Understanding the molecular mechanisms regulating the early placental development using Next-Generation Sequencing (NGS) datasets

Ashish Jain
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Recommended Citation
Jain, Ashish, "Understanding the molecular mechanisms regulating the early placental development using Next-Generation Sequencing (NGS) datasets" (2020). Graduate Theses and Dissertations. 18335. https://lib.dr.iastate.edu/etd/18335

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Understanding the molecular mechanisms regulating the early placental development using Next-Generation Sequencing (NGS) datasets

by

Ashish Jain

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Geetu Tuteja, Co-major Professor
Julie A. Dickerson, Co-major Professor
Karín S. Dorman
Xun Gu
Heike Hofmann

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

Iowa State University
Ames, Iowa
2020

Copyright © Ashish Jain, 2020. All rights reserved.
I would like to dedicate my research to my parents, Shashi Jain and Kuldeep Rai Jain, for their unconditional love and support throughout my life.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>Page</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>Page</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>Page</td>
<td>xii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>Page</td>
<td>xiv</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>Page</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Human placental development</td>
<td>Page</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Trophoblast cells</td>
<td>Page</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Models to study early human placental development</td>
<td>Page</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1 Animal Models</td>
<td>Page</td>
<td>5</td>
</tr>
<tr>
<td>1.3.2 Cell lines</td>
<td>Page</td>
<td>2</td>
</tr>
<tr>
<td>1.4 Complications due to placental disorders</td>
<td>Page</td>
<td>9</td>
</tr>
<tr>
<td>1.5 Multi-omics analysis to study Preeclampsia</td>
<td>Page</td>
<td>10</td>
</tr>
<tr>
<td>1.6 Multi-omics data integration using matrix factorization</td>
<td>Page</td>
<td>11</td>
</tr>
<tr>
<td>1.6.1 iCluster, iClusterPlus, and iClusterBayes</td>
<td>Page</td>
<td>12</td>
</tr>
<tr>
<td>1.6.2 MoCluster</td>
<td>Page</td>
<td>13</td>
</tr>
<tr>
<td>1.6.3 Multi-Omics Factor Analysis (MOFA)</td>
<td>Page</td>
<td>14</td>
</tr>
<tr>
<td>1.7 Dissertation Organization</td>
<td>Page</td>
<td>15</td>
</tr>
<tr>
<td>1.8 References</td>
<td>Page</td>
<td>16</td>
</tr>
<tr>
<td>CHAPTER 2. DECIPHERING TRANSCRIPTIONAL REGULATION IN HUMAN EMBRYONIC STEM CELLS SPECIFIED TOWARDS A TROPHOBLAST FATE</td>
<td>Page</td>
<td>25</td>
</tr>
<tr>
<td>2.1 Abstract</td>
<td>Page</td>
<td>25</td>
</tr>
<tr>
<td>2.2 Introduction</td>
<td>Page</td>
<td>26</td>
</tr>
<tr>
<td>2.3 Methods</td>
<td>Page</td>
<td>28</td>
</tr>
<tr>
<td>2.3.1 RNA-Seq data processing</td>
<td>Page</td>
<td>28</td>
</tr>
<tr>
<td>2.3.2 Human Protein Atlas analysis</td>
<td>Page</td>
<td>28</td>
</tr>
<tr>
<td>2.3.3 Differential expression analysis</td>
<td>Page</td>
<td>29</td>
</tr>
<tr>
<td>2.3.4 Functional enrichment analysis</td>
<td>Page</td>
<td>29</td>
</tr>
<tr>
<td>2.3.5 Weighted co-expression network construction</td>
<td>Page</td>
<td>29</td>
</tr>
<tr>
<td>2.3.6 Randomization Tests</td>
<td>Page</td>
<td>30</td>
</tr>
<tr>
<td>2.3.7 PPI network analysis</td>
<td>Page</td>
<td>31</td>
</tr>
<tr>
<td>2.4 Results</td>
<td>Page</td>
<td>31</td>
</tr>
<tr>
<td>2.4.1 Genes highly expressed in ESCd are enriched for placental development genes</td>
<td>Page</td>
<td>31</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Global expression profile in ESCd are not similar to expression profile in mesoderm cells</td>
<td>32</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Invasion genes are upregulated in ESCd compared to commonly used TB cell lines</td>
<td>33</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Identification of placental gene clusters in ESCd by co-expression network analysis</td>
<td>34</td>
</tr>
<tr>
<td>2.4.5</td>
<td>Protein-Protein Interaction (PPI) network analysis of co-expressed modules</td>
<td>35</td>
</tr>
<tr>
<td>2.5</td>
<td>Discussion</td>
<td>36</td>
</tr>
<tr>
<td>2.6</td>
<td>Supplementary Material</td>
<td>39</td>
</tr>
<tr>
<td>2.7</td>
<td>References</td>
<td>40</td>
</tr>
</tbody>
</table>

CHAPTER 3. TISSUEENRICH: TISSUE-SPECIFIC GENE ENRICHMENT ANALYSIS... 51

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Abstract</td>
<td>51</td>
</tr>
<tr>
<td>3.2</td>
<td>Introduction</td>
<td>51</td>
</tr>
<tr>
<td>3.3</td>
<td>Methods</td>
<td>52</td>
</tr>
<tr>
<td>3.3.1</td>
<td>RNA-Seq Datasets</td>
<td>52</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Tissue Naming conventions</td>
<td>52</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Defining tissue-specific genes</td>
<td>53</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Tissue-specific gene enrichment</td>
<td>54</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Background Genes</td>
<td>54</td>
</tr>
<tr>
<td>3.4</td>
<td>Results</td>
<td>55</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Web Application</td>
<td>55</td>
</tr>
<tr>
<td>3.4.2</td>
<td>R Package</td>
<td>55</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Case study 1: Defining tissue-specificity of genes expressed in differentiated embryonic stem cells (ESCs)</td>
<td>56</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Case study 2: Annotation of cell clusters from single-cell RNA-Seq data</td>
<td>56</td>
</tr>
<tr>
<td>3.5</td>
<td>Conclusions</td>
<td>57</td>
</tr>
<tr>
<td>3.6</td>
<td>References</td>
<td>57</td>
</tr>
</tbody>
</table>

CHAPTER 4. FACTOR ANALYSIS-CLUSTERING MODEL FOR INTEGRATING AND CLUSTERING MULTI-OMICS DATA... 62

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Abstract</td>
<td>62</td>
</tr>
<tr>
<td>4.2</td>
<td>Introduction</td>
<td>63</td>
</tr>
<tr>
<td>4.3</td>
<td>Methods</td>
<td>65</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Factor analysis-clustering model</td>
<td>65</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Model inference</td>
<td>69</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Initialization of model parameters</td>
<td>71</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Simulation Studies</td>
<td>72</td>
</tr>
<tr>
<td>4.3.5</td>
<td>Data processing</td>
<td>74</td>
</tr>
<tr>
<td>4.3.6</td>
<td>Downstream analysis</td>
<td>75</td>
</tr>
<tr>
<td>4.4</td>
<td>Results</td>
<td>76</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Comparison with other tools using simulations</td>
<td>76</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Colorectal cancer subtype analysis</td>
<td>78</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Preeclampsia subtype analysis</td>
<td>80</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
<td>81</td>
</tr>
<tr>
<td>4.6</td>
<td>Supplementary Material</td>
<td>82</td>
</tr>
</tbody>
</table>
C.3.4 Invasion Assay .......................................................... 134
C.3.5 Immunoassays ........................................................... 135
C.3.6 Illumina TruSeq RNA Library Preparation and Sequencing ........ 135
C.3.7 Differential expression analysis ..................................... 136
C.3.8 Functional enrichment analysis ....................................... 136
C.3.9 Weighted co-expression network analysis ........................... 136

C.4 Results ............................................................................. 137
C.4.1 Phenotypes of the Primary Fibroblast and Induced Pluripotent Stem Cells. 137
C.4.2 Invasive features of iPSC-derived TB under low and high $O_2$ ....... 138
C.4.3 Effect of low and high $O_2$ on relative growth rates of differentiating CTL- and EOPE-iPSC. .................................................. 140
C.4.4 Production of hCG and progesterone by CTL- and EOPE-TB under 5% and 20% $O_2$ conditions. .................................................. 141
C.4.5 Production of angiogenic factors PGF and FLT1 by CTL- and EOPE-TB under 5% and 20% $O_2$ conditions. ................................ 142
C.4.6 Gene Expression Profiles of EOPE-TB and CTL-TB ................ 143
C.4.7 Co-expressed gene modules associated with invasion .............. 143

C.5 Discussion .......................................................................... 145
C.6 Conclusion ......................................................................... 151
C.7 Supplementary Material ..................................................... 151
C.8 References .......................................................................... 151

APPENDIX D. PLACENTACELLENRICH: A TOOL TO CHARACTERIZE GENE SETS USING PLACENTA CELL-SPECIFIC GENE ENRICHMENT ANALYSIS ........ 165
D.1 Abstract ............................................................................ 165
D.2 Introduction ....................................................................... 165
D.3 Methods ............................................................................. 166
D.3.1 scRNA-Seq Datasets ...................................................... 166
D.3.2 Defining genes with cell-specific expression ......................... 167
D.3.3 Cell-specific gene enrichment ........................................ 167
D.3.4 Background Genes .......................................................... 168
D.4 Results .............................................................................. 168
D.4.1 Carrying out placenta cell-specific gene enrichment analysis .... 168
D.4.2 Visualizing expression patterns of genes across placenta cell types 170
D.4.3 Example usage scenarios of PlacentaCellEnrich .................... 171
D.5 Discussion ......................................................................... 172
D.6 References .......................................................................... 173
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>List of bottleneck genes in the PPI networks of the ESCd co-expression</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>modules with their function in placental development.</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Parameters used for simulation study</td>
<td>72</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison of the performance of MOFA-Cluster with other tools in terms of</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>mean ARI on simulated data</td>
<td></td>
</tr>
<tr>
<td>A.1</td>
<td>Summary of the differential expression analysis of the placental cell lines</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>with ESCd</td>
<td></td>
</tr>
<tr>
<td>A.2</td>
<td>Detailed information about the RNA-Seq datasets used for analysis.</td>
<td>103</td>
</tr>
<tr>
<td>C.1</td>
<td>Patient information for UC samples collected from CTLs (1 to 10) and from</td>
<td>164</td>
</tr>
<tr>
<td></td>
<td>infants whose mothers suffered EOPE (A to S).</td>
<td></td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Differentiation of the trophoblast cells during pregnancy.</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>ESCd are associated with placental functions and placenta specific genes.</td>
<td>45</td>
</tr>
<tr>
<td>2.2</td>
<td>Comparison of ESCd to mesoderm cells.</td>
<td>46</td>
</tr>
<tr>
<td>2.3</td>
<td>Comparison of ESCd with different placental cell lines.</td>
<td>47</td>
</tr>
<tr>
<td>2.4</td>
<td>ESCd co-expression modules.</td>
<td>48</td>
</tr>
<tr>
<td>2.5</td>
<td>PPI Network showing bottleneck gene connections.</td>
<td>49</td>
</tr>
<tr>
<td>3.1</td>
<td>Bar charts showing the enrichment of (a) Heart Muscle (using the HPA dataset) and (c) Heart (using the Mouse ENCODE dataset) in the 1,000 most highly expressed genes in cardiomyocytes differentiated from ESCs. Heatmaps showing the expression of (b) Heart Muscle and (d) E14.5-Heart specific genes enriched in the 1,000 most highly expressed genes.</td>
<td>59</td>
</tr>
<tr>
<td>3.2</td>
<td>Bar chart showing the enrichment (-Log(_{10})(P-Value)) of placenta-specific genes (using the Mouse ENCODE dataset) in the 1,000 most highly expressed genes in trophoblast-like cells differentiated from ESCs.</td>
<td>60</td>
</tr>
<tr>
<td>3.3</td>
<td>Bar charts showing the enrichment (-Log(_{10})(P-Value)) of visceral organ specific genes in the marker genes of cluster 1 (a), placenta-specific genes in the marker genes of cluster 2 (b), and brain and olfactory bulb specific genes in the marker genes of cluster 3 (c) (using the Mouse ENCODE dataset). Marker genes belonging to each cluster were obtained from Scialdone et al.</td>
<td>61</td>
</tr>
<tr>
<td>4.1</td>
<td>Results comparing the performance of MOFA-Cluster and other tools in terms of ARI scores to identify true clusters using the simulated data.</td>
<td>89</td>
</tr>
<tr>
<td>4.2</td>
<td>Results comparing the performance of MOFA-Cluster and other tools in terms of the mean distance between the estimated model parameters and true parameters using the simulated data.</td>
<td>90</td>
</tr>
<tr>
<td>4.3</td>
<td>Results comparing the performance of MOFA-Cluster and other tools in estimating the significant features in terms of area under precision-recall curves using the simulated data.</td>
<td>91</td>
</tr>
</tbody>
</table>
Figure 4.4  Heatmap showing the percentage of variance shown explained by the different factors in the colorectal cancer subtype analysis ......................................................... 92
Figure 4.5  Boxplot showing the average number of mutations per patients in the different colorectal cancer subtypes .......................................................... 93
Figure 4.6  The boxplots showing the mean expression of the immune checkpoint genes in the colorectal cancer subtype C3 including PD1, PDL1, CTLA4, LAG3, IDO, and SLC43A3 .................................................................................. 94
Figure 4.7  Survival curves comparing the overall survival probability of the identified colorectal cancer subtypes ................................................................. 95
Figure 4.8  Survival curves comparing the overall survival probability of the male and female colorectal cancer patients .................................................. 96
Figure 4.9  Heatmap showing the percentage of variance shown explained by the different factors in the preeclampsia subtype analysis .............................................. 97
Figure A.1  a-e) Barplots showing the enrichment (-Log$_{10}$(P-Value)) of tissue-specific genes in ESCd 40-70, ESCd <40, PHTu, PHTd, and ESCu cells using protein atlas database. Placenta-specific genes are enriched in all cell types except ESCu. .................................................. 104
Figure A.2  Heatmap showing the significant DEGs (P-Adjusted ≤ 0.01 and abs(Fold-Change) ≥ 2) between the ESCu and ESCd groups. ESCd groups have few differentially expressed genes between one another. ........................................ 105
Figure A.3  Plot showing the enrichment of mesoderm development terms in DEGs over-expressed in ESCd or mesoderm. The dotted lines represent the thresholds for FDR (0.05) and fold enrichment (1.5) of the terms. ......................... 106
Figure A.4  PCA plot based on the 1,000 most highly expressed genes in ESCd. PCA was carried out using ESCd, ESCu, placental cell lines, term placental cells, and mesoderm cells. The first three principal components are plotted (cumulative variance of 81%). ................................................................. 106
Figure A.5  Heatmap showing the relationship between the co-expressed modules and cell type. Each block in the heatmap has the correlation score and adjusted p-values between the respective co-expressed module and cell type. .......... 107
Figure A.6  Box plots (left panels) showing expression values in ESCd and ESCu for genes in modules M6 (a), M14 (b), and M16 (c). Gene expression is significantly higher in ESCd. The corresponding right panels show the distribution of the number of genes upregulated in ESCd in 10,000 sets of random gene modules. ................................................................. 108
Figure A.7  Network showing the protein-protein interaction between the bottleneck genes and their neighbors for module M6. The color of the node represents the betweenness score of the genes.  

Figure A.8  PCA plot based on the 1,000 most highly expressed genes in ESCd. The first three principal components are plotted with a cumulative variance of 90%. PCA was carried out using the ESCd, ESCu, and term placenta cells.  

Figure A.9  Plot showing the expression of the identified novel genes in E7.5 and E9.5 mouse placenta. The error bars represent the variance in the samples. The dotted line represents the expression level of 5 Fragments Per Kilobase per Million mapped reads (FPKM).  

Figure A.10 Distribution of $\log_2$(TPM) for genes that are trophectoderm specific (left) and genes that have a highly significant bias in expression in polar trophectoderm (right).  

Figure A.11  Plots showing the $R^2$ values (left panel) and mean connectivity (right panel) of various soft threshold values required for making the co-expression network scale free. The threshold for the $R^2$ value is 0.9 shown as the red horizontal line.  

Figure A.12  Dendrogram showing clustering of the modules based on the correlation between the module’s eigenene expression.  

Figure A.13  Heatmap showing the gene overlap between co-expressed modules based on the module membership (KME) of genes.  

Figure C.1  (A) Cell doubling times performed on randomly selected cell lines (four CTL iPSC lines: MRuc4, MRuc7, MRuc8, MRuc10; six EOPE iPSC lines: MRucB, MRucE, MRucI, MRucJ, MRucK, MRucN). (B and C) PCA of transcriptomes of UC fibroblasts from EOPE and CTL pregnancies and iPSC lines derived from them. The sexes of the samples are designated by circles (female) and triangles (male).  

Figure C.2  (A) Comparison of TB invasion of 10 CTLs (CTL-TB, left: H1 and H9, lines 3 to 10) and 14 EOPE-TB (right: lines A to S) after exposure to BAP conditions for 6 d. (B) Combined TB invasion results for all CTL TB and EOPE TB lines. The log-transformed least-square means were used to determine significance. (C) TB cell proliferation as assessed by total DNA content. Two ESC (H1, H9*; *female line), four CTL TB (5*, 6, 7, 8*), and four EOPE TB (A*, J, M, R*) were measured for total DNA concentration after 6 d of BAP treatment. (D) The DNA content of TB cells (C) was log-transformed and plotted to determine whether there was a correlation with invasion cell counts (log-transformed values; counts) from A (values also shown in Table S4).
Figure C.3  Expression of (A) hCG, (B) P4, (C) PGF, and (D) sFLT-1 in 24-h culture between day 5 and day 6 following initiation of BAP-induced differentiation. 161

Figure C.4  PCA of RNAseq data from CTL TB and EOPE TB (GSE119265). Data are based on the log-transformed TPM from the top quartile of the most variable genes. ................................................................. 162

Figure C.5  Gene expression modules showing correlations in relation to invasive capacity, sex, O2 treatment, and disease status (EOPE vs. CTL) as determined by WGCNA. ................................................................. 163

Figure D.1  Overview of the PlacentaCellEnrich website and options available for analysis. 177

Figure D.2  Bar chart showing placenta cell-specific enrichment of the 1,000 most highly expressed genes in EVT (a) and syncytiotrophoblast (b) differentiated from TB stem cells. Data were analyzed in PlacentaCellEnrich using Vento-Tormo et al. dataset. ................................................................. 178

Figure D.3  Bar chart showing placenta cell-specific enrichment of the 1,000 most highly expressed genes in EVT (a) and syncytiotrophoblast (b) differentiated from TB stem cells. Data were analyzed in PlacentaCellEnrich using Suryawanshi et al. dataset. ................................................................. 179

Figure D.4  Bar chart showing the placenta cell-specific enrichment of the 1,000 most highly expressed genes in E7.5 (a), and E9.5 (b) mouse placenta. Data were analyzed in PlacentaCellEnrich using the Vento-Tormo et al. dataset. ....... 180

Figure D.5  Bar chart showing the placenta cell-specific enrichment of the 1,000 most highly expressed genes in E7.5 (a), and E9.5 (b) mouse placenta. Data were analyzed in PlacentaCellEnrich using the Suryawanshi et al. dataset. ....... 181

Figure D.6  Bar chart showing the placenta cell-specific enrichment of genes upregulated in antimesometrial (a), and mesometrial (b) mouse decidua. Data were analyzed in PlacentaCellEnrich using the Vento-Tormo et al. dataset. ....... 182

Figure D.7  Bar chart showing the placenta cell-specific enrichment of genes upregulated in antimesometrial (a), and mesometrial (b) mouse decidua. Data were analyzed in PlacentaCellEnrich using the Suryawanshi et al. dataset. ....... 183
ACKNOWLEDGMENTS

First and foremost, I would like to thank my major professor Dr. Geetu Tuteja for her guidance, support, and encouragement throughout my graduate life at Iowa State. I am fortunate to have a guide who always encourages new ideas and provides the freedom to explore and learn diverse research topics that helped me grow and develop as an independent researcher. I would also like to thank her for being patient with me and letting me learn by mistakes. I would also like to extend my gratitude towards my committee members Dr. Julie A. Dickerson, Dr. Karin S. Dorman, Dr. Heike Hofmann, and Dr. Xun Gu, for providing valuable suggestions and support. I would particularly like to thank Dr. Karin S. Dorman for providing me support and guidance throughout, especially during the last two years. Thank you for having belief in me more than myself and always encouraging me during our meetings.

I would like to thank our collaborators at the University of Missouri, Dr. R. Michael Roberts and Dr. Toshihiko Ezashi. I would also like to express my gratitude to the several staff members at Iowa State. I would like to thank the past and the current coordinators of the Bioinformatics and Computational Biology program, Trish Stauble and Carla Harris, for making my graduate journey smooth. Thank you, Trish, for always listening to all my problems and giving me insightful advice. I would also like to thank the Genetics, Development, and Cell Biology department staff members, including Danise Jones and Diane Jespon, for their support.

I would like to thank Tuteja lab members Dr. Haninder Kaur, Rebekah Starks, Anilisa Biswas, Ha Vu, Ashwini Rangaraj, and Kelby Kies for their constant support and help. I would also like to thank the former lab members, including Majd Abdulghani, Rabab Abu Alhasan, and Amanda Loke. During my graduate life, I met a lot of amazing people who supported me throughout my journey. Thank you, Urminder, Priyanka, Sagnik, Sambit, Gokul, Arun, and Manjil, for making my time in Ames pleasant.
Lastly, I would like to thank my family members for always encouraging me and supporting me. I don’t think that I would have reached this point in my life without their support. I would like to thank my family members Shashi Jain, Kuldeep Rai Jain, Ankush Jain, Ankit Jain, Nidhi Jain, and Shagun Jain. I also want to mention how the little ones in my family, my nephews and niece, Samkit, Sprash, and Nishtha, always cheered me up over the video calls whenever I felt low. My dissertation is dedicated to my family for their encouragement and support throughout this journey.
ABSTRACT

The placenta is a temporary organ that plays a significant role in the transportation of vital nutrients, gases, pregnancy hormones, and other essential products between the mother and the fetus during pregnancy. It is also involved in various other critical functions, including anchoring the conceptus and fetus immune protection. Throughout pregnancy, the placenta alters its function to cater to the needs of the growing fetus. Aberration in the expression of placental development genes during pregnancy leads to its abnormal structure and function. These abnormalities result in several pregnancy-related complications, including intrauterine growth restriction (IUGR), preterm labor, low birth weight, and preeclampsia (PE). PE causes severe damage to both mother and baby. Many studies suggest that abnormalities leading to early-onset PE start during the first few weeks after conception but can only be detected after 20 weeks. Therefore, it is crucial to study regulatory mechanisms during early placental development. However, this area is not explored very well due to restrictions and lack of early placental cell model. In my thesis, I studied the regulatory mechanisms behind early placenta development by developing computational methods and tools to analyze various next-generation sequencing (NGS) datasets.

First, I analyzed transcriptomics data generated in early trophoblast cells derived from human embryonic stem cells and identified a gene regulatory network specific for the early trophoblast development using co-expression network analysis. Second, I developed a bioinformatics tool that carries out tissue-specific gene enrichment using the hypergeometric test. Third, I developed a statistical method to integrate and cluster gene expression and DNA methylation datasets simultaneously to identify PE subtypes using factor analysis and Gaussian mixture models. Altogether, these studies and methods will help understand the molecular mechanisms regulating critical processes during early placental development.
CHAPTER 1. INTRODUCTION

The placenta is a transient organ required for fetal development and maintenance of pregnancy. In all placental mammals, it plays a significant role in the well-being of both the mother and fetus during pregnancy. It acts as an interface between the mother and the baby and helps in the transport of nutrients and gases and waste elimination (Gude et al., 2004). The placenta is also involved in metabolism, producing a wide variety of essential products, including hormones and growth factors required by both mother and fetus (Benirschke, 1994). It also anchors the fetus to the uterine wall and provides immune protection (Maltepe et al., 2010)(Marin et al., 2003). In primates and rodents, the placenta is hemochorial, which means that it is in the direct contact with the maternal blood and acts as an exchange membrane between mother and fetus (Carter, 2007)(Rossant and Cross, 2001). Problems in the proper development of the placenta cause pregnancy complications, including preeclampsia, intrauterine growth restriction, and miscarriage (Goldenberg et al., 2008)(Monk and Moore, 2004)(Williams and Broughton Pipkin, 2011). Despite the importance of the placenta in pregnancy, this organ has not widely been studied.

1.1 Human placental development

The placenta is a temporary organ that is important for the development of fetus during pregnancy. In humans, the placenta develops from the trophectoderm, the outer layer of the blastocyst, approximately five days after conception. After the implantation of the blastocyst to the uterine wall after ~ 7 days, the initial syncytium invades the uterine wall inducing the epithelium cells to form cavities, called lacunae, which help nutrient transport by facilitating in the movement of maternal blood (Hertig et al., 1956). These cavities are further transformed into a specialized tissue called decidua during pregnancy, which is essential for the development of the embryo and helps in its attachment to the uterine wall (Schlafke and Enders, 1975). The conceptus moves into the stro-
mal region att about 12th day of pregnancy through the uterine epithelium. This syncytial mass
and underlying cytotrophoblasts (CTB) surround the embryo proper and are believed to serve as a
primitive placenta (Gude et al., 2004)(James et al., 2012). After a few days, the columns of CTB
push through the syncytial layer to establish primary villi, which will eventually branch, acquire
cores of blood vessels and connective tissue, and create the early villous placenta (Blair et al.,
2013)(James et al., 2012). At the anchoring villi tips, villous CTB (vCTB) continue to divide to
form an invasive extravillous trophoblast (EVT) population that invade further into the uterine
wall. Some of the vCTBs also enter maternal spiral arteries to alter the blood flow and oxygen ten-
sion in the placenta. These changes alter the expression of the transporter proteins involved in the
transportation of various nutrients and hormones between the mother and the fetus (Gude et al.,
2004)(James et al., 2012). The vCTB at the tip of the anchoring villi proliferate, differentiate, and
fuse to form the syncytiotrophoblasts (STB), the outer layer of placental chorionic villi (Benirschke
et al., 2012). This syncytial cellular layer of the placental chorionic villi is in direct contact with
the maternal blood, mediating the exchange of materials from the mother to the fetus.

1.2 Trophoblast cells

Trophoblast (TB) cells are specialized cells which are the structural and functional components
of the placenta. They invade the maternal decidua and uterus and form the passage to transport
nutrients between the mother and fetus. They are also involved in the production and secretion
of various hormones and growth factors in placental development (Knofler and Pollheimer, 2013).
The term “trophoblast” originates from a Greek word “threphein”, meaning “to feed” and was
first used by Dutch embryologist Dr. Hubrecht to describe cells involved in nutrient transportation
between the mother and baby during pregnancy (Pijnenborg and Vercriuysse, 2013). In his research,
Dr. Hubrecht also found these cells to be highly invasive, and their growth is supported by the
decidua (Hubrecht, 1889). These initial findings have led to many different studies identifying the
various subtypes of the TB cells. Currently, there are three significant subtypes of trophoblast
cells, including CTB, STB, and EVT.
The CTB are the primary cells that enclose the embryo during the early stages of pregnancy. During the early stages, these cells invade the early syncytiotrophoblast to give rise to the villous CTB (vCTB) that forms the placenta villi. These cells readily proliferate throughout pregnancy and give rise to various types of TB cells (Figure 1.1). Due to their continuous proliferation, similar to cancer cells, they are often called the “germinative” TB layer (Simpson et al., 1992). These cells also express various known TB transcription factors including, GATA2, GATA3, TEAD4, and TFAP2C/A (Krendl et al., 2017). In the first 2-3 weeks of pregnancy, these vCTB differentiate to give rise to the EVT. The STB are the multinucleated cells responsible for the transportation of nutrients, hormones, and gases between the mother and baby. These cells are reported to be differentiated from the vCTBs through the fusion of the cells. During differentiation, the vCTBs are thought to exit the cell cycle to fuse together and give rise to STBs (James et al., 2012). It was found that during this differentiation, the expression of the fusion genes syncytin 1 and syncytin 2.

Figure 1.1 Differentiation of the trophoblast cells during pregnancy.
is induced by the upregulation of the GCM1 gene through the protein kinase A (PKA) signaling (Liang et al., 2010). The STBs form the maternal-fetal interface through which the exchange of the essential elements takes place for fetal growth and development. They are in direct contact with the maternal blood and form the outer lining of the placental villi. They also carry out endocrine functions by producing various proteins and hormones that help maintain metabolic and physiological balances during pregnancy (Rossant and Cross, 2001). In addition to that, STBs also protect the fetus from the circulating maternal immune cells by forming a protective immunological barrier (Delorme-Axford et al., 2014).

The EVTs are the invasive cells that are differentiated from CTBs and migrate towards the decidua. The EVTs are non-proliferative cells and are bigger in size than CTBs due to the endoreduplication process. The EVTs have a very different transcriptional profile with the upregulation of many endothelial marker genes, including VE-cadherin and MCAM (Zhou et al., 1997). There is also an upregulation of metalloproteinases, including MMP2, MMP3, and MMP9, that helps in the invasion and migration of these cells (Xu et al., 2000). The upregulation of the invasive genes is accompanied by the downregulation of many receptors of the growth factors in EVTs, leading to the loss of proliferation (Genbacev et al., 2000). There are two types of EVT cells based on their migration pathways. The EVTs that migrate towards the maternal arteries through the decidua stroma are called interstitial EVT (iEVT). Those that migrate into the maternal arteries are called endovascular EVT (eEVT). The iEVTs invade the myometrium up to its third layer, and then they fuse together to form the multinucleated TB giant beds. These cells also produce placental lactogen that helps in the attachment to the uterine wall (Al-Lamki et al., 1999). The eEVT invading the blood vessels first transform the spiral arteries to increase the surface area, then migrate towards the arterial opening and plug it so that the maternal blood cannot circulate inside the placenta. These plugs remain until the end of the first trimester, after which they are removed, and the flow of maternal blood is established into the placenta (Kam et al., 1999).
1.3 Models to study early human placental development

Most current human studies are based on data from the term placenta, due to restrictions on working with human placenta and the previous lack of availability of proliferating early placental cells. Therefore, most of our understanding of early placental development has come from the rodent models (including mouse and rat) (Rossant and Cross, 2001)(Carter, 2007) and immortalized cell lines representing the different TB (Graham et al., 1993).

1.3.1 Animal Models

There is great diversity among mammals in terms of placental development, particularly in terms of the degree of TB cell invasion into uterine tissues (Roberts et al., 2016). The laboratory mouse and primates have a hemochorial placenta, similar to humans. In contrast, in hemochorial placentation, TB cells invade the uterine tissue such that trophoblasts are bathed with maternal blood. In endotheliochorial placentation, TB invasion is not deep enough to penetrate the maternal blood vessels, and in the epitheliochorial placenta, no erosion of uterine epithelium occurs at all (Carter, 2007)(Roberts et al., 2016). There is no perfect experimental model system representing human placentation, even in species with a hemochorial placenta. Many pregnancy disorders, including PE, caused by placental defects, are only found in humans and possibly in apes (Carter, 2011). Due to these differences, multiple models are used to study placental development depending on the research question. The laboratory mouse model has a hemochorial placenta, and many of the TB subtype cells are similar to that of humans. The mouse model also has the advantage of a short gestation period, small size, and easy availability (Rossant and Cross, 2001) (Cox et al., 2009). However, there are also many dissimilarities between the mouse and human placenta. In humans, decidualization of the stromal cells is cyclic and is not dependent on pregnancy, whereas, in mice, it starts when the embryo is present (Gellersen and Brosens, 2014). Also, in humans, the polar trophectoderm is attached to the uterine wall. In contrast, in mice, the mural trophectoderm is attached to the uterine first, followed by the polar one resulting in the formation of the ectoplacental cone (Georgiades et al., 2002). Due to these differences in the TB development between human
and mouse, findings in mouse studies require further verification in the human model system. Despite all these differences, much progress in deciphering of molecular mechanisms of placental development have been made by studying mouse models (Rossant and Cross, 2001) (Woods et al., 2018). Although, rat is often a preferred model for studying maternal blood vessel transformation, as the TB cells invade deeper into the uterine tissue compared to other rodent models (Soares et al., 2012).

1.3.2 Cell lines

Due to various restrictions and high costs of working on the human placenta, different TB cell lines have been used for studying the placenta development. The cell lines are mostly used for studying the invasion and differentiation of the TB cells during the progression of pregnancy. Some of the TB cell lines used are described below.

1.3.2.1 Choriocarcinoma cell lines

Many of the TB cell lines are derived from malignant tumors or choriocarcinomas, including BeWo, JEG3, and JAR cell lines (Kohler and Bridson, 1971) (Apps et al., 2007) (Hertz, 1959). These cell lines are thought to be good models for TB cells as they have high expression of TB marker genes, including GATA3, KRT7, and TFP2C2 (Lee et al., 2016). They also show a unique pattern of expression for various histocompatibility complex (HLA) family genes (Lee et al., 2016). BeWo cells are used to study molecular mechanisms behind syncytialization, whereas the JEG3 cells are used to understand TB invasion (Hannan et al., 2010) (Rothbauer et al., 2017). BeWo cells are derived from a metastatic gestational carcinoma, cultured in the cheek of a hamster around 300 times. The JEG3 lines are further derived from the BeWo cells. Both are multinucleated cells used as a model for STB cells (Hertz, 1959) (Kohler and Bridson, 1971). JAR cells are also derived from the human placental choriocarcinoma but these have no expression of HLA genes similar to that of vCTBs (Apps et al., 2007). Still, these cell lines have significantly different genome-wide DNA methylation profiles compared to the first trimester TB cells (Apps et al., 2011). Although these
cell lines are cost-effective and readily available, findings from any of these cell lines require further validation.

1.3.2.2 Placental explants and primary TB cells

The development of TB cells during the early placental development has also been studied using explants of the placenta and the isolated primary TB cells. Placental villi are used to prepare the placental explant, which is adhered to plastic or an extracellular matrix (Male et al., 2012). The primary TB cells are generally isolated from early pregnancy timepoints, not from the term placenta, due to the low concentration of STB and VCT cells in the placental villi. During the first trimester, the placental villi have high concentrations of STB at the tip and vCTB in the inner layer (Male et al., 2012). These cells can be isolated using enzymatic digestion from the placenta villi. Although these cells are a good model for studying TB development, their use is limited because they cannot proliferate in vitro, are sensitive to contamination, are highly variable between samples, and ethical restrictions. To overcome these limitations, a TB cell model known as HTR-8/SVneo has been developed from primary TB cells transfected with the gene encoding the simian virus 40 large T antigen (Graham et al., 1993). The HTR-8/SVneo cells are morphologically similar to TB cells and express TB marker genes (Chakraborty et al., 2002). These cells are readily used to study the regulation of TB cell invasion and migration. However, Abou-Kheir et al. (2017) have raised questions about the purity of these cells and reported that the HTR-8/SVneo cells contain a mixture of TB and stromal cells.

