OOCYTE AND PREIMPLANTATION EMBRYO CROSS-SPECIES TRANSCRIPTOME META-ANALYSIS REVEALS DIVERGENCE AT GENE LEVEL BUT CONSERVATION IN FUNCTIONS

By

Peter Zachary Schall

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Comparative Medicine and Integrative Biology – Doctor of Philosophy

2021

ABSTRACT

OOCYTE AND PREIMPLANTATION EMBRYO CROSS-SPECIES TRANSCRIPTOME META-ANALYSIS REVEALS DIVERGENCE AT GENE LEVEL BUT CONSERVATION IN FUNCTIONS

By

Peter Zachary Schall

Two of the most critical stages in early development occur during the maturation of oocytes and during the first lineage specification during morula-to-blastocyst transition. The accurate regulation of the transcriptome during these essential events is necessary for the development of a healthy embryo. This thesis presents the culmination of custom pipelines developed to produce three meta-analyses: 1) transcriptome changes during oocyte maturation across four mammalian species (human, rhesus monkey, cow, and mouse), 2) predictive modeling of RNA binding proteins and microRNAs binding to the 3' UTR, impacting stability during oocyte maturation across four mammalian species (human, rhesus monkey, cow, and mouse), and 3) transcriptome changes during the morula-to-blastocyst transition and the establishment of the inner cell mass and trophectoderm across five mammalian species (human, rhesus monkey, cow, pig, and mouse). The results of these studies reveal that there are relatively few individual transcripts regulated commonly across species, while there are greater shared features at the pathway and functional level. This underscores that different species may utilize a different cohort of genes to accomplish a given outcome. Additionally, the pipelines developed for this thesis are highly applicable across many areas of biology.

ACKNOWLEDGMENTS

It may seem strange for some, while for myself, this personal achievement has been more about the journey than the destination. At the risk of sounding cliché, I would have never reached this point in my academic career were it not for the love and support of those around me. While having this space afforded to me, I would like to say a heartfelt thank you to those who impacted me the greatest:

Dr. Vilma Yuzbasiyan-Gurkan. Thank you for being the individual who first introduced me to the world of academic research, mentoring me through my Master's and beyond. None of this could have been accomplished without your support.

Dr. Patrick Venta. Thank you for your much appreciated support and advice over the years. Stretching unofficially from Master's, to officially as one of my committee members during my PhD, your knowledge and advice has helped form me to the scientist I am today.

Dr. Cedric Gondro & Dr. Dalen Agnew. Thank you for investing your time and effort in serving on my guidance committee. Your input, guidance, and advice have been essential for my success.

My Latham lab mates. Dr. Uros Midic, Dr. Meghan Ruebel, and Kailey Schoen. Thank you all for your support, guidance, and friendship over the years we were in the Latham Lab. My skills and knowledge were ever expanded due to your efforts.

Dr. Keith E. Latham. I count myself immensely lucky to have had you as a mentor during my PhD. It cannot be overstated the amount of knowledge you have bestowed upon me. Every facet essential to being a proficient scientist has been broadened and honed due to your mentorship. Thank you for all that you have done for me over the years.

My parents Dr. William D. Schall and Melanie V. Schall, and wife Chloe Schall. I have been truly blessed by the love and support afforded to me by my parents and my wife. Thank you, Mom and Dad, for raising me to always seek out and accomplish my goals, and the providing love and support throughout my life. Thank you to my wife, Chloe, for always being supportive through this rather long journey, proofreading manuscripts you know not the material, and most importantly, your love. Not all aspects of aid and assistance rendered unto me were of a scientific basis. The love and support afforded to me by those in my life, truly, have made this all possible.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

KEY TO ABBREVIATIONS xi

CHAPTER 1. INTRODUCTION 1
- 1.1 Introduction 1
- 1.2 Oocyte Maturation and posttranscriptional regulation via the 3' UTR 2
- 1.3 Morula-to-blastocyst transition 3
- 1.4 Putting the Bio in Bioinformatics 4
- 1.5 Discussion 7
- REFERENCES 8

CHAPTER 2. ESSENTIAL SHARED AND SPECIES-SPECIFIC FEATURES OF MAMMALIAN OOCYTE MATURATION-ASSOCIATED TRANSCRIPTOME CHANGES IMPACTING OOCYTE PHYSIOLOGY 10
- 2.1 Abstract 10
- 2.2 Introduction 11
- 2.3 Materials and Methods 14
  - 2.3.1 Data Selection Processing 14
  - 2.3.2 Human Data Processing 15
  - 2.3.3 Rhesus Data Processing 15
  - 2.3.4 Cow Data Processing 15
  - 2.3.5 Mouse Data Processing 16
  - 2.3.6 Differential Expression Calculation and Gene Homology 16
  - 2.3.7 Differential Expression Calculation and Gene Homology 17
  - 2.3.8 Gene Group Classification 18
  - 2.3.9 Correlating Stability and Translational Changes during Early Oocyte Maturation 18
  - 2.3.10 IPA Core Analysis 19
- 2.4 Results 20
  - 2.4.1 Identification of mRNA Sets According to Cellular mRNA Stability during Maturation 21
  - 2.4.2 Shared and Species-Specific Members of Different mRNA Stability Classes 21
  - 2.4.3 IPA Analysis of Shared and Overall Species Changes in mRNA Abundance 23
  - 2.4.4 IPA Analysis of Primate-Specific Stable and Highly Degraded mRNAs 25
  - 2.4.5 IPA Analysis of Species-Specific Changes in mRNA Abundance 26
  - 2.4.6 Regulation of mRNAs Related to Oxidative Phosphorylation 27

2.4.7 Relationship between Stability Classes and Early Maturational Changes in mRNA Translation 27
2.5 Discussion 29
2.6 Acknowledgements 35
2.7 Funding 35
APPENDIX 36
REFERENCES 45

CHAPTER 3. REGULATION OF MRNA STABILITY VIA THE 3' UTR DURING OOCYTE MATURATION 50
3.1 Abstract 50
3.2 Introduction 50
3.3 Materials and Methods 52
3.3.1 Study Selection 52
3.3.2 Sample Processing 52
3.3.3 Statistical Difference in Abundance & Stability Classification 53
3.3.4 3' UTR Identification and Extraction 54
3.3.5 Identification of RNA Binding Protein Motifs within the 3' UTR 54
3.3.6 Statistical Analysis and Machine Learning on 3' UTR Motifs 56
3.3.7 Ingenuity Pathway Analysis 58
3.3.8 Generation of Figures 58
3.4 Results 58
3.4.1 Correlation of 3' UTR Length and Stability 59
3.4.2 RBPs regulation MmRNA Stability vis 3' UTR binding 59
3.4.3 Stable mRNA targets poly(U) RBPs CPEB2, CPEB4, and U2AF2 60
3.5 Discussion 61
3.6 Acknowledgements 63
APPENDIX 64
REFERENCES 68

CHAPTER 4. CROSS-SPECIES META-ANALYSIS OF TRANSCRIPTOME CHANGES DURING THE MORULA TO BLASTOCYST TRANSITION: METABOLIC AND PHYSIOLOGICAL CHANGES TAKE CENTER STAGE 74
4.1 Abstract 74
4.2 Introduction 75
4.3 Materials and Methods 77
4.3.1 Overview of study design 77
4.3.2 Data set selection and data processing 77
4.3.3 Human embryos and data processing 79
4.3.4 Rhesus embryos and data processing 80
4.3.5 Mouse embryos and data processing 81
4.3.6 Cow embryos and data processing 81
4.3.7 Pig embryos and data processing 82
4.3.8 Differential expression calculation and gene homology 82
4.3.9 Differential meta-analysis 84
4.3.10 IPA 84

4.3.11 DEG and IPA Figures 85
4.4 Results 85
4.4.1 Overview of Datasets and Limitations 85
4.4.2 Shared DEGs observed for MBT and ICMTE DEG lists 87
4.4.3 Shared IPA Features for the Morula-to-Blastocyst Transition 87
4.4.4 Shared IPA Features for ICM-Enhanced and TE-Enhanced DEGs 90
4.4.5 Affected IPA Upstream Regulators Associated with ICM-TE Divergence 91
4.4.6 Taxonomic Differences in Gene and Pathway Regulation 92
4.5 Discussion 93
4.7 Supplemental Data 101
4.8 Funding 101
APPENDIX 102
REFERENCES 116

CHAPTER 5. OVERALL CONCLUSIONS AND FUTURE DIRECTION 125
REFERENCES 129

LIST OF TABLES

Table 1- Summary of different study parameters used to obtain embryos 113

Table 2 – Number and proportion of DEGs (full method) associated with indicated pathways 115

LIST OF FIGURES

Figure 2.1 – Flowchart of analysis. 37

Figure 2.2 – Gene regulation groups during maturation. 38

Figure 2.3 - Ingenuity Pathway Analysis (IPA) features during maturation from shared differentially expressed genes (DEGs). 39

Figure 2.4 - Additional Ingenuity Pathway Analysis (IPA) features during maturation shared by all species 40

Figure 2.5 - Ingenuity Pathway Analysis (IPA) features during maturation from primate-specific regulated mRNAs 41

Figure 2.6 - Overlap of species regulation of differentially expressed genes (DEGs) in the oxidative phosphorylation pathway. 42

Figure 2.7 - Connecting stability and translational classification. 43

Figure 2.8 - Key Ingenuity Pathway Analysis (IPA) features of translation-stability classified groups 44

Figure 3.1 – Flowchart of Analysis 65

Figure 3.2 – 3’ UTR Length versus mRNA Stability 66

Figure 3.3 - Identification of RNA binding proteins and predictive output on mRNA stability 67

Figure 4.1 - Quantification of numbers of identified MBT and ICMTE DEGs using hull and homology methods. 103

Figure 4.2 – Integration of MBT and ICMTE DEGs 104

Figure 4.3 - IPA Canonical Pathways for the MBT 105

Figure 4.4 - IPA Biological Functions for the MBT 106

Figure 4.5 - Overlap of IPA results from W.S. MBT and ICMTE 107

Figure 4.6 - IPA Canonical Pathways: ICM-enhanced 108

Figure 4.7 - IPA Biological Functions: ICM-enhanced.....109

Figure 4.8 - IPA Canonical Pathways: TE-enhanced 110

Figure 4.9 - IPA Biological Functions: TE-enhanced. 111

Figure 4.10 - IPA Upstream Regulator: ICM & TE-enhanced.....112

KEY TO ABBREVIATIONS

GV – Germinal vesicle, immature oocyte

MII – Metaphase II, mature oocyte

DEG – differentially expressed genes

FDR – False discovery rate

mRNAs – messenger RNAs

IPA – Ingenuity Pathway Analysis

CP – Canonical Pathways

UR – Upstream regulators

BF – Biological functions

UTR – Untranslated region

MBT – Morula to blastocyst transition

TE – Trophectoderm

ICM – inner cell mass

CHAPTER 1.

INTRODUCTION

1.1 Introduction

"A mouse is not a Cow" Toronto stem cell biologist Dr. Janet Rossant wrote in 2011 (1). On its face, this is an obvious statement. A scientifically trained mind is not required to visually comprehend that there are massive differences between adults of these two species. At the microscopic level, however, during the earliest stages of development all mammals traverse strikingly similar stages of development. Oocyte and preimplantation stage embryos of different species look very similar and later embryos also look similar due to evolutionary constrains. The oocyte undergo maturation, fertilization, cell division, gastrulation, the morula-to-blastocyst transition, cell lineage formation, and implantation. The convergence in appearance is somewhat misleading, such that a closer examination at the subcellular level reveals marked divergence in composition as revealed by the following meta-analyses. The seeming divergence, however, is not the end of the saga as some degree of convergence is hiding in that substantial functional similarities emerge that we are able to expose. During the course of early development, there are core events oocytes and preimplantation embryos must undertake and accomplish, regardless of species. According to the Ingenuity Pathway Analysis database, the functional category "meiosis" contains 419 member molecules (genes, proteins, or endogenous chemicals). Similarly, when examining a less apparent functional category, such as "function of mitochondria", there are 117 member molecules. With this great magnitude of molecules present within just these two functions, the number of possible combinations any given species can utilize to reach a common end point is monumental. To ascertain an understanding of the various

components a specific species utilize to regulate these functions can only by elucidated by employing a modern meta-analysis.

This thesis includes three chapters: 1) Essential shared and species-specific features of mammalian oocyte maturation-associated transcriptome changes impacting oocyte physiology, 2) Predictive modeling of RNA binding proteins binding motifs in oocyte mRNA 3'UTRs for five mammalian species reveals novel candidate regulators of mRNA stability during oocyte maturation, and 3) Cross-species meta-analysis of transcriptome changes during the morula to blastocyst transition: metabolic and physiological changes take center stage.

1.2 Oocyte Maturation and posttranscriptional regulation via the 3' UTR

Preceding maturation, mammalian oocytes are in a state of arrest at the prophase I stage. The maturational process occurs during each reproductive cycle, after the pre-ovulatory luteinizing hormone (LH) surge which initiates the resumption of meiosis. These immature oocytes (germinal vesical or GV), undergo a breakdown of the nuclear envelope (germinal vesicle breakdown or GVBD), condensation of chromosomes, spindle formation, followed by the extrusion of the first polar body. Upon completion of meiosis II the oocyte undergoes arrest again, halting as a mature oocyte (metaphase II or MII), awaiting fertilization. The ultimate success of fertilization and embryonic development, requires proper maturation. These mature oocytes contain the maternal genetic material and other essential factors required for preimplantation development and embryonic genome activation.

During oocyte maturation, the cell is transcriptionally inactive and no new transcripts are being produced. However, it has been found that there are thousands of mRNAs that have significant changes in expression during maturation. If the relative abundance of mRNAs is significantly changing without the addition of new transcripts, this modulation of expression

must be primarily a consequence of posttranscriptional modes of regulation. The traditional nomenclature of “up-regulated” and “down-regulated” simply do not apply. The changes in relative abundance are essentially a factor of mRNA stability. Those mRNAs exhibiting a relative increase in abundance are being preferentially stabilized, while those showing a decrease have undergone substantial degradation. The oocyte’s ability to modulate transcripts of mRNAs is varied and can be acted upon via RNA binding proteins (RBPs), micro-RNAs (miRNAs), nonsense-mediated decay, and other RNA degradation factors.

One of the primary modes of post-transcriptional regulation of mRNA processes (localization, stability, and translation), and by proxy stability, involves the 3’ untranslated region (UTR), which is the section of mRNA following the translation termination codon. There are a number of regulatory elements within the 3’ UTR that have been found to impact polyadenylation, translation, and the stability of mRNA. Additionally, there are complementary sequence motifs where RNA binding proteins (RBPs) can bind and likewise influence mRNA stability and translational state (2).

There are many factors that can impact human fertility, and an estimated 1/10 females of reproductive age suffer from some form of infertility (3). By first examining the regulation of mRNAs during maturation and their putative roles in pathways and function, followed by deciphering which RBPs are impacting mRNA stability, will allow for a greater understanding of which factors are potentially essential during oocyte maturation.

1.3 Morula-to-blastocyst transition

Post-fertilization, genome activation, and multiple rounds of cell division, the preimplantation embryo undergoes the morula-to-blastocyst transition (MBT). This transition involves the first cell lineage formation: inner cell mass (ICM) and trophectoderm (TE).

Undergoing this transition and cell fate determination requires an intricate coordination of physiological, morphological, and metabolic changes. In humans and domestic animals, embryonic mortality is a major underlying factor causing infertility. In cattle, 37% of embryonic mortality occurs within one-week post-insemination (4). Increasing the understanding of genetic changes and key molecular pathways and functions are essential to address these issues.

1.4 Putting the Bio in Bioinformatics

One of the major challenges in generating meaningful results from a meta-analysis of high-throughput sequencing data, is accurately ascertaining a biological narrative relevant to the study at hand. In short, getting the "bio" from bioinformatics. There is no simple solution to this multi-faceted problem. Researchers have to mine through thousands of datapoints and identify significant changes at the mRNA level, and then further ascertain the impact of those genetic changes at a functional level. Further compounding the difficulty, a meta-analysis by its nature, will necessitate the inclusion of multiple studies, and potentially multiple species.

Before even reaching a point where a narrative can be developed from resultant data, there are multiple preceding steps required: 1) identification of public data relevant to one's study and accessing said data; 2) integrating complex data of potentially different sources and then interpreting those thousands of data points, and; 3) succinctly present the data in a digestible format. Only once these goals have been completed, should a researcher attempt to discern a biological meaning, often, steps 2 and 3 must be repeated with different comparisons and iterations.

While the overarching goal of this research was to explore the biological imperatives during early development, nearly half of the efforts were centered around tackling these questions. To answer these questions, public repositories of sequencing data were accessed,

custom pipelines were developed to leverage techniques from various fields, integrating data from multiple different sources.

These three aims constituted reprocessing 31 public studies, comprised of 346 samples, totaling 4.392 TB of sequencing data, across five mammalian species. The initial steps in these studies mirrored those following a standard protocol of analyzing RNA sequencing of two conditions/stages: identify differentially expressed genes (DEGs). Upon completion of the standard section of the protocol, is where my methods diverge. Many legacy techniques would ascertain shared regulation of genes, within species, by finding the intersection of those with significant difference. However, as highlighted in the following chapters, and in other reviews (5), different techniques of generating sequencing data imbue batch effects (culture medium, library preparation kit, sequencing platform, etc.). Meta-analyses of RNA-seq (and in general, expression data) can be utilized to elucidate complex biological questions by integrating datasets of similar phenotypes, thereby increasing power. Specifically, the methods herein used the R package metaRNASeq (6), to integrate studies with comparable samples within species. This software package utilizes the p-values from input datasets, applies a Fisher's combination method to said p-values, thus deriving a unified species list of changes. While the inherent differences in input studies may initially seem detrimental, with the application of the metaRNAseq package, they become strengths. Using simple intersections of DEGs can vastly underestimate the number of changes. By utilizing the metaRNAseq method, no single study where a gene fails to meet significance, can preclude its inclusion.

Likewise, when comparing across species, the approach of gene intersection was expanded by taking into consideration sequence similarity. This is necessary due to the potential impact imbued from disparate qualities of genome builds, annotation, and evolutionary

divergence. Using the public database MetaPhOrs (7), which applies a tree-based method incorporating multiple other databases to derive a consistency score of gene orthologs/paralogs, allowed for the identification of genes with high sequence homology across all input species. It should be stated, that since both gene methods have limitations and strengths, therefore, the data was presented in tandem: 1) all genes for each species, and 2) only those genes meeting consistency score of sequences similarity. This, in our opinion, generates a more complete picture of transcriptome changes and increases the power of these analyses.

Upon identifying a list of genes, the next logical step is to explore what functions and pathways those genes are regulating. Previous public databases, such as DAVID (8), KEGG (9), and GO (10), have been adequate over the years in identifying pathways and functions with significant overlap to a list of genes, but lack plasticity in comparisons and custom analyses and do not provide information about direction of activity. Therefore, I opted to utilize the IPA (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis). software for this purpose. The IPA software not only outputs entries with significant enrichment (p-value), but also calculates a z-score which provides a measure of predicted activity. With two statistical measures (p-value and z-score), a standard operating procedure was developed, decreasing the difficulty in rank ordering results: 1) limit results to those with significant overlap ($p<0.05$), 2) remove results with a single gene overlap, 3) rank atop those with significant z-scores ($|z|>1.96$), and 4) sort by number of genes per entry. Even with this rank ordered data, the number of entries can seem insurmountable. Therefore, graphical representation was used to ascertain trends and important entries. Over the course of these analyses, numerous iterations of figures were developed to properly present data, such that a reader would find them intuitive and

easily understandable. Additionally, these developed formats aid in the process of identifying those results most meaningful and conserved within an analysis.

Simply listing the top 10 results and expanding on their relation to the model of study, is not sufficient. Interpretation of results and selection of entries should be grounded in the model of study, and effort should be given to integrate the results into a cohesive story. Simply relying on statistical outputs can blind one to potentially important results as those statistics are inherently reliant upon the data and databases.

While the difficulty of developing a cohesive biological narrative is lessened through the application of these steps (data acquisition, metaRNAseq, MetaPhOrs, IPA, and modified rank ordering of results), it is still an arduous and time-consuming task. Nevertheless, the efforts are worthwhile, and the resultant outputs allow for greater extrapolation across study and species. While the focus of these derived methodologies was developed and instituted in a specific use case, they can be applied to additional stages, models, and species.

1.5 Discussion

In the following three chapters, I present the meta-analyses of oocyte maturation, regulation of maternal mRNA stability via the 3’ UTR during oocyte maturation, the morula-to-blastocyst transition, and inner cell mass and trophectoderm (ICMTE) lineage specification. These works were aimed to identify shared and species-specific changes at both the gene and functional level. The results highlight the relatively few shared significant changes at the transcriptome level, while identifying the comparatively more shared features at a functional level. In all, these works underscore the importance of evaluating both the similarities and differences between mammalian species during early development and further the field in understanding the complexity of these essential functions.

REFERENCES

REFERENCES

1. Rossant J. Developmental biology: A mouse is not a cow. Nature. 2011 Mar 24;471(7339):457-8. doi: 10.1038/471457a. PMID: 21430771.

2. Mayr C. What Are 3' UTRs Doing? Cold Spring Harb Perspect Biol. 2019 Oct 1;11(10):a034728. doi: 10.1101/cshperspect.a034728. PMID: 30181377; PMCID: PMC6771366.

3. Mascarenhas MN, Flaxman SR, Boerma T, Vanderpoel S, Stevens GA. National, regional, and global trends in infertility prevalence since 1990: a systematic analysis of 277 health surveys. PLoS Med. 2012;9(12):e1001356. doi: 10.1371/journal.pmed.1001356. Epub 2012 Dec 18. PMID: 23271957; PMCID: PMC3525527.

4. Thomas E. Spencer, Early pregnancy: Concepts, challenges, and potential solutions, Animal Frontiers, Volume 3, Issue 4, October 2013, Pages 48–55, https://doi.org/10.2527/af.2013-0033

5. Song Y, Milon B, Ott S, Zhao X, Sadzewicz L, Shetty A, Boger ET, Tallon LJ, Morell RJ, Mahurkar A, Hertzano R. A comparative analysis of library prep approaches for sequencing low input translatome samples. BMC Genomics. 2018 Sep 21;19(1):696. doi: 10.1186/s12864-018-5066-2. PMID: 30241496; PMCID: PMC6151020.

6. Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics. 2014 Mar 29;15:91. doi: 10.1186/1471-2105-15-91. PMID: 24678608; PMCID: PMC4021464.

7. Chorostecki U, Molina M, Pryszcz LP, Gabaldón T. MetaPhOrs 2.0: integrative, phylogeny-based inference of orthology and paralogy across the tree of life. Nucleic Acids Res. 2020 Jul 2;48(W1):W553-W557. doi: 10.1093/nar/gkaa282. PMID: 32343307; PMCID: PMC7319458.

8. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5):P3. Epub 2003 Apr 3. PMID: 12734009.

9. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4;45(D1):D353-D361. doi: 10.1093/nar/gkw1092. Epub 2016 Nov 28. PMID: 27899662; PMCID: PMC5210567.

10. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330-D338. doi: 10.1093/nar/gky1055. PMID: 30395331; PMCID: PMC6323945.

CHAPTER 2.

ESSENTIAL SHARED AND SPECIES-SPECIFIC FEATURES OF MAMMALIAN OOCYTE MATURATION-ASSOCIATED TRANSCRIPTOME CHANGES IMPACTING OOCYTE PHYSIOLOGY

2.1 Abstract

Oogenesis is a complex process resulting in the production of a truly remarkable cell—the oocyte. Oocytes execute many unique processes and functions such as meiotic segregation of maternal genetic material, and essential life-generating functions after fertilization including posttranscriptional support of essential homeostatic and metabolic processes, and activation and reprogramming of the embryonic genome. An essential goal for understanding female fertility and infertility in mammals is to discover critical features driving the production of quality oocytes, particularly the complex regulation of oocyte maternal mRNAs. We report here the first in-depth meta-analysis of oocyte maturation-associated transcriptome changes, using eight datasets encompassing 94 RNAseq libraries for human, rhesus monkey, mouse, and cow. A majority of maternal mRNAs are regulated in a species-restricted manner, highlighting considerable divergence in oocyte transcriptome handling during maturation. We identified 121 mRNAs changing in relative abundance similarly across all four species (92 of high homology), and 993 (670 high homology) mRNAs regulated similarly in at least three of the four species, corresponding to just 0.84% and 6.9% of mRNAs analyzed. Ingenuity Pathway Analysis (IPA) revealed an association of these shared mRNAs with many shared pathways and functions, most prominently oxidative phosphorylation and mitochondrial function. These shared functions were reinforced further by primate-specific and species-specific differentially expressed genes

(DEGs). Thus, correct downregulation of mRNAs related to oxidative phosphorylation and mitochondrial function is a major shared feature of mammalian oocyte maturation.

2.2 Introduction

Rodent models, particularly mice, comprise the predominant animal models in biomedical research, owing to small size, ease of manipulation and husbandry, available tools for genetic manipulation, and an ever-increasing legacy of genomics, genetic, and other data to enable rapid hypothesis testing. However, although rodent models are highly valuable for some basic studies of mammalian biology, significant differences across species limit the value of rodent models, particularly in reproductive biology, where litter-bearing rodents have clearly different modes of regulation compared with mono-ovular species. In addition, it is well-established that even some of the most fundamental developmental events in the life of every mammal, such as early cell lineage commitment of cells to inner cell mass or trophectoderm, can differ across species in key mechanistic respects (1–4). This suggests that a substantial amount of variation may exist in controlling mechanisms relevant to reproductive biology, and understanding that variation is important for understanding the limits to which any given model organism informs us about human reproductive biology.

That mice are useful models for some aspects of human reproduction, whereas other species (e.g., cow) are more useful for other aspects was noted nearly two decades ago (5). The implicit lessons are that there is much to be learned by taking advantage of multiple mammalian model species to better understand the human embryo or embryos of any given species. In addition, it is important to understand which aspects of each species are shared, which are species-specific, and which are relevant to understanding human biology. Despite these obvious

conclusions, relatively little headway has been made to date on the incorporation of different mammalian models into our quest to understand human reproduction.

There are likely a variety of reasons underlying the limited use of diverse mammalian species to understand human reproduction, such as feasibility, cost, type of study (in vivo vs. in vitro), and biased perceptions of relevance. Despite such limitations, these other species have been extensively employed, including in recent studies using more current technologies such as transcriptome analysis (e.g., see Refs. 6–17). But efficient use of data from diverse species has been limited. Most data have been used for addressing immediate and narrowly focused questions of interest. Differences in developmental timing, assay platforms, and interlaboratory variations in methodology have presented barriers to the broader use of published data, and as a result, very few meta-analyses have been attempted for mammalian oocytes or preimplantation stage embryos.

Our goal here was to gain deeper insight into the fundamental mechanisms, pathways, and processes that contribute to mammalian oocyte maturation. Our strategy was to apply a novel combination of methods to complete a meta-analysis of transcriptome changes during oocyte maturation and compare these changes across multiple species.

Such an analysis must take into account the relationships between maternal mRNA storage, translation, and degradation. The controlling mechanisms of these processes are complex (18, 19). Within the cytoplasmic compartment, mRNAs can variably be translated, stored, or degraded, depending upon the actions of diverse RNA binding proteins, micro-RNAs, nonsense-mediated decay factors, and RNA degradation complexes (18–20). mRNA decay can occur without translation, during translation, or after translation. Deposition in storage granules or other depots can greatly extend mRNA half-life, particularly in oocytes (21), and exit from

storage can lead to faster degradation. mRNA degradation is achieved by both 5′- and 3′-directed exonucleases. Inhibiting translation initiation can enhance the rates of 5′ degradation by exposing the mRNA to de-capping. Conversely, stress-mediated inhibition of translation initiation or elongation can inhibit decapping and stabilize mRNAs. Poly(A) tail lengthening can enhance translation, whereas poly(A) tail shortening can enhance 3′ degradation. Degradation can also be coupled to translation or to translational stalling. Because maternal mRNA degradation can be coupled to translational recruitment (i.e., removal from storage) and translation, one can infer that for many mRNAs, a high rate of degradation during maturation indicates production of protein within the cell. Increased degradation may also reflect a shift to translation inhibition and less protein production during maturation. To distinguish between such possibilities and capture information about mRNA translation, we coupled whole oocyte transcriptome meta-analysis with data from a previous analysis of changes in mRNA translation during the first 8 h of in vitro oocyte maturation (22). The combination of these datasets allows maternal mRNAs to be characterized according to both pattern of stability/degradation and translational regulation.

This analysis revealed mRNAs that are regulated similarly, mRNAs that are regulated species-specifically, and the pathways and cell physiological functions that are associated with these classes of mRNAs. The results reveal for the first time that only a limited number of mRNAs are regulated similarly across all four species examined, but certain pathways and functions nevertheless emerge that are regulated in all species, marking these pathways and functions as fundamental to the overall process of oocyte maturation. In addition, the data reveal cell physiological functions associated with cohorts of mRNAs that are stable, moderately degraded, or highly degraded during maturation, indicating apparent roles before and after fertilization.

2.3 Materials and Methods

A summary of the process flow of the computational analysis for this study is depicted in Fig. 2.1. This included sample set processing for different species, identification of mRNAs of different stability classes, processing gene lists through QIAGEN Ingenuity Pathway Analysis (IPA; QIAGEN, Hilden, Germany), and interspecies comparisons of mRNA expression classes and associated IPA results. Differentially expressed gene (DEG) lists and IPA results are described in Supplemental Information (all Supplemental material is available at https://doi-org.proxy2.cl.msu.edu/10.6084/m9.figshare.14226368.v1).

2.3.1 Data Selection Processing

We identified four mammalian species (human, mouse, cow, and rhesus monkey) for which RNAseq datasets could be identified that contained both germinal vesicle, immature oocyte (GV) and metaphase II, mature oocyte (MII) stage oocytes, at least three biological replicates at each stage, and meeting other quality parameters. To access these datasets, we used The European Nucleotide Archive. Study parameters are listed in Supplemental Table S1, including sequencing platform, sequencing read format/length, and RNA sequencing preparation kit. Unless otherwise noted, each study was processed with the following methods. Raw sequencing data in FASTQ format were downloaded for processing. Initial quality metrics were conducted using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimming was conducted with Fastp (v0.20.0) (23): minimum quality threshold of 20, minimum length of 20, and removal of low complexity/mononucleotide reads. Genomes index and abundance quantified with Kallisto (v0.44.0) (24), using standard settings.

2.3.2 Human Data Processing

Two human studies were identified meeting criteria: PRJNA377237 and PRJNA293908 (12, 25). FastQC identified aberrant nucleotide distribution in the first 13 bp in both studies. Therefore, the Fastp settings were set to include a hard trim of 13 bp from start of reads. The human cDNA genome (GRCh38, build 100) was downloaded from Ensembl. Quantification and differential expression were conducted as detailed under Dataset Selection and Data Processing. MII stage oocytes were matured in vivo.

2.3.3 Rhesus Data Processing

Two rhesus studies were identified meeting criteria: PRJNA343030 and PRJNA448148/PRJNA448150 (13, 17). FastQC identified aberrant nucleotide distribution in the first 13 bp in PRJNA343030 and 6 bp in PRJNA448148/PRJNA448150, which were hard trimmed accordingly. The rhesus cDNA genome (Mmul_10, build 100) was downloaded from Ensembl. Quantification and differential expression were conducted as detailed under Dataset Selection and Data Processing. The two studies used two distinct rhesus monkey populations (Indian-origin and Chinese-origin), providing for inclusion of genetic diversity in this analysis. MII stage oocytes were matured in vivo.

2.3.4 Cow Data Processing

Two cow studies were identified meeting criteria: PRJNA261946 and PRJNA228235 (26, 27). FastQC identified aberrant nucleotide distribution in the first 6 bp for PRJNA228235, which were hard trimmed. The cow cDNA genome (ARS-UCD1.20, build 100) was downloaded from Ensembl. Quantification and differential expression were conducted as detailed under Dataset Selection and Data Processing. MII stage oocytes were matured in vitro.

