AN INTEGRATIVE GENOMICS APPROACH TO IDENTIFYING GENETIC REGULATORS OF PREMATURITY

By

Haley Rae Eidem

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biological Sciences

May 11, 2018

Nashville, Tennessee

Approved:

Antonis Rokas, Ph.D.
Douglas K. Abbot, Ph.D.
John A. Capra, Ph.D.
James G. Patton, Ph.D.
William E. Ackerman III, M.D.

ACKNOWLEDGEMENTS

First, I'd like to thank Jim Patton and the rest of the Vanderbilt University Interdisciplinary Graduate Program. Jim was instrumental in bringing me to Vanderbilt and, for that, I'll always be grateful.

I'd also like to thank the Vanderbilt University Department of Biological Sciences, including Leslie Maxwell, Angela Titus, and Christopher Patterson for their administrative assistance, LaDonna Smith for guiding me through the graduation process, Carol Wiley for grant application expertise, Alicia Goostree for helping whenever I needed it, Steve Baskauf for his teaching mentorship, and Kathy Friedman for keeping me on track.

My work has been generously supported by Vanderbilt University, the Gisela Mosig Travel Fund, the March of Dimes Prematurity Research Center Ohio Collaborative Transdisciplinary Scholar Award, and the Burroughs Wellcome Fund. Furthermore, I've had the opportunity to work with many talented collaborators, both at Vanderbilt and elsewhere, including coauthors Bill Ackerman, Kris McGary, Patrick Abbot, Antonis Rokas, Tony Capra, David Rinker, Irina Buhimschi, Catalin Buhimschi, Caitlin Dunn-Fletcher, Suhas Kallapur, Mihaela Pavlicev, Lou Muglia, Jacob Steenwyk, and Jen Wisecaver as well as colleagues Alison Paquette and Ge Zhang.

The members of my thesis committee have provided enormous support and guidance in all of my projects. Thank you to Patrick Abbot, Jim Patton, Tony Capra, and Bill Ackerman for your time, insight, and encouragement over the years. My research has greatly benefitted from your involvement.

I'd like to recognize my labmates in the Rokas and Abbot labs who have been there for both help and camaraderie, including Jen Wisecaver, Julie Phillips, Matt Mead, Abigail Lind, Mara Kim, Michelle Moon, Juan Ortiz, Kris McGary, Xiaofan Zhou, Yuyu Wang, Xingxing Shen, Mingshuan Wang, Jibril Hirbo, Stephanie Birnbaum, Jacob Steenwyk, David Rinker, Leo Salichos, Patricia Soria, Kitae Kim, Sean King, Brian Cooper, Rebecca Burke-Aguero, Alex Borowsky, Sam Smith, Zack Ely, Maddison Johnson, Rohit Venkat, Ken Polzin, George Greene, Holly Elmore, Abbe LaBella, and Youhuang Bai. Jen, I'll never forget the lattes, happy hours, and everything in between. Julie, I'm so thankful for your friendship on this literal and figurative rollercoaster.

I am especially grateful to my advisor, Antonis Rokas, who has been an outstanding mentor in science and in life. Antonis gave me the freedom to follow my passions while also trusting my abilities as a scientist. He is a constant source of optimism and can create a story out of any results. His family became my friends and his enthusiasm, patience, and support made this work possible. Thank you.

I don't think I'll ever be able to fully express the depth of my gratitude to Jonna Whitman, Deana Hasbrooke, Student Health, and Vanderbilt University Medical Center for helping me survive pulmonary emboli. Dr. Whitman and Deana, thank you for taking such wonderful care of me; I wouldn't be where I am today without you.

Finally, I am forever grateful to my family and friends for their endless encouragement. To my best friend Anna, thank you for 25 years of laughs. To Christine and Micaela (and Otto and Keeper), thank you for being there through the highs and lows. To Brittany (and Charlie and Lily), thank you from the bottom of my heart. Lastly, to Mom, Dad, and Taylor (and Birdie), thank you for your support, pride, and love. It means the world.

Boom. Roasted.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER II

CHAPTER III

LIST OF FIGURES

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER V

CHAPTER I

INTRODUCTION

Incidence

The World Health Organization defines preterm birth (PTB) in humans as birth before 37 completed weeks of a typically 40-week gestation (Figure 1.1). Globally, 15 million babies, or about 1 in 10, are born preterm every year and that number is rising[1]. Prematurity is the leading cause of death in children under 5 worldwide and its complications are responsible for the deaths of approximately 1 million children annually[1]. Rates vary both between and within countries, but incidence continues to rise across most of the developed world despite advances in biological knowledge. Public health and medical interventions designed to delay elective cesarean sections have also failed to reduce the number of preterm births[2]. Babies born too soon are predisposed to suffer from neurodevelopmental, respiratory, gastrointestinal, and other complications throughout their lives. Furthermore, medical care for mothers and babies before, during, and long after preterm births is costly, adding to a yearly total of $26 billion[3].

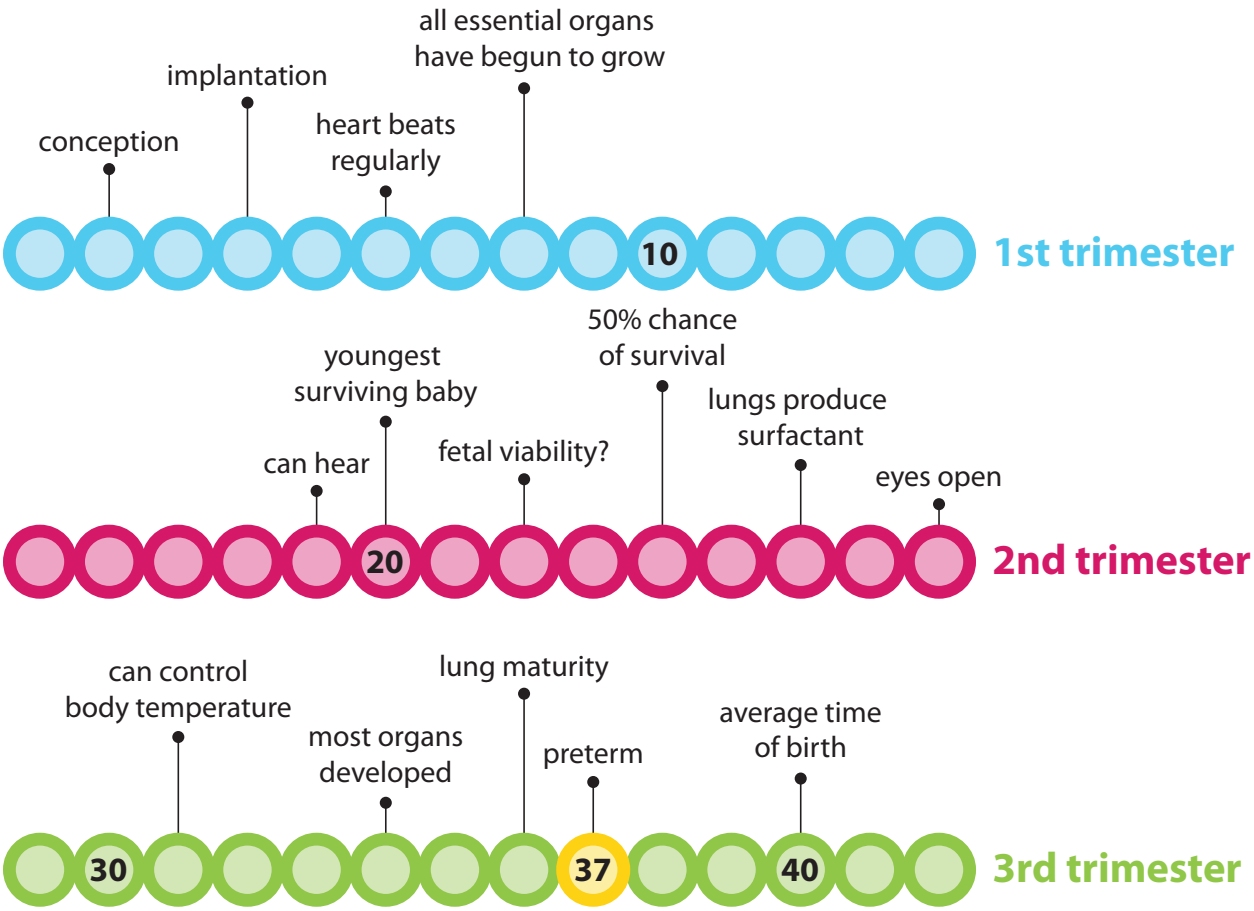

Figure 1.1. Human pregnancy timeline. Normal human pregnancy lasts approximately 40 weeks. Delivery is considered preterm before 37 completed weeks of gestation.

Subtypes

PTB can be caused by 1) the medically indicated induction of preterm delivery due to either maternal or fetal complications, 2) spontaneous, idiopathic preterm labor with intact fetal membranes (sPTB), or 3) preterm premature rupture of membranes (PPROM). Approximately 30% of all preterm births are medically indicated due to preeclampsia (PE), intrauterine growth restriction (IUGR), gestational diabetes mellitus (GDM), chorioamnionitis, or other

complications. The other 70% of preterm births occur spontaneously, with 45% due to sPTB and 25% due to PPROM[1].

Disparities

PTB is a global problem with generally higher rates in low-income countries (~12%) and lower rates in high-income countries (~9%), although rates can vary from ~5% in some northern European countries to ~18% in Malawi[2,4]. Despite this general trend, several high-income countries have much higher rates of PTB than expected. In the United States, for example, the rate of preterm birth has increased by 30% since 1981, reaching its peak at 12.5% in 2006, decreasing for several years after that, and increasing again in 2015 and 2016, with the rate currently reported at 9.6%[2]. Of the 1.2 million preterm births that occur in high-income countries, the United States contributes more than 0.5 million (42%) of them[2].

Major racial and ethnic disparities are also evident in preterm birth. In the United States, the PTB rate among black women is twice as high as the rate among white women, even after adjusting for other confounding factors[5-7]. Furthermore, black women are 3-4 times more likely to have a very early PTB (before 32 weeks) than women from any other ethnic groups[1,8]. Rates of PTB are also higher among black women born in the United States than they are among black women born outside of the United States, but both are higher than the rate among white women[9,10]. Conversely, East Asian and Hispanic women have relatively low rates

of PTB, while South Asian women have increased risk of low birth weight with no connection to PTB risk [1].

Interestingly, PTB rates for many US immigrant groups are positively correlated with the length of time they have lived in the United States[11]. Epidemiological studies have linked poverty, education, age, marital status, prenatal care, and other factors with PTB incidence and, although the mechanisms for their involvement are unknown, these factors presumably influence PTB risk across ethnic groups. Preterm birth rates also vary by state, and sometimes even by neighborhood, in the United States and these differences could reflect the effects of socioeconomic factors, as well[6,12]. Differences in PTB risk are also evident in maternal age, where women under 20 and over 40 are more likely to deliver preterm, although these disparities differ between ethnic groups, with the PTB risk for black women rising at a younger age than it does for women of other ethnicities[13].

Risk factors

In addition to racial and ethnic disparities, many other risk factors have been linked to PTB incidence (Figure 1.2). Pregnancy history has been shown to have an impact on PTB risk, for example, with risk increasing as inter-pregnancy interval decreases[14]. Although the mechanism for recurrence is unclear, it could be related to persistent infection, nutrition, or chronic health conditions. Multiple gestations also increase risk of PTB where almost 60% of twin births and nearly all of higher order births are delivered preterm[15]. Many maternal

medical conditions have been shown to increase PTB risk including thyroid disease, asthma, hypertension, diabetes, history of cervical loop procedures, and depression[1]. Moreover, tobacco use increases PTB risk at least 2-fold[1].

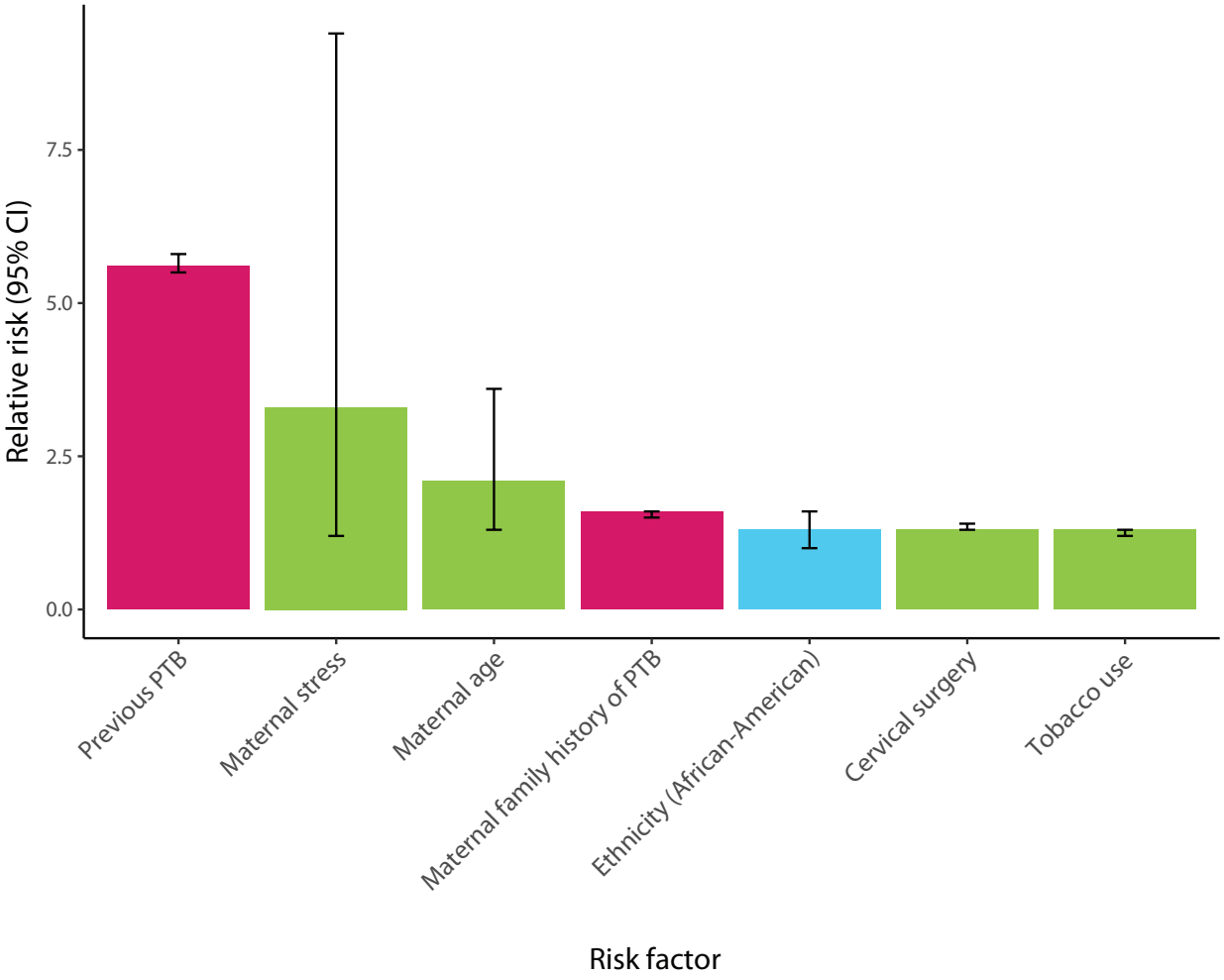

Figure 1.2. PTB risk factors. Risk factors for PTB stem from genetics (pink), environmental stress (green), and ethnicity (blue). Redrawn from Bezold et al. where relative risk and confidence intervals were obtained from previous independent PTB risk studies applying a variety of methods.[16-22]

A variety of infections have been linked to PTB including urinary tract infections, malaria, bacterial vaginosis, HIV, syphilis, and chorioamnionitis[23]. In fact, 1 out of every 4 preterm

births occur in the presence of an often subclinical intra-amniotic infection[23]. These infections typically present as chorioamnionitis, defined as inflammation of the chorion, amnion, and placenta[24]. Intrauterine infections can cause PTB by activating the innate immune system and triggering the release of inflammatory cytokines and chemokines that can then stimulate the production of prostaglandins, activating either uterine contractility and eventually sPTB or matrix-degrading enzymes and eventually PPROM[1].

Genetics

Several lines of evidence suggest that genetics also plays a role in prematurity. sPTB is a trait that appears to be maternally transmitted, where a woman's risk of delivering preterm is increased if her mother, full sisters, or maternal half-sisters have delivered preterm[17]. In fact, women with sisters who delivered preterm are 80% more likely to experience PTB[25]. Women who were born preterm themselves are also more likely to deliver preterm[26,27]. Interestingly, women who deliver post-term are also at an increased risk to deliver post-term again, suggesting that genetics plays a role in birth timing, in general[28]. Twin studies have demonstrated both maternal and fetal genetic contributions to variation in birth timing, with heritability estimated between 15% and 40%[29-32].

Candidate gene studies have been performed on pathways involved in immunity and inflammation as key regulators of sPTB pathogenesis. Most of this research focuses on inflammatory cytokines like tumor necrosis factor (*TNF*) as well as interleukins and their

receptors (*IL-1B*, *IL-2*, *IL-4*, *IL-6*, *IL-10*, *IL-1R*, *IL6R*), but results have been mixed and have generally failed to replicate[16,26,33-35]. Other studies have focused on matrix metalloproteinases involved in extracellular matrix metabolism (*MMP1*, *MMP9*), but again, results have failed to replicate across study populations[26]. Single candidate gene studies have also linked oxytocin (*OXY*) and vascular endothelial growth factor (*VEGF*) to sPTB, but neither have been replicated in larger cohorts[36,37]. Furthermore, family-based linkage studies have implicated *IGF1R*, *AR*, and *IL-2* in sPTB[38,39].

Genome-wide association studies have attempted to uncover polymorphisms associated with sPTB, but, until recently, had also failed to replicate in validation cohorts[40]. A 2017 GWAS, however, identified variants at the *EBF1*, *EEFSEC*, and *AGTR2* loci significantly associated with sPTB[41]. This was the largest PTB-related GWAS to date and, because of its size, it is the only of its kind to identify significant associations.

Despite this growing body of research, however, the common pathways influencing sPTB pathogenesis and their genetic regulators remain poorly understood.

Biomarkers

In addition to genetic factors contributing to sPTB, researchers have also identified several predictive biomarkers through protein and gene expression studies. In fact, a recent meta-analysis identified 42 biomarkers associated with sPTB through a variety of multiplex assays[42].

Like in candidate gene studies (and as no coincidence), the majority of these biomarkers are cytokines, chemokines, metalloproteinases, and growth factors involved in inflammatory pathways. Some of these molecules include RANTES, IL-10, Eotaxin, and TNF-RI identified in maternal serum, ANGPT2 identified in amniotic fluid, and ICAM-1, IGF-I, IL-1B, IL-1Ra, IL-8, MCP-3, MIP-1a, PDGF-BB, TGF-a, TGF-B1, TIMP1, TNFa, TNFR-I, TNFR-II, and VEGF identified in fetal plasma[42].

In general, PTB is now thought to be a complex, multifactorial syndrome associated with multiple mechanisms of disease and can be caused by the interaction of many of the aforementioned racial, ethnic, socioeconomic, environmental, and genetic risk factors. Although 45% of all preterm births are spontaneous, this subtype remains relatively understudied compared to preeclampsia and other diseases of pregnancy[43]. Therefore, the following dissertation focuses on identifying the genetic regulators of sPTB, specifically.

Evolution

PTB has been observed in many other species, indicating that it is not a human-specific phenomenon[44]. The use of animal models has shed additional light on the mechanisms regulating parturition, but this insight is limited because many pregnancy-associated traits evolve rapidly and are different between species. The placenta, for example, is highly variable across mammals in invasiveness, shape, and interdigitation, even in closely related primate species (Figure 1.3)[45-47]. Furthermore, placental morphology and other pregnancy-associated

traits may have evolved rapidly alongside bipedalism and increased cranial size in the human lineage[48]. This evolution could be influenced by maternal-fetal conflict, maternal energy restriction, or cephalopelvic constraint[49].

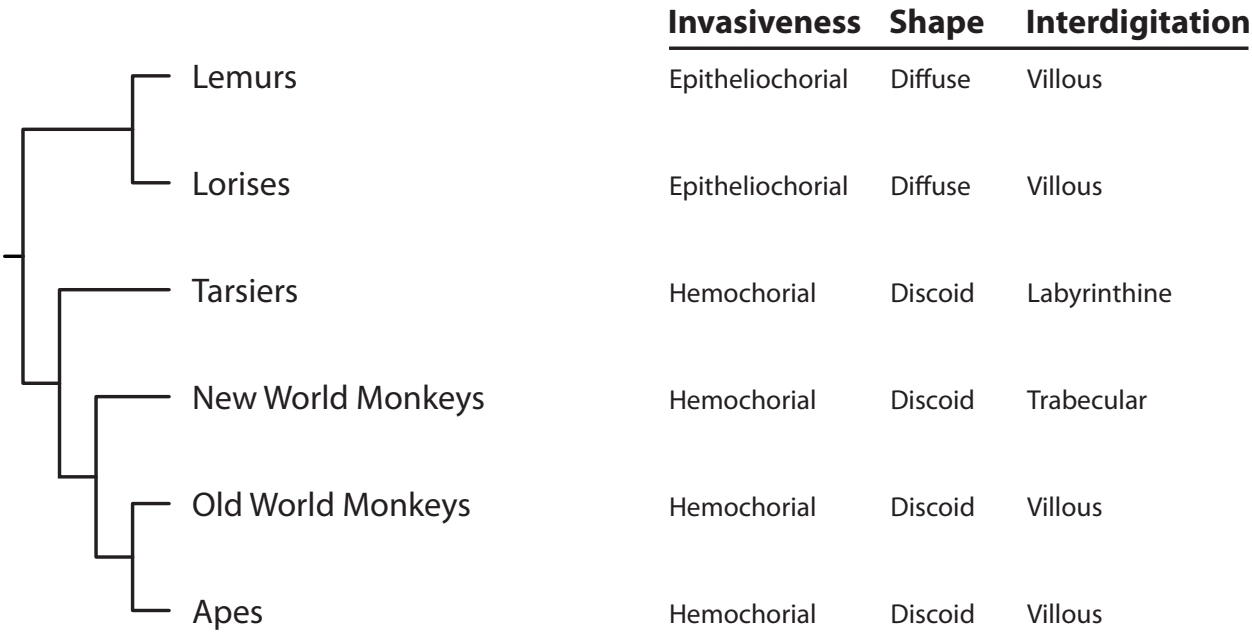

Figure 1.3. Primate placental diversity. A high degree of variation exists in invasiveness, shape, and interdigitation of placenta, even across closely related primate species.

The changing hormonal landscape during pregnancy differs across mammals, as well, adding another layer of complexity to utilizing model species in the study of PTB[50]. Progesterone, in particular, is key in facilitating and maintaining pregnancy, and its levels rise until the time of delivery in humans and great apes[51]. In contrast, Old World monkeys have lower, unchanging levels of progesterone throughout pregnancy and New World monkeys show progesterone withdrawal[51]. Despite these numerous differences, models in sheep, mouse, rat, guinea pigs, and a handful of other species have been used to further our understanding of the events leading up to parturition and PTB[16,52].

Challenges

High levels of heterogeneity exist in the study of sPTB associated with both etiology (e.g., multiple subtypes) as well as pathophysiology (e.g., multiple causative biological pathways), making it difficult to identify and replicate genetic risk factors or proteomic biomarkers. Multiple tissues in the mother (e.g., decidua, myometrium, cervix, etc.) and fetus (e.g., amnion, chorion, umbilical cord, etc.) must interact in order to facilitate nutrient transfer and maintain pregnancy[47]. Moreover, both the maternal and fetal genomes interact to confer sPTB risk[47]. Pregnancy is a uniquely challenging biological system, with time-sensitive dynamics that change dramatically over a period of about 9 months, and this challenge is reflected in the relative lack of biological insight into pathways whose dysregulation lead to sPTB and other pregnancy pathologies. Advances in high-throughput 'omics experiments have allowed for the analysis of sPTB pathogenesis in terms of genomic, transcriptomic, proteomic, metabolomic, and epigenomic variation. However, when employed in isolation, any single 'omics approach can only provide limited insight into what we know is a highly complex biological system. Therefore, understanding the pathways that lead to sPTB pathogenesis will likely require the integration of information across 'omics layers and the study of their combined influences.

Chapter Previews

In this dissertation, I use meta-analysis, comparative transcriptomics, and integrative genomics approaches to map the pathways leading to prematurity, especially in sPTB.

In chapter II, I synthesize the current landscape in transcriptomics research across all subtypes of PTB. Because PTB research is so heterogeneous, a systematic approach to aggregating relevant data is essential to summarizing our current understanding of gene expression in PTB. To accomplish this, I first conducted a systematic literature review of all genome-wide pregnancy studies across gestational tissue types and pathologies, including sPTB. Second, I performed a meta-analysis of all publicly available, genome-wide transcriptomics analyses to identify genes commonly reported as differentially expressed across all represented phenotypes. This analysis uncovered gaps in PTB research and examined the set of currently available PTB candidate genes. A key conclusion of this chapter is that, although sPTB makes up the majority of all PTB cases, only 18% of transcriptomic research focuses on that subtype.

In chapter III, I seek to begin filling that gap in sPTB-specific research by comparing gene expression profiles from placental tissue collected after human sPTB and term deliveries to identify genes whose dysregulation may contribute to sPTB pathogenesis. These types of comparisons are complicated, however, by the necessary difference in gestational age of the tissue samples (e.g., <37 weeks vs. 40 weeks). In humans, it is impossible to collect 'healthy' placental tissue from ~37 weeks of gestation to compare with placental tissue from sPTB

cases. Therefore, differentially expressed genes could reflect sPTB pathology or simply differences due to gestational age. To disentangle these components, I also compare gene expression profiles from a closely related species with similar placental and pregnancy characteristics, Rhesus macaque. By using data from placental tissue collected at a similar gestational age (80% completed gestation) in macaque, I identify sPTB-specific differentially expressed genes as well as those related to gestational age differences.

In chapter IV, I expand beyond transcriptomics to discuss sPTB as a multifactorial syndrome influenced by diverse biological pathways and likely to benefit from an integrative approach that synthesizes equally diverse data. I review the complexities introduced by many gestational tissues and multiple genomes (e.g., maternal and fetal) that must interact to maintain a healthy pregnancy. I also discuss the rapid evolution of pregnancy-related traits as well as the challenges and opportunities this presents for integrative research.

In chapter V, I develop a software approach (integRATE) for integrating the types of diverse 'omics data mentioned in chapter IV and apply it to sPTB studies. This desirability function-based method allows for the integration of heterogeneous 'omics data collected from diverse platforms and samples, as is the case across sPTB research. integRATE uses desirability functions to rank genes both within and across studies based on their cumulative weight of evidence provided by the different types of 'omics data at hand. By applying this framework to sPTB 'omics data (including the transcriptomics data discussed in chapter III as well as

GWAS, proteomics, and epigenomics data), I refine our understanding of sPTB candidate genes supported by the current literature as well as advocate for further functional testing on biological pathways containing these genes.

Collectively, my dissertation uses transcriptomics as well as integrative genomics approaches to identify genetic regulators of sPTB. This work outlines unique challenges in studying sPTB and proposes next steps based on integrating the heterogeneous 'omics data at our disposal.

References

1. Goldenberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R. Epidemiology and causes of preterm birth. *Lancet* **371,** 75–84 (2008).
2. Blencowe, H. *et al.* Born too soon: the global epidemiology of 15 million preterm births. *Reprod Health* **10 Suppl 1,** S2 (2013).
3. Blencowe, H. *et al.* National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet* **379,** 2162–2172 (2012).
4. Liu, L. *et al.* Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals. *Lancet* **388,** 3027–3035 (2016).
5. Adams, M. M. *et al.* Preterm delivery among black and white enlisted women in the United States Army. *Obstetrics & Gynecology* **81,** 65–71 (1993).
6. Collins, J. W., David, R. J., Simon, D. M. & Prachand, N. G. Preterm birth among African American and white women with a lifelong residence in high-income Chicago neighborhoods: an exploratory study. *Ethn Dis* **17,** 113–117 (2007).
7. Kistka, Z. A. F. *et al.* Racial disparity in the frequency of recurrence of preterm birth. *American journal of obstetrics and gynecology* **196,** 131.e1–131.e6 (2007).
8. Muglia, L. J. & Katz, M. The enigma of spontaneous preterm birth. *N. Engl. J. Med.* **362,** 529–535 (2010).
9. Howard, D. L., Marshall, S. S., Kaufman, J. S. & Savitz, D. A. Variations in low birth weight and preterm delivery among blacks in relation to ancestry and nativity: New York City, 1998-2002. *Pediatrics* **118,** e1399–405 (2006).
10. Rosensweig, N. Differing birth weight among infants of U.S.-born blacks, African-born blacks, and U.S.-born whites. *J Nurse Midwifery* **43,** 129–131 (1998).
11. Guendelman, S. & English, P. B. Effect of United States residence on birth outcomes among Mexican immigrants: an exploratory study. *American journal of epidemiology* **142,** S30–8 (1995).
12. Farley, T. A. *et al.* The relationship between the neighbourhood environment and adverse birth outcomes. *Paediatr Perinat Epidemiol* **20,** 188–200 (2006).
13. Boardman, J. P. Preterm Birth: Causes, Consequences and Prevention. *Journal of Obstetrics and Gynaecology* **28,** 559–559 (2009).
14. Ananth, C. V. & Vintzileos, A. M. Epidemiology of preterm birth and its clinical subtypes. *J Matern Fetal Neonatal Med* **19,** 773–782 (2006).
15. Romero, R. *et al.* The preterm parturition syndrome. *BJOG* **113 Suppl 3,** 17–42 (2006).
16. Bezold, K. Y., Karjalainen, M. K., Hallman, M., Teramo, K. & Muglia, L. J. The genomics of preterm birth: from animal models to human studies. *Genome Med* **5,** 34 (2013).
17. Boyd, H. A. *et al.* Maternal contributions to preterm delivery. *American journal of epidemiology* **170,** 1358–1364 (2009).
18. Whitehead, N. & Helms, K. Racial and ethnic differences in preterm delivery among low-risk women. *Ethn Dis* **20,** 261–266 (2010).

19. Wadhwa, P. D. *et al.* Placental corticotropin-releasing hormone (CRH), spontaneous preterm birth, and fetal growth restriction: a prospective investigation. *The American Journal of Obstetrics & Gynecology* **191,** 1063–1069 (2004).
20. Aghamohammadi, A. & Nooritajer, M. Maternal age as a risk factor for pregnancy outcomes: Maternal, fetal and neonatal complication. *African Journal of Pharmacy and Pharmacology* **5,** 264–269 (2011).
21. Shah, N. R. & Bracken, M. B. A systematic review and meta-analysis of prospective studies on the association between maternal cigarette smoking and preterm delivery. *The American Journal of Obstetrics & Gynecology* **182,** 465–472 (2000).
22. Castanon, A. *et al.* Risk of preterm birth after treatment for cervical intraepithelial neoplasia among women attending colposcopy in England: retrospective-prospective cohort study. *BMJ* **345,** e5174–e5174 (2012).
23. Romero, R. *et al.* Infection and labor. V. Prevalence, microbiology, and clinical significance of intraamniotic infection in women with preterm labor and intact membranes. *International Journal of Gynecology & Obstetrics* **32,** 187–187 (2017).
24. Galinsky, R., Polglase, G. R., Hooper, S. B., Black, M. J. & Moss, T. J. M. The consequences of chorioamnionitis: preterm birth and effects on development. *J Pregnancy* **2013,** 412831–11 (2013).
25. Winkvist, A., Mogren, I. & Högberg, U. Familial patterns in birth characteristics: impact on individual and population risks. *Int J Epidemiol* **27,** 248–254 (1998).
26. Plunkett, J. & Muglia, L. J. Genetic contributions to preterm birth: implications from epidemiological and genetic association studies. *Ann. Med.* **40,** 167–195 (2008).
27. Porter, T. F., Fraser, A. M., Hunter, C. Y., Ward, R. H. & Varner, M. W. The risk of preterm birth across generations. *Obstetrics & Gynecology* **90,** 63–67 (1997).
28. Kistka, Z. A. F. *et al.* Risk for postterm delivery after previous postterm delivery. *American journal of obstetrics and gynecology* **196,** 241.e1–241.e6 (2007).
29. Clausson, B., Lichtenstein, P. & Cnattingius, S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG* **107,** 375–381 (2000).
30. Kistka, Z. A. F. *et al.* Heritability of parturition timing: an extended twin design analysis. *American journal of obstetrics and gynecology* **199,** 43.e1–5 (2008).
31. Treloar, S. A., Macones, G. A., Mitchell, L. E. & Martin, N. G. Genetic influences on premature parturition in an Australian twin sample. *Twin Res* **3,** 80–82 (2000).
32. York, T. P., Strauss, J. F., Neale, M. C. & Eaves, L. J. Estimating Fetal and Maternal Genetic Contributions to Premature Birth From Multiparous Pregnancy Histories of Twins Using MCMC and Maximum-Likelihood Approaches. *Twin Research and Human Genetics* **12,** 333–342 (2012).
33. Crider, K. S., Whitehead, N. & Buus, R. M. Genetic variation associated with preterm birth: a HuGE review. *Genet Med* (2005). doi:10.1097/01.gim.0000187223.69947.db
34. Menon, R. *et al.* Ethnic differences in key candidate genes for spontaneous preterm birth: TNF-alpha and its receptors. *Hum. Hered.* **62,** 107–118 (2006).
35. Menon, R. *et al.* Multilocus interactions at maternal tumor necrosis factor-alpha, tumor

necrosis factor receptors, interleukin-6 and interleukin-6 receptor genes predict spontaneous preterm labor in European-American women. *American journal of obstetrics and gynecology* **194,** 1616–1624 (2006).
36. Papazoglou, D., Galazios, G., Koukourakis, M. I., Kontomanolis, E. N. & Maltezos, E. Association of -634G/C and 936C/T polymorphisms of the vascular endothelial growth factor with spontaneous preterm delivery. *Acta Obstet Gynecol Scand* **83,** 461–465 (2004).
37. Kim, J. *et al.* Sequence variants in oxytocin pathway genes and preterm birth: a candidate gene association study. *BMC Medical Genetics* **14,** 77 (2013).
38. Alleman, B. W. *et al.* No observed association for mitochondrial SNPs with preterm delivery and related outcomes. *Pediatr. Res.* **72,** 539–544 (2012).
39. Karjalainen, M. K. *et al.* A Potential Novel Spontaneous Preterm Birth Gene, AR, Identified by Linkage and Association Analysis of X Chromosomal Markers. *PLoS ONE* **7,** e51378 (2012).
40. Zhang, H., Baldwin, D. A. & Bukowski, R. K. A Genome-Wide Association Study of Early Spontaneous Preterm Delivery. *Genetic …* (2015). doi:10.1002/gepi.21887
41. Zhang, G., Jacobsson, B. & Muglia, L. J. Genetic Associations with Spontaneous Preterm Birth. *N. Engl. J. Med.* **377,** 2401–2402 (2017).
42. Polettini, J. *et al.* Biomarkers of spontaneous preterm birth: a systematic review of studies using multiplex analysis. *Journal of Perinatal Medicine* **45,** 71–84 (2017).
43. Eidem, H. R., Ackerman, W. E., McGary, K. L., Abbot, P. & Rokas, A. Gestational tissue transcriptomics in term and preterm human pregnancies: a systematic review and meta-analysis. *BMC Medical Genomics* **8,** 27 (2015).
44. Phillips, J. B., Abbot, P. & Rokas, A. Is preterm birth a human-specific syndrome? *Evol Med Public Health* **2015,** 136–148 (2015).
45. Mossman, H. W. *Vertebrate fetal membranes: comparative ontogeny and morphology; evolution; phylogenetic significance; basic function; research opportunities*. (London: Macmilian, 1987).
46. Wildman, D. E. Review: Toward an integrated evolutionary understanding of the mammalian placenta. *Placenta* **32 Suppl 2,** S142–5 (2011).
47. Eidem, H. R., McGary, K. L., Capra, J. A., Abbot, P. & Rokas, A. The transformative potential of an integrative approach to pregnancy. *Placenta* **57,** 204–215 (2017).
48. Fischer, B. & Mitteroecker, P. Covariation between human pelvis shape, stature, and head size alleviates the obstetric dilemma. *PNAS* **112,** 5655–5660 (2015).
49. Rosenberg, K. & Trevathan, W. Birth, obstetrics and human evolution. *BJOG* **109,** 1199–1206 (2002).
50. Swaggart, K. A., Pavlicev, M. & Muglia, L. J. Genomics of preterm birth. *Cold Spring Harb Perspect Med* **5,** a023127 (2015).
51. Ratajczak, C. K., Fay, J. C. & Muglia, L. J. Preventing preterm birth: the past limitations and new potential of animal models. *Dis Model Mech* **3,** 407–414 (2010).
52. Elovitz, M. A. & Mrinalini, C. Animal models of preterm birth. *Trends Endocrinol. Metab.* **15,** 479–487 (2004).

CHAPTER II

GESTATIONAL TISSUE TRANSCRIPTOMICS IN TERM AND PRETERM HUMAN PREGNANCIES: A SYSTEMATIC REVIEW AND META-ANALYSIS[1]

Haley R. Eidem[1], William E. Ackerman IV[2], Kriston L. McGary[1], Patrick Abbot[1], and Antonis Rokas[1]

[1] *Department of Biological Sciences, Vanderbilt University, Nashville, TN 37234, USA*

[2] *Department of Obstetrics and Gynecology, The Ohio State University, Columbus, OH 43210, USA*

[1] This chapter was published in BMC Medical Genomics. 2015 Jun 5;8:27. doi: 10.1186/s12920-015-0099-8.