1.3.2.3 Pluripotent human embryonic stem cells

The first study that reported the differentiation of human embryonic stem cells (hESCs) into TB cells using Bone morphogenic Protein 4 (BMP4) was in 2002 (Xu et al., 2002). Since then, several other studies have found that culturing with just BMP4 leads to a mixed population of mesoendoderm and TB cells (Yu et al., 2011)(Zhang et al., 2008). A two-step system was also proposed in which hESCs can be first differentiated into primary CTB like cells that can further
be differentiated into STB and EVT like cells (Horii et al., 2016). After trying several media combinations, it was found that the hESCs can be differentiated into TB cells by culturing them in the presence of Bone morphogenic Protein 4 (BMP4), inhibitors of fibroblast growth factor-2 (FGF2) and Activin A signaling, which also inhibits the formation of mesoendoderm cells (Amita et al., 2013). On further investigation of these cells, it was found that the differentiated cells have physical features similar to syncytioTBs and also show high expression of TB marker genes, including KRT7, GATA2, and TCFAP2A (Yabe et al., 2016). It was also reported that these cells express human chorionic gonadotrophin that is specific to syncytioTBs and also an EVT marker gene HLA-G (Amita et al., 2013)(Yabe et al., 2016). Although the differentiated cells show expression of many TB markers and exhibit TB cell morphology, the exact identity of the cells is still unclear(Yabe et al., 2016). However, one hypothesis that was proposed based on the present findings states that these cells represent early post-implantation TB cells (Jain et al., 2017)(Roberts et al., 2018).

1.3.2.4 Human trophoblast stem cells (hTSCs)

Many attempts have been made to develop an in-vitro model to study TB development by isolating the human TSCs from early placental development (Kunath et al., 2014)(Soncin et al., 2015). Okae et al. (2018) has made very promising progress in this field by isolating the hTSC cell lines from both the embryo and the first-trimester placental cells. The isolated TSCs cells seem to be stable in terms of their genetic properties, show the physical properties of TB cells, and have high expression of TB gene markers. These cells are quite specific to early placental development and could not be derived from the term placenta due to the loss of invasive capability of the term TB cells and the low number of EVT cells (Mayhew, 2014). In their research, Okae et al. (2018) also show that the hTSCs can be cultured with specific conditions that give rise to the syncytial and extravillous lineages. It was also found that the hTSCs have similar expression patterns to that of the VCT, but the exact identity and location of these cells during pregnancy is still unknown (Okae
et al., 2018). Further investigation is needed to properly explore the identity and relationship of these cells with early TB cells.

1.4 Complications due to placental disorders

Most complications during pregnancy are reported to originate from the abnormal development of the placenta during early pregnancy (Smith, 2010). These complications include preeclampsia (PE), Intrauterine growth restriction (IUGR), preterm labor, and low birth weight (Monk and Moore, 2004)(Burton et al., 2009)(Lim et al., 1997). These diseases cause a high percentage of maternal and fetal mortality in the United States (Berg et al., 2010)(Mol et al., 2016). The major causes of pregnancy complications include poor TB invasion into the uterus, decidual vasculopathy, aberrant transformation of uterine blood vessels, and defective development of chorionic villi (Mol et al., 2016)(Drăgușin et al., 2018)(Zhang, 2018). These abnormalities are associated with the dysregulation of gene expression in TB during early development (Smith, 2010). In normal placental development, the TB cells first invade into the decidua and remodeled the maternal arteries to establish the maternal blood flow. The remodeling also increases the surface area better for the transportation of the essential nutrients, oxygen supply, and waste removal (Gude et al., 2004)(Knofler and Pollheimer, 2013). Improper remodeling of the maternal blood vessels can lead to unbalanced blood flow into the intervillous space, which results in insufficient transportation of the nutrients and oxygen to the placenta, reducing the surface area for exchange and villous tree branching, leading to IUGR and stillbirth (Monk and Moore, 2004)(Burton et al., 2009). Impaired blood vessel transformation can also cause premature separation of the chorionic membranes leading to premature labor or placental abruption (Buhimschi et al., 2010). It has also been reported that a lot of angiogenic factors are detected in the maternal blood that are released by the placenta during PE and are responsible for vascular dysfunction (Burton et al., 2019). These factors include an increase in the antiangiogenic proteins such as soluble fms-like tyrosine kinase 1 (sFlt-1) and soluble endoglin, that are involved in the inhibition of the proangiogenic proteins such as placental growth factor (PIGF) (Zeisler et al., 2016).
1.5 Multi-omics analysis to study Preeclampsia

PE is one of the major pregnancy complications that affects around 4% to 5% of pregnancies in the United States, with a fetal mortality rate of about 15% (Berg et al., 2010). The symptoms include hypertension and a high level of proteinuria (Mol et al., 2016). It has previously been proposed that early-onset PE is caused by shallow TB invasion during the early stages of pregnancy, causing low maternal blood transportation to the placenta (Huppertz, 2008). These defects cause a deficiency of essential nutrients and hormones, leading to many different pregnancy complications (Huppertz, 2008) (Mol et al., 2016). Due to the exponential growth in the usage of NGS techniques in the last decade, various studies have tried to elucidate the underlying regulatory networks and mechanisms behind PE. This section reviews the studies that integrate NGS datasets using different methods to decipher molecular mechanisms behind PE.

Many of the studies investigating PE use one type of omics data to make a general conclusions and then verify the results either on a small subset of samples or with omics data from other samples (Liu et al., 2013) (Pennings et al., 2011) (Than et al., 2018) (Xiang et al., 2013). In one of those studies, Than et al. (2018) used gene expression data to identify specific target genes important in PE using differential expression analysis and co-expression network analysis. They further verified the importance of the discovered genes in PE by checking the protein levels in maternal serum and methylation level using DNA methylation data. There are few PE studies that have collected multiple omics data from the same placenta samples, as done extensively in The Cancer Genome Atlas (TCGA) (Blair et al., 2013) (Leavey et al., 2018) (Martin et al., 2015) (Xuan et al., 2016). In one study, Blair et al. (2013) identified methylation sites specific for early-onset PE using gene expression microarrays and verified concomitant changes in gene expression of the target genes (Blair et al., 2013). Martin et al. (2015) identified 123 genes with differential expression at both the transcript and DNA methylation level. They found enrichment for the transforming growth factor-beta (TGF-β) signaling pathway, a known TB invasion and migration pathway. Finding the molecular mechanisms of PE is difficult because it is a very heterogenous disease. In fact, some have hypothesized the existence of multiple subtypes in PE depending upon its progression during
pregnancy (Cox et al., 2011)(Huppertz, 2008)(Redman, 2014). Only a small number of DEGs are typically found between the control and PE samples further suggesting great disease heterogeneity (Kleinrouweler et al., 2013)(Tsai et al., 2011). In their study, Leavey et al. (2016) identified three different PE subtypes by clustering samples based on the gene expression data. Examination of DNA methylation patterns found correlations with the differentially expressed genes and provided more information about the PE subtypes (Leavey et al., 2018). Although this is a good start to identify PE subtypes, the results can be further strengthened by identifying PE clusters after data integration. Next we will discuss such methods, that carry out integrative clustering of multi-omics data.

1.6 Multi-omics data integration using matrix factorization

During the last decade, there has been an exponential growth in the usage of NGS techniques due to the drop in the sequencing costs and development of better analysis methods (cos, 2020). These changes have enabled researchers to comprehensively study genome-wide regulatory mechanisms by collecting multi-omics datasets, multiple types of data observed on the same sampled individuals (Chakraborty et al., 2018)(Harber et al., 2019). There have been various statistical approaches developed for integrating multi-omics datasets while dealing with several challenges, including the multiple datasets, the high dimension of the datasets, the small sample size, and interactions between different omics datasets (Meng et al., 2016b)(Pierre-Jean et al., 2019). Matrix factorization is one such method that integrates multi-omics data by reducing the high dimensional data into a few latent variables or factors (Meng et al., 2016b). In this technique, each matrix of input data (from M different high throughput assays) is decomposed into a weight matrix $W_m$ and a common latent variables matrix $Z$. The general approach can be represented by the equations:

$$Y_m \approx Z \ast W_m, \text{ for } m = 1 \ldots M,$$

where $Y_m$ is the input matrix with dimensions $N$ (number of samples) x $D_m$ (number of features), $W_m$ is the weight matrix with dimensions $K$ (latent factors) x $D_m$, and $Z$ is the latent variable matrix with dimensions $N \times K$. As seen from the equation, matrix factorization attempts to
explain the data through linear combinations of a few \((K)\) shared factors. One implication is that the approach is not capable of detecting non-linear relationships between the features and underlying factors within or across assays. Although the approach has several limitations, it is still very effective in integrating multi-omics data and has been used in many data integration tools (Pierre-Jean et al., 2019).

After data integration, the factors \((Z)\) have also been used to find clusters in the datasets to identify disease, tissue or other subtypes. Various tools have been developed that carry out integrative clustering of the multi-omics datasets (Pierre-Jean et al., 2019). They all carry out integrative clustering sequentially, first identifying the latent factors and then, conditional on the found latent variables, carrying out clustering. This approach is considerably easier than clustering the data in the original high dimensional space. There are a multitude of clustering methods (Aggarwal and Reddy, 2014) that can cluster the discovered factors in the low \(K\)-dimensional factor space.

1.6.1 iCluster, iClusterPlus, and iClusterBayes

The initial iCluster method was developed to identify subtypes in the TCGA dataset by integrating multiple sequencing datasets. Shen et al. (2009) developed a method that integrates the continuous datasets, including gene expression, protein expression, and DNA methylation, by capturing the correlation between them in a low-dimensional space. Specifically the model assumes that each of the genomic datasets is conditionally independent given the latent variables. The model parameters are estimated using the Expectation-Maximization (EM) algorithm, and sparsity is induced by applying a soft threshold. The iCluster method is improved further in iClusterPlus to jointly model both continuous and discrete data (Mo et al., 2013). The method uses a modified Monte Carlo (MC) Newton-Raphson algorithm to estimate the model parameters with sparsification by using a lasso penalty (Mo et al., 2013). The main problem with iClusterPlus is the computational cost of choosing appropriate shrinkage parameters. The search involves running the
model many times. To improve this method, Mo et al. (2018) developed iClusterBayes, which takes a fully Bayesian approach to estimate model,

\[ Y_m = X \ast \Gamma_m \ast \beta_m + \epsilon_m, \]

where \( X \) is matrix of \( K \) factors plus a first column of 1s for the mean, \( \beta_m \) is a loading matrix of the coefficients of each feature in \( m \)th dataset, and \( \Gamma_m \) is a diagonal matrix where the first diagonal component is 1 and others are indicator variables (0 or 1) for feature selection and sparsification in the Bayesian model. iClusterBayes assumes factor matrix \( X \) follows a standard multivariate normal distribution, implying a classic factor analysis model (Anderson, 1984). The datasets in this model are conditionally dependent on the latent variables, which helps in the inference of the model parameters using Markov Chain Monte Carlo (MCMC) (Mo et al., 2018). Like iClusterPlus, iClusterBayes can model datasets with both continuous and discrete data. However, as this is a fully Bayesian model, there is no need to search for optimal shrinkage parameters, making it more computationally efficient (Mo et al., 2018). Finally, clustering is carried out on the estimated latent variable by \( k \)-means. The iCluster methods are available as R packages iCluster and iClusterPlus in Bioconductor.

### 1.6.2 MoCluster

MoCluster is another method that uses a matrix factorization technique to integrate multi-omics data (Meng et al., 2016a). It uses a modified consensus PCA (Westerhuis et al., 1998) to integrate the multi-omics data \( (Y_m) \) by defining joint latent variables (JLVs). JLVs are linear combinations of the principal components (PCs) identified from each of the datasets. In MoCluster, the JLVs are identified iteratively. At the \( k \)th iteration, the \( k \)th PC \( (Z_{km}) \) is identified for each of the \( M \) datasets. Then, the \( k \)th JLV is found by maximizing the covariance between the linear combination of the identified PCs and the \( k \)th JLV,

\[
\max_{w_{km}, Z_k} \text{Cov} \left( \sum_{m=1}^{M} w_{km} Z_{km}, Z_k \right).
\]
The maximization over JLVs is carried out using the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm (Helland et al., 1992). MoCluster requires the datasets to be pre-processed before carrying out the integration, including feature centering and scaling. MoCluster differs from traditional consensus PCA as it introduces sparsity in the feature coefficient using a soft thresholding operator on each of the JLV components. This helps to ease biological interpretation of the results from the data integration results. The JLVs are further used to identify clusters by using traditional clustering methods, including \( k \)-means and hierarchical clustering. MoCluster is available as R package mogsa on CRAN (Meng et al., 2016a).

1.6.3 Multi-Omics Factor Analysis (MOFA)

MOFA is one of the latest methods developed to integrate multi-omics data using the MF technique (Argelaguet et al., 2018). This method uses the earlier matrix factorization,

\[
Y_m = Z * W_m + \epsilon_m,
\]

with loading matrix \( W_m \) and latent variable matrix \( Z \) but further assumes normal error \( \epsilon_m \). The latent variables represent the individual datasets in the lower \( K \)-dimensional space, the loading matrix represents the weights of the omics features in the lower dimensional space, and the residual noise accounts for measurement error in the features from different omics data. Very similar to iClusterBayes, MOFA uses a Bayesian framework to estimate the model parameters, but instead of using MCMC to estimate the posterior distribution, it uses a Variational Bayes algorithm (Fox and Roberts, 2011). MOFA adds sparsity in both the loading and latent variable matrices using the spike-and-slab and Automatic Relevance determination priors (Mitchell and Beauchamp, 1988)(MacKay, 1996). In addition to modeling the continuous data using the Gaussian distributions, MOFA also supports binary and count data using Poisson and Bernoulli distributions. MOFA is available as an R package on Bioconductor (Argelaguet et al., 2018).

All of these tools first carry out data integration by reducing the data dimensionality and then cluster the samples using the low-dimension data. The data reduction step ignores the cluster structure in the data, which may lead to bias in the factor estimates and may remove the cluster
signal during data reduction (Ghahramani and Hinton, 1996). There is an opportunity for a tool that can simultaneously estimate the low-dimensional representation and cluster multi-omics data.

1.7 Dissertation Organization

Chapter 1 provides a general background of the work that is presented in this thesis. It includes a general introduction about human placental development, models to study placenta development, pregnancy complications related to the placenta, previous multi-omics studies in PE, and methods to integrate multi-omics data.

Chapter 2 consists of a published manuscript titled “Deciphering transcriptional regulation in human embryonic stem cells specified towards a trophoblast fate” published in Scientific Reports in 2017 (Jain et al., 2017). In this work, the gene expression profile of the BMP4 treated hESC cell model is compared with mesoderm and other placental cell lines using RNA-Seq data to determine if they truly represent a trophoblast subtype. The transcriptional regulation of these cells was also investigated by using co-expression network analysis.

Chapter 3 consists of a published manuscript titled “TissueEnrich: Tissue-specific gene enrichment analysis” published in Bioinformatics in 2018 (Jain and Tuteja, 2018). In this work, TissueEnrich, a tool to carry out tissue-specific gene enrichment in an input gene set, was developed. It was also demonstrated that TissueEnrich can assign tissue identities to differentiated embryonic stem cells and single-cell clusters identified from single-cell RNA-Seq data.

Chapter 4 consists of a manuscript in preparation titled “Factor analysis-clustering model for integrating and clustering multi-omics dataset” for Bioinformatics. In this work, a statistical method is developed that simultaneously integrates and clusters multi-omics data using matrix factorization and a Gaussian mixture model. This method uses the Variational Bayes algorithm to identify the latent factors and clusters simultaneously in the samples from the multi-omics dataset. This tool is used to identify the PE subtypes by integrating gene expression and DNA methylation datasets.

Chapter 5 consists of the conclusions of the thesis, highlighting the essential takeaways from the thesis along with suggestions for future work.
In addition to that, the dissertation includes some appendices consisting of supplementary materials of the chapters and some of the other work that was carried out alongside the thesis.

**Appendix A** consists of supplementary material for Chapter 2

**Appendix B** consists of supplementary material for Chapter 4

**Appendix C** consists of a published manuscript titled “Early onset preeclampsia in a model for human placental trophoblast” published in *Proceedings of the National Academy of Sciences* in 2019 (Sheridan et al., 2019). In this work, the effects of oxygen stress on the molecular mechanisms involving trophoblast invasion in normal and early-onset PE (EOPE) cases, using the early trophoblast cell model from Chapter 1. It is found that under stressful oxygen conditions (20% \(O_2\)), the invasiveness of EOPE TB is reduced compared to the control. A weighted correlation network analysis also revealed two gene modules that, in control samples, are significantly linked to the extent of TB invasion but not in EOPE. This is collaborative work with Dr. Roberts’ lab from University of Missouri. I contributed to the design and implementation of the computational analysis in this study. I carried out various computational analyses, including RNA-Seq data processing, differential gene expression analysis, ontology enrichment analysis, and co-expression analysis. I also contributed to the interpretation of the results and writing the manuscript.

**Appendix D** consists of a published manuscript titled “PlacentaCellEnrich: A tool to characterize gene sets using placenta cell-specific gene enrichment analysis” published in *Placenta* (Jain and Tuteja, 2021). In this work, PlacentaCellEnrich, a tool to carry out the placenta cell-specific gene enrichment analysis, is developed. It has also been demonstrated that PlacentaCellEnrich can be used to understand the expression patterns of placental cells in other model systems using human placenta single-cell RNA-Seq data.

### 1.8 References


CHAPTER 2. DECIPHERING TRANSCRIPTIONAL REGULATION IN HUMAN EMBRYONIC STEM CELLS SPECIFIED TOWARDS A TROPHOBLAST FATE

Modified from a manuscript published in *Scientific Reports*

Ashish Jain$^{1,2}$, Toshihiko Ezashi$^3$, R. Michael Roberts$^{3,4}$, and Geetu Tuteja$^{1,2}$

$^1$Bioinformatics and Computational Biology, Iowa State University, Ames, IA, USA

$^2$Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, USA

$^3$Division of Animal Sciences, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

$^4$Department of Biochemistry, University of Missouri, Columbia, MO, USA

2.1 Abstract

Differentiated human embryonic stem cells (hESC) continue to provide a model for studying early trophoblast cells (TB), but many questions have been raised regarding their true identity. Therefore, we carried out a global and unbiased analysis on previously published transcriptomic profiles for hESC differentiated to TB by means of bone morphogenetic protein-4 and inhibitors of activin A and fibroblast growth factor-2 signaling (BAP treatment). Our results confirm that BAP treated hESC (ESCd) lack a mesoderm signature and are a subtype of placental cells unlike those present at term. ESCd display a high level of expression of genes implicated in migration and invasion compared to commonly used, immortalized TB cell lines and primary cells from term placenta. Co-expression network analysis also identified gene modules involved in cell migration and adhesion, processes that are likely critical during the beginning stages of placentation. Finally, protein-protein interaction analysis predicted several additional genes that may play important
roles in early stages of placental development. Together, our analyses provide novel insights into the transcriptional programs that are active in ESCd.

2.2 Introduction

The placenta is a transient organ required for fetal development and maintenance of pregnancy. In all placental mammals, it plays a major role in the transport of nutrients, gases, waste and hormones between the mother and fetus (Rossant and Cross, 2001). The placenta also anchors the fetus to the uterine wall and provides immune protection (Rossant and Cross, 2001). Trophoblast cells (TB), a cell lineage that first emerges as a simple epithelium, called trophectoderm, at the blastocyst stage of development, is involved in each of these functions. In the case of the human, implantation quickly follows blastocyst attachment to the uterine wall (James et al., 2012) and appears to involve invasive syncytial TB formed ahead of a layer of progenitor TB (Carter, 2007)(Enders, 2000). By about day 12 of pregnancy, the conceptus has moved through the uterine epithelium and into the stromal region. This syncytial mass and underlying cytotrophoblast (cytoTB) completely surround the embryo proper and are believed to serve as a primitive placenta (James et al., 2012). Within days, however, columns of cytoTB have pushed through the syncytial layer to establish primary villi, which will eventually branch, acquire cores of blood vessels and connective tissue, and create the early villous placenta (Gude et al., 2004). These villi are covered by a different kind of syncytium, which consists of a thin multinuclear cellular layer formed from fusion of underlying cytoTB (Benirschke et al., 2012)(Huppertz et al., 1998). Some of these columns of cytoTB form anchoring villi. At their tips, cells continue to divide to form an invasive extravillous TB (EVTB) population that invade further into the uterine wall. Some also enter maternal spiral arteries to alter their blood flow characteristics. Aberrant gene expression in TB during early development is associated with abnormal placental function, which can potentially lead to pregnancy-related complications including the early onset form of preeclampsia (PE), intrauterine growth restriction (IUGR), preterm labor, and low birth weight (Lim et al., 1997)(Monk and Moore, 2004)(Burton et al., 2009)(Huppertz, 2008). Human TB from first trimester placenta are difficult to obtain and
As a result, several other model systems have been used to study TB development, including rodent models (Rossant and Cross, 2001) and immortalized cell lines established from choriocarcinoma cells and first trimester EVTB (Graham et al., 1993). Although these models are extensively used, they each have their limitations and may not be appropriate for studying early human TB function (Bilban et al., 2010) (Soncin et al., 2015). To address this, over the last decade many groups have tried to reprogram human pluripotent cells into TB. Xu et al. first reported that human embryonic stem cells (hESC) can be differentiated into TB after being treated with bone morphogenetic protein-4 (BMP4) (Xu et al., 2002). Since then, many groups have studied this differentiation by varying the hESC culture conditions (Xu et al., 2002) (Li et al., 2013) (Amita et al., 2013) (Horii et al., 2016). Although most of the studies concluded that the differentiated cells are a subtype of TB, the true identity of these cells has been debated (Roberts et al., 2014) (Bernardo et al., 2011) (Lee et al., 2016b). To try to address this controversy, Yabe et al. generated RNA-Seq data from BAP treated (cells treated with BMP4, inhibitors of activin A signaling, and inhibitors of fibroblast growth factor-2 (FGF2) signaling) hESC (herein referred to as ESCd); as well as cells derived from primary cultures prepared from term placentas (Amita et al., 2013) (Yabe et al., 2016). Using principal component analysis (PCA) and a panel of TB and mesoderm markers, they determined that, while ESCd do not represent TB of term placenta, they do express a full array of most known TB marker genes (Yabe et al., 2016). Because there are limitations to using specific gene markers to determine the identity of ESCd (Roberts et al., 2014) (Lee et al., 2016b), we decided that it would be important to carry out an unbiased analysis of the transcriptome profile of these cells. Accordingly, we have used the RNA-Seq data generated by Yabe et al. to carry out a global and unbiased analysis of ESCd and compared the transcriptional profiles of these cells with publicly available RNA-Seq data from cells of mesodermal lineage, as well as TB cell lines derived from choriocarcinomas considered to represent villous cytoTB (BeWo and JEG-3) and EVTB (HTR-8/SVneo), to better understand the identity of the ESCd. Our analyses, which included functional enrichment analysis, PCA, and differential gene expression analysis, strongly indicate that ESCd are representative of invasive TB cells. We further analyzed
the relationship between the genes in ESCd by carrying out co-expression network analysis, and used protein-protein interaction networks to look deeper into the gene regulatory mechanisms and predict novel genes that may be important for early placental development.

2.3 Methods

2.3.1 RNA-Seq data processing

We used publicly available RNA-Seq datasets downloaded from the Gene Expression Omnibus (see Supplementary Table A.1). First, the quality and the adapter content of each dataset was evaluated using Fastqc https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. The low-quality reads and the adapter content identified from Fastqc were filtered using Trimmomatic (Bolger et al., 2014). The filtered reads were aligned to the reference human genome (hg19) using HISAT2 (Kim et al., 2015), and were further filtered to remove reads that map to the mitochondrial genome. The number of reads that aligned to each protein coding gene were counted using the htseq-count tool from HTseq software package (Anders et al., 2015). RNA-Seq data generated from PHTu and PHTd using the same culture conditions on the same day were treated as technical replicates, and combined by adding raw read counts. For JEG-3 RNA-Seq data, each data set from wild-type samples were considered a biological replicate. For other samples, we combined the technical replicates by adding raw read count. We normalized gene counts for each biological replicate by converting them into log transformed TPM values. TPM values were calculated by normalizing the read count first by the gene length and then by the sequencing depth (Li and Dewey, 2011).

2.3.2 Human Protein Atlas analysis

To evaluate the number of placenta-specific genes that are highly expressed in our samples, we used the human protein atlas database (V16.1). This database contains mRNA expression data across different tissues including placenta (Uhlén et al., 2015). The genes in the database are categorized based on tissue-specific expression, and we used their representational state transfer
application programming interface to extract this information. The database contains the expression details of 19,628 genes of which 7,835 are categorized to have some level of tissue specificity (protein atlas categories: tissue enriched, group enriched, and tissue enhanced groups). Out of those 7,835 tissue-specific genes, 354 have tissue specificity in the placenta. We took the total number of genes as the population size, number of placenta-specific genes as number of successes in the population, number of input genes as the number of draws, and the number of placenta-specific genes found in the input genes as number of observed successes to estimate the significance of each set of tissue-specific genes using hypergeometric test.

2.3.3 Differential expression analysis

DESeq2 was used for all of the differential expression analysis (Love et al., 2014). Genes were considered upregulated if they had an absolute fold-change $\geq 2$ and adjusted p-value $\leq 0.01$.

2.3.4 Functional enrichment analysis

Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010) was used for functional enrichment analysis of gene sets using the following ontologies: GO Biological Process, Mouse Phenotype, and Disease Ontology. Because we were investigating gene sets, significance was assessed using the hypergeometric statistic (McLean et al., 2010). Terms were considered significant if they had an FDR $\leq 0.05$, a fold-change $\geq 1.5$, and at least 5 genes from the input set.

2.3.5 Weighted co-expression network construction

To construct the weighted co-expression network we used the WGCNA package in R (Langfelder and Horvath, 2008). We used gene expression data for biological replicates from ESCd, PHTu, PHTd, and the placental cell lines to construct the network. The mapping information of the cell type and dataset were stored in a binary mapping file in which the rows consist of the data points and the columns represent the cell types. After that, we retained the genes with a TPM $> 1$ (19,261 genes) and selected the top quartile genes (4,815 genes) based on the variance for
network construction. The selected genes were used to calculate the adjacency matrix by using
the signed hybrid Pearson correlation method with a soft thresholding power. The power was
selected to achieve a scale free topology for the co-expression network by plotting the scale free
fit and mean connectivity at different soft thresholding powers. We used a power of 16 resulting
in a scale-free network \( R^2 = 0.9 \) (Zhang and Horvath, 2005) that retains a good number of
connections (Supplementary Fig. A.11). The interconnectedness (topological overlap), which is
used to cluster the genes by hierarchical clustering was calculated. We subsequently applied the
dynamic tree cut method to obtain the clusters and merged the closely related clusters based on
the correlation between them (correlation > 0.75) to get the final modules (Langfelder et al., 2008).
The modules were decomposed such that each module was represented by its weighted expression
(module eigengene) in the form of its first principal component. The relationship between the
modules was identified by calculating the correlation between them (Supplementary Fig. A.12).
We also calculated the module membership (KME) for each gene, which is the correlation between
the gene expression and the module eigengene. Based on the KME, we assigned genes to modules
when the correlation was greater than 0.75, allowing a gene to be a part of multiple modules or
multiple regulatory pathways (Supplementary Fig. A.13). The significance of the module in the
cell type was then estimated by calculating the correlation between the module eigengene and the
gene expression in the cell data types which was mapped in the traits file. Using this method, the
gene modules representing the different cell types were identified (Supplementary Fig. A.5).

2.3.6 Randomization Tests

Randomization tests were used to determine if modules associated with ESCd had a significant
number of genes upregulated in ESCd compared to ESCu. For each ESCd module, we generated
10,000 random gene modules that matched in size. For each random gene module, we counted
the number of genes that were upregulated in ESCd vs ESCu. The p-value was calculated as the
number of random modules that have more genes upregulated in ESCd than in the actual module
(divided by 10,000).
2.3.7 PPI network analysis

We used the STRING database (Version 10.5) to construct PPI networks from the co-expressed genes (Szklarczyk et al., 2015). A threshold weight of 0.7 was used for selecting the connection between two proteins. We extracted the largest connected sub-network from the PPI network and computed the betweenness using Cytoscape (Shannon et al., 2003). We defined the top 5% genes in each network as bottleneck genes based on their betweenness score (Score ≥ 95 Percentile).

2.4 Results

2.4.1 Genes highly expressed in ESCd are enriched for placental development genes

RNA-Seq data was previously generated for three groups of ESCd (ESCd < 40µm, ESCd 40-70µm, and ESCd > 70µm) based on the size of the differentiated cells (Yabe et al., 2016). Whereas the latter was comprised largely of sheets of syncytial cells, the ESCd < 40µm and ESCd 40-70µm fractions contained mainly mononuclear cells and a mixture of single cells and small clumps of syncytium, respectively. Similar RNA-Seq data had been generated in parallel on cytoTB cells isolated from term placentas (PHTu) and on syncytiotrophoblast (syncytioTB) generated from such cells after 48 h culture in the presence of fetal bovine serum (PHTd) (Yabe et al., 2016). These experiments revealed that both the ESCd fractions and the TB derived from term placentas were highly enriched for a large number of TB markers, yet there were major differences in their respective RNA profiles, suggesting that they possessed distinct functional attributes.

In the present study, gene expression in the above TB fractions, as well as in a number of other cell types were analyzed by procedures designed to detect functional categories of genes. The data were first normalized by using the established transcripts per million (TPM) normalization method (Li and Dewey, 2011), which is designed to eliminate bias between and within samples caused by variations in transcript length, sequencing depth, and read length (Li et al., 2015). In order to determine if the genes most highly expressed in ESCd were associated with placental functions, the 1,000 most highly expressed genes from these cells were selected and analyzed by using the Genomic
Regions Enrichment of Annotations Tool (GREAT) to determine which functional categories were enriched in each gene set (McLean et al., 2010). The analysis revealed a significant enrichment of placental development terms, including, “abnormal placenta morphology” and “abnormal angiogenesis” in all three ESCd fractions (Figure 2.1). These terms also showed significant enrichment in top 1,000 gene lists generated from PHTu and PHTd from term placental cytoTB, but not in undifferentiated embryonic stem cells (ESCu).

We then determined if typically up-regulated placental transcripts were enriched in the list of top 1,000 genes using data from the human protein atlas database (Uhlén et al., 2015), a source providing gene expression data across different tissues, including the placenta. A gene is categorized as tissue-specific if its expression level is five-fold higher in one tissue, or a small group of tissues, compared to the expression of other tissues (see Methods). For each set of tissue-specific genes, we calculated the significance of occurrence of those genes in the top 1,000 gene lists described above. We found that only placenta-specific genes are significantly enriched in each of the ESCd group gene lists (Figure 2.2 and Supplementary Fig. A.1). As expected, the PHTu and PHTd top 1,000 gene lists also showed a significant enrichment of placenta-specific genes, while the ESCu did not (Supplementary Fig. A.1). Together, these analyses provide further support that after BAP treatment, hESC differentiate into a subtype of placental cells.

2.4.2 Global expression profile in ESCd are not similar to expression profile in mesoderm cells

To determine if ESCd share similarities with mesoderm cells as has been suggested previously, the gene expression profiles of these cell types were compared. Mesoderm cell transcriptome profiles were previously generated by differentiating hESC towards cardiomyocytes, and harvesting cells on day two, obtaining cells representative of mesoderm (Szabo et al., 2015). We performed PCA on the 1,000 most highly expressed genes from the three ESCd groups and plotted the first three principal components, comprising of 93% cumulative variance. The principal components show that ESCd
groups cluster more closely with term placental cells (PHTu and PHTd) than with the mesoderm cells and ESCu (Figure 2.2).

Differential expression analysis between ESCd and mesoderm cells was then carried out using DESeq2 (Love et al., 2014). For this analysis, we increased statistical power by treating the three ESCd groups (ESCd < 40µm, ESCd 40-70µm, and ESCd > 70µm) as biological replicates, because the groups only had between 5 and 44 differentially expressed genes (DEGs) between them (Supplementary Fig. A.2). This analysis identified a total of 7,614 DEGs between the mesoderm and ESCd groups, of which 4,441 were upregulated in ESCd and 3,173 were upregulated in mesoderm cells (Figure 2.2). When functional enrichment analysis was performed, many pregnancy-associated terms were specifically enriched in genes upregulated in ESCd, including “pre-eclampsia”, “pregnancy complication”, and “female pregnancy” (Figure 2.2). Conversely, genes upregulated in mesoderm cells were specifically enriched for mesoderm development terms, such as “abnormal mesoderm development” and “somitogenesis” (Supplementary Fig. A.3). The genes from the “pregnancy complication” and “somitogenesis” terms that were upregulated in ESCd and mesoderm cells, respectively, are individually labeled in (Figure 2.2). These results strongly support the argument that ESCd exhibit functional similarities to placental cells and little functional resemblance to mesoderm.

2.4.3 Invasion genes are upregulated in ESCd compared to commonly used TB cell lines

Yabe et al. hypothesized that the ESCd, i.e. hESC differentiated under BAP conditions, most likely corresponded to the primitive, invasive TB that surround the embryo proper as the conceptus establishes itself in the wall of the uterus (Yabe et al., 2016). We therefore compared the gene expression profiles of ESCd to TB cell lines that have been classically used to model aspects of TB invasion (Hannan et al., 2010). Surprisingly, when compared to JEG-3 (Ferreira et al., 2016), BeWo (Renaud et al., 2015), and HTR-8/SVneo cells (Lee et al., 2016a), genes upregulated in ESCd were enriched for cell migration and invasion terms (Figure 2.3 and Supplementary Table A.2).
These terms include “vasculature development”, “cell adhesion”, “angiogenesis”, and “regulation of cell motility”. While most of these genes implicated in invasion were expressed in the immortalized placental cell lines, they were expressed to a lesser extent than in the ESCd. These results are also supported by PCA (cumulative variance $\sim81\%$) in which ESCd and placental cell lines did not cluster together (Supplementary Fig. A.4).

2.4.4 Identification of placental gene clusters in ESCd by co-expression network analysis

The previous analyses demonstrated that genes associated with placental development and cell invasion were highly expressed in ESCd, but they did not provide information on how the various genes were linked together in molecular networks. To study this aspect of the ESCd phenotype, we performed a weighted correlation network analysis (WGCNA) (Langfelder and Horvath, 2008) to distinguish gene expression networks that were either shared between all the placental cell models, including PHTu and PHTd, or were unusually upregulated in ESCd. A total of 26 biological replicates (after processing and combining 38 RNA-Seq datasets, see Methods 2.3.1) from ESCd, PHTu, PHTd, BeWo, JEG-3, and HTR-8/SVneo were used to carry out co-expression network analysis. The top quartile of genes (4,815 genes) were selected based on the variance in gene expression across datasets, permitting the identification of 16 gene co-expression clusters (M1 through M16). Then, the first principal component, also known as the module eigengene, of each gene module was calculated and used to identify cell type specific modules, of which six (M5, M6, M13, M14, M15, and M16) were significantly and positively enhanced in the three ESCd fractions (Supplementary Fig. A.5). Genes for each of these six modules are provided in Supplementary Data S1.