2.3.5 Mouse Data Processing

Two mouse studies were identified meeting criteria: PRJNA342001 and PRJNA464431 (15, 28). FastQC identified aberrant nucleotide distribution in the first 13 bp in PRJNA342001 and 6 bp in PRJNA464431, which were hard trimmed accordingly. The mouse cDNA genome (GRCm38, build 100) was downloaded from Ensembl. The study PRJNA464431 consisted of three different mouse strains (B6, D2, and BDF1), which were processed independently, therefore resulting in effectively four mouse datasets. Quantification and differential expression were conducted as detailed under Dataset Selection and Data Processing. MII stage oocytes were matured in vivo.

2.3.6 Differential Expression Calculation and Gene Homology

Kallisto output was imported into R (v4.0) and processed with DESeq2 (v1.30.0) (29), and transcript abundance was collapsed to gene using Ensembl identifiers converted with biomartR (v2.45.8), summing transcript isoform abundances for each gene. Two different lists of genes were processed through DESeq2 for each study: unfiltered gene lists and only genes with high level of homology across species. Because normalization and expression threshold selection were performed independently on each gene list, the homology lists were not simple subsets of the full gene lists. As the IPA gene set enrichment software is primarily based on human and mouse data, there is some concern in mapping genes incorrectly for other species. To address this potential concern, the metaPhOrs database (30) was used to identify genes with a high degree of homology, and IPA was repeated on this set of homologous genes. All pairwise species comparisons were retrieved from the database. Genes with high homology scores across all species were retained, numbering 11,272. Differential expression calculations were then repeated on this list of homologous genes. Both methods of filtering were conducted in parallel to

ascertain the impact on mRNA overlap and gene set enrichment within IPA. For both methods, genes with an FPKM (fragments per kilobase of transcript per million mapped read) above 1 in at least one sample were included for differential expression calculation. For each study, DESeq2 (29) was used to calculate differentially expressed genes (DEGs) between GV and MII, where a positive log2(fold-change) indicates a higher expression in MII as compared with GV; the level of significance for genes was set at an adjusted P value (false discovery rate: FDR) below 0.05.

2.3.7 Differential Expression Calculation and Gene Homology

Within each species, the R package metaRNASeq (v1.0.3) (31) was used to calculate a meta *P* value between studies via the Fisher's combination method. In short, the Fisher's combination method assumes that gene counts follow a negative binomial distribution within each included study. For each gene in each study, the null hypothesis was tested; that each gene is not differentially expressed. Whereupon the Fisher's exact test is applied to calculate gene- and study-wise P values.

In a review of different library preparation kits used in RNA-seq (32), there are inherent differences and biases. This was confirmed in this analysis and presented in Supplemental Table S1, such that within species, there is a level of variability in the number of captured genes with expression and number of mRNAs by stability classification. By leveraging the metaRNAseq method, these differences become a strength by allowing the integration of different preparation and sequencing methods for the derivation of a cohesive transcriptome. Volcano plots, plotting pre- and post-metaRNAseq log2(fold-change) versus FDR, for all species and datasets, can be found in Supplemental Figs. S1, S2, S3, and S4.

2.3.8 Gene Group Classification

During oocyte maturation, transcription is inactive and there are no new mRNAs being produced. Therefore, it is incorrect to state that a gene is “upregulated” in MII compared with GV. What is occurring is that many mRNAs are degraded during maturation, thereby drastically changing the “background” mRNA population being used to identify DEGs. Thus, ”upregulated” mRNAs are preferentially stabilized (i.e., display longer half-lives), degrading at a lesser rate compared with the background population. mRNAs calculated to be “downregulated” are those that undergo an elevated rate of degradation (i.e., become destabilized or have shorter half-lives). In addition, mRNAs exhibiting no significant change in expression undergo a moderate amount of degradation, which is less than that impacting the “degraded” class of mRNAs. This situation requires the reframing of directional classification of gene classes post-metaRNASeq. Genes were thus classified by actual change during maturation: upregulated (MII > GV, termed “stable”), no significant change (termed “moderately degraded”), and downregulated (MII < GV, termed “highly degraded”). These trifurcated lists were then compared across species, based on gene symbol, deriving all possible distinct gene group overlaps. There were a number that were found to have discordant directionality: genes regulated in opposite directions across species. As the underpinning goal of this study was to identify core shared features, these genes were not included in the analysis nor appear in the counts of overlaps.

2.3.9 Correlating Stability and Translational Changes during Early Oocyte Maturation

Raw sequencing data in FASTQ format were downloaded for processing for polysome-associated mRNAs at germinal vesicle stage and at metaphase I of meiosis for mouse oocytes (22). Processing of samples was conducted as described under Dataset Selection and Data Processing, with the inclusion of the Fastp parameter of a minimum length of 36 bp for reads,

matching the original publication. Differential expression was calculated between 0h and 8h post maturation induction, corresponding to GV and metaphase I stages. Although not equivalent to GV versus MII comparisons, this analysis nevertheless provides some insight into whether mRNAs are recruited to or depleted from polyribosomes in response to maturation induction. Three classifications of temporal translation pattern were derived: activated (higher in 8 vs. 0h), repressed (lower in 8 vs. 0h), and constitutive (no significant difference). These classifications were intersected with the three stability classifications, resulting in nine gene groups.

2.3.10 IPA Core Analysis

Gene lists were analyzed through the use of Ingenuity Pathway Analysis (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis), focusing on Canonical Pathway (CP) and Diseases and Functions (DF) analysis tools (IPA database v. 11/2020) (33). IPA is a software suite that allows for the enrichment of Canonical Pathways (CP) and Biological Functions (BF, a manually selected subset of Diseases and Functions) and the development of novel networks, based on submitted gene lists. IPA was selected due to the robustness of their database (>7M interactions, >700 pathways, >800,000 expression datasets, and >30 integrated third-party databases), and because it is manually curated and has the ability to compare multiple datasets. As a typical gene set enrichment methodology, submitted gene lists are compared with the genes in each CP/BF to calculate a level of significant overlap (P value; significance set at 0.05). In addition, with the known impact of up- or downregulating a gene on CP or BF, the IPA software can calculate a direction (activated or inhibited) as indicated by a positive or negative z-score (significance set at $z > |1.96|$). It should be noted that the magnitudes of gene expression changes do not factor into the calculations; only the direction of change is used. Each derived gene group was submitted to IPA and the CP/BF results were retrieved.

2.4 Results

We first identified shared and species-specific DEGs (mRNAs that change during maturation from GV to MII stage) and nonchanging mRNAs, as well as shared and species-specific Ingenuity Pathway Analysis (IPA) CPs and BFs associated with gene sets. Through the analysis of these mRNAs and associated IPA results, we then assessed shared and species-specific aspects of maternal mRNA regulation during oocyte maturation.

We analyzed the transcriptomes of the included mammalian species by two methods. The first method used the full gene list, mapping Ensembl gene identifiers to gene symbols, and the second method limited the analysis to those genes with a high level of homology across all four species. This second method addresses disparities in species and associated genome builds (gaps in sequencing, unannotated genes, evolutionary divergence, etc.) and possible impacts on results. Although the utilization of the full unfiltered gene lists may include genes not annotated in all species, the homology-based analysis can result in a decrease in power for the detection of DEGs. Presenting the outputs for both methods provides the most complete view of the analysis. Both methods are valuable and ultimately displayed highly similar results.

In addition to accounting for gene homologies, we leveraged the metaRNAseq method. This allows for the integration of multiple sequencing studies based on their respective generated P values based on a Fisher's combination method. Because of different sequencing platforms, library preparation kits used, and breed/strain/ethnicity variation, and other methodological differences between datasets, this metaRNAseq method helps account for these variables. A total of 94 sequencing libraries were processed for this study (12 Cow, 38 Mouse, 18 Human, and 26 Rhesus). An average of 14,380 genes were captured per study (Supplemental Table S1).

2.4.1 Identification of mRNA Sets According to Cellular mRNA Stability during Maturation

We categorized mRNAs for each species as highly degraded (MII < GV), moderately degraded (MII not significantly different from GV), and stable (MII > GV), and analyzed total ("full") mRNAs detected and highly homologous mRNAs (Supplemental Tables S2–S5). Hereafter, numbers of mRNAs will be given in the format: "full mRNA number (highly homologous mRNA number)." Across the four species analyzed, we observed a median of 2,151 (1,316) stable, 11,962 (6,362) moderately degraded, and 2,048 (1,295) highly degraded mRNAs (Fig. 2.2 and Supplemental Tables S1, S2, S3, and S5).

2.4.2 Shared and Species-Specific Members of Different mRNA Stability Classes

Our next step was to identify mRNAs that changed in abundance with a similar pattern across species. We examined transcriptomes for mRNAs regulated similarly across all four species (Fig. 2.2, designated hereafter as "All-4" mRNAs). We also identified mRNAs that were regulated across three of the four species (Fig. 2.2, designated as "3 of 4" mRNAs). The combined set of the All-4 plus the 3-species mRNAs is referred to as "4&3" mRNAs (Fig. 2.2). We identified 993 (670) "3 of 4" DEGs include 645 (423) not in rhesus monkey, 33 (23) not in human, 174 (119) not in mouse, and 141 (105) not in cow for the full and homology methods. Including the use of the 4&3 mRNA sets provided a less stringent look at shared mRNAs that considers possible impact of differences in genome annotation completeness across species. Such differences could artificially underestimate the degree of conservation of mRNA temporal expression profiles and impacts on cellular functions, pathways, and processes during oocyte maturation. We reasoned, therefore, that allowing a single species exception to a pattern reduced

the risk of such an artifact impacting conclusions of the study. We therefore examined the mRNAs and associated IPA results for the 4&3, as well as the All-4 mRNAs alone.

A prominent result of our analysis was the limited number of All-4 mRNAs (Fig. 2.2, B and C, middle column "All-4" group). Although thousands of mRNAs changed in abundance during maturation within any one species, only 121 (92) were regulated similarly, comprised of 40 (24) stable and 81 (68) highly degraded, across all four species. These 121 (92) mRNAs accounted for an average of only 1.22% (full method) and 1.58% (Homology method) of the total number of 9,878 (5,825) DEGs identified across all four species combined. Even allowing for a single species exception, the number of 3-species mRNAs was still comparatively limited (Fig. 2.2, column "3 of 4" group in Shared mRNAs). These encompassed 993 (670), consisting of 521 (345) stable and 472 (325) highly degraded mRNAs (Supplemental Table S6). Further evidence of the dramatic interspecies differences in mRNA regulation was seen with respect to the moderately degraded mRNAs. Although a median number of 11,962 (6,362) mRNAs were categorized as moderately degraded (i.e., unchanged during maturation), the tremendous variation across species in mRNAs being stabilized or degraded resulted in a very limited number of mRNAs being classified as moderately degraded across all four species or even three of four species (Fig. 2.2).

Aspects of shared mRNA regulation specific to primate species were observed by considering mRNAs displaying changes in human and rhesus monkey only (Fig. 2. 2 and Supplemental Table S6). Genes with shared changes in mRNA relative abundance specific to rhesus and human, exclusive of 3-species or All-4 mRNAs, included a total of 248 (237), consisting of 125 (85) stable and 236 (152) highly degraded mRNAs. There were an additional 881 (390) primate-specific mRNAs classified as moderately degraded.

We observed a range of species-specific changes for mRNAs across the species, 200–1,474 (165–809) stable and 332–1,692 (220–837) highly degraded (Fig. 2.2 and Supplemental Table S7). The rhesus monkey had the fewest species-specific changes, and the human had the most. A range of species-specific moderately degraded mRNAs was found for full and homology analyses (819–4,024 and 154–777, respectively). We also note that some mRNAs encoded by the mitochondrial genome appeared in the species-specific DEG lists but not in the shared All-4 and 4&3 DEG lists.

2.4.3 IPA Analysis of Shared and Overall Species Changes in mRNA Abundance

During maturation, some mRNAs are dramatically degraded in abundance, suggesting translation to produce cognate proteins contributing to the maturation process or terminating production of those proteins to downmodulate associated functions. Conversely, some mRNAs do not undergo degradation, and thus may be reserved to contribute to later functions or may become translationally silenced to downmodulate certain functions. The IPA terms “inhibition” and “activation” applied to the highly degraded and stable cohorts of mRNAs, respectively, thus could indicate which biological pathways and functions are used/terminated and which are reserved/sustained for later use. Defining the biological functions of the stable and highly degraded classes of mRNAs could thus provide insight into a core set of essential features of oocyte maturation that are shared across species, namely pathways and functions that are directly associated with the process of oocyte maturation and those that are associated with maternal mRNAs roles in the early embryo. To ascertain the overall functional systems enriched per species and the core shared features, IPA was applied to the stable and highly degraded whole species (WS) DEG lists for each species, the All-4 DEGs, the 4&3 DEGs, and the primate-specific DEGs. In addition, moderately degraded mRNAs seen specifically in primates were

analyzed using IPA. The other moderately degraded gene lists were not processed through IPA due to the large number of gene members.

To identify the core shared processes of mammalian oocyte maturation, the IPA results of the All-4, 4&3, and the four whole species DEG lists were compared for stable and highly degraded mRNA classes (Fig. 2.3 and Supplemental Tables S8 and S9). The ERK/MAPK signaling pathway was significantly affected in IPA results for stable mRNAs identified in the All-4, 4&3, and individual whole species (WS)-DEG lists, with predicted activation in all of these except the All-4 set (Fig. 2.3 and Supplemental Table S8). Twenty additional pathways displayed significant activation (positive z-score) with stable mRNAs from the 4&3 and all individual WS-DEG sets, including IGF-1 signaling, ephrin receptor signaling, estrogen receptor signaling, IL-3 signaling, and Fms-related receptor tyrosine kinase 3 (FLT3) signaling in hematopoietic progenitor cells in all species (full and homologous DEGs; Fig. 2.3 and Supplemental Table S8).

When comparing the IPA CP results on the highly degraded genes, the most prominent result present across gene sets was for oxidative phosphorylation, which was inhibited for the All-4, 4&3, all individual WS-DEG lists, and for the primate-specific moderately degraded mRNAs, with additional effects observed among species-specific DEG sets (Fig. 2.3 and Supplemental Table S8). In addition, NRF2 (NFE2L2, nuclear factor erythroid 2-like 2)-mediated oxidative stress response, assembly of RNA polymerase II complex, and fatty acid beta-oxidation I had significantly inhibited z-scores for the 4&3 and all individual WS-DEG sets (Fig. 2.3 and Supplemental Table S8).

Additional similarities were seen comparing results obtained by IPA analysis of each individual WS-DEG list that were not present in the All-4 and 4&3 analyses. The shared CPs

identified using the full method included ones with activation in all species, such as PEDF (SERPINF1, serpin peptidase inhibitor clade F member 1)-mediated and IL-8 signaling, and the superpathway of inositol phosphate compounds. The results obtained with the homology method included CPs with significant activation z-scores such as superpathway of inositol phosphate compounds, endotehlin-1, thrombin, and Gaq signaling (Fig. 2.4 and Supplemental Table S8).

2.4.4 IPA Analysis of Primate-Specific Stable and Highly Degraded mRNAs

The comparison of rhesus monkey and human oocyte maturation is of interest for better understanding processes that may be unique to primates, and which may cooperate with those features identified from examining the shared mRNA lists. In addition, common features of these two closely related species would provide corroboration of results obtained for each species. Primate-specific stable CPs included: role of cytokines in mediating communication between immune cells and IL-12 signaling and production of macrophages. Primate-specific moderately degraded mRNAs were significantly enriched in oxidative phosphorylation and eukaryotic translation initiation factor 2 (EIF2) pathway, both activated. The assembly of RNA POL-II complex, glutamate receptor signaling, and the superpathway of methionine degradation were found in the highly degraded DEG results (Fig. 2.5). The BF analysis (Fig. 2.5) showed no overlap between methods; the full method consisted of RNA processing and lipid synthesis, whereas the homology method identified hypoplasia and gamete/gametogenesis. A similar result was found for the moderately degraded mRNAs, with the full gene list method yielding effects on translation of RNA and transport of ion/metals. The homology method results included MAPKKK cascade and phosphorylation of l-tyrosine. Analysis of the highly degraded DEGs yielded effects associated with contractility of muscle, exocytosis of dense core granules, and conversion of isocitric acid (Fig. 2.5).

2.4.5 IPA Analysis of Species-Specific Changes in mRNA Abundance

The aforementioned analysis focused on features of maternal mRNA regulation that were shared across species or that were primate specific. Because shared mRNAs accounted for a small fraction of the total number of mRNAs that were analyzed, and because each individual species displayed dynamic regulation of thousands of mRNAs, oocyte maturation is accompanied by species-specific modulations of maternal mRNAs. These species-specific modulations may signify species-specific requirements for oocyte function or early embryo development, which in turn may signify species-specific sensitivities to exogenous influences such as maternal health or environmental factors. In addition, species-specific mRNA modulations may act cooperatively with the shared changes. We therefore applied IPA analysis to the species-specific stable and highly degraded DEG sets to gain insight into pathways and functions associated with mRNAs that are regulated in a species-specific manner (Supplemental Tables S8 and S9).

Of note were those CPs/BFs overlapping from the All-4, 4&3, and species-specific datasets. Using both the full and homology DEGs, mouse-specific highly degraded DEGs were enriched for oxidative phosphorylation, mitochondrial dysfunction, and sirtuin signaling. These same pathways were also found for cow-specific highly degraded DEGs from the full method. In addition, human-specific highly degraded DEGs were enriched for the NRF2-mediated oxidative stress response. When comparing the BF results from the species-specific stable DEGs, the function organismal death was reinforced for all species, and cell cycle progression for human, rhesus, and mouse. There were no overlapping functions found from the highly degraded DEG functions.

2.4.6 Regulation of mRNAs Related to Oxidative Phosphorylation

One of the most prominent results identified was the shared inhibition of the oxidative phosphorylation pathway across all species and datasets from the highly degraded DEGs. From the IPA database, the oxidative phosphorylation pathway has 109 member molecules. Interestingly, while an average of 47 of those 109 genes were among the DEGs for any given species (Human = 49, Rhesus = 27, Cow = 49, and Mouse = 73), 31 were shared by three of the four species and eight were shared by All-4 species. There were 24 DEGs shared between any two species and 17 species-specific DEGs. When splitting the DEGs by mitochondrial complexes (I–V), the majority of shared DEGs (n = 5) were from complex I and most of the species-specific DEGs were in complex IV (Fig. 2.6). In addition, when cross-referencing the mouse DEGs with the translational state, we found 68 entries with 19 constitutively translated and 49 repressed.

2.4.7 Relationship between Stability Classes and Early Maturational Changes in mRNA Translation

As previously stated, during oocyte maturation, transcription is inactive and stored mRNAs are either degraded, translated, or stabilized for use later. Luong et al. (22) employed the RiboTag method to explore the early maturation-related changes in mRNA translation during the first half of in vitro maturation period for mouse oocytes, from the GV to first meiotic metaphase stages. They defined three distinct groups of mRNAs based on changes in polyribosome abundances during this interval: activated (increased representation in polyribosomes), constitutive (constant representation in polyribosomes), and repressed (reduced representation in polyribosomes). The availability of this analysis in the mouse provides an opportunity for better understanding how initial changes in translation status relate to the stability classes identified

here. This allowed comparison of the shared 4&3 class of mRNAs to the different translation classes defined on events during early oocyte maturation (Fig. 2.7).

Comparing the three translational groups with the three stability groups identified here for the mouse revealed that a large fraction of the stable DEGs (n = 959; 38.15%)) were classified as translationally activated. Most (n = 1,311, 60%) highly degraded mRNAs were in the constitutively translated class with another 813 (37.38%) of the highly degraded mRNAs in the translationally repressed class, and very few in the activated class. Most (n = 7,166, 82.99%) of moderately degraded mRNAs were constitutively translated (Fig. 2.7). Similar patterns were seen when comparing the All-4 and the 4&3 DEG sets to the mouse results, although the proportion of activated highly degraded mRNAs was much lower, as was the proportion of stable-repressed mRNAs (Fig. 2.7).

Subjecting the different translation-stability groups to IPA revealed prominent pathways and functions associated with particular combinations (Fig. 2.8 and Supplemental Tables S10 and S11). The stable-repressed category could only be examined for the mouse DEGs due to the small number of genes for the 4&3 group. The mouse stable-repressed group yielded significant associations with many signaling pathways (vascular endothelial growth factor VEGF, ciliary neurotrophic factor CNTF, insulin-like growth factor 1 IGF1, and Ephrin) and a prominent association with superpathway of inositol phosphate compounds, along with numerous entries for myoinositol, inositol, and phosphoinositide signaling and a relevant significant effect on calcium signaling. The stable-activated category for the 4&3 shared DEGs yielded significant effects for many CPs and BFs related to cell cycle and cell division such as G2/M DNA damage checkpoint, mitotic roles of polo-like kinase, kinetochore metaphase signaling, and cyclins and cell cycle regulation, as well as associations with cell viability, organismal death (inhibited), and

mRNA degradation. These results were also seen for the mouse DEGs, along with effects on many other CPs and BFs. One striking result of this analysis was that the top results obtained for highly degraded-repressed category for both 4&3 and mouse DEGs were for strong inhibition of oxidative phosphorylation, and an effect on mitochondrial function, both with very strong P values [-log10(P) > 20] and associated with effects on numerous mRNAs encoding mitochondrial proteins. Fatty acid beta-oxidation was also inhibited. Affected BFs included inhibition of ATP synthesis and oxidative phosphorylation and activation of oxidative stress. This result was accompanied by predicted activation of sirtuin signaling, a result that may emerge from IPA due to the role for sirtuin signaling in regulating mitochondrial functions (34). The highly degraded-repressed DEGs were also associated with inhibition of NRF2-mediated oxidative stress response and effects on multiple BF entries related to protein synthesis.

2.5 Discussion

An essential question in reproductive biology is what constitutes a high-quality oocyte in mammals. To answer this question, it is instructive to consider data obtained from different species to identify fundamental characteristics of normal oocyte maturation and strategies for managing the rich oocyte endowment to ensure not only oocyte maturation but also preservation of the requisite endowment of mRNAs to support early embryogenesis. We provide here the first meta-analysis to convey a comprehensive cross-species comparison of oocyte transcriptome changes associated with mammalian oocyte maturation. We applied a methodology that was designed to account for differences in library quality, molecular reagents, genome annotations, and sequencing platforms.

One main conclusion from this analysis is that, although each individual species displays many thousands of mRNAs that change in abundance, there emerged a small set of just 121 (92)

mRNAs regulated in common (i.e., highly degraded, or stable) across all four species, and just 993 (670) additional mRNAs that changed in at least three of the four species analyzed, which averaged to just over one quarter of DEGs observed between GV and MII stage oocytes for each species. Thus, the degree of species conservation for transcriptome change during oocyte maturation is limited. This discovery highlights a surprising degree of divergence, given the presumed central importance of maternal mRNA regulation in oocyte function and early embryo development, and suggests that essential functions may not be strictly enforced at the individual gene level but rather at the level of overall pathway and function.

Indeed, despite the limited number of shared DEGs, our analysis was successful in highlighting many pathways and functions that are either used/terminated (highly degraded mRNAs) or reserved/sustained (stable mRNAs) in common across four mammalian species by maternal mRNA regulation during oocyte maturation, denoted by the IPA terms inhibition and activation, respectively. Shared aspects of transcriptome regulation during early maturation were observed both for shared DEG lists and attendant IPA results, and by comparing the IPA analysis obtained for WS-DEG lists for each individual species. This suggests species divergence in the regulation of stability of specific mRNAs, but with an underlying adherence to an essential set of functional outcomes; i.e., different mRNAs may be regulated to achieve effects on shared pathways or biological functions.

The most prominent shared functional outcomes to emerge from examining pathways and functions associated with the shared DEGs was the downregulation of mitochondrial function, reflected in inhibition of oxidative phosphorylation. Inhibition of NRF2-mediated oxidative stress response was also prominent. A majority of the shared DEGs related to oxidative phosphorylation encode components of complex I, whereas many of the species-specific DEGs

related to oxidative phosphorylation encode proteins in complex IV. The highly degraded, translationally repressed (reduced in polyribosomes) subset of 4&3 DEG mRNAs were strongly associated with inhibition of oxidative phosphorylation and ATP synthesis, involving many mRNAs encoding mitochondrial proteins (Fig. 2.8). Such early maturational mRNA degradation coupled with reduced polyribosomal abundance during the first 8h of maturation suggests that this downmodulation of oxidative phosphorylation is an early, shared, regulated process that begins with translational repression followed by transcript degradation. This result indicates that across all four species there is a dramatic exit of mRNAs associated with mitochondrial function and ATP synthesis from the polyribosomes followed by degradation. We note that downmodulation of oxidative phosphorylation and mitochondrial function may be protective by reducing mitochondrial activity and limiting reactive oxygen species production (35). The simultaneous shared degradation-translational repression of the NRF2-mediated oxidative stress response pathway further suggests the importance of downregulating oxidative phosphorylation. Indeed, a deficiency in the degradation of mRNAs related to oxidative phosphorylation is a key feature of human and rhesus monkey oocytes that fail to mature (13, 36). mRNAs related to nucleotide excision repair, fatty acid beta oxidation, and assembly or RNA polymerase II complex were also seen for degraded mRNAs. This suggests that these functions are also used/terminated across species.

To our knowledge, this is the first study to examine the connection between maternal mRNA stability and translation status during early oocyte maturation. Conservation of these relationships is seen across species. Across stability categories, constitutively translated (mouse) mRNAs comprised the bulk of mRNAs detected. mRNAs that are moderately degraded across all four species are mostly in the mouse constitutively translated class. A large proportion of the

shared (All-4 and 4&3) stable mRNAs are in the mouse translationally activated category. Conversely, a large proportion of the shared highly degraded mRNAs are in the mouse translationally repressed and constitutively translated categories. This indicates that the regulation of degradation at least during early oocyte maturation is connected to translation for large cohorts of mRNAs.

Stable-repressed mRNAs in the mouse are associated with multiple signaling pathways, particularly inositol and calcium signaling effects (Fig. 2.8). Early translational repression of this subset of stable mRNAs may sequester them to support later functions, such as oocyte activation. The stable-activated subset of mRNAs were associated with numerous G2/M and M-phase pathways and functions, checkpoint controls, transcription, as well as shared activation of pathways related to stress response, endocrine and cytokine signaling, pluripotency, and microtubule dynamics, but apparent shared inhibition of functions related to death and chromosomal instability (Fig. 2.8). We also found that stable mRNAs were strongly associated with inhibition of apoptosis and organismal death and activation of cell survival and viability functions. These were accompanied by predicted activation of other basic functions such as cytoskeletal and cytoplasmic organization and DNA replication. These observations indicate that mRNAs that are stable and translationally activated during early maturation may endow the oocyte with numerous essential proteins that support key signaling processes and cell survival and proteins that support early embryogenesis.

Beyond the early oocyte maturation period, other maternal mRNAs may be reserved/sustained for later use in the embryo. Stable mRNAs across species were associated with a variety of signaling pathways such as ERK/MAPK, IGF1, Ephrin, Estrogen, IL3, and EGF, as well as DNA damage checkpoint regulation, and several biological functions related to

viability, cytoskeleton regulation, and transcription. Several of these functions were previously associated with mRNAs present on both MII stage oocyte and 1-cell stage polysomes (37). Earlier studies revealed maternal mRNAs encoding transcription regulators that act postfertilization (35, 38, 39) and both DNA repair and checkpoint control can be vital processes following fertilization (40–42). DNA repair and checkpoint control functions were also associated with polysomes in MII stage oocytes and fertilized embryos (37). Successful expression of proteins related to inhibiting apoptosis was previously proposed as an important embryo quality surveillance mechanism after fertilization (43).

Overall, this meta-analysis of oocyte maturation-associated transcriptome changes across four mammalian species used an improved approach that identified key shared functions that are driven by limited numbers of shared DEGs. Chief among these is the downmodulation of mRNAs related to mitochondrial activity, oxidative phosphorylation, and ATP synthesis. This is the first study to conduct such a meta-analysis in mammalian oocytes. Previous studies compared expressed mRNAs between oocytes of different species (16, 44), but a detailed look at maturation-related changes in the transcriptome conserved across species has not been reported. One previously published meta-analysis explored published microarray oocyte datasets for human, mouse, rhesus, and cow (44). Our study employed a more sensitive mRNA expression detection method (RNAseq vs. microarray), thereby quantifying a larger number of mRNAs, and employed a more rigorous approach to functional interpretation of the DEGs. These combined approaches extend insights gained previously. Indeed, we note that no DEG overlaps were observed across all species during oocyte maturation in the earlier study, whereas we identified 121 (92) shared DEGs. However, we also note that the previously reported Gene Ontology enrichment of DEGs associated with oocyte developmental competence shared by at least two of

the included species were also revealed in our findings of effects on mitochondrial function. This further supports the conclusion that the regulation of the oocyte transcriptome related to mitochondrial function and oxidative phosphorylation is a key aspect of oocyte maturation.

The limited degree of conservation of transcriptome changes (i.e., mRNAs displaying significant change in relative abundance) across species echoes recent observations comparing different mouse inbred strains and their F1 hybrids (15). In that study, substantial variation was also observed for transcriptome changes during maturation. These observations indicate that the dynamic process of transcriptome change during oocyte maturation is subject to considerable genetic variability. Our data reveal a core set of features that are indeed shared across species. However, there is a vast amount of maturational change in the oocyte transcriptome that is highly specific to individual species or strains. Even looking at the level of IPA pathways and functions, there are many differences across species. This remarkable species divergence poses significant challenges for efforts to identify molecular markers of oocyte quality, as well as endeavors to optimize in vitro manipulation systems, because oocytes and early embryos of different species may accordingly have very different optima for in vitro culture and other procedures. In addition, the finding that such differences exist in maternal mRNA regulation during oocyte maturation indicates that there are likely multiple strategies for the generation of high-quality oocytes employed in different mammalian species or even different strains. This discovery sets the stage for many interesting future studies to understand why such divergent strategies exist and what exogenous factors and forces have driven the emergence of such diversity among mammals. In addition, the results provide an important baseline against which to judge the extent to which transcriptome alterations impacting essential features emerge under conditions that compromise

oocyte quality. The understanding gained here of essential shared and species-specific aspects of oocyte maturation may thus be useful for designing novel methods for predicting oocyte quality.

2.6 Acknowledgements

We thank Lane Christenson for constructive comments on the manuscript.

2.7 Funding

This work was supported in part by grants from the National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development (T32HD087166), MSU AgBioResearch, and Michigan State University. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

APPENDIX

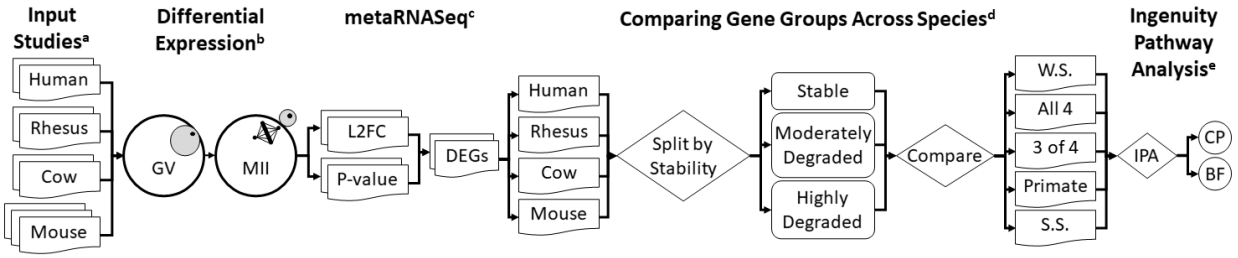

Figure 2.1 – Flowchart of analysis.

Input read counts for each study (a) were imported into DESeq2 and significant differences are calculated between germinal vesicle, immature oocyte (GV) and metaphase II, mature oocyte (MII) (b). Within each species, the P values for all included studies were input into metaRNASeq to calculate differentially expressed genes (DEGs) between GV and MII, resulting in whole species (WS) DEGs (c). Comparison of WS DEGs and derivation of gene groups (d). Each gene list was trifurcated based on direction (stable, moderately degrade, and highly degraded): 1) individual species results, 2) mRNAs regulated in the same direction from "All-4" species, 3) mRNAs regulated in the same direction by "3 of 4" species, 4) mRNAs regulated in the same direction by only human and rhesus, 5) mRNAs regulated in a species-specific (SS) manner, and each gene list was sub mitted to IPA (e) for Canonical Pathways (CP) and Biological Functions (BF) analysis.