ABSTRACT

Background: Preterm birth (PTB), or birth before 37 weeks of gestation, is the leading cause of newborn death worldwide. PTB is a critical area of scientific study not only due to its worldwide toll on human lives and economies, but also due to our limited understanding of its pathogenesis and, therefore, its prevention. This systematic review and meta-analysis synthesizes the landscape of PTB transcriptomics research to further our understanding of the genes and pathways involved in PTB subtypes.

Methods: We evaluated all published genome-wide pregnancy studies across gestational tissues and pathologies, including those that focus on PTB, by performing a targeted PubMed MeSH search and systematically reviewing all relevant studies.

Results: Our search yielded 2,361 studies on placenta, decidua, myometrium, maternal blood, cervix, fetal membranes (chorion and amnion), umbilical cord, fetal blood, and basal plate. Selecting only those original research studies that measured transcription on a genome-wide scale and reported lists of expressed genetic elements identified 96, 21, and 21 gene expression, microRNA, and methylation studies, respectively. Although 30% of all PTB cases are due to medical indications, 90% of these 138 studies focused on them. In contrast, only 7% of these studied focused on spontaneous onset of labor, which is responsible for 45% of all PTB cases. Furthermore, only 23 of the 11,007 unique genetic elements reported to be transcriptionally active were recovered 10 or more times in these 138 studies. Meta-analysis of the 96 gene expression studies across 9 distinct gestational tissues and 29 clinical phenotypes showed limited overlap of genes identified as differentially expressed across studies.

Conclusions: Overall, profiles of differentially expressed genes were highly heterogeneous both between as well as within clinical subtypes and tissues as well as between studies of the same clinical subtype *and* tissue. These results suggest that large gaps still exist in the transcriptomic study of specific clinical subtypes as well in the generation of the transcriptional profile of well-studied clinical subtypes; understanding the complex landscape of prematurity will require large-scale, systematic genome-wide analyses of human gestational tissues on both understudied and well-studied subtypes alike.

INTRODUCTION

In humans, gestation typically lasts 40 weeks; preterm birth (PTB) is defined as birth before 37 weeks of gestation and is the leading cause of newborn death worldwide. More than 15 million babies are born too soon every year and rates of PTB had been increasing until 2006 when changes in obstetrical practices regarding early cesarean sections led to a recent decrease in deliveries before term[1]. Nevertheless, 10% of pregnancies still end before 37 weeks across the world and this high incidence of PTB is problematic because premature babies are at higher risk for lifelong health and developmental problems[2,3]. For example, almost half of all children born premature suffer from vision or hearing loss and learning disabilities at some point in their life[4,5]. The combined medical costs stemming from care during the labor and delivery process as well as from care later in life are estimated to be near $26 billion annually[6].

PTB is a critical area of scientific study not only due to its worldwide toll on human lives and economies, but also due to our limited understanding of its pathogenesis and, therefore, its prevention. PTB is a complex, multifactorial syndrome comprised of multiple clinical subtypes, which often occur at different gestational ages and can be defined as either 'spontaneous' or 'medically indicated.' Medically indicated preterm deliveries account for 30% of PTB cases and are often preceded by complications including preeclampsia (PE), intrauterine growth restriction (IUGR), gestational diabetes mellitus (GDM), and chorioamnionitis[7]. The remaining 70% of PTB cases are idiopathic; 45% is due to the

spontaneous onset of labor (iPTB) and the remaining 25% is due to the preterm premature rupture of membranes (PPROM). Regardless of PTB subtype, however, current therapies are not successful in prolonging time to birth once labor has been initiated and the most effective therapy, progesterone supplementation, is only effective in a small number of high-risk cases[8]. It is critical that we gain greater insight into the genes and pathways that regulate birth timing in humans in order to develop effective prevention and treatment strategies, including for cases of iPTB.

A number of environmental risk factors have been associated with iPTB including infection, nutrition, socioeconomic status, and stress but the pathways through which these risk factors act remain unclear[9]. Recent evidence from family, twin, and case-control studies suggests that genetics also plays an important role in birth timing, and the heritability of PTB is estimated to be approximately 30%[1,6,7]. Thus, PTB tends to run in families and women who were born preterm are also more likely to deliver preterm themselves. Interestingly, however, fathers born prematurely do not appear to pass on this risk to offspring[1]. Furthermore, one of the strongest predictors of PTB is previous preterm birth and, in subsequent pregnancies from the same woman, birth timing tends to occur around the same gestational age for each pregnancy[10]. Candidate gene studies have targeted genes with known biological roles potentially related to processes occurring during pregnancy but, in general, teasing apart the complex genetic architecture of pregnancy and PTB has proved challenging.

Further complicating our understanding of PTB genetic architecture are the numerous maternal and fetal gestational tissues that must interact to facilitate parturition[8,11]. These tissues include decidua, myometrium, cervix and maternal blood originating from the mother and villous placenta, fetal membranes (chorion and amnion), umbilical cord, and fetal blood originating from the fetus (Figure 2.1). Furthermore, the basal plate is a region at the maternofetal interface that is commonly biopsied for the study of PTB and includes cells from both the decidua and villous placenta. The decidua, myometrium, and cervix act to house the fetus as well as expel it during labor and delivery, the chorion and amnion act as membranes separating the fetus from the mother, and the umbilical cord allows for efficient nutrient transfer. Together, these tissues share a general functionality in the efficient maternofetal exchange of nutrients, gas, and waste.

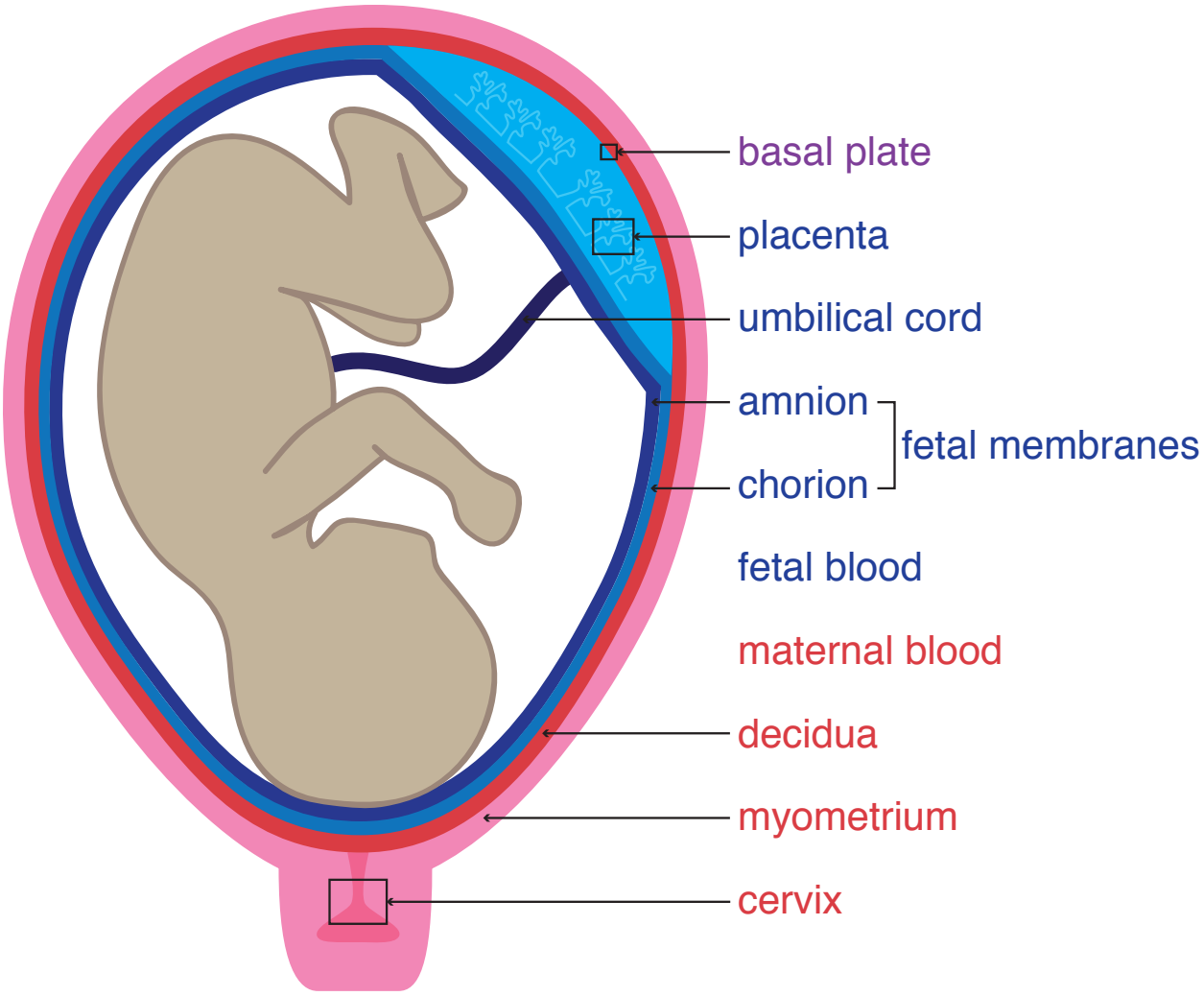

Figure 2.1. The tissues of pregnancy. Our systematic literature review surveyed a total of 9 distinct gestational tissue types including 4 of maternal origin (cervix, myometrium, decidua, and maternal blood; shown in red), 4 of fetal origin (fetal blood, fetal membranes, umbilical cord, and placenta; shown in blue), and 1 of mixed maternal and fetal origin (basal plate; shown in purple).

Although little is known about the complex etiology of PTB, many studies have generated pregnancy-related transcriptomes in various tissue types and pathologies. Because of the diversity of tissues and clinical subtypes involved as well as the large number of questions examined, few studies have attempted to synthesize any dimension of the admittedly

complex transcriptional landscape of this multifactorial syndrome. To synthesize what is known about PTB transcriptomics, we analyzed all published genome-wide studies of gestational tissues (placenta, decidua, myometrium, maternal blood, cervix, basal plate, fetal membranes, umbilical cord, and fetal blood) in both healthy and diseased human pregnancies to identify all statistically supported candidate genetic elements in PTB subtypes.

Our meta-analysis identified 138 genome-wide studies of pregnancy and PTB, of which 96 studied gene expression, 21 studied microRNA activity, and 21 studied methylation. Examination of these studies showed that very few were focused on idiopathic PTB (7%) even though iPTB accounts for 45% of all PTB cases; the majority of PTB research focused on PE. Moreover, there was limited overlap in the identity of candidate genes across studies; for example, the 138 studies reported a total 11,007 unique candidate genetic elements, but only 23/11,007 (0.2%) of those elements were found in 10 or more studies. An in-depth meta-analysis of differentially expressed genes in the 96 studies focused on gene expression showed that overlap between the sets of differentially expressed genes identified in the different studies was limited. In placenta (n=53), for example, 6,291 differentially expressed unique genes were identified but only 2, *LEP* and *FLT1*, were present in more than 10 studies. Similarly, in PE studies (n=27), 5,526 differentially expressed unique genes were identified but only 7 were found in 5 or more studies. The limited overlap of differentially expressed genes across studies of the same tissue or clinical subtype as well as the highly uneven coverage of studies targeting highly prevalent clinical subtypes suggest that larger-scale, systematic

studies aimed at understanding the transcriptional profiles of the diverse clinical PTB subtypes and characterizing their disease-relevant transcriptional differences will be necessary to identify genes whose dysregulation contributes to this complex, multifactorial syndrome.

RESULTS

A systematic review identified 138 transcriptomic studies on 9 gestational tissues and 29 different phenotypes

Of the 2,361 studies identified in our PubMed search, 138 genome-wide transcriptomic studies in human gestational tissue samples were, based on a number of selection criteria, deemed eligible for systematic review (Additional File 2.1)[12-124]. Platform-wise, 130/138 (94%) were microarray studies, 5/138 (4%) were bisulfite-sequencing studies, and 3/138 (2%) were RNA-sequencing studies. All studies were published between 1999 and 2014, primarily in the journals *Placenta* and *The American Journal of Obstetrics and Gynecology*. The phenotypes examined in these studies were quite diverse; 18/138 (13%) studies examined preterm pregnancies, 83/138 (60%) term pregnancies, and 37/138 (27%) both preterm and term pregnancies. One non-clinical phenotype (healthy pregnancies) and 28 distinct clinical phenotypes were represented. Finally, 21/138 (15%) were microRNA studies, 21/138 (15%) were methylation studies, and the remaining 96/138 (70%) were gene expression studies. A total of 11,007 unique genetic elements were reported to be transcriptionally active across all 138 studies (Additional File 2.2), but only 23/11,007 (0.2%) were reported in 10 or more studies.

The 138 studies analyzed 9 distinct gestational tissues, namely placenta, decidua, myometrium, maternal blood, cervix, fetal membranes (chorion and amnion), umbilical cord, fetal blood, and basal plate. The three most common tissues studied were placenta (84/138;

61%), fetal membranes (18/138; 13%), and myometrium (17/138; 12%), whereas each of the other six tissues was sampled in 7 or fewer studies (Figure 2.2).

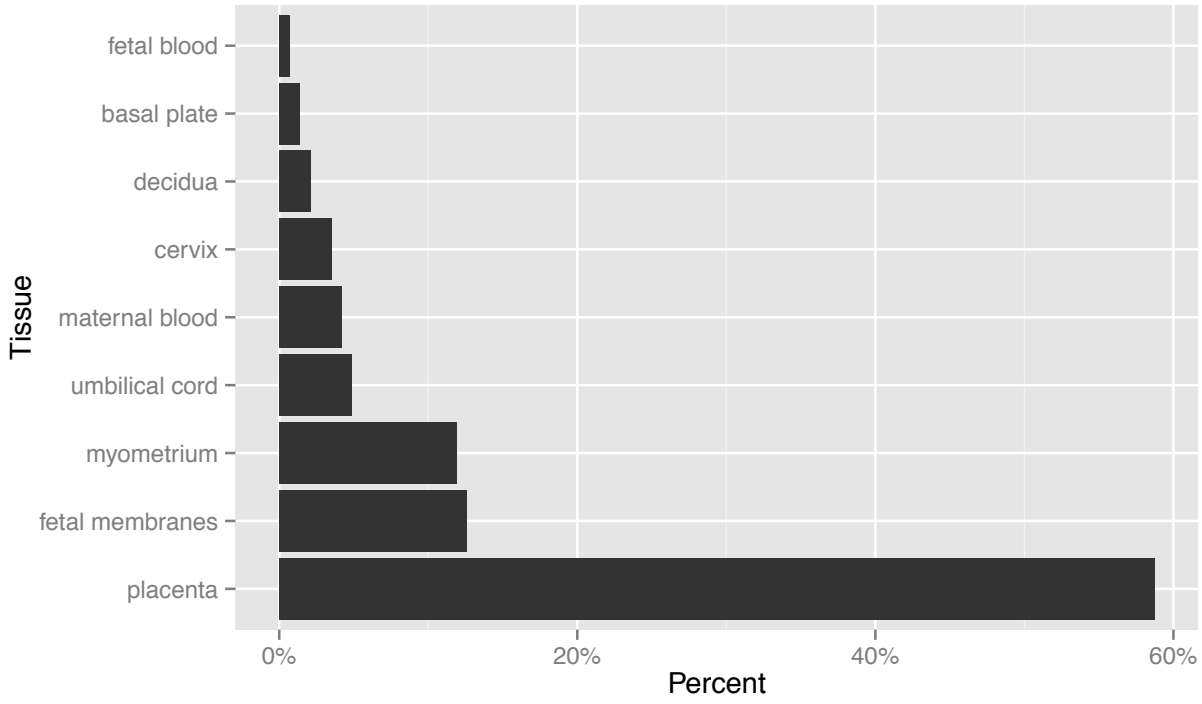

Figure 2.2. The vast majority of genome-wide transcriptomic studies on gestational tissues have focused on the placenta. A targeted PubMed search for genome-wide transcriptomic studies yielded a total of 138 studies focusing on 9 distinct gestational tissue types. Placental research accounted for 58% of all studies in the meta-analysis, followed by fetal membranes (13%) and myometrium (12%).

The 138 studies analyzed 29 distinct phenotypes (Figure 2.3). 16/138 (12%) studies focused on healthy pregnancies, while the remaining 122/138 (88%) studies focused on clinical phenotypes. The most common phenotypes studied were PE (40/138; 29%), labor (16/138; 12%), and iPTB (10/138; 7%). Definitions for all phenotypes are provided in Additional File 2.3.

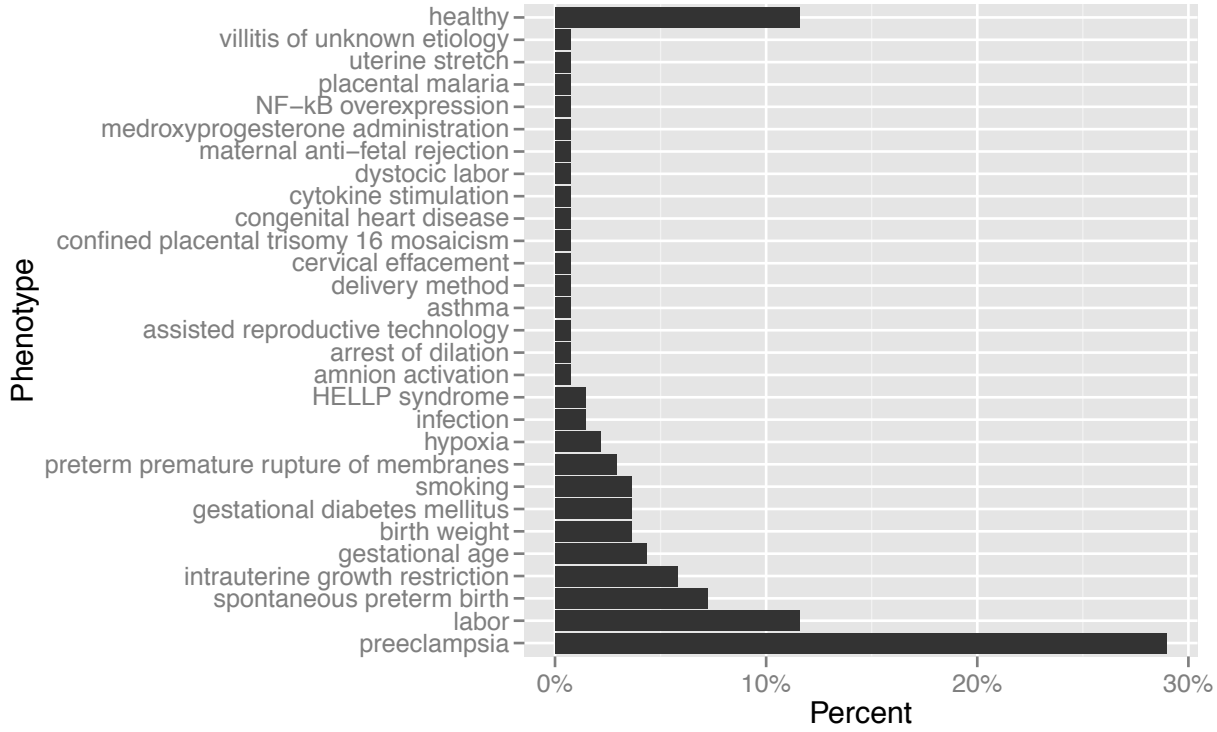

Figure 2.3. Gestational tissue transcriptomic studies in term and preterm human pregnancies organized by phenotype. A targeted PubMed search for genome-wide transcriptomic studies yielded a total of 138 studies focusing on 29 distinct phenotypes. PE research accounted for 29% of all studies in the meta-analysis, followed by labor (12%) and iPTB (7%). Phenotype definitions are provided in Additional File 2.3.

PTB research focus does not reflect PTB subtype epidemiological prevalence

To evaluate whether the proportion of transcriptomic studies devoted on different PTB subtypes reflects their clinical prevalence, we compared the frequencies of the three major clinical etiologies (iPTB at 45%, PPROM at 25%, and medically indicated PTB at 30%) to the frequency of transcriptomic studies devoted to these etiologies (Figure 2.4). We found that although only 30% of all PTB cases are due to medical indications, such as PE, IUGR, or GDM, 124/138 (90%) of the studies in our systematic review focused on them; 40/138 (29%) of these

studies focused on PE alone. In contrast, although iPTB is responsible for 45% of all cases, only 10/138 (7%) of the studies in our systematic review studied this clinical subtype.

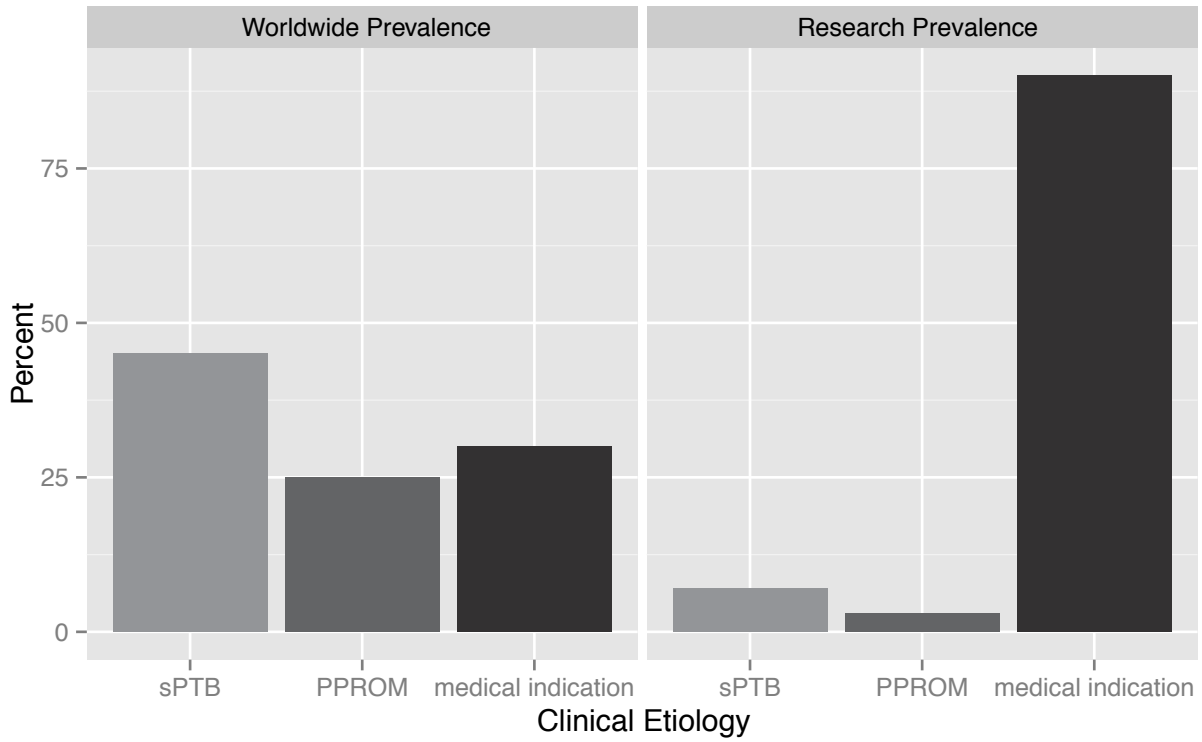

Figure 2.4. Proportion of transcriptomic research does not correspond to PTB subtype prevalence. Although only 30% of all PTB cases are due to medical indications, such as PE, IUGR, or GDM, 90% of the studies in our systematic review focused on them. In contrast, only 7% of the studies focused on iPTB, even though this clinical subtype accounts for the majority (45%) of PTB cases.

A meta-analysis of 96 gene expression studies across 9 distinct gestational tissues showed limited overlap of candidate genes

To perform an aggregated meta-analysis, we focused on the 96/138 gene expression studies. These 96 gene expression studies analyzed all 9 distinct gestational tissues, namely placenta,

decidua, myometrium, maternal blood, cervix, fetal membranes (chorion and amnion), umbilical cord, fetal blood, and basal plate. The three most common tissues studied for differential gene expression were placenta (53/96; 55%), myometrium (17/96; 18%), and fetal membranes (13/96; 14%), whereas each of the other six tissues was sampled in 4 or fewer studies. Genome-wide gene expression profiling studies of the three most commonly studied gestational tissues, i.e., placenta, myometrium, and fetal membranes, identified a total of 8,329 unique differentially expressed genes, of which only 2,021 (24%) were found in two or more studies (Figure 2.5, Additional File 2.4). This examination also showed that only 23 candidate genes were differentially expressed two or more times in studies of all three tissues (Additional File 5). Among the genes present in this overlap were interleukin 1 beta, a proinflammatory cytokine shown to be involved in infection-related PTB and PE, and superoxide dismutase 2, an antioxidant enzyme shown to be involved in oxidative stress associated with PTB[14,19,30,45,61,125-129].

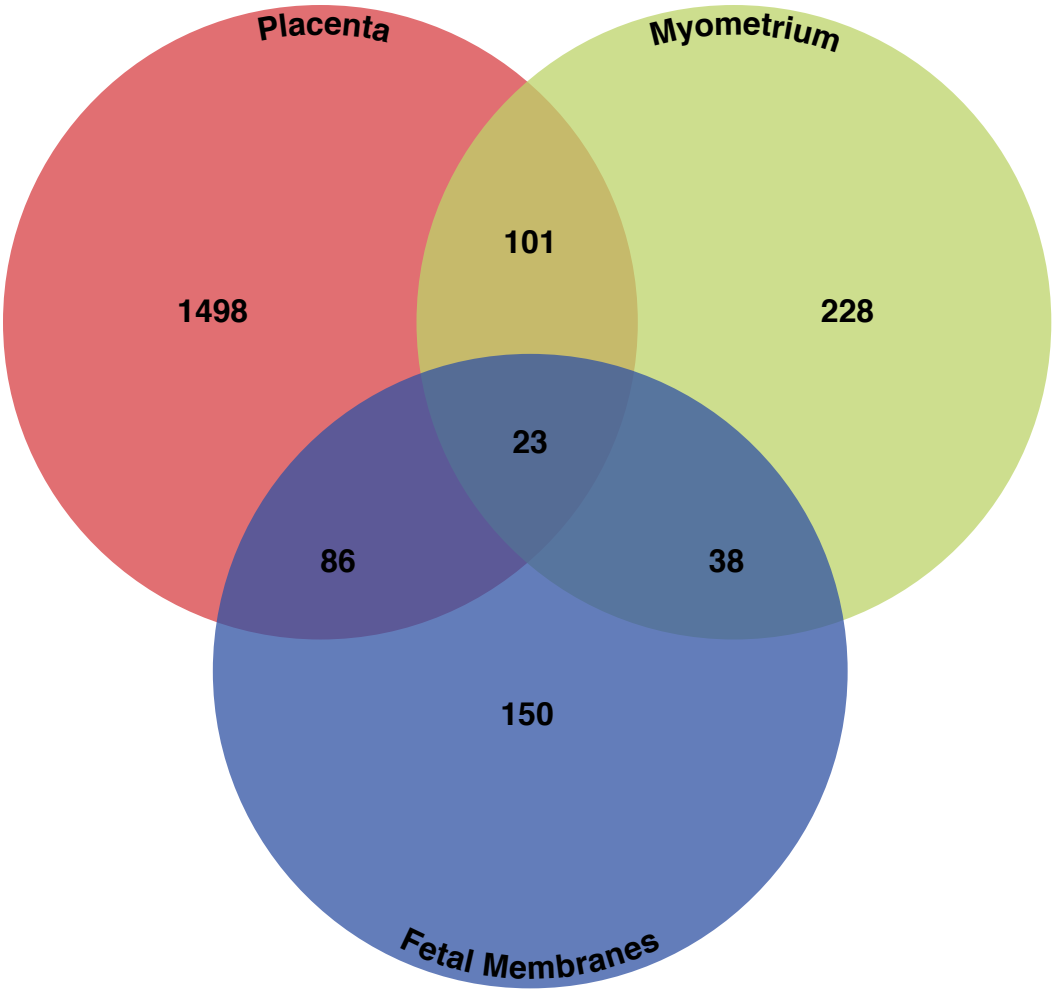

Figure 2.5. Overlap of differentially expressed genes across gestational tissues. Differentially expressed genes present in two or more gene expression studies categorized by tissue were compared across the three most commonly studied tissues (placenta, myometrium, and fetal membranes). Out of 2,021 genes identified to be differentially expressed in at least two studies, only 23 genes were shared across all three tissues.

Although gene expression profiles are available for 29 distinct phenotypes, PTB research is dominated by studies focused on select phenotypes

The 96 gene expression studies analyzed 29 distinct phenotypes. From these studies, 8/96 (8%) studies focused on a non-clinical phenotype (healthy pregnancies), with the remaining

88/96 (92%) studies focused on clinical phenotypes. Among studies focused on clinical phenotypes, the three most common phenotypes investigated were PE (27/96; 28%), labor (14/96; 15%), and IUGR (8/96; 8%); each of the other 26 clinical phenotypes was studied in 5 or fewer studies. Genome-wide gene expression studies of the three most commonly studied clinical phenotypes identified a total of 7,471 unique genes, of which only 1,261 (17%) were present in two or more studies (Figure 2.6, Additional File 2.6). No candidate genes were found two or more times in studies of all three phenotypes. Generally, overlap of differentially expressed genes was more limited across clinical phenotypes than across gestational tissues.

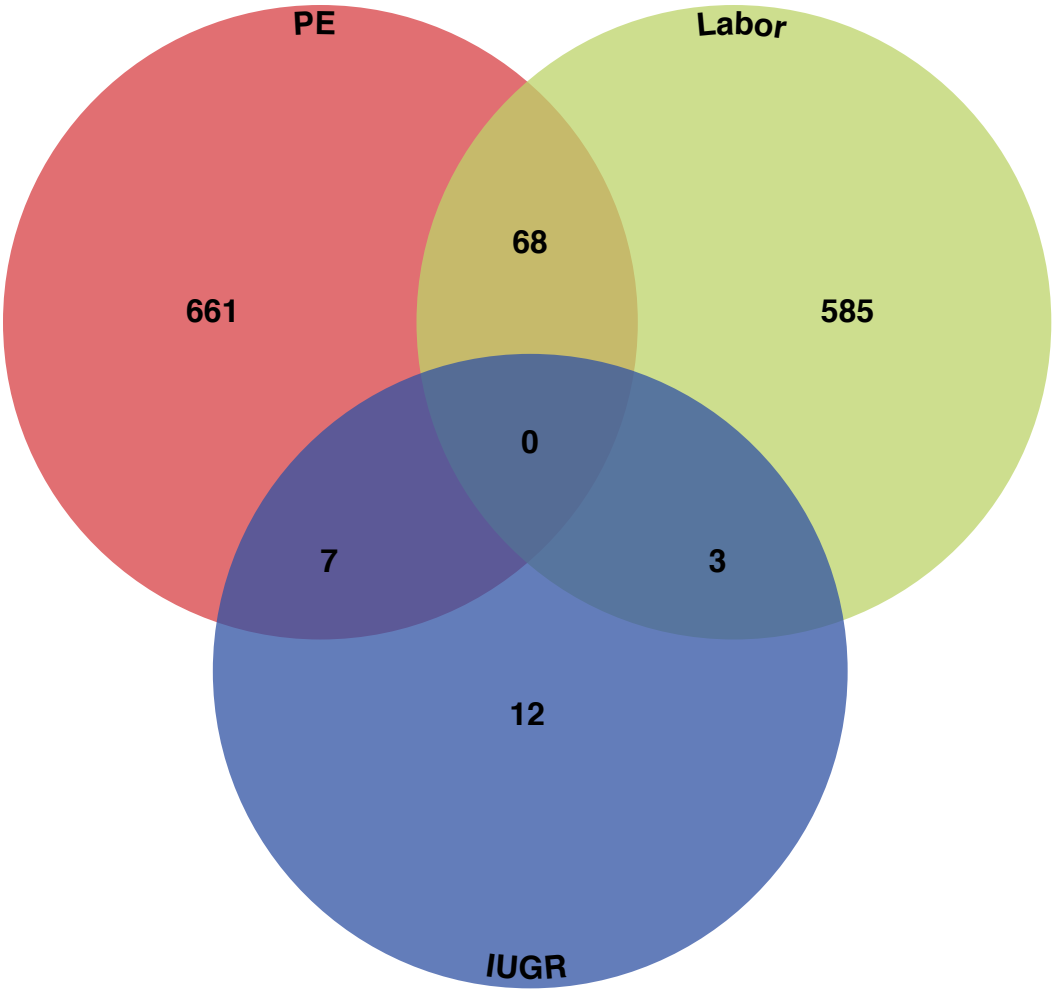

Figure 2.6. Overlap of differentially expressed genes across phenotypes. Differentially expressed genes identified in two or more gene expression studies categorized by phenotype were compared across the most commonly studied phenotypes (PE, labor, and IUGR). Out of 1,261 genes identified to be differentially expressed in at least two studies, none were shared across all three phenotypes.

Overlap of differentially expressed genes identified across PTB studies is limited

Studies of placenta, myometrium, and fetal membranes, the three most commonly studied tissues, focused on a total of 25 distinct phenotypes (Figure 2.7a, Additional File 2.7). The clinical phenotype studied, however, differed between tissues, with PE dominating placental

research (23/53 placental studies or 43%), labor dominating myometrial research (9/17 myometrial studies or 53%), and PPROM dominating fetal membrane research (4/13 fetal membrane studies or 31%). Likewise, the range of tissues studied differed between phenotypes. PE was studied across 4 distinct gestational tissues (placenta, decidua, basal plate, and maternal blood), labor was studied across 4 distinct gestational tissues (myometrium, fetal membranes, placenta, and cervix), and PPROM was studied across only 1 distinct gestational tissue (fetal membranes) (Figure 2.7b, Additional File 2.8).

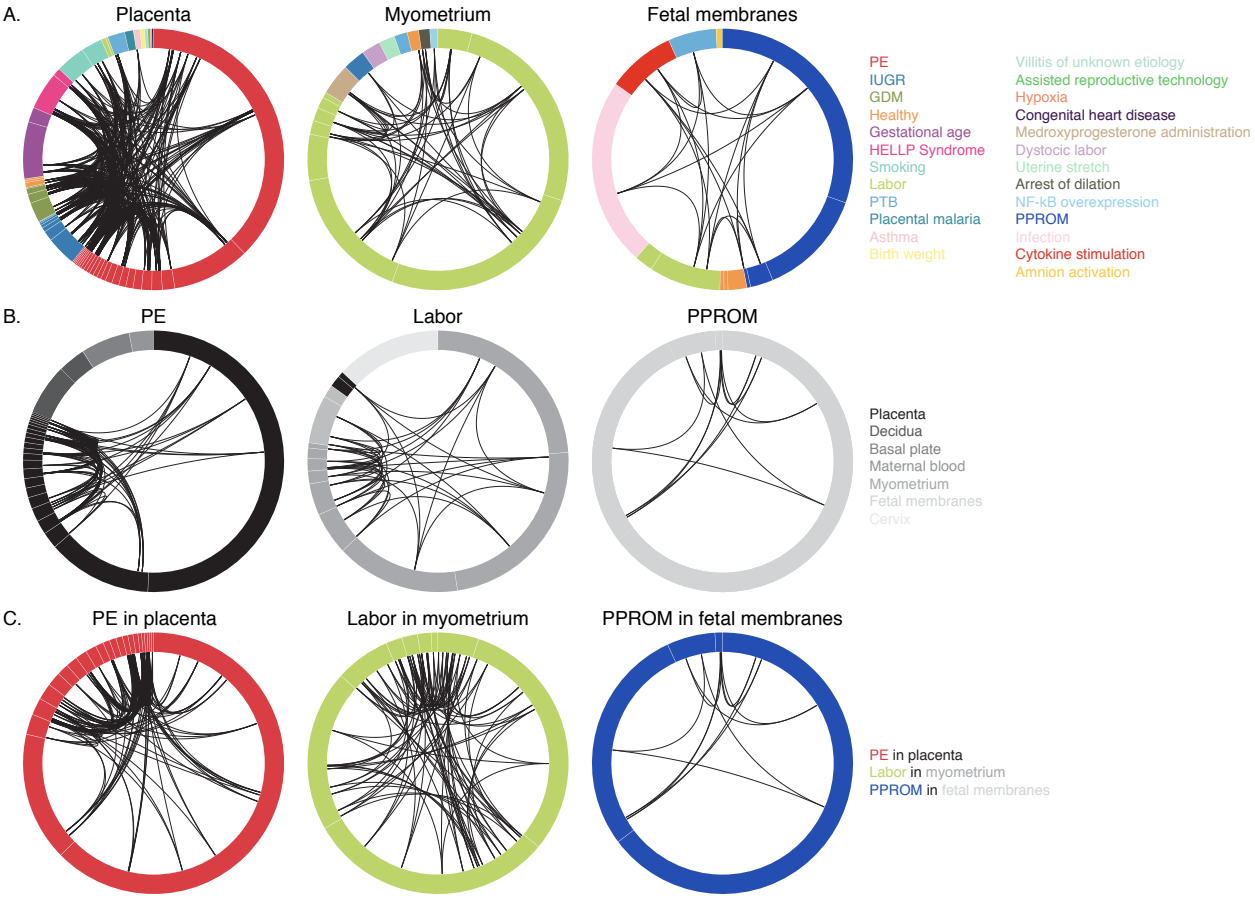

Figure 2.7. Representation of overlap in differentially expressed genes across the most commonly studied tissues, phenotypes, and tissues & phenotypes. Studies are

represented as distinct wedges in the outermost track, colored by phenotype and sized by number of genes reported. Genes that show a high degree of overlap across studies appear as black links connecting each study reporting the gene. In general, the scarcity of links illustrates the considerable lack of overlap in the genes identified as differentially expressed across PTB studies. (A) Representation of overlap in differentially expressed genes across the most commonly studied tissues. Studies of placenta, myometrium, and fetal membranes, the three most commonly studied tissues, focused on a total of 25 distinct phenotypes with PE dominating placental research, labor dominating myometrial research, and PPROM dominating fetal membranes research. (B) Representation of overlap in differentially expressed genes across the most commonly studied phenotypes. PE was studied across 4 distinct gestational tissues (placenta, decidua, basal plate, and maternal blood), labor was studied across 4 distinct gestational tissues (myometrium, fetal membranes, placenta, and cervix), and PPROM was studied across only 1 distinct gestational tissue (fetal membranes). (C) Representation of overlap in differentially expressed genes across the most commonly studied tissue *and* phenotype combinations. The most studied combinations were PE in placenta (n=23), labor in myometrium (n=9), and PPROM in fetal membranes (n=4). Examination of PE in placenta studies identified 16 genes that were present in 4 or more studies, examination of labor in myometrium studies identified 15 genes that were present in 4 or more studies, and examination of PPROM in fetal membranes studies identified 4 genes that were present in 3 or more studies.