Interestingly, three of these modules (M6, M14, and M16), when analyzed by GREAT, showed significant enrichment of terms related to placental development (Figure 2.4 and Supplementary Data S2). Modules M14 and M16 were also enriched for migration and invasion associated terms, such as “cell adhesion” (M14), and “regulation of cell migration” (M16), and for a significant number of genes implicated in the preeclampsia phenotype based on the Osborne disease ontology (Osborne
et al., 2009), which is included in GREAT (False Discovery Rate (FDR) = 0.009 in M14 and FDR = 0.0005 in M16). Finally, the eigengene of module M14 had a positive and significant correlation only with the ESCd fractions, whereas M16 also had a positive correlation with HTR-8/SVneo and term placenta cells (Figure 2.4 and Supplementary Fig. A.5). This information suggests that genes in module M14 might be involved in a novel gene regulatory network important for early placental development.

Module M6 had enrichment for hormone and metabolic processes terms, which are some of the processes associated with syncytiotrophoblast (Chen and Jansson, 2017). In addition to that, the M6 module eigengene also showed a positive correlation with gene expression in BeWo cells, a cell line that has been extensively used to model events leading to the formation of syncytiotrophoblast (Wice et al., 1990) (Figure 2.4 and Supplementary Fig. A.5). Together, these analyses support the hypothesis that ESCd represent precursors of, as well as differentiated syncytiotrophoblast, and demonstrate the ability to capture different trophoblast functions in one model system (Yabe et al., 2016).

To further investigate the cell-type specificity of the genes in ESCd-associated modules, we compared their expression levels in ESCd and ESCu. We found that the expression of genes in each of modules M6, M14, and M16 is significantly higher in ESCd (P-Value < 2.2e-16) compared to ESCu (Supplementary Fig. A.6). We also found a significant enrichment (P-Value < 1e-4) for genes upregulated in ESCd (ESCd vs. ESCu) in these modules by using a randomization test (Supplementary Fig. A.6) (see Methods 2.3.6). These results provide additional evidence that genes in modules M6, M14, and M16 are important in ESCd.

2.4.5 Protein-Protein Interaction (PPI) network analysis of co-expressed modules

In order to determine how genes within each of the placental co-expression modules may interact with one another, and to identify the key regulators within each module, the STRING database was used to construct PPI networks (Szklarczyk et al., 2015) (see Methods D). The concept of “betweenness centrality” was employed to identify those genes most likely to be essential components of the network. Betweenness centrality for a gene (node) in a biological network is defined
as the number of shortest paths (between two other nodes) that pass through that node (Yu et al., 2007). Because genes with a high betweenness control flow of information in a network, they have been called “bottleneck” genes (Yu et al., 2007). Here, we ranked genes in the PPI networks by their betweenness score, and then defined bottleneck genes as representing the top 5% for each network (Supplementary Data S3). The bottleneck genes and the genes with which they are predicted to interact are shown in Figure 2.5 and Supplementary Fig. A.7. Most of the bottleneck genes identified by this analysis have been implicated in placental development, especially as it is related to TB, migration, and invasion (Table A.11). Many bottleneck genes, such as ERBB2, EDNRA, NOS2, POMC, CD44, FAS, MET, EDN1, SMPD1 have been implicated in TB invasion or in preeclampsia, but do not have a well-characterized function in early placental development. Finally, we also identified five genes (PLCB1, LUM, ADCY7, IRF7, and EHHADH) that we predict to be important for pregnancy disorders and placenta development, although such links remains to be established.

2.5 Discussion

In this study, we analyzed RNA-Seq data from BAP treated hESC to further understand functional properties of the cells. Previous studies largely focused on sets of gene markers to establish the identity of ESCd, which could lead to biased data interpretation. In the present study, we carried out several analyses to identify significant gene groupings, in an attempt to better understand how the ESCd might be used to model TB cells as they exist during the formation of the human placenta. Our results showed that the 1,000 most highly expressed genes in the ESCd are enriched with multiple placental development terms, and are also enriched for placenta-related disease ontology terms that included “pregnancy complication” and “pre-eclampsia”. Furthermore, we showed that previously annotated placenta-specific genes are significantly enriched among the highly expressed gene sets from ESCd and that no other tissue-specific genes indicative of the three main germ cell lineages are so enriched. Through use of PCA, we found that ESCd are somewhat related to other placental cell types but bear no resemblance to mesoderm cells derived from
hESC. To confirm that, while ESCd are closer to term placental cells than mesoderm cells, they are substantially different, we also carried out PCA on ESCd, ESCu, PHTu, PHTd, and syncytioTB microdissected from term placenta (Pavlíčev et al., 2017) (Figure A.8). Finally, when we compared ESCd to three commonly used placental cell lines: HTR-8/SVneo, BeWo, and JEG-3, we found that genes upregulated in ESCd were enriched for terms associated with cell invasion. These results provide strong evidence that after BAP treatment hESC differentiate into a subset of invasive placental cells, and may provide good model for studying TB functions, including invasion, as has been demonstrated in a number of other studies (Lee et al., 2015)(Sheridan et al., 2017)(Telugu et al., 2013)(Marchand et al., 2011) (Liu et al., 2017).

Co-expression network analysis has provided an insight into gene regulatory mechanisms operating in ESCd by identifying groups of genes that are likely to work together to regulate specific functional processes. Genes that made up two (M14 and M16) of the six modules we identified were enriched with different GO Biological process terms related to placental cell migration and invasion. It should also be noted that the M14 module eigengene is positively correlated only with ESCd, indicating that the genes in this module may be a part of a unique gene regulatory network that cannot be captured by studying either term placental cells, such as the primary PHTu and PHTd, or the BeWo, JEG-3, and HTR-8/SVneo cell lines. We also found that the eigengene of the M6 module has a positive correlation with BeWo cells, which have been used to model syncytial fusion (Wice et al., 1990). This module is also enriched with “peptide hormone processing” and “hormone metabolic process” terms which are some of the processes associated with syncytioTB. Together, the cell adhesion, cell migration, and hormone process gene co-expression networks identified support previous studies that speculated that ESCd may represent primitive, invasive EVT of the early stage conceptus (Yabe et al., 2016).

We further explored these co-expressed modules using the PPI data from STRING-db. This analysis showed very dense PPI networks for modules M14 and M16, which contain genes functioning in cell adhesion and cell invasion, processes that have been extensively studied in cancer cells. The PPI network for module M6 was much less dense, possibly because there are many understud-
ied genes in this module. Based on betweenness centrality, we found that many bottleneck genes in the PPI network either play an important role in the development and maintenance of early TB cells or are implicated in preeclampsia (Table 2.1). Moreover, we discovered five novel genes which have not been directly implicated in placental development or pregnancy disorders, but are known to be involved in relevant pathways including the endothelin 1 (EDN1) signaling pathway (adenylate cyclase 7; ADCY7) (Malek et al., 1993) and immune response (interferon regulatory factor 7; IRF7) (Wang and Fish, 2012). Interestingly, we found that four out of the five novel genes have high expression (Fragments per kilobase of transcript per million mapped reads (FPKM) > 5) in E7.5 or E9.5 mouse placental tissue (Tuteja et al., 2016) (Supplementary Fig. A.9), providing further support for the possibility that these novel genes identified by co-expression analysis should be further investigated for their putative role in placental development and pregnancy disorders.

The above discussion raises the question as to the precise nature of ESCd and their relationship to trophectoderm, the source material for all placental trophoblast. Our assumption is that ESCd represent the syncytial and cytoTB layers that surround the embryo proper soon after implantation begins, i.e. from about days 8-12, during the second week of pregnancy before the placental villi have emerged, a stage that is no longer available for experimentation (James et al., 2012). If this supposition is correct, we would predict that the differentiation pathway passes through a trophectoderm-like state early after the onset of differentiation (Horii et al., 2016), while the ESCd would possess a transcriptome somewhat intermediate to the trophectoderm of the blastocyst before it implants at around day 7 and the villous trophoblast that begins to emerge towards the end of the second week post-coitus. Although transcriptome data are not available for the latter, they are available from single cell RNA-Seq analyses performed on human blastocysts on days 5, 6, and 7 of preimplantation development, i.e. from the time the blastocyst first forms until the time it would normally initiate implantation (Petropoulos et al., 2016). At day 5, PCA of the sequencing data enabled cells to be assigned to either the inner cell mass (epiblast plus emerging extraembryonic endoderm) or the trophectoderm. Over 2,000 genes were differentially expressed between the two classes of cells, and, by day 7, this value increased to over 3,000. By using the differential expression
analysis of TE against other lineages, a list of 100 TE-specific genes was generated (Petropoulos et al., 2016). Of these TE-specific genes, most are expressed in ESCd (Supplementary Fig. A.10 and Supplementary Data S4). A few TE-associated genes (e.g. CYP26A1, SLC34A2, and MYLPF) had low expression in ESCd relative to trophectoderm. Genes encoding two transcription factors associated with trophoblast specification, GATA2, GATA3, were among those close to the top of the trophectoderm ranking and were also expressed robustly in ESCd. There were also two distinguishable populations of trophectoderm cells in day 6 and day 7 human blastocysts, one representing mural trophectoderm, and the other polar trophectoderm (Petropoulos et al., 2016). The latter, neighboring the inner cell mass, represents the region that attaches to the uterine epithelium and the source of the invasive cells, believed to be syncytial, that may allow the conceptus to implant in primates (Enders, 1989)(Enders et al., 1997)(Hertig et al., 1956). Again, genes with a highly significant bias in expression in polar trophectoderm (top 100 at day 7) are also expressed in ESCd (Supplementary Fig. A.10 and Supplementary Data S4). These data suggest that both the polar trophectoderm and the ESCd are active in placental hormone production (CGA and placental growth factor, PGF) and formation of syncytioTB (GCM1, OVOL1, ERVV-1, and ERVV-2). Together, these two sets of data do not prove, but are consistent with a close ontological relationship between implanting trophectoderm and ESCd.

In summary, our analyses provide evidence in support of the hypothesis that BAP treated hESC represent early invasive syncytial TB. The gene co-expression analysis highlighted networks in ESCd that may provide insight into protein-protein interactions relevant for early placental development. The genes identified from this analysis should be further studied to understand their role in placental development.

### 2.6 Supplementary Material

The supplementary figures are in Appendix A and the supplementary data can be found online (Supplementary Data).
2.7 References


Figure 2.1 ESCd are associated with placental functions and placenta specific genes. (a) Heatmap showing the enrichment (-Log_{10}(FDR)) of various placental development terms. Placental terms are significantly enriched in ESCd and TB cell types. (b) Barplot showing the enrichment (-Log_{10}(P-Value)) of tissue-specific genes in ESCd >70 cells using data from the protein atlas database. Placenta-specific genes are highly enriched in the ESCd >70 group.
Figure 2.2  Comparison of ESCd to mesoderm cells. (a) PCA plot based on the 1000 most highly expressed genes in ESCd. The ESCd samples are closer to term placental cells than to the mesoderm cells or ESCu. (b) A volcano plot showing the significantly upregulated genes (P-Adjusted ≤ 0.05) in ESCd (fold-change ≤ -2) and mesoderm (fold-change ≥ 2). The “pregnancy complication” and “somitogenesis” term genes upregulated in ESCd and mesoderm, respectively, are also marked in the plot. (c) Plot showing the enrichment of placental development terms in genes upregulated in ESCd or mesoderm. The dotted lines represent the thresholds for FDR (0.05) and fold enrichment (1.5) of the terms.
Figure 2.3 Comparison of ESCd with different placental cell lines. Volcano plots showing the differentially expressed genes between the ESCd and HTR-8/SVneo cells (a), JEG-3 cells (b), and BeWo cells (c). The gene ontology terms related to invasion are also plotted for the upregulated genes in the respective cell lines in the right panels. The dotted lines represent the thresholds for FDR (0.05) and fold enrichment (1.5) of the terms.
Figure 2.4  ESCd co-expression modules. (a) Barplot showing the top 5 enriched gene ontology terms in modules M6, M14, and M16. Hormone and metabolic terms are enriched in module M6, whereas invasion and adhesion terms are enriched in modules M14 and M16. (b) Heatmap showing the expression of genes in co-expression modules M6, M14, and M16. The purple and brown colors represent the high and low expression of the genes in the datasets respectively. The barplots below each heatmap show the eigengene expression.
Figure 2.5  PPI Network showing bottleneck gene connections. Networks showing the predicted protein-protein interactions of the bottleneck genes and their neighbors for module M14 (a) and M16 (b). The color of the node represents the betweenness score of the gene.
Table 2.1  List of bottleneck genes in the PPI networks of the ESCd co-expression modules with their function in placental development.

<table>
<thead>
<tr>
<th>Bottleneck Genes</th>
<th>Module</th>
<th>Function in Placenta development</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERBB2</td>
<td>M6</td>
<td>Implicated in TB invasion and differentiation.</td>
</tr>
<tr>
<td>MMP2</td>
<td>M6</td>
<td>Involved in the invasion and proliferation of TB cells during first trimester of pregnancy.</td>
</tr>
<tr>
<td>EGF</td>
<td>M14</td>
<td>Increases invasive capacity of first trimester cytoTB cultures</td>
</tr>
<tr>
<td>PTGS2</td>
<td>M14</td>
<td>Hypomethylated in term placenta.</td>
</tr>
<tr>
<td>KDR</td>
<td>M14</td>
<td>Receptor for Vascular Endothelial Growth Factor. Reduced expression in preeclamptic and preterm birth placentas.</td>
</tr>
<tr>
<td>PLCB1</td>
<td>M14</td>
<td>No known role in placenta.</td>
</tr>
<tr>
<td>EDNRA</td>
<td>M14</td>
<td>Endothelin and its receptors are implicated in preeclampsia.</td>
</tr>
<tr>
<td>WNT5A</td>
<td>M14</td>
<td>Regulates the growth and development of early TB cells.</td>
</tr>
<tr>
<td>HSPG2</td>
<td>M14</td>
<td>Involved in TB cell invasion.</td>
</tr>
<tr>
<td>ITGAV</td>
<td>M14</td>
<td>Mutant mice have abnormal labyrinth layer development.</td>
</tr>
<tr>
<td>COL3A1</td>
<td>M14</td>
<td>High expression level during placental development in mouse.</td>
</tr>
<tr>
<td>PDGFB</td>
<td>M14</td>
<td>Regulates development of the labyrinthine layer in mouse placenta.</td>
</tr>
<tr>
<td>THBS1</td>
<td>M14</td>
<td>Hyperexpression in placenta is associated with disorders in placental villi maturation and branching in gestosis.</td>
</tr>
<tr>
<td>NOS2</td>
<td>M14</td>
<td>Promotes TB invasion.</td>
</tr>
<tr>
<td>STAT6</td>
<td>M14</td>
<td>Part of the signaling pathway that is involved in the proliferation of TB cells during pregnancy.</td>
</tr>
<tr>
<td>LUM</td>
<td>M14</td>
<td>Involved in cell migration, proliferation, and differentiation. No known role in placenta.</td>
</tr>
<tr>
<td>KIT</td>
<td>M14</td>
<td>Expressed in placental tissue during human pregnancy.</td>
</tr>
<tr>
<td>POMC</td>
<td>M14</td>
<td>Differentially methylated in preeclamptic placenta.</td>
</tr>
<tr>
<td>LRP2</td>
<td>M14</td>
<td>Implicated in regulation of maternal-fetal transport during pregnancy.</td>
</tr>
<tr>
<td>JUN</td>
<td>M16</td>
<td>It is a major component of activator protein 1 which helps in the invasion of the TB cells.</td>
</tr>
<tr>
<td>CD44</td>
<td>M16</td>
<td>Implicated in TB invasion and angiogenesis in the placenta.</td>
</tr>
<tr>
<td>FAS</td>
<td>M16</td>
<td>A polymorphism in this gene is associated with preeclampsia.</td>
</tr>
<tr>
<td>MET</td>
<td>M16</td>
<td>Implicated in TB differentiation.</td>
</tr>
<tr>
<td>EDN1</td>
<td>M16</td>
<td>Implicated in pathogenesis of hypertension in preeclampsia.</td>
</tr>
<tr>
<td>DCN</td>
<td>M16</td>
<td>Involved in TB cell migration.</td>
</tr>
<tr>
<td>ADCY7</td>
<td>M16</td>
<td>Involved in Endothelin 1 signaling pathway. No known role in placenta.</td>
</tr>
<tr>
<td>IRF7</td>
<td>M16</td>
<td>Involved in innate immune response to viral infections. No known role in placenta.</td>
</tr>
<tr>
<td>EHHADH</td>
<td>M16</td>
<td>No known role in placenta.</td>
</tr>
<tr>
<td>SMPD1</td>
<td>M16</td>
<td>Reduced activity in preeclamptic placentas.</td>
</tr>
</tbody>
</table>
CHAPTER 3. TISSUEENRICH: TISSUE-SPECIFIC GENE ENRICHMENT ANALYSIS

Modified from a manuscript published in *Bioinformatics*

Ashish Jain$^{1,2}$ and Geetu Tuteja$^{1,2}$

$^1$Bioinformatics and Computational Biology, Iowa State University, Ames, 50010, USA
$^2$Genetics, Development, and Cell Biology, Iowa State University, Ames, 50010, USA

3.1 Abstract

**Summary:** RNA-Seq data analysis results in lists of genes that may have a similar function, based on differential gene expression analysis or co-expression network analysis. While tools have been developed to identify biological processes that are enriched in the genes sets, there remains a need for tools that identify enrichment of tissue-specific genes. Therefore, we developed TissueEnrich, a tool that calculates tissue-specific gene enrichment in an input gene set. We demonstrated that TissueEnrich can assign tissue identities to single cell clusters and differentiated embryonic stem cells. **Availability:** The TissueEnrich web application is freely available at [http://tissueenrich.gdcb.iastate.edu/](http://tissueenrich.gdcb.iastate.edu/). The R package is available through Bioconductor at [https://bioconductor.org/packages/TissueEnrich](https://bioconductor.org/packages/TissueEnrich). Both the web application and R package are for non-profit academic use under the MIT license.

3.2 Introduction

The development of RNA-Seq technology has enabled large-scale comparison of gene expression in a multitude of developmental stages, cell-types, and conditions. RNA-Seq data analysis identifies genes that likely have a shared function, either through differential gene expression analysis or co-expression network analysis. Gene Ontology (GO) enrichment analysis is widely used to assign
function to the gene sets and to gain insights into the biological processes they are involved in. While GO analysis identifies enriched processes in sets of genes, it does not determine enrichment of tissue-specific genes. Understanding which groups of genes are tissue-specific is valuable, as tissue-specific genes are more likely to be associated with human disease (Winter et al., 2004). To this end, tools have been developed that include tissue enrichment, or tissue-specific enrichment calculations (http://genetics.wustl.edu/jdlab/tsea/)(Xu et al., 2014)(Watanabe et al., 2017)(Komljenovic et al., 2016). However, these tools are limited in the flexibility of parameters for defining tissue specificity, or the ability to allow users to define their own tissue-specific gene sets to use for enrichment analysis. Therefore, we developed “TissueEnrich”, a tool to carry out tissue-specific gene enrichment. TissueEnrich is available as an interactive web application, allowing the user to visualize tissue-specific gene enrichment and visualize expression of genes from the input set that were determined to be tissue-specific. We also developed an R package that further allows users to define tissue-specific genes in custom datasets, which they can then use for tissue-specific analysis.

### 3.3 Methods

**3.3.1 RNA-Seq Datasets**

To calculate tissue-specificity of genes, we used RNA-Seq data from GTEx (Ardlie et al., 2015), HPA (Uhlén et al., 2015) (https://v18.proteinatlas.org/about/download), and the mouse ENCODE project (Shen et al., 2012). We used data from tissues with > 1 biological replicate. Processed GTEx data were downloaded from the Expression Atlas (E-MTAB-5214) (Petryszak et al., 2016), and the samples from sub-tissues were grouped and considered to be biological replicates. The one-to-one orthologous genes between human and mouse were downloaded from Ensemble (Version 91) (Zerbino et al., 2018). TissueEnrich only uses protein-coding genes for analysis.

**3.3.2 Tissue Naming conventions**

We used tissue names provided by each of the consortia for their datasets. These sample names are also used in other databases, such as the EBI Expression Atlas, allowing users to easily
compare TissueEnrich results to the original databases, as well as other databases using these data. The tissues isolated from embryonic stages are prefixed with 'E' followed by the timepoint. For example, in Mouse ENCODE data, the placenta tissue isolated from embryonic day 14.5 is named as E14.5-Placenta. All the other tissues are isolated from adults.

### 3.3.3 Defining tissue-specific genes

Multiple metrics exist for defining tissue-specificity (Xu et al., 2014) (Watanabe et al., 2017) (Uhlén et al., 2015) (Kryuchkova-Mostacci and Robinson-Rechavi, 2016) (Dougherty et al., 2010). Count-based methods have been described as simple, and perform well if thresholds are chosen properly (Kryuchkova-Mostacci and Robinson-Rechavi, 2016). For these reasons, TissueEnrich defines tissue-specific genes based on the algorithm used in the HPA (Uhlén et al., 2015). This algorithm uses a count-based method, and our R package provides users the flexibility of adjusting thresholds. The grouping method described by the HPA also allows users to identify groups of tissues within which a particular gene is enriched. Briefly, the genes were divided into six groups. These groups are:

- **Not expressed**: Genes with an expression level less than 1 (TPM or FPKM) across all the tissues.
- **Tissue enriched**: Genes with an expression level greater than or equal to 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a particular tissue compared to all other tissues.
- **Group enriched**: Genes with an expression level greater than or equal to 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a group of 2-7 tissues compared to all other tissues, and that are not considered Tissue Enriched.
- **Tissue enhanced**: Genes with an expression level greater than or equal to 1 (TPM or FPKM) that also have at least five-fold higher expression levels in a particular tissue compared to
the average levels in all other tissues, and that are not considered Tissue Enriched or Group Enriched.

- Expressed in all: Genes with an expression level greater than or equal to 1 (TPM or FPKM) across all of the tissues that are not in any of the above 4 groups.

- Mixed: Genes that are not assigned to any of the above 5 groups.

Genes from the “Tissue Enriched”, “Group Enriched”, and “Tissue Enhanced” groups are classified as tissue-specific genes.

### 3.3.4 Tissue-specific gene enrichment

We used the hypergeometric test to calculate the enrichment of tissue-specific genes in the input gene set.

\[
P(X > k) = \sum_{i=k+1}^{n} \binom{K}{i} \binom{N-K}{n-i} \binom{N}{n}
\]

and the fold-change is calculated as:

\[
Fold - Change = \frac{\binom{k}{n}}{\binom{K}{N}}
\]

N is the total number of genes, K is the total number of tissue-specific genes for a tissue, n is the number of genes in the input gene set, and k is the number of tissue-specific genes in the input gene set. Multiple hypothesis correction is done using the Benjamini and Hochberg correction.

### 3.3.5 Background Genes

Instead of using all the genes as background users can use a custom background gene set for carrying out tissue-specific gene enrichment. In this case, instead of using all the genes in the dataset, a background gene set is being used to carry out the enrichment analysis. It should be noted that the background gene set must have all the genes of the input gene set. The p-value is calculated as:
\[ P(X > k) = \sum_{i=k+1}^{n} \frac{\binom{K_b}{i} \binom{N_b-K_b}{n-i}}{\binom{N_b}{n}} \]

and the fold-change is calculated as:

\[ Fold - Change = \frac{\binom{k}{n}}{\binom{K_b}{N_b}} \]

\(N_b\) is the total number of background genes, \(K_b\) is the total number of tissue-specific genes for a tissue in background genes, \(n\) is the number of genes in the input gene set, and \(k\) is the number of tissue-specific genes in the input gene set.

### 3.4 Results

#### 3.4.1 Web Application

We used R Shiny (McPherson et al., 2017) to develop a user-friendly web application to calculate tissue-specific gene enrichment in a user-provided gene set. From the “Tissue Enrichment” tab, the user can select the organism their data is from and the RNA-Seq dataset to use for tissue-specificity information. The output is an interactive bar chart, depicting the significance of tissue-specific gene enrichment, plotted as –\(\log_{10}(P\text{-value})\) on the y-axis, across each tissue. The user can view the expression values of the genes that were part of a tissue-specific group in an interactive heatmap by clicking the corresponding bar for that tissue in the bar chart. Furthermore, users can search for the expression values of individual genes along with their tissue-specific groups under the “Tissue Specific Genes” tab. The “Help” tab provides detailed usage instructions.

#### 3.4.2 R Package

The TissueEnrich R package contains similar functions to those in the web application. The package includes the “teGeneRetrieval” function, which can be used to define tissue-specific genes in any given dataset. In this function, users can adjust thresholds for calculating tissue-specificity. The resulting tissue-specific genes can be used to carry out tissue-specific gene enrichment, using the “teEnrichmentCustom” function.
3.4.3 Case study 1: Defining tissue-specificity of genes expressed in differentiated embryonic stem cells (ESCs)

We used two RNA-Seq datasets generated from differentiated ESCs to test TissueEnrich. The first dataset is from ESCs differentiated into cardiomyocytes (Szabo et al., 2015). We ran TissueEnrich on the 1,000 most highly expressed genes (processed as previously described in (Jain et al., 2017)) and found enrichment for heart-specific genes using both the HPA and mouse ENCODE datasets (Figure 3.1). These results validate the robustness of TissueEnrich and highlight heart-specific genes that may be interesting for follow up experiments. The second RNA-Seq dataset is from ESCs differentiated into trophoblast-like cells (Yabe et al., 2016). While Yabe et al. (2016) concluded that the differentiated cells are of trophoblast origin, the origin has been debated (Roberts et al., 2014). We previously used an approach similar to what we have now packaged into TissueEnrich to determine that the 1,000 most highly expressed genes in these cells have strong enrichment for placenta-specific genes defined through the HPA (Jain et al., 2017). Here, we used TissueEnrich for the same genes and found that they also show enrichment for placenta-specific genes defined through the mouse ENCODE project (Figure 3.2), indicating trophoblast-like cells may have a conserved function in mouse.

3.4.4 Case study 2: Annotation of cell clusters from single-cell RNA-Seq data

Next, we used single-cell RNA-Seq data from mouse embryos at e6.5 (Scialdone et al., 2016). They clustered the cells using gene expression data and assigned tissue identities to the cell clusters based on marker genes that were upregulated in each cluster compared to all other clusters. Here, we ran TissueEnrich for each set of genes that were specifically up-regulated in cell clusters associated with e6.5 to determine if more information could be obtained about the cell identities. We found that cluster 1 showed strongest tissue-specific enrichment for intestine, liver, and kidney (tissues derived from visceral endoderm); cluster 2 for placenta (tissue derived from extraembryonic ectoderm); and cluster 3 for brain and olfactory bulb tissue (Figure 3.3). These results are in agreement with the tissue identities assigned in Scialdone et al., and provide additional detail
on tissue-specific gene expression that may further help understand cell specification during early development.

3.5 Conclusions

We developed the TissueEnrich web application and R package, and demonstrated that it provides an unbiased way to carry out tissue-specific gene enrichment in mouse and human RNA-Seq data.

3.6 References


Figure 3.1  Bar charts showing the enrichment of (a) Heart Muscle (using the HPA dataset) and (c) Heart (using the Mouse ENCODE dataset) in the 1,000 most highly expressed genes in cardiomyocytes differentiated from ESCs. Heatmaps showing the expression of (b) Heart Muscle and (d) E14.5-Heart specific genes enriched in the 1,000 most highly expressed genes.
Figure 3.2  Bar chart showing the enrichment (-Log_{10}(P-Value)) of placenta-specific genes (using the Mouse ENCODE dataset) in the 1,000 most highly expressed genes in trophoblast-like cells differentiated from ESCs.
Figure 3.3 Bar charts showing the enrichment (-Log$_{10}$(P-Value)) of visceral organ specific genes in the marker genes of cluster 1 (a), placenta-specific genes in the marker genes of cluster 2 (b), and brain and olfactory bulb specific genes in the marker genes of cluster 3 (c) (using the Mouse ENCODE dataset). Marker genes belonging to each cluster were obtained from Scialdone et al.
CHAPTER 4. FACTOR ANALYSIS-CLUSTERING MODEL FOR INTEGRATING AND CLUSTERING MULTI-OMICS DATA

Manuscript in preparation to be published in *Bioinformatics*

Ashish Jain\textsuperscript{1,2}, Geetu Tuteja\textsuperscript{1,2}, and Karin S. Dorman\textsuperscript{1,2,3}

\textsuperscript{1}Bioinformatics and Computational Biology, Iowa State University, Ames, 50010, USA
\textsuperscript{2}Genetics, Development, and Cell Biology, Iowa State University, Ames, 50010, USA
\textsuperscript{3}Department of Statistics, Iowa State University, Ames, 50010, USA

4.1 Abstract

Modern sequencing technologies have made it possible to comprehensively study regulatory mechanisms by simultaneous measurement of multiple ‘omes on the same sampled individuals, producing multi-omics datasets. One important goal is to use the very rich and high-dimensional data to cluster the sampled individuals into groups, which might represent disease subtypes or other important subpopulations with a role in the studied biological process. Existing clustering methods for multi-omics data work by first reducing the dimension of the data and then clustering the reduced-dimensional dataset, but there is a loss of information in this two-step approach. The clusters are not considered during data reduction, though they obviously contribute to the variation, and information learned during clustering is not used to update the data reduction, which often is used to identify important biological features. Here, we present a mixture of factor scores model, MOFA-Cluster, that carries out simultaneous integrative clustering. It integrates the multi-omics data by identifying a reduced number of clustered latent factors using a mixture model. In simulation, MOFA-Cluster is better than MOFA, moCluster, and iClusterBayes for estimating clusters, significant features, and latent variables. We also applied MOFA-Cluster to a colorectal cancer data cohort from The Cancer Genome Atlas (TCGA). We identified three subtypes
of colorectal cancer and various marker genes that may have an important role in the progression of the identified cancer subtypes. Overall, we showed that MOFA-Cluster is a good method to carry out the integrative clustering of the multi-omics data to identify subtypes of various diseases and cancers.

4.2 Introduction

Advances in high throughput sequencing have led to a surge in the amount of sequencing data to study molecular and developmental mechanisms. Various sequencing-based methods can interrogate different aspects of the tightly regulated mechanisms of development and disease (Yugi et al., 2016). For example, RNA-sequencing (RNA-Seq) provides information about the level of gene expression (Ozsolak and Milos, 2011), DNA sequencing explains the regulation of gene expression (Metzker, 2010)(Laird, 2010), and proteomics provides information about protein expression levels (Altelaar et al., 2013). Individually, each dataset provides important information about biological processes, but only the integrated analysis of all datasets reveals the coordinated mechanisms underlying the biology.

The task of taking data from $N$ individuals and identifying previously unknown groups or subtypes is solved using unsupervised clustering (Witten et al., 2016). There are many methods with varying approaches for clustering multivariate data (Xu and Tian, 2015)(Rodriguez et al., 2019). The most popular and commonly used clustering methods are $k$-means and hierarchical clustering (Lloyd, 1957)(MacQueen, 1967)(Johnson, 1967). Both of these methods are dependent on distance or similarity measures to cluster the objects. However, with the increase in the dimensionality of the datasets, the distance between the datasets becomes uniform making these measures useless for clustering. This phenomenon is called as “Curse of dimensionality” (Steinbach et al., 2004). Due to this, the direct clustering using these methods on the high-dimensional sequencing data is not effective. The problem can be solved by reducing the data to a lower dimension before the clustering using Principal Component Analysis (PCA) or other feature extraction methods. However, the dimension reduction step may lead to a loss of information. Due to this drawback, several
methods have been proposed to carry out high dimensional data clustering effectively (Bouveyron and Brunet-Saumard, 2014). One of the earliest effective models to carry out the clustering of high-dimensional data is a mixture of factor analyzers (MFA) based on the subspace clustering method. It carries out the concurrent estimation of factors and a mixture of feature loadings and performs better than the simple dimensionality reduction methods (Ghahramani, Zoubin and Hinton, 1996). MFA assumes the feature loadings vary with cluster, but the factors are shared across all clusters. Other methods were also built to optimize the MFA model and lower its complexity. The above described methods have so far only been used for a single dataset and not been updated for multi-omics datasets.

Multi-omics data are challenging to cluster because they consist of multiple datasets collected on the same sample of \( N \) individuals, where each dataset is very high dimensional. A popular solution to integrate multiple types of data is to use exploratory factor analysis. In this method, the multiple, high dimensional datasets are explained by a few shared latent variables or factors. Many tools have been developed to integrate different omics data using latent variables. iCluster and iClusterPlus carry out the transformation of the high dimensional data to low dimensional data with data integration (Mo et al., 2013)(Shen et al., 2009). A faster and more accurate solution, iClusterBayes performs Bayesian inference for a similar model (Mo et al., 2018). MoCluster uses consensus PCA (CPCA) (Westerhuis et al., 1998) to separately calculate the latent variables for each of the omics data. After that, the tool combines the latent variables from the individual omics data by maximizing the correlation between them, resulting in the final joint latent variables (Meng et al., 2016). Multi-Omics Factor Analysis (MOFA) carries out the data integration using factor analysis (Argelaguet et al., 2018). It uses a probabilistic Bayesian framework to integrate the omics data. The unobserved latent variables are estimated from the data using a Variational Bayes algorithm. They recently updated the tool to work with single-cell omics datasets (Argelaguet et al., 2020).

All of these tools first reduce the data dimension and then cluster the resulting low-dimensional latent variables. No current tool can simultaneously estimate the low-dimensional representation
and cluster the multi-omics data. The presence of cluster structure in the factors violates the assumptions of Factor Analysis and other dimension reduction models. Ignoring clusters during data reduction must at least bias the parameter estimates, such as factor loadings, and may at worst remove the signal of the clusters during data reduction. Simultaneous estimation of reduced dimension and clusters should provide better estimation of both clusters and factor loadings. To identify clusters, including cancer subtypes, in high-dimensional multi-omics data, we developed a mixture of factor scores model and an accompanying tool that can estimate the model using a variational Bayes algorithm. Our tool, MOFA-Cluster, can simultaneously integrate multi-omics datasets and identify clusters with enhanced accuracy over existing methods.

