HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 3 (2023)

Non-Invasive Biomarker Detection for Pregnancy Complications Using Cell-Free RNA

Meghana Somu

Monta Vista High School

ABSTRACT

Pregnancy complications pose a significant threat to maternal health as they may result in a higher risk for issues during pregnancy or labor relative to the risk for these issues in a typical pregnancy. Many cases of maternal deaths and complicated pregnancies can be avoided with a richer understanding of maternal health early on in pregnancy. Genetic analysis of fetal DNA in maternal blood is becoming increasingly common[1], and while genetic screening efforts have progressed substantially in recent years, they have focused on fetal health rather than the health of the mother[2]. This work focuses on the detection of common complications of pregnancy including preeclampsia, gestational diabetes, and chronic hypertension using non-invasive circulating cell-free RNA data. We developed interpretable supervised machine learning methods that had high performance in identifying pregnancy complications from healthy pregnancies (AUC = 0.86). Using our models, we found various relevant transcripts, related to pregnancy biology. These included S100A9, which encodes for a protein involved in inflammation and was elevated in complicated pregnancies, as well as two small RNAs involved in cell proliferation and body mass, RNY4 and RNY3, which were reduced in preeclampsia and GDM and have previous roles in pregnancy. Our findings highlight several promising non-invasive biomarkers for the early diagnosis of complications of pregnancy that have the potential to be easily integrated into existing clinical workflows.

Introduction

Each year in the United States, 50,000 to 60,000 mothers experience complications from pregnancy and delivery that can have severe health impacts. 650 to 750 maternal deaths occur in the United States yearly[3]. Pregnancy complications may result in a higher risk for issues during pregnancy or labor relative to the risk for these issues in a typical pregnancy. Many cases of maternal deaths and complicated pregnancies can be avoided with a richer understanding of maternal health early on in pregnancy[3].

Cell-free DNA (cfDNA) refers to all non-cellular DNA and nucleic acid fragments that enter the bloodstream during necrosis or apoptosis. Cell-free DNA molecules were first identified in 1948, and it was subsequently found that cfDNA was generally present in higher levels among patients with certain diseases compared to healthy patients[4]. Recent work on cfDNA has shown its potential as a biomarker in the fields of non-invasive cancer detection and monitoring, autoimmune disease detection and monitoring, organ transplantation monitoring, prenatal genetic testing, and pathogen detection[5]. Similarly, cell-free RNA (cfRNA) presents an opportunity for non-invasive disease detection and monitoring, particularly in the case of overexpression for tissue-specific transcripts.

Non-invasive prenatal testing (NIPT), a commonly used method of understanding fetal health, uses data from a blood draw. While these genetic screening efforts have advanced significantly in recent years, they have focused on fetal health rather than maternal health[2]. Samples used for NIPT can also be used to obtain cfRNA. Since cfRNA has been shown to be useful in pregnancy contexts[2], samples collected from this commonly performed procedure could also be utilized in testing for complications of pregnancy, thereby reducing the number of procedures that pregnant women must endure to monitor their health. The collection of this data is non-invasive, as well as more accessible.

Several of the most common complications of pregnancy such as gestational diabetes (GDM) and preeclampsia (PE) are tested for and diagnosed in the second trimester using current technologies; therefore, our work focuses on early-stage testing before traditional tests could diagnose these diseases.

In this study, we aim to identify early-trimester cell-free biomarkers specific to patients with pregnancy complications. We developed computational tools to explore the transcriptomes of healthy patients and patients with complicated pregnancies. Taken together, these results suggest an opportunity to detect these diseases early in pregnancy using non-invasive methods.

Results

Data

Our work used high-throughput cell-free RNA-seq data because RNA-seq is a powerful and adaptable technique for measuring gene expression at a genome-wide level[6]. We chose this particular dataset because they take samples from patients at each point in pregnancy. This allowed us to observe the changes in expression as a disease progresses, along with the fact that the relatively short life of cfRNA is convenient for observing fluctuating expression levels in real time. This covered a limitation of many bioinformatics studies that use data from biopsies, which are difficult to obtain and static.

We did not use cord plasma sample data in our models because it was determined to be irrelevant to finding biomarkers of pregnancy complications early on in the pregnancy. We also did not include samples from non-pregnant women. The original study that produced this dataset used samples from non-pregnant patients as their control case, however, we chose not to do so as using non-pregnant samples as a control case is not optimal for building a model to distinguish complicated pregnancies from healthy pregnancies. Instead, we chose samples from healthy pregnant patients to be the controls (Figure 1).

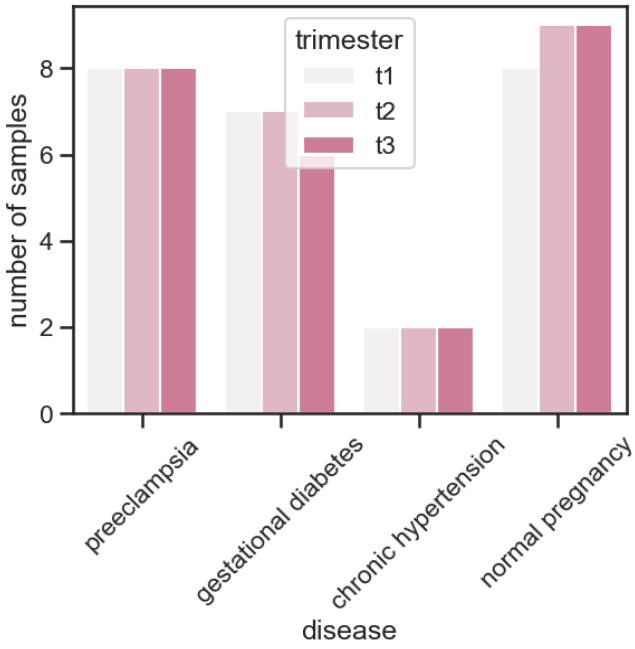

Figure 1. Bar plot of the number of downloaded samples per condition and trimester.

Pregnancy complication differential expression

We performed differential expression analysis in edgeR to examine large-scale transcriptomic differences. We began by identifying significantly differentially expressed transcripts with a Benjamin-Hochberg adjusted p-value of less than or equal to 0.05 (p-value <= 0.05) in each two-way analysis. The expression levels of these transcripts were further examined in plots that detailed how expression differentiated between patients with different complications as well as the controls. For example, we identified two transcripts with elevated expression during the first trimester in patients with chronic hypertension (Figure 2). One encodes for SCARA3, which is expressed due to oxidative stress[7], and the other encodes for MALSU1, a protein involved in mitochondrial translation and ribosomal large subunit biogenesis[8].