To identify common differential gene expression signatures, we looked for overlap between differentially expressed genes reported in studies of the same phenotype *and* tissue. The most studied phenotype-tissue combinations were PE in placenta (n=23), labor in myometrium (n=9), and PPROM in fetal membranes (n=4) (Figure 2.7c, Table 2.1). Examination of PE in placenta studies identified 16 genes that were present in 4 or more studies including *LEP*, a fat-regulating hormone commonly shown to be differentially expressed in gestational tissues of women with PE and HELLP Syndrome, and *FLT1*, a growth factor known to be highly expressed in preeclamptic placental trophoblast cells[17,28,40,44,49,71,81,87,130]. Examination of labor and myometrium studies identified 15 genes that were present in 4 or more studies including *PTGS2*, a cyclooxygenase involved in

inflammation and commonly upregulated in myometrium during labor[14,22,36,60,62,127,131]. Finally, 4 genes were present in 3 or more fetal membranes & PPROM studies including *IL8*, a proinflammatory chemokine often associated with PTB[32,33,51,80,85,132].

Table 2.1. The most often recovered differentially expressed genes in placenta and PE, myometrium and labor, and fetal membranes and PPROM.

Placenta & PE		
Entrez gene ID	**Official gene symbol**	**# studies**
3952	*LEP*	7
2321	*FLT1*	6
3623	*INHA*	6
3624	*INHBA*	6
2022	*ENG*	5
6647	*SOD1*	5
10148	*EBI3*	5
604	*BCL6*	4
1082	*CGB*	4
3972	*LHB*	4
10272	*FSTL3*	4
10544	*PROCR*	4
54210	*TREM1*	4
60676	*PAPPA2*	4
93659	*CGB*	4
94115	*CGB*	4
Myometrium & Labor		
Entrez gene ID	**Official gene symbol**	**# studies**
165	*AEBP1*	4
366	*AQP9*	4
861	*RUNX1*	4
2354	*FOSB*	4
3164	*NR4A1*	4
3576	*IL8*	4
3976	*LIF*	4
5054	*SERPINE1*	4
5292	*PIM1*	4
5334	*PLCL1*	4

5743	*PTGS2*	4
6401	*SELE*	4
9123	*SLC16A3*	4
51129	*ANGPTL4*	4
117247	*SLC16A10*	4
Fetal membranes & PPROM		
Entrez gene ID	**Official gene symbol**	**# studies**
972	*CD74*	3
1117	*CHI3L2*	3
3576	*IL8*	3
7805	*LAPTM5*	3

To examine whether the sets of genes that were most prevalent in each of the three tissue and phenotype pairs (placenta & PE, myometrium & labor, and fetal membranes & PPROM) disproportionally represented particular functions, we examined whether any Gene Ontology functional category was statistically significantly enriched ($p < 0.0001$) in each of the three gene sets (Additional File 2.9). Candidate genes identified in placenta & PE studies were enriched for regulation of cell death (GO:0010941) and apoptosis (GO:0042981), candidate genes identified in myometrium & labor were enriched for wounding (GO:0009611) and inflammatory response (GO:0006954), and candidate genes identified in fetal membranes & PPROM were enriched for defense response (GO:0006952) and immune response (GO:0006955).

DISCUSSION

PTB is a complex, multifactorial syndrome with high prevalence worldwide, whose pathogenesis remains poorly understood, especially for cases of early spontaneous labor. To provide an overview as well as a synthesis of the current landscape of PTB transcriptomics we conducted an in-depth systematic review of the literature as well as a meta-analysis of 96 gene expression studies on a wide diversity of gestational tissues and clinical phenotypes. Examination of our results identifies two key findings. First, the correspondence between PTB subtype prevalence and proportion of transcriptomic research devoted to these subtypes is weak. Second, the overlap between differentially expressed genes identified in different studies is quite small, even on studies aimed on the same phenotypes and tissues. Below, we discuss the possible factors that underlie these two key findings and their implications for research on PTB.

In general, transcriptomic studies on placental tissue samples from women with preeclampsia dominate PTB research. Furthermore, there are very few studies focusing on iPTB, a subtype responsible for 45% of all PTB cases. Although genes commonly associated with PTB clinical subtypes (i.e., *LEP* and *FLT1*) are identified in many of the gene expression studies to be differentially expressed, the overlap between the differentially expressed genes identified across studies is generally very limited. This is not surprising in comparisons between tissues (Figure 5), because these often involve examinations of different clinical subtypes, although it does suggest that there is little overlap in tissue-specific transcriptional profiles of different

clinical subtypes. Similarly, it is not surprising that comparisons between clinical subtypes do not show a high degree of overlap (Figure 2.6), because these often involve examinations of different tissues. Nevertheless, it should be noted that differentially expressed genes with substantial overlap across studies appear to be biologically meaningful. For example, genes involved in hormone regulation (i.e., *CGB*, *CRH*, *INHA*, and *GH2*), which have been previously shown to be key in the maintenance of pregnancy, show substantial overlap in preeclampsia studies. Genes involved in inflammation (i.e., *IL8*), which have been previously shown to be dysregulated in PPROM and other clinical PTB subtypes, are also identified to be differentially expressed in multiple studies.

The observed minimal overlap between the differentially expressed genes identified across studies focused on the same tissue *and* clinical phenotype (Figure 2.7) is possibly more serious. One potential explanation may be the difficulty in obtaining appropriate controls important in pregnancy research; comparing studies that differ with respect to the presence of labor, gestational age, and fetal sex is challenging, since all of these factors are thought to influence the gene expression landscape in gestational tissues. Even though matching of samples with respect to all these factors is very challenging, the reporting of a standard list of such factors as required metadata in transcriptomic studies would facilitate further examination of their importance and likely influence on transcriptomic profiles.

Unfortunately, different studies also follow different guidelines with respect to data availability. For example, some studies do not report the full list of differentially expressed genes identified or do not make them easily available for subsequent analysis (e.g., reporting tables that contain differential expression data on hundreds or thousands of genes in PDF format), therefore limiting and biasing the data available for subsequent analyses. The publishing of the data for *all* genes with differential expression above an explicit significance threshold is crucial in order to carefully analyze aggregated results and draw meaningful conclusions.

CONCLUSIONS

This study synthesizes all high-quality transcriptomic studies on gestational tissues to examine the landscape of PTB as well as to identify genes and genomic elements associated with it. We found that highly prevalent PTB subtypes, such as iPTB, are not well studied as well as that differentially expressed genes identified in different studies are often non-overlapping. Thus, the identification of the genes whose dysregulation contributes to this complex and multifactorial syndrome will require many more large-scale, systematic studies aimed at understanding the transcriptional profiles of these diverse clinical PTB subtypes across gestational tissues and characterizing their disease-relevant transcriptional differences.

METHODS

Search strategy

This systematic review and meta-analysis followed guidelines set by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Additional Files 10-12)[133]. The electronic search was performed on August 16, 2014 in PubMed with no restrictions to identify all articles relating to differentially expressed or methylated genes and microRNAs in human gestational tissues. The search strategy was constructed based on related MeSH terms:

"Pregnancy"[mh] AND "Humans"[mh] AND ("Gene Expression Profiling"[mh] OR "Gene Expression Regulation"[mh]) AND ("Placenta"[mh] OR "Decidua"[mh] OR "Myometrium"[mh] OR "Cervix Uteri"[mh] OR "Extraembryonic Membranes"[mh] OR "Blood"[mh] OR "Plasma"[mh] OR "Umbilical Cord"[mh])

Systematic review

We collected abstracts for all 2,361 studies identified from this search and annotated eligibility based on 6 inclusion criteria:

1. Published in English
2. Full text available
3. Original research
4. Human placental tissue samples
5. Genome-wide analysis

6. Candidate gene list assembled

138 studies met all 6 criteria and were included in the systematic review.

Meta-analysis

Studies were included in our meta-analysis if they met an additional 3 inclusion criteria:

1. Studied differential gene expression
2. Provided candidate gene list
3. DAVID ID conversion successful

Of the 138 studies included in our systematic literature review, 96 gene expression studies met these criteria and were further analyzed. All differentially expressed genes reported in these studies were first extracted and then converted to Entrez ID format using the DAVID online tool, selecting the smallest Entrez ID number if multiple IDs mapped to single genes. We extracted all reported significantly differentially expressed genes based on each study's significance threshold for differential expression. Overlap was determined simply by the presence of the same gene in the gene lists from different studies. DAVID was used to assay functional enrichment according to Gene Ontology categories. All analyses were performed using Python and visualizations were performed using ggplot2 and Circos[134,135].

COMPETING INTERESTS

The authors declare no competing interests.

AUTHORS' CONTRIBUTIONS

HRE and AR designed the study with input from WEA, KLM, and PA. HRE carried out the study and drafted the manuscript with subsequent contributions and revisions from AR. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We thank Louis J. Muglia for offering invaluable advice on experimental design at an early stage of this experiment. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. HRE was supported by the Graduate Program in Biological Sciences at Vanderbilt University. Research on this project was supported by the March of Dimes through the March of Dimes Prematurity Research Center Ohio Collaborative.

REFERENCES

1. Muglia, L. J. & Katz, M. The enigma of spontaneous preterm birth. *N. Engl. J. Med.* **362,** 529–535 (2010).
2. Blencowe, H. *et al.* National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet* **379,** 2162–2172 (2012).
3. Goldenberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R. Epidemiology and causes of preterm birth. *Lancet* **371,** 75–84 (2008).
4. Esplin, M. S. Overview of spontaneous preterm birth: a complex and multifactorial phenotype. *Clin Obstet Gynecol* **57,** 518–530 (2014).
5. Chang, H. H., Larson, J., Blencowe, H. & Spong, C. Y. Preventing preterm births: analysis of trends and potential reductions with interventions in 39 countries with very high human development index. *The Lancet* (2013). doi:10.1016/S0140-6736(12)61856-X
6. Plunkett, J. *et al.* An evolutionary genomic approach to identify genes involved in human birth timing. *PLoS genetics* **7,** e1001365 (2011).
7. Menon, R. Spontaneous preterm birth, a clinical dilemma: etiologic, pathophysiologic and genetic heterogeneities and racial disparity. *Acta Obstet Gynecol Scand* **87,** 590–600 (2008).
8. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: one syndrome, many causes. *Science* **345,** 760–765 (2014).
9. Bezold, K. Y., Karjalainen, M. K., Hallman, M., Teramo, K. & Muglia, L. J. The genomics of preterm birth: from animal models to human studies. *Genome Med* **5,** 34 (2013).
10. Lengyel, C., Muglia, L. J. & Pavličev, M. Genetics of Preterm Birth. *eLS* doi:10.1002/9780470015902.a0025448
11. Iams, J. D. Preterm birth categories-labels with consequences. *American journal of obstetrics and gynecology* **210,** 97–98 (2014).
12. Arcelli, D. *et al.* Identification of circulating placental mRNA in maternal blood of pregnancies affected with fetal congenital heart diseases at the second trimester of pregnancy: implications for early molecular screening. *Prenat. Diagn.* n/a–n/a (2010). doi:10.1002/pd.2443
13. Bethin, K. E. *et al.* Microarray Analysis of Uterine Gene Expression in Mouse and Human Pregnancy. *Molecular Endocrinology* **17,** 1454–1469 (2003).
14. Bollopragada, S. *et al.* Term labor is associated with a core inflammatory response in human fetal membranes, myometrium, and cervix. *American journal of obstetrics and gynecology* **200,** 104.e1–104.e11 (2009).
15. Brennan, D. J. *et al.* Identification of a myometrial molecular profile for dystocic labor. *BMC pregnancy and childbirth* **11,** 74 (2011).
16. Bruchova, H., Vasikova, A., Merkerova, M. & Milcova, A. Effect of maternal tobacco smoke exposure on the placental transcriptome. *Placenta* (2010). doi:10.1016/j.placenta.2009.12.016
17. Buimer, M. *et al.* Seven placental transcripts characterize HELLP-syndrome. *Placenta*

29, 444–453 (2008).
18. Bukowski, R., Hankins, G. D. V., Saade, G. R., Anderson, G. D. & Thornton, S. Labor-Associated Gene Expression in the Human Uterine Fundus, Lower Segment, and Cervix. *PLoS Med.* **3,** e169 (2006).
19. Centlow, M., Wingren, C., Borrebaeck, C., Brownstein, M. J. & Hansson, S. R. Differential Gene Expression Analysis of Placentas with Increased Vascular Resistance and Pre-Eclampsia Using Whole-Genome Microarrays. *J Pregnancy* **2011,** 1–12 (2011).
20. Chaemsaithong, P. *et al.* Characterization of the myometrial transcriptome in women with an arrest of dilatation during labor. *Journal of Perinatal Medicine* **41,** 665–681 (2012).
21. Chan, E. C. *et al.* Human myometrial genes are differentially expressed in labor: A suppression subtractive hybridization study. *J. Clin. Endocrinol. Metab.* **87,** 2435–2441 (2002).
22. Chan, Y.-W., van den Berg, H. A., Moore, J. D., Quenby, S. & Blanks, A. M. Assessment of Myometrial Transcriptome Changes Associated with Spontaneous Human Labour by High Throughput RNA-seq. *Exp. Physiol.* **0,** 1–15 (2014).
23. Chang, S.-D. *et al.* Analyses of placental gene expression in pregnancy-related hypertensive disorders. *Taiwanese Journal of Obstetrics and Gynecology* **50,** 283–291 (2011).
24. Charpigny, G. A Functional Genomic Study to Identify Differential Gene Expression in the Preterm and Term Human Myometrium. *Biol. Reprod.* **68,** 2289–2296 (2002).
25. Chim, S. S. C. *et al.* Systematic Identification of Spontaneous Preterm Birth-Associated RNA Transcripts in Maternal Plasma. *PLoS ONE* **7,** e34328 (2012).
26. Cordeaux, Y., Tattersall, M., Charnock-Jones, D. S. & Smith, G. C. S. Effects of Medroxyprogesterone Acetate on Gene Expression in Myometrial Explants from Pregnant Women. *The Journal of clinical endocrinology and metabolism* **95,** E437–E447 (2010).
27. Dunk, C. E. *et al.* A distinct microvascular endothelial gene expression profile in severe IUGR placentas. *Placenta* **33,** 285–293 (2012).
28. Enquobahrie, D. A. *et al.* Differential placental gene expression in preeclampsia. *American journal of obstetrics and gynecology* **199,** 566.e1–11 (2008).
29. Enquobahrie, D. A., Williams, M. A., Qiu, C., Meller, M. & Sorensen, T. K. Global placental gene expression in gestational diabetes mellitus. *American journal of obstetrics and gynecology* **200,** 206.e1–13 (2009).
30. Esplin, M. S. *et al.* The use of cDNA microarray to identify differentially expressed labor-associated genes within the human myometrium during labor. *American journal of obstetrics and gynecology* **193,** 404–413 (2005).
31. Gack, S. *et al.* Preeclampsia: increased expression of soluble ADAM 12. *J Mol Med* **83,** 887–896 (2005).
32. Haddad, R. *et al.* Human spontaneous labor without histologic chorioamnionitis is characterized by an acute inflammation gene expression signature. *American journal of obstetrics and gynecology* **195,** 394–405.e12 (2006).

33. Han, Y. M. *et al.* Region-specific gene expression profiling: novel evidence for biological heterogeneity of the human amnion. *Biol. Reprod.* **79,** 954–961 (2008).
34. Hansson, S. R. *et al.* Gene expression profiling of human placentas from preeclamptic and normotensive pregnancies. *Molecular human reproduction* **12,** 169–179 (2006).
35. Hassan, S. S. *et al.* The transcriptome of cervical ripening in human pregnancy before the onset of labor at term: Identification of novel molecular functions involved in this process. *J Matern Fetal Neonatal Med* **22,** 1183–1193 (2009).
36. Havelock, O. C. *et al.* Human myometrial gene expression before and during parturition. *Biol. Reprod.* **72,** 707–719 (2005).
37. Heikkilä, A. *et al.* Tumor suppressor and growth regulatory genes are overexpressed in severe early-onset preeclampsia--an array study on case-specific human preeclamptic placental tissue. *Acta Obstet Gynecol Scand* **84,** 679–689 (2005).
38. Heng, Y. J., Pennell, C. E., Chua, H. N., Perkins, J. E. & Lye, S. J. Whole Blood Gene Expression Profile Associated with Spontaneous Preterm Birth in Women with Threatened Preterm Labor. *PLoS ONE* **9,** e96901 (2014).
39. Hiden, U. *et al.* Insulin control of placental gene expression shifts from mother to foetus over the course of pregnancy. *Diabetologia* **49,** 123–131 (2006).
40. Hoegh, A. M., Borup, R., Nielsen, F. C., Sørensen, S. & Hviid, T. V. F. Gene expression profiling of placentas affected by pre-eclampsia. *J. Biomed. Biotechnol.* **2010,** 787545–11 (2010).
41. Jarvenpaa, J. *et al.* Altered expression of angiogenesis-related placental genes in pre-eclampsia associated with intrauterine growth restriction. *Gynecol. Endocrinol.* **23,** 351–355 (2007).
42. Junus, K. *et al.* Gene expression profiling of placentae from women with early- and late-onset pre-eclampsia: down-regulation of the angiogenesis-related genes ACVRL1 and EGFL7 in early-onset disease. *Molecular human reproduction* **18,** 146–155 (2012).
43. Kang, B. Y., Tsoi, S., Shan Zhu, Shenghui Su & Kay, H. H. Differential Gene Expression Profiling in HELLP Syndrome Placentas. *Reproductive Sciences* **15,** 285–294 (2008).
44. Kang, J. H. *et al.* Preeclampsia leads to dysregulation of various signaling pathways in placenta. *Journal of hypertension* **29,** 928–936 (2011).
45. Khanjani, S. *et al.* NF-κB regulates a cassette of immune/inflammatory genes in human pregnant myometrium at term. *Journal of Cellular and Molecular Medicine* **15,** 809–824 (2011).
46. Kim, M. J. *et al.* Villitis of Unknown Etiology Is Associated with a Distinct Pattern of Chemokine Up-Regulation in the Feto-Maternal and Placental Compartments: Implications for Conjoint Maternal Allograft Rejection and Maternal Anti-Fetal Graft-versus-Host Disease. *The Journal of Immunology* **182,** 3919–3927 (2009).
47. Kim, J. *et al.* Transcriptome landscape of the human placenta. *BMC genomics* **13,** 115 (2012).
48. Lapaire, O. *et al.* Microarray screening for novel preeclampsia biomarker candidates. *Fetal. Diagn. Ther.* **31,** 147–153 (2012).
49. Lee, G. S. R., Joe, Y. S., Kim, S. J. & Shin, J. C. Cytokine-related genes and oxidation-

related genes detected in preeclamptic placentas. *Arch Gynecol Obstet* **282,** 363–369 (2010).
50. Lee, J. *et al.* Characterization of the Fetal Blood Transcriptome and Proteome in Maternal Anti-Fetal Rejection: Evidence of a Distinct and Novel Type of Human Fetal Systemic Inflammatory Response. *Am J Reprod Immunol* **70,** 265–284 (2013).
51. Li, R. *et al.* Inflammatory Gene Regulatory Networks in Amnion Cells Following Cytokine Stimulation: Translational Systems Approach to Modeling Human Parturition. *PLoS ONE* **6,** e20560 (2011).
52. Lian, I. A. *et al.* Matrix Metalloproteinase 1 in Pre-eclampsia and Fetal Growth Restriction: Reduced Gene Expression in Decidual Tissue and Protein Expression in Extravillous Trophoblasts. *Placenta* **31,** 615–620 (2010).
53. Lim, S. *et al.* Nuclear Factor Kappa B Activation Occurs in the Amnion Prior to Labour Onset and Modulates the Expression of Numerous Labour Associated Genes. *PLoS ONE* **7,** e34707 (2012).
54. Liu, Y. *et al.* HSP70 is associated with endothelial activation in placental vascular diseases. *Molecular medicine (Cambridge, Mass.)* **14,** 561–566 (2008).
55. Løset, M. *et al.* A transcriptional profile of the decidua in preeclampsia. *American journal of obstetrics and gynecology* **204,** 84.e1–27 (2011).
56. Mayor-Lynn, K., Toloubeydokhti, T., Cruz, A. C. & Chegini, N. Expression profile of microRNAs and mRNAs in human placentas from pregnancies complicated by preeclampsia and preterm labor. *Reproductive sciences (Thousand Oaks, Calif.)* **18,** 46–56 (2011).
57. McCarthy, C. *et al.* Altered gene expression patterns in intrauterine growth restriction: potential role of hypoxia. *American journal of obstetrics and gynecology* **196,** 70.e1–6 (2007).
58. McMinn, J. *et al.* Unbalanced Placental Expression of Imprinted Genes in Human Intrauterine Growth Restriction. *Placenta* **27,** 540–549 (2006).
59. Meng, T. *et al.* Identification of differential gene expression profiles in placentas from preeclamptic pregnancies versus normal pregnancies by DNA microarrays. *OMICS* **16,** 301–311 (2012).
60. Mittal, P. *et al.* Characterization of the myometrial transcriptome and biological pathways of spontaneous human labor at term. *Journal of Perinatal Medicine* **38,** 617–643 (2010).
61. Muehlenbachs, A., Fried, M., Lachowitzer, J., Mutabingwa, T. K. & Duffy, P. E. Genome-wide expression analysis of placental malaria reveals features of lymphoid neogenesis during chronic infection. *J. Immunol.* **179,** 557–565 (2006).
62. Nhan-Chang, C.-L. *et al.* Characterization of the transcriptome of chorioamniotic membranes at the site of rupture in spontaneous labor at term. *American journal of obstetrics and gynecology* **202,** 462.e1–462.e41 (2010).
63. Nishizawa, H. *et al.* Microarray analysis of differentially expressed fetal genes in placental tissue derived from early and late onset severe pre-eclampsia. *Placenta* **28,** 487–497 (2007).

64. Nishizawa, H. *et al.* Comparative gene expression profiling of placentas from patients with severe pre-eclampsia and unexplained fetal growth restriction. *Reprod. Biol. Endocrinol.* **9,** 107 (2011).
65. Ogita, K. *et al.* Differential Expression and Localization of Decorin in Human Choriodecidual Membrane during Preterm and Term Pregnancy. *Am J Reprod Immunol* **51,** 204–210 (2004).
66. Okamoto, A. *et al.* IGFBP1 and Follistatin-like 3 Genes are Significantly Up-regulated in Expression Profiles of the IUGR Placenta. *Placenta* **27,** 317–321 (2006).
67. Osei-Kumah, A., Smith, R., Jurisica, I., Caniggia, I. & Clifton, V. L. Sex-specific differences in placental global gene expression in pregnancies complicated by asthma. *Placenta* **32,** 570–578 (2011).
68. Peng, H.-H. *et al.* The effects of labor on differential gene expression in parturient women, placentas, and fetuses at term pregnancy. *The Kaohsiung Journal of Medical Sciences* **27,** 494–502 (2011).
69. Radaelli, T., Varastehpour, A., Catalano, P. & Haugeul-de Mouzon, S. Gestational diabetes induces placental genes for chronic stress and inflammatory pathways. *Diabetes* **52,** 2951–2958 (2002).
70. Rehman, K. S., Yin, S. & Mayhew, B. A. Human myometrial adaptation to pregnancy: cDNA microarray gene expression profiling of myometrium from non-pregnant and pregnant women. *Molecular human …* **9,** 681–700 (2003).
71. Reimer, T. *et al.* Microarray analysis of differentially expressed genes in placental tissue of pre-eclampsia: up-regulation of obesity-related genes. *Molecular human reproduction* **8,** 674–680 (2002).
72. Roh, C. R., Budhraja, V., Kim, H. S., Nelson, D. M. & Sadovsky, Y. Microarray-based identification of differentially expressed genes in hypoxic term human trophoblasts and in placental villi of pregnancies with growth restricted fetuses. *Placenta* **26,** 319–328 (2005).
73. Saben, J. *et al.* A comprehensive analysis of the human placenta transcriptome. *Placenta* **35,** 125–131 (2014).
74. Shankar, R. *et al.* Molecular markers of preterm labor in the choriodecidua. *Reproductive sciences (Thousand Oaks, Calif.)* **17,** 297–310 (2010).
75. Sitras, V., Fenton, C., Paulssen, R., Vårtun, Å. & Acharya, G. Differences in gene expression between first and third trimester human placenta: a microarray study. *PLoS ONE* **7,** e33294 (2012).
76. Soleymanlou, N. *et al.* Molecular Evidence of Placental Hypoxia in Preeclampsia. *The Journal of clinical endocrinology and metabolism* **90,** 4299–4308 (2005).
77. Sood, R., Zehnder, J. L., Druzin, M. L. & Brown, P. O. Gene expression patterns in human placenta. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 5478–5483 (2006).
78. Tashima, L. S. & Millar, L. K. Genes upregulated in human fetal membranes by infection or labor. *Obstetrics & …* (1998).
79. Textoris, J. *et al.* Evaluation of Current and New Biomarkers in Severe Preeclampsia: A

Microarray Approach Reveals theVSIG4 Gene as a Potential Blood Biomarker. *PLoS ONE* **8,** e82638 (2013).
80. Tromp, G. *et al.* Genome-wide expression profiling of fetal membranes reveals a deficient expression of proteinase inhibitor 3 in premature rupture of membranes. *American journal of obstetrics and gynecology* **191,** 1331–1338 (2004).
81. Tsai, S. *et al.* Transcriptional profiling of human placentas from pregnancies complicated by preeclampsia reveals disregulation of sialic acid acetylesterase and immune signalling pathways. *Placenta* **32,** 175–182 (2011).
82. Uusküla, L. *et al.* Mid-gestational gene expression profile in placenta and link to pregnancy complications. *PLoS ONE* **7,** e49248 (2012).
83. Várkonyi, T. *et al.* Microarray profiling reveals that placental transcriptomes of early-onset HELLP syndrome and preeclampsia are similar. *Placenta* **32 Suppl,** S21–9 (2011).
84. Votavova, H. *et al.* Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. *Placenta* **32,** 763–770 (2011).
85. Weiner, C. P. *et al.* Human effector/initiator gene sets that regulate myometrial contractility during term and preterm labor. *The American Journal of Obstetrics & Gynecology* **202,** 474.e1–474.e20 (2010).
86. Winn, V. D. *et al.* Gene Expression Profiling of the Human Maternal-Fetal Interface Reveals Dramatic Changes between Midgestation and Term. *Endocrinology* **148,** 1059–1079 (2007).
87. Winn, V. D. *et al.* Severe preeclampsia-related changes in gene expression at the maternal-fetal interface include sialic acid-binding immunoglobulin-like lectin-6 and pappalysin-2. *Endocrinology* **150,** 452–462 (2009).
88. Yan, Y. H. *et al.* Screening for preeclampsia pathogenesis related genes. *Eur Rev Med Pharmacol Sci* **17,** 3083–3094 (2013).
89. Zhang, Y. *et al.* Altered global gene expressions of human placentae subjected to assisted reproductive technology treatments. *Placenta* **31,** 251–258 (2010).
90. Zhao, Y. H. *et al.* Genomic expression profiles of blood and placenta reveal significant immune-related pathways and categories in Chinese women with gestational diabetes mellitus. *Diabet. Med.* **28,** 237–246 (2011).
91. Zhou, R. *et al.* Genomewide oligonucleotide microarray analysis on placentae of pre-eclamptic pregnancies. *Gynecol. Obstet. Invest.* **62,** 108–114 (2006).
92. Zhou, Y. *et al.* Reversal of gene dysregulation in cultured cytotrophoblasts reveals possible causes of preeclampsia. *J. Clin. Invest.* **123,** 2862–2872 (2013).
93. Barbaux, S. *et al.* A genome-wide approach reveals novel imprinted genes expressed in the human placenta. *Epigenetics* **7,** 1079–1090 (2012).
94. Blair, J. D., Langlois, S., McFadden, D. E. & Robinson, W. P. Overlapping DNA methylation profile between placentas with trisomy 16 and early-onset preeclampsia. *Placenta* **35,** 216–222 (2014).
95. Chu, T. *et al.* Comprehensive Analysis of Preeclampsia-Associated DNA Methylation in the Placenta. *PLoS ONE* **9,** e107318 (2014).

96. Cruickshank, M. N. *et al.* Analysis of epigenetic changes in survivors of preterm birth reveals the effect of gestational age and evidence for a long term legacy. *Genome Med* **5,** 96 (2013).
97. Das, R. *et al.* DNMT1 and AIM1 Imprinting in human placenta revealed through a genome-wide screen for allele-specific DNA methylation. *BMC genomics* **14,** 685 (2013).
98. Gordon, L. *et al.* Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome research* **22,** 1395–1406 (2012).
99. Kim, S. Y. *et al.* Methylome of Fetal and Maternal Monocytes and Macrophages at the Feto-Maternal Interface. *Am J Reprod Immunol* **68,** 8–27 (2012).
100. Lun, F. M. F. *et al.* Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clin. Chem.* **59,** 1583–1594 (2013).
101. Ruchat, S.-M. *et al.* Gestational diabetes mellitus epigenetically affects genes predominantly involved in metabolic diseases. *Epigenetics* **8,** 935–943 (2013).
102. Schroeder, D. I. *et al.* The human placenta methylome. *PNAS* **110,** 6037–6042 (2013).
103. Suter, M. *et al.* Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics* **6,** 1284–1294 (2011).
104. Turan, N. *et al.* DNA methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease? *BMC Medical Genomics* **5,** 10 (2012).
105. Yuen, R. K. C., Chen, B., Blair, J. D., Robinson, W. P. & Nelson, D. M. Hypoxia alters the epigenetic profile in cultured human placental trophoblasts. *Epigenetics* **8,** 192–202 (2013).
106. Choi, S. Y. *et al.* MicroRNA expression profiles in placenta with severe preeclampsia using a PNA-based microarray. *Placenta* **34,** 799–804 (2013).
107. Elovitz, M. A. *et al.* Distinct cervical microRNA profiles are present in women destined to have a preterm birth. *American journal of obstetrics and gynecology* **210,** 221.e1–11 (2014).
108. Enquobahrie, D. A. *et al.* Placental microRNA expression in pregnancies complicated by preeclampsia. *American journal of obstetrics and gynecology* **204,** 178.e12–21 (2011).
109. Gu, Y., Sun, J., Groome, L. J. & Wang, Y. Differential miRNA expression profiles between the first and third trimester human placentas. *Am. J. Physiol. Endocrinol. Metab.* **304,** E836–43 (2013).
110. Guo, L. *et al.* A Comprehensive Survey of miRNA Repertoire and 3′ Addition Events in the Placentas of Patients with Pre-Eclampsia from High-Throughput Sequencing. *PLoS ONE* **6,** e21072 (2011).
111. Hassan, S. S. *et al.* MicroRNA expression profiling of the human uterine cervix after term labor and delivery. *American journal of obstetrics and gynecology* **202,** 80.e1–8 (2010).
112. Higashijima, A. *et al.* Characterization of placenta-specific microRNAs in fetal growth

restriction pregnancy. *Prenat. Diagn.* **33,** 214–222 (2013).
113. Hu, Y. *et al.* Differential expression of microRNAs in the placentae of Chinese patients with severe pre-eclampsia. *Clinical Chemistry and Laboratory Medicine* **47,**
114. Ishibashi, O. *et al.* Hydroxysteroid (17-β) dehydrogenase 1 is dysregulated by miR-210 and miR-518c that are aberrantly expressed in preeclamptic placentas: a novel marker for predicting preeclampsia. *Hypertension* **59,** 265–273 (2012).
115. Luo, S. S. *et al.* Human Villous Trophoblasts Express and Secrete Placenta-Specific MicroRNAs into Maternal Circulation via Exosomes. *Biol. Reprod.* **81,** 717–729 (2009).
116. Montenegro, D. *et al.* Differential expression of microRNAs with progression of gestation and inflammation in the human chorioamniotic membranes. *American journal of obstetrics and gynecology* **197,** 289.e1–289.e6 (2007).
117. Montenegro, D. *et al.* Expression patterns of microRNAs in the chorioamniotic membranes: a role for microRNAs in human pregnancy and parturition. *J. Pathol.* **217,** 113–121 (2009).
118. Morales-Prieto, D. M. *et al.* MicroRNA expression profiles of trophoblastic cells. *Placenta* **33,** 725–734 (2012).
119. Mouillet, J.-F., Ouyang, Y., Coyne, C. B. & Sadovsky, Y. MicroRNAs in placental health and disease. *American journal of obstetrics and gynecology* **213,** S163–72 (2015).
120. Noack, F. *et al.* miRNA expression profiling in formalin-fixed and paraffin-embedded placental tissue samples from pregnancies with severe preeclampsia. *Journal of Perinatal Medicine* **39,** 267–271 (2011).
121. Pineles, B. L. *et al.* Distinct subsets of microRNAs are expressed differentially in the human placentas of patients with preeclampsia. *American journal of obstetrics and gynecology* **196,** 261.e1–261.e6 (2007).
122. Wang, W. *et al.* Preeclampsia Up-Regulates Angiogenesis-Associated MicroRNA (i.e., miR-17, -20a, and -20b) That Target Ephrin-B2 and EPHB4 in Human Placenta. *The Journal of clinical endocrinology and metabolism* **97,** E1051–E1059 (2012).
123. Xu, P. *et al.* Variations of MicroRNAs in Human Placentas and Plasma From Preeclamptic Pregnancy. *Hypertension* **63,** 1276–1284 (2014).
124. Zhu, X.-M., Han, T., Sargent, I. L., Yin, G.-W. & Yao, Y.-Q. Differential expression profile of microRNAs in human placentas from preeclamptic pregnancies vs normal pregnancies. *American journal of obstetrics and gynecology* **200,** 661.e1–7 (2009).
125. Burleigh, D. W. *et al.* Microarray Analysis of BeWo and JEG3 Trophoblast Cell Lines: Identification of Differentially Expressed Transcripts. *Placenta* **28,** 383–389 (2007).
126. Chen, H.-W. *et al.* Global analysis of differentially expressed genes in early gestational decidua and chorionic villi using a 9600 human cDNA microarray. *Molecular human reproduction* **8,** 475–484 (2001).
127. Chevillard, G., Derjuga, A., Devost, D., Zingg, H. H. & Blank, V. Identification of interleukin-1 regulated genes in uterine smooth muscle cells. *Reproduction* **134,** 811–822 (2007).
128. Fukushima, K. *et al.* Gene Expression Profiles by Microarray Analysis during Matrigel-induced Tube Formation in a Human Extravillous Trophoblast Cell line: Comparison

with Endothelial Cells. *Placenta* **29,** 898–904 (2008).
129. Popovici, R. M. *et al.* Gene expression profiling of human endometrial-trophoblast interaction in a coculture model. *Endocrinology* **147,** 5662–5675 (2006).
130. Sitras, V. *et al.* Differential Placental Gene Expression in Severe Preeclampsia. *Placenta* **30,** 424–433 (2009).
131. Soloff, M. S. *et al.* Effects of Progesterone Treatment on Expression of Genes Involved in Uterine Quiescence. *Reproductive sciences (Thousand Oaks, Calif.)* **18,** 781–797 (2011).
132. Brou, L. *et al.* Dysregulated biomarkers induce distinct pathways in preterm birth. *BJOG* **119,** 458–473 (2012).
133. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G.PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* **6,** e1000097 (2009).
134. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19,** 1639–1645 (2009).
135. Wickham, H. *ggplot2*. (Springer, 2016). doi:10.1007/978-3-319-24277-4

CHAPTER III

COMPARING HUMAN AND MACAQUE PLACENTAL TRANSCRIPTOMES TO DISENTANGLE PRETERM BIRTH PATHOLOGY FROM GESTATIONAL AGE EFFECTS[2]

Haley R. Eidem[a], David C. Rinker[a,b], William E. Ackerman IV[c], Irina A. Buhimschi[d], Catalin S. Buhimschi[c], Caitlin Dunn-Fletcher[e], Suhas G. Kallapur[d], Mihaela Pavličev[e], Louis J. Muglia[e], Patrick Abbot[a], and Antonis Rokas[a,b]

[a] *Department of Biological Sciences, Vanderbilt University, Nashville, TN 37234, USA*

[b] *Program in Human Genetics, Vanderbilt University, Nashville, TN 37235, USA*

[c] *Department of Obstetrics and Gynecology, The Ohio State University, Columbus, OH 43210, USA*

[d] *Center for Perinatal Research, Nationwide Children's Hospital, Columbus, OH 43210, USA*

[e] *Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH 45229, USA*

[2] This chapter was published in Placenta. 2016 May;41:74-82. doi: 10.1016/j.placenta.2016.03.006.