4.3 Methods

4.3.1 Factor analysis-clustering model

In this model, the input data consist of $M$ views representing different high throughput assays, $Y^m = (y^m_{nd}) \in \mathbb{R}^{N \times D_m}$, where the $m$th view is an $N \times D_m$ matrix of data from the same $N$ sampled individuals observed at $D_m$ features. Each view represents different assays with non-overlapping features. The input data is then factored as

$$Y^m = 1\mu^T + ZW^mT + \epsilon^m,$$

where $1$ is a vector of $1$’s, $\mu$ are the feature means, $Z = (z_{nk}) \in \mathbb{R}^{NxK}$ is a matrix that contains the low-dimensional latent variables or factors that are shared across views, $W^m = (w^m_{dk}) \in \mathbb{R}^{D_m \times K}$ are loading matrices that provide the link between the high-dimensional feature space to the low dimensional factor space, and $\epsilon^m = (\epsilon^m_{nd}) \in \mathbb{R}^{N \times D_m}$ denotes residual noise. The data supplied to the model is scaled, such that $\mu = 0$. The residual noise is assumed to follow the Gaussian distribution,

$$p(\epsilon^m_{nd}) = \mathcal{N}(\epsilon^m_{nd} | 0, \Psi^m)$$

for $\Psi^m = \text{diag}(1/\tau^m_1, 1/\tau^m_2, \ldots, 1/\tau^m_{D_m})$ in standard factor analysis models, where $\tau^m_d$ is the error
precision for the $d$th feature in the $m$th dataset. For standard factor analysis models, the $n$th row of the latent factor matrix $z_{n*}^T$ is assumed to follow a multivariate normal with

$$p(z_{n*}^T) = \mathcal{N}(z_{n*}^T \mid 0, \Phi),$$

so the observed data follow normal density

$$p(y_{n*}^{mT}) = \mathcal{N}(y_{n*}^{mT} \mid 0, W^m\Phi W^m + \Psi^m).$$

For this model to be identifiable, constraints must be applied to the $\Phi$, $\Psi^m$, and $W^m$ matrices. Standard factor analysis requires the $z_{n*}$ be orthonormal, i.e., $\Phi = I_K$. Then, the solutions is unique up to an orthogonal transformation, for which it is most common to use $W^m (\Psi^m)^{-1} W^m = \Gamma = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_K)$, where $\lambda_i > \lambda_j$ for all $i > j \in \{1, 2, \ldots, K\}$ (Anderson, 1984). In the Bayesian approach used here, however, the rotation is chosen during estimation to balance the information in data against sparsity-inducing priors.

To account for possible subgroups among the samples, we assume that the $N$ samples are divided into $L$ clusters, distinguished by their factors. This model is distinct from the mixture of factor analyzers (McLachlan and Peel, 2000b), where there are $L$ sets of loadings for shared factors. In the mixture model, we assume

$$p(z_{n*}^T) = \sum_{l=1}^{L} \pi_l \mathcal{N}(z_{n*}^T \mid \nu_l, \Phi_l),$$

with constraints $\sum_{l=1}^{L} \pi_l \nu_l = 0$ and $\Phi_l = I_K$ for all $l$. This model on the factors is a mean-constrained version of the Gaussian mixture model with homogeneous spherical clusters (McLachlan and Peel, 2000a), which in turn is the soft-clustering mixture model most closely related to the hard-clustering $k$-means algorithm (Celeux and Govaert, 1992). Now, the density of the observed data is a mixture,
\[ p(y_{mT}^{n*}) = \sum_{l=1}^{L} \pi_l \int \cdots \int N(y_{mT}^{n*} | W^m z_{n*}^T, \Psi^m) \times N(z_{n*}^T | \nu_l, I_K) \, dz_{n*} \]
\[ = \sum_{l=1}^{L} \pi_l N(W^m \nu_l, W^m W^m + \Psi^m) , \]
which is identifiable if \( W^m \nu_l \) are distinguishable for all \( l = 1, 2, \ldots, L \).

For a Bayesian treatment, we supplement the data with a matrix of latent variables \( V = (v_{nl}) \), specifying the source mixture components of each data point, using a “one-of-\( L \)” vector representation within each row, where component \( v_{nl} = 1 \) if observation \( n \) is in cluster \( l \). In addition, we place prior distributions on all the variables. We use the standard priors for the precisions and mixture proportions,

\[ p(\tau^m_d) = \text{Gamma} (\tau^m_d | a_0^\tau, b_0^\tau) \]
\[ p(\pi) = \text{Dirichlet} (\pi | \alpha_{01}^\pi, \alpha_{02}^\pi, \ldots, \alpha_{0L}^\pi) , \]
with \( a_0^\tau, b_0^\tau = 1 \times 10^{-14} \) as in Argelaguet et al. (2018) and \( \alpha_{0l}^\pi = 1, l = 1, 2, \ldots, L \) to obtain diffuse or uninformative priors.

The constraint \( \sum_{l=1}^{L} \pi_l \nu_l = 0 \) implies \( \nu_{k1}, \nu_{k2}, \ldots, \nu_{kL} \) are not independent for each \( k \). Define \( L \)-dimensional vector \( \nu_{k*}^T \). We impose a degenerate Gaussian prior (Anderson, 1984),

\[ p(\nu_{k*}^T) = N (\nu_{k*}^T | 0, DD^T) , \]
where

\[ D = \begin{bmatrix} 0 & \vdots & 0 \\ \pi_{L-1} I_{L-1} & \vdots & \pi_{L-1} \end{bmatrix} \]

The density \( p(\nu_{k*}'^T) \) does not exist in \( L \)-space because of the degeneracy, but it does exist in \( (L - 1) \)-space, where it is simply
\[ p(\nu_{k+}^T) = \mathcal{N}(\nu_{k+}^T \mid \mathbf{0}, \mathbf{I}_{L-1}) , \]

with \( \nu_{k+} = (\nu_{k1}, \nu_{k2}, \ldots, \nu_{k,L-1}) \), i.e., the density of \( L - 1 \) iid standard normal random variates.

For the weights \( W^m \), we used the same regularization method used in MOFA (Argelaguet et al., 2018) to enforce sparsity in the weight matrix \( (w_{d,k}^m) \) at both view-/factor and feature-wise levels. At the view-/factor level, the weight matrix \( (w_{d,k}^m) \) is shrinked to zero if a factor is not capturing any variation in a view. At the feature level, the weights of the individual features on active factors are made zero, under the biological assumption that only a small number of features participate, with non-zero weight, in any given factor. To achieve this regularization, Argelaguet et al. combines an Automatic Relevance Determination (ARD) prior for view-/factor-wise sparsity with a spike-and-slab prior for feature-wise sparsity. For the spike-and-slab prior, the weights \( w_{d,k}^m \) are reparameterizes as \( w_{d,k}^m = \hat{w}_{d,k}^m s_{d,k}^m \). Thus,

\[ p(\hat{w}_{d,k}^m, s_{d,k}^m) = \mathcal{N}(\hat{w}_{d,k}^m \mid 0, 1/\alpha_k^m) \text{ Bernoulli}(s_{d,k}^m \mid \theta_k^m) , \]

with uninformative priors,

\[ p(\theta_k^m) = \text{Beta} \left( \theta_k^m \mid a_0^\theta, b_0^\theta \right) \]

\[ p(\alpha_k^m) = \text{Gamma} \left( \alpha_k^m \mid a_0^\alpha, b_0^\alpha \right) , \]

and hyper-parameters \( a_0^\theta, b_0^\theta = 1 \) and \( a_0^\alpha, b_0^\alpha = 1 \times 10^{-14} \).
The resulting joint probability density function of the data and parameters, supplemented with the Bernoulli variables \( S \), factors \( Z \), and cluster indicators \( V \), is given by

\[
p(Y, Z, S, V, \hat{W}, \pi, \nu, \theta, \alpha, \tau) = \prod_{m=1}^{M} \prod_{n=1}^{N} N(y_{ns}^{mT} | (S_{m}^{m} \odot \hat{W}_{ns}^{m})z_{ns}^{m}, \tilde{y}^{m}) \\
\times \prod_{n=1}^{N} \prod_{l=1}^{L} [\pi_{l}N(z_{ns}^{m} | \nu_{l}, I_{k})]^{n_{l}} \\
\times \prod_{m=1}^{M} \prod_{d=1}^{D_{m}} \prod_{k=1}^{K} \mathcal{N}(w_{dk}^{m} | 0, 1/\alpha_{k}^{m}) \text{Ber}(s_{d,k}^{m} | \theta_{k}^{m}) \\
\times \prod_{m=1}^{M} \prod_{k=1}^{K} \text{Beta}(\theta_{k}^{m} | a_{0}^{\theta}, b_{0}^{\theta}) \\
\times \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}(\alpha_{k}^{m} | a_{0}^{\alpha}, b_{0}^{\alpha}) \\
\times \prod_{m=1}^{M} \prod_{d=1}^{D_{m}} \mathcal{G}(\tau_{d}^{m} | a_{0}^{\tau}, b_{0}^{\tau}) \\
\times \prod_{k=1}^{K} \mathcal{N}(\nu_{k}' | 0, DD') \\
\times \text{Dir}(\pi | L, a_{0}^{\pi}),
\]

where \( \odot \) represents elementwise multiplication.

### 4.3.2 Model inference

Letting \( x_{n} = (z_{n}, v_{n})^{T} \) be the hidden variables accompanying the \( n \)th observation \( y_{ns} \) and \( (S, \hat{W}, \pi, \nu, \alpha, \theta, \tau) \) the unknown parameters, our goal is to estimate the log posterior distribution \( \ln p(x, | y) \). We extend the Variational Bayes (VB) estimation method proposed in Argelaguet et al. (2018). VB seeks a substitute \( q(x, y) \) for \( p(x, y) \) that maximizes the right side of Jensen’s inequality

\[
\ln \int ddq(x, y) \frac{p(x, y)}{q(x, y)} \geq \int ddq(x, y) \ln \frac{p(x, y)}{q(x, y)},
\]

also known as the Evidence Lower Bound (ELBO). The equality is met and the maximum achieved when \( q(x, y) \) is the true posterior \( p(x, y) \). When the true posterior distribution is not known, VB
assumes \( q(x, \cdot) \) follows some known functional form and chooses its parameters to maximize the ELBO. The resulting estimate \( \hat{q}(x, \cdot) \) will be as close to the true posterior as such functional form can get and will also hopefully be a good approximation to the true posterior distribution. The mean-field approximation to the true posterior distribution assumes \( q(x, \cdot) \) factors, we assume as
\[
q(x, \cdot) = \prod_{n=1}^{N} q_{v_n}(v_n) \prod_{k=1}^{K} q_{z_{nk}}(z_{nk})
\]
and
\[
q_{\theta}(\theta) = q_{\pi}(\pi) \left[ \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{d=1}^{D_m} q_{s_{dk}}(s_{dk}) q_{\nu_k^m}(\nu_k^m) \right] \left[ \prod_{m=1}^{M} \prod_{k=1}^{K} q_{\nu_k^m}(\theta_k^m) \right] \left[ \prod_{m=1}^{M} \prod_{k=1}^{K} q_{\alpha_m^k}(\alpha_k^m) \right] \left[ \prod_{m=1}^{M} \prod_{d=1}^{D_m} q_{\tau_m^d}(\tau_m^d) \right].
\]
In other words, we assume these components of \( \cdot \), vectors \( v_n \), and factor scores \( z_{nk} \) are approximately a posteriori independent. The advantage is that under the posterior independence assumption, the \( q(\cdot) \) adopt particularly simple forms, such as Normal, Multinoulli, etc.

The new terms in our model compared to that of Argelaguet et al. (2018) are
\[
q_{v_n}(v_n) = \text{Multinoulli}(v_n | \mu_{v_{n1}}, \mu_{v_{n2}}, \ldots, \mu_{v_{nL}}), \ n = 1, 2, \ldots, N
\]
\[
q_{\nu_l}(\nu_l) = \mathcal{N}(\nu_l | \mu_{\nu_l}, \sigma_{\nu_l}^2 I_K), \ l = 1, 2, \ldots, L
\]
\[
q_{\pi}(\pi) = \text{Dirichlet}(\pi | \alpha_1^\pi, \alpha_2^\pi, \ldots, \alpha_L^\pi).
\]
In addition, \( q_{z_{nk}}(z_{nk}) \) remains normal, but has a new mean and variance. The parameters of these distributions and the calculations necessary to compute the ELBO are provided in Supplementary Material §B.3.

Argelaguet et al. starts the training with a large number of factors and prunes them when the variance explained drops below a threshold during iterations of VB. We do not prune factors in our method, relying on good initialization. We do drop clusters if there is no element associated with them.
4.3.3 Initialization of model parameters

Initialization of parameters in a cluster model is very difficult, and poor initialization can seriously affect inference (Maitra, 2009) (Baudry and Celeux, 2015) (Fränti and Sieranoja, 2019). The VB approach requires initial values for the parameters of the $q(. \cdot)$ variational distributions. The random or fixed initialization previously used in the MOFA (Argelaguet et al., 2018) model fails, leading to the loss of all factors and clusters during VB iteration. We find it is critical to obtain initial estimates of the factor scores, weights, and initial clusters. In what follows, it is important to consider the update order for parameters is $S^m \hat{W}^m, Z, \alpha, \theta, \pi, \nu,$ and $\tau$.

For initializing the posterior mean $E_q[Z_{nk}]$ of the entries in the latent variable matrix, we concatenated the multivariate data into a combined $N \times \sum_m D_m$ matrix and used the fit $\hat{Z}_0$ from the fad package for maximum likelihood estimation of factor analysis models (Dai et al., 2020). Although fad estimates the latent factors without considering the clusters, these are good estimates for initialization. We can also initialize the latent factors using the estimates from other integration tools, including iClusterBayes and MOFA. We choose the optimum number of latent variables by selecting the $K$ with minimum BIC when running fad with 1 to $N/5$ factor variables. Although this method is defined for unclustered data, like the other factor analyzers (MOFA, MoCluster, iClusterBayes), it is often able to find the latent structure anyway, and it is fast. Next, we apply the NbClust method to $\hat{Z}_0$ to choose the optimum number of clusters $\hat{L}_0$ between 2 and 20 (user-defined). NbClust calculates 30 cluster indices and selects the number of clusters with the maximum number of votes (Charrad et al., 2014). During $L$ selection, we perform $k$-means, with 500 random initializations with a maximum of 1000 iterations to find the best clustering at each choice of $L$.

We take the best $k$-means partition of the $\hat{Z}_0$ matrix with $\hat{L}_0$ clusters to initialize the expected cluster mean matrix $E_q[\nu]$ to the cluster sample means and the expected mixture proportions $E_q[\pi]$ from the estimated labels. For initializing the expected weight matrix (without sparsity) and the precision parameters, we regress the centered input data $Y$ on the initial factor matrix $\hat{Z}_0$ using ordinary least square regression. The coefficients and mean square error values from the regression are used to initialize the expected weight $E_q[\hat{w}_{dk}^m | s_{dk}^m = 1]$ and noise matrices $E_q[\tau_d^m]$, which are
used to update $Z$, $S$ and $\hat{W}$ during the first round of VB. All other parameter initialization is as in (Argelaguet et al., 2018).

Table 4.1 Parameters used for simulation study

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Number of Factors</th>
<th>Cluster Proportion</th>
<th>Dataset</th>
<th>Feature Proportion</th>
<th>Noise Precision (Mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0.25,0.25,0.25,0.25</td>
<td>Data 1</td>
<td>10%</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data 2</td>
<td>15%</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.25,0.25,0.25,0.25</td>
<td>Data 1</td>
<td>5%</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data 2</td>
<td>15%</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.25,0.25,0.25,0.25</td>
<td>Data 1</td>
<td>10%</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data 2</td>
<td>15%</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.25,0.25,0.25,0.25</td>
<td>Data 1</td>
<td>5%</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data 2</td>
<td>15%</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.4,0.15,0.10,0.35</td>
<td>Data 1</td>
<td>10%</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data 2</td>
<td>15%</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0.4,0.15,0.10,0.35</td>
<td>Data 1</td>
<td>5%</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data 2</td>
<td>15%</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.4,0.15,0.10,0.35</td>
<td>Data 1</td>
<td>10%</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data 2</td>
<td>15%</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>0.4,0.15,0.10,0.35</td>
<td>Data 1</td>
<td>5%</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data 2</td>
<td>15%</td>
<td>2</td>
</tr>
</tbody>
</table>

4.3.4 Simulation Studies

4.3.4.1 Data generation

We used the MixSim package to generate mixtures of latent factors (Maitra, 2010)(Melnykov et al., 2012). Function \texttt{MixGOM} simulates cluster means and covariance matrices so that clusters overlap a user-specified amount (Maitra, 2010). We updated the \texttt{MixGOM} function such that the simulated data follows our model constraints $\sum_{l=1}^L \pi_l \nu_l = 0$ and $\Phi_l = I_K$. For that, we simulated $L-1$ clusters of $K$-dimensional factor data with spherical, homogeneous variances and generalized overlap value of 0.2 to achieve moderate clustering difficulty. We post-hoc scaled the simulated data so they had identity covariance matrices and added a $L$th cluster mean $\nu_L$ satisfying the constraint on $\nu_l, l = 1, 2, \ldots, L$. The number of samples per cluster were generated from a Multinomial distribution with the supplied proportions.
Our simulated data consist of two datasets \((M = 2)\) of 100 samples each with 2,000 or 2,500 features. To generate the observed data \(Y^m\), we first selected a small proportion \((\rho)\) of features to associate with each factor and then generated the non-zero entries in the weight matrix from a Gaussian distribution (Dataset 1: \(\mathcal{N}(1, 0.5)\) and Dataset 2: \(\mathcal{N}(2, 1)\)). The data matrix was generated by matrix multiplication of the simulated factor matrix and the simulated weight matrix plus noise generated from a Gaussian distribution with mean zero and random precision for each feature. The precisions were sampled from a Gamma distribution with different means \((\beta)\).

In our simulation study, we varied the proportion of samples in each cluster, the proportion of loaded features \(\rho\), and the noise \(\beta\) in eight different simulation settings. We simulated four clusters with sizes drawn from a Multinomial distribution with homogenous (all 0.25) or heterogenous \((0.40, 0.15, 0.10, 0.35)\) proportions (Table 4.1). To understand whether the methods are sensitive to the amount of sparsity, we varied the proportion of linked features to each factor, \(\rho \in \{5\%, 10\%, 15\%\}\). To assess the role of noise in the data, we varied the mean precision of the features, \(\beta \in \{0.5, 1, 2\}\). Table 4.1 reports the parameters used for the eight simulation settings. We simulated 20 replicates per simulation setting.

### 4.3.4.2 Method evaluation

We evaluated the performance of the different tools by comparing their ability to correctly infer the true clusters, recover important features, and estimate model parameters.

We used the adjusted RAND index (ARI) (Hubert and Arabie, 1985) to compare the estimated clusters to the true clusters. The ARI score measures the agreement between two partitions of the same data, with values near 0 indicating random partitions and values near 1 indicating good agreement. Next, we compared the significant genes using the estimated loadings to compare the methods’ ability to recover the important features. For MOFA-Cluster and MOFA, we use the weights of the features which are in the weight matrix of the final object. For MoCluster and iClusterBayes, the weights of the features are in the loadings of the final object. We use those scores to compare the truly linked features with the estimated significant features using Precision-
Recall (PR) curves (He and Garcia, 2009). We summarized each PR curve by calculating the area under the curve (AUC) for each simulation. A method achieving AUC near 1 is perfectly identifying the important features, while an AUC near 0 indicates a random selection of features.

Finally, we assessed the accuracy of parameter estimation by computing the distance between the estimated and true gene correlation matrices. The distance is calculated as:

\[ d_{\hat{R}} = \frac{||\hat{R} - R||_F}{||R||_F}, \]

where \( \hat{R} \) is the model-estimated correlation matrix, \( R \) is the true correlation matrix, and \( ||\cdot||_F \) is the Frobenius norm. The correlation matrix is calculated as

\[ R = W^m W^m + \Psi^m, \]

substituting estimated posterior expectations of weights and precisions for \( \hat{R} \). An accurate method will estimate the correlation matrix with smaller distance \( d_{\hat{R}} \).

4.3.5 Data processing

4.3.5.1 Colorectal cancer data

We used the colorectal cancer datasets from The Cancer Genome Atlas (TCGA) study (Muzny et al., 2012). We used the processed RNA-Seq and DNA methylation data downloaded from the UCSC’s Xena database (Goldman et al., 2020). A total of 278 samples had both the RNA-Seq and DNA Methylation data and were used in our analysis. The metadata of the samples were also downloaded including the survival data of the patients.

The processed RNA-Seq data consist of the \( \log_2 \) transformed Transcripts per Million (TPM) values. The processed DNA methylation data consist of the bimodal beta values that indicate whether the genes are hyper- or hypo-methylated with values between 0 and 1. Both gene and probes that were linked to the sex chromosomes were filtered out before the analysis. We made the beta values approximately normal by applying the logit transformation (Du et al., 2010). For our analysis, we selected the top quartile genes from the RNA-Seq data and the top 5,000 probes from the DNA methylation data based on the sample variance.
4.3.5.2 Preeclampsia data

To identify subtypes of preeclampsia using data integration, we identified three studies that collected both gene expression and DNA methylation data. We downloaded the raw data from GEO (accessions GSE44712 (Blair et al., 2013), GSE73377 (Martin et al., 2015), and GSE98224 (Leavey et al., 2018)). The microarray gene expression data was processed using the XPS package (Stratowa et al., 2016). After background correction, probes were detected, quantile normalized, and mapped to genes. The genes common to all three datasets were selected, resulting in 11,053 genes. The data were further corrected for batch effects using the ComBat method in the SVA package (Johnson et al., 2007). For processing the DNA methylation data, we first calculated the beta values using the methylated and unmethylated values for the datasets. Then, we used the ChAMP package for probe detection, filtering probes on sex chromosomes, and quantile normalization (Morris et al., 2014). The probes from the three datasets are then combined and corrected for batch effects using the ComBat method (Johnson et al., 2007). We retained the top 5,000 genes from the expression dataset and the top 5,000 probes from the DNA methylation dataset, ranked by variance.

4.3.6 Downstream analysis

4.3.6.1 Features linked to the clusters

We linked features to clusters by first using Analysis of Variance (ANOVA) to compute the \( p \)-value for the hypothesis of equal means across the clusters. After correcting for multiple testing (Benjamini and Yekutieli, 2001), we kept genes with \( q \)-value below 0.05. Among the kept genes, we compared their expression levels in the two clusters with highest mean expression levels using Student’s \( t \)-test. Finally, we assigned a gene to the cluster with higher expression if the \( t \)-test \( p \)-value \( \leq \) 0.05 and fold-change \( \geq \) 1.5.

4.3.6.2 Functional enrichment analysis

Genomic Regions Enrichment of Annotations Tool (GREAT) was used for functional enrichment analysis of gene sets using the following ontologies: GO Biological Process, and Disease
Ontology (McLean et al., 2010). Because we were investigating gene sets, significance was assessed using the hypergeometric statistic. Terms were considered significant if they had an FDR \( \leq 0.05 \), a fold-change \( \geq 2 \), and at least 5 genes from the input set.

4.3.6.3 Survival analysis

The associations of the identified disease subtypes and the phenotypic features of the patients were calculated using the survival R package (Therneau, 2020)(Terry M. Therneau and Patricia M. Grambsch, 2000). The survival probability for the different disease subtypes is calculated based on the overall survival time and last update status of the patients (dead or alive). The final survival plots and the significance of the disease subtypes were calculated using the log-rank test in the survminer package (Kassambara et al., 2020).

4.4 Results

4.4.1 Comparison with other tools using simulations

We compared MOFA-Cluster with other integrative clustering tools by using simulated data from the MixSim package and our own code. We vary several aspects of the data that affect the clustering and model estimation, including the cluster proportions, noise in the datasets, and the proportion of linked features. We ran MOFA, iClusterBayes, MoCluster, and lastly, our tool MOFA-Cluster, on the simulated data. MOFA was not able to find clusters for most of the simulated datasets. Suspecting MOFA’s random initialization of the latent factors as the culprit, we produced another method (MOFA-FAD) where the expected latent factors were initialized with fad, similar to the initialization strategy for MOFA-Cluster. After running each of the tools on the simulated data, we compared their ability to recover clusters, important features, and the correlation structure between features.

MOFA-Cluster is better able to estimate the true cluster membership. In Figure 4.1, MOFA-Cluster has the highest mean ARI for all the simulation conditions, including a statistically significant increase over MOFA (\( p \)-values <0.001) and MoCluster (\( p \)-values <0.05).
Correlation between features is a consequence of genes and other cellular components working together to carry out biological processes. Accurate estimation of this correlation is important for several downstream inferences, including network estimation, prediction of feature function and more. Factor analysis models produce an estimate of the feature correlation matrix based on the estimated loadings and feature variances. As seen in Figure 4.2, MOFA-Cluster better captures the true correlation between features than the other factor analysis models. It has an estimated correlation structure significantly closer to the true correlation structure than MOFA-FAD and iClusterBayes in most simulations. Because MoCluster doesn’t provide the feature variance estimates in the final results, we couldn’t produce an estimate of the feature correlation for MoCluster.

Finally, we compared the methods on their ability to identify which features are linked to the factors. Genes linked to each factor can be predicted by their loading values; larger absolute loadings convey a greater role of the gene in the corresponding factor. We compared these predictive absolute loadings with the true features using Precision-Recall curves. Assessed by the area under the curve, MOFA-Cluster better partitioned truly linked features compared to the other tools (Figure 4.3).

These results indicate that our approach of simultaneously learning factors and clusters works better than the ad-hoc approach of the other tools.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>MOFA Cluster</th>
<th>MOFA (FAD Init)</th>
<th>MOFA</th>
<th>MoCluster</th>
<th>iClusterBayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.56</td>
<td>0.46</td>
<td>0.03</td>
<td>0.11</td>
<td>0.48</td>
</tr>
<tr>
<td>2</td>
<td>0.56</td>
<td>0.41</td>
<td>0</td>
<td>0.35</td>
<td>0.51</td>
</tr>
<tr>
<td>3</td>
<td>0.63</td>
<td>0.41</td>
<td>0.10</td>
<td>0.04</td>
<td>0.50</td>
</tr>
<tr>
<td>4</td>
<td>0.53</td>
<td>0.44</td>
<td>0.03</td>
<td>0.20</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>0.72</td>
<td>0.69</td>
<td>0.10</td>
<td>0.12</td>
<td>0.62</td>
</tr>
<tr>
<td>6</td>
<td>0.71</td>
<td>0.64</td>
<td>0.03</td>
<td>0.40</td>
<td>0.57</td>
</tr>
<tr>
<td>7</td>
<td>0.64</td>
<td>0.62</td>
<td>0.10</td>
<td>0.02</td>
<td>0.62</td>
</tr>
<tr>
<td>8</td>
<td>0.67</td>
<td>0.58</td>
<td>0.20</td>
<td>0.27</td>
<td><strong>0.67</strong></td>
</tr>
</tbody>
</table>
4.4.2 Colorectal cancer subtype analysis

We downloaded the processed TCGA (Llosa et al., 2015) multi-omics colorectal cancer data from the UCSC’s Xena database (Goldman et al., 2020). Specifically, we retained 278 samples for which both RNA-Seq and DNA methylation data were available. Since MOFA-Cluster currently only handles Gaussian-distributed data, the beta values (Du et al., 2010) of the processed DNA methylation data were made approximately normal via the logit transformation. To reduce technical and other sources of noise in the data, we selected the top quartile of genes (4,156 genes) from the gene expression data and the top 5,000 probes from the DNA methylation data based on the variance across the samples.

As mentioned in the methods, we used fad to identify the initial factors after concatenating the datasets. Also, based on previous studies (Guinney et al., 2015) (Meng et al., 2016), we set four as the initial number of clusters. After running MOFA-Cluster, we identified three clusters explained by nine factors. As seen from the variance plot (Figure 4.4), the identified factors explain 42% and 52% of the variance in the gene expression and DNA methylation data, respectively. Using the fitted model to partition the observations using the posterior probabilities of membership in each cluster, we assign 104, 135, and 39 samples to each of the three clusters. (Our default initializer, using NbClust, finds two clusters, with the members of the third cluster distributed between the first two clusters.)

Previous studies have broadly divided colorectal cancer into two subtypes: microsatellite unstable (MSU) and microsatellite stable (MSS) but chromosomally unstable tumors (Llosa et al., 2015). Below, we outline evidence that cluster C3 likely represents the MSU subtype. MSU tumors are located in the right colon and are hypermutated. Using the accompanying TCGA mutation data, we found that patients in cluster C3 have a significantly higher mutation frequency (Figure 4.5). We further explored the three subtypes by identifying genes and probes with significantly higher expression and methylation in each subtype (see Methods). The identified DNA methylation probes were linked to genes by using the Illumina 450K microarray annotations. Based on the set of associated genes, we found that cluster C3 is significantly enriched for GO biological processes related
to the immune response and defense mechanisms, which agrees with the finding that the MSU subtype is associated with an immune microenvironment (Llosa et al., 2015). We also found that the five immune checkpoint genes, PD1, PDL1, CTLA4, LAG3, and IDO have significantly higher expression in cluster C3. These genes were reported to be upregulated in the MSU subtype (Llosa et al., 2015) (Figure 4.6). Several colorectal cancer drugs target these genes, including Nivolumab, which targets PD-1 (Overman et al., 2018). We also found that the gene SLC43A3, which codes for a member of the solute carrying family of proteins, is highly expressed in cluster C3 (Figure 4.6). This gene has previously been identified to play a significant role in the telomeric DNA damage in lung cancer and might be important for colorectal cancer progression (Mender et al., 2020).

Using the genes with significantly higher expression in cluster C1, we found that it is significantly enriched for genes involved in metabolic and biosynthetic processes. An increase in the expression of the metabolic genes has been associated with tumorigenesis and has previously been found in colorectal cancer subtypes identified after integrating the gene expression data from approximately three thousand samples (Guinney et al., 2015). However, in our results, we found that all the samples in cluster C1 are from female patients, whereas in Guinney et al. (2015), the subtype consists of samples from both the genders. These results are also significant as during processing, we already filtered out data coming from the sex chromosomes. We also found that this colorectal subtype also has the highest overall survival probability (Figure 4.7). In cluster C2, we found that the probes with significantly high methylation are linked with many genes from the Protocadherin family. This gene family is a subfamily of cadherin superfamily known to be involved in the homophilic cell adhesion (Hulpiau and van Roy, 2009). The high methylation of these genes indicates that these genes are downregulated in the cancer subtype C2. Downregulation of the cell adhesion genes has previously been reported to propagate the cancer cell to other tissues (Paschos et al., 2009). Therefore, the lower survival of subtype C2 may be due to the metastatic ability of this subtype. We also found that all the samples in this cancer subtype come from male patients. From the survival plot, we also found that the overall survival probability of subtype C2 is significantly lower than subtype C1. However, we couldn’t find any significant differences in the overall survival
probability between males and females (Figure 4.8). These results indicate that the lethality of colorectal cancer in males reported in a previous study (Yang et al., 2017) might be attributed to the different molecular mechanisms in subtype C2.

Overall, our results provide some new insights that can further support colorectal cancer research in the future.

### 4.4.3 Preeclampsia subtype analysis

Previous genome-wide gene expression comparisons have failed to find a significant difference between normal and PE placenta samples, likely due to the high variance among PE samples (Brew et al., 2016)(Redman, 2014). Researchers have hypothesized the possible existence of PE subtypes, as in cancers, with variable progression during pregnancy to explain the high intra-PE variability (Cox et al., 2011)(Roberts and Bell, 2013). Leavey et al. used the gene expressions from a large placental microarray dataset (330 samples) to find three PE subtypes using unsupervised clustering (Leavey et al., 2016). To see if we could verify these results, we used MOFA-Cluster to perform integrative clustering using gene expression and DNA methylation. Specifically, we downloaded placental datasets from GEO, consisting of 57 PE and 43 control samples from three different studies. The gene expression and DNA methylation data for these samples were processed and integrated, as explained in the methods. We selected the top 5,000 genes and DNA methylation probes based on the variance across all the samples for this analysis.

After identifying four factors and two clusters according to our initialization scheme (see Methods), MOFA-Cluster found two PE subtypes explained by four factors. The variance plot (Figure 4.9) shows that the identified factors explained 32% and 10% of the variance in the gene expression and DNA methylation data, respectively. However, cluster 1 consisted of 94 samples, with only six samples relegated to cluster 2. Either there are no recognizable subtypes in PE, the subtypes do not affect the variables measured in these data, or the data are too noisy to detect the clusters and more samples are needed to detect PE subtypes.
4.5 Discussion

In the last decade, significant technological advances have led to a myriad sequencing protocols that researchers can employ to interrogate the regulatory processes underlying disease. The resulting increase in multi-omics datasets has revealed the presence of disease subtypes, whose elucidation is a first step to greater treatment personalization. Multiple tools have been developed to find these subtypes from integrated omics data by first reducing the dimension and then cluster the samples on the reduced dimensions. The failure of current methods to consider the clusters during the dimension reduction step can lead to the loss of clustering information in the latent variables, and biased or misestimation of both the reduced dimension latent factors and the feature loadings, which link features to latent variables. To overcome these issues, we developed a mixture of factor score model that simultaneously performs data integration, reduction and clustering. The underlying model of MOFA-Cluster is closely connected to the recent factor analysis tool for data integration, MOFA (Argelaguet et al., 2018). The main difference between our tool, MOFA-Cluster, and MOFA is the simultaneous clustering of the samples by modeling the latent variables as a Gaussian mixture model.

We have demonstrated through simulations that MOFA-Cluster works significantly better than MOFA in terms of identifying clusters and estimating parameters, even after we improved MOFA initialization to be cluster-aware (MOFA-FAD). We also compared MOFA-Cluster with other integrative clustering tools based on an underlying factor analysis model, including MoCluster (Meng et al., 2016) and iClusterBayes (Mo et al., 2018), which had been previously found to work better than other tools (Pierre-Jean et al., 2019). We found that MOFA-Cluster works better than these methods, particularly in the estimation of model parameters. Among these tools, MOFA-FAD (our version of MOFA) and iClusterBayes, achieved the closest performance to our method. However, we found that the performance of MOFA-Cluster in identifying clusters and significant features, and estimating model parameters is higher than them even after 20 iterations for each simulation.

We also applied our method to a high dimensional multi-omics dataset from 278 colorectal cancer patients from TCGA (Llosa et al., 2015). We identified three subtypes of colorectal cancer
by integrating the gene expression and DNA methylation data. We linked our clusters to previously known genes that are current targets of drugs used to treat colorectal cancer, but we also identified some novel marker genes known to play a role in other cancers that may be future drug targets for colon cancer. These results show that MOFA-Cluster can simultaneously identify disease clusters and the important linked features.

Although MOFA-Cluster represents a demonstrable improvement over existing methods for multi-omics clustering, it also has remaining limitations. In our current model, MOFA-Cluster only works with the continuous data that can be modeled with the Gaussian distribution. There is no support in MOFA-Cluster for non-Gaussian datasets, including copy number variations and mutation counts. However, it should be possible to update MOFA-Cluster in the future to approximate non-Gaussian data with Gaussian pseudo data using second-order Taylor expansions similar to MOFA (Seeger and Bouchard, 2012)(Argelaguet et al., 2018). As in all clustering methods, initialization of MOFA-Cluster is particularly challenging. Poor initialization often leads to loss of all factors or clusters during iterations of the variational Bayes algorithm. Initialization was perhaps the biggest obstacle in getting MOFA-Cluster to work reliably, and providing the initializer partial truths about the data, such as the true number of clusters, could increase MOFA-Cluster performance dramatically (data not shown). We expect improved initialization methods, perhaps using ensemble clustering methods (Wu et al., 2018), or repeated random initialization, could improve MOFA-Cluster performance further.

In summary, MOFA-Cluster is the first tool to carry out simultaneous integration, data reduction and clustering of multi-omics datasets using factor analysis. The simulations and case studies demonstrate that MOFA-Cluster can benefit the research community by improving the discovery of disease subtypes and the features that distinguish them.

4.6 Supplementary Material

The supplementary material is available in Appendix B.
4.7 References


Roberts, J. M. and Bell, M. J. (2013). If we know so much about preeclampsia, why haven’t we cured the disease?