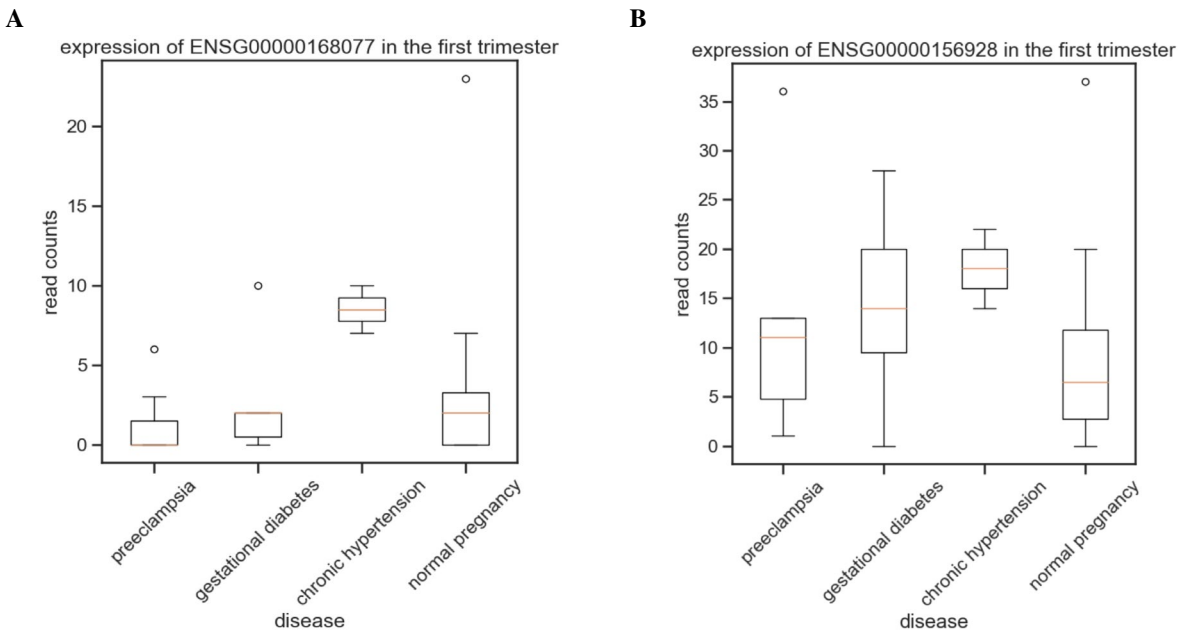

Figure 2. Box plots detailing the expression of differentially expressed transcripts in patients with each condition.

Development of biomarker machine learning protocols

After conducting differential expression analysis, we developed predictive machine learning models to take into consideration genome-wide transcriptome changes in pregnancy complications. Furthermore, we examined whether these models could prioritize transcripts as early-trimester cell-free biomarkers specific to patients with pregnancy complications.

We conducted several two-way analyses between diseased samples and controls, PE and controls, GDM and controls, and chronic hypertension and controls. The goal of using multiple different machine learning methods was to figure out which model was best suited to recognizing patterns in this particular dataset. Multiple feature selection methods were also explored to find the most effective and appropriate dimensionality reduction method for this data. For these analyses, we built and evaluated logistic regression models that used only differentially expressed genes as input, lasso regression models, binary neural networks that used differentially expressed genes as input as well as L1-based feature selection, random forest classifier models, and svm models that employed linear kernels. All models

created used supervised machine learning and all models employing cross-validation used leave-one-out cross-validation.

Model performance

Logistic models

We created 12 logistic models in total with the goal of determining a reasonable starting point for more in-depth machine learning analyses. Models were evaluated using leave-one cross-validation. The average accuracy was computed for each model (Figure 3). Overall, the average accuracy for each model was between 50% and 85%, suggesting relatively poor performance. However, the accuracy was higher, 72% and 85%, for the chronic hypertension samples in trimesters 1 and 2, although this category had substantially fewer samples than other conditions.

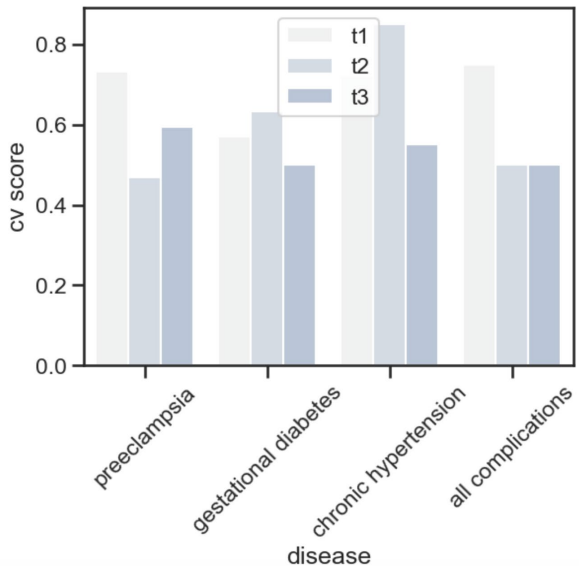

Figure 3. Bar plot showing the cross-validated accuracy scores of logistic models created to predict each disease at each point in pregnancy.

Neural networks

Neural networks have the ability to recognize more complex, non-linear trends within data and use very different methods from that of logistic or lasso regression. We created three binary neural network models in total with the goal of determining whether a more complex machine-learning approach to this data would be more proficient. Since neural networks generally perform best with many more samples than input variables, we reformulated our problem to focus on differentiating complicated pregnancies from healthy pregnancies rather than creating models to predict each individual disease. Differential expression and feature selection were used to reduce the dimensionality of the data, as neural networks tend to perform better with simpler data.

To optimize model performance and remove uninformative features, we tested several methods of feature selection on the binary neural networks to optimize performance. We created and ran a model using trimester one read count data with no feature selection and got an accuracy score of 0.75. After considering the accuracy score and results from the confusion matrix, we tried employing various methods of feature selection to improve performance (supplementary Figures S1-S5). We observed that L1-based feature selection and differential expression did not result in

overfitting and subsequently combined differential expression with L1-based feature selection (accuracy = 0.86). Following the development of this model, we constructed a binary neural network model that used only second-trimester data (accuracy 0.71) as well as a model that used third-trimester data (accuracy 0.71) with the same dimensionality reduction techniques. Overall, we found that model performance was best when differential expression and L1-based feature selection were applied. We tested further models as well, however, performance was not as successful as with lasso and neural networks (see Supplemental Note for more details).

Lasso models

Similarly, to the logistic models, we created 12 lasso models. Lasso regression models were used because the shrinkage of coefficients allows for built-in feature selection, and it reduces variance and minimizes bias. One of the main goals of creating these models was to examine if models could perform more favorably on all the read count data rather than just the data from differentially expressed genes. Our other goal in trying lasso models was to develop an explainable model that would allow us to understand and investigate the coefficients that were being chosen.

We found that lasso models had low AUC when differential expression was used (Figure 4), so we decided to build lasso models that did not utilize this technique. We found that lasso models had high cross-validated accuracy without using differential expression (Figure 5). The accuracy ranged from 46% to 83.33%, with the highest-performing model predicting all pregnancy complications during the first trimester. We noted that performance was worst for models that predicted GDM and chronic hypertension, likely because these were the cases with the least number of samples. Models that predicted PE or all pregnancy complications performed the best, most likely due to the dataset including more samples that fit these cases.

We observed that model accuracy tended to decline over each trimester, likely because many genes tend to increase in expression over the course of pregnancy[9] regardless of pregnancy complications or disease. However, this indicates that our models may be well suited to identifying early-stage complications.