ABSTRACT

Introduction: A major issue in the transcriptomic study of spontaneous preterm birth (sPTB) in humans is the inability to collect healthy control tissue at the same gestational age (GA) to compare with pathologic preterm tissue. Thus, gene expression differences identified after the standard comparison of sPTB and term tissues necessarily reflect differences in both sPTB pathology *and* GA. One potential solution is to use GA-matched controls from a closely related species to tease apart genes that are dysregulated during sPTB from genes that are expressed differently as a result of GA effects.

Methods: To disentangle genes whose expression levels are associated with sPTB pathology from those linked to GA, we compared RNA sequencing data from human preterm placentas, human term placentas, and rhesus macaque placentas at 80% completed gestation (serving as healthy non-human primate GA-matched controls). We first compared sPTB and term human placental transcriptomes to identify significantly differentially expressed genes. We then overlaid the results of the comparison between human sPTB and macaque placental transcriptomes to identify sPTB-specific candidates. Finally, we overlaid the results of the comparison between human term and macaque placental transcriptomes to identify GA-specific candidates.

Results: Examination of relative expression for all human genes with macaque orthologs identified 267 candidate genes that were significantly differentially expressed between preterm and term human placentas. 29 genes were identified as sPTB-specific candidates and 37 as GA-specific candidates. Altogether, the 267 differentially expressed genes were

significantly enriched for a variety of developmental, metabolic, reproductive, immune, and inflammatory functions. Although there were no notable differences between the functions of the 29 sPTB-specific and 37 GA-specific candidate genes, many of these candidates have been previously shown to be dysregulated in diverse pregnancy-associated pathologies.

Discussion: By comparing human sPTB and term transcriptomes with GA-matched control transcriptomes from a closely related species, this study disentangled the confounding effects of sPTB pathology and GA, leading to the identification of 29 promising sPTB-specific candidate genes and 37 genes potentially related to GA effects. The apparent similarity in functions of the sPTB and GA candidates may suggest that the effects of sPTB and GA do not correspond to biologically distinct processes. Alternatively, it may reflect the poor state of knowledge of the transcriptional landscape underlying placental development and disease.

INTRODUCTION

Preterm birth (PTB), or birth before 37 completed weeks of gestation in humans, is a global health issue affecting at least 15 million newborns every year[1-3]. This complex, multifactorial syndrome accounts for around 1 million neonatal deaths annually and surviving neonates often require lifelong care for common comorbidities including developmental, visual, and digestive problems[4,5]. 30% of PTB cases are indicated by medical conditions such as preeclampsia or intrauterine growth restriction, while the remaining 70% are caused by the spontaneous onset of labor either with (25%) or without (45%) premature membrane rupture[6,7].

Spontaneous, idiopathic preterm birth (sPTB), much like most other complex human genetic diseases, is augmented by environmental risk factors (e.g., stress, infection, and socioeconomic status) as well as by genetics. Several studies have shown that women are more likely to deliver preterm if a sister delivered preterm, if a previous child was born preterm, if they were born preterm themselves, or if they have African American ancestry[8-11]. In recent years, studies have also highlighted the importance of gene expression regulation in complex genetic diseases[12]. Thus, analysis of the genetic elements that are active or dysregulated in gestational tissues harbors great potential to identify candidate genes for sPTB and several genome-wide studies have already started to outline its genomic, transcriptomic, and methylomic architecture[13].

Nevertheless, a major obstacle in the transcriptomic study of sPTB in humans is the inability to collect gestational age-matched healthy control tissue to compare with pathologic preterm tissue. Without safe, non-invasive procedures to sample healthy preterm tissues destined for healthy term births, the most common approach is to use healthy term tissues as the control for pathologic preterm tissues[14-16]. This complicates downstream data analysis, though, because observed differences in gene expression reflect not only differences in pathology, but also differences in gestational age (GA).

One potential solution is to use GA-matched controls from a closely related species to distinguish genes dysregulated during sPTB from genes expressed differently at different points in pregnancy. The decoding of the rhesus macaque (*Macaca mulatta*) genome and subsequent comparison with that of human and chimpanzee revealed that these 3 primate species share about 93% of their DNA[17]. Thus, macaque is an ideal species for transcriptional comparison with humans not only because the two species share a close evolutionary affinity but also because of similarities with respect to key pregnancy-related traits. For example, even though placental morphology is highly variable across mammals, human and macaque placentas share the same discoid shape, hemochorial invasiveness, and villous interdigitation[18,19]. Similarly, the relationship between pelvis and fetal head size in humans is more akin to the relationship in macaques than it is to any other primates[20]. This is particularly important as it would alleviate any effects that cephalopelvic constraints might have on birth

timing[21]. Finally, several human pregnancy pathologies have been recorded in macaques including stillbirth, PTB, placenta previa, and placental abruption[22].

In this study, we compare transcriptomes from human sPTB placentas, human term placentas, and macaque placentas at 80% completed gestation to distinguish between sPTB-specific and GA-specific candidate genes. Specifically, candidate genes that are differentially expressed between human sPTB and human term as well as between human sPTB and macaque are potentially sPTB-specific. In contrast, candidate genes that are differentially expressed between human sPTB and human term as well as between human term and macaque are potentially GA-specific. This novel comparative approach disentangles the confounding effects of sPTB and GA differences and allows for the educated prioritization of candidate genes for future studies of pregnancy and prematurity.

RESULTS

RNA sequencing and transcriptome analysis

Analysis of 311.9 million reads from 5 term and 5 preterm human placental samples resulted in the successful mapping of 298.4 million reads (95%). 265.9 million (85%) of these mapped to a single position in the human genome and 96.9 million (32%) mapped to human genes that share 1:1 orthology with macaque genes. Analysis of 29.8 million reads from 2 macaque placental samples at 80% completed gestation resulted in the successful mapping of 28.7 million read pairs (96%). 26.6 million (89%) of these mapped to a single position in the macaque genome and 25.1 million (87%) mapped to macaque genes that share 1:1 orthology with human genes. Notably, the correlation coefficient between human sPTB samples ranges from 0.84 to 0.95, between human term samples ranges from 0.95 to 0.98, and between rhesus samples is 0.89 (Additional File 3.2). This degree of variation between samples is on par with the degree of variation reported between samples from other much more homogeneous tissues[23].

Distinguishing sPTB-specific and GA-specific candidate genes

To identify sPTB-specific and GA-specific candidate human genes, we performed three pairwise differential expression comparisons: human sPTB vs. human term, human sPTB vs. macaque (GA-matched control), and human term vs. macaque (healthy, early gestation comparison) (Figure 3.1a). We identified 267 genes that were differentially expressed between human sPTB and term (149 over-expressed and 118 under-expressed), 12,379 genes

that were differentially expressed between human sPTB and macaque (7,193 over-expressed and 5,186 under-expressed), and 12,566 genes that were differentially expressed between human term and macaque (7,285 over-expressed and 5,281 under-expressed) (Additional File 3.5).

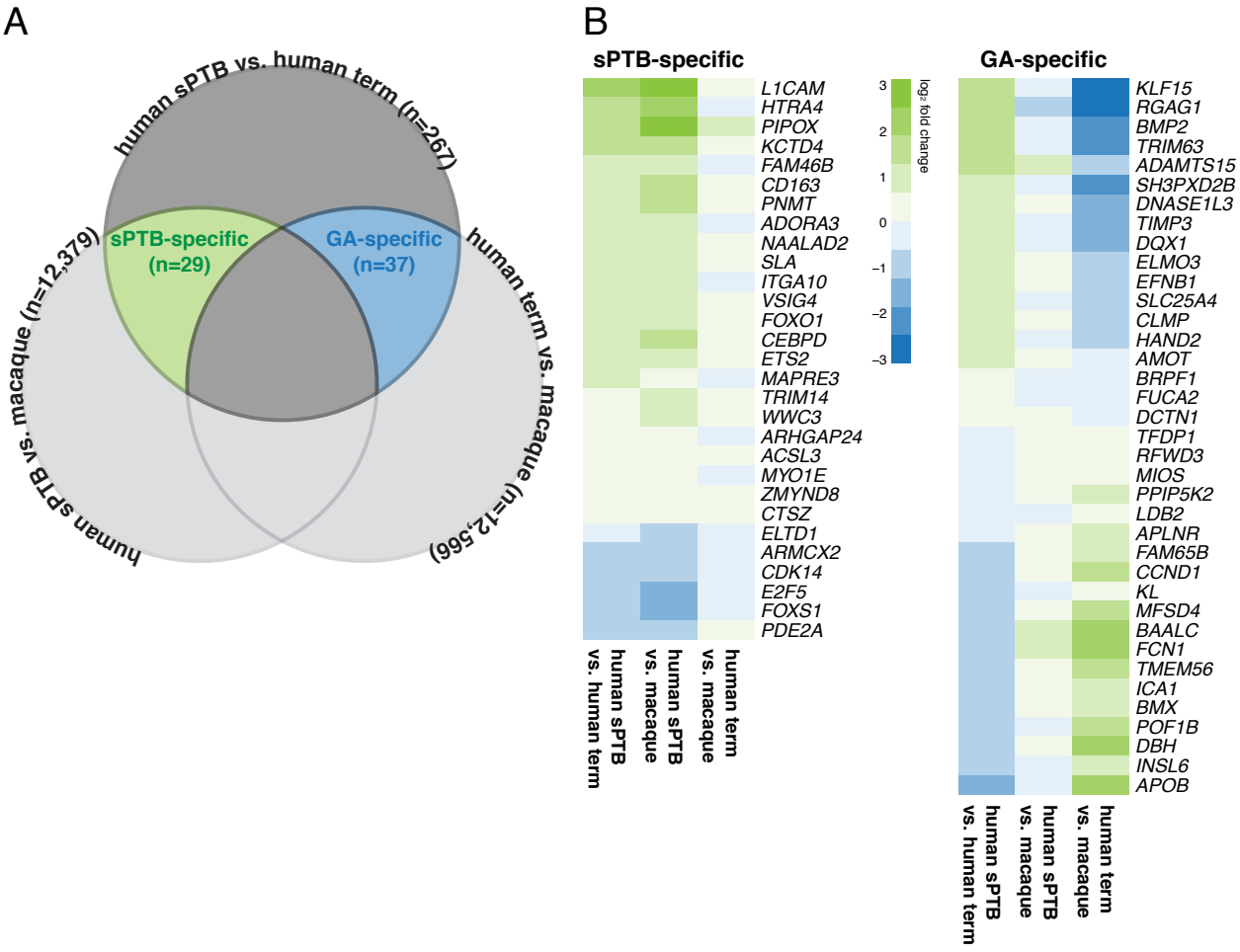

Figure 3.1. A comparative approach to disentangling the effects of PTB and GA in human transcriptomic studies. (A) RNA-seq data from macaque placentas collected at 80% gestation serves as healthy nonhuman primate gestational age controls to disentangle the involvement of sPTB from that of gestational age when comparing human sPTB and term placentas. Differentially expressed genes were first identified between human sPTB and human term placentas and then intersected with differentially expressed genes from 2 other

pairwise comparisons (human sPTB vs. macaque and human term vs. macaque). This intersection allowed for the categorization of the initially identified differentially expressed human sPTB vs. human term genes (dark grey, n=267) as sPTB-specific (green, n=29) or GA-specific (blue, n=37). (B) For each of the 3 pairwise transcriptomic comparisons, log_2 fold change values were plotted corresponding to each of the 29 sPTB-specific gene candidates as well as the 37 GA-specific gene candidates. For consistency, the macaque transcriptomes were used as controls in both inter-species comparisons and, therefore, the direction of fold changes in the human term vs. macaque comparison for GA-specific genes is reversed.

Taken alone, differentially expressed genes from the human sPTB and human term comparison presumably represent expression differences that may be attributable to either sPTB pathology or GA. To further distinguish the 267 differentially expressed genes between those that represented sPTB-specific candidates or GA-specific candidates, we intersected the candidate gene results of all three differential expression experiments. This approach allowed for the identification of 29 sPTB-specific candidate genes and 37 GA-specific candidate genes (Figure 3.1b, Table 3.1, Table 3.2, Additional File 3.6). 23/29 (70%) of the sPTB-specific genes were over-expressed in sPTB and only 6/29 (21%) were under-expressed. In contrast, 18/37 (49%) of the GA-specific genes were over-expressed in sPTB and 19/37 (51%) were under-expressed (Figure 3.1b, Additional File 3.6).

Table 3.1. sPTB-specific genes.

Symbol	Description	Fold change (log2)*	Adjusted p-value*
L1CAM	L1 cell adhesion molecule	2.00	5.63E-05
PNMT	phenylethanolamine N-methyltransferase	1.02	3.58E-04
MYO1E	myosin 1E	0.39	5.42E-04
ARMCX2	armadillo repeat containing, X-linked 2	-0.66	1.24E-03

PDE2A	phosphodiesterase 2A, cGMP-stimulated	-1.03	1.44E-03
MAPRE3	microtubule-associated protein, RP/EB family, member 3	0.65	2.47E-03
CD163	CD163 molecule	1.13	5.98E-03
ETS2	v-ets avian erythroblastosis virus E26 oncogene homolog 2	0.70	1.22E-02
CTSZ	cathepsin Z	0.32	1.22E-02
CDK14	cyclin-dependent kinase 14	-0.76	1.22E-02
VSIG4	V-set and immunoglobin domain containing 4	0.81	1.50E-02
NAALAD2	N-acetylated alpha-linked acidic dipeptidase 2	0.89	1.75E-02
E2F5	E2F transcription factor 5, p130-binding	-0.92	2.44E-02
HTRA4	HtrA serine peptidase 4	1.50	2.63E-02
KCTD4	potassium channel tetramerization domain containing 4	1.26	2.83E-02
TRIM14	tripartite motif-containing 14	0.60	3.19E-02
ZMYND8	zinc finger, MYND-type containing 8	0.38	3.72E-02
ARHGAP24	Rho GTPase activating protein 24	0.54	3.79E-02
FOXO1	forkhead box 01	0.80	4.06E-02
FOXS1	forkhead box S1	-1.01	4.43E-02
ITGA10	integrin, alpha 10	0.83	4.93E-02
SLA	Src-like-adaptor	0.85	4.97E-02
ACSL3	acyl-CoA synthetase long-chain family member 3	0.41	4.97E-02
FAM46B	family with sequence similarity 46, member B	1.18	6.46E-02
ELTD1	adhesion G protein-coupled receptor L4	-0.57	6.51E-02
WWC3	WWC family member 3	0.55	6.53E-02
CEBPD	CCAAT/enhancer binding protein (C/EBP), delta	0.80	7.03E-02
ADORA3	adenosine A3 receptor	1.01	7.09E-02
PIPOX	pipecolic acid oxidase	1.27	7.40E-02

*values from human sPTB vs. human term comparison

Table 3.2. GA-specific genes.

Symbol	Description	Fold change (log2)*	Adjusted p-value*
ICA1	islet cell autoantigen 1, 69kDa	-1.08	2.49E-04
EFNB1	ephrin-B1	0.78	3.58E-04
TMEM56	transmembrane protein 56	-1.05	3.58E-04
KL	klotho	-0.83	1.62E-03
BMX	BMX non-receptor tyrosine kinase	-1.09	2.47E-03
BMP2	bone morphogenetic protein 2	1.53	2.49E-03
SH3PXD2B	SH3 and PX domains 2B	1.19	2.87E-03
KLF15	Kruppel-like factor 15	1.76	3.05E-03
TIMP3	TIMP metallopeptidase inhibitor 3	0.93	5.69E-03
DCTN1	dynactin 1	0.25	5.83E-03
PPIP5K2	diphosphoinositol pentakisphosphate kinase 2	-0.46	5.98E-03
RGAG1	retrotransposon gag domain containing 1	1.54	8.67E-03
ADAMTS15	ADAM metallopeptidase with thrombospondin type 1 motif, 15	1.29	1.29E-02
ELMO3	engulfment and cell motility 3	0.81	1.29E-02
INSL6	insulin-like 6	-1.26	1.33E-02
APOB	apolipoprotein B	-1.55	1.75E-02
LDB2	LIM domain binding 2	-0.57	2.02E-02
APLNR	apelin receptor	-0.62	2.41E-02
POF1B	premature ovarian failure, 1B	-1.12	2.66E-02
MIOS	missing oocyte, meiosis regulator, homolog (Drosophila)	-0.38	2.97E-02
MFSD4	major facilitator superfamily domain containing 4	-0.84	3.83E-02
RFWD3	ring finger and WD repeat domain 3	-0.37	4.01E-02
AMOT	angiomotin	0.69	4.30E-02
DQX1	DEAQ box RNA-dependent ATPase 1	0.92	4.32E-02
DBH	dopamine beta-hydroxylase (dopamine beta-monooxygenase)	-1.26	4.37E-02
TRIM63	tripartite motif containing 63, E3 ubiquitin protein ligase	1.40	4.43E-02

SLC25A4	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 4	0.78	4.63E-02
BRPF1	bromodomain and PHD finger containing, 1	0.31	5.22E-02
FCN1	ficolin (collagen/fibrinogen domain containing) 1	-0.99	5.50E-02
DNASE1L3	deoxyribonuclease I-like 3	1.14	6.31E-02
BAALC	brain and acute leukemia, cytoplasmic	-0.92	6.70E-02
CCND1	cyclin D1	-0.74	7.40E-02
TFDP1	transcription factor Dp-1	-0.29	8.20E-02
HAND2	hand and neural crest derivatives expressed 2	0.71	8.61E-02
CLMP	CXADR-like membrane protein	0.74	8.69E-02
FAM65B	family with sequence similarity 65, member B	-0.71	8.69E-02
FUCA2	fucosidase, alpha-L-2, plasma	0.25	9.45E-02

*values from human sPTB vs. human term comparison

Differentially expressed genes were enriched for developmental and metabolic functions

All together, these 267 genes were significantly enriched for involvement in a wide variety of developmental processes (e.g., GO:0005515) (Additional File 3.7). Furthermore, the gene set as a whole was enriched for function in metabolism, reproduction, immunity, inflammation, and cell signaling. Although statistical significance is limited due to small gene set size, the 29 sPTB-specific genes were involved in transcription factor activity and binding (e.g., GO:0008134 and GO:0005515) and the 37 GA-specific genes were involved in development and response to stimuli (e.g., GO:0007275 and GO:0009628).

Placental sPTB-specific and GA-specific candidate genes are heterogeneously expressed

Given the high heterogeneity of gene expression that has been previously reported in placental transcriptome studies, we evaluated the transcriptional profile of sPTB-specific and GA-specific candidate genes by comparing their relative expression to publicly available expression data from Protein Atlas[13,24]. 27 of the 29 sPTB-specific candidates and 34 of the 37 GA-specific candidates had been previously identified for placental gene expression at varying levels in Protein Atlas (Figure 3.2, Additional File 3.8). Although mean expression levels of these candidate genes in human sPTB, human term, and macaque placentas were generally comparable to expression levels presented by Protein Atlas, our data also reflect the known variability in placental gene expression both within and between individuals[24]. For example, *HTRA4* was highly expressed across our sPTB transcriptomes (FPKM=75.4) and lowly expressed across our term transcriptomes (FPKM=11.7), but Protein Atlas reported an intermediate expression level (FPKM=43.8).

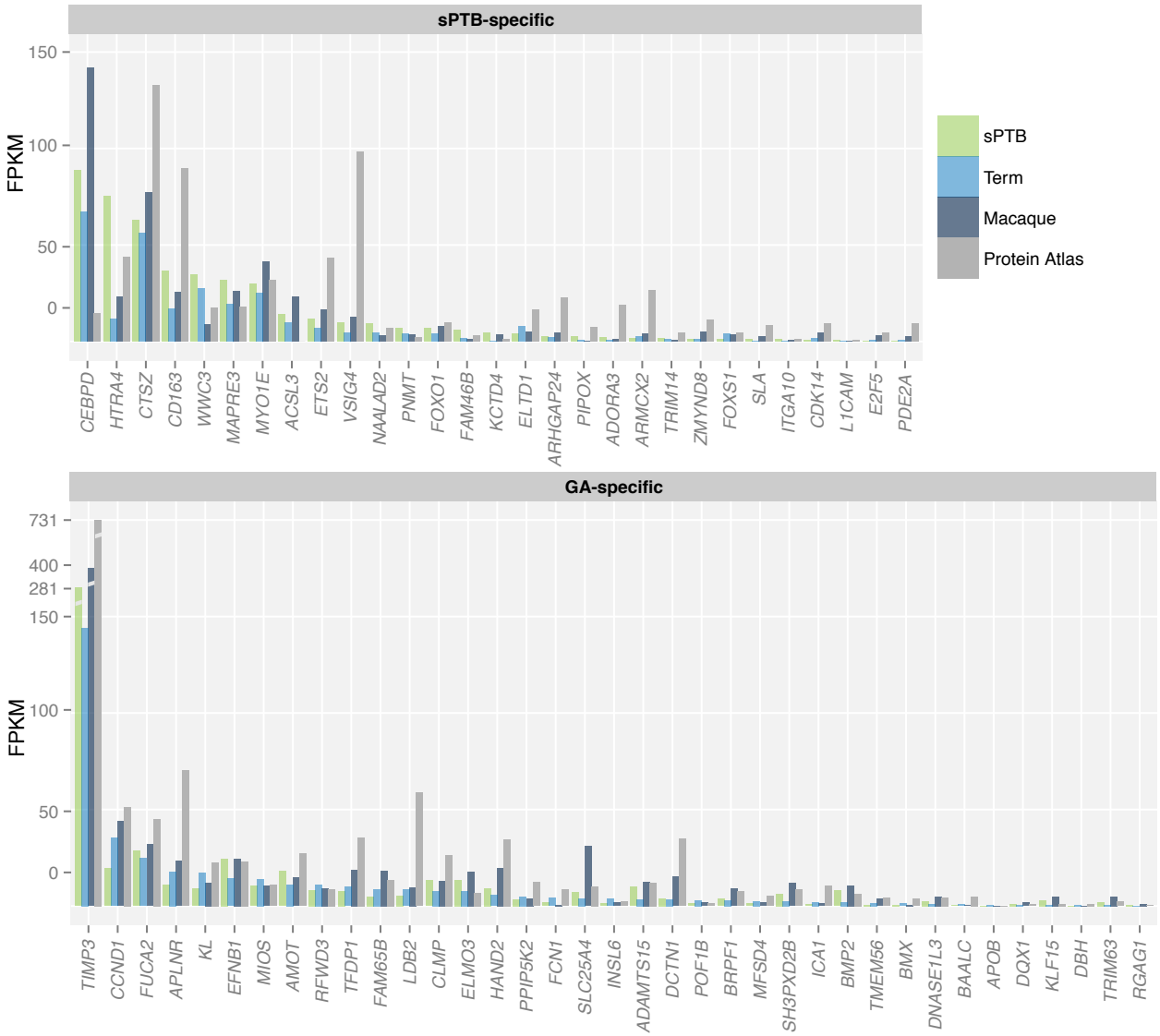

Figure 3.2. sPTB-specific and GA-specific candidate genes show heterogeneous expression patterns in placenta. For sPTB-specific (top panel) and GA-specific (bottom panel) candidate genes, mean mRNA expression in Fragments Per Kilobase of transcript per Million mapped reads (FPKM) was compared between the 5 human sPTB, 5 human term, and 2 macaque placental samples as well as to data from Protein Atlas.

DISCUSSION

A major obstacle in the transcriptomic study of PTB in humans is the inability to collect healthy, GA-matched control tissue samples that facilitate the comparison of preterm diseased tissue to healthy tissue from a corresponding point in pregnancy. This complicates interpretation of results since candidate genes necessarily reflect both sPTB pathology as well as GA differences. Our comparative approach utilizing the transcriptomic profiles of rhesus macaque placental samples from 80% gestation overlaid with transcriptomic profiles of human preterm and term placental tissue allows for the disentanglement of these variables and, thus, the identification of genes with roles specific to sPTB pathology or GA.

Generally, most genes in the sPTB-specific and GA-specific categories have previously been annotated for placental tissue expression in Protein Atlas, sometimes uniquely so, although the level of expression often differed, in line with previous work on the high heterogeneity of placental gene expression (Figure 3.2)[24]. Moreover, the 29 genes in the sPTB-specific category included several previously identified for involvement in pregnancy pathologies. For example, *HTRA4*, a serine peptidase, is over-expressed during early-onset preeclampsia[25]. *CD163*, a hemoglobin scavenger receptor expressed exclusively in macrophages, is over-expressed in the preterm preeclamptic decidua and has also been identified as a predictor of preterm birth in maternal serum[26,27]. The presence of *CD163* among differentially expressed genes in our data may indicate that some maternal decidual tissue (and thus, macrophages) was inadvertently captured during placental biopsy. *ADORA3*, an adenosine receptor, is over-

expressed in preeclamptic trophoblasts and also modulates secretion of matrix metalloproteinases that serve as important components in PPROM signaling pathways[28]. *VSIG4* encodes a protein with an immunoglobulin domain that has been characterized as a maternal biomarker of preeclampsia[29,30]. *PDE2A* is a phosphodiesterase and the gene contains a SNP associated with idiopathic recurrent miscarriage[31]. Finally, *NAALAD2* has been shown to be under-expressed in decidua during preeclampsia[32]. Our results indicate that pathways previously identified as involved in other pregnancy pathologies may also be involved in sPTB pathogenesis.

The 37 genes in the GA-specific category also included several previously identified for involvement in pregnancy pathologies. For example, *BMP2*, an extracellular growth factor, is over-expressed in fetal membranes during spontaneous term labor and preterm labor with chorioamnionitis[33,34]. *TIMP3*, a matrix metallopeptidase inhibitor, is expressed in fetal membranes during labor and has also been shown to be hypomethylated during preeclampsia[35-37]. *APLNR*, a G-protein coupled receptor, is a key receptor of apelin, a gene that is under-expressed during term and preterm labor in amnion as well as during preeclampsia in placenta[38,39]. *KL*, or klotho, is under-expressed in pregnancies where the neonate is small-for-gestational-age[40]. *INSL6* is a member of the relaxin family of peptide hormones and, although little is known about *INSL6* specifically, relaxin expression at the fetomaternal interface has been linked to PPROM pathogenesis and serum relaxin concentration has been identified as a potential PTB biomarker[41,42]. *EFNB1* may play a role in cell adhesion and has

been shown to be differentially expressed during preeclampsia[43]. *ADAMTS15* is a member of a disintegrin and metalloproteinase with thrombospondin motifs protein family and has been previously identified as dysregulated during PPROM[44]. *KLF15* has been shown to be under-expressed in the preeclamptic decidua[32]. Finally, *APOB* is a low density lipoproteins and variation in the gene sequence has been associated with preterm delivery[45]. Our identification of these genes as GA candidates raises the hypothesis that they might be differentially expressed due to differences in gestational age of the tissues being compared rather than due to the underlying pathology. Alternatively, these genes might be involved in both disease *and* development.

In addition to involvement in pregnancy pathologies, some sPTB-specific candidate genes have been annotated for involvement in more general biological processes related to healthy pregnancy, labor, and placentation. For example, *FOXO1*, is a potential modulator of inflammatory events in the myometrium during labor and its expression in endometrium is regulated by progesterone[46,47]. *CEBPD* is a transcription factor and, through its interaction with *CEBPA*, is involved in the regulation of immune and inflammatory responses in various gestational tissues as well as in the development of fetal lungs[48,49]. *ETS2* has been shown to mediate matrix metalloproteinase activity and trophoblast invasion[50]. *ARHGAP24* is a RHOGTPase activator involved in myometrium contractility and shows increased mRNA expression in myometrium during labor[51]. *E2F5* is a member of the *E2f* transcription factor family and has been shown to help coordinate placental development in mice[52]. *FAM46B* has

been shown to be under-expressed in myometrium and cervix during labor[48]. Finally, *PNMT* is an enzyme known to be expressed in myometrium and fetal membranes that shows decreased enzymatic activity in myometrium during labor[53-55]. GA-specific candidate genes have similarly been annotated for involvement in biological processes related to healthy pregnancy and labor. For example, *HAND2* is a progesterone-regulated transcription factor that is expressed in myometrium and essential for embryo implantation and *FCN1* is involved in innate immune defense and over-expressed in myometrium and cervix during labor[40,48].

Comparison of the functions of sPTB-specific candidate genes with those of GA-specific candidate genes, however, does not identify any notable differences. For example, both sPTB-specific and GA-specific candidate genes include developmental genes (GO:0032502) (e.g., *E2F5*, *ARHGAP24*, *ETS2*, *ACSL3*, *MYO1E*, *PDE2A*, *FOXS1*, *CTSZ*, *L1CAM*, and *FOXO1* for sPTB, and *CLMP*, *FAM65B*, *CCND1*, *KL*, *KLF15*, *BMX*, *INSL6*, *LDB2*, *EFNB1*, *HAND2*, *APOB*, *BMP2*, *DNASE1L3*, *SH3PXD2B*, *TIMP3*, *APLNR*, *AMOT*, and *DCTN1* for GA, respectively), genes involved in immunity-related functions (GO:0002376) (e.g., *L1CAM*, *TRIM14*, *FOXO1*, and *VSIG4* for sPTB, and *KL*, *DBH*, *BMX*, *APOB*, *DCTN1*, and *FCN1* for GA, respectively), and genes involved in stress response (GO:0006950) (e.g., *TRIM14*, *CD163*, *FOXO1*, and *VSIG4* for sPTB, and *FCN1*, *CCND1*, *BMP2*, *KL*, *DBH*, *KLF15*, *BMX*, *MIOS*, *RFWD3*, and *DCTN1* for GA, respectively).

Similarity between the functions of sPTB-specific and GA-specific candidates may genuinely reflect the idea that dysregulation during sPTB pathology is not a biologically separate and

distinct process from GA. Alternatively, the overlap of functions may reflect our poor state of knowledge of the transcriptional landscape underlying placental development and disease. For example, the vast majority of transcriptomic studies on placental tissues have traditionally focused on identifying the impact of environmental factors or known clinical subtypes in placental gene expression, heavily biasing any knowledge of the functions of placental genes toward disease (as opposed to developmental or physiological) phenotypes[13]. Furthermore, only a handful of studies have mapped the transcriptional changes occurring during placental development and, given the inability to collect healthy tissue from multiple time points, these studies are typically limited to comparing the transcriptional landscapes of first and third-trimester placentas[56-58]. Interestingly, only 2 of the 37 GA-specific genes in this study (*LDB2* and *BMP2*) have been previously annotated as differentially expressed between early and late pregnancy[56,57].

Although our comparative analysis allowed for the more detailed categorization of otherwise general candidate genes, comparison of transcriptomes across species has several important caveats. For example, biological and analytical differences such as those stemming from alternative splicing, annotation heterogeneity, and genetic variation are potential sources of 'noise'[59-61]. The development and increasingly wide-spread use of RNA-seq, however, has facilitated a more straightforward inter-species transcriptome comparison due to the breadth and depth of expression data generated by this approach[23,61-63]. Like previous inter-species transcriptomics experiments, our comparative analysis was limited only to 1:1 orthologs in

order to facilitate sensible comparison and, even further, the human-macaque transcriptome comparisons were overlaid with a direct human-human comparison as an attempt to filter out potentially large amounts of expected gene expression differences. Still, the results of inter-species transcriptome comparisons must be carefully interpreted due to the many inherent differences in gene expression and its regulation across species.

Additionally, this study would have benefited from the inclusion of term macaque placental tissue samples. Although unavailable at the time of analysis, these samples would allow for an understanding of gene expression in term macaque placentas that could then be compared with the other groups in our analysis. Furthermore, gestation length and the timing of some early developmental events differ between human and macaque, although little is known about comparative placentation towards the end of pregnancy[22]. Therefore, the use of placenta collected at 80% completed gestation in macaque may not be the ideal time point for comparison with human sPTB.

These caveats notwithstanding, our results demarcate how a comparative transcriptomics approach to the study of human sPTB allows for the identification and prioritization of candidate genes and pathways involved specifically in sPTB pathogenesis and GA changes during pregnancy. Although a handful of recent studies have analyzed genome-wide gene expression in the human term (and occasionally preterm) placenta, ethical issues prevent comparisons of sPTB tissue to a healthy tissue at the same GA in humans. Our novel,

comparative approach is the first to utilize GA-matched control placental tissue from a closely related species for comparison with human term and preterm placental tissue samples. Despite the challenges inherent in inter-species transcriptome comparisons, the use of RNA-seq data and well-annotated reference genomes makes possible the human-macaque comparison and ultimate prioritization of otherwise convoluted differentially expressed gene sets as promising sPTB-specific candidates.

METHODS

Tissue collection, RNA isolation, and RNA sequencing

Human placentas were collected immediately after delivery and the decidua basalis layer from a central cotyledon was dissected and discarded. Approximately 1g of underlying villous tissue was then biopsied for further analysis. The 5 term (GA 38-39 weeks, mean 38 or 95% completed gestation) human placental tissue biopsies were all collected after cesarean delivery. Of the 5 preterm (GA 29-33 weeks, mean 32 or ~80% completed gestation) human placental tissue biopsies that were collected, 4 were collected after cesarean delivery and 1 after vaginal delivery. Each of the biopsies was flash frozen in liquid nitrogen and stored at -80°C[64]. Total RNA was isolated using TRIzol and Illumina libraries were constructed using the TruSeq Stranded Total RNA Sample Prep Kit with Ribo-Zero Gold. RNA sequencing (RNA-seq) was performed on an Illumina HiSeq 2500 machine using HiSeq version 3 sequencing reagents. The samples were sequenced using a single-end approach with 50bp reads, generating approximately 30 million reads per sample. Raw count data have been deposited in the NCBI Gene Expression Omnibus (GEO) database and are accessible through GEO Series accession number GSE73714 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73714).

Macaque placentas were collected immediately after delivery and full thickness biopsies (~2cmx2cm) free of clots and debris were taken midway between the attachment of the umbilical cord and the placenta edge. Both placental tissue samples were collected after

cesarean delivery via hysterotomy (GA 128-131 days, mean 129.5 or ~80% completed gestation)[65]. Total RNA was isolated from 100µg of frozen tissue using TRIzol and suspensions were stored at -80°C. RNA-seq was performed on an Illumina HiSeq machine after passing initial quality control metrics. The two samples were sequenced using a paired-end approach with 50 bp reads, generating approximately 15 million paired reads per sample.

Data processing and differential expression analysis

RNA-seq data was analyzed from the 5 term and 5 preterm human placental tissue samples as well as the 2 macaque placental samples from 80% completed gestation. Raw sequence read files were first quality checked using FastQC and then trimmed for quality and adapter sequences using Trimmomatic with the default parameters (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36)[66]. Across all 10 human cDNA libraries, a total of 311.9 million reads were sequenced and subsequently mapped to the hg19 human reference genome using TopHat2 (Additional File 3.1, Additional File 3.2)[67]. Across the 2 macaque cDNA libraries, a total of 29.8 million read pairs were sequenced and subsequently mapped to the Mmul 1.0 macaque reference genome using TopHat2 (Additional File 3.1, Additional File 3.2).

Given our interspecies transcriptome comparisons, we restricted all analyses to 19,063 one-to-one (1:1) human-macaque orthologs obtained from Zimin et al.[68]. Specifically, only sequence reads that uniquely aligned to these 19,063 human or macaque genes were counted using HTSeq, resulting in 18,879 genes with expression values in both species

(Additional File 3.3, Additional File 3.4). DESeq2 was used to quantify relative gene expression differences in terms of fold change (log2) and statistical significance (Benjamini Hochberg-corrected p-values), where positive fold change represents over-expression in human sPTB compared to human term (or human sPTB/term compared to macaque) and negative fold change represents under-expression in human sPTB compared to human term (or human sPTB/term compared to macaque)[69]. Genes were annotated as differentially expressed if the adjusted p-value was < 0.1.

Functional enrichment analysis

GO Biological Process term enrichment was calculated using the Cytoscape plugin BiNGO[70]. Terms were considered significant if the adjusted p-value was < 0.1 after the Benjamini-Hochberg multiple testing correction. FPKM for each gene was calculated across all human sPTB, human term, and macaque transcriptomes and associated RNA and protein expression data was extracted from Protein Atlas using GEneSTATION[71-73].

AUTHORS' CONTRIBUTIONS

HRE and AR designed the study. WEA, IAB, and CSB provided human RNA-seq data and MP, CDF, SGK and LJM provided macaque RNA-seq data. HRE performed the analysis with assistance from DCR. HRE drafted the manuscript with contributions and revisions from all other authors. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Jennifer Wisecaver for providing custom functional annotation Perl scripts. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. HRE was supported by the Graduate Program in Biological Sciences at Vanderbilt University and by a 2016 March of Dimes Transdisciplinary Scholar Award. Research on this project was supported by the March of Dimes through the March of Dimes Prematurity Research Center Ohio Collaborative.