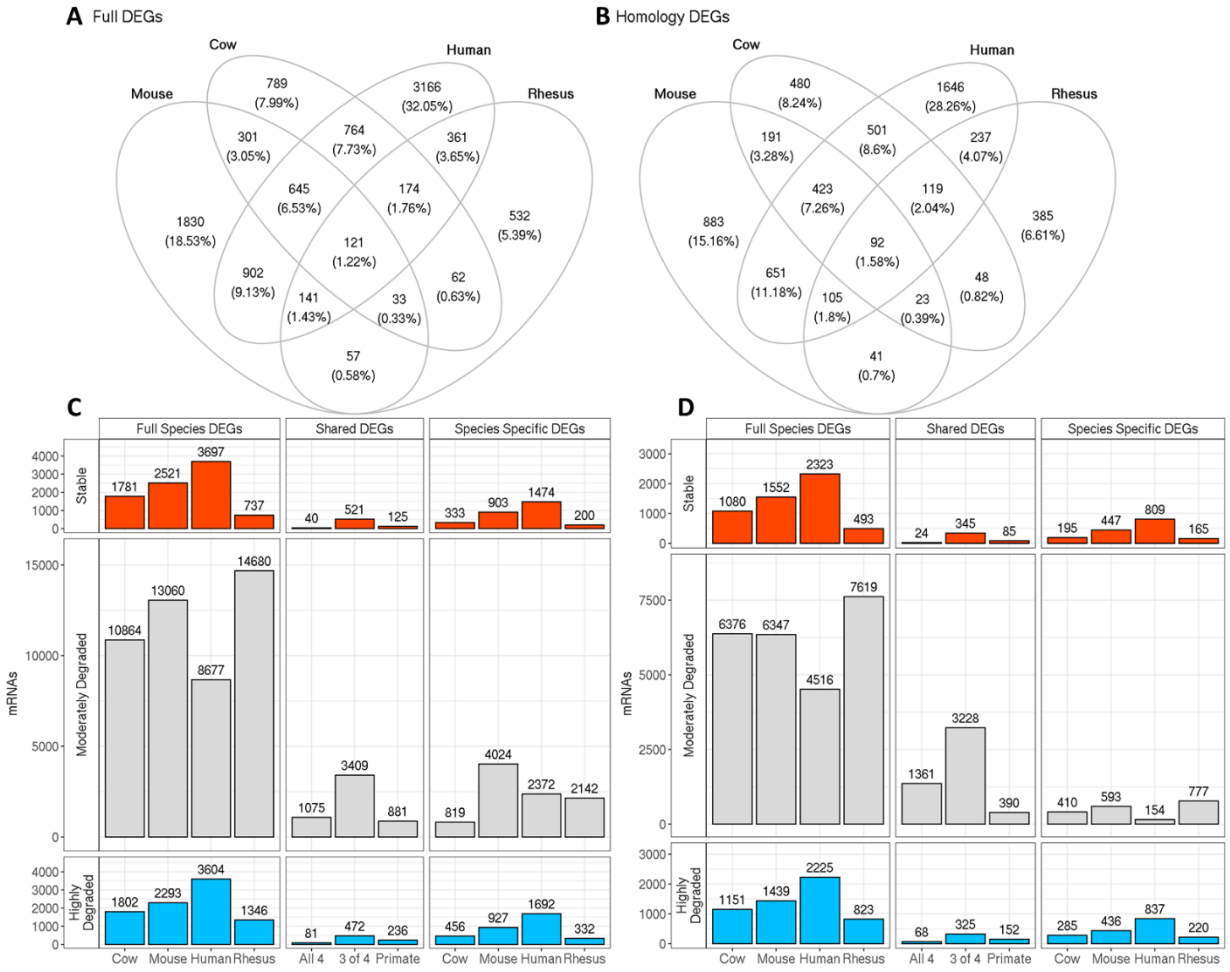

Figure 2.2 – Gene regulation groups during maturation.
Gene regulation groups during maturation. Visual representation of stability classes, numerological depiction of mRNAs, classified by regulation and species. A: Venn diagram overlap of stable + highly degraded mRNAs across the four species using the full gene method. B: Venn diagram overlap of stable + highly degraded mRNAs across the four species using the homology method. C and D: gene counts for full and homology method analyses, respectively, for mRNAs identified by stability classification. Column 1, WS mRNAs, denotes global classification of changes for each whole species gene lists. Column 2 (Shared mRNAs) shows selected mRNAs overlaps of: "All-4" species, "3 of 4" species, and primate-specific mRNAs. Column 3 (SS mRNAs) shows the number of mRNAs regulated in a species-specific manner. The associated mRNA numbers and gene lists are found in Supplemental Tables S2, S3, S4, S5, S6, and S7.

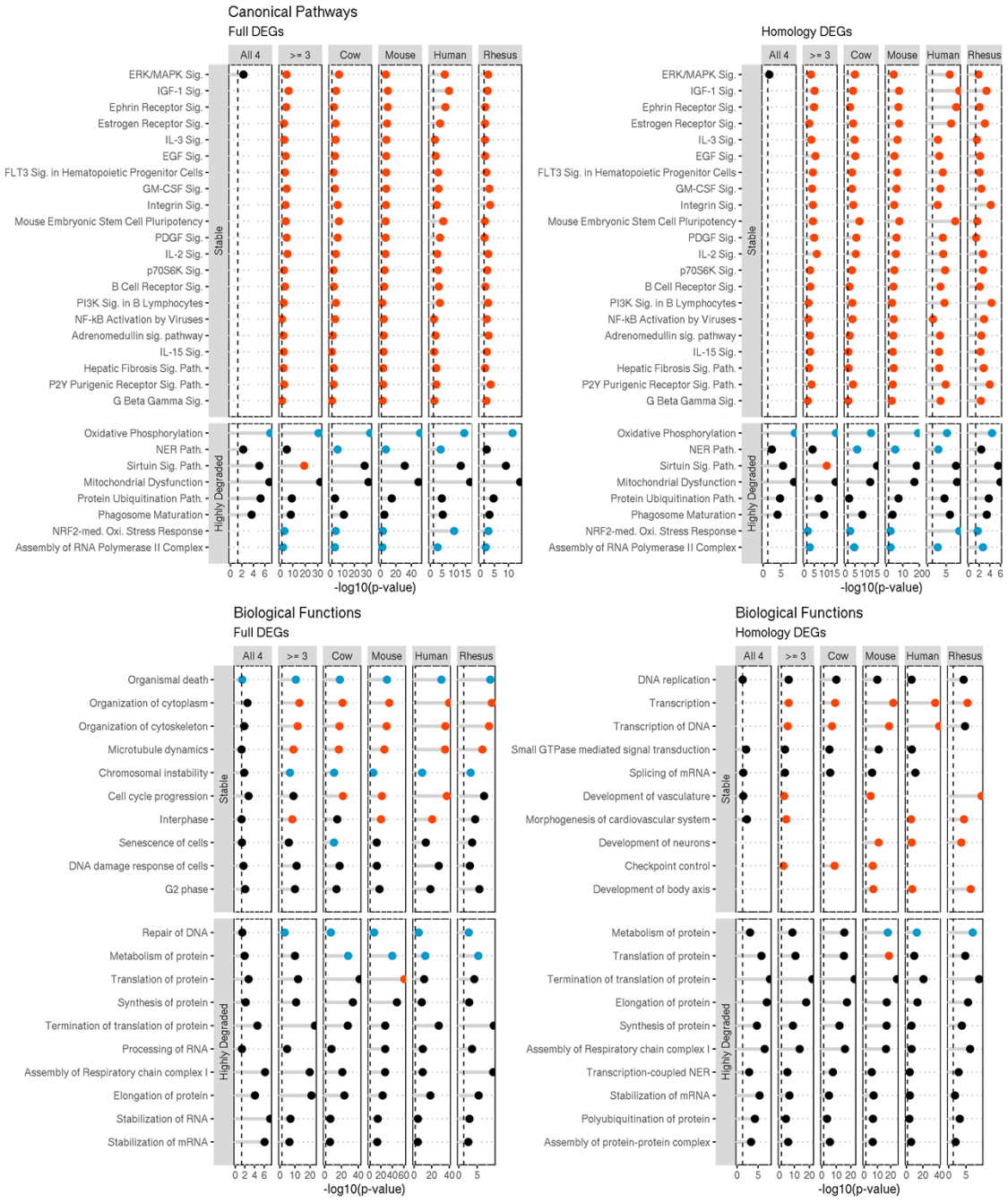

Figure 2.3 - Ingenuity Pathway Analysis (IPA) features during maturation from shared differentially expressed genes (DEGs).

The top selected IPA entries pertinent to oocyte maturation, derived from DEGs shared by all four ("All-4") and at least three species ("4&3"). The four panels represent IPA CP and BF results obtained for DEGs identified by the full and homology methods. Each panel has two vertical facet plots for each mRNA stability classification: stable and highly degraded. For each plot, the x-axis denotes the -log10(P value) and the y-axis are the IPA entries. The color of each points represents the z-score: activated = red, inhibited = blue, no-significant = black. Vertical dashed lines at 1.3 equates to a P value of 0.05. Associated data and additional IPA entries are listed in Supplemental Tables S8 and S9.

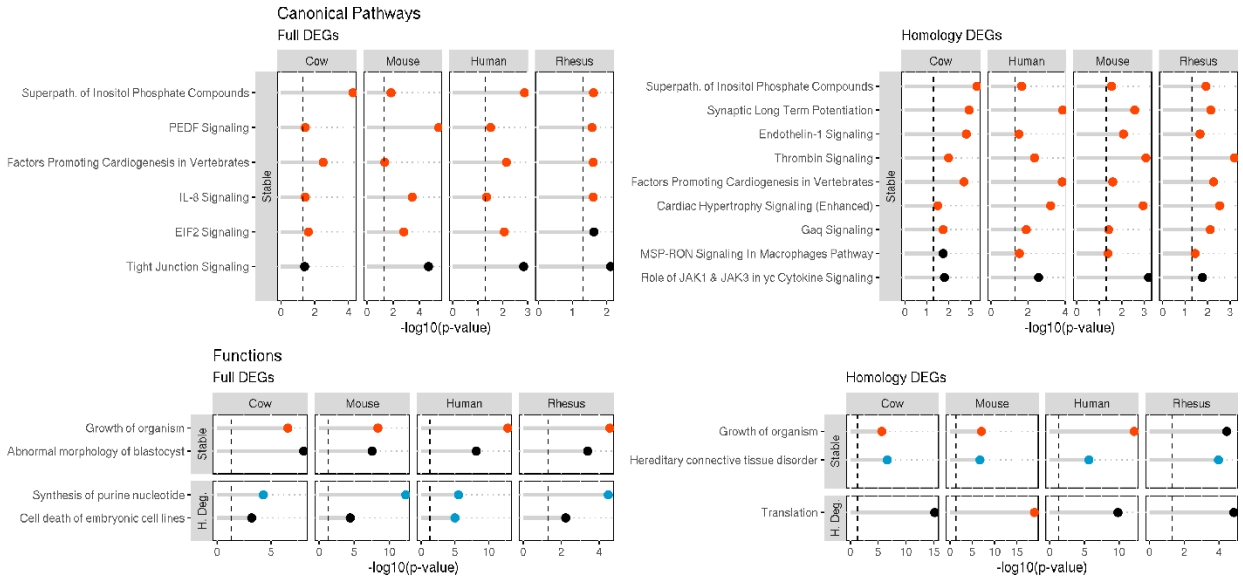

Figure 2.4 - Additional Ingenuity Pathway Analysis (IPA) features during maturation shared by all species.

IPA entries pertinent to oocyte maturation shared by all species while not present in All-4 and 4&3 analyses. The four panels show IPA CP and BF results obtained for DEGs identified by the full and homology methods. Each panel has two vertical facet plots for each mRNA stability classification: stable and highly degraded. For each plot, the x-axis denotes the -log10(P value) and the y-axis are the IPA entries. The color of each points represents the z-score: activated = red, inhibited = blue, not-significant = black. Vertical dashed lines at 1.3 equates to a P value of 0.05. Associated data and additional IPA entries are listed in Supplemental Tables S8 and S9.

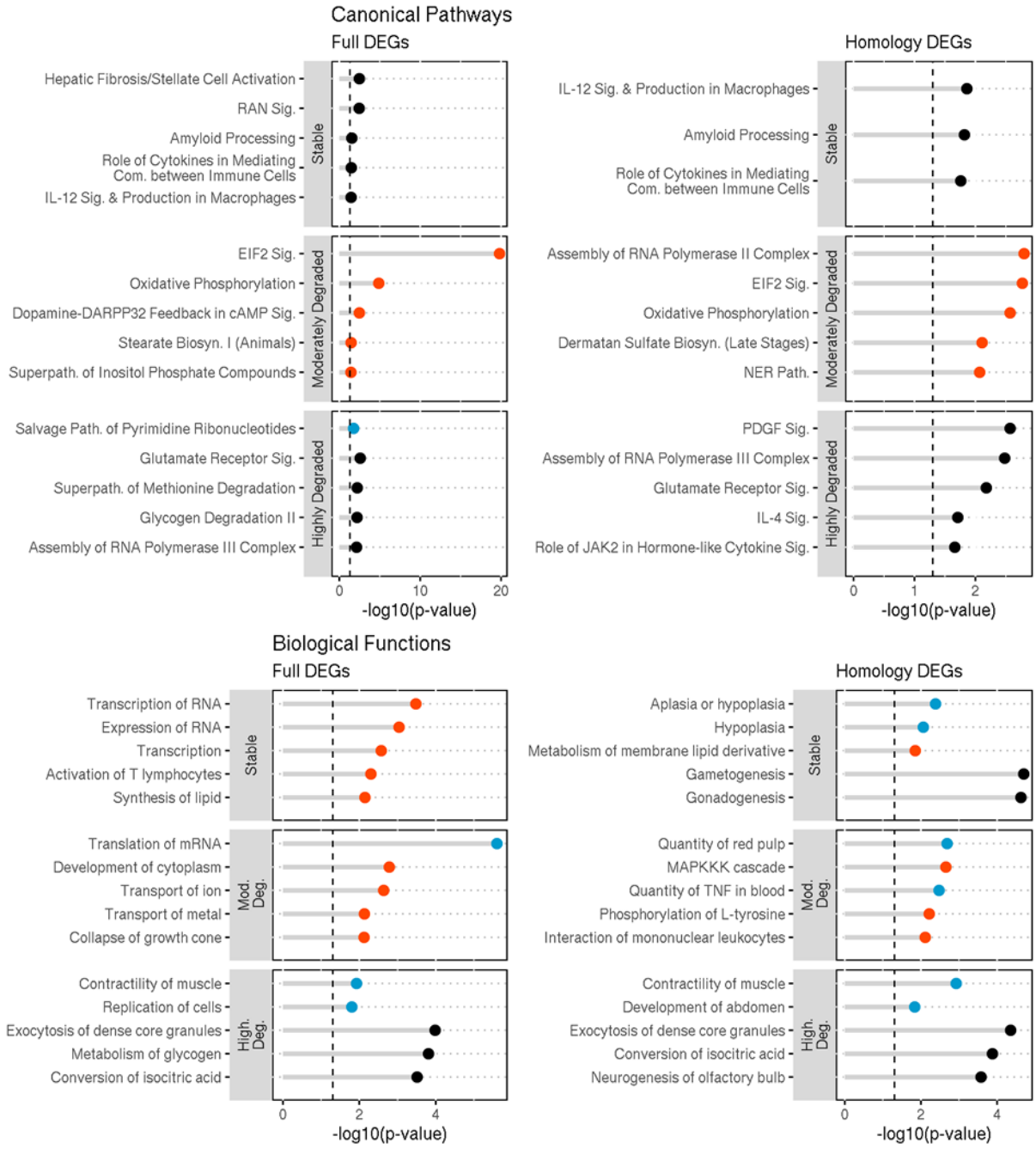

Figure 2.5 - Ingenuity Pathway Analysis (IPA) features during maturation from primate-specific regulated mRNAs.

The top selected IPA entries pertinent to oocyte maturation from primate-specific regulated mRNAs. The four panels represent IPA CP and BF results obtained for DEGs identified by the full and homology methods. Each panel has three facet plots for each mRNA stability classification: stable, moderately degraded, and highly degraded. For each plot, the x-axis denotes the -log10(P value) and the y-axis are the IPA entries. The black color of the points denotes no significant z-score. Vertical dashed lines at 1.3 equates to a P value of 0.05.

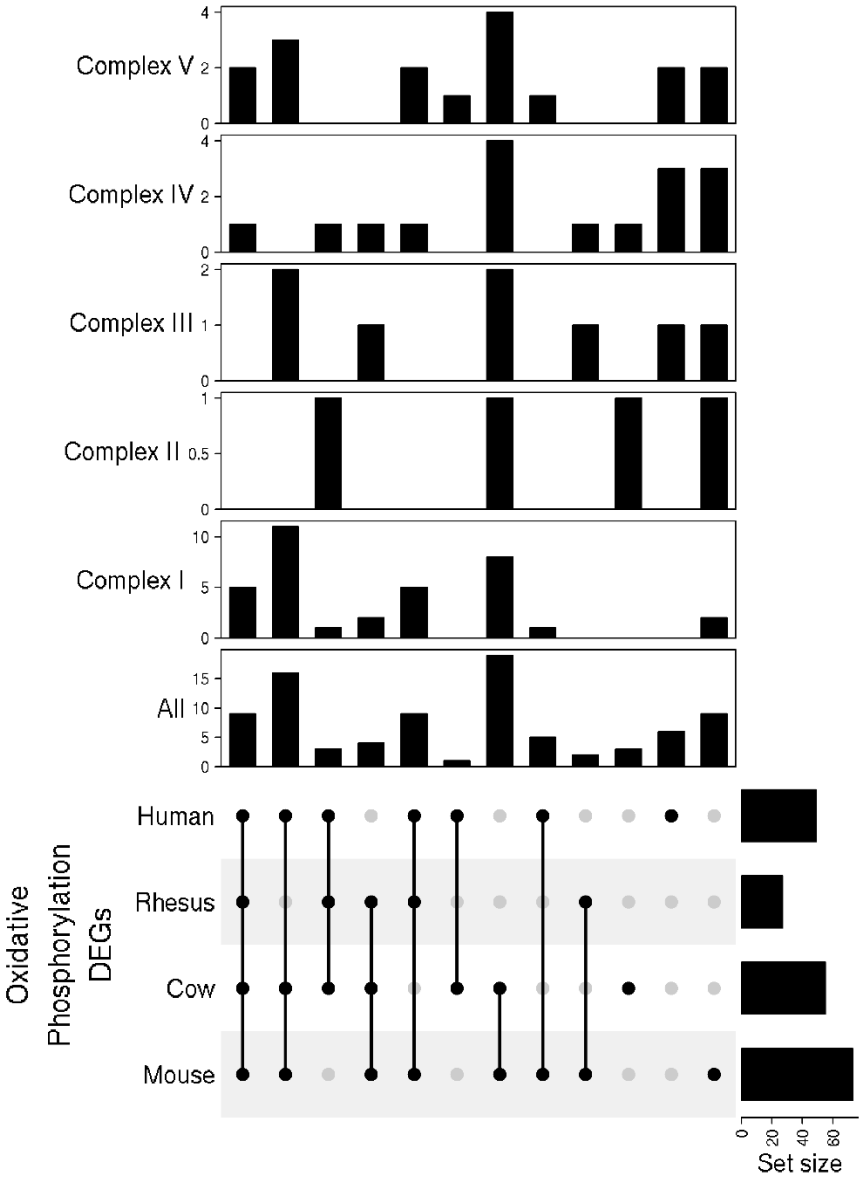

Figure 2.6 - Overlap of species regulation of differentially expressed genes (DEGs) in the oxidative phosphorylation pathway.
UpSet plot depicting the overlap of highly degraded DEGs from each species that were found in the oxidative phosphorylation pathway. Figure depicts the overlaps for the entire pathway and DEG membership by mitochondrial complexes.

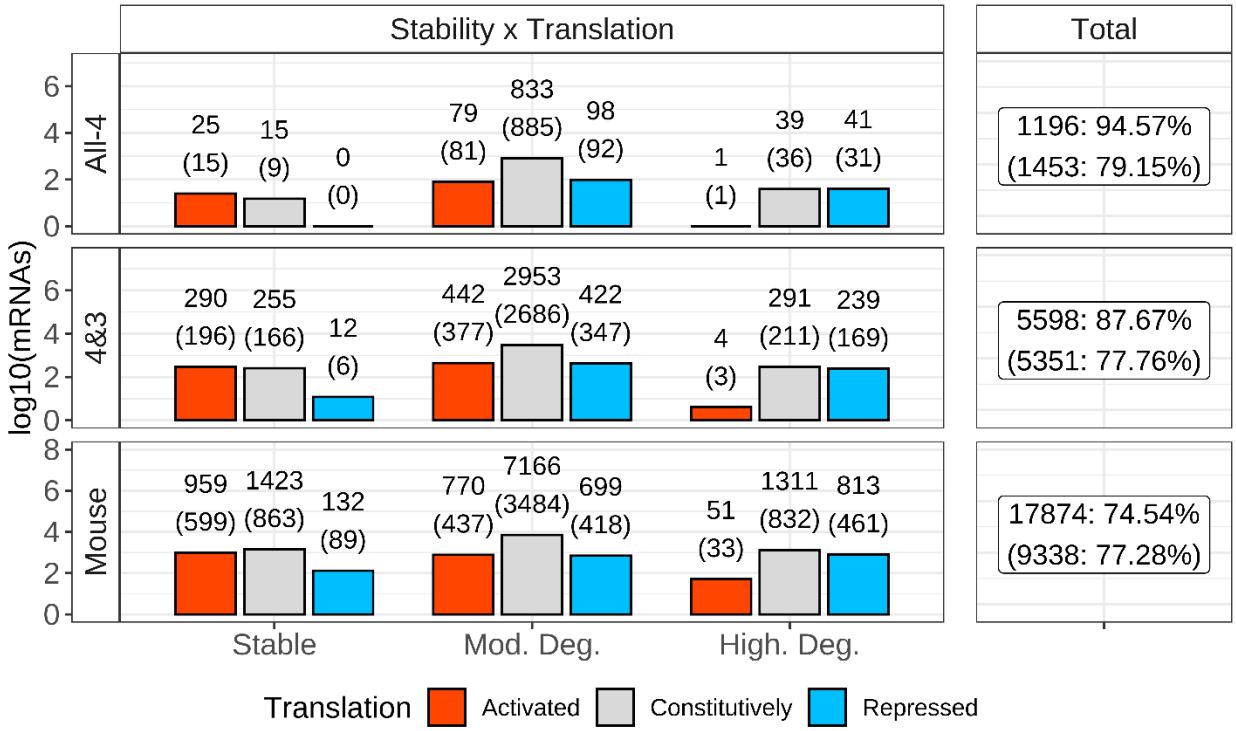

Figure 2.7 - Connecting stability and translational classification.
For the three stability classifications (stable, moderately degraded, and highly degraded), mRNAs were intersected with translational classifications (activated, constitutively, and repressed). The x-axis represents the three stability classes, and the y-axis the number of mRNAs identified. The left facet of the figure is split into three rows, representing gene list origin: All-4, 4&3, and whole species (WS) mouse. Labels above bars denote number of genes from the full method on top, with the number generated from the homology method within parenthesis. The summation of the total number of identified genes, per gene list, are shown in the right facet. Coloring denotes translational classifications: red = activated, gray = constitutively, and blue = repressed.

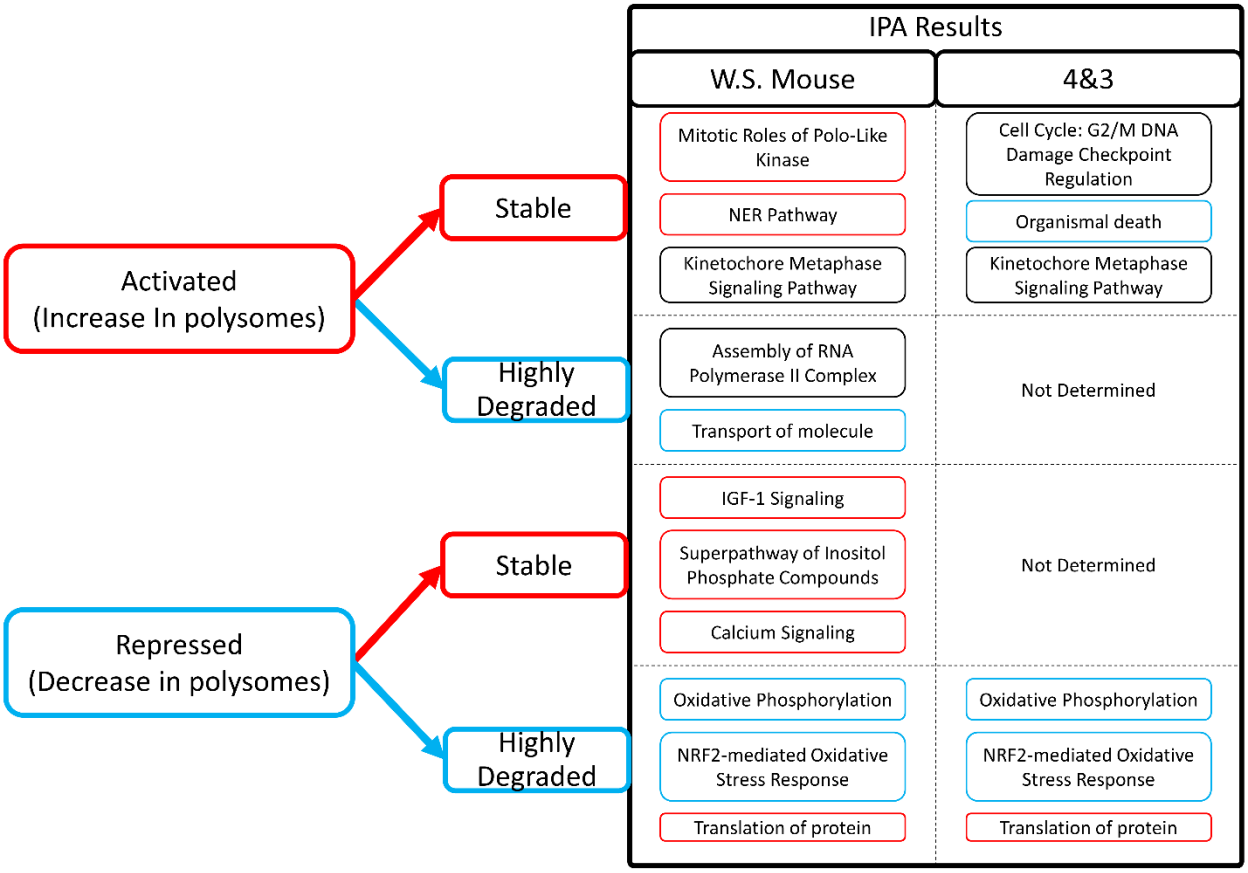

Figure 2.8 - Key Ingenuity Pathway Analysis (IPA) features of translation-stability classified groups.

Key IPA features for the stable and highly degraded mRNAs subdivided into activated or repressed translational categories. IPA results are presented for two submitted gene lists: whole species (WS) mouse and the 4&3 gene group. Exterior textbox coloring denotes the predicted z-score for each listed Canonical Pathways (CP)/Biological Functions (BF). Red = activated, Blue = inhibited. IPA entries are grouped by classified translation-stability groups.

REFERENCES

REFERENCES

1. Berg DK, Smith CS, Pearton DJ, Wells DN, Broadhurst R, Donnison M, Pfeffer PL. Trophectoderm lineage determination in cattle. Dev Cell 20: 244–255, 2011. doi:10.1016/j.devcel.2011. 01.003.

2. Daigneault BW, Rajput S, Smith GW, Ross PJ. Embryonic POU5F1 is required for expanded bovine blastocyst formation. Sci Rep 8: 7753,2018.doi:10.1038/s41598-018-25964-x.

3. Ozawa M, Sakatani M, Yao J, Shanker S, Yu F, Yamashita R, Wakabayashi S, Nakai K, Dobbs KB, Sudano MJ, Farmerie WG, Hansen PJ. Global gene expression of the inner cell mass and trophectoderm of the bovine blastocyst. BMC Dev Biol 12: 33, 2012. doi:10.1186/1471-213X-12-33.

4. Rossant J. Developmental biology: a mouse is not a cow. Nature 471:457–458, 2011. doi:10.1038/471457a.

5. Menezo YJ, Herubel F. Mouse and bovine models for human IVF. Reprod Biomed Online 4: 170–175, 2002. doi:10.1016/s1472-6483(10) 61936-0.

6. Assidi M, Montag M, Van der Ven K, Sirard MA. Biomarkers of human oocyte developmental competence expressed in cumulus cells before ICSI: a preliminary study. J Assist Reprod Genet 28: 173–188,2011. doi:10.1007/s10815-010-9491-7.

7. Cagnone G, SirardM-A.The impact of exposure to serum lipids during in vitro culture on the transcriptome of bovine blastocysts. Theriogenology 81: 712–722.e1-3, 2014. doi:10.1016/j.theriogenology. 2013.12.005.

8. Chitwood JL, Burruel VR, Halstead MM, Meyers SA, Ross PJ. Transcriptome profiling of individual rhesus macaque oocytes and preimplantation embryos. Biol Reprod 97: 353–364, 2017. doi:10. 1093/biolre/iox114.

9. Chu T, Dufort I, Sirard MA. Effect of ovarian stimulation on oocyte gene expression in cattle. Theriogenology 77: 1928–1938, 2012. doi:10.1016/j.theriogenology.2012.01.015.

10. Khan DR, Landry DA, Fournier E, Vigneault C, Blondin P, Sirard MA. Transcriptome meta-analysis of three follicular compartments and its correlation with ovarian follicle maturity and oocyte develop-mental competence in cows. Physiol Genomics 48: 633–643, 2016. doi:10.1152/physiolgenomics.00050.2016.

11. Labrecque R, Sirard MA. The study of mammalian oocyte competence by transcriptome analysis: progress and challenges. Mol Hum Reprod20:103–116,2014. doi:10.1093/molehr/gat082.

12. Reyes JM, Silva E, Chitwood JL, Schoolcraft WB, Krisher RL, Ross PJ. Differing molecular response of young and advanced maternal age human oocytes to IVM. Hum Reprod 32: 2199–2208, 2017. doi:10.1093/humrep/dex284.

13. Ruebel ML, Schall PZ, Midic U, Vincent KA, Goheen B, VandeVoortCA, Latham KE. Transcriptome analysis of rhesus monkey failed-to-mature oocytes: deficiencies in transcriptional regulation and cytoplasmic maturation of the oocyte mRNA population. MolHumReprod24:478–494,2018.doi:10.1093/molehr/gay032.

14. Schall PZ, Ruebel ML, Midic U, VandeVoort CA, Latham KE. Temporal patterns of gene regulation and upstream regulators contributing to major developmental transitions during Rhesus macaque preimplantation development. Mol Hum Reprod 25: 111–123, 2019. doi:10.1093/molehr/gaz001.

15. Severance AL, Midic U, Latham KE. Genotypic divergence in mouse oocyte transcriptomes: possible pathways to hybrid vigor impacting fertility and embryogenesis. Physiol Genomics 52: 96–99, 2020. doi:10.1152/physiolgenomics.00078.2019.

16. Sylvestre EL, Robert C, Pennetier S, Labrecque R, Gilbert I, Dufort I, Leveille MC, Sirard MA. Evolutionary conservation of the oocyte transcriptome among vertebrates and its implications for under-standing human reproductive function. Mol Hum Reprod 19: 369–379,2013.doi:10.1093/molehr/gat006.

17. Wang X, Liu D, He D, Suo S, Xia X, He X, Han JJ, Zheng P. Transcriptome analyses of rhesus monkey preimplantation embryos reveal a reduced capacity for DNA double-strand break repair in primate oocytes and early embryos. Genome Res 27: 567–579, 2017. doi:10.1101/gr.198044.115.

18. Huch S, Nissan T. Interrelations between translation and general mRNA degradation in yeast. Wiley Interdiscip Rev RNA 5: 747–763, 2014. doi:10.1002/wrna.1244.