Biomarker characterization

Differentially expressed genes were investigated using Gene Ontology Enrichment Analysis[10], however, no significant results were identified by the algorithm. Given that the lasso models performed well and lasso model coefficients indicating the importance of a transcript in making a disease prediction can be retrieved, we studied the genes with large coefficients. Interestingly, there was a high number of overlaps between the transcripts identified as important for each condition (number of overlapping transcripts = 186). However, to identify potential early-stage disease-specific biomarkers, we focused on the trimester 1 model in which the two categories were healthy patients or patients with pregnancy complications.

For each important transcript, we looked at GTEx tissue-specific expression (dbGaP accession number phs000424.vN.pN on 02/25/2023) to identify which tissues these genes were related to, and performed a literature search on how these transcripts may be related to pregnancy, pregnancy complications, or adverse pregnancy outcomes (supplementary Table S1). A few key transcripts supported by relevant research are displayed in Table 1. These include S100A9 which was elevated in PE and GDM pregnancies (Figure 6), as well as FHDC1 which was elevated in patients with chronic hypertension (Figure 6), and two small RNAs involved in cell proliferation and body mass. S100A9 is highly expressed in the cervix, esophagus mucosa, and vagina and is associated with cervical cancer[11], early pregnancy loss[12], and is strongly related to inflammation[13] which is a common symptom in PE. FHDC1 is highly expressed in the thyroid and is associated with GDM and preterm birth[14], as well as hyperthyroidism[15] which can lead to high blood pressure, poor fetal growth, and premature delivery. RNY4 and RNY3, which were reduced in PE and GDM (Figure 6) and have previous roles in pregnancy.

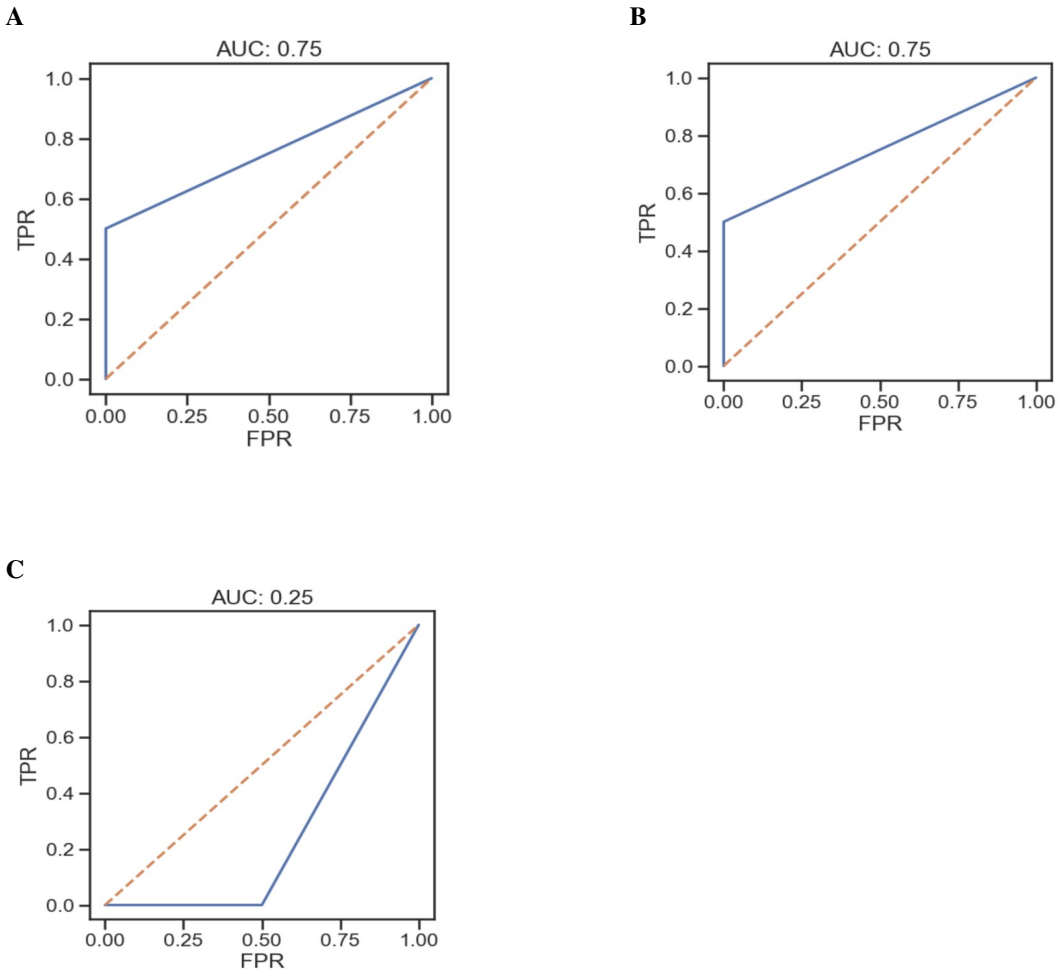

Figure 4. AUC plots for each trimester model for preeclampsia using differential expression.

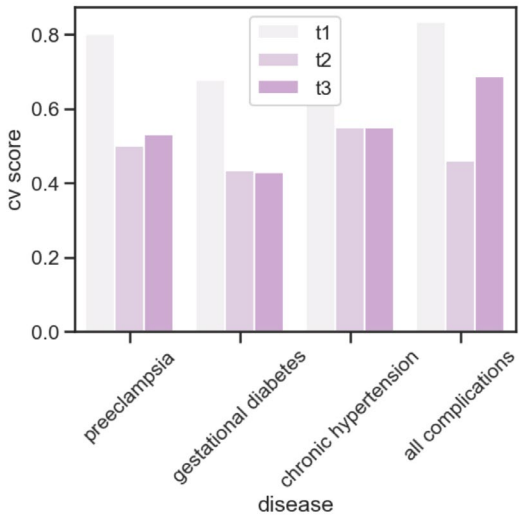

Figure 5. Cross-validated accuracy scores for each lasso model created.