REFERENCES

1. Muglia, L. J. & Katz, M. The enigma of spontaneous preterm birth. *N. Engl. J. Med.* **362,** 529–535 (2010).
2. Bezold, K. Y., Karjalainen, M. K., Hallman, M., Teramo, K. & Muglia, L. J. The genomics of preterm birth: from animal models to human studies. *Genome Med* **5,** 34 (2013).
3. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: one syndrome, many causes. *Science* **345,** 760–765 (2014).
4. Blencowe, H. *et al.* National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet* **379,** 2162–2172 (2012).
5. Liu, L. *et al.* Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *The Lancet* (2014). doi:10.1016/S0140-6736(14)61698-6
6. Rubens, C. E. *et al.* Prevention of preterm birth: harnessing science to address the global epidemic. *Sci Transl Med* **6,** 262sr5–262sr5 (2014).
7. Swaggart, K. A., Pavlicev, M. & Muglia, L. J. Genomics of preterm birth. *Cold Spring Harb Perspect Med* **5,** a023127 (2015).
8. Crider, K. S., Whitehead, N. & Buus, R. M. Genetic variation associated with preterm birth: a HuGE review. *Genet Med* (2005). doi:10.1097/01.gim.0000187223.69947.db
9. Plunkett, J. & Muglia, L. J. Genetic contributions to preterm birth: implications from epidemiological and genetic association studies. *Ann. Med.* **40,** 167–195 (2008).
10. Plunkett, J. *et al.* An Evolutionary Genomic Approach to Identify Genes Involved in Human Birth Timing. *PLoS genetics* **7,** e1001365 (2011).
11. Lengyel, C., Muglia, L. J. & Pavlicev, M. *Genetics of Preterm Birth*. **283,** (John Wiley & Sons, Ltd, 2001).
12. Emilsson, V., Thorleifsson, G., Zhang, B. & Leonardson, A. S. Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
13. Eidem, H. R., Ackerman, W. E., McGary, K. L., Abbot, P. & Rokas, A. Gestational tissue transcriptomics in term and preterm human pregnancies: a systematic review and meta-analysis. *BMC Medical Genomics* **8,** 27 (2015).
14. Elovitz, M. A. *et al.* Distinct cervical microRNA profiles are present in women destined to have a preterm birth. *American journal of obstetrics and gynecology* **210,** 221.e1–11 (2014).
15. Chim, S. S. C. *et al.* Systematic Identification of Spontaneous Preterm Birth-Associated RNA Transcripts in Maternal Plasma. *PLoS ONE* **7,** e34328 (2012).
16. Mayor-Lynn, K., Toloubeydokhti, T., Cruz, A. C. & Chegini, N. Expression profile of microRNAs and mRNAs in human placentas from pregnancies complicated by preeclampsia and preterm labor. *Reproductive sciences (Thousand Oaks, Calif.)* **18,** 46–56 (2011).
17. Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.* Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science* **316,** 222–234 (2007).

18. Wildman, D. E. *et al.* Evolution of the mammalian placenta revealed by phylogenetic analysis. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 3203–3208 (2006).
19. Wildman, D. E. Review: Toward an integrated evolutionary understanding of the mammalian placenta. *Placenta* **32 Suppl 2,** S142–5 (2011).
20. Schultz, A. H. *The life of primates*. (1969).
21. Wells, J. C. K., DeSilva, J. M. & Stock, J. T. The obstetric dilemma: An ancient game of Russian roulette, or a variable dilemma sensitive to ecology? *Am. J. Phys. Anthropol.* **149,** 40–71 (2012).
22. de Rijk, E. P. C. T. & Van Esch, E. The Macaque Placenta—A Mini-Review. *Toxicologic Pathology* **36,** 108S–118S (2008).
23. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).
24. Hughes, D. A. *et al.* Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome biology* **16,** 54 (2015).
25. Singh, H. *et al.* Human HtrA4 Expression Is Restricted to the Placenta, Is Significantly Up-Regulated in Early-Onset Preeclampsia, and High Levels of HtrA4 Cause Endothelial Dysfunction. *The Journal of clinical endocrinology and metabolism* **100,** E936–45 (2015).
26. Schonkeren, D. *et al.* Differential distribution and phenotype of decidual macrophages in preeclamptic versus control pregnancies. *Am. J. Pathol.* **178,** 709–717 (2011).
27. Vogel, I. *et al.* Preterm delivery predicted by soluble CD163 and CRP in women with symptoms of preterm delivery. *BJOG* **112,** 737–742 (2005).
28. Kim, Y.-H., Hwang, H.-S., Kim, Y.-T., Kim, H.-S. & Park, Y.-W. Modulation of matrix metalloproteinase secretion by adenosine A3 receptor in preeclamptic villous explants. *Reproductive sciences (Thousand Oaks, Calif.)* **15,** 939–949 (2008).
29. Chaiworapongsa, T. *et al.* Differences and similarities in the transcriptional profile of peripheral whole blood in early and late-onset preeclampsia: insights into the molecular basis of the phenotype of preeclampsiaa. *Journal of Perinatal Medicine* **41,** (2013).
30. Textoris, J. *et al.* Evaluation of Current and New Biomarkers in Severe Preeclampsia: A Microarray Approach Reveals theVSIG4 Gene as a Potential Blood Biomarker. *PLoS ONE* **8,** e82638 (2013).
31. Kolte, A. M. *et al.* A genome-wide scan in affected sibling pairs with idiopathic recurrent miscarriage suggests genetic linkage. *Molecular human reproduction* **17,** 379–385 (2011).
32. Løset, M. *et al.* A transcriptional profile of the decidua in preeclampsia. *American journal of obstetrics and gynecology* **204,** 84.e1–27 (2011).
33. Haddad, R. *et al.* Human spontaneous labor without histologic chorioamnionitis is characterized by an acute inflammation gene expression signature. *American journal of obstetrics and gynecology* **195,** 394–405.e12 (2006).
34. Kim, G. J. *et al.* Expression of bone morphogenetic protein 2 in normal spontaneous labor at term, preterm labor, and preterm premature rupture of membranes. *The*

American Journal of Obstetrics & Gynecology **193,** 1137–1143 (2005).
35. Fortunato, S. J., Menon, R. & Lombardi, S. J. Presence of four tissue inhibitors of matrix metalloproteinases (TIMP-1, -2, -3 and -4) in human fetal membranes. *Am. J. Reprod. Immunol.* **40,** 395–400 (1998).
36. Xiang, Y. *et al.* Promoter hypomethylation of TIMP3 is associated with pre-eclampsia in a Chinese population. *Molecular human reproduction* **19,** 153–159 (2013).
37. Yuen, R. K., Penaherrera, M. S., Dadelszen, von, P., McFadden, D. E. & Robinson, W. P. DNA methylation profiling of human placentas reveals promoter hypomethylation of multiple genes in early-onset preeclampsia. *Eur. J. Hum. Genet.* **18,** 1006–1012 (2010).
38. Lim, R., Barker, G., Riley, C. & Lappas, M. Apelin is decreased with human preterm and term labor and regulates prolabor mediators in human primary amnion cells. *Reproductive sciences (Thousand Oaks, Calif.)* **20,** 957–967 (2013).
39. Inuzuka, H. *et al.* Decreased expression of apelin in placentas from severe pre-eclampsia patients. *Hypertension in Pregnancy* **32,** 410–421 (2013).
40. Miranda, J. *et al.* The anti-aging factor α-klotho during human pregnancy and its expression in pregnancies complicated by small-for-gestational-age neonates and/or preeclampsia. *The Journal of Maternal-Fetal & Neonatal Medicine* **27,** 449–457 (2013).
41. Anand-Ivell, R. & Ivell, R. Regulation of the reproductive cycle and early pregnancy by relaxin family peptides. *Mol. Cell. Endocrinol.* **382,** 472–479 (2014).
42. Weiss, G., Goldsmith, L. T., Sachdev, R., Hagen, Von, S. & Lederer, K. Elevated first-trimester serum relaxin concentrations in pregnant women following ovarian stimulation predict prematurity risk and preterm delivery. *Obstetrics & Gynecology* **82,** 821–828 (1993).
43. Trifonova, E. A. *et al.* Analysis of the placental tissue transcriptome of normal and preeclampsia complicated pregnancies. *Acta Naturae* **6,** 71–83 (2014).
44. Luo, X. *et al.* Epigenetic regulation of lncRNA connects ubiquitin-proteasome system with infection-inflammation in preterm births and preterm premature rupture of membranes. *BMC pregnancy and childbirth* **15,** 35 (2015).
45. Steffen, K. M. *et al.* Maternal and fetal variation in genes of cholesterol metabolism is associated with preterm delivery. *J Perinatol* **27,** 672–680 (2007).
46. Lappas, M. Forkhead box O1 (FOXO1) in pregnant human myometrial cells: a role as a pro-inflammatory mediator in human parturition. *Journal of Reproductive Immunology* **99,** 24–32 (2013).
47. Labied, S. *et al.* Progestins regulate the expression and activity of the forkhead transcription factor FOXO1 in differentiating human endometrium. *Mol. Endocrinol.* **20,** 35–44 (2006).
48. Bollopragada, S. *et al.* Term labor is associated with a core inflammatory response in human fetal membranes, myometrium, and cervix. *American journal of obstetrics and gynecology* **200,** 104.e1–104.e11 (2009).
49. Yang, Z., Zhu, X., Guo, C. & Sun, K. Stimulation of 11β-HSD1 expression by IL-1β via a C/EBP binding site in human fetal lung fibroblasts. *Endocrine* **36,** 404–411 (2009).
50. Staun-Ram, E., Goldman, S. & Shalev, E. Ets-2 and p53 mediate cAMP-induced MMP-2

expression, activity and trophoblast invasion. *Reprod. Biol. Endocrinol.* **7,** 135 (2009).
51. O'Brien, M., Flynn, D., Mullins, B., Morrison, J. J. & Smith, T. J. Expression of RHOGTPase regulators in human myometrium. *Reprod. Biol. Endocrinol.* **6,** 1 (2008).
52. Ouseph, M. M. *et al.* Atypical E2F repressors and activators coordinate placental development. *Dev. Cell* **22,** 849–862 (2012).
53. Morimoto, T. *et al.* Effect of labor and prostaglandins on phenylethanolamine N-methyltransferase in human fetal membranes. *Endocr. J.* **40,** 179–183 (1993).
54. Hobel, C. J., Parvez, H., Parvez, S., Lirette, M. & Papiernik, E. Enzymes for epinephrine synthesis and metabolism in the myometrium, endometrium, red blood cells, and plasma of pregnant human subjects. *The American Journal of Obstetrics & Gynecology* **141,** 1009–1018 (1981).
55. Casimiri, V. *et al.* Characterization of enzymes of catecholamine synthesis and metabolism in human fetal membranes at birth. *The American Journal of Obstetrics & Gynecology* **164,** 599–603 (1991).
56. Winn, V. D. *et al.* Gene Expression Profiling of the Human Maternal-Fetal Interface Reveals Dramatic Changes between Midgestation and Term. *Endocrinology* **148,** 1059–1079 (2007).
57. Sitras, V., Fenton, C., Paulssen, R., Vårtun, Å. & Acharya, G. Differences in gene expression between first and third trimester human placenta: a microarray study. *PLoS ONE* **7,** e33294 (2012).
58. Uusküla, L. *et al.* Mid-gestational gene expression profile in placenta and link to pregnancy complications. *PLoS ONE* **7,** e49248 (2012).
59. Rogers, J. & Gibbs, R. A. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature reviews. Genetics* **15,** 347–359 (2014).
60. Peng, X. *et al.* Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). *Nucleic Acids Research* **43,** D737–42 (2015).
61. Perry, G. H. *et al.* Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome research* **22,** 602–610 (2012).
62. Khan, Z. *et al.* Primate Transcript and Protein Expression Levels Evolve Under Compensatory Selection Pressures. *Science* **342,** 1100–1104 (2013).
63. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome research* **20,** 180–189 (2010).
64. Ackerman, W. E. *et al.* Comprehensive RNA profiling of villous trophoblast and decidua basalis in pregnancies complicated by preterm birth following intra-amniotic infection. *Placenta* **44,** 23–33 (2016).
65. Kallapur, S. G. *et al.* Intra-amniotic IL-1β induces fetal inflammation in rhesus monkeys and alters the regulatory T cell/IL-17 balance. *The Journal of Immunology* **191,** 1102–1109 (2013).
66. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 1–7 (2014). doi:10.1093/bioinformatics/btu170
67. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of

insertions, deletions and gene fusions. *Genome biology* **14,** R36 (2013).
68. Zimin, A. V. *et al.* A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol. Direct* **9,** 20 (2014).
69. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 550 (2014).
70. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21,** 3448–3449 (2005).
71. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).
72. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 6062–6067 (2004).
73. Kim, M. *et al.* GEneSTATlON 1.0: a synthetic resource of diverse evolutionary and functional genomic data for studying the evolution of pregnancy-associated tissues and phenotypes. *Nucleic Acids Research* **44,** D908–16 (2016).

CHAPTER IV

THE TRANSFORMATIVE POTENTIAL OF AN INTEGRATIVE APPROACH TO PREGNANCY[3]

Haley R. Eidem[1,3], Kriston L. McGary[1,3], John A. Capra[1,2], Patrick Abbot[1], and Antonis Rokas[1]

[1] *Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA*

[2] *Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37235, USA*

[3] *Co-first authors*

[3] This chapter was published in Placenta. 2017 Sep;57:204-215. doi: 10.1016/j.placenta.2017.07.010.

Pregnancy is an ensemble of complex traits subject to substantial genetic and environmental variation

A highly-interconnected network of physiological, cellular, and molecular pathways supports the development of a healthy fetus by maintaining homeostasis through pregnancy despite variation in maternal diet, stress, medical care, and other factors. When genetic and/or environmental variation of this network cannot be buffered to maintain a healthy state, complications arise. Like diseases of other complex traits, the most common complications of pregnancy – preeclampsia (PE), spontaneous preterm birth (sPTB), preterm premature rupture of membranes (PPROM), intrauterine growth restriction (IUGR), and spontaneous recurrent pregnancy loss (RPL) – involve multiple genetic loci and environmental factors[1-5].

Understanding the genetic basis of such complex traits is challenging. For example, although many pregnancy-related traits and pathologies, such as birth timing[1,6], birth weight[6,7], and propensity to develop PE[8], have substantial heritabilities, they are likely governed by numerous genetic variants with small effect sizes and that epistatically interact with each other[2]. Furthermore, pregnancy-related traits are also influenced, to varying degrees, by multiple environmental factors. For example, gestational diabetes[9,10], PE[10,11], and sPTB[9,10,12] are well known for their association with maternal obesity, and sPTB may also be associated with certain environmental exposures, such as bisphenol A[13]. Similarly, chronic and acute stress is thought to reduce birth weight and alter methylation levels of genes involved in the

hypothalamic–pituitary–adrenocortical (HPA) axis in the placenta, cord blood, and maternal blood[14].

Further complicating matters is that the genetic and environmental factors underlying pregnancy-associated traits and diseases do not act independently; rather, they exhibit gene by environment (GxE) effects, where the pathological phenotype is only observed with specific combinations of genetic variants and environmental conditions. For example, PPROM is often associated with inflammation due to bacterial infection, but recent studies argue that the fetal genotype also influences susceptibility[15]. Specifically, human fetuses with a null *SIGLEC14* genotype were more likely to be born prematurely, but only in conjunction with Group B *Streptococcus* (GBS) infection; in its absence, the *SIGLEC14* null variant did not appear to influence prematurity.

But what makes pregnancy-associated traits and diseases extremely, or maybe even singularly, complex is that they involve three additional dimensions. The first dimension of complexity is associated with the fact that pregnancy-associated traits require coordination and communication across many different tissues and organs in two individuals: the mother and the fetus. The interplay of tissues and organs from two individuals creates many of the distinctive complexities of pregnancy including immunosuppression, entwined physiology (respiration and metabolism), and shared endocrinology. For example, the production of progesterone, which maintains gestation in most placental mammals, must successfully shift

from the ovary to the placenta[16]. Similarly, primary human trophoblasts have been shown to release exosomes containing microRNAs, proteins, and phospholipids with antiviral properties, facilitating communication between maternal and fetal tissues[17,18].

The second dimension of complexity in pregnancy is that it involves multiple genomes (maternal, paternal, and fetal), which gives rise to the potential for conflicts of interest over parental investment[19,20]. Parent-of-origin effects on gene expression or genomic imprinting, for example, may have evolved as a result of differences in the consequences of resource investment for paternally and maternally-derived alleles[21,22]. When females mate more than once and offspring are half-sibs, paternally-derived alleles in the fetus may be evolutionarily favored to sequester more resources than optimal from the mother's perspective, favoring imprinting of the maternal allele[23,24].

The third and final dimension of complexity is that of rapid evolutionary change. Pregnancy and its associated tissues evolve rapidly in mammals[25-28], and the placenta is arguably the most diverse mammalian organ[29]. Out of this history of rapid evolutionary change emerged human pregnancy, which is distinctive in its own right[28], as a consequence of several evolutionary events and processes spanning the course of mammalian evolution, including the existence of genetic conflict[30], the primate-specific expansion of cranial size[31], and the human-specific evolution of bipedalism[32].

The importance of considering these additional dimensions of complexity is apparent in hypotheses proposed to explain puzzling facts of human pregnancy. Birth in humans appears to occur sooner than would be expected, given development of the neonate. A much larger fraction of brain growth occurs postnatally in humans than in any other primate. Why? The "obstetrical dilemma" (OD) hypothesis aims to explain gestation length based on two observations unique to human pregnancy – labor that poses risks to both mother and fetus, as well as birth at a point when fetal brain size is only 30% of adult size. The OD hypothesis holds that bipedal locomotion and large cranial capacity, both of which evolved in recent human history, act in opposing ways on the human pelvis, with the result being selection for shortened gestation lengths that preclude cranial expansion beyond pelvic capacity[23,24,33-38]. An alternative hypothesis, known as the "energetic and metabolic constraints on fetal growth and gestation" (EGG) hypothesis, aims to explain gestation length by invoking physiological limitations to metabolic provisioning *in utero*. Here, the primary controlling factor is physiological limits to the transfer of energy and metabolites between the mother and an encephalized fetus[39]. At some point, it is simply more efficient to transfer resources outside of the womb than within.

As these scenarios make clear, neither the genetic basis of pregnancy-associated traits and pathologies nor the proximate or ultimate hypotheses that explain them can be adequately understood from a single experimental approach, data source, or perspective. The history of efforts to decipher the genetic basis of a wide variety of complex traits and diseases offers

numerous examples of the perils associated with reliance on a single experimental vantage point[40-42], and the complexities of pregnancy only amplify the need for integrative approaches that combine multiple data types, approaches, or model systems[25,43,44]. In this review, we examine some of the complexities that make human pregnancy-associated traits and pathologies unique and synthesize recent progress in integrative efforts to understand their genetic and environmental dimensions.

Integrating multiple maternal and fetal tissues

Pregnancy is singular among human processes in involving coordination of many different tissues and organs from two individuals (Figure 4.1), including maternal pregnancy-specific (e.g., decidua, myometrium, and cervix), maternal non-specific (e.g., immune system, metabolism, and endocrine system), and fetal (e.g., lungs, adrenal glands, and fetal membranes). These interactions are responsible for many of pregnancy's unique physiological features, such as immune system modulation[45], entwined respiration and metabolism[46], and shared endocrinology (e.g., progesterone production). True for most placental mammals, the classic example of coordination between maternal and fetal tissues is ensuring that progesterone – a key hormone required for maintenance of gestation – is continuously produced during pregnancy even though the underlying tissue responsible for its production shifts from the ovary to the placenta[16].

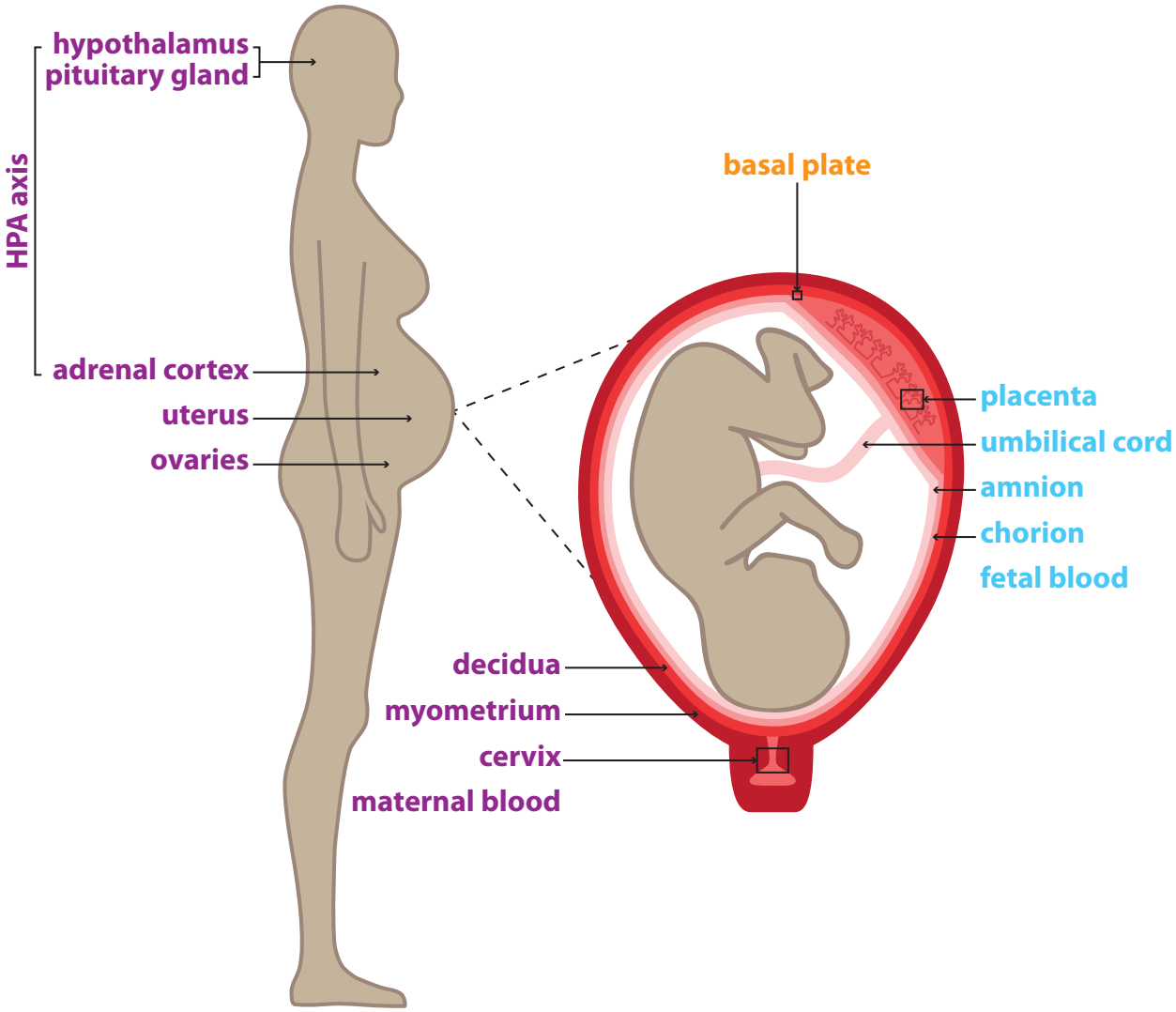

Figure 4.1. Pregnancy uniquely requires communication and coordination across multiple tissues in two individuals. Multiple maternal tissues (purple) and fetal tissues (blue) as well as tissues comprised of both maternal and fetal cells (orange) must interact to facilitate a healthy pregnancy. The placenta serves as the nexus of communication that links multiple tissues in the mother and fetus both locally and at a distance. For example, interactions between NK cells in the decidua and fetal trophoblast cells in the placenta shape the degree of placental invasiveness and rate of the exchange of nutrients and oxygen. Similarly, the hypothalamic–pituitary–adrenocortical (HPA) axis communicates maternal and fetal stress levels across multiple tissues through cortisol shared through blood exchange in the placenta.

The placenta is the nexus of this network of communicating tissues and dedicates a large fraction of its energy budget to secretion and coordination of maternal and fetal needs[47], while also providing the functions of the kidney, the lungs, and the liver for the fetus. Much of this communication occurs locally at the maternal-fetal interface and appears to have evolved early during the evolution of placental mammals by regulatory rewiring of the cAMP signaling pathway in endometrial stromal cells to facilitate decidualization and implantation[48-50]. There are many examples of the importance of communication at the maternal-fetal interface. For example, interaction between maternal immune cells and trophoblasts modulates macrophage inflammatory responses in human pregnancy, which in turn may play a role in implantation and proliferation[51]. Later in gestation, remodeling of the spiral arteries requires successful communication between maternal endothelium and migrating interstitial trophoblasts, which involves both chemotaxis and shifts in cytokine production[52]. Surprising recent work has demonstrated that placental secretion of microRNAs in exosomes at this interface appears to directly increase the resistance of maternal and fetal cell types against viruses implicated in perinatal infections by boosting autophagy[18,53,54].

In addition to local effects at the apposition of maternal-fetal tissues, inter-tissue communication also affects non-adjacent maternal and fetal tissues during gestation. For example, placental production of the neurotransmitter serotonin is crucial for embryonic brain development[55,56]. The placenta also influences maternal physiology by secreting high levels of corticotrophin releasing hormone (*CRH*), which leads to higher levels of maternal

cortisol, and may be involved in post-partum depression through desensitization to endogenous maternal *CRH*[57]. Medical interventions making use of these communication pathways can potentially be exploited for restoring proper signaling and growth in pathological conditions. Case in point, one recent study has identified synthetic peptides that target liposomes to the maternofetal interface, which restored proper signaling and appropriate placental growth by delivery of *IGF2* in a fetal growth restriction mouse model[58]. This important study provides a proof of principle that, in a mouse model, it is possible to take advance of selective binding at the maternofetal interface for the targeted delivery of therapeutics that improve pregnancy outcome.

Pregnancy thus involves complex crosstalk between multiple interdependent tissues. A comprehensive understanding of adverse pregnancy outcomes is unlikely to emerge from an approach that atomizes pregnancy-associated processes into discrete units of function. More promising are approaches that account for the numerous organs and tissues involved plus all of their interactions in both healthy and pathological states. However, the challenge is that in most cases, sampling of all relevant tissues has not been possible. For example, a recent meta-analysis of 93 global transcriptomic studies across 9 gestational tissues and 29 clinical subtypes showed the paucity of studies that capture multiple tissues with the same clinical phenotype (e.g., many data sets for placental gene expression during PE, but few samples from other tissues)[59]. Even if focus is restricted to examination of single tissues, an additional challenge arises in that independently obtained samples from the same tissue type and

clinical subtype show substantial heterogeneity in their gene expression profiles, with only minimal replication of significant genes across studies. This heterogeneity has both technical and biological explanations. Some of the heterogeneity stems from differences in tissue sampling protocols and from differences in the interaction with tissues that were not collected concurrently. Additionally, recent studies have shown that placental gene expression patterns cluster into distinct molecular categories that correlate with maternal symptomology[60,61]. This clustering has also shown improved homogeneity in placental gene expression and illustrates the need for improved molecular phenotyping and experimental design that takes these clusters into account. Nevertheless, much of the observed diversity likely reflects genuine heterogeneity in gene expression within and among individuals, which is evident in global gene expression profiles of placental tissue obtained from healthy pregnancies[62] as well as in comparisons of transcriptional profiles of placental cell types[63].

The extent and magnitude of the observed interactions necessitates the simultaneous sample collection and analysis of multiple tissues, rendering data collection and integration across tissues one of the pressing challenges for understanding pregnancy. An alternative, complementary approach to the simultaneous examination of multiple tissues is *in vitro* reconstruction of tissue interactions, which may become possible as placenta-on-a-chip technology matures[64,65]. A recent study by Blundell et al.[65] demonstrated a model of the human placental barrier using trophoblasts and endothelial cells that accurately reproduces the formation of microvilli and the syncytialization of trophoblasts, and matches the glucose

transfer rate of perfused *ex vivo* human placentas. Furthermore, cell lines of endometrial stromal cells[66], decidual cells[50], and myometrial cells[67] are available, which may lead to more comprehensive pregnancy-on-a-chip approaches. New data from single-cell transcriptomes will also help to detect the effects of local tissue-tissue interactions[63]. Such advances in *in vitro* technology will greatly aid the capture of multiple different data types (e.g., gene expression, protein abundance, phosphorylation, etc.) from interacting tissues simultaneously and in an integrated fashion, which can be used to generate models of tissue interactions. Predictions from such models could in turn be validated through *in vivo* collection of the same data types from the same tissues.

Integrating maternal and fetal genomes

The involvement of both the maternal and the fetal genome in pregnancy complicates its study in three related ways. First, a maternal allele can impact pregnancy through either phenotypic expression in the fetus if transmitted or in the mother as either the untransmitted or transmitted maternal allele (Figure 4.2). Second, as described above, the phenotypic impact of a fetal allele can vary depending on whether it is maternally or paternally derived, a phenomenon known as parent-of-origin effects associated with genomically imprinted genes[68]. Third, and as a consequence of these two, natural selection may act on alleles differently depending on which genome they find themselves in (maternal or fetal) or which they stem from (maternal or paternal). Therefore, alleles may increase the frequency of their transmission to the next generation by different strategies – improving maternal, fetal, or

paternal fitness – which can give rise to genetic conflict since reproductive and/or survival success is different for each organism involved[21].

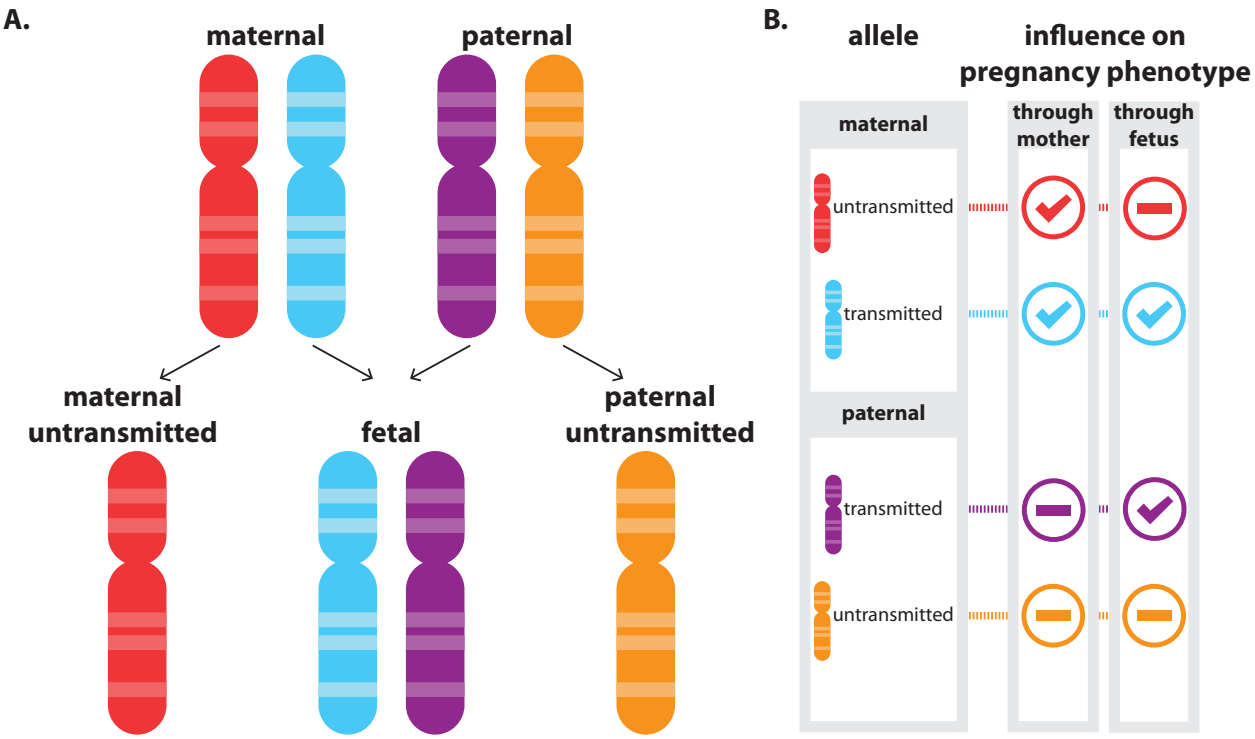

Figure 4.2. Involvement of three genomes complicates genetic analyses aimed at linking genes with pregnancy-associated phenotypes. Unlike genetic analyses of most other traits, studies of pregnancy typically require genotyping both mother and fetus, since variants can influence pregnancy through impacting either the maternal or the fetal phenotype or both. Specifically, non-transmitted maternal alleles (red) affect pregnancy only through impact on maternal phenotype; transmitted maternal alleles (blue) can potentially affect pregnancy through impact on maternal and/or fetal phenotype; transmitted paternal alleles (purple) only impact pregnancy through fetal phenotype; finally, non-transmitted paternal alleles (orange) do not affect pregnancy directly. Genetic conflict arises when natural selection differentially favors alleles when they are maternally or paternally expressed, which may contribute to pathologies of pregnancy, including preeclampsia.

The complexities inherent in analyses of parental and fetal genetics can be illustrated by considering the well-established correlation of maternal height with gestational age at birth

and fetal size[69-74]. Is this correlation causal? If yes, is it driven by phenotypic expression of alleles in the mother or the fetus? To address these questions, Zhang and colleagues[75] used a novel genetic analysis method to disentangle the causal influence of maternal phenotype and genetics from fetal genetics on gestation outcomes. They found that birth length and birth weight were significantly associated with maternal transmitted haplotype whereas gestational age was significantly associated with the maternal non-transmitted haplotype. On the basis of these results, Zhang and colleagues inferred that fetal genetics drives the association between maternal height and fetal growth, but that maternal height, rather than fetal genetics, drives gestational age at birth. This implies that physical and anatomical constraints may contribute to birth timing, as suggested by the OD hypothesis.

Parent-of-origin effects are typically associated with genomic imprinting, where expression of a gene in the fetus or extraembryonic tissue, including the placenta, is only from either the maternally or the paternally inherited allele. Thus, depending on the parent-of-origin, specific variants will be expressed and their effects will be accentuated, whereas other variants won't and their effects will be obscured. Although imprinting was thought to affect most tissues relatively consistently, a recent study revealed a new class of imprinted genes that are specific to the placenta and exhibit paternal allelic expression[76]. These placenta-specific imprinted genes suggest a novel mechanism of imprinting, as these loci are not methylated, the typical mechanism of imprinting, in either sperm or embryonic stem cells[76].

Genomic imprinting and gene expression only from either the maternally or the paternally inherited allele likely arose due to a tug-of-war in fetal resource allocation between the genomes[21,22]; specifically, whereas the selectively favored *fetal* genetic variants will be those that optimize resource allocation to the fetus, the selectively favored *maternal* genetic variants will be those that optimize resource allocation across both the current pregnancy but also future ones[23,24]. Where these two optima are in conflict, the kinship theory predicts that maternally-inherited allelic expression (paternal imprinting) will tend to favor slower fetal growth whereas paternally-inherited allelic expression (maternal imprinting) will tend to favor increased fetal growth[22]. Experimental evidence from functional studies of several imprinted genes provide empirical support for these predictions. For example, in mice, the imprinted *Ipl* gene is maternally expressed in both the placenta and the yolk sac, and functions to restrain placental growth and size[77]. In contrast, the imprinted *Igf2* gene is paternally expressed in the mouse placenta, and functions to increase the supply of maternal nutrients to the fetus, augmenting fetal growth[78].

Genetic conflict between maternal, fetal, and paternal genomes has many possible phenotypic outcomes, but two of the most clinically significant, low fetal birth weight and PE, are closely related and mediated by interactions between the maternal immune system and the placental trophoblasts. It might be expected that fetal birth weight is optimized by natural selection , as a healthy weight is associated with increased perinatal survival[79] and reduces the risk of a wide range of adverse adult outcomes (e.g., cognitive development and

function, chronic health conditions, disability)[80,81]. Unrestricted fetal growth is undesirable since very large babies also have increased morbidity and mortality. Thus, stabilizing selection acts on birth weight to maintain this balance[82]. However, the maternal genome is favored to optimize resource allocation across multiple offspring[21]. One illustration of this parent-offspring conflict and its influence in fetal birth weight regulation comes from the well-studied interaction between two loci that contain highly polymorphic gene clusters: the *KIR* locus, expressed in maternal natural killer cells, and the *HLA* locus, expressed in implanting placental trophoblasts. A recent study has shown that when a mother has a *KIR* haplotype containing the *KIR2DS1* receptor gene, fetal birth weight increases when the fetus paternally inherits the *HLA-C2* ligand genotype but not when the same allele is maternally inherited[83]. This combination of a maternal genotype and a paternally-inherited fetal genotype has a substantial effect on birth weight, increasing it by almost 10% (~250 grams), which is strongly suggestive of *HLA-C2* favoring fetal growth and fitness when paternally-inherited.

An outcome of substantial mismatch in the relative influence of maternal and fetal variants is increased morbidity and/or mortality for both the mother and fetus[84]. Evolutionary biologists have long argued that some pathologies of pregnancy stem precisely from situations in which maternal and fetal reproductive success are not aligned[21,85]. For example, risk for PE is thought to start when the placenta is insufficiently perfused due to inadequate invasion of the decidua by trophoblasts, which leads to increased fetal signals of distress and higher blood pressure[86]. At the genetic level, mismatch occurs when a fetus inherits the paternal

HLA-C2 genotype but the mother has a *KIR AA* genotype since the interaction of these two loci in this specific combination has been shown to increase risk for PE[87]. Interestingly, protective *KIR B* variants have evolved independently in Sub-Saharan African and European populations and are found at different locations within the *KIR* gene cluster[88].

Processes that mediate genetic conflict, including immune interactions and imprinting, introduce unique selective pressures on the maternal, fetal, and paternal genomes. Both the OD and the EGG hypothesis point to the possibility of ongoing genetic conflict between maternal and fetal alleles. Although the OD and EGG hypotheses have been presented as alternative explanations for human fetal development and birth timing, the two views may be complementary and reinforcing. Under both hypotheses, the peak growth of an energy-demanding fetal brain coincides with birth and imposes a substantial maternal cost in energy and risk, which may be compensated for by boosting long-term cognitive development and, presumably, fetal fitness[81,89]. Long-term stability of gestation length may be the result of a compromise between individually optimal maternal, paternal, and fetal fitness. However, the conflict between genomes also plays out on shorter time-scales as illustrated by differences in gestation length and fetal maturity among African, Asian, and European populations[90]. Understanding either timescale requires the integration of maternal and fetal phenotype with maternal and fetal genetic variants and an understanding of evolutionary dynamics.