19. Shyu AB, Wilkinson MF, van Hoof A. Messenger RNA regulation: to translate or to degrade. EMBO J 27: 471–481, 2008. doi:10.1038/sj. emboj.7601977.

20. Chalabi Hagkarim N, Grand RJ. The regulatory properties of the Ccr4-notcomplex. Cells9:2379, 2020.doi:10.3390/cells9112379.

21. Clarke HJ. Post-transcriptional control of gene expression during mouse oogenesis. Results Probl Cell Differ 55: 1–21, 2012. doi:10.1007/978-3-642-30406-4_1.

22. Luong XG, Daldello EM, Rajkovic G, Yang CR, Conti M. Genome-wide analysis reveals a switch in the translational program upon oocyte meiotic resumption. Nucleic Acids Res 48: 3257–3276, 2020.doi:10.1093/nar/gkaa010.

23. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34: i884–i890, 2018. doi:10.1093/ bioinformatics/bty560.

24. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34: 525–527, 2016. doi:10.1038/nbt.3519.

25. Hendrickson PG, Dorais JA, Grow EJ, Whiddon JL, Lim JW, Wike CL, Weaver BD, Pflueger C, Emery BR, Wilcox AL, Nix DA, Peterson CM, Tapscott SJ, Carrell DT, Cairns BR. Conserved roles of mouse DUXandhumanDUX4inactivatingcleavage-stagegenes and MERVL/HERVL retrotransposons. Nat Genet 49: 925–934, 2017. doi:10.1038/ng.3844.

26. Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc Natl Acad Sci USA 111: 4139–4144, 2014. doi:10. 1073/pnas.1321569111.

27. Reyes JM, Chitwood JL, Ross PJ. RNA-Seq profiling of single bovine oocyte transcript abundance and its modulation by cytoplasmic polyadenylation. Mol Reprod Dev 82: 103–114, 2015. doi:10.1002/ mrd.22445.

28. Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, Kuzman M, Fulka H, Cernohorska M, Urbanova J, Svobodova E, Ma J, Suzuki Y, Aoki F, Schultz RM, Vlahovicek K, Svoboda P. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. Genome Res 27: 1384–1394,2017.doi:10.1101/gr.216150.116.

29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550,2014. doi:10.1186/s13059-014-0550-8.

30. Chorostecki U, Molina M, Pryszcz LP, Gabaldon T. MetaPhOrs 2.0: integrative, phylogeny-based inference of orthology and paralogy across the tree of life. Nucleic Acids Res 48: W553–W557, 2020. doi:10.1093/nar/gkaa282.

31. Rau A, Marot G, Jaffrezic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics 15: 91, 2014. doi:10.1186/1471-2105-15-91.

32. Song Y, Milon B, Ott S, Zhao X, Sadzewicz L, Shetty A, Boger ET, Tallon LJ, Morell RJ, Mahurkar A, Hertzano R. A comparative analysis of library prep approaches for sequencing low input translatome samples. BMC Genomics 19: 696, 2018. doi:10.1186/s12864-018-5066-2.

33. Kramer A, Green J, Pollard J Jr, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics 30: 523– 530,2014.doi:10.1093/bioinformatics/btt703.

34. Verdin E, Hirschey MD, Finley LW, Haigis MC. Sirtuin regulation of mitochondria: energy production, apoptosis, and signaling. Trends BiochemSci35:669–675,2010. doi:10.1016/j.tibs.2010.07.003.

35. Ruebel ML, Latham KE. Listening to mother: long-term maternal effects in mammalian development. Mol Reprod Dev 87: 399–408, 2020. doi:10.1002/mrd.23336.

36. Ruebel ML, Zambelli F, Schall PZ, Barragan M, VandeVoort CA, Vassena R, Latham KE. Shared aspects of mRNA expression associated with oocyte maturation failure in humans and rhesus monkeys indicating compromised oocyte quality. Physiol Genomics 53: 137–149,2021.doi:10.1152/physiolgenomics.00155.2020.

37. Potireddy S, Vassena R, Patel BG, Latham KE. Analysis of polysomal mRNA populations of mouse oocytes and zygotes: dynamic changes in maternal mRNA utilization and function. Dev Biol 298: 155–166,2006. doi:10.1016/j.ydbio.2006.06.024.

38. Midic U, Vincent KA, Wang K, Lokken A, Severance AL, Ralston A, Knott JG, Latham KE. Novel key roles for structural maintenance of chromosome flexible domain containing 1 (Smchd1) during preim-plantation mouse development. Mol Reprod Dev 85: 635–648, 2018. doi:10.1002/mrd.23001.

39. Ruebel ML, Vincent KA, Schall PZ, Wang K, Latham KE. SMCHD1 terminates the first embryonic genome activation event in mouse two-cell embryos and contributes to a transcriptionally repressive state. Am J Physiol Cell Physiol 317: C655–C664, 2019. doi:10.1152/ ajpcell.00116.2019.

40. Menezo Y, Dale B, Cohen M. DNA damage and repair in human oocytes and embryos: a review. Zygote 18: 357–365, 2010. doi:10.1017/s0967199410000286.

41. Adiga SK, Toyoshima M, Shiraishi K, Shimura T, Takeda J, Taga M, Nagai H, Kumar P, Niwa O. p21 provides stage specific DNA dam-age control to preimplantation embryos. Oncogene 26: 6141–6149, 2007. doi:10.1038/sj.onc.1210444.

42. Song Y, Li Z, Wang B, Xiao J, Wang X, Huang J. Phospho-Cdc25 correlates with activating G2/M checkpoint in mouse zygotes fertilized with hydrogen peroxide-treated mouse sperm. Mol Cell Biochem396:41–48,2014.doi:10.1007/s11010-014-2140-1.

43. Jurisicova A, Latham KE, Casper RF, Casper RF, Varmuza SL. Expression and regulation of genes associated with cell death during murine preimplantation embryo development. Mol Reprod Dev 51: 243–253, 1998. doi:10.1002/(SICI)1098-2795(199811)51:3<243:: AID-MRD3>3.0.CO;2-P.

44. Biase FH. Oocyte developmental competence: insights from cross-species differential gene expression and human oocyte-specific functional gene networks. OMICS 21: 156–168, 2017. doi:10.1089/ omi.2016.0177.

CHAPTER 3.

REGULATION OF MRNA STABILITY VIA THE 3' UTR DURING OOCYTE MATURATION

3.1 Abstract

During oocyte maturation, the transition from immature (germinal versicle or GV) to mature (metaphase II or MII), thousands of transcripts change in relative abundance. However, during this process, the cell is transcriptionally inactive and the resulting change in relative abundance chiefly comes down to posttranscriptional mechanisms. Specifically of note are RNA binding proteins (RBPs) that bind to the 3' untranslated region and control storage, translation, and stability of transcripts. Utilizing publicly available oocyte maturation RNAseq data, across five mammalian species (human, rhesus monkey, cow, pig, and mouse), and applying a machine learning feature selection and regression algorithm, RBPs were identified that are associated with mRNA stability. To further highlight RBPs with similar motif and overall predictive impact on stability, a clustering algorithm was applied to the RBP motifs. This resulted in a group of RBPs binding to AU rich motifs, either shared across species, a subset of species, or specific, impacting mRNA stability during oocyte maturation.

3.2 Introduction

Oocyte maturation results in the changes of relative abundance of thousands of maternal mRNAs that undergo a complex pattern of polyadenylation, degradation, and translation. These transcripts are essential for the oocyte to meet the physiological demands and are required for later events such as genome activation. The time from the breakdown of the germinal vesicle stretching to embryonic genome activation encompasses a range of physiological and

developmental demands that must be met, and that require the production of proteins at the necessary times (1,2). Posttranscriptional regulation is an essential mode for regulating protein production and expression, chiefly conducted via *cis*-regulatory elements within the 3' untranslated region (3' UTR). These 3' UTR elements are acted upon by *trans*-acting factors, such as RNA binding proteins (RBPs).

Current research into RBPs and their respective roles in modulating maternal mRNAs is limited, and this deficiency is increased when comparing across species. This is a multi-layered problem. Evolutionary divergence can occur at the mRNA sequence level impacting binding sequence sites, secondary structure, and ultimately altering the binding affinity of RBPs. As noted in chapter 1, a meta-analysis of mRNA regulation in four mammalian species (3), the number of shared similarly regulated maternal mRNAs during oocyte maturation is highly limited. The second layer can occur at the protein level, resulting in protein conformation changes and motif recognition sites. To date, more than 1500 RBPs have been cataloged in human cells (4). The binding sites, localization, and function of >350 RBPs have been reviewed, finding 28% are involved in splicing, 46% with more than one function, and 23% lacking an annotated function (5). Therefore, the difficulty in extrapolating species conservation of RBP functions and their associated targets, as they relate to maternal mRNA regulation, is compounded.

Acknowledging the limitations in high-resolution data regarding RBP binding and regulation of maternal mRNAs, an unbiased cross-species computational analysis was undertaken. Whole transcriptome changes during oocyte maturation across five mammalian species (human, rhesus monkey, cow, pig, and mouse) was calculated and mRNAs were binned into statistically different stability categories (stable and highly-degraded). Public databases

containing RBP motif sequences were retrieved, and their binding sites were identified via computational predictive software amongst the two stability groups. This was followed by overlaying RBP mRNA and proteomic expression, and their annotated functions. A total of 120 RBPs were interrogated in this analysis, 105 of which had identified expression in at least one species. The RBP binding site data was then subjected to machine learning feature selection and regression algorithms, identifying high-value RBPs. By further clustering the high-value RBP motif sequences, a core group of RBPs targeting AU rich motifs were identified. The result was a novel collection of shared RBPs identified as likely to regulate mRNAs in multiple species. This analysis thus provides valuable new insight into the mechanisms controlling mRNA utilization and how this affects oocyte and embryo biology.

3.3 Materials and Methods

3.3.1 Study Selection

The datasets used in this study were based on those interrogated previously (3) with the inclusion of one additional study (6). The parameters for the inclusion of RNA sequencing datasets were to contain both GV and MII stage oocytes with at least three replicates per stage. This resulted in a collection of datasets for five mammalian species (human, mouse, cow, rhesus monkey, and pig).

3.3.2 Sample Processing

Pre-processing of the raw sequencing data, quantification of transcripts, and calculation of differentially expressed genes, used the same methods as before (3). The processed studies included: human (PRJNA377237 and PRJNA293908), rhesus (PRJNA343030 and PRJNA448148/PRJNA448150), cow (PRJNA261946 and PRJNA228235), and mouse

(PRJNA342001 and PRJNA464431). Briefly, initial QC metrics were found using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc; RRID:SCR_014583), trimming was performed using Fastp (RRID:SCR_016962, v0.20.0) (7), transcript quantification was performed with Kallisto (RRID:SCR_016582, v0.44.0) (8), and DEG calculations were performed with DESeq2 (RRID:SCR_015687, v1.30.1) (9). For the pig study (6), all software utilized standard settings, excluding trimming: initial QC metrics detected abnormal nucleotide distribution in the first 9 basepairs, and reads were subsequently hard trimmed.

While Kallisto was used for the quantification of transcripts, physical alignments of reads were required for the exploration of the 3' UTR. This was conducted using STAR (v2.7.3a, RRID:SCR_004463) (10) (Figure 3.1, Phase 1). The genomes (GRCh38, Mmul_10, ARS-UCD1.20, GRCm38, Sscrofa11) were retrieved from Ensembl (build 102, RRID:SCR_002344) (11) and were indexed with STAR, the parameter *sjdbOverhang* was the only variable, matching the requirement based on the sequencing length per each study (Supplemental Table S1). STAR SAM outputs were converted and sorted to BAM with SAMtools (v1.10, RRID:SCR_002105) (12).

3.3.3 Statistical Difference in Abundance & Stability Classification

As noted previously (3), the changes in mRNA abundances between the GV and MII stages are a matter of modulating relative stability during a time of transcriptional silence. Therefore, we opted to drop the terms "up-regulated" or "down-regulated", instead using classifications based on stability. In short, transcript quantification was conducted by Kallisto, and statistical significance (FDR<0.05) determined by comparing GV and MII relative mRNA abundances for each individual study via DESeq2. Studies within species were integrated via metaRNAseq (v1.0.3, RRID:SCR_002174) (13), deriving a unified p-value via a Fisher

combination method. This yielded a single list of genes and their associated statistical classification per species. These lists were then classified by stability: GV>MII termed highly-degraded, GV~MII moderately-degraded, and GV<MII stable. It is important to note that this approach avoids artificial suppression of DEG numbers when using multiple datasets for each species, as could otherwise occur by looking for simple intersects between individual study DEG lists.

3.3.4 3’ UTR Identification and Extraction

To accurately explore the impact of mRNA 3’ UTRs on their stability during oocyte maturation, we first determined the sequence coordinates delineating the 3’ UTRs based on the acquired sequencing data (Figure 3.1, Phase 1). To do this, we used APAtrap (14). Preparation of data for APAtrap required the quantification of genome coverage. From the sorted and indexed BAM files, genome coverage was calculated with BEDTools (v2.29.2, RRID:SCR_006646) (15). From the genome coverage files, the APAtrap function *identifyDistal3UTR,* using standard settings was applied, identifying the 3’ UTR coordinates.

The coordinates of all 3’ UTRs identified from APAtrap were imported into R (v4.1.0) for further analysis. For each species, the fasta sequences, DNA coded, of all 3’ UTRs were extracted with the package BSgenome (v1.58). As a comparison, the annotated 3’ UTR coordinates and fasta sequence were retrieved using the biomaRt (v2.46.3) package for each species.

3.3.5 Identification of RNA Binding Protein Motifs within the 3’ UTR

Identification of RBA binding protein motifs within the 3’ UTR were limited to the statistically significant different stability groups: stable and highly-degraded. This was conducted

using the FIMO (16) tool from the MEME suite (RRID:SCR_001783) (17) (Figure 3.1, Phase 2). The FIMO tool (Find Individual Motif Occurrences), identifies all occurrences of motifs within supplied sequences.

RBP motifs were retrieved from the CISBP database for each of the input species (18), and the CISBP IDs were mapped to protein symbols. For situations where motifs for a given RBP was not present for a given species, the human equivalent motif was utilized. The AtTRACT database (19) was cross-referenced for additional RBPs not present in the CISBP database and were previously found to be involved in oocytes and/or mRNA stability. RBP motifs were converted to MEME minimal format using custom scripts. These concentrated RBP lists were used to find all occurrences of RBP motifs within species, with a E-value threshold of 1e-4, for both stable and highly degraded mRNA classes. The location of all RBP motifs for each species were reformatted to tabulate the total occurrences of each RBP motif for each gene. For each species, the resultant matrices for stable and degraded classes, consisted of three columns: 1) gene, 2) RBP, and 3) number of sites. This long-data format was then transformed to wide format: rows consisting of genes, columns consisting of RBPs, and the interior matrix data representing number of RBP sites per gene. An additional column was appended, consisting of the maximal Log2Foldchange for each gene from the input studies. This resulted in a single matrix for each species containing the frequency of motifs for all stable and highly-degraded mRNAs and their respective magnitudes of change.

For the selected RBPs, publicly available data were retrieved to identify their respective expression from three different sources: RNA transcriptome data, protein/proteomics data, and mRNA presence on polysomes. RNA expression was mapped from the input studies, containing GV and MII samples. For protein expression, human data was extracted from (20), mouse from

(21) and (22), and pig from (23). For mRNAs bound to polysomes, cow data was extracted from (24), and mouse from (25).

3.3.6 Statistical Analysis and Machine Learning on 3' UTR Motifs

To determine which RBPs are most predictive of mRNA stabilization or degradation during oocyte maturation, we utilized a machine learning approach to identify 3'UTR motifs that are highly predictive of mRNA degradation or stability during maturation (Figure 3.1, Phase 3). Machine learning (ML) is a wide-ranging computational science field consisting of statistics, information theory, and artificial intelligence, able to handle large amounts of data. Current techniques include clustering, feature selection, classification, and regression models (26). Specific ML practices relevant to this present study, are feature selection and regression models.

Feature selection is a step in ML applications that aims to reduce the number of input variables, arising at the minimally optimum number, increasing the accuracy of ML algorithms (27). For this analysis, features are represented as the RBPs, and the feature selection algorithm utilized is Boruta. The Boruta feature selection is built around the random forest algorithm and employs a method that does not compete features against other features, rather a randomized 'shadow' version of themselves. A threshold is derived using the maximal importance score from the 'shadow' features, and only those non-shadow features exceeding the threshold are deemed important. The Boruta function thereby trims the input features (RBPs), leaving those that exceed the defined maximal importance score of their respective shadow features.

With the post-feature selected dataset, the next step is to apply some form of statistical modeling, e.g., classification or regression. Utilizing the values of Log2FoldChange, I opted to utilize the xgboost regression methodology (28).

To that end, the RBP motif matrices generated for each species were analyzed in python (v3.8.5). mRNAs with a fold-changes below 1.0 (log2foldchange=0.5849625) were excluded. Boruta feature selection was conducted using BorutaShap package (1.0.16) (29). The BorutaShap function was called as follows: *BorutaShap(model=XGBRegressor(),importance_measure='shap',classification=False,percentile=75)* with the xgboost (v1.4.2) regression and the XGBRegressor parameters set as: *XGBRegressor(learning_rate=0.02, subsample=0.2, colsample_bytree=0.5, n_estimators=5000*).

Shapley Additive Explanations (SHAP) (30) values were estimated from the regression output from the python package Shap (v0.39) (31), using the TreeExplainer (32) function. SHAP values, developed from a game theory basis, quantify the contribution of each feature (RBP) on observations (genes), and the overall impact on model predictions (Log2Foldchange, and by proxy mRNA stability). SHAP values are not calculated on a per feature (RBP) basis, rather they are calculated on a per observation (gene) basis.

In summary, the machine learning portion of this analysis takes the number of RBP or miRNA motif sites per gene, selects relevant features (RBP), applies a regression analysis upon the Log2FoldChange of genes, deriving a SHAP scores. This allows for the interrogation of RBPs and visually comprehend the overall predictive correlation between motif frequency and mRNA stability. The output from this machine learning analysis were exported to R for plotting.

From the list of selected RBP, their respective motifs were analyzed with the software package GimmieMotifs (33), aiming to cluster similar motif sequences, without using the complementary motif sequences.

3.3.7 Ingenuity Pathway Analysis

RBP mRNA targets were analyzed though the use of Ingenuity Pathway Analysis (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis), focusing on Canonical Pathway (CP) and Diseases and Functions (DF) analysis tools (IPA database v. 11/2021). The IPA software is a suite of tools that can calculate the overlap between submitted gene lists and their database containing pathways and functions (disease and cancer related functions were excluded from the analysis). Similar to a typical gene set enrichment analysis, IPA calculates which pathways are enriched from submitted gene lists (p-value, significance set at 0.05) and a predictive measure of pathway/function activity (activated or inhibited, as a z-score, significance set at $z>|1.96|$).

3.3.8 Generation of Figures

Figures were produced in R (v4.1.0) with the R package ggplot2 (v3.3.3, RRID:SCR_014601). SHAP plots were generated using ggplot2, motif clusters, expression, and IPA results with ComplexHeatmap (v2.6.2, RRID:SCR_017270) (34).

3.4 Results

All the essential genetic material needs to be present in the oocyte for proper fertilization and embryonic growth, along with a sufficiently rich mRNA endowment. During oocyte maturation, the chief mechanism regulating protein expression is the modulation of transcript stability and translation. As previously stated, stability is primarily regulated via mechanisms involving the 3’ UTR. Layered upon those factors, are RNAs and miRNAs that bind the 3’ UTR, enhancing stability or the increased rate of degradation. Therefore, elucidating the primary RBPs that are essential across mammalian species can shed further light in those essential factors that

lead to proper and successful embryonic development. To that end, we sought out to compare oocyte mRNA 3' UTR sequences and features to derive this information.

3.4.1 Correlation of 3' UTR Length and Stability

For the five included species (human, rhesus, mouse, cow, and pig) we successfully identified and extracted the 3' UTR sequences based on oocyte sequencing data. Application of APAtrap allowed for the identification of a number of 3' UTRs that are not annotated in the Ensembl genome build, a mean of 1743 (412-2550) genes per species. When trifurcating the identified transcripts by stability classification (highly-degraded, moderately-degraded, and stable) we identified a trend: on average, longer 3' UTR transcripts are more stable than their degraded counterparts (Figure 3.2).

3.4.2 RBPs regulation MmRNA Stability vis 3' UTR binding

With the understanding that changes in transcript abundance during oocyte maturation is not a result of new transcript production, rather a stability modulation, ascertaining which factors impact this phenomenon are essential to further the understanding of proper mammalian oocyte maturation. We therefore applied a combination of Boruta feature selection with xgboost regression to predict which RBP motifs have the greatest predictive value upon mRNA stability. When applying the clustering algorithm on the selected RBPs, the most prominent cluster of RBPs were those binding to AU rich motifs, including 8 identified proteins (KHDRBS1, CPEB2, PABPC5, SART3, U2AF2, KHDRBS2, PTBP1/3, and CPEB4) (Figure 3.3). None of these RBPs were populated by the algorithm across all five species, however, by using this grouping method, at least one RBP was identified in every species.

KHDRBS1 was identified among human, cow, pig, and mouse. For human and cow, the majority of the high-feature value targets were populated among the stable mRNAs, pig among the highly-degraded, and a heterogenous mixture for mouse. CPEB2 (human, pig, and mouse) showed stable mRNAs with high-feature value targets for human and mouse, with the inverse for pig. While rhesus did not have CPEB2 selected, CPEB4 was a selected RBP and resulted in a preferential targeting of stable mRNAs. With a similarly structured poly(U) motif, PTBP1/3 was selected for cow, again signaling high-feature value targets among stable mRNAs. U2AF2 selection was limited to the two ungulate species (cow and pig), and both species highlighted stable mRNAs with high feature values. Those results identifying high-feature value targets among the stable mRNAs, indicate those RBPs could be potentially binding the 3' UTR and increasing mRNA stability.

3.4.3 Stable mRNA targets poly(U) RBPs CPEB2, CPEB4, and U2AF2

To explore shared targets and functions populated by stable mRNAs targeted by RBPs, three RBPs with poly(U) binding motifs (CPEB2, CPEB4, and U2AF2) were selected for further analysis. Mouse and human data were limited to CPEB2, Rhesus for CBPE4, and Cow and Pig for U2AF2. These were selected due to the prominent trend of stable mRNAs exhibiting high-feature values (Figure 3.3). From these stable mRNAs, the population was further limited to those transcripts with a feature value greater than 0.33. Across the three RBPs and five species, a total of 370 genes were identified, none were shared by all species, and only 10 overlapping in more than one species. The total number of targets per species were 135, 35, 84, 56, and 71, for human, rhesus, cow, pig, and mouse, respectively.

The stable mRNA targets, with feature values greater than 0.33, of the RBPs for each species (human and mouse, CPEB2; rhesus, CPEB4; cow and pig, U2AF2) were submitted to

IPA for pathway and function results. These results included a few overlapping entries across a subset of species. The canonical pathways PPRa/RXRa activation was found for human and cow, ERK/MAPK signaling in rhesus and pig, and a few inositol biosynthesis pathways for rhesus and cow. When comparing biological functions, the entry organization of cytoplasm was significant for cow, pig, and mouse. Expression of RNA was populated amongst human, rhesus, and cow, and several cell cycle functions for cow and pig.

3.5 Discussion

Ascertaining a further understanding regarding the complex interplay between RBPs and mRNA stability is an essential step in elucidating which RBP factors are essential for oocyte maturation. By applying a novel bioinformatics pipeline and machine learning feature selection and regression, a core group of RBPs targeting AU rich motifs were identified across five mammalian specie (human, rhesus monkey, cow, pig, and mouse). The machine learning results show that there is a heterogenous targeting of both stable and highly-degraded mRNAs by these RBPs. This echoes the trend seen above with cross-species variation at the individual gene level but conservation in key processes, which applies here both limited numbers of shared RBPs, limited numbers of shared target mRNAs, but some conserved cellular functions, nonetheless.

Independent of transcriptional mechanisms, posttranscriptional regulation is an essential facet of an oocyte's ability to modulate protein expression. These mechanisms can be both macro and micro, at the transcript level via 3' UTR regulatory elements and chiefly through the interaction of RBPs. There are several well characterized protein domains that interact with the 3' UTR, primarily amongst them include the RNA recognition motif (RRM) and the K homology (KH). The cytoplasmic polyadenylation element (CPE) is one of the most well characterized elements within the 3' UTR. In mice, translationally activated mRNAs have been

found to contain two or more CPEs, whereas downregulation of translation was found for those mRNAs without CPEs (39). Another well characterized 3' UTR element is AU-rich elements (AREs), which, when bound by RBPs can result in either stabilization or destabilization of transcripts (36).

KHDRBS1 (KH RNA Binding Domain Containing, Signal Transduction Associated 1), also known as Sam68, is a member of the KH domain containing protein family. It has been found to be involved in the posttranscriptional mRNA metabolism, RNA splicing, and the translation regulation of maternal mRNAs, and is released from the cytoplasm upon the resumption of meiosis (37). In one-cell embryos, KHDRBS1 accumulates to the nucleus, and its inhibition via cycloheximide or puromycin resulted in the localization at cytoplasmic granules. These granules were found to be populated with other proteins essential for the initiation of mRNA translation (37). Additionally, mutations within KHDRBS1 have been found to result in aberrations in alternative splicing, resulting the decreased fertility (38).

CPEB2 (Cytoplasmic Polyadenylation Element Binding Protein 2) and CPEB4 (Cytoplasmic Polyadenylation Element Binding Protein 2) are proteins that have a high sequence similarity to CPEB, a protein that regulated cytoplasmic polyadenylation of mRNAs. CPEB2 has been found to be essential for proper meiotic maturation in the porcine model and binds to CPE containing transcripts with homopolymeric poly(U) RNAs (39). In Xenopus oocyte, CPEB4 forms a positive feedback loop with CPEB1, which causes the metaphase I to metaphase 2 transition (40). Interestingly, as CPEB2 was selected for all species excluding rhesus, while CPEB4 was selected for rhesus, indicating a possible species-specific adaptation to regulating mRNA stability during oocyte maturation.

The application of this unbiased novel bioinformatics pipeline has allowed for the identification of a core group of AU-rich binding RBPs impacting mRNA stability during oocyte maturation. Further exploration of this rich dataset can identify conserved mRNA targets and prediction of the canonical pathways and functions these RBPs impact. This developed framework is highly applicable to a number of other developmental stages, cell models, and species. Its robustness can further the field in identifying key features that are both shared and species-specific.

3.6 Acknowledgements

A thank you to Grant Miller for his consultation on the machine learning applications.

APPENDIX

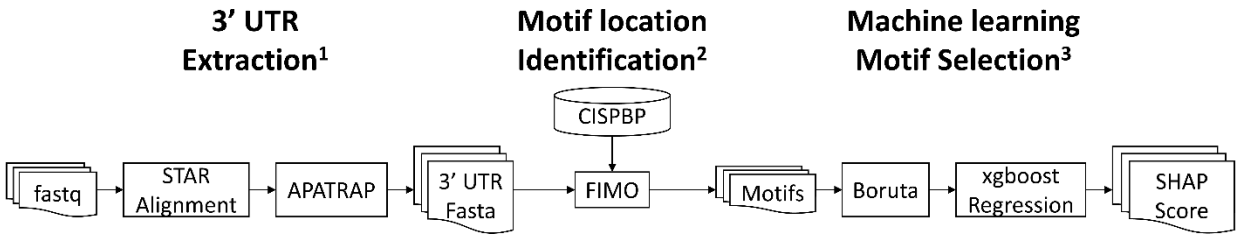

Figure 3.1 – Flowchart of Analysis
Simplified diagram depicting the flow of analysis used to identify miRNA and RBP binding in the 3' UTR. This pipeline was utilized for each study and species. Figure is divided into three sections: 1) 3' UTR extraction, 2) Motif identification, and 3) Motif selection. For section 1: fastq files from each study, within species, were aligned to their respective genomes using STAR, the 3' UTR was extracted using APATrap. Section 2: RBP motifs for each species were downloaded from CISBP and sites identified with FIMO. Section 3: the RBP binding sites were processed for feature selection with the Boruta algorithm and a regression applied with xgboost, with the output of SHAP importance score.

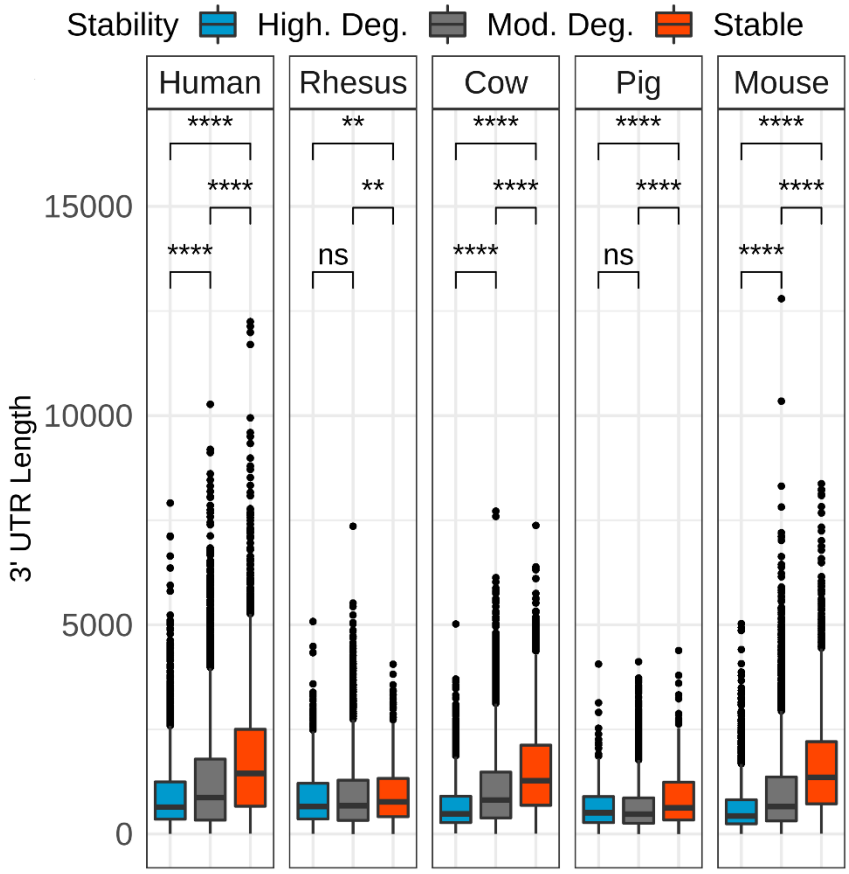

Figure 3.2 – 3' UTR Length versus mRNA Stability
Boxplot comparing mRNA stability classification versus 3' UTR length for five mammalian species during oocyte maturation. Figure consists of five vertical facets for each of the five species, y-axis denotes the length of the 3' UTR, color denotes mRNA stability (blue = highly degraded, grey = moderately degraded, and red = stable). Significance was calculated comparing log10 normalized 3' UTR lengths with a t-test. Significance key: p<=0.01 (**), p<=0.001 (***) and p<=0.0001 (****).