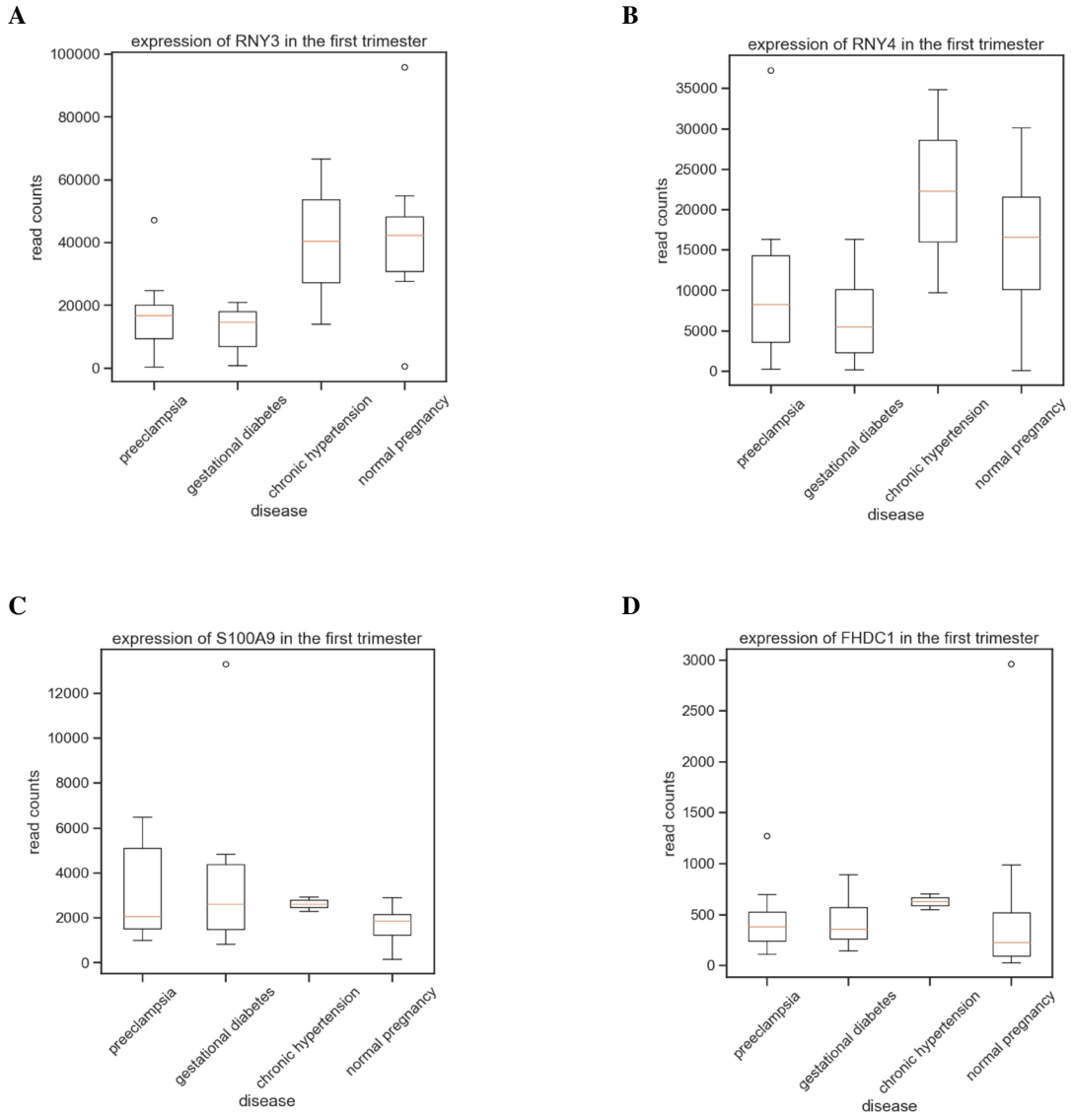

Figure 6. First-trimester expression box plots for a few key transcripts of interest.

Table 1. Selected lasso genes and their roles in pregnancy complications.

Biomarker candidate	GTEx expression	Role
FHDC1	Increased in thyroid	Associated with GDM, preterm birth[14] & hyperthyroidism[15] which can lead high blood pressure, poor fetal growth, and premature delivery
S100A9	Increased in the cervix, esophagus mucosa, vagina	Associated with cervical cancer[11], early pregnancy loss[12], strongly associated with inflammation[13]
RPL8	Increased in ovary	Regulates iron metabolism[16], associated with T2D[17]
MTRNR2L12		Differentially expressed in PE[18], associated with maternal obesity and T2D[19, 20]
RPS18	Increased in ovary	Associated with PE[21, 22]
EEF1A1	Inceased in spleen, ovary	Differentially expressed in PE[23], elevated in GDM and PE placentas[24], related to inflammation[25]

Discussion

Pregnancy complications pose a significant threat to maternal health, but many cases of complicated pregnancies can be avoided with a richer understanding of maternal health early on in pregnancy. In this work, we focus on identifying pregnancy complications in early-trimester pregnancies, using non-invasive circulating cell-free RNA data. We developed an interpretable supervised lasso model that had high performance in identifying pregnancy complications from healthy pregnancies (AUC = 0.86). We conducted further research on the lasso model coefficients to gain a deeper understanding of the transcripts these models identified to be relevant in classifying complicated pregnancies. Our findings highlight several promising biomarkers for the early detection of pregnancy complications.

The design of our study was devised with the goal of avoiding introducing bias in our work to the best of our ability. We chose not to use samples from non-pregnant women as our controls because the goal of our models was to distinguish healthy pregnancies from complicated pregnancies and not to determine the differences between the transcriptome of a pregnant patient and a non-pregnant patient. Using samples from non-pregnant patients as the control samples were likely to introduce bias to the models created in the original experiment. We also chose not to use cord blood data because it was determined to not be as relevant to detecting early-trimester biomarkers. We focused on early-trimester biomarker detection because common pregnancy complications like PE and GDM are diagnosed during the second trimester using current technologies. The use of cfRNA in this work is also important to note because the process of cfRNA collection is non-invasive. The relatively short life of cfRNA also allowed the observation of fluctuating expression levels in real-time, as opposed to static data from biopsies. We used a non-biased, interpretable ML model, which provided valuable insight into which transcripts were relevant to classifying a sample as healthy or complicated. This allowed us to identify biomarkers for complicated pregnancies as opposed to just

determining if machine learning techniques would be able to predict complications of pregnancy. Biomarker identification will be important for possible clinical applications of this work.

Though the study that produced this dataset claims that they collected samples from a diverse group of women, they did not list any details about their patients regarding age or ancestral background. This prevented us from further examining trends in the data that may have been influenced by age or ethnicity. Understanding the correlation between pregnancy complications and maternal age would have been a particularly important factor to examine because of the known link between the two[26]. Because older women are particularly at risk for pregnancy complications, it would be worthwhile to see if certain potential biomarkers are more specific to older women and, if so, study their ontology.

One of the largest limiting factors of this work was the small sample size. Samples from less than 30 women were used, which was an issue because it likely hindered the learning ability of the machine learning models. It is generally known that too little training data can result in a poor approximation and may cause the model to overfit. In the future, to validate our candidate biomarkers, a larger sample size could be used to improve model training. We could also do this by exploring a whole genome bisulfite sequencing dataset produced from the same experiment. This would involve creating several models and optimizing them for this dataset, then validating previously found potential biomarkers and exploring newly identified transcripts of interest. Because our work focuses on biomarker identification, this information can be used in a clinical setting using targeted sequencing methods or qPCR tests. Both of these methods are cheaper and more readily available than whole-genome sequencing, a commonly used method in similar bioinformatics studies.

In this work, we focused on identifying pregnancy complications in early-trimester pregnancies using non-invasive circulating cell-free RNA data. We developed an interpretable supervised lasso model that had high performance in identifying pregnancy complications from healthy pregnancies (AUC = 0.86). Our findings highlight several promising biomarkers for the early detection of pregnancy complications. This work will be important in medicine for monitoring maternal health in a comfortable, convenient way as well as detecting pregnancy complications before they become serious.

Methods

RNA-seq data

The dataset selected for analysis in the present study is publicly available from the Gene Expression Omnibus[27] under the accession number GSE154377. The original study that produced this dataset performed RNA sequencing from healthy pregnant patients, non-pregnant patients, and pregnant patients with certain pregnancy complications. The complications studied included GDM, preeclampsia (also referred to as gestational hypertension), and chronic hypertension. The original study defined GDM as any degree of glucose intolerance with an initial recognition during pregnancy. They defined preeclampsia as new-onset hypertension with new onset of thrombocytopenia, renal insufficiency, impaired liver function, pulmonary edema, and cerebral or visual symptoms while chronic hypertension was defined as blood pressure (BP) of 140/90 mm Hg or higher, that either predates pregnancy or develops before 20 weeks of gestation.

