape of early human development, showcasing distinct single-cell clusters identified by high-variability genes. The plot, generated using Uniform Manifold Approximation and Projection (UMAP) [17], positions cells in a two-dimensional space based on their transcriptional similarity. Different colors represent eleven distinct cell types, ranging from early developmental stages like Epiblast and Primitive Streak to differentiated germ layer derivatives such as Advanced Mesoderm, Endoderm, and Non-Neural Ectoderm, as well as intermediate and progenitor cell states like Nascent Mesoderm and Hemogenic Endothelial Progenitors. The clear separation and clustering of these cell types indicate discrete transcriptional states, providing a comprehensive overview of the cellular diversity and developmental trajectories during human gastrulation.

3.3 Integration of Metadata and Diffusion Mapping

Cell barcodes were matched with their respective metadata, allowing the integration of cluster annotations such as inferred lineage stages. To examine the developmental progression, cells were reordered based on a manually curated lineage hierarchy ranging from epiblast to differentiated lineages like endoderm and erythroblasts. A diffusion map [18] was constructed using the top diffusion components (DC1 and DC2) to capture the latent trajectories that underlie lineage bifurcations. This method preserved the manifold geometry of the dataset and

allowed visualization of continuous transitions between intermediate cell states. Custom color palettes were applied to reflect developmental identities consistently across plots.

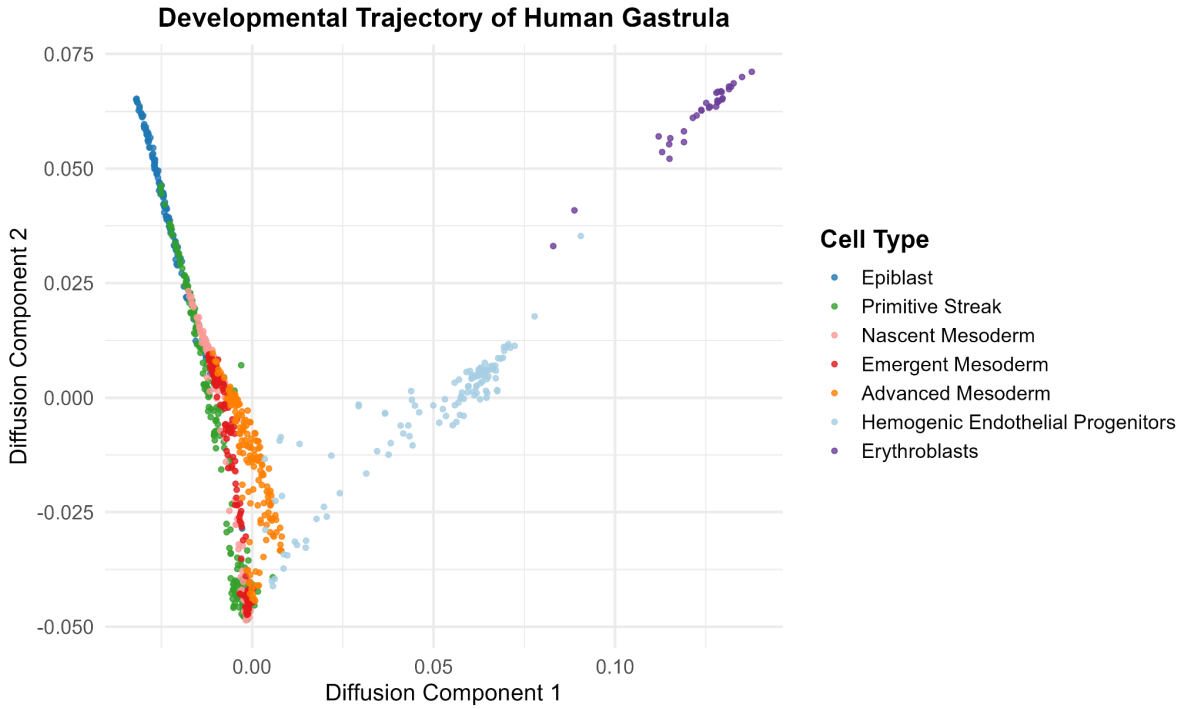


Figure 3.2: Human gastrulation manifold visualized by diffusion map (DC1 and DC2).

This plot presents a diffusion map, a sophisticated non-linear dimensionality reduction technique. It's employed here to visualize the developmental manifold underlying human gastrulation. The axes, labeled Diffusion Component 1 and Diffusion Component 2, correspond to the dominant eigenvectors of a diffusion operator. This operator is constructed from the pairwise similarities between individual cells. These components effectively capture the principal modes of variation and pseudo-temporal progression inherent in the dataset.

3.4 RNA Velocity Estimation and Pseudotemporal Analysis

To estimate RNA velocity, spliced and unspliced transcript counts were extracted from the .loom file. Dynamic modeling was performed using the scVelo framework [2], which captured the directionality of gene expression changes, offering insights into cell fate commitment. The analysis generated velocity vectors that were overlaid on the UMAP embedding, reinforcing the developmental trajectories inferred from clustering. To further quantify transcriptional programs, gene lengths and raw count matrices were extracted from the loom structure, enabling the calculation of FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values. These normalized expression matrices served as input for pathway and gene set enrichment analyses.

3.5 Functional Characterization Using UniPath

The UniPath tool [10] was employed to derive adjusted gene set scores that reflect dynamic regulatory changes during gastrulation. FPKM-normalized single-cell expression data served as the input for gene set scoring. For each cell, UniPath computed p-values using the binorm function by comparing observed expression distributions against null models derived from known human cell type-specific gene signatures. These p-values were then adjusted to account for multiple testing and stochastic background activity, resulting in normalized scores. The final output was stored as gene set score matrices, providing a pseudo-temporal functional landscape of pathway activity across developmental stages. These adjusted scores served as a foundation for downstream trajectory analysis and regulatory inference.

3.6 Stage-Specific DEG Selection and Velocity Integration

To link transcriptional changes with developmental dynamics, differentially expressed genes (DEGs) were identified across major gastrulation stages: Epiblast (Epi), Primitive streak (PS), Nascent mesoderm (NM), Emergent mesoderm (EM), Advanced mesoderm (AM), Hematopoietic progenitors (HEP), and Erythroblasts (Ery). For each stage transition, the DEG list corresponding to the preceding stage was used to extract RNA velocity features from the downstream population, e.g., PS DEGs were used to estimate velocity in Epi cells, NM DEGs in PS cells, and so forth. The extracted stage-specific velocities were then combined using Method 3, as defined in Chapter 2. This approach integrates velocity signals by adjusting combined p-values through a null model framework, followed by aggregation across gene markers weighted by directional consistency. This results in a composite velocity vector for each cell, encapsulating the overall transcriptional push derived from sequential lineage transitions.

This UMAP-based visualization displays the transcriptional landscape of early human development, with individual cells clustered and colored by their identified cell type. Overlaid arrows depict hypothesized lineage trajectories, specifically illustrating a developmental progression starting from the Epiblast, moving through the Primitive Streak, then to Nascent Mesoderm, followed by Emergent Mesoderm, and subsequently Advanced Mesoderm. The trajectories then continue to Hemogenic Endothelial Progenitors, ultimately leading to Erythroblasts, thereby tracing a potential developmental pathway from early pluripotent cells to a specific differentiated blood lineage. This visualization offers insights into the dynamic process of cellular differentiation during human embryogenesis.

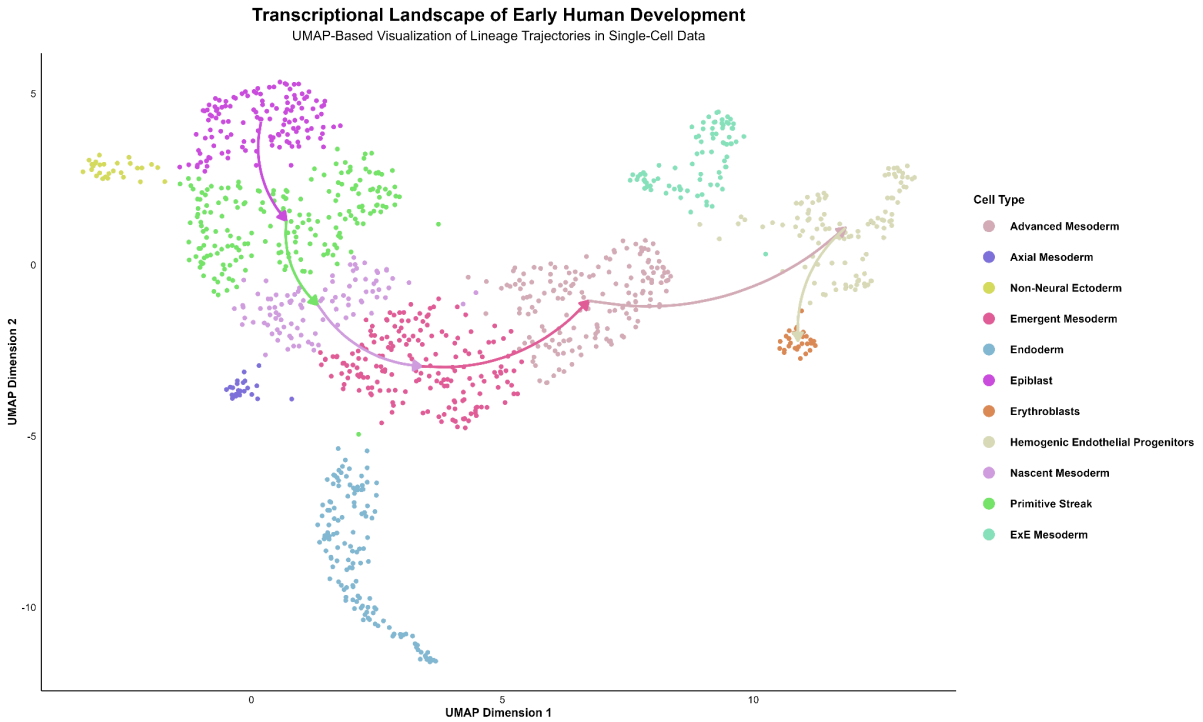


Figure 3.3: UMAP-based visualization of lineage trajectories in single-cell data, depicting hypothesized developmental paths between various cell types during early human development.

Table 3.1: Lineage Types and Associated Genes

Lineage Type	**Associated Genes**
Epiblast to PS	ERRFI1, MCOLN3, LINC01356, SYT6, L1TD1, ROR1, ADGRL2, LINC00467, DTL, CENPF, EFEMP1, BUB1, ID2, VSNL1, UGP2, GAD1, SLCO2A1, PCOLCE2, TDGF1, CD200, LDB2, CENPE, HPGD, SLIT2, LIMCH1, SELENOP, HMGCS1, SERINC5, GFPT2, SHISAL2B, VCAN, ADGRV1, GABRP, MSX2, ZNF204P, DST, SAMD3, TBXT, AKAP12, HPAT5, SEMA3A, PODXL, PTPRZ1, CLU, SFRP1, GRHL2, CORO2A, ABLIM1, CADM1, DDX25, LRRIQ1, LINC00458, SLC7A8, ABHD12B, GABRB3, MNS1, NIPA1, NUSAP1, CRYM, MT1H, CDH1, VAT1L, TOP2A, CDH2, NDC80, RAB31, CRLF1, ZNF723, CYP2S1, SALL4, DNMT3B, PCAT14, PCDH11X

Table 3.1 – Continued from previous page	
Lineage Type	**Associated Genes**
PS to NM	ORC1, LINC01356, CTSK, PLXNA2, SIPA1L2, DNALI1, CDCA8, KIF2C, L1TD1, PGM1, ADGRL2, MPZL1, C1orf112, LINC00467, DTL, CENPF, DISP1, WDR35, EFEMP1, BUB1, CFC1, MCM6, ZEB2, RND3, CCDC141, HJURP, VSNL1, GYPC, GAD1, PCOLCE2, TDGF1, TAGLN3, CD200, MME, LDB2, CENPE, TTC29, SLIT2, LIMCH1, PDGFRA, CDH10, DAB2, SELENOP, HMGCS1, SERINC5, FBN2, GFPT2, SHISAL2B, VCAN, PDLIM4, KIF20A, MSX2, KCNK17, DST, SAMD3, MOXD1, TBXT, RIPPLY2, ENPP1, PPP1R14C, AKAP12, COL1A2, PTPRZ1, CLU, SFRP1, HAS2, SDC2, TTC39B, DDX58, CORO2A, ITGA8, PPP1R3C, ABLIM1, MKI67, HK1, TENM4, CADM1, DDX25, JAM3, EPS8, ACSS3, TMEM132B, LINC00458, GABRB3, ALDH1A2, CA12, NUSAP1, KIF23, MESP2, CDH11, GPT2, MMP2, CDH1, TOP2A, IGFBP4, COLEC12, CDH2, NDC80, RAB31, TNFSF14, CRLF1, ZNF850, DLL3, TPX2, DNMT3B, CDC45, ANOS1, PCDH19, GPC4, MSN
NM to EM	LINC01356, SYT6, RXRG, PLXNA2, FAM89A, DNALI1, L1TD1, PGM1, ROR1, ADGRL2, MPZL1, CR1L, DTL, CENPF, FSHR, CFC1, MCM6, ZEB2, RND3, CCDC141, GYPC, ROBO1, PCOLCE2, P3H2, TDGF1, KDR, MAPK10, GYPB, SPOCK3, SLIT2, RASL11B, PDGFRA, PTPN13, CPE, CDH18, ADAMTS12, DAB2, LHFPL2, FBN2, ANXA6, GFPT2, SHISAL2B, VCAN, DST, CEP85L, SAMD3, KCNQ5, RIPPLY2, ENPP1, ELMO1, BMPER, COL1A2, DYNC1I1, PTPRZ1, DLC1, CLU, ZMAT4, SFRP1, CALB1, HAS2, SDC2, CER1, TTC39B, ITGA8, NEBL, PPP1R3C, ABLIM1, HHEX, TENM4, CADM1, NCAM1, DDX25, GRIN2B, LRIG3, TESC, ACSS3, ALX1, TMEM132B, LINC00458, OTX2, NTRK3, MESP2, CDH11, MMP2, COLEC12, CDH2, ALPK2, PTPRM, RAB31, TNFSF14, CRLF1, PLAUR, DLL3, CD177, ADA, SALL4, ANOS1, MID1, GPC4, MSN

Table 3.1 – Continued from previous page	
Lineage Type	**Associated Genes**
EM to AM	ERRFI1, ECE1, FHL3, MCOLN3, TGFβR3, LINC01356, CREG1, KIFAP3, TNNT2, FAM89A, SIPA1L2, L1TD1, PGM1, ROR1, ADGRL2, DDR2, MPZL1, DCAF6, ATP1B1, PRRX1, RGL1, RGS13, FSHR, CFC1, ZEB2, RND3, COBLL1, LRP2, COL5A2, FN1, COL6A3, ID2, CLIP4, GYPC, NRP2, CYP27A1, ADAMTS9, CCDC80, P3H2, PXK, STX18, PROM1, KDR, MAPK10, UNC5C, PITX2, LRBA, SLIT2, RASL11B, PDGFRA, ODAM, ANXA3, PTPN13, GUCY1A1, CPE, ADAMTS12, DAB2, LHFPL2, SERINC5, HAPLN1, FBN2, PITX1, CD74, ANXA6, GFPT2, ISL1, SHISAL2B, VCAN, PDLIM4, MSX2, BMP5, EPHA7, CEP85L, TMEM200A, AKAP12, SEMA3C, SEMA3A, RELN, PODXL, PTN, BMPER, TPST1, COL1A2, DYNC1I1, CPED1, CALD1, DLC1, ANGPT1, HAS2, PPP3CC, ASH2L, SDC2, BAALC, CTHRC1, FREM1, BNC2, SVEP1, TNC, TEK, GCNT1, ITGA8, NEBL, NRP1, PPP1R3C, ABLIM1, ADAM12, BAMBI, ARID5B, HTRA1, TRIM5, PAMR1, TENM4, CADM1, PARVA, EXT2, NAALAD2, PKP2, LRIG3, DCN, LGR5, URAD, FLT1, EDNRB, COL4A1, SLC7A8, NID2, OTX2, CGNL1, TPM1, FAH, EMP2, CDH11, FENDRR, ATF7IP2, MMP2, VAT1L, PMP22, ENO3, DNAH2, MYL4, COLEC12, LAMA1, CDH2, PSTPIP2, ALPK2, CCBE1, ARHGAP28, RAB31, DOK6, TNFSF14, CPXM1, ADA, FAM210B, COL6A2, APOBEC3G, FAM118A, FBLN1, MID1, MPP1, GYG2, MSN, KLHL4