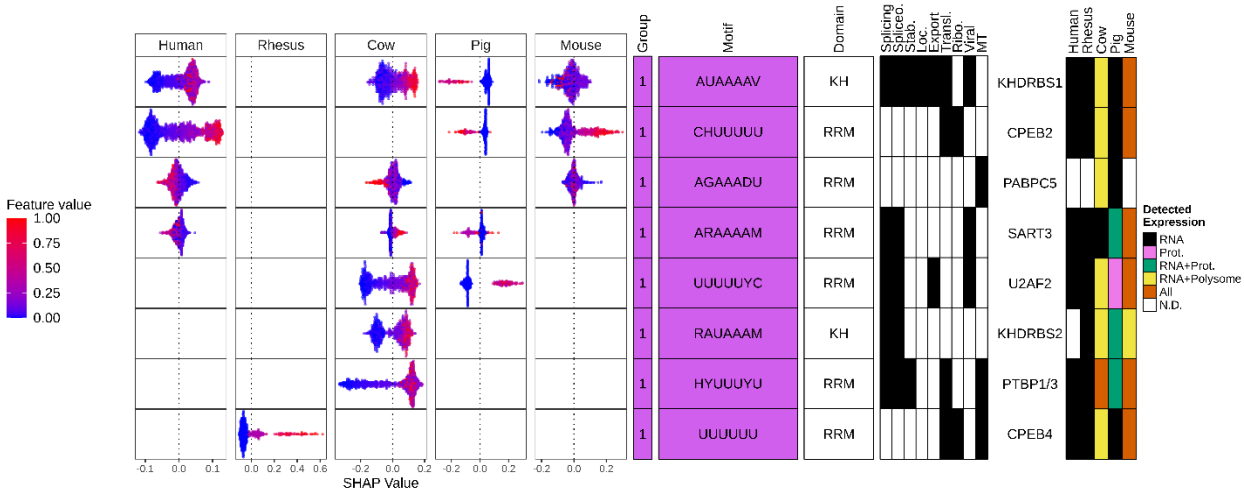

Figure 3.3 - Identification of RNA binding proteins and predictive output on mRNA stability

Figure depicting the selected RBPs among five mammalian species, their predictive SHAP scores on mRNA stability, motif group, IUPAC consensus sequence, domain group, functional categories, and detected expression. SHAP importance plot consists of five vertical facets for each species, x-axis denotes SHAP value (which represents the impact on the model), y-axis denotes each of the identified RBPs, and the color of each dot (an individual gene) denotes the feature value. The single vertical column titled “Group” denotes the identified clusters of RBP motif. For each RBP, the position weight matrix was converted to IUPAC consensus sequence. The domain group for each RBP was retrieved from UniProt, as well as their functional categories. RBP expression was identified from three different data sources: RNA (RNA-seq transcriptome), Prot. (protein expression), Polysome (presence on polysome from sequencing or microarray).

REFERENCES

REFERENCES

1. Latham KE, Garrels JI, Chang C, Solter D. Quantitative analysis of protein synthesis in mouse embryos. I. Extensive reprogramming at the one- and two-cell stages. Development. 1991 Aug;112(4):921-32. PMID: 1935701.

2. Li L, Baibakov B, Dean J. A subcortical maternal complex essential for preimplantation mouse embryogenesis. Dev Cell. 2008 Sep;15(3):416-425. doi: 10.1016/j.devcel.2008.07.010. PMID: 18804437; PMCID: PMC2597058.

3. Schall PZ, Latham KE. Essential shared and species-specific features of mammalian oocyte maturation-associated transcriptome changes impacting oocyte physiology. Am J Physiol Cell Physiol. 2021 Jul 1;321(1):C3-C16. doi: 10.1152/ajpcell.00105.2021. Epub 2021 Apr 21. PMID: 33881934; PMCID: PMC8321790.

4.Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014 Dec;15(12):829-45. doi: 10.1038/nrg3813. Epub 2014 Nov 4. PMID: 25365966.

5. Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Chen JY, Cody NAL, Dominguez D, Olson S, Sundararaman B, Zhan L, Bazile C, Bouvrette LPB, Bergalet J, Duff MO, Garcia KE, Gelboin-Burkhart C, Hochman M, Lambert NJ, Li H, McGurk MP, Nguyen TB, Palden T, Rabano I, Sathe S, Stanton R, Su A, Wang R, Yee BA, Zhou B, Louie AL, Aigner S, Fu XD, Lécuyer E, Burge CB, Graveley BR, Yeo GW. A large-scale binding and functional map of human RNA-binding proteins. Nature. 2020 Jul;583(7818):711-719. doi: 10.1038/s41586-020-2077-3. Epub 2020 Jul 29. Erratum in: Nature. 2021 Jan;589(7842):E5. PMID: 32728246; PMCID: PMC7410833.

6. Du ZQ, Liang H, Liu XM, Liu YH, Wang C, Yang CX. Single cell RNA-seq reveals genes vital to in vitro fertilized embryos and parthenotes in pigs. Sci Rep. 2021 Jul 13;11(1):14393. doi: 10.1038/s41598-021-93904-3. PMID: 34257377; PMCID: PMC8277874.

7. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884-i890. doi: 10.1093/bioinformatics/bty560. PMID: 30423086; PMCID: PMC6129281.

8. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 May;34(5):525-7. doi: 10.1038/nbt.3519. Epub 2016 Apr 4. Erratum in: Nat Biotechnol. 2016 Aug 9;34(8):888. PMID: 27043002.

9. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PMID: 25516281; PMCID: PMC4302049.

10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

11. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, El Houdaigui B, Fatima R, Gall A, Garcia Giron C, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Gonzalez Martinez J, Marugán JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, De Silva N, Flint B, Frankish A, Hunt SE, IIsley GR, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR, Flicek P. Ensembl 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D884-D891. doi: 10.1093/nar/gkaa942. PMID: 33137190; PMCID: PMC7778975.

12. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008. doi: 10.1093/gigascience/giab008. PMID: 33590861; PMCID: PMC7931819.

13. Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics. 2014 Mar 29;15:91. doi: 10.1186/1471-2105-15-91. PMID: 24678608; PMCID: PMC4021464.

14. Ye C, Long Y, Ji G, Li QQ, Wu X. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. Bioinformatics. 2018 Jun 1;34(11):1841-1849. doi: 10.1093/bioinformatics/bty029. PMID: 29360928.

15. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2. doi: 10.1093/bioinformatics/btq033. Epub 2010 Jan 28. PMID: 20110278; PMCID: PMC2832824.

16. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017-8. doi: 10.1093/bioinformatics/btr064. Epub 2011 Feb 16. PMID: 21330290; PMCID: PMC3065696.

17. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015 Jul 1;43(W1):W39-49. doi: 10.1093/nar/gkv416. Epub 2015 May 7. PMID: 25953851; PMCID: PMC4489269.

18. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M,

Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014 Sep 11;158(6):1431-1443. doi: 10.1016/j.cell.2014.08.009. PMID: 25215497; PMCID: PMC4163041.

19. Giudice G, Sánchez-Cabo F, Torroja C, Lara-Pezzi E. ATtRACT-a database of RNA-binding proteins and associated motifs. Database (Oxford). 2016 Apr 7;2016:baw035. doi: 10.1093/database/baw035. PMID: 27055826; PMCID: PMC4823821.

20. Virant-Klun I, Leicht S, Hughes C, Krijgsveld J. Identification of Maturation-Specific Proteins by Single-Cell Proteomics of Human Oocytes. Mol Cell Proteomics. 2016 Aug;15(8):2616-27. doi: 10.1074/mcp.M115.056887. Epub 2016 May 23. PMID: 27215607; PMCID: PMC4974340.

21. Cao S, Huang S, Guo Y, Zhou L, Lu Y, Lai S. Proteomic-based identification of oocyte maturation-related proteins in mouse germinal vesicle oocytes. Reprod Domest Anim. 2020 Nov;55(11):1607-1618. doi: 10.1111/rda.13819. Epub 2020 Oct 14. PMID: 32920902.

22. Pfeiffer MJ, Taher L, Drexler H, Suzuki Y, Makałowski W, Schwarzer C, Wang B, Fuellen G, Boiani M. Differences in embryo quality are associated with differences in oocyte composition: a proteomic study in inbred mice. Proteomics. 2015 Feb;15(4):675-87. doi: 10.1002/pmic.201400334. Epub 2015 Jan 3. PMID: 25367296.

23. Jia B, Xiang D, Fu X, Shao Q, Hong Q, Quan G, Wu G. Proteomic Changes of Porcine Oocytes After Vitrification and Subsequent in vitro Maturation: A Tandem Mass Tag-Based Quantitative Analysis. Front Cell Dev Biol. 2020 Dec 23;8:614577. doi: 10.3389/fcell.2020.614577. PMID: 33425922; PMCID: PMC7785821.

24. Scantland S, Tessaro I, Macabelli CH, Macaulay AD, Cagnone G, Fournier É, Luciano AM, Robert C. The adenosine salvage pathway as an alternative to mitochondrial production of ATP in maturing mammalian oocytes. Biol Reprod. 2014 Sep;91(3):75. doi: 10.1095/biolreprod.114.120931. Epub 2014 Jul 30. PMID: 25078684.

25. Luong XG, Daldello EM, Rajkovic G, Yang CR, Conti M. Genome-wide analysis reveals a switch in the translational program upon oocyte meiotic resumption. Nucleic Acids Res. 2020 Apr 6;48(6):3257-3276. doi: 10.1093/nar/gkaa010. PMID: 31970406; PMCID: PMC7102970.

26. Mjolsness E, DeCoste D. Machine learning for science: state of the art and future prospects. Science. 2001 Sep 14;293(5537):2051-5. doi: 10.1126/science.293.5537.2051. PMID: 11557883.

27. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. Volume 97, Issues 1–2, 1997, Pages 273-324, ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(97)00043-X.

28. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). 2016. Association for Computing Machinery, New York, NY, USA, 785–794. doi: https://doi.org/10.1145/2939672.2939785

29. Keany E. BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. (1.1). 2020. Zenodo. https://doi.org/10.5281/zenodo.4247618

30. Lipovetsky S, Conklin M, 2001. "Analysis of regression in game theory approach," Applied Stochastic Models in Business and Industry, John Wiley & Sons, vol. 17(4), pages 319-330, October.

31. Lundberg SM, Lee S, A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30. 2017.

32. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020 Jan;2(1):56-67. doi: 10.1038/s42256-019-0138-9. Epub 2020 Jan 17. PMID: 32607472; PMCID: PMC7326367.

33. van Heeringen SJ, Veenstra GJ. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. Bioinformatics. 2011 Jan 15;27(2):270-1. doi: 10.1093/bioinformatics/btq636. Epub 2010 Nov 15. PMID: 21081511; PMCID: PMC3018809.

34. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016 Sep 15;32(18):2847-9. doi: 10.1093/bioinformatics/btw313. Epub 2016 May 20. PMID: 27207943.

35. Chen J, Melton C, Suh N, Oh JS, Horner K, Xie F, Sette C, Blelloch R, Conti M. Genome-wide analysis of translation reveals a critical role for deleted in azoospermia-like (Dazl) at the oocyte-to-zygote transition. Genes Dev. 2011 Apr 1;25(7):755-66. doi: 10.1101/gad.2028911. PMID: 21460039; PMCID: PMC3070937.

36. Otsuka H, Fukao A, Funakami Y, Duncan KE, Fujiwara T. Emerging Evidence of Translational Control by AU-Rich Element-Binding Proteins. Front Genet. 2019 May 2;10:332. doi: 10.3389/fgene.2019.00332. Erratum in: Front Genet. 2021 Jun 28;12:715196. PMID: 31118942; PMCID: PMC6507484.

37. Paronetto MP, Bianchi E, Geremia R, Sette C. Dynamic expression of the RNA-binding protein Sam68 during mouse pre-implantation development. Gene Expr Patterns. 2008 May;8(5):311-22. doi: 10.1016/j.gep.2008.01.005. Epub 2008 Feb 2. PMID: 18321792.

38. Wang B, Li L, Zhu Y, Zhang W, Wang X, Chen B, Li T, Pan H, Wang J, Kee K, Cao Y. Sequence variants of KHDRBS1 as high penetrance susceptibility risks for primary ovarian insufficiency by mis-regulating mRNA alternative splicing. Hum Reprod. 2017 Oct 1;32(10):2138-2146. doi: 10.1093/humrep/dex263. PMID: 28938739.

39. Prochazkova B, Komrskova P, Kubelka M. CPEB2 Is Necessary for Proper Porcine Meiotic Maturation and Embryonic Development. Int J Mol Sci. 2018 Oct 12;19(10):3138. doi: 10.3390/ijms19103138. PMID: 30322039; PMCID: PMC6214008.

40. Igea A, Méndez R. Meiosis requires a translational positive loop where CPEB1 ensues its replacement by CPEB4. EMBO J. 2010 Jul 7;29(13):2182-93. doi: 10.1038/emboj.2010.111. Epub 2010 Jun 8. PMID: 20531391; PMCID: PMC2905248.

CHAPTER 4.

CROSS-SPECIES META-ANALYSIS OF TRANSCRIPTOME CHANGES DURING THE MORULA TO BLASTOCYST TRANSITION: METABOLIC AND PHYSIOLOGICAL CHANGES TAKE CENTER STAGE

4.1 Abstract

The morula-to-blastocyst transition (MBT) culminates with formation of inner cell mass (ICM) and trophectoderm (TE) lineages. Recent studies identified signaling pathways driving lineage specification, but some features of these pathways display significant species divergence. To better understand evolutionary conservation of the MBT, we completed a meta-analysis of RNA sequencing data from five model species and ICM versus TE differences from four species. Although many genes change in expression during the MBT within any given species, the number of shared differentially expressed genes (DEGs) is comparatively small, and the number of shared ICMTE DEGs is even smaller. DEGs related to known lineage determining pathways (e.g., POU5F1) are seen, but the most prominent pathways and functions associated with shared DEGs or shared across individual species DEG lists impact basic physiological and metabolic activities, such as TCA cycle, unfolded protein response, oxidative phosphorylation, sirtuin signaling, mitotic roles of polo-like kinases, NRF2-mediated oxidative stress, estrogen receptor signaling, apoptosis, necrosis, lipid and fatty acid metabolism, cholesterol biosynthesis, endocytosis, AMPK signaling, homeostasis, transcription, and cell death. We also observed prominent differences in transcriptome regulation between ungulates and nonungulates, particularly for ICM- and TE-enhanced mRNAs. These results extend our understanding of shared mechanisms of the MBT and the formation of ICM TE and should better inform the selection of model species for particular applications.

4.2 Introduction

Preimplantation development in mammals entails a set of unique events that culminate in the production of a conceptus that is competent to implant into the uterus and continue embryogenesis. The transition from morula to blastocyst (MBT) entails a complex combination of morphological, physiological, and metabolic changes associated with formation of a fluid-pumping polarized epithelium to drive cavitation, in concert with the first cell lineage specification event specifying the inner cell mass (ICM) and trophectoderm (TE) lineages, followed soon thereafter by the separation of primitive endoderm and epiblast lineages (1–8). Important insights into the mechanisms driving these unique features of the MBT have been elucidated in model species, particularly the mouse, for which numerous induced gene mutations affect key molecular pathways related to cell polarization, cell lineage specification, modulations in genomic imprinting, and progress of X chromosome inactivation (2, 3, 9–15). However, fundamental differences exist between species, including differences in genomic imprinting, X chromosome regulation, the expression patterns and roles of essential cell-lineage factors and regulation of signaling pathways related to linage formation (16–28). The existence of such interspecies differences, particularly in early processes that are considered fundamental to mammalian embryogenesis, suggests that there is value in understanding developmental mechanisms in multiple mammalian model species, to better understand the human embryo, and embryos of any given species. Through such studies, shared mechanisms and processes driving important developmental events should be revealed.

Whole transcriptome analysis provides a powerful approach that can be readily applied to mammalian oocytes and embryos, avoiding technical and cost limitations of many other approaches. Indeed, numerous whole transcriptome datasets have been described for mammalian

embryos of different species and stages, and efforts have been made to organize and consolidate these data into online databases (e.g., Ref. 29). Such data have revealed many details of temporal profiles of gene regulation and some of the processes associated with those changes (e.g., Refs. 20, 21, 30–32). However, translating data across species can be challenging due to interspecies differences in development, interlaboratory variations in assay platforms, and variations in embryological methods, as noted previously (e.g., Ref. 20), as well as less efficient methods of identifying overlaps, such as those that rely heavily on gene list intersects. Recently developed computational tools have provided new capabilities for undertaking cross-species meta-analyses of transcriptome data, both at the level of identifying shared DEGs and at the level of identifying relevant functional associations and predicted effects on cells. We recently used such methods to complete a meta-analysis of oocyte maturation-associated transcriptome changes to discover essential shared features of mRNA regulation, combining data for changes in mRNA abundance with changes in mRNA translation status (33). Exemplifying the power of cross-species transcriptome meta-analysis, that study yielded several new insights, most notably the discovery that modulation of the abundances of mRNAs related to oxidative phosphorylation is a major feature of oocyte maturation, signifying a key role for modulating mitochondrial function in regulating oocyte physiology.

The goal of this study was to complete a detailed meta-analysis of whole transcriptome RNA sequencing data for five mammalian species to discover shared changes in gene expression and associated pathways and processes that drive the MBT. The meta-analysis employed a novel analysis pipeline to account for interspecies and interlaboratory differences associated with the individual datasets. Using this novel pipeline, we discovered that although thousands of genes display significant changes in relative mRNA expression levels during this MBT in each species,

the proportion of such genes shared across species is comparatively small. However, these shared DEGs are associated with shared pathways and biological functions that predominantly relate to essential cell physiological and metabolic processes. The existence of such highly shared pathways and functions for the MBT highlights essential molecular changes that underlie this important embryonic process. We also find that the number of shared DEGs that distinguish ICM from TE is much more restricted than those distinguishing morula from blastocyst, as is the number of associated shared pathways and functions. We propose that these prominent, highly conserved changes in physiology and metabolism create a permissive state supporting key developmental events such as lineage formation, which themselves are driven largely by posttranscriptional mechanisms.

4.3 Materials and Methods

4.3.1 Overview of study design

Datasets for each species (Table 1) were processed, followed by the identification of differentially expressed genes (DEGs) and then analysis of associated pathways, functions, and regulators using QIAGEN Ingenuity Pathway Analysis (IPA; Qiagen, Hilden, Germany; https://www.qiagenbioinformatics. com/products/ingenuity-pathway-analysis; RRID:SCR_ 008653) (34), and subsequent interspecies comparisons of DEGs and associated IPA results.

4.3.2 Data set selection and data processing

We identified five species for which datasets were available for comparing morula and blastocyst stages and meeting other quality parameters (Supplemental Table S1; all Supplemental material is available at https://doi.org/10.6084/m9.figshare.15031854.v2). We identified four species for which ICM and TE could be compared (Supplemental Table S1).

Datasets were accessed via the European Nucleotide Archive (RRID:SCR_006515). Study parameters are listed in Table 1 and Supplemental Table S1, including sequencing platform, sequencing read format/length, and RNA sequencing preparation kit. Unless otherwise noted, each study was processed by downloading raw sequencing data in fastq format. Initial sequencing quality metrics were conducted using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc; RRID:SCR_014583). Trimming was conducted using Fastp (RRID:SCR_016962, v0.20.0) (35) with the following parameters: minimum quality threshold of 20, minimum length of 20 bps, and removal of low complexity/mononucleotide reads. Genome indexing and mRNA abundance quantitation were performed with Kallisto (RRID:SCR_ 016582, v0.44.0) (36), using standard settings. For studies containing both morula versus blastocyst and ICM versus TE, the two comparisons were processed and analyzed independently.

We applied two complementary methods for analyses. The "full method" used the full gene lists from each species for identifying DEGs, mapping Ensembl gene identifiers to gene symbols, and then using those DEG lists for subsequent analyses. An average of 14,332 genes were captured per study using the full method (Supplemental Table S1). The second method, denoted here as the "homology method," limited the analysis to orthologous genes with a high level of homology across all five species, using a list of genes selected from the MetaPhOrs repository of phylogeny-based orthologs and paralogs (37). The MetaPhOrs database utilizes information from PhylomeDB (RRID:SCR_007850) (38), Ensembl Compara (39), EggNOG (RRID:SCR_002456) (40), TreeFam (RRID:SCR_013401) (41), Evolclust (42), Hogenom (43), and OrthoMCL (RRID:SCR_ 007839) (44). A score is assigned to each orthology and paralogy prediction based on its level of consistency across the different sources. The consistency score is

the ratio of the number of trees confirming given relationship over the total number of trees that were used to infer the relationship, with a recommended consistency score of 0.5. All pairwise species comparisons were retrieved from the MetaPhOrs database. Using homologous gene lists was intended to minimize impact on results from interspecies differences in genome builds (gaps in sequencing, unannotated genes, evolutionary divergence, etc.). The full method may include genes not annotated in all species. The application of sequence-based consistency scores across species can result in the exclusion of well-studied genes from the study list, even apparent gene homologs that share a gene symbol annotation. Consequently, the more restricted repertoire of genes used with the homology method can limit the number of DEGs identified for downstream analyses and can thus underestimate conservation of gene expression differences and associated effects on pathways and functions. Because the two methods present complementary strengths and weaknesses, we present outputs for both methods to provide a complete view of the analysis. Results are presented in the text in the format of "full (homology)" values. We do not view one method as more correct or more reliable than the other. Rather, the two methods of analysis are complementary, seeking to account for the possible impact of interspecies genetic differences on subsequent analyses.

4.3.3 Human embryos and data processing

Three human studies were identified for inclusion: PRJNA153427 (45) (MBT), PRJNA291062 (46) [MBT and inner cell mass and trophectoderm (ICMTE)], and PRJNA293908 (47) (ICMTE). ICM and TE separation for both PRJNA291062 and PRJNA293908 was conducted via laser cutting. FastQC identified aberrant nucleotide distribution in the first 13 basepairs for PRJNA153427 and PRJNA293908, and the first 3 for PRJNA291062. The Fastp settings were set to include a hard trim to remove those basepairs

from start of reads. The human cDNA genome (GRCh38, annotation build 102) was downloaded from Ensembl, whereupon quantification and differential expression were conducted as detailed in the data processing and differential expression calculation sections.

Study PRJNA153427 utilized a single-cell sequencing-based approach. The authors endeavored to identify ICM and TE samples by using the expression of selected genes and clustering. However, the submitted metadata does not include these derived cell types; rather they are labeled as blastocyst. Due to there being a total of 29 putative blastocyst samples, for the purpose of this analysis, we used the blastocyst label and the averaging of the 29 samples for calculating the transcriptome differences between morula and blastocyst. Study PRJNA291062 employed a whole embryo sequencing approach and contained blastocysts with staging classified as early, mid, and late. We tested the impact of using all three blastocyst stages, exclusion of early, and then the inclusion of only late. The resultant impact on the metaRNASeq output when integrating with PRJNA153427 was less than a 10% difference across the three selection methods. We therefore included all PRJNA291062 blastocyst stage samples in the analysis.

4.3.4 Rhesus embryos and data processing

Two rhesus monkey studies were selected that met inclusion criteria: PRJNA448149 (32) and PRJNA343030 (48), both for the MBT comparison. Both studies showed aberrant nucleotide distribution in the first 13 basepiars via FastQC. Therefore, the Fastp settings were set to include a hard trim to remove those basepairs from start of reads. The rhesus cDNA genome (Mmul_10, annotation build 102) was downloaded from Ensembl, whereupon quantification and differential expression were conducted as detailed in the data processing and differential expression calculation sections.

4.3.5 Mouse embryos and data processing

Three mouse studies were identified for inclusion: PRJNA231896 (49) (MBT), PRJNA289146 (50) (MBT and ICMTE), and PRJNA246056 (51) (ICMTE). The PRJNA246056 study employed immunosurgery for isolating ICM and concanavalin A-conjugated magnetic beads to isolate TE cells. PRJNA289146 separated the ICM and TE cells via pipetting in calcium-free medium. FastQC identified aberrant distribution in the first 13 basepairs for PRJNA231896 and the first 9 for PRJNA291062. Therefore, the Fastp settings were set to include a hard trim to remove those basepairs from start of reads. The mouse cDNA genome (GRCm38, annotation build 102) was downloaded from Ensembl, whereupon quantification and differential expression were conducted as detailed in the data processing and differential expression calculation sections.

4.3.6 Cow embryos and data processing

Four cow studies were identified for inclusion: PRJNA 228235 (52) (MBT), PRJNA254699 (53) (MBT), PRJNA286918 (54) (ICMTE), and PRJNA228235 (55) (ICMTE). ICM and TE samples from PRJNA286918 were extracted via magnetic microbeads conjugated to mouse anti-FITC IgG1. PRJNA228235 ICM and TE samples were collected via dissection. Study PRJNA254699 was ABSOLiD based and required transformation from csfastq to fastq via the Perl script “csfq2fq.pl” (obtained from https://gist.github. com/pcantalupo). FastQC identified aberrant nucleotide distribution in the first 13 basepairs in PRJNA656838 and PRJNA254699. Therefore, the Fastp settings were set to include a hard trim to remove those basepairs from start of reads. The cow cDNA genome (ARS-UCD1.20, annotation build 102) was downloaded from Ensembl, whereupon quantification and differential expression were conducted as detailed in the data processing and differential expression calculation sections.

4.3.7 Pig embryos and data processing

Four pig studies were identified for inclusion: PRJNA 648324 (56) (MBT), PRJNA580004 (57) (ICMTE), PRJNA656843 (55) (ICMTE), and PRJNA307541 (58) (ICMTE). Study PRJNA 648324 contained samples of morula and blastocyst stage embryos that developed in vivo or in vitro. For this analysis, the two types of embryos were processed as separate datasets, yielding two datasets for comparing morula and blastocyst. The PRJNA580004 study separated the ICM and TE cells via manual pipetting and flushing. PRJNA307541 obtained ICM and TE samples via an ultrasharp splitting blade with a stereomicroscope. The PRJNA656843 ICM and TE samples were collected by dissection. Study PRJNA307541 was ABSOLiD-based and required transformation from csfastq to fastq via the Perl script "csfq2fq.pl" (obtained from https://gist.github.com/pcantalupo). FastQC identified aberrant in the first 13 basepairs in PRJNA656843 and PRJNA580004, 9 basepairs in PRJNA648324, and 3 basepairs for PRJNA307541. Therefore, the Fastp settings were set to include a hard trim to remove those basepairs from start of reads. The pig cDNA genome (Sscrofa11, annotation build 102) was downloaded from Ensembl, whereupon quantification and differential expression were conducted as detailed in the data processing and differential expression calculation sections.

4.3.8 Differential expression calculation and gene homology

All Kallisto outputs were imported into R and processed with DESeq2 (RRID:SCR_015687, v1.30.1) (59), and transcript abundance was collapsed to gene using Ensembl gene identifiers, converted with biomaRt (RRID:SCR_019214, v2.45.8) (60). Two different lists of genes were processed through DESeq2 for each study: unfiltered gene lists and only genes with high level of homology across species as defined by MetaPhOrs consistency threshold.

Because normalization and expression thresholding were performed independently on both the full and homology gene lists, the homology lists were not simple subsets of the full method gene lists. For both methods, genes with a fragments per kilobase per million mapped fragments (FPKM) above 1 in at least one sample were included for differential expression calculation. DESeq2 natively applies independent filtering, resulting in the maximum number of genes for multiple test correction. The threshold is dependent on the mean of normalized counts from all samples, set at the lowest quantile wherein the number of rejections is within one residual standard deviation. Coupling the DESeq2 native independent filtering with our manually applied 1 FPKM threshold minimizes the chance of identifying low-quality DEGs. Additionally, we applied the zFPKM package (61) for filtering, which was developed using human cell lines for the accurate detection of biologically relevant genes in RNAseq datasets; we found our threshold of 1 FPKM a more stringent threshold than the zFPKM method imposes (data not shown). For each study, DESeq2 (45) was used to calculate differentially expressed genes (DEGs) between morula and blastocyst and ICM and TE, where a positive log2(fold-change) indicates a higher expression in blastocyst as compared with morula and TE as compared with ICM; the level of significance for genes was set at an adjusted P value [false discovery rate (FDR)] below 0.05. For comparison across species, the Ensembl gene identifiers were converted to gene symbols with biomaRt. Care was taken to identify and rectify gene symbols mapping to multiple Ensembl gene identifiers via a tiered approach: 1) exclusion of genes present on non-chromosomal scaffolds/contigs and 2) if remaining genes showed the same direction of change, the entry with the lowest FDR was selected. This approach allowed the removal of all duplicated entries.

4.3.9 Differential meta-analysis

For each species and cell comparison, the R package metaRNASeq (RRID:SCR_002174, v1.0.3) (62) was used to integrate included studies to calculate a meta P value via the Fisher combination method. As described in Schall and Latham (33), the Fisher's combination method assumes that gene counts follow a negative binomial distribution within each included study. The null hypothesis, that each gene is not differentially expressed, was tested for each study, and then Fisher's exact test was applied to calculate gene- and study-wise P values. As there are different biases inherent for specific library preparation kits and sequencing platforms, leveraging metaRNASeq mitigates these differences and in turn adds strength by integrating different studies for the derivation of a cohesive transcriptome and associated significant changes. Volcano plots for each included study, pre-metaRNASeq [log2(fold-change) versus log10(FDR)] and post-metaRNASeq [log2(fold-change) versus log10(Fisher)], can be found in Supplemental Figs. S1 and S2.

4.3.10 IPA

As in an earlier study (33), the biological significance of observed shared and species-restricted DEGs was assessed using Ingenuity Pathway Analysis (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathwayanalysis). We focused this analysis on canonical pathways (CPs), disease and functions (DFs), and upstream regulators (URs; IPA database content as of May, 2021) (34). IPA was selected due to the robustness of its manually curated knowledgebase, which contains >7 M observations (Qiagen. com, March 2021) including molecular interactions organized into >700 pathways and reported associations of molecules with diseases and biological functions, and >30 integrated third-party databases (Qiagen IPA in-program description), and its ability to compare multiple datasets. Similar to

standard gene set enrichment methodology, submitted gene lists are compared with the genes associated with each CP/DF/UR to calculate a level of significant overlap (P value; significance set at 0.05). With the known impact of up- or downregulating genes on a given IPA CP or biological functions (BFs) entry, the software can also calculate a direction of CP or BF modulation (activation or inhibition), denoted as positive and negative z-scores, respectively (significance set at z> j1.96j). For the upstream regulator analysis, the activity of a UR is predicted based on the direction of change for the downstream DEG targets. It should be noted that the magnitudes of gene expression changes do not factor into the calculations, only the direction of change. For the purposes of this analysis, DF entries were filtered to remove disease/cancer-related entries, and the term biological functions (BFs) was applied. Additionally, for all IPA results, only those with more than one DEG present were included. For both the morula versus blastocyst and ICM versus TE comparisons, the CP and BF results were retrieved. The UR analysis was only applied to the ICM versus TE comparison.

4.3.11 DEG and IPA Figures

Generated barplots quantifying results were produced in R (v4.0.2) with the R package ggplot2 (RRID:SCR_014601, v3.3.3) (63). Heatmaps of DEGs and IPA results were generated using the R package ComplexHeatmap (RRID:SCR_017270, v2.6.2) (64) and arranged with cowplot (RRID:SCR_018081, v1.1.1) (65).

4.4 Results

4.4.1 Overview of Datasets and Limitations

We identified five species for which published datasets were available for comparing morula and blastocyst stages and four species for which published datasets were available for

comparing ICM and TE. Some studies provided data for morula and whole blastocysts, samples for morula and separated ICM and TE samples, or samples for all these stages/cell types together, as indicated in Supplemental Table S1. Cross-species comparisons for differences between morulae and blastocyst used only those studies that contained both morulae and blastocyst samples, or a mixture of single cell samples for the two stages and not specifically divided by cell lineage. Studies comparing ICM and TE used only those studies with separated ICM and TE samples.

Using these datasets, we identified DEGs between morula and blastocyst stages (denoted as MBT DEGs) and between ICM and TE (denoted ICMTE DEGs) for each species (whole species, W.S. sets, Fig. 4.1) using a false discovery rate (FDR) < 0.05. W.S. DEG lists comparing morula and blastocyst stages ranged in size from 3,142 to 5,773 (1,559– 2,802; Supplemental Tables S2–S6). Comparisons between ICM and TE cells were not performed for monkey due to lack of data availability. For the other four species, the total number of genes expressed more highly in ICM, or TE (ICMTE DEG lists) ranged in size from 280 to 2,927 (91 to 1,101), with pig and cow having numbers of DEGs similar to each other but fewer than human and mouse (Fig. 4.1; Supplemental Tables S7–S10). We noted that the ICMTE DEG lists contained some genes that were not classified as detected in whole embryo samples, possibly due to an overall low level of expression less discernible at the whole embryo level.