Figure 4.1  Results comparing the performance of MOFA-Cluster and other tools in terms of ARI scores to identify true clusters using the simulated data
Figure 4.2 Results comparing the performance of MOFA-Cluster and other tools in terms of the mean distance between the estimated model parameters and true parameters using the simulated data.
Figure 4.3 Results comparing the performance of MOFA-Cluster and other tools in estimating the significant features in terms of area under precision-recall curves using the simulated data.
Figure 4.4 Heatmap showing the percentage of variance shown explained by the different factors in the colorectal cancer subtype analysis
Figure 4.5  Boxplot showing the average number of mutations per patients in the different colorectal cancer subtypes
Figure 4.6  The boxplots showing the mean expression of the immune checkpoint genes in the colorectal cancer subtype C3 including PD1, PDL1, CTLA4, LAG3, IDO, and SLC43A3
Figure 4.7  Survival curves comparing the overall survival probability of the identified colorectal cancer subtypes
Figure 4.8  Survival curves comparing the overall survival probability of the male and female colorectal cancer patients
Figure 4.9  Heatmap showing the percentage of variance shown explained by the different factors in the preeclampsia subtype analysis
CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS

In this thesis, we utilized various publicly available NGS datasets to better understand the molecular mechanisms of early placenta development. We used transcriptomics data from a cell culture system that represents early placental development to study the gene regulatory networks. We also identified some novel genes that might play a significant role in placental development. In the process, we have also developed bioinformatics tools to help analyze NGS datasets. We integrated DNA methylation and gene expression data to investigate different subtypes of PE. To help with this analysis, we developed a statistical method to simultaneously integrate and cluster multi-omics datasets using factor analysis. In this chapter, I will go through the critical findings of the research in this dissertation and will also provide suggestions for future directions for the research in this field.

5.1 Specific Findings and Contributions

5.1.1 Deciphering transcriptional regulation in human embryonic stem cells specified towards a trophoblast fate

We compared the expression of BMP4 treated hESCs, which are used as an early TB cell model, with mesoderm and other placental cell lines to determine the true identity of the cells. We found that the expression profile of the early TB cell model is quite different from other placental cell lines and trophoblast cells from term placenta. We also found that the early TB cell model has high expression of many invasive genes compared to placental cell lines. We utilized co-expression network analysis to investigate the transcriptional networks of these cells and identified gene modules involved in cell migration and adhesion, processes that are critical during the beginning stages of placentation. We also identified several novel genes using a protein-protein interaction network that might play an important role in early placental development.
5.1.2 TissueEnrich: Tissue-specific gene enrichment analysis

We developed TissueEnrich, a tool to carry out tissue-specific gene enrichment using a gene set as input. We utilized various publicly available RNA-Seq data comprising of a large number of tissues (Human Protein Atlas, GTEX, and ENCODE) and used an algorithm developed by the Human Protein Atlas to identify tissue-specific genes. Using the processed data, we developed both a web tool and R package to carry out tissue-specific gene enrichment. We used the hypergeometric test to carry out enrichment analysis. We also demonstrated that TissueEnrich can assign tissue identities to cell clusters from single-cell RNA-Seq and differentiated embryonic stem cells.

5.1.3 Factor analysis-clustering model for integrating and clustering multi-omics dataset

We developed MOFA-Cluster, a statistical model that simultaneously integrates and clusters multi-omics data using factor analysis and a Gaussian mixture model. MOFA-Cluster has several advantages over the existing integrative clustering methods that carry out data integration and clustering separately. In MOFA-Cluster, the dimensional reduction process guides the process of cluster formation, allowing it to identify clusters with low signals more efficiently. We have shown that MOFA-Cluster performs better than other integrative clustering tools using simulation data. We also integrated the DNA Methylation and gene expression data from colorectal cancer and identified gender-specific subtypes. However, we couldn’t find any PE subtypes, which might be due to limited data availability, phenotypic variability within the data, or the fact that gene expression data was from microarrays.

5.2 Future Directions

5.2.1 Experimental validation of the identified novel genes at early placental development

While investigating the gene regulatory networks using the transcriptomic data from the early TB cell model, we identified five novel genes (ADCY7, EHHAD, IRF7, LUM, PLCB1) that might
have an important role in the early placental development. ADCY7 is known to be involved in
the endothelin 1 signaling pathway (Malek et al., 1993) and IRF7 in immune response pathways,
(Wang and Fish, 2012) which are relevant to placental development. We also found four of those five
genes also have high expression in placenta tissue at early developmental time points (Tuteja et al.,
2016). Based on these findings, these genes are good candidates to carry out further investigations
in placental development.

5.2.2 Method to integrate the data from different studies in TissueEnrich

TissueEnrich uses RNA-Seq data from various studies separately to carry out the tissue-specific
gene enrichment. TissueEnrich currently supports three different large-scale RNA-Seq datasets for
enrichment analysis, and users have to look through the results from them individually. This could
be made more user friendly by combing the tissue-specific gene information from the three datasets
before carrying out enrichment analysis. One method we could use to combine data could be to
calculate a confidence score for the tissue-specific genes based on the occurrence in different studies
and the number of tissues in those studies.

5.2.3 Support for non-Gaussian datasets in the MOFA-Cluster model

The MOFA-Cluster model is currently supported only to integrate and cluster NGS datasets
that can be modeled using a Gaussian distribution. The statistical model of MOFA-Cluster needs
to be updated to provide support for non-Gaussian NGS datasets, including copy number variations
and mutation counts. For achieving that, we can use the idea from Seeger and Bouchard (2012) to
use a second-order Taylor expansion to dynamically approximate the non-Gaussian data by using
Gaussian pseudo data.

5.3 References

stress is transcriptionally mediated and independent of protein kinase C and cAMP (vascular en-
dothelium/mechanical stress/signal transduction/protein synthesis/transcriptional control). Cell
Biology, 90:5999–6003.


APPENDIX A. SUPPLEMENTAL MATERIAL OF CHAPTER 2

Ashish Jain$^{1,2}$, Toshihiko Ezashi$^{3}$, R. Michael Roberts$^{3,4}$, and Geetu Tuteja$^{1,2}$

Authors and Affiliations

$^1$Bioinformatics and Computational Biology, Iowa State University, Ames, IA, USA

$^2$Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, USA

$^3$Division of Animal Sciences, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

$^4$Department of Biochemistry, University of Missouri, Columbia, MO, USA

Modified from a manuscript published in *Scientific Reports*

Table A.1 Summary of the differential expression analysis of the placental cell lines with ESCd.

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>DEGs</th>
<th>Genes upregulated in ESCd</th>
<th>Genes downregulated in ESCd</th>
</tr>
</thead>
<tbody>
<tr>
<td>BeWo</td>
<td>5478</td>
<td>3866</td>
<td>1612</td>
</tr>
<tr>
<td>HTR-8/SVneo</td>
<td>6516</td>
<td>4081</td>
<td>2435</td>
</tr>
<tr>
<td>JEG-3</td>
<td>6793</td>
<td>4756</td>
<td>2037</td>
</tr>
</tbody>
</table>
Table A.2  Detailed information about the RNA-Seq datasets used for analysis.

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>Study Accession</th>
<th>Read Length</th>
<th>HISAT2 Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESCd &gt;70mm</td>
<td>GSE72712</td>
<td>50bp</td>
<td>82.31%</td>
</tr>
<tr>
<td>ESCd 40-70mm</td>
<td>GSE72712</td>
<td>50bp</td>
<td>82.09%</td>
</tr>
<tr>
<td>ESCd &lt;40mm</td>
<td>GSE72712</td>
<td>50bp</td>
<td>82.28%</td>
</tr>
<tr>
<td>ESCu</td>
<td>GSE72712</td>
<td>50bp</td>
<td>82.56%</td>
</tr>
<tr>
<td>PHTu</td>
<td>GSE73016</td>
<td>50bp</td>
<td>82.10%</td>
</tr>
<tr>
<td>PHTd</td>
<td>GSE73016</td>
<td>50bp</td>
<td>81.49%</td>
</tr>
<tr>
<td>Mesoderm Cells (Day 2 of</td>
<td>GSE64417</td>
<td>100bp</td>
<td>93.90%</td>
</tr>
<tr>
<td>differentiation)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BeWo Cells</td>
<td>GSE66962</td>
<td>100bp</td>
<td>93.90%</td>
</tr>
<tr>
<td>JEG-3 Cells</td>
<td>GSE79779</td>
<td>100bp</td>
<td>96.30%</td>
</tr>
<tr>
<td>HTR-8/SVneo Cells</td>
<td>GSE85995</td>
<td>75bp</td>
<td>96.40%</td>
</tr>
<tr>
<td>Term Syncytiotrophoblast</td>
<td>GSE87725</td>
<td>75bp</td>
<td>93.5%</td>
</tr>
</tbody>
</table>
Figure A.1  a-e) Barplots showing the enrichment (-Log_{10}(P-Value)) of tissue-specific genes in ESCd 40-70, ESCd <40, PHTu, PHTd, and ESCu cells using protein atlas database. Placenta-specific genes are enriched in all cell types except ESCu.
Figure A.2 Heatmap showing the significant DEGs (P-Adjusted ≤ 0.01 and abs(Fold-Change) ≥ 2) between the ESCu and ESCd groups. ESCd groups have few differentially expressed genes between one another.
Figure A.3  Plot showing the enrichment of mesoderm development terms in DEGs over-expressed in ESCd or mesoderm. The dotted lines represent the thresholds for FDR (0.05) and fold enrichment (1.5) of the terms.

Figure A.4  PCA plot based on the 1,000 most highly expressed genes in ESCd. PCA was carried out using ESCd, ESCu, placental cell lines, term placental cells, and mesoderm cells. The first three principal components are plotted (cumulative variance of 81%).
### Module−Cell Type Significance

<table>
<thead>
<tr>
<th>Module</th>
<th>Significance</th>
<th>Significance</th>
<th>Significance</th>
<th>Significance</th>
<th>Significance</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>−0.33</td>
<td>−0.19</td>
<td>−0.18</td>
<td>0.94</td>
<td>0.16</td>
<td>−0.27</td>
</tr>
<tr>
<td></td>
<td>(0.1)</td>
<td>(0.5)</td>
<td>(0.6)</td>
<td>(7e−11)</td>
<td>(0.5)</td>
<td>(0.2)</td>
</tr>
<tr>
<td>M2</td>
<td>−0.69</td>
<td>0.26</td>
<td>0.26</td>
<td>0.36</td>
<td>0.5</td>
<td>−0.62</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.4)</td>
<td>(0.4)</td>
<td>(0.1)</td>
<td>(0.03)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>M3</td>
<td>−0.44</td>
<td>−0.2</td>
<td>−0.2</td>
<td>0.92</td>
<td>−0.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.5)</td>
<td>(0.5)</td>
<td>(7e−10)</td>
<td>(0.2)</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>0.29</td>
<td>0.13</td>
<td>0.26</td>
<td>0.19</td>
<td>−0.052</td>
<td>−0.36</td>
</tr>
<tr>
<td></td>
<td>(0.2)</td>
<td>(0.3)</td>
<td>(0.4)</td>
<td>(0.1)</td>
<td>(0.9)</td>
<td>(3e−14)</td>
</tr>
<tr>
<td>M5</td>
<td>0.51</td>
<td>−0.46</td>
<td>−0.46</td>
<td>0.27</td>
<td>0.37</td>
<td>−0.52</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.3)</td>
<td>(0.1)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>M6</td>
<td>0.7</td>
<td>−0.37</td>
<td>−0.09</td>
<td>0.42</td>
<td>−0.4</td>
<td>−0.51</td>
</tr>
<tr>
<td></td>
<td>(2e−04)</td>
<td>(0.2)</td>
<td>(0.1)</td>
<td>(0.1)</td>
<td>(0.2)</td>
<td></td>
</tr>
<tr>
<td>M7</td>
<td>−0.37</td>
<td>0.57</td>
<td>0.11</td>
<td>−0.41</td>
<td>−0.3</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.01)</td>
<td>(0.0)</td>
<td>(0.08)</td>
<td>(0.2)</td>
<td>(0.00005)</td>
</tr>
<tr>
<td>M8</td>
<td>0.29</td>
<td>0.48</td>
<td>0.82</td>
<td>−0.42</td>
<td>−0.45</td>
<td>−0.45</td>
</tr>
<tr>
<td></td>
<td>(0.2)</td>
<td>(0.04)</td>
<td>(0.02)</td>
<td>(0.08)</td>
<td>(0.05)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>M9</td>
<td>−0.49</td>
<td>0.64</td>
<td>0.64</td>
<td>−0.23</td>
<td>−0.046</td>
<td>−0.27</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.3)</td>
<td>(0.9)</td>
<td>(0.2)</td>
</tr>
<tr>
<td>M10</td>
<td>−0.27</td>
<td>−0.18</td>
<td>−0.2</td>
<td>−0.23</td>
<td>0.029</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>(0.2)</td>
<td>(0.5)</td>
<td>(0.5)</td>
<td>(0.3)</td>
<td>(0.9)</td>
<td>(2e−15)</td>
</tr>
<tr>
<td>M11</td>
<td>−0.37</td>
<td>−0.27</td>
<td>−0.91</td>
<td>0.28</td>
<td>0.53</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.3)</td>
<td>(0.008)</td>
<td>(0.3)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>M12</td>
<td>0.41</td>
<td>−0.58</td>
<td>−0.58</td>
<td>−0.22</td>
<td>0.32</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.01)</td>
<td>(0.009)</td>
<td>(0.3)</td>
<td>(0.2)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>M13</td>
<td>0.71</td>
<td>−0.34</td>
<td>−0.34</td>
<td>−0.46</td>
<td>−0.29</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>(2e−04)</td>
<td>(0.2)</td>
<td>(0.2)</td>
<td>(0.08)</td>
<td>(0.2)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>M14</td>
<td>0.95</td>
<td>−0.16</td>
<td>−0.12</td>
<td>−0.42</td>
<td>−0.23</td>
<td>−0.42</td>
</tr>
<tr>
<td></td>
<td>(5e−13)</td>
<td>(0.5)</td>
<td>(0.6)</td>
<td>(0.08)</td>
<td>(0.3)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>M15</td>
<td>0.72</td>
<td>−0.14</td>
<td>−0.12</td>
<td>−0.012</td>
<td>−0.84</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>(2e−04)</td>
<td>(0.5)</td>
<td>(0.6)</td>
<td>(1)</td>
<td>(5e−07)</td>
<td>(0.2)</td>
</tr>
<tr>
<td>M16</td>
<td>0.53</td>
<td>0.2</td>
<td>0.15</td>
<td>−0.89</td>
<td>−0.57</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.5)</td>
<td>(0.6)</td>
<td>(7e−26)</td>
<td>(0.01)</td>
<td>(0.2)</td>
</tr>
</tbody>
</table>

Figure A.5  Heatmap showing the relationship between the co-expressed modules and cell type. Each block in the heatmap has the correlation score and adjusted p-values between the respective co-expressed module and cell type.
Figure A.6  Box plots (left panels) showing expression values in ESCd and ESCu for genes in modules M6 (a), M14 (b), and M16 (c). Gene expression is significantly higher in ESCd. The corresponding right panels show the distribution of the number of genes upregulated in ESCd in 10,000 sets of random gene modules.
Figure A.7  Network showing the protein-protein interaction between the bottleneck genes and their neighbors for module M6. The color of the node represents the betweenness score of the genes.

Figure A.8  PCA plot based on the 1,000 most highly expressed genes in ESCd. The first three principal components are plotted with a cumulative variance of 90%. PCA was carried out using the ESCd, ESCu, and term placenta cells.
Figure A.9  Plot showing the expression of the identified novel genes in E7.5 and E9.5 mouse placenta. The error bars represent the variance in the samples. The dotted line represents the expression level of 5 Fragments Per Kilobase per Million mapped reads (FPKM).

Figure A.10  Distribution of Log$_2$(TPM) for genes that are trophectoderm specific (left) and genes that have a highly significant bias in expression in polar trophectoderm (right).
Figure A.11  Plots showing the $R^2$ values (left panel) and mean connectivity (right panel) of various soft threshold values required for making the co-expression network scale free. The threshold for the $R^2$ value is 0.9 shown as the red horizontal line.
Figure A.12  Dendrogram showing clustering of the modules based on the correlation between the module’s eigengene expression.
Figure A.13  Heatmap showing the gene overlap between co-expressed modules based on the module membership (KME) of genes.
APPENDIX B. SUPPLEMENTAL MATERIAL OF CHAPTER 4

Ashish Jain$^{1,2}$, Geetu Tuteja$^{1,2}$, and Karin S. Dorman$^{1,2,3}$

Authors and Affiliations

1Bioinformatics and Computational Biology, Iowa State University, Ames, 50010, USA
2Genetics, Development, and Cell Biology, Iowa State University, Ames, 50010, USA
3Department of Statistics, Iowa State University, Ames, 50010, USA

Manuscript in preparation to be published in *Bioinformatics*

**B.1 Introduction**

MOFA-Cluster is a mixture of factor score model that carry out the simultaneous integration and clustering of the multi-omics dataset. As specified in section 4.3.1, we used a Bayesian approach to estimate the latent factors from the model, similar to MOFA$.^3$ However, we assumed the latent variables ($Z$) to be a Gaussian homogenous Mixture density instead of the standard multivariate Normal density. Similar to MOFA, we used the Variational Bayes (VB) approach to estimate the model parameters$.^3$ We also monitor the training and asses the convergence using Evidence Lower Bound (ELBO), as mentioned in 4.3.2. Here, we derived the posterior distributions and ELBO of the model parameters used to update the model parameters and assessing the convergence of the VB algorithm.

**B.2 Updated posterior equations of model parameters**

**B.2.1 Latent Variables ($V, Z$)**

We focus on obtaining the next iterate of $q^{t+1}_x(x) = \prod_{n=1}^N q^{t+1}_n(x_n)$ for each set of hidden variables $x_n$, consisting of factors $z_{ns}$ and cluster assignments $v_{ns}$.
Assuming more posterior independence. If we further assume

\[
q_n(z_{nk}, \nu_{nk}) = q_n(\nu_{nk}) \prod_{k=1}^{K} q_n(z_{nk}),
\]

then

\[
\begin{align*}
\ln q_n(z_{nk}) &= \mathbb{E}_q \left\{ \sum_{m=1}^{M} \sum_{d=1}^{D_m} - \frac{\tau_d^m}{2} \left( y_{nd}^m - \sum_{k'=1}^{K} s_{dk'}^m \hat{w}_{dk'}^m z_{nk'} \right)^2 + \sum_{l=1}^{L} v_{nl} \left[ \ln \pi_l - \frac{1}{2} \ln(2\pi) - \frac{1}{2} (z_{nk} - \nu_{kl})^2 \right] \right\} \\
&\quad + C_{z_{nk1}} \\
&= \mathbb{E}_q \left\{ \sum_{m=1}^{M} \sum_{d=1}^{D_m} - \frac{\tau_d^m}{2} \left[ y_{nd}^m + \sum_{k'=1}^{K} s_{dk'}^m \hat{w}_{dk'}^m z_{nk'}^2 \right]^2 - 2 y_{nd}^m \sum_{k'=1}^{K} s_{dk'}^m \hat{w}_{dk'}^m z_{nk'} \right\} \\
&\quad + \sum_{l=1}^{L} v_{nl} \left[ -\frac{1}{2} \left( z_{nk}^2 + \nu_{kl}^2 - 2 z_{nk} \nu_{kl} \right) \right] + C_{z_{nk2}} \\
&= \mathbb{E}_q \left\{ -\frac{1}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} (\tau_d^m s_{dk}^m \hat{w}_{dk}^m)^2 z_{nk}^2 + \sum_{l=1}^{L} v_{nl} z_{nk}^2 \right\} + C_{z_{nk3}} \\
&= \mathbb{E}_q \left\{ -\frac{1}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} (\tau_d^m s_{dk}^m \hat{w}_{dk}^m)^2 z_{nk}^2 - 2 \sum_{m=1}^{M} \sum_{d=1}^{D_m} s_{dk}^m \hat{w}_{dk}^m z_{nk} \right\} + C_{z_{nk4}} \\
&= \mathbb{E}_q \left\{ -\frac{1}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} (\tau_d^m s_{dk}^m \hat{w}_{dk}^m)^2 z_{nk}^2 + \sum_{l=1}^{L} v_{nl} z_{nk} \right\} + C_{z_{nk4}} \\
&= -\frac{1}{2} \mathbb{E}_q \left\{ \sum_{m=1}^{M} \sum_{d=1}^{D_m} \mathbb{E}_q \left[ \tau_d^m s_{dk}^m \hat{w}_{dk}^m \right] \mathbb{E}_q \left[ v_{nl} \right] \right\} \\
&\quad - 2 z_{nk} \left( \sum_{m=1}^{M} \sum_{d=1}^{D_m} \mathbb{E}_q \left[ \tau_d^m s_{dk}^m \hat{w}_{dk}^m \right] \left( y_{nd}^m - \sum_{j\neq k} \mathbb{E}_q \left[ \hat{w}_{dj}^m \right] z_{nj} \right) + \sum_{l=1}^{L} \mathbb{E}_q \left[ v_{nl} \right] \mathbb{E}_q \left[ v_{kl} \right] \right) + C_{z_{nk4}}
\end{align*}
\]

where \( \mathbb{E}_q[\cdot] \) is the variational posterior expectation over parameters \( \theta \) and all latent states except \( z_{nk} \). Thus, we observe that
\[ q_n(z_{nk}) = \mathcal{N}(z_{nk} | \mu_{znk}, \sigma^2_{znk}), \]

with

\[
\sigma_{znk}^{(t+1)} = \left( \sum_{m=1}^{M} \sum_{d=1}^{D_m} \mathbb{E}_q \left[ \tau_d^m s_d^m z_{dk}^m \right] + \sum_{l=1}^{L} \mathbb{E}_q[v_{nl}] \right)^{-1/2}
\]

\[
\mu_{znk}^{(t+1)} = \left( \sigma_{znk}^{(t+1)} \right)^2 \left[ \sum_{m=1}^{M} \sum_{d=1}^{D_m} \mathbb{E}_q \left[ \tau_d^m s_d^m z_{dk}^m \right] \left( \nu_{ml}^m - \sum_{j \neq k} \mathbb{E}_q[s_d^m \hat{w}_d^m] \mathbb{E}_q[z_{nj}] \right) + \sum_{l=1}^{L} \mathbb{E}_q[v_{nl}] \mathbb{E}_q[\nu_{kl}] \right]
\]

Meanwhile,

\[
\ln q_n(v_{n*}) = \sum_{l=1}^{L} v_{nl} \mathbb{E}_q \left[ \ln \pi_l - \frac{1}{2} \sum_{k=1}^{K} (z_{nk} - \nu_{kl})^2 \right] + C_{v_{n*}} = \sum_{l=1}^{L} v_{nl} \ln \rho_{nl} + C_{v_{n*}},
\]

where

\[
\ln \rho_{nl} = \mathbb{E}_q[\ln \pi_l] - \frac{1}{2} \sum_{k=1}^{K} \left( \mathbb{E}_q[z_{nk}^2] + \mathbb{E}_q[\nu_{kl}^2] - 2 \mathbb{E}_q[z_{nk}] E_{\theta}[\nu_{kl}] \right)
\]

clearly has a Multinoulli(\(\mu_{v_{n1}}, \mu_{v_{n2}}, \ldots, \mu_{v_{nL}}\)) distribution with probabilities

\[
\mu_{vn_l} = \frac{\rho_{nl}}{\sum_{j=1}^{L} \rho_{nj}}.
\]

**B.2.2 Cluster Means (\(\nu\))**

The variational distribution of \(\nu_l\) for \(l < L\) is

\[
\ln q_{\theta}(\nu_l) = \mathbb{E}_q \left\{ -\frac{1}{2} \sum_{n=1}^{N} v_{nl} \left( z_{n*}^T - \nu_l \right) \left( z_{n*}^T - \nu_l \right) - \frac{1}{2} \nu_l^T \nu_l \right\} + C_{bv_{l1}}
\]

\[
= \mathbb{E}_q \left\{ -\frac{1}{2} \sum_{n=1}^{N} v_{nl} \left( z_{n*}^T z_{n*} + \nu_l^T \nu_l - 2 z_{n*} \nu_l \right) - \frac{1}{2} \nu_l^T \nu_l \right\} + C_{bv_{l1}}
\]

\[
= \mathbb{E}_q \left\{ -\frac{1}{2} \left[ \nu_l^T \nu_l \left( \sum_{n=1}^{N} v_{nl} + 1 \right) - 2 \nu_l^T \sum_{n=1}^{N} v_{nl} z_{n*}^T \right] \right\} + C_{bv_{l2}}
\]

\[
= -\frac{1}{2} \nu_l^T \nu_l \left( \sum_{n=1}^{N} \mathbb{E}_q[v_{nl}] + 1 \right) - 2 \nu_l^T \sum_{n=1}^{N} \mathbb{E}_q[v_{nl}] \mathbb{E}_q[z_{n*}]^T \right\} + C_{v_{l2}}
\]

We recognize the multivariate normal distribution
\( \nu_1 \sim \mathcal{N}(\mu_{\nu_1}, \sigma_{\nu_1}^2 I_K) \),

where

\[
\sigma_{\nu_1}^2 = \left( \sum_{n=1}^{N} E_q[v_{n1}] + 1 \right)^{-1}
\]

\[
\mu_{\nu_1} = \sigma_{\nu_1}^2 \sum_{n=1}^{N} E_q[v_{n1}] (E_q[z_{n*}])^T.
\]

There is independence, so

\( \nu_{k1} \sim \mathcal{N}\left(\sigma_{\nu_1}^2 \sum_{n=1}^{N} E_q[v_{n1}] E_q[z_{nk}], \sigma_{\nu_1}^2\right) \).

For \( l = L \), we have

\[
\ln q_{\theta}(\nu_L) = E_q \left\{ -\frac{1}{2} \sum_{n=1}^{N} v_{nL} (z_{n*}^T - \nu_L)^T (z_{n*}^T - \nu_L) \right\} + C_{\nu_L1}
\]

\[
= E_q \left\{ -\frac{1}{2} \left( \nu_L^T \nu_L \sum_{n=1}^{N} v_{nL} - 2 \sum_{n=1}^{N} v_{nL} z_{n*} \nu_L \right) \right\} + C_{\nu_L2}
\]

\[
= -\frac{1}{2} \left[ \nu_L^T \nu_L \sum_{n=1}^{N} E_q[v_{nL}] - 2 \left( \sum_{n=1}^{N} E_q[v_{nL}] E_q[z_{n*}] \right) \nu_L \right] + C_{\nu_L2}
\]

Once again, we recognize

\( \nu_L \sim \mathcal{N}(\mu_{\nu_L}, \sigma_{\nu_L}^2 I_K) \),

where

\[
\sigma_{\nu_L}^2 = \left( \sum_{n=1}^{N} v_{nL} \right)^{-1}
\]

and

\[
\mu_{\nu_L} = \sigma_{\nu_L}^2 \sum_{n=1}^{N} E_q[v_{nL}] (E_q[z_{n*}])^T.
\]

Again, there is independence, so

\( \nu_{kL} \sim \mathcal{N}\left(\sigma_{\nu_L}^2 \sum_{n=1}^{N} E_q[v_{nL}] E_q[z_{nk}], \sigma_{\nu_L}^2\right) \).
B.2.3 Mixture Proportions ($\pi$)

The variational distribution for $\pi$ is

$$\ln q_\theta(\pi) = E_q \left[ \sum_{n=1}^{N} \sum_{l=1}^{L} v_{nl} \ln \pi_l + (\alpha_0^\pi - 1) \sum_{l=1}^{L} \ln \pi_l \right] + C_\pi = \sum_{l=1}^{L} \left( \alpha_0^\pi + \sum_{n=1}^{N} E_q [v_{nl}] - 1 \right) \ln \pi_l + C_\pi,$$

which we recognize as Dirichlet($\alpha_1^\pi, \alpha_2^\pi, \ldots, \alpha_L^\pi$), where

$$\alpha_l^\pi = \alpha_0^\pi + \sum_{n=1}^{N} E_q [v_{nl}].$$

B.2.4 Spike and Slab Weights ($\hat{W}$ and $S$)

We assume the following factorization of the spike and slab weights,

$$q_\theta(\hat{W}, S) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} q_\theta(\hat{w}_{dk}^m, s_{dk}^m) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} q_\theta(\hat{w}_{dk}^m | s_{dk}^m) q_\theta(s_{dk}^m),$$

so the variational distribution is

$$\ln [q_\theta(\hat{w}_{dk}^m | s_{dk}^m) q_\theta(s_{dk}^m)]$$

$$= E_q \left\{ -\frac{1}{2} \sum_{n=1}^{N} \sum_{l=1}^{L} \left( \frac{\theta_k}{\alpha_k^m} \hat{w}_{dk}^m \hat{s}_{dk}^m \sum_{j \neq k} s_{dj}^m \hat{w}_{dj}^m \hat{z}_{nk} - \frac{K}{2} \sum_{j \neq k} s_{dj}^m \hat{w}_{dj}^m \hat{z}_{nj} \right)^2 + \hat{s}_{dk}^m \ln \theta_k^m + (1 - \hat{s}_{dk}^m) \ln (1 - \theta_k^m) - \frac{1}{2} \alpha_k^m \hat{w}_{dk}^m - \frac{1}{2} \ln(2\pi/\alpha_k^m) \right\} + C_{w_{dk}^m s_{dk}^m 1}$$

$$= E_q \left\{ -\frac{1}{2} \sum_{n=1}^{N} \sum_{l=1}^{L} \left( \frac{\theta_k}{1 - \theta_k^m} \hat{w}_{dk}^m \sum_{j \neq k} s_{dj}^m \hat{w}_{dj}^m \hat{z}_{nk} - 2 \hat{s}_{dk}^m \hat{w}_{dk}^m \hat{y}_{nd} - \frac{K}{2} \sum_{j \neq k} s_{dj}^m \hat{w}_{dj}^m \hat{z}_{nj} \right) \hat{w}_{dk}^m \right\}

+ \hat{s}_{dk}^m \ln \theta_k^m - \frac{1}{2} \alpha_k^m \hat{s}_{dk}^m \hat{w}_{dk}^m \hat{w}_{dk}^m + C_{w_{dk}^m s_{dk}^m 2}$$

$$= -\frac{1}{2} \hat{s}_{dk}^m E_q [\tau_d^m] \sum_{n=1}^{N} \left[ E_q [\hat{z}_{nk}^m] \hat{w}_{dk}^m - 2 E_q [\hat{z}_{nk}^m] \hat{w}_{dk}^m \hat{y}_{nd} - \sum_{j \neq k} E_q [s_{dj}^m \hat{w}_{dj}^m] E_q [\hat{z}_{nj}^m] \hat{w}_{dk}^m \right]$$

$$+ \hat{s}_{dk}^m \ln E_q \left[ \frac{\theta_k^m}{1 - \theta_k^m} \right] - \frac{1}{2} \hat{s}_{dk}^m \hat{w}_{dk}^m \hat{w}_{dk}^m + C_{w_{dk}^m s_{dk}^m 2}$$
\[
= -\frac{1}{2} s_{dk}^m \mathbb{E}_q [\tau_d^m] \left[ \left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] \right) + \frac{\mathbb{E}_q [\alpha_k^m]}{\mathbb{E}_q [\tau_d^m]} \right] \hat{w}_{dk}^m
- 2 \left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] y_{nd}^m - \sum_{j \neq k} \mathbb{E}_q [s_{dj}^m \hat{w}_{dj}^m] \sum_{n=1}^N \mathbb{E}_q [z_{nk}] \mathbb{E}_q [z_{nj}] \right) \hat{w}_{dk}^m
+ s_{dk} \ln \mathbb{E}_q \left[ \frac{\theta_k^m}{1 - \theta_k^m} \right]
\]
\[
C_{w_{dk}^m s_{dk}^m}
\]
\[
= s_{dk}^m \left[ -\frac{1}{2} \mathbb{E}_q [\tau_d^m] \left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] + \frac{\mathbb{E}_q [\alpha_k^m]}{\mathbb{E}_q [\tau_d^m]} \right) \right]
\times \left( \hat{w}_{dk}^m - \frac{\left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] y_{nd}^m - \sum_{j \neq k} \mathbb{E}_q [s_{dj}^m \hat{w}_{dj}^m] \sum_{n=1}^N \mathbb{E}_q [z_{nk}] \mathbb{E}_q [z_{nj}] \right)}{\sum_{n=1}^N \mathbb{E}_q [z_{nk}] + \frac{\mathbb{E}_q [\alpha_k^m]}{\mathbb{E}_q [\tau_d^m]}} \right)^2
- \frac{1}{2} \ln \left( \frac{2\pi}{\mathbb{E}_q [\tau_d^m] \left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] + \frac{\mathbb{E}_q [\alpha_k^m]}{\mathbb{E}_q [\tau_d^m]} \right)} \right)
\]
\[
+ \frac{\mathbb{E}_q [\tau_d^m]}{2} \left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] y_{nd}^m - \sum_{j \neq k} \mathbb{E}_q [s_{dj}^m \hat{w}_{dj}^m] \sum_{n=1}^N \mathbb{E}_q [z_{nk}] \mathbb{E}_q [z_{nj}] \right)^2
\]
+ \frac{1}{2} \ln \left( \frac{2\pi}{\mathbb{E}_q [\tau_d^m] \left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] + \frac{\mathbb{E}_q [\alpha_k^m]}{\mathbb{E}_q [\tau_d^m]} \right)} \right) + \ln \mathbb{E}_q \left( \frac{\theta_k^m}{1 - \theta_k^m} \right)
\]
\[
C_{w_{dk}^m s_{dk}^m}
\]