Table 3.1 – Continued from previous page	
Lineage Type	**Associated Genes**
AM to HEP	GALE, IFI6, PTAFR, CSF3R, INPP5B, FHL3, P3H1, TAL1, VAV3, HENMT1, CTSS, SELP, KIFAP3, PTGS2, PTPN7, AGT, FAM89A, SMIM1, TNFRSF1B, RPS6KA1, TIE1, MPL, GSTM5, IFI16, FCGR2A, FCGR2B, ATF6, RCSD1, RGL1, RGS18, BTG2, CR1L, TRAF3IP3, CNST, ZEB2, MFSD2B, EHD3, RASGRP3, ARHGAP25, DYSF, SULT1C4, GYPC, NOSTRIN, GAD1, NRP2, SP100, INPP5D, USP4, UBA7, CD47, HCLS1, MYLK, SLC9A9, MECOM, KLHL6, BHLHE40, TGFβR2, TCTA, PRKCD, PXK, GYG1, KDR, IGFBP7, CXCL2, HPSE, ABCG2, GYPB, LIMCH1, KIT, ANXA3, TMEM131L, GUCY1A1, GUCY1B1, KLHL2, FYB1, DAB2, HAPLN1, MEF2C, IRF1, ECSCR, CD74, LCP2, SDHA, NPR3, IQGAP2, F2R, DOCK2, RHAG, ARHGAP18, SGK1, GMPR, PKIB, MYB, STXBP5, ICA1, ELMO1, CD36, PRKAR2B, CPED1, TBXAS1, INSIG1, DLC1, MTUS1, ANGPT1, MTSS1, ST3GAL1, TNFRSF10C, SLC25A37, CPQ, BAALC, CTHRC1, HEMGN, ALAD, DOCK8, TEK, OSTF1, SYK, TLR4, GSN, GFI1B, KLF6, PRKCQ, ITGA8, NRP1, LIPA, RGS10, MRC1, APBB1IP, BAMBI, ZEB1, RASSF4, ARID5B, SRGN, HK1, HHEX, PTPRE, RHOG, SLC43A3, CLPB, KCNE3, ARRB1, TRPC6, JAML, TSPAN32, CD44, CD82, SERPING1, ZBTB16, NNMT, FLI1, JAM3, ACRBP, CD69, EPS8, ASB8, ARHGAP9, PTPRB, CORO1C, GIT2, TESC, PDE3A, FAR2, NCKAP1L, GLIPR1, DRAM1, ATP6V0A2, FLT1, ELF1, LCP1, RCBTB2, TBC1D4, RGCC, OSGEP, RABGGTA, EGLN3, CDKL1, PYGL, RHOJ, FUT8, RMDN3, ATP8B4, HERC1, THBS1, FAH, IL16, COTL1, ATF7IP2, PRKCB, ORAI3, CPNE2, CDH5, RIPOR1, IRF8, P2RX1, ASGR2, ITGA2B, PECAM1, ADORA2B, TRPV2, TNFAIP1, IGFBP4, MYL4, ITGB3, MILR1, LINC00674, COLEC12, PSTPIP2, CCDC68, ZADH2, RAB31, LDLRAD4, RAB27B, MOB3A, PLAUR, VAV1, STXBP2, ANKLE1, CYP2S1, FOSB, CPXM1, RNF24, GSS, TGM2, ADA, SLC24A3, RIN2, HCK, CASS4, SAMSN1, NRIP1, SLC37A1, RASL10A, TYMP, TANGO2, APOBEC3G, PARVG, ARHGAP6, EFHC2, SYTL4, BTK, ARHGEF6, MPP1, MSN, TMEM164, SASH3

Table 3.1 – Continued from previous page	
Lineage Type	**Associated Genes**
HEP to Ery	TAL1, ORC1, SPTA1, CREG1, KIFAP3, EIF2D, SLC30A1, SLC30A10, SMIM1, CDCA8, KIF2C, DCAF6, BTG2, CD55, CR1L, DTL, CENPF, MCM6, ZEB2, CERKL, ABCB6, HJURP, MFSD2B, LTBP1, GYPC, PLCD1, CD47, GYG1, STX18, NOA1, GYPB, ANXA3, TMEM131L, HMGCS1, MEF2C, SDHA, IQGAP2, SLC35A4, RHAG, GMPR, PPP1R14C, SNX8, TFR2, RELN, CD36, PRKAR2B, CPED1, INSIG1, ANK1, MTSS1, SLC25A37, ASH2L, MCM4, CA2, CPQ, HEMGN, ALAD, UBAC1, TMOD1, GFI1B, KLF6, RGS10, MKI67, HK1, PDCD4, SLC43A3, DHCR7, CLPB, KCNE3, FDXACB1, TSPAN32, MICAL2, CAT, CD82, HMBS, ART4, CORO1C, TESC, DRAM1, ATP6V0A2, RGCC, OSGEP, METTL3, RMDN3, EPB42, NUSAP1, SLC27A2, BBS4, FAH, NPRL3, TRAP1, CRYM, ATF7IP2, REXO5, KIF22, MT1H, SLC4A1, MYL4, B4GALNT2, COG1, C17orf99, PSTPIP2, NDC80, KLF1, DNAJB1, STXBP2, TGM2, ADA, ELMO2, TPX2, FAM210B, SAMSN1, AGPAT3, CDC45, TANGO2, MCM5, APOBEC3G, FAM118A, ALAS2, SYTL4, BTK, MPP1

3.7 Correlation Analysis and Bayesian Network Inference

To assess the regulatory concordance between functional activity and transcriptional kinetics, Spearman correlation was computed between the UniPath-adjusted gene set scores and the combined RNA velocity profiles. This analysis revealed the degree to which velocity-based dynamics aligned with functionally enriched pathways, providing a quantitative metric for trajectory-pathway coupling. Subsequently, Bayesian network [12] analysis was performed to infer probabilistic dependencies among functionally active gene sets and regulatory modules. Using the bnlearn package in R, a series of bootstrapped Hill-Climbing (HC) runs (n = 5) were executed to construct Bayesian network structures. The resulting models were aggregated using the boot.strength() function to estimate arc strengths and consensus edges. High-confidence edges (strength > 0.5) were retained, and the final regulatory network was visualized using the igraph package. This network elucidated potential causal relationships between gene set activities, revealing underlying logic that governs lineage progression during mesodermal and hematopoietic specification.

CHAPTER 4

Results

4.1 Unraveling Early Lineage Commitment in Human Embryonic Stem Cells

This study investigated molecular pathways governing early lineage commitment of human embryonic stem cells (hESCs) towards ectoderm (ECTO), mesoderm (MESO), definitive endoderm (DE), and intermediate progenitor states: posterior primitive streak (PPS) and anterior primitive streak (APS). Two computational methods, Pathway-Poising Correlation via UniPath (Method 1) and Resolving Bifurcations via Poising Ratios (Method 2), identified pathways influencing these differentiation trajectories, with findings benchmarked against established literature.

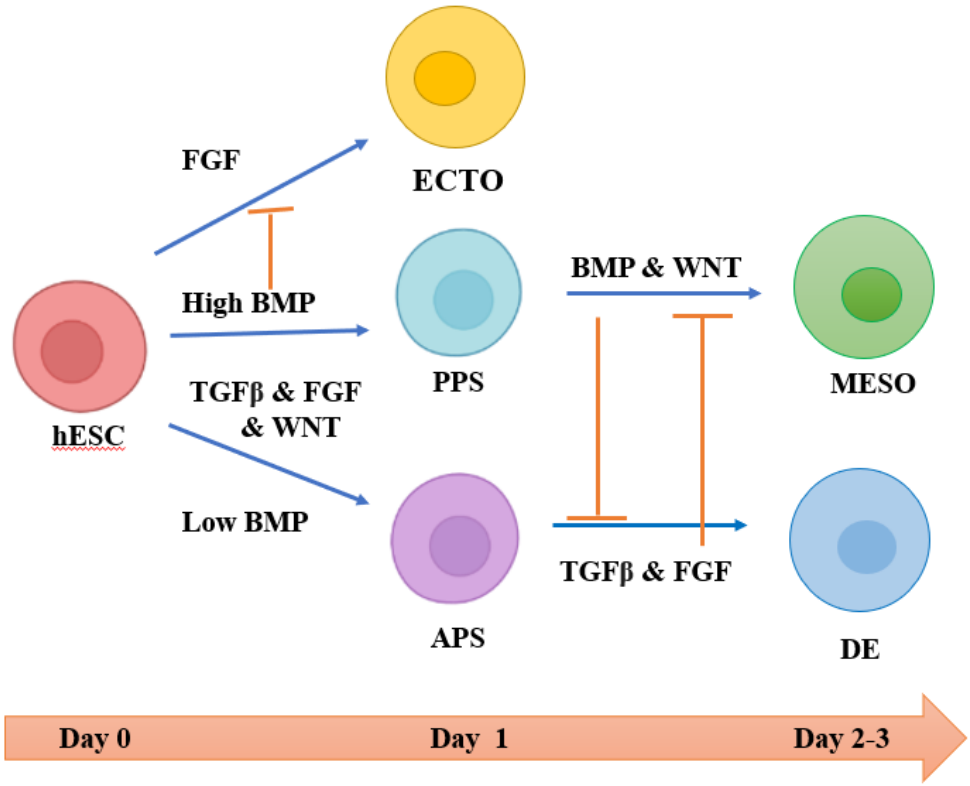


Figure 4.1: Schematic Representation of Early Human Embryonic Stem Cell Differentiation and Key Cytokine Influence.

4.1.1 Overview of hESC Differentiation Trajectories and Key Signaling Pathways

As shown in Figure 4.1 [19] [20], hESCs differentiate over 2-3 days into distinct early lineages. High Bone Morphogenetic Protein (BMP) signaling promotes PPS differentiation, while low BMP combined with Transforming Growth Factor β (TGFβ), Fibroblast Growth Factor (FGF), and Wnt signaling drives APS. Ectodermal commitment (ECTO) [20] is influenced by FGF and inhibited by high BMP. Further differentiation from PPS to MESO [19] is modulated by BMP and Wnt, with inhibition by TGFβ and FGF. Conversely, APS differentiation to DE [20] is influenced by TGFβ and FGF, with inhibition by BMP and Wnt. These interactions highlight the critical role of precise cytokine and growth factor gradients in guiding initial lineage decisions.

This diagram illustrates hESC differentiation into early intermediate lineages (PPS: Posterior Primitive Streak, APS: Anterior Primitive Streak) and subsequent germ layers (ECTO: Ectoderm, MESO: Mesoderm, DE: Definitive Endoderm) over approximately 2-3 days. Key signaling molecules promoting (blue arrows) or inhibiting (orange blunt lines) specific lineage commitments are indicated.

4.1.2 Identification of Lineage-Associated Pathways via Pathway-Poising Correlation (Method 1)

Method 1, using Pathway-Poising Correlation via UniPath, identified pathways whose activity correlated directly with hESC poising levels towards ECTO, PPS, APS, MESO, and DE. Spearman correlation coefficients between pathway scores and differentiation status were computed, and Bayesian networks filtered indirect associations.

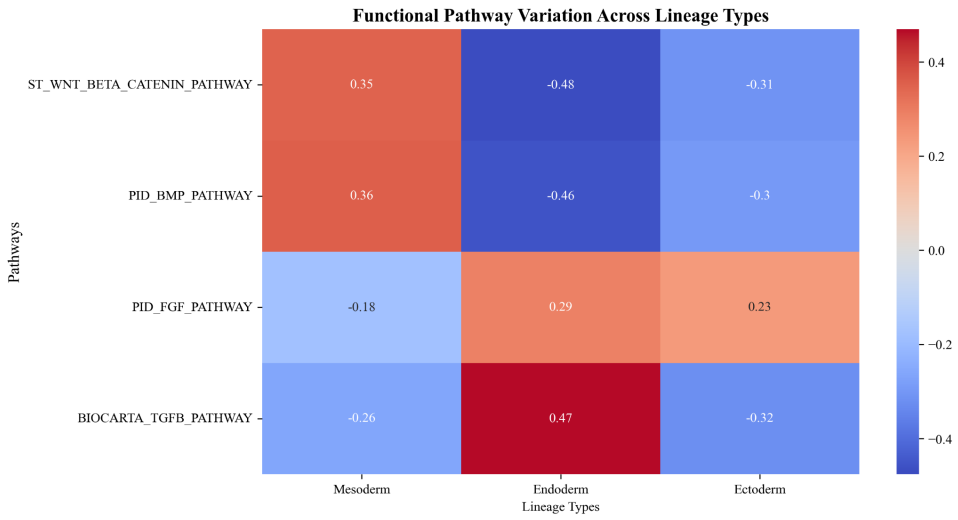


Figure 4.2: Pathway activity correlations with lineage poising using Method 1 (2.12.2).

Using Method 1 as described in section 2.12.2, we computed pathway activity scores and

correlated them with lineage-specific poising levels. The heatmap reveals that WNT and BMP pathways are strong promoters of mesodermal fate, showing positive correlations with mesodermal poising and negative correlations with endodermal and ectodermal poising. In contrast, TGFβ signaling is predominantly associated with endodermal commitment, showing a positive correlation with endodermal poising and negative correlations with mesodermal and ectodermal poising. The FGF pathway displays a dual role, with positive correlations to both endodermal and ectodermal poising, while being inversely correlated with mesodermal bias [19]. These findings support a model where distinct signaling cascades directly influence early lineage bifurcations, highlighting the power of Method 1 in delineating regulatory axes that govern stem cell fate decisions.

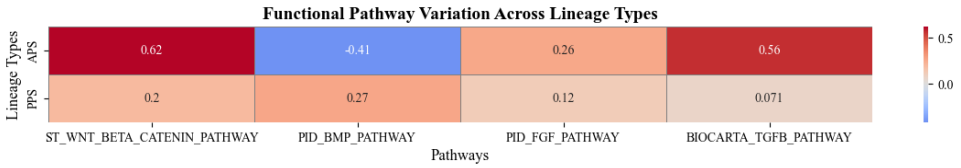


Figure 4.3: Functional Pathway Variation and Correlation with APS and PPS Lineages.

This heatmap displays the Spearman correlation coefficients between the activity of selected key signaling pathways and the poising levels for the APS and PPS intermediate lineages, as identified by Method 1. Red shades indicate positive correlation, while blue shades indicate negative correlation. The intensity of the color corresponds to the strength of the correlation, with values ranging from -0.5 to 0.5 as per the color bar. Values within each cell represent the computed correlation coefficient.

4.1.3 Resolving Early Bifurcations and Fate Divergence via Poising Ratios (Method 2)

Method 2, focused on Resolving Bifurcations via Poising Ratios, provided enhanced resolution for understanding early fate divergence in cells exhibiting partial poising. A poising ratio between two lineages (e.g., endoderm vs. mesoderm) quantified relative bias. Correlation of pathway scores with these ratios elucidated pathways influencing fate divergence.