To evaluate to what extent DEGs expressed more highly in ICM or TE were up- or downmodulated at the whole embryo level during the MBT, we assessed overlaps between the MBT DEG and the ICMTE DEG lists for each species (Fig. 4.2). The largest number of DEGs for which expression was M < B and higher in ICM or TE was seen for human, followed by mouse, cow, and then pig. Interestingly, a large majority (74% full and homology methods) of

such genes were expressed more highly in the TE in humans, and 87% (86%) in pig, whereas only 46% (48%) and 40% (38%) of such genes were expressed more highly in TE for mouse, and cow, respectively. For DEGs that were downregulated during the MBT and different between ICM and TE, a majority were expressed preferentially in the ICM with 74% (72%), 86% (81%), 91% (78%), and 87% (87%) for human, mouse, cow, and pig, respectively. For all four species, a majority of ICMTE DEGs displayed no significant difference in expression between stages, with 57% (57%), 72% (72%), 56% (59%), and 43% (47%) for human, mouse, cow, and pig, respectively.

4.4.2 Shared DEGs observed for MBT and ICMTE DEG lists

A cross-species comparison of MBT DEGs revealed 78 (37) DEGs seen in all five species, and 408 (242) additional DEGs shared in four of the five species (Fig. 4.1; denoted as "All 5" and "4 of 5"; Supplemental Table S11). Comparing ICMTE DEGs for human, mouse, pig, and cow revealed a limited number of shared DEGs (Fig. 4.1; designated as "All 4" and "3 of 4"; Supplemental Table S12), numbering just 21 (15) expressed more highly in ICM (ICM > TE DEGs) and 11 (7) expressed more highly in TE (TE > ICM DEGs).

4.4.3 Shared IPA Features for the Morula-to-Blastocyst Transition

We next evaluated the degree to which the shared pathways and functions were associated with the MBT using two approaches. First, we applied IPA to the shared DEGs identified as intersections between the W.S. MBT DEG lists (Fig. 4.1). Second, we applied IPA to each individual W.S. MBT DEG list and then identified between-species overlaps for the IPA results. This second method complements the first method because effects on pathways and

functions can emerge through modulations in different sets of DEGs associated with a pathway or function.

Applying IPA to the shared MBT DEGs ("All 5", and combined all 5 plus 4 of 5, i.e., "4&5") identified a number of key associated pathways (Fig. 4.3, first two columns; Supplemental Table F3 S13). The "All 5" DEGs garnered 9 (4) significant CPs (none with significant z-scores), whereas the "4&5" DEGs resulted in 54 (26) CPs including 3 (4) inhibited, and 10 (3) activated. Seven CPs were significant in both groups from the full method: TCA cycle, unfolded protein response, NRF2-mediated oxidative stress response, mitotic roles of polo-like kinase, phagosome maturation, clathrin-mediated endocytosis signaling, and role of OCT4 in mammalian ESC pluripotency. The "4&5" DEGs also indicated the inhibition of sirtuin and RHOGDI signaling and the activation of oxidative phosphorylation, RHOA signaling, ILK signaling, telomerase signaling, and superpathway of cholesterol biosynthesis (Fig. 4.3, first two columns; Supplemental Table S13).

The shared MBT DEGs were further associated with significant effects on 111 (76) BFs [3 (0) activated, 1 (1) inhibited] from the "All 5" and 218 (213) from the "4&5" [27 (6) activated, 7 (6) inhibited], of which 64 were shared (Fig. 4.4, first two columns; Supplemental Table S14). Prominent results included the shared feature in both full method datasets (significant in the homology without significant z-score) for the inhibition organismal death and the activation of metabolism of membrane lipid derivative. The "4&5" DEGs, for both the full and homology methods, found activation of invasion of cells, cell survival, cell viability, synthesis/metabolism of cholesterol, various lipid, sphingolipid, and fatty acid metabolism functions, concentration of ATP and steroid metabolism and inhibition of accumulation of lipid,

quantity of sphingolipid, accumulation of glycosylceramide, and organ degeneration (Fig. 4.4, first two columns; Supplemental Table S14).

We next applied IPA to each individual W.S. MBT DEG list and assessed overlaps in IPA results between species. To do this, it was necessary to reduce the number of W.S. MBT DEGs submitted for IPA in accordance with application guidelines by limiting the submitted DEGs to FDR < 0.001. The total of number of significantly affected CPs shared between "All 5" species was 20 (48) and for "4 of 5" 46 (62) and the total number of shared BFs was 48 (58) and 28 (35) for "All 5" and "4 of 5," respectively (Figs. 3–5; Supplemental Tables S13 and S14). Many of the CPs and BFs associated with the shared "4&5" DEGs were also associated with the W.S. DEGs observed for all five species, 4 of 5 species, or smaller subsets of species using the full method (Figs. 3 and 4; Supplemental Tables S13 and S14). Many of these were also seen with the homology method, with or without significant z-scores for some species (Figs. 3 and 4; Supplemental Tables S13 and S14). Some shared CPs and BFs were seen with the homology method only. Additionally, some CPs and BFs were significantly affected in all five species but not associated with the shared DEGs. These included signaling pathways for integrins, RHO family GTPases, CXCR4, as well as ferroptosis and myo-inositol synthesis, and several functions related to growth, cell death, morbidity and mortality (inhibited in three species), metabolism and synthesis of reactive oxygen species, and necrosis, among others. Additional CPs were associated with 4 of 5 species but not the shared DEGs, including actin cytoskeleton, RHGDI (inhibited in three species), mTOR, and HIF1a signaling, protein ubiquitination, and epithelial adherens junctions, among others.

4.4.4 Shared IPA Features for ICM-Enhanced and TE-Enhanced DEGs

Although the number of shared ICM > TE and TE > ICM DEGs (Fig. 4.1) was comparatively small, IPA yielded informative results for pathways and functions associated with shared DEGs ("All 4" plus 3 of 4 combined, i.e., "3&4") DEGs in the two cell types (Figs. 6, 7, 8, and 9; Supplemental Tables S15–S18). We note that z-scores were not calculated because between-stage comparisons do not apply.

For the "3&4" DEGs expressed more highly in the ICM, IPA revealed 12 (0) affected CPs and 110 (0) affected BFs (Figs. 6 and 7; Supplemental Tables S15 and S16). Prominent affected CPs for the full method in the ICM included regulation of epithelial-mesenchymal transition, multiple stem cell pluripotency entries including role of NANOG, and signaling through IGF1, IL-15, STAT3, TGF-b, and PI3/AKT, and regulation of the epithelial-mesenchymal transition (EMT; Supplemental Table S15). For shared "3&4" DEGs expressed more highly in TE cells, IPA revealed a single affected CP (glucocorticoid receptor signaling) for the full method (0 results for homology method) and 57 (0) affected CPs (Fig. 4.8; Supplemental Table S17). Prominent affected BFs from the TE full method IPA results included entries related to cell death, lipid metabolism, cell migration, endocytosis, and cell invasion, among others (Fig. 4.9; Supplemental Table S18). Because of the small numbers of genes, we did not apply IPA to subsets of ICM > TE or TE > ICM DEG sets distinguished by MBT expression comparisons.

To compare overlaps between IPA results for individual species (W.S.) ICMTE DEG lists, we were able to retain the FDR < 0.05 threshold for inclusion in the uploaded DEG lists. For ICM-enhanced genes, there were 33 (27) shared CPs and 54 (65) shared BFs (Figs. 5–7;

Supplemental Tables S15 and S16). For TE-enhanced DEGs, there were 11 (2) shared CPs and 38 (29) shared BFs (Figs. 5, 8, and 9; Supplemental Tables S17 and S18).

Affected CPs associated with ICM-enhanced DEGs in all four species included IGF1 signaling and transcription regulatory network of embryonic stem cells (ESCs). Using the homology method, CPs shared across species included multiple entries related to pluripotency as well as JAK, IL-6, IL9, and FLT3 signaling (Fig. 4.6; Supplemental Table S15). A number of affected CPs and BFs were associated with ICM-enhanced DEGs in at least three species with either method (full or homology) but not seen for shared ICM DEGs (Figs. 7 and 8; Supplemental Tables S15 and S16). Three BFs were associated with all four W.S. lists and the shared DEG "3&4" list (apoptosis, organismal death, and necrosis).

CPs associated with TE-enhanced DEGs of at least three species but not with shared DEGs included ferroptosis, endocytosis signaling, mTOR signaling, ILK signaling (full method), RhoA signaling (both methods), phagosome maturation, and superpathway of cholesterol biosynthesis (homology method; Fig. 4.8; Supplemental Table S17). Most of the BFs associated with at least three species of TE-enhanced DEGs were also seen for the shared DEGs, particularly using the homology method (Fig. 4.9; Supplemental Table S18).

4.4.5 Affected IPA Upstream Regulators Associated with ICM-TE Divergence

In addition to the CP and BF analyses, the upstream regulator (UR) analysis was applied to the ICMTE DEGs. Most of the URs associated with the shared DEGs were likewise associated with at least 3 of the W.S. DEG lists, and many were shared in all 4 species using either full or homology method, including many well-known regulators of ICM-TE lineage divergence such as POU5F1, GATA6, and Let-7 (Fig. 4.10; Supplemental Tables S19 and S20). Some URs implicated in ICM with either or both methods were themselves ICM-enhanced DEGs across

multiple species, such as STAT3, NANOG, SOX17, and KIT. Interestingly, many of the URs implicated in human ICM were DEGs only in human. URs implicated in TE using either or both methods included FOS, ESR1, and TGFB1. An additional three URs were found in all datasets from the full method (FSH, NFKBIA, and ESR2) and 8 URs present in all W.S. lists from the full method (TP53, YAP1, PTEN, MTOR, CDKN1A, PPARA, SREBF1, and WT1). Two URs (MYC, KLF5) implicated with either or both methods were themselves TE-enhanced DEGs in at least two species. Interestingly, MYC was ICM-enhanced in the mouse but TE-enhanced in human and pig.

4.4.6 Taxonomic Differences in Gene and Pathway Regulation

We examined set overlaps for evidence of features that distinguish ungulate species that form epitheliochorial placentae (cow and pig) and non-ungulate species that form hemochorial placentae (human, mouse, and monkey). There was little evidence of such taxonomic separation at the level of overall MBT (Figs. 3 and 4, Supplemental Tables S13 and S14). Indeed, from both the full and homology methods, only three CPs were exclusive to cow + pig (endothelin-1 signaling, melatonin signaling, and heme biosynthesis from uroporphyrinogen-III-I) and a single pathway was limited to human + mouse + rhesus (PTEN signaling; Supplemental Table S13). However, numerous CPs and BFs were associated solely with human + mouse or cow + pig W.S. ICMTE DEG lists (Figs. 6, 7, 8, and 9). Notable CPs associated solely with human and mouse ICM-enhanced DEGs included HIPPO signaling, role of OCT4 in pluripotency, and other signaling pathways including NRF2-mediated oxidative stress response, signaling through mTOR, ATM, TNFR1, IL3, IL10, insulin, ERK5, and PTEN, and HOTAIR regulatory pathway (Fig. 4.6; Supplemental Table S15). Several BFs were associated only with cow and pig ICM-enhanced DEGs including multiple immune cell-related functions, development of stem cells,

binding of connective tissue (Fig. 4.7; Supplemental Table S16). For TE-enhanced DEGs, many CPs were only seen for human and mouse, including adherens junction, RHOGDI, RHOGTPase, actin cytoskeleton and other actin-related CPs, EIF2, integrin, ephrin, RAC, necroptosis, estrogen receptor, and IL-1 signaling, and functions related to DNA repair (Fig. 4.8; Supplemental Table S17). Only ATM signaling was associated specifically with cow and pig TE-enhanced DEGs. These differences were largely echoed for BFs, but with the notable addition of prostaglandin metabolism, BFs related to inflammation and cell proliferation and endocytosis for cow and pig, and many BFs related to protein metabolism, actin and microtubule functions, and cell proliferation for human and mouse (Fig. 4.9; Supplemental Table S18).

We also observed artiodactyl ungulate versus non-ungulate differences for genes related to eight cell lineage formation pathways, for which member genes were identified in the IPA database and compared with the W.S. DEG lists (Table T2 2). Differential expression between ICM and TE was prominent for these genes in human and mouse, but much less so for cow and pig (Table 2, Supplemental Tables S7–S10). Some mRNAs for genes that are widely characterized as ICM- or TE-specific did not display conserved differences in such a manner. For example, CDX2 was only TE specific in humans, despite its well-characterized role in TE-specific functions (3). Other genes displayed ICM- or TE-enhanced expression in human and mouse but not cow or pig. Some genes (e.g., STAT3, WASF1, ACTG1, ID2, MYC, FGFR2, and ASH2L) displayed opposite enrichment in human and mouse, and many other genes were lineage-enhanced in a single species.

4.5 Discussion

This meta-analysis is the first to offer a comprehensive between-species comparison of gene regulation during the morula-to-blastocyst transition for five experimental mammalian

model species, addressing both stage-specific changes in expression as well as the emergence of differential expression between ICM and TE cells. The main outcome of this meta-analysis is that, at the level of mRNA expression, there is a very limited repertoire of shared DEGs comparing between stages, and even fewer shared DEGs distinguishing ICM and TE cells. These limited groups of shared DEGs for whole embryos, ICM, and TE are associated with limited numbers of shared pathways and functions. Additional overlaps between W.S. DEG lists echoed many of the pathways and functions associated with the shared DEGs for whole embryos, but not those distinguishing ICM and TE. These results collectively indicate substantial species divergence in gene regulation discernible at the mRNA level during the MBT and the specification of ICM and TE cells.

One striking result from our analysis was the nature of the most prominent conserved effects on pathways and functions associated with shared DEGs and those observed as overlaps between pathways and functions associated with individual W.S. DEG lists comparing whole morulae and blastocysts. These included basic pathways related to cell physiology and metabolism, such as TCA cycle, unfolded protein response, oxidative phosphorylation, sirtuin signaling, mitotic roles of polo-like kinases, NRF2-mediated oxidative stress, metabolism, apoptosis, necrosis, lipid and fatty acid metabolism, multiple carbohydrate metabolism pathways, cholesterol biosynthesis, endocytosis, AMPK signaling, cellular homeostasis, transcription, and cell death. The broad conservation of regulated changes in expression of genes associated with such general CPs and BFs reveal key roles for changes in fundamental metabolic, physiological, and cellular features in supporting the MBT, possibly providing a permissive environment to enable cell lineage formation. These were accompanied by functions and pathways expected to accompany the MBT, such as OCT4 (POU5F1) function, RhoA signaling, estrogen receptor

signaling, cytoskeleton organization, tight junction signaling, adherens junction remodeling, and cell invasion.

Changes related to increased oxidative phosphorylation were associated with shared DEGs and four of five W.S. DEG lists. Additionally, sirtuin signaling was a downregulated CP observed for shared DEGs and all five species W.S. DEG lists. Sirtuin signaling and mitochondrial dysfunction pathways were also associated with the TE-enhanced DEGs for human and mouse. Because one function of SIRT3 is negative regulation of oxidative phosphorylation (66), these results are reminiscent of previous observations that oxidative phosphorylation increases during the MBT (67–70). A previous study also observed diminishing expression of all sirtuins from zygote to blastocyst stages but an important embryo-protective role for maternally expressed SIRT3 in protecting against reactive oxygen species production and p53-mediated demise (71). Continued modulations in metabolism may play a key role in the elaboration of pluripotency by balancing energy production with other needs such as providing metabolic intermediates for anabolic purposes (6). Intricate and correct timing of changes in mitochondrial function and oxidative phosphorylation may thus be key for formation of healthy blastocysts and proper formation of ICM and TE lineages. Our results indicate that such dynamic regulation may be among the most conserved features of mammalian blastocyst biology.

Our analysis also highlighted NRF2 oxidative stress response as a shared CP for the MBT and affected in the TE of mouse and human blastocysts. An earlier study noted that NRF2-mediated oxidative phosphorylation was associated with the MBT in pig embryos (56). NRF2 function was proposed to protect bovine blastocysts from reactive oxygen species damage (72). Another recent study noted NRF2 function as a possible early marker of TE formation (73). Our data further highlight NRF2 function as playing an important role across mammalian species,

possibly protecting embryos from oxidative damage at a time of increasing metabolic energy production demands.

Estrogen receptor signaling was another prominent shared CP for the MBT using the W.S. and "4&5" DEG lists, and both ESR1 and ESR2 were implicated as shared affected URs in both the ICM and TE cells. Estrogen receptor mRNA increases in abundance during the transition to blastocyst stage in mice (74), and, additionally, estrogen receptor protein is detected in blastocysts (75), Although ESR1 deficiency is not embryo lethal, its downregulation occurs in conjunction with blastocyst activation after delayed implantation to facilitate implantation (76). Estrogen stimulation increases calcium concentration in dormant blastocysts (77) via GPR30 signaling (78) and inhibits hyaluronan expression needed for blastocyst attachment (79). Interestingly, estradiol also regulates *Wnt* gene expression in blastocysts in conjunction with uterine factors (80). One study reported that estradiol promotes the MBT in pigs (81) and another reported negative effects of estrogen on human blastocysts adhesion in vitro (82). The observations here provide evidence for a conserved role for estrogen signaling in blastocysts, possibly as a shared feature of embryo-maternal communication, attachment, or implantation and could be useful for dissecting the stage- and/or concentration-dependent actions of estrogens during the MBT and beyond.

There were fewer conserved DEGs related to ICM and TE divergence than comparing whole embryos, but conservation was nevertheless seen for multiple affected pathways and functions, including multiple entries related to pluripotency, as well as signaling through STAT3, PI3/AKT, RHOA, and TGFb1. Some pathways normally associated with ICMTE divergence were not as well conserved. Interestingly, we observed a limited degree of shared regulation of individual genes associated with eight key pathways that are widely viewed as serving key,

conserved functions in ICM and TE specification. Even some genes strongly implicated for roles in ICM and TE delineation were only differentially regulated in a subset of species or some even within a single species. One example of this is the regulation of CDX2 function. CDX2 is initially expressed from maternal mRNA and ubiquitously throughout early cleaving embryos, becoming restricted to the TE where it suppresses ICM-specific functions (3). But CDX2 displayed TE-specific mRNA expression only in the human. Previous transcriptome studies have concluded phylogenetic differences in the regulation of genes associated with pluripotency and lineage divergence as well (20, 21, 30). Because the transcriptome data are limited to the mRNA level of analysis, such divergence in individual gene regulation indicates that the essential mechanisms that delineate the ICM and TE lineages may employ key elements of posttranscriptional control impacting protein expression, localization, and function, or that distinct sets of pathway member genes operate within each species to achieve common outcomes via a limited number of shared pathways. For the example of CDX2, other posttranscriptional mechanisms may drive the elaboration of a TE-specific mode of CDX2 function in other species, whereas the human relies more heavily on regulation at the level of mRNA expression.

Species differences in the nuances of lineage-determining gene regulation and function have been described for between-species comparisons involving fewer species (8, 20, 21, 30). For example, mouse and cow embryos differ in the mechanisms regulating YAP1 nuclear localization (83) and in other aspects of HIPPO pathway signaling such as the temporal regulation of TAZ nuclear localization and differences in the regulation of CDX2 by TEAD4 (8). Differences have also been reported for effects of SMARCA5inhibition comparing mice and cattle (84). Spatial and temporal regulation of POU5F1 (OCT4) and its control of CDX2 differ between species. By analyzing data for five species in a single study, our data indicate that such

divergence may be more extensive than previously appreciated. One striking observation to emerge is the differences between ungulates and non-ungulates for the CPs and BFs associated with ICM- and TE-enhanced DEGs. These include many prominent pathways that play key roles in ICM and TE lineage formation, including HIPPO signaling, OCT4 role in pluripotency, adherens junction formation, RHOGDI signaling, among others. Such differences in DEGs, CPs, and BFs indicate that the species differences shown here for the spatial and temporal regulation of these key pathways may be particularly significant between ungulates and non-ungulates and may reflect, at least in part, later differences in placentation (epitheliochorial, hemochorial).

The small number of shared DEGs comparing either whole morulae and blastocysts or isolated ICM and TE cells could be considered to reflect inherent variation in timing and kinetics of developmental progression or variations between studies in how the embryos or cells were obtained (Table 1), particularly given the potential impact of embryo culture and interspecies differences in timing of gene expression changes relative to morphological changes previously observed using array technologies (20). However, we consider this explanation unlikely to account for the extent of differences in transcriptome regulation observed, for the following reasons. First, other studies have described significant phylogenetic differences between species, particularly with respect to lineage-dependent DEGs (20, 21, 30). For example, comparing epiblast versus primitive endoderm DEGs revealed just 23.2% and 17.8% overlap of human DEGs with marmoset and mouse, respectively (30). This compares favorably with the 11.2% overlap of ICMTE DEGs between human and mouse found here, although the amount of overlap may vary with stage. Second, regarding differences in developmental kinetics or timing, we note that comparing morula and blastocyst stage embryos and comparing ICM and TE cells entail comparisons of the clearly distinguishable and clearly definable endpoints, before and after key

developmental events and in embryos that have fulfilled a minimum number of changes to accomplish these developmental transitions. This greatly minimizes the potential impact of species differences in developmental timing, which would predominantly affect intermediate and less definable points during these transitions. Although transcriptomes may continue to change to some degree after cavitation, for example, this would not eliminate the need for sufficient transcriptome change to allow a blastocyst to form. Our analysis reveals the degree of conservation and nature of pathways and functions associated with shared DEGs and these essential transitions. Third, with regard to potential effects of embryo culture previously observed, particularly for ICMTE DEGs (20), we note that data for three of the species (mouse, pig, cow) included both flushed and cultured embryos, thereby mitigating the chance that any single species would represent an outlier due to embryo culture effects. Fourth, we note that for study where blastocysts of different degrees of expansion were sampled separately (PRJNA291062), we assessed the impact on results of pooling or not pooling these samples and found <10% effect on identified DEGs, suggesting that transcriptome changes between blastocysts of different degrees of expansion are modest and unlikely to dramatically impact the DEG list for a given species. It should also be noted that we avoided using elongated blastocysts in the ungulate species, so that blastocyst samples across species were similar morphologically. Finally, we emphasize the power of the metaRNASeq approach to overcome many artifacts that can arise with interstudy variations. The metaRNASeq method is not based on a simple intersection or union of identified DEGs between input studies. By integrating the P values from input studies, the potentiality of individual study changes or non-changes is mitigated via this applied statistical test. The results, therefore, are DEGs that are indeed fundamental differences between stages or between cell types and supersede minor interstudy variations within species.

The result for each species was the identification of large and very similar numbers of DEGs comparing morulae and blastocysts, providing ample resolution for discovering cross-species shared DEGs. And although the number of ICMTE DEGs is lower for ungulates than non-ungulates, this does not negate the finding that many ICMTE DEGs were restricted to human or mouse, which displayed similar numbers of ICMTE DEGs. Additionally, we imposed stringent criteria for inclusion of studies in the meta-analysis and rejected studies due to low level of bioreplication. Although TE-ICM samples were utilized from Kong et al. (20) (PRJNA580004), we opted to exclude the MBT comparison due to the study including only two blastocyst replicates. Similarly, Chitwood et al. (85; PRJNA401876), exploring preimplantation embryos in rhesus monkeys, was likewise excluded due to having just two replicates at both morula and blastocyst stages. These considerations collectively support the conclusion that the degree of conservation of transcriptome changes associated with the MBT or ICM-TE divergences is limited and associated prominently with changes related to basic metabolic and physiological functions.

Our meta-analysis included the development and refinement of novel methods of data processing and analysis that should be broadly applicable in diverse biological contexts. Previous forays into the integration of publicly available sequencing data have often been limited to reduced species populations, and most often, singular studies from respective species. Acknowledging that there are variations inherent from the multitude of different methodologies (e.g., embryo staging, culture components, RNA isolation, library preparation kits sequencing platforms, and processing software), the application of methodologies that can integrate disparate studies is essential to further the field. Herein, our applied computational methods accomplish this goal by integrating studies using the metaRNASeq package to derive a more complete

assessment of changes at both the mRNA and functional (IPA) levels, and when coupled with a simplified graphical representation of results (e.g., heat maps, summary bar plots), provides a more cohesive understanding of shared and specific species features with greater mechanistic insight. By applying these methods to a complicated array of datasets spanning both two embryo stages and two cell populations within the blastocyst within a total of five species, we accomplished the most comprehensive assessment, to date, of species conservation of patterns of regulation of genes and associated canonical pathways and biological functions during the MBT. The collection of methods used here should be broadly applicable and can yield substantial new insights when applied to datasets with suitable degrees of species coverage, bioreplication, and richness of sampling. Our analysis demonstrates that as methods for library construction and deep sequencing achieve greater degrees of sensitivity, concurrent refinements in bioinformatics data processing and analysis will have the potential to provide dramatically improved understanding of phylogenetic impacts of gene regulation, species divergence, and evolution.

4.7 Supplemental Data

Supplemental Tables S1–S20 and Supplemental Figs. S1 and S2: https://doi.org/10.6084/m9.figshare.15031854.v2.

4.8 Funding

This work was supported in part by grants from the National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development (T32HD087166), and by MSU AgBioResearch, and Michigan State University. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

APPENDIX

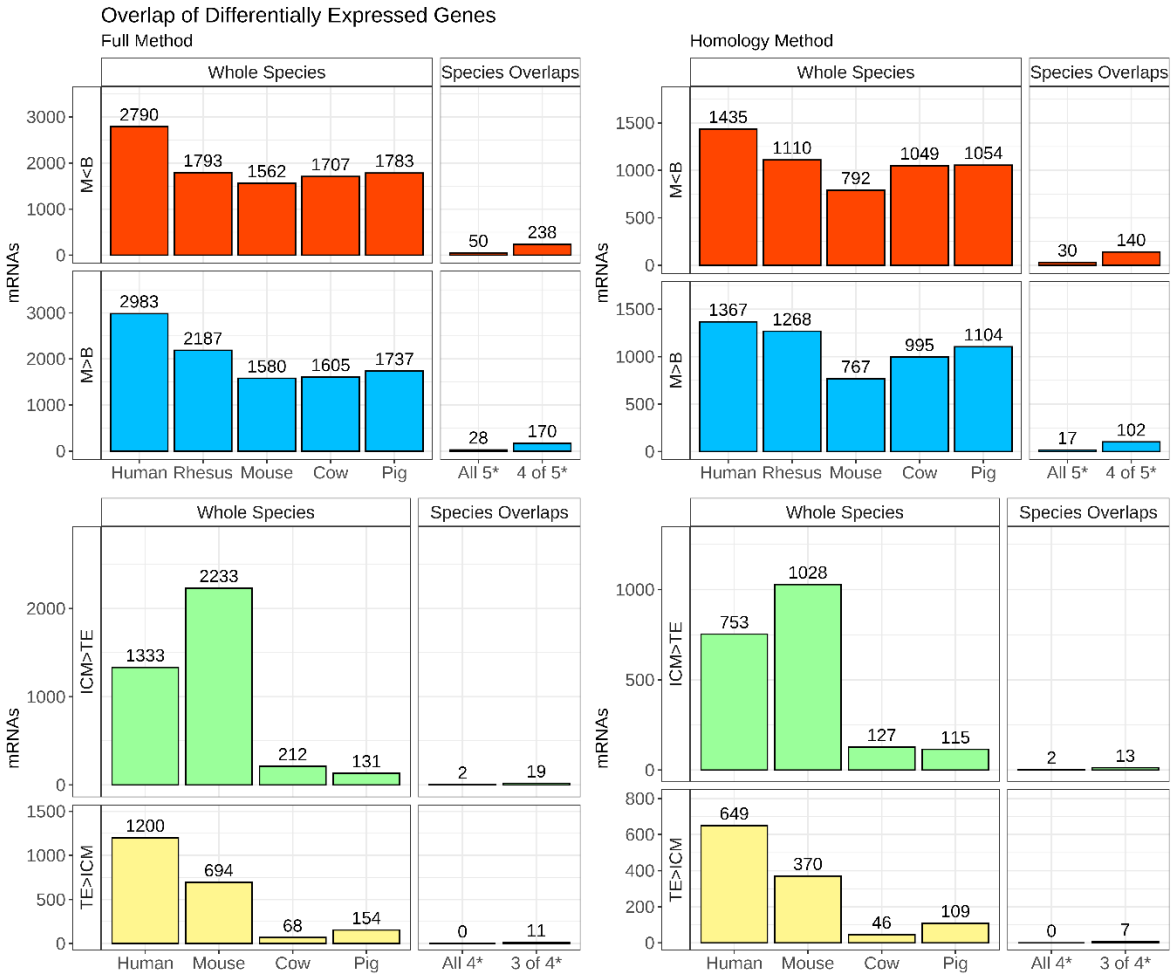

Figure 4.1 - Quantification of numbers of identified MBT and ICMTE DEGs using hull and homology methods.

Quantification of the number of identified differentially expressed genes for the MBT and ICMTE comparisons, for both the full and homology methods. Figure is split into two vertical panels: left panel for the Full method, and the right for the homology method. For both full and homology methods, the panels are split into two rows: top row for the MBT comparison, bottom for the ICMTE comparison. The MBT comparison DEGs are split by direction: top row with red fill are those DEGs with higher expression at Blastocyst, bottom row with blue fills for those with decreased expression in Blastocyst. Similarly, the ICMTE comparison is split by direction: green for ICM>TE, yellow for TE>ICM. For both the MBT and ICM,TE overlaps of the 'Whole Species' DEGs were derived for those shared by all species and those when allowing one species to drop. W.S. lists and intersections of W.S. groups are listed along the x-axis, and the number of identified DEGs are along the y-axis. Comparisons were performed as described in methods. Level of significance was set at FDR<0.05. Data generated from Supplemental Tables S2-12. DEGs, differentially expressed genes; FDR, false discovery rate; ICM, inner cell mass; TE trophectoderm; ICMTE, ICM versus TE; MBT, morula to blastocyst transition.

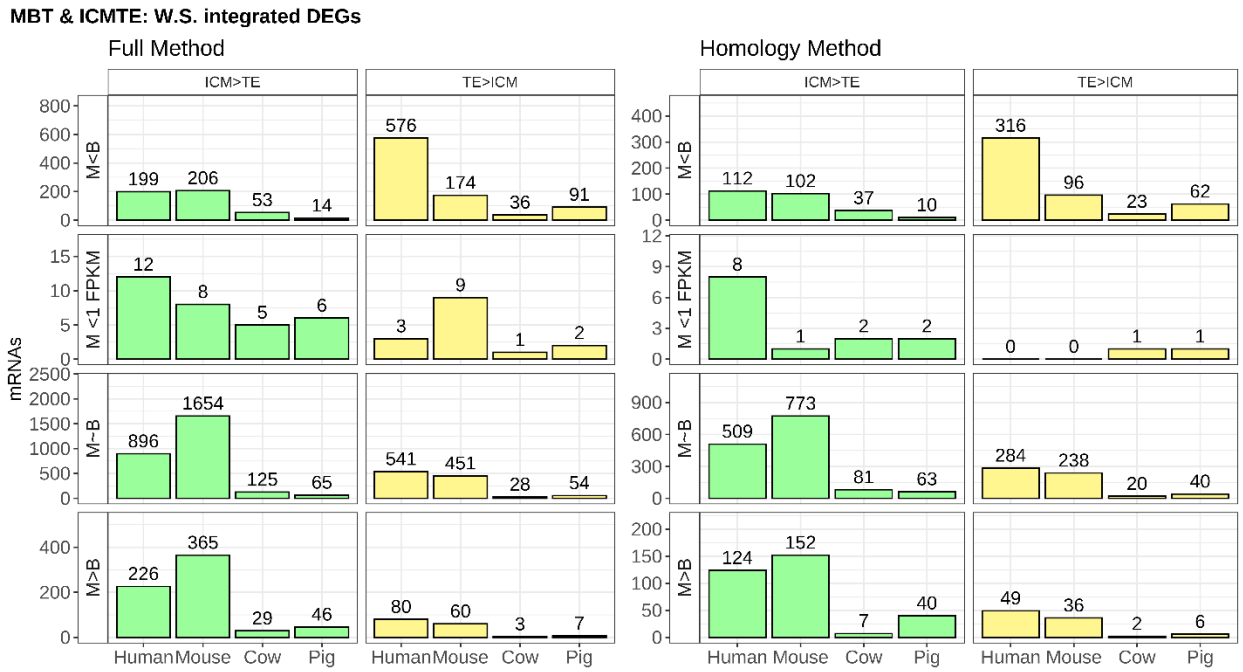

Figure 4.2 – Integration of MBT and ICMTE DEGs

Integration of the MBT and ICMTE DEGs, based on their respective regulation for four species: human, mouse, cow, and pig. Figure is split into two vertical panels for each of the gene derivation methods: full and homology. For each method, their respective panels are bifurcated by ICM-enhanced (ICM>TE, green fill) and TE-enhanced (TE>ICM, yellow fill). For each of the ICM and TE-enhanced genes, their mode of regulation was identified in the MBT comparison. Row 1: M<B, genes increasing in expression. Row 2: <1 FPKM in MBT, genes with no detected expression in the MBT comparison. Row 3: M~B, no significant difference between morula and blastocyst. Row 4: M>B, genes decreasing in expression. Individual species are listed along the x-axis, and the number of identified genes per MBT regulation group are quantified along the y-axis. Comparisons were performed as described in Methods. Significance for all in included DEGs were set at FDR<0.05. Data generated from Supplemental Tables S2-10. DEGs, differentially expressed genes; FDR, false discovery rate; ICM, inner cell mass; TE trophectoderm; ICMTE, ICM versus TE; MBT, morula to blastocyst transition.