We have used the fact that \( \hat{w}_{dk}^{m^2} = s_{dk}^m \hat{w}_{dk}^m \). We see \( s_{dk}^m \) has a Bernoulli distribution. Let

\[
\lambda_{dk}^m = \ln \left( \frac{\theta_k^m}{1 - \theta_k^m} \right) + \frac{1}{2} \ln \left( \frac{\alpha_k^m}{\tau_d^m} \right)
- \frac{1}{2} \ln \left( \sum_{n=1}^N z_{nk}^2 + \frac{\alpha_k^m}{\tau_d^m} \right)
+ \frac{\tau_d^m}{2} \left( \sum_{n=1}^N y_{nd}^m z_{nk} - \sum_{j \neq k} \mathbb{E}_q [s_{dj}^m \hat{w}_{dj}^m] \sum_{n=1}^N z_{nk} z_{nj} \right)^2
\]

\[
\lambda_{dk}^m = \mathbb{E}_q \left[ \ln \left( \frac{\theta_k^m}{1 - \theta_k^m} \right) \right] + \frac{1}{2} \ln \frac{2\pi}{\mathbb{E}_q [\tau_d^m]}
- \frac{1}{2} \ln \left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] + \frac{\mathbb{E}_q [\alpha_k^m]}{\mathbb{E}_q [\tau_d^m]} \right)
\]
\[
+ \frac{\mathbb{E}_q [\tau_d^m]}{2} \left( \sum_{n=1}^N \mathbb{E}_q [z_{nk}] y_{nd}^m - \sum_{j \neq k} \mathbb{E}_q [s_{dj}^m \hat{w}_{dj}^m] \sum_{n=1}^N \mathbb{E}_q [z_{nk}] \mathbb{E}_q [z_{nj}] \right)^2
\]
\[
\sum_{n=1}^N \mathbb{E}_q [z_{nk}] + \frac{\mathbb{E}_q [\alpha_k^m]}{\mathbb{E}_q [\tau_d^m]}
\]
Since the other coefficient is 0, we have

$$\mu_{s_{dk}}^m = \frac{e^{\lambda_{dk}}}{1 + e^{\lambda_{dk}}}.$$

Update for $q_\theta(\hat{w}_{dk}^m)$:

$$q_\theta(\hat{w}_{dk}^m \mid s_{dk}^m = 0) = N(\hat{w}_{dk}^m \mid 0, 1/\alpha_k^m)$$

$$q_\theta(\hat{w}_{dk}^m \mid s_{dk}^m = 1) = N(\hat{w}_{dk}^m \mid \mu_{w_{dk}}^m, \sigma_{w_{dk}}^m),$$

where

$$\mu_{w_{dk}}^m = \frac{\sum_{n=1}^N y_{nd} \mathbb{E}_q[z_{nk}] - \sum_{j \neq k} \mathbb{E}_q[s_{dkj} \hat{w}_{dj}^m] \sum_{n=1}^N \mathbb{E}_q[z_{nk}] \mathbb{E}_q[z_{nj}]}{\sum_{n=1}^N \mathbb{E}_q[z_{nk}] + \mathbb{E}_q[\alpha_k^m]}$$

$$\sigma_{w_{dk}}^2 = \frac{1}{\mathbb{E}_q[\tau_d^m]} \left( \sum_{n=1}^N \mathbb{E}_q[z_{nk}^2] + \mathbb{E}_q[\alpha_k^m] \mathbb{E}_q[\tau_d^m] \right).$$

Taken together this means that we can update $q_\theta(\hat{w}_{dk}^m, s_{dk}^m)$ using:

$$q_\theta(\hat{w}_{dk}^m \mid s_{dk}^m) q_\theta(s_{dk}^m) = \mathcal{N} \left( \hat{w}_{dk}^m \mid s_{dk}^m \mu_{w_{dk}}^m, s_{dk}^m \sigma_{w_{dk}}^m + (1 - s_{dk}^m)/\alpha_k^m \right) \times \left( \mu_{s_{dk}}^m \right)^{s_{dk}^m} \left( 1 - \mu_{s_{dk}}^m \right)^{1 - s_{dk}^m}$$

### B.2.5 Sparsity Parameter ($\theta$)

Variational distribution:

$$\ln q_\theta(\theta_k^m) = \mathbb{E}_q \left\{ \sum_{d=1}^{D_m} [s_{dk}^m \ln \theta_k^m + (1 - s_{dk}^m) \ln (1 - \theta_k^m)] + (\alpha_k^m - 1) \ln \theta_k^m + \left( b_0^\theta - 1 \right) \ln (1 - \theta_k^m) \right\}$$

$$+ C_{\theta_k^m}$$

$$= \sum_{d=1}^{D_m} \mathbb{E}_q [s_{dk}^m] \ln \theta_k^m + \sum_{d=1}^{D_m} (1 - \mathbb{E}_q [s_{dk}^m]) \ln (1 - \theta_k^m) + (\alpha_k^m - 1) \ln \theta_k^m + \left( b_0^\theta - 1 \right) \ln (1 - \theta_k^m)$$

$$+ C_{\theta_k^m}$$

$$= \left( \alpha_k^m + \sum_{d=1}^{D_m} \mathbb{E}_q [s_{dk}^m] - 1 \right) \ln \theta_k^m + \left( b_0^\theta + D_m - \sum_{d=1}^{D_m} \mathbb{E}_q [s_{dk}^m] - 1 \right) \ln (1 - \theta_k^m) + C_{\theta_k^m},$$
which we recognize to be a Beta \( \hat{a}_{\theta mk}, \hat{b}_{\theta mk} \) with

\[
\hat{a}_{\theta mk} = a_0^\theta + \sum_{d=1}^{D_m} E_q [s_{dk}^m] \\
\hat{b}_{\theta mk} = b_0^\theta + D_m - \sum_{d=1}^{D_m} E_q [s_{dk}^m].
\]

**B.2.6 ARD Precision \((\alpha)\)**

Variational distribution:

\[
\ln q_\theta(\alpha_k^m) = E_q \left\{ \sum_{d=1}^{D_m} \left[ \frac{1}{2} \ln \alpha_k^m - \frac{1}{2} \alpha_k^m \hat{w}_{dk}^m \right] + \left( a_0^\alpha - 1 \right) \ln \alpha_k^m - b_0^\alpha \alpha_k^m \right\} + C_{\alpha_k^m}
\]

\[
= \frac{1}{2} D_m \ln \alpha_k^m - \frac{\alpha_k^m}{2} \sum_{d=1}^{D_m} E_q [\hat{w}_{dk}^m] + \left( a_0^\alpha - 1 \right) \ln \alpha_k^m - b_0^\alpha \alpha_k^m + C_{\alpha_k^m}
\]

\[
= \left[ \left( a_0^\alpha + \frac{D_m}{2} - 1 \right) \ln \alpha_k^m - \frac{1}{2} \sum_{d=1}^{D_m} E_q [\hat{w}_{dk}^m] + b_0^\alpha \right] + C_{\alpha_k^m},
\]

which we recognize as Gamma(\( \hat{a}_{\alpha mk}^\alpha, \hat{b}_{\alpha mk}^\alpha \)), where

\[
\hat{a}_{\alpha mk}^\alpha = a_0^\alpha + \frac{D_m}{2}
\]

\[
\hat{b}_{\alpha mk}^\alpha = b_0^\alpha + \frac{1}{2} \sum_{d=1}^{D_m} E_q [\hat{w}_{dk}^m].
\]

**B.2.7 Noise Precision \((\tau)\)**

Variational distribution:

\[
\ln q_\theta(\tau_d^m) = E_q \left\{ \sum_{n=1}^{N} \left[ \frac{1}{2} \ln \tau_d^m - \frac{\tau_d^m}{2} \left( y_{nd}^m - \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk} \right)^2 \right] + \left( a_0^\tau - 1 \right) \ln \tau_d^m - b_0^\tau \tau_d^m \right\} + C_{\tau_d^m}
\]

\[
= \frac{N}{2} \ln \tau_d^m - \tau_d^m \sum_{n=1}^{N} E_q \left[ \left( y_{nd}^m - \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk} \right)^2 \right] + \left( a_0^\tau - 1 \right) \ln \tau_d^m - b_0^\tau \tau_d^m + C_{\tau_d^m}
\]

\[
= \left( a_0^\tau + \frac{N}{2} - 1 \right) \ln \tau_d^m - \tau_d^m \left\{ b_0^\tau + \frac{1}{2} \sum_{n=1}^{N} E_q \left[ \left( y_{nd}^m - \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk} \right)^2 \right] \right\} + C_{\tau_d^m},
\]
which we recognize to be a Gamma \( \Gamma(\hat{a}_{md}^\tau, \hat{b}_{md}^\tau) \) distribution, where

\[
\hat{a}_{md}^\tau = a_0^\tau + \frac{N}{2} \\
\hat{b}_{md}^\tau = b_0^\tau + \frac{1}{2} \sum_{n=1}^{N} E_q \left[ \left( y_{nd}^m - \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk} \right)^2 \right].
\]

### B.3 Evidence Lower Bound (ELBO)

The Evidence Lower Bound (ELBO) at iteration \( t \) is the right hand side of ELBO Equation in 4.3.2. Evaluated using the current \( q^{(t)}_{x,\theta}(x, \theta) \) it is

\[
E_q [\ln p(y \mid x, \theta)] + E_q [\ln p(x \mid \theta)] + E_q \left[ \ln p(S, W) \right] + E_q [\ln p(\theta)] + E_q [\ln p(\alpha)] + E_q [\ln p(\Psi)] + E_q [\ln p(\nu)] + E_q [\ln p(\pi)] - E_q [\ln q_x(x)] - E_q [\ln q_\theta(\theta)].
\]

Here, we derive the expected log likelihood and expected log prior distributions.

#### B.3.1 Expected likelihood

\[
E_q [\ln p(y \mid x, \theta)]
\]

\[
= E_q \left[ -\frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \left( y_{nd}^m - (S^m \odot W^m) z_{n}^T \right)^T (\Psi^m)^{-1} \left( y_{nd}^m - (S^m \odot W^m) z_{n}^T \right) + \ln \left( (2\pi)^{D_m} |\Psi^m| \right) \right]
\]

\[
= E_q \left[ -\frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{d=1}^{D_m} \left( y_{nd}^m - 2y_{nd}^m \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk} + \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk}^2 \right) + \ln \left( (2\pi)^{D_m} |\Psi^m| \right) \right]
\]

\[
= -\frac{1}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} E_q [\tau^m_d] \sum_{n=1}^{N} \left( y_{nd}^m - 2y_{nd}^m \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m \hat{w}_{dk}^m z_{nk} + \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m \hat{w}_{dk}^m z_{nk}^2 \right) + \sum_{k=1}^{K} E_q \left[ (s_{dk}^m \hat{w}_{dk}^m)^2 \right] E_q [\hat{w}_{dk}^m] E_q [\hat{w}_{dk}^m] + 2 \sum_{k_1 < k_2} E_q \left[ s_{dk_1}^m \hat{w}_{dk_1}^m \right] E_q \left[ s_{dk_2}^m \hat{w}_{dk_2}^m \right] E_q [z_{nk_1}] E_q [z_{nk_2}]
\]

\[
- \frac{N \ln(2\pi)}{2} \sum_{m=1}^{M} D_m + \frac{N}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} E_q [\ln \tau^m_d]
\]
B.3.2 Latent Variables \((V, Z)\)

The expected log priors:

\[
\mathbb{E}_q [\ln p(x | \theta)] = \mathbb{E}_{x, \theta} \left\{ \sum_{n=1}^{N} \sum_{l=1}^{L} v_{nl} \left[ \ln \pi_l - \frac{1}{2} (z_{nk}^T - \nu_l)^T (z_{nk}^T - \nu_l) - \frac{K}{2} \ln (2\pi) \right] \right\}
\]

\[
= \mathbb{E}_{x, \theta} \left\{ \sum_{n=1}^{N} \sum_{l=1}^{L} v_{nl} \left[ \ln \pi_l - \frac{1}{2} (z_{nk}^T z_{nk}^T - 2 z_{nk} \nu_l + \nu_l^T \nu_l) - \frac{K}{2} \ln (2\pi) \right] \right\}
\]

\[
= \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbb{E}_x [v_{nl}] \left[ \mathbb{E}_{\theta} [\ln \pi_l] - \frac{1}{2} \left( \sum_{k=1}^{K} \mathbb{E}_{z_{nk}} [z_{nk}^2] - 2 \sum_{k=1}^{K} \mathbb{E}_{z_{nk}} [z_{nk}] \mathbb{E}_{\theta} [\nu_{kl}] + \sum_{k=1}^{K} \mathbb{E}_{\theta} [\nu_{kl}^2] \right) - \frac{K}{2} \ln(2\pi) \right],
\]

The expected log of posterior distribution of \(x\) is:

\[
\mathbb{E}_q [\ln q(x | \theta)] = \mathbb{E}_q \left\{ \sum_{n=1}^{N} \sum_{k=1}^{K} - \frac{1}{2} \left[ \ln 2\pi \sigma_{z_{nk}}^2 + \frac{1}{\sigma_{z_{nk}}^2} (z_{nk} - \mu_{z_{nk}})^2 \right] + \sum_{n=1}^{N} \sum_{l=1}^{L} v_{nl} \ln \mu_{v_{nl}} \right\}
\]

\[
= \mathbb{E}_q \left\{ \sum_{n=1}^{N} \sum_{k=1}^{K} - \frac{1}{2} \left[ \ln 2\pi \sigma_{z_{nk}}^2 + \frac{1}{\sigma_{z_{nk}}^2} \left( \mu_{z_{nk}}^2 + \sigma_{z_{nk}}^2 + 2 \mu_{z_{nk}} \mu_{z_{nk}} \right) \right] + \sum_{n=1}^{N} \sum_{l=1}^{L} v_{nl} \ln \mu_{v_{nl}} \right\}
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} - \frac{1}{2} \left[ \ln 2\pi \sigma_{z_{nk}}^2 + \frac{1}{\sigma_{z_{nk}}^2} \left( \mathbb{E}_q (z_{nk}^2) + \mu_{z_{nk}}^2 - 2 \mathbb{E}_q (z_{nk}) \mu_{z_{nk}} \right) \right] + \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbb{E}_q (v_{nl}) \ln \mu_{v_{nl}}
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} - \frac{1}{2} \left[ \ln 2\pi \sigma_{z_{nk}}^2 + \frac{1}{\sigma_{z_{nk}}^2} (\mu_{z_{nk}}^2 + \sigma_{z_{nk}}^2 + \mu_{z_{nk}}^2 - 2 \mu_{z_{nk}} \mu_{z_{nk}}) \right] + \sum_{n=1}^{N} \sum_{l=1}^{L} \mu_{v_{nl}} \ln \mu_{v_{nl}}
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{2} \left[ \ln 2\pi \sigma_{z_{nk}}^2 + 1 \right] + \sum_{n=1}^{N} \sum_{l=1}^{L} \mu_{v_{nl}} \ln \mu_{v_{nl}}
\]

\[
= -\frac{NK}{2} (1 + \ln 2\pi) - \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{2} \ln \sigma_{z_{nk}}^2 + \sum_{n=1}^{N} \sum_{l=1}^{L} \mu_{v_{nl}} \ln \mu_{v_{nl}}
\]
B.3.3 Cluster Means \((\nu)\)

The expected log priors:

\[
\mathbb{E}_q [\ln p(\nu)] = \mathbb{E}_q \left\{ \sum_{k=1}^{K} \left[ -\frac{1}{2} \sum_{l=1}^{L-1} \nu_{kl}^2 - \frac{L-1}{2} \ln (2\pi) \right] \right\} \\
= -\frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{L-1} \mathbb{E}_q [\nu_{kl}^2] - \frac{K(L-1)}{2} \ln (2\pi),
\]

The expected log of posterior distribution of \(\nu\) is:

\[
\mathbb{E}_q [\ln q(\nu)] = \mathbb{E}_q \left\{ \sum_{k=1}^{K} \left[ -\frac{1}{2} \sum_{l=1}^{L-1} \sigma^2_{\nu_{kl}} (\nu_{kl} - \mu_{\nu_{kl}})^2 - \frac{1}{2} \sum_{l=1}^{L-1} \ln (2\pi \sigma^2_{\nu_{k}}) \right] \right\} \\
= \mathbb{E}_q \left\{ \sum_{k=1}^{K} \left[ -\frac{1}{2} \sum_{l=1}^{L-1} \mathbb{E}_q [\nu_{kl}^2] + \mu_{\nu_{kl}}^2 - \frac{1}{2} \ln (2\pi \sigma^2_{\nu_{k}}) \right] \right\} \\
= \sum_{k=1}^{K} \sum_{l=1}^{L-1} \left[ -\frac{1}{2} \mathbb{E}_q [\nu_{kl}^2] + \mu_{\nu_{kl}}^2 - \frac{1}{2} \ln (2\pi \sigma^2_{\nu_{k}}) \right] \\
= \sum_{k=1}^{K} \sum_{l=1}^{L-1} \left[ -\frac{1}{2} \mathbb{E}_q [\nu_{kl}^2] + \mu_{\nu_{kl}}^2 - \frac{1}{2} \ln (2\pi \sigma^2_{\nu_{k}}) \right] \\
= -\frac{K(L-1)}{2} \ln (2\pi) + \frac{K}{2} \sum_{l=1}^{L-1} \ln \sigma^2_{\nu_{k}}
\]

B.3.4 Mixture Proportions \((\pi)\)

The expected log priors:

\[
\mathbb{E}_q [\ln p(\pi)] = \mathbb{E}_q \left[ \sum_{l=1}^{L} (a_0^\pi - 1) \ln \pi_l + \ln \Gamma \left( \sum_{l=1}^{L} a_0^\pi \right) - \sum_{l=1}^{L} \ln \Gamma (a_0^\pi) \right] \\
= \sum_{l=1}^{L} (a_0^\pi - 1) \mathbb{E}_q [\ln \pi_l] + \ln \Gamma (La_0^\pi) - L \ln \Gamma (a_0^\pi).
\]

The expected log of posterior distribution of \(\pi\) is

\[
\mathbb{E}_q [\ln q(\pi)] = \mathbb{E}_q \left[ \sum_{l=1}^{L} (a_l^\pi - 1) \ln \pi_l + \ln \Gamma \left( \sum_{l=1}^{L} a_l^\pi \right) - \sum_{l=1}^{L} \ln \Gamma (a_l^\pi) \right] \\
= \sum_{l=1}^{L} (a_l^\pi - 1) \mathbb{E}_q [\ln \pi_l] + \ln \Gamma \left( \sum_{l=1}^{L} a_l^\pi \right) - \sum_{l=1}^{L} \ln \Gamma (a_l^\pi).
\]
B.3.5 Spike and Slab Weights ($\hat{W}$ and $S$)

The expected log priors:

\[
\mathbb{E}_q \left[ \ln p(S, \hat{W}) \right] = \mathbb{E}_q \left[ \sum_{m=1}^{M} \sum_{d=1}^{D_m} \sum_{k=1}^{K} \left\{ s^m_{dk} \ln \theta^m_k + (1 - s^m_{dk}) \ln (1 - \theta^m_k) - \frac{1}{2} \left[ \alpha_k^m \hat{w}^m_{dk} + \ln(2\pi/\alpha_k^m) \right] \right\} \right] \\
= \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{d=1}^{D_m} \left\{ (1 - \mathbb{E}_q[s^m_{dk}]) \mathbb{E}_q[\ln (1 - \theta^m_k)] + \mathbb{E}_q[s^m_{dk}] \mathbb{E}_q[\ln \theta^m_k] - \frac{1}{2} \mathbb{E}_q[\alpha_k^m] \mathbb{E}_q[\hat{w}^m_{dk}^2] \right\} \\
+ \frac{1}{2} \sum_{m=1}^{M} D_m \sum_{k=1}^{K} \mathbb{E}_q[\ln \alpha_k^m] - \frac{K}{2} \ln(2\pi) \sum_{m=1}^{M} D_m,
\]

The expected log of posterior distribution of $S, \hat{W}$ is:

\[
\mathbb{E}_q \left[ \ln q(S, \hat{W}) \right] = \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{d=1}^{D_m} \left\{ (1 - \mathbb{E}_q[s^m_{dk}]) \mathbb{E}_q[\ln (1 - \mathbb{E}_q[s^m_{dk}])] - \mathbb{E}_q[s^m_{dk}] \ln (\mathbb{E}_q[s^m_{dk}]) \right\} \\
- \frac{1}{2} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{d=1}^{D_m} \ln \left( \mathbb{E}_q[s^m_{dk}] \mathbb{E}_q \left[ \frac{\sigma_{w^m_{dk}}^2}{\mathbb{E}_q} \right] \right) + (1 - \mathbb{E}_q[s^m_{dk}]) / \mathbb{E}_q[\alpha_k^m] \\
+ \frac{1}{2} \sum_{m=1}^{M} D_m \sum_{k=1}^{K} \mathbb{E}_q[\ln \alpha_k^m] - \frac{K}{2} \ln(2\pi) \sum_{m=1}^{M} D_m,
\]

B.3.6 Sparsity Parameter ($\theta$)

The expected log priors:

\[
\mathbb{E}_q \left[ \ln p(\theta) \right] = \mathbb{E}_q \left[ \sum_{m=1}^{M} \sum_{k=1}^{K} (a^0_k - 1) \ln \theta^m_k + (b^0_k - 1) \ln (1 - \theta^m_k) + \ln \Gamma(a^0_k + b^0_k) - \ln \Gamma(a^0_k) - \ln \Gamma(b^0_k) \right] \\
= \sum_{m=1}^{M} \sum_{k=1}^{K} \left( a^0_k - 1 \mathbb{E}_q[\ln \theta^m_k] + (b^0_k - 1) \mathbb{E}_q[\ln (1 - \theta^m_k)] \right) + \ln \Gamma(a^0_k + b^0_k) - \ln \Gamma(a^0_k) - \ln \Gamma(b^0_k)
\]

The expected log of posterior distribution of $\theta$ is:

\[
\mathbb{E}_q \left[ \ln q(\theta) \right] = \mathbb{E}_q \left[ \sum_{m=1}^{M} \sum_{k=1}^{K} (\hat{a}^\theta_{mk} - 1) \ln \theta^m_k + (\hat{b}^\theta_{mk} - 1) \ln (1 - \theta^m_k) + \ln \Gamma(\hat{a}^\theta_{mk} + \hat{b}^\theta_{mk}) - \ln \Gamma(\hat{a}^\theta_{mk}) - \ln \Gamma(\hat{b}^\theta_{mk}) \right] \\
= \sum_{m=1}^{M} \sum_{k=1}^{K} \left( \hat{a}^\theta_{mk} - 1 \mathbb{E}_q[\ln \theta^m_k] + (\hat{b}^\theta_{mk} - 1) \mathbb{E}_q[\ln (1 - \theta^m_k)] \right) + \ln \Gamma(\hat{a}^\theta_{mk} + \hat{b}^\theta_{mk}) - \ln \Gamma(\hat{a}^\theta_{mk}) \\
- \ln \Gamma(\hat{b}^\theta_{mk})
\]
B.3.7 ARD Precision (\(\alpha\))

The expected log priors:

\[
\mathbb{E}_q[\ln p(\alpha)] = \mathbb{E}_q \left[ \sum_{m=1}^{M} \sum_{k=1}^{K} a_0^\alpha \ln b_0^\alpha + (a_0^\alpha - 1) \ln \alpha_k^m - b_0^\alpha \alpha_k^m - \ln \Gamma(a_0^\alpha) \right]
\]

\[
= \sum_{m=1}^{M} \sum_{k=1}^{K} \{(a_0^\alpha - 1) \mathbb{E}_q[\ln \alpha_k^m] - b_0^\alpha \mathbb{E}_q[\alpha_k^m]\} + MKa_0^\alpha \ln b_0^\alpha - MK \ln \Gamma(a_0^\alpha),
\]

The expected log of posterior distribution of \(\alpha\) is:

\[
\mathbb{E}_q[\ln q(\alpha)] = \mathbb{E}_q \left[ \sum_{m=1}^{M} \sum_{k=1}^{K} \hat{a}_{mk}^\alpha \ln \hat{b}_{mk}^\alpha + (\hat{a}_{mk}^\alpha - 1) \ln \alpha_k^m - \hat{b}_{mk}^\alpha \alpha_k^m - \ln \Gamma(\hat{a}_{mk}^\alpha) \right]
\]

\[
= \sum_{m=1}^{M} \sum_{k=1}^{K} \{(\hat{a}_{mk}^\alpha - 1) \mathbb{E}_q[\ln \alpha_k^m] - \hat{b}_{mk}^\alpha \mathbb{E}_q[\alpha_k^m]\} + MK\hat{a}_{mk}^\alpha \ln \hat{b}_{mk}^\alpha - MK \ln \Gamma(\hat{a}_{mk}^\alpha),
\]

B.3.8 Noise Precision (\(\tau\))

The expected log priors:

\[
\mathbb{E}_q[\ln p(\Psi)] = \mathbb{E}_q \left[ \sum_{m=1}^{M} \sum_{d=1}^{D_m} a_0^\tau \ln b_0^\tau + (a_0^\tau - 1) \ln \tau_d^m - b_0^\tau \tau_d^m - \ln \Gamma(a_0^\tau) \right]
\]

\[
= \sum_{m=1}^{M} \sum_{d=1}^{D_m} \{(a_0^\tau - 1) \mathbb{E}_q[\ln \tau_d^m] - b_0^\tau \mathbb{E}_q[\tau_d^m]\} + \sum_{m=1}^{M} D_m \left[ a_0^\tau \ln b_0^\tau - \ln \Gamma(a_0^\tau) \right],
\]

The expected log of posterior distribution of \(\Psi\) is:

\[
\mathbb{E}_q[\ln q(\Psi)] = \mathbb{E}_q \left[ \sum_{m=1}^{M} \sum_{d=1}^{D_m} \hat{a}_{md}^\tau \ln \hat{b}_{md}^\tau + (\hat{a}_{md}^\tau - 1) \ln \tau_d^m - \hat{b}_{md}^\tau \tau_d^m - \ln \Gamma(\hat{a}_{md}^\tau) \right]
\]

\[
= \sum_{m=1}^{M} \sum_{d=1}^{D_m} \{(\hat{a}_{md}^\tau - 1) \mathbb{E}_q[\ln \tau_d^m] - \hat{b}_{md}^\tau \mathbb{E}_q[\tau_d^m]\} + \sum_{m=1}^{M} D_m \left[ \hat{a}_{md}^\tau \ln \hat{b}_{md}^\tau - \ln \Gamma(\hat{a}_{md}^\tau) \right]
\]
APPENDIX C. EARLY ONSET PREECLAMPSIA IN A MODEL FOR HUMAN PLACENTAL TROPHOBLAST

Modified from a manuscript published in *Proceedings of the National Academy of Sciences*

Megan A. Sheridan\textsuperscript{a,h,c}, Ying Yang\textsuperscript{b,d,e}, Ashish Jain\textsuperscript{f,g}, Alex S. Lyons\textsuperscript{a,b}, Penghua Yang\textsuperscript{b,h}, Sambasiva R. Brahmasani\textsuperscript{b,i}, Aihua Dai\textsuperscript{b}, Yuchen Tian\textsuperscript{b}, Mark R. Ellersieck\textsuperscript{j}, Geetu Tuteja\textsuperscript{f,g}, Danny J. Schust\textsuperscript{k}, Laura C. Schulz\textsuperscript{k}, Toshihiko Ezashi\textsuperscript{b,d}, and R. Michael Roberts\textsuperscript{a,b,d}

\textsuperscript{a}Department of Biochemistry, University of Missouri, Columbia, MO 65211
\textsuperscript{b}Bond Life Sciences Center, University of Missouri, Columbia, MO 65211
\textsuperscript{c}Department of Pathology, Centre for Trophoblast Research, University of Cambridge, Cambridge, United Kingdom
\textsuperscript{d}Division of Animal Sciences, University of Missouri, Columbia, MO 65211
\textsuperscript{e}Department of Molecular Pharmacology and Physiology, University of South Florida, Tampa, FL 33620
\textsuperscript{f}Bioinformatics and Computational Biology, Iowa State University, Ames, 50010, USA
\textsuperscript{g}Genetics, Development, and Cell Biology, Iowa State University, Ames, 50010, USA
\textsuperscript{h}Department of Obstetrics, Gynecology, and Reproductive Sciences, School of Medicine, University of Maryland, Baltimore, Baltimore, MD 21201
\textsuperscript{i}Centre for Cellular and Molecular Biology (CCMB), Hyderabad 500007, India
\textsuperscript{j}Agriculture Experiment Station, University of Missouri, Columbia, MO 65211
\textsuperscript{k}Department of Obstetrics, Gynecology and Women’s Health, University of Missouri, Columbia, MO 65212
C.1 Abstract

We describe a model for early onset preeclampsia (EOPE) that uses induced pluripotent stem cells (iPSC) generated from umbilical cords of EOPE and control (CTL) pregnancies. These iPSC were then converted to placental trophoblast (TB) representative of early pregnancy. Marker gene analysis indicated that both sets of cell differentiated at comparable rates. The cells were tested for parameters disturbed in EOPE, including invasive potential. Under 5% \( O_2 \), CTL-TB and EOPE-TB lines did not differ, but, under hyperoxia (20% \( O_2 \)), invasiveness of EOPE-TB was reduced. RNAseq analysis disclosed no consistent differences in expression of individual genes between EOPE-TB and CTL-TB under 20% \( O_2 \), but, a weighted correlation network analysis revealed two gene modules (CTL4 and CTL9) that in CTL-TB were significantly linked to extent of TB invasion. CTL9, which was positively correlated with 20% \( O_2 \) (\( P = 0.02 \)) and negatively correlated with invasion (\( P = 0.03 \)), was enriched for gene ontology terms relating to cell adhesion and migration, angiogenesis, preeclampsia, and stress. Two EOPE-TB modules, EOPE1 and EOPE2, also correlated positively and negatively, respectively, with 20% \( O_2 \) conditions, but only weakly with invasion; they largely contained the same sets of genes present in modules CTL4 and CTL9. Our experiments suggest that in EOPE the initial step precipitating disease is a reduced capacity of placental TB to invade caused by a dysregulation of \( O_2 \) response mechanisms and that EOPE is a syndrome, in which unbalanced expression of various combinations of genes affecting TB invasion provoke disease onset.

C.2 Introduction

Preeclampsia, whether of the early or later onset form (Lisonkova and Joseph, 2013), is characterized by gestational hypertension and proteinuria, with onset of symptoms in the second half of pregnancy (Huppertz, 2008)(Raymond and Peterson, 2011)(Redman and Sargent, 2005). Overall, it affects approximately 4% of pregnancies. However, the more severe, early onset form of preeclampsia (EOPE), which can be diagnosed as early as 20 weeks of pregnancy, occurs in approximately 0.4% of pregnancies in the U.S.A. and more often leads to fetal growth restriction (Brosens et al., 2011)(Lisonkova and Joseph, 2013)(Paruk and Moodley, 2000)(von Dadelszen et al., 2011). It has
been estimated to lead to 3 million preterm births every year, and is a leading cause of maternal and perinatal morbidity and mortality worldwide (Amaral et al., 2015)(George and Granger, 2010)(Jeyabalan, 2013)(Pennington et al., 2012). Deemed the “disease of theories”, the origins of either form of preeclampsia remain enigmatic as the causes are likely multifactorial, with multiple proposed risk factors and complex inheritance patterns (Oudejans et al., 2007)(Roberts and Cooper, 2001). Removing the placenta is the only known cure for preeclampsia, suggesting that factors released by TB acting on a susceptible mother are responsible for disease symptoms (Sibai et al., 2005). EOPE, in particular, has been attributed to deficient remodeling of the uterine spiral arteries by the invasive extravillous trophoblast (EVTB) (Pijnenborg et al., 2011), which begins about mid-way through the first trimester of pregnancy before disease symptoms are evident (Brosens et al., 2002)(Khong et al., 2015). In turn, the unmodified arteries cause erratic perfusion of the placenta as it matures (Bilodeau, 2014)(Burton et al., 2009b)(Jauniaux et al., 2006)(Jauniaux et al., 2000). Ischemia-reperfusion then leads to oxidative stress (Bilodeau, 2014)(Burton et al., 2009b)(Jauniaux et al., 2006)(Jauniaux et al., 2000). It has also been proposed that EOPE-TB has an inherently impaired response to oxidative stress (Caniggia et al., 2000)(Rajakumar and Conrad, 2000)(Rolfo et al., 2010), causing general TB dysfunction and an increased release (by placental TB cells) of antiangiogenic factors that provoke endothelial dysfunction and inflammation in the maternal vessels (Roberts and Gammill, 2005)(Saito and Nakashima, 2014). In a normal pregnancy, upregulation of vascular endothelial growth factor (VEGF) and placental growth factor (PGF) are important for proper angiogenesis and vasodilation (Burton et al., 2009a)(Torry et al., 2003), while in EOPE, in particular, PGF is released in reduced quantities (Tidwell et al., 2001) and an antagonist of VEGF, known as placenta-derived soluble FMS-like tyrosine kinase-1 (sFLT1), is typically upregulated (Amaral et al., 2015)(Sanchez-Aranguren et al., 2014)(Tripathi et al., 2008).

Studying the etiology of all forms of preeclampsia, including EOPE has been hampered by its delayed clinical presentation and its restriction to humans, with only rare occurrences in non-human primates (Gille et al., 1977)(Hennessy et al., 1997)(Palmer et al., 1979). While rodent models have demonstrated features of EOPE, none encompass the full range of symptoms and nearly all lack
the expected disease progression to eclampsia (Pennington et al., 2012). In vitro models that use primary cells from a mature human placenta are probably inadequate for several reasons. The insults leading to EOPE almost certainly arise early in the first trimester when EVTB is colonizing the endometrium and before onset of extensive maternal blood perfusion, whereas term placentae lack an invasive component. Additionally, term placentae from preeclampsia pregnancies show signs of dysfunction and structural damage resulting from the disease, which are difficult to distinguish from underlying causes (Benirschke et al., 2006)(Kaufmann and Castellucci, 1997). While it is possible to obtain primary cells from the first trimester of pregnancy, PE cannot yet be diagnosed at this stage.

An alternative to animal models or primary tissues derived from placentae, we employ “a disease in a culture dish approach” to study TB development and conditions such as EOPE. Our laboratory has developed a model system for studying TB in which pluripotent stem cells are exposed to bone morphogenetic protein 4 (BMP4) in combination with signaling inhibitors of ACTIVIN-A (A83-01) and FGF2 (PD173074) (BAP treatment) (Amita et al., 2013)(Yabe et al., 2016). These BAP-derived TB are thought to represent the highly invasive cells of the primitive placenta, and therefore provide an advantageous model to study defects of TB invasion (Jain et al., 2017)(Roberts et al., 2018)(Telugu et al., 2013)(Yabe et al., 2016). To capture potential genetic or epigenetic features that might characterize EOPE, fibroblast cells were cultured from explants from umbilical cords (UC) of babies born to mothers who had experienced EOPE during their pregnancies as well as homologous cells from umbilical cords of infants born to mothers following a normal pregnancy to act as controls (CTL) (Yang et al., 2014). It was noted that establishing cultures under 20% $O_2$ conditions, particularly with high glucose in the medium, proved significantly more challenging from EOPE than CTL umbilical cord explants. Moreover the EOPE mesenchyme (fibroblast) cultures, even when established under low (4-5%) $O_2$ conditions, were more susceptible to oxidative stress than the CTLs, suggesting that there were genetic/epigenetic differences that distinguished the two and that some of the features of EOPE, though subtle, might be conserved in the umbilical cord fibroblasts. Presumably these cord cells had a poorer ability than CTL
counterparts to respond appropriately to oxidative stress, which many consider to be a major contributor to the pathophysiology of EOPE (Burton et al., 2009b)(Yung et al., 2014). For the present study, both the EOPE and CTL fibroblast cultures generated from umbilical cords were reprogrammed to create induced pluripotent stem cell (iPSC) lines, with the goal of converting the latter to early stage placental TB by the BAP protocol described above. In this manner, we sought to determine whether hypersensitivity to high $O_2$ remained evident in the TB derived from EOPE pregnancies and whether such cells demonstrated other disparities relative to CTL-TB, especially in gene expression, that might provide clues to the underlying causes of the disease. The goal of this reprogramming approach was, therefore, to attempt to recapitulate events as they occur in the early stages of TB development when EOPE is believed to arise.