The heatmap derived from Method 2, as described in section 2.12.3 (Poising Ratios), illustrates how early progenitor states resolve lineage choices through distinct signaling inputs. The WNT pathway exhibits a strong positive correlation with mesodermal poising, while showing negative correlations with ectodermal and endodermal poising, indicating its role in promoting mesodermal commitment [19] and repressing alternative lineages. Similarly, the BMP pathway also shows positive correlation with mesodermal poising and negative correlations with ectodermal and endodermal poising, reinforcing its involvement in mesodermal lineage specification. In contrast, the FGF pathway displays negative correlation

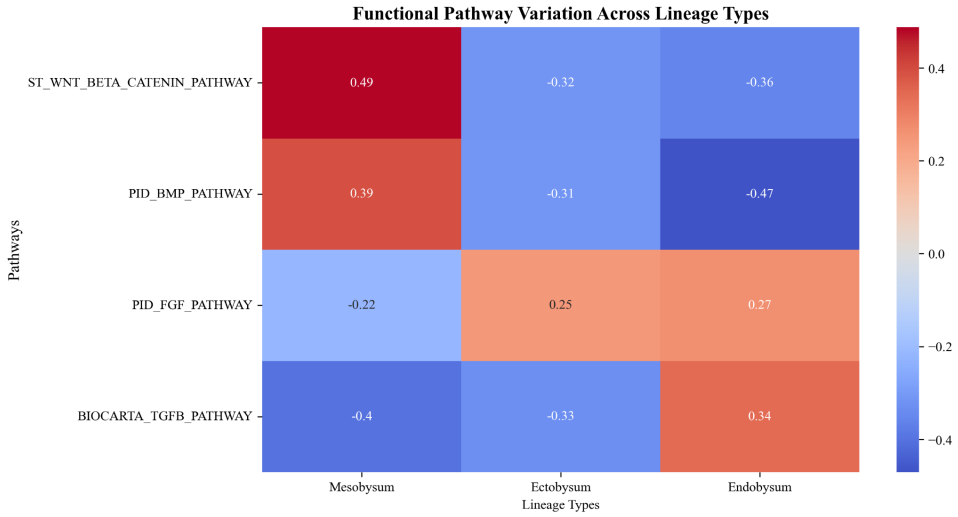


Figure 4.4: Poising ratio correlations with lineage poising using Method 2 (2.12.3).

with mesodermal poising but positive correlations with ectodermal and endodermal poising [20], suggesting a supportive role in non-mesodermal lineage fates. Notably, the TGFβ signaling pathway has negative correlations with mesodermal and ectodermal poising but a positive correlation with endodermal poising [20], indicating a more complex role that may include repression of certain lineage commitments or maintenance of specific early progenitor states. These findings underscore the utility of poising ratios in capturing pathway-specific regulation during early fate bifurcations.

4.1.4 Benchmarking and Consistency of Findings

	BMP	FGF	WNT	TGFβ
APS	-	+	+	+
PPS	+	+	+	+
MESO	+	-	+	-
DE	-	+	-	+
ECTO	-	+	-	-

Table 4.1: Summary of Key Signaling Pathway Regulation in Early hESC Lineage Differentiation.

Key significant pathways identified by both Method 1 and Method 2 showed substantial overlap and consistency, strengthening the robustness of our findings. Critically, these identified pathways are extensively supported by existing literature on hESC differentiation and germ layer formation. The consistent emergence of Wnt, BMP, TGFβ, and FGF signaling pathways as central regulators, with roles in promoting or restricting specific lineages (e.g., Wnt and BMP in mesoderm induction,TGFβ and FGF in endoderm specification, and FGF in ectoderm), directly aligns with current developmental biology understanding [19] [20]. This strong literature support reinforces the biological relevance and accuracy of the identified pathways as direct influencers of hESC lineage commitment.

This table summarizes the observed regulatory effects of Bone Morphogenetic Protein (BMP), Fibroblast Growth Factor (FGF), Wnt, and Transforming Growth Factor β (TGFβ) signaling pathways on the differentiation towards various hESC lineages. A '+' indicates an upregulated pathway, while a '-' indicates a downregulated pathway, correlating with the specified lineage commitment.

4.2 Post-Definitive Endoderm Differentiation and Organ Specification

Beyond the initial germ layer specification, this study also explored the molecular cues driving further differentiation from the Definitive Endoderm (DE) into specialized endodermal progenitor lineages and ultimately organ-specific cell types. As illustrated in Figure 4.3, the DE, typically established by Day 3 of differentiation, serves as a crucial progenitor for both foregut (PFG) and mid-hindgut (MHG) endoderm [20].

The differentiation of DE towards PFG is promoted by the presence of Retinoic Acid (RA), while the activity of RA is indicated to inhibit the formation of MHG. Conversely, the transition from DE to MHG is supported by the combined action of FGF, BMP, and Wnt signaling pathways. From the PFG stage, lineage divergence leads to the specification of liver and pancreas. Liver differentiation from PFG is promoted by MAPK and BMP signaling, while TGFβ signaling acts as an inhibitor. Similarly, pancreas differentiation from PFG is promoted by TGFβ signaling, with MAPK and Hedgehog signaling playing inhibitory roles. These intricate regulatory networks, involving a balance of promoting and inhibitory signals, are critical for directing DE progenitors towards their specific organ fates over time, generally reaching mature stages by Day 7 [20].

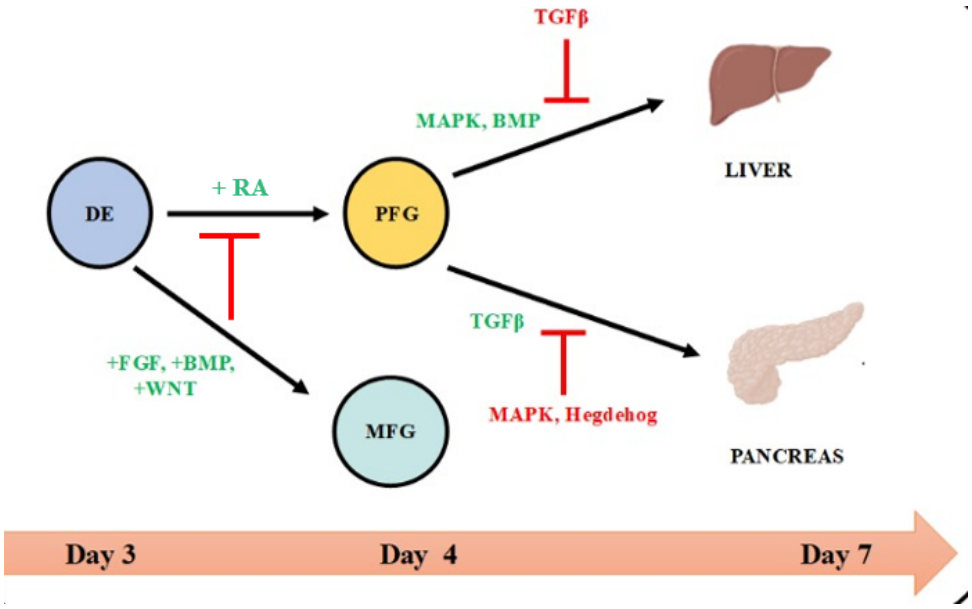


Figure 4.5: Schematic of Definitive Endoderm Differentiation into Foregut/Mid-Hindgut Progenitors and Organ-Specific Lineages.

This diagram illustrates the sequential differentiation of Definitive Endoderm (DE) into specialized progenitor populations: Posterior Foregut (PFG) and Mid-Hindgut (MHG) endoderm [20]. Subsequent differentiation from PFG towards Liver and Pancreas is also depicted.

The timeline indicates key developmental stages (Day 3 to Day 7). Blue arrows represent promoting signals (e.g., +RA, +FGF, +BMP, +WNT, MAPK, BMP, TGFβ), while red blunt lines indicate inhibitory signals (e.g., T for RA, MAPK, Hedgehog, TGFβ).

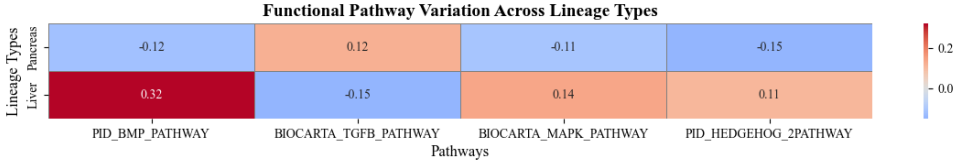


Figure 4.6: Functional Pathway Variation and Correlation with Liver and Pancreas Lineages (Method 1).

This heatmap illustrates the Pearson correlation coefficients between the activity of key signaling pathways and the poising levels for Liver and Pancreas lineages, derived from definitive endoderm, as identified by Method 1. Red shades denote positive correlation, while blue shades indicate negative correlation. The intensity of the color reflects the strength of the correlation, with coefficients provided within each cell.

This table summarizes the observed regulatory effects of Bone Morphogenetic Protein (BMP), Transforming Growth Factor β (TGFβ), Mitogen-Activated Protein Kinase (MAPK), and Hedgehog (Hh) signaling pathways on the differentiation towards Pancreas and Liver lineages

	BMP	TGFβ	MAPK	Hb
PANCREAS	-	+	-	-
LIVER	+	-	+	+

Table 4.2: Summary of Key Signaling Pathway Regulation in Pancreas and Liver Differentiation.

from foregut progenitors [20]. A '+' indicates a pathway with a promoting or upregulated role, while a '-' indicates an inhibitory or downregulated role for the respective lineage commitment.

4.2.1 Benchmarking and Consistency of Findings

Furthermore, our findings regarding the differentiation from Definitive Endoderm (DE) to specialized foregut and mid-hindgut progenitors, and subsequently to liver and pancreas, are highly consistent with established developmental principles. The observed roles of Retinoic Acid (RA) in anterior-posterior patterning of the endoderm, favoring foregut development, are well-documented in developmental biology. Similarly, the pro-liver inductive roles of MAPK and BMP signaling, and the pro-pancreatic roles of TGFβ, with inhibitory effects of MAPK and Hedgehog signaling on pancreatic development, align precisely with the mechanisms described for endodermal organogenesis from pluripotent stem cells. For example, studies by Wells and Melton (2009) [21] have extensively characterized the precise interplay of these signaling pathways, including the roles of FGF, BMP, Wnt, TGFβ, and RA, in patterning the definitive endoderm and specifying organ lineages like liver and pancreas. This strong literature support reinforces the biological relevance and accuracy of the identified pathways as direct influencers of hESC lineage commitment and subsequent organ specification.

4.3 Mesenchymal Cell Differentiation

Beyond germ layer formation and endodermal organogenesis, our analysis also explored the differentiation of mesenchymal cells into their distinct mesodermal derivatives: adipocytes (fat cells), osteocytes (bone cells), and chondrocytes (cartilage cells). As depicted in Figure 4.7, this differentiation is tightly regulated by a complex interplay of promoting and inhibitory signaling pathways [22].

Adipocyte differentiation is positively influenced by factors such as Platelet-Derived Growth Factor (PDGF), Fibroblast Growth Factor (FGF), Peroxisome Proliferator-Activated Receptor gamma (PPARA), and Sonic Hedgehog (SHH). In contrast, Wnt, Bone Morphogenetic Protein (BMP), and Transforming Growth Factor β (TGFβ) signaling pathways exert inhibitory effects on adipogenesis [23].

For osteocyte differentiation, Wnt, BMP, PDGF, and FGF signaling pathways play crucial

promoting roles. However, TGFβ, PPARA, and SHH pathways act as inhibitory influences on osteogenesis, potentially diverting differentiation toward alternative fates.

Cartilage differentiation is actively promoted by BMP, PDGF, FGF, and TGFβ signaling pathways, highlighting their significant contributions to cartilage formation [22].

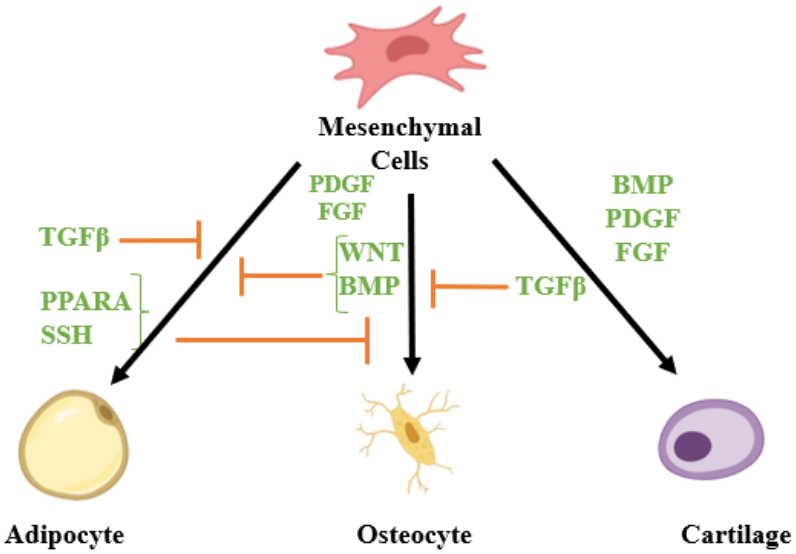


Figure 4.7: Signaling Pathway Regulation of Mesenchymal Cell Differentiation.

This diagram illustrates the differentiation of Mesenchymal Cells into Adipocytes, Osteocytes, and Cartilage.

This heatmap displays the Pearson correlation coefficients between the activity of selected key signaling pathways and the poising levels for Adipocyte, Cartilage, and Osteocyte lineages, as identified by Method 1. Red shades indicate positive correlation, while blue shades indicate negative correlation, with the intensity of the color corresponding to the strength of the correlation as per the color bar.

	BMP	WNT	TGFβ
ADIPOCYTE	-	-	+
OSTEOCYTE	+	+	-
CARTILAGE	+	+	+

Table 4.3: Summary of Key Signaling Pathway Regulation in Mesenchymal Differentiation

This table succinctly summarizes the impact of key signaling pathways on the differentiation of three distinct cell types. A '+' indicates a pathway with a promoting or upregulated role, while a '-' indicates an inhibitory or downregulated role for the respective lineage commitment.

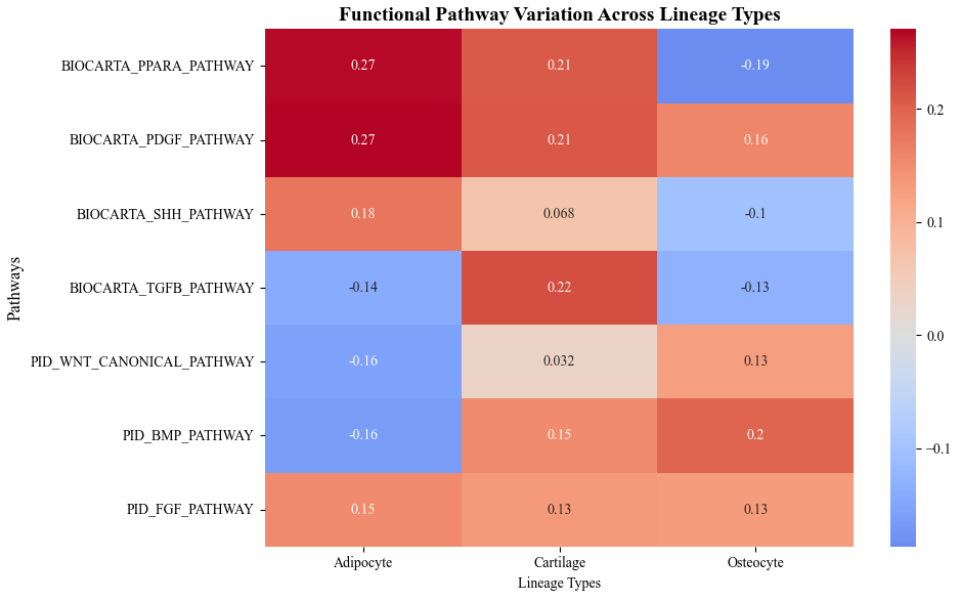


Figure 4.8: Functional Pathway Variation and Correlation Across Adipocyte, Cartilage, and Osteocyte Lineages.

Adipocyte differentiation is uniquely promoted by TGFβ signaling, while being inhibited by both BMP and WNT pathways. Conversely, osteocyte formation is strongly supported by BMP and WNT signaling, but negatively regulated by TGFβ. Cartilage differentiation stands out as it is positively influenced by all three pathways: BMP, WNT, and TGFβ. Overall, the table highlights the intricate and often contrasting roles of these major signaling cascades in directing cell fate within mesenchymal lineages.

4.3.1 Benchmarking and Consistency of Findings

The regulatory mechanisms identified for mesenchymal cell differentiation into adipocytes, osteocytes, and cartilage (Figure 4.8, Table 4.3) are highly consistent with established literature on mesodermal differentiation. Key roles for PDGF, FGF, Wnt, and BMP in promoting osteogenesis and chondrogenesis were observed, while TGFβ showed inhibitory effects in certain contexts. Our finding that Wnt signaling suppresses adipogenesis, but promotes osteogenesis, is strongly corroborated by de Winter and Nusse (2021) [23]. Furthermore, the involvement of PPARA and SHH in adipogenesis aligns with their known roles, and the broader influences of BMP and Wnt on these mesenchymal fates are supported by comprehensive reviews [22]. This consistency reinforces the biological accuracy of our identified pathways as direct influencers of mesenchymal lineage commitment.