Genetic conflict in pregnancy appears to lead to faster evolutionary rates across mammals and may lead to the divergence of many pregnancy processes between humans and common model organisms. A recent integrative study found that genes enriched for placental expression evolve faster than genes enriched for expression in most other tissues[25]. Why is this so? Conflict between maternal and fetal genomes has been proposed to result in the "Red Queen" effect, where both genomes must continually evolve to maintain balance[84]. Some of the conflict is thought to be mediated by genomic imprinting[91]. For example, fixation of a new variant of a paternally imprinted gene favoring fetal growth may be countered by fixation of a variant of a maternally imprinted gene restraining fetal growth (and vice versa); repeated cycles of fixations of variants influencing fetal growth would result in acceleration of evolutionary substitution rates in the genes involved, a hallmark of the "Red Queen" effect at the molecular level. Although early examination of the *Igf2* gene and its receptor *Igfr2* as well as of a small number of other imprinted genes in a few mammals did not show evidence of evolutionary rate acceleration[92], a more recent study that examined a broader set of imprinted genes inferred that imprinted genes show increased evolutionary rates relative to non-imprinted genes, consistent with the expectation that genetic conflict results in "Red Queen" effects[26]. Furthermore, the set of human imprinted genes differs substantially from the set of mouse imprinted genes, which suggests ongoing turnover of the proximal mechanisms mediating long-term genetic conflicts[76].

Integrating function and evolution of pregnancy

Multiple pregnancy-associated traits evolve rapidly (Figure 4.3). For example, the placenta is highly variable across mammals, showing tremendous diversity in structure, shape, invasiveness[27,93]. Along the human lineage, some pregnancy-associated traits may have evolved in conjunction with the rapid evolution of greater cranial capacity and bipedalism, resulting in faster evolutionary rates than would be expected based on genetic conflict alone. Interestingly, one recent study argued that humans may have evolved genetic architecture that allows pelvic size and cranial capacity to co-vary[94]. Women with large heads give birth to babies with larger than average heads, which would seemingly lead to extreme cases of the obstetric dilemma and mortality; however, pelvic size co-varies with cranial size, greatly reducing the risk. Strong selection for this co-variation is presumably less likely in other primates that do not face the OD, although we are unaware of experiments that explicitly test this hypothesis.

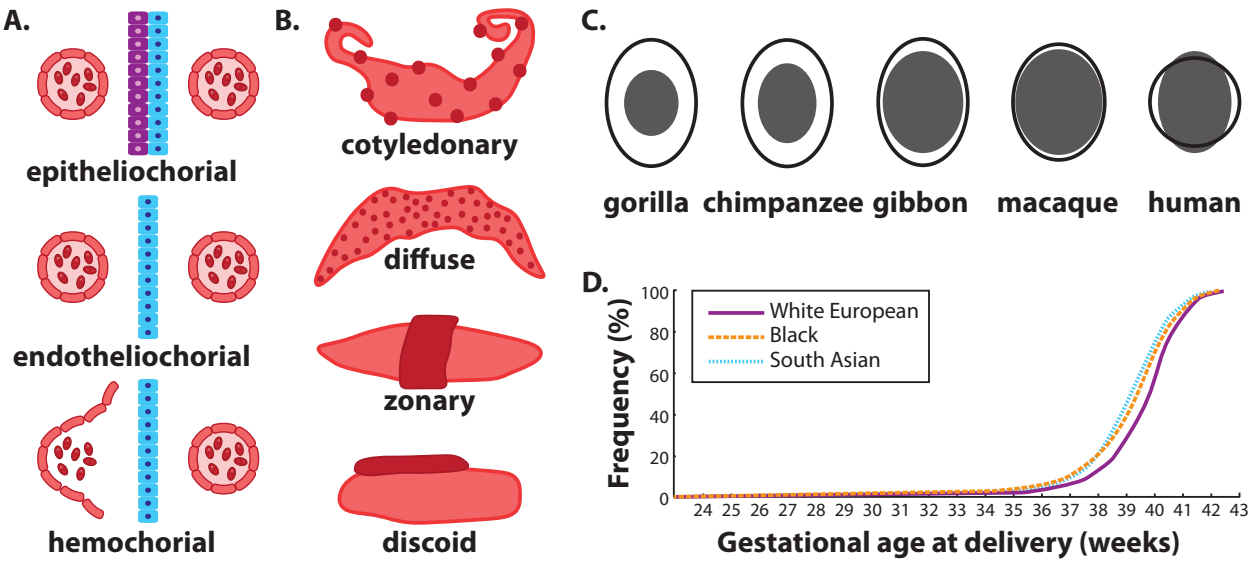

Figure 4.3. Pregnancy evolves rapidly across mammals, primates, and human populations. A) Variation in mammalian placentation and fetal access to the maternal bloodstream. This variation can be broadly categorized into three commonly observed modes: epitheliochorial (minimally invasive), endotheliochorial (intermediate), and hemochorial (very invasive). Maternal uterine epithelium (purple) is present only in epitheliochorial placentae; thus, the fetal chorion (blue) is the only epithelium in both endotheliochorial and hemochorial placentae. In hemochorial placentae, the maternal endothelium of blood vessels (pink) is eroded, leading to the formation of blood sinuses (red). As with many pregnancy traits, placentation does not follow a simple evolutionary pattern; rather, varying levels of invasiveness have repeatedly evolved throughout mammals. B) Likewise, the shape of the placenta (pink) and its areas of contact (red) with the underlying uterus varies widely across mammals and includes cotyledonary, diffuse, zonary, and discoid morphologies. C) Primates vary in the relative sizes of the maternal pelvis (open oval) and the fetal cranium (filled oval) at the time of birth, with both macaques and humans sharing a tight fit that appears to have evolved independently in each of the two lineages. Redrawn from Schultz 1969 and Rosenberg 2002. D) Human populations have diverged in gestation length with populations of European ancestry (purple) delayed about one week relative to populations with African (orange) or South Asian (blue) ancestry. Redrawn from Scioscia 2009.

A consequence of the fast evolutionary pace of change in mammalian pregnancy is that it is more challenging to predict whether traits observed in one species are also likely to be observed in another[27,95]. For example, it has been commonly assumed that humans are unique in giving birth to "babies born alive before 37 weeks of pregnancy are completed", the

definition of PTB given by the World Health Organization[96]. Although this definition is obviously specific to our species, generalizing the human-based definition of PTB as 'parturition prior to 92.5% (37 / 40 weeks) completed gestation' and applying to mammals in general show that substantial variation in the length of gestation is widespread across mammals and that human gestation length is very similar to what would be expected of a mammal of our body mass[97]. Nevertheless, the dramatic long-term negative effects of preterm birth have not been reported extensively in other species. The rate of human fetal brain expansion peaks at the time of normal parturition and infants are born in a secondarily altricial state, unable to care for themselves. Other primates experience peak rates of brain development much earlier and are born precociously[98-100], which Phillips et al. suggest reduces the fitness consequences for earlier birth[97]. Thus, humans may be unique relative to other mammals due primarily to the timing and importance of brain development, but not with respect to variation in gestation length. More generally, by disentangling of studies of genetic contributors to gestation timing from studies of the medical consequences of premature birth across mammals, the possibility emerges that some mammal models may be useful for the study of gestation timing, while others may be useful for the study of specific PTB-associated pathologies. For example, it could be argued that bats, which have gestation lengths much longer than other mammals of similar size, may be great models for understanding the genetics of gestation timing, even though they may be poor models for studying the pathology of PTB, as prematurely born bats are not known to exhibit any of the pathologies associated with human preterm birth[97,101,102].

The rapid evolution of pregnancy makes the integration of evolutionary analyses and functional data critical to translating discoveries from one species to another, e.g., from the mouse to humans[95]. The first step toward this goal is understanding how the biological systems of pregnancy have diverged between two species. One illustrative example is a recent study that integrated diverse types of data, including gene expression from multiple species, multiple genomic signatures (e.g., CpG island density, sequence conservation, and nucleotide substitution rates), histone modification, copy number variation, TF binding motifs, and ChIP-seq, in order to understand how placental development and gene expression evolve quickly[103]. This integrated analysis revealed that cell lineage-specific enhancers from endogenous retroviruses from the RLTR13D5 family led to substantial divergence of gene regulation in trophoblast stem cells between mouse and rats during 25 million years of separation, a time interval substantially smaller than the 80 million year one demarcating the divergence of rodents and primates[103]. Interestingly, other integrative 'omics analyses have pointed to the importance of transposable elements (TEs) in the evolution of pregnancy[104,105]. For example, a cross-species integration of myometrial RNA-Seq data showed that ancient TEs have been coopted into hormone-responsive regulatory elements coordinating uterine gene expression[104]. Thus, DNA sequences derived from ancient repetitive elements such as endogenous retroviruses and TEs appear to have been repeated coopted into the regulatory landscape of several different tissues associated with mammalian pregnancy.

The rapid evolution of gestation may make clinical translation of animal models more difficult, but it can also be exploited to identify candidate genes based on those evolutionary differences[25,106-109]. For example, a recent study identified over 1,000 mammalian genes that repeatedly experienced selection (i.e., accelerated protein sequence evolution) during the evolution of reduced placental invasiveness in other mammals[107]. These genes significantly overlap with genes known to be involved in disorders of pregnancy, particularly PE, and include several associated via GWAS analyses (e.g., *F5*, a coagulation factor; *IL6*, an inflammatory cytokine; and *APOE*, an apolipoprotein) as well as several that are differentially expressed in PE relative to normal gestation (e.g., *S100A8*, a calcium-binding protein involved in inflammatory and immune responses; *CD97*, an adhesion G protein-coupled receptor; and *FLT1*, a vascular endothelial growth factor receptor). By further integrating evolutionary analyses that predict phenotypic impact of mutations and known genetic and physical interactions, this unique study also narrowed down the original list of candidate genes likely involved in the three independent evolutionary transitions towards reduced placental invasion to 199 genes, all of which are novel candidates for involvement in PE.

Some aspects of pregnancy, such as birth timing[90], evolve so quickly that they have even diverged between modern human populations, raising the possibility that searching for genes exhibiting strong population differentiation may be a fruitful approach to identify gene candidates[108,110]. For example, the leptin gene (*LEP*), which has been previously associated with PE[111-113], shows strong differentiation between human populations[110]. The same is true

for the reductase *DHCR7*, which has been associated with both melanoma and preterm birth[108,114].

Accelerating integration to facilitate a dynamic understanding of pregnancy

Typically, integrative studies use either multiple data types in a single tissue or the same data type across multiple tissues[42]. In an excellent recent example illustrating the potential power of the first approach, Chu and colleagues integrated high-throughput microRNA, mRNA, and protein expression data to infer an integrated regulatory network that responds to oxidative stress in human placental trophoblast cells[115]. Similarly, another recent study integrated previous GWAS data on PTB with genes differentially expressed between term and preterm samples of myometrial tissue to identify a significant association between parts of two transcription factor networks and PTB[116]. Examples of the second approach are meta-analyses that synthesize knowledge from a single data type across multiple studies investigating different tissues to identify commonalities as well as gaps in current knowledge. One recent comprehensive meta-analysis[59] of genome-scale gene expression studies found a substantial mismatch between currently available data sets related to various pathologies of pregnancy and their worldwide incidence, identifying tissues and pathology combinations (e.g., placenta and sPTB) that are relatively understudied. Specifically, although only 30% of PTB cases are medically indicated, 76% of global gene expression research focuses on these cases. In contrast, 45% of PTB cases occur spontaneously and only 18% of global gene expression research focuses on this major contributor to PTB incidence. Finally, this meta-analysis also

revealed that replication of results across studies of the same tissue and pathology is very low, with only 23/10,993 unique genetic elements replicated 10 or more times across 134 studies, reflecting both the high heterogeneity of global gene expression profiles across cell types[63], individuals[62], and populations[62], as well as variability in the design and implementation of pregnancy-related gene expression analyses in general.

As the pace of research accelerates, systematic data integration is moving beyond single publications. For example, Uzun and colleagues have developed dbPTB, which systematically collects genes reported to be associated with preterm birth and is regularly updated, creating a dynamic synthesis of genetic associations[117]. Comprehensive resources like these can provide the foundation for additional analyses that incorporate other kinds of data. For example, integration of preterm birth-related genes reported in dbPTB with data from a previous GWAS study that failed to identify any genome-wide significant candidates allowed for the identification of several pathways (e.g., regulation of blood pressure, smooth muscle contraction, and general metabolism) potentially involved in PTB but missed by the GWAS study alone[118].

Although necessary for advancing our understanding of pregnancy and its pathologies, integration of data is a challenge for many reasons. First, data sets relevant to pregnancy can be difficult to obtain. Existing curation of pregnancy literature has focused more on reporting individual candidate genes (e.g., dbPTB) rather than genome-scale data sets. A meta-analysis

of transcriptomes[59] has shown that many papers report genome scale data sets that are not readily available through standard repositories like the Gene Expression Omnibus (GEO), a behavior that is objectionable by present-day standards of practice[119]. Second, re-analysis and quality control of genome scale data sets often require substantial domain-specific knowledge as well as considerable computational expertise. For example, even though RNA-seq data sets can be distilled down to per-gene expression levels (e.g. read counts), a simplification and compression of orders of magnitude, original data sets can involve multiple samples may consist of terabytes of raw data that require substantial computational time. Similarly, integration of multiple types of data requires expertise that may not be readily available in many research labs. Finally, once the raw data has been distilled, integrating data effectively using appropriate statistics, algorithms, and data visualizations requires additional expertise and perspective (Figure 4.4).

Figure 4.4. The multi-faceted complexities of pregnancy require integration of multiple types and sources of data. Pregnancy involves two individuals with two distinct genomes (maternal: light red/light green, fetal: dark red/dark green), multiple tissues, and rapid evolution. Each layer of complexity corresponds to a different type of data and successful integration requires careful attention not only to phenotype and tissue, but also to a variety of molecular and evolutionary modulators. Ideally, several different types of 'omics data from the same mothers and babies can be collected across tissues and intersected with relevant evolutionary and functional data to better illuminate the presumably complex pathways leading to pregnancy pathologies.

Open discussion at a recent NIH meeting for the Human Placenta Project[120] highlighted the importance of making data publicly available and providing the bioinformatics tools needed to analyze it. Fortunately, research supporting the integration of pregnancy data and analyses is accelerating. For example, ImmPort (immport.org) is a new platform that makes very large raw datasets easily available for analysis and integration[121]. Adoption of ImmPort by the pregnancy research community will ensure rapid access to the most recent datasets, including ones that are otherwise impractical for individual labs to host; a recent data release (March 2016) included a very large case-control microbiome study of preterm birth[122], which has over 4,000 individual metagenomics samples. In addition, following a recent public request for feedback on how to integrate placental images from various technologies and molecular data with physiology and anatomy[123], the NIH is in the process of developing a comprehensive electronic placental atlas tool.

Other databases with more general biological data applicable to the study of pregnancy and its pathologies include Protein Atlas[124], GTeX[125], OMIM[126], and similar resources with information about 'omics in the context of general expression patterns or disease associations. Of note, however, is the limited tissues, phenotypes, and species annotated in each of these resources. For example, although Protein Atlas contains dense information on placental FPKM and genes exhibiting placental-specific expression patterns, these data stem only from healthy, term placentas. Similarly, GTeX contains gene expression information for reproductive tissues like uterus, ovary, and cervix, but lacks data for placenta. Therefore, all of

these data are important when considering an integrative approach to pregnancy and its pathologies, but must be utilized as supplementary to other, more targeted 'omics data.

One example of a database designed to facilitate integrative pregnancy research is one developed by the authors, GEneSTATION (Figure 4.5A; genestation.org), which provides access to a wide variety of pre-processed data for on-the-fly exploration across multiple data types, species, and human populations[127]. Datasets include gene and protein expression for multiple tissues and pregnancy phenotypes, gene ages and evolutionary rates, allele frequencies across populations, as well as functional annotations and can be analyzed with novel tools like SynTHy (Synthesis and Testing of Hypotheses, Figure 4.5B), a framework for the aggregation and cross-filtering of candidate genes according to user-determined filters applied to GEneSTATION data, or integRATE, a desirability function-based data integration software that prioritizes candidate genes in terms of their weight of evidence across relevant research.

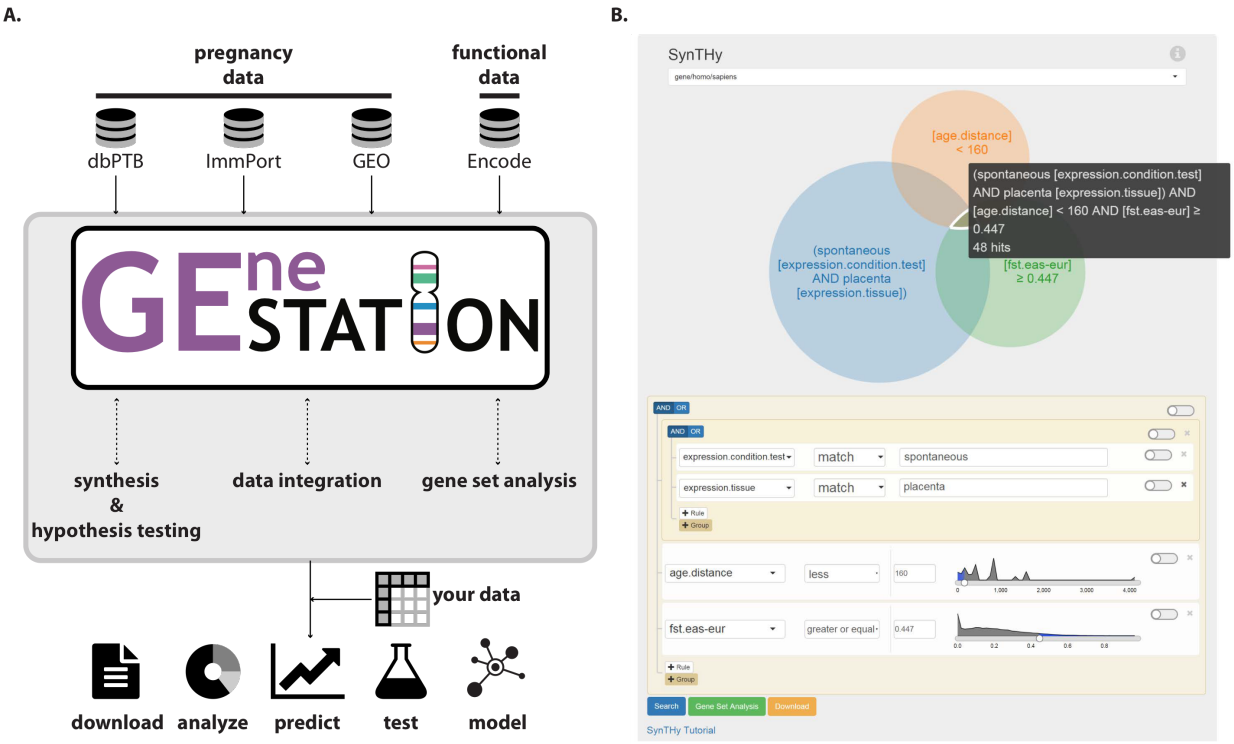

Figure 4.5. GEneSTATION is a new platform for integration of diverse *'omics* and evolutionary data types that facilitates rapid analyses. A. GEneSTATION integrates pregnancy-specific data from a number of public sources with functional data that is not specific to disease. The platform provides users with rapid, simple tools for generating and testing hypotheses, integrating data, and analyzing sets of candidate genes using both existing data and user submitted data. Results from all analyses can be easily downloaded and used to prioritize candidate genes for functional involvement in a trait or pathology. **B.** One example visual analytical tool from GEneSTATION, SynTHy, allows integration of many data types, including differential gene expression in pregnancy tissues for various pathologies, gene ages, and genetic differences between populations (F_{ST}). In the example analysis, the user has selected genes that are differentially expressed in spontaneous preterm birth in the placenta, arose in the lineage leading to placental mammals, and associated with genetic variants that show substantially different frequencies between South Asian and European population ($F_{ST} \geq 0.447$).

As the amount and diversity of 'omics data associated with pregnancy phenotypes and pathologies continues to increase, it will be important to keep in mind that numerous computational methods exist for the integration of 'omics data including genomic variation

analysis, domain knowledge-guided analysis, concatenation-based integration, transformation-based integration, and model-based integration[128]. These tools aim to predict phenotypic traits, identify biomarkers, and illuminate genetic underpinnings of complex diseases, like PTB and other pregnancy-associated pathologies. However, the majority of these methods requires the availability of multiple 'omics data types from the same patient cohort and functionally validated genes known to be involved in a given pathology for model training; in our view, generation of such multi-omics data sets from the same set of patients should be a priority for future research.

Perspective

In this review, we have outlined the complexities associated with the genetic dissection of pregnancy traits and pathologies and presented several different integrative approaches on how can these can be overcome. Going forward, we believe that integrative approaches will yield the greatest understanding of pregnancy-associated phenotypes when they are joined with research platforms that enable pregnancy researchers from multiple fields to readily access diverse types of pregnancy-related data, quickly obtain and examine syntheses of current data, and efficiently explore the connections between their own data and the existing syntheses. Such platforms will increase the pace and efficiency of discovery by reducing effort duplication in basic data processing, quickly identifying datasets relevant to specific questions, translating knowledge from model organisms into humans and vice versa through the integration of an evolutionary perspective and analyses, and providing scientists with

intuitive tools that help them confidently and effortlessly integrate data and analyses from multiple disciplines.

Given the multi-faceted complexities of pregnancy, it is encouraging that the agencies funding pregnancy research have shifted their funding strategies to prioritize approaches that integrate new approaches with existing data and knowledge. In light of the substantial challenges involved in reanalyzing a wide variety of data (request from authors, validation/curation, and basic analysis), future integrative approaches will require data and analysis platforms that make diverse kinds of existing pregnancy data readily available and interoperable, along with algorithms to easily integrate and interpret new data.

The diverse complexities of pregnancy make integrative approaches a necessary part of all future pregnancy research. Rapid discovery and improved clinical care will be the fruit of community efforts to improve data access, to facilitate powerful multi-disciplinary analyses by non-experts, and to develop platforms that promote collaboration across disciplines.

ACKNOWLEDGEMENTS

HRE was supported by a Transdisciplinary Scholar Award from the March of Dimes Prematurity Research Center Ohio Collaborative. This research was supported by the March of Dimes through the March of Dimes Prematurity Research Center Ohio Collaborative.

REFERENCES

1. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: one syndrome, many causes. *Science* **345,** 760–765 (2014).
2. Romero, R., Kuivaniemi, H., Tromp, G. & Olson, J. The design, execution, and interpretation of genetic association studies to decipher complex diseases. *The American Journal of Obstetrics & Gynecology* **187,** 1299–1312 (2002).
3. Pennington, K. A., Schlitt, J. M., Jackson, D. L., Schulz, L. C. & Schust, D. J. Preeclampsia: multiple approaches for a multifactorial disease. *Dis Model Mech* **5,** 9–18 (2012).
4. Hendrix, N. & Berghella, V. Non-placental causes of intrauterine growth restriction. *Semin. Perinatol.* **32,** 161–165 (2008).
5. Lengyel, C., Muglia, L. J. & Pavlicev, M. *Genetics of Preterm Birth*. **283,** (John Wiley & Sons, Ltd, 2001).
6. Clausson, B., Lichtenstein, P. & Cnattingius, S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG* **107,** 375–381 (2000).
7. Magnus, P., Gjessing, H. K., Skrondal, A. & Skjaerven, R. Paternal contribution to birth weight. *J Epidemiol Community Health* **55,** 873–877 (2001).
8. Salonen Ros, H., Lichtenstein, P., Lipworth, L. & Cnattingius, S. Genetic effects on the liability of developing pre-eclampsia and gestational hypertension. *Am. J. Med. Genet.* **91,** 256–260 (2000).
9. Torloni, M. R. *et al.* Prepregnancy BMI and the risk of gestational diabetes: a systematic review of the literature with meta-analysis. *Obesity Reviews* **10,** 194–203 (2009).
10. Marchi, J., Berg, M., Dencker, A., Olander, E. K. & Begley, C. Risks associated with obesity in pregnancy, for the mother and baby: a systematic review of reviews. *Obesity Reviews* **16,** 621–638 (2015).
11. Wang, Z. *et al.* Maternal adiposity as an independent risk factor for pre-eclampsia: a meta-analysis of prospective cohort studies. *Obesity Reviews* **14,** 508–521 (2013).
12. Lengyel, C. S., Ehrlich, S., Iams, J. D., Muglia, L. J. & DeFranco, E. A. Effect of Modifiable Risk Factors on Preterm Birth: A Population Based-Cohort. *Matern Child Health J* **21,** 777–785 (2016).
13. Patel, C. J. *et al.* Investigation of maternal environmental exposures in association with self-reported preterm birth. *Reproductive Toxicology* **45,** 1–7 (2014).
14. Kertes, D. A. *et al.* Prenatal Maternal Stress Predicts Methylation of Genes Regulating the Hypothalamic–Pituitary–Adrenocortical System in Mothers and Newborns in the Democratic Republic of Congo. *Child Development* **87,** 61–72 (2016).
15. Ali, S. R. *et al.* Siglec-5 and Siglec-14 are polymorphic paired receptors that modulate neutrophil and amnion signaling responses to group B Streptococcus. *Journal of Experimental Medicine* **211,** 1231–1242 (2014).
16. Csapo, A. I., Pulkkinen, M. O., Ruttner, B., Sauvage, J. P. & Wiest, W. G. The significance of the human corpus luteum in pregnancy maintenance. I. Preliminary studies. *The American Journal of Obstetrics & Gynecology* **112,** 1061–1067 (1972).

17. Ouyang, Y., Mouillet, J. F., Coyne, C. B. & Sadovsky, Y. Review: placenta-specific microRNAs in exosomes - good things come in nano-packages. *Placenta* **35 Suppl,** S69–73 (2014).
18. Delorme-Axford, E. *et al.* Human placental trophoblasts confer viral resistance to recipient cells. *PNAS* **110,** 12048–12053 (2013).
19. Ishida, M. *et al.* Maternal Inheritance of a Promoter Variant in the Imprinted PHLDA2 Gene Significantly Increases Birth Weight. *Am. J. Hum. Genet.* **90,** 715–719 (2012).
20. Moore, G. E. *et al.* The role and interaction of imprinted genes in human fetal growth. *Phil. Trans. R. Soc. B* **370,** 20140074–20140074 (2015).
21. Haig, D. Genetic Conflicts in Human-Pregnancy. *Q Rev Biol* **68,** 495–532 (1993).
22. Moore, T. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics* **7,** 45–49 (1991).
23. Schultz, A. H. *The life of primates*. (1969).
24. Rosenberg, K. & Trevathan, W. Birth, obstetrics and human evolution. *BJOG* **109,** 1199–1206 (2002).
25. Hirbo, J., Eidem, H., Rokas, A. & Abbot, P. Integrating Diverse Types of Genomic Data to Identify Genes that Underlie Adverse Pregnancy Phenotypes. *PLoS ONE* **10,** e0144155 (2015).
26. Hutter, B., Bieg, M., Helms, V. & Paulsen, M. Divergence of imprinted genes during mammalian evolution. *BMC evolutionary biology* **10,** 116 (2010).
27. Mossman, H. W. *Vertebrate fetal membranes: comparative ontogeny and morphology; evolution; phylogenetic significance; basic function; research opportunities*. (London: Macmilian, 1987).
28. Abbot, P. & Rokas, A. Mammalian pregnancy. *Curr. Biol.* **27,** R127–R128 (2017).
29. Garratt, M., Gaillard, J.-M., Brooks, R. C. & Lemaître, J.-F. Diversification of the eutherian placenta is associated with changes in the pace of life. *PNAS* **110,** 7760–7765 (2013).
30. Spencer, H. G., Clark, A. G. & Feldman, M. W. Genetic conflicts and the evolutionary origin of genomic imprinting. *Trends Ecol. Evol. (Amst.)* **14,** 197–201 (1999).
31. Jeffery, N. Brain expansion and comparative prenatal ontogeny of the non-hominoid primate cranial base. *J. Hum. Evol.* **45,** 263–284 (2003).
32. Schmitt, D. Insights into the evolution of human bipedalism from experimental studies of humans and other primates. *Journal of Experimental Biology* **206,** 1437–1448 (2003).
33. Krogman, W. M. The Scars of Human Evolution. *Scientific American* **185,** 54–57 (1951).
34. Washburn, S. L. Tools and Human Evolution. *Scientific American* **203,** 62–75 (1960).
35. Rosenberg, K. & Trevathan, W. Bipedalism and human birth: The obstetrical dilemma revisited. *Evolutionary Anthropology: Issues, News, and Reviews* **4,** 161–168 (1995).
36. Wittman, A. B. & Wall, L. L. The Evolutionary Origins of Obstructed Labor: Bipedalism, Encephalization, and the Human Obstetric Dilemma. *Obstet Gynecol Surv* **62,** 739–748 (2007).
37. Trevathan, W. Primate pelvic anatomy and implications for birth. *Phil. Trans. R. Soc. B* **370,** 20140065–20140065 (2015).

38. Gruss, L. T. & Schmitt, D. The evolution of the human pelvis: changing adaptations to bipedalism, obstetrics and thermoregulation. *Phil. Trans. R. Soc. B* **370,** 20140063–20140063 (2015).
39. Dunsworth, H. M., Warrener, A. G., Deacon, T., Ellison, P. T. & Pontzer, H. Metabolic hypothesis for human altriciality. *PNAS* **109,** 15212–15216 (2012).
40. Mootha, V. K. *et al.* Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 605–610 (2003).
41. Akavia, U. D. *et al.* An Integrated Approach to Uncover Drivers of Cancer. *Cell* **143,** 1005–1017 (2010).
42. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome biology* **18,** 83–15 (2017).
43. Gracie, S. *et al.* An integrated systems biology approach to the study of preterm birth using '-omic' technology--a guideline for research. *BMC pregnancy and childbirth* **11,** 71 (2011).
44. Lackritz, E. M. *et al.* A solution pathway for preterm birth: accelerating a priority research agenda. *The Lancet Global Health* **1,** e328–e330 (2013).
45. Mor, G. & Cardenas, I. REVIEW ARTICLE: The Immune System in Pregnancy: A Unique Complexity. *Am J Reprod Immunol* **63,** 425–433 (2010).
46. Mele, J., Muralimanoharan, S., Maloyan, A. & Myatt, L. Impaired mitochondrial function in human placenta with increased maternal adiposity. *Am. J. Physiol. Endocrinol. Metab.* **307,** E419–E425 (2014).
47. Carter, A. M. Placental Oxygen Consumption. Part I: In Vivo Studies—A Review. *Placenta* **21,** S31–S37 (2000).
48. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature genetics* **43,** 1154–1159 (2011).
49. Wagner, G. P., Kin, K., Muglia, L., 269, M. P.ev. Evolution of mammalian pregnancy and the origin of the decidual stromal cell. *Int. J. Dev. Biol.* **58,** 117–126 (2014).
50. Kin, K., Nnamani, M. C., Lynch, V. J., Michaelides, E. & Wagner, G. P. Cell-type Phylogenetics and the Origin of Endometrial Stromal Cells. *Cell Rep* **10,** 1398–1409 (2015).
51. Fest, S. *et al.* Trophoblast–Macrophage Interactions: a Regulatory Network for the Protection of Pregnancy. *Am J Reprod Immunol* **57,** 55–66 (2007).
52. Aldo, P. B. *et al.* A Novel Three-Dimensional In Vitro System to Study Trophoblast–Endothelium Cell Interactions. *Am J Reprod Immunol* **58,** 98–110 (2007).
53. Delorme-Axford, E., Bayer, A., Sadovsky, Y. & Coyne, C. B. Autophagy as a mechanism of antiviral defense at the maternal–fetal interface. *Autophagy* **9,** 2173–2174 (2013).
54. Bayer, A. *et al.* Human trophoblasts confer resistance to viruses implicated in perinatal infection. *American journal of obstetrics and gynecology* **212,** 71.e1–71.e8 (2015).
55. Bonnin, A. *et al.* A transient placental source of serotonin for the fetal forebrain. *Nature* **472,** 347–350 (2011).

56. McKay, R. Developmental biology: Remarkable role for the placenta. *Nature* **472,** 298–299 (2011).
57. Thomson, M. The physiological roles of placental corticotropin releasing hormone in pregnancy and childbirth. *J Physiol Biochem* **69,** 559–573 (2013).
58. King, A. *et al.* Tumor-homing peptides as tools for targeted delivery of payloads to the placenta. *Science Advances* **2,** e1600349–e1600349 (2016).
59. Eidem, H. R., Ackerman, W. E., McGary, K. L., Abbot, P. & Rokas, A. Gestational tissue transcriptomics in term and preterm human pregnancies: a systematic review and meta-analysis. *BMC Medical Genomics* **8,** 27 (2015).
60. Leavey, K., Bainbridge, S. A. & Cox, B. J. Large Scale Aggregate Microarray Analysis Reveals Three Distinct Molecular Subclasses of Human Preeclampsia. *PLoS ONE* **10,** (2015).
61. Leavey, K. *et al.* Unsupervised Placental Gene Expression Profiling Identifies Clinically Relevant Subclasses of Human Preeclampsia. *Hypertension* **68,** 137–+ (2016).
62. Hughes, D. A. *et al.* Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome biology* **16,** 54 (2015).
63. Pavlicev, M. *et al.* Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface. *Genome research* **27,** 349–361 (2017).
64. Lee, J. S. *et al.* Placenta-on-a-chip: a novel platform to study the biology of the human placenta. *The Journal of Maternal-Fetal & Neonatal Medicine* **29,** 1046–1054 (2015).
65. Blundell, C. *et al.* A microphysiological model of the human placental barrier. *Lab Chip* **16,** 3065–3073 (2016).
66. Krikun, G. *et al.* A Novel Immortalized Human Endometrial Stromal Cell Line with Normal Progestational Response. *Endocrinology* **145,** 2291–2296 (2004).
67. Fitzgibbon, J., Morrison, J. J., Smith, T. J. & O'Brien, M. Modulation of human uterine smooth muscle cell collagen contractility by thrombin, Y-27632, TNF alpha and indomethacin. *Reprod. Biol. Endocrinol.* **7,** 2 (2009).
68. Lawson, H. A., Cheverud, J. M. & Wolf, J. B. Genomic imprinting and parent-of-origin effects on complex traits. *Nature reviews. Genetics* **14,** 609–617 (2013).
69. Dougherty, C. R. S. & Jones, A. D. The determinants of birth weight. *American journal of obstetrics and gynecology* **144,** 190–200 (1982).
70. Catalano, P. M. & Kirwan, J. P. Maternal factors that determine neonatal size and body fat. *Curr Diab Rep* **1,** 71–77 (2001).
71. Han, Z., Lutsiv, O., Mulla, S., McDonald, S. D. & on behalf of the Knowledge Synthesis Group. Maternal Height and the Risk of Preterm Birth and Low Birth Weight: A Systematic Review and Meta-Analyses. *Journal of Obstetrics and Gynaecology Canada* **34,** 721–746 (2012).
72. Addo, O. Y. *et al.* Maternal Height and Child Growth Patterns. *The Journal of Pediatrics* **163,** 549–554.e1 (2013).
73. Myklestad, K., Vatten, L. J., Magnussen, E. B., Salvesen, K. Å. & Romundstad, P. R. Do parental heights influence pregnancy length?: a population-based prospective study,

HUNT 2. *BMC pregnancy and childbirth* **13,** 33 (2013).
74. Morken, N.-H., Källen, K. & Jacobsson, B. Predicting Risk of Spontaneous Preterm Delivery in Women with a Singleton Pregnancy. *Paediatr Perinat Epidemiol* **28,** 11–22 (2014).
75. Zhang, G. *et al.* Assessing the Causal Relationship of Maternal Height on Birth Size and Gestational Age at Birth: A Mendelian Randomization Analysis. *PLoS Med.* **12,** e1001865 (2015).
76. Court, F. *et al.* Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome research* **24,** 554–569 (2014).
77. Frank, D. *et al.* Placental overgrowth in mice lacking the imprinted gene Ipl. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 7490–7495 (2002).
78. Constância, M. *et al.* Placental-specific IGF-II is a major modulator of placental and fetal growth. *Nature* **417,** 945–948 (2002).
79. Valen, L. V. & Weiss, R. Selection in natural populations V. Indian Rats (Rattus Rattus). *Genet. Res.* **8,** 261 (2009).
80. Richards, M. Birth weight and cognitive function in the British 1946 birth cohort: longitudinal population based study. *BMJ* **322,** 199–203 (2001).
81. Figlio, D., Guryan, J., Karbownik, K. & Roth, J. The effects of poor neonatal health on children's cognitive development. *The American Economic Review* **104,** 3921–3955 (2014).
82. Ulizzi, L. & Terrenato, L. Natural selection associated with birth weight. VI. Towards the end of the stabilizing component. *Ann. Hum. Genet.* **56,** 113–118 (1992).
83. Hiby, S. E. *et al.* Maternal KIR in Combination with Paternal HLA-C2 Regulate Human Birth Weight. *J. Immunol.* **192,** 5069–5073 (2014).
84. Pijnenborg, R., Vercruysse, L. & Hanssens, M. Fetal-Maternal Conflict, Trophoblast Invasion, Preeclampsia, and the Red Queen. *Hypertension in Pregnancy* **27,** 183–196 (2009).
85. TRIVERS, R. L. Parent-Offspring Conflict. *Integr Comp Biol* **14,** 249–264 (1974).
86. Burton, G. J. & Fowden, A. L. The placenta: a multifaceted, transient organ. *Phil. Trans. R. Soc. B* **370,** 20140066–20140066 (2015).
87. Hiby, S. E. *et al.* Maternal activating KIRs protect against human reproductive failure mediated by fetal HLA-C2. *J. Clin. Invest.* **120,** 4102–4110 (2010).
88. Nakimuli, A. *et al.* A KIR B centromeric region present in Africans but not Europeans protects pregnant women from pre-eclampsia. *PNAS* **112,** 845–850 (2015).
89. Cunnane, S. C. & Crawford, M. A. Survival of the fattest: fat babies were the key to evolution of the large human brain. *Comparative Biochemistry and Physiology* (2003). doi:10.1016/S1095-6433(03)00048-5
90. Scioscia, M., Paine, M. A., Gumaa, K., Rodeck, C. H. & Rademacher, T. W. Release of inositol phosphoglycan P-type by the human placenta following insulin stimulus: A multiple comparison between preeclampsia, intrauterine growth restriction, and

gestational hypertension. *The Journal of Maternal-Fetal & Neonatal Medicine* **21,** 581–585 (2009).
91. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366,** 362–365 (1993).
92. McVean, G. T. & Hurst, L. D. Molecular evolution of imprinted genes: no evidence for antagonistic coevolution. *Proceedings of the Royal Society of London B: Biological Sciences* **264,** 739–746 (1997).
93. Wildman, D. E. Review: Toward an integrated evolutionary understanding of the mammalian placenta. *Placenta* **32 Suppl 2,** S142–5 (2011).
94. Fischer, B. & Mitteroecker, P. Covariation between human pelvis shape, stature, and head size alleviates the obstetric dilemma. *PNAS* **112,** 5655–5660 (2015).
95. Ratajczak, C. K., Fay, J. C. & Muglia, L. J. Preventing preterm birth: the past limitations and new potential of animal models. *Dis Model Mech* **3,** 407–414 (2010).
96. World Health Organization. *Preterm birth. Fact sheet No. 363. Geneva: WHO, 2014*. (2014). doi:10.1046/j.1469-0705.1998.12050312.x/full
97. Phillips, J. B., Abbot, P. & Rokas, A. Is preterm birth a human-specific syndrome? *Evol Med Public Health* **2015,** 136–148 (2015).
98. Dobbing, J. The later growth of the brain and its vulnerability. *Pediatrics* **53,** 2–6 (1974).
99. Dobbing, J. & Sands, J. Comparative aspects of the brain growth spurt. *Early Hum. Dev.* **3,** 79–83 (1979).
100. Sakai, T. *et al.* Fetal brain development in chimpanzees versus humans. *Curr. Biol.* **22,** R791–2 (2012).
101. Martin, R. D. & MacLarnon, A. M. Gestation period, neonatal size and maternal investment in placental mammals. (1985). doi:10.1038/313220a0
102. Clauss, M., Dittmann, M. T., Müller, D. W. H., Zerbe, P. & Codron, D. Low scaling of a life history variable: Analysing eutherian gestation periods with and without phylogeny-informed statistics. *Mammalian Biology* **79,** 9–16 (2014).
103. Chuong, E. B., Rumi, M. A. K., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature genetics* **45,** 325–329 (2013).
104. Lynch, V. J. *et al.* Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep* **10,** 551–561 (2015).
105. Kapusta, A. *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS genetics* **9,** e1003470 (2013).
106. Plunkett, J. *et al.* An Evolutionary Genomic Approach to Identify Genes Involved in Human Birth Timing. *PLoS genetics* **7,** e1001365 (2011).
107. Elliot, M. G. & Crespi, B. J. Genetic recapitulation of human pre-eclampsia risk during convergent evolution of reduced placental invasiveness in eutherian mammals. *Phil. Trans. R. Soc. B* **370,** 20140069–20140069 (2015).