\section*{C.3 Methods}

\subsection*{C.3.1 Primary tissue and iPSC generation}

Fresh human umbilical cord (UC) tissues were collected in an aseptic manner at the Women’s and Children’s Hospital (University of Missouri, Columbia, MO). A total of 10 controls (CTL) and 19 infants whose mothers suffered early onset PE (EOPE) were assessed in these experiments (Table C.1). The UC tissue collection (project #1201132 and #1209459) has been described elsewhere (Yang et al., 2014) and approved by the University of Missouri Health Sciences Institutional Review Board. Briefly, the tissues were washed to remove blood cells and minced into fragments to deliver adherent cells by the explant method (Ishige et al., 2009). Outgrowths appeared at the periphery of the minced tissues after one week of culture. After 10-11 days, the fibroblasts were passaged from the 48-well plate into T25 flasks by using TrypLE (Invitrogen). The cells reached confluence in the flask by 14 days and were expanded for reprogramming to iPSC. All cultures used to provide stocks of primary cells and iPSC were performed under a 5% $O_2$/5% $CO_2$/80% N2 gas atmosphere.

Generation of iPSCs from umbilical cord fibroblasts with episomal vectors has also been described elsewhere (Lee et al., 2014). In brief, a protocol developed by Okita et al (Okita et al.,
2011) with episomal vectors carrying shRNA for p53 suppression and non-transforming L-MYC, in addition to the usual reprogramming genes POU5F1, SOX2, KLF4 and LIN-28, was employed to reprogram the fibroblasts. Three micrograms of the developed Y4 combination (POU5F1, SOX2, KLF4, LIN28, L-MYC and p53 shRNA) of the episomal plasmids was electroporated with a Nucleofector II device (Lonza, Basel, Switzerland) and Amaxa NHDF Nucleofector kit (Lonza) into the CTL and EOPE fibroblast cells collected from umbilical cords according to the manufacturer’s instructions. After 2-4 days of recovery, cells were placed into 100 mm dishes previously coated with Matrigel (BD Bioscience, San Jose, CA). The following day the culture medium was switched to mTeSR1 (StemCell Technologies, Vancouver, Canada). Colonies resembling human ESC were mechanically isolated (around day 20 of culture) and expanded. A total of 10 CTL iPSC and 19 EOPE iPSC lines were established.

RNA was extracted from each of the primary fibroblast cultures and the respectively generated iPSC lines (cultured under both O2 concentrations, 4% and 20%) in STAT60 (Tel-Test) and further purified according to the manufacturer’s instructions. RNA samples were submitted to the University of Texas Southwestern Medical Center Microarray Core Facility (https://microarray.swmed.edu/), and microarray analysis was performed with Illumina HumanHT-12 v4 expression BeadChips. This analysis used the entire 47,159 probes on the BeadChip. Raw intensity data were background-subtracted using BeadStudio software to generate transcriptomes for each of the cell lines (GeneSpring 12.6 software; Agilent Technologies, Inc.). A raw intensity was then assigned to each gene by taking the average intensity of all probes corresponding to a gene. The raw intensities were normalized by using the quantile normalization method. Principal component analysis was then applied to the top quartile of protein coding genes (4,186 genes) based on the variance across samples in order to provide an unsupervised clustering and visualization approach to the data. Transcriptome data from the gene expression arrays were also used to evaluate pluripotency of the hESC and iPSC lines using the Pluritest (Muller et al., 2011) (www.pluritest.org/).
C.3.2 Authentication of the cell lines

Cell line authentication was conducted by short tandem repeat (STR) profiling to reveal misidentified cell lines and lines that had become contaminated by cells from other cultures (American Type Culture Collection Standards Development Organization Workgroup et al., 2010)(Yu et al., 2015). Genomic DNA samples from the different iPSC/hESC cell lines and also from the progenitor umbilical cord fibroblasts and tissue were isolated with the Wizard genomic DNA purification system (Promega, A2360). The purified DNA was subjected to STR analysis at DNAcore (University of Missouri) or at IDEXX BioResearch (Columbia, MO) by using AuthentiFiler™ PCR Amplification Kit (ThermoFisher) or Promega CELL ID™ System, respectively. The STR profiles for the H1 (WA01) and H9 (WA09) ESC lines were compared with DNA fingerprinting data provided by Wi-Cell (www.wicell.org) for those cells and, for iPSC cells, with the profiles obtained for the original progenitor umbilical cord materials, respectively (Table S2). Tests for Mycoplasma contamination were performed approximately every three months for active cultures by MycoAlert™ Mycoplasma Detection Kit (LONZA). All results obtained in the present paper were from Mycoplasma-free cultures.

C.3.3 Cell culture and differentiation

Human ESC (H1/WA01 and H9/WA09) and iPSC were cultured in six-well tissue culture plates (Thermo Scientific) coated with Matrigel (BD Bioscience) under an atmosphere of 5% CO₂/air at 37°C in mTeSR1 medium (STEMCELL Technologies). Cells were passaged every 5–6 days. Briefly, colonies were exposed to dispase (1 mg/mL) (STEMCELL Technologies) for 5-7 min at 37°C and then were broken into small clumps with the Stempro EZpassage (Invitrogen) cutting tool. The method for TB differentiation has been described elsewhere with slight modifications (Amita et al., 2013). Briefly, cells were passaged onto Matrigel coated dishes into DME/F12 medium (Thermo Scientific) with knock-out serum replacement (KOSR, Invitrogen) that had been conditioned by mouse embryonic fibroblasts for 24 h (MEF-CM) and supplemented with FGF2 (4 ng/ml). After 24 h, the conditioned medium was replaced with daily changes of non-conditioned DME/F12/KOSR.
medium lacking FGF2 (hESC medium), but containing BMP4 (10 ng/ml), A83-01 (1 μM) and PD173074 (0.1 μM) (BAP treatment) for up to 6 days. Control cultures (undifferentiated) were maintained in MEF-CM containing 4 ng/ml FGF2. After 6 days of BAP treatment, spent medium was collected for immunoassays, total DNA was isolated to determine cell density (Promega), and total RNA (STAT60, Tel-Test Inc.) isolated for RNA sequencing.

C.3.4 Invasion Assay

ESC and iPSC cultures were mechanically dispersed by exposing colonies to dispase (1 mg/ml) for 7 min at 37°C and then broken into small squares of approximately the same size (100 cells) by using the Stempro EZpassage cutting tool. A small aliquot of the cell suspension was mechanically dispersed to single cells via pipetting and total cell number was counted by using a TC20 automated cell counter (BioRad). Cells were seeded at a density of 5x10⁴ cells/well on BD Matrigel-coated membranes with 8.0 μm pores within commercial invasion chambers in MEF-CM + FGF2 (4 ng/ml) under either 5% or 20% O₂ conditions. The next day medium was changed to hESC + BAP to induce TB differentiation, while controls remained in MEF-CM + FGF2. Medium was changed daily, and after 6 days of BAP treatment cells were fixed in 4% (w/v) paraformaldehyde. The tops of the wells were imaged to document overall cell density and morphology prior to analysis. The cells remaining on the top of the membrane were wiped off with a cotton swab and the bottom of the membrane stained with DAPI (4’,6-diamidino-2-phenylindole) and mounted. The number of invaded cells was quantified by taking 12 random field images of the membrane under 10 x magnification. The number of nuclei in each field was quantified by using ImageJ software (https://imagej.nih.gov/ij/docs/guide/user-guide). Three or four independent experiments each with three internal experimental replicates were run for each cell line under both O₂ conditions (Table S4).
C.3.5 Immunoassays

After 6 days of BAP treatment, spent culture medium was collected to measure hormone and growth factor levels. Total DNA was isolated (Promega, A2360) from the respective cultures in order to normalize immunoassay results to cell density levels. ELISAs were performed by following the protocols recommended by the manufacturers. The subsequent immunoassays were used: hCG (human chorionic gonadotrophin) (Genway bio, GWB-BQK0F2), PGF (human placental growth factor; ThermoFisher, EHPGF), P4 (progesterone; Genway bio, GWB-BQK0FC), and human VEGF-R1 (sFLT1; ThermoFisher, BMS268-3). Samples were collected in three independent experiments for each cell line under both $O_2$ conditions.

C.3.6 Illumina TruSeq RNA Library Preparation and Sequencing

High-throughput sequencing was performed at the University of Missouri DNA Core Facility. Libraries were constructed following the manufacturer’s protocol with reagents supplied in Illumina’s TruSeq mRNA stranded sample preparation kit. Briefly, the poly-A containing mRNA is purified from total RNA, the RNA was fragmented, double-stranded cDNA generated from fragmented RNA, and the index containing adapters ligated to the ends. The amplified cDNA constructs were purified by addition of Axyprep Mag PCR Clean-up beads (Fischer Scientific). The final construct of each purified library was evaluated by using the Fragment Analyzer automated electrophoresis system, quantified with the Qubit fluorometer in conjunction with the Qubit HS dsDNA assay kit, and diluted according to Illumina’s standard sequencing protocol for sequencing on the NextSeq 500. Raw read counts for each of the 48 DNA samples analyzed are listed in Table S3. The reads were aligned to the reference human genome (hg19) by using the program HISAT2 (https://ccb.jhu.edu/software/hisat2) (Kim et al., 2015) an overall alignment of >97% was achieved in all samples (Table S3). The raw read counts for each biological replicate were normalized by gene length and then by the sequencing depth to calculate the transcripts per million (TPM values) (Li and Dewey, 2011). The TPM values were then log-transformed for further analysis.
C.3.7 Differential expression analysis

EdgeR was used for differential expression analysis (McCarthy et al., 2012)(Robinson et al., 2010). The reads were normalized by using the trimmed mean of M values (TMM) normalization method (Robinson and Oshlack, 2010). The differentially expressed genes (DEG) were defined as the genes with an absolute fold-change $\geq 1.5$ and adjusted p-value $\leq 0.05$.

C.3.8 Functional enrichment analysis

Functional enrichment analysis of gene sets was carried out by using the Genomic Regions Enrichment of Annotations Tool (GREAT) (Robinson and Oshlack, 2010). The GO Biological Process and Disease Ontology were used for enrichment analysis using the hypergeometric statistic. The ontology terms with an FDR $\leq 0.05$ and a fold-change $\geq 2$ were considered significant.

C.3.9 Weighted co-expression network analysis

Weighted correlation network analysis (WGCNA) (Langfelder and Horvath, 2008) was used to construct a co-expression network from the gene expression data from CTL and EOPE samples. The data from the invasion assays were used to calculate the correlation between co-expressed gene modules and invasion cell counts. The oxygen treatment, disease group information, and finally sex were also incorporated into the evaluation. Analysis was carried out as described previously (Jain et al., 2017) with minor modifications. Genes that had an expression level lower than 1 across all samples were removed from the analyses. Then, the top quartile of genes based on variance were selected. After that, the adjacency matrix for the selected genes was calculated by means of the signed hybrid Pearson correlation method with a soft thresholding power. The power was used to make a scale-free network ($R^2 = 0.8$) that retains a good number of connections. The adjacency matrix was then used to calculate the interconnectedness score (topological overlap) that was used to bundle the genes by hierarchical clustering. The hierarchical tree was further cut into gene modules by using the dynamic tree cut tool. Closely related gene clusters (Correlation $> 0.75$) were merged together to obtain the final gene modules (Langfelder et al., 2008). The identified
gene modules were decomposed such that each module was represented by its weighted expression (module eigengene) in the form of its first principal component. The correlation between the gene expression and the module eigengene defined as module membership (KME) was also calculated for each gene. Based on the KME, the genes were assigned to modules with the correlation value greater than 0.75, thereby allowing a gene to be a part of multiple modules or multiple regulatory pathways. The significance of the association between the module and different features, including invasion cell count and oxygen treatment, was also estimated by calculating the correlation between the module eigengene and the corresponding feature.

C.4 Results

C.4.1 Phenotypes of the Primary Fibroblast and Induced Pluripotent Stem Cells.

The primary fibroblast lines listed in Table C.1 were generated from CTL and EOPE umbilical cords and subsequently reprogrammed by introducing episomal plasmids expressing POU5F1, SOX2, KLF4, MYCL, and LIN28 (Okita et al., 2011). A total of 29 iPSC lines (10 CTL, MRuc1-10; 19 EOPE, MRucA-S) were successfully produced (Table C.1). Each displayed a morphology comparable to epiblast-type human embryonic stem cells (hESC). Cell line authentication by short tandem repeat (STR) profiling (Table S1) verified the identity of the two ESC lines H1 and H9 by comparison with STR data available from WiCell. Each iPSC line exhibited a STR profile identical to that obtained from the DNA of the original progenitor cells. No cross-contamination with other human or animal cells was detected. Each of the iPSC lines possessed a transcriptome characteristic of a pluripotent cell as signified by the bioinformatics tool Pluritest (Muller et al., 2011)(Table S2). Three representative iPSC lines (MRuc7, -8, and -O) as well as H1 and H9 ESC lines have previously been evaluated by their ability to form teratomas in immune-deficient animals (Lee et al., 2014)(Yang et al., 2015).

The population doubling times of selected iPSC lines (both CTL four lines (MRuc4, -7, -8, -10) and EOPE six lines (MRucB, -E, -I, -J, -K, -N) were measured under both 5 % and 20 % $O_2$ conditions of culture. No differences across groups were apparent, although there was clearly
sufficient variance in the data to reveal that growth rates of individual cell lines varied significantly (Fig. C.1A). This experiment provided a control to ensure that high $O_2$ did not greatly influence mean population doubling times of CTL versus EOPE lines as a whole and hence initial cell numbers prior to inducing differentiation to TB. Further, a principle component analysis (PCA) based on microarray data from 29 iPSC lines (10 CTL, 19 EOPE), also could not readily distinguish the CTL versus EOPE iPSC either relative to disease status (CTL versus EOPE) (Table C.1) or to the responses of the cell lines to culture under 20% $O_2$ (Fig. C.1B, right). This mingling of transcriptome data in response to 20% $O_2$ was in stark contrast to the primary fibroblast cells from which the iPSC had been generated, where there was a clear separation according to the $O_2$ atmosphere under which the cell lines had been cultured (Fig. C.1B, left). Together, these experiments suggest that the $O_2$ susceptibility of the primary fibroblast cultures, which were used as the progenitors of the iPSC, was erased as a result of reprogramming (Yang et al., 2014). Finally, there was no indication that male (triangles) and female (circles) fibroblast or iPSC lines were distinguishable as assessed by PCA of global gene expression under either $O_2$ condition (Fig. C.2).

C.4.2 Invasive features of iPSC-derived TB under low and high $O_2$

The $O_2$ sensitivity observed in primary cultures from EOPE patients (Yang et al., 2014), was lost when the fibroblasts were reprogrammed to iPSC (Fig. C.2) prompting the question as to whether measurable differences existed between CTL- and EOPE-iPSC lines after reprogramming. One of the hallmarks of EOPE is reduced EVTBB invasion. Therefore, the ability of TB derived from CTL and EOPE iPSC lines to invade through Matrigel was assessed under both 5% $O_2$, mimicking concentrations in the female reproductive tract, and oxidative stress conditions (20% $O_2$) (Cartwright et al., 2010)(Jauniaux et al., 2003)(Saito and Nakashima, 2014)(Soares et al., 2017). To address this question, CTL and EOPE derived iPSC cell lines, which had been generated and expanded under 5% $O_2$ conditions before cryopreservation, were cultured under non-differentiating conditions in both 5% and 20% $O_2$ for at least three passages in order to acclimate them to the respective gas atmospheres and to minimize acute responses to initial $O_2$ shock. Next, these CTL-
and EOPE- iPSC were plated on Matrigel invasion chambers and exposed to BAP conditions for 6 days to induce TB differentiation under the contrasting $O_2$ conditions.

In total, invasion assays employed two hESC lines (H1 and H9), 8 CTL-TB and 14 EOPE-TB (Fig. C.2A). Some cell lines were omitted from the invasion experiments because they did not readily make the transition from the mTeSR1 culture medium to the medium employed in the differentiation protocol (DME/F12 with knock out serum replacement). Others failed to attach firmly to the Matrigel-coated membranes in the invasion chambers. Each of the cell lines that could be differentiated with ease were tested on at least three separate occasions, and on each occasion the invasion assays were run in triplicate or quadruple (Table S3). Before statistical analysis was performed, the data were log transformed to account for the unequal variance among the cell lines. While the magnitude of invasion varied amongst cell lines, reproducibility was generally consistent across experimental replicates (Table S4). As illustrated in Fig. C.2A and B, under 5% $O_2$, invasiveness of CTL-TB and EOPE-TB lines was not different ($P = 0.668$); 2) Under 20% $O_2$, invasiveness of EOPE-TB was lower than CTL-TB ($P = 0.024$); 3) Invasiveness of CTL-TB was not influenced by 20% $O_2$ ($P = 0.513$); 4) Invasiveness of EOPE-TB as a group was significantly inhibited by 20% $O_2$ ($P = 0.008$). No sex differences were noted. In pairwise comparisons (Fig. C.2A), all but three of the 14 EOPE-TB lines displayed a significant reduction in invasion under 20% $O_2$ conditions when compared to 5% conditions. Of those three, line D and K (MRucDi and MRucKi) approached significance ($P = 0.053$ and $P = 0.051$, respectively), while line Q (MRucQi) lacked any indication of the invasion phenotype seen in the other EOPE-TB lines. One CTL-TB line (9; MRuc9i), showed reduced invasion under 20% $O_2$ conditions, which approached significance ($P = 0.08$) (Fig C.2A, Table S4). Of note, this CTL displayed reduced growth in the invasion chambers under conditions designed to maintain pluripotency prior to initiating differentiation and attached less well to the Matrigel substratum than most other iPSC lines. This unpredictability led to uneven results in the invasion assays and suggests that the decreased invasion of BAP-differentiated line 9 might have been an anomaly and possibly unrelated to $O_2$ responsiveness. Taken together, however, these results indicated a remarkably
consistent phenotypic difference in TB invasiveness between CTL and EOPE cell lines when the assays were performed under 20% \( O_2 \). Additionally it was clear that sensitivity to \( O_2 \) displayed by the progenitor EOPE fibroblasts (Yang et al., 2014) was at least partially regained when the iPSC were differentiated into TB. Of note, CTL lines 7 and 8 (MRuc7i and MRuc8i) were generated from infants delivered prematurely at 34 and 33 weeks, respectively, and the mother of the infant who provided line 7 had received betamethasone, similar to the delivery conditions of the EOPE cell donors. The responses of these cell lines to 20% \( O_2 \) was typical of the rest of the CTL group, namely invasion was not impaired (Fig C.2A, Table S4).

C.4.3 Effect of low and high \( O_2 \) on relative growth rates of differentiating CTL- and EOPE-iPSC.

To assure that the reduced number of invading TB cells displayed by EOPE-TB under 20% \( O_2 \) conditions was not an indirect effect of reduced proliferation, the total DNA concentration of CTL and EOPE lines cultured under low and high \( O_2 \) conditions was measured at d 6 of differentiation, the equivalent time point used for the invasion assays. DNA concentration was selected as a gauge of cell proliferation because counting cells at d 6 is complicated by fusion of TB into multinucleated syncytia (Amita et al., 2013)(Yabe et al., 2016). Two ESC lines (H1 and H9), four CTL lines (MRuc5, -6, -7, and -8) and four EOPE lines (MRucA, -J, -M and -R) were selected. Except to ensure that there was equal sex representation within groups, the selection of lines was made arbitrarily. The cell colonies were differentiated to TB by 6 days BAP treatment in an identical manner to that described in the invasion assays except that the cells were cultured on 6-well dishes rather than in the invasion chambers. In pairwise comparisons, no cell line exhibited a significant difference in total DNA amount when grown under the two contrasting \( O_2 \) conditions (Fig. C.2C). However, the DNA content, and hence population doubling time, was significantly affected by cell line (\( P < 0.0001 \), two-way ANOVA) (Fig C.2C). As cell lines also differ in the number of invading cells, we sought to determine whether this feature was correlated with the DNA content of the cultures. No such correlation was observed (Pearson \( r^2 = 0.16 \), P-Value = 0.50) (Fig. C.2D). Overall, these
results indicate that the differences in invaded cell numbers of the CTL-TB and EOPE-TB at 20% $O_2$ were not explained by the overall proliferation rates of individual iPSC lines under BAP differentiation conditions.

C.4.4 Production of hCG and progesterone by CTL- and EOPE-TB under 5% and 20% $O_2$ conditions.

A feature of BAP treated iPSC/ESC is their ability to secrete the placental hormones hCG and progesterone (P4) (Amita et al., 2013). These hormones are also secreted in greater amounts under 20% $O_2$ conditions than under the lower % conditions (Das et al., 2007). Additionally, serum concentrations of hCG (Heikkila et al., 2001)(Wenstrom et al., 1994) and P4 (Salas et al., 2006)(Walsh, 1988) have been reported to be higher in preeclampsia patients relative to controls as pregnancy progresses. To assess whether hCG and P4 production differed between the two ESC lines (H1 and H9), the same selected CTL lines (MRuc5, -6, -7, and -8) and the EOPE lines (MRucA, -J, -M and -R), media samples from the 24 h of culture between d 5 and d 6 were collected, and the concentrations of hCG and P4 produced were determined by ELISA (Fig. C.3A and B). To account for any proliferation differences between lines, these values were normalized to the DNA concentrations of the respective cultures. The CTL-TB and EOPE-TB displayed no significant differences in hCG production. As anticipated (Amita et al., 2013), 20% $O_2$ significantly increased hCG and P4 production relative to 5% conditions (P = 0.02 and P-Value = 0.008, respectively). Additionally, sex affected hCG production, with females releasing more than males under 20% $O_2$ conditions (P = 0.007), which is consistent with what had been reported previously for H1 and H9 cells, which are male and female respectively (Amita et al., 2013). Of note, after only 6 days of BAP treatment, the sex effect was not demonstrated by the two ESC lines (H1 and H9). This is consistent with overall hormone production when ESC lines are lower than iPSC lines at this time point, therefore a sex effect is only noticeable between ESC lines after 7 or 8 days of BAP treatment (Fig. C.3A) (Amita et al., 2013). Female lines on average demonstrated higher production of P4 than males (P-value = 0.02), but this inference was complicated by a significant interaction between
sex and disease status (P-value = 0.02), such that the sex difference was more pronounced among EOPE lines. More studies based on additional cell lines will be needed to determine whether these relationships of P4 production to sex and disease are authentic. Collectively, the hCG and P4 data indicate that EOPE iPSC display no obvious defect in BAP-directed TB differentiation relative to CTLs, as they produce placental hormones at a comparable magnitude and respond to $O_2$ concentrations similarly.

C.4.5 Production of angiogenic factors PGF and FLT1 by CTL- and EOPE- TB under 5% and 20% $O_2$ conditions.

As EOPE is commonly associated with a dysregulation of angiogenic factors (Tidwell et al., 2001)(Torry et al., 2003)(Whyte et al., 2007) and specifically an increased ratio of sFLT1 to PGF (O’Brien et al., 2017), the production of PGF and the soluble form of FLT1 (sometimes known as VEGFR-1) was determined in the same cell lines as used for hCG and P4 assays above (Fig. C.3C and D). As before, these values were normalized to the total DNA concentration of the respective cultures. While as a group, EOPE-TB and CTL-TB cultures did not significantly differ in their overall PGF production, PGF was slightly increased in CTL-TB in response to high $O_2$ (P = 0.05), while EOPE-TB lacked such a response to $O_2$ (P-Value = 0.55) (Fig. C.5C). Individual pairwise comparisons indicated that three out of the four CTL-TB and TB from both ESC lines upregulated PGF in response to high (20%) $O_2$ conditions, while PGF production was not responsive to high $O_2$ in any of the EOPE cultures (P = 0.13). In contrast to PGF, there were no differences between the EOPE and CTL cultures in sFLT1 production, which was low and quite variable (Fig. C.5D). Further, there were no significant differences in sFLT1 associated with culture in high $O_2$ in either EOPE-TB or CTL-TB. However, there was a slight effect of sex (P-Value = 0.04), with females somewhat producing more sFLT1 than males.
C.4.6 Gene Expression Profiles of EOPE-TB and CTL-TB

Next, we asked whether there were differences in gene expression that distinguished EOPE-TB and CTL-TB after culture under 5% and 20% O$_2$ conditions. Each cell line (8 CTL and 14 EOPE), plus the additional H1 and H9 ESC controls was differentiated to TB under each O$_2$ condition (5% and 20%). After 6 days of BAP treatment, RNA was extracted and processed for RNAseq. The raw data were normalized to TPM and log-transformed (Li and Dewey, 2011). Next, PCA was performed on the top quartile of genes based on their variance across all the samples. The first three principal components, comprising 67% of the total variance, were then plotted (Fig. C.4). This PCA analysis indicated that the samples clustered by O$_2$ condition rather than by CTL-TB and EOPE-TB identity, an outcome not dissimilar to that displayed by the progenitor UC primary fibroblast cells but not by the iPSC lines (Fig. C.2). Next, differential expression analysis was performed by using EdgeR (McCarthy et al., 2012)(Robinson et al., 2010) to compare (EOPE-TB and CTL-TB under each O$_2$ condition). Only two genes (RPS17, FDR = 0.0005; MTRNR2L2, FDR = 0.005) differentially expressed between CTL-TB and EOPE-TB at 20% O$_2$ (Table S5). No other differentially expressed genes distinguished EOPE-TB and CTL-TB under either O$_2$ condition.

C.4.7 Co-expressed gene modules associated with invasion

Of the parameters tested to distinguish EOPE and CTL-TB, the strongest phenotype was the reduced invasive capacity of EOPE-TB under high (20%) O$_2$ conditions (Fig. C.2A and B). As it is clear that this loss of invasiveness under 20% O$_2$ is unlikely to be attributable to differential expression of any single gene, it is plausible to assume that there are multiple genes, most probably working in a combinatorial manner, contributing to this phenotype. Accordingly, a method (WGCNA) to identify gene groups that were co-expressed and correlated with cell invasion count (Fig. C.2A, Table S4), was used in order to determine if there was a correlation between co-expressed gene modules and the respective invasive capacity, disease status (EOPE or CTL), sex, and O$_2$ treatment for each cell line. In our first analysis, we used all of the datasets, comprising both disease states and both O$_2$ levels, and identified five gene modules (A1-A5; Fig. 5A). Although
there were two modules, A1 and A3, that were strongly linked to $O_2$ sensitivity, the correlation of A1 and A3 with invasion was quite modest ($P$-Value = 0.05). Moreover, there was no significant association of these two modules with disease. Because the lack of correlation with disease in all of the five modules could be due to the variance across all of the samples, WGCNA was performed for the CTL-TB and EOPE-TB datasets separately to again identify modules that might significantly correlate with the invasion count and $O_2$ levels. This analysis revealed 10 co-expressed gene modules in CTL-TB (Fig. C.5B; CTL1-10), of which two (CTL4 and CTL9) had a significant correlation with both invasion and $O_2$ conditions. The CTL4 module correlated positively with invasion potential ($P = 0.03$) and negatively with exposure to 20% $O_2$ ($P = 0.02$). This module was enriched with ontology terms such as “stem cell differentiation” (q-value: 5.19e-3), and “blood vessel morphogenesis” (q-value: 7.80e-4). By contrast to the CTL4 module, the CTL9 module showed a negative correlation with both invasion potential ($P = 0.03$) and a positive correlation with exposure to 20% $O_2$ ($P = 0.02$). This module was enriched with ontology terms such as “blood vessel development” (q-value: 7.53e-5), “pre-eclampsia” (q-value: 1.23e-3), “regulation of cell-cell adhesion” (q-value: 3.88e-3), and “negative regulation of cell migration” (q-value: 2.96e-2). The top 100 genes from the CTL9 module, sorted by their module membership value (KME), which calculates the significance of an individual gene within a module, are listed in Table S6. Contained within this Table are a number of genes associated with integrin signaling (SEMA7A, RAP1B), cell migration and adhesion (SDC1, PVRL3), and genes reported to play a role in the cellular stress response (PPM1D, HSPB1, DNAJB9). Although not in the top 100, another interesting gene found in the CTL9 module was PGF (KME = 0.947), which, at the protein level, was also upregulated in response to 20% $O_2$ in CTL-TB only (Fig. C.3C).

Unlike the CTL-TB, where ten modules of co-expressed genes were identified, there were only four co-expressed gene modules identified in the EOPE-TB (EOPE1-4; Fig. C.5C), of which none displayed a significant correlation with invasion. There were two modules (EOPE1 and EOPE2) that were significantly correlated with 20% $O_2$ and 5% $O_2$, respectively. These modules are composed largely of genes similar to those identified as correlated with $O_2$ levels and invasion in the
CTL samples. For example, the EOPE1 gene module in EOPE-TB, which had a significant positive correlation with 20% $O_2$, has 83.74% CTL9 module genes (Fig. C.5B). Similarly, the EOPE2 module in EOPE-TB, which had a significant positive correlation with 5% $O_2$, has 66.7% CTL4 module genes. These transcript level outcomes may indicate a lack of an $O_2$-regulated, invasive response in the EOPE-TB, or they may also be an indication that more statistical power is needed for analyzing the disease samples. Finally, when we added sex as a variable to the WGCNA heatmaps, none of the gene modules identified in the CTL and EOPE-TB were significantly correlated with sex (Figs. C.5A-C). A list of genes identified in modules from all WGCNA analyses are provided in Supplementary Data File 1.

C.5 Discussion

Our goal in this study has been to recapitulate the early stages of placental development as they might occur in a pregnancy complicated with EOPE, a disease where defective placentation is believed to arise as a result of inadequate TB development during the early stages of pregnancy. To investigate the basis of EOPE, we utilized TB cells generated from reprogrammed primary cells containing the genetic, but possibly also the epigenetic features of pregnancies that had manifested EOPE. To this end, we created 29 independent iPSC lines from umbilical cords of EOPE and CTL pregnancies, a potentially valuable resource (Table C.1). Our study has assessed features of TB differentiation and function for 22 (8 CTL, 14 EOPE) of the 29 lines generated.

Difference in invasive phenotype between EOPE and CTL TB. Despite considerable variation in the rates at which different iPSC differentiated to TB in response to BAP treatment, one feature of the study was highly consistent, namely there were clear differences between the EOPE and CTL lines in their invasive properties when they were cultured in non-physiological 20% $O_2$. In fact, for what is believed to be a multifactorial disease, it was surprising to us that 11 of the 14 EOPE-TB lines (with two of three outliers approaching significance) demonstrated reduced TB invasion under these high $O_2$ condition (Fig. C.3), while the controls did not. This outcome is consistent with the central hypothesis that poor TB invasion and subsequent impaired spiral artery
remodeling and placental perfusion, underpin the fundamental insult (or at least the first “step”) leading to EOPE (Roberts and Redman, 1993)(Sones and Davisson, 2016). Although inadequate responses to $O_2$ may underpin the reduced invasiveness noted here, there may also be an enfeebled reaction to other stressors, a possibility we are presently testing. It should also be emphasized that the TB generated by BAP treatment of iPSC and ESC likely represents that formed during the implantation phase of placentation (Roberts et al., 2018), a period when conceptus loss in humans is thought to be particularly high. Whether, these early losses are exaggerated in pregnancies that might otherwise culminate in EOPE is unknown but should be a consideration. As concluded by Burton and Jauniaux (Burton and Jauniaux, 2004), “miscarriage, missed miscarriage, and early and late onset PE represent a spectrum of disorders secondary to deficient TB invasion”. Additionally we noted that the two ESC lines behaved like CTls, and that male and female lines within each grouping, did not differ significantly in invasive behavior.

Production of angiogenic factors and placental hormones by CTL-TB and EOPE-TB. If the first step in establishment of EOPE is inadequate TB invasion, the second is generally recognized to be the responses of the maternal system to this compromised state. Thus, EOPE is recognized by early presentation of maternal symptoms in the form of hypertension, proteinuria and endothelial dysfunction, even though the only cure is removal of the placenta (Fisher, 2015), suggesting that placental TB is the source of the aggravating factors. The maternal endothelial dysfunction in both early and late onset froms of PE seems, in particular, to stem from altered balance in the release of pro- and anti-angiogenic factors from the affected TB cells (Karumanchi and Bdolah, 2004)(Maynard and Karumanchi, 2011)(O’Brien et al., 2017)(Schaarschmidt et al., 2013)(Tidwell et al., 2001). Among these factors are PGF and sFLT1. While concentrations of the latter tend to rise in maternal serum, particularly during the second and third trimesters in pregnancies destined to present as preeclampsia, PGF levels usually fall, such that the ratio of the two has been exploited as a means to predict disease ahead of the onset of physical symptoms. Cultures of primary first trimester TB cells have also indicated that PGF and sFLT1 are reciprocally regulated by $O_2$ levels (Nagamatsu et al., 2004). In addition, with increasing $O_2$ concentrations (2% to 8% to 20%) the
levels of secreted PGF increase, whereas sFLT1 production falls. The $O_2$ mediated regulation of PGF appeared to be due to hypoxia-inducible factor 2 alpha (HIF2A) expression, as a siRNA of HIF2A abrogated the hypoxia-induced suppression of PGF (Fujii et al., 2017). In our experiments (Fig. C.3C), the output of PGF by CTL-TB was also regulated by $O_2$ in five of six CTL lines tested, with the levels significantly increasing under 20% $O_2$ conditions when compared to 5 %. By contrast, the EOPE-TB cells failed to upregulate PGF under high $O_2$ conditions suggesting an absence of the anticipated $O_2$ regulation. However, the same $O_2$ regulation was not displayed by CTL-TB or EOPE-TB with regard to sFLT1 production (Fig. C.3D), although release of this factor was low and quite variable. Nor were sFLT1 levels influenced by the disease origin of the cultures, although the study was relatively under-powered. There is a possibility that this low production of sFLT1 was a consequence of incomplete differentiation of the cells at d 6 of BAP exposure (Jain et al., 2017)(Roberts et al., 2018)(Yabe et al., 2016), and was not necessarily representative of sFLT1 production by TB throughout gestation. It should be recalled that serum sFLT1 and PGF levels have not been useful clinical predictors of EOPE until early in the 2nd trimester and onward, and even then cannot provide a completely reliable marker of disease (Chaiworapongsa et al., 2005)(Levine et al., 2004)(Thadhani et al., 2004). Nonetheless, the failure of 20% $O_2$ to stimulate PGF synthesis in cultures of EOPE-TB (Fig. C.3C), like the reduced invasiveness of EOPE-TB under high 20% $O_2$, is consistent with the disease origins of the cultures and suggests that these features of the disease are patent early in pregnancy.