4.4 Uncovering Cell Fate Regulatory Networks through Integrated Probabilistic Modeling

Our understanding of cell fate decisions necessitates moving beyond simple correlations to uncover the underlying regulatory architecture. To achieve this, we developed an integrated analytical pipeline that combines robust quantification of pathway activity and lineage poising with advanced probabilistic network modeling. Our workflow involved an initial phase of data processing and preliminary association identification using Method 1: Pathway-Poising Correlation via UniPath (Section 2.12.2) and Method 2: Resolving Bifurcations via Poising Ratios (Section 2.12.3). The refined data and insights generated from these steps then served as the foundational input for constructing comprehensive Bayesian Networks (Section 2.13), from which key regulatory modules, such as Markov Blankets [13], were extracted.

4.4.1 HESC

Building upon this integrated methodological framework, we next applied our analytical pipeline to investigate the regulatory drivers underlying the differentiation of human embryonic stem cells (hESCs) towards the three primary germ layers. In the subsequent subsections, we present detailed insights into the regulatory architecture governing Endodermal, Ectodermal, and Mesodermal commitment, respectively.

4.4.1.1 Regulatory Architecture of Endodermal Commitment

To precisely delineate the multi-faceted regulatory landscape governing endodermal lineage commitment, we utilized the comprehensive Bayesian Network constructed from our processed data, which integrated insights derived from both Method 1's pathway-poising correlations and Method 2's bifurcation-specific associations. This probabilistic graphical model allowed us to infer complex, directed influences between pathway activities and lineage poising, distinguishing true regulatory drivers from indirect associations for specific lineage commitment events like endoderm differentiation.

Insights from Method 1-Informed Network

The Bayesian Network analysis, particularly its ability to delineate the Markov Blanket for Endoderm, provided a mechanistic refinement beyond broad pathway-poising correlations. This approach unveiled a distinct set of immediate, direct regulatory influences (blue nodes) on endodermal lineage commitment. Key among these were pathways associated with cell cycle control (PID_PLK1_PATHWAY, PID_AURORA_A_PATHWAY, PID_AURORA_B_PATHWAY), underscoring the necessity of specific proliferative states for proper differentiation. Furthermore, pathways governing cell adhesion and cytoskeletal

regulation (PID_INTEGRIN2_PATHWAY, BIOCARTA_RANMS_PATHWAY) highlighted the critical role of physical cell-matrix interactions and cellular morphology in fate specification.

Crucially, the Markov Blanket revealed direct regulation by PID_TOLL_ENDOGENOUS_PATHWAY and, significantly, PID_HNF3B_PATHWAY [24]. Given your clarification, this latter pathway likely reflects the pivotal role of HNF3B (FOXA2) or its direct upstream regulators/effectors. HNF3B is a well-established master regulator of endoderm development [24], and its direct inclusion in the Markov Blanket underscores its immediate control over endodermal fate.

Beyond these direct dependencies, the network elegantly captured vital indirect regulatory axes. The observed connection of WNT_SIGNALING to Endoderm via PID_HNF3B_PATHWAY (HNF3B) [24] is particularly insightful and highly validated by developmental literature. This illustrates that while canonical signaling pathways like Wnt might operate upstream, their influence on endodermal commitment is critically mediated by central transcription factors such as HNF3B, which then directly orchestrate gene expression programs. This multi-layered regulatory architecture, distinguishing between direct effectors and key indirect modulators, significantly deepens our mechanistic understanding of endoderm specification and provides targeted insights for guiding differentiation.

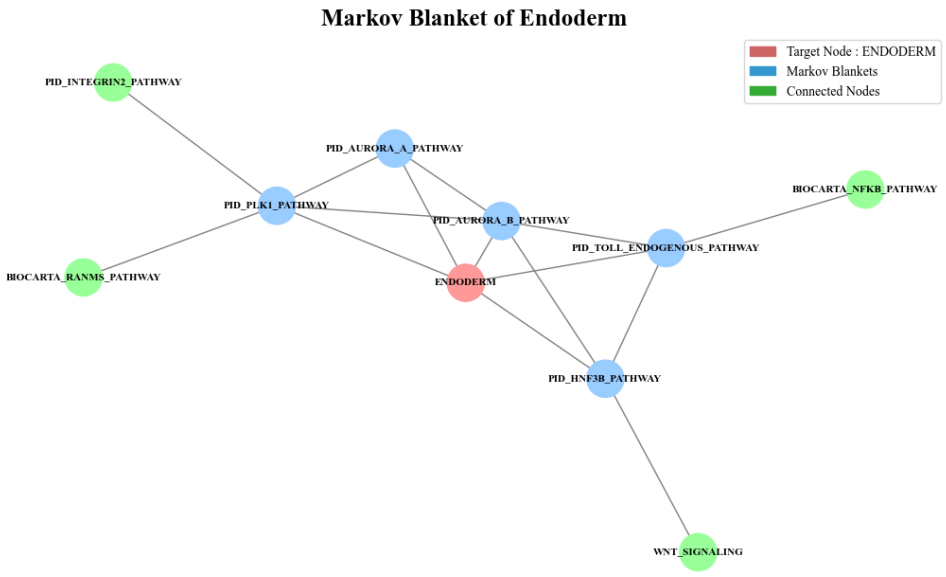


Figure 4.9: Markov Blanket of Endoderm - Method 1

Insights from Method 2-Informed Network

To resolve the intricate regulatory landscape, especially in contexts involving early bifurcation events and subtle shifts in relative lineage biases, the Bayesian Network analysis, utilizing data processed via Method 2's poising ratios, yielded complementary and distinct insights. As depicted in the Markov Blanket for Endoderm, the central "ENDODERM" node is directly

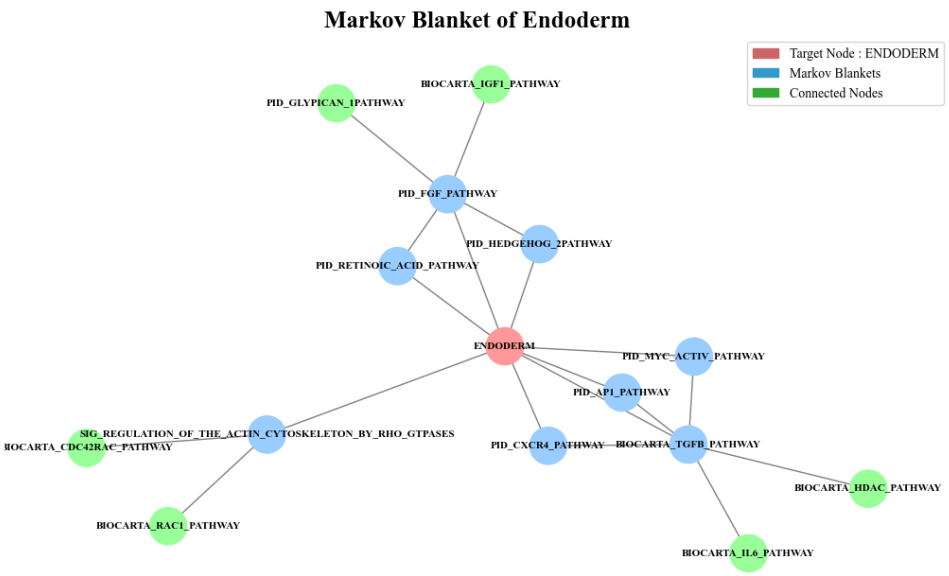


Figure 4.10: Markov Blanket of Endoderm - Method 2

influenced by a set of fundamental pathways. Prominent among these direct regulators are canonical developmental signaling cascades such as PID_FGF_PATHWAY, PID_HEDGEHOG_2PATHWAY, PID_RETINOIC_ACID_PATHWAY, and BIOCARTA_TGFB_PATHWAY, underscoring their immediate roles in initiating and sustaining endodermal lineage specification [20]. Complementing these are pathways reflecting critical cellular functions, including PID_MYC_ACTIV_PATHWAY and PID_API_PATHWAY (involved in growth and transcriptional control), BIOCARTA_CXCR4_PATHWAY (implying migratory or positional cues), and SIG_REGULATION_OF_THE_ACTIV_CYTOSKELETON_BY_RHO_GTPASES (highlighting morphogenetic processes). These findings represent a dynamic set of drivers crucial for establishing endodermal fate, captured by their robust associations with overall endodermal poising and their direct connectivity within the inferred network.

4.4.1.2 Regulatory Architecture of Ectodermal Commitment

Building on our ectodermal commitment strategy, we applied an integrated analytical pipeline to dissect the regulatory framework guiding ectoderm lineage specification. By leveraging pathway-poising correlations from Method 1 and bifurcation-specific associations from Method 2, we provided informative inputs to our overarching Bayesian Network model. This enabled us to infer conditional dependencies and pinpoint key regulatory drivers directly influencing the Ectoderm lineage.

Insights from Method 1-Informed Network

The Bayesian Network analysis of the Ectoderm Markov Blanket reveals direct regulatory influences from pathways critical for cell morphology, adhesion, and growth, including

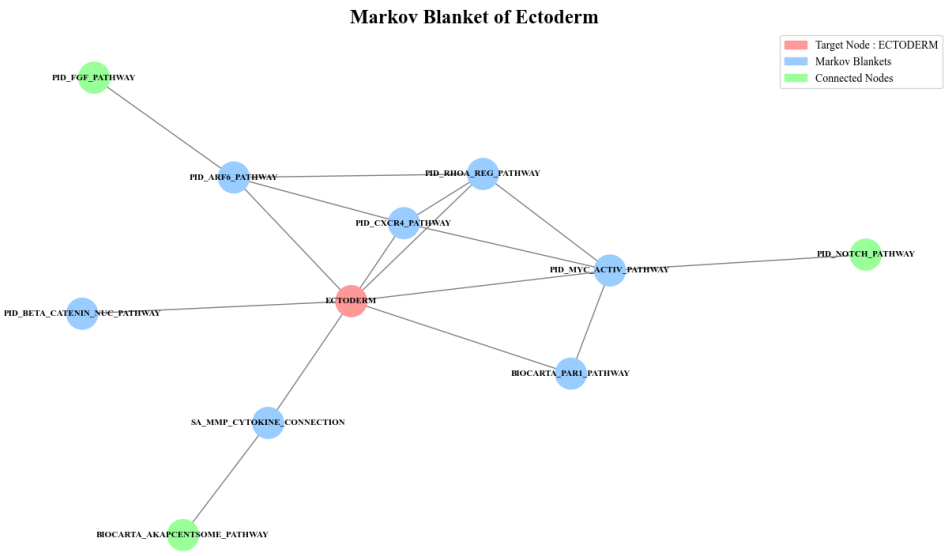


Figure 4.11: Markov Blanket of Ectoderm - Method 1

PID_ARFB_PATHWAY, PID_RHOA_REG_PATHWAY, PID_CXCR4_PATHWAY, PID_MYC_ACTIV_PATHWAY, BIOCARTA_PAR1_PATHWAY, and significantly, PID_BETA_CATENIN_NUC_PATHWAY. This direct inclusion of nuclear Beta-Catenin highlights its immediate role in orchestrating ectodermal gene expression. Additionally, the network differentiates these direct effectors from vital indirect modulators, such as PID_FGF_PATHWAY and PID_NOTCH_PATHWAY [20], whose influence on ectodermal fate is mediated via their connections to pathways within the Markov Blanket, specifically PID_ARFB_PATHWAY and PID_MYC_ACTIV_PATHWAY, respectively, thereby providing a multi-layered mechanistic understanding of ectoderm specification.

Insights from Method 2-Informed Network

The Bayesian Network analysis of the Ectoderm Markov Blanket reveals a distinct set of immediate, direct regulatory influences (blue nodes) on ectodermal lineage commitment. Key among these are canonical developmental signaling cascades such as PID_BMP_PATHWAY [20], ST_WNT_BETA_CATENIN_PATHWAY, NOTCH_SIGNALING [25], and PID_FGF_PATHWAY, alongside pathways involved in cell cycle control and morphology like SA_REG_CASCADE_OF_CYCLIN_EXPR and PID_RHOA_REG_PATHWAY. This underscores their pivotal, direct control over ectodermal fate. Crucially, the network elegantly captures vital indirect regulatory axes; for example, pathways such as PID_PI3KCI_AKT_PATHWAY and PID_P38_ALPHA_BETA_PATHWAY influence Ectoderm indirectly via their connections to pathways within the Markov Blanket, such as PID_BMP_PATHWAY [20] and PID_RHOA_REG_PATHWAY. This multi-layered regulatory architecture, distinguishing between direct effectors and key indirect modulators, significantly deepens our mechanistic understanding of ectoderm specification and provides targeted

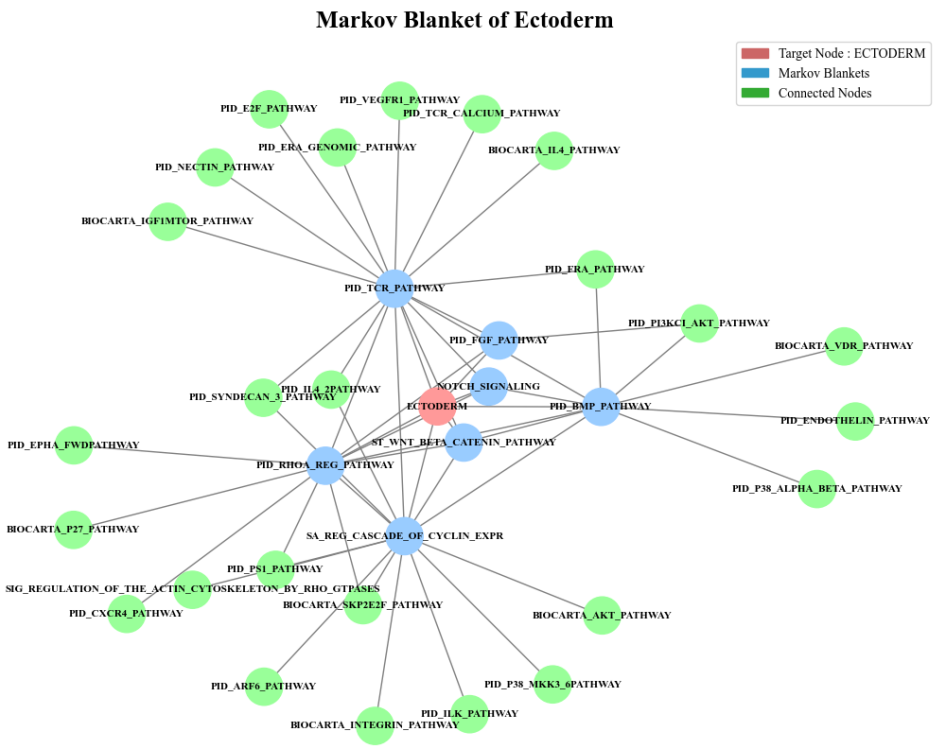


Figure 4.12: Markov Blanket of Ectoderm - Method 2

insights for guiding differentiation.

4.4.1.3 Regulatory Architecture of Mesodermal Commitment

Extending our analytical framework to the mesodermal lineage, we similarly leveraged our integrated pipeline to delineate its complex regulatory architecture. The overarching Bayesian Network was constructed by integrating data processed through Method 1's pathway-poising correlations and Method 2's bifurcation-specific associations. This probabilistic framework allowed us to infer conditional dependencies and identify direct regulatory drivers for the MESODERM lineage.