108. Huang, M. *et al.* Evolutionary triangulation: informing genetic association studies with evolutionary evidence. *BioData mining* **9,** 12 (2016).
109. Crosley, E. J., Elliot, M. G., Christians, J. K. & Crespi, B. J. Placental invasion, preeclampsia risk and adaptive molecular evolution at the origin of the great apes: Evidence from genome-wide analyses. *Placenta* **34,** 127–132 (2013).
110. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature reviews. Genetics* **10,** 639–650 (2009).
111. Ouyang, Y., Chen, H., Chen, H. & Chen, H. Reduced plasma adiponectin and elevated leptin in pre-eclampsia. *International Journal of Gynecology & Obstetrics* **98,** 110–114 (2007).
112. Mise, H. *et al.* Augmented placental production of leptin in preeclampsia: possible involvement of placental hypoxia. *J. Clin. Endocrinol. Metab.* **83,** 3225–3229 (1998).
113. Bartha, J. L., Romero Carmona, R., Escobar Llompart, M. & Comino Delgado, R. The relationships between leptin and inflammatory cytokines in women with pre-eclampsia. *BJOG* **108,** 1272–1276 (2001).
114. Bream, E. N. A. *et al.* Candidate gene linkage approach to identify DNA variants that predispose to preterm birth. *Pediatr. Res.* **73,** 135–141 (2012).
115. Chu, T., Mouillet, J.-F., Hood, B. L., Conrads, T. P. & Sadovsky, Y. The assembly of miRNA-mRNA-protein regulatory networks using high-throughput expression data. *Bioinformatics* **31,** 1780–1787 (2015).
116. Brubaker, D. *et al.* Finding lost genes in GWAS via integrative-omics analysis reveals novel sub-networks associated with preterm birth. *Human molecular genetics* **25,** 5254–5264 (2016).
117. Uzun, A. *et al.* dbPTB: a database for preterm birth. *Database* **2012,** bar069–bar069 (2012).
118. Uzun, A., Dewan, A. T., Istrail, S. & Padbury, J. F. Pathway-based genetic analysis of preterm birth. *Genomics* **101,** 163–170 (2013).
119. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research* **41,** D991–5 (2013).
120. Guttmacher, A. E. & Spong, C. Y. The human placenta project: it's time for real time. *American journal of obstetrics and gynecology* **213,** S3–5 (2015).
121. Bhattacharya, S. *et al.* ImmPort: disseminating data to the public for the future of immunology. *Immunol Res* **58,** 234–239 (2014).
122. DiGiulio, D. B. *et al.* Temporal and spatial variation of the human microbiota during pregnancy. *PNAS* **112,** 11060–11065 (2015).
123. *NOT-HD-15-005: Request for Information (RFI): An Internet Resource for Placental Research - Placental Atlas Tool, (n.d.).*
124. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
125. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45,** 580–585 (2013).
126. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org:

Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **43,** D789–98 (2015).
127. Kim, M. *et al*. GEneSTATION 1.0: a synthetic resource of diverse evolutionary and functional genomic data for studying the evolution of pregnancy-associated tissues and phenotypes. *Nucleic Acids Research* **44,** D908–16 (2016).
128. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics* **16,** 85–97 (2015).

CHAPTER V

INTEGRATE: A DESIRABILITY-BASED DATA INTEGRATION FRAMEWORK FOR THE PRIORITIZATION OF CANDIDATE GENES ACROSS HETEROGENEOUS 'OMICS AND ITS APPLICATION TO PRETERM BIRTH[4]

Haley R. Eidem[1], Jacob Steenwyk[1], Jennifer Wisecaver[1,2], John A. Capra[1,3,4], Patrick Abbot[1], and Antonis Rokas[1,3,4]

[1]*Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA*

[2]*Department of Biochemistry, Purdue University, West Lafayette, IN, USA*

[3]*Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA*

[4]*Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA*

[4] This chapter has been submitted for publication in BMC Medical Genomics.

ABSTRACT

The integration of high-quality, genome-wide analyses offers a robust approach to elucidating genetic factors involved in complex human diseases. Even though several methods exist to integrate heterogeneous omics data, most biologists still manually select candidate genes by examining the intersection of long lists of candidates stemming from analyses of different types of omics data that have been generated by imposing hard (strict) thresholds on P-values, fold changes, and other quantitative variables, increasing the chance of missing potentially important genes. To better facilitate the unbiased integration of heterogeneous omics data collected from diverse platforms and samples, we propose a desirability function framework for identifying candidate genes with strong evidence across data types as targets for follow-up functional analysis. Our software, integRATE, uses desirability functions to rank genes both within and across studies, identifying well-supported candidate genes according to the cumulative weight of biological evidence rather than based on imposition of hard thresholds of key variables. Our approach is targeted towards disease systems with sparse, heterogeneous omics data, so we test integRATE on one such pathology: spontaneous preterm birth (sPTB). Integrating 10 sPTB omics studies identifies both genes in pathways suspected to be involved in disease pathogenesis as well as novel genes never before linked to this syndrome. integRATE is available as an R package on CRAN (https://cran.r-project.org/web/packages/integRATE/).

INTRODUCTION

Biological processes underlying disease pathogenesis typically involve a complex, dynamic, and interconnected system of molecular and environmental factors[1]. Advances in high-throughput omics experiments have allowed for the collection of data corresponding to the genomic, transcriptomic, epigenomic, proteomic, and metabolomic elements that contribute to variation in these biological processes[2]. However, each of these omics approaches, when employed in isolation, can only capture variation within a single layer of a much more complicated biological system[3,4]. For example, even though the thousands of single nucleotide polymorphisms (SNPs) that have been linked to complex diseases or traits via genome-wide association studies (GWAS) have greatly contributed to our understanding of complex disease, we still lack in depth knowledge of the molecular mechanisms underlying the vast majority of these associations[5]. Similarly, transcriptomics studies routinely identify hundreds to thousands of differentially expressed genes between diseased and healthy tissue samples, but disentangling the disease-causing changes in gene expression from its byproducts can be far more challenging[6]. Given the limitations of each omics approach but also that each one focuses on a different layer of the biological system that each complex disease represents, integration of different types of omics data and study of their combined influences to identify the key biological pathways involved has emerged as a promising avenue for research[4].

One optimal design for an integrative omics study is to obtain diverse types of omics data from the same tissue samples or patient cohorts. The resulting data can then be vertically integrated (Figure 5.1, top left) to identify candidate genes and pathways involved in complex disease. Alternatively, a single type of omics data can be collected from a variety of tissue samples or patient cohorts, facilitating their horizontal integration across many samples, which can substantially increase the experiment's power (Figure 5.1, top right). In both vertical and horizontal integration experiments where diverse types of omics data are available from the same samples, a variety of multi-staged and meta-dimensional statistical integration approaches can be utilized (Figure 5.1, bottom)[7]. These multi-omics integration methods allow for the comprehensive modelling of complex traits and phenotypes that ultimately deepen our understanding of the key genomic factors, pathways, and interactions involved in pathogenesis or other biological outcomes. For example, multi-staged integration uses multiple steps to first identify associations between different data types and then identify associations between data types and the phenotype of interest[8], whereas meta-dimensional integration combines data simultaneously based on concatenation, transformation, or model building[9].

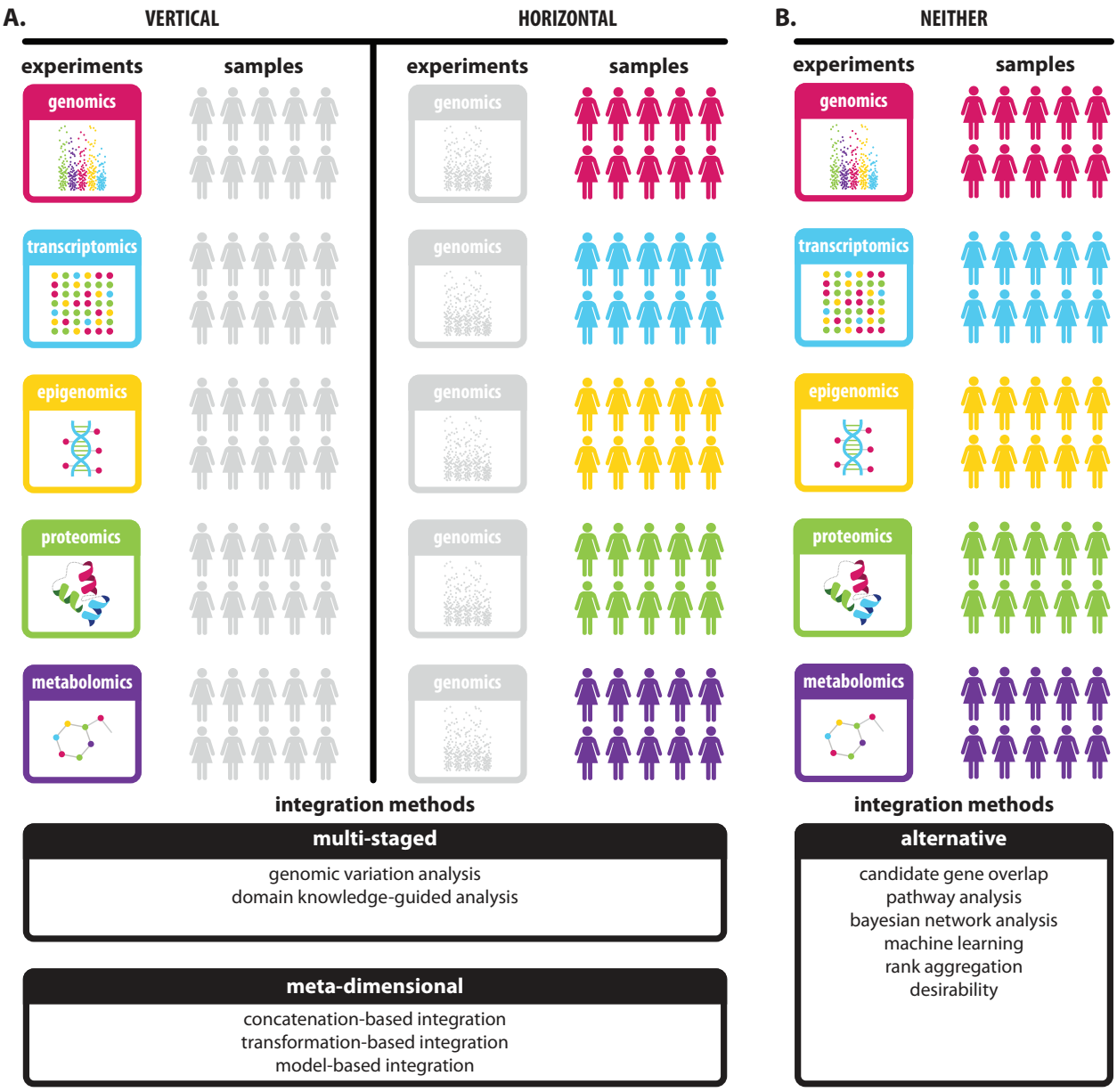

Figure 5.1. Selecting a data integration strategy depends on the structure of accessible multi-omics data. (A, left) If multiple types of omics data ('multi-omics') are available for the same cohort of patients, vertical integrative analysis can be performed to combine information across data types. This integration can be achieved using a variety of multi-staged and meta-dimensional statistical approaches that identify disease subtypes, regulatory networks, and driver genes. (A, right) If the opposite is true and a specific type of omics data is available across a number of different patient cohorts, horizontal meta-analysis can be performed to increase statistical power and identify disease-associated perturbations. (B) In some cases, however, experimental data are only available for different omics data types from

different cohorts of patients and neither vertical nor horizontal data integration can be performed. In these situations, integration relies on mapping data to common units (e.g., genes or pathways) and then either integrating transformed data or simply overlapping candidate sets. The software approach presented here (integRATE) utilizes desirability functions to transform and integrate heterogeneous data allowing for the prioritization of candidate genes for functional analysis.

Although multi-omics data sets capable of being vertically or horizontally integrated are becoming increasingly common, such data sets are lacking for many complex diseases[10-14]. Often, heterogeneous omics data are collected study by study, for a limited set of tissue samples and across only 1 or 2 omics data types at a time (Figure 5.1B, top). In these cases, the statistical methods developed for vertical and horizontal integration are not applicable without consistent sampling (of the same tissue samples or patient cohorts) or gold standards (i.e., genes known to be involved in the complex disease under investigation). For each study, a long list of genes or genomic regions with associated data are produced, depending on the omics experiment, and sorted based on effect size (e.g., fold change), significance (e.g., P-value), or some other criterion. Hard thresholds can then be imposed on P-values, for example, to bin the genes or genomic regions and identify significant candidates for further analysis; this type of approach can then be applied across multiple, heterogeneous omics studies.

Several problems exist with the imposition of hard thresholds, however. Including (or excluding) genes or genomic regions as 'candidates' based P-value, fold change, expression level, and odds ratio cutoffs introduces biases, lowers effect sizes and relationship

information, and can be conflated when combining multiple cutoffs from several criteria[15-17]. These cutoffs can sometimes even be arbitrary, like selecting the top n or n% from each data set. Additionally, statistical significance is not always equivalent to biological significance, meaning that non-statistically significant genes may still be involved in disease pathogenesis, or vice versa. Moreover, while selecting the top n genes might limit the scope of further functional analysis, the alternative approach of selecting all significant hits could mean that thousands of genes are identified as candidates. A final consideration in analyzing heterogeneous omics data is that we sometimes do not know any genes, pathways, or networks that have already been shown to be involved in complex disease. Some integration methods, especially those based on prediction (e.g., machine learning, network analysis), depend on the availability of such knowledge for algorithm training and cannot be performed in their absence[7,8,18-21].

Desirability functions represent one way to integrate heterogeneous omics data in systems where gold standards do not yet exist (Figure 5.1B, bottom). Originally developed for industrial quality control, desirability functions have been successfully used in chemoinformatics to rank compounds for drug discovery and have been proposed as a way to integrate multiple selection criteria in functional genomics experiments[22-26]. In the context of integrating diverse but heterogeneous omics data, desirability functions allow for the ranking and prioritizing of candidate genes based on cumulative evidence across data types and their variables, rather than within-study separation of significant and non-significant

genes based on single variables in single studies. For example, a 2015 study initially proposed the use of desirability functions to integrate multiple selection criteria for ranking, selecting, and prioritizing genes across heterogeneous biological analyses and demonstrated its use by analyzing a set of microarray-generated gene expression data[22].

To facilitate data integration in the presence of largely heterogeneous multi-omics data and when prior biological knowledge is limited, we propose a desirability-based framework to prioritize candidate genes for functional analysis. To facilitate implementation of our framework, we implemented it into a user-friendly software package called integRATE, which takes as input data sets from any omics experiment and generates a single desirability score based on all available information. This approach is targeted towards systems with particularly sparse or heterogeneous data, so we test integRATE on a set of 10 omics data sets related to spontaneous preterm birth (sPTB), a complex disease where multi-omics data are limited.

DESIGN

Variable Transformation

First, relevant studies need to be identified for integration based on any number of characteristics including tissue(s) sampled, disease subtype, or experimental designs (Figure 5.2, step 1). Then desirability functions are fit to each variable within a study (e.g., P-value, odds ratio, fold change, etc.) according to whether low values are most desirable (d_{low}, i.e., P-value), high values are most desirable (d_{high}, i.e., odds ratio), or extreme values are most desirable ($d_{extreme}$, i.e., fold change) (Figure 5.2, step 2). These desirability scores can be calculated by applying one of the following equations:

$$d_{low} = \begin{cases} 0 & Y > B \\ \left[\frac{Y-B}{A-B}\right]^s & A \leq Y \leq B \\ 1 & Y < A \end{cases} \quad (1)$$

$$d_{high} = \begin{cases} 0 & Y < A \\ \left[\frac{Y-A}{B-A}\right]^s & A \leq Y \leq B \\ 1 & Y > B \end{cases} \quad (2)$$

$$d_{extreme} = \begin{cases} \left[\frac{Y-A}{C-A}\right]^s & A \leq Y \leq C \\ \left[\frac{Y-B}{C-B}\right]^s & C \leq Y \leq B \\ 0 & else \end{cases} \quad (3)$$

In these equations, Y is the variable value and s is the scale coefficient affecting the function's rate of change. For d_{low} and d_{high}, A is the low cut point and B is the high cut point where the function changes. For $d_{extreme}$, A is the low cut point, C is the intermediate cut point, and B is the high cut point where the function changes. The user can customize these cut points based on numerical values (e.g., P-value < 0.05) or percentile values (e.g., top 10%). The resulting values, ranging from 0 to 1 (or the minimum and maximum values specified are transformed desirability scores based on information from each variable.

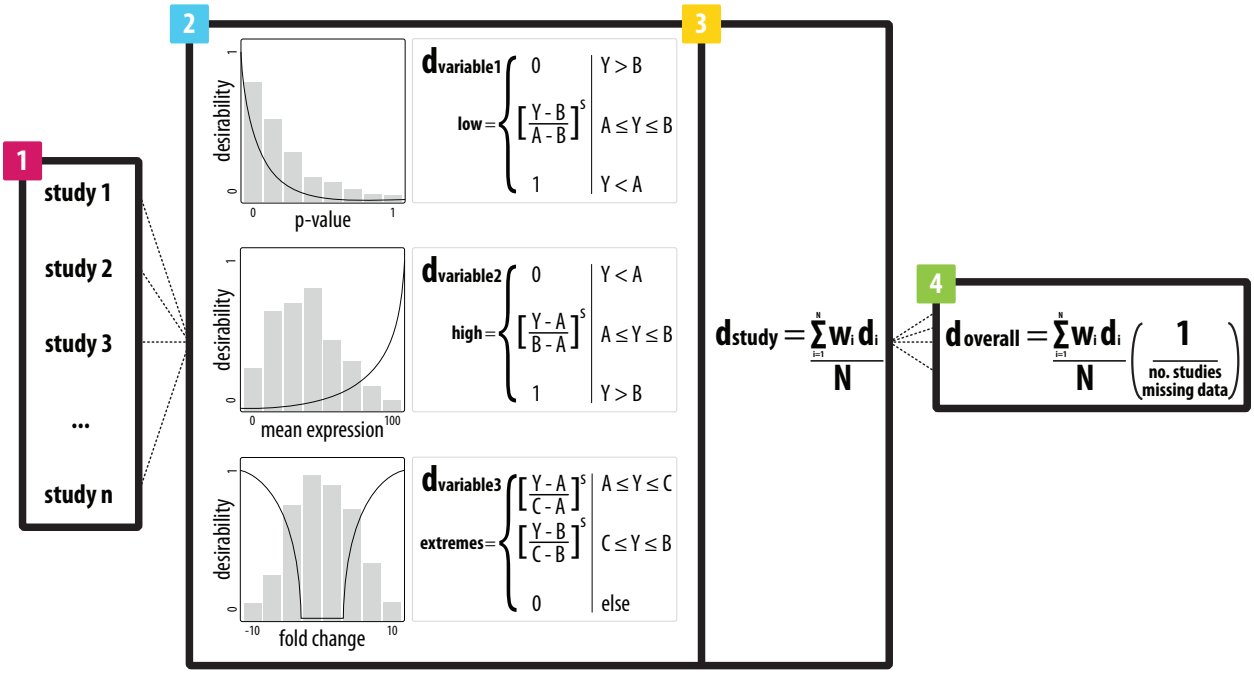

Figure 5.2. integRATE relies on three main steps to identify studies, integrate data, and rank candidate genes. (1) Relevant studies must first be identified for integration based on any number of features including, but not limited to: phenotype, experimental design, and data availability. (2) Data corresponding to all variables in each study are then transformed according to the appropriate desirability function. In this step, the user assigns a function based on whether low values are most desirable (d_{low}), high values are most desirable (d_{high}), or extreme values are most desirable ($d_{extreme}$) and can customize the shape of the function with other variables like cut points (A, B, C), scales (s), and weights (w) to better reflect the

data distributions or to align with user opinion regarding data quality and relevance. (3) These variable-based scores are integrated (d_{study}) with a straightforward arithmetic mean (where weights can also be applied) to produce a single desirability score for each gene in each study containing information from all variables simultaneously. (4) Finally, study-based desirability scores are integrated to produce a single desirability score for each gene ($d_{overall}$) that includes information from all variables in all studies and reflects its cumulative weight of evidence from each data set identified in step 1. These scores are normalized by the number of studies containing data for each gene and can be used to rank and prioritize candidate genes for follow up computational and, most importantly, functional analyses.

Variable Integration

Next, desirability scores for each variable within a study are combined using an arithmetic mean so that data points (e.g., genes) with desirability scores of zero for any given variable remain in the analysis (Figure 5.2, step 3). Desirability for data points within a study can be calculated by:

$$d_{study} = \sum_{i=1}^{N} \frac{w_i d_i}{N} \quad (4)$$

In this equation, w_i is the weight parameter (assigned to each variable), d_i is desirability score for each data point in each variable, and N is the total number of transformed variables. This step produces a single desirability score (d_{study}) for each data point in the study containing information from all transformed variables. Here, the user is also able to include variable weights (w_i) when integrating their desirability scores in case certain variables are considered more informative or accurate than others.

Study Integration

Finally, the d_{study} values can be integrated using the arithmetic mean to produce a single desirability score ($d_{overall}$) for each data point, representing its desirability as a candidate according to the weight of evidence from all variables in all studies that were integrated (Figure 5.2, step 4). The overall score used to prioritize candidates can be calculated by:

$$d_{overall} = \sum_{i=1}^{N} \frac{w_i d_i}{N} \left(\frac{1}{no.studies\ missing\ data + 1} \right) \quad (5)$$

In this equation, w_i is the weight parameter (assigned to each study), d_i is desirability score for each data point in each study, and N is the total number of studies integrated. Importantly, the overall desirability score is normalized by the number of studies missing data for each data point to account for the number of values contributing to each overall desirability score. This normalization factor can be used to calculate a soft cutoff for the most desirable candidates that is equivalent or higher than the desirability score that would be achieved by a data point with a perfect desirability score of 1 in a single study but missing from all other studies. We call genes achieving desirability scores equal to or above this cutoff 'desirable.'

Software

This methodology is implemented in our software, integRATE, available on CRAN as an R package. Although we focus on using desirability functions to integrate heterogeneous omics

data corresponding to complex human diseases, integRATE can be applied to data sets from any phenotype, species, and data type (provided that the units can all be mapped to a common set of elements, like genes). Functionality is provided for the application of customizable desirability functions as well as data visualization.

Implementation

One human complex genetic disease where the omics data available are heterogeneous is preterm birth (PTB). Defined as birth before 37 weeks of completed gestation, PTB is the leading cause of newborn death worldwide despite medical advances that continue to improve outcomes for babies born too early[27]. Evidence from family, twin, and case-control studies suggests that genetics plays a role in determining birth timing and a recent GWAS identified a handful of genes linked to prematurity[28]. Nevertheless, the pathogenesis of PTB and its many subtypes remains poorly understood[29-31].

Although 30% of preterm births are medically indicated due to complications including preeclampsia (PE) or intrauterine growth restriction (IUGR), the remaining 70% occur spontaneously either due to the preterm premature rupture of membranes (PPROM) or idiopathically (sPTB). Further complicating factors are that multiple maternal and fetal tissues are involved (e.g., placenta, fetal membranes, umbilical cord, myometrium, decidua, etc.) as well as multiple genomes (maternal, paternal, and fetal)[32].

The publicly available data for sPTB consist of several different independently conducted omics analyses that cannot be analyzed using statistical approaches developed for vertical and horizontal integration[28,33,34]. Although these omics data have been analyzed in isolation, integration of their information using the desirability-based platform implemented in integRATE might better reflect what we know is a complex system regulated by a complex interplay between many layers of biological regulation and allow for the identification and prioritization of novel candidate genes for further functional and targeted analyses.

Study Identification

Studies were initially identified based on the following PubMed search strategy:

"Pregnancy"[mh] AND "Humans"[mh] AND "Preterm birth"[mh] AND ("Gene Expression Profiling"[mh] OR "Gene Expression Regulation"[mh]) AND ("Placenta"[mh] OR "Decidua"[mh] OR "Myometrium"[mh] OR "Cervix Uteri"[mh] OR "Extraembryonic Membranes"[mh] OR "Blood"[mh] OR "Plasma"[mh] OR "Umbilical Cord"[mh])

Papers that seemed to conduct a genome-wide omics analysis of sPTB from a preliminary scan of the abstract were downloaded for full-text assessment. Furthermore, a thorough investigation was conducted of their associated reference lists to identify papers not captured via PubMed.

Additionally, each study had to meet the following inclusion criteria:

1) Experimental group consisted of sPTB cases only and wasn't complicated by other pregnancy phenotypes (e.g., preeclampsia)
2) Analysis was genome-wide and not targeted to any specific subset of genes or pathways
3) Full data set was publically available (not just top n%)

We identified 54 studies through the first phase of our literature search, but only 10 data sets met all inclusion criteria. All excluded studies are listed in Additional File 5.1 with reasons for exclusion and the 10 data sets used in our pilot analysis are outlined in Table 5.1[33-46].

Table 5.1. 10 sPTB 'omics analyses were identified for integration.

First Author	Year	Experiment	Control	Tissue	'omics Type
Zhang	2017	sPTB	term	maternal blood	genomics (GWAS)
Ackerman	2015	sPTB	term	placenta	transcriptomics (RNA-seq)
Heng	2014	sPTB	term	maternal blood	transcriptomics (microarray)
Chim	2012	sPTB	term	maternal blood	transcriptomics (microarray)
Mayor-Lynn	2011	sPTB	term	placenta	transcriptomics (microarray)
de Goede	2017	sPTB	term	cord blood	epigenomics (microarray)
Fernando	2015	sPTB	term	cord blood	epigenomics (microarray)
Parets	2015	sPTB	term	maternal blood	epigenomics (microarray)
Cruickshank	2013	sPTB	term	fetal blood	epigenomics (microarray)
Heng	2015	sPTB	term	maternal blood	proteomics (mass spec)

Data Transformation

Each of the 10 data sets was transformed to a gene-based format. This step was necessary because integRATE applies desirability functions both within and *across* studies and, in order for that integration to be possible, the genetic elements of each study have to match.

Genomics. SNP-based data containing P-values and effect sizes were mapped to genes with MAGMA, as outlined in the Zhang et al. supplementary methods (http://ctg.cncr.nl/software/magma)[35,47,48].

Transcriptomics. Gene expression data from microarray experiments were accessed via GEO (https://www.ncbi.nlm.nih.gov/geo/) and re-analyzed using the GEO2R plugin (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html)[36-39]. Raw RNA-seq data from Ackerman et al. were analyzed in-house with custom scripts[33].

Epigenomics. Methylation data were accessed via GEO (https://www.ncbi.nlm.nih.gov/geo/) and re-analyzed using the GEO2R plugin (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html)[40-44].

Proteomics. Protein expression data were downloaded from supplementary files associated with each publication and the protein IDs were mapped to genes using Ensemble's BioMart tool (https://www.ensembl.org/info/data/biomart/index.html)[34,45].

Application of integRATE

After mapping results from all 10 omics studies to genes, we applied integRATE and calculated desirabilities for all variables within studies (d_{study}) and all genes across studies ($d_{overall}$). We ran four different sPTB pilot analyses:

1) In the first analysis (**iR-none**), we ran integRATE with no added customizations (e.g., no cut points, no scales, no minimum or maximum desirabilities, etc.) (Figures 5.3-5.5, Additional File 5.2).
2) In the second analysis (**iR-num**), we ran integRATE using numerical cut points (Additional files 5.4-5.7).
3) In the third analysis (**iR-per**), we ran integRATE using percentile cut points (Additional files 5.8-5.11).
4) In the fourth analysis (**HardThresh**), we pulled statistically significant genes from each study to represent the results that would have been obtained if the typical approach based on hard thresholds and intersection of significant genes across studies outlined earlier was applied (Additional files 5.12 and 5.13). All genes with adjusted P-values < 0.1 or unadjusted P-values < 0.05 were deemed significant in each study and intersected to compare with the results from integRATE[49].

Additionally, we performed a permutation test shuffling desirabilities for all genes 1,000 times to test whether our integration strategy produced results different from what might occur at random (Additional File 5.15).

RESULTS

In total, our sPTB pilot analyses of the integRATE pipeline integrated gene-based results from 10 omics studies (1 genomics, 4 transcriptomics, 4 epigenomics, and 1 proteomics, Table 5.1) and included data sets ranging from 422 genes to 20,841 genes.

iR-none. First, the software was run without any added cuts, weights, or scales, resulting in a list of 26,868 genes with data points from one or more of the 10 omics studies (Additional File 5.2). Normalized desirabilities for these 26,868 genes ranged from 8.04E-16 to 0.46 (mean = 0.08±0.05) (Figure 5.3). Furthermore, 7,977 genes (29.7%) had desirabilities ≥ 0.1 corresponding to values equal to or higher than what would be achieved if a given gene achieved perfect desirability in one study but was absent from all others. These top 7,977 genes were enriched for 70 unique GO-Slim Biological Process categories, including pathways involved in metabolic processes, immunity, and signal transduction (Additional File 5.3)[50]. The top 10 genes (Figures 5.4 and 5.5) have desirabilities ranging from 0.46 (*CAPZB*) to 0.38 (*ACTN1*) and all are represented in each of the 10 omics data sets analyzed. This analysis applied integRATE without cut points, allowing for a straightforward, linear transformation of data across all variables and studies.

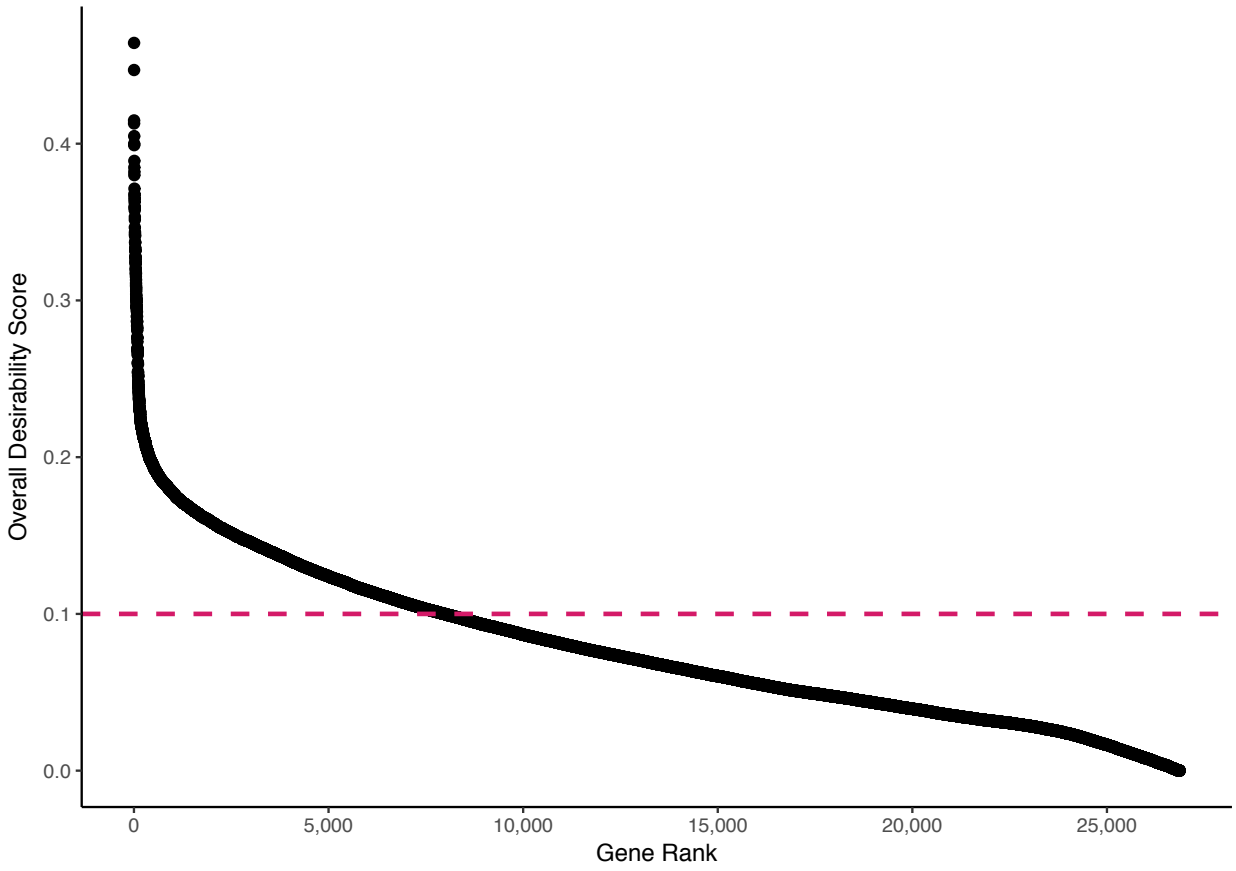

Figure 5.3. After integration, 612/26,868 genes were identified as highly desirable. All genes in the analysis were sorted from most desirable (rank = 1) to least desirable (rank = 26,868) and plotted according to their overall desirability scores, ranging from 8.04E-16 to 0.46. Because this analysis included 10 omics studies, the lower bound for our set of 'desirable' candidate genes is 0.1 (pink dashed line) and 7,977 genes achieved scores greater than or equal to that value. All desirability scores for the entire data set are available in Additional File 5.2 (and in Additional files 5.4 and 5.8 for iR-num and iR-per, respectively).