The study collected blood samples from nine patients with healthy pregnancies, seven with GDM, eight with preeclampsia, and two with chronic hypertension. It also examined a group of seven control samples from non-pregnant women, resulting in a total of 134 samples collected. A total of 127 samples, including only pregnant patients, were downloaded from the dataset. This included samples from each trimester and cord plasma samples, though the latter was not used as input for our machine-learning models. We directly downloaded the samples as TXT files and stored them using iCloud Drive, a cloud service developed by Apple.

The read counts in each TXT file were compiled into three separate data frames in python[28] using pandas (v1.4.4). Each trimester of pregnancy corresponded with one data frame that contained the read counts of each sample that was collected during the said trimester. Each data frame consisted of integer read count data sorted into columns that represented patient IDs and rows that represented a total of 57,736 transcripts. These were then downloaded as TXT files.

Differential expression analysis

Differential expression (DE) analysis was conducted in R (v4.2.2) using edgeR (v3.40.2)[29], a package that specializes in detecting relative changes in expression levels between conditions, and limma (v3.54.0)[30], a package originally developed for RNA-sequencing and DE analysis of microarray data. We conducted nine separate two-way analyses between all the different two-way combinations of the four conditions during each trimester: GDM and healthy pregnancy, preeclampsia and healthy pregnancy, and chronic hypertension and healthy pregnancy. To prepare a DGEList for analysis in edgeR, each manually-created TXT file with read count data frames was loaded into R and the data was stored in three separate DGELists in which the counts were a table of integer read counts and groups corresponded to one of the four mentioned conditions. The data in each trimester was then filtered by counts per million (CPM), so genes were only kept if their CPM was in the top 10% of the CPMs of all listed genes. The data was subsequently normalized and the dispersion was estimated. We then tested for differential expression by conducting tagwise tests and using Benjamini and Hochberg's algorithm[31] to control the false discovery rate (FDR). Results were visualized through volcano plots using the plotSmear function in R. Finally, We generated TXT files of each two-way analysis using estimated count values.

Gene ontology

Differentially expressed genes were investigated using Gene Ontology Enrichment Analysis[10].

Machine learning models

Model formulation

The models created were either binary or multiclass classifiers. Multiclass models involved data from all samples, while each binary model involved only two cases. An alternative model was used where the case was any disease state while the control was healthy pregnancies. Complicated pregnancy samples were characterized by the patient having one of the three diseases discussed; GDM, preeclampsia, and chronic hypertension. In other binary models, the two cases used were healthy pregnancy samples and a specific disease. These models evaluated whether a sample was from a pregnant patient who was healthy or a pregnant patient with a disease. Each model created used read count data from only one trimester to avoid the interference of transcriptional noise.

Machine learning algorithms

All machine-learning analyses were conducted in Python with the pandas and sklearn packages. To test our hypothesis, we created a variety of models that used different methods of data filtering and feature selection. We created nine binary logistic regression models to predict whether a patient had a specific pregnancy complication or was healthy. Three models were created with data from each trimester, and each model only considered genes that were found to be differentially expressed.

Similarly, we created 9 binary lasso regression models to predict whether a patient had a specific pregnancy complication or was healthy. Each lasso model had a test size of 0.25. We also created three additional binary lasso regression models that evaluated whether a patient was healthy or had any of the three pregnancy complications. Lasso regression models did not use any additional methods of feature selection.

We created three binary random forest classifier models and three binary SVM models that evaluated whether a patient was healthy or had any of the three pregnancy complications. The first, second, and third-trimester random forest classifiers had a max depth of five, eight, and eight respectively as well as a min sample split of two. Our SVM models used linear kernels. None of these models used any additional methods of feature selection. We built three binary neural network algorithms (Table 2, Table 3, Table 4) that functioned similarly to the random forest and SVM models. Each binary neural network worked with the differentially expressed genes from that trimester and used L1-based feature selection.

Table 2. Binary neural network architecture used for Trimester 1.

Layer	Node size	Activation function	Param #
Layer 1	8	relu	120
Layer 1	8	relu	72
Output	1	sigmoid	9

Table 3. Binary neural network architecture used for Trimester 2.

Layer	Node size	Activation function	Param #
Layer 1	8	relu	240
Layer 1	8	relu	72
Output	1	sigmoid	9

Table 4. Binary neural network architecture used for Trimester 3.

Layer	Node size	Activation function	Param #
Layer 1	8	relu	176
Layer 1	8	relu	72
Output	1	sigmoid	9

Evaluating the models

Binary models were evaluated with accuracy scores, cross-validation, and area under the curve (AUC) scores. AUC curve plots as well as confusion matrix plots were generated to gain a better understanding of how well each algorithm

was doing. Our multiclass neural networks were evaluated using accuracy scores as well as confusion matrix plots and model accuracy and loss plots. Confusion matrices were plotted using the heatmap function from seaborn[32].

Feature selection and optimization

We used differentially expressed genes and L1-based feature selection to simplify the input for each binary neural network. As the trimester one model was the first one to be built, we first ran this model with no adjustments to the input. We then tried tree-based feature selection, univariate feature selection, L1-based feature selection, and differentially expressed genes to simplify the input. L1-based feature selection and differentially expressed genes allowed the model to perform the best. When the two methods were combined, the model performed better than when these methods of feature selection were applied individually. We ran the trimester two and three models for the first time with no adjustments to the basic model structure. We then ran them using L1-based feature selection and differentially expressed genes and found that they both performed better. We used accuracy scores and observed model loss plots to determine how well the model was doing.

Feature importance

We conducted further research on the lasso regression coefficients that were weighted to be greater than zero. To do this, we researched each gene found by the lasso models that classified whether a pregnancy was complicated or not (i. e. the two cases were healthy or diseased). We also used the GTEx Portal (dbGaP accession number phs000424.vN.pN on 02/24/23) to examine what tissues these genes were largely expressed in and what roles these tissues have in adverse pregnancy outcomes.

Supplementary Materials

Supplementary note

We created three random forest classifier models in total; one from each trimester with each one using the cases of complicated pregnancy (PE, GDM, and chronic hypertension samples) and normal pregnancy. Random forest classifiers were chosen because of their ability to handle large datasets efficiently. They were not combined with any form of feature selection because these models have built-in feature importance methods. We also built three svm models; one from each trimester with each one using the cases of complicated pregnancy (PE, GDM, and chronic hypertension samples) and normal pregnancy. svm models were chosen because they are known to be effective in cases where the dimensions are greater than the number of samples, which applies to the data we used. The goal of creating random forest and svm models was to determine if a different machine-learning approach to this data would be more proficient.

References

"Malsu1 Mitochondrial Assembly of Ribosomal Large Subunit 1 [Homo Sapiens (Human)] - Gene - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, 29 Mar. 2023, https://www.ncbi.nlm.nih.gov/gene/115416.