Insights from Method 1-Informed Network

The Bayesian Network analysis of the Mesoderm Markov Blanket reveals that mesodermal lineage commitment is orchestrated by a tightly integrated set of direct and indirect regulatory influences. At the core, direct effectors such as PID_NODAL_PATHWAY, PID_BMP_PATHWAY, PID_FGF_PATHWAY, and ST_WNT_BETA_CATENIN_PATHWAY represent the canonical signaling axes universally recognized as essential for mesoderm induction and patterning. Additional direct pathways, including PID_HEDGEHOG_2PATHWAY, PID_IGF1_PATHWAY, BIOCARTA_KERATINOCYTE_PATHWAY, and BIOCARTA_ECM_PATHWAY, reflect the importance of cellular growth, extracellular matrix interactions, and tissue morphogenesis in

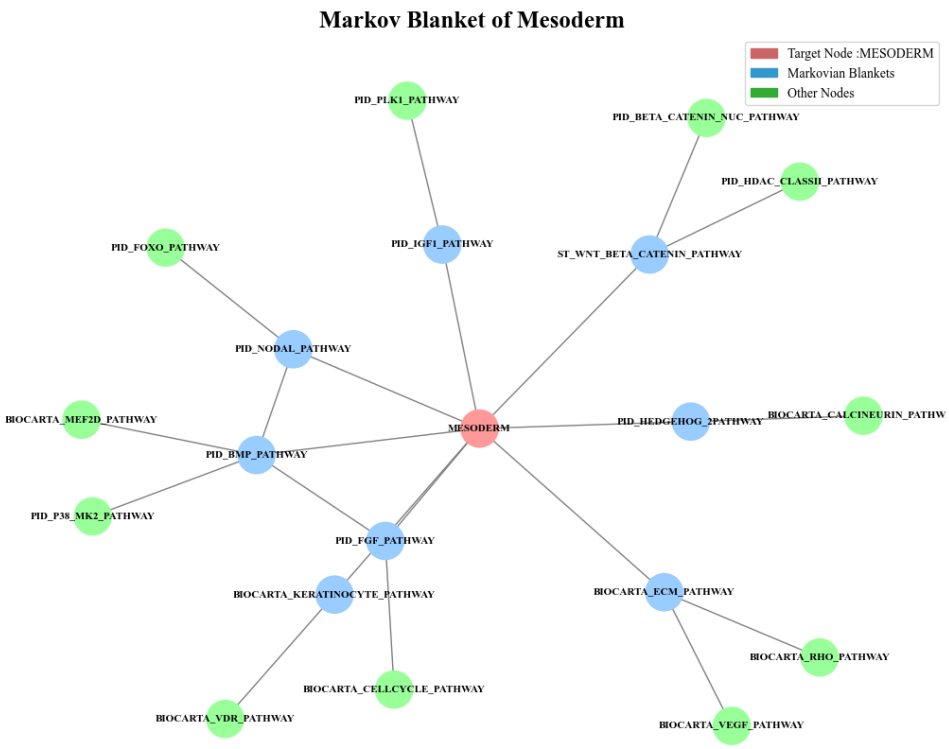


Figure 4.13: Markov Blanket of Mesoderm - Method 1

establishing mesodermal fate.

Crucially, the network also highlights a suite of indirect regulatory relationships that are increasingly recognized as vital for robust mesoderm development. For example, PID_FOXO_PATHWAY modulates mesoderm specification via its interaction with the Nodal axis, while BIOCARTA_MEF2D_PATHWAY [26] and PID_P38_MK2_PATHWAY act through BMP signaling to influence downstream transcriptional programs and cellular differentiation. BIOCARTA_CELLCYCLE_PATHWAY and PID_PLK1_PATHWAY support mesodermal proliferation and maturation indirectly through FGF [19] and IGF1 signaling, respectively. The Wnt/β-catenin module is further refined through direct connections to PID_BETA_CATENIN_NUC_PATHWAY and PID_HDAC_CLASSII_PATHWAY, reflecting the need for coordinated transcriptional and epigenetic regulation. Similarly, BIOCARTA_CALCINEURIN_PATHWAY exerts its effects via the Hedgehog pathway, and both BIOCARTA_VEGF_PATHWAY and BIOCARTA_RHO_PATHWAY shape mesodermal tissue organization and vascular development through the extracellular matrix.

Insights from Method 2-Informed Network

This Mesoderm Markov Blanket, visualized via a refined probabilistic network, reveals a precise hierarchy of regulatory cues driving mesoderm specification. The central node, MESODERM, is surrounded by key direct regulators representing pivotal developmental signals and cellular processes. Prominent among these are Wnt/Beta-Catenin, BMP, and

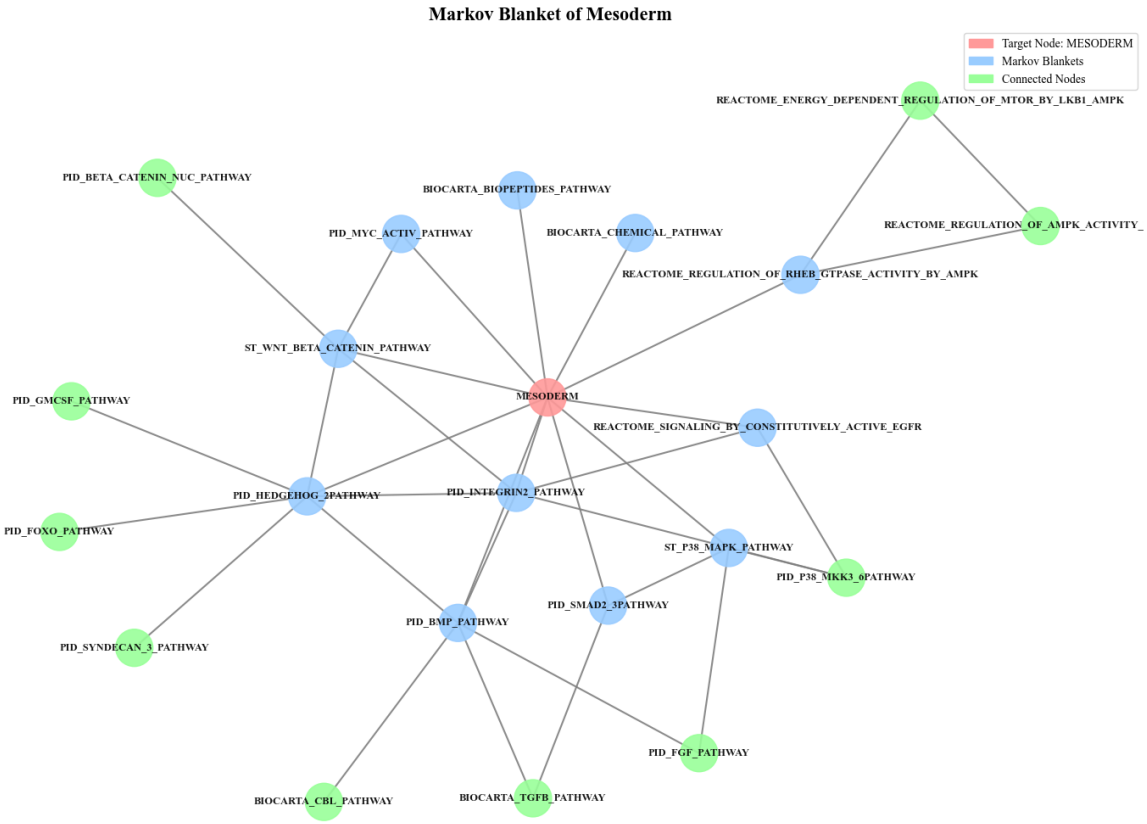


Figure 4.14: Markov Blanket of Mesoderm - Method 2

Hedgehog pathways, which act as primary developmental inducers orchestrating mesodermal commitment. These signals are supported by essential cellular modules including cell adhesion (Integrin2), proliferation control (MYC activation), and metabolic sensing (AMPK-RHEB-GTPase axis), establishing a robust contextual framework for differentiation.

A particularly compelling aspect of this network is the role of SMAD2/3 and p38 MAPK pathways, functioning as direct signal integrators. These nodes bridge upstream modulatory input to mesodermal induction outputs. Notably, the TGFβ (BIOCARTA_TGFβ _PATHWAY) and FGF (PID_FGF_PATHWAY) pathways [19] emerge as key upstream regulators, exerting indirect yet potent control via activation of SMAD2/3 and p38 MAPK, respectively.

For instance, attenuation of upstream TGFβ signaling would lead to diminished SMAD2/3 pathway activity, compromising the fidelity of mesodermal induction. Similarly, FGF-dependent modulation of both p38 MAPK and BMP [19] pathways reflects its dual role in patterning and feedback.

4.4.2 Mesenchymal

Beyond the primary germ layer formation, we also applied our integrated analytical pipeline to decipher the regulatory mechanisms driving mesenchymal stem cell differentiation into

specialized connective tissue lineages. The following subsections detail our findings on the regulatory architecture governing the commitment of mesenchymal stem cells towards adipocyte, cartilage, and osteocyte fates, respectively. Our workflow involved an initial phase of data processing and preliminary association identification. The refined data and insights generated from these steps then served as the foundational input for constructing comprehensive Bayesian Networks (Section 2.13), from which key regulatory modules, such as Markov Blankets, were extracted.

4.4.2.1 Regulatory Architecture of Adipocyte Commitment

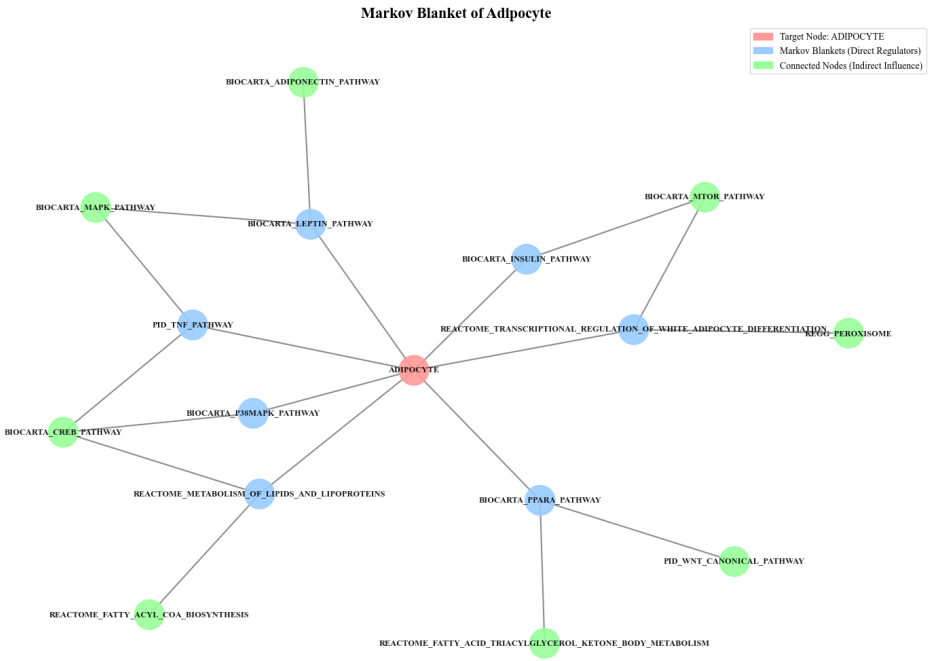


Figure 4.15: Markov Blanket for Adipocyte

The Adipocyte Markov Blanket reveals a convergent regulatory network orchestrating adipocyte differentiation and metabolic control. Core pathways such as Leptin, Insulin, PPARα, p38 MAPK, and transcriptional regulators directly govern adipogenic fate [22]. These are complemented by lipid metabolic processes and hormonal signaling via TNF and mTOR, which integrate nutrient and inflammatory cues.

A key insight is the downregulation of canonical Wnt signaling, a known inhibitor of adipogenesis. This repression facilitates the activation of PPARα [22] and other pro-adipogenic transcriptional programs, promoting terminal differentiation. Downstream effectors like CREB, MAPK, and mTOR emerge as central integrators of external signals into transcriptional and metabolic responses.

Overall, the network illustrates a multi-layered control system, where the suppression of anti-adipogenic signals like Wnt, and reinforcement of metabolic and transcriptional drivers,

ensure precise adipocyte fate specification.

4.4.2.2 Regulatory Architecture of Cartilage Commitment

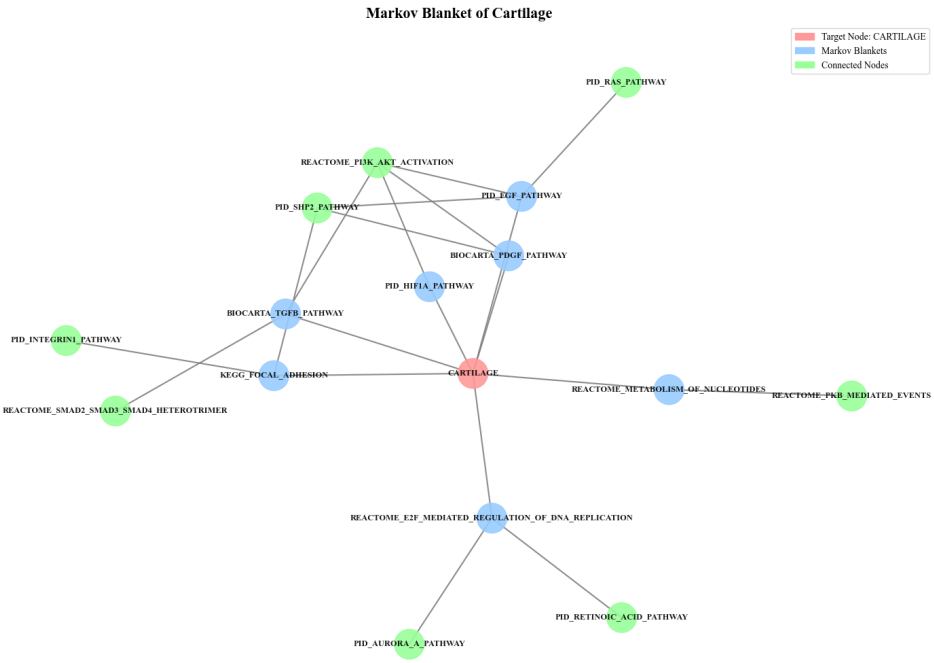


Figure 4.16: Markov Blanket of Cartilage

The Markov Blanket constructed around the CARTILAGE node reveals a structured regulatory framework highlighting critical signaling axes that contribute to cartilage differentiation and maintenance. The first-order nodes directly connected to CARTILAGE represent key upstream pathways that include focal adhesion signaling (KEGG_FOCAL_ADHESION), metabolic regulation (REACTOME_METABOLISM_OF_NUCLEOTIDES), and cell cycle control (REACTOME_E2F_MEDIATED_REGULATION_OF_DNA_REPLICATION). In addition, growth factor pathways such as PDGF (BIOCARTA_PDGF_PATHWAY), FGF (PID_FGF_PATHWAY), and TGF-β (BIOCARTA_TGFB_PATHWAY) [22], along with the hypoxia-responsive HIF1A pathway (PID_HIF1A_PATHWAY), provide direct regulatory input, underscoring the integration of extracellular matrix interactions, proliferation, morphogen signaling, and oxygen sensing in cartilage development.

These first-order signals are further supported by second-order nodes that capture downstream transduction and feedback loops. Notably, the PI3K-AKT signaling cascade (REACTOME_PI3K_AKT_ACTIVATION) acts as a central integrative hub, receiving input from PDGF, FGF, and HIF1A, [22] and mediating cellular survival, metabolic adaptation, and growth responses essential for chondrocyte function. The TGF-β pathway signal is transduced via the SMAD2/3/4 complex

(REACTOME_SMAD2_SMAD3_SMAD4_HETEROTRIMER), highlighting a canonical mechanism of transcriptional control. Similarly, the FGF pathway activates both RAS-MAPK (PID_RAS_PATHWAY) and p38 MAPK signaling axes (PID_P38_MAPK_PATHWAY), reflecting its role in proliferative and differentiation control.

Additional nodes such as SHP2 (PID_SHP2_PATHWAY) [27] and integrin-linked kinase pathways (PID_INTEGRIN1_PATHWAY) represent critical mediators of adhesion and receptor cross-talk, linking mechanical signaling with growth factor input. Cell cycle dynamics and retinoic acid signaling are represented through the AURORA-A (PID_AURORA_A_PATHWAY) and retinoic acid (PID_RETINOIC_ACID_PATHWAY) modules, respectively. Together, this network illustrates a tightly coordinated and layered signaling landscape essential for cartilage biology.