Distinctions in gene expression profiles between CTL-TB and EOPE-TB. EdgeR analysis of RNAseq data to identify differentially expressed genes indicated that under 5% $O_2$, EOPE-TB and CTL-TB were essentially indistinguishable (Table S5). Even under 20% $O_2$ only two differentially regulated genes (RPL17 and NTRNR2L2) distinguished EOPE-TB and CTL-TB cultured in high (20%) $O_2$, with both genes significantly upregulated in the CTL cultures (Table S5). RPL17 (Ribosomal Protein 17) is typically reduced in concentration and its gene down-regulated when $O_2$ levels are low (Herrmann et al., 2013) and upregulated in response to stress (Yang et al., 2002). Similarly, NTRNR2L2 (often known as humanin-2) appears to be protective against hypoxia (Karu
et al., 2015). These data are consistent with the concept that EOPE-TB, unlike CTL-TB, fails to adjust appropriately to stress caused by culture under a 20\% \textit{O}_2 atmosphere. However, lack of up-regulation of RPL17 and NTRNR2L2 cannot in itself readily explain the reduced invasiveness of EOPE-TB. Therefore, given the paucity of differentially expressed genes between EOPE-TB and CTL-TB, we turned to WGCNA to determine whether particular groups of genes were correlated with loss of EOPE-TB invasiveness under 20 \% \textit{O}_2. However, we were unable to find a significant correlation between any co-expressed gene modules and invaded cell count for the EOPE-TB (Fig. C.5C), whereas there were strong correlations found for CTL-TB (Fig. C.5B). Importantly, however, one EOPE-TB gene module (EOPE1), which correlated positively with high (20\%) \textit{O}_2 level but not significantly with invasion had high overlap with the module CTL9. Similarly, EOPE2 had significant overlap in gene content with CTL4, which had a positive association with invasion count and negative association with high (20\%) \textit{O}_2 level (Fig. C.5B). Although, the weak association for EOPE modules EOPE1 and EOPE2 with invasion counts relative to the strong association noted for CTL4 and CTL9 might have a rather mundane explanation, namely that due to phenotypic variability more statistical power is needed to identify gene modules associated with invasion in EOPE-TB than in CTL-TB. For example there is a higher variance in log transformed invasion counts across EOPE-TB (0.26 at 5\% \textit{O}_2 and 0.35 at 20\% \textit{O}_2) compared to the variance of log transformed invasion counts across CTL-TB (0.11 at 5\% \textit{O}_2 and 0.18 at 20\% \textit{O}_2). On the other hand, the consistency of the \textit{O}_2 effects on invasion (Fig. C.2A) would appear to counteract this explanation. Accordingly, because EOPE modules EOPE1 and EOPE2 show a weak, though not significant, correlation with invasion count, the data likely reflect the enfeebled invasive phenotype of EOPE-TB in 20 \% \textit{O}_2. We therefore speculate that the regulatory networks that control invasion lose robustness in EOPE-TB as \textit{O}_2 levels rise, yet remain strong in CTL-TB. Whether this defect is because of an inability of EOPE cells to sense \textit{O}_2 (De Marco and Caniggia, 2002) and therefore respond appropriately, or due to disruption of EOTB’s ability to counteract the destructive effects of increased \textit{O}_2 levels still remains elusive. Possibly, the failure of EOPE-TB to up-regulate RPL17 and NTRNR2L2 in 20\% \textit{O}_2 is more consistent with the latter than the former explanation.
The failure of EOPE-TB to upregulate the proangiogenic factor, PGF, in response to changing O$_2$ conditions also strongly supports the premise that O$_2$ sensing is impaired, at least in the case of certain pathways. It supports a hypothesis that PGF production and TB invasion of EOPE-TB will fall progressively in vivo as extravillous TB approaches a more oxygenated or otherwise more stressful and challenging environment. It should be recognized, however, that CGA and CGB expression and the genes associated with P4 production during BAP-driven differentiation are both upregulated by increasing O$_2$ conditions (Amita et al., 2013) and that these events seem not to distinguish CTL-TB and EOPE-TB (Fig. C.3A and B), implying that not all O$_2$ sensing pathways are dysregulated. In fact, hCG secretion is regulated by oxygen-sensitive K+ channels in primary human placental syncytiotrophoblast (Diaz et al., 2016). Together, these findings favor the idea that the EOPE-TB cell lines retain a limited capacity to respond to changing O$_2$ concentrations in a manner similar to CTL-TB, despite the fact that other aspects of their phenotype, including their invasive capacity, becomes impaired as O$_2$ concentrations rise to potentially damaging levels.

Sexual dimorphism. According to at least one large population study EOPE occurs slightly more frequently in women carrying a male fetus (Lisonkova and Joseph, 2013). Curiously, of the first 15 EOPE placentae from July 2010 through February 2012, we collected 14 at 93.3% that were male, while among CTLs, half were females (Table C.1). In an attempt to balance for sex in our subsequent experiments, EOPE placentae P-S were deliberately selected from pregnancies that provided a female child. Although we did not pursue this apparent skewing towards males further, the observation raised the possibility that might be a bias towards male EOPE fetuses in the population we studied and that sexual dimorphism might exist in our cell lines. Differences in gene expression based on sex of the fetus have been reported for both human (Gonzalez et al., 2018) and mouse (Mao et al., 2010) placentas. Certain female iPSC have also reported to present a distinct transcriptome relative to male lines and to other female lines (Anguera et al., 2012), although these differences appear to be due to reprogramming differences, discordant expression of XIST, and the relative degree of X-inactivation among the female lines occurring during reprogramming. By contrast we obtained no evidence for large scale differences in gene expression comparable to those
observed by Anguera et al. (Anguera et al., 2012) in our transcriptome analyses of either primary fibroblast cultures or iPSC lines (Fig. C.1B). Moreover, sex showed no significant correlation with gene modules linked to either invasiveness or responsiveness of TB to \(O_2\) (Fig. C.5 A-C). In other words, there was no obvious relationship revealed between sex and EOPE phenotype. To be clear, however, there were certain features of TB physiology, e.g. production of hCG, that differed between males and females (Fig. C.3A), but we have not yet pursued this aspect of sexual dimorphism further.

Is the EOPE phenotype observed in iPSC-generated TB due to epigenetic carryover? A potential limitation to our study is the possibility that there were epigenetic marks placed on the DNA of the progenitor EOPE fibroblast cells that were retained through initial explant culture, expansion, reprogramming to iPSC, and ultimately differentiation. In other words, the EOPE-TB might possess an epigenome reflective of a previously stressed state of umbilical cord tissue at the time of collection rather than providing a basis for studying EOPE. This concern is not easy to address. One reassuring observation is that the \(O_2\) sensitivity displayed by the EOPE-fibroblast progenitors was lost once these cells were reprogrammed to iPSC and was only regained after they were differentiated to TB. Reprogramming to permit de-differentiation is known to remove most but not necessarily all epigenetic marks characteristic of a differentiated cell (Kim et al., 2010)(Roberts et al., 2017)(Shao et al., 2013). The relatively low sensitivity of ESC cells to \(O_2\), which has been studied previously in this laboratory (Westfall et al., 2008), and the apparent insensitivity of the iPSC lines described here, suggest that any epigenetic marks associated with responses to oxidative stress had been erased during reprogramming. On the other hand, it remains possible that the iPSC were protected from the deleterious effects of high \(O_2\) because of a shift from aerobic to anaerobic metabolism during reprogramming. This phenomenon, often called the Warburg effect (Gorlach et al., 2015)(Kondoh, 2008), may have been reversed during the subsequent differentiation of the iPSC to TB.
C.6 Conclusion

This study demonstrates the practical value of an in vitro pluripotent stem cell model designed to examine early TB development in pregnancies that had been complicated by EOPE. The basis of this model is the generation of iPSC from umbilical cords of infants born to mothers who experienced EOPE during the course of their pregnancies. These cell lines may be a useful resource for others wishing to study this enigmatic disease. Data generated by using this model are consistent with the concept that in EOPE the initial step precipitating disease is a reduced capacity of placental TB to invade, possibly brought about by a dysregulation of either $O_2$ sensing or protection mechanisms (Caniggia and Winter, 2002). The experiments tend to confirm that EOPE is a quantitative trait and that no one gene or even a small group of genes can be singled out as causing the disease.

C.7 Supplementary Material

The supplementary material (methods, figures, tables, and data) is available online (Supplementary Material).

C.8 References


Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., Hong, H., Nakagawa, M., Tanabe, K., Tezuka, K., Shibata, T., Kunisada, T., Takahashi, M., Takahashi, J., Saji, H.,


Figure C.1 (A) Cell doubling times performed on randomly selected cell lines (four CTL iPSC lines: MRuc4, MRuc7, MRuc8, MRuc10; six EOPE iPSC lines: MRucB, MRucE, MRucI, MRucJ, MRucK, MRucN). Values are the means ± SE for three individual experiments in each line. The O2 concentrations employed are represented by open (4%) and closed (20%) bars. (B and C) PCA of transcriptomes of UC fibroblasts from EOPE and CTL pregnancies and iPSC lines derived from them. The sexes of the samples are designated by circles (female) and triangles (male). (B) Neither primary UC fibroblasts (left; CTL, cyan; EOPE, magenta) nor the iPSC lines (right; CTL, silver; EOPE, gold) show separation by disease status. (C) UC fibroblasts cluster separately according to whether they were cultured under either 20% O2 (black) or 5% (red) O2. The iPSC lines (right; green and blue, respectively) did not separate by O2 condition. Data were generated with microarrays (GSE54400).
Figure C.2 (A) Comparison of TB invasion of 10 CTLs (CTL-TB, left: H1 and H9, lines 3 to 10) and 14 EOPE-TB (right: lines A to S) after exposure to BAP conditions for 6 d. Each letter or number corresponds to the sample ID in Table C.1. Values are the mean number of invaded cells ± SE in 12 randomly selected fields for three or four independent experiments, each in triplicate wells. For statistical comparisons, a complete randomized design was performed as a 2 X 2 X 2 factorial with CTL TB and EOPE TB, $O_2$, and sex as factors on log-transformed invaded cell counts to control for unequal variance across cell lines. Open bars indicate physiological (5%) $O_2$, and closed bars represent 20% $O_2$ conditions. Significance is highlighted (*P < 0.05; #P ≤ 0.054; pairwise comparison 5% vs. 20%). Cell lines outlined in red (5, 8, 9, and 10 CTL; A, P, Q, R, and S EOPE) are female. (B) Combined TB invasion results for all CTL TB and EOPE TB lines. The log-transformed least-square means were used to determine significance. (C) TB cell proliferation as assessed by total DNA content. Two ESC (H1, H9*; *female line), four CTL TB (5*, 6, 7, 8*), and four EOPE TB (A*, J, M, R*) were measured for total DNA concentration after 6 d of BAP treatment. Values are the mean DNA amount ± SE for three individual experiments. A two-way ANOVA indicated no significant differences in DNA concentration in response to $O_2$ concentration, and Bonferroni posttests determined no significance in pairwise comparisons. Overall DNA concentration was significantly affected by cell line (P < 0.0001). (D) The DNA content of TB cells (C) was log-transformed and plotted to determine whether there was a correlation with invasion cell counts (log-transformed values; counts) from A (values also shown in Table S4). There was no significant correlation between DNA concentration and the number of invaded cells of each line (Pearson r = 0.16, P = 0.50). Such lack of correlation is also observed in CTL lines (Pearson r = -0.02, P = 0.94) and EOPE lines (Pearson r = 0.07, P = 0.88) when analyzed separately.
Figure C.3  Expression of (A) hCG, (B) P4, (C) PGF, and (D) sFLT-1 in 24-h culture between day 5 and day 6 following initiation of BAP-induced differentiation. The experiment used two ESCs (H1, H9*; *female line), four CTL TB (5*, 6, 7, 8*), and four EOPE TB (A*, J, M, R*) lines. The concentrations of the various proteins measured by ELISA were normalized to DNA content of the cultures. Values are means ± SE for three individual experiments. Pairwise comparisons were determined by using the Student t test, with significance (P < 0.05) shown by asterisks.
Figure C.4  PCA of RNAseq data from CTL TB and EOPE TB (GSE119265). Data are based on the log-transformed TPM from the top quartile of the most variable genes. (A) Each dot represents an individual sample from either CTL TB 5% O2 (red), CTL TB 20% (black), EOPE TB 5% (blue), or EOPE TB 20% (green). The first three principal components show a separation by O2 treatment but not by disease (CTL vs. EOPE). (B) The same analysis as in A but showing the sex of each line. Each dot represents an individual sample from either female CTL TB 5% O2 (green), female CTL TB 20% (black), male CTL TB 5% O2 (blue), male CTL TB 20% (red); female EOPE TB 5% (yellow), female EOPE TB 20% (cyan), male EOPE TB 5% (gray), or male EOPE TB 20% (magenta). Shown on both plots are the locations of the CTL cultures derived from patients 7 and 8 (Table C.1), where the infant was born prematurely, and EOPE cultures E and S, which represent their nearest neighbors in the PCA plot. Also shown are locations of two male EOPE lines, J and I, derived from UC of dichorionic twins (Table C.1).
Figure C.5  Gene expression modules showing correlations in relation to invasive capacity, sex, O\textsubscript{2} treatment, and disease status (EOPE vs. CTL) as determined by WGCNA. (A) A total of 3,845 genes from combined CTL TB and EOPE TB samples identified five coexpressed gene modules. Out of the five modules, two (A1 and A3) showed a significant correlation with invasion and O\textsubscript{2} levels, but none correlated with disease status. (B) WGCNA analysis of a total of 3,682 genes for CTL TB. Ten coexpressed gene modules were identified. (C) WGCNA analysis of a total of 3,789 genes for EOPE TB. Four coexpressed gene modules were identified. Each module name is shown on the left, and each column represents the correlation value with the corresponding adjusted P value in parentheses. Positive correlations are shown in red, and negative correlations are shown in blue, with white signifying no correlation.
Table C.1  Patient information for UC samples collected from CTLs (1 to 10) and from infants whose mothers suffered EOPE (A to S). Note that the majority of EOPE lines were derived from infants after Cesarian delivery. Half of CTL lines and two EOPE lines were generated following spontaneous vaginal delivery of the infants. All mothers in the EOPE group, with the exception of B and E, received from one to four injections of betamethasone before delivery. Of the CTLs, only the mother of infant 7 received betamethasone.

<table>
<thead>
<tr>
<th>CTL or EOPE</th>
<th>Sample ID</th>
<th>sex</th>
<th>Gestational age (week)</th>
<th>Clinical condition</th>
<th>Mode of delivery</th>
<th>Betamethasone (B) (x # of injections)</th>
<th>Microarray analysis of cells grown in 4%, 20% O₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTL</td>
<td>1</td>
<td>M</td>
<td>39 3/7</td>
<td>Normal</td>
<td>SVD</td>
<td>no</td>
<td>no no</td>
</tr>
<tr>
<td>CTL</td>
<td>2</td>
<td>F</td>
<td>37</td>
<td>Normal</td>
<td>CD</td>
<td>no</td>
<td>yes yes</td>
</tr>
<tr>
<td>CTL</td>
<td>3</td>
<td>M</td>
<td>39 2/7</td>
<td>Normal</td>
<td>SVD</td>
<td>no</td>
<td>yes yes</td>
</tr>
<tr>
<td>CTL</td>
<td>4</td>
<td>M</td>
<td>40</td>
<td>Normal</td>
<td>CD (breech)</td>
<td>no</td>
<td>yes yes</td>
</tr>
<tr>
<td>CTL</td>
<td>5</td>
<td>F</td>
<td>39 3/7</td>
<td>Normal</td>
<td>(Repeat) CD</td>
<td>no</td>
<td>yes yes</td>
</tr>
<tr>
<td>CTL</td>
<td>6</td>
<td>M</td>
<td>40</td>
<td>Normal</td>
<td>SVD</td>
<td>no</td>
<td>yes yes</td>
</tr>
<tr>
<td>CTL</td>
<td>7</td>
<td>M</td>
<td>34</td>
<td>Preterm Premature Rupture of Membranes (PPROM)</td>
<td>SVD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>CTL</td>
<td>8</td>
<td>F</td>
<td>33</td>
<td>PPROM</td>
<td>SVD</td>
<td>no</td>
<td>yes yes</td>
</tr>
<tr>
<td>CTL</td>
<td>9</td>
<td>F</td>
<td>39</td>
<td>Normal</td>
<td>CD</td>
<td>no</td>
<td>no no</td>
</tr>
<tr>
<td>CTL</td>
<td>10</td>
<td>F</td>
<td>39 1/7</td>
<td>Normal</td>
<td>CD</td>
<td>no</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>A</td>
<td>F</td>
<td>30 5/7</td>
<td>HELLP syndrome</td>
<td>CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>B</td>
<td>M</td>
<td>29 5/7</td>
<td>severe PE</td>
<td>CD</td>
<td>no, due to emergent delivery</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>C</td>
<td>M</td>
<td>32 3/7</td>
<td>severe PE</td>
<td>CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>D</td>
<td>M</td>
<td>32 4/7</td>
<td>severe PE</td>
<td>CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>E</td>
<td>M</td>
<td>34 1/7</td>
<td>HELLP syndrome</td>
<td>CD</td>
<td>no, due to age &gt;34 weeks</td>
<td>no no</td>
</tr>
<tr>
<td>EOPE</td>
<td>F</td>
<td>M</td>
<td>32 6/7</td>
<td>severe PE, IUGR</td>
<td>CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>G</td>
<td>M</td>
<td>30</td>
<td>severe PE on top of maternal hypertension</td>
<td>(Repeat) CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>H</td>
<td>M</td>
<td>31</td>
<td>severe PE</td>
<td>(Repeat) CD</td>
<td>B (x 3)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>I</td>
<td>M</td>
<td>29</td>
<td>severe PE, dichorionic/diamniotic twins</td>
<td>(Repeat) CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>J</td>
<td>M</td>
<td>29</td>
<td>severe PE, dichorionic/diamniotic twins</td>
<td>(Repeat) CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>K</td>
<td>M</td>
<td>27 3/7</td>
<td>severe PE</td>
<td>CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>L</td>
<td>M</td>
<td>26 6/7</td>
<td>severe PE</td>
<td>CD (breech)</td>
<td>B (x 2)</td>
<td>no no</td>
</tr>
<tr>
<td>EOPE</td>
<td>M</td>
<td>M</td>
<td>33 5/7</td>
<td>severe PE</td>
<td>CD (breech)</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>N</td>
<td>M</td>
<td>28 5/7</td>
<td>severe PE</td>
<td>CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>O</td>
<td>M</td>
<td>28 5/7</td>
<td>severe PE</td>
<td>CD</td>
<td>B (x 2)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>P</td>
<td>F</td>
<td>31 5/7</td>
<td>severe PE, IUGR</td>
<td>CD</td>
<td>B (x 1)</td>
<td>yes yes</td>
</tr>
<tr>
<td>EOPE</td>
<td>Q</td>
<td>F</td>
<td>27</td>
<td>severe PE</td>
<td>CD</td>
<td>B (x 2)</td>
<td>no no</td>
</tr>
<tr>
<td>EOPE</td>
<td>R</td>
<td>F</td>
<td>34 1/7</td>
<td>severe PE</td>
<td>SVD</td>
<td>B (x 2)</td>
<td>no no</td>
</tr>
<tr>
<td>EOPE</td>
<td>S</td>
<td>F</td>
<td>32 6/7</td>
<td>HELLP syndrome</td>
<td>SVD</td>
<td>B (x 4)</td>
<td>no no</td>
</tr>
</tbody>
</table>
APPENDIX D. PLACENTACELLENRICH: A TOOL TO CHARACTERIZE GENE SETS USING PLACENTA CELL-SPECIFIC GENE ENRICHMENT ANALYSIS

Modified from a manuscript published in Placenta

Ashish Jain$^{1,2}$ and Geetu Tuteja$^{1,2}$

$^1$Bioinformatics and Computational Biology, Iowa State University, Ames, 50010, USA
$^2$Genetics, Development, and Cell Biology, Iowa State University, Ames, 50010, USA

D.1 Abstract

Single-cell RNA-Sequencing (scRNA-Seq) has improved our understanding of individual cell types in the human placenta. However, placental scRNA-Seq data has not been analyzed to identify genes with cell-specific gene expression patterns, which would be useful to understand how expression patterns in other model systems correspond to those in humans. Therefore, we developed PlacentaCellEnrich, a tool that takes a gene set as input, and then reports if the input set is enriched for genes with placenta cell-specific expression patterns, based on human placenta scRNA-Seq data. The PlacentaCellEnrich tool is freely available at https://placentacellenrich.gdcb.iastate.edu/ for non-profit academic use under the MIT license.

D.2 Introduction

The placenta is derived from extraembryonic tissue and undergoes changes in structure and function during pregnancy to cater to the needs of the developing fetus (Burton and Fowden, 2015)(Knöfler et al., 2019). The multiple functions of the placenta are coupled with the development of different trophoblast (TB) cells (Rossant and Cross, 2001)(Turco and Moffett, 2019). Due to the heterogeneous nature of the placenta, bulk RNA-Sequencing (RNA-Seq) data is insufficient
for understanding how the expression of different TB cells contribute to the establishment and maintenance of pregnancy. Single-cell RNA-Sequencing (scRNA-Seq) helps address this, by allowing quantification of gene expression levels in the individual cell types of the placenta (Suryawanshi et al., 2018)(Vento-Tormo et al., 2018)(Liu et al., 2018)(Tsang et al., 2017)(Pique-Regi et al., 2019).

Due to restrictions on human placental research, mammalian and cell culture models are often used to study placental development (Turco and Moffett, 2019). Despite the availability of scRNA-Seq data in human placenta, an understanding of how gene expression patterns from model systems correspond to specific human placenta cells is not always clear. This is due to the difficulty in mapping human placental scRNA-Seq data to other organisms. Therefore, we developed “PlacentaCellEnrich” (https://placentacell.enrich.gdcb.iastate.edu/), a user-friendly, interactive web application that can be used to compare gene lists generated in placenta model systems to gene expression in human cells, by carrying out a gene enrichment analysis.

D.3 Methods

D.3.1 scRNA-Seq Datasets

We used scRNA-Seq data for placental and decidua cells from Suryawanshi et al. (Suryawanshi et al., 2018) and Vento-Tormo et al. (Vento-Tormo et al., 2018) consisting expression data of 22 and 32 cell types respectively. The raw data was normalized into transcript per million (TPM) like values by dividing the raw reads with sequencing depth and multiplying by a scaling factor of 10,000. We used only protein-coding genes for cell-specific gene enrichment. The PlacentaCellEnrich tool uses one-to-one orthologous gene mappings from Ensembl (Zerbino et al., 2018)(Aken et al., 2016) to carry out cell-specific gene enrichment of gene lists from mouse, rat, and rhesus macaque, using human scRNA-Seq datasets.
D.3.2 Defining genes with cell-specific expression

Genes with cell-specific expression are defined based on the algorithm used in the Human Protein Atlas (HPA), that has also been implemented by the TissueEnrich tool (Uhlén et al., 2015)(Jain and Tuteja, 2018). Briefly, the genes were divided into six groups. These groups are:

- Not Expressed: Genes with an expression level less than an expression threshold specified by user across all the cells.

- Cell Enriched: Genes with an expression level greater than or equal to the expression threshold that also have at least five-fold higher expression levels in a particular cell type compared to all other cells.

- Group Enriched: Genes with an expression level greater than or equal to the expression threshold that also have at least five-fold higher expression levels in a group of 2-7 cell types compared to all other cells, and that are not considered Cell Enriched.

- Cell Enhanced: Genes with an expression level greater than or equal to the expression threshold that also have at least five-fold higher expression levels in a particular cell type compared to the average levels in all other cells, and that are not considered Cell Enriched or Group Enriched.

- Expressed in all: Genes with an expression level greater than or equal to the expression threshold across all of the cells that are not in any of the above 4 groups.

- Mixed: Genes that are not assigned to any of the above 5 groups.

Genes from the “Cell Enriched”, “Group Enriched”, and “Cell Enhanced” groups are classified as cell-specific genes.

D.3.3 Cell-specific gene enrichment

We used the hypergeometric test to calculate the enrichment of cell-specific genes in the input gene set. The p-value is calculated as:
\[
P(X > k) = \sum_{i=k+1}^{n} \binom{K_i}{i} \binom{N-K_i}{n-i} \binom{N}{n}
\]

and the fold-change is calculated as:

\[
Fold - Change = \frac{k}{K}
\]

\(N\) is the total number of genes, \(K\) is the total number of cell-specific genes for a cell type, \(n\) is the number of genes in the input gene set, and \(k\) is the number of cell-specific genes in the input gene set. Multiple hypothesis correction is done using the Benjamini & Hochberg correction.

### D.3.4 Background Genes

We provide an optional feature of using background genes for enrichment analysis. It should be noted that the background genes must contain all the genes in the input gene set. The p-value for the enrichment is calculated as:

\[
P(X > k) = \sum_{i=k+1}^{n} \binom{K_b}{i} \binom{N_b-K_b}{n-i} \binom{N_b}{n}
\]

and the fold-change is calculated as:

\[
Fold - Change = \frac{k}{K_b}
\]

\(N_b\) is the total number of background genes, \(K_b\) is the total number of cell-specific genes for a cell type in background genes, \(n\) is the number of genes in the input gene set, and \(k\) is the number of cell-specific genes in the input gene set.

### D.4 Results

#### D.4.1 Carrying out placenta cell-specific gene enrichment analysis

We used two human scRNA-Seq data sets to identify genes with cell-specific expression, as described in the methods section (Suryawanshi et al., 2018)(Vento-Tormo et al., 2018). We then developed a user-friendly web application that analyzes a gene list to calculate placenta cell-specific
gene enrichment using R Shiny (Version 1.4.0) (McPherson et al., 2017). Given a list of genes (e.g. differentially expressed genes from bulk RNA-seq data generated in placenta or genes in a scRNA-Seq cell cluster based on data generated in placenta model systems), the tool can be used to determine if the genes have cell-type specific expression in the human scRNA-Seq data sets. Below is a step-by-step tutorial for carrying out placenta cell-specific enrichment analysis, that is also outlined in Figure D.1. A guided tutorial is also available by clicking the “Placenta Cell Gene Enrichment” tab on the homepage.

1. After navigating to the homepage (https://placentacellenrich.gdcb.iastate.edu/), click on the “Placenta Cell Gene Enrichment” tab.

2. Select the type of Gene Identifier for the input gene set. Users can enter the input genes as Gene Symbols or Ensembl IDs.

3. Select the organism of the input gene set. Currently, PlacentaCellEnrich supports input data for commonly used organisms with a hemochorial placenta including humans, mouse, rat, and rhesus macaque.

4. Select the scRNA-Seq data used for the enrichment analysis. Users can choose either of the two available human placental single-cell studies. We used orthologous genes to carry out enrichment analysis for mouse, rat, and rhesus macaque using these datasets.

5. Select the type of cell-specific genes used for the enrichment analysis. The recommended setting is “All”. The description of cell-specific genes is described in methods section.

6. Enter the expression threshold to define whether the gene is expressed or not. By default, this is set to the recommended value of 1.

7. Select the type of histogram plot. Users can plot the enrichment analysis in terms of either Adjusted P-Values or the Fold-change, calculated during the enrichment analysis.

8. Enter the background gene set. This is optional, and users can use this feature if they want to carry out enrichment analysis using a custom gene set instead of all the genes in the datasets.
Users can either copy the gene list in the text area or upload it in a file using the browse button. The genes should be one per line for both cases. More information about background sets is available in the methods section.

9. Enter the input gene set. Users can either copy the gene list in the text area or upload it in a file using the browse button. The genes should be one per line for both cases. Users can also click the “Sample List” to test the tool. As described on the help page, the sample list is comprised of the 1,000 most highly expressed genes from EVTs described in Okae et al.

10. Click the submit button to carry out enrichment analysis.

11. Click the bar of any cell type in the histogram to see the expression of cell-specific genes in a heatmap.

12. Click the download button to download the enrichment values in tab-separated text values.

D.4.2 Visualizing expression patterns of genes across placenta cell types

PlacentaCellEnrich also enables users to check the expression of genes across various placental cell types in the two human scRNA-Seq datasets (Suryawanshi et al., 2018)(Vento-Tormo et al., 2018). Below is a step-by-step tutorial on how to visualize the expression of a genes across placental cells (Figure D.1).

1. After navigating to the homepage (https://placentacellenrich.gdcb.iastate.edu/), click on the “Placenta Cell-Specific Genes” tab.

2. Select the organism of the placental scRNA-Seq data to use. Currently, PlacentaCellEnrich only has human datasets, so the only option is Homo Sapiens (default).

3. Select the scRNA-Seq dataset. Users can choose either of the two human placental single-cell RNA-Seq studies.

4. Enter the expression threshold to define whether the gene is expressed or not. By default, this is set to the recommended value of 1.
5. Enter the gene in the input search box. The search option suggests the genes that are present in the dataset while the user is typing.

6. Click the submit button to view the expression of the input gene across placenta cell types.

D.4.3 Example usage scenarios of PlacentaCellEnrich

D.4.3.1 Identifying genes with cell-type specific expression in a human TB cell culture system

Okae H. et al. isolated TB stem cell lines from human blastocysts, cultured them, and differentiated them into syncytiotrophoblast (syncytiotB) and extravillous trophoblast (EVT) cells (Okae et al., 2018). To demonstrate the use of our tool and the similarity of cells differentiated in Okae H. et al. to human placental cells, we used PlacentaCellEnrich on the 1,000 most highly expressed genes from syncytiotB and EVT cells differentiated in culture (Okae et al., 2018). As expected, PlacentaCellEnrich shows that differentiated EVT and syncytiotB cells are enriched for genes specifically expressed in first trimester placenta EVTs and syncytiotB cells, respectively (Figure D.2 and Figure D.3). Furthermore, PlacentaCellEnrich identified genes that are specifically expressed in first trimester EVT (or syncytiotB), and also expressed in the cells differentiated in culture (Supplementary Data). Many groups use cell culture models to recapitulate cell types from the first trimester placenta (Yabe et al., 2016)(Jain et al., 2017)(Sheridan et al., 2019)(Hadjantonakis et al., 2020)(Horii et al., 2016), and PlacentaCellEnrich can be used to determine the similarity between the gene expression profiles of the culture model cells to those from human placenta.

D.4.3.2 Comparing mouse and human placental gene expression profiles

In the mouse, TB giant cells proliferate rapidly around embryonic day (E) 7.5, prior to the establishment of blood flow, and by E9.5 blood flows between the mother and baby, transporting nutrients and oxygen (Woods et al., 2018). To demonstrate how PlacentaCellEnrich can be used to compare the expression profiles in mouse placenta with human TB cells, we used RNA-Seq data generated from the mouse fetal placental tissue at E7.5 and E9.5 (Tuteja et al., 2016). We
used the 1000 most highly expressed genes from these datasets. PlacentaCellEnrich shows the enrichment of both EVT and syncytiotrophoblast-specific genes at E7.5 and E9.5, with higher enrichment of syncytiotrophoblast-specific genes at E9.5 than E7.5 (Figure D.4 and Figure D.5). This is expected, since syncytiotrophoblasts help in nutrient transport, which is established by E9.5 (Turco and Moffett, 2019). PlacentaCellEnrich results also indicate that there is minimal decidual contamination in the microdissected mouse fetal placentas. Finally, using PlacentaCellEnrich, we identified EVT and syncytiotrophoblast specific genes that are expressed at each time point in mouse, and can be used to identify candidate genes for future studies (Supplementary Data).

D.4.3.3 Comparing mouse and human decidual gene expression profiles

Decidualization of the endometrium is an essential process in the establishment of pregnancy. In mice, this process is first initiated at the antimesometrial (AM) side of the implantation site, and then spreads to the mesometrial (M) pole (Das, 2010). Zhao et al. investigated the molecular differences between these regions in the mouse decidua, and using RNA-seq found a total of 1,423 differentially expressed genes (DEGs) between the AM (811 upregulated genes) and M (612 upregulated genes) regions (Zhao et al., 2017). Using PlacentaCellEnrich, we found that the DEGs upregulated in the AM region are most highly enriched for 1st trimester human epithelial glandular cell-specific genes (Figure D.6a and Figure D.7a), whereas the DEGs upregulated in the M region are most highly enriched with for Natural Killer cell-specific genes (Figure D.6b and Figure D.7b). These results help link bulk RNA-Seq data from mouse decidua to specific human placental cell types based on gene expression information.

D.5 Discussion

We developed PlacentaCellEnrich, a user-friendly web application that carries out placenta cell-specific gene enrichment using scRNA-Seq data from first-trimester human placenta cells. The enrichment analysis in PlacentaCellEnrich is carried out using the hypergeometric test, which has
also been used in Gene Ontology (GO) enrichment analysis. In our analysis, however, we calculate
the enrichment of cell-specific genes instead of the GO linked gene sets (Hahne et al., 2008).

The two human scRNA-Seq data used in PlacentaCellEnrich were selected based on specific
criteria. We used studies that made their processed data publicly available, and that also had at
least ten cell clusters (Suryawanshi et al., 2018)(Vento-Tormo et al., 2018). Due to the different
processing pipelines, the two datasets we used were not merged, and it is recommended that users
run analysis and view results using both datasets.

PlacentaCellEnrich allows users to input genes from model organisms, including mouse, rat, and
rhesus macaque. We mapped the orthologous genes between humans and other model organisms
using the Ensembl database. The Ensembl database annotates orthologous genes using BLAST
(Aken et al., 2016). To make the ortholog mapping more stringent, we only mapped the one-to-one
orthologs. However, there are alternative ways to map orthologs between species, including using
coding Sequence, Markov’s clustering, and other hybrid methods (Emms and Kelly, 2015)(Li et al.,
2003)(Miller et al., 2019). If a user prefers one of the alternate methods, they can first obtain the
orthologous gene list using the method of their choice, and then use PlacentaCellEnrich.

We provided three usage scenarios that demonstrate how PlacentaCellEnrich can relate gene
expression data from model cell culture systems or model organisms to gene expression data from
first-trimester human placental cells. Furthermore, PlacentaCellEnrich can be used to obtain gene
expression patterns across placental cell types for individual genes of interest. In summary, Pla-
centaCellEnrich is a valuable tool for the placenta research community, providing insights into
single-cell expression profiles from first-trimester human placenta.

D.6 References

Aken, B. L., Ayling, S., and Barrell, D. (2016). The Ensembl gene annotation system. Database,
2016.

sophical Transactions of the Royal Society B: Biological Sciences, 370(1663).


Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cumi- 
46(D1):D754–D761.

Figure D.1 Overview of the PlacentaCellEnrich website and options available for analysis.
Figure D.2 Bar chart showing placenta cell-specific enrichment of the 1,000 most highly expressed genes in EVT (a) and syncytiotrophoblast (b) differentiated from TB stem cells. Data were analyzed in PlacentaCellEnrich the using Vento-Tormo et al. dataset.
Figure D.3  Bar chart showing placenta cell-specific enrichment of the 1,000 most highly expressed genes in EVT (a) and syncytioTB (b) differentiated from TB stem cells. Data were analyzed in PlacentaCellEnrich the using Suryawanshi et al. dataset.
Figure D.4  Bar chart showing the placenta cell-specific enrichment of the 1,000 most highly expressed genes in E7.5 (a), and E9.5 (b) mouse placenta. Data were analyzed in PlacentaCellEnrich using the Vento-Tormo et al. dataset.
Figure D.5  Bar chart showing the placenta cell-specific enrichment of the 1,000 most highly expressed genes in E7.5 (a), and E9.5 (b) mouse placenta. Data were analyzed in PlacentaCellEnrich using the Suryawanshi et al. dataset.
Figure D.6  Bar chart showing the placenta cell-specific enrichment of genes upregulated in antimesometrial (a), and mesometrial (b) mouse decidua. Data were analyzed in PlacentaCellEnrich using the Vento-Tormo et al. dataset.
Figure D.7 Bar chart showing the placenta cell-specific enrichment of genes upregulated in antimesometrial (a), and mesometrial (b) mouse decidua. Data were analyzed in PlacentaCellEnrich using the Suryawanshi et al. dataset.