"PPBP pro-Platelet Basic Protein" *National Center for Biotechnology Information*, U.S. National Library of Medicine, https://www.ncbi.nlm.nih.gov/gene/57349.

"Scara3 Scavenger Receptor Class A Member 3 [Homo Sapiens (Human)] - Gene - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, 29 Mar. 2023, https://www.ncbi.nlm.nih.gov/gene/51435.

A. A. Saleh, S. F. Bottoms, A. M. Farag, M. P. Dombrowski, R. A. Welch, G. Norman, E. F. Mammen. Markers for endothelial injury, clotting and platelet activation in preeclampsia. *Arch Gynecol Obstet.* **251**, 105–110 (1992). https://doi.org/10.1007/BF02718370

A. Ahmed, M. Liang, L. Chi, Y. Q. Zhou, J. G. Sled, M. D. Wilson, P. Delgado-Olguín. Maternal obesity persistently alters cardiac progenitor gene expression and programs adult-onset heart disease susceptibility. *Molecular metabolism.* **43**, 101116 (2021). https://doi.org/10.1016/j.molmet.2020.101116

A. K. Knight, A. L. Dunlop, V. Kilaru, D. Cobb, E. J. Corwin, K. N. Conneely, A.K. Smith. Characterization of gene expression changes over healthy term pregnancies. *PloS one*. **13**, e0204228 (2018). https://doi.org/10.1371/journal.pone.0204228

A. Kulyté, A. Aman, R. J. Strawbridge, P. Arner, I. A. Dahlman. Genome-Wide Association Study Identifies Genetic Loci Associated With Fat Cell Number and Overlap With Genetic Risk Loci for Type 2 Diabetes. *Diabetes*. **71**, 1350–1362 (2022). https://doi.org/10.2337/db21-0804

Basavaraj Vastrad and Chanabasayya Vastrad. Identification of differentially expressed genes and signaling pathways in gestational diabetes mellitus by integrated bioinformatics analysis. *bioRxiv*. (2021). https://doi.org/10.1101/2021.11.24.469869

C. Couture, M. Brien, I. Boufaied, C. Duval, D. D. Soglio, E. A. L. Enninga, B. Cox, S. Girard. Proinflammatory changes in the maternal circulation, maternal–fetal interface, and placental transcriptome in preterm birth. *American Journal of Obstetrics and Gynecology*. **228**, 332.e1-332.e17 (2023). https://doi.org/10.1016/j.ajog.2022.08.035

C. Gebhardt, J. Németh, P. Angel, J. Hess. S100A8 and S100A9 in inflammation and cancer. *Biochemical pharmacology*. **72**, 1622–1631 (2006). https://doi.org/10.1016/j.bcp.2006.05.017

C. W. Ives, R. Sinkey, I. Rajapreyar, A. T. N. Tita, S. Oparil. Preeclampsia—Pathophysiology and Clinical Presentations: JACC State-of-the-Art Review. *Journal of the American College of Cardiology*. **76**, 1690-1702 (2020). https://doi.org/10.1016/j.jacc.2020.08.014

C. Zhao, E. Lu, X. Hu, H. Cheng, J. Zhang, X. Zhu. S100A9 regulates cisplatin chemosensitivity of squamous cervical cancer cells and related mechanism. *Cancer Management and Research*. **10**, 3753-3764 (2018). DOI: 10.2147/CMAR.S168276

D. A. Enquobahrie, M. A. Williams, C. Qiu, D. S. Siscovick, T. K. Sorensen. Global maternal early pregnancy peripheral blood mRNA and miRNA expression profiles according to plasma 25-hydroxyvitamin D concentrations. *The journal of maternal-fetal & neonatal medicine: the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians*. **24**, 1002–1012 (2011). https://doi.org/10.3109/14767058.2010.538454

D. Kaudewitz, P. Skroblin, L. H. Bender, T. Barwari, P. Willeit, R. Pechlaner, N. P. Sunderland, K. Willeit, A. C. Morton, P. C. Armstrong, M. V. Chan, R. Lu, X. Yin, F. Gracio, K. Dudek, S. R. Langley, A. Zampetaki, E. D. Rinaldis, S. Ye, T. D. Warner, A. Saxena, S. Kiechl, R. F. Storey, M. Mayr. Association of MicroRNAs and YRNAs With Platelet Function. *AHA Journals*. **118** (2015). https://doi.org/10.1161/CIRCRESAHA.114.305663

E. Declercq, L. Zephyrin. Severe Maternal Morbidity in the United States: A Primer. *Commonwealth Fund.* (2021). https://www.commonwealthfund.org/publications/issue-briefs/2021/oct/severe-maternal-morbidity-united-states-primer.

E. Jurewicz, A. Filipek. Ca2+- binding proteins of the S100 family in preeclampsia. *Placenta*. **127**, 43-51 (2022). https://doi.org/10.1016/j.placenta.2022.07.018

E. Pandur, I. Szabó, E. Hormay, R. Pap, A. Almási, K. Sipos, V. Farkas, Z. Karádi. Alterations of the expression levels of glucose, inflammation, and iron metabolism related miRNAs and their target genes in the

hypothalamus of STZ-induced rat diabetes model. *Diabetol Metab Syndr*. **14**, 147 (2022). https://doi.org/10.1186/s13098-022-00919-5

G. D. Vecchio, Q. Li, W. Li, S. Thamotharan, A. Tosevska, M. Morselli, K. Sung, C. Janzen, X. Zhou, M. Pellegrini, S. U. Devaskar. Cell-free DNA Methylation and Transcriptomic Signature Prediction of Pregnancies with Adverse Outcomes. *Epigenetics*. **16**, 642–661(2021). https://doi.org/10.1080/15592294.2020.1816774

G. Fan, Y. Gu, J. Zhang, Y. Xin, J. Shao, F. Giampieri, M. Battino. Transthyretin Upregulates Long Non-Coding RNA MEG3 by Affecting PABPC1 in Diabetic Retinopathy. *International Journal of Molecular Sciences*. **20**, 6313 (2019). https://doi.org/10.3390/ijms20246313

H. Chen, I. Aneman, V. Nikolic, N. K. Orlic, Z. Mikovic, M Stefanovic, Z. Cakic, H. Jovanovic, S. E. L. Town, M. P. Padula, L. McClements. Maternal plasma proteome profiling of biomarkers and pathogenic mechanisms of early-onset and late-onset preeclampsia. *Scientific reports*. **12**, 19099 (2022). https://doi.org/10.1038/s41598-022-20658-x

H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, P. D. Thomas. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. **47**, D419-D426 (2019). https://doi.org/10.1093/nar/gky1038

H. S. Gammill, R. Chettier, A. Brewer, J. M. Roberts, R. Shree, E. Tsigas, K. Ward. Cardiomyopathy and Preeclampsia. *Circulation*. **138**, 2359–2366 (2018). https://doi.org/10.1161/CIRCULATIONAHA.117.031527

I. A. Lian, M. Langaas, E. Moses, A. Johansson. Differential gene expression at the maternal-fetal interface in preeclampsia is influenced by gestational age. *PloS one*. **8**, e69848 (2013). https://doi.org/10.1371/journal.pone.0069848