4.4.2.3 Regulatory Architecture of Osteocyte Commitment

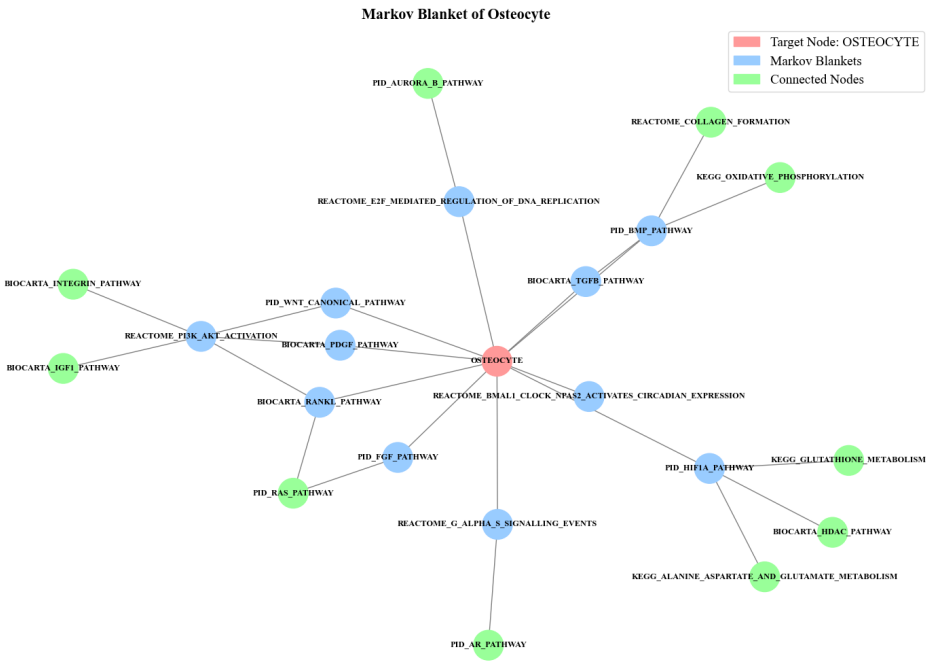


Figure 4.17: Markov Blanket of Osteocyte

The osteocyte Markov Blanket network presents a structured map of signaling interactions that govern osteocyte differentiation and function. At its core, the osteocyte is directly linked to key regulatory pathways, including TGF-beta and BMP signaling [22], which are critical for osteogenic commitment and connect to downstream processes such as collagen synthesis and oxidative phosphorylation. Canonical Wnt signaling also connects to the osteocyte and propagates through the PI3K-AKT axis, underscoring its role in survival and lineage-specific gene expression. PDGF and FGF pathways [22] contribute proliferative cues, converging on PI3K-AKT and RAS signaling, thus integrating mitogenic signals with differentiation.

Mechanosensory inputs are incorporated via RANKL (directly) and integrin (indirectly through PI3K-AKT), reflecting the influence of mechanical stimuli in bone remodeling.

The network also incorporates hypoxia-responsive HIF1A signaling, linking the osteocyte to amino acid and glutathione metabolism, thereby highlighting metabolic adaptation. Circadian control via BMAL1 and hormonal regulation through Gαs and androgen receptor pathways emphasize the role of temporal and endocrine inputs. Cell cycle regulation is represented by E2F-mediated control and Aurora B kinase, suggesting coordination between proliferation and differentiation. This architecture demonstrates how diverse upstream stimuli converge on core pathways such as PI3K-AKT, SMAD, and MAPK [28], which regulate the transcriptional and metabolic programs essential for osteocyte identity.

Overall, the Markov Blanket reflects a layered and integrative signaling framework, where specificity and robustness in cell fate decisions emerge from the convergence of developmental, mechanical, metabolic, and hormonal cues. The inclusion of both direct and indirect regulators highlights the complex yet coordinated control underlying osteocyte maturation.

4.5 Transcription Factors Orchestrating Germ Layer and Mesenchymal Lineage Differentiation Dynamics

To identify key transcription factors (TFs) involved in early lineage specification, we performed Spearman correlation analyses between TF expression levels and combined RNA velocity-derived scores across single-cell samples. These RNA velocity scores reflect the dynamic state transitions of differentiating cells and were computed for both primary germ layers, ectoderm, mesoderm, and endoderm, as well as mesenchymal-derived lineages, including adipocytes, cartilage, and osteocytes. This integrative approach allowed us to uncover TFs likely playing pivotal roles in regulating cell fate transitions during human embryonic stem cell (hESC) differentiation.

4.5.1 Transcription Factors Associated with Germ Layer Specification

The correlation analysis revealed distinct transcriptional signatures for each germ layer. A notable set of TFs including ZSCAN10, STAT3, MYB, ESRRA, OTX2, and CEBPZ, demonstrated strong positive correlations with endoderm and ectoderm RNA velocity scores, but negative correlations with mesoderm. These findings suggest that these TFs act as activators of endodermal and ectodermal lineages, while suppressing mesodermal differentiation. For instance, STAT3 is known to support pluripotency and neural (ectodermal) commitment, while OTX2 plays a critical role in anterior neuroectoderm specification [29].

MYB and ESRRA are associated with early differentiation and maintenance of a progenitor state. ZSCAN10, although less characterized, consistently showed strong positive associations with both endoderm and ectoderm, suggesting a potential regulatory role during early lineage priming.

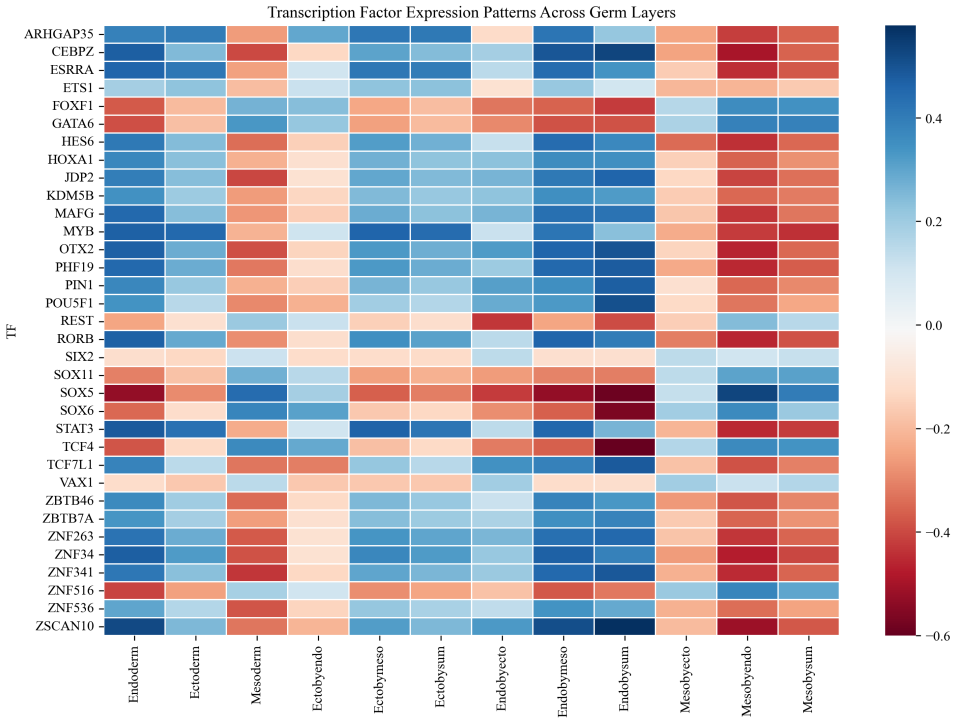


Figure 4.18: TF expression patterns in hESC germ layer specification (Spearman correlation)

In contrast, TFs such as SOX5, SOX6, TCF4, GATA6, and FOXF1 [29] exhibited strong positive correlations with mesodermal scores, while being negatively correlated with ectoderm and endoderm. These are established mesodermal regulators. For example, SOX5 and SOX6 are critical for mesodermal and chondrogenic differentiation, while GATA6 and FOXF1 are central to mesendodermal development and mesoderm patterning. TCF4, a component of the Wnt signaling pathway, further supports the known role of Wnt activity in mesoderm induction.

A subset of TFs, including HOXA1, PHF19, PIN1, and JDP2, showed moderate positive correlations across multiple germ layers, indicating potential roles in early developmental plasticity or intermediate states of lineage priming. In hybrid or transitional cell populations, such as ectoderm by endoderm (Ectobyendo) or endoderm by mesoderm (Endobymeso),TFs like ZSCAN10, STAT3, MYB, and CEBPZ maintained strong correlations, suggesting their involvement in lineage transitions or stabilization of multipotent states.

Interestingly, TFs such as REST and ZNF516 showed weak or negative correlations with endoderm and ectoderm but were mildly associated with mesoderm, indicating potential roles in repressing alternative fates to promote mesodermal identity. Collectively, these results align with known developmental roles of several TFs while also highlighting underexplored

candidates like ZSCAN10 and CEBPZ for future functional validation.

4.5.2 Transcription Factors Associated with Adipocyte, Cartilage, and Osteocyte Differentiation

Expanding our analysis to mesenchymal lineage specification, we assessed TF correlations with RNA velocity-derived scores for adipocyte, cartilage, and osteocyte fates. Several TFs showed clear lineage-specific associations consistent with known differentiation hierarchies.

Strong positive correlations with adipocyte lineage scores were observed for PPARG, KLF4, STAT5A, and RXRA. PPARG [22] is widely recognised as the master regulator of adipogenesis, acting in partnership with RXRA to drive lipid accumulation and adipocyte identity. KLF4 plays a role in early adipogenic commitment, while STAT5A contributes through growth hormone-mediated pathways.

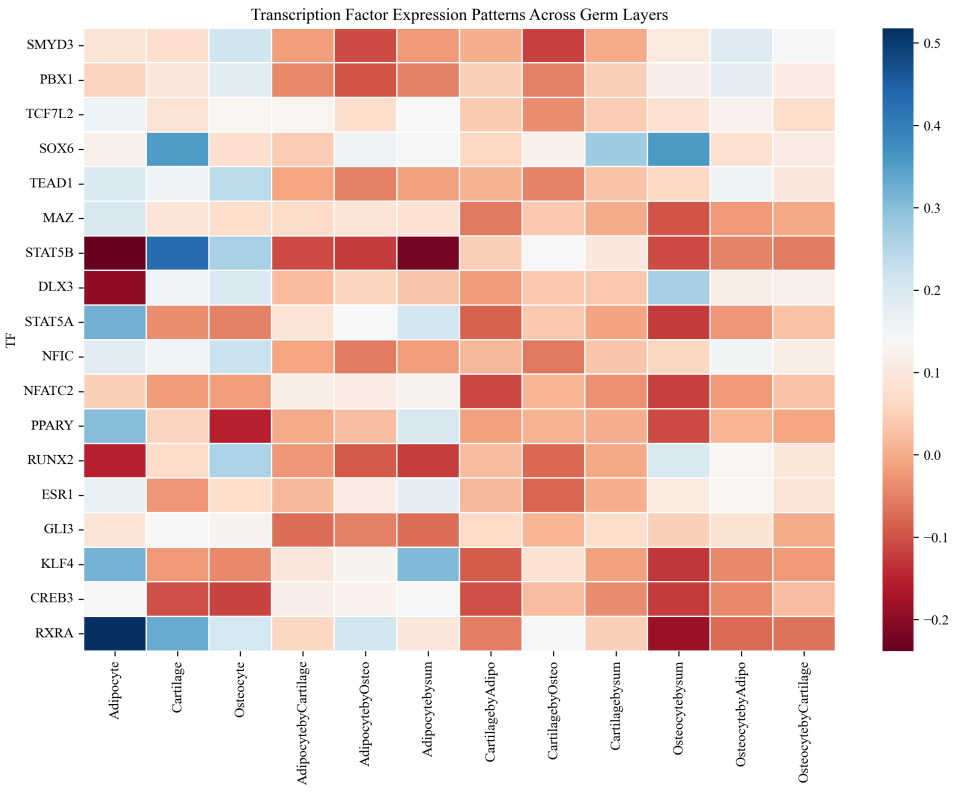


Figure 4.19: TF expression patterns in Mesenchymal specification (Spearman correlation)

For chondrogenic (cartilage) differentiation, SOX6 emerged as a prominent TF, showing high correlation with cartilage scores. Its role as a key regulator of chondrogenesis is well established, often in combination with SOX5 and SOX9. STAT5B, also positively correlated, may influence cartilage development via its role in chondrocyte proliferation.

In the osteocyte lineage, RUNX2 demonstrated a strong positive correlation, consistent with its central function in osteocyte maturation and skeletal development [22]. PBX1 and NFIC

also showed positive associations, suggesting auxiliary roles in osteogenic specification and bone morphogenesis.

Certain TFs, including TEAD1, MAZ, and GLI3, showed moderate correlations across multiple mesenchymal lineages. These TFs may contribute to lineage plasticity or early mesenchymal priming. Similarly, CREB3 and NFATC2 demonstrated context-dependent correlations, which may reflect their roles in cross-lineage regulatory networks or transitional cell states.

In hybrid populations, such as adipocyte-by-cartilage or cartilage-by-osteocyte, factors like SOX6, RXRA, and TEAD1 maintained positive associations, indicating that they may support lineage flexibility and transitions between closely related mesenchymal fates. In contrast, TFs like SMYD3 and CREB3 showed weak or negative correlations across all lineages, suggesting limited involvement or potential repressive roles in these differentiation pathways.

In conclusion, this analysis confirms the relevance of canonical regulators such as PPARG, SOX6, and RUNX2 in mesenchymal differentiation, while identifying additional TFs like TEAD1, PBX1, and MAZ as possible modulators of lineage transitions. These findings underscore the value of integrating RNA velocity and TF expression data to elucidate dynamic regulatory networks governing cell fate decisions during early development.

4.6 Identification of Key Signaling Pathways Driving Human Gastrulation

To identify key regulatory pathways involved in early human development, we applied a computational pipeline combining Bayesian network inference and Random Forest analysis. First, a Bayesian network was constructed to model dependencies among pathway activity profiles, from which the Markovian blanket for each cell fate transition was extracted. The pathways within these blankets were then used as features in a Random Forest classifier, and their importance scores were computed. This integrative approach allowed us to isolate pathways that are most predictive of transitions between key stages of human gastrulation, including Epiblast (Epi), Primitive Streak (PS), Nascent Mesoderm (NM), Emergent Mesoderm (EM), Advanced Mesoderm (AM), Hematopoietic Progenitors (HEP), and Erythroblasts (Ery). The following sections describe the dominant pathways identified for each developmental step.

4.6.1 Key Pathways Driving Epiblast to Primitive Streak Differentiation

The transition from epiblast to primitive streak marks the onset of gastrulation and germ layer specification. Pathway importance analysis using Random Forest on features selected through

Bayesian Markovian blanket extraction revealed several key regulators of this developmental step.

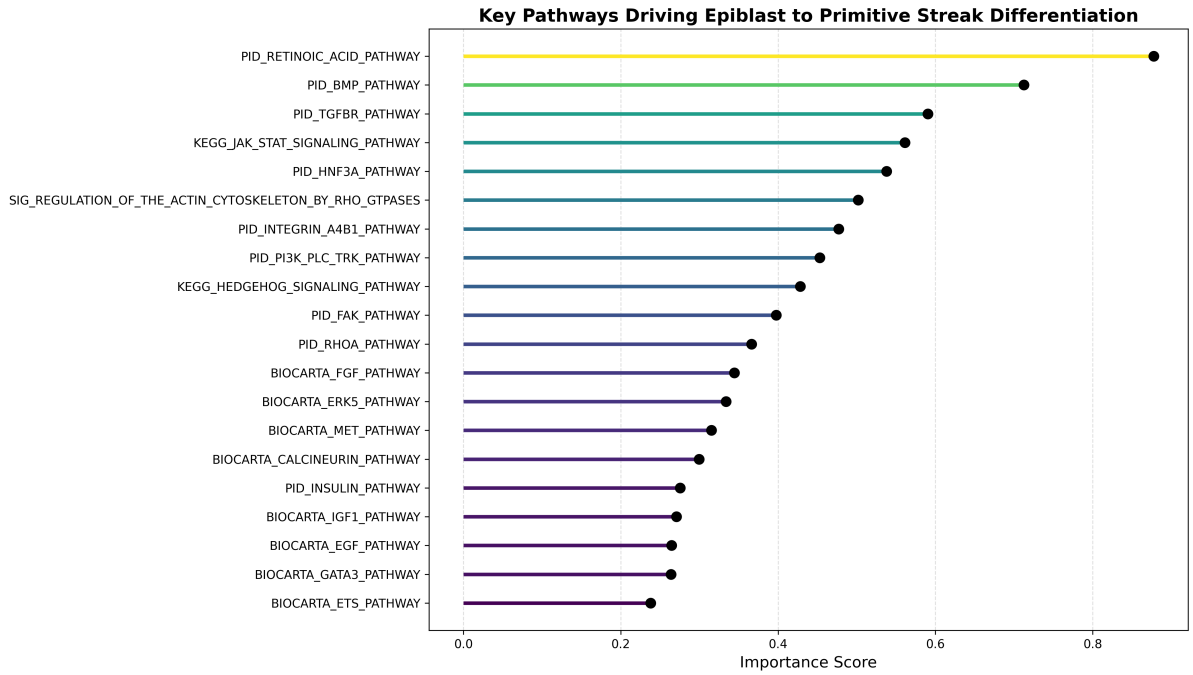


Figure 4.20: Key Pathways Driving Epiblast to Primitive Streak Differentiation.