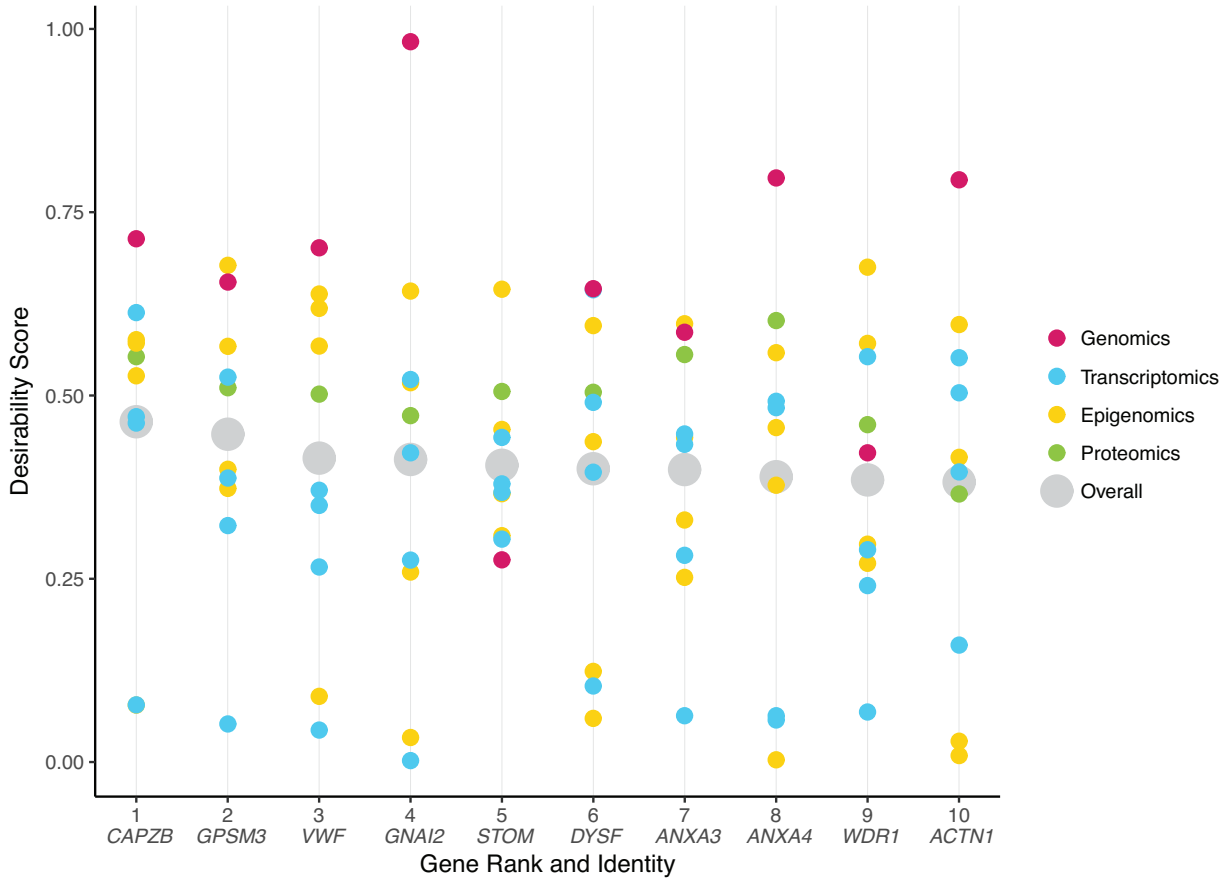

Fig 5.4. The top 10 most desirable genes have a wide range of desirabilities across data types. The top 10 genes in our analysis have overall desirabilities ranging from 0.38 (*ACTN1*) to 0.46 (*CAPZB*), but the d_{study} values range, even when organized by data type. Some genes, like *STOM*, appear to be highly ranked not because of any extremely high d_{study} value, but rather due to a lack of low d_{study} values in any data type. In other words, this gene is likely not identified as particularly important in any individual study but shows a consensus of relatively strong evidence across all 10 studies. Contrastingly, other genes, like *CAPZB*, appear to be highly ranked due to one very high desirability score in a single data type (GWAS) that overpowers underwhelming evidence in other studies.

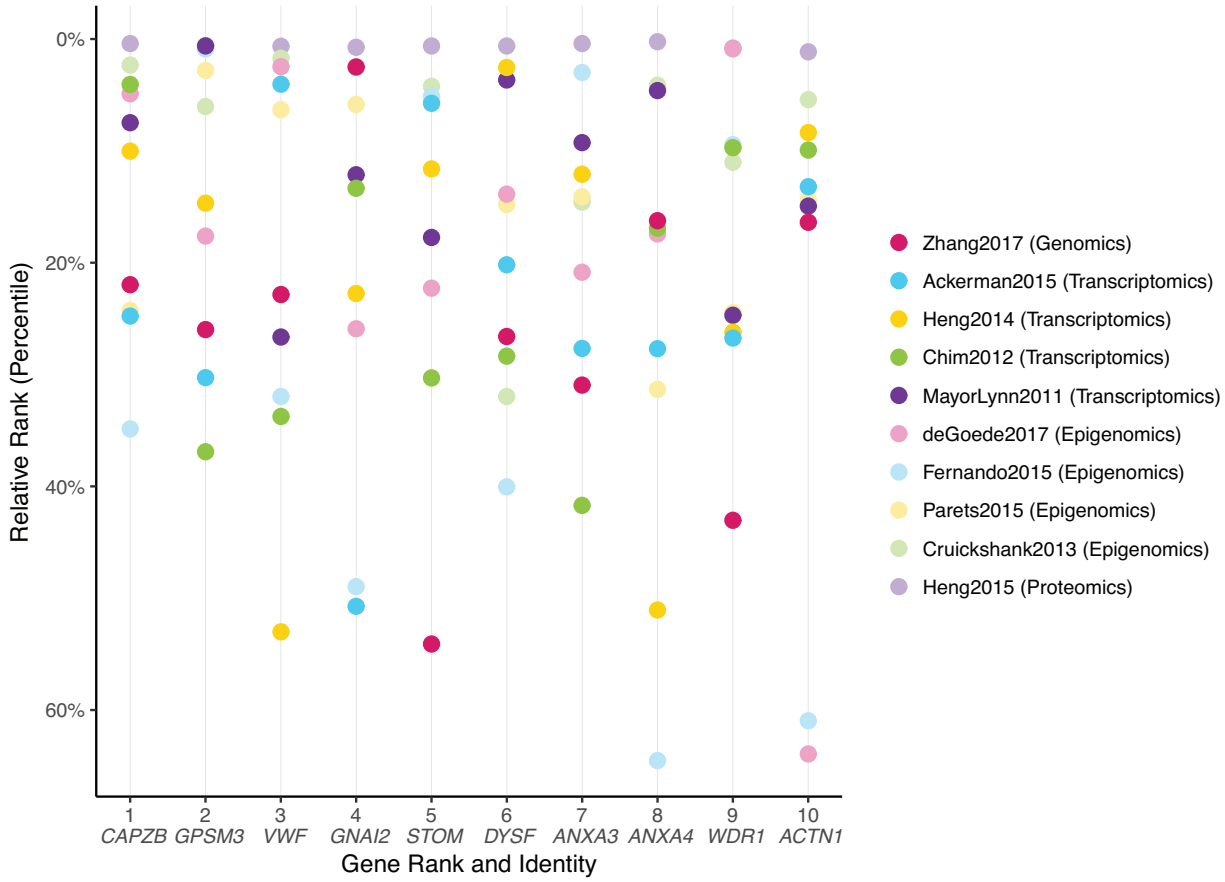

Fig 5.5. The top 10 most desirable genes also show a large discrepancy in their percentile ranks across studies. After ranking the genes in each study by desirability and calculating their percentiles based on the number of unique ranks, the top 10 most desirable genes appear to show even greater variability in relative ranking across not just data type, but individual studies. All 10 genes are in the top 25% of the (smaller) proteomics study, but their relative rankings vary significantly in all other studies. Furthermore, while none of the genes are in the top 25% of the GWAS study (Zhang2017), other studies, like one of the transcriptomics analyses (MayorLynn2011), show a large range in relative rankings, with certain highly desirable genes ranked very high and others ranked very low.

iR-num. We next applied cut points based on numerical values (Additional File 5.4). P-values such that values smaller than 0.0001 received the maximum desirability score of 1 and values larger than 0.1 received the minimum desirability score of 0. All P-values between 0.0001 and 0.1 were transformed according to the d_{low} function. For $d_{extreme}$ functions, 4 cut points are

assigned and we chose commonly used values of 0.5 and 1.5 (or their equivalents if the values were log transformed). Therefore, fold changes below -1.5 or above -1.5 (or below log2(1/3) or above log2(3)) received the maximum desirability score of 1 and fold changes between -0.5 and 0.5 (or between log2(1/1.5) and log2(1.5)) received the minimum desirability score of 0. Intermediate values were transformed according to the $d_{extreme}$ function. This approach mirrors what was applied in a previous implementation of the desirability framework for omics data, and takes into account prior knowledge of typical P-value and fold change distributions[22].

iR-per. Finally, we applied cut points based on percentiles (Additional File 5.8). P-values were cut such that those in the top 5% received the maximum desirability score of 1 and those in the bottom 5% received the minimum desirability score of 0, with all values in between transformed according to the d_{low} function. Fold changes were cut such that those in the top 5% and bottom 5% received the maximum desirability score of 1 and those in the middle 50% received the minimum desirability score of 0, with all other values transformed according to the $d_{extreme}$ function.

While the top most desirable genes in iR-num appeared to be better candidates in each individual study (Additional File 5.7), using these cut points corresponding to standard significant P-value and fold change cut offs greatly reduced the amount of data (Additional File 5.4), since some studies did not exhibit sufficiently low P-values or extreme fold changes, for example. The top 10 most desirable genes in iR-num were supported by only 4 or 5 studies

(Additional File 5.6), as opposed to in iR-none and iR-per where all studies provided data regardless of P-value or fold change.

HardThresh. For comparison, we also manually selected candidate genes by imposing a hard threshold on P-value (P-value < 0.05 if unadjusted and P-value < 0.1 if adjusted) (Additional File 5.12). After binning data into 'significant' gene lists, we intersected these lists to pull out genes that would have been identified simply by selecting the intersection of all significant genes. Although 18,727 genes were considered significant' in at least 1 study, no genes were identified as significant in all 10 studies. The best candidate gene (*KIAA0040*) was considered significant in 6/10 studies and 15 other genes were identified in 5/10 studies (Additional File 5.13). Interestingly, none of these genes appear in the top 10 of our most desirable candidates after integration and, even more generally, none are specifically discussed in any of the studies, either.

Finally, we performed a permutation test to compare our results to a null hypothesis. Desirabilities for all 26,868 genes were randomly shuffled 1,000 times, averaged to produce a null sampling distribution, and then plotted against our empirical iR-none data (Additional File 5.15). Mean desirabilities in our sample distribution ranged from 0.056 to 0.062, with an average of 0.059 (95% CI [0.058, 0.061]). In our empirical iR-none analysis, 15,285/26,868 (56.9%) genes achieved desirabilities greater than the permutation mean of 0.059.

DISCUSSION

By using desirability functions to rank genes within studies and combine results across studies, integRATE allows for the identification of candidate genes supported across experimental conditions and omics data types. This is especially important when heterogeneous sets of omics data, like those available for sPTB, where the statistical approaches developed for vertical or horizontal integration are largely inapplicable. We have shown that integRATE can map any omics data to a common [0, 1] scale for linear integration and produce a list of the most desirable candidates according to their weight of evidence across available studies. These candidates then become interesting targets for follow-up functional testing depending on where in the data their desirability signals come from. Analysis of 10 heterogenous omics data sets on sPTB showed that the gene candidates identified using desirability functions appear to be much more broadly supported than those identified by the intersection of all significant genes across all studies and contain both genes that have been previously associated with sPTB as well as novel ones (Figures 5.4 and 5.5, Additional File 5.13).

integRATE will identify both known and novel candidate genes associated with a complex disease, including ones that are not be among the top candidates in any single omics study but are consistently (i.e., across studies) recovered as significantly (or nearly significantly) associated. For example, genes that are significantly differentially expressed at an intermediate to high level across *many* studies will have high desirability scores. Furthermore,

integRATE can identify such genes across omics types, tissues, patient groups, and any other variable condition. Although integRATE allows for this kind of synergistic, desirability-based analysis, it is important to note that integRATE is not a statistical tool nor is it intended to be the end point of any analysis. integRATE is a straightforward framework for the identification of well-supported candidate genes in any phenotype where true multi-omics data is unavailable and, importantly, it serves as a springboard for future functional analysis, an essential next-step in testing whether the candidates are actually involved in the biology of the disease or phenotype at hand.

One aspect of the desirability framework that is arbitrary is the decision of what may be appropriate cut points for the different functions. In our sPTB pilot analyses (iR-none, iR-num, and iR-per), we observed that the imposition of cut points corresponding to accepted values (e.g., P-value < 0.0001) has the potential of greatly affecting the resulting gene prioritization. On this basis, we propose that desirability functions are best used to integrate highly heterogeneous omics data without imposed numerical cut points for P-values, fold changes, and other variables. Implemented this way, one can maximize the information from the analysis of each omics data set used in prioritizing candidate genes. But users may also have reasons to want to put more weight on data sets that are of higher quality or on data types that may be more informative. In such instances, the weight parameter can be used to reflect study quality instead of imposing cut points (e.g., studies that fail to achieve P-values as low

as others in the integrative analysis can be weighted less to reflect potentially lower experimental quality).

In our sPTB pilot analyses, members of the annexin family (*ANXA3*, *ANXA4* and *ANXA9*) appear in the top 10 most desirable candidate gene sets regardless of analysis approach (e.g., without cut points as well as with numerical *and* percentile cut points). This family is involved in calcium-dependent phospholipid binding and membrane-related exocytotic and endocytotic events, including endosome aggregation mediation (*ANXA6*). In a previous proteomic analysis, *ANXA3* was found to be differentially expressed in cervicovaginal fluid 26-30 days before the eventual onset of sPTB as compared to before healthy, term deliveries[51]. Furthermore, members of the annexin family are known to be involved in coagulation (*ANXA3*, *ANXA4*). Coagulation has been previously suggested to be involved in PTB and, even though the mechanism of such involvement is still a mystery, it is interesting that several genes involved in coagulation or blood disorders appear in our top candidate lists[52]. In addition to *ANXA3* and *ANXA4*, *VWF* (or Von Willebrand Factor) is a gene encoding a glycoprotein involved in homeostasis that has been found to be expressed significantly more in preterm infant serum as compared to term[53,54]. Finally, another highly desirable candidate, *STOM*, encodes an integral membrane protein that localizes to red blood cells, the loss of which has been linked to anemia[55]. These results suggest that homeostasis and coagulation might play a key role in the development of sPTB.

In addition to homeostasis and coagulation, another biological process represented across our results is actin regulation and muscle activity. The most notable gene associated with this biological process is *CAPZB*, which encodes part of an actin binding protein that regulates actin filament dynamics and stabilization and is present in the top 10 most desirable candidate gene list in all three analyses. Although *CAPZB* has never been linked to sPTB or other pregnancy pathologies, its role in muscle function could be linked to myometrial and uterine contractions that, when they occur prematurely, might be directly involved in the development of sPTB[56,57]. Another one of our top candidates, *ACTN1*, is also involved in actin regulation and, even more interestingly, has also been linked to blood and bleeding disorders[58,59]. Finally, several other highly desirable genes identified in one or more of our integrative analyses include *GPSM3*, *WDR1*, and *DYSF*, are all involved in the development and regulation of muscle or in the pathogenesis of muscle-related diseases[60-62].

Even outside the top 10 most desirable genes across our integrative analyses, we found genes both previously identified as being involved in pregnancy or sPTB pathology as well as involved in pathways potentially relevant to sPTB (Additional File 2). For example, one gene falling just outside the top 10 most desirable candidates in all analyses is *MMP9*, a matrix metalloproteinase. Interestingly, *MMP9* has been linked not only to sPTB, but also to preterm premature rupture of membranes (PPROM) and preeclampsia (PE) across a number of fetal and maternal tissues and at a variety of time points during pregnancy[63-67]. *MMP9* gene expression has been observed as significantly higher during preterm labor than during term

labor in maternal serum, placenta, and fetal membranes[68-70]. Even in the first trimester, levels of *MMP9* in maternal serum were higher in PE cases than in healthy controls, suggesting that increased *MMP9* protein expression is linked to the underlying inflammatory processes governing PE pathogenesis[66]. Finally, fetal plasma *MMP9* concentration has been found to be significantly higher in fetuses with PPROM than in early and term deliveries with intact membranes, implicating *MMP9* in the membrane rupture mechanism controlling early delivery due to membrane rupture[67]. We see similar evidence of *MMP9* as a desirable sPTB candidate maintained across omics and tissue types in our integRATE analyses, raising the hypothesis that its role in inflammation and extracellular matrix organization relates to sPTB even in the absence of PPROM or PE.

A recent GWAS analysis, the largest of its kind across pregnancy research, identified several candidate genes with SNPs linked to PTB[28]. This study linked *EBF1*, *EEFSEC*, and *AGTR2* to preterm birth and *EBF1*, *EEFSEC*, *AGTR2*, and *WNT4* to gestational duration (with *ADCY5* and *RAP2C* linked suggestively). By analyzing 43,568 women of European ancestry, this large study is the first to identify variants and genes that are statistically associated with sPTB. Interestingly, our integrative analysis identified *EBF1* as a desirable candidate ($d_{overall}$ = 0.15 [top 3%] in iR-none and $d_{overall}$ = 0.23 [top 1%] in iR-per), suggesting that this gene, in addition to GWAS, might also be functionally linked to sPTB pathogenesis across transcriptomics, epigenomics, and proteomics studies. Even when analyzing the 9 other omics studies without this GWAS data set, *EBF1* still achieved an overall desirability of 0.17, placing it in the top 2%

of all genes (Additional File 5.14). While our integrative analysis supports the identification of *EBF1* as an interesting candidate gene for follow up, the lack of signal for any of the other GWAS-identified hits also reinforces the need to approach complex phenotypes like sPTB from a variety of omics perspectives, since sequenced-based changes may impact the phenotype in indirect and complicated functional ways.

In summary, integRATE is a software solution to desirability-based data integration, most applicable in biological research areas where omics data is especially heterogeneous and sparse. Our approach combines information from all variables across all related studies to calculate the total weight of evidence for any given gene as a candidate involved in disease pathogenesis, for example. Although not a statistical approach, this method of data integration allows for the prioritization of candidate genes based on information from heterogeneous omics data even without known 'gold standard' genes to test against and can be used to inform more targeted downstream functional analyses.

AUTHORS' CONTRIBUTIONS

Conceived and designed experiments: HRE AR. Performed experiments: HRE. Developed scripts: HRE JS. Analyzed data: HRE. Wrote paper: HRE. Assisted with project development: AR JW. Provided feedback: JW PA JAC AR.

ACKNOWLEDGEMENTS

We thank Dr. Louis J. Muglia and members of the March of Dimes Prematurity Research Center Ohio Collaborative for invaluable discussion and support in designing and applying this approach to sPTB. We also thank Dr. Ge Zhang for providing access to preprocessed GWAS data. HRE was supported by a Transdisciplinary Scholar Award from the March of Dimes Prematurity Research Center Ohio Collaborative. This research was supported by the March of Dimes through the March of Dimes Prematurity Research Center Ohio Collaborative and by the Burroughs Wellcome Fund.

REFERENCES

1. Gohlke, J. M. *et al.* Genetic and environmental pathways to complex diseases. *BMC Systems Biology* **3,** 46 (2009).
2. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome biology* **18,** 83–15 (2017).
3. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148,** 1293–1307 (2012).
4. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nature reviews. Genetics* **108,** 1111 (2018).
5. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics* **11,** 446–450 (2010).
6. Casamassimi, A., Federico, A., Rienzo, M., Esposito, S. & Ciccodicola, A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. *Int J Mol Sci* **18,** 1652 (2017).
7. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics* **16,** 85–97 (2015).
8. Holzinger, E. R. & Ritchie, M. D. Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics* **13,** 213–222 (2012).
9. Kim, D., Shin, H., Song, Y. S. & Kim, J. H. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform* **45,** 1191–1198 (2012).
10. Peng, C., Li, A. & Wang, M. Discovery of Bladder Cancer-related Genes Using Integrative Heterogeneous Network Modeling of Multi-omics Data. *Sci Rep* **7,** 15639 (2017).
11. Pavel, A. B., Sonkin, D. & Reddy, A. Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity. *BMC Systems Biology* **10,** 16 (2016).
12. Zhu, J., Shi, Z., Wang, J. & Zhang, B. Empowering biologists with multi-omics data: colorectal cancer as a paradigm. *Bioinformatics* **31,** 1436–1443 (2015).
13. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
14. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17,** 98–110 (2010).
15. Cohen, J. The Cost of Dichotomization. *Applied Psychological Measurement* **7,** 249–253 (2016).
16. Streiner, D. L. Breaking up is Hard to Do: The Heartbreak of Dichotomizing Continuous Data. *The Canadian Journal of Psychiatry* **47,** 262–266 (2016).
17. Barnwell-Ménard, J.-L., Li, Q. & Cohen, A. A. Effects of categorization method, regression type, and variable distribution on the inflation of Type-I error rate when categorizing a confounding variable. *Statistics in Medicine* **34,** 936–949 (2014).

18. Reif, D. M., White, B. C. & Moore, J. H. Integrated analysis of genetic, genomic and proteomic data. *Expert Rev Proteomics* **1,** 67–75 (2004).
19. Hamid, J. S. *et al.* Data integration in genetics and genomics: methods and challenges. *Hum Genomics Proteomics* **2009,** 1–13 (2009).
20. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18,** 389–401 (2007).
21. Hawkins, R. D., Hon, G. C. & Ren, B. Next-generation genomics: an integrative approach. *Nature reviews. Genetics* **11,** 476–486 (2010).
22. Lazic, S. E. Ranking, selecting, and prioritising genes with desirability functions. *PeerJ* **3,** e1444 (2015).
23. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. 1–9 (2012). doi:10.1038/nchem.1243
24. Harrington, E. C., Jr. The Desirability Function. *Industrial Quality Control* **21,** 494–498 (1965).
25. Derringer, G. C. & Suich, R. Simultaneous optimization of several response variables. *Journal of Quality Technology* **12,** 214–219 (1980).
26. Derringer, G. C. A balancing act: optimizing a products properties. (1994).
27. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: one syndrome, many causes. *Science* **345,** 760–765 (2014).
28. Zhang, G., Jacobsson, B. & Muglia, L. J. Genetic Associations with Spontaneous Preterm Birth. *N. Engl. J. Med.* **377,** 2401–2402 (2017).
29. Plunkett, J. & Muglia, L. J. Genetic contributions to preterm birth: implications from epidemiological and genetic association studies. *Ann. Med.* **40,** 167–195 (2008).
30. Muglia, L. J. & Katz, M. The enigma of spontaneous preterm birth. *N. Engl. J. Med.* **362,** 529–535 (2010).
31. Lengyel, C., Muglia, L. J. & Pavlicev, M. *Genetics of Preterm Birth*. **283,** (John Wiley & Sons, Ltd, 2001).
32. Eidem, H. R., McGary, K. L., Capra, J. A., Abbot, P. & Rokas, A. The transformative potential of an integrative approach to pregnancy. *Placenta* **57,** 204–215 (2017).
33. Ackerman, W. E. *et al.* Comprehensive RNA profiling of villous trophoblast and decidua basalis in pregnancies complicated by preterm birth following intra-amniotic infection. *Placenta* **44,** 23–33 (2016).
34. Heng, Y. J. *et al.* Albumin decrease is associated with spontaneous preterm delivery within 48 h in women with threatened preterm labor. *J. Proteome Res.* **14,** 457–466 (2015).
35. Zhang, G. *et al.* Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. *N. Engl. J. Med.* **377,** 1156–1167 (2017).
36. Makieva, S. *et al.* The preterm cervix reveals a transcriptomic signature in the presence of premature prelabor rupture of membranes. *The American Journal of Obstetrics & Gynecology* **216,** –602.e21 (2017).
37. Heng, Y. J., Pennell, C. E., Chua, H. N., Perkins, J. E. & Lye, S. J. Whole Blood Gene Expression Profile Associated with Spontaneous Preterm Birth in Women with

Threatened Preterm Labor. *PLoS ONE* **9,** e96901 (2014).
38. Chim, S. S. C. *et al.* Systematic Identification of Spontaneous Preterm Birth-Associated RNA Transcripts in Maternal Plasma. *PLoS ONE* **7,** e34328 (2012).
39. Mayor-Lynn, K., Toloubeydokhti, T., Cruz, A. C. & Chegini, N. Expression profile of microRNAs and mRNAs in human placentas from pregnancies complicated by preeclampsia and preterm labor. *Reproductive sciences (Thousand Oaks, Calif.)* **18,** 46–56 (2011).
40. de Goede, O. M., Lavoie, P. M. & Robinson, W. P. Cord blood hematopoietic cells from preterm infants display altered DNA methylation patterns. *Clin Epigenetics* **9,** 39 (2017).
41. Hong, X. *et al.* Genome-wide DNA Methylation Associations with Spontaneous Preterm Birth in US Blacks: Findings in Maternal and Cord Blood Samples. *Epigenetics* **13,** 0–00 (2017).
42. Fernando, F. *et al.* The idiopathic preterm delivery methylation profile in umbilical cord blood DNA. *BMC genomics* **16,** 736 (2015).
43. Parets, S. E. *et al.* Fetal DNA Methylation Associates with Early Spontaneous Preterm Birth and Gestational Age. *PLoS ONE* **8,** e67489 (2013).
44. Cruickshank, M. N. *et al.* Analysis of epigenetic changes in survivors of preterm birth reveals the effect of gestational age and evidence for a long term legacy. *Genome Med* **5,** 96 (2013).
45. Saade, G. R. *et al.* Development and validation of a spontaneous preterm delivery predictor in asymptomatic women. *American journal of obstetrics and gynecology* **214,** 633.e1–633.e24 (2016).
46. Menon, R. *et al.* Amniotic fluid metabolomic analysis in spontaneous preterm birth. *Reproductive sciences (Thousand Oaks, Calif.)* **21,** 791–803 (2014).
47. Zhang, G. *et al.* Assessing the Causal Relationship of Maternal Height on Birth Size and Gestational Age at Birth: A Mendelian Randomization Analysis. *PLoS Med.* **12,** e1001865 (2015).
48. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11,** (2015).
49. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33,** 2938–2940 (2017).
50. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45,** D183–D189 (2017).
51. Heng, Y. J. *et al.* Human cervicovaginal fluid biomarkers to predict term and preterm labor. *Front Physiol* **6,** 151 (2015).
52. Velez, D. R. *et al.* Preterm birth in Caucasians is associated with coagulation and inflammation pathway gene variants. *PLoS ONE* **3,** e3283 (2008).
53. Cowman, J. *et al.* Dynamic platelet function on von Willebrand factor is different in preterm neonates and full-term neonates: changes in neonatal platelet function. *J. Thromb. Haemost.* **14,** 2027–2035 (2016).
54. Strauss, T. *et al.* Activity of Von Willebrand factor and levels of VWF-cleaving protease

(ADAMTS13) in preterm and full term neonates. *Blood Cells Mol. Dis.* **67,** 14–17 (2017).
55. Zhu, Y. *et al.* Stomatocytosis is absent in 'stomatin'-deficient murine red blood cells. *Blood* **93,** 2404–2410 (1999).
56. Littlefield, R., Almenar-Queralt, A. & Fowler, V. M. Actin dynamics at pointed ends regulates thin filament length in striated muscle. *Nat. Cell Biol.* **3,** 544–551 (2001).
57. Caldwell, J. E., Heiss, S. G., Mermall, V. & Cooper, J. A. Effects of CapZ, an actin capping protein of muscle, on the polymerization of actin. *Biochemistry* **28,** 8506–8514 (1989).
58. Bottega, R. *et al.* ACTN1-related thrombocytopenia: identification of novel families for phenotypic characterization. *Blood* **125,** 869–872 (2015).
59. Kunishima, S. *et al.* ACTN1 mutations cause congenital macrothrombocytopenia. *Am. J. Hum. Genet.* **92,** 431–438 (2013).
60. Zhao, P. & Chidiac, P. Regulation of RGS5 GAP activity by GPSM3. *Mol. Cell. Biochem.* **405,** 33–40 (2015).
61. Ono, S. Functions of actin-interacting protein 1 (AIP1)/WD repeat protein 1 (WDR1) in actin filament dynamics and cytoskeletal regulation. *Biochem. Biophys. Res. Commun.* (2017). doi:10.1016/j.bbrc.2017.10.096
62. Liu, J. *et al.* Dysferlin, a novel skeletal muscle gene, is mutated in Miyoshi myopathy and limb girdle muscular dystrophy. *Nature genetics* **20,** 31–36 (1998).
63. Athayde, N. *et al.* Matrix metalloproteinases-9 in preterm and term human parturition. *J Matern Fetal Med* **8,** 213–219 (1999).
64. Chen, J. & Khalil, R. A. Matrix Metalloproteinases in Normal Pregnancy and Preeclampsia. *Prog Mol Biol Transl Sci* **148,** 87–165 (2017).
65. Xu, P., Alfaidy, N. & Challis, J. R. G. Expression of matrix metalloproteinase (MMP)-2 and MMP-9 in human placenta and fetal membranes in relation to preterm and term labor. *J. Clin. Endocrinol. Metab.* **87,** 1353–1361 (2002).
66. Poon, L. C. Y., Nekrasova, E., Anastassopoulos, P., Livanos, P. & Nicolaides, K. H. First-trimester maternal serum matrix metalloproteinase-9 (MMP-9) and adverse pregnancy outcome. *Prenat. Diagn.* **29,** 553–559 (2009).
67. Romero, R. *et al.* Fetal plasma MMP-9 concentrations are elevated in preterm premature rupture of the membranes. *The American Journal of Obstetrics & Gynecology* **187,** 1125–1130 (2002).
68. Tency, I. *et al.* Imbalances between matrix metalloproteinases (MMPs) and tissue inhibitor of metalloproteinases (TIMPs) in maternal serum during preterm labor. *PLoS ONE* **7,** e49042 (2012).
69. Sundrani, D. P., Chavan-Gautam, P. M., Pisal, H. R., Mehendale, S. S. & Joshi, S. R. Matrix metalloproteinase-1 and -9 in human placenta during spontaneous vaginal delivery and caesarean sectioning in preterm pregnancy. *PLoS ONE* **7,** e29855 (2012).
70. Yonemoto, H. *et al.* Changes in matrix metalloproteinase (MMP)-2 and MMP-9 in the fetal amnion and chorion during gestation and at term and preterm labor. *Placenta* **27,** 669–677 (2006).

CHAPTER VI

CONCLUSION

'omics research across PTB subtypes

By performing a systematic review and meta-analysis of gene expression studies in pregnancy and its pathologies in chapter II, I first showed that large gaps exist in the transcriptomic analysis of PTB subtypes, particularly in sPTB. Second, I showed that gene expression profiles are highly heterogeneous both *across* tissues and phenotypes as well as *within* the same tissue and phenotype, indicating that more high quality, large-scale data will be necessary to understand the complex landscape of gene expression and its regulation during pregnancy, parturition, and prematurity. Although this research focused primarily on gene expression, data were also collected from microRNA and methylation studies and similar conclusions were drawn about the sparsity of genome-wide research and heterogeneity of expression and methylation patterns[1].

Other meta-analyses have been published focusing on sPTB proteomics or pathology-specific biomarkers[2,3]. Similarly, a meta-analytical approach has also been taken to better understand gene-gene interactions conferring sPTB risk[4]. Due to the heterogeneity found between pathologies as well as the sparsity of pregnancy-related 'omics data, future research should combine the gene expression, methylation, and microRNA studies from my meta-analysis

with these proteomics and biomarker studies as well as with other publicly available 'omics studies (e.g., metabolomics) for use in future integrative analyses.

Our lab has been developing a database, GEneSTATION, designed to accomplish this goal, among many others[5]. Currently, GWAS, RNA-seq, gene expression microarray, and methylation data sets related to sPTB and other PTB subtypes are described in GEneSTATION with important metadata associated about experimental and control group designation, gestational age, and tissue type. Future work will expand this resource to include proteomic and metabolomic data as well as add functionality for sorting and downloading data sets based on phenotype. A combination of automated as well as expert-curated approaches to maintaining this resource will be necessary to build an up-to-date resource of appropriately organized, subtype-specific PTB 'omics data.

Gene expression in sPTB

In chapter III, I analyzed both human and macaque RNA-seq data to identify transcripts that are differentially expressed in placenta between sPTB and term pregnancies. Comparing gene expression from human placental tissue at preterm and term identified significantly differentially expressed transcripts that, in addition to sPTB pathology, also likely represented differences in gestational age of the samples. By using gestational age-matched control samples from macaque, I was able to parse this list of differentially expressed genes and categorize them as either sPTB-specific or gestational age-specific. Some of our sPTB-specific

candidates are involved in inflammatory and immune processes (*L1CAM*, *TRIM14*, *FOXO1*, *VSIG4*, *CEBPD*) and others exhibit matrix metalloproteinase activity (*ETS2*). Several sPTB-specific candidates have also been linked to sPTB or other pregnancy pathologies in previous research (*HTRA4*, *CD163*, *ADORA3*, *PDE2A*, *NAALAD2*). These are obvious candidates for future functional analyses in placental cell lines or mouse models.

Experimentally, our collaborators have recently accessed additional placental tissue samples from macaque who delivered at term. Once sequenced for RNA expression, these samples can be used to weed out normal differences between preterm and term macaque placenta and, thus, refine our comparison with human data. Furthermore, placental gene expression from preterm and/or term deliveries in other primate species (e.g., lemur) might even be considered to expand this comparative analysis.

In light of the high heterogeneity in placental genes expressed both within and across human populations, more research is necessary to identify the common pathways regulating parturition and prematurity in this tissue[6]. To this end, another collaborator collected single cell RNA-seq data from placenta to help uncover key genes and regulators involved in communication at the maternal-fetal interface[7]. This data showed strong signals from growth factors and immune pathways, similar to results from our transcriptomic and integrative analyses. Nevertheless, more work is required in tissue samples as well as single cells to

compare gene expression profiles over time, in different pathologies, and across populations and identify pathways dysregulated leading up to PTB.

The promise of data integration in sPTB

While chapter IV outlined why data integration is so crucial in a complex, multifactorial syndrome like sPTB, chapter V offered an approach to accomplish this task with the heterogeneous data at hand. The desirability-based method used by integRATE is intended to prioritize candidate genes based on their weight of evidence across high-quality data for follow-up functional testing. Therefore, future research could closely investigate the roles of some of our most promising candidates (*ANXA3*, *ANXA4*, *ANXA9*, *VWF*, *STOM*, *CAPZB*, *ACTN1*, *GPSM3*, *WDR1*, *DYSF*, *MMP9*) in gestational tissues during pregnancy. Many of these candidate genes are involved in homeostasis or muscle activity, suggesting that future analyses might investigate these pathways in greater detail to determine whether their dysregulation contributes directly to sPTB pathogenesis.

Until recently, no GWAS had been successful in identifying SNPs associated with sPTB. After combining data from several birth cohorts as well as from direct-to-consumer genetic testing, our collaborators identified several loci significantly associated with sPTB (*EBF1*, *EEFSEC*, *AGTR2*) and we were able to utilize this data in our integrative analysis (chapter V)[8]. Sample sizes are continuing to grow with the establishment of large cohorts, biobanks, and consortia, and these larger samples will allow future GWAS to identify other common maternal and fetal

polymorphisms associated with increased sPTB risk and further refine our understanding of its pathogenesis.

As more and more data sets are collected across tissue types, pathologies, and 'omics platforms, we become increasingly able to utilize statistical integration strategies like network-based and machine learning approaches to identify genetic regulators of sPTB. These methods have successfully been used to discover driver genes in cancer, but they require multi-omics data not currently available in sPTB or any other pregnancy pathology[9]. The collection of this type of data is a critical next step in advancing PTB research and will require the careful establishment of cohorts where samples are collected and sequenced across time and tissues, but within the same patients. For example, an ideal sPTB study would follow a large cohort of women through pregnancy and collect tissue samples (e.g., maternal blood) during each trimester for multi-omic sequencing. This type of data could shed light on the interconnected genomic, transcriptomic, epigenomic, proteomic, and metabolomic components leading to sPTB pathogenesis.

Summary

Together, my dissertation work has contributed novel methods for the study of sPTB as well as insights into its pathogenesis. This research demonstrates the utility of meta-analytical (chapter II), comparative transcriptomic (chapter III), and integrative genomic (chapters IV and V) approaches to better understanding a complex phenotype with sparse, heterogeneous

data. As PTB rates around the world continue to rise despite all we've learned about its pathogenesis, continued research is essential to discover predictive biomarkers and drug targets that lessen the burden of prematurity.

References

1. Eidem, H. R., Ackerman, W. E., McGary, K. L., Abbot, P. & Rokas, A. Gestational tissue transcriptomics in term and preterm human pregnancies: a systematic review and meta-analysis. *BMC Medical Genomics* **8,** 27 (2015).
2. Conde-Agudelo, A., Papageorghiou, A. T., Kennedy, S. H. & Villar, J. Novel biomarkers for the prediction of the spontaneous preterm birth phenotype: a systematic review and meta-analysis. *BJOG* **118,** 1042–1054 (2011).
3. Liu, Y., Liu, Y., Zhang, R., Zhu, L. & Feng, Z. Early- or mid-trimester amniocentesis biomarkers for predicting preterm delivery: a meta-analysis. *Ann. Med.* **49,** 1–10 (2017).
4. Uzun, A., Dewan, A. T., Istrail, S. & Padbury, J. F. Pathway-based genetic analysis of preterm birth. *Genomics* **101,** 163–170 (2013).
5. Kim, M. *et al.* GEneSTATION 1.0: a synthetic resource of diverse evolutionary and functional genomic data for studying the evolution of pregnancy-associated tissues and phenotypes. *Nucleic Acids Research* **44,** D908–16 (2016).
6. Hughes, D. A. *et al.* Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome biology* **16,** 54 (2015).
7. Pavlicev, M. *et al.* Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface. *Genome research* **27,** 349–361 (2017).
8. Zhang, G., Jacobsson, B. & Muglia, L. J. Genetic Associations with Spontaneous Preterm Birth. *N. Engl. J. Med.* **377,** 2401–2402 (2017).
9. Yang, H., Wei, Q., Zhong, X., Yang, H. & Li, B. Cancer driver gene discovery through an integrative genomics approach in a non-parametric Bayesian framework. *Bioinformatics* **33,** 483–490 (2017).

APPENDIX

All additional files available at: https://doi.org/10.6084/m9.figshare.5930743.v4