J. Camunas-Soler, E. P. S. Gee, M. Reddy, J. D. Mi, M. Thao, T. Brundage, F. Siddiqui, N. L. Hezelgrave, A. H. Shennan, E. Namsaraev, C. Haverty, M. Jain, M. A. Elovitz, M. Rasmussen, R. M. Tribe. Predictive RNA profiles for early and very early spontaneous preterm birth. *American Journal of Obstetrectrics and Gynecology*. **227**, (2022). https://doi.org/10.1016/j.ajog.2022.04.002

J. Jin, C. Zhu, J. Wang, X. Zhao, R. Yang. The association between ACTB methylation in peripheral blood and coronary heart disease in a case-control study. *Frontiers in Cardiovascular Medicine*. **9**, 972566 (2022). https://doi.org/10.3389/fcvm.2022.972566

K. H. Tan, S. S. Tan, S. K. Sze, W. K. R. Lee, M. J. Ng, S. K. Lim. Plasma biomarker discovery in preeclampsia using a novel differential isolation technology for circulating extracellular vesicles. *American Journal of Obstetrics and Gynecology*. **211**, 380.e1-380.e13 2014. https://doi.org/10.1016/j.ajog.2014.03.038.

K. Kristensen, D. Wide-Swensson, C. Schmidt , S. Blirup-Jensen , V. Lindstro, H. Strevens, A. Grubb. Cystatin C, β-2-Microglobulin and β-Trace Protein in Pre-Eclampsia. *Acta Obstetricia et Gynecologica*. **86**, 921926 (2007). https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1080/00016340701318133.

K. Murata, Y. Miyamura, N. Toyoda, Y. Ikeda, Y. Kozuka, Y. Sugiyama. Nihon Sanka Fujinka Gakkai zasshi. **33**, 1669–1674 (1981).

L. A. Corchete, E. A. Rojas, D. Alonso-López, J. D. L. Rivas, N. C. Gutiérrez, F. J. Burguillo. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci Rep.* **10**, 19737 (2020). https://doi.org/10.1038/s41598-020-76881-x

L. Baxi, E. A. Reece, D. Barad, R. Farber, A. Williams. Glycosylated hemoglobin (HbA_1) and hemoglobinopathies in pregnancy. *De Gruyter*. **12**, 133-136 (1984). https://doi.org/10.1515/jpme.1984.12.3.133

L. Y. Liu, T. Yang, J. Ji, Q. Wen, A. A. Morgan, B. Jin, G. Chen, D. J. Lyell, D. K. Stevenson, X. B. Ling, A. J. Butte. Integrating multiple 'omics' analyses identifies serological protein biomarkers for preeclampsia. *BMC medicine*. **11**, 236 (2013). https://doi.org/10.1186/1741-7015-11-236

M. Alcaide, M. Cheung, J. Hillman, S. R. Rassekh, R. J. Deyell, G. Batist, A. Karsan, A. W. Wyatt, N. Johnson, D. W. Scott, R. D. Morin. Evaluating the quantity, quality and size distribution of cell-free DNA by multiplex droplet digital PCR. *Sci Rep*. **10**, 12564 (2020). https://doi.org/10.1038/s41598-020-69432-x

M. D. Pertile, M. Halks-Miller, N. Flowers, C. Barbacioru, S. L. Kinnings, D. Vavrek, W. K. Seltzer, D. W. Bianchi. Rare autosomal trisomies, revealed by maternal plasma DNA sequencing, suggest increased risk of feto-placental disease. *Science translational medicine*. **9**, eaan1240 (2017). https://doi.org/10.1126/scitranslmed.aan1240

M. D. Robinson, D. J. McCarthy, G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*. **26**, 139-140 (2010). doi:10.1093/bioinformatics/btp616.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*. **43**, e47 (2015). https://doi.org/10.1093/nar/gkv007

M. K. Farag, N. A. E. Maksoud, H. M. Ragab, K. R. Gaber. Predictive value of cystatin C and beta-2 microglobulin in preeclampsia. *Journal of Genetic Engineering and Biotechnology*. **9**, 133-136 (2011). https://doi.org/10.1016/j.jgeb.2011.09.003

M. N. Moufarrej, R. J. Wong, G. M.Shaw, D. K. Stevenson, S. R. Quake. Investigating Pregnancy and Its Complications Using Circulating Cell-Free RNA in Women's Blood During Gestation. *Frontiers in pediatrics*. **8**, 605219 (2020). https://doi.org/10.3389/fped.2020.605219

M. V. Dijk, C. B. Oudejans. STOX1: Key player in trophoblast dysfunction underlying early onset preeclampsia with growth retardation. *Journal of pregnancy*. **2011**, 521826 (2011). https://doi.org/10.1155/2011/521826.

M.N. Moufarrej, S.K. Vorperian, R.J. Wong, A. A. Campos, C. C. Quaintance, R. V. Sit, M. Tan, A. M. Detweiler, H. Mekonen, N. F. Neff, C. Baruch-Gravett, J. A. Litch, M. L. Druzin, V. D. Winn, G. M. Shaw, D. K. Stevenson, S. R. Quake. Early prediction of preeclampsia in pregnancy with cell-free RNA. *Nature*. **602**, 689–694 (2022). https://doi.org/10.1038/s41586-022-04410-z

N. Yang, Q. Wang, B. Ding, Y. Gong, Y. Wu, J. Sun, X. Wang, L. Liu, F. Zhang, D. Du, X. Li. Expression profiles and functions of ferroptosis-related genes in the placental tissue samples of early- and late-onset preeclampsia patients. *BMC Pregnancy Childbirth*. **22**, 87 (2022). https://doi.org/10.1186/s12884-022-04423-6

P. A. Cavazos-Rehg, M. J. Krauss, E. L. Spitznagel, K. Bommarito, T. Madden, M. A. Olsen, H. Subramaniam, J. F. Peipert, L. J. Bierut. Maternal age and risk of labor and delivery complications. *Matern Child Health J*. **19**, 1202-1211 (2015). https://doi.org/10.1007/s10995-014-1624-7

P. Csaicsich, J. Deutinger, G. Tatra. Platelet-specific proteins (beta-thromboglobulin and platelet factor 4) in normal pregnancy and in pregnancy complicated by preeclampsia. *Archives of gynecology and obstetrics*. **244**, 91–95 (1989). https://doi.org/10.1007/BF00931379

R. Boufermes, D. Haddad. Correlation between the Diabetic Marker (Hba1c) and the Anemia Marker (Hba2) In Type 2 Diabetes. *Journal of Geriatric Research*. **4**, (2020). https://www.imedpub.com/abstract/correlation-between-the-diabetic-marker-hba1c-and-the-anemia-marker-hba2-in-type-2-diabetes-29532.html.

R. Navajas, F. Corrales, A. Paradela. Quantitative proteomics-based analyses performed on pre-eclampsia samples in the 2004–2020 period: a systematic review. *Clinical Proteomics*. **18** (2021). https://doi.org/10.1186/s12014-021-09313-1.