Retinoic acid signaling emerged as a primary driver, suggesting its critical role in initiating differentiation. Canonical developmental pathways such as BMP, TGF-beta receptor [20], and JAK-STAT signaling were also highly ranked, consistent with their known involvement in primitive streak formation and mesoderm induction. The presence of cytoskeletal regulation pathways, including Rho GTPase and FAK signaling, highlights the importance of cell shape changes and migration during epiblast regression.

Additional pathways such as Integrin A4B1, PI3K-PLC-TRK [19], and Hedgehog signaling suggest roles in adhesion, survival, and patterning. The involvement of FGF [19], ERK/MAPK, and MET signaling points to mitogenic and morphogenetic cues, while insulin, IGF1, and EGF pathways reflect metabolic and proliferative support. Transcriptional regulators such as HNF3A (FOXA1), GATA3, and ETS pathways were also identified, indicating early gene expression reprogramming required for lineage commitment.

Collectively, these pathways reflect a tightly regulated interplay of differentiation signals, morphogenic gradients, cytoskeletal remodeling, and transcriptional activation that drive the epiblast toward a primitive streak identity.

4.6.2 Key Pathways Driving Primitive Streak to Nascent Mesoderm Differentiation

The emergence of nascent mesoderm from the primitive streak represents a critical step in mesodermal lineage commitment. Applying the Markovian blanket-guided Random Forest pipeline to this transition revealed key signaling mechanisms driving this developmental change.

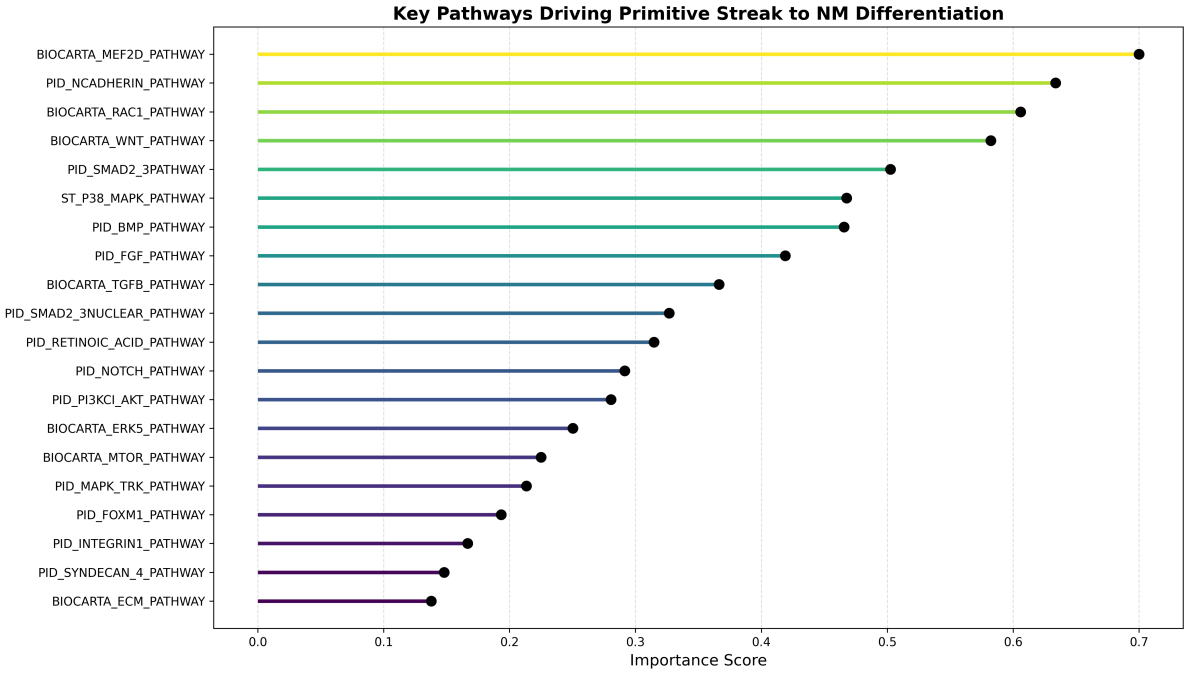


Figure 4.21: Key Pathways Driving Primitive Streak to Nascent Mesoderm Differentiation.

Among the most prominent was MEF2D signaling, pointing to early activation of muscle and mesodermal differentiation programs. Adhesion and EMT-related pathways like N-cadherin and Rac1 further emphasize the extensive cell migration and cytoskeletal reorganization required during this phase. Canonical developmental signals,WNT, TGFβ/SMAD2/3, BMP, FGF [19], and Notch, were consistently ranked high, underscoring their role in mesoderm induction, patterning, and specification.

Stress response and differentiation signaling through p38 MAPK [19], along with downstream modulators like ERK, mTOR, and PI3K/AKT, suggest an integrated response coordinating environmental signals with cell fate decisions. Pathways related to extracellular matrix remodeling and adhesion, such as Integrin, Syndecan-4, and ECM signaling, further highlight the dynamic tissue restructuring underway.

Collectively, this pathway profile reflects a tightly coordinated program involving morphogenic signaling, transcriptional regulation, and biomechanical adaptation essential for nascent mesoderm formation from primitive streak progenitors.

4.6.3 Key Pathways Driving Nascent Mesoderm to Emergent Mesoderm Differentiation

As cells progress from nascent to emergent mesoderm, they undergo further lineage commitment, expansion, and patterning. The feature importance analysis revealed a prominent role for canonical developmental signaling cascades, including WNT, Notch, BMP, and FGF [19], which continue to guide mesodermal maturation.

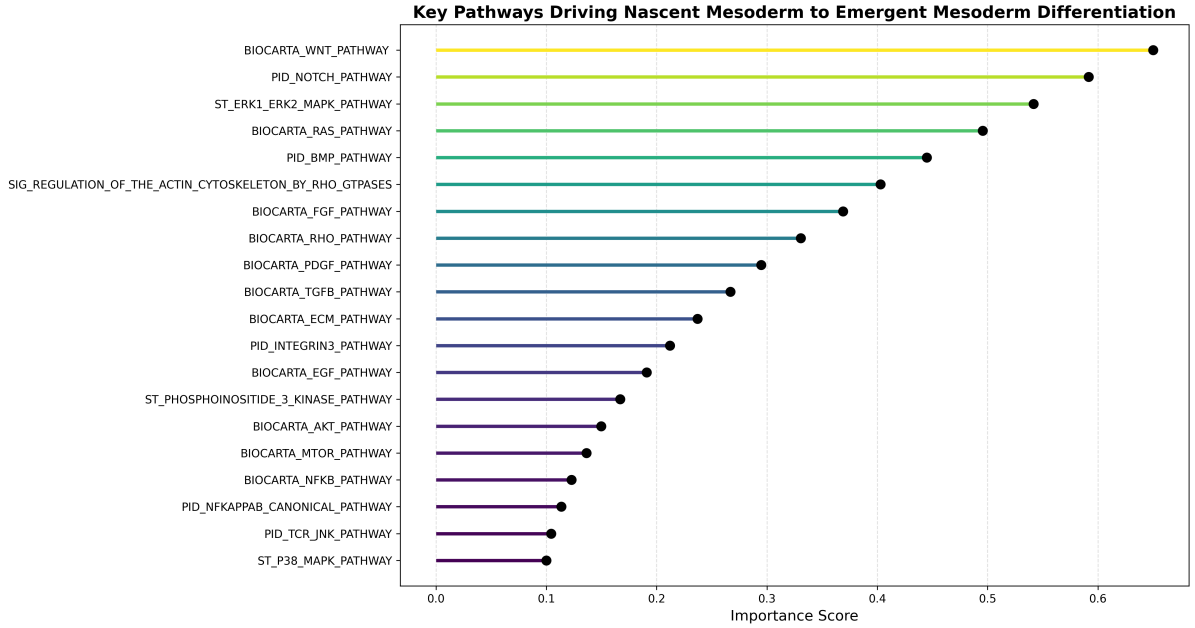


Figure 4.22: Key Pathways Driving Nascent Mesoderm to Emergent Mesoderm Differentiation

MAPK/ERK [19] and RAS signaling pathways were strongly represented, indicating enhanced proliferative and differentiation cues during this transition. Cytoskeletal remodeling and migration, essential for tissue morphogenesis, were reinforced by the involvement of Rho GTPase, actin regulation, and ECM-associated pathways. Notably, integrin signaling (particularly INTEGRIN3) and ECM interaction pathways suggested increasing complexity in cell–matrix communication. The PI3K/AKT/mTOR axis emerged as a key regulator, integrating growth factor signaling to support cell survival and anabolic processes.

Additionally, the presence of NF-κB signaling (both canonical and broader pathways) and stress-response MAPK branches [19] (p38, JNK) highlighted the role of inflammatory and mechanical cues in refining mesodermal cell identity and positioning.

Altogether, the identified pathways reflect a tightly coordinated transition involving morphogen signaling, adhesion, cytoskeletal dynamics, and metabolic control as nascent mesoderm cells acquire emergent mesodermal features.

4.6.4 Key Pathways Driving Emergent Mesoderm to Advanced Mesoderm Differentiation

The transition from emergent to advanced mesoderm marks further specification into defined mesodermal subtypes with increased structural and signaling complexity. Pathway importance analysis highlights the central role of ERK/MAPK, FGF, and TGFβ signaling [19], reinforcing their sustained involvement in mesodermal maturation and proliferation.

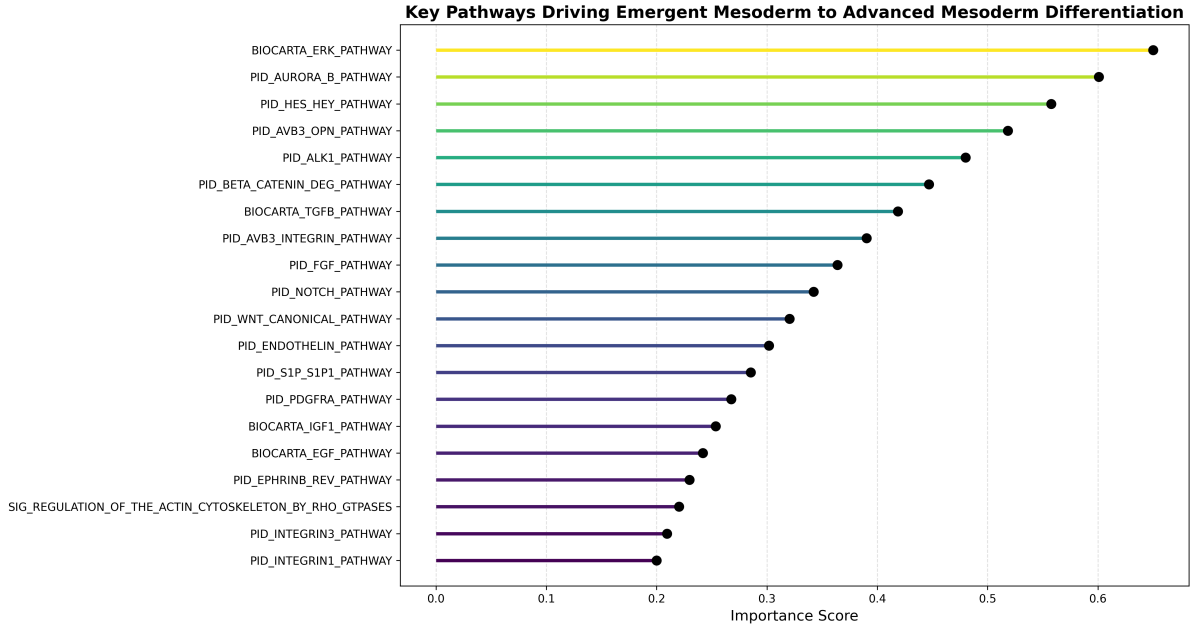


Figure 4.23: Key Pathways Driving Emergent Mesoderm to Advanced Mesoderm Differentiation

Notably, Notch, Wnt (canonical) [19], and ALK1/Endothelin signaling pathways emerged as critical regulators, suggesting active coordination of angiogenic, cardiac, and vascular mesoderm development. The appearance of PDGFRA and IGF1 pathways also points toward roles in mesodermal expansion and lineage stabilization through growth factor-mediated cues. Aurora B kinase and HES/HEY [30] (Notch effectors) suggest increasing regulation of cell cycle dynamics and transcriptional control, vital for precise mesodermal patterning. Meanwhile, integrin signaling (AVB3, Integrin 1 & 3) and actin cytoskeleton regulation emphasize continued roles for adhesion, migration, and mechanical integration during tissue organization.

Additionally, beta-catenin degradation hints at fine-tuned control over Wnt activity [19], likely essential to restrict or resolve earlier patterning signals. The S1P signaling pathway, involved in cell migration and vascular development, further underlines the onset of mesodermal sublineage commitment.

Altogether, these pathways represent a refined signaling environment that balances proliferation, structural organization, and fate specification as mesodermal cells progress toward more committed states.

4.6.5 Key Pathways Driving Advanced Mesoderm to Hematopoietic Progenitor Differentiation

The transition from advanced mesoderm to hematopoietic progenitors (HEPs) marks the onset of blood lineage specification and vascular development. Pathway importance analysis highlights Notch, FGF, BMP, and Wnt (canonical) [31] signaling as top contributors,each known to play integral roles in early hematopoietic induction, hemogenic endothelium specification, and vascular patterning.

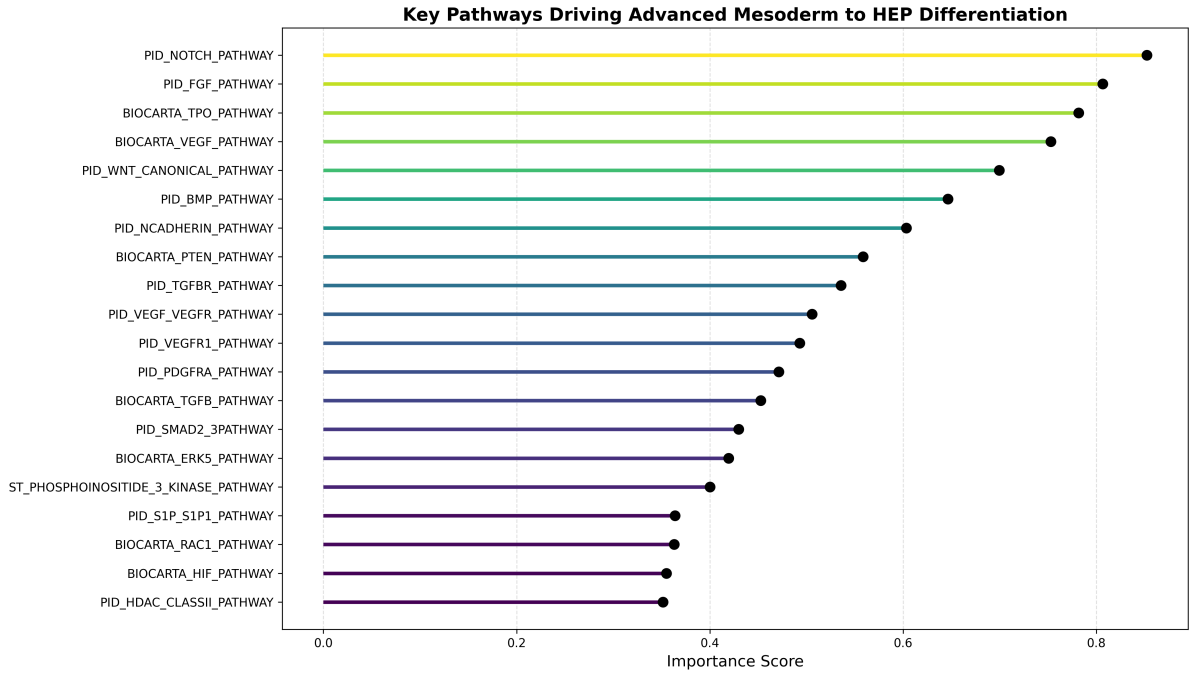


Figure 4.24: Key Pathways Driving Advanced Mesoderm to Hematopoietic Progenitor Differentiation

The prominence of the BIOCARTA_TPO_PATHWAY and VEGF/VEGFR [31] signaling underscores the activation of blood progenitor fate programs and angiogenic cues required for establishing hematopoietic niches. Signaling, via TGFβR, SMAD2/3, and TGFβ pathways, supports lineage priming and tightly regulates the balance between self-renewal and differentiation.