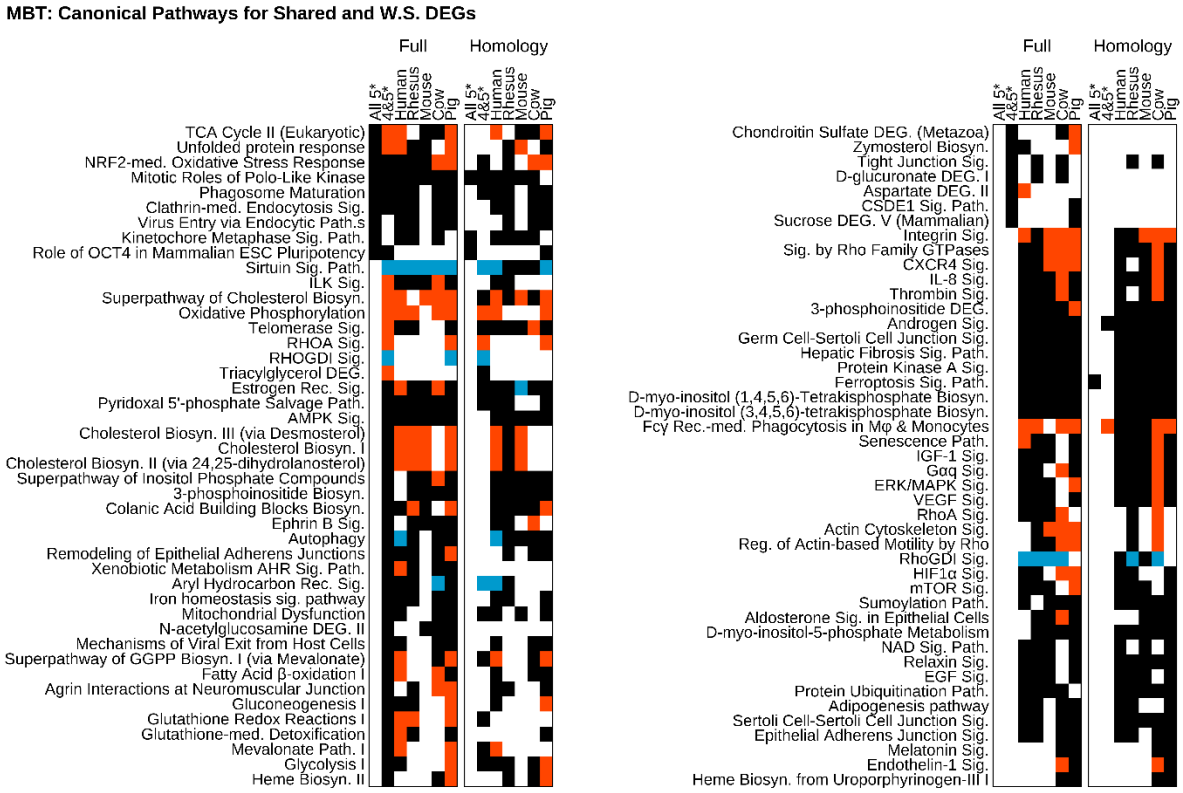

Figure 4.3 - IPA Canonical Pathways for the MBT.
Heatmap depicting IPA canonical pathways for the MBT comparison; entries were limited to those present in at least three datasets in the full analysis. Figure consists of two wrapped columns for the full and homology methods. Along the top x-axis, each column is further divided by dataset, 4&5* DEGs shared by at least four species and the whole species DEG lists for human, rhesus, mouse, cow, and pig. Along the y-axis are the identified canonical pathways. Level of significant overlap was set at $P < 0.05$; entries with white color denote those not meeting significance. Color denotes z score and predicted activity: red = activated, blue = inhibited, black = nonsignificant z score. Data generated from Supplemental Table S13. DEGs, differentially expressed genes; IPA, Ingenuity Pathway Analysis; MBT, morula to blastocyst transition.

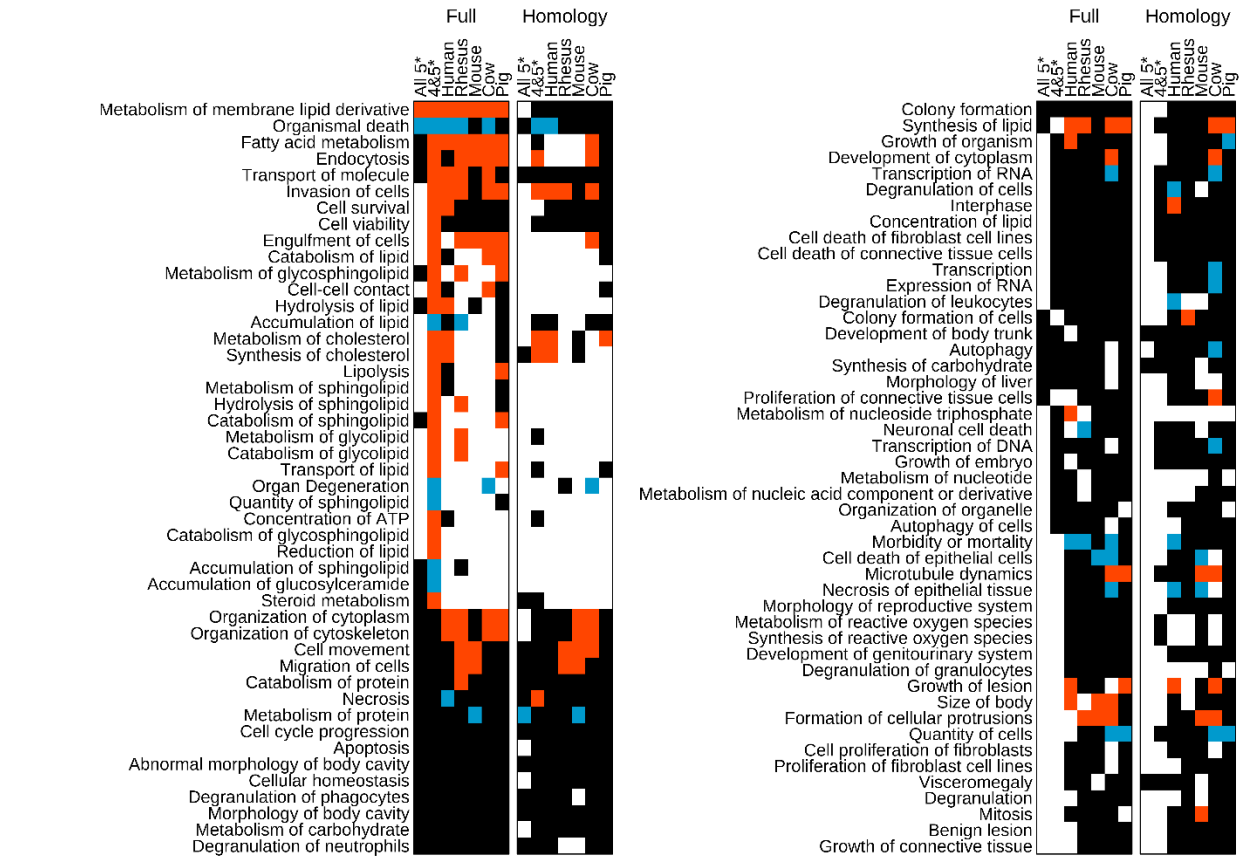

Figure 4.4 - IPA Biological Functions for the MBT.

Heatmap depicting IPA biological functions for the MB comparison; entries were limited to those present in at least three datasets in the full analysis. Figure consists of two wrapped columns for the full and homology methods. Along the top x-axis, each column is further divided by dataset, 4&5* DEGs shared by at least four species and the whole species DEG lists for human, rhesus, mouse, cow, and pig. Along the y-axis are the identified biological functions. Level of significant overlap was set at $P < 0.05$; entries with white color denote those not meeting significance. Color denotes z score and predicted activity: red = activated, blue = inhibited, black = nonsignificant z score. Data generated from Supplemental Table S14. DEGs, differentially expressed genes; IPA, Ingenuity Pathway Analysis; MBT, morula to blastocyst transition.

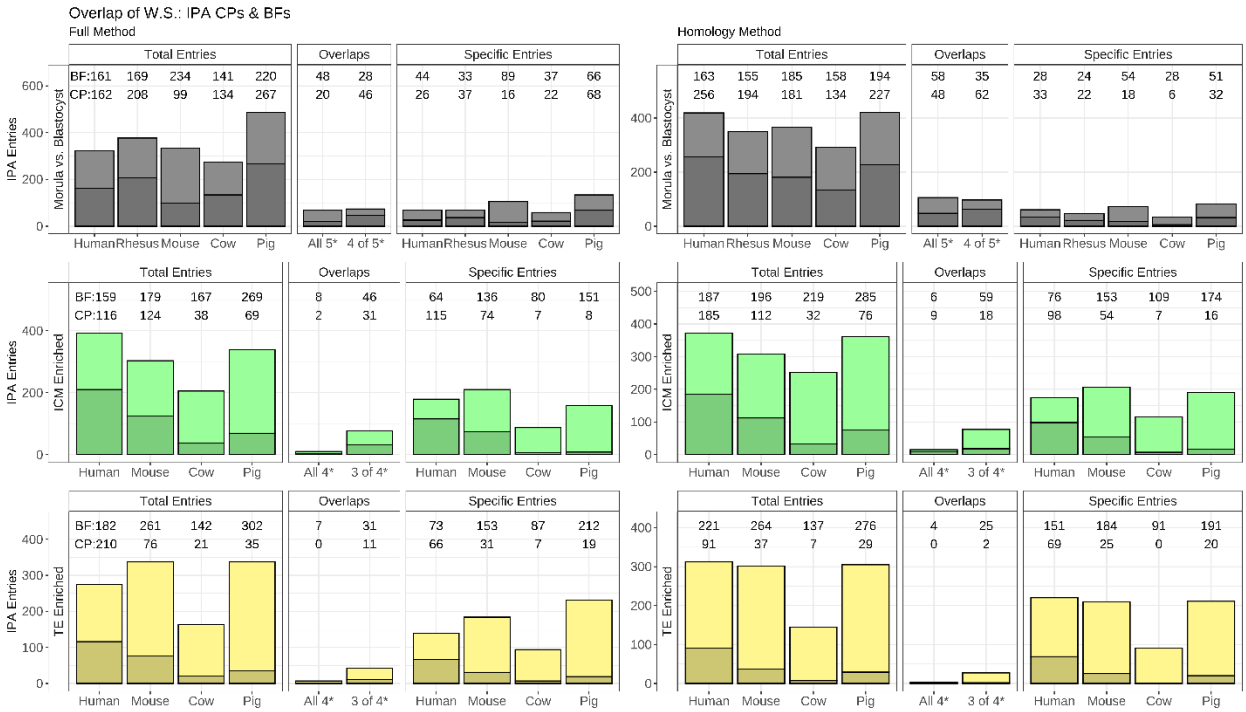

Figure 4.5 - Overlap of IPA results from W.S. MBT and ICMTE

Quantification of the number of pathways/function identified from the IPA analysis for the "whole species" DEG lists for both the MBT and ICMTE comparisons and for both the full and homology methods. Figure consists of two vertical panels for the two different methods: full and homology. Each method panel is further split into three rows: morula vs. blastocyst (gray fill), ICM-enhanced (green fill), and TE-enhanced (yellow fill). Overlaps of IPA results were derived for those shared by all species and allowing for one species to drop. Additionally, pathways and functions present in only one species were quantified. Each bar within the figure depicts the number of pathways (darker fill on bottom) and functions (lighter fill on top). Level of significance was set at $P < 0.05$. Data generated from Supplemental Tables S13–S18. BFs, biological functions; CPs, canonical pathways; DEGs, differentially expressed genes; ICM, inner cell mass; ICMTE, ICM and TE; IPA, Ingenuity Pathway Analysis; MBT, morula to blastocyst transition; TE, trophectoderm; W.S., whole species.

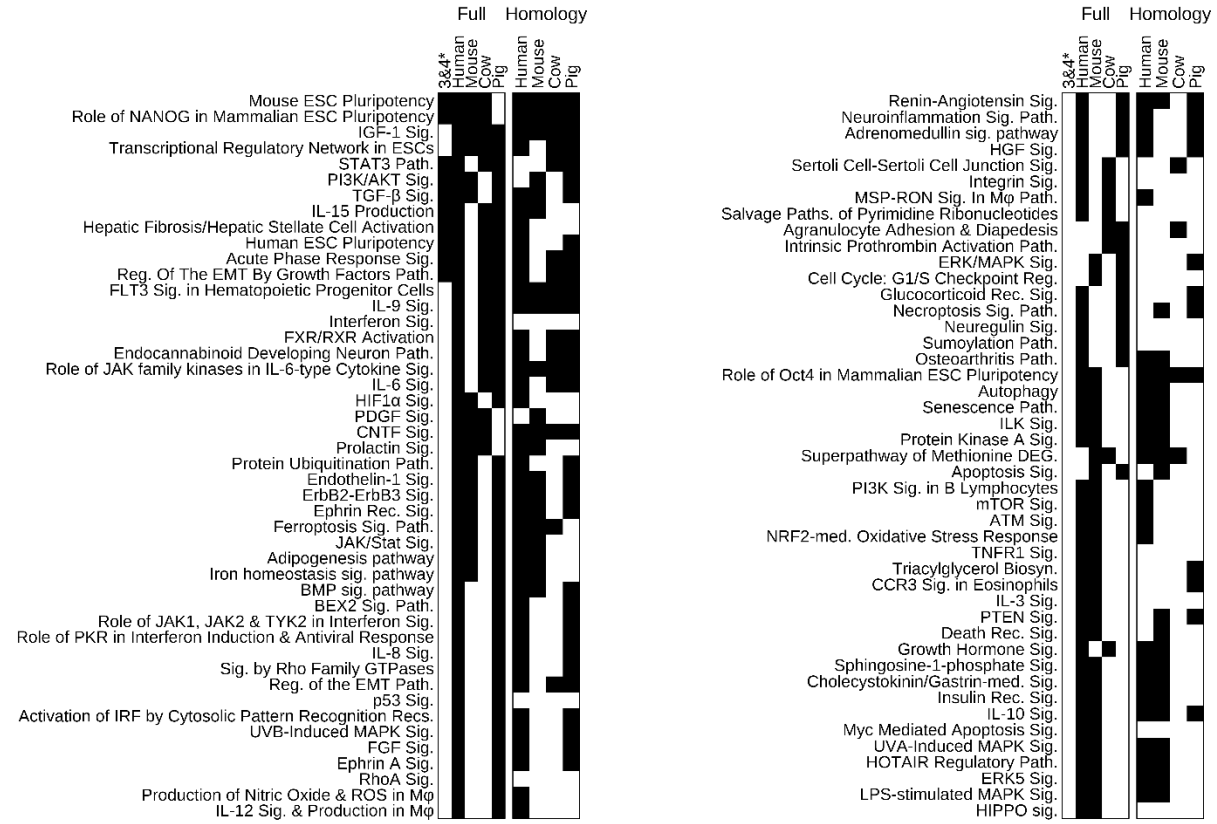

Figure 4.6 - IPA Canonical Pathways: ICM-enhanced

ICM-enhanced heatmap depicting IPA canonical pathways for ICM-enhanced genes; entries were limited to those present in at least two datasets in the full analysis. Figure consists of two wrapped columns for the full and homology methods. Along the top x-axis, each column is further divided by dataset: 3&4* DEGs shared by at least three species and the whole species DEG lists for human, mouse, cow, and pig. Along the y-axis are the identified canonical pathways. Level of significant overlap was set at $P < 0.05$; entries with white color denote those not meeting significance. Data generated from Supplemental Table S15. DEG, differentially expressed gene; ICM, inner cell mass; IPA, Ingenuity Pathway Analysis.

Figure 4.7 - IPA Biological Functions: ICM-enhanced

Heatmap depicting IPA biological functions for ICM-enhanced genes; entries were limited to those present in at least two datasets in the full analysis. Figure consists of two wrapped columns for the full and homology methods. Along the top x-axis, each column is further divided by dataset: 3&4* DEGs shared by at least three species and the whole species DEG lists for human, mouse, cow, and pig. Along the y-axis are the identified biological functions. Level of significant overlap was set at $P < 0.05$; entries with white color denote those not meeting significance. Data generated from Supplemental Table S16. DEG, differentially expressed gene; ICM, inner cell mass; IPA, Ingenuity Pathway Analysis.

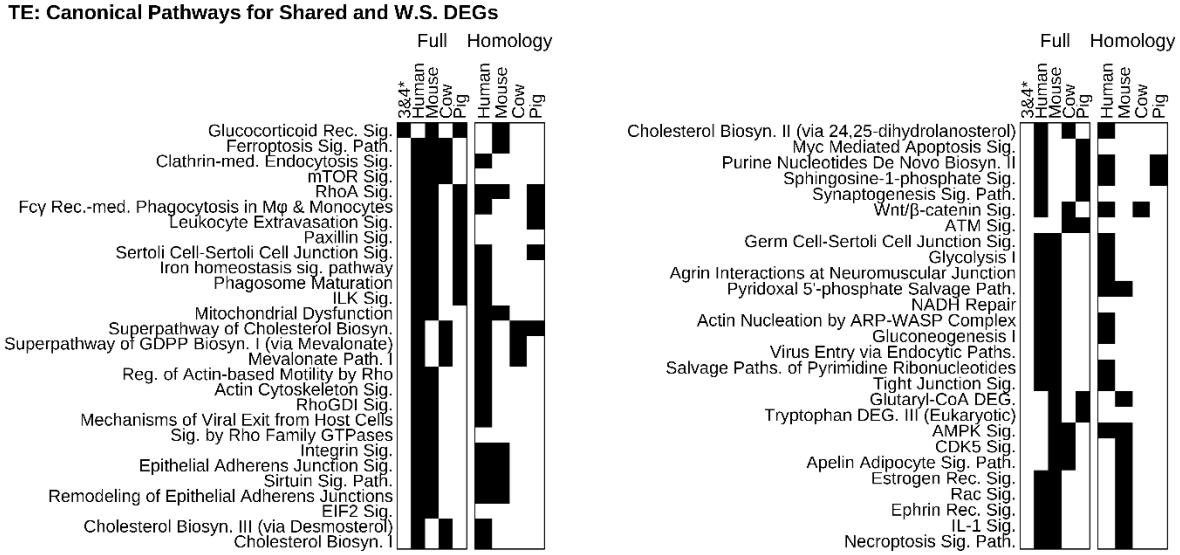

Figure 4.8 - IPA Canonical Pathways: TE-enhanced

Heatmap depicting IPA canonical pathways for TE-enhanced genes; entries were limited to those present in at least two datasets in the full analysis. Figure consists of two wrapped columns for the full and homology methods. Along the top x-axis, each column is further divided by dataset: 3&4* DEGs shared by at least three species and the whole species DEG lists for human, mouse, cow, and pig. Along the y-axis are the identified canonical pathways. Level of significant overlap was set at $P < 0.05$; entries with white color denote those not meeting significance. Data generated from Supplemental Table S17. DEG, differentially expressed gene; IPA, Ingenuity Pathway Analysis; TE, trophectoderm.

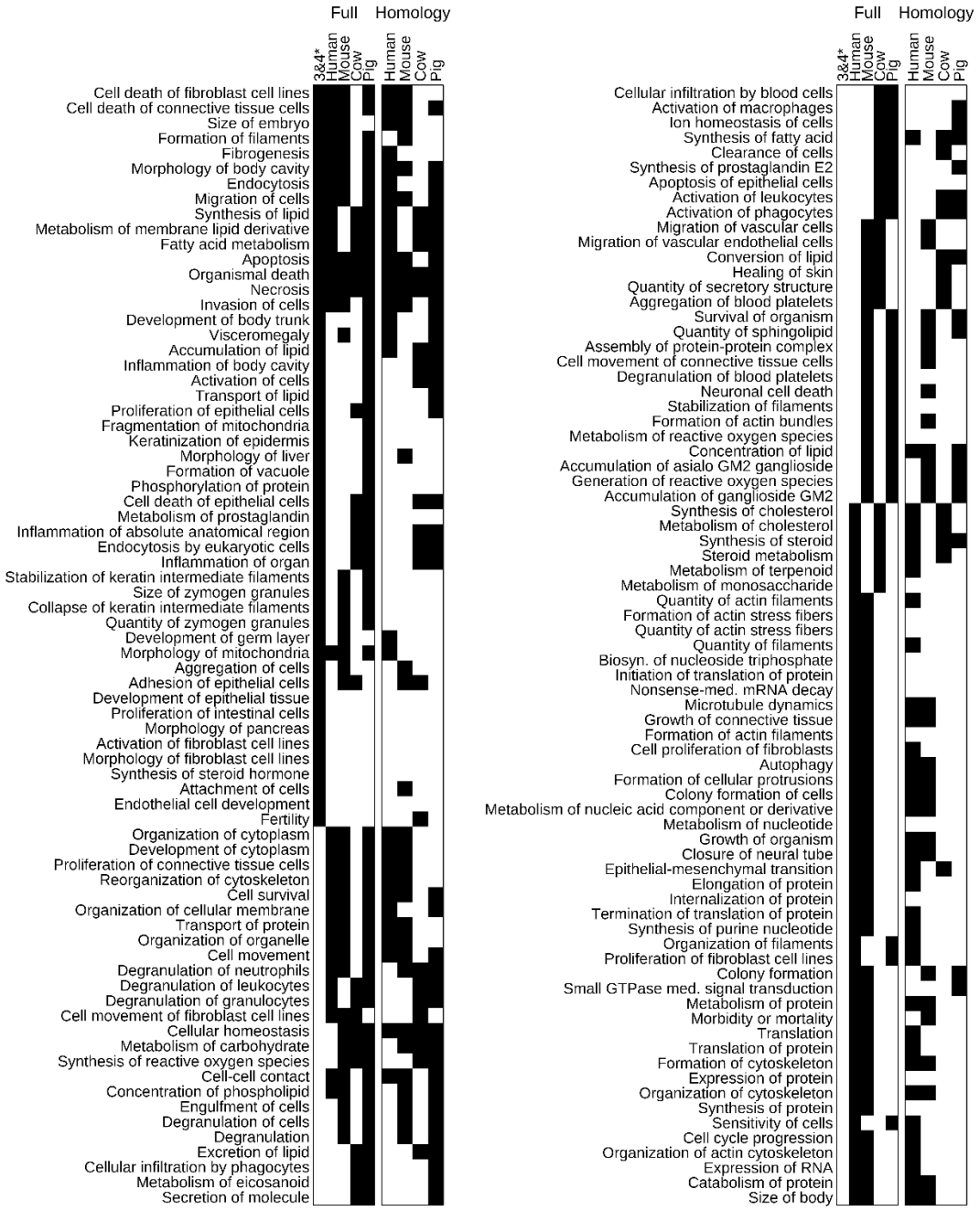

Figure 4.9 - IPA Biological Functions: TE-enhanced.

Heatmap depicting IPA biological functions for TE-enhanced genes; entries were limited to those present in at least two datasets in the full analysis. Figure consists of two wrapped columns for the full and homology methods. Along the top x-axis, each column is further divided by dataset: 3&4* DEGs shared by at least three species and the whole species DEG lists for human, mouse, cow, and pig. Along the y-axis are the identified biological functions. Level of significant overlap was set at $P < 0.05$; entries with white color denote those not meeting significance. Data generated from Supplemental Table S18. DEG, differentially expressed gene; IPA, Ingenuity Pathway Analysis; TE, trophectoderm.

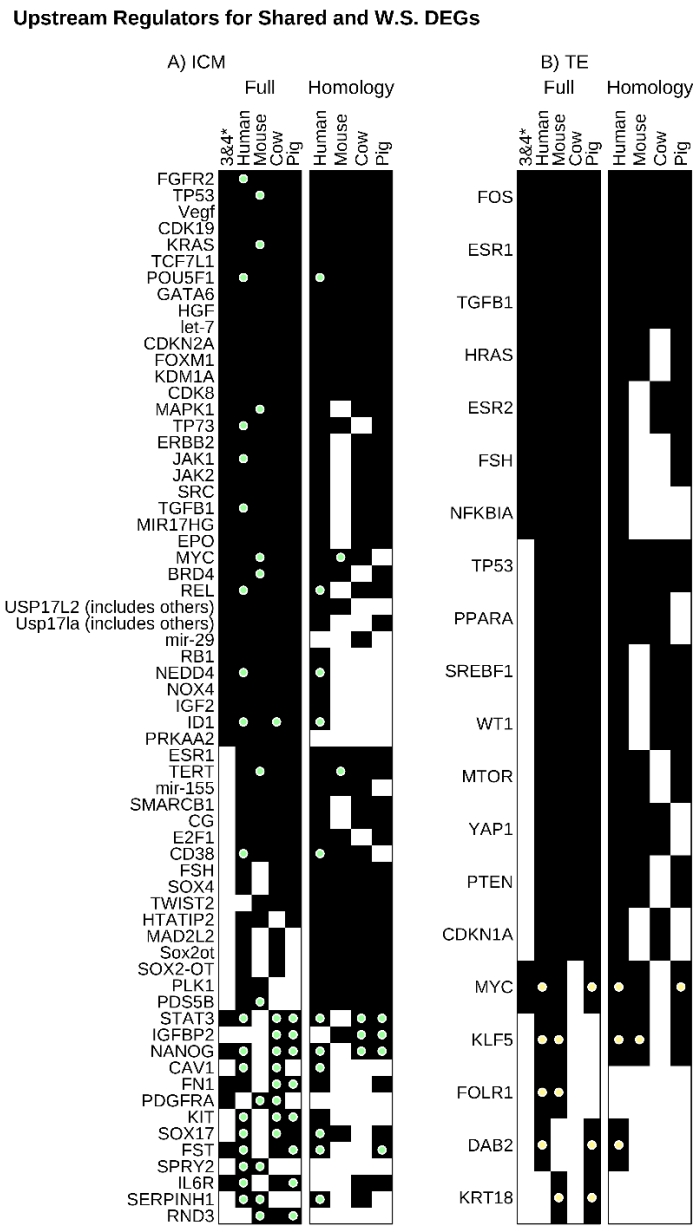

Figure 4.10 - IPA Upstream Regulator: ICM & TE-enhanced

Heatmap depicting IPA upstream regulators for ICM- and TE-enhanced genes, limited to those with significance in all four species in full method. Figure consists of two panels: ICM-enhanced (A) and TE-enhanced (B), with two columns for the full and homology methods. Along the top x-axis, each column is further divided by dataset for three species: human, mouse, and cow. Along the y-axis are the identified biological functions. Level of significant overlap was set at $P < 0.05$; entries with white color denote those not meeting significance. Green circles denote those upstream regulators that are also ICM-enhanced. As described in methods, the identification of DEGs in the full method but not the homology method does not cast doubt on the validity of the DEG call; differences arise due to genes falling below the MetaPhOrs consistency score or impact in changes in normalization using the homology method. Data generated from Supplemental Tables S19 and S20. DEG, differentially expressed gene; ICM, inner cell mass; IPA, Ingenuity Pathway Analysis; TE, trophectoderm.

Table 1- Summary of different study parameters used to obtain embryos

Study Information			Embryo Production Information				Staging	
Species	Study ID	Use	Embryo Production	Pre-embryo production	Culture Medium	Atmosphere	Morula	Blastocyst
Human	PRJNA153427	MBT	IVM-ICSI	GnRH, FSH, hCG, collection 36h post-hCG	G1.3, PBS with 20% HAS, G2	N.S.	Day 4	Day 6
Human	PRJNA291062	MBT, ICMTE	IVM-ICSI	GnRH, FSH, hCG, collection 36h post-hCG	G1.3, G2	6% CO2	N.S.	Early, mid, hatched
Rhesus	PRJNA343030	MBT	IVF-IVC	rhFSH, rhCG, collection 32-35h post-rhCG	HECM-9 + 10% FBS	5% CO2	N.S.	N.S.
Rhesus	PRJNA448149	MBT	IVF-IVC	hFSH, hCG, collection 30h post-hCG	HECM-9	5% CO2, 10% O2, 85% N2	Day 5	Day 6
Mouse	PRJNA289146	MBT, ICMTE	IVF-IVC	PMSG, hCG, collection 14h post-hCG	G1, G2	5% CO2	N.S.	N.S.
Mouse	PRJNA231896	MBT	In vivo, flush (1)	PMSG, hCG	KSOM	N.S.	64h	64+20h culture
Cow	PRJNA228235	MBT	IVM-IVF	N.S.	Synthetic oviduct fluid + 5% ECS + BME AA + MEM Non-essential AA	5% CO2, 5% O2, 90% N2	N.S.	N.S.
Cow	PRJNA254699	MBT	In vivo, flush (1)	FSH, PGF	N.S.	N.S.	Day 6	Day 7
Pig	PRJNA648324	MBT	in vivo	Altrenogest ReguMate, PMSG, hCG	N.S.	5% CO2, 5% O2, 90% N2	Day 4	Day 6
			in vitro	N.S.	hormone free maturation medium, Porcine Zygote medium-5	5% CO2, 5% O2, 90% N2	100 h	174 h
Human	PRJNA293908	ICMTE	IVM-ICSI	ovarian stim. cycle, long agonist protocol	N.S.	N.S.	N.S.	N.S.
Mouse	PRJNA246056	ICMTE	In vivo, flush	PMSG+hCG or PG600	N.S.	N.S.	N.S.	94 h
Cow	PRJNA286918	ICMTE	IVM-IVF	Abattoir	IVM medium, Charles Rosenkrans + CR1aa + 10%FBS	5% CO2	N.S.	Day 7
Cow, Pig	PRJNA656838	ICMTE	IVM-parth.	Abattoir	IVM medium + PMSG, TALP + 6-DMAP	N.S.	N.S.	Day 9 (cow), Day 7 (pig)
Pig	PRJNA580004	ICMTE	In vivo flush (2)	N.S.	N.S.	N.S.	N.S.	132-140 h
Pig	PRJNA307541	ICMTE	in vivo	PMSG, hCG	N.S.	N.S.	N.S.	N.S.

Table 1 (cont'd)

(1) Stimulated cycle, (2) unstimulated cycle. DEGs, differentially expressed genes; FSH, follicle stimulating hormone; GnRH, gonadotropin-releasing hormone; hCG, human chorionic gonadotropin; hMG, human menopausal gonadotropin; ICM, inner cell mass; ICMTE, ICM and TE; MBT, morula-to-blastocyst transition; N.A., not applicable; N.S., not stated; PG600, PMSG þ hCG; PGF, prostaglandin F2a; PMSH, pregnant mare serum gonadotropin; TE, trophectoderm.