R. Verma, P. Verma, S. Budhwar, K. Singh. S100 proteins: An emerging cynosure in pregnancy & adverse reproductive outcome. *The Indian journal of medical research*. **148(Suppl)**, S100–S106 (2018). https://doi.org/10.4103/ijmr.IJMR_494_18

R.R. Nair, A. Khanna, K. Singh. Role of inflammatory proteins S100A8 and S100A9 in pathophysiology of recurrent early pregnancy loss. *Placenta*. **34**, 824-827 (2013). https://doi.org/10.1016/j.placenta.2013.06.307.

S. A. Abdelazim, O. G. Shaker, Y. A. H. Aly, M. A. Senousy. Uncovering serum placental-related non-coding RNAs as possible biomarkers of preeclampsia risk, onset and severity revealed MALAT-1, miR-363 and miR-17. *Sci Rep*. **12**, 1249 (2022). https://doi.org/10.1038/s41598-022-05119-9

S. Wei, D. Wang, H. Li, L. Bi, J. Deng, G. Zhu, J. Zhang, C. Li, Min Li, Y. Fang, G. Zhang, J. Chen, S. Tao, X. Zhang. Fatty acylCoA synthetase FadD13 regulates proinflammatory cytokine secretion dependent on the NF-

κB signalling pathway by binding to eEF1A1. *Cellular Microbiology*. **21**, e13090 (2019). https://doi.org/10.1111/cmi.13090

Siobán B. Keel, Susan Phelps, Kathleen M. Sabo, Monique N. O'Leary, Catherine B. Kirn-Safran, Janis L. Abkowitz. Establishing Rps6 hemizygous mice as a model for studying how ribosomal protein haploinsufficiency impairs erythropoiesis. *Experimental Hematology*. **40**, 290-294 (2011). https://doi.org/10.1016/j.exphem.2011.12.003

T. Lekva, R. Lyle, M. C. P. Roland, C. Friis, D. W. Bianchi, I. Z. Jaffe, E. R. Norwitz, J. Bollerslev, T. Henriksen, T. Ueland. Gene expression in term placentas is regulated more by spinal or epidural anesthesia than by late-onset preeclampsia or gestational diabetes mellitus. *Sci Rep*. **6**, 29715 (2016). https://doi.org/10.1038/srep29715

T. Meng, H. Chen, M. Sun, H. Wang, G. Zhao, X. Wang. Identification of differential gene expression profiles in placentas from preeclamptic pregnancies versus normal pregnancies by DNA microarrays. *Omics: a journal of integrative biology*. **16**, 301–311 (2012).. https://doi.org/10.1089/omi.2011.0066

V. Alur, V. Raju, B. Vastrad, A. Tengli, C. Vastrad, S. Kottursheti. Integrated bioinformatics analysis reveals novel key biomarkers and potential candidate small molecule drugs in gestational diabetes mellitus. *Bioscience reports*. **41**, BSR20210617 (2021). https://doi.org/10.1042/BSR20210617

V. Alur, V. Raju, B. Vastrad, C. Vastrad, S. Kottursheti. Analysis of key genes and pathways associated with the pathogenesis of Type 2 diabetes mellitus. *bioRxiv*. **12**, 456106 (2021). https://doi.org/10.1101/2021.08.12.456106

Van Rossum, G. & Drake, F.L., 2009. *Python 3 Reference Manual*, Scotts Valley, CA: CreateSpace.

W. Siwen, S. Rui, W. Ziyi, J. Zhaocheng, W. Shaoxiong, M. Jian. S100A8/A9 in Inflammation. *Frontiers in Immunology*. **9**, 1664-3224 (2018). https://doi.org/10.3389/fimmu.2018.01298

Waskom, Botvinnik, Olga, Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, Qalieh, Adel. (2017). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. https://doi.org/10.5281/zenodo.883859

X. Li, F. Cheng, G. Cao. Expression of S100 calcium-binding protein A8 in peripheral blood of patients with preeclampsia during pregnancy. *European Journal of Inflammation*. **17** (2019). https://doi.org/10.1177/2058739219858527

X. Yang, T. Meng. Long Noncoding RNA in Preeclampsia: Transcriptional Noise or Innovative Indicators? *BioMed research international*. **2019**, 5437621 (2019). https://doi.org/10.1155/2019/5437621

X. Yang, Y. Ding, L. Sun, M. Shi, P. Zhang, Z. Huang, J. Wang, A. He, J. Wang, J. Wei, M. Liu, J. Liu, G. Wang, X. Yang, R. Li. Ferritin light chain deficiency-induced ferroptosis is involved in preeclampsia pathophysiology by disturbing uterine spiral artery remodelling. *Redox Biology*. **58**, 2213-2317 (2022). https://doi.org/10.1016/j.redox.2022.102555.

Y. Benjamini, Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*. **57**, 289-300 (1995). https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Y. Huang, Z. Li, E. Lin, H. Pei, R. Gaizhen. Oxidative damage-induced hyperactive ribosome biogenesis participates in tumorigenesis of offspring by cross-interacting with the Wnt and TGF-β1 pathways in IVF embryos. *Experimental & Molecular Medicine*. **53**, 1792–1806 (2021). https://doi.org/10.1038/s12276-021-00700-0

Y. I. Elshimali, H. Khaddour, M. Sarkissyan, Y. Wu, J. V. Vadgama. The clinical utilization of circulating cell-free DNA (CCFDNA) in blood of cancer patients. *International journal of molecular sciences*. **14**, 18925–18958 (2013). https://doi.org/10.3390/ijms140918925

Y. Lu, Y. Li, G. Li, H. Lu Identification of potential markers for type 2 diabetes mellitus via bioinformatics analysis. *Molecular Medicine Reports*. **22**, 1868-1882 (2020). https://doi.org/10.3892/mmr.2020.11281

Y. Wen, M. Cheng, L. Qin, W. Xu. TNFα-induced abnormal activation of TNFR/NF-κB/FTH1 in endometrium is involved in the pathogenesis of early spontaneous abortion. *Journal of cellular and molecular medicine*. **26**, 2947–2958 (2022). https://doi.org/10.1111/jcmm.17308

Z. Chen, J. Gan, M. Zhang, Y. Du, H. Zhao. Ferroptosis and Its Emerging Role in Pre-Eclampsia. *Antioxidants (Basel, Switzerland)*. **11**, 1282 (2022).. https://doi.org/10.3390/antiox11071282

Z. Li, X. Pan, Y. D. Cai. Identification of Type 2 Diabetes Biomarkers From Mixed Single-Cell Sequencing Data With Feature Selection Methods. *Frontiers in bioengineering and biotechnology*. **10**, 890901 (2022). https://doi.org/10.3389/fbioe.2022.890901

Z. Masoumi, G. E. Maes, K. Herten, Á. Cortés-Calabuig, A. G. Alattar, E. Hanson, L. Erlandsson, E. Mezey, M. Magnusson, J. R. Vermeesch, M. Familari, S. R. Hansson. Preeclampsia is Associated with Sex-Specific Transcriptional and Proteomic Changes in Fetal Erythroid Cells. *International journal of molecular sciences*. **20**, 2038 (2019). https://doi.org/10.3390/ijms20082038

Z. Ou, Q. Li, W. Liu, X. Sun. Elevated Hemoglobin A2 as a Marker for β-Thalassemia Trait in Pregnant Women. *The Tohoku Journal of Experimental Medicine*. **223**, 223-226 (2011). https://doi.org/10.1620/tjem.223.223