Pathways involving N-cadherin, Rac1, and S1P signaling reflect the cytoskeletal remodeling and migratory behaviors essential for hematopoietic progenitor emergence and niche localization. Meanwhile, PTEN, PI3K, and HDAC class II signaling suggest complex integration of intracellular signaling, metabolic reprogramming, and chromatin-level regulation necessary for HEP commitment.

The combined influence of angiogenic (VEGF) [31], inflammatory (HIF), and transcriptional regulation (HDAC, SMAD, Notch) pathways emphasizes the multifaceted molecular environment orchestrating this critical stage of early hematopoietic development.

4.6.6 Key Pathways Driving Hematopoietic Progenitor to Erythroblast Differentiation

The differentiation of hematopoietic progenitors (HEPs) into erythroblasts represents a crucial commitment to the erythroid lineage. Pathway analysis reveals a strong enrichment of signaling modules known to drive erythropoiesis, survival, and proliferation of red blood cell precursors.

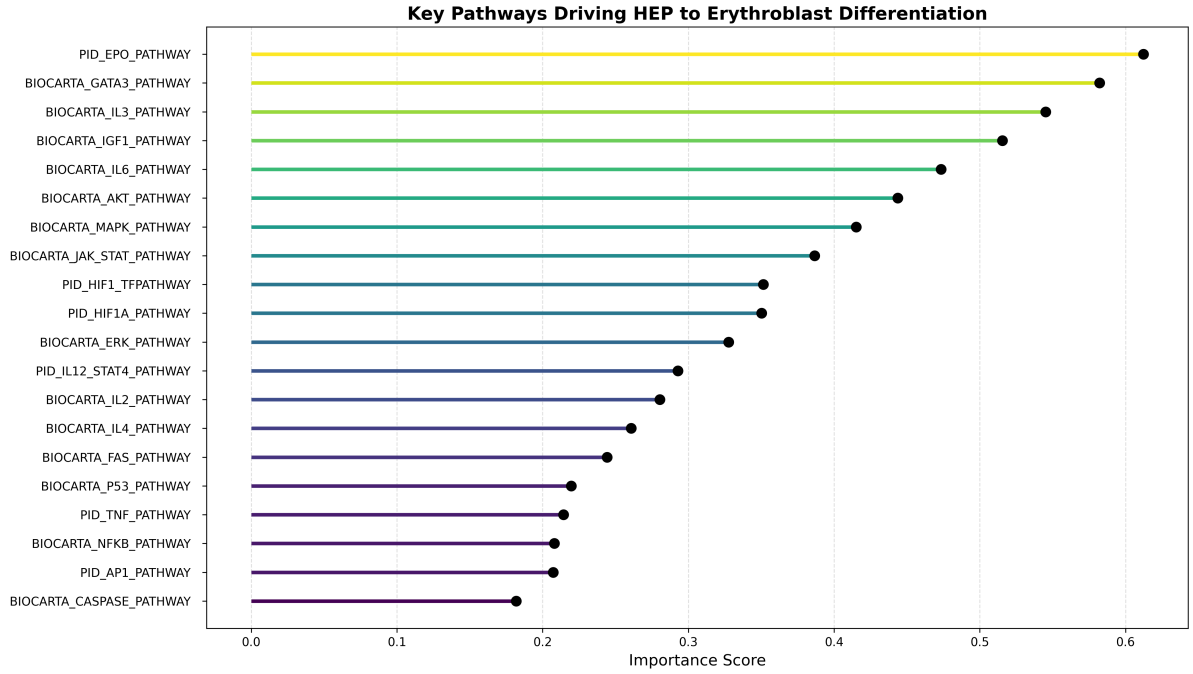


Figure 4.25: Key Pathways Driving Hematopoietic Progenitor to Erythroblast Differentiation

At the forefront is the Erythropoietin (EPO) pathway [32], a central driver of erythroid differentiation and maturation. Supporting transcriptional regulators include the GATA3 pathway, pivotal for hematopoietic lineage commitment, and JAK/STAT signaling, a key mediator of cytokine-driven erythropoiesis.

Pro-growth and survival cues are also prominent: IGF1, IL-3, IL-6, and AKT [33] pathways promote progenitor cell survival and proliferation, while MAPK and ERK cascades contribute to erythroid cell cycle progression and terminal maturation. The activation of HIF1A and HIF1_TF [32] pathways indicates a hypoxic response element, aligning with known roles of hypoxia in promoting erythropoietin signaling and red cell production.

Additionally, immune-related pathways such as IL-2, IL-4, IL-12/STAT4, and TNF [33] signaling reflect cytokine crosstalk that may modulate erythroid development under inflammatory or stress conditions. Apoptotic regulators, including the Fas, p53, AP-1, NF-κB, and Caspase pathways [33], suggest fine-tuned control over erythroid progenitor survival and elimination during differentiation.

Overall, the transition from HEPs to erythroblasts is orchestrated by a synergistic network of cytokine, transcriptional, hypoxic, and apoptotic signals, ensuring robust commitment and maturation of the erythroid lineage.

CHAPTER 5

Conclusion and Future Aspects

5.1 Conclusion

This thesis successfully introduces and validates a novel computational framework that significantly advances the understanding and control of stem cell differentiation. By integrating single-cell RNA sequencing, RNA velocity analysis, and sophisticated probabilistic modeling techniques (UniPath and Bayesian Networks), we have established a robust pipeline for systematically identifying key signaling pathways and transcription factors that orchestrate lineage commitment in both human embryonic stem cells (hESCs) and mesenchymal stem cells (MSCs).

A core strength of our approach lies in its ability to estimate lineage "poising levels" and resolve differentiation bifurcations, moving beyond static transcriptomic snapshots to capture the dynamic essence of cell fate decisions. The comprehensive application of this framework to human gastrulation data revealed intricate regulatory networks, confirming the established roles of major developmental pathways such as Wnt, BMP, TGFβ, and FGF, while also highlighting the importance of more nuanced players like Retinoic Acid and specific cell adhesion pathways in precise lineage transitions. Critically, our methodology identifies direct regulatory drivers and distinguishes them from indirect associations, providing a deeper mechanistic understanding.

Furthermore, the identification of a diverse set of transcription factors, including both well-known master regulators (e.g., PPARG, SOX6, RUNX2) and underexplored candidates (e.g., ZSCAN10, CEBPZ), provides valuable molecular targets for future experimental manipulation. The consistency of our computational predictions with extensive prior biological literature across hESC germ layer specification, endodermal organogenesis, and mesenchymal differentiation underscores the accuracy and reliability of the proposed framework. By eliminating the need for genetic modification, this research offers a safer, more precise, and potentially more clinically viable alternative to current regenerative medicine strategies. Ultimately, this work represents a crucial step towards developing highly controlled, scalable, and reproducible protocols for stem cell-based therapies, addressing long-standing challenges in the field and paving the way for targeted tissue regeneration.

5.2 Future Aspects

The computational framework developed in this thesis opens several exciting avenues for future research and clinical translation:

Causal Inference and Experimental Validation

While Bayesian Networks infer probabilistic dependencies and suggest direct regulatory influences, establishing true causality remains a critical next step. Future work could integrate the identified pathways and transcription factors into in vitro perturbation experiments (e.g., using CRISPR/Cas9 for gene knockout/activation, or small molecule inhibitors/activators) followed by scRNA-seq to definitively validate their causal roles in directing specific lineage fates. This would translate computational predictions into empirically verifiable biological insights.

Pharmacological Modulators and Small Molecule Screening

The identification of key signaling pathways and transcription factors lays the groundwork for the discovery of specific pharmacological modulators. A significant future direction involves leveraging the identified pathways to conduct high-throughput in silico and in vitro screens for novel small molecule modulators. This could lead to the discovery of highly potent and specific chemical compounds capable of inducing desired cell fates with unprecedented precision, facilitating cost-effective and scalable production of specialized cells for therapeutic applications.

Integration with Multi-Omics Data

The current framework primarily uses scRNA-seq data. Future enhancements could incorporate other single-cell multi-omics data, such as single-cell ATAC-seq (for chromatin accessibility), single-cell proteomic data, or spatial transcriptomics. This multi-modal integration would provide a more holistic view of cellular states, regulatory landscapes, and spatial organization, further refining the prediction of lineage commitment and regulatory interactions.

Development of Predictive Models for Disease

The established lineage-specific regulatory networks and identified TFs/pathways could be used to build predictive models for developmental disorders or diseases linked to aberrant differentiation. Understanding how these networks are perturbed in disease states could identify novel diagnostic biomarkers and therapeutic targets for interventions.

Personalized Regenerative Medicine

The framework's ability to account for intrinsic variability among stem cells, particularly

those from different sources (e.g., patient-specific iPSCs), lays a strong foundation for personalized regenerative medicine. Future work could focus on applying the framework to patient-derived stem cells to design individualized differentiation protocols, optimizing therapeutic outcomes and minimizing off-target effects.

Real-Time Differentiation Monitoring and Control

Advances in live-cell imaging and biosensors could be combined with the computational framework to enable real-time monitoring of differentiation trajectories. Integrating these data into a closed-loop system could facilitate dynamic, feedback-controlled differentiation processes, allowing for adaptive adjustments to optimize cell yields and purity.

Expansion to Broader Developmental Contexts and Organoids

Applying this integrated framework to more complex organoid systems or in vivo developmental contexts would further test its generalizability and uncover additional layers of regulatory complexity, ultimately contributing to the bioengineering of more functional tissues and organs.

By pursuing these future directions, the computational framework presented in this thesis holds the potential to profoundly impact the fields of stem cell biology, developmental biology, and regenerative medicine, bringing us closer to realizing the full therapeutic promise of stem cell-based therapies.

References

[1] Alan Trounson and Natalie D DeWitt. Pluripotent stem cells progressing to the clinic. *Nature Reviews Molecular Cell Biology*, 17(3):194–200, 2016.

[2] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Efrat Braun, Hadas Hochgerner, Viktor Petukhov, Kristina Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

[3] Tanya Barrett, Sarah E Wilhite, Pierre Ledoux, Cara Evangelista, Irene F Kim, Marina Tomashevsky, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[4] NCBI. Ncbi sra toolkit. `https://github.com/ncbi/sra-tools`. Accessed: 2025-06-17.

[5] Simon Andrews. Fastqc: a quality control tool for high throughput sequence data. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`, 2010. Accessed: 2025-06-17.

[6] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

[7] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, et al. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[8] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[9] Volker Bergen, Marius Lange, Shila Peidli, Fabian Alexander Wolf, and Fabian J Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, 2020.

[10] Shaurya Chawla, Suresh Samydurai, Shun Long Kong, Zhenxun Wu, Zhen Wang, Wai Leong Tam, Debarka Sengupta, and Vibhor Kumar. Unipath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. *Nucleic Acids Research*, 49(3):e13, 2021.

[11] Jerrold H. Zar. *Biostatistical Analysis*. Pearson Education, Upper Saddle River, NJ, 5th edition, 2010.

[12] Marco Scutari and Jean-Baptiste Denis. *Bayesian Networks with Examples in R*. Chapman and Hall/CRC, Boca Raton, FL, 2014.

[13] Elena R Palacios, Adeel Razi, Thomas Parr, Michael Kirchhoff, and Karl Friston. On markov blankets and hierarchical self-organisation. *Journal of Theoretical Biology*, 486:110089, 2020.

[14] Fabian A Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

[15] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.

[16] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[17] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. arXiv:1802.03426.

[18] Ronald R Coifman and Stephane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[19] Kyle M Loh, Angela Chen, Pang Wei Koh, Tianda Z Deng, Rahul Sinha, Jonathan M Tsai, Amira A Barkal, Kimberle Y Shen, Rajan Jain, Rachel M Morganti, Shyh-Chang Ng, Nathaniel B Fernhoff, Benson M George, Gerlinde Wernig, Rachel EA Salomon, Zhenghao Chen, Hannes Vogel, Jonathan A Epstein, Anshul Kundaje, William S Talbot, Philip A Beachy, Lay Teng Ang, and Irving L Weissman. Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell*, 166(2):451–467, 2016.

[20] Kyle M Loh, Lay Teng Ang, Jingyao Zhang, Vibhor Kumar, Jasmin Ang, Jun Qiang Auyeong, Kian Leong Lee, Siew Hua Choo, Christina YY Lim, Massimo Nichane, Junru Tan, Monireh Soroush Noghabi, Lisa Azzola, Elizabeth S Ng, Jens Durruthy-Durruthy, Vittorio Sebastiano, Lorenz Poellinger, Andrew G Elefanty, Edouard G Stanley, Qingfeng Chen, Shyam Prabhakar, Irving L Weissman, and Bing Lim. Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell*, 14(2):237–252, 2014.

[21] James M Wells and Douglas A Melton. Vertebrate endoderm development. *Annual Review of Cell and Developmental Biology*, 25:321–346, 2009.

[22] Y Zhang, D Khan, J Delling, and E Tobiasch. Mechanisms underlying the osteo- and adipo-differentiation of human mesenchymal stem cells. *The Scientific World Journal*, 2012:793823, 2012. Epub 2012 Mar 12.

[23] TJJ de Winter and Roel Nusse. Running against the wnt: How wnt/beta-catenin suppresses adipogenesis. *Frontiers in Cell and Developmental Biology*, 9:627429, 2021. Published 2021 Feb 9.

[24] SL Ang, A Wierda, D Wong, KA Stevens, S Cascio, J Rossant, and KS Zaret. The formation and maintenance of the definitive endoderm lineage in the mouse: involvement of hnf3/forkhead proteins. *Development*, 119(4):1301–1315, 1993.

[25] Junying Yu, Jizhou Zou, Zhaohui Ye, Henry Hammond, Guokai Chen, Akitsu Tokunaga, Prashant Mali, Yuming M Li, Curt Civin, Nicholas Gaiano, and Linzhao Cheng. Notch signaling activation in human embryonic stem cells is required for embryonic, but not trophoblastic, lineage commitment. *Cell Stem Cell*, 2(5):461–471, 2008.

[26] A Kolpakova, S Katz, A Keren, A Rojtblat, and E Bengal. Transcriptional regulation of mesoderm genes by mef2d during early xenopus development. *PLoS One*, 8(7):e69693, 2013.

[27] Margot E Bowen, Ugur M Ayturk, Kyle C Kurek, Wentian Yang, and Matthew L Warman. Shp2 regulates chondrocyte terminal differentiation, growth plate architecture and skeletal cell fates. *PLoS Genetics*, 10(5):e1004364, 2014.

[28] Qiang Chen, Peng Shou, Cheng Zheng, Min Jiang, Guodong Cao, Qing Yang, Jing Cao, Na Xie, Tania Velletri, Lixin Zhang, Yufang Zhang, Jian Hou, You Wang, Anping Wang, and Yufang Shi. Fate decision of mesenchymal stem cells: adipocytes or osteoblasts? *Cell Death & Differentiation*, 23:1128–1139, 2016.

[29] Alex Tsankov, Hongcang Gu, Vladimir Akopian, Michael J Ziller, Jennifer Donaghey, Ido Amit, David K Gifford, and Alexander Meissner. Transcription factor binding dynamics during human es cell differentiation. *Nature*, 518:344–349, 2015.

[30] Ana G Freire, Abhinav Waghray, Fabiana Soares-da Silva, Tatiana P Resende, Daphne F Lee, Catarina F Pereira, Dália S Nascimento, Ihor R Lemischka, and Patrícia Pinto-do Ó. Transient hes5 activity instructs mesodermal cells toward a cardiac fate. *Stem Cell Reports*, 9(1):136–148, 2017. Epub 2017 Jun 22.

[31] Flora F Bruveris, Elizabeth S Ng, Edouard G Stanley, and Andrew G Elefanty. Vegf, fgf2, and bmp4 regulate transitions of mesoderm to endothelium and blood cells in a human model of yolk sac hematopoiesis. *Experimental Hematology*, 103:30–39.e2, 2021. Epub 2021 Aug 23.

[32] Asterios S. Tsiftsoglou, Ioannis S. Vizirianakis, and John Strouboulis. Erythropoiesis: Model systems, molecular regulators, and developmental programs. *IUBMB Life*, 61(8):800–830, Aug 2009.

[33] Zuzana Tóthová, Martina Šemeláková, Zuzana Solárová, Jozef Tomc, Natasa Debeljak, and Peter Solár. The role of pi3k/akt and mapk signaling pathways in erythropoietin signalization. *International Journal of Molecular Sciences*, 22(14):7682, 2021.