Table 2 – Number and proportion of DEGs (full method) associated with indicated pathways

			Human		Rhesus	Mouse		Cow		Pig	
Pathway	Total Genes	Overlaps	MB	ICMTE	MB	MB	ICMTE	MB	ICMTE	MB	ICMTE
Adherens Junction	170	DEGs	65	37	49	28	39	38	3	38	2
		Percent	38%	22%	29%	16%	23%	22%	2%	22%	1%
Human ESC	162	DEGs	50	42	47	22	20	23	7	26	5
		Percent	31%	26%	29%	14%	12%	14%	4%	16%	3%
Mouse ESC	102	DEGs	40	26	34	24	22	16	5	27	4
		Percent	39%	25%	33%	24%	22%	16%	5%	26%	4%
Nanog ESC	117	DEGs	44	30	39	26	17	19	4	27	2
		Percent	38%	26%	33%	22%	15%	16%	3%	23%	2%
Oct4 ESC	46	DEGs	19	18	17	15	14	11	2	12	2
		Percent	41%	39%	37%	33%	30%	24%	4%	26%	4%
PCP	60	DEGs	18	11	9	9	4	9	0	12	1
		Percent	30%	18%	15%	15%	7%	15%	0%	20%	2%
HIPPO	84	DEGs	37	14	27	15	23	18	1	22	2
		Percent	44%	17%	32%	18%	27%	21%	1%	26%	2%
NOTCH	38	DEGs	9	7	3	7	4	8	1	6	0
		Percent	24%	18%	8%	18%	11%	21%	3%	16%	0%
Average of 8 pathways		Percent	36%	24%	27%	20%	18%	19%	3%	22%	2%

DEGs, differentially expressed genes; ICM, inner cell mass; TE, trophectoderm; ICMTE, ICMTE DEGs, MB, MBT DEGs.

REFERENCES

RFERENCES

1. Alarcon VB. Cell polarity regulator PARD6B is essential for trophectoderm formation in the preimplantation mouse embryo. Biol Reprod 83: 347–358, 2010. doi:10.1095/biolreprod.110.084400.

2. Gerri C, McCarthy A, Alanis-Lobato G, Demtschenko A, Bruneau A, Loubersac S, Fogarty NME, Hampshire D, Elder K, Snell P, Christie L, David L, Van de Velde H, Fouladi-Nashta AA, Niakan KK. Initiation of a conserved trophectoderm program in human, cow and mouse embryos. Nature 587: 443–447, 2020. doi:10.1038/s41586-020-2759-x.

3. Karasek C, Ashry M, Driscoll CS, Knott JG. A tale of two cell-fates: role of the HIPPO signaling pathway and transcription factors in early lineage formation in mouse preimplantation embryos. Mol Hum Reprod 26: 653–664, 2020. doi:10.1093/molehr/gaaa052.

4. Kwon J, Kim NH, Choi I. ROCK activity regulates functional tight junction assembly during blastocyst formation in porcine parthenogenetic embryos. PeerJ 4: e1914, 2016. doi:10.7717/peerj.1914.

5. Marikawa Y, Alarcon VB. RHOA activity in expanding blastocysts is essential to regulate HIPPO-YAP signaling and to maintain the trophectoderm-specific gene expression program in a ROCK/actin filament-independent manner. Mol Human Reprod 25: 43–60, 2019. doi:10.1093/molehr/gay048.

6. Mathieu J, Ruohola-Baker H. Metabolic remodeling during the loss and acquisition of pluripotency. Development 144: 541–551, 2017. doi:10.1242/dev.128389.

7. Saini D, Yamanaka Y. Cell polarity-dependent regulation of cell allocation and the first lineage specification in the preimplantation mouse embryo. Curr Top Dev Biol 128: 11–35, 2018. doi:10.1016/bs. ctdb.2017.10.008.

8. Sharma J, Antenos M, Madan P. A comparative analysis of HIPPO signaling pathway components during murine and bovine early mammalian embryogenesis. Genes (Basel) 12: 281, 2021. doi:10.3390/genes12020281.

9. Borensztein M, Syx L, Ancelin K, Diabangouaya P, Picard C, Liu T, Liang JB, Vassilev I, Galupa R, Servant N, Barillot E, Surani A, Chen CJ, Heard E. Xist-dependent imprinted X inactivation and the early developmental consequences of its failure. Nat Struct Mol Biol 24: 226–233, 2017. doi:10.1038/nsmb.3365.

10. Hemberger M. Epigenetic landscape required for placental development. Cell Mol Life Sci 64: 2422–2436, 2007. doi:10.1007/s00018-007-7113-z.

11. Le Bin GC, Munoz-Descalzo S, Kurowski A, Leitch H, Lou X, Mansfield W, Etienne-Dumeau C, Grabole N, Mulas C, Niwa H, Hadjantonakis AK, Nichols J. Oct4 is required for lineage priming in the developing inner cell mass of the mouse blastocyst. Development 141: 1001–1010, 2014. doi:10.1242/dev.096875.

12. Moore JM, Rabaia NA, Smith LE, Fagerlie S, Gurley K, Loukinov D, Disteche CM, Collins SJ, Kemp CJ, Lobanenkov VV, Filippova GN. Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos. PLoS One 7: e34915, 2012. doi:10.1371/journal. pone.0034915.

13. Morgan HD, Santos F, Green K, Dean W, Reik W. Epigenetic reprogramming in mammals. Hum Mol Genet 14 Spec No 1: R47– R58, 2005. doi:10.1093/hmg/ddi114.

14. Ralston A, Rossant J. Cdx2 acts downstream of cell polarization to cell-autonomously promote trophectoderm fate in the early mouse embryo. Dev Biol 313: 614–629, 2008. doi:10.1016/j.ydbio. 2007.10.054.

15. Yagi R, Kohn MJ, Karavanova I, Kaneko KJ, Vullhorst D, DePamphilis ML, Buonanno A. Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development. Development 134: 3827–3836, 2007. doi:10.1242/dev.010223.

16. Berletch JB, Yang F, Disteche CM. Escape from X inactivation in mice and humans. Genome Biol 11: 213, 2010. doi:10.1186/gb-2010-11- 6-213.

17. Cheong CY, Chng K, Ng S, Chew SB, Chan L, Ferguson-Smith AC. Germline and somatic imprinting in the nonhuman primate highlights species differences in oocyte methylation. Genome Res 25: 611– 623, 2015. doi:10.1101/gr.183301.114.

18. Daigneault BW, Rajput S, Smith GW, Ross PJ. Embryonic POU5F1 is required for expanded bovine blastocyst formation. Sci Rep 8: 7753, 2018. doi:10.1038/s41598-018-25964-x.

19. Hisey E, Ross PJ, Meyers SA. A review of OCT4 functions and applications to equine embryos. J Equine Vet Sci 98: 103364, 2021. doi:10.1016/j.jevs.2020.103364.

20. Hosseini SM, Dufort I, Caballero J, Moulavi F, Ghanaei HR, Sirard MA. Transcriptome profiling of bovine inner cell mass and trophectoderm derived from in vivo generated blastocysts. BMC Dev Biol 15: 49, 2015. doi:10.1186/s12861-015-0096-3.

21. Hu Y, Huang K, Zeng Q, Feng Y, Ke Q, An Q, Qin LJ, Cui Y, Guo Y, Zhao D, Peng Y, Tian D, Xia K, Chen Y, Ni B, Wang J, Zhu X, Wei L, Liu Y, Xiang P, Liu JY, Xue Z, Fan G. Single-cell analysis of nonhuman primate preimplantation development in comparison to humans and mice. Dev Dyn 250: 974–985, 2021. doi:10.1002/dvdy.295.

22. Kong Q, Yang X, Zhang H, Liu S, Zhao J, Zhang J, Weng X, Jin J, Liu Z. Lineage specification and pluripotency revealed by transcriptome analysis from oocyte to blastocyst in pig. FASEB J 34: 691– 705, 2020. doi:10.1096/fj.201901818RR.

23. Monk D. Genomic imprinting in the human placenta. Am J Obstet Gynecol 213: S152–S162, 2015. doi:10.1016/j.ajog.2015.06.032.

24. Niakan KK, Eggan K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. Dev Biol 375: 54–64, 2013. doi:10.1016/j.ydbio.2012.12.008

25. Okamoto I, Patrat C, Thepot D, Peynot N, Fauque P, Daniel N, Diabangouaya P, Wolf JP, Renard JP, Duranthon V, Heard E. Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. Nature 472: 370–374, 2011. [Erratum in Nature 474: 239–240, 2011]. doi:10.1038/nature09872.

26. Tachibana M, Ma H, Sparman ML, Lee HS, Ramsey CM, Woodward JS, Sritanaudomchai H, Masterson KR, Wolff EE, Jia Y, Mitalipov SM. X-chromosome inactivation in monkey embryos and pluripotent stem cells. Dev Biol 371: 146–155, 2012. doi:10.1016/j.ydbio.2012.08.009.

27. Wu YQ, Zhao H, Li YJ, Khederzadeh S, Wei HJ, Zhou ZY, Zhang YP. Genome-wide identification of imprinted genes in pigs and their different imprinting status compared with other mammals. Zool Res 41: 721–725, 2020. doi:10.24272/j.issn.2095-8137.2020.072.

28. Xu D, Zhang C, Li J, Wang G, Chen W, Li D, Li S. Polymorphic imprinting of SLC38A4 gene in bovine placenta. Biochem Genet 56: 639–649, 2018. doi:10.1007/s10528-018-9866-5.

29. Hu B, Zheng L, Long C, Song M, Li T, Yang L, Zuo Y. EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. Open Biol 9: 190054, 2019. doi:10.1098/rsob.190054.

30. Boroviak T, Stirparo GG, Dietmann S, Hernando-Herraez I, Mohammed H, Reik W, Smith A, Sasaki E, Nichols J, Bertone P. Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. Development 145: dev167833, 2018. doi:10.1242/dev.167833.

31. Fan X, Tang D, Liao Y, Li P, Zhang Y, Wang M, Liang F, Wang X, Gao Y, Wen L, Wang D, Wang Y, Tang F. Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. PLoS Biol 18: e3001017, 2020. doi:10.1371/journal.pbio.3001017.

32. Schall PZ, Ruebel ML, Midic U, VandeVoort CA, Latham KE. Temporal patterns of gene regulation and upstream regulators contributing to major developmental transitions during Rhesus macaque preimplantation development. Mol Hum Reprod 25: 111–123, 2019. doi:10.1093/molehr/gaz001.

33. Schall PZ, Latham KE. Essential shared and species-specific features of mammalian oocyte maturation-associated transcriptome changes impacting oocyte physiology. Am J Physiol Cell Physiol 321: C3–C16, 2021. doi:10.1152/ajpcell.00105.2021.

34. Kramer A, Green J, Pollard J Jr, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics 30: 523– 530, 2014. doi:10.1093/bioinformatics/btt703.

35. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34: i884–i890, 2018. doi:10.1093/bioinformatics/bty560.

36. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34: 525–527, 2016. doi:10.1038/nbt.3519.

37. Pryszcz LP, Huerta-Cepas J, Gabaldon T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res 39: e32, 2011. doi:10.1093/nar/gkq953.

38. Huerta-Cepas J, Bueno A, Dopazo J, Gabaldon T. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res 36: D491–D496, 2008. doi:10.1093/nar/gkm899.

39. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327–335, 2008. doi:10.1101/gr.073585.107.

40. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44: D286–D293, 2016. doi:10.1093/nar/gkv1248.

41. Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res 42: D922–D925, 2014. doi:10.1093/nar/gkt1055.

42. Marcet-Houben M, Gabaldon T. EvolClust: automated inference of evolutionary conserved gene clusters in eukaryotes. Bioinformatics 36: 1265–1266, 2020. doi:10.1093/bioinformatics/btz706.

43. Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perriere G. Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10, Suppl 6: S3, 2009. doi:10.1186/1471-2105-10-S6-S3.

44. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189, 2003. doi:10.1101/gr.1224503.

45. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, Huang J, Li M, Wu X, Wen L, Lao K, Li R, Qiao J, Tang F. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol 20: 1131–1139, 2013. doi:10.1038/nsmb.2660.

46. Dang Y, Yan L, Hu B, Fan X, Ren Y, Li R, Lian Y, Yan J, Li Q, Zhang Y, Li M, Ren X, Huang J, Wu Y, Liu P, Wen L, Zhang C, Huang Y, Tang F, Qiao J. Tracing the expression of circular RNAs in human pre-implantation embryos. Genome Biol 17: 130, 2016. doi:10.1186/s13059-016-0991-3.

47. Hendrickson PG, Doráis JA, Grow EJ, Whiddon JL, Lim JW, Wike CL, Weaver BD, Pflueger C, Emery BR, Wilcox AL, Nix DA, Peterson CM, Tapscott SJ, Carrell DT, Cairns BR. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat Genet 49: 925–934, 2017. doi:10.1038/ng.3844.

48. Wang X, Liu D, He D, Suo S, Xia X, He X, Han JJ, Zheng P. Transcriptome analyses of rhesus monkey preimplantation embryos reveal a reduced capacity for DNA double-strand break repair in primate oocytes and early embryos. Genome Res 27: 567–579, 2017 [Erratum in Genome Res 27: 1621.2., 2017]. doi:10.1101/gr.198044.115.

49. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, Huang Y. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol 16: 148, 2015. doi:10.1186/s13059-015-0706-1.

50. Liu W, Liu X, Wang C, Gao Y, Gao R, Kou X, Zhao Y, Li J, Wu Y, Xiu W, Wang S, Yin J, Liu W, Cai T, Wang H, Zhang Y, Gao S. Identification of key factors conquering developmental arrest of somatic cell cloned embryos by combining embryo biopsy and single-cell sequencing. Cell Discov 2: 16010, 2016. doi:10.1038/celldisc. 2016.10.

51. Biase FH, Cao X, Zhong S. Cell fate inclination within 2-cell and 4- cell mouse embryos revealed by single-cell RNA sequencing. Genome Res 24: 1787–1796, 2014. doi:10.1101/gr.177725.114.

52. Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc Natl Acad Sci USA 111: 4139–4144, 2014. doi:10.1073/pnas.1321569111.

53. Jiang Z, Sun J, Dong H, Luo O, Zheng X, Obergfell C, Tang Y, Bi J, O'Neill R, Ruan Y, Chen J, Tian XC. Transcriptional profiles of bovine in vivo pre-implantation development. BMC Genomics 15: 756, 2014.

54. Zhao XM, Cui LS, Hao HS, Wang HY, Zhao SJ, Du WH, Wang D, Liu Y, Zhu HB. Transcriptome analyses of inner cell mass and trophectoderm cells isolated by magnetic-activated cell sorting from bovine blastocysts using single cell RNA-seq. Reprod Domest Anim 51: 726–735, 2016. doi:10.1111/rda.12737.

55. Kajdasz A, Warzych E, Derebecka N, Madeja ZE, Lechniak D, Wesoly J, Pawlak P. Lipid stores and lipid metabolism associated gene expression in porcine and bovine parthenogenetic embryos revealed by fluorescent staining and RNA-seq. Int J Mol Sci 21: 6488, 2020. doi:10.3390/ijms21186488.

56. van der Weijden VA, Schmidhauser M, Kurome M, Knubben J, Floter VL, Wolf E, Ulbrich SE. Transcriptome dynamics in early in vivo developing and in vitro produced porcine embryos. BMC Genomics 22: 139, 2021. doi:10.1186/s12864-021-07430-7.

57. Liu X, Hao Y, Li Z, Zhou J, Zhu H, Bu G, Liu Z, Hou X, Zhang X, Miao YL. Maternal cytokines CXCL12, VEGFA, and WNT5A promote porcine oocyte maturation via MAPK activation and canonical WNT inhibition. Front Cell Dev Biol 8: 578, 2020. doi:10.3389/fcell.2020.00578.

58. Zhong L, Mu H, Wen B, Zhang W, Wei Q, Gao G, Han J, Cao S. Long non-coding RNAs involved in the regulatory network during porcine pre-implantation embryonic development and iPSC induction. Sci Rep 8: 6649, 2018. doi:10.1038/s41598-018-24863-5.

59. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550, 2014. doi:10.1186/s13059-014-0550-8.

60. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 4: 1184–1191, 2009. doi:10.1038/nprot.2009.97.

61. Hart T, Komori HK, LaMere S, Podshivalova K, Salomon DR. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 14: 778, 2013. doi:10.1186/1471-2164-14-778.

62. Rau A, Marot G, Jaffrezic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics 15: 91, 2014. doi:10.1186/1471-2105-15-91.

63. Wickham H. ggplot2 Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag, 2016.

64. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32: 2847–2849, 2016. doi:10.1093/bioinformatics/btw313.

65. Wilke C, Fox SJ, Bates T, Manalo K, Lang B, Barrett M, Stoiber M, A P, Denney B, Hesselberth WJ, van der Bilj W, Grenie M, Selker R, Uhlitz F. cowplot (v1.1.1) wilkelab/cowplot: 1.1.1 (Version 1.1.1). Zenodo. http://doi.org/10.5281/zenodo.4411966. 2021, p. wilkelab/cowplot: 1.1.1 (Version 1.1.1). Zenodo.

66. Lombard DB, Tishkoff DX, Bao J. Mitochondrial sirtuins in the regulation of mitochondrial activity and metabolic adaptation. Handb Exp Pharmacol 206: 163–188, 2011. doi:10.1007/978-3-642-21631-2_8.

67. Gardner DK, Harvey AJ. Blastocyst metabolism. Reprod Fertil Dev 27: 638–654, 2015. doi:10.1071/RD14421.

68. Houghton FD, Thompson JG, Kennedy CJ, Leese HJ. Oxygen consumption and energy metabolism of the early mouse embryo. Mol Reprod Dev 44: 476–485, 1996. doi:10.1002/(SICI)1098-2795(199608) 44:43.0.CO;2-I.

69. Lopes AS, Madsen SE, Ramsing NB, Løvendahl P, Greve T, Callesen H. Investigation of respiration of individual bovine embryos produced in vivo and in vitro and correlation with viability following transfer. Hum Reprod 22: 558–566, 2007. doi:10.1093/humrep/del404.

70. Trimarchi JR, Liu L, Porterfield DM, Smith PJ, Keefe DL. Oxidative phosphorylation-dependent and -independent oxygen consumption by individual preimplantation mouse embryos. Biol Reprod 62: 1866–1874, 2000. doi:10.1095/biolreprod62.6.1866.

71. Kawamura Y, Uchijima Y, Horike N, Tonami K, Nishiyama K, Amano T, Asano T, Kurihara Y, Kurihara H. Sirt3 protects in vitro-fertilized mouse preimplantation embryos against oxidative stress-induced p53-mediated developmental arrest. J Clin Invest 120: 2817–2828, 2010. doi:10.1172/JCI42020.

72. Khadrawy O, Gebremedhn S, Salilew-Wondim D, Rings F, Neuhoff C, Hoelker M, Schellander K, Tesfaye D. Quercetin supports bovine preimplantation embryo development under oxidative stress condition via activation of the Nrf2 signaling pathway. Reprod Domest Anim 55: 1275–1285, 2020. doi:10.1111/rda.13688.

73. Meistermann D, Bruneau A, Loubersac S, Reignier A, Firmin J, François-Campion V, Kilens S, Lelievre Y, Lammers J, Feyeux M, Hulin P, Nedellec S, Bretin B, Castel G, Allegre N, Covin S, Bihouee A, Soumillon M, Mikkelsen T, Barriere P, Chazaud C, Chappell J, Pasque V, Bourdon J, Freour T, David L. Integrated pseudotime analysis of human pre-implantation embryo single-cell transcriptomes reveals the dynamics of lineage specification. Cell Stem Cell 28: 1625–1640.e6, 2021. doi:10.1016/j.stem.2021.04.027.

74. Hou Q, Gorski J. Estrogen receptor and progesterone receptor genes are expressed differentially in mouse embryos during preimplantation development. Proc Natl Acad Sci USA 90: 9460–9464, 1993. doi:10.1073/pnas.90.20.9460.

75. Hou Q, Paria BC, Mui C, Dey SK, Gorski J. Immunolocalization of estrogen receptor protein in the mouse blastocyst during normal and delayed implantation. Proc Natl Acad Sci USA 93: 2376–2381, 1996. doi:10.1073/pnas.93.6.2376.

76. Saito K, Furukawa E, Kobayashi M, Fukui E, Yoshizawa M, Matsumoto H. Degradation of estrogen receptor a in activated blastocysts is associated with implantation in the delayed implantation mouse model. Mol Hum Reprod 20: 384–391, 2014. doi:10.1093/molehr/gau004.

77. Yu LL, Zhang JH, He YP, Huang P, Yue LM. Fast action of estrogen on intracellular calcium in dormant mouse blastocyst and its possible mechanism. Fertil Steril 91: 611–615, 2009. doi:10.1016/j.fertnstert. 2007.11.072.

78. Yu LL, Qu T, Zhang SM, Yuan DZ, Xu Q, Zhang JH, He YP, Yue LM. GPR30 mediates the fast effect of estrogen on mouse blastocyst and its role in implantation. Reprod Sci 22: 1312–1320, 2015. doi:10.1177/1933719115578921.

79. Hadas R, Gershon E, Cohen A, Elbaz M, Ben-Dor S, Kohen F, Dekel N, Neeman M. Production of hyaluronan by the trophectoderm is a prerequisite for mouse blastocyst attachment. bioRxiv, 2020. doi:10.1101/2020.03.27.012880.

80. Mohamed OA, Dufort D, Clarke HJ. Expression and estradiol regulation of Wnt genes in the mouse blastocyst identify a candidate pathway for embryo-maternal signaling at implantation. Biol Reprod 71: 417–424, 2004. doi:10.1095/biolreprod.103.025692.

81. Niemann H, Elsaesser F. Evidence for estrogen-dependent blastocyst formation in the pig. Biol Reprod 35: 10–16, 1986. doi:10.1095/biolreprod35.1.10.

82. Valbuena D, Martin J, de Pablo JL, Remohí J, Pellicer A, Simon C. Increasing levels of estradiol are deleterious to embryonic implantation because they directly affect the embryo. Fertil Steril 76: 962– 968, 2001. doi:10.1016/s0015-0282(01)02018-0.

83. Yamamura S, Goda N, Akizawa H, Kohri N, Balboula AZ, Kobayashi K, Bai H, Takahashi M, Kawahara M. Yes-associated protein 1 translocation through actin cytoskeleton organization in trophectoderm cells. Dev Biol 468: 14–25, 2020. doi:10.1016/j.ydbio.2020.09.004.

84. Shi Y, Zhao P, Dang Y, Li S, Luo L, Hu B, Wang S, Wang H, Zhang K. Functional roles of the chromatin remodeler SMARCA5 in mouse and bovine preimplantation embryos. Biol Reprod 105: 359–370, 2021. doi:10.1093/biolre/ioab081.

85. Chitwood KL, Burrel VR, Halstead M, Meyers SA, Ross PJ. Transcriptome profiling of individual rhesus macaque oocytes and preimplantation embryos. Biol Repord. 97:353-364. 2017. Doi:10.1093/biolre/iox114

CHAPTER 5.

OVERALL CONCLUSIONS AND FUTURE DIRECTIONS

In these works, the results find that at the transcriptome level, there are relatively few shared features across mammalian species, particularly at the level of individual genes with respect to mRNA regulation. When comparing the impact of species-specific and shared mRNAs major changes, however, there is substantial conservation at the pathway and functional level. Specifically, during oocyte maturation, any one of the four input mammalian species (human, rhesus, mouse, and cow) has a minimum of 2000 mRNAs with significant changes in relative abundance, but with roughly 100 shared between species (1). And even when loosening the restrictions to being shared in just 3 of 4 species, the total is still near just 1000. However, when processing individual species mRNAs with significant changes in relative abundance, there are a great number of pathways and functions conserved. Of note, from the highly-degraded mRNAs, there is conserved inhibition for pathways relating to mitochondrial function, oxidative phosphorylation, and NRF2 mediating oxidative stress. Several shared DEGs were especially populated within the mitochondrial complex I, and species-specific DEGs within complex IV (1). These findings suggest that the decreased function of oxidative phosphorylation could be protective in nature, limiting mitochondrial activity and thereby decreasing reactive oxygen species.

When analyzing the 3’ UTR and associated RBPs impacting stability, the results continue the trend of sparse conservation at an individual RBP and gene level, but greater conservation at the level of function. Specifically, the poly(U) binding RBPs (CPEB2, CPEB4, and U2AF2) targeted high feature value mRNAs in a species-specific manner, but signaling pathways,

functions relating to cytoplasm, inositol metabolism, and cell cycle regulation had significant enrichment across at least a subset of the species.

A similar trend was seen during the MBT; few shared DEGs, greater shared functional categories. Reciprocal to the findings during oocyte maturation, oxidative phosphorylation was increased in four of the five mammalian species, essentially reversing the diminishment of activity that emerged during oogenesis (2). This activation may indicate the importance of mitochondrial function and oxidative phosphorylation in energy metabolism for the formation of blastocysts, cellular reprogramming, and preparation for implantation.

While the three chapters in concert provide a rich resource covering two key developmental events in mammalian reproduction, there are limitations. On the level of the transcriptome, the human and mouse genomes are well characterized, but there is a certain level of deficiency in genome build and annotation across the other species. With the advancements of third-generation long-read sequencing (3) (e.g., PacBio (4) and oxford nanopore (5)), the genome build quality will increase in resolution, which would warrant revisiting these studies. As sequencing data of additional species are deposited in the public domain, the cross-species comparisons can increase. Indeed, after the completion of the chapter on oocyte maturation, porcine data became available, as noted in its inclusion in the chapter covering RBPs and 3' UTR binding. This methodology can also be extrapolated to integrate non-mammalian species with similar cell staging and developmental events. Beyond the same staging, the meta-analysis can also be leveraged to compare additional developmental stages, such as fertilization (MII versus 1-cell), embryonic genome activation, morula compaction, etc. This methodology, however, is not constrained to developmental stages. Assuming sufficient data across species,

treatment and conditional based studies can be processed (e.g., tumor versus normal, treated versus non-treated, etc.).

A major point of advancement is the interrogation of RBP binding via the application of machine learning, feature selection, regression algorithms, and SHAP scores. A cursory search of PubMed (as of 11/2021), with the terms “RNA sequencing” + “SHAP or SHAPley”, garnered only three results. These three studies, covering the years 2019-2021, pertain to computational predictions of ribosomal entry (6), tissue classifier (7), and longevity-association (8). This is clearly an untapped computational resource in the sequencing field. Beyond the previously described transcriptome deficiencies, the RBP analysis was also not without its limitations. The included proteomic studies had publication years ranging from 2010-2020 (9-13), and while there have been advancements in proteomic analysis, the resolution is still lacking behind RNA sequencing (14). Updated proteomic data pertaining to the included species, and expansion to missing species would be beneficial to ascertain which RBPs are expressed in the oocyte. Currently, the data presented for the RBP study was limited to singular RBPs. This can be expanded to explore combinations of RBP sites within the 3’ UTR, possibly leveraging a Market Basket Analysis (MBA) (15). MBAs were developed for the retail and restaurant space, identifying products and items often purchased together. One could easily imagine adopting this method to identify which RBPs are often found together for a given mRNA and then extrapolate across stability classes. An additional wrinkle would be to factor in both proximity of RBP sites to each other and the physical location of the sites, i.e., closer to the transcription start site or closer to the polyadenylation site. This would necessitate using a combination of computational methods: the SHAP analysis, MBA, and a sliding window feature to capture all available information.

Another future analysis would entail tracking the identified stabilized maternal mRNAs during oocyte maturation and exploring their further relative abundances through fertilization and embryonic genome activation. Which of these transcripts continue to exhibit increases in relative abundance and which show future degradation? During fertilization and genome activation, which RBP factors, new or previously identified, continue to imbue stabilizing or destabilizing effects?

Outside the computational field, the binding sites of high value RBPs can be further interrogated with the application of RNA immunoprecipitation sequencing (RIP-seq) (16) to verify and augment the computational predicted data. With the advancements of machine learning algorithms in the data science fields, the possibilities to port their applications to the biological field, and specifically transcriptome data, is nearly limitless. The results and methods of studies reported in this thesis lay the groundwork, provide a proof on concept, and a rich resource upon which further hypothesis testing projects can be developed.

REFERENCES

REFERENCES

1. Schall PZ, Latham KE. Essential shared and species-specific features of mammalian oocyte maturation-associated transcriptome changes impacting oocyte physiology. Am J Physiol Cell Physiol. 2021 Jul 1;321(1):C3-C16. doi: 10.1152/ajpcell.00105.2021. Epub 2021 Apr 21. PMID: 33881934; PMCID: PMC8321790.

2. Schall PZ, Latham KE. Cross-species meta-analysis of transcriptome changes during the morula-to-blastocyst transition: metabolic and physiological changes take center stage. Am J Physiol Cell Physiol. 2021 Dec 1;321(6):C913-C931. doi: 10.1152/ajpcell.00318.2021. Epub 2021 Oct 20. PMID: 34669511.

3. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018 Sep;34(9):666-681. doi: 10.1016/j.tig.2018.05.008. Epub 2018 Jun 22. PMID: 29941292.

4. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278-89. doi: 10.1016/j.gpb.2015.08.002. Epub 2015 Nov 2. PMID: 26542840; PMCID: PMC4678779.

5. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. Dev Growth Differ. 2019 Jun;61(5):316-326. doi: 10.1111/dgd.12608. Epub 2019 Apr 29. PMID: 31037722.

6. Wang J, Gribskov M. IRESpy: an XGBoost model for prediction of internal ribosome entry sites. BMC Bioinformatics. 2019 Jul 30;20(1):409. doi: 10.1186/s12859-019-2999-7. PMID: 31362694; PMCID: PMC6664791.

7. Yap M, Johnston RL, Foley H, MacDonald S, Kondrashova O, Tran KA, Nones K, Koufariotis LT, Bean C, Pearson JV, Trzaskowski M, Waddell N. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. Sci Rep. 2021 Jan 29;11(1):2641. doi: 10.1038/s41598-021-81773-9. PMID: 33514769; PMCID: PMC7846764.

8. Kulaga AY, Ursu E, Toren D, Tyshchenko V, Guinea R, Pushkova M, Fraifeld VE, Tacutu R. Machine Learning Analysis of Longevity-Associated Gene Expression Landscapes in Mammals. Int J Mol Sci. 2021 Jan 22;22(3):1073. doi: 10.3390/ijms22031073. PMID: 33499037; PMCID: PMC7865694.

9. Virant-Klun I, Leicht S, Hughes C, Krijgsveld J. Identification of Maturation-Specific Proteins by Single-Cell Proteomics of Human Oocytes. Mol Cell Proteomics. 2016 Aug;15(8):2616-27. doi: 10.1074/mcp.M115.056887. Epub 2016 May 23. PMID: 27215607; PMCID: PMC4974340.

10. Cao S, Huang S, Guo Y, Zhou L, Lu Y, Lai S. Proteomic-based identification of oocyte maturation-related proteins in mouse germinal vesicle oocytes. Reprod Domest Anim. 2020 Nov;55(11):1607-1618. doi: 10.1111/rda.13819. Epub 2020 Oct 14. PMID: 32920902.

11. Pfeiffer MJ, Taher L, Drexler H, Suzuki Y, Makałowski W, Schwarzer C, Wang B, Fuellen G, Boiani M. Differences in embryo quality are associated with differences in oocyte composition: a proteomic study in inbred mice. Proteomics. 2015 Feb;15(4):675-87. doi: 10.1002/pmic.201400334. Epub 2015 Jan 3. PMID: 25367296.

12. Jia B, Xiang D, Fu X, Shao Q, Hong Q, Quan G, Wu G. Proteomic Changes of Porcine Oocytes After Vitrification and Subsequent in vitro Maturation: A Tandem Mass Tag-Based Quantitative Analysis. Front Cell Dev Biol. 2020 Dec 23;8:614577. doi: 10.3389/fcell.2020.614577. PMID: 33425922; PMCID: PMC7785821.

13. Peddinti D, Memili E, Burgess SC. Proteomics-based systems biology modeling of bovine germinal vesicle stage oocyte and cumulus cell interaction. PLoS One. 2010 Jun 21;5(6):e11240. doi: 10.1371/journal.pone.0011240. PMID: 20574525; PMCID: PMC2888582.

14. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH. Proteomics: Technologies and Their Applications. J Chromatogr Sci. 2017 Feb;55(2):182-196. doi: 10.1093/chromsci/bmw167. Epub 2016 Oct 18. PMID: 28087761.

15. Blattberg R.C., Kim BD., Neslin S.A. (2008) Market Basket Analysis. In: Database Marketing. International Series in Quantitative Marketing, vol 18. Springer, New York, NY. https://doi.org/10.1007/978-0-387-72579-6_13

16. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell. 2010 Dec 22;40(6):939-53. doi: 10.1016/j.molcel.2010.12.011. PMID: 21172659; PMCID: PMC3021903.